Big Data or the renewal of the economic analysis

Summary:

- The granularity and multimensionality of Big Data offer several advantages to economists, from identifying economic trends as they occur (“nowcasting”) to testing agents’ behavior theories or creating a set of tools to manipulate and analyze these data.
- Economists face three challenges: the accessibility to these new data, the capacity to replicate them, and the development of technical competencies to handle them.
- The reinforced integration of courses in computer sciences and advanced statistics appears as a priority of public policy, as well as the development of state-funded research laboratories oriented towards Big Data
- In the same path, a narrower collaboration between the companies owning the massive datasets and the economists working on them would be a big advance in the economics research discipline.

"Big Data". Under this "media hype" term lies the emergence of unprecedentedly new rich datasets during the last decade, from digital processes to social media exchanges to the Internet of things (systems, sensors, mobile devices, etc.).

- To put this phenomenon into numbers, the capacity to store information has grown exponentially (indeed, the world's technological per-capita capacity to store information has doubled every 40 months[1] since the 1980's) and as of 2012, 2.5 exabytes (2.5×10^{18}) of data are generated every day[2].

- Unprecedented data already open incommensurable opportunities in domains ranging from genomics (drastic reduction of the human genome sequencing time) to social sciences (through social network data) to business analytics and predictive algorithms (Google's search results, Apple's auto-complete function, online advertising, insurers' risk scores, credit card companies' underwriting activities, etc. are a handful of examples).

- A concrete example taken by Linar Einav and Jonathan Levin[3] (Stanford University) shows the dramatic rise of new datasets in the Internet era: in the retail stores' sector for instance, before the emergence of the Internet, data collection was often limited to daily sales – by product in the best of cases; nowadays, thanks to scanner data, online inventories and retails, all the consumer's track and behavior can be traced, from purchase histories and locations to search queries to advertising exposure.

- Similar examples can be given about inventories, online transactions or public services data (tax filings, social insurance programs, etc.), as well as employment data (Internet giants such as LinkedIn or Monster.com can provide job titles, what skills people have, what employers they worked at, occupational level detail, etc. which provides a brand new level of granularity compared to the classical survey data).

But behind the debates on what this "new asset class" (as the World Economic Forum[4] calls it) will bring to the economy, we can also wonder how these massive new datasets can improve the way we measure, track and describe the economic activity, but also how the development of new advanced data analytics methods and predictive modeling tools that have emerged in statistics and computer science may prove useful in economics.

- Indeed, as Taylor, Shroeder and Meyer[5] point out, the place of the economic science at the intersection between academic and applied knowledge used for business purpose as well as

[1] http://science.sciencemag.org/content/332/6025/60

[2] https://www.ibm.com/big-data/us/en/

[3] http://web.stanford.edu/~jdlevin/Papers/BigData.pdf

[4] https://www.weforum.org/reports/personal-data-emergence-new-asset-class

[5] http://journals.sagepub.com/doi/pdf/10.1177/2053951714536877

its strong body of theory and methodology make it an interesting candidate in using bigger and richer datasets while keeping the reliability and representativeness that characterize this discipline.

- As Schroeder[6] describes it, big data in economics corresponds to a step change in the scale and scope of the sources of materials (and tools for manipulating these sources) available in relation to a given object of interest; this differs from the practical definition in businesses where the "volume, variety and velocity" of data can help constitute an advantage on the competitors.

Big Data in economics

Big Data in economics could be associated to the dimensions of:

1) "multidimensionality" (in terms of the number of variables per observation, the number of observations, or both)
2) "granularity" (Big Data series are often described as micro-level detailed data relating to human behavior).

8 advantages for the economic research and policymaking

1) **Better track and forecast economic activity**. Indeed, general and local governments collect vast amounts of microlevel administrative data, in areas such as tax collection, social programs, education or demographics among others.

2) **Extend the possibilities of panel studies**. The statistical power of these data can be shown in the influence gained by the research articles which use them, such as Piketty and Saez[7] (2003) whose analysis of Internal Revenue Service (IRS) data to shed light on income distribution has raised many political debates on economic inequalities.

3) **A higher level of frequency and granularity than traditional survey data.** Massive data for tracking private sector economic activity – even in real time (such as the MIT's Billion Prices Project[8] which gathers prices from hundreds of online retail websites to give an accurate proxy of inflation, or Master Card's SpendingPulse[9] which tracks consumer spending through credit cards payments) also constitute a powerful tool to track economic activity with a higher level of frequency and granularity than traditional survey data.

[6] http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199661992.001.0001/acprof-9780199661992-chapter-11

[7] http://piketty.pse.ens.fr/fichiers/public/PikettySaez2003.pdf

[8] http://www.thebillionpricesproject.com/

[9] https://www.mastercardadvisors.com/en-us/solutions/spendingpulse.html

4) **Proxies of economic indicators**. Indirect measures such as search queries or social media posts can also be used as accurate proxies of economic indicators such as employment or consumer confidence (see the example of Choi and Varian[10] on the use of Google trends to "predict the present", suggesting that Google queries for some specific products closely mirrors the demand for that product). The availability of data in "real-time" can thus offer an advantage in terms of "nowcasting", or identifying economic trends as they are occurring.

5) **A larger size of data which would contribute to a significant improvement of measurements**. The progressive availability of large-scale administrative data sets and new private-sector data should allow better measurements of economic effects and outcomes, thanks to more granular and comprehensive data, especially when it comes to understanding individual actions (what MIT's Brynjolffson[11] calls "nano-data"); the large size of the datasets becoming available also resolves the statistical problem of limited observations and makes analysis more powerful and potentially more accurate.

6) **A better perception of the effects of economic policies and events.** These new data could also encourage economists to pose new sorts of questions and enable novel research designs, as areas ranging from labor market dynamics (Choi and Varian[12]), to the effect of early education on earnings (Chetty et al., see below), stock market dynamics (Moat et al.[13]) and the workings of online markets (Einav et al.[14]). The possibility to combine various datasets broadens the array of research, as shown for instance by the study of Chetty, Friedman and Rockoff[15] (2011) which combines administrative data on 2.5 million New York City schoolchildren with their earnings as adults twenty years later to show the "value added" of having a good teacher; in this case, the high level of granularity in the data makes it possible to link test score data and subsequent tax records for a large number of students in such a way that aggregate data or a small random sample would not be able to do. Several aspects of human behaviors, such as social connections (through data on social networks) or geolocations could also become much easier to observe and analyze; the example of Pew Research Center's Scott Keeter[16], who raises the idea of using Big Data derived from social media as a supplement or even as a substitute for, government statistical data gathered using traditional survey-based methods, is a proof of this idea.

[10] http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/full

[11] http://digital.mit.edu/bigdata/agenda/slides/Brynjolfsson%20Big%20Data%20MIT%20CDB%202012-12-12.pdf

[12] http://onlinelibrary.wiley.com/doi/10.1111/j.1475-4932.2012.00809.x/full

[13] https://www.nature.com/articles/srep01801

[14] http://siepr.stanford.edu/sites/default/files/publications/10-033_Paper_Einav_10.pdf

[15] http://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.104.9.2633

[16] http://www.pewresearch.org/2012/05/24/survey-research-its-new-frontiers-and-democracy/

7) **Allow for "natural experiments".** For instance, moving from weekly data to much higher frequency (up to minute-by-minute data), or to data on individual consumers and units can allow to detect very specific details or micro-level variations that would be hard to isolate and exploit with more aggregated data. The study by Einav, Farronato and Levin[17] (2016), which analyzes online pricing and sales strategies, is a sharp example of this advantage of benefitting from granular data to obtain rich information on the individuals being studied and to explore a variety of consequences from a given experiment, for instance, substitution to different items in the event of a price change. These assumptions become interesting when we apply them to the case of the companies, and especially online platforms for whom it has become easier and more cost effective to experiment when they have more customized and granular pricing strategies and easier and cheaper automated methods to capture the results of an experiment and (if successful) implement it.

8) New opportunities might also come with the **new statistical and machine-learning techniques**[18] which can help develop stronger predictive models, especially in empirical microeconomics. The study by Einav, Jenkins and Levin[19] (2012) is one example of how to use predictive modelling with big data to incorporate heterogeneity into econometric models and analyses; in this study, predictive modelling techniques allow for the construction of "credit-risk scores" that help researchers understand consumer borrowing behavior and how lenders should set loan prices and credit limits for different segments of borrowers as stratified by their default risk. Embracing heterogeneity through big data and improved statistical techniques could even become feasible and common in many other sectors, as detailed data could help embrace not only "average effects" but estimating mappings from measurable heterogeneity into treatment effects and optimal policies; the example of the grocery company Safeway[20], which offers customized individual-specific discounts as a function of individual price elasticities, shows the progressive ability of companies to go beyond simple elasticities in their price studies and develop algorithms that estimate elasticity and optimal prices that are customized to each type of customers; same for governments in their policy creation, with the ability to find the optimal policies depending on the users (health policies that depend on healthcare environment and the physician and patient characteristics, education policies tailored on the grade, the school, the teacher, or the student mix, etc.).

[17] https://web.stanford.edu/~leinav/pubs/AR2016.pdf

[18] The readers interested by these new techniques can consult Hal Varian's paper "Big Data: New Tricks for Econometrics"(http://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.3), in which he describes thoroughly new tools to analyze and manipulate Big Data, such as new data manipulation tools and software, new variable selection methodologies (as there are more potential predictors), as well as new ways to model complex relationships (through machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, etc.).

[19] http://web.stanford.edu/~leinav/pubs/ECMA2012.pdf

[20] https://retailleader.com/pulling-more-meaning-big-data

Some challenges and some risks

However, even though these massive datasets and statistical techniques open many new opportunities, **several challenges** also lie ahead for the economists.

1) **Data accessibility**: an important part of the new data researchers work on belongs to private companies (who store data from their customers), and the benefits of learning from outsiders do not always outweigh the costs of data disclosure for the companies that accept to share data with researchers.

2) **Unstructured nature of the data**, which often presents an econometric challenge in terms of untangling the various dependencies in the data.

3) **The necessary development of new capacities by the economists** – especially some of the new tools of the computer scientists (SQL, R) as the standard computational resources are often outreached, and the standard machine-learning algorithms – to be able to combine the conceptual framework of economics with the ability to actually implement ideas quickly and efficiently on large-scale data; the famous "data scientists" all companies seem to look at, and whose job is to analyze data to look for empirical patterns, are exactly at the crossroads of these computer sciences and econometrics patterns. Extracting and summarizing different variables, and exploring relationships between them will become an increased part in the economists' work and require new competencies in computer science and data management.

4) As this article has described it, we can wonder that the emergence of big data will change sharply the landscape of economic policy and economic research. However, it will not substitute to economic theory; as Sascha Becker[21] puts it, the usual practice of economic forecasting (theory – simulation – calibration – predictions) will not be changed in that "you will need theory to understand the mechanisms or even to suggest what you might hope to find in the first place". In fact, even though big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, it never tells us which correlations are meaningful; also the magnitude of big data can create some purely "bogus" correlations between series that do not have anything in common. In the same vein, big data does not substitute to research designs and asking the right questions; indeed, no problem can be solved by crunching data alone, and there is always a need to understand the issue we want to analyze.

5) The ability of big data and the associated new statistical and machine-leaning techniques to reduce anything to a single number is only an appearance of exactitude and **does not replace a thorough scientific analysis**.

[21] http://journals.sagepub.com/doi/full/10.1177/2053951714536877

The overreliance on big data can even lead to misleading effects, as collections of big data often merge data that were collected in different ways and different purposes; this risk is very often observed on data collected from web search queries, as shown by the example of Google Flu Trends, for which the responsibility of data collection in its failure has been pointed by Harvard's statistician Kaiser Fung[22]; the overreliance of web data has even led to what Marcus and Davis[23] call a risk of "echo-chamber effect" as materialized by Google Translate's reliance on Wikipedia articles' translation … and vice-versa. Another danger stems from the ever-increasing difficulty to replicate the data and programs of economic studies when data sets become ever-larger, as pointed out by Project Syndicate's Barry Eichengreen[24], who calls for a deeper stress put on historical analysis of economic phenomena, rather than the development of increasingly sophisticated statistical methods.

Recommendations

As this article has described it, the advantages of using Big Data for economic analysis are numerous. In terms of recommendations for public policies and education, the "gold mine" of Big Data fits completely into the exponential development of the NICTs in the daily life and represents an additional argument for the development of the courses of computer sciences and programming, especially in the university degrees in economy and sociology; the recent integration of a "Big Data" module to the CFA® exam is the perfect illustration of this phenomenon[25]. The development of state-funded laboratories concentrated on Big Data could also represent a solution to the lack of representativeness this discipline suffers to the researchers.

On the same path, a narrower collaboration between researchers and the companies owning these massive data could benefit all the actors and would allow on the one side the companies to benefit from external viewpoints and decision supports, and one the other side the researchers de benefit from massive data for the development of new models and for testing new theories.

Sacha TENEBAUM

[22] https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data

[23] https://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html

[24] https://www.theguardian.com/business/economics-blog/2013/may/17/economic-big-data-rogoff-reinhart

[25] https://www.bloomberg.com/news/articles/2017-05-09/cfa-exam-to-include-big-data-artificial-intelligence-as-topics