

Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?

Linnet Taylor¹, Ralph Schroeder² and Eric Meyer²

Big Data & Society

July–December 2014: 1–10

© The Author(s) 2014

DOI: 10.1177/2053951714536877

bds.sagepub.com



Abstract

Although the terminology of Big Data has so far gained little traction in economics, the availability of unprecedentedly rich datasets and the need for new approaches – both epistemological and computational – to deal with them is an emerging issue for the discipline. Using interviews conducted with a cross-section of economists, this paper examines perspectives on Big Data across the discipline, the new types of data being used by researchers on economic issues, and the range of responses to this opportunity amongst economists. First, we outline the areas in which it is being used, including the prediction and ‘nowcasting’ of economic trends; mapping and predicting influence in the context of marketing; and acting as a cheaper or more accurate substitute for existing types of data such as censuses or labour market data. We then analyse the broader current and potential contributions of Big Data to economics, such as the ways in which econometric methodology is being used to shed light on questions beyond economics, how Big Data is improving or changing economic models, and the kinds of collaborations arising around Big Data between economists and other disciplines.

Keywords

Big Data, economics, econometrics, interdisciplinarity, epistemology, business

Introduction

Big Data is increasing in importance as a source of information about the social world. A variety of social science disciplines have experimented with these new sources and types of data, with perhaps communications studies in the lead at the moment. However, the discipline of economics appears so far to have been fairly slow to pick up on the promise of this new category of data. In this paper, we focus on the field of economics as a case study to examine the adoption of Big Data approaches and epistemologies in the social science disciplines. Although every discipline is different in its reasons for adopting or rejecting Big Data analysis, economics may be a useful case study because, as we argue below, it occupies an interesting space at the intersection between academic and applied knowledge used for business purposes (see also Savage and Burrows, 2007, 2009) and it may therefore have a distinct trajectory in making inroads in the social sciences and in the uses of bigger and richer datasets. At the

same time, economics also has a strong body of theory and methodology which may make economists sceptical of Big Data sources and approaches and may pose unique challenges in this discipline concerning reliability and representativeness. For these reasons, exploring how economics is encountering Big Data may offer insights into the question of how Big Data is shaping – or not – the direction of the social sciences.

Our working definition of Big Data is that there is 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

¹University of Amsterdam, Amsterdam, The Netherlands

²Oxford Internet Institute, Oxford, UK

Corresponding author:

Linnet Taylor, University of Amsterdam, Plantage Muidersgracht 14, Amsterdam 1018TV, The Netherlands.

Email: l.e.m.taylor@uva.nl



(Schroeder, 2014). This definition is different from the terminology used in industry (Laney, 2001), which revolves around the ‘volume, variety and velocity’ of data, a definition also adopted by some of the academic economists that we focus on in this paper – even though they invariably use ‘Big Data’ that fits our definition, whether they are aware of definitions or not. Other researchers from our sample, particularly if they are connected to industry, may also use the terminology of computational methods to set their work apart from previous studies within the academic sphere (e.g. Tambe, 2012). Indeed, relying purely on the terminology of ‘Big Data’ is problematic, as economists and others working in this area may or may not actually use the term even if they are clearly operating within our definition or within a broader conception of computational methods. This can be seen in the number of papers which name ‘Big Data’ as a particular feature of their analysis, which is relatively small: a Scopus search for papers with ‘Big Data’ in the title, abstract or keywords currently (in January 2014) yields 2034 articles, of which only 32 are categorised as ‘economics, econometrics and finance’. Of course, this is just one indicator, which is not able to find uses of Big Data that do not mention the term, but in addition to this, to the best of our knowledge, there is only one current publication which explicitly advocates Big Data as an important force in the future of economics (Einav and Levin, 2013).

The lack of a clear adoption of terminology is not surprising in a new and still emerging area, even if various characteristics of Big Data clearly make it an important resource for economics. Einav and Levin (2013) have pointed out three of these main characteristics. First, that Big Data sources are frequently available in real-time, which can offer an advantage in terms of ‘nowcasting’, or identifying economic trends as they are occurring. The second relates to the scale of the data: the large size of the datasets becoming available resolves the statistical problem of limited observations and makes analysis more powerful and potentially more accurate, while their granularity (a characteristic Michael Zhang, Assistant Professor at Hong Kong University of Science and Technology, terms ‘nano-data’, following Erik Brynjolfsson (M. Zhang, interviewed 10 May 2013)) increases their power in terms of understanding individual actions. Third, such data often involve aspects of human behaviour which have previously been difficult to observe, for example personal connections (such as those within Facebook) or geolocation (such as the place from which a ‘tweet’ was sent via Twitter). However, Einav and Levin (2013) also point out some drawbacks which may have led to economists’ comparative reluctance to adopt Big Data so far. The main one is the unstructured nature of such

data and the complexity of the linkages often contained within it, which upset the usual econometric assumption that data points are not interdependent, or at least are interdependent in certain defined ways. As Einav and Levin point out, this complexity presents an econometric challenge in terms of untangling this dependence structure and understanding the data.

Besides the characteristics of sources of Big Data which make them suitable for economists, there are also certain ways in which economists are well suited to being users of Big Data. Big Data analysis demands technical skills in terms of statistics and coding which are part of the standard training for most economists. The econometric challenge of working with Big Data using statistical techniques appropriate for entire populations is part of a continuum of such challenges faced by economists as data sources have become larger and more complex over time, and the rewards of solving such problems, in terms of advancing the discipline, are potentially significant. Perhaps most importantly, there is a considerable amount of Big Data found within the traditional territory of economics: financial transactions of all kinds, including increasingly granular sources such as loyalty card data and online purchases, labour market data, and detailed population data. All these concerns suggest that Big Data is a potential goldmine for economists and that there may be a demonstrable opportunity cost for many economists in not engaging with this type of research. Yet, as we shall see, there are also limits to the uses of Big Data in economics, and these shed interesting light on its role among the social sciences and beyond.

Research questions and methods

This paper uses a series of interviews ($n = 17$) conducted with economists who have been working with Big Data, or data scientists working on questions within the economics or business fields, to examine the issues involved and the challenges and rewards of this type of data. These interviews are part of a larger project funded by the Alfred P. Sloan Foundation for which we have interviewed more than 125 social scientists over the period 2012–2014 (and still ongoing) using a semi-structured interviewing approach designed to elicit information about their engagement with Big Data, the tools and skills they use to work with data and learn more about how they gain access to data sources. The questions we ask here and the conclusions we draw out concerning economists’ use of Big Data are also informed by this larger study, which also relies on desk research, scientometrics, participation in various fora such as conferences about Big Data, and our own engagement in research in this area. The economists, like our other interviewees, were selected using

purposive sampling to find those working at the research front of Big Data in social science disciplines, and thus do not constitute a representative sample of any given discipline. However, given the newness of Big Data approaches to research, focusing on those engaged at the research front makes more sense than trying to understand Big Data use with random sampling techniques or measuring their impact via citations. Thus, this qualitative study is not necessarily representative, but is intended as an exploratory effort to uncover the motivations, practices and challenges encountered by social scientists, and thus to inform future directions in this area.

Two research questions guide this paper, as follows:

1. For what purposes are Big Data used (prediction and/or nowcasting, marketing research, substituting new or cheaper datasets for older ones, or other factors)?
2. Which type of economic or other knowledge advance is this use of data contributing to? In other words, which subdisciplines, economic methods, models, and motivations are apparent among early adopters of Big Data approaches in economics?

What constitutes Big Data within economics?

As already mentioned, finding a consistent definition of 'Big Data' in the field of economics is difficult, while understandings in the social sciences are still emerging around what constitutes 'big' versus 'not big'. An incremental rise in the number of data points does not serve as a definition per se – as a discipline, economics has generally aspired to the most extensive and detailed datasets possible, and has a history of adapting and evolving statistical techniques to deal with new types of data; nor does the need for programming skills – unlike some other social science disciplines, economists tend to learn to code in order to use analytical software such as R and SAS. Within the group of economists interviewed, there were a range of opinions on what constitutes Big Data, with the agreement that the specific terminology is fairly recent – although some were working with what is now being termed 'Big Data' a decade ago, most had not heard the term until around 2010, and agreed that it has not gained much traction within academic economics in particular. They did agree that it was possible to identify a class of data which was particular in terms of its size and complexity, although there were several different points of view as to which features rendered it genuinely new.

One common starting point amongst the economists interviewed was that the emergence of Big Data can be

situated within a continuum of developments in the discipline, and that the practices and perspectives which define economics are not particularly responsive to new levels of size or complexity in the datasets available. Within economics, Big Data cannot be characterised mainly as a shift in the sources of data, as is possible, for example, where new social media provide these sources in other social sciences such as media and communication studies. For instance, Professor David Hendry notes that there is a difference between macro- and micro-economics in terms of the number of observations commonly accessible to the researcher, so that 'in cross-sections relevant to macroeconomics about 1000 would be seen as Big Data and needing a lot of different methods of analysis' (D. Hendry, interviewed 24 April 2013).

However, many respondents did identify some aspects of Big Data which have epistemological or pragmatic implications for those economists who choose to engage with it. On the pragmatic side, Big Data can be characterized as highly multidimensional in terms of the number of variables per observation, the number of observations, or both, given the accessibility of more and more data – what Professor Hal Varian, Chief Economist at Google, referred to as 'fat data, long data, extensible data and cheap data' (H. Varian, interviewed 29 January 2013). Professor Liran Einav, an economist at Stanford, similarly identifies a trend towards data sources where 'you just know a lot of stuff on every observation... [such as] histories and stuff like that from which you could construct a very broad set of potential variables' (L. Einav, interviewed 20 February 2013).

This multidimensionality is also important because it necessitates new approaches and training. Nathaniel Hilger, an economics PhD at Harvard, defines it functionally in terms of the need for new or adapted analytical tools: 'It [Big Data] starts when you can't use Stata, I think' (N. Hilger, interviewed 29 March 2013). Similarly, Alberto Cavallo, an Assistant Professor at MIT, defines it in terms of a messiness that challenges current skillsets:

I think to me the big challenge now has become having people who have enough skills to be able to jump into a very messy data set that has been built for other purposes and then knowing what to look for and how to clean that data, and transform it into meaningful information, and I think that is going to be the big challenge. (A. Cavallo, interviewed 15 November 2012)

For Prasanna Tambe of NYU's Stern School, granularity is the defining feature of the new datasets. He offers an analogy with van Leeuwenhoek's invention of the microscope in the 17th century:

[With a microscope] you can basically look at one organism at a whole new level of detail. And I like that analogy for Big Data as well. That in a lot of ways we're looking at questions that people have looked at before, but you're just turning up the microscope. I think that's a pretty apt description when it comes to consumer spending, labour markets, crowd funding, there are so many examples I can think of where the questions are old but they will need to look at them with this new level of analysis that just, sort of explodes the number of policy implications and things like that you can get from them. (P. Tambe, interviewed 26 April 2013)

Other respondents offered a definition of Big Data as datasets relating to human behaviour, i.e. the by-products of people's use of technology and behaviour as consumers in a technologically-enabled market. For example, Duncan Simester, a Professor at MIT, takes the view that Big Data is 'micro-level detailed data describing some type of consumer behaviour. ... I could imagine that if I was in operations management it might be machine cycles...but it's a behavioural response measure' (D. Simester, interviewed 11 February 2013).

Sascha Becker, a Professor at the University of Warwick, defines Big Data from a methodological perspective as universal with regard to the phenomenon of study ('N=all'), and in turn to its characteristic of stretching computational resources:

[It is Big Data] in the sense that it's the universe, so literally all firms that are multinationals. That, for me, would be one definition of really Big Data as opposed to some sample. And that ... we linked up with the universe of all German workers, so we crossed 32 million German workers with 6000 multinational firms and also non-multinational firms, domestic firms. And that was, I guess, the first instance where simply computing power set certain limits in my research work. (S. Becker, interviewed 23 May 2013)

Finally, Einav suggests that the advent of the terminology of 'Big Data' in the economic sphere may be largely driven by industry. He sees corporations collecting ever more extensive and intensive data from their customers, and offering access to economists when they realise that their different interests in unlocking the data's value may align:

Except for maybe the more sophisticated companies out there, many of them just sit on their datasets and they realise they have potentially a gold mine of data but they have no idea what to do with it. So in that sense maybe what happens with Big Data is that more

and more, private and academic enterprises came along to say, 'Well, you know, you guys are all sitting on huge datasets and it's time for you to actually potentially get some value out of this'. (L. Einav, interviewed 20 February 2013)

These different viewpoints are not incompatible: access to Big Data appears linked at least partly to the corporate connections in the field of economics and in business in particular, but also to a desire on the part of many economists to find new perspectives on enduring questions. The next section outlines some of the main patterns in the uses of Big Data among the economists in our interviews.

Rationales for the adoption of Big Data approaches

As with any new technological development, the adoption of Big Data approaches has depended on various factors (such as data availability) and is taking place at different rates among different groups. The process of adoption may be top-down (institutionally driven), bottom-up (based on individuals' perceptions of an advantage) or, as is usually the case, a mixture of the two (Rogers, 2003). Economic analysis using Big Data has slowly been gaining social scientific traction on both these levels. On the institutional level, both the American Economic Association and the US National Bureau of Economic Research (NBER) held panels or workshops in 2012 and 2013 to discuss Big Data's potential in economic analysis, and the head of Pew Survey Research has described how interest is rising around the idea of using Big Data derived from social media and other transactional sources to supplement, and possibly in some cases as a substitute for, government statistical data gathered using traditional survey-based methods (Keeter, 2012).

Meanwhile, on the individual level, economists have adopted Big Data approaches where these can offer a new take on traditional economic questions such as labour market dynamics (Choi and Varian, 2012), the effect of early education on earnings (Chetty et al., 2011), stock market dynamics (Moat et al., 2013) and the workings of online markets (Einav et al., 2011). Some of these papers involve, or are even led by, computer scientists (Antenucci et al., 2013) or behavioural scientists (Moat et al., 2013), but with economists as co-authors. Beyond this, a minority trend is also emerging where 'Big Data economics' is effectively adopted from outside the field entirely, for example where computer scientists use Big Data to look at questions bearing on issues that are central to economics, such as Bollen et al. (2011) who studied the relationship between Twitter and the stock market.

Furthermore, economists appear to be engaging with different aspects of Big Data approaches depending on their priorities. For instance, ‘velocity’ is frequently named as an identifying characteristic of Big Data (Laney, 2001), but not all economists using Big Data are engaging with this aspect. Many are using a real-time feed of some kind, such as the Billion Prices Project at MIT (Cavallo, 2011), or the MIT project run by Duncan Simester, where real-time Twitter data will be used in combination with transaction data from stores to compare consumer sentiment with actual purchasing behaviour (D. Simester, interviewed 11 February 2013). The majority, however, draw samples from a Big Dataset within the company which owns the data, often using technology such as Hadoop or Apache Pig, and then analyse the data with ordinary statistical tools such as Matlab or STATA which do not take advantage of the data’s real-time aspect. Professor Hal Varian, Chief Economist at Google, pointed out that this way of working with otherwise unmanageably large datasets has analytic advantages in the context of economic analysis:

In a lot of cases drawing a signal from that data is just as good as using the data itself. So there are cases where the Big Data advantage can be exaggerated, and where sampling is the best procedure. ... And the advantage of sampling of course is that you can draw a repeated sample, so you can see how your results vary with the sampling distribution. (H. Varian, interviewed 29 January 2013)

Some economists see themselves as non-adopters, suggesting that there is nothing new in Big Data. This belief that Big Data represents just a point along a continuum of more or less extensive datasets is epitomised by the econometrician Professor David Hendry when he says that ‘whether the dataset’s big or small doesn’t actually matter in establishing change, but if it’s big and the system is complex the only way to establish change is to model that complexity’ (D. Hendry, interviewed 24 April 2013).

However, many of the economists interviewed who did see Big Data as a step change in the kinds of analysis that were possible said that using this type of data allowed them to address problems in innovative ways, and this also relates to their interest in new technical approaches. There was a consensus that the aspects of Big Data which seem to attract economists – that it is granular, population-level data with multiple dimensions that allow researchers to analyse cases along many variables – allow economic researchers to test theories of behaviour that were previously untestable, creating a new set of metrics for issues of economic

interest which were previously in the realm of theory. For example, Nathaniel Hilger, a former Harvard PhD student in economics and now an Assistant Professor at Brown University, has worked on several projects involving large-scale administrative data from the US Internal Revenue Service, and believes that this kind of population-level data is potentially revolutionary.

I think the essential feature of all these [Big Data] projects is that it uses a very large amount of sand to get enough gold to do causal inference on a question that hadn’t previously been able to be analysed as convincingly. ... another benefit of having Big Data is once you get the essential causal effect you’re looking for, if you have enough gold, you can then parse the gold to look at the effect on different subgroups and learn even more about what’s driving the causal effect. (N. Hilger, interviewed 29 March 2013)

One important rationale for using ‘born-digital’ data is that, in contrast to the classic survey-based datasets which have been the basis for much applied economics over the last century, economists can often collect it themselves. The ability to collect large-scale data independently can be especially powerful with regard to questions which have previously been the preserve of governments. One example of this is the Billion Prices Project, devised by Alberto Cavallo, now an Assistant Professor at MIT. The project, which involves programming a web scraper to gather online prices for goods and using them to compile an inflation index, was devised as a way to create a more transparent and accurate inflation measure in Argentina. The ability to access real-time price data has effectively positioned Cavallo’s research as an alternative to national governments’ inflation measures. Cavallo notes that the project seems to illustrate the case for more independent data collection among economists:

We [economists] have been using the same data sets over and over again, and since we wanted new answers, we have been developing new econometric techniques to try to transform the data, and get more meaningful information out of them. But it was reaching a point where there is nothing else you can do on that side, and just having a fresh, new data set brings a whole new perspective, and I think people are starting to realise that, and gradually people are becoming more interested in data collection itself. (A. Cavallo, interviewed 15 November 2012)

Besides the two extremes of survey versus scraped data, the option of on-demand data such as that provided by Google Analytics is also proving a reason to explore

new opportunities. A currently high-profile example is in the field of ‘nowcasting’ (Choi and Varian, 2012), which uses what might be termed curated synopses of huge datasets, such as people’s web searches through Google, to make highly accurate short-term predictions. The well-known example from Choi and Varian examines consumer and labour market trends, suggesting that the changing volume of queries about given products or services on Google closely mirrors demand. The project illustrates how adopting Big Data approaches may not involve learning new computational techniques, or necessarily challenging the discipline’s methodological bounds. The important question may be whether the data gathered using these new sources raises new epistemological concerns, and whether it takes economists outside their comfort zone in terms of reliability and replicability. We will address these questions in the section that follows.

The challenges of interpreting Big Data

Although economists generally have sophisticated statistical skills and plenty of expertise in coding large datasets, the new sources of data described here present challenges which highlight issues in how Big Data is becoming part of social science research. For example, the size of Big Data may render the idea of statistical significance, a mainstay of hypothesis-testing, useless. Varian says, ‘when you have a billion observations, everything’s significant’. Varian and another senior economist, David Hendry, have very different approaches to the interpretative issues highlighted by Big Data. Varian is prompted to ask whether it is time to officially separate the statistical notion of significance from its more general meaning – a discussion which has been underway in the natural sciences, notably medical statistics, for several decades (e.g. Gardner and Altman, 1986). For Varian, the significance problem highlights existing weaknesses in economic practice which can be resolved by taking a broader view of what is worth reporting:

you really do have to address what we should be addressing all along, the importance – unless we use significance in a phenomenal sense, or an operational sense, not the statistical sense. Because after all, statistics was designed to deal with datasets of a hundred or so observations, when you look at it. So we’ve developed some bad habits, I think, in terms of misusing statistical terms. (H. Varian, interviewed 29 January 2013)

David Hendry is concerned with the argument, as voiced by Mayer-Schönberger and Cukier (2013), that

much work using Big Data is essentially descriptive, dealing with correlation rather than causality. He notes that if economics cannot seek causality, it similarly loses one of its mainstays:

it applies in epidemiology, it applies in sociology, political science and in economics that you get large datasets, and under the null [hypothesis] that there’s no connection you will find lots of connections unless you’re extremely careful about how you analyse it. Many of the methods of analysis that I see people using, even through to genetics and studies of DNA, are using methods that I think are seriously flawed in terms of picking up things that are not there. (D. Hendry, interviewed 24 April 2013)

In contrast to Varian’s approach of adopting more of a phenomenal lens, Hendry advocates sharpening economists’ modelling techniques (chief of which is what is known to economists as the ‘LSE/Hendry approach’) to make economic analysis more powerful, regardless of the size of the dataset.

Despite these problems, which are both methodological and epistemological, the debate about whether access to extensive, highly granular data heralds the ‘end of theory’ (Anderson, 2008) has found only limited resonance in economics. Professor Sascha Becker of the University of Warwick suggests, like Varian, that Big Data will cause economists to reevaluate their assumptions, but rather entails a more iterative interaction between theory and empirical data:

I think theories [in the light of Big Data], they don’t have the same value. It’s more about in the past maybe we would have theory and then would do simulations and calibrations and then make predictions about what might happen. And Big Data allows you to really go out there and measure stuff. But still you will need theory to understand the mechanisms or even to suggest what you might hope to find in the first place. (S. Becker, interviewed 23 May 2013)

If Big Data is causing these economists to reevaluate the explanatory power of economic methodology, it is also causing some to reevaluate the explanatory power of economists themselves. The Council of Economic Advisors in the USA makes annual predictions of unemployment rates and other indicators, and an experiment at the University of Michigan is being conducted by a team of economists in collaboration with Mike Cafarella, a Professor of Computer Science, to test whether Twitter may outperform the Council as a predictive tool. The premise of the study, Cafarella explains, is to assume that the economists’ errors are

random and therefore (to the economists) not predictable, and to see whether Twitter can quantify them:

If the social media data is actually carrying some brand new information in the universe, something that we didn't have previously, then we should be able to predict [professional economic advisors'] error. ... and at least in the case of unemployment using Twitter we were able to predict about one third of the error. (M. Cafarella, interviewed 2 July 2013)

This is a particularly interesting project because social media has been criticised as having unknown bias and therefore being of questionable reliability as a social scientific tool (González-Bailón et al., 2014, forthcoming), yet the Michigan project sets it against human judgement with the aim of quantifying error margins.

Another issue is the emerging uses in the social sciences of data mining. A term which used to denote 'bad' quantitative social science which lacked a clear hypothesis, with the advent of Big Data this is becoming a more credible form of research. A study by Michael Zhang published in the *American Economic Review* demonstrates how attitudes to data mining are changing: Zhang, who moved into economics with a background in computer science, data-mined Wikipedia content in order to develop his research questions, noting that he and his collaborator Zhu 'needed time to process the data before we actually came to the research question ... basically, all these questions came after we had the data'. The data mining led to two papers on behavioural economic questions (Zhang and Zhu, 2006, 2011), the latter a natural experiment made possible by the Chinese government's on-off blocking of Wikipedia.

Similarly, Einav and Levin's (2013) work with eBay auction data has involved data mining in order to search for the right questions. They describe how rather than seeking out a particular dataset to answer an established question, as is common with economists who work on survey data which is curated and therefore has more predictable contents, a windfall of 'data in the wild' such as the by-products of consumers' eBay use may require a very different strategy to seek the right question. It also may require a different timescale from curated data, weighted towards question development rather than model-based analysis:

So we kind of came to it not having a particular idea of what exactly we want to do. We just wanted to formulate reasonable questions that could kind of leverage the idea that you have the Big Data rather than some sort of a smallish portion of it. So initially we were basically for six months just playing with the data,

trying to understand, you know, what we could do with it and what could be interesting. (L. Einav, interviewed 20 February 2013)

Einav and Levin's work illustrates how economists may know that a dataset contains great analytical value without being able to specify that value in advance. Along with Zhang and Zhu's research, it suggests that rather than being exclusively hypothesis-led, economics research using Big Data may need to work towards a different, or broader, definition of methodological rigour to take into account data where most of the uncertainty is weighted towards the pre-analysis phase, and once the data is understood through a particular question, the extensiveness of the dataset makes the process more of a snapshot than an excavation.

The challenge of access

Einav and Levin (2013) argue that Big Data approaches have the potential to allow economists to ask a great variety of new questions. Arguably, however, these questions depend largely on researchers' ability to access new sources of data, most of which are proprietary. The challenge of access to appropriate data for one's research is not new, and given the proprietary nature of much Big Data, similar hierarchies are likely to emerge to those already existing in the discipline of economics, where senior researchers have the resources, influence and networks to gain access to the 'best' data. Corporate data in particular presents similar problems regardless of its size, since it is proprietary and tends to be offered to researchers only subject to non-disclosure agreements which may limit the replicability of studies. The disciplinary expectations around replicability and access to data may have to relax as more researchers use Big Data for their studies – or in an alternate scenario may grow more stringent as datasets invisible to all but the author become more common. Michael Zhang suggests that a new politics of data will emerge, but that either scenario is possible:

The usual practice is to sign some NDA [non-disclosure agreement] kind of arrangement and then by the time when the paper can be published sometimes, you know, some companies, they don't care about the data anymore. ... so far no journal has a policy to say that you cannot publish if you don't share, so, there's no threat to authors – but in future I would imagine people will. (M. Zhang, interviewed 5 October 2013)

Getting access to the 'best' sources of data is, as noted above, traditionally an issue of hierarchy. The research team working with Raj Chetty at Harvard, for example, has access through him to US Internal Revenue Service

data of unprecedented size and detail on individuals' employment history, which they have used to produce groundbreaking analysis of, for example, how early childhood education affects people's life chances (Chetty et al., 2011). However, Nate Hilger, Chetty's research team member, described the process of obtaining and sustaining access to such a huge and detailed dataset as a significant investment of time and effort. The research team could only access the IRS data in secure data rooms authorised by the IRS central office, they had to get what he described as 'fairly, I think, high level security clearance', which 'took months', and involved the team submitting information on 'everywhere we'd lived for the last ten years' (N. Hilger, interviewed 29 March 2013).

The investment of time and effort is no less when accessing a highly restricted dataset such as the employment histories stored in LinkedIn's database in the firm's Mountain View headquarters. However, the difference between IRS data and born-digital data from an internet firm is that younger researchers can, with the right contacts, gain access as easily as senior ones. They may even have more chance of access if they have a high level of relevant technical skills, as did Prasanna Tambe, who worked on LinkedIn data to produce a study of employment dynamics (Tambe, 2012). Tambe originally self-funded his research as a summer project. He got access to the data through a combination of having fairly advanced programming skills (learned during a masters in Computer Science and a PhD in Economics) and via connections in California which allowed him to spend a significant period at the company's headquarters working with the data, which (similarly to the IRS data) could not be accessed outside the firm's building. The effort involved in getting access was similar to that with IRS data, but significantly more informal:

There's not an easy answer in the sense that what you'd like to be able to give is sort of a blueprint as how you could do this...[There are] various companies, and there's maybe half a dozen of the big ones, all have their sort of own incentives. ... I knew somebody who knew somebody or just reached out randomly and got a response and... it took usually multiple conversations or contacts, visits. So it wasn't that it was that easy or direct, a direct interface through which you could access the data. (P. Tambe, interviewed 26 April 2013)

However, the return for his effort was a more detailed dataset than the IRS could offer. Job websites such as LinkedIn, Monster.com and Careerbuilder.com collect individual-level sequential employment histories at a level of detail not offered by national administrative

data. Tambe describes it as 'job titles...what skills people have, what employers they worked at, occupational level detail, all those things were sort of, you might call a new level of granularity for labour data'. Tambe's experience suggests that a generational and geographic divide between more traditional research based on large survey-based datasets and economists using datasets from the big internet firms, often based in Silicon Valley.

The challenge of data access is driving significant career changes in the field of economics. Senior economists such as Hal Varian (at Google) and Bernardo Huberman (at Hewlett Packard) are working outside the academy for global corporations where they have access to new and privileged sources of data, and such moves are becoming more common amongst mid-career economists. So far, many are keeping an affiliation within academia: Patrick Bajari, Vice President and the Chief Economist at Amazon.com, remains a Professor of Economics at the University of Washington, and Steven Tadelis, Distinguished Economist at eBay, also holds the post of Associate Professor at UC Berkeley's Haas School of Business. Over time, this increasing corporate affiliation on the part of some of the best-known economists in the field may even contribute to a scenario where having access to corporate data becomes a factor in universities' hiring decisions.

Although economists with coding skills such as Cavallo have acquired important datasets such as online price data through computational methods alone, most Big Data remains proprietary. Google's publicly available datasets such as Insights and Trends are curated, and it is unlikely that the company will make the full extent of search data public any time soon. Varian says that the company does not like the prospect of negotiating access with individual researchers 'dealing with things on a case by case basis', and therefore has decided to 'make data available to everybody or to nobody'.

Conclusion: The implications of Big Data for economics

We have sought to answer two main questions in this paper. First, we have outlined the purposes for which Big Data is being used, and demonstrated that Big Data applied to economic questions has the potential to be disruptive both methodologically and epistemologically. It may reframe some questions that are, or should be, important to economists, and may do so in ways that lead to new styles of thinking and investigation. Furthermore, given that the development of methodologies for the analysis of Big Data presents various challenges, econometrics may provide a useful bridge

between computer science, which can access and manipulate the data and do the calculations but does not traditionally contend with the questions of representativeness or validity, and the social sciences, which are interested in these questions.

Big Data has emerged as a strong fit with behavioural economics and part of the economics discipline related to industry in particular, but it may also present some interesting challenges for economics within academia. Porter (2010) has written that ‘statistical reason is the beacon of an ideal of impersonal rationality achieved through technical methods’ (pp. 45–46). This paper has suggested that Big Data has the potential to challenge this notion of rationality – for example, by addressing the messy problem of sentiment analysis in social media in the cause of predicting economists’ error rate on particular questions. It may also have the power, however, to stimulate new thinking on issues such as external validity, ways to address sampling bias, and the ways in which, rather than allowing the end of theory, exponentially larger datasets can be used in combination with sophisticated modelling strategies to produce more detailed and accurate explanations of social processes.

Second, we have addressed the question of how Big Data is contributing to advances in knowledge. We have shown that the intersection of economics and Big Data poses some important questions for economics and social science in general. One can see the shift towards Big Data as being a change in methodological emphasis, a change in data management and analytic tactics, but one can also see it on another level as necessitating a more fundamental shift in perspective from a science based on the notion of the mean and the standard deviation from the ‘normal’ to one based on individual particularity – an epistemological change which brings into question some of the fundamental tenets of economics as a discipline. If one takes this last perspective and looks at Big Data as a qualitative as well as a quantitative change, one can also see challenges for economists in conceptualising these new datasets and methods.

Besides the epistemological challenge, there are pragmatic issues to be resolved if the discipline is to engage with Big Data more broadly. Several questions arise from our interviews: will Big Data democratise access to the most valuable data for economists, or does it make less even the playing field for those in less well-resourced positions or institutions? Will it lead to new hierarchies forming around different forms of access and new sources? Second, will economists’ statistical and computational skills enable them to participate in developing new methodologies and analytical approaches for Big Data, and will their concern with generalisability and reliability prompt new

approaches which address Big Data’s uncertain biases? Finally, our interviews also raise the question of replicability. We have outlined a picture of Big Data that is largely proprietary, with open-access data already curated by firms in ways which are inaccessible to researchers. If Big Data holds great promise for economics, as many of our interviewees believe, will the way economists present and publish their results have to change, and will the field have to accept less access to datasets and limited or no replicability?

For these reasons, the Big Data turn, if such it is, may possibly be disruptive in economics, with possible analogous ruptures in other social sciences. We have seen this before: the invention of statistics in the 1800s allowed the emerging discipline of economics to take a new turn in the analysis of social dynamics. It is also possible that instead of Big Data becoming an accepted stream within economics, it could give rise to a sub-field of its own with separate disciplinary and methodological norms, with the implication that those who practise this type of analysis will become separated from or balkanised within the larger discipline. So far, there are enough prominent economists engaging with Big Data studies (Brynjolfsson, Varian, Huberman, Poterba) such that Big Data can be seen as an emerging specialism rather than a break from the discipline. However, the emergence of conferences or sessions concentrating on Big Data suggests, if not marginalisation, then at least that Big Data is a specialisation rather than the future of the mainstream.

This paper has outlined the challenges and potential rewards of using Big Data in the field of economics. Overall, the evidence presented here suggests that the value of Big Data to the discipline may lie partly in creating a stimulus for new ways of thinking, but specifically in challenging economists to imaginatively apply an economic perspective to the evolving digital landscape. This work may be anchored in strong lines of existing inquiry – most of those interviewed for this paper are applying new datasets to economic issues they have been interested in their whole careers – but often involve a new way of seeing existing information and inventive methods to separate the signal from the noise. The most innovative work is being done by those who, as Nathaniel Hilger put it, can devise ways to see the gold amidst the sand.

Declaration of conflicting interest

The authors declare that there is no conflict of interest.

Funding

This research was supported by a grant from the Alfred P. Sloan Foundation.

References

- Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*, 16 July. Available at: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed on 1 April 2014).
- Antenucci D, Cafarella MJ, Levenstein MC, et al. (2013) Ringtail: Feature selection for easier nowcasting. In: *WebDB '13: Sixteenth International Workshop on the Web and Databases*. New York, NY.
- Bollen J, Mao H and Zeng X (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2(1): 1–8.
- Cavallo A (2011) Scraped data and sticky prices. Paper presented at the American Economic Association annual conference, Denver, CO. Available at: <http://www.aeaweb.org/aea/2011conference/program/retrieve.php?pdfid=403>.
- Chetty R, Friedman JN, Hilger N, et al. (2011) How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics* 126(4): 1593–1660.
- Choi H and Varian H (2012) Predicting the present with Google trends. *Economic Record* 88(S1): 2–9.
- Einav L, Kuchler T, Levin JD, et al. (2011) Learning from seller experiments in online markets. NBER Working Paper Series No. 17385. Available at: <http://www.nber.org/papers/w17385>.
- Einav L and Levin JD (2013) The data revolution and economic analysis. NBER Working Paper Series No. 19035. Available at: <http://www.nber.org/papers/w19035>.
- Gardner MJ and Altman DG (1986) Confidence-intervals rather than P-values: Estimation rather than hypothesis-testing. *British Medical Journal* 292(6522): 746–750.
- González-Bailón S, Wang N, Rivero A, et al. (2014, Forthcoming). Assessing the bias in samples of large online networks social networks. *Social Networks*.
- Keeter S (2012) Presidential address: Survey research, its new frontiers, and democracy. *Public Opinion Quarterly* 76(3): 600–608.
- Laney D (2001) 3D data management: Controlling data volume, variety and velocity (META Group File 949). Available at: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that Will Transform How We Live, Work, and Think*. New York: Houghton Mifflin Harcourt.
- Moat HS, Curme C, Avakian A, et al. (2013) Quantifying Wikipedia usage patterns before stock market moves. *Scientific Reports* 3(1801): 1–5.
- Porter TM (2010) Statistics and the career of public reason: engagement and detachment in a quantified world. In: Crook T and O'Hara G (eds) *Statistics and the Public Sphere: Numbers and the People in Modern Britain*. New York: Routledge, pp. 32–47.
- Rogers E (2003) *Diffusion of Innovations*, 5th ed. New York: Free Press.
- Savage M and Burrows R (2007) The coming crisis of empirical sociology. *Sociology* 41(5): 885–899.
- Savage M and Burrows R (2009) Some further reflections on the coming crisis of empirical sociology. *Sociology* 43(4): 762–772.
- Schroeder R (2014) Big Data: Towards a more scientific social science and humanities? In: Graham M and Dutton WH (eds) *Society and the Internet: How Networks of Information are Changing our Lives*. Oxford: Oxford University Press.
- Tambe P (2012) *How the IT workforce affects returns to IT innovation: Evidence from Big Data analytics*. NYU Stern School of Business. Available at: http://www.krannert.purdue.edu/faculty/kkarthik/wise12/papers%5Cwise12_submission_56.pdf.
- Zhang X and Zhu F (2006) Intrinsic motivation of open content contributors: The case of Wikipedia. In: *Workshop on Information Systems and Economics*.
- Zhang X and Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101(4): 1601–1615.