BANCA D'ITALIA
EUROSISTEMA

Bank of Italy and Bank for International Settlements Workshop on 'Computing Platforms for Big Data and Machine Learning'

Opening Remarks by the Deputy Governor of the Bank of Italy

Luigi Federico Signorini

Rome, 15 January 2019

Ladies and Gentlemen,

It is a pleasure to open this workshop and welcome all the participants.

Everyone is aware that modern societies are generating an unprecedented amount of digital information. The latest ***Data never sleeps*** annual report, published last June, claims that mankind now generates more than 2.5 exabytes of data every day. The increase is bound to continue: by 2020, an estimated 1.7 MB of data will be created every second for every person on earth.

The enormous wealth of information that has thus become available has great potential value if effectively captured and aggregated. However, it also poses fundamental challenges that are easily overlooked. Let me open this workshop by briefly mentioning three that I think deserve more attention than they usually get.

First, big data are worth little if they are used in a methodologically unsound way. Such data are typically representative only of selective strata of the population; increasing the volume of data will not improve the accuracy of estimates if the estimation procedure does not correct for this. Social media data for example, like many other internet-based sources of big data, are usually based on a biased sample of the population; now-fashionable media-based indicators (of confidence, say, or political trends) are often employed with scarcely a thought for this fact. Relatedly, the careless use of massive amounts of rough data will sometimes result in an exceptionally good in-sample fit, especially with non-parametric or machine learning approaches, but a poor performance out of sample (this is sometimes called the 'overfitting plague'). Reference texts should place much more emphasis on these facts in order to provide a more rigorous guide to practitioners and, ideally, educate the public at large as the ultimate consumer of processed information.

Second, while the world produces a relentlessly growing mountain of data, it has not yet settled on a reliable set of criteria for making (some of) it effectively accessible in the more or less distant future. Large volumes

of internet data are casually stored, overwritten and discarded without any consideration for what the next generations will need or want to know about the world today: and this not just in the general pursuit of knowledge, but also for more immediately practical purposes, such as clarifying legal rights. Long-term conservation of data is also hampered by changes in physical storage devices, IT platforms and software. The theory and practice of building archives in the big-data era is in its infancy; it needs to grow up fast.

Third, the emergence of big data has raised the importance of data integrity, confidentiality and privacy to a level never seen before. Protection of personal data is central to our societies; one could say that it helps to define them, to the extent that it is inextricably linked to the protection of individual rights more generally. The sheer amount of personal data now available, and the growing ease of connecting individual information across databases, have far-reaching implications for fairness and freedom; they have thus spurred much digital-privacy regulation. The connections with the previous point are, I think, rather obvious (data integrity, for instance, is central to both); so are the implied trade-offs. For instance, the right to be forgotten, increasingly (and, I think, rightly) granted to individuals, conflicts with society's desire to keep records. No wonder that rules are sometimes inconsistent across different jurisdictions, data domains or legal purposes. A systematic way of weighing conflicting principles against one another should be found. Ideally, one would like to have global standards; international cooperation is therefore to be encouraged as far as possible. However, there are limits to this, at least as long as different societies protect individual freedom to a different extent. For the time being, we have to live with this limitation.

The challenges of big data require continuous organisational and technological innovation in the institutions that want to use them effectively and fairly. At the Bank of Italy we have taken several steps in the past few years to enhance our ability to collect, store and exploit huge amounts of data, using state-of-the-art hardware and software. Other institutions are surely doing the same. This workshop is a good opportunity to share experiences.

Preserving the confidentiality of individual data is one specific challenge that many institutions represented here will often face. The issue is not just that the social value of statistical analysis cannot override the legal obligation to refrain from infringing on personal privacy, but also that the quality of the data itself is often premised on confidentiality being

guaranteed: without it, proprietary data would be impossible to collect, and much sensitive data, such as data on personal wealth, would be next to useless. We are cooperating with other institutions to find ways to reap the value of merging data from different sources, and to make them available – insofar as possible – to independent researchers, while avoiding confidentiality breaks.

Let me conclude by thanking once again all the speakers and participants for joining us today. Special thanks go to those who have contributed to organising the workshop. It brings together highly professional data scientists, IT architects and business specialists from central banks, providing views of excellent quality and a broad variety of perspectives. I am sure that you can count on having an interesting and productive day.

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