Article:
Lovelace, R (2016) Book Review: The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences, by Rob Kitchin. Journal of Regional Science, 56 (4). pp. 722-723. ISSN: 0022-4146

https://doi.org/10.1111/jors.12293

Review of “The Data Revolution: Big Data, Open Data, Data Infrastructures & their Consequences”, Rob Kitchen

By Robin Lovelace, University of Leeds

Created using open source software.

At the outset I must declare a vested interest in this review. I will soon begin a 5 year fellowship in 'Transport and Big Data' at the Leeds Institute for Transport Studies (ITS). I will be expected to have decent grasp of the 'data revolution' not only in terms of technology for handling large datasets but also in terms of the wider questions it raises. Like most people I am learning as I go along and this book greatly helped place the technical work in its wider context. The data revolution is undeniably driven by technology, meaning the discussion of ethics, for example, often takes a back seat, especially in the so-called field of 'data science' which is overwhelmingly male dominated and technocentric.

For the impatient reader: The Data Revolution does a great job of raising the wider issues associated with the dramatic and still accelerating 'datarisation' of life. As the book points out, 'small data' is not going to disappear and to some extent Big Data is just an extension of smaller datasets. However, as a disclaimer, this review is primarily undertaken in relation to its treatment of Big Data. It does not intend to provide a synopsis of each chapter. Why? The book has already been rightly well reviewed in general terms. Big Data is probably the fastest moving branch of the revolution with major implications for Regional Science. It is also an aspect of the data revolution that I have direct experience with, as lead developer of the Propensity to Cycle Tool, an online planning support tool has been deployed nationally. The review is inevitably shaped by this context.

A dataset's lifecycle can roughly be summarised by what is done to it, in the following order: collection → cleaning → processing → analysis → modelling → visualisation → policy relevance. My work focusses on these technical aspects of the brave new so-called 'Big Data' world. It's fair to say that I've only thought informally about the ethical consequences of the data driven research wave that I'm riding. The Data Revolution opened my eyes to a new world of literature and insight that lurks just below the surface of a topic that can seem dull and solely technical to outsiders. My worldview opening the book was that data is best understood narrowly as a raw material for generating new knowledge in the domain that most interests me: energy use in transport and how to reduce it, as one small part of a grand transition away from fossil fuels in all sectors of the economy (Smil 2008). The Data Revolution has challenged this utilitarian worldview, proving new vocabulary and concepts that will be of real world use to researchers and practitioners using Big Data, especially when communicating the methods and results to others.

In this context most people implicitly know that that data relates to power (and by extension that

Big Data implies great power – just look at Google), but lack the language to articulate the wider economic/political/ethical issues that working with nonstandard datasets entails. The first Commandment of Big Data (see below) is that you should only use the term if you know precisely what you mean by it. Kitchen is frustratingly relativist here. I would argue that people stepping (or more commonly stumbling) into the Data Revolution are better served by pithy and clear guidance rather than an overview of what different people has said in the literature over the years. In the same way that it is the science journalist's duty to tell us of the best evidence, rather than report all sides of the story (some of which muddy the water), I would argue that it should be the role of the scholar of data to provide normative guidance.

The etymology of Big Data is discussed, running through the usual 'three V's' as well as 'exhaustivity' (which Kitchen uses in the sense of covering everything rather than in the sense that Big Data is a biproduct emitted like gas from a car's exhaust), resolution, 'relationality' and flexibility. I was dissapointed to see little on the fact that Big Data is often a biproduct of another process – i.e. it is not hypothesis driven – an important limitation compared with conventional datasets.

Another approach, which is to define large datasets primarily as a technical challenge,f was also given only passing reference. From a Quantitative Geography perspective myself and colleagues defined Big Data in terms of the tools that are needed to make sense of them. We acknowledged that in the social sciences Big Data is used as a 'catch-all' phrase: Datasets “that are difficult to analyze using established methods” (Lovelace et al., 2016).

Another issue I have with the book is that it provides almost no guidance to a non technical reader explaining how to actually 'do' Big Data. This is a major oversight based on Kitchin's own insight that open data is only empowering to those who know how to use it. However, with the parallel revolution of open source software and the emergence of online communities and software products that make Data Science accessible to the masses such as RopenSci and Rstudio, everyday citizens can empower themselves to make sense of the data deluge. I would point people who are interested in processing large datasets towards a list of free Data Science resources here: http://datasciencemasters.org/

Despite these specific issues surrounding the discussion of Big Data (which should be understood in the context of my worldview, outlined above) the book overall is fantastic. Although Rob Kitchen has written entire papers on many of the topics covered, he does an excellent job of avoiding the academic's curse of going into too much detail at the expense of the Big Picture. And any issues one may have about the coverage of Big Data in the book will more than compensated for by the discussion of data overall. Big Data is after all simply an extension of 'small data' (which Kitchen argues cogently should be called 'capta' in a fascinating passage) so it is crucial to get to harness these ubiquitous condensed and abstracted chunks of information before trying to tame their comparatively monstrous Big brothers!

The Data Revolution is an elegantly written synthesis of the wider context in which data is becoming to the 21st Century what oil was to the 20th. Too often datasets are seen in purely technical terms. Although technology is crucial for actively participating in the Data Revolution (and I would have liked to have seen more about the open source communities such as those rapidly coalescing around R and Python) technology without direction is ammoral. In summary, The Data Revolution

can be used as a meta-guidebook for empowering oneself with the conceptual tools needed to understand and use the data deluge unfolding worldwide for the greater good.

Appendix: The 5 commandments of Big Data for the Greater Good (from the Obesity Network seminar, Oxford, 16th March, 2016). These are designed to stimulate debate and provide practical guidance to researchers newly immersed in the Big Data revolution:

1) thou shalt remember the purpose of thine research regardless of the size of thine dataset

2) thou shalt not spend excessive amounts of time making visualising big data for the sake of it (or social media clickbait)

3) thou shalt not do big data until thou has done 'small data' first

4) thou shalt not hide thine ideas behind complex terminology associated with the terms 'big data' or 'data science', the meaning of which has not been clearly identified.

5) if thou wants to be a data scientist thou must program ... "for documentation, sharing and scientific repeatability" ([mount 2016](http://www.r-bloggers.com/some-programming-language-theory-in-r/)).