Focus: Big Data, Little Questions?
By Emma Uprichard, University of Warwick

Big data. Little data. Deep data. Surface data. Noisy, unstructured data. Big. The world of data has gone from being analogue and digital, qualitative and quantitative, transactional and a by-product, to, simply, BIG. It is as if we couldn't quite deal with its omnipotence and just ran out of adjectives. BIG. With all the data power it is supposedly meant to entail, one might have thought that a slightly better descriptive term might have been latched onto. But, no. BIG. Just BIG.

For those who may have missed the data obsessed world, 'big data' is causing a bit of storm. To be fair, it is more a future storm, with organisations, public and private firms and governments preparing for all that it will bring. Some say big data is already here and always has been, since we have always had more data than we know what to do with. Others say it is unlike anything that has been before because its v-dimensionality makes it different, new and powerful. The four big Vs are: Velocity, because it's 'live' and coming in all the time, e.g. Twitter, Flickr; Variety, because there's so many different kinds, from images (YouTube), to text (blogs), to numbers (transactions, automated logs); Veracity, because it's uncertain or imprecise and we don't always know what's there; and Volume, because there's so much of it, it's big. In a nutshell: high volume, high variety and high velocity. To this list, big data has also been discussed in relation to its clear versatility, volatility, virtuosity, vitality, visionary, vigour, viability, vibrancy, and even virility. (The letter 'v' must have increased its value due the recent hysteria related to it.)

Other less enthusiastic versions of this v-dimensionality might be that big data is also: valueless, vampire-like, venomous, vulgar, violating and very violent. I am not a fan of the term. It is too full of commercial management-speak for my liking. It misses what is important to me as a social researcher, which is about making more visible where the power networks lie - pun intended; detecting misinformation is just as important as being able to make sense of the information we have at hand. Rather annoyingly, it is being presented as the future 'problem solver' for all things, from breast cancer to low cost governance, from better security to predictive systems, from smart cities and better traffic and water systems, to an end to urban squalor. It is full of as much promise as it is warning: the promise for better societies, but unless we are fully prepared, our societal progress is doomed; the promise of better health systems, but unless we can cope with the big live digital imaging, people are going to die, because we are not going to be able to detect illness accurately. It promises cleaner, more sustainable renewable energy; better banking; better governance; better education for all; more efficient, faster, more cost effective everything. The promises and warnings go on. If we thought utopian authors were full of hope for better futures, the public discourse of big data is akin to a future Fairy God-Mother with a magic wand, granting wishes to help solve some of the most 'wicked problems'.

Factual – Measured - Critical

What's the big deal? Well, more and more data will be automatically collected and generated through everyday interaction. So much more of everything will be simultaneously data-producing and data-driven. Therefore, individuals, organisations, institutions and governments will be able to be turn to data to help answer their respective questions. One of the key ideas is that machine learning and pattern detecting data mining will increasingly help sieve through this humungous ever growing amount of data. The production of data, as well as these live data exploration techniques, supposedly create less bias, more objective analysis, more systematic data driven problem solving. Social problems can be solved more 'scientifically' – or so the story goes.

Let me very clear. I am not 'against' big data per se. I understand that the world is changing and we are generating more and more data the more synced up and digital our everyday lives become. I understand we are all automatically and mundanely plugged into a cyborg-grid of networked data points, which recursively make the world and shape the choices we have before us. I actually like data. I believe we should turn to data as much as possible to answer social questions. But we need to be careful about what is beginning to happen and big data social science we want to be part of.

Big data cannot deal with big questions

Let's face it, big data is not going to solve our big social problems, such as global warming, violence, genocide, war, social divisions, sexism, racism, disability, homophobia, water and food security, homelessness, global poverty, health and educational inequality, infant mortality, care for the elderly, and so on. It may help to describe them, to picture them in new ways, to visualise the available data differently, and this may help to communicate the problems to more people. This is certainly an important part to causing change. However, there are significant difficulties in using it to be able to tell us anything especially interesting.

Ironically, a key weakness of big data lies in its scale – scale in terms of the unit of analysis and scale in terms of time. In terms of the unit of analysis, big data can be excellent at being able to say what you are doing and/or what everyone is doing, but it is going to be very difficult to get it to say much else. So, for example, it can tell us what you are doing on the Twitter, or what you search for on Google, how you shop at Tesco's or how you use water, electricity, credit cards, and other such thing; it can also say these things for everyone, together, aggregately, overall. It will also be able to say how you are networked to your friends or contacts. It will also be able to pin point where big clumps of missing data are. (This is partly why the security and crime organisations are very interested in big data: if you are very different to everyone else, it may be possible to 'find' you amidst all the data.)

But big data is not – at least at the moment – very good at telling us what you or anyone will do. In fact, it will not tell us much about anything about what to do, what decisions are needed to make things different in the future – or even how to keep

things the same in the future; it doesn't seem as though it is going to be able to make very good predictions about the medium to long-term future either, and even short term futures can only ever be uncertain. Big data won't be able to tell us how to design local, regional and global policies and it will certainly not be able to do what we need policies of all kinds to do: be appropriate for some people sometimes and in some places. Social systems are not well modelled or known through universal laws. Social systems tend to be too dynamic for that kind of modelling, not least because we are reflexive beings and will remember things we don't even know we'll remember, and we react to the very models we use to model ourselves. So there are real limits to what we can do with big data.

What big data will be good at – and is already very good at – is enabling us to capture a snap shot of 'now'. Like the old Polaroid pictures, we will get excited because the image that is produced relatively instantaneously, and like the days when laser photography was new, the scale of the capture will be mesmerizing and we want to see how much we can see, how much more we can describe differently. These things are important and they are necessary. Anything that helps us to see the world a bit differently is interesting in my view as it can potentially help to nurture a healthy 'sociological imagination'. But the frame will remain on the relative present – the 'plastic present' to use a phrase I've used before - and that is unlikely to be enough to help us address the big social problems in the world today and make any substantive changes to them tomorrow. As Heffernan (2013) recently put it, 'Big data will never give you big ideas… Big data doesn't facilitate big leaps of the imagination. It will never conjure up a PC revolution or any kind of paradigm shift. And while it might tell you what to aim for, it can't tell you how to get there.'

Big data as methodological genocide?

As Wallerstein's (2000) 'Racist Albatross' explains so well, the social sciences have always been caught in the middle between the Sciences and the Humanities, and have been torn about by torn apart by the Methodenstreit, the epistemological debates about concerned with how they seek to do 'objective' and/or 'good' research. In many ways, British Sociology and even Political Science, has tended to develop more into the humanist camp, with qualitative methods and social theory winning out. This isn't necessarily a problem in itself, at least not yet, but it definitely will be where big data are concerned. After all, most big data is and will continue to be social data.

At the moment, the physical, engineering, computational, and mathematical sciences tend to be leading the way in terms of big data analysis, mainly because they are among the few to have the analytical skills to do so. But just as I said to my former colleague, Noortje Marres, in conversation in a bit of an outcry: 'Just because they are looking at social data, doesn't mean that what they are doing is social!' We are all, whether we like it or not, slowly but surely, becoming complicit to a deeply positivist, reductionist kind of social science, where variables are the be all and end all, where causality is devoid of meaning, and where non social scientists are the

ones ruling the roost in terms of access, collection and analysis – of big data, which is social data.

At the risk of sounding a bit melodramatic, the big data hype is generating, for want of a better term, a methodological genocide. To my mind, it even has a flavour of being a disciplinary genocide. It is fierce and it is violent, and social scientists – and especially sociologists – need to fight back. Certainly, if we are going to meaningfully interrogate the social systems and structures that make up the social world, we will need to improve our quantitative skills. I know, I'm sorry to say it, I know this doesn't always go down well among many social scientists, especially among sociologists in the UK. But whilst I do think that one of the ways we will need to fight back is to increase our quantitative skills – we need to be clear about the kind of social science we move forward to.

After all, increasing quantitative skills doesn't just mean increasing our statistical skills. We need good philosophers of science and social science too. We absolutely still need excellent social theory about what the data represent and we also need excellent qualitative methods to reinterpret and rethink the units of analysis we are observing. We need to be able to challenge what is being done with our data and that requires a basic understanding about how variables are created, how codes are made, and how these are being constantly used, modelled and reworked into everyday life. We need to think about what it means to measure the social world and how our models of causality are constructed. Importantly, we also need to know who is doing the counting. Who is making the decisions? Who is deciding what is counted and measured and how these counts and measurements are used and for whom? These answers are not trivial and social scientists need to be part of those conversations.

Many new statistical techniques used to crunch through big data involve 'shrinking' the data. This not only 'dilutes' the importance of extreme cases – the outliers – within large datasets, but also focuses the analysis on the masses in the middle. One of the key strengths of social research and sociological research in particular is a sensibility to social divisions, minority groups, oppressed and silenced voices. In order to remain strong in these areas, we must absolutely remain attentive to the methodological techniques that go some way to erase extreme cases, pockets of extreme difference. Another big way of organising data is through data mining, machine learning and pattern recognition. At the core of those approaches, there are issues such as classification – who or what goes into which group and how are units of analysis measured as 'similar' or 'different'? How should we count in a way that allows for meaningful counts over time? How we shape the social through our counting and classifying are highly political and ethical issues.

Social scientists know how to deal with data that is too big to handle!

Social scientists are not powerless by any means. The concept of data being too big to handle is far from new for most social scientists. Most are well accustomed to

having too much data and learn to live with that horrible overwhelming feeling that we get during most empirical projects at just how much data we have to organise, synthesise and make sense of. It is what we get trained very early on to do, because we always have too much data. Qualitative researchers in particular have important lessons to tell the big data world here. Having too much data to handle is the norm, as is having a lot of 'junk' we don't need, want or event know about until we get closer to it. It's just the way social data is. Theoretical sampling and analysing to the point of theoretical saturation, which are core to a qualitative researcher's general repertoire, are excellent ways of dealing with too much data. Likewise, having too much data is a taken for granted *a priori* position by digital methods researchers.

Quantitative researchers too tend to have too many datasets to explore, too many variables to choose from and yet rarely the variables they want or need for the questions they are interested. And those involved in simulation approaches such as multi agent based simulation know only too well the challenge of simplifying complex interactions down to simple rules. Social scientists have a range of important tools and techniques, theories and sampling techniques for dealing with data that is too bid and messy to handle. We need to find a way of voicing our capacity to deal with big data. We can afford to be more confident in our ability to have something important to say here.

Indeed, qualitative skills are highly valuable and in a world of big data, they may need to be shaken up a bit, reawaken, made stronger, so that we can capitalise on their strengths. What we can measure may certainly help us to know more about certain aspects of the social world, but we must not make the mistake of conflating data with the world it represents, models or is produced by. Of course, there will always be recursivity between models and what is modelled, what is measured and processes of measuring. We need to measure and learn to model and have a voice in the big data debates. But we must not make the mistake of assuming that the bigger the dataset, the bigger the sample, the better we will know the world. Tukey (1997:21) was right when he pointed out that, 'no data set is large enough to provide complete information about how it should be analysed!' I find it remarkable that we describe the world's most 'wicked problems' and we are then surprised that that we fail to make any substantive changes in the world, even though we have also tended to turn to the same data, use somewhat similar variables, analysing them using mostly similar methods - all the same things that went into creating those problems in the first place!

If we take C. Wright Mills' quest for a 'sociological imagination' seriously, then ideally we need to also turn to big data to help us think differently, to see differently and re-en/act the world differently. So much social theory has gone into arguing and discussing these very issues and we cannot afford to let big data run away without good social theories about what to do with the masses of data we are producing. Bourdieu (1990:64) warned us about the limits of change when we become complicit to our 'structuring structures' that tend to make us 'cut our coats according to our

cloth', and so we become 'the accomplices of the processes that tend to make the probable a reality'. If we are creating a mess by generating so many haystacks of big data that we are losing all the needles, then we need to figure out a different kind of way of doing things, as we cannot sew new cloth without any needles. Whatever else we make of the 'big data' hype, it cannot and must not be the path we take to answer all our big global problems. On the contrary, it is great for small questions, but may not so good for big social questions. Social scientists need to find a way not to be complicit in the new wave of Methodenstreit that is intrinsic to what big data brings.

References

Bhaskar, R. (1979) The Possibility of Naturalism : A Philosophical Critique of the Contemporary Human Sciences. Brighton: Harvester.

Bourdieu, P. (1990) The Logic of Practice. Stanford, CA: Stanford University Press.

Heffernan, M. (2013) 'Big data, big risk', Moneywatch, July 18, 8:39 AM, http://www.cbsnews.com/8301-505125_162-57593647/big-data-big-risk/, accessed July 19 2013.

Tukey, J.W. (1997) More Honest Foundations for Data Analysis. Journal of Statistical Planning and Inference, 57:21-28.

Wallerstein, I. (2000) 'The Racist Albatross: Social Science, Jörg Haider, and Widerstand', Lecture at the Universität Wien, Mar. 9. http://www.iwallerstein.com/the-racist-albatros/

Dr Emma Uprichard is a member of the Centre for Interdisciplinary Research at the University of Warwick. She has a longstanding interest in the methodological challenge of applying complexity theory in social science. She is especially concerned with issues of time and temporality and the ways in which different scales of time impact on change and continuity in the world.