Dialogues in Human Geography

Commentary

Big data and human geography: Opportunities, challenges and risks

Dialogues in Human Geography
3(3) 262–267

DOI: 10.1177/2043820613513388
dhg.sagepub.com
SAGE

Rob Kitchin
National University of Ireland, Ireland

Abstract

We are entering an era of big data – data sets that are characterised by high volume, velocity, variety, exhaustivity, resolution and indexicality, relationality and flexibility. Much of these data are spatially and temporally referenced and offer many possibilities for enhancing geographical understanding, including for post-positivist scholars. Big data also, however, poses a number of challenges and risks to geographic scholarship and raises a number of taxing epistemological, methodological and ethical questions. Geographers need to grasp the opportunities whilst at the same time tackling the challenges, ameliorating the risks and thinking critically about big data as well as conducting big data studies. Failing to do so could be quite costly as the discipline gets left behind as others leverage insights from the growing data deluge.

Keywords

big data, data deluge, human geography, methodology, praxis, theory

Like many terms used to refer to the rapidly evolving use of technologies and practices, there is no agreed definition of big data. A survey of the emerging literature, however, denotes a number of key features (Boyd and Crawford, 2012; Dodge and Kitchin, 2005; Laney, 2001; Marz and Warren, 2012; Mayer-Schonberger and Cukier, 2013; Zikopoulos et al., 2012) – big data are:

- huge in *volume*, consisting of terabytes or petabytes of data;
- high in *velocity*, being created in or near real time;
- diverse in *variety*, being structured and unstructured in nature;
- *exhaustive* in scope, striving to capture entire populations or systems;
- 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 data sets and
- *flexible*, holding the traits of extensionality (can add new fields easily) and scalability (can expand in size rapidly).

Given the drive to digitize and scale traditional small data into digital archives that are voluminous and varied, it is velocity and these additional characteristics that set big data apart from other data repositories and infrastructures.

Sources of big data can be broadly divided into three categories: directed, automated and volunteered. Directed data are generated by digital forms

Corresponding author:
Rob Kitchin, National University of Ireland, Maynooth, Ireland.
Email: rob.kitchin@nuim.ie

of surveillance, wherein the gaze of the technology is focussed on a person or place by a human operator. Automated data are generated as an inherent, automatic function of the device or system and include traces from digital devices, such as smart phones that record and communicate the history of their own use; transactions and interactions across digital networks; clickstream data that records how people navigate through a website or an app; sensed data generated by a variety of sensors and actuators (that measure levels of light, humidity, temperature, gas, electrical resistivity, acoustics, air pressure, movement, speed, etc) embedded into objects or environments; scanning of machine-readable objects such as travel passes (e.g. Oyster card on the London underground); passports or barcodes on parcels that register payment and movement through a system; machine-to-machine interactions across the Internet of things and capture systems, in which the means of performing a task captures data about that task. In contrast, volunteered data are gifted by users and include interactions across social media and the crowdsourcing of data wherein users generate data and then contribute them to a common system, such as OpenStreetMap.

These three forms of data generation now mean that more data are being produced every 2 days than in all of history prior to 2003 (Varian, cited in Smolan and Erwitt, 2012). In less than a decade, we have moved from a situation of seeking to understand the world with scarce and costly data to an overabundance of data, albeit not without issues of access, quality and scope. So what does this mean for geographic scholarship?

Opportunities

To date, those in the knowledge business have been operating in data deserts, seeking to extract information and draw conclusions from a small numbers of observations (Miller, 2010). This is particularly the case in the social sciences and humanities, where studies might comprise a fairly small number of interviews or surveys, or a handful of ethnographies or case studies. Where larger data sets have been produced, such as national censuses, the number of variables is limited, the data are generated infrequently and the data are released at a generally coarse spatial scale.

Big data holds the promise of a data deluge – of rich, detailed, interrelated, timely and low-cost data – that can provide much more sophisticated, wider scale, finer grained understandings of societies and the world we live in. It offers the possibility of shifting from data-scarce to data-rich studies; static snapshots to dynamic unfoldings; coarse aggregations to high resolutions; relatively simple hypotheses and models to more complex, sophisticated simulations and theories. For positivistic social scientists, big data offers the potential for a new era of computational social science; a new paradigm of 'data-driven science' that challenges established epistemologies through its blending of abductive, inductive and deductive approaches (Batty et al., 2012; Lazer et al 2009; Miller, 2010). For post-positivist scholars, the era of big data opens up a massive amount of unstructured data, much of it new (e.g. social media) and many of which have heretofore been difficult to access (e.g. millions of books, documents, newspapers, photographs, art works, material objects etc, from across history), along with new tools of data curation, management and analysis. Given that much of these big data have spatial attributes, they represent a huge potential opportunity for human geography. On the one hand, they provide an abundant source of data for geographical analysis and, on the other hand, they offer the possibility to promote the value of geography scholarship to a wider audience.

Imagine, for example, the human geography and broader social science research that could be undertaken with the data set put together by President Obama's team for his 2008 and 2012 election campaigns. This included hundreds of randomized, large-scale polling experiments, cookies that tracked visitors to their websites and data assembled from a variety of sources including registration data, census and other government data, commercial data aggregators, credit ratings agencies, cable television companies and social media sites (Issenberg, 2012). The result was a set of interrelated, massive databases about every voter in the country consisting of a minimum of 80 variables (Crovitz, 2012), and often many more, relating to a potential voter's

demographic characteristics and location, their voting history, their social and economic history, their patterns of behaviour and consumption and expressed views and opinions. Obama's data, admittedly assembled at a cost of over US$100 million, would provide an incredibly rich social, political and economic insight into the US society if made available for analysis.

Challenges

Whilst big data offers many opportunities, it also poses a number of challenges for human geography. At a fundamental level is the need for new methods of handling and analysing data sets that consist of millions or billions of observations that are being generated on a dynamic basis in a variety of forms (Batty et al., 2012). Traditional statistical methods are designed with respect to data-scarce science; to identify significant relationships from small, clean sample sizes with known properties. Whilst there has been much recent progress in devising new data analytics that can make sense of massive data sets, new forms of data science are in their infancy. In human geography, it is fair to say we are largely underprepared for the era of big data beyond a handful of scholars and centres. Certainly our undergraduate methods courses are hopelessly out of date with respect to both structured and unstructured data handling and analysis. A browse of methods textbooks suggests that little has changed in general methods training in the discipline since the early 1990s and the wide-scale roll-out of geographic information system (GIS). The data revolution underway demands a wider appreciation of the variety of emerging data sources and types, and a wider set of skills, including those being developed in the digital humanities, as well as basic coding, modelling and simulation (DeLyser and Sui, 2013a, 2013b; Sui and DeLyser, 2012).

As Floridi (2012) notes, big data raises fundamental epistemological questions about the organisation and practices of science: certainly, coping and extracting useful, valid information from the data deluge and making sense of it is not simply a technical issue that can be dealt with by technological solutions alone. Rather it requires careful rethinking with respect to the philosophy of science (Leonelli, 2012). Such a rethinking will undoubtedly have to take place within human geography and will do so in the context of the rise of new forms of empiricist and positivist thinking outside of the social sciences and their encroachment in examining social, political and spatial issues.

With respect to the latter, it is clear that academics from across a broad range of fields and disciplines are engaging with big data to make pronouncements about geographical processes and phenomena. This is driven by a certain level of naivety that big data can speak for themselves and does not require contextual or domain-specific knowledge with regard to analysis and interpretation (Anderson, 2008; Strasser, 2012) but also opportunism to capture new lines of research funding. For example, the emerging field of social physics, where physicists, computer and data scientists and others make pronouncements about social and spatial processes based on big data analysis, especially relating to cities and the supposed laws that underpin their formation and functions, often wilfully ignores a couple of centuries of social science scholarship, including rich traditions of urban quantitative analysis and model building (e.g. Bettencourt et al., 2007; Lehrer, 2010; Lohr, 2013). The result is an analysis of cities that are reductionist and fails to take account of the effects of culture, politics, policy, governance and capital, as well as a rich tradition of work that has sought to understand how cities function socially, culturally, politically and economically. The challenge for human geography is to push back against such naive forms of predatory science and, at the same time, to be able to compete in the same channels for research funding.

It is somewhat of a paradox that despite the emerging data deluge, access to such data is highly limited. This is because big data is mainly generated by privately owned businesses and government. Unless access can be negotiated to such data, geographers will be cut-off from a rich seam of potential studies. Where access is gained, there are a number of ethical and security challenges of working with such data, given that they are highly resolute, providing fine-grained detail on people's everyday lives. We are only just starting to think

through the various issues related to dataveillance more broadly (Dodge and Kitchin, 2007; Andrejevic, 2007) and have barely begun to consider in detail the big data ethical implications of our own scholarship, including our working relationships with business and government.

Risks

Given these challenges, there are a number of risks to human geography stemming from the rise of big data. The first is that the discipline is presently ill-prepared to embrace both methodologically and theoretically the era of big data. Indeed, the social sciences more broadly have been remarkably slow to react, with other fields filling the void. The risk is that it will be difficult to recover ground with respect to these fields, especially after they have benefitted from the significant allocation of new research funds and the support of policymakers. At the same time, we need to guard against two other risks to the integrity of geographical scholarship: the rise of empiricism and psuedo-positivism and the marginalisation of small data studies.

One disturbing development with big data and associated analytics has been the argument that with enough volume data can speak for themselves. Such empiricist thinking is illustrated through the claims of Chris Anderson (2008) who argues that big data signals 'the end of theory' with 'the data deluge mak[ing] the scientific method obsolete'. He contends that:

> Petabytes allow us to say: 'Correlation is enough'. We can stop looking for models. 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.

Some data analytics software is sold on precisely this notion. For example, the data mining and visualization software, Ayasdi, claims to be able to 'automatically discover insights – regardless of complexity – *without asking questions*. Ayasdi's customers can finally learn the answers to questions that they did not know to ask in the first place' (Clark, 2013; my emphasis).

There is a powerful and attractive set of ideas at work in these arguments. First, that big data can capture a whole of a domain and provide full resolution. Second, that there is no need for a priori theory, models or hypotheses. Third, that through the application of agnostic data analytics, the data can speak for themselves free of human bias or framing and that any patterns and relationships within big data are inherently meaningful and truthful. Fourth, that meaning transcends context or domain-specific knowledge. This has been accompanied by pseudo forms of positivism that are little more than data dredging riddled with the associated ecological fallacies. Such thinking seems to have gained purchase in business circles, and, to some degree, parts of data science, though it has also been resisted by those seeking to adapt the scientific method. Having gone through a tumultuous set of debates to shift human geography from empiricist forms of regionalism to a more scientific footing and post-positivist approaches, it would be a mistake to let such empiricist approaches drift back into the discipline.

The hype surrounding big data suggests that it is superior to traditional 'small data' studies due to its inherent characteristics. The danger is that funders and policymakers shift their emphasis to big data at the expense of small data, marginalising the worth of small data studies. Such a move misunderstands both the nature of big data and the value of small data. Big data may seek to be exhaustive, but as with all data they are both a representation and a sample. What data are captured is shaped by the technology used, the context in which data are generated and the data ontology employed. The world is vastly complex, and it is impossible to capture a whole domain and all of its nuances, contradictions and paradoxes. Big data generally captures what is easy to ensnare – data that are openly expressed (what is typed, swiped, scanned, sensed etc; people's actions and behaviours; the movement of things), which it takes at face value. It is much weaker at capturing complex emotions, values, beliefs and opinions; the varied, contextual, rational and irrational ways in which people interact and make sense of the world. In

general, these can only be inferred from big data, and they require a different set of research tools to be more fully explored. Small data studies can be much more finely tailored to answering specific research questions.

Conclusion

Big data are undoubtedly going to become a key part of geographic scholarship. Such data present a number of opportunities for human geography analysis offering rich insights into our social and spatial world. Big data also, however, poses a number of challenges and risks. Traditional geographical methods are designed for small and scarce data, and there is a need to engage with and develop new forms of data handling and analysis. This is going to require a redesigning of methods courses fit for a new era – drawing on both the digital humanities and the computational social sciences. Other fields and disciplines are increasingly going to undertake spatial analysis, looking to set the research agenda, compete for research funding and dominate access to various kinds of data and the policy space. In so doing, they are going to promote particular ways of undertaking social science, in many instances empiricist and pseudo-positivist in nature, and geographical analyses that are naive and reductionist. At the same time, we need to guard against small data studies being marginalised by funders and policymakers. Big data poses a number of epistemological, methodological and ethical questions. Given the philosophical debates within human geography over the past 60 years, the discipline is well positioned to engage in such reflection and to plot a path forward that draws on thinking in critical and qualitative GIS, radical statistics and mixed-method approaches. As we enter the age of big data, it is clear that we need critical reflection and research about big data as well as studies using big data.

References

Anderson C (2008) The end of theory: the data deluge makes the scientific method obsolete. *Wired*, June 23, 2008, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 12 October 2012).

Andrejevic M (2007) *iSpy: Surveillance and Power in the Interactive Era*. Lawrence, KS: University of Kansas.

Batty M, Axhausen KW, Giannotti F, Pozdnoukhov A, Bazzani A, Wachowicz M, et al. (2012) Smart cities of the future. *European Physical Journal Special Topics* 214: 481–518.

Bettencourt LMA, Lobo J, Helbing D, Kühnert C and West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences* 104(17): 7301–7306.

Boyd D and Crawford K (2012) Critical questions for big data. *Information, Communication and Society* 15(5): 662–679.

Clark L (2013) No questions asked: big data firm maps solutions without human input. *Wired*, 16 January 2013, http://www.wired.co.uk/news/archive/2013-01/16/ayasdi-big-data-launch (accessed 28 January 2013).

Crovitz LG (2012) Crovitz: Obama's 'big data' victory. *Wall Street Journal*, 18 November http://online.wsj.com/article/SB10001424127887323353204578126671124151266.html (accessed 19 November 2012).

DeLyser D and Sui D (2013a) Crossing the qualitative-quantitative divide II: inventive approaches to big data, mobile methods, and rhythmanalysis *Progress in Human Geography* 37 (2): 293–305.

DeLyser D and Sui D (2013b) Crossing the qualitative-quantitative chasm III: enduring methods, open geography, participatory research, and the fourth paradigm. *Progress in Human Geography*. Epub ahead of print 19 March 2013. DOI: 10.1177/0309132513479291

Dodge M and Kitchin R (2005) Codes of life: identification codes and the machine-readable world. *Environment and Planning D: Society and Space* 23(6): 851–881.

Dodge M and Kitchin R (2007) Outlines of a world coming in existence: pervasive computing and the ethics of forgetting. *Environment and Planning B* 34(3): 431–445.

Floridi L (2012) Big data and their epistemological challenge. *Philosophy and Technology* 25(4): 435–437.

Issenberg S (2012) *The Victory Lab: The Secret Science of Winning Campaigns*. New York, NY: Crown.

Laney D (2001) 3D data management: controlling data volume, velocity and variety. *META Group*, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed 16 January 2013).

Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, et al. (2009) Computational social science. *Science* 323: 721–733.

Lehrer J (2010) A physicist solves the city. *New York Times*, 17 December 2010, http://www.nytimes.com/2010/12/19/magazine/19Urban_West-t.html?pagewanted=all&_r=0 (accessed 1 April 2013).

Leonelli S (2012) Introduction: making sense of data-driven research in the biological and biomedical sciences. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 1–3.

Lohr S (2013) SimCity, for real: measuring an untidy metropolis. *New York Times*, 23 February 2013, http://www.nytimes.com/2013/02/24/technology/nyu-center-develops-a-science-of-cities.html?pagewanted=all (accessed 1 April 2013).

Marz N and Warren J. (2012) *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. MEAP edition. Westhampton, NJ: Manning.

Mayer-Schonberger V and Cukier K (2013) *Big Data: A Revolution that will Change How We Live, Work and Think*. London, UK: John Murray.

Miller HJ (2010) The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science* 50(1): 181–201.

Smolan R and Erwitt J (2012) *The Human Face of Big Data*. New York, NY: Sterling.

Strasser BJ (2012) Data-driven sciences: from wonder cabinets to electronic databases. *Studies in History and Philosophy of Biological and Biomedical Sciences* 43: 85–87.

Sui D and DeLyser D (2012) Crossing the qualitative-quantitative chasm I: hybrid geographies, the spatial turn, and volunteered geographic information (VGI) *Progress in Human Geography* 36(1): 111–124.

Zikopoulos PC, Eaton C, deRoos D, Deutsch T and Lapis G (2012) *Understanding Big Data*. New York, NY: McGraw Hill.