BANK FOR INTERNATIONAL SETTLEMENTS

BIS Working Papers

No 930

Big data and machine learning in central banking

by Sebastian Doerr, Leonardo Gambacorta and Jose Maria Serena

Monetary and Economic Department

March 2021

JEL classification: G17, G18, G23, G32.

Keywords: big data, central banks, machine learning, artificial intelligence, data science.

BIS Working Papers are written by members of the Monetary and Economic Department of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in them are those of their authors and not necessarily the views of the BIS.

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

© *Bank for International Settlements 2021. All rights reserved. Brief excerpts may be reproduced or translated provided the source is stated.*

ISSN 1020-0959 (print)
ISSN 1682-7678 (online)

Big data and machine learning in central banking

Sebastian Doerr, Leonardo Gambacorta and Jose Maria Serena*

Abstract

This paper reviews the use of big data and machine learning in central banking, leveraging on a recent survey conducted among the members of the Irving Fischer Committee (IFC). The majority of central banks discuss the topic of big data formally within their institution. Big data is used with machine learning applications in a variety of areas, including research, monetary policy and financial stability. Central banks also report using big data for supervision and regulation (suptech and regtech applications). Data quality, sampling and representativeness are major challenges for central banks, and so is legal uncertainty around data privacy and confidentiality. Several institutions report constraints in setting up an adequate IT infrastructure and in developing the necessary human capital. Cooperation among public authorities could improve central banks' ability to collect, store and analyse big data.

JEL classification: G17, G18, G23, G32.

Keywords: big data, central banks, machine learning, artificial intelligence, data science.

* Sebastian Doerr, Leonardo Gambacorta and Jose Maria Serena are in the Monetary and Economic Department (MED) at the Bank for International Settlements (BIS). We would like to thank Fernando Perez Cruz for his technical advice and input. For comments and suggestions we also thank Gianni Amisano, Douglas Araujo, Claudio Borio, Agustin Carstens, Stijn Claessens, Jon Frost, Julian Langer, Michel Jiullard, Juri Marcucci, Luiz Awazu Pereira, Rafael Schmidt, Hyun Song Shin and Bruno Tissot. Giulio Cornelli provided excellent statistical assistance. The views expressed are those of the authors and not necessarily those of the BIS.

1. Introduction

The world is changing and so is the way it is measured. For decades, policymakers and the private sector have relied on data released by official statistical institutions to assess the state of the economy. Collecting these data require substantial effort and publication often happens with a lag of several months or even years. However, the last years have seen explosive growth in the amount of readily available data. New models of data collection and dissemination enable the analysis of vast troves of data in real time. We now live in the "age of big data".[1]

One major factor in this development is the advent of the information age, and especially the smart phone and cloud computing: individuals and companies produce unprecedented amounts of data that are stored for future used in servers of technology companies. For example, billions of Google searches every day reveal what people want to buy or where they want to go for dinner. Social media posts allow market participants to track the spread of information in social networks. Companies record every step of their production or selling process, and electronic payment transactions and e-commerce create a digital footprint.

An additional catalyst in the creation of big data, especially financial data, has been the Great Financial Crisis (GFC) of 2007-09. The GFC laid bare the necessity of more disaggregated data: a relatively small but interconnected bank such as Lehman Brothers could bring down the financial system because it was highly interconnected. The regulation and reporting requirements set up after the GFC have increased the data reported to central banks and supervisory authorities – and further work to enhance central bank statistics is in progress (Buch (2019)).

The advent of big data coincides with a quantum leap in technology and software used to analyse it: artificial intelligence (AI) is the topic *du jour* and enables researchers to find meaningful patterns in large quantities of data. For example, natural language processing (NLP) techniques convert unstructured text into structured data that machine-learning tools can analyse to uncover hidden connections. Network analysis can help to visualise relations in these high-dimensional data. For the first time in history, it is possible to produce a real-time picture of economic indicators such as consumer spending, business sentiment or people's movements.

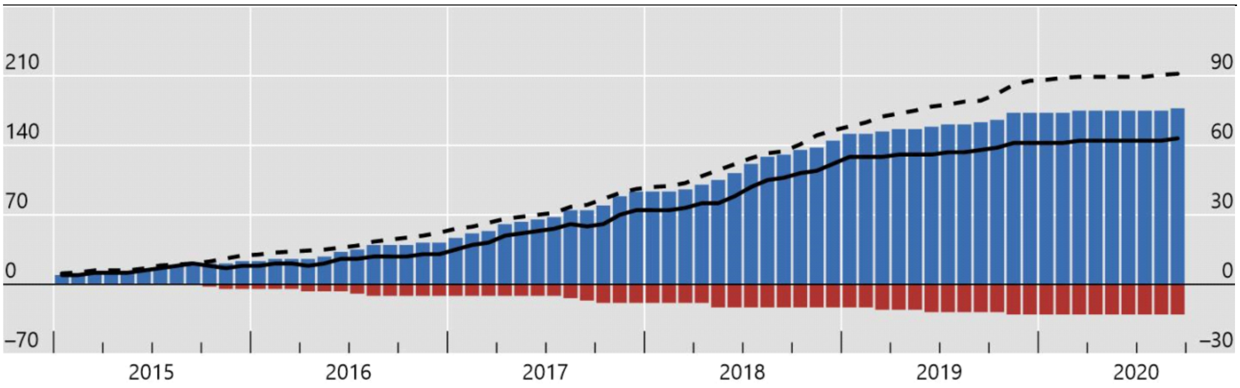
These developments have spurred central banks' interest in big data. Rising interest is reflected in the number of central bank speeches that mention big data and do so in an increasingly positive light (Graph 1). And yet, big data and machine learning pose challenges – some of them more general, others specific to central banks and supervisory authorities.

This paper reviews the use of big data and machine learning in the central bank community, leveraging on a survey conducted in 2020 among the members of the Irving Fischer Committee (IFC). The survey contains responses from 52 respondents from all regions and examines how central banks define and use big data, as well as which opportunities and challenges they see.

[1] Forbes (2012): "The Age of Big Data", accessed 12 June 2020.

Central banks' interest in big data is mounting[1]

Number of speeches | Graph 1



Cumulative count of speeches:

Lhs: Total   Rhs:[2] Net, positive–negative   Positive stance   Negative stance

[1] Search on keyword "big data". [2] The classification is based on the authors' judgment. The score takes a value of –1 if the speech stance was clearly negative. It takes a value of +1 if the speech stance was clearly positive or a project/pilot using big data has been conducted. Other speeches (not displayed) have been classified as neutral.

Sources: Central bankers' speeches; authors' calculations.

The survey uncovers four main insights, First, central banks define big data in an encompassing way that includes unstructured non-traditional data sets, as well as structured data sets from administrative sources or those collected due to regulatory reporting requirements.

Second, central banks' interest in big data has markedly increased over the last years. Comparing answers from the 2020 survey to its 2015 vintage, around 80% of central banks discuss the topic of big data formally within their institution, up from 30% in 2015. Further, over 60% of respondents report a high level of interest in the topic of big data at the senior policy level. The discussion on big data in central banks focuses on a wide range of topics. A key topic of discussion is the availability of big data and tools to process, store and analyse it. The design of legal frameworks, for example in defining access rights to confidential data, or aspects of cyber security are also at the centre of central bankers' interest.

Third, beyond discussions, there is also action: in contrast to 2015, the vast majority of central banks are now conducting projects that involve big data. Among the institutions that currently use big data, over 70% use it for economic research, while 40% state that they use it to inform policy decisions. Several institutions use big data in the areas of financial stability and monetary policy, as well as for suptech and regtech applications. Around two thirds of respondents want to start new big data-related projects in 2020/21.

And fourth, the advent of big data poses new challenges. Several central banks report that cleaning the raw data (eg in the case of data obtained from newspapers or social media), sampling and representativeness (eg in the case of data based on Google searches or employment websites), or matching new data to existing sources are obstacles to the usefulness of big data for central banks. Another often-cited challenge refers to legal aspects around privacy and confidentiality, especially with respect to data from non-traditional sources such as web pages. For example, central banks grapple with ethics and privacy issues that accompany the use of potentially sensitive data acquired through public sources or via web scraping. Finally, central banks also need to tackle more practical problems: the vast majority report that they face budget constraints and have difficulty in training existing or hiring new staff to work on big data-related issues. They also report that setting up the adequate IT infrastructure proves challenging.

The existing literature investigating central banks' use of big data mainly focuses on individual countries. A notable exception is the survey conducted by the IFC (2015) among its central bank members in 2015.[2] Against this backdrop, this paper provides an updated assessment of how central banks define big data, how their interest in and use of big data has evolved over the last years also using machine learning techniques and which challenges central banks face in collecting, storing, and analysing it.

The rest of the paper is organised as follow. Section 2 provides an overview of how central banks define big data. Sections 3 illustrates in which fields central banks use or plan to use big data and discusses specific use cases. Section 4 discusses opportunities and challenges for central banks and supervisory authorities in the use of machine learning and big data. Section 5 discusses how cooperation among public authorities could relax the constraints on collecting, storing and analysing big data. Section 6 concludes.

2. How do central banks define big data?

Big data is commonly defined in terms of volume, velocity and variety (the so-called 3Vs). For data to be "big", they must not only have high volume and high velocity, but also come in multiple varieties.[3]

Central bank definitions of big data reflect these characteristics. Around one third of the respondents in the survey define big data exclusively as large non-traditional or unstructured data that require new techniques for the analysis (Graph 2, left-hand panel). The remaining two thirds also include traditional and structured data sets in their definition of big data. No central bank considers traditional data alone as big data. These proportions are similar across advanced and emerging market economies, denoted in blue and red.

The encompassing definition is reflected in the variety of raw data sources used for analysis. The right-hand panel in Graph 2 shows a word cloud with the most-frequently used sources, as reported by central banks in the survey. These range from structured administrative data sets such as credit registries to non-traditional data obtained from newspapers and online portals or by scraping the web. A promising avenue for central bankers and policymakers is to complement traditional data sources with non-traditional data sources to inform policy decisions.

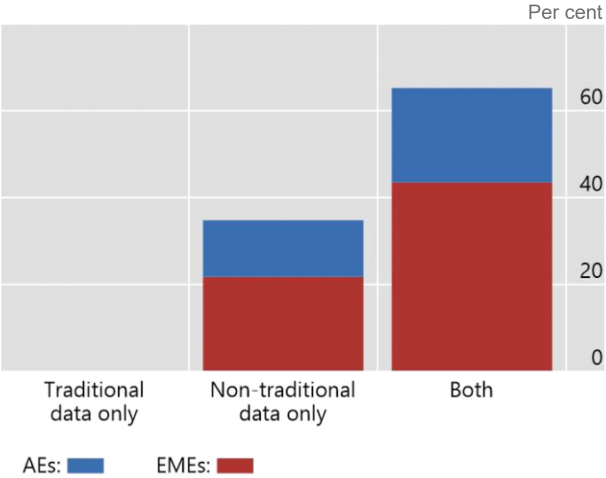
[2] See IFC (2018) for a collection of country experiences. See also Cœuré (2017) and Tissot (2015).

[3] Occasionally, veracity is added as a fourth V, as big data is often collected from open sources (Tissot (2019)).

Central bank definitions of big data and main sources

Graph 2

How does your institution define big data?[1]



Word count on sources[2]

[1] The graph reports the share of respondents that selected each respective answer to the question "How does your institution define big data?". Respondents could select multiple options. Specifically, 35% of respondents consider only non-traditional data as big data. Non-traditional data include unstructured data sets that require new tools to clean and prepare, data sets that have not been part of your traditional pool and data sets with high-frequency observations and/or a large number of cross-sectional units. The remaining 65% additionally consider structured traditional databases as big data. [2] The word cloud highlights the most-frequent terms mentioned by central banks in their current and future big data projects. In a first step, open answers are transformed by removing special characters, white spaces or stopwords (such as "the", "a" or "we"). A text-mining algorithm then counts the frequency of individual words. Words mentioned more frequently appear larger.

Source: IFC (2020); authors' calculations.

Central banks have substantial experience with large structured data sets, typically of a financial nature, but have only recently started to explore unstructured data. Financial data sets are analysis-ready, as they are generally collected for regulatory purposes and adhere to reporting requirements. A catalyst in the creation of financial data has been the Great Financial Crisis (GFC), which laid bare the necessity of disaggregated data (IFC (2011)). The ensuing regulation and reporting requirements have increased the data reported to central banks and supervisory authorities – and further work to enhance central bank statistics is in progress (Buch (2019)). By contrast, unstructured data are often the by-product of corporate or consumer activity. Before they are analysed, they must be cleaned and curated, ie organised and integrated into existing structures.

A short introduction to machine learning

Machine learning is a subfield of artificial intelligence that focuses on learning and extracting knowledge from data. By leveraging data collected by central banks or available from other sources, machine learning can provide real-time insights about eg inflation or consumer spending. Broadly speaking, machine learning involves the development of algorithms that use large amounts of data to autonomously infer their own parameters. Machine learning is usually classified into supervised and unsupervised learning.

Supervised learning algorithms take an input (eg the characteristics of a house) to predict its most likely outcome (eg house price). An algorithm is a "classifier" if the outcome is countable and a "regression" otherwise.[4] To predict these outcomes, the machine learning algorithm relies on the so-called labelled data, which consists of pairs of inputs and outputs that have been sampled from the underlying data. The parameters of any machine learning algorithm are tuned to fit the labelled data by minimizing a loss function, eg the classification loss or R^2.

Machine learning algorithms are either non-parametric, ie the number of parameters linearly grows with the number of training points, or overparametrised, ie the models have more parameters than available data. In the latter case, these parameters are typically meaningless and machine learning algorithms are evaluated on a test set. The test set is a labelled data set that has the same distribution as the training set, but has not been used in training the algorithm itself.

Given the overcapacity of machine learning algorithms, an unconstrained minimization of the desired loss function would yield an loss of zero in the training data (ie a perfect fit). When this occurs, the model is usually poor at predicting the outcome for data points outside of the training sample, ie its results do not generalise beyond the training data, rendering the model useless. This problem is known as overfitting. Machine learning algorithms combat overfitting in several ways. First, they add a penalty to the loss function, called a regulariser, that enforces the solution to be smooth or to use fewer parameters, thus seeking to uncover more general patterns from the data. A further approach is to divide the training data into two parts: a training and a validation set. While the former is used to train the algorithm (subject to some regulator), the latter is used to verify its fit.

For example, let's assume we want to predict the house prices based a set of 100 indicators and we have 100,000 samples of those indicators and the corresponding prices of those houses. We could divide the available data into 80% for training, 10% for validation and 10% for testing. We will train our different models with the training data and use the 10% validation data to select the model that yields the best fit on this data set. Finally, we use the test set to have an independent evaluation of the model, because the validation set error will be biased (overoptimistic).

Classic machine learning algorithms, like support vector machines or random forests, typically require to preprocess the inputs before they could be used. This process is referred to as feature engineering and it relies on human experts that understand the problem at hand for designing those features that could subsequently be exploited by the machine learning algorithm to provide accurate solutions. In the house price example above, feature engineering could involve merging the location of the property with other data sets to include a feature related to proximity to schools. When the human knowledge is deficient or incomplete, the performance of classic machine learning algorithm would reflect these initial shortcomings.

Deep Learning (DL) algorithms are a subset of machine-learning techniques that solve the problem of relying on human experts by learning itself the features of the data together with the classification or regression task. Deep learning architecture is biologically inspired by the structure of the brain, especially the visual cortex, in which

[4] Supervised learning is sometimes distinguished into *supervised learning for classification and supervised learning for prediction*. Simply put, if the outcome is categorical (face yes/no, colour blue yes/no), then it is a classification problem. If the outcome is continuous (house prices, height, etc), then it is a regression problem (Athey and Imbens (2019)).

layers or neurons process the data sequentially using non-linear functions to extract insights from data. In each layer of these neural networks, larger patches of the images are processed, and the last layer has access to all the relevant features to make a decision.[5] The structure of the layers of the neural networks are typically tuned to the kind of data that is being used. One of the limitations of deep learning is the need for large number of training examples, because besides finding the optimal classifier it also needs to find the fitting features for it.

Supervised machine learning is useful when the problem needs to be solved is already known. The bottleneck of supervised machine learning is collecting a sufficient amount of label data, which need to be hand-labelled by experts. Supervised machine learning is useful for automatising a task that a human can solve but that it is tedious, and a machine can do much faster and reliable than a human, given a sufficiently accurate algorithm. Supervised machine learning thus solves a problem with a known solution, so the performance of the algorithm is relatively straightforward to evaluate.

On the other hand, unsupervised machine learning algorithms only rely on the input data, without a labelled target dataset, and the goals is understanding the underlying structure of the data independently. Clustering, latent variable models or submanifold mapping are prototypical algorithms in unsupervised learning. These algorithms cannot be straightforwardly evaluated, because they require a human to analyze the extracted features and bring meaning to the low dimensional representations that have been achieved. In this sense, typical unsupervised learning models require deep knowledge of the available data to be able to create models that can extract meaning from it, as well as to interpret their output. Over the last years, there has been efforts to further develop neural networks for unsupervised machine learning that avoid the need for human input.[6]

Natural Language Processing, a common set of applications of machine learning, is used to extract information from written texts. NLP is solving tasks such as sentiment analysis or machine translation, which are posed as a supervised-learning problem. For example, texts are categorised as being positive or negative instance, or texts from two languages matched to be translated.[7] Unsupervised machine learning is used for topic modeling that can then be interpreted by humans when analysing big collection of documents like newspaper articles or scientific papers. An important impact that machine learning is having in natural language processing is creating language models, in which words are mapped to a space in which the Euclidean distance captures the meaning of and relation among words.

---

[5] Current artificial neural networks structures have been optimised for speed and the application at hand and they are far from biological neural networks functioning and structure.

[6] Generative Adversarial Networks (GANs), Variational Autoencoder (VAEs) and Normalizing Flows (NFs) promise to deliver a universal simulator based on training data. Even though they have brought some successes in some applications, widespread applicability in finance is not in the horizon yet.

[7] For an application, see amonsgst others, Amstad et al (2021) who develop a novel trade sentiment index that assesses the positive or negative tone of the Chinese media coverage, and evaluates its capacity to explain the behaviour of 60 global equity markets.

3. How do central banks use big data?

Over the last decade, big data and machine learning have permeated almost all sectors of society, including economic and financial analysis. Private financial institutions have rapidly incorporated big data applications into their toolkit. Insurance companies routinely use AI to better measure client characteristics and improve their pricing schemes. Within credit institutions, applications are frequent and range from credit risk analysis to fraud detection and compliance. Big data is also used by private research desks - eg to forecast inflation or output, especially in countries that lack reliable official statistics.

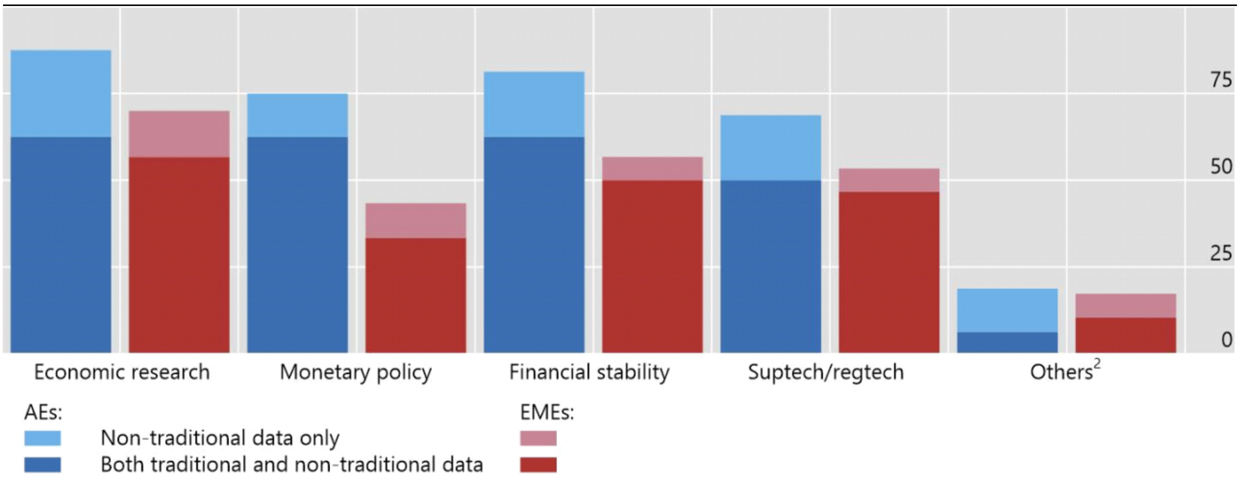
Central banks and supervisory authorities are also making active use of big data. According to the 2020 IFC survey, the fraction of central banks currently using big data has risen to 80%, up from just 30% in 2015. Over four-fifth of respondents indicate that they discuss big data issues extensively. 60% of respondents report a high to very-high level of interest at the senior policy level.

Big data is used in a variety of areas, including research as well as monetary policy and financial stability. Central banks also use it for supervision and regulation (suptech and regtech), often reflecting their specific mandates. Advanced economy central banks (represented by the blue bars in Graph 3) appear to use big data more than their peers in emerging market economies (red bars).

Purposes for which central banks use big data[1]

In per cent

Graph 3



[1] The graph reports the share of respondents that selected each respective answer to the question "For what general purposes does your institution use big data?". Respondents could select multiple options. The different shades used in the bars indicate how institutions define big data (see note to left-hand panel of Graph 2). [2] Includes "monitoring crypto assets", "cyber security" and "network analysis".

Source: IFC (2020).

A large and increasing number of central banks support their economic analysis with nowcasting models drawing on big data. These models produce high-frequency forecasts that can be updated almost in real time. For example, the GDPNow forecast of US real GDP growth by the Federal Reserve Bank of Atlanta is updated up to seven times a month (Higgins (2014)). The Weekly Economic Index of the Federal Reserve Bank of New York (Lewis et al (2020)) provides a weekly estimate of economic activity based on a large number of series (eg railroad traffic or retail sales).

While nowcasting is employed primarily to forecast GDP or private consumption, its use is gradually advancing in other areas. For example, some models predict inflation dynamics from online retail sales and others unemployment from online job portals. The models often rely on both structured and unstructured data. For example, D'Amuri and Marcucci (2017) nowcast unemployment using Google search data. Kalamara et al (2020) show that newspaper text can improve forecasts of macroeconomic variables. Nowcasting and real-time economic indicators are particularly useful in times of heightened uncertainty or economic upheaval, such as the Covid-19 crisis (Box A).

Central banks also use big data techniques to measure other aspects of economic activity. For example, they have used natural language processing to produce economic or policy uncertainty indices from textual data (Baker et al (2016)) or to gauge sentiment in response to monetary policy announcements, including those for unconventional policy measures (Hansen and McMahon (2016)).

Financial big data are widely used to support financial stability analysis (Cœuré (2017), Draghi (2018)). These data include large proprietary and structured data sets, such as those from trade repositories for derivatives transactions, or from credit registries for loans or individual payments. For instance, trade repositories have helped identify networks of exposures. Credit registries support the assessment of credit quality, eg by improving estimates of default probabilities or loss-given-default.[8] Real-time gross settlement system data help to show bank-firm interconnections through their payments.

Many suptech and regtech applications are still exploratory, but could become standard (Broeders and Prenio (2018), Financial Stability Board (2020)). Natural language processing augments traditional credit scoring models, drawing on information from news media or financial statements. It also helps validate compliance with disclosure requirements. Regulators also use machine learning for consumer protection. For example, they assess misconduct risk among financial institutions or screen contracts for suspicious terms and conditions. Big data algorithms are also starting to be used for the detection of fraudulent payment transactions and to combat money laundering.

Of course, economic agents adjust to new technologies. For example, Cao et al (2020) show that firms are aware that their filings are parsed and processed for sentiment via machine learning. Consequently, they avoid words that computational algorithms perceive as negative. This will bias any analysis based on them.

---

8 Credit registries also help banks to better assess credit qualtity and extend loans to safe borrowers who had previously been priced out of the market, resulting in higher aggregate lending (Pagano and Jappelli 1993), and furthering financial inclusion. Credit registries also reduce moral hazard problems by increasing borrowers' cost of default, thus increasing debt repayment (Padilla and Pagano 2000). Conversely, sharing of credit-related information has the benefit of reducing the information monopoly a lender has on its borrowers.

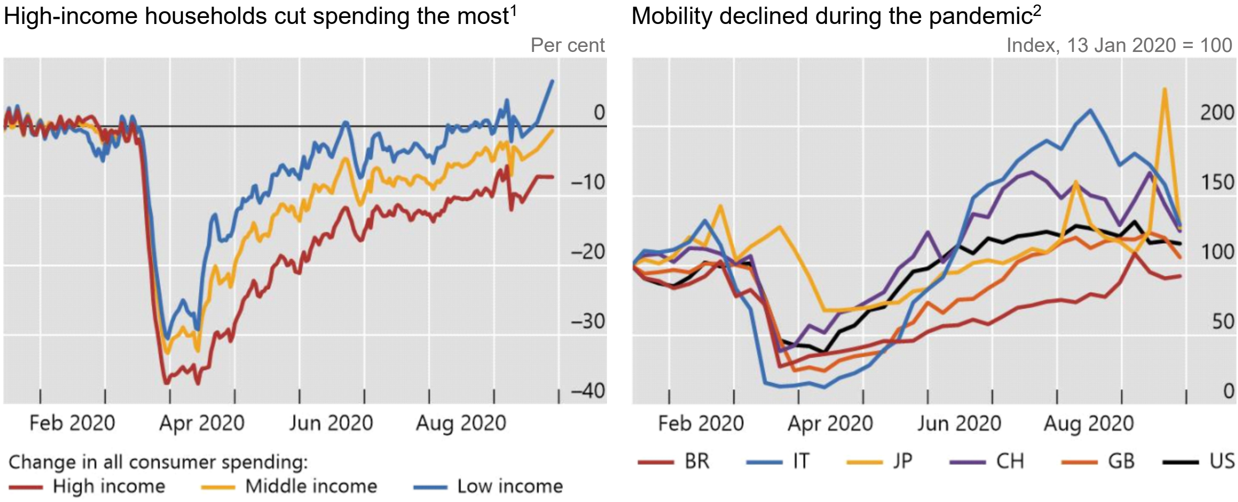
Box A. Using big data during the Covid-19 pandemic

To understand the real-time impact of the Covid-19 pandemic on economic activity and to guide policy decisions, researchers are experimenting with new data and methods. GDP nowcasting has been extensively used during the Covid-19 crisis (Foroni et al (2020)). Examples include nowcasts by the Federal Reserve Banks of New York and Atlanta, which incorporate a range of traditional macroeconomic data, often from monthly surveys, as well as unstructured data. The Central Bank of Brazil uses unstructured information from news or businesses to support nowcasting. These high-frequency forecasts allow policymakers to get a more accurate picture of economic and financial developments, as compared with forecasts based on traditional data that are usually published with a lag.

A set of academic papers have examined spending patterns across regions and population subgroups. Often containing billions of observations, the data are obtained from private companies, such as banks or app providers. For the United States, Chetty et al (2020) construct daily indices on consumer spending and other indicators disaggregated by zip code, industry and income based on data from private companies. They show that high-income households have reduced spending by more than low-income households (Graph A1, left-hand panel). These spending cuts have contributed to job losses among low-income households working for businesses that cater to high-income households. Baker et al (2020) show that greater social distancing coincides with larger drops in spending. Chakrabarti et al (2020a,b) at the Federal Reserve Bank of New York use high-frequency data to investigate changes in consumer spending and business revenue in response to state re-openings. Similar studies have found that household spending declines by more among Danish households with higher health risks (Andersen et al (2020)), or that consumption baskets converge towards the goods basket of low-income households in Spain (Carvalho et al (2020)).

Other studies use publicly available information from newspapers, internet searches and mobility indices to assess the impact of the pandemic. Uncertainty indices developed by the Atlanta Fed and the Bank of England, among others, are based on a variety of sources, such as newspapers, and show that uncertainty spiked during the pandemic to a different extent across countries (Altig et al (2020)). Using earnings call transcripts, Hassan et al (2020) develop measures of the risks that listed firms in more than 80 countries associate with the spread of Covid-19. They show that, initially, firms' primary concerns related to the collapse in demand, as well as increased uncertainty and disruption in supply chains. As the pandemic-induced recession deepened, financing concerns became more prominent. Doerr and Gambacorta (2020a,b) and Wolski and Wruuck (2020) examine data on internet search queries from Google Trends to investigate the impact of the pandemic on local labour markets in the United States and Europe. For their part, Chen et al (2020) use data on mobility indices, based on mobile phone geolocation information provided by Apple and Google, to assess the impact of Covid-19 on people's movements in various countries (Graph A1, right-hand panel).

Consumer spending and mobility during Covid-19 — Graph A1



[1] Change in average credit and debit card spending for the United States. Based on data from Affinity Solutions. Income groups are classified by household zip code. [2] Apple Covid-19 Mobility Trends Reports for the category "driving". The Mobility Trends index reflects daily requests for directions in Apple Maps and is standardised to 100 on 13 January 2020.

Sources: Tracktherecovery.org; covid19.apple.com/mobility; authors' calculations.

4. Challenges in the use of big data

As the wide variety of examples illustrates, big data offers vast opportunities: central banks and supervisory authorities already use big data and machine learning for research purposes, to inform monetary policy decisions and for regulation and supervision. However, the use of big data poses various challenges for central bankers and policymakers. Graph 4 shows the main topics under discussion in central banks. These include the availability of IT infrastructure and human capital, legal and privacy issues, as well as the availability and strategic use of big data. These topics are discussed to a similar extent among central banks in advanced economies (in blue) and emerging market economies (in red).

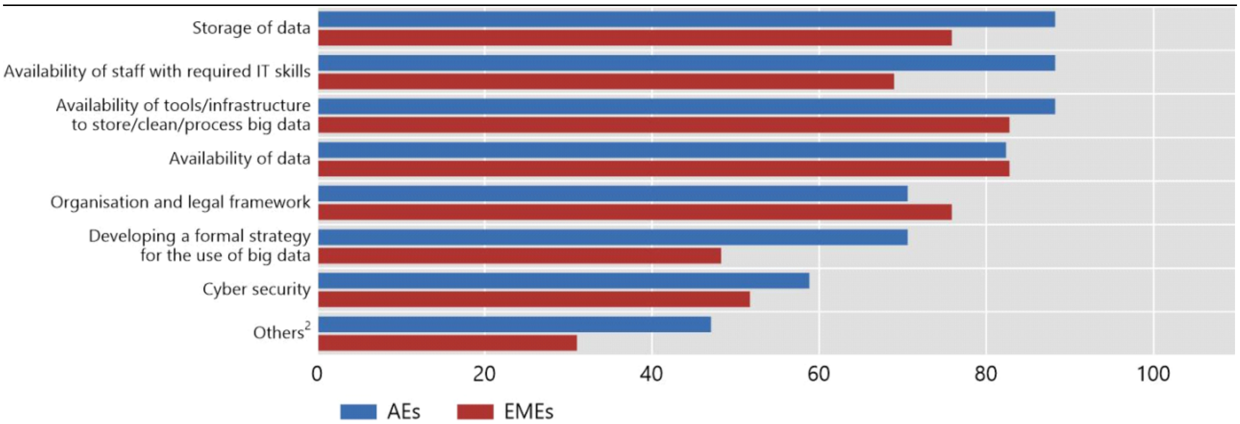
A key challenge for central banks is setting up the necessary IT infrastructure. Providing adequate computing power and software, as well as training existing or hiring new staff, involves high up-front costs. The same holds for creating a data lake, ie pooling different data sets that are curated for future use. Yet a reliable and safe IT infrastructure is a prerequisite not only for big data analysis, but also to prevent cyberattacks.

A related challenge is the legal underpinning for the use of private and confidential data. Traditionally, most data were collected and hosted within public institutions and hence readily available for analysis with clearly defined access rights. More recently, data creation has migrated to the private sector. Companies own vast troves of granular data, some of which are confidential and not accessible to central banks, while others are publicly available. For example, central banks can web-scrape information from market platforms or social media.

However, various terms and conditions may restrict the use of these data. In addition, certain forms of web-scraping are illegal in some jurisdictions. For example, in the United States, web crawlers may not obtain data from sites that require authentication. Further, companies often use techniques such as "rate-throttling" or CAPTCHA ("Completely Automated Public Turing test to tell Computers and Humans Apart") to prevent large-scale automated web-scraping.

What is the focus of the discussions on big data within your institution?[1]

In per cent — Graph 4



[1] The graph reports the share of respondents that selected each respective answer to the question "What is the focus of the discussions on big data within your institution?". Respondents could select multiple options. [2] Includes "data quality and reliability", "data interpretation" and "data governance".

Source: IFC (2020).

General challenges to the use of big data and machine learning

A more general challenge concerns ethics and privacy. Citizens might feel uncomfortable with the idea that central banks, in particular, and governments and large corporation, in general, are scrutinising their search histories, social media postings or listings on market platforms. While these concerns are not new, the amount of data produced in a mostly unregulated environment makes them more urgent (Jones and Tonetti (2020), Boissay et al (2020)).

Fundamentally, these considerations indicate that citizens value their privacy and might be unwilling to share their data if given a choice. Privacy concerns vary both across and within jurisdictions. For example, Chen et al (2021) find that women are less willing than men to share their data with fintechs for better offers. They are also less likely to use fintech products and services. The gender gap is present in several countries and the willingness to share data is lower among individuals in richer countries. To varying degrees, policymakers need to convince the public that the use of data will not infringe their right to privacy – while it takes a long time to build trust, it can easily evaporate if there is, for example, a data breach (Cantù et al 2020).

A further challenge is "algorithmic fairness".[9] Considerations of algorithmic fairness are less relevant for some tasks (eg nowcasting), but they may matter greatly for others (eg evaluating the suitability of regtech applications), in general any application of machine learning that effects individuals needs to be subject to fairness validations (MacCarthy (2019)). Algorithms train on preclassified data sets that can be subject to biases, including related to gender and ethnicity. As discussed above, algorithms require a manually labelled training data set to "learn" what represents a positive or negative stance before they can eg classify political speeches independently. The initial classification of speeches by human operators necessarily contains subjective elements, as not all words are unambiguously positive or negative. These decisions could lead to an algorithm that "misclassifies" speeches from the perspective of somebody who has different word associations in mind (Narayanan (2019)).

In general, the use of an algorithm does not necessarily make a judgment objective, because the problem might lie with the available data, ie censored data. For instance, data on past loan applications could reflect any discriminatory decisions on the part of loan officers vis-à-vis minorities or women (Angwin et al (2016), Ward-Foxton S (2019)). Likewise, unrepresentative data could lead an algorithm to wrongly infer attributes about underrepresented segments of the population or perpetuate any previous biases.[10]

Machine learning algorithms are trained to perform in some error metric. For example, the machine learning introduction above made the point that parameters values are irrelevant in most machine learning models. Financial institutions report the "difficulty of explaining processes" as a major obstacle to the application of big data in stress testing (Institute of International Finance (2019)). Even if suptech

---

[9] IIF (2020) finds that there is no "one-size-fits-all" approach to machine learning governance, and there are interesting regional differences, many of which can be attributable to existing non-discrimination and data protection laws.

[10] Consider the use of suptech applications for stress testing. Since suptech applications train and refine machine learning algorithms on historical data, any biases or unintended consequences of previous stress testing rounds could hence be perpetuated. For example, previous rounds of CCAR have been linked to a disproportionate decline in small businesses lending (Doerr (2019)).

applications were to excel in predictive accuracy, the fact that they cannot necessarily "explain" why a decision was made (eg why risk weights should be X% and not Y%) would limit their use.

Finally, data quality represents another challenge. This is especially the case for unstructured data, which are often a by-product of economic or social activity and not curated for the purpose of further analysis. Major problems include data cleaning (eg in the case of media, social media or financial data), sampling and representativeness (eg in the case of Google searches or employment websites) and matching new data to existing sources.

Graph 5 shows a cloud of words with the most-frequently mentioned challenges mentioned by central banks in the IFC survey. Larger words are mentioned more often. Central banks' answers reflect the preceding discussion. Terms related to IT infrastructure and staff clearly stand out, for example *hardware*, *software*, *skill* and *data scientist*. Other challenges in the realms of *privacy* and *legal* also feature prominently.

Challenges in the use of big data: most-mentioned words

Word cloud | Graph 5

Source: IFC (2020); authors' calculations.

Is machine learning the silver bullet for prediction?

Machine learning can yield large benefits in terms of the accuracy of predictions. The breakthrough in machine learning came when researchers started to let the data decide what model works best for prediction, rather than approaching the data with an algorithm based on the specific model through which the researcher understands the data (based on priors). However, this approach implies significant differences between machine learning and traditional regression that may compromise machine learning's suitability for economic analysis, which will be discussed in the following paragraphs.

First, unlike in traditional regression analysis, in machine learning the coefficients on explanatory variables are not usually of intrinsic interest and do not come with standard errors. They hence cannot be interpret in a classical sense. For example, just because a LASSO regression[11] chooses to include a dummy for gender, but not for race, in predicting borrower default, one must not conclude that race does not matter.

Second, the underlying specification that maximises predictive power need not be unique (Mullainathan and Spiess (2017)). Some explanatory variables are correlated and interacted with other variables in a complex and non-linear way, so two specifications with a different set of variables can be equally successful in predicting an outcome. While in principle it is possible to measure the importance of individual variables for obtaining the outcome (their so-called *feature impact*), quantifying their impact gets increasingly difficult as data and algorithms become more complex (Bolón-Canedo et al (2015)). Moreover, quantifying feature impact does not allow for causal interpretation: gender might be a sufficient statistic for a host of other (unobservable) variables but is by itself not causing the observed outcome.

Third, while years of experience with traditional models such as Ordinary Least Squares (OLS) regressions have led to a best practice when performing analysis and presenting results to economists and policymakers, no such standards exist for machine learning yet. There is a bewildering amount of algorithms to choose from that is growing each day. Algorithms differ along several dimension and contain subjective elements, such as "empirical tuning", eg the choice of information criteria by the researcher. A best practice that prescribes which algorithm is suited for which task is still developing. Researchers at times cannot interpret coefficients or judge their importance in the prediction process. All of this implies that good prediction can come at the cost of accepting that the underlying model is a black box or might hinge on the chosen algorithm.

Fourth, machine learning models do not seek to identify causal relationships, which constitutes a critical aspect in economic analysis. Indeed, the underlying model is irrelevant in many instances (such as identifying which pictures contain dogs and which contain cats, or which credit card transaction is classified as fraudulent), but it is certainly of high relevance for policymakers . A highly sophisticated deep learning model may accurately predicts a recession in the next month, but the prediction is of little help in preventing the downturn unless policymakers can identify where the troubles originate. Is credit expanding too fast in some sectors? Are housing prices too high with respect to fundamentals? Is a slowdown in growth associated with a reduction in consumption, investment or foreign trade? Policy responses would drastically differ under each scenario.[12]

---

[11] A Least absolute shrinkage and selection operator (Lasso) regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (ie models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muticollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

[12] A similar scenario is possible in regtech and suptech applications. Suppose several banks report that they fail a stress test scenario, based on results from their preferred regtech application. Unless banks can identify why they fail – ie where the problems originate (mortgage lending? SME lending? The interbank market?) – they will not be able to adequately address them; and neither will the supervisor. In a recent IIF survey financial institutions cite the "difficulty of explaining processes" and "supervisory understanding of or consent to use new processes" as major obstacles to the application of RegTech in stress testing (Institute of International Finance (2019)).

Fifth, some machine learning methods may not be suited to predict outcomes in an environment that needs some elements of judgement, as it is often the case in economic analysis (see Narayanan (2019)). For example, everybody agrees what a car is, so as long as we provide an algorithm with enough training data of cars, it will eventually become almost perfect in recognizing cars in pictures. However, machine learning is less reliable in tasks that aim to automate judgement. The reason is that there usually is no unanimous judgement. For example, interpretations of a monetary policy stance are inherently subjective, and algorithms are bound to make decisions that are "wrong" from somebody's perspective. It is hence paramount to remember that using an algorithm does not make any decision more objective.[13] Similarly, while machine learning excels when one is "searching for a well-defined profile, when there are a reasonable number of events per year, and when the cost of false alarms is low" (Schneier (2015)), it is less successful in predicting low-frequency events such as recessions or terrorist attacks. Succinctly put, a cat today is the same as a cat yesterday and a cat tomorrow – and there are plenty of cat-pictures to train on. But a recession tomorrow need not look anything like a recession today, and recessions are infrequent events.[14]Further, data quality can comprise the usefulness of machine learning techniques. As for any analysis, the golden rule of "garbage in, garbage out" applies. Incorrect or poor-quality input will produce faulty output. While this is true for traditional as well as machine learning methods, big data exacerbates the problem: most big data is produced as a by-product of other applications and hence often unstructured and not representative, so not randomly sampled (Tissot (2018)). This implies that big data is not big just because it has many observations. What matters are its size relative to the total population and its representativeness. For big data to be useful, it is paramount to validate every step of the pre-analysis: it needs to be cleaned, verified, and checked for its internal and external validity.

Finally, training algorithms not only requires huge amounts of data, but also enormous computing power and storage capacity. In fact, computational demands are increasing exponential, which is why the world's most powerful machine learning algorithms are developed by a handful of large technology firms. In fact, researchers worry that academic institutions will be priced out of research on AI and machine learning (Strubell et al (2019)). This high concentration of resources also creates significant operational risks for private sector companies that use these services. In principle, nothing would keep a central bank from buying a ready-trained algorithm, instead of spending resources on training its own algorithm. However, every algorithm is only as good as its underlying data: even if the underlying data is of excellent quality, US or Chinese data on borrower defaults need not be representative of default in Germany or Cameroon.

These considerations highlight a general trade-off between accuracy and interpretability. Accuracy refers to the number of correctly predicted outcomes, eg the share of identified credit card frauds that are actual frauds. Explainability refers to

---

[13] A related point is that, while an algorithm itself has no ulterior motives, algorithms are developed by humans that are subject to biases, and that algorithms crunch data that is the result of human action and influence. For example, algorithms could perpetuate any previous biases present in data on loan application.

[14] According to Schneier (2015), "data mining works best when you're searching for a well-defined profile, when there are a reasonable number of events per year, and when the cost of false alarms is low". This is why machine learning works wonders in predicting credit card fraud, but does a poor job in preventing terrorist attacks, which are highly idiosyncratic. It is also why (as of now) machine learning is likely ill-suited to predict economic recessions.

the ease with which the prediction outcome can be attributed to a specific cause or factor. Central banks hence might face a trade-off between sophisticated machine learning methods that excel at prediction but are weak in terms of explainability, vis-a-vis simpler models that allow researchers to understand factor impacts, but do worse in prediction. However, the emerging field of interpretable machine learning is working on predictive algorithms that are both accurate and explainable (Corbett-Davies and Goel (2018)).

5. A role for policy cooperation?

Big data can lead to data gravity: companies that already have an edge in collecting, storing, and analysing data attract ever more data over time. Heavy up-front investment is required to build data storage facilities, hire and train staff, gather and clean data, and develop or refine algorithms. However, once the infrastructure is in place, the cost of adding each extra unit of data is minimal. Data gravity has led to a high concentration among companies that provide big data services. For example, in 2018, the global market share of the five biggest cloud service providers reached almost 80%. Amazon alone had a global market share of almost 50% (The Economist (2020)).

Central banks could decide to purchase data and algorithms from external providers instead of developing big data capacities in-house. This could result in significant cost savings. Several private companies collect data from non-traditional sources (such as satellite images or web-scraped prices from online shops) and develop readily available deep learning algorithms. It would further allow central banks to draw on the expertise of private sector data scientist, which are in high demand and short supply.

Purchasing external data or applications is not unusual, but such decisions come with significant legal and other risks. Data from non-traditional sources are often not externally validated, eg by accounting companies or regulatory bodies. The combination of poor quality and limited user experience could result in incorrect inferences, with the associated risks. The same argument holds for purchasing external algorithms. Deep learning algorithms often require several months of training on large quantities of data. When purchased, the algorithms cannot be readily understood without detailed knowledge of the underlying code. This turns analysis into a black box. Central banks might thus consider building up sufficient expertise in-house – at least initially – before turning to external providers.

Cooperation among public authorities could relax the constraints on collecting, storing and analysing big data. Many of these issues result from an acute shortage of data scientists, who are also much sought-after in the private sector (Cœuré (2020)). Smaller jurisdictions face additional problems, as they do not benefit from economies of scale when investing in hardware and software. Instead, they could share in the setup costs or "rent" the necessary storage capacity and computing power, as well as staff resources, from larger jurisdictions.

A further reason for cooperation among central banks could be to contribute to mitigate the environmental costs of big data. Training algorithms and storing large amounts of data consume enormous amounts of energy, thereby expanding an institution's carbon footprint (Strubell et al (2019)). Indeed, recent initiatives aim at fostering "green AI" (Schwartz et al (2019)). With central banks' increasingly focusing on the "green swan", ie the financial stability risks resulting from climate change, they should strive for minimizing their big data carbon footprint. Put succinctly, ten central banks sharing one data centre is likely better for the climate than ten central banks operating ten individual data centres.

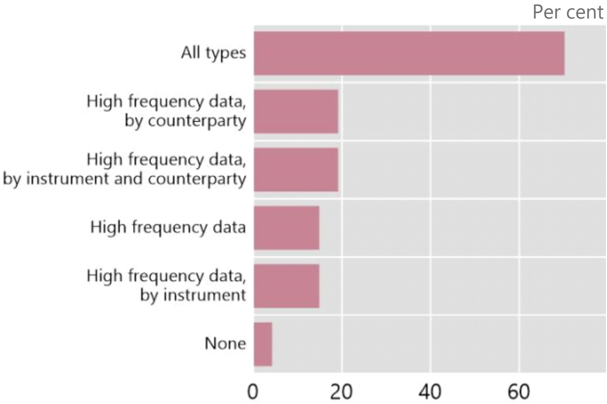
Examples of successful cooperation include the ECB's AnaCredit database or the BIS's international banking statistics and International Data Hub. AnaCredit (short for "analytical credit data sets") collects harmonised data from euro area member states in a single database to support decision-making in monetary policy and macroprudential supervision. Likewise, the BIS collects and processes confidential banking data in cooperation with central banks and other national authorities. That said, legal and practical aspects may constrain cooperation across jurisdictions. Legal obligations to store the raw data within national boundaries could restrict the sharing of confidential data. There would be a need to agree on, say, cloud computing contracts, as well as on rules for data use and the protection of confidentiality. Protocols to ensure algorithmic fairness would have to be developed and ratified.

Looking ahead, a promising new area for collaboration could be global payments data. Around two thirds of respondents in the survey consider high-frequency payments data of all types useful, less than 5% see no use for these data (Graph 7, left-hand panel). Over 90% would be willing to contribute to a pilot study on their use (right-hand panel).

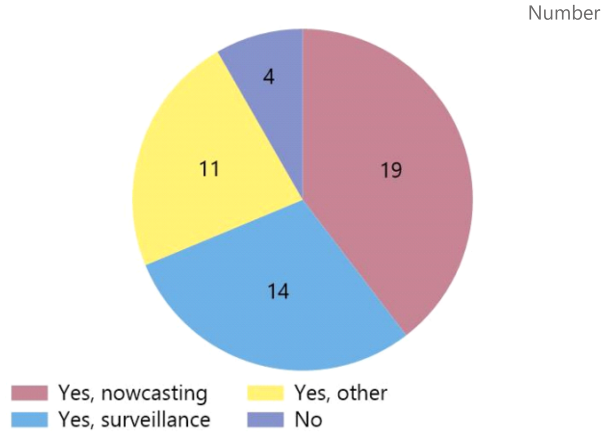
Payment data: interest in central banking

Graph 7

Which types of payments data are useful for your institution?



Would your institution be willing to contribute to a pilot study on the use of payments data?



Source: IFC Big data survey (2020)

International financial institutions can help to foster cooperation. In this regard, the BIS Innovation Hub has identified as strategic priorities, among others, effective supervision (including regtech/suptech) and data platforms/open finance that could draw on big data. It is currently developing its work programme in these fields, with a view to producing proofs-of-concept that can benefit the central banking community. [15] The BIS Innovation Network, which the Innovation Hub has established, will bring together experts on innovation from BIS member central banks and, as a key priority, will focus on central banks' problems/prototypes in early stages of development to kick-start collaboration on practical problem statements and solutions. Members of the network will also exchange views on nascent technologies, potentially including big data and algorithms.

6. Conclusion

Big data and machine learning are now being used in almost every sector of the economy. Central banks are also increasingly using big data for research purposes and to inform policy decisions. In 2020, over 80% of central banks report that they use big data, up from just 30% five years ago. Among the institutions that currently use big data, over 70% use it for economic research, while 40% state that they use it to inform policy decisions.

These numbers suggest that big data and machine learning offer many useful applications and can help central banks in fulfilling their mandate. Nowcasting GDP and inflation or examining spending patterns across regions and population subgroups in real time provide just two of many examples.

Yet, central banks also face challenges to unlock the full potential of big data and machine learning. A key topic of discussion is the availability of big data and tools to process, store and analyse it. The design of legal frameworks or aspects of cyber security are also at the centre of central bankers' concerns. More practical problems are budget constraints and the difficulty in training existing or hiring new staff to work on big data related issues.

Central banks are willing to join forces to reap the benefits of big data, the IFC survey shows. Indeed, half of them reported an interest in collaborating in one or more specific project, with three types of cooperation envisaged. First, by sharing knowledge among those institutions that have developed specific expertise that can be reused in other jurisdictions. These expertise include general big data techniques (eg data visualisation, network analysis, machine learning tools), more general information management issues (eg development of open-source coding, data sharing protocols, encryption and anonymisation techniques for using confidential data) as well as specific applications that are more devoted to the central bank community (eg suptech and regtech areas). Second, by using big data to work on global issues such as international spillovers, global value chains and cross-border

[15] The Innovation Hub implements a one-year work programme that sets its key focus themes, their associated projects, the centres where they will be conducted and partnerships. The programme is endorsed by the Economic Consultative Committee (ECC) and is developed under the direction of the Head of the Innovation Hub and the centre heads.

payments. Third, by developing joint exploratory projects to benefit from economies of scale and collectively share (limited) financial and human resources.

International financial institutions can greatly support these cooperative approaches. They can facilitate innovation by promoting technological solutions to harmonise data standards and processes among jurisdictions. With this spirit, the BIS Innovation Hub has been established to identify and develop insights into critical trends in financial technology of relevance to central banks, explore the development of public goods to enhance the functioning of the global financial system, and serve as a focal point for a network of central bank experts on innovation. Such a network could undoubtedly play an important role in facilitating international cooperation to exploit big data sources and techniques.

References

Altig, D, S Baker, J Barrero, N Bloom, P Bunn, S Chen, S Davis, J Leather, B Meyer, E Mihaylov and P Mizen (2020): "Economic uncertainty before and during the Covid-19 pandemic", *Journal of Public Economics*, vol 191, September, pp 104–274.

Angwin, J, J Larson, S Mattu and L Kirchner (2016): "Machine bias", *ProPublica*, 23 May.

Amstad, M, L Gambacorta, C He and D Xia (2021): "Trade sentiment and the stock market: new evidence based on big data textual analysis of Chinese media", *BIS Working Paper*s, no 917.

Andersen, A, E Hansen, N Johannesen and A Sheridan (2020): "Consumer responses to the Covid-19 crisis: evidence from bank account transaction data", *CEPR Discussion Papers*, no 14809.

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

Bailey, M, R Cao, T Kuchler and J Stroebel (2018): "The economic effects of social networks: evidence from the housing market", *Journal of Political Economy*, vol 126, no 6, pp 2224–76.

Baker, S, N Bloom and S Davis (2016): "Measuring economic policy uncertainty", *Quarterly Journal of Economics*, vol 131, no 4, pp 1593–636.

Baker, S, R Farrokhnia, S Meyer, M Pagel and C Yannelis (2020): "How does household spending respond to an epidemic? Consumption during the 2020 Covid-19 pandemic", *NBER Working Papers*, no 26949.

Battaglini, M and E Patacchini (2019): "Social networks in policy making", *Annual Review of Economics*, vol 11, pp 473–94.

Boissay, F, T Ehlers, L Gambacorta and H S Shin (2020): "Big techs in finance: on the new nexus between data privacy and competition", in R Rau, R Wardrop and L Zingales (eds), *The Handbook of Technological Finance*, Palgrave Macmillan.

Bolón-Canedo, V, N Sánchez-Maroño and A Alonso-Betanzos (2015): "Recent advances and emerging challenges of feature selection in the context of big data", *Knowledge-Based Systems,* vol 86, September, pp 33–45.

Broeders, D and J Prenio (2018): "Innovative technology in financial supervision (suptech) – the experience of early users", *FSI Insights*, no 9.

Buch, C (2019): "Welcoming remarks International Seminar on Big Data", Building Pathways for Policy-Making with Big Data, Bali, 26 July 2018, also in *IFC Bulletin* (2015).

Cantú, C, S Doerr, G Cheng, J Frost and L Gambacorta (2020): "On health and privacy: technology to combat the pandemic", *BIS Bulletin*, no 17.

Carvalho, V, S Hansen, A Ortiz, J Garcia, T Rodrigo, S Rodriguez Mora and P Ruiz de Aguirre (2020): "Tracking the Covid-19 crisis with high-resolution transaction data", *CEPR Discussion Papers*, no 14642.

Cao, S, W Jiang, B Yang and A Zhang (2020): "How to talk when a machine is listening: corporate disclosure in the age of AI", SSRN.

Chakraborty C and A Joseph (2017): "Machine learning at central banks", *Bank of England Working Paper*, no 674.

Chakrabarti, R, S Heise, D Melcangi, M Pinkovskiy and G Topa (2020a): "Did state reopenings increase consumer spending?", *Liberty Street Economics*, 18 September.

——— (2020b): "How did state reopenings affect small businesses?", *Liberty Street Economics*, 21 September.

Chen, S, S Doerr, J Frost, L Gambacorta and H S Shin (2021): "The fintech gender gap", *BIS Working Papers*, fothcoming.

Chen, S, D Igan, N Pierri and A Presbitero (2020b): "Tracking the economic impact of Covid-19 and mitigation policies in Europe and the United States", *IMF Working Papers*, no 20/125.

Chetty, R, J Friedman, N Hendren and M Stepner (2020): "How did COVID-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data", *NBER Working Papers*, no 27431.

Cœuré, B (2017): "Policy analysis with big data", speech at the conference on Economic and Financial Regulation in the Era of Big Data, organised by the Bank of France, Paris, 24 November.

——— (2020): "Leveraging technology to support supervision: challenges and collaborative solutions", speech at the Financial Statement event series, Peterson Institute, 19 August.

Corbett-Davies, S and S Goel (2018): "The measure and mismeasure of fairness: a critical review of fair machine learning", arXiv preprint arXiv:1808.00023.

Cornelli, G, S Doerr, L Gambacorta and O Merrouche (2020): "Inside the regulatory sandbox: effects on fintech funding", *BIS Working Papers*, no 901.

D'Amuri, F and J Marcucci, (2017) "The predictive power of Google searches in forecasting unemployment", *International Journal of Forecasting*, vol 33, no 4, pp 801–16.

Doerr, S (2019): "Unintended side effects: stress tests, entrepreneurship, and innovation", *BIS Working Papers*, no 823.

Doerr, S and L Gambacorta (2020a): "Identifying regions at risk with Google Trends: the impact of Covid-19 on US labour markets", *BIS Bulletin*, no 8.

——— (2020b): "Covid-19 and regional employment in Europe", *BIS Bulletin*, no 16.

Draghi, M (2018): "Welcome remarks at the third annual conference of the ESRB", September.

Financial Stability Board (2020): "The use of supervisory and regulatory technology by authorities and regulated institutions. Market developments and financial stability implications", Report to the G20, October.

Foroni, C, M Marcellino and D Stevanovic (2020): "Forecasting the Covid-19 recession and recovery: lessons from the financial crisis", *ECB Working Papers*, no 2468.

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

Hansen, S and M McMahon (2016): "Shocking language: understanding the macroeconomic effects of central bank communication", *Journal of International Economics*, vol 99, no 1, pp 114–33.

Harvey, R (2019): "Deep learning: miracle or snake oil?", mimeo.

Hassan, T, S Hollander, L van Lent and A Tahoun (2020): "Firm-level exposure to epidemic diseases: Covid-19, SARS, and H1N1", *NBER Working Papers*, no 26971.

Higgins, P (2014): "GDPnow: a model for GDP nowcasting", *Federal Reserve Bank of Atlanta Working Paper Series*, no 2014-7.

Institute of International Finance (2019): "Machine learning in credit risk", July.

——— (2020): "Machine learning governance", December.

Irving Fisher Committee (2011): "Initiatives to address data gaps revealed by the financial crisis", *IFC Bulletin,* no 34.

——— (2015): "Central banks' use of and interest in big data", *IFC Report,* no 3.

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

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

Jones, C and C Tonetti (2020): "Nonrivalry and the economics of data", *American Economic Review,* vol 110, no 9, pp 2819–58.

Kalamara, E, A Turrell, C Redl, G Kapetanios and S Kapadia (2020): "Making text count: economic forecasting using newspaper text", *Bank of England Staff Working Paper*, no 865.

Lewis, D, K Mertens and J Stock (2020): "Tracking the Covid-19 economy with the weekly economic index (WEI)", *Liberty Street Economics*, 4 August.

McCarthy (2019): "Fainess in algorithmic decision-making", Brookings Institution, 6 December.

Mullainathan, S and J Spiess (2017): "Machine learning: an applied econometric approach", *Journal of Economic Perspectives*, vol 31, no 2, pp 87–106.

Narayanan, A (2019): "How to recognize AI snake oil", speech at the MIT panel on The Promise and Peril of Artificial Intelligence, 19 November.

Padilla, A and M Pagano (2000): "Sharing default information as a borrower discipline device", *European Economic Review*, vol 44, no 10, pp 1951–80.

Pagano, M and T Jappelli (1993): "Information sharing in credit markets", *Journal of Finance*, vol 48, no 5, pp1693–18.

Schneier B (2015): *Data and Goliath. The hidden battles to collect your data and control your world*, W W Norton & Company.

Schwartz, R, J Dodge, N Smith and O Etzioni (2019): "Green AI", arXiv preprint arXiv:1907.10597.

Strubell, E, A Ganesh and A McCallum (2019): "Energy and policy considerations for deep learning in NLP", arXiv preprint arXiv:1906.02243.

The Economist (2020): "Should data be crunched at the centre or at the edge?", Special Report, 20 February.

Tissot, B (2015): "Big data for central banks", in IFC (2015).

——— (2019): "Financial big data and policy work: opportunities and challenges", Eurostat, *Statistical Working Papers*, KS-TC-19-001-EN-N.

Ward-Foxton S (2019), "Reducing bias in AI models for credit and loan decisions, *DesignLines*, 30 March.

Wolski, M and P Wruuck (2020): "Covid-19 and the EU labour market: corporate health matters", *VoxEU*, 5 August.

Previous volumes in this series

929 March 2021	Greening (runnable) brown assets with a liquidity backstop	Eric Jondeau, Benoit Mojon and Cyril Monnet
928 February 2021	Debt specialisation and diversification: International evidence	Gregory R. Duffee and Peter Hördahl
927 February 2021	Do macroprudential policies affect non-bank financial intermediation?	Stijn Claessens, Giulio Cornelli, Leonardo Gambacorta, Francesco Manaresi and Yasushi Shiina
926 February 2021	Answering the Queen: Machine learning and financial crises	Jérémy Fouliard, Michael Howell and Hélène Rey
925 January 2021	What 31 provinces reveal about growth in China?	Eeva Kerola, and Benoît Mojon
924 January 2021	Permissioned distributed ledgers and the governance of money	Raphael Auer, Cyril Monnet and Hyun Song Shin
923 January 2021	Optimal bank leverage and recapitalization in crowded markets	Christoph Bertsch and Mike Mariathasan
922 January 2021	Does regulation only bite the less profitable? Evidence from the too-big-to-fail reforms	Tirupam Goel, Ulf Lewrick and Aakriti Mathur
921 January 2021	Firm-specific risk-neutral distributions with options and CDS	Sirio Aramonte, Mohammad R Jahan-Parvar, Samuel Rosen and John W Schindler
920 January 2021	Investing like conglomerates: is diversification a blessing or curse for China's local governments?	Jianchao Fan, Jing Liu and Yinggang Zhou
919 January 2021	Understanding bank and nonbank credit cycles: A structural exploration	C Bora Durdu and Molin Zhong
918 January 2021	Sovereign credit and exchange rate risks: Evidence from Asia-Pacific local currency bonds	Mikhail Chernov, Drew Creal and Peter Hördahl
917 January 2021	Trade sentiment and the stock market: new evidence based on big data textual analysis of Chinese media	Marlene Amstad, Leonardo Gambacorta, Chao He and Dora Xia
916 January 2021	Firm-level R&D after periods of intense technological innovation: The role of investor sentiment	Sirio Aramonte and Matthew Carl

All volumes are available on our website www.bis.org.