Original Research Article

BIG DATA & SOCIETY

Big Data & Society
April–June 2014: 1–12

DOI: 10.1177/2053951714528481
bds.sagepub.com
SAGE

Big Data, new epistemologies and paradigm shifts

Rob Kitchin

Abstract

This article examines how the availability of Big Data, coupled with new data analytics, challenges established epistemologies across the sciences, social sciences and humanities, and assesses the extent to which they are engendering paradigm shifts across multiple disciplines. In particular, it critically explores new forms of empiricism that declare 'the end of theory', the creation of data-driven rather than knowledge-driven science, and the development of digital humanities and computational social sciences that propose radically different ways to make sense of culture, history, economy and society. It is argued that: (1) Big Data and new data analytics are disruptive innovations which are reconfiguring in many instances how research is conducted; and (2) there is an urgent need for wider critical reflection within the academy on the epistemological implications of the unfolding data revolution, a task that has barely begun to be tackled despite the rapid changes in research practices presently taking place. After critically reviewing emerging epistemological positions, it is contended that a potentially fruitful approach would be the development of a situated, reflexive and contextually nuanced epistemology.

Keywords

Big Data, data analytics, epistemology, paradigms, end of theory, data-driven science, digital humanities, computational social sciences

Introduction

> Revolutions in science have often been preceded by revolutions in measurement. Sinan Aral (cited in Cukier, 2010)

> Big Data creates a radical shift in how we think about research . . .. [It offers] a profound change at the levels of epistemology and ethics. Big Data reframes key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorization of reality . . . Big Data stakes out new terrains of objects, methods of knowing, and definitions of social life. (boyd and Crawford, 2012)

As with many rapidly emerging concepts, Big Data has been variously defined and operationalized, ranging from trite proclamations that Big Data consists of datasets too large to fit in an Excel spreadsheet or be stored on a single machine (Strom, 2012) to more sophisticated ontological assessments that tease out its inherent characteristics (boyd and Crawford, 2012; Mayer-Schonberger and Cukier, 2013). Drawing on an extensive engagement with the literature, Kitchin (2013) details that Big Data is:

- 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 (n = all);

National Institute for Regional and Spatial Analysis, National University of Ireland Maynooth, County Kildare, Ireland

Corresponding author:
Rob Kitchin, National Institute for Regional and Spatial Analysis, National University of Ireland Maynooth, County Kildare, Ireland.
Email: Rob.Kitchin@nuim.ie

- fine-grained in *resolution* 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 *scaleability* (can expand in size rapidly). (see boyd and Crawford, 2012; Dodge and Kitchin, 2005; Laney, 2001; Marz and Warren, 2012; Mayer-Schonberger and Cukier, 2013; Zikopoulos et al., 2012).

In other words, Big Data is not simply denoted by volume. Indeed, industry, government and academia have long produced massive data sets – for example, national censuses. However, given the costs and difficulties of generating, processing, analysing and storing such datasets, these data have been produced in tightly controlled ways using sampling techniques that limit their scope, temporality and size (Miller, 2010). To make the exercise of compiling census data manageable they have been produced once every five or 10 years, asking just 30 to 40 questions, and their outputs are usually quite coarse in resolution (e.g. local areas or counties rather than individuals and households). Moreover, the methods used to generate them are quite inflexible (for example, once a census is set and is being administered it is impossible to tweak or add/remove questions). Whereas the census seeks to be exhaustive, enumerating all people living in a country, most surveys and other forms of data generation are samples, seeking to be representative of a population.

In contrast, Big Data is characterized by being generated continuously, seeking to be exhaustive and fine-grained in scope, and flexible and scalable in its production. Examples of the production of such data include: digital CCTV; the recording of retail purchases; digital devices that record and communicate the history of their own use (e.g. mobile phones); the logging of transactions and interactions across digital networks (e.g. email or online banking); clickstream data that record navigation through a website or app; measurements from sensors embedded into objects or environments; the scanning of machine-readable objects such as travel passes or barcodes; and social media postings (Kitchin, 2014). These are producing massive, dynamic flows of diverse, fine-grained, relational data. For example, in 2012 Wal-Mart was generating more than 2.5 petabytes (2^{50} bytes) of data relating to more than 1 million customer transactions *every hour* (Open Data Center Alliance, 2012) and Facebook reported that it was processing 2.5 billion pieces of content (links, comments, etc.), 2.7 billion 'Like' actions and 300 million photo uploads *per day* (Constine, 2012). Handling and analysing such data is a very different proposition to dealing with a census every 10 years or a survey of a few hundred respondents.

Whilst the production of such Big Data has existed in some domains, such as remote sensing, weather prediction, and financial markets, for some time, a number of technological developments, such as ubiquitous computing, widespread internet working, and new database designs and storage solutions, have created a tipping point for their routine generation and analysis, not least of which are new forms of data analytics designed to cope with data abundance (Kitchin, 2014). Traditionally, data analysis techniques have been designed to extract insights from scarce, static, clean and poorly relational data sets, scientifically sampled and adhering to strict assumptions (such as independence, stationarity, and normality), and generated and analysed with a specific question in mind (Miller, 2010). The challenge of analysing Big Data is coping with abundance, exhaustivity and variety, timeliness and dynamism, messiness and uncertainty, high relationality, and the fact that much of what is generated has no specific question in mind or is a by-product of another activity. Such a challenge was until recently too complex and difficult to implement, but has become possible due to high-powered computation and new analytical techniques. These new techniques are rooted in research concerning artificial intelligence and expert systems that have sought to produce machine learning that can computationally and automatically mine and detect patterns and build predictive models and optimize outcomes (Han et al., 2011; Hastie et al., 2009). Moreover, since different models have their strengths and weaknesses, and it is often difficult to prejudge which type of model and its various versions will perform best on any given data set, an ensemble approach can be employed to build multiple solutions (Seni and Elder, 2010). Here, literally hundreds of different algorithms can be applied to a dataset to determine the best or a composite model or explanation (Siegel, 2013), a radically different approach to that traditionally used wherein the analyst selects an appropriate method based on their knowledge of techniques and the data. In other words, Big Data analytics enables an entirely new epistemological approach for making sense of the world; rather than testing a theory by analysing relevant data, new data analytics seek to gain insights 'born from the data'.

The explosion in the production of Big Data, along with the development of new epistemologies, is leading many to argue that a data revolution is under way that has far-reaching consequences to how knowledge is produced, business conducted, and governance enacted (Anderson, 2008; Bollier, 2010; Floridi, 2012; Mayer-Schonberger and Cukier, 2013). With respect to knowledge production, it is contended that Big Data presents

Table 1. Four paradigms of science.

Paradigm	Nature	Form	When
First	Experimental science	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical science	Modelling and generalization	pre-computers
Third	Computational science	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory science	Data-intensive; statistical exploration and data mining	Now

Compiled from Hey et al. (2009).

the possibility of a new research paradigm across multiple disciplines. As set out by Kuhn (1962), a paradigm constitutes an accepted way of interrogating the world and synthesizing knowledge common to a substantial proportion of researchers in a discipline at any one moment in time. Periodically, Kuhn argues, a new way of thinking emerges that challenges accepted theories and approaches. For example, Darwin's theory of evolution radically altered conceptual thought within the biological sciences, as well as challenging the religious doctrine of creationism. Jim Gray (as detailed in Hey et al., 2009) charts the evolution of science through four broad paradigms (see Table 1). Unlike Kuhn's proposition that paradigm shifts occur because the dominant mode of science cannot account for particular phenomena or answer key questions, thus demanding the formulation of new ideas, Gray's transitions are founded on advances in forms of data and the development of new analytical methods. He thus proposes that science is entering a fourth paradigm based on the growing availability of Big Data and new analytics.

Kuhn's argument has been subject to much critique, not least because within some academic domains there is little evidence of paradigms operating, notably in some social sciences where there is a diverse set of philosophical approaches employed (e.g. human geography, sociology), although in other domains, such as the sciences, there has been more epistemological unity around how science is conducted, using a well defined scientific method, underpinned by hypothesis testing to verify or falsify theories. Moreover, paradigmatic accounts produce overly sanitized and linear stories of how disciplines evolve, smoothing over the messy, contested and plural ways in which science unfolds in practice. Nevertheless, whilst the notion of paradigms is problematic, it has utility in framing the current debates concerning the development of Big Data and their consequences because many of the claims being made with respect to knowledge production contend that a fundamentally different epistemology is being created; that a transition to a new paradigm is under way. However, the form that this new epistemology is taking is contested. The rest of this paper critically examines the development of an emerging fourth paradigm in science and its form, and explores to what extent the data revolution is leading to alternative epistemologies in the humanities and social sciences and changing research practices.

A fourth paradigm in science?

Whilst Jim Gray envisages the fourth paradigm of science to be data-intensive and a radically new extension of the established scientific method, others suggest that Big Data ushers in a new era of empiricism, wherein the volume of data, accompanied by techniques that can reveal their inherent truth, enables data to speak for themselves free of theory. The empiricist view has gained credence outside of the academy, especially within business circles, but its ideas have also taken root in the new field of data science and other sciences. In contrast, a new mode of data-driven science is emerging within traditional disciplines in the academy. In this section, the epistemological claims of both approaches are critically examined, mindful of the different drivers and aspirations of business and the academy, with the former preoccupied with employing data analytics to identify new products, markets and opportunities rather than advance knowledge per se, and the latter focused on how best to make sense of the world and to determine explanations as to phenomena and processes.

The end of theory: Empiricism reborn

For commentators such as Chris Anderson, former editor-in-chief at *Wired* magazine, Big Data, new data analytics and ensemble approaches signal a new era of knowledge production characterized by 'the end of theory'. In a provocative piece, Anderson (2008) argues that 'the data deluge makes the scientific method obsolete'; that the patterns and relationships contained within Big Data inherently produce meaningful and insightful knowledge about complex phenomena. Essentially arguing that Big Data enables an empiricist mode of knowledge production, he contends:

> 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) argues:

> 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 argues that 'mining Big Data reveals relationships and patterns that we didn't even know to look for.' Likewise, Steadman (2013) argues:

> The Big Data approach to intelligence gathering allows an analyst to get the full resolution on worldwide affairs. Nothing is lost from looking too closely at one particular section of data; nothing is lost from trying to get too wide a perspective on a situation that the fine detail is lost.... The analyst doesn't even have to bother proposing a hypothesis anymore.

The examples used to illustrate such a position usually stem from marketing and retail. For example, Dyche (2012) details the case of a retail chain that analysed 12 years' worth of purchase transactions for possible unnoticed relationships between products that ended up in shoppers' baskets. Discovering correlations between certain items led to new product placements and a 16% increase in revenue per shopping cart 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. Whilst it might be desirable to explain why associations exist within the data and why they might be meaningful, such explanation is cast as largely unnecessary. Siegel (2013: 90) thus argues with respect to predictive analytics: 'We usually don't know about causation, and we often don't necessarily care... the objective is more to predict than it is to understand the world... It just needs to work; prediction trumps explanation'.

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 didn't know to ask in the first place. Simply stated, Ayasdi is 'digital serendipity'. (Clark, 2013)

Further, it purports to have totally removed

> the human element that goes into data mining – and, as such, all the human bias that goes with it. Instead of waiting to be asked a question or be directed to specific existing data links, the system will – undirected – deliver patterns a human controller might not have thought to look for. (Clark, 2013)

There is a powerful and attractive set of ideas at work in the empiricist epistemology that runs counter to the deductive approach that is hegemonic within modern science:

- Big Data can capture a whole 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 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, though Big Data may seek to be exhaustive, capturing a whole domain and providing full resolution, it is both a representation and a sample, shaped by the technology and platform used, the data ontology employed and the regulatory environment, and it is subject to sampling bias (Crawford, 2013; Kitchin, 2013). Indeed, all data provide oligoptic views of the world: views from certain vantage points, using particular tools, rather than an all-seeing, infallible God's eye view (Amin and Thrift, 2002; Haraway, 1991). As such, data are not simply

natural and essential elements that are abstracted from the world in neutral and objective ways and can be accepted at face value; data are created within a complex assemblage that actively shapes its constitution (Ribes and Jackson, 2013).

Second, Big Data does 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. As such, an inductive strategy of identifying patterns within data does not occur in a scientific vacuum and is discursively framed by previous findings, theories, and training; by speculation that is grounded in experience and knowledge (Leonelli, 2012). New analytics might present the illusion of automatically discovering insights without asking questions, but the algorithms used most certainly did arise and were tested scientifically for validity and veracity.

Third, just as data are not generated free from theory, neither can they simply speak for themselves free of human bias or framing. As Gould (1981: 166) notes, 'inanimate data can never speak for themselves, and we always bring to bear some conceptual framework, either intuitive and ill-formed, or tightly and formally structured, to the task of investigation, analysis, and interpretation'. Making sense of data is always framed – data are examined through a particular lens that influences how they are interpreted. Even if the process is automated, the algorithms used to process the data are imbued with particular values and contextualized within a particular scientific approach. Further, patterns found within a data set are not inherently meaningful. Correlations between variables within a data set can be random in nature and have no or little causal association, and interpreting them as such can produce serious ecological fallacies. This can be exacerbated in the case of Big Data as the empiricist position appears to promote the practice of data dredging – hunting for every association or model.

Fourth, the idea that data can speak for themselves suggests that anyone with a reasonable understanding of statistics should be able to interpret them without context or domain-specific knowledge. This is a conceit voiced by some data and computer scientists and other scientists, such as physicists, all of whom have become active in practising social science and humanities research. For example, a number of physicists have turned their attention to cities, employing Big Data analytics to model social and spatial processes and to identify supposed laws that underpin their formation and functions (Bettencourt et al., 2007; Lehrer, 2010). These studies often wilfully ignore a couple of centuries of social science scholarship, including nearly a century of quantitative analysis and model building. The result is an analysis of cities that is reductionist, functionalist and ignores the effects of culture, politics, policy, governance and capital (reproducing the same kinds of limitations generated by the quantitative/positivist social sciences in the mid-20th century). A similar set of concerns is shared by those in the sciences. Strasser (2012), for example, notes that within the biological sciences, bioinformaticians who have a very narrow and particular way of understanding biology are claiming ground once occupied by the clinician and the experimental and molecular biologist. These scientists are undoubtedly ignoring the observations of Porway (2013):

> Without subject matter experts available to articulate problems in advance, you get [poor] results . . . . Subject matter experts are doubly needed to assess the results of the work, especially when you're dealing with sensitive data about human behavior. As data scientists, we are well equipped to explain the 'what' of data, but rarely should we touch the question of 'why' on matters we are not experts in.

Put simply, whilst data can be interpreted free of context and domain-specific expertise, such an epistemological interpretation is likely to be anaemic or unhelpful as it lacks embedding in wider debates and knowledge.

These fallacious notions have gained some traction, especially within business circles, because they possess a convenient narrative for the aspirations of knowledge-orientated businesses (e.g. data brokers, data analytic providers, software vendors, consultancies) in selling their services. Within the empiricist frame, data analytics offer the possibility of insightful, objective and profitable knowledge without science or scientists, and their associated overheads of cost, contingencies, and search for explanation and truth. In this sense, whilst the data science techniques employed might hold genuine salience for practioners, the articulation of a new empiricism operates as a discursive rhetorical device designed to simplify a more complex epistemological approach and to convince vendors of the utility and value of Big Data analytics.

Data-driven science

In contrast to new forms of empiricism, 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 guided 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. As such, how data are generated or repurposed is directed by certain assumptions, underpinned by theoretical and practical knowledge and experience as to whether technologies and their configurations will capture or produce appropriate and useful research material. Data are not generated by every means possible, using every kind of available technology or every kind of sampling framework; rather, strategies of data generation and repurposing are carefully thought out, with strategic decisions made to harvest certain kinds of data and not others. Similarly, how these data are processed, managed and analysed is guided by assumptions as to which techniques might provide meaningful insights. The data are not subject to every ontological framing possible, or every form of data-mining technique in the hope that they reveal some hidden truth. Rather, theoretically informed decisions are made as to how best to tackle a data set such that it will reveal information which will be of potential interest and is worthy of further research. And instead of testing whether every relationship revealed has veracity, attention is focused on those – based on some criteria – that seemingly offer the most likely or valid way forward. Indeed, many supposed relationships within data sets can quickly be dismissed as trivial or absurd by domain specialists, with others flagged as deserving more attention (Miller, 2010).

Such decision-making with respect to methods of data generation and analysis are based on abductive reasoning. Abduction is a mode of logical inference and reasoning forwarded by C. S. Peirce (1839–1914) (Miller, 2010). It seeks a conclusion that makes reasonable and logical sense, but is not definitive in its claim. For example, there is no attempt to deduce what is the best way to generate data, but rather to identify an approach that makes logical sense given what is already known about such data production. Abduction is very commonly used in science, especially in the formulation of hypotheses, though such use is not widely acknowledged. 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; Loukides, 2010; Miller, 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 – applying new data analytics and being able to connect together large, disparate data together in ways that were hitherto impossible, and which produce new valuable data and 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 that it will lead to more holistic and extensive models and theories of entire complex systems rather than elements of them (Kelling et al., 2009).

For example, it is contended that data-driven science will transform our understanding of environmental systems (Bryant et al., 2008; Lehning et al., 2009). It will enable high-resolution data being generated from a variety of sources, often in real-time (such as conventional and mobile weather stations, satellite and aerial imagery, weather radar, stream observations and gauge stations, citizen observations, ground and aerial LIDAR, water-quality sampling, gas measures, soil cores, and distributed sensors that measure selected domains such as air temperature and moisture) to be integrated together to provide very detailed models of environments in flux (as opposed to at freeze-points in time and space) and to identify specific relationships between

phenomena and processes that generate new hypotheses and theories that can then be tested further to establish their veracity. It will also help to identify and further understand connection points between different environmental spheres – such as the atmosphere (air), biosphere (ecosystems), hydrosphere (water systems), lithosphere (rocky shell of the Earth) and pedosphere (soils) – and aid in the integration of theories into a more holistic theoretical assemblage. This will provide a better comprehension of the diverse, interrelated processes at work and the interconnections with human systems, and can be used to guide models and simulations for predicting long-term trends and possible adaptive strategies.

Computational social sciences and digital humanities

Whilst the epistemologies of Big Data empiricism and data-driven science seem set to transform the approach to research taken in the natural, life, physical and engineering sciences, their 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. Those using the scientific method in order to explain and model social phenomena, in general terms, draw on the ideas of positivism (though they might not adopt such a label; Kitchin, 2006). Such work tends to focus on factual, quantified information – empirically observable phenomena that can be robustly measured (such as counts, distance, cost, and time), as opposed to more intangible aspects of human life such as beliefs or ideology – using statistical testing to establish causal relationships and to build theories and predictive models and simulations. Positivistic approaches are well established in economics, political science, human geography and sociology, but are rare in the humanities. However, within those disciplines mentioned, there has been a strong move over the past half century towards post-positivist approaches, especially in human geography and sociology.

For positivistic scholars in the social sciences, Big Data offers a significant opportunity to develop more sophisticated, wider-scale, finer-grained models of human life. Notwithstanding concerns over access to social and economic Big Data (much of which is generated by private interests) and issues such as data quality, Big Data offers the possibility of shifting 'from data-scarce to data-rich studies of societies; from static snapshots to dynamic unfoldings; from coarse aggregations to high resolutions; from relatively simple models to more complex, sophisticated simulations' (Kitchin, 2014: 3). The potential exists for a new era of computational social science that produces studies with much greater breadth, depth, scale, and timeliness, and that are inherently longitudinal, in contrast to existing social sciences research (Lazer et al., 2009; Batty et al., 2012). 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 positivistic scholarship to date, especially those of reductionism and universalism, by providing more finely-grained, 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 (Kitchin, 2013). Further, given the extensiveness of data, it is possible to test the veracity of such theory across a variety of settings and situations. In such circumstances, it is argued that knowledge about individuals, communities, societies and environments will become more insightful and useful with respect to formulating policy and addressing the various issues facing humankind.

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 much of which has 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 also 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. The first group believes that new digital humanities techniques – counting, graphing, mapping and distant reading – bring methodological rigour and objectivity

to disciplines that heretofore have been unsystematic and random in their focus and approach (Moretti, 2005; Ramsay, 2010). In contrast, the second group argues that, rather than replacing traditional methods or providing an empiricist or positivistic approach to humanities scholarship, new techniques complement and augment existing humanities methods and facilitate traditional forms of interpretation and theory-building, enabling studies of much wider scope to answer questions that would be all but unanswerable 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. Culler (2010: 22) notes that close reading involves paying 'attention to how meaning is produced or conveyed, to what sorts of literary and rhetorical strategies and techniques are deployed to achieve what the reader takes to be the effects of the work or passage' – something that a distant reading is unable to perform. His worry is that a digital humanities approach promotes literary scholarship that involves no actual reading. Similarly, Trumpener (2009: 164) argues that a 'statistically driven model of literary history . . . seems to necessitate an impersonal invisible hand', continuing: 'any attempt to see the big picture needs to be informed by broad knowledge, an astute, historicized sense of how genres and literary institutions work, and incisive interpretive tools' (pp. 170–171). Likewise, 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 and, Marche argues, treat 'all literature as if it were the same'. He continues:

> [t]he algorithmic analysis of novels and of newspaper articles is necessarily at the limit of reductivism. The process of turning literature into data removes distinction itself. It removes taste. It removes all the refinement from criticism. It removes the history of the reception of works.

Jenkins (2013) thus concludes:

> the value of the arts, the quality of a play or a painting, is not measurable. You could put all sorts of data into a machine: dates, colours, images, box office receipts, and none of it could explain what the artwork is, what it means, and why it is powerful. That requires man [sic], not machine.

For many, then, 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 are 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, in press). Moreover, the analysis is riddled with assumptions of social determinism, as exemplified by Pentland (2012): 'the sort of person you are is largely determined by your social context, so if I can see some of your behaviors, I can infer the rest, just by comparing you to the people in your crowd'. In contrast, human societies, it is argued, are too complex, contingent and messy to be reduced to formulae and laws, with quantitative models providing little insight into phenomena such as wars, genocide, domestic violence and racism, and only circumscribed insight into other human systems such as the economy, inadequately accounting for the role of politics, ideology, social structures, and culture (Harvey, 1972). People do not act in rational, pre-determined ways, but rather live lives full of contradictions, paradoxes, and unpredictable occurrences. How societies are organized and operate varies across time and space and there is no optimal or ideal form, or universal traits. Indeed, there is an incredible diversity of individuals, cultures and modes of living across the planet. Reducing this complexity to the abstract subjects that populate universal models does symbolic violence to how we create knowledge. Further, positivistic approaches wilfully ignore the metaphysical aspects of human life (concerned with meanings, beliefs, experiences) and normative questions (ethical and moral dilemmas about how things should be as opposed to how they are) (Kitchin, 2006). In other words, positivistic approaches only

focus on certain kinds of questions, which they seek to answer in a reductionist way that seemingly ignores what it means to be human and to live in richly diverse societies and places. This is not to say that quantitative approaches are not useful – they quite patently are – but that their limitations in understanding human life should be recognized and complemented with other approaches.

Brooks (2013) thus contends that Big Data analytics struggles with the social (people are not rationale and do not behave in predictable ways; human systems are incredibly complex, having contradictory and paradoxical relation); struggles with context (data are largely shorn of the social, political and economic and historical context); creates bigger haystacks (consisting of many more spurious correlations, making it difficult to identify needles); has trouble addressing big problems (especially social and economic ones); favours memes over masterpieces (identifies trends but not necessarily significant features that may become a trend); and obscures values (of the data producers and those that analyse them and their objectives). In other words, whilst Big Data analytics might provide some insights, it needs to be recognized that they are limited in scope, produce particular kinds of knowledge, and still need contextualization with respect to other information, whether that be existing theory, policy documents, small data studies, or historical records, that can help to make sense of the patterns evident (Crampton et al., 2012).

Beyond the epistemological and methodological approach, part of the issue is that much Big Data and analysis seem to be generated with no specific questions in mind, or the focus is driven by the application of a method or the content of the data set rather than a particular question, or the data set is being used to seek an answer to a question that it was never designed to answer in the first place. With respect to the latter, geotagged Twitter data has 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.

Whereas most digital humanists recognize the value of close readings, and stress how distant readings complement them by providing depth and contextualization, positivistic forms of social science are oppositional to post-positivist approaches. The difference between the humanities and social sciences in this respect is because the statistics used in the digital humanities are largely descriptive – identifying and plotting patterns. In contrast, the computational social sciences employ the scientific method, complementing descriptive statistics with inferential statistics that seek to identify associations and causality. In other words, they are underpinned by an epistemology wherein the aim is to produce sophisticated statistical models that explain, simulate and predict human life. This is much more difficult to reconcile with post-positivist approaches. Advocacy then rests on the utility and value of the method and models, not on providing complementary analysis of a more expansive set of data.

There is a potentially fruitful alternative to this position 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 research is not a neutral, objective activity that produces a view from nowhere, and that there is an inherent politics pervading the datasets analysed, the research conducted, and the interpretations made (Haraway, 1991; Rose, 1997). As such, the researcher is acknowledged to possess 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. Moreover, it is recognized that how the research is employed is not ideologically-neutral but is framed in subtle and explicit ways by the aspirations and intentions of the researchers and funders/sponsors, and those that translate such research into various forms of policy, instruments, and action. In other words, within such an epistemology the research conducted is reflexive and open with respect to the research process, acknowledging the contingencies and relationalities of the approach employed, thus producing nuanced and contextualized accounts and conclusions. 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 that are situated and reflexive.

Conclusion

There is little doubt that the development of Big Data and new data analytics offers the possibility of reframing the epistemology of science, social science and humanities, and such a reframing is already actively taking place across disciplines. Big Data and new data analytics enable new approaches to data generation and analyses to be implemented that make it possible to ask and answer questions in new ways. Rather than seeking to extract insights from datasets limited by scope, temporality and size, Big Data provides the counter problem of handling and analysing enormous, dynamic, and varied datasets. The solution has been the development of new forms of data management and analytical techniques that rely on machine learning and new modes of visualization.

With respect to the sciences, access to Big Data and new research praxes has led some to proclaim the emergence of a new fourth paradigm, one rooted in data-intensive exploration that challenges the established scientific deductive approach. At present, whilst it is clear that Big Data is a disruptive innovation, presenting the possibility of a new approach to science, the form of this approach is not set, with two potential paths proposed that have divergent epistemologies – empiricism, wherein the data can speak for themselves free of theory, and data-driven science that radically modifies the existing scientific method by blending aspects of abduction, induction and deduction. Given the weaknesses in the empiricist arguments it seems likely that the data-driven approach will eventually win out and over time, as Big Data becomes more common and new data analytics are advanced, will present a strong challenge to the established knowledge-driven scientific method. To accompany such a transformation the philosophical underpinnings of data-driven science, with respect to its epistemological tenets, principles and methodology, need to be worked through and debated to provide a robust theoretical framework for the new paradigm.

The situation in the humanities and social sciences is somewhat more complex given the diversity of their philosophical underpinnings, with Big Data and new analytics being unlikely to lead to the establishment of new disciplinary paradigms. Instead, Big Data will enhance the suite of data available for analysis and enable new approaches and techniques, but will not fully replace traditional small data studies. This is partly due to philosophical positions, but also because it is unlikely that suitable Big Data will be produced that can be utilized to answer particular questions, thus necessitating more targeted studies. Nonetheless, as Kitchin (2013) and Ruppert (2013) argue, Big Data presents a number of opportunities for social scientists and humanities scholars, not least of which are massive quantities of very rich social, cultural, economic, political and historical data. It also poses a number of challenges, including a skills deficit for analysing and making sense of such data, and the creation of an epistemological approach that enables post-positivist forms of computational social science. One potential path forward is an epistemology that draws inspiration from critical GIS and radical statistics in which quantitative methods and models are employed within a framework that is reflexive and acknowledges the situatedness, positionality and politics of the social science being conducted, rather than rejecting such an approach out of hand. Such an epistemology also has potential utility in the sciences for recognizing and accounting for the use of abduction and creating a more reflexive data-driven science. As this tentative discussion illustrates, there is an urgent need for wider critical reflection on the epistemological implications of Big Data and data analytics, a task that has barely begun despite the speed of change in the data landscape.

Acknowledgements

Evelyn Ruppert and Mark Boyle provided some useful comments on an initial draft of this paper. The research for this paper was funded by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636).

References

Amin A and Thrift N (2002) *Cities: Reimagining the Urban*. London: Polity.

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

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

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

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

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

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

Brooks D (2013) What data can't do. *New York Times*, 18 February 2013. Available at: http://www.nytimes.com/2013/02/19/opinion/brooks-what-data-cant-do.html (accessed 18 February 2013).

Bryant R, Katz RH and Lazowska ED (2008) Big-data computing: Creating revolutionary breakthroughs in commerce, science and society. In: *Computing Research Initiatives for the 21st Century, Computing Research Association, Ver. 8*. Available at: http://www.cra.org/ccc/docs/init/Big_Data.pdf (accessed 12 October 2012).

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

Cohen D (2008) Contribution to: The promise of digital history (roundtable discussion). *Journal of American History* 95(2): 452–491.

Constine J (2012) How big is Facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day, 22 August 2012. Available at: http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/ (accessed 28 January 2013).

Crampton J, Graham M, Poorthuis A, et al. (2012) *Beyond the Geotag? Deconstructing 'Big Data' and Leveraging the Potential of the Geoweb*. Available at: http://www.uky.edu/~tmute2/geography_methods/readingPDFs/2012-Beyond-the-Geotag-2012.10.01.pdf (accessed 21 February 2013).

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

Cukier K (2010) Data, data everywhere. *The Economist*, 25 February (accessed 12 November 2012).

Culler J (2010) The closeness of close reading. *ADE Bulletin* 149: 20–25.

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*. 20 November. Available at: http://blogs.hbr.org/cs/2012/11/eureka_doesnt_just_happen.html (accessed 23 November 2012).

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

Gould P (1981) Letting the data speak for themselves. *Annals of the Association of American Geographers* 71(2): 166–176.

Han J, Kamber M and Pei (2011) *Data Mining: Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann.

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

Harvey D (1972) *Social Justice and the City*. Oxford: Blackwell.

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer.

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*. Redmond: Microsoft Research, pp. xvii–xxxi.

Jenkins T (2013) Don't count on big data for answers. *The Scotsman*, 12 February 2013. Available at: http://www.scotsman.com/the-scotsman/opinion/comment/tiffany-jenkins-don-t-count-on-big-data-for-answers-1-2785890 (accessed 11 March 2013).

Kelling S, Hochachka W, Fink D, et al. (2009) Data-intensive Science: A new paradigm for biodiversity studies. *BioScience* 59(7): 613–620.

Kitchin R (2006) Positivistic geography and spatial science. In: Aitken S and Valentine G (eds) *Approaches in Human Geography*. London: Sage, pp. 20–29.

Kitchin R (2013) Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography* 3(3): 262–267.

Kitchin R (2014) The real-time city? Big data and smart urbanism. *GeoJournal* 79: 1–14.

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

Laney D (2001) 3D data management: Controlling data volume, velocity and variety. Meta group. Available at: 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, et al. (2009) Computational social science. *Science* 323: 721–733.

Lehning M, Dawes N, Bavay M. et al. (2009) Instrumenting the earth: Next-generation sensor networks and environmental science. In: Hey T, Tansley S and Tolle K (eds) *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond: Microsoft Research, pp. 45–51.

Lehrer J (2010) A physicist solves the city. *New York Times*, 17 December. Available at: http://www.nytimes.com/2010/12/19/magazine/19Urban_West-t.html (accessed 23 December 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): 1–3.

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

Manovich L (2011) Trending: The promises and the challenges of big social data. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf (accessed 9 November 2012).

Marche S (2012) Literature is not data: Against digital humanities. *Los Angeles Review of Books*, 28 October 2012. Available at: http://lareviewofbooks.org/article.php?id=1040&fulltext=1 (accessed 4 April 2013).

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

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

Miller HJ (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*. London: Verso.

Open Data Center Alliance (2012) *Big Data Consumer Guide*. Open Data Center Alliance. Available at: http://www.opendatacenteralliance.org/docs/Big_Data_Consumer_Guide_Rev1.0.pdf (accessed 11 February 2013).

Pentland A (2012) Reinventing society in the wake of big data. *Edge*, 30 August 2012. Available at: http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data (accessed 28 January 2013).

Porway J (2013) You can't just hack your way to social change. *Harvard Business Review Blog*, 7 March 2013. Available at: http://blogs.hbr.org/cs/2013/03/you_cant_just_hack_your_way_to.html (accessed 9 March 2013).

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

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

Ribes D and Jackson SJ (2013) Data bite man: The work of sustaining long-term study. In: Gitelman L (ed.) *'Raw Data' is an Oxymoron*. Cambridge, MA: MIT Press, pp. 147–166.

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

Rose G (1997) Situating knowledges: Positionality, reflexivities and other tactics. *Progress in Human Geography* 21(3): 305–320.

Ruppert E (2013) Rethinking empirical social sciences. *Dialogues in Human Geography* 3(3): 268–273.

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

Seni G and Elder J (2010) *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. San Rafael: Morgan and Claypool.

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

Steadman I (2013) Big data and the death of the theorist. *Wired*, 25 January 2013. Available at: http://www.wired.co.uk/news/archive/2013-01/25/big-data-end-of-theory (accessed 30 January 2013).

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.

Strom D (2012) Big data makes things better. *Slashdot*, 3 August. Available at: http://slashdot.org/topic/bi/big-data-makes-things-better/ (accessed 24 October 2013).

Trumpener K (2009) Critical response I. Paratext and genre system: A response to Franco Moretti. *Critical Inquiry* 36(1): 159–171.

Wyly E (in press) Automated (post)positivism. *Urban Geography*.

Zikopoulos PC, Eaton C, DeRoos D, et al. (2012) *Understanding Big Data*. New York: McGraw Hill.