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).

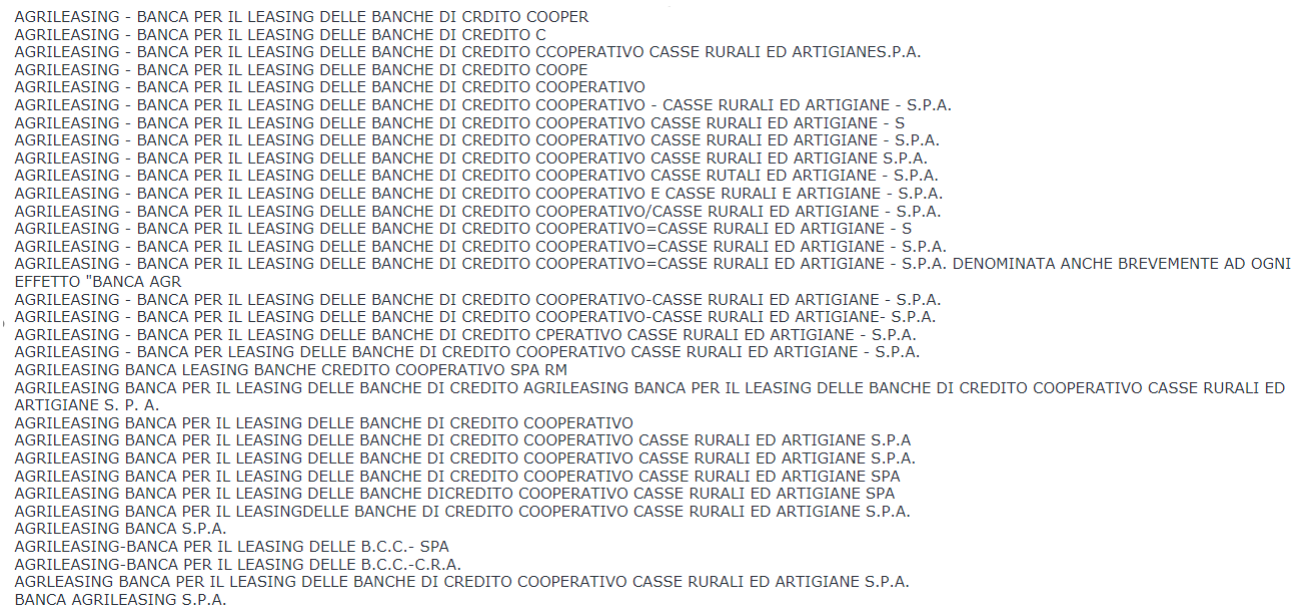
Combining the new granular data in a coherent framework would allow for an even better understanding of the dynamics of the European financial system. Here, some challenges remain. We list three of the main challenges we have encountered and show how we coped with them.

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.



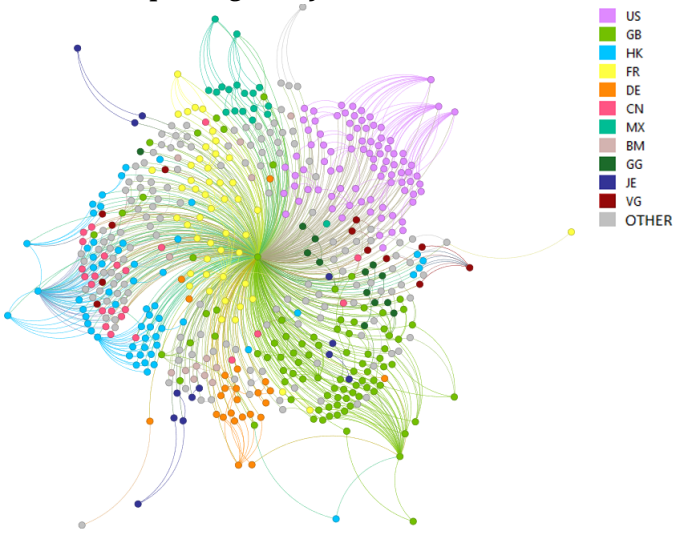
Source: DNB DSH project documentation.

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.

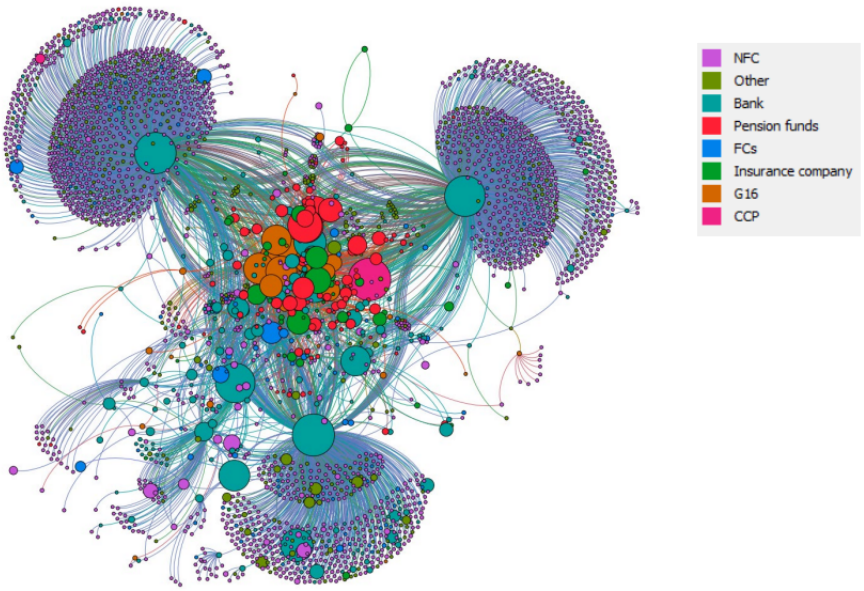
terconnected with three large Dutch banks that facilitate most clients and have access to Central Clearing Parties (CCPs). The cluster in the middle indicates that Dutch pension funds and insurance companies do not trade exclusively with Dutch banks but also use large international banks and CCPs to 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.

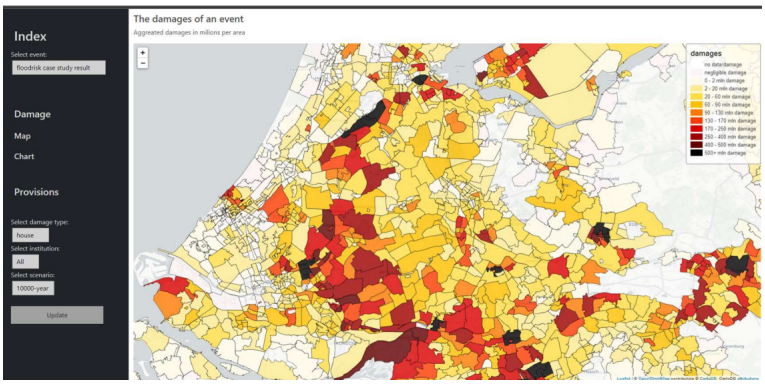
[5] More background information on this project can be found in the summary of the workshop on the role of technology in finance (BIS, 2023).

[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.

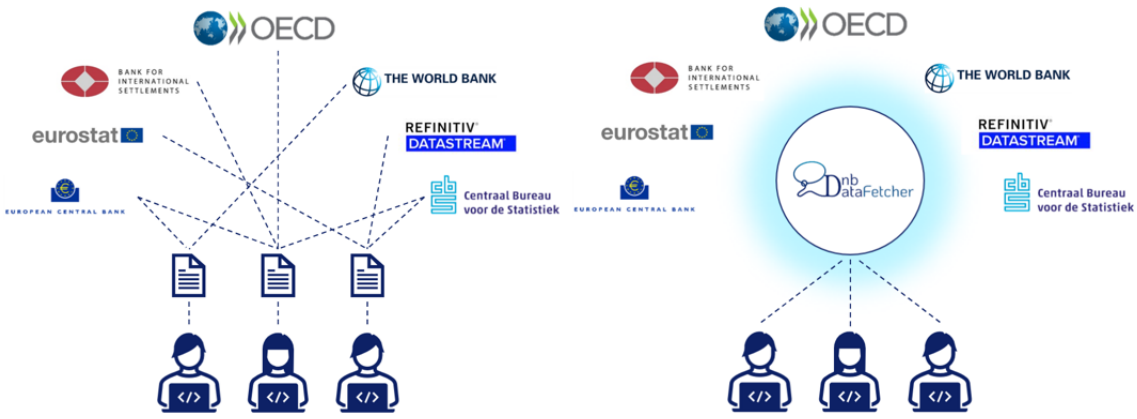
2.3 Automating data processes

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

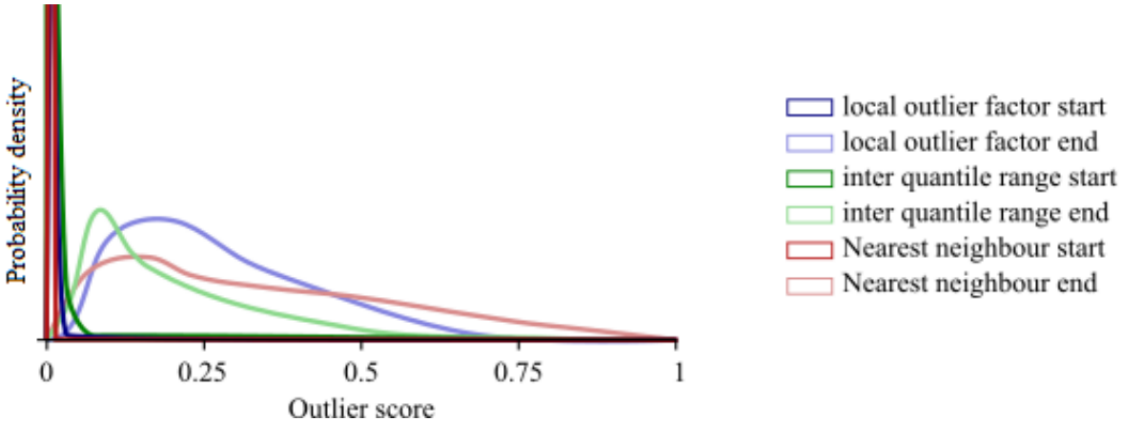
There is a lot to gain from machine learning, but when setting up the DSH we soon realized that starting with fancy machine learning models would not convince end-users to use it. For people without a lot of knowledge of or experience with data science, more complex models may soon be seen as a black 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.

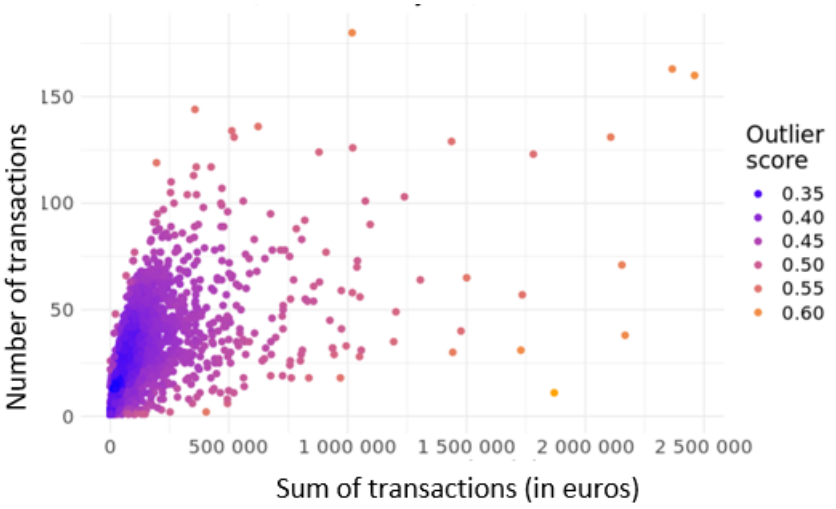
At the Data Science Hub, we are currently implementing the reinforcement learning outlier detection approach for granular data reported by Dutch insurers and pension funds under the Solvency II and FTK frameworks. This application shows that outliers can be identified by the ensemble learner. Moreover, 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-

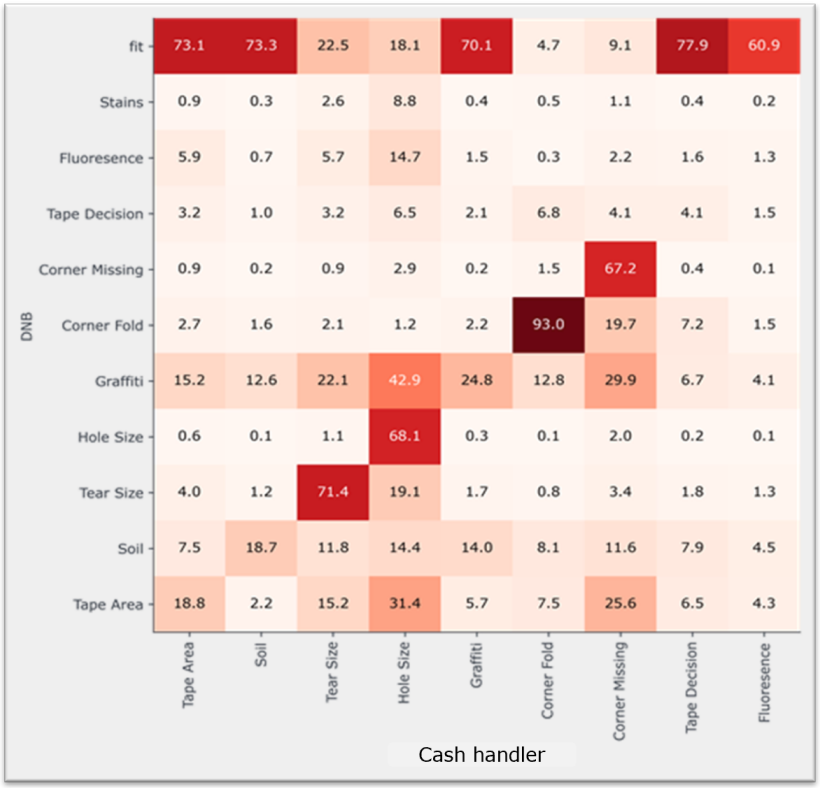
ple accounts on a single address" and "a single deposit per month and immediate withdrawal". Seen separately, these are relatively innocent. Together, however, they can indicate human trafficking of seasonal workers. The combination identifies subcontractors who organize housing for seasonal workers, 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

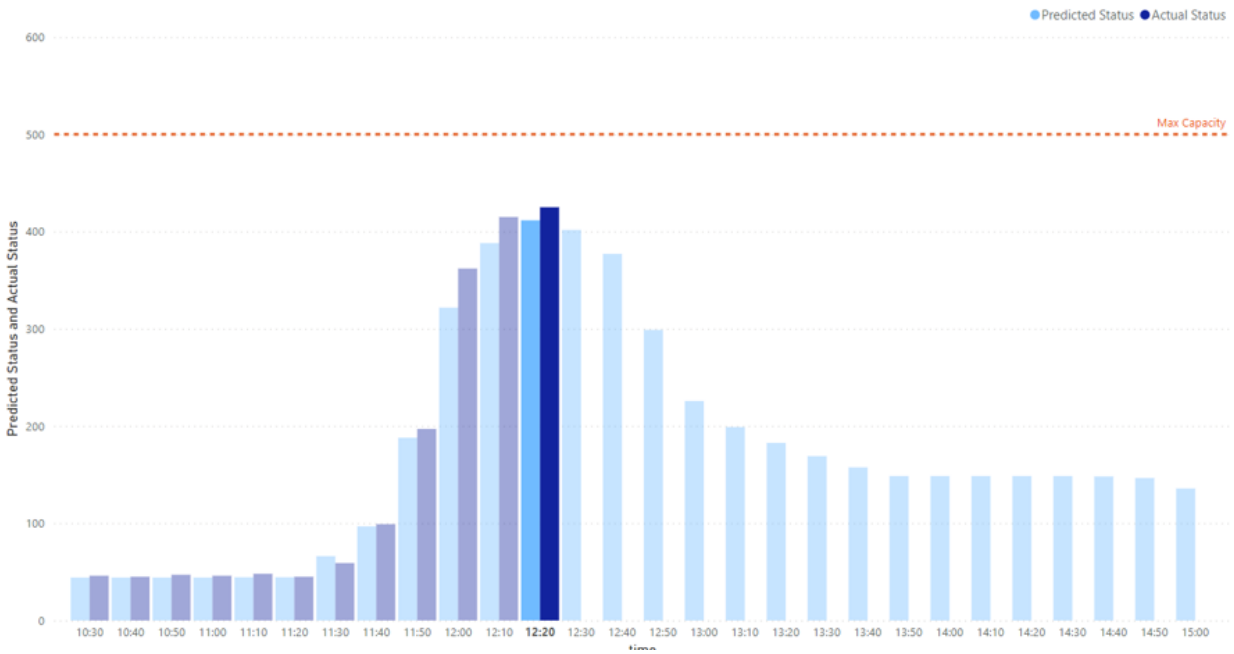
on the size of the project, this takes (a lot of) time. One of our first projects, Dataloop, provides a perfect example of this. Dataloop was initiated as a collaboration between DNB's Statistics Division and the DSH with the goal of improving data quality in supervisory reporting. Dataloop does so by 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

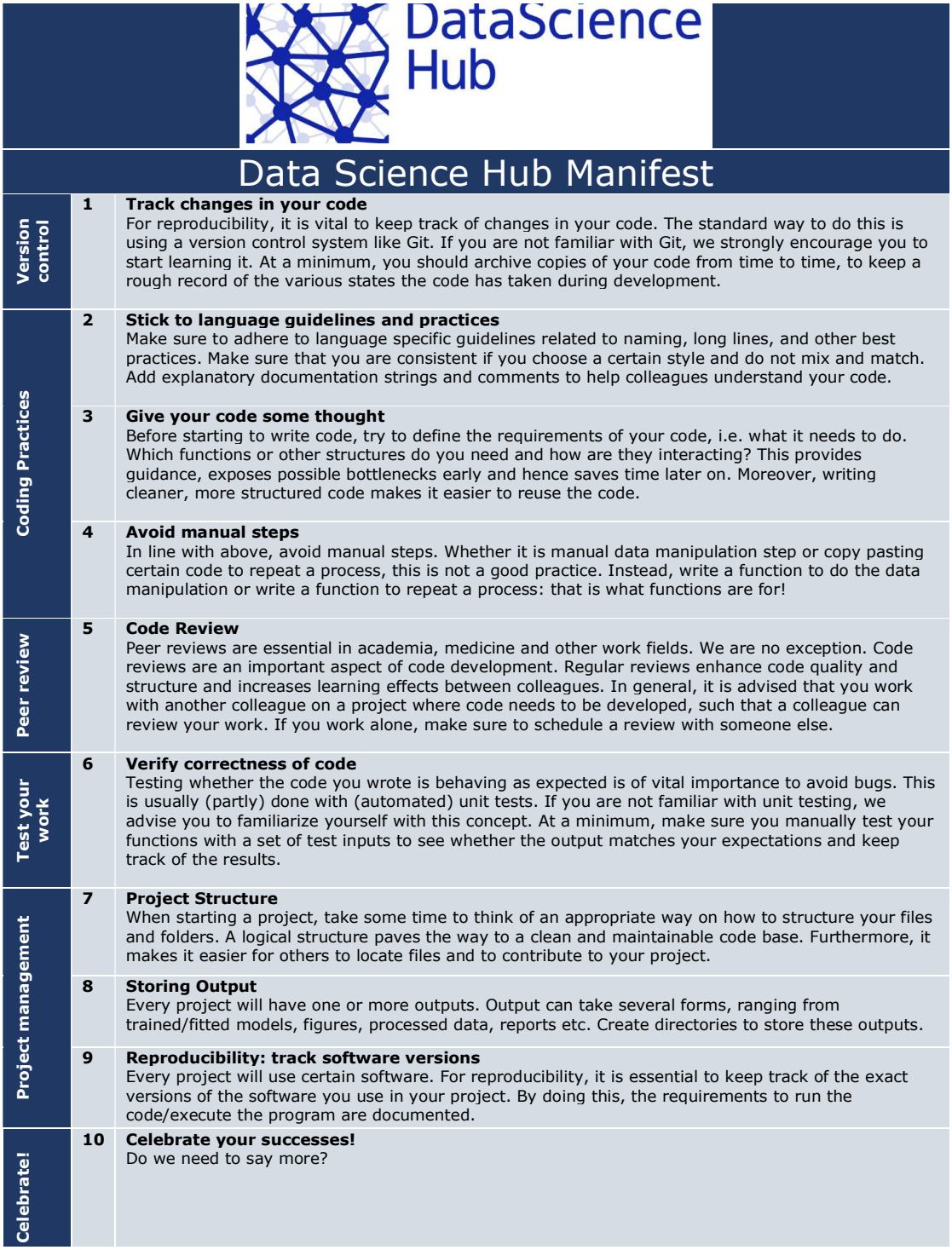
The job title "Data Scientist" is relatively new, and the function that a data scientist fulfills in an organization can vary quite a lot depending on the type of organization. In fully-fledged AI-minded organizations, such as tech start-ups, data scientists may have well-defined responsibilities, such 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

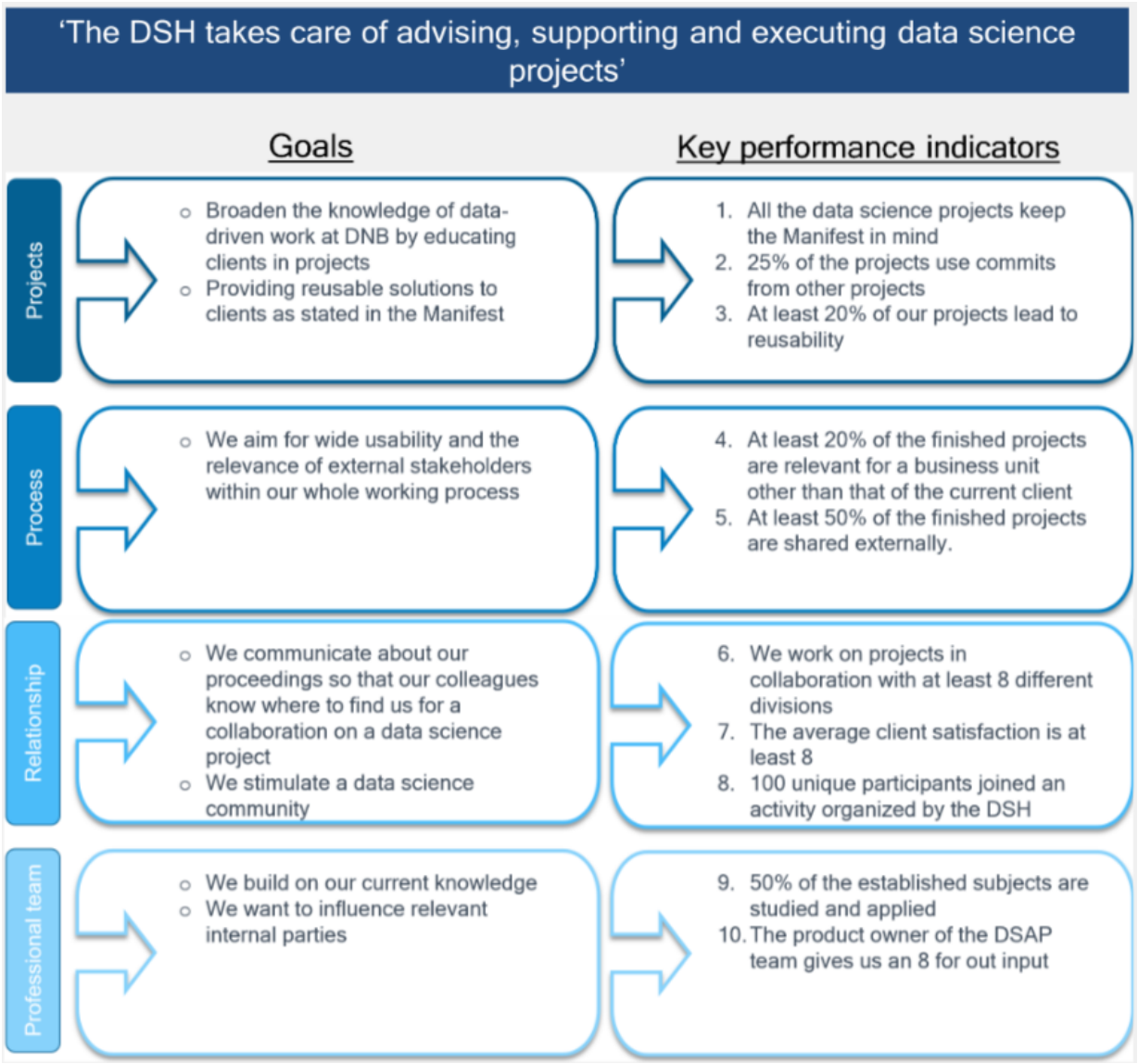
Having a clear role helps to define the mandate of the DSH and position the DSH within the organization. Our mandate is to take care of advising, supporting, and executing data science projects throughout the Dutch Central Bank. As pure data scientists, we love to measure what we do. Therefore, we translated 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**

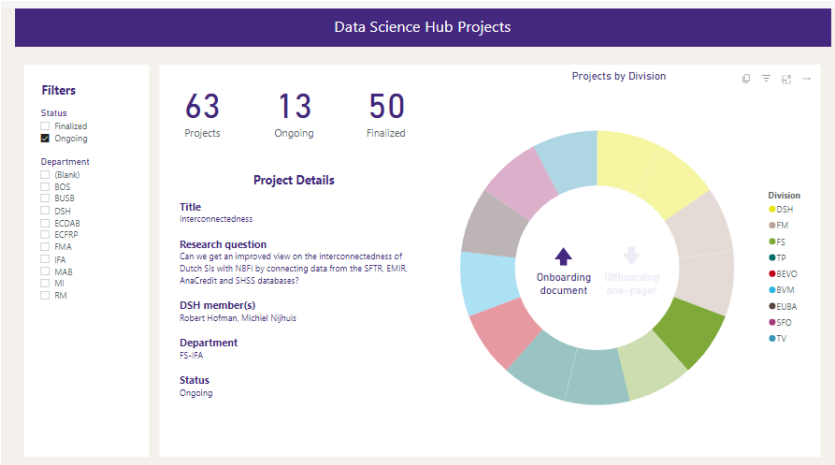


Source: DSH Annual Report 2022

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.

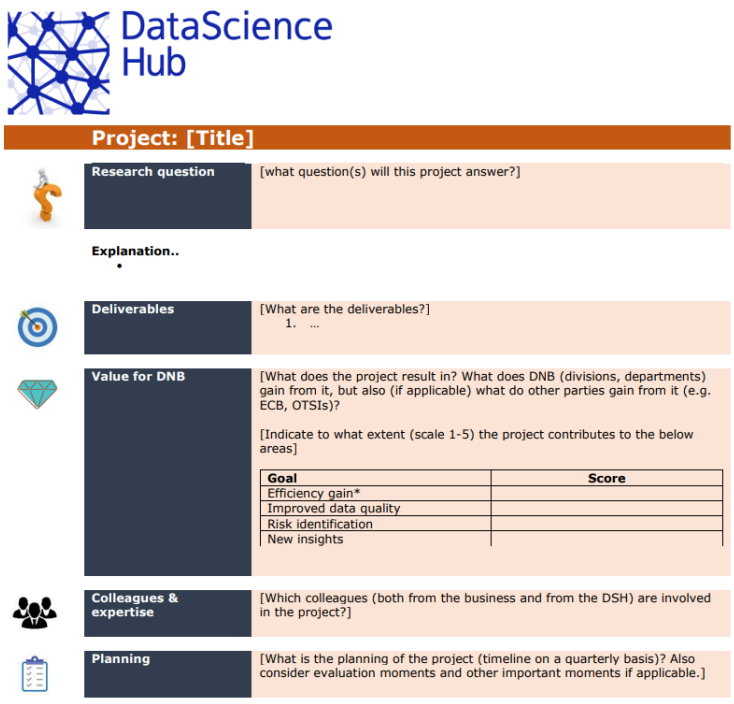
Third, we also communicate our successes externally. A project does not end with celebrating its completion (see Rule number 10 of our Manifest in Figure 10). We actively seek ways to communicate the results to the outside world since projects

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]*

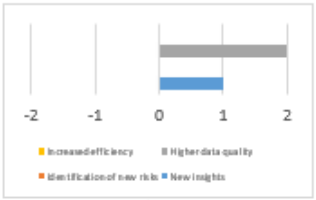
Do you want more information? Contact data_science@dnb.nl

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**.

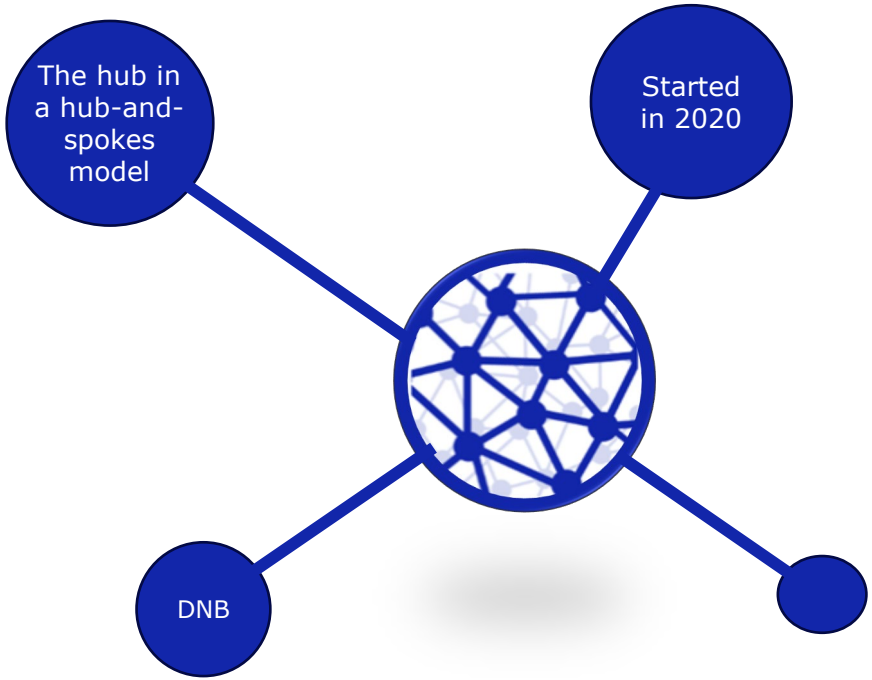
World Economic Forum (2018) The New Physics of Financial Services.

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

Data Science Hub



This presentation

7 Experiences

5 Essentials

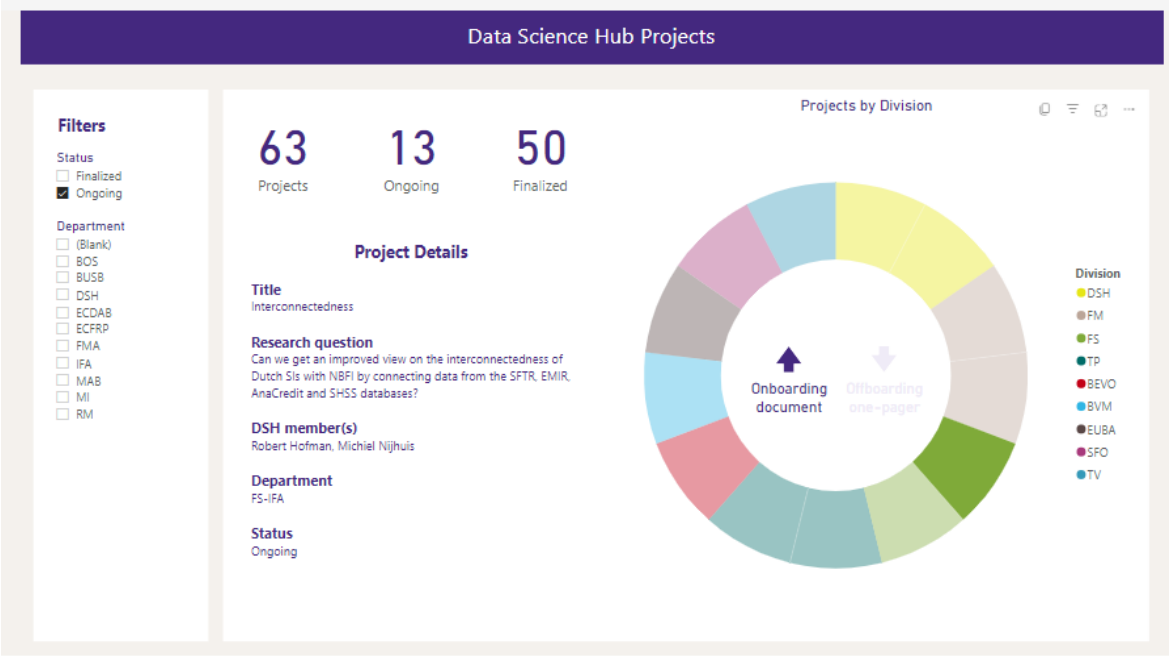
Way of Working

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]

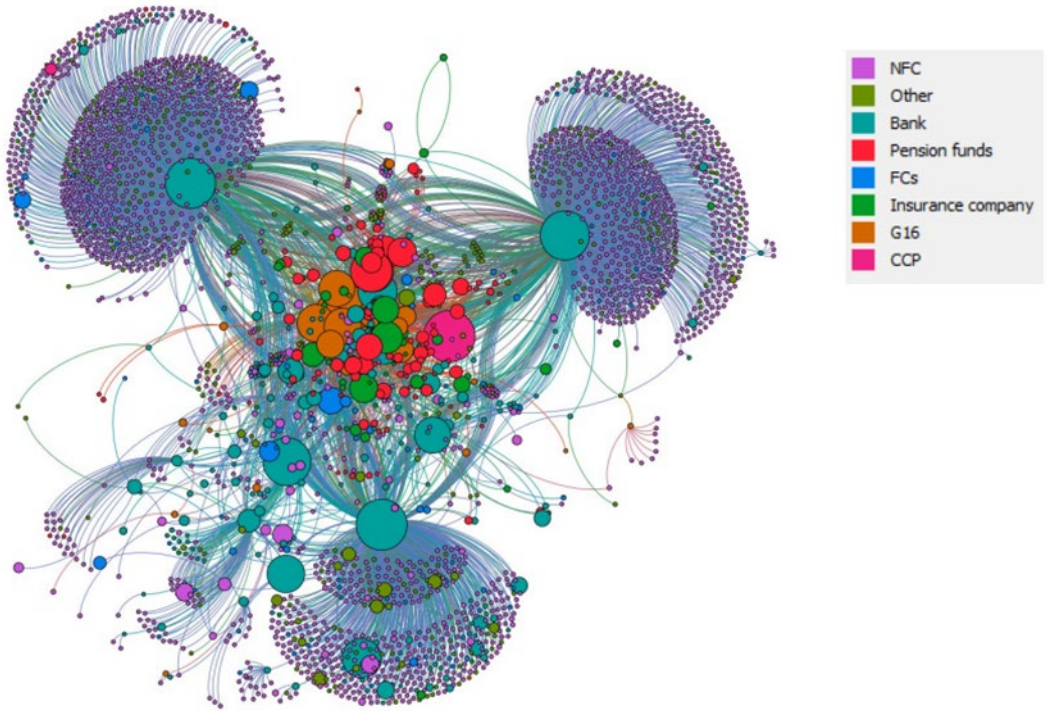
Granular data

*"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

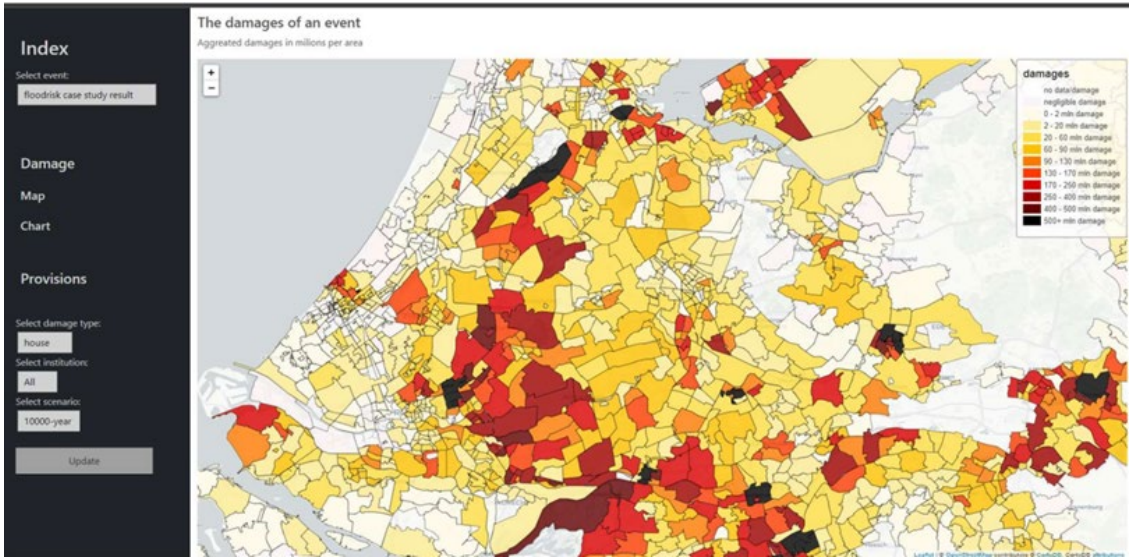
Combining internal and external sources

"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.

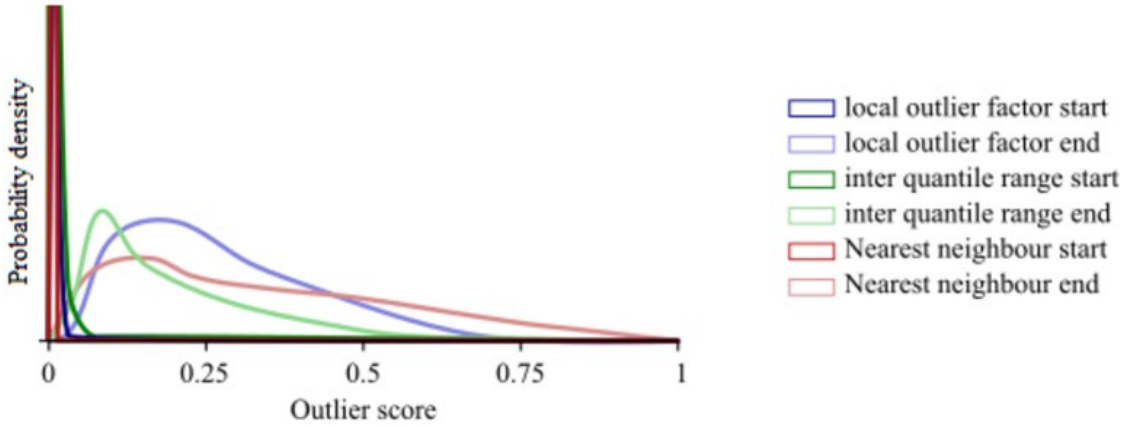
DeNederlandscheBank EUROSYSTEEM

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.

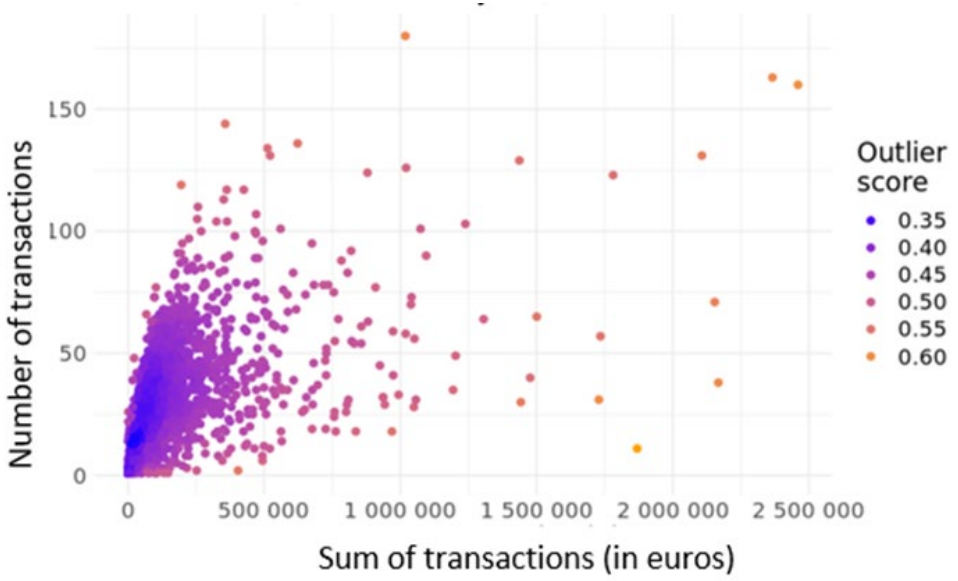
The project is described in more detail in the paper "Outlier Detection with Reinforcement Leanring for Costly to Verify Data" (Nijhuis and van Lelyveld, 2023)

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.

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.

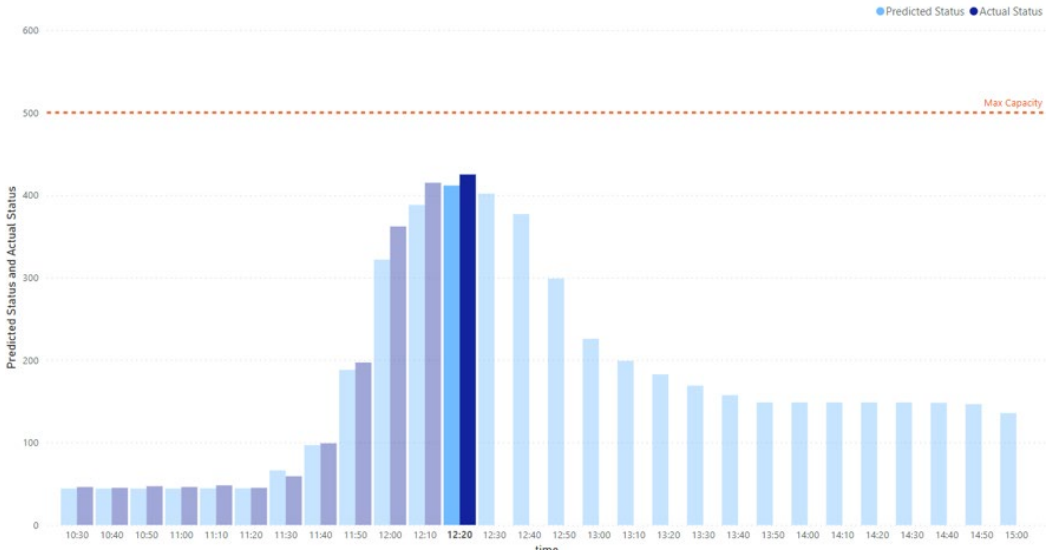
DeNederlandscheBank EUROSYSTEEM DataScience Hub

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.



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.

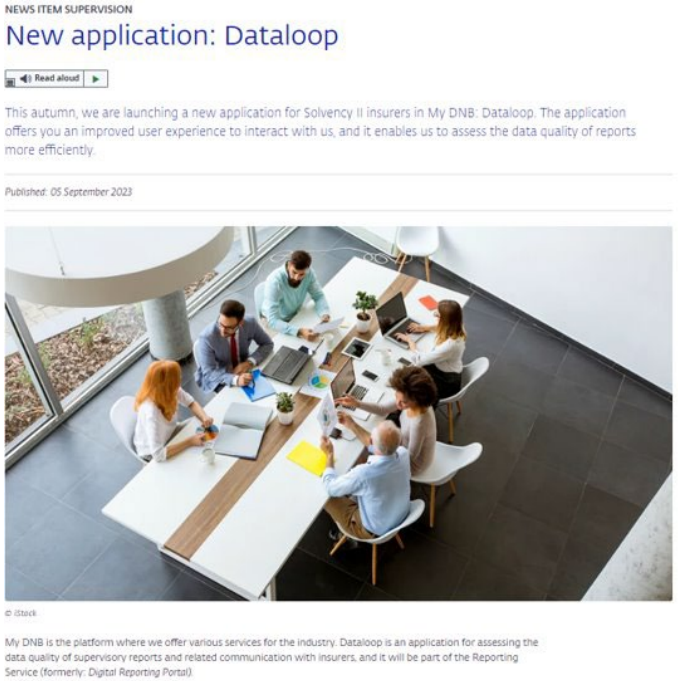
DeNederlandscheBank EUROSYSTEEM DataScience Hub

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

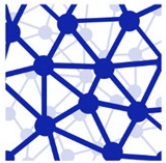
"Responsible coding is needed to ensure the replication and reproducibility of the work."

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."

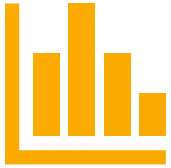


DataScience Hub

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.

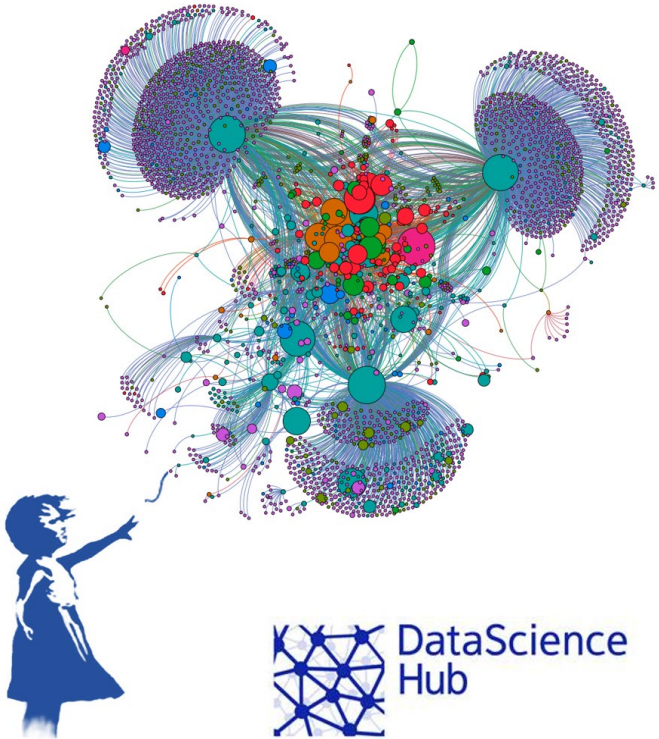
Activities and trainings
Build a data science community by offering regular and ad-hoc events.

DeNederlandscheBank
EUROSYSTEEM

Conclusion

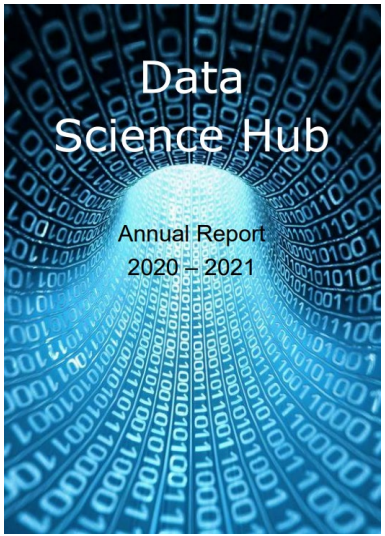
- A well-defined strategy, mandate, goals, and a structured way of working definitely helped us to apply data science within the organization.
- Just start! And don't think that getting data science to work for an organization can be achieved by hiring a few smart data geeks and having them develop "AI" 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