BANCA D'ITALIA
EUROSISTEMA

Harnessing Big Data & Machine Learning Technologies for Central Banks

Closing Remarks by the Deputy Governor of the Bank of Italy

Luigi Federico Signorini

Rome, 27 March 2018

Ladies and gentlemen,

This two-day workshop on 'Harnessing Big Data & Machine Learning Technologies for Central Banks' is now drawing to a close. Let me begin by thanking all the speakers, discussants and participants who have contributed to the success of this initiative. I have the pleasure of sharing with you some thoughts on the main issues.

The papers presented during this two-day workshop have left no doubt as to the actual and potential value of big data and their importance for economic analysis. The present wave of data warehousing and business analytics confirms this, with big data poised to deliver topline research and statistical applications in a cost-effective way.

Huge amounts of data can provide substantial insights for the private and public sectors, enabling them to transform these data into new products and services for customers and citizens. Big data are already changing some aspects of business in the financial industry with regard to banks, hedge funds, broker-dealers, and other firms.

Central banks and supervisors are also interested.

The widespread adoption of digital technology has increased the number and depth of information sources of interest to economic analysis and financial stability, two of the core concerns of financial authorities. We are not only voracious data users, we are also big data producers. We collect terabytes of granular data in the fields of banking supervision, oversight of financial markets and payment systems.

Central banks can and should harvest the benefits of new technologies. We should select the best technology and exploit its power in turning ideas into actual applications and improved statistical and computing efficiency.

The potential is huge, but there are, as usual, pros and cons. There are important trade-offs that must be kept in mind. I see two main issues, or rather groups of issues.

One is technical, and some papers in this workshop have pointed to it. We need to keep analysing and experimenting in order to be able to assess the real information value of big data bases.

‘N = All’, as in the popular definition by Mayer-Schönberger and Cukier[1], may sometimes be illusory. Sometimes big data turn out not to be representative of the whole population. For example, Groves (2016)[2] argues that big data often contain few variables and extracting their value requires linkages to other data; furthermore, they lack representativeness. Social media or any other internet-based sources of big data also usually lack full coverage of the population of interest.

The availability of massive amounts of data increases the chances of finding an exceptionally good fit in sample when one estimates any model, but a very poor performance out of sample. Big data are harvested from different technical sources and change with consumers’ preferences. Therefore, the parameters estimated using these data can be subject to different forms of instability. This, however, is not unique to big data.

Large volumes, variety, and the short-lived nature of data available on web pages make it difficult to accumulate and preserve historical big data; their preservation is hampered by changes in physical storage devices, IT platforms and software. This point deserves attention. Poor preservation or the loss of data could jeopardise the accountability of decisions and research reproducibility.

The other key issue with big data is the protection of integrity, confidentiality and privacy of data. Exploitation of data must be respectful of this principle, in legal and ethical terms.

The effective application of the relevant laws and international coordination are essential in this field.

Let me now conclude by touching on some principles that we follow in the Bank of Italy as large producers, consumers and providers of data.

We adopt an integrated approach to collecting and processing banking and financial data. As I have said on other occasions, we have always gathered information from reporting agents following a single protocol and have then used this information for

[1] V. Mayer-Schönberger, and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt, 2013.

[2] R. M. Groves, ‘Nonresponse Rates and Nonresponse Bias in Household Surveys’, *Public Opinion Quarterly*, Vol. 70(5): 646-675.

multiple purposes.[3] This ensures consistency across datasets and intermediaries and reduces the need for burdensome ex-post data reconciliation. We have followed this approach since the 1980s.

The Bank of Italy has developed a shared corporate statistical data dictionary and a corporate statistical data warehouse to ensure the uniqueness of reporting models. Both were planned with a view to managing the different areas of statistical and supervisory information as parts of a single system.

Although the supervision and central banking functions require their own analytical approaches, their decision-making processes draw on the aforementioned data dictionary and warehouse. The same holds for other uses of the data, such as for research and official statistics.

Some of the data we manage are confidential and/or sensitive. The protection of the integrity and confidentiality of data is therefore a key concern. Access to data is suitably restricted within the Bank, and data management follows clear rules on procedures and responsibility.

Central banks must also be ready to cope with the increasing demand by the public for access to granular data. As a data producer, the Bank of Italy has always strived to make its statistics available to the widest possible audience. We already share some data with researchers and other institutions, with appropriate and adequate protection of confidentiality, of course. Within the Eurosystem's Household Finance and Consumption Survey (HFCS), Italian data are available on request alongside those of other euro-area countries. For firms' data, which are much more difficult to anonymize, the Bank has developed a remote access system called BIRD (**B**ank of **I**taly **R**emote Access to **D**ata), which enables users to perform their analyses online, thereby preserving data confidentiality. Particularly confidential data might require on-site access. The Bank of Italy has started to design a Research Data Centre which will have a suitable taxonomy of confidentiality areas.

[3] Cfr. comment by L.F. Signorini to the paper "*Ways to improve the use of banking statistics by policy-makers: what is reasonable, what is feasible and what the SSM and the banking union are calling for*" by Fernando Restoy. Seventh ECB conference on statistics 'Towards the banking union –opportunities and challenges for statistics' ECB (2015).

Other European countries have already started to work on this (the Centre d'accès sécurisé aux données in France, the Administrative Data Research Network in the UK and the Deutsche Bundesbank Research Data and Service Centre). In addition, we need to find solutions to allow us to link microdata belonging to different institutions without compromising individual confidentiality.

The explosion of granular data provides many new research opportunities. For all the issues I have highlighted, big data have big potential. In the past two days we have just scratched the surface. More research is surely needed.

I wish to thank you all very much for your participation. Special thanks to all the speakers, discussants, chairs and participants for making the sessions lively and thought-provoking. Finally, I would also like to take this opportunity to thank all those involved in preparing this event.

Designed by
the Printing and Publishing Division of the Bank of Italy