Irving Fisher Committee on Central Bank Statistics

BIS

Data science in central banking: applications and tools

Douglas Araujo, Giuseppe Bruno, Juri Marcucci, Rafael Schmidt, Bruno Tissot[1]

Executive summary

The Irving Fisher Committee on Central Bank Statistics (IFC) periodically organises workshops on "Data science in central banking" with a diverse audience of practitioners and technicians. The most recent one took place in 2022 and focused on **the broad spectrum of data science applications/tools used in central banks**.

The concept of data science refers to the study of data and therefore includes the various techniques for extracting insights from them. **Yet data science is fundamentally different from traditional data analysis, as it typically applies to large, complex and/or unstructured information sets.**

A key factor supporting the development of central banks' data science projects in recent years has been **the sheer volume and complexity of financial data available** in today's societies. This requires more sophisticated techniques for data management and analysis, a trend reinforced by the new opportunities opened up by artificial intelligence (AI) and machine learning (ML). 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. Additionally, advances in statistical computing infrastructure and enhanced training have enhanced the data skills of official sector staff.

These elements have accelerated efforts to advance data science, helping central banks to quickly adapt to the swiftly evolving financial landscape. In this endeavour, 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.**

From the outset, IT has absorbed a great deal of attention and resources. Central banks are increasingly aware that a modern IT architecture is crucial in reliably

[1] Respectively, Economist, Monetary and Economic Department, Bank for International Settlements (BIS) (Douglas.Araujo@bis.org); Director, Economics and Statistics Directorate, Bank of Italy (Giuseppe.Bruno@bancaditalia.it); Economist, Economics and Statistics Directorate, Bank of Italy (Juri.Marcucci@bancaditalia.it); Head of IT, Monetary and Economics Department, BIS (Rafael.Schmidt@bis.org); and Head of Statistics and Research Support, BIS, and Head of the Secretariat of the Irving Fischer Committee on Central Bank Statistics (IFC) (Bruno.Tissot@bis.org).

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

We thank Olivier Sirello for helpful comments and suggestions.

and securely dealing with data. A key objective 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. The emphasis is on adopting advanced IT tools and engineering practices – including cloud computing, software containers, automation tools, and continuous software integration and delivery pipelines. In particular, there has been a growing interest among central banks on using and producing software that can be shared as open source, either with their peers or with the general public. Such an open source software (OSS) strategy can be instrumental for honing their own IT development and strengthening their data science capabilities.

Once the IT infrastructure is able to support the development and deployment of data science applications, **the focus is on performing the various mathematical and statistical operations that are needed to deal with the raw data**. Data scientists need not only to access very large and complex information sets, but also to compile statistics via multiple sequential tasks (signal extraction, quality management, dissemination) before using them to extract relevant insights. Many different AI techniques can be used for this purpose, including for conducting textual analysis, reflecting the increasing opportunities offered by natural language processing (NLP) tools and large language models (LLMs).

A third lesson is that data science projects require a good understanding of the business cases and therefore a close cooperation with subject-matter experts. One obvious reason is that economic indicators such as GDP are more than just numbers: analysing them calls for an understanding of the way the statistics have been compiled as well as the complex factors that drive them – say, fiscal policy or geopolitical tensions – and their real-world implications. Moreover, this expertise is essential to support informed policy decisions: in particular, 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. This is even more important with data science applications that may need to be adapted when used to answer economic questions, for instance when analysing causal relationships. Finally, business area expertise can help central banks prioritise effectively between concurrent data science initiatives especially in view of resources constraints.

1. Introduction: data science to support central bank operations

The scope of data science

Central banks have been actively reviewing the ongoing adoption of data science as well as developments in the big data ecosystem in recent years (IFC (2017)). A number of projects have been launched on a pilot basis, of which some have already become permanent processes supporting current operations. In this context, central banks have realised the importance of sharing experience, not least to develop in-house knowledge and reduce reliance on external services providers.

To support these initiatives, the IFC has organised recurrent workshops on "Data science in central banking" with a broad audience of practitioners and technicians.[2] The most recent one, organised on a virtual basis with the Bank of Italy at the BIS in 2022, focused on the **broad spectrum of data science applications/tools used in central banks**. Several hundred participants representing more than 120 institutions took this opportunity to present and discuss novel data science applications of particular relevance that are under development or already in production.

But what is data science? The concept **refers to the study of data, and therefore includes the various techniques for extracting insights from them**. Yet in practice its scope is elusive and continuously evolving. As a starting point, data science should cover all the features related to *data analysis,* including the "procedures for analysing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data […], and all the machinery and results […] which apply to analysing data" (Tukey (1962)). From this perspective, it requires the performance of multiple tasks, such as data-gathering, preparation and exploration; data representation and transformation; computing with data; visualisation and presentation; data modelling; and using data to study science itself (Donoho (2017)). The goal of data science is therefore to turn *data analysis* into *actionable knowledge*, whether that means better decision-making, identifying new patterns and trends, optimising processes, or even creating new methods of data-driven research.

Data science is, however, fundamentally different from traditional analysis as it typically applies to large or complex data, including different types of unstructured information such as text (Bholat (2020)). In fact, as argued by Jörg Osterrieder in his keynote address, the AI "revolution" may not be as new as people usually think.[3] But what is new is the current combination of greater computing power, the increased availability of big data sources and advanced analytical techniques, and the multiple business applications of interest for which these data and tools can bring useful insights.

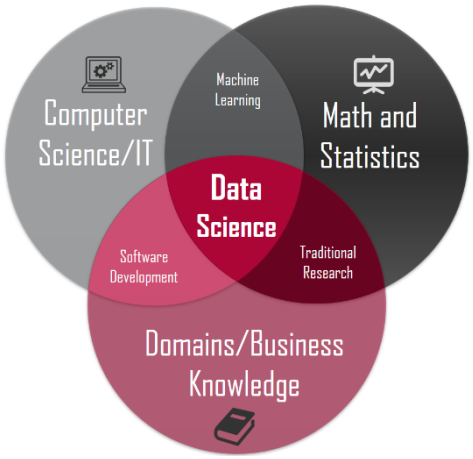
These considerations mean that the range of methods potentially covered by data science is considerable and requires diverse fields of expertise. As a result, an effective approach to data science should encompass the large variety of relevant business cases as well as multiple types of data. Data science could thus be usefully defined as an "*interdisciplinary field* that uses scientific method, processes, algorithms and systems to extract knowledge and insights from data in various forms" (Walczak (2019)). In view of these interrelationships, the multifaceted role of data scientists would appear to sit at the intersection of **information technology; mathematical and statistical methods; and business, or "subject-matter" expertise** (Graph 1).

[2] The material from the last workshops can be found at www.bis.org/ifc/events.htm.

[3] The term was first coined in the academia during the 1950s (Moor (2006)).

What is data science?

Graph 1



Source: N Cheng, "Data analyst primer: the essential guide", Medium, 27 Aug 2020, after D Conway, "The data science Venn diagram", 30 September 2010.

Data science in central banking

The approach to data science outlined above would seem to be well suited for central banks, whose activities range from statistical production to economic analysis, monitoring and policymaking. Recent projects have in fact involved the three dimensions at the intersection of data science. First, significant investment in software engineering tools, especially to support big data analytics, has improved the underlying IT and data infrastructure. Second, more diverse and mature econometric and data visualisation tools have been deployed to leverage mathematical/statistical knowledge and extract more actionable and timely insights from data. And third, deep subject-matter expertise in central banks (eg on economic and financial information) has ensured that the business needs and related priorities are fully understood.

A number of factors have supported these developments. First is the **skill set that is increasingly available internally**. Reflecting the variety of the tasks they perform, central banks have at their disposal a large variety of staff competencies, a key condition for ensuring the involvement of multidisciplinary teams and making the most of the new data landscape.[4] Moreover, while central banks have traditionally had a large number of statisticians and subject-matter experts in their ranks, their staff have become increasingly proficient in IT tools and techniques in recent years. The upskilling of internal staff, combined with fresh hires with new IT abilities, has helped the adoption of best practices from computer science fields, in particular software engineering. In addition, the methodology of prototyping tools in business areas, experimenting quickly and then developing them over time as their usage

[4] See the Banco de Portugal's aim to have specialised staff, represented by different "colours", comprising a blend of business and technology skills and working together in composite groups of "purple people", ie specialised staff whose skills overlap (Teles Dias (2021)).

matures has proved a pragmatic approach to further increasing data science usage in existing teams (Araujo et al (2022)).

A second supportive factor has been the **decisive action of central banks to enhance the underlying global statistical infrastructure** to make it more flexible and efficient, for instance by developing global registers and identifiers and by promoting data-sharing and access to new sources of information. One particularly important dimension in supporting data-based applications in this context has been the development of global initiatives to standardise statistical information, eg the Statistical Data and Metadata eXchange (SDMX; IFC (2016)) or the Legal Entity Identifier (LEI). Such standardisation initiatives provide a common language to deal with the data of interest, on top of which more comprehensive applications can be developed. And, in fact, they can be instrumental for setting up advanced data architectures. For instance, the SDMX information model allows organisations to automate data processing and collection, based on a metadata-driven approach that strengthens their data quality management – a valuable benefit for central banks that are key producers of economic and financial statistics.

A third factor has been the **general push for developing innovation in central banking** to better address the rapidly changing financial landscape, as observed for instance by the creation of the BIS Innovation Hub.[5] This research initiative aims to develop public goods in the technology space to support central banks and improve the functioning of the financial system. These goals are realised in the form of specific projects, for instance Projects Rio and Ellipse, which, respectively, facilitate market monitoring and the development of early warning indicators based on big data analytics. More generally, innovation has proved essential to support central banks not only as statistical producers but also as users of the data needed for conducting their activities, depending on their mandates and specific requirements.

Drawing on the various central bank contributions included in this IFC Bulletin, this overview sheds light on the **three main dimensions at the intersection of data science**. Section 2 reviews the IT infrastructure elements that can support data science projects in central banks, such as the setup of a data platform, the provision of the software and tools for data management, and the new opportunities offered by modern IT tools and engineering practices, including cloud computing and the use of software containers. Section 3 discusses the role played by mathematical and statistical techniques and access to new information sources that allow central banks to extract insights from more data, including unstructured ones eg text. Lastly, Section 4 emphasises the importance of leveraging subject-matter expertise to support data science use cases in the economic and financial sphere.

2. Modern IT architectures

The first main area at the intersection of the multidisciplinary data science concept relates to IT. Central banks, as well as other public institutions and international organisations, are increasingly aware of the **importance of having a modern IT architecture** to deal with the data they need to fulfil their missions. This reflects the rapid pace of technological advance and the mission-critical nature of many of their data-intensive processes.

[5] For more on the BIS Innovation Hub's projects, see www.bis.org/about/bisih/projects.htm.

The IT architecture has to cover the **various processes, software and hardware that underpin the multiple activities when dealing with data** – eg sourcing, collecting, storing, managing, analysing, sharing and making data available to the appropriate users. These "plumbing elements" have to be linked together in a way that is comprehensive and consistent with the organisational priorities that guide day-to-day activities and investments (OECD (2019)). Ideally, this integration would be organised in the context of a dedicated data governance framework supporting the institution's overall strategy (Križman and Tissot (2022)). Yet the implementation of new IT platforms and software stacks can be complex, requiring considerable investment in terms of resources and time.

While no single solution can fit all organisations and use cases, **a number of useful insights can be derived from the recent projects implemented in the central banking community**. In particular, three points deserve to be highlighted. First, a modern data platform is essential for central banks that want to fully tap the potential from the information available. This means for example that, to be efficiently organised, data should be put in a common place available to all potential users subject to access rules. Second, the IT environment should provide all the IT tools and processes needed to manage data consistently and efficiently, and in a user-friendly way. Third, there is a premium on keeping this infrastructure up to date to allow the continuous deployment of novel applications and tools, which usually require a large quantity of information to be reliably sourced and delivered. One way to deal with this issue is to move some IT operations to a cloud environment, an option that is increasingly under consideration in the central banking community.

A common data platform?

A first lesson is that there can be **value in implementing a unique and powerful platform** to manage all the different types of information (including big data sets) of interest to the organisation. Such multi-tenant data platforms, which may be centralised (enterprise data lake) or decentralised (data mesh architecture), can bring several benefits, especially in terms of the ease of governance and economies of scope and scale offered by a common infrastructure. They can be instrumental when dealing with multiple business areas, as is often the case for financial stability (macroprudential) analyses, which typically draw on multiple information sources. It thus puts a premium on developing an institution-wide, unified data model – comprising for instance a set of common identifiers for financial institutions and instruments, consistent definitions and a comprehensive metadata schema.

A second lesson is that the **platform should be flexible enough to manage all the various types of data** of interest. For instance, a key requirement for the system for data collection, transformation and analyses implemented at the European Union's Single Resolution Board (ESRB) was to be able to deal with both large, highly structured data (eg granular supervisory reporting from financial institutions) and unstructured information collected from narrative texts. This reflects the need to use advanced analytical tools for checking banks' crisis resolution plans (ResTech; Loiacono and Rulli (2022)). Given the heterogeneity of the information at stake, quality was ensured by designing the data ecosystem on unified standards such as the LEI and the eXtensible Business Reporting Language (XBRL) global framework for exchanging business information. Similarly, the various processes followed by the BIS for the collection, production and dissemination of its own statistics – eg for

validating, transforming and mapping different types of macro as well as micro data sets – are based on a (SDMX) metadata-driven environment.

Experience shows that there are significant **challenges** when implementing an institutional data platform, namely the difficulty of dealing with diverse technological landscape across business units; their varied requirements, for instance as regards the trade-offs faced in terms of agility/robustness or security/innovation; and different organisational setups – with the key issue of adequately balancing the centralised governance of the platform against the provision of sufficient user freedom to design their (evolving) IT requirements.

A third lesson is to **offer sufficient IT self-service capacities to the business units**, so that they are adequately empowered with the resources to manage their data projects on their own rather through a centralised point. In fact, the development of the BIS centralised multi-tenant platform was accompanied by the creation of an analytical lab to facilitate the launch and maintenance of the various data-based initiatives. Certainly, one issue was the time and close collaboration required by the upskilling of the staff located in different units. But the implementation of self-service capacities proved to be a catalyst for innovation, stimulating the efficient design and delivery of the business areas' projects. Self-service was also a key element of the Banco de Portugal's information strategy to become a more data-driven institution. Under its advanced analytics pillar, a data science lab was set up to offer to the various units tools for code versioning, dedicated storage, grid computing and multiple coding languages used in data science and econometrics. These facilities let statisticians develop automatic quality control checks based on traditional techniques as well as ML approaches to identify erroneous reports in the credit registry. The central bank's strategy was to promote the automation and standardisation of the new projects to follow good practices in software development.

Making available efficient IT tools and processes

Modern IT architectures offer two key benefits to central banks. The first is to provide access to better-quality data, delivered more promptly to staff with the appropriate permissions to use them. The second is to **make available to each business unit the necessary tools to deal with this information**. The obvious reason is that data science work requires the use of advanced analytical tools and the necessary software so that users can (i) deal with a wide range of data types; (ii) have access to large spectrum of functionalities; and (iii) rely on efficient operational processes to save time and resources (Wibisono et al (2019)).

As regards the first aspect, the aim of a modern IT platform is to **facilitate users' access to a large and diverse range of data types**, ie structured or unstructured, coming from different sources (eg commercial vendors, reporting entities) and with different formats (eg generic spreadsheets "pushed" by reporters, data "pulled" from websites). Certainly, such diversity can raise practical difficulties, which are reinforced by the increasing demand for very granular information observed in recent years. In addition, the conditions for sharing/accessing these new data sets may require important legal and security work, for instance to deal with confidentiality and licensing issues.

With respect to IT aspects more specifically, the continuous **access to new information sources calls for setting up agile, structured and scalable data ingestion processes**. For instance, the BIS has developed a microdata ingestion utility

that allows the structured loading of millions of data points from a variety of commercial and public data providers. This utility usually requires minimal configuration to import a new data set, ensuring a fast turnaround for users. It has been in production since 2020, leveraging open source tools such as Python/Pandas and SQLAlchemy to download, parse and store data (files) on an SQL server. This tool and underlying software are available for interested central banks as part of the BIS's promotion of international software collaboration. Similarly, the ECB has developed a framework, based on Apache Spark distributed computing and other open source tools, to integrate granular banking data from multiple sources such as the European credit registry (Anacredit), the European Market Infrastructure Regulation (EMIR), money market statistical reporting (MMSR), the Central Securities Database (CSDB), and the securities holdings statistics (SHS). The aim is to analyse in a comprehensive and timely way the systemic importance of European banking groups by simultaneously considering different financial instruments and the related interconnections.

Second, making available more diverse and well managed information is not enough since business area users need sufficient capabilities to deal with it. This calls for **having a sufficiently varied palette of IT software, advanced analytical tools and accessible programming languages**. To address this demand, the free open-source library *gingado* has been created at the BIS to facilitate the internal use of ML in economic and finance use cases (Araujo (2023)). This package uses the SDMX standard to help users find and obtain high-quality data to augment their particular data set; provides convenient functions that train benchmark ML models; and promotes proper modelling documentation. And because this library is written in Python, it can also be easily used in other programming environments based in R, Stata and others. Relatedly, central banks' analytical capabilities can be improved more generally by both the use and production of OSS (see Box 1).

Box 1

Central banks as users and providers of open-source software

Douglas Araujo, Stratos Nikoloutsos, Rafael Schmidt, Olivier Sirello

In a world increasingly shaped by collaborative creativity, central banks are engaging more intensively with open-source software (OSS), ie software whose source code can be freely edited, reproduced and redistributed.① As software users, they may extensively rely on OSS, including ML tools, to perform their operations.② As providers, they can share publicly or with other central banks their source codes, such as macroeconomic models or tools for data dissemination. This box discusses how central banks can exploit the value added of OSS.

As a starting point, OSS can be defined against its opposite, "closed-source" or proprietary software whose source code is not disclosed. Private companies traditionally choose closed-source models, which can be the most straightforward option, especially if the source code they develop is a trade secret. Yet many firms, including big techs, have been increasingly open-sourcing the tools developed in-house, and a number of newer firms have even built their entire business models around OSS. The private sector experience shows that there is no single correct choice; each model has its own advantages and disadvantages.

As regards public sector entities, including central banks, they tend not to disclose their source code. Apart from the need to protect intellectual property rights, a key reason is privacy protection and confidentiality settings, since some codes might contain sensitive information (eg insights into policy). Hence, even the central banks that are most active in OSS generally

prefer to keep some of their source codes confidential. Still, a number of them see value in gradually open sourcing their software and are taking gradual steps in this direction.

Compared with the closed-source model, OSS can bring four main benefits to central banks, in terms of costs, customisation, security and usage base:

- First, from a financial perspective, OSS is virtually always free of charge, unlike most closed-source software. This is an important feature for central banks wishing to control costs while selecting software from a flexible external palette.

- Second, OSS enables central banks to tailor the software they use to their specific needs. In contrast, closed software developed by third parties typically prevents customers from sharing, modifying or using it beyond a narrow set of purposes.

- Third, the disclosure of the source code can make it easier for software users and producers to identify and correct any vulnerability. Indeed, it is usually more straightforward for central banks to test the security and robustness of an OSS than of closed-source programs, thereby reducing operational and security risks.

- Finally, OSS often benefits from a strong collaboration between developers and users. The code is typically shared on accessible "repositories", which promotes software improvements thanks to collective programming and frequent feedback. In fact, many OSS have user communities that provide efficient support and share learning resources, especially for data science applications, in turn improving the quality of the tools made available. Alternatively, if a central bank decides to produce an internal application itself and make it open source, it can more easily share the related development burden with peer institutions, in turn helping to build collective capacity and promote best practices.

Reflecting these strengths, OSS solutions have proved able to address a variety of needs for software users in central banks. Well known use cases include tools for data management (eg DuckDB, MySQL, Apache Cassandra), and big data (eg Hadoop), data analysis (eg pandas, tidyverse), and data visualisation (eg Shiny, Plotly) as well as more advanced data science applications such as ML and deep learning (eg scikit-learn, PyTorch, TensorFlow, Keras). Furthermore, central banks also use OSS for source control (eg git), integrated development environment (IDE) (eg RStudio, Spyder, VS Code) and programming languages (such as Python, R and Julia). Central banks' growing interest in these examples stems from the increasingly dominant role that OSS plays in software development, with virtually all commercial applications now having one OSS alternative. In addition, central banks' use of OSS can help them attract data scientists and other technology experts.

Moreover, a growing number of central banks act as providers of OSS, reflecting their higher degree of technological maturity. For instance, the Bank of England publishes the source code behind its research publications,③ and the Central Bank of Brazil does the same for the application programming interfaces (APIs) that support its fast retail payment system PIX.④ Further, various central banks disclose the code of their macroeconomic models⑤ or of their data management processes, as in the case of De Nederlandsche Bank for the quality rules used to improve supervisory reports.⑥ Lastly, some institutions also share the codes of their prototype central bank digital currencies (CBDCs). Notable examples are the Federal Reserve Bank of Boston's partnership with MIT on the OpenCBDC project and the Central Bank of Norway CBDC sandbox.⑦

However, central banks face important challenges in using and even more in producing OSS. One key issue is security: just as OSS is easier to audit than is closed-source software, malicious actors may be better able to understand how an OSS application could be compromised. Another important drawback is that making a software open source can require considerable financial and human resources. For example, maintaining an OSS in a fully fledged public repository will typically involve dedicated staff to address changing user needs and/or to adapt the software to the evolving technological environment.

Despite such challenges, the rising number of central banks adopting new open source tools confirms that OSS benefits are greater than their limitations. One main factor supporting this trade-off is that open source is not only about code-sharing for development purposes, as it also drives resource-sharing across the user community and active collaboration for the enhancement and development of new projects. In the age of accelerated innovation in the financial space, source code can thus be considered as one of the new frontiers of international cooperation.

The BIS has taken several steps to promote OSS. First, the BIS supports sdmx.io, a platform for managing and exchanging statistical data and metadata.⑧ This platform has become an ecosystem of OSS tools and components that allow organisations to fully exploit the SDMX open standard⑨ for collecting, producing and disseminating statistics. A key software made available by the BIS on this platform is the Fusion Metadata Registry (FMR), which enables the management and sharing of SDMX metadata and is built on open-source components shared with other organisations. Another OSS partnering with sdmx.io is the Stat Suite, a platform for the efficient production and dissemination of high-quality statistical data, which was developed in partnership with the OECD, Eurostat and the SIS-CC community. Looking forward, sdmx.io seeks to onboard more tools and components in collaboration with interested partner organisations, including the engine developed for the SDMX-based Validation and Transformation Language (VTL) by the Bank of Italy.

Another BIS contribution has been its Open Tech initiative to promote and support the development and adoption of open-source technology in official statistics and the financial sector. This initiative aims to address the growing demand expressed by central banks, commercial banks and technology providers, with the aim of collaborating together on the development and implementation of open-source solutions as public goods so as to enhance efficiency, security, best practices, knowledge-sharing and innovation. The first BIS Open Tech project was the Project Ellipse integrated regulatory data and analytics platform, which was launched in 2021 by the BIS Innovation Hub in collaboration with the Monetary Authority of Singapore.

Overall, OSS has been already adding value to central banks and is paving the way for greater innovation as well as technical collaboration with their main stakeholders.

① The Open Source Initiative (OSI) specifies internationally recognised criteria for OSS, available at opensource.org/osd. ② D Araujo, G Bruno, J Marcucci, R Schmidt and B Tissot "Machine learning applications in central banking", *IFC Bulletin*, no 57, 2022. ③ github.com/Bank-of-England. ④ github.com/bacen/pix-api. ⑤ D Araujo, *Open-sourced macroeconomic models*, 2023, github.com/dkgaraujo/OpenSourcedMacroModels. ⑥ github.com/DeNederlandscheBank/data-quality-rules. ⑦ Respectively available at github.com/mit-dci/opencbdc-tx and github.com/norges-bank/cbdc-sandbox-frontend. ⑧ The platform provides tools for cleaning, transforming, and publishing data to make them more easily accessible and usable, www.sdmx.io. ⑨ SDMX is a data exchange standard used by international organisations and national statistical systems, and the platform aims to make working with this standard more user-friendly and efficient, sdmx.org.

A third important area is to **ensure that data scientists' operations are based on sound and efficient IT processes**, especially when they rely on self-service capacities. For instance, and as mentioned above, a key benefit of the *gingado* package is indeed to foster the dissemination of good practices in ML use cases. It also promotes efficiency through its consistent and simple application programming interface (API) that makes it easy to plug into existing code or to reuse ML code by other teams. In addition, other best practices in software engineering are finding their way into central bank processes. One key example is the use of containers – ie a fully self-contained operating environment on which a user application can be run in an

isolated and portable way.[6] Containerisation ensures that specific analyses can be "packaged" and therefore run on different computers. It thus provides important opportunities to make business operations more efficient, fosters the reproducibility and portability of the projects involved (for instance in economic research; Vilhuber (2021)), and promotes internal collaboration as well as cooperation with academia and external organisations. These benefits were highlighted by the recent containerisation project (based on the Docker tool) undertaken by the Bank of Canada with Emory University, Ryerson University and Stanford University.

Using cloud computing services?

Cloud computing has emerged as one of the key technological innovations in data architecture, offering advantages such as "scalability" – ie the ability to quickly adapt the IT hardware to perform well a larger number of analyses and processes, operational efficiency, the possibility to use a wider range of tools and computing resources, and continuity – meaning the possibility to keep the software stack up to date (IFC (2020)).

Certainly, **a number of central banks remain sceptical**. In a recent survey of senior IT managers at 25 central banks, Edmond et al (2022) report a reluctance to migrate to the cloud due to concerns about data protection and privacy, security and other operational risks; and even when a cloud strategy exists, a private cloud environment will often be favoured over a public cloud.

Nevertheless, the increasing experience in using cloud computing services has highlighted **a number of important benefits for central banks. One relates to more efficient data ingestion processes.** A cloud-based environment can flexibly allow for the onboarding of large data volumes at speeds that are multiples of those seen with more traditional IT infrastructure. For instance, the cloud-based data lake solution explored by the Bank of Canada accelerated the loading of a data set with millions of observations. One reason was the reduced need for computing memory allowed by the common platform, compared with the previous scenario where each user would require a copy of the data. In another example, the Bank of Canada has been using cloud services to automatically ingest on a weekly basis expenditure information at commercial addresses collected by SafeGraph, a private data vendor. The data set appears to usefully complement other official statistical sources, especially in terms of timeliness and details. For instance, it provides information on businesses openings and closures, facilitating the organisation of the Bank of Canada's payments surveys.

A second main benefit provided by cloud environments relates to the manipulation of the data once they are ingested. This takes advantage of massive parallel computing, which is the ability to "divide-and-conquer" calculations to accomplish tasks much more quickly than if they were done sequentially. The resulting gains in speed and productivity can be observed for many data management processes, such as data pre-processing and visualisation, but also for analytical tasks. The example provided by Nvidia shows that the parallelised calculation capacity offered in the cloud through graphics processing units (GPU) chips have enabled recent breakthrough advances in sustainable finance, LLMs and

[6] In other words, containers encapsulate all dependencies of a certain application to ensure it runs the same way in different environments (Mouat (2015)).

other fields relying on the use of AI/ML tools – noting that the training of most of these tools requires considerable computing power.

Further, and despite recurrent security concerns as mentioned above, **cloud-based solutions can also help to maintain or increase the security posture**. First, many cloud providers offer advanced threat detection and response capabilities, which can automatically identify and mitigate potential security breaches, something that might be resource-intensive for a central bank to manage on its own infrastructure. A related example of cloud usage for its security focus is to enable secure collaboration with external researchers. Moreover, a cloud environment can, generally speaking, provide efficient processes for securing software, such as by implementing frequent updates and patching to close any known vulnerabilities. Of course, this does not mean that central banks should entirely outsource cyber security management, but that cloud service providers can be important partners in maintaining an efficient and secure working environment.

In any cases, there are substantial training and other costs associated with the implementation of cloud platforms, although the efficiencies gained might offset these, as discussed by Handel et al (2022). Hence it is important to rigorously **evaluate the resources implications before making any decision to transition to the cloud**. One key element is that on-premise solutions often demand significant expenditures, encompassing hardware acquisition, the setup and maintenance of a data centre, and the associated staff costs. In contrast, cloud solutions usually adopt an operational expenditure model, implying that the recurrent costs supported by client institutions are determined on an ongoing basis and aligned with their usage patterns. A second important point is that, as time progresses, on-premise IT setups would experience higher costs due to hardware ageing and related upgrade requirements. This is less of an issue for cloud infrastructures, which leverage economies of scale and continuous innovation and do usually not pass direct incremental hardware costs to user institutions. Nevertheless, while cloud-based solutions may appear advantageous by smoothing user costs and keeping available IT tools updated over time, it remains crucial to take a comprehensive view of the total expenses involved as well as to keep other strategic considerations in mind.

3. Leveraging statistical and mathematical tools to extract data insights

The second main dimension of data science relates to the **ability to perform various mathematical and statistical operations to deal with the raw data available**. From this perspective, a sound IT infrastructure that supports the required analytical tools and software (see Section 2) is a necessary but not sufficient condition. Data scientists must have the skills to apply various scientific methods and algorithms to access very large and complex data sets, prepare them for further analyses, and conduct inference exercises.

Accessing new, big data sources

Accessing new large and multifaceted "big data" sources has become increasingly important for central banks. One key reason is that, in an increasingly complex financial environment, policy institutions require more information and more

promptly ("actionable knowledge") if they are to fulfil their mandates – see IFC (2020; 2021a). Fortunately, central banks' thirst for data can be effectively quenched by leveraging new information sources, ie so-called alternative data. Their usage has increasingly supported a wider range of monitoring and policymaking tasks, as was evident during the Covid-19 pandemic. Looking ahead, these new sources are likely to continue to provide useful value as a complement to the traditional ones in the "new normal" landscape for official statistics (Jahangir-Abdoelrahman and Tissot (2023)).

A key factor, as observed by the Bank of France and the French Prudential Supervision and Resolution Authority (ACPR), is that **alternative sources can bring more timely and higher-frequency data**, which can be very useful in uncertain and fast-moving situations such as a financial crisis. Moreover, the new types of source considered, ranging from Google searches and social media posts to satellite images or mobile phone data, may offer unique insights. For instance, compared with "traditional" statistics, they can provide an almost real-time view of economic developments, often with high granularity in terms of sectors and/or geographic locations. One example has been the use by the International Monetary Fund of payments data from M-Pesa, one of the largest mobile money services in Africa, to analyse recent trends in mobile money – eg the relative importance of peer-to-peer (P2P) transactions and international money transfers and the impact on broad money (Shirono et al (2021)). Another example has been the use of Google Maps data to geolocate financial access points such as ATMs and mobile money agents in Kenya, complementing information collected through more traditional financial inclusion surveys.

A second factor supporting the access to alternative data sets is that they often represent "low-hanging fruit", since they usually exist as an organic by-product of existing processes and activities. In particular, authorities are increasingly realising the potential value for economic and financial analyses of the wide range of administrative registers that are generated by public sector activities. For instance, electronic payments data have been used in Portugal to study consumer behaviour (Carvalho et al (2020)); similarly, cargo ship identification information was helpful to better track global supply chain activity in real time (Cerdeiro et al (2020)).

Needless to say, data scientists need to be able to **deploy various techniques and tools to efficiently make use of all the various data sources of interest**. This is needed for data collation, ie to properly query, integrate, store, clean, prepare and process the raw data. Cases in point relate to SQL for querying relational databases; the usage of Hadoop for processing vast amounts of data; applying packages such as tidyverse in R or pandas for data manipulation in Python; using integration tools like Talend or Apache NiFi to combine disparate data sources; and leveraging Tableau Prep, OpenRefine or other modern business intelligence (BI) software for cleaning and preparing the data (IFC (2019)). All these steps are essential to make the information ready for further use. And, once the raw data are properly organised and prepared, various platforms such as Jupyter Notebooks, RStudio and Quarto can be mobilised for conducting deeper – and reproducible – analysis.

Preparing raw data as the basis for further analyses

Once the data are accessible, the next step is to prepare them for knowledge extraction. Advanced analytical tools can be instrumental from this perspective,

especially as regards three main tasks inherent to central banks' statistical activities: the selection of the data features that are relevant; the management of data quality; and the dissemination of data to users.

As regards **signal extraction**, modern analytical tools are needed to select the indicators of interest, especially when information is buried in a sea of granular data points. For instance, Bank Indonesia has extracted over 600,000 user comments on the TripAdvisor platform, covering more than 1,000 touristic destinations in the country. This information was used to better measure tourism dynamics, for instance to assess the impact of Covid-19 pandemic. In parallel, another language-based project was related to the compilation of government expenditure statistics, by automating the classification of individual payment transactions.

Meanwhile, the Bank of Spain has developed a specific application to create a database of sustainability information on Spanish corporations. The tool extracts a structured database from the vast amount of environmental, social and governance indicators disclosed by corporates so as to enable regulators to better monitor them. Certainly, the underlying components of the application required significant open source inputs, starting with Streamlit,[7] an open-source application for ML work, and Dash, a Plotly Python framework for creating interactive web applications.[8] But in addition to the IT side, the overall data science initiative also required the involvement of statisticians to select the information of relevance. Similarly, the Banco do Portugal has designed a web application to ingest IMF data on the Coordinated Direct Investment Survey (CDIS) and the Coordinated Portfolio Investment Survey (CPIS) to conduct network analysis and derive countries' influence in global direct investment and portfolio flows. The aim was also to allow users to interact with the data, for instance to track individual countries' positions in global investment networks over time.

The second area, **data quality management**, is an imperative for central banks that are important producers of top-quality statistics. But they are also heavy users of data that need to be reliable to support their analyses and ultimately policy decisions. Hence, it is essential for them to detect any quality problems adequately and on time, especially when confronted with spurious data reports. One example was during the financial turmoil induced by the Covid-19 pandemic in March 2020, when central counterparties' (CCPs) initial margins increased dramatically (Boudiaf et al (2023)). Authorities had to analyse whether this reflected a real development, ie the demand for additional protection against a potential default of CCPs' clearing members, or was simply a matter of misreported data.

Fortunately, **advanced analytical tools can be a great help in ensuring the quality of the data sets compiled,** even when these are highly granular and so large that they cannot be manually checked. For instance, the supervisory (big) data set collected from trade repositories (TRs) under the auspices of EMIR presents significant quality challenges due to the high volume of data as well as to obstacles in aggregating inputs across TRs, since a similar transaction can be reported multiple times and in different ways (ESMA (2021); IFC (2018)). The ECB's strategy has been to automate quality controls, allowing for the production of timely analyses. Another example relates to the ECB CSDB, where data on all individual securities in Europe are

[7] www.streamlit.io.

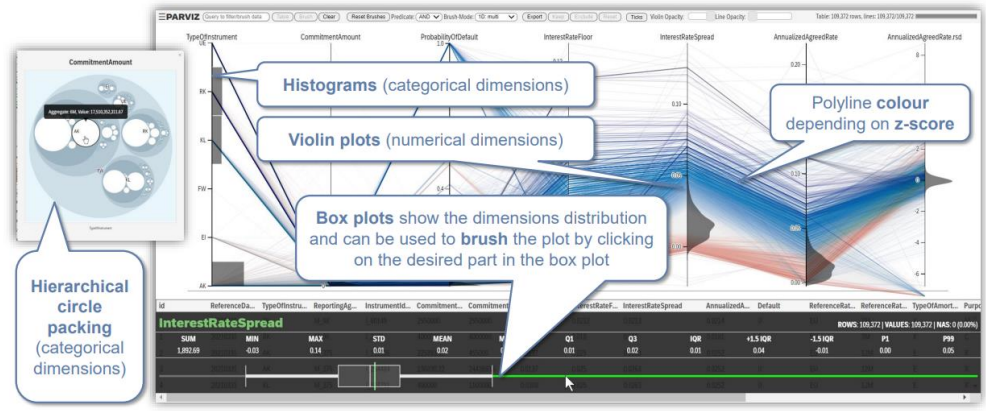
[8] plotly.com/dash/.

compiled from multiple sources. Various algorithms – eg logistic regression, decision trees with the use of the ML gradient-boosting algorithm CatBoost (Prokhorenkova et al (2018)) – are used to select the data points that would require intervention by a data quality manager. The project has also focused on making the results explainable to end users (by using Shapley values for interpreting the ML modelling) and ensuring their relevance over time as new data come in (by checking the model performance over different periods). The Bank of Italy has also leveraged ML algorithms for data quality controls applied to its credit registry, with the combination of advanced techniques including an autoencoder, a type of artificial neural network to predict data anomalies.

The above examples underline the key contribution of data science in supporting the efficient production of statistics by central banks, helping to increase timeliness while respecting resource constraints and without compromising on data quality. It is also worth noting that **the new tools used for quality management can help authorities liaise with data reporters so that they improve their data submissions**. The ECB, for example, has worked closely with the industry to create the Banks' Integrated Reporting Dictionary (BIRD), a data dictionary to support the standardisation of data transmissions for regulatory or statistical purposes. This has helped to reduce banks' reporting burden, improved the data collected thanks to a better understanding of the reporting guidelines, and facilitated statistical calculations. In addition, regular quality feedback reports are sent to the original data owners of the TR data collected in the EMIR context.

Visual detection of outlier observations in credit data collected by the Austrian central bank

Graph 2



Source: P Reisinger, T Kemetmüller and C Leitner, "Interactive visualisation tool: outlier detection in large multidimensional data sets", *IFC Bulletin*, no 59, October 2023.

Another important comment is that **data quality management can be facilitated by statistical visualisation tools**, particularly to detect and analyse specific data points of interest. For instance, the Banco de Portugal has been using PowerBI, a BI visualisation tool, to check the information contained in its repository of balance sheet data for non-financial corporations. The tool allows for different data sources to be rapidly connected, as well as for consistency checks and drilling down into aggregates so that analysts can better spot individual data points that might need further investigation. In a similar vein, the Central Bank of the Republic of Austria

has developed a tool-based framework to visualise and detect outliers in its credit register. The approach is dynamic as it facilitates the visualisation of the various dimensions of the data (more than 100 variables) while also allowing users to change their search criteria (**Error! Reference source not found.**). Similarly, the BIS makes use of the Tableau dynamic visualisation tool to support data management tasks. The approach is multidisciplinary, involving IT technical features and advanced statistical methodologies, as well as communication aspects (eg as regards the cognitive content of the graphs and related visual perceptions); hence a key lesson of the project was the need to involve the various community of data scientists from different units.

Data science applications have also found their way into the **third main area of data dissemination**. For instance, central banks often receive information requests associated with the data published, eg from the media and the general public. Addressing these requests can be resource-intensive, not least because they need to be properly assigned to the business areas in charge so that they can respond. The ECB's mail robot (MailBot) system was developed to automate the sending of such messages and replace manual intervention. The application allows the classification of inquiries by business areas, the identification of similar queries to avoid duplication of work and the automation of some reply processes. The model was trained on previous requests and answers provided by the business areas, using a specific classification algorithm (extremely randomised trees; Geurts et al (2006)) and estimating the degree of similarity between text queries through their vectoral representations.

Extracting analytical insights from data…

Much of what is usually considered as "data science" encompasses the **various mathematical methods that can be used to extract relevant insights from the data once they have been properly collected and prepared**.

In general terms, these methods are based on AI, ie "the various computer systems that can perform tasks that traditionally have required human intelligence" (FSB (2017)). This includes the important subset of ML, "a method of designing a sequence of actions to solve a problem that optimise automatically through experience and with limited or no human intervention". There is also a growing interest in so called generative AI, ie AI models learning from their input ("training data") and able to generate new data with similar characteristics (for instance new texts produced with LLMs).[9] The approach can be top-down, with humans designing how the data should be processed by the systems that can mimic human-like calculations, but much faster and on a vastly greater scale. It can also be bottom-up, using algorithms adapted to fit the data and with a focus on optimising the intrinsic performance of the model instead of mimicking human behaviour.

In practice, **data insights can be extracted in multiple ways**. The diversity of ML algorithms provides a large number of practical alternatives for exploring data (Athey and Imbens (2019)). When the task can be performed with the help of a particular subsample of the data set for which the outcome is known (and on which the model can be trained), "supervised learning" algorithms can be used, such as

[9] Examples of LLMs include OpenAI's GPT models (used in ChatGPT), Google's PaLM (used in its conversational AI tool Bard), Meta's Llama as well as BigScience's open model BLOOM; see Box 2.

regularised regressions, random forests, gradient boosting trees and regression or classification neural networks (LeCun et al (2015)). In other cases – for instance when the explored data points need to be grouped automatically or summarised into fewer dimensions, or for other cases where the algorithm has to identify the patterns in the data autonomously – then unsupervised learning ML models such as clustering techniques and manifold learning can come into play. Other ML models, less used in data science currently but with arguably considerable potential in supporting decision-making, are part of reinforcement learning, ie ML algorithms that follow an optimisation rule to maximise a specific objective. Importantly, models of the same or different types can be combined in more or less sophisticated ways, as shown in the examples discussed by Araujo et al (2022).

… including unstructured data like text

An increasing number of central banks are working on the use of NLP techniques to support a whole range of applications when dealing with new information sources (eg Gentzkow et al (2019); Araujo et al (2022)). This interest reflects **the large amount of text data, often unstructured, that are available to them as part of their routine activities.** Another important development has been technological advance, which has facilitated the development of NLP models based on multiple languages and thereby their global use.[10] Moreover, NLPs' capabilities have been expanding, spurring central banks' interest in LLMs (see Box 2).

Box 2

Harnessing recent breakthroughs in large language models (LLMs)

Douglas Araujo, Stephan Probst, Rafael Schmidt, Boris Vitez, Markus Zoss

In recent years, the field of natural language processing (NLP) has made significant progress, primarily due to the emergence of LLMs (Graph 3) such as ChatGPT – the language model developed by OpenAI, which is capable of generating text based on context and past conversations. This box discusses some of the ways LLMs might be helpful for central banks and the associated challenges.

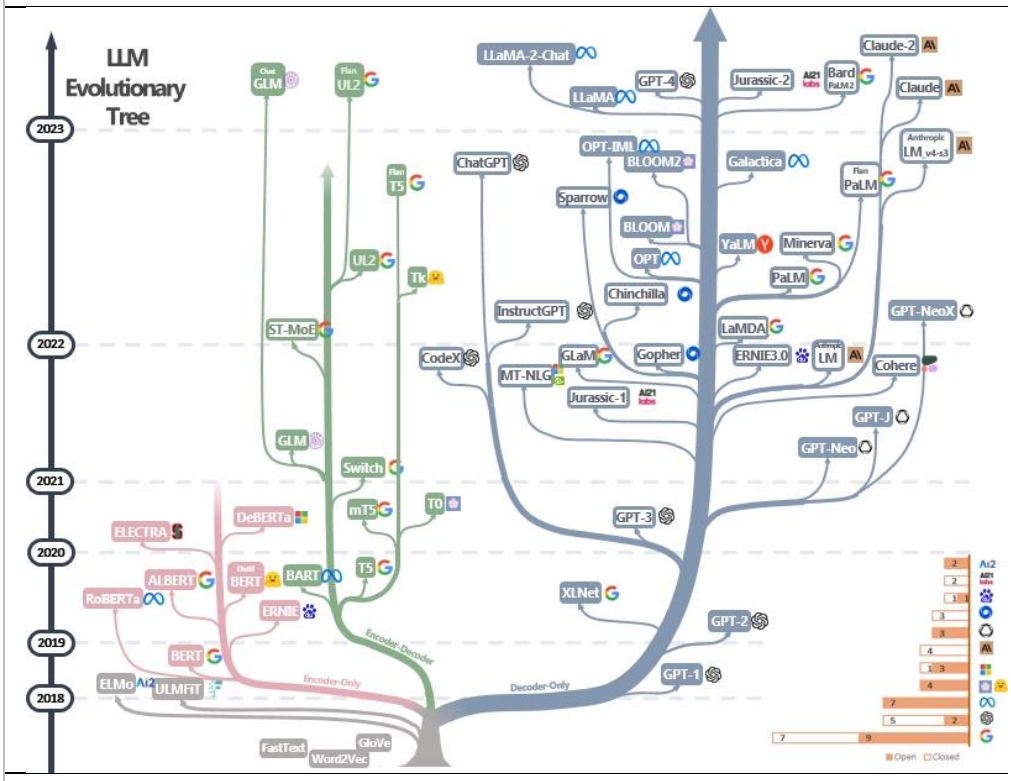
These models are built with a ground-breaking neural network architecture known as the "transformer",① which enhances models' ability to grasp nuances of word meanings in their context and enables them to expand in size and process vast amounts of text. As a result, LLMs now achieve human-like proficiency in a variety of tasks that involve language, such as generating various text formats, text summarisation, sentiment analysis, translation etc. Sophisticated post-training methods, such as supervised fine-tuning and reinforcement learning from human feedback, have further refined the more recent LLMs, providing them with more advanced reasoning capabilities. These developments open up new potential use cases.

[10] Initially, the development of large NLP models concentrated in the English language and the availability of advanced NLP tools was more limited for non-English languages. To address this issue, the Bank of Thailand has developed three AI-powered NLP applications so that it can analyse documents written in the Thai language. Looking ahead, the growing availability of various models even for languages with relatively fewer speakers, such as Google's MADLAD-400 (Kudugunta et al (2023)), is likely to spur central banks' exploration of various applications using local language(s).

A combination of factors contributes to the superiority of LLMs over other language models. The first is their architecture: unlike previous models, which look at text data sequentially, transformers can capture long-range dependencies and better differentiate the meaning and importance of words in different contexts. Also, this architecture can be significantly scaled up, leading to the second factor behind LLMs' success: their size. These models are much larger, allowing them to learn more intricate patterns and representations from the massive amounts of text data they are trained with. As a result, their output better resembles general language and specialised areas of written knowledge. A third factor is how they are trained. While generally all LLMs follow lengthy "classical" training processes, albeit with a larger data set, more recent models add more human-intensive steps that seek to improve their helpfulness, harmlessness and truthfulness. With this, LLMs can more easily adapt to new tasks after being given a few examples or even without any examples. These factors support enhanced performance and, crucially, the ability to transfer knowledge from one task to another, which is a key advantage of LLMs.

The evolutionary tree of large language models (LLMs)

Updated as of 6 August 2023 Graph 3



Source: J Yang, Q Feng, X Han, X Hu, H Jiang, H Jin, R Tang and B Yu, "Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond", arXiv, April 2023.

LLMs have led to breakthroughs in various NLP applications, such as in conversational AI, leading to significant quality improvements. For instance, new computer dialogue systems can provide more coherent and contextually relevant responses to human inquiries. For text summarisation, LLMs generate accurate and concise summaries of lengthy documents in a way that can be calibrated by the user to reach different audiences (eg a technical abstract, or a higher-level executive summary). Further, these models can accurately perform sentiment analysis of texts, by capturing nuances and contextual information better than previous models. Other examples are automated translation between languages, where LLMs can produce more fluent and contextually accurate output, and software engineering, where these models may help design, test and document code.

Because of these capabilities, LLMs are becoming useful tools in economic research, echoing the achievements they have brought in other domains. At the inception of research projects, they may assist in ideation by providing brainstorming possibilities and informed counterarguments. In the writing phase, LLMs can aid in text synthesis, editing and crafting compelling headlines, making the dissemination of research more impactful. Background research could also be streamlined with LLMs by supporting literature reviews, reference formatting, and even explaining complex concepts in simpler terms. On the technical front, LLMs can help in coding and debugging, data reformatting, or data classification. The versatility of LLMs may therefore lead to important possibilities for reshaping economic research looking forward.②

Central banks are also increasingly recognising the potential of LLMs to support their various activities.③ Economic data analysis is a major area of focus, as LLMs can help to parse vast amounts of economic reports, documents and news to efficiently extract, summarise and present information. In particular, the analysis of financial news, social media and other public fora facilitates the gauging of public sentiment, for instance as regards current and future economic and financial conditions, or the conduct of specific policies. The models also offer valuable support for the various internal exercises (eg risk assessments, economic forecasts) as required by the conduct of monetary and financial stability policies. Lastly, some central banks have started deploying LLMs in public relations tasks, for instance to handle inquiries from the public or financial institutions or to explain complex economic topics, regulatory requirements as well as policy decisions.

Looking ahead, the range of central banking applications that could benefit from LLMs is likely to continue to expand, especially in the policy area. For example, these models could support more complex financial monitoring exercises by tracking in a granular way market segments, institutions or instruments that show signs of instability or policy concerns. Regulators could leverage LLMs to facilitate compliance monitoring, eg by analysing internal reports to assess the implementation of financial regulations. Similarly, supervisors could detect suspicious patterns and potential misbehaviour or illicit activities by scrutinising transaction-level data and related textual content. Lastly, LLMs might assist in training central bank staff as well as the broader financial community and the public at large, by making complex topics more accessible, enhancing financial literacy, and supporting data queries at scale through more intuitive interfaces.

However, with great potential comes great responsibility. Central banks are rightfully cautious as they integrate these technologies. Transparency is paramount in understanding how the new models derive their conclusions and in communicating to the public the reasoning behind policy actions. A related issue is algorithmic unfairness, ie the risk that (non-explicit) unfair discrimination or bias entailed in pre-existing material used for training the model can be systematically perpetuated or even amplified, especially as it interfaces directly or indirectly with people. The trust in high-quality official statistics hinges upon addressing this pivotal question.④ Consequently, a number of authorities have worked recently on documenting ethical considerations for guiding ML development and usage.⑤ Data privacy also remains a top concern, given the risk that sensitive information could be inadvertently processed by LLMs and then disseminated in unexpected ways. Lastly, as for other AI-based tools, there is always the danger of overreliance on automated processes, with the need to strike the right balance between machine assistance and human judgment, especially for policy decisions.

In summary, LLMs have already ushered in a new era of automation and experimentation in central banking operations. Their ability to understand language and context in a massive and automated way has led to breakthroughs in several real-world applications. These advances can hold great promise in the central banking domain. But, as AI's role evolves, continuous research, collaboration and transparency will be essential to harness its benefits responsibly.

① A Gomez, L Jones, Ł Kaiser, N Parmar, I Polusukhin, N Shazeer, J Uszkoreit and A Vaswani, "Attention is all you need", *NeurIPS Proceedings*, 2017. ② A Korinek, "Generative AI for economic research: use cases and

implications for economists", September version submitted to the *Journal of Economic* Literature, 2023. ③ D Araujo, G Bruno, J Marcucci, R Schmidt and B Tissot, "Machine learning applications in central banking", *IFC Bulletin*, no 57, 2022. ④ C Julien, "*Machine Learning project report*", UNECE, September, 2020. ⑤ UK Statistics Authority, "*Ethical considerations in the use of Machine Learning for research and statistics*", October 2021, https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-machine-learning-for-research-and-statistics/.

As a result, various **NLP-based data science projects aim to leverage the considerable amount of texts available from various different sources**. This material is being increasingly used to support internal processes in central banks, reflecting the fact that they routinely analyse very large volumes of textual information, usually in unstructured format – ranging from internal documents (eg financial stability or monetary policy reports, governors' speeches), reports from supervised entities (eg board documents, loan documentations), or external publications (eg social media, financial press, economic papers) – that often need to be analysed by various teams with different skills and background in parallel.

Central banks are taking these opportunities to explore and eventually reap benefits from a variety of texts, including regulatory and market data. Crucially, these data sets can be used by themselves, or merged with traditional data, to obtain new policy insights, eg by offering a more timely and granular assessment of the economy and by allowing a better understanding of how market participants are affected by shocks in financial networks. These use cases have benefited from the **recent breakthrough provided by language-based ML models**, which cover a broad range of applications, both for internal processes and to directly support policy. But of course, LLMs also have downsides, including questions about how to ensure the suitability of training data, the need for significant investments to preserve security, their high energy consumption during training, and the often unsatisfactory levels of transparency (Chen et al (2023)) – a point of particular importance for public authorities. The more experience central banks gain in exploring these issues, the more they will be in a position to address the associated challenges while keeping risks under control.

Experience shows that **there is no one-size-fits-all approach**. Advanced NLP models can address a wider range of use cases, but ML techniques can also allow for more nuanced, quantitative analyses of textual information. Moreover, simpler techniques can also deliver important analytical benefits. This puts a premium on combining various techniques. One example is the Bank of Italy's assessment of whether information from newspapers can be useful for forecasting the stock market index and the relative performance of Italian banks. A sentiment analysis dictionary was developed to assess the predictive power of news, as broken down by topics using the Lasso regression analysis method (Tibshirani (1996)). The approach underlined the importance of using local language dictionaries in text analyses and also the fact that **interesting results could be obtained with simpler techniques such as dictionary-based search analyses compared with more sophisticated and resource-intensive NLP models**.

4. Incorporating subject matter expertise to support well defined analytical use cases

The use of advanced analytics, to be fully efficient, **requires a good understanding of the projects involved and therefore a close interaction with the business areas, including their subject-matter experts and statistical methodologists**. In other words, a data science process developed in a specific domain cannot be blindly applied to another completely different area. That third dimension of data scientists' work is essential, not least because it is key to ensure that the insights extracted from existing indicators represent useful knowledge (Drozdova (2017)). This aspect is even more important in policy institutions such as central banks, since their actions are taken on the basis of available data and have to be clearly explained to the public.

To illustrate this point, **three main areas deserve to be highlighted among the various data science projects undertaken by central banks**. First, the new approaches can help to make a better sense of the vast amount of granular information available in today's societies.[11] Second, they tend to provide useful insights on the state of the economy and its prospects, which is a key input for central banks in pursuing their mandates in the areas of monetary and financial stability. Third, they may help to assess how their policies are communicated and can be made more effective.

Making sense of the wealth of granular financial information

An important issue for central banks and financial supervisors is to deal with the vast amount of granular data collected on the financial system and to extract relevant indicators (IFC (2021b)). The problem here is to be able to detect signals at a very detailed level (eg a specific institution or a market segment) without being overwhelmed by the vast amount of data points to be considered. The key is to see "the forest as well as the trees" (Borio (2013)) which calls for both automated analytical techniques and a good understanding of user information needs.

Network analysis methods can be useful in addressing these aspects, as shown by the Bank of Japan in monitoring transactions in the Japanese government bond (JGB) repo market. The approach helped to extract two main insights from the large and granular data set considered: (i) transaction relationships are built around a small number of leading institutions acting as market intermediaries; and (ii) the Covid-19 pandemic significantly reduced the importance of securities lenders relative to other intermediaries. In another study, the Bank of Japan assessed the role of currency swaps as a source of US dollar funding based on transaction-level data collected on the over-the counter derivatives market. The approach proved helpful in analysing the characteristics of the cross-currency swap market in Japan; and in monitoring market liquidity and the trading behaviour of market participants in a timely and detailed way and at a high frequency. Network analysis and NLP techniques have also been applied in a number of jurisdictions to identify fraud or misconduct. The aim is typically to extract from the vast amount of information

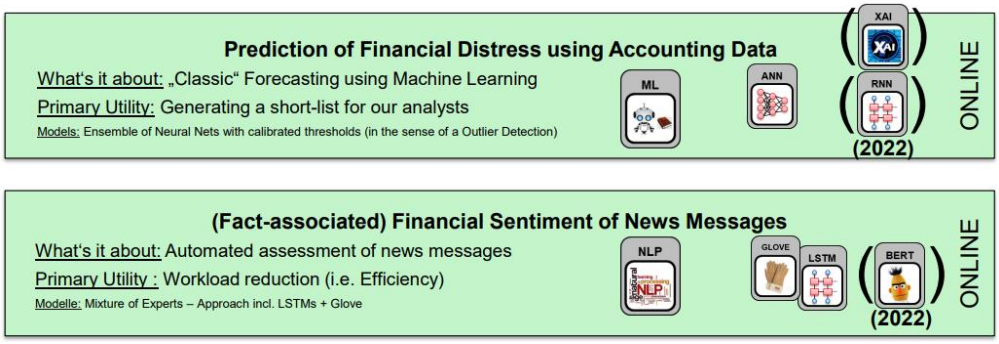
[11] In addition to the basic data editing tasks that can greatly benefit from data science applications when dealing with large data sets, as argued in Section 3.

(including text) generated by financial activities signals that need to be further analysed by experts and that may warrant on-site examination by supervisors.

Other techniques can also be applied in isolation or as a complement to facilitate micro-level investigations. For example, the Deutsche Bundesbank has in production a number of tools for risk management purposes. One key objective was to improve the assessment of the credit risk posed by its market counterparties, using ML-based forecasts of potential financial distress (estimated from accounting data) complemented by an automatic NLP-based algorithm so as to select incoming news for further investigation (**Error! Reference source not found.**). In a similar way, the Bank of Thailand's supervision of domestic financial institutions has been facilitated by an NLP-based analysis of the meeting minutes of boards of directors – using a combination of tools to support word segmentation, topic modelling and name-entity extraction (**Error! Reference source not found.**).

Use cases in production for machine learning algorithms at the Deutsche Bundesbank

Graph 4



Source: B Sahamie ,"Natural language processing for risk management: discussion of use cases", *IFC Bulletin*, no 59, October 2023.

Nowcasting and modelling the economy

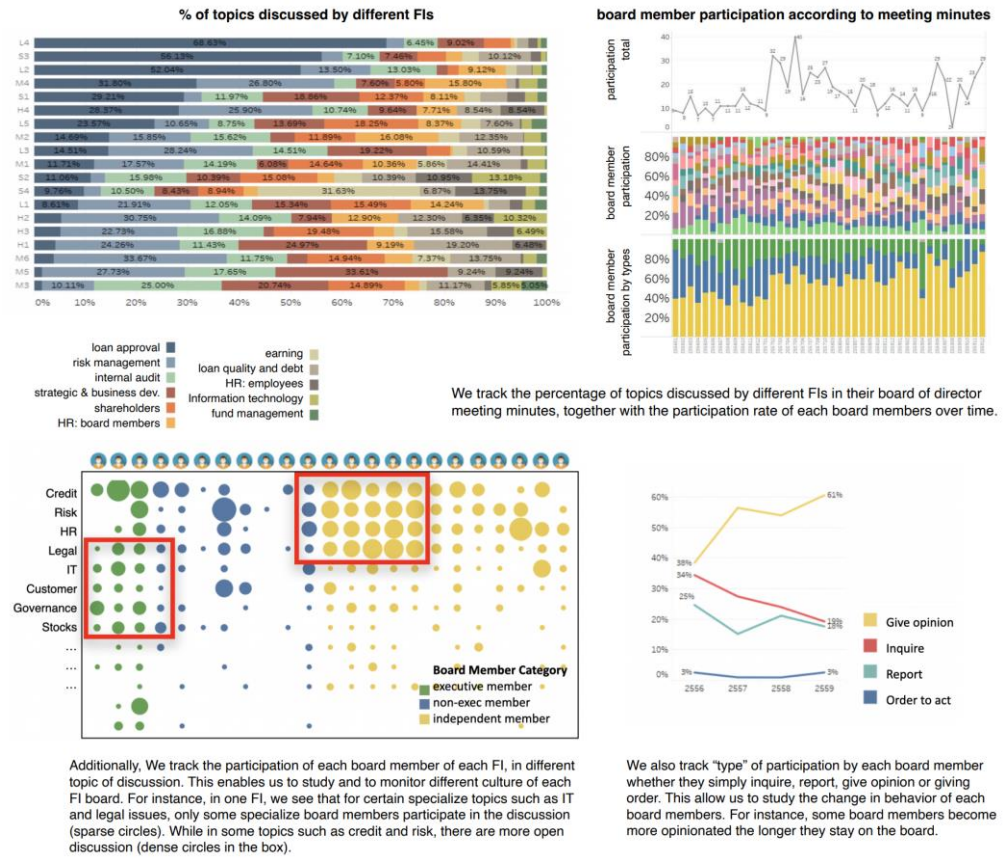
In pursuing policy goals such as monetary and financial stability, central banks have to rely on a sound analysis of the state of the economy (eg GDP growth, inflation) and potential developments. **Nowcasting techniques have therefore been increasingly used for supporting the associated monitoring and forecasting exercises** (see Giannone et al (2008); Bańbura and Modugno (2014); Kronenberg et al (2023)), with typically two main approaches.

The first type of approach is data-driven, by agnostically "letting the data speak" before validating the results with expert judgment. A large number of variables will often be considered by the model, with due consideration for the different frequencies available and/or data vintages. For instance, the BIS has an initiative to nowcast countries' business and financial cycles based on a fully automated infrastructure so as integrate a large pool of indicators (ie various macroeconomic and financial time series as well as social media text) as they come in. Similarly, an ECB and Delft University of Technology project has combined traditional vector autoregression (VAR) modelling with the use of a deep learning architecture based on neural networks to estimate relationships in the economy. The results suggest that

there are important non-linearities in the interactions between variables and across time periods that could be better explored for macroeconomic modelling.[12]

Analyses of discussions of boards of directors at financial institutions supervised by the Bank of Thailand

Graph 5



Source: Treeratpituk, P and J Kerdsri, "Integrating natural language processing to central bank operations at Bank of Thailand", *IFC Bulletin*, no 59, October 2023. FIs = financial institutions; IT = information technology.

Input data can also be non-quantitative, such as text. For instance, the Bank of Korea has put together 450,000 news articles from the web to build a news sentiment index (NSI) for the Korean economy using neural networks techniques. This NSI has proved to be a good leading indicator for GDP growth, allowing policymakers to assess the economic situation in a timely manner and with little cost (in comparison, for instance, with business surveys). Another analysis by the Delft University of Technology showed that extracting textual information from economic and financial news can help to anticipate oil price uncertainty. Based on the Baker et al (2016) methodology, an index was built to reflect the degree of uncertainty involved in news texts and the topical characterisation of the spikes observed in oil price uncertainty. The project involved the embedding of the texts, by mapping their words into numerical vectors (using the neural network-based approach doc2vec; Le

[12] Lenza et al (2022) have also documented the importance of considering the non-linear relationship between inflation and its determinants in Europe, using a quantile regression forest approach.

and Mikolov (2014)), an assessment of the similarity between the articles considered, their classification in clusters based on the Louvain algorithm (Blondel et al (2008)), and the identification of each cluster by a specific topic using the Latent Dirichlet Allocation method (Blei et al (2003)).

A second approach is to select ex ante a set of specific variables and assess their usefulness for economic analyses. A case in point relates to payments data, which have been collected for many years by several central banks and are now increasingly tapped to extract relevant insights. For instance, Bank Indonesia leverages data on retail payments transactions collected from the National Clearing System and classified by specific keywords to construct a real-time measure of household consumption. One issue, however, is that the model performance is not a given and may vary significantly over time, even though it appears to have improved significantly since the pandemic. In parallel, Bank Indonesia also nowcasts sectoral corporate activity using transaction-level payments data between more than 100 corporations spanning nine sectors, as collected in its real-time gross settlement system. One challenge was to clean entity names to ensure that the same firm is identified even in cases where the spelling changes slightly across different data points. Another issue was the different model performance observed across economic sectors. A third example is the Bank of Japan's work with personal mobility data collected from smartphones to nowcast economic activity in the first stages of the Covid-19 pandemic. Business data with GPS information from Agoop was matched with administrative data sets to measure business activities in labour-intensive industries and the services sector.

Of course, **the two types of approach can be combined**. For instance, the Narodowy Bank Polski has used a vector error correction model (VECM) to model the non-performing loan (NPL) portfolios of Polish banks. The project first looked agnostically at the various potential explanatory variables that might drive NPL dynamics with a view to selecting the most important ones, such as economic growth. This was combined with a parallel study that was focused on the specific impact of foreign direct investment (FDI) on economic growth and NPLs.

In any case, a major lesson from the above examples is, first, that there are increasingly available information sources that one can tap to assess the evolving macroeconomic landscape in a timely way. Second, as for other data science applications, **nowcasting exercises require the combination of strong IT computing capacities, advanced analytical techniques, and deep expert judgment** – either to validate ex post the automated results obtained, or ex ante to select the variables of interest before testing their information content.

Policy communication

An important issue is to **assess the way policy is communicated to the public, an area to which central banks are paying increasing attention** (IFC (2022)). For instance, the Reserve Bank of Australia has analysed how various audiences perceive communication quality in terms of the degree of readability and reasoning attributed to the messages published. These were processed by an NLP model containing a part-of-speech (PoS) tagger to identify the contribution of each word to a particular part of the text, coupled with a syntax tree parser to take into account its grammatical structure. The processed messages were then classified with random forests depending on their characteristics such as their degree of readability or reasoning

and types of audience (eg economists vs others). The analysis showed that there could be some trade-offs, for instance that simplicity may improve the readability of a text but not enhance the clarity of its reasoning – suggesting that central banks may be better off by producing different texts, each tailored to a specific audience category.

Moreover, **language-based models have proved helpful for assessing the effectiveness of central bank policies**. As an example, the Central Bank of Malaysia has used automated content analysis to extract the sentiment from each of its monetary policy statements published between 2004 to 2020. The initiative required the development of a specific monetary policy dictionary to classify words as "dovish" or "hawkish". This allowed the Bank to test the prediction content of its statements in terms of interest rate developments and other financial market movements. Another approach followed by Bank Indonesia aimed to capture public opinion, as expressed on social networks, on issues related to central bank activities. In particular, the project deployed language analytical tools both to measure public perception of the credibility of central bank monetary policy actions based on multiple daily articles collected from varied domestic news source and to test its relevance for analysing the level and stability of inflation expectations.

References

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

Araujo, D, G Bruno, J Marcucci, R Schmidt and B Tissot (2022): "Machine learning applications in central banking", *Journal of AI, Robotics & Workplace Automation*, vol 2, issue 3, pp 271–93.

Athey, S and G Imbens (2019): "Machine learning methods that economists should know about", *Annual Review of Economics*, vol 11, pp 685–725.

Baker, S, N Bloom and S Davis (2016): "Measuring economic policy uncertainty", *Quarterly Journal of Economics*, vol 131, pp 1593–636.

Bańbura, M and M Modugno (2014): "Maximum likelihood estimation of factor models on data sets with arbitrary pattern of missing data", *Journal of Applied Economics*, vol 29, no 1, pp 133–60.

Bholat, D (2020): "The impact of Covid on machine learning and data science in UK banking", *Bank of England Quarterly Bulletin*, Q4.

Blei, D, M Jordan and A Ng (2003): "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, no 3, pp 993–1022.

Blondel, V, J-L Guillaume, R Lambiotte and E Lefebvre (2008): "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, vol 10.

Boudiaf, I, M Scheicher and F Vacirca (2023): "CCP initial margin models: A peek under the hood", *SUERF Policy Brief*, no 624, June.

Borio, C (2013): "The Great Financial Crisis: setting priorities for new statistics", *Journal of Banking Regulation*, vol 14, pp 306–17. Also published as *BIS Working Papers*, no 408, April.

Carvalho, B, S Peralta and J dos Santos (2020): "What and how did people buy during the Great Lockdown? Evidence from electronic payments", *Covid Economics*, Centre for Economic and Policy Research, vol 28.

Cerdeiro, D, A Komaromi, Y Liu and M Saeed (2020): "World seaborne trade in real time: a proof of concept for building AIS-based nowcasts from scratch", *IMF Working Papers*, no 7.

Chen, L, M Zaharia and J Zou (2023): "How is ChatGPT's behavior changing over time?", arXiv:2307.09009 [cs.CL], August, https://doi.org/10.48550/arXiv.2307.09009.

Donoho, D (2017): "50 years of data science", *Journal of Computational and Graphical Statistics*, vol 26, no 4, pp 745–66.

Drozdova, A (2017): "Modern informational technologies for data analysis: from business analytics to data visualisation", *IFC Bulletin*, no 43, March.

Edmond, D, S Bawa, L Garg and V Prakash, (2022): "Adoption of cloud services in central banks: hindering factors and the recommendations for way forward", *Journal of Central Banking Theory and Practice*, vol 11, no 2, pp 123–43, May.

European Securities and Markets Authority (ESMA) (2021): *EMIR and SFTR data quality report*.

Financial Stability Board (FSB) (2017): Artificial intelligence and machine learning in financial services: market developments and financial stability implications, www.fsb.org/wp-content/uploads/P011117.pdf.

Gentzkow, M, B Kelly and M Taddy (2019): "Text as data", *Journal of Economic Literature*, vol 57, no 3, pp 535–74.

Geurts, P, D Ernst, and L Wehenkel (2006): "Extremely randomized trees", *Machine Learning*, no 63, pp 3–42.

Giannone, D, L Reichlin and D Small (2008): "Nowcasting: the real-time informational content of macroeconomic data", *Journal of Monetary Economics*, vol 55, no 4, pp 665–76.

Handel, D, A Ho, K Huynh, D Jacho-Chavez and C Rea (2022): "Cloud computing research collaboration: an application to access to cash and financial services", *IFC Bulletin*, no 57, November.

Irving Fisher Committee on Central Bank Statistics (IFC) (2016): "Central banks' use of the SDMX standard", *IFC Report*, no 4.

——— (2017): "Big data", *IFC Bulletin*, no 44.

——— (2018): "Central banks and trade repositories derivatives data", *IFC Report*, no 7.

——— (2019): "Business intelligence systems and central bank statistics", *IFC Report*, no 9.

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

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

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

——— (2022): "How central banks communicate on official statistics", *IFC Report*, no 15, February.

Jahangir-Abdoelrahman, S and B Tissot (2023): "The post-pandemic new normal for central bank statistics", *Statistical Journal of the IAOS*, vol 39, no 3, pp 559–72.

Križman, I and B Tissot (2022): "Data governance frameworks for official statistics and the integration of alternative sources", *Statistical Journal of the IAOS*, vol 38, no 3, pp 947–55.

Kronenberg, P, M Bannert, H Mikosch, S Neuwirth and S Thöni (2023): "The Nowcasting Lab: live out-of-sample forecasting and model testing", available at SSRN.

Kudugunta, S, A Bapna, I Caswell, C A Choquette-Choo, O Firat, X Garcia, A Kusupati, K Lee, R Stella, D Xin and B Zhang, (2023): "MADLAD-400: A multilingual and document-level large audited dataset", arxiv.org/pdf/2309.04662.pdf.

Le, Q and T Mikolov (2014): "Distributed representations of sentences and documents", *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pp 1188–96, Beijing.

LeCun, Y, Y Bengio and G Hinton (2015): "Deep learning", *Nature*, vol 521, pp 436–44, May.

Lenza, M, I Moutachaker and J Paredes (2023): "Density forecasts of inflation: a quantile regression forest approach", *ECB Working Paper Series*, no 2830.

Loiacono, G and E Rulli (2022): "ResTech: innovative technologies for crisis resolution", *Journal of Banking Regulation*, vol 23, pp 227–43.

Moor, J (2006): "The Dartmouth College Artificial Intelligence Conference: the next fifty years", *AI Magazine*, vol 27, no 4, pp 87–91.

Mouat, A (2015): "Using docker: developing and deploying software with containers", *O'Reilly Media, Inc*.

Organisation for Economic Co-operation and Development (OECD) (2019): "The path to becoming a data-driven public sector", *OECD Digital Government Studies*, OECD Publishing, November.

Prokhorenkova, L, A Dorogush, A Gulin, G Gusev and A Vorobev, (2018): "CatBoost: unbiased boosting with categorical features", *Advances in Neural Information Processing Systems (NeurIPS),* vol 31.

Shirono, K, H Carcel-Villanova, E Chhabra, B Das and Y Fan (2021): "Is mobile money part of money? Understanding the trends and measurement", *IMF Working Papers*, no 117.

Teles Dias, L (2021): "Post-crisis skills landscape: the emergence of 'purple people", *IFC Bulletin*, no 53, April.

Tibshirani, R (1996): "Regression shrinkage and selection via the Lasso", *Journal of the Royal Statistical Society*, Series B (Methodological), vol 58, no 1, pp 267–88.

Tukey, J (1962): "The future of data analysis", *The Annals of Mathematical Statistics*, no 33, pp 1–67.

Vilhuber, L (2021): "Use of Docker for reproducibility in economics", Office of the American Economic Association Data Editor, https://aeadataeditor.github.io/posts/2021-11-16-docker.

Walczak, E (2019): "Data science in the Bank of England: who are we and what do we do?", lecture at Taras Shevchenko National University of Kyiv, May.

Wibisono, O, H Ari, B Tissot, A Widjanarti and A Zulen (2019): "Using big data analytics and artificial intelligence: a central banking perspective", "Data analytics", *Capco Institute Journal of Financial Transformation*, 50th edition, pp 70–83.