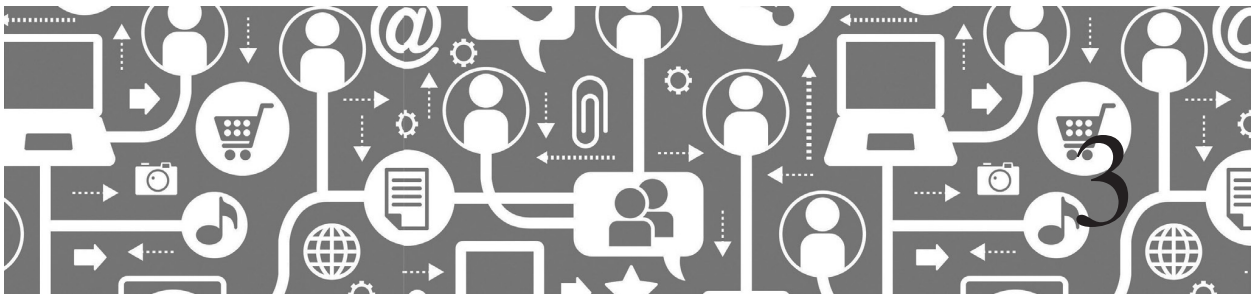

3

Big Data – Hype or Revolution?

Rob Kitchin

INTRODUCTION

The etymology of 'big data' can be traced to the mid-1990s, first used to refer to the handling and analysis of massive datasets (Diebold, 2012). Laney (2001) refined the definition to refer to data characterized by the now standard 3Vs, with big data being:

- huge in *volume*, consisting of terrabytes or petabytes of data;
- high in *velocity*, being created in or near real-time;
- diverse in *variety* in type, being structured and unstructured in nature.

More recently, these characteristics have been refined further to include:

- *exhaustive* in scope, striving to capture entire populations or systems (n=all);
- 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;
- *flexible*, holding the traits of extensionality (can add new fields easily) and scalability (can expand in size rapidly).

(boyd and Crawford, 2012; Dodge and Kitchin, 2005; Marz and Warren, 2012; Mayer-Schonberger and Cukier, 2013.)

Data holding all or most of these qualities have existed in a number of domains, such as remote sensing, weather forecasting, and financial markets, for some time. By the turn of the new millennium they were starting to become more common given the development and convergence of new technological developments such as ubiquitous computing, widespread internetworking, Web 2.0. and the creation of social media, No-SQL database designs and cloud storage solutions, and data analytics designed to cope with data abundance (Kitchin, 2014a). From 2008 onward the term started to gain traction, quickly rising up a hype cycle aided by a strong boosterist discourse that contended big data was set to revolutionize how business is conducted and

governance enacted. Not long after the term drifted into common academic use accompanied by an argument that big data would transform how research would be conducted.

This chapter examines the latter contention and the extent to which big data and its associated data analytic represents a genuine revolution with respect to how we make sense of the world or whether it has been over-hyped and is merely a new inclusion amongst a suite of options in the academy's research armoury. The chapter starts by detailing how big data differs from traditional datasets used by social scientists. It then examines the argument that it is leading to the creation of new research paradigms across disciplines, what have been termed data-driven science, computational social science and digital humanities. In particular, it focuses on the extent to which social media data, in combination with other big social data, offers the possibility for a different kind of social science.

BIG DATA AND NEW DATA ANALYTICS

There is some scepticism within the literature as to the extent to which big data is anything new. Critics, usually focusing on the characteristic of volume, suggest that we have long possessed very large datasets that have been challenging to process and analyze. In part this is a reaction to the term 'big' which tends to emphasize the volume aspect of the 3Vs. However, it is the total sum of the characteristics noted above, especially the qualities of velocity and exhaustivity (see Kitchin and McArdle, 2016, for an examination of the ontological characteristics of 26 datasets drawn from seven domain: mobile communication; websites; social media/crowdsouring; sensors; cameras/lasers; transaction process generated data; and administrative), that make the nature of big data differ from traditional data, or what might be termed 'small data' (see Table 3.1). The distinction is apparent if one compares a national census with a social media site such as Facebook.

Table 3.1: Comparing small and big data

Characteristic	*Small data*	*Big data*
Volume	Limited to large	Very large
Exhaustivity	Samples	Entire populations
Resolution and indexicality	Coarse & weak to tight & strong	Tight & strong
Relationality	Weak to strong	Strong
Velocity	Slow, freeze-framed	Fast
Variety	Limited to wide	Wide
Flexible and scalable	Low to middling	High

Source: Kitchin (2014a: 28)

While a national census has a large volume and attempts to be exhaustive (it seeks to sample all people resident in a country), it has very weak velocity (carried out once every ten years in most countries), weak variety (restricted to generally 30–40 highly structured questions), and no flexibility or scalability (once the census is initiated there is no opportunity to alter the questions or format). Moreover, while the raw data has high resolution and indexicality (individuals and households) it is released to researchers in an aggregated form. Other small data datasets are typically produced using a tightly controlled method using sampling techniques that limit their scope (non-exhaustive), temporality and size in order to produce high quality, representative data and make the process manageable and less costly. In contrast, Facebook has over a billion registered users globally and in 2014 was processing 10 billion messages (and associated comments and links), 4.5 billion 'Like' actions, and 350 million photo uploads *per day* (Marr, 2014). All that content and associated metadata is linked indexically to all individual users and through friending and tagging they are interlinked between users. Moreover, Facebook is a dynamic environment with the company constantly tweaking its platform and experimenting with different versions of its algorithms.

While the census is producing voluminous 'small data', Facebook is producing data that are qualitatively different in nature. In fact, Facebook is producing a data deluge – a constantly flowing torrent of rich, highly informative information about people, their lives, and what is happening in different societies in places around the world. The same is true of Twitter, Whatsapp, Snapchat, Foursquare and other social media platforms. When we compare Facebook to the data produced in most social science studies through surveys, political polls, interviews, or focus groups – where the number of respondents might be in the order of 10s or 100s and rarely exceeds 1000, the data are generated at a single point in time (usually over a couple of weeks or months), and are limited in variety – the difference becomes more stark. As detailed below, however, it should be noted that while the data produced within Facebook or Twitter is exhaustive, the data made available to researchers external to those companies might be sampled (though the sample generally consists of tens of thousands of records).

This kind of qualitative difference in the nature of data is happening across domains – health, education, work, consumption, finance, policing, public administration, science, etc. – in which new socio-technical systems are producing data through algorithmically-controlled and automated cameras, sensors, scanners, digital devices such as smart phones, clickstreams, and networked interactions such as online transactions (e.g., shopping) and communication (e.g., social media) (Kitchin, 2014a). For example, if we consider the developing areas of urban informatics a wealth of urban big data are being generated, much of it at the level of the individual: digital CCTV footage with facial/clothes recognition, automatic number plate recognition, sensor networks that track mobile phone unique signatures, and travel passes such as the London Oyster card (Kitchin, 2016). Other kinds of real-time data include the locations of buses and trains, how many bikes/spaces are in bike stands, road speeds on different segments, the number of spaces in car parks, general CCTV footage, air traffic, air quality, pollution readings, water levels, sound levels, current weather – all of which are increasingly becoming open in nature and underpin a diverse apps economy (e.g., see the Dublin Dashboard – http://www.dublindashboard.ie). To this we can add geo-referenced social media data (such as Twitter or Foursquare), crowdsourced data such OpenStreetMap, and live citizen city reporting (e.g., 311 services in the US and websites such as fixmystreet.ie), and citizen science data such as personal weather stations.

These data are systematic and continuous in operation and coverage, verifiable and replicable, timely and traceable over time, and relatively easy to visualize and to compare across locales through graphs/maps (though they are not straightforward to plug into modelling, profiling and simulations). They offer the potential to shift from 'data-scarce to data-rich analysis, static snapshots to dynamic unfoldings, coarse aggregations to high resolution, and relatively simple hypotheses and models to more complex, sophisticated theories' (Kitchin, 2013: 263). How we come to know and understand cities, and how we can govern and operate their various systems, then is being transformed through access to big data streams (Batty, 2013; Townsend, 2013; Kitchin, 2014b; Kitchin *et al.*, 2015). These big data also raise a whole series of ethical questions with respect to their use in dataveillance (surveillance through data records), social sorting (differential treatment to services), anticipatory governance (predictive profiling), control creep (data generated for one purpose being used for another) and the extent to which their systems make the city hackable, brittle and buggy (Townsend, 2013; Kitchin, 2014b, 2016).

Importantly, the development of the data deluge has been accompanied by the creation of new analytical methods suited to trying to extract insights from massive datasets using machine learning techniques, wherein the power of computational algorithms are used to process and analyze data. Again, there has been

much hype concerning these new data analytics for three reasons. First, until recently, data analysis techniques were designed to extract insights from scarce, static, clean and poorly relational datasets, that were scientifically sampled and adhere to strict assumptions (such as independence, stationarity, and normality), whereas new data analytics can cope with a deluge of variable quality data (Miller, 2010). Second, whereas data was traditionally generated with a specific question in mind, new data analytics can repurpose data, detect and mine patterns, and identify potential questions that the data might answer (Kelling *et al.*, 2009; Prensky, 2009). In other words, the hypotheses can be generated from the data. Third, an ensemble approach can be adopted in which, rather than selecting a single approach to analyze a phenomena, can apply hundreds of different algorithms to a dataset to determine the best explanatory model (Franks, 2012; Siegel, 2013). These new analytical techniques have been in development since the start of computing but have become significant area of recent research investment in order to increase the big data toolkit in four main areas: data mining and pattern recognition; data visualization and visual analytics; statistical analysis; and prediction, simulation, and optimization (National Science Foundation, 2012; Kitchin, 2014a). For many, big data and new data analytics will inevitably challenge dominant paradigms across the academy, ushering in new epistemologies in all disciplines and it is to this issue the chapter now turns.

A DATA REVOLUTION?

In Thomas Kuhn's (1962) well-known explanation as to how science periodically transforms from one dominant paradigm (an accepted way of interrogating the world and synthesizing knowledge) to another, an established body of knowledge is challenged and destabilized by a new set of ideas, eventually reaching a tipping point wherein the latter replaces the former. An example would be the shift from creationism to evolution, or Newtonian laws of physics to Einstein's theories of relativity. In Kuhn's account a paradigm shift occur because the dominant mode of science cannot account for particular phenomena or answer key questions. In contrast, Jim Gray (Hey *et al.*, 2009) proposed that the transitions between paradigms can also be founded on advances in data production and the development of new analytical methods. Underpinning this view is the observation that '[r]evolutions in science have often been preceded by revolutions in measurement' (Sinan Aral, cited in Cukier, 2010). Gray thus proposed that science was entering a fourth paradigm (exploratory science) based on the growing availability of big data and new analytics (his first paradigm was 'experimental science' that operated pre-Renaissance, the second was 'theoretical science' operating pre-computers, and the third was 'computational science' operating pre-big data) (Hey *et al.*, 2009).

The idea of academic paradigms has been subject to much critique, not least because within some disciplines there is little evidence of paradigms operating (notably some social sciences) and the idea tends to produce overly linear stories about how disciplines evolve, smoothing over the messy, contested and plural ways in which they unfold in practice. Nevertheless, the idea has utility here for considering whether the creation of big data has initiated a revolution in how academic research is being conducted. In particular, I explore three developments: (a) the notion that big data gives rise to the end of theory enabling a new form empiricism in which data can speak for themselves; (b) the creation of data-driven rather than knowledge-driven science; and (c) the formation of the digital humanities and computational social sciences.

The end of theory?

For Chris Anderson (2008), big data, new data analytics and ensemble approaches

signalled a new era of knowledge production characterized by 'the end of theory'. He argued that 'the data deluge makes the scientific method obsolete', with the patterns and relationships contained within big data inherently producing meaningful and insightful knowledge about phenomena. He continued:

> 'There is now a better way. Petabytes allow us to say: "Correlation is enough." … 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. There's no reason to cling to our old ways.'

Similarly, Prensky (2009) contends: 'scientists no longer have to make educated guesses, construct hypotheses and models, and test them with data-based experiments and examples. Instead, they can mine the complete set of data for patterns that reveal effects, producing scientific conclusions *without* further experimentation.' Dyche (2012) thus states that 'mining big data reveals relationships and patterns that we didn't even know to look for'. Dyche's example is a retail chain which analyzed 12 years' worth of purchase transactions for possible unnoticed relationships between products. Discovering correlations between certain items in shoppers' baskets led to new product placements and a 16 percent increase in revenue in the first month's trial. There was no hypothesis that Product A was often bought with Product H that was then tested. The data were simply queried to discover what relationships existed that might have previously been unnoticed. Similarly, Amazon's recommendation system produces suggestions for other items a shopper might be interested in without knowing anything about the culture and conventions of books and reading; it simply identifies patterns of purchasing across customers in order to determine if Person A likes Book X they are also likely to like Book Y given their own and others' consumption patterns.

There are a powerful and attractive set of ideas at work in this empiricist epistemology that run counter to the deductive approach that is hegemonic within modern science: big data can capture a whole of a domain and provide full resolution; there is no need for *a priori* theory, models or hypotheses; 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; meaning transcends context or domain-specific knowledge, thus can be interpreted by anyone who can decode a statistic or data visualization. These work together to suggest that a new mode of science is being created, one in which the *modus operandi* is purely inductive in nature. Whilst this empiricist epistemology is attractive, it is based on fallacious thinking with respect to the four ideas that underpin its formulation. First, big data are not exhaustive being both a representation and a sample, shaped by the technology and platform used, the data ontology employed, the regulatory environment, and are subject to sampling bias (Crawford, 2013; Kitchin, 2013). Second, big data do not arise from nowhere, free from the 'the regulating force of philosophy' (Berry, 2011: 8). Contra, systems are designed to capture certain kinds of data and the analytics and algorithms used are based on scientific reasoning and have been refined through scientific testing. Third, just as data are not generated free from theory, neither can they simply speak for themselves free of human bias or framing. Making sense of data is always cast through a particular lens that frames how they are interpreted. Further, patterns found within a data set are not inherently meaningful and correlations between variables within a data set can be random in nature and have no or little casual association. Fourth, whilst data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is likely to be anemic

or unhelpful as it lacks embedding in wider debates and knowledge.

Data-driven science

In contrast, data-driven science seeks to hold to the tenets of the scientific method, but is more open to using a hybrid combination of abductive, inductive and deductive approaches to advance the understanding of a phenomenon. It differs from the traditional, experimental deductive design in that it seeks to generate hypotheses and insights 'born from the data' rather than 'born from the theory' (Kelling *et al.*, 2009: 613). In other words, it seeks to incorporate a mode of induction into the research design, though explanation through induction is not the intended end-point (as with empiricist approaches). Instead, it forms a new mode of hypothesis generation before a deductive approach is employed. Nor does the process of induction arise from nowhere, but is situated and contextualized within a highly evolved theoretical domain. As such, the epistemological strategy adopted within data-driven science is to use guide knowledge discovery techniques to identify potential questions (hypotheses) worthy of further examination and testing. The process is guided in the sense that existing theory is used to direct the process of knowledge discovery, rather than simply hoping to identify all relationships within a dataset and assuming they are meaningful in some way. Any relationships revealed within the data do not then arise from nowhere and nor do they simply speak for themselves. The process of induction – of insights emerging from the data – is contextually framed. And those insights are not the end-point of an investigation, arranged and reasoned into a theory. Rather, the insights provide the basis for the formulation of hypotheses and the deductive testing of their validity. In other words, data-driven science is a reconfigured version of the traditional scientific method, providing a new way in which to build theory. Nonetheless, the epistemological change is significant.

Rather than empiricism and the end of theory, it is argued by some that data-driven science will become the new paradigm of scientific method in an age of big data because the epistemology favoured is suited to extracting additional, valuable insights that traditional 'knowledge-driven science' would fail to generate (Kelling *et al.*, 2009; Miller, 2010; Loukides, 2010). Knowledge-driven science, using a straight deductive approach, has particular utility in understanding and explaining the world under the conditions of scarce data and weak computation. Continuing to use such an approach, however, when technological and methodological advances mean that it is possible to undertake much richer analysis of data and to identify and tackle questions in new and exciting ways, makes little sense. Moreover, the advocates of data-driven science argue that it is much more suited to exploring, extracting value and making sense of massive, interconnected data sets; fostering interdisciplinary research that conjoins domain expertise (as it is less limited by the starting theoretical frame); and will lead to more holistic and extensive models and theories of entire complex systems rather than elements of them (Kelling *et al.*, 2009).

Computational social sciences and digital humanities

Whilst the epistemologies of big data empiricism and data-driven science seems set to transform the approach to research taken in the natural, life, physical and engineering sciences, its trajectory in the humanities and social sciences is less certain. These areas of scholarship are highly diverse in their philosophical underpinnings, with only some scholars employing the epistemology common in the sciences. For scholars in the social sciences who employ quantitative approaches big data offers a significant opportunity to develop more sophisticated, wider-scale, finer-grained models of human life.

Moreover, the variety, exhaustivity, resolution, and relationality of data, plus the growing power of computation and new data analytics, address some of the critiques of such scholarship to date, especially those of reductionism and universalism, by providing more sensitive and nuanced analysis that can take account of context and contingency, and can be used to refine and extend theoretical understandings of the social and spatial world (Lazer *et al.*, 2009; Batty *et al.*, 2012; Kitchin, 2013). Further, given the extensiveness of data (e.g., all social media posts of a society, all movements within a city) it is possible to test the veracity of such theory across a variety of settings and situations.

For post-positivist scholars, big data offers both opportunities and challenges. The opportunities are a proliferation, digitization and interlinking of a diverse set of analogue and 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 that have been rendered into digital form over the past couple of decades by a range of organizations (Cohen, 2008); and the provision of new tools of data curation, management and analysis that can handle massive numbers of data objects. Consequently, rather than concentrating on a handful of novels or photographs, or a couple of artists and their work, it becomes possible to search and connect across a large number of related works; rather than focus on a handful of websites or chat rooms or videos or online newspapers, it becomes possible to examine hundreds of thousands of such media (Manovich, 2011). These opportunities are most widely being examined through the emerging field of digital humanities.

Initially, the digital humanities consisted of the curation and analysis of data that are born digital and the digitization and archiving projects that sought to render analogue texts and material objects into digital forms that could be organized and searched and be subjected to basic forms of overarching, automated or guided analysis such as summary visualizations of content (Schnapp and Presner, 2009). Subsequently, its advocates have been divided into two camps. Those that believe new digital humanities techniques – counting, graphing, mapping and distant reading – will bring methodological rigour and objectivity to disciplines that heretofore been unsystematic and random in their focus and approach (Moretti, 2005; Ramsay, 2010). And those that argue the new techniques complement and augment existing humanities methods and facilitate traditional forms of interpretation and theory building, enabling studies of much wider scope and to answer questions that would all but impossible without computation (Berry, 2011; Manovich, 2011).

The digital humanities has not been universally welcomed with detractors contending that using computers as 'reading machines' (Ramsay, 2010) to undertake 'distant reading' (Moretti, 2005) runs counter to and undermines traditional methods of close reading. Marche (2012) contends that cultural artefacts, such as literature, cannot be treated as mere data. A piece of writing is not simply an order of letters and words, it is contextual and conveys meaning and has qualities that are ineffable. Algorithms are very poor at capturing and deciphering meaning or context. For many, the digital humanities is fostering weak, surface analysis, rather than deep, penetrating insight. It is overly reductionist and crude in its techniques, sacrificing complexity, specificity, context, depth and critique for scale, breadth, automation, descriptive patterns and the impression that interpretation does not require deep contextual knowledge.

The same kinds of argument can be levelled at computational social science. For example, a map of the language of tweets in a city might reveal patterns of geographic concentration of different ethnic communities (Rogers, 2013), but the important questions are who constitutes such concentrations,

why do they exist, what were the processes of formation and reproduction, and what are their social and economic consequences? It is one thing to identify patterns; it is another to explain them. This requires social theory and deep contextual knowledge. As such, the pattern is not the end point, but rather a starting point for additional analysis, which almost certainly is going to require other data sets. As with earlier critiques of quantitative and positivist social sciences, computational social sciences is taken to task by post-positivists as being mechanistic, atomizing, and parochial, reducing diverse individuals and complex, multidimensional social structures to mere data points (Wyly, 2014).

There is a potentially fruitful middle ground to this debate that adopts and extends the epistemologies employed in critical GIS and radical statistics. These approaches employ quantitative techniques, inferential statistics, modelling and simulation whilst being mindful and open with respect to their epistemological shortcomings, drawing on critical social theory to frame how the research is conducted, how sense is made of the findings, and the knowledge employed. Here, there is recognition that there is an inherent politics pervading the datasets analysed, the research conducted, and the interpretations made (Haraway, 1991). As such, it is acknowledged: that the researcher possesses a certain positionality (with respect to their knowledge, experience, beliefs, aspirations, etc.); that the research is situated (within disciplinary debates, the funding landscape, wider societal politics, etc.); the data are reflective of the technique used to generate them and hold certain characteristics (relating to sampling and ontological frames, data cleanliness, completeness, consistency, veracity and fidelity); and the methods of analysis utilized produce particular effects with respect to the results produced and interpretations made. Such an epistemology also does not foreclose complementing situated computational social science with small data studies that provide additional and amplifying insights (Crampton *et al.*, 2012). In other words, it is possible to think of new epistemologies that do not dismiss or reject big data analytics, but rather employ the methodological approach of data-driven science within a different epistemological framing that enables social scientists to draw valuable insights from big data.

THE LIMITS OF SOCIAL MEDIA BIG DATA

The discussion so far has argued that there is something qualitatively different about big data from small data and that it opens up new epistemological possibilities, some of which have more value than others. In general terms, it has been intimated that big data does represent a revolution in measurement that will inevitably lead to a revolution in how academic research is conducted; that big data studies will replace small data ones. However, this is unlikely to be the case for a number of reasons.

Whilst small data may be limited in volume and velocity, they have a long history of development across science, state agencies, non-governmental organizations and business, with established methodologies and modes of analysis, and a record of producing meaningful answers. Small data studies can be much more finely tailored to answer specific research questions and to explore in detail and in-depth the varied, contextual, rational and irrational ways in which people interact and make sense of the world, and how processes work. Small data can focus on specific cases and tell individual, nuanced and contextual stories.

Big data is often being repurposed to try and answer questions for which it was never designed. For example, geotagged Twitter data have not been produced to provide answers with respect to the geographical concentration of language groups in a city and the processes driving such spatial autocorrelation.

We should perhaps not be surprised then that it only provides a surface snapshot, albeit an interesting snapshot, rather than deep penetrating insights into the geographies of race, language, agglomeration and segregation in particular locales. Moreover, big data might 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 field of view/sampling frame (where data capture devices are deployed and what their settings/parameters are; who uses a space or media, e.g., who belongs to Facebook); the technology and platform used (different surveys, sensors, lens, textual prompts, layout, etc. all produce variances and biases in what data are generated); the context in which data are generated (unfolding events mean data are always situated with respect to circumstance); the data ontology employed (how the data are calibrated and classified); and the regulatory environment with respect to privacy, data protection and security (Kitchin, 2013, 2014a). Further, big data generally capture 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) – as well as data that are the 'exhaust', a by-product, of the primary task/output.

Small data studies then mine gold from working a narrow seam, whereas big data studies seek to extract nuggets through open-pit mining, scooping up and sieving huge tracts of land. These two approaches of narrow versus open mining have consequences with respect to data quality, fidelity and lineage. Given the limited sample sizes of small data, data quality – how clean (error and gap free), objective (bias free) and consistent (few discrepancies) the data are; veracity – the authenticity of the data and the extent to which they accurately (precision) and faithfully (fidelity, reliability) represent what they are meant to; and lineage – documentation that establishes provenance and fit for use; are of paramount importance (Lauriault, 2012). In contrast, it has been argued by some that big data studies do not need the same standards of data quality, veracity and lineage because the exhaustive nature of the dataset removes sampling biases and more than compensates for any errors or gaps or inconsistencies in the data or weakness in fidelity (Mayer-Schonberger and Cukier, 2013). The argument for such a view is that 'with less error from sampling we can accept more measurement error' (p.13) and 'tolerate inexactitude' (p. 16).

Nonetheless, the warning 'garbage in, garbage out' still holds. The data can be biased due to the demographic being sampled (e.g., not everybody uses Twitter) or the data might be gamed or faked through false accounts or hacking (e.g., there are hundreds of thousands of fake Twitter accounts seeking to influence trending and direct clickstream trails) (Bollier, 2010; Crampton *et al.*, 2012). Moreover, the technology being used and their working parameters can affect the nature of the data. For example, which posts on social media are most read or shared are strongly affected by ranking algorithms not simply interest (Baym, 2013). Similarly, APIs structure what data are extracted, for example, in Twitter only capturing specific hashtags associated with an event rather than all relevant tweets (Bruns, 2013), with González-Bailón *et al.* (2012) finding that different methods of accessing Twitter data – search APIs versus streaming APIs – produced quite different sets of results. As a consequence, there is no guarantee that two teams of researchers attempting to gather the same data at the same time will end up with identical datasets (Bruns, 2013). Further, the choice of metadata and variables that are being generated and which ones are being ignored paint a particular picture (Graham, 2012). With respect to fidelity there are question marks as to the extent to which social media posts really represent peoples' views and the faith that should be placed on them. Manovich (2011: 6) warns that '[p]eoples' posts, tweets, uploaded photographs, comments, and other

types of online participation are not transparent windows into their selves; instead, they are often carefully curated and systematically managed'.

There are also issues of access to both small and big data. Small data produced by academia, public institutions, non-governmental organizations and private entities can be restricted in access, limited in use to defined personnel, or available for a fee or under license. Increasingly, however, public institution and academic data are becoming more open. Big data are, with a few exceptions such as satellite imagery and national security and policing, mainly produced by the private sector. Access is usually restricted behind pay walls and proprietary licensing, limited to ensure competitive advantage and to leverage income through their sale or licensing (CIPPIC, 2006). Indeed, it is somewhat of a paradox that only a handful of entities are drowning in the data deluge (boyd and Crawford, 2012) and companies such as mobile phone operators, app developers, social media providers, financial institutions, retail chains, and surveillance and security firms are under no obligations to share freely the data they collect through their operations. In some cases, a limited amount of the data might be made available to researchers or the public through Application Programming Interfaces (APIs). For example, Twitter allows a few companies to access its firehose (stream of data) for a fee for commercial purposes (and have the latitude to dictate terms with respect to what can be done with such data), but with a handful of exceptions researchers are restricted to a 'gardenhose' (c. 10 percent of public tweets), a 'spritzer' (c. one percent of public tweets), or to different subsets of content ('white-listed' accounts), with private and protected tweets excluded in all cases (boyd and Crawford, 2012). The worry is that the insights that privately owned and commercially sold big data can provide will be limited to a privileged set of academic researchers whose findings cannot be replicated or validated (Lazer *et al.*, 2009).

Given the relative strengths and limitations of big and small data it is fair to say that small data studies will continue to be an important element of the research landscape, despite the benefits that might accrue from using big data such as social media data. However, it should be noted that small data studies will increasingly come under pressure to utilize the new archiving technologies, being scaled-up within digital data infrastructures in order that they are preserved for future generations, become accessible to re-use and combination with other small and big data, and more value and insight can be extracted from them through the application of big data analytics.

CONCLUSION

There is little doubt that much of the rhetoric concerning big data is hyped and is boosterist, especially that produced by companies seeking to push new big data products, or research centres seeking to capture grant income. At the same time, there is no doubt that big data are qualitatively different to traditional small data and it does offer the potential to change how business is conducted, societies are governed, and academic research conducted. Big data and new data analytics do offer the possibility of reframing the epistemology of science, social science and humanities (though it will not lead to the 'end of theory'), and such a reframing is already actively taking place across disciplines. Nonetheless, small data studies will continue to be valuable because they have a tried and tested track record of producing insights by working a narrow seam and due to the various shortcomings of big data. As such, one can argue that there is a revolution underway, and that it will have profound effects, but that it will not lead to full-scale regime change. With respect to social media data then, its analysis will no doubt have a strong and positive impact on sociological

and geographical research, providing a very rich, extensive, longitudinal set of data and studies, but these are most likely to complementary to a plethora of other studies.

ACKNOWLEDGEMENTS

The research for this chapter was funded by a European Research Council Advanced Investigator award (ERC-2012-AdG-323636-SOFTCITY). The chapter draws heavily on a previously published paper – Kitchin, R. (2014) Big data, new epistemologies and paradigm shifts. *Big Data and Society* 1 (April-June): 1–12 – and also Chapter 2 of Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE, London.

REFERENCES

Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*, 23rd June, http://www.wired.com/science/discoveries/magazine/16–07/pb_theory (last accessed 12th October 2012).

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y. (2012) Smart cities of the future. *European Physical Journal Special Topics* 214: 481–518.

Batty, M. (2013) *The New Science of Cities*. MIT Press: Cambridge, MA.

Baym, N.K. (2013) Data not seen: The uses and shortcomings of social media metrics. *First Monday* 18(10), http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752 (last accessed 3 January 2014).

Berry, D. (2011) The computational turn: Thinking about the digital humanities. *Culture Machine* 12, http://www.culturemachine.net/index.php/cm/article/view/440/470 (last accessed 3rd December 2012).

Bollier, D. (2010) *The Promise and Peril of Big Data*. The Aspen Institute. http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf (last accessed 1st October 2012).

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

Bruns, A. (2013) Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday* 18(10), http://firstmonday.org/ojs/index.php/fm/article/view/4879/3756 (last accessed 3rd January 2014).

CIPPIC (2006) *On the Data Trail: How detailed information about you gets into the hands of organizations with whom you have no relationship. A Report on the Canadian Data Brokerage Industry*. The Canadian Internet Policy and Public Interest Clinic, Ottawa. http://www.cippic.ca/uploads/May1–06/DatabrokerReport.pdf (last accessed 17th January 2014).

Cohen, D. (2008) Contribution to: The Promise of Digital History (roundtable discussion), *Journal of American History* 95(2): 452–491.

Crampton, J., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W. and Zook, M. (2012) *Beyond the Geotag? Deconstructing "Big Data" and leveraging the Potential of the Geoweb*. http://www.uky.edu/~tmute2/geography_methods/readingPDFs/2012-Beyond-the-Geotag-2012.10.01.pdf (last accessed 21st February 2013).

Crawford, K. (2013) The hidden biases of big data. *Harvard Business Review Blog*. April 1st. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/ (last accessed 18th September 2013).

Cukier, K. (2010) Data, data everywhere. *The Economist*, February 25th. http://www.economist.com/node/15557443 (last accessed Nov 12 2012).

Diebold, F. (2012) *A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline*. http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf (last accessed 5th February 2013).

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.

Dyche, J. (2012) Big Data "Eurekas!" Don't Just Happen. *Harvard Business Review Blog*,

20th November. http://blogs.hbr.org/cs/2012/11/eureka_doesnt_just_happen.html (last accessed 23th November 2012).

Franks, B. (2012) *Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics*. Wiley: Hoboken, NJ.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holtoefer, J. and Moreno, Y. (2012) Assessing the Bias in Communication Networks Sampled from Twitter. Working Paper. http://arxiv.org/abs/1212.1684 (last accessed 17th January 2014).

Graham, M. (2012) Big data and the end of theory? *The Guardian*, 9th March. http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory (last accessed 12th November 2012).

Haraway, D. (1991) *Simians, Cyborgs and Women: The Reinvention of Nature*. Routledge: New York.

Hey, T., Tansley, S. and Tolle, K. (2009) Jim Grey on eScience: A transformed scientific method, in Hey, T., Tansley, S. and Tolle, K. (eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research; Redmond, Washington. xvii-xxxi.

Kelling, S., Hochachka, W., Fink, D., Riedewald, M., Caruana, R., Ballard, G. and Hooker, G. (2009) Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* 59(7): 613–620.

Kitchin, R. (2013) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 79(1): 1–14.

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

Kitchin, R. (2014b) The real-time city? Big data and smart urbanism. *Geojournal* 3(3): 262–267.

Kitchin, R. (2016) *Getting smarter about smart cities: Improving data privacy and data security*. Data Protection Unit, Department of the Taoiseach, Dublin, Ireland.

Kitchin, R. and McArdle, G. (2016) What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3: 1–10.

Kitchin, R., Lauriault, T. and McArdle, G. (2015) Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. *Regional Studies, Regional Science* 2: 1–28.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

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 (last accessed 16th January 2013).

Lauriault, T.P. (2012) *Data, Infrastructures and Geographical Imaginations: Mapping Data Access Discourses in Canada*. PhD Thesis, Carleton University, Ottawa.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) Computational Social Science. *Science* 323: 721–733.

Loukides, M. (2010) What is data science? *O'Reilly Radar*, 2nd June 2010, http://radar.oreilly.com/2010/06/what-is-data-science.html (last accessed 28th January 2013).

Manovich, L. (2011) *Trending: The Promises and the Challenges of Big Social Data*. http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (last accessed 9th November 2012).

Marche, S. (2012) Literature is not Data: Against Digital Humanities. *Los Angeles Review of Books*, 28th October 2012, http://lareviewofbooks.org/article.php?id=1040&fulltext=1 (last accessed 4th April 2013).

Marr, B. (2014) Big Data: The 5 Vs Everyone Must Know. Mar 6, https://www.linkedin.com/pulse/20140306073407–64875646-big-data-the-5-vs-everyone-must-know (last accessed 4 Sept 2015)

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

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

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

Moretti, F. (2005) *Graphs, Maps, Trees: Abstract Models for a Literary History.* Verso, London.

National Science Foundation (2012) *Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA)*. Programme solicitation NSF 12–499, http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf (last accessed 25th February 2013).

Prensky, M. (2009) H. sapiens digital: From digital immigrants and digital natives to digital wisdom. *Innovate* 5(3), http://www.innovateonline.info/index.php?view=article&id=705 (last accessed 12th October 2012).

Ramsay, S. (2010) *Reading Machines: Towards an Algorithmic Criticism*. University of Illinois Press: Champaign, IL.

Rogers, S. (2013) Twitter's languages of New York mapped. *The Guardian*, 21st February 2013 http://www.guardian.co.uk/news/datablog/interactive/2013/feb/21/twitter-languages-new-york-mapped (last accessed 3rd April 2013).

Schnapp, J. and Presner, P. (2009) *Digital Humanities Manifesto 2.0*. http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (last accessed 13 March 2013).

Siegel, E. (2013) *Predictive Analytics*. Wiley: Hoboken, NJ.

Townsend, A. (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. W.W. Norton & Co: New York.

Wyly, E. (2014) Automated (post)positivism. *Urban Geography* 35(5): 669–690.