CENTRAL BANKING
FOCUS REPORT

ECB INTERVIEW
Per Nymand-Andersen

BIG DATA SURVEY
Central banks make significant changes to how big data is governed, managed and processed

ONLINE FORUM
Experts discuss harnessing big data

Big Data in Central Banks

In association with
BearingPoint®

Contents

The rapid evolution of data

Report editor: **Daniel Hinge**
daniel.hinge@infopro-digital.com

Contributor: **Emma Glass**
emma.glass@infopro-digital.com

Chairman: **Robert Pringle**

Editor: **Christopher Jeffery**
chris.jeffery@infopro-digital.com

Publisher: **Nick Carver**
nick.carver@infopro-digital.com

Commercial director: **John Cook**
john.cook@infopro-digital.com

Commercial editorial manager: **Stuart Willes**
stuart.willes@infopro-digital.com

Commercial subeditor: **Alex Hurrell**
alex.hurrell@infopro-digital.com

Global corporate subscriptions manager: **Samima Danga**
samima.danga@infopro-digital.com

Cover image:
Johnason/Shutterstock

Central Banking Publications
Infopro Digital Services
Haymarket House
28–29 Haymarket
London SW1Y 4RX, UK
Tel: +44 (0)20 7316 9000
Fax: +44 (0)20 7316 9935
Email: info@centralbanking.com
Website: www.centralbanking.com

Published by Infopro Digital Services Ltd

Much can change in a year – especially where technology is concerned. *Central Banking* conducted its first survey of big data use by central banks in 2016, when thinking on the topic was in flux. This year's survey updates the original findings, highlighting the rapid evolution of the field and a process of coalescence into concrete applications.

One of the most striking results revealed in this year's survey is the shift of big data analysis into the mainstream. In 2016, 22% of central banks described big data as a "core input" into policymaking, while this year 36% said this was the case. In the 2017 survey, 58% of respondents said they were using big data as at least an auxiliary input to policy.

The past year has been marked by growing collaboration among central bank data users, notably in forums such as the Irving Fisher Committee. Central banks have also devoted considerable effort to the question of how data – big and small – should be gathered, stored and organised within their organisations to generate the most powerful results.

Big data techniques such as machine learning are becoming much more widely used in economics, as our feature on the topic finds. Applications from early warning systems to trend forecasting and picking out patterns from unstructured data are being adopted across the central banking community. Deep learning methods bring the potential for dramatic change in the years ahead.

Our Q&A with the European Central Bank's Per Nymand-Andersen reveals the ways in which central banks are harnessing big data, providing insights from one of the leading institutions in the field. Participants in our online forum held in September also revealed the diversity of applications and flagged several issues for central banks to be aware of.

As big data enters the mainstream, it remains critical to understand the ways in which it can transform our thinking, but also to be realistic about its shortcomings and biases. This report is designed as a small step in that direction. ❑

Daniel Hinge
Report editor

Teaching machines to do monetary policy

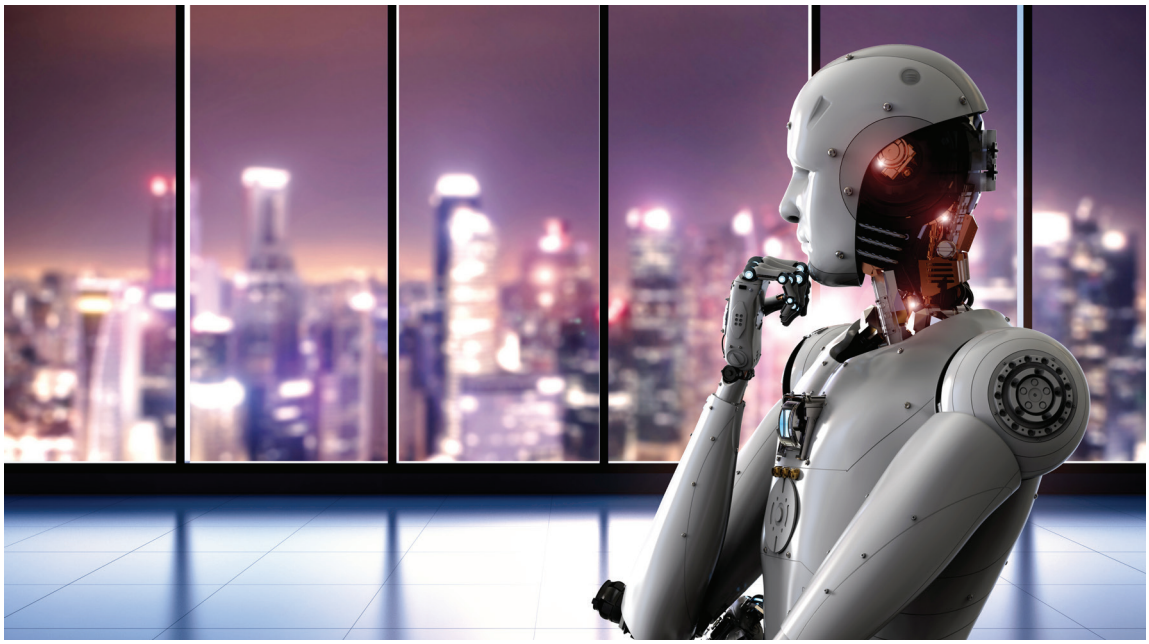
Machine learning may not yet be at the stage where central bankers are being replaced with robots, but the field is bringing powerful tools to bear on big economic questions. By Daniel Hinge.

The idea of artificial intelligence – and machine learning, a subset of the genre – conjures images of shiny metal robots trooping to work on Threadneedle Street to set the UK's monetary policy. International Monetary Fund (IMF) chief Christine Lagarde envisaged such a scenario in a recent speech in London, though she concluded robots wouldn't make for good central bankers because machines follow rules, whereas a good central banker requires discretion to respond to surprises.[1] "In 2040, the governor walking into the Bank will be of flesh and bones, and behind the front door she will find people – at least a few," Lagarde predicted.

Uses for machine learning have abounded in recent years as companies, academics and central banks have embraced the technology, supported by increasingly powerful computers. The most eye-catching examples of machine learning in action see robots taking on humanoid roles, whether it is defeating the world champion at *Go*, navigating a car along a busy street or holding something that resembles a conversation.

Economists who use machine learning tend to stress its more prosaic side, however. It is a powerful tool if used to answer the right questions, but also a statistical technique that should be viewed as just one element of the econometrician's toolkit, rather than an entirely new way of thinking about statistics.

Stephen Hansen, an associate professor at the University of Oxford, is making use of machine learning to analyse text. The technique he is applying – latent Dirichlet allocation (LDA) – resembles factor models employed by central banks. "Conceptually, it is not so different from the kinds of dimensionality-reducing techniques that central banks are already using, it's just that it's being applied to text data," he says. One of the challenges of working with text is its large number of dimensions – machine learning helps to pick out the ones that matter.



Pinning down the concept

At its heart, machine learning is a tool for the automated building, selection or refinement of statistical models. Computers are uniquely well suited to picking through vast quantities of data in search of patterns – which is why machine learning is normally applied to large, often unstructured datasets. It can also work well on smaller datasets that may have other awkward features, such as lack of structure or high dimensions.

While there is a multitude of possible approaches, in general, machine learning proceeds through training, validation and testing phases, with the dataset carved up into three pieces accordingly. The system trains on the first section – often around 60% of the data – before the model is refined.

A big part of the power of machine learning comes from optimisation – computers are very good at choosing parameters to minimise a loss function based on the training data. Where the computer thinks a variable is irrelevant or duplicated, its parameter can be set at zero to eliminate it. The problem is that the model may fit the training data so closely that it fails as soon as it encounters anything new – a problem known as overfitting. The validation phase, working on around 20% of the data, helps to avoid this. A common technique is "regularisation", in which more complex models are penalised. By trading off goodness-of-fit against simplicity, the computer can find a model that is most likely to succeed out of sample.

In the testing phase, the model is run on the final 20% of the dataset to make its predictions. If the results are good, with low errors, the model may be used in further analysis. Otherwise it will be returned for refinement.

Machine learning is generally split into supervised and unsupervised learning. In supervised learning, the system is trained on a set of known outputs – an image recognition program may be trained using images of cats to categorise photos it has not come across before into those that contain cats and those that don't. Unsupervised learning deals with "clustering", or asking the computer to find any pattern in the dataset, without the researcher imposing a model.

Putting it to work

Working with high-dimensional models – those with more parameters to estimate than there are observations – is challenging. "It is problematic because I am trying to get a lot of information from a very small amount of information," says Jana Marečková, a PhD candidate at the University of Konstanz, who has been putting various types of machine learning into practice.

One of her projects has been to detect structural breaks in time-series data. She makes use of a model where the parameter vector can change at any moment, requiring estimation of the number of parameters multiplied by the number of time periods – a high-dimensional problem. The assumption, however, is that, until there is a structural break, the parameters will be constant. Her choice of regularisation method simultaneously finds the position of the structural breaks and the parameter estimates.

A second project makes use of clustering analysis to find patterns in a survey of non-cognitive skills, mapping these on to labour market outcomes. The *1970 British Cohort Study* tracks a group of people born in 1970, asking them a set of questions every four or five years throughout childhood and into adulthood, yielding a rich dataset on broad aspects of their lives, including non-cognitive skills and economic outcomes. "When I compared my results with the psychology literature, I was able to label the groups by the measures that are already known in psychology," Marečková says. "That was a really nice result for me. The machine-learning technique found the right grouping."

The results, set out in a paper co-authored with Winfried Pohlmeier,[2] suggest non-cognitive skills – those relating to an individual's personality or temperament – have a significant impact on how likely a person is to find employment. Marečková and Pohlmeier also seek evidence of an impact on wages, but the relationship proves weaker.

FOMC transcripts

Hansen, of the University of Oxford, similarly uses an unsupervised learning technique to analyse transcripts of Federal Open Market Committee (FOMC) meetings. While others have used dictionary techniques – picking out keywords for a machine to find, for example – his approach, with co-authors Michael McMahon and Andrea Prat, employs LDA, a Bayesian factor model designed to find patterns in text without human guidance.[3]

The researchers identify around 10,000 unique words across every transcript produced during Alan Greenspan's tenure as chair, which LDA is able to boil down into about 40 topics. Each FOMC meeting can then be represented as the percentage of time spent on each topic. In this way, the authors are able to construct a unique dataset from information that at one time only a human could process by reading. With the study comprising 149 meetings, 46,502 unique interjections, and 5.5 million words, there is only so much value a human could extract – even with a lot of time on their hands.

"A lot of machine-learning literature developed with prediction in mind, so the question central banks tend to ask themselves is: can I draw on this toolkit to improve my forecasting?" Hansen says. "But machine-learning techniques can also be used to represent new data."

While LDA is much like other factor models, Hansen notes that text throws up unusual challenges. First, it is high-dimensional: "The dimensionality is really an order of magnitude greater than in quantitative time series." Second, it is very sparse, meaning each transcript will use only a subset of the 10,000 words.

Hansen says the Bayesian nature of the model helps handle situations where an FOMC member only says a few words – it will still allocate shares of the 40 topics, but you should not take the data "too seriously", he says. "These Bayesian models allow you to not overfit the model based on these limited data points."

He and his co-authors exploit a natural experiment, since during Greenspan's tenure he realised tapes of the FOMC meetings were not being erased once minutes were written up, as members of the committee had previously thought. The decision was later taken to publish the transcripts with a lag, allowing researchers to examine how the quality of discussion changed before and after committee members knew their exact phrasing was the subject of historical record. The authors find that the quality of discussion did shift: "We show large behavioural responses to transparency along many dimensions," they write. "The most striking results are that meetings become less interactive, more scripted and more quantitatively oriented."

FOMC members also change their voting patterns as they become more experienced, the researchers find – becoming more likely to challenge the consensus view, and speaking more broadly and less quantitatively. The authors attribute this to the reduced concern each member has over their career later on in their terms.

Going deeper

All of the more "sci-fi" applications of machine learning, from self-driving cars to AlphaGo – the system that beat Fan Hui at the game he'd spent his life studying – are based on an idea called "deep learning". The technique utilises artificial neural networks – so called because they mimic the patterns of neurons and synapses in the human brain.

A deep learning system is structured in layers, each a set of nodes, connected to the nodes in the next layer via a series of links. Input data is processed through any number of "hidden" layers to produce a final output. For instance, data on driving conditions – the input – is processed into a decision – the output – by a self-driving car.

The technique can be very powerful, allowing robots to mimic human behaviour and decision-making. But moral dilemmas reminiscent of Isaac Asimov stories also emerge. A car could easily find itself in a version of the philosophical "trolley problem" – should it kill one to save many? Swerve into a lamppost to avoid the schoolchildren who just stepped out into the road? Engineers training their cars might end up deliberately teaching them to kill their passengers in certain situations.

Economists making use of deep learning are less likely to encounter such knotty moral issues, but there are still plenty of challenges. Peter Sarlin, an associate professor at the Hanken School of Economics, used to build early warning models based on machine learning for the European Central Bank. One of his recent projects with Samuel Rönnqvist uses neural networks in two stages to build an early warning indicator for financial distress.[4] In the first stage, Reuters articles are processed to reduce the number of dimensions from millions to the few hundred that contain meaningful information for financial stability. In the second stage, these few hundred inputs are processed into a signal that is represented by just two nodes: distress or tranquillity.

Black boxes

Working with so many dimensions, computers often go beyond human comprehension. "Humans are not necessarily going to understand it, but that is not the value proposition," says Sarlin. "The value proposition is that we are capable of understanding and computationally analysing human input, and relating that to those events that we want to pinpoint."

An issue for those making use of neural networks, particularly central banks trying to set policy, is that it is hard to know exactly what is going on in the hidden layers – the models are "black boxes". That means, however powerful the program is, and even if it is right 100% of the time, it will be difficult for a policymaker to justify a course of action based on the model output. "What we have done here is a lot more advanced than what we have done with central banks around early warning models, in general," says Sarlin. "At central banks, we have had to end up with models that are fully interpretable."

Nevertheless, he notes that, even with very complex models, it can be possible to trace the reasoning throughout. "The fact we don't understand how precisely we came up with the model doesn't mean we cannot interpret the results," he says.

Systemic risk

Sarlin believes a particularly valuable use for deep learning is in building more realistic models of systemic risk in the financial system. He says, broadly, there have been two branches of research in the area, running on parallel tracks. Some researchers have concentrated on features of the banks and other financial institutions, such as how particular balance-sheet characteristics might make them more vulnerable. Others have built network models to map how interconnections can transmit shocks through the system.

Sarlin's research is now focusing on ways machine-learning techniques can be used to model risks at the level of individual financial institutions before network theory connects them. In this way, supervisors can build up a much more complete picture of the financial system. "I think that is how, in future, you should be looking at interconnected risk," he says.

Prediction versus causation

Much of statistics is devoted to establishing evidence of causation, rather than simple correlation, but many of the standard techniques do not work in machine learning – it can be difficult to employ randomised control trials, for example. Chiranjit Chakraborty and Andreas Joseph explore some of these questions in a wide-ranging Bank of England working paper on the topic.[5] Machine learning has often been developed for use in the market, with a "what works is fine" attitude, they say.

"Concretely, machine-learning tools often ignore the issue of endogeneity and its many possible causes," Chakraborty and Joseph write. "Additionally, there are few known asymptotic or small-sample properties of estimators to draw on. These are serious issues when making decisions and more research is needed."

Susan Athey and Guido Imbens are working on one method of establishing causal inference.[6] Their approach is to estimate heterogeneity in causal effects in experimental studies. The challenge is that some elements of the population have received treatment while others have not, but clearly none can have received both. Athey and Imbens divide up the population into subgroups, then use validation techniques similar to those used to refine standard machine-learning models, to explain differences in treatment effects between subgroups.

The technique can be applied to gain additional insights into randomised controlled trials that have already been conducted. "A researcher can apply our methods and discover subpopulations with lower-than-average or higher-than-average treatment effects, and can report confidence intervals for these estimates without concern about multiple testing," the researchers say.

Future applications

Research into machine learning and the many challenges it poses is clearly active, and many major questions remain unanswered. But with a proliferation of research and the creation of many private sector firms working with machine learning, the ease of putting the techniques into practice is growing rapidly. Google has released a package of open-source tools for deep learning called TensorFlow[7], which works via the Python programming language. Other free machine-learning tools abound – SciKit-Learn[8] is another Python-based toolkit that supports methods including clustering, dimensionality reduction and model selection.

As such, machine learning gives central banks a suite of new tools they can put to use at relatively low cost. For example, Rendell de Kort, an economist at the Central Bank of Aruba, presented a machine-learning model for forecasting tourism demand at a conference hosted by the Irving Fisher Committee earlier this year.[9] Using neural networks and another method known as "random forests", de Kort found that machine learning yielded "fairly accurate" estimates of tourism demand without a severe computational burden. However, both techniques do represent black boxes, he noted.

The IMF's Lagarde may find it unlikely that machines will sit on monetary policy committees any time soon, but the advances in deep learning are opening up areas of artificial intelligence that previously seemed highly speculative. Various companies claim they will have self-driving cars on the market in a few years' time, and robots are already taking on managerial roles in industries such as asset management. Most applications will likely remain in the narrower econometric space, further from the public eye, but robot central bankers – or at least central bank analysts – do not seem far off. As Marečková says of deep learning systems, there is a long way to go, but "they are going to be really powerful once they start doing what they promise". ❑

Notes

1. Christine Lagarde, *Central banking and fintech – A brave new world?*, International Monetary Fund, September 2017.
2. Jana Marečková and Winfried Pohlmeier, *Noncognitive skills and labor market outcomes: A machine learning approach*, May 2017, *https://tinyurl.com/ya4xerxo*
3. Stephen Hansen, Michael McMahon and Andrea Prat, *Transparency and deliberation within the FOMC: A computational linguistics approach*, June 2017, *https://tinyurl.com/yajst9qm*
4. Samuel Rönnqvist and Peter Sarlin, *Bank distress in the news: Describing events through deep learning*, *2016, https://tinyurl.com/yddc3bz9*
5. Chiranjit Chakraborty and Andreas Joseph, *Machine learning at central banks*, September 2017, *https://tinyurl.com/y87vm5kz*
6. Susan Athey and Guido Imbens, *Machine learning methods for estimating heterogeneous causal effects*, April 2015, *https://tinyurl.com/y77qyjzy*
7. TensorFlow, *www.tensorflow.org*
8. Scikit-learn.org*, Machine learning in Python*, *https://tinyurl.com/cuttxkx*
9. *Rendell E de Kort,* Forecasting tourism demand through search queries and machine learning, *March 2017, www.bis.org/ifc/publ/ifcb44f.pdf*

Making the most of big data

Per Nymand-Andersen, adviser to senior management at the European Central Bank, discusses how central banks can benefit from embracing big data and what this could mean for the industry in the near future.

The concept of big data can be hard to pin down – how would you define it?

Per Nymand-Andersen: Big data can be defined as a source of information and intelligence resulting from the recording of operations, or from the combination of such records. There are many examples of recorded operations – records of supermarket purchases, robot and sensor information in production processes, satellite sensors, images, as well as behaviour, event and opinion-driven records from search engines, including information from social media and speech recognition tools. The list seems endless, with more and more information becoming public and digital as a result – for example, the use of credit and debit payments, trading and settlement platforms, and housing, health, education and work-related records.

Should central banks take advantage of big data?

Per Nymand-Andersen: Yes, though central banks do not have to be ahead of the curve. They should not miss this opportunity to extract economic signals in near real time and learn from the new methodologies. Big data can help to enhance economic forecasts and obtain more precise and timely evaluations of the impact of policies.

We may also have to manage expectations to strike a balance between the perceived big data hype and the reality of the future. Big data is part of the 'data service evolution' – it is borderless and impacts the structure and functioning of financial markets, our economies and societies.

Per Nymand-Andersen is an adviser to senior management at the European Central Bank (ECB) and a lecturer at Goethe University Frankfurt. Nymand-Andersen has worked as an economist and statistician with the ECB since 1995, becoming principal market infrastructure expert in 2007 before taking up his current role in 2010. He completed an MBA at Copenhagen Business School.

Therefore, central banks must monitor closely, assess and adapt. We must explore and experiment with new technologies to avoid falling behind the technological curve, and to anticipate their impact on central banks' policies and their transmission throughout the economy.

Central banks may need to join forces in exploring and assessing the usefulness of selective big data sources and their relevance to central banking, as has been initiated by the Irving Fisher Committee on Central Bank Statistics (IFC), for instance.

What are some of the most promising uses for big data at central banks?

Per Nymand-Andersen: The availability and accessibility of large data sources is a new and rich field for statisticians, economists, econometricians and forecasters, and is relatively unexploited for central banking purposes. It would be of particular interest if such sources could help to detect trends and turning points within the economy, providing supplementary and more timely information compared with central banks' traditional toolkits.

Central banks already have access to a large amount of statistics, intelligence, structured data and information, which are regularly fed into their decision-making processes in the course of fulfilling their mandates. Central banks therefore appear well positioned to apply their existing models and econometric techniques to new datasets and develop innovative methods of obtaining timelier or new statistics and indicators. These supplementary statistics may provide further insights that support central bankers' decision-making

processes, and assess the subsequent impact and associated risks of decisions on the financial system and real economy. Big data could help central bankers obtain a near real-time snapshot of the economy, as well as providing early warning indicators to identify turning points in the economic cycle.

Additionally, there are new methods and techniques being developed by academic and private researchers to deal with new big data sources. For instance, text-mining techniques open up new possibilities to assess what John Maynard Keynes referred to as "animal spirits", which cannot be captured in standard economic equations and quantitative variables. Sentiment indexes harvested from internet articles, social media and internet search engines may, by applying adequate statistical algorithms, provide useful and timely insight into consumer sentiment, market uncertainty or systemic risk assessments. Furthermore, new machine-learning techniques and tools are used to provide predictions on the basis of large, complex datasets.

Turning to banking supervision and regulation, there is a clear drive for regulatory authorities to obtain more micro-level data. Since the financial crisis, regulators have been keen to expand their data collections to improve their ability to monitor financial risks and vulnerabilities. New big data sources may support these supervisory tasks – such sources include online operations in trading platforms, credit card payment transactions, mobile banking data, records related to securities settlement and cash payment systems, clearing houses, repurchase operations and derivatives settlement, and commercial and retail transactions.

How is the European Central Bank (ECB) putting this to work?

Per Nymand-Andersen: Collaborative efforts among central banks have been initiated by the IFC, bringing central banks together in showcasing pilot projects on the use of big data for central banking purposes by applying the same methodological framework and working with a similar timetable. The ECB is a member of the IFC and is actively contributing to this work. These pilots are focused around four main themes focused on extracting insight from the following: the internet, including search engines and price information; commercial sources, such as credit card operations; administrative sources, such as fiscal and corporate balance-sheet data; and financial markets, looking at liquidity, transactions and prices. Further work is also being undertaken on text analysis of media reports.

Likewise, the ECB has the Money Market Statistics Regulations and AnaCredit providing high-volume intraday transactional micro-level data on intra-bank loans and banks' loans to corporates and households, and we are experimenting with text analysis and search machines.

Target2Securities – which provides a unique platform for settlement in near real time of securities transactions – is a one-stop shop for securities settlement operated by central banks, with the central securities depositories as core customers. This could likewise be a potential source of data for exploring intraday and daily transactions of securities trading.

There's currently a lot of discussion around machine learning – is its impact being exaggerated?

Per Nymand-Andersen: No, though it may take some time for it to enter the

mainstream. Central banks are well equipped to experiment with these new techniques. Machine-learning – or artificial intelligence – techniques may be utililised to find new and interesting patterns in large datasets, visualise such datasets, provide summary statistics and predictions and even generate new hypotheses and theories derived from the new patterns observed.

How can central banks overcome the communication challenges presented by black-box models?

Per Nymand-Andersen: Communication is key as part of conveying and obtaining support for your policies. This includes providing the underlying evidence and assessments that lead to these decisions. Off-the-shelf models and their components should be transparent and available for replication. They should comply with similar statistical quality standards as those that already prevail, such as transparency of sources, methodology, reliability and consistency over time. Big data and models remain tools to assist with the decision-making process, but are only of substantial value if they are appropriately understood and analysed. There is no automaticity in complex decision-making.

What other pitfalls exist in the use and interpretation of big data?

Per Nymand-Andersen: One misperception of big data I hear of frequently is that we do not need to worry about sample bias and representativeness, as large volumes of information will supersede standard sampling theory given that the big data sources provide *de facto* census-type information – this is incorrect.

For instance, access to all tweets would mean access to the characteristics of the entire tweeting population – corporates and members of the general public using a Twitter account. But the characteristics of this population may differ from those who do not tweet and are therefore excluded from the sample dataset. Thus, not all groups are represented by data sourced from Twitter and their characteristics may vary across countries, cultures, nationalities and ages. Therefore additional information is required to adjust the figures and to gross these up to the entire population as part of securing the quality of data – particularly if the aim is to extract signals and indicators on household sentiment, or to start producing household indexes using Twitter.

Secondly, statistical corrections will still have to be made for other features relating to unit measurements, double-counting – re-tweeting the same message – over-representativeness and over-fitting of models. In an event-driven context – such as tweets or internet searches – volume changes do not necessarily refer to reporting units or to changes in demand. Take, for instance, the increased focus on the emissions scandal in the automotive industry – and the subsequent expected increase in internet searches and tweets at that time. These potential increases in searches and tweets may be driven by concerns and wanting to observe the impact of the scandal, rather than an increased interest in purchasing cars. Therefore, the data has to be adjusted if it is to be used as an indicator.

Thirdly, another pitfall refers to the misperception that correlation means causation. A high correlation between variables does not necessarily mean causation. Thus, no conclusions can be drawn purely on the basis of correlations

between two variables. The similarity could be coincidental – additional controls therefore need to be conducted.

A fourth – and equally important – pitfall refers to ensuring sufficient quality. Statistical quality cannot be taken for granted and needs to be taken seriously to provide an accurate reflection of the structure and dynamics of our economies. Large datasets do not speak for themselves – they have to be described and contextualised before they can provide useful insights. Similarly, it is important that new big data sources are transparent in terms of their methodology and how data is generated. Otherwise, the value of policy advice and forecasting using big data will be seriously undermined.

Data governance seems to be a challenge at many central banks – how successful has the ECB been in ensuring the right people have access to the right data?

Per Nymand-Andersen: Data governance is vital for credibility and trust in institutions. Strict confidentiality protection is crucial, and is well defined in European legislation – council regulation 2533/98 and ECB/1998/NP28.

When moving from macro-level to micro-level data and statistics, governance components must be revisited to ensure they apply to granular data. This means finding flexible methods to clearly define the roles and responsibilities of each actor, and organising access profiles, controls and audit trails at each point of the production process. This is to manage new incoming data and metadata – including linking and mapping data dictionaries – organise updates and revisions, enrich data and micro-aggregation methods, prove representative statistics and ensure users can benefit from the disaggregation of statistics. Data governance is challenging and remains important for central banks. Several central banks and organisations acknowledge that data and statistics are strategic assets for the institution and are initiating organisational changes to create chief data officer functions as part of streamlining and enhancing data governance across the organisation.

When it comes to private data sources, one could ask whether big data sources on individual behaviours and patterns could be commercialised, or if they should become a public commodity that complies with statistics confidentiality and privacy rules. My vision would be for the availability of new private data sources in public domains to be fostered. The wealth of information and derived knowledge should be a public commodity and made freely available – at least to universities, researchers and government agencies.

How must central banks adapt their operations to cope with larger volumes of data?

Per Nymand-Andersen: Central banks' IT environments must be significantly more flexible to accommodate and manage multiple and large volumes of data streams, and provide the necessary tools for data explorations. I believe this is well under way within the European System of Central Banks. More importantly, central banks need to attract data scientists with the ability to structure and process unstructured big datasets and swiftly perform statistical and analytical data exploration tasks. These new skills are in high demand, and central banks must compete with attractive private employers. Central banks will also need to invest in training and reschooling existing

staff to acquire these new skill sets. Pooling available resources within the central banking community or creating partnerships are interesting options in this regard.

What areas of big data analysis are still closed off because of computational limits?

Per Nymand-Andersen: In today's IT world, storage and processing power may no longer be the main bottleneck. IT environments must become significantly more flexible as a vital supporting tool for generating knowledge and value.

How might central banks' use of data change over the next 10 years?

Per Nymand-Andersen: The data service evolution is changing our society, the way we communicate, socialise, date, collaborate, work and use and share data and information. Applying technological enhancements to valuable services – such as mobile devices, cloud services, artificial intelligence and the internet of things – will also change central banks' data usage. Ten years from now is a lifetime, and becomes a vision.

My vision would be that public authorities move from data owners to data sharers. This requires mindset changes, collaboration and trust. I am a great believer in linking and sharing micro-level datasets among public authorities in such a way that it is only collected once, then made available to other authorities, while ensuring the necessary privacy and confidentiality protection. In our central banking world, this would relate to relevant national and European authorities in banking, markets and supervision. A precondition for managing, linking and sharing micro-level datasets is the use of standards and mapping datasets, so we have a common semantic understanding of how to describe financial markets, actors, instruments and their characteristics.

When one has worked with data for more than 20 years, one must conceptualise and structure the pool of unstructured data. Therefore, financial market actors and financial regulators must collaborate and intensify their working relationships beyond institutional boundaries and agree on standards to become an integrated part of the digital transformation. We need a kind of "Schengen agreement for data". For instance, we must develop and adapt common identifiers of institutions – legal entity identifiers; instruments – unique product identifiers; transactions – unique transaction identifiers; and decision-makers – enabling large datasets to be linked together, irrespective of where they are physically stored. Authorities can then slice and dice multiple datasets irrespective of where the data is physically stored. It provides the ability to identify ownership structures between legal entities and their decision-makers, and clarifies the relationships between creditors, instruments, guarantees, trades and decision makers in near real time. This transparent way of showing close relationships between borrowers and lenders will mitigate excessive risks exposures within the financial system, and avoid negative spillover effects to the real economy and citizens. This is of great interest to the financial actors and will provide significant cost savings for the industry, as well as efficiency gains for public and international authorities and regulators. ❑

The views expressed are those of the interviewee and may not necessarily represent those of the European Central Bank or the European System of Central Banks.

Big data in central banks

As an active area for new projects, big data is becoming a fixture in policymaking, with an increasing number of central banks carving out a budget for data handling, writes Emma Glass.

This article reports the findings of a new survey conducted by Central Banking in association with BearingPoint during August and September 2017. Central Banking first surveyed how central banks governed, managed and processed big data in 2016, and this new survey sets out to expand on the original results. This work has only been possible with the support and co-operation of the central bankers who agreed to take part. They did so on the condition that neither their names nor those of their central banks would be mentioned in this report.

Key findings

- Big data is an active area for new projects in central banks. As a result of these, central banks are making significant changes in three areas: reforming departments, implementing new technology and upgrading systems.
- Development of credit registers is a key focus for many central banks' big data projects, closely followed by administrative sources and consolidation of internal systems.
- Big data has become a fixture in the policymaking process for central banks. Over half of survey respondents said it was an input into processes, with just over one-third saying it was a core component.
- Data collection is the greatest challenge central bankers face as the use and application of big data becomes more prevalent.
- A significant data governance gap exists in central banks. Over half of respondents said they did not have clear data governance – although work is under way to address this.
- Expertise in big data is typically shared across departments. Very few central banks have a designated big data department.
- Central banks are increasingly carving out a budget for data handling. The number of central banks with such a budget is twice that of the 2016 survey.
- An overwhelming majority of central banks have strategies in place for dealing with collecting and managing data.
- Self-developed data platforms are the most popular method for regulatory data collection, followed by Excel and document-based handling and commercial solutions from the market.

- Ensuring data quality is the greatest challenge to central bankers in the process of data collection.
- Social media is an emerging source of data for central banks: just under one-fifth of respondents collect data for policy purposes from social media.
- Central banks see benefits from collaborating on big data projects, chiefly in terms of efficiency and cost reduction.
- Considerable investments are being made in new technology to support developments in big data.
- Data mining and trend forecasting are the most popular methods for data analytics.
- Shared internal platforms are integral to working on big data, with three-quarters of respondents using them.
- Central bankers typically avoid external data storage such as the cloud. Many central bankers view external data storage as risky, particularly in terms of the security and confidentiality of data.

Profile of respondents

Central Banking received responses from 50 central banks, with 27 of those that took part in the 2016 survey[1] taking part again this year. The average staff size was 1,338 and 40 respondents had fewer than 2,000 employees. The average staff size of a central bank is 2,698.[2]

Geography	% of respondents
Europe	56
Americas	18
Asia	10
Middle East	8
Africa	6
Oceania	2
Total respondents	50

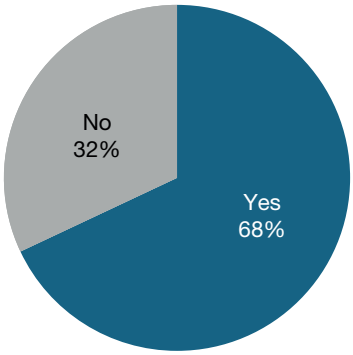
Staff size	% of respondents
<500	28
500–999	28
1,000–1,999	24
2,000–4,999	16
>5,000	4
Total respondents	50

Economic classification	% of respondents
Industrial	32
Emerging-market	30
Developing	24
Transition	14
Total respondents	50

Department	% of respondents
Statistics	58
IT	38
Other	4
Total respondents	50

Percentages in some tables may not total 100 due to rounding.

Has your central bank undertaken any new projects involving big data in the past 12 months?



Big data is an active area for new projects in central banks. These new projects have three aims: reforming central bank departments, implementing new technology systems and upgrading systems to better process big data. Just over two-thirds of respondents said they have undertaken new projects involving big data in the past 12 months. This group of 34 was dominated by central banks from Europe. Just under half this group were from industrial economies, and the remainder were evenly spread between transition,[3] developing and emerging-market economies.

Respondents explained how big data-related projects were impacting their institutions, which can be grouped around three themes. First, central banks are changing the structure of their institutions to accommodate this new area of research. Of the 34 central banks that answered positively, five referred to institutional changes, such as the creation of a data unit, in their comments. A large central bank described the structural change that has taken place: "We have created an internal team across different departments with different skills to work on specific projects involving big data and central bank issues." An Asian central banker said big data projects had been outsourced: "Our bank outsourced a research project to enhance the quality of economic forecasting using big data information." Internal discussion plays a key part in institutional change. This is recognised by one European central bank as a means to facilitate new structural changes and encourage communication on the topic. This bank launched a "big data forum", which is "intended to be a forum for internal discussion".

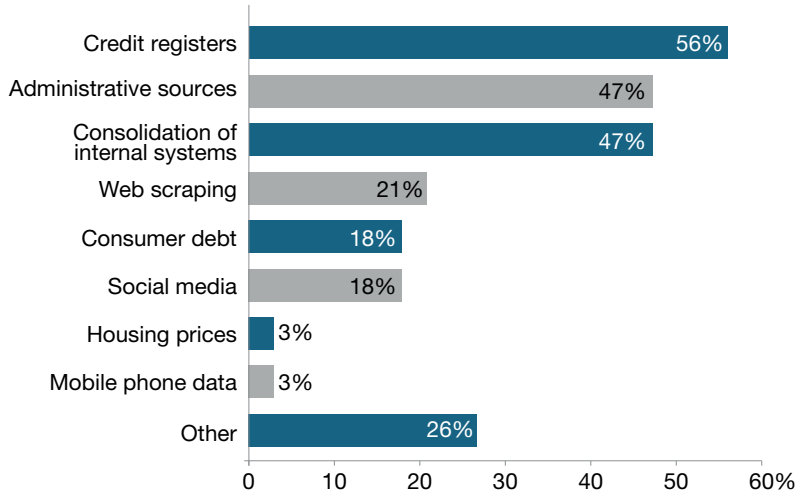
Second, central banks are implementing new technologies that draw on big data, something four central banks referred to in their comments. A European respondent referred to their own database and "common registry entities (in collaboration with national statistical services)". An Asian respondent listed four new projects that are under way: "data catalogue (internal and external), private cloud, exchange market surveillance and data collection from financial institutions".

Third, central banks are upgrading the systems they have in place to process big data. Two central banks made reference to this, with one central banker from Europe providing this extended comment: "The bank is running methodological and technological upgrades of the statistical business process under its mandate. A new innovative project that is intended to collect, validate, process and disseminate big amounts of data from different sources is under way."

Sixteen respondents said they had not undertaken new projects involving big data. Emerging-market economies made up half of this group, with only two drawn from industrial economies. The respondents who commented suggested that small steps were being made in this area, but nothing substantial. A respondent from the Americas said: "There are individual efforts by some

experts, but not in a project context." A European central banker pointedly said: "We do not intend to use big data."

If yes, in which area(s) were the projects focused?



One respondent did not reply. Thirty-three respondents provided at least one answer.

Development of credit registers is the most important focus for many central banks in big data projects. This was closely followed by administrative sources[4] and the consolidation of internal systems. Credit registers were almost always selected with another option – only one respondent checked this option alone. A European respondent who selected credit registers, web scraping and administrative sources detailed their institution's plans: "We started to plan a new credit register database suitable for AnaCredit purposes. We collected administrative data from the National Pension Insurance Institution for research purposes, and price data from retail trade companies' websites." A smaller central bank with fewer than 100 employees said: "Credit registers service will be operative from the last quarter of 2017. The present small quantities of data will increase over the next few years."

Administrative sources also proved popular among respondents. Sixteen central banks chose this option, with over half of these from developing and emerging-market economies. A common pairing was with credit registers or consolidation of internal systems. Eleven central banks chose administrative sources with credit registers, while seven chose it with consolidation of internal systems.

A respondent from the Americas who selected administrative sources and the consolidation of internal systems has centralised economic and financial databases to provide reliable, up-to-date and accurate data including Bloomberg and ASYCUDA, and its financial institution report management system comprising a data collection module, a supervisory data centre and data visualisation reporting tool.

Consolidation of internal systems was selected by 32% of respondents. Three central banks – two of which were European and from industrial economies – chose this option in isolation, while the remaining 13 chose it in

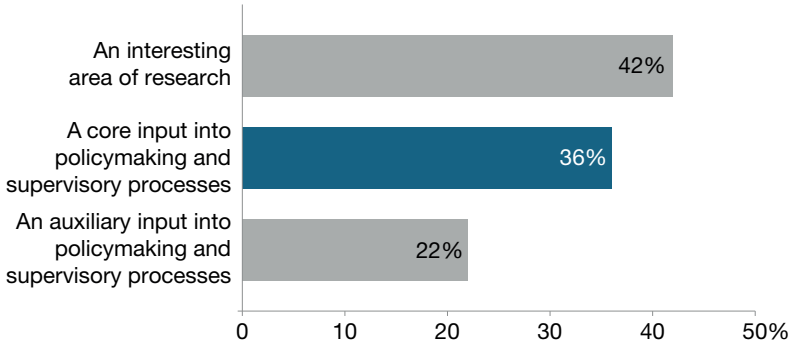
tandem with at least one other answer. Six central banks chose four or more options. These central banks gave greater priority to big data work at their institution, with one central bank in Asia pointing out that it was now part of their strategic plan: "The bank has enhanced its data analytics capability, especially the use of micro data. This initiative is part of the bank's three-year strategic plan."

A respondent in an industrial economy noted their priorities: "The considered administrative sources include information from several databases, namely securities statistics integrated systems; simplified corporate information (central bank balance-sheet databases), database accounts and other sources, which refers to payments data, which are obtained from the payment systems department."

The "other" category garnered various examples. An institution with more than 2,000 staff said it was using Google search data, while another – also in Europe – commented that its project was focused on "the broad range of financial and statistical data, as well as macroeconomic institutions". A central bank in a transition economy is looking to alternative sources for its data, primarily in commercial banks.

Consumer debt, social media, mobile phone data and house prices proved to be less popular with respondents. A central banker from the Americas said: "So far we have used social media (Twitter) data to assess public and market perceptions on central bank policies. We are working on RSS feeds and sources (mainly newspapers) to complement this analysis." All respondents who chose social media in this field also chose web scraping as an option.

Which best represents your central bank's view of big data?



Big data has become a fixture in central bank policymaking. Over half of survey respondents said it was an input into these processes, with just over one-third saying it was a core component. The number of central banks that described big data as a core input into policymaking is double that from the 2016 survey. Eighteen central banks – one-third of all respondents – reported big data as a core input in comparison to eight in Central Banking's 2016 survey.[5]

Of the 18 who view big data as a core input, developing economies made up one-third, with almost half from European central banks and just over one-fifth (22%) from the Middle East. Respondents who commented stressed the application in the area of financial stability and supervision. A European central banker explained the role big data played in supervision: "The organisation sees

data as a key enabler in meeting supervisory requirements and driving insight generation. It recognises there are gaps in the infrastructure and processes that drive the optimal use of data in the organisation."

A respondent from Asia was expanding its use: "The bank has always used data in its policymaking and supervisory processes. We are continuing to expand our data capabilities, including applications in new areas of process automation and machine learning that would have impact not just on internal operations but also on the finance sector."

In addition, 11 central banks said big data was an auxiliary input into policymaking and supervisory processes. The group was evenly split between industrial and emerging-market economies. A European respondent said: "We have acknowledged the importance of these new data sources, which might help central banks grasp a more accurate picture of economic reality in almost real time." An industrial-economy central bank said: "The use of micro-data brings flexibility to data management in a way that we can readily adjust and satisfy *ad hoc* requests, in some cases tailor-made to our customers' needs."

For a minority of respondents, big data remains an interesting area of research. However, comments made by these respondents typically noted how they expected big data to enter the policymaking processes. A central banker from the Americas said: "It is currently an area of keen interest. We expect this view to be upgraded to 'an auxiliary input to policymaking and supervisory processes' within the next 24 to 36 months." A central banker from an emerging-market country acknowledged that their view of big data had changed: "Before there was not much interest. As a first stage there is a clear interest to view big data as a research topic, answering new questions and producing new business-cycle indicators. Ultimately, our goal is to explore its potential as an auxiliary monetary policy tool."

In which area of big data have you encountered the greatest challenges to increasing the use of datasets in your central bank?

	1		2		3		4		Total
	No.	%	No.	%	No.	%	No.	%	No.
Data collection	16	37	7	16	9	20	12	27	44
Data management	11	26	6	14	12	29	13	31	42
Data processing	11	26	17	40	11	26	3	8	42
Data analysis	6	14	12	29	11	26	13	31	42
Total	44	100	42	100	43	100	41	100	–

Votes were cast using a scale of 1–4, where 1 denotes the most significant and 4 the least significant. Six respondents did not reply.

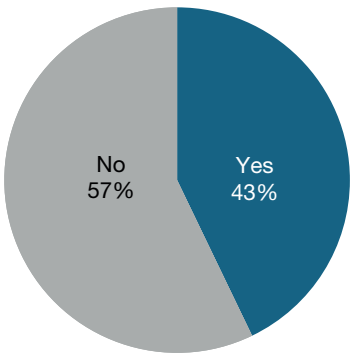
Data collection is seen as the greatest challenge central bankers face as the use of big data grows. Sixteen central banks – 37% of respondents – see this as most significant. This group was dominated by industrial economies from Europe, with the Americas, Africa and the Middle East also featuring. Two common themes emerged from respondents' comments; those regarding outsourcing data collection and those with concerns about inefficient systems. A respondent from the Americas typified the former sentiment: "We have been experiencing difficulties in collecting data from private resources (credit card companies)." In Europe, a central banker said: "We have found the greatest challenges in gathering data – not only using web-scraping techniques but also trying to buy the data directly from the private companies that own that kind of data." Equally, an Americas-based counterpart's comment spoke to the second theme: "The technological issue is always the most challenging one because of its resources and cost consumption."

Almost two-thirds of respondents found data processing challenging, ranking it either first or second in terms of significance – 28 respondents selected this option. From their comments, it is clear the issue with data processing for many central bankers is having the technical capabilities available to process the data. A central bank in the Americas said: "There isn't an infrastructure of technology that allows experts the use of datasets available, in terms of big data."

Data management was the most popular third choice, although it should be noted that 26% of respondents said it was the most significant challenge they face. Almost half of this group of 12 who placed it third were from African and Asian institutions. One of the biggest concerns raised regarding big data is managing the volume and variety of the datasets. An Asian respondent noted this issue: "Most of our current efforts are focused on consolidating internal systems to provide data consumers with a single view of data. This approach will go a long way to addressing the data management issues that arise from the volume and variety of big data."

Fifty-seven per cent of respondents selected data analysis in third or fourth place. Twenty-four respondents – more than half of whom were from Europe – gave this response.

Does your central bank have clear data governance, with defined roles and responsibilities?



One respondent did not reply.

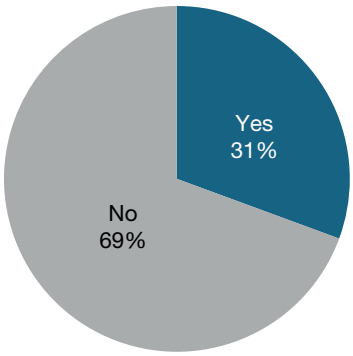
A significant data governance gap exists in central banks. Over half of the survey respondents (57%) said they did not have clear data governance, although many noted work is under way to address this. Industrial economies featured prominently in this group, and comments from these respondents stressed work is also in progress. A respondent from the Americas referred to altering policy: "We are working on designing the necessary policies to

implement institutional data governance, as well as on the proposal to create an organisational structure that supports this process." For one central bank from Asia, change had already taken place, with the establishment of a new division in May 2017: "We are planning to involve external consultants in order to organise the data management system in the central bank, develop standards, regulations, policy and definitions of key roles."

Just under half of respondents said their central bank has clear data governance in place. This group comprised ten European central banks, four from the Americas, three from Asia, three from the Middle East and one from Africa. Developing and emerging-market economies were in the majority, with two-thirds of the results. One common theme in respondents' comments was the departmental assignment of data management. A European central banker said: "The statistics department is responsible for the data collection [for] both statistical and supervisory purposes, but the increasing need for micro data requires us to rethink our current processes and create a strategy in data management."

A respondent from an industrial economy described their approach to data governance, which is the responsibility of the statistics department in co-ordination with IT: "Both departments should aim to promote the organisation of information architectures, the definitions of concepts, and the creation of metadata and catalogues, dictionaries and repositories of information. Each department of the bank has a data steward responsible for the content and management of the data it produces. There is also a master data steward, who co-ordinates the activity and oversees general guidelines."

Does your central bank have a specific function working on big data?



One respondent did not reply.

Expertise in big data is typically found across departments in a central bank – less than one-third of institutions have dedicated big data functions. Of the 34 who did not have a specific function, two-thirds were from Europe. A common theme in comments from central bankers was the dispersal of big data specialists throughout the institution. A central banker from the Americas captured this sentiment: "This function is scattered in different areas. This is a project to build multiple area task groups to deal with these issues in a unified and more efficient way." Similarly, a European respondent said: "So far we only have an internal team of experts from different fields trying to solve some policy-relevant issues using big data with the continuous help and support of the IT department."

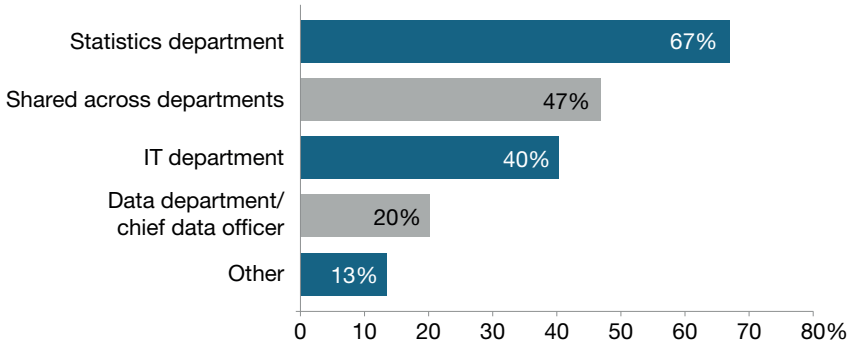
A European central banker said their institution did not have a specific function working on big data; however, a new unit has been carved out within the statistics department: "[The unit was created] to ensure the integration and sharing of data across the organisation, with operational responsibilities within

the scope of the integrated model of information management. This unit will also be responsible for bridging IT and statistical producing units regarding any big data research or developments, and is also involved in improving the corporate data warehouse and refining the data catalogue."

Just under one-third of respondents reported having a specific function working on big data at their institution. Of these 15 central bankers, two-thirds were from developing and emerging-market economies. A respondent at a small European central bank said: "In general, the monetary statistics division is responsible for data production at the bank. However, this division does not have specific function [formally] working on big data." In contrast, a large European central bank with more than 1,500 staff said: "The data operations function was set up to manage all data initiatives across the organisation. Its remit is not limited to big data but focuses on all data challenges."

Data analytics units proved to be common in Asia. Two respondents referred to them in their comments: "We established a data governance and analytics unit in 2015, which subsequently was expanded in March 2017 to a data analytics group (DAG). This serves a wide range of data-centric functions such as collecting data analytics techniques, including using big data technologies."

If yes, in which of the following departments is it located?



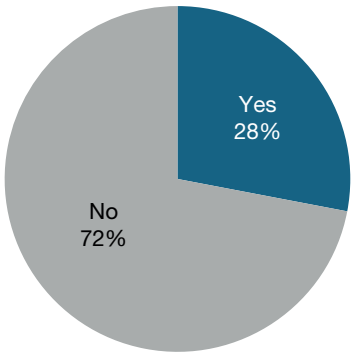
Fifteen central banks replied, seven respondents provided multiple answers.

Where central banks have a big data function, the statistics department is typically its base, though it may share some responsibility. Six central banks selected the statistics department along with another department, and six central banks said their big data work was shared across departments. A respondent from the Americas described their collaboration with the IT department: "The big data set is compiled from the banking system's institutions, then it's processed from the IT department for the statistics department's use."

Six central banks located their big data function in the IT department. Only one chose this option in isolation, with the remainder selecting a combination of options. These were, in the main, small institutions. A central banker who typified this said: "It was initially proposed to have this team more closely aligned with the business but they report to the chief information officer." Several respondents stressed the links between the IT department and their data analytics teams. A respondent from a central bank in Asia said: "The data analytics unit reports directly to the assistant governor responsible for group-

wide IT." Another respondent commented on the collaboration between the two: "DAG is a separate department, but works very closely with the IT department in implementing the necessary systems and architecture to support data collection and analysis."

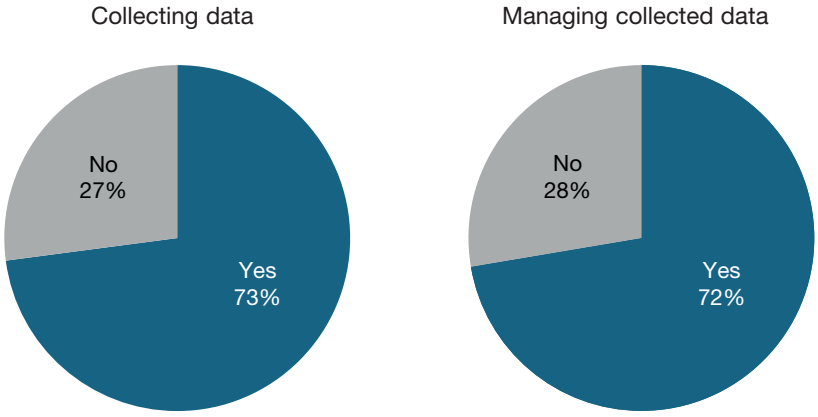
Does your central bank have a single allocated budget for handling data – including big data?



Central banks are allocating budget to data, although this is still the preserve of a minority. Those with a single allocated budget for the handling of data is double that reported in the 2016 survey.[6] Fourteen central banks – 28% of respondents – have a budget assigned for big data. Developing and emerging-markets countries figured prominently in this group, making up two-thirds. In some instances, this budget is supplied for specific big data projects. A European central banker said it was part of the bank's protocol to allocate budget to projects: "All data projects that require organisational resources need approval from the organisational investment committee." Of the 50 respondents, only two central bankers revealed their budget amount.

Thirty-six respondents do not have a single allocated budget. This group included all Middle East respondents and two-thirds of the European central bankers who took part in the survey. As many of these central banks do not have a specific big data function, budget is taken from departmental resources. A medium-sized institution said the statistics department was given a specific percentage of the budget.

Does your central bank have a strategy for collecting and managing collected data?



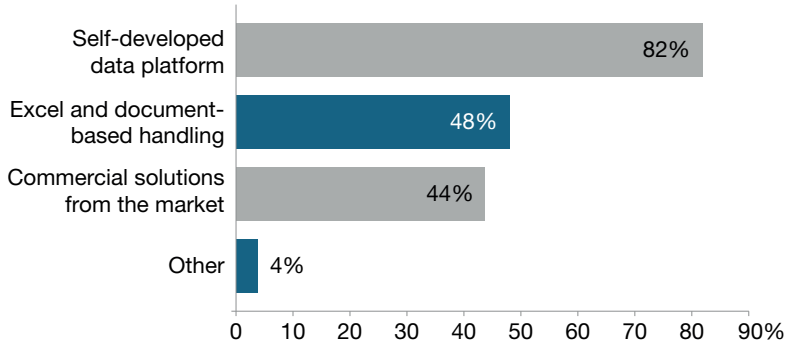
Two respondents did not reply and one respondent chose only one option.

An overwhelming majority of central banks have strategies in place for dealing with collecting data and managing the collected data. Thirty-four central banks said yes to both options, with only one central bank having a strategy for collecting data. Two-thirds of respondents who answered yes to both were European. In comments, two themes prevailed: strategies in place for IT departments and outsourcing data collection and management. A central bank in the Americas referred to the first theme: "The central bank has an infrastructure of technology that supports database management in a planned and organised manner." A European respondent epitomised the latter theme: "Data processing and collection is outsourced via regular procurement." A central banker from Europe said data was collected via an online reporting portal: "Collected data is managed differently, depending on whether it is required from an analysis perspective or to satisfy mandatory reporting requirements."

Thirteen central banks do not have a strategy for dealing with collecting and managing data. Transition and developing-economy central banks featured strongly in this group, which were typically smaller institutions. A respondent from Europe said: "The central bank has established the practice of collecting all input data through the 'single entry point' – through the department of financial statistics and review."

Of the 50 survey respondents, only one had a strategy for collecting data, but not for managing it. In their comment, they said this was something the central bank is working towards changing: "To receive the information that the bank required from financial institutions, we have the organisation, processes and infrastructure necessary for data collection and its processing. Internal consolidation processes also include data processes that allow the integration of information. However, since we have just begun to implement big data solutions, we have not yet decided how to take data to the server cluster."

What methods does your central bank employ for regulatory data collection?



One respondent did not reply. Respondents provided multiple answers.

Central banks typically use a suite of techniques to manage regulatory data collection. Self-developed data platforms are the most popular method, with almost all respondents from Europe employing this approach. While common,

it tends not to be the only method used: 29 respondents chose it in combination with at least one other. The most popular combination was all three methods. One central banker said: "The vast bulk of data collected from a regulatory collection perspective is done through a self-developed collection platform." A respondent from a large central bank in an industrial economy rejected commercial solutions: "We do not rely on external commercial solutions for confidentiality issues. However, we use both proprietary visual analytics software and open-source solutions to handle this kind of data."

Commercial solutions were indeed a delicate issue for some central banks due to confidentiality concerns. One central banker from the Americas said: "We use commercial tools for extracting, transforming and loading (ETL). Data stage is the most used tool; however, we have some internally developed applications that use Pentaho for validations and online ETL processes."

The second most popular choice was Excel and document-based handling, though this too was often employed in combination with another option. Only three central banks chose this option alone, two of which were from the Middle East.

Which represents the greatest challenge to data collection?

	1		2		3		Total	
	No.	%	No.	%	No.	%	No.	%
Processing large volumes of data	11	22	13	27	25	51	49	100
Data quality	28	57	17	35	4	8	49	100
Timely analysis of collected data	10	20	19	39	20	41	49	100

Votes were cast using a scale of 1–3, where 1 denotes the most significant and 3 the least. One respondent did not reply.

The quality of the data collected is the greatest challenge in the process of data collection. Twenty-eight respondents (57%) selected this option. Developing and emerging-market economies were in the majority here, with 10 respondents from an industrial economy and only two from transition countries. Respondents from two large central banks expanded on the issues they were grappling with. A respondent from a large institution with more than 2,000 staff said: "The biggest challenge is dealing with unstructured data and semantics."

For a central banker whose institution employs more than 3,000 staff the issue was cleaning data: "The development of risk models implies the need to implement data cleaning processes that allow the results of the algorithms to be more reliable. For this reason, data cleaning is one of our top priorities, continuing the need to process high volumes of information with better response times. Therefore, we are working to install a layer of statistical servers that will interact with the server cluster to reduce processing time and calculate complex metrics."

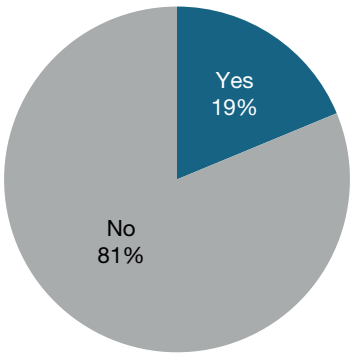
Resourcing is also an issue for central bankers, as a respondent from Asia

explained: "In collecting data from financial institutions, ensuring data quality – beyond simple validation checks such as correct summation – is paramount. However, the process of querying anomalies and assessing financial institutions' responses is labour-intensive, especially as assessing financial institutions' responses requires domain knowledge and experience."

The second most significant challenge was processing large volumes of data, said 22% of respondents. Eleven respondents were from industrial and emerging-market economies. Just over half were European, with an even representation globally. A central banker from Europe said: "The greatest challenge is in processing large volumes of unstructured data to filter out the noise and obtain the right signal. This affects the timely analysis of the collected data. Data quality is also an important issue when we deal with policy because we might have a false sense of precision and take the wrong action due to sampling composition bias and other statistical challenges connected with big data."

The third most significant challenge was the timely analysis of collected data. Thirty-nine per cent of respondents – five central bankers from the Americas and nine from Europe – chose this option. The remaining five were from Africa, Asia and the Middle East. A central bank from Asia pointed out that priorities varied by department: "Statistics-relevant departments see timely analysis of collected data as the greatest challenge, and processing large volumes of data as the least challenging. However, the information management department ranks data quality as the most significant, and timely analysis of collected data as the least."

Does your central bank collect data from social media to gauge sentiment on key issues related to central bank policy?

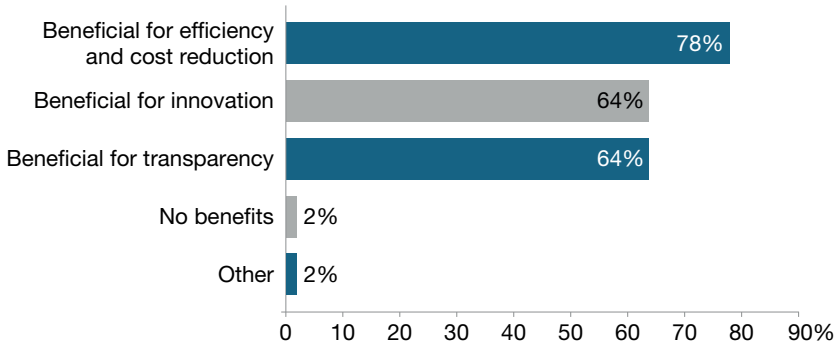


Social media is emerging as a source of data on sentiment for central banks, though this is a minority pursuit. Less than one-fifth of respondents collect data from social media to gauge sentiment on key issues related to policy. Respondents from central banks in developing and emerging-market economies in particular used this source. A respondent from Asia said: "The bank uses social listening services that collect data from social media (such as Facebook and Twitter) to gauge sentiment on key issues related to central bank policy." One emerging-market central banker said their central bank was in the process of developing a Twitter-based sentiment tool.

Thirty-nine central banks do not use social media to gauge public sentiment. More than three-quarters of respondents (81%) gave this answer, with all respondents from the Middle East and the Americas choosing this option. Several comments referenced future projects, however. For example, a central bank in Europe said: "There are some examples of this within the organisation but it does not have widespread traction at this point." Another central banker – also from Europe – said: "Some projects with that focus are already in the pipeline." In

some cases, central banks made reference to using news articles rather than social media to gauge sentiment. This European respondent did so in their comment: "Recently we have initiated a project for construction of a policy uncertainty index, based on news articles."

There are trends towards collaborative approaches in the market – what is your central bank's view on data collaboration?



Three respondents did not reply. Respondents provided multiple answers.

Central bankers see benefits from collaboration with others and those in the industry. Most see efficiency and cost reduction as the main areas in which the benefit will be felt. This was the most popular option with more than three-quarters of respondents. European central banks dominated here, making up two-thirds of the total responses. One-third (13) of the 39 central banks were from industrial economies, and 21 were from developing and emerging-market economies.

For one-third of respondents, the most popular combination was all three benefits. A common theme was the implementation of new schemes and committees that encourage collaboration. A European central bank referred to the Irving Fisher Committee, where members work on "pilot projects".

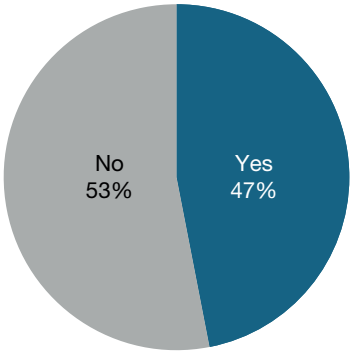
Similarly, a respondent in the Americas expressed the benefits a similar committee would provide for them: "Definitely, the collaboration – mainly of clean data and with the necessary quality – allows a greater efficiency of the processes, but to have a scheme that allows us to provide information to society in an accessible format has many benefits for society. Although, to reach this level, it is necessary to make an adequate analysis and an infrastructure design that allows collaboration with society without affecting the bank's operations."

Following closely behind this, 64% of respondents said that data collaboration is beneficial for innovation. All of the respondents from the Americas chose this option, whereas only half of the respondents from Europe prioritised innovation. Data collaboration being beneficial for transparency was also selected by 64% of respondents.

Has your central bank invested in any new technology in the past 12 months that supports developments in big data?

Considerable investment is being made in new technology to support developments in big data, with just under half of respondents investing in the

past year. Twenty-three central banks noted a change in technology in the past 12 months – a group in which large central banks from industrial economies featured prominently. A budget allocation allowed many central banks to invest in this area, with some developing their own data warehouses. A central banker from the Americas explained: "We have recently purchased a big data solution that is in the process of installation and configuration." A respondent from a large institution commented: "More powerful servers and databases have been purchased for big data."



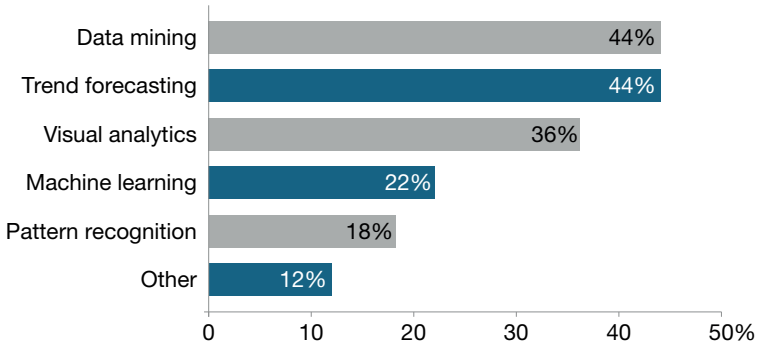
One respondent did not reply.

A respondent from an industrial economy said: "The central bank developed a proof of concept for the AnaCredit project to explore the use of new infrastructures and software solutions, namely Hadoop, for the storage and analysis of big data. This technology may also be used in the corporate data warehouse technology stack."

A respondent from Europe had been given the green light: "Budget approval and implementation process are under way." Another European central bank was ahead of the curve: "We already invested in some new technology two years ago. We should improve our big data platforms in the near future, according to the demand of new indicators from the business units." More investment is being made in this area, as exhibited by a central banker from Europe: "A new framework on reporting has been developed, covering all reporting institutions."

Twenty-six respondents indicated their central bank had not made an investment in new technology in the past 12 months. More than two-thirds of respondents were from Europe. Common themes in the responses to this question referred to investments made before the reference period and the possibility of future investment in this area.

Which of the following new methods does your central bank use to analyse big data?



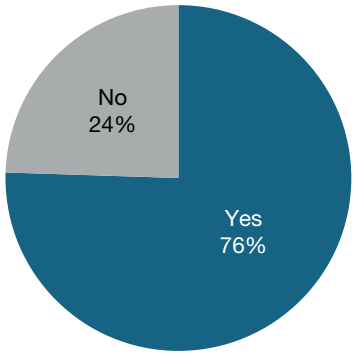
Three respondents did not reply. Respondents provided multiple answers.

Central banks employ a range of tools to analyse big data. Data mining and trend forecasting are the most popular methods for data analytics, with visual analytics and machine learning also enjoying support. Typically, central banks employ a range of methods, with just nine respondents choosing a single option. A respondent from the Americas described the multiple methods they use: "We actually use visual tools and pattern recognition in our analysis. We are also exploring machine-learning techniques for forecasting purposes." In the main, the larger the central bank, the more methods employed. This was the case for two very large central banks in Europe with more than 5,000 staff members: "We use almost all the techniques listed above as they are all important in analysing big data."

The second central bank was pleased with the resources available to them: "There is a significant analytial skillset available within the organisation."

Visual analytics was the third most popular option. Just over one-third of respondents said they made use of this. Developing and emerging-market economies made up half of the result here. A central banker from the Americas said: "Since our first challenge is the development of the risk model, we are using pattern recognition and visual analytics."

A respondent from a small European institution said the central bank has no need for big data analytics methods as "the central bank doesn't actually analyse stock of data requiring the use of the previously mentioned technologies".



One respondent did not reply.

Does your central bank have a shared internal platform to enable different departments to access data resources?

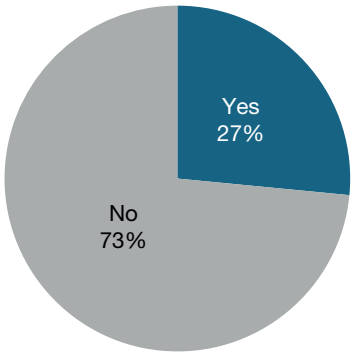
Shared internal platforms are integral to working on big data, with three-quarters of respondents using them to enable different areas of the central bank to access data resources. The same percentage gave this answer in 2016.[7]

Thirty-seven respondents have a shared internal platform at their central bank. More than half were from developing and emerging-market economies, with only one-third from industrial economies. The comments made by respondents can be divided into two categories: central bankers using commercial software, and central bankers creating their own internal platforms. A central banker whose comments reflected the first category said: "We use commercial business intelligence and data storage software, such as Teradata and MicroStrategy." A European respondent said their integrated reporting platforms used financial reporting. In the second category, central bankers were making use of internal systems.

A large central bank said: "We have a shared internal platform that allows all the members of our big data team to work on different projects based on big data." A European central banker said: "Existing financial and statistical database is available for internal users from different units of the central bank." Access to the

internal platforms is restricted in some cases, as mentioned in this comment: "The information is in an institutional database management system so that employees with access privileges can consult the information. Our information policy is that the data must be available to all who require it for the fulfilment of their functions." A similar sentiment was shared by a European central banker: "We apply a 'sandbox' environment that replicates production data and centralises it for organistional wide use where legally possible."

A minority do not have a shared internal platform. Comments indicated that central banks were planning projects in this area. A respondent from Asia said: "The central bank plans to introduce a data warehouse to collect and store statistical data. As a result, we plan to concentrate all data in one place on a single platform." A relatively large central bank said: "Information is shared internally on a need-to-know basis, but a single platform does not yet exist."



One respondent did not reply.

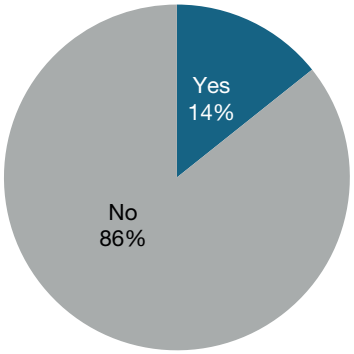
Has your central bank launched any new cyber-security initiatives in light of developments in big data?

A considerable minority of central bankers have launched new cyber-security initiatives in light of developments in big data. However, this is an area in which more respondents are looking to invest further. Thirteen central banks said they have launched a new cyber-security initiative. Comments from central bankers tended to stress concern for the management of data and frameworks in place. One of the largest central banks that responded to the survey is restructuring its institution to alleviate concerns: "We are analysing the issues related to cyber security and are designing a computer emergency response team in the IT department." A much smaller institution made a similar point: "We are in the process of defining and establishing new cyber-security frameworks for the bank, which will also include points of big data." A European central banker said: "A number of initiatives have been launched, not solely in relation to big data but in relation to data management as a whole."

Almost three-quarters of central bankers have not considered new cyber-security initiatives at their institution. A majority were from European central banks; however, there was an equal spread between developing, emerging-market and industrial economies. There were very few comments here. A central banker from the Americas, however, gave a forward-looking comment: "In the next dates we will work with computer security specialists to design the necessary measures. In this first phase the server cluster will not be exposed."

Does your central bank use external data storage, such as the cloud?

Central bankers tend to avoid external data storage – many view external data



One respondent did not reply.

storage as a risk, particularly regarding security and confidentiality of data. All respondents from Asia and the Middle East were placed in this group, along with three-quarters of European respondents. Central banks from emerging-market economies dominated, making up almost half the responses. Respondents' comments revealed two key concerns over external data storage: conflict with central bank policy and the confidentiality of data.

Several central bankers said current IT or security policies prohibit the use of the cloud. Two central bankers expressed their concerns with confidentiality in external data storage – one of whom said their concern meant they would always store internally: "Even if external data storage can offer more flexibility and other advantages, it is better to develop big data platforms internally for confidentiality issues."

An Asian central banker noted the conflict external data storage has with central bank policy: "Internal regulations restrict the use of external cloud resources."

The other was looking to solve this issue for future use: "By the nature of our operation, there is sensitive or confidential information that we must safeguard; however, the possibility of information being stored in the cloud is being studied, if this favours collaboration with society."

This view was shared by a central banker for the Americas: "We are studying the possibility of using the cloud as well as the roadmap to achieve this." A European central banker did not feel the need for external data storage: "Internal storage is enough for existing databases."

Seven central bankers said they use external storage. One European said this was introduced in 2017 for open data. A respondent from the Americas said: "We externally host non-confidential data for public use. Confidential data is only stored on servers owned and managed by the central bank." Microsoft Azure, a cloud-computing software, proved popular as a storage facility to central bankers, with several making reference to it in their comments. ❑

Notes

1. Emma Glass, Big data in central banks: 2016 survey, *Central Banking Journal* 27 (2) November 2016, last modified November 7, 2016, www.centralbanking.com/2474825
2. Emma Glass, ed., *Central Bank Directory 2017,* London; Risk Books 2016, xix.
3. A transition economy is an economy changing from a centrally planned economy to a market economy.
4. Administrative sources is the organisational unit responsible for implementing an administrative regulation (or group of regulations), for which the corresponding register of units and the transactions are viewed as a source of statistical data: OECD, *Glossary of statistical terms*, last updated on February 4, 2004, https://stats.oecd.org/glossary/detail.asp?ID=7
5. Emma Glass, Big data in central banks: 2016 survey, op. cit. page 72.
6. Emma Glass, ibid. page 75.
7. Emma Glass, ibid. page 76.

Collaboration is key to central banks making the most of data

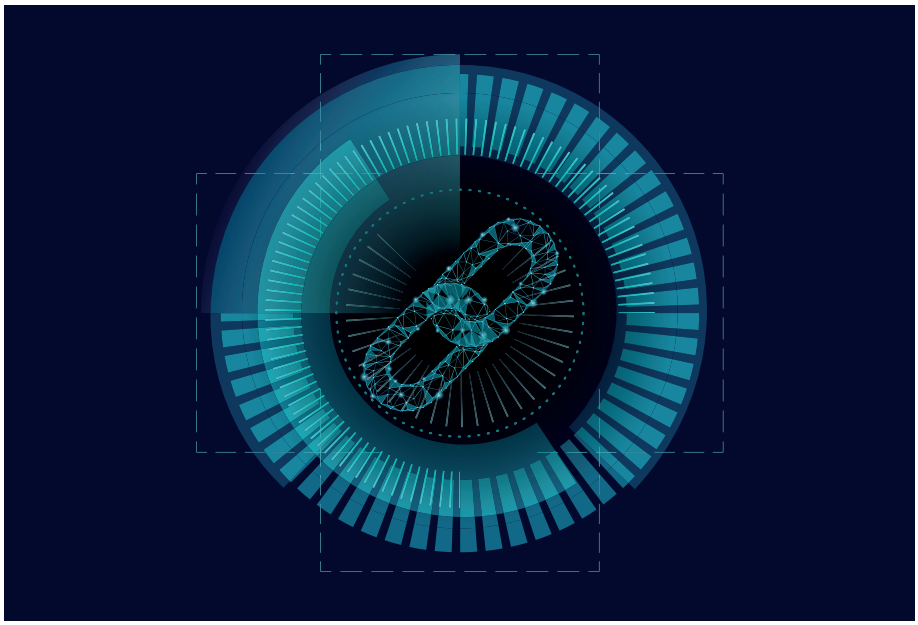
As data becomes increasingly significant for central banks, working with the right people and employing the right approaches is central to meeting objectives, says Maciej Piechocki of BearingPoint.

BearingPoint®

The financial crisis exposed central banks' and regulators' need for high-quality, comparable and timely data on the financial system. Since then, central bankers, policymakers, supervisory authorities and standard-setters have been collaborating to harmonise and standardise regulatory reporting for banks and insurance companies. There are ongoing debates about how the financial services industry could be better and less onerously supervised through a smarter approach, how central banks are dealing with data – big data in particular – and how they are becoming data-driven enterprises.

In view of these developments, it is mission-critical for central banks to reshape their data management and automate processes, as well as find innovative approaches to deal with data and big data. Automation helps minimise risk, reduce errors, increase transparency and deliver a better basis for decision-making, and regulatory and risk technology – regulatory technology (regtech) and supervisory technology (suptech) – support central banks with this.

The Basel Committee on Banking Supervision (BCBS) recently issued a consultation paper examining the benefits of regtech for the financial industry, as well as for regulators and – in particular – banking supervisors.[1] According to the paper, the effective development and application of regtech offers an opportunity to automate regulatory reporting and compliance requirements, and facilitate cross-sectoral and cross-jurisdictional co-operation for improved compliance – regtech can enable effective risk-data aggregation and risk management. Quoted in BCBS 415, BearingPoint – with its Abacus platform as a representative example of such regtech/suptech – signifies developments in this area. In bringing its regtech/risktech to the market, BearingPoint facilitates the conception and realisation of innovative approaches to regulatory reporting, such as shared utilities, integrated platforms for data management and analytics, and Regulatory-as-a-Service. Although unique in its conception, the Austrian approach could be the future of regulatory reporting.



The Global Partnership for Financial Inclusion also referred to the AuRep project in its report, *Digital financial inclusion: Emerging policy approaches*, which gives directions for regulatory reporting and supervision on a global stage. The authors name AuRep as an example of the "input-based" approach that "leverages new technologies – or regtech – which enable regulators to capture more granular data on financial sector activity, including activity by new market entrants, or related to new digitally enabled delivery mechanisms or products, while reducing the reporting burden on regulated institutions."[2]

Because of the vast amount of structured and unstructured data from various sources, decision-making is now more difficult than before, and central banks need adequate solutions to analyse this data. A crucial point is how to mine all this information from different sources exhaustively and at reasonable cost.

The specific challenge for central banks in the sense of an effective 360-degree risk-based supervision is to rapidly access and effectively manage, process and analyse the increasing amounts of supervisory, statistical and markets data and big data in a timely manner. The real-time or near-real-time access and efficient processing are especially regarded as critical factors due to limitations in human and IT resources.

Data and big data are key topics for central banks worldwide. But first, combining data with the right people, technology, processes and collaboration models – such as those in Austria – will allow central banks to leverage it for their missions and objectives. ❑

Notes

1. Bank for International Settlements, *Implications of fintech developments for banks and bank supervisors – Consultative document*, BCBS 415, August 2017, *www.bis.org/bcbs/publ/d415.htm*
2. Global Partnership for Financial Inclusion, *Digital financial inclusion: Emerging policy appproaches,* 2017, *https://tinyurl.com/y84efpqj*

Tapping into big data's potential

Central Banking convened a panel of experts to discuss how central banks can harness big data for their needs, hopefully without falling foul of some of the many pitfalls that await.

BearingPoint®

The Panel

Alastair Firrell
Data Scientist, Advanced Analytics, Bank of England

Maciej Piechocki
Financial Services Partner, BearingPoint

Leif Anders Thorsrud
Senior Researcher, Monetary Policy Research, Norges Bank

Bruno Tissot
Head of Statistics and Research Support, Bank for International Settlements, and Secretary, Irving Fisher Committee on Central Bank Statistics

Big data has emerged as a hot topic in central banking circles over the past few years, with many official institutions reviewing their data policies while increasing their staffing and systems capabilities in a bid to unlock big data's potential. Public attention has often focused on the scraping, cleaning and analysis of unstructured publicly available data, which offers the potential for macroeconomic nowcasting, for example, or securing real-time communications feedback.

Central banks also hold large amounts of structured, sometimes confidential, data that they must categorise, hold securely and process in an appropriate manner – sometimes using big data techniques. Then there are the multi-faceted challenges associated with making use of all these different combinations of data, including textual data. In this forum, our panel discusses the results of the *Big data in central banks* 2016 survey, infrastructure support for large datasets, the effectiveness of central banks' efforts, the real impact of machine learning, changing regulatory views, support from the executive level, and staffing and systems.

From left: Bruno Tissot, Leif Anders Thorsrud, Maciej Piechocki and Alastair Firrell

Central Banking: According to last year's survey on big data in central banks, most central banks think of big data as large, unstructured, external data – is this an accurate perception?

Leif Anders Thorsrud, Norges Bank: Big data is something of a buzz-word. Data is as useful as the questions you ask of it – big data is like any other type of data in that respect. Maybe the reason this came up was because most central bankers are more familiar with structured data, big or small. Unstructured data is more exotic, so that's why people tend to think big data is something new – and therefore must be unstructured. Clearly though, big data can be both structured or unstructured.

Alastair Firrell, Bank of England: In a lot of cases, big data is just data that your current systems don't cope with very well. It's bigger, more varied or stranger than the data that you already deal with. Central banks have traditionally been good at middle-sized – particularly aggregated – numeric data, so we are seeing a lot of interest in textual data and data that is unstructured, semi-structured or variably structured – so it's not white noise, it's not that unstructured and we're seeing some properly large datasets. Although in many cases what we would call large is not the same as a lot of the data community would.

Maciej Piechocki, BearingPoint: With big data you have different aspects, and there is relevance to how central banks deal with the data in general. When you look into the responses to the survey, they clearly show that, although it is unstructured data as far as the research is concerned, it could be structured and voluminous for other purposes – such as the credit register. I think there is a question about what the data is used for, and not so much the size or the structured versus unstructured demarcation.

Bruno Tissot, Bank for International Settlements (BIS): There are two camps. Firstly, there are those who say big data is primarily the type of unstructured data the private sector is dealing with. According to a recent BIS review, central banks are clearly interested too, for example, in looking at internet searches for nowcasting. A second area that is really key for central banks is in dealing with


Bruno Tissot

very large administrative and financial datasets. It is not simply because it is large that makes it big data, but because it is large and complex. In addition, you sometimes need big data techniques to facilitate/improve the analysis of relatively simple structured datasets.

Central Banking: There has been a lot of use of big data in research at central banks, but has there also been an important role for actual policy-making?

Leif Anders Thorsrud: In terms of big data, my impression is that most of it has been going on in research departments in central banks, but I think we are now seeing more research being taken into policy analysis. Norges Bank has two different big data pillars – one more focused on structured, granular data types, and one where we use unstructured data types. Both motivated by policy questions, and with the goal of making research policy relevant.

Maciej Piechocki: That is an interesting topic from the timing perspective because for policy proposals there are often topics that start with the research department – and we should not forget that central banks can mandate certain data collections. Let's say you find it interesting to look at certain granular data on the housing market, what follows are things like the European credit register, for example, so central banks are realising they need it on a more regular basis.

Bruno Tissot: Regarding the private sector type of big data, it is fair to say that, on average, central banks are just starting to explore the internet of things, it's not really in production on a large scale. But then there is the big part – dealing with administrative, commercial and financial micro-level datasets. What has really increased since the financial crisis is the need to manage this large and expanding amount of information and go beyond the aggregates, by making use of available micro datasets. This is a key factor driving central banks' interest in big data techniques.

Central Banking: What are some of the uses of big data – filling in for a void in statistics or nowcasting requirements, for example? What progress is being made there, and are there concrete examples of where that really adds value?

Leif Anders Thorsrud: There are several projects where we have used unstructured textual data from newspapers, and crunched that using machine-learning techniques to put it into a nowcasting framework. We have found really good results when we do that. I've been working a lot with standard time-series models, model combinations and forecast combination techniques – what we typically see is that we have different models, they work reasonably well, but it's the data you put into it that determines how well you'll do in terms of predicting the present.

Maciej Piechocki Financial Services Partner, BearingPoint
Dr Maciej Piechocki is a partner at BearingPoint, where he is responsible for regulatory and risk technology. Over the past 10 years he has specialised in digitalisation, regulation and big data, especially in the financial services sector. Piechocki is responsible for delivering services and solutions to clients in regulatory reporting, management and analytics. During his career, he has worked with a number of regulators worldwide, such as Deutsche Bundesbank, the European Banking Authority, the European Central Bank, the Polish Central Bank, the US Securities and Exchange Commission, Japan's Financial Services Agency and the Chinese Ministry of Finance. Piechocki has also worked with a number of large banks, insurers and listed companies.

Central Banking: Are there any areas where big data isn't living up to its full potential?

Alastair Firrell: There are certainly areas in which it is very challenging to get value. We've utilised Twitter in a few scenarios, looking at specific events or for particular quantities of information or trends, and there's a lot of noise in there. If you're happy that you're looking for the presence of a particular firm, then maybe that's fine as they will always be tagged in the same way. But trying to get sentiment out of something like that can be wobbly.

That can even just be down to a lack of context. For example, the reuse of Mervyn King – there's Mervyn King, the governor of the Bank of England; Mervyn King, the bowls player; Mervyn King, the darts player, and so on – you get it at the wrong time and suddenly there's a spike, but people are talking about a different Mervyn King. These are not necessarily big data problems, just data problems – but if you are receiving a constant stream and don't have a big context around them and someone sitting there cleaning them, this can create problems quite quickly.

We have had a lot of good results from news and our own internal textual data. For example, we have regional agents who talk to firms and write up their interviews, using that information to look over a lengthy time period for indicators of different discussions and hot topics. There is an awful lot of value to be had, but as with all data analysis you have to know the context – whether what it is telling you is really meaningful.

Bruno Tissot: It seems to me that we are just at the beginning of making sense of the increasing volume and variety of data we can access. Micro-level datasets can be very complex. Sometimes you are merging information from different and inconsistent sources; so you have to choose among them, and this choice may depend on circumstances. You may also want to aggregate granular information in a way that can also evolve over time.

A good example is what we do at the BIS to compute international debt issuance statistics. We aggregate micro-level information based on the residency concept, to compute all the debt issued by the economic agents that are located in a given country (in line with System of National Accounts principles). But we can also aggregate the same statistics on a so-called "nationality basis", by looking at all the debt issued not only by a resident firm of the country but also by the foreign entities controlled by this national firm – these affiliates are located outside

Maciej Piechocki

of the country and are therefore not captured by a residency-based framework. Constructing such nationality-based statistics can be quite challenging: one has to identify the perimeter of global firms, reclassify their individual units and consolidate granular information at the group level.

Central Banking: A lot is going on that is not this internet unstructured data. There is a lot of work with BIS involved too, so in terms of regulators holding large pools of data, are they likely to remain granular with regard to data analysis?

Bruno Tissot: We are facing a revolution in financial statistics that perhaps in the future people will compare to what happened in the 1930s for the real economy. At that time, the Great Depression influenced the development of the national accounts framework. Similarly, the recent financial crisis has triggered unprecedented efforts to collect more information on the financial sector – especially in the context of the Data Gaps Initiative endorsed by the Group of 20. Large micro datasets have become in high demand in this context. For instance, you cannot just look at a group of various financial institutions in an aggregated way, you must also look at those that are systemic on an individual basis. Or you need to have a sense of the distribution of macro aggregates and look at "fat tails", and so on.

Central Banking: How are you using big data techniques with regard to textual data – what projects do you have running?

Leif Anders Thorsrud: Norges Bank has an unstructured big data project, mostly using business newspaper data for nowcasting applications, but also for more structural oriented analysis. We are working on US data to do the same there, constructing business cycle indicators, pricing returns on the stock exchange and using basically the same type of raw data throughout all of these applications – we get really consistent results. In terms of techniques, we use a blend of the traditional toolkit and things from the machine-learning literature – it is about dynamic verbal selection techniques and clustering algorithms.

Central Banking: How can individual-level payments systems data be used to answer practical questions at a central bank?

Alastair Firrell: There is a range, so some questions that can be answered are primarily operational: How is your payment system doing? Which scenarios lead to system blockages? Is there a way to change the way banks and institutions interact with your payments systems to free that up, or to inform your IT personnel going forward?

There is information within the payments that is not just numeric bank-to-bank payment values, but whole streams from where the payment comes and where it

Alastair Firrell Data Scientist, Advanced Analytics, Bank of England

Alastair Firrell is a data scientist in the advanced analytics division of the Bank of England. He specialises in analytical programming, data modelling and handling, data visualisation and reproducible research. His current research includes text analysis of central bank minutes and speeches and the use of graph databases for payment analysis. Firrell has more than 15 years of experience in central banking technology – primarily in data and application development. He has designed and built analytical systems and databases for many large datasets, including the Capital Requirements Directive IV banking regulatory collections and high-value payments data. Firrell earned an MSc in data science from Dundee University in 2017.

goes to, and you have at least a certain amount of information about the kind of institutions or people at each end of those. From this you can start looking at the impact of decisions – changes in sectoral trading, for example, and geographic dispersion. So there's a reasonable amount to be done – as well as looking at things such as anti-money laundering and financial crime, there is the imperative for anyone working on payment systems to be checking what's going on.

Bruno Tissot: Apart from the traditional role played by central banks in payments systems, there are also new uses of these data for economic analysis. In Portugal, where tourism is important, payments systems data has been used to assess its impact on the economy. But the use of specific datasets will often depend on the policy question you have in mind, so not all countries will have the same practices.

Central Banking: In last year's survey, many central bankers cited a lack of support from policymakers as the most significant challenge to increasing the use of datasets in their institution. What needs to change for big data analysis to gain more support?

Maciej Piechocki: With topics on data – especially big data – there are two issues. One is that regulators and central banks are getting more data – there's a lot coming up, and you need to install the governance that can handle it properly. The second issue is accountability – the more data you have, the more accountability you need to figure out the information and get valuable insights.

Bruno Tissot: Our recent Irving Fisher Committee work shows that there is strong policy support for exploring big data within central banks. But this involves a lot of costs, resources and time. More importantly, a holistic approach is needed to ensure a continuum between the IT organisation, the collection of data, the statistical processing and policy use. Ensuring a vast amount of information is not just collected and prepared, that it is useful, really is key.

Alastair Firrell: There is the desire from senior personnel to use the data – seeing there is benefit whether it is the larger or the more varied datasets, to complement or completely take over the traditional sets. If you don't see the value, you're never going to sponsor it. Then there is the desire to actually manage it

Alastair Firrell

properly – to put the structures in place, instead of just saying "that's great, let's use that", hoping they magically appear.

Leif Anders Thorsrud: It is important to clearly separate what is in production and what is in the development phase. There are different requirements in these environments – in the development phase the researcher or the analyst working on the data may be more in charge of the data, but if something useful comes out of it and it goes into the production phase then issues of ownership and data governance become more important.

Maciej Piechocki: The survey is clear about executive sponsorship, and I think this is regardless of whether it's a research type of project or a regular production project. I agree on the kind of operational governance of these topics, but the executive sponsorship – if it is the chief data officer or head of statistics – is less relevant but helps to get these topics moving forward.

Leif Anders Thorsrud: You need some support, but in terms of how many bells and whistles you put onto the governance of a short-term or exploratory project, I think it is important to have a fair degree of flexibility. If not, it will be years before you are actually able to do something with that data.

Central Banking: Eighty-five per cent of survey respondents said there was no single allocated budget for data, and there is also the issue of whether to have a chief data officer – someone who flies the flag for data at an institution. Do you think the fact there is no single budget is part of the problem?

Bruno Tissot: It is key to extract information from the data collected. To this end you need an IT infrastructure, adequate statistical applications, sometimes legal and HR support – there is a full production chain to get from data to information. This requires good co-ordination. But whether central banks should have a specific way of organising themselves – set up a data framework or a "data lake", appoint a "chief data officer" – depends on circumstances. And it is not the key issue; what matters is not as much the organisational structure than the coherence of the information management process to transform "data" into (useful) "information".

Central Banking: A lot of key developments in big data have come from IT personnel – not necessarily from the front-end parties that have been calling for use of it. They are saying that some traditional economists' mindsets may be obsolete. Have you encountered this?

Alastair Firrell: Because there is an awful lot of blather around what big data is and the value of machine learning, many don't see concrete examples of what

Leif Anders Thorsrud Senior Researcher, Monetary Policy Research, Norges Bank

Leif Anders Thorsrud is a senior researcher in monetary policy research at Norges Bank, and a researcher at the Center for Applied Macro and Petroleum Economics at BI Norwegian Business School, where he earned a PhD in 2014. His areas of expertise include time-series analysis, factor models, business cycles, energy economics, monetary and fiscal economics. Thorsrud's research agenda currently centres on how unstructured data sources can be used to understand macroeconomic fluctuations. Prior to obtaining his PhD, Thorsrud accrued more than six years of experience as a model builder and forecaster at Norges Bank and the Reserve Bank of New Zealand.

can be done. These ideas buzz around but don't necessarily make it to where they need to. I think it is incumbent on technicians and data people to try things out and ask the economists, statisticians and policy people the right questions: if we did something a bit like this, does that spark any ideas?; if I show you clustering, does that make you think?; if I show you anomaly detection, does that capture anything?; or if I show you topic modelling out of text, does that grab you in any way?

Otherwise, people will say: "These policymakers never ask us for anything, just the same old stats", and the policymakers will say: "I don't really know what data we've got, and I don't know how we can use it, so I'm just going to ask for the same old stats". With this dialogue, is not always easy to marry the burning question with the person who knows how they might be able to answer it.

Maciej Piechocki: It's also a generation question. Some skilled IT graduates are coming from universities where they have access to all types of open-source data. They are coming to a central bank and are used to working with these tools, and they have innovative ideas to take top-notch technologies and to popularise them within central banks.

Central Banking: How useful has machine learning been? Leif has already described it as a new buzzword, but what real impact is it making?

Leif Anders Thorsrud: Machine learning comes with a new set of tools, and it is always better to have more tools in the toolkit. I think it is useful – we are applying it in most of our work these days alongside more standard tools.

Alastair Firrell: We use anomaly detection in trading patterns and the like, and this would be a very traditional machine-learning thing. We're using it in a more fuzzy way for more fuzzy tasks around extracting, for example. Extracting information from job advertisements to find out truly what the job is, for example – which is harder than you would think. It's about a toolbox full of tools, and at different times you'll use different ones.

Bruno Tissot: These techniques can address things that we saw as important during the crisis but are difficult to model – non-linearities and network

Leif Anders Thorsrud

analysis, for example. But the choice of a given technique will depend on the questions you face. It's very good to have new, sophisticated tools available, but the risk is to develop black boxes, which cannot deliver meaningful messages. It is essential to explore these techniques, work on specific projects and, perhaps more importantly, to define exactly the question you want to answer. This exploratory work can be shared within and across institutions. Central banks want to see precisely what other central banks and authorities are doing in terms of big data projects.

Central Banking: Big data is not always representative data – how should central banks ensure they have adequate accuracy, confidentiality, responsiveness and representativeness?

Bruno Tissot: This is an important issue that is sometimes overlooked. People tend to think that, because it's a very large dataset, by definition it is a reliable source of information. But you cannot really judge the accuracy of a dataset if you don't know its coverage bias, which can be significant. Even extremely large big data samples can compare unfavourably with (smaller) traditional probabilistic samples – that are, in contrast, designed to be representative of the population of interest. We should be mindful of these limitations: the risk is to have misguided policy decisions if they are based on inaccurate data.

Central Banking: What are the challenges people have to deal with when coping with unstructured open-source, textual and confidential data?

Maciej Piechocki: I think that proper data management and data strategies are key here; for many central banks the challenges lie in matching different datasets – matching findings from research on unstructured data with semi-automated analyses that you can run on the granular datasets. The matching of data is still at the beginning, but we are starting to see a lot of streams on big data in the research area and a lot of dynamics in terms of the granular datasets going beyond aggregate. What I have not seen much is this being brought together.

Central Banking: What is the optimal way of storing data? Most central banks still use their own data platforms rather than commercial platforms, so what in-house solutions do they have and why are they superior?

Alastair Firrell: Central banks face a big security challenge regarding their ability to use cloud infrastructure. Various other organisations might be able to farm a greater proportion of their data out to cloud infrastructure – at least for storage, but ideally for processing, analytics and all sorts of other things.

Bruno Tissot Head of Statistics and Research Support,
Bank for International Settlements (BIS) and Secretary,
Irving Fisher Committee on Central Bank Statistics

Bruno Tissot has worked at BIS since 2001 as a senior economist and secretary to the markets committee of central banks in the monetary and economic department, and later as an adviser to the general manager and secretary to the BIS Executive Committee. Between 1994 and 2001, he worked for the French Ministry of Finance. Tissot is currently head of BIS's Statistics and Research Support and is a graduate from Paris-based École Polytechnique and the French Statistical Office National Institute of Statistics and Economic Studies.

Maciej Piechocki: You should never answer a question about storing data before you know what you are going to do with it – but interestingly for many central banks the question of storage comes first. If you want to run big data techniques there will be different storage, if you want to do simple querying on very structured data then a structured query language database will do well.

Central Banking: We've seen databases experiencing problems recently and there are concerns around whether there are techniques that could detect issues such as these in the future. Are central banks on top of potential breaches?

Bruno Tissot: It is at the top of the agenda – because of the importance of the information the central bank has access to, you may have privacy issues, you may have a legal issue if there is a leak of confidential information, and you may face financial consequences. Key is perhaps reputation risk. If authorities collect confidential private information and this information is not protected adequately, it can be very damaging for their reputation and credibility.

Maciej Piechocki: It's a critical piece of the infrastructure, especially on the collection side. The US Securities and Exchange Commission uses a collection portal that collects data from listed and regulated identities, which is what every central bank is doing as well – and there is lots of sensitive data. If there is also personal data that will come into the game that is extremely exposed to cyber risk.

Alastair Firrell: The techniques and the ability to apply them are there, for central banks to spot a lot of these issues. Security is paramount, even if it constrains our use of commercial platforms for bigger data analysis. ❑

This is a summary of the forum that was convened by Central Banking and moderated by Central Banking's editor, Christopher Jeffery. The commentary and responses to this forum are personal and do not necessarily reflect the views and opinions of the panellists' respective organisations.

Watch the full big data in central banks webinar, Tapping into big data's potential*, at www.centralbanking.com/3309881*