Emma Uprichard: Most big data is social data – the analytics need serious interrogation

blogs.lse.ac.uk/impactofsocialsciences/2015/02/12/philosophy-of-data-science-emma-uprichard/

In the final interview in our Philosophy of Data Science series, **Emma Uprichard***, in conversation with* **Mark Carrigan***, emphasises that big data has serious repercussions to the kinds of social futures we are shaping and those that are supporting big data developments need to be held accountable. This means we should also take stock of the methodological harm present in many big data practices. It doesn't matter how much or how good our data is if the approach to modelling social systems is backwards*.

This interview is the last installment of our series on the Philosophy of Data Science*. Previous interviews:* Rob Kitchin*,* Evelyn Ruppert*,* Deborah Lupton*,* Susan Halford*,* Noortje Marres*, and* Sabina Leonelli*.*

You've argued that we're witnessing a 'methodological genocide'. Is this a matter of the hype surrounding data science or something intrinsic to it?

I think it's a bit of both – and a bit more. 'Methodological genocide' is certainly a very strong term. However, I used it deliberately and also very carefully; 'genocide' in whatever context it is used is serious. One of the editors asked me whether I was sure this was the right term given the connotations that 'genocide' raises. Perhaps the term is overstating the situation and no doubt other ones might have been more appropriate. But what I was trying to get at was that there is something non-trivial, something about the violence, being committed under the guise of 'big data' which is happening at a *methodological* level that is not being discussed. Many people tend to focus on the ethical, privacy and security issues intrinsic to big data, or on big data's capacity to transform particular areas of social life, such as health, education and cities. These issues are certainly important and they are of course closely linked with methodology too. However, there are some even more fundamental questions about what big data are ontologically and epistemologically that are largely being ignored.

I'm not saying that everything about big data is bad, negative or terrible. Far from it. I think a lot of good is already coming from it. Making more visible the data systems that are already making and shaping everyday life is a good thing. We do need to pay attention to the interconnected systems of data systems. Who is generating the most or least data? Who owns it? Or who owns part/s of it? Who *can* access and process (parts of) the data and are these the same people who actually do? To whom do the analytics and findings go to and for which purposes? Who is profiting the most and least from big data? These are questions that need urgent attention and they also need constant 'updating' since, as we know, personnel within organisations changes, data circuits morph as well as split and merge, and new laws and cultures around what is or is not acceptable in relation to data keep changing.

Image credit: Karen Eliot (Flickr, CC BY-SA)

The issue is, though, not whether we have small or big or even bad or good data, although these may be contingent necessary conditions for the kinds of questions that may be asked about the possibilities of data in general. A more difficult question to answer well is: how are we to study the social world empirically given that all social systems – including big data systems – are open, historical, complex, dynamic nonlinear and tend to towards states far from equilibrium as they evolve through time and space? Moreover, how might we produce empirical descriptions and explanations about/within these system, which are also adequate at the level of both cause and meaning? And how might we do this in a way that is useful for policy and planning purposes? How precisely our use of big data is going to be able to contribute to understanding the social from the perspective that big data is both ontologically and epistemologically part of the social structures from with it emerges is far from clear.

My worry is that there are already modes of practice that are already being laid down, attitudes and cultures that are being normalised, laws being set up, and global networked infrastructures that are being created with little thought to given to what is 'social' about them – and indeed often by those with little or no training in how the social is considered to work. To many the question of what is 'social' about the data is not even a necessary question because there seems to be an assumption in some circles that the 'social bit' doesn't really matter; instead what matters are the data. But in the case of big data, the data usually are the social bit! I cannot emphasis this point enough: most big data is *social* data. Yet the modes of analysis applied to big data tend to mirror approaches that have been long been rejected by social scientists. And importantly, they have been rejected *not* because of the 'discomfort' that comes from the idea of modelling human systems using modes of analysis useful to model and think about atoms, fluid dynamics, engine turbulence or social insects etc., although that may be part of it. These kinds of big data methodologies might well be used meaningfully to address certain kinds of questions and ones that we haven't even asked before. But one of the main reason these 'social physics' approaches tend to be rejected by social scientists is that these methodological approaches don't so much as even a nod to how important social meaning, context, history, culture, notions of agency or structure might be – and yet these matter enormously to how we use data to study social change and continuity.

At the end of the day, it doesn't matter how much or how good our data is if the epistemological approach to modelling social systems is all backwards. To use big data to simply drive forward positivist notions of the social without accounting the important of history, culture, meaning, context, agency and structure in shaping social life is on a hiding to nowhere good. Yet so far the signs are precisely that big data analytics are going down that path. So

when I say that there is a methodological genocide going on, what I am getting at is that at a methodological level, the data analytics need serious interrogation particularly in terms of exactly they can or are going to improve our lives. As Shaw argues (2007), there are a number of 'sides' to genocide, and likewise here the methodological genocide that I am hinting at is epistemological, disciplinary, technological, urbanised, elitist, commercialised and, I dare say, often gendered. There may be a view that these things are not so important because what really matters is about finding tools that fit the data. But unless the tools also fit the social that you want to re-produce then it is unlikely that they will be much help in transforming anything for the better. If you produce bad findings – and bad findings can simply be being economic with the messiness of the social systems that are being explored – and these bad findings are picked up and used, then somewhere along the line the issue of liability needs to be raised. I am sure that, with a bit of brainstorming, we could develop a number of ways to ensure that those driving big data futures in the present were held more accountable both in the present and future. Big data has potentially serious repercussions to the kinds of social futures we are each complicit in shaping and those with the analytical capacity or institutions that are supporting big data developments need to be held accountable – and this means that we also take stock of the methodological harm that is already present in many big data practices.

You've suggested that 'big data' cannot deal with 'big questions': what do you mean by this?

For me, the promise of big data rests on being able to tackle big nasty global problems that we haven't yet been able to do enough about, e.g. hunger, poverty, inequality, racism, sexism, etc. There are so many of 'wicked social problems' that I can't even list them as I'll be missing some out and by the time I'll have finished this sentence, some new ones will be in the making anyway. It is not that I do not think that big data cannot help to answer some things we could not have answered before or that it does not offer a way into thinking about things differently. I think the level of granularity – a new kind of macro, if you like, is interesting. I can't help but find some of the 'live' traffic modules mesmerising. But to me it is a bit like observing ants and saying, 'Hurray! We've cracked it. Ants are good at following each other and they also build ant hills'. Ask any experienced taxi driver in any city and he or she is likely to be able to give you a much better and more useful image of where traffic tends to being jammed and indeed will also be able to suggest ways probably about how to un-jam the roads. Perhaps I'm missing the point about models like these, but in terms of research design, my money would be on collecting information from taxi drivers rather than big traffic GPS analytics. It would be great to use both of course – not least because many taxi drivers are also relying on GPS systems too!

At the end of the day, what matters is not whether data are big or small, but rather what kinds of questions we ask generally about the kinds of interventions we want to instigate. I'm a sociologist by background and so I've been trained to think critically about social divisions – what makes them, what shapes them, where they are, how they are maintained? My methodological interests are primarily to do with how to design and conduct empirical research that might help to destabalise the causal processes that re-produce social divisions over time and space. Essentially I bring ontological and epistemological assumptions (about complex systems) to how social change and continuity happens the way it does (and not another way) and I am interested in developing methodological repertoires that might help to un-do – at the very least try not to replicate and reproduce – some of the wicked problems mentioned above. In other words, I am interested in general questions such as: How do we produce empirical models that account for our desired futures and not merely projected ones? How might individual and collective forms of agency through generations and the life course be part of those models? How do we account for context and acknowledge that social systems will change as they move towards a changing attractor?

I know there are exceptions and am sure I have missed many good examples – indeed, the Masters students I'm teaching on the new module on 'Big Data Research: Hype or Revolution?' are exposing me to excellent new readings, which is very helpful! But so far the way that big data is being used or discussed tends to be about how to make more money or how to do things more cheaply; or how big data will ultimately provide those who own and access our personal data to perform experiments on us or manipulate 'driver nodes', either out of curiosity or to make further financial profits. I am sure that big data might be able to be used to answer the big problems, but in order to do so, it will have to be used alongside and as part of other *social* theories of change and continuity more

generally.

Is it possible to sustain a methodological pluralism in an 'age of big data'? Or do too many interests mitigate against this?

It is not only possible but it is fundamental! I don't think big data changes that. And in fact many big data practices actually involve the use a range of techniques, so I think methodological pluralism is firmly built into big data already. We might dispute whether or not it is the right kind of methodological pluralism and I certainly have my own concerns. I've certainly always been keen to promote the value of methodological pluralism. I think it's important to remain open to the strengths (and weaknesses!) that any method brings. But it is not merely a question of using multiple methods, we also need to ask sensible kinds of questions about the data in the first place. The other related issue to this though, crudely speaking, is that, across the disciplines and different data driven enterprises, the 'default' methodologically is – whether one or more approaches is used – is to assume that linearity works for the most part and that we can tweak the rest. In other words, variable based analyses which depend on correlations and forms of linear modelling can basically do a good enough job. I am not totally against those methods and indeed have been teaching them to undergraduate and postgraduate students my entire career. (It is not for nothing that over the past two years I have been spending most of my time on Q-Step related activities, which is one of many of the UK's initiative for increasing quantitative training precisely as part of the preparation for big data futures.) But my own position, as I've hinted at in various ways above already, is that the empirical challenge of modelling the social sensibly lies not in whether or not one or more methods are used, but rather the causal assumptions inscribed in the methods and methodologies that we use. After all, if the same modes of change and continuity are built into and assumed within big data analytics are also the more or less the same as those that have helped to create some of the 'wicked problems' mentioned earlier, then why would we expect that big data is going to transform anything?

It is important to appreciate that much of methodological drive to do with big data is coming from commercial entities, which quite rightly are interested in targeting their products and services to *most* of their clients and customers. However, for social policy and planning purposes, especially around issues to do with social divisions of various kinds, we not only need to know what is *mostly* happening, we also need to examine the 'odd cases', the outliers, the different minority trends, and so on. The methodological practices are related, but the focus of interpretation, particularly in relation to understanding causality, is often different. From the point of view of developing sensible approaches for the purpose of improving social divisions, we need new configurations of methods and alternative interdisciplinary methodologies, which not only recognise the diversity of the causal assumptions inscribed within different methods and methodologies, as well as the different ways that methods and methodologies themselves help to re-produce different kinds of descriptions and causal explanations. So methodological plurality will get us some way, but not far enough!

References

Shaw, M. (2007) *What is Genocide?* Cambridge: Polity Press.

This interview is part of an ongoing series on the Philosophy of Data Science. Previous interviews in the series: Rob Kitchin, Evelyn Ruppert, Deborah Lupton, Susan Halford, Noortje Marres, and Sabina Leonelli.

Note: This article gives the views of the author, 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

Dr Emma Uprichard *is Associate Professor and Deputy Director at the Centre for Interdisciplinary Methodologies and Co-Director of the Warwick Q-Step Centre, at the University of Warwick. She has recently completed the ESRC project 'Food Matters', which explored food hates and avoidances through the life course and is currently on the*

ESRC Seminar Series on 'Complexity and Methods in the Social Sciences: An interdisciplinary approach'.

About the Interviewer

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](#)*.*