Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC-Bank Indonesia Satellite Seminar on *"Big Data"* at the ISI Regional Statistics Conference 2017

Bali, Indonesia, 21 March 2017

Central banks' use of and interest in "big data"[1]

Jens Mehrhoff, Eurostat

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

DEUTSCHE BUNDESBANK
EUROSYSTEM

Central banks' use of and interest in "big data"

Session 1: "Big Data for Central Banks"

Jens Mehrhoff*, currently on secondment to the European Commission

* This presentation represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank, the European Commission or its staff.

Structure of the presentation

1. Introduction
2. Identifying sources
3. Joint conceptual framework and roadmap
4. Statistical paradises and paradoxes
5. One way forward for official statistics

"*But the 'big data' that interests many companies is what we might call 'found data', the digital exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast."* Tim Harford, Financial Times, 28 March 2014.

1. Introduction

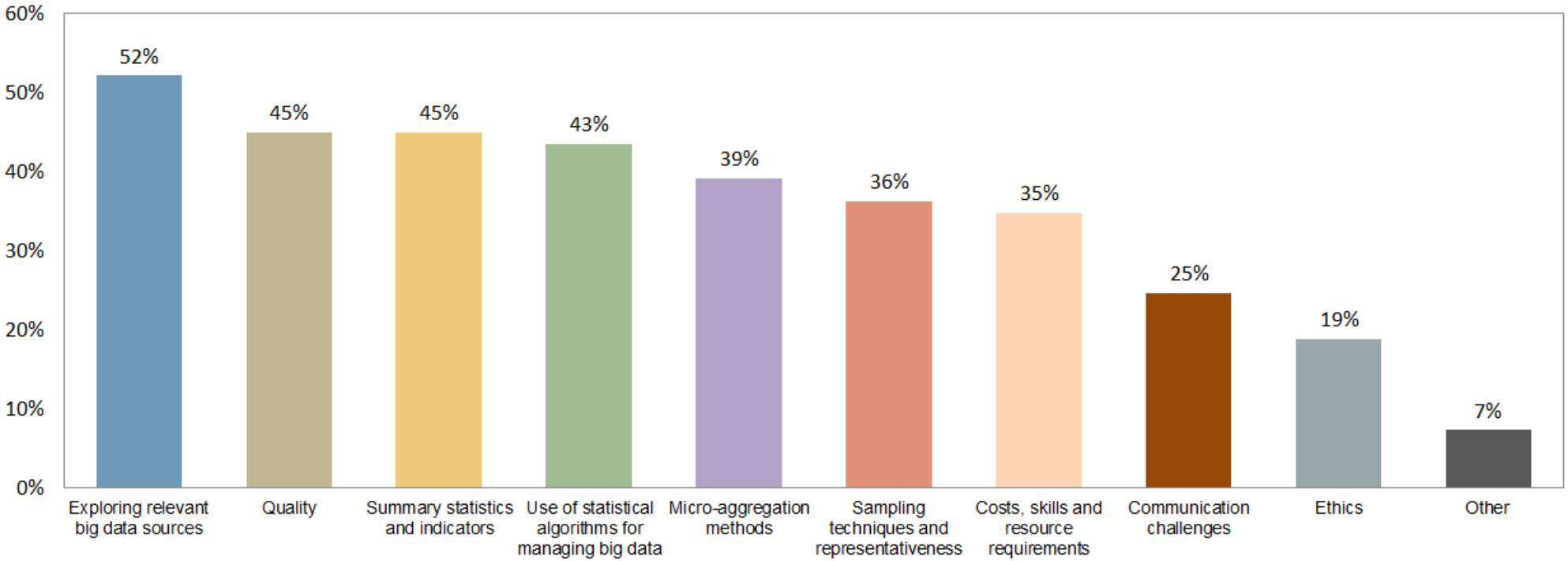
- In January 2015 the Irving Fisher Committee on Central Bank Statistics launched an **online survey on central banks' use of and interest in "big data"**.
- The aim of this survey was twofold:
  - To **take stock of central banking experience** in the use of big data; and
  - To **explore central banks' interest** in this topic with a view to defining a roadmap for further action.
- The **vast majority (69) of IFC member central banks responded**, representing a response rate of 83%.
- The **main conclusions** of the survey are the following. http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf

1. Introduction

- There is **strong interest in big data** in the central banking community, in particular at **senior policy level**.
- Central banks actual involvement in the **use of big data is currently limited**.
- Big data can be **useful for conducting central bank policies**.
- Big data are perceived as a **potentially effective tool** in supporting macroeconomic and financial stability analyses.
- Big data may also **create new information/research needs**.
- **International cooperation can add value.**
- Exploring big data is a **complex, multifaceted task**.
- **Regular production** of big data-based information **will take time, especially because of resource issues**.

1. Introduction

Which statistical topics would be of interest to you as part of the big data subject?



Note: multiple responses possible.

2. Identifying sources

- Despite the **limited current experience** in the use of “big data”, there is a **strong interest** within the central banking community to **cooperate and share experiences** on the use of big data.

- The IFC Executive decided to select **few case studies for piloting the usefulness** of “big data”, as a potentially effective tool-kit, in supporting central banks’ monetary policy, financial stability and/or banking supervision policies and invited the **IFC community to cooperate in the piloting phase**.

- These **supplementary statistics** may provide further insights contributing to **guiding central bankers’ policy actions** as well as to **assessing the subsequent impact and associated risks** of these policy decisions on the financial system and real economy.

- The way forward may be to take **small steps in developing and applying** a structural approach for piloting the use of big data, or non-official sources, for central banking purposes.

2. Identifying sources

– The following **four of group of sources and tentative pilot projects** for showcasing could be envisaged:

1. **Administrative dataset:** internal central banks and statistics offices databases and other public databases

2. **Internet dataset:** patterns and behaviour of internet activities; examples could be internet search machines, social media consumer internet purchases

3. **Commercial dataset:** relate to for instance transactional data from payments, settlements or trading systems or mobile banking and operators

4. **Financial market data:** financial market data or data relating to individual financial instruments

3. Joint conceptual framework and roadmap

- One benefit of conducting joint pilot studies would be to have a **similar overall structure and approach** in managing the work, processes and expected outcome of the pilot projects within a certain time frame.

- Despite of differences of the “big data” sources, it is important that each group of participating central banks describe in details the **characteristics of the supplier** according to a **standardised set of key information** for each of the **five statistical production processes** covering

  - **“Input”**,
  - **“Quality”**,
  - **“Production”**,
  - **“Results”** and
  - **“Assessment”**

  as part of exploring its relevance for central banking tool-kits.

3. Joint conceptual framework and roadmap

- **Input:** e.g. Who are the source provider and what type of relevant information is available? Include an example of the information content of the source.
- **Quality:** e.g. How transparent is the source on its methodology? Please describe the sample and its representativeness.
- **Production:** e.g. Please describe the production process required. What types of modelling, statistical algorithms, machine learning techniques, text mining and semantic analysis is required?
- **Results:** e.g. Which types of statistics information/indicators can be made available? Please describe how the indicators could be used for central banking purposes.
- **Assessment:** e.g. Please provide a short overview/summary of the pilot study. Can the source easily be used for statistics production purposes?

4. Statistical paradises and paradoxes

Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- "*Is an 80% non-random sample 'better' than a 5% random sample in measurable terms? 90%? 95%? 99%?*" (Wu, 2012)

- Let us consider a case where we have an **administrative record** covering f_a percent of the population, and a **simple random sample (SRS)** from the **same population** which only covers f_s percent, where $f_s \ll f_a$.

- How large should f_a / f_s be before an estimator from the **administrative record dominates** the corresponding one from the **SRS, say in terms of MSE**?

- $\bar{x}_a = \frac{1}{n_a} \sum_{i=1}^{N} x_i R_i \, , R_i = \begin{cases} 1 & \text{if } x_i \text{ is recorded,} \\ 0 & \text{otherwise.} \end{cases}$

- The **administrative record has no probabilistic mechanism** imposed by the data collector.

4. Statistical paradises and paradoxes

Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- Expressing the **exact error**, where $f_a = n_a/N$:

$$\bar{x}_a - \bar{X}_N = \frac{\mathrm{E}[xR]}{\mathrm{E}[R]} - \mathrm{E}[x] = \frac{\mathrm{Cov}[x,R]}{\mathrm{E}[R]} = \underbrace{\rho_{x,R}}_{\text{Data Quality}} \cdot \underbrace{\sigma_x}_{\text{Problem Difficulty}} \cdot \underbrace{\sqrt{\frac{1-f_a}{f_a}}}_{\text{Data Quantity}}.$$

- The **MSE** of $\bar{x}_a$ is more complicated, mostly because R_i depends on x_i:
  $\mathrm{MSE}[\bar{x}_a] = \mathrm{E}\left[\rho_{x,R}^2\right] \cdot \sigma_x^2 \cdot \left(\frac{1-f_a}{f_a}\right)$.

- For **biased estimators** resulting from a large self-selected sample, the **MSE is dominated (and bounded below) by the squared bias term**, which is **controlled by the relative sample size** f_a.

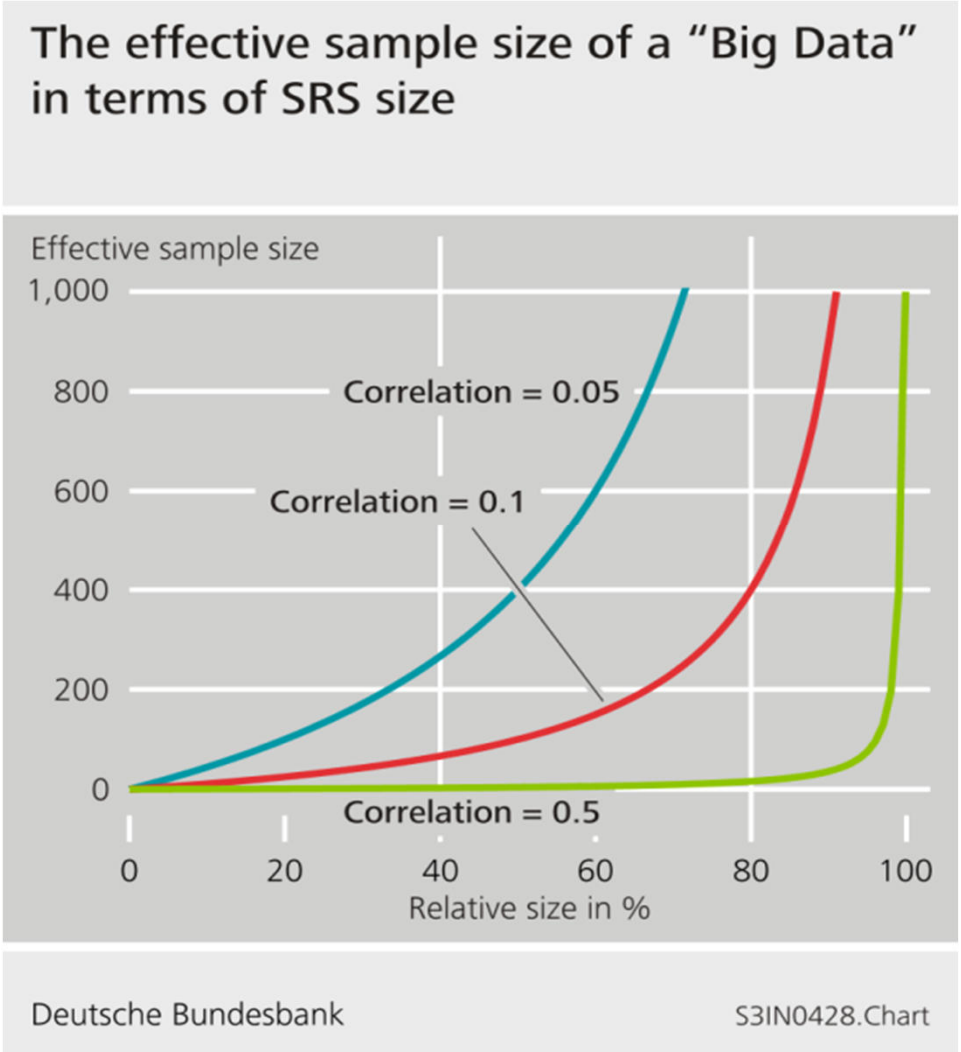
- The **non-sampling errors** can be made arbitrarily **small only when the relative size** f_a **is made arbitrarily large**, that is $f_a \to 1$; just **making the absolute size** n_a **large will not do the trick**.

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- A **key message** here is that, as far as statistical inference goes, what makes a **“big data”** set big is typically **not its absolute size**, but its **relative size to its population**.
- Therefore, the question **which data set one should trust more** is unanswerable without knowing N.
- But the general message is the same: when dealing with self-reported data sets, **do not be fooled by their apparent large sizes** or by common wisdom from studying probabilistic samples.
- This reconfirms the **power of probabilistic sampling** and reminds us of the **danger in blindly trusting that “big data”** must give us better answers.
- **Lesson 1:** What matters **most is the quality**, not the quantity.

4. Statistical paradises and paradoxes

Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.



- Imagine that we are given a **SRS** with $n_s = 400$.
- If $\rho_{x,R} = 0.05$ and our intended **population is the USA**, then $N \approx 320{,}000{,}000$, and hence we will need $f_a = 50\%$ or $n_a \approx 160{,}000{,}000$ to place more trust in $\bar{x}_a$ than in $\bar{x}_s$.
- If $\rho_{x,R} = 0.1$, we will need $f_a = 80\%$ or $n_a \approx 256{,}000{,}000$ to dominate $n_s = 400$.
- If $\rho_{x,R} = 0.5$, we will need over 99% of the population to beat a SRS with $n_s = 400$.

4. Statistical paradises and paradoxes
Meng, X.L. (2014), in: Past, Present, and Future of Statistical Science.

- However, the **availability of both small random sample(s) and large non-random sample(s)** opens up many possibilities. The following (non-random) sample of questions touch on this:

  - Given **partial knowledge of the collection/response mechanism** for a (large) biased sample, what is the **optimal way to create an intentionally biased sub-sampling scheme** to counter-balance the original bias so the resulting sub-sample is guaranteed to be **less biased** than the original biased sample in terms of the sample mean, or other estimators, or predictive power?

  - What should be the **key considerations when combining small random samples with large non-random samples**, and what are the sensible **"corner-cutting" guidelines when facing resource constraints**?

- **Lesson 2:** Do not ignore seemingly tiny probabilistic datasets when combining data sources.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- What's the **difference between "data" and "information"**?
- We're entering a world where **data will be the cheapest commodity around**, simply because the society has created **systems that automatically track transactions** of all sorts.
- Collectively, the society is **assembling data on massive amounts** of its behaviours.
- Indeed, if you think of these **processes as an ecosystem**, it is **self-measuring in increasingly broad scope**.
- Indeed, we might **label these data as "organic"**, a now-natural feature of this ecosystem.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- **Information is produced from data by uses.** Data streams have no meaning until they are used.
- The user finds meaning in data by **bringing questions to the data** and **finding their answers in the data**.
- An old quip notes that **a thousand monkeys at typewriters** will eventually produce the **complete works of Shakespeare**.
- The **monkeys produce "data" with every keystroke**. Only we, as **"users", identify the Shakespearian content**.
- **Data without a user** are merely the jumbled-together **shadows of a past reality**.

5. One way forward for official statistics

Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- **What's this got to do with official statistics?** For decades, **official statistics has created "designed data"** in contrast to "organic data."
- The questions we ask of businesses and households **create data with a pre-specified purpose**, with a use in mind.
- Indeed, designed data through surveys and censuses are **often created by the users**.
- This means that the **ratio of information to data (for those uses) is very high**, relative to much organic data.
- **Direct estimates are made from each data item** – no need to search for a Shakespearian sonnet within the masses of data.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- What has changed is that the **volume of organic data produced now swamps the volume of designed data**. The **risk of confusing data with information** has grown exponentially.

- We must **collectively figure out the role of organic data** in extracting useful information about the society.

- The **challenge is to discover how to combine designed data with organic data**, to produce resources with the most efficient information-to-data ratio.

- This means we **need to learn how surveys and censuses can be designed to incorporate transaction data** continuously produced by the internet and other systems in useful ways.

- Combining data sources to **produce new information not contained in any single source is the future**. The **biggest payoff will lie in new combinations** of designed data and organic data, not in one type alone.

5. One way forward for official statistics
Groves, R.M. (2012), in: Director's Blog – Census Bureau.

- To continue the monkey-typewriter metaphor, the **internet and other computer systems are like typewriters that have an unknown set of keys disabled**.
- Some keys are missing **but we don't know which ones are missing**. They're **not capturing all behaviours in the society**, just some.
- The Shakespearian library may or may not be result of the monkeys pounding on the keys. In contrast to the beauty of the bard's words, **we may only find pedestrian jingles and conclude that's as good as it gets**.
- We **need designed data for the missing keys**; then we **need to piece them together** with the masses of organic data from the present keys.
- **The combination of designed data with organic data is the ticket to the future.**

Contact

JENS MEHRHOFF

European Commission
Directorate-General Eurostat
Price statistics. Purchasing power parities. Housing statistics

BECH A2/038
5, Rue Alphonse Weicker
L-2721 Luxembourg
+352 4301-31405
Jens.MEHRHOFF@ec.europa.eu