EDITORIAL

Putting big data in its place: a *Regional Studies and Regional Science* perspective

Introduction

This second Editorial for *Regional Studies, Regional Science* is timed to coincide with the publication of Rob Kitchin et al.’s contribution and associated commentaries (Batty, 2014; Holden & Pires, 2014; Kitchin, Lauriault, & McArdle, 2014; Wilson, 2014). We focus on the topic of big data and aim to offer some perspective in relation to recent critiques and its utility within regional studies and regional science, whereas Kitchin et al. consider its epistemological connotations in the context of urban indicators, city benchmarking and real-time dashboards. Although this commentary focuses on big data within a social science context, much of what we write could be applied more widely. The next section considers definitions of big data before we turn our attention to infrastructure and investment, and conclude with some reflections on what the future might hold for big data in regional studies and regional science.

Putting the big data backlash in perspective

Big data is a topic that in recent years has been the source of much enthusiasm, hype and a fair amount of cynicism. Such reactions are not uncommon within an academy that is simultaneously curious and critical, but the response to the emergence of ‘big data’ and its associated proclamations leads us to conclude that the debate often lacks clarity, direction and reason – from proponents and critics alike.

It can be argued to lack clarity partly because of its interdisciplinarity and partly because advocates of big data are often too busy with analysis to consider the theoretical underpinnings of their research (Kitchin, 2014a). It often lacks direction because individuals and groups of researchers from different disciplines are struggling to locate ‘big data’ within wider shared historical and theoretical contexts. It also often lacks reason because critics too quickly dismiss big data as ‘just another fad’ and proponents do a poor job of answering key questions about big data. For example, issues such as ‘is big data actually new?’ ‘does big data mean the end of theory?’ (cf. Anderson, 2008), and ‘what can big data not do?’ have too often been sidestepped in favour of presenting the results of complex analytical procedures, often in visual form, and often without adequate exposition of the methods. We therefore think it is time to pause and reflect in order to consider how big data approaches might help advance knowledge in regional studies and regional science.

The critiques of big data alluded to above have been explored elsewhere at greater length (e.g. Kitchin, 2014b; Lazer, Kennedy, King, & Vespignani, 2014) so in this Editorial we wish to position the contribution in relation to the big data backlash that has emerged in recent months (e.g. *The Economist*, April 20, 2014) and which – if we are...
not careful – will undermine attempts by social scientists to capitalize upon the richness, variety and availability of vast new data sets with which we can, potentially, begin to understand better many of society’s grand challenges.

**What is big data? An indefinite definition**

In many ways, trying to define big data is a thankless task since instead of adding clarity such attempts typically muddy the waters (e.g. Ward & Barker, 2013). For computer scientists, most of what is being done under the rubric of ‘big data’ in the social sciences might be considered rather trivial, since it does not typically enter the realm of the petabyte (1 million gigabytes). In the humanities, the situation is more complex still, with the meaning of the term ‘data’ often being far from clear (Schöch, 2013). Diebold (2012) attempted to trace the origins of the term and understand its development, but most discussions of what big data actually is at some point refer to the 2001 Gartner definition of volume, velocity and variety, and often also add veracity as an additional facet (Laney, 2012). Diebold himself appears to have been the first to use these terms in an academic context (Diebold, 2003), although others are cited elsewhere at an earlier date (Laney, 2001). Added to this confusion is the commonly held view that big data is essentially any data that are too big to handle or analyse using existing tools. Combined with the interdisciplinary nature of the term, and the fact that it is used within academia, business and media, one quickly reaches the conclusion that big data cannot adequately be ascribed a fixed definition and trying to do so is often unhelpful.

Given this inherent fuzziness of big data definitions, we prefer to define ‘big data’ as a fluid, user-centred concept that emerges as a result of a relative imbalance between the data themselves and the constraints on collection, management and then synthesis by the analyst. In this user-centric view, big data might also be seen as arising from a problem of ‘small tools’, where the difficulty in analysis in one field (e.g. economics) might be solved by adopting the methods and tools from another discipline (e.g. computer science). We think that big data could therefore benefit from greater interdisciplinarity and pragmatism, so we take the view here that inherently interdisciplinary fields such as regional studies and regional science have much to offer big data, and much to gain by adopting big data methods. However, making such progress does of course require infrastructure and investment, and this is the subject of the next section.

**Infrastructure and investment for big data**

Handling and analysis of big data often requires significant investment in storage, transfer and processing architecture. Furthermore, the use of systems that have been designed to manage and query big data are often complex, which from the perspective of the social sciences may require additional supporting human resources with specialized skills that lie outside of traditional disciplines. Arguably, this may encourage interdisciplinarity, although such facilitation will also likely involve a cost, and this may become acute given that such technical support skills are in high demand but suffer from low supply (Harris, Shetterley, Alter, & Schnell, 2013). The required investments for big data analytics can therefore often remain out of reach of the lone scholar, and require funding streams larger than might be required to facilitate more traditional data analysis. In future this will likely create an inequity between well-funded and less well-funded research.
Some research agencies have, however, responded in an attempt to mitigate such issues. For example, in a UK context there have been a series of large-scale investments from the Economic and Social Research Council (ESRC) in centres supporting service and research into and using big data;\textsuperscript{1} although, internationally, such centralized initiatives are not visible across all jurisdictions.

A further hurdle to big data reaching its full potential is that availability of big data is often limited, leaving analyses only available to those groups of organizations or researchers with privileged access. In the context of social media content, for example, very few researchers would have access to a full database of live or archived content, perhaps instead having to work with small samples or restricted extracts. We therefore see one of the main tasks of proponents of big data is to encourage openness and transparency wherever possible so that the richness of data and the benefits it can bring can be shared more widely. If big data is to achieve anything near its full potential, then our view is that interdisciplinary investment and infrastructure must improve, but this will take time. If we expect paradigm-changing results from big data analyses in the short-term, we are likely to be disappointed.

The future of big data in regional studies and regional science

The syntheses of data into information are hallmarks of both regional studies and regional science research. Huge volumes of data are generated within regions daily, such as through the use or management of public services (e.g. global positioning satellite (GPS) tracking of law enforcement officials and use of healthcare facilities), or captured from transport-related activities (e.g. road flow information gathered by networks of traffic cameras). Other data are also generated by the private sector, including transaction data associated with consumption or the use of social media, where content are georeferenced by mobile devices. The ability to capture, store and link such temporally rich data within regions remain largely to date an unmet opportunity, and one which we implore the regional studies and regional science communities to meet head on.

A good reference point for such approaches is perhaps the numerous open data stores that have emerged within municipalities across the world, such as London’s DataStore, OpenData Paris, New York’s NYC Open Data platform and the Dublin data dashboard example in Rob Kitchin et al.’s paper. However, the geographic extent of these repositories are typically limited within city boundaries and do not necessarily encompass their wider metropolitan regions. The delineation of data collection and dissemination within official administrative boundaries makes their use for regional studies and regional science more problematic, and this represents another challenge that is yet to be met. In future, then, regional big data will have to overcome the problems of administrative boundaries that have beset many previous smaller scale studies. This will require a form of cross-jurisdictional working that reflects the underlying functional structure of cities and regions across the world, rather than more convenient political extents. To date, however, our contention is that the data dashboard approaches seen in many cities serve as excellent test cases for the utility of big data approaches across regions.

At a supra-national level, the European Union’s Cohesion Policy Open Data platform, while newer and less fully developed, offers an insight into the future of data dissemination in the European Union. Although connections between data collection regimes may be possible, it is likely that compatibility differences would emerge in relation to the format, nature and type of data disseminated, thus making comparison more complex and prone to caveat. Although there are numerous national data stores, these
typically enable bulk extracts, without the ability to refine for particular regional extents, and, furthermore, they often do not capture some of the richness of attributes of those more localized data stores. In addition to these issues, more generally there are few examples of data stores that provide access to ‘big data’, instead mainly favouring more aggregate and temporally less frequent data.

These fledgling big data examples serve both as reminders of the massive potential of big data for enlightening citizens and improving our understanding of critical urban and regional processes, but also that we need to think more carefully about the theoretical and epistemological aspects of big data if we are truly to advance knowledge and contribute to better public policy. With these points in mind, we offer three final reflections on big data from a regional studies and regional science perspective. We would encourage contributions to the journal that respond to any (or all) of these:

(1) We need to think more, not less, about big data. Contrary to Anderson’s provocation that big data could lead to the ‘end of theory’, we see a real need for analysts to pause for thought. Where Anderson stated that ‘Who knows why people do what they do? The point is they do it. […] With enough data, the numbers speak for themselves’, we instead see the opportunity to ask the ‘why’ question as well as the ‘what’ question. This would involve a combination of ‘small data’ and big data approaches of the kind recently advocated by Lazer et al. (2014)

(2) In relation to method, infrastructure and knowledge, we think there is a real need to increase interdisciplinary thinking and working. The methods of one discipline cognate to regional studies and regional science (e.g. economics) could greatly benefit from the insights, methods and thinking of another more distant one (e.g. computer science), but at present the barriers to such working persist, both from an institutional perspective and a research-funding one. If big data is to solve big problems then our contention is that we need to assemble large interdisciplinary teams that include input from a wide range of disciplines. This is an obvious point, but one too often ignored in the pursuit of the results-driven world of big data research.

(3) Finally, we think there is a problem with a lack of successful big data exemplars, in both the fields of regional studies and regional science, but also more widely. This is particularly the case in the social sciences, where big data research is still in its infancy and where critiques have been most trenchant. Answering some fundamental questions, such as ‘how has big data helped cities and regions?’ and ‘how can big data improve the lives of ordinary citizens?’, have too often been overlooked or, when they are asked, unsatisfactorily answered. We therefore look to the regional studies and regional science communities to help fill this gap and to continue the tradition of excellent big data research and critique offered by Kitchin et al. and others in this journal.

Note
1. See http://www.adrn.ac.uk/.

References
Anderson, C. (2008). The end of theory. Wired magazine, 16, 16 July.
Batty, M. (2014). A Perspective in City Dashboards. Regional Studies, Regional Science, doi:10.1080/21681376.2014.987540.
Diebold, F. X. (2003). ‘Big Data’ Dynamic factor models for macroeconomic measurement and forecasting. In M. Dewatripont, L. P. Hansen, & S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society* (pp. 115–122).

Diebold, F. (2012). *On the Origin (s) and Development of the Term “Big Data”*. Penn Institute for Economic Research. Pier Working Paper, 12–037 (pp. 1–8). Available from: http://economics.sas.upenn.edu/sites/economics.sas.upenn.edu/files/12-037.pdf

Harris, J. G., Shetterley, N., Alter, A. E., & Schnell, K. (2013). The Team Solution to the Data Scientist Shortage. Accenture; Retrieved September 8, 2014 from http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Team-Solution-Data-Scientist-Shortage.pdf.

Holden, M., & Pires, S. (2014). Commentary on Rob Kitchin et al.’s ‘Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science*, doi:10.1080/21681376.2014.987541.

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1–12, 2053951714528481.

Kitchin, R. (2014b). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

Kitchin, R., Lauriault, T., & McArdle, G. (2014). Knowing and governing cities through urban indicators, city benchmarking, and real-time dashboards. *Regional Studies, Regional Science*, doi:10.1080/21681376.2014.983149

Laney, D. (2001). *3D-Data Management Controlling Data Volume Velocity and Variety*. Application Delivery Strategies. Stamford: The Meta Group. Retrieved September 8, 2014 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Laney, D. (2012). The Importance of ‘Big Data’: A Definition. Gartner, Inc.

Lazer, D. M., Kennedy, R., & King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science, 343*(6176), 123–1205.

Schöch, C. (2013). Big? Smart? Clean? Messy? Data in the Humanities. *Journal of Digital Humanities, 2*, 2–13.

Ward, J. S., & Barker, A. (2013). Undefined By Data: A Survey of Big Data Definitions. *arXiv preprint arXiv:1309.5821*.

Wilson, M. (2014). Flashing lights in the quantified self-city-nation. *Regional Studies, Regional Science*, doi:10.1080/21681376.2014.987542

Alasdair Rae and Alex Singleton
*Regional Studies, Regional Science*
a.j.rae@sheffield.ac.uk
alex.singleton@liverpool.ac.uk