The experience of 10 years of data in Central Banking - from gathering real-time data and big data to challenges like storage or skills[1]

di Piero Cipollone

I am delighted to open this Session on "**Central bank as a pool of real-time data: the 'Whys' and the 'Hows'**" jointly organized by Lietuvos bankas and Bank for International Settlements on the occasion of the 100th Anniversary of Lietuvos bankas. I would like to thank the organizers for the kind invitation.

1. Introduction

I would like to share the Bank of Italy's experience in central banking, using real-time data – also known as big data, Nontraditional data or Alternative data – for policy purposes with all the challenges involved.

As we all know, we live in a data-empowered era where we can plan a trip and Google can estimate our travel time by recommending the best route based on both current and past traffic data or where Netflix can suggest us movies or shows we might like based on its data from people with similar preferences. Our lives have become not only more data-driven but also more and more data-producers, generating more data than ever before. This combined with greater and greater computing power and ad-hoc technology enables private companies and government institutions to use that new data for different purposes efficiently.

2. The role of big data in central banking activities

Central banks have always used market data and macroeconomic data based on surveys to make projections about economic activity, inflation and unemployment, to then guide their monetary policy decisions. The Bank of Italy has always given a very high priority to the collection and use of granular data for economic analysis. For example, it started conducting a well-structured Survey on household income and wealth (SHIW) in the early sixties; at a time, when only in the US a similar effort was done. Bank of

[1] Intervento alla conferenza internazionale "Future of Central Banking" Vilnius, Lithuania, 29 settembre 2022

Italy was also one of the pioneers in Europe in creating a Central Credit Register, the information archive on household and firms' debts at a loan by loan granular level. Besides the collection of data and production of statistics on banks and the broader financial sector for which it is legally responsible, the Bank runs a large number of surveys and collects granular data from firms, households and the public administration.

We do have a sound history of basing our decisions on data. It is not a surprise therefore that the Bank of Italy was early on very eager to look at the potential off by the huge increase in the availability of information coming from the ICT revolution with the Web at its center, i.e. the phenomenon of digitalization. Some activity in big data started in the early 2000's; for example, we were using data from Google Trends as soon as Google made them available.[2] However, it was overall a very scattered activity, mainly performed at the initiative of individual researchers. The real step forward was made in 2016 when this activity was elevated to a strategic priority of the Bank. We set up a multidisciplinary team to address the potential benefits and hidden risks of embracing the technological challenges of artificial intelligence (AI), machine learning (ML) and natural language processing (NLP) fueled by the advances in big data, which continue to evolve at an incredible speed.

It is then that we started to collect in a systematic manner data from a variety of non-traditional sources, such as social media, newspapers, and credit card transactions. These new sources of data have changed the data landscape and enriched economic analysis with more disaggregated and more timely economic information.

It is important to stress that we see these sources of data as complementary to traditional sources of structured data, based on surveys, which remain of foremost importance since they allow us to collect high-quality and reliable data within a clear methodological and theoretical framework built to analyze specific phenomena. In any case, the role of the non-traditional and unstructured data has been growing over time and certainly, the pandemic crisis was a big push for us to use it even more, given the exceptional circumstances and the impossibility to run surveys to gather data by national institutions. It was great that we had already some experience and some alternative data in our hands to be able to run some analyses without using survey data. At this stage, I would say that big data and ML techniques have already transitioned from a supporting role to a symbiotic

[2] See D'Amuri and Marcucci (2017), "The predictive power of Google searches in forecasting US unemployment", International Journal of Forecasting, Volume 33, Issue 4, https://doi.org/10.1016/j.ijforecast.2017.03.004, which was published as a working paper version in 2009

relationship with more traditional statistical analysis. Still, we do recognize that we are only at the beginning of the journey and that the potential remains huge and largely unexploited.

3. Data Science at the Bank of Italy

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. Albeit 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 ML algorithms to extract the information hidden among them.

In the age of Big Data, economic analysis should be addressed with different tools. Among the data we are going to use, I would mention the following: 1) government data, such as electronic invoices and Tax records; 2) corporate data, such as data from Google or other private companies such as retailers; 3) unstructured data, such as textual data from social media, newspapers, job searching platforms, people and goods mobility, FinTech apps, etc.

The necessity to exploit nontraditional data with special algorithms is unavoidable for addressing the present economic conundrums. For example, Raj Chetty[3] shows how big data gathered from private companies in the US can be fruitfully used to understand and solve some of the most important social and economic problems in a particular period of stress like the recent pandemic. Taking advantage of big data can ultimately improve macroeconomics policymaking: i) by answering new questions and producing new, accurate and more granular indicators; ii) by offering a painstaking and detailed description of the economic scenario through innovative data sources; iii) slashing the time lags in statistics production, therefore, contributing to a timelier nowcasts/forecasts of existing indicators. Again, Raj Chetty shows that with real-time data like card payment transaction data securely merged with other individual data, we can design a new system of real-time national accounts that can be useful for diagnosing issues in

[3] A few days ago Raj Chetty from Harvard gave a talk titled "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data" at AMLEDS (Applied Machine Learning, Economics, and Data Science) webinar (https://sites.google.com/view/amleds/home) presenting the latest results of his work on the economic tracker https://opportunityinsights.org/wp- content/uploads/2020/05/tracker_paper.pdf

the economy. Big data can open new pathways for macro policy and macro modelling. We can fine-tune our policies based on the current state of the economy and evaluate the observed impacts of those policies in real-time.

The spread and usage of Bbg data, right now, cannot replace official statistics. The two data sources should be seen as complementary. Both outputs should be compared to ensure the robustness of new indicators vis-à-vis existing time series that can serve as a benchmark to validate those new indicators.

At the Bank of Italy, we use big data and machine learning to support our routine economic and statistical analysis and to carry out research projects. So far, this activity has focused on three main areas: 1) indicators for now-casting and forecasting, 2) expectations of households and firms and 3) sentiment and confidence indicators. The main purpose has been to enrich the information set and create new real-time indicators and models to improve our analysis of the economy and the ability to anticipate future trends.

Some prominent examples are the following.

Angelico et al.[4] (2022) use social media such as Twitter to measure inflation expectations for Italy. Using NLP techniques, the authors isolate the signal from the noise due to advertisements on tweets related to price(s) and inflation. They show that Twitter-based indicators of inflation expectations are not only significantly correlated with the traditional survey-based or market-based measures, but for the survey-based measures they are also informative in sample and they have predictive power even out of sample. We also use Twitter to create sentiment indicators towards banks and analyze their relationship with deposit growth. Indeed, Accornero and Moscatelli (2018) show that a Twitter-based indicator of sentiment toward banks improves the predictions of a standard benchmark model of depositor discipline based on financial data.[5] Astuti et al. (2022) use Twitter to analyze the most actively discussed topics related to the COVID-19 pandemic and build a real-time indicator for the average

[4] Angelico C., J. Marcucci, M. Miccoli, e F. Quarta (2022), "Can We Measure Inflation Expectations Using Twitter?", Journal of Econometrics, Volume 228, Issue 2, 259-277, https://doi.org/10.1016/j.jeconom.2021.12.008

[5] Accornero M. and Moscatelli M., "Listening to the Buzz: Social Media Sentiment and Retail Depositors' Trust" (2018). Bank of Italy Temi di Discussione (Working Paper) No. 1165, http://dx.doi.org/10.2139/ ssrn.3160570

sentiment of the users.[6]

Aprigliano et al. (2022) analyze around 2 million articles from four main Italian newspapers and compute a text-based sentiment indicator that closely mimics the confidence indicators for households and businesses from our NSO.[7] They also compute a daily Economic Policy Uncertainty (EPU) index both for the general economy and for specific sectors or topics. They use these indicators to nowcast the economic activity in Italy showing that using a Bayesian Model Averaging technique their indicators greatly reduce the uncertainty surrounding the short-term predictions of the main macroeconomic aggregates, especially during recessions. They also employ these indices in a weekly GDP tracker, achieving sizeable gains in forecasting accuracy.

We also use credit and debit card transaction records to build models and indicators for the short-term forecast of economic activity. Ardizzi et al. (2019) analyze the reaction of consumer expenditure in Italy (measured using daily data on debit card payments at the POS) to daily EPU, built using textual data from news and the social network Twitter. The authors show that an increase in EPU temporarily reduces debit card payments and raises the ratio of ATM withdrawals to POS payments, signaling an increase in the preference for cash.

We make large use of online ads on the real estate from the largest platform in Italy (the website www.immobiliare.it, the equivalent of Zoopla or Zillow in the UK and US), to follow developments and prices in the housing market. Loberto et al. (2022) show the potential of this new database[8] to study the real estate market, while Guglielminetti et al. (2021) show the impact of COVID-19 on housing demand for Italian households.[9]

---

[6] Astuti, V., Crispino M., Langiulli M., and Marcucci, J., (2022), "Textual analysis of a Twitter corpus during the Covid-19 pandemics", Bank of Italy Occasional Paper No. 692, http://dx.doi.org/10.2139/ ssrn.4154474.

[7] Aprigliano V., Emiliozzi S., Guaitoli G., Luciani A., Marcucci J, Monteforte L., (2022), "The power of text-based indicators in forecasting Italian economic activity", International Journal of Forecasting, https://doi.org/10.1016/j.ijforecast.2022.02.006

[8] Loberto, M., Luciani, A., and Pangallo, M., «What do online listings tell us about the housing market?» (2022). Forthcoming in the International Journal of Central Banking (https://dx.doi.org/10.2139/ ssrn.3176962).

[9] 5 Gulielminetti E., Loberto M., Zevi G., and Zizza R. (2021). "Living on my own: the impact of the Covid-19 pandemic on housing preferences", Occasional Paper 627, Bank of Italy (https://dx.doi.org/10.2139/ ssrn.3891671).

Benetton et al. (2022) give new empirical evidence on house prices and climate change adaptation concentrating on the city of Venice.[10]

We also use ML algorithms for statistical production and for data quality management. Recently, La Serra and Svezia (2022) received the 2022 IFC Young Statistician award for implementing a Machine Learning algorithm for anomaly detection in insurance assets granular reporting.[11] La Ganga et al. (2022) apply supervised ML to spot quality degradation on Non-Performing Loans data.[12]

4. The challenges and risks of Data Science

Using big data and nontraditional data we need to cope with many challenges.

First, big data imply significant costs. The management of big data requires investments: you need special IT infrastructures to ingest, store and process nontraditional data together with traditional one that has become too big to efficiently handle in a classical data warehouse. There are organizational costs because you have to put together different skills and hire or create those skills internally.

Second, there is always an issue of representativeness and some form of selection bias because big data are the result of data collection for a purpose usually unrelated to a specific research question. Very often big data are related to an unknown population where some strata are heavily represented, while others are not represented at all. That is why, it is always important to validate externally the results obtained from big data, using what is known from surveys and other traditional sources. At this stage, we should still gravitate toward new measures that are not too different in the aggregate from the official measures, but that have other dimensions that are not found in official statistics, such as being timelier, higher frequency, having greater geographic and sectoral details, in two words, more granularity. However, this is something that is going to change in the future, as it is very likely that what we now consider non-traditional data will play an increasingly relevant role into official statistics.

---

[10] Benetton M., Emiliozzi S., Guglielminetti E., Loberto M. and Mistretta A, (2022), "Do House Prices
Reflect Climate Change Adaptation? Evidence from the City on the Water", Bank of Italy, mimeo.
[11] La Serra V. and Svezia E., (2022), "Statistical matching for anomaly detection in insurance assets granular reporting", IFC Conference - Basel 25-26 august 2022
(https://www.bis.org/ifc/events/ifc_11thconf/ ifc_11thconf_young_statistician.pdf).
[12] La Ganga B., Cimbali P., De Leonardis M., Fiume A, Meoli L., and Orlando M., (2022), "A decision-making rule to detect insuffi data quality: an application of statistical learning techniques to the non-performing loans banking data" Bank of Italy Occasional paper 666
(https://dx.doi.org/10.2139/ssrn.4032815).

Big data often include sensitive data like personal data or privately owned data. Data collection is a big issue as access to data is not always easy, in particular, because most central banks are not endowed with a broad mandate to collect data, beyond that of the financial sector for which we have legal responsibility. This issue is of key importance for the future: there is not a general framework governing the access of public authorities and especially statistical authorities to privately held data. The result of this regulatory gap is fragmentation of approaches, and a lack of standards and clarity about the rights and obligations of private and public counterparties. It is something that needs to be discussed.

Concerning the issues of privacy and confidentiality, we have very well-established methods and processes derived from decades of experience in dealing with personal or firm-level data, which are related to a large extent to our supervisory and financial stability tasks. The Bank of Italy runs a very sound Data Risk Analysis on each new dataset to safeguard the security and privacy of its sensitive data to avoid data breaches and intellectual property theft. For example, we ran a gender wage inequality internal research relying on widely agreed cryptographic protocols.

Finally, a big challenge concerns the preservation of data over time. While we keep producing a huge amount of data, we have not yet agreed on a shared set of criteria for making these data accessible in the future. Technology-neutral standards must be established to ensure data availability for future generations. Cooperation and, whenever possible, coordination between statistical agencies, governments, academia and private companies is equally crucial, especially across different jurisdictions.

5. International cooperation and collaboration

Dealing with big data will require not only further investment from both the public and the private sector but also tighter cooperation between the private sector, which typically owns most of the new nontraditional data, and the public sector, which uses such data for policy reasons and the common good. It will be also important to cooperate with other public authorities, which have many administrative data that can be helpful to measure the state of the economy, and with academia and other central banks to set up common standards.

We are currently involved in many projects and activities with other central banks and academia to foster collaboration and create a network of researchers interested in data

science, ML and big data for economic policy. Since 2019 we have organized a series of conferences with the Federal Reserve Board and the Bank of Canada on "Non-traditional Data, Machine Learning and Natural Language Processing in Macroeconomics" where both researchers from academia and from CBs and NSOs can present their work on these topics. At these conferences, we have both the academic and scientific programs where research papers are presented and discussed, in addition to a closed-door day, where central bankers and researchers from NSOs, government agencies and international organizations talk about different issues related to nontraditional data and ML applications. This year's conference is jointly organized with FRB, Bank of Canada and Sveriges Riksbank and it will be held in Stockholm on October 3-5.[13] We are also collaborating with the Bank of England and with the BIS through the Irving Fisher Committee to organize workshops and conferences on the topic.[14]

To create a network and foster scientific collaboration, jointly with the FRB, the Sveriges Riksbank, the University of Pennsylvania and the Imperial College of London we are also organizing a series of webinars on "Applied Machine Learning, Economics and Data Science" (AMLEDS). From September to June we host a monthly webinar, where researchers working on data science and ML applied to economics and finance present their work.[15] More than 3,000 researchers from academia, CBs, and NSOs around the globe and from different disciplines are registered. Finally, yet importantly, we have organized a special issue for the Journal of Econometrics on "Machine Learning for Economic Policy" jointly with the European Central Bank, the Federal Reserve Board, the Bank of England, the Bank of Canada, and the King's College London. The special issue will gather a selection of papers, which use machine learning and big data for policy purposes.

6. Conclusions

[13] The program is available at the link https://www.riksbank.se/en-gb/press-and-published/conferences/2022/conference-on-non-traditional-data-machine-learning-and-natural-language-processing-in-macroeconomics/

[14] See the IFC and Bank of Italy Workshop on "Data Science in Central Banking" which was held in Rome in October 2021 (https://www.bis.org/ifc/events/211019_ifc_bdi.htm) and in Basel in February 2022 (https://www.bis.org/ifc/events/220214_ifc.htm).

[15] So far, we have invited among others Serena Ng (Columbia), Francis X. Diebold (University of Pennsylvania), Matthew Gentzkow (Stanford), Nick Bloom (Stanford), Héléne Rey (LBS), Bryan Kelly (Yale), Jianqing Fan (Princeton). On September 23 2022, we hosted Raj Chetty (Harvard) talking about "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data".

The Bank of Italy is at the forefront in this area and it is firmly committed to researching these issues and to cooperating with other national and international institutions with the aim of better serving our societies.

Let me conclude my talk by thanking, once again, all the organizers for having me here and all participants for joining us today.