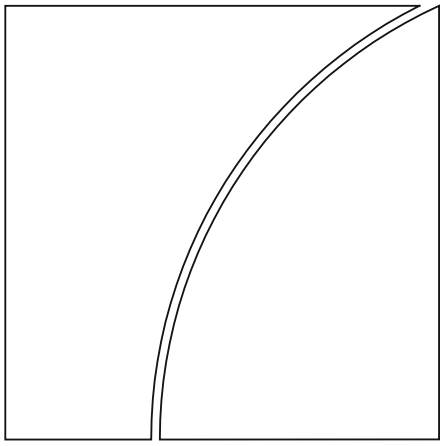
Irving Fisher Committee on Central Bank Statistics

BIS



IFC Bulletin

No 64

Data science in central banking: enhancing the access to and sharing of data

May 2025

Contributions in this volume were prepared for the proceedings of the workshop on "Data science in central banking: enhancing the access to and sharing of data" co-organised by the Bank of Italy and the Irving Fisher Committee (IFC) in Rome, Italy, 17-19 October 2023.

The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the Bank of Italy, the IFC and its members and the other institutions represented at the meeting.

This publication is available on the BIS website (www.bis.org).

© *Bank for International Settlements 2025. All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1991-7511 (online)

ISBN 978-92-9259-850-1 (online)

Data science in central banking: enhancing the access to and sharing of data

IFC Bulletin No 64

May 2025

Proceedings of the workshop on "Data science in central banking: enhancing the access to and sharing of data" co-organised by the Bank of Italy and the Irving Fisher Committee (IFC) in Rome, Italy, 17-19 October 2023.

Overview

Data science in central banking: unlocking the potential of data

Douglas Araujo, Economist, Monetary and Economic Department (MED), Bank for International Settlements (BIS)

Adam Cap, Senior Macroeconomic Analyst, MED, BIS

Ilaria Mattei, Senior Financial Market Analyst, MED, BIS

Rafael Schmidt, Head of IT, MED, BIS

Olivier Sirello, Statistical Analyst, MED, BIS

Bruno Tissot, Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat

Opening remarks

Bruno Tissot, Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat

Keynote speeches

Project Aurora: the power of data, technology and collaboration to combat money laundering across institutions and borders

Beju Shah, BIS

Why SDMX matters? A community journey towards SDMX as an AI-enabler and a data mesh enabler

Eric Anvar, Organisation for Economic Co-operation and Development (OECD)

1. Data science transforming central banks

Data science in the context of international banking statistics

Jacob Ewertzh and John Svanäng, Sveriges Riksbank

Experiences, essentials and perspectives for data science in the hearts of central banks and supervisors: a case study of the Dutch central bank

Patty Duijm and Iman van Lelyveld, De Nederlandsche Bank

Digital transformation of financial regulators and the emergence of supervisory technologies (SupTech): a case study of the UK Financial Conduct Authority

Pavle Avramović, UK Financial Conduct Authority

2. Making data open

Building a database on cryptocurrencies

Anamaria Illes and Ilaria Mattei, BIS

European single access point as a blueprint for global financial and green data hubs

Pawel Martyniuk and Michal Piechocki, Business Reporting - Advisory Group (BR-AG)

Data sharing using a global data registry: on a place to discover global structured time series, macro and micro data

Matt Nelson and Glenn Tice, BIS

3. Advanced analytical tools for data-sharing

Overcoming data-sharing challenges in central banking: federated learning of diffusion models for synthetic mixed-type tabular data generation

Timur Sattarov, Deutsche Bundesbank, and Marco Schreyer, University of St Gallen

Exploring aggregation strategies for federated learning in national statistics

Mauro Bruno, Erika Cerasti, Massimo De Cubellis and Francesco Pugliese, Italian National Institute of Statistics (ISTAT), Rafik Chemli, Benjamin Santos and Julian Templeton, Statistics Canada, and Matjaz Jug, Statistics Netherlands

4. Accessing big data

EMIR data for financial stability analysis and research

Michele Leonardo Bianchi, Bianca Sorvillo, Dario Ruzzi, Federico Apicella, Luigi Abate and Leonardo Del Vecchio, Bank of Italy

Unveiling the interconnectedness of banks in payment system: methodology, utilization, and data governance considerations

Renardi Ardiya Bimantoro, Irfan Sampe and Mohammad Khoyrul Hidayat, Bank Indonesia

Big data platform (FinPulse) initiative

Tsenddorj Dorjpurev, California Polytechnical State University

5. Data access arrangements

New strategy of data sharing and data access in statistics: the view from Banco de Portugal

Ana R Gonçalves, Mário Lourenço, Daniel V Sousa and Thomas Verheij, Banco de Portugal

Individual data access and sharing protection policy: definition and case study of Bank Indonesia

Johanes Iman Anugrah, Akhmad Zacky Nugraha and Sapto Widyatmiko, Bank Indonesia

6. Open source as avenues for collaboration and sharing

Open-sourced central bank macroeconomic models

Douglas Araujo, BIS

Leveraging open-source software and data standards as the backbone of your open data strategy

Darran Hodder and Matthew Nelson, Regnology

Collaborating on SDMX APIs and open-source software

Brian Buffet, Stratos Nikoloutsos and Xavier Sosnovsky, BIS

7. Time series analysis

Future of time series: preliminary results from a BIS-IFC survey of central banks and statistical agencies

Ryland Thomas, Bank of England

Coding time series with machine learning

Ayoub Mharzi, International Monetary Fund (IMF)

8. Extracting information with machine learning tools: anomaly and error detection

Siamese neural networks for detecting banknote printing defects

Katia Boria, Andrea Luciani, Sabina Marchetti and Marco Viticoli, Bank of Italy

Error spotting with gradient boosting: a machine learning-based application for central bank data quality

Csaba Burger, Magyar Nemzeti Bank, and Mihály Berndt, Clarity Consulting

9. Machine learning powered data access, tools and applications

A machine learning approach for the detection of firms infiltrated by organised crime in Italy

Pasquale Cariello, Marco De Simoni and Stefano Iezzi, Bank of Italy

Research for all: exploring machine learning applications in generating synthetic datasets

Carmelita Esclanda-Lo, Gabriel Masangkay, Chelsea Anne Ong and Rossvern Reyes, Bangko Sentral ng Pilipinas

Invoices rather than surveys: using ML to build nominal and real indices

Pablo Acevedo, Dagoberto Quevedo and Marco Rojas, Central Bank of Chile, Emiliano Luttini, World Bank, and Matías Pizarro, Pontificia Universidad Católica - Chile

10. New insights from natural language processing

Leveraging large language models to extract data citations

Sebastian Seltmann, Emily Kormanyos and Hendrik Christian Doll, Deutsche Bundesbank

Let's talk about sentiment: natural language processing using machine learning on bank earnings transcripts

Seung Jung Lee, Sriram Nagaraj, Dylan Saez, Victors Stebunovs, Cindy Vojtech and Karl Wirth, Board of Governors of the Federal Reserve System

Constructing high-frequency and thematic economic sentiment indicators from online news articles: applications in the Philippine context

Alan Chester Arcin, Carmelita Esclanda-Lo, Chelsea Anne Ong and Rossvern Reyes, Bangko Sentral ng Pilipinas

11. Natural language processing on central bank communication

Do anecdotes matter? Exploring the beige book through textual analysis from 1970 to 2023

Shengwu Du, Karen Guo, Flora Haberkorn, Abby Kessler, Isabel Kitschelt, Seung Jung Lee, Anderson Monken, Dylan Saez, Kelsey Shipman and Sandeep Thakur, Board of Governors of the Federal Reserve System

Central bank communication on economic activity

Sercan Eraslan and Eniko Gabor-Toth, Deutsche Bundesbank

From the ML model to practice: case study on NLP-based decision-making on the eligibility of security prospectuses

Janek Blankenburg, Maximilian König, Philipp Rothhaar and Bernd Rusitschka, Deutsche Bundesbank

Closing remarks

Alessandra Perrazzelli, Deputy Governor, Bank of Italy

Irving Fisher Committee on Central Bank Statistics

BIS

Data science in central banking: unlocking the potential of data

Douglas Araujo, Adam Cap, Ilaria Mattei, Rafael Schmidt, Olivier Sirello, Bruno Tissot[1]

Executive summary

Since 2021, the Irving Fisher Committee on Central Bank Statistics (IFC) has established a series of workshops on "Data science in central banking", gathering statisticians and economists from central banks, international organisations and national statistical offices (NSOs) in addition to other stakeholders. The aim has been to review the adoption of data science in the central banking community and beyond, identify the associated opportunities and challenges as well as exchange best practices. The third edition focused on **how data science can effectively unlock the potential and value of data** by fostering their use, reuse, access and sharing.

Against this setting, **four main takeaways** stand out.

First, innovative techniques can play a central role in **leveraging traditional as well as emerging data sources and types**, with benefits for both producers and users of economic and financial information in central banks. For producers, they offer various tools to streamline statistical processes and improve data quality, especially in terms of accuracy, frequency, timeliness and granularity. For users, they enable the analysis of large, complex and/or high-dimensional data sets to extract meaningful insights or detect new patterns.

Second, data science can support **adequate and secure sharing of data sets** – particularly granular ones – without disclosing sensitive information, for instance by using tools such as privacy-enhancing technologies.

Third, **challenges persist when tapping into the increasing and various amounts of information available** today. A prominent one is the need for robust yet costly information technology (IT) resources, such as those required to handle and share large data sets. Moreover, organisational barriers can prevent the successful integration of new data techniques into existing processes, for example, because of "silos" or a limited coordination between subject matter experts, IT and data scientists. Efficient access to and use of data may also be constrained by still limited standardisation. This is particularly the case for alternative or secondary data sources,

[1] Respectively, Adviser on data science in economics (douglas.araujo@bis.org), Senior Macroeconomic Analyst (adam.cap@bis.org), Senior Financial Market Analyst (ilaria.mattei@bis.org), Head of IT, Monetary and Economic Department (rafael.schmidt@bis.org), Statistical Analyst (olivier.sirello@bis.org), Head of BIS Statistics and Head of the IFC Secretariat (bruno.tissot@bis.org).

The views expressed here are those of the authors and do not necessarily reflect those of the BIS, the IFC or any of those institutions represented at the conference.

We thank Bilyana Bogdanova, Magdalena Erdem, Mário Lourenço and Luís Teles Dias for helpful comments and suggestions.

which often lack common definitions, methodological consistency and alignment with well established standards, in turn limiting their use for compiling official statistics.

The fourth takeaway is that, fortunately, **central banks** as both users and producers of statistics **are well equipped to address these challenges**, notably thanks to their long-standing experience in big data analytics, information standards and, more broadly, data management and governance.

Looking ahead, facilitating effective data access and sharing to make the most of the information available to central banks may call for further progress in the following key areas:

- **Data management and governance** to ensure the quality of statistical information, including its availability, transparency and usability.
- **Interoperability capabilities and statistical standards** to enable effective and accurate use of data, for instance by fostering their integration, findability, access and sharing for reuse, especially by data scientists.
- **Modern, metadata-driven and easily accessible (big) data platforms** within organisations complemented by single access points for the public and data spaces more generally.
- **Strengthened collaboration** among peers and counterparts to advance data science projects, by developing IT solutions for secure and adequate data access and sharing as well as by promoting a structured exchange of experiences across interested stakeholders.

1. Introduction

The ongoing data revolution offers central banks many opportunities to inform their policy decisions with sound economic and financial analyses. As producers of official statistics, they can leverage the increasing and diverse supply of information sources to support their statistical function and fill gaps (De Beer and Tissot (2021); IFC (2023a)). Special attention has been placed on secondary data – that is, information which is not primarily collected for statistical purposes, such as geospatial data, private records or administrative registers (MacFeely (2020)). As data users, central banks can benefit from more real-time, detailed and multidimensional insights to support their mandates (IFC (2024)). To this end, they have been working on better integrating micro data into macroeconomic aggregates which they have traditionally focused on (IFC (2021a)).

Yet **harnessing this wealth of information effectively can be challenging**. Large and/or complex data sets tend to require additional processing before being able to be used for statistical purposes, reflecting their inherently limited quality and poor standardisation, as noted by the Bank of Italy's Deputy Governor Alessandra Perrazzelli. In addition, making insights available to a broader range of users, including researchers, may encounter important operational, technical, legal and ethical barriers – such as confidentiality restrictions, commercial limitations, resource shortages and IT requirements (IFC (2023b)).

Fortunately, **innovation can help overcome many of these constraints**. In particular, data science – which combines statistical, mathematical, IT and subject matter expertise – provides a wide range of tools such as big data analytics, machine learning (ML) and artificial intelligence (AI). These can be instrumental in tapping into large or complex data – such as unstructured information containing text, images, audio and video (IFC (2023c)). Innovative techniques can for instance improve data quality, extract useful insights and securely share them. Recent central bank experience also shows that these tools **can help access and share effectively the information available** to support their statistical function as well as their policy analyses.

However, while data science can be a powerful enabler, **maximising data use calls for strengthening cooperation**, knowledge-sharing and the exchange of best practices, as emphasised by the Head of BIS Statistics Bruno Tissot. Such efforts can take place at various levels, including among central banks, within national statistical systems and internationally. From an operational perspective, they may also require further improvements to the existing global statistical infrastructure, especially by developing common standards, unique identifiers and proper data management frameworks.

The overview of this *IFC Bulletin* draws on the various contributions presented in the third edition of the "Data science in central banking" IFC-Bank of Italy workshop and focuses on **how data science approaches can effectively unlock the potential of data by fostering their use, reuse, access and sharing**. Section 2 discusses the related benefits for central banks in their dual role as data producers and users. Section 3 turns to the key obstacles faced in this endeavour, especially regarding emerging data types and secondary sources. Section 4 concludes with a number of key priorities moving forward, highlighting the importance of international collaboration.

2. Accessing and making data available for further analysis

Data science can enhance the use of existing as well as novel data sources, with benefits for both producers and users of economic and financial information in central banks. For producers, it offers various tools to improve overall data quality, especially in terms of accuracy, frequency, timeliness and granularity, while also streamlining statistical processes. For users, it enables the analysis of large, complex and/or high-dimensional data sets, for instance to extract meaningful insights or detect new patterns. Finally, it supports both producers and users in adequately and safely sharing information, such as through privacy-enhancing technologies.

Improving data quality, especially in terms of accuracy, timeliness and processing

A prominent area where data science and ML can be instrumental relates to **data quality**[2] and accuracy specifically, with common applications dealing with **anomaly and error detection**. For example, the Bank of Italy employs Siamese neural networks to identify manufacturing defects and automate quality controls for the production of banknotes. Another example is from the Magyar Nemzeti Bank and Clarity Consulting, which follow a gradient-boosting ML approach to optimise error identification modelling that can be applied to both aggregate and granular data sets. In a similar vein, other work at the Bank of Italy shows a further promising use case of gradient boosting to support anti-money laundering and prudential supervision, by uncovering patterns in company financial statements that can suggest association with criminal activities.

ML can also enable the production of **more granular and timely statistics while also maintaining high accuracy**, for instance by leveraging sources such as big financial data sets or web-based indicators collected in real time. The Central Bank of Chile uses ML to analyse daily transactions from millions of electronic invoices based on their text descriptions and derive more timely consumption and investment metrics. More broadly, ML algorithms can significantly improve the speed and accuracy for processing large amounts of granular information (Liu et al (2008)), in turn facilitating the expansion of central banks' statistical offerings (Brault et al (2024)).

Furthermore, the advanced capabilities provided by **innovative tools can significantly foster the automation of statistical processes**. One prominent area is data **collection**, particularly to extract and structure information from various sources and formats, such as webpages, geodata or textual documents. For instance, the Deutsche Bundesbank uses large language models (LLMs) to extract citations and assess the associated impact of academic papers. Another project showcases how ML can facilitate the identification of financial securities eligible for specific policy purposes, by classifying the textual information extracted from their prospectus. More generally, innovative techniques can streamline processes across the entire data life cycle. For instance, the Sveriges Riksbank is developing a set of automated tools to test, integrate, deploy and orchestrate software to support the compilation of BIS international banking statistics. This has in particular helped automate data **aggregation** and **evaluation**, with consequent benefits in terms of operational continuity and efficiency. Similarly, the UK Financial Conduct Authority has been leveraging innovative supervisory technologies (SupTech) to support data **integration**, for instance by linking multiple sources to detect market abuse.

[2] "Quality" is a multifaceted concept spanning the various characteristics sought for in official statistics and captured by the Fundamental Principles of Official Statistics (UN (2014)). These include accuracy, integrity, security, transparency and trustworthiness. The concept also covers the user side, for instance to ensure that the data are fit for purpose, easily accessible, findable, traceable and reusable, and that they can eventually be deleted adequately. Lastly, it includes ethical aspects, such as assuring impartiality, objectivity, professional independence and social acceptability (ie information collected is not misused) to secure trust in official statistics (IFC (2021b)).

Dealing with large, multidimensional and novel data for analysis

Data science techniques can facilitate the analysis of large and multidimensional data sets. Clustering, neural networks or network analysis can identify uncovered patterns among many variables – a task that traditional statistical methods may struggle with (Mittal et al (2019)). A practical application is demonstrated by the Bank of Italy, which builds network graphs based on a centrality measure to show interconnections between participants in the derivatives market using granular data collected under the European Market Infrastructure Regulation (EMIR).[3] Bank Indonesia presents a similar use case deploying big data analytics and network analysis – particularly the core periphery model – to visualise interbank relationships using granular payments.

Data science can also help **tap into unstructured data such as text through natural language processing** (NLP) and LLMs. Thanks to their ability to capture relationships between words, these tools can assist users in analysing financial reports, news articles and social media (Araujo et al (2024a)).[4] A widely used application in central banking is **sentiment analysis**, a set of methods to quantify qualitative components of textual data to gain better, real-time insights into market mood, economic expectations or public perceptions of policy actions. For example, the Bangko Sentral ng Pilipinas uses FinBERT (Araci (2009)) – an encoder-based model for sentence identification finetuned specifically for financial texts – to construct sentiment indices, effectively complementing and predicting forward-looking indicators such as purchasing managers' indices. Similarly, the Board of Governors of the Federal Reserve System combines tree-structured topic modelling with BERT-based sentiment analysis to analyse developments in equity and credit default swap markets depending on the specific content released during banks' earnings calls. This research underscores the importance of sentiment analysis in understanding market reactions and the broader financial implications of communication by financial institutions.

Making use of textual data is particularly relevant for central banks for at least two reasons. First, they are themselves a key source of rich qualitative statements, for instance through various monetary policy decisions, reports, speeches and press releases. One important example is the BIS database comprising most of the speeches delivered by Governors and other central bank officials worldwide.[5] Second, textual information can be useful for assessing the impact of central bank communication on public expectations and, in turn, for forecasting economic indicators. For instance, the Deutsche Bundesbank has measured the influence of the central bank's narrative on underlying economic conditions and associated risks. Such novel techniques can also improve the forecast accuracy of a number of macroeconomic indicators, at least to some degree. Recent research suggests for

[3] New data techniques may facilitate the use of derivatives data sets, which can be challenging to analyse due to their high granularity and large volume (IFC (2018a)). For additional details on EMIR, see ec.europa.eu.

[4] For additional information on how LLMs can support the compilation of official statistics, see IFC (2025a) and UNECE (2023a).

[5] The database can be accessed at bis.org and through the open source python library *gingado* (Araujo (2023)).

instance that the analysis of the impact of European Central Bank's (ECB) monetary policy statements can improve euro area forecasts (Araujo et al (2024b)). Lastly, **nowcasting** applications often rely on the integration of data from various sources and types to "forecast" the present state of the economy (IFC (2021c)). One example is shown by the Board of Governors of the Federal Reserve System that combines the use of "standard" macroeconomic indicators with more qualitative information – especially the regular commentary in the Beige Book from US Federal Reserve banks, which provide anecdotal information on economic conditions.

Supporting adequate data access and sharing

Another area where data science can yield significant benefits relates to access to and sharing of data. This is relevant because **various obstacles still persist in making statistical information available to a broader audience**. Reasons include operational and resource constraints for producers, on one hand, and difficulties for users to effectively find the information they are looking for, on the other hand. Innovative techniques can be instrumental in both cases by allowing automatic curation and dissemination of the indicators available and, conversely, by facilitating data discovery through state-of-the-art search engines and user-friendly open data portals.[6]

The added value of data science is particularly significant for micro data, which are increasingly used by central banks.[7] Access to and sharing of disaggregated data often involve many challenges, due to their complexity and highly confidential nature.[8] Hence, there is typically a need to balance protection of sensitive information with its secure use by larger audiences, including for non-statistical purposes such as economic policy and research.[9]

Fortunately, novel data techniques, such as **privacy-enhancing technologies (PETs)**, **can mitigate these constraints** by sharing information that would otherwise be unavailable due to data protection protocols (OECD (2023)). First, **data obfuscation** can alter the data at their source by removing individual identifiers through various techniques. Second, **encrypted data processing** allows calculations to be performed while the data remain encrypted, by using techniques such as

[6] For example, the BIS Data Portal leverages the Statistical Data and Metadata eXchange (SDMX) standard to improve dissemination efficiency for producers and data discoverability for users; see Lambe and Park (2024).

[7] Micro data are "data below the level of aggregation and with a higher likelihood of identifying individual reporting units than in the aggregated data" (Israël and Tissot (2021)). Central banks have been increasingly expanding their use, including for greater precision, flexibility and timeliness compared with traditional macro indicators (Carstens (2016); IFC (2024)).

[8] For instance, the sixth principle of the Fundamental Principles of Official Statistics reads "individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes" (UN (2014)).

[9] The Principles governing international statistical activities recommend developing "a framework describing methods and procedures to provide sets of anonymous micro data for further analysis by bona fide researchers, maintaining the requirements of confidentiality"; see unstats.un.org.

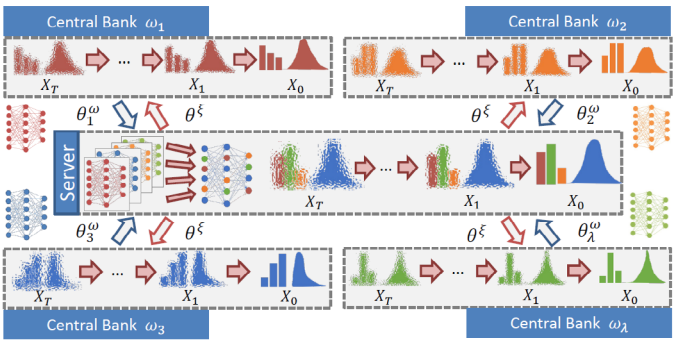
homomorphic encryption and multi-party secure private computing.[10] A third approach features **coordinated and distributed analytics**, such as federated learning that allows for the training of ML models without directly accessing the data (see Box A). Lastly, **data accountability** tools, such as personal data stores and threshold secret sharing, can help information owners and holders to retain greater control over the shared data.

In practice, **central banks have been actively adopting PETs**. A typical **data obfuscation** application is the generation of **synthetic data** to produce "fake" data that mimic the statistical properties of the original sample (Jordon et al (2022)). This technique is indeed increasingly used in official statistics and research as a way to reuse or share anonymised data (UNECE (2023b); Ravn (2025)). As analysed in a Bangko Sentral ng Pilipinas study, various tools can be used for this purpose (eg Gaussian Mixture Model, Tabular Variational Autoencoder, Generative Adversarial Networks). The best choice typically depends on various criteria, such as the quality of the synthetic data generated, the level of privacy protection and the ability to handle large amounts of information.

Central banks have also been applying other types of ML-based PETs. One example is Project Aurora by the BIS Innovation Hub Nordic Centre which relies on **network analysis** of payment data to improve the detection of money laundering activities.[11] In particular, the NSOs of Italy (Istat), Canada (Statistics Canada) and the Netherlands have explored **homomorphic encryption** techniques for analysing sensitive information collected from mobile devices. Finally, the Deutsche Bundesbank and the University of St Gallen illustrate how **federated learning** can be used to train local ML models without sharing the full data set. This approach safeguards privacy and lowers the typically high costs associated with data transfers (Graph 1).

FedTabDiff: training ML models with federated learning — Graph 1

1. Each central bank trains a **local generative model θ_i^{ω}.**
2. The Server aggregates all models into a **global generative model θ^{ξ}** accumulating knowledge from local datasets D_i^{ω}.
3. The global model θ^{ξ} is used to **generate high quality tabular data** without sharing the actual data.



Source: T Sattarov and M Schreyer, "Overcoming data-sharing challenges in central banking: federated learning of diffusion models for synthetic data generation", *IFC Bulletin*, no 64, May 2025.

[10] See Ricciato (2024) for a recent application of multi-party secure private computing in official statistics.

[11] See also Auer et al (2025) for a review of PETs in the context of digital payments data.

Box A

Sharing data with and without accessing them: exploring data science solutions

Data sharing refers to the process of making information available to others, such as individuals, entities or systems, typically through dissemination or transfer platforms. In contrast, **access** is a sharing modality and refers to the process by which authorised users obtain and use data for specific purposes, often governed by legal or commercial agreements.

The varying degrees of combining these two concepts give rise to distinct scenarios. The first occurs when **sharing data aligns with accessing them**. In practice, this option often entails secure transfers, encrypted techniques and fine-grained control of access rights. Yet, this solution is not always feasible due to scalability limitations and security concerns. A second scenario involves **sharing the insights or outputs** derived from data analysis instead of the raw data themselves. One traditional example is aggregation methods where information is shared in a summarised form, such as through averages or totals. However, this often requires centralising the data with a single holder, potentially creating concentration risks. A third scenario enables the **sharing of computations** among data holders, allowing them to retain full control over their data without needing access to others' data. The last scenario may entail **sharing information on the data themselves** (or metadata). A concrete case is making available data catalogues without granting access to the assets. Another example is exchanging information on corporate groups, for instance integrating accounting, fiscal and statistical data to streamline collections and enhance the quality of statistical information (IFC (2023b)).

Data science offers new methods to facilitate sharing without granting access to the underlying data. Federated learning, for instance, allows multiple parties to collaboratively train a machine learning model without exchanging raw data. Instead, only model updates or aggregated parameters are shared, preserving privacy and confidentiality. Another promising technology, actively explored by statistical organisations, is multi-party secure private computing, where parties can jointly compute a result without revealing their individual data, hence ensuring privacy and security throughout the process.

The above developments illustrate how **innovative techniques can support central banks in promoting efficient and secure data sharing**, with two key implications. First, they underscore the increasing focus on exchanging insights or computations without the need to share raw data, allowing for more effective use of sensitive information with significantly reduced privacy risks. Second, they point to the importance of adopting advanced encryption methods, robust cyber security measures and privacy-preserving technologies. At the same time, the emergence of quantum computing may challenge existing cryptographic practices, making early investments in **quantum-resistant solutions** essential.

3. Key obstacles in accessing, sharing and using data

While innovative data techniques can enable better access to and sharing of vast and diverse amounts of information, various challenges remain, including the need for substantial investments in IT resources, organisational barriers and the still limited information standardisation typically associated with novel or secondary sources.

The need for substantial IT resources

Obviously, **effective use of data science techniques requires state-of-the-art IT resources**. This is particularly the case when handling large and complex data sets

and facilitating their access by many stakeholders. Fortunately, central banks appear to have significantly upgraded their IT and data infrastructures over the past few decades, notably in response to the data revolution as well as to the Great Financial Crisis of 2007–09 (IFC (2020)). In particular, they have been deploying modular, scalable and metadata-driven solutions, such as data lakes, hubs and big data platforms, to efficiently store, manage and share novel or complex – including unstructured – data (Azzabi et al (2024)).

Upgrading IT infrastructures has often entailed significant transitory investments, not least because of the need to both upgrade existing technological stacks and maintain legacy solutions. For instance, a joint Bank of England and IFC survey showed that moving away from "traditional" proprietary systems such as FAME[12] has led to significant short- and medium-run costs (IFC (2023d)). One prominent reason for this is that transitioning from integrated tools can be difficult without disrupting existing setups that already work well. Another is that migrating to new solutions may require, at least initially, **additional maintenance and technical expertise** compared with more traditional applications implemented by central banks. Yet, migrations also bring important benefits, including solutions better suited for data science applications, especially those based on open source software (OSS) such as R and Python.[13] Such OSS solutions can indeed yield greater customisation, less need for (costly) user licensing requirements and reduced vendor lock-in (Araujo et al (2023)).

More fundamentally, new data science techniques require a thorough reassessment of IT resources to support central banks' activities in the longer run. One notable reason is the growing demand for **high-performance computing and efficient platforms for managing structured and unstructured data** (BIS (2024)). In particular, the use of resource-intensive applications, such as ML and big data analytics, often requires high-performance hardware (IFC (2020)).[14] One possible strategy – increasingly explored by central banks – is to rely on **cloud services**, at least in a hybrid approach combined with on-premises solutions tailored to business and security requirements, as shown by Bank Indonesia. The actual design of the IT infrastructure will in practice depend on a number of factors. For instance, cloud services offer flexible, modular and scalable solutions, especially for using customised hardware on demand (UNECE (2024a)). They can also provide **centralised storage solutions**, which may facilitate data sharing within and across organisations. Yet, the cloud may also present various limitations, starting with increased dependency on external providers. Other challenges include arguably weak data protection and sovereignty, especially when hosting facilities are located in foreign jurisdictions, raising potential legal and geopolitical risks (IFC (2025b)).

[12] FAME (Forecasting Analysis and Modelling Environment) is a proprietary statistical software and database for time series.

[13] As opposed to closed source or proprietary software, the source code of OSS can be freely edited, reproduced and redistributed; see opensource.org/osd.

[14] Such hardware includes notably graphics, tensor or neural processing units (GPUs, TPUs and NPUs, respectively) as well as clusters of computers to efficiently distribute and parallelise computations.

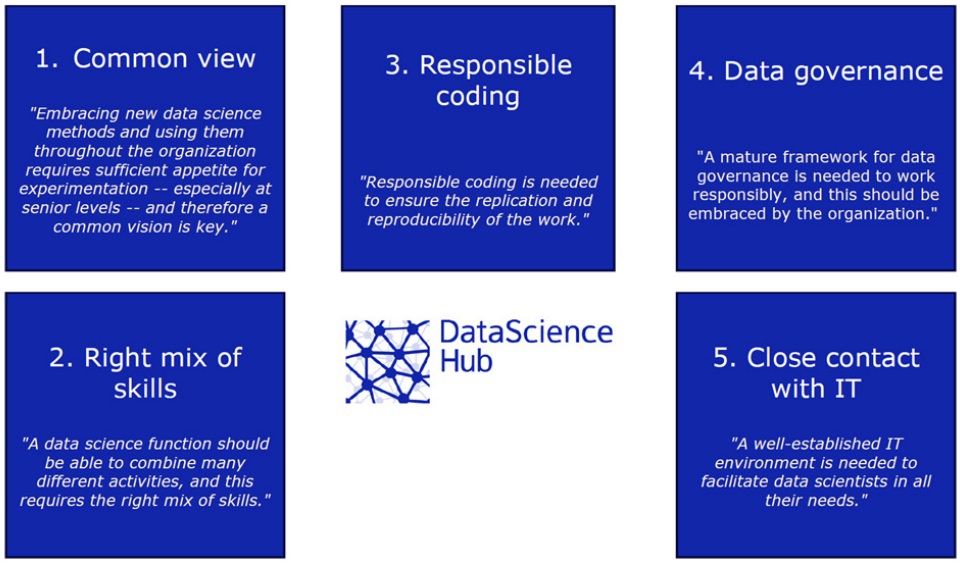
Organisational barriers

Despite the opportunities provided by innovative techniques, **the use, access to and sharing of data may face various organisational barriers** across different areas of central banking. A first challenge comes from inherent "**data silos**" (OECD (2019)). These arise because data collections and ownership are often structured organically and closely aligned with separate functional domains, such as statistics, payments, market surveillance and supervision. As a result, data tend often to be managed independently by individual business units in isolated infrastructures (IFC (2023b)). This, in turn, can limit internal user access to data and reduce awareness of available information sources (UNECE (2021)).

Limited internal coordination may be another issue. Data science fundamentally requires multidisciplinary teams (IFC (2023c)). Yet, in practice, many organisations face difficulties in promoting collaboration on projects involving subject matter experts, IT and data science teams (UNECE (2022, 2024b)). Hence, effectively meeting end users' information requirements calls for fostering synergies between business areas, which may be challenging given the diverse objectives pursued by central banks. The experience of De Nederlandsche Bank's Data Science Hub suggests that success will often depend on the following key aspects: (i) a common strategy to embrace data science; (ii) the right mix of skills; (iii) adequate IT practices, including "ethical/responsible coding" by developing software that, for instance, ensures replicability, privacy protection or accessibility; (iv) a sound data governance framework; and (v) close collaboration between stakeholders, especially data scientists, IT developers and business users (Graph 2).

Essential elements for supporting data science integration in the organisation

Graph 2



Source: P Duijm and I van Lelyveld, "Experiences, essentials and perspectives for data science in the heart of central banks and supervisors", *IFC Bulletin*, no 64, May 2025.

Finally, **reluctance to adopt data science tools in various corners of the organisation** may also raise issues. This resistance can be attributed to the diversity of profiles, with some arguably having less exposure to and/or interest in innovation. Risk aversion may also play a role, as experimenting with new data techniques can raise various challenges without yielding the expected benefits. Moreover, low levels of buy-in can result from the perceived poor transparency associated with ML techniques, often referred to as a "black box". As analysed by the International Monetary Fund (IMF), these challenges emphasise the importance of **continuous monitoring**, **upskilling** and **training** of staff. They also call for addressing complexity through measures such as **human review of AI/ML outputs** ("human-in-the-loop"). These considerations imply that adoption of data science in the organisation takes time and persistence, putting a premium on a clearly communicated, strategic and long-term approach (Duijm and van Lelyveld (2025)).

Limited standardisation hindering user access to data

Despite their benefits, the **deployment of new data science techniques is often hindered by limitations in the data themselves**. In fact, central bank users may typically need to access large granular data sets "as is", which often present complex issues (eg high multidimensionality, uncoded variables) and may **lack formal structures** compared with aggregated statistics (IFC (2024)). Moreover, many big data sets are characterised by variables formatted as free text. This can make their integration with other information sources difficult, since matching free-text variables is harder than using coded identifiers.

These difficulties are reinforced by the **limited standardisation and volatility of novel data**, at least in comparison with more traditional macroeconomic aggregates which leverage well established standards, both in terms of modelling – such as the SDMX (IFC (2025c))[15] – and methodology – such as the System of National Accounts (SNA). As noted by Banco de Portugal, micro data typically require greater modelling flexibility and customisation to handle uncoded patterns, manage complex entity relationships and navigate hierarchies to drill down from aggregates ("zoom in") or the reverse ("zoom out"). Another example relates to geospatial information, which is increasingly used by NSOs despite the lack of a fully harmonised approach across the entire data life cycle (UNECE (2024c)). Turning to methodological aspects, novel data might face sudden changes in their compilation and sources, raising inconsistencies and causing potential disruptions in the statistical chain (Benedetto et al (2025)).

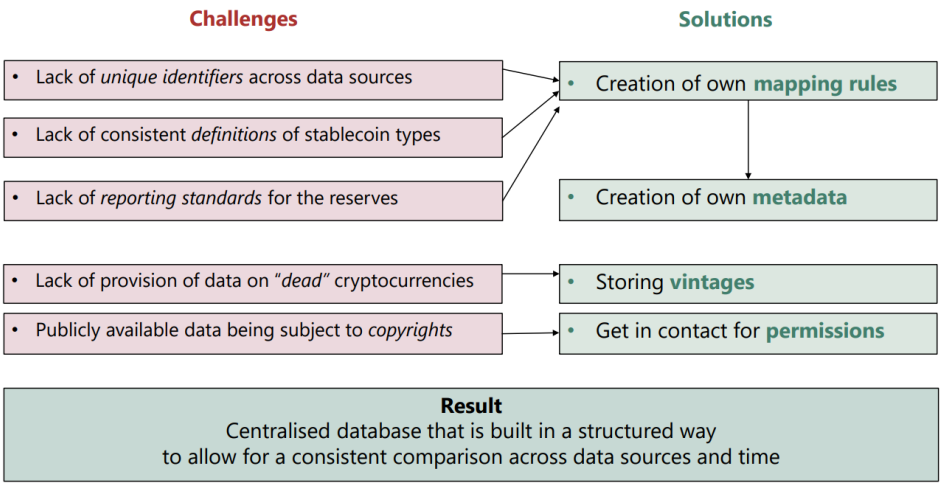
An additional challenge arises when tapping into secondary data sources to complement reference data sets, something that has become increasingly important for official statisticians. These sources often lack **unique identifiers** that are consistent across providers and domains, even though these identifiers are crucial for

[15] SDMX is an ISO-standard that facilitates the exchange, production and dissemination of statistical data and metadata between organisations. It is sponsored by eight international organisations (BIS, European Central Bank, Eurostat, International Labour Organisation, IMF, Organisation for Economic Co-operation and Development, United Nations and World Bank) and widely adopted by central banks, statistical offices and international organisations worldwide; see IFC (2025c), ECOSOC (2025) and sdmx.org.

supporting many tasks, including quality assurance and data integration. One example documented by the BIS is the absence of **common definitions and identifiers** for cryptocurrencies, which compromises consistency and comparability across multiple data sources (Graph 3). Fortunately, a number of efforts are under way to address these challenges, for instance by developing adequate metadata as well as promoting the adoption of up-to-date and easily accessible global unique identifiers, such as the Legal Entity Identifier (LEI; FSB (2024)).[16]

Linking data sets: the experience of building a database on cryptocurrencies

Graph 3



Source: A Illes and I Mattei, "Building a database on cryptocurrencies", *IFC Bulletin*, no 64, May 2025.

4. Making data access and sharing more effective

While data science can surely help, achieving effective and adequate data access and sharing may require further progress in four key areas: (i) data management and governance; (ii) interoperable information standards; (iii) common data platforms; and (iv) increased collaboration among the various stakeholders involved.

[16] The LEI is an ISO international standard to uniquely identify institutions involved in financial activities; see leiroc.org.

Advance data governance and management frameworks

Robust data governance is an essential condition to access the sheer volume of complex information available today and share insights appropriately.[17] It not only helps to ensure data quality, availability, usability, integrity and security by laying out clear sets of principles, policies and organisational frameworks. It also contributes to the adequate management of resources, particularly through the appropriate use of data science techniques. Three important benefits are worth highlighting from this perspective: access to high-quality information, its usability and appropriate sharing.

First, ensuring **access to high-quality information** is crucial, especially in terms of accuracy, completeness, lineage, traceability and transparency (IFC (2021b)). In practice, this may involve a number of solutions where data science can help. These include the development of trustworthy machine-actionable documentation of data and metadata to inform users about the sources, transformation processes, tools and methods employed. Another approach could be developing a citation standard for data, similar to current practices for software or academic research (Schwalbach and Mauer (2025)).

The second aspect is **data usability**. An organisation-wide or holistic approach to data can comprise solutions such as data inventories or catalogues to support users, including data scientists, to efficiently explore and reuse available information. For example, Banco de Portugal has adopted a master data management approach to harmonise statistical concepts across data sets in a common data dictionary, such as lists of unique codes and concepts. This strategy can also help with setting up a centralised data warehouse and a unified statistical production platform. Another important focus is to align data dissemination practices with well established guidance, such as the FAIR principles.[18]

A third consideration is to enable proper **data sharing** between interested stakeholders **while safeguarding data security and privacy**. Protecting sensitive information is very often a regulatory requirement but it is also key to securing trust among data reporters, compilers and users. Data science techniques can be useful in this regard, as illustrated by Bank Indonesia's project to protect individual data in the face of digitalisation and cyber vulnerabilities. This initiative highlighted the merits of developing a comprehensive data access framework, with adequate controls and a detailed set of information protection criteria.

[17] Data governance can be defined as "a system of decision rights and accountabilities for the management of the availability, usability, integrity and security of the data and information to enable coherent implementation and co-ordination of data stewardship activities as well as increase the capacity (technical or otherwise) to better control the data value chain, and the resulting regulations, policies and frameworks that provide enforcement"; see UNECE (2024d).

[18] These principles promote information discovery and reuse by requiring data to be findable (detailed and searchable metadata), accessible (open and standardised), interoperable (common vocabularies) and reusable (clear licences, provenance and attributes); see Wilkinson et al (2016).

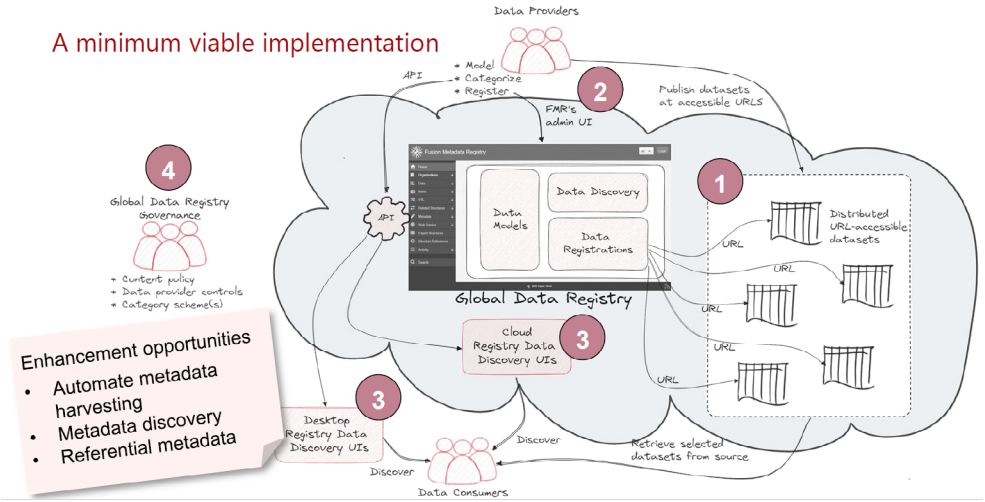
Promote interoperable standards

The development of interoperable information standards is another key factor in enabling effective and accurate use and reuse of data, especially by data scientists.

A first step is to **set up adequate data and metadata standards** to ensure accurate interpretation and seamless information exchange (UNECE (2024e)). Indeed, initiatives such as the SDMX standard, actively promoted by the central bank community, are instrumental in providing a common governance framework for data exchange, management, dissemination and discovery. In particular, and as shown by the BIS, a key strength of SDMX is the availability of a unified global metadata registry – a repository that serves as a unique reference point and source of cross-domain concepts at the international level (Graph 4; Tice et al (2025)).[19]

The SDMX global data registry to support cross-domain statistical work

Graph 4



Source: G Tice and M Nelson, "Data sharing using a global data registry, on a place to discover global structured time series, macro and micro data, *IFC Bulletin*, no 64, May 2025.

A second related consideration is to **promote and ensure interoperability among the various data standards and systems**. Key benefits include smoother data integration and reuse, reduced reporting burden and overall greater process efficiency. Central banks along with statistical offices and international institutions have been actively developing solutions, such as converters between SDMX – largely used for modelling and exchanging time series – and other standards such as XBRL – often employed for financial and micro supervisory reporting. Another progress relates to improved support for increasingly used data types, such as micro, geospatial and unstructured information.[20] Lastly, ongoing efforts are exploring how common implementation standards, such as SDMX, the Data Documentation

[19] For example, to train ML models, as done through the *gingado* library (Araujo (2023)).

[20] For example, SDMX has included support for micro and geospatial data with its version 3.0 (Nikoloutsos and Sirello (2023)).

Initiative (DDI) and Validation and Transformation Language (VTL), can help statistical organisations design metadata-driven production systems by using reference information standards, such as the Generic Statistical Business Process Model (UNECE (2024f)).

Relatedly, **interoperability within and across standards can also facilitate the implementation of decentralised data architectures** to better address business requirements ("**data mesh**"). This paradigm of data management enables domain-oriented decentralisation of analytical data operations, with distributed access, applications and responsibilities across teams (Dehghani (2022)). In addition, business areas can better tailor data processes and products to their needs while leveraging a common data modelling language. For this to be effective, standards must not only be compatible with each other but also inherently designed for interoperability, for example by ensuring that different versions of a standard can be employed to work seamlessly across teams and systems.

Finally, **standards can act as AI enablers and facilitate data findability**, as argued by Eric Anvar (OECD) in his keynote speech. For instance, the interactions between SDMX and LLMs are especially useful to enhance data discovery via natural language search while ensuring a consistent understanding across domains and business units and over the entire data life cycle.[21] Perhaps more significantly, standards can support AI systems in achieving greater accuracy through the provision of metadata or documentation about the data, hence mitigating the so-called "garbage in, garbage out" issue often faced by AI/ML solutions (IFC (2025c)).[22] Finally, standardised data could also facilitate machines in accessing information, for instance to feed AI models and digital twins (IFC (2025a); Hassani and MacFeely (2025)).

Develop common data platforms and access points

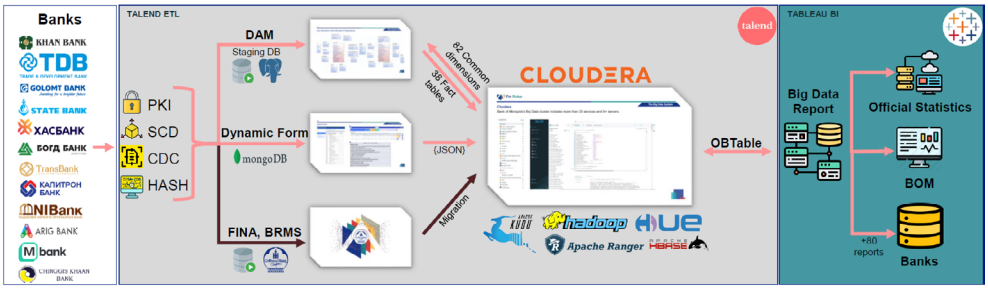
The above considerations suggest that making the most of large, diverse and complex data sets calls for setting up **modern, metadata-driven and easily accessible data platforms** within the organisation. For example, the Bank of Mongolia has developed such a platform (FinPulse) to monitor and analyse the country's financial system for regulatory purposes (Graph 5). By creating a standardised, unified regulatory database, this initiative helped **reduce reporting lags and improve timeliness** and **efficiency**. Another advantage was the strengthening of users' data access and analytical capabilities by allowing a real-time view of business intelligence reports and key indicators. Yet, building and maintaining such platforms often require massive investments in both technology and expertise. This highlights that a **strategic, institution-wide approach is essential** for successfully adopting, implementing and managing big data infrastructure, not least to address business requirements and resource constraints effectively (IFC (2020)).

[21] As a recent example, the IMF has developed an AI chatbot "StatGPT", enabling users to search for statistical data using natural language; see Kroese (2024).

[22] One related question is how to make data FAIR-R (FAIR^2) or findable, accessible, interoperable, reusable and "ready for AI"; that is, how to ensure their responsible use in AI applications for public good. See Verhulst et al (2025).

Data platform architecture: the example of the FinPulse project at Bank of Mongolia

Graph 5



The Big Data platform uses staging databases (i.e., parallel warehouse) that are kept on the bank's premises to hold raw, granular-level data using the BoM's Data Acquisition Model (DAM) format which is encrypted and transmitted to the BoM through a closed-circuit connection, and offers a single, unified source of information for both supervisors and banks to access in near real-time. Besides DAM, the platform also collects data from commercial banks and licensed institutions through manually filled Dynamic form.

1. **Data Acquisition Model (DAM):** The DAM generates an automated report daily with access in near real-time, eliminates data redundancy and inconsistency both interdepartmentally and system-wide, and can acquire and assimilate more forms of data.
2. **Dynamic Form:** Data that is not readily available through DAM will be collected using a web-based dynamic questionnaire management solution connected to the big data system. Dynamic form uses customizable data templates at a fixed frequency and is easy to make additional changes and updates to the reports.

Source: T Dorjpurev, "Big data platform (FinPulse) initiative", *IFC Bulletin*, no 64, May 2025.

Offering single access points can be another option to facilitate access, especially when data ownership is split across various organisations and domains. The European Single Access Point (ESAP) is a concrete example that connects national data platforms via a common data registry.[23] As the analysis by the Business Reporting – Advisory Group shows, the ESAP – whose initial rollout is currently planned for 2027 – offers the advantage of providing data in a standardised and machine-readable format. Relatedly, the development of "**data spaces**" is another solution for enhancing data access and reuse through common data infrastructures and governance frameworks with a clear structure for participants to access, share and collaborate on their data assets. One key example is the Common European Data Spaces (European Commission (2024)).

Lastly, central banks in particular have developed specific data access solutions to improve academic researchers' use of their micro data (IFC (2024)). Examples of such **research data centres** include on-site infrastructures such as the Deutsche Bundesbank's Research Data and Service Centre or the Bank of Italy's LabBI, which provide secure environments for accessing confidential information.[24] Other possibilities offer **remote execution systems**, such as the Bank of Italy's REX, allowing external participants to perform statistical operations on granular records without having direct access to or visualisation of the underlying data.

Foster collaboration through the sharing of tools and experience

Central banks have been leading the way to promote collaboration among their peers and counterparts to advance their data science projects as well as develop solutions

23 See the ESAP regulation at consilium.europa.eu.

24 The ECB maintains a list of the research data centres and a catalogue of available micro data sets within the European System of Central Banks; see ecb.europa.eu.

for better data access and sharing. Two main approaches have typically been followed, namely a technical one – by developing commonly available IT solutions – and an organisational one – by promoting a structured exchange of experience across interested stakeholders.

First, at the technical level, **making software available as an open source** can play a key role in fostering collaboration among developers, data scientists and, broadly, users. Tangible benefits include economies of scale, more creativity, innovation and transparency (Araujo et al (2023)). It can also help reduce the risk of bias or errors in the methods and algorithms used, although it might increase complexity.

In practice, **central banks have been increasingly producing OSS**, including for economic modelling and official statistics. As documented by the BIS, they are already sharing software codes by **open sourcing macroeconomic models**. Their experience provides a useful benchmark for new projects, based on criteria such as accessibility, documentation and replicability (Graph 6). This work also offers practical advice for enhancing code availability. In parallel, a number of **open source tools for official statistics** have been developed through the BIS OpenTech and sdmx.io initiatives.[25] Key ones are the Fusion Metadata Registry, Fusion Workbench and SDMX Dashboard Generator, which support SDMX processes across the data life cycle. More generally, leveraging OSS and common data standards can be key for central banks as producers of official statistics to develop a comprehensive open data strategy, as analysed by Regnology.

Benchmarking macroeconomic model open sourcing — Graph 6

Core criteria

- Open access
  - Easy to find
  - Free access
- Documentation
  - Academic description
  - API documentation
  - Input data documentation
  - Output documentation
- Replication
  - End-to-end execution
  - Instructions/tutorial/vignette
  - Minimum requirements

Additional criteria

- Software
  - Software open availability
  - Testing
  - Technology neutrality
  - Availability of past versions
  - Explicit code versioning
- Contribution
  - Contact with software maintainers
  - 3rd party contributions possible
  - Contribution guidelines
- License
  - Explicit open source license

Source: D Araujo, "Open sourced central bank macroeconomic models", *IFC Bulletin*, no 64, May 2025.

Second, from an organisational perspective, collaboration may require a coordinated involvement of the stakeholders willing to share data, knowledge and best practices. To this end, central banks have been actively developing a **community of practice** to exchange ideas continuously, not least in the context of the IFC. A number of initiatives have also aimed to extend such knowledge-sharing beyond the

[25] See bis.org and sdmx.io, respectively.

central banking community, including to academia and international organisations. An example is the **International Network for Exchanging Experience on Statistical Handling of Granular Data** (INEXDA), which facilitates the exchange of information on micro data issues (IFC (2018b)).

Collaboration can be strengthened by the ongoing development of **common international data frameworks**. Important and promising developments in this regard are the revisions of core statistical methodologies (such as the newly revised SNA 2025 and Balance of Payments Manual - BPM7), the data collections initiated in the context of the **Data Gaps Initiative** (DGI)[26] and the various actions undertaken to strengthen the global statistical infrastructure – including in terms of registers, identifiers and data standards. Taken together, these initiatives can be instrumental in facilitating better access to alternative data for official statistics and fostering adequate sharing of economic and financial information for the public good.

[26] The DGI is an international initiative endorsed by the G20 and supported by several countries. Launched in the aftermath of the Great Financial Crisis, it entered its third phase in 2022 to specifically address four statistical areas: (i) climate change; (ii) household distributional information; (iii) fintech and financial inclusion; and (iv) access to private sources of data and administrative data, and data sharing. See IMF et al (2023) and imf.org.

References

Araci, D (2009): "FinBERT: Financial sentiment analysis with pre-trained language models", arXiv.

Araujo, D (2023): "gingado: a machine learning library focused on economics and finance", *BIS Working Papers*, no 1122, September.

Araujo, D, S Doerr, L Gambacorta and B Tissot (2024a): "Artificial intelligence in central banking", *BIS Bulletin*, no 84, January.

Araujo, D, N Bokan, F Comazzi and M Lenza (2024b): "Word2Prices: embedding central bank communications for inflation prediction", *CEPR Discussion Paper*, no 19784, December.

Araujo, D, S Nikoloutsos, R Schmidt and O Sirello (2023): "Central banks as users and providers of open-source software", Box 1 in "Data science in central banking: applications and tools", *IFC Bulletin,* no 59, October, pp 8–10.

Auer, R, R Böhme, J Clark and D Demirag (2025): "Privacy-enhancing technologies for digital payments: mapping the landscape", *BIS Working Papers*, no 1242, January.

Azzabi S, Z Alfughi and A Ouda (2024): "Data lakes: a survey of concepts and architectures", *Computers*, vol 13, no 7, 183.

Bank for International Settlements (BIS) (2024): "Artificial intelligence and the economy: implications for central banks", *BIS Annual Economic Report*, chapter III, pp 91–127.

Benedetto, C, S Crestini, A de Gregorio, M de Leonardis, A del Monaco, D Gulino, P Massaro, F Monacelli and L Rubeo (2025): "Applying artificial intelligence to support regulatory reporting management: the experience at Banca d'Italia", *Bank of Italy Occasional Papers*, no 927, April.

Brault, J, M Haghighi and B Tissot (2024): "Granular data: new horizons and challenges for central banks", *IFC Bulletin,* no 61, July.

Carstens, A (2016): "Micro-data as a key input to designing macro-prudential policy: the Mexican experience", *ECB Conference on Statistics*, July.

De Beer, B and B Tissot (2021): "Official statistics in the wake of the Covid-19 pandemic: a central banking perspective", *Theoretical Economics Letters*, vol 11, no 4, August.

Dehghani, Z (2022): *Data mesh: delivering data-driven value at scale*, O'Reilly Media.

Duijm, P and I van Lelyveld (2025): "Data science for central banks and supervisors: how to make it work, actually", *Harvard Data Science Review*, vol 7, no 1.

European Commission (2024): *Commission staff working document on common European data spaces*, SWD (2024) 21, January.

Financial Stability Board (FSB) (2024), *Implementation of the Legal Entity Identifier: progress report*, October.

Hassani, H and S MacFeely (2025): "Digital twin and official statistics: a paradigm shift?", *Statistical Journal of the IAOS*, April.

International Monetary Fund (IMF), Inter-Agency Group on Economic and Financial Statistics and Financial Stability Board Secretariat (2023): *People planet economy: delivering insights for action*, G20 Data Gaps Initiative 3 Workplan, March.

Irving Fisher Committee on Central Bank Statistics (IFC) (2018a): "Central banks and trade repositories derivatives data", *IFC Report*, no 7, October.

——— (2018b): "INEXDA – the granular data network", *IFC Working Papers*, no 18, October.

——— (2020): "Computing platforms for big data analytics and artificial intelligence", *IFC Report*, no 11, April.

——— (2021a): "Micro data for the macro world", *IFC Bulletin*, no 53, April.

——— (2021b): "Issues in data governance", *IFC Bulletin*, no 54, July.

——— (2021c): "Use of big data sources and applications at central banks", *IFC Report*, no 13, February.

——— (2023a): "Post-pandemic landscape for central bank statistics", *IFC Bulletin*, no 58, June.

——— (2023b): "Data-sharing practices", *IFC Guidance Note*, no 3, March.

——— (2023c): "Data science in central banking: applications and tools", *IFC Bulletin*, no 59, October.

——— (2023d): "Central banks' use of time series products", *IFC Guidance Note*, no 4, December.

——— (2024): "Granular data: new horizons and challenges", *IFC Bulletin*, no 61, July.

——— (2025a): "Governance and implementation of artificial intelligence in central banks", *IFC Report*, no 18, April.

——— (2025b): "Cloud services and AI in central banks: balancing benefits and risks", *IFC Report*, no 18, Box C, February.

——— (2025c): "SDMX adoption and use of open source tools", *IFC Report*, no 17, February.

Israël, J-M and B Tissot (2021): "Incorporating micro data into macro policy decision-making", *IFC Bulletin*, no 53, April.

Jordon, J, L Szpruch, F Houssiau, M Bottarelli, G Cherubin, C Maple, S Cohen and A Weller (2022): "Synthetic data – what, why and how?", *ArXiv*.

Kroese, B (2024): "Exploring official statistics using generative AI and SDMX", presentation at the 72nd plenary session of the Conference of European Statisticians, Geneva, 20–21 June.

Lambe, E and T Park (2024): "The BIS Data Portal project – delivering the next generation platform for BIS statistics", *IFC Bulletin*, no 60, April.

Liu, F, K Ting and Z-H Zhou (2008): "Isolation forest", *Eighth IEEE International Conference on Data Mining*.

MacFeely, S (2020): "In search of the data revolution: has the official statistics paradigm shifted?", *Statistical Journal of the IAOS*, vol 36, no 4, November.

Mittal, M, L Goyal, D Hermanth and J Sethi (2019): "Clustering approaches for high-dimensional databases: a review", *WIREs Data Mining and Knowledge Discovery*, vol 9, no 3, e1300, January.

Nikoloutsos, S and O Sirello (2023): "SDMX, an international standard for micro data", presentation at the SDMX Global Conference 2023, November.

Organisation for Economic Co-operation and Development (OECD) (2019): *Enhancing access to and sharing of data: reconciling risks and benefits for data re-use across societies*, OECD Publishing, Paris.

——— (2023): "Emerging privacy-enhancing technologies: current regulatory and policy approaches", *OECD Digital Economy Papers*, no 351, March.

Ravn, L (2025): "The fabrication of synthetic data promises: tracing emerging arenas of expectations and boundary work", *Big Data & Society*, vol 12, no 1, January.

Ricciato, F (2024): "Steps toward a shared infrastructure for multi-party secure private computing in official statistics", *Journal of Official Statistics*, vol 40, no 1, March.

Schwalbach, J and R Maurer (2025): "Sharing digital trace data: researchers' challenges and needs", *Big Data & Society*, vol 12, no 1, March.

Tice, G, M Nelson, S Nikoloutsos and X Sosnovsky (2025): "SDMX metadata-driven approach: the BIS experience", Box C, *IFC Report*, no 17, February, p 11.

United Nations (UN) (2014): "Resolution adopted by the General Assembly on 29 January 2014: 68/261 – Fundamental Principles of Official Statistics", January.

United Nations Economic and Social Council (ECOSOC) (2025): *Report of the Statistical Data and Metadata eXchange (SDMX) sponsors*, 56th session of the United Nations Statistical Commission, New York City, 4–7 March.

United Nations Economic Commission for Europe (UNECE) (2021): *Guide to sharing economic data in official statistics*, February.

——— (2022): *Machine learning for official statistics*, March.

——— (2023a): "Large language models for official statistics", *High-Level Group for the Modernisation of Official Statistics White Paper*, December.

——— (2023b): *Synthetic data for official statistics – a starter guide*, January.

——— (2024a): *Cloud for official statistics*, March.

——— (2024b): *Organisational aspects of implementing ML based data editing in statistical production*, High-Level Group for the Modernisation of Official Statistics, September.

——— (2024c): "INGEST Task Force on Standards Issues: a path towards the use of common standards to support the integration of geospatial and statistical information", *Working Paper Series on Statistics*, no 11, August.

——— (2024d): *Data stewardship and the role of national statistical offices in the new data ecosystem*, ECE/CES/STAT/2023/4, April.

——— (2024e): *Data governance framework for statistical interoperability (DAFI)*, High-Level Group for the Modernisation of Official Statistics, March.

——— (2024f): *Statistical implementation standards in the context of GSBPM*, High-Level Group for the Modernisation of Official Statistics, October.

Verhulst, S, A Zahuranec and H Chafetz (2025): "Moving toward the FAIR-R principles: advancing AI-ready data", *SRRN*, no 5164337, March.

Wilkinson, M et al (2016): "The FAIR guiding principles for scientific data management and stewardship", *Scientific Data*, vol 3, 160018, March.

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Opening remarks[1]

Bruno Tissot,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BANCA D'ITALIA
EUROSISTEMA

Data Science in Central Banking: Enhancing the access to and sharing of data

Opening remarks by Bruno Tissot
Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat

3rd IFC and Bank of Italy Workshop

Rome, 17 October 2023

Ladies and Gentlemen, dear colleagues,

I am very happy to welcome you to today's event on Data Science in Central Banking hosted by the Bank of Italy.

As you know, the Irving Fisher Committee on Central Bank Statistics (IFC) of the Bank for International Settlements (BIS) decided a few years ago to organise recurrent workshops on "Data science in central banking". The aim has been to review the adoption of data analytics and business intelligence techniques and developments in the big data ecosystem. How can this be achieved? We are convinced that showcasing projects and sharing experience can provide a key opportunity to develop in-house knowledge and reduce the reliance on external services providers.

The most recent IFC workshop took place in 2022 and focused on the broad spectrum of data science applications/tools used in central banks. And indeed there are many projects currently going on in our community. A key supporting factor has been the sheer volume and complexity of financial data available in today's societies. It calls for more sophisticated techniques for data management and analysis, a trend reinforced by the new opportunities opened up by artificial intelligence and machine learning. Another factor has been the greater focus on real-time, evidence-based policymaking, which requires authorities to rely on better analytical and forecasting capacities to support their decisions. It also puts a premium on accessing data and disseminating them to interested users: international institutions, other domestic authorities, academia, the general public etc. These aspects are particularly relevant for central banks as both producers and users of official data – a quite unique perspective in fact.

But we should start by trying to define what data science really means for us. This concept generally refers to the study of data and therefore includes the various techniques for extracting insights from them. But data science is fundamentally different from traditional data analysis, as it typically applies to large, complex and/or unstructured information

sets. Hence the role of data scientists lies at the intersection of three areas: information technology (IT); mathematical and statistical methods; and business, or "subject-matter" expertise.

Let me discuss these three aspects in turn, noting that more details can be found in the IFC Bulletin no 59 just published by the BIS last week on "*Data science in central banking: applications and tools*".

IT infrastructure and tools

From the outset, IT has absorbed a great deal of attention and resources to support data science projects, as underscored by the IFC Report no 11 published in 2020 on "*Computing platforms for big data analytics and artificial intelligence*". This report showed that central banks are increasingly aware that a modern IT architecture is crucial to reliably and securely deal with data. In addition, a key objective for them is to facilitate access to a large and diverse range of sources as well as relevant IT software and tools in a user-friendly way.

But implementing such an IT architecture can be challenging, calling as it does for careful implementation, clear governance frameworks, and the application of common standards. Exploratory work has sometimes barely started, and substantial staff and IT investment as well as business adjustments will continue to be needed. Looking ahead, the primary focus will thus be on adopting advanced IT tools and engineering practices – including cloud computing, software containers, automation tools etc.

Leveraging mathematical and statistical techniques

Once a sufficient IT infrastructure is available, the next stage is to perform the various mathematical and statistical operations that are needed to deal with the raw data at hand. One needs not only to access very large and complex information sets, but also to derive statistics compiled via multiple sequential tasks (signal extraction, quality management, dissemination) before using them to extract relevant insights. Many different approaches can be leveraged for this purpose, including textual analysis, with the support of natural language processing (NLP) tools and large language models (LLMs).

This area is obviously key for us statisticians in central banks. Last year the Committee published the IFC Bulletin no 57 that took stock of how central banks are developing machine learning across a variety of use cases. It showed that the new techniques can help improving the efficiency and effectiveness their operations, including through the dealing of larger and new sources of information in a more automated way. This Bulletin also underlined the diversity and maturity of innovative approaches already developed and used.

In fact, the new techniques being contemplated can facilitate the gathering of more and better information, which is key for central banks that rely heavily on data. They can help respond to their information demand by enhancing statistical quality, providing richer context, and making sense of the abundance of data available to derive useful conclusions. This is particularly evident as regard the three main central bank policy areas, namely

monetary stability, micro supervision and macroprudential tasks… as well as for broader financial stability-related goals, including the functioning of the payment system, financial inclusion, consumer protection, and anti-money laundering as we will hear today.

The experience of central banks has highlighted several lessons. Let me highlight a few:

First, there is a wealth of alternative information sources that have barely been tapped and which can provide new and useful insights. Financial granular data, administrative registers as well as text-based information (a type of data that had been available well before the advent of big data techniques…) represent the new frontier for central bank statisticians.

Second, complementarity is essential: new methods can bring additional insights but have to be blended with other types of exercise.

Third, while data science can be instrumental in dealing with complexity, there are important challenges faced by users as the functionality of the new tools is rarely intuitive.

Subject-matter expertise

A third key aspect of data science is to rely on a good understanding of the business cases, requiring therefore a close cooperation with subject-matter experts. This is because analysing data calls for an awareness of the way the statistics have been compiled as well as of the complex factors that drive them.

Moreover, subject-matter expertise is essential to support informed policy decisions: translating data insights into actionable recommendations for central banks cannot be communicated as a "black box" and requires transparent explanations to the various stakeholders involved – from other authorities to the general public.

From this perspective, the focus should be on:

Transparency ("metadata", as we statisticians usually refer to);

Ethics and governance (noting that central banks are in a rather good position, being able to build on the strong foundations provided by the international fundamental principles of official statistics); and

Public communication (putting a premium on enhancing economic and statistical literacy in society).

Accessing alternative information sources and sharing data

Let me conclude by looking at the topic of todays' event to deal, in particular, with issues related to the access to private and administrative sources and data-sharing.

As you know, one important development for central bank statisticians has been related to the international cooperation framework under the Data Gaps Initiative (DGI) endorsed

by the G20 and supported by many countries. This work was decided in response to the Great Financial Crisis (GFC) of 2007–09 to help to coordinate national work on financial data sets, including through enhanced data-sharing. It comprised several phases. The third one, initiated in 2022, calls for better data to understand today's key challenges and to make official statistics more detailed and timely. Main international financial organisations, including the BIS with the support of the IFC which represents more than 100 central banks globally, are actively involved.

The new five-year phase of the DGI covers priority policy needs in the areas of (i) climate change; (ii) distribution of household income and wealth; (iii) fintech and financial inclusion; and, particularly relevant for our event today, (iv) access to private and administrative sources of information and data-sharing. One needs, in this context, to emphasise the instrumental role played by the Bank of Italy in pushing this last topic on the agenda when chairing the G20 in 2021.

These international initiatives underscore that the topic of today's workshop, ie data access and sharing, is crucial. In fact, a key lesson of the recent Covid-19 crisis is that coping with the shifts in user demands triggered by unprecedented shocks such as pandemics requires tapping into new information sources, often held by private actors and government agencies. Another lesson is that the use of artificial intelligence in the compilation phase could transform the whole production chain of official statistics, making it faster and more resilient and user-friendly. Furthermore, it is also important to have standardised data publication and international data-sharing in the distribution phase. Let me remind you in this context that the BIS is an active sponsor of the Statistical Data and Metadata eXchange (SDMX) standard, and that these issues will be dealt extensively not only during this workshop but also on the occasion of the upcoming 9th SDMX Global Conference on "*Empowering Data Communities*" to be held in Manama, Bahrain, on 29 October-2 November 2023.

Looking ahead, we believe that the way forward should be based on cooperation, to spur knowledge-sharing and the exchange of best practice. This can be instrumental to facilitate **access** to privately-held data and administrative registers. For instance, recent pilot projects in central banks have showed that innovative text-mining techniques can greatly help to extract relevant information that policy makers are needing the most. Similarly, AI approaches can be used to better monitor and understand complex, granular developments in financial markets. Needless to say, substantial investments, especially in terms of IT hardware, software and staff skills will be required to replicate such initiatives more broadly.

Cooperation should also lead to better **data-sharing**, especially with other producers of official statistics, the private sector and academia – noting that several central banks are, together with a number of national statistical offices and international organisations, actively involved in INEXDA, the International Network for Exchanging Experience on Statistical Handling of Granular Data, to in particular facilitate the use of granular data for research purposes. Data-sharing across jurisdictions is also an essential element supporting global public goods. Fortunately, significant progress has been already achieved by the international community of central banks and supervisory authorities

in the sharing of financial and regulatory data, with the strong support of the BIS. Lastly, data-sharing can more generally help to ensure the broadest possible data availability in the most efficient way, while keeping the reporting burden under control.

Let me stop here. The fact that more than 100 institutions are represented in today's event is a clear indication of the relevance of the selected presentations. It also shows that the various issues covered are particularly topical for central banks and all statisticians more generally.

Designed by the Printing and Publishing Division of the Bank of Italy

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Project Aurora: the power of data, technology and collaboration to combat money laundering across institutions and borders[1]

Beju Shah,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BIS Innovation Hub



Project Aurora

The power of data, technology and collaboration to combat money laundering

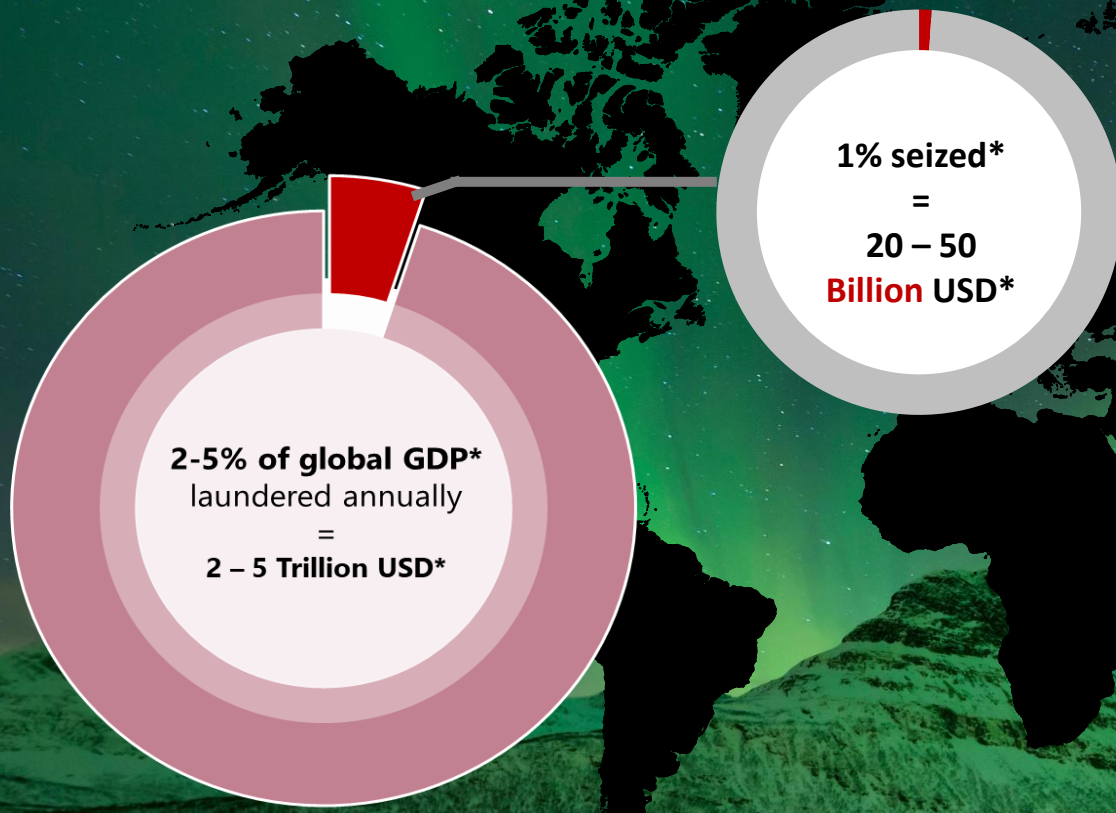
Irving Fisher Committee Data Science Workshop – Rome – October 2023

Beju Shah, Head of BIS Innovation Hub Nordic Centre

BIS Innovation Hub

Aurora

Money laundering is a global and costly problem



Increasing costs of financial crime compliance

2022:
$274 billion

28% increase
from 2020

80% increase
From 2019

Source: 1) UN Office of Drugs & Crime - https://www.unodc.org/unodc/en/money-laundering/overview.html and 2) https://www.unodc.org/documents/data-and-analysis/Studies/Illicit_financial_flows_2011_web.pdf

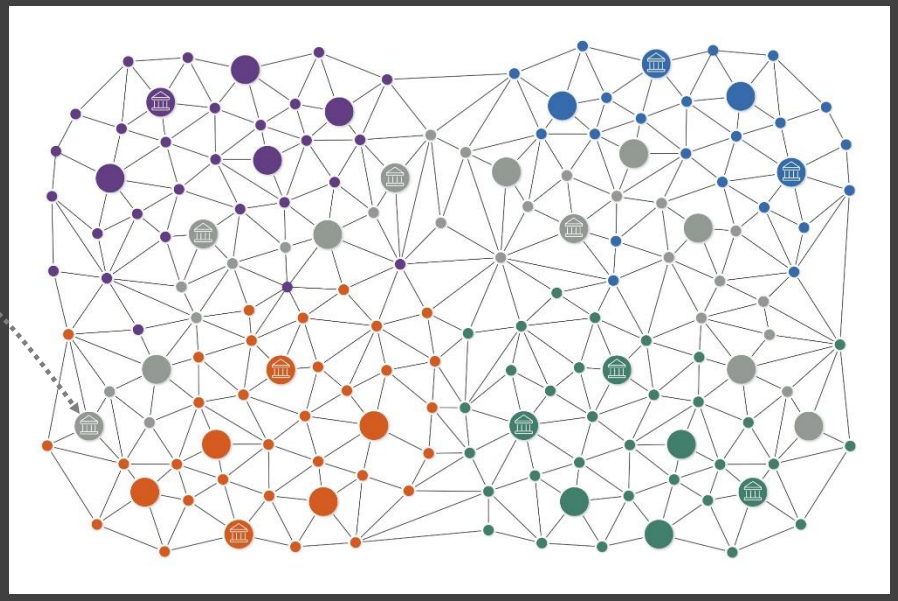
* Estimates are calculated relative to recent GDP

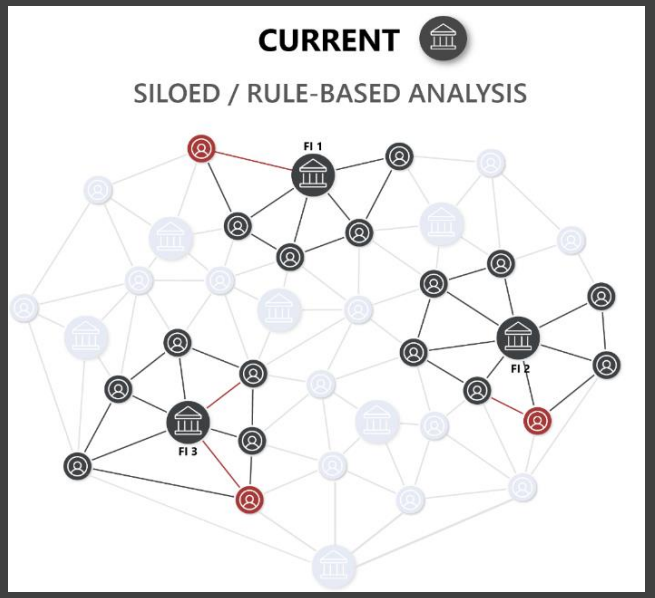
** True Cost Financial Crime Compliance 2022 | LexisNexis Risk Solutions. **Note: an increase of 28% from 2020 to 2022 in compliance costs, 23%-80% increase between 2019 and 2022.**

Challenges with today's approach(es)

- Money launderers create a **complex network of transactions across financial institutions and borders**, while in contrast, these institutions themselves often have a **limited and siloed view of the network**

money laundering pattern



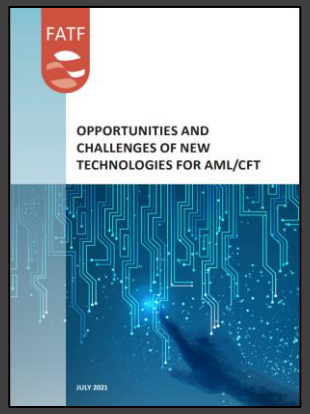


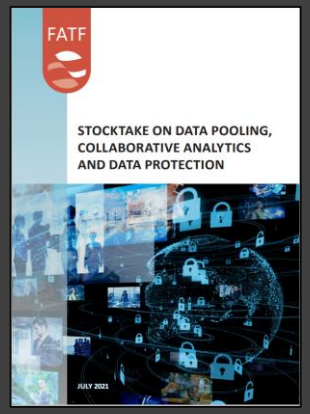
- **Non-standardised and / or readily and easily consumable data**, making collaboration and analysis across institutions (and borders) **more difficult**
- Rule base approaches are not suitable for analysing behavioural pattern
- Data required (based on the typology) for effective **AML/CFT measures have to be balanced with the need to protect individuals' privacy, personal data and human rights**.

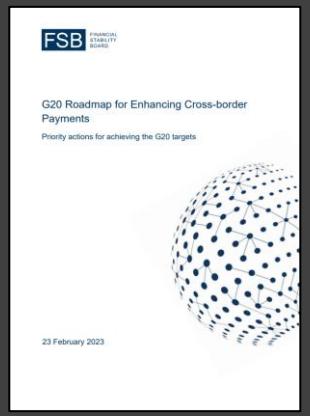
BIS Innovation Hub

Aurora

Leveraging data and technology







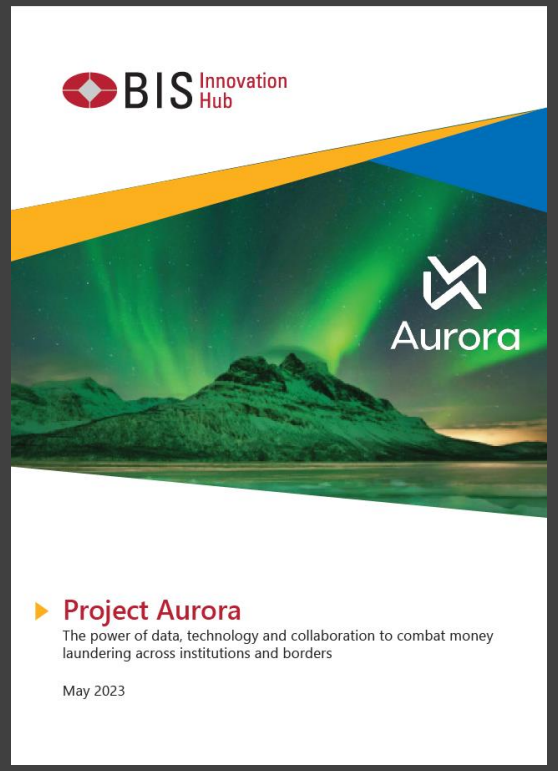


- FATF stocktake highlights several approaches / technologies to improve AML efforts:
  - **Data sharing / data pooling**
  - **Network analytics**
  - **Machine learning / Federated learning**
  - **Privacy enhancing technologies (PETs)**
- **Payments data could be utilised more.** Analysis of this data where **privacy is protected, sensitive information is encrypted** and, in some cases, where such **data need not leave a jurisdiction or institution** could be **game changing**.
- Cross border payments modernisation, ISO20022, legal entity identifier (LEI), FATF Recommendation 16, public private partnerships, beneficial ownership standards and further data standardisation initiatives could help.
- Network nature of payments data is well suited to graph data structures.
- When combined with the latest technologies could unlock the power of this data in AML efforts.

BIS Innovation Hub

Aurora

Project overview



PROBLEMS

- Money laundering is a **cross institutional, payment system and borders problem.** It is a complex data challenge
- Siloed and rule-based monitoring is **ineffective when a network view of behavioural patterns is required**
- Compliance costs > **$274 billion***

  < **1%** laundered money seized (**$20-$50 billion****)
- **Holistic network view needed** but FIs **cannot share data**

OBJECTIVES

- Show how to **improve the performance and effectiveness** of AML analysis while **protecting privacy and sensitive information**
- **Showcase the potential of AI and PETs** with connected payments data to detect money laundering
- **Test and compare the performance** of advanced technologies and different analysis approaches
- Underscore **the power of public private partnerships, and collaboration** when enabled by technology

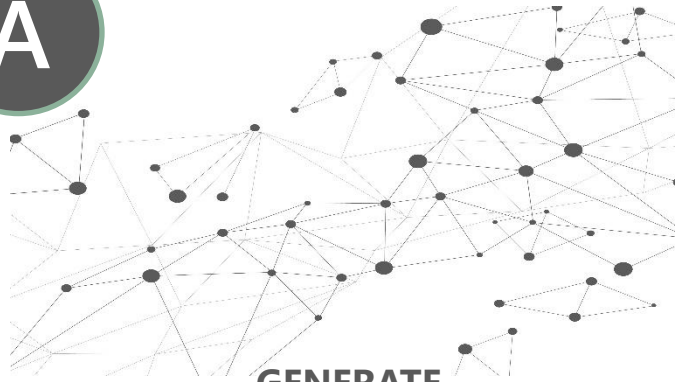
* True Cost Financial Crime Compliance 2022 | LexisNexis Risk Solutions. **Note: an increase of 28% from 2020 to 2022 in compliance costs, 23%-80% increase between 2019 and 2022.**

**UN Office of Drugs & Crime - https://www.unodc.org/unodc/en/money-laundering/overview.html & https://www.unodc.org/documents/data-and-analysis/Studies/Illicit_financial_flows_2011_web.pdf

BIS Innovation Hub

Aurora

A



GENERATE SYNTHETIC DATA

DATA SET

6 COUNTRIES

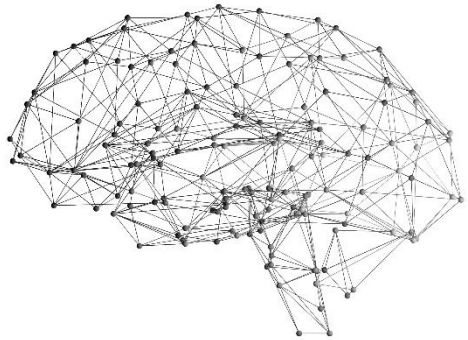
29 FINANCIAL INSTITUTIONS

>155,000 ACCOUNTS

MONEY LAUNDERING PATTERNS EMBEDDED INCLUDE:

- **COMPLEX LAYERING SCHEMES**
- **SMURFING**

B



TESTING MACHINE LEARNING MODELS

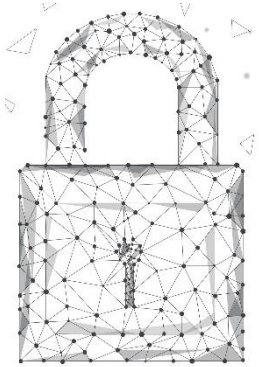
MONITORING SCENARIOS

SILOED

NATIONAL

CROSS-BORDER

C



TESTING PRIVACY-ENHANCING TECHNOLOGIES

COLLABORATIVE ANALYSIS & LEARNING

CENTRALISED

CENTRALISED XB

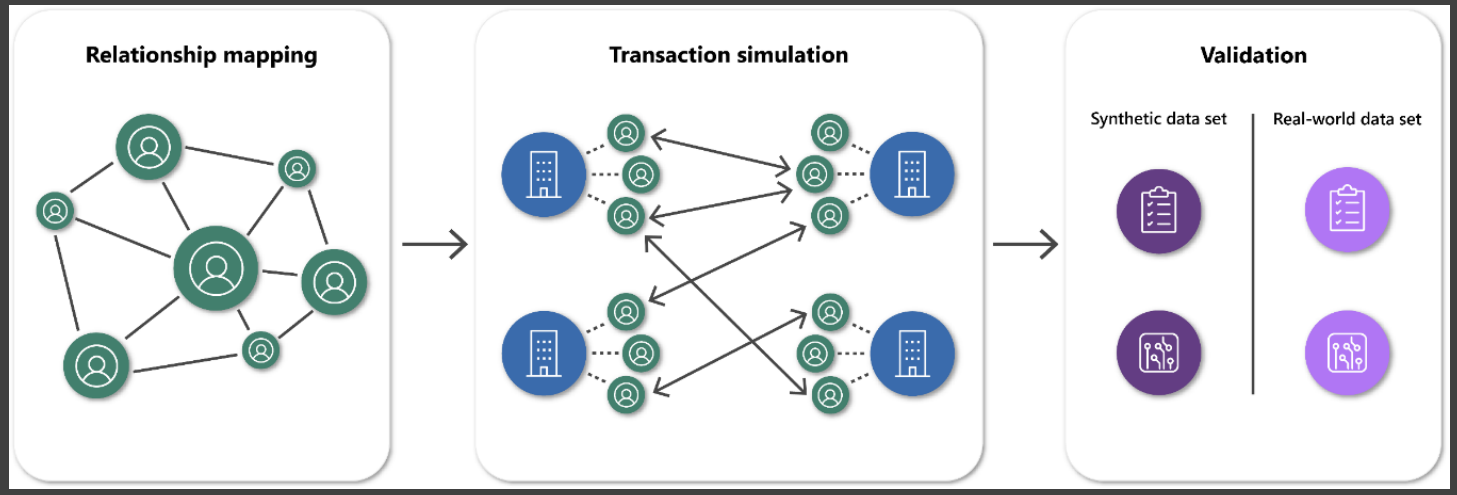
DECENTRALISED

HYBRID

Part A – Generating the synthetic data

The synthetic data are generated in three steps:

1. **Relationship mapping** generates a network graph that represents relationships between individuals and businesses.
2. **Transaction simulation** of transaction flows based on the relationships in the relationship map.
3. **Validation** of the generated synthetic data

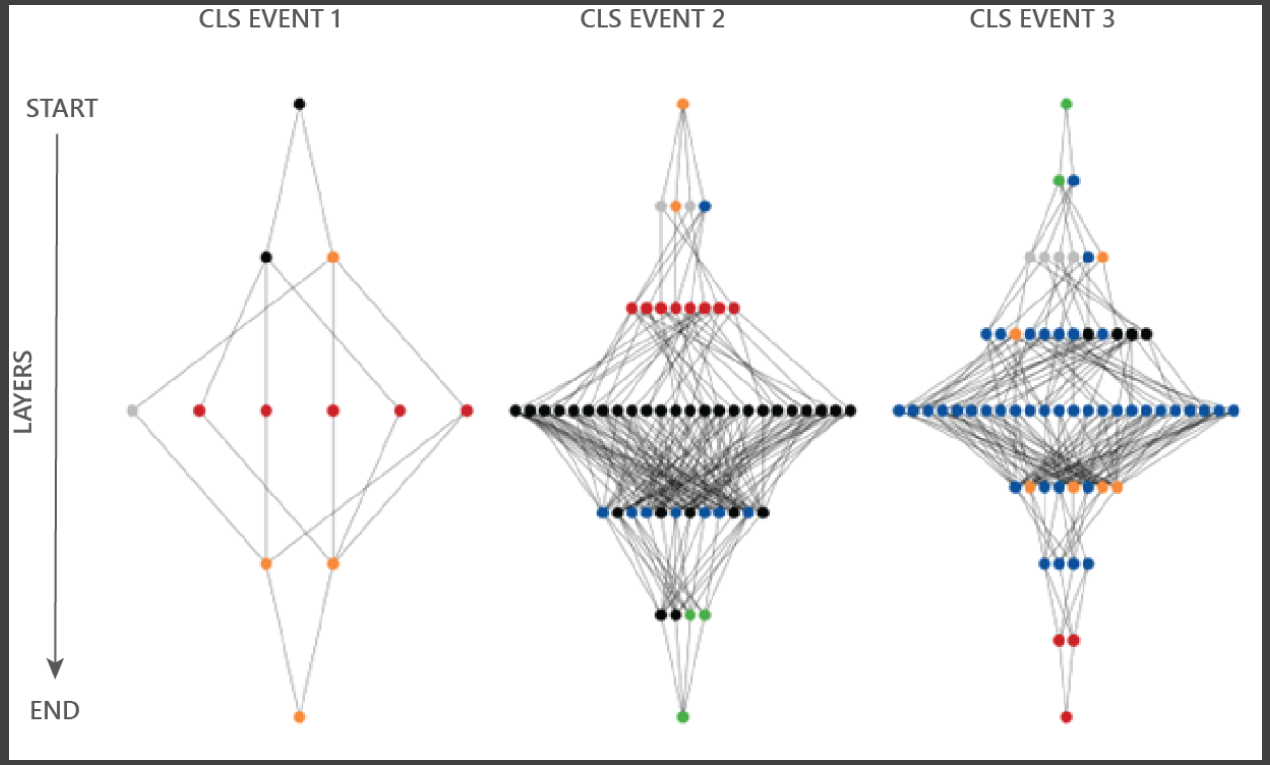


BIS Innovation Hub

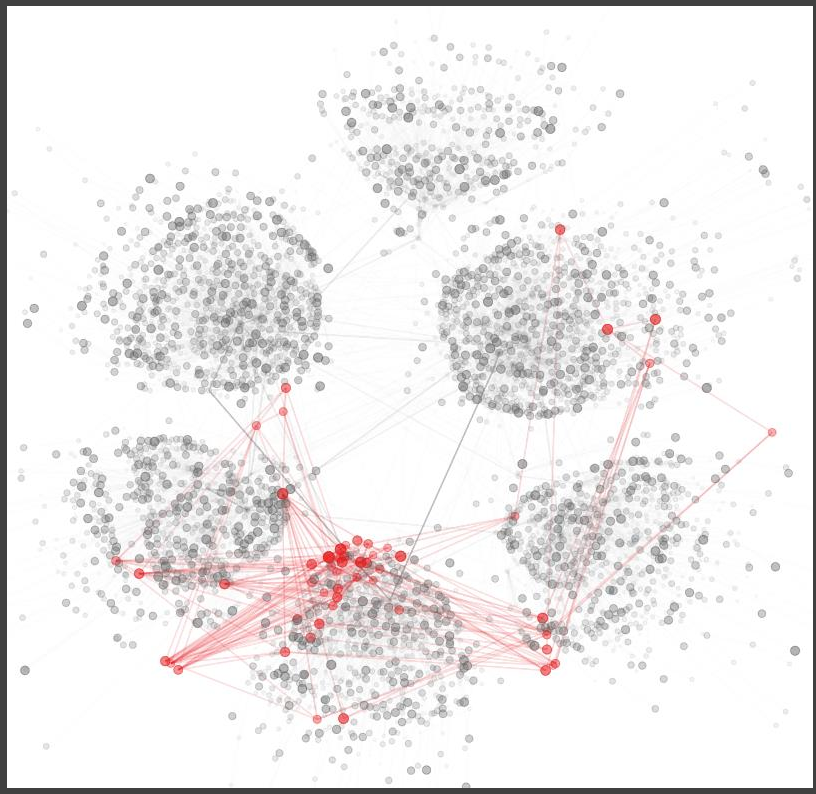
Aurora

Part A - Synthetic data

Examples of complex layering scheme (CLS) patterns in the synthetic data



Example of transactions over 1 day with money laundering activities highlighted in red



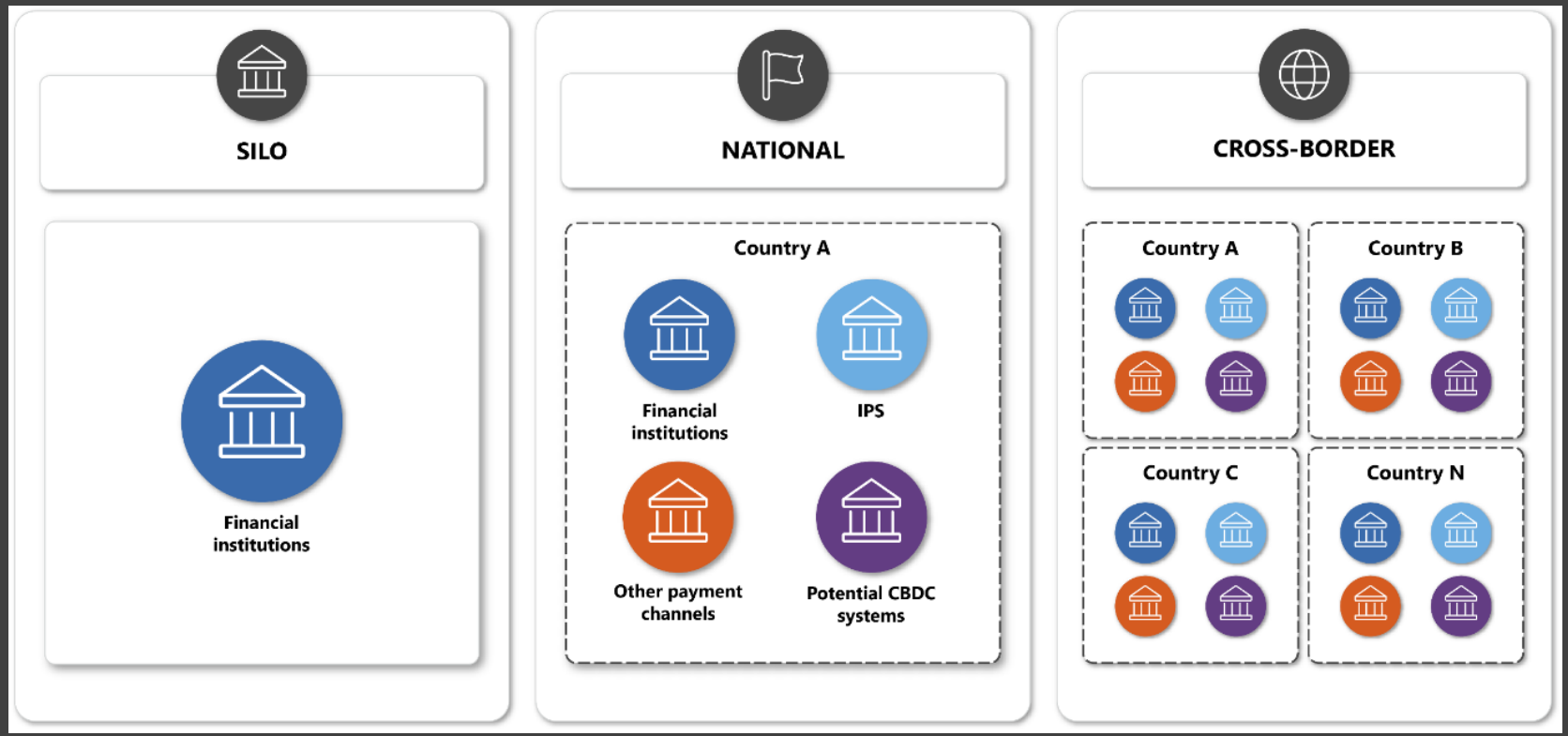
BIS Innovation Hub

Aurora

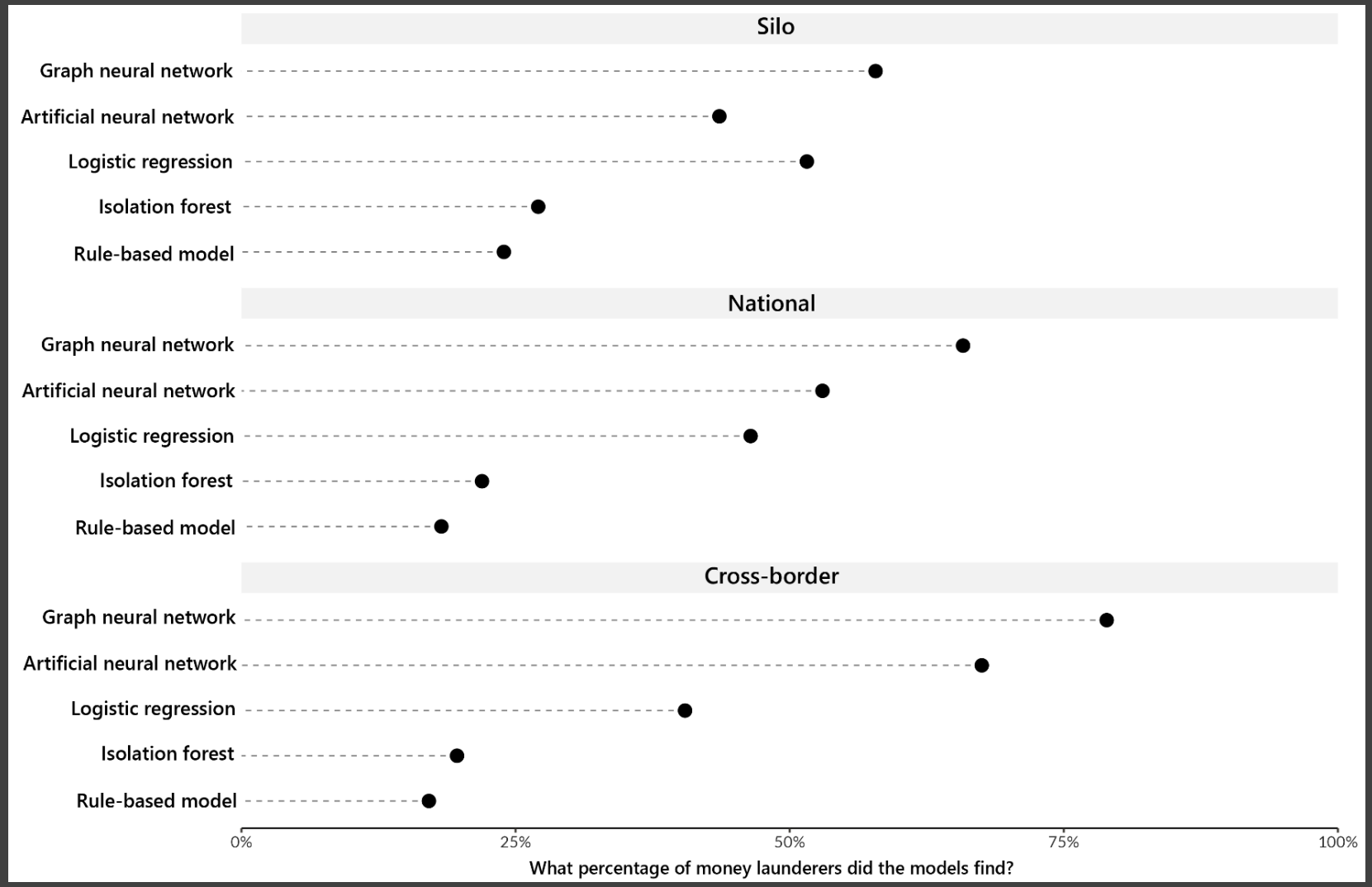
Part B – Applying machine learning models

Several machine learning models tested against different views of data for performance /effectiveness

- **Graph neural-networks**
- **Artificial neural networks**
- **Logistic regression**
- **Isolation forest**
- **Rule based as baseline**



Part B – Results

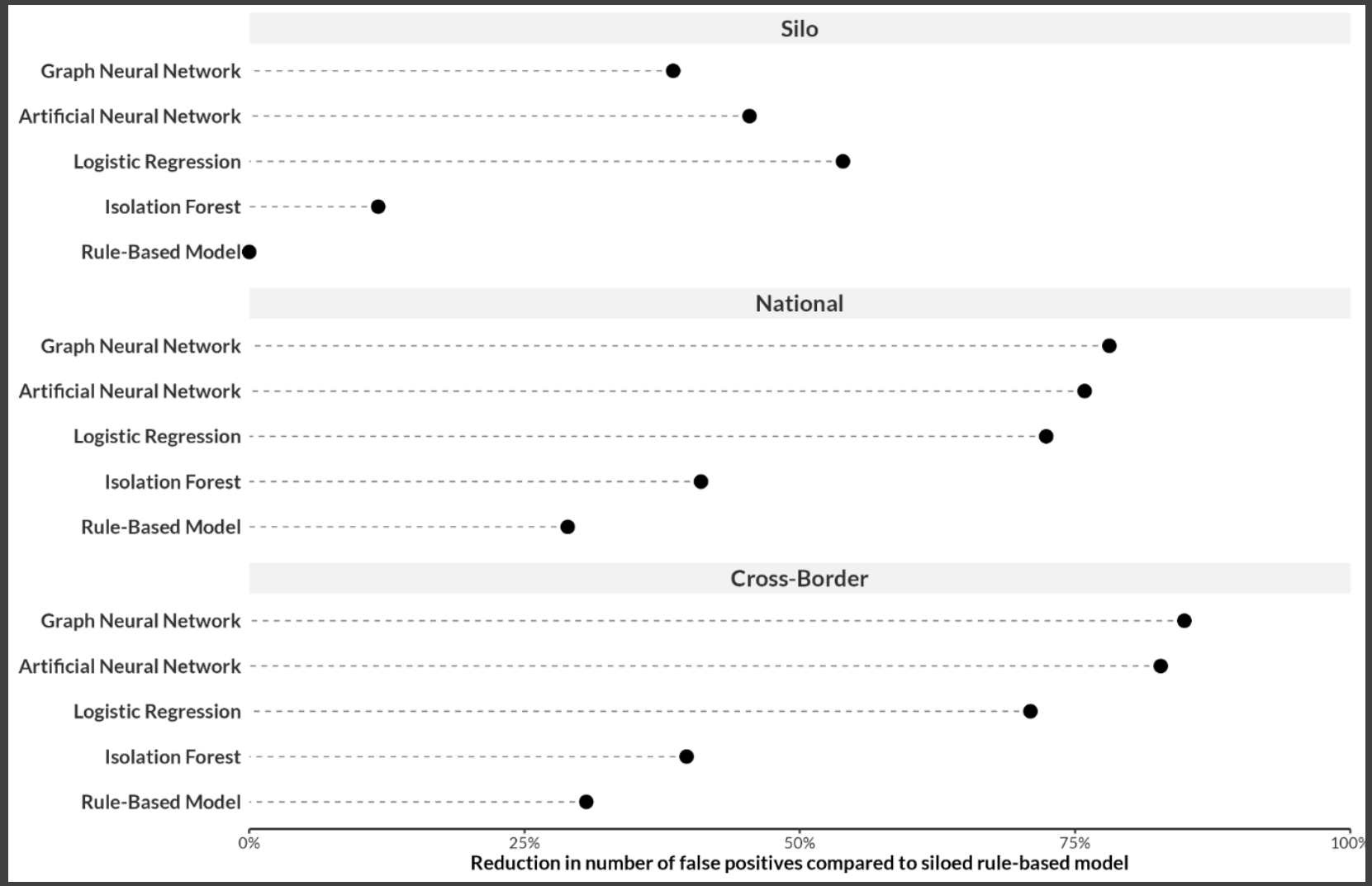


2 – 3x increase in detection of money launderers

BIS Innovation Hub

Aurora

Part B – Results



80% reduction in false positives

BIS Innovation Hub

Aurora

Part C – Applying PETs



Part B demonstrated the potential **benefits of using machine learning models on shared transaction data to detect complex money laundering**.



Implementing such arrangements could be challenging due to **concerns relating to data protection, privacy, security, competition and legal compliance**.



Part C explores the use of **privacy-enhancing technologies (PETs) to facilitate secure and privacy-preserving CAL (Collaborative Analysis and Learning) arrangements**, and assessed the effectiveness of machine learning models in combination with different PETs.

BIS Innovation Hub

Aurora

Part C – CAL arrangements explored

COLLABORATIVE ANALYSIS & LEARNING (CAL)

CAL 1 (centralised national)

- Data are encrypted and obfuscated using PETs
- Data are shared in a centralised national system.
- Machine learning models are trained and applied at a national-level data.

CAL 2 (centralised cross-border)

- Data are encrypted and obfuscated using PETs
- Data are shared in a centralised cross-border system.
- Machine learning models are trained and applied on data at a cross-border level.

CAL 3 (hybrid national and cross-border)

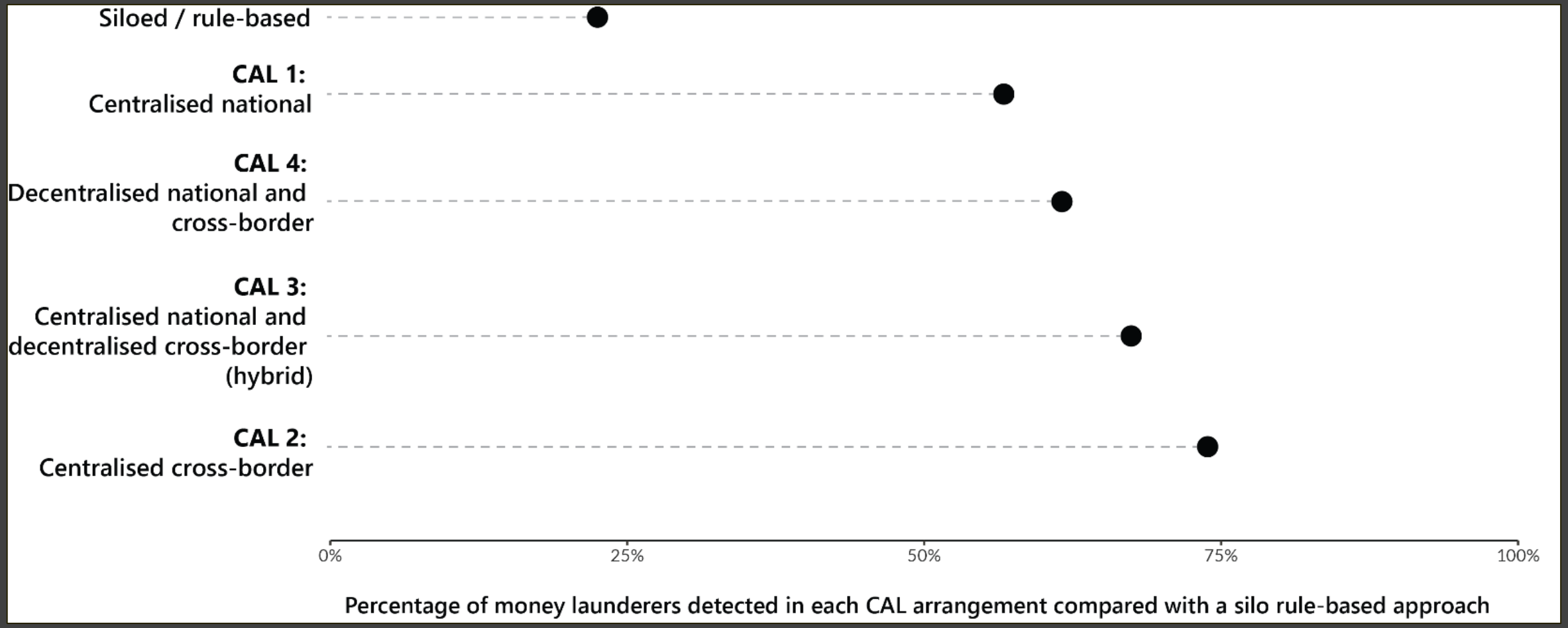
- Data are encrypted and obfuscated using PETs.
- Data are shared into a centralised national system.
- Countries collaborate on training a machine learning model using **federated learning** without the need to share transaction data across borders.

CAL 4 (decentralised national and cross-border)

- participants collaboratively train a machine learning model using **federated learning** locally on their own data
- share the model updates to a common global model

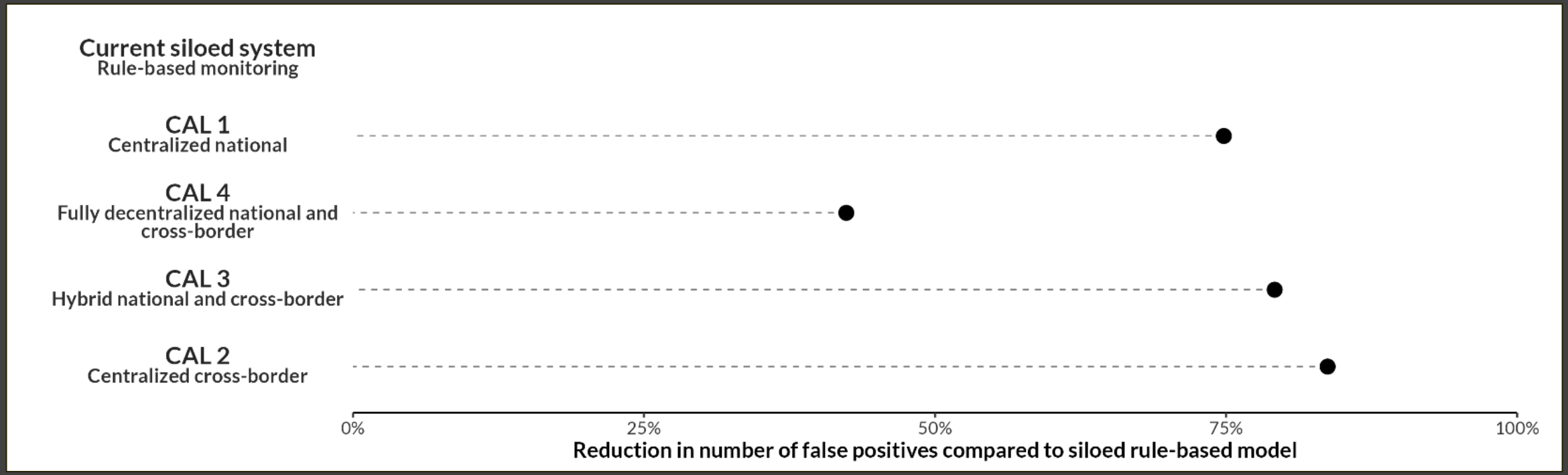
Part C – Results

When **machine learning is combined with PETs** that **performance and effectiveness is maintained**

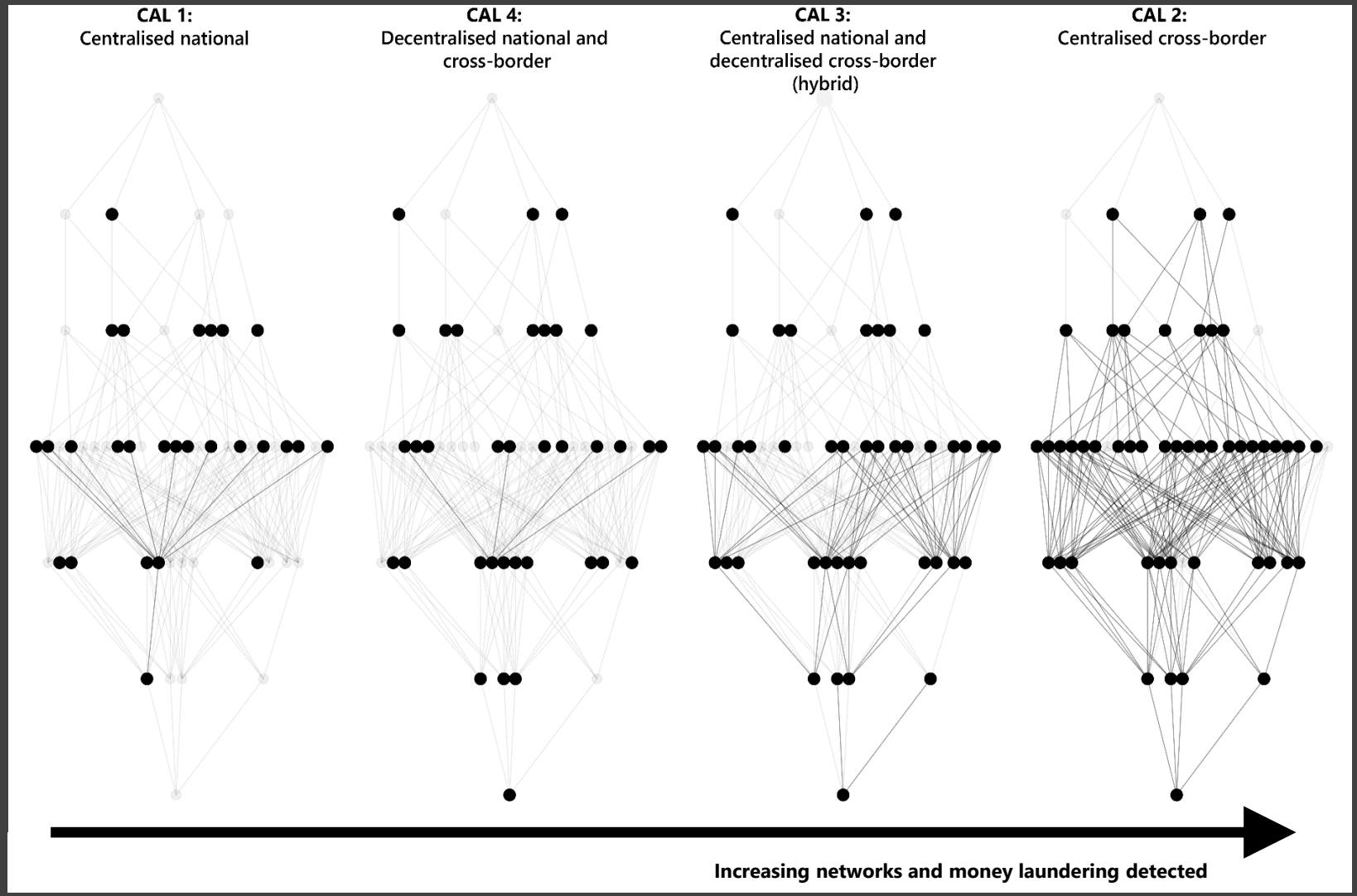


Part C – Results

When **machine learning is combined with PETs** that **performance and effectiveness is maintained**



Part C – Results



The broadest view of data is most performant and effective

CALS vary in performance and effectiveness but all are better than today's silo / rule based

Different data points and sources = different typologies

- Not all money laundering typologies can be detected solely through transaction data analysis using a minimum set of data points.
- Depending on the typology, additional data sources and data points may be required to aid detection
- A small number of examples of this are:

N.B. there are > 11,000 typologies

- **Personal use of business accounts:** detected by monitoring know-your-customer (KYC) data such as ultimate beneficial owner (UBO) labels, tax reports, business expense receipts or through the observation of large transactions from legal entities to individuals.
- **Transacting with politically exposed persons (PEPs):** detected by flagging entities and counterparties of transactions as PEPs. In a collaborative data aggregation setting, this would require the sharing of this flag among the relevant parties.
- **Off-ramping into crypto assets:** money launderers can use these channels to hide their trail of money. Additional information from the VASPs (eg KYC data) and analysis of the relevant blockchain are necessary to detect these typologies.
- **Trade-based money laundering (TBML):** TBML involves the exploitation of international trade to move value around the world, often using complex financial

Data standards example – Legal Entity Identifiers (LEI)

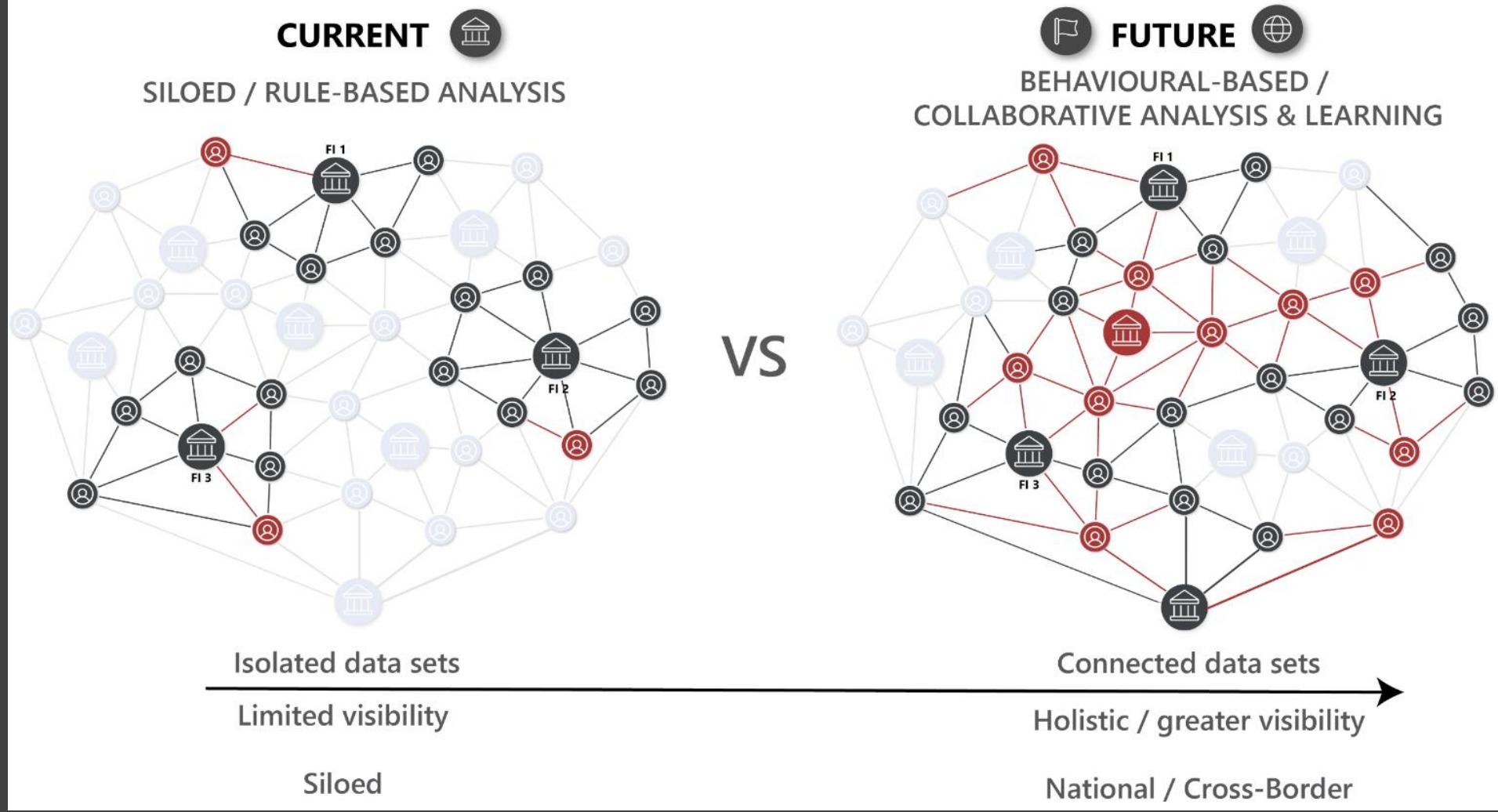
- Legal entities were represented by **name, entity number in the country, country, and address**
- Data constructed to allow **identification and linkage of legal entities**, while **LEIs can easily be utilized as unique identifiers for representing legal entities** as opposed to a generic entity number
- Approach showed the potential benefits of **using unique identifiers such as LEIs to enable effective analysis and detection of money laundering typologies and networks** involving legal entities

- **Enhanced Entity Resolution:** By utilizing unique identifiers like LEIs, network analysis could potentially accurately link and track entities across multiple transactions and financial networks. This could enable a comprehensive view of the connections between different entities involved in money laundering activities.
- **Uncovering Complex Relationships:** Graph structures built using identifiers could allow for the visualization and exploration of complex relationships within a network. These structures reveal hidden patterns, connections, and dependencies that can aid in identifying intricate money laundering networks that may be difficult to detect through traditional analysis methods.
- **Real-Time Monitoring and Predictive Analytics:** Identifiers could enable proactive detection of emerging money laundering networks and predictive insights.

BIS Innovation Hub

Aurora

What the future could be



BIS Innovation Hub

Aurora

Conclusion

CAL approaches using AI / machine-learning models and PETs showed greater effectiveness in detecting money launderers and suspicious networks compared to siloed and rule-based monitoring. Could be **game changing**.

PETs could enable secure and privacy-preserving CAL arrangements. Investment and development of PETs and PPC is advancing exponentially and in time commoditized and scalable services will be available. Many applications.

Collaboration between the private and public sectors is crucial for innovation in AML efforts. **Cross-functional dialogue is needed** to support the development and / or adoption of data standards (e.g. LEI, ISO20022, UBO)

Real-world proofs of concept (or pilots) would be needed to test the feasibility and effectiveness of the technologies and CAL approaches at scale. Necessary to surface legal, regulatory, data protection, and technical issues and questions.

New technologies, such as **Knowledge Graphs,** could be explored further with public and private sector experts as a public good exercise, which could enable more better standardised information sharing between the private the public sectors.

A data driven approach is essential. Though some data would be common, **different typologies require different data fields / data sets** – some of which may come from other public or private sources. **Collaboration is key.**


BIS Innovation Hub
Aurora
Questions

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Why SDMX matters? A community journey towards SDMX as an AI-enabler and a data mesh enabler[1]

Eric Anvar,
Organisation for Economic Co-operation and Development (OECD)

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Why SDMX matters? A community journey towards SDMX as an AI-enabler and a Data Mesh enabler

IFC Workshop "Data science in central banking: Enhancing the access to and sharing of data"

Rome, October 2023

Keywords
#OpenSource #OpenKnowledge #Community #SDMX #MetadataDriven #AI #DataMesh



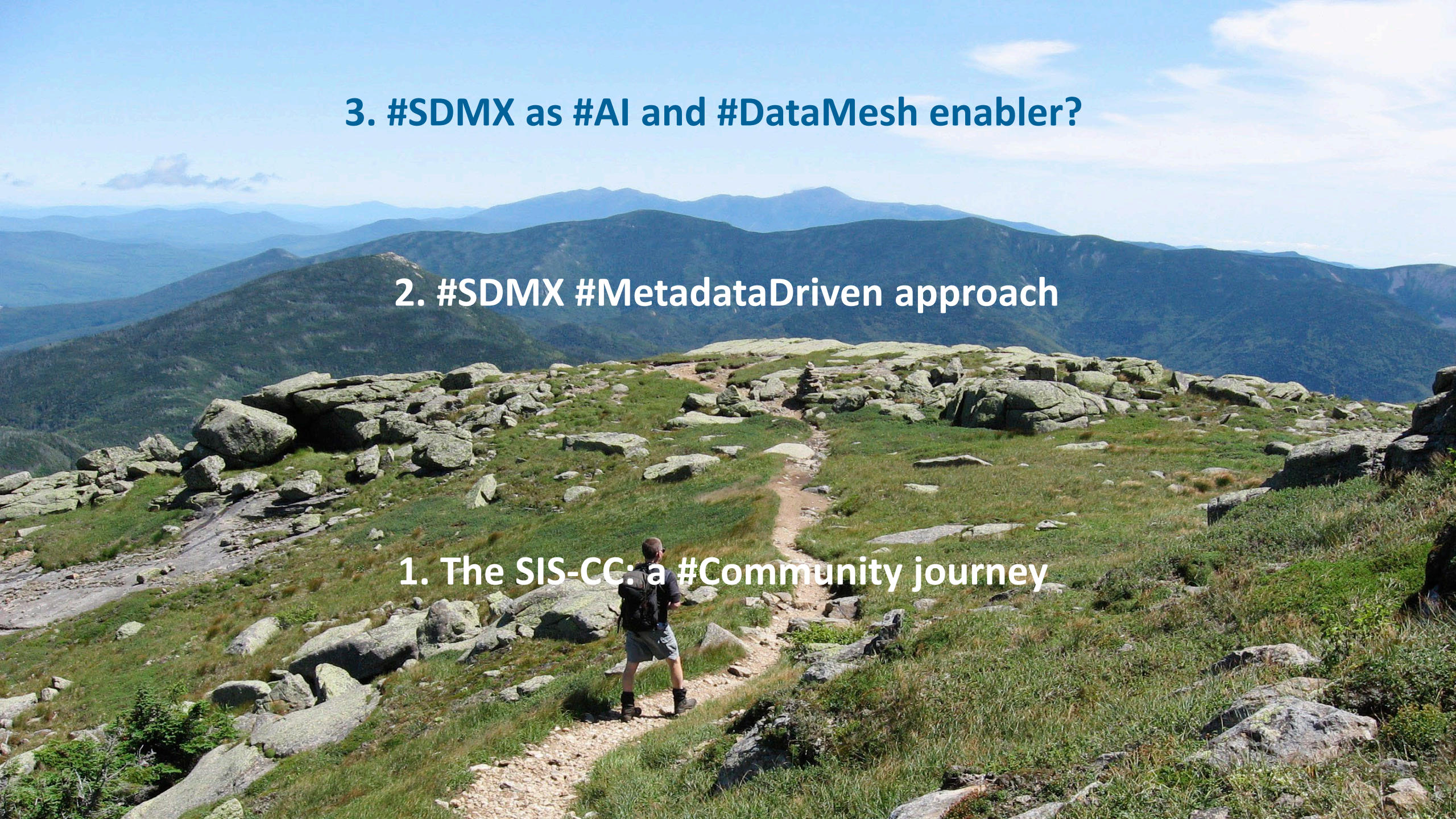
SIS-CC
Statistical Information System Collaboration Community



Perhaps there is more to SDMX than just a tool to report to international organisations?



Source: Why SDMX Matters, https://siscc.org


3. #SDMX as #AI and #DataMesh enabler?
2. #SDMX #MetadataDriven approach
1. The SIS-CC: a #Community journey

SIS-CC
Statistical Information System
Collaboration Community

THE SIS-CC: A #COMMUNITY JOURNEY

SIS-CC: 12 years, 20 Members and Partners

SIS-CC ("the Community") is a reference open-source community for official statistics, focusing on product excellence and **delivering concrete solutions to common problems through co-investment and co-innovation**.

Statistics Canada / Statistique Canada

Federal Competitiveness and Statistics Centre

Pacific Community / Communauté du Pacifique

INE Instituto Nacional de Estadísticas · Chile

SDMX.IO

Food and Agriculture Organization of the United Nations

STATEC

إحصائيات تونس / Statistiques Tunisie

ILO

UK Data Service

United Nations Statistics Division

BIS

National Bank of Belgium / Eurosystem

unicef

Australian Bureau of Statistics

Stats NZ / Tatauranga Aotearoa

OECD Better Policies for Better Lives

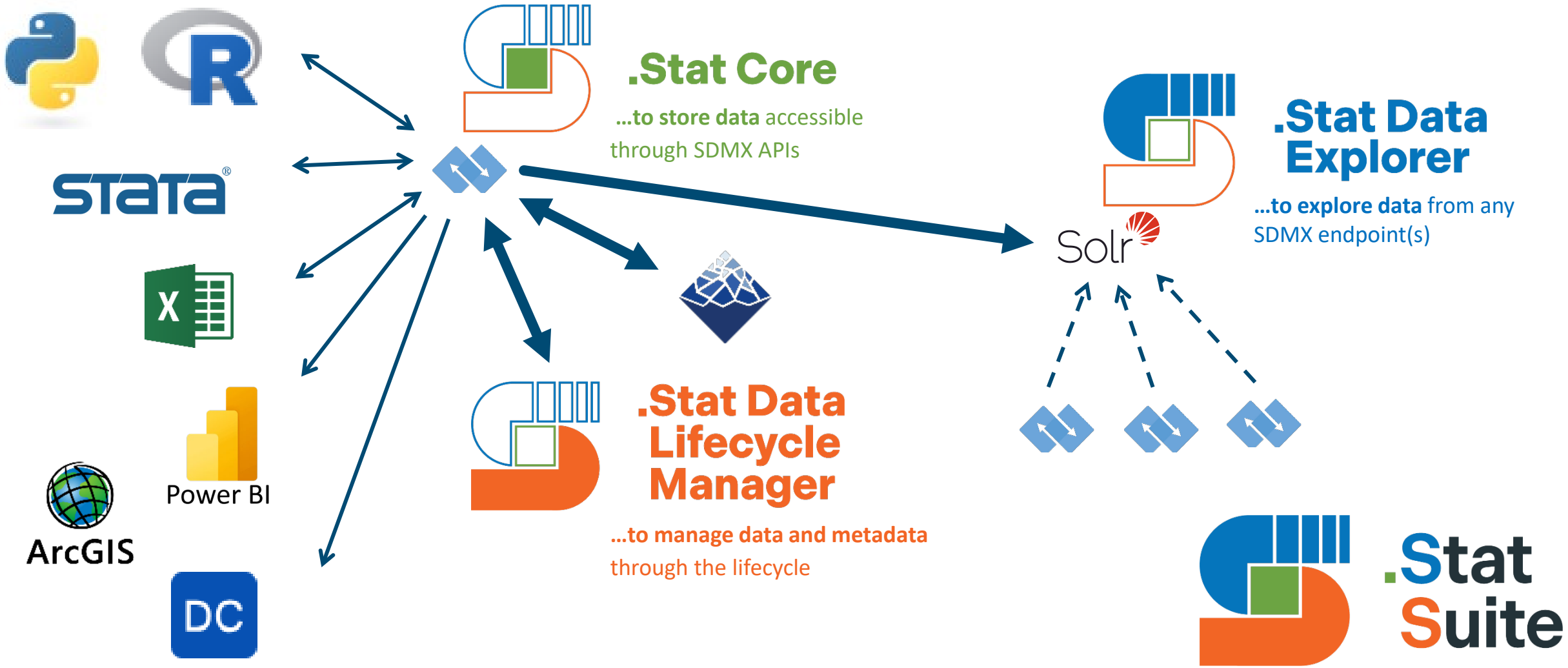
eurostat

Members

Partners

The #OpenSource .*Stat Suite*

SIS-CC
Statistical Information System Collaboration Community



.Stat Core

...to store data accessible through SDMX APIs

.Stat Data Explorer

...to explore data from any SDMX endpoint(s)

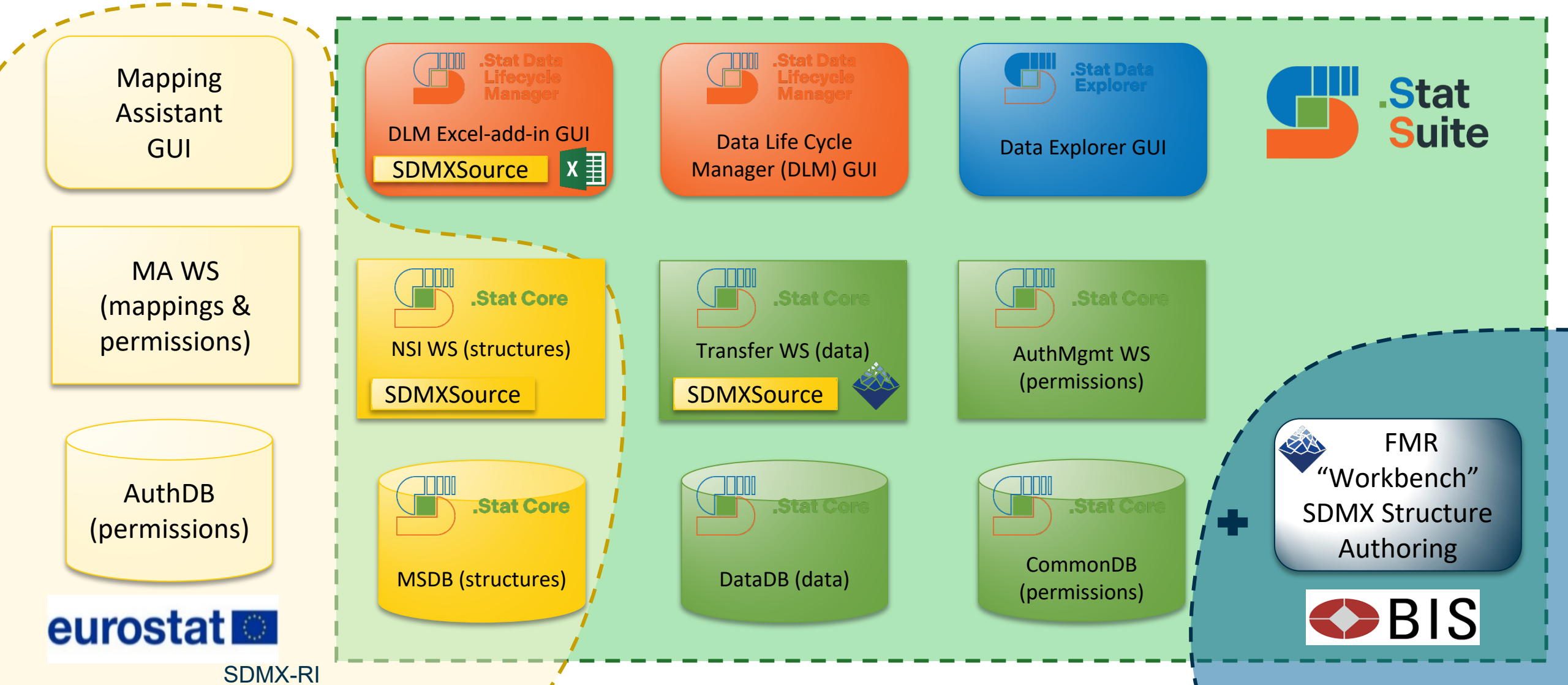
.Stat Data Lifecycle Manager

...to manage data and metadata through the lifecycle

.Stat Suite

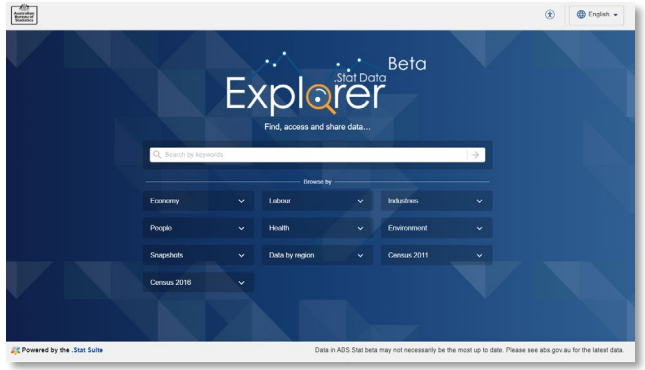
Component-based Architecture ➔ Cooperative "Business Model"

SIS-CC Statistical Information System Collaboration Community

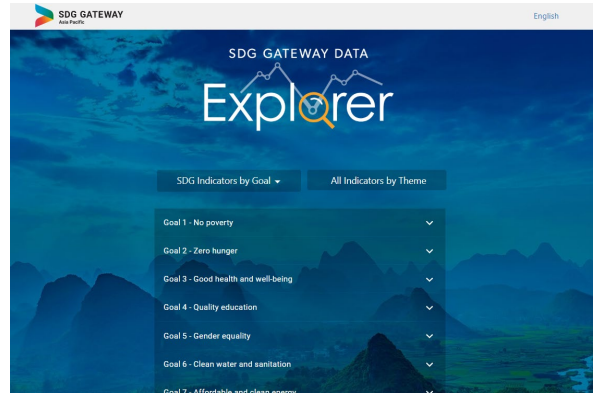


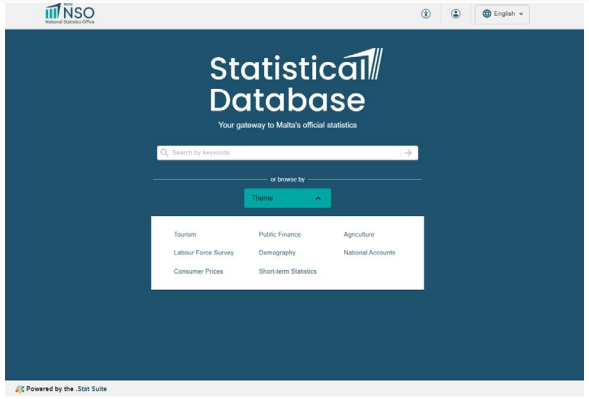
Examples of implementations

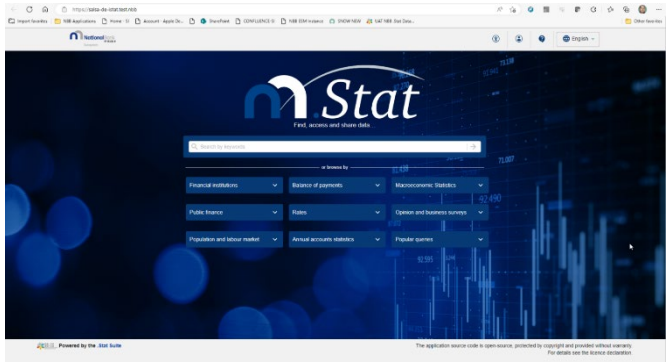
SIS-CC
Statistical Information System Collaboration Community

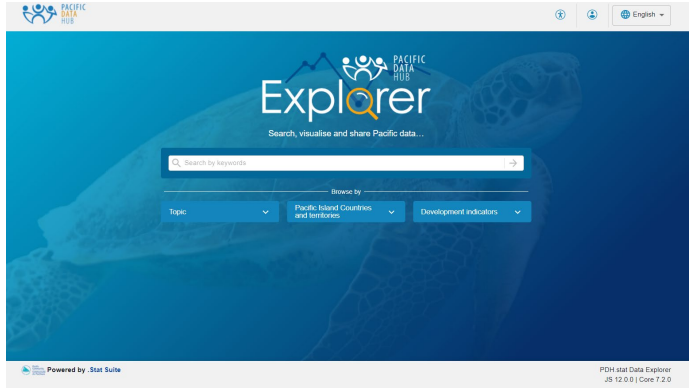




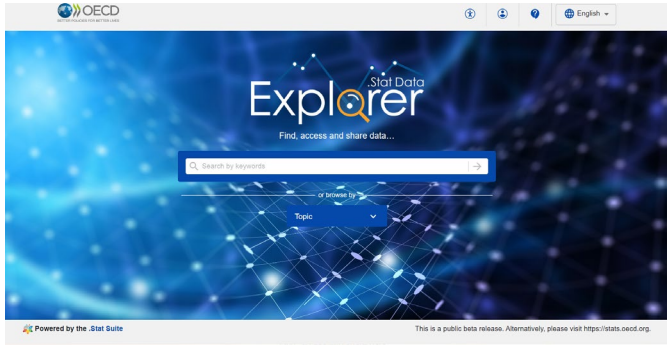












.Stat Suite

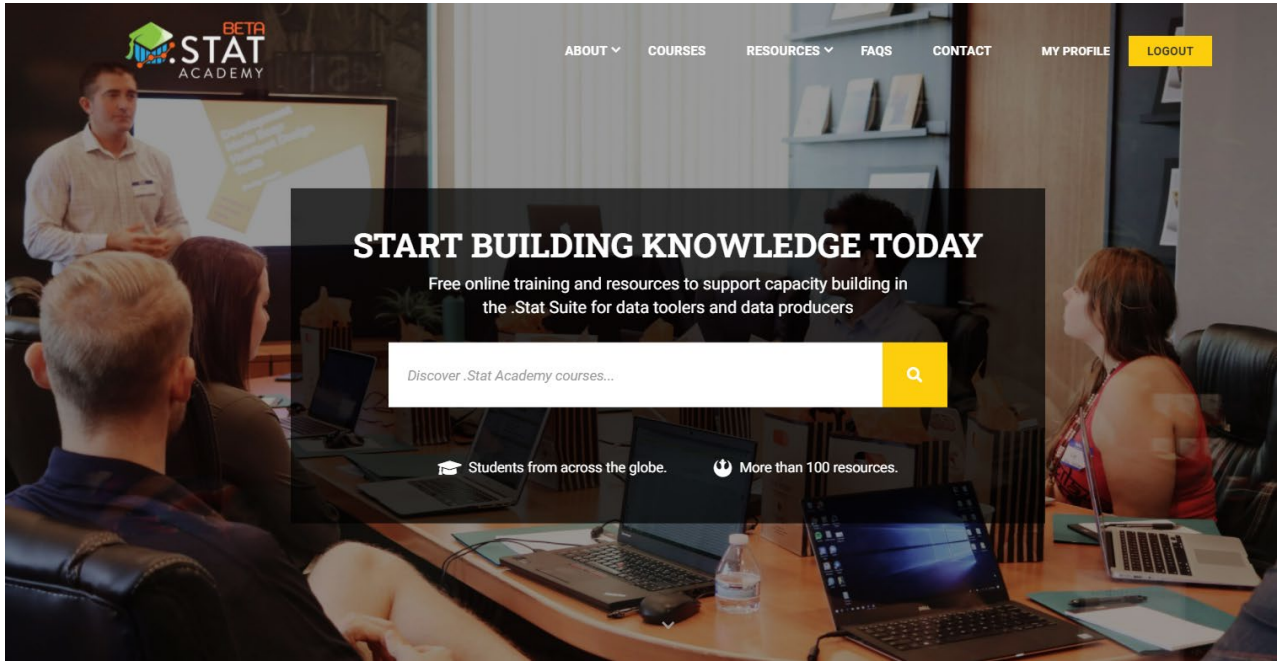


Search, visualise and share Pacific data...

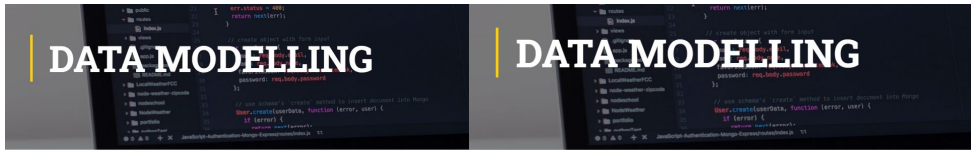
Pacific Data Hub .STAT

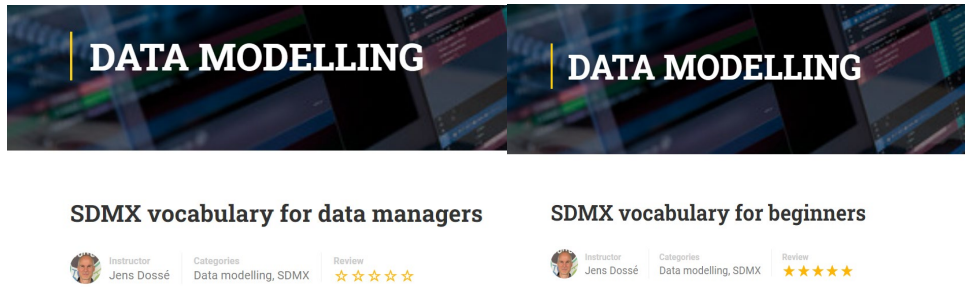
The #OpenKnowledge .Stat Academy

SIS-CC
Statistical Information System Collaboration Community



https://academy.siscc.org





Starting point: Build capacity in SDMX data modelling

.STAT ACADEMY

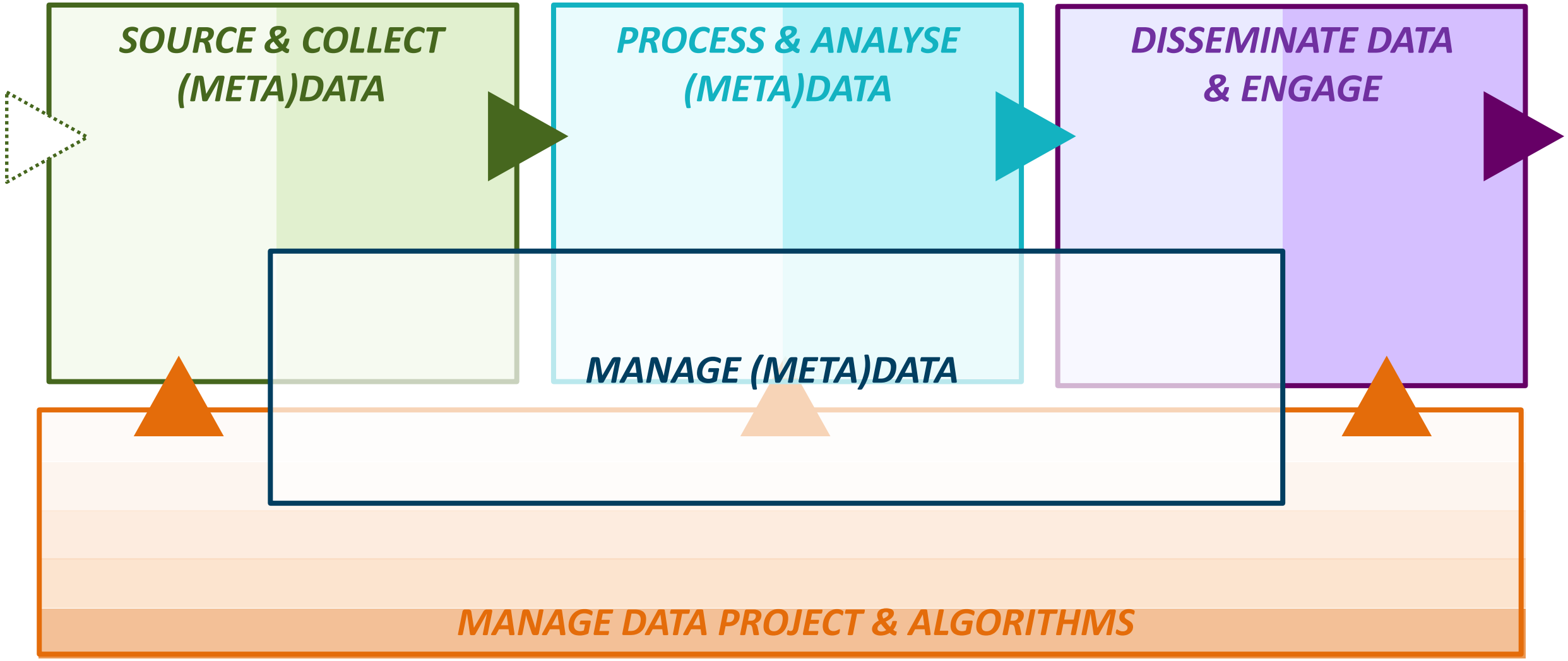
SIS-CC
Statistical Information System
Collaboration Community

THE #SDMX METADATA-DRIVEN APPROACH



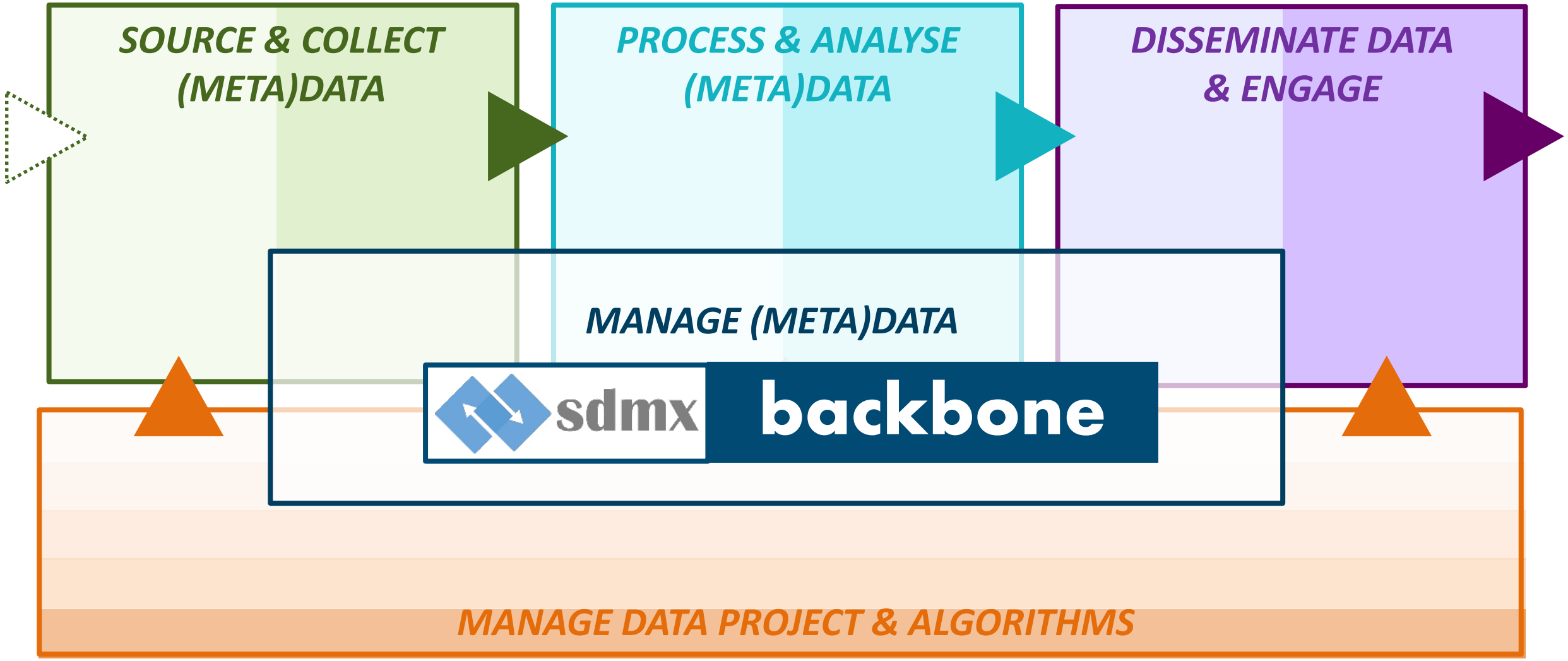
The Data Platform

SIS-CC
Statistical Information System Collaboration Community

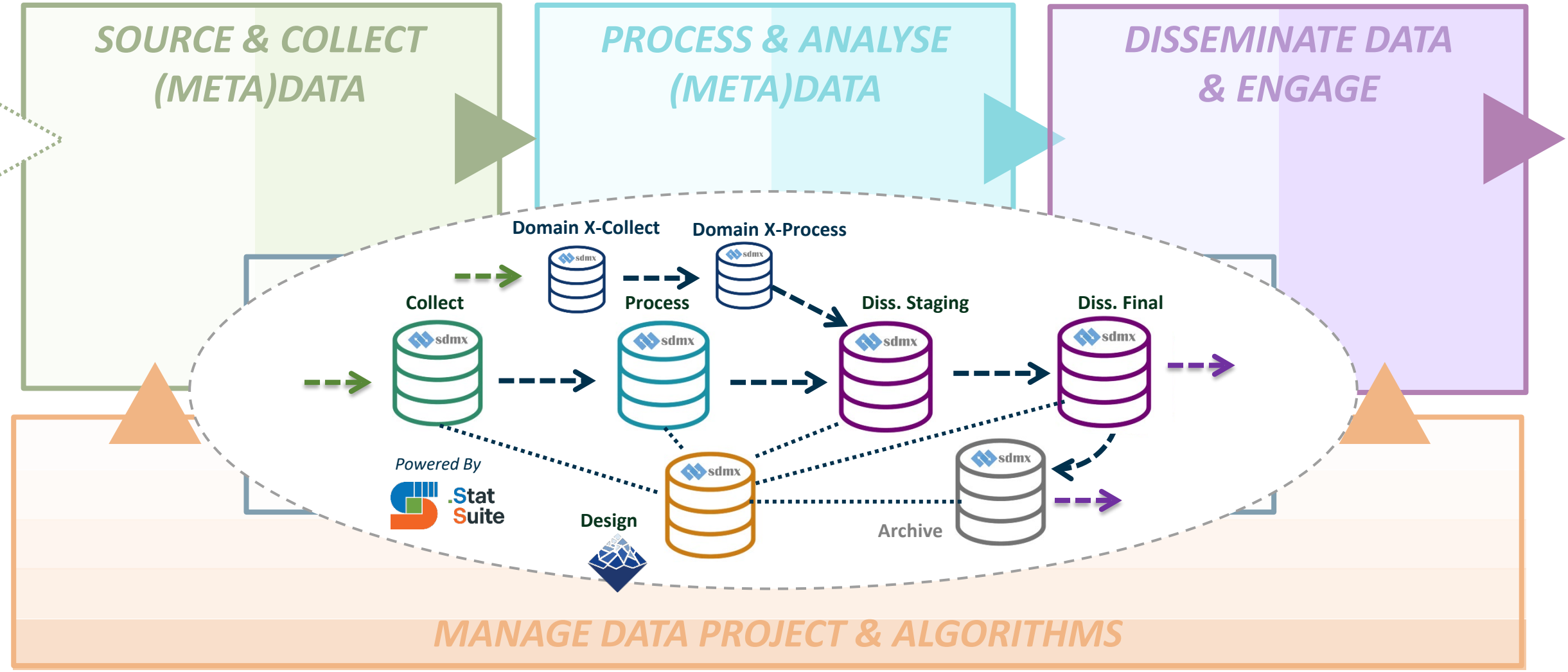


The #SDMX Backbone

SIS-CC
Statistical Information System Collaboration Community



The #SDMX Backbone

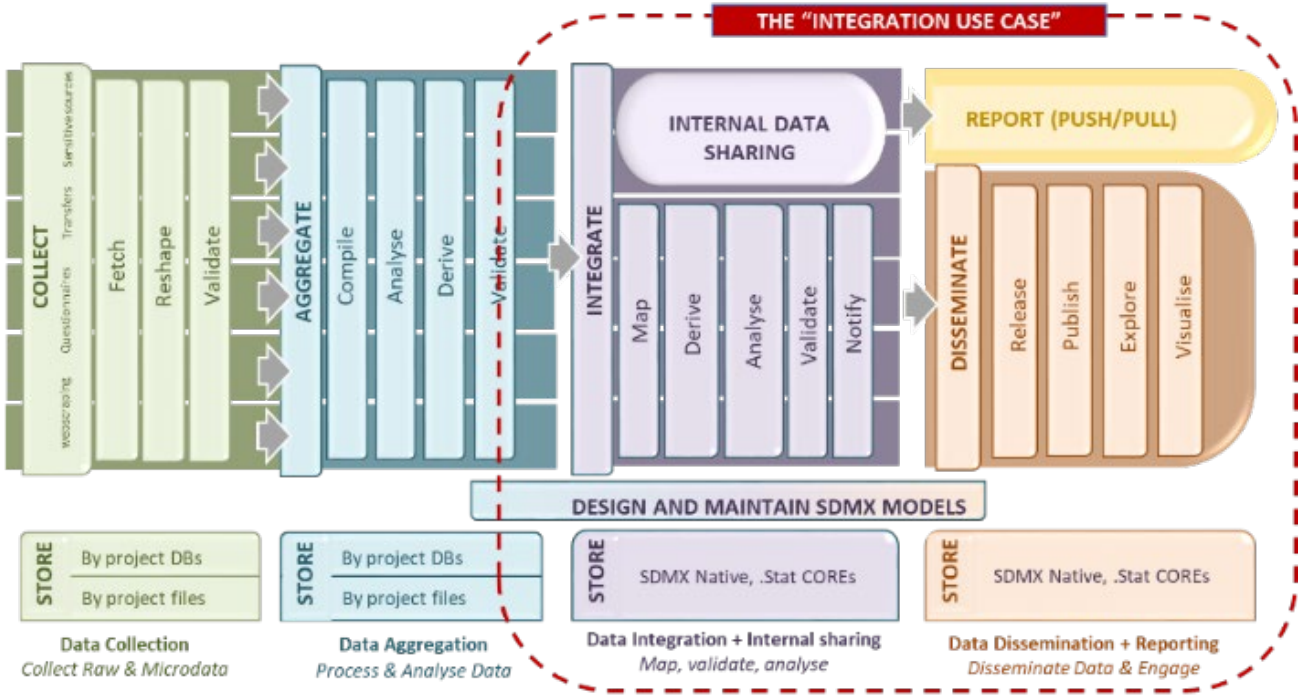


The #SDMX Backbone

SIS-CC Statistical Information System Collaboration Community

The "Data Integration" use case at the National Bank of Belgium

National Bank of Belgium Eurosystem

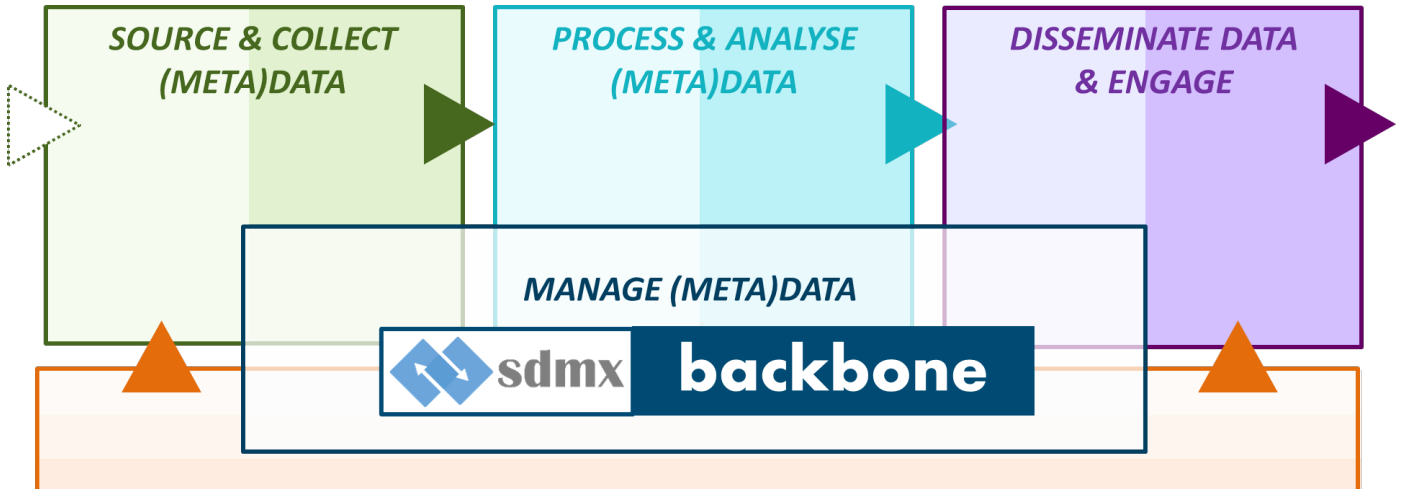


European Commission | OECD BETTER POLICIES FOR BETTER LIVES | BIS

Source: SALSA Project – National Bank of Belgium

The #SDMX Backbone – Wrap-up

- **sdmx data storage becomes a commodity**: on premise and over the cloud, and across the entire data lifecycle and business domains.
- **sdmx APIs enable universal data interoperability**: microservice orientation, component-based architecture, cooperative business model
- **sdmx semantics enables "metadata-driven" data user / manager experience:** structural & referential metadata, presentation & process metadata



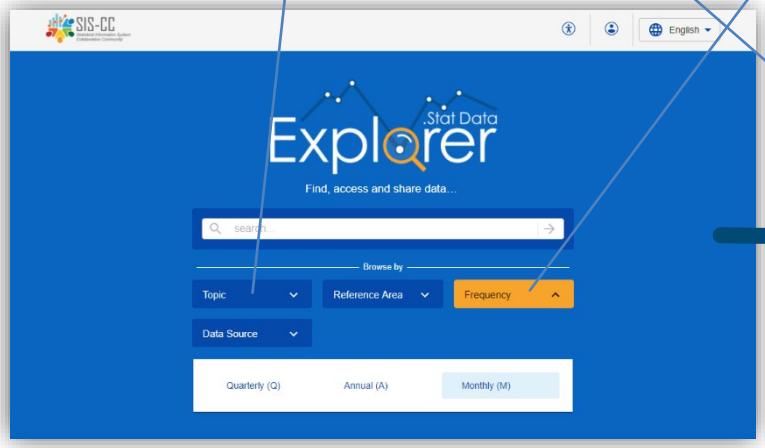
#SDMX #MetadataDriven Data User Experience

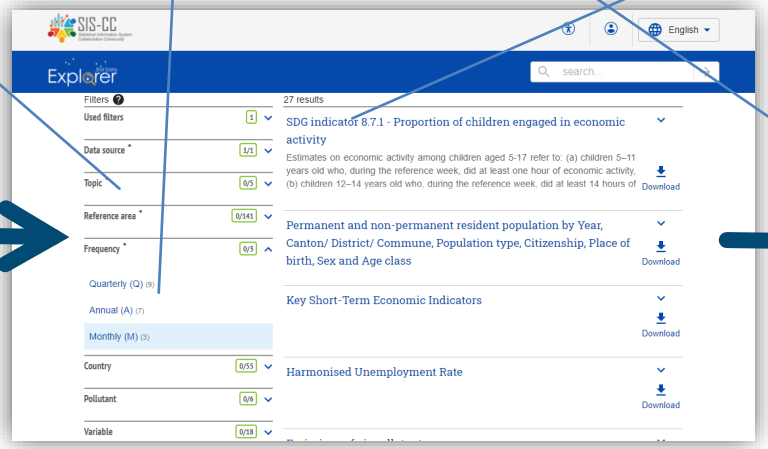
SIS-CC
Statistical Information System Collaboration Community

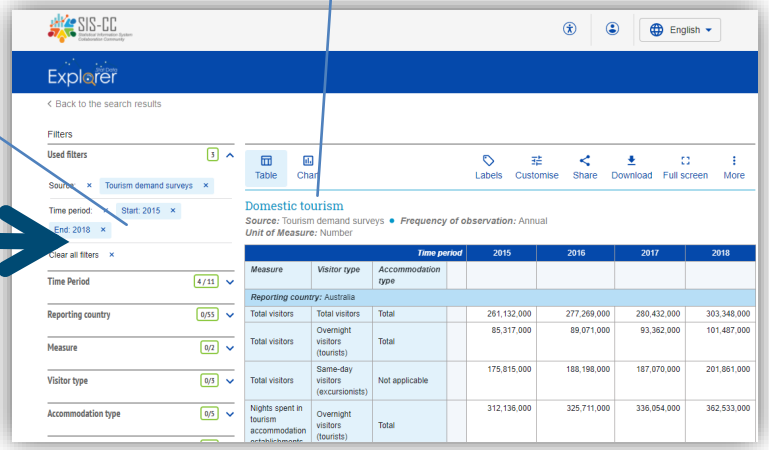
category scheme(s)

dimensions:
concepts + codelists

dataflow
+ presentation annotations





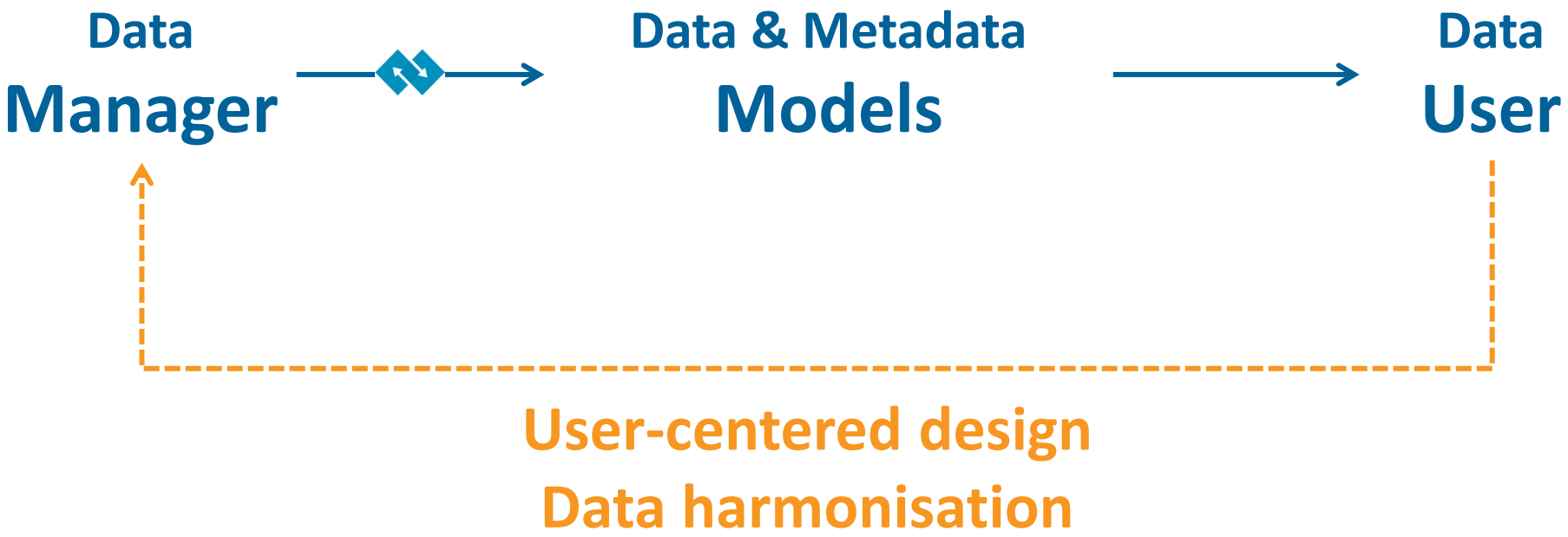


Step 1 Home page
Lexical search or Topics

Step 2 Search results
Filtering on dynamic facets

Step 3 Visualisation page
Selecting dimensions / variables

#SDMX #MetadataDriven Data User Experience

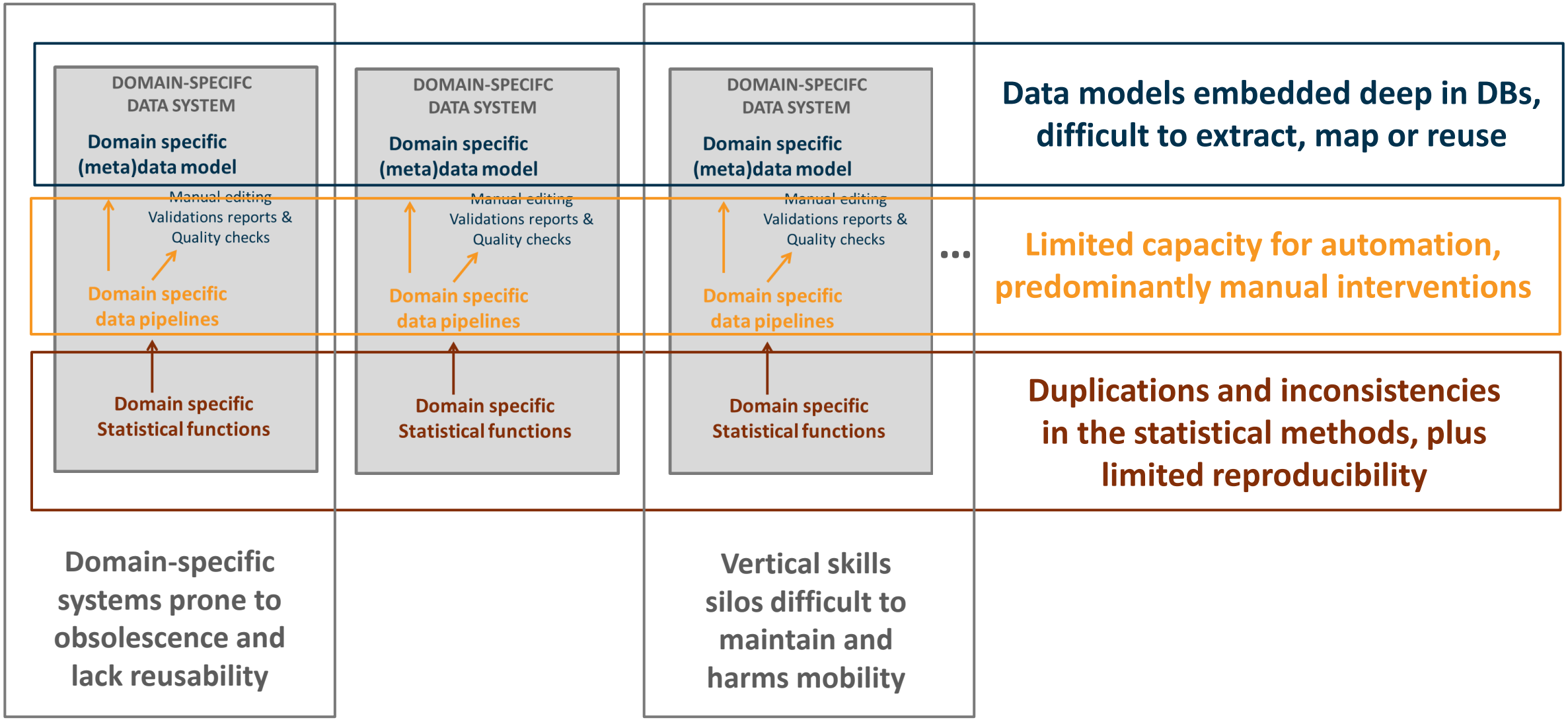


Establish a systematic User Research



Photo by Celpax on Unsplash

The Data Platform is fragmented



#SDMX #MetadataDriven Data Manager Experience

SIS-CC Statistical Information System Collaboration Community

MANAGE DATA & METADATA

MANAGE DATA PROJECT & ALGORITHMS

Data Analysis

sdmx **Data Product** ----> Power BI

Discoverable, Understandable, Interoperable, Globally addressable

Data Management

sdmx **Structural Metadata**

Categories, DSDs/MSDs, Dataflows

sdmx **Referential Metadata**

Data quality, Methodology

sdmx **Process Metadata (?)**

Extraction, Transformation Metadata

Validations reports & Quality checks

Data Engineering

Metadata-driven pipelines

Manipulating SDMX data

Extraction, Harmonisation, Chained calculations, Derived indicators

Data Transformation

Metadata-driven statistical functions

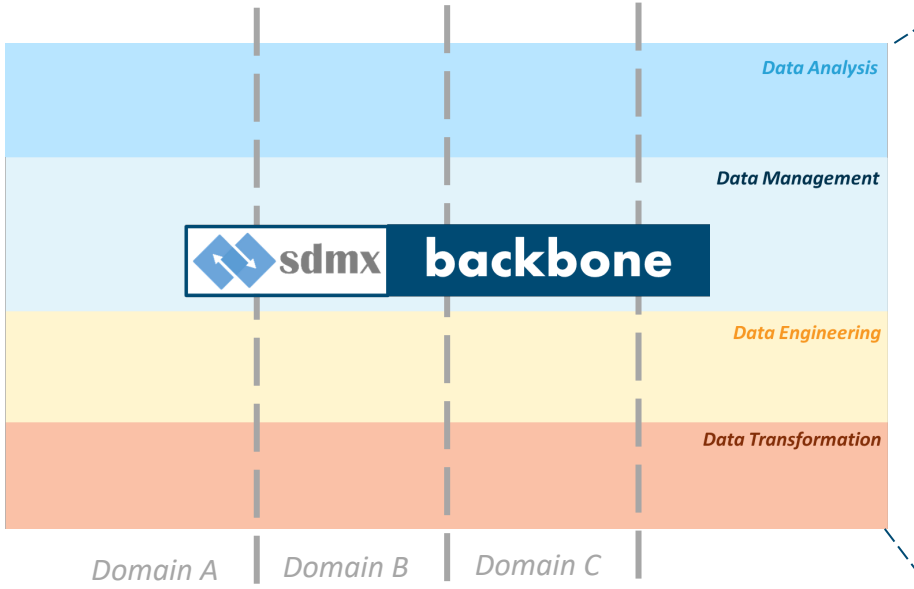
Manipulating SDMX data

Interpolation, Chain linking, Aggregation functions, Seasonal adjustments, Smoothing, Rebasement, Backcasting, Forecasting, Growth rates

(?) feature that exists partially / peacemeal in SDMX and that ought to be specified

#SDMX #MetadataDriven Data Manager Experience

The #SDMX #MetaDataDriven approach combines (centrally managed) data infrastructure with the possibility of (decentralized) data operations, creating efficiencies at both ends



- Decoupled roles and specializations
- Data mgt by non coding-experts
- Facilitation of staff mobility
- Possibility of data tasks mutualization
- Facilitation of platform migration
- Domain specific coding language
- Domain specific interfaces
- Capitalize on shared methods / reuse
- Capitalize on common infrastructure

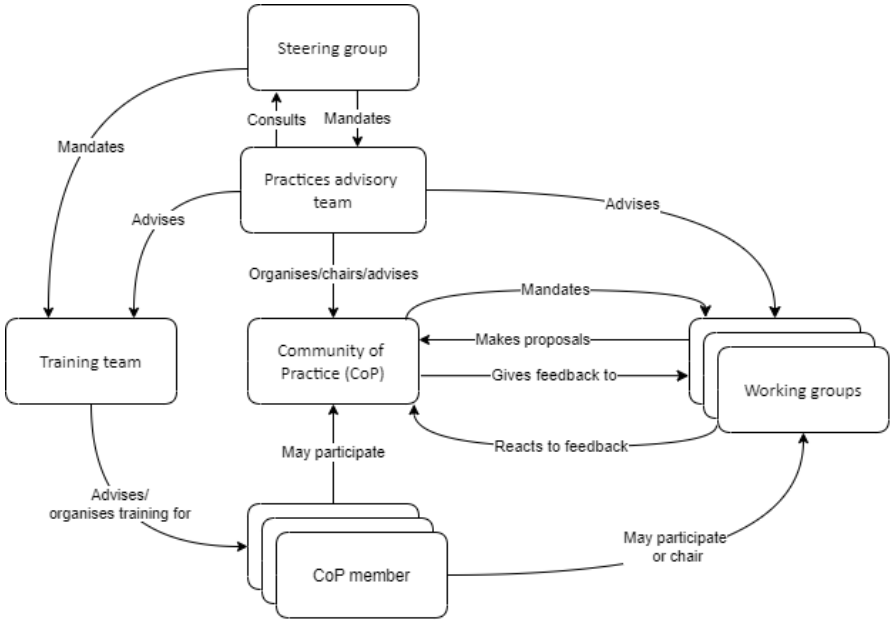


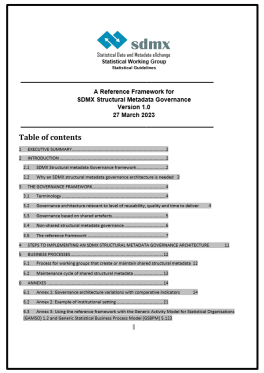
Structure your Data governance

Photo by UX Indonesia on Unsplash

#SDMX enables Data Governance

SDMX provides the tools (conceptual and technical) to operate your governance and progressively harmonise data and metadata models





A Reference Framework for SDMX Structural Metadata Governance
Version 1.0
27 March 2022

Table of contents

A Reference Framework for SDMX Structural Metadata Governance, March 2023

#SDMX #MetadataDriven approach – Wrap-up

sdmx governance

- Systematic user research – enabled by SDMX
- Structured data governance – enabled by SDMX

- SDMX data storage becomes a commodity
- SDMX APIs enable universal data interoperability
- SDMX semantics enables “metadata-drivenness”

sdmx backbone

SDMX metadata drives
Data User experience
Data Manager experience



Interested in developing an #SDMX Backbone & Governance?

Contact me – eric.anvar@oecd.org

SIS-CC
Statistical Information System
Collaboration Community

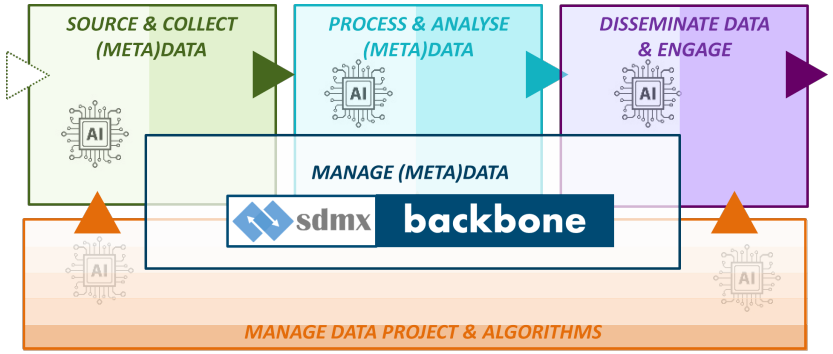
#SDMX #AI
SDMX AS AN AI ENABLER?

Opportunities #SDMX #AI

SIS-CC
Statistical Information System Collaboration Community

Nowcasting
Federated Learning

Outlier detection
Clustering

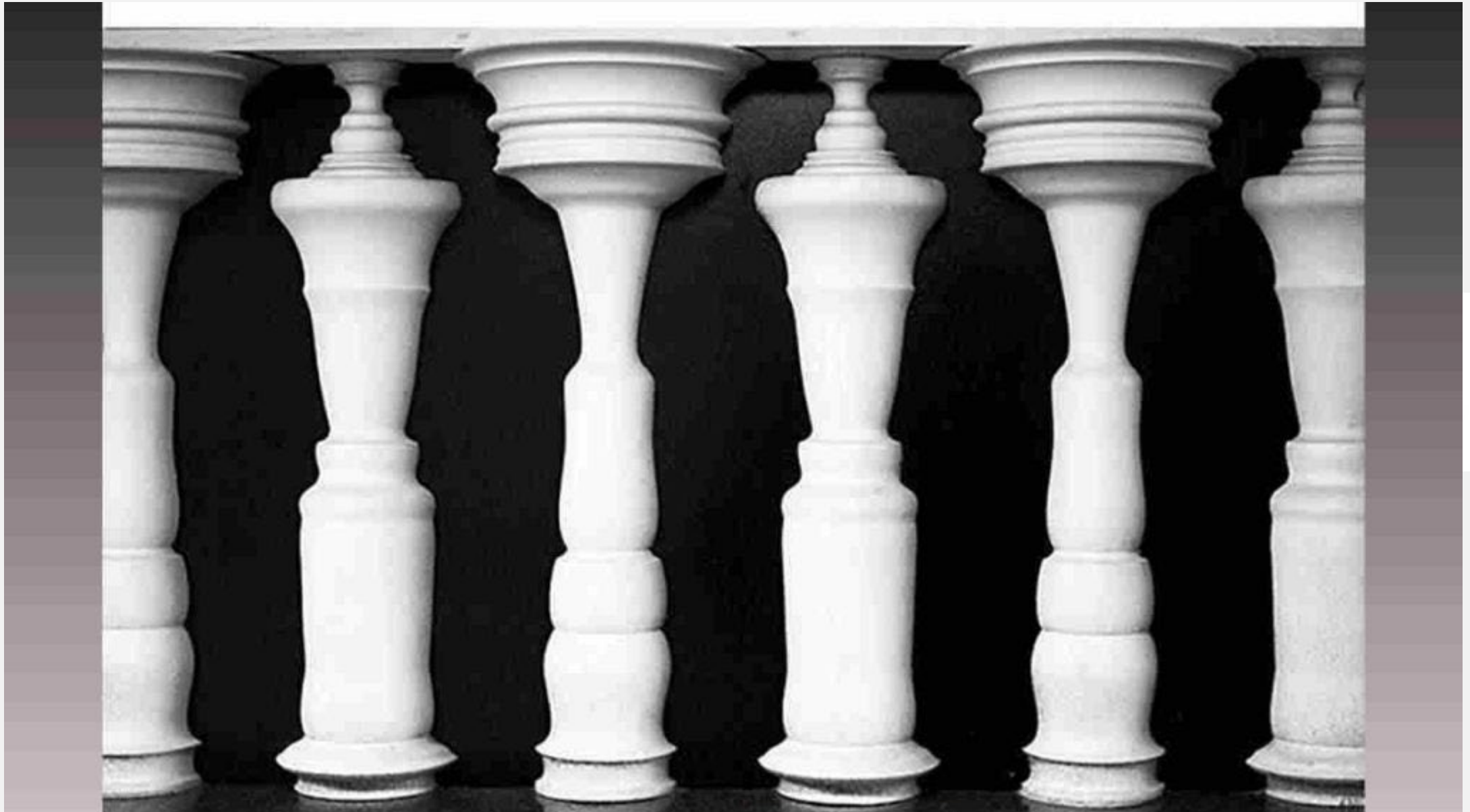


Natural Language
Conversational
Data Access

Code generation
Data Modelling

The cognitive gap

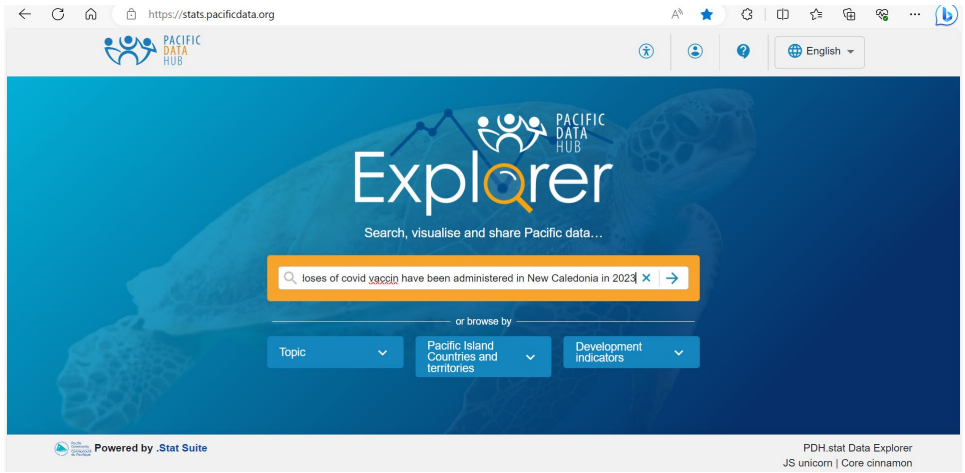
SIS-CC
Statistical Information System
Collaboration Community

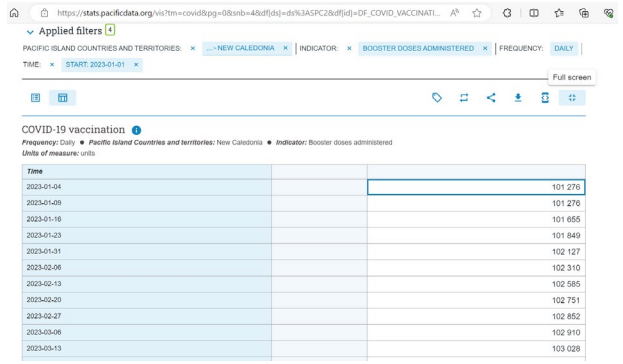


Augmenting the Data Explorer with #SDMX #AI (I)

SIS-CC Statistical Information System Collaboration Community

How many booster doses of covid vaccins administered in New Caledonia in 2023?





Step 2.5 or 3.5

Natural Language query transformed into search queries
+ Pre-selection of dimensions on target dataflow(s)

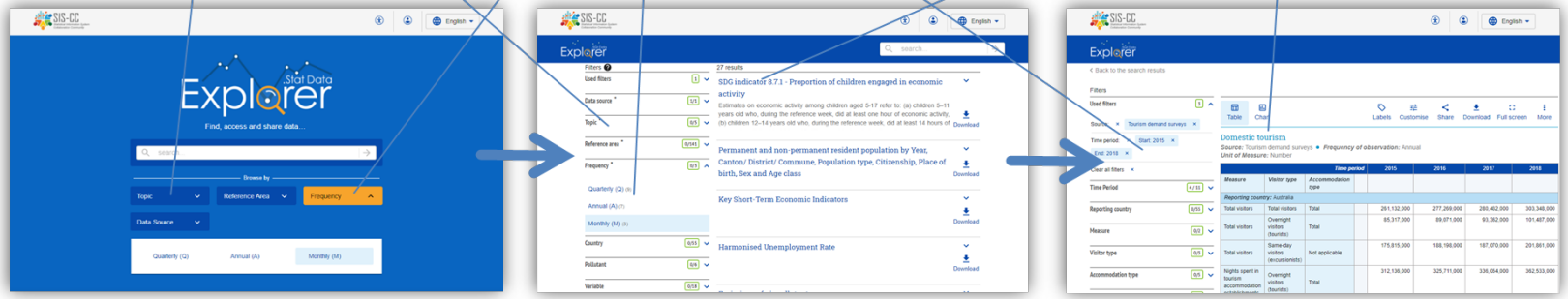
Augmenting the Data Explorer with #SDMX #AI (I)

SIS-CC
Statistical Information System Collaboration Community

category scheme(s)

dimensions:
concepts + codelists

dataflow
+ presentation annotations



Step 1 Home page
Lexical search or Topics

Step 2 Search results
Filtering on dynamic facets

Step 2.5

Natural Language query transformed into engine search queries + initial filter selection

Step 3 Visualisation page
Selecting dimensions / variables

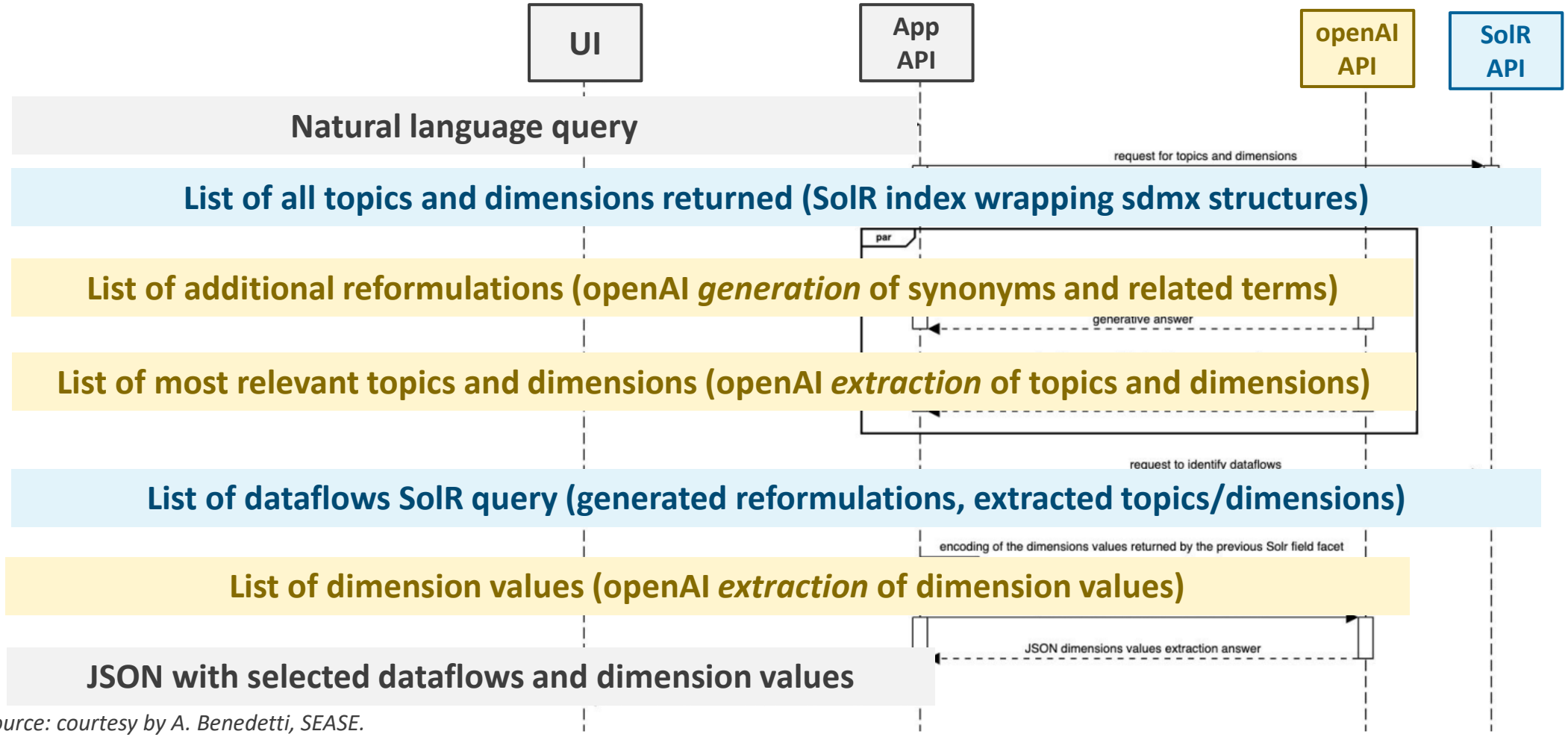
Step 3.5

Direct identification of the relevant dataflow with a pre-selection of dimensions

Augmenting the Data Explorer with #SDMX #AI (I)

SIS-CC Statistical Information System Collaboration Community

Proof of Concept ongoing, mixing Generative AI & SolR techniques.

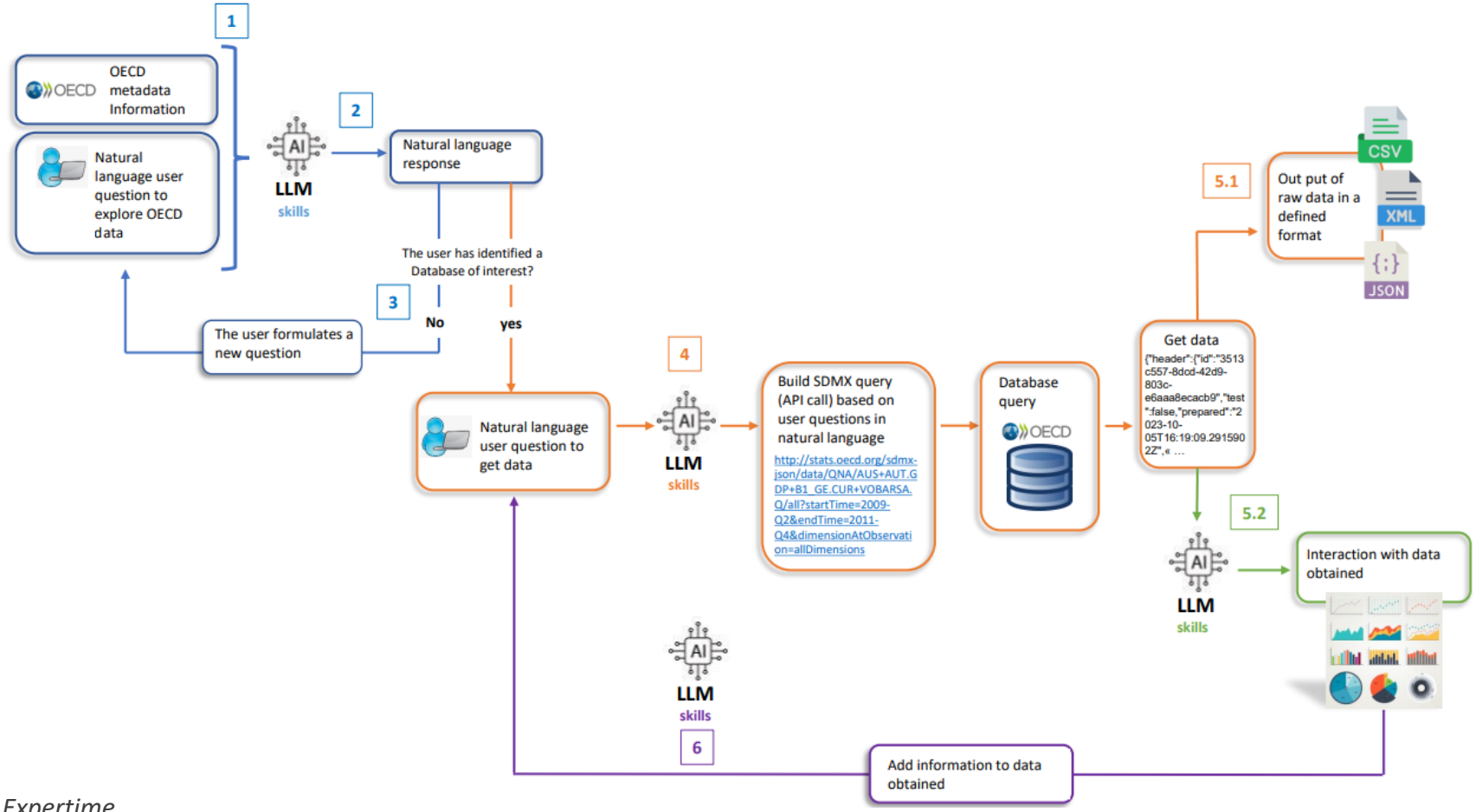


Source: courtesy by A. Benedetti, SEASE.

Replacing the Data Explorer with #SDMX #AI (II)

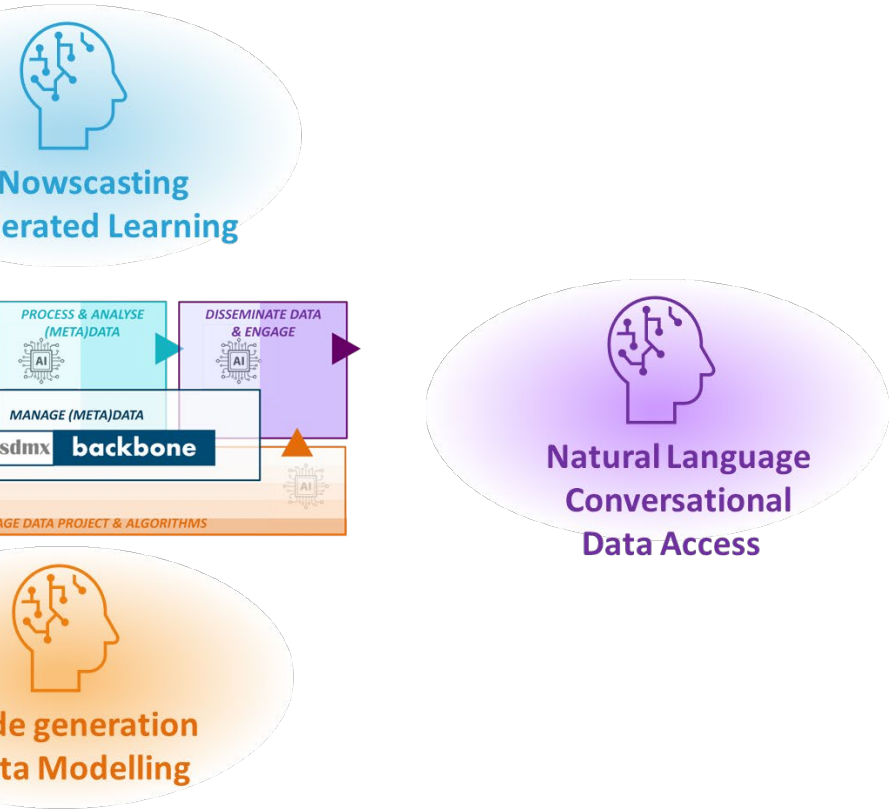
SIS-CC
Statistical Information System Collaboration Community

The StatsBot complements / replaces the Data Explorer with a conversational data assistant



Source: courtesy by K. Castillo Perez, Expertime.

Creating #SDMX #AI open and reusable assets



- ***Natural language capable Data Explorer (I)***
- ***Universal StatsBot to access any SDMX source through a conversation (II)***
- ***Capacity building – Prompt engineering practice in SDMX context***
- ***Specialized, "SDMX-literate" open source LLM***

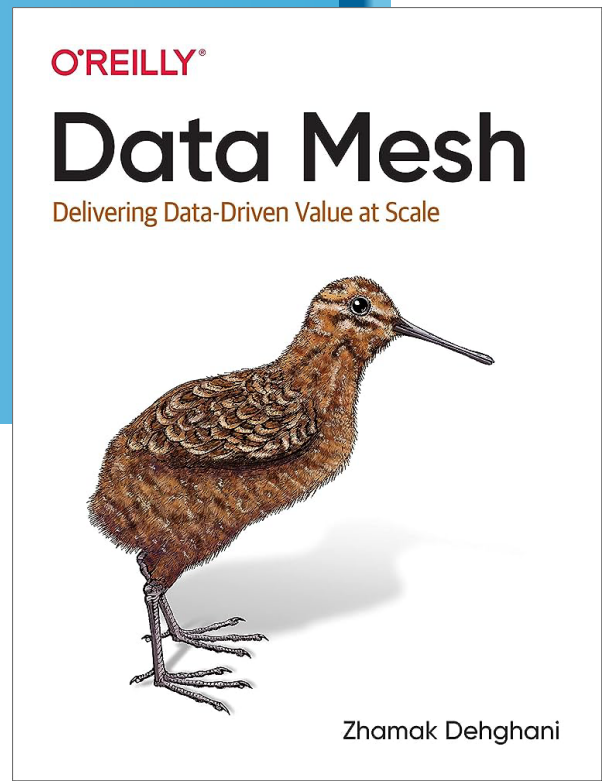


Interested in co-investing in #SDMX #AI open assets?

Contact me – eric.anvar@oecd.org

SIS-CC
Statistical Information System Collaboration Community

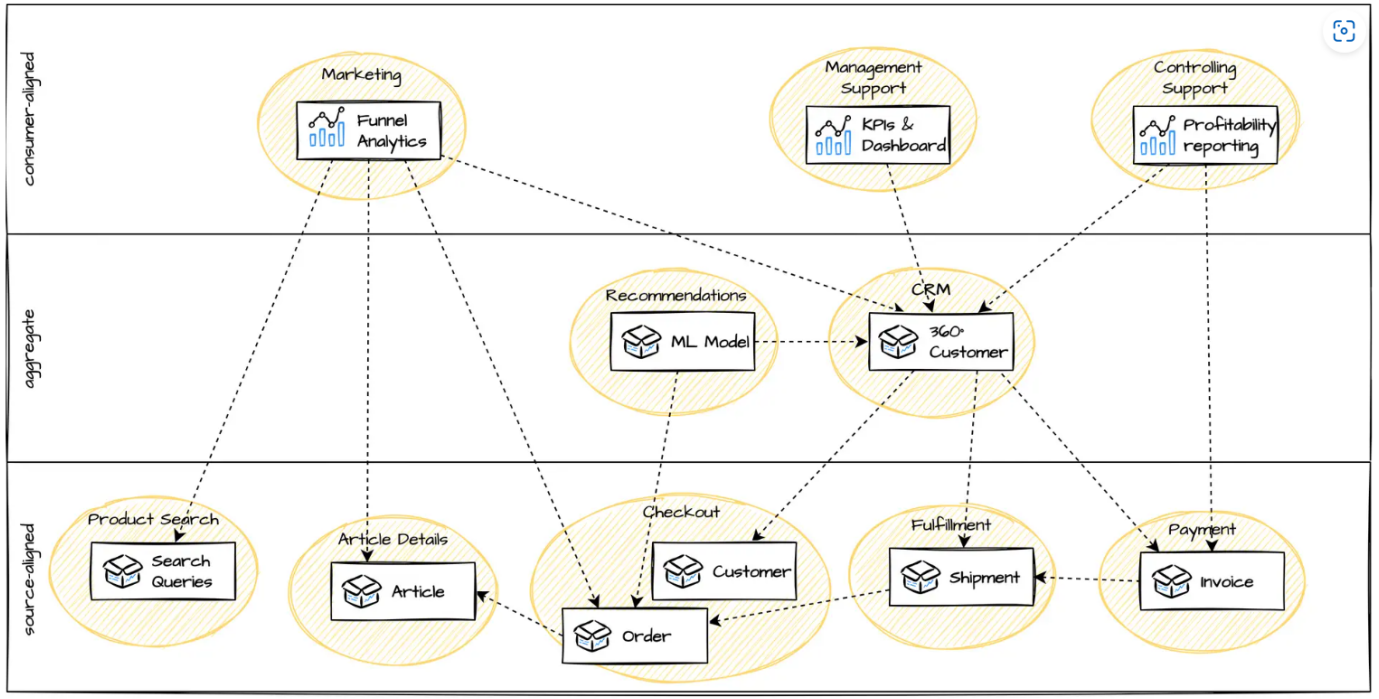
#SDMX #DATAMESH
SDMX AS A DATA MESH ENABLER?



What is a #DataMesh?

Address the problem with the centralized organization behind the DWH/Data Lake approach to analytical data: centralized expertise often leads to bottleneck and failure.

The #DataMesh is the socio-technical paradigm to switch from a purely centralized model to a model that enables domain-oriented decentralization of analytical data operations.



Source: Data Mesh Architecture (datamesh-architecture.com)

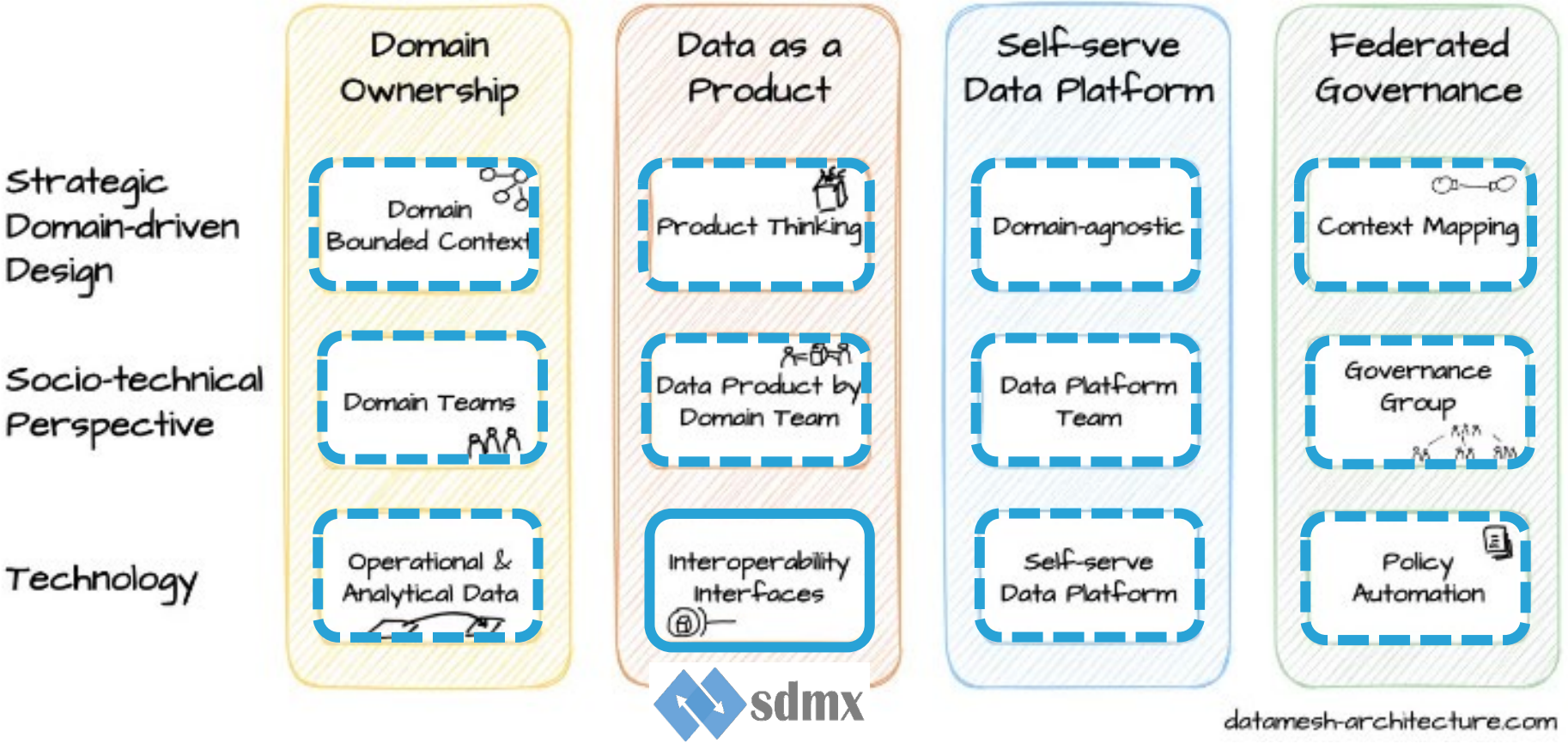
#DataMesh requires common data semantics

SIS-CC
Statistical Information System Collaboration Community

*"Data mesh attempts to **strike a balance between team autonomy and interteam interoperability and collaboration**. It gives domain teams autonomy to take control of their local decision making, for example, **choosing the best data model for their data products**, while it uses computational governance policies to impose a consistent experience across all data products, for example, **standardising the data modelling language that all domains utilize**. Data mesh gives domain teams autonomy to build and maintain their data products, while it places a **domain-agnostic data platform in place for teams** to do so in a consistent and cost-effective way."*

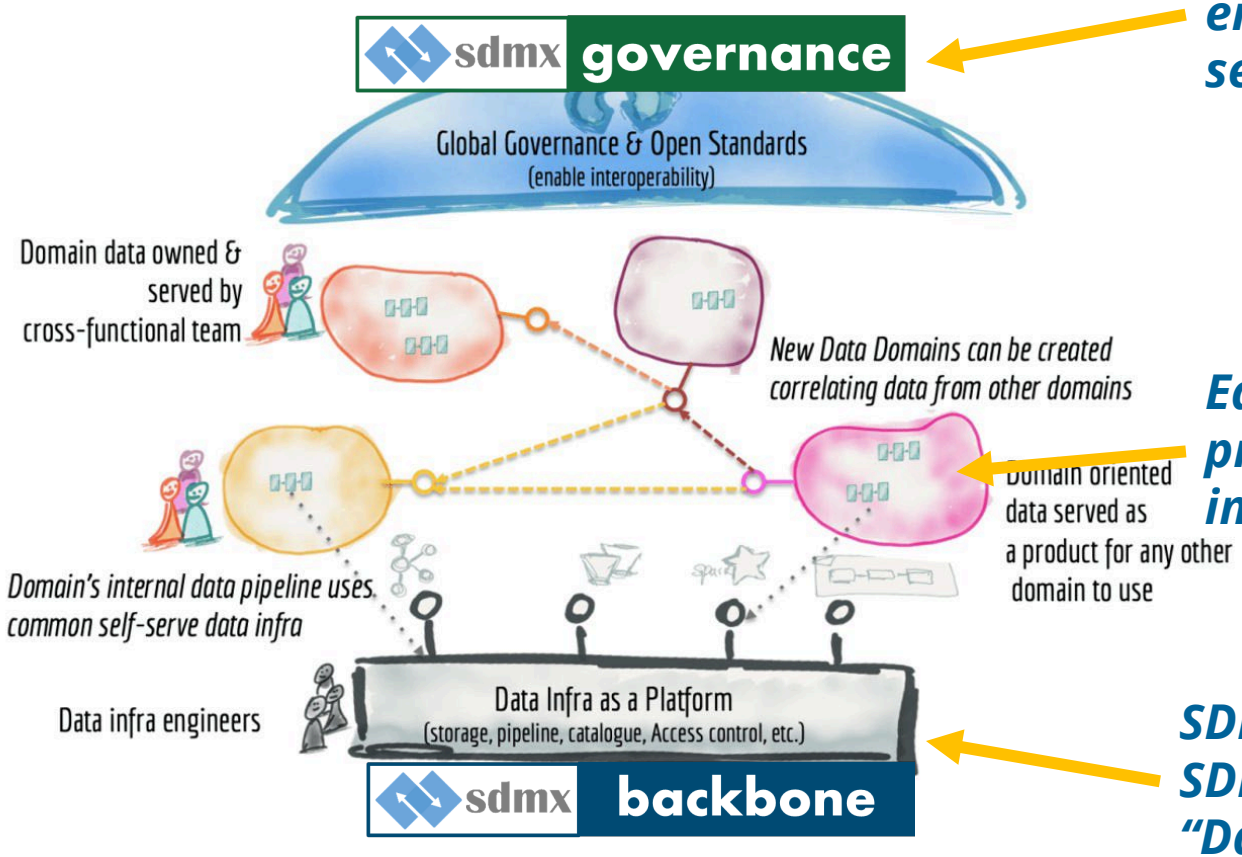
Zhamak Dehghani, Data Mesh

#SDMX as a #DataMesh enabler?



Source: Data Mesh Architecture (datamesh-architecture.com)

#SDMX as a #DataMesh enabler?



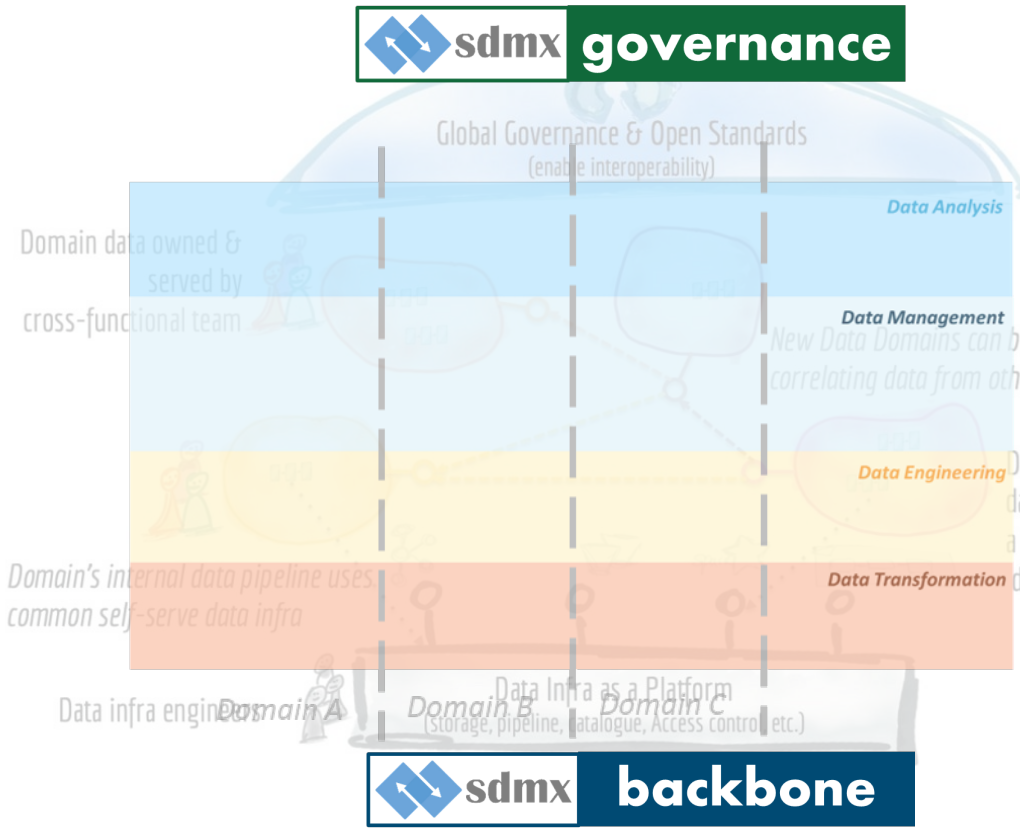
Source: Data Mesh, O'Reilly, Z. Dehghani

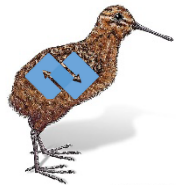
SDMX governance as cornerstone of data governance, enabling a progressive harmonisation of data semantics across domains and the data lifecycle.

Each business domain to build their data process and products in autonomy, drawing on common information models and technical building blocks.

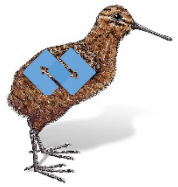
SDMX backbone brings commoditised SDMX storage, SDMX API-centricity, and SDMX semantics as part of the "Data infrastructure as a Platform".

Creating #SDMX #DataMesh open and reusable assets



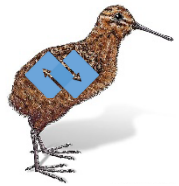


Develop the architecture vision & business case for #DataMesh in Official Statistics

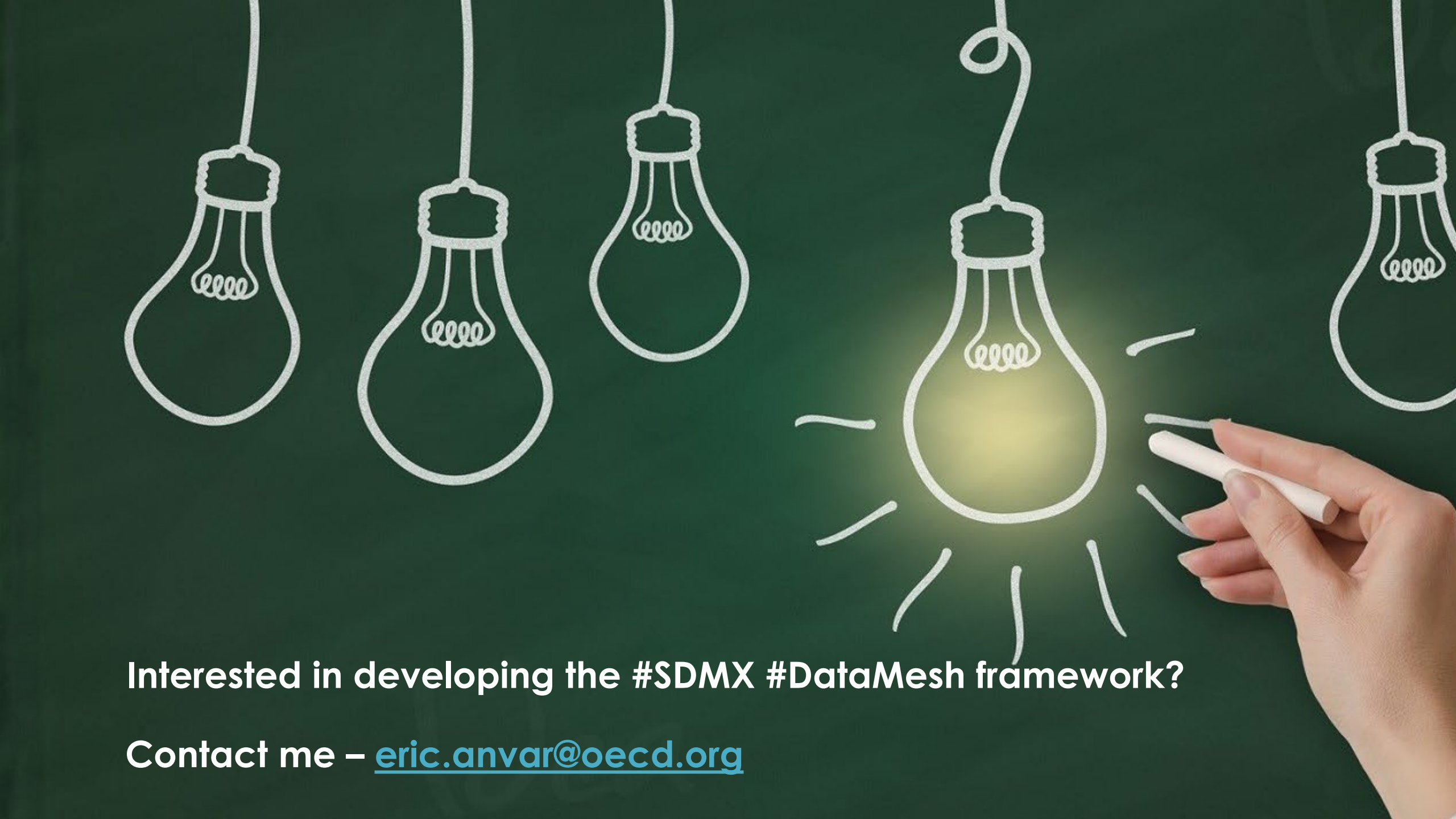


Adapt and augment #SDMX (and #VTL?) with the #DataMesh paradigm

(data product, data contract, DDD, computational governance…)



Develop implementation guidelines for the #SDMX #DataMesh & experiemnt



Interested in developing the #SDMX #DataMesh framework?

Contact me – eric.anvar@oecd.org

Reference resources

SIS-CCommunity

- SIS-CC Community
- SIS-CC 2020-25 Strategy
- Why SDMX Matters

SDMX.IO Ecosystem

- The SDMX.IO ecosystem

SDMX Community

- SDMX Community
- SDMX 2021-25 Strategy
- The business case for SDMX
- SDMX 3.0 packages under review
- SDMX Guideline: *Modelling statistical domains and exchange frameworks in SDMX*
- SDMX Guidelines: *A Reference Framework for SDMX Structural Metadata Governance*

.Stat Suite, open source platform

- .Stat Suite – High level product overview and flight planner
- .Stat Suite – Kanban Board
- .Stat Suite – Changelog
- .Stat Suite – Extensive documentation
- .Stat Suite – Code repository

.Stat Academy, open knowledge platform

- .Stat Academy – Learning Paths for Data Producer; Data Tooler
- .Stat Academy – Courses
- .Stat Academy – Webinars
- .Stat Academy – Become an instructor

3. #SDMX as #AI and #DataMesh enabler?



2. #SDMX #MetadataDriven approach

sdmx governance

sdmx backbone

1. The SIS-CC: a #Community journey

SIS-CC
Statistical Information System Collaboration Community

.Stat Suite

.STAT ACADEMY

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Data science in the context of international banking statistics[1]

Jacob Ewertzh and John Svanäng,
Sveriges Riksbank

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Data science in the context of international banking statistics

Rome (2023-10-17)

Jacob Ewertzh, Data scientist

John Svanäng, Economist

Sveriges riksbank

Background to international banking statistics in Sweden

- New producers of this statistical product
- International banking statistics ”pilot” for data science at the Riksbank
- Why this transformation?
  - Where did we start?
  - What are the benefits?
  - What is the long-term objective?

Results of the project

- Fast and assured
- Rapid implementation of new ideas
- Lower cost than expected



Ideas for enhancing data handling

- Open source approach



- Continuous development



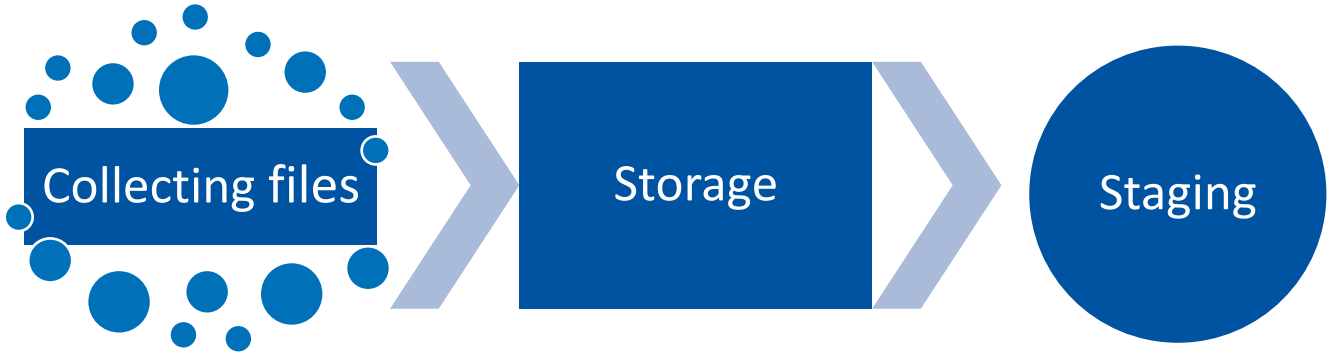
- Collaboration



How we produce international banking statistics

- Data collection
- Processing, Review & Dissemination
- Evaluation

How we produce international banking statistics - Data collection



- Validation rules - *Data quality*
- Feedback loops - *Fast response*

- Original files stored - *No "manual" processing*
- Version history - *Reproducibility*

Fast deployment of changes

- Updating schemas - *Continuous development*

How we produce international banking statistics – Processing, Review & Dissemination

Staging
- Convert microdata
- Enrich with metadata

Inter-mediate

Various calculations

Marts
- Publication
- External deliveries
- End users

Automated output:
- For internal users
- Graphs & tables
- Webpage
- External stakeholders



- Audit & Control
- Confidentiality

Micro-data
- Reuse of data: i.e. loans and securities.

How we produce international banking statistics - Evaluation

- Agile way of working
- Small single team
- New products for end users
- Continuous development
  - Ability to create and modify code
  - Fast, secure and independent
  - Will effectively implement new ideas



Tools for continuous development

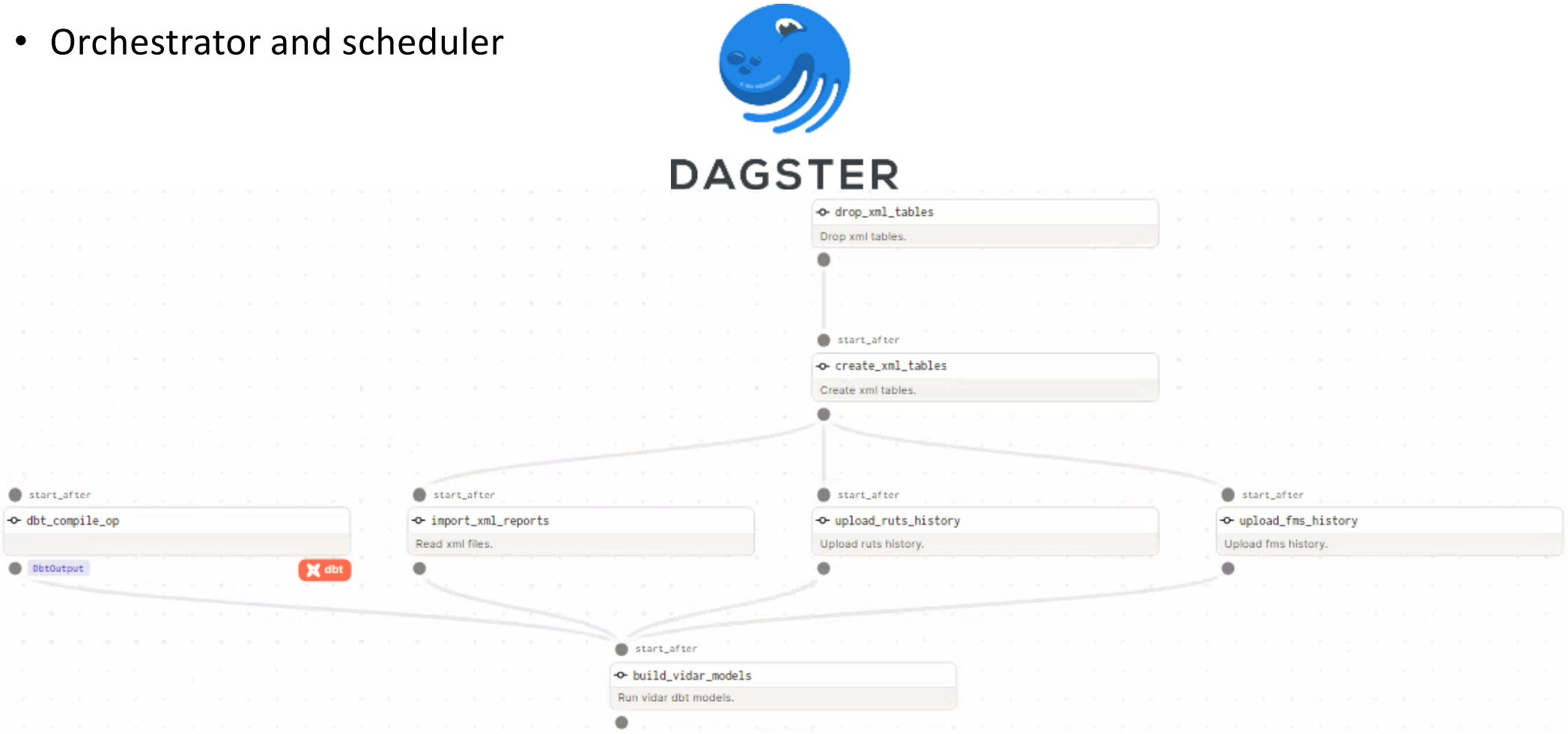
- Platform for project management
  - Code sharing and review
  - Pipeline for test and deployment
  - Package management
  - Planning and timeline



SVERIGES RIKSBANK

Tools for continuous development

- Orchestrator and scheduler



Tools for continuous development

- Data transformation
- Testing
- Access management
- Publication of documentation



dbt

pytest

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Experiences, essentials and perspectives for data science in the hearts of central banks and supervisors: a case study of the Dutch central bank[1]

Patty Duijm and Iman van Lelyveld,
De Nederlandsche Bank

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Experiences, essentials and perspectives for Data Science in the hearts of central banks and supervisors: A case study of the Dutch central bank*

Patty Duijm and Iman van Lelyveld

December 2023

*Patty Duijm (p.duijm@dnb.nl) and Iman van Lelyveld (iman.van.lelyveld@dnb.nl) are both at the De Nederlandsche Bank; Van Lelyveld is also at VU Amsterdam. The contents of this paper reflect the opinions of the individual authors and do not necessarily reflect the views of De Nederlandsche Bank. Throughout the paper we refer to some of the many data science projects conducted at the Data Science Hub. All of this work is with great thanks to the data scientists involved.

Abstract

Central banks and supervisory authorities are venerable institutions not known for their ability to quickly adopt new techniques in a rapidly changing world. However, these authorities play a central role in the financial system. Moreover, they have an advantage that should not be ignored: In many cases, they receive confidential information about individual firms that is not available elsewhere, and they have a solid understanding of the domain. We discuss how to leverage the potential of data science, Artificial Intelligence (AI), and Machine learning (ML) using the example of experiences and use cases at one of these organizations: DNB, the Dutch central bank. The dual role of DNB as central bank and prudential supervisor, the cloud-first strategy, and the successful implementation of a Data Science Hub ensure that the lessons learned are of wider relevance.

Keywords: data science, project implementation, central banking, supervision

JEL classifications: C8, E58

1 Introduction

New data sources and new techniques are rapidly providing new possibilities for central bankers and supervisors to improve the tools they have at their disposal. In this article, we present a case study on how to effectively ingrain data science based on our experience at a central bank and supervisor (De Nederlandsche Bank (DNB), the Dutch Central Bank). These data science techniques are also commonly known as Machine Learning (ML) or Artificial Intelligence (AI).[1] Following a trial in the Statistics division, data science in the Dutch Central Bank was formalized by the establishment of the Data Science Hub (DSH) in 2020. The DSH is the hub in a hub-and-spoke model, working with the various divisions across the central bank, supervision, and resolution. It is tasked with promoting data-driven ways of working and fostering the data science community.

The goal of our study is three-fold. First, we demonstrate the huge potential of data science in seven experiences – all supported by our own projects. Second, based on our experience, we draw up a list of five must-haves – the "essentials" – for fruitful data science work in an organization. It is a common misconception that getting data science to work for an organization can be achieved by hiring a few smart data geeks and having them develop "AI" in a remote corner of the organization. We will argue that AI should become part and parcel of daily work

[1] Following the definition of Nasution et al., 2021, we define data science as the extraction of knowledge from high volume data, using skills in computing science, statistics, and specialist domain knowledge of experts. In the field of supervision, such tools are known as "Suptech" and they allow supervision to become more efficient or have more comprehensive risk capture. As for tooling, we almost exclusively use open source coding languages (mostly Python, some R). These languages are developing extremely fast and are designed to collaborate (also with other frameworks). Internally, code is, in principle, free to share through Azure DevOps. DNB has decided to treat code as data and apply the existing sensitivity framework. Externally, the DSH operates the DNB Github that hosts our publicly available packages. Although for real development, we use an integrated development environment (IDE, e.g.,Visual Studio Code), Jupyter Notebooks are invaluable to let people interact and experiment with code and data. As for statistical methods, we take a pragmatic approach and try to solve the issue at hand with the simplest possible method rather than the most fancy one (see Chakraborty and Joseph, 2017 for an excellent overview of relevant methods). We are in close contact with teams that focus on Business Analytics (i.e., dashboarding in Power BI) or Robotic Process Automation (RPA).

processes to reap the full benefits. Third, we share how we work at the DSH. This last section is meant to provide practical guidance and inspiration to other organizations that are thinking about implementing data science in their organization.[2] We thus leave out much to the technical details of the – sometimes highly technical – solutions we have provided to our clients.

Throughout our study, our own data science projects are used to support our findings and make the lessons as concrete as possible. However, for a full overview of all the projects we have undertaken over the last years, we advise you to consider our Annual Reports published on our web page.

The experiences we have gained over the last years are the following:

1. The combination of – sometimes novel – **granular data** offers new insights useful for both supervisors and policymakers.

2. **Combining internal data with external resources** increases the information value of the data, allowing supervisors and policymakers to improve the incorporation of external factors and risks.

3. **Automating data processes** results in more efficient, accurate and highly frequent economic (prediction) models and allows policymakers to focus on modeling rather than preparing the data.

4. Machine learning has great potential and **outlier detection models** have already proven to be effective innovations in supervision. Complex machine learning approaches, such as neural networks, are, however, not a prerequisite for a successful data science project. In many cases, a simple model is a great place to start.

[2] Of course, the impact of big data and AI on the industry is much broader, also impacting central banks or regulators. For example, how do you keep up with the rapid pace of change that creates significant regulatory uncertainty for financial institutions seeking to use new techniques? On top, the increased use of AI at financial institutions may also expose financial institutions more to cyber-risks. The study by the World Economic Forum, 2018 and more recently Bank of England, 2023 provides a very complete overview of how AI will change the financial sector. In our study, we focus on how to incorporate it within the organization of a central bank and supervisor.

5. The real value of data science lies in the combination of data science techniques and **domain knowledge**, and therefore perfectly illustrates the importance of domain experts in data science.

6. The value of data science applications depends on the **adoption by the business** and therefore user-friendly interfaces to integrate the data science solution into the daily workflow are as important, if not more important, as technical excellence.

7. Last but not least, data science has value for the **entire organization**, including, for example, HR and business operations, and should therefore be in the 'heart of the organization'.

We identify the following five essentials for organizations:

1. Embracing new data science methods and using them throughout the organization requires sufficient appetite for experimentation – especially at senior levels – and therefore a **common vision** is key.

2. A data science function should be able to combine many different activities, and this requires the **right mix of skills**.

3. **Responsible coding** is needed to ensure the replication and reproducibility of analyses and policy decisions.

4. A mature framework for **data governance** is needed to work responsibly, and this should be embraced by the organization.

5. A well-established IT environment is needed to facilitate data scientists in all their needs, and even more important is a **close cooperation with IT.**

2 DNB's experiences

2.1 Granular data sets

The 2008-2010 crisis revealed that authorities were missing crucial information to accurately identify risks in the financial system. This realization led to a significant increase in the volume and granularity of data that financial institutions are required to report. For Europe, for example, granular information on credits (AnaCredit), money market transactions (MMSR, SFTR), derivative trades (EMIR), security holdings (SHS) and trading (MiFId) is being collected. Although data quality issues remain, these very granular data allows for an unprecedented coverage of all major activities in the financial sector (Ullersma and Van Lelyveld, 2022).

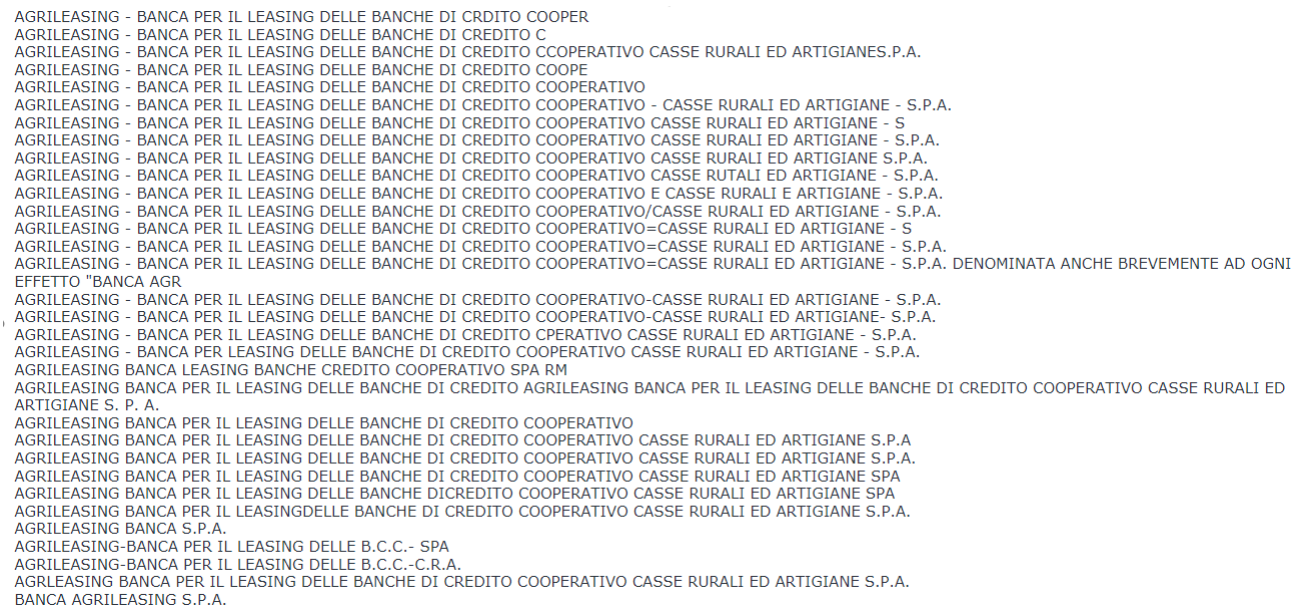
Combining the new granular data in a coherent framework would allow for an even better understanding of the dynamics of the European financial system. Here, some challenges remain. We list three of the main challenges we have encountered and show how we coped with them.

First, in many cases, reporting agents are free to submit counterparty names as free-form text. As Figure 1 shows, exposures to the same agent can therefore be labeled slightly – or completely – differently, underestimating the concentration of risks. To map a differently spelled counterparty to a single and unique identifier, we have developed a "fuzzy name matching" package which is available on Github (Nijhuis, 2022).

Second, an issue in merging granular datasets is that not all entities included in the data have a single unique identifier. After the global financial crisis, the Legal Entity Identifier (LEI) was introduced. The potential of the LEI for supervisors, financial markets, and institutions is significant. It not only introduces unique firm identifiers (Level 1 data) but also contains information on ownership structures (Level 2 data). Thus, one is able to, for example, plot intra-firm networks as is shown in Figure 2. At the DSH, we attempted to measure intra-firm complex-

Figure 1: **The case for Fuzzy Name Matching**

The figure shows how one unique institution may show up in datasets with many differently spelled names, providing a clear case for our fuzzy name matching package.



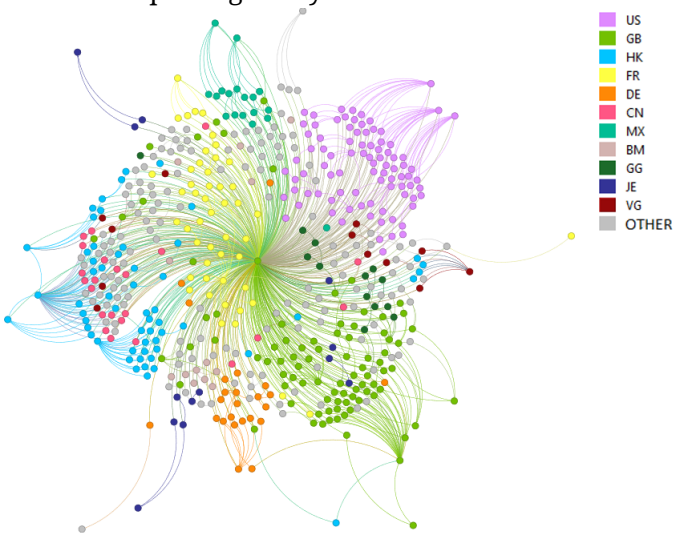
Source: DNB DSH project documentation.

ity using LEI data, but soon found that substantial data issues, among which is the current low coverage of the data, impede its use (Rietveld, Lange, and Duijm, 2023).

Third, by focusing on one specific topic, one may simply not get the complete picture. For example, due to the over-the-counter (OTC) nature of derivative markets there is no centralized overview of the market. Participants only observe their own volumes and exposure concentrations. The major US investment banks therefore did not realize that jointly they were massively exposed to a single entity, the lightly regulated insurer AIG. In setting their capital buffers and implementing other risk mitigating procedures, they were thus ignoring an important yet unobserved concentration risk. In one of our projects, we have measured the degree of interconnectedness of derivative positions between institutions for derivative contracts for which at least one counterparty is established in the Netherlands (Van den Boom et al., 2021). Figure 3 shows that the system is highly in-

Figure 2: **Intra-firm networks using LEI data**

The figure plots the intra-firm network of HSBC Holdings PLC. Every dot represents an entity with a LEI code that belongs to HSBC Holdings PLC. The colors of the nodes represent the country of the reporting entity.



Source: Rietveld, Lange, and Duijm, 2023.

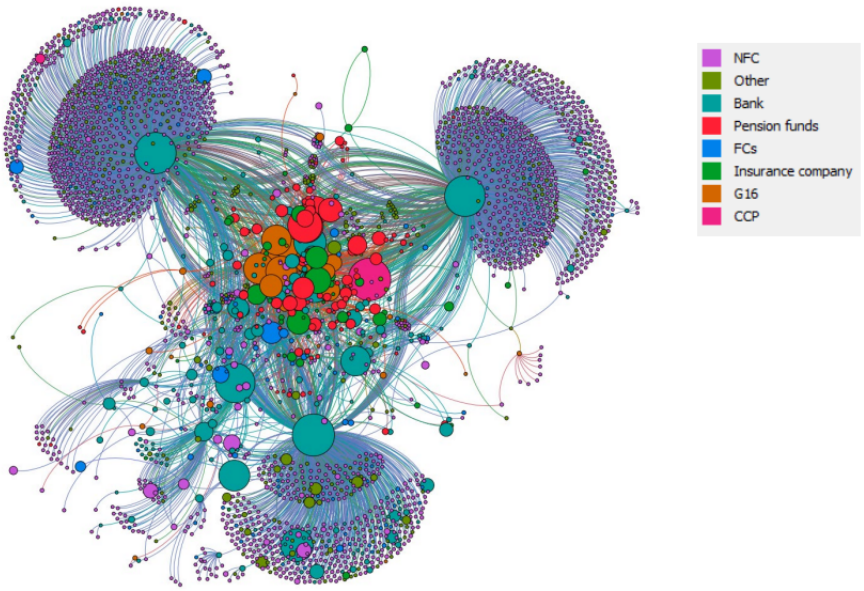
terconnected with three large Dutch banks that facilitate most clients and have access to Central Clearing Parties (CCPs). The cluster in the middle indicates that Dutch pension funds and insurance companies do not trade exclusively with Dutch banks but also use large international banks and CCPs to enter the derivative market. These kinds of insights are pretty valuable. However, this network is based only on interconnections via derivative trades. To get a complete view of how interconnected the financial system is, one should actually consider different types of linkage. Currently, in one of our projects we have combined several granular data sets with the aim of coming to a comprehensive view of exposures of Dutch banks on non-bank financial institutions (NBFIs).[3] The data included comes from AnaCredit, Securities Holdings Statistics (SHS), Securities Financing Transactions Regulation (SFTR) and European Market Infrastructure Regulation (EMIR). We have combined these data sets by using the aforementioned LEI information

[3] We are not the first to do this, cf. Hüser and Kok, 2019. Furthermore, see the survey by Hüser, 2015 for an overview of the multilayer financial networks literature.

and data obtained from the Register of Institutions and Affiliates Database (RIAD). Given the confidentiality of some of these data sources, we do not show the output here. However, we can say that the established network clearly shows that Dutch banks are exposed to NBFIs via multiple linkages. Hence, for concentration risk, one should not focus solely on a single source of exposure but take into account different sources of (preferably) granular data.

Figure 3: **Network analysis of interest rate derivatives**

The figure shows a network based on interest rate derivatives. The colors of the dots indicate the different sectors and the sizes reflect the (logarithmic) aggregate size of the derivative positions.



Source: Van den Boom et al., 2021

2.2 Combining internal and external sources

The data you own is much more valuable to you if it is augmented with data owned by others (Mewald, 2023). This is also the case for central banks. The data a central bank receives through regular reporting can become even more informative if we add additional nontraditional data. For example, Van Dijk and De Winter extract topics from a large corpus of Dutch financial news (spanning January 1985 to

January 2021) and investigate whether these topics are useful for monitoring the business cycle and nowcasting GDP growth in the Netherlands (Van Dijk and De Winter, 2023). Their newspaper sentiment indicator has a high concordance with the business cycle and increases the accuracy of DNB's nowcast of GDP growth, especially in periods of crisis. Therefore, tone-adjusted newspaper topics seem to contain valuable information not embodied in traditional monthly indicators from statistical offices.

Of course, adding other data is not a new idea. Hedge funds, for example, have been using "alternative data" for decades. One of the first companies to use alternative data like satellite imagery, web scraping, and other creatively sourced datasets was Renaissance, a hedge fund looking for an edge in trading. A big bank like UBS uses satellite imagery of big retailers' parking lots and correlates car traffic with quarterly revenue, generating accurate predictions of earnings before they are released.[4] Another great example is one from Banque de France (Bricongne, Meunier, and Pical, 2021). They use freely available granular daily satellite-based data for air pollution to predict industrial production. These cases clearly show that alternative data, i.e. external data without a direct link to the own business, can help.

In cooperation with the BIS Innovation Network, the Data Science Hub has developed a digital twin pilot of climate risks.[5] Here, a digital twin is defined as a digital representation of a real-world entity or system.[6] In this pilot, the digital twin was developed to measure the effects of climate events on the financial system via real estate exposures of financial institutions. For the Netherlands, a flood risk case has been assessed. Insight into the spread of flood risk was obtained based on existing research on damages caused by specific water depths. Zipcode

[4] The founder of Walmart, Sam Walton, would fly over parking lots in the 1950s in person to do pretty much the same thing.

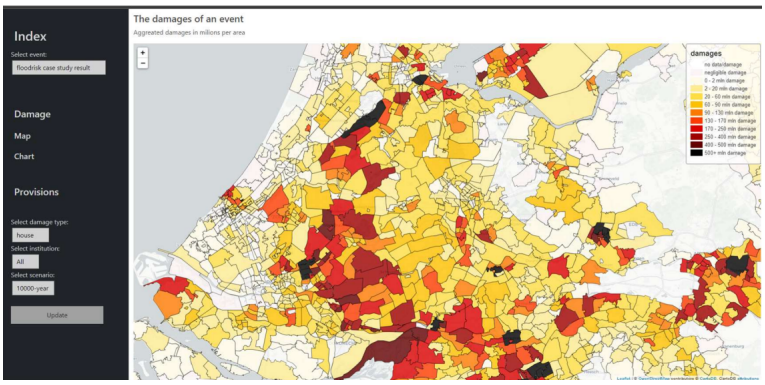
[5] More background information on this project can be found in the summary of the workshop on the role of technology in finance (BIS, 2023).

[6] The definition of the digital twin is obtained from Gartner. See Jones et al., 2020 for a detailed explanation of Digital Twins.

maps in combination with basic housing information and housing price statistics were in turn used to map real estate exposures to the flood map and determine loss estimates for the financial industry, i.e., banks and insurance companies (see Figure 4).

Figure 4: **A Digital Twin pilot for climate risk**

The figure shows estimated losses for the Dutch financial industry in the case of a flood risk scenario (with an estimated probability of once in 10.000 years).



Source: DNB DSH project documentation.

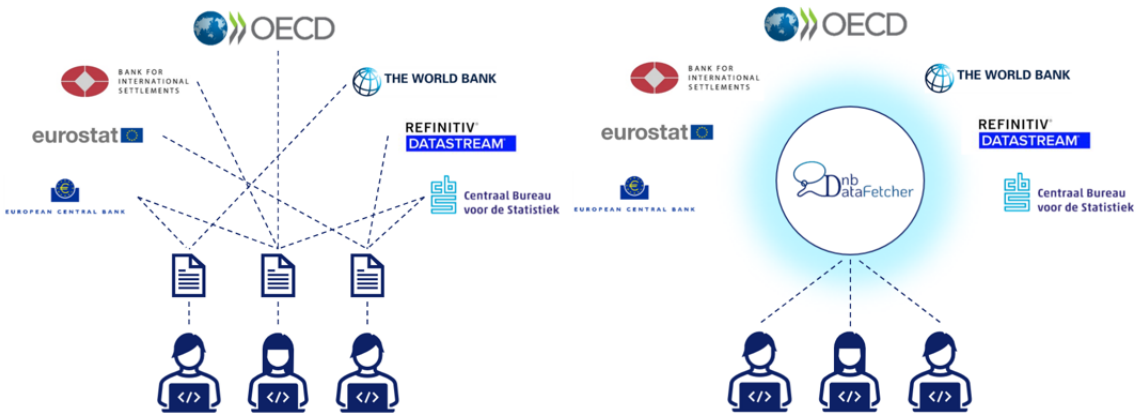
2.3 Automating data processes

Until relatively recently the typical workflow was that data was collected manually from either internal or external sources. Often the wrangling of data was a labor-intensive job in Excel. Such manual processes are not only expensive, but also prone to human error. For example, for DNB's internal inflation prediction model, external data was collected from various sources on a regular basis as input for the model. In fact, multiple processes within the central bank use external data sources, resulting in colleagues collecting (the same) data manually or via ad hoc scripts. This may also result in cases in which different (or even outdated) versions of the same data set are used in DNB. The left-hand panel of Figure 5 shows this situation.

In an ideal situation, i.e., the right-hand panel of Figure 5, colleagues in the

same institution have immediate access to the same data, while restrictions dictated by privacy and confidentiality should be respected. Therefore, in addition to opening a discussion on how to modernize data workflows, we developed DataFetcher.[7]

Figure 5: **The DNB DataFetcher**



Source: DNB DSH project documentation.

The DataFetcher is a Python package that acts as a wrapper on top of publicly available Application Programming Interfaces (APIs) granting access to various data sets (e.g., IMF, OECD, and ECB). Users no longer need to understand all the separate APIs but can download data using unified syntax. Working closely with users in development allowed us to establish trust and leave users in control. Once the DataFetcher was established, we started on infrastructure to collect the necessary data and fill a database on Azure – our cloud provider. Again, working closely with the modeling department allowed for skill transfer and the establishment of a sense of comfort with this new way of working. This approach is known as BizDevOps (developing and operating close to or by the users) and is especially effective if requirements are fluid or to be fleshed out in the process.

The next step we are working on now is to be able to automatically run models

[7] The package is available to ESCB NCBs.

in the cloud. The ultimate goal here is to be able to easily interact with forecasts for example from a smartphone. It is not our ambition to conquer the market with this app, but the ability to quickly and painlessly change some part of the process allows for more flexibility in the development process. For example, it will be much easier to change the inflation forecasting model to incorporate unanticipated energy crises or pandemics. This, in turn, will improve policymakers' ability to react to unforeseen events in a timely manner.

2.4 Start simple

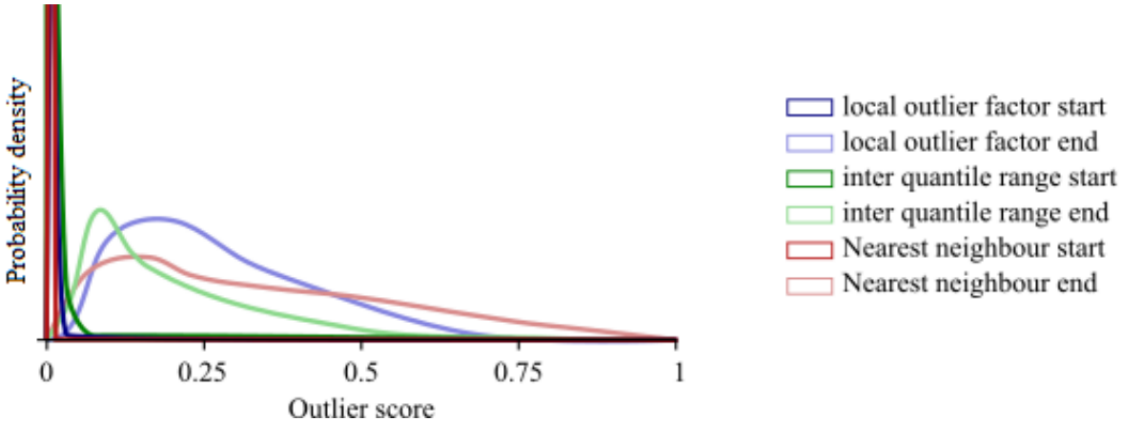
There is a lot to gain from machine learning, but when setting up the DSH we soon realized that starting with fancy machine learning models would not convince end-users to use it. For people without a lot of knowledge of or experience with data science, more complex models may soon be seen as a black box. During our first year, we therefore have put more effort in more simple projects, for example guiding colleagues to use Python instead of Excel. Here the main lesson is: start simple. Of course, we are still the Data Science Hub and get most enthusiastic about the real data science. An important part of both the compilation of statistics and supervision is to identify observations that are out of the ordinary. In this section we cover two of these projects, that focus on outliers: first, an approach in which we implement reinforcement learning in granular prudential reporting, and, second, a Know Your Customer (KYC) use case.

The first use case is one in which we implement a *reinforcement learning algorithm* (Nijhuis and Van Lelyveld, 2023). Outliers are often present in data, and many algorithms exist to find them. Often we can verify outliers to determine whether or not they are data errors. Traditionally, outliers are identified using 'business rules' – ground truths that are valid by definition or results from experience. Assets should equal liabilities, for example. However, the definition and hardcoding of business rules are both cumbersome. Also, in some use cases we

have not yet established strong priors for what is 'normal'. Unfortunately, checking such points is time-consuming and underlying issues leading to the data error can change over time. An outlier detection approach should therefore be able to optimally use the knowledge gained from verification of the ground truth and adjust accordingly. With advances in machine learning, this can be achieved by applying reinforcement learning in a statistical outlier detection approach. The approach uses an ensemble of proven outlier detection methods in combination with a reinforcement learning approach to tune the coefficients of the ensemble with each additional bit of data.[8] Figure 6 plots the distributions of outlier scores for the three different methods in the ensemble. In a reinforcement learning approach, an algorithm is not just trained and applied, but in each iteration, the algorithm gets feedback on its performance. In this case, analysts manually check a set of extreme values identified by the algorithm and record their evaluations. The algorithm then takes this feedback into account and presents a new list of outliers (possibly also incorporating fresh data).

Figure 6: **Outlier detection with reinforcement learning**

The figure shows the distribution of the outlier scores for the first and last iteration for the different parts of the ensemble.



Source: DNB DSH project documentation.

[8] Ensembles combine the strengths of different types of algorithms to get better performance.

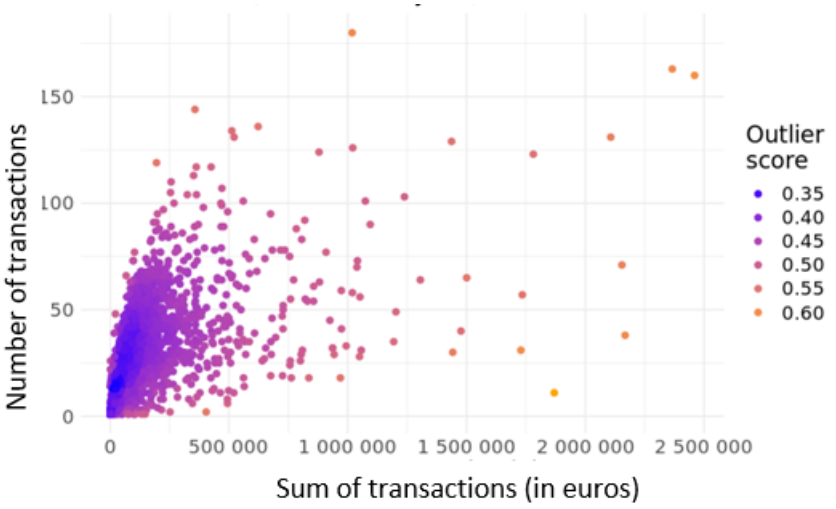
At the Data Science Hub, we are currently implementing the reinforcement learning outlier detection approach for granular data reported by Dutch insurers and pension funds under the Solvency II and FTK frameworks. This application shows that outliers can be identified by the ensemble learner. Moreover, applying the reinforcement learner on top of the ensemble model can further improve the results by optimizing the coefficients of the ensemble learner.

The second use case is KYC. KYC is a mandatory customer due diligence process that requires financial institutions to verify customer identity and assess and monitor their activities to prevent fraud. Since larger banks often have millions of clients and billions of financial transactions, data science has huge potential to help to monitor customers and identify potentially fraudulent transactions. In fact, it is already applied. For example, Anzo (Cambridge Semantics) provides flexible knowledge graphs that allow institutions to connect customer information from structured and unstructured data and thus provides a data-driven solution for KYC processes. Of course, the use of data science to monitor customers comes with additional challenges, such as discussions on consumer trust in technology and privacy. However, as Elliott et al., 2022 stress, without integrated and innovative contributions from the industry resulting in improved services, it will be impossible to shape a path towards more substantial technological innovations (Elliott et al., 2022). While financial institutions have to comply with KYC guidelines and regulations, supervisors are in charge of assessing whether they do so. Based on samples of client data from supervised entities, the Data Science Hub in cooperation with our colleagues in integrity supervision have therefore developed an outlier detection model to do risk assessments of these clients and map it with the risk classification of the supervised entity. With the model, we were able to effectively select clients with abnormal transaction profiles. More specifically, we applied an Isolation Forest outlier detection algorithm to millions of profiles (Section 5.4 in the Cambridge State of SupTech Report 2022 provides a more extensive overview of our case). Figure 7 shows a graph with outlier detection scores

for bank clients plotted against two client characteristics. The results of the outlier detection model resulted in identification of new risks and efficiency gains, since supervisors are now able to consider all transactions instead of considering small samples. The model and results have been shared with the supervised banks to ensure transparency. This example clearly shows that the real value of data science lies in the combination of the domain knowledge of the supervisor and the computational power of a computer to analyze millions of client transactions. The importance of domain knowledge is the topic of the next section.

Figure 7: **Using outlier detection for integrity risk**

The figure shows a plot with outlier scores for bank clients plotted on two characteristics of those transactions, i.e. the sum of transactions (in euro) and the number of transactions by the client



Source: DNB DSH project documentation.

2.5 Domain Knowledge

As stated, the KYC project is a perfect example that shows the importance of domain knowledge in data science projects. This was also stressed by DNB board member Steven Maijoor in his speech at the Data Science Conference organized by the Data Science Hub in 2022 using the following example. To detect outliers in client transaction data, we traditionally define telltale identifiers such as “multi-

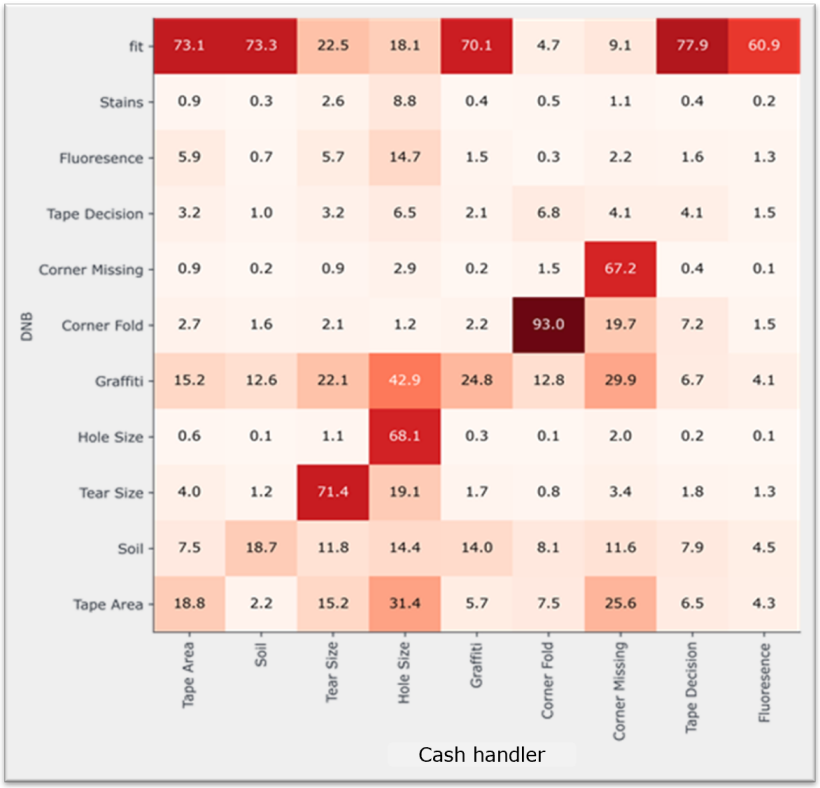
ple accounts on a single address" and "a single deposit per month and immediate withdrawal". Seen separately, these are relatively innocent. Together, however, they can indicate human trafficking of seasonal workers. The combination identifies subcontractors who organize housing for seasonal workers, which is a perfectly legal activity. But if at the same time there is an immediate withdrawal of the wages deposited with only a fraction of the wage paid to the worker, it is clearly an illegal activity. However, the combination could also be consistent with student housing: a large inflow when student grants and loans arrive and a relatively quick withdrawal rate. While these examples are just based on two dimensions, in practice there are many more dimensions, and these can interact in multiple and non-linear ways. With the use of data science techniques, we can identify those. Exactly for this reason, data scientists should be in close contact with colleagues with domain knowledge, not only to provide input for the model but also to interpret model outcomes.

Another example of a data science project that shows the importance of domain knowledge is our False Unfit Banknotes project. Commercial cash handlers send banknotes that they consider unfit for circulation to DNB. Cash handlers also manage ATMs in the Netherlands. Unfit banknotes are checked again at DNB because DNB has specific authentication sensors to determine whether a banknote is unfit for circulation. During the sorting process at DNB, a surprisingly large percentage of these unfit banknotes are evaluated as fit. We call these banknotes "false unfit".

In cooperation with colleagues from the Payments division, the Data Science Hub investigated the high percentage of false unfit banknotes and how this percentage could be reduced. By looking at the data on the matched banknotes, it can be seen where the classification differs between DNB and the cash handler and specific rules that do not add up can be pinpointed. Figure 8 shows the percentage of cases in which DNB and the cash handler classifications are in line. For example, in 93% of the cases both DNB and the cash handler decide to classify a

Figure 8: **Detecting False Unfit Banknotes**

The figure shows, on the first row, the percentage of bills deemed fit. The other cells show the (dis)agreement of DNB and the cash handler.



Source: DNB DSH project documentation.

banknote as unfit due to a folded corner, and hence in 7% of the cases the cash handler decides to classify a banknote as unfit due to a folded corner while DNB does not. Hence, for fully compatible measurement, the diagonal of the matrix from the bottom left to the top right would be filled with dark squares, as the cash handler's trigger would be identical to the DNB's trigger. The number of DNB fit classifications if the cash handler detects a problem shows the extent of the false fit problem. Only hole size, tear size, and corner defects are regularly triggered for the same banknote by both the cash handler and DNB. The settings on the cash handler's machine could be adjusted to reduce the number of false unfits. While it is easy to compare the consequences of adjusting just one of the rules (e.g., tape decision or dirt), it quickly becomes more complicated once multiple rule settings are adjusted simultaneously. We therefore applied machine learning to arrive at the optimal combination of multiple rule adjustments. Reducing the number of unfits can save much effort and expense, and this project resulted in a set of recommendations for our Payments division to achieve these cost reductions.

2.6 Adoption by the business

The real value of data science lies in the adoption by the business. We can work on fancy models, but if they are not used in practice, the business value is less than zero. Data science can generate value in multiple ways, and here we distinguish two main types.

First, data science projects can be a one-off project. The project discussed in the previous section, our False Unfit Banknotes project, is a good example of such a project: it generated clear value for the end-users, but will not be implemented further or be followed up.

The second type of projects are data science projects that find their way to implementation to be used in daily processes. This often comes with more challenges. We have seen countless promising Proofs-of-Concepts (PoCs) received with

much enthusiasm that fail to make their way into production. Note that the success of any algorithm is crucially dependent on how seamlessly we can integrate innovation in the existing workflow.

The concept of "in production" is a source of much confusion between IT and the average user. Typically, an analysis is used to set policy as soon as the policymaker is convinced that the results are sufficiently solid. The analysis is often somewhat a journey and invariably involves manual steps. Reproducibility is often not the first concern and it is ensured because the analyst is closely involved and has intimate knowledge of how to replicate the results. For an IT department that is asked to bring such an analysis into production, the standards need to be much higher: the process needs to run without (much) manual intervention or knowledge of the subject matter. This involves programming to catch all kinds of eventualities and extensive unit testing. The challenge is to find an organizational form that allows abstracting away typical IT housekeeping tasks (e.g., ensuring proper backups) while allowing the analyst sufficient flexibility in further developing the tool.

At this stage – bringing the data science application "in production" – frictions do not just result from the collaboration between the data scientists and the IT department. On top of the technical challenges comes the interaction with the end-users. In developing a data science application, receiving feedback from the end-users to improve an algorithm turned out difficult, since data science environments were kept too separate from what the average user could access or is comfortable with. In other cases, end-users underestimate the considerable effort that is needed to train and tune a model. Based on smooth experiences with consumer apps, they have unrealistic expectations of what bespoke algorithms can do in the short run.

For a data science application to create value, adoption by both the IT department and end-users is key. Therefore, one should be able to align the often divergent perspectives and desires of both sides. In our experience and depending

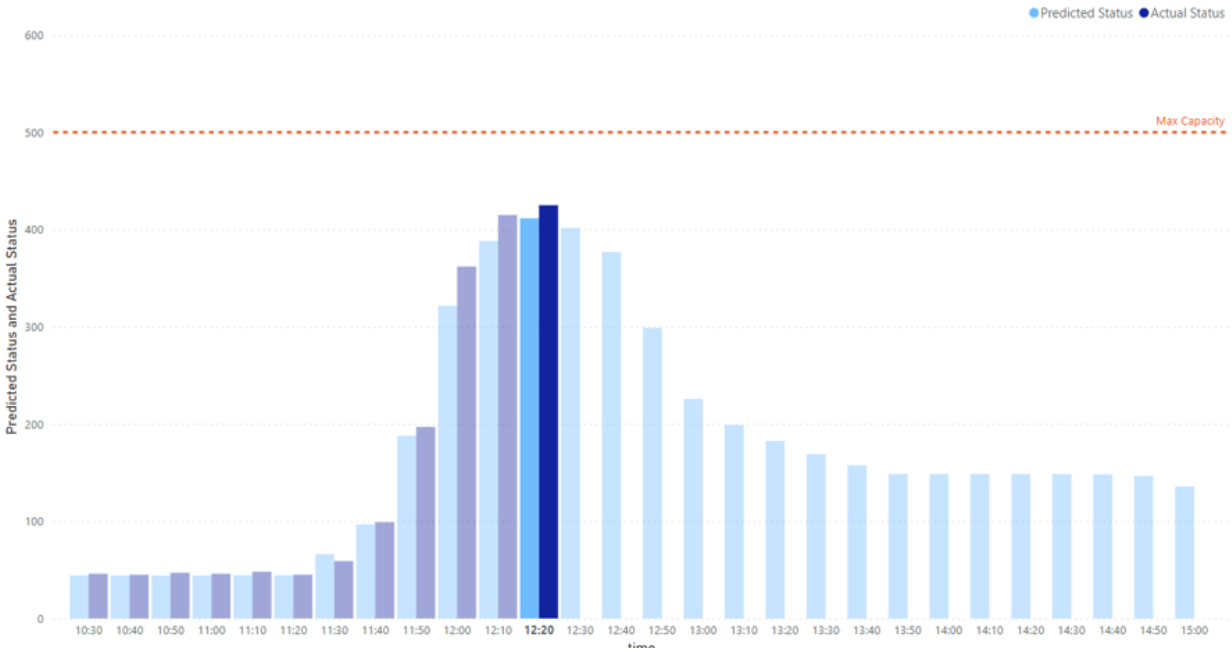
on the size of the project, this takes (a lot of) time. One of our first projects, Dataloop, provides a perfect example of this. Dataloop was initiated as a collaboration between DNB's Statistics Division and the DSH with the goal of improving data quality in supervisory reporting. Dataloop does so by centralizing and visualizing data from different sources. It also offers several feedback loops, for example, between analysts and data validation tools. The project started in 2018, and a lot has happened since then. Only recently, in September 2023, Dataloop was publicly launched as an application for Solvency II insurers under DNB's supervision after a successful pilot. This is just to show that big data science projects like Dataloop take time, especially in this case where the end-users are both internal clients, i.e. supervisors, and external parties, i.e. the institutions under supervision of DNB.

2.7 Data science for the entire organization

Thinking about data science in central banking and supervision, the easiest and probably the most straightforward link is to financial data. However, data science can be applied anywhere in the organization. While the focus in the first years of the Data Science Hub has been on the more straightforward topics, we now also touch on other less traditional topics. For example, we are experimenting with motion sensors in our office building to predict how busy our cafeteria will be. Figure 9 shows the results of our prediction model in the current state.[9] Such forecasts can help our catering service plan capacity and our staff make a more informed choice to time their lunch. Sensor data can, however, be used for other purposes, for example, to monitor the no-shows for meeting room reservations. Another project we tackled is one with Human Resources (HR). The HR data are quite rich, and with these data we were able to perform an analysis on employees' resignations to test which factors impact this decision.

[9] While we can predict the lunch crowd, there are still some issues to fix, since in our model people do not seem to leave the restaurant after lunch; a clear flaw in the calibration of the sensors.

Figure 9: **Predict restaurant visitors using sensordata**



Source: DNB DSH project documentation.

Other departments that are now also starting to get involved are ones that work mainly with text. A large amount of information flows into DNB as text. Natural Language Processing (NLP) and recent advances in Large Language Models (LLMs) such as ChatGPT and BARD have great promise for data science applications in all parts of a central bank. One hurdle we face is that document storage and retrieval have not evolved at the same pace. Documents are scattered in different systems, are not stored in a consistent format, and are difficult to access from our analytics platform. Notwithstanding these hurdles, we see more and more initiatives to make new data science techniques work for less traditional departments.

3 The essentials

As showcased in the previous section, data science has given us much in the last 3.5 years. Starting in 2020 we did not have bank-wide center of excellence like the Data Science Hub at DNB and at the time of writing this paper, we have worked on more than 60 data science projects throughout the whole organization. However, setting up a DSH is not a standalone business. Integrating it in the organization requires more; in this section, we discuss the five essentials that we believe are important for an organization to become a more data-driven one that embraces data science. Note that the section is based on our experience and is dependent on how a data science team is organized within an organization (see Section 4 to get a more detailed understanding of how the DSH is embedded in DNB and how it operates on a daily basis).

3.1 Common vision

Data science is evolving quite fast, putting pressure on the traditional model of generating knowledge and insights from data. Just a few decades ago, much of the empirical analysis was based on manual manipulation of data points. Developing software to analyze data was very slow, costly, and required specialized knowledge. Since then, the granularity, timeliness, and volume of data has shown exponential growth. At the same time, computational power has exploded, compute costs have dropped significantly, and software development has become much easier. This combination opens many opportunities for supervisors and central banks. Putting data science into practice requires, however, equal effort from both the end users of data science tooling – supervisors and policymakers – and IT. In our approach, we try to bring these worlds together.

It is key to understand each other's approach. Often the users of data science tooling expect a smooth user experience given what they are used to consumer

apps. However, central banks and supervisors, dealing with highly sensitive information, need to tread carefully. On the other side is the IT department: they often know what is and is not feasible in the existing IT landscape, but may sometimes not fully understand what the end-users are looking for. Finding the productive middle ground might be hard, but not impossible. Often the solution is communication. Therefore, for an agile approach to working with data, development should be as close to the business as possible. This implies the willingness to allow activities traditionally reserved for IT departments to take place across the institution. At the Dutch Central Bank, we are currently developing a "built-by-business" (3B) policy detailing how to enable and maintain applications developed by departments other than our IT department.

3.2 Right mix of skills

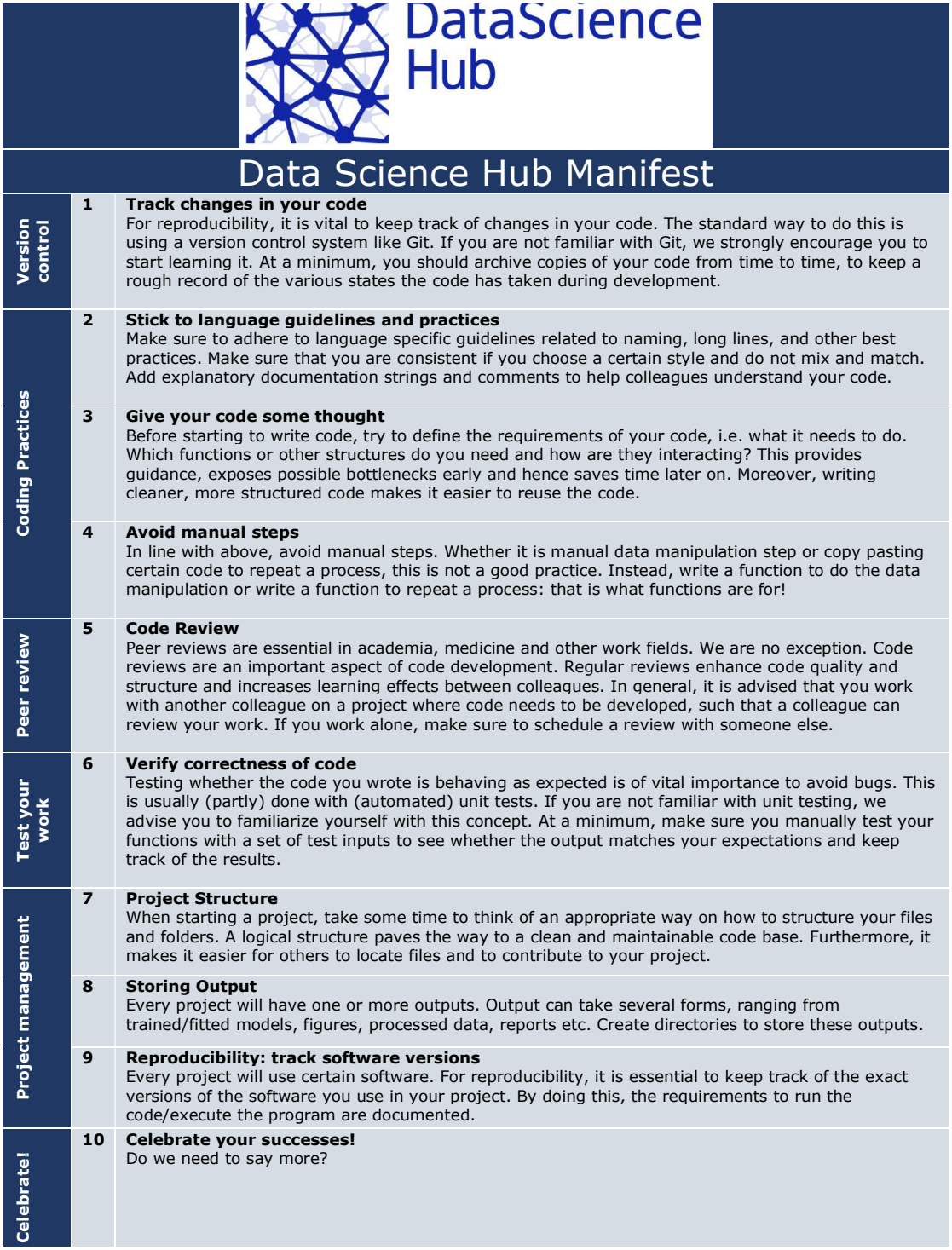
The job title "Data Scientist" is relatively new, and the function that a data scientist fulfills in an organization can vary quite a lot depending on the type of organization. In fully-fledged AI-minded organizations, such as tech start-ups, data scientists may have well-defined responsibilities, such as being responsible for a specific part of a model. In organizations where data science is not (yet) integrated in the organization as a whole, the role of a data scientist may be less clear. In sum, if the essentials discussed here have not reached maturity, the data scientist is expected to attempt the almost impossible: excel at mathematics, statistics, machine learning, coding, data engineering, marketing, governance and navigating the policies drafted for a different time. As mentioned above, setting up the DSH in 2020 is not just about hiring a few data scientists. Of course, a necessary skill for a data scientist is coding. But another skill that turned out to be a need-to-have is the ability to communicate complex subject matter to a broad range of people, many of whom may never have heard of the term data science. The aim of the DSH is to work demand driven to ensure that the things we do are not a result

of available data and an eagerness to work on complex coding, but have potential value for the organization. However, generating demand takes some effort. Policymakers and supervisors need to get a sense of how data science solutions may help them. This first requires a basic understanding of what data science can and cannot do. In Section 4.2 we provide more information on how we did this at DNB. The next skill that turned out to be important is much more related to data engineering. Since the DSH works throughout the organization, the type of data we deal with varies from confidential supervisory data to financial market data sourced from a rating agency. This means that data we work with is stored at different locations in very different formats. Therefore, a bit of knowledge is needed on the data engineering side. Lastly, as discussed in Section 2.6, for a data science project to land in the organization, adoption by both the IT department and end-users is key. In an organization where data science is relatively new, this requires the ability of data scientists to quickly find their way through the organization. Needless to say; the ability to integrate data science in the business is not only dependent on the skills set of data scientists, but may as well require additional skills from, for example, managers and business analysts. As stressed by Harris, 2012 , they need to become i) ready and willing to expirement; ii) adept at mathematical reasoning; and iii) capable to see the big (data) picture.

3.3 Responsible coding

Responsible coding does not only imply writing code that is clean and readable, but also involves things like regular reviews and unit tests. This, in turn, is the key to being transparent about policy processes and work consistently over time. At the DSH we formulated several core principles that are laid down in our Manifest. Figure 10 shows the ten core principles. With these principles, we primarily target a non-technical audience. In addition we have fleshed out the resulting guidance on a much more technical level.

Figure 10: **DSH Manifest**



Source: DNB DSH project documentation.

3.4 Data governance

The increasing volume and use of data results in more attention to the privacy, protection, and security of data. On the one hand this is covered by (recent) regulations regarding data.[10] On the other hand, for organizations with a long history, this requires a new way of working with respect to, for example, data ownership. Data should be carefully managed throughout the organization, with data owners responsible for the use and management of their data. For a data scientist this implies that there is an important initial step before actually starting to work with the data: Request formal approval to use the data for a specific period, providing information on the goal, timeline, and output of the project for which the data is used. At the same time, there should be enough flexibility in the data governance process for the organization to be able to analyze data in a timely manner and respond to developments and changing business needs. To run things smoothly, it is therefore important that both the data owner and the data scientist know what they can expect from each other in their respective roles.

Data governance is present in each step of a data science project – also during the modeling phase. For example, one concern related to foundational models[11] is that they often require confidential information to be sent to the model (hosted outside of the organization) or, even more concerning, that this information is used to further train the model. To alleviate these concerns, OpenAI – one of the leading companies developing foundational models – is working towards developing trained models that can be used off-line. This would allow organizations to implement a foundational model within their own infrastructure.

[10] Over the last years, quite some countries have developed regulatory (national) frameworks. Recently, the EU-wide Data Governance Act entered into application on 24 September 2023. At the same time, there is a strong case for having a global standard for data governance, since many of the data governance issues cannot be fully resolved at the national level and require international cooperation (Chief Executives Board for Coordination, 2023).

[11] Foundational models are models that are pre-trained on massive datasets, and adapted for multiple downstream tasks. Large Language Models (LLMs) are a subset of foundation models that can perform a variety of natural language processing (NLP) tasks.

3.5 Close contact with IT

Data science projects often require multiple steps and different processes, from data cleaning to merging and analyzing the data, building models, and visualizing results. This requires an IT environment that allows for the variety of tasks and operations and offers the most important integrations, such that the data scientist is offered a seamless experience without having to switch to a different environment. In our experience, this requires quite some effort from IT. The environment, of course, needs to meet the requirements arising from internal data governance and security policies. However, the second most important thing is that the environment is user-friendly. The complexity lies in the fact that they have to serve different users with different desires with respect to the data, software, end-products, etc. Again, here there is a clear difference between fully-fledged AI organizations that could build their data science platforms from scratch and the more traditional organizations that start integrate data science in their existing IT environment. The latter is much more complex, but also much needed to keep data scientists happy.

We have mentioned it a few times; the value of data science applications depends on the adoption by the business, and as such the move to the production stage is key. Close cooperation with IT is crucial in this step to ensure that the transition to production runs smoothly and that the application meets internal policies and security standards. However, this does not start at the end of a data science project. As mentioned by Davenport and Malone, 2021, rather than thinking of deployment as the last step in a linear set of activities, a data scientist – or at least key members of data science teams – should consider factors that have a strong influence on deployment throughout the data science project. In our experience, this comes with quite a difficult trade-off. On the other hand, you want the data scientist to think about the deployment stage upfront, while – on the other hand – you don't want to limit the data scientist in experimenting with various

data science solutions. Hence, short lines of communication with IT throughout the data science project are valuable.

Lastly, the production stage isn't the final stage: it should be clear where the responsibilities with respect to maintenance and continuous improvement lie. For us, this is not always clear. Of course, there is an argument to have the responsibilities there where the application lands, i.e. close to the end-users. However, the end-user is not always capable of doing this, and the distance to IT may simply be too much. In this case, a move towards a BizDevOps way of working - where development is close to the business - could be a solution.

4 Way of Working at the Data Science Hub

Functioning as a hub, the DSH undertakes data science projects in collaboration with other departments (the 'spokes') throughout the organization. This section explains how we work and is meant to be a practical guide and inspiration for other organizations that are thinking about or running a data science department.

4.1 Goals and KPIs

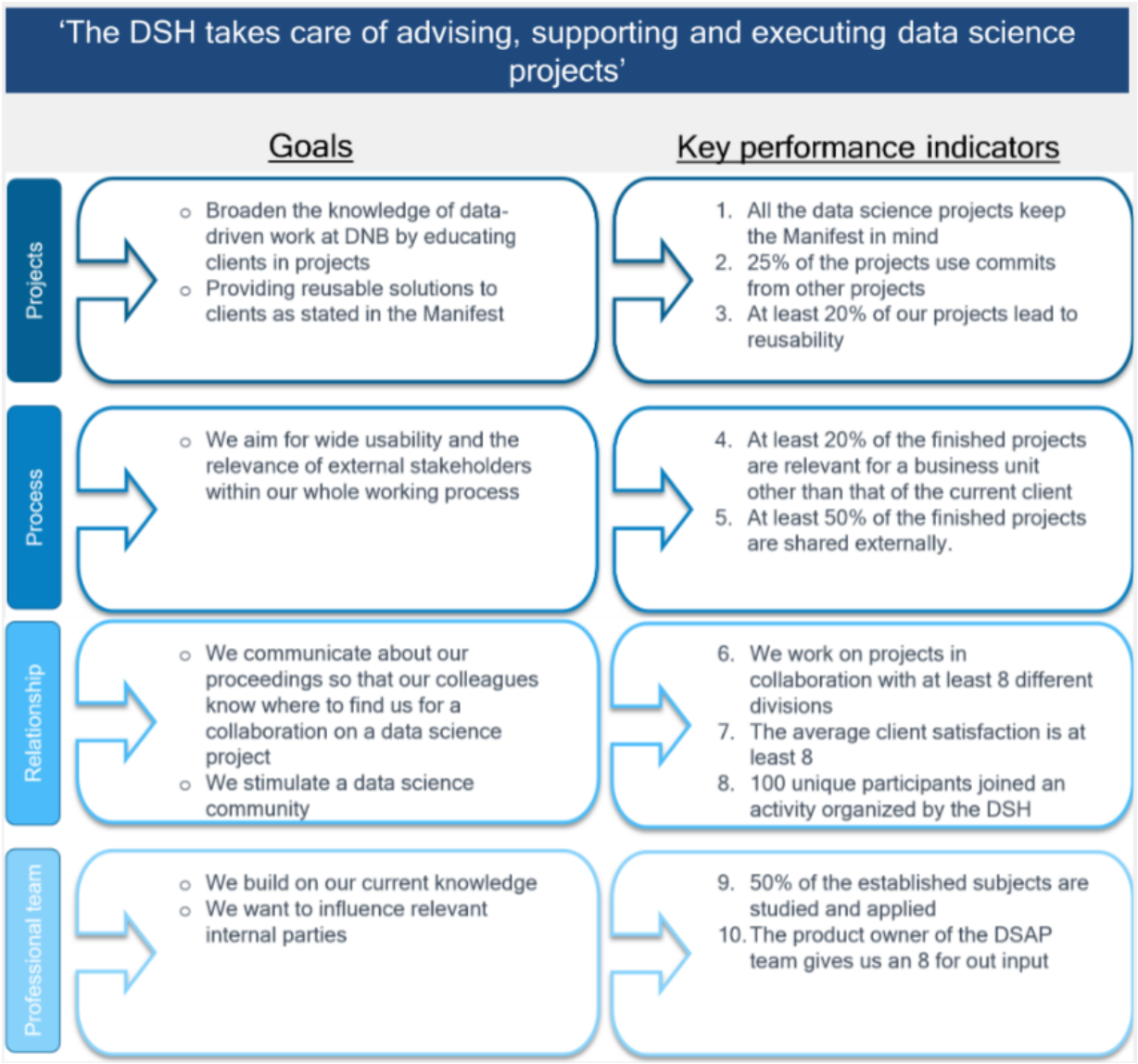
Having a clear role helps to define the mandate of the DSH and position the DSH within the organization. Our mandate is to take care of advising, supporting, and executing data science projects throughout the Dutch Central Bank. As pure data scientists, we love to measure what we do. Therefore, we translated our goals into measurable KPIs. Figure 11 shows our goals and Key Performance Indicators (KPIs) for 2022 and in our Annual Report 2022 we provide more information on the results.

4.2 Inspiration and communication

Setting up the DSH we aimed to work demand-driven, implying that – in the ideal situation – policymakers and supervisors come to us with their data science question. In practice, this requires a lot of work. First of all, they are busy with their own work and should be able to find you at the right moment in time. Second, after they have found you, their expectations often need to be aligned with what is reasonable to expect from a data scientist. We soon discovered that passively waiting for requests is not working and started actively promoting data science throughout the organization.

The first step is to inspire people: What can data science offer you (and what not)? We are giving presentations at all levels throughout the organization and

Figure 11: **Goals and KPIs**

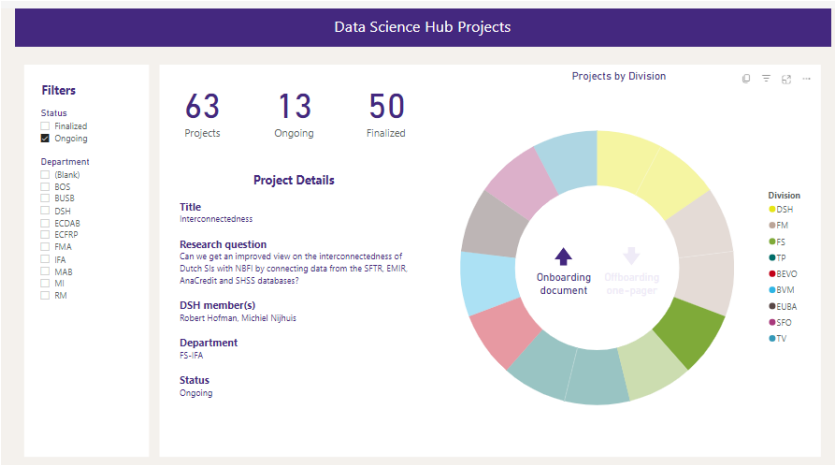


Source: DSH Annual Report 2022

actively seek collaborations, for example, with our innovation office and HR. Regarding the latter, every 3 months we organize a Data Party for new employees at the Dutch Central Bank. The Data Party is a co-creation with our colleagues from Data Office. In less than three hours we let them experience the wonderful world of data and data science using a realistic case in an escape room setting. It is not rocket science, but it turns out to be an approachable and effective way to promote data science to new colleagues.

The second step is making available all project documentation internally. Finalized data science projects provide a very good – if not the best – example for colleagues. Sometimes a data science project in one department can even be replicated for use by another department using a different data set. Therefore, at the DSH we try to maximize the output of our finalized projects by promoting them internally. We therefore have all project documentation (see Section 4.3) from all finalized and ongoing data science projects presented in an approachable PowerBI dashboard (see Figure 12).

Figure 12: **Project overview DSH 2023**



Source: DNB DSH project documentation.

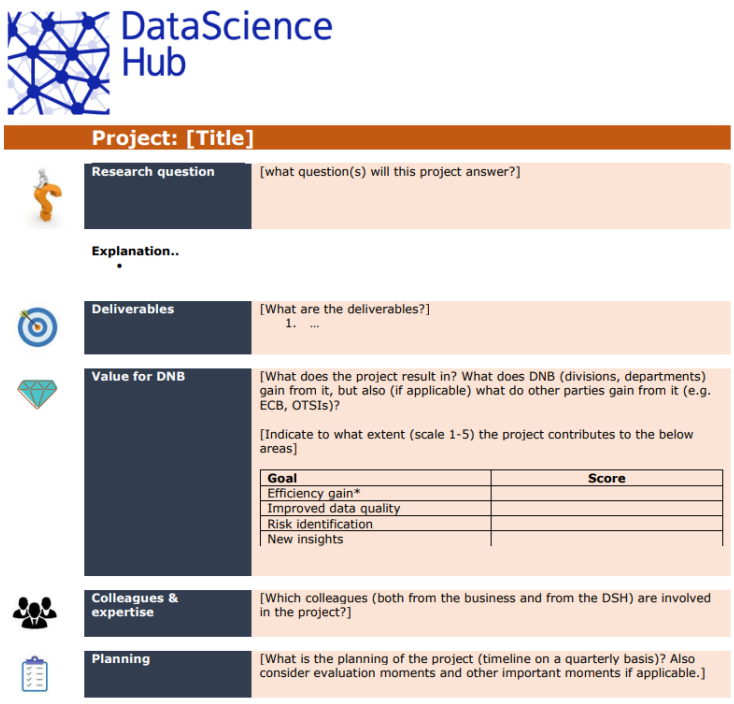
Third, we also communicate our successes externally. A project does not end with celebrating its completion (see Rule number 10 of our Manifest in Figure 10). We actively seek ways to communicate the results to the outside world since projects

may be of interest to other central banks as well. Each year, we present an overview of our projects in our Annual Report published on our website. And if confidentiality allows, we share our code via the Github of the Dutch Central Bank.

4.3 Project design

Data science projects can last forever, and going from the experimentation phase to the implementation phase can take some time. The aforementioned Dataloop project is a perfect example of this. Therefore, it is important to clearly define a data science project upfront - before the work starts. At the DSH, we take care of this with the onboarding procedure but aim to do this as lean and mean as possible. At the start of a project, we agree on the research question, deliverables, value for the Dutch Central Bank, responsibilities, and planning (see Figure 13).

Figure 13: **Project onboarding**



Source: DNB DSH project documentation.

As shown in Figure 12, we run multiple projects at the same time. As mentioned above, projects are always a collaboration between the DSH and the client (end user). In the ideal situation, the client is involved in the coding part. Especially if the project results in an application that is used in daily business, the client should have a thorough understanding of how to properly code. However, in practice, this is not always the case. This is simply because the client is tasked with a completely different job that does not require any coding skills. In such cases, more support from the DSH is required and given. In the longer term, however, and to move towards the data-driven organization, one would like to have the data science knowledge covered in the business in every field of expertise.

Throughout the project, we have regular meetings with the client and always aim to present the ongoing work during department meetings (both within the DSH and at the client department) and other sessions, such as the Open Source Lunches the DSH organizes (see next Section 4.4).

At the end of a project, we offboard the project. Like the onboarding, this is done in cooperation with the client. We summarize the project output in the offboarding form (see Figure 14). At the end of the project, we also send a survey to our client asking for feedback on our work. In addition to this, the survey also provides input to track our KPIs.

Figure 14: **Project offboarding**

DataScience Hub

Offboarding one-pager

Brede Welvaart GIS kaarten (BWGIS) was a project to enrich the CBS microdata on Dutch households with climate risk indicators

Can we use publicly available GIS maps to enrich the CBS microdata on Dutch households with climate risk indicators of living environments?

With the 'Brede Welvaart' project, the department Algemeen Economisch Beleid conducts research on prosperity of Dutch households with microdata from the Dutch Central Bureau of Statistics (CBS). A useful addition to this dataset would be information on climate risk indicators for households' living environments. The Klimaateffectatlas provides publicly available GIS files for such climate risk indicators. This project aimed to use the Geospatial Data Abstraction Library (GDAL) package in Python to match all dwellings in the Netherlands from the BAG to specific climate risks of interest, based on their geographical location. These matched climate risk indicators could then be used to enrich the CBS microdata for research on the relation of climate risk indicators and prosperity of Dutch households.

Result

The objective was to match households with three climate risk indicator .tiff-files: "Water depth during flooding", "Apparent temperature on a summer day", and "Water depth during heavy rainfall". Geographic information for all dwellings in the Netherlands was obtained from the BAG dataset to match on a household level. GDAL was used to translate dwelling coordinates to the conversions of the climate risk maps, hereby extracting the climate risk indicators for each dwelling. As a final step, the climate risk indicators were then merged with CBS microdata using the dwelling ID.



This resulted in:

- Code for converting BAG VBO data instances to specific .tiff-file mappings.
- Code to merge multiple climate risk indicators on VBO level.
- Enriched microdata of Dutch households with climate risk indicators.

Quote from business

"DSH helped us very timely and efficiently with retrieving data on climate risks from GIS maps. The collaboration was very pleasant and the result was exactly what we aimed for. "

Collaborators: *[Names]*

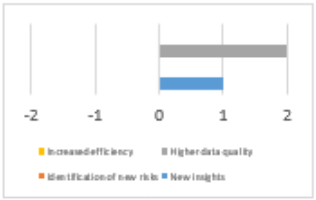
Do you want more information? Contact data_science@dnb.nl

Project review

Customer satisfaction

Overall	9/10
Professional expertise	5/5
Timeliness	4/5
Result-oriented	4/5
Accessibility	5/5

Project Targets



The numbers indicate if targets were delivered as expected; zero is as expected, negative (positive) values imply below (above) expectations

Technical Details & Open Source

Programming language: Python
Data: BAG & Klimaateffectatlas
Method: Geospatial Data Abstraction Library (GDAL)

OSGeo Project

Source: DNB DSH project documentation.

4.4 Activities and training

The largest part of our time is devoted to the execution of data science projects. Part of our mandate (see Section 4.1) is also to advise and support the data science community throughout the organization. We do so by organizing training courses and a variety of activities.

On the regular program, we have the Open Source Lunches (OSLs) and Open Source Workshops (OSWs). OSLs are organized once every two months. In this event, we generally have two presentations about data science related projects from DSH or other DNB colleagues. Sometimes we have a demo of a software package or an analysis platform. Attendance for the OSL is open to all DNB colleagues, however, a certain basic understanding of coding is needed to be able to fully get the message. Therefore, the OSLs typically attract more advanced users of code and modeling.

Depending on the topic, the Open Source Workshops (OSWs) may serve a broader group of colleagues. As the name already suggests, this activity takes the form of a real workshop. In about three hours, participants are introduced to a new data science technique or package. In the past, we have experimented a bit and have offered OSWs ranging from “Getting started with Pandas in Python” to “Webscraping”. However, setting up a completely new workshop every three to six months is quite time-consuming. Therefore, we recently changed our focus and now offer three topics in cooperation with the DNB Academy; “Version control with GIT”, “Clean and Responsible Coding” and “Explainable AI”. The rationale behind this is that third party providers can provide standard training for Python and other languages. However, for some topics, such as GIT, we discovered that there is a large demand for this to be offered by people who are familiar with the DNB IT environment. The workshops are provided by our data scientists and are conducted in small, interactive groups ($\pm$ 8 people).

Another event that we have hosted two times now is called “Word Datapreneur

(WDP)" (in English: "Become a Datapreneur"). This initiative is part of DNB's digital ambition and is organized with the help of many other departments. Participants work on their own data science project for five months and learn how to improve their work with data science-driven techniques. No prior coding experience is required. The only two conditions are that the project has to be in line with the work of all group members involved, and that the participants will do the data science themselves. The WDP program offers support, such as personal coaching and data science workshops during this period.

On top of these regular events, every now and then we organize ad hoc events, varying from a virtual inspiration session (during Covid times) to visits to peer institutions. In 2021, we have organized the conference Central Bankers go Data Driven: Applications of AI and ML for Policy and Prudential Supervision.

5 Conclusion and way forward

New granular data sources combined with new techniques provide unique opportunities for central bankers and supervisors. We have shared how we take advantage of these opportunities at DNB's Data Science Hub. A well-defined strategy, mandate, goals, and a structured way of working definitely helped us apply data science in DNB.[12]

To conclude, we have three messages going forward; i) just start doing it; ii) next to the technical solutions do not forget the people since they are crucial for the transition towards a data-driven organization; and iii) join forces and share.

First, whether our data science projects would produce value for organization was not certain at the start. Without implementation this will remain an open question. There is certainly still room for improvement in the availability of big data and tools to process and analyze it, in the implementation of data science solutions, and in convergence of the views on and expectations of data science. These obstacles are also experienced by other central banks and supervisors (Di Castri et al., 2022; Araujo et al., 2023). While at DNB we are also still working on realizing the five essentials listed in our paper, we hope that this article inspires to just start and simply work with what you have. It may be frustrating to know that the full potential of a data science solution may not (yet) be unlocked due to factors beyond your influence. However, sometimes you only really know what you need untill you are actually working on it.

Second, the challenges with respect to the IT environment have evolved over time. At first, most of our efforts were directed at aligning the divergent desires of different users in combination with internal policies and security standards. More recently, our efforts have shifted to realizing technically more challenging solutions that can stand the test of time. But even an IT environment that supports

[12] Based on early suptech users' experiences, Broeders, D. and Prenio, J., 2018 also stress well-defined suptech strategies with i) goals, ii) data evaluation, and iii) a step-by-step action plan.

an advanced data science platform with boundless opportunities is not a guarantee for data science success. People are as important, and a transition toward a data-driven organization requires equal effort from both the end users of data science tooling – supervisors and policymakers – and IT. At the moment, these worlds are still quite far apart. Especially when more data science solutions will find their way to implementation and will be used (and preferably maintained) by colleagues throughout the bank, it may be desirable to have the IT knowledge closer to the business and consider a BizDevOps way of working.

Finally, organizations like central banks and supervisors are all experiencing similar challenges in their journey towards becoming a more data-driven organization. In his study on the emergence of SupTech, Avramović, 2023 finds that the usage of technology to support supervisory processes is also connected to competition between regulators. That is, regulators are duplicating each other's efforts, attempting to create the best SupTech solutions to keep up with the developments in financial markets. Avramović notes that awareness of such inefficiencies might result in more open information sharing and collaboration between regulators. Open-source tooling facilitates sharing functionality. In the words of our Board Member Steven Maijoor during his speech at our Data Science conference: "*instead of sharing shiny PowerPoint presentations, we could share the functionality that allows us to replicate the analyses of others with our own data*". International coordination and collaboration could help accelerate this and central banks are willing to join forces to reap the benefits of big data as surveys show (Doerr, Gambacorta, and Serena, 2021; Di Castri et al., 2019; Araujo et al., 2023).

There are already some initiatives, like the BIS Innovation Hub, to exchange information and collaborate. A quick win is to start sharing code (cf. DNB GitHub. Similarly, Araujo made a start by collecting open-sourced macroeconomic models run by central banks and other official sector agencies on GitHub.

To conclude, data science has not only great potential but is already valuable for central bankers and supervisors.

References

Araujo, D. (2024) Open-sourced central bank macroeconomic models, *Working Paper*.

Araujo, D., Bruno, G., Marcucci, J., Schmidt, R., and Tissot, B. (2023) Data science in central banking: applications and tools, *IFC Bulletin* **59**.

Avramović, P. (2023) Digital Transformation of Financial Regulators and the Emergence of Supervisory Technologies (SupTech): A Case Study of the U.K. Financial Conduct Authority, *Harvard Data Science Review* **5**.

Bank of England (2023) Artificial intelligence and machine learning, *Discussion Paper* **5**.

Bricongne, J.-C., Meunier, B., and Pical, T. (2021) Can satellite data on air pollution predict industrial production?, *SSRN Electronic Journal*, 1–35.

Broeders, D. and Prenio, J. (2018) Innovative technology in financial supervision (suptech) - the experience of early users, *FSI Insights on Policy Implementation* **9**.

Chakraborty, C. and Joseph, A. (2017) Machine learning at central banks, *Bank of England Working Paper* **674**.

Chief Executives Board for Coordination (2023) International Data Governance – Pathways to Progress.

Davenport, T. and Malone, K. (2021) Deployment as a Critical Business Data Science Discipline, *Harvard Data Science Review* **3**.

Di Castri, S., Grasser, M., Ongwae, J., Mestanza, J., Daramola, D., Apostolides, A., Christofi, K., Rowan, P., Wu, T., and Zhang, B. (2022) The state of suptech report, *University of Cambridge*.

Di Castri, S., Hohl, S., Kulenkampff, A., and Prenio, J. (2019) The suptech generations, *FSI Insights on Policy Implementation* **19**.

Doerr, S., Gambacorta, L., and Serena, J. M. (2021) Big data and machine learning in central banking, *BIS Working Paper* **930**.

Elliott, K., Coopamootoo, K., Curran, E., Ezhilchelvan, P., Finnigan, S., Horsfall, D., Ma, Z., Ng, M., Spiliotopoulos, T., Wu, H., and Van moorsel, A. (Oct. 2022) Know your customer: balancing innovation and regulation for financial inclusion, *Data & Policy* **4**.

Harris, J. (2012) Data Is Useless Without the Skills to Analyze It, *Harvard Business Review*.

Hüser, A.-C. (2015) Too Interconnected to Fail: A Survey of the Interbank Networks Literature, *Journal of Network Theory in Finance* **1**.

Hüser, A.-C. and Kok, K. (2019) Mapping bank securities across euro area sectors: comparing funding and exposure networks, *ECB Working Paper* **2273**.

Jones, D., Snider, C., Nassehi, A., Yon, J., and Hicks, B. (2020) Characterising the Digital Twin: A systematic literature review, *CIRP Journal of Manufacturing Science and Technology*.

Mewald, C. (July 2023) Why Data Is *Not* the New Oil and Data Marketplaces Have Failed Us, *Medium*.

Nasution, M. K. M., Sitompul, O. S., Elveny, M., and Syah, R. (June 2021) Data science: A Review towards the Big Data Problems, *Journal of Physics: Conference Series* **1898**.

Nijhuis, M. (2022) Company Name Matching, *Medium*.

Nijhuis, M. and Van Lelyveld, I. (2023) Outlier Detection with Reinforcement Learning for Costly to Verify Data, *Entropy* **25**.

Rietveld, G., Lange, N., and Duijm, P. (2023) Measuring intra-bank complexity by (not) connecting the dots with LEI, *DNB Analysis*.

Ullersma, C. and Van Lelyveld, I. (2022) Granular data offer new opportunities for stress testing. In: *Handbook of Financial Stress Testing*. Ed. by D. Farmer, T. Schuermann, A. Kleinnijenhuis, and T. Wetzer. Cambridge University Press.

Van den Boom, B., Hofman, R., Jansen, K., and Van Lelyveld, I. (2021) Estimating Initial Margins – The COVID-19 market stress as an application, *DNB Analysis*.

Van Dijk, D. and De Winter, J. (2023) Nowcasting GDP using tone-adjusted time varying news topics: Evidence from the financial press, *DNB Working Paper* **766**.

World Economic Forum (2018) The New Physics of Financial Services.

Experiences, essentials and perspectives for Data Science in the heart of central banks and supervisors

A case study of the Dutch Central Bank

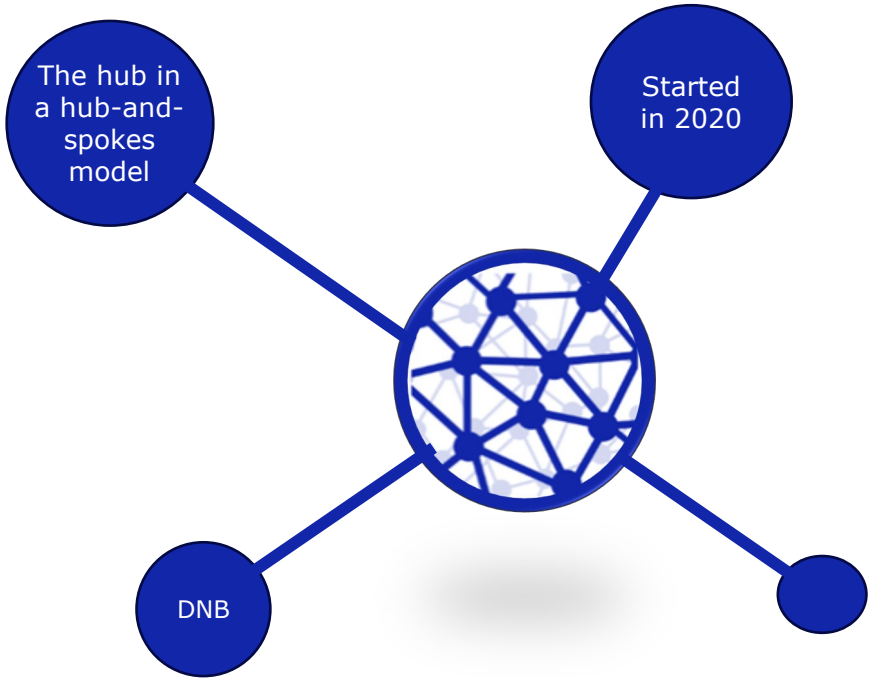
Patty Duijm and Iman van Lelyveld

DeNederlandscheBank

EUROSYSTEEM

DataScience Hub

Data Science Hub



This presentation

7 Experiences

5 Essentials



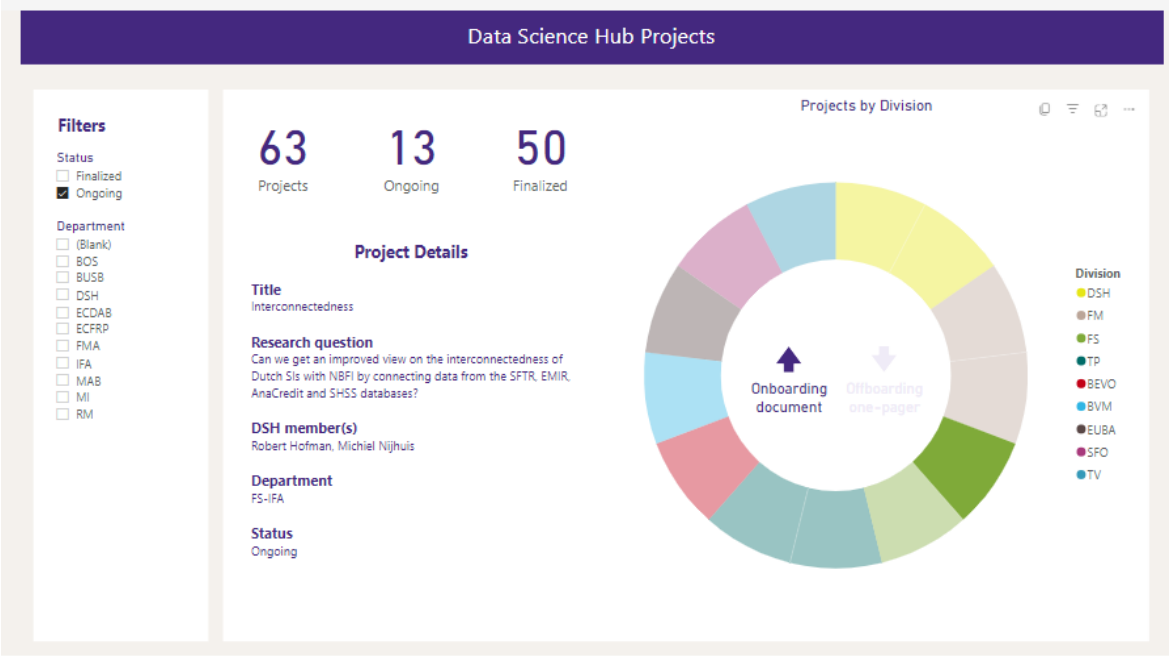
Way of Working

DeNederlandscheBank
EUROSYSTEEM

Seven experiences

Since the start of the DSH in 2020 we have worked on 63 data science projects.

Our experiences are all illustrated with examples from selected projects.



For more information on all our (finalized) project see our Annual Reports

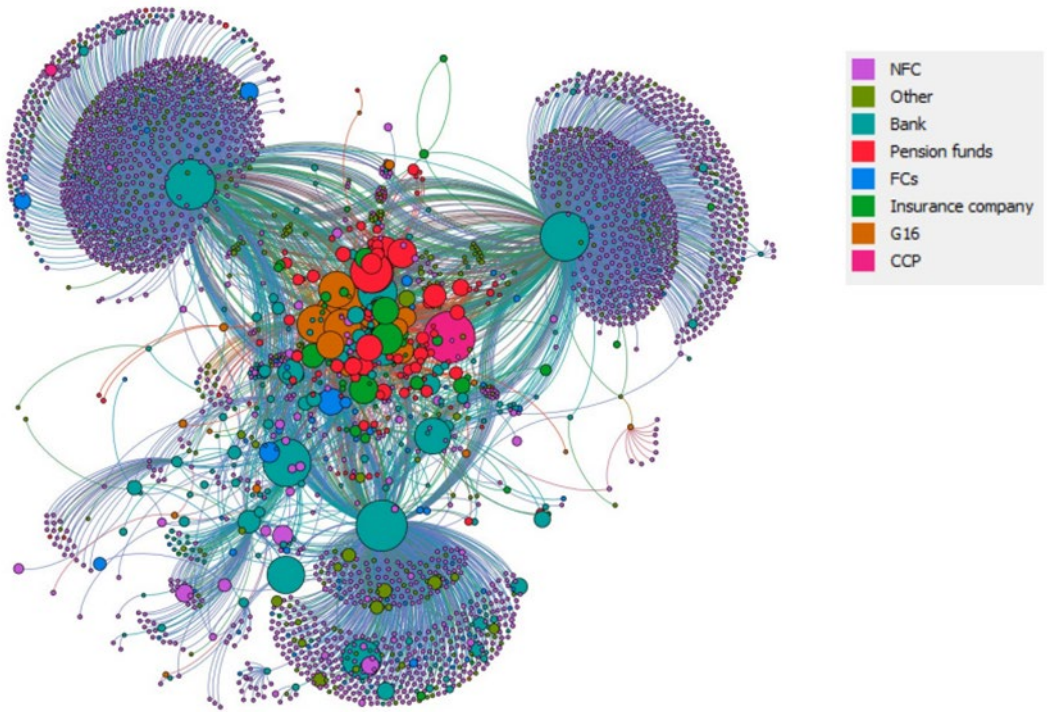
Granular data

*"The **combination of - sometimes novel - granular data** offers new insights useful for both supervisors and policymakers."*

Examples are SHS, Anacredit, EMIR, MMSR.

Three challenges:

- Counterparty names as free-format text
- No single unique identifier
- One needs to consider different sources of granular data



Project: Network analysis of interest rate derivatives

The colours of the dots indicate the different sectors and the sizes reflect the (logarithmic) aggregate size of the derivative positions.

The project is described in more detail in DNB Analysis "Estimating Initial Margins – The COVID-19 market stress" (Van den Boom et al., 2021)

DeNederlandscheBank EUROSYSTEEM

DataScience Hub

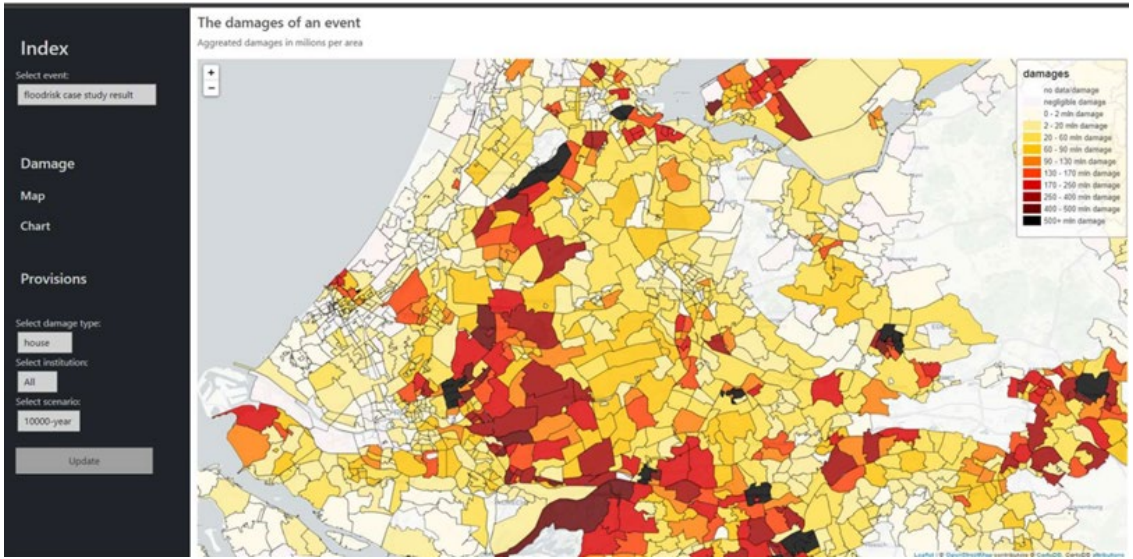
Combining internal and external sources

"Combining internal data with external resources increases the information value of the data, allowing supervisors and policymakers to improve the incorporation of external factors and risks."

The data you own is much more valuable to you if it is augmented with data owned by others
(Mewald, 2023)

Especially these "alternative" sources of data may generate unique insights.

DeNederlandscheBank EUROSYSTEEM
DataScience Hub



Project: Digital Twin of Climate Risks

In cooperation with the BIS Innovation Network, the Data Science Hub has developed a digital twin pilot of climate risks. The digital twin was developed to measure the effects of climate events on the financial system via real estate exposures of financial institutions.

More information can be found in the BIS note
"Summary of the workshop on the role of technology in finance" (BIS, 2023)

Automating data processes

"Automating data processes results in more efficient, accurate and highly frequent economic (prediction) models and allows policymakers to focus on modeling rather than preparing the data."

Manual data wrangling is a labor intensive job, prone to human errors.

Automating data collection processes results in more efficiency and accuracy.



Project: the DataFetcher

The DataFetcher is a package developed by the DSH and the figure illustrates its purpose. Multiple processes within the central bank use external data sources, resulting in colleagues collecting (the same) data manually or via ad hoc scripts. This may also result in cases where different (or even outdated) versions of the same data set are used within DNB. This is the left panel. In an ideal situation, i.e., the right panel, colleagues within the same institution have immediate access to the same data while restrictions dictated by privacy and confidentiality should be respected.

The package is available to ESCB NCBs.

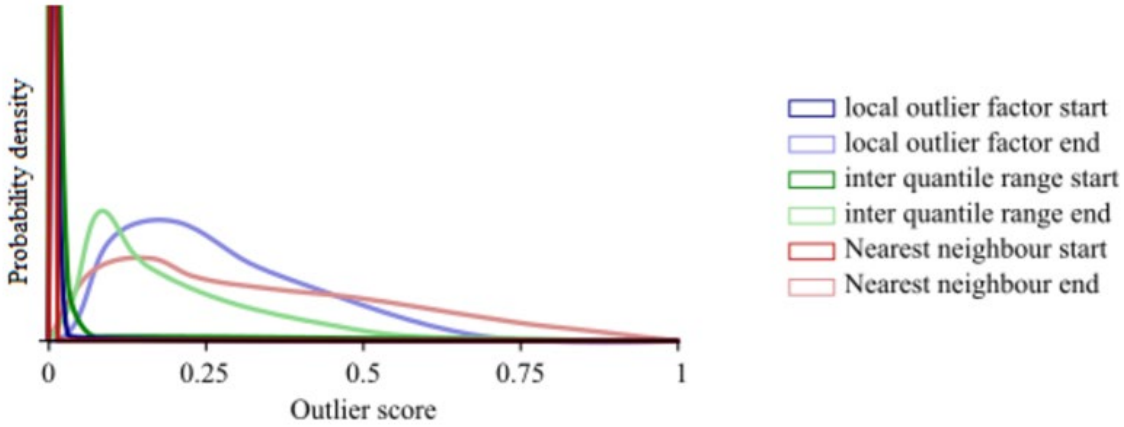
DeNederlandscheBank EUROSYSTEEM

DataScience Hub

Start simple

"Machine learning has great potential and outlier detection models have already proven to be effective innovations in supervision. Machine learning is, however, not a prerequisite for a successful data science project. In many cases, a simple model is a great place to start."

An outlier detection approach should be able to optimally use the knowledge gained from the verification of the ground truth and adjust accordingly.



Project: Outlier detection with reinforcement learning

In this project we applied a reinforcement learning on a statistical outlier detection approach. Using an ensemble of proven outlier detection models in combination with reinforcement learning we were able to tune the coefficients of the ensemble with each additional bit of data. The figure shows the distribution of outlier scores for the three different methods. In a reinforcement learning approach, an algorithm is not just trained and applied, but in each iteration, the algorithm gets feedback on its performance.

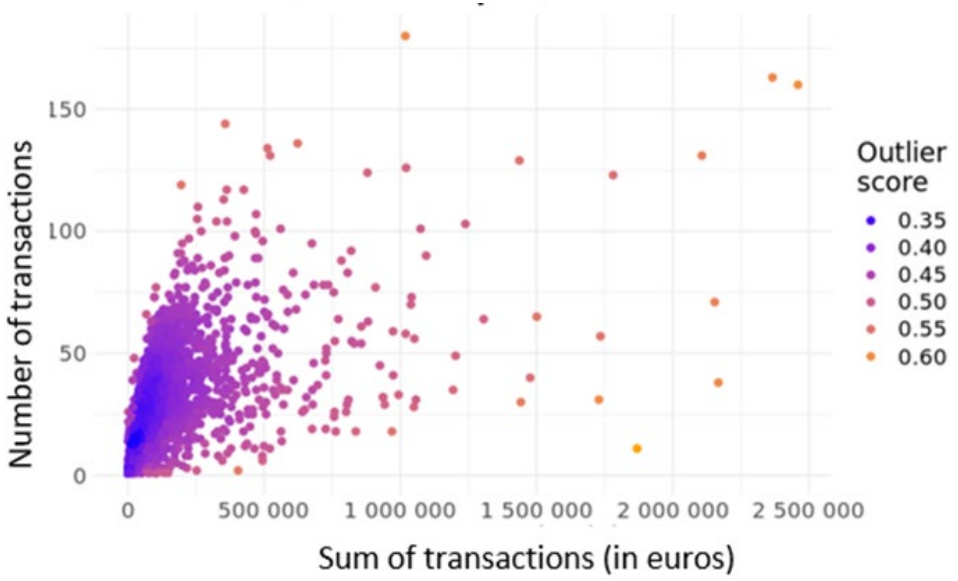
The project is described in more detail in the paper "Outlier Detection with Reinforcement Leanring for Costly to Verify Data" (Nijhuis and van Lelyveld, 2023)

DeNederlandscheBank EUROSYSTEEM
DataScience Hub

Domain knowledge

"The real value of data science lies in the combination of data science techniques and domain knowledge, and therefore perfectly illustrates the importance of domain experts in data science."

Domain knowledge is not only important as input for the model, but also to interpret model outcomes.



Project: Know Your Customer

In this project outlier detection is used to identify potentially fraudulent client transactions from a sample of bank clients. With the model, we were able to effectively select clients with an abnormal transaction profile. This project clearly shows the importance of domain knowledge. What are fraudulent transactions? It is relatively easy to classify an outlier as "fraudulent", while it is not. The graph shows outlier detection scores for bank clients, plotted against two of the client characteristics. The results of the outlier detection model resulted in the identification of new risks and efficiency gains, since supervisors are now able to consider all transactions instead of considering samples.

More information on this DSH project can be found in Cambridge Suptech Lab's "State of Suptech Report".

DeNederlandscheBank EUROSYSTEEM

DataScience Hub

Adoption by the business

"The value of data science applications depends on the adoption by the business and therefore user-friendly interfaces to integrate the data science solution into the daily workflow are as important, if not more important, as technical excellence."

Successful data science solutions can be both be one-off solutions or applications implemented in the workflow.

DeNederlandscheBank
EUROSYSTEEM

DataScience Hub



Project: False Unfit Banknotes

The aim of this project was to reduce the number of False Unfit Banknotes (i.e. banknotes classied as unfit by the cash handler, while deemed fit by DNB). While it is easy to compare the consequences of adjusting just one of the rules (e.g., tape decision or dirt), it quickly becomes more complicated once multiple rule settings are adjusted simultaneously. Machine Learning was applied to arrive at the optimal combination of multiple rule adjustments. Reducing the number of unfits can save a lot of effort and expense, and this project resulted in a set of recommendations for our Payments division to achieve these cost reductions.

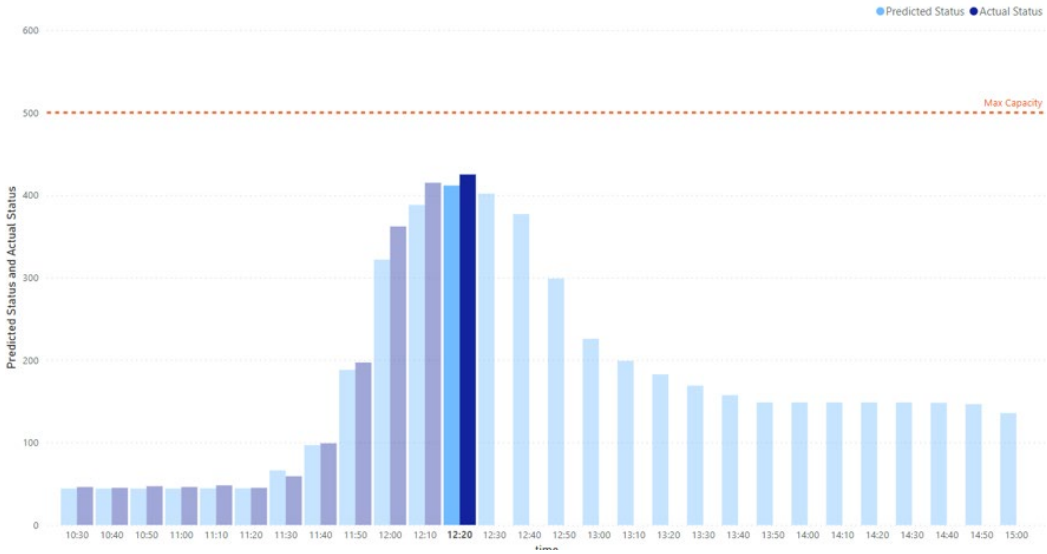
Data science for the entire organization

"Last but not least, data science has value for the entire organization, including HR and business operations, and should therefore be in the 'heart of the organization."

Don not only consider the traditional banking topics.

Data science can bring value for, for example, HR, internal services or the legal department as well.

DeNederlandscheBank
EUROSYSTEEM
DataScience Hub



Project: Sensor data

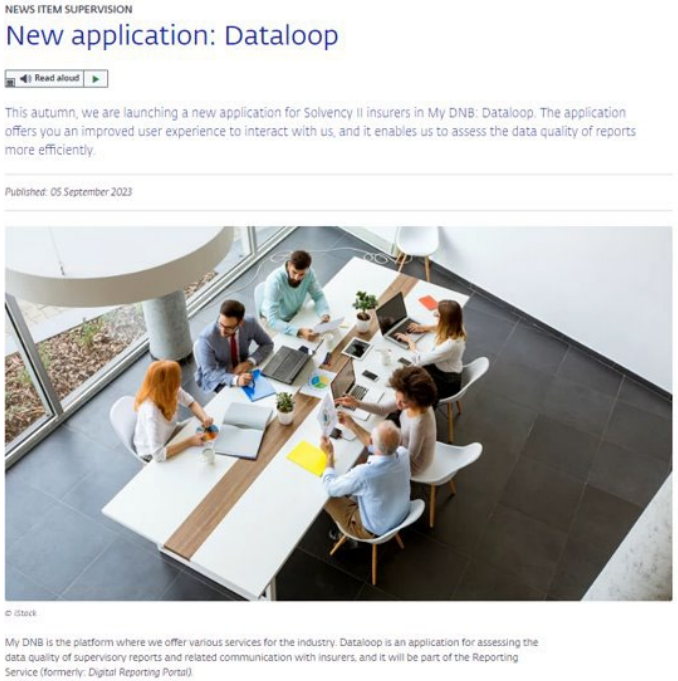
We are currently experimenting with motion sensors in our office building to predict how busy our cafeteria will be. Figure 9 shows the results of our prediction model in the current state.9 Such forecasts can help our catering to plan capacity and our staff make a more informed choice to time their lunch. Sensor data can, however, be used for other purposes, for example, to monitor the no-shows for meeting room reservations.

Five essentials

Dataloop

- Started as a pilot in 2018...
- Launced as an application to be used by both supervisors and supervised entities in 2023.

What are the essentials to integrate data science in the organization?



NEWS ITEM SUPERVISION

New application: Dataloop

Read aloud

This autumn, we are launching a new application for Solvency II insurers in My DNB: Dataloop. The application offers you an improved user experience to interact with us, and it enables us to assess the data quality of reports more efficiently.

Published: 05 September 2023

© iStock

My DNB is the platform where we offer various services for the industry. Dataloop is an application for assessing the data quality of supervisory reports and related communication with insurers, and it will be part of the Reporting Service (formerly: *Digital Reporting Portal*).

1. Common view

"Embracing new data science methods and using them throughout the organization requires sufficient appetite for experimentation -- especially at senior levels -- and therefore a common vision is key."

3. Responsible coding

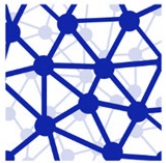
"Responsible coding is needed to ensure the replication and reproducibility of the work."

4. Data governance

"A mature framework for data governance is needed to work responsibly, and this should be embraced by the organization."

2. Right mix of skills

"A data science function should be able to combine many different activities, and this requires the right mix of skills."



DataScience Hub

5. Close contact with IT

"A well-established IT environment is needed to facilitate data scientists in all their needs."

Way of Working at the Data Science Hub



Measurable goals
Have a clear mandate and measurable KPIs



Inspiration & communication
Inspire people at all levels, make available all project documentation and communicate about successes.



Project design
Work in a structured way and agree upon division of tasks with onboardings and offboardings.



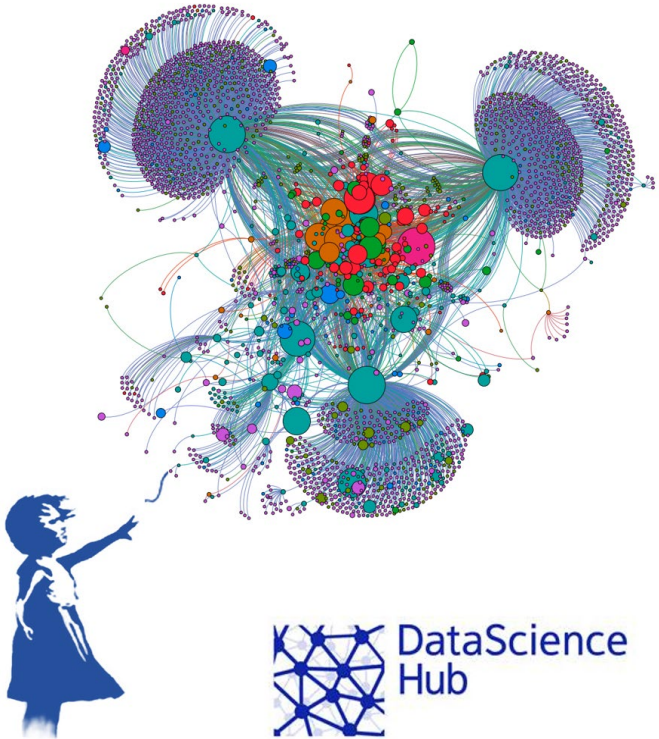
Activities and trainings
Build a data science community by offering regular and ad-hoc events.

DeNederlandscheBank
EUROSYSTEEM

Conclusion

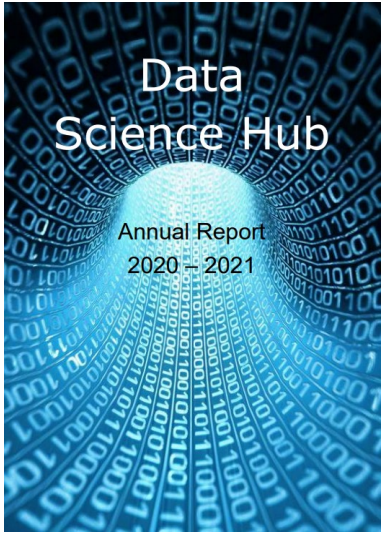
- A well-defined strategy, mandate, goals, and a structured way of working definitely helped us to apply data science within the organization.
- Just start! And don't think that getting data science to work for an organization can be achieved by hiring a few smart data geeks and having them develop "AI" in a remote corner of the organization.
- Share! "*Instead of sharing shiny PowerPoint presentations, we could share the functionality that allows us to replicate the analyses of the others with our own data" (Maijoor, 2022)*
    - *A quick win is to start sharing code, as we do through DNB's GitHub.*

Data science has not only great potential but is already valuable for central bankers and supervisors.



DeNederlandscheBank
EUROSYSTEEM

Relevant links



[Annual Report 2020-2021]



[Annual Report 2022]

GitHub

[DNB GitHub]

[data_science@dnb.nl]

DeNederlandscheBank
EUROSYSTEEM

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Digital transformation of financial regulators and the emergence of supervisory technologies (SupTech): a case study of the UK Financial Conduct Authority[1]

Pavle Avramović,
UK Financial Conduct Authority

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Digital Transformation of Financial Regulators and the Emergence of Supervisory Technologies (SupTech): A Case Study of the U.K. Financial Conduct Authority

Pavle Avramović

The ability of financial regulators to supervise the fast-evolving and technologically enabled world of financial services is uncertain. Scars from the 2007–2008 crisis remain and, with novel technological approaches being deployed at an unprecedented rate, the question of regulatory ability is often overlooked. How can financial regulators meet the challenges of this growing ecosystem and prevent harmful outcomes for consumers and markets? Could supervisory technologies (SupTech) be the answer?

This article provides a case study of the use of SupTech by the U.K. Financial Conduct Authority (FCA). Through 16 qualitative interviews with senior colleagues at the FCA and other regulators, this study focuses on how the field of SupTech has emerged and how the FCA developed its SupTech tool (BLENDER) to monitor market abuse in financial markets.

Novel institutional drivers of SupTech emergence are identified, centering on 'regulatory competition' and how regulators' mimetic behaviors result in competition and potentially contribute to further SupTech growth. Through a dynamic capability lens (Teece et al., 1997), the research also provides insights into the development journey of a SupTech tool.

The article provides an important foundation for the research of SupTech and calls for further explorations within this field to provide regulators with the insight needed to deliver their objectives in the interest of consumers, firms, and markets.

Keywords: digital transformation, SupTech, financial regulation, innovation

---

Media Summary

This article provides a case study of the use of supervisory technologies (SupTech) by the U.K. Financial Conduct Authority (FCA), one of the leading financial regulators in the world. Through 16 qualitative interviews with senior colleagues at the FCA and other international regulators, the study focuses on how the field of SupTech has emerged and how the FCA developed its SupTech tool (BLENDER) to monitor market abuse in financial markets. Novel institutional drivers of SupTech emergence are identified, centering on 'regulatory competition' and how regulators' mimetic behaviors result in competition and could contribute to further SupTech growth. Through a dynamic capability lens (Teece et al., 1997), the research provides a rare insight into the development journey of a SupTech tool and highlights key factors of sensing opportunities, visionary leadership, and flexible delivery frameworks as crucial for SupTech creation. The article provides an important foundation for the research of SupTech and calls for further explorations within this field to

provide regulators with the insight needed to deliver their objectives in the interest of consumers, firms, and markets.

---

1. Introduction

The financial services industry has historically been an early adopter of information technology (IT), due to its data-centric focus and close interrelationships between consumers, financial institutions, and intermediaries (Zhu et al., 2004). This adoption of IT fueled the increased digitization and digitalization of the sector, as firms looked for opportunities to improve processes and build new offerings (Alt et al., 2018). These factors together have resulted in the rise of 'FinTech,' "the fusion of finance and technology" (Goldstein et al., 2019, p. 1647). Financial services firms are rapidly digitally transforming and adopting technologies such as artificial intelligence (AI) and cloud computing to develop efficiencies and improve their service offerings (World Economic Forum, 2020).

On the other side of the fast-evolving financial services sector are financial regulatory bodies, organizations with crucial regulatory responsibilities to ensure well-functioning markets. These organizations with restricted budgets, limited technological ability, and growing mandates are expected to be able to regulate efficiently and effectively complex and often-changing financial markets (Financial Stability Board [FSB], 2017).

With such novel technological innovations, an important question raised is, are financial regulatory bodies able to fulfill their duties and ensure consumers are protected and financial collapses are avoided?

Digital transformations of government functions have been often explored, such as e-portals (Luna-Reyes & Gil-Garcia, 2014) and business-to-government reporting (Troshani et al., 2018). However, there is a research gap in information systems (IS) literature regarding the digital transformation of regulatory bodies, particularly within financial services regulation. Within the "Senior Scholars' Basket of IS Journals" (Association for Information Systems, 2011), a limited number of articles were found using a mix of keywords, including 'digital transformation,' 'regulatory authorities,' 'regulation,' 'regulatory agencies,' and others. Other, less recognized journals contained literature on this topic, but this demonstrable gap leaves regulators potentially ill-informed and underprepared for the challenges they are facing.

The question of overall regulatory ability, however, has been recognized in industry research. Global institutions such as the Bank for International Settlements (BIS, 2018), the Cambridge SupTech Lab (Cambridge SupTech Lab, 2022), consultancies (Oliver Wyman, 2017), and media (Gilman, 2021; Holden, 2020) explored how regulators have been slow to respond and institutionally bound in their activities, thus increasing the need to digitally transform in order to successfully regulate financial services. The crisis of 2007–2008 demonstrates the impact of regulatory failures; in the context of today's interconnected global economy, the impact of technologically driven failure could be significantly more severe.

One emerging response to this question is the development of the supervisory technologies (SupTech) field, which is defined by the World Bank (2021) as "the use of technology to facilitate and enhance supervisory processes from the perspective of supervisory authorities." Regulators are increasingly turning to technology to assist their supervisory activities.

This research article seeks to assess how the SupTech field emerges and develops within financial regulatory agencies to provide an adequate response to the challenges faced by the regulatory system. By focusing on the two elements of emergence and development, the aim is to understand and inform regulators of the driving factors of this novel field and how such technological approaches could be developed to support regulators. To achieve this, this article will focus on the experience of the UK Financial Conduct Authority (FCA), one of the more technologically advanced regulators in the world (Friend, 2021). In a case study format, it seeks to understand how the FCA has approached the development of a SupTech solution (BLENDER) through a dynamic capabilities lens (Teece et al., 1997), and what institutional factors have contributed to the growth of the SupTech field within the organization and more widely in the global ecosystem of financial services.

The article starts with an overview of the relevant literature to provide an understanding of the topic and a reflection of the theoretical foundation for further analysis. After exploring the research design and question, it provides key findings and sets out the BLENDER development journey. Finally, it considers the implications of these findings, namely, what these institutional drivers mean for global SupTech emergence, learnings for SupTech development, and important lessons for research, regulators, and industry.

2. Context

2.1. Digitalization of Financial Services and Regulatory Gap

Financial regulatory bodies refer to organizations responsible for delivering objectives 'in the public interest,' which often consist of protecting consumers and ensuring markets work well (Baldwin et al., 2011). Historically, these organizations have been perceived as key actors in the successful functioning of the financial services ecosystem. However, after the 2007–2008 crisis, there has been an increased focus on their role, remit, and activities, and increased attention on how they affect the lives of consumers and firms, particularly around the potential risks and effects of regulatory failures (Moshirian, 2011). This newly developed scrutiny has resulted in regulators being increasingly in the public eye and having progressively higher expectations of delivering their objectives, with technology playing a growing role in enabling their work (Cambridge Centre for Alternative Finance [CCAF], 2019).

Similarly, technology continues to be a key enabler for financial services firms, supporting the creation of new products and services at unprecedented rates (Organisation for Economic Co-operation and Development [OECD], 2018). The introduction of technology-enabled opportunities like Open Banking (Deloitte, 2017), the prospect of improving financial inclusion (Consultative Group to Assist the Poor, 2020), and the availability of fast

payments, which could be critical for time-sensitive incidents (BIS, 2016), provide manifold benefits for consumers and markets.

Nevertheless, technology drives complex challenges as well. For example, the lack of explainability of AI algorithms (Ostmann & Dorobantu, 2021), algorithmic trading and impact on market monitoring (Kunz & Martin, 2015), and the challenge of crypto-asset regulation (FSB, 2022) are all examples of difficult issues for regulators.

Therefore, it is to be expected that regulators, just like industry, are also turning to technology—for support in managing their growing portfolios of work and increasing their insight, efficiency, and effectiveness.

2.2. SupTech as an Emerging Field

Digital transformation has been a widespread effort within the public sector and government. The potential for technology to advance the delivery of public sector goals and objectives has been extensively researched (Dada, 2006; Fountain, 2001; Twizeyimana & Andersson, 2019; West, 2007). For financial regulators, however, the area is only now starting to be explored.

Industry research captures the digital transformation of regulators through the SupTech field. Multiple definitions currently exist (Zeranski & Sancak, 2020). These range from one of the original definitions by the Financial Stability Institute at the BIS as "the use of innovative technologies by supervisory agencies to support supervision" (Broeders & Prenio, 2018) to broader classifications. The rationale behind the introduction of this new term is to capture the extensive efforts by financial regulators across the world who use technology to assist their work programs.

In terms of types of regulators that are utilizing these tools, central banks and market regulators are often included and there is a broad regional split, with examples of more than 130 international regulators participating in global surveys on their use of SupTech (Cambridge SupTech Lab, 2022). SupTech has been mainly recognized within financial services; however, other sectoral regulators are exploring the applications of such approaches. For example, the Digital Regulation Cooperation Forum is creating a network to enable regulators to join up on the use of SupTech in the United Kingdom across communications, data privacy, competition, and financial services (Digital Regulation Cooperation Forum, 2022).

From a use-case perspective, SupTech represents a broad and complex field with a breadth of applications split across different regulatory functions. Research by the World Bank (2021) captured the diversity of potential SupTech applications, outlining the key regulatory activities and technologies being applied, with data being at the 'heart' of all efforts as a critical way for regulators to understand what is occurring in their respective markets. Figure 1 outlines the four primary sectors of AML/CFT (anti-money laundering/combatting the finance of terrorism) supervision, macro-prudential supervision (market monitoring), micro-prudential supervision (firm monitoring), and licensing (record management). These regulatory responsibilities are enabled through the combination of SupTech tools and data analysis. For example, prudential SupTech tools

have been critical during the COVID-19 pandemic, with more than 20 regulators experimenting with more than 70 different prudential tools to enable oversight of companies (Beerman et al., 2021).

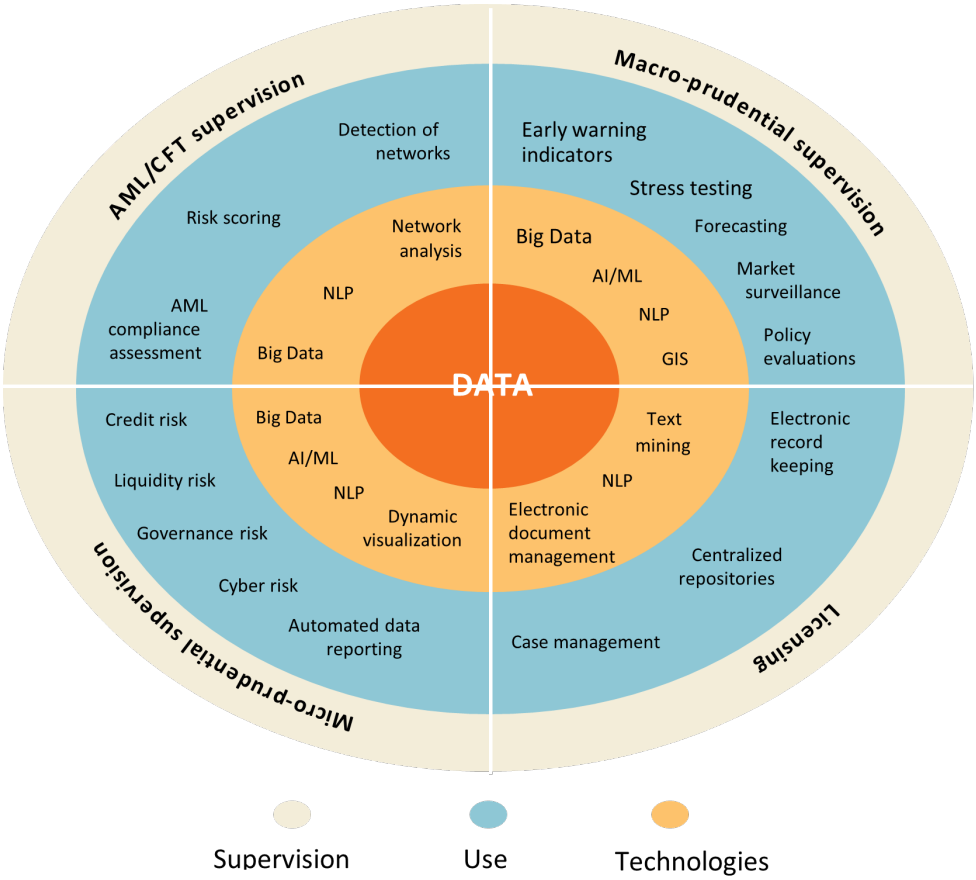


Figure 1. **SupTech taxonomy with SupTech use cases**. From Appaya et al. (2020).

Di Castri et al. (2019) capture the historical journey of SupTech in the influential "suptech generations" paper outlining the different maturity of technologies and regulatory processes. In their view, SupTech captures 'innovative' applications of technology (third and fourth generations), however, they acknowledge previous technological efforts and their contribution (Figure 2).

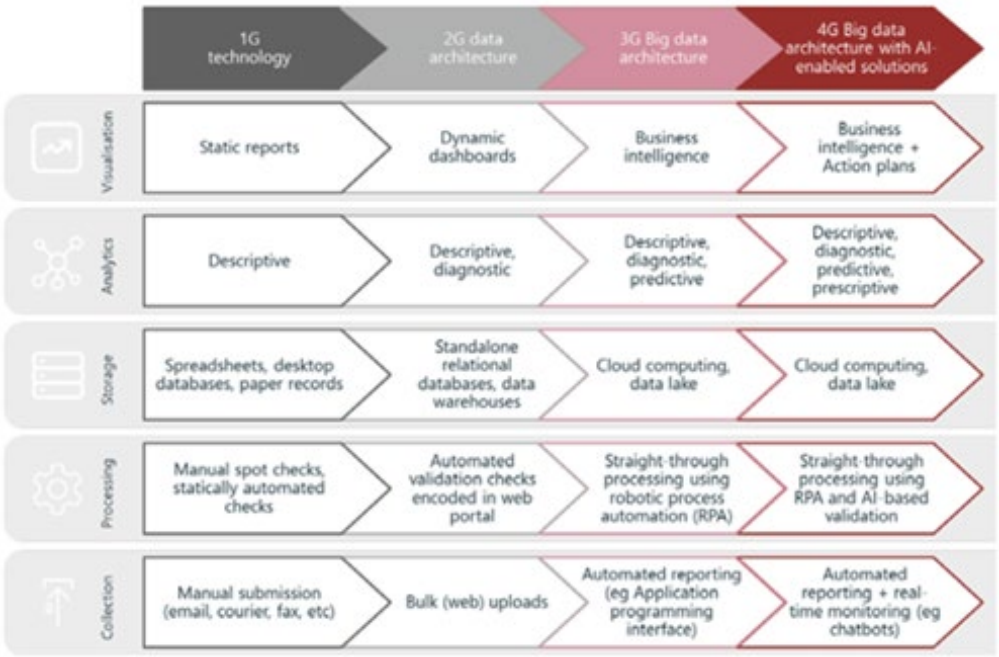


Figure 2. Evolution of technology within SupTech (Di Castri et al., 2019).

Industry and think tank research demonstrates that regulatory agencies around the world are working more closely to understand how this field is developing and how the usage of technology is assisting regulators to meet increasing regulatory supervision needs. Nevertheless, it should be acknowledged that regulators are still pursuing existing supervisory approaches to monitor markets and some technological applications could be considered out of scope for SupTech, and this should be addressed as the field matures and is further researched (FinCoNet [International Financial Consumer Protection Organisation], 2020; FSB, 2020; International Association of Insurance Supervisors and Access to Insurance Initiative, 2019; Toronto Centre, 2018; World Bank, 2018).

Despite the growing industry literature, there is partial information on how any one particular SupTech solution is developed. The focus of the industry research has been on showcasing the breadth and diversity in case studies of financial regulators. However, due to potential market risks of sharing too much information held by regulators, there have been limited explorations of SupTech development.

2.3. Theoretical Foundations—Institutional Theory and Dynamic Capabilities

Given the inherent novelty, complexity, and breadth of SupTech and the digital transformation lifecycle, the focus of this article is on both the emergence and development aspects of SupTech through the research question of 'how does SupTech emerge and develop within financial regulatory agencies?' From a theoretical perspective, this article uses a combined approach to answer this question, looking at institutional factors driving the emergence of SupTech and how SupTech is developed through a dynamic capabilities lens.

Firstly, it explores institutional factors that are affecting the emergence process, using literature on institutional theory and information systems (Currie & Swanson, 2009; Hinings et al., 2018; Orlikowski & Barley, 2001). As previously highlighted, SupTech's growth is on a global scale and most of the literature discusses the common and often duplicative approaches and challenges faced by regulators in SupTech development (Cambridge SupTech Lab, 2022; FinCoNet, 2020; FSB, 2020; OECD, 2019). This commonality provides a connection to new institutionalism and the isomorphic pressures that drive organizations to behave similarly. In this case, the isomorphic pressures drive regulators to pursue analogous approaches and use technology for the supervision of markets.(DiMaggio & Powell, 1983).

DiMaggio and Powell (1983) discuss in depth how institutions tend to copy and mimic each other, motivated by different pressures. These include meeting social expectations (coercive), tackling uncertainty (mimetic), or building legitimacy (normative). A relevant example is the mimetic behavior portrayed by financial services firms and regulators to create a crisis environment during the financial crash. It could be argued that a case of 'contagion of legitimacy' occurred between these two types of institutions in the United States (Riaz, 2009; Wolfson, 2013). The successes of firms caused a 'reverse-legitimacy' of the system and creation of 'blind spots' for financial regulators, resulting in the subsequent crash. Considering the similar approaches regulators are taking toward SupTech development, it could be valuable to explore the drivers and pressures for SupTech and understand the overlaps.

Initial comparison between the isomorphic pressures and identified drivers of SupTech development from literature highlight that both coercive and mimetic pressures exist, with limited normative pressures identified. Interestingly, mimetic behaviors in technology are already identified, but there is no underlying detail around the level of solution or approach similarity.

Table 1. Isomorphic pressures and SupTech; adapted from DiMaggio and Powell (1983).

Types of Pressures	**Description**	**Regulatory SupTech Examples/Equivalents**
Coercive Pressures	Result from both formal and informal pressures applied on organizations by other organizations upon which they are dependent and by societal cultural expectations	• Regulatory obligations and duty of operating effectively and efficiently (FSB, 2020) • Consumer attitudes have changed, and they have higher expectations of financial regulators and of their role in protecting consumers (OECD, 2019)
Mimetic Pressures	Result from uncertainty and encourage organization to imitate each other in order to manage the ambiguity of the	• Common supervisory needs of automating supervisory procedures and analysing datasets result in 'greater similarity' of SupTech tools

Types of Pressures	Description	Regulatory SupTech Examples/Equivalents
	environment	being developed (FinCoNet, 2020) • Overall availability of technology at an affordable price drives regulators to pursue similar technological approaches, for example, the move to cloud computing (FSB, 2020)
Normative Pressures	Result from the need for professionalization and a group wanting to define the conditions and methods of their work	• Not applicable to the SupTech field due to the inherent novelty. Potential future certifications and courses could be created (e.g., the Cambridge Innovation Leadership Programme [CSL1], 2022) as well as further efforts on SupTech standards. However, it is too early to assess at this stage due to lack of industry and academic evidence of such definitions of conditions and methods

Secondly, from a development perspective, dynamic capability theory (Teece, 2007; Teece et al., 1997) is used to understand the development process and how regulators enable digital transformation and creation of SupTech tools. Dynamic capabilities are organizational capabilities driven by internal and external factors (a dynamic environment). They require dynamic efforts by organizations to reconfigure their processes and assets to address the opportunities and risks (Eisenhardt & Martin, 2000). Although some challenge the dynamism of regulatory agencies and the public sector (Piening, 2013), the dynamic capabilities processes of sensing, seizing, and transforming align well with regulatory processes of how regulation is created and implemented.

The fast evolution of the financial services ecosystem contributes to the appropriateness of dynamic capability theory (Figure 3). The dynamism of the industry proportionally affects the dynamism of the capabilities being created. We can identify important links between the novel products and services being created by financial services firms, and financial regulators' opportunities to develop themselves to address these advances (Salvato & Vassolo, 2017). Aligning dynamic capabilities with 'technological capabilities' enabled by IT (rather than solely people or process-related capabilities) ensures the required breadth and alignment with the literature and the framework to interpret the development of the BLENDER tool (Li & Chan, 2019; Mikalef & Pateli, 2017).

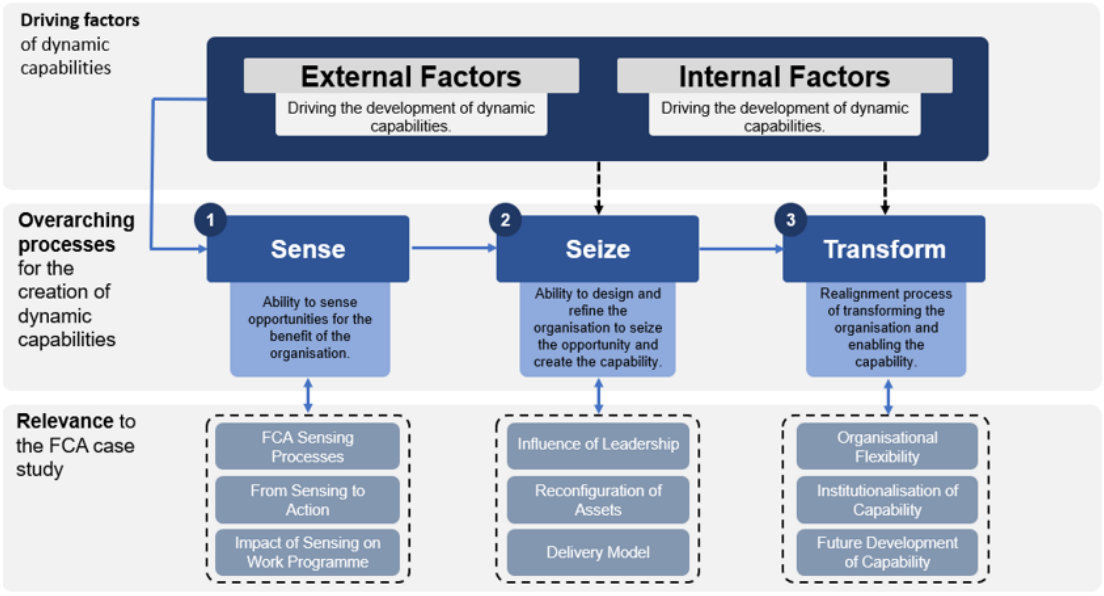


Figure 3. Dynamic Capability Framework. Adapted from Teece et al. (1997), Teece (2007), Teece (2018).

From a theoretical limitation perspective, institutional theory approaches in information systems research have been often applied with success. However, dynamic capability theory has been academically challenged due to its lack of specificity (Arend & Bromiley, 2009; Barreto, 2010). Also, as a strategic management framework, it was originally created with profit-making organizations in mind and research shows some limitations for using it for nonprofit/government bodies due to the lack of profit incentives (Bryson et al., 2007; Nelson & Winter, 2002).

Given these limitations, this approach combines these two theories to create a new theoretical framework (Figure 4). Between these two lenses, primary focus is given to the macro analysis to provide valuable background and scene-setting for the topic. The micro analysis offers the technical detail needed to understand the development process (which would not be possible through a singular theory approach). To validate usage of the theory for non-profit bodies, credible examples of research in public sector have been identified, with mentions of how the theory allows for the development of novel insights (Kattel & Mazzucato, 2018; Piening, 2013).

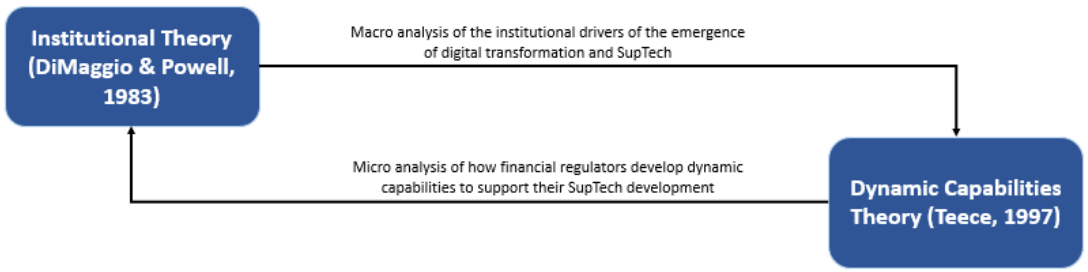


Figure 4. Research theoretical framework.

Even with the aforementioned industry explorations, there are limited academic studies to understand the drivers of SupTech. Similarly, there is a lack of detailed exploration of its development. A combined framework should provide a value-adding way of understanding the institutional isomorphic drivers of change and the dynamic capabilities creation process.

2.4. BLENDER History

This case study focuses on a singular project, the 'BLENDER' project within the FCA Market Oversight Division. It looks at the journey from its inception in 2013 to its current usage and future plans. BLENDER is a technology tool that assists supervisors to identify market abuse. Market abuse is behavior such as market manipulation, where market participants artificially affect the price of a stock, and these activities are enforceable by the regulator (FCA, 2020). BLENDER acts as 'middleware' and brings together data from different data sources (trading venues), 'blends' the data, and feeds it into a market surveillance tool to support supervisors to see transactions from multiple sources and identify patterns and diverse types of market abuse.

BLENDER was selected because it is a technology tool that enables the FCA to carry out one of its primary objectives of ensuring market integrity (FCA, 2016). It is also a finalized project, providing a complete 'picture,' and there was agreement to provide sufficient information from a confidentiality perspective. There are other SupTech tools at the FCA, however, due to market risks, limited information can be shared.

The BLENDER tool is an integral part of the market monitoring process of the FCA, without which it would not be able to have a holistic view of the market and see market manipulation occurring. It is a relevant example of a macro-prudential solution within the World Bank (2021) SupTech taxonomy and a late second generation / early third generation SupTech tool, according to Di Castri et al. (2019). Tackling market abuse is one of the primary objectives of many international financial regulators (European Securities Markets Authority, 2016), therefore, development of the BLENDER tool can be representative of SupTech tool development more broadly and provide insights. Lastly, it meets criteria from a dynamic capabilities perspective of being a capability that is driven by dynamic external and internal factors.

For context, BLENDER started in 2013 as a new project to create a technology application to help merge (blend) multiple data sources to a single data set. This was a result of the Markets in Financial Instruments Directive (MiFID) directive, which was created to increase competition for equity trading and stimulate innovation in the market (CFA Institute, 2009). As a result of MiFID, several new trading venues were being formed, creating a potential risk where the FCA is unable to have a 'helicopter view' of the market and the trading happening within it.

The Market Oversight Division, responsible for monitoring markets, agreed to run a proof of concept (PoC). This tested out how they could bring in the different data sources, merge them into a standardized format, and feed them into the overall FCA markets surveillance tool SMARTS to visualize the data and identify market abuse (NASDAQ, 2013).

Importantly, the PoC was delivered in-house, led by technology staff from the Market Oversight division itself, with limited input by the FCA IT division (allowing for the needed flexibility and speed to develop the tool, while also saving on cost). From 2013 to 2015, the PoC evolved to a pilot system with multiple venues feeding into it, followed by entering full production on the cloud in 2017. In 2018, BLENDER was migrated to the needed format to meet MIFiD 2 regulations. Today, it is an active part of the Market Oversight technology stack with potential plans of evolving it further.

3. Approach

3.1. Data Collection

The data collection approach for this research is qualitative, focused on interview data. The interviews allowed for a broad understanding of people's experiences and perspectives of digital transformation and SupTech, within the required structured method (Weiss, 1995). Interviews were conducted with 16 senior stakeholders from the FCA and other international regulators, using a specific interview protocol to understand their experiences in a consistent way with the conceptual theoretical framework as a foundation.

The interviews were delivered in a semistructured manner. This provided the opportunity for interviewees to share their experiences and dive deeper into areas that seemed more relevant for the research, while the open interview format put them at ease. The focus was on the interviewees' experiences, rather than their personal views and opinions. This was crucial to gain more tangible and credible insights into their role and how they partake in the digital transformation and SupTech work of the FCA.

This two-way approach provided a comprehensive view of the SupTech journey of the FCA by i) analyzing senior leaders' experiences and perspectives of how and why institutional factors drive the development of SupTech, and ii) understanding how the BLENDER project becomes developed through a dynamic capability lens.

3.2. Data Analysis

To support the research, a qualitative interview analysis approach was applied (Creswell & Poth, 2017; Gioia et al., 2013; Robson & McCartan, 2016). Through the interview data, themes were identified (Attride-Stirling, 2001; Braun & Clarke, 2020), focusing on how the information shared by the interviewees relates back to the literature and provides more information on the emergence factors and development journey of SupTech at the FCA.

A similar thematic approach was followed for both levels of the data analysis. Following transcription of the interviews, the data was codified, identifying the basic, organizing, and global themes (Attride-Stirling, 2001). By juxtaposing the relevant themes and wider context, several principal thematic networks were created to provide a consistent way of analyzing the data using literature and the theoretical framework. An example of a thematic network can be found in Appendix A.

For the contextual analysis of the institutional drivers of the emergence of SupTech, the focus was on identifying areas where the data is 'overflowing' the literature, looking for the novel factors that could be impacting the environment and growth of this field. The themes were analyzed in relation to the isomorphic drivers (DiMaggio & Powell, 1983) to understand which drivers were already recognized in literature (and which ones were not), while taking into account perspectives from FCA interviewees and international stakeholders to see potential overlaps and differences.

For the BLENDER tool, a dynamic capability lens (Teece, 2007; Teece et al., 1997) was applied, assessing how the thematic networks and interview data overlap with the key concepts. Overarching processes of sensing, seizing, and transforming were considered in the BLENDER context, in addition to understanding the external and internal drivers that motivated its development.

4. Findings

4.1. Regulatory Competition—A SupTech Institutional Driver

DiMaggio and Powell (1983) discuss the three isomorphic pressures that drive organizations into action. Table 2 outlines relevant data points in comparison to previously identified literature and how interview participants perceived these pressures. Most of the captured types of pressures are commonly found in literature and do not represent novel knowledge. Participants discussed how the focus on digital transformation and SupTech stemmed from a variety of factors—including the impact of statutory obligations, the need for regulators to be proactive in protecting consumers (Friend, 2021), and the availability of technology to support regulatory activities when it is untenable to do so manually and when it could enable better decision-making (Di Castri et al., 2019).

Table 2. Isomorphic pressures and findings; adapted from DiMaggio and Powell (1983).

Types of Pressures	**Description**	**Regulatory SupTech Examples/Equivalents**	**Findings Equivalents**
Coercive Pressures	Result from both formal and informal pressures applied on organizations by other organizations upon which they are dependent and by societal cultural expectations	• Regulatory obligations and duty of operating effectively and efficiently (FSB, 2020) • Consumer attitudes have changed, and they have higher expectations of financial regulators and of their role in protecting	• *So, I think it is a big piece about effectiveness; how we use data and automation to be a more effective regulator and then also a more efficient regulator (Interview 1)*

Types of Pressures	Description	Regulatory SupTech Examples/Equivalents	Findings Equivalents
		consumers (OECD, 2019)	• *The increased ability to produce and derive information from data… All of that leads to more efficiencies and new possibilities, new opportunities. And also leads to a change in expectations from institutional and retail actors and consumers (Interview 13)*
Mimetic Pressures	Result from uncertainty and encourage organizations to imitate each other in order to manage the ambiguity of the environment	• Common supervisory needs of automating supervisory procedures and analyzing data sets result in 'greater similarity' of SupTech tools being developed (FinCoNet, 2020) • Overall availability of technology at an affordable price drives regulators to pursue similar technological approaches, for example, the move	• *I see, I see a lot of similarities. As I mentioned before, some of the international regulators are further advanced than we are, and some are less advanced. But that being said, when we do speak to them, it tends to be that we're all following a similar approach, and*

Types of Pressures	Description	Regulatory SupTech Examples/Equivalents	Findings Equivalents
		to cloud computing (FSB, 2020)	*facing similar challenges (Interview 5)* • *I think that the main reason is that the technologies are there now. I mean, it's available, readily available. And, you know, financial authorities, regulators recognize the benefits these technologies bring (Interview 14)*
Normative Pressures	Result from the need for professionalization and a group wanting to define the conditions and methods of their work	• Not applicable to the SupTech field due to novelty. Potential future certifications and courses could be created, as well as standards	• *None*

However, one of the institutional factors that stood out is *regulatory competition*: a novel factor not found in literature. Participants noted the focus on SupTech and usage of technology to support supervisory processes was also connected to competition between regulators. Regulators could be seen as copying each other's efforts (mimetic behavior) and attempting to create the best SupTech solutions to keep up with the evolving market. Through an advanced SupTech portfolio, regulators can benefit from a well-functioning market, reduce regulatory burden for firms, and improve the overall attractiveness of the market itself. They are able to highlight an innovative national financial services ecosystem where they have a reputation of being advanced and able to deal with the fast-developing financial services and FinTech market with potential national benefit.

These perspectives came from FCA participants:

I think inevitably that there is competition, this is a quirk of regulators, but it is definitely not hostile competition. (Interview 8)

… a market that is open for business and a market that is good to do business, in an increasingly global world, increasingly, in a global world, that has to be, I think, a competitive advantage for a lot of regulators. (Interview 3)

And participants from international regulators also reflected:

I think some countries are much further ahead than others, some will put more emphasis on it, they may see it as a competitive advantage. (Interview 15)

And so, we all work together, but if one (a regulator) is doing something cool, then there is immediate pressure on all the others to do something similar. (Interview 13)

This is not the only driving factor. Many of the participants spoke about the potential for regulators to more actively collaborate and how this could also be observed as 'healthy' competition. Nevertheless, it is interesting to consider how mimetic behavior by regulators, in the face of uncertainty, could be driven by competition, and whether such competition could either lead to divisions and lack of sharing or lead to higher collaboration due to common goals.

4.2. BLENDER as a Dynamic Capability

The overarching finding based on the data is that the BLENDER tool can be perceived as a dynamic capability for the FCA because it was driven by external factors (e.g., change in regulations) and internal factors (e.g., senior management decision-making). A detailed breakdown can be found in Appendix B.

4.2.1. Sensing

Within the dynamic capabilities framework, sensing is an important starting point, and it is the ability of organizations to sense opportunities and potential risks in dynamic markets. Within the wider FCA interview participants, their views on FCA's capability to sense opportunities and risks were mixed. However, for the BLENDER project, the consensus was that the FCA's sensing ability worked better than usual. The risk of cross-market manipulation was identified, and with the United Kingdom having the largest market fragmentation and highest number of venues emerging (CFA Institute, 2009), there was a recognition that something needed to be done.

We at the FCA recognized the risk that people could do transactions on multiple venues, and nobody would ever join the dots. BLENDER was the organizational response to that. (Interview 10)

The fragmentation in the markets opened up opportunities for new types of abuse… so the unique thing BLENDER did was to bring those venues, the data from those venues into one place, so we could supervise them in one place. (Interview 9)

4.2.2. Seizing

After the successful sensing process, the more complex question was the seizing element and how the opportunity could be addressed through products, processes, and services (Teece, 2007). Multiple factors contributed to the FCA's ability to seize the opportunity: the 'visionary' leader, a strong investment case, and a flexible governance framework. Within the team, an individual drove the development of BLENDER and ensured that the opportunity was realized and actioned:

> *We had somebody who spotted the opportunity and brought people in specifically to find that to exploit those. (Interview 9)*

> *'XYZ' was sort of a maverick genius, a visionary who had a lot of incentive to get things done. He wanted to do things that nobody else had done. And he was willing to give people free rein to let them experiment… (Interview 11)*

Furthermore, the team came to an agreement to develop the tool with limited input from the IT division, resulting in faster working and more customization and flexibility. From a dynamic capabilities perspective, it was critical that the structural assets (governance), alignment, and organizational boundaries between the two teams were clear and respected:

> *The good thing about being developed in-house meant there wasn't the level of governance around it. It was run as a proof of concept and that did allow it to remain very flexible, very nimble, and managed by a business team fundamentally, rather than an IT support team. (Interview 12)*

4.2.3. Transforming

BLENDER was successfully deployed and delivered and is used by a number of FCA staff to support daily market monitoring.

> *BLENDER exists now as a part the wider ecosystem. (Interview 8)*

BLENDER as a capability can be seen to have gone through the full cycle, with a team using the data through the complementary visualization tools and providing important insight into the organization. However, it was clear from the participants that there is no endpoint and that the goal for the team is continuous development of the BLENDER tool and ways of working. This is aligned with the overall view from Teece that an organization needs to be able to continuously sense and seize opportunities, to be able to react when new threats and opportunities occur (Teece, 2007).

4.3. Imitation, Collaboration, and Institutionalization

Seeing the institutional drivers of SupTech and the evolution of BLENDER as a dynamic capability, it is important to consider the overlaps between these areas, and what novel findings they might contribute to the field of SupTech.

It can be argued that there is a relationship between mimetic drivers, the development of dynamic capabilities, and their inimitability principles. Teece (1997) sees inimitability of

dynamic capabilities as a crucial factor. This includes the development of tacit knowledge to limit the ability to copy work and the use of intellectual property to protect them. From the broader perspective of SupTech development, this raises important questions, as regulatory participants often discussed how copying each other's approaches were important ways for developing their own work programs, for example:

> *The FCA is repeatedly lionized and lauded and referenced as having been emulated and copied by other regulators. What is less frequently mentioned is the FCA has copied and emulated others, maybe not in the regulatory space, but in the retail space or in the financial markets themselves, and the academic space. So, there is no shame about borrowing or copying. (Interview 2)*

The answer to this challenge of inimitability could potentially be through collaboration — regulators can explore how they can co-create solutions and work together. Research has shown that regulators often seek support from other financial regulators in developing their approach to innovation (World Bank & CCAF, 2020). Participants often spoke about the great opportunity there is for closer partnerships between regulators:

> *I suppose that collaboration is key. In this area and a new emerging field, we will be facing the same challenges and we can learn from each other and collaboration with knowledge sharing can be beneficial. (Interview 5)*

Seeing the interaction between emergence and development, and the organizational focus to continuously improve, we can also reflect on the following step and how SupTech approaches become institutionalized in organizational ways of working and gain legitimacy (Hinings et al., 2018; Tolbert & Zucker, 1996). Participants noted that the FCA is still in the early phases of adopting technology. Exposure to data and digital technologies vary depending on factors such as the digitization of the sector and firm types. With the breadth of the SupTech field itself, it may prove difficult to conceptualize the legitimization and institutionalization of SupTech. However, it is still early to judge this, and participants noted that further scaling could drive adoption.

> *The future is already here, but it's unevenly distributed. And that's where we are at the moment, right? Like, there are examples of future practice and process occurring, but it is not, by any means, uniform across the organization. And that's the key thing, scaling up and across. (Interview 11)*

The institutionalization of wider SupTech is uncertain, however, if we reflect on BLENDER, it provides functionality and crucial data that support the operation of the team. Without the tool, it would not be possible to monitor markets. This dependence on the tool could be perceived as it being legitimized in the eyes of the users. Time will tell what the tool's future evolution and developments will have on its legitimization.

5. Discussion

This research focuses on seeing how SupTech emerges and specifically develops within financial regulatory agencies, in this case, within the FCA. To answer this question, we will

reflect on what the findings mean for the literature and the theoretical framework used, to understand the impacts on industry and academic research.

5.1. Regulatory Competition Expanded

Looking at the research question 'How does SupTech emerge and develop within financial regulatory agencies?' the approach taken was to look at areas where the findings provided novel insights not previously explored by the literature and identify what additional factors are driving the development of SupTech from an institutional lens. FSB (2020) carried out one of the more detailed explorations of the drivers of SupTech, in their case, what are the driving factors of a SupTech strategy (Figure 5). There were many overlaps with the interview data, from the drive for efficiency and effectiveness, to improving insights and availability of technology such as artificial intelligence and machine-readable data.

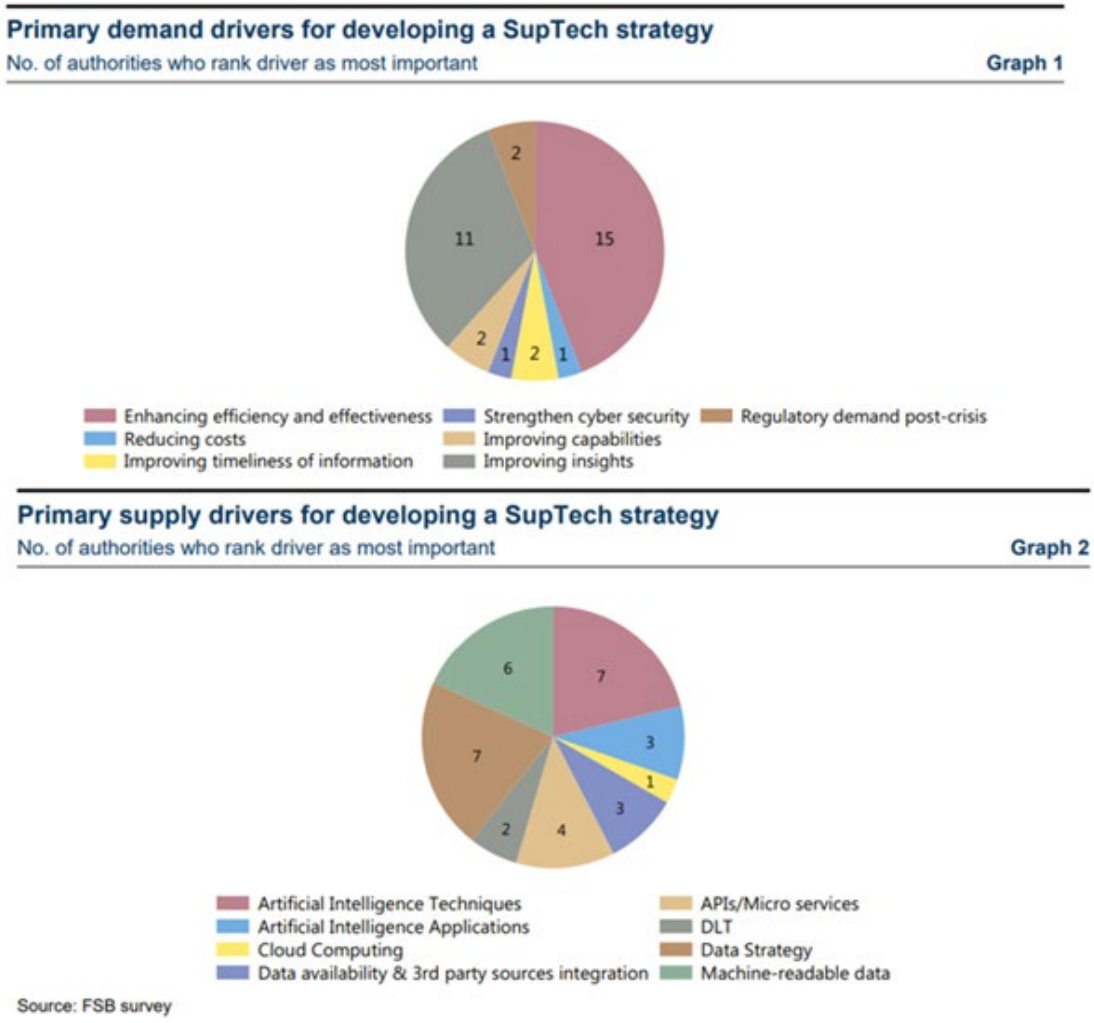


Figure 5. Demand and supply drivers of a SupTech strategy (FSB, 2020).

As outlined previously (Table 2), these drivers could be assessed from an institutional perspective (Di Maggio & Powell, 1983) and the majority would fall under coercive

pressures—regulators have a duty to fulfill, as expected by legal obligations and societal expectations, and they will do what they can to meet them.

However, an area that has not been recognized yet in SupTech literature and has been found in the data is 'regulatory competition,' originating from a combination of two factors. Firstly, aligned to mimetic pressures that cause market participants to copy each other to manage the ambiguity of the environment, interview participants referred to imitative behavior between each other, regulators 'coalescing' around SupTech, and starting multiple initiatives to explore how SupTech can support and improve their work (Cambridge SupTech Lab, 2022).

Secondly, participants reflected on how there is a potentially healthy dose of competition between them as a part of this SupTech development. This could stem from different reasons, for example, regulators wanting to improve the reputation of their organization, which has historically proven to be important for regulatory agencies and a way to build trust with industry and consumers (Gilad & Yogev, 2021; Yeung, 2009). Equally, it could also be driven by regulators wanting to create well-functioning markets to potentially drive investment and attractiveness. Some evidence exists that well-regulated markets could improve investment (Sinha, 2012). In the context of the United Kingdom, regulation has historically been seen as a way to ensure a well-performing market (Competition Markets Authority, 2020). This gains more relevance with the United Kingdom having left the European Union, and the newly assigned secondary international growth and competitiveness objective for the FCA, which will require the FCA to have regard to support the government's ambition to encourage economic growth (FCA, 2021; Norton Rose Fulbright, 2021; UK Gov, 2022). This poses an important question regarding the relationship between political will and the potential impact on regulatory behavior and competition between markets.

Furthermore, regulatory competition and mimetic behavior might provoke other types of reactions by regulators. For the field of SupTech, there is limited public information, both in literature and in industry reports. This could stem from regulatory opaqueness and limited transparency, due to the potential risks of outlining how specific technological approaches function (World Bank, 2019). This was also reflected in the data and could impact the amount of copying and mimetic behavior that can be carried out.

In turn, this potentially results in a higher spirit of collaboration between regulators. Being limited with how much information regulators can share externally may inspire greater levels of working together. As regulators spend more resources on developing SupTech, there could be an opportunity of collaborating either bilaterally or through international networks such as the recently launched Cambridge SupTech Lab (CSL2, 2022), Global Financial Innovation Network's RegTech & SupTech workstream (GFIN1, 2021), or the Bank of International Settlements Innovation Hubs (BIS, 2022). Similarly, regulators and industry have also reflected on the opportunity of opensource and how it could aid SupTech development globally (FINOS, 2022]; GFIN2, 2021).

5.2. Sensing, Visionary Leadership, and Delivery Model

SupTech is a novel and complex area with multiple factors contributing to its development. Focusing on the FCA and the BLENDER project, several more detailed intricacies are provided on how SupTech develops. As we saw in the findings, the progress of the BLENDER project (and the tool itself) could be interpreted through the overarching processes of dynamic capabilities of sensing, seizing, and transforming (Teece, 2007; Teece et al., 1997).

In analyzing BLENDER, it appears that the tool's development was linear. There were particular elements that enabled the BLENDER tool to function as a capability for the FCA.

5.2.1. Sensing

Firstly, important processes that enable successful regulation rely on horizon scanning and anticipatory regulation, the ability for regulators to be able to identify issues early and react (National Audit Office, 2021; Regulatory Horizons Council, 2022). These processes could be paired with the sensing aspect of dynamic capabilities because they allow for the identification of opportunities to improve the organization in hand. Interview data demonstrated mixed feedback for the FCA's overall ability for this process, however, for BLENDER, the sensing elements of a proactive attitude, in-house regulatory expertise, and adoption of academic insight made a difference and allowed for the FCA to react appropriately.

5.2.2. Seizing

Similarly, for the seizing process, the investment case and senior buy-in allowed for the further development of BLENDER. The more important factor was the governance model of allowing the in-house build to be driven by the Market Oversight Department, rather than the IT division. West (2007) reflects on the importance of flexibility and speed in government digital transformation, and it has proven crucial in the case of BLENDER as well. Moreover, for the successful seizing process, Teece (2018) discusses the managerial impact and the ability to seize the resources appropriately; with BLENDER, we saw the impact of the project leader and how their visionary drive and capability was instrumental in developing BLENDER.

5.2.3. Transforming

BLENDER was completed and is now a regularly updated tool and enabler for the FCA to collect and 'blend' data from multiple venues to identify potential market abuse. As mentioned, it has become a 'part of the ecosystem' and has achieved legitimacy in the eyes of its users (Hinings et al., 2018), but questions are raised around the impact of continuous transformation and development on its legitimization.

5.3. Implications for Regulators, Academia, and Industry

Regulatory competition is a known concept within policy development literature, and it has been explored how it can affect regulatory–government relations, standards development,

and regulatory arbitrage (Frantz & Instefjord, 2018; Haufler & Maier, 2019; Radaelli, 2004). Nevertheless, the acknowledgment of regulatory competition within SupTech development could act as a positive driver of organizational change for regulators. This could take form as increased information sharing between regulators to support knowledge building and increased collaboration between regulatory bodies through the mentioned different networks to advance SupTech (for example, potential efficiencies through the sharing of code).

In regard to BLENDER, due to the U.K. specific position and high number of trading venues, there have not been similar tools identified in other markets, therefore, not all lessons will be applicable to other international regulators. However, key insights of (i) regulatory recognition of technology as an enabler, and (ii) innovative and agile in-house development approaches, which supported the iterative build of BLENDER, are still relevant. These insights can be applied to other regulators and combined with the understanding of the criticality of horizon scanning, visionary leadership, and flexible delivery models, a version of a SupTech roadmap is provided. With the FCA's advanced technology portfolio, this can inform both advanced financial regulators and those at the start of their digital transformation and SupTech journeys.

From an academic research perspective, one of the important implications is the recognition that SupTech and digital transformation of financial regulators are compelling areas to pursue further research with high benefits. More research can support the development of SupTech and inform much-needed standards. For example, as technological adoption evolves in industry, it can be valuable to understand how the evolution of SupTech is impacted by technological progress made by firms.

Furthermore, the dynamic capabilities framework provides a useful lens to interpret this phenomenon and reinforces previous research of dynamic capabilities in the public sector (Kattel & Mazzucato, 2018; Piening, 2013). In relation to industry, an important lesson for industry firms is to continue engaging with regulators and explore what different opportunities exist either to support regulators in their SupTech journey or to share insights to inform regulatory efforts. We are already seeing the FCA driving forward its 'RegTech-SupTech bridge' initiative, which explores how to jointly develop solutions with industry and many regulators are increasing their engagement efforts to hear about innovation in the industry (Australian Transaction Reports and Analysis Centre, 2018; FinCEN, 2022; Rusu, 2022).

6. Conclusion

The experience of one regulator will not be the 'golden truth' for all regulators and it will not highlight every single opportunity or challenge that could inform other organizations' SupTech development. However, this research highlights two important discoveries that could be impactful for the industry.

The first of the findings is that SupTech emergence is partially driven by regulators being competitive organizations, who want to deliver positive outcomes for a plethora of reasons,

some regulatory and some institutional. The realization of its existence might result in more open information sharing and more collaboration between regulators. In addition, it might support the establishment of further certifications and standards that are needed to formalize the field of SupTech and the required focus on assessing the success of SupTech development, ensuring the deployments are cost-effective and impactful.

Secondly, the research provides a comprehensive overview of one regulator's way of developing a SupTech tool and way of operating in the face of changing circumstances. It demonstrates how IS development foundations are applicable to regulators, and that leadership, delivery models, and senior stakeholder buy-in are critical for regulatory IT development and data science. It provides a blueprint for how regulators could think about the changing environment and highlights how the dynamic environment in which financial services regulators operate could result in dynamic efforts by the regulators themselves.

With the constant evolution of markets, regulators will have to leverage technology to support their activities (Cœuré, 2020). Consequently, academia has a critical observer role to play in collaborating with regulators and informing them through research of the potential opportunities, risks, and harms technology can pose.

Acknowledgments

I would like to sincerely thank my parents, Milka and Zoran, my sister Hana, and my friends for their unwavering support in delivering this research. I would also like to thank my interviewees and Professor Susan Scott for her guidance, honesty, and encouragement.

Disclosure Statement

This research does not represent the views or opinions of the Financial Conduct Authority (FCA). The author is an employee of the FCA, but this research was conducted independently.

References

Alt, R., Beck, R., & Smits, M. (2018). FinTech and the transformation of the financial industry. *Electronic Markets*, *28*(3), 235–243.

Appaya, M. S., Dohotaru, M., Ahn, B., Kliatskova, T., Seshan, P., Pascaru, I. (2020). *A Roadmap to SupTech Solutions for Low Income (IDA) Countries (English)*. World Bank Working Paper No. 153566. http://documents.worldbank.org/curated/en/108411602047902677/A-Roadmap-to-SupTech-Solutions-for-Low-Income-IDA-Countries

Arend, R., & Bromiley, P. (2009). Assessing the dynamic capabilities view: Spare change, everyone? *Strategic Organization*, *7*(1), 75–90.

Association for Information Systems. (2011). *Research - Senior Scholars' Basket of Journals*. Retrieved April 23, 2001, from https://aisnet.org/page/SeniorScholarBasket

Attride-Stirling, J. (2001). Thematic networks: An analytic tool for qualitative research. *Qualitative Research*, *1*(3), 385–405.

Australian Transaction Reports and Analysis Centre. (2018, June 26). *AUSTRAC brings industry together with RegTechs to support innovation.* https://www.austrac.gov.au/austrac-brings-industry-together-regtechs-support-innovation

Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding regulation*. Oxford University Press.

Barreto, I. (2010). Dynamic capabilities: A review of past research and an agenda for the future. *Journal of Management*, *36*(1), 256–280.

Beerman, K., Prenio, J., & Zamil, R. (2021). *Suptech tools for prudential supervision and their use during the pandemic*. FSI Insights on Policy Implementation No. 37. Financial Stability Institute. https://www.bis.org/fsi/publ/insights37.pdf

Bank for International Settlements. (2022). *BIS Innovation Hub Centres*. Retrieved July 31, 2022, from https://www.bis.org/about/bisih/about.htm.

Bank for International Settlements. (2016). *Fast payments - Enhancing the speed and availability of retail payments*. Committee on Payments and Market Infrastructures. https://www.bis.org/cpmi/publ/d154.pdf

Bank for International Settlements. (2018). *Sound practices: Implications of FinTech developments for banks and bank supervisors*. https://www.bis.org/bcbs/publ/d431.pdf

Braun, V., & Clarke, V. (2020). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352.

Broeders, D., & Prenio, J. (2018). *Innovative technology in financial supervision (suptech) - The experience of early users*. FSI Insights on Policy Implementation No 9. Financial Stability Institute. https://www.bis.org/fsi/publ/insights9.htm

Bryson, J., Ackermann, F., & Eden, C. (2007). Putting the resource-based view of strategy and distinctive competencies to work in public organizations. *Public Administration Review*, *67*(4), 702–717.

Cambridge SupTech Lab. (2022). *State of SupTech report 2022*. *Cambridge Centre for Alternative Finance (CCAF),* University of Cambridge. www.cambridgesuptechlab.org/SOS

Cambridge Centre for Alternative Finance. (2019). *Early lessons on regulatory innovations to enable inclusive FinTech: Innovation offices, regulatory sandboxes, and RegTech.* Cambridge Centre for Alternative Finance and Office of the UNSGSA.

CFA Institute. (2009). *Market microstructure: The impact of fragmentation under the markets in financial instruments directive*. https://www.cfainstitute.org/-/media/documents/article/position-paper/market-microstructure-impact-of-fragmentation-under-the-markets-in-financial-instruments-directive.ashx

Consultative Group to Assist the Poor. (2020). *Open banking: How to design for financial inclusion.* https://www.cgap.org/sites/default/files/publications/2020_10_Working_Paper_Open_Banking.pdf

Cœuré, B. (2020, August 19). *Leveraging technology to support supervision: Challenges and collaborative solutions* (Innovation Hub speech). Bank for International Settlements. https://www.bis.org/speeches/sp200819.htm

Competition Markets Authority. (2020). *Regulation and competition: A review of the evidence.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/857024/Regulation_and_Competition_report_-_web_version.pdf

CSL1. (2022). *Innovation Leadership Programme*. Cambridge SupTech Lab. Retrieved October 25, 2022, https://ccaf.io/suptechlab/about_us

CSL2. (2022). *About us*. Cambridge SupTech Lab. Retrieved October 25, 2022, https://ccaf.io/suptechlab/about_us

Creswell, J., & Poth, C. (2017). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Sage Publications.

Currie, W., & Swanson, E. (2009). Special Issue on institutional theory in information systems research: Contextualizing the IT artefact. *Journal of Information Technology*, *24*(4), 283–285.

Dada, D. (2006). The failure of e-government in developing countries: A literature review. *The Electronic Journal of Information Systems in Developing Countries*, *26*(1), 1–10.

Deloitte. (2017). *How to flourish in an uncertain future, Open Banking and PSD2*. https://www2.deloitte.com/content/dam/Deloitte/cz/Documents/financial-services/cz-open-banking-and-psd2.pdf

Di Castri, S., Hohl, S., Kulenkampff, A., & Prenio, J. (2019). *FSI Insights on Policy Implementation No 19: The suptech generations*. Financial Stability Institute of the Bank for International Settlements. https://www.bis.org/fsi/publ/insights19.pdf

DiMaggio, P., & Powell, W. (1983). The iron cage revisited: Institutional Isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147–160.

Digital Regulation Cooperation Forum. (2022). *Digital Regulation Cooperation Forum: Plan of work for 2022 to 2023*. Retrieved July 18, 2022, from https://www.gov.uk/government/publications/digital-regulation-cooperation-forum-workplan-2022-to-2023

Eisenhardt, K., & Martin, J. (2000). Dynamic capabilities: What are they? *Strategic Management Journal*, *21*(10–11), 1105–1121.

European Securities Markets Authority. (2016). *Market abuse*. Retrieved April 25, 2021, from https://www.esma.europa.eu/policy-activities/market-abuse#:~:text=The%20Market%20Abuse%20framework%20is,inside%20information%20and%20market%20manipulation

Financial Conduct Authority. (2016). *Enhancing market integrity*. Retrieved April 24, 2021, from https://www.fca.org.uk/about/what-we-do/enhancing-market-integrity#:~:text=We%20aim%20to%20make%20sure,and%20return%20of%20client%20assets

Financial Conduct Authority. (2021). *FCA business plan 2021/2022*. https://www.fca.org.uk/publications/business-plans/2021-22

Financial Conduct Authority. (2020). *Market abuse*. Retrieved April 24, 2021, from https://www.fca.org.uk/markets/market-abuse

FinCEN. (2022). *FinCEN Innovation Initiative.* Financial Crimes Enforcement Network. Retrieved July 31, 2022, from https://www.fincen.gov/resources/fincens-innovation-initiative

FinCoNet. (2020). *SupTech tools for market conduct supervisors.* International Financial Consumer Protection Organisation. http://www.finconet.org/FinCoNet-Report-SupTech-Tools_Final.pdf

FINOS. (2022). *Open RegTech Strategic Initiative*. Retrieved July 31, 2022, from https://www.finos.org/open-regtech

Fountain, J. (2001). *Building the virtual state*. Brookings Institution Press.

Frantz, P., & Instefjord, N. (2018). Regulatory competition and rules/principles-based regulation. *Journal of Business Finance & Accounting, 45*(7–8), 818–838.

Friend, A. (2021). *The Financial Conduct Authority's innovation journey: Moving forward in the face of uncertainty*. Alliance for Innovative Regulation. https://regulationinnovation.org/wp-content/uploads/2021/03/The-FCAs-Innovation-Journey-Moving-Forward-in-the-Face-of-Uncertainty-AIR-Final-2.pdf

Financial Stability Board. (2017). *Financial stability implications from FinTech*. https://www.fsb.org/2017/06/financial-stability-implications-from-fintech/

Financial Stability Board. (2020). *The use of supervisory and regulatory technology by authorities and regulated institutions: Market developments and financial stability implications*. https://www.fsb.org/2020/10/fsb-report-highlights-increased-use-of-regtech-and-suptech/

Financial Stability Board. (2022*). Assessment of risks to financial stability from Crypto-assets*. https://g20.org/wp-content/uploads/2022/02/FSB-Report-on Assessment-of-Risks-to-Financial-Stability-from-Crypto-assets_.pdf

GFIN1. (2021). *GFIN RegTech & SupTech workstream*. Global Financial Innovation Network. Retrieved July 31, 2021, from https://www.thegfin.com/regtech-suptech

GFIN2. (2021). *GFIN RegTech & SupTech Update*. Global Financial Innovation Network. https://static1.squarespace.com/static/5db7cdf53d173c0e010e8f68/t/61e82ddce7f05416ec493909/1642606081591/GFIN+RegTech-SupTech+Update+-+Showcase+-+EXTERNAL+-+December+2021.pdf

Gilad S., & Yogev T. (2012). How reputation regulates regulators: Illustrations from the regulation of retail finance. In M. L. Barnett & T. G. Pollock (Eds.) *The Oxford handbook of corporate reputation* (pp. 321–340). Oxford University Press.

Gilman, M. (2021, July 13). *Where is suptech heading?* TechCrunch+. https://techcrunch.com/2021/07/13/where-is-suptech-heading/

Gioia, D., Corley, K., & Hamilton, A. (2013). Seeking qualitative rigor in inductive research. *Organizational Research Methods*, *16*(1), 15–31.

Goldstein, I., Jiang, W., & Karolyi, G. (2019). To FinTech and beyond. *The Review of Financial Studies*, *32*(5), 1647–1661.

Haufler, A., & Maier, U. (2019). Regulatory competition in capital standards: A "race to the top" result. *Journal of Banking and Finance, 106*, 180–194.

Hinings, B., Gegenhuber, T., & Greenwood, R. (2018). Digital innovation and transformation: An institutional perspective. *Information and Organization*, *28*(1), 52–61.

Holden, A. (2020). *The future of compliance: Digital transformation meets regulation*. Finextra. Retrieved March 23, 2021, https://www.finextra.com/blogposting/18924/the-future-of-compliance-digital-transformation-meets-regulation.

International Association of Insurance Supervisors and Access to Insurance Initiative. (2019). *RegTech and SupTech: Implications for supervision, report of the A2ii – IAIS consultation call*. https://www.a2ii.org/sites/default/files/2019-07/regtech_and_suptech_implications_for_supervisors_consultation_call_report.pdf

Kattel, R., & Mazzucato, M. (2018). Mission-oriented innovation policy and dynamic capabilities in the public sector. *Industrial and Corporate Change*, *27*(5), 787–801.

Kunz, K., & Martin, J. (2015). Into the breech: The increasing gap between algorithmic trading and securities regulation. *Journal of Financial Services Research*, *47*, 135–152.

Li, T., & Chan, Y. (2019). Dynamic information technology capability: Concept definition and framework development. *The Journal of Strategic Information Systems*, *28*(4), Article 101575.

Luna-Reyes, L., & Gil-Garcia, J. (2014). Digital government transformation and internet portals: The co-evolution of technology, organizations, and institutions. *Government Information Quarterly*, *31*(4), 545–555.

Mikalef, P., & Pateli, A. (2017). Information technology-enabled dynamic capabilities and their indirect effect on competitive performance: Findings from PLS-SEM and fsQCA. *Journal of Business Research*, 70, 1–16.

Moshirian, F. (2011). The global financial crisis and the evolution of markets, institutions and regulation. *Journal of Banking & Finance*, *35*(3), 502–511.

National Audit Office. (2021). *Principles of effective regulation*. National Audit Office. https://www.nao.org.uk/wp-content/uploads/2021/05/Principles-of-effective-regulation-SOff-interactive-accessible.pdf

NASDAQ. (2013, May 14). *NASDAQ OMX successfully delivers SMARTS to Financial Conduct Authority (FCA) to enhance market surveillance capabilities*, *NASDAQ* (Press release). https://www.nasdaq.com/about/press-center/nasdaq-omx-successfully-delivers-smarts-financial-conduct-authority-fca-enhance

Nelson, R., & Winter, S. (2002). Evolutionary theorizing in economics. *Journal of Economic Perspectives*, *16*(2), 23–46.

Norton Rose Fulbright. (2021). *Impact of Brexit on financial services*. Retrieved May 15, 2021, from https://www.nortonrosefulbright.com/en/knowledge/publications/e47d24f4/impact-of-brexit-on-financial-institutions

Organisation for Economic Co-operation and Development. (2019). *Effective approaches for financial consumer protection in the digital age: FCP Principles 1, 2, 3, 4, 6 and 9*. Task Force on Financial Consumer Protection, Directorate for Financial and Enterprise Affairs Committee on Financial Markets. https://www.oecd.org/finance/financial-education/Effective-Approaches-FCP-Principles_Digital_Environment.pdf

Organisation for Economic Co-operation and Development. (2018). *Financial markets, insurance and private pensions: Digitalisation and finance*. https://www.oecd.org/finance/private-pensions/Financial-markets-insurance-pensions-digitalisation-and-finance.pdf

Oliver Wyman. (2017). *Supervising tomorrow*. https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2019/feb/Oliver-Wyman_Supervising-Tomorrow.pdf

Orlikowski, W., & Barley, S. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? *MIS Quarterly*, *25*(2), 145–165.

Ostmann, F., & Dorobantu, C. (2021). *AI in financial services*. Alan Turing Institute. https://www.turing.ac.uk/research/publications/ai-financial-services

Piening, E. (2013). Dynamic capabilities in public organizations. *Public Management Review*, *15*(2), 209–245.

Radaelli, C. (2004). Puzzle of regulatory competition. *Journal of Public Policy, 24*(1), 1–23.

Regulatory Horizons Council. (2022). *Closing the gap: Getting from principles to practices for innovation friendly regulation*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1083582/closing-the-gap-regulation-full-report.pdf

Riaz, S. (2009). The global financial crisis: An institutional theory analysis. *Critical Perspectives on International Business*, *5*(1/2), 26–35.

Robson, C., & McCartan, K. (2016). *Real world research: A resource for users of social research methods in applied settings* (4th ed.). Wiley.

Rusu, J. (2022, February 10). *FCA Approach to Innovation & SupTech* (Speech). Financial Conduct Authority. https://www.fca.org.uk/publication/transcripts/jessica-rusu-regtech-forum.pdf

Salvato, C., & Vassolo, R. (2017). The sources of dynamism in dynamic capabilities. *Strategic Management Journal*, *39*(6), 1728–1752.

Sinha, A. (2012). Implications for growth and financial sector regulation. In *Financial sector regulation for growth, equity and stability* (pp. 45–84), BIS Papers No 62. Bank of International Settlements. https://www.bis.org/publ/bppdf/bispap62.pdf

Teece, D. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), 1319–1350.

Teece, D. J. (2018). Business models and dynamic capabilities*. Long Range Planning, 51*(1), 40–49. https://doi.org/10.1016/j.lrp.2017.06.007

Teece, D., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, *18*(7), 509–533.

Tolbert, P. S., & Zucker, L. G. (1996). The institutionalisation of institutional theory. In S. Clegg, C. Hardy, & W. Nord (Eds.), *Handbook of organisation studies* (pp. 175–190). Sage Publications.

Toronto Centre. (2018). *SupTech: Leveraging technology for better supervision*. https://res.torontocentre.org/guidedocs/SupTech%20-%20Leveraging%20Technology%20for%20Better%20Supervision%20FINAL.pdf

Troshani, I., Janssen, M., Lymer, A., & Parker, L. (2018). Digital transformation of business-to-government reporting: An institutional work perspective. *International Journal of Accounting Information Systems*, *31*, 17–36.

Twizeyimana, J., & Andersson, A. (2019). The public value of e-government - A literature review. *Government Information Quarterly*, *36*(2), 167–178.

UK Gov. (2022). *Financial services future regulatory framework review: Proposals for reform – Response to the consultation*. UK Government. Retrieved August 2, 2022, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092499/FRF_Review_-_Proposals_for_Reform__Government_Response_-_July_2022_.pdf

Weiss, R. (1995). *Learning from strangers: The art and method of qualitative interview studies*. Free Press.

West, D. (2007). *Digital government: Technology and public sector performance*. Princeton University Press.

Wolfson, M. H. (2013). An institutional theory of financial crises. In M. H. Wolfson & G. A. Epstein (Eds.), *The handbook of the political economy of financial crises* (pp.172–190). Oxford University Press.

World Bank. (2018). *From spreadsheets to Suptech technology solutions for market conduct supervision*. World Bank Group. https://openknowledge.worldbank.org/bitstream/handle/10986/29952/127577-REVISED-Suptech-Technology-Solutions-for-Market-Conduct-Supervision.pdf?sequence=4&isAllowed=y

World Bank. (2021). *The next wave of Suptech innovation: Suptech solutions for market conduct supervision*. World Bank Group. https://openknowledge.worldbank.org/handle/10986/35322

World Bank & CCAF. (2020). *The global Covid-19 fintech regulatory rapid assessment report*. World Bank Group and the University of Cambridge. https://www.jbs.cam.ac.uk/wp-content/uploads/2020/10/2020-ccaf-report-fintech-regulatory-rapid-assessment.pdf

World Economic Forum. (2020). *Forging new pathways: The next evolution of innovation in financial services.* https://www.weforum.org/reports/forging-new-pathways-the-next-evolution-of-innovation-in-financial-services

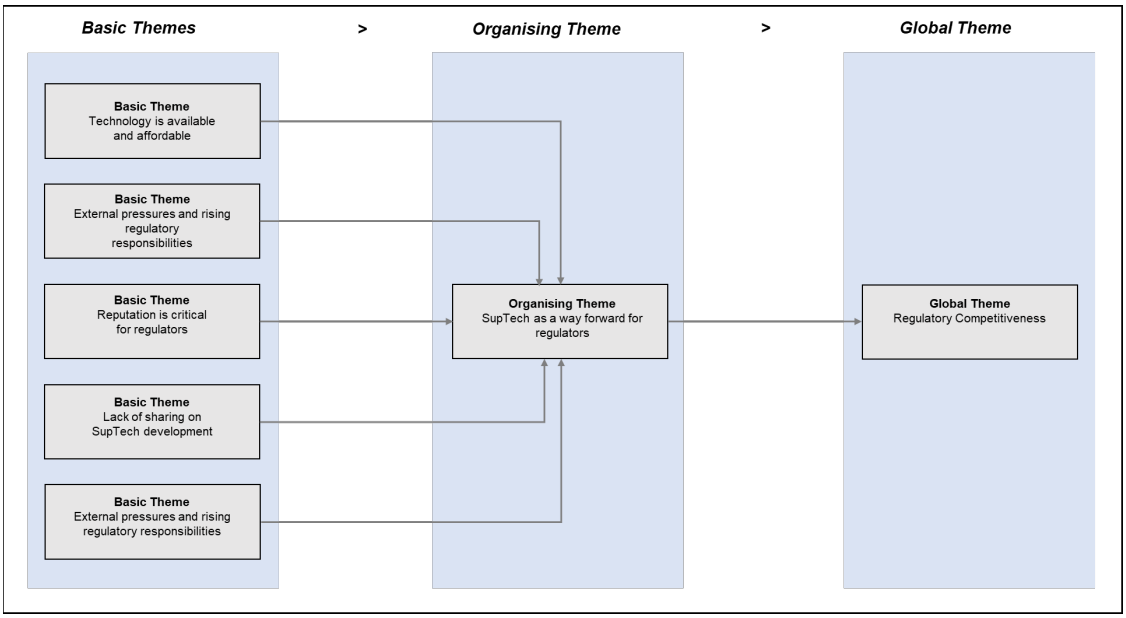
Yeung, K. (2009). Presentational management and the pursuit of regulatory legitimacy: A comparative study of competition and consumer agencies in the United Kingdom and Australia. *Public Administration*, *87*(2), 274–294.

Zeranski, S., & Sancak, I. (2020). Digitalisation of financial supervision with supervisory technology (SupTech). *Journal of International Banking Law & Regulation*, *8*, 309–329.

Zhu, K., Kraemer, K., Dedrick, J., & Xu, S. (2004). Information technology payoff in e-business environments: An international perspective on value creation of e-business in the financial services industry. *Journal of Management Information Systems*, *21*(1), 17–54.

Appendices

Appendix A. Thematic Network Example



Appendix B. Dynamic Capability Detailed Analysis

Overarching Processes	Sensing			Seizing				Transforming				
Description	To identify and shape opportunities, enterprises must constantly scan, search, and explore across technologies and markets, both 'local' and 'distant' (**sensing** opportunities)			Once a new (technological or market) opportunity is sensed, it must be addressed through new products, processes, or services (to be **seized**)				To achieve dynamic capabilities, organizations must be able to continuously renew and seize opportunities, and to periodically **transform** aspects of the organization and culture so as to be able to proactively reposition to address yet newer threats and opportunities as they arise.				
BLENDER Overview	No existent processes recognised for the sensing of regulatory changes. Could be classified under "processes to identify target market segments, changing customer needs and customer innovation			Important to be aware of how the decision making functions in the organization and to receive the necessary buy-in and investment. Buy-in received through governance meetings and an outline of a business case				Complete delivery, creation of a new team to use the new tools and use the data provided by BLENDER and the complementary softw are to visualise the data and provide the insight, achievement of full transformation.				
	Sensing identified with important support by in-house PhD students who provided detailed analysis of the potential impact of the new regulation			Managerial judgment (decision-making skills) take on great significance in such contexts. Visionary drive w as crucial in driving forw ard the BLENDER project				The key is the ability to recombine and to reconfigure assets and organisational structures as the organization grow s. As the team grew and the data sources grew, BLENDER evolved as a tool in order to meet the demand				
	Senior management also identified the potential impact of the new regulation and the potential risks			Management of assets and ensuring that the delivery framew ork/model is present (delivery through the MO team), for the needed flexible delivery				As the FCA moved to the cloud, BLENDER w as also transformed and moved to the cloud from the data centre				
	As it w as a new development, there were not historical examples or comparisons that could have been used			Collaborating with IT department on needed infrastructural aspects, good alignment between departments				Other opportunities are considered also as to whether the BLENDER approach could be used in other teams or BLENDER could feed into other tools				
The below processes are subset of processes that support sensing, seizing and transforming activities (Teece et al., 1997)												
Subset Processes	Process			Position							Path	
	Coordination of different areas	Learning and development	Reconfiguration of assets	Technological Assets	Complementary Assets	Financial Assets	Reputational Assets	Structural Assets	Institutional Assets	Market/Structure Assets	Organizational Boundaries	Path Dependencies	Technological Opportunities
Description	Ability for how effectively internal coordination or integration is achieved	Ability to perform tasks better over time, involves individual skills, new routines in the organization	Ability to sense the need to reconfigure the firm's asset structure and to accomplish the necessary internal and external transformation	Overall technology assets owned by the organization. Mag or mag not be protected by IP	Typically downstream assets that could enhance the uses of products and services and build complementarities	Cash position, degree of leverage and overall financial position (including cash flow)	Intangible assets that enable firms to achieve goals in the market. Could vary between internal and external reputation	Formal and informal structure of an organization, level of hierarchy and governance modes	Broader institutions, regulatory systems, and legal/policy settings	Product/market position and the impact of that	The degree of integration, both from an external perspective and IP protection to internal nature of coordination and ways of working	Impact of previous decisions to the organization's future behaviour and activities	The potential for how far and how fast an organization can move and achieve future opportunities is proportionate to the number of opportunities ahead of it
BLENDER Overview	BLENDER was developed mainly by stakeholders within the Market Oversight team with limited coordination with the IT department	BLENDER was more positioned as the underlying tool and its development was informed with new data sources (venues) as the market evolves	BLENDER was sensed appropriately and the organisation carried out the necessary reconfiguration of assets in order to develop and deliver the solution	FCA owned the technology and tools; the relevant BLENDER code and platform (including the move to the cloud platform that occurred at a later stage)	The tool (SMARTS) where BLENDER data is fed through is critical for the usage of BLENDER data and to support FCA activities	The financial assets that supported the development of BLENDER. Project was developed in-house, therefore, the view was that the project was delivered at a good rate and value for money was achieved	External reputation was not a factor (as it was a confidential project), but there was the internal reputation of the market oversight team and the organizational belief that they can deliver this project independently	For the development of BLENDER, the governance was important and allowed for independence of the Market Oversight team to develop the solution with limited input from the IT division	As the FCA is a regulator, these institutional assets were more influential as a driver of the BLENDER development for the FCA to fulfil its regulatory obligations	Not applicable to the FCA and BLENDER tool as the FCA does not compete or have a market position	Boundaries were limited to the Market Oversight division both for the development and overall usage of BLENDER	More broadly in the FCA, the path dependencies have a strong impact, however, as BLENDER was a new tool, there were not too many influences of previous decisions, solely the choice of using the data centre to host the data	The main opportunity was to move BLENDER to the cloud and improve the efficiencies and effectiveness of the tool
Assessment	Effective coordination within the Market Oversight division for the deployment of the tool	No particular learning and development applicable for the BLENDER tool. Learning and development was present for the usage of the BLENDER data, but that was through a separate tool	Successful reconfiguration of assets for the delivery of BLENDER	Technological assets were crucial in the development and deployment of BLENDER	The complementary assets were key for the successful usage of BLENDER	Financial assets were important and there was sufficient investment to deliver the project	Reputation and belief that the Market Oversight team had the IT capability to deliver this tool was crucial in the delivery of this project	The structure and organization was important and beneficial in how BLENDER was developed and delivered	Important in driving the development of BLENDER and ensuring that the end product is fit for purpose		Did not have an impact on BLENDER and its development (apart from the boundary set between the responsibilities of the Market Oversight Division and IT team)	No major impacts of previous decisions on the development of BLENDER	The cloud opportunity was well received, for future opportunities, it depends on the overall technology stack as BLENDER is a bit of an older tool and depending on future procurement, it might get replaced

Note. FCA = U.K. Financial Conduct Authority; MO = Market Oversight. IP = Intellectual Property. BLENDER = FCA in-house SupTech tool that assist supervisors to identify market abuse.

©2023 Pavle Avramović. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

Digital Transformation of Financial Regulators & Emergence of Supervisory Technologies (SupTech) – A Case Study of the FCA

Pavle Avramovic

Manager, Emerging Tech & Research, UK Financial Conduct Authority

17 October 2023

3rd IFC and Bank of Italy Workshop on "Data Science in Central Banking: Enhancing the access to and sharing of data

Issue 5.2, Spring 2023 Cornucopia **Published on** Apr 27, 2023 **DOI** 10.1162/99608f92.7a329be7 SHOW DETAILS

Digital Transformation of Financial Regulators and the Emergence of Supervisory Technologies (SupTech): A Case Study of the U.K. Financial Conduct Authority

by Pavle Avramović

Published on Apr 27, 2023

last released 1 month ago

FCA Official

Contents

Objectives of the Research

-

Current Landscape

-

SupTech Overview

-

Regulatory Competition

-

BLENDER as a Dynamic Capability

-

Implications for Industry, Academia & Regulators

-

Questions

FCA Official

Learning Objectives

By the end of this presentation, you should be able to:

- **Understand why novel SupTech technologies** are **emerging** in **regulators and central banks**
- **Learn how dynamic capability theory** can be applied to **delivering technology projects with impact**

Current Landscape

The ability of financial regulators to supervise the fast-evolving and technologically enabled world of financial services is **uncertain.**



Scars from the 2007-2008 financial crisis still remain, with future instability potentially occurring



Novel technological deployments in financial services are continuously deployed at an increasing rate





How can financial regulators meet the challenges of this growing ecosystem and prevent harmful outcomes for consumers and markets?

Could SupTech be the answer?

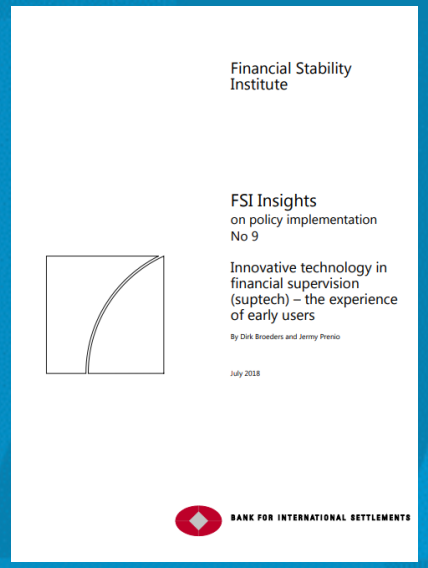
SupTech Explained

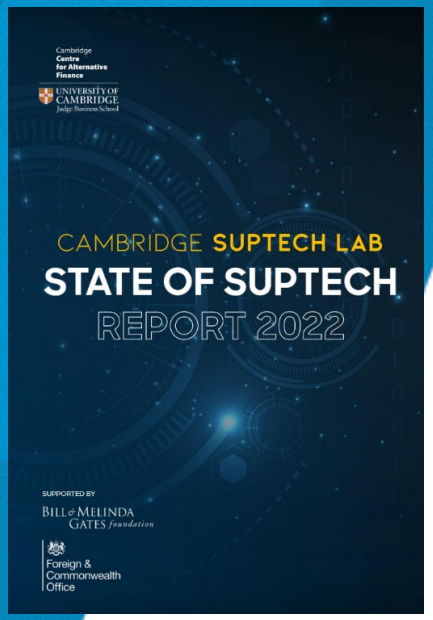
"Use of innovative technologies by supervisory agencies to support supervision" – *Bank for International Settlements*

Catch-all term to capture regulatory initiatives using technology, recognised by many international bodies









FCA Official

SupTech Use Cases

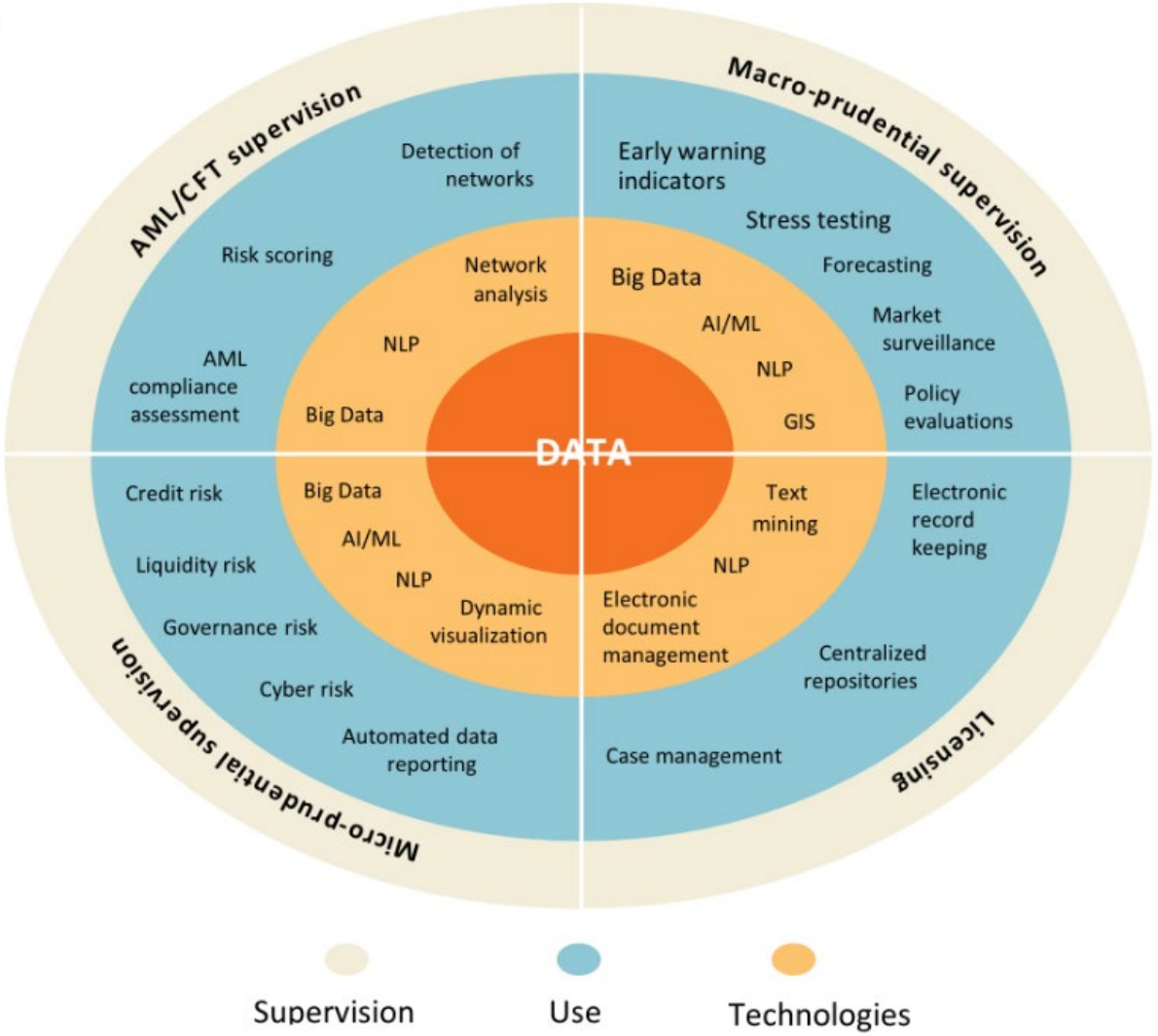
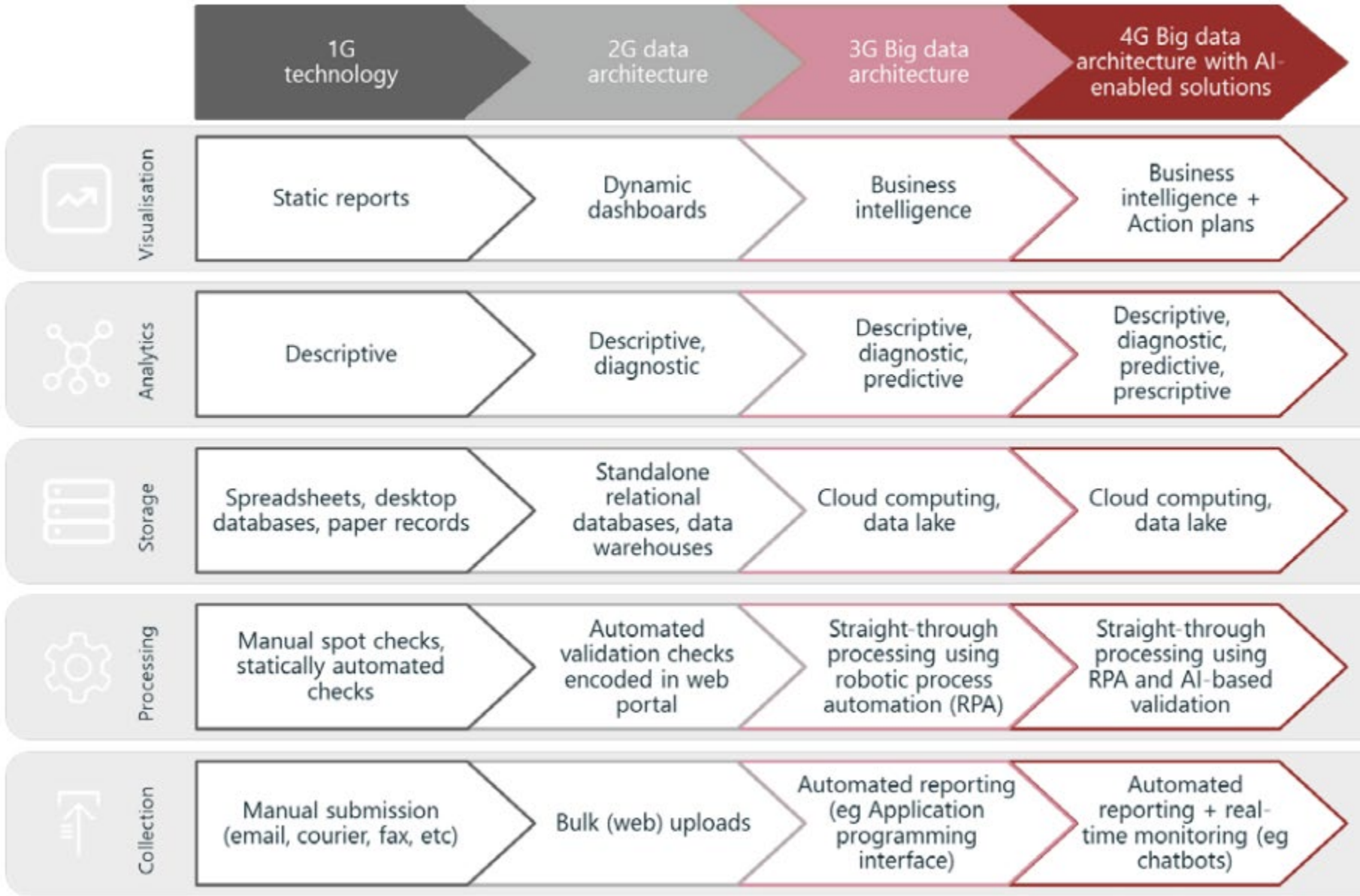


Figure 1. SupTech taxonomy with SupTech use cases. From Appaya et al. (2020).

SupTech Generations



Di Castri et al. (2019)

Research Focus

FCA Official

Why does this matter?

Academic and industry research on regulatory use of technology is **limited**

There is a **lack of case studies** for regulators to learn from on how to **approach the development** of SupTech tools



Research Question

How does SupTech emerge and develop within financial regulatory agencies?

FCA Official

Research Approach

- 16 qualitive, semi-structured interviews with senior colleagues at the FCA and other international regulators (Australia, Canada, Switzerland, USA)

- Thematic analysis approach (Attride-Stirling, 2001); creation of thematic networks, seeing where the data is ‘overflowing’ the literature and identifying novel factors that could impact the environment

- Creation of a new framework to address the research question with a twofold focus

FCA Official

Research Approach

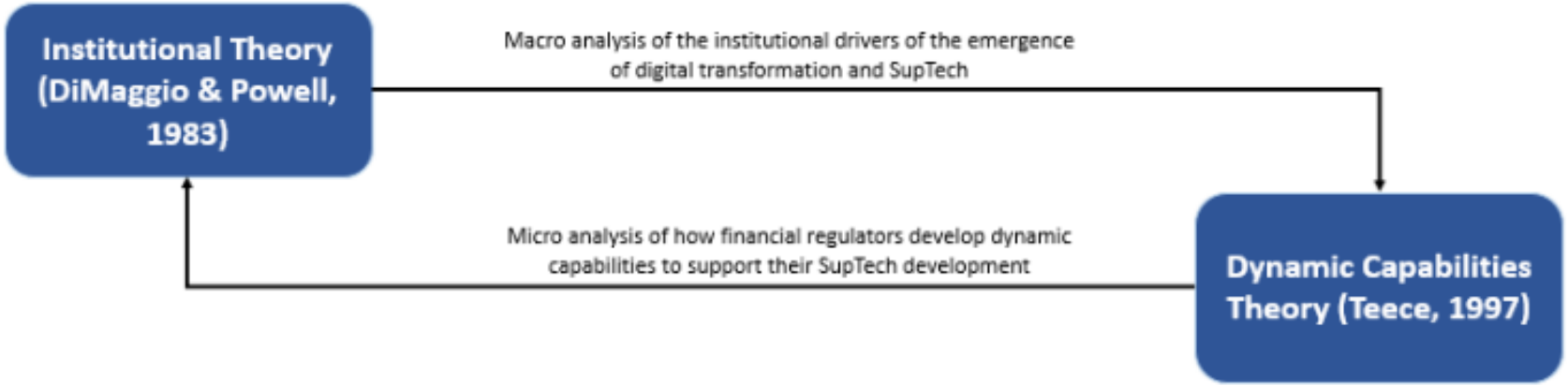


Figure 4. Research theoretical framework.

Key Findings

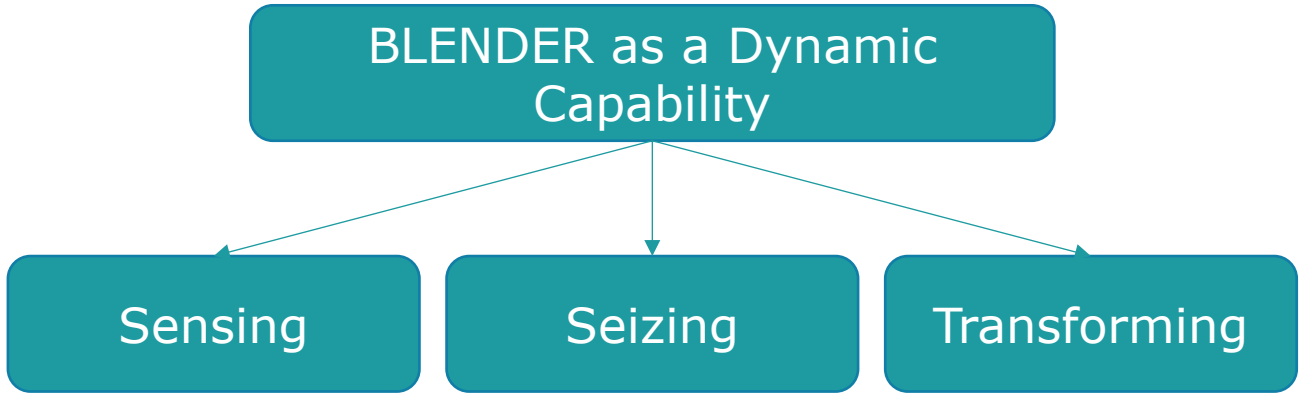
FCA Official

Key Findings

1. Emergence of SupTech

- Drivers of SupTech Growth
- Regulatory Competition – Novel Factor
- Future Impact on SupTech

2. Development of SupTech



Institutional Drivers – SupTech Emergence

DiMaggio and Powell provide an overview of different types of Isomorphic Pressures (those that drive organisations into action):

Types of Pressures	Description	Regulatory SupTech Examples/Equivalents
Coercive Pressures	Result from both formal and informal pressures applied on organizations by other organizations upon which they are dependent and by societal cultural expectations	Regulatory obligations and duty of operating effectively and efficiently (FSB, 2020)
Normative Pressures	Result from the need for professionalization and a group wanting to define the conditions and methods of their work	Not applicable to the SupTech field due to novelty. Potential future certifications and courses could be created, as well as standards
Mimetic Pressures	Result from uncertainty and encourage organizations to imitate each other in order to manage the ambiguity of the environment	Common supervisory needs of automating supervisory procedures and analysing data sets result in 'greater similarity' of SupTech tools being developed (FinCoNet, 2020)

Introducing Regulatory Competition

An area not identified in literature, is the idea of "regulatory competition" in SupTech development.

There are several factors why this type of competition might arise

Market Attractiveness
Well-functioning markets could drive investment and increase international competitiveness

Regulatory Reputation
Important way for regulators to build trust with consumers and industry

Mimetic Pressures
In order to manage the ambiguity of the environment, regulators could imitate each other

Hostile Competition?

I think inevitably that there is competition, this is a quirk of regulators, but it is definitely not hostile competition (Interviewee 12)

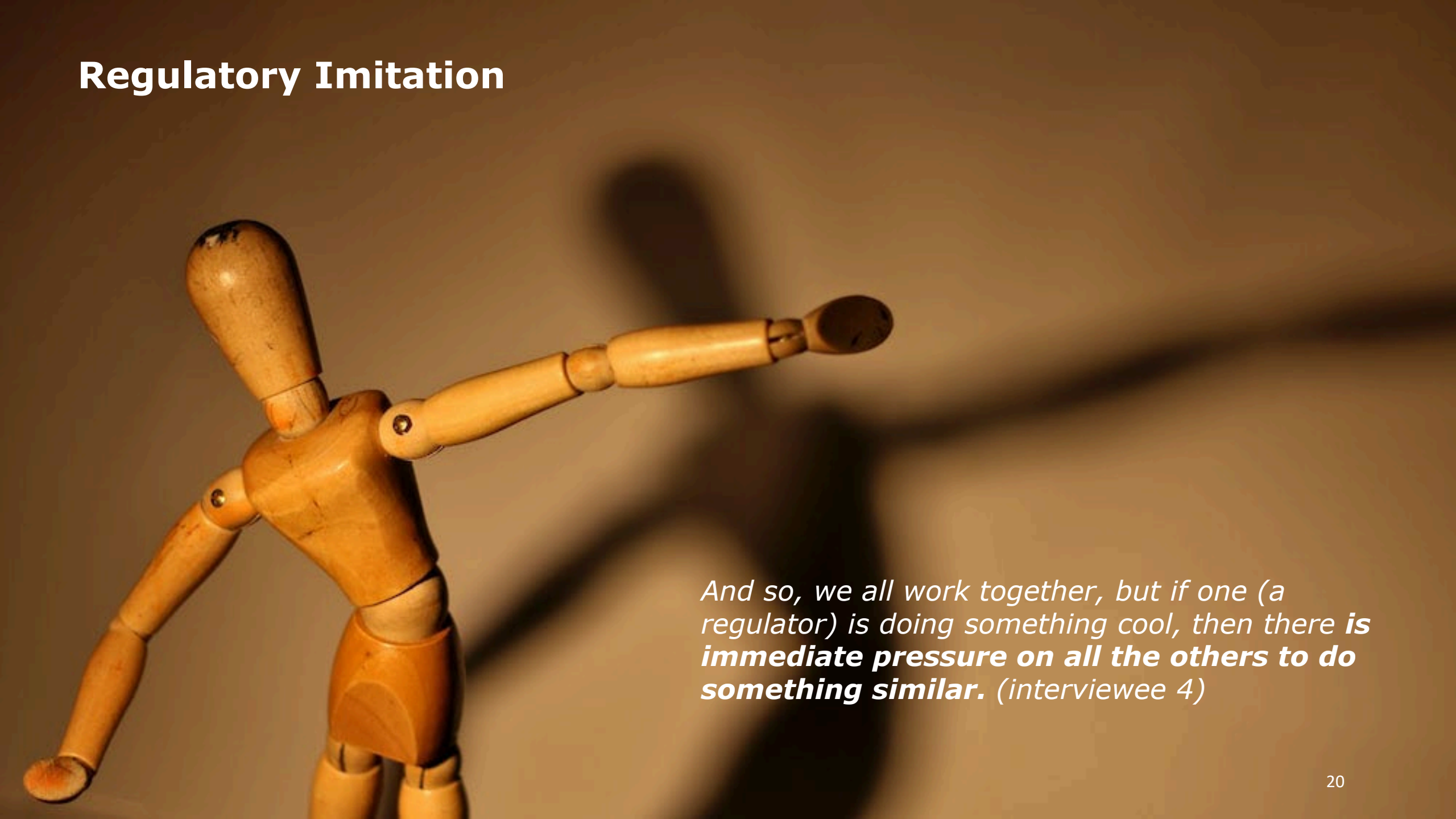


Competitive Advantage



I think some countries are much further ahead than others, some will put more emphasis on it, ***they may see it as a competitive advantage****. (Interviewee 10)*

Regulatory Imitation



And so, we all work together, but if one (a regulator) is doing something cool, then there ***is immediate pressure on all the others to do something similar.*** *(interviewee 4)*

Future Impact on SupTech



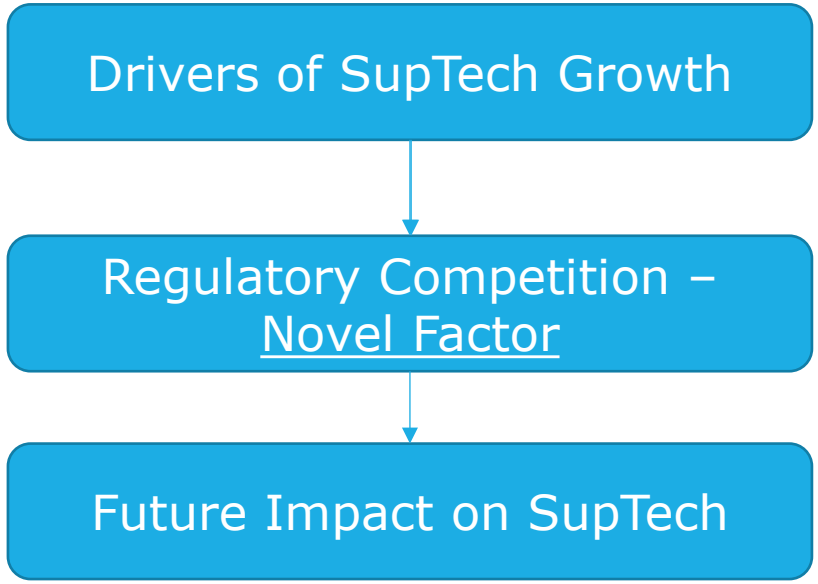
Could the recognition of competition drive more collaboration?

Or will the opaque nature of SupTech result in more competitive behaviour and less joint work?

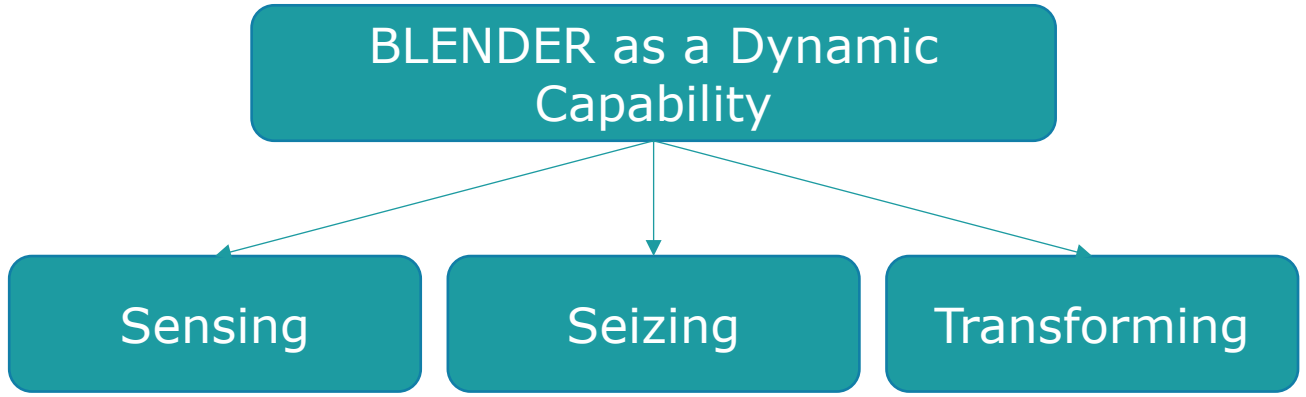
FCA Official

Key Findings

1. Emergence of SupTech



2. Development of SupTech



BLENDER Project

Project Overview	Phase: Complete	FCA Objective: Market integrity

What is it?

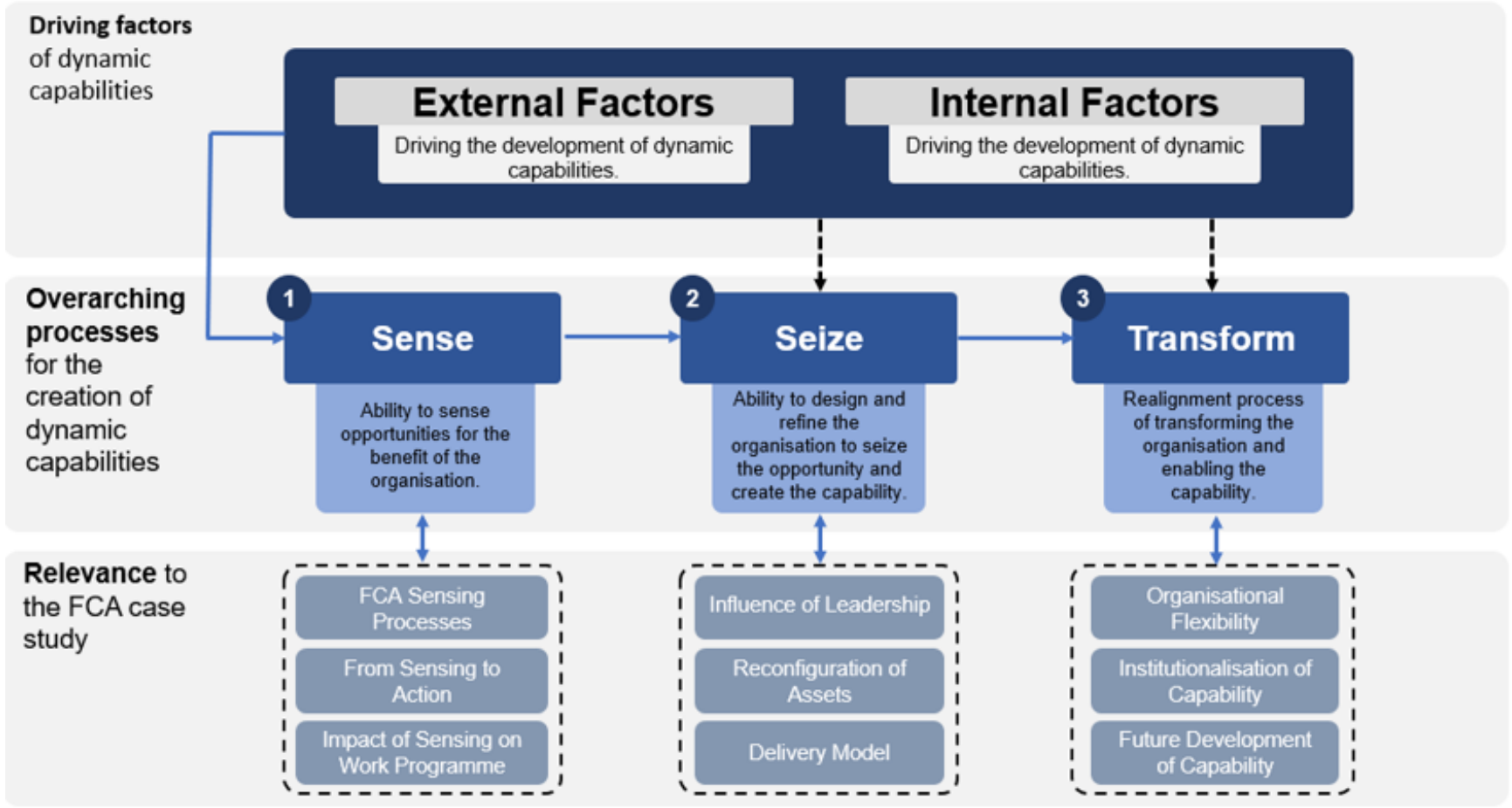
BLENDER is a technology tool that assists supervisors to identify market abuse.

BLENDER acts as 'middleware' and brings together data from different data sources (trading venues), 'blends' the data, and feeds it into a market surveillance tool to support supervisors to see transactions from multiple sources and identify patterns and diverse types of market abuse.

The BLENDER tool is an integral part of the market monitoring process of the FCA, without which it would not be able to have a holistic view of the market and see market manipulation occurring

Dynamic Capability Theory (Teece, 1997)

The organisational ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments.



BLENDER as a Dynamic Capability

Sensing	FCA Sensing practices worked positively, identifying the risk of cross-market manipulation early with the UK having the largest market fragmentation and highest number of venues emerging
Seizing	The right investment case and senior buy-in allowed for the further development of BLENDER. Many factors recognised in literature have occurred, such as the flexibility of delivering in-house, the managerial impact of seizing resources and a driven project leader
Transforming	BLENDER is a completed tool and enabler for the FCA to collect data from multiple venues. It has become part of the ecosystem and has achieved legitimacy in the eyes of its users.

The overarching finding based on the data is that the BLENDER tool can be perceived as a dynamic capability for the FCA .

FCA Official

Implications

1. Recognition of **regulatory competition in SupTech development** could act as a **positive driver for change and stimulate collaboration**.

2. Criticality of **horizon scanning, visionary leadership and flexible delivery models** in the development of SupTech tools (and the usage of the dynamic capabilities framework).

3. **Further academic and industry research** can support the development **of SupTech, including future talent, opensource code and needed standards**

FCA Official

Any questions?

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Building a database on cryptocurrencies[1]

Anamaria Illes and Ilaria Mattei,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



Building a database on cryptocurrencies

Ilaria Mattei and Anamaria Illes, *Bank for International Settlements*

3rd IFC Workshop on "Data Science in Central Banking: Enhancing the access to and sharing of data"

17 October 2023

The views expressed herein are those of the authors and do not necessarily reflect the opinions of the BIS.

Database on cryptocurrencies

- **What is it?**
  - Overview, scope and definitions
- **How did we build it?**
  - Preparatory work
  - Data sources
  - Key challenges and measures taken to tackle these
  - Daily automated updates
  - Residual manual tasks (for stablecoins only)
- **Current applications**
- **What's next**

Overview, scope & definitions

- The cryptocurrencies database is a collection of:
  - **Market variables** (market capitalisation, prices, trading volumes, …) for all cryptocurrencies coming from multiple data sources.
  - **Derived indicators**, mostly on *stablecoins*: market structure, price volatility, price deviations from peg, reserves.

- *Stablecoins* in scope: cryptocurrencies that aim (claim) to maintain a stable value relative to a specified asset, or a pool or basket of assets, and that have a positive market capitalisation.
  - **A. Fiat-backed:** claim to be backed by assets denominated in a fiat currency.
  - **B. Crypto-backed:** claim to be backed by other cryptocurrencies.
  - **C. Commodity-backed:** claim to be backed by commodities.
  - **D. Unbacked:** do not claim to be backed by any reserves and rather seek to maintain a stable value through, for instance, algorithms or protocols.

From the conceptual idea to its materialization

1. Definition of the **dimensions to monitor:**
   - Size and features of the market
   - Quality of the stabilization mechanism (for stablecoins only)
   - Composition of the reserves (for stablecoins only)

2. Identification of the **variables:**
   - Market capitalisation, volumes, prices
   - Attestation reports (for stablecoins only)

3. Selection of the **data sources**: CoinGecko, CCData (CryptoCompare), websites of stablecoins

Data source and variables in details

Sources	Database of market variables	Database of derived indicators
CoinGecko	List of stablecoins	(1) Number of stablecoins
	Market capitalisation	(1) Levels of and changes in market capitalisation (2) Ranking (3) Market concentration (HHI) (4) Turnover
	(Off-chain) volume	(1) Trading volume (2) Turnover
	Closing prices	(1) Price (2) Price volatility (3) Deviations from peg
CCData (CryptoCompare)	OHLC prices	(1) Prices (2) Deviations from peg
Stablecoins' websites	Composition of reserves	(1) Breakdown and timeseries of reserves

Challenges

- Lack of *unique identifiers* across data sources
- Lack of consistent *definitions* of stablecoin types
- Lack of *reporting standards* for the reserves
- Lack of provision of data on "*dead*" cryptocurrencies
- Publicly available data being subject to *copyrights*

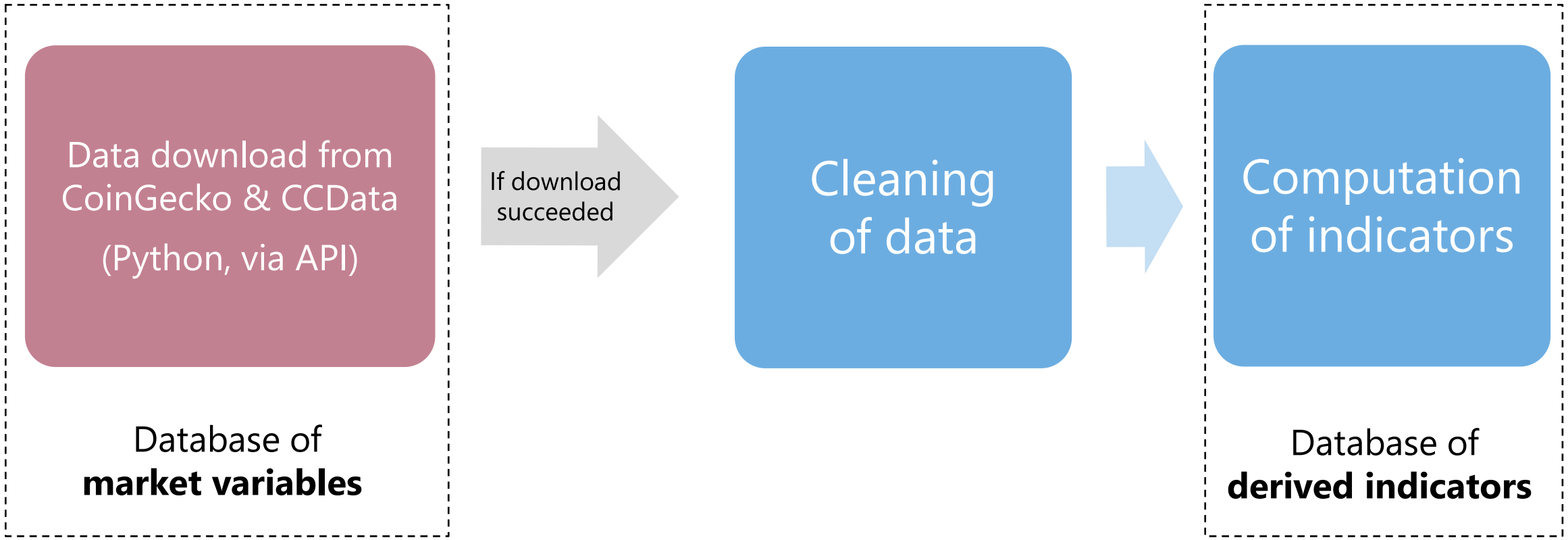
Solutions

- Creation of own **mapping rules**
- Creation of own **metadata**
- Storing **vintages**
- Get in contact for **permissions**

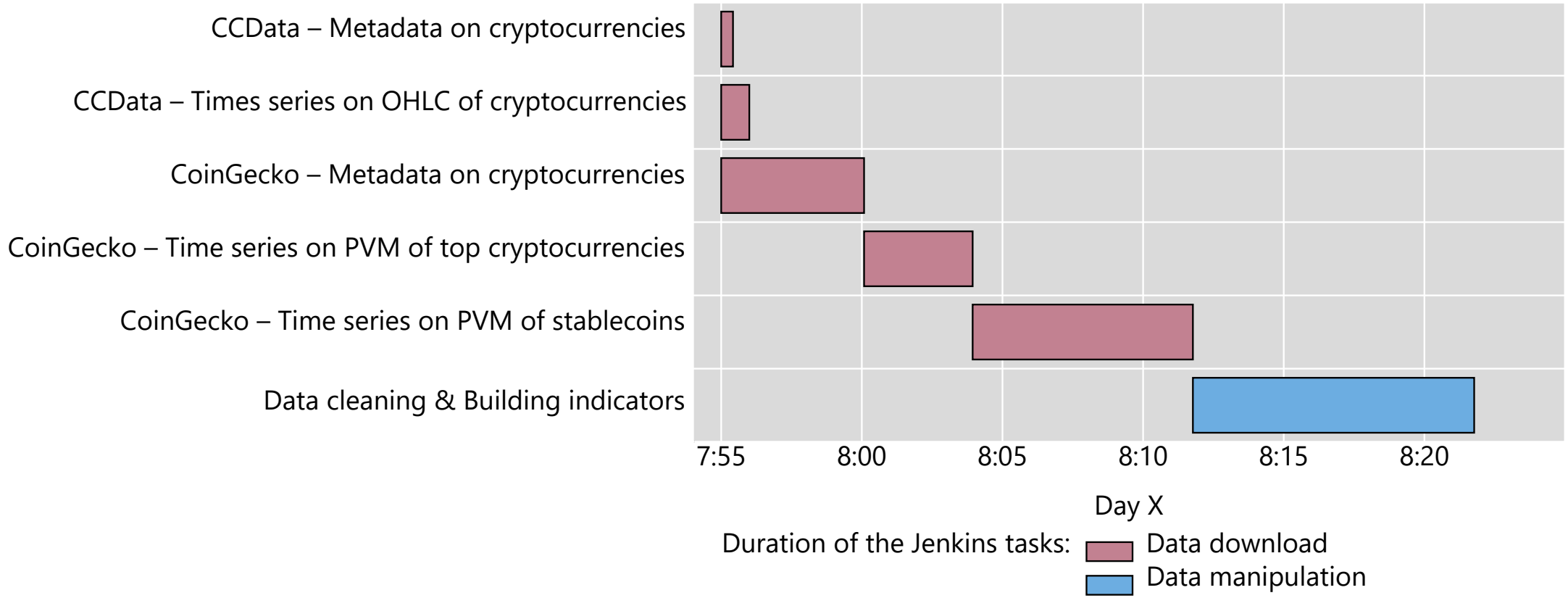
Result

Centralised database that is built in a structured way
to allow for a consistent comparison across data sources and time

The process runs automatically every day and takes less than 30 minutes



In a typical day…



OHLC= open, high, low and close prices; PVM = prices, volumes and market capitalisation.

Residuals manual tasks (for stablecoins only)

- **At the start and on an ad-hoc basis:**
  - Checking of (new) coins against our definition
  - Categorisation of (new) stablecoins by type
  - Mapping of (new) stablecoins among data sources
- **Troubleshooting codes which fail to run, eg due to:**
  - New (stable)coins
  - Temporary access restrictions to public API
  - Change in data structure (rare)
- **Update of breakdown of reserves on a monthly basis:**
  - Manually extract the information from the stablecoins' websites
  - Map the reserve-assets into our own high-level asset types

Current applications

- Daily (internal) newsletter on cryptocurrencies
- Stablecoins Dashboard (internal)
- Forthcoming paper

Ideas for future upgrades

- **Further automation:**
  - String matching of cryptocurrencies' names across data sources
  - Collection of information on reserves
- **Exploration of other sources, eg:**
  - Coinmetrics
  - IntoTheBlock
  - Chainalysis
  - ….
- **Addition of new indicators, eg:**
  - Transactions
  - Inflows to/outflows from exchanges
  - DeFi
  - …

BIS

Questions?

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

European single access point as a blueprint for global financial and green data hubs[1]

Pawel Martyniuk and Michal Piechocki,
Business Reporting - Advisory Group (BR-AG)

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BRAG

European single access point as a blueprint for global financial and green data hubs

Pawel Martyniuk
Michal Piechocki

3rd IFC and Bank of Italy Workshop on "Data Science in Central Banking: Enhancing the access to and sharing of data"
17-19 October 2023

DISCLAIMER

The views presented are those of the author solely. They do not purport to and shall not be understood as views of the European Commission, the Council of Europe, the European Securities and Markets Authority or any other agencies involved in the development of ESAP.

European Single Access Point

The why?



Anne is an investor who is looking for opportunities on pan-European capital market



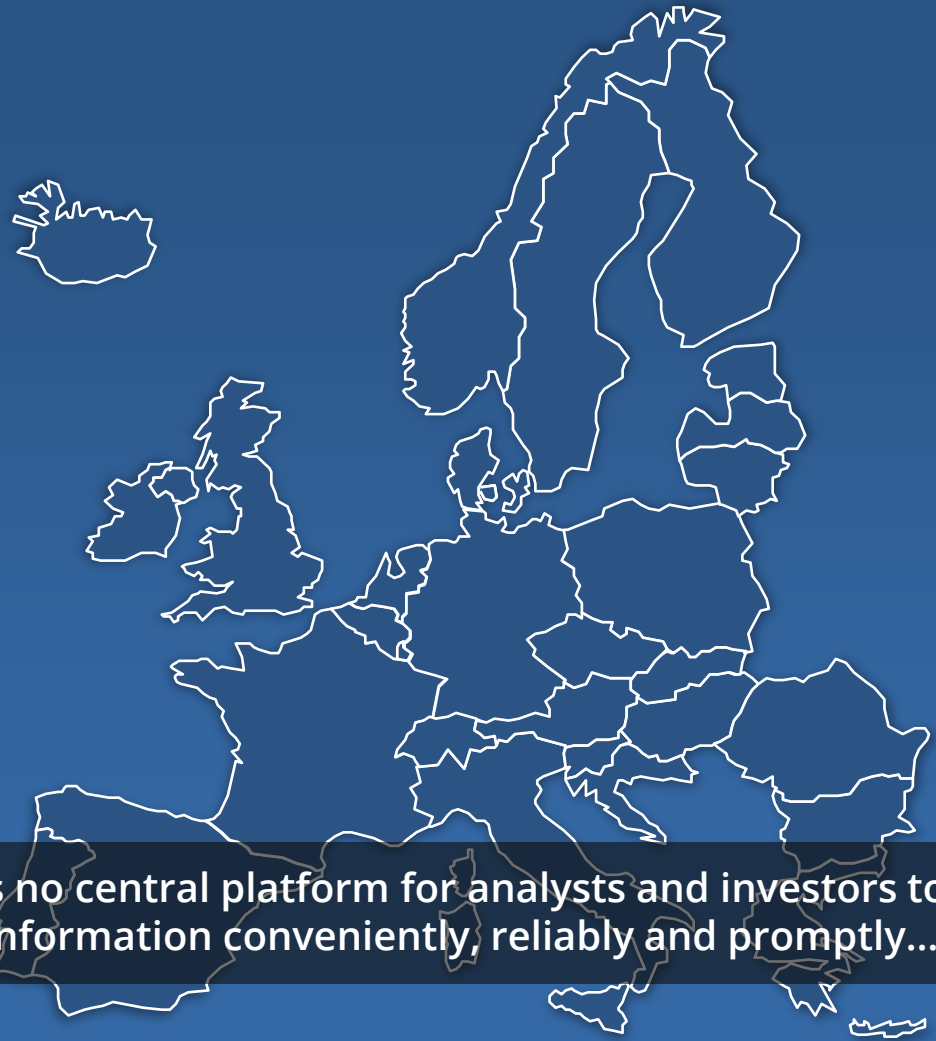


Peter is an M&A manager who is looking for possible synergy opportunities among the European capital market participants



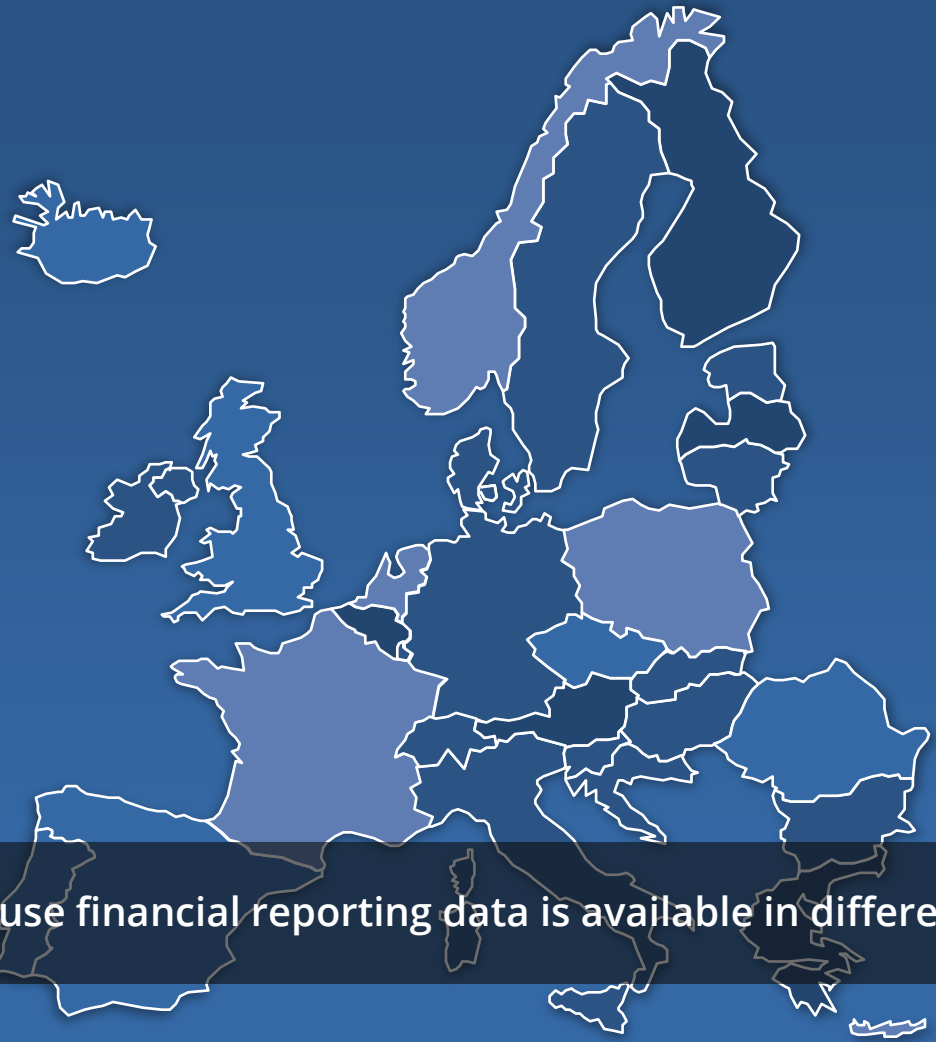


Maria is a university employee who is searching for ESG data for her research



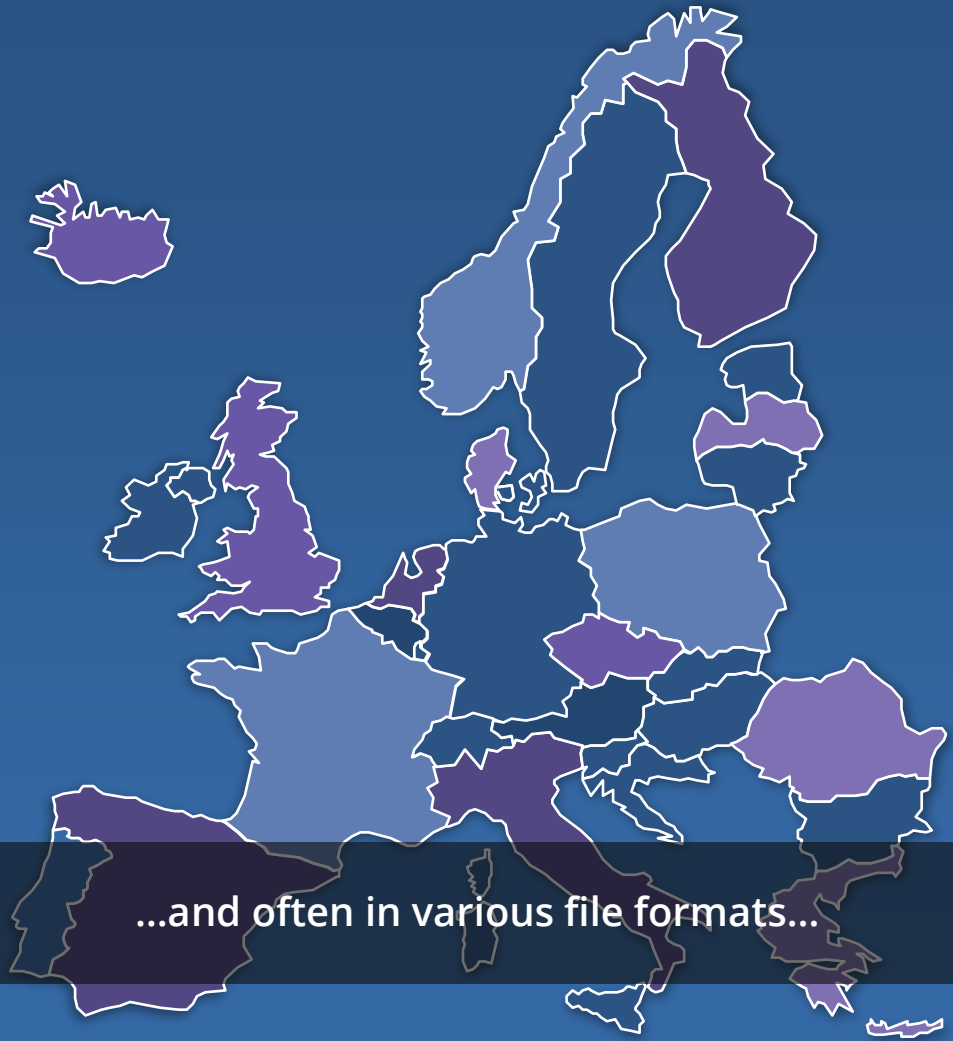


Since there is no central platform for analysts and investors to obtain that information conveniently, reliably and promptly...



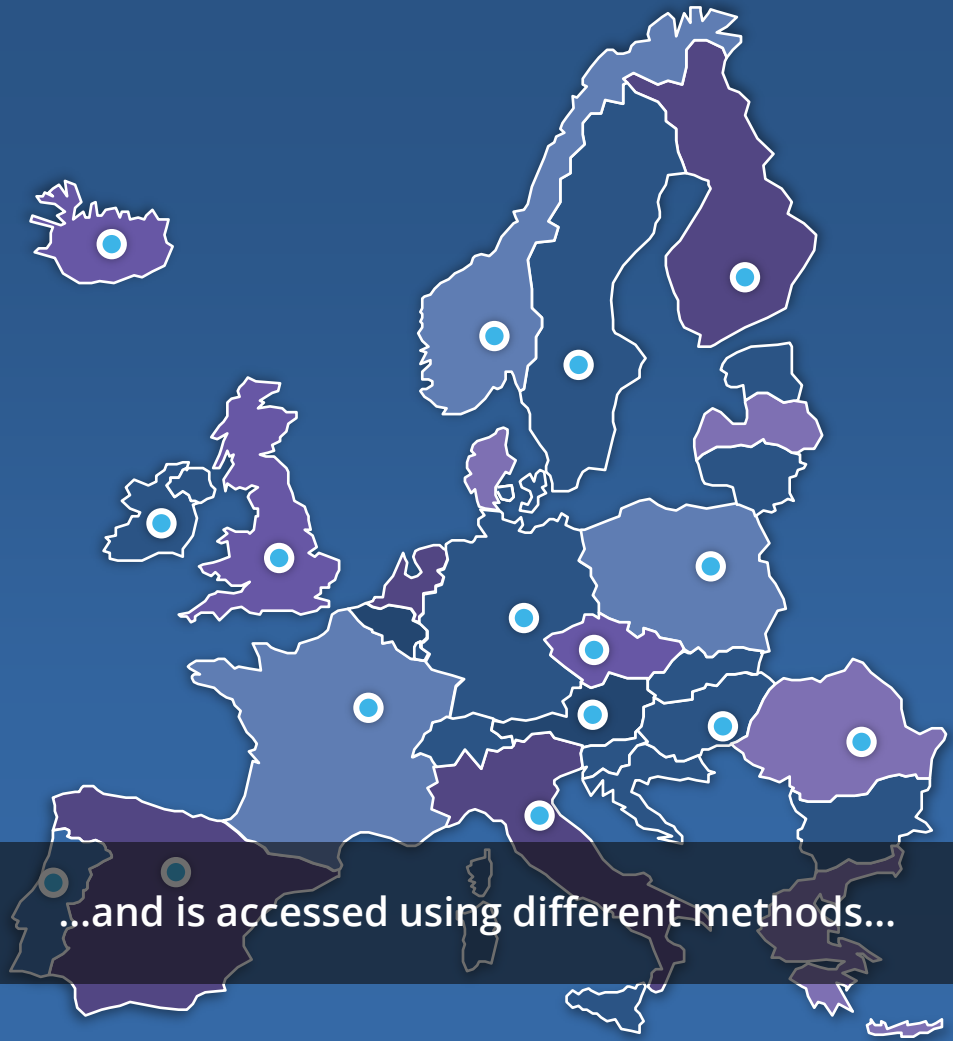


...and also because financial reporting data is available in different languages...



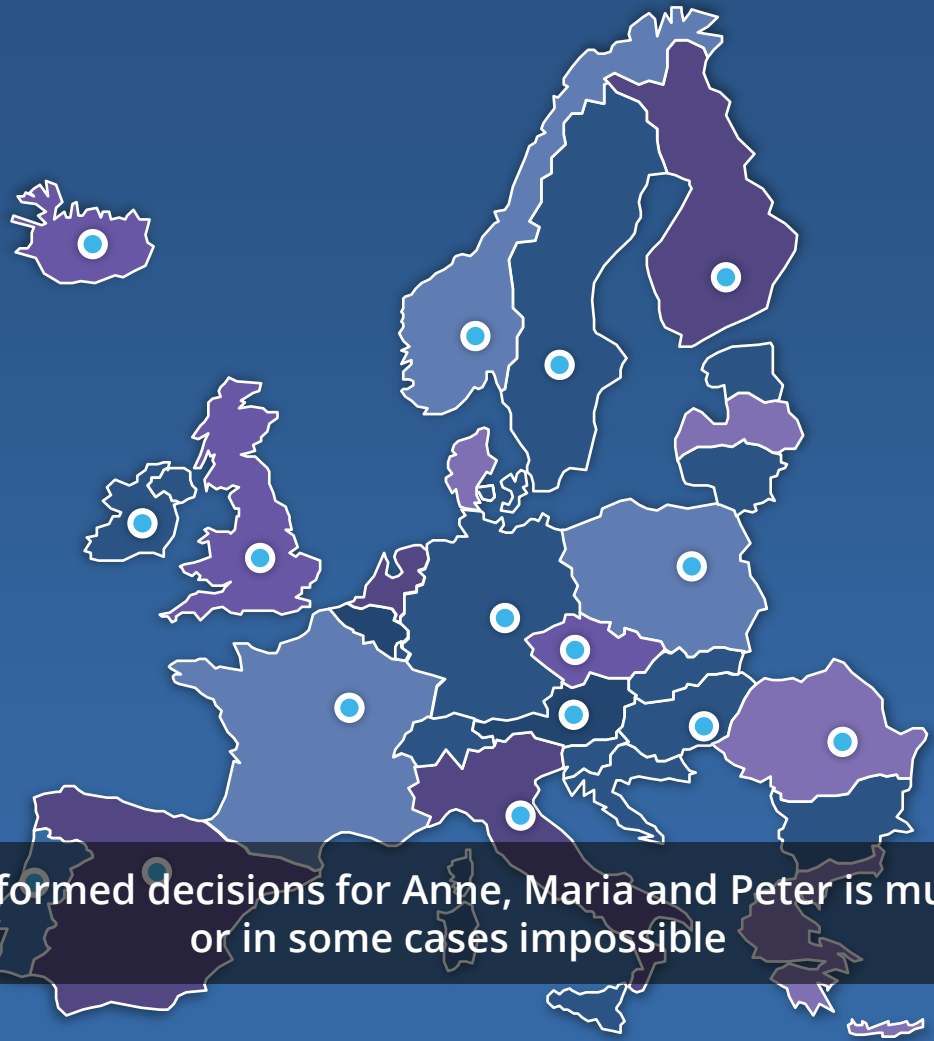


...and often in various file formats...





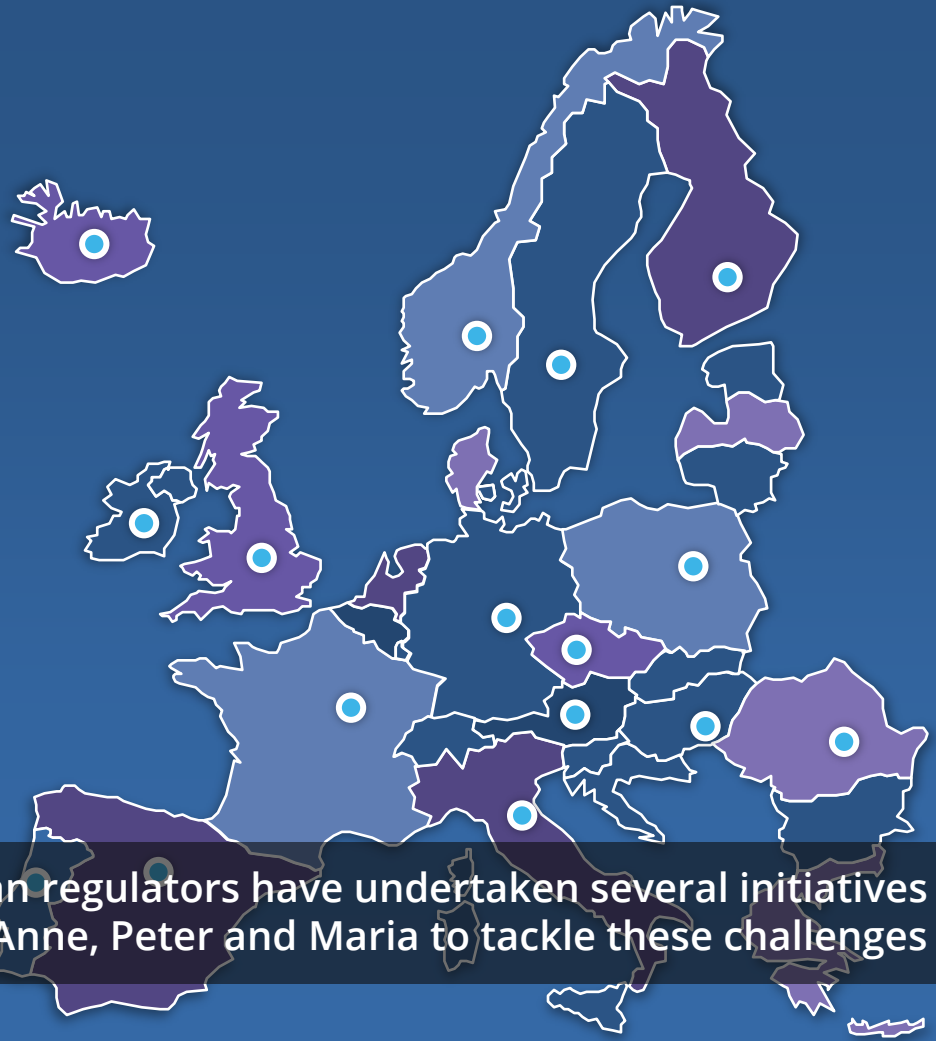
...and is accessed using different methods...





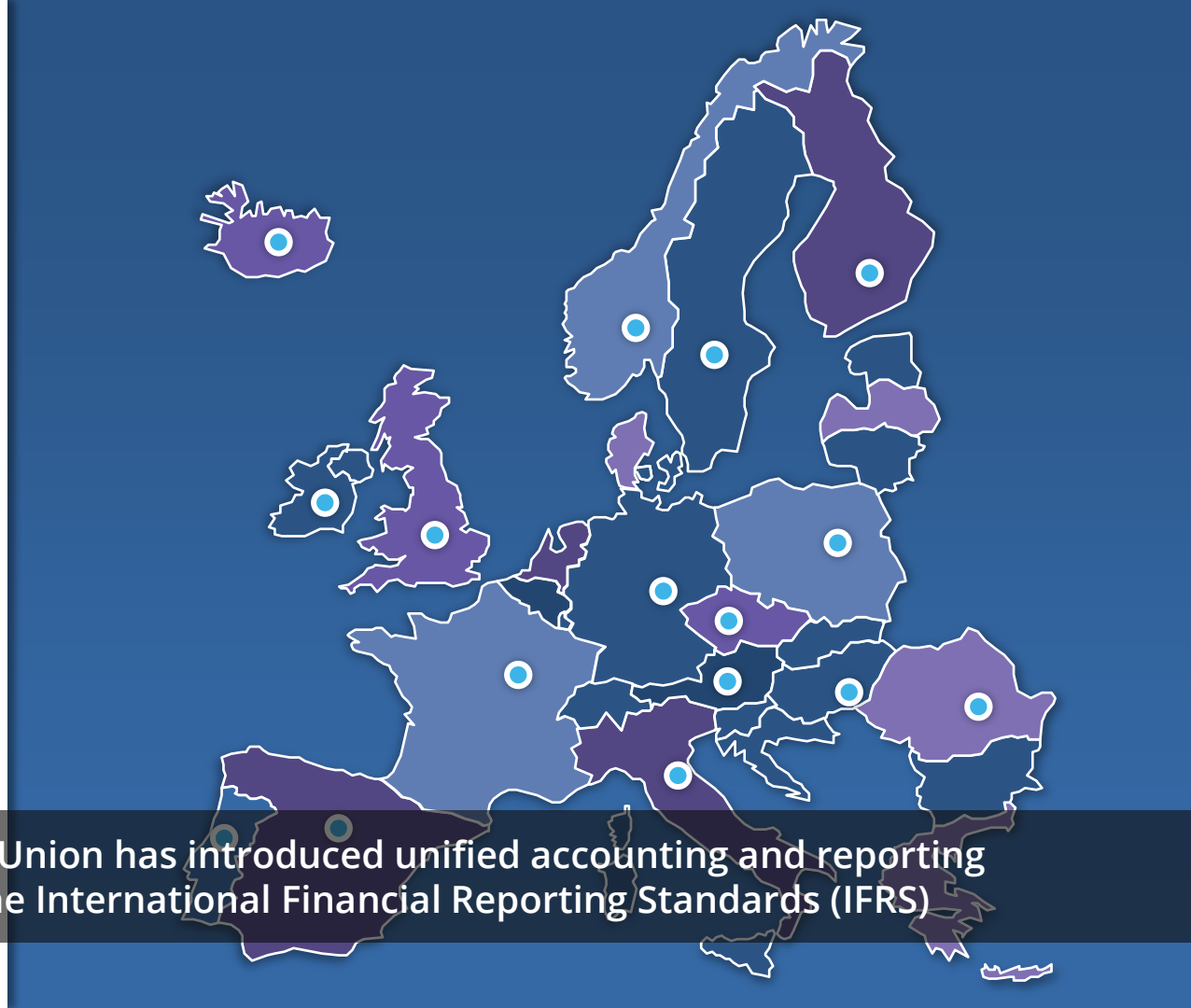


...making informed decisions for Anne, Maria and Peter is much harder,
or in some cases impossible



European regulators have undertaken several initiatives to help Anne, Peter and Maria to tackle these challenges

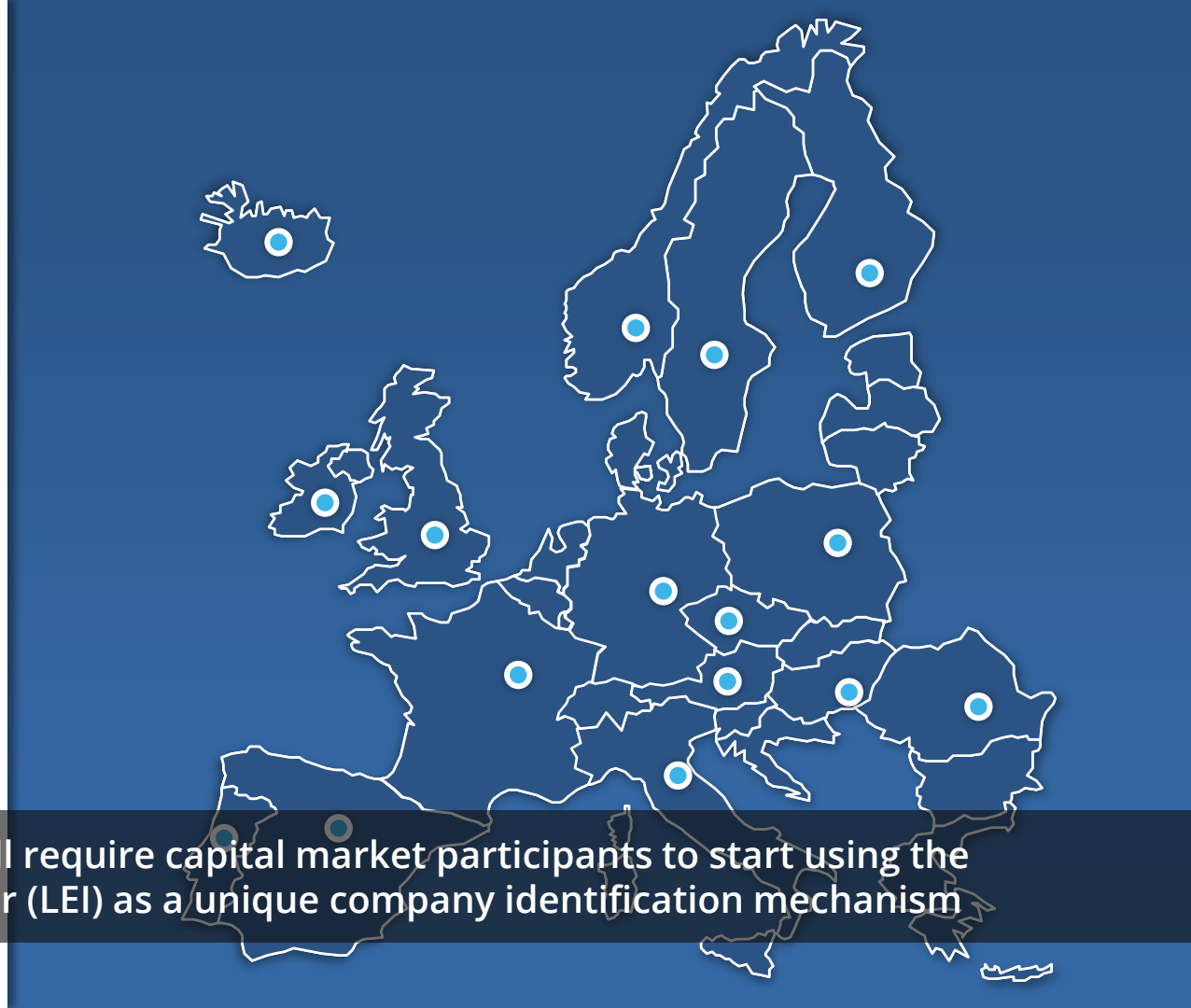
IFRS® 2005



In 2005 the European Union has introduced unified accounting and reporting standards called the International Financial Reporting Standards (IFRS)

IFRS® 2005

LEI GLOBAL LEGAL ENTITY IDENTIFIER FOUNDATION 2017

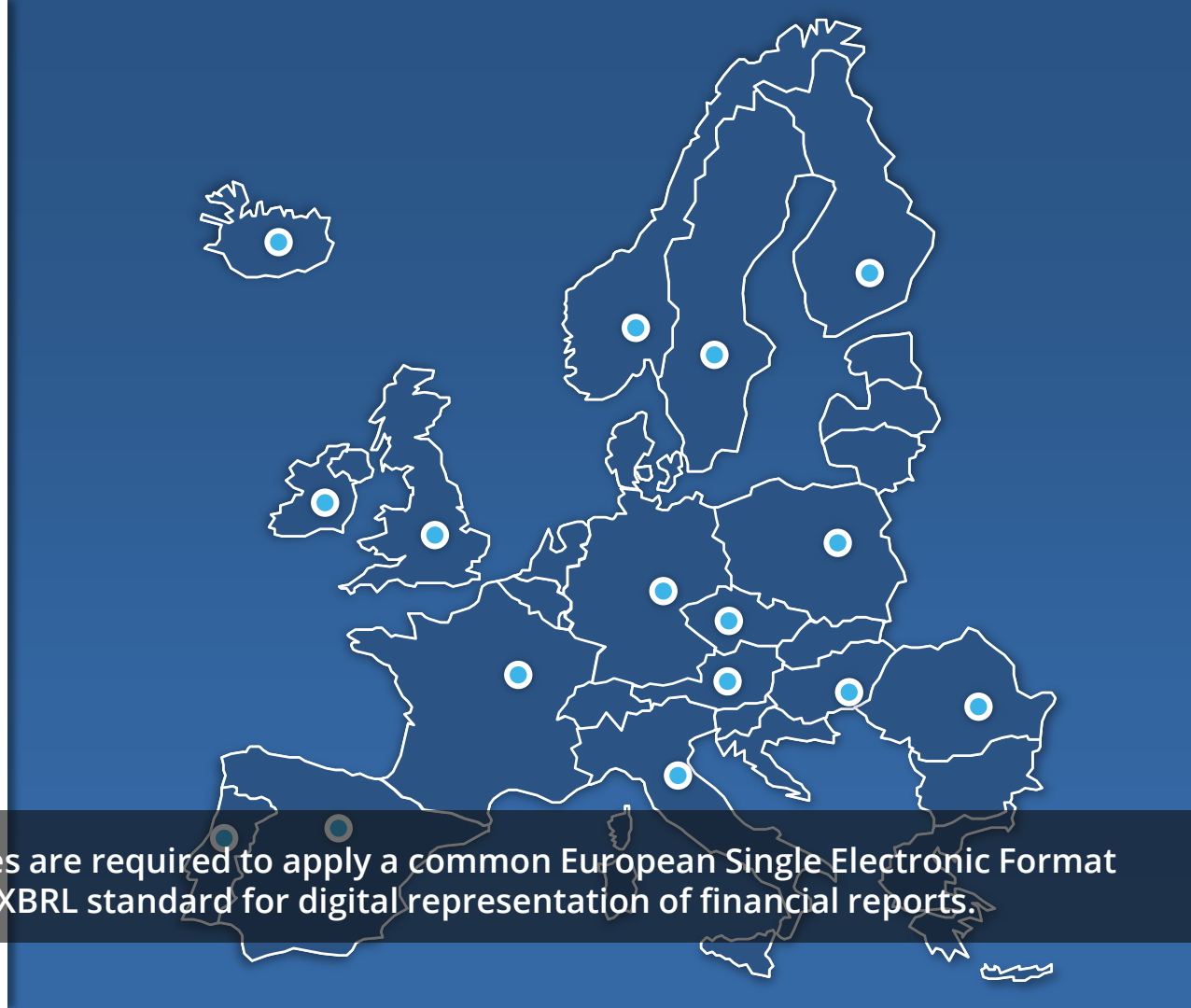


From 2017 the EU will require capital market participants to start using the Legal Entity Identifier (LEI) as a unique company identification mechanism

IFRS® 2005

LEI GLOBAL LEGAL ENTITY IDENTIFIER FOUNDATION 2017

XBRL™ THE BUSINESS REPORTING STANDARD 2020



From 2020 all EU listed companies are required to apply a common European Single Electronic Format which adopts the Inline XBRL standard for digital representation of financial reports.



Yet still, for an investor, the European capital market information is not conveniently accessible

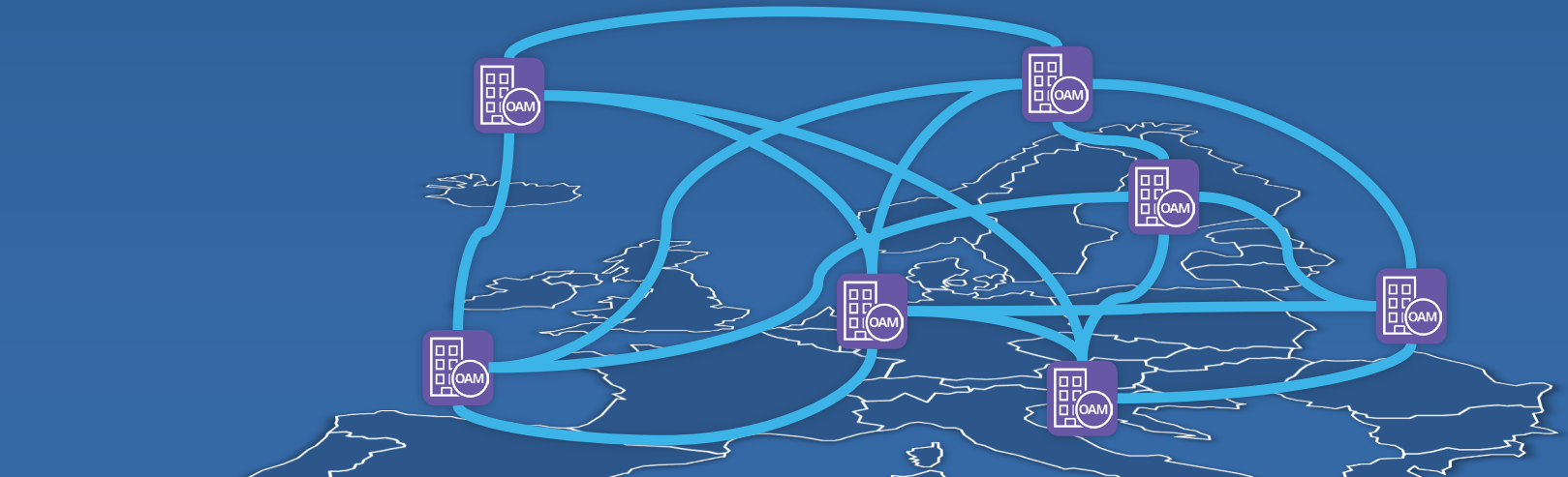
Currently in each country there is an offical registry called the Officialy Appointed Mechanism (OAM) that is responsible for distribution of regulated information received from listed companies



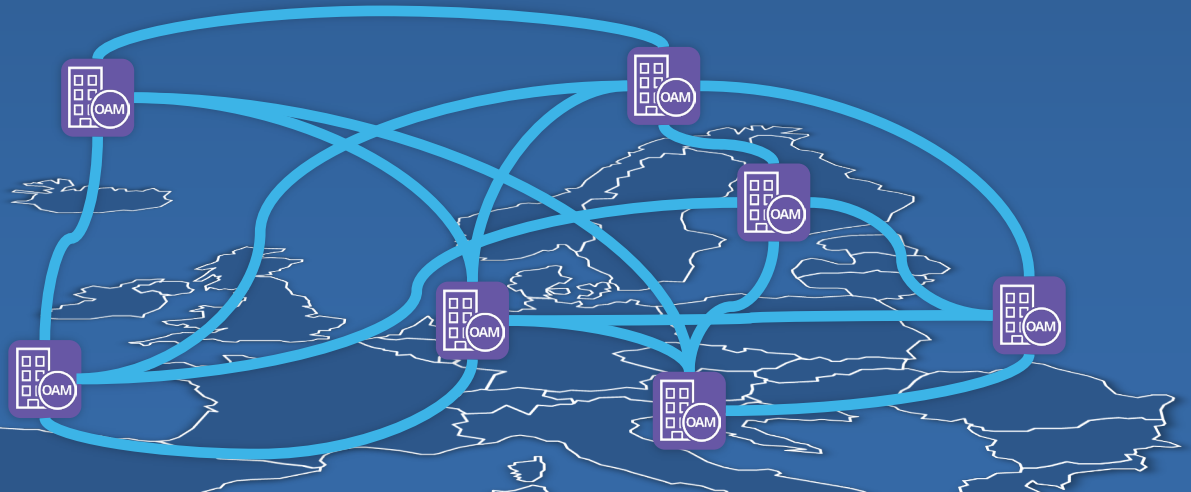
At present these registries are not connected in any usable manner



But what if these registries were connected through a single access point?



What if they became a „cloud-based” commonly accessible registry that synchronises national regulatory information automatically?



This is what we call the European Single Access Point




ESAP
OAM
OAM
OAM
OAM

ESAP

Report Metadata

- Issuer Name
- Unique Issuer ID (LEI)
- Type of regulated information
- Report URL (OAM)
- Report URL (Issuer's IR website)
- Issuer's Home Member State

etc.



ESAP

Report Metadata

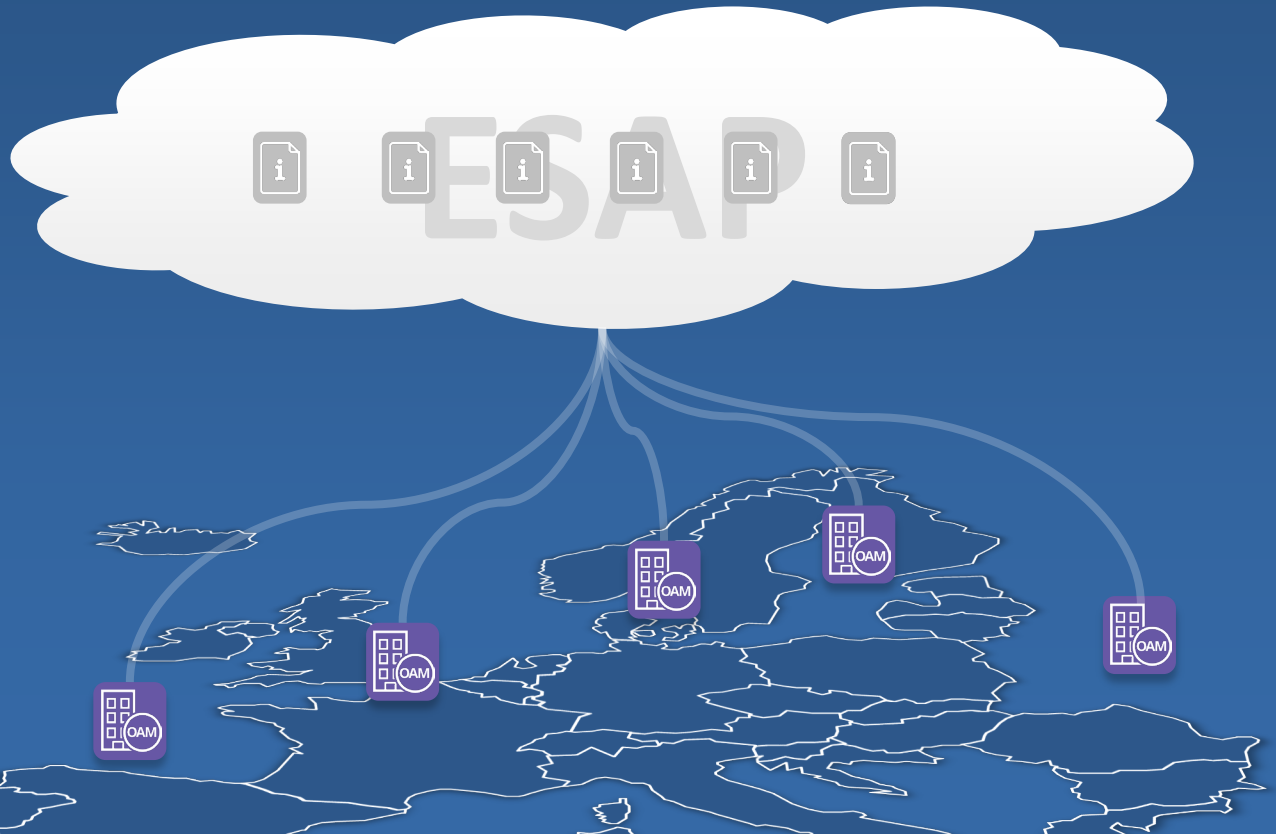
- Issuer Name
- Unique Issuer ID
- Type of regulated information
- Report URL (OAM)
- Report URL (Issuer)
- Issuer's Home Member State

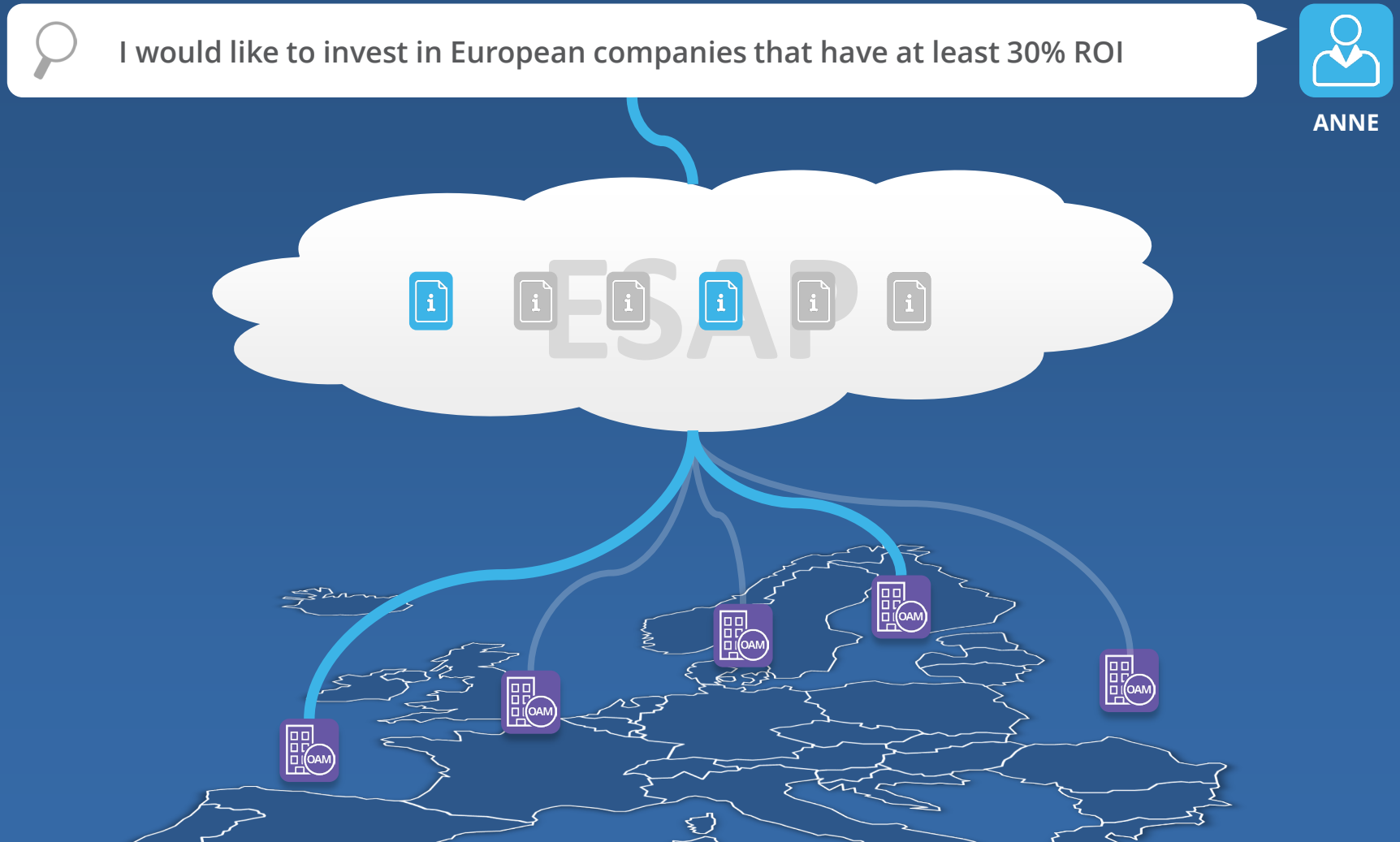
Etc.



I would like to invest in European companies that have at least 30% ROI

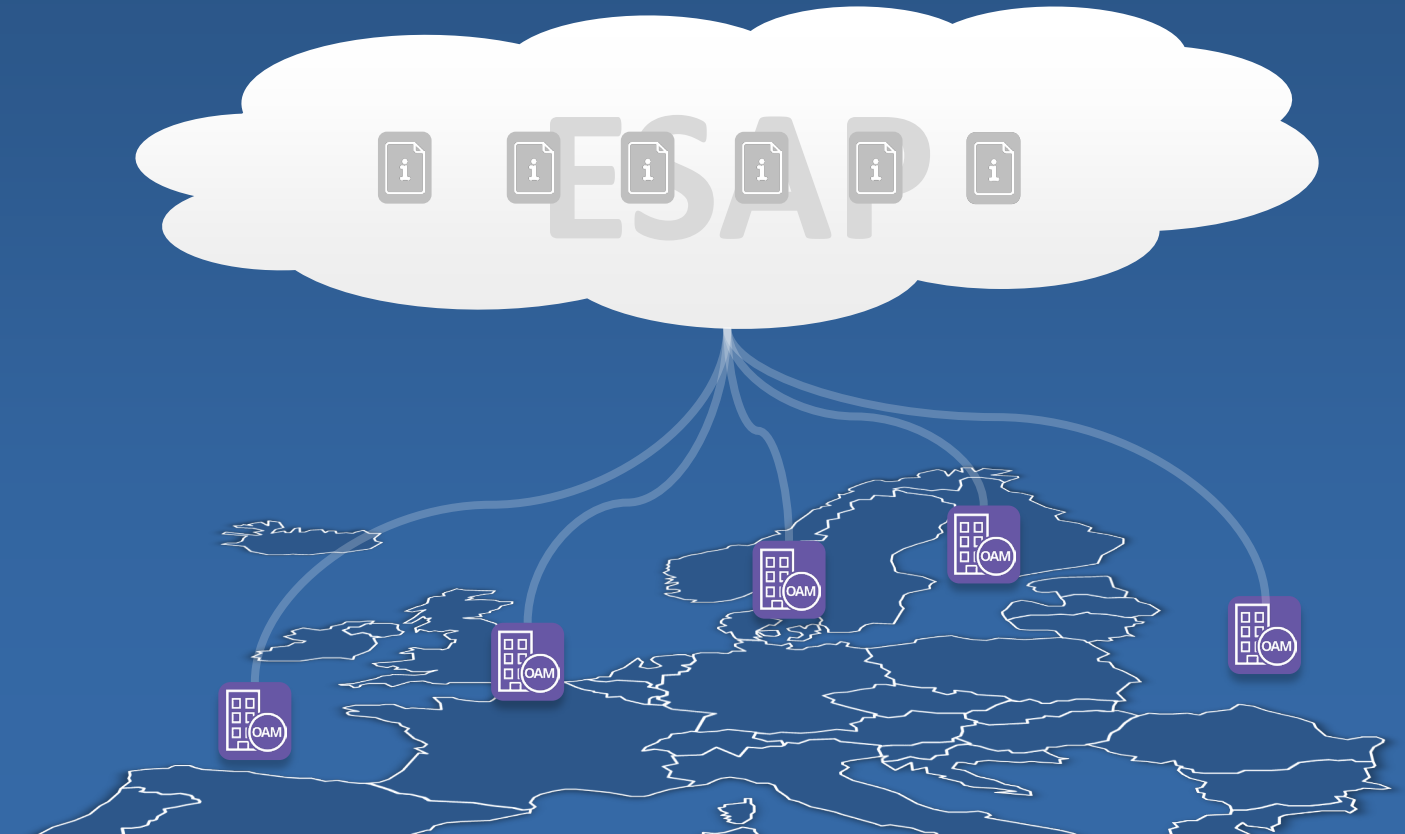
ANNE




I would like to invest in European companies that have at least 30% ROI
ANNE
ESAP
OAM
OAM
OAM
OAM

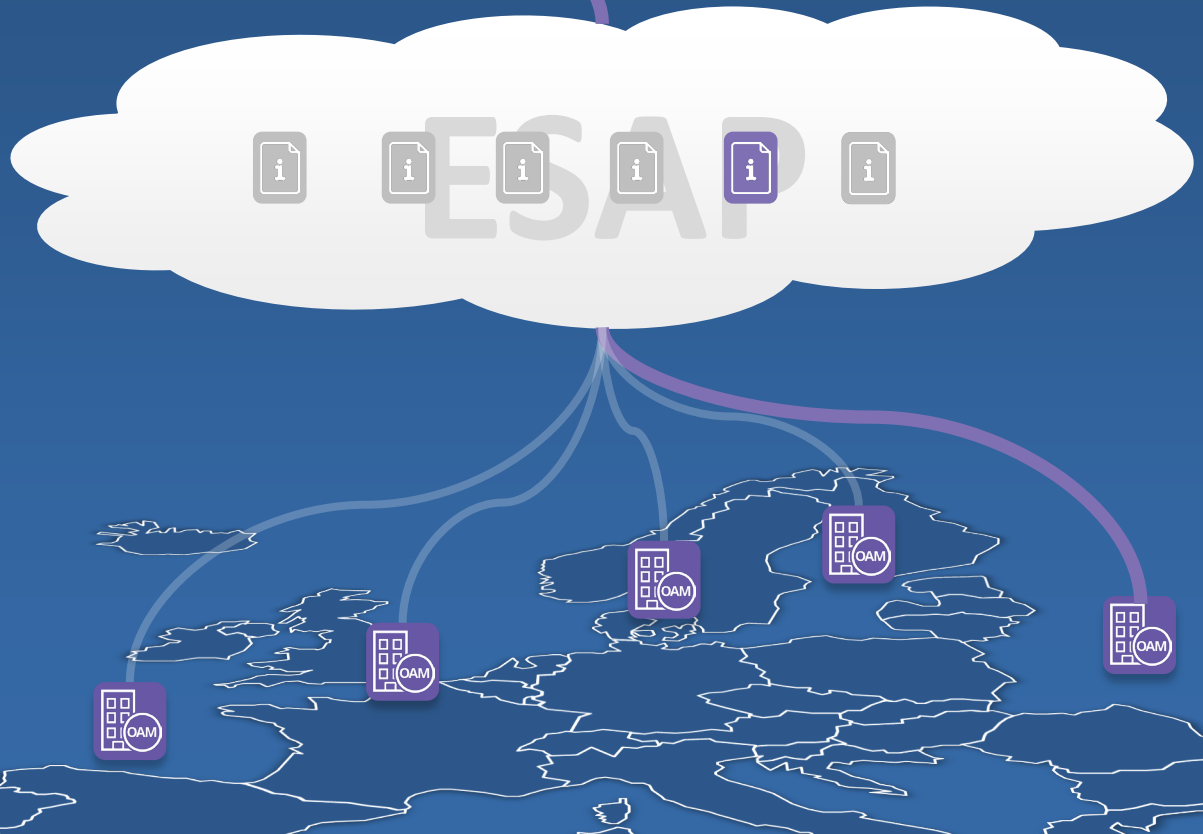
I am looking for steel producer in east Europe with low debt ratio to merge with a shipbuilding company

PETER



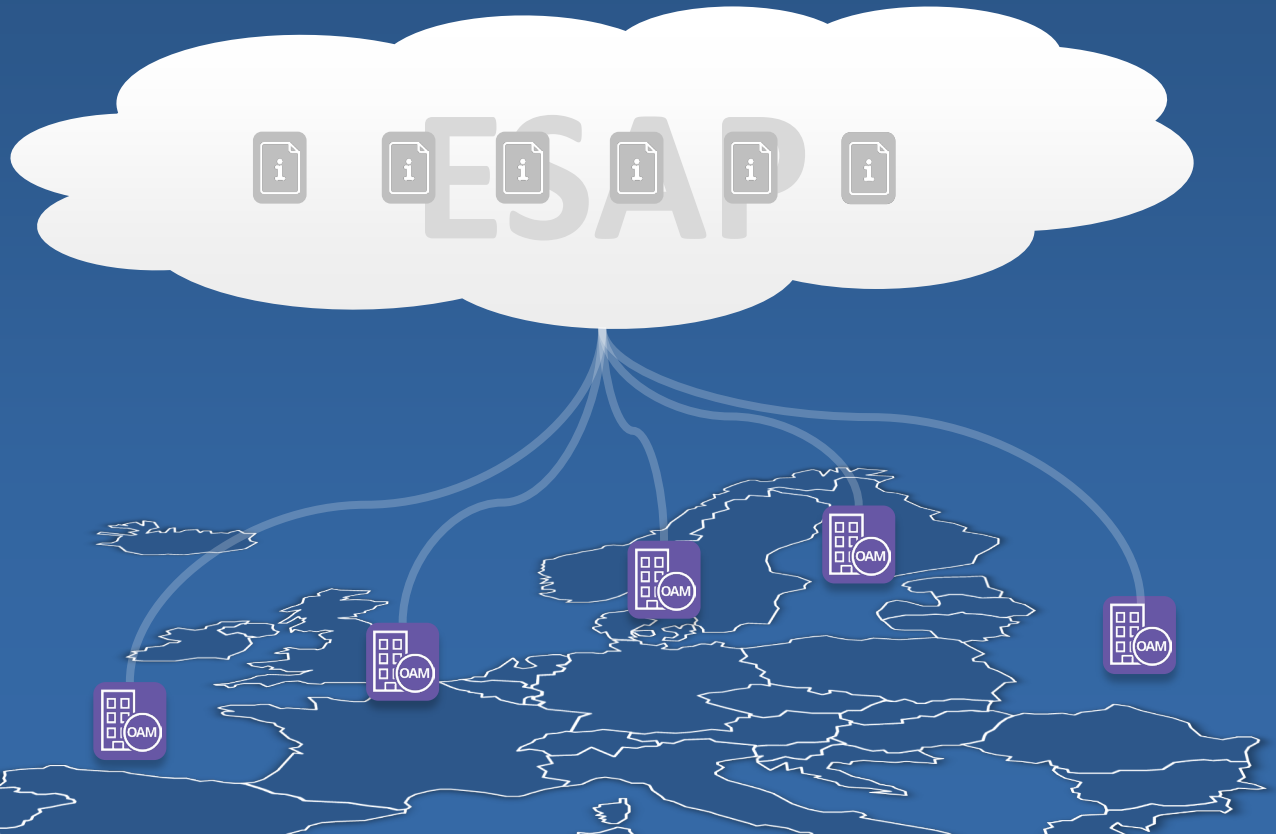
I am looking for steel producer in east Europe with low debt ratio to merge with a shipbuilding company

PETER



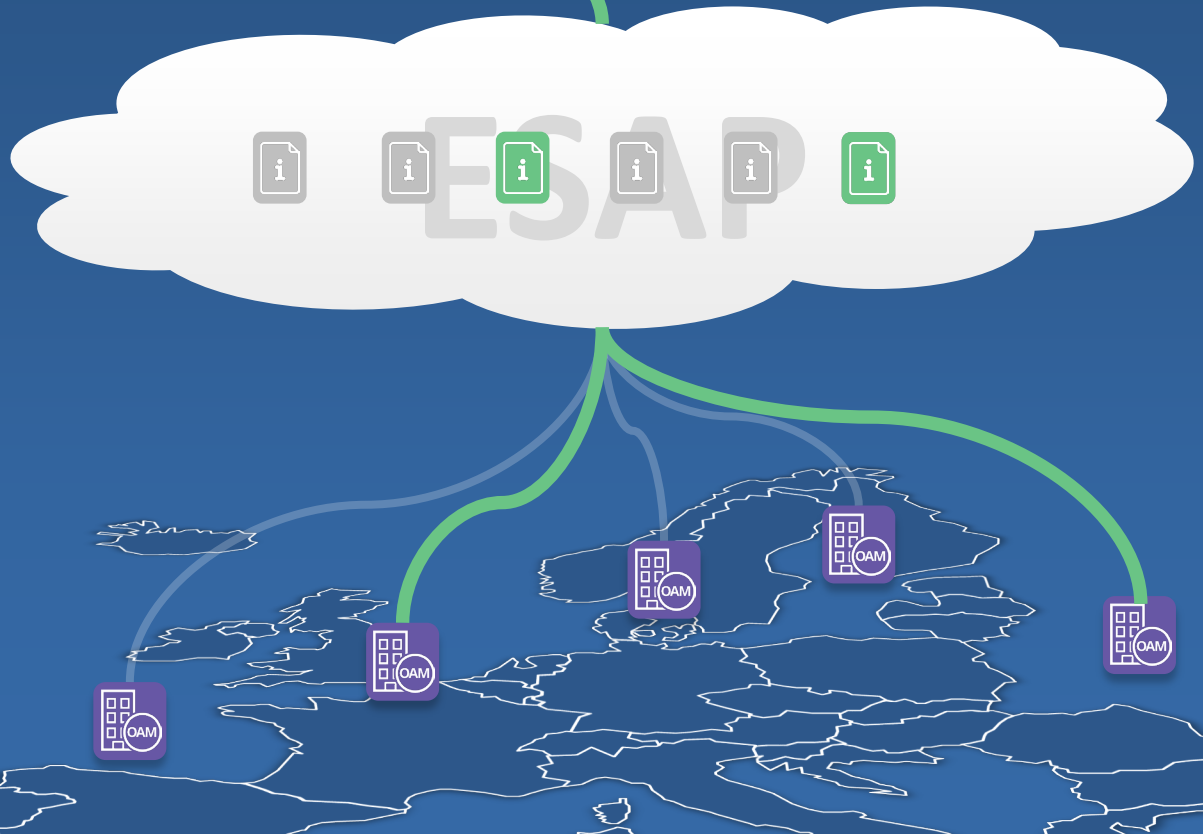
I am looking for companies with less than 1 ton GHG emissions per employee per year

MARIA



I am looking for companies with less than 1 ton GHG emissions per employee per year

MARIA



European Single Access Point

The how?

ESAP key components

- Machine-readable data (iXBRL, ISO 20022)
- Common identifiers (LEI, ISIN)
- Digital accessibility (APIs)



ESMA



Investors associations



Analysts associations



Stock exchanges



OAMs



NCAs



Listed comp. associations



Data vendors



Academia



EFRAG



Banking associations



Business registers



CRAs



Standard setters



SME associations



EU programs



Other stakeholders

Layer of open-source developed applications (analysis, use, deployment) based on ESAP data

Bulk download	Web application	Quality assurance	Linking datasets
Security	Administration & users	Machine learning	eIDAS & LEI
Data submission API	Data storage (cloud)	Metadata management	Innnovation

Layer of open-source developed applications allowing interaction with ESAP via Push API



OAMs and regulated repositories or regulated data disseminators



Registries



SMEs*



Listed companies*

European Single Access Point

The what?

ESAP key data

- Financial reports (IFRS)
- Investment information (Prospectus)
- Sustainability data (CSDR, EU Taxonomy)
- Other (EIOPA, EBA and other)

Corporate data scope

383 disclosure requirements across EU legal acts for listed entities only

Company establishment	Growth and gaining maturity	Company transformation, limitation or discontinuity
Common actions (reporting, registration, announcements, notifications etc.)		
Registration of a company Certification of product/service Application for public grants / funding Exemptions and privileges Tax reporting and registers Social contributions registers Procurement procedures Initial Coin/Token Offering	Initial Public Offering Prospectuses and KIIDs publication Issuance of debt securities Corporate actions publication Registration of related parties and subsidiaries Licences and permissions Application for public grants / funding Tax reporting and registers	Splits / mergers / acquisitions Nationalisation Filing for bankruptcy / insolvency Deregistration Going dormant Sanctions Prosecutions and regulatory warnings Internationalisation
Other actions and market-driven activities		
Advertising Website, social media Investment and business plans publication Forecasts, prospects and plans announcements	Ratings publication Rankings and indices Certificates Corporate sustainability reporting Major internal corporate announcements International cooperation and trade events	Announcements of recovery plans Bankruptcy auctions

European Single Access Point

The when?

2004: Transparency Directive

2017: European Financial Transparency Gateway

2020: High-level Forum on Capital Markets Vision

2020: EC DG FISMA Cost-benefit analysis

2021: European Commission COM/2021/723

2022: Council of the EU Position

2023: Council and Parliament agreement

Summer 2027* Planned launch of ESAP



European Single Access Point

What’s in it for business registers?

What are companies expecting?

1. Access to finance
2. Regional and global promotion
3. Trust (trade, grants, procurement)

Boosting post-pandemic economy

Type of financing	Micro (<= 2m EUR)	Small (<= 10m EUR)	Medium (<=43/50m EUR)	Large (> 43/50m EUR)
Internal funds/founders	Likely	Possible	Unlikely	Unlikely
Bank loans / credit cards / leasing / factoring	Likely	Likely	Likely	Likely
Grants / subsidies	Likely	Likely	Possible	Possible
Crowdfunding	Likely	Possible	Unlikely	Unlikely
Trade credit	Likely	Likely	Likely	Likely
SME Growth Market	Unlikely	Possible	Likely	Unlikely
Corporate bonds	Unlikely	Unlikely	Possible	Likely
Seed funds / Angel investors	Likely	Possible	Unlikely	Unlikely
Financial VCs	Unlikely	Possible	Likely	Likely
Strategic VCs	Possible	Possible	Likely	Likely
Equity listing (IPO)	Unlikely	Possible	Likely	Likely

Short-term financing

Long-term financing

Smaller financing needs

Larger financing needs

Ideas

- EU: Provide listed companies data to ESAP

- Non-EU: Connect to ESAP to promote economy and SMEs
- Non-EU: Additional information for BO/AML checks

- All: Innovate comparison of companies data
- All: Global ESG data sharing and access



BRAG

Thank you for your time

3rd IFC and Bank of Italy Workshop on “Data Science in Central Banking: Enhancing the access to and sharing of data”
17-19 October 2023

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Data sharing using a global data registry: on a place to discover global structured time series, macro and micro data[1]

Matt Nelson and Glenn Tice,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Data sharing using a global data registry
On a place to discover global structured time series, macro and micro data

Matt Nelson and Glenn Tice

Abstract

The increasing availability of data and the importance of evidence-based decision making have led to an urgent need for better data sharing practices. A 'global data registry' may provide a practical solution to enhance data discoverability, accessibility, and interoperability. This paper discusses the specific needs of data scientists and presents a potential solution for a global registry of structured statistical, parametric and unit record datasets using the Statistical Data and Metadata Exchange (SDMX) international standard ISO 17369:2013 and open-source software. Ultimately, we argue that a global registry for structured data can play a key role in enabling more effective use of datasets already available for policy making, research, and innovation by easing the burden of data acquisition on data scientists.

Keywords: data discovery, structured data, data science, SDMX

JEL classification: C80

The views expressed are those of the authors and do not necessarily reflect those of the BIS.

Contents

1. Navigating the labyrinth: Challenges in accessing and utilising public data for data scientists 3
2. Existing landscape of public data portals: A mosaic of regional and domain-specific resources..... 3
3. Data Scientists' needs for structured data discoverability and access 4
4. Single access points 5
5. Data registry: Paving the way for an evolved global data portal 7
6. SDMX 3.0: A standard metadata solution for a global data registry 8
7. A minimum viable implementation using open source software 9
8. Conclusion 10

1. Navigating the labyrinth: Challenges in accessing and utilising public data for data scientists

Data science, a discipline fuelled by the analysis and interpretation of data, relies heavily on access to diverse and comprehensive datasets for modelling, analysis, and innovation. While there exists a vast repository of structured public time series, macro, and micro data, sourced from international, regional, and national institutions, data scientists encounter formidable challenges in harnessing this information for their work.

Primarily, the obstacle lies in the fragmented nature of these datasets, scattered across disparate data portals hosted by different organisations. The data ecosystem lacks a unified platform or central repository, compelling users to navigate through numerous websites, each with its interface, to seek and retrieve relevant data. This dispersion and lack of a centralised resource significantly hampers the efficiency and productivity of data scientists, consuming valuable time and effort in the search for appropriate datasets.

Compounding this issue is the inadequacy of structural metadata accompanying these datasets. Often, the metadata—information describing the dataset's structure, attributes, and relationships—remains insufficient or poorly defined. Ambiguous or missing definitions of variables within the datasets lead to challenges in interpretation and utilisation. This ambiguity not only impedes the understanding of the dataset but also complicates the integration of these datasets into analytical models, hindering the accuracy and reliability of data-driven insights.

Furthermore, the absence of a standardised format or clear taxonomy across these datasets exacerbates the problem. Datasets from different sources may adopt varying conventions, making it challenging for data scientists to harmonise disparate data for meaningful analysis. The lack of a common data language or standardisation amplifies the complexities in data integration and hinders the seamless utilisation of these resources.

Consequently, the absence of a centralised repository, coupled with scattered data across multiple portals, inadequate structural metadata, and a lack of standardised formats, poses a significant barrier for data scientists. This fragmented landscape not only impedes the discovery and accessibility of data but also hampers the efficiency and effectiveness of data analysis, limiting the potential for innovation and insights derived from these public resources.

In essence, the critical need for a unified, easily accessible repository of structured public data, equipped with comprehensive metadata and standardised formats, becomes apparent. Such a platform would streamline the process of discovering and accessing global data, empowering data scientists to focus more on analysis and innovation rather than grappling with the challenges of data collection and interpretation across multifarious sources.

2. Existing landscape of public data portals: A mosaic of regional and domain-specific resources

The data landscape teems with a multitude of data portals, serving as repositories for a wealth of information curated by various entities ranging from international organisations to national governments. These portals, designed to offer access to a diverse array of data, play a pivotal role in facilitating data-driven research, analysis, and decision-making. However, their effectiveness in meeting the structured data needs of data scientists remains somewhat fragmented.

One category of these portals comprises open data repositories, epitomised by stalwarts such as the EU Open Data Portal and the US data.gov, among numerous other national-level services worldwide. These platforms boast extensive repositories spanning a broad spectrum of data types, encompassing everything from highly unstructured data like images and document scans to geospatial information and semi-structured datasets. While these portals offer a vast array of information, they encounter a challenge when it comes to structured data. Despite hosting structured tabular datasets, the metadata outlining the structure of these datasets often remains limited. For data scientists reliant on structured data as their primary working medium, this scarcity of comprehensive metadata poses a hurdle, impeding their ability to interpret data reliably and compare it with other datasets—an essential aspect of rigorous analysis and cross-comparison.

Another category comprises statistical data portals, prominently represented by institutions like the European Central Bank (ECB), Bank for International Settlements (BIS), Organisation for Economic Co-operation and Development (OECD), along with individual National Central Banks (NCBs) and National Statistical Offices (NSOs). These portals excel in providing exclusively structured data accompanied by robust metadata. However, their drawback lies in often being regional or domain specific. While they offer high-quality structured datasets with comprehensive metadata, their focus on specific regions or domains limits the breadth of data accessible to data scientists with diverse research interests or global-scale analyses.

Amidst these established portals, emerging initiatives like Google Data Commons strive to furnish a global data catalogue. However, these initiatives, while promising, are still in their nascent stages, grappling with immaturity in terms of comprehensiveness, usability, and accessibility.
The coexistence of these diverse data portals underscores the rich tapestry of available resources. Yet, their regional, domain-specific nature or their immaturity in catering to structured data needs presents challenges for data scientists seeking comprehensive, globally relevant structured datasets.

In essence, while these data portals serve as valuable repositories of information, the lack of a singular, all-encompassing platform for structured data presents a significant gap. Data scientists, reliant on structured data for their analyses, face the arduous task of navigating this mosaic of portals, often finding themselves constrained by the regional or domain-specific limitations, and grappling with inadequate metadata crucial for robust data interpretation and cross-comparison.

3. Data Scientists' needs for structured data discoverability and access

Delving deeper into the quest for structured data discoverability on a global scale unveils a series of critical prerequisites essential for data scientists. Foremost among these requirements is the necessity for a comprehensive view of global dataset availability. In a digital landscape dominated by ubiquitous search engines like Google and Bing, the rapid discoverability of general web content transcends geographical boundaries. However, attempts to reliably discover datasets in a similar manner remains challenging. While the finer granularity of availability information at series and data point levels is invaluable, a fundamental requirement emerges—a global index of availability at the dataset level. This foundational perspective offers substantial benefits for data scientists and serves as an initial stepping stone for the development of more sophisticated and refined discovery services.

Critical to the success of any discovery service is the establishment of a robust metadata framework. Such a framework must comprehensively describe the structure of each dataset (structural metadata) while also exposing referential metadata provided by the data product provider. For data scientists

seeking datasets pertinent to their applications, this metadata serves as a compass, guiding them towards candidate datasets and aiding in evaluating their suitability. Clarity in metadata is paramount; data scientists require precise definitions of the concepts encapsulated within the dataset variables. A dataset merely labelling an independent variable as 'country' without accompanying metadata diminishes its utility compared to a dataset explicitly defining the variable's conformity to international standards like the ISO 3166 country classification or providing a domain-specific taxonomy supplemented by geographic shape information and outlining precise regional boundaries.

Once datasets are discovered, ease of retrieval becomes paramount. That involves ensuring datasets are not only available in machine-consumable formats but also conform to well-understood schemas. Classically format means CSV, XML, JSON and Excel, but as the scale of granular data increases others such as Apache Parquet become increasingly important. Compatibility with commonly available connectors or data parsers streamlines the process, negating the need for bespoke code to transform incoming data into usable structures. Additionally, the transfer mechanism plays a pivotal role; facilitating easy movement of data from source to the point of computation and analysis with minimal overhead. Data published in formats like PDF or accessible only through a complex download procedure, while not inherently unusable, necessitate unwelcome pre-processing efforts before data scientists can delve into the core aspects of data analysis, impeding efficiency and agility in the data science workflow.

In essence, the efficacy of a global data discovery service for data scientists hinges on the provision of a comprehensive dataset view, robust metadata frameworks, machine-consumable formats, and seamless data transfer mechanisms—factors crucial in facilitating efficient, precise, and unencumbered access to structured datasets across global domains.

4. Single access points

As its name suggests, the planned European Single Access Point solves the problems of data discoverability and accessibility in the European public financial and sustainability domain by consolidating the data into a centralised warehouse and providing a single search and retrieval interface.

For the data consumer, the single access point approach provides a convenient abstraction hiding the complexity of querying possibly multiple heterogeneous data sources and even allowing analytics to be taken to the data, for instance using SQL to execute joins and aggregations within the data warehouse.

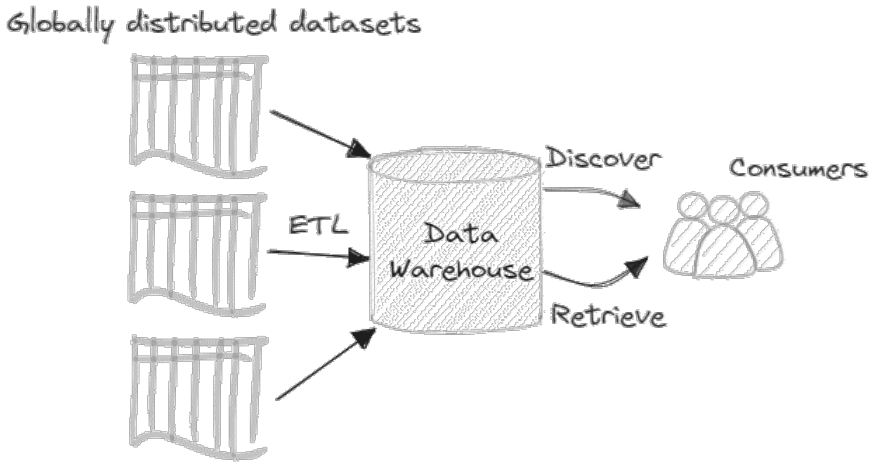


Figure 1 Single access point with data consolidated into a central warehouse.

Federation provides a variation on the approach where the data remains at source but is retrieved, integrated, transformed and transmitted to the data consumer via the single access point on demand.

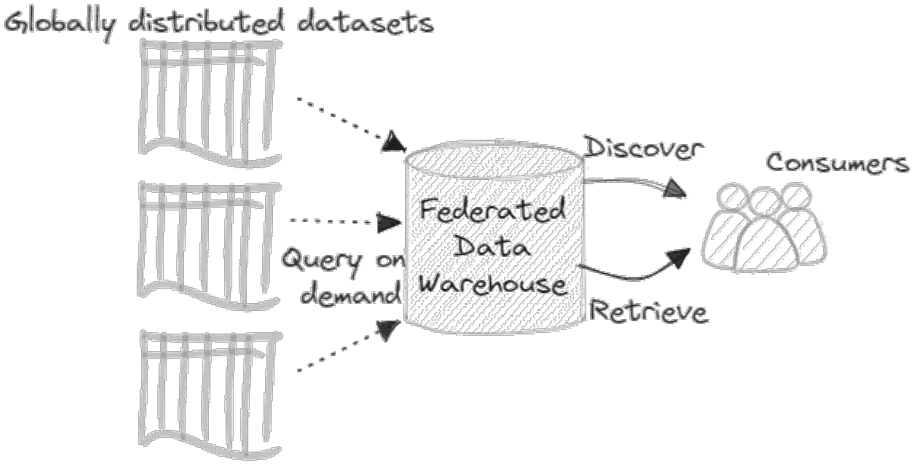


Figure 2 Single access point with a federated data warehouse.

For the single access point provider however, such services are often complex and costly to both build and operate particularly in term of storage and compute demand which invariably grows over time.

Establishing the data collection framework is also a significant challenge. The single access point approach is feasible where robust data collection pipelines are already in place, or at least data exchange protocols are in place limiting the scope of the problem to one of a technical implementation exercise.

5. Data registry: Paving the way for an evolved global data portal

As we have seen, single access points can be justified where the value of the data products and the benefits for data consumers outweigh what can often be the considerable cost of building and operating such services.

In the pursuit of a more streamlined and efficient global data ecosystem, the concept of a data registry emerges as a cost-effective alternative, reshaping the landscape of data access and utilisation. Unlike conventional data repositories that store the entirety of datasets, the data registry approach opts for a more agile and efficient methodology—centralising only the metadata detailing the structure of each dataset, along with pointers directing users to their actual locations.

This approach alleviates the burden of managing and handling the actual data, focusing instead on curating rich metadata that acts as a comprehensive catalogue of available datasets. Essentially, the registry functions as a compass, guiding data scientists to the precise locations of datasets without shouldering the weight of hosting the entire dataset repository.

Conceptually akin to existing services such as the EU Open Data Portal and similar data catalogues, this evolved data registry distinguishes itself by its emphasis on enriched metadata. It transcends mere pointers by providing sufficiently detailed and structured metadata, enabling effective discovery, evaluation, and interpretation of datasets. With comprehensive metadata elucidating the structure, attributes, and relationships within each dataset, data scientists gain the crucial insights necessary for confident and informed data utilisation.

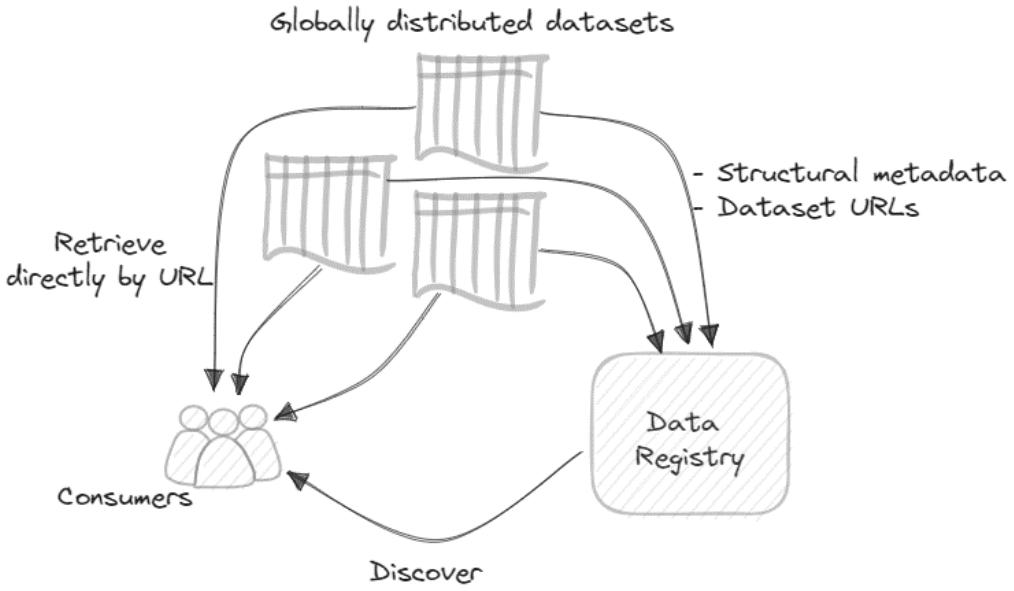


Figure 3 The data registry approach centralises only the data discovery and access metadata

Moreover, this model aligns harmoniously with the emerging paradigm of the Data Mesh—an approach geared towards organising and facilitating access to decentralised data across intricate data ecosystems. Two fundamental principles underpin the Data Mesh concept, resonating with the ethos of the data registry approach:

(a) Domain-Oriented Ownership: In this paradigm, each domain assumes responsibility for managing, curating, and granting access to its data. The data registry, by cataloguing metadata and pointers

rather than actual datasets, empowers individual domains to maintain control over their data while ensuring its discoverability on a global scale.

(b) Data Products as Standardised Interfaces: Embracing the philosophy of treating data as a product, the data registry fosters the creation of standardised interfaces—data products—by dedicated data teams within each domain. These interfaces provide uniform and standardised methods for accessing and consuming data, enhancing interoperability and simplifying the process of utilising diverse datasets across domains.

In essence, the data registry approach embodies a paradigm shift—a departure from the conventional handling of data repositories towards a leaner, metadata-focused model. By cataloguing rich metadata and employing pointers, it not only streamlines data access but also aligns seamlessly with the progressive principles of the Data Mesh, fostering domain-oriented ownership and standardised interfaces for enhanced data utilisation across complex and decentralised data ecosystems.

In summary, a data registry provides a light-weight and agile alternative where the case for a centralised or federated data warehouse and single access point cannot be argued.

6. SDMX 3.0: A standard metadata solution for a global data registry

As already noted, the utility of a global registry for structured data is dependent on a robust metadata framework. Version 3.0 of the ISO 17369 SDMX metamodel is a good candidate in that it provides specific artefacts for (a) data modelling, (b) data discovery and, (c) data registrations – pointers in the form of URL links to the actual data.

Modifications to the SDMX metamodel in version 3.0 improved the ability to model micro-datasets by adding support for multiple measures and array values for measures and attributes. These improvements mean that unlike previous versions, SDMX 3.0 is flexible enough to model most tall and wide microdata, plus macro, star schema and time series. Essentially, any dataset that can be managed as a data frame, but with the backing of the rich structural metadata essential for effective discoverability and interpretability.

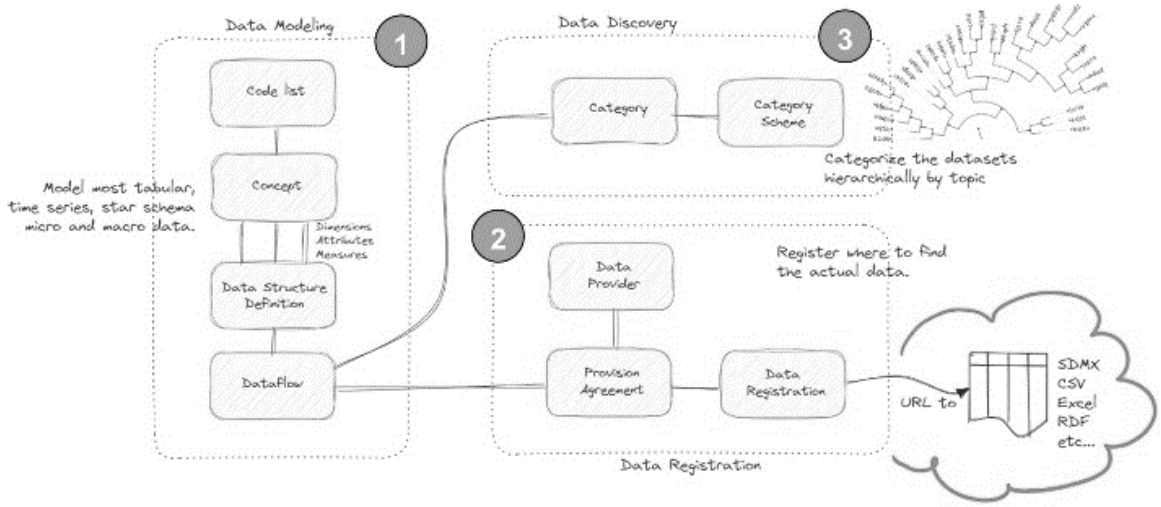


Figure 4 The SDMX 3.0 metamodel supports data modelling, discovery and registration metadata.

Examination of the SDMX 3.0 metamodel reveals several metadata artefacts specifically for data modelling. The four most important are the Code List, Concept, Data Structure Definition (DSD) and the Dataflow as illustrated in group (1) in Figure 4. An explanation of the precise purpose of these artefacts is beyond the scope of this paper, but the main principles are that the Dataflow is analogous to a fact table definition in a star schema data warehouse, and Code Lists define precise enumerations of values for discrete variables.

Data discovery is assisted by categorising datasets using hierarchical tree structures. In SDMX 3.0 the Category and Category Scheme illustrated in group (3) in Figure 4 support this purpose.

Finally, pointers to the actual data are managed as Data Registration artefacts (group (2), Figure 4) materialised as URLs. These are specifically decoupled from the Dataflow in the SDMX 3.0 metamodel to support the use case where multiple providers each contribute separate subsets towards a wider dataset. A common example being a dataset such as 'population' with global coverage where each country contributes observations for their own specific region.

For maximum accessibility, the URLs are expected to use the http:// or https:// protocols, but other more exotic protocols could also be offered, for instance s3:// referring to resources in the Amazon Web Services 'S3' object store.

7. A minimum viable implementation using open source software

While SDMX 3.0 provides a standard theoretical basis for a global data registry, software tooling is needed to make it a reality. For illustrative purposes, we present a reference solution based on the free and open-source Fusion Metadata Registry 11 (FMR) application which implements the SDMX 3.0 metamodel, critically including support for data registrations.

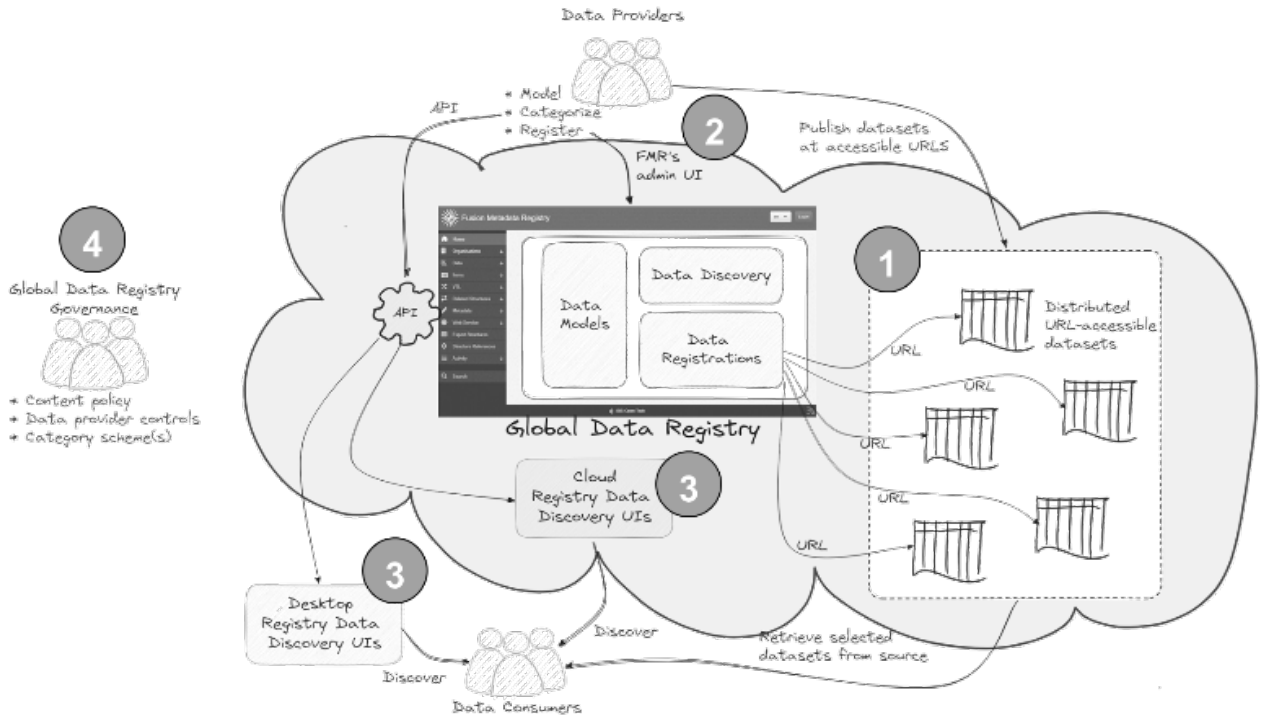


Figure 5 A minimum viable product using FMR 11

Structured datasets are first required, the only hard requirement being that they are individually resolvable by URLs. Next, the datasets need to be modelled, categorised and registered. Generally the data providers or the data product owners are best placed to undertake that task being those in possession of both the definitive structural and referential metadata. In practice the modelling task may best involve independent curators to help ensure consistency and maximise harmonisation of concepts within the global data registry, thus improving interoperability between datasets.

As previously noted, data scientists typically require both user and application programming interfaces. FMR's standardised SDMX REST API adequately supports the latter.

A cloud-hosted web user interface could be offered for simple use cases, although desktop applications perhaps implemented in Python could be conceived for more complex applications where data scientists need seamless integration of datasets into their workflow. Irrespective of the precise nature of the user interfaces, the key architectural principle is that all are loosely coupled to the central global data registry interfacing exclusively through its REST API.

Finally, a governance model is required to exercise control over the category tree(s) and other aspects of the registry including defining and enforcing content policy and administering the community of data providers.

8. Conclusion

While formal research is lacking on the subject, desktop analysis and anecdotal evidence indicates that data scientists would benefit from a place to discover global structured time series, macro and micro data. We observed that a large corpus of structured data is published by institutions globally, but the effort needed to find, retrieve and interpret them could be reduced.

The option of providing a 'single access point' provides attractive benefits for data scientists as the end users of such a discovery service, but the relatively high implementation and operating costs mean it is hard to make an investment case. The 'data registry' approach presents a more practical option in this respect as storage and compute, the main cost drivers for the single access point option, are contained and rise slowly as the service scales.

To test the feasibility, we examined a concrete reference implementation using the SDMX 3.0 international standard for the metadata framework and the Fusion Metadata Registry 11 software tool as the main data registry engine concluding that the key components for a minimum viable product exist.



Data sharing using a global data registry

On a place to discover global structured time series, macro and micro data

3rd IFC Workshop on Data Science in Central Banking

Glenn Tice & Matt Nelson (BIS MED IT)

The requirement for global structured data discoverability and access

The Problem

Public structured data is hard to find

- Data scientists need data
- Large global resource of structured data published by institutions worldwide

But

- No single place to discover what's available and how to get access to it

Existing Landscape

Many excellent data services already exist, but are variously regional, domain specific or not tuned for structured data

- EU open data portal
- US data.gov
- Institutional data portals
- European Single Access Portal (planned)
- Emerging initiatives (e.g. Google Data Commons)

The requirement for structured data discoverability and access

Global

Easily discover what datasets are available

Sufficient information to evaluate the suitability of a dataset

Confidently interpret and use datasets

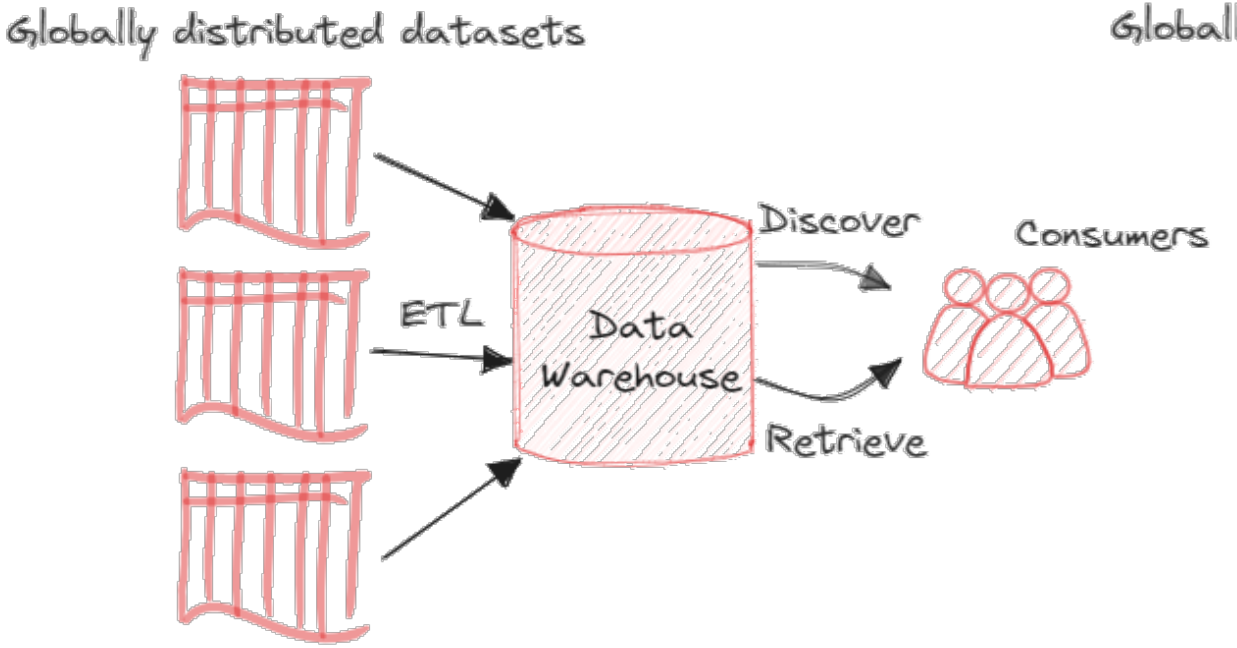
Tuned for structured datasets

Easily retrieve selected datasets

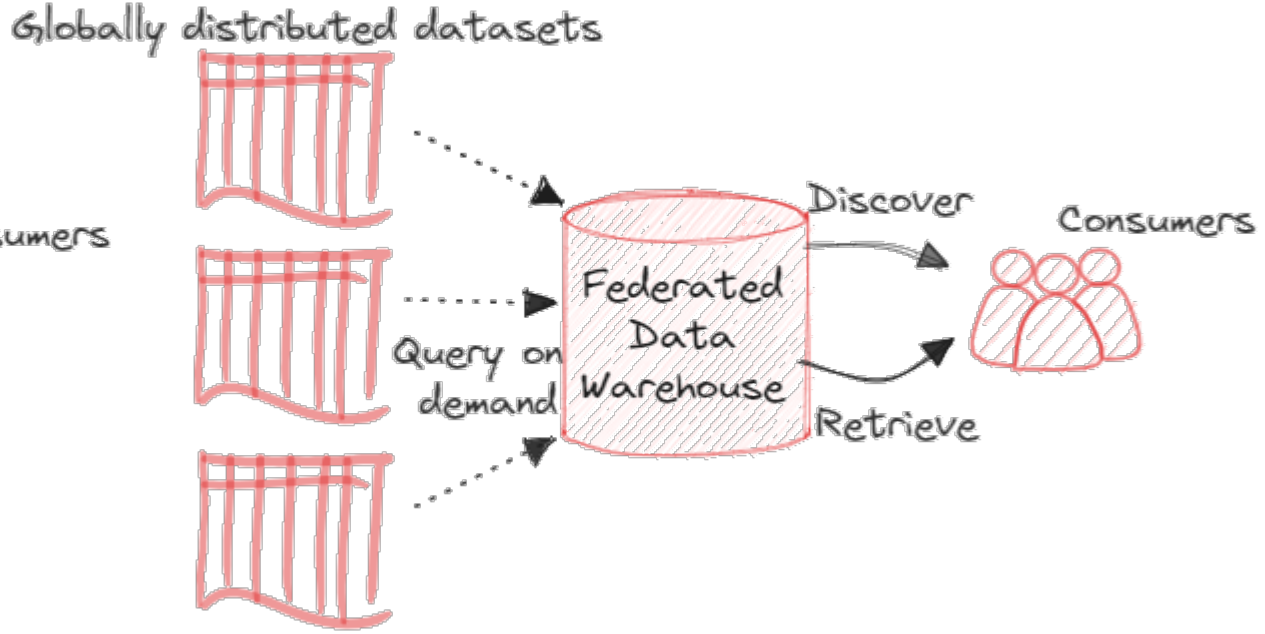
UI for interactive use

API for automation

Options – single access point



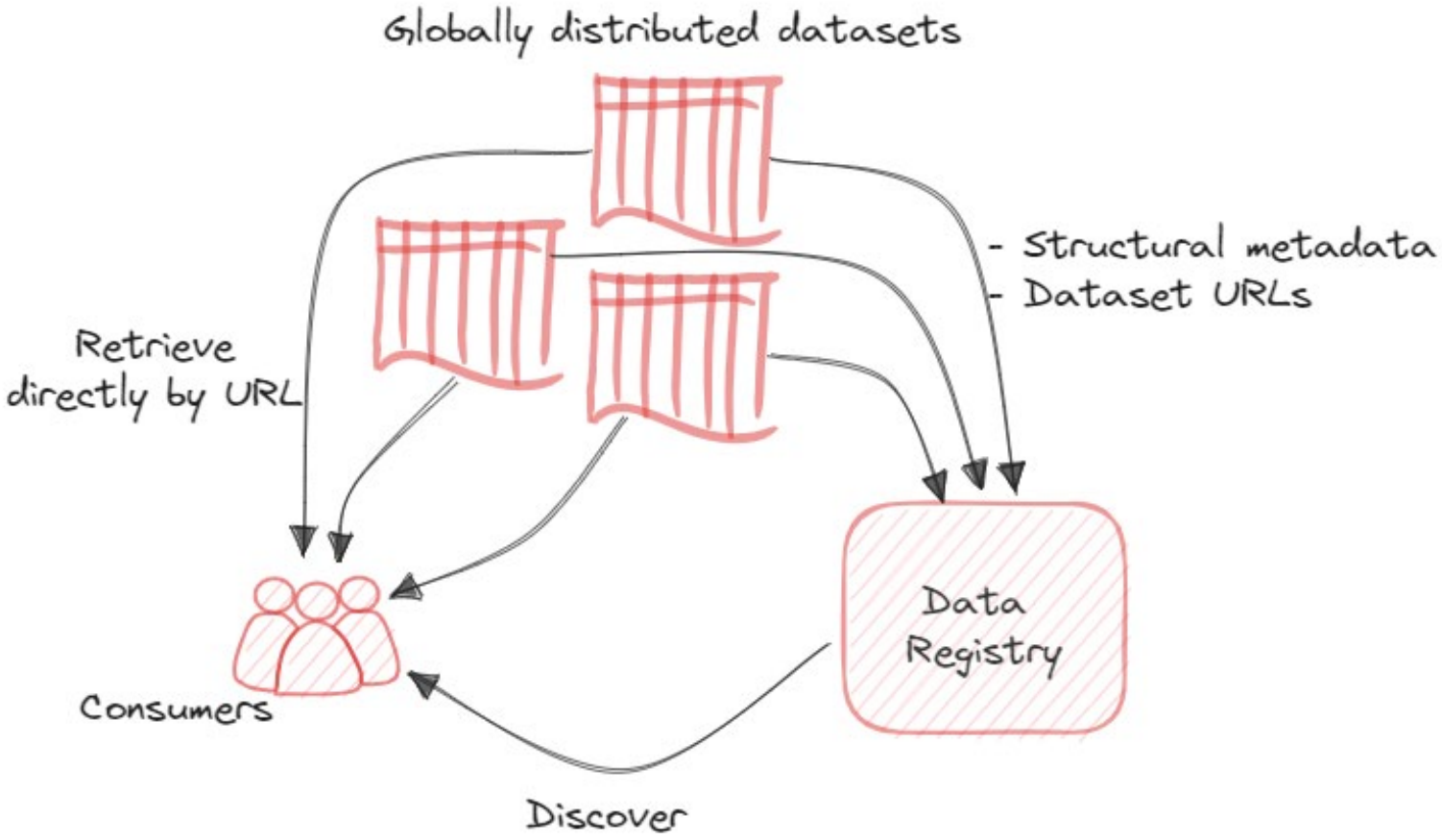
Data consolidated into a central warehouse



Federated data warehouse

A single access point is attractive but potentially complex and costly to implement

A ‘data registry’ provides a simpler alternative approach



- Catalogue of datasets and where to find them
- Each dataset has:
  - Detailed structural model
  - Data URL
- Decentralised – follows Data Mesh principles
- Technically simpler
- Legally simpler - risk of third party IP or personal data privacy claims reduced

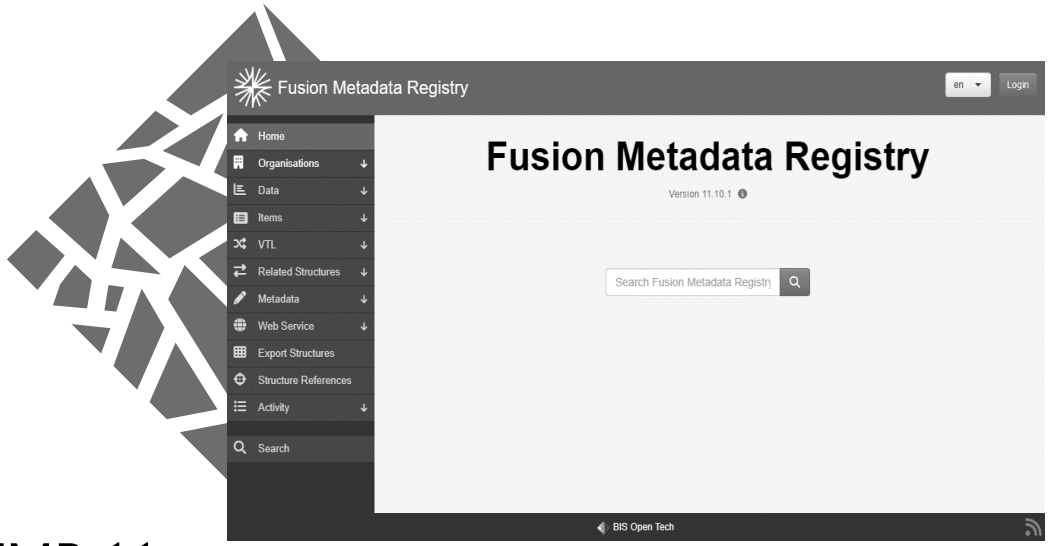
Concrete solutions

Data modelling

SDMX 3.0

- Data modelling
- Data discovery
- Data registrations – URLs link to the data

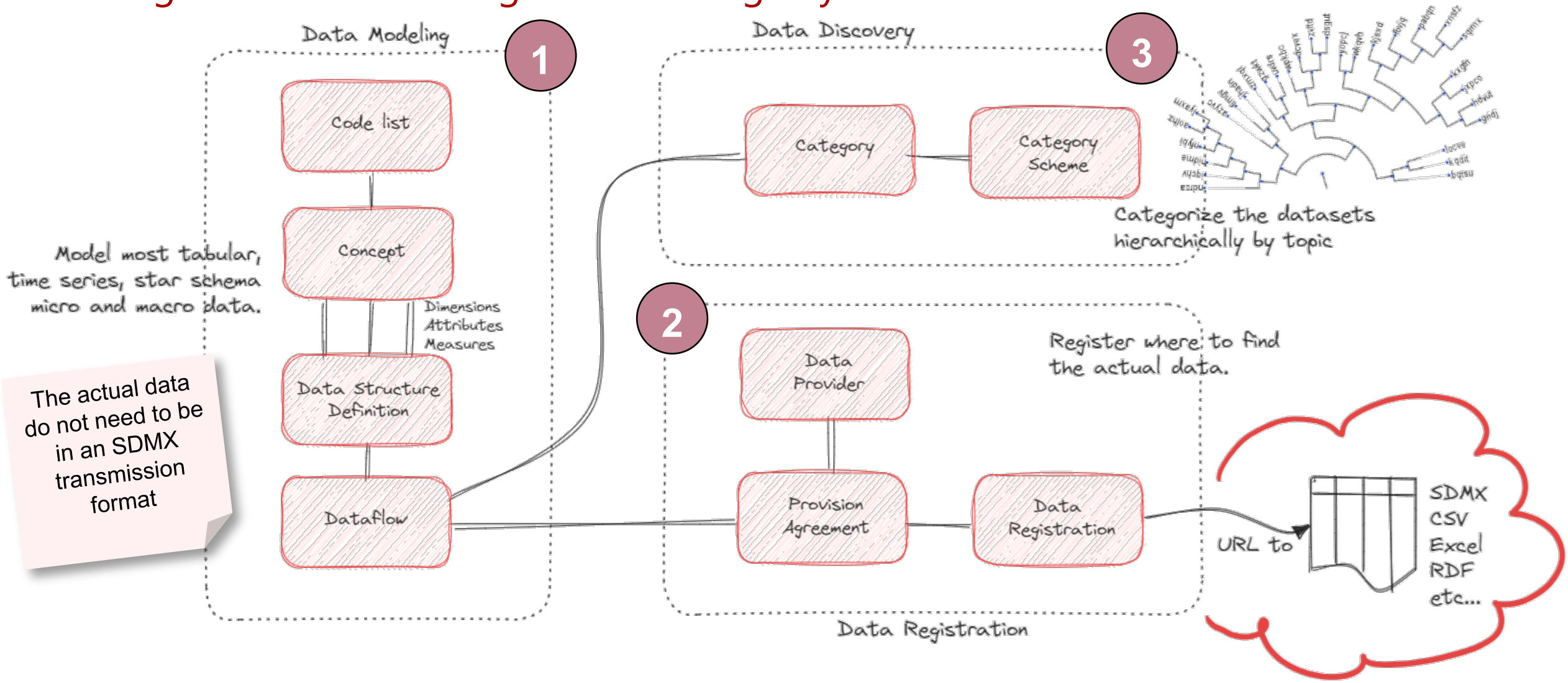
Software



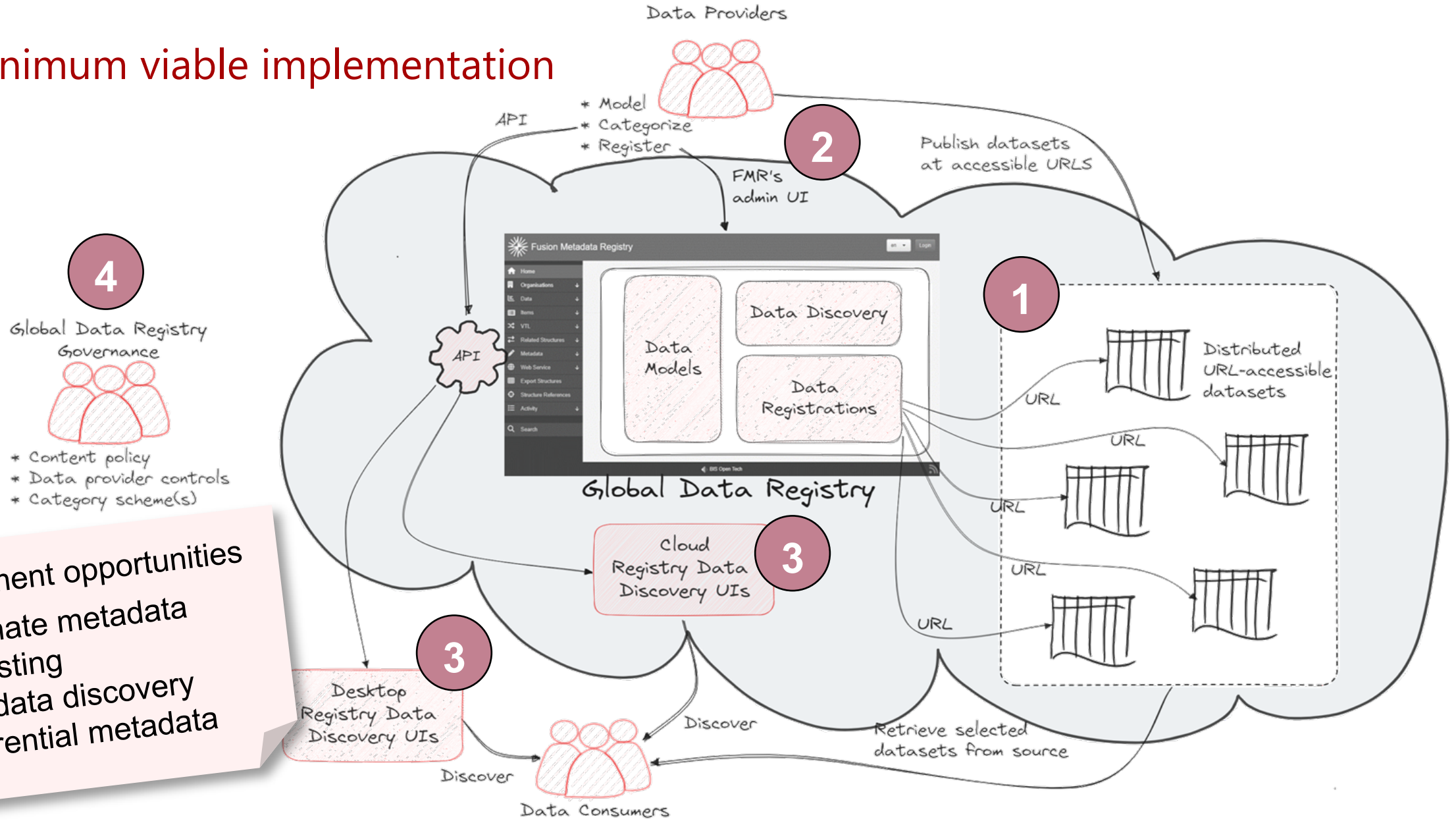
FMR 11

- SDMX 3.0 structural metadata registry
- Free and open source
- Cloud native
- BIS owned and maintained

Using SDMX 3.0 for the global data registry use case



A minimum viable implementation



Enhancement opportunities

- Automate metadata harvesting
- Metadata discovery
- Referential metadata

Conclusion

Plenty of public structured datasets published by institutions globally, but **hard to find and retrieve.**

The **global data registry** approach provides a centralised place to discover and retrieve datasets without the complexity of a single access point.

A practical implementation could be based on **SDMX 3.0** and existing software tools like **Fusion Metadata Registry.**

Is there **demand** from data consumers?
More consultation needed.

Thank you

Glenn Tice glennphilip.tice@bis.org
Matt Nelson matt.nelson@bis.org

BIS MED IT

References

About Fusion Metadata Registry (FMR)	Link
FMR quick start using Docker	Link
European open data portal data.europa.eu	Link
Banca d'Italia public Aggregated Data – Statistical Database	Link
US open data portal data.gov	Link
Google Data Commons	Link
European Single Access Point (ESAP) regulation	PDF Link
SDMX data registrations API query example (returns XML)	Link

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Overcoming data-sharing challenges in central banking: federated learning of diffusion models for synthetic mixed-type tabular data generation[1]

Timur Sattarov,
Deutsche Bundesbank,

Marco Schreyer,
University of St Gallen

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Overcoming data-sharing challenges in central banking: Federated Learning of Diffusion Models for Synthetic Mixed-Type Tabular Data Generation

Timur Sattarov[1,2] Marco Schreyer[2]

Abstract

Realistic synthetic tabular data generation encounters significant challenges in preserving privacy, especially when dealing with sensitive information in domains like finance. In this paper, we introduce FedTabDiff, a novel federated learning framework for generating high-fidelity mixed-type tabular data without centralized access to the original datasets. Leveraging the strengths of Denoising Diffusion Probabilistic Models (DDPM), our approach addresses the inherent complexities in tabular data, such as mixed attribute types and implicit relationships. More critically, FedTabDiff realizes a decentralized learning scheme that permits multiple entities to collaboratively train a generative model while respecting data privacy and locality. We extend DDPM into the federated setting for tabular data generation, which includes a synchronous update scheme and weighted averaging for effective model aggregation. Experimental evaluations on real-world financial and medical datasets attest to the framework's capability to produce synthetic data that maintains high fidelity, utility, privacy, and coverage.

Keywords: federated learning, diffusion models, tabular data, data sharing

JEL classification: C81, C18

[1] Research Data and Service Centre, Deutsche Bundesbank, Frankfurt am Main, Germany (timur.sattarov@bundesbank.de). The views expressed in this article are those of the author and do not necessarily represent those of the Bundesbank or the Eurosystem.

[2] School of Computer Science, University of St. Gallen, St. Gallen, Switzerland (marco.schreyer@unisg.ch)

Contents

Overcoming data-sharing challenges in central banking: Federated Learning of Diffusion Models for Synthetic Mixed-Type Tabular Data Generation 1

1. Introduction 3

2. Related Work 4

3. Federated Learning of Diffusion Models 5

4. Experimental Setup 7

   Datasets and Data Preparation 7

   Model Architecture 8

   Evaluation Metrics 9

5. Experimental Results 10

6. Conclusion 12

References 12

1. Introduction

In the rapidly changing financial regulatory landscape, data analytics has become increasingly vital, with central banks collecting extensive microdata to guide policy, assess risks, and maintain stability globally. The detailed nature of this data introduces unique challenges, notably in data privacy. Real-world tabular data is crucial for developing complex, real-life dynamic models but often contains sensitive information, necessitating strict regulations like the EU's *General Data Protection Regulation* and the *California Privacy Rights Act*. Despite implementing such safeguards, institutions remain wary of deploying Artificial Intelligence (AI) models due to risks of data leakage (Carlini et al. 2020; Bender et al. 2021; Kairouz et al. 2019) and model attacks (Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017; Salem et al. 2018), which could potentially reveal personally identifiable information or confidential training data (Carlini et al. 2020; Bender et al. 2021; Kairouz et al. 2019; Fredrikson, Jha, and Ristenpart 2015; Shokri et al. 2017; Salem et al. 2018). A promising avenue to address these risks originates from generating high-quality synthetic data. In this context, the term '*synthetic data*' denotes data of a generative process grounded in the inherent properties of real data. Modeling these processes provides a nuanced understanding of underlying patterns, unlocking insights, especially in high-stake domains (Assefa et al. 2020; Schreyer et al. 2019). Generating high-fidelity synthetic tabular data is crucial for regulatory-compliant data sharing, fostering collaboration among various entities, and enabling the modeling of rare but impactful events like fraud (Charitou, Dragicevic, and d'Avila Garcez 2021; Barse, Kvarnstrom, and Jonsson 2003). The Irving Fisher Committee (IFC) and the United Nations Economic Commission for Europe (UN-ECE) have both underscored the importance of data sharing and collaboration in the financial sector. The IFC's guidance note[3] discusses the challenges and benefits of cross-border cooperation, advocating for harmonized data practices and the use of advanced technologies. Similarly, the UNECE's guide[4] recommends that national statistical offices expand their data sources beyond national confines, incorporating international statistics to improve economic data. However, data sharing comes with a number of challenges such as data privacy or security concerns which can have severe consequences in case of any breaches of confidential information. In addition, the data transmission plays a big role as moving extensive datasets between central banks can become bandwidth-intensive and expensive. This approach is particularly pertinent as real-world tabular data is often characterized by inherent complexities, including:

1. **Mixed Attribute Types**: Tabular data comprises diverse attribute data types, including categorical, numerical, and ordinal data distributions. Modeling these complexities requires effectively integrating different data types into a cohesive model.

2. **Implicit Relationships**: Tabular data encompasses implicit relationships between individual records and attributes. Modeling these complexities necessitates

[3] https://www.bis.org/ifc/data_sharing_practices.pdf

[4] https://unece.org/sites/default/files/2021-02/Data%20sharing%20guide%20on%20web_1.pdf

disentangling the dependencies into a model representing the relationships between records and attributes.

3. **Distribution Imbalance**: The skewed distributions and imbalances in financial tabular data represent a significant hurdle. These factors demand advanced modeling techniques that precisely encapsulate and represent the nuanced patterns in financial data.

In recent years, deep generative models (Goodfellow, Shlens, and Szegedy 2014) have made impressive strides in the creation of diverse and realistic content, such as images (Brock, Donahue, and Simonyan 2018; Rombach et al. 2022), videos (Chan et al. 2019; Singer et al. 2022; Yu et al. 2023), audio (Wang et al. 2023; Le et al. 2023; Bai et al. 2022), code (Li et al. 2022; Chen et al. 2022; Le et al. 2022), and natural language (OpenAI 2023; Bubeck et al. 2023; Touvron et al. 2023). Lately, ***Denoising Diffusion Probabilistic Models (DDPM)*** (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020) demonstrated the ability to generate synthetic images of exceptional quality and realism (Dhariwal and Nichol 2021; Rombach et al. 2022; Gao et al. 2023). To effectively train these models, characterized by extensive parameters, substantial compute resources, and extensive training data are crucial (de Goede 2023). However, challenges arise when sensitive data is distributed across various institutions, such as financial authorities, and cannot be shared due to privacy concerns (Assefa et al. 2020; Schreyer et al. 2022; Schreyer, Sattarov, and Borth 2022). Recently, the concept of ***Federated Learning (FL)*** (McMahan et al. 2017; McMahan and Ram-age 2017) was proposed in which multiple devices, such as smartphones or servers, collaboratively train AI models under the orchestration of a central server (Kairouz et al. 2019). The training data remains decentralized, providing a pathway to learn a shared model without direct data exchange.

To summarize, the main contributions of this study are:

- Introduction of *FedTabDiff*, a novel federated learning framework for generating synthetic mixed-type tabular data, merging denoising diffusion models with federated learning to enhance data privacy.
- Empirical evaluation of *FedTabDiff* using real-world financial and healthcare datasets, illustrating its efficacy in synthesizing high-quality, privacy-compliant data.

The reference implementation of the code can be accessed through: https://github.com/sattarov/FedTabDiff.

2. Related Work

Lately, diffusion models (Cao et al. 2022; Yang et al. 2022; Croitoru et al. 2023) and federated learning (Aledhari et al. 2020; Li et al. 2021; Zhang et al. 2021a) triggered a considerable amount of research. In the realm of this work, the following review of related literature focuses on the federated deep generative modeling of tabular data:

Deep Generative Models: Xu et al. (Xu et al. 2019) introduced CTGAN, a conditional generator for tabular data, addressing mixed data types to surpass previous models' limitations. Building on GANs for oversampling, Engelmann and Lessmann (Engelmann and Lessmann 2021) proposed a solution for class imbalances by integrating conditional Wasserstein GANs with auxiliary classifier loss. Jordon et al. (Jordon, Yoon, and Van Der Schaar 2018) formulated PATE-GAN to enhance data synthesis privacy, providing differential privacy guarantees by modifying the PATE framework. Torfi et al. (Torfi, Fox, and Reddy 2022) presented a differentially private framework focusing on preserving synthetic healthcare data characteristics. To handle diverse data types more efficiently, Zhao et al. (Zhao et al. 2021) developed CTAB-GAN, a conditional table GAN that efficiently addresses data imbalance and distributions. Kim et al. (Kim et al. 2021) enhanced synthetic tabular data utility using neural ODEs, and Wen et al. (Wen et al. 2022) introduced Causal-TGAN, leveraging inter-variable causal relationships to improve generated data quality. Zhang et al. (Zhang et al. 2021b) offered GANBLR for a deeper understanding of feature importance, and Noock and Guillame-Bert (Nock and Guillame-Bert 2022) proposed a tree-based approach as an interpretable alternative. Kotelnikov et al.(Kotelnikov et al. 2022) explored tabular data modeling using multinomial diffusion models (Hoogeboom et al. 2021) and one-hot encodings, while FinDiff (Sattarov, Schreyer, and Borth 2023), foundational for our framework, uses embeddings for encoding.

Federated Deep Generative Models: De Goede et al. in (de Goede 2023) devise a federated diffusion model framework utilizing Federated Averaging (McMahan et al. 2017) and a UNet (Ronneberger, Fischer, and Brox 2015) backbone algorithm to train DDPMs on the Fashion-MNIST and CelebA datasets. This approach reduces the parameter exchange during training without compromising image quality. Concurrently, Jothiraj and Mashhadi in (Jothiraj and Mashhadi 2023) introduce *Phoenix*, an unconditional diffusion model that employs a UNet (Ronneberger, Fischer, and Brox 2015) backbone to train DDPMs on the CIFAR-10 image database. Both studies underscore the pivotal role of federated learning techniques in advancing the domain.

To the best of our knowledge, this is the first attempt utilizing diffusion models in the federated learning setup for synthesizing mixed-type tabular financial data.

3. Federated Learning of Diffusion Models

This section outlines our proposed *FedTabDiff* model for integrating *Federated Learning* (FL) with *Denoising Diffusion Probabilistic Models* (DDPM), aiming to enhance the privacy of generated mixed-type tabular data.

Gaussian Diffusion Models. The Denoising Diffusion Probabilistic Model (Sohl-Dickstein et al. 2015), (Ho, Jain, and Abbeel 2020) is a latent variable model that uses a forward process to perturb data $x_0 \in \mathbb{R}^d$ incrementally with Gaussian noise ϵ, then restores it through a reverse process. Starting at x_0, latent variables $x_1, \ldots, x_T$ are derived via a Markov Chain, transforming them into Gaussian noise $x_T \sim \mathcal{N}(0,\mathrm{I})$, defined as: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathrm{I})$

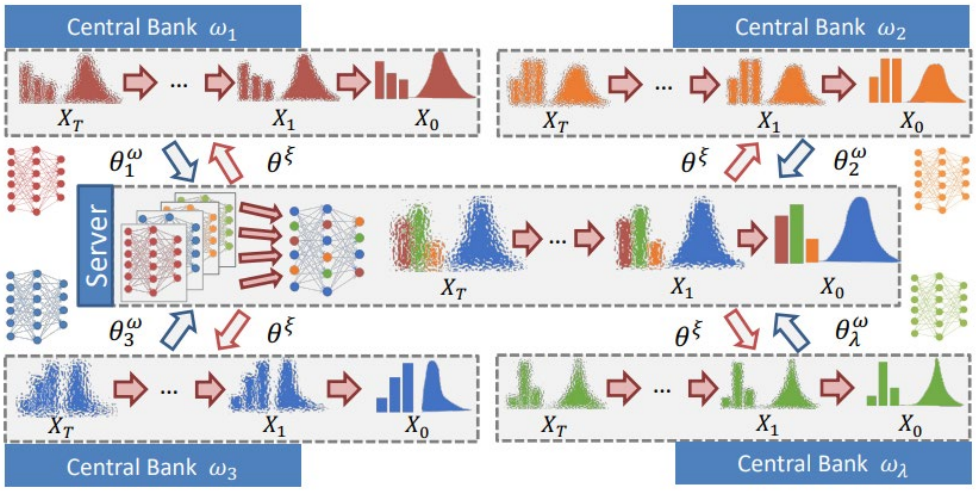


Figure 1. *Schematic illustration of the FedTabDiff model depicting each central bank ω_i training a local diffusion model 'FinDiff'. Each timestep $X_T, \dots, X_1, X_0$ reflects the latent data representations in the reverse diffusion. Model parameters θ_i^ω are aggregated on the server producing model θ^ξ, which is subsequently shared back after every training round r = 1,..., R.*

Here, β_t is the noise level at timestep *t*. Sampling x_t from x_0 for any *t* is expressed as

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{1-\widehat{\beta}_t}x_0, \widehat{\beta}_t \mathrm{I}),$$

where $\widehat{\beta}_t = 1 - \prod_{i=0}^{t}(1-\beta_i)$. In the reverse process, the model denoises x_t to recover x_0. To approximate, a neural network with parameters θ is trained, parameterizing each step as following:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)),$$

where μ_θ and Σ_θ are the estimated mean and covariance. Following Ho et al. (Ho, Jain, and Abbeel 2020), with Σ_θ being diagonal, μ_θ is computed as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\widehat{\alpha_t}}}\epsilon_\theta(x_t, t))$$

Here, $\alpha_t \coloneqq 1 - \beta_t, \widehat{\alpha_t} \coloneqq \prod_{i=0}^{t}\alpha_i$ and $\epsilon_\theta(x_t, t)$ is the predicted noise component. Empirical evidence shows that using a simplified mean-squared error loss yields superior results compared to the variational lower bound $log\ p_{(\theta)}(x_0)$.

$$\mathcal{L}_t = \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2$$

Federated Learning. The DDPM learning process is enhanced by *Federated Learning* (McMahan et al. 2017). FL enables knowledge acquisition from data distributed across *C* clients, denoted as $\{\omega_i\}_{i=1}^{c}$. The training dataset is partitioned into *C* sub-datasets, $D = \{D_i\}_{i=1}^{c}$, each privately accessible only to a single client ω_i. Data subsets D_i and D_j, where $i \neq j$, can have different data distributions. We build on the FinDiff model for mixed-type tabular data generation (Sattarov, Schreyer, and Borth 2023), inspired by the work presented in (de Goede 2023; Jothiraj and Mashhadi 2023). We introduce a central FinDiff model f_θ^ξ with parameters θ^ξ, maintained by a trusted entity and collectively learned by clients $\{\omega_i\}_{i=1}^{c}$. Each client ω_i maintains a decentralized FinDiff

model $f_{\theta,i}^{\xi}$ and contributes to learning the central model. The federated optimization uses a synchronous update scheme over $r = 1, \ldots, R$ communication rounds (McMahan et al. 2017). In each round, a subset of clients $\omega_{i,r} \subseteq \{\omega_i\}_{i=1}^{c}$ is selected. These clients receive the current central model parameters $\theta_{i,r}^{\xi}$, perform optimization, and then send their updated model parameters back for aggregation. The schematic process with four clients is depicted on fig. 1. Following Federated Averaging (McMahan et al. 2017) is used to compute a weighted average over the decentralized model updates, defined as:

$$\theta_{r+1}^{\xi} \leftarrow \frac{1}{D} \sum_{i=1}^{\lambda} |D_i| \, \theta_{i,r+1}^{\omega}$$

where λ denotes the clients, r current communication round, $|D|$ is the total sample count, and $|D_i| \subseteq |D|$ is the number of samples accessible by the client ω_i.

4. Experimental Setup

This section describes the details of the conducted experiments, encompassing the data preparation steps, model architecture, and evaluation metrics.

Datasets and Data Preparation

Two real-world tabular datasets are used for the evaluation:

1. **City of Philadelphia Payments**[5] encompasses a total of 238,894 payments generated by 58 distinct city departments in 2017. Each payment includes 10 categorical and 1 numerical attribute. The feature *doc_ref_no_prefix* definition (document reference number prefix definition) was used for the non-iid split.

2. **Diabetes hospital data**[6] encompasses a total of 101,767 clinical care records collected by 130 US hospitals between the years 1999-2008. Each care record includes 40 categorical and 8 numerical attributes. The feature *age* (a patient's age group) was used for the non-iid split.

Non-iid partition. We aim to substantiate the model's efficacy and effectiveness, ensuring generalizability across non-iid data splits. The data is non-iid partitioned by segregating based on a categorical feature. The five dominant categories within this feature serve as a basis for data allocation to different clients, thereby introducing a realistic non-iid and unbalanced data environment for federated training. We present the descriptive statistics and details on the non-iid data partitioning scheme in the table 1.

[5] https://www.phila.gov/2019-03-29-philadelphias-initial-release-of-city-payments-data

[6] https://www.kaggle.com/datasets/brandao/diabetes

Table 1. *Non-iid and unbalanced data partitioning scheme. Only a subset $D_i \subset D$ is privately accessible to a client ω_i.*

	# samples in D_i	
Client	**Philadelphia**	**Diabetes**
ω_1	40,038	9,685
ω_2	28,521	17,256
ω_3	16,831	22,483
ω_4	93,119	26,068
ω_5	36,793	17,197
all	215,302	92,689

Numeric features were normalized using the Quantile Transformer[7] and categorical attributes were encoded using embeddings as per (Sattarov, Schreyer, and Borth 2023).

Model Architecture

The model architecture for both datasets involves the feed forward neural network with four layers of 1024 neurons each. Models are trained for up to 1000 communication rounds R = 1000 with a mini-batch size of 512 using Adam optimizer (Kingma and Ba 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$. PyTorch v2.0.1 (Paszke and Gross 2019) is used for optimization of the model parameters.

Diffusion Model Training Hyperparameters. We determined that the optimal number of diffusion steps for training the diffusion model is T = 500. We used a linear scheduler with $\beta_{start} = 0.0001$ and $\beta_{end} = 0.02$. Also, each embedding dimension of categorical attributes was assigned to 2.

Federated Learning Hyperparameters. At every communication round $r = 1, \dots, R$ each client ω_i performed 20 model optimization updates (client rounds) on its local model θ_i^{ω} before sharing the model parameters. All five clients $\lambda = 5$ were selected at every communication round for decentralized optimization iteration. The Flower framework v1.4.0 (Beutel et al. 2022) simulates the federated learning scenario and trains the model with multiple clients.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

Evaluation Metrics

We employ a series of evaluation metrics, including fidelity, utility, coverage, and privacy, to evaluate the effectiveness of our model quantitatively. The metrics are selected to represent diverse aspects of data generation quality, ensuring a holistic view of model performance.

Fidelity. Fidelity evaluates the extent to which the synthetic data replicates the real data, a process executed at both column and row levels. The fidelity of columns is investigated by assessing the similarity between each column in the synthetic and real datasets. For numeric attributes, the Kolmogorov-Smirnov statistic (Massey 1951), is applied to quantify the utmost difference between empirical and theoretical cumulative distributions. The Total Variation Distance measures the disparity between the real and synthetic categorical columns. It is defined as $\sum_{c \in C} | p(x^{d_c}) - p(s^{d_c}) |$, with $p(s^{d_c})$ representing the proportion of existing categories c in attribute d. The row fidelity is measured using correlations between a pair of columns. For numeric attributes, a Pearson correlation (Benesty et al. 2009) is computed between a pair of columns. Next, the absolute difference between the correlations of the real and synthetic attribute pairs is used to estimate the coefficient. We used the contingency table for categorical attributes by computing the Total Variation Distance (TVD) on a pair of categories in attributes a and b.

Utility. The evaluation of synthetic data crucially depends on the measure of its utility, which involves quantifying its functional equivalence to real-world data. This metric involves training machine learning models on synthetic datasets and subsequent evaluations on original ones. This criterion measures the applicability and quality of synthesized data in various downstream tasks, like classification or anomaly detection. This study measures utility by training classifiers on synthetic data sharing the same dimensions with the training set, followed by evaluations on the actual test set. The average accuracy denotes the resultant utility, consolidated over all classifiers. The classifiers utilized in this study for a diversified approach include Random Forest, Decision Trees, Logistic Regression, Ada Boost, and Naive Bayes.

Coverage. The coverage metric quantifies how much a synthetic column encompasses all potential categories found within a real column. In dealing with categorical features, this metric initially calculates the number of unique categories inherent in the real column. Subsequently, it determines the presence of those categories within the synthetic column, and returns the proportion of real categories mirrored in the synthetic data.

Privacy. The privacy metric quantifies the degree to which the synthetic data inhibits the identification of original data entries. The Distance to Closest Records (DCR) is utilized to assess the privacy of the generated data. The DCR is determined as the nearest distance from a synthetic data point to the authentic data points. The final score is the median of the DCRs for all synthetic data points.

The implementation of the evaluation metrics originates from the Synthetic Data Vault (SDV) library v1.3.0 (Patki, Wedge, and Veeramachaneni 2016).

Evaluation Process. The effectiveness of the *FedTabDiff* federated approach is assessed by comparing it with a non-federated scenario. Here, each client trains a

FinDiff synthesizer, and the quality of generated data is evaluated against all data partitions, including the entire dataset. This comparison helps determine the added value of the federated setting and whether the global model effectively aggregates comprehensive data knowledge. Furthermore, it reveals whether clients (or central banks), especially those from underrepresented groups with limited data volume, benefit from the global model's capability to generate data reflecting broad characteristics.

5. Experimental Results

This section presents the results of the experiments, demonstrating the efficacy of the *FedTabDiff* model and providing quantitative analyses. The results, detailed in table 2 together with depicted heatmaps in fig. 2, compare the Federated *FedTabDiff* model with Non-Federated FinDiff models across fidelity, utility, coverage, and privacy.

***Table 2**. Comparative analysis of the Federated (FedTabDiff) versus Non-Federated (FinDiff) models, evaluated using the full dataset D. Non-Federated diffusion models are trained individually at each client ω_i with subset $D_i \subset D$ (reflected in column `Split') and evaluated against the entire dataset D.*

			Evaluation Measures			
Dataset	**Client**	**Split D_i**	**Fidelity ↑**	**Utility ↑**	**Coverage ↑**	**Privacy ↓**
Philadelphia	ω_1	19%	0.267 ± 0.03	0.263 ± 0.04	0.689 ± 0.03	3.162 ± 0.19
	ω_2	13%	0.264 ± 0.03	0.325 ± 0.06	0.681 ± 0.02	3.103 ± 0.13
	ω_3	8%	0.207 ± 0.03	0.118 ± 0.04	0.847 ± 0.04	3.178 ± 0.03
	ω_4	43%	0.394 ± 0.01	0.430 ± 0.01	0.863 ± 0.02	2.919 ± 0.14
	ω_5	17%	0.238 ± 0.03	0.197 ± 0.03	0.898 ± 0.01	3.359 ± 0.33
	FedTabDiff		**0.590 ± 0.01**	**0.837 ± 0.03**	**0.944 ± 0.03**	**2.607 ± 0.18**
Diabetes	ω_1	10%	0.217 ± 0.01	0.104 ± 0.03	0.944 ± 0.02	10.261 ± 0.25
	ω_2	18%	0.269 ± 0.01	0.186 ± 0.01	0.943 ± 0.03	10.091 ± 0.38
	ω_3	24%	0.314 ± 0.01	0.242 ± 0.01	**0.946 ± 0.01**	9.895 ± 0.31
	ω_4	28%	0.331 ± 0.01	**0.281 ± 0.01**	0.939 ± 0.01	9.941 ± 0.21
	ω_5	18%	0.269 ± 0.01	0.185 ± 0.01	0.943 ± 0.02	10.139 ± 0.19
	FedTabDiff		**0.720 ± 0.01**	0.265 ± 0.01	0.906 ± 0.01	**3.120 ± 0.09**

*Scores are derived from the averaged results and standard deviations of five experiments, each initiated with distinct random seeds

Fidelity. The *FedTabDiff* model demonstrates superior fidelity, outperforming non-federated models by up to 56% in the Philadelphia dataset and by approximately 118% in the Diabetes dataset. The model's adaptability across varying data distributions is underscored by examining fidelity across different data subsets, D_i, as presented in fig. 2a. The *FedTabDiff* showcases stable fidelity scores across all clients indicating its suitability for the proposed task. Conversely, non-Federated models achieve analogous scores solely on data subsets exposed to local training, a phenomenon visibly depicted through the emphasized diagonal scores in the associated heatmaps. This outcome aligns with expectations, considering that the model, trained only on the local subset D_i, acquaints itself exclusively with the proper-

ties therein, remaining uninformed about the structural intricacies of the comprehensive (global) dataset D.

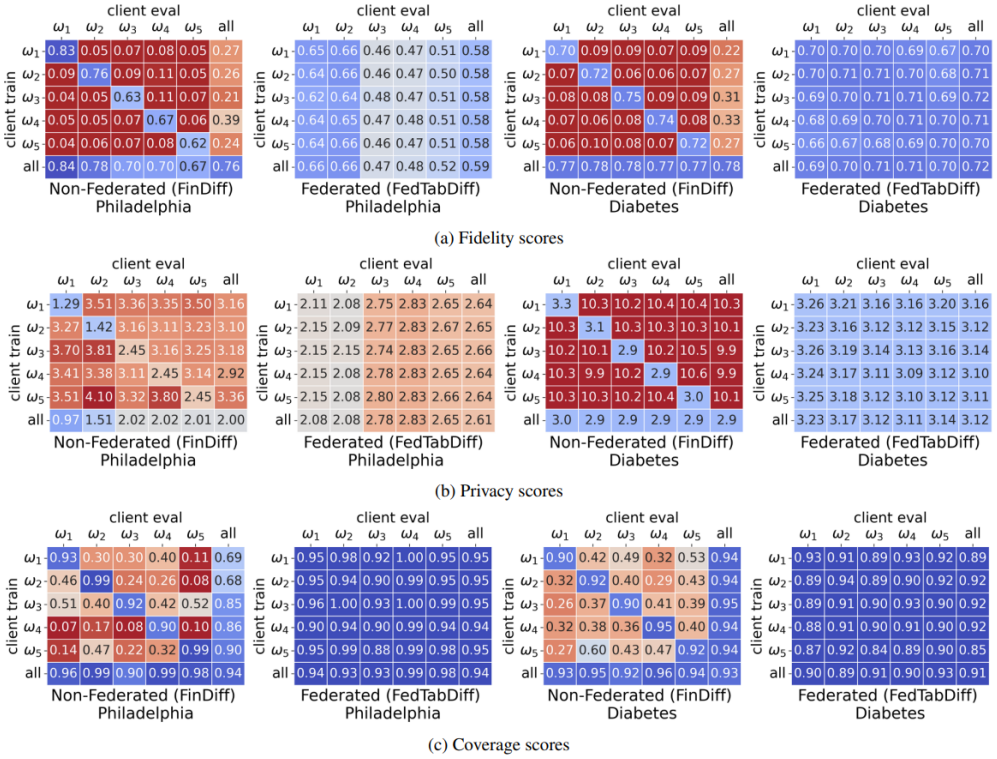


***Figure 2**: Comparison of evaluation metrics between Federated (FedTabDiff) and Non-Federated (FinDiff) diffusion models on individual client ω_i. In the Non-Federated model, each client is trained on its data subset D_i and evaluated across all subsets.*

Utility. The Utility scores from the table 2 indicate a substantial performance enhancement by *FedTabDiff* in the Philadelphia dataset, with an over 218% improvement compared to individual client models. With a score of 0.837, it outperforms all client models, highlighting its superior capacity to synthesize data with broader insights in a federated setting. For the Diabetes dataset, *FedTabDiff* ranks second, slightly behind client ω_4. This variation is attributed to the datasets' partitioning, where the Philadelphia dataset's imbalance is more pronounced, influencing the federated model's relative effectiveness.

Privacy. The privacy metric shows a substantial improvement with *FedTabDiff*, which scores 2.607 for Philadelphia and 3.120 for Diabetes. Although the scores for *FedTabDiff* seem promising, carefully examining the chosen privacy metric is imperative to ensure the validity of the synthetic data generated. Generating samples with substantial deviation from the original might introduce a risk of producing unrealistic data. Simultaneously, generating samples analogous to the original data may result in data replication, potentially breaching privacy safeguards. A balanced strategy, which moderately navigates between these two extremes, may offer a viable path forward.

Coverage. In the Diabetes dataset, *FedTabDiff* scores 0.906, slightly below the highest non-federated score of 0.946, while leading in Philadelphia with 0.944. These results highlight its performance and bring nuances in coverage metrics to light. For instance,

uniform random generation of categorical entries might inflate scores, as seen in Diabetes dataset results in table 2 and the last column of the non-Federated heatmaps in fig. 2c. Conversely, while the *FedTabDiff* gives attention to distribution shape, there is a chance to overlook minority entries, especially within skewed distributions.

Overall, the *FedTabDiff* model consistently surpasses non-federated models across all evaluated metrics. This superior performance of the federated learning model demonstrates its effective learning capabilities from the combined datasets of all clients. Consequently, it affirms that the federated learning setup allows diffusion models to effectively learn the underlying data structures across all clients, thereby generating data of higher quality.

6. Conclusion

This study introduces *FedTabDiff*, a novel federated diffusion-based generative model for the high-fidelity synthesis of mixed-type tabular data, enabling collaborative model training without sharing sensitive data. The model enhances performance metrics across two diverse datasets and maintains data privacy in a federated setting. The methodological strength of *FedTabDiff* is thoroughly examined through extensive experiments rooted in a realistic, non-iid data environment. The model's ability to prevent sharing sensitive information during the distributed training of generative models, specifically for synthesizing tabular data for downstream tasks, is apparent. Finally, through this federated approach, the positions of underrepresented entities with limited data are substantially enhanced, thereby promoting the practice of responsible AI in the financial sector. This phenomenon is particularly profound in financial regulation, where the official statistics are distributed across central banks in an unbalanced and non-iid manner. Models of this nature enable central banks to overcome these challenges by adopting collaborative training of generative models. Developed approach facilitates the sharing of models themselves rather than the direct sharing of sensitive data. Future directions for this preliminary research include more advanced federated learning methods and diffusion models.

References

Aledhari, M.; Razzak, R.; Parizi, R. M.; and Saeed, F. 2020. Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. IEEE Access, 8: 140699–140725.

Assefa, S. A.; Dervovic, D.; Mahfouz, M.; Tillman, R. E.; Reddy, P.; and Veloso, M. 2020. Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls. In Pro-ceedings of the First ACM International Conference on AI in Finance, 1–8.

Bai, H.; Zheng, R.; Chen, J.; Ma, M.; Li, X.; and Huang, L. 2022. A3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing. In International Confer-ence on Machine Learning, 1399–1411. PMLR.

Barse, E.; Kvarnstrom, H.; and Jonsson, E. 2003. Syn-thesizing test data for fraud detection systems. In 19th Annual Computer Security Applications Conference, 2003. Proceedings., 384–394.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Par-rots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. In Noise reduction in speech processing, 37–40. Springer.

Beutel, D. J.; Topal, T.; Mathur, A.; Qiu, X.; Fernandez-Marques, J.; Gao, Y.; Sani, L.; Li, K. H.; Parcollet, T.; de Gusmäo, P. P. B.; and Lane, N. D. 2022. Flower: A Friendly Federated Learning Research Framework. arXiv:2007.14390.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In International Conference on Learning Representations.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.-A.; and Li, S. Z. 2022. A Survey on Generative Diffusion Model. arXiv preprint arXiv:2209.02646.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2020. Extracting Training Data from Large Language Models. arXiv preprint arXiv:2012.07805.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody Dance Now. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 5933–5942.

Charitou, C.; Dragicevic, S.; and d'Avila Garcez, A. 2021. Synthetic Data Generation for Fraud Detection using GANs. arXiv:2109.12546.

Chen, B.; Zhang, F.; Nguyen, A.; Zan, D.; Lin, Z.; Lou, J.-G.; and Chen, W. 2022. CodeT: Code Generation with Gen-erated Tests. arXiv preprint arXiv:2207.10397.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion Models in Vision: A Survey. IEEE Trans-actions on Pattern Analysis and Machine Intelligence.

de Goede, M. 2023. Training Diffusion Models with Federated Learning: A Communication-Efficient Model for Cross-Silo Federated Image Generation.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems, 34: 8780–8794.

Engelmann, J.; and Lessmann, S. 2021. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174: 114582.

Fredrikson, M.; Jha, S.; and Ristenpart, T. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 1322–1333.

Gao, S.; Zhou, P.; Cheng, M.-M.; and Yan, S. 2023. Masked Diffusion Transformer is a Strong Image Synthesizer. arXiv preprint arXiv:2303.14389.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv preprint arXiv:1412.6572.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. Advances in Neural Information Pro-cessing Systems, 33: 6840–6851.

Hoogeboom, E.; Nielsen, D.; Jaini, P.; Forre´, P.; and Welling, M. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. Advances in Neural Information Processing Systems, 34: 12454–12465.

Jordon, J.; Yoon, J.; and Van Der Schaar, M. 2018. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In International conference on learning representations.

Jothiraj, F. V. S.; and Mashhadi, A. 2023. Phoenix: A Federated Generative Diffusion Model. arXiv preprint arXiv:2306.04098.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2019. Advances and Open Problems in Federated Learning. arXiv preprint arXiv:1912.04977.

Kim, J.; Jeon, J.; Lee, J.; Hyeong, J.; and Park, N. 2021. OCT-GAN: Neural ODE-Based Conditional Tabular GANs. In Proceedings of the Web Conference 2021, 1506–1515.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Kotelnikov, A.; Baranchuk, D.; Rubachev, I.; and Babenko, A. 2022. TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv:2209.15421.

Le, H.; Wang, Y.; Gotmare, A. D.; Savarese, S.; and Hoi, S. C. H. 2022. CoderL: Mastering Code Generation Through Pretrained Models and Deep Reinforcement Learning. Advances in Neural Information Processing Systems, 35: 21314–21328.

Le, M.; Vyas, A.; Shi, B.; Karrer, B.; Sari, L.; Moritz, R.; Williamson, M.; Manohar, V.; Adi, Y.; Mahadeokar, J.; et al. 2023. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale. arXiv preprint arXiv:2306.15687.

Li, Q.; Wen, Z.; Wu, Z.; Hu, S.; Wang, N.; Li, Y.; Liu, X.; and He, B. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering.

Li, Y.; Choi, D.; Chung, J.; Kushman, N.; Schrittwieser, J.; Leblond, R.; Eccles, T.; Keeling, J.; Gimeno, F.; Dal Lago, A.; et al. 2022. Competition-Level Code Generation with AlphaCode. Science, 378(6624): 1092–1097.

Massey, F. J. 1951. The Kolmogorov-Smirnov test for good-ness of fit. Journal of the American Statistical Association, 46(253): 68–78.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Artificial Intelligence and Statistics, 1273–1282. PMLR.

McMahan, B.; and Ramage, D. 2017. Federated Learning: Collaborative Machine Learning Without Centralized Training Data. Google Research Blog, 3.

Nock, R.; and Guillame-Bert, M. 2022. Generative Trees: Adversarial and Copycat. arXiv preprint arXiv:2201.11205.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Paszke, A.; and Gross, S. e. a. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d' Alche´Buc, F.; Fox, E.; and Garnett, R., eds., Advances in Neural Information Processing Systems 32, 8024–8035.

Patki, N.; Wedge, R.; and Veeramachaneni, K. 2016. The Synthetic data vault. In IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399–410.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In IEEE/CVF conference on computer vision and pattern recognition, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234–241. Springer.

Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; and Backes, M. 2018. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Ma-chine Learning Models. arXiv preprint arXiv:1806.01246.

Sattarov, T.; Schreyer, M.; and Borth, D. 2023. FinDiff: Diffusion Models for Financial Tabular Data Generation. arXiv:2309.01472.

Schreyer, M.; Hemati, H.; Borth, D.; and Vasarhelyi, M. A. 2022. Federated Continual Learning to Detect Accounting Anomalies in Financial Auditing. In Workshop on Federated Learning: Recent Advances and New Challenges (in Con-junction with NeurIPS 2022).

Schreyer, M.; Sattarov, T.; and Borth, D. 2022. Federated and Privacy-Preserving Learning of Accounting Data in Financial Statement Audits. In Proceedings of the Third ACM International Conference on AI in Finance, 105–113.

Schreyer, M.; Sattarov, T.; Reimer, B.; and Borth, D. 2019. Adversarial Learning of Deepfakes in Accounting. NeurIPS'19 Workshop on Robust AI in Financial Services.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In 2017 IEEE Symposium on Security and Privacy (SP), 3–18. IEEE.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-Video: Text-to-Video Generation Without Text-Video Data. arXiv preprint arXiv:2209.14792.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. In International confer-ence on machine learning, 2256–2265. PMLR.

Torfi, A.; Fox, E. A.; and Reddy, C. K. 2022. Differentially Private Synthetic Medical Data Generation Using Convolutional GANs. Information Sciences, 586: 485–500.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv preprint arXiv:2307.09288.

Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. 2023. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. arXiv preprint arXiv:2301.02111.

Wen, B.; Cao, Y.; Yang, F.; Subbalakshmi, K.; and Chandramouli, R. 2022. Causal-TGAN: Modeling Tabular Data Using Causally-Aware GAN. In ICLR Workshop on Deep Generative Models for Highly Structured Data.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling tabular data using conditional gan. NeurIPS, 32.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. arXiv preprint arXiv:2209.00796.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked Generative Video Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10459–10469.

Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; and Gao, Y. 2021a. A Survey on Federated Learning. Knowledge-Based Systems, 216: 106775.

Zhang, Y.; Zaidi, N. A.; Zhou, J.; and Li, G. 2021b. GAN-BLR: A Tabular Data Generation Model. In International Conference on Data Mining (ICDM), 181–190. IEEE.

Zhao, Z.; Kunar, A.; Birke, R.; and Chen, L. Y. 2021. CTAB-GAN: Effective Table Data Synthesizing. In Asian Confer-ence on Machine Learning, 97–112. PMLR.

DEUTSCHE BUNDESBANK EUROSYSTEM

University of St.Gallen
Institute of Computer Science

Overcoming Data-Sharing Challenges in Central Banking: Federated Learning of Diffusion Models for Synthetic Data Generation

Timur Sattarov, Marco Schreyer

3rd IFC Workshop on Data Science in Central Banking
„Data Sharing and Data Access“
17-19 October 2023, Rome, Bank of Italy

The views expressed in this presentation are those of the author and do not necessarily represent those of the Bundesbank or the Eurosystem.

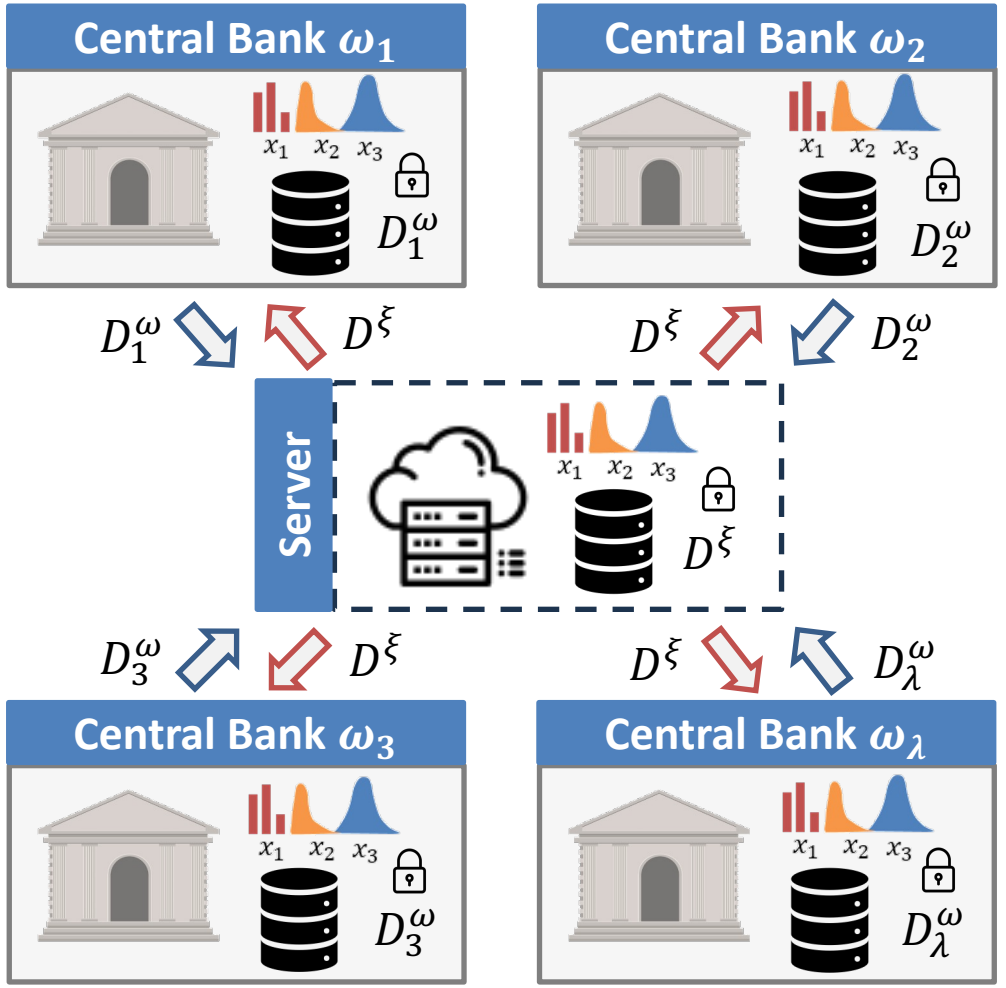
Agenda

- Motivation
- Federated Learning
- Diffusion Models
- FedTabDiff
- Experimental Results
- Conclusion

Motivation

The sharing of financial data between central banks is crucial for managing economic policy and financial sector supervision.
Key advantages:

- **Improved economic policy**: Informed responses to global trends and risks.
- **Boosted financial stability**: Identifies risks and strengthens systems proactively.
- **Better cross-border regulation**: Enables consistent standards and detects misconduct.



Motivation

Challenges:

- **Legal and Regulator Constraints**: varying laws and data protection regulations in different countries can be cumbersome for data sharing.
- **Data Privacy and Security Concerns**: any breaches of confidential information can have severe consequences.
- **Data Transmission**: moving extensive datasets between central banks can become bandwidth-intensive and expensive.



Motivation

Challenges:

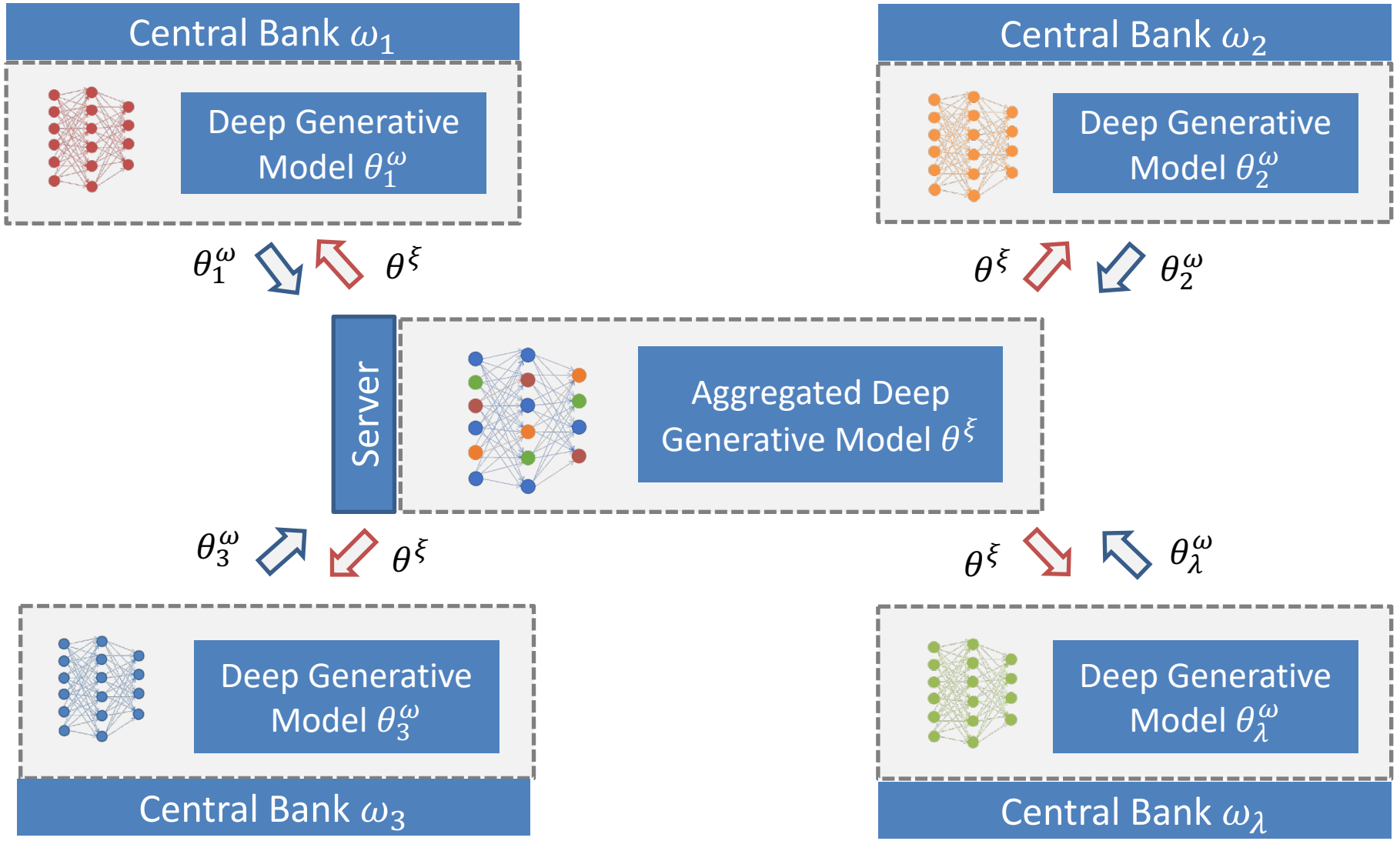
- **Legal and Regulator Constraints**: varying laws and data protection regulations in different countries can be cumbersome for data sharing.
- **Data Privacy and Security Concerns**: any breaches of confidential information can have severe consequences.
- **Data Transmission**: moving extensive datasets between central banks can become bandwidth-intensive and expensive.



Idea: “Federated Learning + Diffusion Models”

- Use **Federated Learning** for decentralized node training without data exchange.
- Utilize **Diffusion Models** to synthesize central bank’s local data; then share the trained synthesizer model as part of the federated training.
- The **global model aggregates** knowledge from central banks to produce high-quality synthetic data.

Data Sharing with Federated Learning



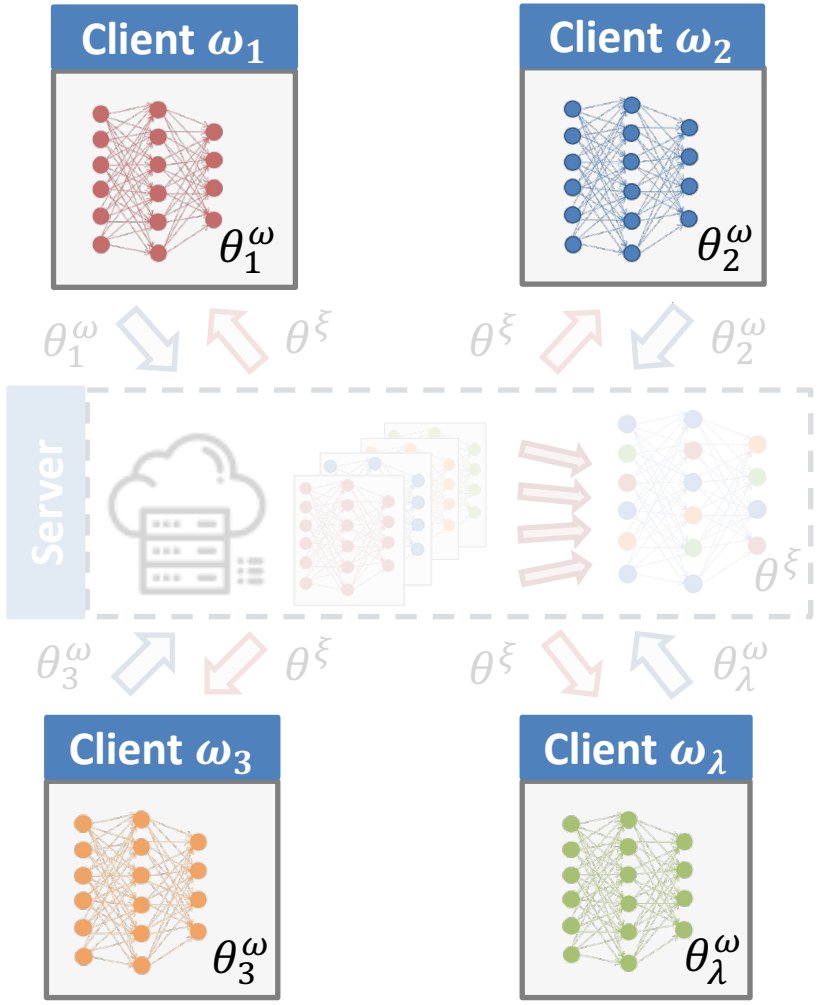
Agenda

- Motivation
- **Federated Learning**
- Diffusion Models
- FedTabDiff
- Experimental Results
- Conclusion

Federated Learning: the mechanics

Main ingredients:

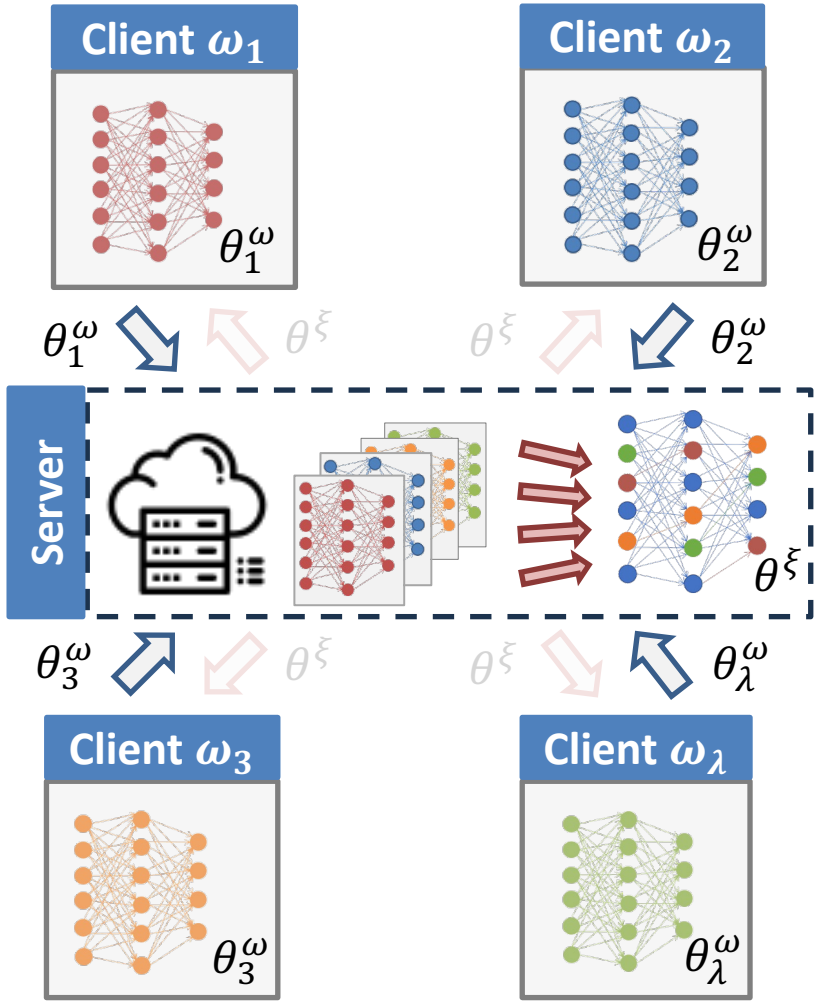
1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.



Federated Learning: the mechanics

Main ingredients:

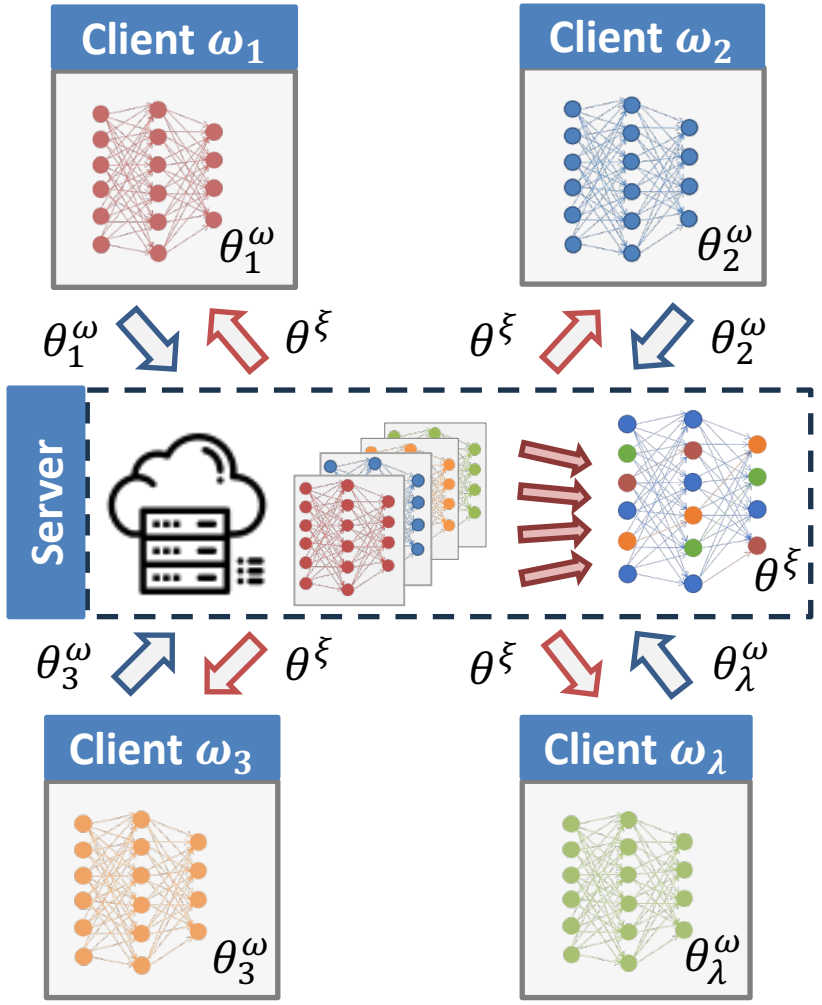
1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.
2. **Model Aggregation at the Server**: after each communication round $r = 1, \dots, R$ only the model parameters are sent to a centralized server, where they are aggregated into a "global" model θ^{ξ}.



Federated Learning: the mechanics

Main ingredients:

1. **Training on the Client**: the initial model θ_i^{ω} is trained locally on each client ω_i using the local data D_i, ensuring data privacy and security.
2. **Model Aggregation at the Server**: after each communication round $r = 1, \dots, R$ only the model parameters are sent to a centralized server, where they are aggregated into a "global" model θ^{ξ}.
3. **Continuous Learning Cycle**: The updated global model is sent back to the clients for further local training, creating a continuous cycle of learning and improvement.



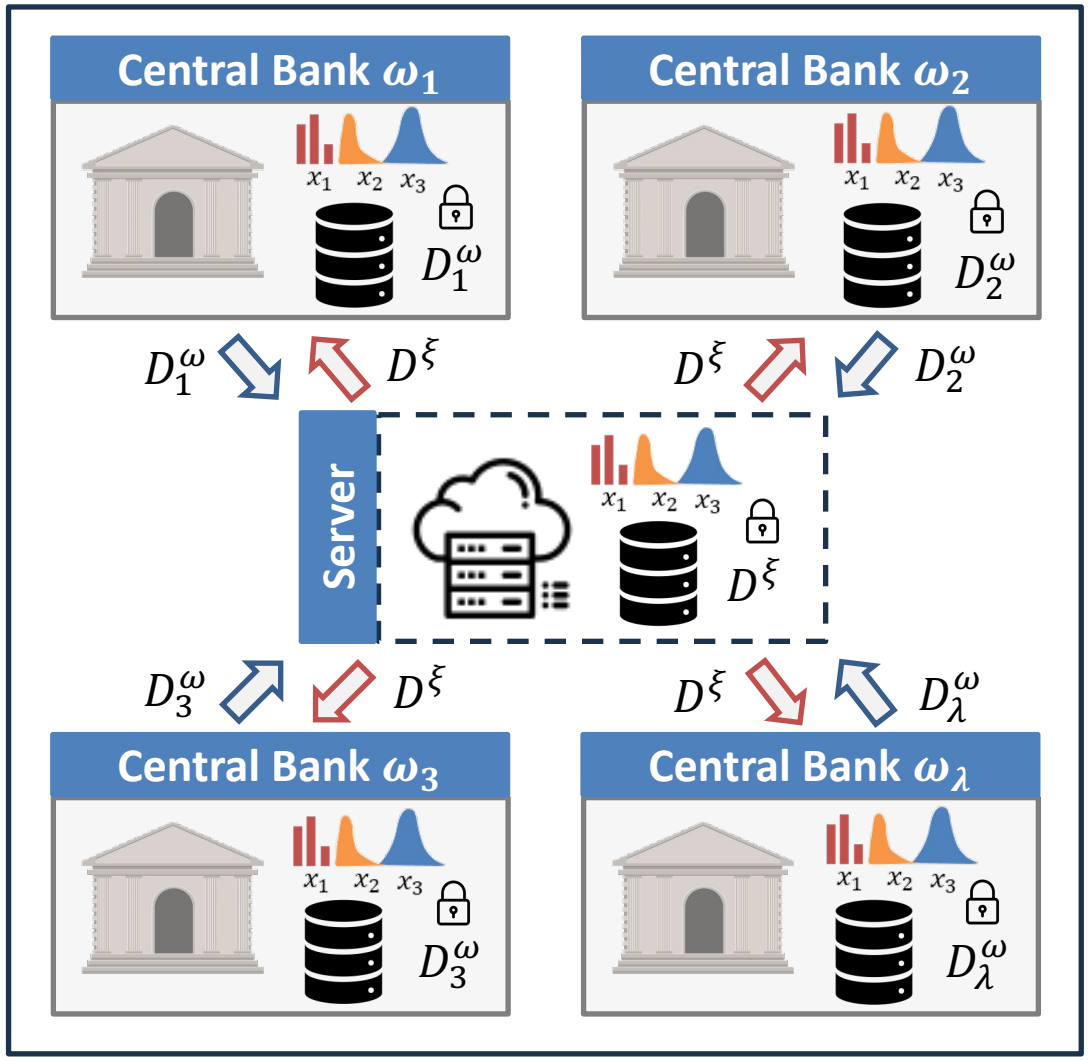
Federated Learning: Benefits

Privacy Preserving: Federated Learning enhances user privacy by keeping all the sensitive data on the local device, never sending raw data to the central server.

Reduced Data Transfer Costs: in the Federated Learning setup only the model parameters are being exchanged therefore avoiding transmission of large data volumes.

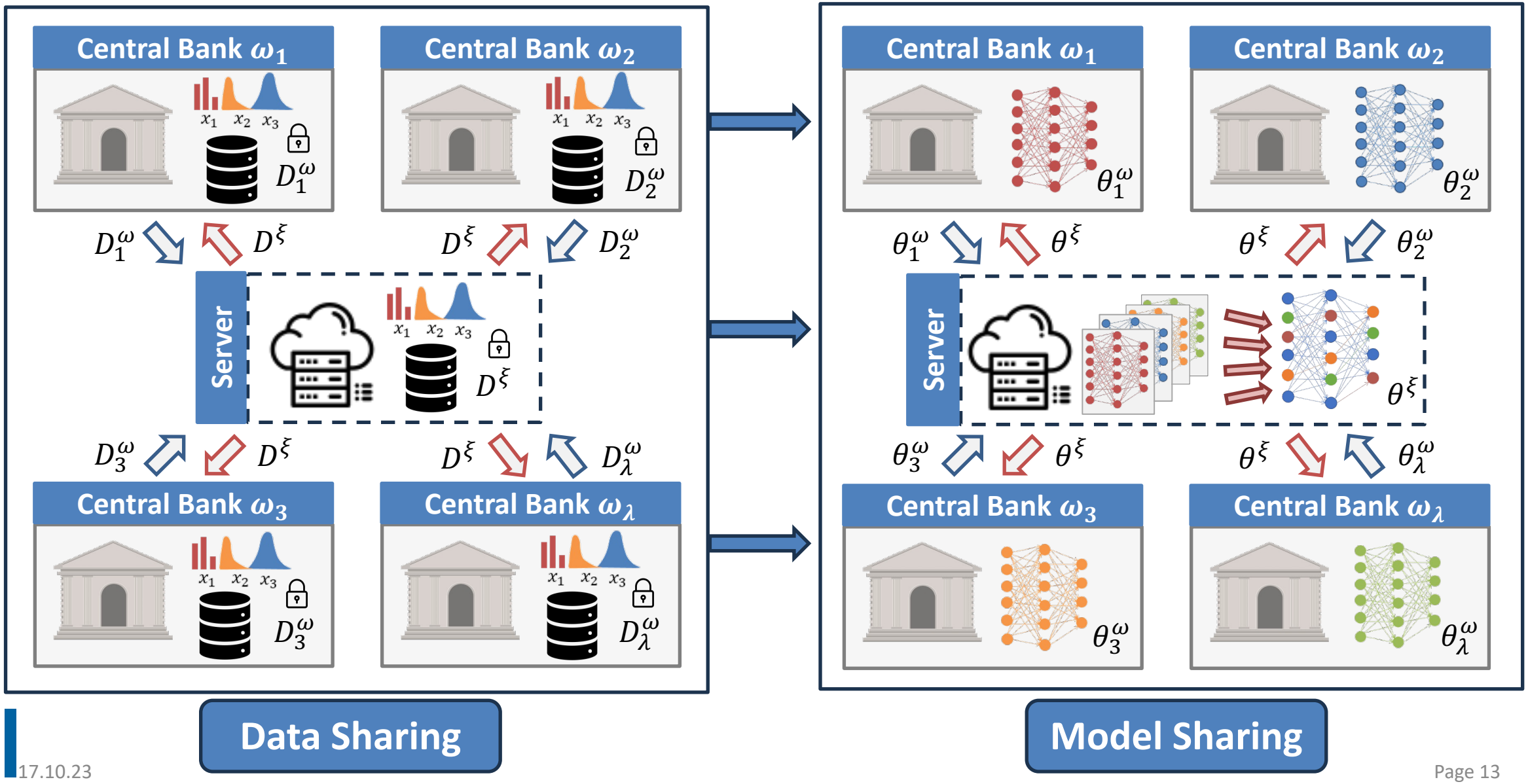
Global Insights: financial institutions can benefit from insights gathered globally across different markets and segments, but applied in a way that is tailored to local market conditions and regulations.

From Data Sharing to Model Sharing

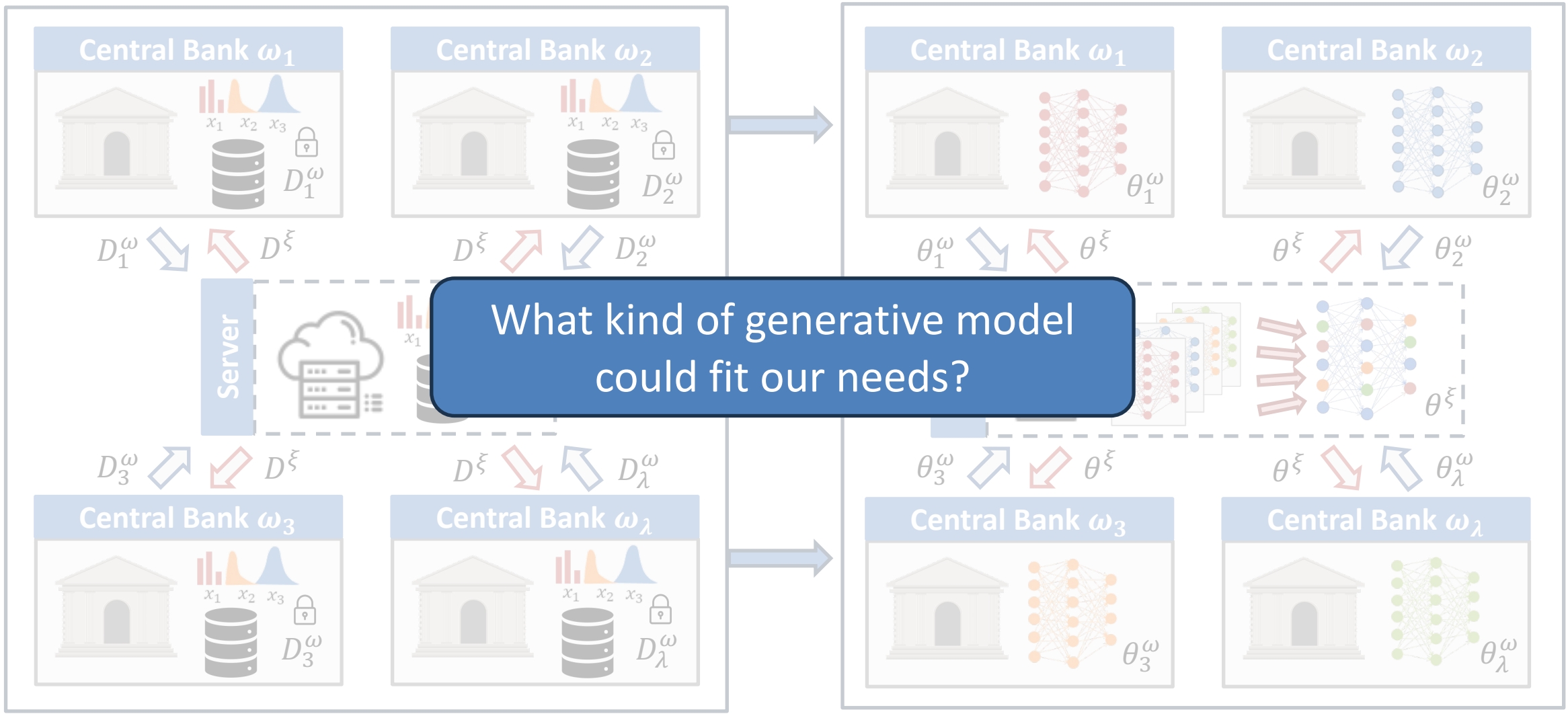


Data Sharing

From Data Sharing to Model Sharing



From Data Sharing to Model Sharing



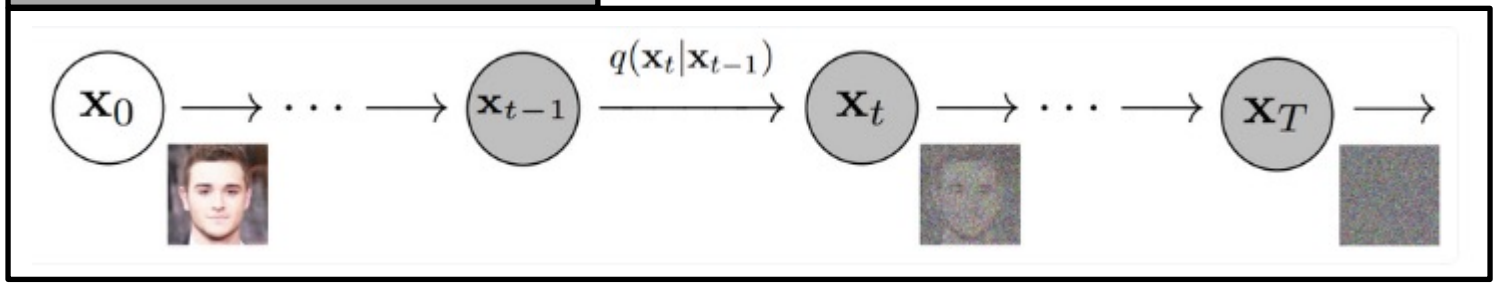
Agenda

- Motivation
- Federated Learning
- **Diffusion Models**
- FedTabDiff
- Experimental Results
- Conclusion

Diffusion Models

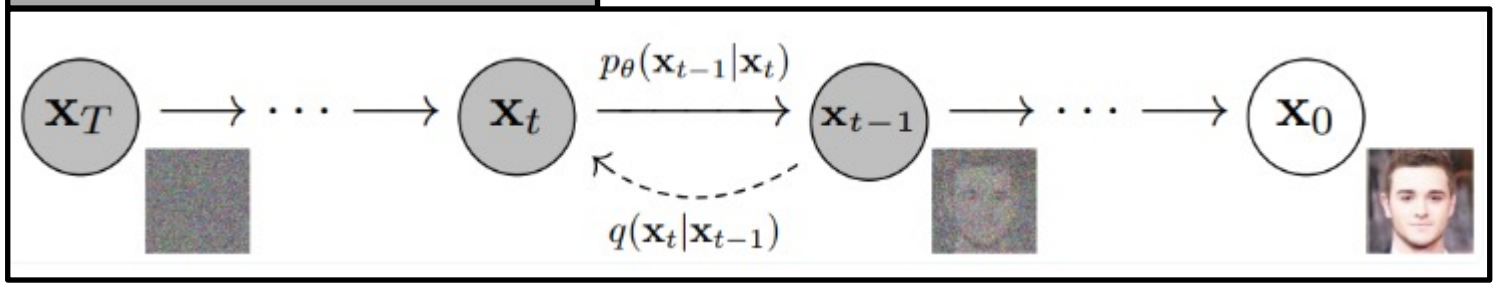
- Diffusion Models – are **generative models** trained with the objective of noise removal and subsequently constructing the desired **data samples from pure noise**.
- First introduced by Jascha Sohl-Dickstein et al. in 2015 with motivation from non-equilibrium thermodynamics. They can be thought as a **sequence of denoising autoencoders**.

Forward Diffusion Process



- Markov chain of diffusion steps.
- Add Gaussian noise in *T* steps.
- When $T \to \infty$, x_t is equivalent to an isotropic Gaussian.
- No learning at this step.

Reverse Diffusion Process



- Reverse process of *T* steps.
- Data generation from Isotropic Gaussian noise.
- Usually is called sampling.
- Learning is required.

Figure is taken from „Denoising Diffusion Probabilistic Models “ by Ho et al. https://arxiv.org/pdf/2006.11239.pdf

Diffusion Models for Financial Tabular Data

FinDiff: Diffusion Models for Financial Tabular Data Generation

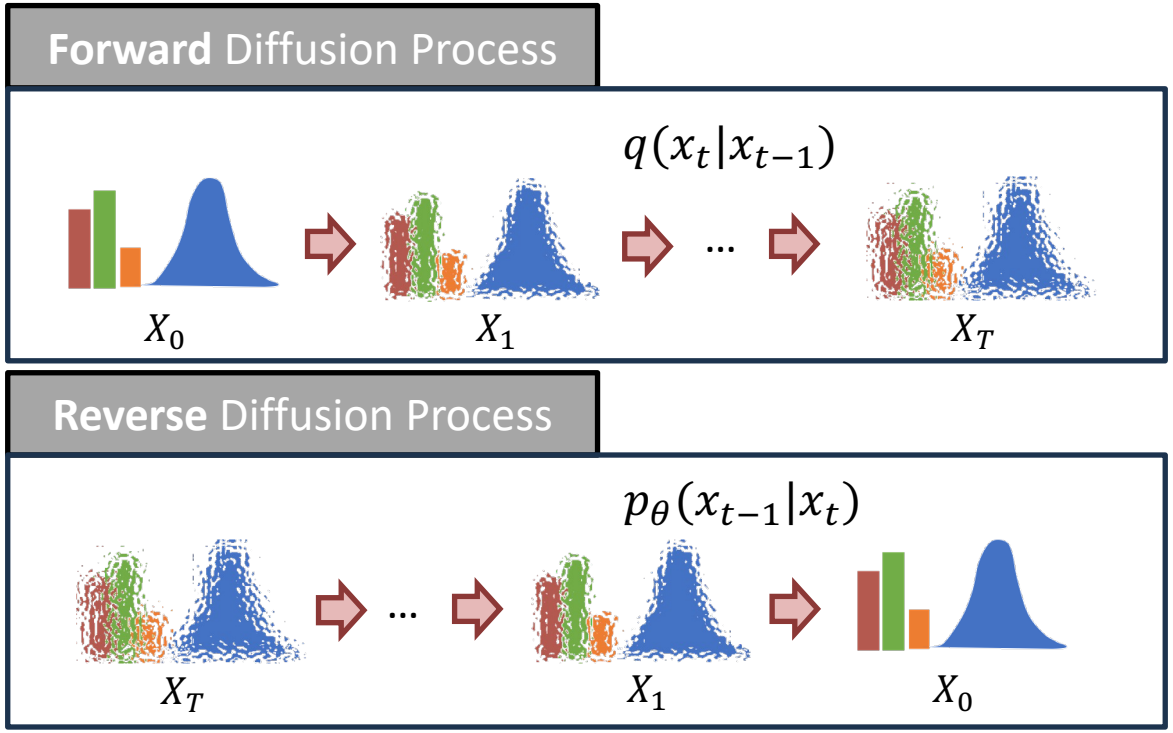
Timur Sattarov[1,2]
timur.sattarov@bundesbank.de

Marco Schreyer[1]
marco.schreyer@unisg.ch

Damian Borth[1]
damian.borth@unisg.ch

[1]University of St. Gallen, St. Gallen, Switzerland

[2]Deutsche Bundesbank, Frankfurt am Main, Germany



https://arxiv.org/abs/2309.01472

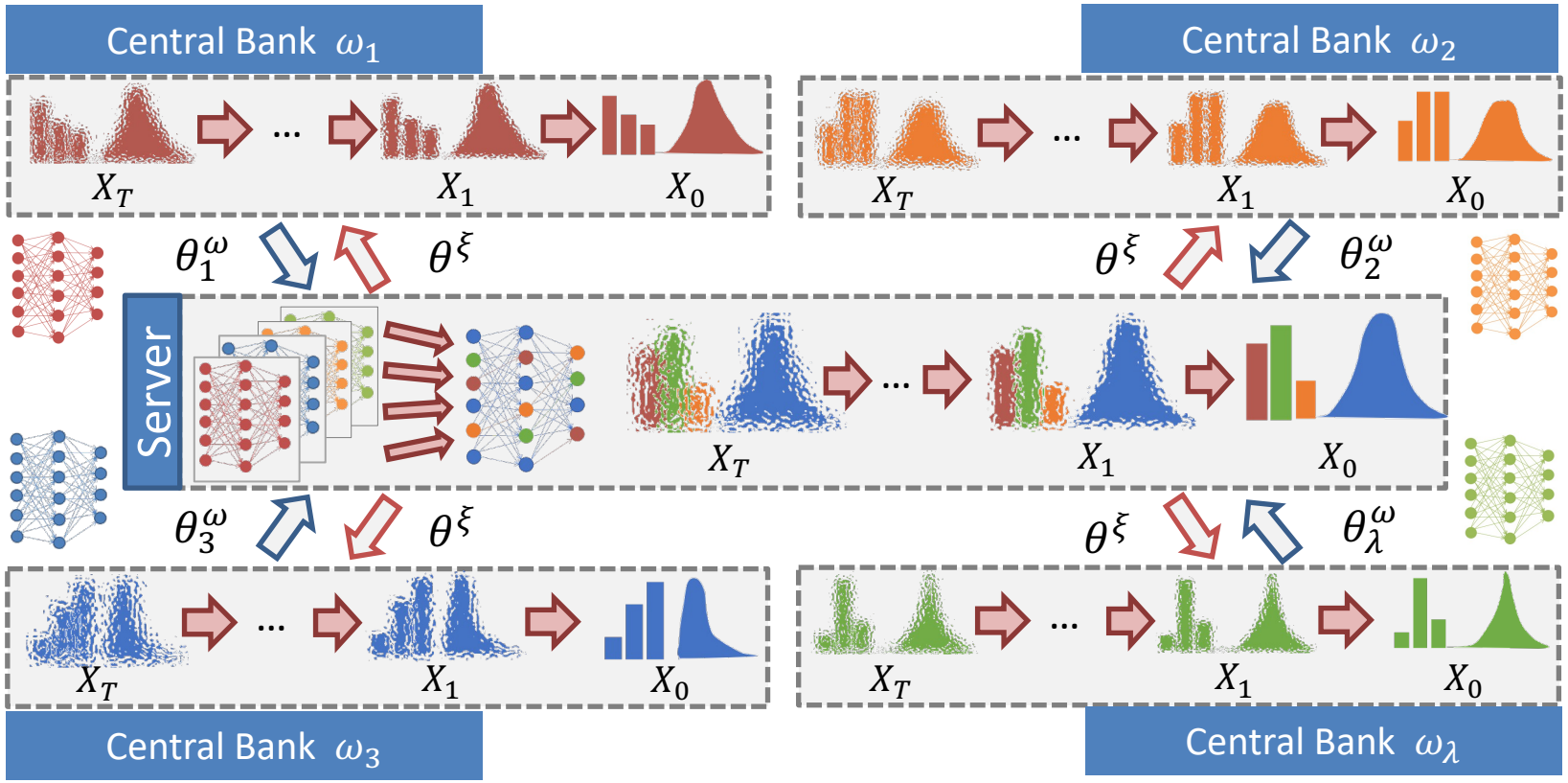
- FinDiff is a diffusion based generative model, that synthesizes financial tabular data for **regulatory downstream tasks**.
- It uses **embeddings** for mixed modality financial data, comprising both **categorical and numeric** attributes.
- Empirical results demonstrate **high fidelity, privacy, and utility** using FinDiff.

Agenda

- Motivation
- Federated Learning
- Diffusion Models
- **FedTabDiff**
- Experimental Results
- Conclusion

FedTabDiff schematic overview

1. Each central bank trains a **local generative model θ_i^{ω}.**
2. The Server aggregates all models into a **global generative model θ^{ξ}** accumulating knowledge from local datasets D_i^{ω}.
3. The global model θ^{ξ} is used to **generate high quality tabular data** without sharing the actual data.



Experimental Setup

- Mixed-type tabular datasets:
  City of Philadelphia Payments – 215,302 payments generated by 58 city departments in 2017. Contains 10 categorical and 1 numeric attributes.
  Diabetes Hospital Data – 92,689 clinical care records collected by 130 US hospitals between 1999-2008. Each record has 40 categorical and 8 numeric attributes.
- The dataset was **non-iid partitioned** $D_i \subseteq D$ across 5 clients ω_i.
- Evaluation **metrics**: fidelity, utility, privacy, and coverage.
- **Federated learning hyperparameters**
  total communication rounds: $R = 1000$
  client model θ_i^{ω} updates: $R_c = 20$
- **Diffusion model hyperparameters:**
  MLP layers: 1024 -> 1024 -> 1024 -> 1024
  activation: leakyRelu
  total diffusion steps: $T = 500$

Dataset	Client	# samples $\mathcal{D}_i$
Philadelphia	ω_1	40,038
	ω_2	28,521
	ω_3	16,831
	ω_4	93,119
	ω_5	36,793
	all	215,302
Diabetes	ω_1	9,685
	ω_2	17,256
	ω_3	22,483
	ω_4	26,068
	ω_5	17,197
	all	92,689

Experimental Results

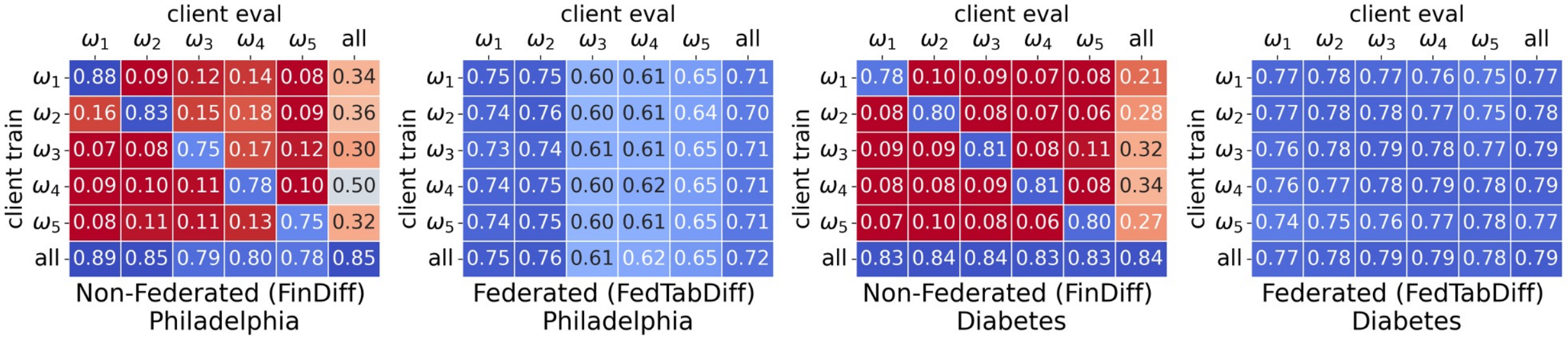
- Comparative analysis of the **Federated** (FedTabDiff) versus **Non-Federated** (FinDiff) models, evaluated using the full dataset D.
- Non-Federated diffusion models are trained individually at each client ω_i with subset $D_i \subseteq D$ (column "Split") and evaluated against the entire dataset D.

			Evaluation Measures			
Dataset	**Client**	**Split $\mathcal{D}_i$**	**Fidelity [8, 32] ↑**	**Utility [50] ↑**	**Coverage [7] ↑**	**Privacy [53] ↓**
Philadelphia	ω_1	19%	0.267 ± 0.03	0.263 ± 0.04	0.689 ± 0.03	3.162 ± 0.19
	ω_2	13%	0.264 ± 0.03	0.325 ± 0.06	0.681 ± 0.02	3.103 ± 0.13
	ω_3	8%	0.207 ± 0.03	0.118 ± 0.04	0.847 ± 0.04	3.178 ± 0.03
	ω_4	43%	0.394 ± 0.01	0.430 ± 0.01	0.863 ± 0.02	2.919 ± 0.14
	ω_5	17%	0.238 ± 0.03	0.197 ± 0.03	0.898 ± 0.01	3.359 ± 0.33
	FedTabDiff		$\mathbf{0.590 \pm 0.01}$	$\mathbf{0.837 \pm 0.03}$	$\mathbf{0.944 \pm 0.03}$	$\mathbf{2.607 \pm 0.18}$
Diabetes	ω_1	10%	0.217 ± 0.01	0.104 ± 0.03	0.944 ± 0.02	10.261 ± 0.25
	ω_2	18%	0.269 ± 0.01	0.186 ± 0.01	0.943 ± 0.03	10.091 ± 0.38
	ω_3	24%	0.314 ± 0.01	0.242 ± 0.01	$\mathbf{0.946 \pm 0.01}$	9.895 ± 0.31
	ω_4	28%	0.331 ± 0.01	0.281 ± 0.01	0.939 ± 0.01	9.941 ± 0.21
	ω_5	18%	0.269 ± 0.01	0.185 ± 0.01	0.943 ± 0.02	10.139 ± 0.19
	FedTabDiff		$\mathbf{0.720 \pm 0.01}$	$\mathbf{0.265 \pm 0.01}$	0.906 ± 0.01	$\mathbf{3.120 \pm 0.09}$

*Scores are derived from the averaged results and standard deviations of five experiments, each initiated with distinct random seeds

Experimental Results

- Fidelity - **similarity of every column** in the synthetic dataset against the real dataset.
- Fidelity score comparison between **Federated** (FedTabDiff) and **Non-Federated** (FinDiff).
- In the Non-Federated model, each client is trained on its data subset $D_i \subseteq D$ and evaluated across all subsets.



Conclusion and Future Work

- Through the adoption of federated learning methodologies central banks may transition **from data sharing to model sharing.**
- FedTabDiff is a **federated diffusion-based generative model** for high-fidelity synthesis of mixed-type tabular data.
- The model **avoids sharing of sensitive information** by training a generative model and sharing it across different authorities without distributing the underlying data.
- Generated tabular data can be used for a variety of **downstream tasks,** such as regulatory compliance, anti-money laundering, fraud detection, risk management and many others.
- Future trajectories: advancement of **privacy-preserving** techniques, mitigation of **information dissemination** risks and evaluation on the **proprietary regulatory financial statistics**.

Thank you

Timur Sattarov
Data Service Centre
Deutsche Bundesbank
Frankfurt am Main, Germany
timur.sattarov@bundesbank.de

Marco Schreyer
School of Computer Science
University of St.Gallen
St.Gallen, Switzerland
marco.schreyer@unisg.ch

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Exploring aggregation strategies for federated learning in national statistics[1]

Mauro Bruno, Erika Cerasti, Massimo De Cubellis and Francesco Pugliese,
Italian National Institute of Statistics - ISTAT

Rafik Chemli, Benjamin Santos and Julian Templeton,
Statistics Canada

Matjaz Jug,
Statistics Netherlands

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Exploring Aggregation Strategies for Federated Learning in National Statistics

Authors: Benjamin Santos, Julian Templeton and Rafik Chemli (Statistics Canada), Erika Cerasti, Francesco Pugliese, Mauro Bruno and Massimo De Cubellis (ISTAT), Matjaz Jug (Statistics Netherlands)

ABSTRACT

Federated Learning (FL) can open opportunities for analytics to be generated on sensitive, distributed data. National Statistical Offices (NSOs) can hold a wealth of information which others could use to generate better statistics. By considering adopting FL, NSOs can explore the feasibility of collaborating with other NSOs to use Machine Learning (ML) models as a method to generate analytics from their distributed data sources. Within an FL environment, the aggregation strategy used by the ML model holder is a key component in ensuring that the ML model is accurately updated after each training round. This work explores a weighted Federated Averaging aggregation scheme which aims to highlight the potential benefits of positively weighing clients who contain more homogeneous data. The preliminary results exhibit that when clients with more homogeneous data are weighed higher, the global model can achieve better performance, but that a sufficiently complex FL task is needed for these effects to be more prevalent. Clients with more heterogeneous data exhibit generally worse performance in the simulations conducted. Furthermore, Homomorphic Encryption can be used as another defense within FL scenarios at high computational and memory costs. While shallow ML models can mitigate this issue to an extent, their performance can also degrade depending on the number of operations performed. Future studies will highlight the effects of data homogeneity within FL experiments on more complex and realistic environments.

Keywords: Federated Learning, Homomorphic Encryption, Privacy Enhancing Technologies

Introduction

In an era increasingly reliant on data, official statistics plays a pivotal role in the fabric of modern society. Published statistics provide essential insights into various aspects of life, including economic performance, population demographics, health trends, and environmental changes. The integrity, accuracy, and reliability of these statistics are crucial for maintaining public trust and enabling governments, businesses, and individuals to make evidence-based decisions that can impact society itself. The advent of big data has significantly impacted official statistics and National Statistical Offices (NSOs). The sheer volume, variety, and velocity of data being generated presents both unprecedented opportunities and formidable

challenges. Big data offers a richer, more granular view of complex phenomena, potentially enhancing the depth and breadth of statistical analysis. However, this paradigm shift requires a re-evaluation and adaption of traditional statistical methodologies. The conventional approaches to data collection and analysis are being tested by the complexities inherent in big data, including issues related to data quality such as relevance, timeliness, and privacy protection.

Amongst these challenges, new methodologies, such as input privacy preservation, have become increasingly important. Input privacy refers to techniques and practices designed to protect the confidentiality and integrity of data sources, particularly when dealing with privately held information. As statistical agencies increasingly turn to privately held data to complement or enhance traditional data sources, the need to safeguard individual and organizational privacy becomes imperative. Input privacy approaches involve a range of techniques, from data anonymization and encryption to more sophisticated approaches such as Homomorphic Encryption (HE) and Secure Multi-Party Computations [1]. NSOs also use output privacy techniques such as differential privacy and statistical disclosure control. The proper adoption of these methodologies, following extensive research of their utility, is essential in ensuring that the analysis of privately held data is conducted responsibly, ethically, and effectively.

Although NSOs would benefit from holding more information from other parties, including other NSOs, this data may be private or legally protected, and its direct usage may be a point of conflict. Although legislation can exist to permit NSOs access to private datasets, avoiding forceful access can help further build trust when possible. NSOs themselves also have a duty to protect data, where laws can enforce this. To get insights from data shared with other organizations, Federated Learning (FL) is a potential solution since it trains models with an organization's data without sharing the data itself. Within the United Nations Privacy Enhancing Technologies (PET) Lab, ISTAT, Statistics Canada, and Statistics Netherlands have been researching the capabilities of FL. Within this work, we have explored different federated aggregation strategies, including weighted strategies which positively or negatively emphasize learning from homogeneous data sources.

This paper will first outline the various techniques being tested. This includes an overview of FL and HE, which can further enhance privacy when used with FL. The experiment methodology is then described, with the key results presented and discussed. Finally, future work this group can explore within the UN PET Lab is outlined.

Background Information

Understanding Federated Learning

FL is a Machine Learning (ML) methodology which enables ML models to be trained across distributed devices while keeping the training data stored locally on each device [2, 3]. Using FL, an organization can train a ML model held by some central authority without requiring the training data to be shared with the authority. This permits analytics to be derived from distributed private data sources. During the training process, the global model held by the central authority is sent to all clients

(individuals or organizations) participating in training that model for the training round. The locally updated models are then sent to and aggregated by the central authority who stores the updated global model for future use or further training. Thus, the training process ensures that the private client data does not leave that client. The central authority's role consists of aggregating the updates from the clients to potentially improve the global model's performance [4]. If the central authority has an isolated testing set, the updated model can be tested before committing it as the main model. The updated models can also be version controlled to ensure that data poisoning attacks or concept drifts are appropriately handled.

The performance of the distributed trained model relies on the careful selection of its hyperparameters, the quality and quantity of the distributed training data, the number of clients participating in training, the ability to evaluate the model updates, and the aggregation strategy utilized to combine local models from each training round. There are a variety of aggregation strategies which can be used in FL, which each vary in complexity and utility. Some popular strategies include Federated Averaging (FedAvg) [5], Federated Adaptive Gradient (FedAdagrad) [6], Federated Adam (FedAdam) [7], and Federated Yogi (FedYogi) [6]. Each of these have pros and cons depending on the type of data being held by the clients. Within this paper we will explore strategies to weight FedAvg, which simply computes the average of each local model's weights, by assigning weights to each local model. This can allow models to be weighed higher based on different heuristics. We consider two scenarios:

1. Where models trained with more homogeneous data are weighed higher.
2. Where models are weighed higher when trained with more heterogeneous data.

The purpose of exploring these two opposite strategies is to understand the impact of prioritizing clients with more homogeneous data during the training rounds. If a client lacks homogeneous data in a round, it may be worth excluding the client for the round or using a subset of the data available for training. These two weighing schemes will be respectively referenced as HWFedAvg (homogeneously weighted FedAvg) and NWFedAvg (non-homogeneously weighted FedAvg). Full details on these schemes will be described later in the paper.

Although FL permits analytics previously impossible to be generated, not all privacy risks are mitigated [8]. In fact, locally trained client models can still be attacked if not appropriately protected. This can lead to privacy leaks such as the training set being reconstructed by the central authority. Fortunately, other PETs can be used alongside FL to further protect local models. Homomorphic Encryption (HE) is a PET that can provide additional protection at the cost of the total compute and memory needed [9].

Understanding Homomorphic Encryption

HE is a public-key encryption scheme which enables certain arithmetic functions to be calculated between encrypted data and another data source without needing to decrypt the encrypted data [10]. The outputs of these functions are also encrypted, where the plaintext is the correct result of the applied function. There are different types of HE algorithms, each offering varying degrees of functionality and security

(at the cost of the corresponding complexity). Partially HE supports either addition or multiplication operations, but not both, to be applied with the encrypted data [11]. Somewhat HE supports a limited number of both addition and multiplication operations on encrypted data, with restrictions on the number of operations that can be performed [12]. Fully HE is the most robust form of HE which supports both addition and multiplication without the previously mentioned restrictions [13].

While we will explore HE in an FL scenario, it has many important uses. For example, it can securely perform computations on encrypted data in delegated computing scenarios (such as if companies need to send data to a competitor's cloud platform for aggregation by an NSO) [14]. Within Secure Multi-Party Computations, HE can enable multiple parties to jointly compute a function over their encrypted inputs [15]. Furthermore, it can allow privacy-preserving data analysis to be performed on encrypted data, which is important in sensitive domains such as healthcare. Other encryption approaches can help with generating data analytics, such as Functional Encryption, however within FL, HE best enhances a local model's privacy when sent for aggregation since its outputs remain encrypted. With HE, the central authority can send an encrypted model to be trained, then securely aggregate the encrypted local model weights to update the model. This ensures that the aggregation step remains secure for all clients. The corresponding memory and computational costs are high when used, but if feasible, the resulting analytics can be worth it.

Importance of Considering Data Homogeneity

A key challenge in FL is that client data cannot be observed directly by the central authority. Thus, there can be quality issues with the data being used for training. This can lead to data poisoning attacks, which attempt to reduce the model's performance. Within this paper we consider the scenarios in which client data is more or less homogeneous. When data is heterogeneously distributed, local models tend to diverge from the optimal global model. In these cases, by aggregating divergent local models, the performance of the global model may degrade. FedAvg does not address the issue of data heterogeneity when aggregating local models. To mitigate the negative impact of data heterogeneity and promote the convergence of the models, several algorithms have been proposed. These can be classified as either *client variance reduction* or *adaptive global model update* [16]. The former aim to improve client-side learning and use a naïve global updating strategy, while the latter focuses on improving server-side learning using a simple Stochastic Gradient Descent-based client learning strategy. This research includes FedYogi in its tests as an adaptive global model algorithm. These types of algorithms differ from FedAvg since they consider the previous training rounds before aggregating the models.

Experiment Methodology

Previously, we compared various aggregation strategies on a public Human Activity Recognition (HAR) dataset in a homogeneous data environment [17]. That work highlighted that FedAvg and FedYogi are effective federated aggregation strategies when applied to a task without hyperparameter tuning. In this work, we explore a weighted aggregation strategy, termed HWFedAvg, which weighs client models

based on the homogeneity of their training data. Furthermore, we analyze the effectiveness of positively weighing client models when trained with heterogeneous datasets with the NWFedAvg strategy. The base simulation environment remains the same as the previous work we conducted, where the Flower library is used for the FL components [18].

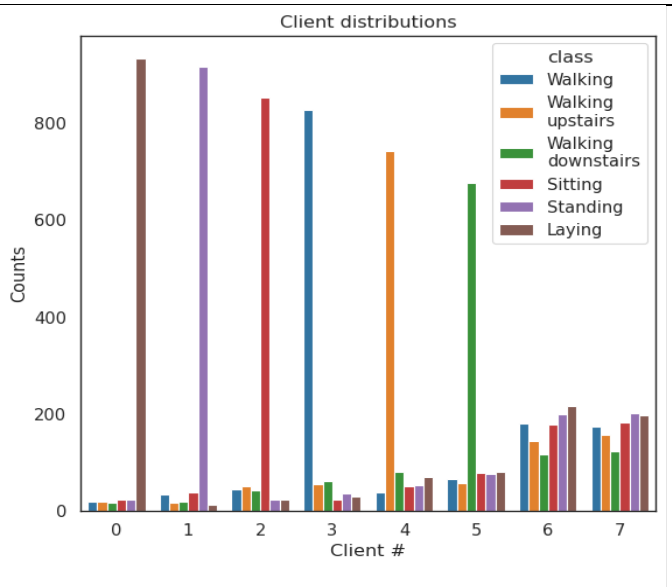
This research utilizes a simplistic HAR dataset comprising of accelerometer and gyroscope smartphone data. The ML model aims to ascertain human activity types from the following six categories: walking, walking upstairs, walking downstairs, sitting, standing, and laying. Within these tests, the HAR dataset has been partitioned among eight clients, aiming to establish a set of clients with a mix of heterogeneous and homogeneous local datasets. We have prepared the following four partitioning methodologies to distribute the data.

- Random: Allocation of samples across clients is randomized.
- Majority Even: The first k clients, where k is the number of classes in the classification task, predominantly holds samples from a unique class. Once all k classes have been uniquely assigned to the first k clients as a majority class, the remaining clients receive a homogeneous distribution of data. Each client holds an equal number of unique records.
- Majority: The same distribution of majority classes as described in Majority Even, however each client can hold a different total number of samples.
- Pick Two: The first $k/2$ clients each hold two unique majority classes (out of k total classes), representing heterogeneous datasets. The remaining clients hold homogeneous datasets with the remaining records. Each client holds an equal number of total records.

The experiments within this paper utilize eight clients, each receiving data using the 'Majority Even' method. It is important to note that each unique class is assigned as a majority class to only one client, where any remaining clients receives a balanced distribution of all classes. A visual representation of the data partitions from the Majority Even method is depicted in Figure 1 below. Note that clients zero through five each have one majority class and their datasets are heterogeneous, whereas clients six and seven have homogeneous datasets.

Human activity class distribution among clients using the 'Majority Even' splitting method

Figure 1



Two different sets of experiments will be presented in this paper. The first will be a comparison between FedAvg, HWFedAvg, NWFedAvg, and FedYogi. The second will be an exploration of different performance costs which can arise when applying HE. Specific details on these experiments are outlined below.

Experiment 1 – Understanding the Importance of Data Homogeneity

This experiment aims to understand the impact of clients holding homogeneous datasets and heterogeneous datasets when participating in training rounds. After utilizing the 'Majority Even' splitting algorithm to define the client datasets, two aggregation strategies are tested. The first gives a greater weight to clients with a homogeneous distribution, HWFedAvg (i.e. clients 6 and 7 in Figure 1), and the second prioritizes clients with a heterogenous distribution, NWFedAvg (i.e. clients from 1 to 5 in Figure 1). The details of the HWFedAvg and NWFedAvg aggregation strategies will be described below.

First, we will outline the key variables utilized by weighting algorithms. Variable i represents a single client, where there are eight clients within this experiment, whereas variable j denotes a single class, where six classes exist within the HAR dataset. Each client i will have an associated weight w_i, which is computed based on the data they hold at a given training round r. Every class j, for each client i, also contains a probability p_j^i, which represents the probability of client i randomly selecting a sample of class j from their training set. The computation for this probability is presented below.

$$p_j^i = \frac{N_{i,j}}{N_i}$$

Where, $\mathrm{N_{i,j}}$ is the total number of samples held by client i of class j to be used for training and $\mathrm{N_i}$ is the total number of samples held by client i for the training round.

Utilizing each normalized probability $\mathrm{p_j^i}$, each client will derive a locally held probability vector $\overrightarrow{\mathrm{p_i}}$, where $\overrightarrow{\mathrm{p_i}} = \{\mathrm{p_1^i}, \ \ldots, \mathrm{p_k^i}\}$ for the k available classes. Similarly, a global vector $\vec{P} = \{P_1, \ \ldots, P_j\}$ is available to all clients, which contains the probabilities P_j of any class j being selected by any participating client. The equation for P_j described below.

$$P_j = \frac{\mathrm{N_j}}{\mathrm{N}}$$

Where $\mathrm{N} = \sum_i \mathrm{N}_i$ and $\mathrm{N_j}$ is the total number of samples of class j being used in round r.

We then begin constructing a set of weights with the probabilities which have been calculated. Let $\tilde{P}$ be the normalized product of the probabilities of picking a class j in the joint dataset (P_j) times the square probability for a client i to have this class j (p_i^j). This is presented below.

$$\widetilde{P_i} = \frac{\sum_j P_j \cdot \left(p_i^j\right)^2}{\sum_i \sum_j P_j \cdot \left(p_i^j\right)^2}$$

These squared probabilities $\widetilde{P_i}$ add importance to majority classes with respect to the minority classes. The normalized weights can then be computed as the inverse of these values, as outlined below.

$$w_i = \frac{1/\widetilde{P_i}}{\sum_i 1/\widetilde{P_i}}$$

These weights can then be scaled by considering sample size N_i, as follows.

$$\alpha_i = \frac{N_i}{N}$$

The final weights can then be computed through the equation below.

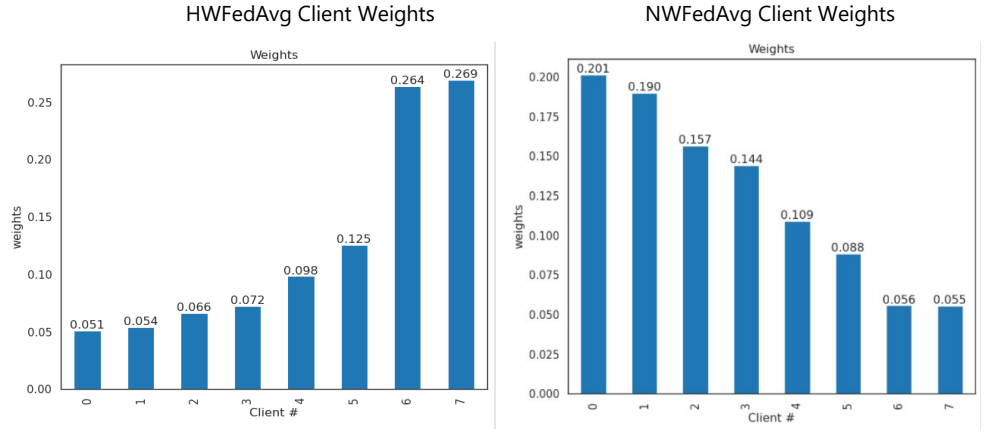
$$W_i = \frac{\alpha_i w_i}{\sum_i \alpha_i w_i}$$

Note that these weights represent adding value to more homogeneous data. Thus, the HWFedAvg strategy uses this equation whereas the NWFedAvg strategy will utilize the inverse of the weights.

With the HWFedAvg and NWFedAvg strategies better understood, examples of how clients weighted using them are presented in Figure 2 below.

Comparison between using HWFedAvg and NWFedAvg

Figure 2



These values are then used to ensure that clients are appropriately weighed on the selected weighting heuristic. We test both homogeneity and heterogeneity to better showcase the importance of data homogeneity in FL scenarios. Full results from tests utilizing these strategies will be presented and discussed later in this paper.

Experiment 2 – Exploring the effects of using Homomorphic Encryption

Previously, we explored using the Paillier public key cryptosystem to securely aggregate the local model weights when using the FedAvg strategy [17]. The Paillier cryptosystem is an additive HE scheme which the central authority can use to average all client model weights when the central authority has access to the public key [19]. Within the Paillier HE scheme, a set of ciphertexts $\{ct_1, ct_2, ..., ct_m\}$ can be added, then decrypted, such that the ciphertext output ct is the encrypted equivalent to adding the corresponding plaintexts $pt = pt_1 + pt_2 + \cdots + pt_m$, where ct_i is the ciphertext of pt_i (i.e., $ct_i = Enc(pt_i)$). Now, we extend secure aggregation to the HWFedAvg and NWFedAvg strategies presented in this paper. To this end, we use the below property of the Paillier cryptosystem.

$$\text{Dec}(ct^k \bmod n^2) = k \times pt \bmod n.$$

This property is used to allow the multiplication of ciphertext $ct_1 = \text{Enc}(pt_1)$ by a plaintext pt_2 and its output decrypts to the multiplication of both plaintexts,

$\text{Dec}(ct_1 \times pt_2) = pt_1 \times pt_2$. With this, we can leave the aggregation weights W_i as plaintext while encrypting the local model weights. Here, each W_i is exposed to the central authority. Alternatively, the CKKS HE scheme can be used to both multiply and add ciphertext [20]. Due to memory constraints when using CKKS and fully encrypting the local models with Paillier HE, this work encrypts only the 2 outermost layers for the two HE scenarios tested. In the first scenario, the local models have four layers: input-561x437, inner1-437x312, inner2-312,6, and output-6. The second scenario uses a shallow model with three layers: input-561x432, inner-432x6, output-6. Both models are derived from tuners and have high performance, where a shallow model is ideal due to the costly time and space complexity brought by using HE.

Experiment Results and Discussions

In this section we outline and discuss the results from running the two experiments detailed in the previous section. First, the impact of a client holding homogeneous data is analyzed by using traditional and weighted aggregation strategies. Next, the experiment results when combining HE with FL are presented.

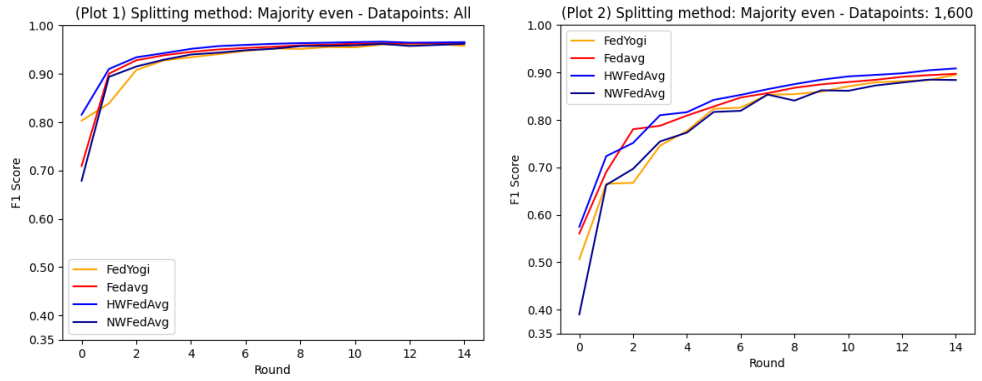
Experiment 1 Results and Discussion

To evaluate the importance of data homogeneity, simulations will be run when using the HWFedAvg, NWFedAvg, FedAvg, and FedYogi aggregation strategies. While other strategies can also be experimented with, these results are preliminary and will be expanded upon with a more complex dataset. None of these strategies are optimized, and further optimizations can positively impact the performance of using a given strategy when applicable. These tests continue the tests from the previous PET Lab work on exploring FL aggregation strategies which indicate that FedAvg and FedYogi will outperform FedAdam and FedAdagrad on this task without proper tuning [17]. The outputs being analyzed are the F1 Scores for the general performance of the central model after each training round.

Since the dataset is insufficiently complex to distinguish performance differences over many rounds, we further test the same FL experiments with only a subset of the data being used. This mimics a scenario in which the distributed clients contain less data a given set of training rounds and helps to better compare the aggregation strategies, and hence better clarifies the importance of considering data homogeneity for distributed clients. Specifically, we first test with all 7,726 records, then test with only 1,600 records. Although different limitations on the dataset size have been tested, reduced or greater sizes did not highlight a difference as strong as when 1,600 is used.

The F1 Scores for both the full and limited dataset tests are presented below in plots 1 and 2 of Figure 3, comparing the four aggregation strategies.

F1 score comparison between the full dataset (7,724 records) and a limited dataset (1,600 records) over all strategies Figure 3



When working with the full dataset, plot 1 of Figure 3 showcases that all strategies can quickly converge in learning the target classification task. Each achieves a very high F1 Score within only two training rounds when using their complete datasets to train the model. Consequently, reducing the number of samples available to each client for training in plot 2 of Figure 3 reduces the rate of learning and increases the time for each strategy to begin achieving higher F1 Scores. This showcases that the original classification task is too simple when using the full dataset and the differences between aggregation strategies becomes more prevalent in more complex scenarios.

There are several key observations which these plots present regarding when data homogeneity is considered. First, both tests highlight that HWFedAvg achieves higher performance in the first round compared to all other strategies whereas NWFedAvg performs the worst in that same round. While the end results slowly begin to converge, this is an important observation which indicates that valuing homogeneous data from clients containing a mix of homogeneous and heterogeneous data distributions can result in quicker overall learning of the target task. While more complex tasks will help further highlight this, this can be important when in a scenario where clients cannot frequently participate in training rounds or where few clients with homogeneous data are able to training rounds. Prioritizing clients with homogeneous data may result in a high-quality central model at a faster rate. Furthermore, if a context drift occurs, this faster learning may help efficiently adapt the model.

Second, we note that strategies which do not consider the heterogeneity problem (i.e. FedAvg) or that attribute greater weight to homogeneous clients (HWFedAvg) provide better results in terms of F1 Score compared to strategies considering the heterogeneity problem (i.e. FedYogi) or which attribute greater weight to heterogeneous datasets (NWFedAvg). This is highlighted by FedAvg and HWFedAvg generally performing better than FedYogi and NWFedAvg in plots 1 and 2 of Figure 3. These findings may not always hold but help exhibit the challenges of working with clients holding heterogeneous datasets in FL scenarios. Figure 3 also showcases that HWFedAvg most closely overlaps with FedAvg, whereas NWFedAvg closely overlaps with FedYogi.

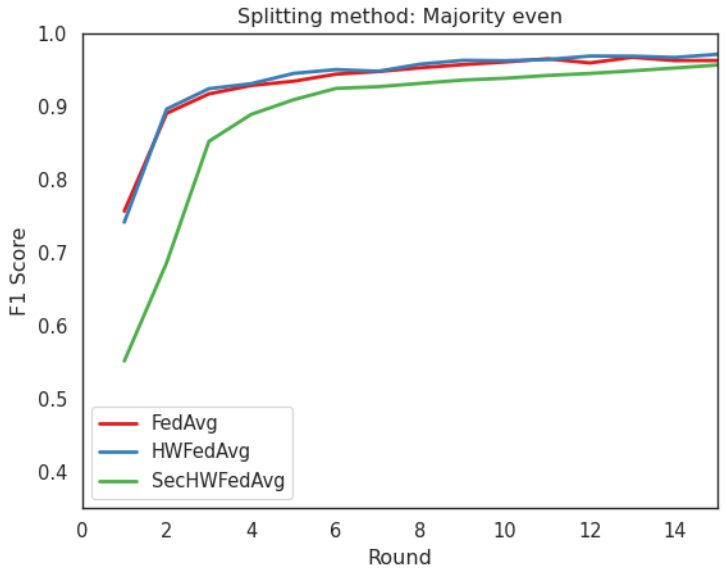
The results suggest that among the analyzed aggregation strategies, those favouring homogeneous dataset training tend to yield the best performance. This indicates that positively weighing clients with homogeneous datasets can result in better overall performance of the model in an FL environment.

Experiment 2 Results and Discussion

Within this experiment we have performed a suite of tests using HE on model architectures of varying complexity, where the full models are not encrypted due to memory constraints. These tests have highlighted that the secure versions of HWFedAvg and NWFedAvg perform well compared to the non-secure counterparts when only the output layer is encrypted. However, as Fig. 4 presents, performance degrades when encrypting more layers (in this scenario the 2 outermost layers are encrypted).

F1 score comparison between FedAvg, HWFedAvg, and the encrypted SecHWFedAvg (outermost two layers are encrypted with Paillier HE)

Figure 4



We hypothesize that this drop in performance is due to increased noise from the additional multiplication operations in the HE process. However, since the F1 Score of the secure aggregated model is approaching the non-secured versions, we intend to further explore this hypothesis in future work. Furthermore, we intend on expanding these tests to fully encrypt the models since both the larger and shallow models could not be fully encrypted due to RAM constraints at the time of testing. Finally, from the tests performed, we argue that in situations with a performance or memory bottleneck, a shallow model is best to use since it can be better secured with far less computational and memory cost than a larger model.

Conclusions and Next Steps

This paper has presented preliminary studies of the effects from applying custom weighting strategies to FedAvg in an FL environment. The simulations indicate that placing value on data homogeneity can result in better overall results whereas prioritizing heterogeneous data can reduce a global model's performance. These preliminary results, favouring a homogeneity-centric approach, warrant further exploration. Future research endeavours will involve more complex classification tasks and utilize benchmarking datasets to substantiate and extend these findings. We expect that in future experiments, the performance of more complex strategies, such as FedYogi, will improve. Furthermore, we hypothesize that a more statistically significant difference will be found when comparing HWFedAvg and NWFedAvg (with HWFedAvg performing better). Within the UN PET Lab, we will also be performing these scenarios in realistic deployments with mock data being used to test the infrastructure requirements needed to consider using FL.

To conclude, this paper has explored weighted FL aggregation strategies, named HWFedAvg and NWFedAvg, with the aim of better understanding the complexity of homogeneous and heterogeneous data scenarios in FL. Simulations have been run to explore the feasibility and utility of the weighting strategies. Paillier HE has also been applied to further test the effects of encrypting model weights within FL scenarios. A performance drop has been observed alongside the increased compute requirements. These tests indicate that data homogeneity is important to consider, but further work needs to be done on more complex tasks to derive more concrete results.

Disclaimer

The content of this article represents the position of the authors and may not necessarily represent that of Statistics Canada, ISTAT, and Statistics Netherlands.

References

[1] United Nations. (2023). *United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*. United Nations Committee of Experts on Big Data and Data Science for Official Statistics, New York. Website: https://unstats.un.org/bigdata

[2] Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., & Dou, D. (2022). From distributed machine learning to federated learning: A survey. *Knowledge and Information Systems*, *64*(4), 885-917.

[3] Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[4] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, *37*(3), 50-60.

[5] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

[6] Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., ... & McMahan, H. B. (2020). Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

[7] Ju, L., Zhang, T., Toor, S., & Hellander, A. (2023). Accelerating Fair Federated Learning: Adaptive Federated Adam. *arXiv preprint arXiv:2301.09357*.

[8] Gosselin, R., Vieu, L., Loukil, F., & Benoit, A. (2022). Privacy and security in federated learning: A survey. *Applied Sciences*, *12*(19), 9901.

[9] Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, *14*(1–2), 1-210.

[10] Gentry, C. (2009). *A fully homomorphic encryption scheme*. Stanford university.

[11] Morris, L. (2013). Analysis of partially and fully homomorphic encryption. *Rochester Institute of Technology*, *10*, 1-5.

[12] Gentry, C., Sahai, A., & Waters, B. (2013). Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In *Advances in Cryptology–CRYPTO 2013: 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part I* (pp. 75-92). Springer Berlin Heidelberg.

[13] Gentry, C. (2009, May). Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing* (pp. 169-178).

[14] Yang, Y., Huang, X., Liu, X., Cheng, H., Weng, J., Luo, X., & Chang, V. (2019). A comprehensive survey on secure outsourced computation and its applications. *IEEE Access*, *7*, 159426-159465.

[15] Acar, A., Aksu, H., Uluagac, A. S., & Conti, M. (2018). A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)*, *51*(4), 1-35.

[16] Nguyen, H., Phan, L., Warrier, H., & Gupta, Y. (2022). Federated Learning for Non-IID Data via Client Variance Reduction and Adaptive Server Update. *arXiv preprint arXiv:2207.08391*.

[17] Santos, B., Templeton, J., Chemli, R., Molladavoudi, S., Pugliese, F., Cerasti, E., ... & Jug, M. (2023, September). Insights into privacy-preserving federated machine learning from the perspective of a national statistical office. In *Conference of European Statistics*.

[18] Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., ... & Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

[19] Paillier, P. (1999, April). Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques* (pp. 223-238). Berlin, Heidelberg: Springer Berlin Heidelberg.

[20] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). Homomorphic encryption for arithmetic of approximate numbers. In *Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23* (pp. 409-437). Springer International Publishing.

3rd IFC Workshop - Federated Learning Data Science in Central Banking

Aggregation strategies for Federated Learning

Erika Cerasti

- Benjamin Santos, Julian Templeton, Rafik Chemli, Saeid Molladavoudi (Statistics Canada)
- **Erika Cerasti**, Francesco Pugliese, Massimo De Cubellis (ISTAT)
- Matjaz Jug (Statistics Netherlands)

17-19 October 2023
Bank of Italy - Rome

International Context

PETs
***privacy-enhancing technologies*:**
Methodologies and approaches
to mitigate privacy risks when using
sensitive or confidential data





NSOs can build greater trust with the public and unlock new opportunities associated with more accurate and complete data collection.

https://unstats.un.org/bigdata/task-teams/privacy/index.cshtml

Are PETs highly reliable and can they be used for accessing sensitive data (health records, tax records or credit card data) by NSOs?

Istat

International Context

UNCEBD : UN Committee of Experts on Big Data and Data Science for Official Statistics

Elements to accelerate the adoption of PETs in the NSO community:

- Experimentation (PET Lab)
- Outreach & Training
- Support Services

UN PET Lab

Created to facilitate experimentation on pilot projects and effective collaboration on "real world" use case.

Objectives: Develop principles, policies and open standards for data sharing, taking full account of data privacy, confidentiality and security issues.

Secure Multi-Party Computation

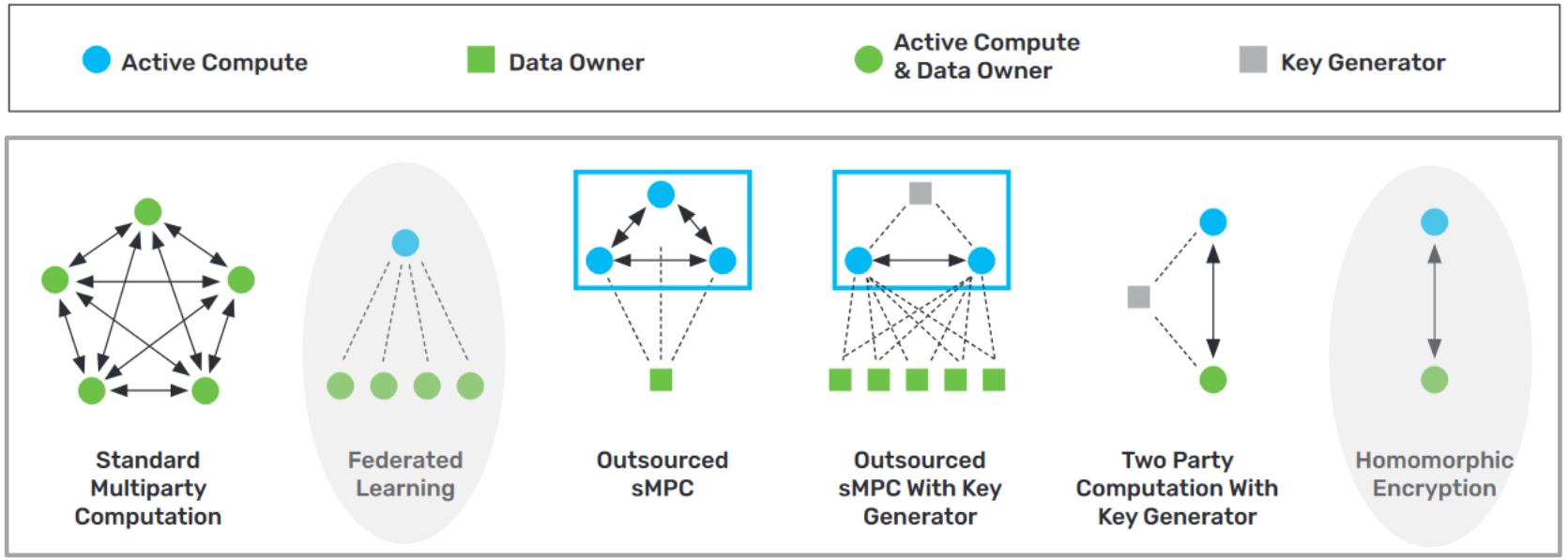
Secure multi-party computation – sMPC

Two or more (mutually distrusting) parties wish to compute an agreed-on function on data that they provide to that computation but are unwilling to disclose to others.

Federated Learning

sMPC protocols use frequent communication among the compute parties.

- available network bandwidth
- network latency between parties



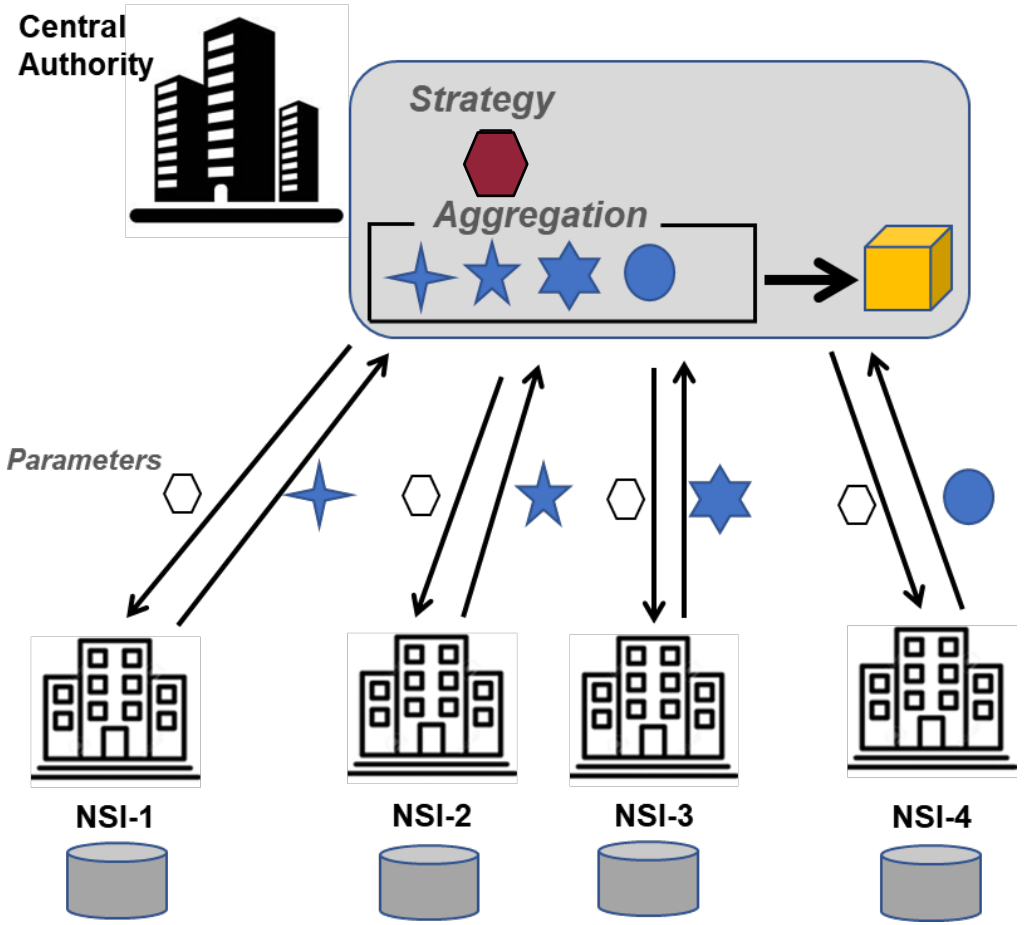
Istat

Federated Learning

Decentralized / Distributed Computation

FL allows a centralized ML model to be trained on data residing on distributed client devices.
After training the model with data locally, a client will send the weights or gradients back to the server to be aggregated.

This allows analytics to be derived without collecting data.



Aggregation strategy

Performance of the trained models can reach similar performance of centralized approaches but a careful selection of the hyperparameters and the aggregation strategy is important.

Istat

Federated Learning and Homomorphic Encryption

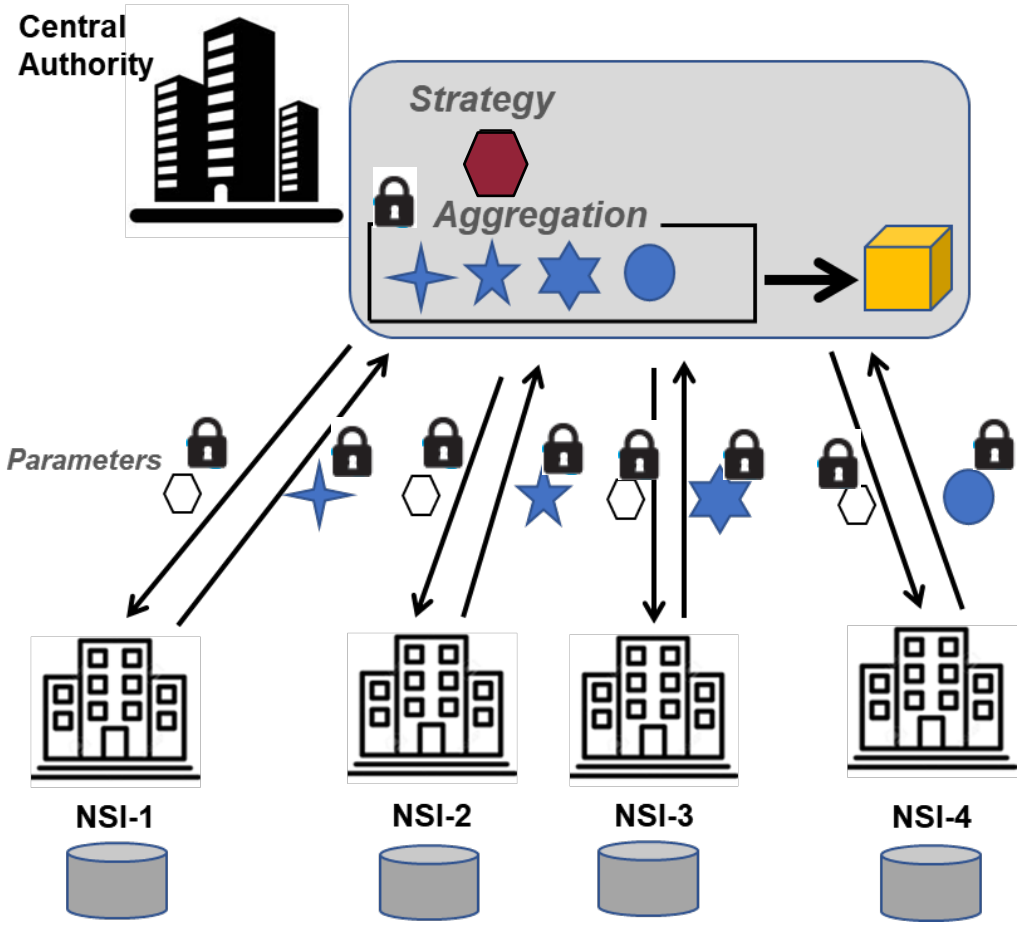
Decentralized / Distributed Computation

Locally trained client models can still be attacked, which can be better protected by using other PETs in conjunction with FL

- Example: Homomorphic Encryption (HE)

HE is a cryptographic technology allowing for the direct computation (addition and multiplication) on encrypted data.

It adds more computational complexity and can result in a high computational overhead.



Aggregation strategies

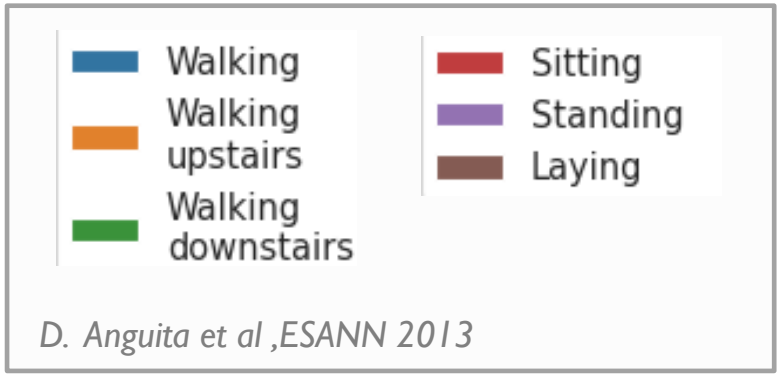
Human Activity Recognition public DATASET

30 volunteers (19-48 years) - 6 class of human activities

Recordings:

Accelerometer and gyroscope data collected by smartphones.
(3-axial linear acceleration and 3-axial angular velocity, rate of 50Hz)

Human activity classes



D. Anguita et al ,ESANN 2013

Work objectives

- Explore different federated aggregation strategies
- Explore the role of dataset **heterogeneity**
- Apply HE to model weights

Federated aggregation Strategies

- FedAvg
- **Weighted FedAvg (WFedAvg)**
- FedAdagrad
- FedAdam
- FedYogi

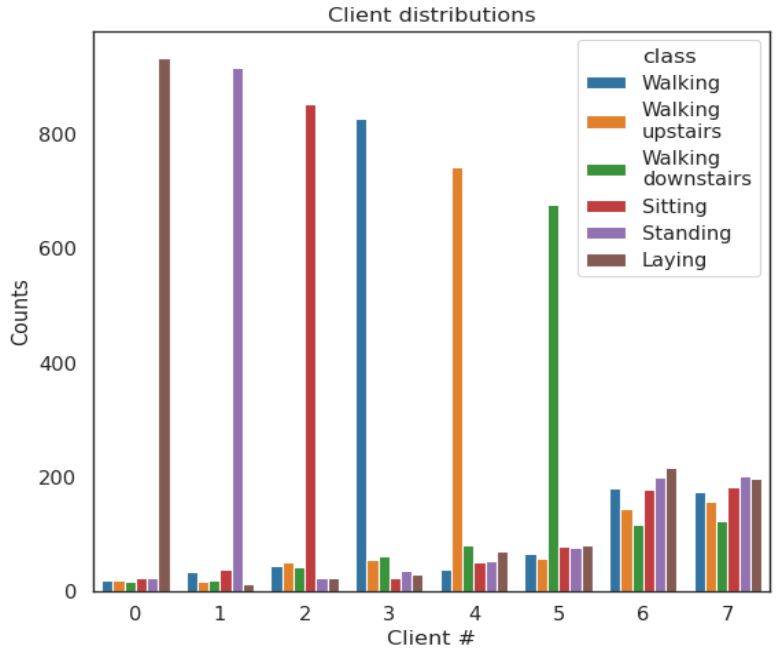
Istat

Data Splitting Methodologies

Splitting methods - heterogeneous datasets – 8 clients

- ***Random***: Samples are randomly distributed among clients
- ***Majority even:*** Each client has one majority class, same number of records
- ***Majority:*** Each client has one majority class, different number of records
- ***Pick two:*** Each client has two majority classes, same number of records

Note that each class can only be assigned as a majority class once, where remaining clients without a majority class are given a distribution of all classes.



Weighted Federated Averaging Strategy

- Index i runs on clients, i.e, 1 to 8 clients
- Index j runs on classes, i.e, 1 to 6 classes
- Properties on i are related to clients: weights w_i, Gini coefficient G_i, entropy H_i
- Properties on j are related to classes: probabilities for picking particular class j for client i.

A vector of probabilities p_i is local for a client i: $\vec{p_i} = p_j^i \rightarrow \{p_1^i, \ldots, p_6^i\}$ with $p_j^i = \frac{\text{count class } j \text{ client } i}{N_i}$ with N_i the total number of samples for client i. p_j^i is normalized.

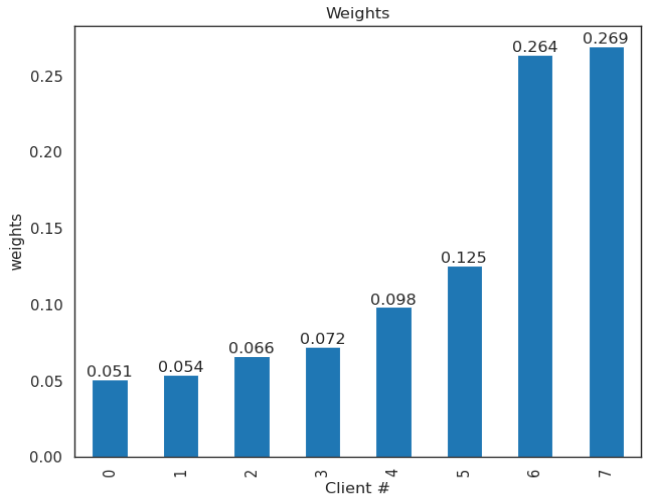
The vector P_j is global (common to all clients) $\vec{P} = P_j \rightarrow \{P_1, \ldots, P_6\}$ with $P_j = \frac{\text{count class j for all clients}}{N}$ with $N = \sum_i N_i$ the total number of samples for all clients. P_j is normalized.

A set of weights could be constructed by multiplying a vector of a local property times a set of local vectors for each client. For instance, let $\tilde{P}$ be the normalized product of the probabilities of picking a class j in the joint dataset times the square probability for a client i to have this class j,

$$\tilde{P_i} = \frac{\sum_j P_j \cdot (p_i^j)^2}{\sum_i \sum_j P_j \cdot (p_i^j)^2}$$

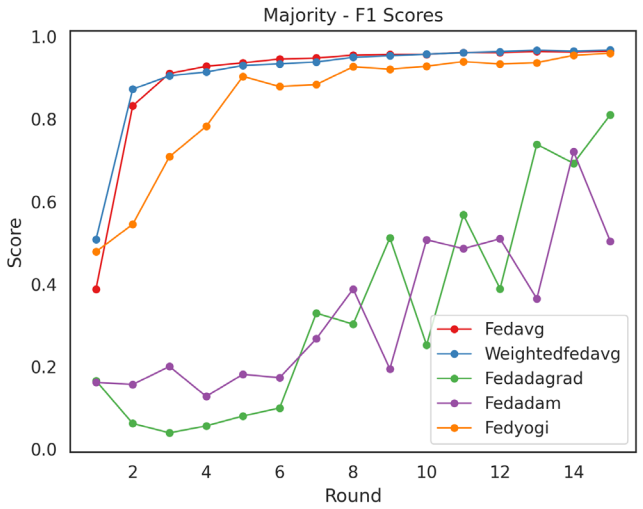
Setting this squared probabilities adds more importance to majority classes respect to the minority classes. Then the normalized weights can be computed as the inverse of this value:

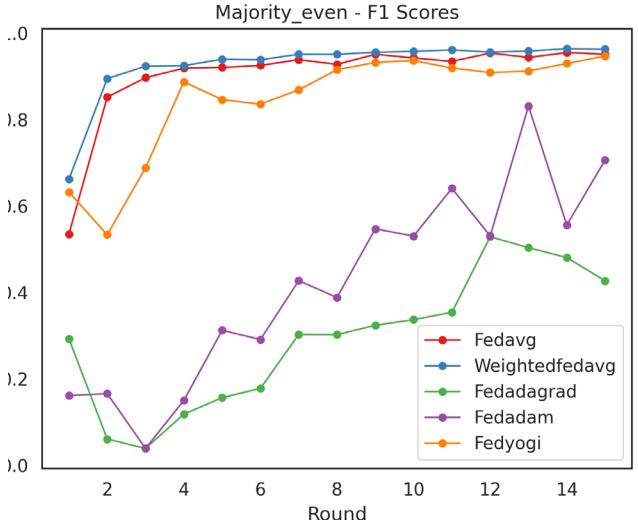
$$w_i = \frac{1/\tilde{P_i}}{\sum_i 1/\tilde{P_i}}$$

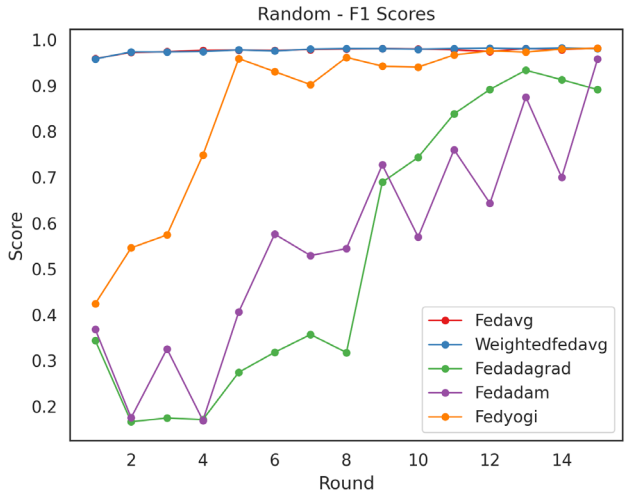


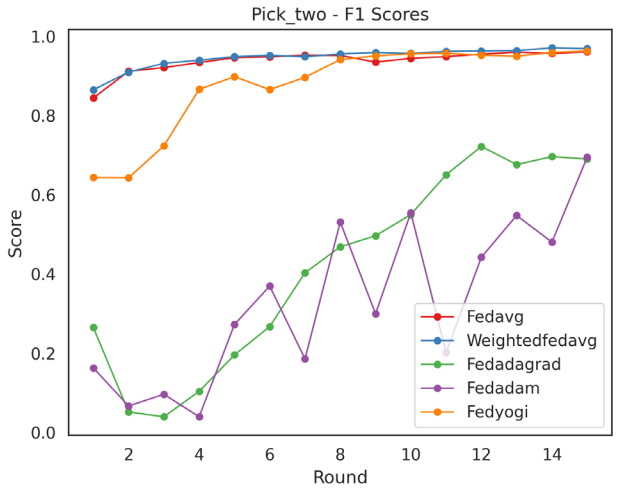
Experiment Results

- FedAvg, WFedAvg and FedYogi perform better than Fedadagrad and Fedadam
  - Fedadam and Fedadagrad would do better if properly optimized
- FedYogi seem to reach convergence a bit later than Fedavg and WFedAvg
- FedAvg and WFedAvg show a very similar behavior (for random split they are expected to be identical)
- The convergence is reached after few rounds

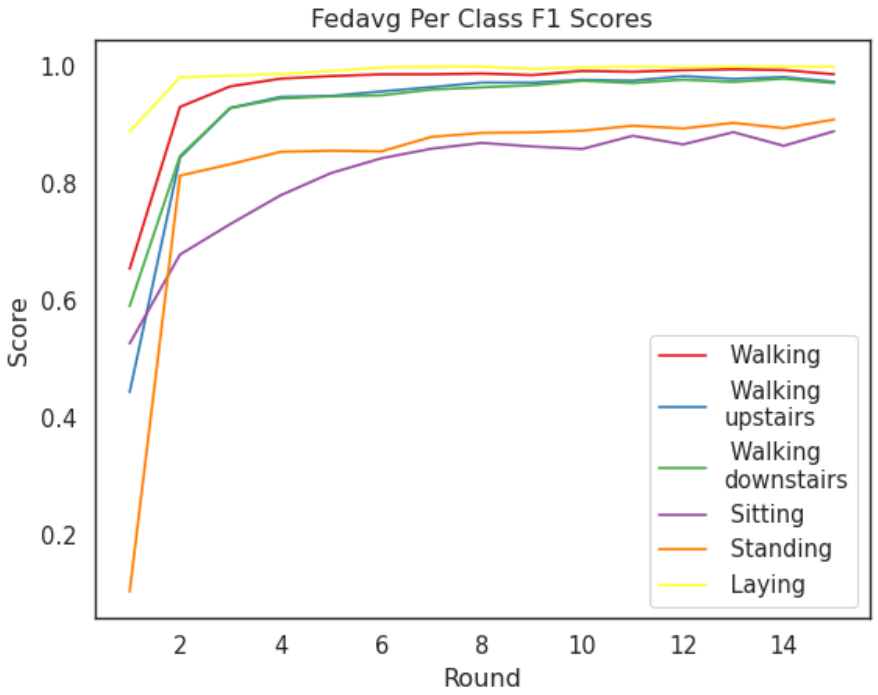


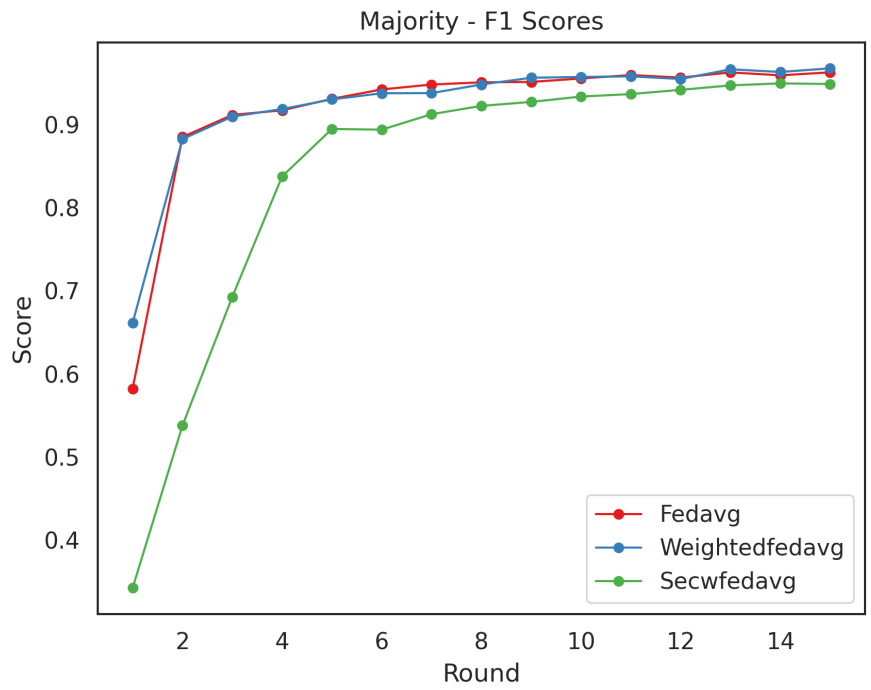






Experiment Results





- FedAvg Performance for each activity class
- For two classes learning is slower

- Performance of FedAvg and WFedAvg compared to FedAvg with Homomorphic Encritption
- HE makes the convergence slower

Istat

Conclusions

- Homomorphic Encryption can add significant time and communication costs, scaling with the amount of encrypted weights/gradients
- Heterogeneity of the training dataset seems to not affect performance in this example:
  - The best models are FedAvg and WFedAvg, which are very similar
  - Adaptive methods (FedYogi) do not show a better performance with heterogenous data
  - It seems that the simpler aggregation technique (FedAvg) work better for this dataset

These results are preliminary

- The task is too easy with the selected dataset. Convergence is reached too quickly
- Next steps are to explore different hyperparameters and use a more complex dataset

Preliminary results indicate faster overall training and faster learning of different classes when using Weighed Federated Averaging (can lead to less communication cost if learning is quicker)

UN PET repository: https://unstats.un.org/wiki/display/UGTTOPPT/Case+study+repository

Istat

Thanks for your attention

For further information please write to *erika.cerasti@istat.it*

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

EMIR data for financial stability analysis and research[1]

Michele Leonardo Bianchi, Bianca Sorvillo, Dario Ruzzi, Federico Apicella, Luigi Abate, Leonardo Del Vecchio, Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

EMIR data for financial stability analysis and research

Bianchi, Michele Leonardo; Sorvillo, Bianca; Ruzzi, Dario; Apicella, Federico; Abate, Luigi; Del Vecchio, Leonardo[1]

Abstract

This work aims to set best practices for the analysis of the novel and extensive European Market Infrastructure Regulation (EMIR) dataset. By comprising trade-level information on millions of derivative transactions every day, this dataset is particularly useful to supervisors, regulators and central banks for monitoring financial markets by means of real-time risk metrics. Data granularity, however, comes at the cost of high complexity compounded by the dual-reporting nature of EMIR (as both 'legs' of a transaction may be subject to reporting obligations). After highlighting the data quality and practical issues faced when collecting and using EMIR data, we propose a framework to move from raw to clean data. Moreover, we provide statistics on notional values, based on which we visualise interconnections among counterparties.

Keywords: EMIR, granular data, derivatives, financial stability

JEL classification: G18, G00

[1] Financial Stability Directorate, Bank of Italy. The authors thank G. Papadia, D. Zambuto and G. Turturiello for technical support. The views expressed are those of the authors only and do not necessarily reflect those of the Banca d'Italia.

Contents

EMIR data for financial stability analysis and research ... 1
Summary ... 3
1. Introduction ... 5
1.2 European Market Infrastructure Regulation (EMIR) ... 5
1.3 Overview of existing research work on EMIR data ... 7
2. Data collection process through trade repositories ... 8
3. EMIR Data ... 10
3.1 Data overview ... 10
3.3 Data processing framework ... 15
3.4 Dealing with collateral portfolios ... 18
3.5 Comparison with other sources ... 19
3.6 Market statistics ... 21
3.7 Some EMIR network snapshots ... 23
4. Conclusion ... 25
References ... 26
Appendix ... 28
A.1 Reporting of centrally cleared trades ... 28
A.2 Data processing and confidentiality issues ... 29
A.3 EMIR introspection for the adjustment of GLEIF data ... 29

Summary

EMIR data can be used in many areas relevant to the work of supervisors, regulators and central banks, which include financial and market supervision, financial stability, research and statistics. In such areas EMIR data are useful to build real-time risk metrics and to visualise and capture interconnections between counterparties, hence supporting authorities' monitoring of financial markets.

The main purpose of this work is threefold: (i) to give a broad picture of the EMIR data, (ii) to highlight some practical issues associated with them, and (iii) to present some methods and solutions that analysts and researchers working in competent authorities may adopt when using these data. The issues in questions are related to data quality, to the two-sided reporting mechanism and to the quantitative information on margins, which are reported at the collateral portfolio level rather than at the contract level. The analyses presented in this paper are based on EMIR data available to Banca d'Italia, which is among the competent authorities with access to them.

Each day the EMIR dataset increases by millions of entries and a big data platform is needed to collect, store, process and analyse the data. Unlike other datasets available to competent authorities, EMIR data are not directly collected by them: data come from four trade repositories (TR) and ESMA provides a single access point to these data. Another key characteristic is that EMIR data are collected on a daily basis: this means that today one has the possibility to examine the derivative positions with reference to normally two trading days ago. This is of practical relevance for financial stability monitoring tasks.

After a brief description of the main aspects of the regulation on reporting and of the data collection process, we provide an overview of existing research works on EMIR data. Then, we move our focus to the data with the aim (1) to have an overview of the available information; (2) to define a framework to process the raw dataset in order to obtain a *proper* dataset for the analysis; (3) to provide some market statistics; (4) to visualise interconnections among counterparties.

More in detail, we first explore the qualitative information in the dataset. This first step is focused on the information on the counterparties, including their nationality, and on the characteristics of the contracts, including the contract type, the asset class of the underlying, the currencies of the legs, the execution venue and the information on clearing and collateralisation.

Then, we analyse one of the most important quantitative fields, which is the notional value. By leveraging on this value and on some other additional fields, we define a data cleaning framework for the overall data set. Additionally, two-sided reported trades are deduplicated[2] according to a pecking-order criterion centred around reporting reliability to avoid double counting of notional values. We then describe how to deal with collateral portfolios[3] to extract initial and variation margins.

The data processing framework allows us to explore the quantitative information related to the notional value, to compare the data with other sources (e.g. FINREP and the supervisory and statistical reports), to provide some derivative market statistics evaluated on the outstanding contracts, and to visualise interconnections between market participants.

Transactions where one of the counterparties is an Italian resident or where the underlying is Italian are only a small part of the EMIR dataset Banca d'Italia has access to. Most of them, in terms of the number of submissions, are reported by Dutch, Cypriot, and German counterparties. In terms of notional amounts, German entities are the major reporting counterparties, before Dutch and Cypriot entities. In their role of reporting, Italian entities are below five per cent of the total notional of outstanding contracts.

[2] If both counterparties submit a report for the same trade, only one is selected on the basis of the deduplication algorithm.

[3] To minimise the costs of collateralisation, the management of the collateral related to derivatives contracts between two counterparties is usually performed by grouping contracts with the same characteristics or underlyings (i.e. collateral portfolios).

The main findings of our analysis conducted on the data between January 2021 and April 2023 are the following:

- there exist thousands of reporting counterparties (considered at an individual level), for some of which there are millions of reports;
- the combination of the cleaning[4] and deduplication processes applied to the raw dataset decreases significantly the number of records and trades to be included in the analyses;
- as far as Italian banking groups are concerned, the notional amounts reported in EMIR are close to those reported in supervisory (FINREP, harmonised at the EU level) and statistical reports;
- a substantial portion of the total notional value is associated with derivatives on interest rates, particularly swaps; equity derivatives show the second largest share of the overall notional value, while credit and currency derivatives provide only modest contributions and commodity derivatives represent the smallest fraction.

Finally, with a new EMIR reporting regime entering into force in April 2024, the number of fields will increase from around 100 to around 200, in order to foster the alignment with the best market practices and to improve the information content of the submitted trades.

[4] Intragroup transactions are discarded in this analysis.

1. Introduction

European Market Infrastructure Regulation (EMIR) data can be used in many areas relevant for supervisors, regulators and central banks. These areas include financial and market supervision, financial stability and statistics. The main purpose of this work is to help analysts and researchers who work in competent authorities and wish to deal with EMIR data by also shedding light on some practical issues they may face when using this dataset. For example, EMIR data can be used to improve the authorities' understanding of how financial markets really work, to build real-time risk metrics, to visualise and capture the interconnections between counterparties.

EMIR dataset comprises millions of entries – each representing a derivative transaction – every day. The dataset complexity has two dimensions: (1) the complex nature of derivative instruments; and (2) the huge amount of data. A big data platform is needed to collect, store, process and analyse these data.

Unlike other datasets available to competent authorities, EMIR data are not directly collected by them (see Section 2 on the data collection process). This means that there is no direct communication channel between reporting entities and most data users, something that instead is helpful when for instance double checks on submissions from reporting entities is needed. As we will see in the section on trade repositories (TRs), the journey of reports from the reporting entity to the analyst working with EMIR data is far from being linear and asking for an explanation of a single submission can be challenging (not to say impossible).

Another key characteristic is that EMIR data are collected on a daily basis. This marks an important difference with the typical supervisory data which are collected only on a monthly (or quarterly) basis (e.g. Banca d'Italia is informed of the exact composition of banks' bond portfolios of supervised entities with a monthly frequency). Accessing EMIR data allows to know the trade-by-trade activity on derivatives of the supervised entities. Additionally, EMIR has also information on derivatives of non-supervised entities (e.g. non-financial companies).

After having briefly described the main aspects of the regulation, Section 1.3 reviews the research work carried out so far on EMIR. Section 2 explains the data collection process, while Section 3.1 provides an overview of the main collected information and on the reporting entities. Data quality issues and the processing framework are discussed in Sections 3.2 and 3.3, respectively. In Section 3.4 we show how to deal with collateral portfolio codes to extract quantitative information on initial and variation margins. By focusing on Italian banks, Section 3.5 compares the notional amounts extracted from EMIR data to those reported in FINREP and in the supervisory and statistical reports. Section 3.6 discusses the main derivative market statistics. In Section 3.7 we show examples of networks built with EMIR data. The analyses presented in Section 3 are based on EMIR data available to Banca d'Italia as one of the relevant competent authorities. In the Appendix we show how centrally cleared trades should be reported, we describe how to access the big data platform, some confidentiality issues and how to extract some information on the group structure from EMIR reports.

1.2 European Market Infrastructure Regulation (EMIR)

With the aim of improving stability, transparency and efficiency in derivatives markets, on 16 August 2012, Regulation (EU) No 648/2012, better known as European Market Infrastructure Regulation (EMIR), entered into force. EMIR is based on the following principles: clearing, margining and reporting. It also provides a framework for authorisation and supervision of CCPs under the responsibility of the national competent authorities (NCAs), as well as registration (or recognition) and supervision of TRs through the European Securities and Markets Authority (ESMA).

Under EMIR, there are three main obligations on market counterparties that hold for any form of derivatives: (1) to report all derivative contracts to a TR recognised by ESMA; (2) to centrally clear some

OTC derivatives; (3) to apply risk mitigation techniques for OTC derivative contracts not cleared by a CCP, such as confirmation, daily mark-to-market, portfolio reconciliation and exchange of collateral.

To reduce regulatory burdens and compliance costs and to improve transparency and data quality, EMIR has been amended by Regulation (EU) No 834/2019, in force from 2019. The most significant changes concern the categorisation of counterparties for the purpose of applying the clearing obligation and the exemption, under some circumstances, from the reporting obligation of intragroup derivatives contracts. Regulation (EU) 2019/363 amends the fields to be reported to TRs and defines the reporting template currently in force. Among the various regulatory documents of the EMIR and the Securities Financing Transactions Regulation (SFTR) package, Regulation (EU) 2019/361 concerns the access to the data held in TRs by the competent authorities. Direct and immediate access to derivatives data is essential to allow authorities to fulfil their responsibilities and mandates (see also Section 4.1.3.2 in ESMA, 2023).

ESMA and the European Systemic Risk Board (ESRB) have access to the full EU-wide dataset. Banca d'Italia has access to those transactions where one of the counterparties is an Italian resident or where the underlying asset is Italian. Additionally, Regulation (EU) 2019/361 grants Banca d'Italia access to a broader perimeter including all transaction data:

- on derivatives in either of the following cases: (i) where the reference entity of the derivative is established within a Member State whose currency is the euro and falls under the supervisory responsibilities and mandates of Banca d'Italia; (ii) where the reference obligation is sovereign debt of a Member State whose currency is the euro (see article 1, paragraph 9);
- on derivatives concluded on trading venues or by CCPs and by counterparties that fall under responsibilities and mandates of Banca d'Italia when monitoring systemic risks to financial stability in the euro area (see article 1, paragraph 10).

The implementation of the regulation seems not homogeneous among different TRs. We detected some issues which resulted in either underreporting and possible overreporting of EMIR data from TRs to Banca d'Italia. We are interacting with TRs and ESMA to solve these issues (see also Section 4.1.3.2 in ESMA, 2023, where the EMIR regulatory access filtering is discussed).

For every derivative contract, both counterparties are required to submit details of who they are trading with and the details of the trade. The reporting is conducted at trade level.[5] The dual nature of reporting (i.e. by both counterparties of each contract) is for data quality purposes and reduces the incentive to omit trades.[6] Reports are submitted to a TR, where the data is aggregated and stored (see Section 2). There are over 100 fields in each report, although not all fields apply to each trade.[7] The information to be reported includes data on counterparties and data about the contract, and is expected to be the same for both counterparties (i.e. *common data*). Details about the characteristics of the contract are also requested, including qualitative and quantitative information on the contract itself, on its underlying and on the collateral (when the contract is collateralised). Additionally, for quantitative information on contract value, price and margins, mandatory reporting on a daily basis is requested up to the maturity or the early termination of the contract. Under EMIR, initial and variation margins are calculated and reported, based on the net positions resulting from a set of contracts (*collateral portfolio*). Margins must

5 It should be noted that is not possible to understand if the derivative is used for trading, hedging or other purposes. There are only two fields (i.e. "*trading capacity*" and *"directly linked to commercial activity or treasury financing"*) that may help to understand if the reporting counterparty has concluded the contract as principal on own account (on own behalf or behalf of a client) or as agent for the account of and on behalf of a client and, if the contract is directly linked to commercial or treasury financing activities. The quality of these two fields does not seem adequate for their use.

6 Firms can choose to delegate the reporting obligation to the other counterparty or a third-party. As usual, the delegating firm remains responsible for the accuracy of the reports.

7 See the Commission Implementing Regulation (Eu) 2019/363 of 13 December 2018.

be computed at the collateral portfolio level, rather than at the contract level. As discussed in Section 3.4, dealing with this reporting framework is challenging from a computational perspective.

A new reporting regime will enter into force in April 2024 with the aim of improving the clarity of the reporting framework.[8] This regulatory package defined in 2022 and known as EMIR Refit amends the details of the data to be reported to TRs by increasing the number of fields from around 100 to around 200 (see the regulatory and implementing technical standards 2022/1855 and 2022/1860).[9] It is to be noted that the major part of the increase in the number of fields is to align the reporting to the best market practices and to improve the information content of the submitted data.

1.3 Overview of existing research work on EMIR data

Empirical studies on granular derivatives data are relatively few, mainly due to limited data availability (see also the Annex in ESMA, 2023). Since 2014, EMIR started to require EU institutions to report their derivative transactions to TRs and data availability in Europe has largely improved since then. TR data are confidential and only ESMA, ESRB, European Central Bank (ECB) and NCAs have access to these transaction-by-transaction data. Additionally, a big data infrastructure is needed to store, manipulate and analyse this large amount of data. Confidentiality rules, together with technical complexities, slow down the development of a specific stream of research and papers exploring EMIR data have been published in peer-reviewed academic journals only recently.

The research on EMIR data can be divided into three main categories: (1) notes on data description (including on data quality issues); (2) studies on specific contract types or particular market events; (3) risk analyses based on network methodologies.

The description of EMIR data is not a simple task. After an overview of the EMIR dataset and of the data cleaning procedures, Abad et al. (2016) provide descriptive statistics of interest rate, FX, and CDS markets. Ascolese et al. (2017), van Lelyveld (2017) and Ullersma and van Lelyveld (2022) explore the potential and challenges of EMIR data for policymakers and regulators. Interesting papers on data cleaning methodology are those by Pérez-Duarte and Skrzypczynski (2019) and Lenoci and Letizia (2021).

Abbassi and Bräuning (2021) study German banks hedging strategies based on FX forward contracts. Numerous analyses have been conducted on credit default swap (CDS) contracts (see Kenny et al., 2016, Levels, 2018, Abraham, 2020, Pedro, 2020, Bianchi, 2021, Mazzacurati et al., 2021). Among CDS contracts, in fact, there are the single-name CDS that are simple to analyse, given that their information is more standardised[10] and the risk direction can be easily identified. Bellia et al. (2018) examine the clearing choice of counterparties stipulating CDS with underlying of Italy, France and Germany's sovereign debt. While Bias et al. (2020) shed light on the derivatives use by equity mutual funds, Kenny (2016) and Fiedor and Killeen (2021) focus on securitisation special purpose entities.

Some researchers investigate specific market events. Cielinskaet al. (2017) and Joseph and Vasios (2022) investigate the impact of the Swiss National Bank's decision to discontinue the floor of 1.20 Swiss francs per euro on the morning of 15 January 2015 on the foreign exchange derivatives market. Schroeder (2020) examine the underlying drivers of the flash crash in Pound Sterling vs US Dollar in October 2016 and the contribution of different market participants to the drying up of liquidity during that flash crash. Bouveret et al. (2022) explore Archegos positions and show that it is possible to track the steep increase in concentrated exposures that the family office undertook in February and March 2021.

8 See the ESMA Guidelines for reporting under EMIR of 14 December 2022 and the Commission Delegated Regulation (EU) 2022/1855.

9 See also ESMA (2022b).

10 Their underlying is usually identified by an International Securities Identification Number (ISIN) or a Legal Entity Identifier (LEI) code.

Some papers examine initial and variation margins (see ESRB, 2020, and Cominetta et al., 2019) also with a focus on risk management practices of central counterparties (Grothe, 2021). D'Errico and Roukny (2021) use bilateral obligations resulting from CDS contracts to explore the mechanics of derivative portfolio compression[11] in terms of both feasibility and efficiency.

There are numerous studies implementing network methodology on EMIR data. Fiedor (2017) presented a comprehensive description and analysis of the centrally cleared interest rate derivative market in the EU. Bardoscia et al. (2019) build a separate network for each derivative market (interest rate, credit, and foreign exchange derivatives) and, by joining them together in a single multiplex network, they assess the systemic vulnerability of individual institutions and develop centrality measures specifically designed for multiplex networks. Rosati and Vacirca (2020) reconstruct the network of relationships in the centrally cleared derivatives market between clearing members and CCPs, as well as those between clearing members and their clients. They analyse the topology of the network through different centrality measures applied to both the aggregate network and the multiplex network. Then, they identify critical nodes and assess their interconnectedness within and across the different market segments. Drawing on derivative margins data and cash buffers, Bardoscia et al. (2021) develop a model to compute liquidity shortfalls in a stress scenario. Kahros et al. (2021) provide a descriptive analysis of client clearing in both OTC and exchange-traded derivative markets and across all products.

Recently a fourth stream of research appeared, more focused on risk assessment. Among others, Jukonis (2022) introduces a model to estimate market risk for equity derivatives exposures in an adverse scenario.

2. Data collection process through trade repositories

In this section, we briefly recall the data collection process. A TR is an entity that centrally collects and maintains records of financial transactions, with the aim of allowing firms to report these transactions to their corresponding regulatory authority. Under Regulations (EU) No 648/2012 (EMIR) and No 2015/2365 (SFTR), TRs collect and maintain records of derivatives transactions. They play a central role in enhancing the transparency of derivative markets and providing detailed data to monitor risks to financial stability.

The EMIR[12] and SFTR[13] regulations require counterparties to report daily to a TR the details of transactions concluded, modified and terminated (life-cycle events) on derivatives and on securities financing transactions, respectively. Once registered, the TR is supervised by ESMA to ensure that it complies on an ongoing basis with all EMIR and SFTR requirements, thereby enabling regulators to

[11] Portfolio compression is a multilateral netting technique that enables market participants to coordinate the replacement of existing contracts in order to reduce the size of their mutual obligations, thereby reducing counterparty risk, while maintaining the same underlying market risks.

[12] Counterparties subject to reporting on the basis of the EMIR regulation (art. 9) are: financial counterparties (including banks, investment firms, insurance companies, UCITS and pension institutions, central counterparties) and non-financial counterparties (companies established in the EU other than financial counterparties) established in the EU (including their branches and non-EU branches) and all EU branches based outside the EU. Regulation (EU) No 834/2019, which entered into force on 17 June 2019, amended the reporting obligation of derivatives for non-financial counterparties (NFCs): an NFC must report to the TR the OTC derivative contracts belonging to the categories for which it exceeds the thresholds (clearing thresholds) indicated by ESMA, or for all contracts, if it does not calculate its position in relation to these thresholds.

[13] Counterparties subject to SFTR reporting (art. 2) are: financial counterparties (including insurance companies, funds and market infrastructures such as central counterparties and central securities depositories) and NFCs established in the EU (including related branches and non-EU branches) and all EU branches of entities based outside the EU.

access data and details of derivative contracts to fulfil their respective mandates. ESMA supervises the TR against which it has authorising, controlling, and sanctioning powers.

The companies can choose the TR to which they report the trades on a daily basis. One company can use more than one TR for EMIR reporting: each counterparty independently reports the concluded transactions to the chosen TR. If other actors are involved (e.g. brokers, CCPs), a single transaction gives rise to multiple reports (see Appendix A.1).

Eight TRs have been registered by ESMA since 2013. After Brexit, only four of them are currently active in the EU market (see Table 1).

TRs registered by ESMA

Table 1

Denomination	Derivative asset class	Country (currency of the fee)	Effective date
DTCC Derivatives Repository Ltd. (DTI)	all	Ireland (USD)	14/11/2013
Krajowy Depozyt Papierów Wartosciowych S.A (KDP)	all	Poland (PLN)	14/11/2013
Regis-TR S.A. (RGS)	all	Luxembourg (EUR)	14/11/2013
UnaVista TRADEcho B.V. (UVT)	all	Netherlands (EUR)	14/11/2013

Source: ESMA.

It is important that market participants report to TRs all details regarding derivative contracts they have entered into. As a result, information on the risks inherent in derivatives markets will be centrally stored and easily accessible to ESMA, the relevant competent authorities, the ESRB and the relevant central banks of the European System of Central Banks[14]. The provision of TR services is characterised by economies of scale, which may hamper competition in the field. At the same time, the imposition of a comprehensive reporting requirement on market participants may increase the value of the information maintained by TRs and also for third parties providing ancillary services, such as trade confirmation, trade matching, credit event servicing, portfolio reconciliation or compression.

The reports are of two types: (1) trade activity reports, to which we refer to as *activities* (A), containing all reports sent the previous day; (2) trade state reports, to which we refer to as *states* (S), containing all pending trades at the end of the reporting day. While activities show each lifecycle event of a transaction (e.g. conclusion, valuation, modification, termination), states, are at a higher level of aggregation.

Even if data users (e.g. NCAs) can request direct access to TR data, ESMA developed a single access point to EMIR transaction data (i.e. the Trace project) to provide a unified access to the transaction data stored by TRs (see Figure 1). Although the Trace project simplifies data collection by competent authorities, the journey of the reports from the reporting entity to the analyst working with data is far from linear. There is no direct connection between data users and reporting entities and asking for an explanation of a single submission can prove very challenging.

[14] The EMIR and SFTR Regulations and the ESMA regulatory technical standards (EU Reg. 2019/357 and 2019/361) ensure NCAs the visibility of data on the basis of their respective institutional prerogatives. For the Bank of Italy, the access powers are attributable to the following purposes: financial stability, monetary policy, supervision of CCPs, banking and financial intermediaries, and crisis resolution.

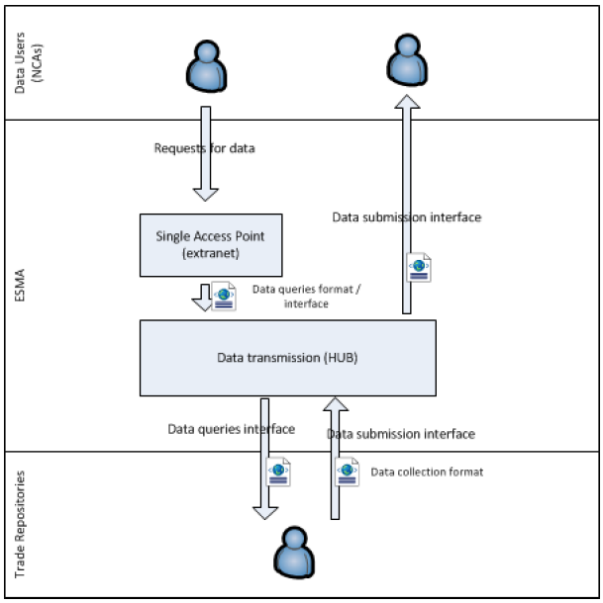


Figure 1: ESMA Trace project. Source: ESMA documentation.

3. EMIR Data

The focus of this section is on EMIR data and the analyses presented are based on data available to Banca d'Italia as one of the relevant competent authorities. Section 3.1 provides an overview of the main information on the contracts and on the reporting entities. From Section 3.2 to 3.4 we discuss data quality issues, the processing framework and how to extract quantitative information on initial and variation margins. In Section 3.5 we compare, for a sample of Italian banks, the notional amounts extracted from EMIR data to those reported in FINREP and in the supervisory and statistical reports. Market statistics in terms of notional value are reported in Section 3.6. Finally, in Section 3.7 we show examples of networks drawn from EMIR data.

3.1 Data overview

Starting from December 29, 2020, Banca d'Italia receives each month through Trace more than half billion daily records collected from four TRs (Figure 2).[15] This means that we have around 30 million records for each trading day; most reports come from Regis. According to Chart 10 of ESMA (2021), Banca d'Italia has access to about half of total EMIR submissions. As reported by ESMA, at the beginning of 2021, the number of EMIR submissions sharply dropped, as UK-based TRs were excluded. Both activities and states are reported for a total of more than 17 billion records in the period from the last days of 2020 to the end of April 2023.

EMIR data are also enriched with other information from Global Legal Entity Identifier Foundation (GLEIF), that allows to identify entities through the legal entity identifier (LEI) code, a 20-digit alphanumeric code that enables clear and unique identification of companies participating in global financial markets.

In more than 70 per cent of the records, both counterparties have an LEI code, that is entities are operating legally as of the last verification. In some cases, however, the LEI code is not available and the reporting entity uses a client code internally defined to identify the counterparties. For instance, this can

[15] In counting the number of reports, we do not consider the updates of the same report but only the last available report.

happen when the counterparty is a retail client (i.e. an individual); there are also cases of firms in which the LEI does not result as *issued* anymore.[16]

In what follows, we provide the broad picture of the EMIR archive stored in the Banca d'Italia big data platform. We focus on the number of submissions (observations) related to *activities* and *states*. Analysing states allows us to have a look at the entire picture of the derivatives reported in EMIR in terms of outstanding contracts. For this reason, Section 3.6 on market statistics deals only with the notional in euro of trade states. In this section instead, we focus on all reports (observations) between January 2021 and April 2023.

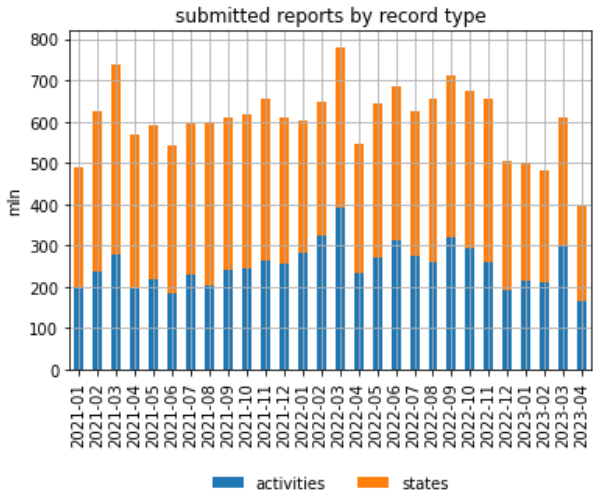


Figure 2: Overall monthly data. Source: EMIR data available at the Banca d'Italia. Number of monthly reports by record type.

For both reporting and other[17] counterparty's countries, Figure 3 shows for each country the number of observations (states and activities). Italy is the seventh country by number of reported records. Less than 10 per cent of the transactions involves Italian counterparties. Most of the transactions are reported by Dutch, Cypriot, and German counterparties. Cypriot counterparties report almost 20 per cent of total transactions, but the notional of these trades is small. Even if British counterparties have been excluded from reporting after Brexit, these are still a large portion of the counterparties resulting as "other".

At an individual level, there are thousands of reporting counterparties but the first 50 counterparties represent more than 80 per cent of the observations available at the Banca d'Italia. The counterparties reported as "other" are less concentrated: the first 50 counterparties represent slightly less than 50 per cent of the total observations.

As far as the contract characteristics are concerned (see Figure 4), equity derivatives account for more than 60 per cent of total submissions while contracts with interest rates as underlying are almost 20 per cent. Commodity and currency asset classes are almost equivalent in terms of total observations (around 10 per cent). At least for the counterparties to which we have access to, credit derivatives are a small fraction of the overall derivative market.

16 Possible LEI registration status are: issued, duplicate, lapsed, merged, retired, annulled, canceled, transferred, pending_transfer, pending_archival.

17 On each trade there are two counterparties, i.e. the reporting and the other.

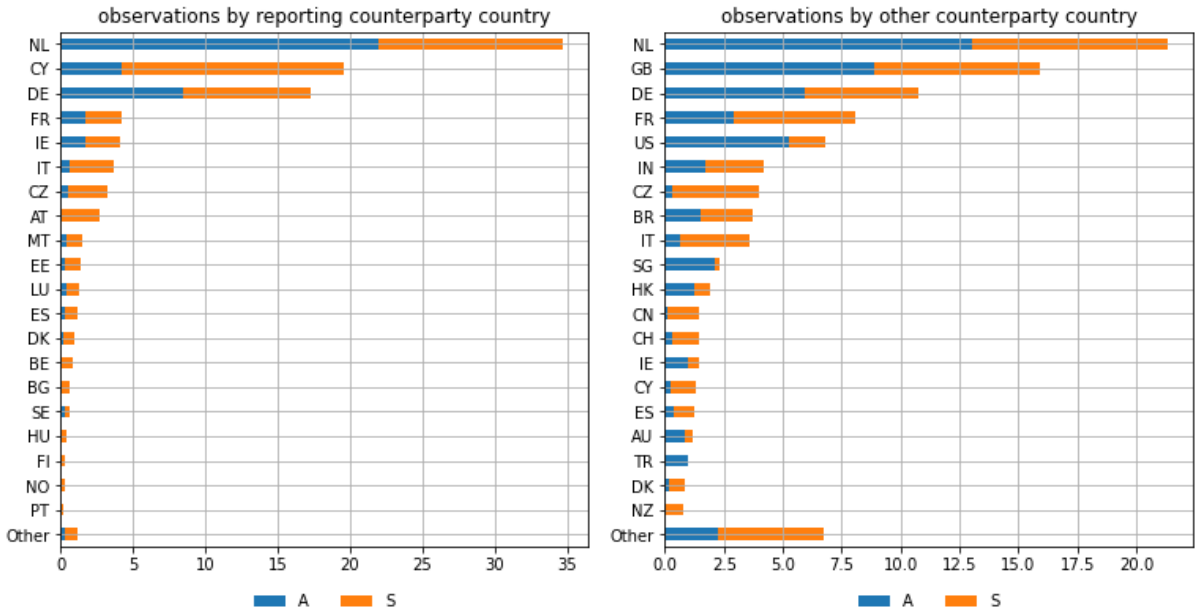


Figure 3: Overall data by country of the reporting (other) counterparty: number of observations. Source: EMIR data available at the Banca d'Italia. Percentage over the total number of observations.

More than one-third of reports refer to futures contracts (see Figure 4). Contracts for differences are more than one quarter of the total reports. This is may also be due to the fact that they are retail products and many trades have negligible notional values (see also Figure 7). While options are around 20 per cent of the total, all other contract types, including swaps, are below 6 per cent. Spread bets represent a market niche.

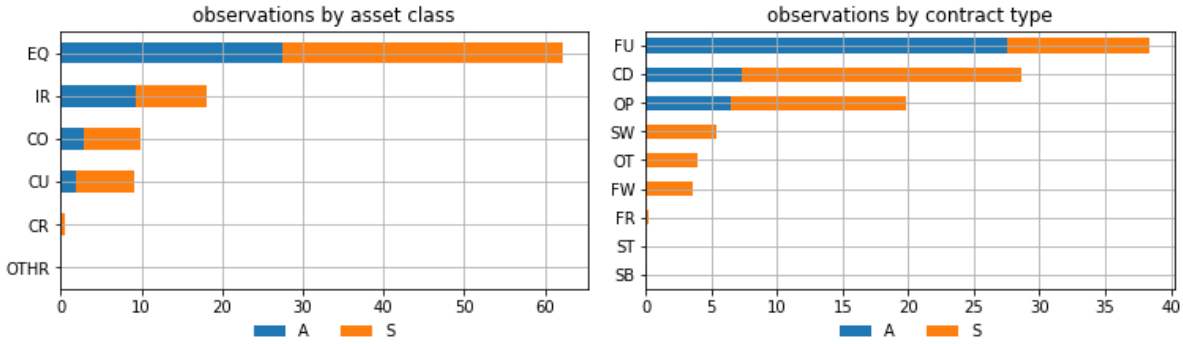


Figure 4: Overall data by contract characteristics: number of observations. Source: EMIR data available at the Banca d'Italia. Percentage over the total number of observations. Asset classes: commodities (CO), credit (CR), currencies (CU), equities (EQ), interest rates (IR), other (OTHR). Contract types: contracts for differences (CD), forward rate agreements (FR), forwards (FW), futures (FU), options (OP), spread bets (SB), swaps (SW), swaptions (ST), other (OT).

The main currencies are the euro and the US dollar (see Figure 5). However, certain other currencies are not negligible, at least in terms of number of submissions.[18]

[18] In the case of an interest rate or currency derivative contract, these will be the notional currency of leg 1 and 2.

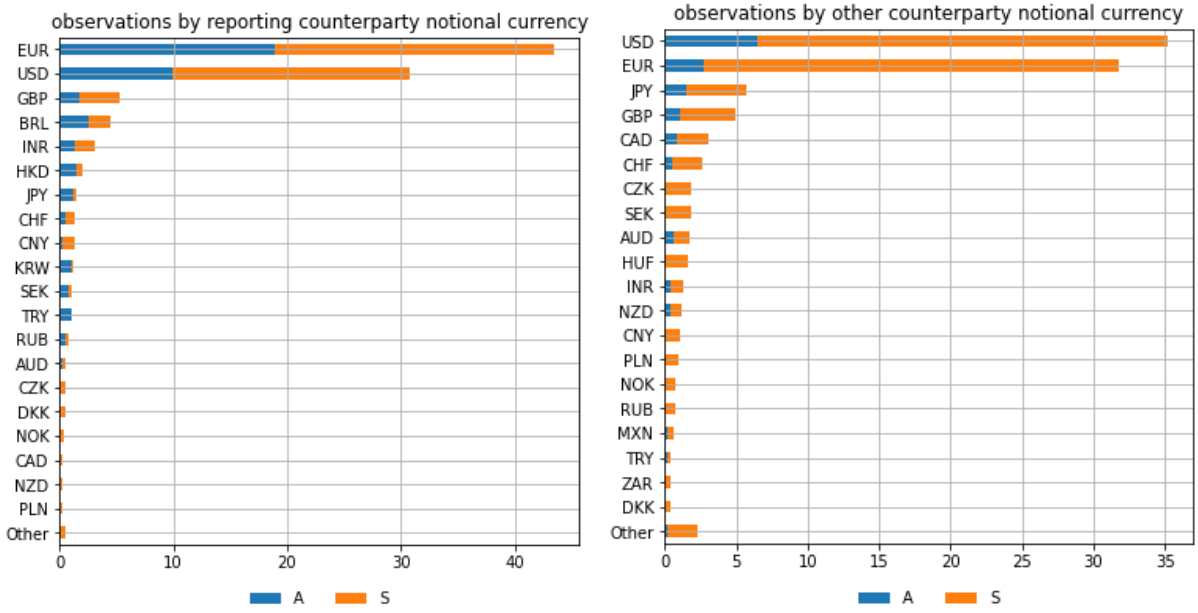


Figure 5: Overall data by notional currency: number of observations. Source: EMIR data available at the Banca d'Italia. Percentage over the total number of observations.

While around 63 per cent of overall records on states refer to over-the-counter (OTC) contracts, the remaining 37 are No-OTC, that is they can be exchange traded and cleared (ETD) or traded on multilateral or organised trading facilities (MTF or OTF) and not necessarily centrally cleared.

Figure 6 distinguishes between markets, cleared and non-cleared transactions and collateralisation type. More than one quarter of transactions are uncollateralised. Most of the collateralised transactions are partially or one-way collateralised. It could be the case that the quality of these fields is not always high.

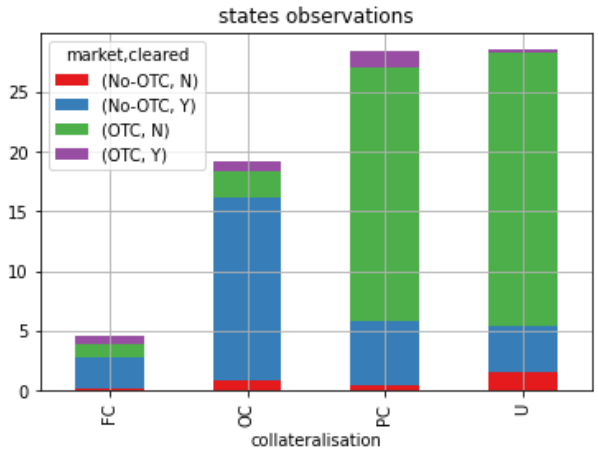


Figure 6: Overall data by market, clearing and collateral information: trade states observations. Source: EMIR data available at the Banca d'Italia. Percentage over the total number of observations. Collateralisation type: fully collateralised (FC), one way collateralised (OC), partially collateralised (PC), uncollateralised (UC).

One of the most important quantitative data is the contract notional value. The size of this value depends on various factors, including the counterparties' sectors, the contract type and the asset class. Figure 7 shows the states notional in euro by size buckets. We report "not a number" (*nan*), when the euro exchange rate is not available.[19] Some entities report negative or huge values, which in most cases are reporting errors. More than one quarter of the trades have a notional between 0 and 1,000 euro. Given

[19] The values of this field are not always available for a technical issue we are currently solving.

the volume of trades of this type, it is not always clear whether they are reporting errors or if they are transactions of retail clients on trading platforms. The amount of trades with notional greater than one billion is negligible.

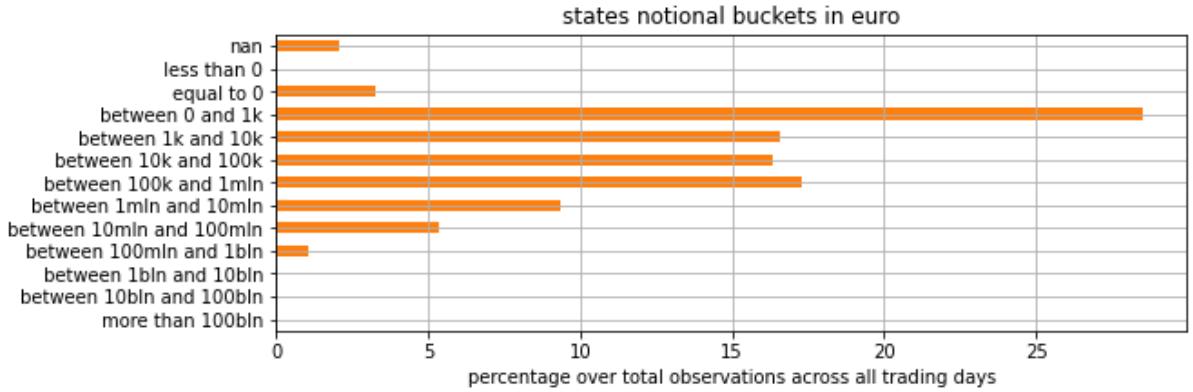


Figure 7: Trade states notional in euro by size buckets. Source: EMIR data available at the Banca d'Italia. Percentage over the total number of observations.

3.2 Data quality issues

Nearly eight years after the beginning of the reporting mandate, the data quality collected under EMIR is still far from perfect and further effort is needed. The annual reviews of the quality of EMIR data published by ESMA show that, despite the progress made in recent years, certain issues are still present (see ESMA, 2022). In particular, the report published in 2022 shows that, while less than 10 per cent of reported derivatives tend to be reported late by the counterparties, more than 20 per cent do not receive updated valuations on a daily basis, and pairing rate continues to be relatively low at 60 per cent.[20] In a recent letter to the European Commission, the ESRB states that persistent data quality issues pose several limitations to the use of EMIR data for the monitoring of financial stability risks.[21]

Among the issues raised by the ESRB two stand out. First, in the monitoring of the post-Brexit clearing landscape, what appeared to be the repatriation of third-country contracts from the UK to the EU proved to be due to misreported notional values. In addition, many reported transactions lacked the important fields to identify the product class (for example, the underlying rates for interest rate swaps). Second, in monitoring liquidity risks arising from margin calls during the COVID-19 crisis and following the Russian invasion of Ukraine, it appeared that many entities wrongly reported collateral as flows instead of stock values. In addition, the amounts and the portfolio codes do not match those reported by the counterparty.

Similar problems have emerged in the work done at Banca d'Italia. The data quality is not homogeneous across reporting entities, derivatives markets and reported fields. These are the major issues:

1. The classification of the type of players (e.g. dealer vs. investor, hedge fund vs. ETF) is problematic because information on counterparties is limited. On the bright side, LEI codes however allow us to merge EMIR data with external data on the counterparty information that is needed.

[20] See ESMA (2022a).

[21] See ESRB (2022). The letter highlights a few examples of when the EMIR data proved to be fundamental for the ESRB in monitoring developments from a systemic risk perspective but substantial data quality issues hampered the ability to perform the analyses and achieve meaningful and reliable results.

2. The identifier of the derivatives underlying assets may be missing or inconsistently reported (e.g. ISIN vs. differing name(s) in text). Therefore, the aggregation of positions that refer to the same underlying asset is cumbersome.

3. The dual-sided reporting regime reveals in a few cases conflicting information provided by the two counterparties regarding the direction and quantity of the trade and the contract specification. Only 0.3 per cent of total trades in EMIR have a discrepancy in the notional amounts reported by the two counterparties greater than 25 per cent, and in only 0.1 per cent of trades the counterparties provide inconsistent information, by stating that they are both buyers or both sellers at the same time.[22]

4. The usefulness of initial and variation margins for financial stability monitoring is also compromised by the fact that margins are reported at the collateral portfolio level, instead of trade level (see Section 9). The fact that collateral portfolios are specific to a pair of counterparties does not guarantee that the two counterparties include exactly the same trades in their collateral portfolios (e.g. misreporting or data quality issues) and that they use the same name to denote those portfolios. Without the same composition and denomination of collateral portfolios used by the pair of counterparties, the risk of reporting inconsistencies grows.

3.3 Data processing framework

In this section, we describe in detail the steps that we take to transition from the raw dataset of "trade state" reports to one that is fit for the analysis. The processing procedure discards erroneous or unnecessary observations (cleaning and filtering), and samples only one report per trade to avoid double counting of a single trade coming from the EMIR dual-sided reported regime (deduplication). We discuss results from the cleaning and deduplication process using "trade state" data sampled at a weekly frequency from December 29, 2020 to August 31, 2022. For each of the weekly data points we select a Friday or the last trading day in the week if it is not Friday. Table 2 outlines the data processing procedure and presents statistics on the median of the weekly results.

The combination of the cleaning and deduplication exercises leaves us with 7.13 million records (42.5 per cent of initial), 7.13 million trades (49.1 per cent of initial), and 86.1 trillion euros of total gross notional (71 per cent of initial). Panel A in Table 2 reports the median daily number of trade state reports (records), number of unique trade identifiers and total gross notional (in euro billions). We call these values *initial*, as they refer to the raw dataset, i.e. before data processing. On an ordinary day, there are almost 17 million records, which refer to approximately 14.5 million distinct trades and have an aggregate notional of 1,261 trillion euro.

In what follows, we assess the impact of the various data processing steps in terms of drop rates computed with respect to the initial values. In order to have a meaningful representation of the effects of the processing procedure on the aggregate notional, we use as initial values the outcome of our first cleaning step, which removes outlier observations characterised by implausible notional amounts (either too small or too large). Net of outliers, the initial total gross notional is just over 121 trillion euros (bottom right corner of Panel A in Table 2).

Panel B in Table 2 illustrates the data cleaning steps and reports the number of records and trade identifiers, as well as the total gross notional that survive and those that are dropped in each step, which are implemented sequentially. Alongside the number of dropped observations, we report its percentage share of the initial value, which we use to gauge the magnitude of the discarded information.

The **first cleaning step** removes erroneous records due to implausible notional values defined as either below 1,000 or above 50 billion euros. These thresholds are based on both standard practices and

[22] The total notional in EMIR affected by each of these inconsistencies is about 1 per cent.

exploratory analysis of the outliers. The choice of the lower bound is in line with previous studies on EMIR data (Abad et al., 2016) and it allows not only to deal with non-positive notional values but also to remove all those trades that given their size might not be important from a systemic risk viewpoint. We note however that these smaller trades account for a substantial part of the raw EMIR dataset. In fact, over 99% of the 5.2 million observations discarded in this first step are records with positive notional value below 1,000 euro and mostly refer to trades that involve online brokerage firms and their retail clientele. While the lower bound is the reason why a third of the initial records and trades are discarded in the first step, the upper threshold is what causes the huge drop in the total gross notional to about a tenth of its initial value. We choose 50 billion, which is bigger than the 10 billion threshold used by Abad et al. (2016), based on evidence that trades with notional amount above that value are never reported by CCPs and are often reported by non-financial counterparties, which makes us confident in treating them as erroneous reporting.

The **second and third cleaning steps** exclude trades that appear inactive. This happens either when the reference date of the reporting is beyond the trade's stated maturity or termination date (Filter 2) or when the valuation of the contract is done more than 100 days from the reference date (Filter 3). The latter filter is more impactful as it drops approximately 7 per cent of both records and trades and 10 per cent of the total gross notional.

The **fourth step** discards reports that do not specify whether the reporting counterparty is buying or selling the derivative. The **fifth step** discards reports where both the reporting entity and its counterparty do not have the LEI code. The impact of these two filters is negligible.

A strong effect on the cleaning procedure comes from the **sixth cleaning step**, which excludes intragroup positions between companies of the same group. We identify these trades using the intragroup flag reported under EMIR, which we further complement with information on group structure and ownership from GLEIF. Intragroup transactions do not constitute standard buy (or sell) market operations; their purpose can rather reflect risk transfer or liquidity management. As such, these positions can be excluded from financial stability analyses. This notwithstanding, they might still be relevant for different kinds of analyses and in this case they might still be looked at. By applying this filter, we drop about 9 per cent of the reports, accounting for almost 8 per cent of the total notional amount in EMIR. It should be noted that this is not a cleaning step but a filter one needs to explore EMIR data from a financial stability perspective, where intragroup positions do not affect the risks of the counterparties.

The **seventh cleaning step** drops redundant records leaving the number of trades unchanged. It ensures that, for a given trade, there exists at most one report per counterparty. This is not data deduplication per se, as we are not choosing which of the two reports is considered, but rather we are focusing on the reports submitted by the same counterparty. The same counterparty can erroneously submit multiple reports for a given trade on a given day for a number of reasons. For instance, this is observed when a report is submitted to flag the origination of a contract and another report is used to provide its valuation.

Based on the list of reported EMIR fields that are recognised as compulsory by the ESMA validation rules, we favour reports with action type[23] of valuation ("V"), i.e. inserted for valuation updates, and drop the others. Whenever the preferred action type "V" is not available, we randomly pick one of the existing reports. [24]

[23] The values of this field can be: new (N), modify (M), error (E), early termination (C), correction (R), compression (Z), valuation update (V) and position component (P). A simpler, from a practical perspective, alternative, giving similar results in terms of total notional values, is to discard all trades having more than one action type for a given reference date and reporting counterparty.

[24] See the Excel file on validation rules at the following link https://www.esma.europa.eu/policy-rules/post-trading/trade-reportingopa.eu

Cleaning process

Table 2

BEFORE CLEANING (A)					
records	initial:	16,795,372			
trades	initial:	14,525,682			
notional (bn)	initial:	1,261,440	initial w/o implausible:		121,179
AFTER CLEANING (B)					
FILTER 1: implausible notional (<1000 or > 50bn)					
records	survive:	16,795,372	drop:	5,244,802	(31.2%)
trades	survive:	14,525,682	drop:	5,464,352	(37.6%)
notional (bn)	survive:	1,261,440	drop:	-	
FILTER 2: maturity or termination date before reference date					
records	survive:	11,472,930	drop:	77,641	(0.5%)
trades	survive:	8,971,797	drop:	89,534	(0.6%)
notional (bn)	survive:	120,884	drop:	295	(0.2%)
FILTER 3: valuation date in excess of 100 days from reference date					
records	survive:	10,341,586	drop:	1,131,343	(6.7%)
trades	survive:	7,956,912	drop:	1,014,884	(7.0%)
notional (bn)	survive:	107,782	drop:	13,102	(10.8%)
FILTER 4: missing counterparty side					
records	survive:	10,341,490	drop:	96	(0.0%)
trades	survive:	7,956,871	drop:	42	(0.0%)
notional (bn)	survive:	107,781	drop:	1	(0.0%)
FILTER 5: no LEI for both counterparties					
records	survive:	10,338,336	drop:	3,154	(0.0%)
trades	survive:	7,952,988	drop:	3,884	(0.0%)
notional (bn)	survive:	107,779	drop:	2	(0.0%)
FILTER 6: intra-group transaction					
records	survive:	8,848,798	drop:	1,489,538	(8.9%)
trades	survive:	7,132,240	drop:	820,748	(5.7%)
notional (bn)	survive:	98,513	drop:	9,266	(7.6%)
FILTER 7: multiple reports of the same trade submitted by the same counterparty					
records	survive:	8,842,711	drop:	6,088	(0.0%)
trades	survive:	7,132,240	drop:	0	(0.0%)
notional (bn)	survive:	98,403	drop:	109	(0.1%)
FILTER 8: unmatched pair of counterparties in double reporting					
records	survive:	8,833,348	drop:	9,364	(0.1%)
trades	survive:	7,130,328	drop:	1,912	(0.0%)
notional (bn)	survive:	98,290	drop:	113	(0.1%)
AFTER CLEANING and DEDUPLICATION (C)					
records	final:	7,130,329	(42.5%)		
trades	final:	7,130,329	(49.1%)		
notional (bn)	final:	86,093	(71.0%)		

The **final cleaning step** excludes trades for which reports submitted by distinct counterparties are inconsistent in the identification of the pair of entities involved. This can happen for instance when both A and B are entities subject to the EMIR reporting obligation but while A reports a trade with B, B reports that the same trade is in place with a third entity C. We treat these cases as erroneous observations to drop. The impact of this filter is very limited.

Panel C in Table 2 shows the results of deduplication of the clean EMIR dataset, by dropping one observation per pair of matched transactions. Deduplication of two-sided reported trades is carried out according to a pecking-order criterion, centred around reporting reliability. More specifically, we assume that CCPs authorised and recognised by ESMA are more reliable than their clearing members, and the latter are more reliable than other entities. After obtaining the list of authorised[25] and recognised[26] CCPs from the ESMA's website, we retrieve their clearing members by sampling the non-CCP counterparties of their trades. We distinguish 3 cases of trades that need deduplication: (i) for trades between a CCP and its clearing member, we take the CCP report; (ii) for trades between a clearing member of one of the CCPs and a counterparty that is neither a CCP nor a clearing member of it, we take the clearing member's report; (iii) for all other trades (i.e. none of the counterparties is a CCP's clearing member or they both are) we take the report of the counterparty that on the date of interest has submitted the largest number of reports.

3.4 Dealing with collateral portfolios

To minimise the costs of collateralisation, the management of the collateral related to derivatives contracts between two counterparties is usually performed by grouping contracts with the same characteristics or underlying (i.e. collateral portfolios). By following this market practice, initial and variation margins are reported at the collateral portfolio level, rather than at the contract level. However, while at the level of a single contract the counterparties submit data by reporting the same trade identifier (*trade id*), this is not the case for the collateral portfolio code. Indeed, there is not a unique identifier, and two counterparties may label the same collateral portfolio with different codes. Data show that two pairs of counterparties may not agree on the number of collateral portfolios that exist between them. It can occur that, for a given trade, the first counterparty reports a collateral portfolio code, while the second counterparty does not. It can also occur that only one counterparty submits the report of a given trade.

To deal with these issues, we leverage the double reporting of the same trade and define, on each reference date, a unique collateral portfolio identifier composed by the LEIs of the counterparties sorted in alphabetic order and a progressive number that denotes and groups together all collateral portfolios that are used by the two counterparties and have some trades in common. On the basis of this identifier, for each counterparty pair it is possible to compute the initial and variation margin posted and received by a given entity at a given reference date.

In order to extract all available information provided by the double reporting, the analysis of margin should be conducted considering both sides of the reporting. For each counterparty pair and each unique collateral portfolio, it is possible to merge, by means of a selected function, the values of margins received by a counterparty with those posted by the other counterparty.[27] In practice this means that an analyst working with EMIR data needs both a deduplicated data set to investigate notional and contract values and a not deduplicated data set to examine margins.

Margin values are not always available in EMIR data. Firstly, according to Regulation (EU) No 148/2013, most non-financial corporations are not required to report collateral of contracts. Secondly, the EMIR

25 https://www.esma.europa.eu/sites/default/files/library/ccps_authorised_under_emir.pdf

26 https://www.esma.europa.eu/sites/default/files/library/third-country_ccps_recognised_under_emir.pdf

27 Average, median or maximum between the two reported values can be selected.

validation rules by ESMA do not always allow retrieving the values of margins posted by an entity from the trade reports submitted by its counterparty. For instance, in the case of partially collateralised derivative contracts, where the reporting counterparty might be the only one receiving the initial margin, the validation rules do not require populating the field of initial margin received. Without this value, it is not possible to know the value of the initial margin posted by the other counterparty unless this submits its own report. For all these reasons, any analysis of margins should hinge on EMIR considering both sides of the reporting and give priority to the reports submitted by the entity whose margins are being assessed.

3.5 Comparison with other sources

For Italian banks EMIR is not the only available source of data on exposure to derivatives (but it is so for non-financial firms). Italian banks have to submit information regarding their exposures to derivatives in FINREP and in the national supervisory and statistical reports (Matrice dei conti).[28] We consider these two datasets to validate the notional amounts extracted from EMIR data available at Banca d'Italia and obtained after the cleaning and deduplication phases.

Supervisory and statistical reports have a quarterly frequency and include information on notional values and carrying amounts of derivatives used for hedging and trading purposes. The values are broken down by type of underlying instrument.[29] FINREP framework requires banks to quantify their exposures both at group and at individual level, while the national supervisory reports collect individual data only.

The cross-check has been carried out on six banks considered at a group level and on a quarterly basis.[30] Each bank has a total notional exposure of more than 100 billion. These banks account for almost the entire derivatives exposure held by the Italian banking system. Figure 8 reports, for each bank, the comparison between the three data sources in terms of notional amounts held at the end of each quarter between Q4 2020 and Q4 2022.

[28] On 1 January 2014, the fourth directive on capital requirements set a regulation of reporting to harmonise the collection of data in the Eurozone. The FINREP report is among those designed to meet the fourth directive legislative requirements.

[29] The derivatives are distinguished among interest rates derivatives, equity derivatives, currency derivatives, credit derivatives and commodity derivatives. There is also a residual class for other types of derivatives.

[30] Given that the national supervisory report is at the individual level, the notional values have been aggregated according to the bank group structure at the relevant date.

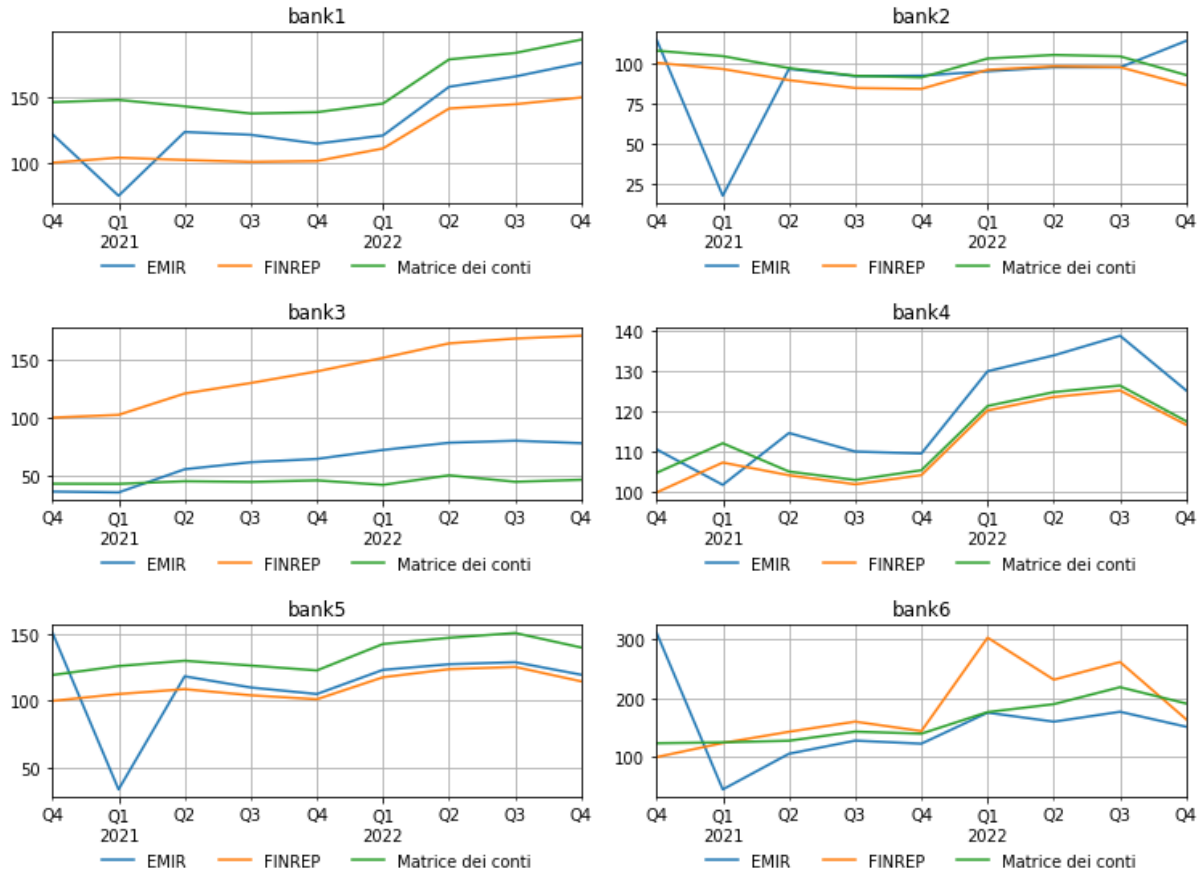


Figure 8: Comparison between different sources between Q4 2020 and Q4 2022. Source: deduplicated EMIR data available at the Banca d'Italia, FINREP and supervisory and statistical reports (Matrice dei conti). All data are rescaled so that, for each bank, the FINREP value is equal to 100 at the Q4 2020. For a bank we consider the reports submitted by the banks including the transactions with the parent company.

Except for bank3, the amounts reported in EMIR are close to those reported in the other two datasets. Moreover, dynamics are very similar. Values reported by bank3 in FINREP are much different from those contained in EMIR and in the national supervisory report. This is due to problems in properly identifying the derivatives held by subsidiaries of the group of this bank.

Table 3 provides an overview of the notional amount (in billion euro) of derivatives, by asset class, at the end of 2022. The values are rescaled for confidentiality reasons: for each bank, the FINREP value is equal to 100. Interest rate derivatives are the most relevant. Without considering bank3, the reported total notional values of the three different sources are similar. The differences by asset class could be due to different classification criteria between EMIR and the other two sources.

Overall, it seems that the total notional amounts of derivatives held by the main Italian banks obtained by using different sources of information are reasonably close to each other. The main differences are therefore related to banks characterised by a high degree of complexity.

EMIR, FINREP and supervisory and statistical reports data as of December 31, 2022

Table 3

	CO	CR	CU	EQ	IR	OTHR	TOT
bank1	-	0.0	3.0	-	114.6	-	117.6
	-	0.0	4.8	0.0	95.2	-	100.0
	-	0.0	4.7	-	124.6	-	129.4
bank2	0.4	0.1	1.0	6.0	124.3	-	131.8
	0.3	1.4	1.3	4.6	92.4	-	100.0
	0.7	1.5	2.7	3.4	98.9	-	107.2
bank3	1.0	0.2	1.3	1.4	41.9	0.0	45.9
	0.5	0.4	6.1	1.7	91.2	0.2	100.0
	0.1	0.0	1.9	0.2	25.2	-	27.4
bank4	0.2	4.3	9.2	1.9	91.6	-	107.3
	0.2	4.3	5.5	1.1	88.9	-	100.0
	0.2	4.3	5.3	1.1	89.9	0.0	100.8
bank5	0.7	0.0	24.8	7.3	71.4	-	104.4
	0.1	0.0	22.2	6.2	71.4	-	100.0
	0.1	0.0	22.6	9.1	90.3	-	122.2
bank6	0.0	17.5	3.2	8.4	63.2	-	92.3
	0.6	17.2	5.8	8.4	68.1	-	100.0
	2.1	29.9	5.2	9.2	69.9	0.0	116.3

Source: EMIR (first row), FINREP (second row) and supervisory and statistical reports (Matrice dei conti, third row) data available at the Banca d'Italia. The total notional of derivatives contracts is reported. All data are rescaled so that, for each bank, the FINREP value is equal to 100 at the Q4 2022. Asset classes: commodities (CO), credit (CR), currencies (CU), equities (EQ), interest rates (IR), other (OTHR).

3.6 Market statistics

Following the data processing framework defined in Section 3.3 we compute market statistics based on state reports.

Figure 9 shows the notional of outstanding derivatives contracts by asset class and contract type. We evaluate the monthly averages of daily information on the cleaned and deduplicated data. Figures are lower than those reported in ESMA (2021), as Banca d'Italia has access to only a portion of EMIR data. However, Banca d'Italia can also access information not involving Italian counterparties (see Section 3.1).

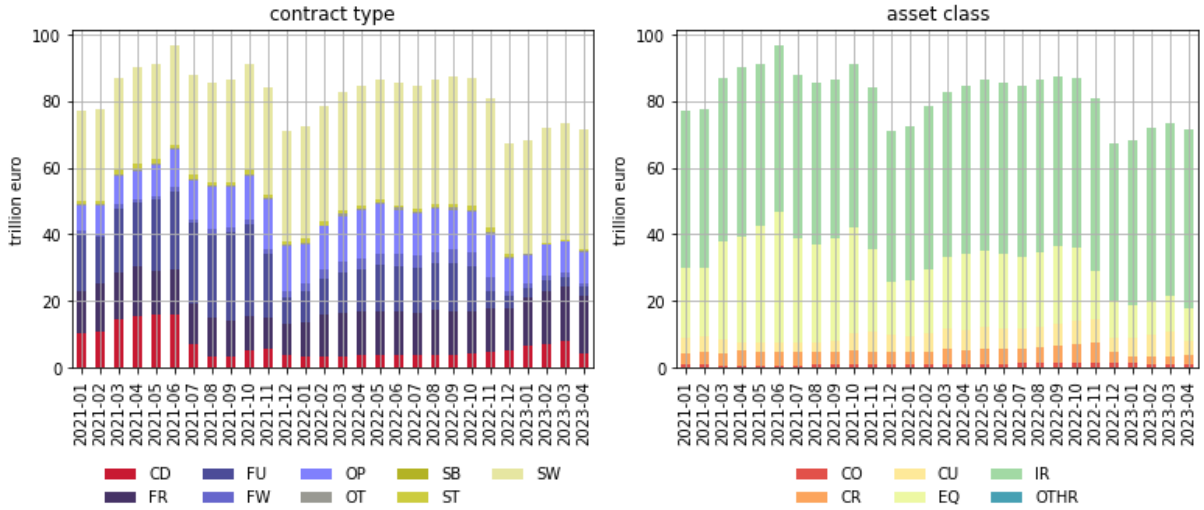


Figure 9: Trade states notional in euro by asset class and contract type. Source: EMIR deduplicated data available at the Banca d'Italia. Monthly averages of daily data. Asset classes: commodities (CO), credit (CR), currencies (CU), equities (EQ), interest rates (IR), other (OTHR). Contract types: contracts for differences (CD), forward rate agreements (FR), forwards (FW), futures (FU), options (OP), spread bets (SB), swaps (SW), swaptions (ST), other (OT).

The market size seems quite stable over the observation period, even if a decrease has been registered in the last few months. Regarding the contract types, excluding contracts for differences, their size does not vary significantly. As for other financial assets, the overall notional of the derivative portfolio of major financial intermediaries, which are the main EMIR contributors, varies slowly.

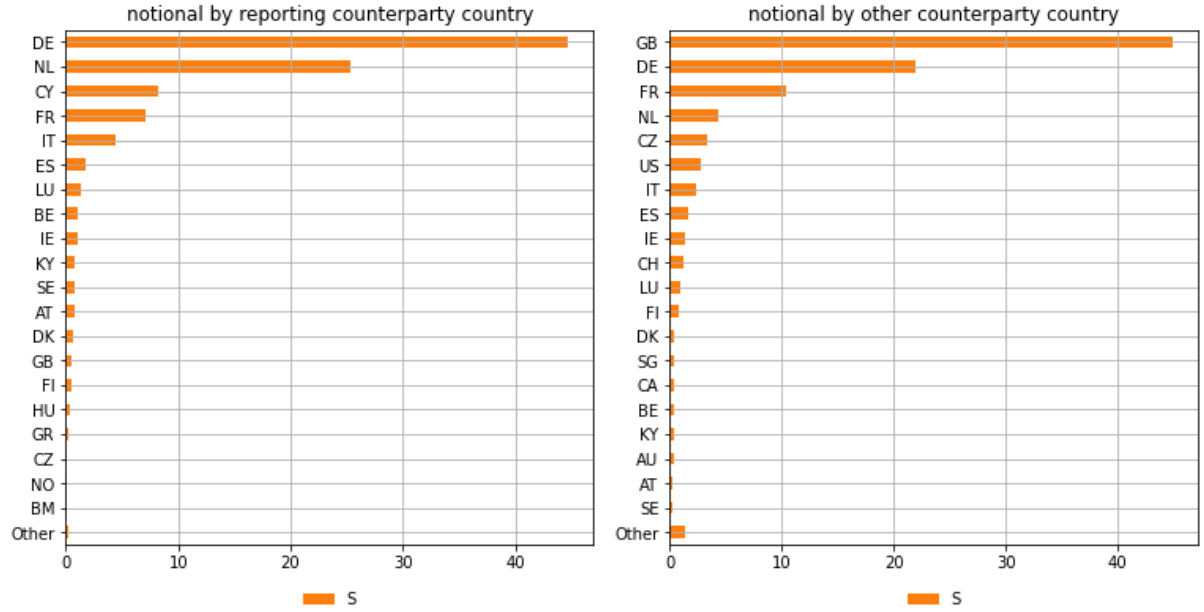


Figure 10: Overall data by country of the reporting (other) counterparty: states notional. Source: EMIR deduplicated data available at the Banca d'Italia. Percentage over the total states notional.

In terms of notional (see Figure 10), German entities are the major players in the market, both as reporting and as other counterparty. Italian entities are below five (three) per cent of the total states notional as reporting (other) counterparty. This may be due to the fact that major CCPs on derivatives are not Italian. The reporting of cleared derivatives involves multiple entities, including the CCP, and for this reason these entities are among the major counterparties both in terms of reports and of notional (see Appendix A.1). According to the data available at the Banca d'Italia, in Europe, the derivative market is concentrated in a few countries and British counterparties are among the major players.

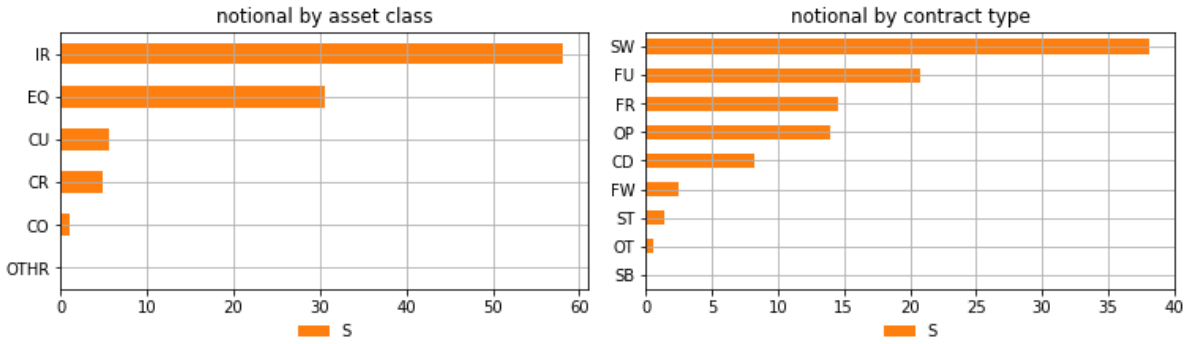


Figure 11: Overall data by contract characteristics: states notional. Source: EMIR deduplicated data available at the Banca d'Italia. Percentage over the total states notional. Asset classes: commodities (CO), credit (CR), currencies (CU), equities (EQ), interest rates (IR), other (OTHR). Contract types: contracts for differences (CD), forward rate agreements (FR), forwards (FW), futures (FU), options (OP), spread bets (SB), swaps (SW), swaptions (ST), other (OT).

Although equity derivatives account for more than 60 per cent of total submissions, these represent less than one-third of the overall notional (see Figure 11). More than half of the total notional refers to contracts on interest rates, mainly swaps. Credit and currency derivatives are almost equivalent in terms of overall notional, but not in terms of observations (see Figure 4). At least for the counterparties to which we have access to, commodity derivatives are a small fraction of the total notional on derivatives. Future contracts represent one-fifth of the market in terms of notional. Options and FRAs are slightly below 15 per cent. Contracts for differences are more than one-quarter of the total reports but account for less than ten per cent of the total notional. The notional of swaptions is small. This is also the case for other derivatives. Spread bets are negligible.

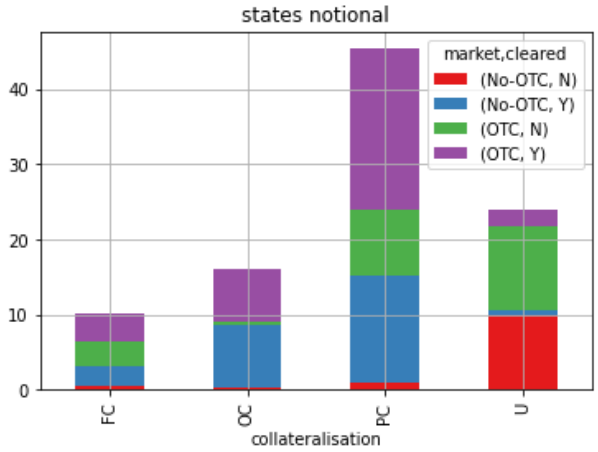


Figure 12: Overall data by market, clearing and collateral information: states notional. Source: EMIR deduplicated data available at the Banca d'Italia. Percentage over the total states notional. Collateralisation type: fully collateralised (FC), one way collateralised (OC), partially collateralised (PC), uncollateralised (UC).

In terms of notional, around 61 per cent of trade states refer to OTC contracts, and the remaining 39 to are not OTC. However, three fourth of the contracts have some form of collateralisation and two third are cleared (see Figure 12).

3.7 Some EMIR network snapshots

Another way to exploit EMIR data is by building network graphs to visualise interconnections between market participants. Figure 13 shows a network based on a centrality measure. First, we select only the

connections (edges) between counterparties greater than 500 million, then we choose only the entities (nodes) with more than 250 billion euro derivative exposures at the end of September 2021 (2022).

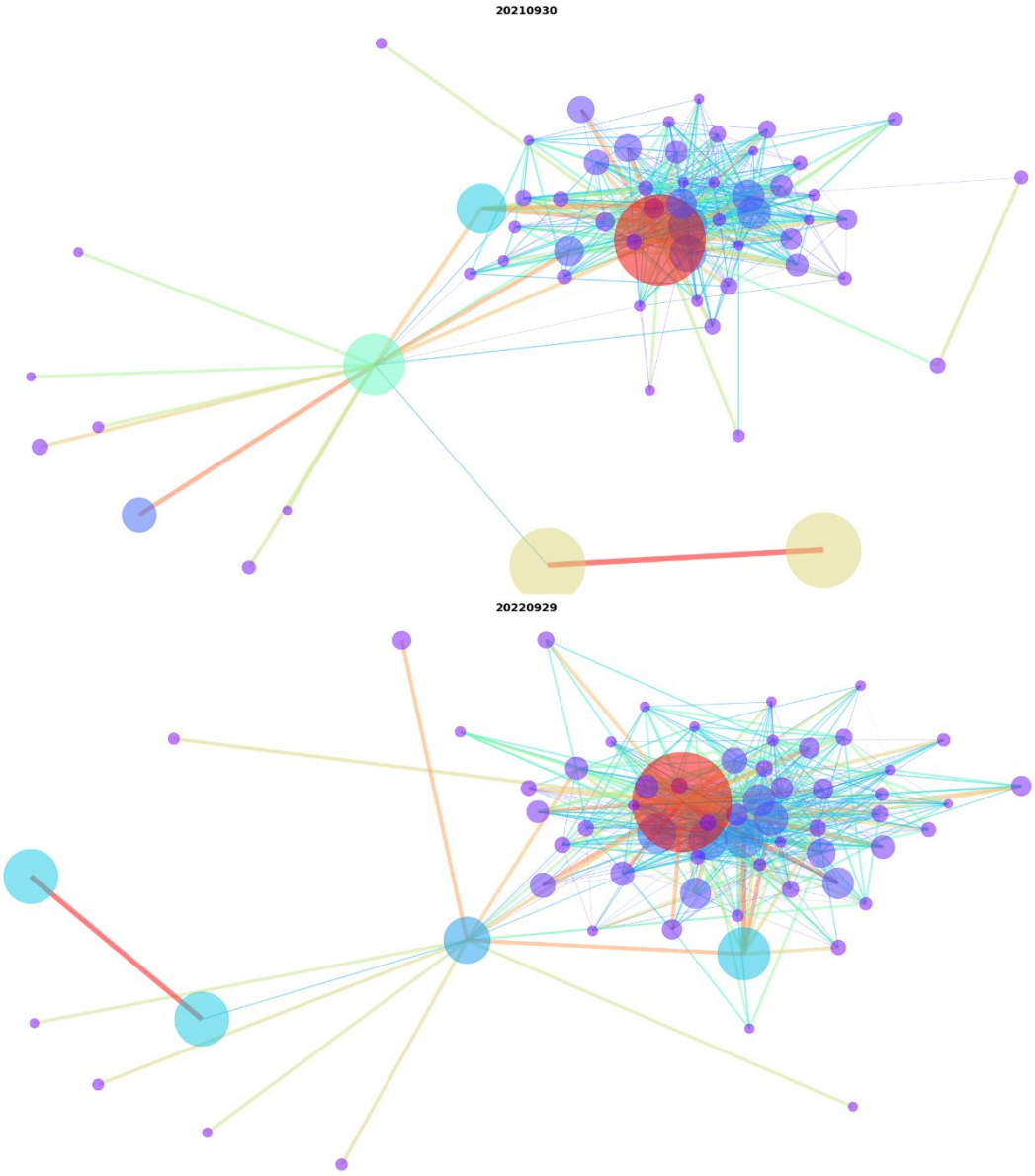


Figure 13: network of the biggest connections among major reporting counterparties at September 30, 2021 and September 29, 2022. Source: EMIR deduplicated data available at the Banca d'Italia.

The size and the colour of the edge represent how large the notional of the transactions is among entities, while the size and the colour of the circle represent the total notional of derivative transactions of the entity. We are considering different contract types and there is more than one CCP in the network. A CCP is the entity with the largest notional (the red circle) and is placed in the centre of the graph. The network varies slowly over time.

4. Conclusion

The complexity of EMIR is twofold: derivatives are complex instruments and there is a huge amount of data available to be organised and analysed. EMIR has required central banks, supervisors and regulators to adopt innovative data analysis approaches: "big data" are to be collected and processed with state-of-the-art technologies. This represents a challenge, but also an opportunity.

After briefly reviewing the main features of the EMIR regulation and the research leveraging this data source, we describe the journey of the data from the reporting entities to the data users and we illustrate the EMIR data available at the Banca d'Italia, with a look towards financial stability analysis and research. We provide a data overview and present a data processing framework to deal with data quality issues on the notional value and to define the identifier needed to compute initial and variation margins from data on collateral portfolios. Similar approaches could be considered to solve possible issues regarding the contract value. Finally, we provide some market statistics and an example of how a simple EMIR network could be visualised.

References

Abad, J., Aldasoro, I., Aymanns, C., D'Errico, M., Rousová, L. F., Hoffmann, P., Langfield, M., Neychev, M., Roukny, T. (2016). Shedding light on dark markets: First insights from the new EU-wide OTC derivatives dataset. ESRB Occasional Paper Series, 11.

Abbassi P., and Bräuning, F. (2021). Demand effects in the FX forward market: Micro evidence from banks' dollar hedging. *The Review of Financial Studies*, 34(9), 4177-4215.

Abraham, L. (2020). Competition analysis on the over-the-counter credit default swap market. *arXiv preprint arXiv:2012.01883*.

Ascolese, M., Molino, A., Skrzypczynski, G., Cerniauskas, J. and Pérez-Duarte, S. (2017), Euro-area derivatives markets: structure, dynamics and challenges, *IFC-National Bank of Belgium Workshop on* "Data needs and Statistics compilation for macroprudential analysis"

Bianchi, B. (2021). Cross-border credit derivatives linkages. ESRB: Working Paper Series 2021/115.

Bardoscia, M., Bianconi, G., and Ferrara, G. (2019). Multiplex network analysis of the UK over-the-counter derivatives market. *International Journal of Finance & Economics*, 24(4), 1520-1544.

Bardoscia, M., Ferrara, G., Vause, N., and Yoganayagam, M. (2021). Simulating liquidity stress in the derivatives market. *Journal of Economic Dynamics and Control*, *133*, 104215.

Bellia, M., Girardi, G., Panzica, R., Pelizzon, L., and Peltonen, T. A. (2019). The demand for central clearing: to clear or not to clear, that is the question. SAFE Working Paper, 193.

Bias, D., Guagliano, C., Haferkorn, M., Haimanna, M., and Kaserera, C.. Mutual funds and derivatives: Evidence from linked fund-trade data. *Preprint*.

Bouveret, A. and Haferkorn, M. (2022). Leverage and derivatives – the case of Archegos. ESMA Report on Trends, Risks and Vulnerabilities Risk Analysis.

Cielinska, O., Joseph, A., Shreyas, U., Tanner, J., and Vasios, M. (2017). Gauging market dynamics using trade repository data: the case of the Swiss franc de-pegging. Bank of England Financial Stability Paper, 41.

Cominetta, M., Grill, M., and Jukonis, A. (2019). Investigating initial margin procyclicality and corrective tools using EMIR data. ECB *Macroprudential Bulletin*, *9*.

Dalla Fontana, S., Holz auf der Heide, M., Pelizzon, L., and Scheicher, M. (2019). The anatomy of the euro area interest rate swap market. ECB Working Paper Series, 2242.

D'Errico, M., and Roukny, T. (2021). Compressing over-the-counter markets. *Operations Research*, 69(6), 1660-1679.

ESMA (2021). EU Derivatives Markets, ESMA Annual Statistical Report 2021.

ESMA (2022a). EMIR and SFTR data quality report 2021.

ESMA (2022b). Guidelines for reporting under EMIR.

ESMA (2023). 2022 report on quality and use of transaction data.

ESRB (2020). Mitigating the procyclicality of margins and haircuts in derivatives markets and securities financing transactions. ESRB Expert Group on the Macroprudential Use of Margins and Haircuts.

ESRB (2022). ESRB's view regarding data quality issues and risks for financial stability.

Fiedor, P., and Killeen, N. (2021). Securitisation special purpose entities, bank sponsors and derivatives. *Journal of International Financial Markets, Institutions and Money*, 75, 101452.

Fiedor, P., Lapschies, S., and Országhová, L. (2017). Networks of counterparties in the centrally cleared EU-wide interest rate derivatives market. ESRB Working Paper Series, 54.

Grothe, M., Pancost, N. A., and Tompaidis, S. (2021). Empirical analysis of collateral at central counterparties. ESRB Working Paper Series, 131.

Kahros, A., Pioli, A., Carraro, T., Gravanis, M., & Vacirca, F. (2021). A Descriptive Analysis of the Client Clearing Network in the European Derivatives Landscape. *Journal of Financial Market Infrastructures*, 9(1).

Kenny, O., Killeen, N., and Moloney, K. (2016). Network analysis using EMIR credit default swap data: micro-level evidence from Irish-domiciled special purpose vehicles (SPVs). *IFC Bulletin*, 41.

Joseph, A., & Vasios, M. (2022). OTC Microstructure in a period of stress: A Multi-layered network approach. *Journal of Banking & Finance*, 138, 106400.

Jukonis. A. (2022). Evaluating market risk from leveraged derivative exposures. ECB Working Paper Series, No. 2722.

Lenoci, F. D., & Letizia, E. (2021). Classifying Counterparty Sector in EMIR Data. In *Data Science for Economics and Finance* (pp. 117-143). Springer, Cham.

Levels, A., van Stralen, R., Kroon, S., and van Lelyveld, I. (2018). CDS market structure and risk flows: the Dutch case. DNB Working Paper, n. 592.

Mazzacurati, J., Guagliano, C., and Spolaore, A. (2021). Funds and single-name CDS: Hedging or Trading, ESMA Working Paper.

Pedro, S. A. B. (2020). An assessment of the portuguese OTC market network structure under EMIR-Bringing light to the portuguese OTC CDS Market. Master thesis.

Pérez-Duarte, S., and Skrzypczynski, G. (2019). Two is company, three's a crowd: automated pairing and matching of two-sided reporting in EMIR derivatives' data. Ninth IFC Conference on "Are post-crisis statistical initiatives completed?".

Rosati, S., & Vacirca, F. (2020). Interdependencies in the euro area derivatives clearing network: A multi-layer network approach. *Journal of Network Theory in Finance* 5(2), 1–41.

Schroeder, F., Lepone, A., Leung, H., and Satchell, S. (2020). Flash crash in an OTC market: trading behaviour of agents in times of market stress. *The European Journal of Finance*, 26(15), 1569-1589.

Ullersma, C., and van Lelyveld, I. (2022). Granular data offer new opportunities for stress testing. *Handbook of Financial Stress Testing*, ed. Farmer, J.D., Kleinnijenhuis, A.M., Schuermann, T., Wetzer, T., pp.185-207.

van Lelyveld, I. (2017), The use of derivatives trade repository data: possibilities and challenges, *IFC-National Bank of Belgium Workshop on* "Data needs and Statistics compilation for macroprudential analysis"

Appendix

A.1 Reporting of centrally cleared trades

Under EMIR, the reporting and the number of entities involved in a trade depend on whether the trade is uncleared (Figure A.1) or centrally cleared (Figure A.2). These are illustrative examples.

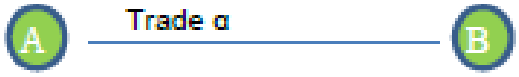


Figure A.1: uncleared derivatives. Source: Eurex documentation "Reporting by Eurex Clearing according to EMIR Article 9".

If a trade α between A and B is uncleared, there is a direct contractual relationship between the two counterparties (Figure A.1). If both counterparties have an obligation under EMIR, the bilateral trade must be reported by both sides, A (B) as the reporting counterparty of the trade and B (A) as the other counterparty. This is the standard case for over-the-counter (OTC) derivatives trades that are not cleared by a central counterparty (CCP).

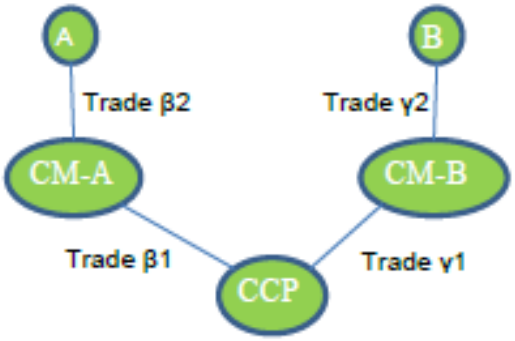


Figure A.2: exchange traded and OTC cleared derivatives. Source: Eurex documentation "Reporting by Eurex Clearing according to EMIR Article 9".

When instead a trade is centrally cleared, the reporting involves multiple entities. In the example of Figure A.2, A and B – both clients – are not clearing members of the CCP. Each of them refers to a clearing member, namely CM-A and CM-B. The latter can trade with the CCP. If the derivative is traded on an exchange (e.g. a futures contract), the orders of A and B are matched on the trading venue immediately, resulting in cleared trades. However, there is neither a direct link between A and B nor a direct contractual relationship between A, B and the CCP. In fact, in order for A (B) to have a centrally cleared trade, CM-A (CM-B) concludes a trade β1 (γ1) with the CCP, which therefore is the counterparty of the trade. CM-A (CM-B) also concludes a back-to-back trade β2 (γ2) with A (B). If all entities are subject to reporting obligations under EMIR, all trades (β1, β2, γ1 and γ2) are reported. All these trades share a field that identifies the presence and the name of the CCP.[31] The same reporting framework as shown in Figure A.2 also applies to OTC cleared derivatives, with the only difference being that there is no trading venue that matches the orders of A and B.

Finally, in the special case of clearing members that want their trades cleared by a CCP, the reporting is similar to that in Figure A.2, which means that the CCP appears as the counterparty of the clearing members in their trades.

[31] In EMIR reporting there exists a field (i.e. reporting tracking number) that links all trades (β1, β2, γ1 and γ2) together. Unfortunately, most of the time the information is missing and therefore a link between the trades cannot be established.

A.2 Data processing and confidentiality issues

At the Banca d'Italia the management of the EMIR database is carried out by the Statistical Data Collection and Processing Directorate (RES) with the help of the IT function. The EMIR database is stored in a Hadoop cluster and can be queried using Spark, a unified analytics engine for large-scale data processing. The data processing can be executed using the Jupyter notebook and implemented in Python (PySpark) or R (SparkR). As an alternative, users can implement their code in SQL. The results of the processing phase can be investigated on the same notebook or can be exported to be analysed through the most common econometric or statistical software.

The users should be aware that EMIR information must be treated as confidential according to art. 83 of the EMIR. Furthermore, this archive is subject to Regulation (EU) No 2016/679 concerning the protection of individuals in relation to the processing of personal data as the possibility that this archive may contain personal data is not excluded, although this is not required by law. Therefore, those who want to access EMIR information must comply with the constraints imposed by the regulations and provisions relating to the protection of personal data.

In particular:

- the users are subject to official secrecy pursuant to art. 7 of the Consolidated Law on Banking and Credit (Legislative Decree no. 385 of 1 September 1993);
- access can only take place for the pursuit of the objectives to perform an institutional function;
- no confidential information can be provided to unauthorised third parties or to external subjects or bodies;
- the data used in the context of analyses, research works and publications can only be disclosed in aggregate form, in order to avoid the possibility that the parties involved in the operations are identified.

A.3 EMIR introspection for the adjustment of GLEIF data

Entities in EMIR are represented by their LEI code. To identify these entities and their group, EMIR data are enriched with the information provided by the Global Legal Entity Identifier Foundation (GLEIF). GLEIF information includes, among other things, the legal name of the entities, the legal address and the group structure (direct and ultimate parent company), if any. The key of the enrichment is the LEI code.

Since we noticed some errors in GLEIF data, particularly on the group structure, we tried to adjust them by leveraging the EMIR intragroup field. In this field, each reporting entity can select 'Y' for intragroup transaction and 'N' for non-intragroup transactions. This means in practice that, given a pair of counterparties, if 'Y' is reported both the reporting counterparty and other counterparty should belong to the same group[32]. In theory, is it possible to identify the group structure (i.e. all LEI codes of entities belonging to the same group) by looking directly on intragroup trades and without considering GLEIF data. We refer to this approach as *EMIR introspection*. However, since usually not all entities in a group are involved in derivative trading, to obtain a more reliable group structure, we consider both the information coming from EMIR and those coming from GLEIF. Finally, for all intermediaries supervised by the Banca d'Italia, we consider the information on the group structure as collected for supervisory purposes.

[32] We noticed some errors in the reporting of this field and for this reason, if an entity results with more than one parent company, we decided to select as parent only the company having with the given entity the largest number of intragroup transaction ('Y'). All possible ultimate parent companies are preliminary selected on GLEIF. Additionally, we exclude all investment funds from the group structure, even if a 'Y' is reported. The list of investment funds is retrieved from the ECB web-site.

EMIR data for financial stability analysis and research

Dario Ruzzi
(joint work with M.L. Bianchi and B. Sorvillo)
Banca d'Italia
Financial Stability Directorate

3rd IFC and Bank of Italy Workshop on
Data Science in Central Banking: Enhancing the access to and sharing of data
October 18, 2023

Outline

- EMIR data
- access to data
- use of data
- data cleaning (including *deduplication*)
- data quality on notional values
- data quality on the value of contract
- research works on EMIR data

EMIR data in a snapshot







EMIR data

The European Market Infrastructure Regulation (EMIR), adopted in 2012, requires all EU counterparties to report their **derivative transactions** to trade repositories (TRs).

The information collected in EMIR can be categorized as follows:

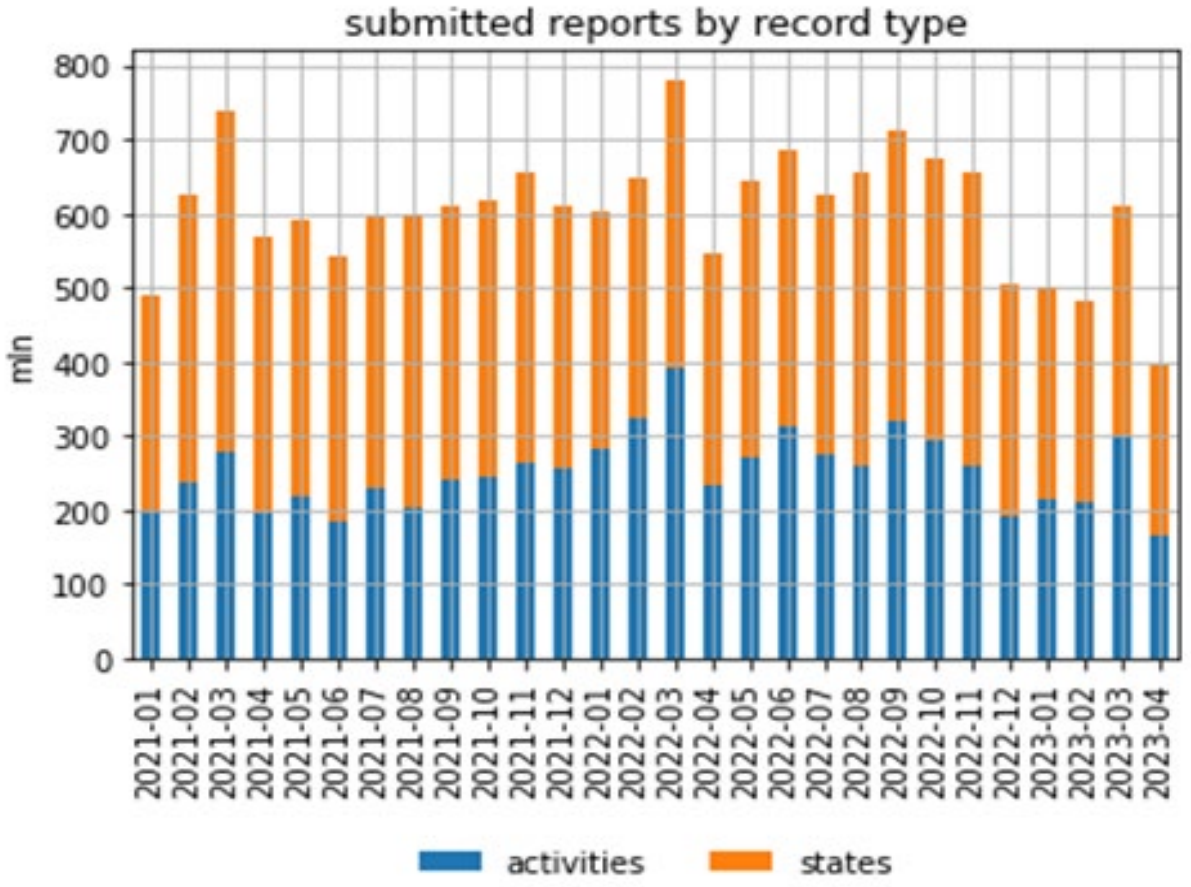
- *trade details*: detailed information related to derivative transactions, including trade identifiers, dates and times of trade execution, product type, underlying assets, notional amounts, contract maturity dates, contract characteristics;
- *counterparty information*: counterparties involved in derivative transactions, including the legal entity identifier (LEI) and their roles in the transaction;
- *valuation and margin data*: valuation of derivative contracts and of the related margins on a daily basis

EMIR data

The reports are of two types:

1) trade activity reports or ***activities*** (*flows*) containing all reports sent the previous day;
2) trade state reports or ***states*** (*outstanding positions)*, containing all pending trades at the end of the reporting day.

For risk analyses purposes “states” are usually considered.



Access to data











- Starting from 29 december 2020, we collect EMIR transaction data from ESMA TRACE (4 trade repositories).
- Around 600 million observations each month (states + activities)
- Data available at *t*+2 (*t*+3)
- EMIR users work with «enriched» data and/or other «big» dataframes (Gleif, Banca d'Italia internal databases on securities or entities, ESMA FIRDS)
- cluster Hadoop/Spark (go live: July 2021) : PySpark (*), SparkR, SAS or SQL
- PySpark via Jupyter «is the way»

Use of data

- Users in various Directorates, mostly in *Financial Stability*
- Deep dives and internal notes on specific counterparties, underlyings or products
- Banca d'Italia **product intervention power** (MiFIR and MiFID II) to assess the financial stability risks associated with financial instruments, structured deposits and related financial activities/practices
- Global and cross country systemic risk monitoring (FSB and NCAs)
- Various research projects (more details in the following slides)
- … market risk sensitivity analysis

Some data quality issues

- **Reporting errors:** terminated transactions, unresonable notional or contract value, wrong sector and intragroup fields
- **Data gaps:** some TRs do not send us all data as per the regulation
- **Inconsistencies**: side and valuation provided by counterparties of the same trade
- **Fair value:** not always representing the position value (e.g. futures)
- **Initial and variation margins (IM and VM)**:
  - reported at *collateral portfolio* level (do we see all trade in a collateral portfolio?)
  - reporting requirements for posted vs received depend on collateralization type
- **Transaction chain**: it is extremely difficult to reconstruct the entire transaction chain of a contract (centrally cleared trades): the report tracking number does not help much

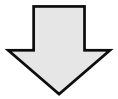
Data cleaning: our steps

1. remove implausible notional < €1000 and > €50 bln (-38% trades)
2. exclude inactive trades (-7.6%)
3. exclude trades where buyer and seller are not clear (≈ 0%)
4. remove trades where both counterparties have no LEI (≈ 0%)
5. exclude intragroup positions based on Gleif, Banca d'Italia supervisory registers, and ECB list of investment funds (-5.7%)
6. keep only one report per date-counterparty-trade (≈ 0%)

Deduplication: our approach

Deduplication of two-sided reported trades is carried out according to a pecking-order criterion, centered around reliability and number of reports:

1) CCPs (more reliable than their clearing members, CMs)
2) CMs (more reliable than their clients)
3) All other counterparties



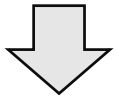
cleaned deduplicated data

ready for the analysis

Dealing with collateral portfolios

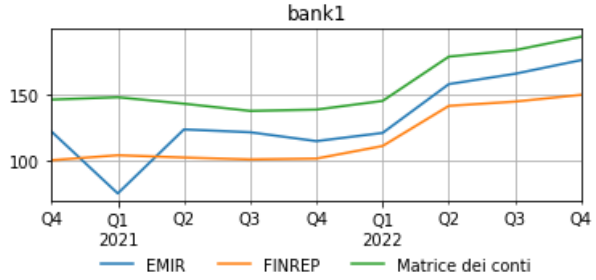
Margins are reported at the collateral portfolio level:

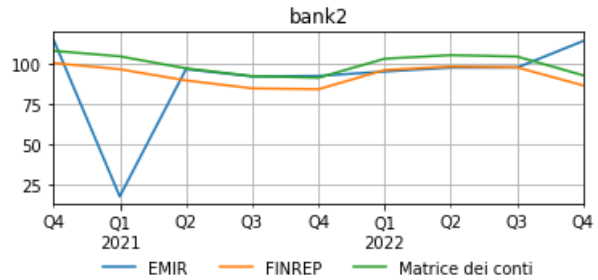
1) no unique portfolio identifier shared by both counterparties
2) the content of portfolios may differ between counterparties
3) some collateralization types (e.g. partial) do not require reporting of received margins

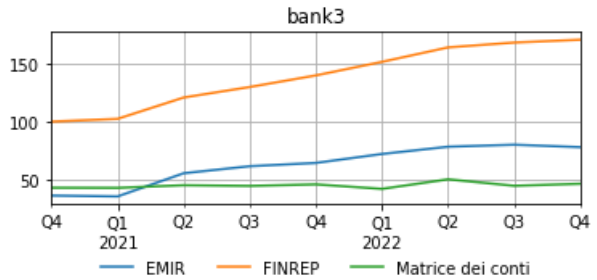


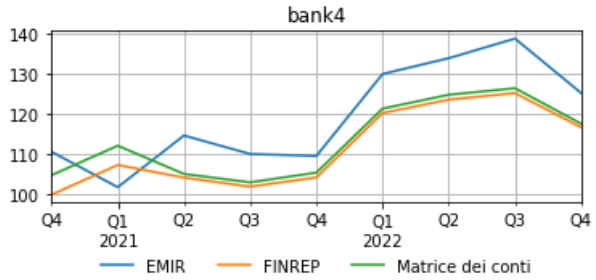
we NEED the double reporting here!

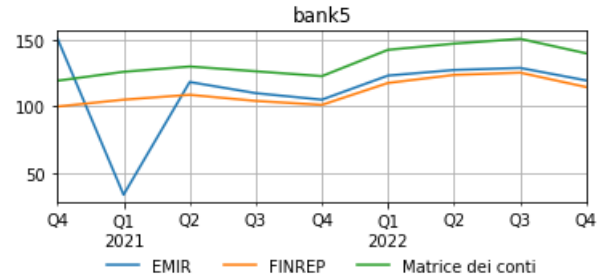
Data quality on notional values: EMIR vs Finrep

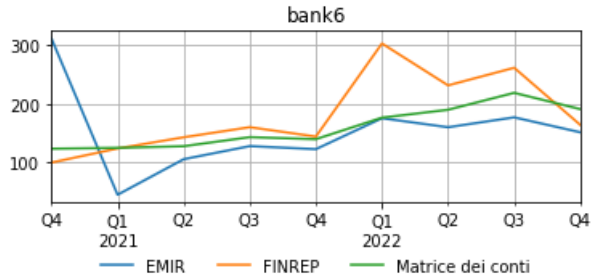






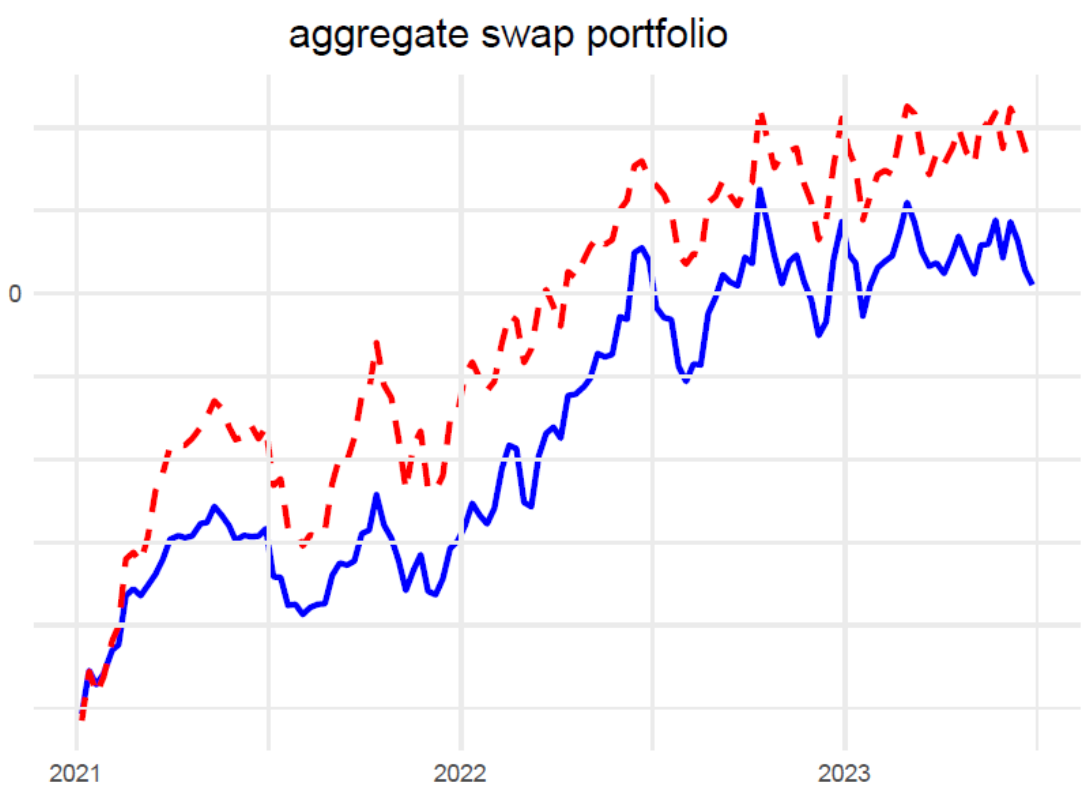




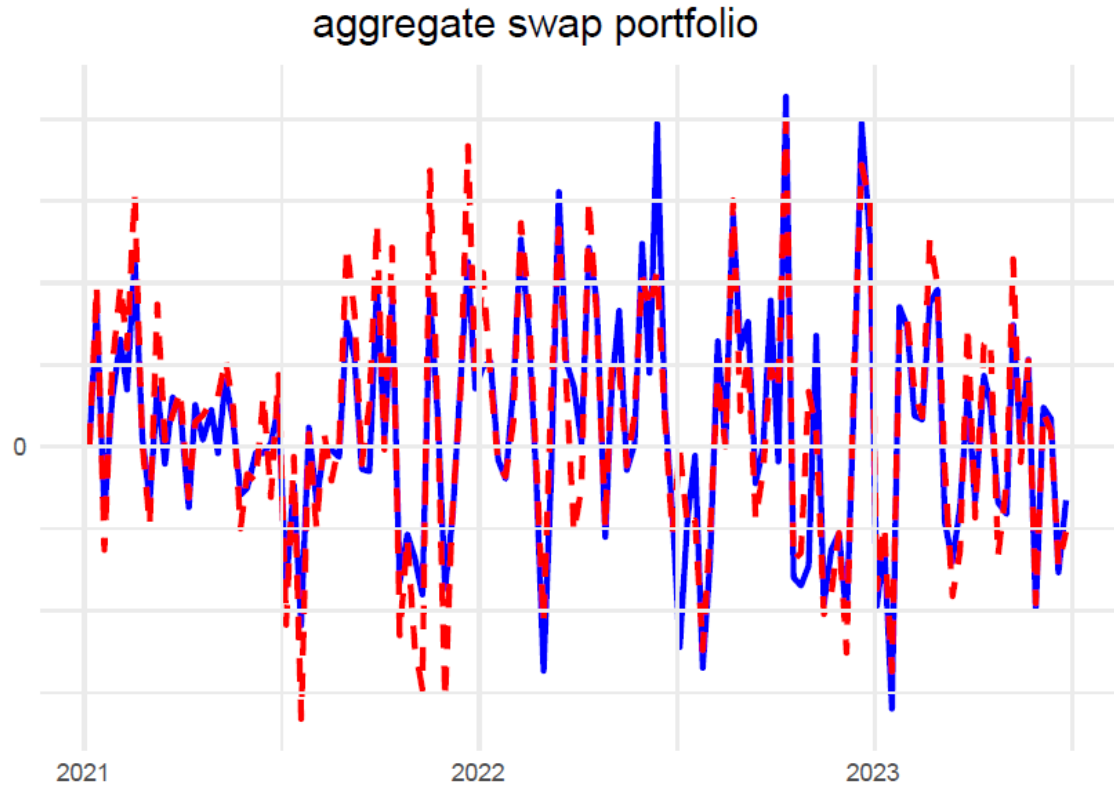


- Source: cleaned deduplicated EMIR data available at the Banca d'Italia, FINREP and supervisory and statistical reports (Matrice dei conti). All data are rescaled so that, for each bank, the FINREP value is equal to 100 at the Q4 2020.

Data quality on the value of contract: IR swaps

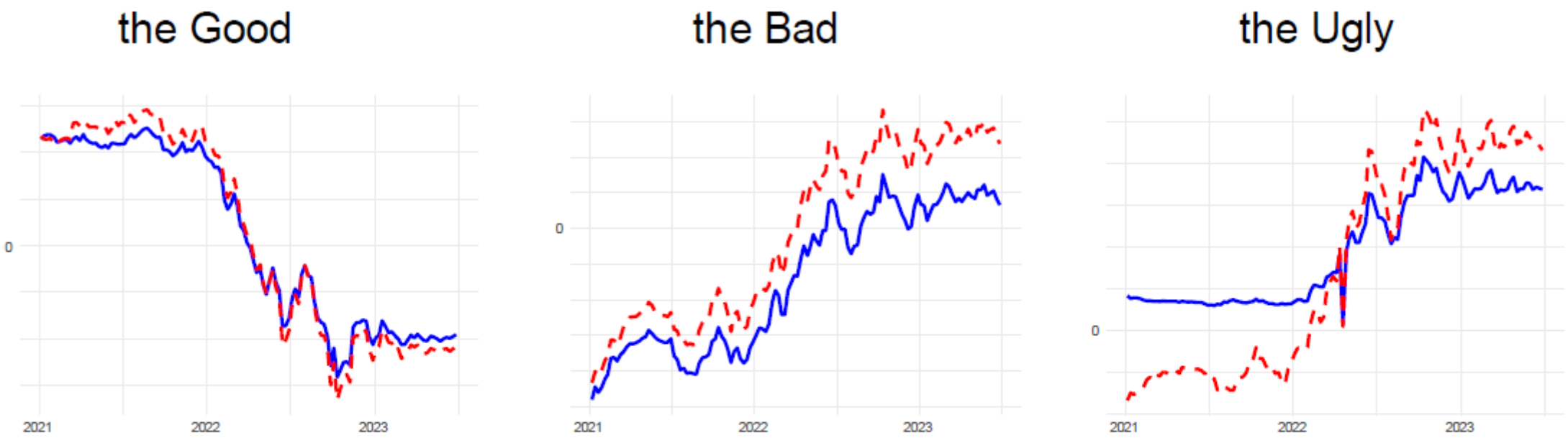


swap portfolio observed in EMIR and implied by the model



swap portfolio observed in EMIR and implied by the model

Data quality on the value of contract: IR swaps



swap portfolio observed in EMIR and implied by the model

Research works on EMIR data at Banca d'Italia

- *Shifting the yield curve for Italian banks*: valuation of interest rate derivative contracts and risk sensitivity analysis using granular information on these derivatives and fixed-rate bonds (*available upon request*)
- *CO_2 emission derivatives In Italy*: analysis of the market of CO_2 emission permits derivatives in Italy (*available upon request*)
- *EMIR networks*: investigation of the structure of EMIR networks (notional and margins), their dynamics over time and the main risk drivers (*work in progress*)
- *Interest rate futures*: analysis on sovereign bond futures (*work in progress*)
- ... a lot of work

Thank you for your attention

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Unveiling the interconnectedness of banks in payment system: methodology, utilization, and data governance considerations[1]

Renardi Ardiya Bimantoro, Irfan Sampe and
Mohammad Khoyrul Hidayat,
Bank Indonesia

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Unveiling the interconnectedness of banks in Payment System - Methodology, Utilization, and Data Governance Considerations

Renardi Ardiya Bimantoro[1], Irfan Sampe[2], Mohammad Khoyrul Hidayat [3],

Abstract

The global financial crisis was a reminder of the importance of understanding the interconnectedness of financial systems. Financial networks often have a small number of main players that are highly connected to each other. Nevertheless, disruptions to one of these players could expose a risk to the rest of the network. Understanding how liquidity shocks propagate throughout the interbank market relies on a network topology that can be constructed using payment system data. Several studies have shown that interbank payment systems and interbank markets tend to have a core-periphery structure, with banks operating in a hierarchy. The network can be exploited to prioritize liquidity monitoring on banks that are identified as core and to analyse contagion effect. On the other hand, while understanding the structure of financial networks is critical in systemic risk assessment, the network structure utilizes the bank's individual data that could pose a data privacy or security risk during its generation or dissemination to users. Therefore, various protection measures need to be taken to ensure the confidentiality of the bank's data. This paper aims to capture the network topology in Bank Indonesia's Real Time Gross Settlement (BI-RTGS) and examine if Indonesia's interbank payment system has a core-periphery structure while ensuring its usage adheres to governance principles.

Keywords: interconnectedness, systemic risk, payment system, core-periphery, support vector machine (SVM), big data analytics, data governance, data privacy

JEL classification: D85, G21, K20

[1] Data Innovation & Digitalization Department – Bank Indonesia; e-mail: renardi_ardiya@bi.go.id

[2] Data Innovation & Digitalization Department – Bank Indonesia; e-mail: irfans@bi.go.id

[3] Statistics Department – Bank Indonesia; e-mail: moh_khoyrul@bi.go.id

The views expressed here are those of the authors and do not necessarily reflect the views of Bank Indonesia

Contents

1. Background 3

2. Literature Review 4

   2.1 Empirical Evidence of Contagion Effect in Financial Market 4

   2.2 Core-Periphery Structure in Financial Network 5

3. Methodology 6

   3.1 Data 6

   3.2 Workflow 7

      3.2.1 Data Preprocessing 8

      3.2.2 Data Extraction 8

4. Result and Analysis 10

5. Conclusion and Future Work 11

   5.1 Conclusion 11

   5.2 Future Work 11

References 12

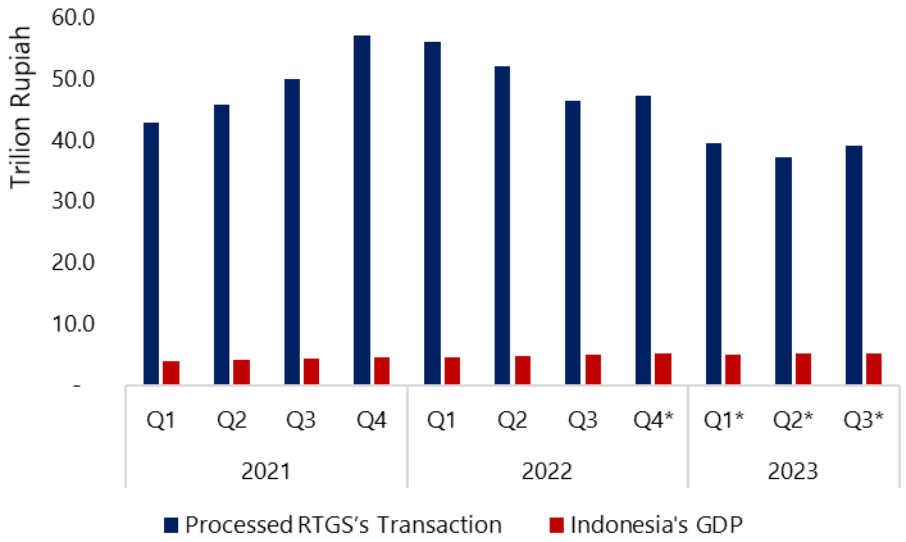
1. Background

Currently, the majority of economic transactions are carried out through the fund transfer payment method. This comprises a wide range of transaction methods, ranging from retail payments such as card-based transactions, QR-based transactions, and internet-based transactions to large-scale gross settlements. Contrary to retail payments, which are usually of smaller value with high frequency, large-scale gross settlements are commonly employed for high-value transactions, albeit with a lower frequency. These settlements play a vital role in facilitating the flow of funds across the economy.

Bank Indonesia, in collaboration with the government, consistently endeavors to develop suitable policies to achieve its three primary objectives: monetary stability, financial system stability, and payment system stability. Bank Indonesia has implemented and operated the BI-RTGS payment system since November 17, 2000. Due to its coverage of about 90% of total payment transactions in Indonesia, BI-RTGS has been designated as a systemically important payment system.

Processed RTGS's Transaction Compared to Indonesia GDP in 2021-2023

Figure 1



Source: Bank Indonesia & BPS

*) Temporary Number

The significant role of the large value payment system (BI-RTGS) in maintaining financial system stability and mitigating systemic risk, makes it crucial for the central bank to thoroughly analyse the structure of the interconnections among participants in the payment system. Furthermore, the global financial crisis served as a reminder of the significance of comprehending the interdependence of financial institutions. Hence, it is imperative to ascertain the configuration of connections among actors (banks) in the BI-RTGS payment system through network analysis.

The network structure in the interbank market can be described as tiered, with a cluster of highly interconnected institutions that also serve as intermediates for other banks with fewer connections. Multiple studies have demonstrated that

interbank markets typically exhibit a core-periphery structure, wherein banks function within a hierarchical system. Any disruption to one of these players could potentially pose a risk to the entire network. The comprehension of how liquidity shocks spread across the interbank market depends on the construction of a network topology utilizing payment system data. Utilizing the network can be advantageous in prioritizing liquidity monitoring at institutions that are recognized as core and in analysing the contagion effect.

However, the utilization of the bank's private data in the network structure could potentially create data privacy or security concerns when it is generated or shared with users. Hence, it is necessary to implement security measures to guarantee the confidentiality of the bank's data. The objective of this article is to analyse the network structure of Bank Indonesia's Real Time Gross Settlement (BI-RTGS) system and determine if the interbank market in Indonesia exhibits a core-periphery structure, while also ensuring compliance with governance standards.

2. Literature Review

2.1 Empirical Evidence of Contagion Effect in Financial Market

The global financial crisis served as a reminder of the significance of comprehending the danger of contagion when major institutions are unable to meet their payment obligations. Kaufman (1992) defines contagion as the transmission of shock effects from one or more firms to others. This issue is more prevalent in the banking business than in other industries. Financial instability and contagion in the financial system can arise from a bankruptcy or other force majeure event (Heijmans & Wendt, 2023).

In their study, Zhang (2021) analyse the influence of bank liquidity production on systemic risk and its varying effect on network interconnectedness. The authors of this study put up three hypotheses: (1) The generation of excessive bank liquidity raises the level of systemic risk. (2) Internal and external liquidity creation have distinct impacts on systemic risk. (3) The influence of liquidity creation on systemic risk intensifies as the interconnectivity of the bank network rises. The study found that the interconnection between banks is crucial in determining the risk of spill over and contagion. This highlights the need for monitoring the financial network within the regulatory framework.

Another supporting evidence of contagion found in banking markets (Ballester, 2016). It is identified through an initial indication of increased co-movement in Credit Default Swap (CDS) returns. The significant rise in prevalence should be seen as a preliminary indication, prompting authorities to act. The contagion occurred in multiple waves, starting in July 2007, and the financial and eurozone crises were separate occurrences. During the financial crisis, contagion was characterized by a systematic aspect. However, with the eurozone crisis, the idiosyncratic component became more prominent.

By using net directional return spillover metrics, the study was able to identify a set of banks in certain countries that act as both sources and recipients of contagion. US banks seem to have been net transmitters throughout the period from 2007 to 2009. Amidst the eurozone crisis, banks in euro-peripheral nations played a significant

role in transmitting idiosyncratic spillovers, while the issues in the eurozone had less impact on US banks. The variations in susceptibility to contagion within Europe, and even within the eurozone, are striking, with the eurozone periphery being more susceptible to systematic transmission.

The examination of the spread of contagion inside the euro-periphery portfolio verifies the presence of elevated levels of the systematic contagion index. Spanish banks transmit both systematic and idiosyncratic risk spillovers to Greek banks, which act as net recipients of these risks.

Collectively, these results suggest that the interconnections between financial markets across countries play a crucial role in the propagation of shocks. This is evident from the fact that the overall spillover index consistently shows a higher contribution from systematic factors compared to idiosyncratic factors. Nevertheless, there have been certain cases during the recent eurozone crisis where unique and unpredictable events have resulted in the transmission of negative effects to other banking markets. This appears to be associated with instances of bank failure, such as the Lehman Brothers' collapse in September 2008, or heightened vulnerability to bank failure, as seen in Spanish banks during the latter two quarters of 2011.

2.2 Core-Periphery Structure in Financial Network

Central bank research conducted in several countries, including Germany, the Netherlands, Italy, the UK, and Korea, has found networks inside the interbank market and interbank payment system. The objective of this research is to comprehend the network structure that exists in these systems. These findings demonstrate that the observed network structure exhibits a hierarchical organization. The majority of the research focuses on the interbank tiering concept that Craig and Von Peter (2010) introduced. In their study, Craig and Peter (2010) suggest that the interbank payment system functions in a hierarchical manner, with banks at lower levels being connected solely through transaction centers known as core. This network configuration is an expansion of the core-periphery structure that Borgatti & Everett (1999) first proposed and which later served as a benchmark for other central banks in determining the layout of the interbank network.

In 2010, Craig and Peter utilized Borgatti and Everett's core-periphery structure approach to examine the network topology within an interbank market. Craig and Peter suggest that the network structure of the interbank market exhibits a hierarchical arrangement, sometimes known as tiering. The interbank tiering notion refers to the formation of a network structure that arises when a set of interconnected banks also serve as intermediates for other non-interconnected banks. The attributes of an ideal core-periphery model in an interbank market are defined as follows:

i. Top-tier banks (core) are interconnected with each other.

ii. Lower-tier banks (periphery) are not interconnected with each other.

iii. Top-tier banks (core) are connected with (some) lower-tier banks (periphery)

Typically, interbank transactions in a payment system are one-way, meaning that bank A can send a transaction to bank B, but the opposite may not be true. The payment system network is seen to be a directed network. Hence, the models suitable for discerning the network structure in the payment system are the asymmetric

discrete model (Craig & Peter, 2010) and the asymmetric continuous model (Boyd et al., 2010).

Bank Indonesia is conducting research to build a network analytics approach that utilizes the core-periphery model and machine learning algorithms. The objective is to discover the connectivity structure of participants (banks) in the BI-RTGS payment system. This can be utilized to enhance the efficacy of systemic impact assessments in the financial system. This methodology is anticipated to be beneficial in identifying banks classified as core banks, which are institutions that could have a substantial impact on the payment system. Additionally, there is a concern for data protection, emphasizing the need for secure handling and confidentiality of banking data in this analysis to prevent unauthorized access or data breaches.

Thus, this study is expected to answer the following research questions:

1. Does Indonesia's RTGS interbank payment system exhibit a core-periphery structure in its network topology?
2. What steps can be taken to keep individual bank data safe and private while studying Indonesia's bank network in the BI-RTGS system?

3. Methodology

3.1 Data

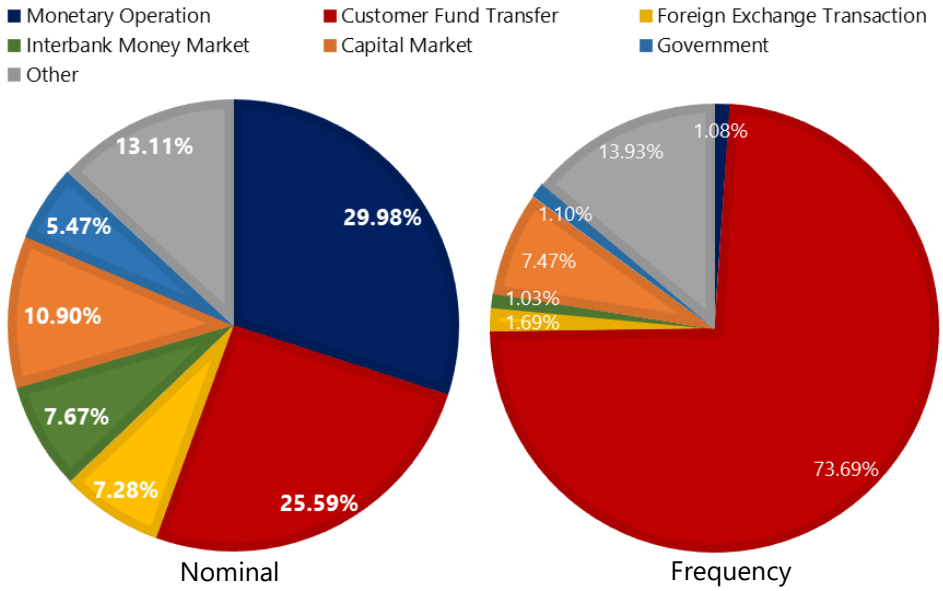
The study utilized data sourced from the Bank Indonesia Real Time Gross Settlement (BI-RTGS). BI-RTGS is a financial infrastructure managed by Bank Indonesia that facilitates high-value fund transfers, debit clearing, regular payments, and regular billing services (as stated in Regulation of Member of Board of Governors Number 24/5/PADG/2022). Since November 2000, the Real Time Gross Settlement (RTGS) system has had the capability to handle payment transactions exceeding Rp. 100,000,000.00 (one hundred million rupiah) in value.

The data included in this study encompasses RTGS transaction data spanning from January 2018 to December 2023. On average, the value of transactions through this service amounts to approximately 13,000 Trillion Rupiah per month. The Real Time Gross Settlement (RTGS) system encompasses various transaction types:

1. Monetary operations, which encompasses expenses and the duration of financial instruments.
2. Interbank Money Market, which encompasses the maturity of transactions inside this market.
3. Government transactions encompassing the issuance of SBN (Sovereign Bonds), redemption of SBN, SBN coupons, and other related activities.
4. Customer Fund Transfer
5. The capital market, which encompasses securities transactions.
6. Foreign exchange transactions

The share for each transaction type is shown in Figure 2, which shows that the largest share of transaction (±29%) is monetary operation, followed by customer fund transfer (±25). The data structure obtained from the RTGS system is as shown in Table 1.

Nominal and Frequency Share of RTGS Transactions Based on Transaction Type in November 2023 Figure 2



Source: BI-RTGS (processed)

RTGS Data Structure Table 1

No.	Data Field
1	SETTLEMENT TIME
2	SENDING PARTICIPANT
3	RECEIVING PARTICIPNAT
4	STATUS
5	TRANSACTION CODE
6	MESSAGE TYPE
7	CURRENCY CODE
8	AMOUNT (NOMINAL)
9	BLOCK4*

*) Block 4 information contains transaction detail related to costumer information.

Despite the BI-RTGS structure data containing quite information, particularly related to the individual details of financial institutions and their customers, our focus on data governance necessitates a prudent approach during data processing and analysis. We adhere to data minimization principle, only collecting and processing the minimum amount of data necessary to identify the core-periphery structure. In this study, we only use settlement time, sending/receiving bank, transaction code, status, and amount to reduce the risk of data breaches and ensures compliance with data protection regulations.

3.2 Workflow

To obtain the classification of core and periphery banks within a payment system network that consider not only the links between banks but also take into account the weight of those links, which is characterized by their transaction nominal and frequency, We propose a novel methodology that examines machine learning and

core periphery model. In general, the workflow consists of data preprocessing and extraction.

3.2.1 Data Preprocessing

The data preprocessing stage is carried out to prepare the raw RTGS fund transfer transaction data so that they can be further processed at the next stage. The process is as follows:

1. Filter out non-bank BI-RTGS participants.

To filter out non-bank BI-RTGS participants names, we cross-referencing the participant list with the official records of authorized bank participants. We then identify and eliminate any names that do not belong to banks. For example, switching company and bank indonesia account are excluded.

2. Map branch unit bank to their parent bank

In order to establish the relationship between branch unit banks and their parent banks, we examine the hierarchical arrangement of the banking system and align each branch unit bank with its parent bank using official records and organizational connections. We perceive the parent and child corporations as a single entity of a network structure.

3. Codify the Bank Name using Bank Code

Before processing and analyzing the data, we ensure that all bank identifiers are anonymized. We replace bank names or any other identifiable information with generic IDs or pseudonyms (e.g. Bank A, B, C). This approach prevents that individual banks cannot be directly identified from the data, protecting their privacy and ensuring compliance with data protection standards.

4. Data Aggregation & Matrix Representation

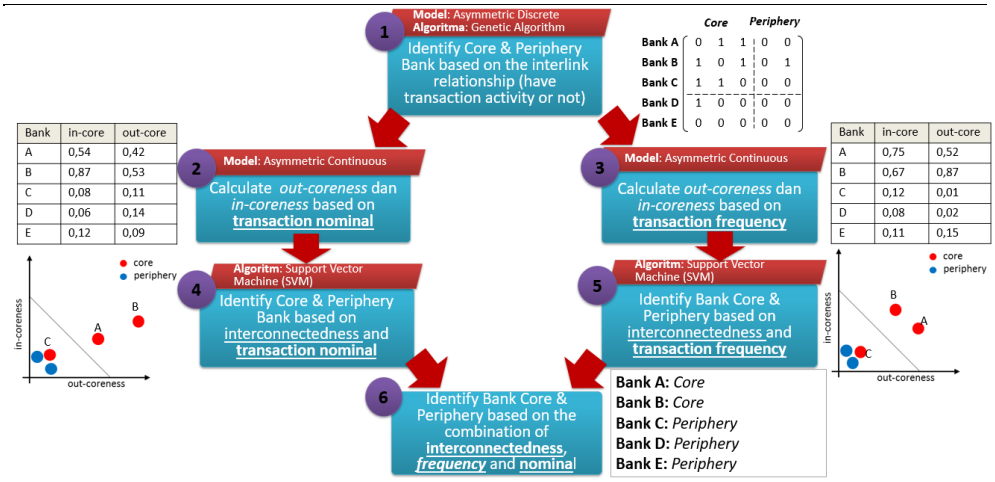
We consolidate the data monthly basis and represent the data in both discrete (binary, 1: connected and 0: not connected) and continuous matrices. This stage is essential for the subsequent data processing technique, in which we will employ the block modeling method on discrete matrices and the matrix decomposition method on continuous matrices.

3.2.2 Data Extraction

In order to incorporate both the quantity and intensity of connections between banks (as indicated by the transaction amount or frequency of transactions), this study aims to merge the concepts of the asymmetric discrete model and the asymmetric continuous model to establish a core-periphery classification. We employ a machine learning technique, specifically Support Vector Machine (SVM), to categorize banks into core and periphery based on the outcomes of both models. Figure 3 presents the structure for the core-periphery bank classification.

The Methodoloy of Core-Periphery Bank Classification

Figure 3



1. Assymmetric Discrete Model

During the initial stage, we employ the Asymmetric Discrete Model to determine the most favorable division of core-periphery banks in the BI-RTGS network. This division is based on the connection data, where 1 represents a link and 0 represents no connection. The model in this study is implemented using a Genetic Algorithm. The nominal transaction data is encoded as a binary transition matrix for each period and kind of transaction. By utilizing a Genetic Algorithm software, we can derive the best core-periphery partition. Borgatti (1999) states that the Genetic Algorithm outperforms other algorithms in identifying core-periphery partitions due to its ability to generate more resilient outcomes.

2. Assymmetric Continous Model

During phases 2 and 3, we employ the asymmetric continuous model to determine the out-coreness and in-coreness values for each bank in the BI-RTGS network. The input consists of a continuous transition matrix that represents the nominal data and frequency of transactions for each time and kind of transaction. Next, we ascertain the values of out-coreness and in-coreness that may generate a transition matrix resembling the original one with the least disparity, hence minimizing the loss in matrix reconstruction.

3. Machine Learning Model (SVM)

Subsequently, during phases 4 and 5, we employ a machine learning technique to categorize banks into core and periphery categories, utilizing the outcomes obtained from the discrete and continuous models. The SVM method is employed to implement this classification, utilizing the discrete model's outcomes as labels. The classification is based on the similarity between the out-coreness and in-coreness values of each core and periphery bank, taking into account nominal and transaction frequency.

4. Result and Analysis

In the context of the core-periphery structure, a bank is considered "core" if it is consistently designated as such every month throughout the observation period. According to the classification results, we have successfully identified a core-periphery structure in the BI-RTGS interbank payment system. This structure consists of a number of banks that have been consistently identified as the core over the whole observation period. Below are the findings from the identification process in BI-RTGS transactions.

Comparison of Core Periphery Classification Model — Table 2

		SVM	
	Classification	**Core**	**Periphery**
DISCRETE	**Core**	**10**	**41**
	Periphery	**0**	**54**

Among the many banks that participate in BI RTGS, there are 10 banks that are consistently categorized as core banks by SVM for a period of 24 months. According to Craig & Peter's (2010) core-periphery model, these 10 banks are considered the core banks in the BI-RTGS network. They are interconnected with each other and serve as intermediaries for other lower-tier banks, making them top-tier banks in the core-periphery structure. While the research did not discover a structure that precisely matches the ideal model proposed by Craig & Peter (such as the presence of periphery banks connected exclusively to other periphery banks), it did effectively identify a stable core-periphery structure throughout the observation period.

The Department of Financial System Surveillance (DSSK) at Bank Indonesia employs these identification results, which are displayed on our network analytics dashboard for supervisory purposes as a part of the Supervisory Technology Framework. They further analyse the results by conducting an assessment of intraday liquidity risk to evaluate how well the identified core banks manage their intraday liquidity. In addition, the application of core-periphery analysis in network mapping can be utilized to examine the spread of contagion.

Moreover, access to the dashboard, where the results are displayed, is strictly available only to users who have agreed to a non-disclosure agreement. We implement access controls to ensure that only authorized user can view and interpret the classification results. This not only safeguards data privacy but also prevents the potential for data misuse. Nevertheless, despite the benefits of this finding as an early warning tool to identify risks within the banking network, it is crucial to recognize that these classifications have significant financial or reputational implications for the banks involved. Consequently, reliance on this finding as the sole basis for decision-making is inadvisable. It is essential to consider the ethical implications of this result of classifying banks while conducting financial stability surveillance.

To strengthen surveillance in the financial sector, we also have develop a system for identifying the primary players (banks) in foreign exchange transactions. Bank Indonesia can systematically monitor the conduct and composition of foreign exchange market players under its supervision by utilizing the technique previously outlined, which focuses on analyzing interbank transaction data in the BI-RTGS system. This evaluation is of utmost importance due to the potential occurrence of exchange rate risk and liquidity risk resulting from the actions of market participants.

5. Conclusion and Future Work

5.1 Conclusion

In this study, we have proposed a novel approach to identifiying core-periphery structure. We examine machine learning to categorize banks into core and periphery banks, using discrete and continuous models and the SVM method based on the similarity between out-coreness and in-coreness values.

The BI-RTGS interbank payment system has identified a core-periphery structure consisting of 10 banks consistently categorized as core banks for 24 months. These banks are interconnected and serve as intermediaries for lower-tier banks, making them top-tier in the core-periphery structure. The Department of Financial System Surveillance (DSSK) at Bank Indonesia uses these identification results for supervisory purposes, assessing intraday liquidity risk and examining contagion spread. To enhance financial sector surveillance, DSSK used the result for identifying primary players in foreign exchange transactions, analyzing interbank transaction data in the BI-RTGS system.

We have implemented various protective measures to ensure the confidentiality of the banks' data. These include data minimization, bank anonymization, implementing stringent access controls to our network analytics dashboard, and utilizing the results by always consider the ethical implications of classifying banks and ensuring these are used responsibly in the context of financial stability oversight.

5.2 Future Work

There are several improvements in the methodology that can be applied for future works.

1. Strengthen the monitoring and analysis (e.g. liquidity risk and contagion) of core and periphery banks at a more frequently level, such as weekly or daily intervals, to detect any occurrences that could influence the network.
2. Enhance the data source to encompass the whole interbank payment system in Indonesia, including but not limited to retail payment system like fast payment and card-based transaction.
3. Incorporate additional economic indicators into the BI-RTGS transactions to enrich understanding of the overall economic ecosystem and its influence on the core-periphery structure e.g. macroeconomic and sectoral indicators.

References

Baek, S., Soramaki, K., & Yoon, J. 2014. "Network indicators for monitoring intraday liquidity in BOK-Wire+". Journal of Financial Market Infrastructures 2(3), 37-66

Boyd, J. P., Fitzgerald, W. J., Mahutga, M. C., &Smith, D. A. 2010. "Computing continuous core/periphery structures for social relations data with MINRES/SVD". Social Networks, 32(2), 125–137 3.

Craig, Ben R. & von Peter, Goetz, 2010. "Interbank tiering and money center banks". BIS Working Papers 322, Bank for International Settlements.

Fricke, D., Lux, T., 2014, "Core-periphery structure in the overnight money market: Evidence from the e-MID trading platform". Computational Economics 45:359–395

Goldberg, David E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning, Kluwer Academic Publishers, Boston, MA.

Untoro, Widodo P.R., dan Yuwana W. 2014. "Kajian Penggunaan Instrumen Sistem Pembayaran sebagai Leading Indicator Stabilitas Sistem Keuangan. Working Paper, Bank Indonesia.

Van Lelyveld, I., in 't Veld, D., 2012, "Finding the core: Network structure in interbank markets". DNB Working Paper 348, DNB.

Abduraimova, K. (2022). Contagion and tail risk in complex financial networks. Durham University Business School, United Kingdom. Available online 29 May 2022. Article history: Received 31 October 2020; Accepted 28 May 2022.

Ballester, L., Casu, B., & González-Urteaga, A. (2016). Bank fragility and contagion: Evidence from the bank CDS market.

Zhang, X., Fu, Q., Lu, L., Wang, Q., & Zhang, S. (2021). Bank liquidity creation, network contagion and systemic risk: Evidence from Chinese listed banks.

Heijmans, R., & Wendt, F. (2023.). Measuring the impact of a failing participant in payment systems. De Nederlandsche Bank, Netherlands; International Monetary Fund, United States of America.

Fricke, D., & Lux, T. (2012). Core-periphery structure in the overnight money market: Evidence from the e-MID trading platform. Kiel Working Papers, No. 1759.

In 't Veld, D., Van der Leij, M., & Hommes, C. (2020). The formation of a core-periphery structure in heterogeneous financial networks.

DIDD
DEPARTEMEN INOVASI & DIGITALISASI DATA

Data at your fingertips.

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

Unveiling the Interconnectedness of Banks in Payment Systems: Methodology, Utilization, and Data Governance Considerations

17 October 2023

Renardi Ardiya Bimantoro, Mohammad Khoyrul Hidayat, Irfan Sampe
Department of Data Innovation and Digitalization, Department of Statistics – Bank Indonesia
Email: renardi.ardiya@bi.go.id; moh.khoyrul@bi.go.id; irfans@bi.go.id

Opinions expressed are solely our own and do not express the views or opinions of Bank Indonesia

Background

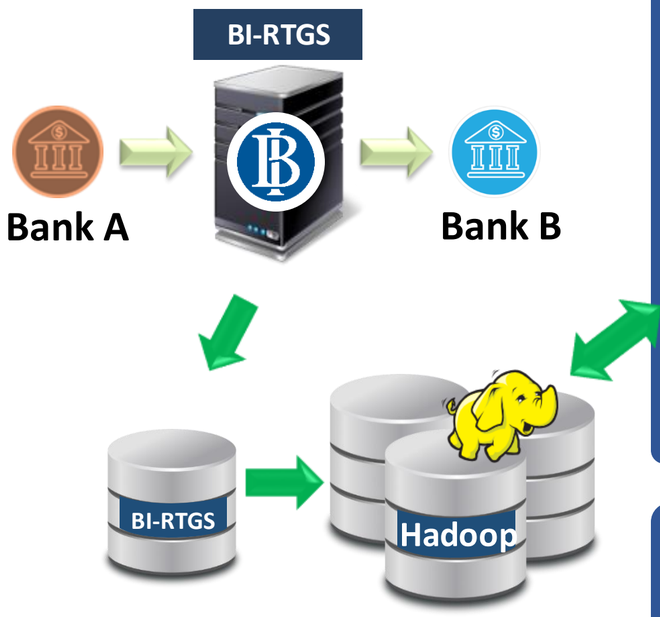
BACKGROUND

- In the payment system, every bank is interconnected with each other. When problems occurs in one bank, they can potentially have ripple effects on other linked institutions. Therefore, it's crucial to identify the interconnected structure of participants in the payment system.
- Leveraging Big Data Analytics through the core-periphery model to visualize the systemic structure (interbank tiering) comprehensively serves to complement the interconnectedness analysis that has been conducted so far.

OBJECTIVE

Discover the interlinked structure of participants in the payment system by applying network analysis method.

DATA SOURCES & METHODOLOGY



Data Preparation

1. Filtering & Mapping

- Filter out non-bank BI-RTGS participants names.
- Map branch unit bank to their parents bank
- Codify the Bank Name using Bank Code

2. Data Aggregation & Matrix Representation

- Aggregate transaction data on a monthly basis.
- Represent transaction data in discrete matrices (connectivity) & continuous matrices (transaction amount/frequency).

Data Extraction

Model *Core - Periphery*

3. Model Asymmetric Discrete

Identify Core & Periphery banks based on connectivity (presence or absence of transactions).

4. Model Asymmetric Continuous

Calculate out-coreness and in-coreness based on transaction amount & frequency.

5. SVM Classification

Classification using Support Vector Machine (SVM) based on the results of discrete and continuous models.

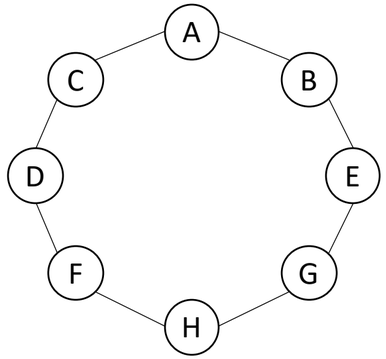
- ✓ Total transaction rows: ± 1 million rows per month
- ✓ Number of banks transacting: 117 banks
- ✓ Transaction amount ≥ Rp.100 million or ≥ USD 6,390

Interbank & Payment System Network

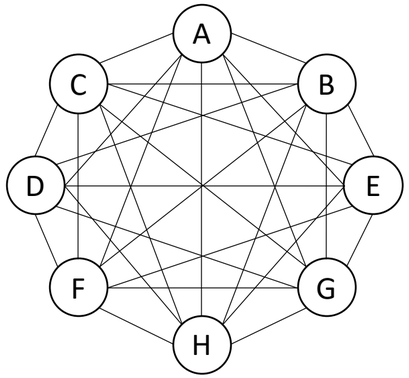
Numerous studies on interbank & payment system networks indicate that the observed structure exhibits a hierarchical pattern, often referred to as interbank tiering.

Interbank tiering describes the interbank structure that emerges when there are interconnected banks also function as intermediaries between non-interconnected banks (Craig & Peter, 2010).

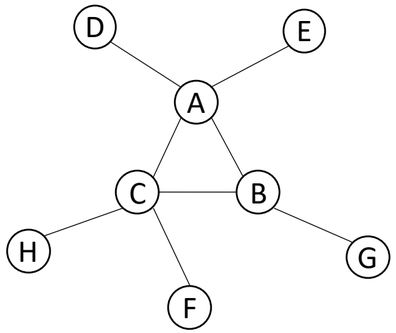
The model often used to identify such network structures is the core-periphery model (Craig & Peter, 2010).



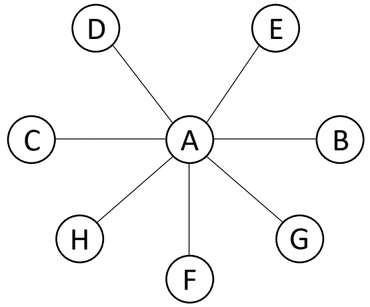
Ring



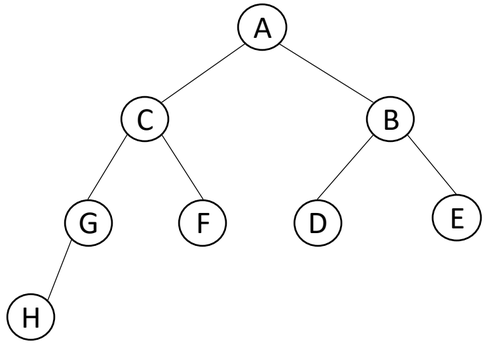
Full Mesh



Core – Periphery



Star



Tree

Core-Periphery Model Type

	A	B	C	D	E
A	0	1	1	0	0
B	1	0	0	1	1
C	1	0	0	1	0
D	0	1	1	0	1
E	0	1	0	1	0

Matriks Biner Simetris

	A	B	C	D	E
A	0	1	1	0	0
B	0	0	0	1	1
C	1	1	0	1	0
D	1	0	0	0	1
E	1	0	1	0	0

Matriks Biner Asimetris

Model *Symmetric Discrete* Borgatti dan Everett, 1999	**Model *Symmetric Continuous*** Borgatti dan Everett, 1999
Objective: To classify nodes in an undirected network into core and periphery nodes.	**Objectives:** To determine the coreness value of each node in an undirected network.
Data: Connectivity between nodes in a network (0: not connected, 1: connected).	**Data :** Edge weights in the network with continuous values (e.g., based on transaction amount or frequency).
Model *Asymmetric Discrete* Craig dan Peter, 2010	**Model *Asymmetric Continuous*** Boyd et al., 2010
Objective: To classify nodes in a directed network into core and periphery nodes.	**Objectives:** To determine the coreness value of each node in a directed network
Data : Connectivity between banks in a network (0: not connected, 1: connected).	**Data :** Edge weights in the network with continuous values (e.g., based on transaction amount or frequency).

	A	B	C	D	E
A	0	10	5	0	0
B	10	0	0	15	25
C	5	0	0	3	0
D	0	15	3	0	30
E	0	25	0	30	0

Matriks Kontinu Simetris

	A	B	C	D	E
A	0	10	18	0	0
B	0	0	0	20	1
C	15	10	0	2	18
D	30	0	0	0	10
E	25	0	3	0	0

Matriks Kontinu Asimetris

Model used to identify BI-RTGS *network*

Model Asymmetric Discrete

Using blockmodelling techniques to identify core and periphery banks based on interbank connectivity patterns. (Craig & Peter, 2010)

Blockmodelling is a network analysis method that partitions a network into several groups (also called blocks), where each group consists of nodes that have similar connectivity patterns with other members.

The ideal blockmodel is achieved when the three properties of the model are met, namely:

- ✓ ***Top-tier banks*** (***core***) are interconnected with each other. (CC)
- ✓ ***Lower-tier banks*** (***periphery***) are not interconnected with each other. (PP)
- ✓ ***Top-tier banks*** (***core***) are connected to (some) ***lower-tier banks*** (***periphery***). $(CP\ \&\ PC)$

$$\begin{bmatrix} CC & CP \\ PC & PP \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \boldsymbol{RR} \\ \boldsymbol{CR} & \mathbf{0} \end{bmatrix}$$

Blok *CC* (*Core to Core*) : Connectivity between core banks
Blok *PP* (*Periphery to Periphery*) : Connectivity between *periphery* banks
Blok *CP* (*Core to Periphery*) : Connectivity from *core bank* → *periphery bank*
Blok *PC* (*Periphery to Core*) : Connectivity from *periphery bank* → *core bank*

$\boldsymbol{RR}$: $At\ least\ 1\ connection\ in\ every\ row$ $\boldsymbol{CR}$: $At\ least\ 1\ connection\ in\ every\ column.$

		Core			Periphery				
		A	***B***	***C***	***D***	***E***	***F***	***G***	***H***
Core	***A***	0	1	1	0	1	0	0	0
	B	1	0	1	0	0	0	1	0
	C	1	1	0	0	0	0	0	1
Periphery	***D***	1	0	0	0	0	0	0	0
	E	0	1	0	0	0	0	0	0
	F	0	0	1	0	0	0	0	0
	G	0	1	0	0	0	0	0	0
	H	0	0	0	0	0	0	0	0

Example of block model with 3 Cores (A, B, C)

Model Asymmetric Discrete

Let's suppose there are n banks labeled 1,2,..., n.

- The interbank connectivity data (transactions) is stored in matrix N where:

$$N_{ij} = \begin{cases} 1, & if\ there\ is\ a\ transaction\ from\ i\ to\ j \\ 0, & if\ there\ is\ a\ transaction\ rom\ j\ to\ i \end{cases}$$

for $i = 1,2,\dots n \quad and \quad j = 1,2,\dots n$

- Define C^* as a set of banks designated as core, and N_{C^*} is the blockmodel of matrix N

Objective Function

$$\arg\min_{C^*} \left[e(N_{C^*}) = \frac{e_{cc} + e_{pp} + (e_{cp} + e_{pc})}{\sum_i \sum_j N_{ij}} \right]$$

Where

$$E = \begin{bmatrix} e_{cc} & e_{cp} \\ e_{pc} & e_{pp} \end{bmatrix} = \begin{bmatrix} c(c-1) - \sum_{i\in C^*} \sum_{j\in C^*} N_{ij} & (n-c) \sum_{i\in C^*} \max\left\{0, 1 - \sum_{j\notin C^*} N_{ij}\right\} \\ (n-c) \sum_{j\in C^*} \max\left\{0, 1 - \sum_{i\notin C^*} N_{ij}\right\} & \sum_{i\notin C^*} \sum_{j\notin C^*} N_{ij} \end{bmatrix}$$

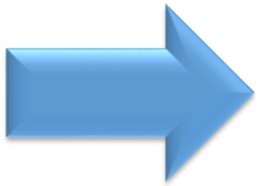
$e(N_{C^*})$ = Total $error$ (*blockmodel* N_{C^*})
e_{CC} = $Error$ at block CC
e_{PP} = $Error$ at block PP
e_{CP} = $Error$ at block CP
e_{PC} = $Error$ at block PC
N_{ij} = Transaction Matrix Element *at* row i and column j

The optimal partition is a set of core banks (C^*) that produces the smallest error value ($e(N_{C^*})$) (closest to the ideal blockmodel).

Model Asymmetric Discrete

	A	B	C	D	E	F	G	H
A	0	1	1	0	1	0	0	0
B	1	0	0	0	0	0	1	0
C	1	1	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	0	1	0	0	0	0	0	0
F	0	0	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0

?



	A	B	C	D	E	F	G	H
A	0	1	1	0	1	0	0	0
B	1	0	**0**	0	0	0	1	0
C	1	1	0	0	0	0	0	1
D	1	0	0	0	0	0	0	**1**
E	0	1	0	0	0	0	0	0
F	0	0	1	0	0	0	0	0
G	0	1	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0

- If the number of nodes in a network is n, then there are $2^n - n - 1$ choices in determining which nodes become the core in order to produce an optimal blockmodel (closest to the ideal blockmodel).
- How to obtain the optimal blockmodel?

- ✓ **Heuristic algorithms can be used to obtain the optimal blockmodel.**
- ✓ **According to Borgatti (1999), the Genetic Algorithm is better for finding the optimal blockmodel as it provides more stable results compared to other algorithms.**

Model Asymmetric Continues

The matrix decomposition method is utilized to measure the 'core' strength of each bank. (Fricke & Lux, 2012)

Pengirim	Penerima	Nominal Transaksi
A	B	143
A	C	95
A	E	61
B	A	61
B	C	70
B	D	72
C	A	66
C	B	86
C	E	50
D	A	20
E	B	29

$$N = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 143 & 95 & 0 & 61 \\ 61 & 0 & 70 & 72 & 0 \\ 66 & 86 & 0 & 0 & 50 \\ 20 & 0 & 0 & 0 & 0 \\ 0 & 29 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

N : Matriks *Adjacency*
nominal transaction data

$$u = \begin{bmatrix} 0.355 \\ 0.340 \\ 0.232 \\ 0.016 \\ 0.042 \end{bmatrix} \begin{array}{l} A \\ B \\ C \\ D \\ E \end{array} \quad , v = \begin{bmatrix} 0.212 \\ 0.396 \\ 0.236 \\ 0.130 \\ 0.113 \end{bmatrix} \begin{array}{l} A \\ B \\ C \\ D \\ E \end{array}$$

- u_i : ***coreness*** bank i as sender
- v_j : ***coreness*** bank j as receiver

The fundamental concept of the **Asymmetric Continuous Model** is to decompose the adjacency matrix into '**out-coreness**' and '**in-coreness**' as follows:

$$N = \begin{cases} uv^T, & off-diagonal \\ 0, & diagonal \end{cases}$$

where u: **out-coreness vector** & v: **in-coreness vector**.

Model Asymmetric Continues

Suppose there are n banks labeled as $1,2,\ldots,n$

- The nominal or frequency of interbank transactions is stored in the matrix N
- where: $N_{ij} = \begin{cases} transaction\ nominal\ or\ frequency, & if\ there\ is\ transaction\ from\ i\ to\ j \\ 0, & if\ there\ is\ no\ transaction\ from\ i\ to\ j \end{cases}$

  for $i = 1,2,\ldots n \quad dan \quad j = 1,2,\ldots n$
- Let u be defined as the out-coreness vector and v as the in-coreness vector.

Objective Function

$$\arg\min_{\boldsymbol{u},\boldsymbol{v}} \sum_{i} \sum_{j \neq i} \left(N_{ij} - \boldsymbol{u}_i \boldsymbol{v}_j^T\right)^2$$

Contoh*

	B	C	H	A	F	D	E	G
B	0	144	59	281	78	114	93	67
C	110	0	34	161	45	65	53	39
H	20	15	0	28	8	12	9	7
A	396	296	120	0	161	234	191	138
F	34	26	10	50	0	20	16	12
D	70	52	21	102	28	0	34	24
E	48	36	14	69	19	28	0	17
G	25	19	8	37	10	15	12	0

Normalisasi

	B	C	H	A	F	D	E	G
B	0	0.144	0.059	0.281	0.078	0.114	0.093	0.067
C	0.11	0	0.034	0.161	0.045	0.065	0.053	0.039
H	0.02	0.015	0	0.028	0.008	0.012	0.009	0.007
A	0.396	0.296	0.12	0	0.161	0.234	0.191	0.138
F	0.034	0.026	0.01	0.05	0	0.02	0.016	0.012
D	0.07	0.052	0.021	0.102	0.028	0	0.034	0.024
E	0.048	0.036	0.014	0.069	0.019	0.028	0	0.017
G	0.025	0.019	0.008	0.037	0.01	0.015	0.012	0

$$\boldsymbol{u} = \begin{bmatrix} 0.85 \\ 0.42 \\ 0.24 \\ 0.24 \\ 0.15 \\ 0.10 \\ 0.07 \\ 0.05 \end{bmatrix} \begin{matrix} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \\ \mathbf{D} \\ \mathbf{E} \\ \mathbf{F} \\ \mathbf{G} \\ \mathbf{H} \end{matrix}, \boldsymbol{v} = \begin{bmatrix} 0.68 \\ 0.46 \\ 0.35 \\ 0.27 \\ 0.22 \\ 0.19 \\ 0.16 \\ 0.14 \end{bmatrix} \begin{matrix} \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \\ \mathbf{D} \\ \mathbf{E} \\ \mathbf{F} \\ \mathbf{G} \\ \mathbf{H} \end{matrix}$$

$$u_A v_B = (0.85)(0.46) = 0.391 \qquad u_B v_A = (0.42)(0.68) = 0.285$$

Model Asymmetric Continues

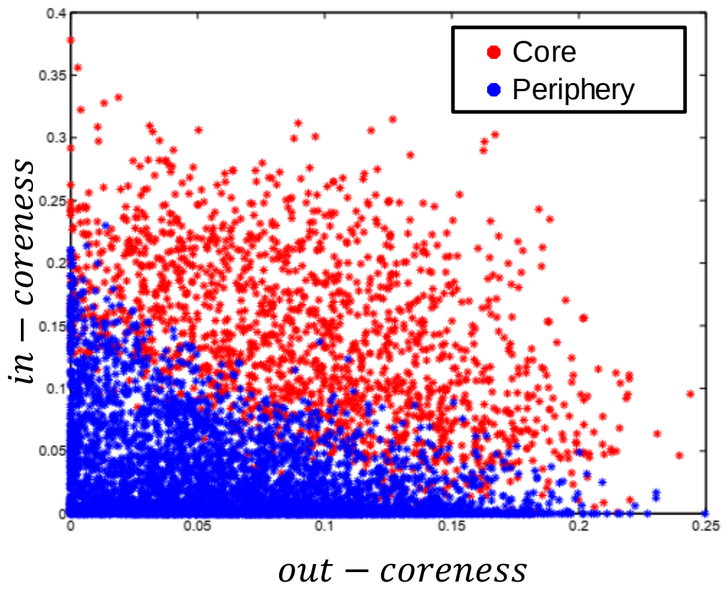
Model *Asymmetric Discrete*

- **Objective:** Classify banks into core and periphery.
- **Criteria:** Only considers the interbank connections in a network. **(0: not connected, 1: connected).**

Model *Asymmetric Continuous*

- **Objective:** Measure coreness, indicating the 'core' strength of each bank in a network.
- **Criteria:** Considers both transaction values and interbank connections when determining the coreness value of each bank.

Scatter plot of Coreness

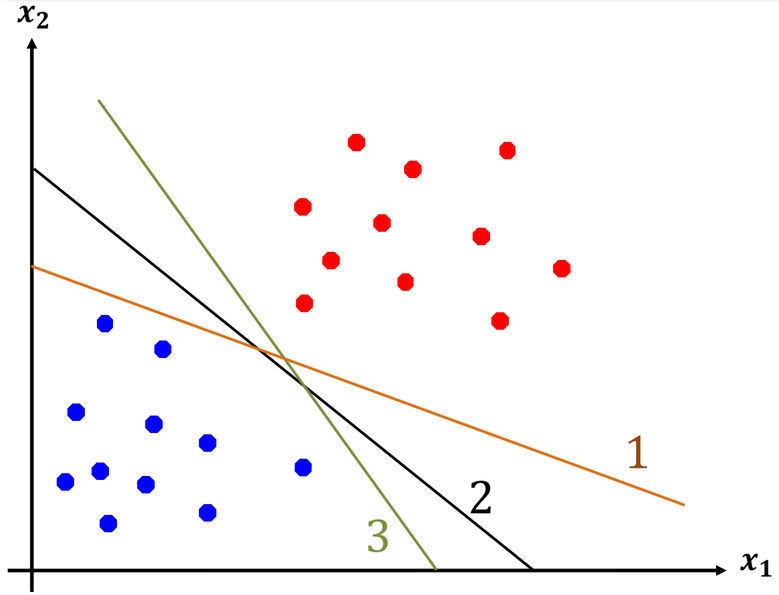


Core and periphery bank characteristic based on coreness value

- **Visualization:** A scatter plot of coreness values can be used to analyze the relationship between the results of the discrete model and the continuous model.
- **Plot Details:** The scatter plot is formed based on out-coreness values (x-axis), in-coreness values (y-axis), and the classification results of the discrete model (color of the dots).
- **Observation:** The out-coreness and in-coreness values of core banks should be significantly higher compared to periphery banks.

Model Asymmetric Continues

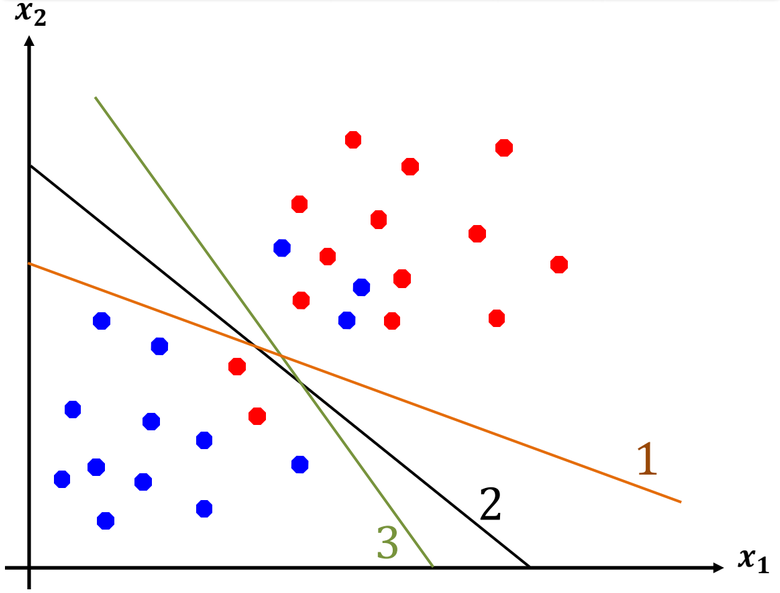
Ideal Cases



- Core
- Periphery

x_1 : *out – coreness*
x_2 : *in – coreness*

Non-Ideal Cases



- ❑ A **line equation** can be used as a **boundary that separates the core and periphery regions**.
- ❑ There **are several possible line equations**; which equation will be chosen or is the most optimal? **1, 2, or 3?**

Using machine learning algorithm to determine the core and periphery threshold area
Example: *Support Vector Machine* (SVM)

Core-Periphery Classification Using SVM

1 **Model**: Asymmetric Discrete
Algoritma: Genetic Algorithm

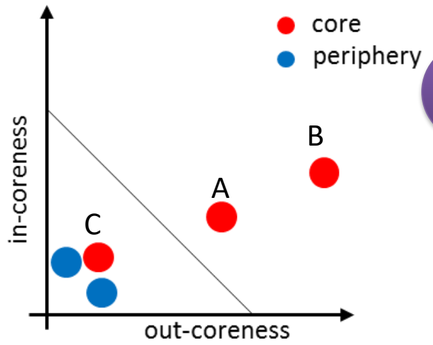
Identigy Core & Periphery Bank based on the interlink relationship (have transaction activity or not)

	A (Core)	B (Core)	C (Periphery)	D (Periphery)	E (Periphery)
Bank A	0	1	1	0	0
Bank B	1	0	1	0	1
Bank C	1	1	0	0	0
Bank D	1	0	0	0	0
Bank E	0	0	0	0	0

2 **Model**: Asymmetric Continuous

Calculate *out-coreness* dan *in-coreness* based on **transaction nominal**

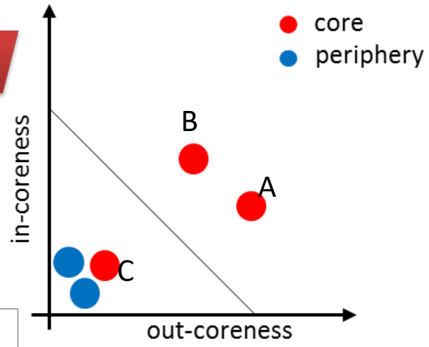
Bank	in-core	out-core
A	0,54	0,42
B	0,87	0,53
C	0,08	0,11
D	0,06	0,14
E	0,12	0,09



3 **Model**: Asymmetric Continuous

Calculate *out-coreness* dan *in-coreness* based on **transaction frequency**

Bank	in-core	out-core
A	0,75	0,52
B	0,67	0,87
C	0,12	0,01
D	0,08	0,02
E	0,11	0,15



4 **Algoritma**: Support Vector Machine (SVM)

Identify Core & Periphery Bank based on interconnectedness and **transaction nominal**

5 **Algoritma**: Support Vector Machine (SVM)

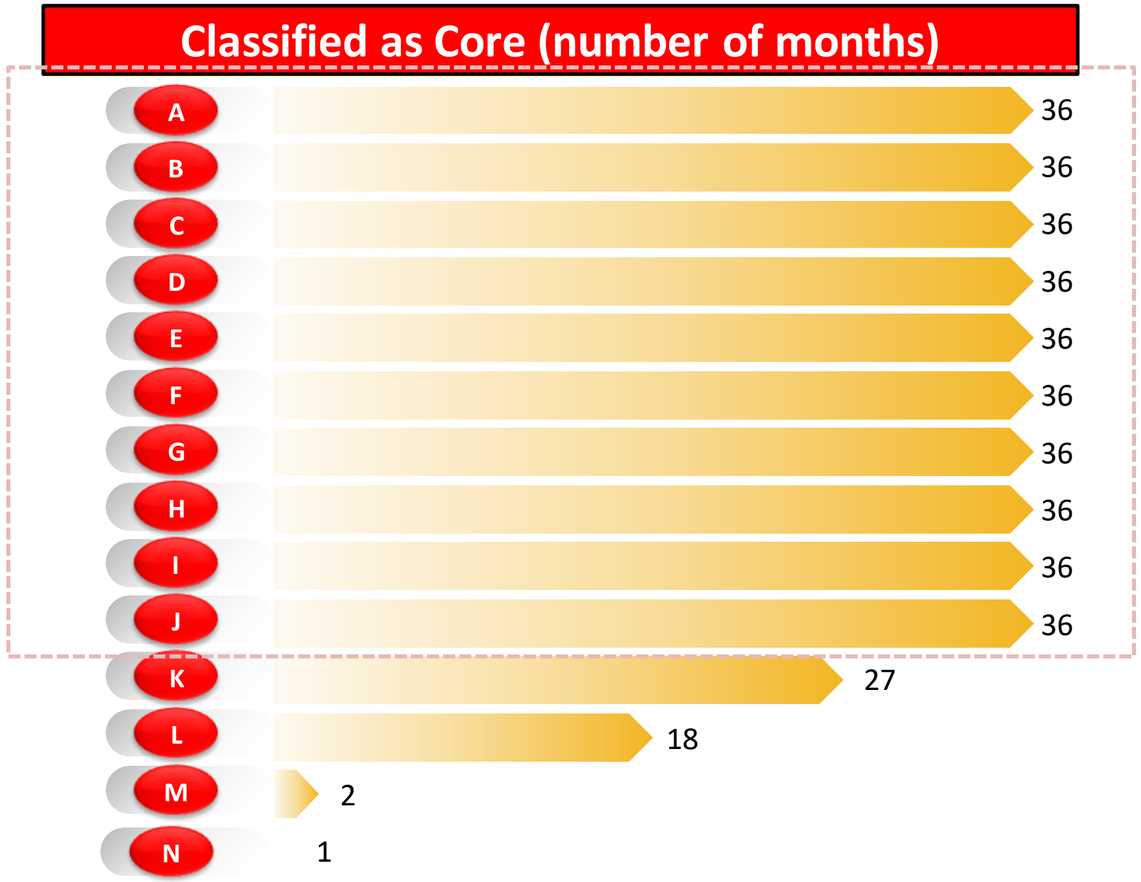
Identify Bank Core & Periphery based on interconnectedness and **transaction frequency**

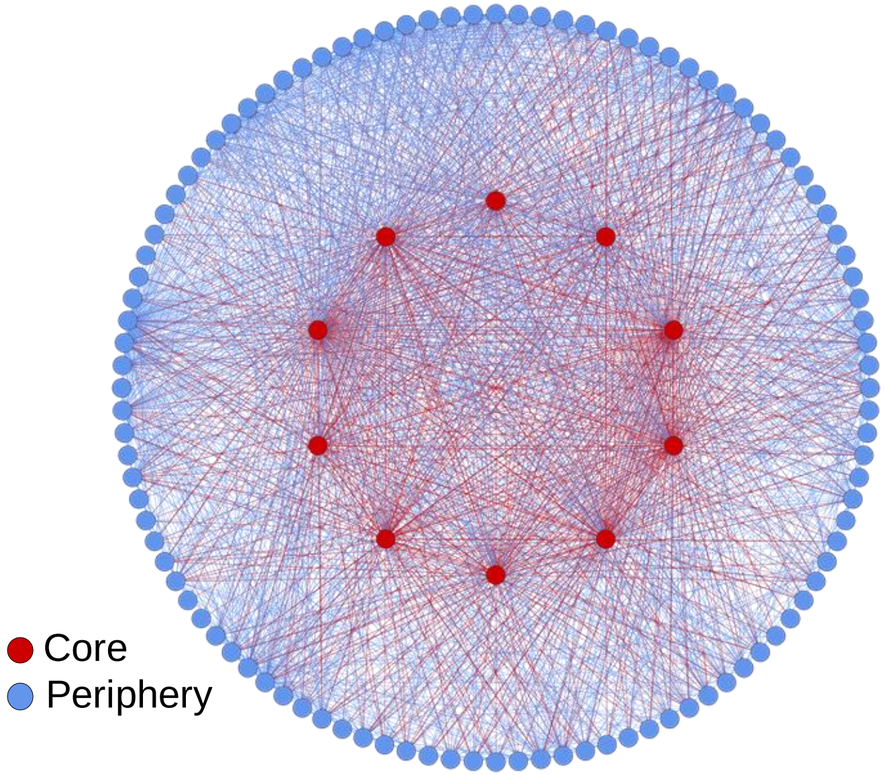
6 Identify Bank Core & Periphery based on the combination of **interconnectedness**, ***frequency*** and **nominal**

Bank A: *Core*
Bank B: *Core*
Bank C: *Periphery*
Bank D: *Periphery*
Bank E: *Periphery*

Core-Periphery Model Result

- The identification of the BI-RTGS network structure with the core-periphery model was conducted every month from January 2016 to December 2018 (36 months).
- Based on the results obtained, a core-periphery structure was identified within the BI-RTGS interbank transactions.
- This is evidenced by the presence of 10 banks that were identified as core banks (out of 117 banks that transacted with each other) throughout the observation period (36 months).

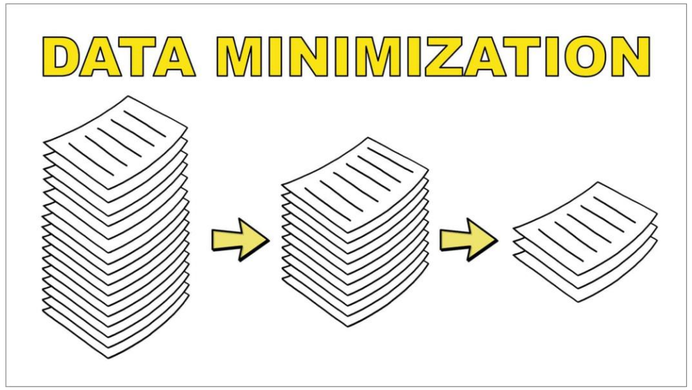




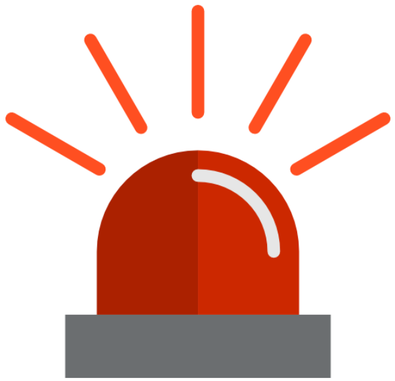
Struktur *Core-Periphery* in BI-RTGS Transaction

Data Governance Considerations

- **Data Minimization:** Only collect and process the minimum amount of data necessary to identify the core-periphery structure. In this case, we only use settlement time, sending/receiving bank, trx status, and amount. This reduces the risk of data breaches and ensures compliance with data protection regulations.
- **Bank Anonymization:** Before processing and analyzing the data, we ensure that all bank identifiers are anonymized. We replace bank names or any other identifiable information with generic IDs or pseudonyms. This ensures that individual banks cannot be directly identified from the data, protecting their privacy and ensuring compliance with data protection standards.
- **User Access Control:** Implement access controls to ensure that only authorized personnel can view and interpret the classification results. This not only ensures data privacy but also prevents potential misuse of the data.
- **Model Usage:** The core-periphery classification model should be used cautiously. For instance, it can be employed as an early warning system to highlight potential risks in the banking network. However, it should not be the sole source for making critical decisions. Always consider the ethical implications of classifying banks, especially if these classifications have significant financial or reputational implications for the banks involved.







DIDD
DEPARTEMEN INOVASI & DIGITALISASI DATA

Data at your fingertips.

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

Thank You

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Big data platform (FinPulse) initiative[1]

Tsenddorj Dorjpurev,

California Polytechnical State University

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Big Data Platform (FinPulse) Initiative

"Reimagining regulatory data collection and analysis"



Speaker:

Tsenddorj Dorjpurev,

California Polytechlical State Univsersity
Former Project Lead at The Bank of Mongolia

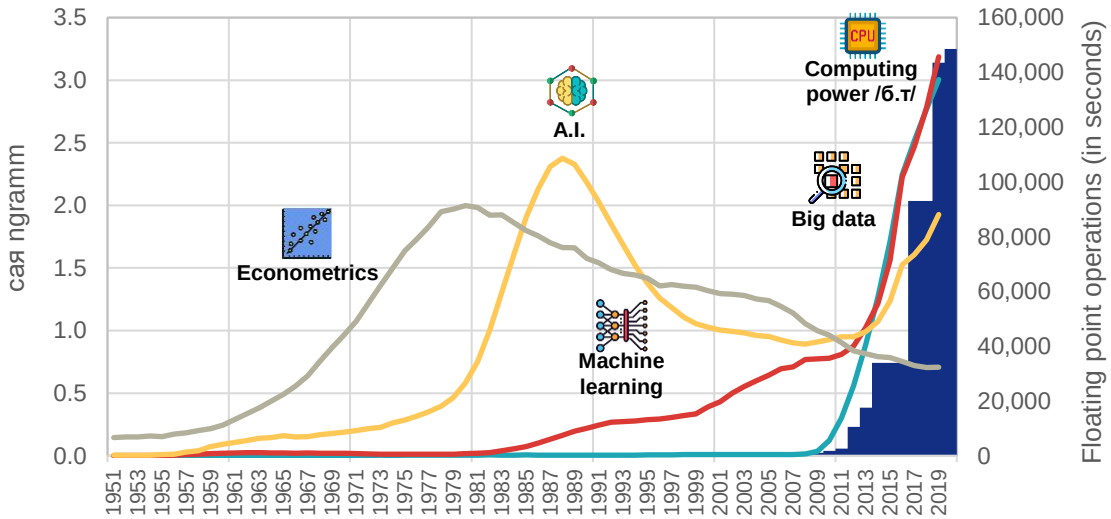
17 OCT 2023

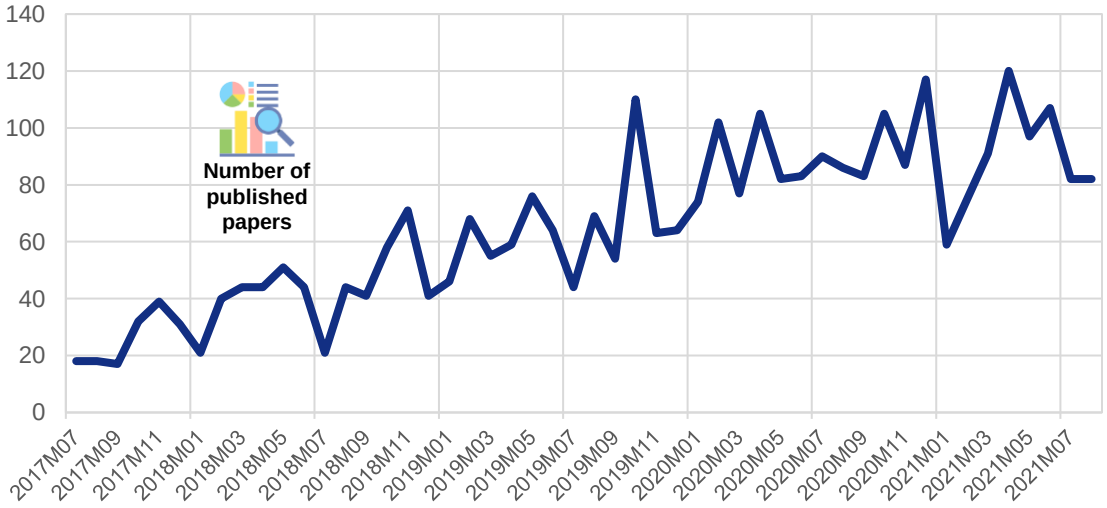
General interest

A strong interest in big data can be seen after 2010, where the term machine learning took up in the second half of the ’80s and artificial intelligence has measurable volume as early as the ’60s, and it has shown a decline through the ’90s 2000s and the first half of 2010s as well

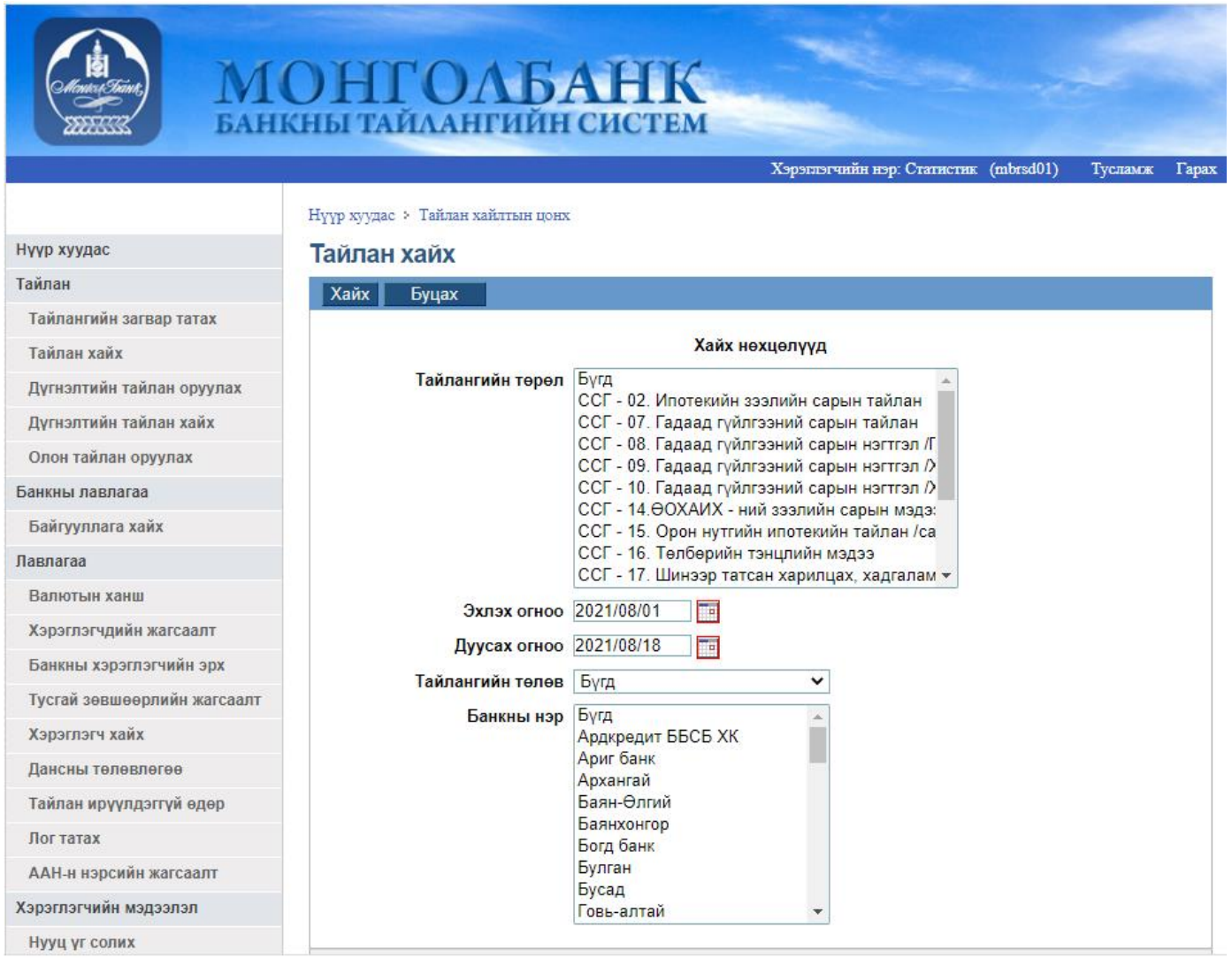
Academic interest

The field of Big Data has **3,315 papers (310 authors)** compared with 22,500 articles published in the field of central banking, from a total of 1,001,932 articles in the archive. In the following figure, we can see that **around 100 articles** are being published in this field every month.





Legacy Systems



+250

+70 templates are received by 5 departments:

RSD	PSD	RMFMD	BRPD	SD

- Duplicated information => Cross check validation
- Increasing frequency of changes in the templates and ad hoc data collection
- Increased fatigue at bank employees due to increasing regulatory data demand
- Demand for higher frequency, cohesive data collection

Project Impact



Creation of a standard, unified database

In Numbers

Condensed **70+ reports** and **250+ sheets** of data templates into a **single standardized data acquisition model**.



Reduced time-lag of official statistics

Reduced the time between data collection and data dissemination from **t+17 days** to **t+1 day.**



Data collection frequency

Increased average data collection frequency from **monthly** to **daily.**



Data volume

Increased recorded annual data volume **x425**, from **40GB** to **17TB** (26GB when data mirroring and more with backup).



Increased labor efficiency (cost reduction and human error)

Enabled **200+ workers** in the banking sector to focus on higher-value tasks such as problem solving and analytics.



Improvements of the Entire Central Bank and Partners

- Increased analytical capabilities that supervisors can use to provide policymakers and upper management with instant access to reports and key indicators.
- The standardized data model and staging database setup allow banks to change systems without disrupting the data-sharing connection.

Fin Pulse

Regulatory Reports

Current Data Collection System

Reserve Management Financial Markets Department

- Mortgage Soft Loan Program (by daily news)
- Mortgage bonds held by the Bank of Mongolia
- Mortgage Soft Loan Program (by resource allocation)
- Bank Reserve Requirement Statement Information
- Mortgage Soft Loan Program (by demand news)
- Loan Outstanding Report (by banks)
- Daily Current and Deposit Balance Information
- Liquidity Ratio Report
- Daily International Transactions News
- Foreign Currency Risk Report (by banks)
- Monthly Interbank Transaction Report
- Foreign Currency Trading Report (by banks)

Research and Statistics Department

- Foreign Trade Statistics
- Oyun Tolgoi Monthly News
- Oyun Tolgoi Monthly Export Forecast
- State Budget News
- Cross-Border Passenger Survey
- Sex Dissaggregated Loan Report (by soum and district)
- Total Loan Report (by soum and district)
- Sex Disaggregated Individuals Current Account and Savings Report (by soum and district)
- Mongolian Embassy's Abroad Expenditure Report
- Government Foreign Loan Report
- Monthly Sales and Transportation Data of Coal Exporting Mongolian Enterprises
- Diplomatic Mission Expenditure Report
- Consumer Loan Report (by debt-to-income ratio)
- Green Loan Report
- Total Current and Savings Report (by soum and district)
- Sex Disaggregated Childrens Savings Report (by soum and district)
- Loan information for changes in loan contracts due to the COVID-19 pandemic
- Cash Replenishment News
- Bank Branch and Personnel Gender Report (by soum and district)
- Balance of Payments
- Loan Report of Banks (by economic activity)
- Rural-Area Loan Report
- SME Loan Report
- Bank Branch and Personnel Report
- Consolidated Balance Sheet (by banks)
- M2 Money Supply (by banks)
- Mortgage Loan Report
- Interest Rate Lending Report (Outstanding)
- Interest Rate Lending Report (New)
- Interest Rate Deposit Report (Outstanding)
- Monthly International Transactions News
- Balance of Payments (by banks)
- Interest Rate Deposit Report (New)
- Declared Loan Rates Information (by banks)
- Nominal and Real Effective Exchange Rate of MNT (NEER, REER)
- International Reserves and Foreign Currency Liquidity Report
- Bank of Mongolia Balance
- Bank of Mongolia Net Domestic Asset News
- Foreign and Domestic Ownership Concentration Survey
- Official Reserve Assets Report

Banking Resolution and Policy Department

- Rural-Area Banks Balance Sheet
- Liquidated Banks Balance Sheet
- Agricultural Support Loan Interest Rate /subsidies/
- Repo Weekly News
- Monthly Banks Balance Sheet
- Detailed Loan Report
- Weekly Banks Balance Sheet

Supervision Department

- Monthly Report on Job Support Loans
- List of All Related Parties of the Bank and Report on Assets Related to Them
- Mortgage Daily News /6% interest rate/
- Financial Derivatives Report
- Calculation of Interest Rate Support from the Government
- Calculation of Job Support Loan Interest Rate /subsidies/

Payment Systems Department

- Settlement Representative Report
- Remote Banking Report
- E-Money Report
- Payment Card Clearing Report
- Payment Card Report
- Continuity Report

28	Email	14	BRMS
16	FINA	13	Others

The New System in Development

Reserve Management Financial Markets Department

- Mortgage Soft Loan Program (by daily news)
- Mortgage bonds held by the Bank of Mongolia
- Mortgage Soft Loan Program (by resource allocation)
- Bank Reserve Requirement Statement Information
- Mortgage Soft Loan Program (by demand news)
- Loan Outstanding Report (by banks)
- Daily Current and Deposit Balance Information
- Liquidity Ratio Report
- Daily International Transactions News
- Foreign Currency Risk Report (by banks)
- Monthly Interbank Transaction Report
- Foreign Currency Trading Report (by banks)

Research and Statistics Department

- Foreign Trade Statistics
- Oyun Tolgoi Monthly News
- Oyun Tolgoi Monthly Export Forecast
- State Budget News
- Cross-Border Passenger Survey
- Sex Dissaggregated Loan Report (by soum and district)
- Total Loan Report (by soum and district)
- Sex Disaggregated Individuals Current Account and Savings Report (by soum and district)
- Mongolian Embassy's Abroad Expenditure Report
- Government Foreign Loan Report
- Monthly Sales and Transportation Data of Coal Exporting Mongolian Enterprises
- Diplomatic Mission Expenditure Report
- Consumer Loan Report (by debt-to-income ratio)
- Green Loan Report
- Total Current and Savings Report (by soum and district)
- Sex Disaggregated Childrens Savings Report (by soum and district)
- Loan information for changes in loan contracts due to the COVID-19 pandemic
- Cash Replenishment News
- Bank Branch and Personnel Gender Report (by soum and district)
- Balance of Payments
- Loan Report of Banks (by economic activity)
- Rural-Area Loan Report
- SME Loan Report
- Bank Branch and Personnel Report
- Consolidated Balance Sheet (by banks)
- M2 Money Supply (by banks)
- Mortgage Loan Report
- Interest Rate Lending Report (Outstanding)
- Interest Rate Lending Report (New)
- Interest Rate Deposit Report (Outstanding)
- Monthly International Transactions News
- Balance of Payments (by banks)
- Interest Rate Deposit Report (New)
- Declared Loan Rates Information (by banks)
- Nominal and Real Effective Exchange Rate of MNT (NEER, REER)
- International Reserves and Foreign Currency Liquidity Report
- Bank of Mongolia Balance
- Bank of Mongolia Net Domestic Asset News
- Foreign and Domestic Ownership Concentration Survey
- Official Reserve Assets Report

Banking Resolution and Policy Department

- Rural-Area Banks Balance Sheet
- Liquidated Banks Balance Sheet
- Agricultural Support Loan Interest Rate /subsidies/
- Repo Weekly News
- Monthly Banks Balance Sheet
- Detailed Loan Report
- Weekly Banks Balance Sheet

Supervision Department

- Monthly Report on Job Support Loans
- List of All Related Parties of the Bank and Report on Assets Related to Them
- Mortgage Daily News /6% interest rate/
- Financial Derivatives Report
- Calculation of Interest Rate Support from the Government
- Calculation of Job Support Loan Interest Rate /subsidies/

Payment Systems Department

- Settlement Representative Report
- Remote Banking Report
- E-Money Report
- Payment Card Clearing Report
- Payment Card Report
- Continuity Report

53	DAM
18	Dynamic Form

Data Acquisition Model of Big Data (FinPulse) System

Regulatory reports

- ❑ Loan Report of Banks
- ❑ SME Loan Report
- ❑ Rural-Area Loan Report
- ❑ Mortgage Loan Report
- ❑ Consolidated Balance Sheet
- ❑ Interest Rate Lending Report
- ❑ Balance of Payments
- ❑ Liquidity Ratio Report

. . .



70+ reports

20+ Standards

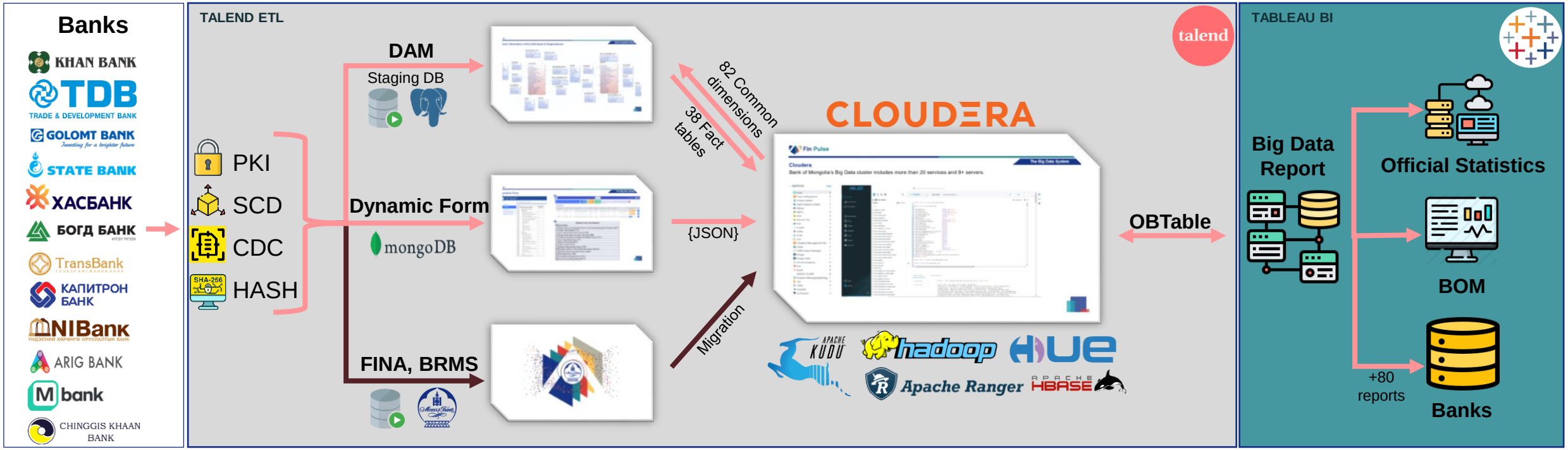
Data Security	Data Quality	Flexibility	New potential
Data Encryption	Data-Driven Culture	Staging DB	Data analytics
Role-Based Access	Metadata Management	DAM	New KPIs
Access Logs and Auditing	Unified Data Management	Adaptable to meet demands	Extra Actionable Insights
Multiple Layers of Security (closed circuit connection)	Automated Testing with Assertion	Supports Scalability	Data Interconnectivity

Official statistics standards & guidelines

- ❑ IFRS
- ❑ IMF SDDS
  - ❑ Monetary and Financial Statistics Guidelines
  - ❑ Balance of Payments Manual
- ❑ The System of National Accounts
- ❑ Green Taxonomy
- ❑ AFI Sex-disaggregated data guidelines
- ❑ The International standard industrial classification of all economic activities ISIC
- ❑ Merchant Category codes
- ❑ Zip postal code

. . .

Architecture

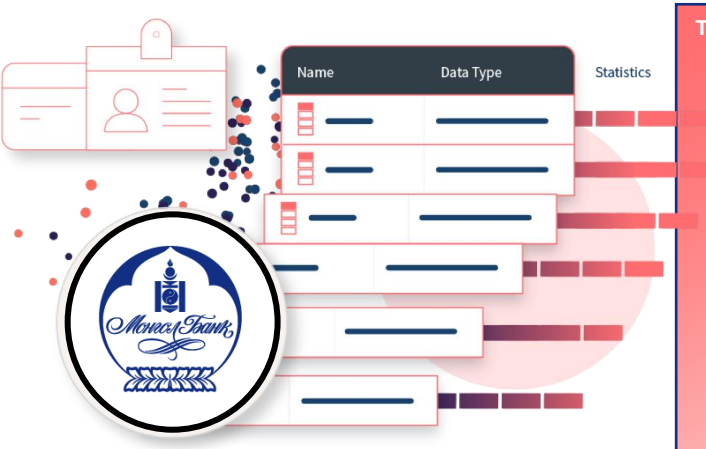


The Big Data platform uses staging databases (i.e., parallel warehouse) that are kept on the bank's premises to hold raw, granular-level data using the BoM's Data Acquisition Model (DAM) format which is encrypted and transmitted to the BoM through a closed-circuit connection, and offers a single, unified source of information for both supervisors and banks to access in near real-time. Besides DAM, the platform also collects data from commercial banks and licensed institutions through manually filled Dynamic form.

1. **Data Acquisition Model (DAM):** The DAM generates an automated report daily with access in near real-time, eliminates data redundancy and inconsistency both interdepartmentally and system-wide, and can acquire and assimilate more forms of data.
2. **Dynamic Form:** Data that is not readily available through DAM will be collected using a web-based dynamic questionnaire management solution connected to the big data system. Dynamic form uses customizable data templates at a fixed frequency and is easy to make additional changes and updates to the reports.

Fin Pulse

ETL “job”



TALEND ETL

12 Banks × **120** Common Dimensions and Fact Tables = **1440** Talend Jobs

38 Merged fact table → **38** Talend Jobs

10 OBT → **10** Talend Jobs

1488 Total Jobs on Talend



STAGING DB SQL

38 x 12 = 456 CDC SQL tables

12 Banks × **120** Common Dimensions and Fact Tables = **1440** SQL tables

1896 SQL tables

talend

CLOUDERA

12 Banks × **120** Common Dimensions and Fact Tables = **1440** SQL tables + **456** CDC tables (log) = **1896** SQL tables

Hue
Oracle
PostGre
SQL

38 Merged fact tables → **38** Merged fact tables’ SQL tables

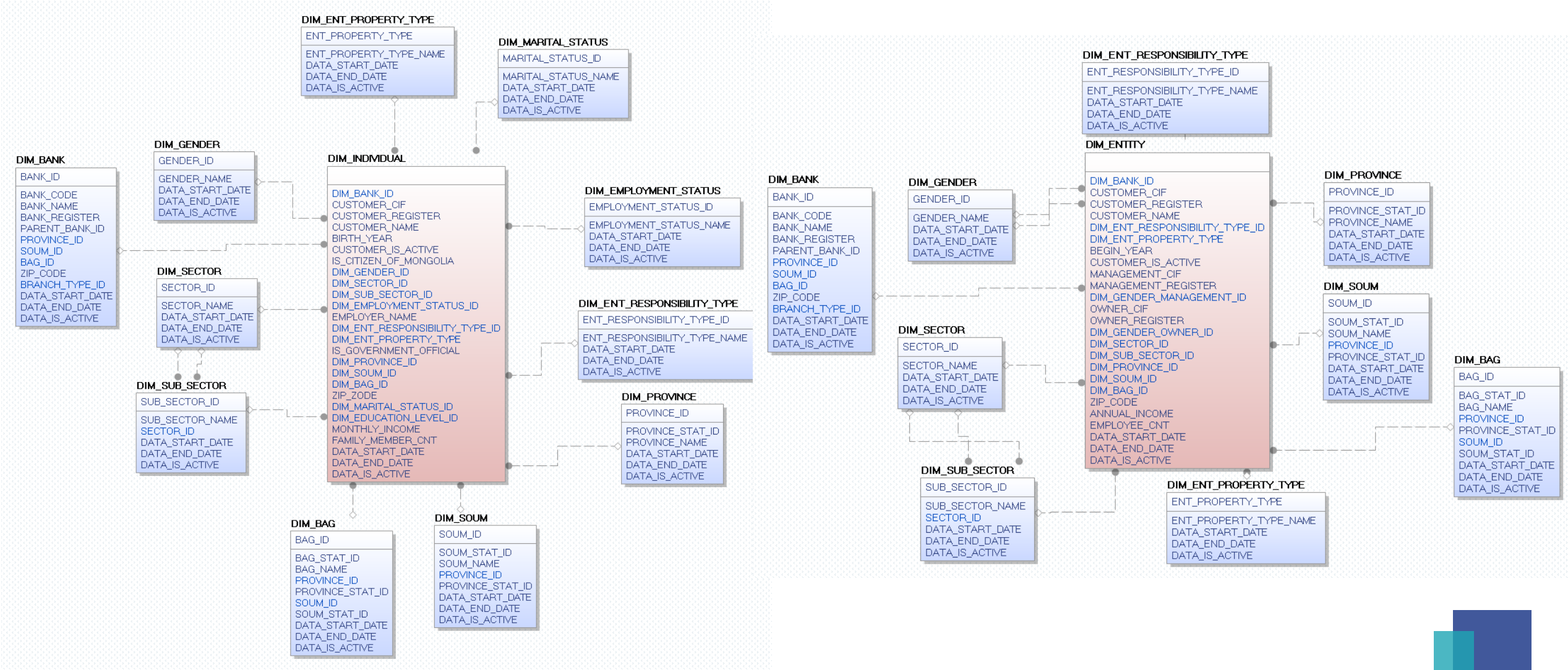
10 OBT → **10** OBT SQL tables

TOTAL SQL

1896 SDB SQL tables + **1896** BDS SQL tables + **38** Merged fact tables’ SQL tables + **10** OBT = **3840 Total SQL tables**

Basic Information of the Individuals & Organizations

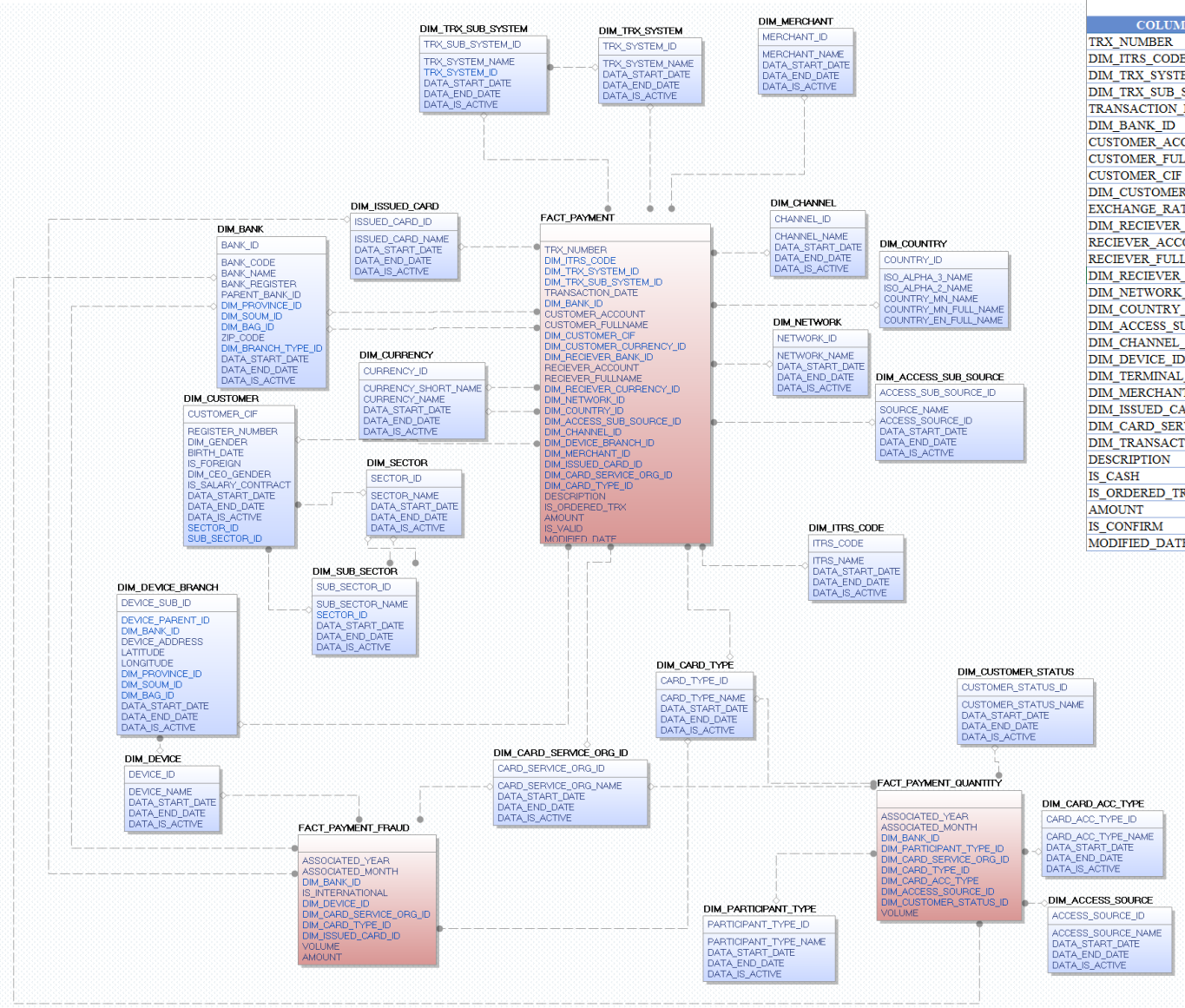
Data Acquisition Model



Fin Pulse

Payment Information

Data Acquisition Model



FACT_PAYMENT table

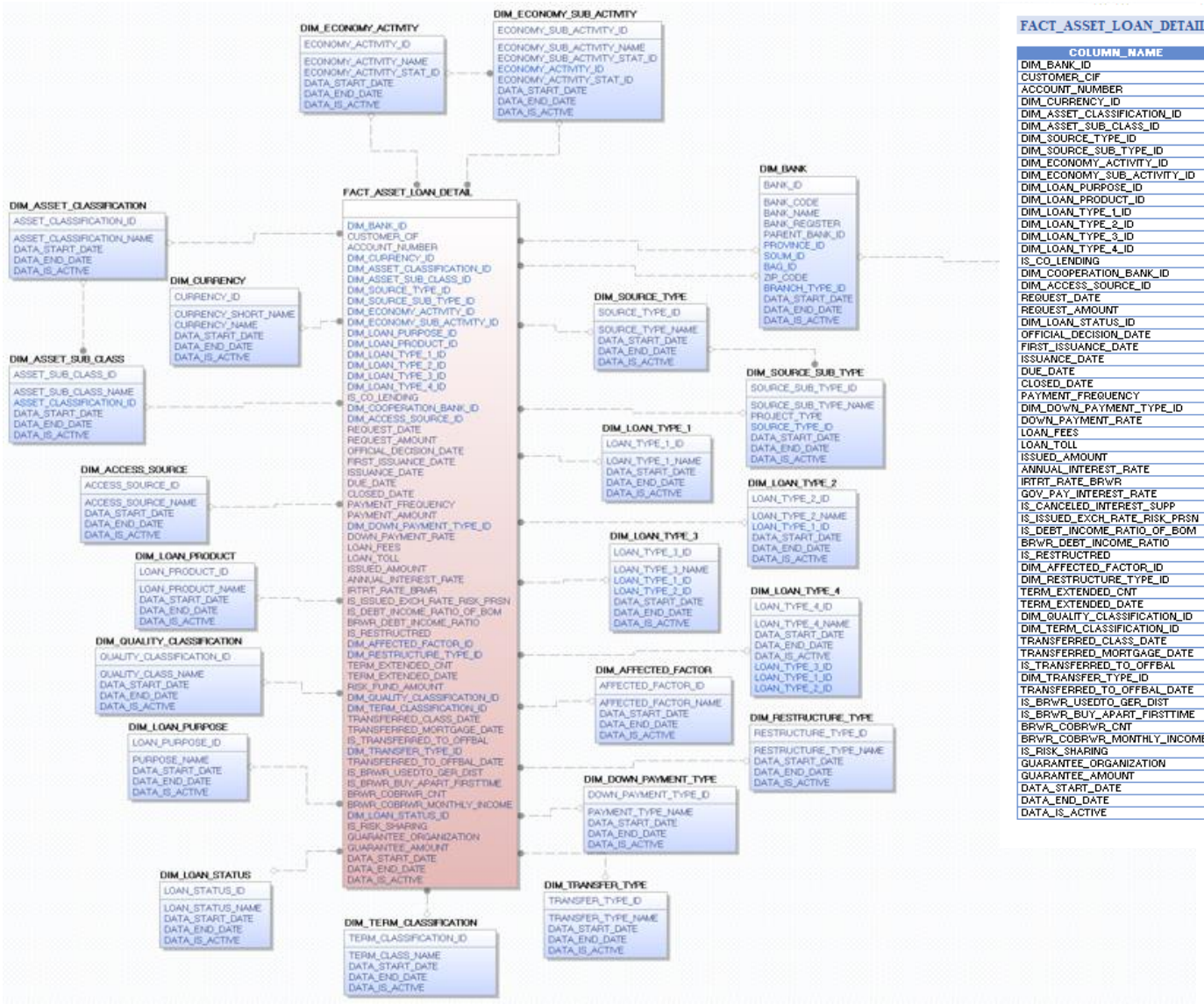
COLUMN_NAME	NAME	DESCRIPTION
TRX_NUMBER	Гүйлгээний дугаар	Банкуудын гүйлгээ бүрийг тодорхойлсон давтагдашгүй ID дугаар
DIM_ITRS_CODE	ITRS дугаар	Гадаад гүйлгээ хийсэн тохиолдолд DIM_ITRS table-н харгалзах ITRS_CODE оруулна.
DIM_TRX_SYSTEM_ID	Төлбөр тооцооны системийн төрөл	Банк доторхи гүйлгээ бол банкууд Inner-bank, Гадаад гүйлгээ Foreign transaction төрлөөр тус тус оруулна.
DIM_TRX_SUB_SYSTEM_ID	Төлбөр тооцооны системийн дэд төрөл	Банк доторхи гүйлгээ бол банкууд Inner-bank, Гадаад гүйлгээ Foreign transaction төрлөөр тус тус оруулна.
TRANSACTION_DATE	Гүйлгээ хийгдсэн өдөр, цаг, минут	YYYY-MM-DD HH:MM:SS форматаар байна
DIM_BANK_ID	Банк, салбар	Харилцагч салбар дээр хийгдсэн гүйлгээ бол салбарын кодоор авна.
CUSTOMER_ACCOUNT	Харилцагчийн дансны дугаар	Илгээгчийн данс ихдээ 12 оронтой тоо оруулна.
CUSTOMER_FULLNAME	Харилцагчийн нэр	Гадаад гүйлгээ хийсэн тохиолдолд илгээгчийн дансны нэрийг текстээр оруулна.
CUSTOMER_CIF	Харилцагчийн CIF	Банканд бүртгэлтэй харилцагчийн дахин давтагшгүй CIF дугаарыг оруулна.
DIM_CUSTOMER_CURRENCY_ID	Харилцагчийн дансны валют	DIM_CURRENCY table-н харгалзах CURRENCY_ID оруулна
EXCHANGE_RATE	Харилцагчийн валютын ханш	Валютын ханшын дүнг оруулна
DIM_RECIEVER_BANK_ID	Хүлээн авагчийн банк	DIM_BANK table-н харгалзах BANK_ID оруулна
RECIEVER_ACCOUNT	Хүлээн авагчийн дансны дугаар	Хүлээн авагчийн данс ихдээ 12 оронтой тоо оруулна.
RECIEVER_FULLNAME	Хүлээн авагчийн нэр	Гадаад гүйлгээ хийсэн тохиолдолд илгээгчийн дансны нэрийг текстээр оруулна.
DIM_RECIEVER_CURRENCY_ID	Хүлээн авагчийн дансны валют	DIM_CURRENCY table-н харгалзах CURRENCY_ID оруулна
DIM_NETWORK_ID	Нэгдсэн болон гадаад сүлжээний төрөл	DIM_NETWORK table-н харгалзах NETWORK_ID оруулна
DIM_COUNTRY_ID	Хүлээн авагчийн улсын дугаар	DIM_COUNTRY table-ийн COUNTRY_ID оруулна.
DIM_ACCESS_SUB_SOURCE_ID	Хандалтын эх үүсвэр	DIM_ACCESS_SUB_SOURCE table-н харгалзах ACCESS_SUB_SOURCE_ID оруулна
DIM_CHANNEL_ID	Гүйлгээний төрөл	DIM_CHANNEL table-н харгалзах CHANNEL_ID оруулна
DIM_DEVICE_ID	Төхөөрөмжийн төрөл	DIM_DEVICE table-ийн харгалзах DEVICE_ID оруулна
DIM_TERMINAL_ID	Төхөөрөмжийн терминал дугаар	DIM_DEVICE_BRANCH table-ийн харгалзах TERMINAL_ID оруулна
DIM_MERCHANT_CODE	Мерчантийн төрөл	DIM_MERCHANT table-н харгалзах MERCHANT_CODE оруулна
DIM_ISSUED_CARD_ID	Банкны картын эзэмшлийн төрөл	DIM_ISSUED_CARD table-н харгалзах ISSUED_CARD_ID оруулна
DIM_CARD_SERVICE_ORG_ID	Картын төрөл	DIM_CARD_SERVICE_ORG table-н харгалзах CARD_SERVICE_ORG_ID оруулна
DIM_TRANSACTION_TYPE_ID	Картын дэд төрөл	DIM_CARD_TYPE table-н харгалзах CARD_TYPE_ID оруулна
DESCRIPTION	Гүйлгээний утга	Текст утгаар оруулна
IS_CASH	Бэлэн мөнгөний гүйлгээ эсэх	1 = Тийм, 0 = Үгүй
IS_ORDERED_TRX	Захиалгат гүйлгээ эсэх	1 = Тийм, 0 = Үгүй
AMOUNT	Гүйлгээний дүн	Тоон утгаар оруулна
IS_CONFIRM	Хүчинтэй эсэх	Банкууд ITRS кодыг зассан тохиолдолд 1 гэж дугаарлана. Өмнөх буруу утгатай гүйлгээг 0 болгож өөрчилнө.
MODIFIED_DATE	Өөрчилсөн өдөр, цаг, минут, секунд	YYYY-MM-DD HH:MM:SS форматаар байна. Гүйлгээний мэдээллийг зассан өдөр цагийг оруулна.

FACT_PAYMENT_QUANTITY table

Хөдөлгөөнт банкны хэрэглэгч, картын тоо болон Гэрээт хамтрагчдийн тоон мэдээллийг FACT_PAYMENT_QUANTITY table-руу оруулна.

COLUMN_NAME	NAME	DESCRIPTION
ASSOCIATED_YEAR	Тайлант жил	Сар бүр нэгдсэн тоогоор татах тул тухайн жилийн тоог оруулна
ASSOCIATED_MONTH	Тайлант сар	Сар бүр нэгдсэн тоогоор татах тул тухайн сарын тоог оруулна
DIM_BANK_ID	Банк, салбар	Хэрэглэгчийн бүртгүүлсэн салбар нэгж
DIM_PARTICIPANT_TYPE_ID	Оролцогчдийн хандах төрөл	Карт, хөдөлгөөнт банк, гэрээт хамтрагч гэсэн ангилал бий
DIM_CARD_SERVICE_ORG_ID	Картын үйлчилгээ үзүүлэгчийн карт	DIM_CARD_SERVICE_ORG table-ийн CARD_SERVICE_ORG_ID оруулна
DIM_TRANSACTION_TYPE_ID	Картын төрөл	Зөвхөн картын хэрэглэгчийн тоон дээр нөхөгдөнө
DIM_CARD_ACC_TYPE_ID	Картын дэд төрөл	Зөвхөн картын хэрэглэгчийн тоон дээр нөхөгдөнө
DIM_ACCESS_SOURCE_ID	Хандалтын эх үүсвэр	Зөвхөн хөдөлгөөнт банкны ангилалийн ID авч ашиглана
DIM_CUSTOMER_STATUS_ID	Идэвхитэй, болон нийт, тайлант сард г.м ангилалыг оруулна	
VOLUME	Тоо хэмжээ	Тоон утгаар оруулна

Asset Loan Balance

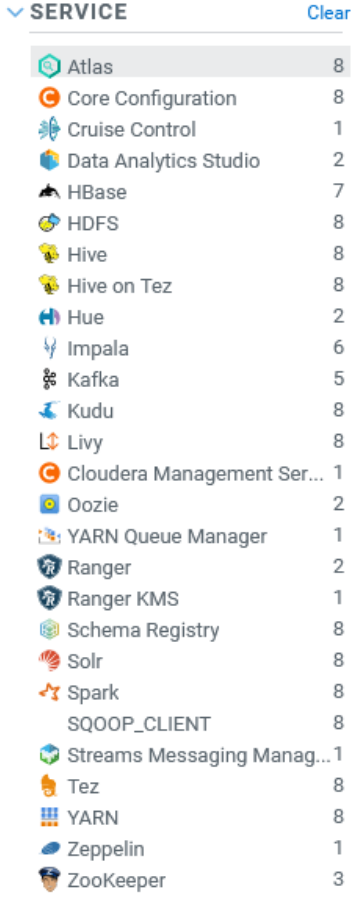


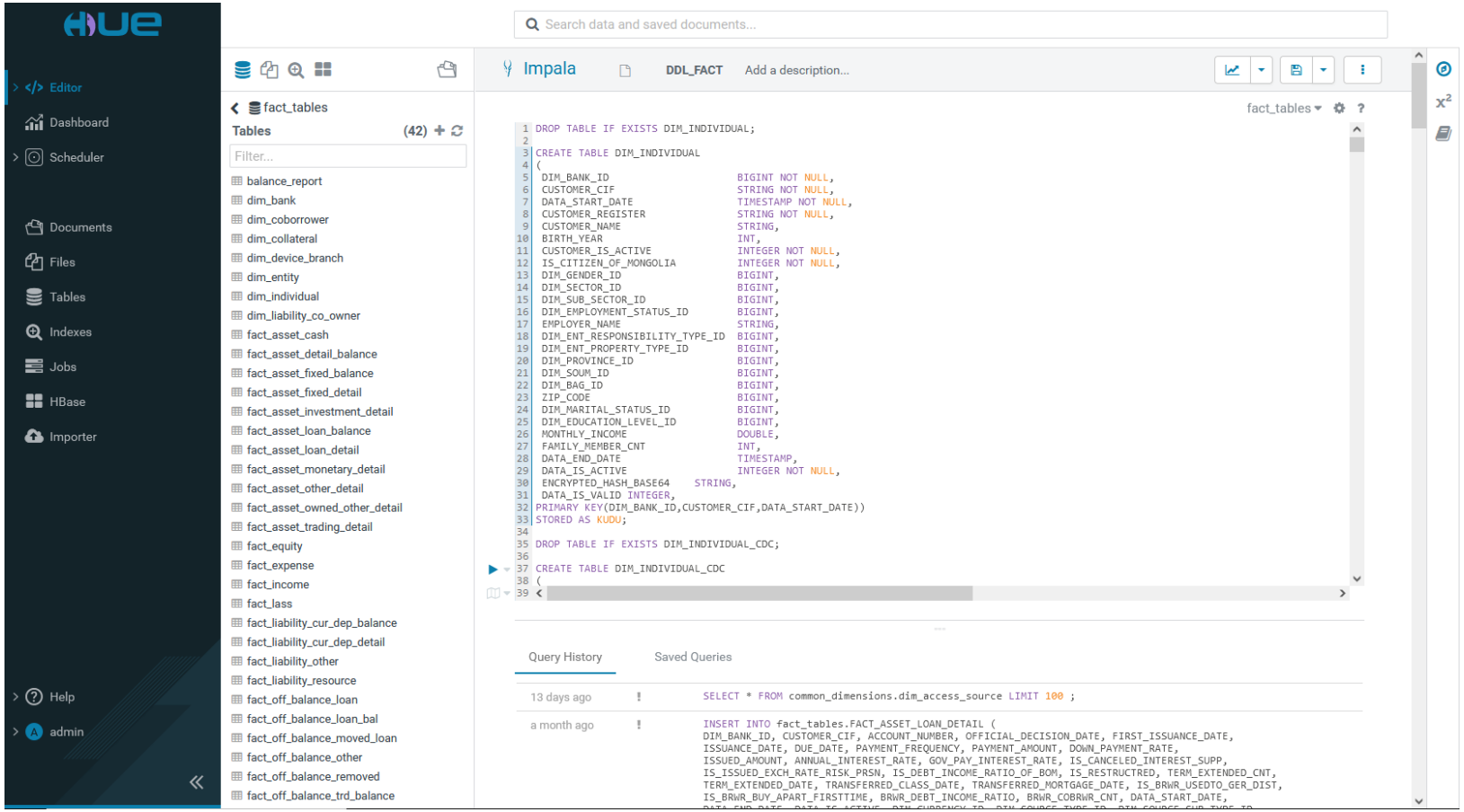
FACT_ASSET_LOAN_DETAIL table

COLUMN_NAME	DATA_TYPE	NULLABLE	NAME	DESCRIPTION
DIM_BANK_ID	number(18,0)	NO	Банкны салбарын код	DIM_BANK table-н банкны BANK_ID оруулна.
CUSTOMER_CIF	number(18,0)	NO	Үндсэн зээлдэгчийн банкны CIF дугаар	Банкны бүртгэлийн харилцагчийн давтагдашгүй дугаар оруулна.
ACCOUNT_NUMBER	number(16,0)	YES	Дансны дугаар	Дансны дугаар оруулна.
DIM_CURRENCY_ID	number(18,0)	NO	Валют төрөл	DIM_CURRENCY table-н харгалзах currency_id-г оруулна.
DIM_ASSET_CLASSIFICATION_ID	number(18,0)	YES	Хөрөнгийн ангилал	DIM_ASSET_CLASSIFICATION table-н харгалзах asset_classification_id-г оруулна.
DIM_ASSET_SUB_CLASS_ID	number(18,0)	YES	Хөрөнгийн дэд ангилал	DIM_ASSET_SUB_CLASS table-н харгалзах asset_sub_classification_id-г оруулна.
DIM_SOURCE_TYPE_ID	number(18,0)	YES	Эх үүсвэрийн төрөл	DIM_SOURCE_TYPE table-н харгалзах source_type_id-г оруулна.
DIM_SOURCE_SUB_TYPE_ID	number(18,0)	YES	Эх үүсвэрийн дэд төрөл	DIM_SOURCE_SUB_TYPE table-н харгалзах source_sub_type_id-г оруулна.
DIM_ECONOMY_ACTIVITY_ID	number(18,0)	YES	ҮА-ны эдийн засгийн салбарын ангилал	DIM_ECONOMY_ACTIVITY table-н харгалзах economy_activity_id-г оруулна.
DIM_ECONOMY_SUB_ACTIVITY_ID	number(18,0)	YES	ҮА-ны эдийн засгийн салбарын дэд ангилал	DIM_ECONOMY_SUB_ACTIVITY table-н харгалзах economy_sub_activity_id-г оруулна.
DIM_LOAN_PURPOSE_ID	number(18,0)	YES	Зээлийн зорилго (Хөрөнгө худалдан авсан, Хөрөнгө барьцаалж зээл авсан эсэх)	DIM_LOAN_PURPOSE table-н харгалзах loan_purpose_id-г оруулна.
DIM_LOAN_PRODUCT_ID	number(18,0)	YES	Зээлийн бүтээгдхүүн	DIM_LOAN_PRODUCT table-н харгалзах loan_product_id-г оруулна.
DIM_LOAN_TYPE_1_ID	number(18,0)	YES	Зээлийн төрөл (зөвхөн нэгэн зээлийн хувьд бүлэглэдэг, бусад үед NULL байна)	DIM_LOAN_TYPE_1 table-н харгалзах loan_type_1_id-г оруулна.
DIM_LOAN_TYPE_2_ID	number(18,0)	YES	Зээлийн төрөл (зөвхөн нэгэн зээлийн хувьд бүлэглэдэг, бусад үед NULL байна)	DIM_LOAN_TYPE_2 table-н харгалзах loan_type_2_id-г оруулна.
DIM_LOAN_TYPE_3_ID	number(18,0)	YES	Зээлийн төрөл (зөвхөн нэгэн зээлийн хувьд бүлэглэдэг, бусад үед NULL байна)	DIM_LOAN_TYPE_3 table-н харгалзах loan_type_3_id-г оруулна.
DIM_LOAN_TYPE_4_ID	number(18,0)	YES	Зээлийн төрөл (зөвхөн нэгэн зээлийн хувьд бүлэглэдэг, бусад үед NULL байна)	DIM_LOAN_TYPE_4 table-н харгалзах loan_type_4_id-г оруулна.
IS_CO_LENDING	number(1,0)	YES	Хамтын зээлнүүлэлт эсэх	тийм бол 1, үгүй бол 0 утга оруулна.
DIM_COOPERATION_BANK_ID	number(18,0)	YES	Хамтын зээлнүүлэлтэнд оролцогч банкны код	DIM_BANK table-н банкны BANK_ID оруулна.
DIM_ACCESS_SOURCE_ID	number(18,0)	YES	Зээл олгосон салбарын ангилал (салбар дээр олгосон зээл, интернэт банкаар олгосон зээл г.м)	DIM_ACCESS_SOURCE table-н харгалзах access_source_id-г оруулна.
REQUEST_DATE	datetime	YES	Өргөдөл/Хүсэлт гаргасан огноо	"YYYY-MM-DD" форматтай утгаар оруулна
REQUEST_AMOUNT	number(26,4)	YES	Өргөдлөөр хүссэн дүн	Тоон утгаар оруулна
DIM_LOAN_STATUS_ID	number(1,0)	NO	Зээлийн статус	DIM_LOAN_STATUS table-н харгалзах loan_status_id-г оруулна.
OFFICIAL_DECISION_DATE	datetime	YES	Зээл/Үүрэг шийдвэрлэсэн огноо	"YYYY-MM-DD" форматтай утгаар оруулна
FIRST_ISSUANCE_DATE	datetime	NO	Анх зээл/үүрэг олгосон огноо буюу гэрээгээр олгосон огноо	"YYYY-MM-DD" форматтай утгаар оруулна
ISSUANCE_DATE	datetime	YES	Зээл/Үүрэг олгосон огноо (Анх зээл/үүрэг олгосон огноо эсвэл бүтэц өөрчлөгдсөн огноо)	"YYYY-MM-DD" форматтай утгаар оруулна
DUE_DATE	datetime	NO	Зээл/Үүрэг төлөлт дуусах огноо	"YYYY-MM-DD" форматтай утгаар оруулна
CLOSED_DATE	datetime	NO	Зээл/Үүрэг төлж дууссан/хаагдсан огноо	"YYYY-MM-DD" форматтай утгаар оруулна
PAYMENT_FREQUENCY	number(8,0)	YES	Төлөлт хийдэх давтамж (жоногоор)	Тоон утгаар оруулна
DIM_DOWN_PAYMENT_TYPE_ID	number(18,0)	YES	Урьдчилгаа төлбөрийн төрөл (бэлэн төлөлт, барьцаа эд хөрөнгө г.м)	Тоон утгаар оруулна
DOWN_PAYMENT_RATE	number(18,0)	YES	Урьдчилгаа төлбөрийн хувь	DIM_DOWN_PAYMENT_TYPE table-н харгалзах down_payment_type_id-г оруулна.
LOAN_FEES	number(26,4)	YES	Зээлийн шимтгэлийн дүн	Бутархай тоон утгаар оруулна
LOAN_TOLL	number(26,4)	YES	Зээлийн хураамжийн дүн	Тоон утгаар оруулна
ISSUED_AMOUNT	number(26,4)	NO	Зээл/Үүрэг анх олгосон дүн	Тоон утгаар оруулна
ANNUAL_INTEREST_RATE	number(26,4)	NO	Зээл/Үүргийн жилийн хүү (нэмэгр)	Тоон утгаар оруулна
IRTRT_RATE_BRWR	number(26,4)	NO	Зээлийн хүүний зээлдэгчид ноогдох хувь	Бутархай тоон утгаар оруулна
GOV_PAY_INTEREST_RATE	number(26,4)	YES	Засгийн газраас төлөгдөх хүүний хувь /жилээр/	Бутархай тоон утгаар оруулна
IS_CANCELED_INTEREST_SUPP	number(1,0)	YES	Хүүний дэмжлэг цуцлагдсан эсэх	Бутархай тоон утгаар оруулна
IS_ISSUED_EXCH_RATE_RISK_PRSN	number(1,0)	YES	Ханшийн эрсдэлтэй этгээдэд олгосон гадаад валютын зээл эсэх	тийм бол 1, үгүй бол 0 утга оруулна.
IS_DEBT_INCOME_RATIO_OF_BOM	number(1,0)	YES	Монголбанкны өр орлогын харьцаагаар тооцогддог зээл эсэх	тийм бол 1, үгүй бол 0 утга оруулна.
BRWR_DEBT_INCOME_RATIO	number(26,4)	YES	Зээлдэгчийн өр орлогын харьцаа (Өр орлогын харьцаагаар ангилагддаг зээлийн хувьд бөглөнө)	тийм бол 1, үгүй бол 0 утга оруулна.
IS_RESTRUCTRED	number(1,0)	YES	Зээлийн бүтцэд өөрчлөлт орсон эсэх	Бутархай тоон утгаар оруулна
DIM_AFFECTED_FACTOR_ID	number(18,0)	YES	Бүтцийн өөрчлөлтөнд нөлөөлсөн хүчин зүйл	тийм бол 1, үгүй бол 0 утга оруулна.
DIM_RESTRUCTURE_TYPE_ID	number(18,0)	YES	Зээлийн бүтцийн өөрчлөлтийн төрөл	DIM_AFFECTED_FACTOR table-н харгалзах affected_factor_id-г оруулна.
TERM_EXTENDED_CNT	number(8,0)	YES	Хугацаа сунгасан тоо	DIM_RESTRUCTURE_TYPE table-н харгалзах restructure_type_id-г оруулна.
TERM_EXTENDED_DATE	datetime	YES	Хугацаа сунгасны дараах төлөгдөх дуусах огноо	Тоон утгаар оруулна
DIM_QUALITY_CLASSIFICATION_ID	number(18,0)	NO	Зээлийн чанарын ангилал	"YYYY-MM-DD" форматтай утгаар оруулна
DIM_TERM_CLASSIFICATION_ID	number(18,0)	NO	Зээлийн хугацааны ангилал	DIM_QUALITY_CLASSIFICATION table-н харгалзах quality_classification_id-г оруулна.
TRANSFERRED_CLASS_DATE	datetime	YES	Зээлийн чанарын ангилал шилжсэн/өөрчлөгдсөн огноо	DIM_TERM_CLASSIFICATION table-н харгалзах term_classification_id-г оруулна.
TRANSFERRED_MORTGAGE_DATE	datetime	YES	Банкны өөрийн эх үүсвэрээс хөтлөөгүй шилжсэн огноо	"YYYY-MM-DD" форматтай утгаар оруулна
IS_TRANSFERRED_TO_OFFBAL	number(1,0)	YES	Шилжсэн зээл эсэх (Off Balance)	"YYYY-MM-DD" форматтай утгаар оруулна
DIM_TRANSFER_TYPE_ID	number(18,0)	YES	Хаашаа шилжсэн	тийм бол 1, үгүй бол 0 утга оруулна.
TRANSFERRED_TO_OFFBAL_DATE	datetime	YES	Шилжсэн огноо (Off Balance)	DIM_TRANSFER_TYPE table-н харгалзах transfer_type_id-г оруулна.
IS_BRWR_USEDTO_GER_DIST	number(1,0)	YES	Зээлдэгч нь гэр хороололд эмэгдэгдэг байсан эсэх	"YYYY-MM-DD" форматтай утгаар оруулна
IS_BRWR_BUY_APART_FIRSTTIME	number(1,0)	YES	Анх удаа орон сууц худалдаж авч байгаа эсэх	тийм бол 1, үгүй бол 0 утга оруулна.
BRWR_COBRWR_CNT	number(4)	YES	Зээлдэгчийн хамтран зээлдэгчийн тоо	тийм бол 1, үгүй бол 0 утга оруулна.
BRWR_COBRWR_MONTHLY_INCOME	number(26,4)	YES	Хамтран зээлдэгчийн сарын орлого	Тоон утгаар оруулна
IS_RISK_SHARING	number(1,0)	YES	Эрсдэл хуваах нөхцөлтэй эсэх	Тоон утгаар оруулна
GUARANTEE_ORGANIZATION	nvarchar2(200	YES	Баталгаа гаргасан байгууллага	тийм бол 1, үгүй бол 0 утга оруулна.
GUARANTEE_AMOUNT	number(26,4)	YES	Баталгааны дүн	Баталгаа гаргасан байгууллагын нэр оруулна.
DATA_START_DATE	datetime	YES	Эхлэх хугацаа	Тоон утгаар оруулна
DATA_END_DATE	datetime	YES	Дуусах хугацаа	4.Тусламж хэсэгт тайлбарласны дагуу оруулна.
DATA_IS_ACTIVE	number(1,0)	NO	Статус	4.Тусламж хэсэгт тайлбарласны дагуу оруулна.
				4.Тусламж хэсэгт тайлбарласны дагуу оруулна.

Cloudera

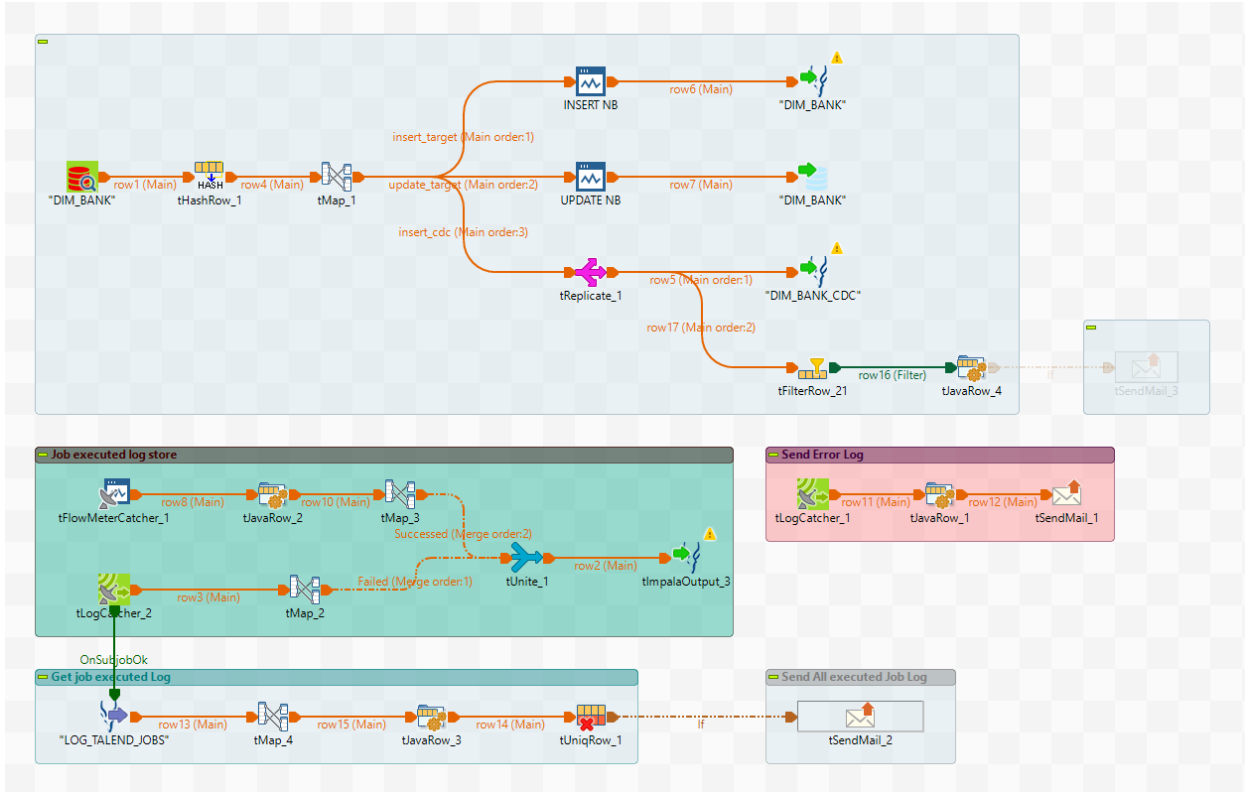
Bank of Mongolia’s Big Data cluster includes more than 20 services and 8+ servers.

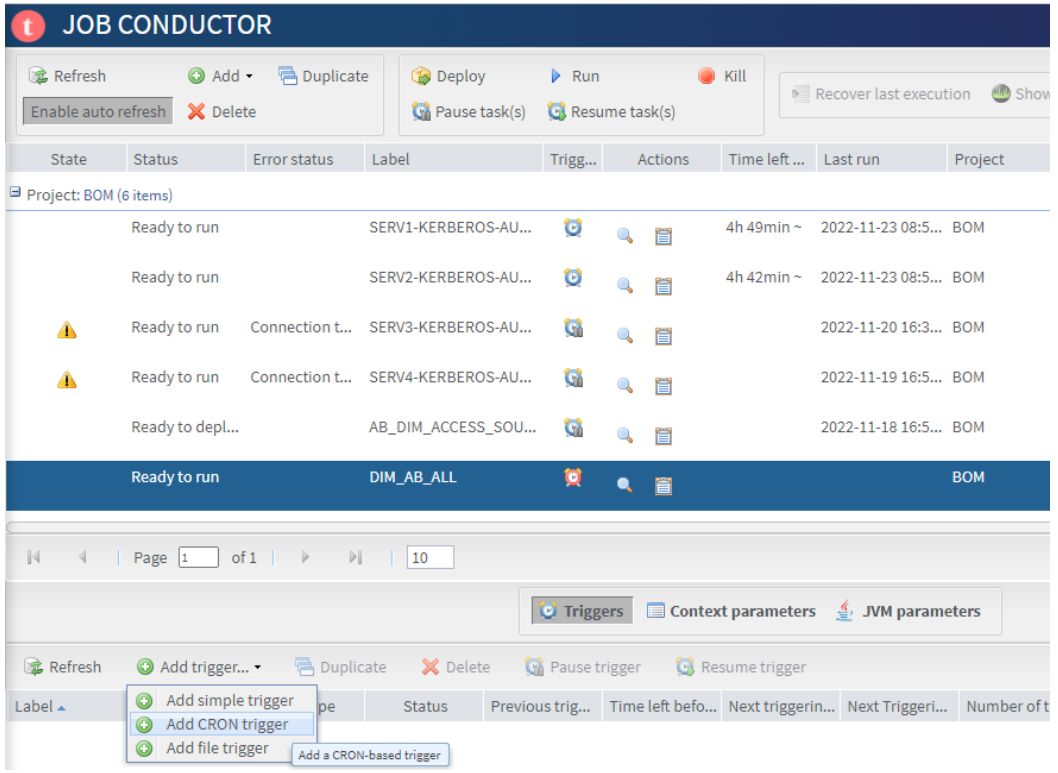




Talend Big Data

Big data system of the Bank of Mongolia has installed the TAC with 4 servers and 10 cores.



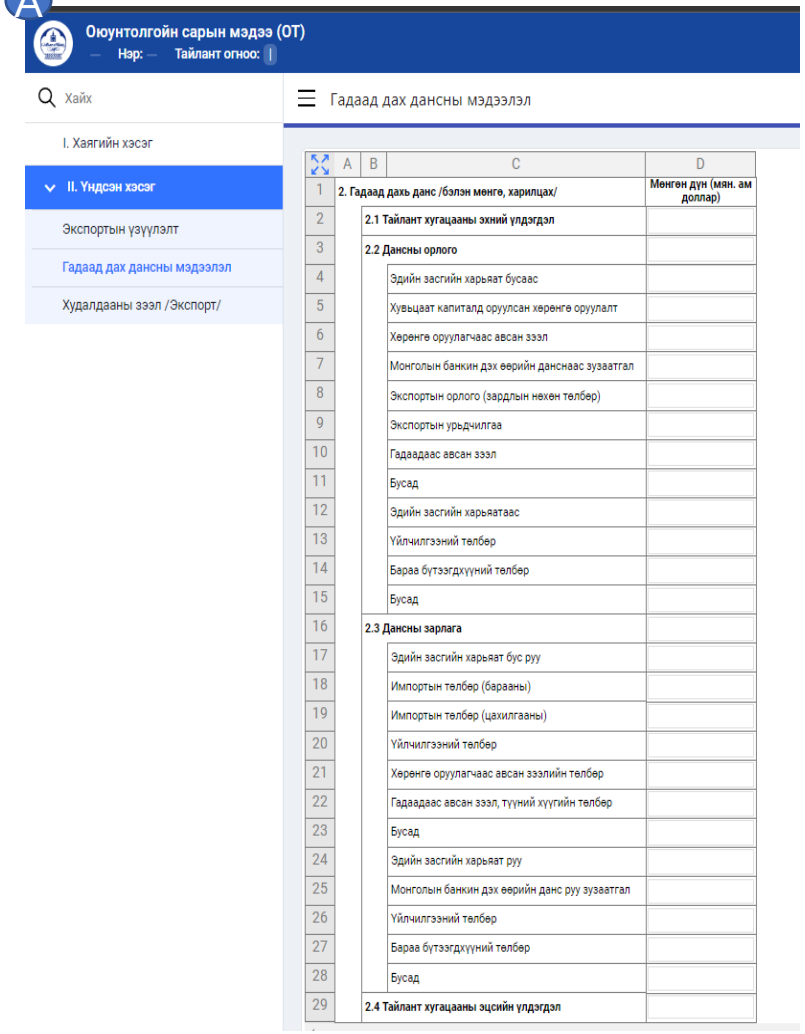


Fin Pulse

The Big Data System

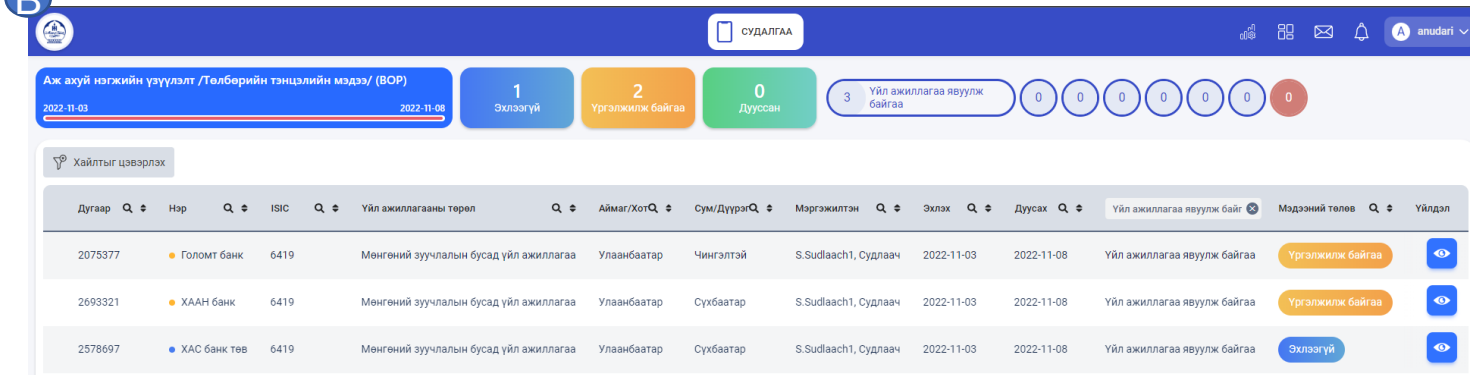
Dynamic Form

A



	A	B	C	D
1	2. Гадаад дахь данс /бэлэн мөнгө, харилцах/			Мөнгөн дүн (мян. ам доллар)
2		2.1 Тайлант хугацааны эхний үлдэгдэл		
3		2.2 Дансны орлого		
4			Эдийн засгийн харьяат бусаас	
5			Хувьцаат капиталд оруулсан хөрөнгө оруулалт	
6			Хөрөнгө оруулагчаас авсан зээл	
7			Монголын банкин дэх өөрийн данснаас зузаатгал	
8			Экспортын орлого (зардлын нөхөн төлбөр)	
9			Экспортын урьдчилгаа	
10			Гадаадаас авсан зээл	
11			Бусад	
12			Эдийн засгийн харьяатаас	
13			Үйлчилгээний төлбөр	
14			Бараа бүтээгдхүүний төлбөр	
15			Бусад	
16		2.3 Дансны зарлага		
17			Эдийн засгийн харьяат бус руу	
18			Импортын төлбөр (барааны)	
19			Импортын төлбөр (цахилгааны)	
20			Үйлчилгээний төлбөр	
21			Хөрөнгө оруулагчаас авсан зээлийн төлбөр	
22			Гадаадаас авсан зээл, түүний хүүгийн төлбөр	
23			Бусад	
24			Эдийн засгийн харьяат руу	
25			Монголын банкин дэх өөрийн данс руу зузаатгал	
26			Үйлчилгээний төлбөр	
27			Бараа бүтээгдхүүний төлбөр	
28			Бусад	
29		2.4 Тайлант хугацааны эцсийн үлдэгдэл		

B



C

"Dynamic Form“ List of Reports	
№	**Report Name**
1	Monthly Sales and Transportation Data of Coal Exporting Mongolian Enterprises (MST)
2	Foreign Trade Statistics (FTS)
3	Oyun Tolgoi Monthly Export Forecast (OME)
4	Mortage Bonds Held by the Bank of Mongolia (MBH)
5	Foreign and Domestic Ownership Concentration Survey (FDO)
6	Oyun Tolgoi Monthly News (OT)
7	Balance of Payments (BOP)
8	E-Money Report (E-MR)
9	Settlement Representative Report (SRR)
10	Mongolian Embassy's Abroad Expenditure Report (MEAER)
11	Diplomatic Mission Expenditure Report (DMEP)
12	State Budget News (SBN)
13	Cross-Border Passenger Survey (CPS)
14	Payment Card Clearing Report (PCCR)
15	Correspondent Bank News (CBN)


МонголБанк

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

New strategy of data sharing and data access in statistics: the view from Banco de Portugal[1]

Ana R Gonçalves, Mário Lourenço, Daniel V Sousa, and Thomas Verheij,
Banco de Portugal

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

New strategy of data sharing and data access in statistics

The view from Banco de Portugal

Ana R. Gonçalves[1], Daniel V. Sousa[2], Mário Lourenço[3], Thomas Verheij[4]

Abstract

Information management assumes a central role in the strategy of every organization. The increasing number of data sources, submitted according to a variety of formats and concepts, treated by different systems and subject to different quality control procedures, determined the need for Banco de Portugal to reflect on how it intends to structure its statistical production systems. Based on pillars such as the use of centralised data repositories, the integration of data acquisition and statistical compilation processes, efficient reference data and data catalogue management, this paper focuses on the harmonisation of statistical concepts as the cornerstone of this strategy, moving towards the goal of having a unified statistical production system based on a broad dictionary of concepts capable of addressing the needs of all statistical outputs.

[1] Banco de Portugal, Statistics Department – anragoncalves@bportugal.pt.

[2] Banco de Portugal, Statistics Department – dvsousa@bportugal.pt.

[3] Banco de Portugal, Statistics Department – mfllourenco@bportugal.pt.

[4] Banco de Portugal, Statistics Department – tjverheij@bportugal.pt.

Contents

New strategy of data sharing and data access in statistics 1

The view from Banco de Portugal 1

1. Introduction 3

2. Where it all started: untangling the web 4

3. The future vision for Banco de Portugal's statistical information systems 5

4. Working towards the definition of a common set of concepts 8

    4.1. Preliminary work and guiding principles 8

    4.2. Defining what to harmonise 9

    4.3. Defining how to harmonise – the harmonisation cycle 11

    4.4. Results 12

    4.5. Project outputs 14

5. Closing remarks – Where do we go from here? 16

6. References 16

1. Introduction

Information is the major asset of any modern organization. Managing information must be, therefore, at the very heart of every organization's business strategy. This is particularly true for National Central Banks, given the increasing need for data to evaluate and provide insights on ever more complex and dynamic economic phenomena.

In Banco de Portugal, the need to deal with increasingly larger datasets, transitioning from "stable" aggregated reports to unstructured microdata, from "controllable" data sources to a stage where virtually any economic agent can be a provider of relevant data, combined with the widespread dissemination of self-service data processing tools, provided the need to evaluate what could be done in order to have more effective information management policies and procedures. In the context of Banco de Portugal Statistics Department, the increasing number of data sources, submitted according to a variety of formats and concepts, treated by different systems and subject to different quality control procedures, determined the need to reflect on how to structure our statistical production systems. Section 2 provides more details on our departing point and why we started to envision the need to change it.

Based on pillars such as the use of centralised data repositories, the integration of data acquisition and statistical compilation processes, as well as efficient reference data and data catalogue management, a vision for the future of our statistical information systems was defined. Section 3 refers to this vision and to the steps we have already taken towards its implementation.

One of the cornerstones of our strategy is the definition of a common set of concepts aiming towards the standardisation of our statistical production systems. For this to happen, these need to progressively adopt the same broad dictionary of common concepts. This dictionary needs to be capable of addressing the needs of all statistical outputs (from more "traditional" ones, such as Monetary and Financial Statistics or Balance of Payments Statistics, to outputs concerning Non-Financial Corporations' Balance Sheet Statistics, for instance). Section 4 provides an in-depth description of the work done towards harmonising statistical concepts and implementing this dictionary of statistical terms.

Section 5 closes and presents the steps we foresee to implement Banco de Portugal's new strategy of data sharing and data access in statistics.

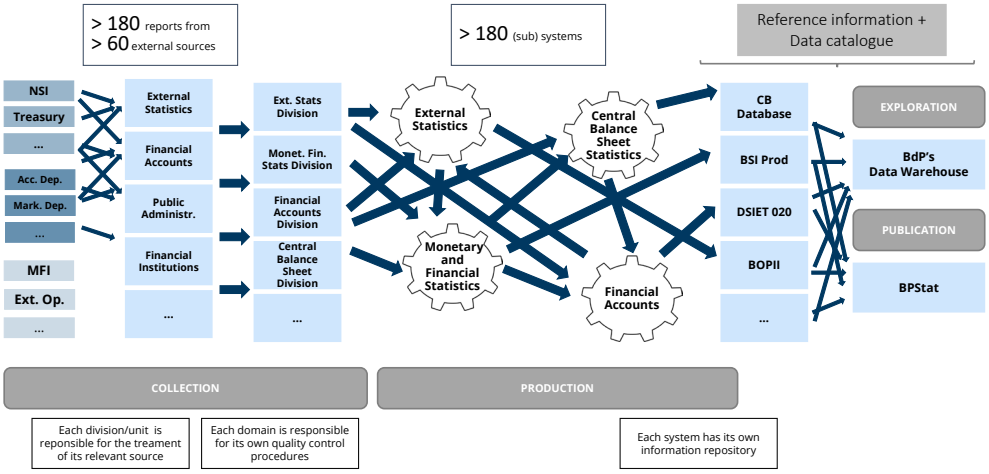
2. Where it all started: untangling the web

The information landscape has shifted dramatically in recent years for all organizations. The need to address additional information requirements or the availability of processes that allow the gathering of unstructured data, produced or sourced by various economic agents, have determined a proliferation of systems specifically designed to deal with each dataset. The global financial crisis also provided the motivation to obtain more detailed data on a wide variety of subjects. The growing use of big data and administrative data sources together with the need to produce statistics and explore datasets from different perspectives and with greater detail, contributed to the growing complexity of statistical production systems.

Banco de Portugal's statistical production systems are no different. In 2020, we listed more than 180 reports of data obtained from over 60 different external sources. These reports were collected, analysed, controlled, and used in production procedures carried out by over 180 statistical systems or subsystems (Figure 1).

Mock-up of Banco de Portugal's statistical production system in 2020

Figure 1



In this scenario, each division or unit assumes the responsibility to deal with the data it considers relevant for its own purposes. Data is, therefore, analysed several times, from different perspectives, according to the goals of each unit.

The organizational setup of Banco de Portugal's Statistics Department also contributed to the proliferation of datasets and data processing procedures. The fact that the same department is responsible for a variety of statistical production systems, with different requirements and a wide number of reporting agents, together with the growing availability of relevant administrative and public data (capable of providing further insights into the economic phenomena object of each divisions fields of expertise) led to the proliferation of datasets and data formats which were handled according to their specific sets of concepts and codes, hence undermining

the possibility to implement a common conceptual framework that would capture this complexity.

Quality control procedures were designed according to the needs of each production system and its specific conceptual framework. This ultimately meant that several units conducted similar quality control checks over the same datasets. These checks, however, could not be standardised in a way that enabled them to be useful for different production systems.

Data was then used within each production system – with several connections established directly between them – (i) to provide relevant data for comparability checks (to assure the consistency of different statistical outputs with similar characteristics) and (ii) to provide input for other statistical production systems.

Statistical outputs were stored in each system's specific database. Statistical series published in our statistics portal (BPstat), as well as shared within Banco de Portugal to be used by other departments (within our corporate data warehouse, BPW) were derived from each system's database. Each statistical system was, in the end, responsible for their own dissemination processes. Most of the production processes were, in conclusion, specifically implemented by each system for its own purposes, albeit some information management principles were already in place (for instance, data shared with internal users was recoded according to a reference information system and a data catalogue was increasingly put into place with metadata that helped users have easier access to our outputs).

The interdependences of our systems while conducting IT maintenance or overall revisions of methodology to address new needs, to fulfil new reporting requirements by international organizations or to define new data inputs became more and more challenging through the years. The fact that each system had its own specific aspects, built according to its own needs, based on its own concepts, resulted in increasing difficulties whenever the teams responsible for each system changed. The potential dissemination of know-how was reduced by these constraints, leading to an overall consciousness that day-to-day activities could be more efficient.

This determined the need to think about how this could change and led to the implementation of a set of pilot experiences towards a new setup, in line with the new vision for Banco de Portugal's statistical production systems.

3. The future vision for Banco de Portugal's statistical information systems

In 2017, Banco de Portugal launched an Integrated Data Management (IDM) Programme as part of its Strategic Plan for the 2017-2020 period, aiming at better using the available data by means of rationalisation of the processes associated with its collection and processing, promoting its effective sharing within the organization (MORENO, 2021).

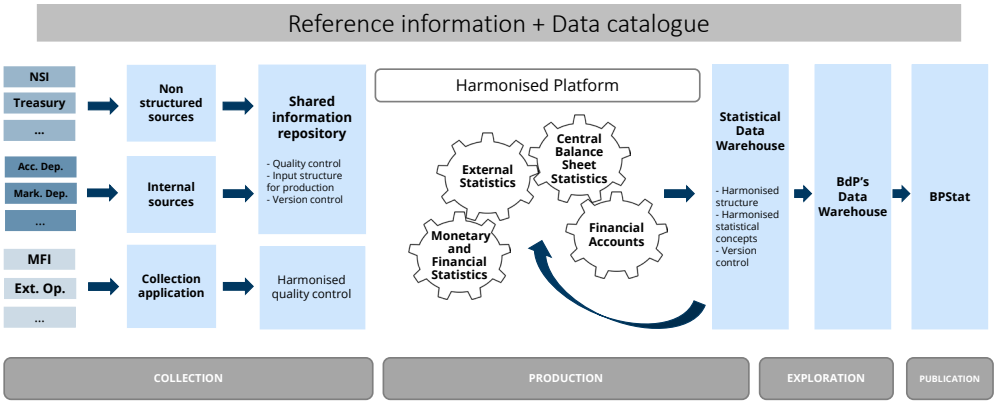
The importance of Data Management was also consecrated in Banco de Portugal's Strategic Plan for the 2021-2025 period, with the need to consolidate the IDM Programme and the evolution of our internal data governance models and data

architecture being explicitly referred as a central goal of the institution, turning Banco de Portugal into an organization focused towards data.

Within the Statistics Department, data management was one of the main concerns when a more streamlined conceptual framework for the future of our systems was designed (Figure 2). This vision highlights the role of our reference information and our data catalogue as cornerstones of our strategy, considering the importance of using concepts, dimensions and terms which are commonly understood by all systems, giving our teams full knowledge of all the data available in-house (where it is stored, how can it be accessed, which relevant terms does it include, who is the expert on a given topic in case of any doubts, etc.).

Future vision for Banco de Portugal's statistical production systems

Figure 2



This vision tries to simplify the statistical production process, basing itself in the traditional stages of any statistical production system, while trying to harmonise procedures and isolating the impacts of any changes that occur over time.

To do so, common data repositories (or warehouses) accessible by different production systems were defined, enabling the elimination of direct links between different production platforms. A primary repository stores data obtained from external sources, as well as from other departments within Banco de Portugal. Each dataset is stored in this repository only once, eliminating duplicated reports. Information can then be used by every statistical production system as they see fit, bearing in mind there is only one unified vision of the data reported and shared.

Statistical production processes will be increasingly defined within a harmonised production platform which will consist of simpler, easier to handle processes, enabling systems to use the same procedure within several production routines. For instance, if there is the need to implement a procedure to estimate a certain statistical phenomenon, the same procedure should be used in more than one system, hence ensuring the same set of tools and methodology is used to perform similar tasks.

This is also true for quality control procedures, given that most quality control checks have generally similar purposes (the identification of outliers or possibly incorrect observations which then need to be confirmed or corrected, for instance).

Another data repository stores data compiled by the different production systems, using standardised data structures defined according to harmonised statistical concepts (a set of principles explained in more detail in the next section). This statistical data warehouse will enable all statistical outputs to be stored in the same infrastructure. Systems that need the output of another system will only need to connect themselves to this data warehouse to gather the most recent information made available regarding any data domain.

Statistical outputs will cover all data made available to other departments within Banco de Portugal through our corporate data warehouse, as well as all statistical series published in our statistics portal, BPstat. These are, usually, subsets of the information produced by our statistical production systems. Storing each unrestricted statistical output in a single warehouse will allow all outputs for dissemination, reporting, etc, to be generated from the datasets stored in this warehouse (basing such reports in the "single truth" each system shares as its output).

This vision will enable us to simplify our systems, harmonise our processes and make them more efficient, giving way to leaner statistical production procedures, consolidated within a single framework, easier to maintain and evolve, ensuring greater homogeneity. This will also allow our processes to be more transparent (eliminating black boxes or situations where team members do not fully know the expected result of a certain action or change in an everyday procedure). Auditability is maintained to keep track of how processes change and how data is handled. It is our goal to make systems more driven towards information sharing, which will help reduce adaption costs for team members.

In 2022, the first Banco de Portugal's Statistical Department's centralised data repository was implemented. It corresponds to a platform where non-structured reported data has been increasingly integrated and shared among internal users.

Working with Banco de Portugal's IT department, the possibility to have more standard, flexible and agile processes which can be applied to different information domains has also been evaluated in the past year, namely through the conduction of pilot experiments or case studies focusing on the development of new statistical production systems with these concerns in mind.

An in-house pilot platform where standardised quality control checks are conducted (applied to different production systems and stages of the statistical production) is also being developed, as well as the repository where statistical production outputs will be stored.

For this to be possible, nonetheless, it became evident that a harmonised set of statistical concepts was needed as a corner stone of our strategy. Therefore, a task force specifically mandated towards this goal was created. The following Section describes the work developed by this task force and the issues addressed when defining Banco de Portugal's Statistics Department's harmonised statistical concepts.

4. Working towards the definition of a common set of concepts

4.1. Preliminary work and guiding principles

In March 2022, a preliminary study[5] regarding the conceptual framework of our systems, developed by members of different business areas within Banco de Portugal's Statistics Department, was presented. This study showed that, to ensure the standardisation and consolidation of our production systems, the harmonisation of "the language" in which datasets are coded is crucial, ultimately leading to:

- Greater efficiency in data processing;
- Synergies in information sharing processes;
- Easier maintenance and implementation of further developments in our information systems;
- Smaller cost of entry for team members when they change their role within the department or when they need to work with different datasets (other than the usual ones).

Between November 2022 and March 2023, a second team[6] developed a set of harmonised concepts to be adopted in Banco de Portugal's Statistics Department. This team's work started by summarising the work developed by the previous group, detailing the harmonisation of concepts and mappings that were to be implemented in a third stage.

Some proposals regarding the guiding principles and methodology for the construction of the codes and for the strategy to achieve harmonisation were considered, namely: i) the need for the codes to have clear designations, (ii) the importance of avoiding the combination of information of several dimensions into a single dimension (e.g. the sector dimension should only have information regarding the sector, excluding concepts that can be obtained by cross-referencing other dimensions), (iii) the importance of sharing the most granular data available as the output of each statistical system, (iv) the preferable use of already structured encodings, prioritizing *Statistical Data and Metadata eXchange* (SDMX) codes, and (v) the need to validate with all business areas the result of the harmonisation procedures. With these principles in mind, to find a method of harmonisation that would facilitate the development of the project, different dictionaries already available within several data catalogues and metadata repositories of Eurosystem working groups were thoroughly analysed, among which *Single Data Dictionary* (SDD), SDMX, *Data Point Modelling* (DPM), *Informal Coordination Group on Integrated Reporting by the Banking sector* (ICG IRB), *Banks' Integrated Reporting Dictionary* (BIRD) and the *Integrated Reporting Framework* (IReF).

[5] This study was conducted by Alexandra Miguel, Ana Colaço, José Alexandre Neves, Lídia Brás and Rita Poiares.

[6] The second team was composed by Diogo Barbosa, Ana R. Gonçalves, Ana Francisco, Elena Bucea and Daniel V. Sousa.

Regarding the methodology, three basic rules were considered: (i) the construction of the codes would, desirably, guarantee the same logic for all common dimensions; (ii) searches and aggregations should be simple, as "automatic" as possible, intuitive and, whenever possible, consider hierarchies; and (iii) the new harmonised codes should begin with letters, following a clear logic in the way they are designed.

4.2. Defining what to harmonise

The definition of a single set of concepts applicable to every statistical output could be done considering either aggregated datasets (usually, statistical time series) or granular datasets (individual data on specific economic agents or operations).

Evaluating the available data from a bottom-up approach (considering granular data as the starting point) would expectedly increase the comparability between datasets, providing greater flexibility regarding data aggregation and dissemination. It would imply, however, that more variables needed to be analysed, introducing more complexity in the process, which necessarily implied a more in-depth analysis of specific aspects of each set of the current concepts used within each statistical production system. Possible difficulties arising from different aggregation procedures would also not be directly addressed at this stage, as they would need to be evaluated individually, when aggregated outputs are analysed.

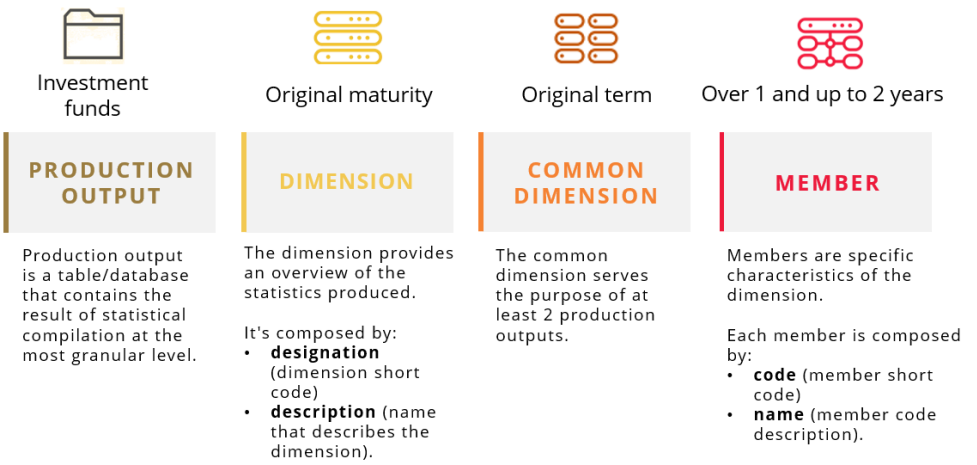
A top-down approach (starting from aggregated datasets, composed by data disseminated through reports to different international organizations or through Banco de Portugal's own publications), on the other hand, would lead to faster results and to an easier implementation, although it would be more difficult to evaluate all relevant variables and statistical concepts basing the analysis solely on aggregate data (which might not include all variables relevant for all statistical outputs). Additionally, hierarchies associated with different statistical outputs can differ, making it difficult to link different datasets. Finally, this approach would also provide additional difficulties when common dimensions were to be recoded according to the new harmonised set of concepts.

This ultimately led to the decision of harmonizing according to a bottom-up approach, starting from the most granular information available at each statistical production system, evaluating concepts autonomously but working closely with data experts to have detailed discussions whenever specific issues needed to be addressed in the definition of harmonised concepts.

The harmonisation procedure was based on a glossary of terms where statistical production outputs were assumed to be the cornerstone of the process, as exemplified in Figure 3.

Example of the application of the glossary of terms

Figure 3

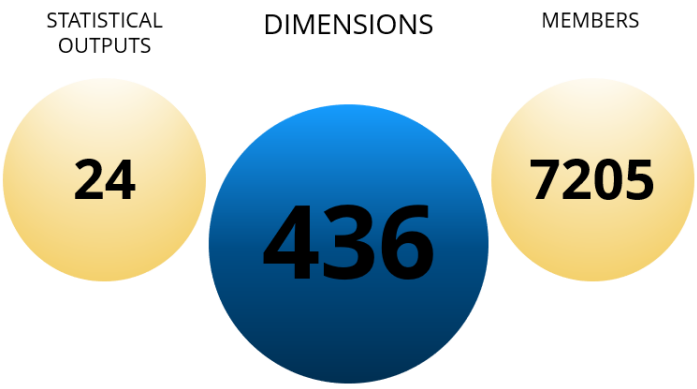


Statistical production outputs were defined as a set of tables or databases containing the results of the statistical compilation process (at the most granular level possible). These tables contain dimensions that stand for the perspective (attribute or variable) of such statistics. These dimensions are characterised using a designation (short dimension code) and a description (name describing the dimension) and contain different members, i.e., characteristics shown in the dimension, with each member being composed by a code (short member code) and a name (description of the member code). The harmonisation procedure might imply that current members could be ultimately split into members of different dimensions (designated as "complex members"). Dimensions that are part of at least two production outputs are defined as common dimensions. Dimensions that are part of only one production output are defined as specific dimensions.

Thus, 24 statistical production outputs (produced by 6 different business areas within Banco de Portugal's Statistics Department) were explored, which translated into a total of 436 dimensions to be evaluated (Figure 4).

Collection of statistical outputs, dimensions and members to be harmonised

Figure 4

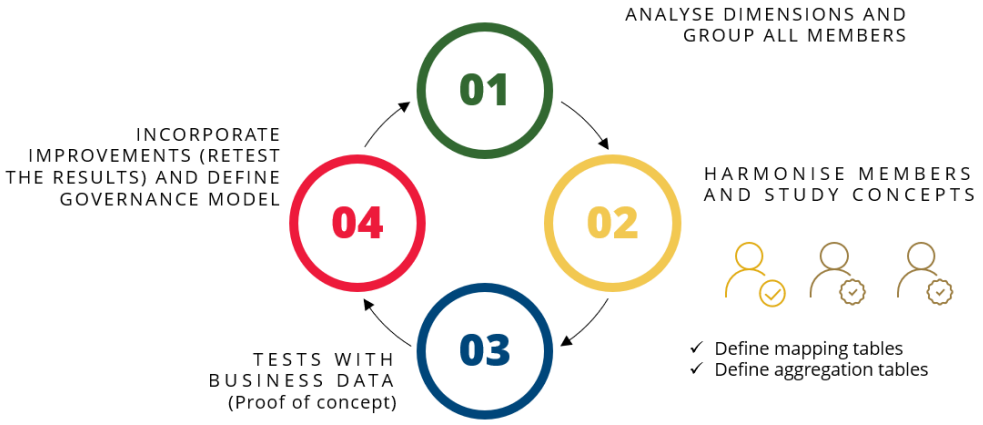


4.3. Defining how to harmonise – the harmonisation cycle

Having defined the level of detail on which the project would be based, with the objective of ensuring an unbiased and efficient process, a harmonisation cycle was defined considering several interactions (Figure 5).

Harmonisation cycle

Figure 5



The first phase of the process consisted of an in-depth analysis of the dimensions and the members of each statistical output to identify opportunities of harmonisation (considering both variables, or dimensions, and observations, or members).

In the second stage, the concepts' nature was thoroughly studied, trying to find similarities which would allow the unequivocal definition of a single set of concepts, broad enough to enclose currently analogous classifications. The outcome of this

analysis was always validated by more than one team member, hence ensuring that it would lead to a uniform and intuitive new set of concepts. At this stage, the team also defined mapping tables (linking old and new concepts), as well as aggregation tables that facilitate comparison between different levels of data granularity.

In a third moment, proofs of concept were conducted to check the validity and appropriateness of the proposed solution, testing the harmonisation procedure applied to the evaluated statistical outputs.

This was relevant to ensure the consistency of the outputs produced based on the harmonised set of concepts (keeping the new dictionary of dimensions and terms permanently updated and organised), while also providing further validation that the new set of harmonised concepts was in fact generating the intended results, namely facilitating the comparability of different statistical outputs.

Finally, in the last stage, all improvements that derived from the previous stages were implemented, redefining the harmonised set of concepts, as well as detailing a governance model for its future application and maintenance.

4.4. Results

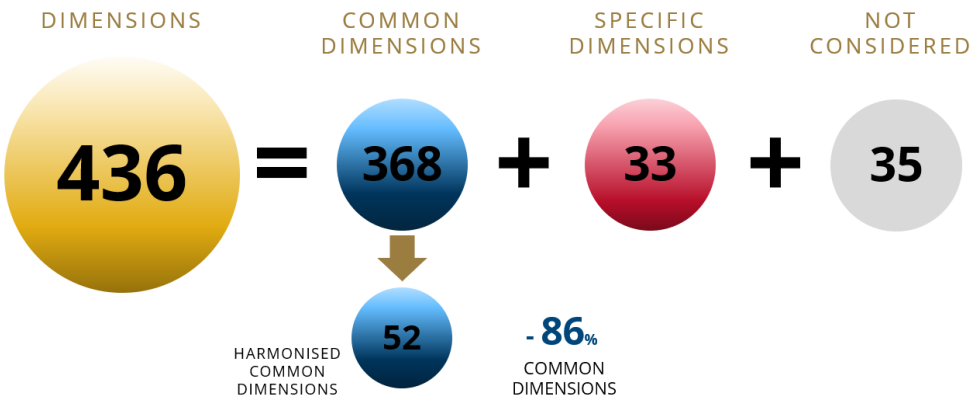
The set of 436 dimensions collected from the 24 statistical production outputs analysed was subject to the harmonisation procedure and recoded into new harmonised dimensions according to the referred guidelines and methodology.

Within the 436 dimensions analysed, 368 were considered common to at least two statistical outputs, 33 were classified as specific dimensions and 35 were deemed redundant (and, hence, excluded from the harmonisation procedure, considering it would be possible to obtain such information from other dimensions).

The set of 368 common dimensions was reduced to 52 dimensions after the harmonisation of its names and descriptions. This meant that, at the start, the 436 dimensions initially listed were reduced to a set of 85 dimensions to be further analysed (52 common dimensions and 33 specific dimensions), a reduction of 79% in the set of dimensions available (86%, if only common dimensions are considered) (Figure 6).

Harmonisation process results regarding the set of dimensions

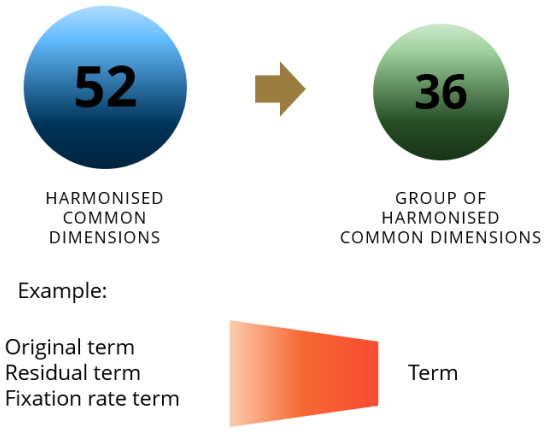
Figure 6



The 52 common dimensions were summarised into 36 groups of dimensions, considering the same set of members were shared between more than one dimension. This was the case, for example, for the dimension "Term", since it included the same members as dimensions "Original term", "Residual term" and "Refixation term".

Group of harmonised dimensions

Figure 7



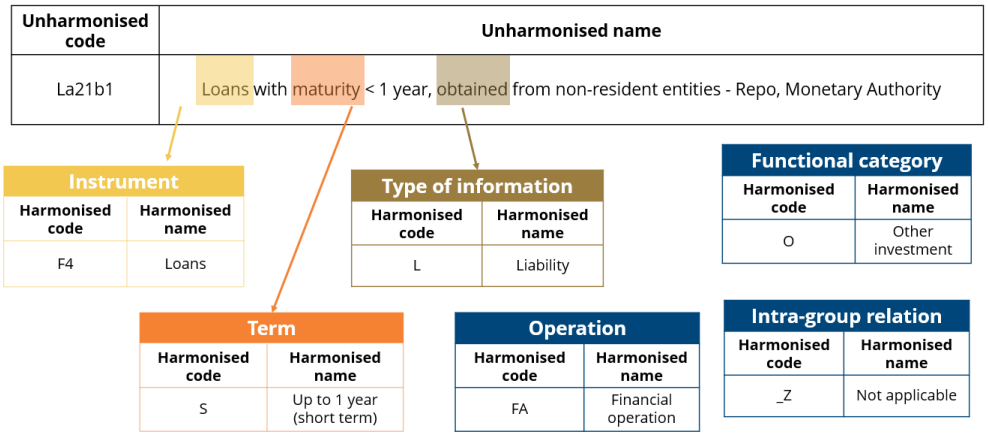
By this stage, the focus shifted into the harmonisation of members, i.e., recoding the list of attributes associated with each dimension. The harmonisation of the members of common dimensions was prioritised. From the 36 groups of common dimensions, 8 were evaluated as “dynamic”, which led to the option to have them harmonised through the definition of good practices regarding its coding and description (as was the case of the “Period” dimension, for instance). The remaining 28 groups of common dimensions were analysed one-by-one, having all members potentially included in the respective lists manually evaluated, combining business

codes and descriptions to fully understand the underlying concepts. Of the remaining 33 specific dimensions, 14 were also analysed at this stage, considering that, although they were deemed specific of a particular statistical output or business area, they shared members with other dimensions (as was the case of dimensions regarding specific uses of territory codes, for instance). The remaining dimensions were to be evaluated whenever needed (i.e., when the need to harmonise the specific outputs that use such dimensions arises).

The full list of members of the various outputs analysed led to the identification of 7205 members to be harmonised, 1610 of which were considered complex (which meant they would be segregated into a combination of several dimensions). Figure 8 exemplifies the analysis of one of these cases.

Harmonisation example for complex members

Figure 8



This analysis added complexity to the harmonisation process, generating new members as part of the attributes of common dimensions.

After the harmonisation procedure, 4104 distinct harmonised members were determined, reflecting a 43% reduction from the starting number of members considered in this exercise.

4.5. Project outputs

The thorough analysis of the dimensions and members resulted into several deliverables relevant for the organization and documentation of the process as a whole, for its effective implementation.

A dimension matrix that centralises the exhaustive survey of the dimensions collected was built, presenting the correspondence between the old dimensions of the statistical production outputs evaluated and the new harmonised dimensions, as well as a classification of whether the dimensions were considered common or specific.

A set of reference tables was also defined, including i) a dictionary of the harmonised terms (dimensions and members); ii) a glossary of terms which includes further detail on the harmonised members (to harmonise several concepts, a trade-off between the detail to be presented in the designation of the member and its specific nature had to be considered, leading to the need to have more detailed definitions documented in this glossary).

One of the main outputs of this project was the definition of mapping tables, which enable the correspondence between information currently available in the statistical production outputs and the new harmonised codes. This is a key component of this exercise, allowing systems to progressively transition from the current situation (where specific concepts are in place). Aggregation tables were also defined, enabling the comparability of the outputs produced according to different levels of granularity.

Although these deliverables are quite relevant, they would be pointless (as the whole harmonization exercise, probably) if nothing was envisioned to enable the permanent update and future implementation of the harmonised concepts and its principles.

This led to the definition of a governance that would enable the harmonisation process to outlast the initial analysis carried in the last couple of years, making it a permanent part of the statistical production in Banco de Portugal's Statistics Department's. Whenever changes in production systems occur or constraints in the current harmonised sets of dimensions and members are identified, these rules enable the permanent update of the harmonisation effort. Generic principles were agreed upon, among which the need to:

1. Follow, whenever possible, the same logic of harmonisation – at dimension (common and specific) and member levels;

2. Continue defining intuitive codes, possibly with an implicit hierarchy that will allow simple and/or automatic searches and aggregations;

3. Have member codes beginning with letters;

4. Keep having the ability to distinguish different concepts.

Specific rules were also defined for each analysed dimension. In general, priority was given to international code lists, such as SDMX codes and SDD dictionaries, adapted whenever needed to fulfil the principles agreed upon. Among others, rules were made for the way new member descriptions should be defined, as well as on how business areas should act when the need to adjust or add new elements to the set of harmonised concepts arises.

This governance clearly states that the harmonisation work conducted would be permanently subject to revisions, redefinitions, and improvements. A timeline for further developments regarding the effective implementation of these principles was also defined, as well as a set of use cases to be tested in the months following the second stage of the conceptual harmonisation process. Additionally, it was determined that a unit within the Data Management Division would be responsible for ensuring the continuity of the harmonization process and the correct application of its governance.

5. Closing remarks – Where do we go from here?

The need for a harmonised set of statistical concepts to be used by all statistical domains was, from the beginning, considered to be the cornerstone of Banco de Portugal's Statistics Department's new strategy for data sharing and data access. Efficient and homogeneous statistical production systems that can rely on centralised data repositories and quality control procedures are only possible if the information management programme is imposed in such fashion that virtually all systems and its outputs speak the same language and are defined according to the same standards.

A thorough analysis of the current business concepts adopted within our systems was conducted and a team was mandated to define a new set of dictionaries which would be the basis of the integration of our statistical outputs into a single repository, enabling its centralised dissemination.

Harmonisation could either be implemented within every single statistical production system from the start, or it could be progressively implemented whenever systems need to be changed. Given the constraints to a "big bang" implementation, a stepwise approach was considered more appropriate, defining the need, at this stage, to increasingly add statistical harmonised outputs to our final data repository. This is being done with no impact in the current production systems which remain virtually unchanged, except for the fact that they need to be able to produce their statistical outputs coded according to the rules defined by this task force. This is being done using a set of tools that allow systems to translate their current concepts into new ones. This was considered determinant to fully grasp the needs of conceptual harmonisation bearing in mind all datasets. The final stage of our production processes was also considered to be the most suitable stage to initiate the harmonisation procedures, given the need to define the data structures that are being materialised in our integrated statistical outputs data warehouse. Once harmonised codes are consolidated, they will progressively be implemented in the remaining stages of our statistical production processes, as the systems are being remodelled and further developments are implemented.

Nonetheless, this does not mean all work is done. Conceptual harmonisation is an essential step towards implementing the vision we have for our statistical production systems in the future, but it must keep being updated and revised to address new needs and bypass unidentified constraints. It is, however, a vital step to fully achieve a situation where data can be shared among different data specialists and users within Banco de Portugal, facilitating data usage and its comparability.

6. References

MORENO, M. Carmo (2021), Data Governance: an orchestra of people, processes, and technology, in IFC Bulletin No 54 (2021), Issues in Data Governance.



NEW STRATEGY OF DATA SHARING AND DATA ACCESS IN STATISTICS

THE VIEW FROM BANCO DE PORTUGAL

OCTOBER 2023

BANCO DE PORTUGAL
EUROSISTEMA

PRESENTATION OVERVIEW

01 | BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

02 | WHERE IT ALL STARTED

03 | FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

04 | DEFINING A COMMON SET OF CONCEPTS

05 | CLOSING REMARKS



BANCO DE PORTUGAL'S STATISTICS DEPARTMENT



01

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT
ORGANIZATIONAL STRUCTURE



MONETARY AND FINANCIAL STATISTICS	EXTERNAL STATISTICS	CENTRAL BALANCE SHEET	FINANCIAL ACCOUNTS	DATA MANAGEMENT	FINANCIAL SYSTEM MICRODATA	COMMUNICATION AND PLANNING	STATISTICS QUALITY

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

OUR STATISTICS

CENTRAL BALANCE SHEET STUDIES

INDEBTEDNESS OF THE NON-FINANCIAL SECTOR

EXCHANGE RATES

INCOME AND SAVING

BUDGET OUTTURN

INTEREST RATES

GENERAL GOVERNMENT FINANCIAL ACCOUNTS

SECURITIES

ECONOMIC AND FINANCIAL INDICATORS

INTERNATIONAL RESERVES

PENSION FUNDS

MFI BALANCE SHEET

GENERAL GOVERNMENT FINANCING

BALANCE OF PAYMENTS

EXCHANGE RATES INDICES

INTERNATIONAL INVESTIMENT POSITION

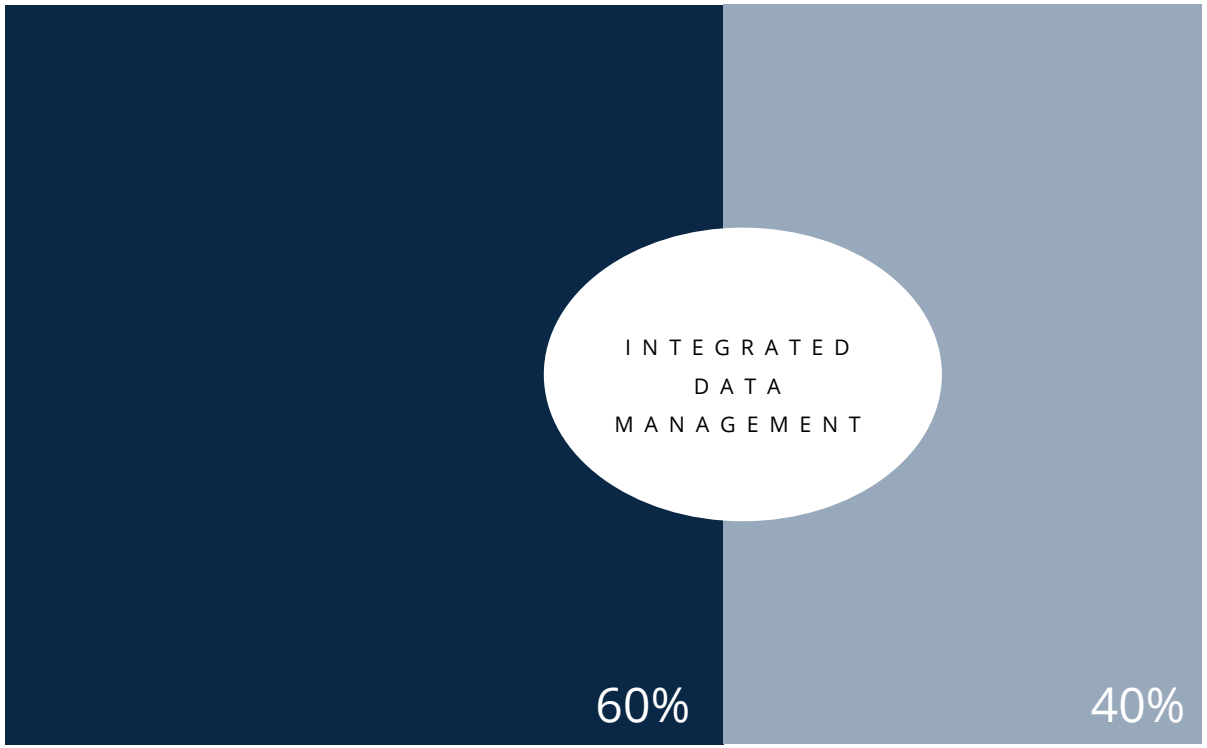
NATIONAL FINANCIAL ACCOUNTS

GENERAL GOVERNMENT DEBT

BANCO DE PORTUGAL'S STATISTICS DEPARTMENT

WHAT DO WE DO?

PRODUCE AND DISSEMINATE STATISTICS



MANAGE OTHER DATABASES

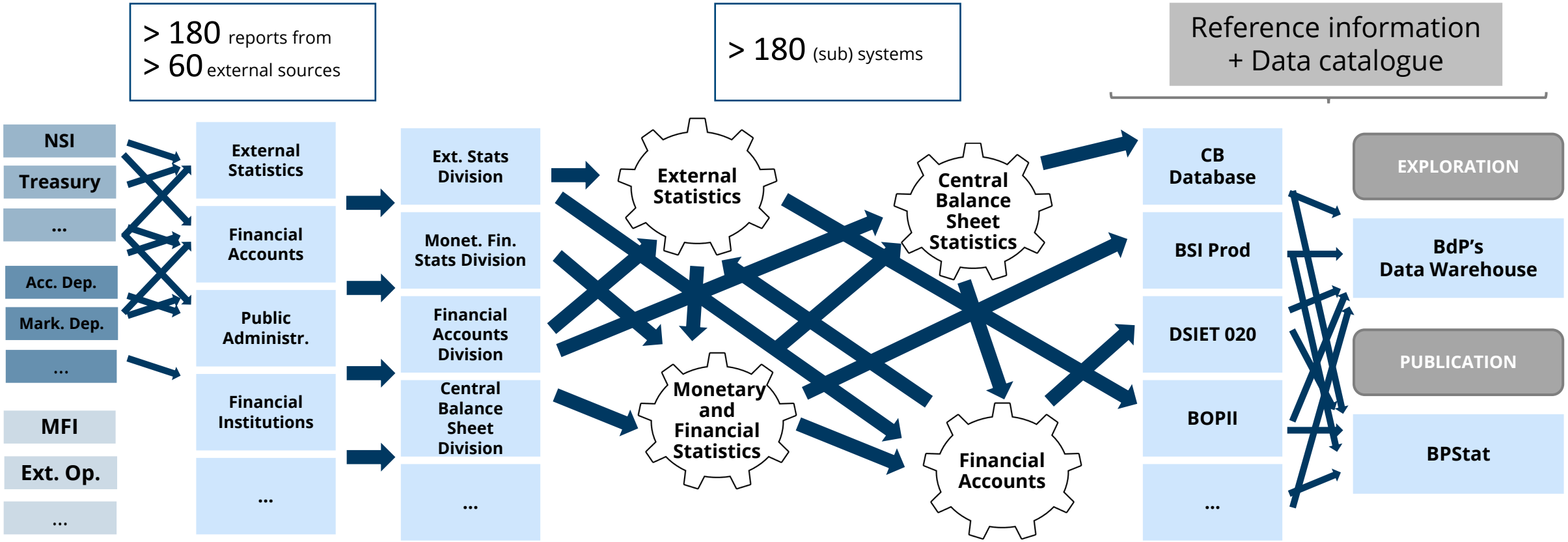
[CENTRAL CREDIT REGISTER, BANKING ACCOUNTS DATABASE, IN-HOUSE CREDIT ASSESSMENT SYSTEM]

WHERE IT ALL STARTED



02

WHERE IT ALL STARTED



COLLECTION

PRODUCTION

Each division/unit is reponsible for the treament of its relevant source

Each domain is responsible for its own quality control procedures

Each system has its own information repository

WHERE IT ALL STARTED

Main thoughts

Each domain was responsible for the treatment of its own relevant set of data.

Adapted to the purpose of each division.

No harmonized concepts, processes or systems.

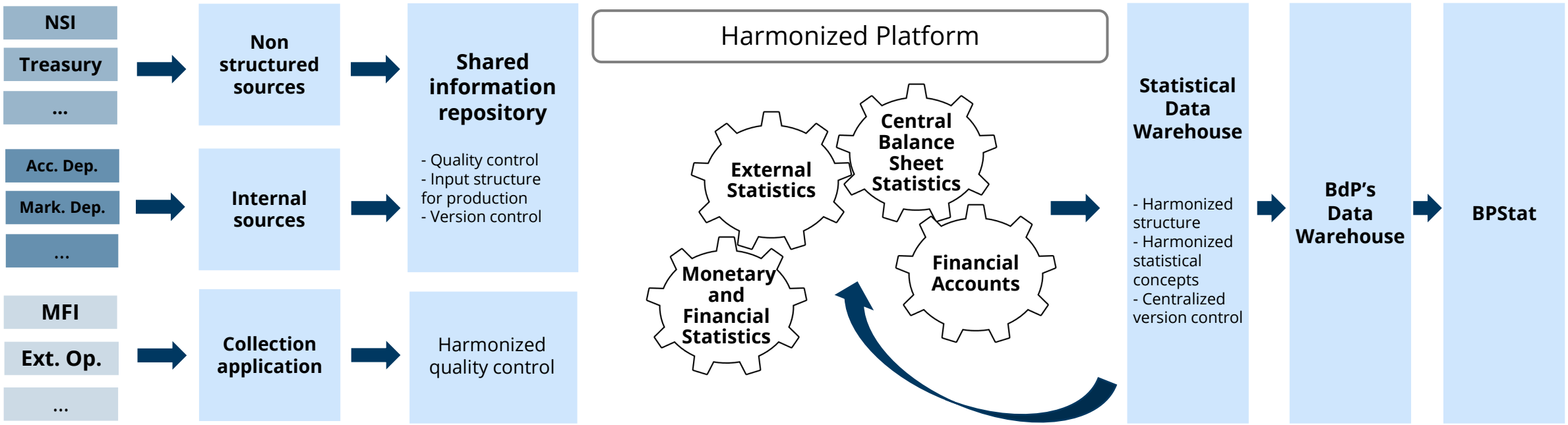
FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS



03

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS
UNTANGLING THE WEB

Reference information + Data catalogue

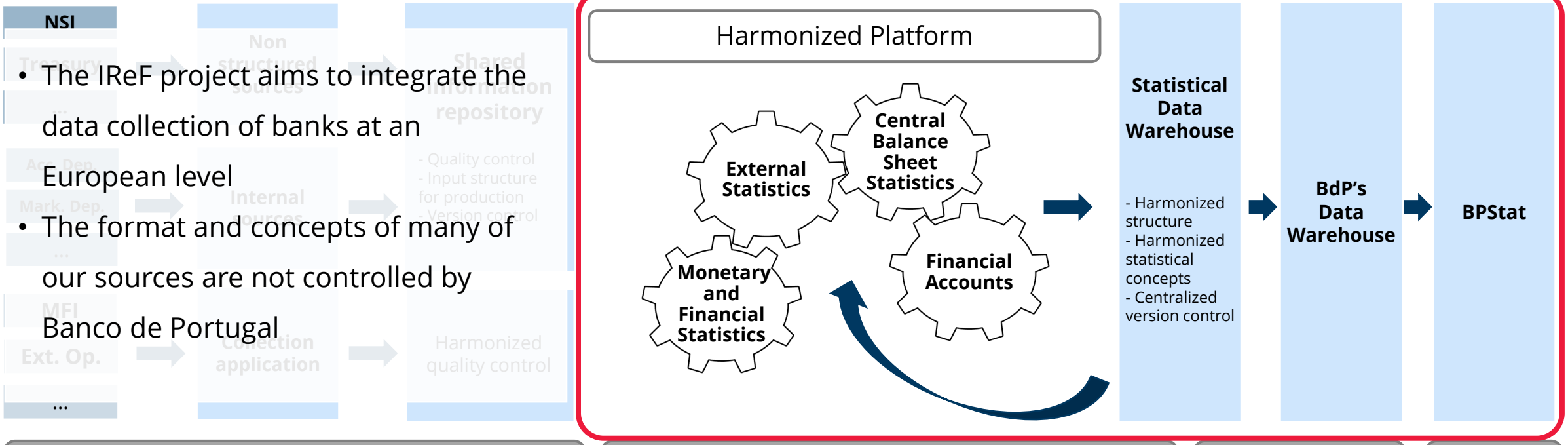


Simplify, Harmonize and Enable.

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS
UNTANGLING THE WEB

Reference information + Data catalogue

- The IReF project aims to integrate the data collection of banks at an European level
- The format and concepts of many of our sources are not controlled by Banco de Portugal



Semantic Integration

FUTURE VISION OF OUR STATISTICAL INFORMATION SYSTEMS

UNTANGLING THE WEB

Main thoughts

Simplify and harmonize our domains.

Enhance our teams' knowledge on where and how to intervene.

Homogeneous and consolidated domains, easier to maintain and evolve.

Promote transparency, auditability and efficiency in our processes.

Information sharing, with less adaption costs.

Semantic integration as a cornerstone of our strategy.

DEFINING A COMMON SET OF CONCEPTS



04

DEFINING A COMMON SET OF CONCEPTS

WHY DOES IT MATTER?

ALL DOMAINS “TALK” THE SAME LANGUAGE.

HARMONIZED CONCEPTS FACILITATE COMPARISON.

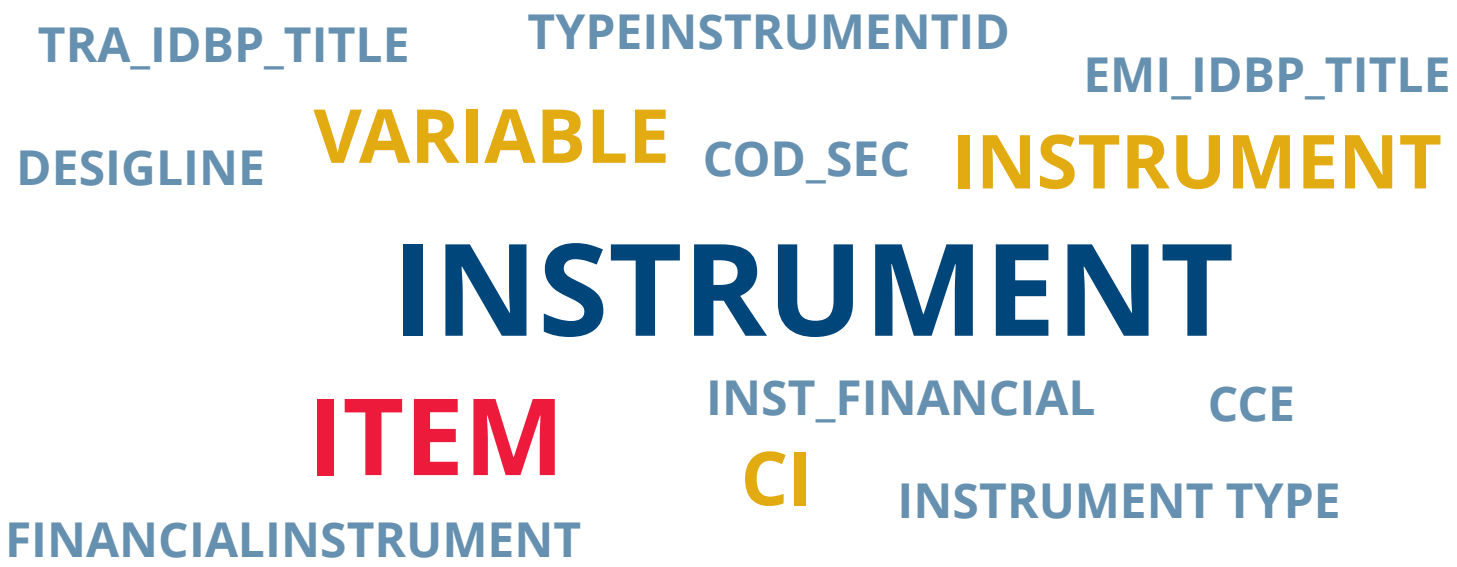
PRODUCTION “COSTS” ARE REDUCED.

DIFFERENT INTERPRETATIONS ARE AVOIDED.

FULL KNOWLEDGE OF ALL THE DATA AVAILABLE IN-HOUSE.

DEFINING A COMMON SET OF CONCEPTS
UNHARMONIZED DATA – THE ISSUE

WHEN WE SEARCH FOR THE DIMENSION **INSTRUMENT**, WHAT DO WE FIND IN THE DATABASES?



DEFINING A COMMON SET OF CONCEPTS
UNHARMONIZED DATA – THE ISSUE

MULTIPLE AND DIFFERENT SOURCES, REGULATIONS AND REPORTS.

CONCEPTS WITH HETEROGENEOUS DESCRIPTIONS.

DIMENSIONS (VARIABLES) WITH MULTIPLE CONCEPTS.

DIFFICULTIES WHEN CROSSING DATA BETWEEN DOMAINS.



DEFINING A COMMON SET OF CONCEPTS
UNHARMONIZED DATA – HOW CAN WE SOLVE IT?

A TASK FORCE TO HARMONIZE CONCEPTS.



DEFINING A COMMON SET OF CONCEPTS
HARMONIZATION STRATEGY

A | DEFINE A GLOSSARY

B | COLLECT STATISTICAL OUTPUTS

C | HARMONIZATION CYCLE

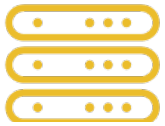
D | RESULTS

DEFINING A COMMON SET OF CONCEPTS

A) DEFINE A GLOSSARY



Investment funds (example)



Original maturity (example)



Over 1 and up to 2 years (example)

PRODUCTION OUTPUT

Production output is a table/database that contains the result of statistical compilation at the most granular level.

DIMENSION

The dimension provides an overview of the statistics produced.

It's composed by:

- **designation** (dimension short code)
- **description** (name that describes the dimension).

MEMBER

Members are specific characteristics of the dimension.

Each member is composed by:

- **code** (member short code)
- **name** (member code description).

DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS

MEMBERS

7205

HUMAN RECOURCES

5

BUSINESS AREAS

6

STATISTICAL OUTPUTS

24

DIMENSIONS

436

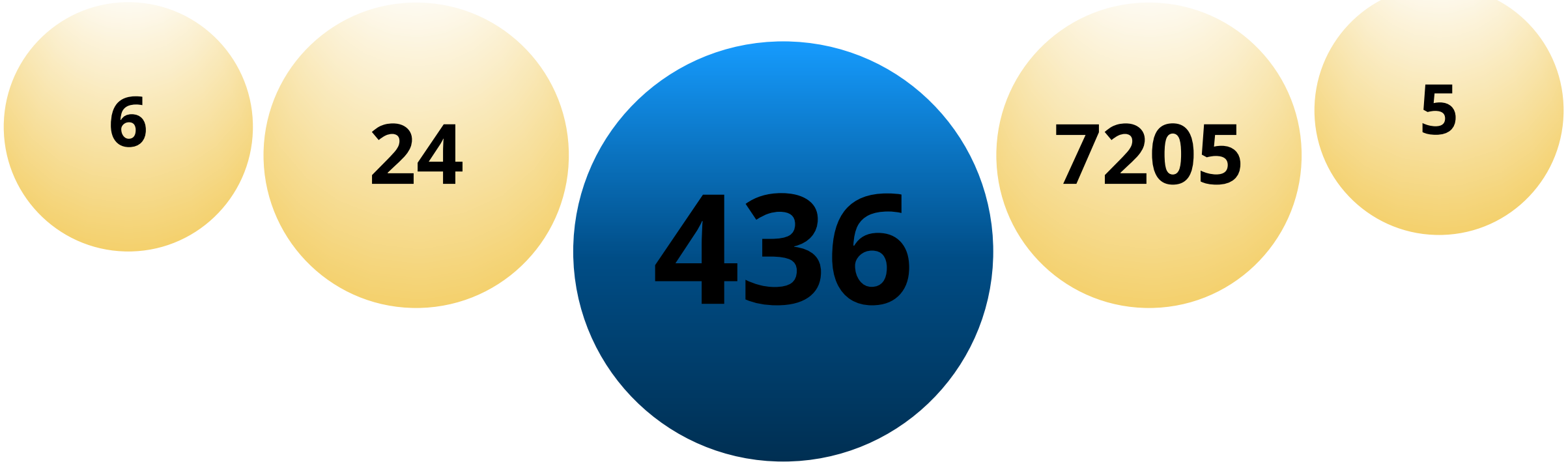
DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



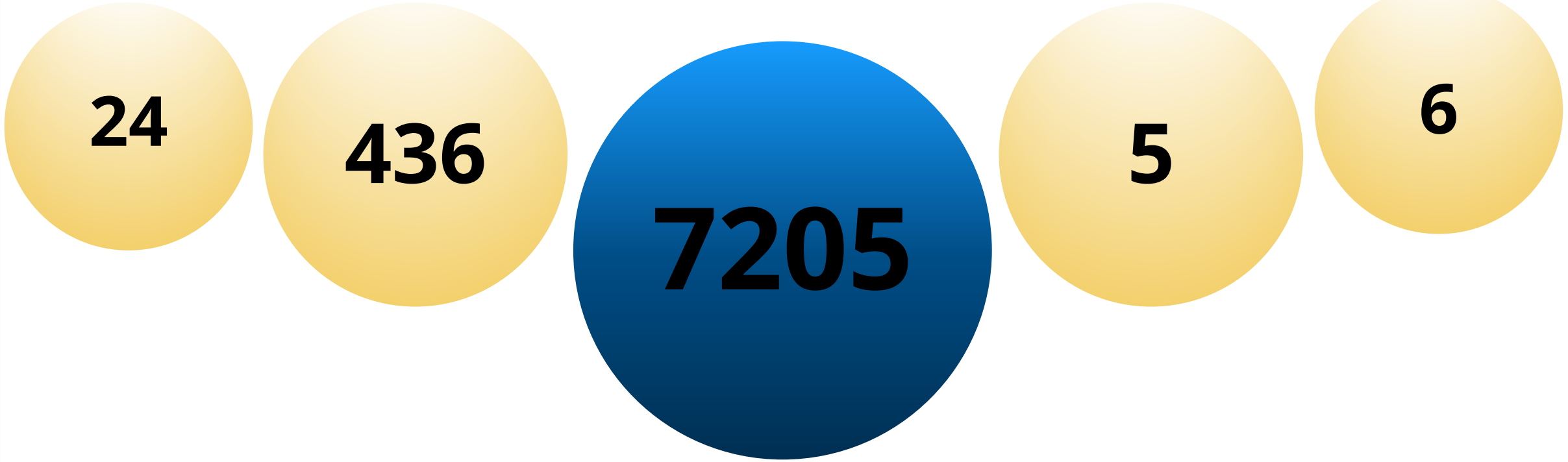
DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



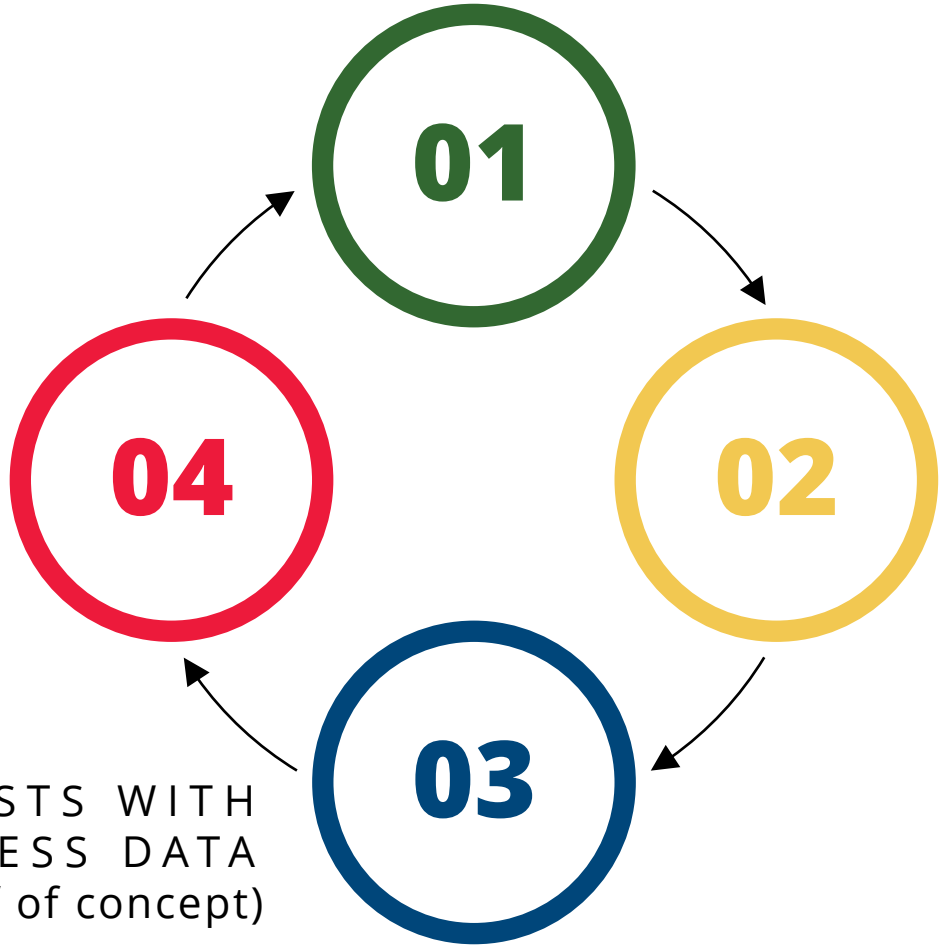
DEFINING A COMMON SET OF CONCEPTS

B) COLLECT STATISTICAL OUTPUTS



DEFINING A COMMON SET OF CONCEPTS

C) HARMONIZATION CYCLE



ANALYSE DIMENSIONS AND GROUP ALL MEMBERS

HARMONIZE MEMBERS AND STUDY CONCEPTS



- ✓ Define mapping tables
- ✓ Define aggregation tables

TESTS WITH BUSINESS DATA (Proof of concept)

INCORPORATE IMPROVEMENTS (RETEST THE RESULTS) AND DEFINE GOVERNANCE MODEL

DEFINING A COMMON SET OF CONCEPTS
D) RESULTS

DIMENSIONS

436

HARMONIZED DIMENSIONS

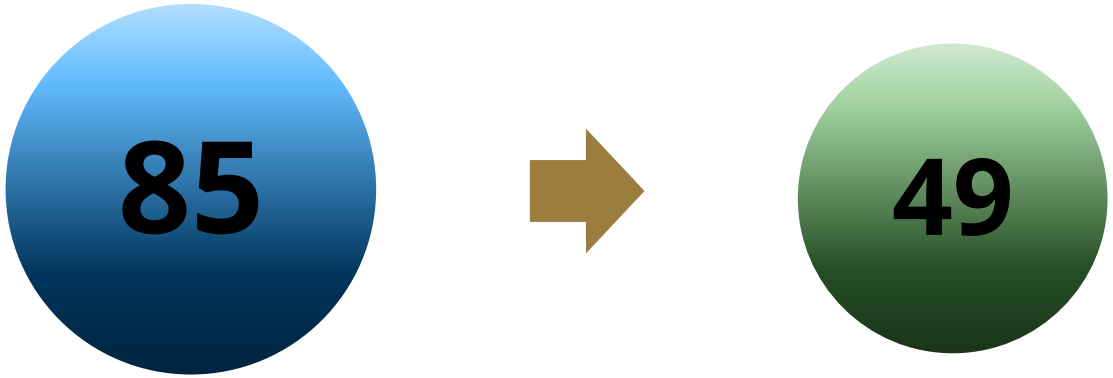
85

- 81%

DIMENSIONS

DEFINING A COMMON SET OF CONCEPTS

D) RESULTS

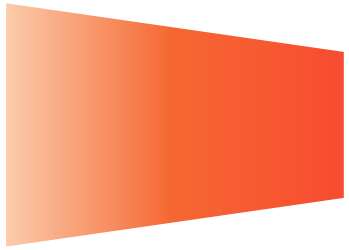


HARMONIZED DIMENSIONS

CODELISTS

Example:

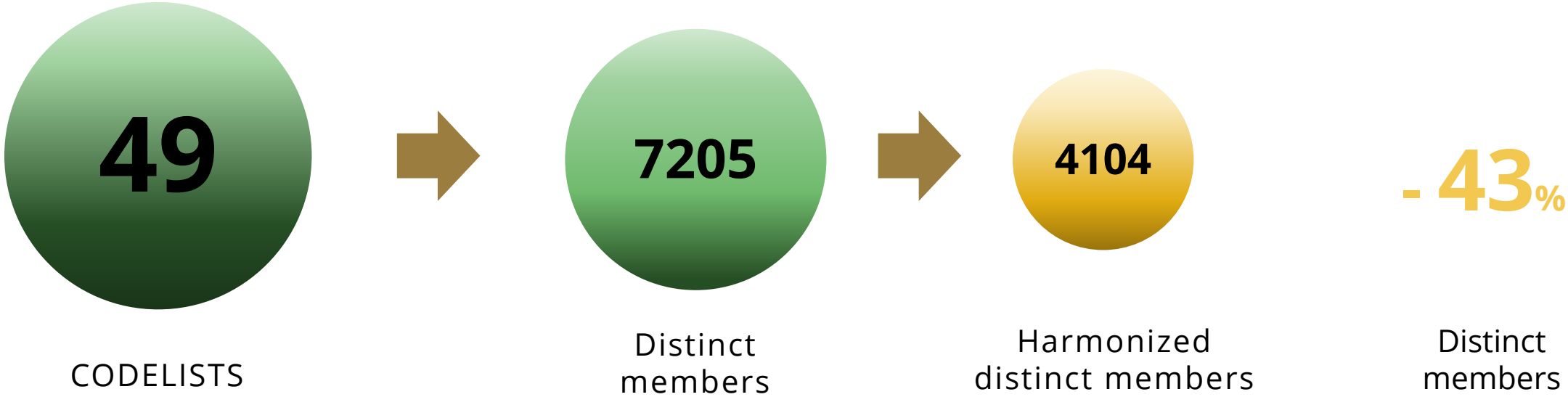
Original maturity
Residual maturity
Fixation rate maturity



Maturity

DEFINING A COMMON SET OF CONCEPTS

D) RESULTS



DEFINING A COMMON SET OF CONCEPTS
GOVERNANCE PRINCIPLES

USE ALPHANUMERIC CODES, STARTED WITH A LETTER.

FOLLOW THE STANDARD SDMX CODELISTS DEFINED IN THE SINGLE DATA DICTIONARY (SDD).

FOLLOW OTHER CODELISTS IN SDD, IF THERE IS NO MATCH FOR THE CONCEPT IN SDMX.

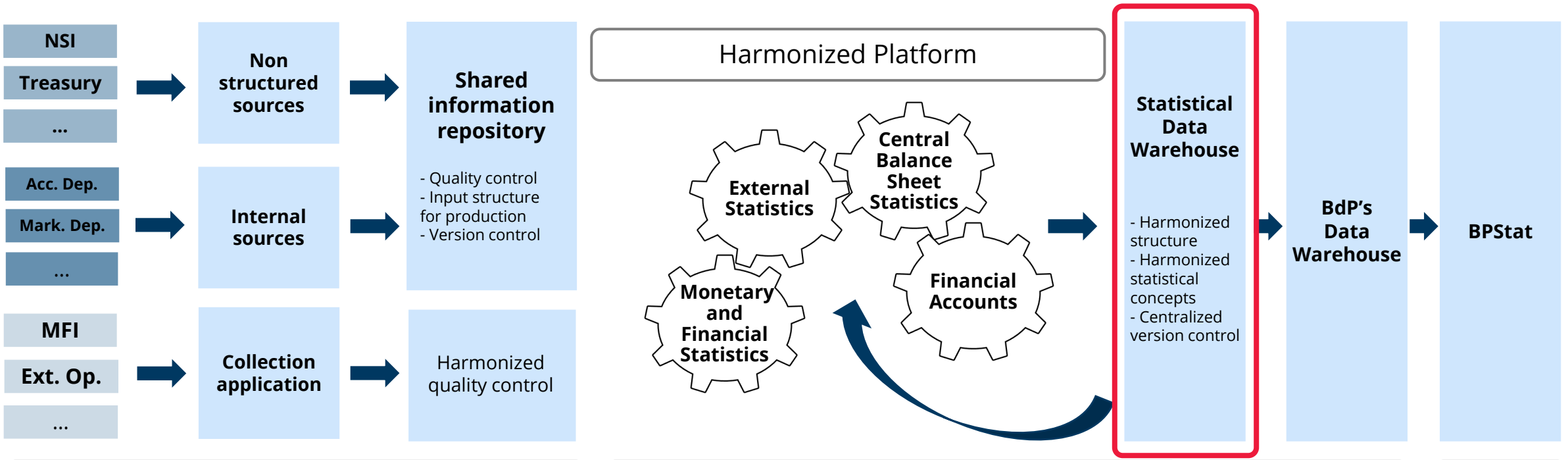
USE INTELLIGIBLE DESIGNATIONS AND CODES.

DESCRIBE THE UNDERLYING RULES FOR THE CODES OF EACH DIMENSION.

LIMIT CODE SIZE.

DEFINING A COMMON SET OF CONCEPTS
GOVERNANCE PRINCIPLES: WHERE TO START?

Reference information + Data catalogue



Simplify, Harmonize and Enable.

DEFINING A COMMON SET OF CONCEPTS

GOVERNANCE PRINCIPLES: WHERE TO START?

STATISTICAL DATA WAREHOUSE.

REFORMULATION OF STATISTICAL PRODUCTION PROCESSES.

PRODUCTION OF NEW STATISTICS.

NEW DIMENSION AND MEMBERS.

DEFINING A COMMON SET OF CONCEPTS
IMPLEMENTATION



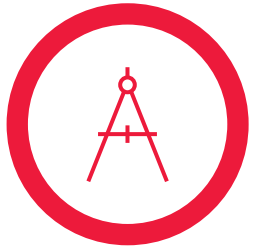
WITHIN PRODUCTION PROCESSES

4 processes already adopted the new set of concepts



WHEN STORING STATISTICAL OUTPUTS IN THE DATAWAREHOUSE

3 statistical outputs translated into the new set of concepts



WORK IN PROGRESS

Planned integration of all statistical outputs translated into the new set of concepts by the end of 2024

CLOSING REMARKS



05

CLOSING REMARKS

SEMANTIC INTEGRATION IS A CORNERSTONE OF OUR STRATEGY.

CENTRALIZED STATISTICAL PRODUCTION PLATFORM.

CROSS-REFERENCE OF DIFFERENT DOMAINS (COHERENCE AND OVERALL QUALITY).

HARMONIZATION IS ALWAYS AN ON-GOING PROCESS.

IMPLEMENTATION IS STILL A CHALLENGE: A STEPWISE APPROACH WAS CONSIDERED MORE FEASIBLE.

“COMMON LANGUAGE”.


QUESTIONS

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Individual data access and sharing protection policy: definition and case study of Bank Indonesia[1]

Johanes Iman Anugrah, Akhmad Zacky Nugraha and Sapto Widyatmiko,
Bank Indonesia

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Individual Data Access & Sharing Protection Policy

Definition and Case Study of Bank Indonesia

Johanes Iman Anugrah, Akhmad Zacky Nugraha, Sapto Widyatmiko

Abstract

The information age presents a fundamental paradox: the unprecedented potential for data-driven progress stands in stark contrast to growing concerns about individual privacy and security. Navigating this tightrope, where efficient access and exchange clash with fundamental rights, demands a robust "Individual Data Access & Sharing Protection Policy" (IDASP). This paper explores the intricacies of IDASP, examining its goals, principles, and practical implementation frameworks.

We begin by outlining the pressing need for data protection in the digital era, highlighting both the benefits of information exchange and the vulnerabilities it exposes. We then delve into the theoretical underpinnings of IDASP, defining "individual data" and exploring the rationale behind its protection. This includes safeguarding personal information, maintaining confidentiality of sensitive corporate data, and upholding fundamental principles like privacy, trust, and accountability.

The core of the paper examines the practical aspects of IDASP implementation. We present a comprehensive data access control framework built on the "5W1H" approach, addressing crucial questions about what data is protected, why access is granted, who can access it, where and when they can access it, and how the process is governed. This framework emphasizes granular control, purpose limitation, robust security measures, and transparency throughout the "Access Life Cycle," encompassing pre-access onboarding, monitored usage, cautious extensions, and swift revocation mechanisms.

Furthermore, we discuss the essential role of the Access Responsibility Matrix (RACI) in defining clear roles and responsibilities for data access and sharing within an organization. This ensures accountability at all levels, promotes responsible data handling, and fosters trust in data governance practices.

This paper argues that IDASP is not merely a regulatory checkbox, but a strategic imperative for navigating the data-driven future with responsibility and integrity. By prioritizing individual control, responsible data utilization, and transparent governance, IDASP empowers organizations to leverage the power of information while preserving trust and building a data-driven world where both individuals and organizations can thrive.

Keywords: Individual data privacy, data protection, data governance, IDASP, data access control, data sharing.

Contents

Introduction/Background 3

Problem Statement 3

Theoretical Basis 4

2.1. Defining Individual Data: 4

2.2. The Necessity of Individual Data Protection: 4

2.3. Risk of Individual Data Disclosure 5

Methodology and Data Framework 6

3.1. Data Access Control By Design 7

3.1.1. Protecting Data with the 5W1H of Access 7

3.1.2. Embedded in Policy, Procedure, and Standard 7

3.1.3. Guidance for Responsible Data Handling 8

3.1.4. The Power of Collaborative Governance 8

3.2. Access Protection Criteria (5W1H Approach) 9

3.3. Access Life Cycle 9

3.4. Access Responsibility Matrix (RACI) 11

3.4.1. The RACI Matrix in Action 11

3.4.2. Benefits of the Access Control RACI Matrix 12

Conclusion and Key Take Aways 12

4.1. Conclusion 12

4.2. Key Takeaways: 13

References 13

Introduction/Background

The data-driven age presents a paradox: the unprecedented potential for information exchange and societal progress stands in stark contrast to the mounting concerns regarding individual data privacy and security. As the global digital landscape expands, blurring geographical boundaries and transforming business processes across sectors, the need for a robust "Individual Data Access & Sharing Protection Policy" (IDASP) becomes increasingly urgent.

On the one hand, global digitalization fuels an interconnected world, fostering faster access to information and resources. This megatrend unlocks unparalleled opportunities for businesses, facilitating streamlined operations, personalized experiences, and innovative products. For governments, granular data can inform evidence-based policymaking, driving economic growth and social welfare. Yet, this very interconnectedness breeds vulnerabilities.

Indonesia's commitment to data protection, exemplified by its latest Personal Data Protection Law and the Financial Sector Development and Reinforcement Law, highlights the growing recognition of this imperative. However, real-world challenges cast a shadow on this commitment. Statista's report in 2022 has revealed that numbers of global organizations became the victim to ransomware attacks in just a year, which serves as a stark reminder of the ever-present threat of data breaches.

Further complicating the issue is the inherent tension between rapid data sharing and individual protection. Businesses thrive on efficient access to and exchange of customer information, often seeking faster processing and streamlined data flows. Yet, these very processes raise concerns about potential misuse, unauthorized access, and the erosion of individual control over personal data. This dichotomy, where process efficiency clashes with privacy rights, lies at the heart of the IDASP dilemma.

Navigating this crossroads demands a nuanced approach. We must find ways to unlock the benefits of data exchange while ensuring robust safeguards for individual privacy. The urgency for a comprehensive IDASP lies not only in mitigating the risks of data breaches and safeguarding fundamental rights, but also in fostering trust and confidence in the digital ecosystem. It is time to bridge the gap between the imperatives of access and protection, paving the way for a future where both individuals and organizations can flourish in the data-driven age.

Problem Statement

In the digital age, individual data has become a valuable commodity, constantly collected, accessed, and shared across the web. While this interconnectedness offers undeniable benefits, it also raises two critical questions: why should we protect individual data, and how can we ensure its security in the face of ubiquitous access and sharing?

Finding solutions to these challenges requires a multi-pronged approach involving individuals, businesses, governments, and civil society. Embracing a comprehensive IDASP framework that prioritizes informed consent, granular control, and robust

security measures is the first step towards building a future where both data sharing and individual data protection can thrive.

Theoretical Basis

The concept of "Individual Data" encompasses a fundamental element of our digital age: the need to protect the information that defines and belongs to individual entities. This theory will explore the "why" and "what" of individual data protection, considering both personal data and certain corporate data falling under this umbrella.

2.1. Defining Individual Data:

Based on our cases, we define the term "individual data" as:

- Natural Persons (Humans): This includes any data directly or indirectly identifiable with a specific person, such as name, address, financial records, health information, biometric data, and online activity, as referred in Indonesia's Personal Data Protection Law as well as GDPR.
- Single Entity of Corporate Data (Data Subject): In certain contexts, specific corporate data can be considered "individual" if it relates to a single identifiable entity within the organization and possesses confidential or non-disclosure characteristics. This could include intellectual property, trade secrets, or sensitive financial information, or any other non-public information that could potentially cause legal, financial, or even reputational risks to the data controller or processor.

2.2. The Necessity of Individual Data Protection:

Protecting individual data transcends mere legal compliance. It's driven by fundamental principles and essential considerations, as also described in different perspective in following framework diagram in section 3.1:

1. Privacy and Autonomy: Protecting personal data safeguards individuals' right to privacy and control over their personal information. This includes preventing unauthorized access, discrimination based on personal data, and the ability to choose how and when their information is shared.
2. Confidentiality and Non-Disclosure: Certain corporate data, even though not directly tied to a specific person, can hold significant confidential or non-disclosure value. Protecting this "individual" data ensures fair competition, prevents intellectual property theft, and safeguards sensitive financial information.
3. Security and Risk Mitigation: Data breaches and unauthorized access can have devastating consequences for both individuals and organizations. Robust data protection measures minimize these risks and safeguard sensitive information.

4. Accountability and Trust: Building trust in the digital ecosystem requires transparency and accountability in data handling practices. Effective protection frameworks ensure responsible data use and empower individuals to hold organizations accountable for breaches or misuse.

Implementing effective protection measures requires a multi-layered approach:

- Granular Access Control: Individuals and organizations should have control over who accesses their data, for what purposes, and for how long. This can involve opt-in/opt-out mechanisms, tiered access levels, and detailed audit logs.
- Data Minimization and Purpose Limitation: Collecting and sharing only the minimum necessary data for specific purposes reduces the risk of exposure and misuse. Clear purpose justification and limitations on secondary data use are crucial.
- Security and Encryption: Implementing robust security measures like encryption, regular backups, and vulnerability assessments is essential to protect data from unauthorized access, loss, or alteration.
- Transparency and Accountability: Organizations handling individual data must be transparent about their practices, inform data subjects about collection and sharing, and be held accountable for breaches or misuse. Clear privacy policies, regular audits, and effective redress mechanisms are key.

The concept of "Individual Data" encompasses both personal and certain corporate data requiring protection due to its confidential nature or non-disclosure right. Recognizing the significance of individual data protection and implementing effective safeguards is crucial to upholding fundamental rights, building trust in the digital ecosystem, and fostering responsible data practices in our data-driven world.

2.3. Risk of Individual Data Disclosure

In today's interconnected world, individual data sits at the heart of countless transactions, interactions, and operations. Unfortunately, the ease of access and sharing often obscures the inherent risks associated with data disclosure. A robust "Individual Data Access & Sharing Protection Policy" (IDASP) acts as a vital shield, minimizing these risks and safeguarding the very fabric of digital trust.

Failing to uphold an effective IDASP exposes organizations to a trio of perilous consequences:

1. Legal Loss:

- Non-Compliance with Data Protection Laws: Breaches of IDASP principles can trigger hefty fines and penalties under relevant data protection regulations. Indonesia's Personal Data Protection Law and Law on Financial Sector

Development and Reinforcement stand as reminders of the legal ramifications of data security lapses.

- Lawsuits and Damages: Disadvantaged parties – individuals whose data has been compromised due to IDASP failure – have the right to seek legal recourse. This can translate into costly lawsuits and compensation claims, draining both financial resources and reputation.

2. Financial Fallout:

- Ransomware Woes: Data breaches perpetrated by malicious actors often lead to ransom demands, forcing organizations to fork over hefty sums to regain control of their sensitive information. The financial impact of such cyberattacks can be crippling, disrupting operations, and impacting profitability.
- Compensation Claims: Legal claims arising from data breaches and IDASP non-compliance can lead to significant financial burdens. Compensation awarded to affected individuals can drain resources and impact an organization's bottom line.

3. Reputational Ruin:

- Public Trust Erosion: When individual data is compromised, the organization responsible faces a steep decline in public trust. Consumers become wary, stakeholders express concern, and the brand image suffers irreparable damage. Rebuilding trust becomes a long and arduous process.
- Policy Credibility Collapse: An IDASP framework that fails to protect individual data effectively becomes hollow and ineffective. Its credibility plummets, rendering it useless in safeguarding sensitive information and fostering responsible data practices within the organization.

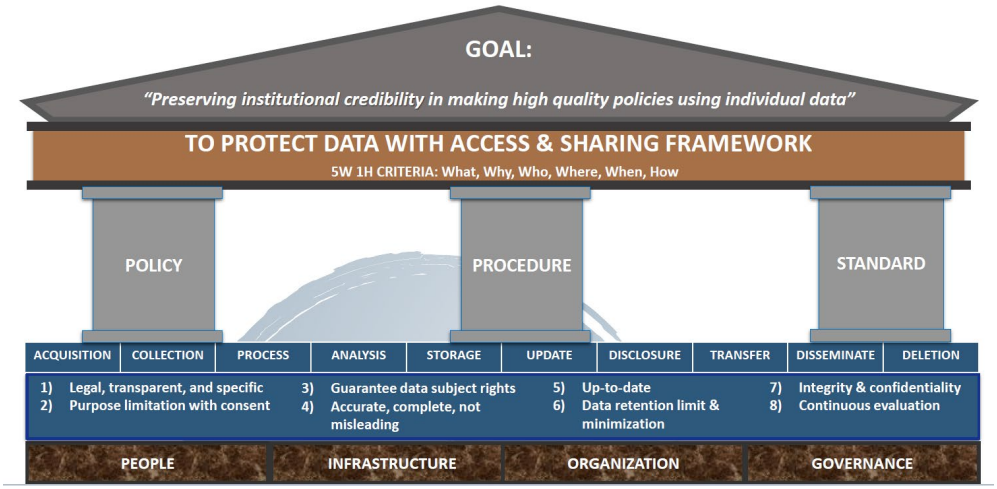
These risks highlight the critical importance of an effective IDASP. It is not merely a regulatory checkbox; it is a vital investment in protecting the organization, its stakeholders, and ultimately, the very foundation of trust in the digital ecosystem. By prioritizing robust data security measures, granular access control, and unwavering transparency, organizations can mitigate these risks and navigate the data-driven world with confidence and responsibility.

The peril of individual data disclosure in the absence of a strong IDASP cannot be overstated. The legal, financial, and reputational repercussions can be far-reaching and devastating. Embracing and effectively implementing an IDASP framework is not just a legal requirement; it is a strategic imperative for navigating the digital age with prudence and integrity.

Methodology and Data Framework

Based on the principles that we defined before, we propose a protection framework particularly in data access arrangement, which will be elaborated in this section.

3.1. Data Access Control by Design



The cornerstone of a robust "Individual Data Access & Sharing Protection Policy" (IDASP) lies in its data access framework. This framework, aptly named "Data Access Control by Design," aims to transcend mere data security by ensuring the preservation of institutional credibility in making high-quality policies based on accurate and responsibly accessed individual data.

3.1.1. Protecting Data with the 5W1H of Access

The framework operates on the principle of providing clear and granular control over data access, addressing the crucial questions of:

- What: All individual data and applications/systems used for processing, storing, or accessing it.
- Why: Users must state their objective and legal basis for access, ensuring data is used only for legitimate purposes.
- Who: Predefined user categories with specific privileges (data users, operators, contributors, etc.).
- Where: Permitted access channels like on-premise applications, internal platforms, cloud sharing, and visualization tools.
- When: Access duration based on data characteristics, urgency, and risks, with minimum of 1 month, maximum of 1 year, and possibility of ad hoc or ongoing access.
- How: A centralized Data Governance Committee oversees access, supported by a three-layer decentralized accountability system within departments.

3.1.2. Embedded in Policy, Procedure, and Standard

The Data Access Control framework is integrated into three key governance elements:

- Policy: Clearly defines objectives, principles, and user categories for data access.

- Procedure: Outlines specific steps for requesting, granting, and monitoring access rights.
- Standard: Establishes technical and security protocols for secure data access and exchange.

3.1.3. Guidance for Responsible Data Handling

This comprehensive framework is further strengthened by applying eight key guidance principles, embedded within policy, procedure, and standard:

- Legal, transparent, and specific: Access aligns with data protection laws, is processed in transparent manner to data subjects, and specifies user roles and purposes.
- Purpose limitation with consent: Access is granted only for documented purposes with informed consent from data subjects whenever necessary.
- Guarantee data subject rights: Individuals have the right to access, rectify, erase, and restrict processing of their data.
- Accurate, complete, not misleading: Data used for policymaking must be accurate, complete, and not misleading.
- Up to date: Regular data updates ensure accuracy and reliability.
- Data retention limit & minimization: Data is retained only for the minimum necessary period and then securely disposed off.
- Integrity & confidentiality: Robust security measures protect data from unauthorized access, alteration, or loss.
- Continuous evaluation: Regular framework review and updates adapt to evolving needs and regulations.

3.1.4. The Power of Collaborative Governance

Effective data access control necessitates the involvement of various entities:

- People: Individuals responsible for data access and policy implementation must be adequately trained and accountable.
- Infrastructure: Secure IT systems and infrastructure are essential for data storage and transmission protection.
- Organization: Establishing a data governance structure within the organization ensures clear lines of responsibility and oversight.
- Governance: Regular audits, assessments, and effective redress mechanisms strengthen the overall data governance framework.

By integrating these elements, the Data Access Control by Design framework fosters a culture of responsible data stewardship and builds trust with individuals. This, in turn, empowers the organization to make informed policy decisions based on accurate and secure individual data, preserving institutional credibility, and fostering continued success in the data-driven era.

3.2. Access Protection Criteria – (5W1H Approach)

This framework empowers individuals with control over their data while enabling responsible data use within the organization. It ensures transparency and trust by clearly defining:

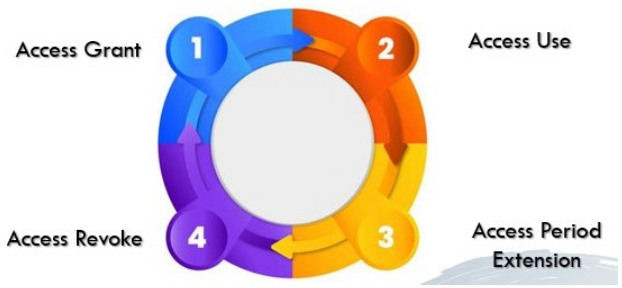
- What: All individual data and applications/systems used for processing, storing, or accessing it.
- Why: Users must state their objective and legal basis for access, ensuring data is used only for legitimate purposes.
- Who: Predefined user categories with specific privileges (data users, operators, contributors, etc.).
- Where: Permitted access channels like on-premise applications, internal platforms, cloud storage sharing, and visualization tools.
- When: Access duration based on data characteristics, urgency, and risks, with minimum of 1 month, maximum of 1 year, and possibility of ad hoc or ongoing access.
- How: A centralized Data Governance Committee oversees access, supported by a three-layer decentralized accountability system within departments.

This comprehensive framework protects individual data, builds trust, and fosters responsible data utilization within the organization. Apart from these criteria, there are still several considerations that will be added to the framework, this will strengthen the previous criteria, including:

- The framework emphasizes legal compliance and internal regulations for data access.
- Access is granted based on legitimate interests and external agreements.
- Anonymization and aggregation techniques can be used for data visualization.
- Regular review and updates ensure the framework adapts to evolving needs.

This comprehensive framework addresses the 5W1H of data access, empowering individuals with control over their data while enabling responsible data utilization for legitimate purposes. By implementing this framework, the organization fosters a culture of data privacy and builds trust with all stakeholders.

3.3. Access Life Cycle



The "Individual Data Access & Sharing Protection Policy" (IDASP) ensures continuous control and oversight over individual data access through a comprehensive "Access Life Cycle." This cycle encompasses four stages:

1. Access Grant: Stringent On-boarding
    - Rigorous identity verification confirms user eligibility and authorized roles.
    - Users articulate clear and documented purpose for requiring access.
    - Mandatory training equips users with data protection and responsible handling practices.
    - A thorough risk assessment evaluates potential issues associated with granting access.
2. Access Usage: Monitored Activity
    - Detailed logs track all user access attempts, data accessed, and actions performed.
    - Time-bound access minimizes exposure and prevents unnecessary data contact.
    - Automatic notifications warn users and administrators of approaching expiry.
    - Close monitoring ensures adherence to the approved purpose and prevents misuse.
3. Access Period Extension: A Cautious Approach
    - Extensions require a compelling justification and renewed purpose statement.
    - A thorough re-assessment of risks and potential impacts is conducted.
    - Approval requires authorization from relevant data governance parties.
4. Access Revoke: Swift and Decisive Action
    - Data subjects have the right to request access revocation under specific circumstances.
    - Security breaches or misuse by users trigger immediate access revocation.
    - Changes in user roles or status may necessitate access modification or revocation.

This cycle prioritizes individual control, responsible data utilization, and transparent monitoring. It safeguards sensitive information throughout its journey, building trust and fostering data security within the organization.

The Access Life Cycle within the IDASP framework emphasizes a proactive and vigilant approach to protecting individual data. By incorporating stringent onboarding procedures, continuous monitoring, and a well-defined process for access extension and revocation, the IDASP safeguards sensitive information

throughout its journey, upholding individual privacy and building trust in the organization's data handling practices.

3.4. Access Responsibility Matrix (RACI)

A robust "Individual Data Access & Sharing Protection Policy" (IDASP) requires not only comprehensive frameworks but also clear lines of responsibility. This is where the Access Responsibility Matrix (RACI) comes into play, delineating roles and ensuring accountability throughout the data lifecycle.

3.4.1. The RACI Matrix in Action

This matrix outlines who is Responsible, Accountable, Consulted, or Informed (RACI) for various aspects of data access and sharing within the organization. Let's delve into the specific roles for each stage:

1. Access Request:
    - Head of Department (HoD): Ultimately responsible and accountable for approving access requests, ensuring alignment with data protection regulations and departmental needs.
    - Information Manager (IM): Responsible for providing consultation and advice to the HoD during review, assessing potential risks and recommending appropriate access levels.
    - Internal Data User (IDU): Responsible for submitting clear and justified access requests, specifying purpose and legal basis for data utilization.
2. Data Usage:
    - Head of Department (HoD): Accountable for overall data usage within their department, staying informed about how data is being utilized and ensuring adherence to approved purposes.
    - Information Manager (IM): Informed about data usage patterns, identifying potential discrepancies or misuse and notifying the HoD if necessary.
    - Internal Data User (IDU): Responsible for using data only for approved purposes, upholding data protection regulations, and reporting any irregularities.
3. Data Sharing:
    - Head of Department (HoD): Accountable for all data sharing activities within their department, authorizing only legitimate and secure sharing practices.
    - Information Manager (IM): Consulted on data sharing decisions, providing guidance on secure sharing methods and potential compliance risks.
    - Internal Data User (IDU): Responsible for sharing data only in accordance with approved purposes and established protocols, ensuring recipient authorization and secure transmission.

3.4.2. Benefits of the Access Control RACI Matrix

The RACI matrix offers several advantages for the IDASP:

- Clear Delegation of Responsibility: Clearly defined roles prevent confusion and ambiguity, ensuring everyone knows who is accountable for each stage.
- Enhanced Accountability: Holding individuals accountable for their actions fosters responsible data handling and minimizes the risk of misuse.
- Improved Communication: The matrix facilitates communication between departments and roles, enabling streamlined workflow and efficient handling of data access and sharing requests.
- Strengthened Trust: By ensuring transparent and responsible data governance, the RACI matrix builds trust with individuals and stakeholders.

The RACI matrix is an invaluable tool for implementing and upholding the IDASP. By clearly defining roles and responsibilities, it fosters accountability, promotes responsible data handling, and ultimately builds trust in the organization's data governance practices.

Conclusion and Key Take Aways

Overall, the IDASP framework provides a valuable roadmap for organizations navigating the complex world of data access and sharing in the digital age. By embracing its principles and implementing its measures, organizations can protect individual data, build trust, and thrive in the data-driven future.

4.1. Conclusion

- Needs of Individual Data in Central Bank

  Central banks nowadays dive deeper into microdata utilization to produce accurate & up-to-date policy.

- Need to Protect Individual Data

  The global issue of data breach and the concern of protecting private data becomes high priority.

- Data Protection by Access Control Policy

  Governing data access and sharing is very important because human aspect is the most vulnerable.

The IDASP framework offers a comprehensive approach to protecting individual data in the digital age. By prioritizing individual control, responsible data use, and transparent monitoring, it fosters trust and minimizes risks associated with data access and sharing. Implementing this framework empowers organizations to leverage the benefits of data while upholding fundamental privacy rights and building trust with stakeholders.

4.2. Key Takeaways:

1. The balance between processing individual data for good policy-making and protecting the data can be achieved by implementing a good data protection framework.
2. A good data access protection framework should define purpose clarity, strategies, and guiding principles.
3. Data Access & Sharing Protection Framework consist of:
    a. Data Access Protection Criteria.
    b. Data Access Lifecycle & Institutional Arrangement.
    c. Data Access Responsibility Matrix.
4. Organizational structure should have a centralized data governance committee that bridges data producer/contributor and the data user.

References

D. Garcia, E. Fernandez, F. Lopez (2022): "The RACI Model: A Practical Approach to Data Protection Accountability", International Journal of Law and Information Technology.

Earley, S., Henderson, D., Sebastian-Coleman, L. (2017): "The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK)", DAMA International.

European Union (2018): General Data Protection Legislation.

Indonesia (2022): National Legislation on Personal Data Protection.

Indonesia (2023): National Legislation on Financial Sector Development and Reinforcement.

Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi and Meiko Jensen (2018): "Privacy issues and data protection in big data: A case study analysis under GDPR", IEEE International Conference on Big Data Big Data 2018, pp. 5027-5033, December.

Stalla-Bourdillon S, Thuermer G, Walker J, Carmichael L, Simperl E (2020): "Data protection by design: Building the foundations of trustworthy data sharing. Data & Policy", Cambridge Press.

Tiwari S, Sharma S, Shetty S, Packer F. (2022): "The Design of a Data Governance System", BIS Paper, no. 124, May.

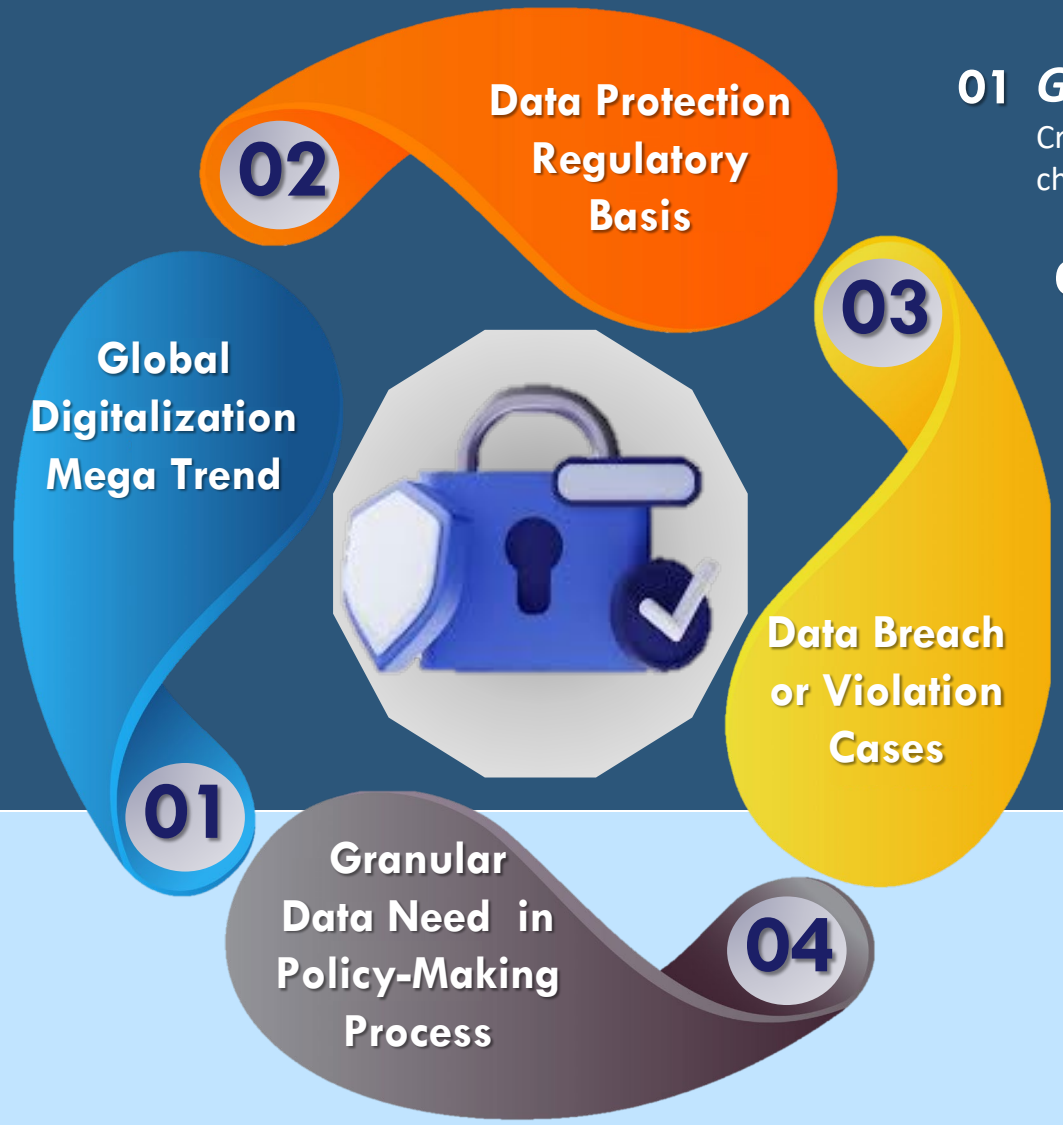


INDIVIDUAL DATA ACCESS & SHARING *PROTECTION POLICY*

DEFINITION & CASE STUDY IN BANK INDONESIA

ROME, OCTOBER 18, 2023

1 BACKGROUND:
CHALLENGES & URGENCY



Challenges in Global & National Environment

01 *Global Digitalization Megatrend*
Creating opportunities to accelerate access, activities that further increase connectivity and change business processes in various sectors.

02 *Data Protection Regulatory Basis*
Indonesia has passed Personal Data Protection Law and Law on Developing and Strengthening Financial Sector as commitment to protect confidential data

03 *Data Breach Cases*
Statista reports that 66% of organizations worldwide were victim of ransomware attacks between March 2022 and March 2023.

04 *Granular Data Need in Policy-Making*
The dynamics of macroeconomic and financial market pressures have led to an increase in the need for granular data as a basis for policy making.

Impact toward Organization's Business Process

Fast Access & Sharing on Individual Data

vs.

Individual Data Protection

2

1 PROBLEM STATEMENT

1) **Why** should we **protect individual data?**
2) **How** can we protect the individual data in terms of **access and sharing arrangement?**



2

INDIVIDUAL DATA CONCEPT

Why do We Need to Protect Individual Data?

INDIVIDUAL DATA

Covering Personal and Single Organizational Data

*'**Individual Data**' refers to…*

1. *natural person (human) and/or*
2. *single entity of corporate data (data subject)*
3. *due to its confidential nature or having non-disclosure right*

Risk of Individual Data Disclosure



Legal

- ❑ **Non-compliance** with data protection laws
- ❑ **Lawsuit** filed by disadvantaged party



Financial

- ❑ Paying ransom for data theft by hacker
- ❑ Paying compensation due to legal claims

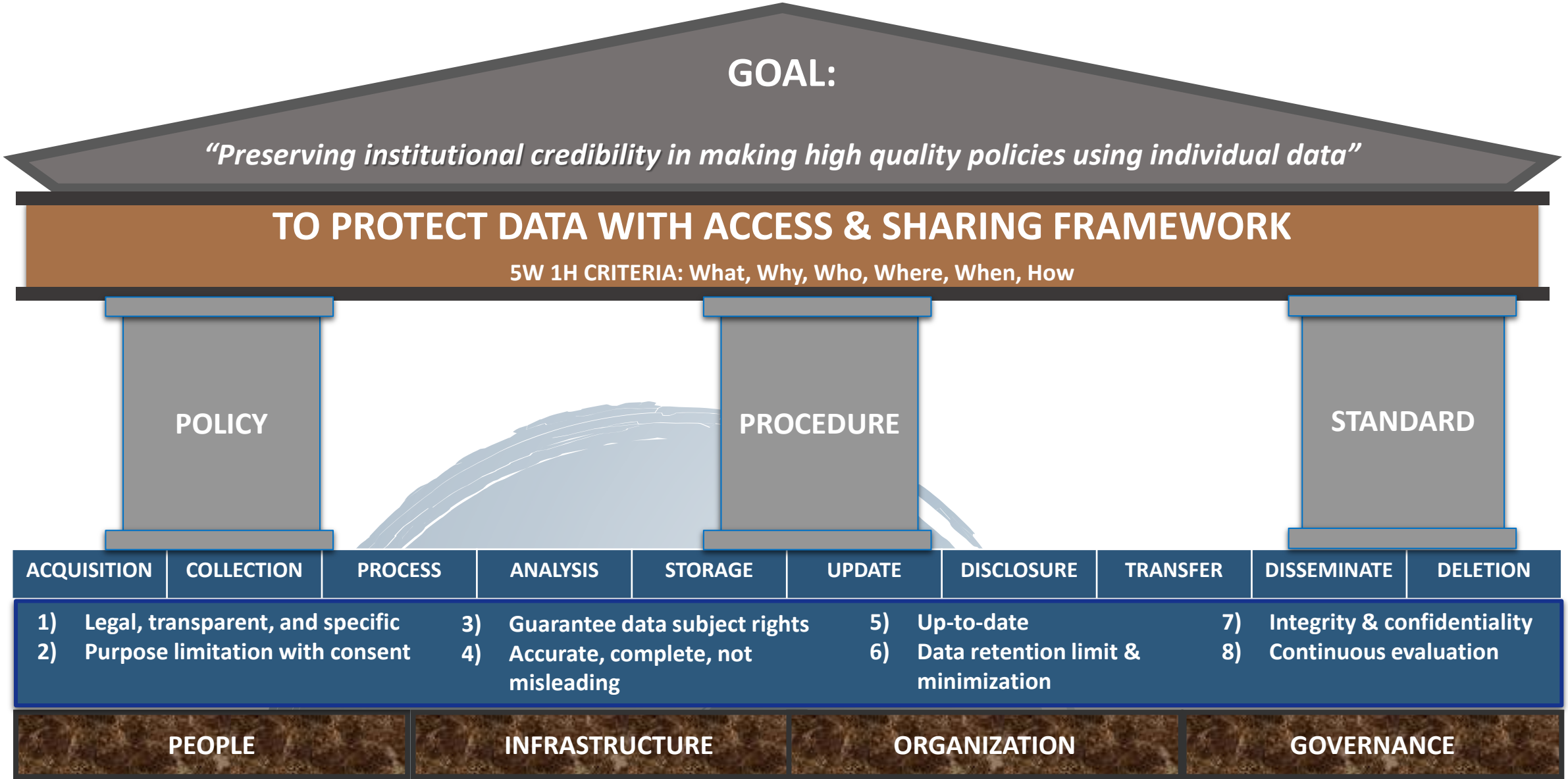


Reputation

- ❑ Damage to public trust
- ❑ May defect policy credibility

DATA ACCESS FRAMEWORK:

DATA ACCESS CONTROL BY DESIGN



4a DATA ACCESS FRAMEWORK:

ACCESS PROTECTION CRITERIA – (5W1H APPROACH)



What

1. Data/Information
2. Applications or system



Why

User should inform **the objective** and **the legal basis** for requesting access.

The basis for the provisions:
(1) National law,
(2) Internal regulation,
(3) Executive level decisions, and
(4) Other legitimate interests (e.g. MoU, and International Standards)



Who

Categories of standardized user who are eligible to be given access:
1. Data User
2. Application Operator
3. Data Contributor
4. Infrastructure Administrator
5. Supervisor
6. Legitimate External Stakeholders



Where

The Access channel may consist of:
(i) On-Premise Application,
(ii) Internal Communication Platform,
(iii) Cloud Sharing Access,
(iv) Visualization Tools.



When

1. Access must be provided depending on the **characteristics**, **urgency** and **risks** of providing access to the data/application.
2. Retention period:
   - **Minimum 1 month**
   - **Maximum 1 year**
3. Can be **ad hoc** or on **regular basis.**



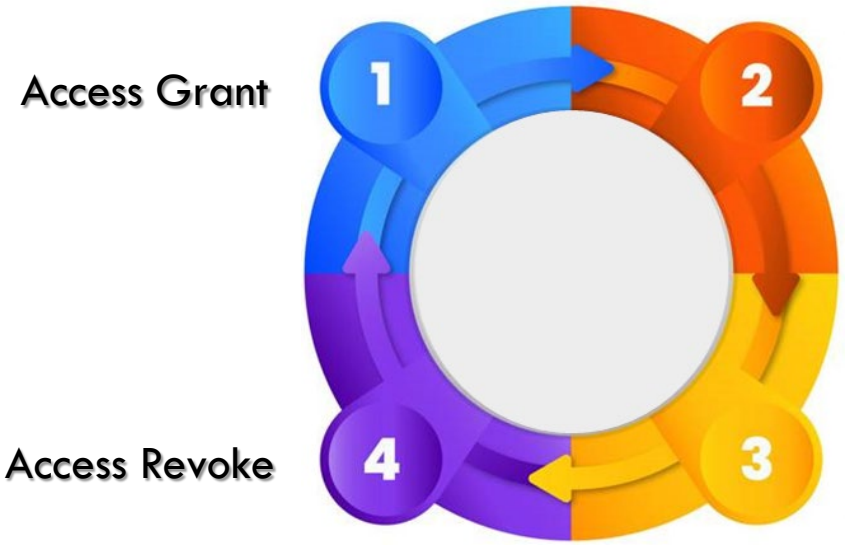
How

- Define a centralized data governance committee focus on individual data access and data sharing.
- Implementing **3-layer decentralized accountability commitment** for:
  1. Data user of a department,
  2. Information manager on the department,
  3. Head of department.

4b

DATA ACCESS FRAMEWORK:
LIFE CYCLE & INSTITUTIONAL ARRANGEMENT

ACCESS LIFECYCLE



1. **Access Grant:** the process of granting new access submitted by prospective users by completing the mandatory procedures**.**
2. **Access Usage:** the period of access usage by approved users for the specific set of time**.**
3. **Access Period Extension:** the process of extending an access retention period due to expiry.
4. **Access Revoke:** the process of revoking a user's access.

INSTITUTIONAL ARRANGEMENT

DATA CONTRI-BUTOR

DATA PRODUCER

DATA APLICATION ADMIN

Coordinator of Information Managers

DATA GOVERNANCE COMMITTEE

Access to Internal Data User

Dept. Head A	*Dept. Head B*
Information Manager	*Information Manager*
User A.1 User A.2 . . .	User B.1 User B.2 . . .

Sharing to External Parties

Individual Person

Private Organization

International Body

4C

DATA ACCESS FRAMEWORK:
ACCESS RESPONSIBILITY MATRIX (RACI)

	Head of Department	*Information Manager*	*Internal Data User*
Access Request	**RESPONSIBLE, ACCOUNTABLE**	**RESPONSIBLE, CONSULT**	**RESPONSIBLE**
Data Usage	**ACCOUNTABLE, INFORMED**	**INFORMED**	**RESPONSIBLE**
Data Sharing	**ACCOUNTABLE**	**CONSULTED**	**RESPONSIBLE**

5 FINAL WRAP-UP:
CONCLUSION & KEY TAKE AWAYS

Needs of Individual Data in Central Bank
Central banks nowadays dive deeper into microdata utilization to produce accurate & up-to-date policy.

Need to Protect Individual Data
The global issue of data breach and the concern of protecting private data becomes high priority.

Data Protection by Access Control Policy
Governing data access and sharing is very important because human aspect is the most vulnerable.

KEY TAKE AWAYS

1. The balance between processing individual data for good policy-making and protecting the data can be achieved by implementing a good data protection framework.

2. A good data access protection framework should define purpose clarity, strategies, and guiding principles.

3. Data Access & Sharing Protection Framework consist of:
   a. Data **Access Protection Criteria**.
   b. Data **Access Lifecycle & Institutional Arrangement**.
   c. Data **Access Responsibility Matrix**.

4. Organizational structure should have a centralized data governance committee that bridges data producer/contributor and the data user.

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA



ANY QUESTION?

Author:

- Johanes Iman Anugrah
- Akhmad Zacky Nugraha
- Sapto Widyatmiko

"Data Should Be Protected Like Our Bodies"

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Open-sourced central bank macroeconomic models[1]

Douglas Araujo,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Open-sourced central bank macroeconomic models

Douglas Araujo (douglas.araujo@bis.org)[1]

Abstract

Central banks and other financial policymakers rely on macroeconomic models to understand transmission channels of policy decisions, forecast the economy under different scenarios, and inform policy stances in a forward-looking way. Over time, some central banks have made the source code of their macroeconomic models publicly available, an interesting development in the decades-long increase in levels of central bank transparency. The advantages of open sourcing model code include enhanced predictability, accountability, streamlined collaboration eg with other central banks, improvements to the quality of the models and raised awareness about models. This paper proposes a framework to benchmark how macroeconomic models are open sourced according to criteria such as code accessibility, level of documentation, existence of version or vintage control, and openness to comments or proposed technical enhancements by the general public. Sharing source code of models appears to be a trend that is likely to pick up over time, as modern software engineering practices and higher levels of awareness of the advantages of open source software permeate more central banks' practices. However, choices on whether and how these models are disclosed may depend on availability of resources and objectives of the central bank. The paper concludes with practical, cost-mindful suggestions for central banks considering opening the codes of their macroeconomic models or improving their current levels of disclosure. Open-sourced macroeconomic models can be the next frontier in central bank communication.

Keywords: macroeconomic models, monetary policy, open source software, central bank communication.

JEL classification: E5, E60, C87.

1. Introduction

Central banks and other authorities have relied for decades on macroeconomic models to provide useful insights for decision-making. These models help to understand transmission channels of policy decisions, forecast the economy under different scenarios, and inform policy stances in a forward-looking way. But *model transparency*, ie the extent to which information about the models themselves is public, varies considerably across central banks, echoing the variation of transparency stances more broadly (Eijffinger and Geraats, 2006). This paper examines practices around one dimension of model transparency by central banks, namely whether and how the software codes for executing these models are publicly available. While the focus of this paper is on central banks and their models, which are mostly used for monetary policy purposes, other financial authorities are also in scope and will also feature in the analysis.

But why would code sharing in particular be a form of transparency? Central banks often publish academic papers describing their models; shouldn't that provide enough insight? A thorough description

[1] Economist, Bank for International Settlements. The author thanks Rodrigo Barbone Gonzalez, John Barrdear, Leon Berkelmans, Ben Cohen, Deniz Igan, Lauri Jantunen, Aaron Mehrotra, Benoît Mojon, Robert Munteanu, Rory Sarten, Olivier Sirello, Koji Takahashi, Ryland Thomas, Alexandre Tombini and seminar participants at the Irving Fisher Committee-Banca D'Italia's 3rd annual workshop on Data Science in Central Banking and at the Bank for International Settlements for helpful comments and suggestions. All errors are my own. This work does not necessarily represent the opinion of the Bank for International Settlements. This version is dated 11 March 2024.

of models' formulas and assumptions, as typically found in such papers, is indeed useful for interested practitioners, and an obvious requirement for model transparency. But publicly sharing the software code advances transparency even more. For one, it significantly lowers the bar for third-party users to reproduce results and more easily do scenario analysis and simulate different macroeconomic environments. Beyond that, opening model code also allows others to explore the model at a more fundamental level, testing alternative assumptions, formulas and parameter calibrations. All of the above would help spread understanding of how the model works, and could even lead to further improvements in macroeconomic modelling in the same way that much open source software benefits from user suggestions. These publicly available model codes are broadly taken as being "open source" for the purposes of this paper, although a more precise definition discussed below will help benchmark model code disclosure and inform suggestions for improving it.

Public sharing of macroeconomic model code is not an isolated phenomenon; far from it. Central banks are increasingly open sourcing some of their software in general (Araujo et al (2023c)), not just macroeconomic models. This trend can be broadly associated with the widespread internalisation of data science practices in central banking (Araujo et al (2023b)). Meanwhile, academia continues its longer-term movement towards code transparency as part of a sound and accountable scientific process. Also, key parts of macroeconomic modelling are often performed with open source software, such as Dynare (Juillard (1996)).[2] Taken together, open sourcing macroeconomic models is both a driver and a reflection of central banks modernising their macroeconomic modelling practice from these two complementary perspectives: the software or data science view (how they compute models, what are the data and computation requirements) and the academic or scientific view (how to reproduce or replicate models).

There are practical advantages to central banks that decide to open source model code. Beyond the transparency gains mentioned above, which accrue to society at large, authorities themselves benefit in several ways. First, model source code can promote enhanced predictability of policy, including from reputational gains as an accountable and transparent central bank, which can support policy credibility (Blinder (2000)). Collaboration with other central banks and international institutions is made easier by the dissemination of code. Improvements in model quality can even be obtained, for example, from macroeconomists – an audience that would be naturally interested in exploring model code – that offer suggestions to help the publishing central bank stay at the frontier of modelling practices. Other potential benefits include improved awareness about the models, leading to a clearer understanding of policy trade-offs by private sector economists, and a higher probability that newly-trained macroeconomists will already be familiar with the underpinnings of the main models.

Importantly, central banks use a suite of macroeconomic models as one input to inform policy, but not all models need to have their source code disclosed for central banks to start benefitting. As pointed out by Blanchard (2018) among others, different types of macroeconomic models are needed for policy and as a result, central banks typically have a portfolio of models.[3] Thus, gradually open sourcing one model at a time can help central banks do so in a strategic way. Once a model is open sourced, the key is really *how* they are made public. Imagine two extremes of model code disclosure: one central bank that simply shares the software with scant documentation in a hidden link in its website, and another one where the code is properly documented, easy to find and with clear channels to receive feedback or

[2] One tangible sign of the influence of Dynare in macroeconomic modelling is the large number of citations: one of its versions has more than 1000 citations as of drafting, with others summing up hundreds of references. Beyond that, the author's observation is that Dynare helped democratise macroeconomic modelling to generations of economists inside and outside of central banks.

[3] Sims (2002) describes how these models are used in the monetary policy process in a few major central banks. Post-Global Financial Crisis overviews of the evolution of macroeconomic models include Pescatori and Zaman (2011), Dou et al (2020) as well as Blanchard (2018), Hendry and Muellbauer (2018) and other papers in the same issue of the *Oxford Review of Economic Policy*. More recently, models exploring agent heterogeneity are increasingly being explored (Costa (2022)).

suggestions from any interested user. It is easy to see that the latter central bank would reap more benefits from code open sourcing. More broadly, how the code is shared influences the extent to which central banks and third-party users benefit, as well as the former's overhead and maintenance costs. A contribution of this paper is precisely to assess current model code disclosure practices, and identify the main actionable low-hanging fruits for central banks to better position themselves in this respect.[4]

But while model code disclosure can benefit authorities, in some cases a particular set of risks could be relevant. One concern is that publicly available models might crowd out private sector forecasts, or unwarrantedly reveal insights into monetary policy that interferes with policymaking. Experience in the US and other countries that have been sharing model code for some time mitigate these concerns. Another risk is that open sourcing model code could open up the central bank to technical criticism of its models, although that problem can be tackled with appropriate investment in model code quality and documentation, which is already to be expected in any case. Ultimately, model sharing might also require clear communication on the fact that particular models can have their results adjusted manually in the process of producing forecasts for monetary policy discussions. Even when these manual adjustments are considered sensitive information and therefore should not be shared, it would still be beneficial to publish its source code of the main model itself if it is not sensitive information, especially if it mechanically reflects economic equations already described in published papers.

This work rests on insights from two different literatures. The first stream, from the industrial organisation and management fields, focuses on open source software as community-based innovation (Lerner and Tirole (2002, 2005b)). This literature discusses strategic openness and open innovation by firms (West (2003), West and Gallagher (2006), Borges et al (2016)) and the motivations and drivers for voluntary participation in open source projects by third-party software developers (Shah (2006), von Krogh et al (2012)). This substantive body of prior work seeks to understand how entities – usually the private sector – decide to create or contribute to open source projects, and how software developers engage in these projects when they are often doing this as a public good (Johnson (2002)) without immediate economic remuneration. While that literature concentrates on analysing the incentives of firms such as software vendors and of private developers to distribute, use and contribute to open source software, the current paper is interested in central banks (and other financial authorities). Also in contrast to most of that literature, this paper is observational, not theoretical.

The other stream of previous work focuses on central banking policymaking, with particular attention to transparency and to macroeconomic modelling practice. Eijffinger and Gerats (2006) and Dincer and Eichengreen (most recently in 2013) show that central banks are steadily more transparent, a movement in tandem with their independence. Interestingly, already back in 2006 Eijffinger and Gerats highlight model code as a more advanced form of transparency when discussing the economic type of transparency (ie, the data, models and forecasts feeding into the policy function): "Only two central banks attain the maximum score of 3 on economic transparency, the Reserve Bank of New Zealand and the Bank of England. *The latter deserves special mention; it provides extensive documentation on its economic models, including the computer code for its macroeconometric model*" (my emphasis). Mishkin (2004) discusses how well-calibrated policy transparency can help central banks achieve their goals during crisis events. This theme has also been the subject of international guidance provided by the International Monetary Fund (IMF) to countries that volunteer for this assessment; see IMF (2020, 2023) for the most recent version of the Central Bank Transparency (CBT) code. This paper adds to this literature by concluding that open sourcing macroeconomic model code can be seen as the next frontier in central bank communication.

The next sections elaborate on the points above. Section 2 introduces concepts around code sharing that are important for the rest of the analysis. Section 3 places the topic of this paper in broader

[4] Similar exercises in central banking disclosure practice include Fracasso et al (2003), who study inflation reports, and the analyses of central banks' financial stability reports by Muñoz et al (2012).

context, including other software codes shared by central banks and how the discussion is evolving in academia. Section 4 elaborates the opportunities and challenges of model code disclosure mentioned in the introduction. Section 5 proposes criteria to benchmark code sharing of macroeconomic models by central banks and other authorities, and Section 6 presents the main results, with Section 7 assessing current central bank practices in light of those criteria. Section 8 draws on these examples to offer practical advice on how to improve the level of code-sharing for those central banks that wish to do so, and the final section concludes. Importantly, this paper does not judge or evaluate the models themselves, and neither does it compare models or their outputs with each other. The present exercise compiles publicly-available information from the macroeconomic models published by some central banks and assess the extent to which information is available and the model is easy to replicate.

2. Defining open source

Arguably the most intuitive definition of what constitutes "open source code" is very simple: code that is publicly available.[5] In the case of macroeconomic models, this would be a text file containing instructions for the computer to important any necessary data when applicable, estimate the model equations, and produce some form of output, which can be analytical (such as impulse response functions or variance decompositions) or quantitative, such as forecasts. As for the instructions themselves, they are made in one more programming languages, defined as a specific set of commands that is both readable by expert humans and understood by modelling software such as Dynare (Juillard (1996)), more generally-applicable scripting languages (such as Julia and Python) or written in compiled languages such as C++ or Rust.

A more complete and useful definition of open source code is provided by the Open Source Initiative (2023), encompassing a range of tenets. The most important points are that the code should be freely shared and distributed, in a way that allows for modifications and derived works. Access to source code should not discriminate against persons or groups, and the software should not be restricted to only certain uses (eg, only academic or policy applications). The source code license apply equally to all users, without specific carve-outs or other ad hoc instructions for particular users. These and other terms apply to promote a broad and fair access to the source code when it is considered "open". In other words, an open model code would enable a third party outside the central bank or agency in question to run their own local copy of the macroeconomic model, without having to request access or inquire for the code and wait a response. To be sure, open sourcing does not grant access to external parties to the central bank's internal systems. The public version of the code itself does not even need to be stored in the central bank servers: there are sites dedicated to hosting other entities' codes.

As noted above, whether a particular software code can be considered open source depends on the terms of its *license*. Software licenses are legal documents that describe the permissible uses, modifications or redistributions of the code. Typically, in the absence of a license these actions by the end-user would normally constitute copyright infringement, but software creators, in this case the central bank, can use licenses to explicitly carve them out – with or without conditions. While there are multiple types of standardised licenses, and software creators are free to create their own, open source is usually taken as a synonym for free software in the sense that its price is zero.

Open source software across all domains has seen widespread growth across countries in the number of unique programmers and the quantity of software code worked (Graph 1). This provides a favourable background for central banks willing to share model code, because the increasing number of

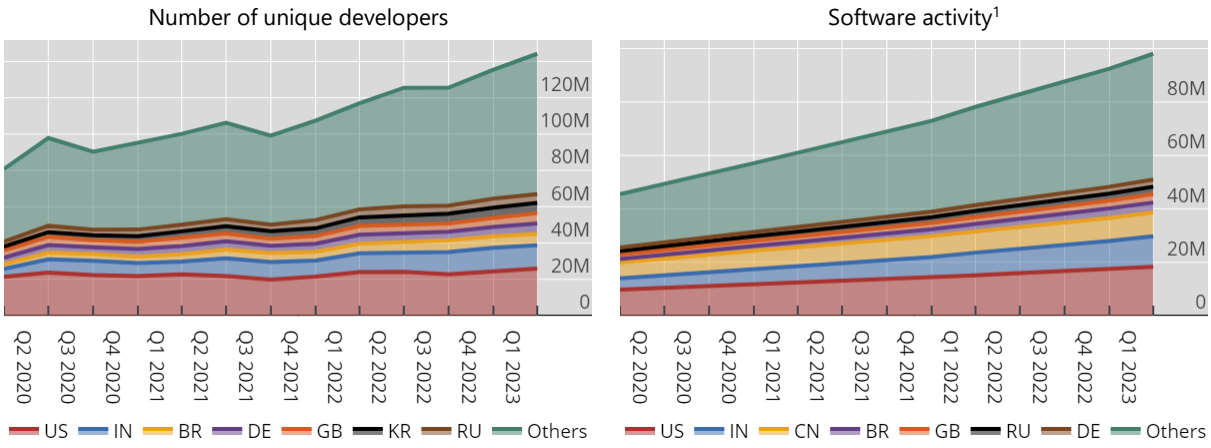
[5] A long-running quip in the computer science lore states in contrast that "all software code is open if you know Assembly", the low-level computer instruction language to which programmes are usually compiled.

people that is familiar with open source suggests that macroeconomists would also be increasingly able to engage with open sourced macroeconomic model code.

Growth in open source software activity, by location

Top 7 jurisdictions according to most recent period; all others aggregated in single series

Graph 1



Information as of 24 November 2023. [1] Software activity is defined as the number of "git pushes", which roughly corresponding to saving in a central repository one or more edited code files (ie, containing added, removed or changed code).

Source: GitHub Innovation Graph Metrics.

3. Broader context of open source and central banking

3.1 Central bank open source software

Many if not all central banks are users of multiple open source software for key IT and data science tasks, and some are also themselves sharing their own code (Araujo et al (2023)). Interestingly, the type of code being shared is heterogenous, as can be seen in the non-exhaustive list of central bank open source repositories found in Table 1.

Central bank public code repositories by type

Best efforts list of official central bank repositories[1]

Table 2

Institution	Internal utilities	Payments, CBDC, open banking	Research replication	Macroeconomic models
Central Bank of Brazil		X		
National Bank of Belgium	X			
Bank of Canada			X	X
Central Bank of Colombia			X	
National Bank of Denmark			X	

European Central Bank				X	X
Bank of France	X				
Bank of England	X			X	
National Bank of Greece		X			
Bank of Italy	X			X	X
Dutch Central Bank	X				
Bank of Norway		X			
Federal Reserve Bank of Boston (collaboration with MIT)		X			
Federal Reserve Bank of New York				X	X

Information as of 22 September 2023. [1] The helpful list of central bank open source repositories kept by the Dutch Central Bank (DNB (2023)) is kindly acknowledged.

Source: Author compilation, authorities' websites.

Repositories that contain internal utilities include formatting and styling software or data cleaning and wrangling tools. The latter in particular can be more easily used by other central banks, which often share similar data science use cases (Araujo et al (2023b)). The repositories including code related to payments, CBDC and open banking feature both experimental and production-level code. An example of the former is the Bank of Norway's sandbox for CBDC smart contracts, while the latter is illustrated by the application programming interface (API) for the Brazilian retail fast payment system Pix (Duarte et al (2022)). The research replication repositories are usually associated with specific publications. And the macroeconomic model repositories are part of the sample studied in this paper; other central banks make model code in other ways rather than in code repositories.

Similar to the repositories above, BIS work in this area also illustrates the more intense role of central banks as providers of open source software. In particular, BIS OpenTech, sdmx.io and the BIS Innovation Hub's Project Ellipse[6] (in partnership with the Monetary Authority of Singapore) all promote sharing of source code, either within the central banking community or with the broader public. Other BIS research work is open source, eg the gingado machine learing library (Araujo (2023)).

3.2 Disclosure of model code in the context of broader academic transparency

Beyond central bank transparency, disclosure of model code is also in line with broader developments in academic transparency. A number of widely-read academic journals in economics enforce some form of transparency and replicability. The definition used in this work follows the National Academies of Sciences, Engineering, and Medicine (2019): *reproducibility* is the ability to obtain consistent results with the same input data, computational steps and conditions of analysis, while *replicability* is the ability to obtain consistent results across studies that look at the same scientific question but each with its own data.

Vilhuber (2020) offers a comprehensive historical review of the evolution of code and data transparency in academic journals, most notably through the early imposition of "data availability policies". One notable evolution in this direction is the creation in December 2022 of the Data and Code Availability Standard (Koren et al (2022)), or DCAS, which defines criteria for availability of data, code, supporting materials and sharing. The DCAS was originally created by data editors of the American Economic

[6] https://github.com/bis-ih-ellipse-dkp.

Association[7], the Royal Economic Society[8], the *Review of Economic Studies*, the *Canadian Journal of Economics* and *Economic Inquiry* and later adopted by the Econometric Society[9]. The latter is also an example of a newly established (since July 2023) role of a Data Editor tasked with enforcing submission of complete reproduction packages, that should be well documented for future use. Interestingly, they go beyond the DCAS in some areas, such as calling for data formats to be readable also in non-proprietary formats.

In other words, open sourcing model code supports central banks alignment with best practices, promoting both reproducibility and replicability. In addition, this practice can help central bank research related to macroeconomic modelling reach wider impact due to alignment with leading journal practices.

3.3 Other macroeconomic model code sharing initiatives

Beyond central banks and government agencies, there are other academic-led compilations of macroeconomic model code that are worth mentioning. These resources are worth consulting for their wealth of model code, and because they can serve as starting material or benchmark models for central banks that are considering to leverage openly available code to add to their model portfolios.

The Macroeconomic Model Data Base (MMB), a product from the Macroeconomic Model Comparison Initiative (MMCI) (Taylor and Wieland (2012)) is a system that combines multiple models in a way that enables a comparison of their results. Interestingly, the website also includes guidelines and tools for authors that want to contribute models and a forum for discussion – both tools that help build a community around the models. Pfeifer (2023) and the Macrosimulation webpage (Prante and Kohler (2024)) also offer compilations of relevant macroeconomic model codes, gathering various papers and their respective Dynare code in the former case and R and Python codes in the latter collection. And because it is hosted in a public code repository system that includes a page for users to describe issues and discuss potential solutions, users can interact with one another and with the repository maintainer. Similarly, Cesa-Bianchi's (2024) webpage contain replications of landmark papers.

4. Opportunities from public disclosure of models

The advantages of open sourcing model code include enhanced predictability; more streamlined collaboration with other central banks and with international finance institutions; potential improvements to the quality of the models; and raised awareness about models' characteristics and limitations. As discussed in more detail below, the multiple ways in which central banks can opt to share code certainly affect the extent to which these opportunities are accrued. But as a community of macroeconomists start to form around a collection of open-sourced models, these opportunities can help shape research in the area and even facilitate central bank economists to incorporate any relevant crowd-sourced marginal improvements to their own models.

Making the model source code available significantly adds to model transparency, which can clarify communication and with this, improve the predictability of monetary policy, as observed by Blinder et al (2008). For example, Fracasso et al (2003) show that the quality of inflation reports lowers the

[7] Publisher of the *American Economic Review*, *AER: Insights*, *American Economic Journal: Applied Economics*, *Economic Policy*, *Macoreconomics*, and *Microeconomics*.

[8] Publisher of the *Econometrics Journal* and the *Economic Journal*.

[9] Publisher of the journals *Econometrica*, *Quantitative Economics* and *Theoretical Economics*.

uncertainty of private agents. It can only follow that sharing model code would further contribute to policy predictability, as it lowers the costs for reproduction and replication. The magnitude of the effects are less clear ex ante, as they probably depend on the starting point: a central bank that is already very transparent with respect to its models' characteristics and outputs, and how they feed into the policy process, might not reap much additional gains in predictability. Still, open sourcing models contribute to reinforce an appropriate level of disclosure and help maintain predictability, including from enhanced credibility (Blinder (2000)) that can arise from the enhanced level of transparency that is to share model code.

Model code disclosure can also directly facilitate collaboration with other central banks and international organisations. A simple use case is model comparison. Smets (1995) compared the monetary policy transmission mechanism across macro models of different central banks. If central banks as a rule made their models available, a broader range of stakeholders could compare models. Collaboration can also benefit the process of calibrating models, ie manually setting the values for some of the structural parameters in a model instead of estimating them from data. Calibration is usually done when a parameter is already well established in the literature, or when it is challenging to estimate, which calls for a better understanding of the effects of these parameters on the data-driven estimates (eg, Iskrev (2019)). For some central banks, it might also be challenging to estimate a structural parameter *de novo* due to shorter data histories, or regime and structural changes. Model code sharing can alleviate these challenges by significantly lower the bar for central bank experimentation with other policy models.

Beyond central banks, allowing third-party users to examine the model by open sourcing its code can lead to improvements in the quality of the model, both due to *ex ante* and *ex post* factors. The former comprise the strong incentives for a central bank that commits to publish model source code to invest resources before the publication to ensure that the code is of high quality and the model itself can withstand more detailed scrutiny. At the very least, the model's shortcomings would be communicated more explicitly, which is also a desirable outcome. As for the latter, model improvement can also occur once its code is published. As argued by Johnson (2006), the quality of an open source project can increase compared to a similar closed source initiative because proper peer review of codes are more prevalent in open codebases, and also because individuals share more freely ideas for improvement, given they will not necessarily be the ones to implement them. Del Negro et al (2015) make a similar point when justifying the rationale for open sourcing their Julia-based DSGE code: it is a form of validating, and improving, model code beyond in-house run tests of code accuracy. More broadly, once a sufficiently numerous pool of stakeholders becomes familiar with its inner workings, contributions by third parties in the form of suggested code modifications, testing routines (to make sure future versions of the model run as expected or "break" visibly), or improved documentation start to gain momentum. This is likely to occur especially if early contributions are constructively dealt with by central banks, either by accepting them (with or without requested revisions) or by clearly even if briefly explaining why the suggestion in question is not taken on board. [10]

But attracting active model contributors might be more challenging, at least at first while a culture of accepting third-party contributions is not widespread. An important obstacle is in the numbers: the set of potential contributors is a niche group of macroeconomists with sufficient knowledge of central bank models and of programming to know where these models can be improved. Given there is not much code in the open or easily accessible even if strictly speaking it is published, many macroeconomists might not

[10] The open source literature also identifies an incentive-related argument. In settings where economic agents compete with each other, the social benefits of open innovation could be associated with free-riding, reducing incentives for the innovator (in this case, the central bank), as highlighted by Schmidt and Schnitzer (2003) and Saint-Paul (2003). But in the case at hand, central banks will continue to use macroeconomic models as key ingredients of monetary policy, regardless of whether the source code is closed or open. Hence, the incentives for continuous improvement of models remain even when it is shared. In fact, the interests of the central bank and of society at large in the achievement of central bank mandates, such as control of inflation, are (or should be) aligned, so opening model code should promote good faith contributions.

be familiar with the practice of central bank modelling (even if they are with the theory). On top of that, central banks lack the track record to accept specific inputs from third parties that do not flow directly from another central bank, international financial institution such as the IMF or the BIS, or consulting economists that are hired for that purpose. In the spirit of Lerner and Tirole (2005b), those limitations are offset by a potentially strong set of private incentives for potential contributors to provide high quality input, if they had the chance. Contributions are individually named and fully transparent, and are therefore seen and assessed by peer macroeconomists. The central bank responsible for the model code needs to approve each contribution, which provides a "seal of approval" to that particular piece of code. And in addition to the central bank in question, private contributions would be potentially seen by other central banks as well as by other interested stakeholders such as private sector economists (eg from private forecasters, bank economic research departments) or from academia, reinforcing the signal about that contributor's ability.

Another advantage from public model disclosure is that stakeholders, such as private sector economists, can have a more concrete understanding of how central banks use models to view the economy and monetary policy transmission mechanisms. This can improve the way that the private sector interprets and receives policy announcements, especially due to the ability to adapt and run their own analyses based on central bank models. Opening model codes can thus be a practical way to help central banks reap some of the expected benefits from disclosure of policy function or predicted policy paths (eg, Rudebusch and Williams (2008)) in a less likely way to be wrongly interpreted as a policy commitment. In particular, sharing model code can contribute de-mystify central bank models as all-knowing black boxes that lead policymaking, and instead promote the more realistic image of mathematical simplifications that offer insights that feed into a larger policy debate.

5. Challenges related to model code disclosure

Naturally, open sourcing model code can also have challenges and even risks. I classify them into four main types, along the economic, policy, operational and reputational perspectives.[11]

One plausible risk as an open source macroeconomic model becomes widely used by forecasters and other private sector economists is similar to the phenomenon modelled by Morris and Shin (2002): the (now publicly available) forecasts or interpretations of policy transmission channels resulting from a model disclosure may be welfare decreasing depending on how noisy they are compared to private information. More specifically, forecasts resulting from open source central bank models could lead private agents to rationally overreact to this new information, even in situations where these agents had more precise information about the economy. This overreaction would stem from the coordinating effect arising from the unique role of central banks, and to paraphrase Morris and Shin, could be even more detrimental if the forecast from the open source model was faulty in the first place. Ideally this can be addressed by appropriate clarity by the central bank on the fact that model forecasts and estimated transmission channels are usually a *starting point*, not a direct ingredient in policy decisions. In other words, central bankers do not mechanistically follow model results. A related coordination risk occurs if the central bank models become the standard way the private sector sees the economy – while this may well occur given a facilitated implementation of the official models, having the code at hand would also enable private agents to more easily experiment with different alternatives for calibration or even model architectures, thus promoting more variation in private forecasts.

[11] Beyond the potential drawbacks of code disclosure, one practical challenge could be related to contractual or other restrictions central banks might face in sharing data that would feed into these models. However, this can be overcome with synthetic data.

From a policy dimension, another risk of open sourcing model code is that it might undesirably reveal insights into central bank thinking that could interfere with the ability to conduct monetary policy. Of course, model outputs are important ingredients of policy decisions, but definitely not the only ones. These macroeconomic models are tools to help policymakers put together how they see the economy working from its various components. Still, in a way, this is related to the Lucas critique (1976): macroeconomic models often make assumptions about market reactions, which could themselves be shaped by model disclosures. However, this does not seem to have been the case in practice, even for open-source models of widely followed economies such as the US.

Operational challenges include the preparation of the code for disclosure in a way that is consistent with central banks' strong reputation in economic analysis, and ways to properly deal with feedback. Code disclosure requires central banks to invest upfront in ensuring that the code is complete, working, properly structured and ideally well documented. Granted, this investment would be ideal even for internal code, but the need to maintain a reputation for high quality analyses could reflect on an added investment if the code is made public.

Another potential concern is that publicly available model code could open up the model for criticism, potentially harming a central bank's technical reputation. While in the end modelling efforts would always benefit from critical input, especially from informed third-parties that were not involved in the original modelling effort, the tolerance for criticism might be different between central banks and over time. This is also related to how central banks react to spontaneous feedback received on the model code. As usual in open source projects of relevance, third-party developers occasionally suggest code improvements, point out bugs or simply ask technical questions related to reproducing the model. The need to respond can impose a burden for many central banks, especially if its jurisdiction counts with specific requirements in analogy with "freedom of information" requests. A remedy to this is to ensure the model is well documented *ex ante*, thus precluding any need for generic responses, while also engaging with technical suggestions in a way that balances staff availability at the central bank with the benefits from leveraging crowd-sourced talent.

6. Criteria

I use two groups of criteria in assessing central bank open source models. The first set of criteria are related to the transparency objective of sharing code: allowing others to see how the model is implemented, understand its various features and inputs, and be able to reproduce it. The second group of criteria go beyond just the transparency objective, helping build and promote a community of users of the code: practices around software choice and versioning, how easy it is to contribute with issues or code suggestions, and whether third party users have clarity on the permitted uses via an explicit license. The following subsections elaborate these criteria.

In this section, each criterion is written in bold, with specific questions to help judge compliance in italics. A normal-faced text follows with more detail, eg its rationale or how it relates to other criteria.

6.1 Core criteria for code transparency

The most important criteria to promote code transparency are related to open access (how easy it is to find and access the code), documentation, and how easy it is to reproduce the model.

Open access

- **Easy to find**: *is the code available in a location on the central bank's website that is easy for interested users to find, such as a dedicated repository or as accompanying material to the paper describing the model?* Learning about the existence of model source code is of primary importance for sharing. Because there isn't yet a culture of sharing code, it is reasonable to assume many potentially interest users of model code might not know it is available, let alone how to find it, unless it is easy to find.
- **Free access**: *can the source code be freely accessed and downloaded by anyone without the need to register with the central bank or ask for its permission?* Even if some aspects of code are not free to run (eg, use of commercial software), access to the code file itself may be free. Still, some central banks might conceivably want to restrict access to the model code in order to have a tighter control on its usage. Even in these cases, the central bank might find it more convenient and a better cost-benefit mix to post the code for free access while fine-tuning its license, as discussed below.

Documentation

- **Academic description**: *is an academic document (either published in a journal or as working paper or occasional paper) available with a full description of the model?* A detailed composition of the model is obviously a pre-requisite for proper model transparency. One common and useful way this can be done is by writing a detailed description of the model in academic format.
- **Application programming interface (API) documentation**: *are all user-facing methods, classes, variables and other programming objects adequately documented?* For models that are executed in environments based on graphical user interfaces (GUI), this criteria amounts only to pointing to the different necessary files, as the code execution itself occurs in the GUI.
- **Input data documentation**: *are the expected data formats well documented, such that a new user can know what data goes in, in which format, and whether any data transformations (seasonal adjustment, logarithm, de-meaning, etc) are expected?* Ideally, documentation includes a complete list of all necessary input data, and would inform users about sources of the data used to estimate the original model, preferably official sources such as central bank themselves or statistical agencies, or vendors when needed. Input data documentation needs to be very precise to avoid ambiguity during reproduction. For example, whether a particular series is seasonally adjusted or not. A very helpful practice related to this is including data that supports a minimal use case. Regardless of whether this data would be real or simulated (eg, to avoid breaching data distribution agreements that might fall on vendor data), offering users a minimally reproducible example not only advances documentation of the model itself but helps users to understand how it is run and to test it in case users make changes to the code.
- **Output documentation**: *are the formats of the model results well described, so that the user can form an expectation on the type of output to be expected?* In cases where the code offers reproduction for a published paper, the output is usually a table or graph and thus can be considered to be well-documented. When this is not the case, the model should be clear about what type of output can be expected.

Reproduction and replication

- **End-to-end execution**: *is it possible to run the model end-to-end with the available code and instructions?* This criteria does not require an exact reproduction of central bank *results*, as it may depend on data that is either available only administratively or on vendor data subject to

distribution limitations. Rather, this criteria points to an ability to *execute* the code beginning-to-end with only the distributed code and any other requirement obtainable through clearly documented instructions, such as data or specific hardware.

- **Reproduction or replication instructions, tutorial or vignette available**: *does the model include explicit instrucitons for reproduction of central bank results or replication with other data, from download to final result?* These documents are meant to guide users, including those that have never interacted with this code before. They include a narrative description, coupled with examples, of all steps necessary to execute the code. For central banks, as for other software creators in general, writing vignettes can be a way to improve the code, as it puts developers in the shoes of first-time users, a point made by Wickham and Bryan (2023).
- **Minimum requirements**: *are the minimum hardware and software requirements clearly stated?* Naturally, a full list of requirements is impractical, but at least a major version of the software language used to run the model is required (it is likely the main constraint in any case).

6.2 Additional criteria for community-based benefits

Software

The choice of programming language has implications not only for the performance of the macroeconomic models (Aruoba and Fernández-Villaverde (2015)), but also on the ability of third-party users to understand and execute the code. Of particular importance is the difference between software that is proprietary and accessible only at a cost compared to free software (usually itself open source).[12] Similarly, models that do not run across the most widely used operating systems (Windows, macOS and Linux) create roadblocks for an important subset of users.

- **Software open availability**: *is the programming language itself free and easy to acquire?* The software might not need to be open source, as long as it is accessible to a broad range of users.
- **Technology neutrality**: *can the model be run in a broad variety of operating systems, including Linux, macOS and Windows?* Some software used for econometric modelling may be executable across multiple platforms (especially macOS and Windows) even when the programming language itself is not free. Alternatively, even a free and open source programming language could in practice not be available for all platforms.
- **Testing**: *does the code contains tests that ensure that new code, including any adjustments made by users, can be tested against to ensure it runs?* Given that model codes are often complex in the sense that they take multiple input series, perform non-trivial mathematical operations on these series, and then generate some output, including testing routines in the code would ensure that all these steps occur as appropriate. This is particularly important as data manipulation can suffer from silent errors, ie those situations where the software executes as expected (ie, the code "does not break") but because of formatting or similar issues the data was not in the expected format, leading the code to execute in a completely different way as intended.
- **Explicit code versioning**: *is the version of the code explicitly stated?* Users reproducing or replicating the model need to be certain which version of the model they are working with, which is especially important as macroeconomic tools evolve and get incorporated into newer version of models.

[12] In some cases, like MatLab, there are suitable open source alternatives (Octave, in this case).

- **Availability of past versions**: *can users freely and easily access past versions of the model in the same website?* Access to versioned code helps users be certain of the model version that they are executing. This is especially important in the case of macroeconomic modelling code, since a number of central banks tend to update their models every few years to take stock of lessons learned from previous years, and potentially of advances in modelling.

Contribution

- **Contact with software maintaners**: *is some form of contacting a person or team responsible for the model software clearly available to all, eg in email format or in a discussion board?* While opening up a contact channel to the broad public can seem daunting for some central banks, it is one important way to ensure contributions come in without the central bank actively exerting effort. Naturally, this form of contact does not need to be an e-mail mailbox – which almost calls for a reply – but a forum or even a commenting field similar to how pictures are shared in various social networks. The point of this criterion is to encourage central banks to offer some effective way for the broader public to provide their written feedback to model software maintainers.
- **Third party contributions possible**: *can third party users independently propose changes to the code and documentation?* An evolution of the communication possibility is to allow third party users to literally propose improvements to the code or to the documentation.
- **Contribution guidelines**: *does the model have explicit guidelines for contributors, or a contributor agreement?* Of course, this is more valid for the codes that are shared in a way that offers opportunities for collaboration. In those cases, it is important to offer consistency between actions and the guidelines, as they often diverge (Elazhary et al (2019)).

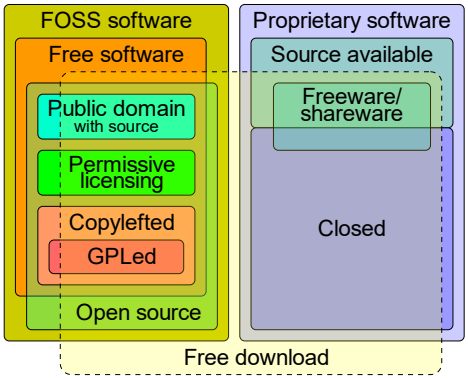
License

- **Explicit open source license**: *is the license explicit and is it compatible with the open source definitions?* Over the years, several standardised open source licenses were created with the purpose of setting different terms that are considered open source. These standard licenses are widely known, usually offer legal teams comfort around their meaning and implications in different circumstances, and thus can more easily be used off-the-shelf.[13] Still, some specific circumstances might call for a customised license. While they might not be as easily understood as the standard licenses, and thus should not be preferred as they impose a cost on each user to check the conditions, custom-made licenses can offer central banks reasonable flexibility in how they open source model code. In any case, it is important to understand different types of software and source code availability that licenses imply. Graph 2 overviews license concepts.

[13] One widely used catalogue of standard licenses is the Linux Foundation's SPDX, available at https://spdx.org/licenses/.

Software license concepts according to the Free Software Foundation

Graph 2



[1] FOSS: free and open source software. GPLed: software under the GNU General Public License (GPL).

Source: Wikipedia.

7. Current central bank practices

A best-efforts stock-take of central banks with open-sourced model code is presented in Table 2. The models listed were sourced mainly from the author's own familiarity with some of these codes; crowd-sourced contributions in a public repository, and from searching the most recent macroeconomic models by the central banks that already share other types of code, listed in Table 1.

Some observations are in order. First, not many countries populate this list: Table 2 is short compared to the population of central banks. This was not a comprehensive stock take, and there could be other central banks that do share code but were hard to find. Still, the models of some central banks that share other, non-model code (Table 1) are not open source. For example, the Central Bank of Brazil's SAMBA model (Fasolo et al (2023)) or the National Bank of Belgium's BEMGIE model (de Walque et al (2023)) have been recently overhauled but their codes are not publicly available. The Danish central bank's MONA model (2003) and the Bank of Norway's NEMO model (2019) also do not seem to have publicly available code. These models and others are well-documented, so the absence of public code might represent either an explicit decision to retain code confidentiality or simply an inertia from past practices. Second, once a country's central bank does decide to open source model code, then it is common for more than one model to be publicly shared. Case in point is the United States, perhaps the first country with a central bank to publish model code approximately one decade ago, has no less than four open source models. Together, these observations point to a possible effect at the jurisdiction level on the decision to share model code: perhaps culture, legal framework, or a combination of these and other factors prevent central banks from benefitting from opening their macroeconomic model code.

Other works in this space that are worth mentioning are the Bank of Canada's StateSpaceEcon.jl Julia package (St-Pierre (2023)) and the Bank of Italy's Python-based Black-it package (Benedetti et al (2022)). They are not macroeconomic models themselves, but well-documented open source libraries for coding such models.[14]

[14] Codes available respectively at github.com/bankofcanada/StateSpaceEcon.jl and github.com/bancaditalia/black-it.

Open-sourced central bank models used in the assessment

Best-efforts list of official models with available code

Table 3

Jurisdiction	Institution	Model name	Model ID[1]	Reference
Australia	Reserve Bank of Australia	MARTIN	AU-01	Ballantyne et al (2019)
Chile	Central Bank of Chile	MSEP	CL-01	Arroyo Marioli et al (2020), BCC (2020)
Chile	Central Bank of Chile	XMAS	CL-02	García et al (2019), García and Guerra-Salas (2020), BCC (2020)
Denmark	Ministry of Finance[2]	MAKRO	DK-01	Bonde et al (2023)
European Union	European Central Bank	BEAR	EU-01	Dieppe and van Roye (2023)
European Union	European Commission[2]	Output Gap Model	EU-02	
Finland	Bank of Finland	Aino	FI-01	Kilponen et al (2016)
France	Ministry of Economics and Finance[2]	Mésange	FR-01	Dufernez et al (2017)
France	Ministry of Economics and Finance[2]	Opale	FR-02	Daubaire et al (2017)
France	Ministry of Economics and Finance[2]	Modèle Saphir	FR-03	Amoureux et al (2018)
Germany	Federal Ministry for Economic Affairs and Climate Action[2]	-	DE-01	BMWK and BMFin (2022)
Iceland	Central Bank of Iceland	QMM	IS-01	Daníelson et al (2019)
Japan	Bank of Japan	Q-JEM	JP-01	Hirakata et al (2019)
Sweden	Riksbank	Ramses II	SE-01	Adolfson et al (2013)
United Kingdom	HM Treasury and Office for Budget Responsibility[2]	OBR model	GB-01	Office for Budget Responsibility (2013)
United States	Federal Reserve Bank of New York	DSGE.jl	US-01	
United States	Federal Reserve Bank of New York	Nowcasting	US-02	Bok et al (2017)
United States	Federal Reserve Board	FRB/US	US-03	Brayton et al (2014), Laforte (2018)
United States	Federal Reserve Board	EDO	US-04	Chung et al (2010)

Information as of 22 September 2023. [1] Model ID is used throughout the text to refer to each model. [2] Other official financial agencies such as Ministries of Finance are also in scope of this compilation.

Source: Author compilation, authorities' websites.

8. Assessment of open-sourced model code

I assessed each code against the criteria identified above (Table 3). When in doubt, I took a positive view of compliance with the criteria – for example, a simple documentation of the usage interface can be taken

as documentation of the API, even if rudimentary. This is because the purpose of this assessment is to take a broad view of the main practices in official macroeconomic model code disclosures. Models judged to fulfil the criteria are therefore also encouraged to continuously improve.

Characteristics of open-sourced macroeconomic model codes

Best efforts list of official models with available code

Table 4

	Core criteria									Additional criteria									
	Open access		Documentation				Reprod./Replic.			Software					Contribution			Lic.
Model ID[1]	Easy to find	Free access	Academic	API	Input data	Output data	End-to-end	Instructions	Requirements	Open software	Tech neutrality	Testing	Explicit version	Past versions	Contact	Third parties	Guidelines	Explicit license
AU-01	X	X	X	X		X	X	X	X									X
CL-01	X	X	X	X	X	X	X	X	X	X	X				X			
CL-02	X	X	X	X		X	X	X	X	X	X				X			
DE-01		X	X	X	X		X			X			X	X				
DK-01	X	X	X	X	X		X		X		X		X	X	X	X		X
EU-01	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
EU-02	X	X		X	X		X		X				X	X				
FI-01	X	X	X	X	X	X	X		X	X	X							
FR-01	X	X	X	X	X	X	X	X	X				X	X	X	X		X
FR-02	X	X	X	X	X	X	X	X	X				X	X	X	X		X
FR-03	X	X				X	X	X	X				X	X	X	X		X
IS-01	X	X	X	X	X	X	X	X										X
JP-01	X	X	X	X	X	X	X	X	X						X			X
SE-01	X	X	X	X			X	X		X	X				X	X		
GB-01	X	X	X	X	X		X		X									
US-01	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X
US-02	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X		X
US-03	X	X	X	X	X		X	X	X	X	X							
US-04	X	X	X	X			X			X	X							X

Information as of 22 September 2023.[1] See Table 2.

Source: Author compilation, authorities' websites.

8.1 Core criteria

Open access

The vast majority of model code can be classified as having open access. Having the code publicly available is used in this paper as a definition of open source, so by design all models comply with this criterium. In terms of the ease of access, DE-01 could be better.

Documentation

In some cases, models list the **input data** used to estimate the model including their sources. Other cases document input data, but dispersed in the academic documentation of the model instead of a single location that is easy to consult. In a few cases, authors embed the data in the distributed model. One helpful practice, observed for example in AU-01, is to distribute simulated data when the original series are under commercial protection, specifying which ones are simulated.

Reproduction and replication

Open source models are already being used in academia by third-party authors as a way to obtain credible counterfactuals for macroeconomic variables. Gross and Leigh (2022) use AU-01, re-estimated up to 2019 and adapted to use only publicly available data, to estimate counterfactuals for nominal cash rate, (trimmed-mean) inflation, real GDP, unemployment gap, wages and exchange rate. Blanchard and Summers (2020) use US-03 to simulate the US economy under a neutral fiscal policy.

8.2 Additional criteria

One general remark about the additional criteria is that model codes stored in dedicated repository websites already benefit from many useful characteristics in terms of open source code. In addition to making the code easy to find and access (core criteria), it automatically includes explicit versioning and availability of past. Moreover, such repositories also provide a useful platform for contact with code maintainers and other code users, and allow for suggestions from third parties to improve code and documentation. These code repositories also facilitate a clear definition of a use license and implementing testing routings (in the form of "continuous integration", a technique that ensures that every new version of the code must be able to execute some testing or production-like routines).

Software

The software choices across models are heterogenous, but with a clear tendency to gravitate to Matlab (which has an open source alternative, GNU Octave) and EViews. In terms of **software open availability**, many models run on software that is freely available. But in some cases central banks use commercial software that is not freely available, such as GAMS (General Algebraic Modelling System)[15] and EViews. Table 4 presents an overview of the software languages used by the models studied in this paper.

Most common programming languages for macroeconomic models

Best efforts list of official models with available code

Table 5

Languages	Open source	Multi-platform	Observations	Models
Matlab (or GNU Octave)	No (Yes)	Yes	Most models with Dynare	CL-01, CL-02, EU-01, SE-01 US-02, US-04

[15] Interestingly, and laudably, GAMS Software GmbH has an explicit and public open source strategy that acknowledges the value of open source to the firm's software, and also mentions that the company seeks to contribute code to open source software.

EViews	No	No		AU-01, GB-01, JP-01[2], US-03[2]
Portable TROLL	No	Yes (raw code; GUI[3] only available for Windows)		FR-01, FR-02
Julia	Yes	Yes		US-01[1]
SAS	No	Yes		FR-03
EUCAM	No	No		EU-02
GAP	No	No		DE-01
GAMS	No	No	APIs[3] available for Python, Java and other languages	DK-01

Information as of 22 September 2023. [1] Based on previous Matlab code. [2] Also has a Python version. [3] GUI: graphical user interface, API: application programming interface.

Source: Author compilation, authorities' websites.

Dynare (Juillard (1996)) appears to be the reason why most models use Matlab. It is a time-tested, well-documented open source software that estimates a wide range of models used by central banks. A whole community of macroeconomists, many of them active in central banks, participate in its forums and in the events organised by the Dynare community. There is even a working paper series dedicated to models developed with this software.

Two purpose-specific languages feature in the studied sample: Portable TROLL[16] and GAP. The former is a C-based software that is reportedly executable across multiple platforms, with a graphical user interface (GUI) in Windows. It is developed and marketed by Intex Solutions, Inc. Some European models (DE-01, EU-02) run on GAP, a programme developed by the European Commission based on the EU output gap methodology. While GAP itself is publicly accessible, it requires Microsoft Excel and MatLab, neither of which are free. And although both have free and open source alternatives, such as LibreOffice for the former and Octave for the latter, it is not clear whether GAP would execute on them.

Using GAP as an example, **technology neutrality** is not really present in all models. For example, the GAP model works only on Windows (since it is assumed or known to be the most widely professionally used OS by European agencies). Similarly, model codes that rely on EViews are only runnable on Windows. On the other side, many models are based on Julia, which is runnable on multiple OSs.

In general, **testing** is not included in the source codes. While model codes are not made for autonomous execution, instead serving the purpose of directly interacting with analysts and economists, the lack of embedded testing routines (such as unit testing) can be problematic because it does not include a key way to check whether code changes or improvements maintain an appropriately executing file. In other words, the model code is not guaranteed to work. As mentioned above, it is possible to set up testing in practical ways when the code is hosted in a repository website. For example, EU-01 and US-02 are tested automatically at every new code contribution that is pushed to the central repository.

For many open source models, **explicit code versioning** and **availability of past versions** (if any) is lacking. While such models are not expected to be updated at a high frequency, over time as third-party use of models increase, explicitly adequately version can be important. One interesting case is the European Central Bank's PERFORM system, which includes a git repository, a suite of software to run models, and schedulers that execute code testing and quality control (see Box 20 of Work Stream on Eurosystem Modelling (2021)).

[16] TROLL stands for Time-shared Reactive OnLine Laboratory.

Contribution

Central banks currently do not score well on **contribution guidelines**, and in fact, on anything that might harness one of the key advantages of having source code accessible by all: the power of contributions.

Unless the academic paper authors are considered to be the code maintainers, which is not always the case, there is often no way to contact the people responsible for model code. This is of course made worse by the fact that in many cases the model white paper is not even identifiably authored. In any case, even with a relevant contact, third parties willing to spend their time reporting bugs, suggesting code improvements, or simply asking questions about the code are often left without a clear, official way to contact maintainers.

Naturally, opening up such contact risks creating a wrong impression about availability of resources to follow up on these forms of contact mentioned above. But it does not have to be this way. The central bank might be upfront about third parties not expecting a response. Or alternatively, the central bank might either set up, or use an existing forum-like system where any third parties can post their comments. Many code repositories make such a forum available by default. Another advantage of these fora for contact between central bank staff code maintainers and third parties interested in the code is that all communication is public, documented, and identified.

License

Model **licenses** are an important area where some models can further develop. There are many standard licenses developed over time by the software community, but the software creator might wish to create a custom license. While describing details of the main open source licenses is outside the scope of this paper, one important message is that absence of a license does not mean that third parties can freely use it. On the contrary, lack of licenses formally precludes broader use of open source software because other people are not granted the permission to copy and adapt the code, for example. In some cases, a "disclaimer" instead of a license is available. When this disclaimer included at least some general notices about permitted usage or about any potential liabilities from the model, it was considered to be similar to a license.

The selection of a licensing model is a critical decision that can significantly impact how widely used the macroeconomic model is. The standard, 'off-the-shelf' licenses mentioned above offer a distinct advantage in this context, as they have undergone extensive scrutiny and acceptance within the software community. This pre-validated status of such licenses mitigates the legal and operational uncertainties typically associated with custom licensing agreements. While concerns about commercialising copies of macroeconomic models may be less pronounced in the context this paper explores, the strategic choice of a well-established open source license can still play a pivotal role. It can ensure legal clarity and broad acceptability, fostering a collaborative and transparent development environment that might attract third-party talent even from outside the central bank jurisdictions' borders. Correspondingly, Lerner and Tirole (2005a) show that the decision about the license both depends on, and influences, the community expected to be developed around the model code.

9. Enhancing availability of code

For those central banks that wish to put existing code in the public domain or to improve their current levels of accessibility of the code they make public, below are practical, cost-mindful suggestions:

- **Set up a code repository**. It could be self-hosted or, more easily, hosted in specialised servers such as GitHub, GitLab, etc. Another advantage is that they facilitate other auxiliary characteristics

of a good software repository, such as explicitly choosing a license and facilitating contribution by creation of issues or even code suggestions.

- **Create a page listing available code**. A central page in the authority's website could concentrate its available model code (and potentially other source code as well, such as for replicating papers). This would be easy to set up and relatively inexpensive to maintain, given that publication of new source code is not a frequent event. It can also help "cross-polinate" ideas, by helping users that are interested in one code find other code made available by the authority.
- **Create and share tutorials on how to replicate and adapt the models**. Ideally these tutorials would include steps from scratch, such as setting up the environment, accessing the necessary data, running the model, and when relevant changing parameters to those desired by the user.
- **Consider adapting code requiring proprietary software to open source solutions**. This would promote higher third party engagement, facilitate sharing with other central banks and external institutions, and even would facilitate new staff being onboarded to projects. Moving the code from one language to another (known as "transpiling") would be mostly an one-off investment. It could be partially or totally offset by the cost of giving up paid licenses, and could be simplified through the use of language artificial intelligence (AI) models, especially those focused in coding.

10. Conclusion

More than two decades ago, Blinder et al (2001) argued that central banks should better disclose how their models worked. Specifically, they suggested "well-chosen words supplemented by a few key numbers" would be more informative then "masses of equations" because the latter are far from transparent. This paper makes the case that model *code* can be even better, lists some central banks that are already doing this, analyses their practices and suggests specific actions to better harness this practice.

If model codes are publicly disclosed, comparisons of model responses such as Fisher et al (1988) and Smets (1995) can be more easily accomplished. This could enrich the literature on comparative analysis of monetary policy: for example, if replications of the models of two central banks respond the same way to similar shocks, then diverging rate decisions between them might suggest different monetary policy stances. At the same time, other views suggest that transparency might at some point be excessive if it interferes with policy effectiveness – along these lines, source code should be viewed, similar to broader policy transparency, as a means to an end (Mishkin, 2004). In any case, episodes like the 2007-09 Global Financial Crisis and more recently the post-Covid inflation forecast errors have prompted macroeconomics into collective soul-searching (Caballero (2010), Stiglitz (2011), Chahad et al (2022), Borio et al (2023), Koch and Noudeldin (2023)). Having a range of model codes in the open could facilitate diagnosing what is needed to prevent future errors, including by backtesting whether improved models would have been helpful ex ante.

Since macroeconomic models embody the academic work of a public institution, open sourcing of model code can also be seen as part of broader transparency and accountability movements in academia and in administration. The first one is "open science": the practice of sharing more openly code, data and even research plans. Open science acceptance and practice in economics has grown substantially over the last decade, following other social sciences (Ferguson et al, 2023). And the other is "open government", the idea that administration data should be made open and accessible to the extent possible (Attard et al, 2015). Part of the prominence of both of these "open" movements is due to confidence crises in the scientific and political worlds. In contrast, open sourcing models can help shield central banks from losses in their analytical capabilities. In addition, it can be seen as a way of transferring back to society

technologies funded with public money, as long as such disclosure does not interfere with the effectiveness of monetary policy.

This work is also related to the influence exercised by central bank forecasts on private agents (eg, Ehrmann et al (2012), Hubert (2015) and Pedersen (2015)). If the publication of central bank forecasts directly affects private agents regardless of performance, then it can be argued that sharing the code can lead to two effects: first, it would make the effect more diffuse since agents would be able to form their own views using the same model as the central bank. And second, whenever forecasts are indeed published, agents would be able to benchmark that against the hard output of the model, thus inferring the "human adjustment" component of the forecast. Whether this is beneficial or not is a matter to be explored in future research.

Public code sharing can be considered the next frontier in central bank communication, building on a multi-decade evolution (Blinder et al (2008)). Sharing model code provides exact, customisable implementations, which offers central banks the opportunity to be better listeners (Macklem and Vardy, (2023)) by standardising the language around its open sourced models and also observing how private sector users adapt these models. Model code also facilitates central banks' outreach to a broader audience, a more technical one, even if not the general public (Haldane and McMahon (2018)). As central banks consider what type of forecasts to provide to the public, with arguments in favour of publishing a range of forecasts based on different scenarios (Goodhart, 2023), open sourcing model code could allow users the possibility of running the models and testing their own scenarios. Model code sharing could also perhaps be seen as contributing to one of the pillars of central bank transparency as codified by the IMF (2020, 2023), namely the second pillar "*Transparency in policies*" through its sub-item "*policy decisions*".[17]

Central banks have been increasingly active open source code providers, and have been relying on open source tools for many of their most advanced data analytical tasks, such as machine learning (Araujo et al (2023a, 2023c)). Open sourcing macroeconomic model code in a way that maximises its benefits for central banks themselves and for societies would be another significant step in this direction. The diversity in model code disclosure practices suggests that there is still ample need for research to understand the drivers, implications and strategies on open sourcing macroeconomic model code.

References

Adolfson, M, S Laséen, L Christiano, M Trabandt and K Walentin (2013): "*Ramses II – model description*", Riksbank Occasional Paper Series No 12.

Amoureux, V, I Benoteau and A Naouas (2018): "*Le modèle de microsimulation Saphir*", DG Trésor Working Paper No 2018/06, September.

Araujo, DKG (2023): "*gingado: a machine learning library focused on economics and finance*", BIS Working paper No 1122.f

Araujo, DKG, G Bruno, J Marcucci, R Schmidt and B Tissot (2023a): "*Machine learning applications in central banking*", Journal of AI, Robotics & Workplace Automation, v.2, n.3, p.271-293.

Araujo, DKG, G Bruno, J Marcucci, R Schmidt and B Tissot (2023b): "*Data science in central banking: applications and tools*", IFC Bulletin No 59.

[17] In more detail, this would be related to subprinciple 2.1.3, ie transparency about supporting analysis for monetary policy.

Araujo, DKG, S Nikoloutsos, R Schmidt and O Sirello (2023c): "*Central banks as users and providers of open-source software*", Box in IFC Bulletin No 59.

Aruoba, SB, J Fernández-Villaverde (2015): "*A comparison of programming languages in macroeconomics*", Journal of Economic Dynamics and Control, v.58, September, p.265-273.

Arroyo Marioli, F, F Bullano, J Fornero and R Zúñiga (2020): "*Semi-structural forecasting model*", Banco Central de Chile Working Paper No 866, January.

Attard J, F Orlandi, S Scerri and S Auer (2015): "*A systematic review of open government data initiatives*", Government Information Quarterly, v.32, n.4, p.399-418.

Ballantyne A, T Cusbert, R Evans, R Guttmann, J Hambur, A Hamilton, E Kendall, R McCririck, G Nodari and D Rees (2019): "*MARTIN has its place: a macroeconometric model of the Australian economy*", Reserve Bank of Australia Research Discussion paper, RDP 2019-07.

Benedetti, M, G Catapano, F De Sclavis, M Favorito, A Glielmo, D Magnanimi and A Muci (2022): "*Black-it: a ready-to-use and easy-to-extend calibration kit for agent-based models*", Journal of Open Source Software, v.7, n.79, p.4622.

Bank of Norway (2019): "*A complete documentation of Norges Bank's policy model NEMO*", 22 March.

Blanchard, OJ (2018): "*On the future of macroeconomic models*", Oxford Review of Economic Policy, v.34, n.1-2, p.43-54.

Blanchard, OJ and LH Summers (2020): "*Automatic stabilizers in a low-rate environment*", American Economic Association Papers and Proceedings, v.110, p.125-130, May.

Blinder, AS (2000): "*Central-bank credibility: why do we care? How do we build it?*", American Economic Review, v.90, n.5, p.1421-1431.

Blinder, AS, C Goodhart, P Hildebrand, D Lipton and C Wyplosz (2001): "*How do central banks talk?*", Geneva Report on the World Economy 3, Centre for Economic Policy Research, London.

Blinder, AS, M Ehrmann, M Fratzscher, J De Haan and D-J Jansen (2008): "*Central bank communication and monetary policy: a survey of theory and evidence*", Journal of Economic Literature, v.46, n.4, p.910-945.

Bok, D, D Caratelli, D Giannone, AM Sbordone and A Tambalotti (2017): "*Macroeconomic nowcasting and forecasting with big data*"

Bonde M, J Ejarque, G Høegh, E Partsch, P Stephensen, T Vasi (2023): "*MAKRO model documentation*", paper available at www.dreamgruppen.dk.

Borio, C, B Hofmann and E Zakrajšek (2023). "*Does money growth help explain the recent inflation surge?*" BIS Bulletin no 67, Bank for International Settlements.

Brayton, F, T Laubach and D Reifschneider (2014): "*The FRB/US model: a tool for macroeconomic policy analysis*", FEDS Notes, April.

Bundesministerium für Wirtschaft und Klimaschutz – BMWK and Bundesministerium der Finanzen – BMFin, Germany (2022): "*Ergänzende Unterlagen zur Bestimmung des gesamtwirtschaftlichen Produktionspotenzials*", 26 January.

Caballero, RJ (2010): "*Macroeconomics after the crisis: time to deal with the pretense-of-knowledge syndrome*", Journal of Economic Perspectives, v.24, n.4, p.85-102, Fall.

Central Bank of Chile – BCC (2020): "*Uso de modelos macroeconómicos en el Banco Central de Chile*", paper available at bcentral.cl.

Cesa-Bianchi, A (2024): "*Replications*", available at: https://sites.google.com/site/ambropo/replications. Access on 11 March 2024.

Chahad, M, A Hofmann-Drahonsky, B Meunier, A Page and M Tirpák (2022): "*What explains recent errors in the inflation projections of Eurosystem and ECB staff?*" ECB Economic Bulletin no 3/2022, European Central Bank.

Chung, HT, MT Kiley, J-P Laforte (2010): "Documentation of the Estimated, Dynamic, Optimization-based (EDO) model of the U.S. economy: 2010 version", Federal Reserve Board Finance and Economics Discussion Series (FEDS) No 2010-29.

Costa, R (2022): "*Heterogeneity in macroeconomics – implications for monetary policy*", Remarks at XXV Annual Conference of the Central Bank of Chile, Santiago, 21-22 November.

Danmarks Nationalbank (2003): "*MONA – a quarterly model of the Danish economy*".

Daubaire, A, G Lefebvre and O Meslin (2017): "*La maquette de prévision Opale2017*", DG Trésor Working Paper No 2017/06, May.

Daníelson, Á, L Elíasson, MF Gudmundsson, SJ Haraldsdóttir, LS Kro, ThG Pétursson, ThS Sveisson (2019): "*QMM – a quarterly macroeconomic model of the Icelandic economy Version 4.0*", Central Bank of Iceland Working Paper No 82, December.

De Nederlandsche Bank (DNB) (2023): "*GitHub pages of institutions*", available at: https://github.com/DeNederlandscheBank/general-information/blob/main/github_pages_institutions.md

De Walque, G, T Lejeune, A Rannenberg and R Wouters (2023): "*BEMGIE: Belgian Economy in a Macro General and International Equilibrium Model*", National Bank of Belgium Working Paper Research, no 435, March.

Del Negro, M, M Giannoni, P Li, E Moszkowski and M Smith (2015): "*The FRBNY DSGE model meets Julia*", Liberty Street Economics, 3 December.

Dieppe, A and B van Roye (2023): "*The BEAR toolbox*", website available at: www.ecb.europa.eu/pub/research/working-papers/html/bear-toolbox.en.html. Accessed on 22 Sept 2023.

Dincer, N and B Eichengreen (2014): "*Central bank transparency and independence: updates and new measures*", International Journal of Central Banking, v.10, n.1, p.189-253.

Dou, WW, AW Lo, A Muley and H Uhlig (2020): "*Macroeconomic models for monetary policy: a critical review from a financial perspective*", Annual Review of Financial Economics, v.12, p.95-140, November.

Duarte, A, J Frost, L Gambacorta, P Koo Wilkens and HS Shin (2022): "Central banks, the monetary system and public payment infrastructures: lessons from Brazil's Pix", BIS Bulletin No 52, 23 March.

Ehrmann M, SCW Eijffinger and M Fratzscher (2012): "*The role of central bank transparency for guiding private sector forecasts*", The Scandinavian Journal of Economics, v.114, n.3, p.1018-1052.

Eijffinger, SCW and PM Geraats (2006): "*How transparent are central banks?*", European Journal of Political Economy, v.22, p.1-21.

Elazhary, O and M-A Storey, N Ernst, A Zaidman (2019): "*Do as I do, not as I say: do contribution guidelines match the GitHub contribution process?*", arXiv, mimeo.

Fasolo, AM, E Araujo, M Valli Jorge, A Kornelius and LSG Marinho (2023): "*Brazilian macroeconomic dynamics redux: shocks, frictions, and unemployment in SAMBA model*", Latin American Journal of Central Banking, in press.

Ferguson, J, R Littman, G Christensen, EL Paluck, N Swanson, Z Wang, D Birke and J-H Pezzuto (2023): "*Survey of open science practices and attitudes in the social sciences*", Nature Communications, v.14, n.5401.

Fisher, PG, SK Tanna, DS Turner, KF Wallis and JD Whitley (1988): "Comparative properties of models of the UK economy", National Institute Economic Review, v.125, p.69-87, August.

Fracasso, A, H Genberg and C Wyplosz (2003): "*How do central banks write? An evaluation of inflation reports by inflation targeting central banks*", Geneva: International Center for Monetary and Banking Studies; London: Centre for Economic Policy Research.

García, B, S Guarda, M Kirchner, R Tranamil (2019): "*XMas: an extended model for analysis and simulation*", Banco Central de Chile Working Paper No 833, May.

García, B, J Guerra-Salas (2020): "*On the response of inflation and monetary policy to an immigration shock*", Banco Central de Chile Working Paper No 872, April.

Goodhart, C (2023): "Should central banks abandon single point forecasts?", VoxEu, 29 August.

Gross, I and A Leigh (2022): "*Assessing Australian monetary policy in the twenty-first century*", Economic Record, v.98, n.322, p.271-295, September.

Haldane, A and M McMahon (2018): "*Central bank communications and the general public*", AEA Papers and Proceedings, v.108, p.578-583.

Hendry, DF and JNJ Muellbauer (2018): "*The future of macroeconomics: macro theory and models at the Bank of England*", Oxford Review of Economic Policy, v.34, n.1-2, p.287-328

Hirakata, N, K Kan, A Kanafuji, Y Kido, Y Kishaba, T Murakoshi and T Shinohara (2019): "*The Quarterly Japanese Economic Model (Q-JEM): 2019 version*", Bank of Japan Working Paper Series, No 19-E-7, June.

Hubert, P (2015): "*Do central bank forecasts influence private agents? Forecasting performance versus signals*", Journal of Money, Credit and Banking, v.47, n.4, p.771-789.

International Monetary Fund (2020): "*The central bank transparency code*", IMF Policy Paper 2020/038, 4 July.

International Monetary Fund (2023): "*Central bank transparency code review – guidance note*", IMF Policy Paper 2023/050, 31 October 2023.

Iskrev, N (2019): "*What to expect when you're calibrating: measuring the effect of calibration on the estimation of macroeconomic models*", Journal of Economic Dynamics and Control, v.99, p.54-81.

Johnson, JP (2004): "*Open source software: private provision of a public good*", Journal of Economics & Management Strategy, v.11, n.4, p.637-662.

Johnson, JP (2006): "*Collaboration, peer review and open source software*", Information Economics and Policy, v.18, n.4, p.477-497.

Juillard, M (1996): "*Dynare: a program for the resolution and simulation of dynamic models with forward variables through the use of a relaxation algorithm*", CEPREMAP Working Paper No 9602.

Koch, C and D Noureldin (2023): "*How we missed the inflation surge: an anatomy of post-2020 inflation forecast errors*", IMF Working Paper no 2023/102.

Koren, M, M Connolly, J Llull and L Vilhuber (2022): "*Data and Code Availability Standard [Version 1.0]*", Website, https://datacodestandard.org https://doi.org/10.5281/zenodo.7436134.

Laforte, J-P (2018): "*Overview of the changes to the FRB/US model*", FEDS Notes, December.

Lerner, J and J Tirole (2002): "*Some simple economics of open source*", The Journal of Industrial Economics, v.50, n.2, p.197-234.

Lerner, J and J Tirole (2005a): "*The scope of open source licensing*", The Journal of Law, Economics, and Organization, v.21, n.1, p.20-56, April.

Lerner, J and J Tirole (2005b): "*The economics of technology sharing: open source and beyond*", Journal of Economic Perspectives, v.19, n.2, p.99-120, Spring.

Lucas, R (1976): "*Econometric policy evaluation: a critique*", in: Carnegie-Rochester Conference Series on Public Policy, v.1, 19-46.

Macklem, T and J Vardy (2023): "*20 years of central bank communications, and lessons for the future*", in: Central banking in the Americas: Lessons from two decades, BIS Papers No 143.

Morris, S and HS Shin (2002): "*Social value of public information*", American Economic Review, v.92, n.5, p.1521-1534.

Mishkin, FS (2004): "*Can central bank transparency go too far?*", NBER Working Paper No 10829.

Muñoz, S, S Jahjah, M Cihak, ST Sharifuddin, KI Tintchev (2012): "*Financial stability reports: what are they good for?*", IMF Working Paper WP/12/1.

National Academies of Sciences, Engineering, and Medicine (2019): "*Reproducibility and replicability in science*", Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

Office for Budget Responsibility, United Kingdom (2013): "*The macroeconomic model*", Briefing paper No 5, October.

Open Source Initiative (2023): "*Open source definition*", available at: https://opensource.org/osd/

Pedersen, M (2015): "*What affects the predictions of private forecasters? The role of central bank forecasts in Chile*", International Journal of Forecasting, v.31, n.4, p.1043-1055.

Pescatori, A and S Zaman (2011): "*Macroeconomic models, forecasting and policymaking*", Federal Reserve Bank of Cleveland Economic Comentary No 2011-19, October.

Pfeifer, J (2023): "*DSGE_mod: a collection of Dynare models*", computer software, available at: https://github.com/JohannesPfeifer/DSGE_mod

Prante, F and K Kohler (2024): "*Macrosimulation*", available at: https://macrosimulation.org. Accessed on 11 march 2024.

Rudebusch, GD and JC Williams (2008): "*Revealing the secrets of the temple: the value of publishing central bank interest rate projections*", In: Campbell, JY (Ed.), *Asset Prices and Monetary Policy*, p.247-289.

Saint-Paul, G (2003): "*Growth effects of non-proprietary innovation*", Journal of the European Economic Association: Papers and Proceedings, v.1, n.2-3, p.429-439.

Schmidt, KM and M Schnitzer (2003): "*Public subsidies for open source? Some economic policy issues of the software market*", Harvard Journal of Law & Technology, v.16, n.2, p.473-505.

Shah, SK (2006): "*Motivation, governance, and the viability of hybrid forms in open source software development*", Management Science, v.52, n.7, p.1000-1014.

Silvo, A and F Verona (2020): "*The Aino 3.0 model*", Bank of Finland Research Discussion Papers No 9/2020.

Sims, C (2002): "The role of models and probabilities in the monetary policy process", Brookings Papers on Economic Activity, 2002, v.2002, n.2, p.1-40.

Smets, F (1995): "*Central bank macroeconometric models and the monetary policy transmission mechanism*", In: *Financial structure and the monetary policy transmission mechanism*, Central Bank, BIS p.394.

St-Pierre, NL, B Bejanov and J Jensen (2023): "*What we learned building a modelling language for macroeconomics*", JuliaCon 2023, presentation, 28 July.

Stiglitz, JE (2021): "*Rethinking macroeconomics: what failed, and how to repair it*", Journal of the European Economic Association, v.9, n.4, p.591-645, August.

Taylor, JB, V Wieland (2012): "*Surprising comparative properties of monetary models: results from a new model database*", Review of Economics and Statistics, v.94, n.3, p.800-816.

von Krogh, G, S Haefliger, S Spaeth and MW Wallin (2012): "*Carrots and rainbows: motivation and social practice in open source software development*", MIS Quarterly, v.36, n.2, p.649-676.

West, J (2003): "*How open is open enough? Melding proprietary and open source platform strategies*", Research Policy, v.32, n.7, p.1259-1285.

West, J and S Gallagher (2006): "*Challenges of open innovation: the paradox of firm investment in open-source software*", R&D Management, v.36, n.3, p.319-331.

Wickham, H and J Bryan (2023): "*R packages*", 2nd edition, O'Reilly Media.

Work stream on Eurosystem modelling (2021): "*Review of macroeconomic modelling in the Eurosystem: current practices and scope for improvement*", European Central Bank Occasional Paper Series No 267, ECB Strategy Review.



Open-sourced central bank macroeconomic models

Douglas Araujo

3rd IFC and Bank of Italy Workshop on "Data Science in Central Banking"

18 Oct 2023

Disclaimer

The views in this presentation do not necessarily represent the views of the Bank for International Settlements or of the central banks and other authorities mentioned.

All errors are my own.

This is work in progress. Additional or better information about central banks or ministries that share macroeconomic model code is welcome.

Overview

- Macroeconomic models help central banks (and other policymakers such as MinFins) to:
  - understand transmission channels of policy decisions
  - forecast the economy under different scenarios, and
  - inform policy stances in a forward-looking way

- Over time, some central banks open sourced the code of their macroeconomic models
  - interesting development in the decades-long increase in levels of central bank transparency

- This paper:
  - benchmarks how macro models are open sourced, and
  - offers practical, cost-mindful suggestions for central banks considering open sourcing models

- *Evaluating* or *categorising* the models themselves are not in scope

What is open source code?

- Simple definition: **code is publicly available**

- More complex definition (according to Open Source Initiative):
  - code is publicly available…
  - … for free…
  - … for anyone to download…
  - … using any technology… *(does not have to work in all technologies though)*
  - … and to use in their own software or derived work.

Why open sourcing model code?

- Models are already described in working papers / white papers
  - assumptions
  - building blocks
  - estimation methods
- Ideally, all relevant information on models should already be public in these papers
  - still, not always clear how to 100% replicate them from scratch just by reading papers
- Code takes transparency to next level
  - major step towards reproducibility
  - enables testing different assumptions, data, scenarios, formulas, etc
  - helps to ensure coding quality

Open source model code in context

- Central bank transparency
  - Moving from explaining decisions to sharing the objective part of “thought process”
- Academic transparency in the economics profession
  - Data and Code Availability Standard – DCAS (Koren et al, 2022)
  - Top journals: data editor, reproducibility, **code**
- Central bank code
  - Public central bank repositories (includes also non-model codes)

Central bank public code repositories by type

Best efforts list of official central bank repositories[1] Table 1

Institution	Internal utilities	Payments, CBDC, open banking	Research replication	Macroeconomic models
Central Bank of Brazil		X		
National Bank of Belgium	X			
Bank of Canada			X	X
Central Bank of Colombia			X	
National Bank of Denmark			X	
European Central Bank			X	X
Bank of France	X			
Bank of England	X		X	
National Bank of Greece		X		
Bank of Italy	X		X	X
Dutch Central Bank	X			
Bank of Norway		X		
Federal Reserve Bank of Boston (collaboration with MIT)		X		
Federal Reserve Bank of New York			X	X

Information as of 22 September 2023. [1] The helpful list of central bank open source repositories kept by the Dutch Central Bank is kindly acknowledged.

Source: Author compilation, authorities' websites.

Implications of model reproducibility

- Central bank models *inform* policy, they don't *decide* policy
  - Policy = **model output** + **qualitative expert knowledge** + **judgement**
  - Model code only informative about the first component

- Some central banks might be more willing to share code for *simulation models* to understand transmission channels rather than production-grade *forecast models*

- Model code can be helpful to other central banks, in particular those in smaller countries:
  - adapting for their own domestic economy; or
  - as workhorse model for major economies, helping them evaluate the external scenario

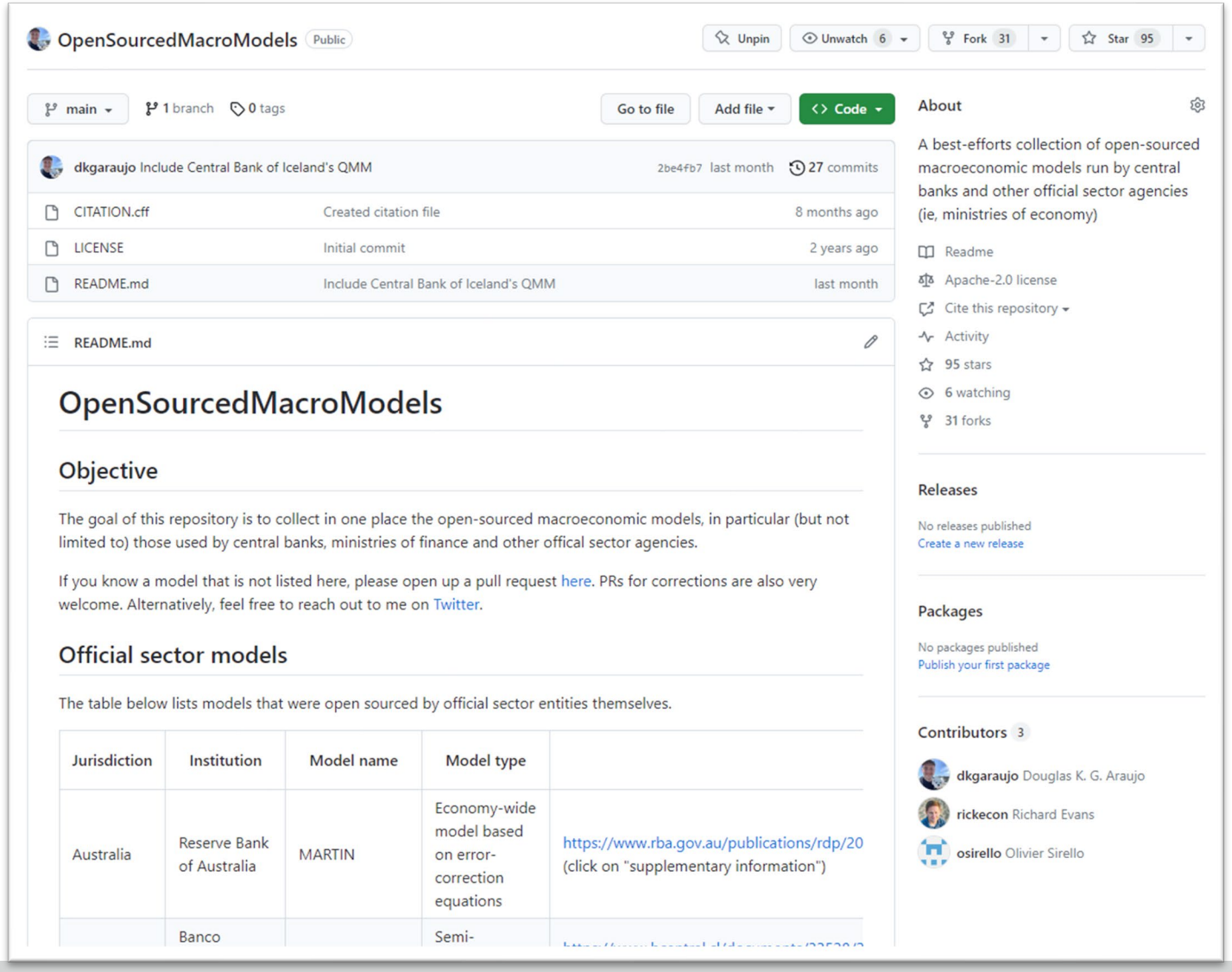
Who shares model code currently?

- Australia – RBA
- *Canada – Bank of Canada
- Chile – Banco Central de Chile
- Denmark – Min Finance
- European Union – ECB, European Commission
- Finland – Bank of Finland
- France - Min of Economics and Finance
- Germany - Min of Economic Affairs and Climate Action
- Iceland – Central Bank of Iceland
- *Italy – Banca D'Italia
- Japan – Bank of Japan
- Sweden - Riksbank
- UK - HM Treasury
- US - FRB, FRBNY

* not a model itself, but a modelling package

If you know another **open source model**, please let me know!

Who shares model code currently?



OpenSourcedMacroModels Public

Unpin | Unwatch 6 | Fork 31 | Star 95

main | 1 branch | 0 tags | Go to file | Add file | Code

dkgaraujo Include Central Bank of Iceland's QMM 2be4fb7 last month 27 commits

CITATION.cff	Created citation file	8 months ago
LICENSE	Initial commit	2 years ago
README.md	Include Central Bank of Iceland's QMM	last month

README.md

OpenSourcedMacroModels

Objective

The goal of this repository is to collect in one place the open-sourced macroeconomic models, in particular (but not limited to) those used by central banks, ministries of finance and other offical sector agencies.

If you know a model that is not listed here, please open up a pull request here. PRs for corrections are also very welcome. Alternatively, feel free to reach out to me on Twitter.

Official sector models

The table below lists models that were open sourced by official sector entities themselves.

Jurisdiction	Institution	Model name	Model type	
Australia	Reserve Bank of Australia	MARTIN	Economy-wide model based on error-correction equations	https://www.rba.gov.au/publications/rdp/20 (click on "supplementary information")
	Banco		Semi-	

About

A best-efforts collection of open-sourced macroeconomic models run by central banks and other official sector agencies (ie, ministries of economy)

Readme
Apache-2.0 license
Cite this repository
Activity
95 stars
6 watching
31 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

dkgaraujo Douglas K. G. Araujo
rickecon Richard Evans
osirello Olivier Sirello

Check out the repo!

Open-sourced central bank models used in the assessment

Best-efforts list of official models with available code

Table 2

Jurisdiction	Institution	Model name	Model ID[1]	Reference
Australia	Reserve Bank of Australia	MARTIN	AU-01	Ballantyne et al (2019)
Chile	Central Bank of Chile	MSEP	CL-01	Arroyo Marioli et al (2020), BCC (2020)
Chile	Central Bank of Chile	XMAS	CL-02	García et al (2019), García and Guerra-Salas (2020), BCC (2020)
Denmark	Ministry of Finance[2]	MAKRO	DK-01	Bonde et al (2023)
European Union	European Central Bank	BEAR	EU-01	Dieppe and van Roye (2023)
European Union	European Commission[2]	Output Gap Model	EU-02	
Finland	Bank of Finland	Aino	FI-01	Kilponen et al (2016)
France	Ministry of Economics and Finance[2]	Mésange	FR-01	Dufernez et al (2017)
France	Ministry of Economics and Finance[2]	Opale	FR-02	Daubaire et al (2017)
France	Ministry of Economics and Finance[2]	Modèle Saphir	FR-03	
Germany	Federal Ministry for Economic Affairs and Climate Action[2]	-	DE-01	BMWK and BMFin (2022)
Iceland	Central Bank of Iceland	QMM	IS-01	Daníelson et al (2019)
Japan	Bank of Japan	Q-JEM	JP-01	Hirakata et al (2019)
Sweden	Riksbank	Ramses II	SE-01	Adolfson et al (2013)
United Kingdom	HM Treasury and Office for Budget Responsibility[2]	OBR model	GB-01	Office for Budget Responsibility (2013)
United States	Federal Reserve Bank of New York	DSGE.jl	US-01	
United States	Federal Reserve Board	FRB/US	US-02	Brayton et al (2014), Laforte (2018)
United States	Federal Reserve Board	EDO	US-03	Chung et al (2010)

Information as of 22 September 2023. [1] Model ID is used throughout the text to refer to each model. [2] Other official financial agencies such as Ministries of Finance are also in scope of this compilation.

Source: Author compilation, authorities' websites.

Benchmarking model open sourcing

Core criteria

- Open access
  - Easy to find
  - Free access
- Documentation
  - Academic description
  - API documentation
  - Input data documentation
  - Output documentation
- Replication
  - End-to-end execution
  - Instructions/tutorial/vignette
  - Minimum requirements

Additional criteria

- Software
  - Software open availability
  - Testing
  - Technology neutrality
  - Availability of past versions
  - Explicit code versioning
- Contribution
  - Contact with software maintainers
  - 3rd party contributions possible
  - Contribution guidelines
- License
  - Explicit open source license

Characteristics of open-sourced macroeconomic model codes

Best efforts list of official models with available code

Table 3

	Core criteria									Additional criteria									
	Open access		Documentation				Replication			Software					Contribution			Lic.	
Model ID[1]	Easy to find	Free access	Academic	API	Input data	Output data	End-to-end	Instructions	Requirements	Open software	Tech neutrality	Testing	Explicit version	Past versions	Contact	Third parties	Guidelines	Explicit license
AU-01	X	X	X	X		X	X	X	X									X
CL-01	X	X	X	X	X	X	X	X	X	X	X				X			
CL-02	X	X	X	X		X	X	X	X	X	X				X			
DE-01		X	X	X	X		X			X			X	X				
DK-01	X	X	X	X	X		X		X		X		X	X	X	X		X
EU-01	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
EU-02	X	X		X	X		X		X				X	X				
FI-01	X	X	X	X	X	X	X		X	X	X							
FR-01	X	X	X	X	X	X	X	X	X				X	X	X	X		X
FR-02	X	X	X	X	X	X	X	X	X				X	X	X	X		X
FR-03	X	X				X	X	X	X				X	X	X	X		X
IS-01	X	X	X	X	X	X	X	X										X
JP-01	X	X	X	X	X	X	X	X	X						X			X
SE-01	X	X	X	X			X	X		X	X				X	X		
GB-01	X	X	X	X	X		X		X									
US-01	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X
US-02	X	X	X	X	X		X	X	X	X	X							
US-03	X	X	X	X			X			X	X							X

Information as of 22 September 2023. [1] See Table 2.

Source: Author compilation, authorities' websites.

Core criteria: general comments

- Most models fulfil core criteria for code sharing
- Main “core” gaps found in:
  - Documentation of data
  - Instruction usage
- These can be expected to improve over time, as interest by third parties increase
  - Could also be good for internal purposes, eg onboarding of new economists

Most common programming languages for macroeconomic models

Best efforts list of official models with available code

Table 4

Languages	Open source	Multi-platform	Observations	Models
Matlab (or Octave)	No (Yes)	Yes	Most models with Dynare	CL-01, CL-02, EU-01, SE-01 US-03
EViews	No	No		AU-01, GB-01, JP-01[2], US-02[2]
Portable TROLL	No	Yes (raw code; GUI is only available for Windows)		FR-01, FR-02
Julia	Yes	Yes		US-01[1]
SAS	No	No		FR-03
EUCAM	No	No		EU-02
GAP	No	No		DE-01
GAMS	No	No	APIs available for Python, Java and other languages	DK-01

Information as of 22 September 2023. [1] Based on previous Matlab code. [2] Also has a Python version.

Source: Author compilation, authorities' websites.

Additional criteria: contribution

- Judging by "revealed preferences", contribution from third-parties does not seem to be a driver of code sharing up until this point
- Understandably, code sharing to improve transparency of CB models (which are arguably the best for each particular economy), rather than to elicit feedback or improvements
- This could be an avenue to explore, but will probably require cultural shifts

Additional criteria: license

- Ideally, publicly available code should have an explicit license – even if it is not open source
  - Licenses protect both contributors and users
- But, so far many model codebases are not licensed
  - Lack of license: no official permission to copy, distribute or modify
  - If lack of license is intentional, should be explicit
- In practice most users will not care, but other official users (eg, central banks) will
- Wide variety of off-the-shelf licenses, and they can also be custom-made

Choose an open source license

An open source license protects contributors and users. Businesses and savvy developers won't touch a project without this protection.

{ Which of the following best describes your situation? }



I need to work in a community.

Use the **license preferred by the community** you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to **add a license**.



I want it simple and permissive.

The **MIT License** is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, **.NET**, and **Rails** use the MIT License.



I care about sharing improvements.

The **GNU GPLv3** also lets people do almost anything they want with your project, *except* distributing closed source versions.

Ansible, **Bash**, and **GIMP** use the GNU GPLv3.

Practical advice to enhance code availability

- Make it easier to find
  - Listing available code or point to repository in one place in the central bank's website
- Make it easier to use
  - Create tutorials and examples
  - Be explicit about the license
- Make it easier to contribute
  - Set up a code repository
- Make it more accessible and future-proof
  - Consider adapting code requiring proprietary software to OS languages (AI can help)

Tentative conclusions

- While some central banks share code for many years now, discussion on its benefits and challenges is still scarce
- This work hopes to shed light on this practice, to help CBs harness the benefits from transparency when that is convenient
- Code-sharing can boost CB cooperation and technical assistance, but it can also be a benefit for internal staff

Many thanks!
Question, feedback, criticisms…?

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Leveraging open-source software and data standards as the backbone of your open data strategy[1]

Darran Hodder and Matthew Nelson,
Regnology

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

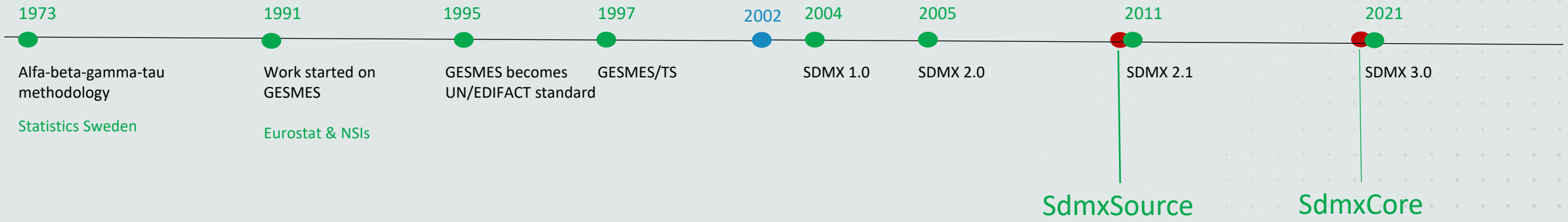
Leveraging open-source software & data standards as the backbone of your Open Data Strategy

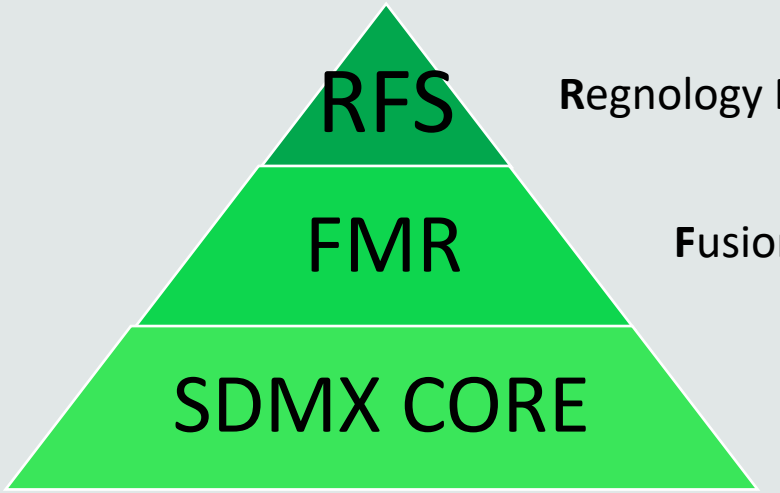
IFC October 2023

Regnology

SDMX History and Open Source

The heads of the statistics organisations sponsoring the Statistical Data and Metadata eXchange (SDMX) met on 14 June 2002 to agree on concrete projects that will advance their joint initiative. The institutions and their respective constituencies favour open and effective international cooperation in maintaining and if necessary developing new e-standards for their data exchanges.





Regnology **F**usion **S**tatistics – Commercial, extending functionality of FMR

Fusion **M**etadata **R**egistry – Open Source, free

SdmxCore - Open-Source Java Software for SDMX

Business case for Open Source

- **Cost Saving** - free
- **Longevity** - reduce the risk
- **Control** – ability to make changes
- **Robust** – many eyes



Business case for Open Source

- **Cost Saving** - free
- **Longevity** - reduce the risk
- **Control** – ability to make changes
- **Robust** – many eyes



Business case for Open Source

- **Cost Saving** - free
- **Longevity** - reduce the risk
- **Control** – ability to make changes
- **Robust** – many eyes



Business case for Open Source

- **Cost Saving** - free
- **Longevity** - reduce the risk
- **Control** – ability to make changes
- **Robust** – many eyes



Business case for Open Source

- **Cost Saving** - free
- **Longevity** - reduce the risk
- **Control** – ability to make changes
- **Robust** – many eyes

There is a strong business case for Open Source, but the decision must be a considered one

There are parallels between standards and open source; both can be used to reduce risk and protect investment

Both custom and commercial software can leverage open source. The relationship can be symbiotic, benefitting both parties



Open Data

What does good look like?

- **Discoverable** – to be able to find information
- **Accessible** – to be able to easily retrieve the information
- **Interoperable** – to be able to exchange and make use of information
- **Interpretable** – to be able to understand, interpret, and use the information for various purposes



Open Data

How is this achieved?

- **Standards** – common framework and structure, consistent formatting
- **Software** – help foster the use of the standard



IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Collaborating on SDMX APIs and open-source software[1]

Brian Buffet, Stratos Nikoloutsos and Xavier Sosnovsky,
Bank for International Settlements

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BIS



Collaborating on SDMX APIs and Open Source Software

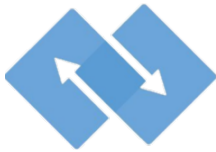
SDMX – the ISO Standard

- **Effective governance**
  - Eight international organizations (sponsors)
- **Continuously evolving**
  - SDMX 3.0 (latest release) – Improved support for micro/big/geo data
- **Collaborative evolution**
  - working groups & task forces, involving national and international organisations (and independent consultants)
  - SDMX 2025 roadmap
- **Comprised of a robust Information model + technical standard for modelling any data & implementing tools.**
- **SDMX in G20 DGI work (Data Gaps Initiative)**
  - International Data Cooperation / Data Sharing & Access
  - Data Structure Definitions (DSDs) established for several domains

BIS
EUROPEAN CENTRAL BANK
EUROSYSTEM
eurostat
ILO
INTERNATIONAL MONETARY FUND
OECD
United Nations Statistics Division
THE WORLD BANK

SDMX – The Vision

- **Improve Quality**
  - **Standardize and harmonize Data and Metadata exchange** through agreed formats**,** ensuring consistency and comparability
  - **Robust information model** supporting the entire data lifecycle
  - Foster Collaboration
- **Promote Efficiency**
  - Enhance Data Sharing, ie **support production, exchange and dissemination of statistics** for public good
  - Enable automated (machine to machine) data sharing and process automation
- **Drive Innovation**
- **Facilitate Data Use**
  - Web services
  - Connectors (Python, R, MS Power BI, …)



Web Services – the SDMX v2 API

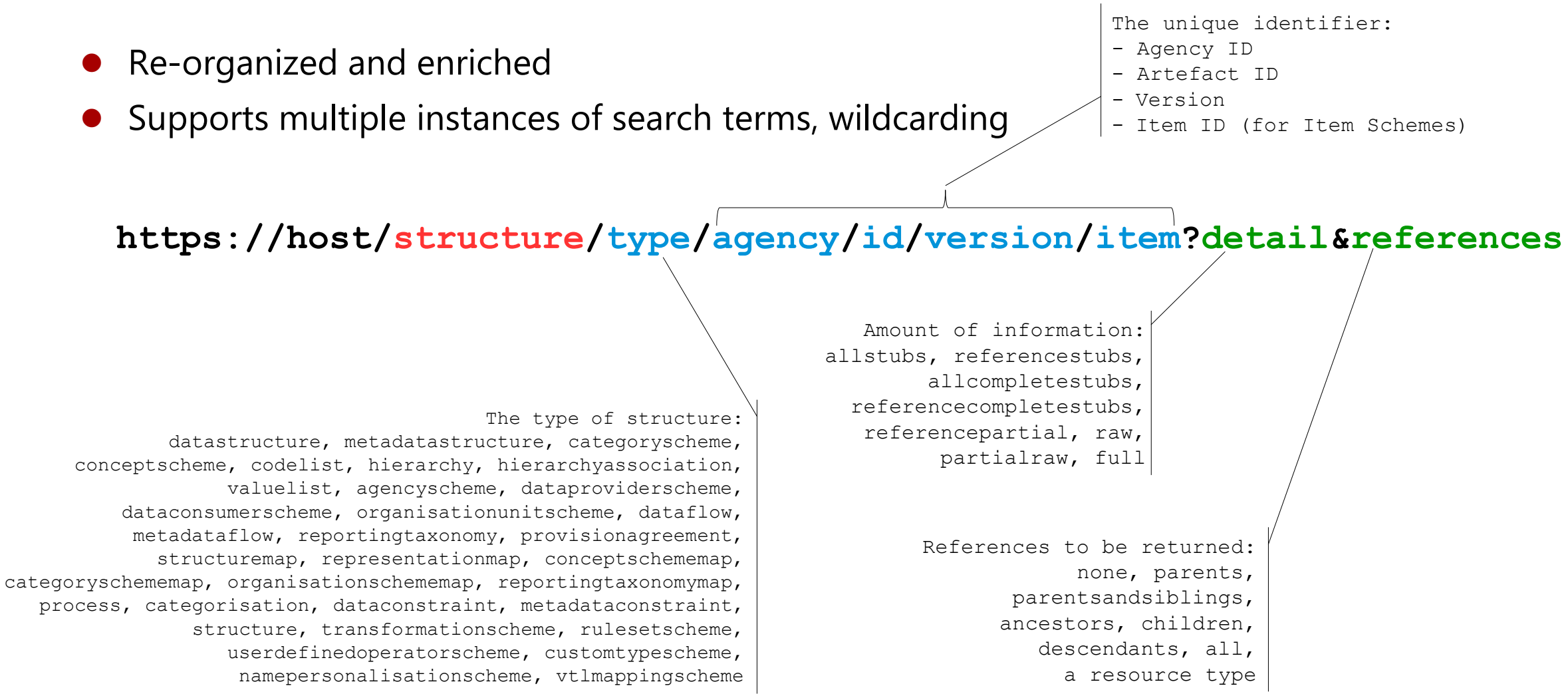
- A RESTful API maintained by SDMX-TWG Task Force 4
  - 9 releases since 2011 (first released with SDMX 2.1)

- The 9th release (with SDMX 3.0 – Sep 2021) features:
  - Enhanced data queries and reference metadata support
  - Restructured resources (breaks backwards compatibility to v1.*)
  - Metadata maintenance (additional HTTP verbs – POST, PUT, DELETE)
  - Schema query "allowed content" in more formats

- Improved clarity and documentation of the API
- Deprecated SOAP API

The structure queries

- Re-organized and enriched
- Supports multiple instances of search terms, wildcarding



The data queries

The unique identifier of the context:
- Agency ID
- Artefact ID
- Version

https://host/data/context/agency/id/version/key?c
&updatedAfter
&firstNObservations
&lastNObservations
&attributes
&measures
&dimensionAtObservation
&includeHistory

The context of data retrieval:
datastructure, dataflow,
provisionagreement

Key(s) of the series to be returned:
eg M.GR.EUR.SP00

with wildcarding:
eg M.*.EUR.SP00

With support for multiple keys:
eg M.GR.EUR.SP00, M.CY.EUR.SP00

Component-based filters (for any Dimension, Attribute or Measure):
eg **c**[REF_AREA]=CH&**c**[CONF_STATUS]=F

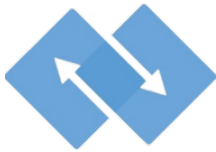
Support for operators:
eg **c**[ICP_ITEM]=**sw**:01&**c**[TIME_PERIOD]=**ge**:2015

eq	Equal
ne	Not equal
lt	Less than
le	Less than or equal to
gt	Greater than
ge	Greater than or equal to
co	Contains
nc	Does not contain
sw	Starts with
ew	Ends with
nd	And
or	Or

The data queries

https://host/data/context/agency/id/version/key?c

- **&updatedAfter** — Retrieves what changed since supplied timestamp. Must be percent-encoded (e.g.: 2009-05-15T14%3A15%3A00%2B01%3A00)
- **&firstNObservations** — Maximum number of observations starting from the first observation
- **&lastNObservations** — Maximum number of observations counting back from the most recent observation
- **&attributes** — Id of the dimension at the observation level
- **&measures** — The attributes to be returned: dsd, msd, dataset, series, obs, all, none, {attribute_id}
- **&dimensionAtObservation** — The measures to be returned: all, none, {measure_id}
- **&includeHistory** — Whether to return vintages



Other queries

- Data validity

  https://host/schema/context/agency/id/version?dimensionAtObservation
  &explicitMeasure

- Data availability

  https://host/availability/context/agency/id/version/key/componentId?c
  &updatedAfter
  &references
  &mode

- Metadata

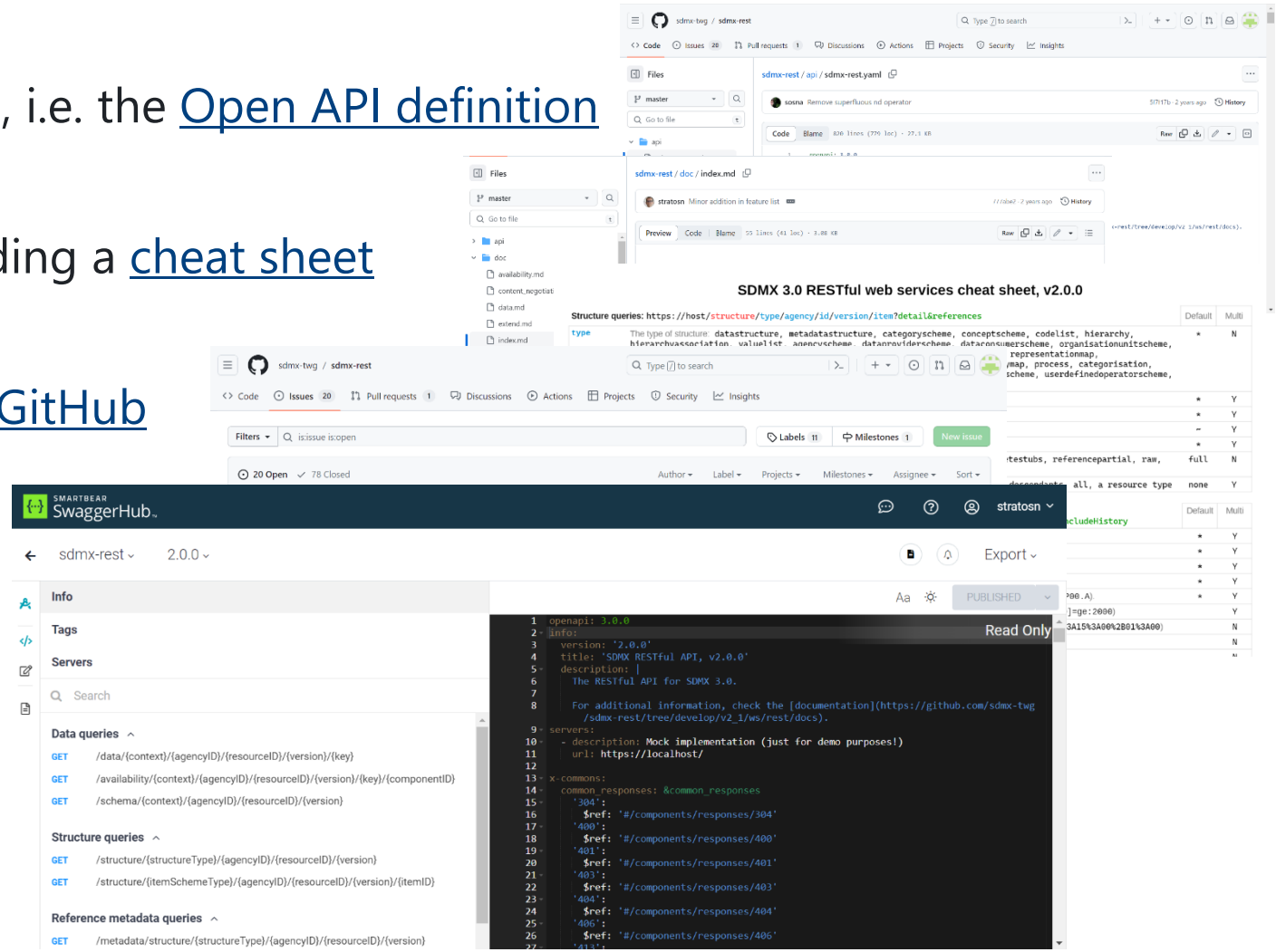
  https://host/metadata/metadataset/provider/id/version?detail

  https://host/metadata/metadataflow/agency/id/version/provider?detail

  https://host/metadata/structure/type/agency/id/version/provider?detail

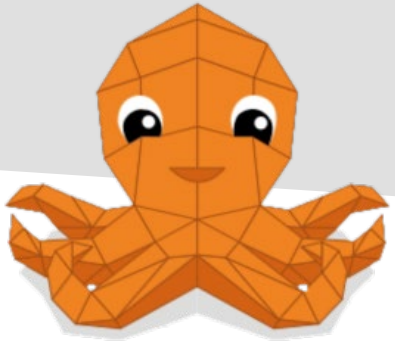
SDMX 3.0 RESTful API v2.0.0

- The normative part of the specification, i.e. the Open API definition
- The Developers' documentation, including a cheat sheet
- Request features and report issues on GitHub
- See the API spec on SwaggerHub
- The SDMX v2 API in action



BIS

sdmx.io





What is sdmx.io?

sdmx.io is not a single project but an **ecosystem of open source tools, patterns, guidance, learning materials and other resources** like pre-configured containerised environments that make the software quick and simple to deploy.

- An **open-source SDMX software and resource portal**
  - Software releases (FMR, FusionXL, etc.)
  - E-learning courses and webinars
  - Knowledge and use-case articles
- **Governance in place, incl. Prioritisation and User Advisory Groups**
- **Officially public**, after the SDMX Sponsors approval

- Share software tools and components based on SDMX
- Complement existing tools in order to achieve statistical use-cases
- Interface SDMX with other standards or tools
- Promote reusability of SDMX software



Collaboration partners and activities

- **IMF**: IMF SDMX Central is powered by FMR11 (since Sep 2023)
- **SIS-CC/Eurostat/BIS**: working closely on achieving SDMX 3.0 compliance
- **OECD/SIS-CC**: .Stat/FMR integration for the National Bank of Belgium (**NBB**) use case
  - `FMR Workbench`
  - `.Stat Suite` on sdmx.io
- **NBB:** Launch of FMR Workbench user group (targeting CBs and SIS-CC members)
- **Eurostat**: SDMX Reference Implementation for SDMX 3.0
  - `sdmx-core` evolution (Java/.Net)
- **Banca d'Italia**: `VTL Engine` & `FMR` → exploring the use cases
  - UNSD working with BdI on the first use case
- **Meaningful Data**: `SDMXthon` & `FMR` → exploring the use cases



The use-cases



Statistics production



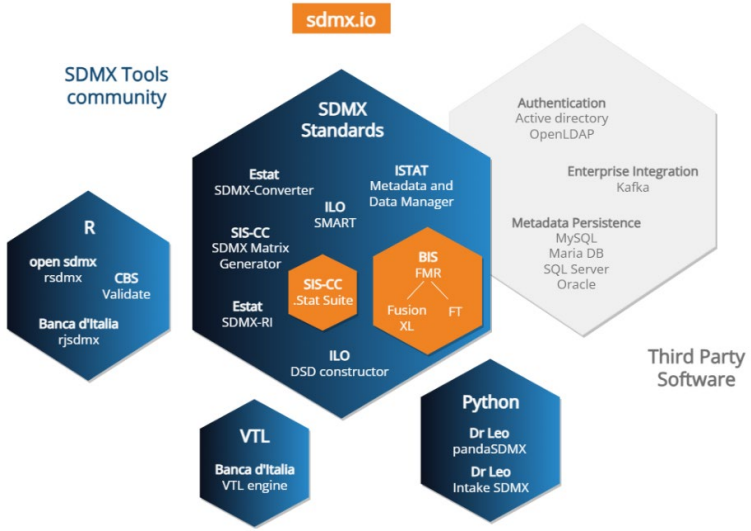
Statistics dissemination



Data domain modeling



Data collection

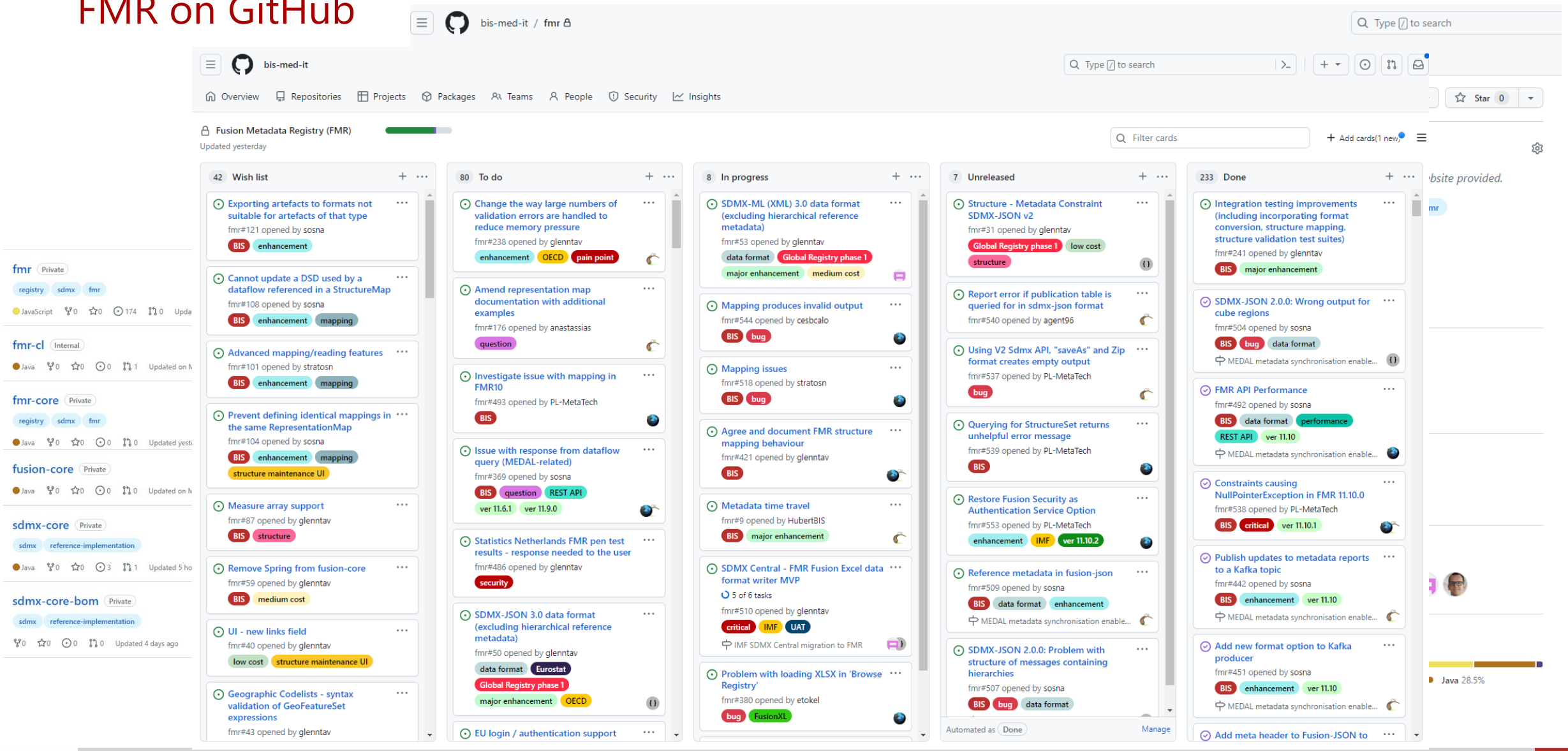




eLearning Catalogue

sdmx.io's library of learning resources covers a diverse range of topics including how to use software tools, SDMX theory and data modelling best practice.

FMR on GitHub



BIS

Thank you!

Brian.Buffett@bis.org Stratos.Nikoloutsos@bis.org Xavier.Sosnovsky@bis.org

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Future of time series: preliminary results from a BIS-IFC survey of central banks and statistical agencies[1]

Ryland Thomas,
Bank of England

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Bank of England

Future of Time Series:
Preliminary Results from a BIS-IFC survey of Central Banks and Statistical Agencies

Ryland Thomas

Bank of England

Data Science in Central Banking: Enhancing the access to and sharing of data

17-19 October 2023, Rome, Italy



Survey questions

1. Are you able to share with us a broad overview or schematic of the end-to-end process for your time series products, from getting data 'in the building' to producing briefing/publications? Do you treat time series data as distinct from other forms of data (as we do at the BoE)?

2. What commercial and/or proprietary systems do you use to support this process? In particular, we would be very interested in your experience of using FAME which we use at the BoE.

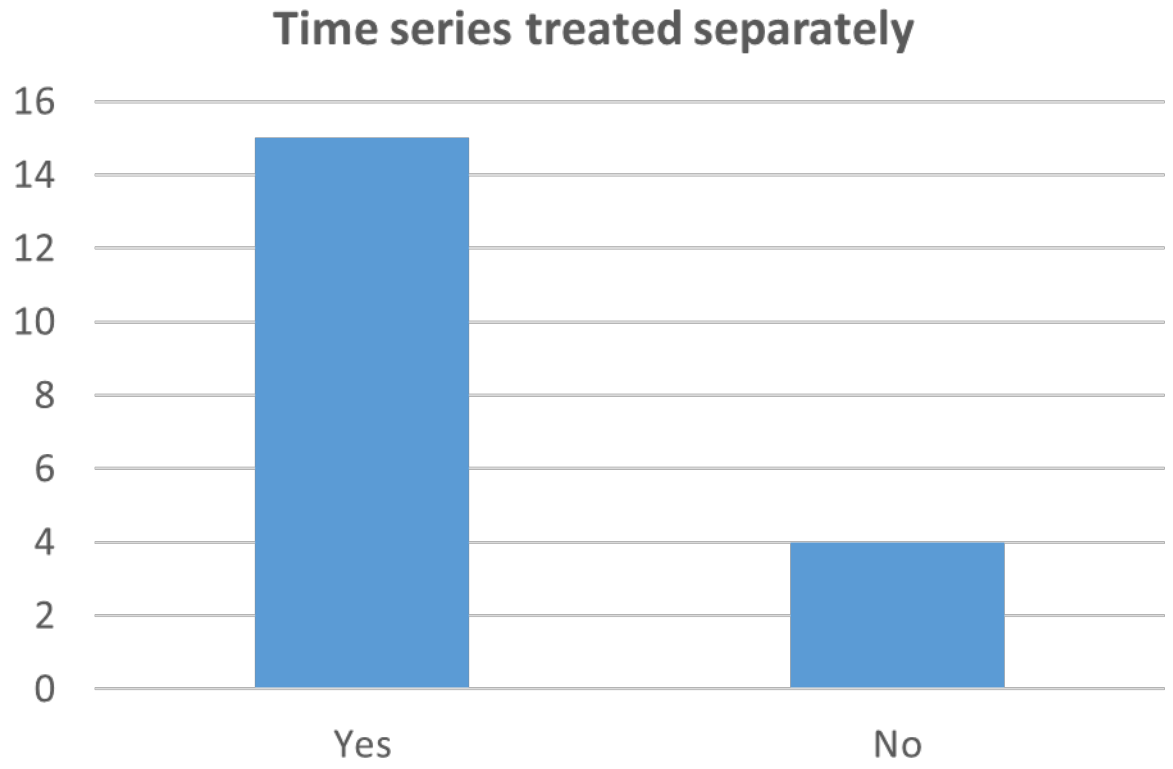
3. What are the main strengths and shortcomings of your current approach? These can include issues such as timeliness, reliability, adaptability and ease of upgrading; or costs such as licencing, development and maintenance.

4. If you could develop your system over again, what would your ideal target state be, and how would it improve on your current set-up? Have you undertaken any market research to assess what is possible in this realm, and if so, what were the key findings?

Respondents

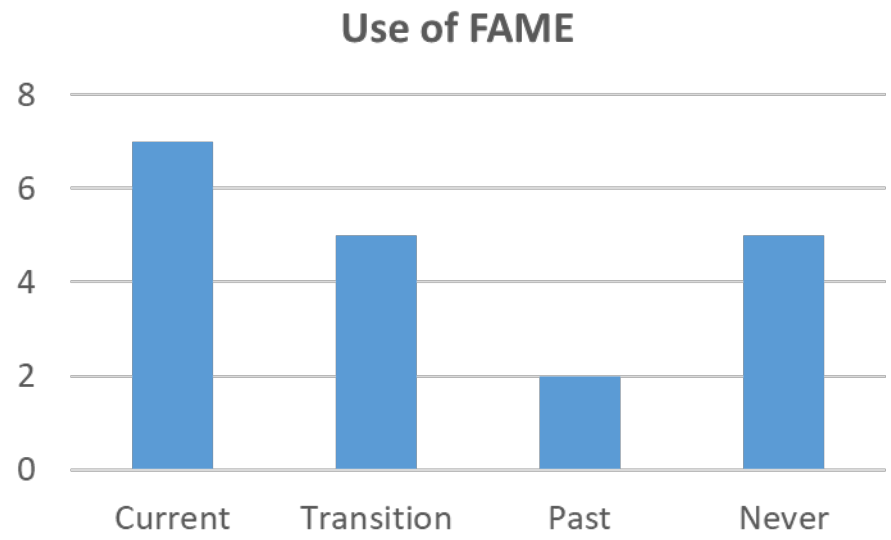
- Feedback from 18 central banks/statistical agency via BIS Irving Fisher Committee network, 19 including the Bank of England
- Detailed responses available in spreadsheet on eBIS

Time series treatment (incl. BoE)



- Most respondents treat time series as distinct from other types of data in at least some key aspects of data management and access

FAME Usage across sample (incl. BoE)



- Many of the respondents still using FAME though some clearly in transition away
- No one using something else was looking to move to FAME
- Of 19 institutions surveyed, 12 currently use FAME of which 5 are in the process of transitioning away
- 2 central banks previously used FAME but have transitioned fully away
- The 5 that have never used FAME use bespoke systems: there is no single, commonly used alternative for time series data

Reasons given to stick with FAME

- It has excellent functionality for time series data
- No other proprietary database handles traditional macroeconomic time series as effectively – e.g. Fame dynamic formulae which can be stored in the same database
- With proper support it is stable and efficient
- It is deeply embedded in analytical systems and moving away would be very costly
- It works well with Excel, through the FAME populator add-in

Reasons given to move away from FAME

- Although some users are expert in FAME, their number is dwindling, and many users do not use FAME directly but through other interfaces
- Excel use is falling, with analysts increasingly using open source analytical tools such as R and Python, and specialist data visualisation tools such as Tableau
  - FAME does not integrate as well with these packages
- It is increasingly difficult to find technology people who are able to support FAME
- FAME licence is expensive (but perhaps not compared to some proprietary databases or costs of developing and maintaining a complex data architecture)
- FAME is not designed to handle large/unstructured data sets, which are increasingly available
- Searchability/metadata are better handled outside of FAME

What systems are replacing FAME?

- No single solution has yet emerged
- Some have moved to SQL data warehouses, one to a NoSQL database, one to HBASE, one is developing a data lakehouse
- Although these solutions are better for handling larger and irregular data sets, a major challenge is to replicate the convenience, speed and efficiency of FAME for traditional time series data
- Those systems also still require work to integrate with R and Python
- Some seeking to de-couple FAME (eg 4GL language) and play to its strengths as a time series database - other tools can now do analysis and visualisation better
- FAME database coupled with open source tools seen by some as efficient low cost solution
  - Large costs and risk in starting from scratch

Other issues

- Aim if possible for “single source of truth”
  - Data is collected once and used multiple times
  - Data is stored in one system
  - “It is better to produce micro-level data and then aggregate at one central institution”
- Use of Cloud Services
- Now have a broader spectrum of end users: conjunctural analyst, forecaster, researcher, data scientist all with different requirements. These “use cases” need to be identified and mapped out.
- Key challenges seem to be:
  - How best to scale up and integrate new elements into a system especially starting from a complex position ?
  - For current users what role does FAME play in that? Is it a help or barrier to fundamental change?
  - How to manage the demands of end users with increasingly different and complex requirements ? How to manage transition of existing users to new state.

Lessons so far

- **FAME isn't the issue *per se***
  - Still performs well as a fast, efficient time series database and some are sticking with it for the time being
  - There is nothing out there that does all the things that FAME currently does well
- It is other elements that cause trade-offs for CBs and SAs
  - **Acquisition/data input** is getting more complex with different types of data
  - **Searchability/metadata** is better served outside FAME
  - **Analytical time series functions and econometrics** better in open source tools than FAME 4GL
  - A range of **visualisation** tools are now available eg Tableau
- **Key challenges seem to be:**
  - How best to plug in and fit different elements together especially starting from a complex position ?
  - How to manage transition of users familiar with old systems to new ways of working
  - How to manage scalable solutions to meet demands of end users with increasingly different needs?
- **Sharing knowledge and collaboration going forward**
  - Sessions at future IFC conferences or bespoke co-development workshops, BIS innovation hub can help



FoTS - Feedback from other Central Banks

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Coding time series with machine learning[1]

Ayoub Mharzi,
International Monetary Fund (IMF)

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



STATISTICS

Coding Time Series With Machine Learning

3RD IRVING FISHER COMMITTEE WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING"

BANCA D'ITALIA - OCTOBER 18, 2023

Ayoub Mharzi – Data Scientist
IMF Statistics Department

Context

Increasingly available text data

Surveys



Data tables



Websites



Challenges and opportunities

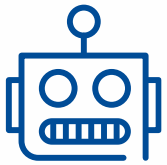
Nonstandard terminology



Outdated processes



New technologies



Objectives

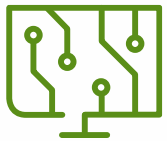
Understand the data



Explore new approaches



Design solutions



What is “coding time series”?

- IMF member countries publish economic time series data in their National Summary Data Pages (NDSPs)
- IMF staff map these data to the internal Catalogue of Time Series (CTS) to ingest in IMF database.
- This is referred to as “coding time series.”

DATASTRUCTURE	IMF:ECOFIN_DSD(1.0)	Datastructure
DATASTRUCTURE_NAME	ECOFIN Data Structure Definition	Datastructure Name
DATA_DOMAIN	CGO	Dataset
REF_AREA	KW	Country
COUNTERPART_AREA	_Z	Counterpart area
UNIT_MULT	6	Scale = Million
FREQ	A	Frequency = Annual
COMMENT	Source: https://www.cbk.gov.kw/en/stati	Source / Observation status

Descriptor_Alt	Descriptor	INDICATOR	2009	2010
	General public services and defence			
	Public order and safety			
	Education affairs and services			
	Health affairs and services			
	Social security and welfare affairs and s			
	Housing and community affairs and serv			
	Recreational, cultural and religious affa			
	Fuel and energy affairs and services			
	Agricultural affairs and services			
	Manufacturing affairs and services			
	Transportation and communication affa			
	Other economic affairs and services			
	Expenditures, not classified by function			

Challenges

Time consuming and manual.

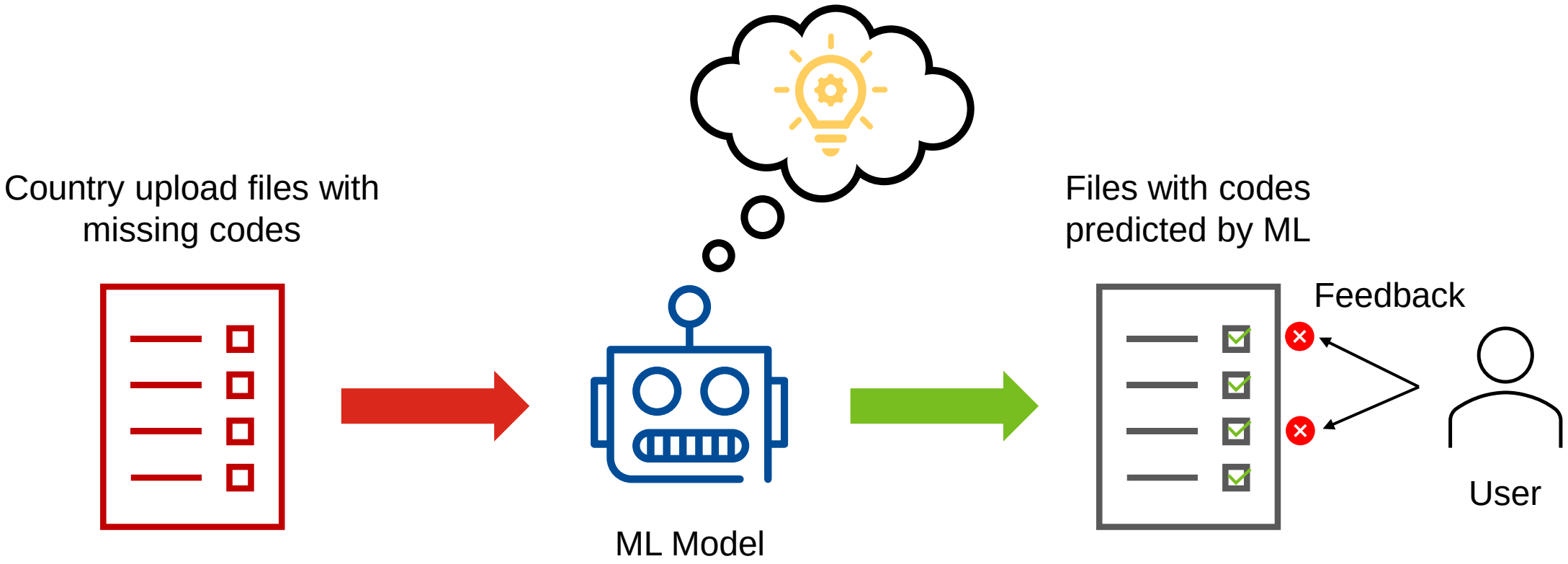
Finding similarities is not simple
when new wording is used for time series descriptors.



Not always straightforward.
E.g., “Health, affairs and services”:
218 indicators with the word “health”;
32 with the word “service.”

Subject to Human Error

Hypothesis


Country upload files with missing codes
ML Model
Files with codes predicted by ML
Feedback
User

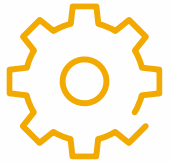
Data processing



~142k coded indicators from 53 countries across 4 macroeconomic sectors



Build a master dataset based on already coded data files to train/test



Light data processing and cleaning approach to allow reproducibility and scalability



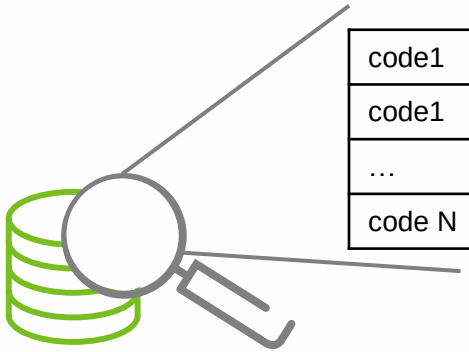
A few data files were in French, Spanish and Portuguese

- **Remove country-specific codes**.
- **Construct descriptors hierarchy.**
- **Keep only alphanumeric characters and remove stop words** (“a”, “the” etc.)
- **Keep available metadata fields** such as data domain, frequency, etc.
- **Train and Test data**: 90/10 split for training and testing/validation

Methodology

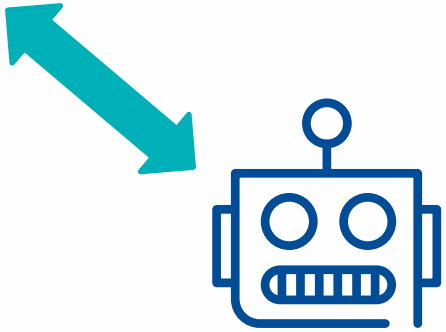
- Our data have specific terminology, a **supervised ML approach** is used.
- Start with proof of concept (PoC) to test the **feasibility and adequacy** of several ML models and feature extraction techniques (TF-IDF, word embeddings, Skip-Gram, logistic regression, etc.)
- Use an open-source text classification library, **fasttext**
- Text classification with Skip-Gram model and 3 n-gram
- Vectors dimension: 100, Epochs: 20

High-level underlying logic

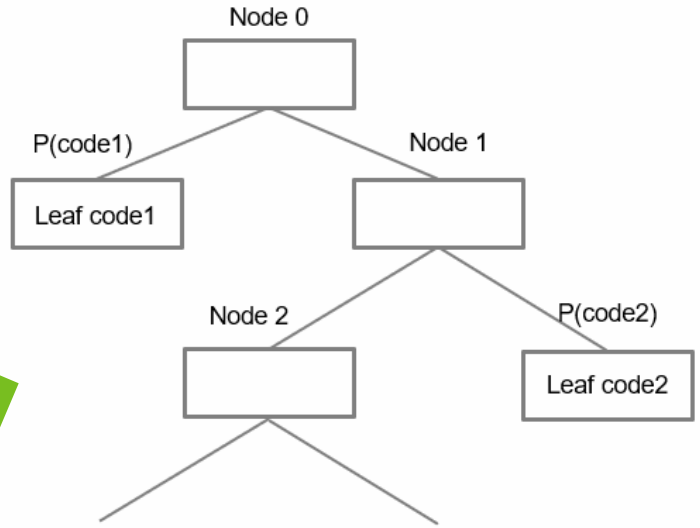


code1	descriptor1
code1	descriptor2
…	…
code N	descriptor M

Constructed hierarchical descriptors with their respective codes



Model learns from training data and creates a decision tree calculating probabilities of codes based on descriptors



Overall process

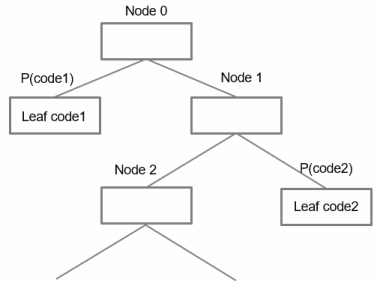
1 Data Processing

Health	Affairs	and	Services

3X

DATA_DOMAIN_CGO	METHODOLOGY_2014_Manual	Health	Affairs	Services

Model 2



3 Rank codes based on their probability

Code N	Probability N
code2	Probability 2
…	…
code 1	Probability 1

4 Returning top 5 codes with highest probability

GEL_G14_CG	Government and Public Sector Finance, Expenditure by COFOG, Central Government, Health [2014 Manual]
…	…
…	…
…	…
…	…

Results

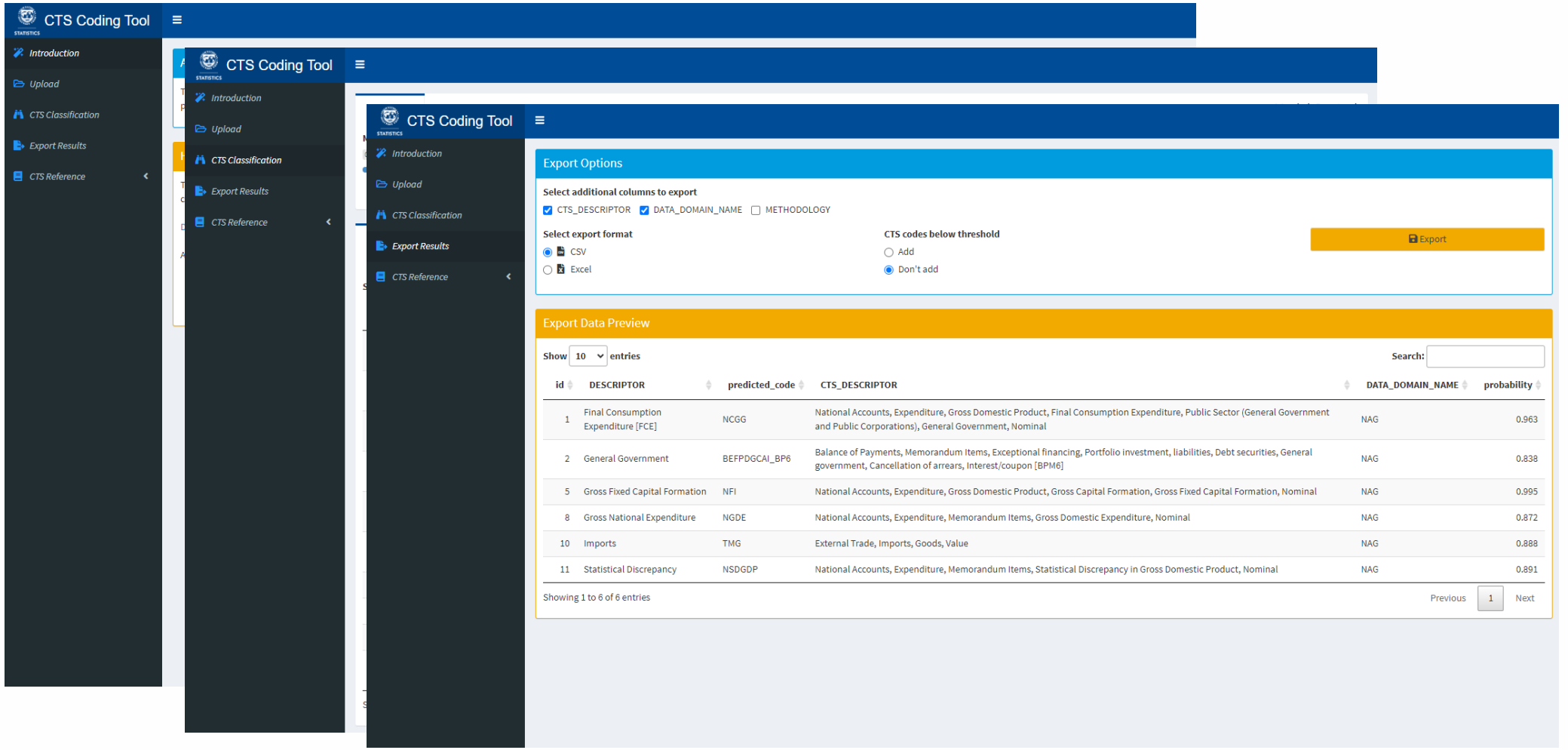
- Precision ranging by data domain, majority of domains precision around 90%
- Recall ranging by data domain, majority of domains recall around 90%
- Cases of low precision and recall due to:
  - Low number of time series for certain domain in training dataset
  - Original data file's structure not allowing to construct full hierarchy indicators
- For a data file with ~500 time series, the model takes about 2 seconds to return predicted codes

Domain	N	precision	recall
balance payments bpm	5,014.0	0.965	0.966
m&b central bank	1,309.0	0.915	0.914
m&b depository corporations	1,293.0	0.92	0.922
international investment position bpm	1,205.0	0.975	0.973
direction trade statistics	771.0	0.966	NA
financial soundness indicators sectoral financial statements	544.0	0.965	0.963
financial soundness indicators institution deposit takers ofcs	533.0	0.32	0.319
national accounts gross domestic product	398.0	0.885	0.872
financial soundness indicators data report form	348.0	0.965	0.963
m&b financial corporations	344.0	0.925	0.924
balance payments	337.0	0.99	0.991
financial access survey	267.0	0.86	0.861
international investment position	247.0	0.98	0.980
merchandise trade	246.0	0.69	0.679
consumer price index	224.0	0.915	0.915

Achievements

- Use open-source tools to build a solution
- Built a specific vocabulary and train the model
- Light data processing allowing reproducibility and scalability of the tool
- Achieved for most domains a high-level Precision and Recall
- Create an **R package** with the documented code available for sharing upon request
- Build an **R-Shiny app** for end users based on received feedback

Shiny App


CTS Coding Tool
Introduction
Upload
CTS Classification
Export Results
CTS Reference
Export Options
Select additional columns to export
CTS_DESCRIPTOR
DATA_DOMAIN_NAME
METHODOLOGY
Select export format
CSV
Excel
CTS codes below threshold
Add
Don't add
Export
Export Data Preview
Show 10 entries
Search:
id
DESCRIPTOR
predicted_code
CTS_DESCRIPTOR
DATA_DOMAIN_NAME
probability
1 Final Consumption Expenditure [FCE] NCGG National Accounts, Expenditure, Gross Domestic Product, Final Consumption Expenditure, Public Sector (General Government and Public Corporations), General Government, Nominal NAG 0.963
2 General Government BEFPDGCAI_BP6 Balance of Payments, Memorandum Items, Exceptional financing, Portfolio investment, liabilities, Debt securities, General government, Cancellation of arrears, Interest/coupon [BPM6] NAG 0.838
5 Gross Fixed Capital Formation NFI National Accounts, Expenditure, Gross Domestic Product, Gross Capital Formation, Gross Fixed Capital Formation, Nominal NAG 0.995
8 Gross National Expenditure NGDE National Accounts, Expenditure, Memorandum Items, Gross Domestic Expenditure, Nominal NAG 0.872
10 Imports TMG External Trade, Imports, Goods, Value NAG 0.888
11 Statistical Discrepancy NSDGDP National Accounts, Expenditure, Memorandum Items, Statistical Discrepancy in Gross Domestic Product, Nominal NAG 0.891
Showing 1 to 6 of 6 entries
Previous
1
Next

Next steps

- Incorporate user feedback into the end-to-end pipeline
- Officially roll out the tool into production and implement monitoring mechanisms
- Test the model and solution for out-of-scope time series
- Make the R package sharable upon request

Lessons learned

- **NLP/ML is a fast-changing field:** methods such as TF-IDF or Random Forest, although still used as baseline, are increasingly challenged by more recent methods such as word embedding, transformers, and other deep learning models.

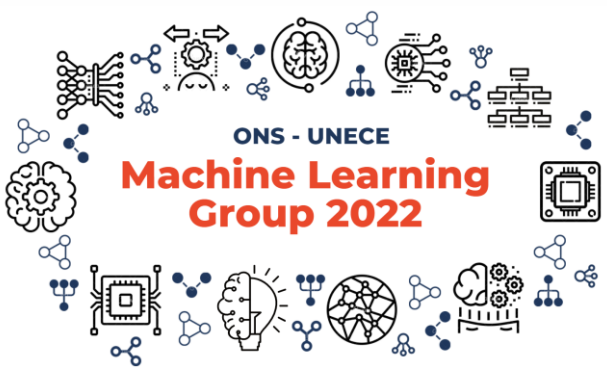
- **Generative AI is changing the landscape of NLP:** considering if it is worth developing in-house solution or opt for off-the-shelf solutions such as ChatGPT.

- **Importance of the maintenance strategy:** Preparing in advance the monitoring and re-training plan is crucial for the sustainability of the model and can increase the acceptance and facilitate the transition from PoC to production.

- **Machine Learning solutions take time to implement to reach robust and high-quality results**.

- **Collaboration and knowledge sharing can have a big impact :** peer organizations can face similar challenges and we believe a solution like ours could be adapted and implemented by others.

Useful References



- Text and classification
  - Model retraining

Measure (2017): use artificial intelligence (AI) to map worker injury to internal coding

Statistics Canada

Evans et al (2021), “Need for Speed: Using fasttext (Machine Learning) to Code the Labor Force Survey”

Thank You

Team: Alberto Sanchez, Alessandra Sozzi, Ayoub Mharzi, Lamya Kejji and Yamil Vargas
Ayoub Mharzi: Amharzi@imf.org
Analytics Team: STA-Analytics@imf.org

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Siamese neural networks for detecting banknote printing defects[1]

Katia Boria, Andrea Luciani, Sabina Marchetti and Marco Viticoli,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Siamese neural networks for detecting banknote printing defects

K. Boria, A. Luciani, S. Marchetti and M. Viticoli [1]

Abstract

The production of banknotes is a complex process, composed of different printing steps, in which various kinds of defects can be generated that, if not adequately monitored, can lead to production waste, significantly impacting productivity and costs. This paper proposes a new approach for identifying defects during banknote production using *one-shot learning* methods. These methods rely on a small number of observations in order to train a Siamese neural network to reproduce the similarities between pairs of samples. The network can then identify defects in new banknote images by comparing them to benchmark samples. The proposed approach allows the correct identification of some specific defects on banknotes, even with limited training data, laying the foundation for the development of a solution for recognition and intelligent classification of defects on banknotes.

Keywords: banknotes production, artificial intelligence, neural networks, one-shot learning, quality control

JEL classification: C45, L15, L69, O39

Contents

1. Introduction 2

2. Defects in banknote printing and production variability 3

3. Methodology 4

   3.1 Data 4

   3.2 Deep learning and convolutional neural networks 5

   3.3 Convolutional neural networks for the study of banknotes 7

   3.4 One-shot learning and Siamese neural networks 8

4. Siamese networks for bite-type defect detection 10

[1] Bank of Italy. The views expressed in this article are those of the authors and do not necessarily reflect those of the Bank of Italy.

4.1 Model architecture and training 10
4.2 Results 14
5. Conclusions and possible extensions 18
References 19

1. Introduction

Similar to other industries, banknote production relies on a comprehensive system of quality controls. In the Eurosystem, such controls are conducted throughout the entire production process (raw materials, semi-finished and finished banknotes) in compliance with the relevant ISO standards and through rigorous procedures defined by the European Central Bank (ECB) and harmonized among the various printing works to ensure a high degree of homogeneity of euro banknotes. With particular reference to quality checks on finished banknotes, the number and type of printing defects acts as a key discriminating parameter for the conformity of each production batch. Today, the assessment of defects is carried out by operators through visual examination of the banknotes. Therefore, although there is an articulated set of safeguards aimed at ensuring the application of objective criteria, the experience and sensitivity of the workers play a decisive role in the evaluation process, which could therefore be affected by subjectivity factors.

In principle, automating quality control of industrial products helps make processes more accurate and efficient. Artificial intelligence, and in particular neural networks in deep learning, are useful in this scenario due to their ability to represent and describe systems characterized by high complexity and variability. At the printing stage, the evaluation and identification of potential defects by neural networks enables a reduction in operational costs by devolving to quality control specialists the detailed analysis of potential defects found and their causes.

This paper reports on what has been observed as part of an exploratory study on the use of neural networks for automated recognition and classification of a particular type of banknote defects in the context of quality control. Section 2 provides an essential overview of the quality control process for detecting print defects on banknotes. Section 3 outlines the main technical features of Siamese neural networks - that is, the class of deep learning models used in the present study - as well as the set of input images used for training and calibration. The main aspects of model training and the results achieved using Siamese networks are reported and discussed in Section 4. Finally, Section 5 outlines possible directions for methodological and operational research and development.

2. Defects in banknote printing and production variability

As part of the production process, strict quality controls are carried out on the produced and semi-finished banknotes, in compliance with the European Central Bank's (ECB) quality benchmark decisions and the requirements of the Integrated Quality, Environment and Occupational Health and Safety Management System.

Together, these checks provide important information about potential process criticalities and final product defects; this information can be profitably used to intervene in the process itself to reduce production waste in the printing process and even marginal defects in the finished product.

The verification of many parameters has already been automated through the introduction of optical measuring systems. However, visual acceptance checks on finished banknotes are still assigned to highly trained staff, who perform these tasks manually on a representative sample of each production batch. Defects are then identified and classified based on ECB documents and reference samples. Classification is done by defect type and size, as well as by area of the surface where different defects may be found. The final conformity assessment of the production batch then depends on the number and classification of defects found on the representative sample of the batch. This process is based on a set of rules and comparison with reference samples, similar to what is performed automatically by *deep learning* techniques.

For the experimentation described in this paper, a pilot application was chosen for the detection of defects on the €50 banknotes of the Europa series[2] currently in production at the Bank of Italy. In this first phase, the study focused on a specific type of defect present on the European flag, the so-called *bite* defect, which consists of a lack of ink on a homogeneous background, and can vary in shape, size and position. The choice was determined by the relevance and number of occurrences referring to this defect type in the production batches examined at the time of experimentation.

The variability of the banknotes was another important factor of analysis that had to be taken into account in the experimental phase. Since banknotes result from different printing stages, they may exhibit appreciable differences between them that reflect normal production variability and hence should not be considered defects. In particular:

1. the position of elements on the banknote surface may vary within fixed margins of tolerance;

2. the density and colour of inks can vary, resulting in different shades and intensities;

3. the acquisition phase, either manual or automatic, can introduce additional image variability.

[2] The Europa series is the second series of Euro banknotes, introduced from 2013 to 2019.

In addition to product variability, it was also necessary to account for the variability of defects to be identified and classified, differing in size, location and colour intensity.

We initially proceeded with manual acquisition of high-resolution (~2656x1467 pixels, 24-bit colour, RGB model) images of each banknote, with the aim to minimize the probability of acquisition defects. In order to reach a sufficient number of samples to be processed, we also considered images acquired automatically during the production process, which had a lower resolution (~675x388 pixels, 24-bit colour, RGB pattern). For both cases, the acquisition of banknote images required operator intervention, although in the second case the procedure was significantly faster. Manual acquisition, on the other hand, allowed for a more careful selection of the defects to be acquired, representative of a longer production time frame (several weeks) and greater production variability. At present, the two different acquisition approaches are to be taken into account when training the network for their effects on the availability, extent and variability of the image sample.

3. Methodology

The proposed application focuses on the detection of bite-type printing defects using Siamese neural networks. This section introduces the reader to the technical aspects of the analysis conducted, which include: i) the type of data representation of digital images, i.e., ordered sets of matrices called *tensors*; ii) the convolutional neural networks used for tensor processing, and the limitations of such models in this specific application of banknote images elaboration; iii) the specialization of deep learning, known as one-shot learning, and the related class of Siamese neural networks.

3.1 Data

From a technical point of view, the images belong to the type of data in "unstructured" format. For the purpose of statistical learning, each image is mathematically represented by an ordered set of three matrices, or three-rank tensor, which captures its relevant characteristics. Specifically, the so-called pixels[3], i.e. the elementary graphic components of each image, are mapped into groups of elements of the matrices. Within a tensor, each element of a matrix assigns a numerical value to its corresponding pixel; in the case of 24-bit colour, values vary between 0 and 255 and indicate the intensity of the colour associated with the matrix according to its placement in the sequence. In the case of the RGB model, the first matrix represents the so-called Red colour channel (*R*), the second matrix represents the Green colour channel (*G*), and the third represents the Blue

[3] Pixels are minimal units of the surface area of a digital image.

colour channel (*B*)[4]. As an example, according to such a representation scheme, a blue pixel in the image is obtained by setting to zero the related matrix values in the R and G channels, while setting to 255 the matrix value in the R channel. In the general case, given an image resolution of NxM pixels encoded via the RGB model, the image will be numerically represented by a three-rank tensor containing NxMx3 integers, each of which can as mentioned vary between 0 and 255.

3.2 Deep learning and convolutional neural networks

Advanced image processing generally makes use of artificial intelligence techniques, particularly deep learning, for automatic recognition or classification of objects represented in an image. Deep learning is a specialization of *machine learning*, extending the process of learning a representation-rule mapping inputs to outputs to sequences of representation-layers (Goodfellow, Bengio & Courville, 2016). The approach is based on the *data-driven* paradigm, according to which the information necessary for the understanding of a given phenomenon. The proper functioning of the related representation model[5] is learned from a set of empirical data, called the *training* set. By design, training of complex machine learning model requires the specification and formalization of a limited number of assumptions about the phenomenon being analysed.

Neural networks are models for deep learning equipped with internal processing components (layers) iteratively delivering a representation of inputs. Within each layers, information processing is based on units called *neurons*. In their simplest formalization, so-called "fully connected", layers are composed of neurons that receive, process, and transmit all information coming from a preceding layer to the next.

In a fully connected neural network, input observation $\boldsymbol{y}^0 \in \mathbb{R}^{n_0}$ is sequentially mapped into projection $\boldsymbol{y}^{l+1} \in \mathbb{R}^{n_{l+1}}$, the projection of the input observation produced by the l-th fully connected layer, with $l=0,\ldots,L$, derived through the recursive formula:

$$\boldsymbol{y}^{l+1} = g^l(\boldsymbol{w}^l \boldsymbol{y}^l + \boldsymbol{b}^l).$$

g^l is a generic (possibly non linear) *activation function* that applies to the linear transformation of the vector $\boldsymbol{y}^l \in \mathbb{R}^{n_l}$. Parameters' values, i.e. the elements from matrix of weights $\boldsymbol{w}^l \in \mathbb{R}^{n_{l+1} \times n_l}$ and from vector of bias terms $\boldsymbol{b}^l \in \mathbb{R}^{n_{l+1}}$, are defined by the learning process, provided a random initialization, based on the training set.

The training of a model is also subject to a "calibration" phase, to validate its performance on an unseen set of observations and to establish optimal values of the model's hyper-parameters, i.e. its architecture - defined by the type of layers

[4] Another frequently used model is the so-called HSV, for which cell values vary between 0 and 360. The first channel corresponds to the hue (*Hue*). The second channel identifies the degree of saturation (*Saturation*). The third channel refers to the brightness of the colour (*Colour Value*).

[5] The proper functioning of the model is expressed by the value of an objective function, which guides its training process.

that regulate transmission of information; the depth of the neural network, i.e., the number of hidden layers; the number of neurons per layer and the activation function - and features of the training optimization routine, including the learning algorithm.

Image processing applications typically make use of neural networks containing convolutional layers (*Convolutional Neural Networks*, CNNs)[6]. CNNs are inspired by the biological mechanism of visual perception. They operate a reduction in the complexity of the input data through targeted extraction of a synthesis to make processing efficient. Technically, convolutional layers process the information represented by a tensor according to a procedure based on recursive filters that slide along surface dimensions, that is, throughout groups of pixels adjacent to each other, or region. From each region, projections called *filter maps* are extracted.

In order to ensure the stability of the information processing system, convolutional layers alternate with *pooling* layers that synthesize and aggregate the contribution of several adjacent regions into a filter map (Boureau *et al.,* 2010). Typically, pooling layers collect either the average value (*average pooling*) or the maximum value (*max pooling*) of adjacent regions from each region. Pooling usually results into a lower-dimension tensor whose elements will serve as input for the next layer.

By construction, the recursive region-based nature of the detection scheme entails replication of information from same pixel across multiple projections. As a consequence, each filter map is sensitive to small shifts of an element within the image area, and thus to small shifts of the pixels that contribute to define it. This is particularly relevant in the use case considered, since banknotes are characterized by an established organization of the elements represented on their surface, although with tolerated variability given by their reciprocal positioning and colour intensity. The use of CNN for classification in image processing can be distinguished into approaches based i) on input segmentation and recognition of different elements (*object recognition*) or ii) on labelling an input image (*image recognition*). Object recognition can either involve image segmentation into specific areas of the surface identified as potentially relevant, to be passed on to ad hoc image recognition models for labeling individual features, or incorporate the two phases within a single neural network (Redmon *et al.,* 2016). CNNs for recognizing different elements of an image are usually characterized by complex architectures, which are associated with significant computational costs for the training and application phases. The literature related to image classification has over time produced CNNs characterized by high performance on large volumes of data, partly due to the refined complexity of the architectures (Krizhevsky *et al.,* 2017). In particular, the accuracy of the classification process is enabled by deep neural networks organized according to hierarchical structures (Simonyan & Zisserman, 2014; Szegedy *et al.,* 2015; Yan *et al.,* 2015) and granular parameterization of layers, in which the size of recursive filters approximates the single pixel (Zeiler & Fergus, 2014).

[6] LeCun *et al.* (1989).

The training process of a CNN consists in the global optimization of the model, through the progressive adaptation of its parameters' value. To handle high dimensionality of the input and the recursive nature of its processing, CNNs used for image recognition are characterized by a significant number of parameters to be estimated, and thus require large training sets to converge. Since acquisition of a large number of training sample is especially costly in our setup, sound training of CNNs represents a challenging task.

Flexibility of models characterized by complex architectures is known to expose classification exercises to so-called *overfitting*. Overfitting consists in the modeling of irreducible error in the training data, that penalizes the model's ability to generalize the prediction performance. To mitigate this risk and enhance convergence of the training process, while addressing the computational complexity of processing, a number of steps can be taken. Our application, in particular, makes use of: i) ad hoc methods for parameters initialization, assigning initial values within optimal ranges (Domínguez, 2020), such as the so-called *Xavier* method for defining the parameters of a uniform distribution (Glorot & Bengio, 2010); ii) optimal calibration of the training algorithm, typically belonging to the *Stochastic Gradient Descent* (SGD) category for CNNs (Qian, 1999; Wijnhoven & De With, 2010); iii) input normalization (via *batch normalization*) to smooth the values of the objective function and its gradient derivative (Ioffe & Szegedy, 2015); iv) adoption of regularization tools such as dropout (Hinton *et al.,* 2012), i.e., random switch-off of links between neurons during training, and quadratic regularization of weights in convolutional layers (Yu *et al.,* 2008); v) re-initialization of the value of parameters when local optima are achieved, to improve exploration of the solution space (Treadgold & Gedeon, 1996; Guo & Li, 2006).

3.3 Convolutional neural networks for the study of banknotes

CNNs can improve banknote image analysis mainly due to two factors: the flexibility of the models and their ability to process data of digital image without resorting to intermediate tools to extract characteristic information, or *features* (Lee *et al.,* 2017).

In the literature addressing processing of banknote images, applications typically make use of statistical-mathematical models and machine learning techniques to recognize denomination (Grijalva *et al.,* 2010; Sharma *et al.,* 2012; García-Lamont *et al.,* 2013), serial number (Feng et al, 2014; Liu *et al.,* 2010; Wenhong *et al.,* 2010; Hasanuzzaman *et al.,* 2011), currency (Manikandan & Sumithra, 2015), wearing (Sun & Li, 2008; Daraee & Mozaffari, 2010; Mousavi *et al.,* 2015) or for the detection of counterfeit specimens (Darade *et al.,* 2016; Suresh *et al.,* 2016). Advanced analytical techniques are also employed for counting, operational support to visually impaired individuals, and quality monitoring of circulating banknotes among individuals, merchants and banks themselves.

Recent literature also highlights publications using CNNs for banknote recognition (Jadhav *et al.,* 2019; Zhang *et al.,* 2019; Jang *et al.,* 2020; Park *et al.,* 2020; Veeramsetty *et al.,* 2020) and counterfeit specimen detection (Kamble *et al.,* 2019; Sawant *et al.,* 2022), achieving an overall increase in performance. In order to overcome the challenges associated with the use of CNNs, such as the high

volume of observations requested for training, and the tendency of CNNs to overfitting given the variability characteristics associated with banknotes, CNNs have in some cases been developed according to *adversarial learning* approaches for generating synthetic images (Ali *et al,* 2019; Desai et al., 2021; Khemiri *et al.,* 2022) or *transfer learning*, i.e., approaches that make use of already trained models and then operate their specialization, or fine-tuning, through a less intensive training process (Laavanya & Vijayaraghavan, 2019; Linkon *et al.,* 2020; Aseffa *et al.,* 2022).

Applications referring to banknote production can also be found in literature (Ke *et al.,* 2016; Pham *et al.,* 2017). In particular, research by Ke and co-authors focuses on the detection of bite-like defects - the same class of defects used in this work - by developing a CNN. Architectural details of the network and information on the training set, however, are not fully exposed by the authors.

The review of methodologies proposed in the literature for image processing of banknotes through deep learning techniques underscored that, for the purposes of our application, training a sufficiently deep CNN would have a number of limitations, given by the aforementioned variability in the arrangement of graphical elements, defects on the image surface, and the limited availability of observations for training. To mitigate the relatively limited availability of observations, a preliminary extension of the training set was considered, by resorting to *data augmentation* solutions (Wang & Perez, 2017). Extensions considered in such preliminary stage included altering existing images through rotation, resampling, and transposition techniques, as well as generating synthetic observations obtained by perturbing available images or training a support model for image generation (Goodfellow *et al.,* 2020).

3.4 One-shot learning and Siamese neural networks

Preliminary analysis suggested that data augmentation, in our case, is subject to the trade-off between the instability of the training process due to limited volumes of observations, and potential bias associated with synthetic data augmentation. In particular, data augmentation tends to expose the neural network to two types of issues. On one hand, additional images that over-represent the occurrence of printing defects with certain characteristics make the model exposed to overfitting those materialization of the defects only: this would undermine the model's ability to detect bite defects exhibiting different characteristics from those observed for training. On the other hand, the generation of novel synthetic defects could introduce relevant forms of algorithmic bias, in case these were unobservable in reality but relevant to steer the model's reasoning.

To curb such limitations, we resorted to CNN architectures following the *one-shot learning* approach (Fei *et al.,* 2006).

One-shot learning represents a specialization of statistical learning, aimed to emulate the ability of the human mind to associate entities unknown to the subject with known ones, based on similarity criteria. The approach thus aims to reproduce the ability of natural intelligence to make visual comparisons based on the knowledge of a ground truth conferred by experience. Given an image of a

generic entity - such as an object, a human face, or a landscape - and a range of possible categories of membership, the assignment of the image to one of these categories occurs from the simultaneous identification of points of commonality and points of difference with representative images of each class.

In its most rigorous formulation, one-shot learning requires that the training set consists of a number of prototypical observations equal to the number of categories of interest. However, the literature generally refers to one-shot learning in its more extended meaning[7] referred to as *k-shot*, in which a (limited) number of examples per class are considered, then differentiating it from *zero-shot* learning (Palatucci *et al.,* 2009) whereby no example image per class is available during training[8].

For one-shot learning, we make use of so-called *Siamese* neural networks (Koch *et al.,* 2015). Siamese neural networks exhibit two parallel sub-architectures (i.e., branches) identical to each other in terms of structure and parameter values (such as in green in Figure 1). Each branch of a Siamese network processes the respective element of the pair of observations that constitutes its input: a "verification" observation, that needs potential attribution to a given class, and a "support" observation, representing the class. The network maps the pair of observations toward "signals", synthetic and separable representations of the input. The intuition is as follows: two observations belonging to the same category are projected by the parallel architectures onto signals that are similar to each other; conversely, observations belonging to different categories will be characterized by orthogonal signals. The verification and support signals, respectively. $\boldsymbol{y}^{Sig,v}, \boldsymbol{y}^{Sig,s} \in \mathbb{R}^m$, are combined and processed by a "comparison" layer (Figure 1, in orange), which quantifies their similarity by the representation $\boldsymbol{y}^{L'} \in \mathbb{R}^{n_{L'}}$, obtained from the transformation:

$$\boldsymbol{y}^{L'} = g^{Sim}\left(\boldsymbol{w}^{Sim} Sim\left(\boldsymbol{y}^{Sig,v}, \boldsymbol{y}^{Sig,s}\right) + \boldsymbol{b}^{Sim}\right)$$

where g^{Sim} and *Sim* are an activation function and a similarity function, respectively, $\boldsymbol{w}^{Sim} \in \mathbb{R}^{n_{L'} \times m}$ is the matrix of weights and $\boldsymbol{b}^{Sim} \in \mathbb{R}^{n_{L'}}$ is the vector of bias terms. In the general case $L' \leq (L+1)$, that is an additional sequence of representation layers could be attached from the comparison layer. In the simplest case, considered by our application, $L' = L + 1$, i.e., the representation obtained from the output of the comparison layer is definitely the output of the Siamese network, such that, in our binary case study, $n_{L'} = 1$ and g^{Sim} is the sigmoid function that projects $\boldsymbol{y}^{L+1} = \boldsymbol{y}^{L'}$ in the interval [0,1]. In the training phase, let N be a pairs of images in the training set; each pair is labeled as containing images - one as verification and one as support - belonging to the same class ($c_i = 1$) or to distinct classes ($c_i = 0$). The values

[7] *One-shot* learning should be distinguished from "*one-shot* transfer" which has to be intended as a specialization of *transfer learning*.

[8] In the literature, there are numerous examples of the application of neural networks based on *zero-shot* learning, as well as in the area of natural language processing. Among the most frequent applications are classification exercises under conditions of extreme class imbalance (Ochal *et al.,* 2021), as well as face recognition tasks, which are required to guarantee a certain level of performance under conditions of high variability, for example in facial expression, brightness or background.

$y_1^{L+1}, \dots, y_N^{L+1}$ produced by the Siamese network therefore correspond to the probabilities that the *N* verification images belong to the same class as their respective support images. Training the model thus consists of minimizing the following *loss function*:

$$FP(\boldsymbol{y}^{L+1}) = -\frac{1}{N}\sum_{i=1,\dots,N} c_i \log\left(y_i^{L+1}\right) + (1-c_i)\log\left(1-y_i^{L+1}\right)$$

where c_i takes as said value 1 when the verification and supporting observations belong to the same class, 0 otherwise, and $y_i^{L+1} = Prob(c_i = 1), i = 1, \dots, N$.

4. Siamese networks for bite-type defect detection

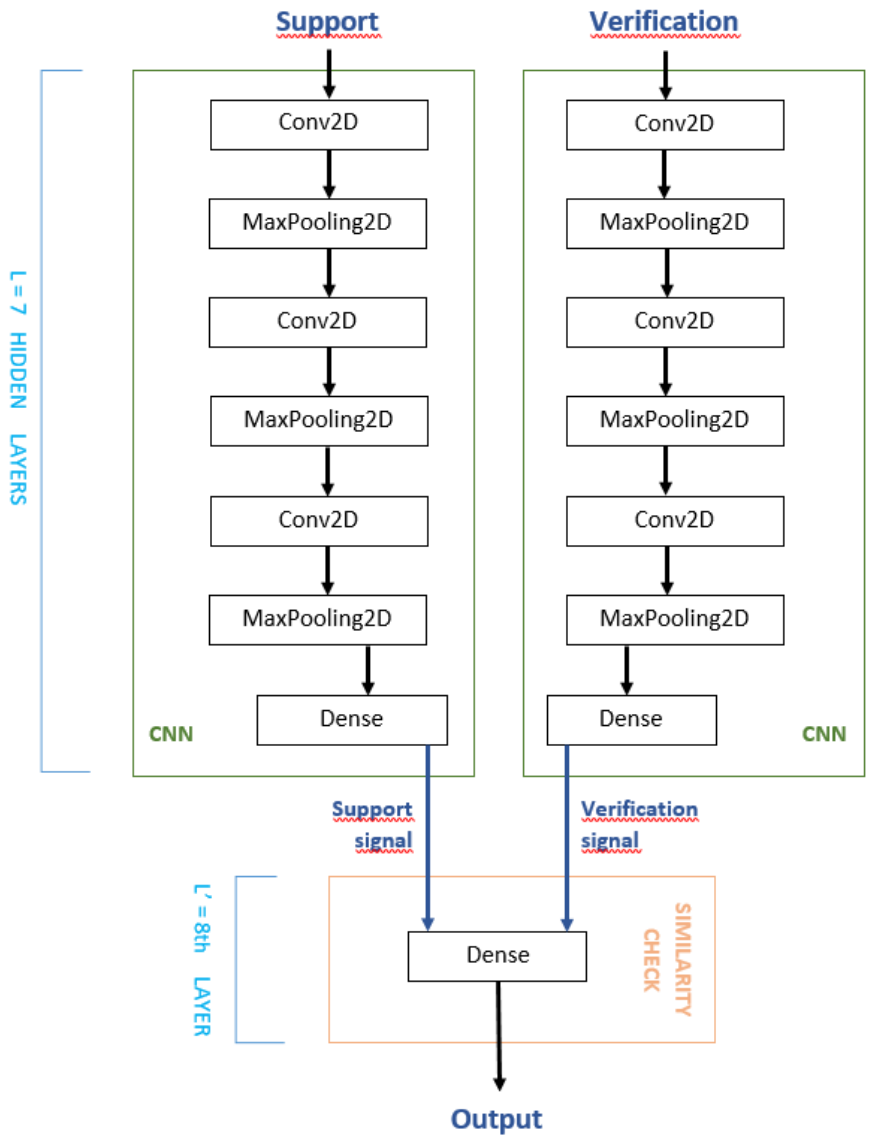
4.1 Model architecture and training

In the developing of a Siamese network for the detection of bite defects on the European flag graphical element, we resorted to an essential architecture consisting of a pair of twin networks as depicted in Figure 1. Each network is characterized by three convolutional layers (*Conv2D*) followed by three collection layers (*MaxPooling2D*), on which a sequence of fully connected layers (*Dense*) is grafted for signal vectors generation. These are combined according to the absolute distance metric and used to derive the output value $\boldsymbol{y}^{L+1}$, corresponding to the probability that the pair belongs to the same category[9].

[9] Additional variants were considered during the experimental phase. For brevity and little contribution in terms of relevance to the final considerations, the results are not reported.

Conceptual representation of the Siamese network architecture used in this work.

Figure 1



To facilitate comparison between results, the architecture of the networks is unchanged in the main blocks, while the number of processing units, i.e. nodes of each layer, depends on the resolution considered. Specifically, after of the manual acquisition of the banknotes high-resolution images (~2656x1467 pixels, 24-bit color, RGB model), the following four sets of images were generated by re-sampling, each of which was used in separate sessions for training and testing of the Siamese network in order to compare the performance obtained in the four cases:

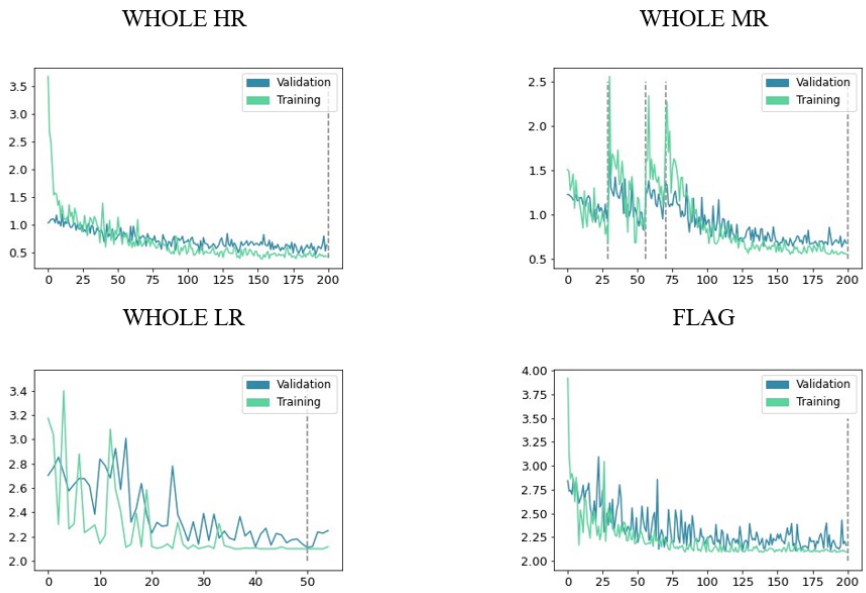
- Whole HR: whole banknote, high resolution (896x528 pixels);
- Whole MR: whole banknote, medium resolution (448x264 pixels);
- Whole LR: whole banknote, low resolution (299x176 pixels);
- Flag: cropping of the banknote's flag[10] in high resolution (224x352 pixels).

The training of the Siamese networks is done as mentioned through the Stochastic Gradient Descent (SGD) algorithm, starting from a set of parameters initialized according to the Xavier method. In order to promote optimal convergence of the process, batch normalization and dropout are also used, as well as re-initialization of the value of the parameters when a local optimum is reached, or the training set is overfitted (Figure 2). Hyper-parameters such as *Leaky ReLu* activation functions (Maas *et al.,* 2013) are also defined using a set of calibration images excluded from the training process.

[10] See Jang *et al.* (2020).

Siamese neural network training process representation for the four different input resolution considered: Whole HR, Whole MR, Whole LR and Flag. The green (blue) curves represent the value of the loss function during the iterations of the SGD optimization routine in the training (validation) set images. The dashed lines indicate the re-initialization of the training process, corresponding to a local optimum convergence.

Figure 2



The high performance of machine learning algorithms often has to contend with reduced interpretability of the learned internal mechanisms, as well as of the explainability of the underlying logic. Useful techniques to support the understanding, and thus the control of the proper logic of a complex analytical model, pertain to the field of so-called *eXplainable Artificial Intelligence* (Molnar, 2020). In the case of Siamese networks, the graphical representation of signals allows visual comparison between the synthesis of the supporting and verification images for a qualitative assessment of the similarity of the vectors extracted from each twin network. However, the logic underlying the signal extraction process is concealed by the large number of parameters used for recursive projection of the information content of the input images, consistent with the poor interpretability of deep learning models (Zhang *et al.,* 2019).

Unlike other types of neural networks, in CNNs the transmission of input images from one convolutional layer to another allows intermediate tensors to be extracted, which can in turn be represented as images, called *attention masks*. In the literature, the use of attention masks is typically used during model training, and is aimed at exogenous control of the process by the analyst (Xu *et al.,* 2015). Intuitively, the weight tensor of the l-th convolutional layer $\boldsymbol{w}^l$ is combined

element by element with an arbitrary mask $\boldsymbol{a}$ to emphasize or not emphasize specific areas of the image, and transformed as seen above according to the activation function g^l. The resulting tensor, $\boldsymbol{w}^{l,a}$, is thus characterized by elements such that higher values will correspond to greater network attention to a particular area of the image and vice versa.

For the purposes of our application, attention mechanisms are divided into *hard* and *soft*. Hard attention actually corresponds to random or deterministic exclusion of part of the surface (*cropping*). In this case $\boldsymbol{a}$ takes binary values: 0 or 1. Hard attention mechanisms can be used to exogenously direct the learning process or to maximize the exploration of the tensor surface and define optimal attention trajectories by the model (Ba *et al.,* 2014). In its simplest form, soft attention instead resorts to an activation function that compresses the values of the weights into a given range (Xu *et al.,* 2015), such as in the range [0,1] in the case of the *softmax* function. Built-in soft attention mechanisms are used to ensure overall control of the learning process[11] (Gregor *et al,* 2015; Kosiorek *et al.,* 2017).

This application considers an alternative use of soft attention masks according to a simplified approach geared toward model explicability. Attention masks are locally activated, after the training, to extract partial images along the twin CNNs, representing the logic adopted by the model along the processing steps. Specifically, for each convolutional layer the corresponding weight tensor is projected into the unit interval [0,1] according to the softmax activation function, which highlights the relative weight of each element. The weights thus transformed are then combined with the input of the convolutional layer. This corresponds to superimposing a soft attention mask on the banknote image. The intuition is as follows: the l-th layer uses weights whose relative values in the softmask are close to 1 at regions considered relevant for signal extraction, and close to 0 vice versa (see Figure 4).

4.2 Results

Having completed the learning phase, the Siamese network classifies each new image using the support image sets whose membership in one of the two categories of bite and no-bite is known by construction. Assignment of a new image to one of the two classes is then made from a comparison of the image and the two support sets.

The selection of the support images to be used for comparison is a critical step in the use of Siamese networks because classification is done from pairs of observations. The choice on the method for the support image selection certainly has an impact on the overall performance of the network; to analyze this impact, a hold-out image set, i.e., a portion of the images in the verification set, was selected, and initially separated from the training set, and finally used for the sole

[11] Specifically, at a given output, such as the classification of an element on the image surface, the model is asked to identify the reference region by operating a bias in the value of the corresponding parameters.

evaluation of the support image selection method to ensure the quality of the analysis. The following approaches are then considered:

1. RND-1: Each verification image is compared with one randomly selected support image per class;

2. SSIM-0: Each verification image is compared with a specific single support image per class: the support selected image is the one with the maximum similarity with the verification image among those available in that particular class, according to the Structure Similarity Index Measure (SSIM)[12];

3. SSIM-1: Each verification image is compared with the second most similar of the available support images candidates for each class. In the SSIM-1 approach, the exclusion of the most similar support image is intended to reduce the impact of not pertinent but similar features that might be misleading for the classification exercise;

4. SSIM-k: Each verification image is compared with k suppor images per class, selected according to similarity (k=10 in the following analysis).

Table 1 shows the results obtained using the four different resolutions described above, and using the four different selection methods for the support images. The test set contains 46 images, including 14 with bite defect of different severity[13]. The performance on the test set is measured from the Recall rate[14], i.e., the rate of images with correctly classified bite defect, and the F1-score, the harmonic mean of the *Recall* and *Precision* score, defined as the rate of banknotes correctly classified as not defective. In our application, false positives, i.e., images of banknotes incorrectly identified as defective, and defect under-reporting, or false negatives, are considered equally costly.

[12] SSIM is a widely used measure for image comparison and image quality assessment, introduced by Wang *et al.* (2004). It is derived from the composition of three comparison runs related to image brightness, contrast and texture.

[13] The classification is based on the documentation provided by the European Central Bank.

[14] Given the number of banknotes classified as defective by the network (*all positives*), *Recall* is defined by the proportion of banknotes correctly classified (*true positives*): *Recall = (true positives)/(all positives)*.

Results for the binary image classification exercise in the test set, for the high (HR), medium (MR) and low (LR) resolution whole banknote cases, and for the European flag crop case. Performance below the threshold value 0.5 is indicated by "-".

Table 1

	Recall				F1-Score			
	RND-1	**SSIM-0**	**SSIM-1**	**SSIM-k**	**SSIM-k**	**SSIM-0**	**SSIM-1**	**SSIM-k**
Whole HR	-	.75	.62	-	-	.67	.55	-
Whole MR	1.	**.87**	**.87**	.5	.84	.82	**.93**	.67
Whole LR	-	.62	.62	-	-	.55	.62	-
Flag	-	-	.5	.98	-	-	**-**	-

We observe that, for the verification set considered, the use of the medium-resolution whole image (Whole MR) allows for greater accuracy in classifying observations into bite and no-bite. In particular, the SSIM-1 selection method appears to perform better in assigning the banknote to the correct category.

Siamese networks are characterized by their ability to identify critical features during the projection of the input into the signal, thus the assignment of the verification image to its correct class, under conditions of high variability. When this variability is quite excessive, as in the case of high resolution images (Whole HR) the granularity in terms of pixels becomes misleading to separate signals obtained from the verification and the support images. This can also be seen from the visual inspection of the signals extracted from the Siamese network, shown in the top left panel of Figure 3, where it can be seen that signals are more similar to each other when compared with the medium-resolution case (top right). Similarly, by excessively reducing the image resolution, the absolute size of the bite defect, the relative size of the field corresponding to the flag, and the overall variability turn out to be attenuated to a point that the Siamese network cannot learn a reliable criterion for their recognition (bottom left). In the case of cropped images containing the flag element only, the combination of the reduction in overall (second-level) variability and the increased relative weight of the region of pixels corresponding to the defect, which is characterized by high (first-level) variability, make a one-shot learning approach inappropriate. Indeed, the

performance measured in quantitative terms, and the graphical inspection of the obtained signals (bottom right) reflect this inadequacy.

Comparison of signals for verification and support images. The images show the signals obtained for pairs of images of type Whole HR, Whole MR, Whole LR, and Flag. Support images were selected using the SSIM-1 method. For each box, the verification signal is obtained from an image with a bite defect, to be compared with the bite and no-bite type support images, respectively. Figure 3

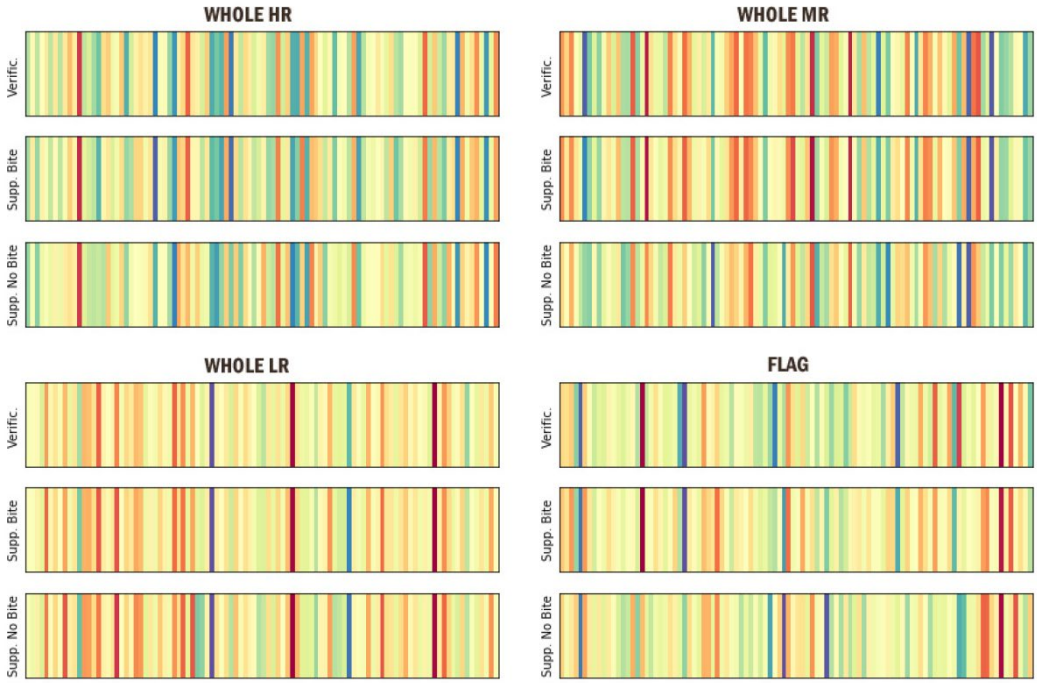
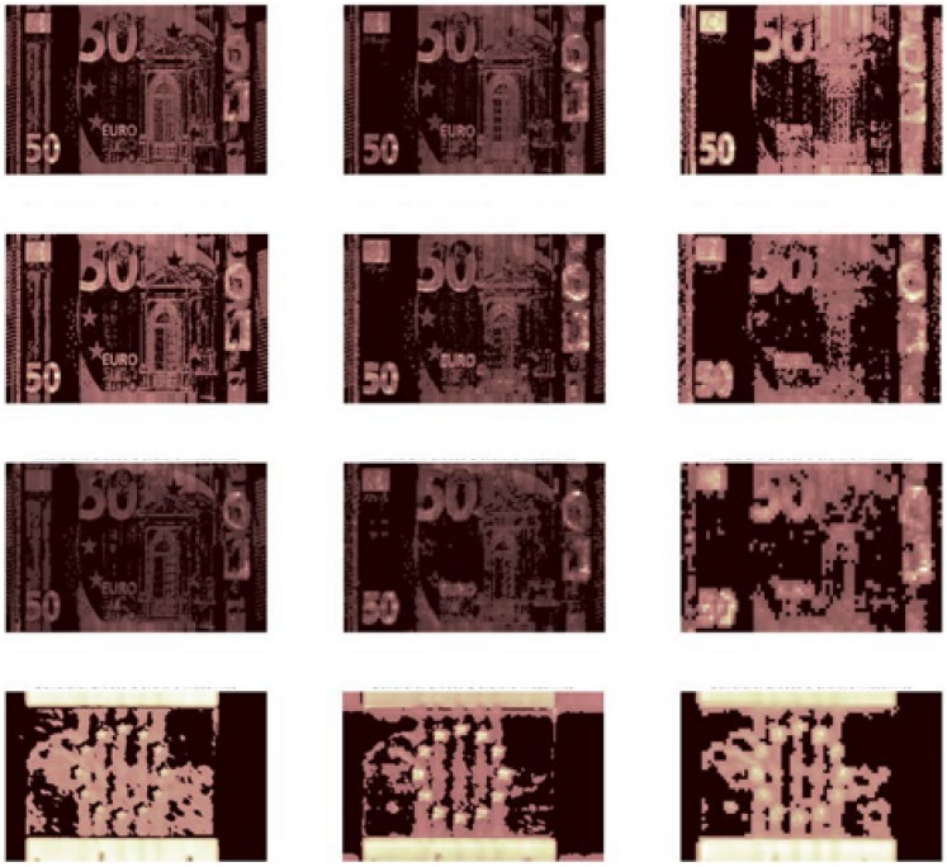


Figure 4 illustrates an example of the attention mechanism operated by the convolutional components of a Siamese networks, obtained for the same bite defect image, and in the four cases Whole HR, MR, LR, and Flag crop. In the different images, brighter shades correspond to a greater attention, or concentration, conferred by the neural network in the three convolutional layers inside the CNN. It can be seen that, in the absence of indications given to the network during its construction, both the architecture for the analysis of high-resolution images and that for medium-resolution images are able to independently identify the area of the critical surface for our purposes of bite defect detection, and subsequent assignment to one class or the other. This virtuous process does not seem to be triggered, except approximately, in the case of the Siamese network calibrated from the low-resolution training set. Similarly, the same behaviour is produced in the cropped flag case: the Siamese network

seems to "focus" on the entire image area, thus making the process of synthesizing the input information overly susceptible to minor variations.

Graphical representation of the attention mechanism operated by the three blocks of the convolutional component of the Siamese networks for a training image with bite defect, in the formats Whole HR, Whole MR, Whole LR, and Flag. Figure 4



5. Conclusions and possible extensions

This paper reports the main results obtained from the exploratory study on the use of neural networks for the autonomous recognition and classification of banknote defects in the context of quality control. The complexity of the problem, mainly due to the high degree of natural product variability and the variability of defect types and locations, imposed a very stringent approach in the selection of experimental scenarios. We restricted the analysis to a specific type of defects

(bites), located on a predetermined portion (Flag) of a single denomination of banknotes (€50 of the second series).

The results support the relevance of these tools for identifying print defects. In particular, the case of medium-resolution whole banknote (Whole MR) allows balancing the needs related to image acquisition, computational efficiency, and quality of model performance. Further work needs to be carried out in order to attest their actual potential as an alternative, or cooperative, use along with current non-automated qualitative techniques. Further study and experimentation activities must extend the analysis to categories of defects other than bite, making use of useful techniques to adapt a previously trained model to a new task (so-called *transfer learning*), while maintaining a binary classification output. Finally, the implementation of a system for the classification of multiple print defects could make use of so-called *ensemble learning* techniques to combine and aggregate the results produced by a multi-network architecture.

However, any extensions considered would be circumscribed by an a-priori choice that keeps their domain of analysis and research within one-shot learning algorithms, and specifically using Siamese networks. This choice is also suggested by the literature produced by the scientific community for problems variously related to what is presented in these pages.

References

Ali, T., Jan, S., Alkhodre, A., Nauman, M., Amin, M., & Siddiqui, M.S. (2019). DeepMoney: counterfeit money detection using generative adversarial networks. *PeerJ Computer Science*, *5*, e216.

Aseffa, D.T., Kalla, H., & Mishra, S. (2022). Ethiopian banknote recognition using convolutional neural network and its prototype development using embedded platform. *Journal of Sensors*, *2022*, 1-18.

Ba, J.L., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention.

Boureau, Y.L., Ponce, J., & LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 111-118).

Darade, S.R., & Gidveer, G.R. (2016). Automatic recognition of fake Indian currency note. In *2016 international conference on Electrical Power and Energy Systems (ICEPES)* (pp. 290-294). IEEE.

Daraee, F., & Mozaffari, S. (2010). Eroded money notes recognition using wavelet transform. In *2010 6th Iranian Conference on Machine Vision and Image Processing* (pp. 1-5). IEEE.

Desai, S., Rajadhyaksha, A., Shetty, A., & Gharat, S. (2021). CNN based counterfeit Indian currency recognition using generative adversarial network. In *2021*

International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 626-631). IEEE.

Domínguez, F.R. (2020). Supermasks and a Good Initialization Are All You Need (*Doctoral dissertation, Pontificia Universidad Catolica de Chile (Chile)*).

Fei-Fei, Li, Rob Fergus e Pietro Perona (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence 28.4*: 594-611.

Feng, B.Y., Ren, M., Zhang, X.Y., & Suen, C.Y. (2014). Part-based high accuracy recognition of serial numbers in bank notes. In *Artificial Neural Networks in Pattern Recognition: 6th IAPR TC 3 International Workshop, ANNPR 2014, Montreal, QC, Canada, October 6-8, 2014. Proceedings 6* (pp. 204-215). Springer International Publishing.

García-Lamont, F., Cervantes, J., López, A., & Rodríguez, L. (2013). Classification of Mexican paper currency denomination by extracting their discriminative colors. In *Advances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II 12* (pp. 403-412). Springer Berlin Heidelberg.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256). JMLR Workshop and Conference Proceedings.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT press*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *International conference on machine learning* (pp. 1462-1471). PMLR.

Grijalva, F., Rodriguez, J.C., Larco, J., & Orozco, L. (2010). Smartphone recognition of the US banknotes' denomination, for visually impaired people. In *2010 IEEE ANDESCON* (pp. 1-6). IEEE.

Guo, Z.H., & Li, S.H. (2006). Feed-Forward neural network using SARPROP algorithm and its application in radar target recognition. In *Advances in Neural Networks-ISNN 2006: Third International Symposium on Neural Networks, Chengdu, China, May 28-June 1, 2006, Proceedings, Part II 3* (pp. 369-374). Springer Berlin Heidelberg.

Hasanuzzaman, F.M., Yang, X., & Tian, Y. (2011). Robust and effective component-based banknote recognition by SURF features. In *2011 20th Annual Wireless and Optical Communications Conference (WOCC)* (pp. 1-6). IEEE.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R.R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). pmlr.

Jadhav, M., kumar Sharma, Y., & Bhandari, G.M. (2019). Currency identification and forged banknote detection using deep learning. In *2019 International Conference on Innovative Trends and Advances in Engineering and Technology (ICITAET)* (pp. 178-183). IEEE.

Jang, U., Suh, K.H., & Lee, E.C. (2020). Low-quality banknote serial number recognition based on deep neural network. *Journal of Information Processing Systems*, *16*(1), 224-237.

Kamble, K., Bhansali, A., Satalgaonkar, P., & Alagundgi, S. (2019). Counterfeit currency detection using deep convolutional neural network. In *2019 IEEE Pune Section International Conference (PuneCon)* (pp. 1-4). IEEE.

Ke, W., Huiqin, W., Yue, S., Li, M., & Fengyan, Q. (2016). Banknote image defect recognition method based on convolution neural network. *International Journal of Security and Its Applications*, *10*(6), 269-280.

Khemiri, W., Tarifa, A., Jaafar, W., & Abderrazak, J.B. (2022). Towards a Hybrid Variant of GAN For Counterfeit Money Detection SSQUGAN: A Semi-Supervised Quadrupled GAN.

Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop* (Vol. 2, No. 1).

Kosiorek, A., Bewley, A., & Posner, I. (2017). Hierarchical attentive recurrent tracking. *Advances in neural information processing systems*, *30*.

Krizhevsky, A., Sutskever, I., & Hinton, G.E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.

Laavanya, M., & Vijayaraghavan, V. (2019). Real time fake currency note detection using deep learning. *Int. J. Eng. Adv. Technol.(IJEAT)*, *9*.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

Lee, J. W., Hong, H.G., Kim, K.W., & Park, K.R. (2017). A survey on banknote recognition methods by various sensors. *Sensors*, *17*(2), 313.

Linkon, A.H.M., Labib, M.M., Bappy, F.H., Sarker, S., Jannat, M.E., & Islam, M.S. (2020). Deep Learning Approach Combining Lightweight CNN Architecture with Transfer Learning: An Automatic Approach for the Detection and Recognition of Bangladeshi Banknotes. In *2020 11th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 214-217). IEEE.

Liu, L., Ye, Y.T., Xie, Y., & Pu, L. (2010). Serial number extracting and recognizing applied in paper currency sorting system based on RBF Network. In *2010 international conference on computational intelligence and software engineering* (pp. 1-4). IEEE.

Maas, A.L., Hannun, A.Y., & Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml* (Vol. 30, No. 1, p. 3).

Manikandan, K., & Sumithra, T. (2015). Currency recognition in mobile application for visually challenged. *Discovery*, *30*, 245-248.

Molnar, C. (2020). Interpretable machine learning. *Lulu.com*.

Mousavi, S.A., Meghdadi, M., Hanifeloo, Z., Sumari, P., & Arshad, M.R.M. (2015). Old and worn banknote detection using sparse representation and neural networks. *Indian J. Sci. Technol*, *8*, 913-918.

Ochal, M., Patacchiola, M., Storkey, A., Vazquez, J., & Wang, S. (2021). Few-shot learning with class imbalance. *arXiv preprint arXiv:2101.02523*.

Palatucci, M., Pomerleau, D., Hinton, G.E., & Mitchell, T.M. (2009). Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, *22*.

Park, C., Cho, S.W., Baek, N.R., Choi, J., & Park, K.R. (2020). Deep feature-based three-stage detection of banknotes and coins for assisting visually impaired people. *IEEE Access*, *8*, 184598-184613.

Pham, T.D., Lee, D.E., & Park, K.R. (2017). Multi-national banknote classification based on visible-light line sensor and convolutional neural network. *Sensors*, *17*(7), 1595.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, *12*(1), 145-151.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

Sawant, V.M., Tupe, R. D., Tawade, A.S., & Bhalerao, S.M. (2022). Fake Currency Identification System Using CNN. *International Journal of Wireless Network Security*, *8*(1), 18-23.

Sharma, B., & Kaur, A. (2012). Recognition of Indian paper currency based on LBP. *International Journal of Computer Applications*, *59*(1).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, B., & Li, J. (2008). The recognition of new and old banknotes based on SVM. In *2008 Second International Symposium on Intelligent Information Technology Application* (Vol. 2, pp. 95-98). IEEE.

Suresh, I.A., & Narwade, P.P. (2016). Indian currency recognition and verification using image processing. *International Research Journal of Engineering and Technology (IRJET)*, *3*(6), 87-91.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).

Treadgold, N.K., & Gedeon, T.D. (1996). The SARPROP algorithm, a simulated annealing enhancement to resilient back propagation. In *Proceedings International Panel Conference on Soft and Intelligent Computing* (pp. 293-298).

Veeramsetty, V., Singal, G., & Badal, T. (2020). Coinnet: platform independent application to recognize Indian currency notes using deep learning techniques. *Multimedia Tools and Applications*, *79*(31-32), 22569-22594.

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, *11*, 1-8.

Wenhong, L., Wenjuan, T., Xiyan, C., & Zhen, G. (2010). Application of support vector machine (SVM) on serial number identification of RMB. In *2010 8th World Congress on Intelligent Control and Automation* (pp. 6262-6266). IEEE.

Wijnhoven, R.G., & de With, P.H.N. (2010). Fast training of object detection using stochastic gradient descent. In *2010 20th International conference on pattern recognition* (pp. 424-427). IEEE.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057). PMLR.

Yan, Z., Zhang, H., Piramuthu, R., Jagadeesh, V., DeCoste, D., Di, W., & Yu, Y. (2015). HD-CNN: hierarchical deep convolutional neural networks for large scale visual

recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2740-2748).

Yu, K., Xu, W., & Gong, Y. (2008). Deep learning with kernel regularization for visual recognition. *Advances in neural information processing systems*, *21*.

Wang Z., Bovik, A.C., Sheikh H.R., Simoncelli, E.P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13.4*: 600-612.

Zeiler, M.D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13* (pp. 818-833). Springer International Publishing.

Zhang, Q., Yan, W.Q., & Kankanhalli, M. (2019). Overview of currency recognition using deep learning. *Journal of Banking and Financial Technology*, *3*, 59-69.

Siamese neural networks for detecting banknote printing defects

K. Boria, A. Luciani, S. *Marchetti* and M. Viticoli
(Bank of Italy)

3rd IFC Workshop on Data Science in Central Banking
Oct. 18, 2023

1 Introduction

2 Data & Methods

3 Empirical Results

4 Conclusions & Next Steps

Introduction

- Banknote production in the Eurosystem relies on strict quality controls throughout the process (ECB quality requirements, ISO standards);
- Similar to any manufacturing process, the printing of banknotes can give rise to various imperfections and defects;
- The number, type, and size of defects on banknotes are critical for the conformity of production batches (with related losses).

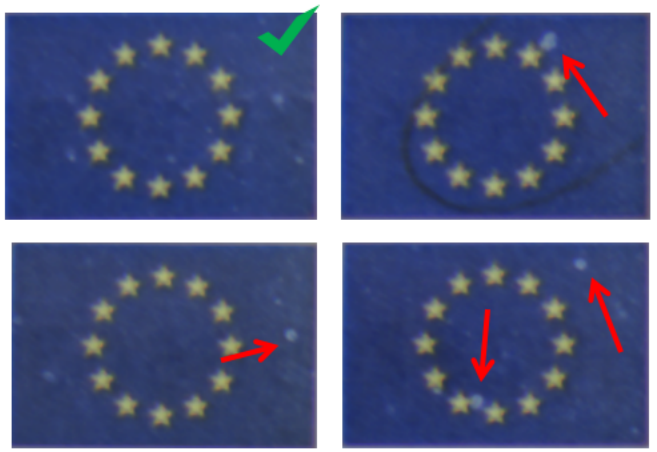
Introduction

- Banknotes exhibit variability in the position of elements, and in the shade and intensity of colors;
- While measurement of many parameters is carried out via automatic optical systems, the validation process cannot be fully automatized and requires highly trained staff (potential subjectivity factors);
- Artificial Intelligence (AI) systems are profitably used across industries to support and automatize quality control;
- Our work focuses on a denomination/defect pair to assess whether neural networks could improve process efficiency.

Data

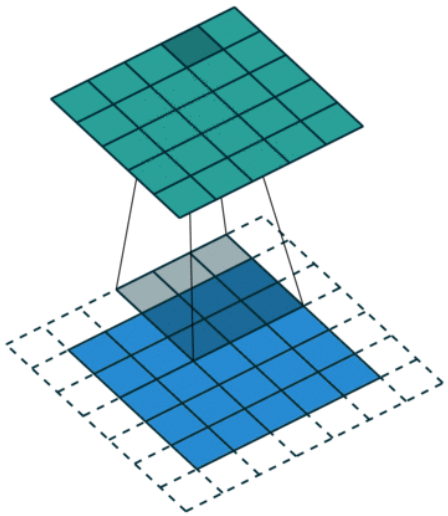
- Since acquisition of high-resolution images is costly:
  - Focus on €50 banknotes of the *Europa* series;
  - Focus on flag-*bite* defects: lack of ink on a homogenous background, with varying shape, size, and position.
- We manually acquire 24-bit color images (RGB model) of banknotes with 2,656 x 1,467 pixels resolution and annotate them as *fit*/*unfit*.





Neural Networks & Image Processing

- Previous research on banknotes analysis mostly resorts to convolutional neural networks (CNNs);
- Applications address recognition of denomination, serial number, currency, state of wear as well as counterfeit detection. On the production phase, we mention Pham et al., 2017; Ke et al., 2016.
- The inherent variability of banknotes requires training of complex CNN architectures, usually addressed via data augmentation.



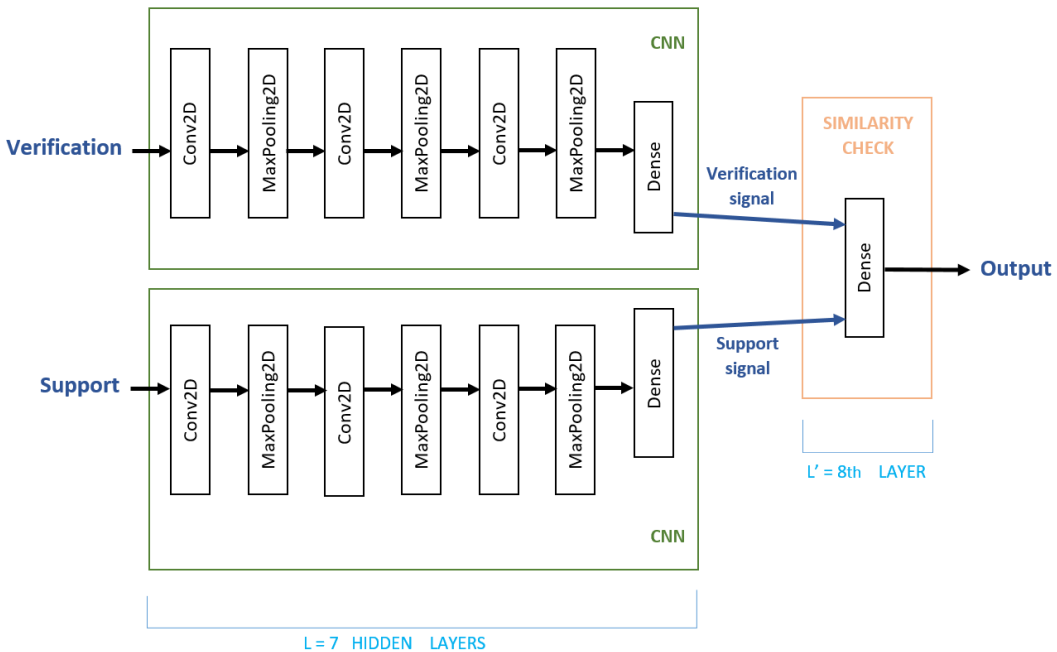
One-shot learning & Siamese Networks

- The traditional way: data augmentation
  - Manual: Overfitting of specific bite defects;
  - AI-enabled: Hallucination-prone behavior;
- The practical way: *one-shot learning* (Fei et al., 2006)
  - Emulates the ability of the human mind to associate entities based on similarity criteria.



Model development & Training

- We build a Siamese neural network architecture (Koch et al., 2015) for the bite-detection task:
  - Each twin branch extracts a signal from its input, either a *verification* or a *support* image;
  - The more similar the signals, the most likely images belong to the same class.

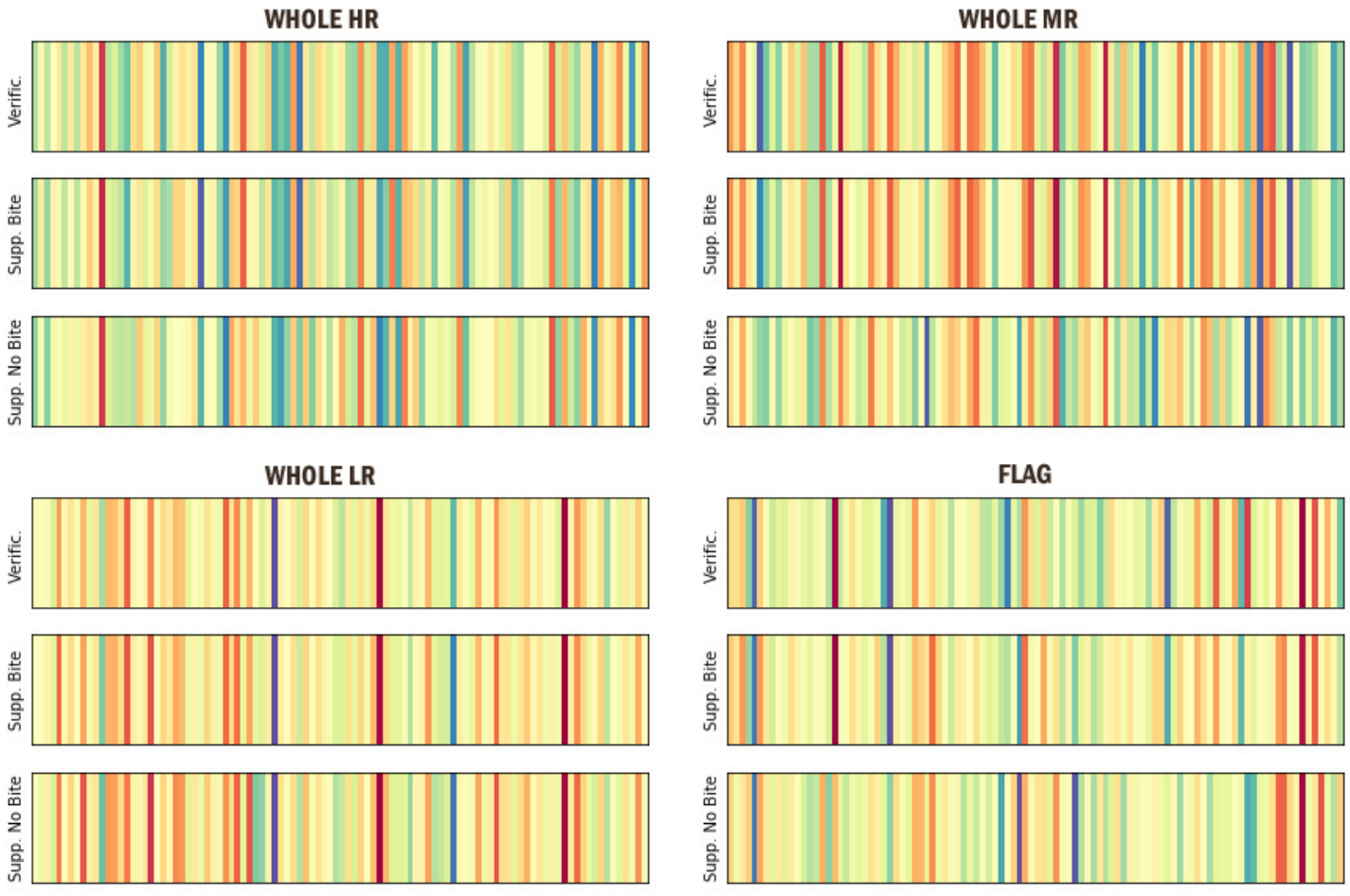


Model development & Training

We train our models on different resolutions of the input set: high, medium and low resolution + cropped flag detail:

- Training set: $N = 200$ images (100 bites, 100 fit; $C(N, 2)$ training pairs);
- We ensure convergence of the training process to global minima via drop-out, batch normalization, random re-initialization of weights;
- We control internal logics of networks via soft masks extracted from convolutional layers.

Empirical Results

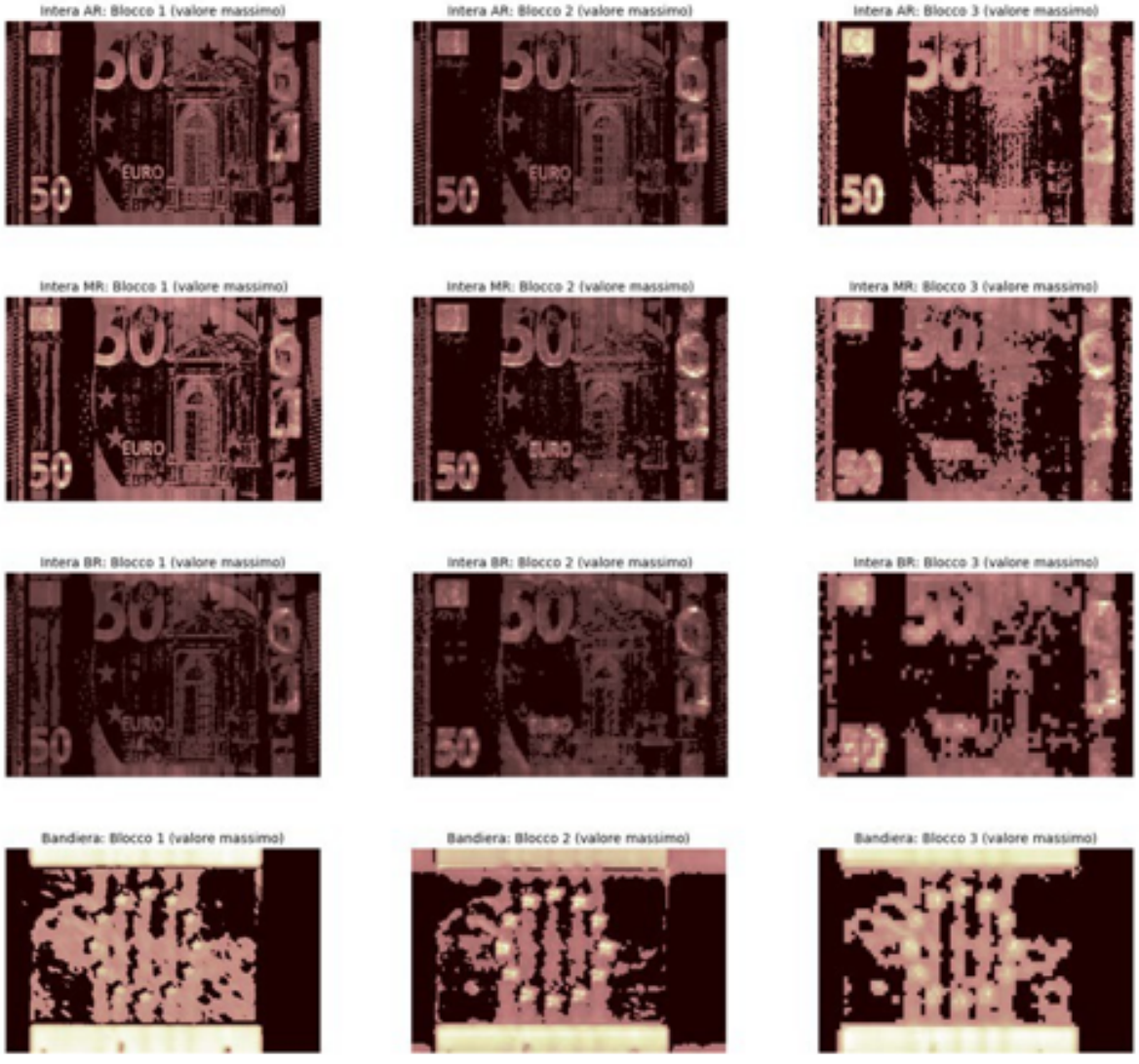


Empirical Results

	Recall				F1-Score			
	RND-1	SSIM-0	SSIM-1	SSIM-k	SSIM-k	SSIM-0	SSIM-1	SSIM-k
WHOLE HR	-	.75	.62	-	-	.67	.55	-
WHOLE MR	1.	.87	.87	.5	.84	.82	.93	.67
WHOLE LR	-	.62	.62	-	-	.55	.62	-
FLAG	-	-	.5	.98	-	-	-	-

Test set size: 46 images (14 bites; 32 fit);
RND-1: Support image is picked at random;
SSIM-j, j=0,1: the (j+1)th most similar image is picked as support;
SSIM-k: the k most similar images are picked as support, with k=10.

Empirical Results



Conclusions & Next Steps

- Our exploratory study provides insights on the detection of banknote defects for quality control via one-shot learning;
- We find medium resolution input allows for greater accuracy based on different metrics and evaluation criteria;
  - Trade-off between resolution and attention scattering due to variability patterns of banknotes;
- We are able to gain insights on the internal logics adopted by the models via soft masks;
- Future work will extend the analysis to additional defects (minor bite type defects already tested) and denominations, and to additional model architectures.

Thank you!

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Error spotting with gradient boosting: a machine learning-based application for central bank data quality[1]

Csaba Burger,
Magyar Nemzeti Bank,

Mihály Berndt,
Clarity Consulting

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Error spotting with gradient boosting

A machine learning-based application for central bank data quality

Csaba Burger, PhD, CFA, data science advisor at the Central Bank of Hungary (MNB)
Mihály Berndt, data scientist at Clarity Consulting

Abstract

Supervised machine learning methods, in which no error labels are present, are increasingly popular methods for identifying potential data errors. Such algorithms rely on the tenet of a 'ground truth' in the data, which in other words assumes correctness in the majority of the cases. Points deviating from the ground truth, outliers, are flagged as potential errors.

This paper implements an outlier-based error-spotting algorithm using gradient boosting, and presents a blueprint for the modelling pipeline. It uses a cross sectional view on the loan-to-value and its related columns of the Credit Registry (Hitelregiszter) of the Central Bank of Hungary (MNB), and introduces a set of synthetic error types. The paper shows that gradient boosting is not materially impacted by the choice of the imputation method, hence, replacement with a constant, the computationally most efficient, is recommended. Second, the Huber-loss function is better in capturing data errors. Finally, errors in the target variable are captured best, while errors in the predictors are hardly found at all. These empirical results may generalize to other cases, depending on data specificities.

Keywords: data quality, machine learning, gradient boosting, central banking, loss functions, missing values

JEL classification: C5, C81, E58

Acknowledgements: the authors would like to express their gratitude to Lívia Réka Ónozó for her insightful comments on a previous version of this paper. However, only the authors should be held responsible for any errors, omissions or opinions expressed herein.

Contents

1. Introduction 3

Supervised learning for data error spotting 4

The sparsity-aware logic of xgboost 5

The machine learning pipeline 7

Data and Methodology 9

- Data 9
- Modelling pipeline 9
- Bayesian optimization for hyperparameter-tuning 12
- Synthetic errors and model evaluation 13

Results 15

- The baseline model 15
- Hypotheses 1 and 2 18
- Testing Hypothesis 3 20

Conclusion 22

References 23

Appendix 25

1. Introduction

Granular central bank data, a relatively recent phenomenon of the last decade, has unique characteristics. First, it is usually delivered by a manageable number of data providers, which produce their data using their own scripts. This may introduce data errors on provider level. Second, such data sets may have several hundred columns, of which some are used only for a subset of its observations. To illustrate: collateral value is only relevant for loans with collateral, hence, it is missing for others. Third, relationships between data columns rely on economic processes, hence, the term 'outlier' or 'error' may win additional interpretations, should an observation not correspond to an expected relationship.

Once rule-based data errors are taken care of (format and constraint violations, values that violate syntactic and semantic constraints, duplicates (Rahm and Do, 2000, Kim et al., 2003), outlier detection algorithms may be applied. While the use of unsupervised outlier detection methods is more widespread, supervised learning algorithms "*empower learning methods with application-specific knowledge so as to obtain application-relevant anomalies*" (Aggarwal, 2017: 219). A specific subset of supervised methods used for data error identification is where there is no ex ante label for data errors. Particularly the 'attribute-wise learning for scoring outliers' (ALSO) (Paulheim and Meusel, 2015) has gained the attention of central bankers (Benatti, 2018), where a target variable of the granular data set is explained with the help of other variables, predictions are calculated, and the residuals of each observation serve as the starting point for outlier detection.

This paper is an empirical study of a gradient boosting error detection mechanism using synthetic errors we introduced to the data. It proposes a modelling pipeline to uncover potential data errors in a central bank granular data set. By doing so, it calculates how two evaluation metrics, calculated on the test dataset, change when certain modelling decisions are made. The metrics were: (1) the share of discovered errors out of all errors, and (2) the relationship between the share of intentional errors within outliers, compared to the share of intentional errors in the whole test data set. The modelling decisions were about the imputation of missing values, the choice of the xgboost loss function and the location of the error. The data used for illustration purposes is the Loan-to-value (LTV) relevant mortgage portfolio within the Credit Registry (*Hitelregiszter*) of the Central Bank of Hungary.

The findings underscore the fact that with xgboost, the using a constant value plus a flag, may be a decision as good as a prediction-based imputation to deal with missing values. Moreover, an xgboost with a Huber loss function seems capture more data errors than using the squared error loss function. Finally, the algorithm finds non-trivial errors in the target variable far better than errors in the explanatory values, hence, it may help in directing human attention towards the location of the actual error. We believe that our paper offers points to consider in future studies.

The paper is structured as follows. The next two sections contain the recommendations the literature holds on the xgboost-based outlier detection

pipeline. The fourth section discusses the data and methodology in detail, while the fifth presents the outcome of the error simulations. The final section concludes.

Supervised learning for data error spotting

Data error may refer to a set of issues, including rule-violations (e.g. a NOT NULL constraint), semantic pattern errors (such as a value outside 0 and 100, when it should be within), duplicates (best captured with record-linkage algorithms), or to outliers (Abedjan et al., 2016). The latter represents a subset, and have to fulfill two requirements. First, they are quantitative or qualitative deviations from an expected true value, and second, they should be confirmed to be errors. Outlier-based machine learning algorithms can do the former, whereas error-confirmation requires additional human intervention.

Hawkins (1980) defined outliers as observations which deviate from others "*as to arouse suspicions that it was generated by a different mechanism*". In that sense, it assumes the existence of a "ground truth", a relationship between variables which are expected to represent the baseline mechanism. For data quality purposes the 'ground truth' hypothesis assumes that most of the data on which the model is trained is right. Since it is hardly possible to limit the error to the test data only, the error will be present in the train data too. This means that any algorithm only flags outliers if they break the learned function, which is not necessarily the ground truth.

Supervised learning for outlier detection may not be the first choice. Unknown data errors are untagged, hence, the target variable cannot be a label stating if an observation contains an error or not. We often do not know a priori how errors look like, they may be heterogeneous and stem from a variety of sources (Heidari et al., 2019). Because of these, unsupervised algorithms (Aggarwal, 2017) are popular. Among these, distance-based methods identify 'densely' located records as normal observations, and rare instances as errors. Such an approach, however, faces problems with the increase in the number of dimensions, and may struggle with extreme values.

To overcome these problems, Paulheim and Meusel (2015) propose a novel way to use supervised learning for identifying outliers. They assume erroneous observations do not reflect the true relationship between the variables. They propose a method called 'attribute-wise learning for scoring outliers' (ALSO), in which a target variable is explained with several explanatory features. Subsequently, the residual, the difference between the prediction and the actual value, is calculated. The authors loop through all columns as target variable, and use the residual values for each instance to calculate a final outlier score[1] for each observation. Their approach translates the

[1] If i'_k is the predicted and i_k the actual value for an observation using the variable k as the target variable, its final, unweighted outlier score as proposed by Paulheim and Meusel (2015) is calculated as follows: $\sqrt{\sum_{k=1}^{n}(i_k - i'_k)^2}$. One may weigh the difference between i_k and i'_k to reflect the strength of each relationship using their relative squared error values. With the weights assigned, the weighted

data points into a new residual feature space. Within it, outliers can be considered as observations that are far away from the origin[2].

Benatti (2018) follows Paulheim and Meusel (2015) in that he also creates a model for each column using the remaining columns and calculates residuals for each instance. At the same time, he does not aggregate the obtained residuals into one single outlier score. Instead, he carries out a clustering of the residual values to identify groups of observations. According to his logic, observations far away from such clusters may be flagged as potential data errors.

The work of Benatti (2018) is relevant for further reasons. First, in his paper he looks for potential data errors in a granular central bank dataset, a similar challenge to ours. Second, his model was specified with a gradient boosting algorithm (xgboost), which provides sufficient flexibility to account for categorical and data provider-level specificities. It is capable of learning highly non-linear relationships and variable interactions, works well with high dimensional data, and it can deal with missing data within the modelling phase (Chen and Guestrin, 2016). In short, xgboost is an excellent tool to learn true, complex relationships within the data.

The sparsity-aware logic of xgboost

The xgboost algorithm entails a sequential development of several decision trees[3]. The first tree attempts to predict the target variable with a simple aggregation[4], the second tree predicts the error of the first. Then the predictions of these two trees are added and a new prediction error is calculated. The following tree is grown on the error of the previous trees, and its prediction is added to the last one. This process is repeated several times. The final number of trees can be determined by cross-validation or Bayesian Optimization, discussed later.

This section presents a simplified overview of the xgboost, based on Chen and Guestrin (2016). A tree ensemble model uses K additive functions to predict the output.

residuals in the k-th model are: $o(i) := \sqrt{\frac{1}{\sum_{k=1}^{n} w_k} \sum_{k=1}^{n} w_k (i_k - i'_k)^2}$, and the weights are calculated by $w_k := 1 - min(1,\ R_k)$, where $R_k := \sqrt{\frac{\sum_{j=1}^{m} \left((i_j)_k - (i'_j)_k\right)^2}{\sum_{j=1}^{m} \left((i_j)_k - \bar{k}\right)^2}}$.

[2] These are not distance-based outliers, but points which do not fit in the relationships between the features, which makes its use particularly appropriate for error spotting.

[3] In short, these are not the usual decision trees (sometimes called xgboost trees), and their purpose can be viewed as clustering the residuals by their directions. Residuals are the differences between the predictions and the actual values, and the direction is determined by whether the prediction is greater (overestimation) or smaller (underestimation) than the actual value. The algorithm clusters the observations along the values of the features according to the values of the residuals.

[4] The aggregation can be specified to be the mean (L2 case of the loss function), or the median (L1 case).

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{K}} \boldsymbol{f}_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{i}}), \;\; \boldsymbol{f}_{\boldsymbol{k}} \in \boldsymbol{\mathcal{F}}, \qquad (1)$$

where $\mathcal{F}$ is a set of regression trees, but the main difference compared to decision trees is that each regression tree contains a continuous score on each leaf node. Each f_k is an independent tree structure, the final prediction of a tree is the sum of the scores of the corresponding leaves. Since the "model includes functions as parameters, the model is trained in an additive manner". At iteration t the objective is the following:

$$\boldsymbol{\mathcal{L}}^{(\boldsymbol{t})} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \boldsymbol{l}(\boldsymbol{y}_{\boldsymbol{i}}, \hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\boldsymbol{t}-\boldsymbol{1})} + \boldsymbol{f}_{\boldsymbol{t}}(\boldsymbol{x}_{\boldsymbol{i}})) + \boldsymbol{\Omega}(\boldsymbol{f}_{\boldsymbol{t}}) \,. \qquad (2)$$

The authors did not use the cost function of the weak learner to fit the residual, but the second order approximation, which can be viewed as a cost/performance trade-off, and because of the Taylor's expansion it is a good approximation at the point we evaluate. The second order extension:

$$\tilde{\boldsymbol{\mathcal{L}}}^{(\boldsymbol{t})} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} \left[\boldsymbol{g}_{\boldsymbol{i}} \boldsymbol{f}_{\boldsymbol{t}}(\boldsymbol{x}_{\boldsymbol{i}}) + \frac{\boldsymbol{1}}{\boldsymbol{2}} \boldsymbol{h}_{\boldsymbol{i}} \boldsymbol{f}_{\boldsymbol{t}}^{\boldsymbol{2}}(\boldsymbol{x}_{\boldsymbol{i}}) \right] + \boldsymbol{\Omega}(\boldsymbol{f}_{\boldsymbol{t}}) \,, \qquad (3)$$

whereby $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are the first and second order gradient statistics on the loss function. At iteration *t* the goal is to achieve the greatest possible reduction of the loss. The authors measure the quality of a fixed tree structure *q*, by the following equation:

$$\tilde{\boldsymbol{\mathcal{L}}}^{(\boldsymbol{t})}(\boldsymbol{q}) = \frac{\boldsymbol{1}}{\boldsymbol{2}} \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{T}} \frac{(\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{i}}} \boldsymbol{g}_{\boldsymbol{i}})^{\boldsymbol{2}}}{\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{i}}} \boldsymbol{h}_{\boldsymbol{i}} + \boldsymbol{\lambda}} + \boldsymbol{\gamma T}. \qquad (4)$$

In an L_2 loss regression case the nominator is the sum of the residuals, squared, and the denominator is the number of residuals. It can be viewed as the algorithm's attempt to cluster the residuals by their direction and magnitude. It is impossible to enumerate all the possible tree structures *q*. Therefore, a greedy algorithm that starts from the single leaf containing every node (or possibly a subset controlled by a hyperparameter subsample) iteratively adds branches to the tree. The loss reduction after a split is calculated by the following equation:

$$\boldsymbol{\mathcal{L}}_{\boldsymbol{split}} = \frac{\boldsymbol{1}}{\boldsymbol{2}} \left[\frac{(\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{L}}} \boldsymbol{g}_{\boldsymbol{i}})^{\boldsymbol{2}}}{\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{L}}} \boldsymbol{h}_{\boldsymbol{i}} + \boldsymbol{\lambda}} + \frac{(\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{R}}} \boldsymbol{g}_{\boldsymbol{i}})^{\boldsymbol{2}}}{\sum_{\boldsymbol{i} \in \boldsymbol{I}_{\boldsymbol{R}}} \boldsymbol{h}_{\boldsymbol{i}} + \boldsymbol{\lambda}} - \frac{(\sum_{\boldsymbol{i} \in \boldsymbol{I}} \boldsymbol{g}_{\boldsymbol{i}})^{\boldsymbol{2}}}{\sum_{\boldsymbol{i} \in \boldsymbol{I}} \boldsymbol{h}_{\boldsymbol{i}} + \boldsymbol{\lambda}} \right] - \boldsymbol{\gamma}, \qquad (5)$$

where I_L and I_R are the instances of the left and right nodes after a split, and $I = I_L \cup I_R$. The gain after a split must be positive, otherwise the branch is removed. Hence, λ directly controls overfitting and reduce the prediction's sensitivity to individual observations, while γ post-prunes the regression trees.

An advantage of xgboost's sparsity-aware split finding is that observations with missing data are classified to the default direction at a split that is determined only

by non-missing data. There are two possible directions at each split for the candidates. The algorithm tries both directions, and determines the default direction by evaluating the splits by Equations (4) and (5).

The machine learning pipeline

In this paper, we discuss (1) missing value treatment options, the (2) use of various loss functions in xgboost, and (3) the identification of the erroneous variable, which helps in locating the data errors not just row-wise but column-wise too.

To deal with missing values among predictors, one rely on the sparsity-aware method of xgboost, which puts missing instances in the direction with the greatest loss reduction. Using this may reduce the bias on the train dataset, but may result in high prediction errors during the prediction phase, since the underlying 'ground truth' is not learnt by the model.

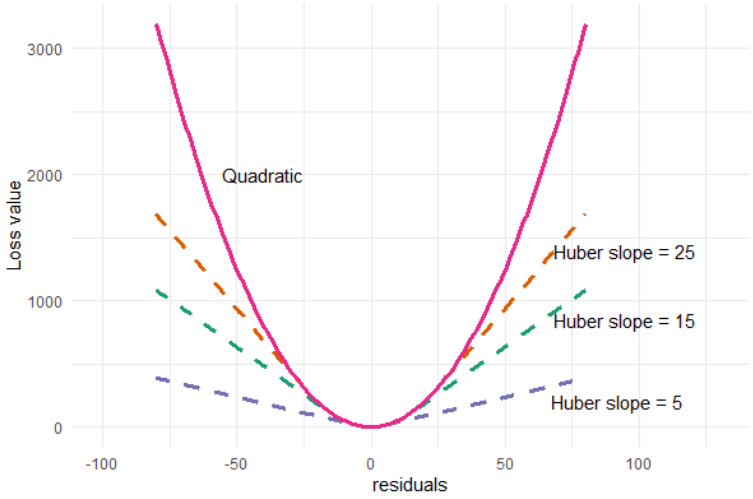
Alternatively, one may use explicit missing value replacement *before* modelling. Emmanuel et al. (2021) describe several estimator functions, stretching from the simple replacement with a constant, to a learner-function based prediction, using cases where values are filled. There is no single one best way to deal with missingness, particularly when it is an issue both during model training and during deployment (Khosravi et al. 2020). Despite that, Twala et al. (2008) do not find convincing difference in the performance of a classification tree when comparing a simple 'missing' flag with imputation. We also argue that a sufficiently flexible tree-based algorithm will not require a sophisticated missing value imputation. **We argue that replacement with a constant and the addition of an imputed value flag variable (1/0) is not inferior to an imputation algorithm. In addition, we recommend being cautious with the xgboost sparsity-aware model fitting, since it may not learn the underlying 'true' relationships within the data (hypothesis 1).**

The second step, referred to as 'dealing with high-leverage point outliers' is an expectation stemming from the data error identification approach. Our estimator function should not be materially impacted by the data errors we are about to identify, which may be achieved with a robust estimator function. Begashaw and Yohannes (2020) provide a solution in a linear regression setting; Sadouk et al. (2020) propose one for deep learning applications. With gradient boosting we propose changing the default objective function which is minimized during the model fitting: we argue for replacing the squared loss function with the Huber loss function. This latter makes sure that the residual (denoted as r) turns from quadratic (as in the case of squared loss) to linear above a certain threshold. This threshold is often referred to as the Huber-slope, denoted as δ, and its value is determined during the

hyperparameter-optimization process, discussed later. The simple[5] Huber-loss function is reflected by **Equation 6** and by **Figure 1**.

$$L_\delta(r) = \begin{cases} \frac{1}{2}r^2 & \text{for } |r| \le \delta \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \quad (6)$$

Figure 1. Loss values of a quadratic and of several Huber-loss functions



The Huber-function ensures that small deviations are squared, but large residuals count in a linear manner only. As a result, parameter tuning will not chase outliers and can learn the underlying ("true") function. **We argue that an xgboost with a Huber-loss function is better suited to uncover data errors than the squared loss objective function (Hypothesis 2).**

Finally, the outliers of one single model are investigated for being potential data errors. **We argue that errors are better identified when they are in the target variable as opposed to the explanatory variables (Hypothesis 3).** This hypothesis relies on the assumption that there is a certain degree of (linear or non-linear) relationship between explanatory variables. If a data error distorts an explanatory variable 'A', the other explanatory column 'B', which is not independent from 'A', may take over its role in explaining Y. As a result, A loses its importance in the final xgboost function, while B's feature importance increases. Because of that, the algorithm may not yield outliers on these errors within 'A', only if they are in the target. In sum, it may be a good practice to look for the data error within the target variable and not in the explanatory variables.

[5] The xgboost implementation uses the pseudo Huber-loss function, which is a smooth approximation of the Huber loss function, meaning continuous derivatives: $L_\delta(r) = \delta^2(\sqrt{1 + \left(r/\delta\right)^2} - 1$

Data and Methodology

Data

The Credit Register (Hitelregiszter) of the Central Bank of Hungary (MNB) is a monthly updated collection of all loan contracts issued in Hungary, containing 22 reporting tables, 482 attributes and 31 identifiers. These together usually yield between 30-60 million new lines each month. The dataset goes through of around two thousand logical quality rules. Aggregate time series are checked with the help of ARIMA- and further filters, whereas static changes are probed on record / cell level. Finally, a cross-sectional algorithm, the subject of this paper, is applied on the data.

This paper uses the loan-to-value (LTV) column and its related fields of the dataset. This limits the analysis to mortgages and home equity loans of the residential sector, issued after 1st October 2021, the date of a recent legal change. The final table is a combination several Hitreg tables, comprising data on the debtor(s), collateral(s) and on the loan contract. Altogether we work with 73 thousand lines of one single reporting date of 30th September 2022, using 274 columns. Metrics and the distribution of the target variable, and of a subset of the explanatory variables, are presented in the Appendix.

The target variable was the LTV-value at credit issuance, which is subject to strict legal limitations, hence its accuracy is of utmost interest. As a consequence, we only kept the columns which were relevant at the time of issuance (to illustrate: current outstanding loan amount was removed, since it has little relevance for the LTV at issuance).

In theory, calculating LTV is straightforward: the loan amount divided by the allocated collateral value. In practice, however, it is rather complex, since many decisions, including the choice of the collateral evaluation method and the allocation process of a collateral to the loan contract, are at the data providers' discretion. As a result, the columns at hand are insufficient to calculate LTV directly. The relationship between LTV and the other columns is complex, which makes xgboost an appropriate choice.

In short, LTV has highly non-linear relationship with the explanatory variables, and it does not contain missing values, as it is an enforced condition of a successful data submission to the central bank. As a result, it may serve as a prime candidate for hypothesis testing.

Modelling pipeline

Although xgboost predictions are not affected by multicollinearity, it may distort feature importance. Since we expect certain degree of interpretability, we kept one out of the pairwise collinear variables only (over 0.90). We also eliminated quasi constant values, where the frequency ratios of the most frequent and the second most

frequent values were extremely high (the variance of the feature close to zero), to avoid overfitting or misleading conclusions.

In a similar manner, we recoded categorical and discrete variables with rare values. Any category below a threshold of 2.5 percent was recoded into an 'other' category, and upper, rare values of discrete numerical values were summarized (e.g. the number of residential real estate collaterals above the value of three was recoded into three). Categorical variables with missing values were given the value of 'missing'. In most cases these were '*missing non at random*' (MNAR) variables, and in such cases, simple flagging is preferred to imputation.

Next, we converted dates into the number of days elapsed between the start date and the reporting date.
During encoding (dealing with discrete non-numeric variables), we found that *label encoding*[6] improves model performance most even for variables that enumerations cannot be sorted (also known as nominal variables). It may be explained by avoiding the "*curse of dimensionality*": as the number of predictors increases, the difference between the maximum and minimum distances[7] approaches to zero. Boosting algorithms tend to suffer from the curse of dimensionality as they tend to overfit, hence label encoding can be recommended.

The next step contained the treatment of missing values for continuous variables (**Figure 2**), the focus of the first hypothesis (**Table 1**). We have not used columns where the share of missing values exceeded 20 percent at all. Among the rest, the feature with the most missing values had a missing value share of 13 percent, followed by several columns with 12 percent. There seems to be some degree of systematic rule for missingness, since certain feature values are missing for the same observations (**Figure 2)**. This is to be expected: some variables are not interpreted for certain loan contracts.

Three approaches were assessed to fill missing fields. First, missing values were recoded into an arbitrary outlier value (e.g. collateral market value of zero HUF). The model may learn that this special value is not necessarily linked to an economic meaning.

[6] Label encoding converts a categorical variable to a numerical value. While at first it may sound surprising, this kind of treatment works well with the tree-based xgboost-method, since it can learn any difference between 1 and 2, and this is not necessarily the same difference as between 2 and 3. In addition, label encoding does not increase dimensionality, as opposed to the one-hot-encoding technique.

[7] Maximum and minimum distances are distances calculated between the furthest and nearest observations respectively.

Figure 2. Missing values among the continuous variables. Please see the Appendix for further details

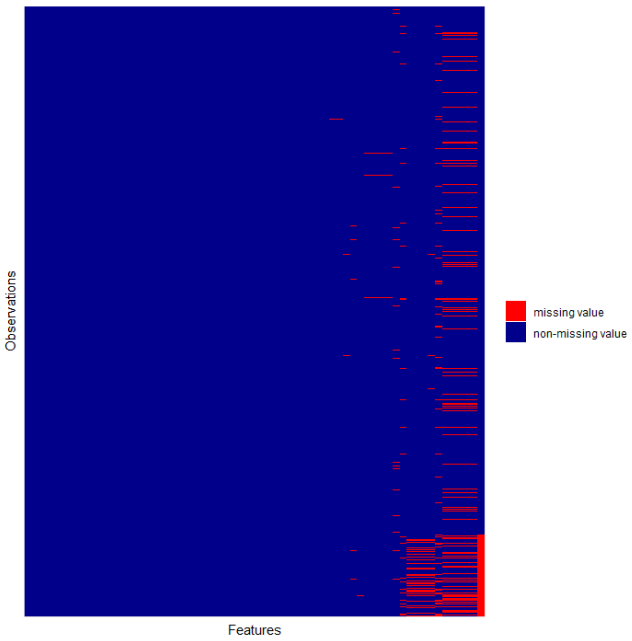


Table 1. Missing value treatment options for **Hypothesis 1**

Missing value treatment	Description	Assumption
Replacement with a constant	The constant (zero) is infrequent or nonexistent in the original data	Xgboost is capable of learning that it is a special value. If the ratio of missing values exceeds 5 percent, an indicator column is added to show the manually filled rows[8].
Imputation with MissRanger	All complete predictors are used (but not the target)	There are existing relationships between predictors
Xgboost sparsity-aware algorithm	The dataset with missing values is fed directly to the xgboost optimization.	This approach puts instances with missing values in the class with the greatest loss reduction when growing a new branch on a tree. This approach delivers a good model on the train data but may not capture the 'ground truth'.

In the second approach, we imputed missing values using the MissRanger package of R (Stekhoven et al. 2012). This is a sophisticated and parallelized, random forest-based algorithm, which uses both categorical and continuous data, and optimizes the imputation process without the need of a test data set. We chose to use all complete predictors to impute the missing values for incomplete predictors. The target variable was not used for the imputation to prevent data leakage. The third method was not to replace missing values at all: we left blank cells blank and let the xgboost deal with them, as discussed earlier.

[8] The presence or absence of the additional flag column is not material: it changes little on model performance or variable interpretation.

Bayesian optimization for hyperparameter-tuning

The last step was the design of the xgboost function. One choice is about the loss function, subject to our second hypothesis and discussed above. The Huber slope parameter, which makes the loss function quadratic up to it, and linear above, is determined with a Bayesian optimization procedure (Snoek et al., 2012), along with all other hyperparameters. This is an efficient method that fits a Gaussian process to the results, looks for hyperparameters which obtain the best evaluation metric (RMSE or the mean pseudo-Huber error) on the test data.

The objective function is the evaluation metric based on the hyperparameter setups, is unknown. The algorithm samples hyperparameter sets, observes the evaluation metric and updates the prior form of the function. There is a predefined similarity measure (Kernel) between observations, whereby similar hyperparameters correspond to similar evaluation metrics. The next samples are determined by the acquisition function, which is based on the previously queried observations, the Kernel function and other hyperparameters, that is related to the exploitation-exploration trade-off, which corresponds to the expected value of the evaluation metric and the variance, based on the previously explored areas and the updated posterior function.

The Bayesian Optimization framework is cited from Shahriari et al. (2016) and Garrido-Merchán et al. (2023). To retrieve the optimum hyperparameters $\boldsymbol{x}^*$ of the unknown objective function, we solve the following:

$$\boldsymbol{x}^* = \boldsymbol{arg}\ \underset{\boldsymbol{x}\in\boldsymbol{X}}{\boldsymbol{min}}\ \boldsymbol{f}(\boldsymbol{x}). \tag{7}$$

The Bayesian Optimization framework in our case can be defined by the following tuple:

$$\boldsymbol{\mathcal{A}} = (\boldsymbol{\mathcal{GP}}, \boldsymbol{\alpha}(\cdot), \boldsymbol{p}(\boldsymbol{f}(\boldsymbol{x})|\boldsymbol{\mathcal{D}})), \tag{8}$$

where $f(x)$ is the optimized function, $\mathcal{GP}$ is a Gaussian Process, $\alpha(\cdot)$ is an acquisition function, and $p(f(x)|D)$ is a predictive distribution (Garrido-Merchán et al, 2023). The task is to provide a predictive distribution for test observations $\{x_i^*\}_{i=1}^M$ with a given dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^N$. A GP is fully defined by the following:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{f}^* \end{bmatrix} \sim \boldsymbol{\mathcal{N}}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \widehat{\mathbf{K}}_{\boldsymbol{X},\boldsymbol{X}} & \mathbf{K}_{\boldsymbol{X},\boldsymbol{X}^*} \\ \mathbf{K}_{\boldsymbol{X}^*,\boldsymbol{X}} & \mathbf{K}_{\boldsymbol{X}^*,\boldsymbol{X}^*} \end{bmatrix}\right), \tag{9}$$

where $y_i = f(x_i) + \varepsilon_i$ are the noisy observations of the optimized function $f(x)$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, the train and test observations are arranged in matrices $\boldsymbol{X}$ and $\boldsymbol{X}^*$, $\boldsymbol{f}$ and $\boldsymbol{f}^*$ are the unobserved true function outputs for all the inputs, $\mathbf{K}_{\boldsymbol{X},\boldsymbol{X}}$ is N-by-N matrix containing the similarities between the observations calculated by the Radial Basis Function $\kappa(\boldsymbol{x_i}, \boldsymbol{x_i'}|\boldsymbol{\tau}) = \sigma^2 \exp\left(-\frac{1}{2}\left(\frac{x-x'}{\ell}\right)^2\right)$, where the hyperparameters of the Kernel denoted by $\boldsymbol{\tau}$, and $\widehat{\mathbf{K}}_{\boldsymbol{X},\boldsymbol{X}} = \mathbf{K}_{\boldsymbol{X},\boldsymbol{X}} + \sigma_\varepsilon^2 \mathbf{I}$. Since the observations and the true

function values at the test points are jointly distributed as a multivariate Normal, the predictive posterior distribution $p(f(x^*)|\mathcal{D})$ are given by:

$$\boldsymbol{\mu_n}(\boldsymbol{x^*}) = \mathbf{K}_{X^*,X}\widehat{\mathbf{K}}_{X,X}^{-1}\boldsymbol{y}, \quad (10)$$

$$\boldsymbol{\sigma_n^2}(\boldsymbol{x^*}) = \mathbf{K}_{X^*,X^*} - \mathbf{K}_{X^*,X}\widehat{\mathbf{K}}_{X,X}^{-1}\mathbf{K}_{X,X^*} \quad (11)$$

The predictive distributions are the estimates of unknown functions areas, and the next samples are determined by the acquisition function. It deals with the exploration-exploitation trade-off, in the sense of the search space and the promising areas (Shahriari et al, 2016). In our analysis we used the upper confidence bound as an acquisition function:

$$\boldsymbol{\alpha_{UCB}}(\boldsymbol{x^*};\boldsymbol{\mathcal{D}_n}) := \boldsymbol{\mu_n}(\boldsymbol{x^*}) + \boldsymbol{\beta_n}\boldsymbol{\sigma_n}(\boldsymbol{x^*}). \quad (12)$$

The approach, preferably parallelized, vastly reduces computation time compared to a brute force-based grid search. The permitted bounds for the hyperparameter-optimization, defined after a set of initial experiments, are reflected by **Table 2**.

Table 2. Permitted hyperparameter ranges

Hyperparameter	Descipriton	Bounds
Eta (learning rate)	Step size used in update	0.01 - 0.1
Max depth	Maximum depth of a tree.	4 - 8
Gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree.	0-8
Alpha	L1 regularization term on weights.	10-20
Lambda	L2 regularization term on weights.	1-12
Huber-slope*	Defines the range for the loss function piecewise.	10-30

* with Huber loss function only. The maximum number of trees was set to 280

Synthetic errors and model evaluation

We introduced three types of synthetic errors to the data (**Table 3)**. Synthetic errors are often used to test data quality-algorithms (Abedjan et al. 2016). We added the following errors to 5 percent of randomly selected cases (both in train and in test datasets):

- LTV divided by 100 (an LTV of 50 percent is to be submitted as '50', but is often submitted as 0,5)
- LTV set to 80 (as if a dummy value was submitted, which also happens in reality). This error type was applied to observations with a true LTV of 60 or less to have a sufficiently sizeable error.

- LTV multiplied by a sample drawn from uniformly distributed ranges of U(0.3, 0.5) and U(1.2, 1.4)[9] random variables. The two ranges are not symmetrical to limit the number of errors above 100 (percent) of LTV.

In a separate experiment, errors were introduced to one of the predictors. We chose to alter the second most important predictor in terms of gain in the feature importance matrix (described later), the allocated collateral value. Data providers have more discretion during the allocation process than it is the case with loan amount (the most important feature) - this latter is explicitly written in a loan contract. Therefore, variations, errors and outliers are more probable in the case of the allocated collateral value. The errors we introduced were:

- Values were multiplied by a random value sampled from the uniformly distributed ranges U(0.3, 0.5) or U(1.2, 1.4). This in essence means a multiplication by a value in any of the mentioned bounds.
- Values were set to 10 mln HUF (approx. 25 thousand EUR). This error type was applied to observations with a true allocated collateral value of 40 mln HUF or more to have a sufficiently sizeable error. Such an error would occur when dummy value is reported to the central bank.

Table 3. Data errors used for algorithm evaluation

Error location	Error description	Error rationale
Response variable	values divided by 100	An LTV of 50 percent is expected to be submitted as '50', it is often submitted as 0,5
Response variable	values set to 80	as if a dummy value was submitted
Response variable	Values were multiplied by a random value. This random value was drawn for each observation from a uniform distribution of between U(0.4, 0.6) and U(1.2, 1.4).	A random error
Predictor (allocated collateral value)	Values set to 10 mln HUF	The use of a dummy value or an internal cap on allocation
Predictor (allocated collateral value)	Values were multiplied by a random value. This random value was drawn for each observation from a uniform distribution of between U(0.4, 0.6) and U(1.2, 1.4).	A random error

The outlier detection approaches were evaluated with two metrics (**Table 4**). First, we look at the share of discovered errors out of all errors. An error was considered as discovered when its absolute residual was above a pre-defined threshold of 20 (since the target variable is expected to be between 0 and 80 (a legal limitation), an absolute threshold does not materially distort the results). Ceteris paribus, the more errors we identify, the better the model is. At the same time, we need an additional metric: the share of intentional errors within outliers, compared to the share of intentional errors

[9] We used the usual abbreviation of the uniform distribution (denoted by U).

in the whole test data set (referred to as 'lift'). Error density should be considerably higher in the outlier set than in the data.

We tested two outlier definitions. First was based on a pre-defined (unstandardized[10]) residual value, above which all observations were treated as outliers. The second definition relied on the top pre-defined share of observations, in our case, 5 percent, which were classified as outliers, when ranked according to the absolute value of its residuals. One argument for this second definition could be the limited time available to follow up outlier-issues with the data providers.

Table 4. Error detection evaluation metrics

Metric	Metric description	Outlier definition
Share of known errors discovered	Synthetic errors in the outlier set as a proportion of all synthetic errors	**Definition 1:** Outlier set defined as the observations with an absolute residual of a pre-defined value or higher **Definition 2:** A pre-defined share of observations with the greatest residual values
Lift	Frequency of synthetic errors in the outlier set, divided by the frequency of synthetic errors in the entire data	

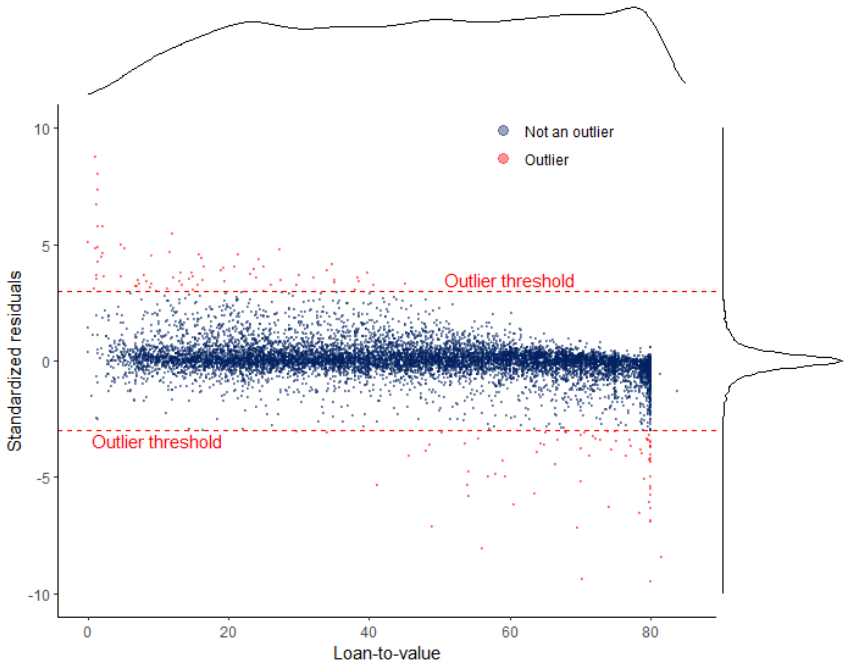
Results

The baseline model

We prepared a baseline xgboost model, imputed missing values using a constant, and a squared loss function (and with no synthetic errors). Although the model's residual plot still reflects a systematic pattern (**Figure 3**) - it overestimates low observations and underestimates high ones - it performs well. The RMSE is 5.6 percent, and the share of outliers is 1.4 percent, using the cutoff value on *standardized* residuals of 3. A sample of our flagged outliers were identified as data errors by domain experts. The algorithm also found intuitive errors (e.g. LTV between 0 and 1 instead of 0 and 100).

[10] We decided on using the unstandardized value for better interpretability, given the fact that LTV is expected to move between 0 and 100. In more heterogeneous cases or if the ALSO-algorithm is to be carried out in its full extent, one may standardize the residuals.

Figure 3. Residual plot of the baseline model with marginal distributions (graph contains a random sample of data points only)



The two most important features in the model are loan amount and allocated collateral value at loan origination (**Table 5**). Since LTV at loan origination is related to the division of these two values, it is promising to see these features on the top of the list. The reason why we say LTV is only related to these two features is that banks are permitted to apply certain rules (caps, modifications) in using these values during LTV-calculation both to the nominator and to the denominator.

Table 5. Feature importance in the baseline model (top 10)

Feature	Gain	Cover	Frequency
Loan amount (in HUF)	0,32	0,13	0,07
Allocated collateral value (at loan origination)	0,25	0,07	0,09
Full collateral market value at loan origination	0,12	0,06	0,06
Collateral value	0,05	0,03	0,04
Monthly repayment amount	0,04	0,07	0,06
Effective interest rate	0,03	0,02	0,03
Financed real estate type[11]	0,02	0,03	0,02

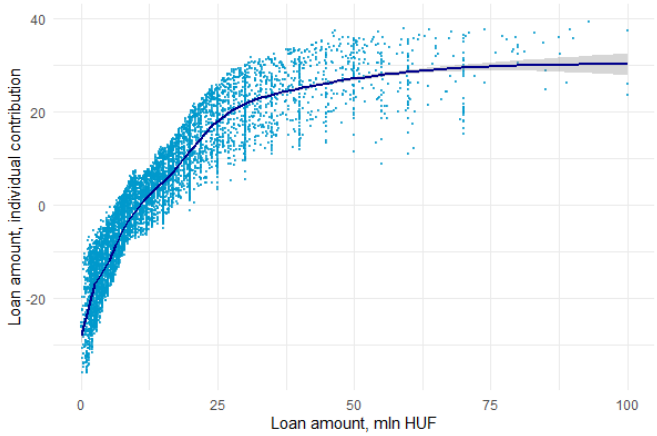
[11] The categorical variable converted to a numerical value using label encoding. While at first it may sound surprising, this kind of treatment works well with the tree-based xgboost-method, and does not increase dimensionality, as opposed to the one-hot-encoding technique.

Days past since collateral value determination[12]	0,02	0,02	0,03
Rate driver[13]	0,02	0,00	0,01
Expected loss	0,01	0,04	0,04

Figure 4 and **Figure 5** depict the individual feature contributions for a randomly selected 10 thousand observations for the two most importance features (by gain). The individual feature contribution is a weight the given value of the predictor has on the target. Since xgboost develops a complex set of trees, including predictor interactions (see section **4.2**), the same predictor value may have varying impacts on the prediction. Each point on the plot gives the impact the actual feature set of the observation on the target.

These plots reflect the division between the 'Loan amount' (which is related to the nominator) and the 'Allocated collateral value' (which is related to the denominator). Note, however, the strange curve of the 'Allocated collaretal value' around zero. This reflects the impact of the missing value treatment. The cases where this value was missing were given decisively heterogeneous and broad individual contribution values by the model. We assume that here the algorithm learned the underlying function along other variables. Beyond zero, the individual contribution plot for this predictor is reminiscent of a hyperbole.

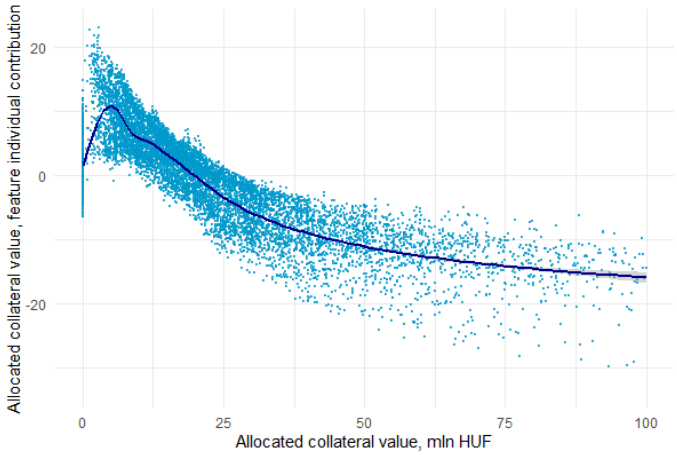
Figure 4. Individual feature contribution of'Loan amount' (with an illustrative LOESS-function)



[12] The maximum amount, in case of multiple collaterals.

[13] The categorical variable converted to a numerical value using label encoding. While at first it may sound surprising, this kind of treatment works well with the tree-based xgboost-method, and does not increase dimensionality, as opposed to the one-hot-encoding technique.

Figure 5. Individual feature contribution of'Allocated collateral value' (with an illustrative LOESS-function)



Hypotheses 1 and 2

We classified all observations with an absolute residual of 20 or higher as outliers, and calculated the evaluation metrics. The 80-dummy and the div_100 error types were discovered with a lift value of 10-12, with the exception the 'none' (no explicit missing value replacement) and Huber-loss combinations (**Table 6)**. In this latter case, the share of outliers was similar, but the models' RMSE and MAE values and the total share of outliers were higher, and therefore lift values were lower. This poor performance of the 'none' – 'Huber-loss' combinations were consistent across all three error-types.

Table 6. Evaluation metrics when using an absolute outlier threshold of 20; with a **synthetic error in the target variable**

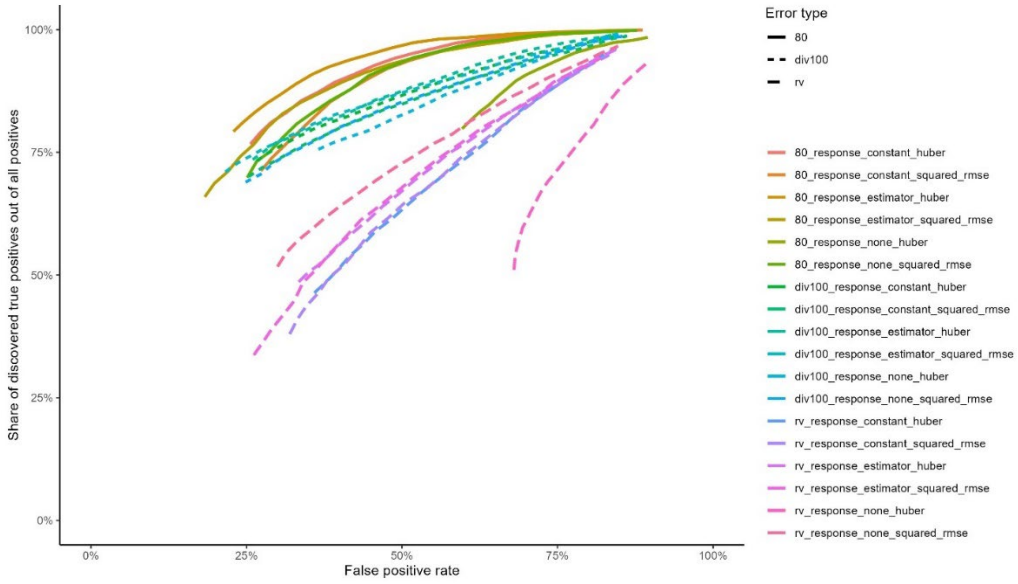
Error type	Missing value replacement	Loss function	Outliers as % of total	Share of discovered errors	Lift	RMSE	MAE
80	estimator	Huber	6,8%	88,9%	13,00	12,10	6,82
80	none	Huber	13,3%	88,0%	6,60	17,90	9,73
80	constant	Huber	6,8%	87,2%	12,80	11,60	6,54
80	constant	squared loss	6,9%	85,3%	12,40	11,30	7,00
80	none	squared loss	6,5%	83,1%	12,80	10,90	6,69
80	estimator	squared loss	5,6%	80,1%	14,30	10,20	6,14
div100	none	Huber	8,1%	82,3%	10,20	13,30	6,97
div100	estimator	Huber	6,4%	81,3%	12,60	12,40	6,32
div100	constant	Huber	6,6%	81,1%	12,40	12,30	6,34
div100	estimator	squared loss	5,9%	79,1%	13,50	11,50	6,26
div100	constant	squared loss	6,2%	78,4%	12,60	11,50	6,37
div100	none	squared loss	6,2%	78,1%	12,60	11,30	6,22

rv	none	squared loss	5,3%	64,2%	12,20	9,30	5,40
rv	none	Huber	11,2%	64,2%	5,76	22,10	9,59
rv	constant	Huber	5,4%	59,0%	10,90	9,24	5,36
rv	estimator	Huber	5,1%	58,8%	11,60	9,11	5,29
rv	constant	squared loss	4,4%	52,0%	12,00	8,25	4,86
rv	estimator	squared loss	3,6%	48,4%	13,30	7,62	4,49

There is no visible tradeoff between discovered error shares and lift values. Moreover, models with imputed missing values do not outperform models with a constant-replacement. Further, the Huber objective function performs better than the squared loss functions do in terms of the first metric, the share of discovered intentional errors. However, the share of outliers as of all observations is not necessarily larger with the Huber loss function, if a missing value imputation was used.

Figure 5 illustrates model performance[14] over the various error types in the response variable when gradually decreasing the absolute outlier threshold value from 25 to 5. The x-axis shows the share of non-errors in the outlier set, in other words, the false positive rate, while the y-axis represents the share of known, outlier errors as a percentage of all synthetic errors. **Figure 6** reinforces model stability and the messages of our research.

Figure 6. False- and true-positive rates when moving the outlier threshold from 25 to 5 (synthetic errors in the target variable)



[14] The difference **Figure 6** and a ROC-curve is that in our case, we do not have a direct probability value, only a cutoff outlier value.

Table 7 contains model metrics when a relative outlier threshold of 5 percent of all observations is used. In a similar vein, **Table 7** confirms that the estimation of missing values as imputation in our case delivers better outlier recognition performance than using a constant.

Regarding our initial Hypotheses: the Huber-loss often delivers better results than the squared loss, and an imputation is most often better than no imputations. These are indications in favor of Hypotheses 1 and 2, but every dataset should be tested for it.

Table 7. Error detection metrics when using a relative outlier threshold of 5 percent (**synthetic error in the target)**

Error type	Missing value replacement	Loss function	Outliers as % of total	Share of discovered errors	Lift	RMSE	MAE
80	estimator	Huber	5,0%	76,5%	15,30	12,10	6,82
80	constant	Huber	5,0%	75,7%	15,10	11,60	6,54
80	estimator	squared loss	5,0%	75,0%	15,00	10,20	6,14
80	none	squared loss	5,0%	73,1%	14,60	10,90	6,69
80	constant	squared loss	5,0%	72,3%	14,50	11,30	7,00
80	none	Huber	5,0%	54,7%	10,90	17,90	9,73
div100	estimator	squared loss	5,0%	74,5%	14,90	11,50	6,26
div100	estimator	Huber	5,0%	73,9%	14,80	12,40	6,32
div100	constant	Huber	5,0%	73,5%	14,70	12,30	6,34
div100	constant	squared loss	5,0%	72,3%	14,50	11,50	6,37
div100	none	squared loss	5,0%	71,8%	14,40	11,30	6,22
div100	none	Huber	5,0%	70,7%	14,10	13,30	6,97
rv	none	squared loss	5,0%	62,6%	12,50	9,30	5,40
rv	estimator	squared loss	5,0%	58,9%	11,80	7,62	4,49
rv	estimator	Huber	5,0%	58,4%	11,70	9,11	5,29
rv	constant	squared loss	5,0%	56,8%	11,40	8,25	4,86
rv	constant	Huber	5,0%	56,5%	11,30	9,24	5,36
rv	none	Huber	5,0%	31,6%	6,33	22,10	9,59

Testing Hypothesis 3

The outcome of the same experiments (recall that we use slightly different errors) is summarized by **Table 8**.

Table 8. Error detection metrics with an absolute outlier threshold of 20 (**synthetic error in the second most important predictor)**

Error type	Missing value replacement	Loss function	Outliers as % of total	Share of discovered errors	Lift	RMSE	MAE
10	none	Huber	3,5%	3,0%	0,87	7,91	4,73
10	estimator	Huber	3,0%	2,1%	0,70	7,47	4,36
10	none	rmse	1,5%	1,6%	1,05	5,83	3,43
10	constant	Huber	2,3%	1,3%	0,57	6,98	4,21
10	constant	rmse	1,3%	1,1%	0,83	5,55	3,25
10	estimator	rmse	1,1%	0,9%	0,83	5,27	3,18
rv	none	rmse	2,0%	10,7%	5,41	6,30	3,69
rv	estimator	Huber	4,5%	7,9%	1,75	9,18	6,11
rv	none	Huber	2,5%	5,8%	2,31	7,10	4,29
rv	constant	Huber	3,0%	4,3%	1,43	7,72	4,75
rv	constant	rmse	1,5%	2,1%	1,41	5,84	3,46
rv	estimator	rmse	1,3%	1,9%	1,51	5,54	3,37

Table 8 shows dismal error recognition metrics with disappointingly low values: the models are not better than randomly selecting observations from the data. The metrics send a similar message with relative outlier threshold (not shown).

To understand the reason for that, we compare the feature importance matrix for the baseline model and for the model with synthetic errors in the predictor 'Allocated collateral value'. While the feature importance matrix can only signal changes and is not a mathematical proof, it indicates change directions. **Table 9** shows that the importance of this predictor decreases slightly for all three aspects (gain, cover and frequency). At the same time, the third predictor, exhibits higher importance values, among other changes. These moves point out the reason why the algorithm is less capable to identify errors in the predictors. Our intuition is that xgboost can approximate the underlying function despite the added noise to 'Allocated collateral value', because it finds a structure using other variables.

Table 9. Feature importance of the baseline model and the model with random variable synthetic errors in the second predictor

Baseline model				Peer model with synthetic errors in the predictor			
Feature	**Gain**	**Cover**	**Frequency**	**Feature**	**Gain**	**Cover**	**Frequency**
Loan amount (in HUF)	0,32	0,13	0,07	Loan amount (in HUF)	0,32	0,15	0,07
Allocated collateral value (at loan origination)	**0,25**	**0,07**	**0,09**	**Allocated collateral value (at loan origination)**	**0,23**	**0,08**	**0,09**
Full collateral market value at loan origination	0,12	0,06	0,06	Full collateral market value at loan origination	0,13	0,07	0,07
Collateral value	0,05	0,03	0,04	Collateral value	0,05	0,03	0,04
Monthly repayment amount	0,04	0,07	0,06	Monthly repayment amount	0,03	0,06	0,05
Effective interest rate	0,03	0,02	0,03	Effective interest rate	0,03	0,03	0,03
Financed real estate type*	0,02	0,03	0,02	Financed real estate type*	0,02	0,04	0,02
Days past since collateral value determination**	0,02	0,02	0,03	Expected loss	0,02	0,03	0,04
Rate driver*	0,02	0,00	0,01	Days past since collateral value determination**	0,02	0,03	0,04
Expected loss	0,01	0,04	0,04	Original maturity (in days)	0,01	0,04	0,04

* the categorical variable converted to a numerical value using label encoding. While at first it may sound surprising, this kind of treatment works well with the tree-based xgboost-method, and does not increase dimensionality, as opposed to the one-hot-encoding technique.
** The maximum amount, in case of multiple collaterals

Conclusion

In this paper, we analyzed technical aspects of a supervised, gradient boosting-based machine learning method, described by Paulheim and Meusel (2015) and Benatti (2018), to identify potential data errors in a central banking dataset. The algorithm relies on the assumption that most of the data points are correct and there are 'ground truth' relationships between the features. Data points deviating from such a relationship, outliers, are flagged as a potential errors, which are to be investigated further. While the results we found may be specific to the data used, the paper provides a detailed overview of the aspects one may consider during a similar modelling pipeline.

We used the Credit Registry (Hitelregiszter) dataset collected by the Central Bank of Hungary (MNB), which is a collection of datasets provided by supervised entities (banks) to the central bank. More specifically, we looked at the loan-to-value (LTV) part of the data for one business date, carrying out a cross-sectional analysis. In order to test three technical hypotheses relevant during model development, we introduced three types of synthetic errors to the data.

First, we tested the treatment of missing values in the data (when the data is missing both during model training and prediction). We found that an imputation method most often proposed by the literature is not necessarily superior to using a constant value as a placeholder for 'missing' label. This is explained by the flexible nature of the xgboost algorithm. Moreover, one should be careful when using the sparsity-

aware splitting method native to the xgboost algorithm, because it may worsen model performance, particularly when used in combination with the Huber loss function.

Second, we compared two loss functions for xgboost. We found that the Huber loss function captures more errors than the traditional, squared loss function. Squared loss-function can be more sensitive to outliers, and when using it, xgboost is more likely to 'learn' data errors as true relationships than it would be the case with the Huber-function.

Finally, we showed that the algorithm captures errors in the target variable, as opposed to errors in the predictors. The reason for that lies in the flexibility of the xgboost and high dimensionality of our dataset. During model training with errors in the predictors, the 'role' of the erroneous predictor is taken over by others, strongly diminishing error identification performance.

Potential future research questions could be directed towards expanding the cross-sectional nature of our investigation by looking at pattern development over time. Another potential research idea relates to providing better – and automated - answers why an outlier seems to be an error. While individual feature contributions seem to be interpretable options, they are still a far cry from being able to formulate an actionable question to the data provider.

References

Aggarwal C. C. (2017) **Outlier Analysis**. Springer, Second Edition. DOI 10.1007/978-3-319-47578-3

Benatti, N. (2018) A machine learning approach to outlier detection and imputation of missing data. Paper presented at the Ninth IFC Conference on "Are post-crisis statistical initiatives completed?" Basel, 30-31 August 2018. Available at: https://www.bis.org/ifc/publ/ifcb49_48.pdf

Chen, T., Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

Daniel J. Stekhoven, Peter Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics*, Volume 28, Issue 1, 1 January 2012, Pages 112–118, https://doi.org/10.1093/bioinformatics/btr597

Garrido-Merchán, E. C., Piris, G. G., Vaca, M. C. (2023) Bayesian Optimization of ESG Financial Investments. https://arxiv.org/pdf/2303.01485.pdf

Khosravi, P., Vergari A., Choi, J., Liang, Y., Van den Broeck, G. (2020) Handling Missing Data in Decision Trees: A Probabilistic Approach. Paper presented at the conference 'The Art of Learning with Missing Values Workshop at ICML, Vienna, Austria, 2020', available at: http://starai.cs.ucla.edu/papers/KhosraviArtemiss20.pdf

Kim, W., Choi, BJ., Hong, EK. et al. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery 7, 81–99 (2003). https://doi.org/10.1023/A:1021564703268

Paulheim, H., Meusel, R. (2015) A decomposition of the outlier detection problem into a set of supervised learning problems. **Mach Learn** 100, 509–531 (2015). https://doi.org/10.1007/s10994-015-5507-y

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Sadouk, L., Gadi, T., Essoufi, E.H. (2020). Robust Loss Function for Deep Learning Regression with Outliers. In: Bhateja, V., Satapathy, S., Satori, H. (eds) Embedded Systems and Artificial Intelligence. Advances in Intelligent Systems and Computing, vol 1076. Springer, Singapore. https://doi.org/10.1007/978-981-15-0947-6_34

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., De Freites, N. (2016) Taking the Human Out of the Loop: A Review of Bayesian Optimization. In: Proceedings of the IEEE, vol. 104, no. 1, pp. 148-175, Jan. 2016, doi: 10.1109/JPROC.2015.2494218.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems, 25.

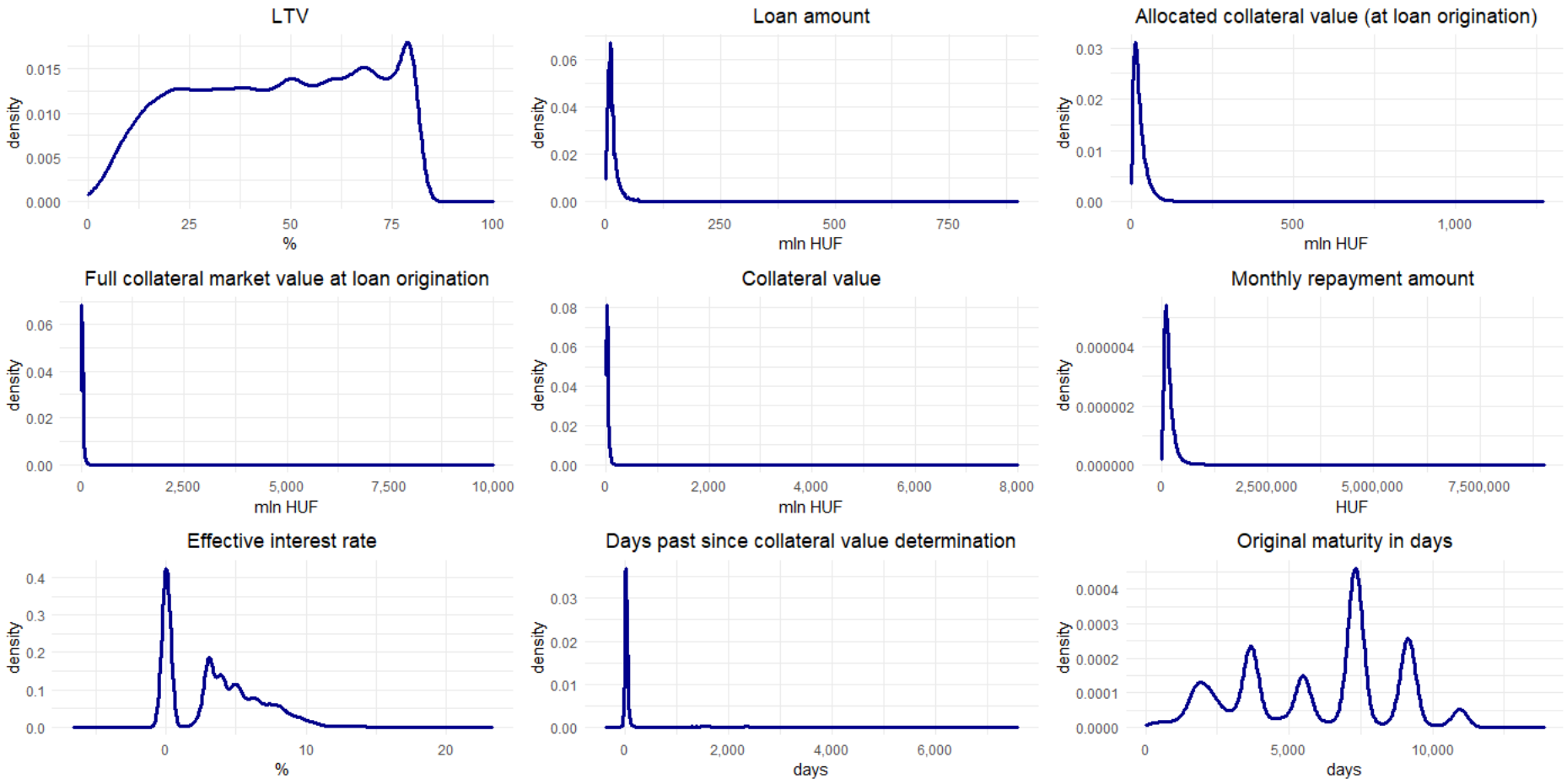
Twala, B. E. T. H.; Jones, M. C. and Hand, D. J. (2008). Good methods for coping with missing data in decision trees. Pattern Recognition Letters, 29(7) pp. 950–956.

Appendix

Table 10. Selected summary statistics of the target and the most important explanatory variables by feature importance gain

n = 72 996	Number of missing values	Mean	Standard deviation	Median	Median absolute deviation from the median	Skewness	Kurtosis
LTV	-	47	22	48	28	-0.13	-1.16
Loan amount	-	14 076 346	13 538 794	10 000 000	7 413 000	8	294
Allocated collateral value (at loan origination)	9 732	26 976 614	25 248 320	20 300 000	15 122 520	6	150
Expected loss	2 803	11 527 849	30 627 914	7 835 219	8 584 562	29	1 145
Debt-to-income percentage	549	30	18	30	13	100	18 497
Monthly repayment amount	544	175 902	139 058	143 876	83 582	9	324
Effective interest rate	5 929	4	3	4	5	0.46	-0.43
Full collateral market value at loan origination	8 933	34 025 618	60 486 728	27 000 000	18 532 500	121	19 042
Collateral value	8 933	25 340 908	48 613 488	20 000 000	15 715 560	120	18 711
Original maturity in days	-	5 967	2 681	7 297	2 731	-0.18	-1.05
Days past since collateral value determination**	8 930	204	644	28	22	5	36
Elapsed days since loan origination	0	204,15	150,86	192	127,5	7,82	170,33
APR	64	5,49	2,61	5,49	2,95	2,18	81,03
Birth year of the main debtor	0	1982,11	10,45	1983	10,38	-1,33	6,08
Loan cost percentage	64	6,47	2,85	6,1	2,33	2,26	56,92
Mortgage rate	3842	5,63	2,46	5,2	2,14	1,07	1,1

Figure 7. Density functions for the target variable (LTV) and for the most important explanatory features (by feature importance gain) **Note**: amounts are in HUF



Dr. Csaba Burger, CFA (MNB)
Mihály Berndt (Clarity Consulting)

3rd IFC and Bank of Italy Workshop on "Data Science in Central Banking: Enhancing the access to and sharing of data"

ERROR SPOTTING WITH GRADIENT BOOSTING

Rome, 18th October 2023

The views expressed are those of the authors and do not necessarily reflect the official view of the Central Bank of Hungary (Magyar Nemzeti Bank).

MOTIVATION, BACKGROUND, SUMMARY OF RESULTS

Background

- MNB's commitment to high data quality
- Machine learning is suitable for large data volumes
- The role of ML in data quality checks is not yet standardized

Results

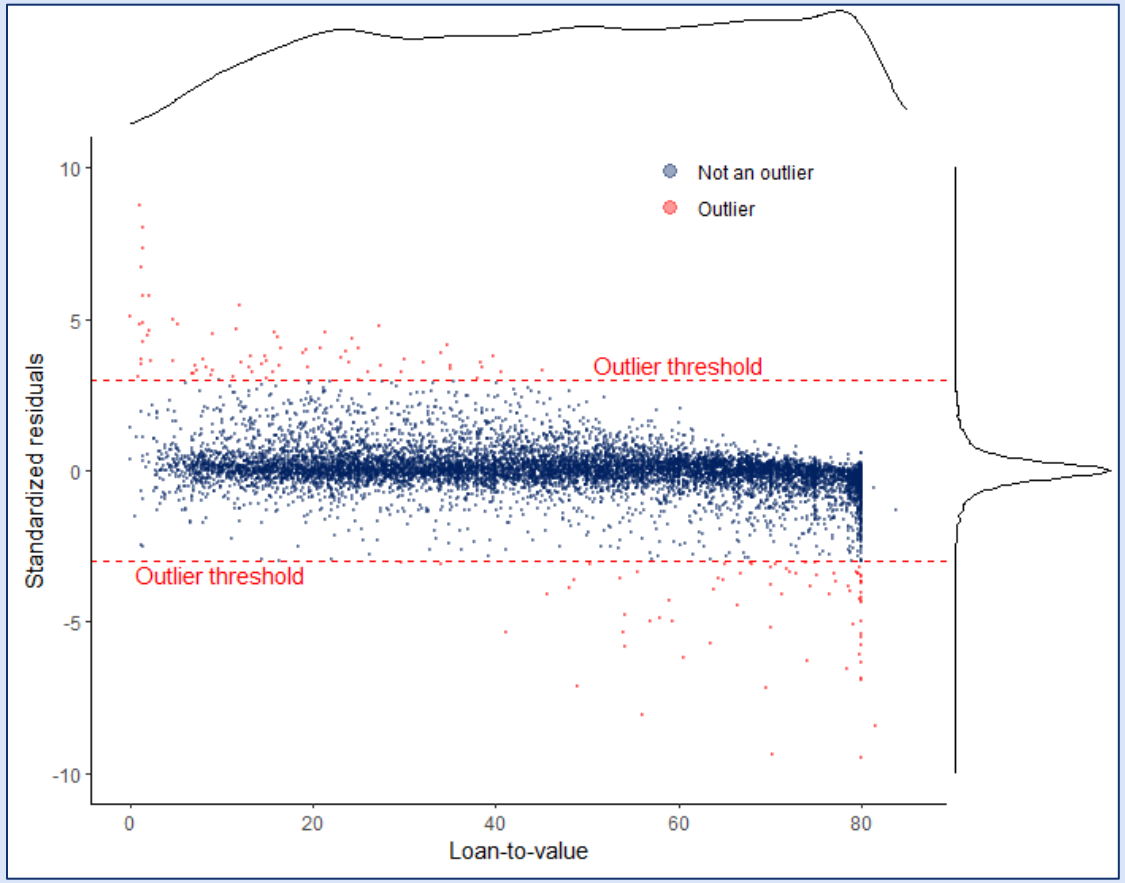
- Un-labelled supervised learning can uncover relationships within the data
- State-of-the-art modelling techniques (XGBoost, Bayesian optimization)
- We present a few recommendations to flag potential data errors

METHODS WE EMPLOY

Unlabelled supervised methods we use

1. **Aggregated time series**
2. **Cross-sectional - granular**
3. **Granular time series**

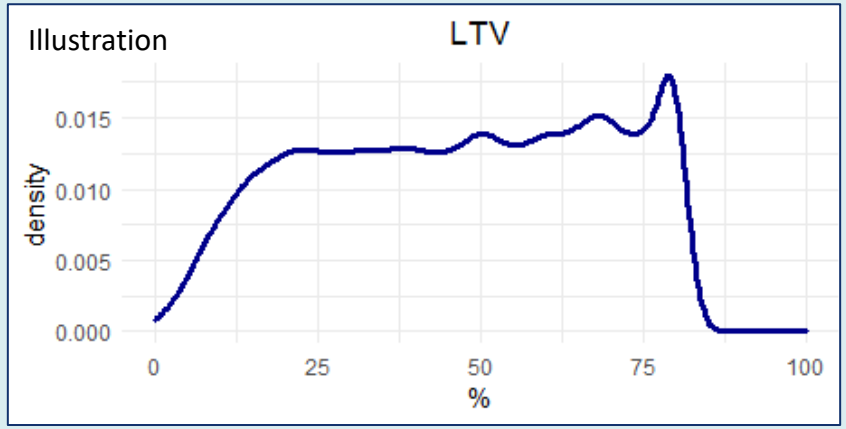
Residual plot in a model explaining a selected target variable



DATA

MNB LTV report

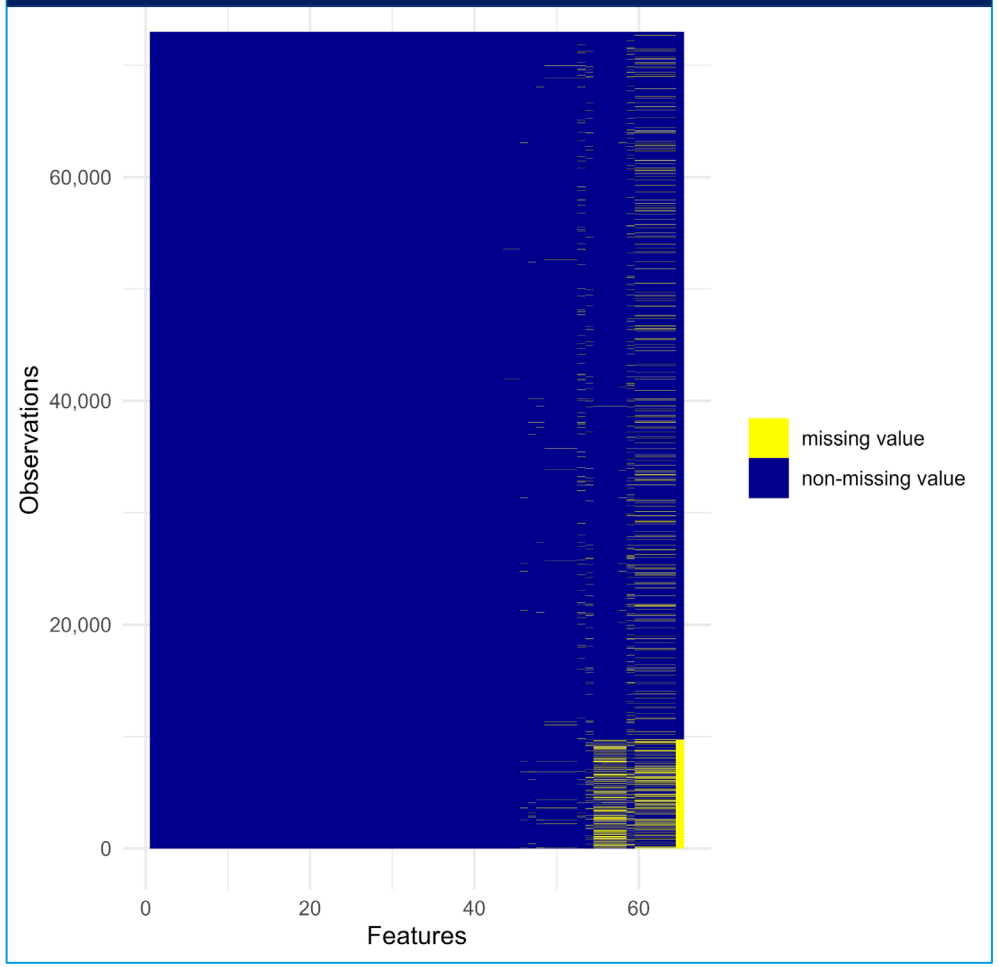
- First ranking mortgages with a start date after 1st Oct 2021
- Approx. 73 thousand lines
- 274 columns → 69 columns (high correlations, missing value share >= 20 percent)



Just a theory

$$\text{LTV} = \frac{\text{Loan amount}}{\text{Allocated collateral value}}$$

Missing values



XGBOOST GROWS DECISION TREES IN A SEQUENTIAL MANNER

Loss reduction calculation

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma.$$

Similarity scores based on:

- residual **direction**
- residual **magnitude**

Many hyperparameters to optimize

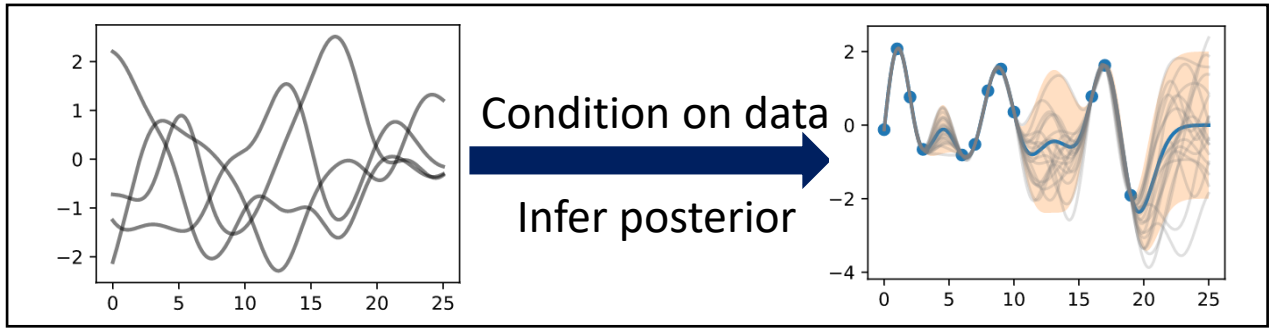
Sparsity-aware split finding

1. Visit only non-missing entries
2. Determine the best split and **default direction** for missing value based on the Similarity score above

BAYESIAN OPTIMIZATION TO FIND THE BEST HYPERPARAMETER CONFIGURATION

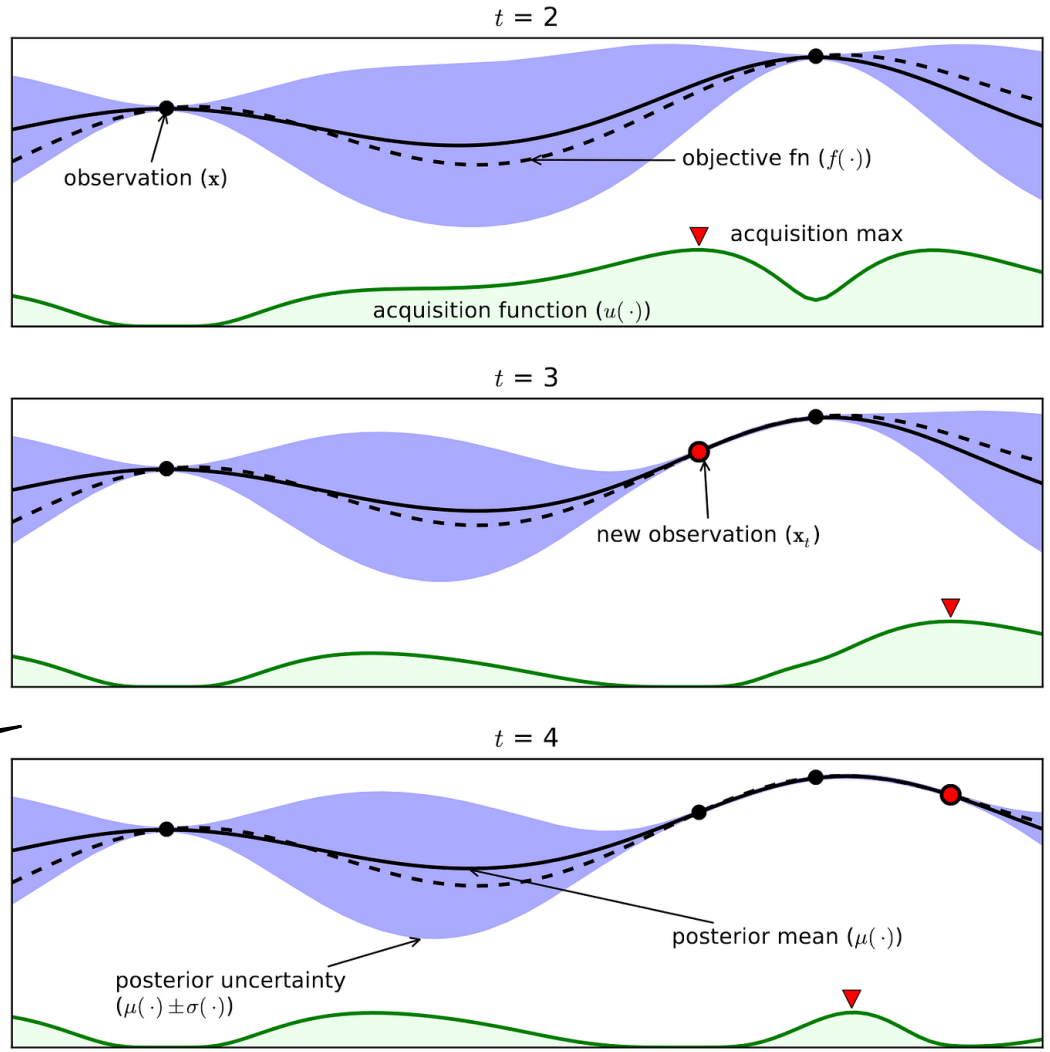
GP: model of the objective function behaviour

- Train and test points are jointly distributed as multivariate normal
- Kernel encodes similarities between data points (shape of the prior)



How to determine new samples?

- Acquisition Function
- Exploration-explotation trade-off



MODELLING PIPELINE AND DESIGN QUESTIONS

Treatment of missing values | **Treatment of rare values** | **Determining the loss function** | **Bayesian optimization** | **Interpreting the results**

Question 1

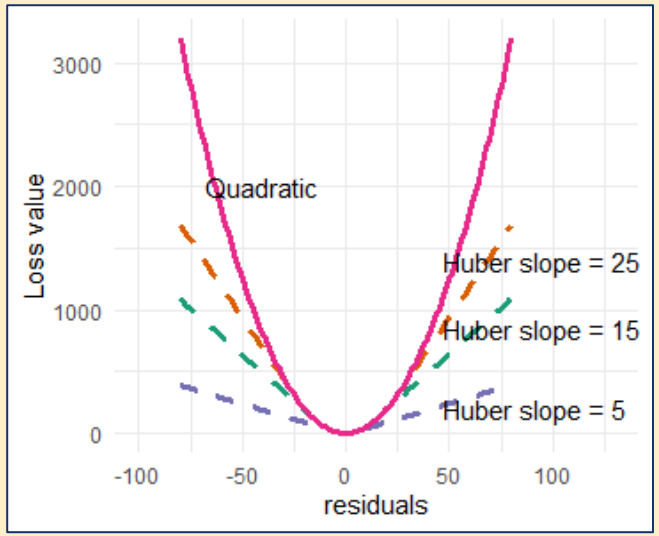
Ways to deal with missing values:

- Estimator (missRanger)
- Constant (unusual dummy)
- Xgboost's sparsity-aware split finding

Question 2

Loss function choice:

- Mitigate the impact of existing errors on finding the ground truth



Question 3

We assume:

- **If**: explanatory column 'B', is not independent from 'A'
- **and** data error distorts an explanatory variable 'A'
- Then B takes over from A

SYNTHETIC ERRORS
INSPIRATION FROM EXISTING ERRORS

Location	Description
Response variable	Values divided by 100.
Response variable	Values set to 80.
Response variable	Values were multiplied by a random value, drawn for each observation from U(0.4, 0.6) and U(1.2, 1.4)
Predictor (2nd most important)	Values set to 10 mln HUF
Predictor (2nd most important)	Values were multiplied by a random value, drawn for each observation from U(0.4, 0.6) and U(1.2, 1.4)

Errors in 5% of all observations, both in train and test sets

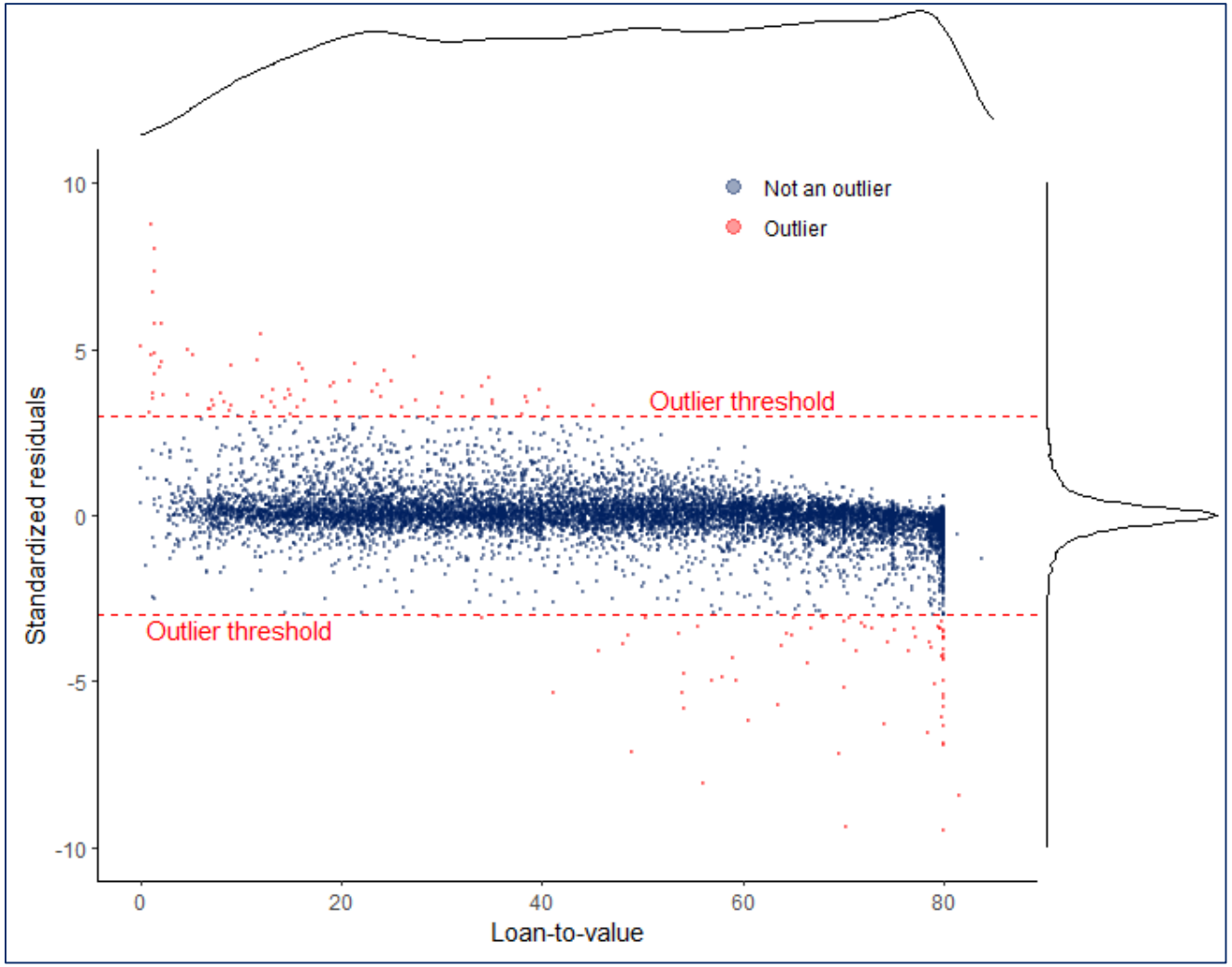
BASELINE MODEL PERFORMANCE
WITHOUT SYNTHETIC ERRORS

The baseline model

- missing values using a constant
- squared loss function
- **no synthetic errors**

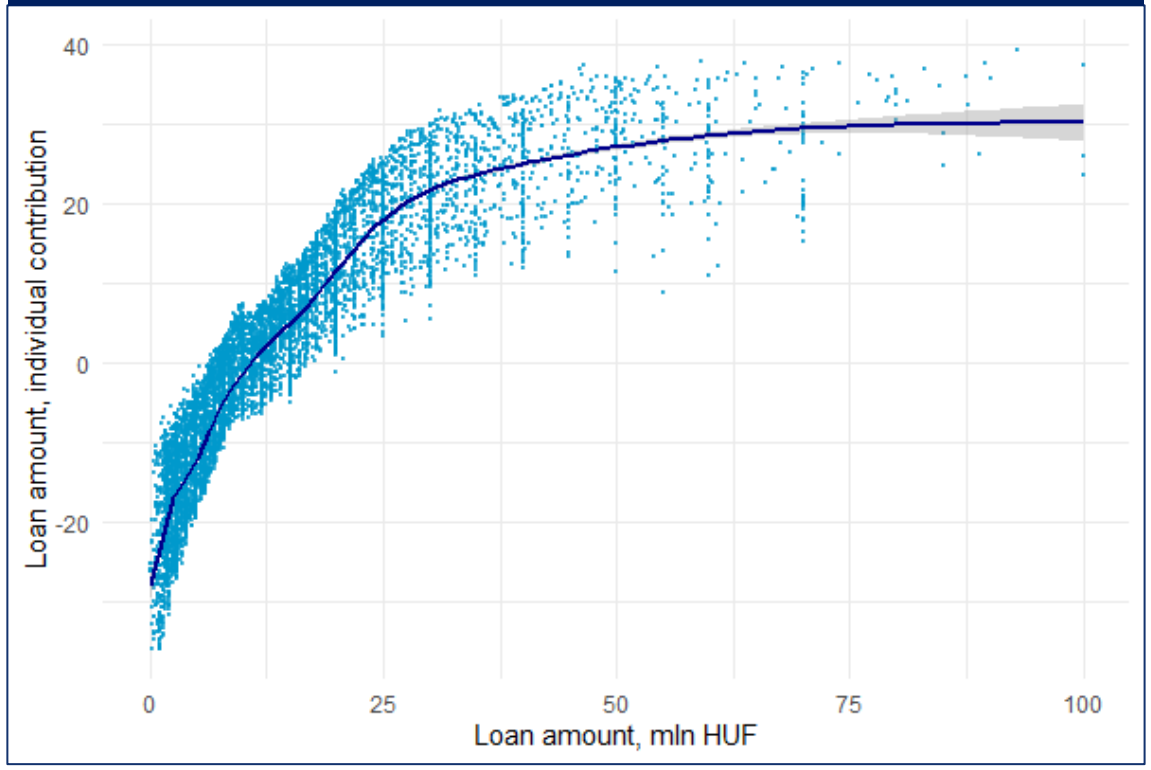
Model performance

- RMSE = 5.6 percent, MAE = 3.2 percent
- the share of outliers is 1.4 percent only (cutoff of *standardized* residuals of 3)
- The algorithm found intuitive errors (LTV as a fraction between 0 and 1)

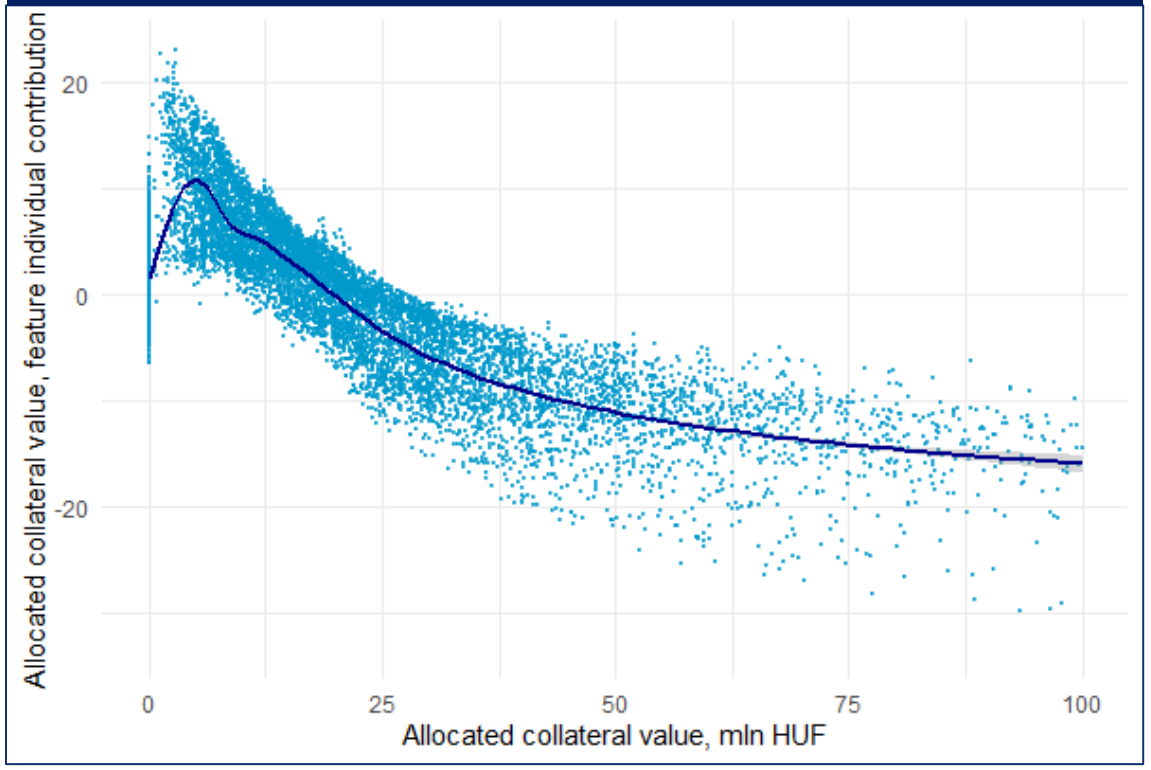


THE BASELINE MODEL – INDIVIDUAL FEATURE CONTRIBUTIONS (IFC)

IFC for Loan amount
+ a LOESS function



IFC for allocated collateral value
+ a LOESS function



$$\text{LTV} = \frac{\text{Loan amount}}{\text{Allocated collateral value}}$$

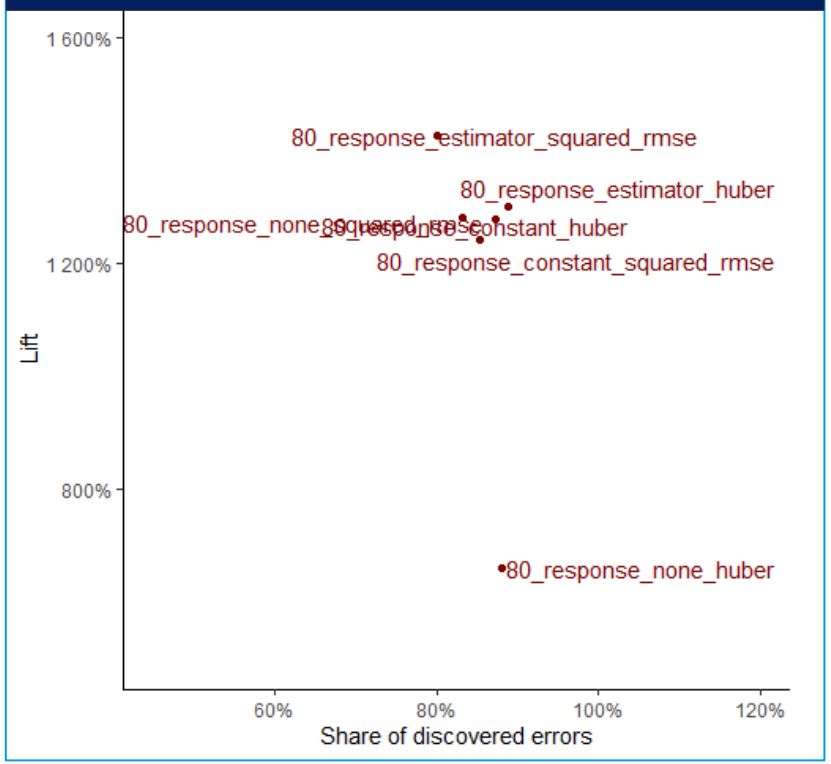
METRICS TO ASSESS MODEL PERFORMANCE

	Share of discovered errors	Lift value
Formula	$Disc.error\ sh. = \frac{Errors\ among\ outliers}{All\ errors}$	$Lift = \frac{Error\ share\ among\ outliers}{Error\ share\ in\ all\ data}$
Rationale	Did we find every synthetic error?	Am I any better off by looking at outliers than going through the raw data?

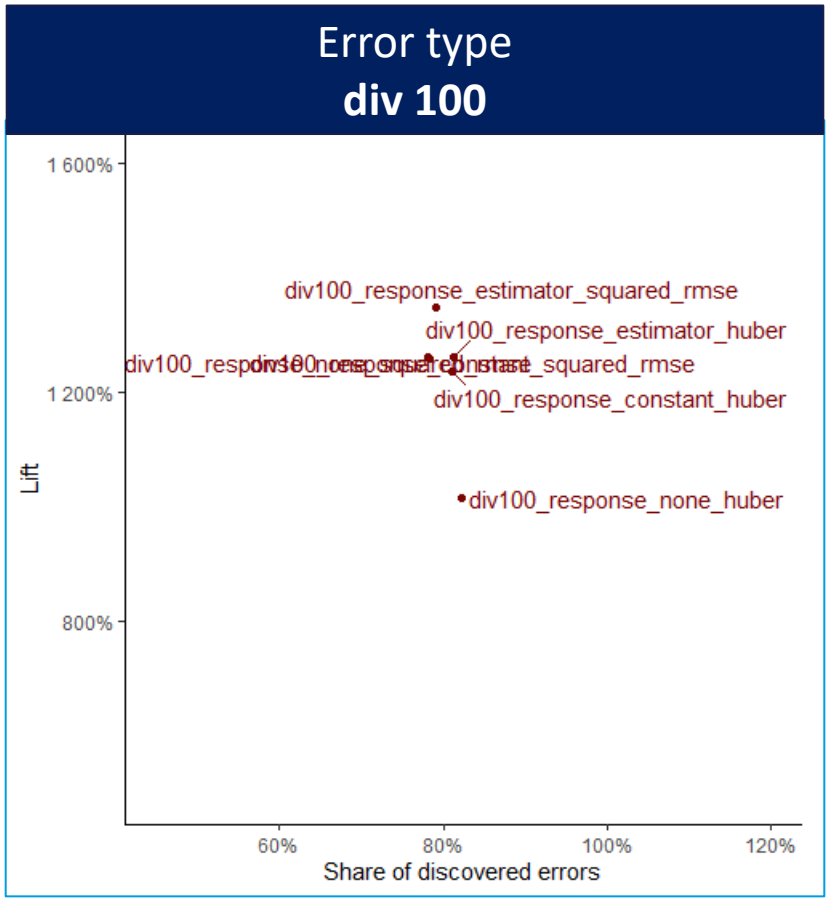
One metric is insufficient

HYPOTHESIS 1 (MISSING VALUE REPLACEMENT) AND HYPOTHESIS 2 (LOSS FUNCTION)

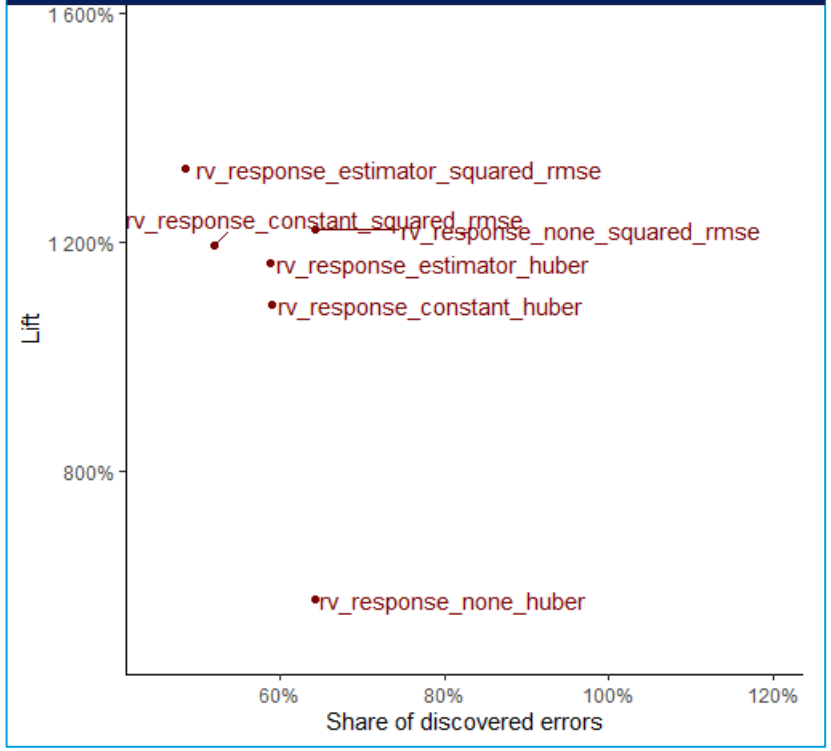
Error type
80



Error type
div 100



Error type
random value multiplication



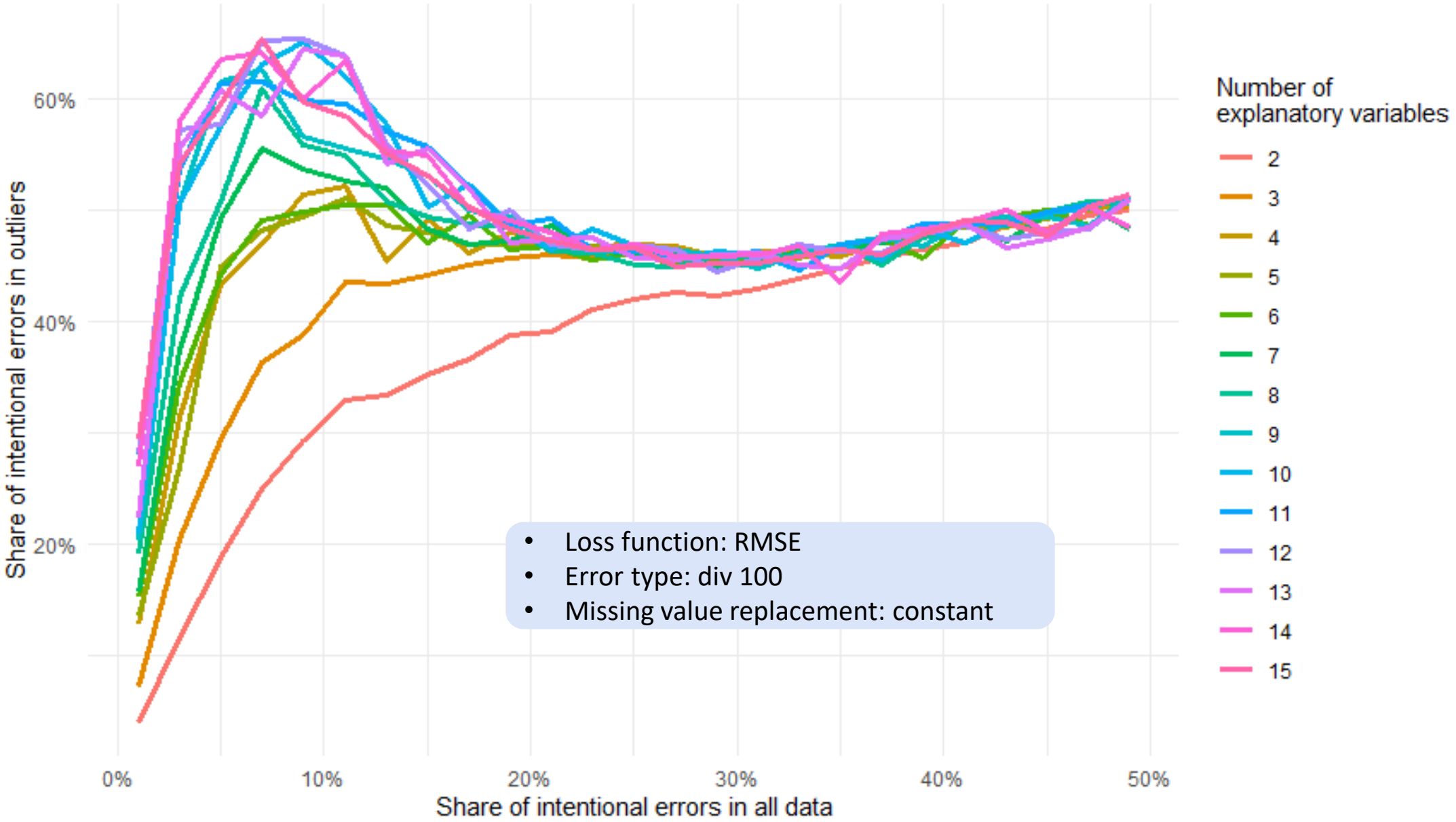
HYPOTHESIS 3 – ERROR IN ALLOCATED COLLATERAL VALUE

Error type	Missing value replacement	Loss function	Outliers as % of total	Share of discovered errors	Lift
10	none	Huber	3,5%	3,0%	0,87
10	estimator	Huber	3,0%	2,1%	0,70
10	none	rmse	1,5%	1,6%	1,05
10	constant	Huber	2,3%	1,3%	0,57
10	constant	rmse	1,3%	1,1%	0,83
10	estimator	rmse	1,1%	0,9%	0,83
rv	none	rmse	2,0%	10,7%	5,41
rv	estimator	Huber	4,5%	7,9%	1,75
rv	none	Huber	2,5%	5,8%	2,31
rv	constant	Huber	3,0%	4,3%	1,43
rv	constant	rmse	1,5%	2,1%	1,41
rv	estimator	rmse	1,3%	1,9%	1,51

Vs. 70-80 % when error in target

Vs. 10-12 when error in target

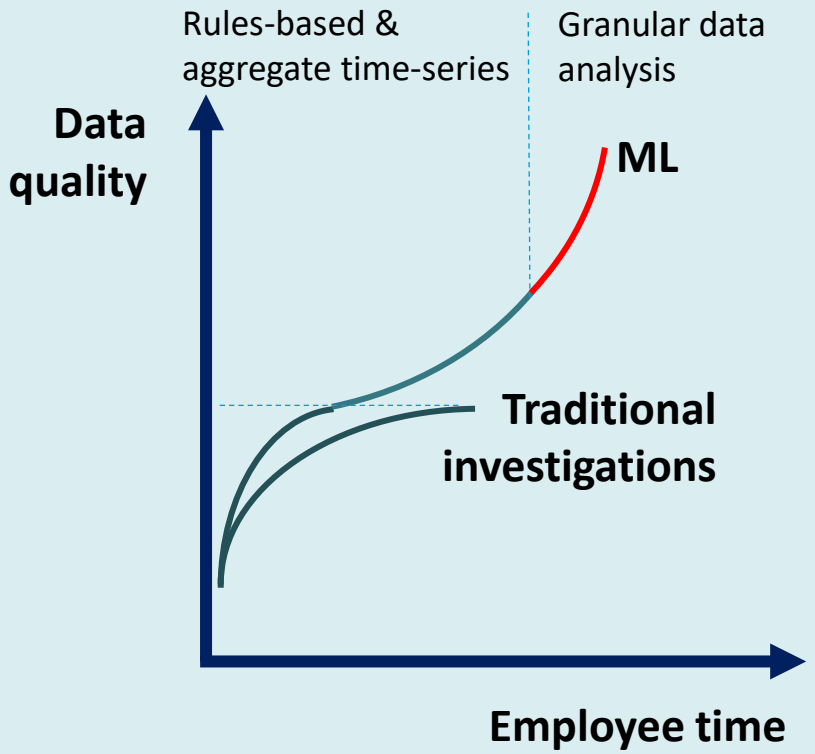
ABOVE NINE EXPLANATORY FEATURES AND AN ERROR SHARE OF AROUND 10 PERCENT EFFICIENCY STARTS TO DROP



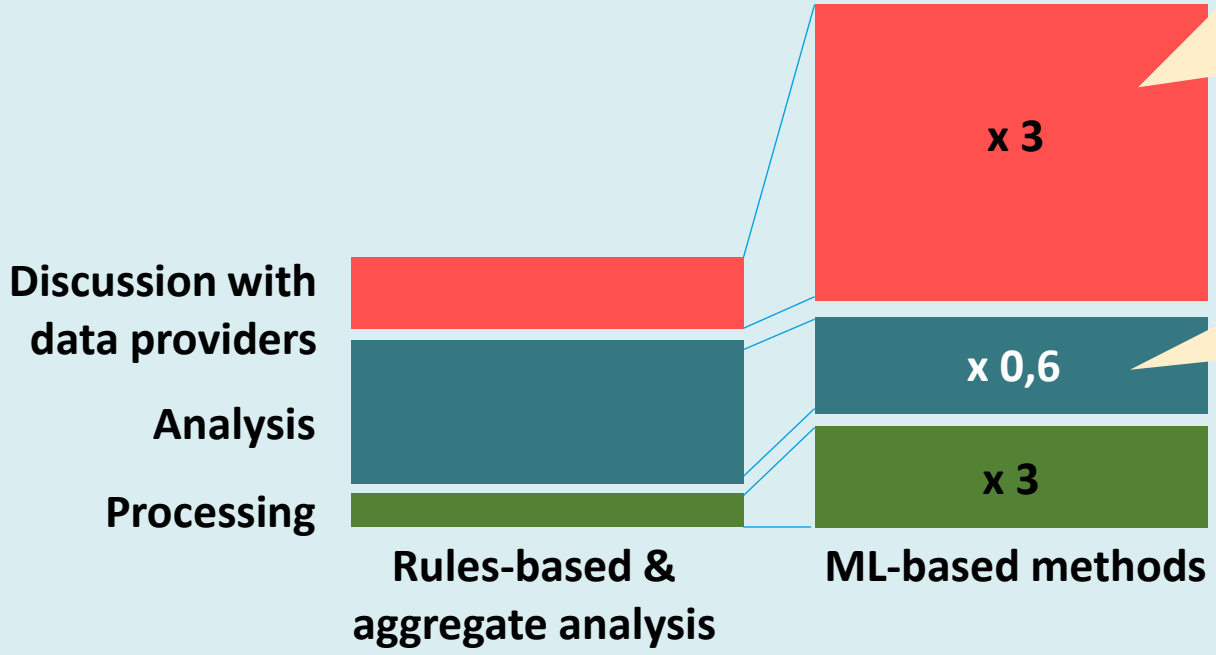
WHY YOUR COLLEAGUES WILL NOT LOVE YOU

ML-BASED DATA QUALITY TESTS CREATE MORE WORK FOR OTHERS

Data quality explodes *but only if you work with it*



Required human labour



New discussions with data providers

Automation

SUMMARY

Findings recap

- A supervised learning algorithm to flag potential data errors
- The method successfully identifies synthetic errors
- It provides hints to their location
- We also analysed various steps during the preprocessing phase (missing values and loss function) which may improve performance

Implications

- Our results helps the data providers
- The ’last mile problem’ is still there: error flags do not provide interpretation
- Our results help modellers: model predictions may be used instead of actual values

Thank you for your attention!

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

A machine learning approach for the detection of firms infiltrated by organised crime in Italy[1]

Pasquale Cariello, Marco De Simoni and Stefano Iezzi,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

BANCA D'ITALIA
EUROSISTEMA

UIF
Unità di Informazione Finanziaria per l'Italia

A machine learning approach for the detection of firms infiltrated by Organised Crime in Italy

Pasquale Cariello, Marco De Simoni, Stefano Iezzi
Financial Intelligence Unit for Italy (UIF) – Bank of Italy

Rome, October 19, 2023

Disclaimer

The views expressed in this presentation are those of the presenter and do not necessarily correspond to those of the UIF or Banca d'Italia.

Outline

1. Research context and objectives
2. The dataset
3. Classification methodology
4. Main results
5. Robustness checks
6. Conclusions and future research

Research context

- Infiltrated firms are legally registered and seemingly legitimate businesses, controlled by organized crime (OC).
- Control is exerted through (Arlacchi, 2007):
  - ownership or management of the company by OC affiliates
  - use of financial resources derived from illegal activities
  - corrupt practices
  - and in some cases, even the adoption of illegal practices such as violence and intimidation.
- (De Simoni, 2022) finds that, in some cases, the use of corporate vehicles serves as a means to launder illegally obtained capital.
- Other studies (Ravenda *et al.*, 2015; Mirenda et al.,2022) consistently indicate that infiltrated firms display distinctive features in their financial statements. Insights from this stream of literature have stimulated the development of statistical models that try to detect infiltrated from non-infiltrated firms using financial reporting data.

Objective of the project

Capitalize on the findings of recent studies to build a machine learning algorithm to identify infiltrated businesses on the basis of financial statements and other information.

Two main innovations:

1. List of about 28,000 companies with a high probability of infiltration status:
   - combination of public sources and proprietary data from the UIF.
   - one of the most comprehensive censuses of companies controlled by OC in Italy.
   - previous studies typically use lists of infiltrated companies of smaller scale and with more uncertain links to OC.
2. A dataset of Italian companies (limited liabilities, joint stock companies, and other firms legally required to deposit financial statements) observed from 2010 to 2021 incorporating data on financial statements, debts to the financial sector, employment, and owners and administrators.

The algorithm provides a "risk" score for each capital company.

Overview of the panel dataset

Year	Infiltrated firms	Non-infiltrated firms	% of infiltrated firms
		Number of records	
2010	13,231	894,738	1.48
2011	13,661	906,929	1.51
2012	13,668	902,023	1.52
2013	13,581	904,407	1.50
2014	13,690	914,611	1.50
2015	13,887	932,143	1.49
2016	13,917	942,194	1.48
2017	13,956	958,721	1.46
2018	14,073	985,884	1.43
2019	13,918	1,006,100	1.38
2020	13,001	997,485	1.30
2021	10,908	920,255	1.19
Total	**161,491**	**11,265,490**	1.43
		Number of firms	
Total	**28,570**	**1,804,278**	**1.58**

Variable selection

Drawing from the most relevant papers in this field,
we select a list of 32 variables/indicators that thoroughly characterize a firm's financial profile:

Economic Size (5 variables)
Total assets, revenues, equity, short-term liabilities, fixed assets

Equity and liquidity (7 indicators)
Cash/assets, short-term assets/short-term liabilities, cashflow, etc.

Indebtedness (4 indicators)
Granted loans over equity, Granted loans over revenues, Net debt over EBITDA, Total debt over assets

Discretionary elements (3 indicators)
Accrued liabilities over assets, Accrued incomes over assets, Inventory over assets

Investment and cost structure (5 indicators)
Tangibles/assets, Cost of rents/ revenues, Net purchases/revenues, Intermediate inputs/revenues, Capex

Employment (3 indicators)
Cost of labour over number of employees, Revenues over number of employees, Added value over number of employees

Profitability (5 indicators)
ROI, ROE, ROA, etc.

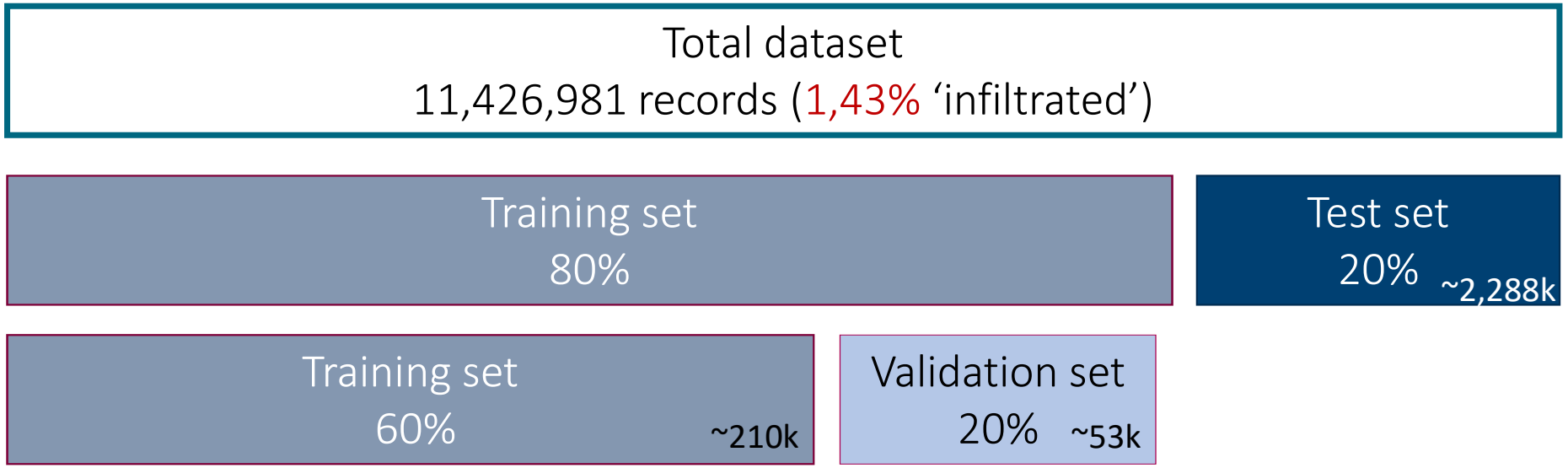
Three opacity indicators
- Ownership
- Management
- Miscellaneous attributes

Structural characteristics
Economic sector of activity (3-digit NACE code), Per capita province value added, Legal form.

Challenges in classification

- Some firms labelled as non-infiltrated may be connected to OC.
  - *Risk of potential bias in model training.*
    - ➢ *Use of the most up-to-date, reliable and comprehensive database to identify potentially infiltrated firms (UIF source; from now, for the sake of brevity, simply 'infiltrated')*
    - ➢ *The population of alleged legal firms is so large that this bias could be inconsequential.*
- Infiltrated firms in the target sample are disproportionately concentrated in Southern regions.
  - *Use of province-level dummies can improve the model's overall accuracy, but could result in an overly localized model and, therefore, less effective in classifying correctly firms in less affected provinces.*
    - ➢ *Use of a dual strategy: 1) per capita province-level value added; 2) no geographical covariate*
- High imbalance of the two classes (records of infiltrated firms 1.4%)
  - ➢ Stratified undersampling strategy with proportional allocation combined with clustering
  - ➢ Strata: combination of region and economic sector

Classification methodology - Data splitting



Data splitting is made with <u>stratified under-sampling with clustering</u>

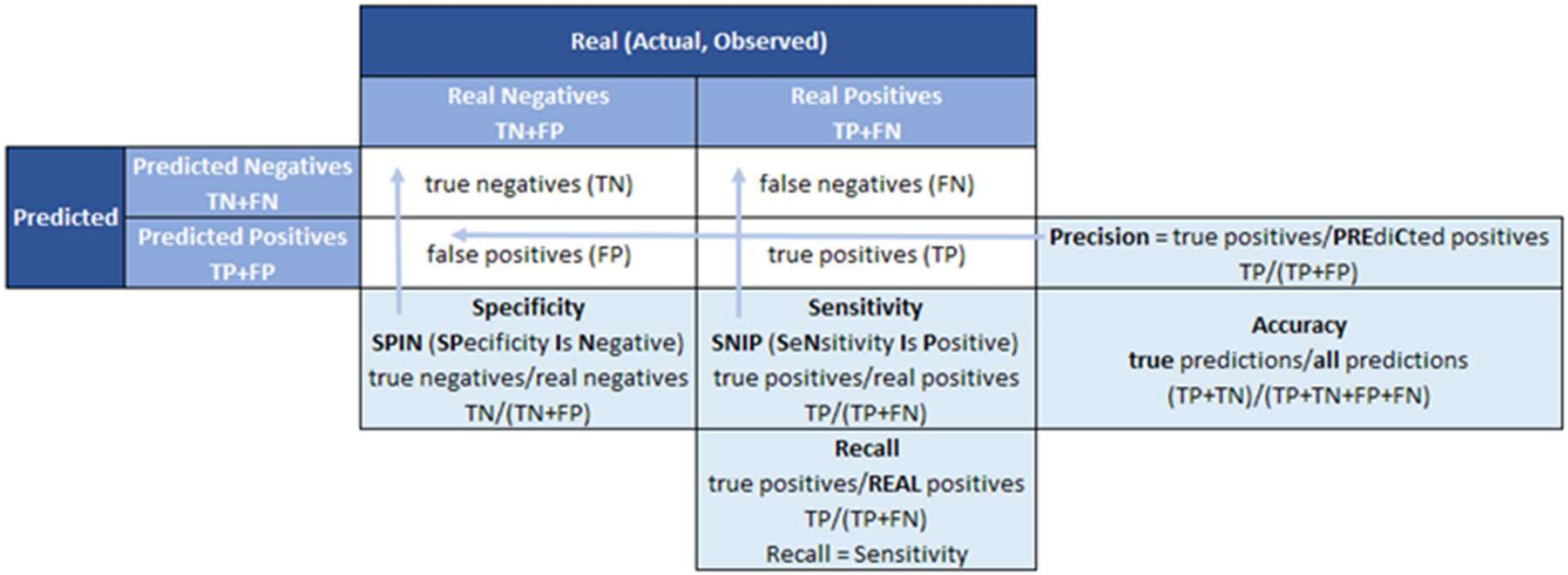
- Strata: infiltration status + region + economic sector
- Cluster: firm

Splitting is repeated 5 times to verify robustness of the results. We pick the 2nd best model.

Data Preparation & Model Calibration

- XGBoost (eXtreme Gradient Boosting) is the chosen algorithm to train the model;
- The dataset has been previously subjected to a very basic cleansing treatment in order to spot and resolve potential data inconsistencies;
- No imputation of missing data (only on Opacity variables), since Xgboost has the capability to effectively manage them;
- No transformation of the variables (only one-hot dummies on SECTOR);
- All monetary variables have been adjusted to 2021 constant prices;
- Use a randomized grids search with 50 random combinations to find the optimal combination of hyperparameters.

Some performance metrics



- Maximization of the performance measured by Recall (Sensitivity or TP rate)
- More tolerance on false positive (Specificity and Precision): it can be signal of potential infiltration in apparently legal firms.

Performance on unbalanced test set

Comparison with 'baseline' models;
(DummyClassifiers makes predictions based only on target distribution, ignoring the input features).

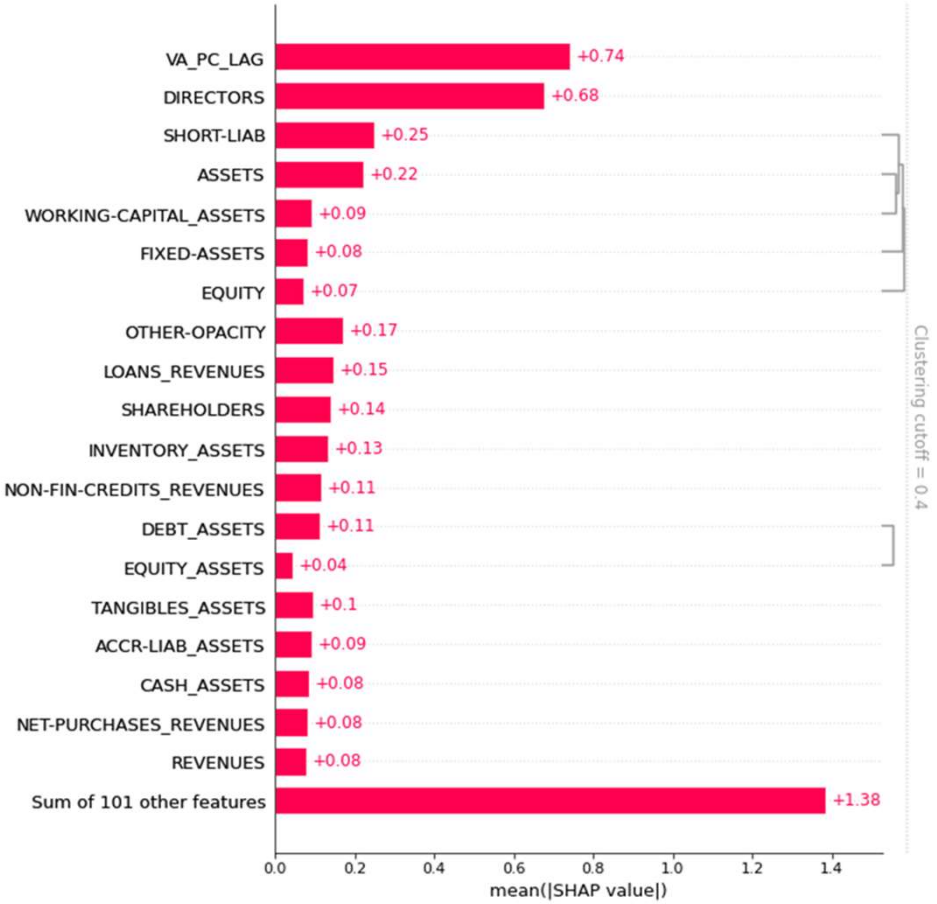
TEST SET	Accuracy	Sensitivity (Recall)	Specificity	PR-AUC	F2
XGB	0.743	0.756	0.742	0.678	0.752
Stratified	0.501	0.490	0.511	0.493	0.491
Most frequent (0)	0.986	0.000	1.000	0.492	0.000
Uniform	0.500	0.499	0.500	0.497	0.494
Less Frequent (1)	0.014	1.000	0.000	0.492	0.829

Risk score computation

Frequency distribution of estimated risk score
Year 2021

Risk score	N	%
Up to 0.50	729,433	78.3
From 0.51 to 0.80	129,799	13.9
From 0.81 to 0.95	54,969	5.9
From 0.95 to 0.99	13,916	1.5
Over 0.99	3,046	0.3
Total	931,163	100.0

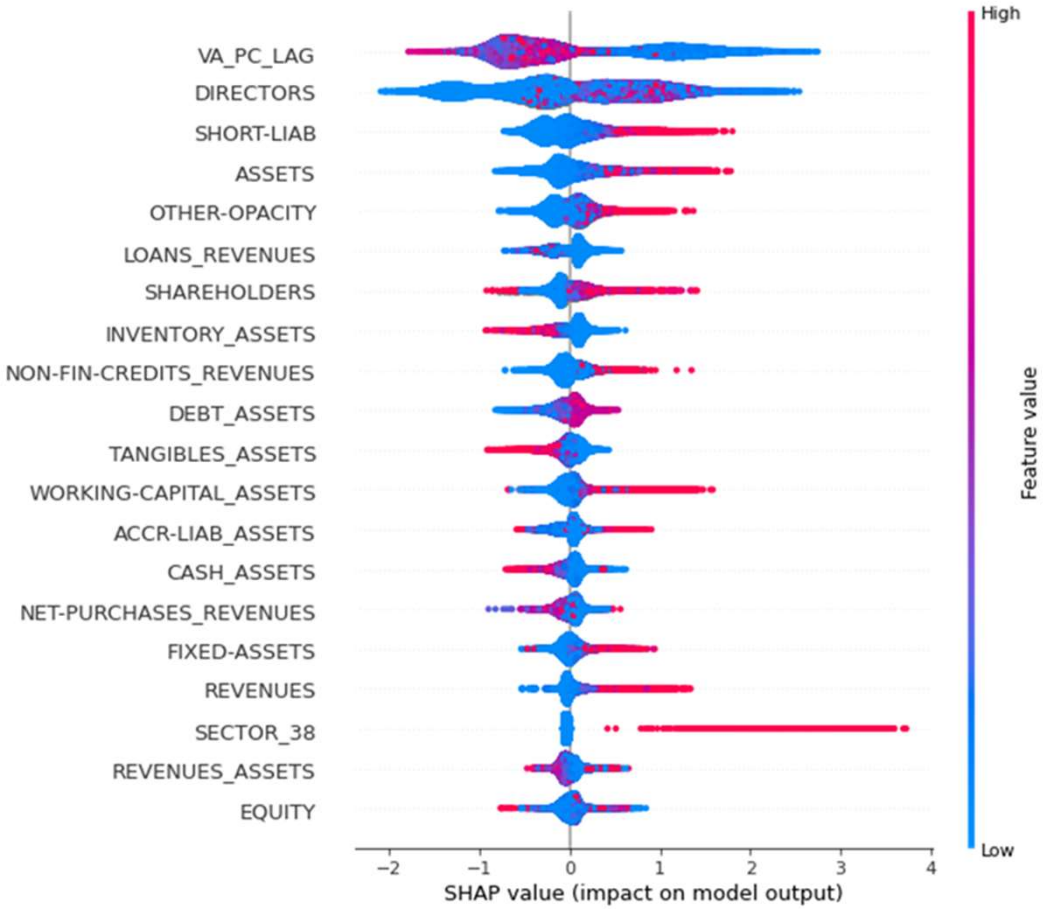
Variable importance - SHAP



Most influencial groups:

- Geo & Sector
- Opacity
- Size (Assets and Revenues)
- Equity and Liquidity

Variable importance - SHAP (2)



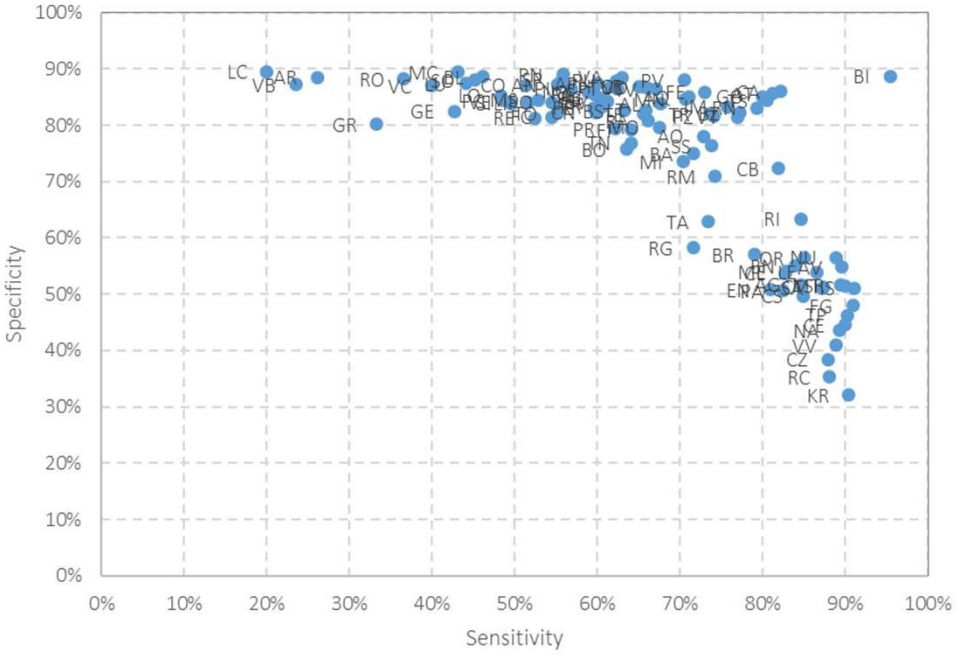
- As the "size" (assets, revenues, fixed assets, and short liabilities) increases, the probability of predicting infiltration also rises.
- Investment indicators (specifically, tangible assets over total assets) and liquidity indicators (namely, cash over total assets) have a negative effect on the likelihood of model prediction.
- Research has highlighted that companies infiltrated by organized crime are, on average, larger, more indebted and less liquid (Bianchi et al., 2022).

Test set performance by geographical breakdown

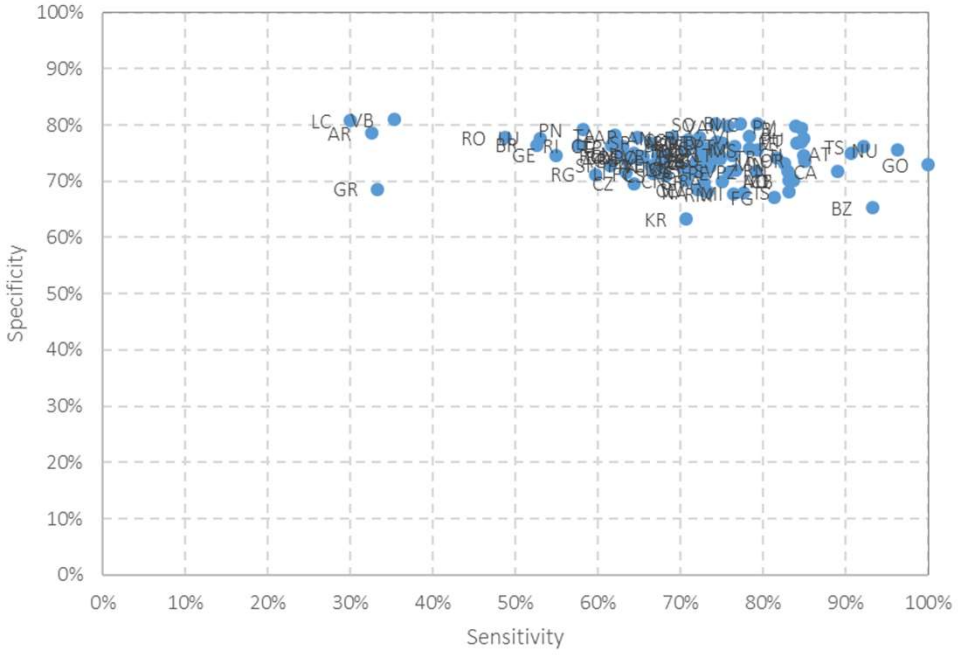
	With per-capita province-level value added		Without any geographical variable	
	Recall	**Specificity**	**Recall**	**Specificity**
North-West	0.650	0.805	0.737	0.727
North-East	0.624	0.830	0.713	0.759
Centre	0.686	0.782	0.742	0.707
South and Islands	0.858	0.582	0.711	0.692
Total	**0.755**	**0.742**	**0.723**	**0.719**

Performance on test set by province

Per capita province-level value added used as a covariate



No geographical information used as a covariate



Conclusions and future development

- We developed of a Machine Learning algorithm to detect legally registered firms *potentially* infiltrated by OC (risk score)
- Sample of about 28 thousand Italian firms considered to be infiltrated with high probability (Primary source is the UIF archives)
- Highly varied list of financial and budget indicators and variables, identified on the basis of the latest literature on criminal infiltration in real economy.
- Use of information on ownership and managers to build 3 opacity indicators.
- Infiltrated firms compared with stratified random samples of alleged legal firms in order to train and test the model.
- Model performance on test set: Recall 75.6%, Specificity 74.2%.
- Risk score resulting from the algorithm as an *additional* red flag indicator for UIF institutional functions.
- Potential scope for sharing the risk score information with other selected actors involved in AML and/or the fight against economic crime.
- Opportunities for future research: alternative methods to manage unbalancing; adopt a dynamic approach; variable transformations to cancel out geographic/sectoral effects.

References

- Arlacchi P. (2007), La Mafia imprenditrice, Il Saggiatore.
- Ravenda, D., Argilés-Bosch, J.M. and Valencia-Silva, M.M. (2015), Detection Model of Legally Registered Mafia Firms in Italy. *European Management Review*, 12: 23-39.
- Mirenda, Litterio, Sauro Mocetti, and Lucia Rizzica (2022), "The Economic Effects of Mafia: Firm Level Evidence." *American Economic Review*, 112 (8): 2748-73.
- De Simoni M. (2022), The financial profile of firms infiltrated by organised crime in Italy. *UIF, Quaderni dell'antiriciclaggio, Collana Analisi e Studi*.

Thank You
For Your Attention

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Research for all: exploring machine learning applications in generating synthetic datasets[1]

Carmelita Esclanda-Lo, Gabriel Masangkay,
Chelsea Anne Ong and Rossvern Reyes,
Bangko Sentral ng Pilipinas

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Research for All: Exploring machine learning applications in generating synthetic datasets

Carmelita Esclanda-Lo, Gabriel Masangkay, Chelsea Anne Ong, Rossvern Reyes[1]

Abstract

Sharing granular data with internal and external parties is becoming more challenging for central banks due to stricter laws, regulations, and growing concerns about data privacy. These factors hinder or slow the progress of collaborative and data-centric research and innovation. However, with the advancements in data generation, it is now possible to generate synthetic datasets, which mimic the statistical characteristics of the actual data but mask or hide any private information contained therein, using artificial intelligence (AI) methods.

In this paper, we compare the results of several AI methods, *i.e., the Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks (GANs), and Tabular Variational Autoencoders (TVAE)*, in generating synthetic granular and tabular datasets, using the 2023 Quarter 1 Consumer Expectation Survey (CES) dataset of the Bangko Sentral ng Pilipinas (BSP).

The results demonstrate that TVAE generates synthetic tabular data with the highest fidelity, privacy, and utility among the other algorithms tested. Conversely, GANs performed poorly in terms of data fidelity and utility.

Keywords: Synthetic Data, Artificial Intelligence, Machine Learning, Data Sharing

JEL classification: C45 C81

[1] This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" which was held on 17-19 October 2023 in Rome, Italy. Carmelita Esclanda-Lo (esclandacg@bsp.gov.ph), Gabriel Masangkay (masangkayga@bsp.gov.ph), Chelsea Anne Ong (ongcs@bsp.gov.ph), and Rossvern Reyes (reyesros@bsp.gov.ph) are from the Department of Economic Statistics of the BSP. The views expressed are those of the authors' only and not those of the BSP.

Table of Contents

1. Introduction 3
    Background 3
    Objectives 3
    Importance of Synthetic Data Generation in Research and Applications 3
2. Synthetic Data Generation Methodologies 4
    Methodologies in Tabular Synthetic Data Generation 4
    Use of Synthetic Data Generation in Central Banks 4
3. Data 5
4. Methodology 6
    Data Collection and Processing 6
    Synthetic Data Generation Libraries 6
    Algorithms for Synthetic Data Generation 7
    Evaluation Metrics 8
    Best Synthetic Data Evaluation Criteria 11
5. Results 11
    Data Fidelity 11
    Data Utility 14
    Data Privacy and Disclosure Risk 16
    Evaluation of Best Synthetic Dataset 17
6. Conclusion and Recommendations 18
References 19
Appendix A 21
Appendix B 25

1. Introduction

Background

In advancing data-centric research and fostering innovation, central banks are progressively embracing the idea of sharing granular data with researchers by launching data-sharing initiatives, which allow data sharing under conditions that protect sensitive information with confidentiality. Currently, data-sharing frameworks are established to protect sensitive information. Still, these processes are often tedious and stringent, burdening data producers (central banks) and requesters due to strict data privacy and confidentiality regulations.

To address data-sharing concerns, the BSP developed its Data Governance Manual, which specifies, among many provisions, the rules on sharing data with external parties and protecting the privacy of its data subjects. As part of its data governance initiatives, the BSP crafted its data masking guidelines outlining rules on obfuscating personal identifiable information (PII) and sharing data with external parties.

Other Philippine government agencies also have methods of disseminating data to the public. Most notably, the Philippine Statistics Authority (PSA) makes the microdata from their surveys available through the PSA Data Archive. As part of the request, data requesters are asked to submit their application to the agency head stating the purpose of their request. The PSA will provide a subset of the requested microdata upon its approval. For some data, access could only be done via onsite access to the PSA's Data Enclave Center.

The emergence of synthetic data through AI offers a potential solution for improving data sharing and addressing these data-sharing concerns. Synthetic datasets produced by AI exhibit identical mathematical or statistical characteristics as their original counterparts while preserving privacy.

Objectives

This study aims to explore and compare various machine learning (ML) methodologies to generate another dataset with mathematical and statistical structures similar to the original dataset. More precisely, this research aims to evaluate the performance, check the limitations, and dissect the practical applicability of various AI-based models, including Gaussian Mixture Models (GMM) and advanced Generative AI models, such as GANs and Variational Autoencoders (VAE), in creating synthetic granular and tabular data that ensures utility and privacy. In addition, this study aims to show a preliminary pipeline for implementing and evaluating the said ML models.

Importance of Synthetic Data Generation in Research and Applications

Assefa et al. (2020) highlighted the importance of effective synthetic data generation in the financial sector. Since financial data includes PII, data sharing is highly

restricted. Synthetic data generation addresses the challenges of publishing stream data by enhancing data privacy and security, facilitating model development and testing, and overcoming actual financial data unavailability. Furthermore, it discusses the complexities of producing realistic synthetic financial data, including the need for data generators to accurately capture the underlying patterns, correlations, and distributions of actual financial data. Addressing these challenges is essential to ensure the utility of synthetic data. For central banks, adopting synthetic data can be a strategic solution for promoting research collaborations while preserving data privacy.

2. Synthetic Data Generation Methodologies

Methodologies in Tabular Synthetic Data Generation

Pathare et al. (2023) generated synthetic data on unbalanced, balanced, numerical-only datasets, categorical-only datasets, and datasets with a mix of numerical and categorical attributes using multiple models, i.e., Conditional Tabular Generative Adversarial Network and Classification and Regression Trees (CART), with results based on different comparison parameters which are accuracy, propensity score, log-cluster, and execution time. For all types of datasets analyzed in the study, CART generates data with the highest quality, while Bayesian networks performed the worst among other models.

One of the overarching objectives of generating synthetic datasets is to protect the PII of data subjects while maintaining the usefulness or utility of the data. Little et al. (2021) assessed the disclosure risk of the synthetic data they generated through CART and GANs. They found that the synthetic data produced via CART had the highest utility but also had the highest risk of disclosure. On the other hand, Table GAN synthetic data produced the lowest risk but also had the lowest utility. This finding corroborates an important concept regarding utility and risk as mentioned in the United Nations Economic Commission for Europe's (UNECE) (2022) Synthetic Data for Official Statistics: *a trade-off exists between utility and disclosure risk*, i.e., as the utility of the synthetic data increases, the disclosure risk increases exponentially.

Use of Synthetic Data Generation in Central Banks

The integration of synthetic data methodologies in central banking is in its nascent stage. Synthetic data generation was mentioned in the first IFC report on data sharing (IFC, 2015), where it was cited that synthetic data could *"transform the original confidential microdata into artificial microdata with the same statistical properties that third parties can use."*

On the other hand, National Statistics Offices (NSOs) worldwide have developed and operationalized synthetic datasets for various use cases (UNECE, 2022). Some NSOs have developed synthetic data to improve the efficiency of data-sharing processes. For instance, Statistics Canada developed a version of a census-based database for testing and running the new dynamic micro-simulation model of the Canadian retirement and income system. The synthetic database would enable

policymakers to experiment and model changes to the Canada Pension Plan. The micro-simulation model would use the original data for the final analysis.

3. Data

We used the First Quarter 2023 Consumer Expectation Survey (CES) dataset, formatted in Microsoft Excel, to generate synthetic datasets. The CES, which the BSP conducts, is a nationwide quarterly survey on consumers' sentiments, i.e., Philippine household sentiments, on family income, financial situation, and economic condition of the country for the current quarter, next quarter, and next 12 months. With over 1,000 columns or variables, which reflects the diverse sets of survey questions, pre-processing was needed to transform the CES dataset into a dataset that can be used with AI methods.

Sample CES variables

A total of 36 variables is analyzed in this study

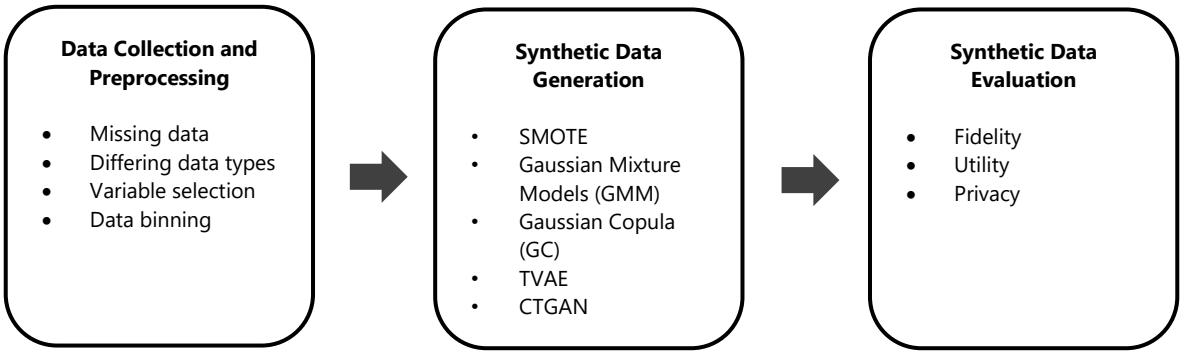
Table 1

Variables	Description	Values
AGE	Age	0 - 100
INCOME	Income Group	Low, Middle, High
SEX	Sex	Male, Female
C5C	Inflation Rate in the Current Quarter	Less than 0%, 0.1% - 1.9%
E1S	Has Family Savings	Yes, No
B1S	Present Family Situation	Better, Same, Worse

4. Methodology

Methodology Overview for Synthetic Data Generation

Pipeline for generating and evaluating synthetic datasets. Figure 1



To achieve our objectives, we created a pipeline for generating and evaluating synthetic datasets, as shown in Figure 1. We first collected and pre-processed the CES dataset, and selected 36 variables based on a simulated research problem on identifying the determinants of households' inflation outlook. Subsequently, we generated the synthetic datasets through various statistical and generative machine learning models using open-source Python Data Synthesizer libraries (e.g., Synthetic Data Vault (SDV) and YData Synthetic). Lastly, we evaluated the generated synthetic datasets using metrics based on three key dimensions or qualities: fidelity, utility, and privacy.

Data Collection and Processing

With reference to the study of Basilio (2010) on the determinants of households' inflation outlook, we limited the number of variables of the CES from over 1,000 to 36 variables for computational efficiency. As listed in Appendix A, these variables include the respondents' demographic characteristics and outlook on various economic and financial indicators.

After reducing the variables, we cleaned and pre-processed the data. We then transformed columns containing more than one data type (e.g., a combination of string, float, integer) to its proper data type. Lastly, we binned the age variable to reduce the effects of outliers.

Synthetic Data Generation Libraries

We used open-source Python libraries, such as the SDV and YData Synthetic libraries, to generate synthetic datasets.

The Synthetic Data Vault (SDV)

The SDV is an ecosystem designed by DataCebo, Inc. for synthetic data generation and evaluation using ML (i.e., ranging from classical statistical to deep learning methods). By using SDV, we could define the constraints for pre-processing the selected variables and compare the resulting synthetic datasets with the original.

In the SDV implementation, metadata creation is required where the data type of each variable must be explicitly specified. Then, the generated model creates samples of synthetic data that retain the format and the mathematical properties of the original or actual dataset. In addition, SDV has a conditional sampling feature that allows the generation of hypothetical scenarios by fixing values to extreme cases or imputing data.

YData Synthetic

YData Synthetic is an open-source Python package that generates synthetic tabular and time-series data using classical and state-of-the-art generative models.

Like in SDV, we defined first the numerical and categorical columns before fitting the data to the model. A sample of the synthetic data is then generated based on the size of the actual dataset. The models chosen were CTGANs and GMMs.

Algorithms for Synthetic Data Generation

We utilized and compared the performance of five algorithms in generating synthetic data of the 36 variables from the CES dataset. These algorithms are discussed below.

Synthetic Minority Over-Sampling Technique

SMOTE is an oversampling approach where new instances of the minority class are created by joining all or any of the *k* minority class nearest neighbors to balance the dataset (Chawla, 2002). Suppose we need to double the size of the dataset. In that case, only two neighbors out of the k-nearest neighbors are chosen, and a synthetic sample is created at a random point somewhere between two examples in the feature space.

Gaussian Copula (GC)

GC relies on the fact that separate distributions can be modeled from the joint distribution. Mathematically, this works by applying the Probability Integral Transform, a mathematical method used to transform any random variable with a cumulative distribution function (CDF) into a uniform distribution.

Although the GC algorithm is designed for only numerical data, this synthesizer converts other data types using Reversible Data Transforms (RDTs). Some controls in synthetic data can be made by adjusting the parameters, such as setting the minimum or maximum boundaries, rounding off the values, and setting the distribution shape of numerical columns similar to the actual data.

Conditional Tabular Generative Adversarial Networks (CTGAN)

CTGAN is a variation of the GAN-based method for modeling tabular data. GAN deep learning models have two main components (a Generator and a Discriminator). They are trained through an adversarial learning process where the generator and discriminator compete and improve iteratively. New data is generated by inputting random noise into the generator, introducing interesting variability to the synthetic data.

This algorithm works ideally for data with complete values. For the SDV implementation, in addition to the parameters of GC models, the CTGAN synthesizer has other parameters like the number of epochs or the number of times to train the GAN to improve the model, batch size, learning rate, and embedding dimensions.

Tabular Variational Autoencoder (TVAE)

A tabular variational autoencoder (TVAE) is a mathematical model for compressing and representing tabular data. It consists of an encoder that transforms the input data into a distribution characterized by mean (μ) and standard deviation (σ) parameters, introducing a degree of randomness. During training, the model minimizes an objective function that encourages accurate reconstructions and diverse representations. This approach enables the TVAE to learn a flexible and probabilistic encoding of tabular data, allowing for nuanced understanding and generation of meaningful representations while considering uncertainties in the information.

Like CTGANs, the TVAE works best for data with complete values. Parameters specific to TVAE are the batch size, hidden layer size, regularization, and loss factor.

Gaussian Mixture Models (GMM)

GMMs assume that the data is represented as a mixture of Gaussian distributions characterized by its mean and covariance. GMMs generate synthetic data via sampling from the learned distribution, ensuring that the new data has the same characteristics as the actual data.

Evaluation Metrics

Data Fidelity

Data fidelity pertains to how well synthetic data captures the information in the real dataset. We followed the framework of Platzer and Reutterer (2021) to measure data fidelity, wherein distributions and correlations are measured and compared for synthetic and actual or real data.

- Histogram Plots and Bray-Curtis Similarity Score

Histogram plots were generated for real and synthetic datasets to illustrate the visual differences. An excellent example of a synthetic dataset is one that keeps the shape and distribution close to the original dataset.

The Bray-Curtis Distance is computed to quantify the visual differences between the histogram plots. This measure is used in biology to measure the distances of two compositions between two sites. In this case, these sites are the real and the synthetic

data. The Bray–Curtis similarity score is bounded between 0 and 1 and is obtained by the following equation:

$$Bray - Curtis\ similarity = \frac{\sum_{i=1}^{n} \| A_i - B_i \|}{\sum_{i=1}^{n} \| A_i + B_i \|}$$

where A_i and B_i represent vectors A and B, respectively. A higher Bray-Curtis similarity score indicates that the count distribution of the synthetic data is closer to the actual data.

- Correlation and Cosine Similarity Score

Meanwhile, the correlation coefficients between the variables were computed to determine whether relationships within the original dataset were preserved. To visually depict the correlation of variables, we plotted the Cramér's V values on the heatmap.

Cramér's V is an extension of Pearson's chi-squared test for independence. It assesses the strength of association or correlation between two categorical variables in a contingency table. A zero value indicates no association, and a one indicates perfect association between the variables. It is calculated by taking the square root of the chi-squared statistic, which is then divided by the product of the sample size and one less than the minimum dimension:

$$V = \sqrt{\frac{\delta^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

where δ^2 is the phi coefficient, $\chi^2$is the chi-squared statistic, *n* is the number of observations, *k* is the number of columns, and *r* is the number of rows.

Additionally, the cosine similarity scores were computed to assess the similarity between the Cramer's V values of the real and synthetic data by computing the cosine of the angle between the two vectors, which fall from -1 and 1. A value closer to 1 denotes higher similarity. The following equation defines cosine similarity:

$$Cosine\ similarity\ (\cos\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} = \frac{\mathrm{A} \cdot \mathrm{B}}{\| \mathrm{A} \| \| \mathrm{B} \|}$$

where $\mathrm{A} \cdot \mathrm{B}$ represents the dot product of vectors A and B, while $\| \mathrm{A} \|$, $\| \mathrm{B} \|$ represent the Euclidean norm (magnitude) of vectors A and B, respectively. A higher cosine similarity score indicates that the distribution and correlation of the synthetic data are closer to the actual data.

- Statistical Similarity Score

To account for the combined histogram and correlation scores, we computed the statistical similarity score as the average of the histogram Bray-Curtis similarity score and the cosine similarity score of the Cramer's V values.

Data Utility

Synthetic data generation methodologies such as removing information and adding noise can reduce the usability of the data. However, the generated synthetic data should yield an ML model performance similar to the actual data. To this end, we

simulated the research problem of Basilio (2010), which identified the determinants of households' inflation outlook.

We developed a multi-class classifier to predict the range of inflation rates in the next 12 months given in the CES data. Python's PyCaret package, an open-source library that automates ML workflows, was utilized for simplicity and ease of implementation.

The algorithms used were K-nearest neighbors, tree-based models (e.g., decision tree, random forest, gradient boosting method), linear models (e.g., logistic regression, linear support vector machines, ridge), and Bayesian models (e.g., Naïve Bayes, Bernoulli). We further improved the top-performing ML model by hyperparameter tuning and assessed this through the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score, more commonly called the Area Under the Curve (AUC). Furthermore, we supplemented the results by getting the model's feature importance score.

Data Privacy

Carlini et al. (2021) discussed membership inference attacks in the context of privacy and ML, aiming to determine whether specific actual data points are part of a model's training dataset, which may potentially compromise privacy. One key metric discussed in the paper is the balanced attack accuracy, which assesses how often an attack correctly predicts membership on a combined dataset where each data point is labeled as either real (0) or synthetic (1). Despite its widespread use in various papers, the accuracy metric has limitations. Since it is an average-case metric, it fails to consider the costs of incorrect predictions. For institutions like central banks, having incorrect predictions could expose the actual data, which poses security and privacy risks. Another evaluation metric is the AUC score. An AUC score near 50% indicates random guessing, suggesting that the model cannot effectively distinguish between the actual and synthetic data, thereby preserving privacy. However, this approach averages all false-positive rates, even in high error rates. As a result, the likelihood of revealing the real data is higher. To assess data privacy more effectively, the study discussed using precision, which represents the percentage of true positives among all predicted positives. A high precision score reduces false positives while simultaneously increasing true positives.

Hence, we thoroughly evaluated our synthetic data using the three metrics: accuracy, AUC score, and precision. Accuracy assesses the overall correctness of the model, AUC measures the model's ability to distinguish between classes, and precision focuses on minimizing false positives. Collectively, these metrics aim to ensure data privacy, especially in scenarios where accurately identifying the actual data points from synthetic ones is essential.

Best Synthetic Data Evaluation Criteria

Synthetic Data Evaluation Matrix

Scores: 3 – Excellent, 2 – Fair, 1 - Poor

Table 2

	Data Fidelity	Data Utility	Data Privacy	Decision
3 – Excellent	Has a Statistical Similarity Score of 0.95 or higher	Overall average difference of 0.05 or less	Privacy Score is higher than 0.9	Synthetic data can be used for research.
2 – Fair	Statistical Similarity Score higher than 0.90 but lower than 0.95	Overall difference of 0.05 to 0.10	Privacy Score is higher than 0.8 but lower than 0.9	Synthetic data can be used for research but with conditions
1 – Poor	Has a Statistical Similarity Score of 0.90 or lower	A difference of more than 0.10 Other conditions not satisfying any of the above	Privacy Score is lower than 0.8	Not for research use. Test another algorithm

[1] This decision matrix is devised for evaluating the best synthetic data.

Sources: asq.org/quality-resources/decision-matrix

Table 2 shows a decision matrix for ranking the algorithms regarding data fidelity, utility, and privacy to determine the best algorithm for generating synthetic data. The rubric ranges from one to three, with three indicating excellent performance in the corresponding evaluation criteria.

5. Results

This section evaluates the synthetic data generated from the CES dataset using various open-source Python Data Synthesizer libraries (e.g., SDV, YData Synthetic, and SMOTE). Using machine learning, we demonstrated the utility of synthetic data for research and provided a quantitative evaluation focusing on the fidelity, utility, and privacy of the generated synthetic data.

Data Fidelity

In assessing the quality of the generated synthetic data in terms of similarity in statistical properties, we calculated the statistical similarity score as the average similarity in the histogram and correlation heatmap.

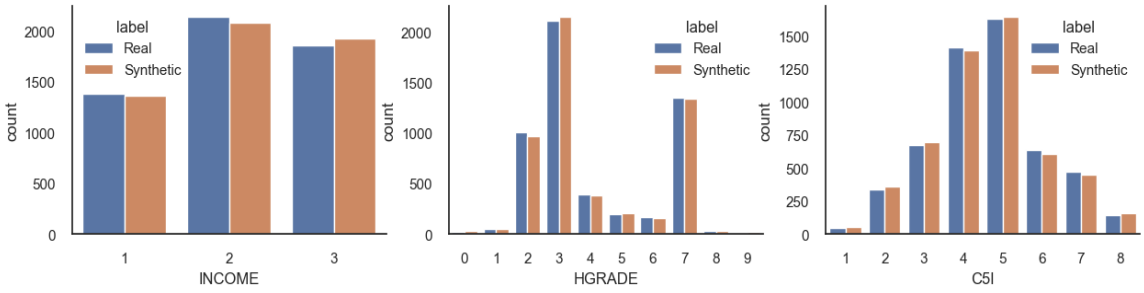
Histogram Plots

We analyzed a total of 36 variables, which sample histogram plots are shown in Figure 2. The GC model best captured the statistical properties of the original dataset. On

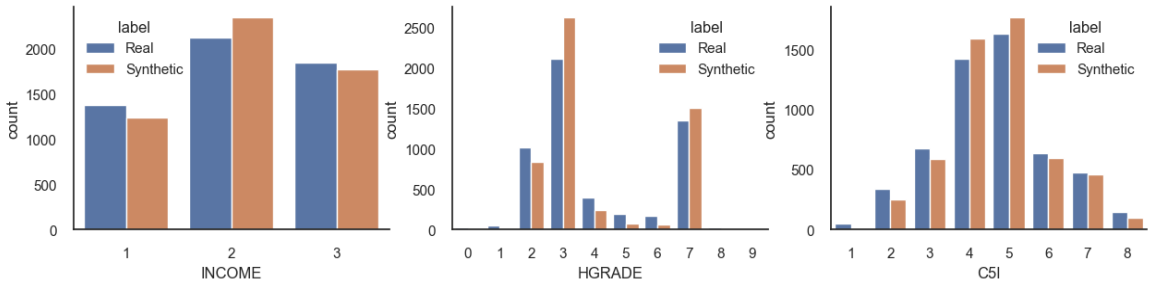
the other hand, SMOTE had the lowest similarity score among the tested algorithms. The full results are shown in Appendix B.

Histogram of Real and Synthetic Data — Figure 2

A. Gaussian Copula (GC)



B. SMOTE



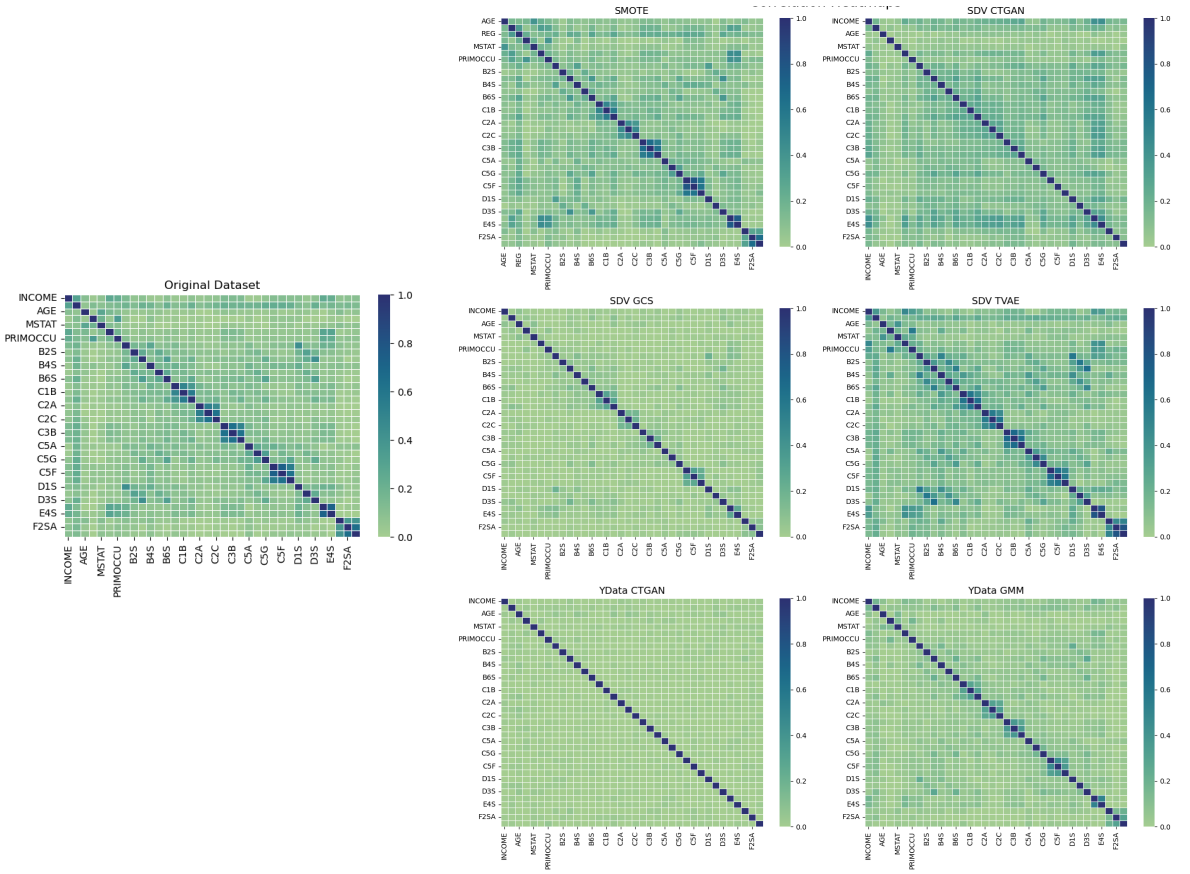
Correlation heatmaps

In addition to the generated histogram plots, we also evaluated the correlation between variables. Figure 3 shows the heatmaps of Cramer's V correlation for the original and synthetic datasets. The heatmaps of SMOTE and SDV TVAE synthetic datasets highly resemble the original dataset, given that these captured specific portions of highly correlated variables (in dark blue). Although slightly faint, the YData GMM and SDV GC could identify correlated variables correctly.

Correlation Heatmaps of Real and Synthetic Datasets — Figure 3

A. Real Dataset

B. Synthetic Datasets



Statistical Similarity Score

We computed the similarity scores to support the histogram plots and the Cramer's V values heatmap for the 36 variables, , as shown in Table 3. Each row in the table shows the similarity of the values generated from the real and the corresponding synthetic data. Then, we computed the mean of the Bray-Curtis and Cosine Similarity scores to derive an overall *statistical similarity score*.

Statistical Similarity Score[1]

Values in parentheses are the standard deviation of the computed scores. Table 3

Name of Python Library	Algorithm	Histogram Similarity	Correlation Similarity	Overall Statistical Similarity Score
YData	GMM	0.9696 (±0.0247)	0.9306	0.9501
SDV	TVAE	0.9381 (±0.0427)	0.9171	0.9276
In-house	SMOTE	0.8882 (± 0.0599)	0.9666	0.9274
SDV	GC	0.9893 (±0.0062)	0.8377	0.9135
SDV	CTGANs	0.8884 (±0.0699)	0.7226	0.8055
YData	CTGANs	0.8974 (±0.0621)	0.4023	0.6499

[1] Statistical Similarity Score is the average of the computed histogram and correlation similarity scores.

Sources: Authors' computations

Most algorithms produced a statistical similarity score above 0.9 except for CTGANs, indicating poor data fidelity and inability to resemble the original data's statistical properties.

In contrast, for histogram similarity, YData GMM datasets yielded the highest similarity score (0.9893), showing that the algorithm best preserved the count distribution of the variables. In contrast, SMOTE and CTGANs presented the lowest histogram similarity, coinciding with the histogram plots.

Meanwhile, SMOTE had the highest correlation similarity (0.9666), corroborating its high cosine similarity scores heatmap vis-à-vis the original.

For the combined statistical similarity, GMM and TVAE presented the most accurate fully synthetic data compared to the rest of the algorithms with scores of 0.9501 and 0.9276. Closely following is SMOTE. However, it presented a low histogram similarity score, making it unfit for exploratory data analysis. On the other hand, GC datasets preserved the histogram well, but the correlation of variables was farther than the actual dataset, with a cosine similarity score of 0.8377.

Data Utility

We measured the usability of the synthetic dataset by evaluating the performance of each algorithm compared to the original data through these metrics: accuracy, AUC,

recall, precision, F1 score, Kappa, and the Matthews Correlation Coefficient (MCC)[2]. We also evaluated the Feature Importance to supplement the machine learning results.

In Table 4, we compared the performance of ML models trained on both synthetic and actual datasets using a Gradient Boosting Classifier (GBC) since this gave the best performance compared to other ML models (e.g., naïve Bayes, k-neighbors, logistic regression), and validated on test data from the actual dataset.

SMOTE, TVAE, and GMM performed well when compared with the actual dataset. Meanwhile, GC- and GAN-based synthetic datasets provided significantly worse performance on all metrics and, thus, is not recommended for research purposes.

Xu et al. (2020) generated synthetic tabular data using GANs and found that TVAE outperforms GANs in classification and regression tasks. However, GANs still offered several favorable attributes, making it easier to learn data distributions better than other models, such as Bayesian networks.

Machine Learning Performance of Synthetic Datasets[1]

Utility is evaluated based on the AUC score — Table 4

Dataset/Python Library	Algorithm	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Actual Data		0.7852	0.9493	0.7852	0.7913	0.786	0.7303	0.7313
Synthetic Data:								
	SMOTE	0.7459	0.9259	0.7459	0.751	0.7456	0.6815	0.6828
SDV	TVAE	0.7247	0.9217	0.7247	0.7255	0.7239	0.6529	0.6531
YData	GMM	0.7171	0.9039	0.7171	0.7125	0.7123	0.6411	0.6419
SDV	GC	0.4276	0.739	0.4276	0.3794	0.3844	0.2539	0.2593
SDV	CTGAN	0.438	0.7121	0.438	0.4174	0.3901	0.2681	0.2846
YData	CTGAN	0.2686	0.5433	0.2686	0.2347	0.2271	0.0071	0.0076

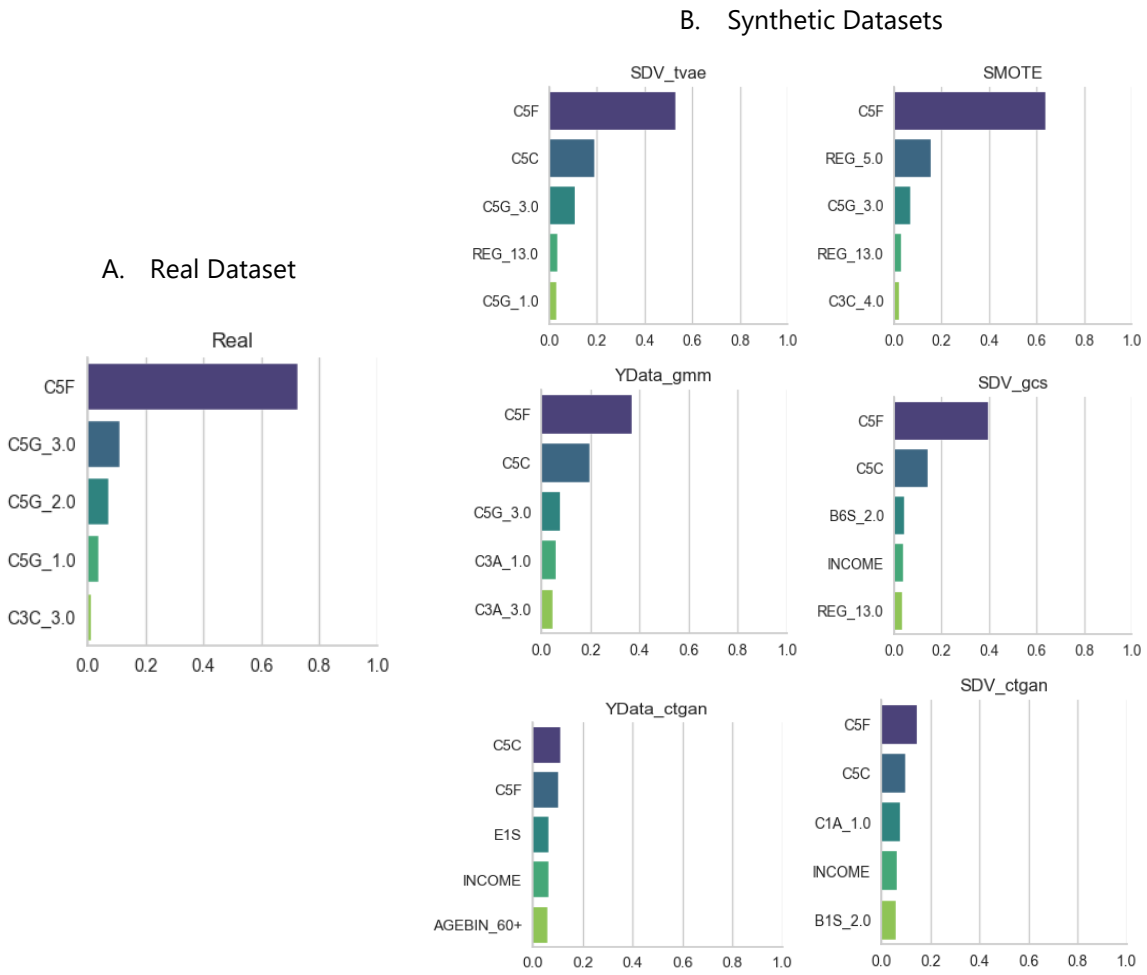
[1] The ML performance is evaluated based on the difference between the actual and synthetic data performances. The lower the difference, the better the performance.

Sources: Authors' computations

Furthermore, the feature importance scores helped interpret machine learning performance by determining the relative significance of each feature within the model. This capability allowed us to pinpoint key variables and enhance our understanding of the problem. We condisder the utility of synthetic data to be high when it successfully preserves both the order and magnitude of feature importance.

[2] The metrics are based on the default PyCaret machine learning performance evaluation. Kappa or Cohen's Kappa is an evaluation metric quantifying the level of agreement between two or more raters (or models) in the classification of categorical data, while considering the agreement between two or more raters. The Matthews Correlation Coefficient or MCC is an evaluation metric measuring the quality of predictions by considering both true positive and true negative results while accounting for the balance between classes.

Feature Importance Scores of Real and Synthetic Datasets Figure 4



TVAE, GMM, and SMOTE matched the top predictor variables (C5F and C5G) of the ML model developed using the actual dataset, aligning with the observed ML performance in Table 4.

In contrast, we found that GC and CTGANs exhibited notably distinct feature distribution patterns compared to the actual and other synthetic datasets examined in this study.

Data Privacy and Disclosure Risk

The membership inference score gauges the susceptibility of individual data points to membership inference attacks. Even without access to the actual dataset, attackers can reveal the data used to create the synthetic data, posing a risk of re-identification and privacy breaches.

Here, we compared three metrics to assess the privacy of synthetic data: accuracy, AUC score, and precision. A low score implies an increased risk of inference, compromising individual record privacy. In contrast, a high score suggests that an attacker is unlikely to determine if a record was part of the original dataset. Table 5 shows the membership inference scores.

Membership Inference Score[1]

Proxy for Privacy score

Table 5

Name of Python Library	Algorithm	Accuracy	AUC Score	Precision
YData	CTGANs	0.9660	0.9972	0.9754
SDV	CTGANs	0.9451	0.9929	0.9574
SDV	GC	0.9148	0.9796	0.9260
SDV	TVAE	0.8181	0.9246	0.8472
YData	GMM	0.7911	0.9032	0.8199
	SMOTE	0.7664	0.7784	0.7134

[1] The privacy score is proxied by the concept of membership inference score, which relies on precision as the primary measure.

Sources: Author's computations

We found that GAN-based models consistently achieved higher scores in this metric, providing compelling evidence that attempts to reverse engineer the model for synthetic dataset generation and reveal the actual records are highly improbable with such models.

However, SMOTE is an exception to this trend. By design, SMOTE generated synthetic data more systematically and straightforwardly, making it relatively easier to replicate the underlying process. Consequently, SMOTE-generated synthetic data were more vulnerable to attacks for exposing actual data points.

It is worth noting that the significant gap between the AUC score and the precision score reveals the impact of false positives during the classification process. Such disparity underscored the importance of minimizing false positives to enhance the overall effectiveness of privacy-preserving measures in ML models.

Evaluation of Best Synthetic Dataset

We devised a straightforward scoring system to comprehensively assess and identify the most suitable synthetic dataset, as discussed in Table 2. We derived these scores from the algorithms' performance across the evaluation metrics discussed in earlier sections.

Synthetic Data Evaluation Matrix[1]

3 – Excellent, 2 – Fair, 1 - Poor Table 6

Metric	Algorithm	Fidelity	Utility	Privacy	Overall
YData	CTGANs	1	1	3	1.7
SDV	CTGANs	1	1	3	1.7
SDV	GC	2	1	3	2.0
SDV	TVAE	3	2	2	2.3
YData	GMM	3	2	2	2.3
In-house	SMOTE	2	3	1	2.0

[1] The evaluation matrix is based on the following resource: asq.org/quality-resources/decision-matrix

Sources: Authors' computations

Based on this scoring system, TVAE and GMM emerged as the top-performing synthetic data generation models, excelling in data fidelity, utility, and privacy, with an average score of 2.3. Meanwhile, GC and SMOTE ranked second with average scores of 2.0. CTGAN for YData and SDV ranked last, consistently scoring poorly across all evaluation metrics except privacy.

High-quality synthetic datasets must be based on an algorithm that performs exceptionally well in all three aspects. Of the six synthetic data generation algorithms tested, only GMM and TVAE algorithms have shown higher-quality synthetic CES data for research. On the other hand, CTGAN-based synthetic data ranked highest in data privacy. Still, CTGAN could not capture the inherent properties of the actual dataset, agreeing with the results of Little et al. (2021) and Pathare et al. (2023).

6. Conclusion and Recommendations

Synthetic data generation is valuable for addressing data-sharing concerns with external entities, especially researchers outside the BSP. In harnessing the potential of synthetic data for research and data-driven decision-making, it is imperative to remain prudent about preserving privacy, maintaining data quality, and adhering to ethical and regulatory standards. By maximizing the potential of synthetic datasets, data-sharing protocols can be streamlined through automated generation and evaluation processes while upholding data privacy and confidentiality.

We explored the generation of synthetic datasets and presented a synthetic tabular data generation pipeline using the CES dataset, primarily composed of categorical variables. Unlike previous studies using cleaned datasets from public repositories (e.g., UCI Machine Learning repository), we utilized a raw and unprocessed dataset, adding complexity in maintaining data integrity during cleaning and pre-processing before employing machine learning algorithms for synthetic data generation.

The algorithms we used in this study encompass traditional statistical methods (e.g., SMOTE, GMM, and GC) and cutting-edge deep learning techniques (e.g., TVAE and GANs) through open-source libraries dedicated to synthetic data generation,

specifically YData Synthetic and SDV. Both methodologies offered ease of use and flexibility, allowing control over data processing parameters to enhance model performance and synthetic data quality.

We found that GMM and TVAE are the most effective algorithms for generating synthetic data from the CES dataset, meeting evaluation criteria for data fidelity, utility, and privacy. Notably, GAN-based algorithms excelled in preserving data privacy but not the statistical properties of the actual dataset.

Future research endeavors may explore synthetic data generation for survey datasets containing various data types (i.e., numerical and a combination of categorical and numerical data), such as the BSP's Consumer Finance Survey. Furthermore, central banks often use time-series data for macroeconomic research. Future studies may explore creating synthetic time-series data to extend short datasets by adding extra data points. Lastly, future works may delve into developing a more systematic approach to the synthetic data generation process. It is worth emphasizing that this study serves as proof of concept and sets the stage for operationalizing the synthetic data generation pipeline in the BSP.

References

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., & Veloso, M. (2020). Generating synthetic data in Finance. Proceedings of the First ACM International Conference on AI in Finance. https://doi.org/10.1145/3383455.3422554

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022). Membership inference attacks from first principles. 2022 IEEE Symposium on Security and Privacy (SP). https://doi.org/10.1109/sp46214.2022.9833649

Cramér, H. (1946). Mathematical Methods of Statistics (PMS-9). In *Princeton University Press eBooks*. https://doi.org/10.1515/9781400883868

Irving Fisher Committee on Central Bank Statistics. (2015). (rep.). Data-sharing: issues and good practices. Retrieved 2023, from https://www.bis.org/ifc/events/7ifc-tf-report-datasharing.pdf.

Heyburn, R., Bond, R., Black, M., Mulvenna, M., Wallace, J., Rankin, D., & Cleland, B. (2018). Machine learning using synthetic and real data: Similarity of evaluation metrics for different healthcare datasets and for different algorithms. *Data Science and Knowledge Engineering for Sensing Decision Support*. https://doi.org/10.1142/9789813273238_0160

Koblents, E. & Megia, A.L. (2023). "Joint secondary anonymisation of categorical and numerical variables in sensitive time series microdata - novel approach for Statistical Disclosure Control of a sensitive microdata set published in BE," IFC Bulletins chapters, in: Bank for International Settlements (ed.), Post-pandemic landscape for central bank statistics, volume 58, Bank for International Settlements.

Little, C., Elliot, M., Allmendinger, R., & Samani, S. S. (2021). Generative Adversarial Networks for Synthetic Data Generation: A Comparative study. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2112.01925

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: a novel SMOTE-Based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, *4*(1), 18. https://doi.org/10.3390/asi4010018

Pan, J., Pham, V., Dorairaj, M., Chen, H., & Lee, J. (2020). Adversarial Validation approach to concept drift problem in user targeting automation systems at Uber. arXiv (Cornell University). https://arxiv.org/pdf/2004.03045.pdf

Pathare, A., Mangrulkar, R. S., Suvarna, K., Parekh, A., Thakur, G., & Gawade, A. (2023). Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *International Journal of Information Management Data Insights*, *3*(2), 100177. https://doi.org/10.1016/j.jjimei.2023.100177

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. *IEEE*. https://doi.org/10.1109/dsaa.2016.49

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). SciKit-Learn: Machine Learning in Python. *HAL (Le Centre Pour La Communication Scientifique Directe)*. https://hal.inria.fr/hal-00650905

Raghunathan, T. E. (2021). Synthetic data. *Annual Review of Statistics and Its Application*, *8*(1), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Platzer, M., & Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.679939

Sallier, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis1. *Statistical Journal of the IAOS*. https://doi.org/10.3233/sji-200682

United Nations Economic Commission for Europe. (2022). *Synthetic Data for Official Statistics: A Starter Guide*. https://unece.org/sites/default/files/2022-11/ECECESSTAT20226.pdf

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.00503

Zhang, Y., Zaidi, N. A., Zhou, J., & Li, G. (2023). Interpretable tabular data generation. *Knowledge and Information Systems*, *65*(7), 2935–2963. https://doi.org/10.1007/s10115-023-01834-5

Zhao, Z., Kunar, A., Van Der Scheer, H., Birke, R., & Chen, L. Y. (2021). CTAB-GAN: Effective table data Synthesizing. *arXiv (Cornell University)*. https://arxiv.org/pdf/2102.08369.pdf

Appendix A

List of Variables

Variable Name	Description	Values
INCOME	Income Class	1 – Low-income 2 – Middle-income 3 – High-income
REG	Region	1 – Region 1 – Ilocos 2 – Region 2 – Cagayan Valley 3 – Region 3 – Central Luzon 4 – Region 4 – CALABARZON 5 – Region 5 – Bicol 6 – Region 6 – Western Visayas 7 – Region 7 – Central Visayas 8 – Region 8 – Eastern Visayas 9 – Region 9 – Western Mindanao 10 – Region 10 – Northern Mindanao 11 – Region 11 – Southern Mindanao 12 – Region 12 – Central Mindanao 13 – Region 13 – National Capital Region 14 – Region 14 – Cordillera Administrative Region 15 – Region 15 – Autonomous Region in Muslim Mindanao 16 – Region 16 – Caraga 17 – MIMAROPA
AGE	Age	None
SEX	Sex	1 – Male 2 – Female
MSTAT	Marriage Status	1 – Single 2 – Married 3 – Common-law/Live-in 4 – Widowed 5 – Divorced 6 – Separated 7 – Annulled 8 - Unknown
HGRADE	Highest Educational Attainment	0 - No Grade Completed 1 - Early Childhood Education 2 - Primary Education 3 - Lower Secondary Education 4 - Upper Secondary Education 5 - Post-Secondary Non-Tertiary Education 6 - Short-Cycle Tertiary Education 7 - Bachelor Level Education or Equivalent 8 - Master Level Education or Equivalent 9 - Doctor Level Education or Equivalent

List of Variables (cont.)

Variable Name	Description	Values
PRIMOCCU	Primary Occupation	0 - Armed Forces Occupations 1 - Managers 2 - Professionals 3 - Technicians and Associate Professionals 4 - Clerical support workers 5 - Service & sales workers 6 - Skilled agricultural, forestry & fishery workers 7 - Craft & related trades workers 8 - Plant & Machine Operators & Assemblers 9 - Elementary Occupations
B1S	Present Financial Situation	1 - Better 2 - Same 3 - Worse
B2S	Financial Situation after 3 months	1 - Better 2 - Same 3 - Worse
B3S	Financial Situation after 12 months	1 - Better 2 - Same 3 - Worse
B4S	Present Economic Condition of the Country	1 - Better 2 - Same 3 - Worse
B5S	Economic Condition of the Country after 3 Months	1 - Better 2 - Same 3 - Worse
B6S	Economic Condition of the Country after 12 Months	1 - Better 2 - Same 3 - Worse
C1A	Number of Unemployed Persons for the Current Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C1B	Number of Unemployed Persons for the Next Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C1C	Number of Unemployed Persons for the Next 12 Months	1 - Increase 2 - Same as this period 3 - Decrease
C2A	Level of Interest Rates for Borrowing Money for the Current Quarter	1 - Increase 2 - Same as this period 3 - Decrease
C2B	Level of Interest Rates for Borrowing Money for the Next Quarter	1 - Increase 2 - Same as this period 3 - Decrease

List of Variables (cont.)

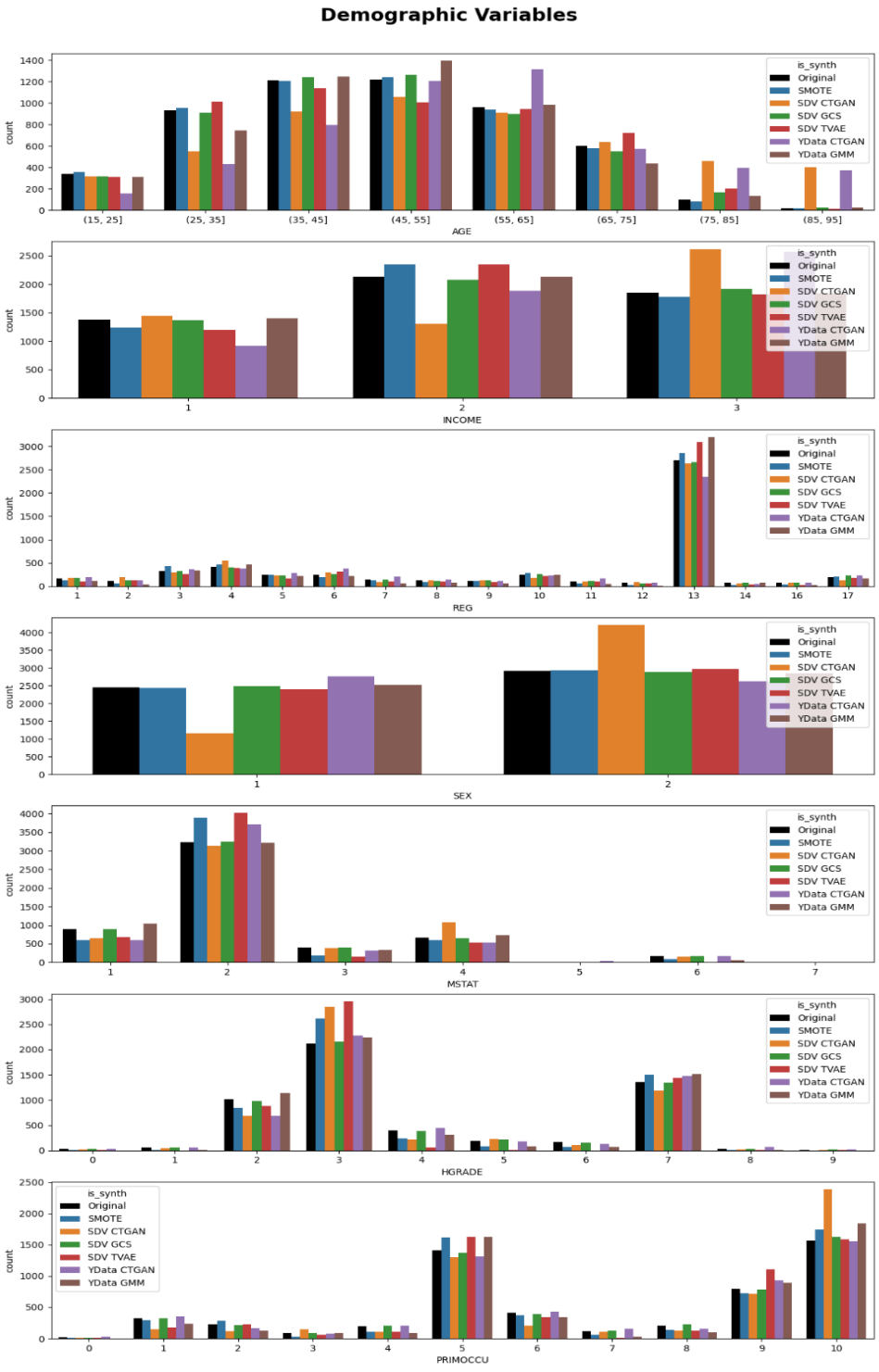
Variable Name	Description	Values
C2C	Level of Interest Rates for Borrowing Money for the Next 12 Months	1 - Increase 2 - Same as this period 3 - Decrease
C3A	Peso-Dollar Exchange Rate for the Current Quarter	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C3B	Peso-Dollar Exchange Rate for the Next Quarter	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C3C	Peso-Dollar Exchange Rate for the Next 12 Months	1 - Appreciate 2 - Same as this period 3 - Depreciate 4 - Don't know
C5A	Inflation rate will rise, remain, unchanged or fall in the current quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down
C5C	Inflation rate for the current quarter	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
C5D	Inflation rate will rise, remain, unchanged or fall in the next quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down
C5F	Inflation rate for the current year	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
C5G	Inflation rate will rise, remain, unchanged or fall in the next quarter	1 - Will go up 2 - Remain unchanged 3 - Will go down

List of Variables (cont.)

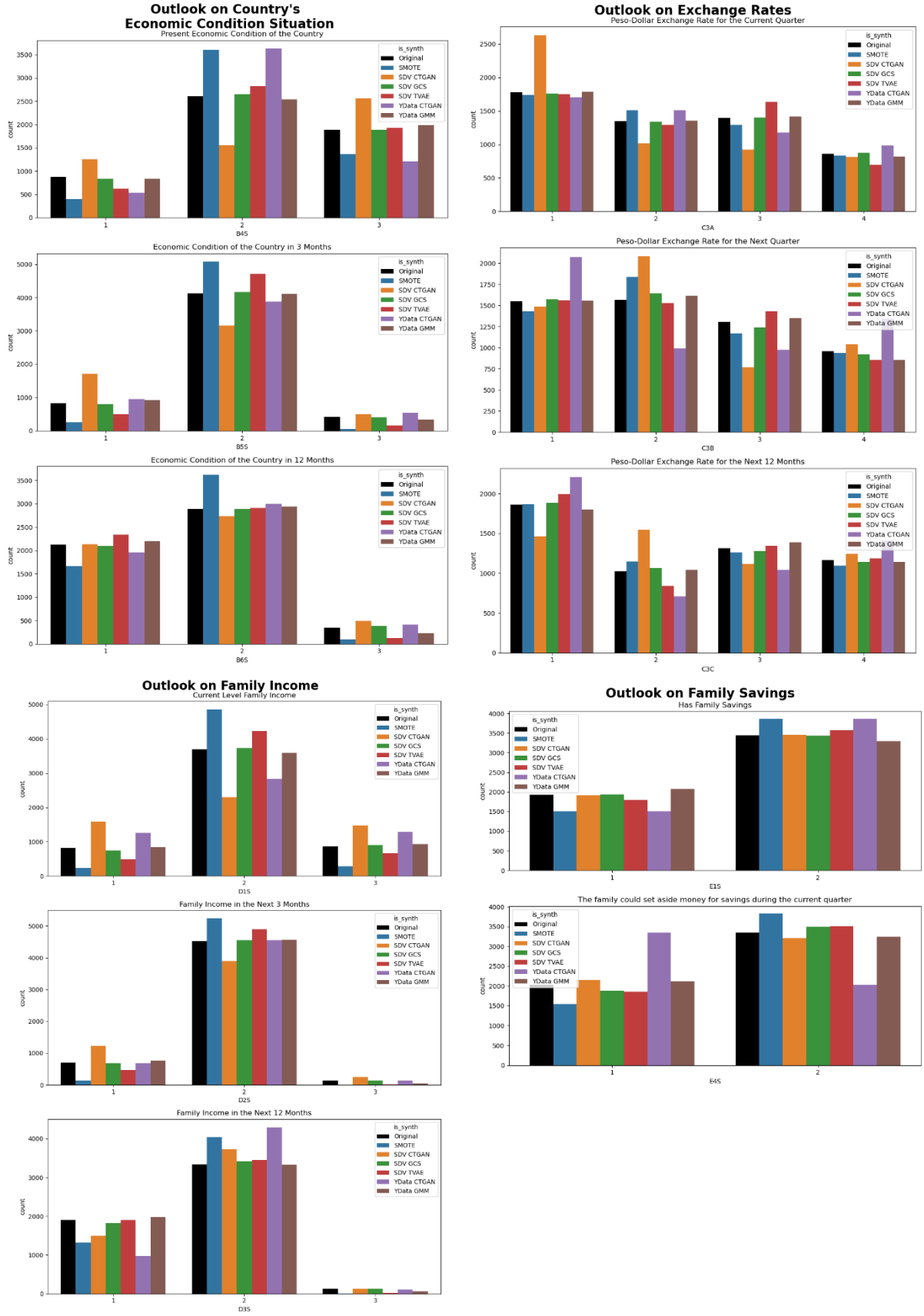
Variable Name	Description	Values
C5I	Inflation rate for the next 12 months	1 - Less than 0% 2 - Equal to 0% 3 - 0.1% to 1.9% 4 - 2% to 3.9% 5 - 4% to 5.9% 6 - 6% to 7.9% 7 - 8% to 9.9% 8 - 10% or more
D1S	Current Level Family Income	1 - Went up 2 - Same as now 3 - Went down
D2S	Family Income Next 3 Months	1 - Will go up 2 - Same as now 3 - Will go down
D3S	Family Income Next 12 Months	1 - Will go up 2 - Same as now 3 - Will go down
E1S	Has family savings	1 - Yes 2 - No
E4S	The family could set aside money for savings during the current quarter	1 - Yes 2 - No
F1SA	Outstanding loan	1 - Yes 2 - No
F2SA	Plan to apply loan in the next quarter	1 - Yes 2 - No
F3SA	Plan to apply loan in the next 12 months	1 - Yes 2 - No

Appendix B

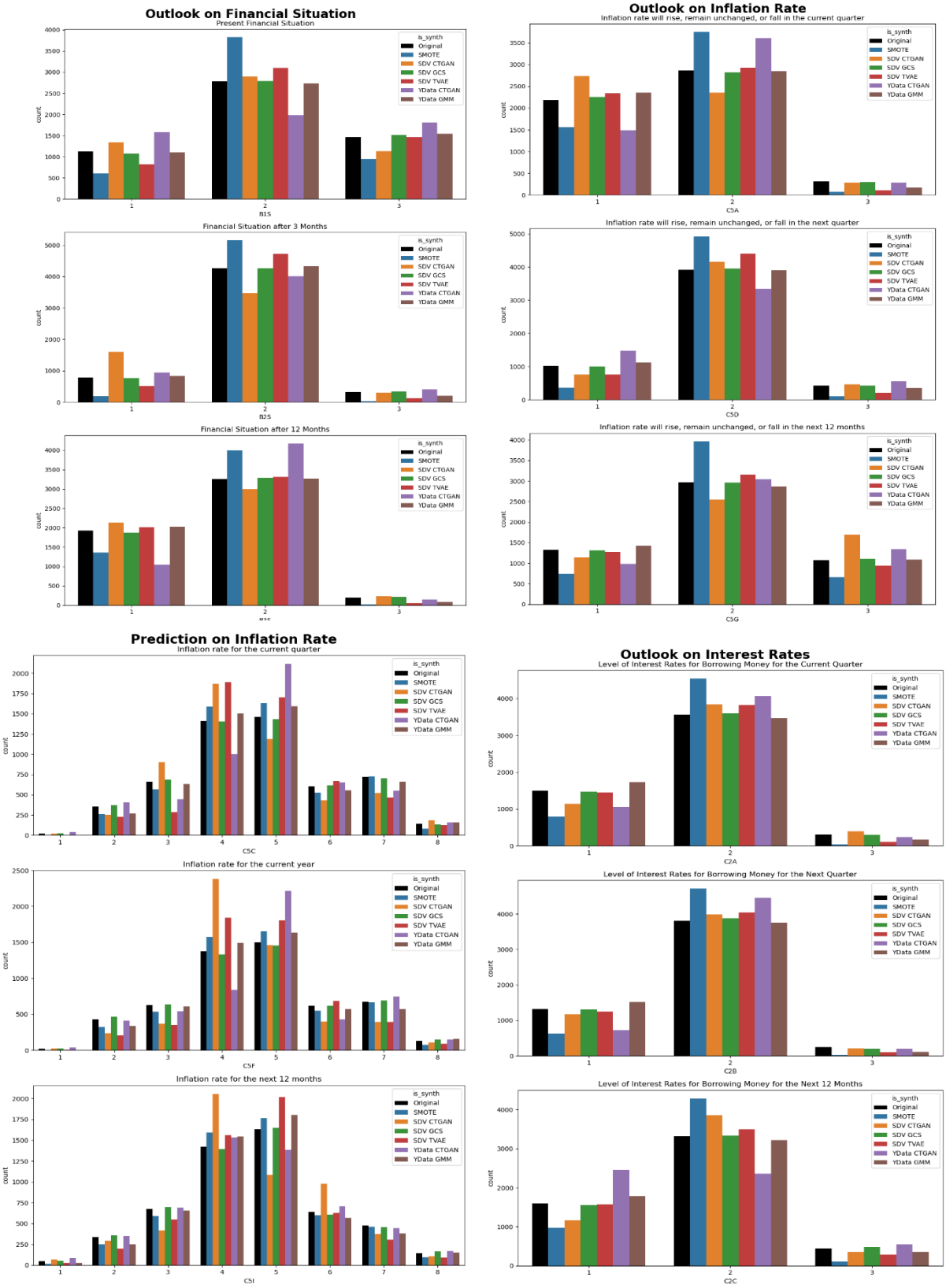
Comparison of histogram plot for each variable for actual vs. synthetic data



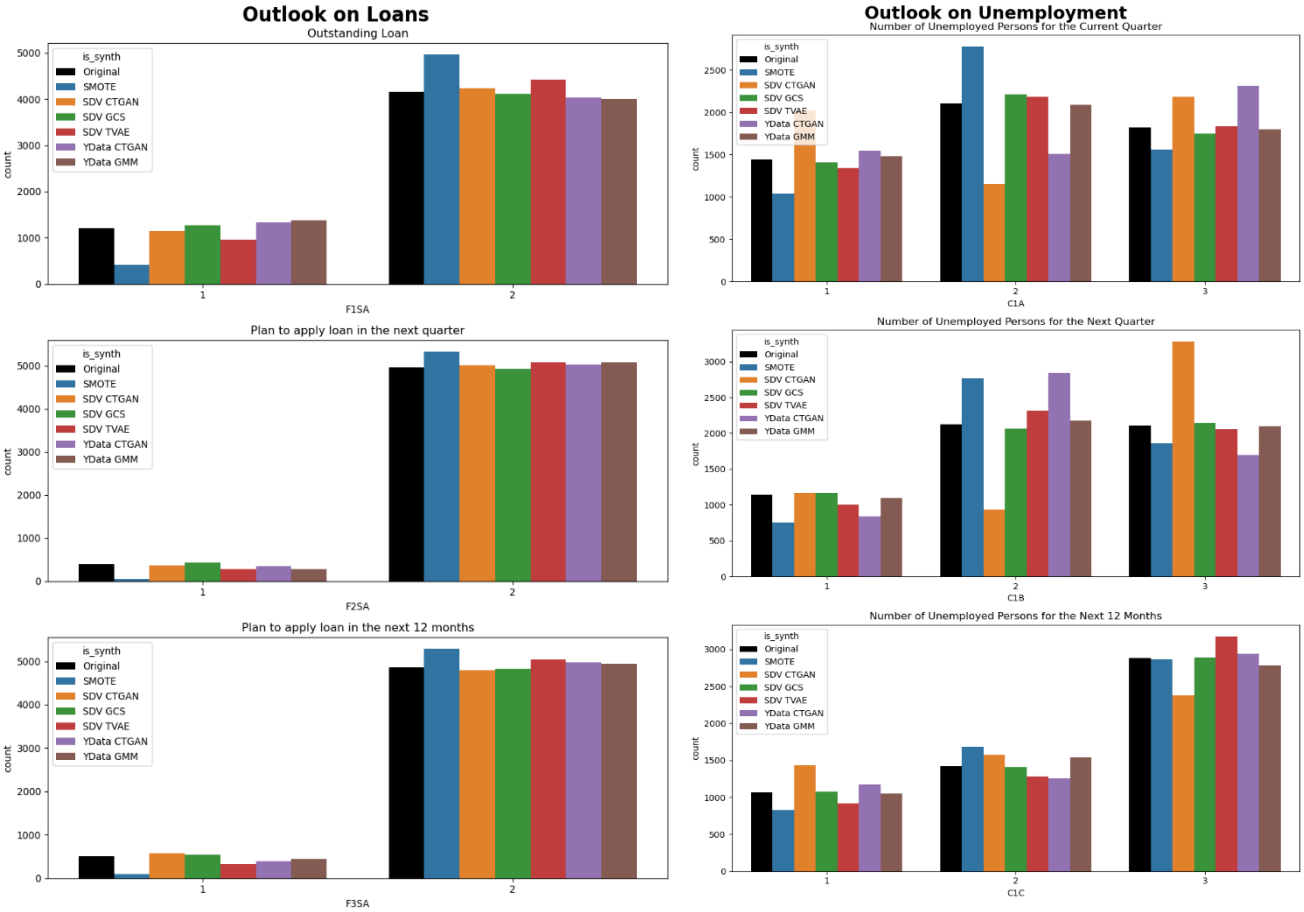
Comparison of histogram plot for actual vs. synthetic data (cont.)



Comparison of histogram plot for actual vs. synthetic data (cont.)



Comparison of histogram plot for actual vs. synthetic data (cont.)



Research for All: Exploring machine learning applications in generating synthetic datasets

CARMELITA ESCLANDA-LO
GABRIEL MASANGKAY
CHELSEA ANNE ONG
ROSSVERN REYES

3RD IFC AND BANK OF ITALY WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING: ENHANCING THE ACCESS TO AND SHARING OF DATA", OCTOBER 17-19, 2023



* The views expressed herein are those of the authors' only and do not necessarily reflect those of the Bangko Sentral ng Pilipinas

Classification: GENERAL

OUTLINE

01 MOTIVATION

02 RELATED STUDIES

03 DATA

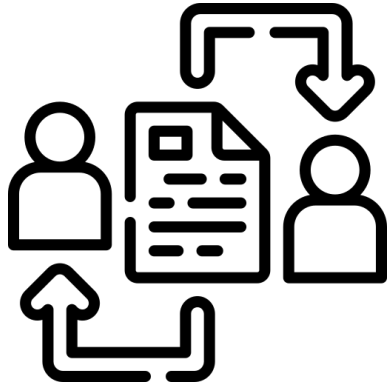
04 METHODOLOGY

05 RESULTS

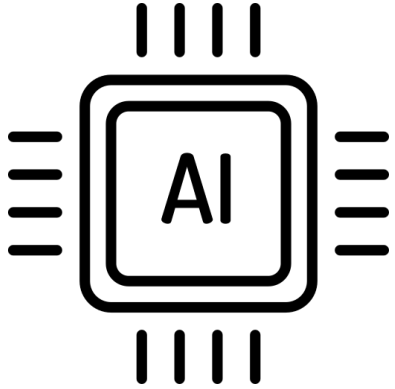
06 KEY TAKEWAYS AND RECOMMENDATIONS

Classification: GENERAL

Motivation and Objectives



Ease data sharing procedures



Explore AI for synthetic data generation



Generate quality and private data for research use

Classification: GENERAL

Related Studies

Data Sharing Practices

Data sharing frameworks are in place and delegated entities enforce these frameworks.

The BSP has developed the **Data Governance Manual** which specifies sharing data to external parties and protecting sensitive information.

Meanwhile, **National Statistics Offices around the world have developed and operationalized synthetic datasets** for public data dissemination, improve efficiency of data sharing processes.

Use of Synthetic Data for Research

For methodologies on generation of synthetic tabular data, most studies explore **Generative Adversarial Networks (GANs) and Tree-based models.**

“As the utility of synthetic data increases, the disclosure risk increases exponentially.”

Classification: GENERAL



Data

CONSUMER EXPECTATIONS SURVEY

Quarterly survey conducted by BSP to gather information from Filipino households regarding **sentiments on various economic indicators.**

SAMPLE VARIABLES

	Description	*Values*
Identifier Variables		
AGE	Age	0-100
INCOME	Income Group	Low, Middle, High
SEX	Sex	Male, Female
.....	.....	.....
Response Variables		
C5C	Inflation Rate in the Current Quarter	Less than 0%, 0.1%-1.9%, ...
E1S	Has Family Savings	Yes, No
B1S	Present Financial Situation	Better, Same, Worse
.....	.....	.....

Methodology



Data Collection and Preprocessing

Processing is done to preselect columns, **address missing data, differing data types**. Options for **variable selection, data binning, partial synthesis** are covered in this study.

Generate Synthetic Data

The following algorithms will be tested using in-house and **open-source packages** (e.g., Synthetic Data Vault (SDV), YData Synthetic):

- **SMOTE**
- **Gaussian Mixture Models (GMM)**
- **Gaussian Copula (GC)**
- **Tabular Variational Autoencoders (TVAE)**
- **Conditional Tabular Generative Adversarial Networks (CTGAN)**

Evaluate Synthetic Data

Assess whether synthetic datasets can be used as an alternative dataset. These shall be evaluated based on three key dimensions: **fidelity, utility, privacy.**

Classification: GENERAL

Synthetic Data Evaluation

Assess whether **synthetic dataset can be used as an alternative dataset for research use.**



Data Fidelity

- Statistical Similarity
  - Histogram
  - Correlation



Data Utility

- Machine Learning Performance
  - Accuracy
  - AUC
  - Recall
  - Precision
  - F1
  - Kappa
  - MCC



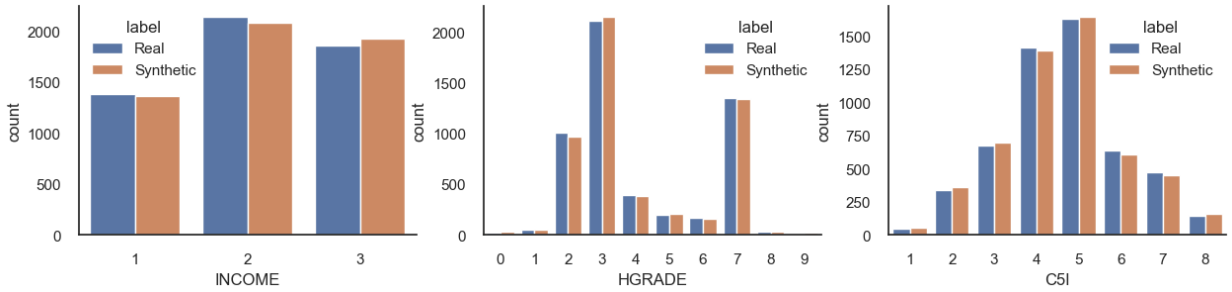
Data Privacy

- Membership Inference
  - Accuracy
  - AUC
  - Precision

Data Fidelity

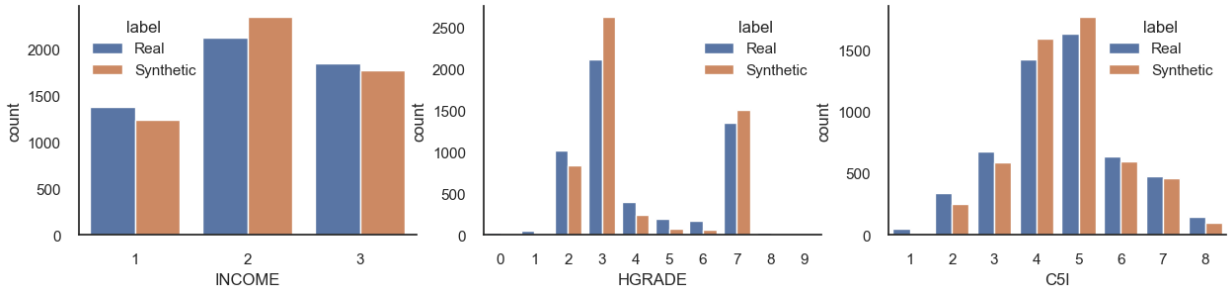
HISTOGRAM COUNTPLOTS

A total of 36 variables is analyzed to **compare the count distribution** for real and synthetic datasets.



Gaussian Copula

Similarity score = 0.9893 (± 0.0062)

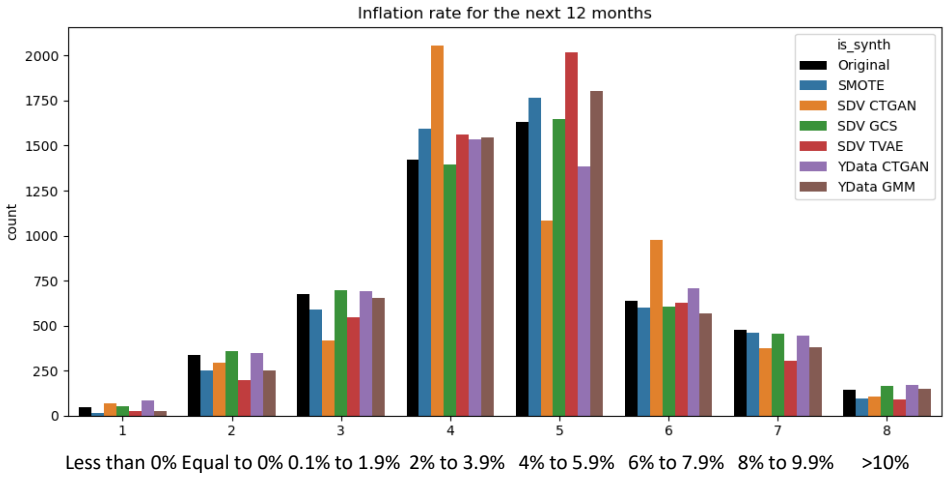


SMOTE

Similarity score = 0.8882 (± 0.0599)

**Values in parentheses are standard deviations*

TARGET VARIABLE (C5I)

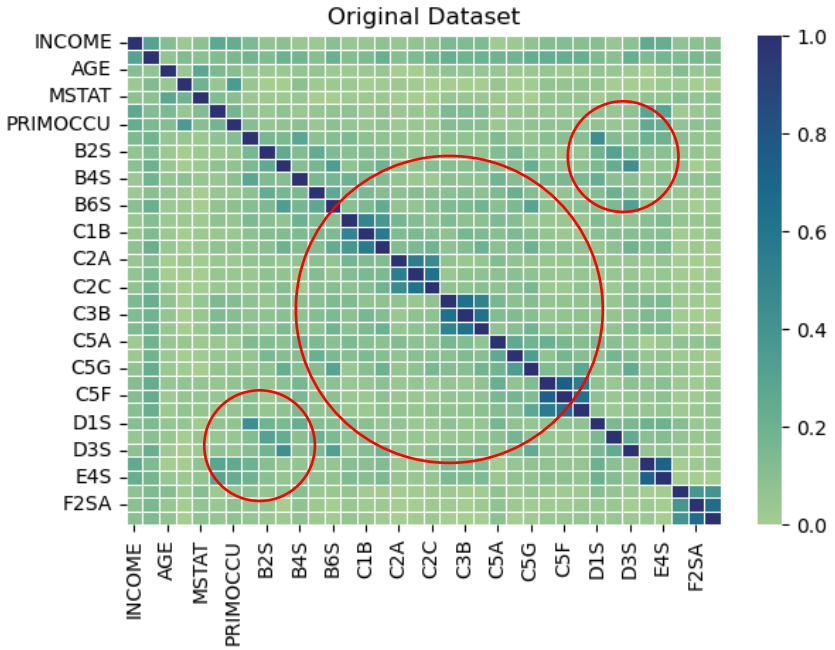


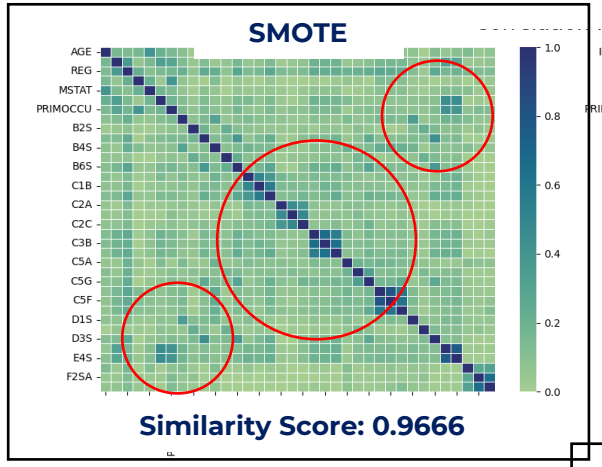
Classification: GENERAL

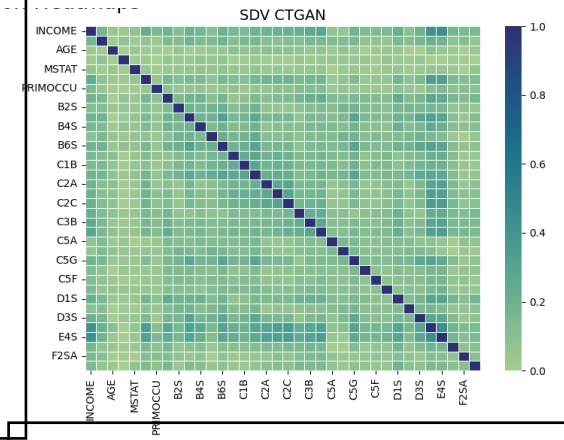
Data Fidelity

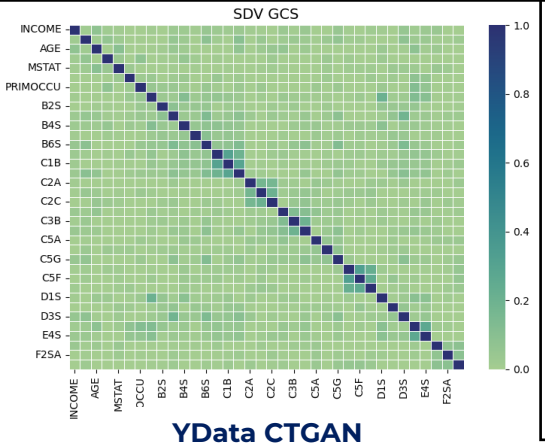
CORRELATION HEATMAP – CRAMER'S V

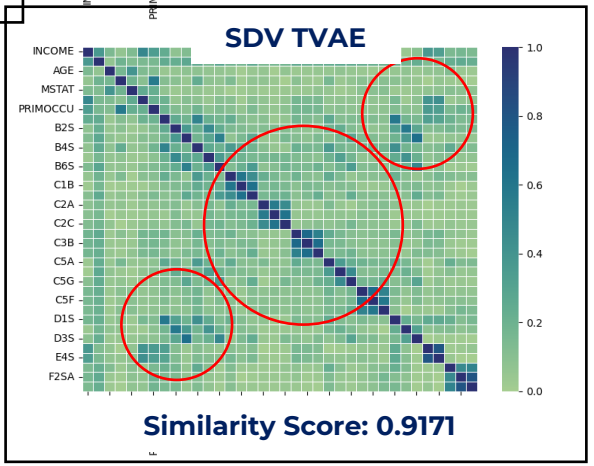
Cramér's V is used to determine whether a **significant relationship exists between two categorical variables.**

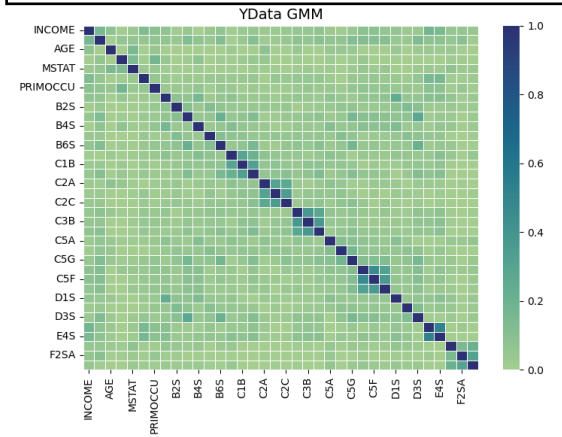












Classification: GENERAL

Data Fidelity

STATISTICAL SIMILARITY

The Statistical Similarity score is computed as the **average of the histogram and correlation similarity scores.**

Python Library	Algorithm	Bray-Curtis Similarity Scores (Histogram)	Cosine Similarity Scores (Correlation)	Statistical Similarity Score
YData	GMM	0.9696 (± 0.0247)	0.9306	0.9501
SDV	TVAE	0.9381 (± 0.0427)	0.9171	0.9276
In-house	SMOTE	0.8882 (± 0.0599)	**0.9666**	0.9274
SDV	GC	**0.9893 (± 0.0062)**	0.8377	0.9135
SDV	CTGANs	0.8884 (± 0.0699)	0.7226	0.8055
YData	CTGANs	0.8974 (± 0.0621)	0.4023	0.6499

**Values in parentheses are standard deviations*

Classification: GENERAL

Data Utility

MACHINE LEARNING PERFORMANCE

A multi-class classifier is built to **predict the range of inflation rate in the next 12 months.** Results presented are in terms of percentage difference against the real dataset.

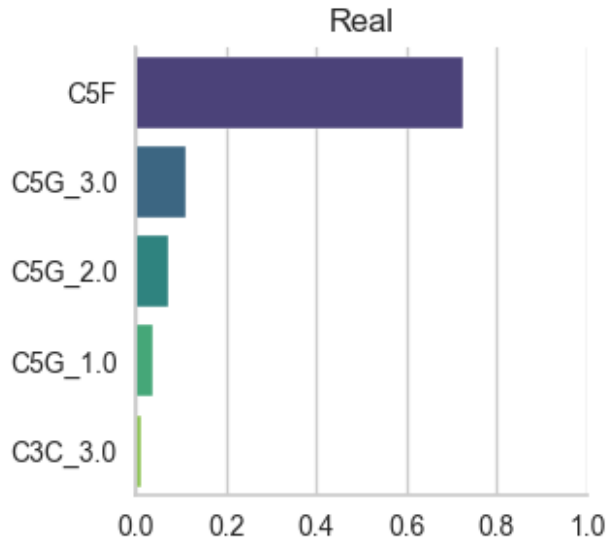
Data		Acc.	AUC	Recall	Prec.	F1	Kappa	MCC	Average Difference
Real		***0.7852***	***0.9493***	***0.7852***	***0.7913***	***0.786***	***0.7303***	***0.7313***	
SMOTE		-0.04	-0.02	-0.04	-0.04	-0.04	-0.05	-0.05	-0.04
SDV	TVAE	-0.06	-0.03	-0.06	-0.07	-0.06	-0.08	-0.08	-0.06
YData	GMM	-0.07	-0.05	-0.07	-0.08	-0.07	-0.09	-0.09	-0.07
SDV	GC	-0.36	-0.21	-0.36	-0.41	-0.40	-0.48	-0.47	-0.38
SDV	CTGAN	-0.35	-0.24	-0.35	-0.37	-0.40	-0.46	-0.45	-0.37
YData	CTGAN	-0.52	-0.41	-0.52	-0.56	-0.56	-0.72	-0.72	-0.57

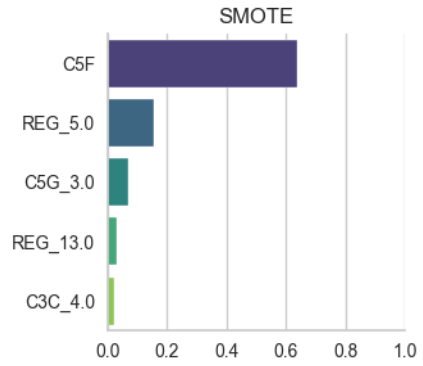
Classification: GENERAL

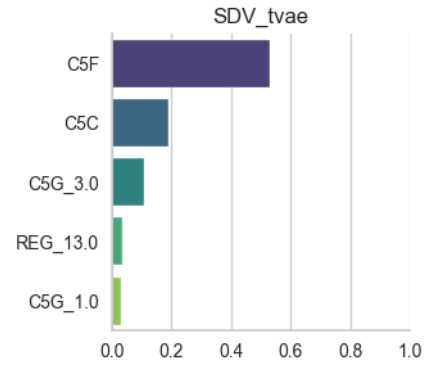
Data Utility

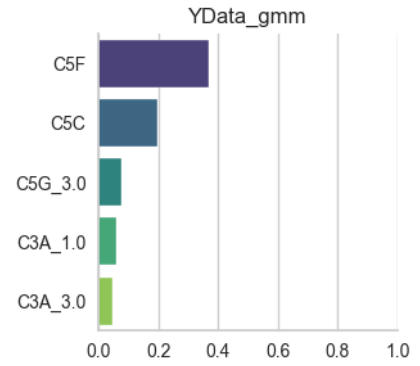
MACHINE LEARNING PERFORMANCE - FEATURE IMPORTANCE

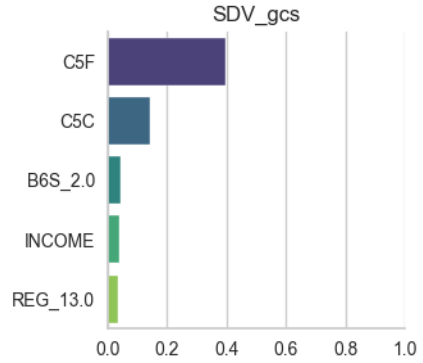
The **top predictors** of inflation rate in the next 12 months are shown using feature importance scores.

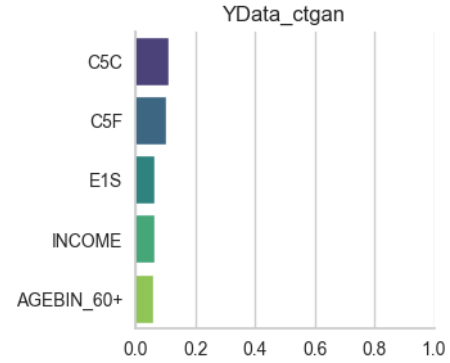


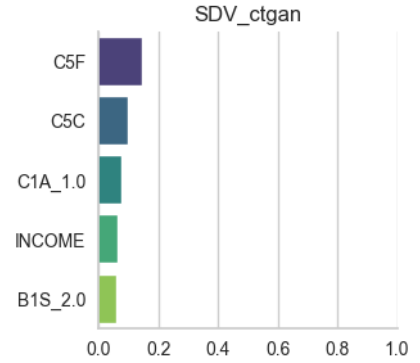












Classification: GENERAL

Data Privacy

MEMBERSHIP INFERENCE

A binary classifier is built to **distinguish real (0) from synthetic (1) data** and **evaluate using precision metric or privacy score.** A low score implies an increased risk of inference, compromising individual record privacy, while a high score suggests that an attacker is unlikely to determine if a record was part of the real dataset.

Python Library	Algorithm	Accuracy	AUC Score	Precision (Privacy Score)
YData	CTGAN	0.9660	0.9972	0.9754
SDV	CTGAN	0.9451	0.9929	0.9574
SDV	GC	0.9148	0.9796	0.9260
SDV	TVAE	0.8181	0.9246	0.8472
YData	GMM	0.7911	0.9032	0.8199
In-house	SMOTE	0.7664	0.7784	0.7134

Classification: GENERAL

Best Synthetic Dataset

To determine the best synthetic dataset, each algorithm shall be evaluated according to this metric.

	Data Fidelity	Data Utility	Data Privacy	Overall
3 - Excellent	Has a Statistical Similarity score of 0.95 and up	Overall average difference of 0.05 or less No difference higher than 0.05 in any metric	Privacy score is higher than 0.9	Synthetic data can be used for research
2 - Fair	Statistical Similarity score is higher than 0.90 but lower than 0.95	Overall difference of 0.05 to 0.10 No difference higher than 0.10 in any metric	Privacy score is higher than 0.8 but lower than 0.9	Can be used for research but with conditions
1 - Poor	Has a Statistical Similarity score of 0.90 and lower	Difference of more than 0.10 Other conditions not satisfying any of the above	Privacy score is lower than 0.8	Not valid for research use Re-evaluate algorithm

Classification: GENERAL

Best Synthetic Dataset

The best synthetic dataset should have a score of 3 on all metrics. A data being produced by an algorithm having a score of 1 in any metric should not be used for research and should be re-evaluated.

3	Excellent	**2**	Fair	**1**	Poor		

Metric		In-house	YData		Synthetic Data Vault		
		SMOTE	**CTGANs**	**Gaussian Mixture Model (GMM)**	**CTGANs**	**Gaussian Copula**	**TVAE**
Data Fidelity	Statistical Similarity	2	1	3	1	2	3
Data Utility	Machine Learning	3	1	2	1	1	2
Data Privacy	Membership Inference	1	3	2	3	3	2
AVERAGE SCORE		**2.0**	**1.7**	**2.3**	**1.7**	**2.0**	**2.3**

Classification: GENERAL

Key Findings and Future Works

Key Takeaways:

- Synthetic data could **replicate real data.** This can serve as an alternative and be shared with external parties. A rubric is created to decide if a synthetic data can be used for research purposes.
- For the CES dataset, synthetic datasets generated using the **TVAE and GMM** algorithm produced the best results. On the other hand, GAN-based models performed poorly in all synthetic evaluation metrics except data privacy.
- By **utilizing open-source libraries**, the implementation of generating synthetic data is much easier.

Future Works:

- Expand this study by adding numerical and time-series datasets
- Explore more algorithms for synthetic data generation
- Operationalize the synthetic data generation pipeline for research use

Classification: GENERAL

Thank you!

Chelsea Anne S. Ong
ongcs@bsp.gov.ph
Department of Economic Statistics
Bangko Sentral ng Pilipinas

Classification: GENERAL

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Invoices rather than surveys: using ML to build nominal and real indices[1]

Pablo Acevedo, Dagoberto Quevedo and Marco Rojas,
Central Bank of Chile,

Emiliano Luttini,
World Bank,

Matías Pizarro,
Pontificia Universidad Católica - Chile

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Invoices rather than surveys: using ML to build nominal and real indices

Pablo Acevedo
Central Bank of Chile

Emiliano Luttini
World Bank

Matías Pizarro
PUC-Chile

Dagoberto Quevedo
Central Bank of Chile

Marco Rojas
Central Bank of Chile

3rd IFC and Bank of Italy Workshop on Data Science in Central Banking
October 19, 2023. Rome, Italy

Disclaimer

The views are those of the authors and do not necessarily reflect the views of the Central Bank of Chile or its board members

Electronic Invoice

- The electronic invoice document is a mandatory digital document that reports the transactions between firms.
- The Central Bank of Chile obtains every day all documents issued in the last 24 hours ($\approx$ 3 millions).[1]
- Each document reports a free/unstructured text that describes each good or service in the transaction.
- However, it is *not standardized* to a product or service classification.

[1] Currently we have approximately 8,000 millions of documents with 19,000 millions free/unstructured text.

Example of an electronic invoice document



AGROSUPER COMERCIAL

AGROSUPER COMERCIALIZADORA DE ALIMENTOS LIMITADA

GIRO: Mayorista de Aves, Cerdos, Vacunos, Cocinas, Balnearios Congelados, Verduras, Hortalizas y Frutas Congeladas, Comercialización de Vinos y Licores.
CASA MATRIZ: La Estrella 401 of. 7 Sector Punta de Cortés - Rancagua - Fono (72) 233440
DIR. CABLEG.: FAX: 5306165 - SANTIAGO CHILE - CASILLA 277 - STGO. / 23 MAIPU
SERVIFONO: 600 600 6001

R.U.T.: 79.984.240 - 8

FACTURA ELECTRONICA

Nº 74

S.I.I. - RANCAGUA

SEÑOR(ES):	Fernanda Chacana Serrano	R.U.T.:	14.205.748-2
DIRECCION:	Punta Cortes s/n	150966 595985 0 1	San Pablo
GIRO:	Almacén		
COMUNA:	Pudahuel	CIUDAD:	Prov Santiago
VENDEDOR:	16030-Hector Iver Moraga Carreño	DIR. ORIGEN:	
FECHA:	05/05/2003		San Pablo N°9500,Pudahuel-Prov Santiago
CONDICION DE VENTA:	Efectivo	FECHA DE VENCIMIENTO:	05/05/2003
LOCAL:	Punta Cortes s/n		Pudahuel Prov Santiago

PLU	CODIGO	PRODUCTO	ENV.	EST.	UNID.	KILOS	PRECIO	VALOR
	04-01-02-062	(Super)Mortadela Jamonada-Fresco	09-09	1	1	3,00		
	04-01-02-064	(Super)Mortadela Salch. Cerveza C-75-Fresco	09-09	1	2	3,30		
	04-01-06-186	(Super)Arrollado lomo con ají-Fresco	09-09	1	1	2,60		
	04-04-02-068	(Super Pollo)Mortadela jamonada de pollo-Fresco	09-09	1	1	3,03		

Figure 1: Example of electronic invoice document. Source: Servicios de Impuestos Internos (SII).

Motivation

- Challenge:

  → We do not know **what** products are actually being sold.

- What we do:

  1. Use cutting-edge machine learning techniques. to classify transactions into product categories.
  2. Measure the economy in real time (e.g., consumption, investment).
  3. Study the differential effect the pandemic (and lockdowns) had on different consumption goods.

Roadmap

1. Classification approach
   - Strategy
   - Classification model
   - Quality metrics

2. Applications

Classification approach

Strategy

To standardize the free-text description, we do the following:

1. Consider two classifiers (COICOP, CUP):
   - Classification of Individual Consumption According to Purpose (COICOP): 303 classes.
   - Unique Code of Chilean Products (CUP): 290 classes.
2. Build a training sample using a human-entity to label items to this classifiers. Detail
3. Train a supervised multi-class text classification model to extrapolate learning to the rest of documents.

Training sample composition[2]

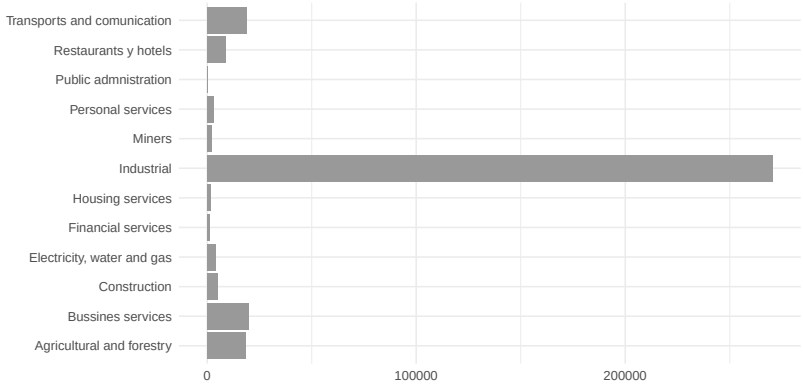


Figure 2: Training sample composition by Division on Unique Code of Chilean Products.

[2]Until December 2022

Expected output

Free/unstructured text	Class code	Class name
Fairlife 2% Reduced Fat Ultra-Filtered Milk, Lactose Free, 52 fl oz	011412	Liquid milk
Fairlife 2% Chocolate Ultra-Filtered Milk, Lactose Free, 52 fl oz	011412	Liquid milk
Great Value 2% Reduced Fat Milk, 128 Fl Oz	011412	Liquid milk
Great Value Fat Free Milk, Gallon, 128 fl oz	011412	Liquid milk
Crystal Whole Vitamin D Milk, Gallon, 128 fl oz	011412	Liquid milk
Horizon Organic Whole Shelf-Stable Milk Boxes, 8 Oz., 12 Count	011412	Liquid milk
Minute Instant Brown Rice, Light and Fluffy, 28 oz	011110	Rice
Minute Instant White Rice, Light and Fluffy, 14 oz	011110	Rice
Ben's Original Ready Rice Fried Flavored Rice, Easy Dinner Side, 8.5 OZ Pouch	011110	Rice
Success Rice Boil-in-Bag White Rice, Family Size, 32 Oz	011110	Rice
Clorox Splash-Less Liquid Bleach, Regular (Concentrated Formula) 77 Ounce	056114	Cleaner
Clorox Clean-Up All Purpose Cleaner with Bleach, Spray Bottle, 32 oz	056114	Cleaner

Word embedding

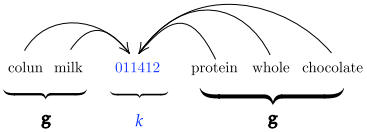
- Vector representation of words
  - Encoding: words that are close in the vector space are also similar in context to labelcode for class k.
- Many pre-compilations are available for Spanish and could be used in other contexts
  - However, it does noy apply in our case:
    $\Rightarrow$ We build a word embedding model is using words from training sample.
- In order to reinforce the fit of model, other features of the transaction are added as complements of the unstructured text:

$$\mathbf{g} = \{g \cup \langle b \rangle \cup \langle s \rangle\}, \tag{1}$$

g = standardized text after to applied a text cleaning strategy, Detail
b = brand,
s = sector that issued the text.

Definition of classification model

- We use FastText (Joulin *et al.*, 2016) as the classification model (similar to the CBoW (Mikolov *et al.*, 2013)), where the middle word is replace by a labelcode for class k:



- The softmax function f gets the probability of observing label for class k given a paragraph $\mathbf{g}$:

$$f(k, \mathbf{g}) \equiv \mathbb{P}(k \mid \mathbf{g}) = \frac{e^{h_{\mathbf{g}}^{\top} \mathbf{v}_k}}{\sum_{l=1}^{K} e^{h_{\mathbf{g}}^{\top} \mathbf{v}_l}},$$

where, $\mathbf{u}_w \in \mathbb{R}^D$ = vector representation of word w in the D documents in training sample,
$\mathbf{v}_k \in \mathbb{R}^D$ = vector representation of class k,
$h_{\mathbf{g}} \in \mathbb{R}^D = \sum_{w \in \mathbf{g}} \mathbf{u}_w$, paragraph features of $\mathbf{g}$.

Fine tuning and execution conditions

- The fastText library was downloaded from GitHub[3] repository and compiled.
- For each classifier we run an exhaustive hyperparameter grid search using the *automatic hyperparameter optimization*[4] option (`autotune`), running for 12 hours, getting the set of hyperparameters for build the final model.
- All execution were run on a machine with the Windows Server 2022 operating system, Intel Xeon Gold 6126 2x2.60 Ghz processor and 1.5 TB RAM, under Python 3.11. `scikit-learn` package was used for run the quality metrics.

[3] github.com/facebookresearch/fastText
[4] fasttext.cc/docs/en/autotune.html

Quality metrics

- A repeated stratified q-fold cross-validation is done, using $q = 5$ and stratifying by class k, then repeated 10 times.
- For each repeated cluster $q \times k$, we compute the metrics of Precision, Recall and F1.
  Detail
- Results:

Table 1: Quality metrics by classifier.

Classifier	Precision	Recall	F1
COICOP	0.955	0.954	0.954
UCP	0.936	0.935	0.935

Notes: A weighted average by number of text in class k is shown for each quality metrics by classifier.

Applications

1. Durable and non-durable consumption

- Our data is firm-to-firm: not consumption of households
  - We solve this by using *purchases* of the retail sector.
- We map the product categories into durable and non-durable consumption
  - Then, we sum across all transactions for durable/non-durable to obtain nominal aggregates.
  - Obtain real aggregates using official deflators.

$\Rightarrow$ Build real-time consumption of durable and non-durable

  - Comparable to National Account series.

- We can do this for finer groups of products.

Durables and non-durables series

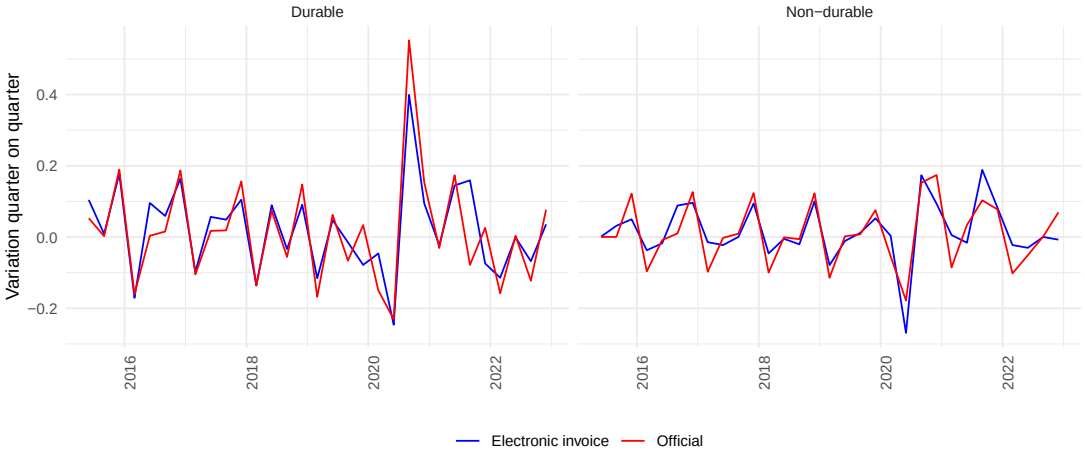


Figure 3: Volumen variation: quarter on quarter delta log changes. Authors own calculation based on Central Bank of Chile data.

2. Consumption pattern before and during lockdown

What we do

- We assess the consumption pattern before and during lockdowns during the COVID-19 pandemic.

How we do it

- We build series from the retail sector purchases.
- Goods are identified using the COICOP classification.
- An empirical approach is proposed to consider purchases variations with respect to the first week of the lockdown. Detail

Heterogenous effects of lockdown on consumption

By macrodivision product

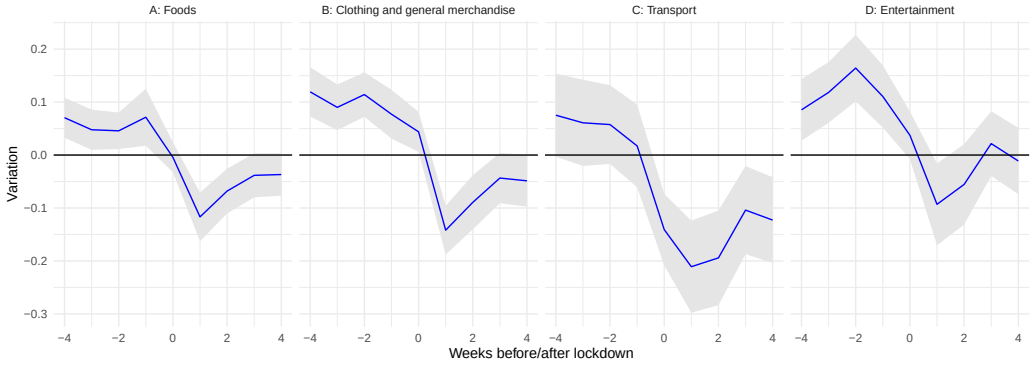


Figure 4: Purchases variations per week before or during lockdown for retail sector by division product. This figure report $\sum_{s \in S} \alpha_s$ of estimate regression define in Equation 4 of the empirical approach. The gray area denote a confidential interval 5-95%.

Conclusion

1. We propose a methodology to go from raw description of goods sold to standardized product categories.
2. We use frontier techniques in machine learning together with other observables from tax documents to carry out the classification.
3. We apply this to build real-time series of consumption (and investment), and to study the effect of lockdowns on different types of consumption.

Appendix

Text cleaning strategy

Return

The text component or features are standardize, removing elements into the text whose inclusion not contribute to the classification process. The following operations are applied:

1. Transform text to lowercase.
2. Standardization to Unicode format NFKD for transform to ASCII encoding, e.g. replace accents.
3. Remove any character distinct to $[a - z]$.
4. Remove stopwords.
5. Apply stemming.
6. Remove words with a length less than two.
7. Replace white space with a length greater than one, and any white spaces in start o end of the text.

Training sample

Return

- A selection criteria statistic is defined for each text $i \in G_s$ in sector s:

$$\gamma_{si} = \frac{1}{3}\left(\frac{m_i}{\sum_{j \in G_s} m_j}\right) + \frac{1}{3}\left(\frac{e_i}{\sum_{j \in G_s} e_j}\right) + \frac{1}{3}\rho_i^s, \tag{2}$$

where,

m_i = total revenue,
e_i = number of events or reports,
ρ_i^s = monthly normal probability occurrence over time,
G_s = set of unique free texts in sector s.

- n free text in sector s with the highest value of γ_{si} are selected.
- A person interprets the text and labels items to a class that represents the good or service.

Quality metrics

Return

- For each run cluster q and class k, we calculate the number of true positives (T_p), of false positives (F_p), and of false negatives (F_n).
- Precision (P): measure of result relevancy,

$$P^{qk} = \frac{T_p^{qk}}{T_p^{qk} + F_p^{qk}},$$

- Recall (R): measure of how many truly relevant results are returned,

$$R^{qk} = \frac{T_p^{qk}}{T_p^{qk} + F_n^{qk}},$$

- F1-score: harmonic mean of precision and recall,

$$F1^{qk} = 2\frac{P^{qk} \times R^{qk}}{P^{qk} + R^{qk}}.$$

Empirical Approach

Consumption pattern before and during lockdown

Return

Weekly purchases change (in relation to the ones done in the first week of lockdown) is determined as:

$$\Delta \log X_{t,c} = \log X_{t,c} - \log X_{\rho(t),c}, \tag{3}$$

where $X_{t,c}$ is the total purchases in the municipality c and period t and $\rho(t)$ is the week of the closest lockdown start. This is incorporated on the following model:

$$\Delta \log X_{t,c} = \lambda_c + \sum_{s \in S} \alpha_s \mathbb{1}\{s(t) = s\} + \sum_{w \in W} \beta_w \mathbb{1}\{w(t) = w\} + \epsilon_{t,c}, \tag{4}$$

where,

S = set of weeks before or during lockdown, $\{-4, \ldots, 0, \ldots, 4\}$,
W = set of weeks number in month, $\{1, \ldots, 5\}$,
$s(t)$ = weeks before or during lockdown in t, $s(t) \in S$,
$w(t)$ = week number in month of t, $w(t) \in W$.

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Leveraging large language models to extract data citations[1]

Sebastian Seltmann, Emily Kormanyos, Hendrik Christian Doll,
Deutsche Bundesbank

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Leveraging Large Language Models to Extract Data Citations

Sebastian Seltmann[1][2], Emily Kormanyos[1], and Hendrik Christian Doll[1][3][4]

Abstract

Empirical researchers and data providing institutions currently face challenges in efficiently understanding data usage in scholarly papers. Tracing data usage in papers relies on human readers and remains time-consuming and error-prone. To address this, we explore the potential of using Large Language Models (LLMs), specifically GPT-3.5, to automate the identification and categorization of research data sources. By employing web scraping, we collect a comprehensive sample of research papers and create a human-labeled validation dataset. We analyze the accuracy of GPT-3.5 in detecting and summarizing data sources in economics and finance papers. We evaluate the detection and prediction accuracy and address the issue of false answers provided by the model. We find that LLMs can advance considerably on the status quo. Results are encouraging in terms of quality of extraction and coverage. Our paper provides a detailed description of the pipeline to implement our solution at data providing institutions, enabling the understanding of data impact and therefore efficient data provision services.

Keywords: Data citations, Automated data source classification, Large Language Models (LLMs), Natural Language Processing (NLP), Research data centers (RDCs).

JEL classification: C81, C82, C88, C87.

[1] Deutsche Bundesbank, Data Service Centre

[2] Contact: sebastian.seltmann@bundesbank.de (corresponding author)

[3] All views expressed in this report are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

[4] We wish to express our gratitude for Shir Frank and Kilian Graef for their excellent student research assistantship. Without their support, this project would not have been possible.

Contents

1. Introduction 3

2. Contribution to the status quo 4

   2.1 Achievements to date 4

   2.2 Our value proposition 6

3. Data 7

   3.1 Sample collection 7

   3.2 Manual labeling of evaluation sample 8

4. Methodology 9

   4.1 Pipeline overview 9

   4.2 Performance measures 12

5. Results 13

6. Conclusions 15

References 17

Appendix 19

1. Introduction

Empirical researchers use data to conduct analyses linking theory to societal questions. To conduct such analyses, generate new research ideas, and navigate available commercial and open-source data sources, researchers need to understand which data sources academic papers analyze to arrive at their conclusions.

While scholarly citations are highly standardized and easy to use in machine-readable workflows, data citations often come in a variety of formats to date. The unstructured nature of data citations has complicated their automated extraction for the longest time. If data citations could be extracted automatically, this would give rise to a variety of valuable applications measuring data impact.

Enabled applications could include data usage networks, where extracted data citations could help researchers trace data usage across papers and domains. Extracted data citations could fuel applications, such as dataset recommender systems. Additionally, knowing where data is used enables data providing institutions to estimate of the value of dataset provision e.g. in terms of citations.

In this paper, we explore the potential of Large Language Models (LLMs) to identify mentions of data sources in research papers. Specifically, we analyze to what extent a widely adopted LLM, GPT-3.5, can accurately detect, describe, and summarize all specific mentions of data sources in research papers in the fields of economics and finance. We describe a proof-of-concept for a pipeline to extract data citations into structured lists.

To this end, we collect a large sample of research papers using web scraping techniques, thereby ensuring a sufficiently comprehensive and unbiased data coverage. We randomly sample and construct a human-labeled validation dataset to compute performance metrics of detection accuracy, i.e., to what extend the LLM is able to (i) detect all mentioned sources and (ii) whether it describes them accurately.

We find that we can advance extraction considerably compared to the status quo. Results are encouraging in terms of recall. In addition to gauging the ability of LLMs to extract data usage in research papers, we provide a guide targeted at implementing our solution at data providing institutions. By data providing institutions in this context, we refer to providers of official and commercial data, as well as research institutions.

Two recent trends enable us to take this step. First, recent technological progress in natural language processing (NLP) provides an opportunity to build a performant AI-driven reading system targeted at detecting, classifying, and connecting data source descriptions in research papers. Specifically, the increase in capabilities of large language models (LLMs), such as generative pre-trained transformer models, as popularized by ChatGPT, has been remarkable for our use-case.

Second, empirical research in economics increasingly utilizes granular data from administrative sources (e.g. Einav and Levin, 2014). Such datasets are often standardized, quality-controlled, and constitute known entities with fixed names, facilitating extraction. There is a trend towards providing this administrative data through dedicated research data centers (RDCs, Card et al., 2010).

RDCs have an interest in tracing their data, since if they can show the impact of the data they provide, this can motivate the use of public funds for data provision. Furthermore, the data providers can then target data provision efforts to impact-generating datasets. With these use-cases in mind, naturally extraction efforts have been numerous to date. In the next section, we outline the current state-of-the art in information extraction from research papers and our contribution.

2. Contribution to the status quo

2.1 Achievements to date

A large and interdisciplinary stream of prior literature has considered the task of automated extraction of dataset mentions from free texts, with or without the aid of NLP techniques. Since the widespread availability and use of LLMs is a relatively recent phenomenon, however, there is no specific prior research on the feasibility and usefulness of LLMs to this task at the time of writing and to the best of our knowledge.

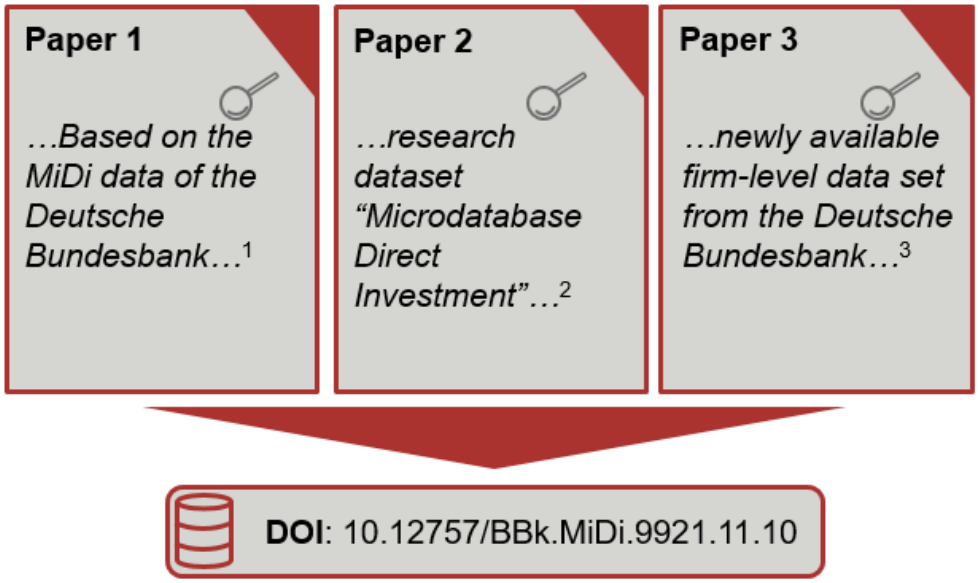


Figure 1: Examples for diverse citations of the same dataset. The dataset can be identified by the digital object identifier (DOI), however the DOI is often not present or cited. Sources: Authors' own depiction based on actual data citations in 1 Lipponer (2006), 2 Blank et al. (2020), and 3 Buch et al. (2005)

While one can achieve the desired outcome of automated information extraction from research papers using either (i) rule-based look-up tables, (ii) predecessor NLP models or (iii) LLMs, the use of the latter offers a hitherto unparalleled quality available to most users. Since there is a large number of a priori unknown ways to cite the same dataset, look-up tables and simple models tend to

have a low recall trying to identify used datasets (Figure 1 gives an example for the complexity of the extraction task with three actual ways a dataset is cited, exemplary for numerous further possibilities).

Data citations do not necessarily have to be in textual, unstructured form. Concepts for unique and persistent identification have been introduced in the literature, for example the Digital Object Identifier (DOI). The DOI can identify any physical or digital object and is used for dataset citations as well. However, DOI usage for data citations (at least in economics and finance) remains piecemeal, therefore text extraction using Natural Language Processing (NLP) methods remain to be the method of choice for the task at hand.

LLMs offer potential of democratizing NLP applications for the wider public and academia without requiring extensive prior knowledge of computer programming and machine learning (ML). Therefore, the arrival of these models, exemplified here by the LLM GPT-3.5, warrants a new inspection of the issue of extracting dataset and methodology information from research papers. Importantly, we aim to show how providers of official statistics can leverage on LLMs with a practical application.

Prior to the popularization of LLMs, several examples from the previous literature have shown efforts for automated information extraction from research papers. The extracted entities covered by the prior literature include citations (e.g. Saier and Färber, 2020, as one recent example), papers' focal points and topic trends (e.g., Chen and Luo, 2019), or tasks and methods (e.g., Houngbo and Mercer, 2012, Kovačević et al., 2012, Eckle-Kohler et al., 2013, Luan, 2018, Jayaram and Sangeeta, 2017, or Yao et al., 2019).

Most relevant to our work are instances studying specifically the automated extraction of analyzed datasets. Such examples include the work of Boland et al. (2012) and Ghavimi et al. (2016). In contrast to the majority of earlier work on the subject, the focus of the literature has recently shifted to the use of NLP and ML techniques to identify and extract entities such as datasets, methods, or results from research papers (e.g., Kardas et al., 2020, Kumar et al., 2021, or Gemelli et al., 2023).

The sub-stream of the prior literature focusing specifically on dataset mentions has itself given rise to the construction of open-access databases for research data usage and networks.[5] Wang et al. (2022) provide a recent systematic review of the literature on extracting and evaluating scientific methods used to gain insights from the ubiquitous data sources, which characterize current empirical research.

Importantly, however, the majority of earlier work on the subject share a potential issue: The authors generally focus on research papers from the data mining, knowledge discovery, data science, computer science, ML, or bibliometric disciplines. Since papers from these disciplines might be more likely to contain particularly well written data and methodology sections compared to other disciplines, the performance gains of the devised tools and methods might conceivably be biased upward.

Closely related to our focus on dataset usages, Kumar et al. (2021) study the extraction of dataset mentions in the scientific literature for reusing and replicating used datasets and ultimately building a dataset recommender system. Such a system

[5] See Otto et al. (2020) for a discussion of several such databases.

can help researchers adequately identify the datasets best suited to answer a given research question. Ikeda et al. (2020) aim to provide a similar pathway to this goal, taking care to study a large and multidisciplinary corpus of papers, both before the widespread availability of LLMs.

Examples from the prior literature, which use LLMs to extract data citations are scarce. Compared to traditional NLP and textual analysis methods, however, LLMs pose the unique potential to facilitate applying state-of-the-art NLP algorithms by removing all barriers to entry except for the users' ability to use application programming interfaces (APIs). Therefore, such models also hold the potential of dramatically simplify complicated information extraction pipelines compared to the more complicated approach of using NLP techniques 'by hand'.

Perhaps closest to our paper is the work of Polak and Morgan (2023), who propose ChatExtract, an automated data extraction method using conversational LLMs. It requires minimal initial effort to extract data from research papers by employing engineered prompts and follow-up questions similar to the approach used in this paper. The authors demonstrate that their tool is able to provide answers with around 90% precision and recall rates on materials data, suggesting that ChatExtract is a simple and transferable tool for data extraction across various fields.

2.2 Our value proposition

Importantly, however, Polak and Morgan (2023) focus on materials data and research papers, and they demonstrate the usefulness of their tool based on known properties. By contrast, we aim to provide a solution for users who do not know the sought properties a priori, i.e., researchers or RDCs for whom reading papers front to back in order to identify data citations is not feasible. Therefore, we go beyond ChatExtract for a wider task in question (identifying unknown datasets from previously unseen research papers).

Where most of the prior literature focuses on the natural sciences or ML/NLP papers specifically, our focus is on economics and finance as two large bodies of academic literature. These academic fields tend to lean increasingly on standardized datasets either licensed from commercial providers, provided by RDCs in public institutions, or owned and disseminated by single researchers within their professional networks. Often times for any given paper, joint data from multiple sources is used to provide novel insights.

This data availability structure implies that the automated identification of datasets in these bodies of the scientific literature is crucial to understand not only the impact of single datasets, but also how they are combined with other data sources. Importantly, these applications extend beyond the work of RDCs to data providers, researchers in search for adequate datasets to answer their research questions, and the literature on academic networks, to only name a few.

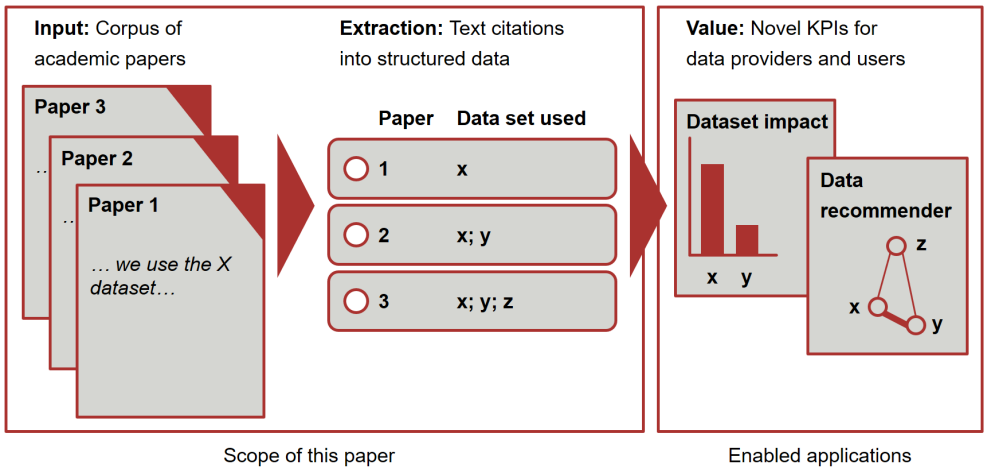


Figure 2: Scope of this paper and value proposition. Source: Authors' own depiction.

With these applications in mind, we aim to inform the informational basis. Specifically, we aim to fuel two use-cases by generating structured lists of data cited per paper. These two use-cases are (i) to be able to recommend data to researchers ("based on your interest, you might also like this data") and (ii) data impact factors ("This dataset justifies investment by generating high research or policy value"). We outline these value propositions in Figure 2**.** In the next section, we outline the data we use to extract the necessary data to inform these use-cases.

3. Data

3.1 Sample collection

In order to provide a feasibility study for Central Banks, official statistics offices and academia to obtain structured citations data from textual data mentions in papers, we use data from the Discussion Paper series of Deutsche Bundesbank. This series comprises papers written by researchers including one or more employees of the Bundesbank.

Sample

Corpus of research papers for automated extraction of data citations and evaluation sample Table 1

	Source	Sampling	Number of papers	Sample period	Labeling
Corpus to be labeled	Bundesbank discussion paper series	Full universe	1,102	02/1995 – 07/2023	Using GPT 3.5
Evaluation sample	Bundesbank discussion paper series	Random sample	103	06/1995 – 04/2023	Manually

Source: Authors

Our input sample for the dataset extraction pipeline comprises all research papers released under the Discussion Paper series of Deutsche Bundesbank between February 1995 and July 2023. The Deutsche Bundesbank discussion paper series covers a variety of different topics, such as monetary policy, banking supervision, and household economic behavior. In total, this corpus includes 1,102 research papers at the time of our data collection, published between 1995 and 2023. This constitutes the corpus to be labeled. Subsequently, it is straightforward to extend our approach to label any corpus of academic papers.

From the corpus to be labeled, we randomly select 103 discussion papers as an evaluation sample. In order to obtain a baseline to evaluate the performance of our approach, we label these papers manually. By labeling in this context we refer to reading a paper and manually noting the datasets used (if any). These manually labeled papers constitute our evaluation sample for the dataset extraction pipeline. This means that the performance evaluation metrics introduced and discussed in Sections 4 and 5 are based on estimated averages across this evaluation sample.

3.2 Manual labeling of evaluation sample

For manual labeling, we randomly assign 20 or 21 papers each to five human readers. Human readers scan the papers for specific dataset mentions. Figure 1 illustrates that dataset mentions differ widely, lacking a standardized format and standardized sections within papers. Therefore, the human readers need to read all papers thoroughly to find dataset citations.

Labeling guide

Step-by-step instructions used for human labeling of evaluation sample Table 2

Step	Instruction
1	Open the folder with the list of pdf assigned to name
2	For each pdf
	- open the relevant pdf
	- read paper for mentions of dataset used for empirical analysis
	- if mention found, record all synonyms used for the dataset(s) that you find in the text
	- if mention found, record the entire passage that describes the dataset(s) that you find in the text (this can range from one sentence to an entire paragraph or more)
3	Save resulting json file

Source: Authors

For all identified datasets in the evaluation sample, we record two types of labels: short labels and long (verbose) labels. The latter corresponds simply to the full text passage where each used dataset is discussed. In this step, we record only the first

mention per dataset, omitting repeated (and differing) descriptions of the same dataset.

For short labels, we record all synonyms of the dataset names for each of the manually labeled papers. If authors refer to the same dataset with different names within or across papers, recording only one strict definition as the ground truth will introduce a downward bias on the evaluation of the algorithm's performance.

If we force our algorithm to find exact matches for the identified datasets, all datasets identified with another name than that specifically mentioned in the paper would be recorded as unidentified, i.e., negatives. This is more severe when definitions are long, since the chance of *not* finding each of, e.g., ten words in the correct order is generally higher than for, say, definitions of three words length.

In addition, the ground truth will likely not be represented in any one paper, i.e., it is unlikely that several researchers use the same definition for the same dataset except for instances where dataset names are extremely short and standardized. This means that we would not be able to aggregate or compare results for the same dataset across papers if we do not undertake some sort of standardization of the dataset mentions.

When manually reading and labeling papers, we notice several ambiguities, proving the complexity of the extraction task not only for machines but also for humans (as described by Bender et al, 2020). Beyond the wide range of synonyms used to describe datasets, further particularities arise. While labeling, it is not always clear, what constitutes a dataset name. This can be present in cases, when there are descriptions in textual form instead of an actual dataset name (compare example in Figure 2). We further decide not to consider single time series or model calibration parameters as datasets.

As a result of the manual labeling process, we end up with a corpus of papers to be labeled and an evaluation sample, where we know (i) which datasets are used within, (ii) synonyms of the datasets used, if any, and (iii) the text passage, where the dataset was mentioned. With this data ready, we proceed in the next section by outlining the methodology used, specifically the extraction pipeline we built.

4. Methodology

To extract data citations automatically from the corpus of papers obtained, we construct a pipeline for assistant-style LLMs. Our goal is to build a flexible process allowing us to (i) test different variations efficiently, (ii) track the results systematically, and (iii) easily transfer code to other interested parties. In the following, we outline the components of this pipeline, including the measures we use to judge performance.

4.1 Pipeline overview

The pipeline takes as input a paper in the form of a PDF file and produces a list of datasets as output. Our pipeline can be segmented into five steps. First, we split papers into smaller segments. Second, we filter out irrelevant segments, before

– third – combining each segment with the instruction for the LLM. In the fourth step, we ask the LLM for the information contained, before finally consolidating the responses in step 5 (compare Figure 3 for a graphical depiction of the dataset citation extraction pipeline used).

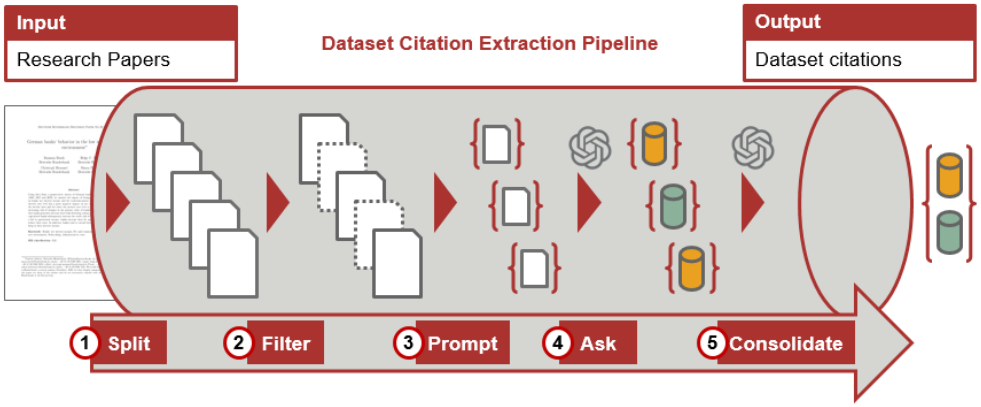


Figure 3: The Dataset Citation Extraction Pipeline used for the purpose of this study. Source: Own depiction.

In the first step, we split the document into smaller segments, or snippets. The main goal here is to ensure that each snippet will be small enough to fit into the limited context window of the LLM. A context window is the zone, where text can be analyzed coherently by the LLM, so to speak the length of the "conversation" per paper we can engage with the model. In our case, we use GPT-3.5.[6] An intuitively simple approach could be e.g. to treat each page of the document as one snippet.

While page-splitting is easy to implement, this has the drawback of splitting an ongoing paragraph into two at the page boundary. In fact, instances where contextually connected paragraphs do not end at the end of a page are most frequent, causing the LLM to consider only part of the true context of the respective page-paragraph combination at a time. A mitigation of this issue is to define an overlap between the chunks. We achieve this with `langchain`, a solution tailored to application building with LLMs.

Alternatively, we evaluate splitting documents by paragraphs by using multiple new-line commands as split separators.[7] This yields a larger number of smaller snippets

[6] More precisely, we use the `gpt-3.5-turbo` model of the OpenAI API. It is technically possible and might be sensible to compare the performance to GPT-4 with a longer context window in the future. The model used at the time of writing allowed a context window of 4,000 tokens, also referred to as 4k (approximately 5 pages of text). A token can be thought of as a piece of a word, one token is approximately 4 characters in English language. Newer versions of GPT-3.5 allow for a 16k token context window (approximately 20 pages). Recently, a 128k token context window was announced for GPT-4 (encompassing approximately 300 pages of text). In the interest of cost efficiency and technological availability at the time of writing, we relied on GPT-3.5 as a trade-off between performance and cost, but note that in the future most likely larger snippets will be used.

[7] e.g., "\n\n" in Python

compared to a page split, which leads to higher API costs but might improve performance.

The second, optional, step is to pre-filter the snippets for informative content. Independently of the splitting approach used, the text is likely to contain several snippets without relevant information for dataset extraction. Examples of such uninformative text snippets include (most) tables containing numbers, the references section or small fragments without textual content.[8] Wherever possible, we remove such snippets from consideration for further processing, prior to involvement of the LLM, based on simple heuristics.

Next, the text snippets from the paper have to be combined with instructions. This creates a prompt for a zero-shot task (compare e.g. Sun et al., 2021).[9] A prompt in this context is the plain English input we send to the model. The exact language model used determines the particulars of how to phrase the instructions. Depending on how the underlying model was designed and trained, it might allow differentiation between "system" and "user" prompts or allow fill-in-the-middle prompts.

The quality of the generated responses also differs significantly depending on the phrasing of the instructions. As it has been noted in the literature, that zero-shot prompts can outperform prompts with examples of task-completion (Reynolds & McDonell, 2021), our pipeline allows the systematic tracking of changes in output quality caused by alternative instructions. The prompts we use for this analysis can be seen in Appendix 1. In short, we ask the model identify a list of data sources, or simply write "None" if it cannot find any.

In the fourth step, we send the prompt to the LLM, consisting of instructions and text content. At this point different sampling parameters can be set. Of particular interest to us is the "temperature" parameter, which rebalances the probability distribution for the next token in the generated sequence, with high temperature corresponding to a more uniform distribution.

High temperature is useful to generate more creative output and to prevent stiff, predictable writing style. In our use-case, where we are interested in simple retrieval of factual information from a given text, we set the temperature as low as possible. Our pipeline also takes care of any connectivity issues, for example due to downtime or throttling. The result of this step is an answer to our inquiry for each individual text snippet.

In the fifth and last step, individual model replies per snippet must be consolidated into a single answer covering the entire input paper. Any single data source is likely to be mentioned multiple times across several sections of a research paper. The identification of the union of all the answer-sets is also left to the LLM. The result is supposed to be a single list of all unique data sources used in the processed research paper. We evaluate, how close the automatically-generated list is to what a human reader would expect in the next section.

[8] Tables could be informative if they contain dataset information. This can be the case, for instance, when an Appendix table includes information on further datasets used.

[9] Even though language models in our context technically are few-shot learners

4.2 Performance measures

Evaluating the consolidated results offers some interest interesting aspects. For once, LLMs like GPT-3.5 have been reported to hallucinate ("confabulation"). Further, our evaluation aims to compare two lists of free text (the consolidated model output list and the manually labeled evaluation sample), requiring fuzzy string comparisons.

While our application of fact-retrieval from a given text is not particularly susceptible to confabulation, we nonetheless need to measure the quality of the responses that our pipeline produces. This is also crucial information to test the impact of changes to our processing pipeline.

Determining the performance of the model and pipeline by comparing its output to a list labeled by a human is not trivial. In essence, we would like to gauge the similarity of the two lists. The generated lists being free text, rather than a subset of predefined labels complicate this. While our approach allows the model to identify data sources beyond our current awareness, the output then necessitates the use of fuzzy string comparisons. In Figure 4, we show the evaluation process pipeline.

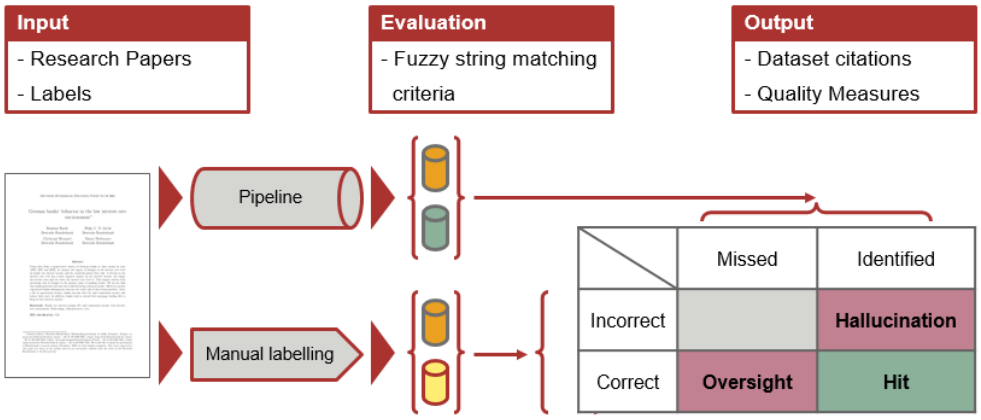


Figure 4: Assessing the quality of the model's output. Source: Own depiction

We explore different versions of this fuzzy comparison, highlighting the tradeoff between the error types based on the strictness of the match. We calculate model recall (fraction of identified "true" data sources, regardless of spurious identifications) and output similarity measured by various criteria.

For the task at hand, the model can make two kinds of errors, oversights and hallucinations. Oversights are cases, where a truly mentioned data source is not recognized as such by the model and thus not listed in the output. Hallucinations are incorrect identifications, where the models output includes some entry that does not objectively count as a data source (either because the model misunderstood or confabulated).

These two kinds of errors are not equally important in our case. For our use case, we prefer the inclusion of an inaccurate entry in the result set to the possibility of overlooking a true data source mention, as the former can more easily be detected

and corrected downstream. In other words, our current focus is predominantly on recall at this step.

5. Results

When evaluating recall of our approach (i.e. the fraction of correctly identified dataset citations as a fraction of all dataset citations), we apply several criteria to compare model predictions with our evaluation sample. We (i) let the LLM judge its own results as described above, (ii) compare exact string matches, (iii) fuzzy string matches, (iv) fuzzy string matches including synonyms, and (v) the Jaccard similarity.

As these criteria are not equally strict in comparing similarity, measures of recall vary widely. In the strictest scenario, when we only count exact matches, we reach a recall of only 12% (17% conditional on the fact, that the model predicted any data to be contained at all). This result comes maybe unsurprisingly, as it signifies the variety of dataset names, i.e. complexity of the extraction task.

When we ask the LLM to evaluate its own results, we obtain a 34% recall (62% when considering only papers, where the model predicted any data). While elegant in the context of our pipeline to leverage the LLM for evaluation, too, this criterion remains a black box, as the exact evaluation metric is subjective to the model and users are unaware a priori of the direction of a potential bias.

The mildest evaluation measure is likely fuzzy synonym search, where any variations of synonyms of the dataset that are found are counted as a hit. Arguably, this may be a valid measure for our use-case, as a synonym of a dataset name is often just as correct as any other. Using this criterion, our approach finds 71% percent of dataset citations in the papers provided. Figure 5 depicts recall results for all evaluation criteria used.

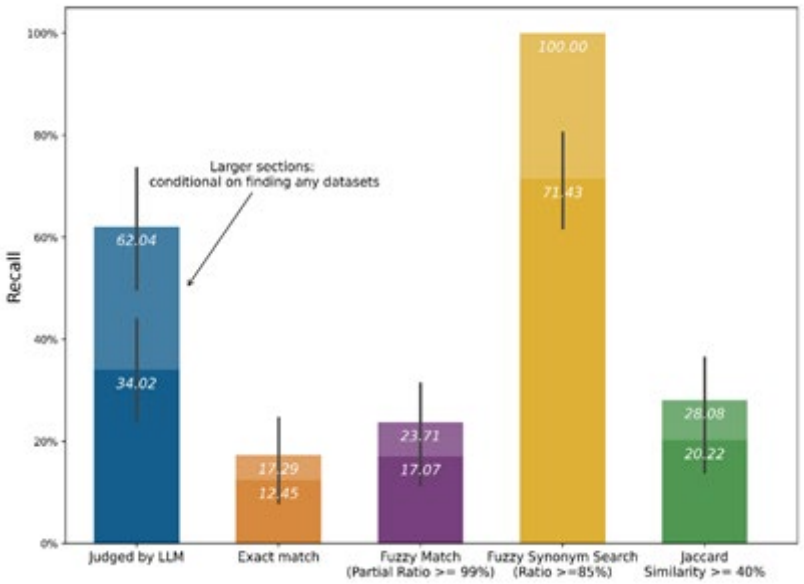


Figure 5: Performance measures of model recall based on different criteria. Source: Author's own calculations.

Note that fuzzy synonym search is not necessarily always feasible to evaluate results, if our pipeline were to be applied in another context, as it requires some domain knowledge to know synonyms of dataset names. However, providers of datasets often possess such domain knowledge. Further note that such domain knowledge is only needed for evaluation using synonyms, not to algorithm performance itself.

As the recall numbers vary considerably depending on the exact criteria used, due to the loosely defined task of comparing lists of strings, we feel compelled to comment briefly on our subjective feeling. When looking at the similarity of lists ourselves, we were surprised by the capability of the LLM to identify datasets correctly, even in the case of complicated names or infrequently used datasets.

Users of the automated dataset extraction pipeline might be interested in inspecting the performance of the LLM assistant on an individual dataset level. Here, we notice a very high ability of the algorithm to identify well-structured and well-defined datasets, in our case due to the corpus of input papers particularly from administrative data sources (datasets originating from Deutsche Bundesbank) and commercial providers. A particularity of these datasets is that they are usually clearly defined, named, and separable entities.

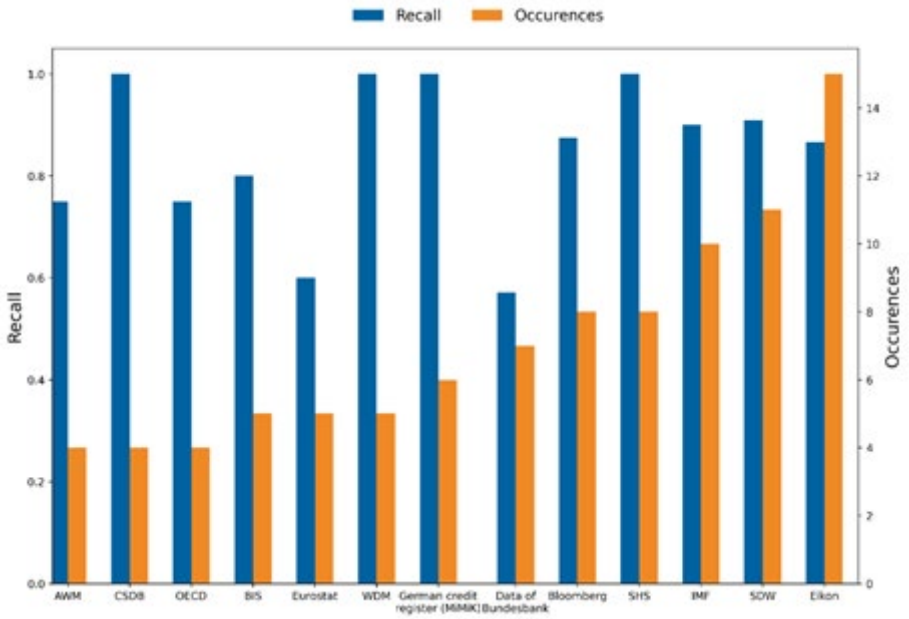


Figure 6: Recall per dataset, judged by LLM. Source: Authors own calculations.

Furthermore, the existence of relatively unique abbreviations seem to support automatic dataset extraction (e.g. "CSDB" and "SHS"[10]). Figure 6 shows recall for a number of selected administrative datasets in our sample. Before deriving ambitious conclusions from this finding, the scope of the extraction would have to be enlarged, as the number of occurrences per dataset in our sample remains limited. However, we take this as an encouraging sign for the potential of our approach.

Even considering the preliminary nature of results as part of the presented proof-of-concept, it seems already efficient to streamline the process of identifying data usage in papers by using an LLM, since the baseline to date is human labeling. This holds true even if model performance itself is deemed insufficient for certain use-cases (whereas we would argue that for our use-case it is already sufficient), as long as pre-labeling by the model increases efficiency of subsequent human labeling. Hence, using our pipeline seems to yield a fairly reliable gauge of data impact, when identifying known datasets in unknown papers.

6. Conclusions

Both, data providing institutions and data users such as empirical researchers have an interest in tracing how datasets are used. If structuredly available, information on how datasets are used in research papers would allow data providers to measure data impact and researchers to inspect data usage in a systematic way. In other words, the whole infrastructure around research paper citations (such as paper and journal

[10] "Central Securities Data Base" and „Securities Holdings Statistics", two large and well-documented European administrative datasets used for non-commercial research.

impact and citation networks) could be recreated for datasets, benefiting providers and users of data.

To date, dataset citations in research papers have not been systematically used, due to the unstructured textual nature of dataset citations. Recent advancements in NLP allow us to progress on the status quo. We present a pipeline to leverage LLMs, specifically GPT-3.5, to extract dataset citations and evaluate performance using a manually labeled evaluation sample.

We find that LLM performance is good in terms of recall. The LLM correctly identifies many of the datasets used in the papers presented to the LLM. Performance in identification is particularly good for known datasets, where domain knowledge in the form of known used synonyms of dataset names can be provided to the algorithm. However, measures of recall vary to a fair degree due to the unclear defined task of comparing two lists.

We argue that model performance is sufficiently high to measure data impact, since the proof-of-concept already offers value-generating use-cases. First, our approach considerably advances on the status quo, where manual labeling to trace data impact is often prohibitive in terms of time and effort. Here, we offer a viable pre-screening option to make tracing for humans feasible. Second, even if there are no human capacities available, our approach allows tracing data impact over time reliably, as long as dataset citation practices remain constant in a field.

Our pipeline is easily transferrable to other institutions interested in tracing their data usage in research or researchers interested in data citation networks. We use GPT 3.5 for the results described here, however the LLM can be exchanged relatively simply. Exchanging the LLM at the core of our pipeline allows interesting extensions and likely quality gains in the near future, with the quick pace of increasing LLM capabilities we currently witness.

Possible extensions include leveraging on-premise LLMs to allow tracking data impact in confidential texts, without concern of sharing confidential data outside the institution. If in a given institution, data impact is characterized predominantly by its usage in internal data-driven decision-making, it could make sense to evaluate data mentions in confidential texts, such as internal briefs, board meeting minutes, etc.

Our proof-of-concept allows important lessons learned for data providers. Data impact can now be traced automatically and efficiently. Therefore, it is now possible to provide data knowing how it is used, i.e. with a more clear user-centric focus. Therefore, data-driven decision-making and knowledge-generation can be enhanced in several key aspects.

By knowing the impact of datasets, (i) resources can be focused on value-creating data, (ii) underused data can be identified and potential usage hurdles removed, (iii) popular datasets can be promoted, (iv) combined with researcher information, tailored datasets can be recommend, and (v) frequently jointly used datasets can be provided together. With these value propositions in mind, our next step is to ingest a larger corpus of research papers to leverage on the next generation of LLMs.

References

Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K. E. & Jaradeh, M. Y. (2020). Improving access to scientific literature with knowledge graphs. *Bibliothek Forschung und Praxis, 44*, 516–529.

Bender S., Doll H. C., & Hirsch, C. (2020). Where's Waldo? Conceptual Issues when Characterizing Data in Empirical Research. In: Lane, J., Mulvany, I., & P. Nathan (Eds), "Rich search and discovery for research datasets: Building the next generation of scholarly infrastructure", Sage Publications.

Blank, S., Lipponer, A., Schild, C. J., & Scholz, D. (2020). Microdatabase Direct Investment (MiDi)–A full survey of German inward and outward investment. German Economic Review, 21(3), 273-311.

Boland, K., & Krüger, F. (2019). Distant Supervision for Silver Label Generation of Software Mentions in Social Scientific Publications. *BIRNDL@ SIGIR*, (pp. 15–27).

Boland, K., Ritze, D., Eckert, K., & Mathiak, B. (2012). Identifying references to datasets in publications. *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings 2*, (pp. 150–161).

Buch, C. M., Kleinert, J., Lipponer, A., Toubal, F., & Baldwin, R. (2005). Determinants and effects of foreign direct investment: evidence from German firm-level data. Economic Policy, 20(41), 52-110.

Card, D., Chetty, R., Feldstein, M. S., & Saez, E. (2010). Expanding access to administrative data for research in the United States. *American economic association, ten years and beyond: Economists answer NSF's call for long-term research agendas*.

Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. *Advanced Engineering Informatics, 42*, 100959.

Eckle-Kohler, J., Nghiem, T.-D., & Gurevych, I. (2013). Automatically assigning research methods to journal articles in the domain of social sciences. *Proceedings of the American Society for Information Science and Technology, 50*, 1–8.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*, 1243089.

Färber, M., Albers, A., & Schüber, F. (2021). Identifying Used Methods and Datasets in Scientific Publications. *SDU@ AAAI*.

Gemelli, A., Vivoli, E., & Marinai, S. (2023). CTE: A Dataset for Contextualized Table Extraction. *arXiv preprint arXiv:2302.01451*.

Ghavimi, B., Mayr, P., Lange, C., Vahdati, S., & Auer, S. (2016). A semi-automatic approach for detecting dataset references in social science texts. *Information Services & Use, 36*, 171–187.

Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2019). Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. *arXiv preprint arXiv:1906.09317*.

Houngbo, H., & Mercer, R. E. (2012). Method mention extraction from scientific research papers. *Proceedings of COLING 2012*, (pp. 1211–1222).

Ikeda, D., Nagamizo, K., & Taniguchi, Y. (2020). Automatic identification of dataset names in scholarly articles of various disciplines. *International Journal of Institutional Research and Management, 4*, 17–30.

Jayaram, K., & Sangeeta, K. (2017). A review: Information extraction techniques from research papers. *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, (pp. 56–59).

Kardas, M., Czapla, P., Stenetorp, P., Ruder, S., Riedel, S., Taylor, R., & Stojnic, R. (2020). Axcell: Automatic extraction of results from machine learning papers. *arXiv preprint arXiv:2004.14356*.

Kovačević, A., Konjović, Z., Milosavljević, B., & Nenadic, G. (2012). Mining methodologies from NLP publications: A case study in automatic terminology recognition. *Computer Speech & Language, 26*, 105–126.

Kumar, S., Ghosal, T., & Ekbal, A. (2021). DataQuest: An Approach to Automatically Extract Dataset Mentions from Scientific Papers. Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23, (pp. 43–53).

Lipponer, A. (2006). Microdatabase direct investment-MiDi. A brief guide. Bundesbank working paper, Frankfurt.

Luan, Y. (2018). Information extraction from scientific literature for method recommendation. *arXiv preprint arXiv:1901.00401*.

Otto, W., Zielinski, A., Ghavimi, B., Dimitrov, D., Tavakolpoursaleh, N., Abdulahhad, K., . . . Dietze, S. (2020). Knowledge extraction from scholarly publications: The GESIS contribution to the rich context competition.

Polak, M. P., & Morgan, D. (2023). Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering–Example of ChatGPT. *arXiv preprint arXiv:2303.05352*.

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7).

Saier, T., & Färber, M. (2020). unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. *Scientometrics, 125*, 3085–3108.

Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models. PLOS Digital Health, 2(11), e0000388.

Sun, X., Gu, J., & Sun, H. (2021). Research progress of zero-shot learning. Applied Intelligence, 51, 3600-3614.

Wang, Y., Zhang, C., & Li, K. (2022). A review on method entities in the academic literature: extraction, evaluation, and application. *Scientometrics, 127*, 2479–2520.

Yao, R., Hou, L., Ye, Y., Zhang, J., & Wu, J. (2019). Method and Dataset Mining in Scientific Papers. *2019 IEEE International Conference on Big Data (Big Data)*, (pp. 6260–6262).

Yoo, S., & Jeong, O. (2020). Automating the expansion of a knowledge graph. *Expert Systems with Applications, 141*, 112965.

Appendix

Box 1

Prompts

The final prompt used reads as follows

```
A dataset is a collection of structured or unstructured data that is organized and grouped together for a specific purpose. It typically consists of multiple data points or observations related to a particular topic or subject. A dataset can include various types of information such as numerical values, text, images, audio, video, or any other form of data. It is often used in the context of data analysis, machine learning, and statistical research, where the data is utilized to extract insights, train models, or draw conclusions. Datasets can be generated through various means, including surveys, experiments, observations, or by gathering existing data from different sources.

    The text after the empty lines is a scientific paper excerpt

    According to the definition on the first line, search for any mentions of datasets or data-sources used in the paper's research.

    If you find any, please compile a list of all mentioned datasets in this excerpt.

    If there aren't any, please reply with 'None'.
```

Consolidation prompt

```
You will be provided one or more lists of datasets.

    Each new list starts with '=>'.

You need to combine all the lists into a single one by removing redundant entries. Delimit the final list with simple '-' bullet points.
```

Recall prompt

```
You will be provided with text delimited by triple quotes that is supposed to be a list of datasets. Check if the following true datasets are directly contained in the answer:

    ...

    For each of these true datasets perform the following steps:

    1 - Restate the true dataset

    2 - Write 'yes' if the true dataset is mentioned in the answer, otherwise write 'no'

Finally, provide a count of how many 'yes' findings there are. Provide this count as {"count":<insert count here>}.
```

DEUTSCHE BUNDESBANK EUROSYSTEM

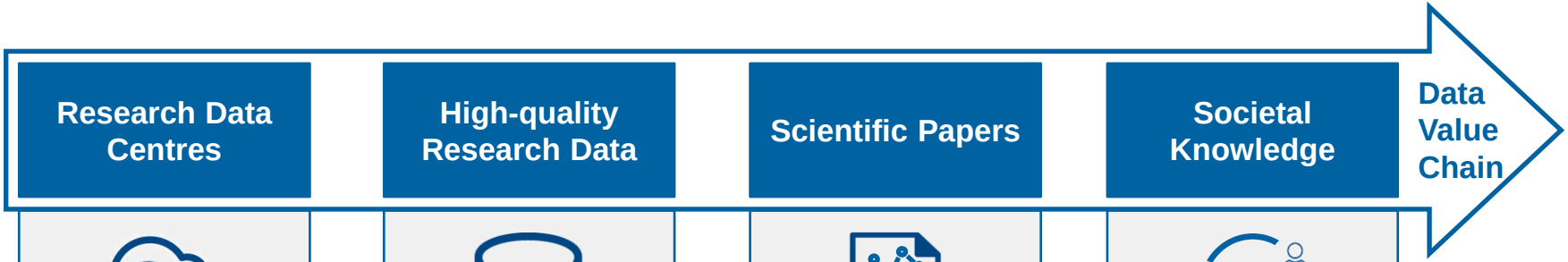
Leveraging Large Language Models to Extract Data Citations

3rd IFC Workshop on Data Science in Central Banking

Sebastian Seltmann, Emily Kormanyos, Hendrik Christian Doll, Shir Frank, and Kilian Graef – Deutsche Bundesbank, Data Service Centre

The authors would like to thank colleagues in Bundesbank's Data Service Centre for valuable comments and feedback.
All views expressed in this report are personal views of the authors and do not necessarily reflect the views of Deutsche Bundesbank or the Eurosystem.

The Data Value Chain



In social sciences, publically funded research data centres (RDCs) have proliferated in recent years*

RDCs provide high-quality granular data from administrative source low-cost or free of charge

Empirical research papers use data to derive insights

Knowledge from empirical research enables public debate and data-driven decision making

Since public funds are used, the public has an interest in knowing the value created by the investment

Issue: It is hard to measure data impact, as data citations in scientific papers are not standardised

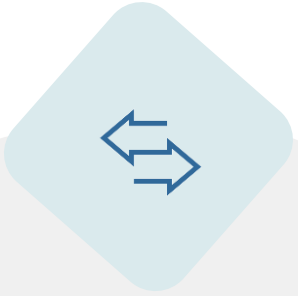
* For details on how RDCs work, compare slides in appendix

Extracting Data Citations from Academic Papers



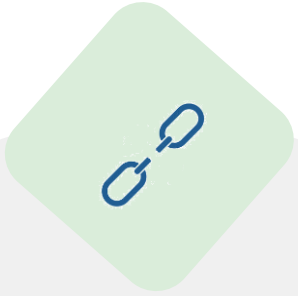
Theory

Blaschke & Hirsch (2023)[1] attempt to **measure the value of an RDC**, leveraging on manually collected info on projects and publications



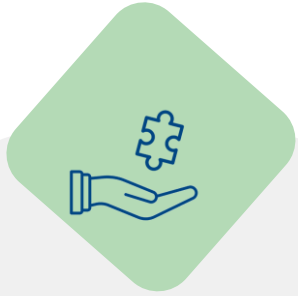
Related Work

Polak & Morgan (2023)[2] propose **ChatExtract to extract materials data in research papers**, leading to accurate results when identifying known properties



Gap in Literature

Most of the literature to date seems to **focus on the natural sciences or ML/NLP papers** specifically

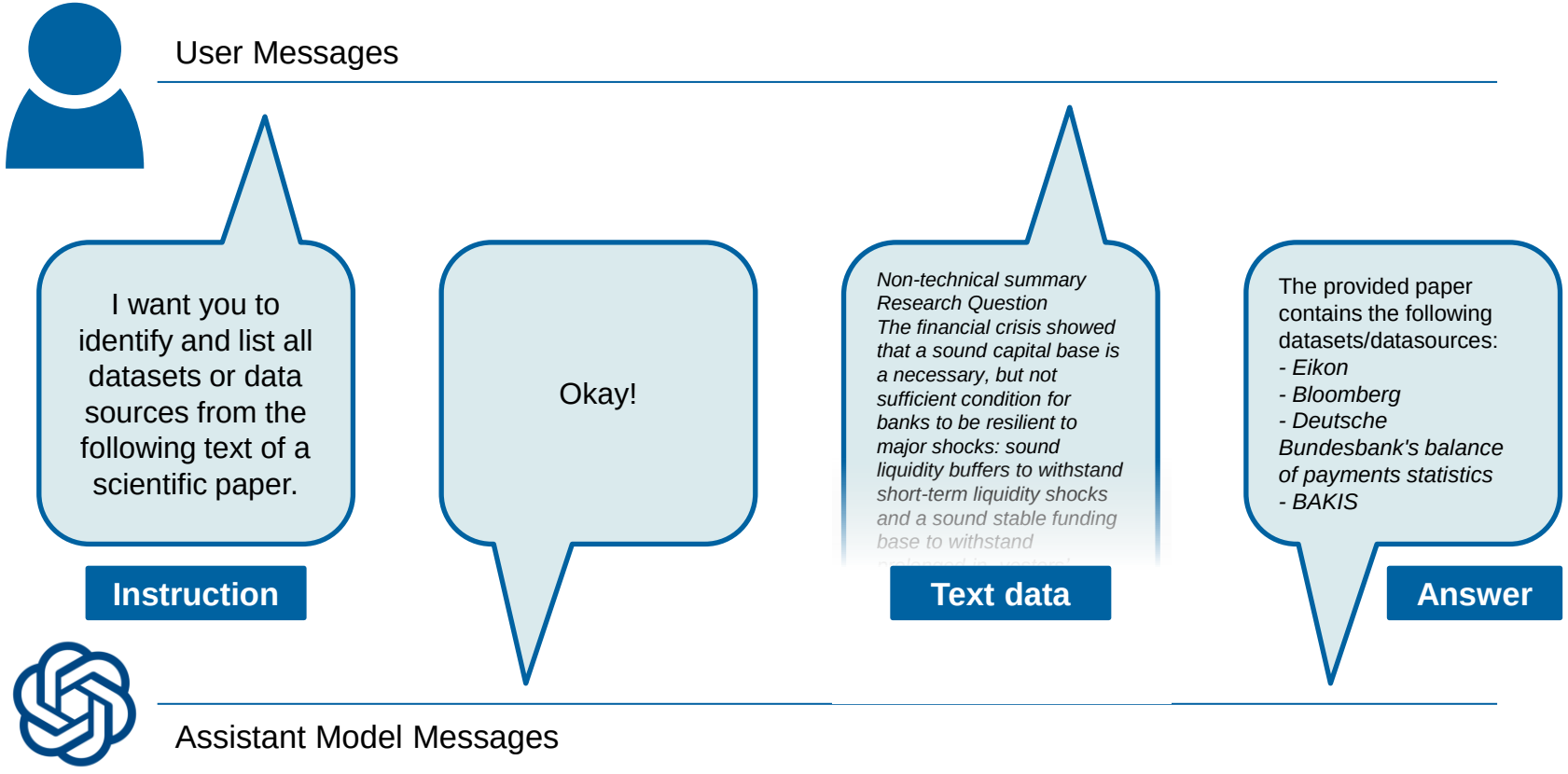


Our Contribution

Recent advances in natural language processing (NLP) enable us to flexibly **detect and connect data source descriptions from academic papers using GPT-3.5**

[1] Blaschke & Hirsch (2023)
[2] Polak & Morgan (2023)

The Assistant-Style Language Model Approach



Paper Corpus and Evaluation Sample



CORPUS TO BE LABELLED

- From **Bundesbank discussion paper** series
- Data sample: **1,103** papers
- Papers focus on economics and finance

EVALUATION SAMPLE

- Focus on **empirical papers using databases** (not data used for model calibrations)
- We draw a **random sample** for labeling to ensure **unbiased** evaluation metrics
- **103** papers manually labelled

Dataset Citation Extraction Pipeline

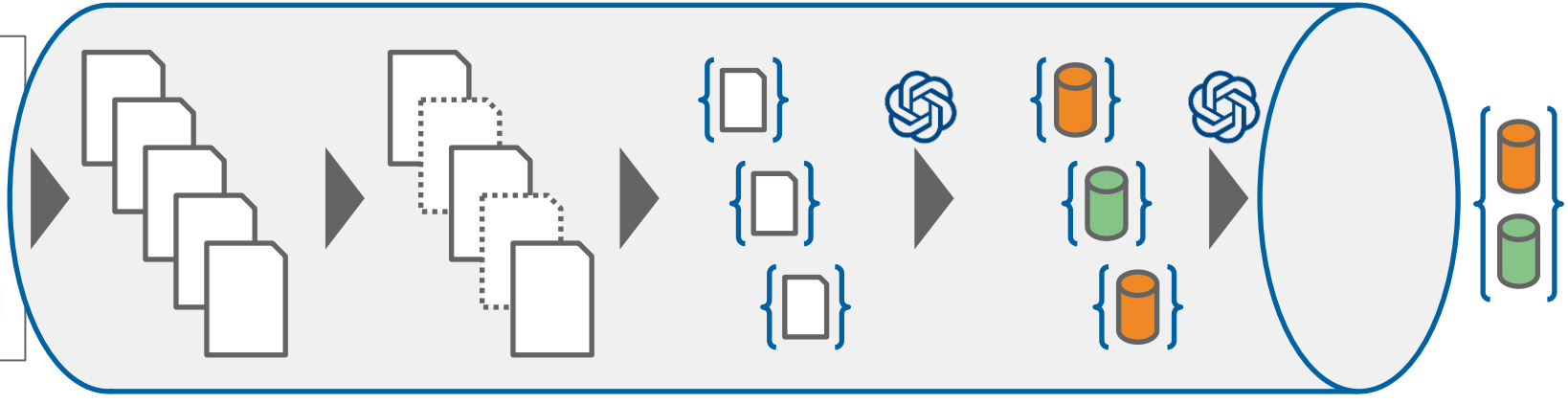
Input

Academic Papers

Output

Data Sources





Split Filter Prompt Ask Consolidate

Assessing the Quality of the Assistant-Model's Output

Input
- Academic Papers
- Labels: Data Sources

Output
- Data Sources
- Quality Measures

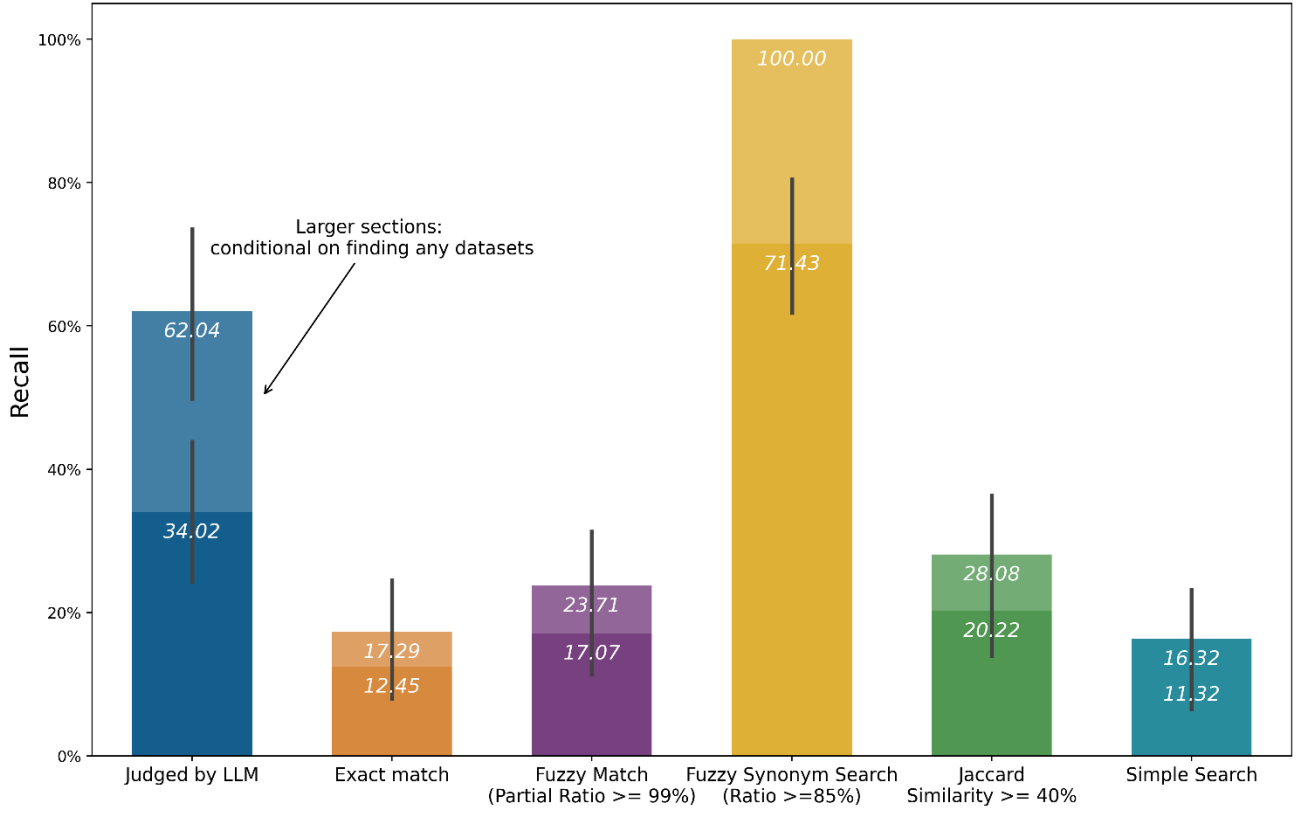


Pipeline

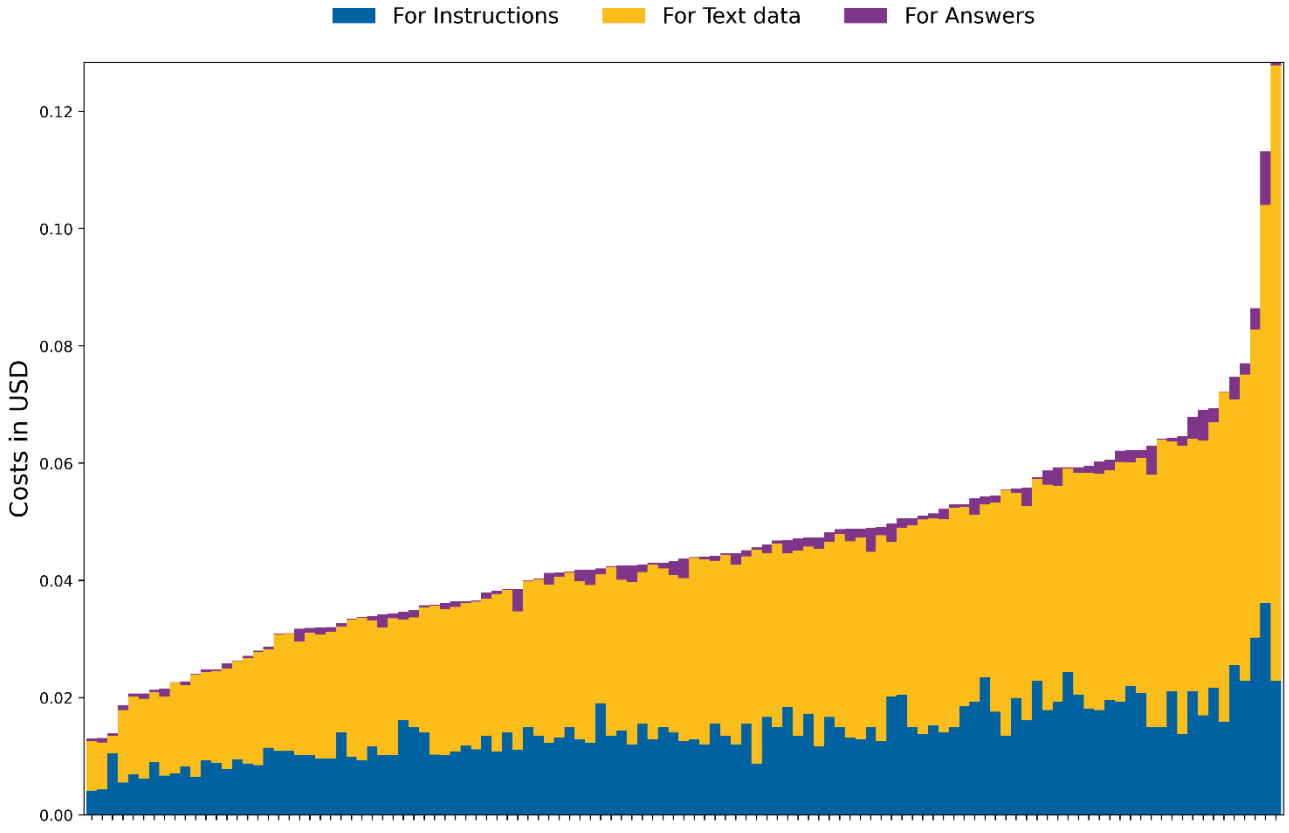
Manual labelling

	Missed	Identified
Incorrect		**Hallucination**
Correct	**Oversight**	**Hit**

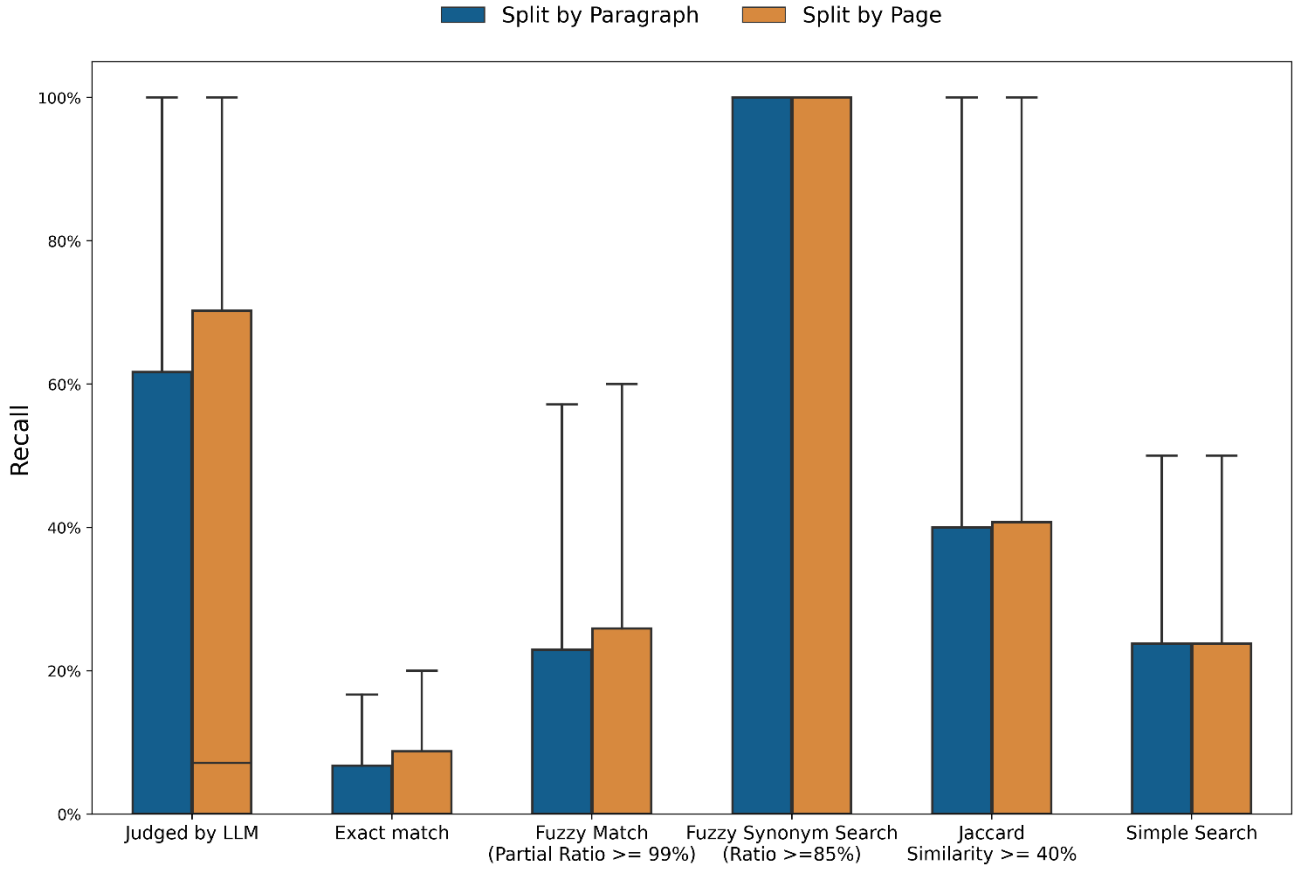
Results
Different measures of apparent performance



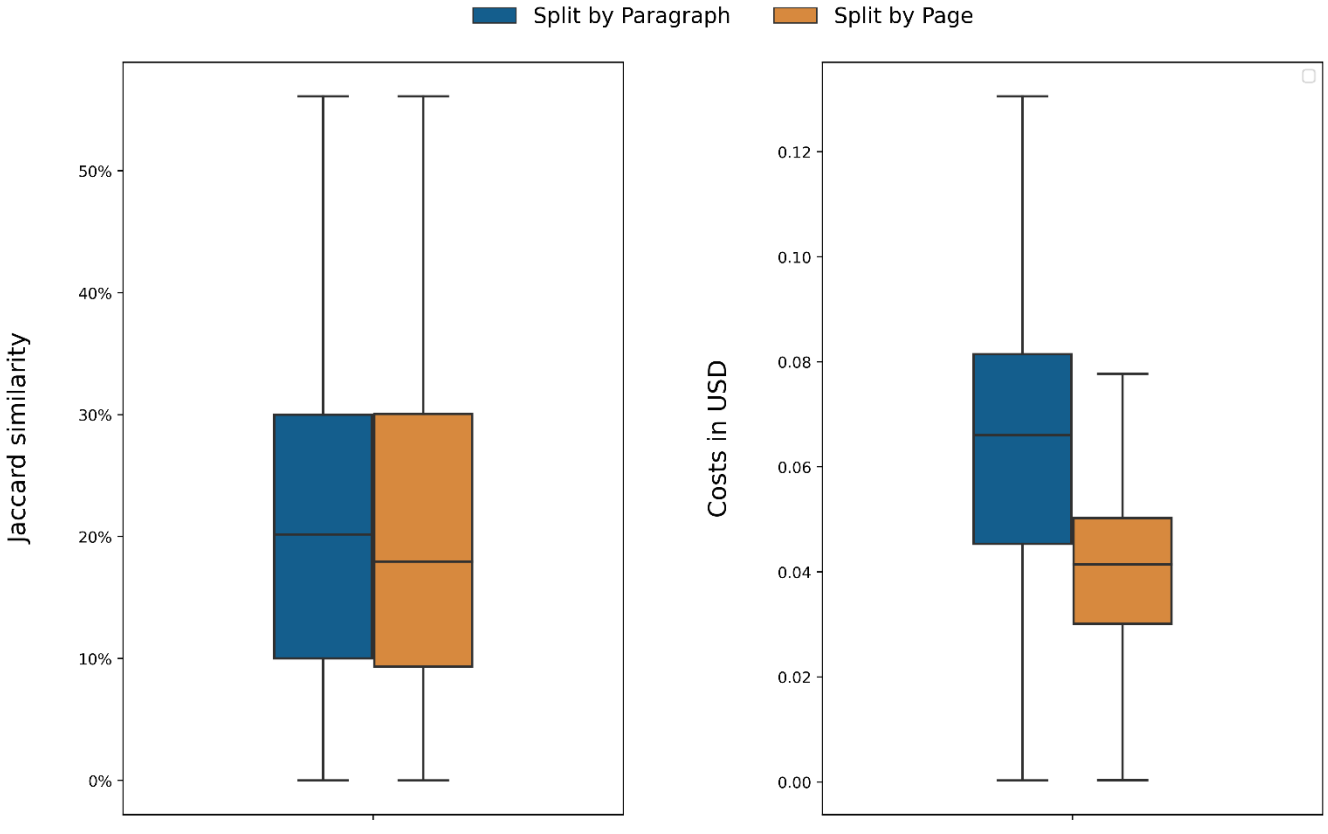
Results
API Costs per paper processed



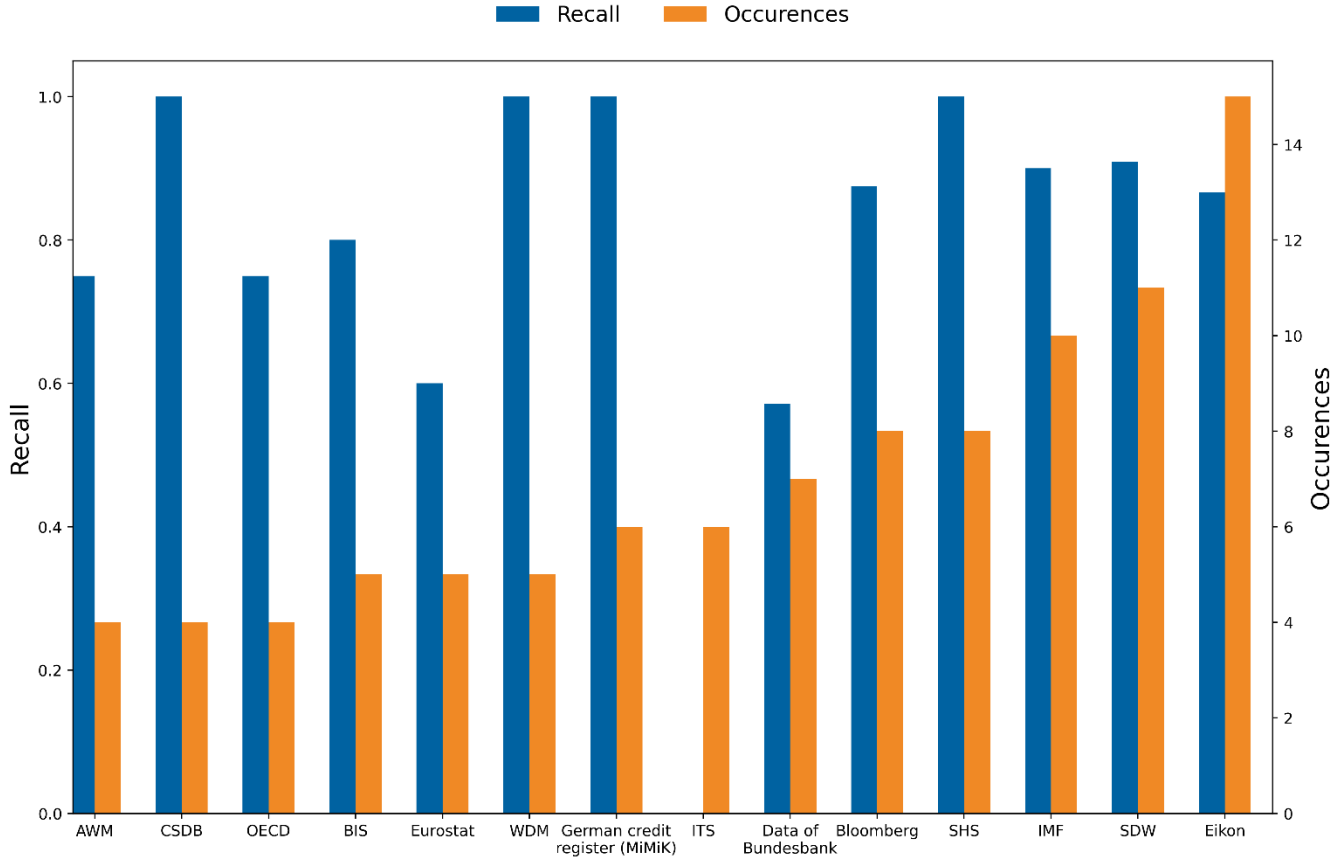
Results
Larger Contexts provide minor performance improvement



Results
Larger Contexts significantly reduce overall costs



Results
Recall for individual datasets



Challenges & Mitigations

DEPENDENCY

Reliance on the OpenAI API is a point of failure
Unforeseen outages are possible

- OpenAI transparently reports system status
- Pipeline can plug-in alternative models

UNSTRUCTUREDNESS

Assistant-Style language models produce plain text, strict output-syntax is merely a suggestion

- Zero “temperature” improves reliability

TESTING

The definition of success itself is up to debate and interpretation

- Pipeline tracks as many potentially relevant dimensions as possible

Future Extensions and Applications

RESULT	EXTENSION	DETAIL
Corpus of academic papers, labelled with data sets used	**Test other LLMs**	Leveraging GPT4 could improve results, while LLaMA can run locally
	Data usage networks	Explore which datasets are commonly used together
	Data recommender	„*Researchers who worked with Dataset A often also used Dataset B*“
	Data impact factor	„*Dataset A has a high impact because it fueled X journals publications*“

Key Take-Aways

1

CAPABILITY

Assistant-style language models can effectively extract specific types of information from large bodies of unstructured text

2

EVALUATION

Think carefully about how to measure success and track experiments systematically

3

GUIDANCE

Give the model clear and simple tasks, yet as much context as possible



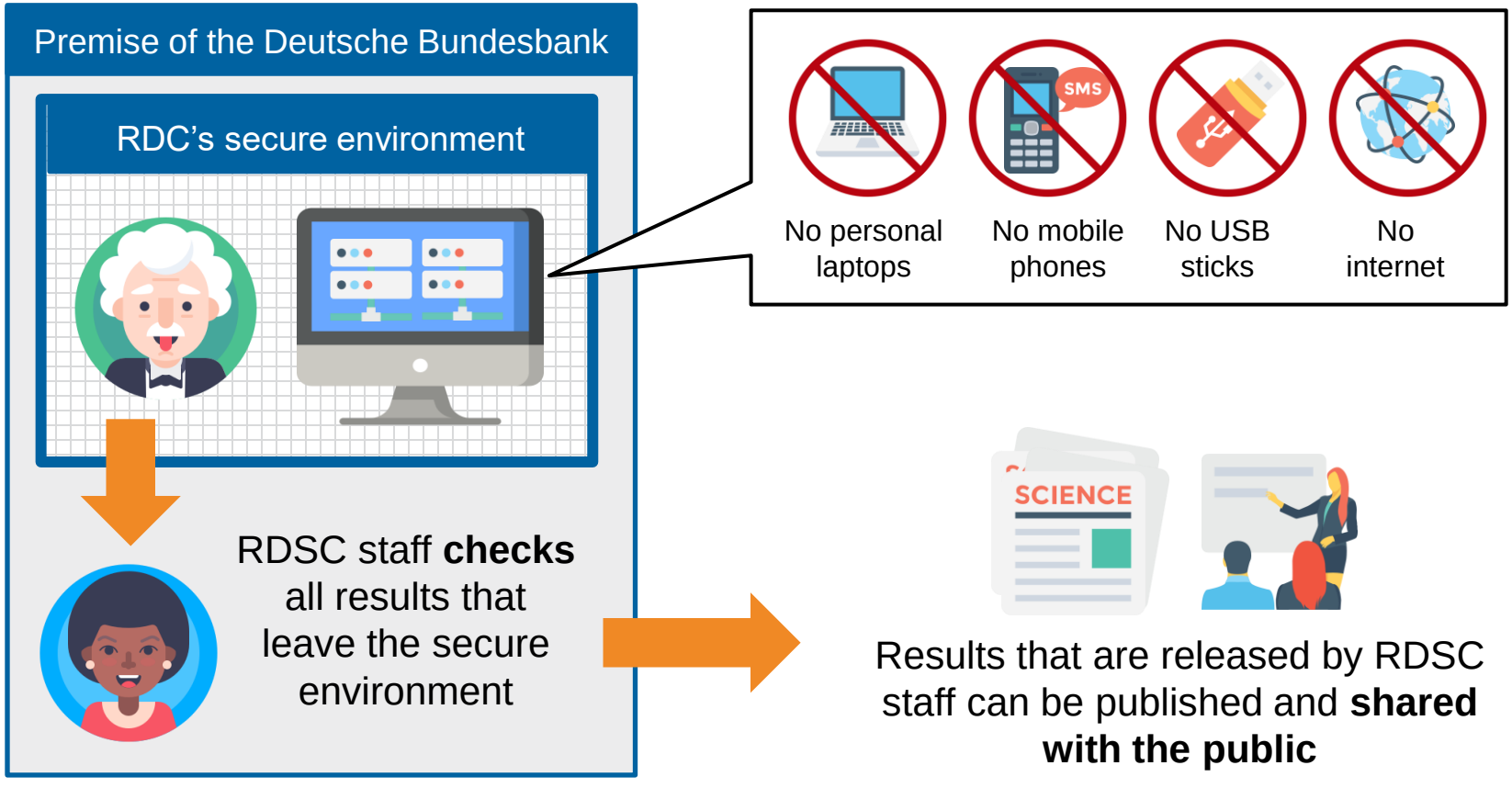
References

- Blaschke, J. & C. Hirsch (2023). On the value of data sharing: Empirical evidence from the Research Data and Service Centre, Technical Report 2023-08 – Version 1.0. Deutsche Bundesbank, Research Data and Service Centre. Retrieved 2023/08/21 from https://www.bundesbank.de/resource/blob/863758/7ebe74476186cd3364a11b3869ada80a/mL/2023-08-value-data.pdf
- Polak, M. P. & D. Morgan (2023). Extracting accurate materials data from research papers with conversational language models and prompt engineering–example of chatgpt. arXiv preprint arXiv:2303.05352



sebastian.seltmann@bundesbank.de
fdsz-data@bundesbank.de
https://www.bundesbank.de/en/bundesbank/research/rdsc

Appendix: RDCs allow to securely share confidential granular data from administrative sources with external researchers



Research Data and Service Centre

Appendix: RDCs proliferate in recent years, both nationally and internationally, enabling high-quality research



Bundesamt für Migration und Flüchtlinge

Office for National Statistics

BANCA D'ITALIA EUROSISTEMA

BANCO DE PORTUGAL EUROSISTEMA

eurostat

BANQUE DE FRANCE EUROSYSTÈME

DESTATIS Statistisches Bundesamt

BANCO DE ESPAÑA Eurosistema

IAB INSTITUT FÜR ARBEITSMARKT- UND BERUFSFORSCHUNG Die Forschungseinrichtung der Bundesagentur für Arbeit

Centraal Bureau voor de Statistiek

** selected examples*

Research Data and Service Centre

Appendix: The Prompts

Identification Prompt

```
A dataset is a collection of structured or unstructured data that is organized and grouped
together for a specific purpose. It typically consists of multiple data points or observations
related to a particular topic or subject. A dataset can include various types of information
such as numerical values, text, images, audio, video, or any other form of data. It is often
used in the context of data analysis, machine learning, and statistical research, where the data
is utilized to extract insights, train models, or draw conclusions. Datasets can be generated
through various means, including surveys, experiments, observations, or by gathering existing
data from different sources.
The text after the empty lines is a scientific paper excerpt
According to the definition on the first line, search for any mentions of datasets or data-
sources used in the paper's research.
If you find any, please compile a list of all mentioned datasets in this excerpt.
If there aren't any, please reply with 'None'.
```

Appendix: The Prompts

Consolidation Prompt

```
You will be provided one or more
lists of datasets.
Each new list starts with '=>'.
You need to combine all the lists
into a single one by removing
redundant entries. Delimit the final
list with simple '-' bullet points.
```

Recall Prompt

```
You will be provided with text delimited by triple
quotes that is supposed to be a list of datasets.
Check if the following true datasets are directly
contained in the answer:

...

For each of these true datasets perform the following
steps:
1 - Restate the true dataset
2 - Write 'yes' if the true dataset is mentioned in
the answer, otherwise write 'no'
Finally, provide a count of how many 'yes' findings
there are. Provide this count as {"count":<insert
count here>}.
```

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Let's talk about sentiment: natural language processing using machine learning on bank earnings transcripts[1]

Seung Jung Lee, Sriram Nagaraj, Dylan Saez,
Victors Stebunovs, Cindy Vojtech and Karl Wirth,
Board of Governors of the Federal Reserve System

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Let's Talk About Sentiment: Natural Language Processing Using Machine Learning on Bank Earnings Transcripts

Seung Jung Lee, Sriram Nagaraj, **Dylan Saez**,
Victors Stebunovs, Cindy Vojtech, and Karl Wirth

3rd IFC and Bank of Italy Workshop on
"Data Science in Central Banking: Enhancing the access to and sharing of data"

19 October 2023

The views expressed herein are those of the authors' and do not necessarily reflect the opinions of the Board of Governors of the Federal Reserve System.

Overview

1. Motivation
   - Bank Earnings Calls
2. Literature
3. Introducing Roget
   - Topic Model: Word2Vec, Clustering, Tree
   - Sentiment Model: FinBERT
4. Data and Summary Stats
5. Econometric Specification
6. Results
7. Concluding Remarks
8. Change in Language Over Time

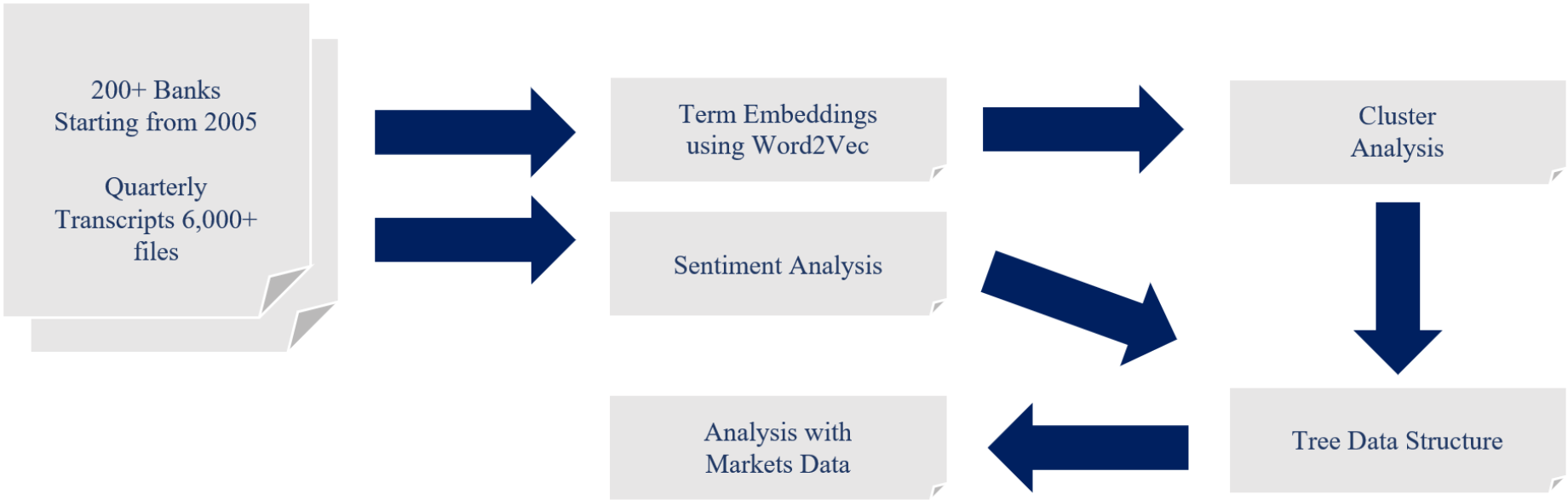
Motivation

- Large publicly-listed banks have conference calls when they announce quarterly earnings.
  - Which topics are of focus or take up the most time?
  - What is the sentiment associated with various topics?
  - Does this matter for market prices, specifically equity and CDS prices?

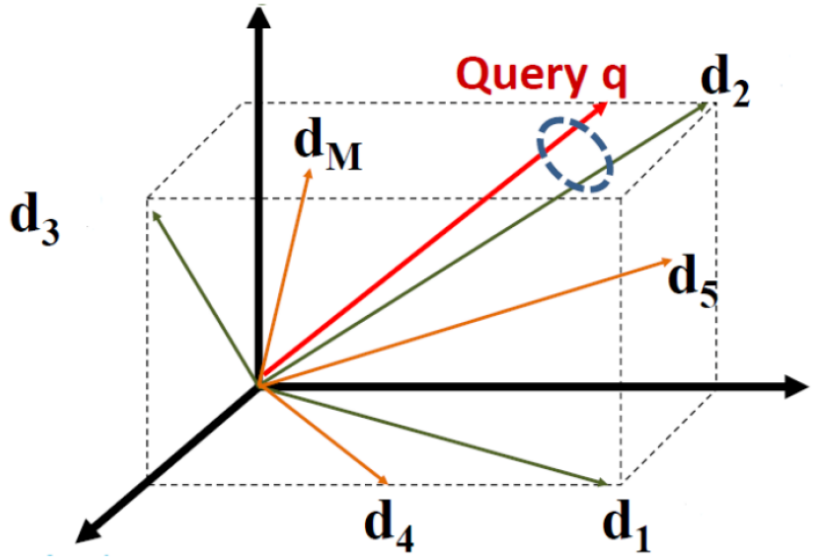
Literature

- Discussions in earnings conference calls or corporate disclosures are useful in understanding market movements, earnings surprises, or future earnings.
  - Sentiment: Tetlock, Saar-Tsechansky, and Macskassy (2008), Loughran and McDonald (2011), Mayew and Venkatachalam (2012), Price, Doran, Peterson, and Bliss (2012), Jiang, Lee, Martin, and Zhuo (2019), De Aicus, Falconieri, and Tastan (2020)
- **Mostly based on entire body of text/words and dictionary approach**.
- Machine learning approaches.
  - Sentiment: Yang, Siy, and Huang (2020), Lu (2021)
  - Topic: Huang, Lehavy, Zang, and Zheng (2018)
- **We combine tree-structure topic modeling with BERT-based sentiment**.

Topic Model: Roget Model Overview



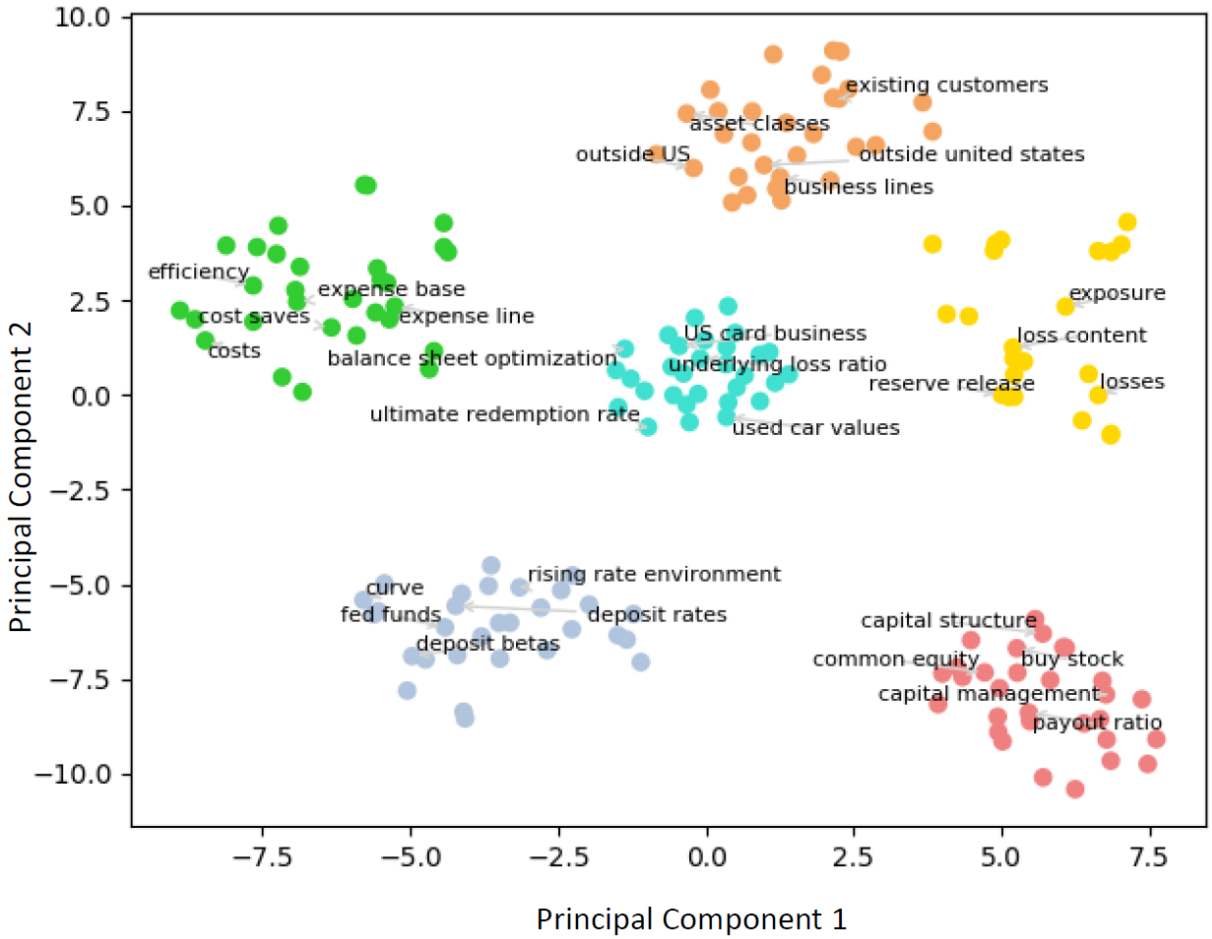
Topic Model: Word2Vec



Assume you have M expressions (words, paragraphs, slides, documents): Each can be represented by a vector and compared to the other via the vector distance:

If a new expression, or query is presented, it can be compared with the existing expressions and the closest one(s) can be found. **Clusters can be found within the existing vectors.**

Topic Model: Clusters

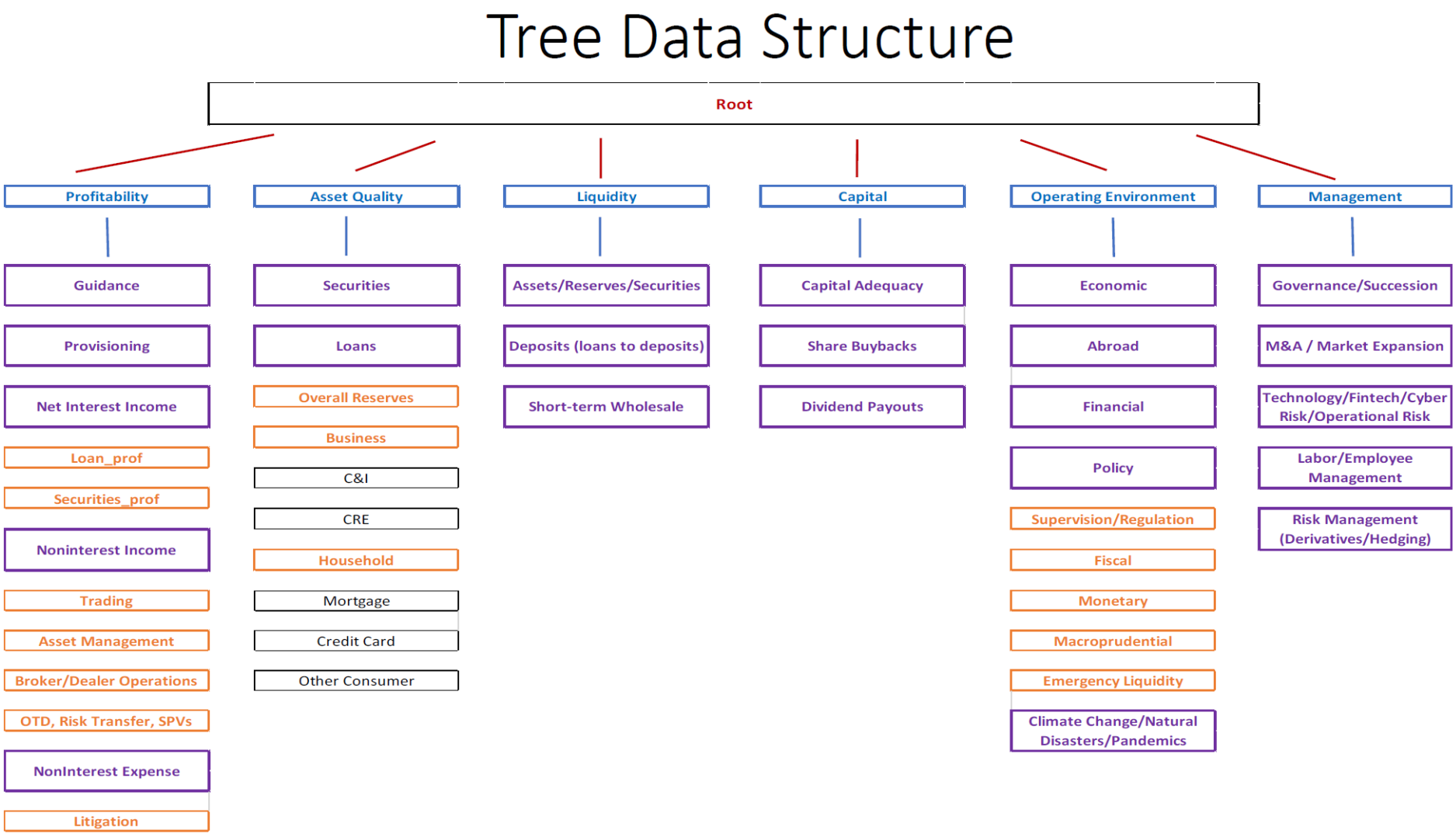


Resulting in interpretable high-level clusters and measurable distance between term representations.

Topic Model: Tree Data Structure

- Topics arranged in a tree structure starting with the "root"
  - Each node contains its name, Word2Vec representation, topic weights, and average topic sentiment
- Parent nodes have multiple children with a sub-topic relationship
- Topic weight/sentiment averaging is bottom-up: starting with the leaf nodes, the topic sentiment and weights ascend the tree to the root
- The model is therefore a semi-supervised (or weakly supervised) approach

Topic Model: Tree Data Structure



Topic Model: Tree Data Structure

Dominant Topics *Down* the Tree – SVB Financial Group (SIVB Q4 2022 Earnings Conference Call Transcript January 19, 2023, 6:00 PM ET

	Topic 1	**Topic 2**
Yes, Casey. I'll just go to the results of the fourth quarter as an indicator of why that statement makes sense. We're looking at venture deployment in the quarter, 35 billion or so think of that is kind of an annualized run with 120 billion to 140 billion venture deployment in the quarter, from a balance sheet perspective on balance sheet. While we did see the decline in deposits, it was much lower than what we saw in the third quarter. And the reason for that gets to what Greg mentioned, as well as Mike, where we're seeing that lower level of cash burn. So even on a much slower venture deployment number, call it in the mid $30 billion range, we started to see that on balance sheet deposit when we look at cash burn versus inflows get to a much more normalized level.	Liquidity	Deposits
So that I think is an indicator with cash burn continuing to slow based on what Mike and Greg just said, that we can get back without going to the 2021 deployment level, to not just deposit of being at the same level, but the potential for deposit growth.		

Sentiment Analysis using FinBERT

- Yang et al. 2020
- Same architecture as BERT, but not BERT based
- Pre-trained on 2.5 Billion tokens (more than BERT):
  - Corporate reports (25 years)
  - Earnings call transcripts (15 years)
  - Analyst reports (13 years)
- Fine-tuned on Analyst Tone - labeled sentences from financial analyst reports
- Results on several financial sentiment datasets

Negative Sentiment Example – JPMorgan Chase (JPM) Q3 2020 Earnings Call Transcript October 13, 2020, 8:30AM ET

Question: I may have missed this, I had to jump off for a minute on the call, but can you give us some color? As you've spoken very well about what's going on in the consumer credit area, but when you go into the commercial side of the business, can you share what the sectors that you're seeing the biggest challenges? And can you give us some color on the rerating process that you're going through on those credits that are in trouble today and what kind of deterioration you're seeing in those specific credits in terms of possibly write-downs or revaluations or if it's collateral, like in a commercial real estate loan?

Answer: Sure. So, the sectors, I think, are ones that you would expect airlines, lodging, restaurants, other T&E, real estate, oil and gas. And those continue to be the sectors under the most pressure. When you look at downgrades here in the third quarter -- or not here in the third quarter -- in the third quarter, we saw downgrades slow a bit because in the second quarter, we saw significant downgrades just on the increased level of debt that companies were taking on. So, we saw downgrades slow a bit in the third quarter, but we do expect downgrades to continue, particularly in real estate. And then elsewhere, in wholesale, I would say CEO sentiment is guarded, but constructive.

Market Data Sources

- **SNL Database**
  - EPS consensus estimates and actual values
- **FRB-FAME Database**
  - CDS and equity prices
- **FRB-Fitch Database**
  - Control variables–ex: total assets, net income, gross loans, etc.

Data: Summary Statistics

Variable Name	Obs. (GSIBs)	Average	Std. Dev.	Minimum	Maximum
Root	2,537	0.145	0.147	-0.530	0.756
Profitability	2,537	0.142	0.180	-0.947	0.974
Asset Quality	2,537	0.073	0.271	-0.999	1
Liquidity	2,537	0.110	0.266	-0.994	1
Capital	2,537	0.109	0.216	-0.922	1
Operating Environment	2,537	0.146	0.189	-0.457	1
Management	2,537	0.222	0.244	-0.994	0.999
Abnormal returns: Equity: 60 days	2,175	-0.001	0.038	-0.441	0.153
Abnormal returns: Equity: 120 days	2,175	-0.001	0.037	-0.407	0.161
CDS (4-Quarter Difference)	2,147	-0.001	0.046	-0.402	0.488

Calculating Equity Abnormal Returns

$$r_{i,t} - rf_t = \alpha + \beta_1 Mkt_t + \beta_2 SMB_t + \beta_3 HML_t$$

- For t-121 through t-1 to calculate betas
- Where:
  - *Mkt*, *HML*, and *SMB* are Fama-French daily factors
  - *rf* is the risk-free rate

$abnormal\ return = r_{i,t} - \hat{r}_{i,t}$ on earnings release date t

Calculating CDS Movement

$$CDS_{i,t} = \alpha + \beta_1 CDS_{s,t}$$

- Where CDSs,t is the change in sovereign CDS for the firm i's country

Sentiment Regressions

$$y_{i,t} = \Theta_i + \lambda_t + \beta_1 S_{c,i,t} + \epsilon_{i,t}$$

for entity i on earnings release quarter t to calculate betas

- Where:
  - $y_{i,t}$ is the abnormal returns or CDS spread for entity i on earnings release date t
  - θ_i is the fixed effect for entity i
  - λ_t is the time fixed effect at quarter date t
  - $S_{c,i,t}$ is the sentiment score for a given category c, for entity i on earnings release quarter t
    - Seven categories: root, profitability, liquidity, asset quality, capital, management, operating environment
    - Run each category of sentiment separately, then all at once
    - Can add a dummy that indicates whether earnings missed analysts' forecasts

GSIB Abnormal Equity Returns and Sentiment

Sentiment | -999 as 0 | GSIB ONLY

	DV: Abnormal Return for Equity Prices - 60 Day									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Root	0.045***									
	(0.007)									
Profitability		0.029***								0.020***
		(0.006)								(0.007)
Liquidity			0.008**							0.003
			(0.004)							(0.004)
Capital				0.006						0.002
				(0.004)						(0.004)
Asset Quality					0.014***					0.009**
					(0.004)					(0.004)
Management						0.011***				0.002
						(0.004)				(0.005)
Operating Environment							0.016***			0.002
							(0.006)			(0.006)
ROA	-0.001	-0.001	-0.0004	-0.0004	-0.0005	-0.0004	-0.0005	0.00004		-0.001
	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)	(0.0004)		(0.0004)
MISSED EST	-0.016***	-0.016***	-0.018***	-0.018***	-0.018***	-0.017***	-0.018***		-0.017***	-0.016***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)		(0.002)	(0.002)
Observations	2,175	2,175	2,175	2,175	2,175	2,175	2,175	2,490	2,215	2,175
R^2	0.073	0.063	0.050	0.048	0.057	0.051	0.052	0.00001	0.042	0.068
Adjusted R^2	-0.482	-0.498	-0.520	-0.524	-0.509	-0.517	-0.516	-0.517	-0.522	-0.496
F Statistic	35.888*** (df = 3; 1359)	30.677*** (df = 3; 1359)	23.816*** (df = 3; 1359)	22.611*** (df = 3; 1359)	27.154*** (df = 3; 1359)	24.544*** (df = 3; 1359)	24.992*** (df = 3; 1359)	0.009 (df = 1; 1641)	61.584*** (df = 1; 1393)	12.441*** (df = 8; 1354)
Significance Levels										*p<0.1; **p<0.05; ***p<0.01

Abnormal Equity Returns and Sentiment

Sentiment | -999 as 0 | GSIB ONLY | Profitability: Children

	DV: Abnormal Return for Equity Prices - 60 Day
General.Profitability.and.Guidance	-0.001
	(0.003)
Provisioning	0.007*
	(0.004)
Net.Interest.Income	0.001
	(0.004)
Noninterest.Income	0.018***
	(0.004)
Noninterest.Expense	-0.0001
	(0.004)
roa_diff_ann	-0.001
	(0.0004)
miss_est	-0.016***
	(0.002)
Observations	2,175
R^2	0.063
Adjusted R^2	-0.503
F Statistic	13.103*** (df = 7; 1355)
Significance Levels	*p<0.1; **p<0.05; ***p<0.01

CDS Spread Movement and Sentiment

Sentiment | -999 as 0 | GSIB ONLY

	DV: Abnormal Return for CDS Prices (% Change)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Root	-0.017**								
	(0.008)								
Profitability		-0.007							-0.006
		(0.006)							(0.007)
Liquidity			0.004						0.007*
			(0.004)						(0.004)
Capital				-0.016***					-0.015***
				(0.005)					(0.005)
Asset Quality					0.003				0.006
					(0.004)				(0.004)
Management						-0.006			-0.003
						(0.004)			(0.005)
Operating Environment							-0.010*		-0.007
							(0.006)		(0.007)
Tier1 Capital Ratio	-0.002*	-0.002*	-0.002*	-0.002*	-0.002*	-0.002*	-0.002*	-0.002*	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	2,147	2,147	2,147	2,147	2,147	2,147	2,147	2,147	2,147
R^2	0.006	0.004	0.003	0.010	0.003	0.004	0.005	0.002	0.015
Adjusted R^2	-0.602	-0.605	-0.606	-0.595	-0.607	-0.605	-0.603	-0.606	-0.592
F Statistic	3.943** (df = 2; 1332)	2.385* (df = 2; 1332)	2.259 (df = 2; 1332)	6.883*** (df = 2; 1332)	1.904 (df = 2; 1332)	2.701* (df = 2; 1332)	3.179** (df = 2; 1332)	3.293* (df = 1; 1333)	2.944*** (df = 7; 1327)
Significance Levels									*p<0.1; **p<0.05; ***p<0.01

Concluding Remarks

- Bank equity prices are sensitive to information conveyed in the profitability part of the bank earnings calls
- CDS premiums are sensitive to what is revealed in the capital discussions
- Application: ROGET

How does the model handle semantic shift and uncommon words?

Examples: COVID-19, Climate, Environment, Weathering (the economic or environment storm?)

- Add to the mapping of words and phrases (semi-supervised)
- For each new block of text, the embedding model is updated to include the current data. The tree is re-measured in the bottom-up fashion
- A weighting function is used in every re-measurement of the tree to mitigate against the distance becoming skewed due to larger/smaller amounts of text in a given block

Thank you!

Contact Information

- Dylan Saez: dylan.c.saez@frb.gov
- Seung Lee: seung.lee@frb.gov

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Constructing high-frequency and thematic economic sentiment indicators from online news articles: applications in the Philippine context[1]

Alan Chester Arcin, Carmelita Esclanda-Lo, Chelsea Anne Ong and Rossvern Reyes,
Bangko Sentral ng Pilipinas

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Constructing High-Frequency and Thematic Economic Sentiment Indicators from Online News Articles

Applications in the Philippine Context

Alan Chester Arcin, Carmelita Esclanda-Lo, Chelsea Anne Ong and Rossvern Reyes[1]

Abstract

The use of natural language processing (NLP) methods in informing macroeconomic policy has risen over the past decade with the increasing availability of alternative data sources. Digital news articles are high-frequency sources of information that economic agents may refer to when making decisions and policies. In this study, we create a News Sentiment Index (NSI), a high-frequency indicator that is based on online business and financial news, to infer market sentiment. Specifically, two NSIs are constructed using the dictionary and machine learning methods, with a sentiment lexicon created from genetic algorithm and Financial Bidirectional Encoder Representations from Transformers (FinBERT), a pre-trained NLP model trained in the finance domain. To examine the relevance of these indicators in capturing key global and Philippine economic events, we evaluate the relationship between the NSIs and other economic indicators through correlation analysis. Furthermore, this study expands to topic modeling using Non-negative Matrix Factorization to explore sentiments by theme. Our findings show that the NSIs track key economic events and are positively and strongly correlated with other high-frequency indicators, such as the Purchasing Managers' Index (PMI) and Philippine Stock Exchange Index (PSEi). Meanwhile, on topic modeling, sentiments from news on company earnings and trade also show strong positive correlations with the PMI, PSEi, and survey-based outlook indicators. We conclude that use of NSIs and topic modeling can potentially complement existing survey-based measures to timely gauge market sentiment.

Keywords: Sentiment Analysis, Opinion Mining, Text Analysis, Dictionary, Machine Learning

JEL Classification: C82 D83

1 This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" held on 17-19 October 2023 in Rome, Italy. Alan Chester Arcin (arcinat@bsp.gov.ph) is from the Department of Economic Research (DER) of the Bangko Sentral ng Pilipinas (BSP), Carmelita Esclanda-Lo (esclandacg@bsp.gov.ph), Chelsea Anne Ong (ongcs@bsp.gov.ph), and Rossvern Reyes (reyesros@bsp.gov.ph) are from the Department of Economic Statistics (DES) of the BSP. The views expressed herein are those of the authors only and do not necessarily reflect those of the BSP. The authors acknowledge Cherrie Mapa and Genna Paola Centeno of the DER-BSP for their contributions to this study. The authors also thank Senior Director Redentor Paolo M. Alegre, Jr. of the DES-BSP and Officer-in-Charge Dennis D. Lapid of the DER-BSP for their unwavering support.

Contents

1. Introduction 3

2. Data 3

3. Methodology 4

Text Pre-processing 4

Approaches in Sentiment Analysis 4

Dictionary-based method 5

Machine learning-based method 5

Sentiment Scoring and Model Selection 6

Topic Modeling 7

Latent Dirichlet Allocation 7

Non-Negative Matrix Factorization 8

Coherence Score 8

4. Results and Discussion 9

Correlation of NSIs with Selected Economic Indicators 9

Topic Modeling Results 12

Correlation of Topic-level NSIs against Economic Indicators 13

8. Conclusion and Recommendations 14

References 16

Appendix A 18

Appendix B 20

Appendix C 22

Appendix D 23

1. Introduction

Sentiment of economic agents can have a considerable impact on real economic activity. For instance, consumer sentiment can influence consumption and saving decisions, while business sentiment can affect investment, hiring, and pricing decisions. As such, a measurement of market sentiment is an important component of macroeconomic surveillance.

Typically, sentiment or people's attitudes or opinions towards a particular topic are measured through surveys. The Bangko Sentral ng Pilipinas (BSP) deploys the Business Expectations Survey (BES) and Consumer Expectations Survey (CES) every quarter to gauge the economic sentiment of firms and households, respectively. However, conducting surveys can be tedious, often time-consuming, and expensive. The lags in the release of survey results imply that information that could be important for economic surveillance and policymaking may not be available on a near-real-time basis.

More recently, the accessibility of possible alternative high-frequency sources of information, such as online newspaper articles and the growing field of sentiment analysis techniques, has opened opportunities for innovative sentiment measurement. This study explores applying sentiment analysis techniques on news data to determine market sentiment and generate news-based sentiment indices (NSIs). Two indices are developed using: (1) dictionary method and (2) machine learning (ML) method.

This study adds to the increasing list of central banks (Federal Reserve Bank of San Francisco (Shapiro et Al., 2020), Bank Negara Malaysia (Chong et al., 2022), Bank of Korea (Seo, 2022), Reserve Bank of Australia (Nguyen & Cava, 2020), and Narowdy Bank Polski (Marszal, 2022)) that performed sentiment analysis of online news data as an economic surveillance tool.

Overall, this research shows that the NSIs match the expected sentiment of key economic events. These indices display the potential to supplement existing surveys and provide a timely measurement of market sentiment. This study also delves into topic modeling to identify key themes that could be useful in determining the behaviour of certain key economic indicators.

2. Data

The data used in this study consists of news articles from multiple sources on the web. Table 1 shows the news sources used in the dataset, sorted by share to total in descending order.[2] The availability of digital archives allows news articles to be retrieved as early as January 2018. We are particularly interested in the sections of economy, business, finance, and related fields (e.g., stock market, banking).

[2] The authors have coordinated with several media outlets to seek permission to scrape their news data. The following media outlets have agreed to have their news articles retrieved online: BusinessWorld, Manila Bulletin, Inquirer.NET, Manila Standard, and BusinessMirror.

Share of News Sources to Total Number of Articles Gathered

January 2018 – December 2022 Table 1

	Average number of articles per month	Share to the total number of articles gathered
BusinessWorld	453	25.9%
Manila Bulletin	520	23.8%
Manila Standard	332	18.7%
BusinessMirror	294	16.8%
Inquirer.NET	260	14.8%

Source: Authors' Estimates

3. Methodology

Text Pre-processing

Pre-processing of text was done to reduce the noise and unwanted parts of the data and subsequently improve the accuracy of models used in the study. We used the following standard NLP practices in cleaning text:

- Standardization of column headers and date format;
- Exclusion of ads and sponsored articles;
- Deletion of the *lede* (the first paragraph summarizing the article) of some news sources as these may be redundant to the content of the news article;
- Removal of extra spaces and unwanted punctuation marks, hyperlinks, and HTML tags;
- Case normalization;
- Conversion of Unicode characters to ASCII;
- Removal of stop words; and,
- Tokenization of articles.

Note that dictionaries might assign a different polarity for base words and their inflections (e.g., inflate and inflation), and stem words interfere with word matching. Hence, we did not use lemmatization and stemming.

Approaches in Sentiment Analysis

Broadly, sentiment analysis is used to determine the tone of text, whether positive, negative, or neutral. We utilized two general approaches in determining sentiment: (a) dictionary-based method and (b) ML-based method.

Dictionary-based method

The dictionary-based approach involves lexicons and a set of rules. While this word-matching method is straightforward, we acknowledged that it trades interpretability and speed with one of its obvious flaws – ignoring word context and complex word usage. We used the following well-known lexicons in literature listed in Table 2:

List of Sentiment Lexicons

Selected Lexicons with various domains — Table 2

	Description	Author
General Inquirer (GI Lexicon)	Created for general-purpose content analysis of text data	Stone et al. (1966)
Loughran-McDonald Sentiment Dictionary (LM Lexicon)	Constructed using corporate 10-K reports (financial articles) between 1994 and 2008.	Loughran and McDonald (2011)
Financial Stability Sentiment Dictionary (FS Lexicon)	Constructed using financial stability reports of 35 central banks between 2005 and 2015.	Correa et al. (2017)
Hu and Liu Opinion Lexicon (HL Lexicon)	Constructed by determining the sentiment scores of e-commerce reviews using Part-of-Speech tagging and feature identification.	Hu and Liu (2004)
VADER	Constructed to determine the intensity of social media text snippets (tweets) with annotated features.	Hutto and Gilbert (2014)
Genetic Algorithm-based Lexicon (PH Lexicon)	Constructed by using an algorithm that finds the combination of words that maximizes the number of correctly classified polarities from annotated text while minimizing the number of words.	Holland (1975) and Castelli et al. (2017)

Machine learning-based method

The ML-based approach predicts the sentiment of text using a pre-trained model. It is shown to capture more nuances in text but is more complex, requires more computational power, and is less transparent. We used the following ML models:

1. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin, 2018) is a transformer-based machine learning technique developed by Google for NLP tasks. BERT can capture the context and meaning of words based on both their preceding and succeeding words. On various sentiment analysis benchmarks and competitions, BERT has achieved state-of-the-art performance. Applications of BERT include sentiment analysis, social media monitoring, and others.

2. Financial Bidirectional Encoder Representations from Transformers (FinBERT)

FinBERT (Huang, A. H., Wang, H., and Yang, Y., 2022) is short for FinancialBERT, a domain-specific variant of BERT designated to understand and analyze financial data. It is specifically trained and fine-tuned on financial texts such as business articles from Reuters and corporate 10-K reports.

Sentiment Scoring and Model Selection

Between the dictionary-based and ML-based approaches in quantifying sentiment in text, we adopted a more straightforward method of scoring for the dictionary-based method, only using integer representations of positive, negative, and neutral sentiment (1, -1, and 0, respectively). The net diffusion of words determines the news sentiment for an article:

$$Sentiment\ Score_{article,dictionary} = \frac{\sum Positive_{w,article} - \sum Negative_{w,article}}{Total_{w,article}}$$

where *w, article* corresponds to the count of words in an article. Following Hutto & Gilbert (2014) and Asmi & Ishaya (2012), we have implemented a linguistic rule known as negation rule. This is where a word switches polarity (-1 to 1 or vice versa) whenever it is preceded by a negation word, such as 'not' or 'never'.

Given that the ML models return a probability score of sentiment, we assign the results to either positive, negative, or neutral. Since the models compute on a per-sentence basis, the news sentiment of an article is derived by the net diffusion of sentences:

$$Sentiment\ Score_{article,ML} = \frac{\sum Positive_{s,article} - \sum Negative_{s,article}}{Total_{s,article}}$$

where *s, article* corresponds to the count of sentences in an article.

Since the number of articles per news source varies per month, we followed the approaches of Chong et al., (2022) and Baker et al., (2016), where the NSIs are calculated from the average sentiment scores of the articles in each month, adjusted by the number of articles per news source. The indices were also standardized to a historical average of zero. Estimates of the index above zero indicate positive sentiment, and estimates below zero indicate otherwise.

A set of 3000 labeled sentences were used as a test dataset. The selected sentences were also taken from the same news data sources and contained at least one word from any of the five lexicons. The same dataset was used in constructing a Philippine-specific lexicon using a genetic algorithm (PH lexicon) that retrieves a combination of keywords that maximizes the number of correctly classified sentences from the test dataset. The words that comprise the PH lexicon can be found in Appendix A.

To select which lexicon to use for the main NSIs for the dictionary-based and machine learning-based approaches (NSI-D and NSI-ML, respectively), we assessed the classification performance of the five off-the-shelf lexicons along with the constructed PH lexicon and the two pre-trained machine learning models vis-à-vis the manually labeled dataset based on accuracy and macro-F1 scores.

Accuracy measures the fraction of the models' correct sentiment predictions. Meanwhile, macro-F1 score combines precision (a measure of false positives) and recall (a measure of false negatives). Macro-F1 is a commonly used metric in NLP literature (Shapiro et. al, 2020) and a function of F1 score with the number of classes taken into consideration:

$$F1\ score = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2TP}{2TP+TP+FN} \text{ and } Macro\ F1\ score = \frac{1}{n}\sum_i^n F1\ score_i$$

where *n* is the number of classes.

While accuracy is a commonly used metric in evaluating the performance of a model, it does not consider the distribution of the data and is adversely affected by false negatives. The macro-F1 score, on the other hand, penalizes models that have too many false positives and false negatives. In addition, macro-F1 score performs better with imbalanced classes. Thus, the macro-F1 score is a more appropriate measure for model evaluation.

The model with the highest accuracy and macro-F1 score is used to generate the NSI-D and NSI-ML. Table 3 shows the accuracy and macro-F1 scores of the various lexicons and ML models. As may be expected, the lexicon created using the genetic algorithm (PH lexicon) outperformed other lexicons, while FinBERT surpasses BERT in terms of accuracy and macro-F1 score. Thus, both the PH lexicon and FinBERT model were chosen to generate the NSI-D and NSI-ML, respectively.

Evaluation Metrics of Lexicons and Machine Learning Models

Table 3

Lexicons	Accuracy	Macro-F1 score
PH Lexicon	0.66	0.64
LM Lexicon	0.52	0.46
FS Lexicon	0.51	0.51
VADER	0.51	0.42
HL Lexicon	0.50	0.46
GI Lexicon	0.45	0.42
ML models		
FinBERT	0.66	0.63
BERT	0.58	0.45

Source: Authors' estimates

Note: These are computed relative to the annotated dataset.

Topic Modeling

We now uncover the themes that best characterize the collection of news articles. Topic modeling involves using an unsupervised ML technique to cluster text based on similarities. We evaluated two of the most used methods for topic modeling: 1) Latent Dirichlet Allocation (LDA) and 2) Non-Negative Matrix Factorization (NMF).

Latent Dirichlet Allocation

LDA (Blei et al., 2003) is a generative model that represents documents as mixtures of topics, and each topic is made up of word probabilities. There are two well-known implementations of LDA: (a) Variational inference (VI) and (b) Markov Chain Monte Carlo-Gibb's sampling/Collapsed Gibb's sampling. Gibb's sampling yields precise results, while VI cannot guarantee approximation even after convergence (Blei et al., 2018). However, this comes with a significant tradeoff - Gibb's sampling is slower and

more computationally expensive than VI. Due to its precision, we chose to use the Gibb's sampling implementation of LDA in evaluating different topic modeling algorithms.

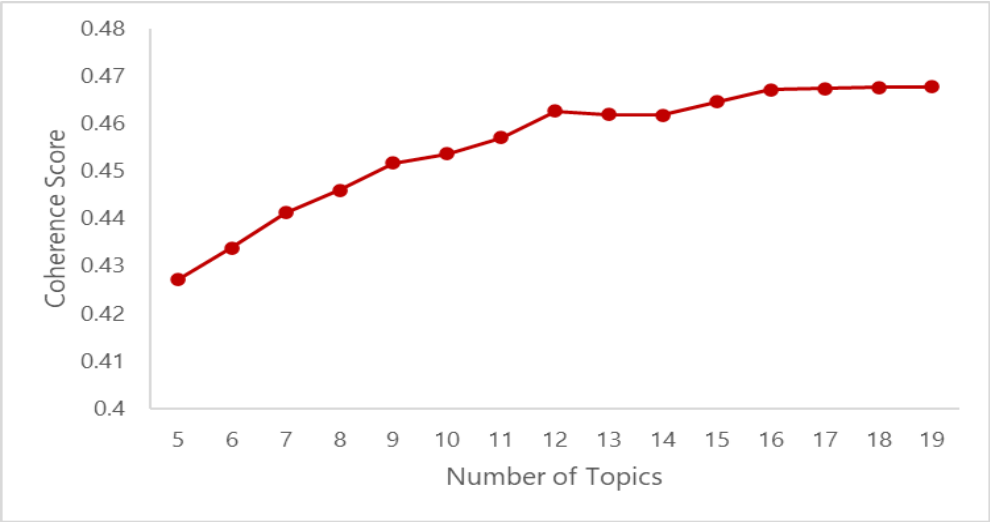
Non-Negative Matrix Factorization

Non-Negative Matrix Factorization (Lee and Seung, 2000) is a statistical method that decomposes a Term-Document matrix into Word-Topic and Topic-Document matrices. Given that NMF is an application of linear algebra, it is a swift and efficient topic modeling algorithm. The NMF model provided similar results to LDA while being less complex and taking less time to execute. Hence, we chose NMF over LDA to retrieve the topics from news articles.

Coherence Score

Coherence score is the degree of similarity between the top words of a certain topic. The graph of coherence score against the number of topics suggests a marginal increase for every additional topic. We computed the coherence score of topics ranging from 5 to 19, averaging the scores of 100 NMF models with different random states per number of topics. The coherence score is computed as the average of all the cosine similarities between each topic word and its cluster through the Python Gensim implementation derived from Roder et al. (2015). Figure 1 shows the coherence score of the overall coherence scores for NMF. Similar to a scree plot, we follow a heuristic where the first inflection point or the "elbow" in the curve is the optimal choice. In this case, we found that 12 is the optimal number of topics. Consequently, we still checked whether the resulting topics were consistent by looking at the keywords per topic.

Figure 1: Coherence scores against Number of Topics



Source: Authors' estimates

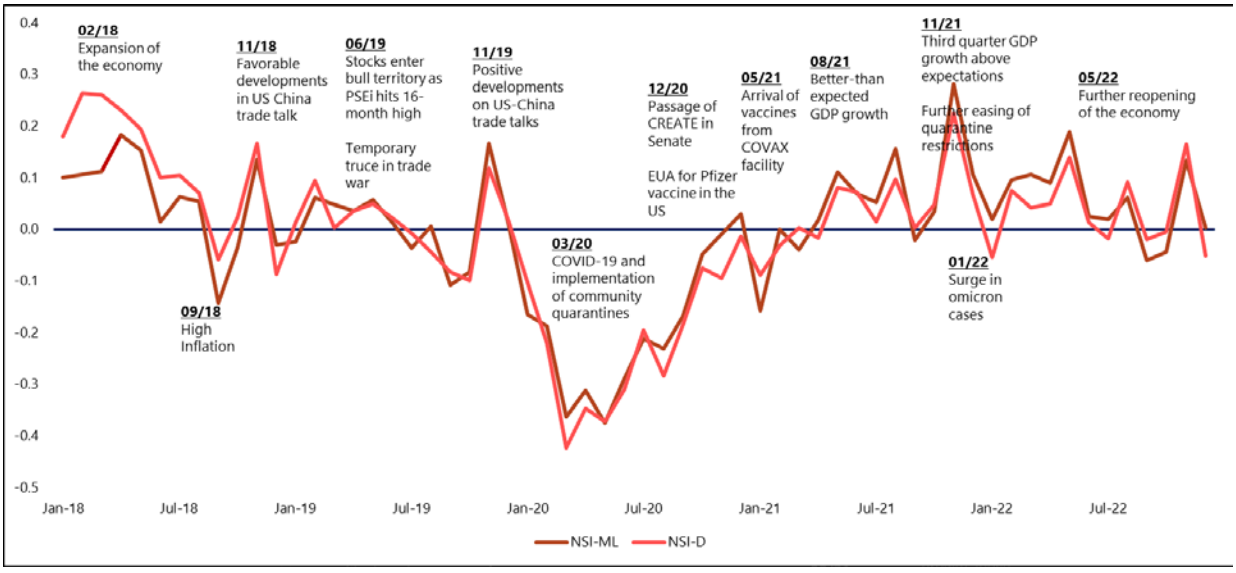
4. Results and Discussion

This section presents an analysis of the NSI-D and NSI-ML estimates. Figure 2 shows that the indices move together, suggesting broadly similar estimates even if these NSIs are derived using two different methodologies. More importantly, Figure 2 suggests that the NSI estimates match the expected market sentiment of key events that may have significant economic repercussions.

For example, sentiment dipped and turned significantly negative following the onset of the COVID-19 pandemic and the implementation of community quarantines. It marginally recovered in December 2020 (the first time the index registered a positive reading since the pandemic), coinciding with the passage of the Corporate Recovery and Tax Incentives for Enterprises (CREATE) bill in the Senate and the issuance of the emergency use authorization of the Pfizer vaccine in the US. Sentiment estimates were also above zero in May 2021 when the vaccines from the World Health Organization-led COVAX facility arrived in the Philippines. Sentiment estimates also settled in the positive region when favorable economic outturns, such as news of actual gross domestic product (GDP) exceeding expectations, were reported. Other key events that match the peaks (positive sentiment) and troughs (negative sentiment) of the NSIs have been identified in Figure 2 for the readers' reference. Overall, the results show that both NSI construction methods (NSI-D and NSI-ML) reflect economic upturns and downturns.

Figure 2: NSIs and Key Economic Events

January 2018 – December 2022



Source: Authors' estimates

Correlation of NSIs with Selected Economic Indicators

Table 4 presents the correlation of NSI-D and NSI-ML against selected economic indicators such as Philippine Stock Exchange Index (PSEi), Purchasing Managers'

Index (PMI) and the confidence indices (CI) from the BSP BES and CES. As expected, the NSI-D and NSI-ML are positively correlated with the PSEi, PMI, BES CI, and CES CI.

Correlation Scores of NSIs against Selected Economic Indicators

January 2018 – December 2022 Table 4

	PSEi	PMI	BES	CES
NSI-D	0.61*	0.71*	0.49**	0.38
	(0.00)	(0.00)	(0.03)	(0.10)
NSI-ML	0.53*	0.67*	0.43***	0.28
	(0.00)	(0.00)	(0.07)	(0.23)

Source: Authors' Estimates

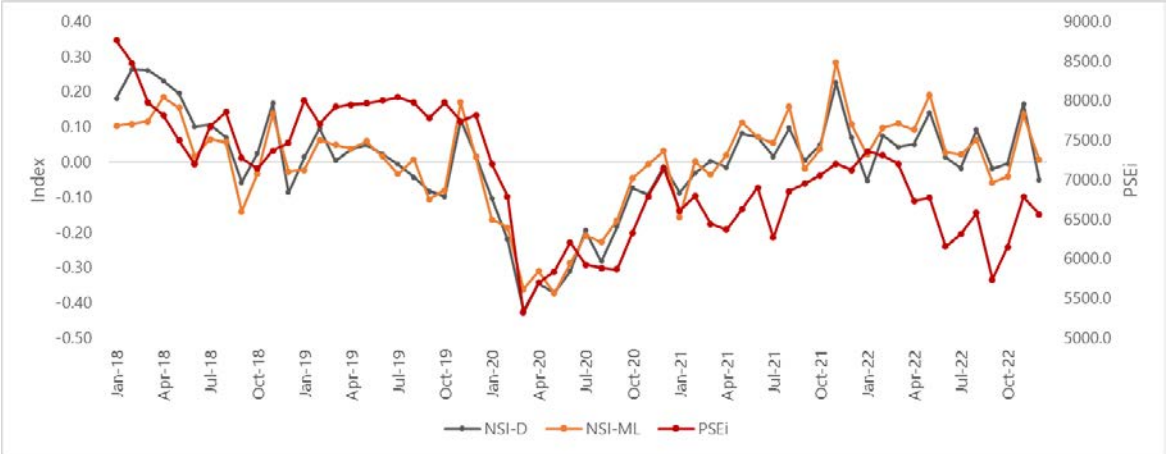
Note: Values in parenthesis are corresponding p-values. Asterisks *, **, and *** denote significance at 1%, 5% and 10% levels, respectively.

The monthly averages of NSI and selected economic indicators, such as the monthly average of daily PSEi and the monthly PMI, are shown in Figures 3 and 4, respectively. Both NSIs correlate positively with the PSEi at 0.61 and 0.53 for the NSI-D and NSI-ML, respectively.

The NSIs are also positively correlated with the PMI, which tracks economic or market conditions as seen by purchasing managers. Similar to the PSEi, NSI-D is more correlated with PMI at 0.71 compared to 0.67 of NSI-ML.

Figure 3: Correlation of NSIs with Average Monthly PSEi

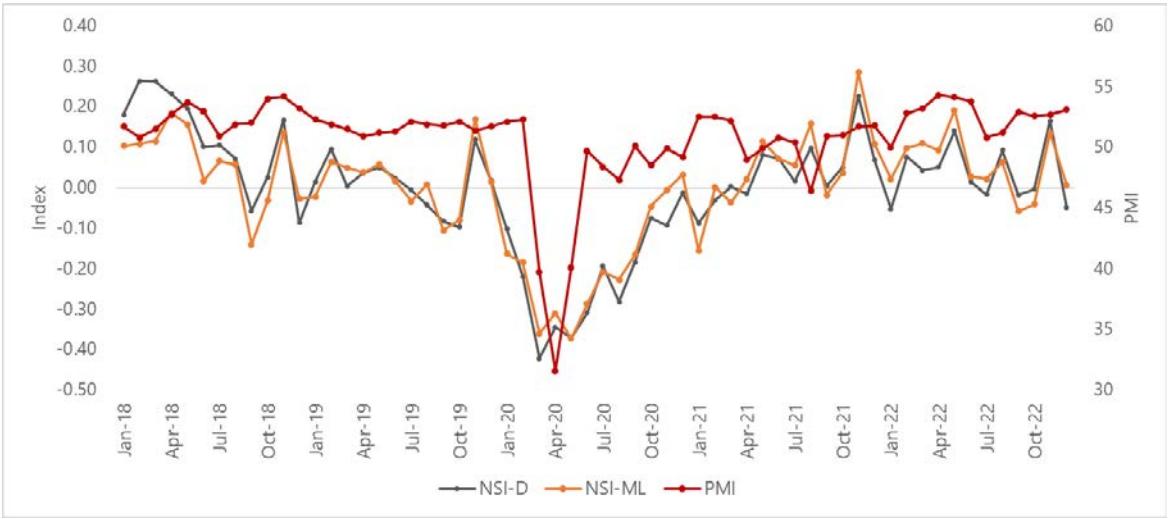
January 2018 – December 2022



Source: Authors' estimates

Figure 4: Correlation of NSIs with Monthly PMI

January 2018 – December 2022

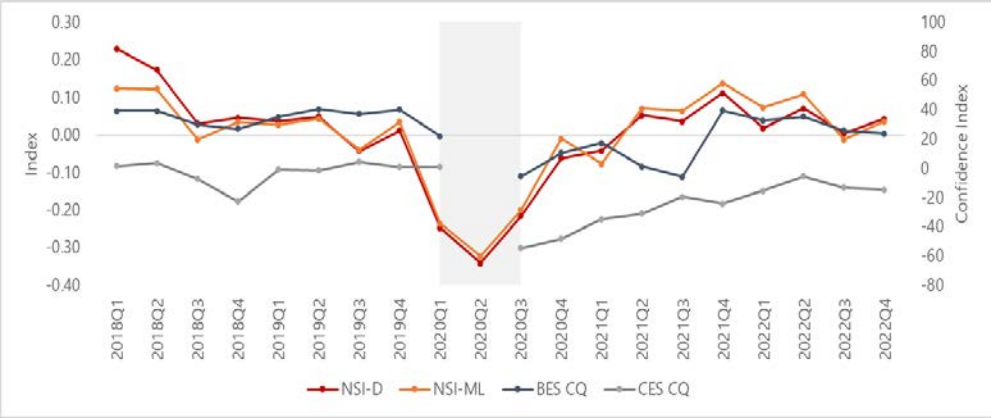


Source: Authors' estimates

To compare the NSIs with the quarterly economic indicators, such as the BES and CES current quarter confidence indices, we compute the quarterly average NSIs from Q1 2018 to Q4 2022, as shown in Figure 5. The quarterly NSIs show a positive correlation with business and consumer confidence indices. For the BES, NSI-D is higher at 0.49 compared to 0.43 with NSI-ML, while for the CES, the correlation with NSI-D is weaker at 0.38 and 0.29 with NSI-ML. Given that the NSIs are available at a higher frequency, this can fill the gaps when surveys are unavailable, like in Q2 of 2020 when survey activities were halted because of the COVID-19 pandemic.

Figure 5: Correlation of NSIs with Confidence Indices

2018Q1 -2022Q4



Source: Authors' estimates

Topic Modeling Results

Based on the top words per topic generated through NMF, the following themes were identified: energy, agriculture, services, corporate news, stock market, micro, small and medium enterprises (MSMEs), banking, trade, government finance, government projects, company earnings, and inflation.

Figure 6 shows some of the word clouds associated with selected topics. Word clouds for all topics can be found in Appendix B. The word clouds and their respective notable keywords are as follows (clockwise starting from upper left): The first word cloud refers to Banking with the words "banking", "lending", "financial institution" alongside words pertaining to the BSP (e.g., "Bangko Sentral", "central bank", and "BSP"). For the next word cloud, emphasis is given to the words "inflation", "Federal Reserve" and "central bank" pertaining to monetary policy. Energy can be assigned as the topic in the next word cloud with the words "electricity", "power", "capacity", and "meralco" (Manila Electric Company) as keywords. Aside from the words "revenue", "sale" and "net income", frequency terms such as "month", "quarter" and specific months of year are usually observed with reports of publicly listed firms. Hence, we label the topic as Company Earnings.

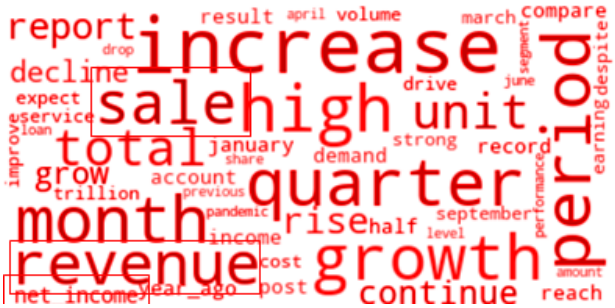
Figure 6: Word clouds of selected topics



Banking



Monetary Policy



Company Earnings



Energy

Correlation of Topic-level NSIs against Economic Indicators

We analyzed how the NSI per topic or theme moves with economic indicators. Due to its relative ease of implementation and higher correlation scores with economic indicators compared to NSI-ML (see Table 2), NSI-D was used to generate separate NSIs for each of the identified topic. Correlation scores were also computed for NSI-D per topic with PSEi, PMI, BES, and CES CI. Table 5 shows the topics with which sentiment scores posted the highest correlation with the monthly PSEi and PMI.

Correlation Scores of Topic-level NSIs against Selected Economic Indicators

January 2018 – December 2022 Table 5

	PSEi	PMI
Company Earnings	0.70	0.54
Services	0.39	0.48
Trade	0.39	0.62
Banking	0.45	0.50
Infrastructure	0.41	0.34

Source: Authors' estimates

Note: Sentiment indices per topic were estimated using NSI-D. The calculated p-values are close to 0 ($p < 0.01$), hence significant at 1% level.

Company earnings (news articles relating to financial disclosures of firms) ranked highest among the topics with correlation against PSEi, followed by services and trade. Thus, it can be inferred that the NSI from news on company earnings can be used to gauge PSEi's movements and serve as a good indicator of market confidence. Meanwhile, market sentiment based on trade-related news (i.e., articles relating to imports and exports) appeared to be the most correlated with PMI, followed by company earnings and banking. The trends of the topic-level NSIs with the highest correlation against monthly PSEi and PMI are shown in Appendix C.

Table 6 shows the topics whose NSIs had the highest correlation scores with BES and CES CI. The NSI of news relating to company earnings ranked highest, followed by infrastructure and trade. This finding implies that market sentiment related to company earnings can also be an alternative measure of consumer and business outlook. Interestingly, these topics also displayed a high positive correlation with the GDP growth rate. The trends of the topic-level NSIs with the highest correlation against quarterly BES, CES, and GDP growth are shown in Appendix D.

Top Topics with NSIs that Exhibited Highest Correlation Scores with quarterly BES CI, CES CI and GDP growth rate

January 2018 – December 2022 Table 6

	BES CI	CES CI	GDP
Company Earnings	0.77	0.79	0.81
Infrastructure	0.60	0.62	0.70
Trade	0.47	0.36	0.78

Source: Authors' estimates

Note: Sentiment indices per topic were estimated using NSI-D. The calculated p-values are close to 0 ($p < 0.01$), hence significant at 1% level

It can also be noted that the above topic-level NSIs display a higher correlation with selected economic indicators compared to the general NSIs. This suggests that topic modeling may bring out specific facets of news that are more significant in explaining co-movement with selected economic indicators, particularly the BES, CES, and GDP growth. Therefore, these topic NSIs show potential as complements to these key economic measures.

8. Conclusion and Recommendations

News-based indices offer a faster and more cost-effective approach to high-frequency sentiment measurement by leveraging widely available news articles and innovations in sentiment analysis techniques.

Evaluation of the indices shows that the constructed NSIs match the expected sentiment for critical economic events. This suggests that NSIs have the potential to complement the BSP's existing BES and CES and provide a timely measurement of market sentiment.

Meanwhile, topic modeling helped identify the key themes in the corpus of news data used in this study. Notably, some topic-level NSIs demonstrated higher correlations with key economic indicators than the broad NSIs. This implies that topic modeling can help uncover specific news themes that may have more predictive content for key economic variables, particularly the BES, CES, and GDP growth rate. While correlation does not imply causality, the high correlation scores between the identified topics and key economic indicators are encouraging and suggest that this strand of research is worth pursuing.

If data limitations are addressed, further studies may generate a more extended series of NSIs to determine how it track economic indicators in the long run. To make our models more robust, we plan to fine-tune our existing models and explore new ones to improve the accuracy of classifying sentiments from texts. For example, more linguistic rules can be added (on top of the negation rule), and Part-of-Speech tagging can be considered for NSI-D. We also intend to improve NSI-ML by retraining FinBERT with the manually annotated dataset and complementing our analysis with other ML models (e.g., PH Lexicon + Support Vector Machines (SVM), FinBERT +

Random Forests). Other topic modeling techniques, such as dynamic topic modeling, may also be explored to account for the changes in topics over time.

References

Asmi, A., & Ishaya, T. (2012, October). Negation identification and calculation in sentiment analysis. In The second international conference on advances in information mining and management (pp. 1-7).

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4), 1593-1636.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518), 859-877.

Chong, E., Ho, C. C., Ong, Z. F., & Ong, H.H, (2022). Using news sentiment for economic forecasting: a Malaysian case study. IFC Bulletins chapters, in: Bank for International Settlements (ed.), Machine learning in central banking, volume 57, Bank for International Settlements.

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2021). Sentiment in central banks' financial stability reports. Review of Finance, 25(1), 85-120.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

Huang, Allen and Wang, Hui and Yang, Yi, FinBERT - A Large Language Model for Extracting Information from Financial Text (July 28, 2020). Contemporary Accounting Research, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3910214 or http://dx.doi.org/10.2139/ssrn.3910214

Holland, John (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press. ISBN 978-0262581110.

Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Kanzari, D., Nakhli, M. S., Gaies, B., & Sahut, J. (2023). Predicting macro-financial instability – How relevant is sentiment? Evidence from long short-term memory networks. Research in International Business and Finance, 65, 101912. https://doi.org/10.1016/j.ribaf.2023.101912

Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems (pp. 556-562).

Li, X., Ding, Y., Chen, X., & Lin, X. (2019). FinBERT: A Pretrained Language Model for Financial Communications. arXiv preprint arXiv:1908.10063. Retrieved from https://arxiv.org/abs/1908.10063

Marszal, Anna (2022). What news can really tell us? Evidence from a news-based sentiment index for financial markets analysis, NBP Working Papers 349, Narodowy Bank Polski.

Castelli, M., Manzoni, L., Vanneschi, L., & Popovič, A. (2017). An expert system for extracting knowledge from customers' reviews: The case of Amazon. com, Inc. Expert Systems with Applications, 84, 117-126.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of finance, 66(1), 35-65.

Nguyen, K., & Cava, G. L. (2020). Start Spreading the News: News Sentiment and Economic Activity in Australia. https://www.rba.gov.au

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. https://doi.org/10.1145/2684822.2685324

Seo, Beomseok (2022). Machine-Learning-Based News Sentiment Index (NSI) of Korea (한국의 기계학습 기반 뉴스심리지수) (2022). Bank of Korea WP 2022-15, Available at SSRN: https://ssrn.com/abstract=4230611 or http://dx.doi.org/10.2139/ssrn.4230611

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring News Sentiment. Federal Reserve Bank of San Francisco Working Paper 2017-01. Available at https://doi.org/10.24148/wp2017-01

Stevens, K., Kegelmeyer, P. W., Andrzejewski, D., & Buttler, D. (2012). Exploring Topic Coherence over Many Models and Many Topics. In Empirical Methods in Natural Language Processing (pp. 952–961). Available at: https://www.aclweb.org/anthology/D12-1087.pdf

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Appendix A

PH Lexicon Word List – First 300 Words

Polarity	Words
Positive[1]	able, absorb, absorbs, absorbing, absorbed, abuzz, acceleration, accustom, accustoms, accustoming, acquire, acquiring, acquired, actual, adherence, advance, advances, advancing, advanced, advantaging, advantaged, afford, affords, affording, afforded, agreeable, ahead expectations, ahead forecast, ally, allying, allied, amass, amasses, amassing, amassed, apex, appealable, appear, appears, appearing, assertion, assure, assures, assuring, attains, attaining, attractive, attractiveness, augment, augments, augmenting, augmented, auspicious, beat, beats, beating, beaten, beautiful, benign, better, betters, bettering, bettered, bloom, blooms, blooming, bloomed, bolster, bolsters, bolstering, bolstered, bonus, boost, boosts, boosting, boosted, brisk, build, builds, built, calm, calms, calming, calmed, canny, capability, careful, casual, ceaseless, champion, champions, championing, championed, charitable, cheer, cheers, cheering, cheered, clarification, clarity, classify, classifies, classifying, classified, clean, cleans, cleaned, collaborate, collaborates, collaborating, collaborated, collaborative, concentrate, concentrates, concentrating, concentrated, concession, conclusive, confident, conserve, conserves, conserving, conserved, constitution, contestability, convenience, conventional, convince, convinces, convincing, convinced, cool, cools, cooling, cooled, coordinate, coordinates, coordinating, coordinated, correction, cost reductions, courteous, creatively, culminate, culminates, culminating, culminated, dedicate, dedicates, dedicating, definite, deposit, depositing, deposited, despite, despites, despiting, despited, develop, develops, developed, difference, domain, dominance, driven strength, easier, ebullient, efficiencies, embrace, embraces, embraced, emerges, emerging, emphasize, emphasizes, emphasizing, emphasized, encouraged, encouragement, enjoys, enjoying, enjoyed, enlighten, enlightens, enlightening, enlightened, enter, entering, entered, entertain, entertains, entertaining, entertained, entices, enticing, enticed, equilibrium, essence, establish, establishes, establishing, established, esteem, esteems, esteeming, esteemed, evolve, evolves, evolving, evolved, excellent, excel, excels, excelling, excelled, exclusiveness, expansion, expect, expects, expecting, expectation, expenditure, expression, extent, faithful, fantastic, fastest, fastgrowing, fate, fates, fating, fated, favoring, firm, firming, flare, flares, flaring, flared, flexibility, freehold, freer, frontrunner, fuse, gain, gains, gaining, gained, generous, giddy, gift, gifts, gifting, gifted, give, gives, globalized, goal, grow, growing, grew, growth, guidance, harvests, harvesting, harvested, hearten, heartens, heartening, heartened, heroic, higher, hire, hires, hiring, hitch, hitches, hitching, hitched, illuminate, illuminates, illuminating, illuminated, imaginable, impenetrable, implore, implores, imploring, implored, improves, improved, include, increase, increased, inexhaustible, innovate, innovates, innovating, innovated, inquire, inquires…

[1] Displays only 300 out of 635 words of the lexicon

PH Lexicon Word List – First 300 Words (continued)

Polarity	Words
Negative[2]	abdicate, abdicates, abdicating, abdicated, aberrant, abet, abets, abetting, abetted, abnormal, abnormally, abolish, abolishes, abolished, abort, aborts, aborting, aborted, abrogate, abrogates, abrogating, abrogated, abrogations, abrupt, absences, abusive, abusively, accidental, accomplice, accuse, acquiesce, acquiesces, acquiescing, acquiesced, addict, addicts, addicting, addicted, addiction, adulteration, adversities, aftermaths, agonize, agonizes, agonizing, agonized, ail, ails, ailed, alarm, alarms, alarming, alarmed, alien, aliening, aliened, allegation, allegations, allege, alleges, amortization, angry, anguish, anguishes, anguishing, anguished, annoy, annoys, annoying, annoyed, annulment, annulments, annul, annuls, annulling, annulled, anomalously, anticompetitive, anxious, appall, appals, appalling, appalled, apprehensive, arbitrarily, arbitrariness, arbitrary, argue, argues, arrear, arrears, arrests, arresting, arrogance, arrogant, assault, assaults, assaulting, assaulted, attrition, avaricious, backbreaking, backdate, backdates, backdating, backdated, backward, bailout, bald, balked, bandit, banish, banishes, banishing, banished, bankrupt, bankrupts, bankrupting, bankrupted, barbaric, batter, batters, battering, battered, below, bemoan, bemoans, bemoaning, bemoaned, bitterness, bizarre, blackspot, bland, blink, blinks, blinking, blinked, blood, bloods, blooding, blooded, bloodbath, blot, blots, blotting, blotted, blunder, blunders, blundering, blundered, blur, blurs, blurring, blurred, bog, bore, bores, boring, bored, bottom, bottoms, bottomed, boycott, boycotts, boycotting, boycotted, breach, breaching, breached, breakages, breakdowns, breaks, broke, breakup, brexit, bribe, bribes, bribing, bribed, brood, broods, brooding, brooded, browbeat, browbeats, browbeating, browbeaten, bum, bums, bumming, bummed, bumpy, burdening, bust, busts, busting, busted, cancellation, cancer, carelessness, catastrophes, catastrophic, cause, causes, caused, cautions, cautioning, cavernous, cease, ceases, ceasing, ceased, charge, charged, chasten, chastens, chastening, chastened, chide, chides, chiding, chided, chill, chills, chilling, chilled, choke, chokes, choking, choked, choosy, circumvent, circumvents, circumventing, circumvented, clash, clashes, clashing, clashed, clawback, climbdown, closure, closures, closuring, closured, clueless, clumsy, coercion, collisions, collude, colludes, colluding, colluded, combat, combats, combating, combated, complain, complains, complaining, complaints, complicate, complicates, complicating, complicated, concern, concerns, concerning, concerned, conciliate, conciliates, conciliating, conciliated, condemnations, condemns, condemning, condemned, confess, confesses, confessing, confessed, confession, confine, confining, confined, confinement, confinements, confiscate, confiscates, confiscating, confiscated, confiscation, confrontations, confronts, confronting, confuse, confuses, confused, consternation, constrict, constricts…

[2] Displays only 300 out of 2573 words of the lexicon

Appendix B

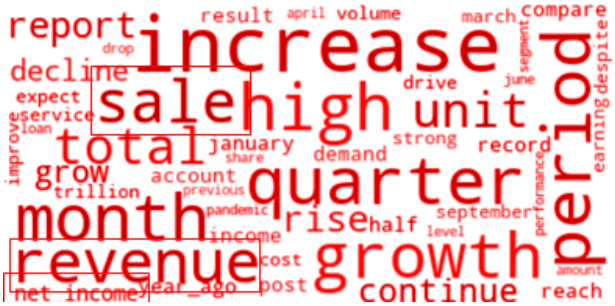
Word Cloud of Topics



Banking



Monetary Policy



Company Earnings



Energy



Agriculture



Services

Word Cloud of Topics (continued)



Company News



Stock Market



MSMEs



Trade



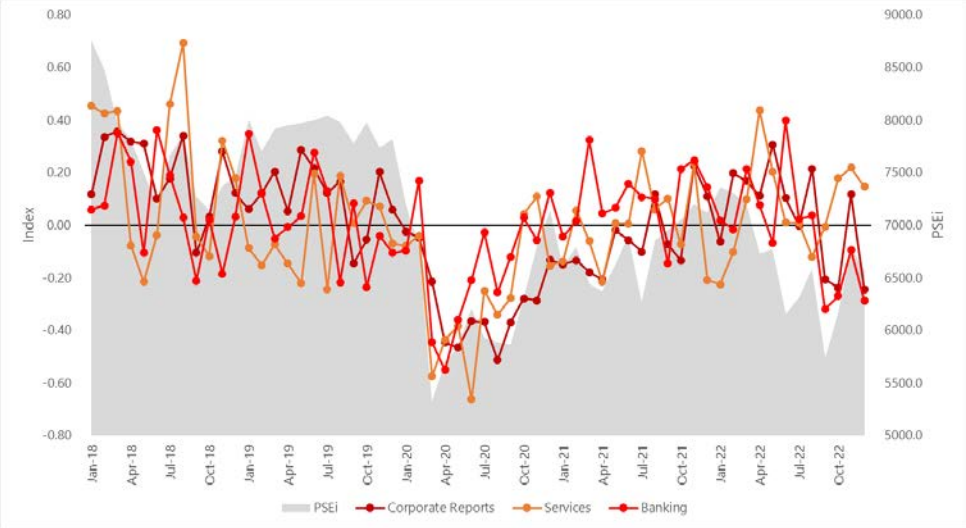
Fiscal Policy



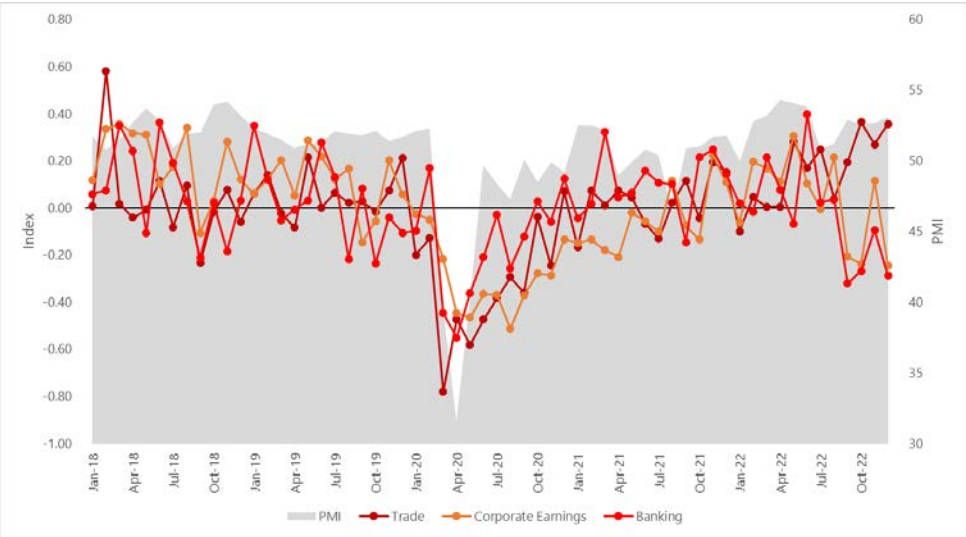
Infrastructure

Appendix C

Topic-level NSIs with Highest Correlation against Monthly PSEi

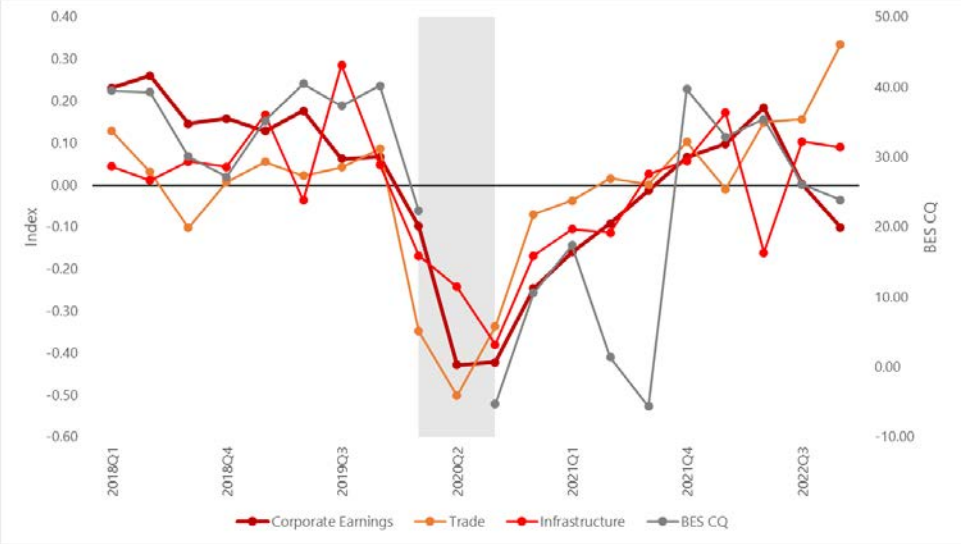


Topic-level NSIs with Highest Correlation against Monthly PMI

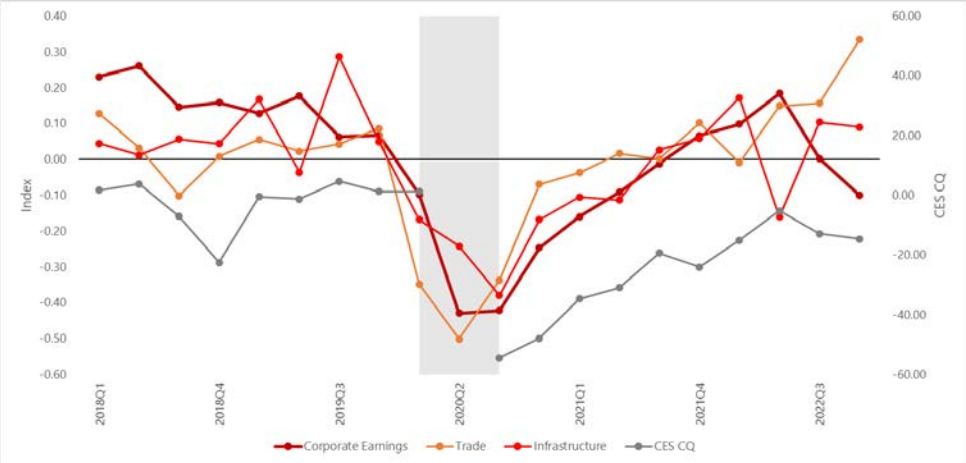


Appendix D

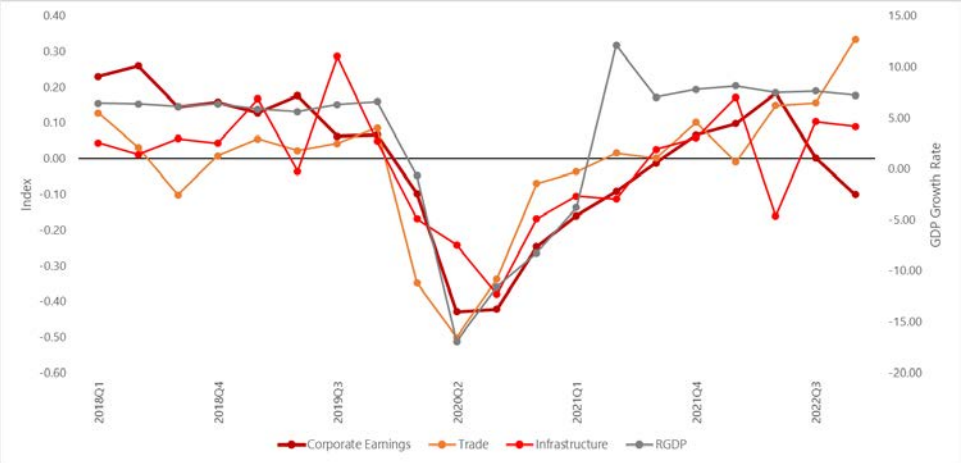
Topic-level NSIs with Highest Correlation against BES Current Quarter



Topic-level NSIs with Highest Correlation against CES Current Quarter



Topic-level NSIs with Highest Correlation against GDP Growth Rate



Constructing High-frequency and Thematic Economic Sentiment Indicators from News Articles

ALAN CHESTER ARCIN, CARMELITA ESCLANDA, CHELSEA ANNE ONG, ROSSVERN REYES

3RD IFC AND BANK OF ITALY WORKSHOP ON "DATA SCIENCE IN CENTRAL BANKING: ENHANCING THE ACCESS TO AND SHARING OF DATA", OCTOBER 17-19, 2023

---

* The views expressed herein are those of the authors' only and do not necessarily reflect those of the Bangko Sentral ng Pilipinas

Outline

1. Motivation
2. Overview, Data Collection and Pre-processing
3. Methodology
4. Model Evaluation
5. Results – Sentiment Indices
6. Topic Modelling
7. Results – Topic modelling
8. Key Takeaways and Future work

Motivation



Sentiment influences economic agents' decisions



Survey-based measures of sentiment are **costly, tedious, and published with a delay**



News-based indices can be **produced at a higher frequency**, allowing for more up to date assessment at a lower cost

Overview

Web scraping



Data cleaning



Sentiment analysis



NSI



Data Collection and Pre-processing

Data were collected from multiple local news sources. The sections of interest were limited to economy, banking, finance and related fields. Text from the articles are pre-processed and further analyzed.

Pre-processing steps include:

- Removal of extra spaces in the text
- Case normalization
- Removal of punctuations and numbers
- Unicode conversion
- Tokenization
- Vectorization

Methodology

Dictionary based method involves a pre-defined list of keywords from some of the most used dictionaries in finance and well-known general dictionaries:

- General Inquirer (GI Lexicon)
- Loughran-McDonald Sentiment Dictionary (LM Lexicon)
- Financial Stability Sentiment Dictionary (FD Lexicon)
- Hu and Liu Opinion Lexicon (HL Lexicon)
- Valence Aware Dictionary for Sentiment Reasoning (VADER)

Using genetic algorithm, we retrieved combinations of keywords that maximizes the number of classified sentences from ~3000 manually labelled sentences (PH Lexicon).

Methodology

Machine learning based method utilizes two pretrained models to predict sentiment of text:

- BERT
  - ML model specifically trained using Wikipedia and BooksCorpus for common language tasks

- FinBERT
  - Built by further training BERT using financial text such as 10-K and 10-Q Corporate Reports, analyst reports and fine tuning it for financial sentiment classification

Model Evaluation

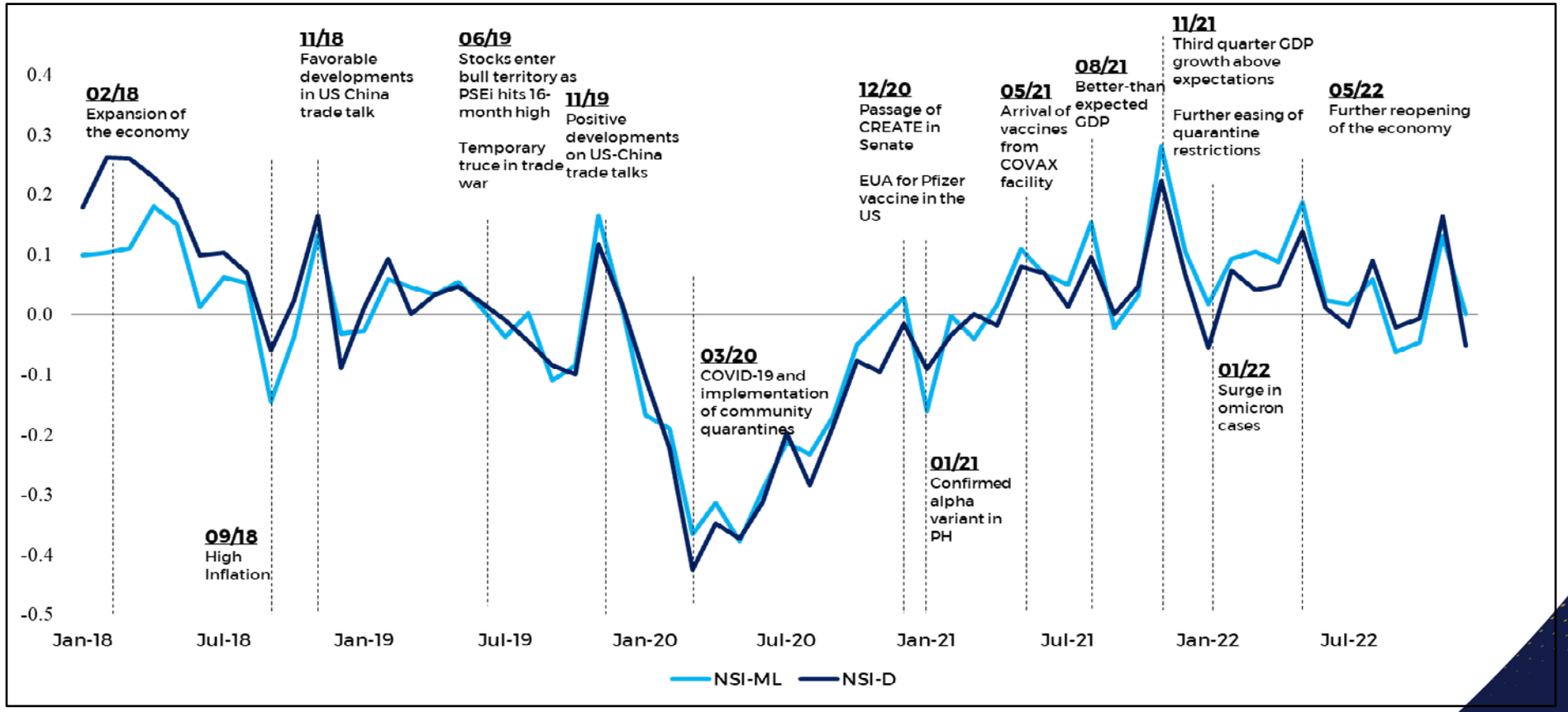
Both dictionaries and machine learning models were evaluated against ~3000 manually labelled sentences sampled from online news articles

Lexicons	Accuracy	Macro F1
PH Lexicon	0.66	0.64
LM Lexicon	0.52	0.46
FD Lexicon	0.51	0.51
VADER	0.51	0.42
HL Lexicon	0.50	0.46
GI Lexicon	0.45	0.42
Machine Learning Models	**Accuracy**	**Macro F1**
FinBERT	0.66	0.63
BERT	0.58	0.45

Accuracy measures the fraction of the model's correct predictions over total predictions. Meanwhile, macro F1 score combines both precision (a measure of the proportion of true positives that were correct) and recall (a measure of the proportion of true positives that were actually predicted).

Results

Estimates of NSIs strongly co-move with key economic events.



Results

Correlation analysis suggest co-movements between NSIs and selected economic indicators

Model	PSEi	PMI	BES	CES
NSI-D	0.61	0.62	0.49	0.38
NSI-ML	0.53	0.58	0.43	0.28

The Business Expectations Survey (BES) is one of the quarterly surveys conducted by the Bangko Sentral ng Pilipinas (BSP) which gathers information about entrepreneurs' views on the general business situation in their own industry and on the national economy. Likewise, the Consumer Expectations Survey (CES) captures the economic outlook of consumers as an indication of the country's future economic condition.

Topic Modeling and Correlation Analysis

Topic Modeling involves unsupervised machine learning techniques to **cluster text based on similarities**.

We considered the two of the commonly used techniques in topic modelling:

- **Non-Negative Matrix Factorization (NMF)**
- Latent Dirichlet Allocation (LDA)

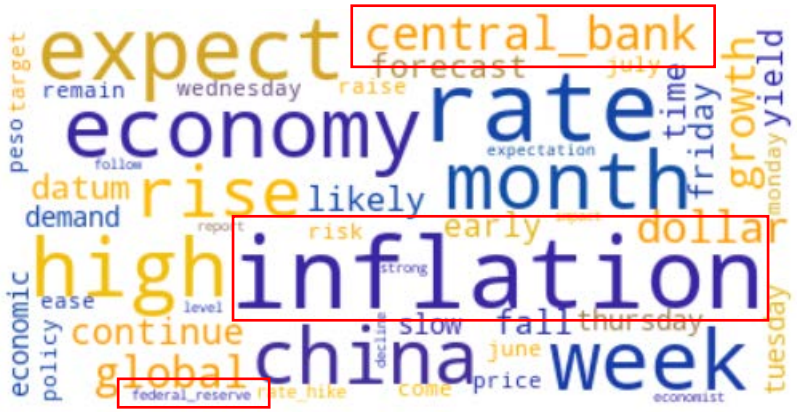
Using the coherence score, we determined **12** to be the optimal number of topics. We then group the articles according to topics, compute the NSI for each and correlate these with selected economic indicators.

Results

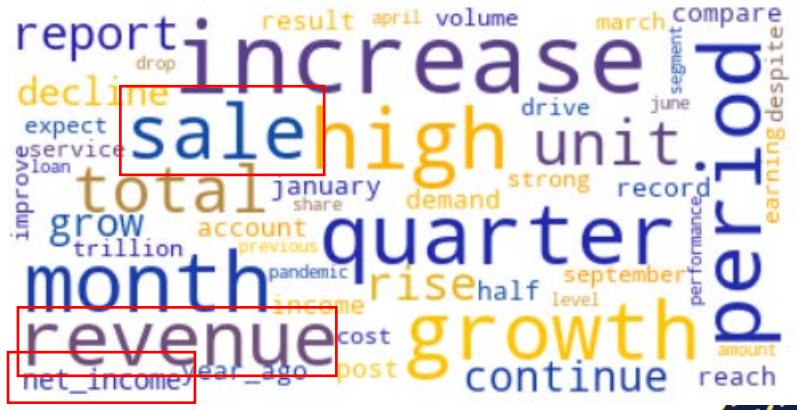
Word clouds of selected topics



Banking



Monetary Policy



Company Earnings



Energy



Agriculture

Results

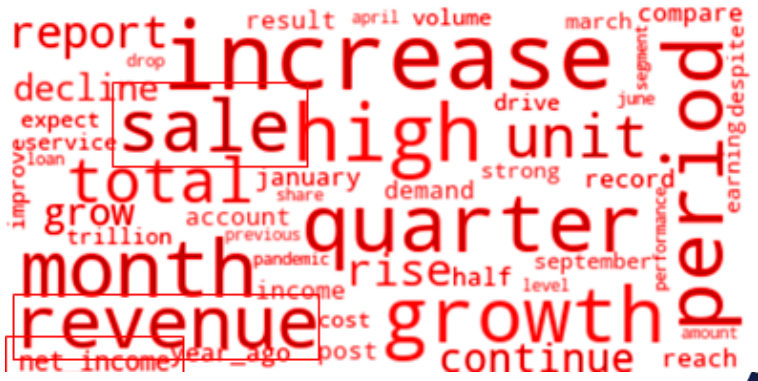
Word clouds of selected topics



Banking



Monetary Policy



Company Earnings

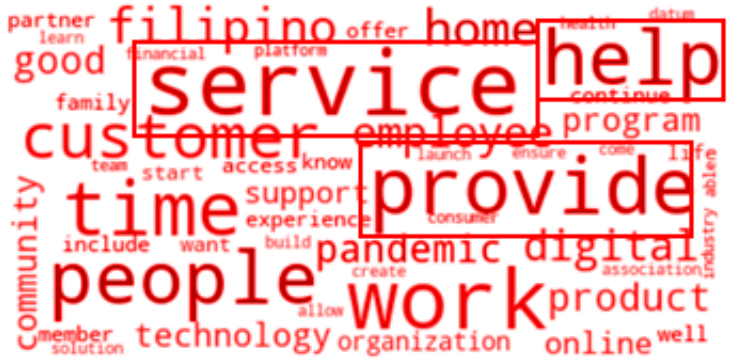


Energy



Agriculture

Results



Services | Corporate News | Stock Market

MSMEs | Philippine Trade | Fiscal Policy | Infrastructure

Results

Correlation analysis show comovements of selected topics against monthly indicators PSEi and PMI...

Topic	PSEi	PMI
Company Earnings	0.72	0.53
Services	0.45	0.47
Trade	0.44	0.63
Banking	0.41	0.49
Government Projects	0.44	0.35

As well as against quarterly indicators BES, CES and GDP growth rate

Topic	BES CI	CES CI	GDP
Company Earnings	0.74	0.79	0.81
Infrastructure	0.60	0.62	0.70
Trade	0.48	0.36	0.78

Key Findings and Future Works

Key Takeaways:

- NSIs are found to **match expected sentiment for key economic events**
- NSIs showed potential in **complementing existing surveys** (BES and CES) and providing **timely measurement of market sentiment**
- Topic modeling can **help uncover specific news themes** that may have more predictive content for specific key economic variables

Future Works:

- Enhance NSI-D (e.g., Add more linguistic rules, Part-of-Speech Tagging)
- Retrain FinBERT using the manually labelled dataset for NSI-ML
- Complement with other ML models (e.g., LM Lexicon + SVM, FinBERT + Random Forests, etc.)

Sources

- Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. M.I.T. Press.
- Loughran, Tim and McDonald, Bill, When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks (March 4, 2010). Journal of Finance, Forthcoming, Available at SSRN: https://ssrn.com/abstract=1331573
- Correa, Ricardo, Keshav Garud, Juan M. Londono, and Nathan Mislang (2017). Sentiment in Central Banks' Financial Stability Reports. International Finance Discussion Papers 1203.
- Zhang, L., Liu, B. (2017). Sentiment Analysis and Opinion Mining. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7687-1_907
- Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805
- Huang, Allen and Wang, Hui and Yang, Yi, FinBERT - A Large Language Model for Extracting Information from Financial Text (July 28, 2020). Contemporary Accounting Research, Forthcoming, Available at SSRN: https://ssrn.com/abstract=3910214 or http://dx.doi.org/10.2139/ssrn.3910214

Thank you!

* The views expressed herein are those of the authors' only and do not necessarily reflect those of the Bangko Sentral ng Pilipinas

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Do anecdotes matter? Exploring the beige book through textual analysis from 1970 to 2023[1]

Shengwu Du, Karen Guo, Flora Haberkorn, Abby Kessler, Isabel Kitschelt, Seung Jung Lee, Anderson Monken, Dylan Saez, Kelsey Shipman and Sandeep Thakur,
Board of Governors of the Federal Reserve System

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Do Anecdotes Matter?: Exploring the Beige Book through Textual Analysis from 1970 to 2023

Shengwu Du, Karen Guo, **Flora Haberkorn**, Abby Kessler, Isabel Kitschelt, **Seung Jung Lee**, Anderson Monken, Dylan Saez, Kelsey Shipman, and Sandeep Thakur

3rd IFC and Bank of Italy Workshop on
"Data Science in Central Banking: Enhancing the access to and sharing of data"

19 October 2023

The views expressed herein are those of the authors' and do not necessarily reflect the opinions of the Board of Governors of the Federal Reserve System or anyone else in the Federal Reserve System.

Overview

1. Motivation
2. Literature
3. Beige Book Modeling
4. Sentiment in Beige Books
5. Topics in Beige Books
6. Conclusion

Motivation - History of Beige Book

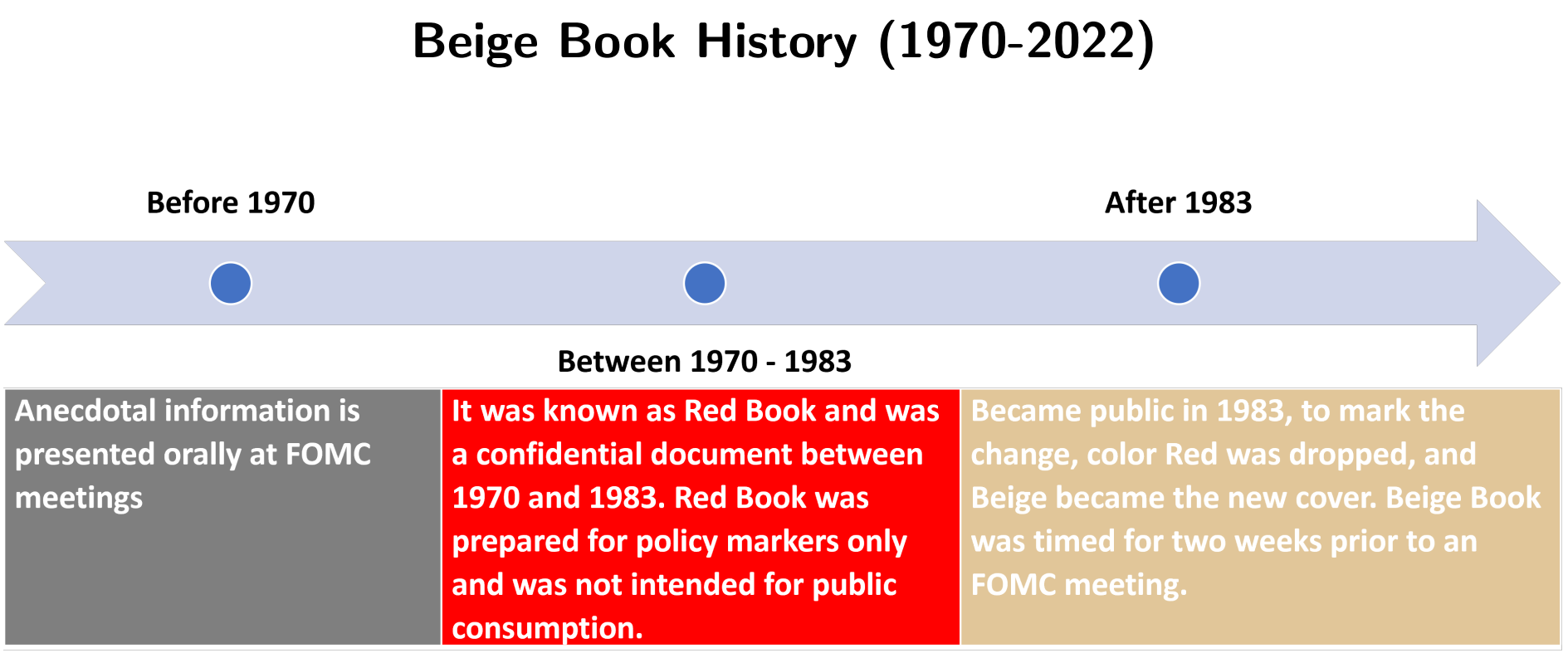


Figure: Roughly 40 years after it was made public, and nearly 50 years after it began as a document meant to mirror the condition of the economy, for some the Beige Book has become an important part of the FOMC process.

Literature

- Discussions in the Beige Book are useful to understand business sentiment in the US economy.
- **Mostly based on entire body of text/words and dictionary approach**.
- Machine learning approaches.
  - *Fettig, Rolnick, and Runkle (1999)*
  - *Balke and Petersen (2002)*
  - *Fulmer and Zhang (2017)*
  - *Gascon and Werner (2022)*
- **We use BERT-based sentiment and topic modeling**.

Beige Book Cleaning and Modeling

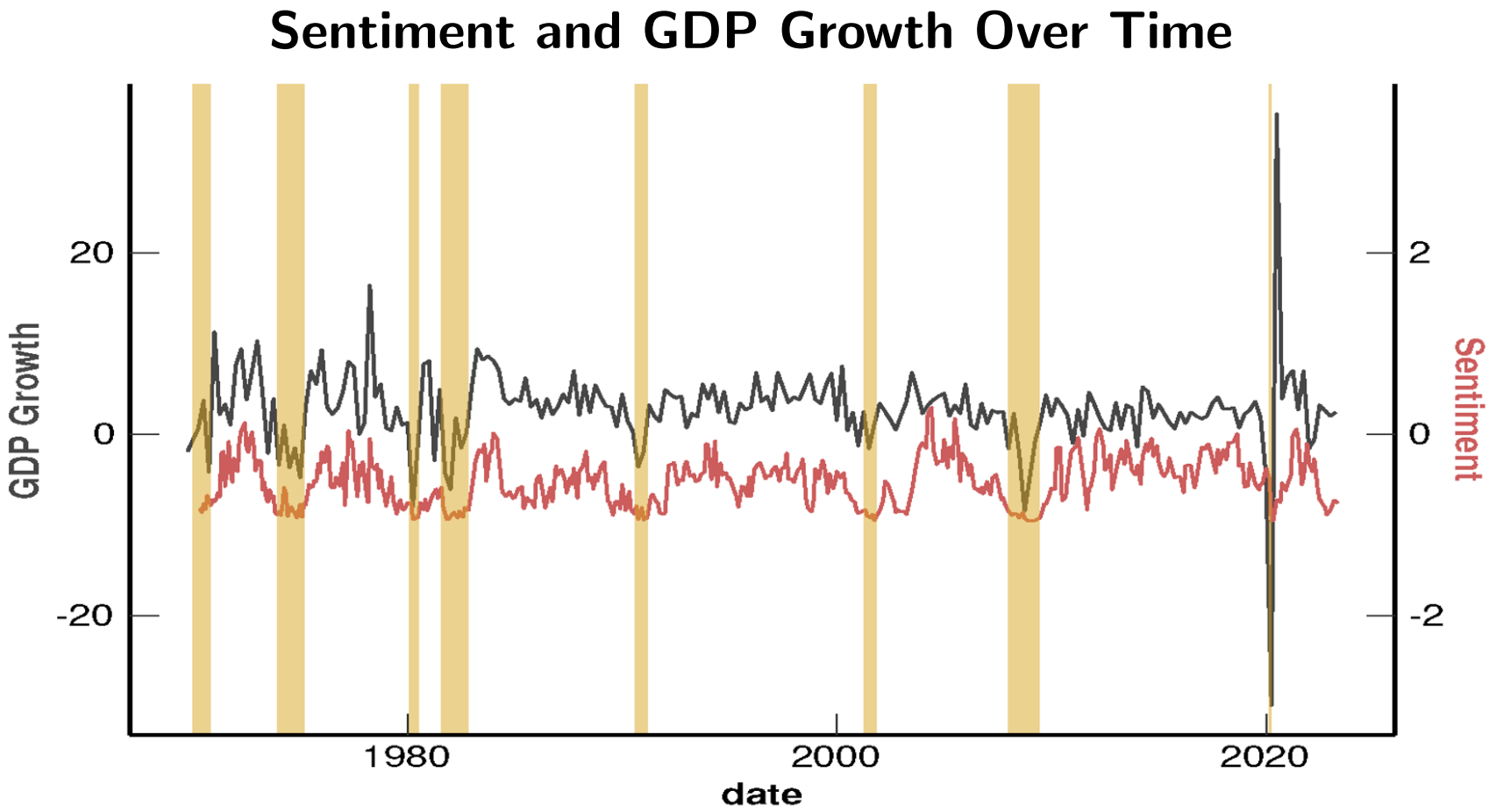
NLP
- Lowercase
- Remove Numbers and Stop words
- Tokenize

FinBERT
- Sentiment Scoring
- Pre-trained NLP model that is trained on financial data

LDA
- Topic Modeling

Figure: The Beige Book Model

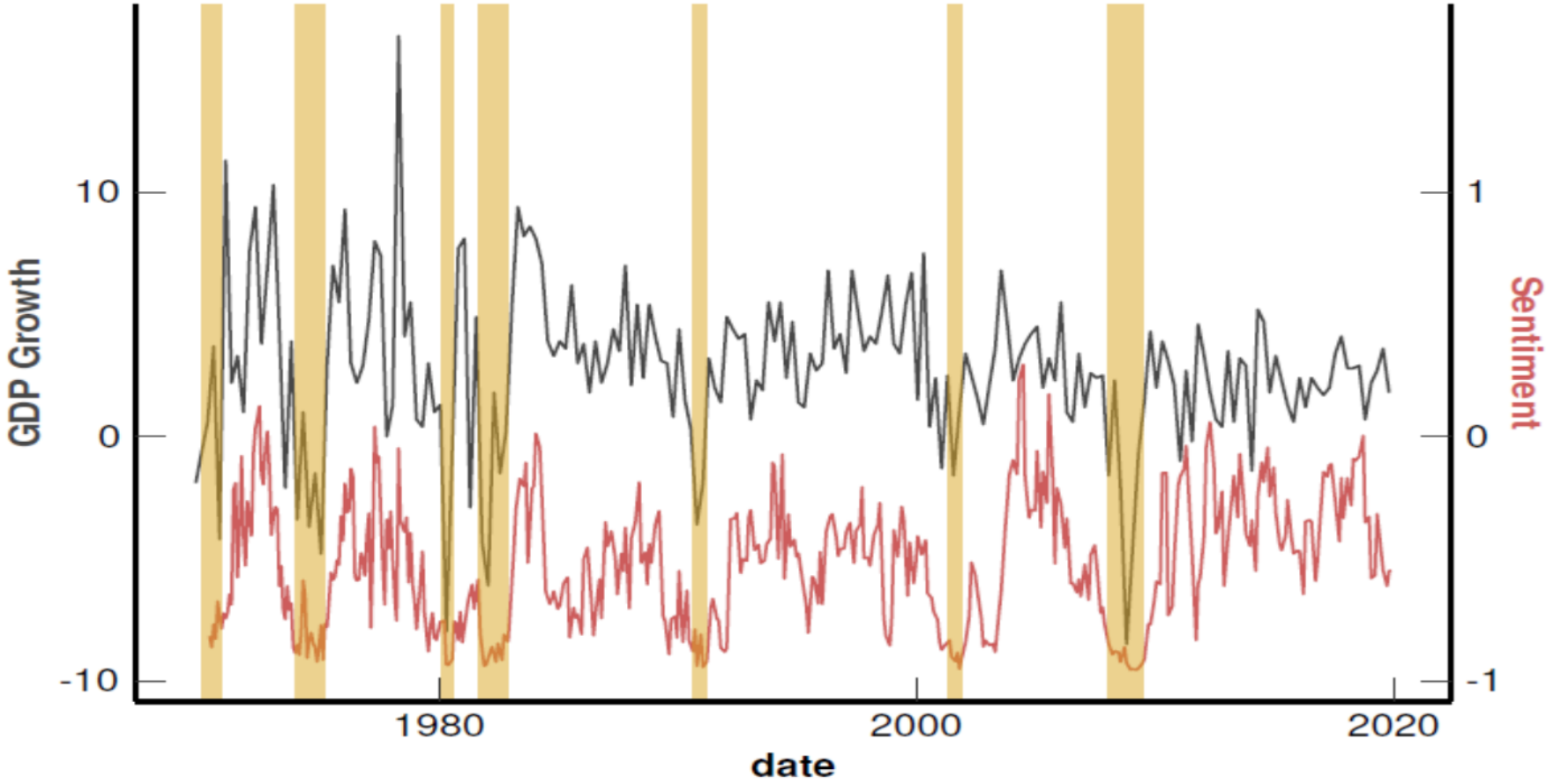
Sentiment and GDP Growth Over Time



Correlation is 0.34

Sentiment and GDP Over Time

Sentiment and GDP Growth Over Time (COVID omitted)



Correlation is 0.44 prior to COVID

Some Regressions

- Nowcast Real GDP Growth.
  - OLS regression on (lagged) GDP growth, yield spread (2 year – 10 year yield on US Treasuries), Beige Book sentiment, and SF-Fed News sentiment (Shapiro, Sudhof, and Wilson, 2020), separately and together.
- Nowcast Recessions.
  - Logit regression on (lagged) GDP growth, yield spread, Beige Book sentiment, and SF-Fed News sentiment, separately and together.
- Can also forecast.

Nowcast: BB Sentiment and Real Activity

Sentiment and GDP Growth Over Time

	Real GDP Growth (seasonally adjusted annual rate)			
	(1)	(2)	(3)	(4)
Lagged Real GDP Growth	0.387***	0.271***	0.289***	0.223***
	(0.074)	(0.077)	(0.074)	(0.077)
Yield Spread	−0.206			−0.248
	(0.236)			(0.237)
News Sentiment		4.926***		3.070**
		(1.240)		(1.380)
FinBERT Beige Book (text concatenated)			3.380***	2.778***
			(0.831)	(0.927)
Observations	159	159	159	159
R^2	0.159	0.232	0.236	0.275
Adjusted R^2	0.148	0.223	0.226	0.256

Note: *p<0.1; **p<0.05; ***p<0.01

Regressions based on 159 observations from 1980:Q2 to 2019:Q4.

Out-of-Sample RMSE Comparison for Nowcasting

	2020 Q1	2020 Q2	2020 Q3	2020 Q4
Spec. 4 vs Spec. 4 (without BeigeBook)	0.977	**0.975**	**0.958**	0.878
	0.159	0.078	0.057	0.112
Spec. 3 vs Spec. 3 (without BeigeBook)	**0.943**	**0.941**	**0.927**	**0.768**
	0.042	*0.038*	*0.049*	0.098
Spec. 4 vs Spec. 3 (without BeigeBook)	**0.902**	**0.903**	**0.887**	**0.636**
	0.025	*0.032*	*0.050*	0.099
Observations	14	13	12	11

Nowcast: BB Sentiment and Recessions

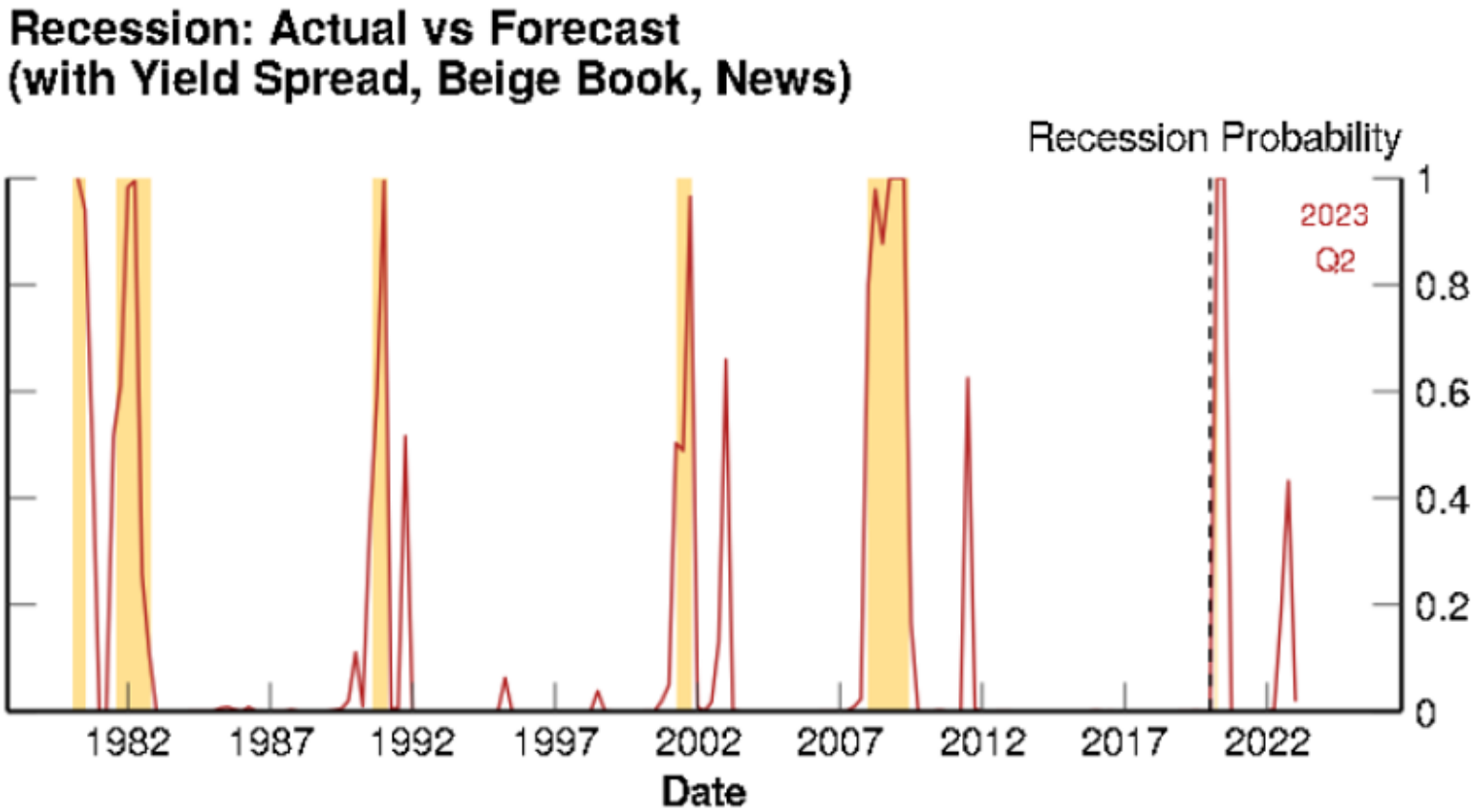
Sentiment and GDP Over Time

	Recession			
	(1)	(2)	(3)	(4)
Lagged Real GDP Growth	−0.685***	−0.489***	−0.491***	−0.552**
	(0.145)	(0.168)	(0.187)	(0.273)
Yield Spread	0.188			−0.039
	(0.343)			(0.540)
News Sentiment		−11.877***		−11.669**
		(3.403)		(5.221)
FinBERT Beige Book (text concatenated)			−28.823***	−29.295**
			(9.531)	(11.625)
Observations	159	159	159	159
Akaike Inf. Crit.	75.614	53.509	37.704	33.762

Note: *p<0.1; **p<0.05; ***p<0.01

Regressions based on 159 observations from 1980:Q2 to 2019:Q4.

Out-of-Sample Comparison Recession



Dashed line marks when our model's out-of-sample forecasts begin, highlighted periods are NBER recessions.

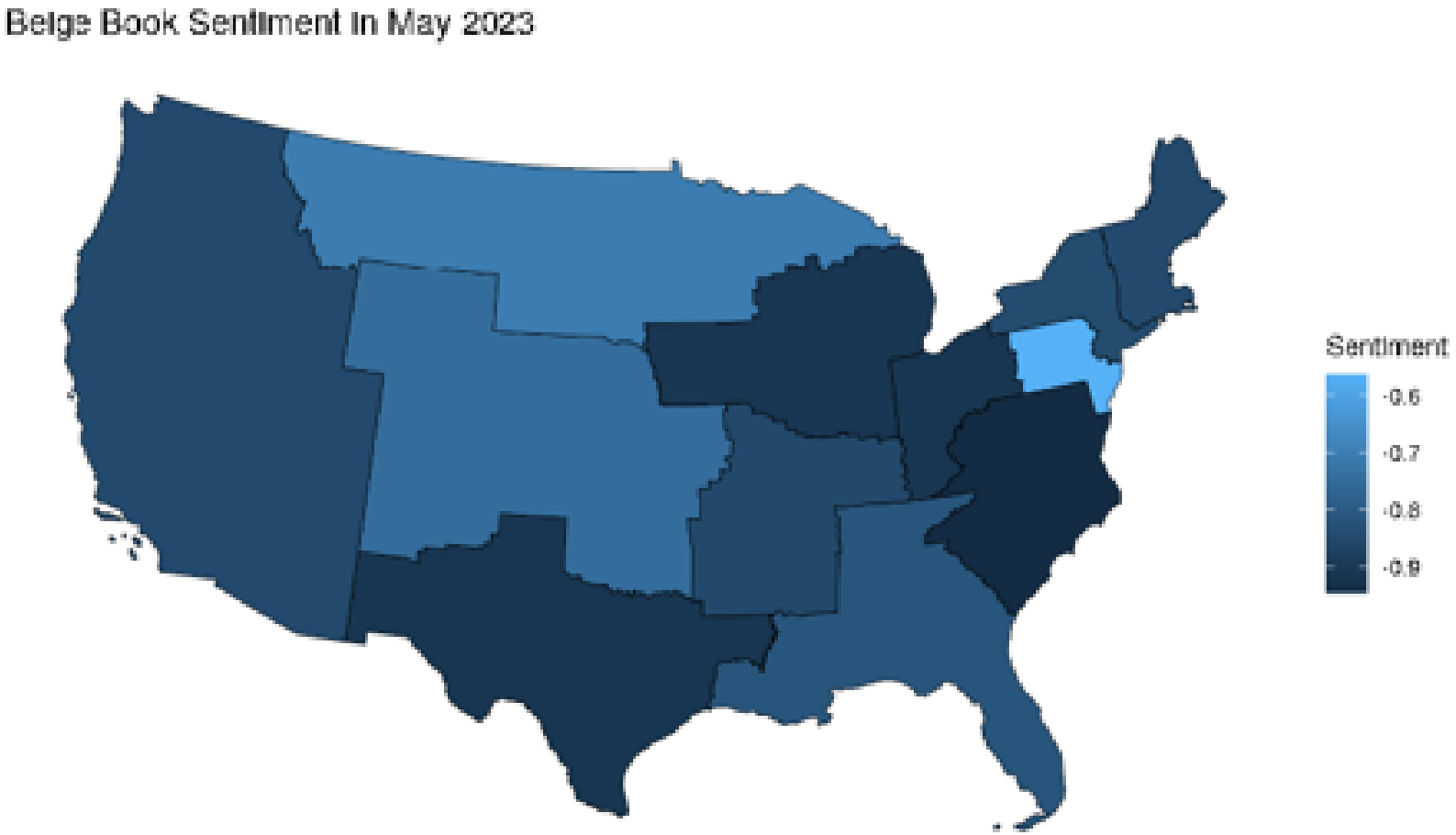
Forecast: BB Sentiment and Real Recessions

Sentiment and GDP Over Time

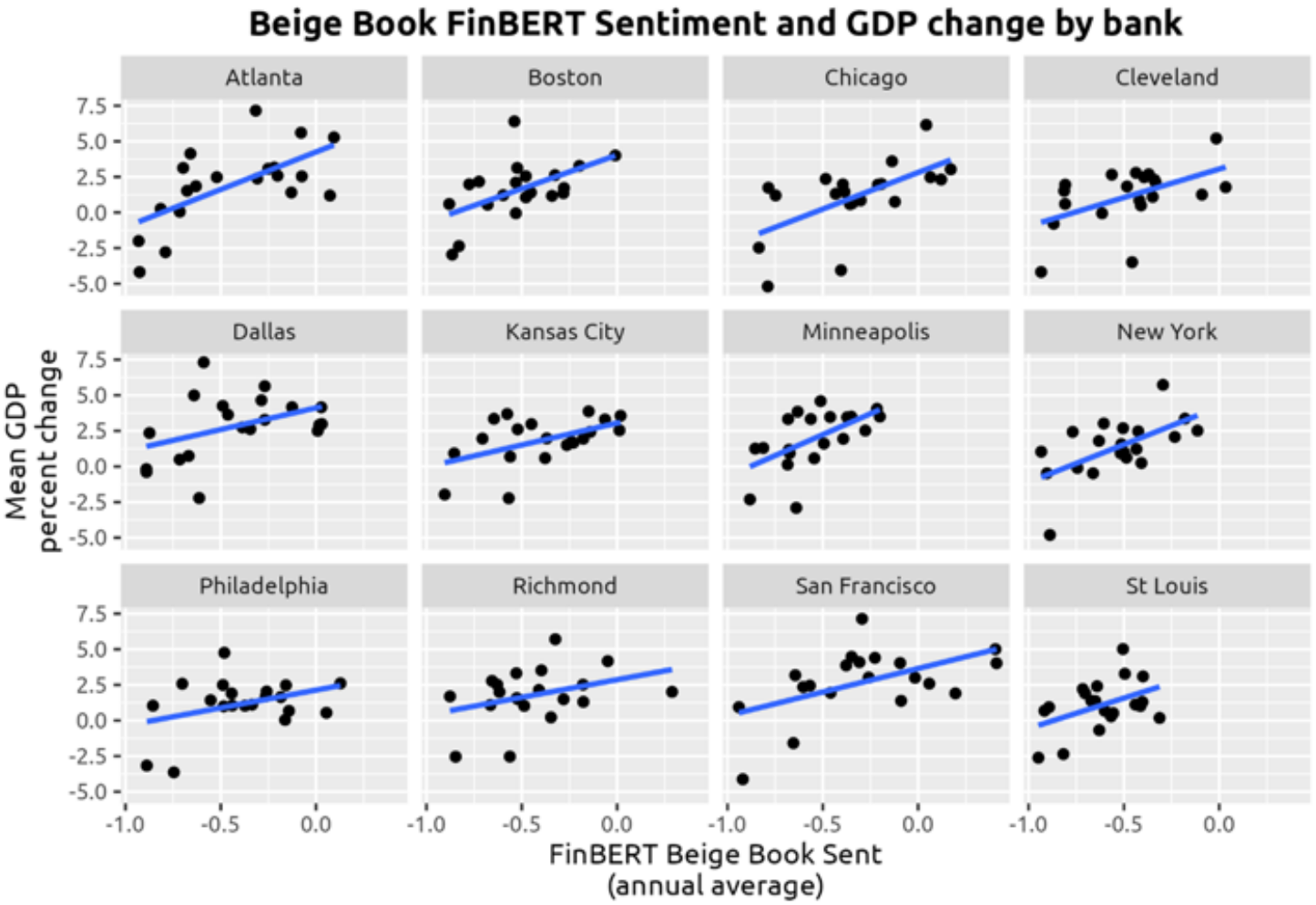
	Recession			
	(1)	(2)	(3)	(4)
(Twice) Lagged Real GDP Growth	−0.367***	−0.144	−0.083	0.104
	(0.095)	(0.104)	(0.109)	(0.143)
Lagged Yield Spread	−0.414			−1.151**
	(0.286)			(0.526)
Lagged News Sentiment		−10.071***		−9.159**
		(2.623)		(3.752)
Lagged FinBERT Beige Book Sentiment (text concatenated)			−17.326***	−21.073***
			(4.777)	(7.990)
Observations	159	159	159	159
Akaike Inf. Crit.	97.696	75.748	60.906	51.764

Note: *p<0.1; **p<0.05; ***p<0.01
Regressions based on 159 observations from 1980:Q2 to 2019:Q4.

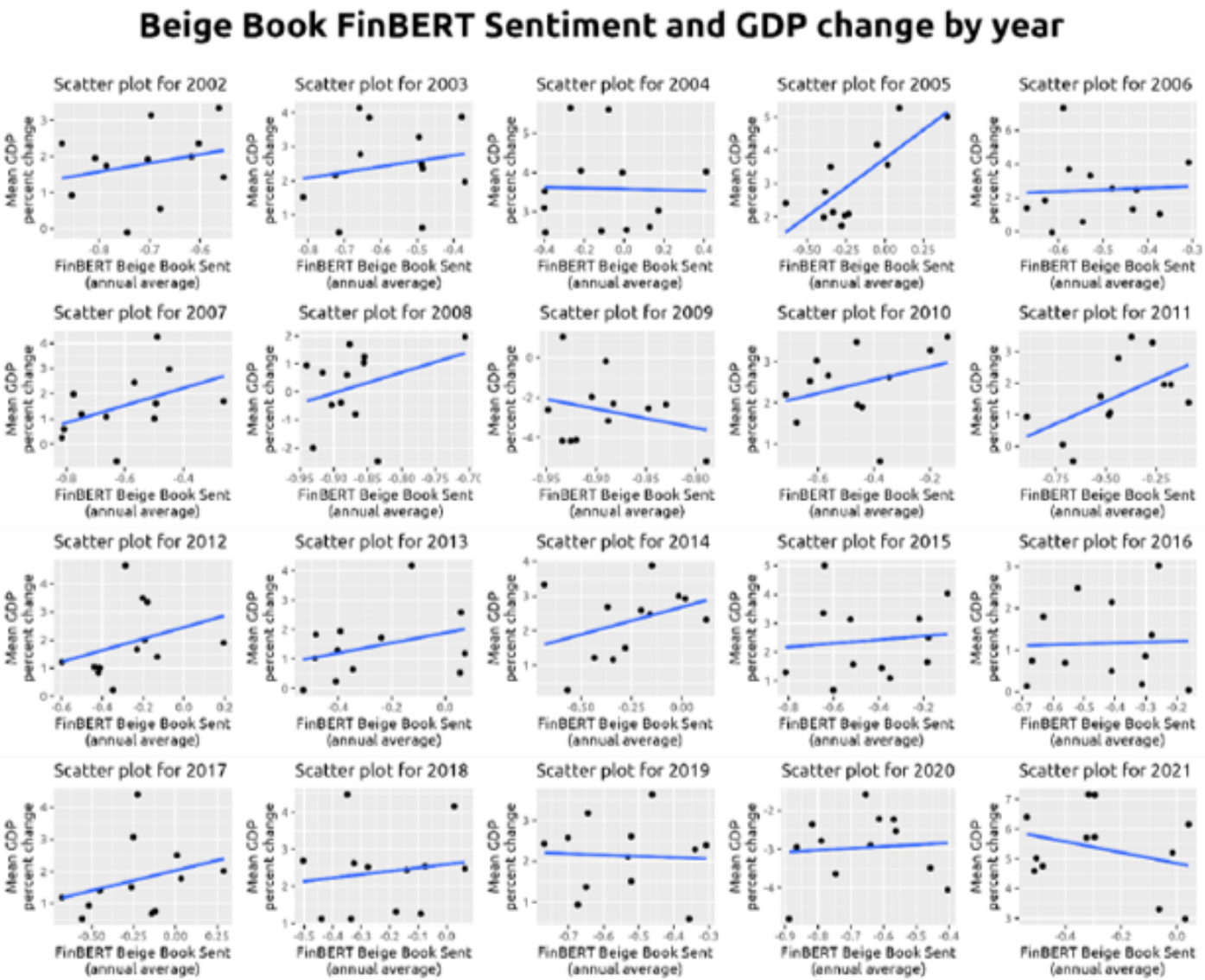
Regional Differences in Beige Book Sentiment



Regional Differences in Sentiment, by District



Regional Differences in Sentiment, by Year



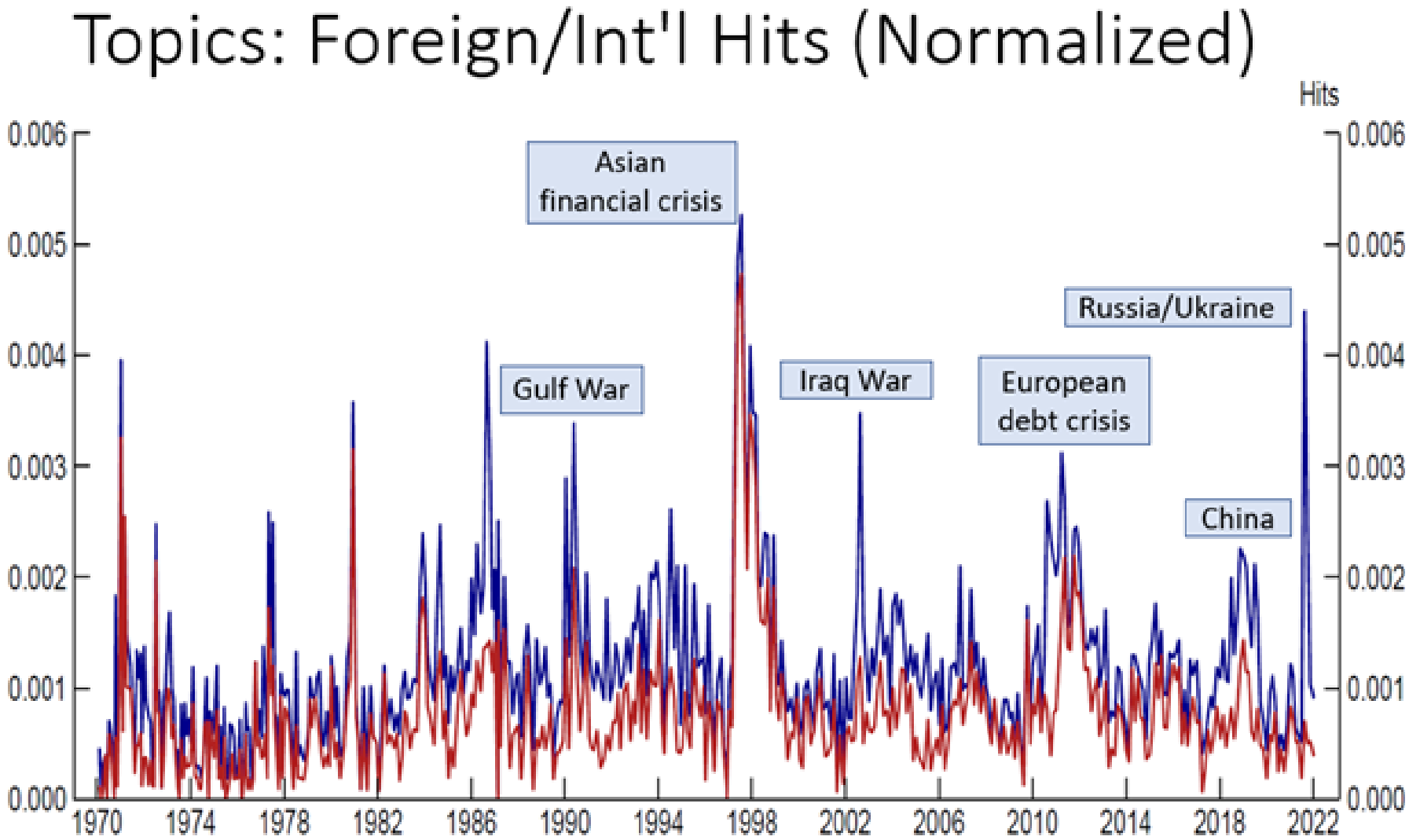
Regional Differences in Sentiment

	Real GDP Growth (percent change)	
	(1)	(2)
FinBERT Sentiment (District level)	3.611***	1.490**
	(0.622)	(0.638)
FinBERT Sentiment (combined excl. District)		3.852***
		(0.761)
Constant	3.296***	4.906***
	(0.529)	(0.460)
District FE	Yes	Yes
Year FE	Yes	Yes
Observations	240	240
R^2	0.482	0.384
Adjusted R^2	0.408	0.349
Residual Std. Error	1.668 (df = 209)	1.749 (df = 226)
F Statistic	6.483*** (df = 30; 209)	10.852*** (df = 13; 226)

Note: *p<0.1; **p<0.05; ***p<0.01

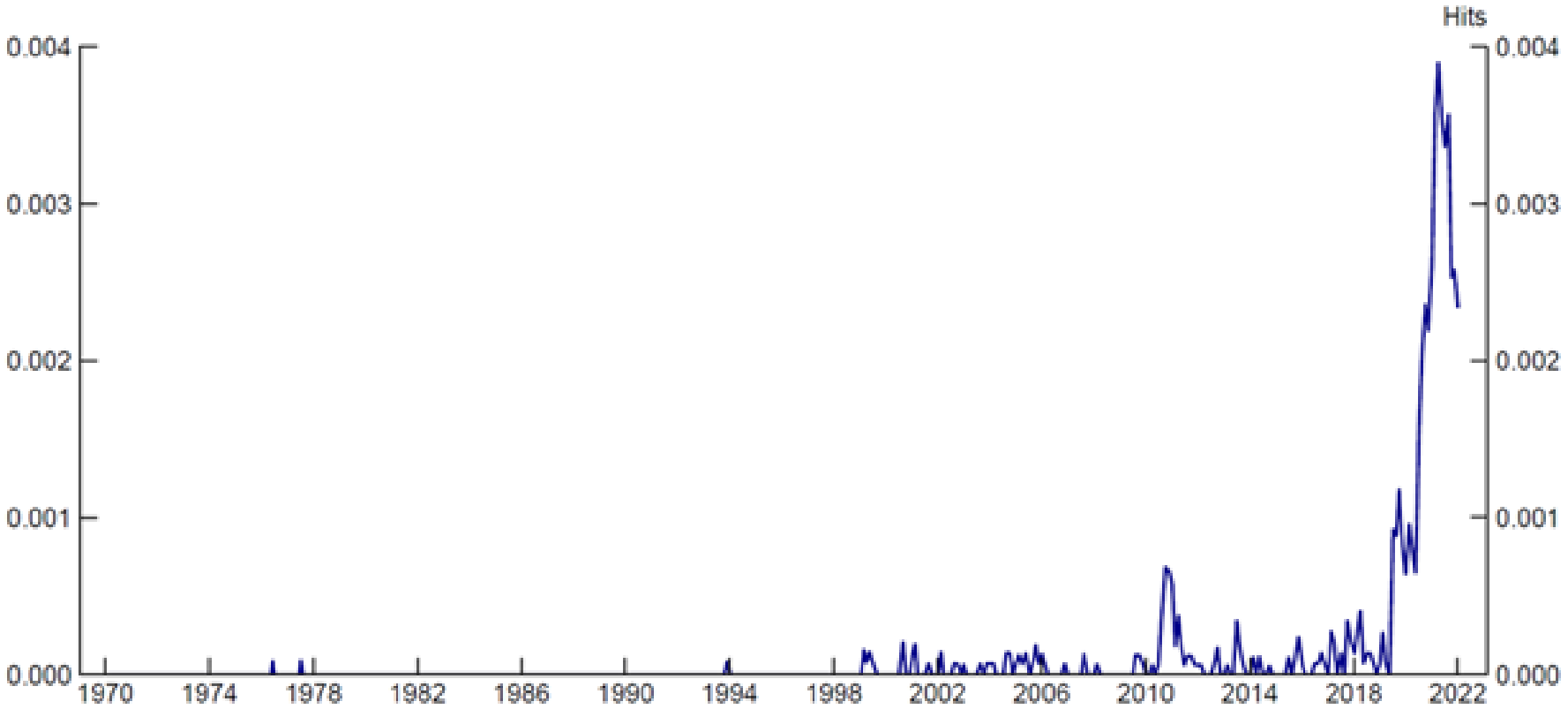
Regressions based on 240 from 2002 to 2021. Method: Concatenated text average.

Beige Book Search Engine - Key Words



Beige Book Search Engine - Key Words

Topics: Supply Chain Hits (Normalized)



Beige Book LDA Analysis

- LDA: ML algorithm that identifies main topics in a group of documents
  - Views each document as a distribution of topics and each topic as as a distribution of words
  - Number of topics chosen via coherence scores
- LDAvis: Visualization tool for LDA output
  - Displays topics and associated words (or words and associated topics) by prevalence

Beige Book LDA Results

Recessionary Period	Topics (Negative Sentiment)	Topics (Positive Sentiment)
1972-1974	• Savings, credit, and loans (1) • Raw materials and prices (2, 5) • Construction (4, 7) • **Fuel and energy prices (8)**	• Agriculture and food prices (1, 7, 8) • Consumer and business confidence (2) • Supply chain (3) • Employment and labor market (4, 6) • Construction and loans (5)
2007-2008	• Commercial real estate and construction (1, 5) • Wages, prices, and costs (2) • **Real estate, loans, and credit quality (3)** • Business activity and employment (4) • Cars, inventories, and sales (6)	• Prices and raw materials (1, 2) • Real estate and construction (3, 5, 6) • Business activity and employment (4)
2022	• Real estate (1) • **Higher prices (2, 4)** • Cars and loans (3) • Wages and labor (5) • Mortgages and credit quality (6) • **Supply chain (7, 8)**	• Real estate (1) • **Travel and consumer activity (4)**

Conclusion

- Anecdotes seem to matter
- Sentiment in Beige Books sheds light on macro activity, both in the time series and in the cross-section
- Topics are relatively easy to extract; large language models (LLMs) can help with more nuanced topical analysis going forward

Thank You!

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

Central bank communication on economic activity[1]

Sercan Eraslan and Eniko Gabor-Toth,
Deutsche Bundesbank

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Central bank communication on economic activity

3rd IFC Workshop on Data Science in Central Banking

Eniko Gabor-Toth[1] Sercan Eraslan [1]

[1]Deutsche Bundesbank

October 19, 2023

This presentation represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem.

Eraslan & Gabor-Toth (2023)

Central bank communication is important for shaping expectations on...

- monetary policy
- inflation
- ... and last but not least on economic activity

This paper:

- Measure the information content of the central bank narrative on underlying economic conditions and risks
- Generate sentiment indices for topics/sections in the monthly reports of the Bundesbank (German and English)
- Forecasting exercise: do the MR sentiment and uncertainty indices improve the forecasting accuracy of key macro indicators in real-time?

Review of the relevant literature

So far most studies using text-based data in business cycle analysis and finance make use of newspaper/press articles...

- Tetlock (2007), Baker et al. (2016), Ulbricht et al. (2017), Ardia et al. (2019), Larsen & Thorsrud (2019), Ellingsen et al. (2020), Kalamara et al. (2020), Thorsrud (2020), Aguilar et al. (2021), Larsen et al. (2021) etc.

Studies on central bank communication have mostly focused on monetary policy/inflation...

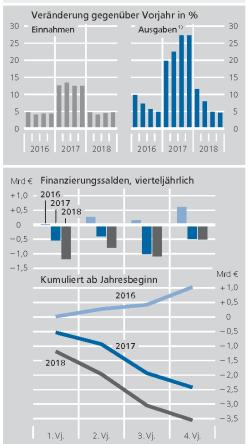
- Gürkaynak et al. (2005), Ehrmann & Fratzscher (2007), Haldane & McMahon (2018), Enders et al. (2019), Coibion et al. (2020), Ehrmann & Talmi (2020), Swanson (2021) etc.

... growing literature on incorporation of text-based data obtained from central banks and research institutes into macroeconomic forecasting...

- Sharpe et al. (2020): Fed's Green Book for the US; Clements & Reade (2020), Jones et al. (2020): BoE's IR for the UK; Müller (2020): Bundesbank and research institutes forecast reports for DE

Textual database

Finanzen der sozialen Pflegeversicherung[1]



Quelle: Bundesministerium für Gesundheit. * Vorläufige Vierteljahresergebnisse. Die endgültigen Jahresergebnisse weichen in der Regel von der Summe der ausgewiesenen vorläufigen Vierteljahresergebnisse ab, da letztere nicht nachträglich revidiert werden. 1 Einschl. der Zuführungen an den Pflegevorsorgefonds.
Deutsche Bundesbank

schnittliche Satz sank zum Jahresbeginn aber nur auf 1,0 %. Selbst mit den neuen Belastungen aus dem Terminservice- und Versorgungsgesetz könnten somit wieder Überschüsse entstehen.

Geforderte Beitragsentlastung von Betriebsrenten mit erheblichen Einnahmenausfällen verbunden

Diskutiert wird derzeit, die Beiträge auf Betriebsrenten und andere Versorgungsbezüge zu senken. Dies würde die Krankenversicherung finanziell stark belasten. Würden die genannten Beiträge halbiert (wie vor 2004 der Fall),[5] würden Einnahmen von etwa 3 Mrd € pro Jahr entfallen.[6] Dies entspricht Mindereinnahmen in Höhe von etwa 0,2 Beitragssatzprozentpunkten.

Begründung für halbierten Beitragssatz nicht überzeugend

Begründet werden die Forderungen nach Beitragsermäßigungen mit dem Argument, dass es derzeit zu einer „Doppelverbeitragung" komme: Beiträge würden sowohl in der Anspar- als auch in der Rentenphase fällig. Hierzu kommt es jedoch lediglich in einigen Ausnahmefällen.[7] In der Regel unterliegen Zuführungen zu Versorgungseinrichtungen oder Aufstockungen von Pensionsrückstellungen keinen GKV-Beiträgen.[8] Insofern erscheint es folgerichtig, Betriebsrenten und andere Versorgungsbezüge wie hinausgeschobene Arbeitsentgelte zu verbeitragen und keine Beitragsermäßigung vorzunehmen.

Perspektivische Verschlechterung der Finanzlage

Die derzeit gute Finanzlage mit Überschüssen und hohen Reserven verdeckt absehbaren demografischen Druck. Leistungsausweitungen oder Begünstigungen mögen zwar kurzfristig finanzierbar erscheinen. Unpopuläre Leistungseinschränkungen oder steigende Beitragssätze, die die gesamtwirtschaftliche Entwicklung belasten, werden aber damit künftig in noch stärkerem Maße erforderlich. Dies sollte bei den aktuellen Diskussionen nicht aus den Augen verloren werden.

Soziale Pflegeversicherung

Gestiegenes Defizit ließ Rücklagen abschmelzen

Das Defizit der sozialen Pflegeversicherung stieg im Jahr 2018 um 1 Mrd € auf 3½ Mrd €.[9] Die Einnahmen wuchsen um 4 %, darunter die Beiträge der Beschäftigten um gut 4½ % und die

5 Mit dem GKV-Modernisierungsgesetz wird seit 2004 der volle GKV-Beitrag erhoben.
6 Vgl.: Bundestags-Drucksache 19/4431 vom 21. September 2018, S. 4.
7 So etwa im Fall von Entgeltumwandlungen für Beiträge an Direktversicherungen, Pensionskassen oder Pensionsfonds, die über 4 % der Beitragsbemessungsgrenze in der gesetzlichen Rentenversicherung (West) hinausgehen (§ 1 Abs. 1 Nr. 1 Sozialversicherungsentgeltverordnung).
8 Vgl.: Deutscher Bundestag, Wissenschaftliche Dienste, Vermeidung der sog. „Doppelverbeitragung" von Betriebsrenten aus Direktversicherungen und sonstiger Versorgungsbezüge in der gesetzlichen Kranken- und Pflegeversicherung – Aktueller Diskussionsstand und Reformpläne, 2016, S. 5 ff.
9 Vorläufige Rechnungsergebnisse gegenüber endgültigem Jahresergebnis. Hier wird die Entwicklung ohne den Pflegevorsorgefonds beschrieben, der zum Vermögensaufbau seit 2015 planmäßig Überschüsse erzielt. Zum Jahresende 2018 betrugen die kumulierten Rücklagen 5 Mrd €.

- Data source: monthly and quarterly reports of the Deutsche Bundesbank: German and English in PDFs
- Corpus: 396 reports from January 1990 through December 2022
- **Challenges**: scanned documents, changes in PDF versions over time, unicode errors, text structure (two columns), reconstruction of words separated at line endings, side notes, footnotes

From files to database

Key Question: How can we obtain a structured, informative, quantitative representation of these texts?

1. Remove graphs and charts, footnotes, side notes, convert scanned pdfs into editable pdfs
2. Extract texts using: R (tabulizer, tesseract, pdftools)
3. Text cleaning and pre-processing, reconstruct words separated at line endings in R, experiment with ChatGPT 3.5 for pre-processing
4. Assign topics

Text extraction: example

Turning raw text files into structured databases is often a challenge:

- June 2000

```
Konjunkturlage\r\nGesamtwirtschaft\r\nErsten Berechnungen des Statistischen
Bun-\r\ndesamtes zufolge ist die gesamtwirtschaftliche\r\nProduktion auch i
m ersten Vierteljahr 2000\r\nrelativ kräftig gestiegen. Saison- und kalen-\
r\nderbereinigt erhöhte sie sich gegenüber den\r\nHerbstmonaten 1999 um 3\0
314 %. Auf ein Jahr\r\nhochgerechnet wächst damit die deutsche\r\nWirtschaf
t seit Jahresmitte 1999 mit fast 3%.\r\nDen vergleichbaren Vorjahrsstand üb
ertraf\r\ndas reale Bruttoinlandsprodukt nach Aus-\r\nschaltung von Kalende
runregelmäûigkeiten\r\num 2,3%.
```

- July 2000

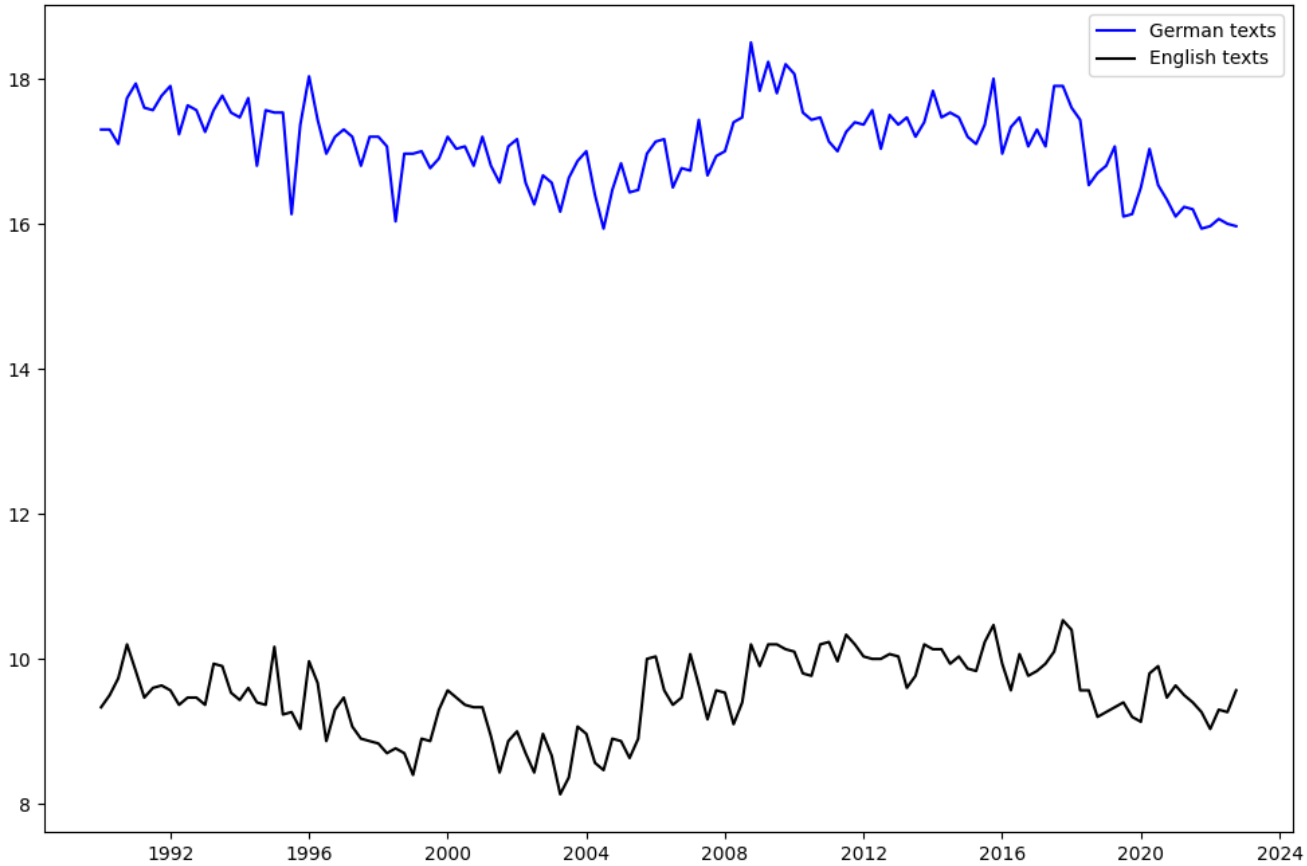
```
\026\017\t\030\003\t\r\004\003\020\023\f\031\002\r\n\032\002\020\f\020\v\002\0
21\004\002\t\n\002\005 \033\002\034\002\020\v\002\r\n\a\017\016 #$%\f\036\020\
031\017\016&'$\035\f$&\020\f\031 !\v\031 \v\f\037\030 \017( )\v\017\r\n%\016\f
\020\n\017\037\030 \v\f*!+\031\020\036 ,\016\031\017\037\030\020\016\020\003 \
r' \016\031\030-\030\020\016 \036\017\037\030\r\n%\017\016 .\v\037\030*\031\v,
\016 $\v\037\030 /\031\"\016\f,$\017\036\036\016$ %\016\036 0\016\0311\r\n\v\0
312\016\017\020\016$%\016$ 3\016!\016\0312\016\036 \036\v\017\036'$2\016\031\0
16\017$\017,\020 \017$\r\n%\017\016\036\016( )'$\v\020 $'\037\030(\v\n\036 \f(
\001\0224\0015\003 6\v\036\036\020\r\n(\v$7 \f( &\f\031\"*\031\017\036\020\017
,\016 \"\f*+\n\n\017,\016 \r\037\030!\v$&\f$,\016$\r\n\v\f\036\"\f,\n\016\017\
037\030\016$7 %\017\016 2\016\017%\016$ 6\0318\030\035\v\030\031\036('$\v\020\
016\r\n9:\031\017\n \f$% )\v\017 \"\f\036\v((\016$7 \036' \016\031,\0172\020 \
036\017\037\030\r\n,\016,\016$82\016\031 %\016( \a\f\031\037\030\036\037\030$\
017\020\020 %\016\036 \022\003 0\017\016\031\020\016\n1\r\n\035\v\030\031\016\
036 \001\004\004\004 \016\017$ \036\016\030\031 &\031+*\020\017,\016\031 9$\03
6\020\017\016, \f(\r\n&$\v:: \0215\003 \a\016\031 .\v\037\030*\031\v,\016\"\f!
\v\037\030\036 \017( 0\016\0311\r\n,\n\016\017\037\030 \"\f\031 \016$\020\036:
\031\016\037\030\016$%\016$ 0'\031\035\v\030\031\036\"\016\017\0207 %\016\031
\017$\r\n%\016$ \016\031\036\020\016$ )'$\v\020\016$ %\016\036 ;\v\030\031\016
\036 2\016\031\016\017\020\036 \"!\016\0171\r\n\036\020\016\n\n\017, ,\016!\01
6\036\016$ !\v\0317 <\016\031,\031-=\016\031\020\016 \036\017\037\030 !\016\01
7\020\016\031\r\n\v\f* \022\002\0234\0025\003\r\n\a\v\030\017$\020\016\031 \03
6\020\016\030\020 <'\031 \v\n\n\016( \016\017$\016 \v$\030\v\n\020\016$% \030'
\030\016\r\n\a>$\v(\017& \017( 9\f\036\n\v$%\036,\016\036\037\030+*\020\003 ?\
017\016\031 \016\031\031\016\017\037\030\020\016\r\n%\016\031 0'\031\035\v\030
\031\036\v2\036\020\v$% \017( 9:\031\017\n\034)\v\017 @\001\0045\003 \a\017\01
6\r\n.\v\037\030*\031\v,\016 \v\f\036 %\016
```

Main themes over time: TF-IDF

- Generate a heatmap for important keywords in German texts.

Readability: Flesch-Kincaid Grade Level

- 1st grade - 5th grade Very easy to read.
- 5th grade - 11th grade Average level.
- 11th grade - 18th grade The text is for skilled readers. For example, an academic paper.



Textual analysis

- The sentiment/tone is simply calculated as the relation between positive and negative words:

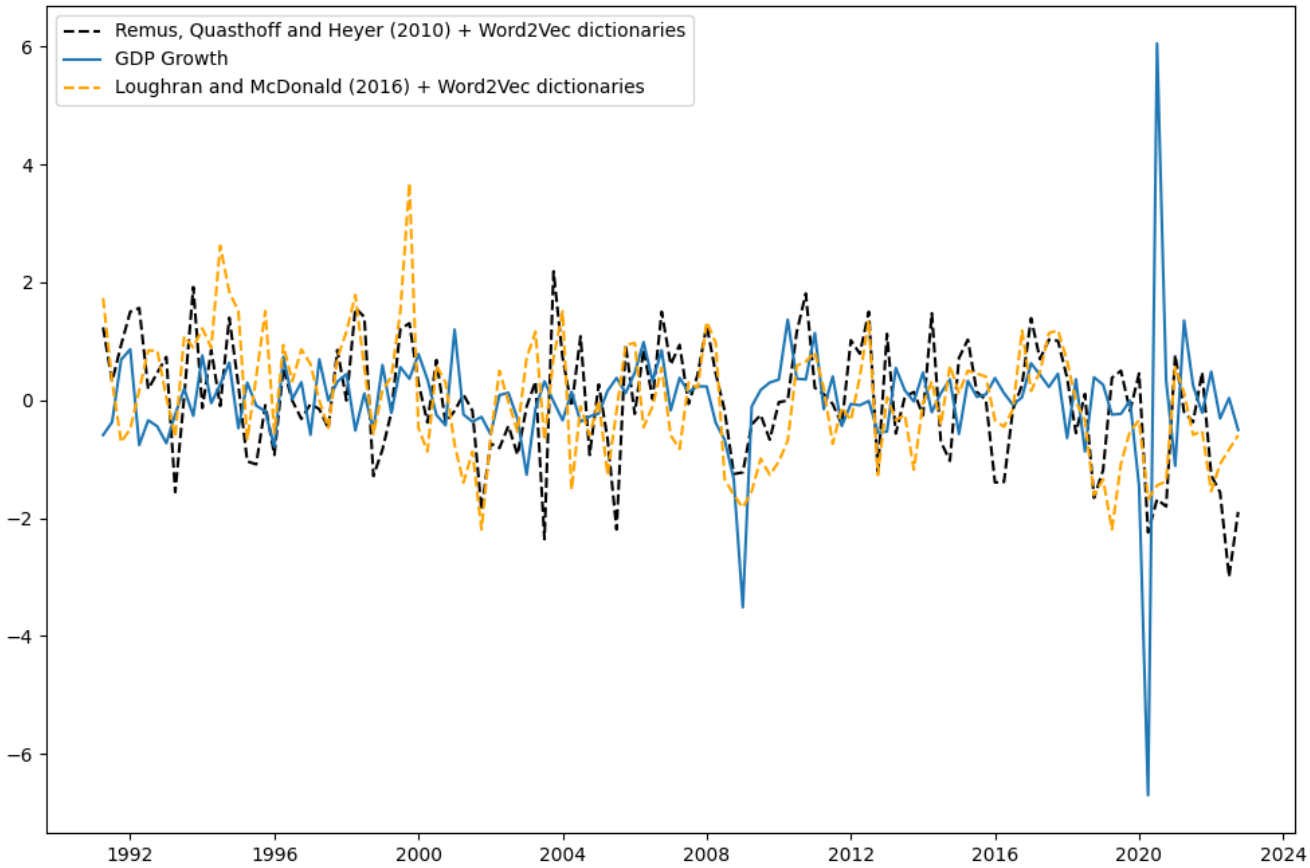
$$s_{n,t} \equiv \frac{W^{+}_{n,t} - W^{-}_{n,t}}{W_{n,t}}$$

where $W^{+}_{n,t}$, $W^{-}_{n,t}$ and $W_{n,t}$ are the numbers of positive, negative and all words in topic n in month t.

Result: sentiment, uncertainty and readability indices...

- for topics as they are written in monthly reports, such as industry, labour market, financial/securities markets, prices etc.
- for the overall economic conditions and the whole monthly report

SI economic conditions vs. GDP growth



Mixed-frequency dynamic factor model

We use a mixed-frequency dynamic factor model (MF-DFM) for the forecasting exercise.

$$y_t = Hs_t + u_t, \qquad u_t \sim N(0, \Sigma_{u_t})$$
$$s_t = Fs_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma_{\varepsilon_t})$$

... and follow Mariano & Murasawa (2003) to deal with the quarterly indicator. Assuming $lnY_t^q = \frac{1}{3}(lnY_t^m + lnY_{t-1}^m + lnY_{t-2}^m)$ and taking three period differences:

$$y_t^q = \frac{1}{3}y_t^m + \frac{2}{3}y_{t-1}^m + y_{t-2}^m + \frac{2}{3}y_{t-3}^m + \frac{1}{3}y_{t-4}^m$$

with $y_t^q := \Delta_3 lnY_t^q$ (observed every third period) and $y_t^m := \Delta_3 lnY_t^m$ (unobserved).

Estimation and forecast setup

Dataset

- Quarterly GDP and 13 monthly macroeconomic and financial indicators (gdp, ip, new orders, construction, retail sales, exports, unemployment, cpi, hwwi, dax30, corp. spread, ifo c, ifo pp, ifo ee)
- monthly sentiment indices (balance of payments and foreign trade, construction, economic conditions, financial/securities markets, industry, labour market, monetary policy, prices, public finance)
- from 1991M1 to 2021M6
- real-time vintages from 2007M7 - 2021M6

Model specifications: MF-DFMs without sentiment indices vs. MF-DFMs with sentiment indices

Estimation: State space form by means of Kalman filter $(n_f = 3, p = 3, q = 2)$, expanding window

Forecast

- up to 2 quarters ahead
- real-time forecast evaluation period 2007Q3−2021Q2
- Evaluation via MAD (calculated with first releases)

Forecast evaluation results

	2Q-ahead			1Q-ahead			Nowcast			Back
$h =$	8	7	6	5	4	3	2	1	0	-1
GDP	0.95	0.96	0.95	0.98	0.96	1.01	0.95	0.88	1.16	1.20
IP	0.91	0.95	0.99	1.01	0.90	0.96	0.89	0.85	1.00	1.05
NewOr	0.92	0.98	0.96	1.01	0.98	1.03	0.95	0.88	0.93	0.98
Const	1.00	1.04	1.08	1.08	1.01	1.02	0.93	0.99	1.07	0.99
Sales	1.00	1.00	1.02	1.00	1.02	0.99	0.87	1.04	0.96	1.04
Expo	0.94	0.97	0.98	1.05	0.99	1.03	0.93	0.93	1.02	1.05
UnEmp	0.96	0.95	1.02	1.00	0.93	0.90	1.00	1.00	1.00	-
CPI	1.00	1.00	1.00	0.98	1.01	0.99	0.95	1.02	0.97	-
HWWI	0.97	0.97	1.00	0.95	0.93	0.97	0.96	0.93	0.93	-
DAX30	0.97	1.04	1.05	1.02	1.06	0.99	1.01	1.07	0.98	-
CorSp	1.00	1.03	1.04	0.98	1.02	0.93	1.00	0.97	0.99	-

Notes: This table plots relative MADs, such that values less (grater) than 1 hint to a better (worse) forecast performance for the MF-DFM with sentiment indices.

Summary and relevance of the results

Key take-aways:

- Sentiment indices appear to lead to (slight) increases in forecast accuracy of some key macroeconomic indicators
- Monthly reports seem to have an additional information content on economic activity on top of those contained in macroeconomic and survey-based indicators
- This helps the private sector better forecast economic developments and thus can enhance the transmission of monetary policy

IFC-Bank of Italy Workshop on "Data science in central banking: enhancing the access to and sharing of data"

17-19 October 2023

From the ML model to practice: case study on NLP-based decision-making on the eligibility of security prospectuses[1]

Janek Blankenburg, Maximilian König, Philipp Rothhaar and Bernd Rusitschka,
Deutsche Bundesbank

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.



From the ML Model to Practice

Case Study on NLP-based Decision-Making on the Eligibility of Security Prospectuses

SCAI
Service & Community Center
for Artificial Intelligence



Maximilian König
AI Solution Architect



Bernd Rusitschka
AI Expert in DG Markets



Janek Blankenburg
AI Application Engineer



Philipp Rothhaar
Expert in DG Markets

In collaboration with further colleagues from DG Markets and Prof. Christian Hänig and Serhii Hamotskyi from Anhalt University of Applied Sciences

Agenda

SCAI
Service & Community Center for Artificial Intelligence

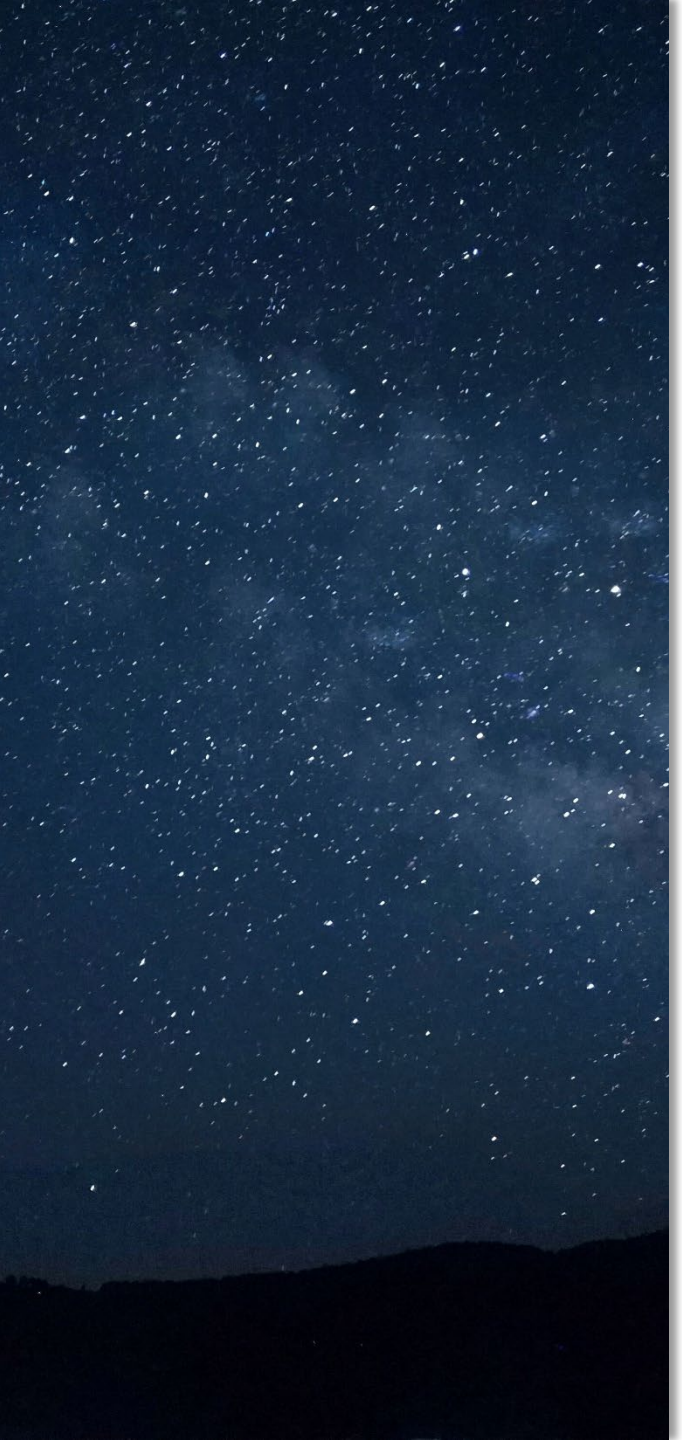
Status quo ante	Deciding the Eligibility of Securities' Prospectuses
Training a model	Proof of Concept with a fine-tuned model
Integration	**…** of the model into the business process
Learnings	… from the process

DEUTSCHE BUNDESBANK EUROSYSTEM

SCAI Service & Community Center for Artificial Intelligence

Status Quo Ante

Deciding the Eligibility of Securities‘ Prospectuses



Status Quo Ante

Deciding the Eligibility of Securities Prospectuses

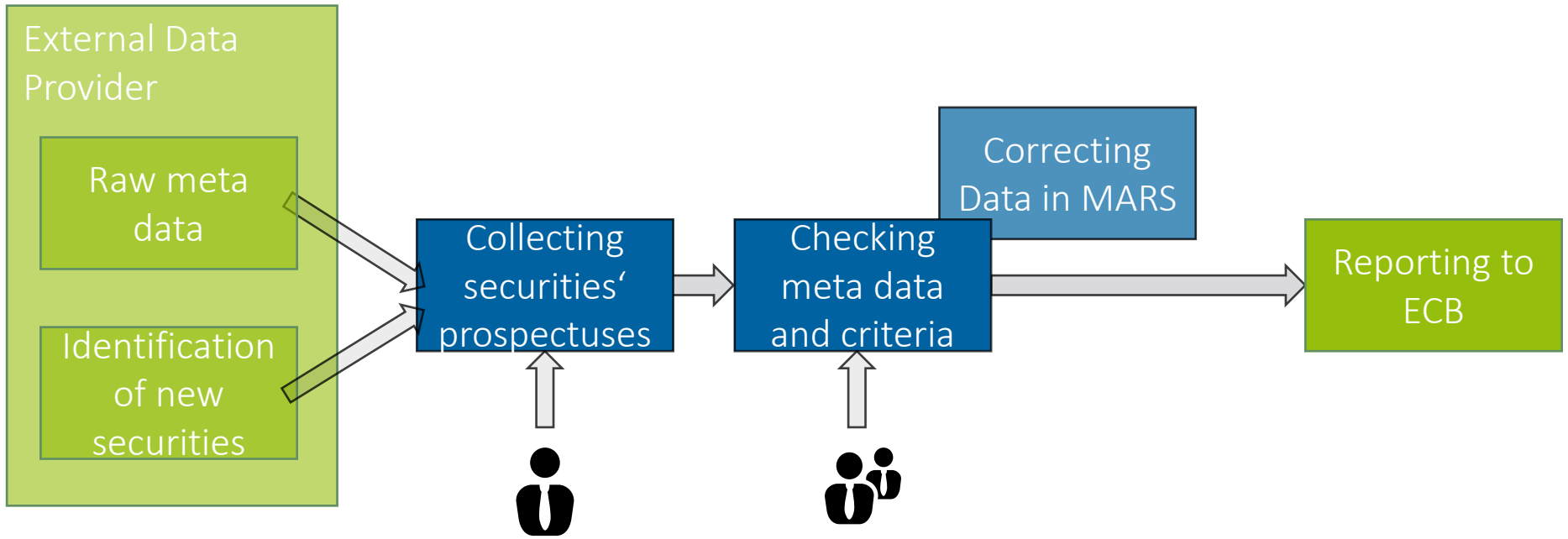
SCAI
Service & Community Center for Artificial Intelligence

- NCBs report daily new **eligible marketable assets** to ECB, which collects them into **EADB (Eligible Assets Database)**
- Checking a security / asset for eligibility is based on harmonized criteria (Guideline (EU) 2015/510)
- The reporting contains the **eligible assets** as well as related **meta data**

- Several eligibility criteria are established based on a security's prospectus
  - So far this is achieved by manually checking / reading the prospectuses in a four-eyes principle
- At Deutsche Bundesbank (BBk) the (BBk-made) application MARS is used for collecting the securities' data and reporting them to ECB

Status Quo Ante

Process Flow



Manual assessment is time-consuming and repetitive

DEUTSCHE BUNDESBANK EUROSYSTEM

SCAI Service & Community Center for Artificial Intelligence

Training a model

Proof of concept with a fine-tuned model

Research Project Automatic Annotation

Proof of Concept using NLP

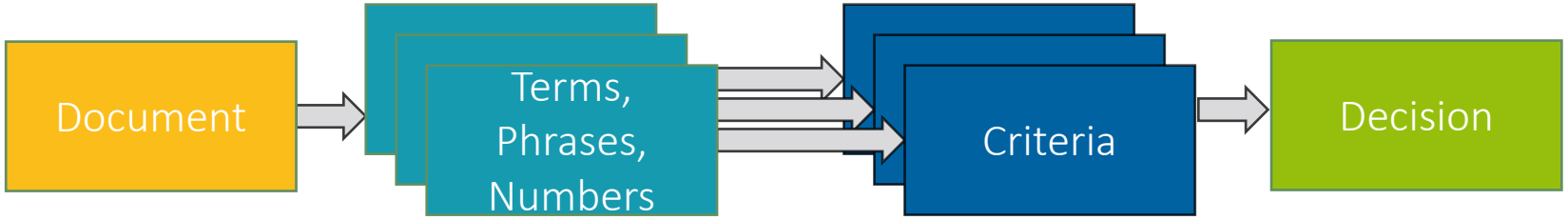
SCAI
Service & Community Center for Artificial Intelligence

§ 5
(Status)

status_nicht_nachrangig_bevorrechtigt

Die Schuldverschreibungen begründen nicht besicherte und nicht nachrangige Verbindlichkeiten der Emittentin. Bei Emission handelt es sich bei den Schuldverschreibungen um bevorrechtigte Schuldtitel **(Senior Preferred Schuldverschreibungen)**, die nicht den durch § 46f Absatz 5 in Verbindung mit Absatz 6 KWG gesetzlich bestimmten niedrigeren Rang haben.

- Task at hand: **Identifying in PDF-Documents a given number of terms, phrases, numbers etc. that form the basis for the decision**

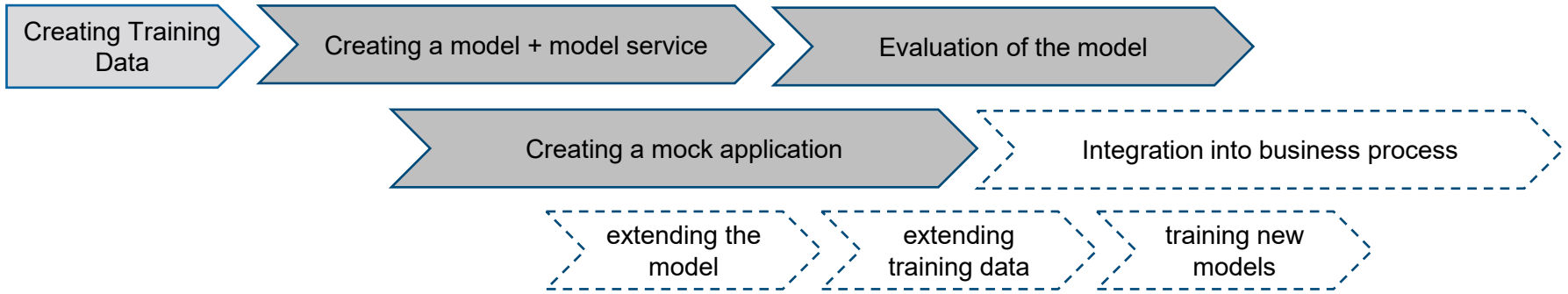


- In ML terms: **Multiclass/multilabel Classification Task (≈ 20 categories)**

Research Project Automatic Annotation

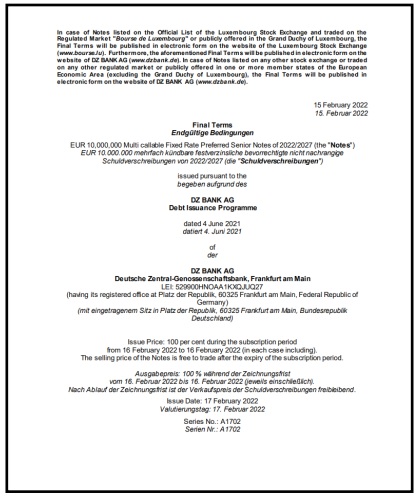
Starting Point

- At the start of the project (early 2022, pre „GPT breakthrough“):
  - No German-language domain specific (i.e. financial) language model available
  - Hence 2-Step modelling process:
    - (1) Fine-tuning a language model for German financial documents
    - (2) Training a multilabel classifier on top of the language model
  - No public dataset available -> creating training data is the first step



Creating Training Data

Data collection



Number of prospectus: 413
Issuing period: 2021 - 2022
Eligible documents: 369
Ineligible documents: 44
Training set: 272
Test set: 141 + 141

Data annotation



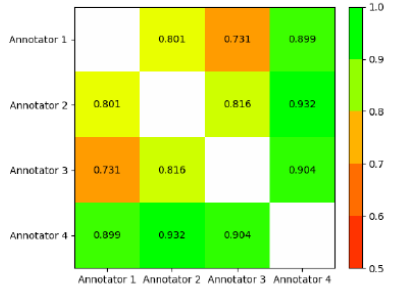
Annotation tool: Konfuzio
Annotation types: ~40

Disregarding of pages without annotations for training and validation purposes.

Annotation statistics

Target type	Train	Test
coupon fixed	431	375
coupon variable index	56	84
coupon variable margin	38	42
coupon variable operator	37	43
coupon variable tenor	45	75
currency	514	577
early redemption amount	64	52
early redemption	177	108
isin	421	417
principal amount	784	800
redemption at maturity amount	26	42
redemption at maturity	370	347
special termination	96	109
special termination amount	61	63
status non preferred	56	47
status senior non preferred	488	333
type of instrument	431	422

Inter-annotator agreement

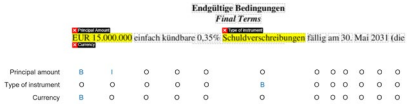


Average inter-annotator agreement scores

Test set was used to measure IAA. Therefore, every prospectus in the test set was annotated by a second analyst. 4 analysts seved as annotators in total.

Data preprocessing

1st step: Extraction of JSON-formatted raw data containing the annotations from the annotation tool
2nd step: Conversion and transformation of extracted data into dataset for token classification (BIO encoding)



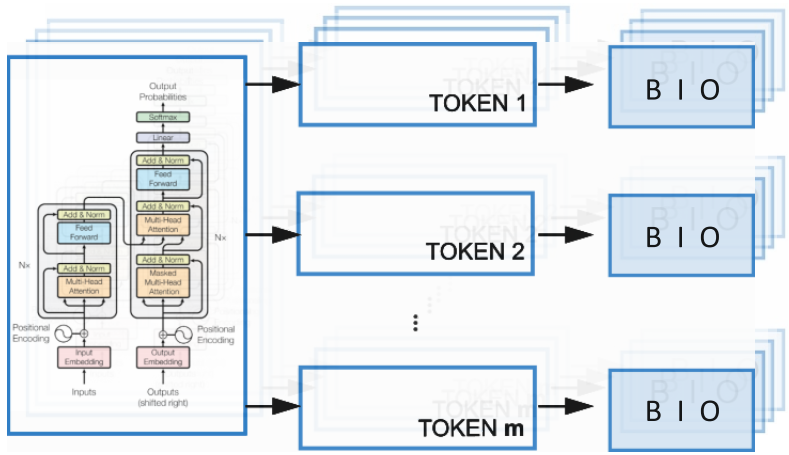
- Implementation of dataset classes using Hugging Face Datasets framework
- Challenge: overlapping text sequences belonging to different annotation types

Model Architecture / Document Processing

SCAI
Service & Community Center for Artificial Intelligence

1. Conversion PDF -> Text (including OCR)
2. Text processing and clean-up (e.g. extraction of German parts of bilingual docs, analysis of textboxes, …)
3. *Embedding (Text to vectors) using fine tuned language model*
4. *Labelling with multilabel classifier*
5. Decision based on deterministic rules (derived from EU Guideline)

3 + 4



DEUTSCHE BUNDESBANK EUROSYSTEM

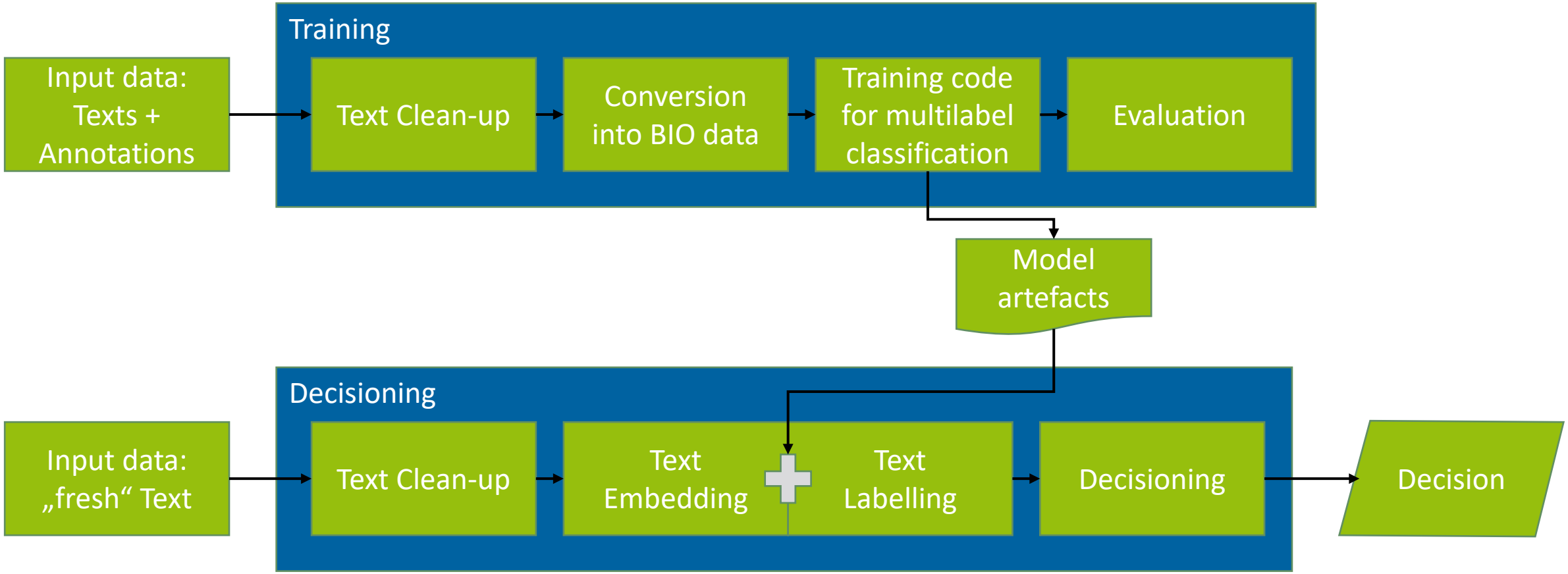
SCAI Service & Community Center for Artificial Intelligence

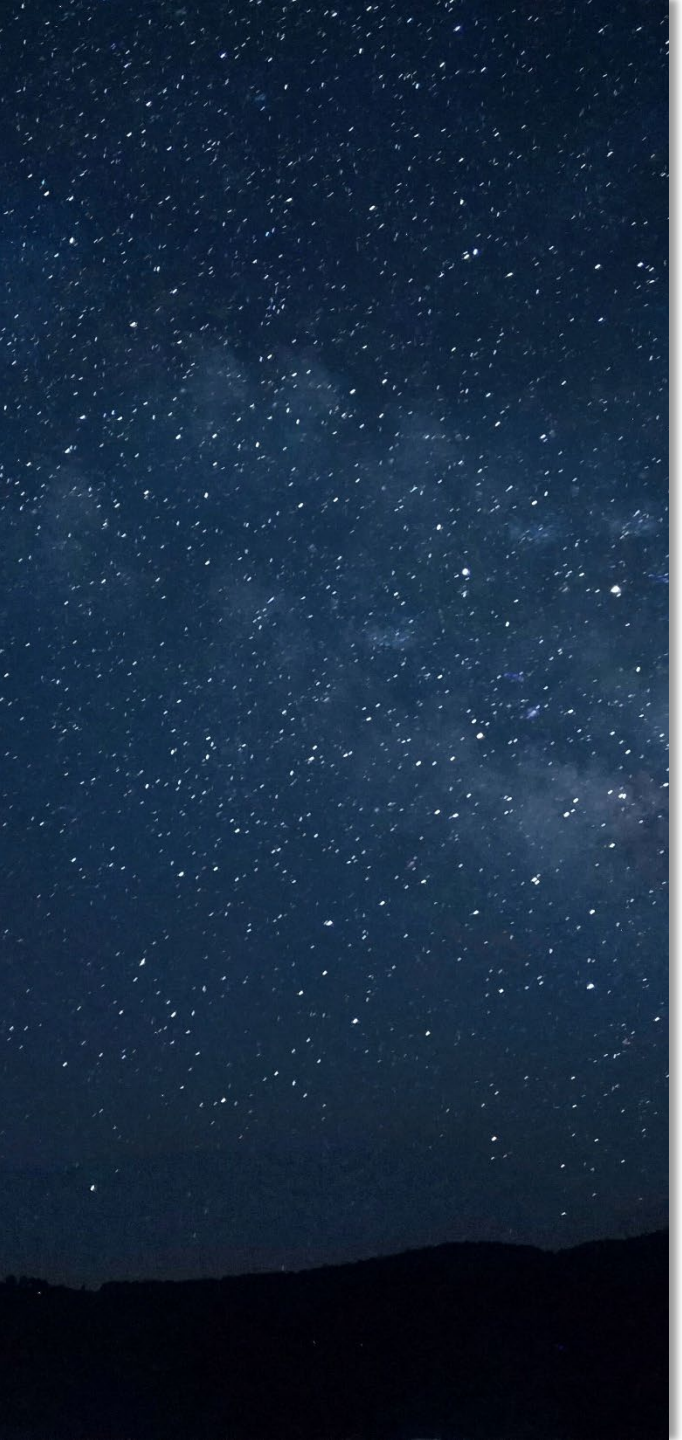
Integration

of the model into the business process

Operating the Model

Model Training and Decisioning





Integration into Business Process

Process Needs

- Given a document the experts needs:
  a. the decision of the model,
  b. the criteria causing that decision and
  c. (optimally) the relevant passages in the document (or relevant meta data) to
- check the validity of the ML decision.

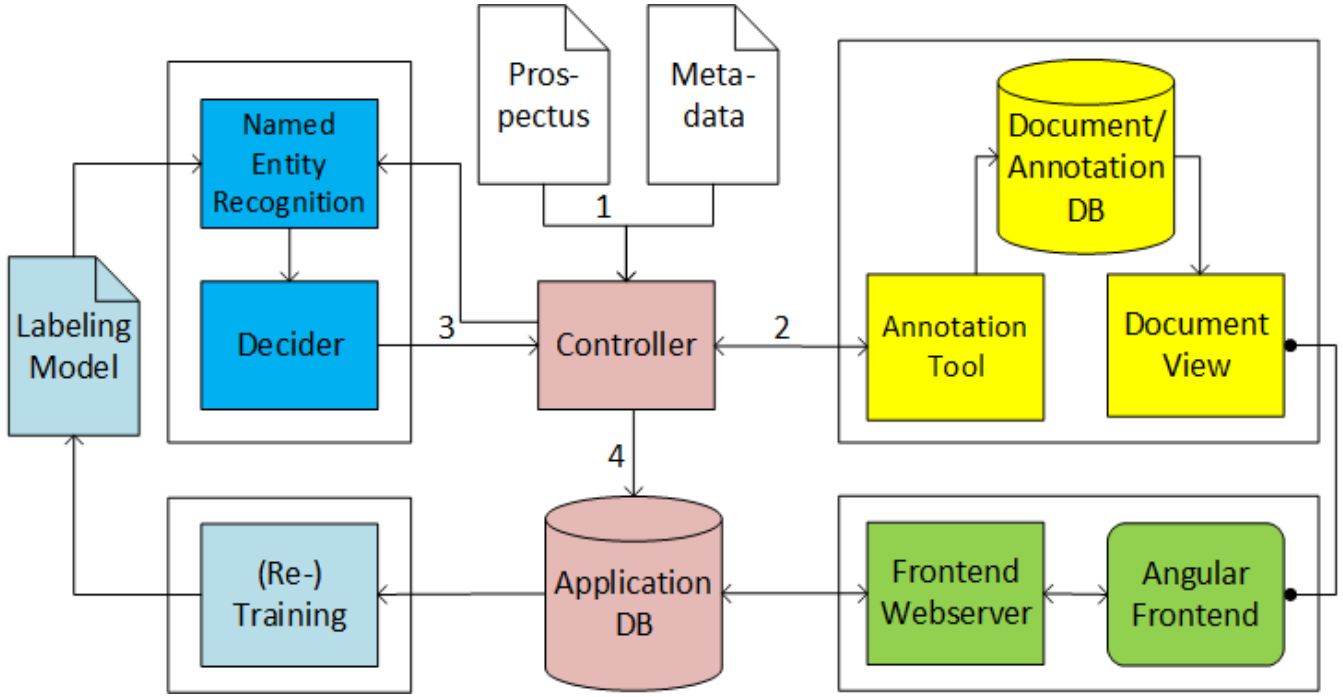
- If the model makes a mistake, the expert needs to **overwrite that decision** and
- (optimally) collect the data for future model improvements (retraining)

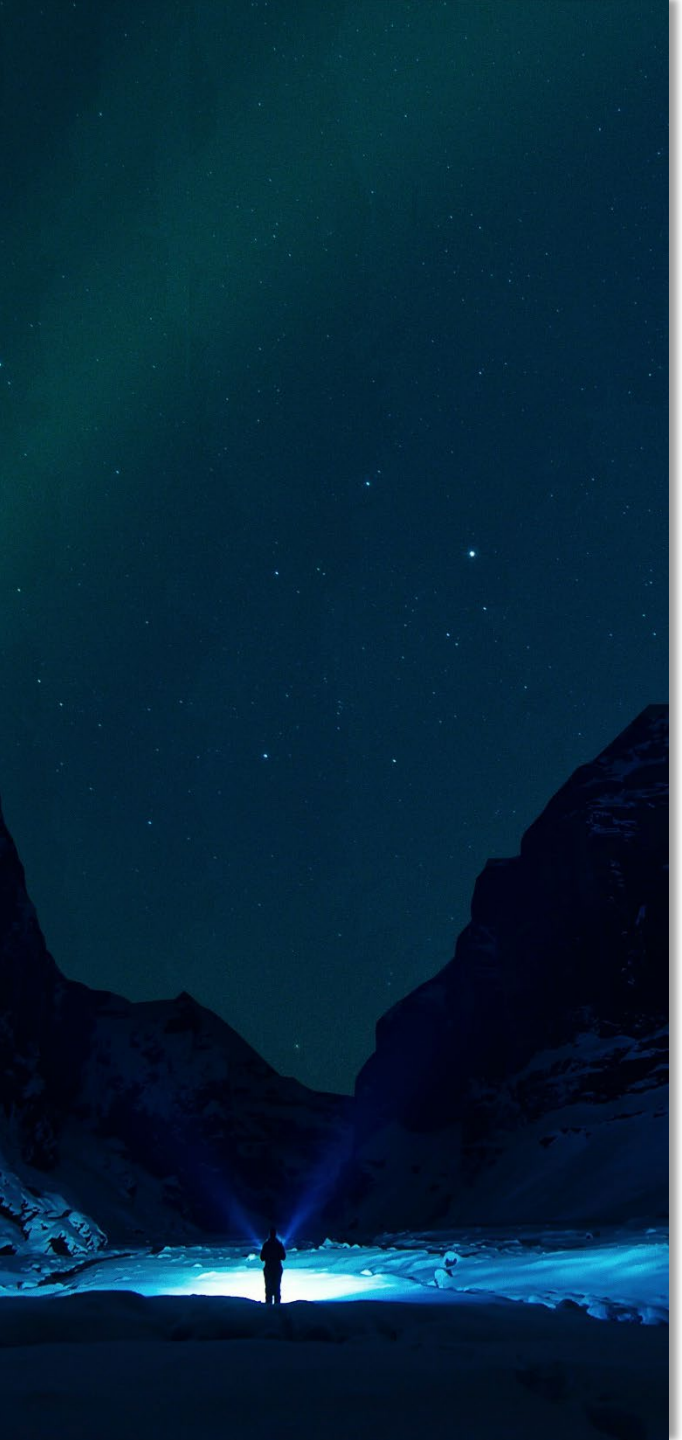
- If retraining is undertaken, we need both valid as well as invalid model decisions.

Integration into Business Process

Overview of Application Architecture

- Containerized application with communication via REST
- Integration into the actual business application (MARS) open as of yet





Implications of Using ML in the Process

SCAI
Service & Community Center for Artificial Intelligence

- Using an ML model can reduce processing time by
  replacing manual reading with reviewing found passages

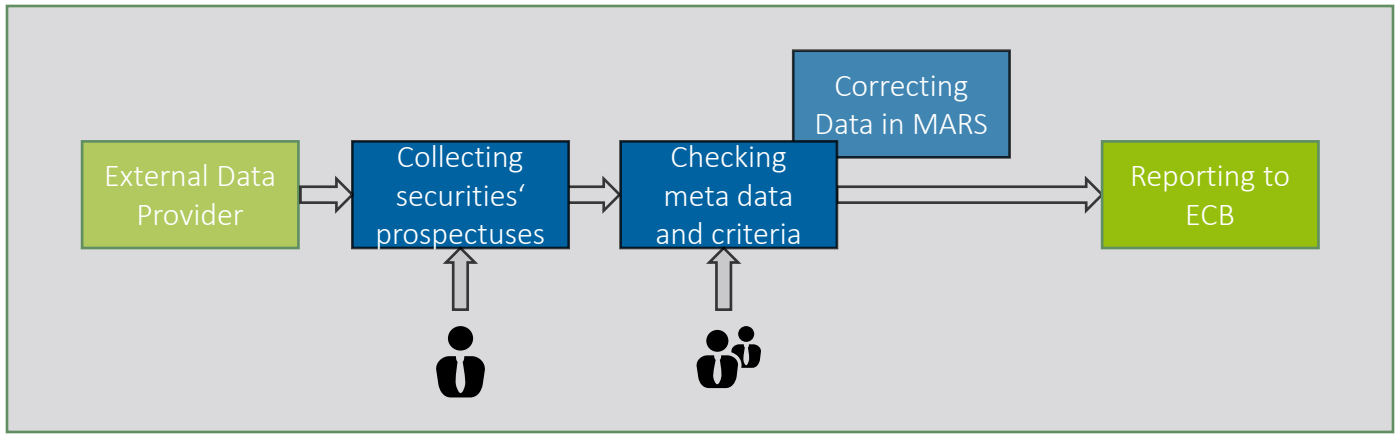
- An ML model will always have a chance for error, but
  the **accuracy can reach the Inter Annotator Agreement (IAA)** at the least

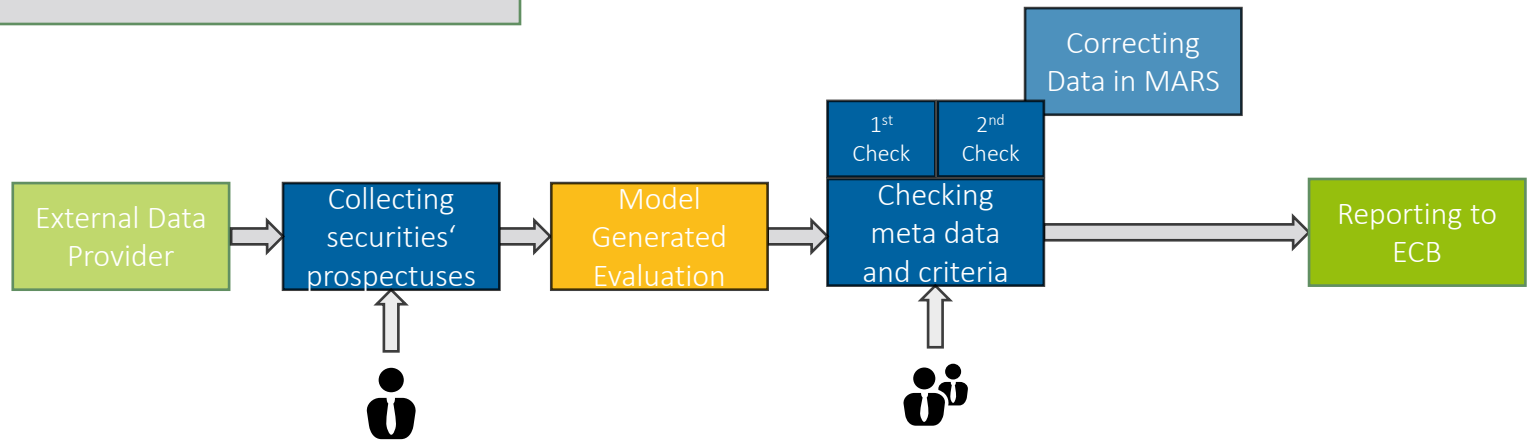
- Current legal environment requires a „**human in the loop**“
  - If model accuracy is (acceptably) high, the four eyes principle (as well as the review by two experts) could be replaced by a simple review
    (2 pairs of human eyes ⇒ „AI eyes“ + 1 pair of human eyes)

- Using an ML model will require:
  - **continuous monitoring** of model performance
  - **continuous improvement** of model mistakes and training data

Evolving the „4 Eyes Principle“

New Process Flow – Proof of Concept





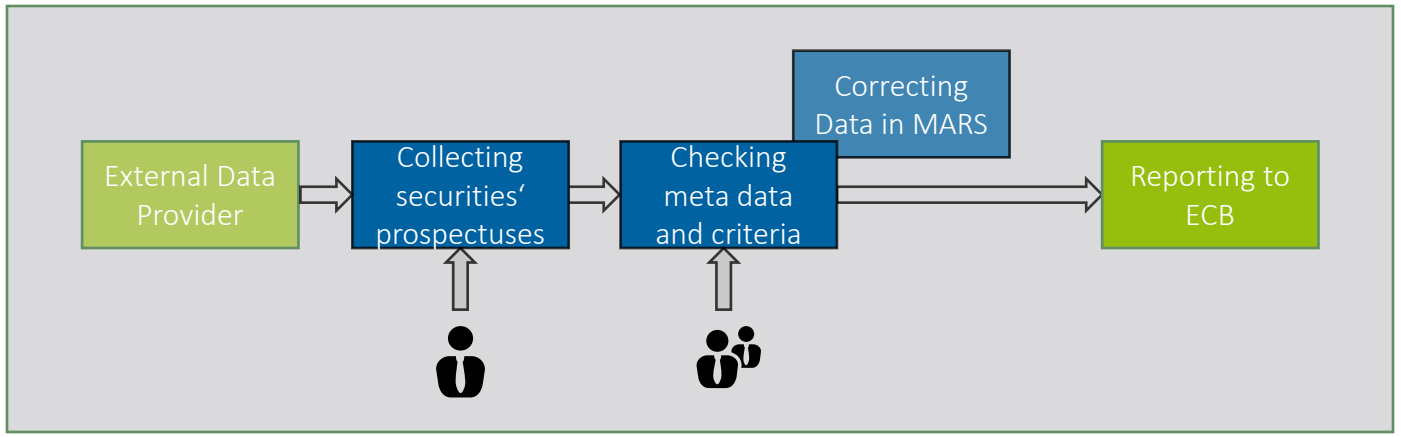
External Interface

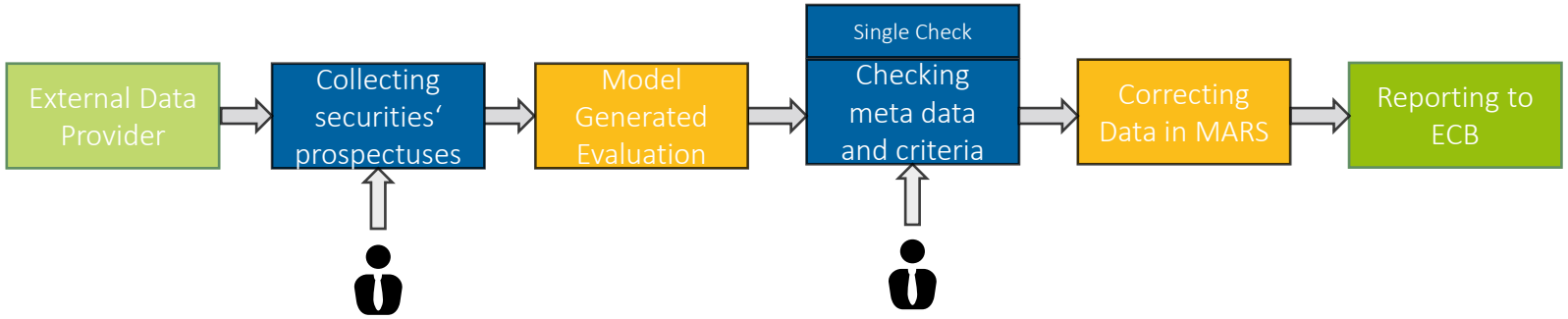
Human Interaction

Automatic Process

Evolving the „4 Eyes Principle“

New Process Flow – 1st Evolution





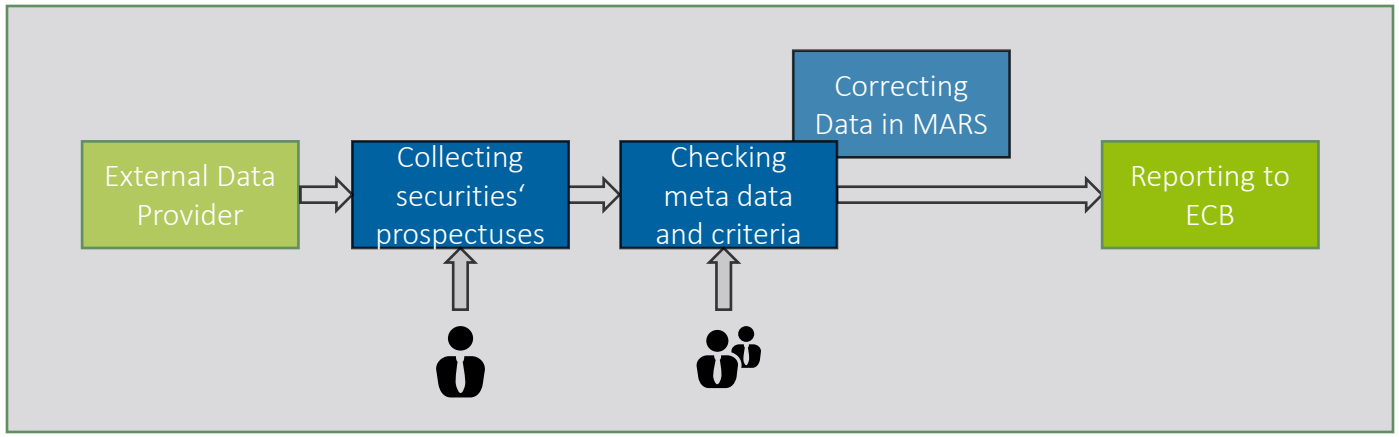
External Interface

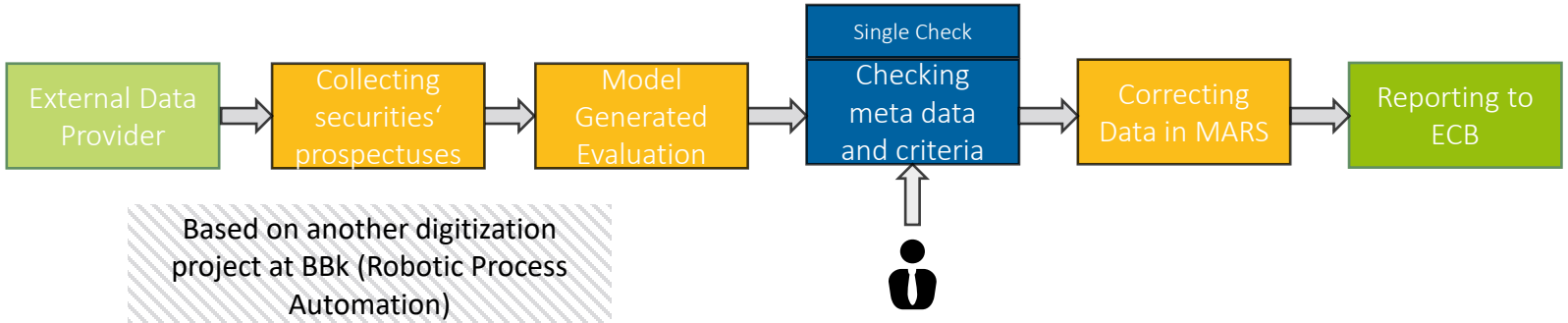
Human Interaction

Automatic Process

Evolving the „4 Eyes Principle“

New Process Flow – 2nd Evolution





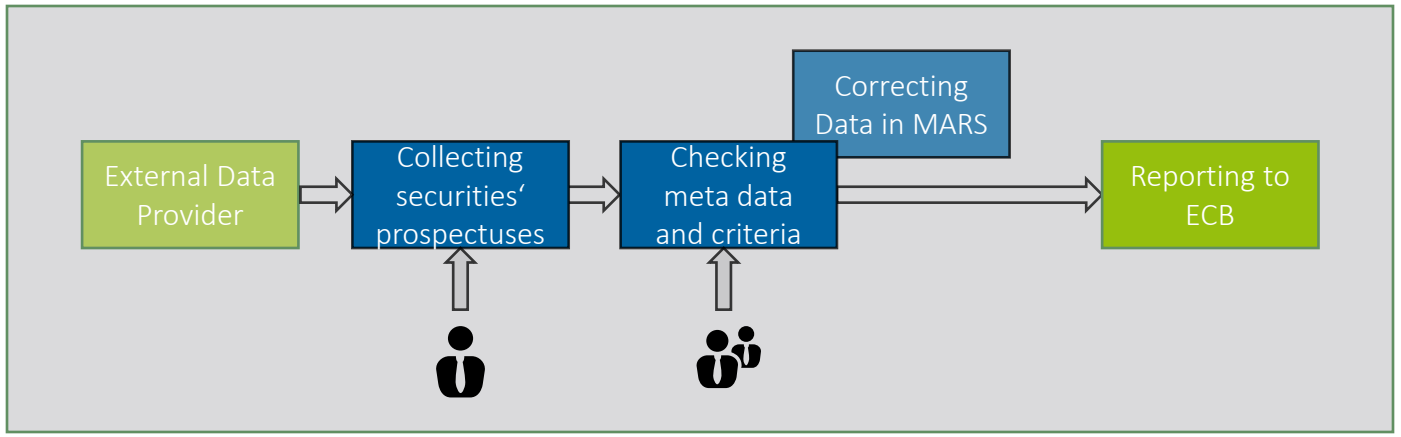
External Interface

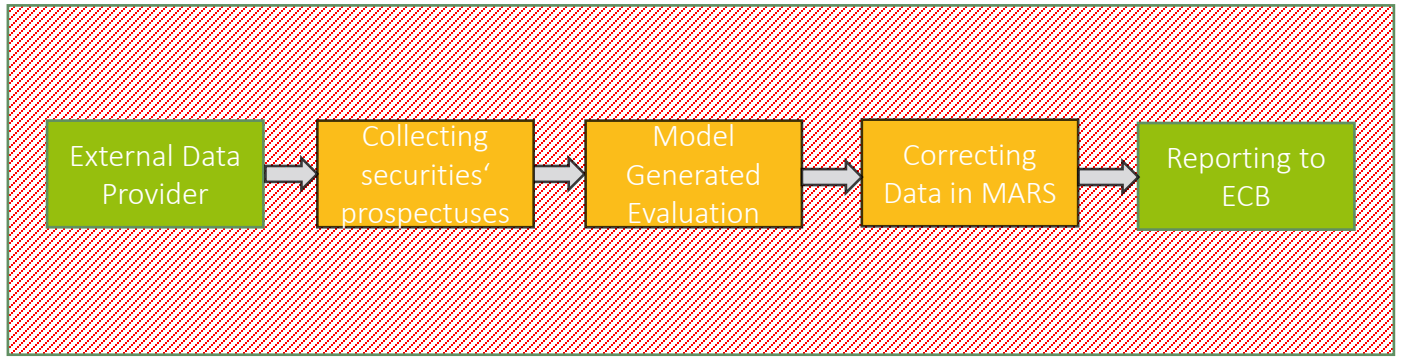
Human Interaction

Automatic Process

Evolving the „4 Eyes Principle“

Currently not Possible: Fully Automated Process – No Human in the Loop





External Interface

Human Interaction

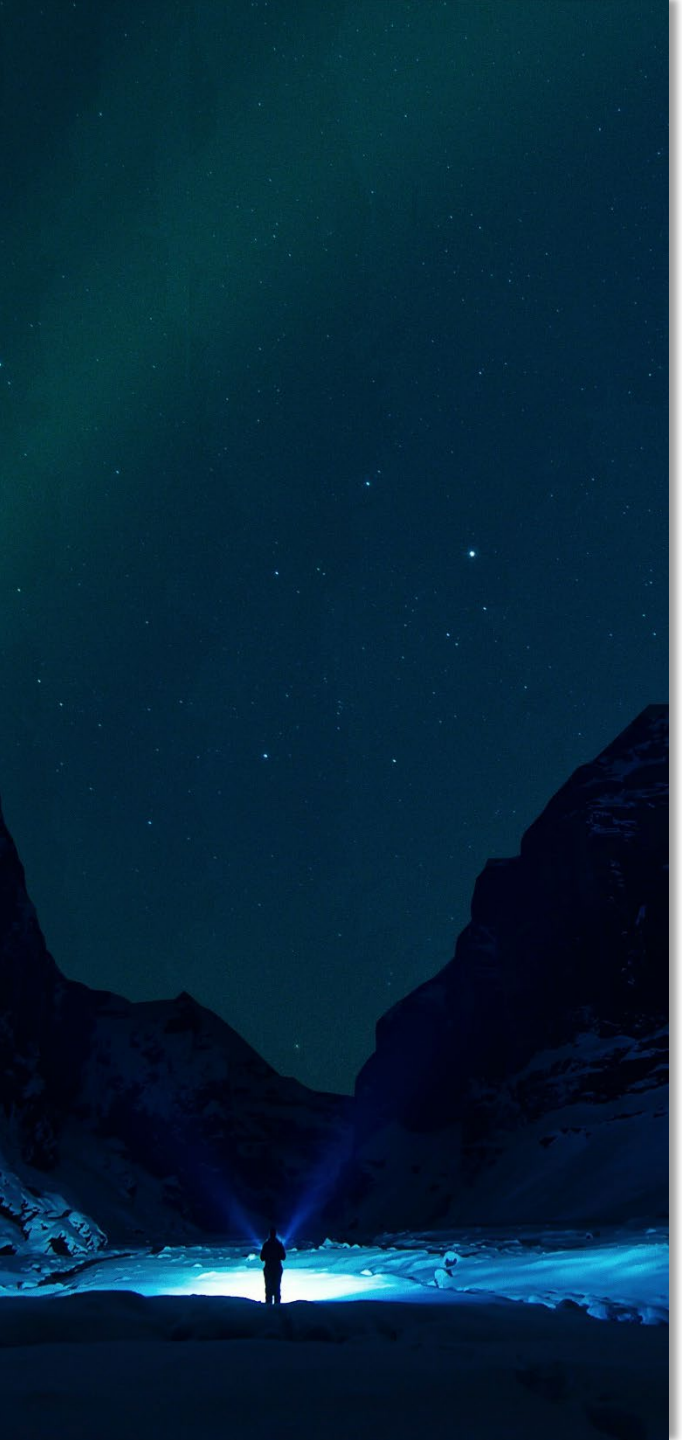
Automatic Process

DEUTSCHE BUNDESBANK EUROSYSTEM

SCAI Service & Community Center for Artificial Intelligence

Learnings

from the Process



Learnings from the Project

SCAI
Service & Community Center for Artificial Intelligence

- Creating training data is highly costly

- Understanding the business process is key
  - if only part of the process is automated, **the benefit may not outweigh the complexity**

- Building the necessary environment is highly complex
  - The codebase of the proof of concept easily reaches **10‘000 lines of code**

- **Integration into production is hard**, in particular if it necessitates new components, e.g.
  - Application for creating and storing text annotations
  - ML model monitoring and model archives (MLOps)
  - GPUs for model training



SCAI
Service & Community Center for Artificial Intelligence

Questions?

SCAI
Service and Community Center for Artifical Inteligence
scai@bundesbank.de

BANCA D'ITALIA
EUROSISTEMA

Data Science in Central Banking: Enhancing the access to and sharing of data

Closing remarks by Alessandra Perrazzelli
Deputy Governor of Bank of Italy

3rd IFC and Bank of Italy Workshop

Rome, 19 october 2023

Good afternoon, ladies and gentlemen,

I am delighted to see so many distinguished experts gathered here today to conclude this impressive workshop on data science in central banking jointly organized by the Bank of Italy and the Irving Fisher Committee of the Bank for International Settlements.

Over the past three days, we have embarked on an educational journey that has enriched our understanding of how data science is reshaping the landscape of the financial industry and of central banking in particular. The knowledge we have shared, the insights we have gained, and the connections we have forged are invaluable, and they will undoubtedly play a central role in the future of our institutions and of the financial sector as a whole.

Before I bid farewell to this event, I would like to take this opportunity to express my sincere gratitude to all of the participants, speakers, and organizers for their contributions.

Food for thought

Over the last three days, we have delved deep into the world of data science, exploring its applications, challenges, and the transformative potential it holds for central banking. We have learned about cutting-edge techniques in data analysis, machine learning (ML), and artificial intelligence (AI) that can be applied to monetary policy, risk management, financial stability, and many other areas that are crucial to central banks' mandates.

Data Science is revolutionizing the banking industry. The presentations of these past few days testify to the special and now important role of data science in crafting central banking operations. The Bank of Italy is committed to being at the forefront of data science innovation in the central banking community. We are constantly seeking better

ways to collect and harness new forms of data in order to improve our policy decisions and operational efficiency. In recent years, we have started our journey through Data Science by establishing a multidisciplinary team to consider the benefits and hidden risks of tackling the technological challenges of artificial intelligence and machine learning fuelled by the advances in big data, which continue to evolve at incredible speed. In our competitive world, we are always in search of innovation. The Bank of Italy has shown a knack for technological innovation, realizing its potential for growth and success. In 2021, as a great way to inspire our employees and promote creativity, collaboration, and out-of-the-box thinking, the Bank of Italy sponsored a hackathon, a worldwide competition on applying big data, natural language processing and artificial intelligence techniques to green and sustainable finance, that was jointly organized with the BIS Innovation Hub of Singapore under the Italian G20 Presidency. In 2020, we created Milano Hub, a technological hub supporting the development of innovation and the digital transformation of the Italian financial system. After the great success of the first two calls for proposals, we will launch a third one in the next months

In the last few years, we have redesigned our recruitment and hiring process by seeking out new skillsets such as Big Data, Machine Learning and Artificial Intelligence. This is just the beginning. We need to rethink our processes and foster the adoption and development of the right skills. To achieve these goals, we have sponsored special data science training programmes in some Italian universities. The new cohort of hires have already started their journey, which aims to create a flatter organizational structure.

In 2020, the total amount of data created, captured, copied, and consumed globally was around 100 zettabytes (an astounding value of 10^{23} bytes), and it is expected to rapidly increase, reaching 180 zettabytes in 2025.[1]

The mind-boggling amount of data that is now available to us, provided we are able to analyse it effectively, can give us a better picture of the economy at both the micro and the macro level. The Bank of Italy has constantly striven to be at the cutting edge in developing software and hardware platforms, enabling big data analytics[2] for statistical and economic applications.

One of the most significant takeaways from this workshop is recognizing that data is not just a resource: it is essential for effective decision-making in central banking. The quality, quantity, and timeliness of data can make all the difference in crafting policies that safeguard our economies and maintain financial stability. By harnessing the power of data science, central banks can enhance their capabilities, anticipate trends, and respond swiftly to emerging challenges.

---

1 Data taken from https://www.statista.com/statistics/871513/worldwide-data-created/ on September 25 2023.

2 See, for example, 'Big data processing: Is there a framework suitable for economists and statisticians?', 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 2804-2811, doi: 10.1109/BigData.2017.8258247. 'Weaving Enterprise Knowledge Graphs: The Case of Company Ownership Graphs.

Risks and Precautions in applying Data Science methodology

Extreme care is required when employing these new tools. Over the course of the workshop, a number of findings have been highlighted. The ethical dilemmas that institutions and practitioners must wrestle with in order to protect individuals from the consequences of blindly adopting AI-based solutions are numerous and definitely profound.

Let me quickly mention some of them.

First of all, data science relies on massive amounts of information that leverage large quantities of third-party data, and complex algorithms to generate answers. These methodologies are often opaque and unreliable, an issue that we must address in our work. We should always strive to pick and adopt solutions with sound and accurate explainability.

Moreover, the kind of big data we employ typically entails selection bias because of the peculiar characteristics of the population; increasing the number of observations does not reduce the sampling error unless corrected for. Data should be used in a fair and unbiased way, avoiding discrimination or harm to individuals or groups.

Second, safety and security: data and systems should be secure to protect against unauthorized access, misuse or abuse. Human oversight is also essential in the development and use of data and new technologies.

Third, the availability of much more detailed personal data increases the importance of its integrity, confidentiality and privacy. Providing the right is central to preserving the democratic values of our societies. Furthermore, individuals should exercise full sovereignty over their own data and be able to give or deny consent to their collection and use. This has prompted the development of regulations concerning the treatment of digital data (think of the level of data protection granted under the GDPR in the EU or the CCPA in California).

A few years ago, the Bank of Italy started using advanced data security tools to process individual wage data for analysing the gender gap (see 'Women, labour markets and economic growth' by F. Carta, M. De Philippis, L. Rizzica and E. Viviano, Banca d'Italia, 2023), a topic recently recognized by the awarding of the 2023 Noble Prize in Economics to Prof. Claudia Goldin 'for having advanced our understanding of women's labour market outcomes'.

Technologies for processing personal data privately are already available, such as differential privacy, developed in 2006 by Microsoft researchers[3], or fully Homomorphic Encryption, in 2009, which has been touted as a possible solution for processing individual data.[4] The Bank of Italy has begun experimenting with these technologies.

---

3 See 'Calibrating noise to sensitivity in private data analysis**',** Cynthia Dwork Frank McSherry, Kobbi Nissim, 2006.

4 See 'A Fully Homomorphic Encryption Scheme', Craig Gentry, 2009.

It is important to note that there is no one-size-fits-all approach to data ethics. The specific ethical considerations will vary depending on the type of data being collected and used, the purpose for which it is being used, and the potential impact on individuals and society.

Looking Ahead

As we are nearing the conclusion of this workshop, let us bear in mind that our journey with data science in central banking has just begun. The knowledge and skills we have acquired here are tools that we can wield to drive innovation and excellence in our institutions. It is our responsibility to apply what we've learned, adapt to the ever-evolving landscape of technology and data, and lead the way in shaping the future of central banking.

I encourage all of you to stay connected, collaborate, and continue the dialogue beyond this workshop. The relationships we have built here can be the foundation for future partnerships and collaborations that will drive progress in our field. The Bank of Italy is firmly committed to carrying out further research on these issues and cooperating with other institutions to strengthen methodologies and applications.

Finally, let us not forget the significance of our work. Central banking plays a key role in the stability and prosperity of our countries. Data science can help us navigate the complex waters of modern finance. Together, we can steer our institutions toward greater resilience, transparency, and effectiveness.

We have had the privilege of learning from distinguished experts and practitioners in the field, who have shared their knowledge, experiences, and insights. These interactions have enriched our perspectives and deepened our understanding of the intricate relationship between data science and central banking. I would like to extend my heartfelt gratitude to all the speakers, instructors, and organizers who have made this workshop a resounding success.

Thank you all for your dedication, your commitment, and your thirst for knowledge. It has been an honour and a privilege for me to be a part of this workshop with you. I wish you all continued success in your endeavours, and may the insights gained here guide us toward a brighter future in our profession.

Designed by the Printing and Publishing Division of the Bank of Italy