Data Science in Central Banking

Welcome address by Piero Cipollone
Deputy Governor of the Bank of Italy
Banca d'Italia and the Irving Fischer Committee (BIS)
Rome 19 October 2021

Ladies and Gentlemen,

I am delighted to open this virtual conference on Data Science in Central Banking, jointly organized by Banca d'Italia and the Irving Fischer Committee of the Bank for International Settlements. I would like to welcome all the participants joining us today from some sixty countries.

1. Introduction

For the last two years, we have lived through a dramatic period as the COVID-19 pandemic has swept the globe. Never before as in these dark days has Data Science in the form of big data and machine learning (ML) algorithms proved so helpful in the war against Coronavirus. The lack of biological knowledge on this virus has spurred the data science community to step up and contribute to the fight against COVID-19. Scientists from many different disciplines and public organizations have acknowledged the importance of data analytics by open sourcing the virus genome and other datasets in the hope of a swift data-driven solution.

In this four-day conference we will endeavour to share among institutions and academia the newest and most interesting applications of Data Science and machine learning to sharpen our analytical capacity to cope with new and rapidly evolving economic equilibria.

2. The role of Data Science in central banking activities

Data Science is an interdisciplinary field that combines computer science, statistics and business domain knowledge aimed at generating insights from noisy and often unstructured data. It integrates mathematics with scientific methods and computing platforms. Still a young field, it has quickly developed over the last few years. Its main driver is the astounding volume of data stored by private companies and public authorities, which can now be treated more easily with new algorithms to extract the information hidden among them.

Nonetheless, at the Bank of Italy, Data Science is not completely new. We do have a sound history of basing our decisions on data. In 2016, we established a multidisciplinary team to address the potential benefits and hidden risks of embracing the technological challenges of artificial intelligence (AI) and machine learning (ML) fuelled by the advances in big data, which continue to evolve at an incredible speed.

A gargantuan amount of digital activity is occurring every day. In fact, data are constantly generated by our internet activities. This explosion stems from the aggregate actions of about 4.7 billion active internet users worldwide[1]. These numbers are projected to rise even further in the coming years. According to SeedScientific[2], at the dawn of 2020 the total amount of data in the world was around 44 zettabytes, tantamount to $44 \cdot 10^{21}$ bytes, which is the number of cells we could count in more than 1,400 human beings.

This astonishing amount of data can give us a better understanding of the state of the economy at both the micro and the macro level, provided that we able to extract the signal from the noise.

[1] As of January 2021.

[2] See https://seedscientific.com/how-much-data-is-created-every-day/

Banca d'Italia has constantly striven to be at the cutting edge in developing software and hardware platforms, enabling big data analytics[3] for statistical and economic applications.

The rise of Data Science started at the beginning of 2010 when high-quality models for image recognition were created, computational power achieved sufficient growth, and people in many scientific areas realized the full potential of such an approach.

Data Science glues together machine learning and data processing. The former is a collection of tools, which allows us to learn from the given data and to extract patterns and interactions between series and values. The latter describes the possible set of actions in relation to the data itself: collection, manipulation, preparation, and visualization.

It is important to flag a few differences between Data Science and the classical Econometrics we study at University. Unlike Econometrics, which concentrates on solving non-linearity bias problems in a typical linear framework, Data Science is about improving our ability to work with non-linear relationships in the system. Another difference is that Econometrics concentrates on methods such as robustness, while machine learning algorithms became popular for their outstanding predictive performance[4].

3. Data Science at the Bank of Italy

Banca d'Italia is organizing, along with the Federal Reserve Board, the Sveriges Riksbank, the University of Pennsylvania and the Imperial College, a series of webinars on 'Applied Machine Learning, Economics, and Data Science' (AMLEDS)[5]. The aim of these webinars is to foster the integration of data science tools into economics and policy-related issues and to promote closer cooperation on issues related to data science, big data and machine learning techniques applied to policy questions. Such cooperation has intensified during the pandemic, leading to many important initiatives such as a series of conferences on non-traditional data organized by the Federal Reserve Board and the Bank of Italy (for example, this year's conference is going to be hosted by the Bank of Canada in November). This new world has, of course, set many challenges for central banks and public institutions: on one level, central banks have had to devise specific organizational structures to have a consistent and more efficient approach to the big data and machine learning tools used by each institution. At the same time, they have had to increase their data production, leveraging on non-traditional data to get more timely and high-frequency indicators of economic activity, which are important during unprecedented shocks like the COVID-19 pandemic.

4. The challenges and risks of Data Science

We need to exert extreme caution when employing these new tools.

Let me briefly go over some of them.

First, big/web data might lose their statistical relevance when they are employed in an unsound way. Such data typically entail selection bias because of the features of the population; increasing

[3] See, for example, 'Big data processing: Is there a framework suitable for economists and statisticians?', 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 2804-2811, doi: 10.1109/BigData.2017.8258247. 'Weaving Enterprise Knowledge Graphs: The Case of Company Ownership Graphs.'

[4] I am thinking of Extreme Gradient Boosting (XGBoost), which owes its wide popularity to its dominance over several machine learning algorithms rather than to its mathematical properties.

[5] AMLEDS are a series of webinars open to all those in the world who are interested in applied machine learning, big data, and natural language processing for economics and in how these techniques and data science can be applied to social science.

the sample size will not shrink the sampling error if the estimation algorithm does not correct for this kind of distortion.

Second, the availability of a huge amount of data raises the importance of its integrity, confidentiality and privacy. Personal and company data protection is central to our societies.

The sheer amount of personal data now available, and the growing ease with which individual information can be merged across databases, have far-reaching implications for privacy, competition and freedom. Indeed, this has prompted the development of regulations concerning the treatment of digital data (think of the GDPR or the CCPA in California).

Rules on data management often differ across jurisdictions and data domains. Therefore, international cooperation is to be encouraged as far as possible.

Technologies enabling the processing of personal data are already available. In 2019, Google made available its differential privacy library, which allows sensitive data to be processed privately.

To reach these welfare-improving goals, further investment from both the public and the private sector are required. Close cooperation between the private sector, which typically owns most of the new data, and the public sector, which uses (or would like to use) such data for policy reasons and for the common good, will also be essential.

This is why central banks have always made great efforts when it comes to collecting and analysing data. Throughout its history, Banca d'Italia has drawn extensively on data published by the National Statistical Institute and other national and international agencies. It has also been an active producer of statistics, not only on banking, financial and fiscal variables, but also on firms and households.

Banca d'Italia has been collecting micro-level statistical information on companies since the early 1950s. These data are now enriched with non-traditional sources such as social media, blogs, newspapers, and private company datasets[6].

5. Conclusions

Let me conclude my talk by thanking, once again, all the speakers and participants for joining us today, even in this virtual fashion. We hope to welcome you here in Rome in person in the near future. Special thanks go to those who have helped to organize this workshop, which brings together leading economists, statisticians, artificial intelligence and machine-learning specialists, data scientists from about fifty-five central banks and fifteen participants from Universities and government agencies. This guarantees a broad variety of perspectives and a lively discussion.

I am sure that you are going to have a very interesting and productive workshop.

[6] Such as the real estate website, immobiliare.it, and the mortgage website, mutuionline.