Big data and central banking

Overview of the IFC satellite meeting

Bruno Tissot[1]

1. Introduction – Big data issues for central banks

Background

"Big data" is a key topic in data creation, storage, retrieval, methodology and analysis. The private sector is already using data patterns from micro data sets to produce new and timely indicators. For central banks, the flexibility and real-time availability of big data open up the possibility of extracting more timely economic signals, applying new statistical methodologies, enhancing economic forecasts and financial stability assessments, and obtaining rapid feedback on policy impacts.[2] Yet the public and private sector may have different areas of concern, not least on account of data quality.[3]

As confirmed by a recent IFC survey,[4] the central banking community is indeed taking a strong interest in big data, particularly at senior policy level. A key message of the survey was that big data could prove useful in conducting central bank policy, and that it was perceived as a potentially effective tool in supporting micro- and macroeconomic as well as financial stability analyses.[5]

Yet central banks' actual use of big data is still limited, due mainly to resource and IT constraints, but also to more fundamental challenges. In particular, exploring big data is a complex, multifaceted task, and any regular production of big data-based information would take time, given the lack of transparency in methodologies and the poor quality of some data sources. From this perspective, big data may also

[1] Head of Statistics and Research Support, BIS, and Head of the IFC Secretariat (Bruno.Tissot@bis.org). The views expressed here are those of the author and do not necessarily reflect those of the Bank for International Settlements (BIS) or the Irving Fisher Committee on Central Bank Statistics (IFC). This overview benefited from comments by K Hennings, R Kirchner and J Mehrhoff.

[2] As evidence for central banks' increasing interest in using big data, see D Bholat, "Big data and central banks", Bank of England, *Quarterly Bulletin*, March 2015, https://ssrn.com/abstract=2577759.

[3] For instance, while online retailers targeting potential customers based on past web searches might find it acceptable to be "wrong" once out of five times, official statisticians would usually consider such an 80% accuracy level as inadequate.

[4] See Irving Fisher Committee on Central Bank Statistics, "Central banks' use of and interest in 'big data'", October 2015.

[5] See in particular the various techniques presented at the ECB Workshop on *Using big data for forecasting and statistics*, organised in cooperation with the International Institute of Forecasters in April 2014, www.ecb.europa.eu/pub/conferences/html/20140407_workshop_on_using_big_data.en.html.

create new information/research needs, and international cooperation could add value in this context.

One caveat is to define what big data really is.[6] In general terms, one can think of extremely large data sets that are often a by-product of commercial or social activities and provide a huge amount of granular information at the level of individual transactions. This form of data is available in, or close to, real time and can be used to identify behavioural patterns or economic trends. Yet, there is no formally agreed definition that would cover all possible cases.[7] For instance, it may not be sufficient for a data set to be large to qualify as "big data" – indeed, national statistical authorities have been dealing with large data sets covering millions of records (for instance census data) for many decades without branding them as "big data". In particular, a key factor to consider is whether the data set is structured and can be handled with "traditional" statistical techniques, or if it is unstructured and requires new tools to process the information.

In practice, it is usually referred to "big data" when (i) the data set is unstructured (and often quite large), as a by-product of a non-statistical activity – in contrast to traditional data sets, produced for statistical purposes, which are, by design, clearly structured; or (ii) large volumes of records that are relatively well structured but nevertheless difficult to handle because of their size, granularity or complexity – and which could benefit from the application of big data tools (eg IT architecture, software packages, modelling techniques) to process the information more efficiently. In any case, assessing what is big data leaves room for judgment, and depends on a number of criteria such as the following "Vs":[8] **volume** (ie number of records and attributes); **velocity** (speed of data production eg tick data); and **variety** (eg structure and format of the data set). Some observers have added other "Vs" to this list, such as **veracity** (accuracy and uncertainty of big data sets that usually comprise large individual records), valence (interconnectedness of the data) and **value** (the data collected are often a by-product of an activity and can trigger a monetary reward; hence they are usually not available as a public good, due either to commercial considerations or confidentiality constraints).[9]

[6] See P Nymand-Andersen, "Big data – the hunt for timely insights and decision certainty: Central banking reflections on the use of big data for policy purposes", *IFC Working Paper*, no 14, 2015.

[7] Following the work conducted under the aegis of the United Nations (see Meeting of the Expert Group on International Statistical Classifications, "Classification of types of big data", United Nations Department of Economic and Social Affairs, ESA/STAT/AC.289/26, May 2015), big data can be classified into three types: (1) social networks (human-sourced information, such as blogs, videos, internet searches); (2) traditional business systems (process-mediated data, such as data produced in the context of commercial transactions, e-commerce, credit cards); and (3) internet of things (machine-generated data, such as data produced by pollution or traffic sensors, mobile phone locational information, and logs registered by computer systems).

[8] For these Gartner "3Vs", see D Laney, *3D data management: controlling data volume, velocity, and variety*, META Group (now Gartner), 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[9] Not all observers agree on the precise list and definitions of the Vs one should consider; for a presentation, see P Nymand-Andersen, op cit, as well as C Hammer, D Kostroch, G Quiros and STA Internal Group, "Big data: potential, challenges, and statistical implications", *IMF Staff Discussion Note*, SDN/17/06, September 2017.

Moreover, big data raises a number of challenges for central banks, especially as regards its handling and use for policymaking. The handling of big data requires significant resources (not least on the IT side) and proper arrangements for managing the information.[10] Turning to its policymaking use, big data creates opportunities but is not without risks, such as that of generating a false sense of certainty and precision. From this perspective, the apparent benefits of big data (in terms, say, of lower production costs or speed in producing information) should be balanced against the potential large economic and social costs of misguided policy decisions that might be based on inadequate statistics.

The IFC satellite meeting

In view of the various challenges, and given the strong interest expressed by the central banking community, IFC members joined forces to monitor developments and issues related to big data – such as the methodologies for its analysis, its value compared with "traditional" statistics, and the specific structure of the data sets. This was done by focusing on a few pilot projects on which central banks have been invited to cooperate so as to share specific experiences. The pilots were intended to cover four main areas: (i) internet data; (ii) administrative data; (iii) commercial data; and (iv) financial market data. A key milestone was the presentation of this initial work at this IFC Satellite meeting on big data co-organised with Bank of Indonesia in March 2017 on the occasion of the Regional Statistics Conference of the International Statistical Institute (ISI).

Opening the meeting, Yati Kurniati, Bank Indonesia, underlined the traditional importance played by data in central banks' day-to-day work. This has been reinforced by the ongoing "big data" revolution: public authorities have access to an increasing supply of information, in terms of volume, velocity and variety; in turn, this information can be used to produce new types of indicator to support policy. Yet this raises new, sometimes acute challenges, with many of them relating to the statistical production process itself – eg collecting, cleaning and storing the new data sets, as well as extracting meaningful information with adequate technologies etc.

Aurel Schubert, European Central Bank (ECB) and Vice Chair of the IFC, also acknowledged in his opening remarks the importance of these challenges. But he felt that there was a clear opportunity for policymakers to access new and complementary information sources.[11] This puts a premium on collaborative work in the central banking community to explore the synergies and benefits of using big data.

The keynote speech by Agus Sudjianto, Executive Vice President and Head of Corporate Model Risk at Wells Fargo, was an opportunity to learn from commercial banks' experiences in dealing with large data sets. Their risk management work and the related use of data has clearly expanded since the Great Financial Crisis (GFC) of 2007–09, not least because of the need to comply with more stringent regulation. In particular, the production of stress tests requires an increasing amount of information and sophisticated quantitative tools.[12] Commercial banks now amass large data

[10] For an overview of the challenges posed by using big data for official statistics more generally, see C Hammer et al, op cit.

[11] See also J Mehrhoff, "Demystifying big data in official statistics – it is not rocket science!", presentation at the Second Statistics Conference of the Central Bank of Chile, October 2017.

[12] See Basel Committee on Banking Supervision, "Making supervisory stress tests more macroprudential: Considering liquidity and solvency interactions and systemic risk", Working Paper, no 29, November 2015.

volumes at a highly disaggregated level, for instance to measure changes in their portfolios over time, capture the drivers of their risk profiles with sufficient sensitivity, and validate their modelling tools. To do that, they have to deal with big data sets and use "big data algorithms", eg new machine learning techniques. And a key aspect was the very considerable amount of work required in terms of data cleaning and data reconciliation.

Another consequence was that financial institutions need highly skilled staff, especially graduates in mathematics, finance and statistics. As highlighted in the address by Vijay Nair, former ISI President and University of Michigan, this testifies to the emergence of data science as a key mode of scientific discovery, on a par with experimental, theoretical, and computational analysis.

The meeting was fruitful in offering various perspectives on these issues. The first session discussed the importance of big data for central banks in general. The second and third sessions focused on specific big data sets, ie internet data sets and financial, administrative and commercial data sets, respectively. The last session reviewed the implications of big data sources in central bank communication. The event ended with a panel discussion on big data governance, the related challenges and the implications in terms of resources, especially in HR and IT.

2. Big data for central banks

The first session, chaired by Robert Kirchner of the Deutsche Bundesbank, discussed the importance of big data for central banks. The initial presentation, by Eurostat, highlighted the need for central bank statisticians to carefully consider the characteristics of big data sets. In particular, the quality of an apparently large non-random sample of data is determined not by its absolute number of records but by its relative size compared to the population of interest.[13] Moreover, the statistical representativeness of a sample depends on its possible coverage bias – that is, the extent to which the structure of the sample is representative of the structure of the entire population studied given the methodology used. From this perspective, a key problem with large, non-random big data sets is their organic nature: the data are often self-reported or is the by-product of social activities (eg financial transactions, internet clicks). As a result, the coverage bias of these samples is unknown and can be significant. For instance, social media sources will yield information whose quality depends on differences in the usage intensity of these social medias; that is, the less one uses them the less one is represented. Such big data sets (eg internet-based) may thus be much lower in statistical quality than (comparatively smaller) probabilistic samples that are designed to be representative of the population of interest. In other words, using (very) large amounts of data is no guarantee of accuracy, and there is a key misperception of the intrinsic value of big data from this perspective.

Perhaps more fundamentally, it is important to distinguish between "data" and "information"; the latter depends on processing the former.[14] Traditional official

[13] See X Meng, "A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it)", in X Lin, C Genest, D Banks, G Molenberghs, D Scott and J-L Wang (eds), *Past, present, and future of statistical science*, Chapman and Hall, 2014, pp 537–62.

[14] See R Groves, "Designed data and organic data", in the Director's Blog of the US Census Bureau, 2011, www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.

statistics can be described as "designed data", since they are collected for a specified statistical purpose through adequate statistical processes such as surveys and censuses; the collection of these designed data sets is almost by definition organised in order to extract meaningful information, a key difference to "organic" big data. With the increasing supply of organic data relative to that of designed data, the risk of confusion between "data" and "information" is clearly on the rise. The challenge is thus to complement, instead of replace, the collection of designed data with organic big data so as to maximise information-to-data ratios.

The second presentation, by the consulting firm BearingPoint, also underlined the importance of big data for central banks with a focus on the actual challenges posed by data collection and analytics.[15] One key lesson was that central banks are more and more developing their own data platforms and reporting frameworks, in particular to handle regulatory data collection. Such data are increasingly important, given central banks' greater post-crisis involvement in financial stability issues and /or supervisory functions. Thus, decisions on data have become of strategic importance to central banks. Many are now rethinking their IT system architecture and data governance framework to access big data sources or use big data techniques. Some have appointed chief data officers and have put in place coherent data strategies.

Yet the view from central banks was that existing processes should be enhanced to effectively handle the new and increasing amounts of supervisory, statistical and financial markets information. A critical factor from this perspective is the limitations faced in terms of human and IT resources. It is also essential to set up a new information value chain to replace traditional template-driven data collections; to access granular, micro-level data from various different sources and at reasonable cost; to develop process automation; and to facilitate the link between micro data points and aggregated macro indicators so as to "go beyond the aggregates". But this requires a greater harmonisation of data sets and the integration of various IT systems, both among reporters and between supervisory authorities and supervised entities. One way forward is to introduce automated secure data transfer mechanisms based on standards such as XBRL[16] and SMDX[17] and to explore innovations such as blockchain and distributed ledger technology (DLT).[18]

The third presentation, by the Business Reporting Advisory Group, highlighted the large number of big data pools generated by various regulations. This information can be very rich for central banks, with potential applications for eg financial supervision, inflation assessment, the monitoring of specific market participants.

[15] See also E Glass, "Survey analysis – Big data in central banks", Central Banking Focus Report, 2016, www.centralbanking.com/central-banking/content-hub/2474744/big-data-in-central-banks-focus-report.

[16] eXtensible Business Reporting Language.

[17] Statistical Data and Metadata eXchange; see IFC, "Central banks' use of the SDMX standard", March 2016.

[18] See Committee on Payments and Market Infrastructures, "Distributed ledger technology in payment, clearing and settlement – An analytical framework", February 2017, especially p 2. DLT refers to the processes and related technologies that enable nodes in a network (that is, a computer participating in the operation of a DLT arrangement) to securely propose, validate and record state changes (or updates) to a synchronised ledger that is distributed across the network's nodes. Financial transactions are recorded in a "block" (or batch of transactions), which is added to a chain comprising a history of transactions and is known as a "blockchain".

While individually these data pools may not fall into the category of big data analysis, together they constitute a "data lake" that can be usefully exploited via big data algorithms. To this end, it is important to develop adequate data standards, identifiers and dictionaries. Fortunately, central banks (like other financial institutions) are increasingly relying on a number of data standards (in particular SDMX, XBRL as well as ISO 20022 for payments standards) and identifiers when collecting and processing the data, and this should be strongly supported.

3. Internet data sets

The second session, chaired by Gülbin Şahinbeyoğlu, Central Bank of the Republic of Turkey, reviewed central banks' use of internet data sets. This web-based information comprises a variety of indicators, such as search queries, the recording of clicks on specific pages, the display of commercial information (see the third session of the seminar) and text posted online (see the fourth session).

The first presentation by the Central Bank of Armenia described its recent experience in collecting various kinds of web-based data for different purposes. A key message was that the increasing amount of information generated by web and electronic devices can be effectively used to complement official statistics. One example is the collection of prices posted online by supermarkets on a daily basis, which allows advanced estimates of consumer inflation to be computed. A second example is the collection of housing prices displayed by real estate agencies on their websites, which has helped to create a housing price index (there is no such index based on traditional statistics in Armenia). In addition, the fact that one can easily capture the various housing characteristics related to these web announcements has facilitated the application of hedonic price methodologies. A third example is the collection of job announcements posted by employment agencies on the web, which has led to the computing of a leading indicator of business activity. Yet these experiences highlighted a number of challenges for the central bank. One was the need to use new techniques (eg web-scraping tools) and methodologies. Another was the difficulty of automatically and easily accessing the data, as well as collecting the information consistently over time (a particular issue when collecting the prices of goods that are kept identical on the web only for a short period). Lastly, experience points to the limited quality of the data collected, especially when announcement prices captured on the web differ from actual transaction prices (an issue of particular importance for house prices). Despite these challenges, the central bank was actively seeking to expand its big data work to exploit administrative and social media sources.

These findings were echoed by Sveriges Riksbank, which also uses internet data sets for its inflation forecasts, although on a more limited scale. The central bank scrapes from the internet sales price information as regularly posted by grocery retailers with an online presence. However, this collection focuses on the specific component of the price index covering the consumption of fruit and vegetables. The reason is the high volatility of this price component and the related difficulties in forecasting it. Overall, the approach followed appears to be a useful way of enhancing short-term forecasts of inflation patterns. Moreover, the data collection process has proved to be robust and scalable, and requires little in the way of maintenance costs.

A presentation by the Central Bank of Aruba focused on using internet search queries to forecast tourism receipts, something of particular importance for this island economy. The approach tackles three key issues. One is the lag involved in producing the official tourism statistics, which can be "nowcasted" using online data sources. The second advantage is the possibility of capturing unsuspected patterns in the data: instead of inferring statistical relationships, as with "traditional" statistical modelling, big data algorithms such as machine learning models allow a wide range of effects to be incorporated (for instance, seasonal patterns, non-linearities, lagged effects) without the need to make ex ante assumptions. Third, these new techniques appear efficient in terms of predictive capacity, and can be implemented easily and in an automated way using standard packages developed by the industry. Yet one drawback is the limited interpretability of such relationships derived from "black box" calculations.

The final presentation, by the Bundesbank, described the use of web data to capture depositors' expectations. This work highlighted three important features of internet data sets for policymakers. One is that they can be used as a proxy when no available data exist: in that case, information on internet queries related to the term "*deposit insurance*" proved to be a valuable proxy for depositor concerns about funds held in banks. Second, internet-based information can be usefully complemented with other, more traditional data sources – in this case, the Bundesbank's statistics on interest rates and the balances of overnight banking deposits. Third, the (near) real time availability of web data allows trends in deposits to be anticipated and the risk of bank runs to be analysed, depending on the causality patterns found in the relevant variables. From this perspective, the Bundesbank is set to expand its approach to other big data sets, eg social media.

4. Financial, administrative and commercial data sets

The third session, chaired by Aurel Schubert, considered the broader range of data sets that qualify as "big data", in particular those comprising financial, administrative and commercial records. Certainly, the distinction between these data sets and web-based data sets can be artificial, since a significant part of the information collected on the web can be the result of commercial activities (eg the examples presented in the second session).

One example was the presentation by Bank Indonesia on collecting price lists to construct property market indicators. Previously, this information used to be posted by property agents or in newspaper ads. But most of it has now moved to the web, and in particular to a small number of property online websites – the three largest covered by Bank Indonesia's data collection represent a market share of above 50%. One advantage for the Bank is to complement its traditional statistics on property prices, which are derived from surveys available only quarterly and for a limited number of large cities. Moreover, the data are relatively easy to access, representing a significant amount of information – more than 2 million data points per month. However, significant data quality issues have arisen. One reason is the questionable accuracy of the information that individuals input to the web, which can be prone to errors and typos. Another limitation is that the information is not well structured, particularly as to property locations. Moreover, values are duplicated, since the same property can be advertised in various places. Furthermore, information accuracy

varies over time, since previous advertisements can be re-posted after the expiration time, or because the announcement can continue to be posted even after the property is sold. To address these challenges, the central bank had to set up a clear and precise information process distinguishing four main phases: data acquisition (ie downloading the information); data preparation (ie detection of the characteristics such as the location of the property, removal of duplicated advertisements); data processing (ie removal of outliers and extraction of indices); and data validation.

Another presentation, by the Central Bank of Chile, was based on the collection of a fiscal database covering more than 10 years of (anonymised) tax records for roughly 25 million taxpayers. A key advantage was the richness of this data set, which allows the computation of various indicators over time and, in turn, the analysis of the factors driving business demography (eg survival rates). Yet a key challenge, apart from confidentiality constraints and the need to anonymise data, is the lack of quality control as well as the significant number of missing values. Hence cleaning the data requires significant preparatory work, for instance, to delete extreme values as well as deal with missing records.

Another example presented by the Bank of Portugal focused on credit registries, which have become the largest data sets maintained by some central banks. These data are well structured, but they qualify as "big data sets" since the information is highly granular (covering most individuals and corporations applying for a credit), contains multiple characteristics (eg on the debtor, the credit extended, the instrument used etc) and is often complex to analyse. In Europe, for instance, the AnaCredit[19] project is leading to the collection of almost 200 attributes per data point on a monthly basis (and on a daily basis for a subset). Reflecting the complexity of this information as well as its sheer importance for central bank functions, the project has triggered a full rethink of central bank information management frameworks. In particular, attention has focused on the rationalisation of data collection and management; the need to harmonise the underlying statistical concepts and ensure that consistent data can be used for multiple purposes; the set up of a single entry point for reporting and accessing the information; and the willingness to limit the associated reporting burden as possible. All in all, the project has proved instrumental in steering central banks' attention to the need to manage information in an integrated way across units.

5. Central bank communication

The fourth session, chaired by Toh Hock Chai, Central Bank of Malaysia, reviewed the implications of big data sources in central bank communication. The first presentation by the BIS recalled that that finding text similarities across a large sample of documents can be very difficult. Big data techniques, in particular text-mining technologies, can facilitate such work. One way is to build a semantic similarity database: all the words are first extracted from the textual information of interest (in the BIS study, the speeches delivered by the Federal Reserve Board members over two decades); these words are then characterised by attributes covering various dimensions in a vectoral space; and similarities between two words can be measured by the proximity of these attributes. For instance, the exercise showed that the word

[19] See also A Schubert, "AnaCredit: banking with (pretty) big data", Central Banking Focus Report, 2016.

"*forward*" appears to have close similarity with "*guidance*" and "*communicate*". Interestingly, the techniques allow these relationships to be tracked over time. For instance, and not surprisingly, the word "*systemic*" was deemed to be associated with "*macroprudential*" in the post-GFC period, but less so before 2007.

The second presentation, by the ECB, showed how such techniques could be used to assess the impact of policy communication and expectations for policy decisions. This is a relatively new topic of interest for central banks. In the past, attention focused mainly on comparing outcomes in financial markets with central bank intentions, for instance, by conducting "event studies" around the times of policy decisions. The new techniques now allow the perception of public messages by the various stakeholders to be assessed, thus providing a possible way of fine-tuning policy communication. For instance, the ECB applies computational linguistic techniques to select the words used in its statements that have the highest discriminative power and can thereby gauge the tone of its communications. Based on a global news database covering the ECB press conferences, the index allows communication phases with a "hawkish tone" to be distinguished from those "dovish" ones. The usefulness of this index can be checked by looking at its correlation with other variables (eg interest rates, to see whether actual policy changes are correctly anticipated by media reports). This experience suggests that a quantitative approach to central bank communication is possible and can provide useful insights.

The third presentation by the Bank of Italy noted that most information available on the web is textual and can therefore be exploited through ad hoc techniques. Moreover, it was particularly important to have an objective measure of the central bank's communication and of its impact on the sentiment of stakeholders. To this end, textual information is used to create a heatmap showing the usage of specific words over time. The approach can also provide insights to assess the readability and formality of central bank communication. Lastly, this allows for semantic analysis, by extracting the contextual usage of specific words and analysing similarities across documents.

6. Big data governance

The seminar ended with a panel discussion on issues related to big data governance, chaired by Katherine Hennings, Central Bank of Brazil and IFC Vice Chair. The discussions reviewed big data work in central banks and covered related resource issues, especially in HR and IT.

One view was that central banks are relatively new in exploiting big data, in contrast to the long-standing experience of national statistical offices (NSOs) in handling large and confidential data sets such as censuses and administrative records. A key reason was that central banks have traditionally been data users rather than data producers. They have been catching up rapidly, especially since the GFC, with more and more central banks being called on to collect, produce and use large data sets. Yet the lessons learned are mainly tentative, as big data sources of information are still under evaluation in most central banks.

What is unknown was whether these new data developments will lead to a change in the institutions' business models. A commonly shared view was that this would not fundamentally change what central banks do, given their role as data users

– unlike most NSOs, which are usually not obliged to use their data for policy purposes. Yet there were specific areas in which central banks' processes may have to change significantly with the advent of big data – for instance, short-term forecasting and nowcasting, IT security etc. This puts a premium on enhancing the governance of related data sets, in particular by establishing formal *Memoranda of Understanding* with the related data providers. In any case, central banks have multifaceted mandates, particularly in the post-crisis era. Dealing with the increasing supply of big data while combining old and new roles may thus prove challenging.

Another governance issue is to maximise the use of new information available to support central bank policy functions while managing the associated risks. This calls for existing data governance frameworks to be revamped. For instance, use of internet-based information raises several difficulties, as compared with the production of traditional statistics. First are the legal, financial and ethical issues posed by accessing (often private) information that is a by-product of commercial activities. Then there are the operational, legal and reputational risks entailed in dealing with transaction-level information that is potentially confidential; the responsibility for authorising such data collections, not least as regards aspects such as confidentiality protection, data ownership and privacy; the degree of statistical accuracy of these data and the level of confidence in their sources; and even the information content of data derived from self-generated activities – for instance, the information value of the number of clicks on a specific topic will vary as these clicks are influenced by the search engines and based on users' past searches. One view was that the complexity of these issues might well increase as central banks move from experimentation to the actual regular production and use of big data-based information.

From this perspective, the consensus was that cooperation (both internationally among central banks and domestically among statistical authorities) should and would indeed expand to facilitate the exploration of the big data area. This reflects the fact that there was a lot to learn from each other. Yet one view was that such cooperation may prove temporary and may well recede once big data has become a more mature area and sufficient work expertise has been developed in central banks.

Turning to resource challenges resulting from handling big data, these mainly reflect the sheer size of the data sets, their lack of structure and the often poor quality of raw data obtained from internet streaming, large administrative records or other sources. Moreover, sophisticated statistical techniques are often required to derive meaningful information from such data.

A major area is IT. The implications of big data for central banks' information systems are potentially huge. There are large IT processing costs and difficult and expensive technology choices have to be made. One risk is to spend more time and resources on cumbersome activities – cleaning the data, organising the underlying platforms etc – rather than on actually analysing and using the data. One way to address this risk is to focus on specific use cases and resist the temptation to collect various large data sets covering an excessively wide range of purposes.

Another issue was that public statisticians have a tendency to be "cloud computing-adverse", mainly because of the disclosure risks posed for confidential information. Most prefer to operate in a "secluded" data environment. This may well reduce the scope for public authorities to benefit from new big data techniques developed in the marketplace – for instance, some applications may be available only as part of a cloud-based solution. Nevertheless, it was also recognised that a number

of big data sources have a public nature, implying that central banks may have sufficient opportunities to make use of private sector solutions. And, within central banks, important organisational changes were also expected to better deal with big data, including the creation of internal centres for big data statistics, data lakes, internal clouds etc. In any case, experimentation will shed more light on these aspects.

A second key area is staff. The necessary skills may not be available in-house, especially in IT, data science and methodology as well legal expertise. Given the limited supply of graduates, central banks may well face a "war for talent". This could be a key obstacle, since skilled staff are a prerequisite if central banks are to benefit from big data opportunities as well as manage the associated risks. Moreover, the skills shortage also raises questions around compensation and staff career paths, as well as management issues – one view, for instance, was that the relatively important role played by economists in central banks' managerial positions might well be called into question by these developments.

In any case, these challenges highlighted the likelihood that significant time and effort will be needed before a regular production of big data-based information can be undertaken to support central bank statistical and analytical work on a large scale.