Rob Kitchin: “Big data should complement small data, not replace them.”

blogs.lse.ac.uk/impactofsocialsciences/2014/06/27/series-philosophy-of-data-science-rob-kitchin/

6/27/2014

Over the coming weeks we will be featuring a series of interviews conducted by **Mark Carrigan** *on the nature of ‘big data’ and the opportunities and challenges presented for scholarship with its growing influence. In this first interview,* **Rob Kitchin** *elaborates on the specific characteristics of big data, the hype and hubris surrounding its advent, and the distinction between data-driven science and empiricism.*

What is ‘big data’? Is this a term we can meaningfully use given the hype surrounding it, or does it obscure the continuities between contemporary fashions and longer standing trends within the social sciences?

As with many terms referring to new, diverse and quickly developing technological phenomena big data has been variously defined. Some definitions, whilst simple and clear — such as big data being any dataset that is too large to fit in an Excel spreadsheet — have limited and misleading utility as they do not get to the heart as to what is different ontologically and epistemologically about big data. And there is a significance difference, which is why there is so much hype surrounding these data.

For me, big data has seven traits — they are:

- huge in *volume*, consisting of terrabytes or petabytes of data;
- high in *velocity*, being created in or near real-time;
- diverse in *variety* in type, being structured and unstructured in nature, and often temporally and spatially referenced.
- *exhaustive* in scope, striving to capture entire populations or systems (n=all)
- fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely *indexical* in identification;
- *relational* in nature, containing common fields that enable the conjoining of different datasets;
- *flexible*, holding the traits of extensionality (can add new fields easily) and *scalable* (can expand in size rapidly).

Big data then are not simply very large datasets, they have other characteristics. This is why a census dataset does not constitute big data as I define it. True, census data are huge in volume, seeks to be exhaustive, and has high resolution (though they are usually aggregated for release), indexicality and relationality. However, a census has very slow velocity being generated once every ten years, very weak variety consisting of 30 to 40 structured questions, no flexibility or scalability (once formulated a census is set and is being administered it is impossible to tweak or add/remove questions). Another traditionally large dataset, the national household survey, is more timely, usually administered quarterly, but at the sacrifice of exhaustivity (it is sampled), but likewise lacks variety and flexibility.

Image credit: Paul L Dineen (Flickr, CC BY)

In contrast, in 2012 Facebook was processing 2.5 billion pieces of diverse content (links, comments, audio and video clips, etc), 2.7 billion 'Like' actions and 300 million photo uploads *per day* and Wal-Mart was generating more than 2.5 petabytes (2^{50} bytes) of data relating to more than 1 million customer transactions *every hour*. These are systems that can be altered on the fly, with new forms of content and metadata fields added, and can cope with large ebbs and flows in data generation.

Early forms of big data include continuous streams of remote sensing, weather, and financial markets data, but in recent years their production has exploded to include data generated by digital devices such as smart phones, transactions and interactions across digital infrastructures (for example, email or online banking/purchasing), clickstream data that records navigation through websites and apps, measurements from sensors embedded into objects or environments, the scanning of machine-readable objects such as travel passes or barcodes, digital surveillance technologies such as CCTV, and social media postings. Most of these data hold all or most of the characteristics of big data, with velocity being the key attribute, and there are no doubt varieties of big data. The key issue is that handling, storing and analyzing such data is a very different proposition to dealing with a census and that has only become possible with advances in computation, widespread internetworking, new database structures, and new forms of data analytics.

Viewed in this way I do think we can use the term meaningfully, regardless of the hype and hubris surrounding it. Indeed, I think it is the job of academics to unpack and conceptualise such terms in order to think through the characteristics of a phenomenon that has rapidly gained discursive and material traction and is having major effects in shaping company and government policy, investment and practices. It is in this way we can start to get a handle on what big data means for social sciences research and praxes and what it means for how government and business are conducted. In other words, the term only obscures if we allow it to.

Is there a risk that critiquing the excesses so readily identifiable in the discourse surrounding 'big data' could lead to a failure to recognise the immense value of new computational techniques and other innovations?

Nothing is, or should be, beyond critique. Moreover, to date the hype and hubris surrounding big data has by far outweighed the critique that has been levelled against it. Funding agencies and corporate R&D departments are

queuing up to pump money into big data and data analytics projects and training programmes. Yet, if we want big data to deliver valuable insights that are scientifically robust and valid then we need to work through the various critiques that are being levelled at them and to make appropriate adjustments to methods and approaches. This makes absolute sense from a scientific perspective – science progresses by improving its ideas, methods and theories in response to critique and internal reflection. It shouldn't hide from or deny critique, but rather should embrace and confront it. At the same time, we shouldn't let the hype of big data and the value of new computational techniques detract from the value of other approaches. Big data should complement small data, not replace them.

You've distinguished between a new empiricism, in which 'big data' is understood to speak for themselves, and a data-driven science which has a more nuanced understanding of data and how it is constituted. Could you say a little more about this distinction and why it matters?

In the initial hype about the explosion in big data and its potential in business, government and science there were proclamations about how big data could revoluntionize how we make sense of the world. Big data was seen to offer the possibility of exhaustive, timely, resolute data across all domains of human endeavour and, along with new sophisticated machine learning techniques, created a situation wherein data could seemingly speak for themselves. In the words of Chris Anderson, the Editor of Wired, "We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot. […] Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

In other words, our understanding of the world could emerge from the data, not from theory; from induction not deduction. This is a problematic position for a number of reasons. Big data, despite its attempts to be exhaustive is always partial, with gaps, biases and uncertainties. Moreover, the data do not come from nowhere, but are produced by systems that are designed and tested within scientific frameworks, and are surrounded by an assemblage of different contexts and interests. The production of big data is thus driven by knowledge and other factors which mean the data never simply speak for themselves.

In reaction, a number of scientists have forwarded the notion of data-driven science. This approach uses a guided approach to mining the data, using established knowledge and abduction to direct the exploratory analytics employed. The resulting data are then used formulate hypotheses that are then tested using a traditional deductive approach. Here, there is a modification to scientific method, with a new stage inserted between inception and testing. By using a guided approach, a controlled and contextual approach can be undertaken that can assess whether patterns in the data are meaningful or random/trivial, and the emerging hypotheses can be evaluated against existing knowledge/theories.

The reason why this distinction matters is because each proposed approach has a different means of seeking to extract value and insight from big data. Whilst data-driven science has its strengths and weaknesses it seems to have far more benefits that the empiricist approach. In fact, the empiricist approach seems like an intellectual cul-de-sac that will lead to erroneous conclusions and I doubt few serious scientists would subscribe to its tenets.

Do we need a philosophy of data science? Is this something that data scientists would actually be influenced by? Does it matter?

There are inherent philosophies of data science, whether data scientists recognize this or not. Even if a scientist claims to have no philosophical position, she is expressing a conceptual position about how she makes sense of the world. On questioning, their position with respect to epistemology, ontology, ideology and methodology can be teased out (though it might be slightly confused and not well thought through). Philosophy is important because it provides the intellectual framework that shapes and justifies what kinds of questions are asked, how they are asked, how the answers are made sense of, and what one does with the resulting knowledge. Avoiding it weakens the intellectual rigour of a project and widens the scope of potential critique. Quite often scientists avoid the difficult work

of thinking through their philosophical position by simply accepting the tenets of a dominant paradigm, or by operating merely at the level of methodology. Generally, this consists of a claim to using the 'scientific method' which tries to position itself as a commonsensical, logical, and objective way to approach understanding the world that is largely beyond question.

As I've already discussed, the philosophy of science is not fixed and does change over time with new ideas about how to approach framing and answering questions. This is clearly happening with debates concerning how big data and new forms of data analytics is and can alter the scientific method, and also debates over the approach of the digital humanities and computational social sciences. And even if data scientists do not want to engage in such debates, their work remains nonetheless open to philosophical critique. In my view, the intellectual rigour of data science would be significantly improved by working through its philosophical underpinnings and engaging in debate that would strengthen its position through evolution in thought and practice and which rebuffs and challenges critique. Anything less demonstrates a profound ignorance of the intellectual foundations upon which science is rooted. So, yes, philosophy does and should matter to data science.

Image credit: Wired, Richard Barrett-Small (Wikimedia, CC BY)

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Prof. Rob Kitchin *is an ERC Advanced Investigator on The Programmable City project at the National Institute for Regional and Spatial Analysis at the National University of Ireland Maynooth. He's PI for two data infrastructures – the All-Island Research Observatory and the Digital Repository of Ireland – and is author or editor of 22 books. His webpage can be found here. He tweets* @RobKitchin.

Mark Carrigan *is a sociologist based in the Centre for Social Ontology at the University of Warwick. He edits the Sociological Imagination and is an assistant editor for Big Data & Society. His research interests include asexuality studies, sociological theory and digital sociology. He's a regular blogger and podcaster.*