Closing remarks for the 2020 Banca d'Italia and Federal Reserve Board Joint Conference on "Nontraditional Data & Statistical Learning with Applications to Macroeconomics"

November 12, 2020

This has been a wonderful conference with a tremendous number of interesting papers. I've very much enjoyed the presentations and the discussions. As I understand it, there were 700 registrants this year, and attendance has been quite high. This conference is clearly filling an important gap.

I'd like to start by thanking the organizing committee from the Bank of Italy and the Federal Reserve Board—Giuseppe Bruno, Juri Marcucci, Ricardo Correa, and Chris Kurz—as well as other individuals at both banks that made this conference possible. Of course, the success of this conference importantly reflects the active engagement of people from our global community of researchers from central banks, statistical agencies, and academia, so thank you to all of our presenters and session chairs.

Work at this conference broadly fell into three buckets: Projects with big or nontraditional data, applications of machine learning methods, and applications of NLP methods. Not surprisingly, many presentations discussed Covid-19, and the role of these data and methods in the pandemic. I think there is little doubt that we've all benefitted from the availability of timely, high-frequency data these past 9 months.

I thought the presenters did an excellent job over the past two days employing tools in a manner that made their results understandable and interpretable. That is commendable, because some of tools can yield results that are difficult to explain to policymakers or the public.

The range of questions that were considered is impressive, and the answers were equally provocative. I know I've got a list of ideas and topics that I want to follow up on as a result of the presentations, so I am personally thankful for all the work encompassed by this conference.

In the time I have left, I thought it might be useful to step back and discuss three general challenges we face, particularly when moving from individual research projects to things that can be used on a recurring basis either in official statistics or in central bank analysis. Much of what I will say will mirror themes you heard from the panel a few moments ago. Each of these challenges will benefit from a cross-disciplinary set of clever minds working together to develop creative solutions.

Challenge 1: The first challenge is to **close the gap between traditional and nontraditional data**. When developing a nontraditional measure that has a counterpart in official statistics, there exists a tension between having your new measure be too similar to the official measure (as that limits its value added) and having it be too different from the official measure (as that suggests potential problems with the methodology or with the data). We can all think of examples of big data that, despite their scale, suffer from selection bias. In our economic measurement work with nontraditional data at the Board, we've long grappled with this tension, and I recently heard Matthew Shapiro from the University of Michigan highlight it as well.

In practice, I think there has been a tendency to gravitate toward new measures that aren't too different from the official measures but that have another dimension that is not found in official statistics—such as being timelier, higher frequency, having greater geographic or industrial detail, and so on. By

matching an official measure along key dimensions—often a fairly aggregate summary at a monthly or quarterly frequency—the nontraditional measure effectively "inherits" the official measure's credibility. Of course, just because the aggregate measures match, it does not mean the more granular information in the nontraditional measure is accurate, but by satisfying that aggregate constraint it does boost one's confidence at least somewhat.

There are other ways to resolve the tension between being too similar vs. being too different.

- One common approach is to not compete and instead blend nontraditional data into official indicators. Indeed, our colleagues in statistical agencies around the world have active agendas to do just that. For example, some groups are using scanner data or web-scraped data to help measure consumer prices. Others are using point-of-sale information to improve measures of consumer spending.
- Another common way to resolve the tension is to use official indicators to reweight nontraditional data. Indeed, that is something that has been done with web-scraped prices, and it is something that we do at the Board with the employment measures we derive from payroll processor data and consumer spending measures we derive from aggregated card transactions. Importantly, that reweighting requires high-quality, low-frequency measures, which is something only statistical agencies can do in a convincing fashion.
- Yet another way to resolve the tension is to go even bigger. If we can expand nontraditional data to be close to a population measure, then that eliminates selection bias and, provided the methodology is defensible, should engender confidence in the resulting measures. For example, suppose (counterfactually) that all workers were paid through a third-party payroll provider, then acquiring data from *all* payroll providers would eliminate the risk of selection bias that comes from working with one provider's data. In practice, of course, similar data will tend to be highly correlated, and so it is hard to justify the cost of getting a measure for the population if it involves bilateral negotiations with lots of parties and bespoke programs to make the data comparable.

There is a lot of interesting work going on related to this challenge of closing the gap between traditional and nontraditional data, including work by many at this conference, but we've only scratched the surface.

Challenge 2: The second challenge is **ensuring the long-term viability of work with nontraditional data**. Establishing agreements with providers of nontraditional data can be time consuming, complicated, and costly. I fully agree with Stephen's comment during the panel about investing in relationships—that's been a key component of our work at the Fed. But even once an arrangement is successfully established, there are a multitude of risks, including hold-up problems in contract negotiations, providers refusing to renew, and risks of disruptions in the receipt of data for technology or other reasons. For some economic measures, the nontraditional or big data we need are often quite concentrated, so diversification is rarely a means to mitigate these risks. Because the data often cover a large share of the economy, disruptions are more acute than when a survey struggles with response rates.

As researchers, you understand the costs and risks associated with using nontraditional data and new techniques. From an institutional perspective, nontraditional data arrangements require investments of money and time for the data to be used productively, and those investments are largely sunk if the arrangement falls through. Consequently, if an institution depends on a particular data source for a

statistical release or for recurring analysis, then some contingency planning may be appropriate. I suspect contingency planning for data availability has not been high on most of our to-do lists.

Hold-up risks and the risks of other disruptions may be greater for public-private arrangements, so one way to avoid this challenge is to focus on government sources of administrative or nontraditional data, where possible. In the U.S., a great example of leveraging administrative data comes from the Census Bureau's Business Formation Statistics release, which provides, at a weekly frequency, an early indication of business formations based on applications for employer identification numbers that are needed for tax purposes.

Of course, sourcing data from the government often will not be an option. This is where I think there is room for some creative thinking to harness private data for the public good while still protecting privacy and confidentiality. Perhaps someone at this conference will figure out a way to rethink the fundamental value proposition for potential data providers in order to get their buy-in for economic measurement. Or perhaps someone will devise an institutional arrangement inside or outside of governments that facilitates the creation of public statistical products. Solving this challenge could be transformative for the economic measurement community and those that depend on timely, granular information. In that vein, I was intrigued by the "Development Data Partnership" that Marco mentioned during this afternoon's panel, as it has a lot of features that I think are important.

Challenge 3: The last challenge relates back to Senior Deputy Governor Franco's opening remarks, and that is the importance of **engaging on issues of privacy and confidentiality**. In the U.S., the Census Bureau has been a leader in this area with its work to apply formal privacy methods to the decennial population census. As formal privacy methods are further developed, and as understanding of these methods grows and diffuses, it is easy to imagine a future where it has an effect on a lot of nontraditional data work. I won't pretend to be an expert on this topic, but for those of you that are not familiar with formal privacy methods, they effectively make explicit the trade-off between privacy and the accuracy of statistics derived from a database. As more statistics are generated, they have to become less accurate to protect privacy. If you know ahead of time what statistics you want to generate from a database, then the privacy protecting algorithms can be tuned to allocate more of the scarce privacy budget to those statistics. All of us—data producers and data users—may need to be a part of the conversations to ensure we develop data access and use methods that are practical and that help us get the most possible value out of data will still protecting privacy.

Conclusion

The progress we make on these three broad challenges will have implications for many of the interesting nontraditional data efforts and the applications of ML and NLP techniques discussed at this conference. I can't think of a group better poised to make progress on these and other questions, and I look forward to continued collaboration among those represented here. Hopefully future conferences can be held in person.

Thank you.