Title
Big Data: why (oh why?) *this* computational social science?

Author
O'Sullivan, David

Publication Date
2017

Peer reviewed

Big Data: why (oh why?) *this* computational social science?

David O'Sullivan,

Department of Geography,

University of California, Berkeley

dosullivan@berkeley.edu

Many others, in this volume and elsewhere, have and will comment on the political, social, economic, cultural and ethical implications of Big Data.(1) I strongly agree that those implications are important. Indeed, I believe they are the most urgent aspects of Big Data with which critically engaged social science must grapple. There is a good fight to be fought in the public arena over the many worrying directions in which the political-economic impulses driving Big Data are pointing. The complicated ways in which the Big Data movement (if we can call it that) is entangled with a burgeoning, authoritarian and surveillant state, simultaneously enabling and in thrall to a rhetoric of free-market 'disruption', demand our close attention, if they are to be countered by more humane alternatives. There is, of course, a substantial literature that refuses to roll over before the technological triumphalism (or is it fatalism?) of work such as *Too Big To Know* (Weinberger, 2011) and *Big Data* (Mayer-Schönberger and Cukier, 2013). Many of the concerns raised by the current moment are ably dissected in *Digital Disconnect* (McChesney, 2013), *The Filter Bubble* (Pariser, 2011), *You Are Not A Gadget* (Lanier, 2010) *Who Owns The Future?* (Lanier, 2014), and *To Save Everything, Click Here* (Morozov, 2013) among others. Given the centrality of geographical data of one kind or another to the data deluge, it is surely important that geographers become more visible in this public conversation (Farmer and Pozdnoukhov, 2012). Substantial contributions like *Code/Space* (Kitchin and Dodge, 2011)

remain firmly academic in tone, but nevertheless provide a foundation for future contributions that tackle more specifically spatial aspects of Big Data and its impacts.

Recognizing the importance of these wider debates, I nevertheless want to focus on narrower methodological concerns. Whatever else Big Data has accomplished, it has placed quantitative and computational methods firmly on the social science agenda (Barnes, 2009; Barnes and Sheppard, 2009; Burrows and Savage, 2014; Johnston et al., 2014; Wyly, 2009). I welcome that development, and see it as providing an opening for a more plural vision of geography and other social sciences. However, it is unlikely that opening will lead us anywhere new if we persist in understanding Big Data as primarily about the novelty of the data themselves. Data however 'big' are severely limited in how they represent *processes*. Given the centrality of process to developing any sophisticated understanding of how the world works, this is more than a limitation of Big Data. If understanding, explaining and effectively intervening in the world are the goals then we must ask questions about the style of computational social science we ought to be aiming for. Yet there is every sign that (over-)excitement and hype around Big Data is in danger of causing us to lose sight of such matters. This would be unfortunate both for opponents *and* proponents of the potential of Big Data for social science; drawing attention to these issues is therefore my aim.

In the next section, I set out my understanding of the epistemology of Big Data, and suggest why Big Data has been so successful; successful, that is, as a widely adopted technology, not necessarily as a way to understand the world. From there, I move on to consider a persistent dualism in how computational tools have been deployed in the sciences, namely a distinction between top-down, aggregate or statistical approaches to explanation (among which Big Data can be placed), and bottom-up, emergentist approaches, often associated with complexity science

(Coveney and Highfield, 1995; Manson, 2001; Mitchell, 2008; O'Sullivan, 2004; Thrift, 1999; Waldrop, 1992). While these two traditions share substantial elements in their intellectual heritage, they yield sharply divergent perspectives on explanation, understanding, and prediction, and suggest very different intellectual styles and methodological directions for computational social science. Further, the two traditions tackle the central issue of process very differently. The current openness to greater use of computers and (secondary) quantitative data is an opportunity for better, more effective social science that we are in danger of missing if the Big Data paradigm remains dominant. That danger has both scientific and ethical dimensions, the latter returning the argument to the broader context considered at the outset.

The mysterious rise of Big Data

Big Data has seemingly come out of nowhere, very quickly, but this is illusory. Iconic magazine covers on the topic such as *Nature*'s 'Big Data: Science in the Petabyte Era' (September 2008) and *The Economist*'s 'Data Deluge' (February-March 2010) popularized the term 'Big Data', but were testament to developments already well under way. Even so, the speed with which such a media-detected (and inevitably, amplified) 'trend' has morphed into a prime directive for all of science has been surprising. Living in New Zealand until the end of 2013 somewhat shielded me from this juggernaut, but even there, by early 2013, Big Data was unavoidable as a series of Royal Society of New Zealand and National Library sponsored discussions on the topic broadcast by the state-funded National Radio NZ makes clear.[(2)] As has happened elsewhere, national science funding and infrastructure initiatives were quickly hitched to this latest, urgent strategic imperative, opening profitable opportunities for private companies building New Zealand's Ultra-Fast and Rural Broadband Initiatives. As has happened in many other jurisdictions these developments are often explicitly connected to the parallel evolution of

'smart cities'.[3] The details of New Zealand's experience of the Big Data and smart cities movement are geographically, politically and culturally specific (as Kitchin, 2014 reminds us we should expect), but the commonalities with what is unfolding elsewhere are striking.

The New Zealand case immediately brings to the fore the oft-discussed question of what exactly it is that makes Big Data 'big'. Certainly, the *volume* of data generated in New Zealand would not qualify as 'big' in many other contexts. The by now overly-familiar three 'V's—volume, velocity, and variety—supposedly definitive of Big Data, were appropriately enough purloined from a business intelligence report (Laney, 2001). They have proven insufficiently descriptive for many tastes, leading to an academic cottage industry proposing and debating additional attributes (preferably ones starting with the letter 'V'). Rather than add to that debate, I consider Big Data to be primarily an epistemological stance (boyd and Crawford, 2012). That stance, can be crudely sketched as a claim that, given sufficient data, the world can be known (if not completely, then well enough for any particular purpose), and that we are currently on the threshold of an era where the technological capacity to assemble such complete datasets has (at last!) arrived (Anderson, 2008). As such, this moment heralds the realization of a dream of certain kinds of deterministic positivism (Wyly, 2014b). This is, of course, philosophically extremely shaky ground (Pigliucci, 2009), although it appears that epistemological difficulties are not a deterrent to adoption of the approach. In many contexts, where Big Data are deployed the point is not, after all to understand the world, but simply to know it well enough to make a profit (Wyly, 2014a); or, an even lower bar, to be plausibly able to claim that a profit might some day be made using insights gleaned from data (Wilson, 2012). Suffice to say, along with many geographers, while intrigued by the possibilities such datasets may open up, I am unpersuaded by the grandiose claims made for Big Data. Understanding the world still demands that we carefully

develop theories, consider the implications of those theories for what we expect to observe in the world, and subject those expectations to scrutiny through empirical observation, using multiple methods, only a few of which are enhanced by the dragnet of Big Data collection.

In spite of its unconvincing epistemological claims, how is it that this particular computational approach to studying the world has come to dominate so much recent thinking about using computation to learn about the world? At least three answers come to mind. First, it (retroactively) justifies data collection that rests on questionable ethical and legal foundations. Never mind how we came to be in possession of these vast data repositories, just think about what we can do with them! A case of ends justifying means on a societal scale.

Second, the Big Data techno-social regime is feasible in a context where collecting more data became necessary as a matter of everyday business practice. It is not clear that what has emerged was deliberately planned by any of the leading protagonists. For example Brin and Page in an early paper on the Google search engine note that "we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers" (Brin and Page, 1998, page 18), suggesting that building a data gathering, advertising company was not their original intention. However, in a business environment where building an audience took priority over the difficult task of selling often not obviously useful services, it was imperative for those services to be free at the point of use, leaving targeted advertising and the attendant surveillance as one of the few viable, sustainable business models. [(4)] This path is one premised on a financial speculation that the data will eventually pay off (Wilson, 2012), but regardless of the eventual correctness of that speculation, once a company starts down this path, more data about users can only be better, and available technology has made the assembly of vast data sets possible.

Third, from the perspective of making a profit, there is little doubt that Big Data *can* work. Indeed, from this perspective, profit (or more generally efficiency) is the only metric that matters. “The capitalist correlation imperative is clear: spurious correlation is fine, so long as it is *profitable* spurious correlation” (Wyly, 2014a, page 681 emphasis in the original). This leads to a stance on knowledge that is unconcerned with explanation: just as I don't need to understand how my phone works to use it, corporations and governments don't need to understand how or why the algorithms they use work to operate them, or at any rate, not until they plainly *don't* work (see Flood, 2011; and earlier work by Wallace and Wallace, 2001 on which he draws).

Big Data as method

So much for a schematic explanation of the rise of Big Data. In practical terms, what does Big Data *as method* consist of? Given the loose way in which the term is used, it is not easy to pin this down, but given the salience of the quantity of data in the approach, the emphasis is on large scale data analysis methods, of various statistical kinds. It bears emphasizing that this orientation immediately places data ahead of theory, since data and the world they are assumed to illuminate come before any consideration of the questions to be addressed. In any case, speaking very generally, statistical models are fit to data to identify factors accounting for variations in the data. To be sure a wider range of methods are available now than at any earlier time, including machine learning, data mining and pattern recognition methods, alongside more exploratory approaches, particularly interactive visualization (see Miller and Han, 2009 for a survey of these developments in geography). In an ideal case, a single researcher, analyst, or small team of analysts might use exploratory visualization methods to develop familiarity with a dataset. This in turn might prompt some ideas about the patterns there to be found, the methods most likely to emphasize those patterns, and from there the statistical models most suited to

advancing understanding of the phenomena represented by the dataset. For many scientists working today with much larger datasets than ever before, this is a reasonable description of how they would proceed. It is also not very new or different from how they would have proceeded in the past.

But data volume and velocity *do* matter. When datasets are very large and rapidly changing, then the scope for an exploratory approach is limited, since the computational demands are potentially limitless. Problems must be rendered tractable by *pre-defining* what constitutes a pattern of interest in terms of the data known to be available. In corporate, or other environments where timely, 'actionable intelligence' is prized, much more constrained, automated approaches are likely to prevail. In these contexts, much of the decision-making about what patterns to attend to must be delegated to the diagnostic statistics generated by whatever methods are deployed. In a visionary piece, Stan Openshaw (1994) (with some excitement) anticipated the type of continuous monitoring that this approach implies, along with a major drawback: the identification of many spurious patterns and correlations. Such considerations demand that the criteria specifying what is interesting and what is not be narrowed further, forestalling the open-ended search for patterns that might inspire the collection of extensive data in the first place.

It would be absurd to argue that there is no potential in more detailed, more frequently updated data, for describing better how social change unfolds over time. Perhaps previously unknown social phenomena can be observed as a result of the improved temporal resolution in such data. Loose analogies with the advances made possible in other fields by the invention of the telescope or microscope (Brynjolfsson and McAfee, 2011) do not seem completely misguided. Particularly when used in conjunction with other approaches, there are surely grounds for (guarded) optimism about the social scientific possibilities of Big Data. In a

specifically geographical context, Miller and Goodchild (2014) identify some of what is exciting about this approach, and point to interesting continuities with previous work in quantitative geography. Geography's longstanding challenge of bridging from the local and particular to the global and general is central to their argument that when thoughtfully and carefully deployed the Big Data approach holds promise for a 'data-driven geography' particularly in the abductive early 'discovery' phase of research when the aim is to develop good ideas and candidate explanations.

Even so, a larger point here is that the methods associated with Big Data start from the aggregate level and deploy statistical methods to identify relationships among data attributes, in more or less traditional ways. The mode of explanation is inferential statistical, based on a constant conjunction model of causality, rather than on a realist, mechanistic or process-oriented account (Sayer, 1992). Contemporary large datasets, particularly those which are frequently updated, give an impression of dynamism and by extension may be considered to offer us a rich representation of process. In truth this is little more than an impression. Large datasets, even frequently updated ones embody no concept of process. Sometimes, it is implied that the *velocity* of such datasets, their currency, and frequency of update somehow captures process. In fact, most such data are simply rapidly updating snapshots of events. Nothing recorded in the data captures the processes or mechanisms that drive the changes occurring in the data. Process and change are thus rendered as 'one damn thing after another' with no notion of process or mechanism in the data themselves.

Instead, data impose rather rigid concepts of identity on people and places, reducing process from 'becoming' to mere change in attribute values associated with otherwise unchanged social entities (whether individuals or institutions). This distinction, and the need for a different approach to data that taking process, time and change seriously entails has been a focus in the

geographical information science community for many years (O'Sullivan, 2005). It is hard to conceive of any means by which process can be 'retrofitted' to Big Data *as data*. Rather, it is in creative use and practice, through analysis from a theoretically informed perspective that concepts of process are 'added' to data. Ironically, such explicitly theoretically informed analysis is one of the approaches most loudly eschewed by the more aggressive advocates of Big Data.

Bifurcated computation

Regardless of its epistemological limitations and process-blindness, Big Data is clearly in the ascendant, at least for now. For my present purposes, it is instructive to consider Big Data as the latest development in the deployment of computation in commerce and government, and to pay particular attention to *alternative* computational approaches to understanding the world, not currently so strongly favored. The origins of modern computation lie in the Second World War (Ceruzzi, 2003; Dyson, 2012). That context saw computers and the closely related field of operations research applied to the solution of the practical production and logistical challenges of mounting modern warfare on a large scale (DeLanda, 1991 offers a critical historical overview; while Shrader, 2006, in a three volume official history, gives a feel for the scope). Of particular importance to the development of computing were the demanding mathematical problems that arose in these contexts, such as code-breaking, complex optimization problems, and the simulation of nuclear reactions.

From their origins, we can identify two broad applications of computers. First, are applications of computation to datasets too large for calculation by hand, to produce closed-form solutions to mathematically well-defined problems, using various types of numerical analysis. Such calculations rely on algorithms for manipulating large matrices, on interpolation and approximation methods, and on the mathematics of linear algebra and optimization (Aspray and

Gunderloy, 1989). While routine, such computation is more demanding of computing resources particularly as datasets grow in size, when the associated computational requirements may grow with the square, cube or even higher-order powers of the problem size (Cormen, 2013). In principle, such calculations are not difficult, but they are computationally intensive. This domain of computation is associated with the more efficient management of logistical systems, the optimization of resource allocations in productions systems, and the field generally known as operations research (Light, 2003). The Big Data phenomenon sits squarely in this tradition.

A second broad area of applications can be identified where computation is iteratively applied to prospectively simulate real or hypothetical systems. Here the applications themselves may not be very data intensive, but repeatedly iteratively performing (often) simple calculations over time and/or space leads to substantial computational demands in the aggregate. Such computation is deployed in simulating target systems of interest, in meteorology, product design, military applications (flight and combat simulators), and perhaps most familiar to a general audience, in computer gaming (Crogan, 2011).

It is useful to consider these two computational approaches as they have played out in a particular discipline (geography) to provide a more specific account of the differences between them. Geography's quantitative revolution witnessed its own somewhat related bifurcation, in the divergence between theoretical model-oriented methods (Chorley and Haggett, 1967) and more pragmatic applications of inferential statistics to primary and secondary data, typified by the coverage of texts such as Leslie King's *Statistical Analysis in Geography* (King, 1969). It was not long before the peculiar challenges of spatial data complicated and compromised the latter enterprise considerably (Gould, 1970). Meanwhile, beyond a few specialized areas and subfields such as urban modeling, the model-oriented approach found only limited acceptance, before the

dramatic epistemological upheavals of the 1970s and 1980s. As Thrift notes, this “ghettoizing of complexity theory in geography was a tragedy, since the potentialities for much wider interaction were there.” (1999, page 60 note 2). The marginalization of complexity oriented approaches in geography, within *quantitative* geography, is instructive, because it emphasizes the extent to which any attempt to map methodological approaches onto political or other predispositions is doomed to failure.

Even from a computational perspective, the distinction I am drawing is somewhat artificial, since many of the same computational tools and algorithms are equally applicable to either Big Data or complexity science, the point of general purpose computing being precisely its mutable, reprogrammable nature. The truth is more complicated and nuanced than any simple binary account would suggest. On the one hand, simulation depends on repetitive, often routine calculation, not achievable by hand. On the other, applying closed-form solutions to *small* datasets can enable iterative and interactive exploration of many possible solutions, and lead from there to the concept of a solution space, and ultimately, to a more nuanced understanding of the original problem. When numerical analysis is applied in this way it can transform the questions asked of data. This more exploratory stance toward datasets has initiated the trend toward interactive visualization of larger, more complicated datasets.

Two cultures of computation?

It is tempting to map these two computational styles (closed solutions and open-ended exploration) directly onto two cultural manifestations of computing: authoritarian, corporate, statist, big brother, Big Data on the one hand, and liberatory, individual-empowering, personal computing on the other.(5) The dualism is deeply etched into many accounts of the history of computing (see, for example Barbrook, 2007; Levy, 1984 or consider the 1984 Apple Superbowl ad; see Golumbia, 2009). The duality is particularly emphasized by self-styled 'revolutionary' or

'disruptive' Silicon Valley startups, deploying their own version of the complex cultural politics of the post-1960s counter-culture (Frank, 1998) These contradictions are brilliantly dissected by Fred Turner (2008) in his *From Counterculture to Cyberculture*, and many of the contradictory oddities of high tech's self-consciously liberal (often libertarian) yet conservative elite are entertainingly recounted in Pauline Borsook's enduringly relevant *Cyberselfish* (2000).

A similarly odd clash of cultures is evident in the contrast between a 'new age' holism in the language and iconography around chaos and complexity science, and the more authoritarian, establishment and business agendas both funding and consuming this science (Thrift, 1999). Thus, one important center of complexity science has been the Santa Fe Institute (SFI), which according to Helmreich "is sometimes considered the good twin of Los Alamos, concerned with the technology of life, rather than the technology of death" (1998, page 43),[(6)] and established in part through the efforts of George Cowan a former director of the Los Alamos laboratory (see Waldrop, 1992, pages 53–69). Or again: "[m]ost scientists at SFI are wary of any association with New Age movements" (page 41); and yet books such as Stuart Kauffman's *At Home in the Universe* (1995) struggle (for the most part failing) to stay on the scientific side of a surprisingly fine line between 'new age-y' flakiness and detached scientific rigor, when it comes to the more grandiose claims of complexity science. Such odd intellectual (if not cultural) contradictions may be the inevitable outcome when reductive scientific methods, so successful at explaining phenomena that are timeless at human scales (e.g., the Solar System, evolution), are applied to systems where historical modes of explanation with their attendant contingencies and chance events have been predominant.

A simplistic mapping of the cultural origins of the two computational 'traditions' under discussion onto particular political or economic agendas is plainly unsustainable. Science and

technology studies in numerous fields have repeatedly and convincingly demonstrated the highly contingent nature of the relationships between technologies and the politics they embed and produce (Feenberg, 1991; Latour and Woolgar, 1979; Winner, 1986). So, while it is tempting to suggest that complexity-oriented, bottom-up modeling is inexorably associated with anti-authoritarian and more open approaches to knowledge, while Big Data, top-down, classificatory and inferential statistical approaches are aligned with powerful interests, it is demonstrably untrue. There is nothing intrinsic to either approach that determines the ends to which they can or should be deployed. Closed-form calculation might be used to optimize the efficient production and equitable distribution of medical or other public services, while simulation can be (and almost certainly has been) used to explore possible strategies for the illegal invasion and occupation of another country.

Contrasting computational epistemologies

The lack of a one-to-one mapping from computational approach onto particular political or economic agendas notwithstanding, it should nevertheless be clear that these two broad approaches as distinctive scientific practices embed different thinking about process. They also each sit more easily with contrasting attitudes to the use of computation in furthering understanding of socioeconomic, political and cultural systems. Indeed it is my argument here that we ought to *choose* which approaches to the use of computation are more likely to advance our understanding of the world, and adopt them for that reason. It is likely that such pragmatism would see Big Data de-emphasized in favor of complexity models and other computational approaches more attuned to process and explanation, as in the digital humanities (Burdick et al., 2012). This contrast between complexity science and Big Data is schematically elaborated in Table 1.

Table 1 The differing methodological, representational and epistemological approaches of complexity science and Big Data

Complexity science	**Big Data**
Embeds theory in models	Correlation and classification
Represents process	Temporal snapshots
Open-ended exploration of process implications	Exploration of already-collected data
Bottom up	Top down
Multiple levels and scales	Two levels: aggregate and individual
Many alternative histories (or futures)	'Just the facts' (or optimal solutions)

Complexity-oriented model-based approaches are precisely about process. It is a representation of process (however limited) that drives the dynamics of such models, and open-ended investigation of model behaviors can be considered as an exploration of the conditions of possibility of the system being modeled. Before getting too excited about this it should be acknowledged that in many cases the notion of 'process' embodied in such models is not greatly richer than that implied by the historical snapshots of Big Data. Change is most often cast in terms of changing attribute values of otherwise fixed and stable entities. Nevertheless, the focus is on change, and the circumstances that produce change, a perspective that forces users of models to consider processes and mechanisms directly. Interesting model structures that combine both attribute change and systemic structural change (Gross and Blasius, 2008) are one possible advance in this regard.

Open-ended exploration of dynamic models[7] engenders a different, more humble and provisional attitude to knowledge, compared to pre-defining and then identifying 'optimal' solutions or patterns of interest. Simulative computation, posits 'possible worlds' (in the form of simulation results under different scenarios or model configurations) and implicitly acknowledges the speculative nature of the exercise (Casti, 1997). Speculative computer modeling in this vein has led to the recognition in parts of the mathematical sciences (and more widely) of limits to knowledge and prediction, in the form of dynamic effects such as chaos (Gleick, 1987), and properties such as emergence, path dependence, positive feedback and adaptivity, all of which are likely to preclude reliable prediction of a system's behavior over time. The recognition of these system characteristics under the banner of complexity science (Coveney and Highfield, 1995; Manson, 2001; O'Sullivan, 2004), calls, at least potentially, for greater humility on the part of scientists, and a recognition of limits to knowledge *inherent in the nature of the systems under study* (Cilliers, 1998; Richardson et al., 2001). Understood this way, complexity science underwrites a pluralistic approach to knowledge, that acknowledges the importance of understanding systems at multiple levels, from multiple perspectives, and using a variety of methods (Harvey and Reed, 1996). Recognizing that social systems are composed of complex individuals, organized into households, at the same time playing multiple roles in a range of institutions of varying organizational structure, that themselves have a range of aims and goals, when seen through a complexity science lens ought to provoke realization that *no* single-level, top-down understanding of how society works will do; and furthermore, different methods are likely to be appropriate for getting at what is going on in each of these diverse contexts (Manson and O'Sullivan, 2006).

It is important to note that Big Data and complexity science are not as far removed from one another as they at first appear. Both are about fitting simple models to observations: statistical models derived from observational data on the one hand, synthetic simulation models on the other. At the same time, the important differences in epistemology sketched here, are real, particularly if care is not . A complexity-oriented modeling approach to knowledge allows us to think of data not as hard, precise evidence of reality, but as a set of patterns that constrain a space of plausible, speculative models whose structure and mechanisms can account for those patterns (Grimm et al., 2005), and which may therefore be useful in building process-oriented, theoretical, explanations for the existence of those patterns. Data in this context become an intermediate step in the development of explanations. By contrast, under the model most often adopted in the world of Big Data, data themselves *are* the phenomena, and explanation is less about understanding processes and mechanisms—that is, *explaining* the world—and more about describing the data, at which point the phenomena themselves are taken to have been understood.

Conclusions

As I have tried to show an orientation to process is absent from the epistemology of Big Data, yet is surely central to any coherent approach to explanation in geography and the social sciences more generally. Other computational approaches offer more in this regard, but have been less prominent in recent years. As much as anything this may be symptomatic of intellectual 'fashions' in science. Chaos and complexity theory both had their own times 'in the sun', and complexity remains a much-trafficked buzzword. Perhaps, these approaches were tried but failed to deliver on their initial promise, as I am suggesting Big Data is likely to? There may be some truth in this view, although it depends on an odd, fashion-conscious perspective on how we should evaluate scientific method. More seriously, it fails to appreciate how great a challenge

taking seriously the uncertainties introduced by complexity poses, for conventional modes of scientific explanation. The complexity enterprise points to a much wider remit for historical and narrative modes of explanation, which go against the grain of dominant modes of scientific explanation. The tension is exacerbated by the 'simple rules, complicated behavior' mantra so often used to 'sell' complexity science, which present simple models as the end point of complexity-oriented approaches, when, really they are only the beginning. Just as it is foolish to believe that data-mining Big Data can provide answers to every social science question, it would be foolish to argue that simple complexity science models can answer every question.

Recognizing and valuing pluralism in methods is key to the complexity-oriented computational approaches I favor. That implies two things. First, that there is, of course, a place for Big Data (Miller and Goodchild, 2014). It would be absurd to argue otherwise. Without doubt, when contemporary datasets and data-mining methods are applied to questions of genuine social scientific interest, new phenomena will be identified, and new perspectives on old questions will emerge. But *understanding* those new phenomena will demand approaches other than those of Big Data. Which leads immediately to the second point, that other approaches to geography and social science remain vital to any coherent way forward. What is disturbing about much of the hype around Big Data is the apparent desire to advance on all fronts simultaneously: Big Data, not content with being a 'revolutionary' approach to social science, must become a whole 'system for living', a societal lifestyle choice, a new mode of governance, of business, and of science.

The tragedy is that this stance toward Big Data could easily discredit all computational approaches to the social world. Not only by getting the science wrong (Lazer et al., 2014), but also by becoming a pervasive social surveillance system, the necessity for which is unclear,

beyond the desire for profit of large corporate interests, and the data anxiety of a surveillant state (Crampton, 2014; Crawford, 2014). The psychology of Big Data holds out an entirely false promise: that if only the data were bigger, we would know even more. There are certainly contexts where this might be true (astronomy's Square Kilometre Array, for example, see Taylor, 2012), but they are not social ones. Much of what is revealed by social Big Data we either already know, or is accessible in other ways that can place human actions and decisions in much richer social contexts. It seems likely that we would lose very little of genuine scientific interest by *not* recording and storing every person-machine micro-interaction. There are no scientific grounds for 'collecting it all', only commercial imperatives (and even those are founded on a wild speculation), which returns us to the important political-economic issues I opened with.

There is a danger in focusing as I have on method, on means over ends. It is ethical positions, not methodological choices that most affect the impact of research (Lake, 2014). Ultimately, the two cannot be disentangled, and if the undoubted potential of computational methods in the social sciences is to be realized it is important that we learn what can be learned from past mistakes, recognize the limitations of all our data, and focus on developing computational approaches to geography and social science better aligned with handling those problems. Such a social science will not only be better science *qua* science, but will also be more ethically defensible as a direct result of recognizing the explanatory limitations of data. In sum, we should not only be challenging the political economy of Big Data, we ought to be deeply (and vocally) suspicious of its epistemology, not only from within critical traditions skeptical of quantification anyway, but from the perspective that more interesting quantitative and computational methods are available.

Notes

1. Big Data is a troublesome phrase to use correctly given its willful (and annoying) mixing of a singular adjective and a plural countable noun—'numerous data' would have been a more accurate, if less compelling coinage. Further, the phrase 'Big Data' has come to stand for a complex mix of technologies, ideas and practices, such that it may be considered a singular noun phrase. In a perhaps futile attempt to hold back the tide of offenses against grammar, where I consider the meaning to be 'many data' I treat 'data' as plural, whereas when the meaning is Big Data (the idea), I treat it as singular.

2. See http://www.radionz.co.nz/national/programmes/bigdata (accessed June 2, 2015) where the series is archived at the time of writing.

3. For information about the infrasructure investments, see http://www.med.govt.nz/sectors-industries/technology-communication/fast-broadband (accessed June 2, 2015). Some sense of the high priority given to Big Data in national science and innovation priorities is provided by the appointment and reports of the New Zealand Data Futures Forum http://www.nzdatafutures.org.nz/ (accessed June 2, 2015).

4. Some insight into how such intentions drifted is provided by Zuckerman (2014).

5. It is particularly tempting if you live in Northern California, perhaps even more so if you have only recently moved there!

6. Los Alamos is, of course, home to one of the United States National Laboratories most closely associated with past and ongoing development of nuclear weapons.

7. Of course, model building science is not always an open-ended exercise. A considerable amount of economic theory is built on reversing the sequence of model-building and exploration leading to the refinement of testable theory. Instead, models provide rhetorical

support for theoretical positions already firmly established, and are refined to fit theory, rather than the other way around. In the process the models become more, not less divorced from reality (Lawson, 1997; Keen, 2011). I am concerned here with an approach to the use of simulation models, discussed in my recent book (O'Sullivan and Perry, 2013). The proper use of simulation in science is a difficult philosophical area, which remains under-explored by philosophers of science (but see Winsberg, 2010; Weisberg, 2013).

References

Anderson C, 2008, "The end of theory: the data deluge makes the scientific method obsolete" *Wired* **16**(07) 2-6.

Aspray W, Gunderloy M, 1989, "Early computing and numerical analysis at the National Bureau of Standards" *Annals of the History of Computing* **11**(1) 3-12.

Barbrook R, 2007 *Imaginary Futures: From Thinking Machines to the Global Village* (Pluto Press, London).

Barnes T J, 2009, "'Not only … but also': quantitative and critical geography" *The Professional Geographer* **61**(3) 292-300.

Barnes T J and Sheppard E, 2009, "'Nothing includes everything': towards engaged pluralism in Anglophone economic geography" *Progress in Human Geography* **34**(2) 193-214.

Borsook P, 2000 *Cyberselfish: A Critical Romp through the Terribly Libertarian Culture of High Tech* (Public Affairs, New York).

boyd d and Crawford K, 2012, "Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon" *Information, Communication & Society* **15**(5) 662-679.

Brin S and Page L, 1998, “The anatomy of a large-scale hypertextual Web search engine” in: *Seventh International World-Wide Web Conference* (WWW 1998), April 14-18, 1998, Brisbane, Australia(Brisbane, Australia). Available at http://ilpubs.stanford.edu:8090/361/ (accessed June 2, 2015).

Brynjolfsson E and McAfee A, 2011, “The Big Data Boom Is the Innovation Story of Our Time” *The Atlantic.* Available at http://www.theatlantic.com/business/archive/2011/11/the-big-data-boom-is-the-innovation-story-of-our-time/248215/ (accessed June 2, 2015).

Burdick A, Drucker K, Lunenfeld P, Presner T and Schnapp J, 2012 *Digital_Humanities* (MIT Press, Cambridge, MA)

Burrows R and Savage M, 2014, “After the crisis? Big Data and the methodological challenges of empirical sociology” *Big Data & Society* **1**(1) 1-6.

Casti J L, 1997 *Would-be Worlds: How Simulation is Changing the Frontiers of Science* (John Wiley & Sons, New York).

Ceruzzi P E, 2003 *A History of Modern Computing* second edition edition (The MIT Press, Cambridge, MA).

Chorley R J and Haggett P eds, 1967 *Models in Geography* (Methuen, London).

Cilliers P, 1998 *Complexity and Postmodernism: Understanding Complex Systems* (Routledge, London).

Cormen T H, 2013 *Algorithms Unlocked* (The MIT Press, Cambridge, MA).

Coveney P and Highfield R, 1995 *Frontiers of Complexity: The Search for Order in a Complex World* (Ballantine Books, New York).

Crampton J W, 2014, “Collect it all: national security, Big Data and governance” *GeoJournal*, http://link.springer.com/article/10.1007/s10708-014-9598-y (accessed June 2, 2015).

Crawford K, 2014, "The Anxieties of Big Data" *The New Inquiry.* Available at http://thenewinquiry.com/essays/the-anxieties-of-big-data/ (accessed June 2, 2015).

Crogan P, 2011 *Gameplay Mode: War, Simulation, and Technoculture* (University Of Minnesota Press, Minneapolis, MN).

DeLanda M, 1991 *War in the Age of Intelligent Machines* (Swerve Editions, New York).

Dyson G, 2012 *Turing's Cathedral: The Origins of the Digital Universe* (Pantheon Books, New York).

Farmer C J and Pozdnoukhov A, 2012, "Building streaming GIScience from context, theory, and intelligence", in *Proceedings of the Workshop on GIScience in the Big Data Age 2012*, pages 5-10. Available at http://ncg.nuim.ie/content/staff/staff/downloads/apozdnoukhov/GIScience2012.pdf (accessed June 2, 2015).

Feenberg A, 1991 *Critical Theory of Technology* (Oxford University Press, New York).

Flood J, 2011 *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City--and Determined the Future of Cities* (Riverhead Books, New York)

Frank T, 1998 *The Conquest of Cool: Business Culture, Counterculture, and the Rise of Hip Consumerism* (University Of Chicago Press, Chicago, IL).

Gleick J, 1987 *Chaos: Making a New Science* (Viking Penguin, New York).

Golumbia D, 2009 *The Cultural Logic of Computation* (Harvard University Press, Cambridge, MA).

Gould P, 1970, "Is *statistix inferens* the geographical name for a wild goose?" *Economic Geography* **46** 439-448.

Grimm V, Revilla E, Berger U, Jeltsch F, Mooij W M, Railsback S F, Thulke H-H, Weiner J, Wiegand T, DeAngelis D L, 2005, "Pattern-oriented modeling of agent-based complex systems: lessons from ecology" *Science* **310** 987-991.

Gross T, Blasius B, 2008, "Adaptive coevolutionary networks: a review" *Journal of The Royal Society Interface* **5**(20) 259-271.

Harvey D L and Reed M, 1996, "Social science as the study of complex systems", in *Chaos Theory in the Social Sciences: Foundations and Applications* Eds L D Kiel and E Elliott (University of Michigan Press, Ann Arbor, MI), pages 295-323.

Helmreich S, 1998 *Silicon Second Nature: Culturing Artificial Life in a Digital World* (University of California Press, Berkeley, CA).

Johnston R J, Harris R, Jones K, Manley D, Sabel C E, Wang W W, 2014, "Mutual misunderstanding and avoidance, misrepresentations and disciplinary politics: spatial science and quantitative analysis in (United Kingdom) geographical curricula" *Dialogues in Human Geography* **4**(1) 3-25.

Kauffman S A, 1995 *At Home in the Universe: The Search for Laws of Complexity* (Penguin, London).

Keen S, 2011 *Debunking Economics: The Naked Emperor Dethroned* (Zed, London).

King L J, 1969 *Statistical Analysis in Geography* (Prentice Hall, London).

Kitchin R, 2014 *The Data Revolution* (SAGE Publications Ltd, Thousand Oaks, CA).

Kitchin R and Dodge M, 2011 *Code/Space: Software and Everyday Life* (MIT Press, Cambridge, MA).

Lake R W, 2014, "Methods and moral inquiry" *Urban Geography* **35**(5) 657-668.

Laney D, 2001, "3D Data Management: Controlling Data Volume, Velocity and Variety", http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (accessed June 2, 2015).

Lanier J, 2010 *You Are Not a Gadget: A Manifesto* (Allen Lane, London and New York).

Lanier J, 2014 *Who Owns the Future?* (Simon & Schuster, New York).

Latour B and, Woolgar S, 1979 *Laboratory Life: The Construction of Scientific Facts* (Sage Publications, Beverly Hills, CA).

Lawson T, 1997 *Economics and Reality* (Routledge, London).

Lazer D, Kennedy R, King G and Vespignani A, 2014, "The Parable of Google Flu: Traps in Big Data Analysis" *Science* **343**(6176) 1203-1205.

Levy S, 1984 *Hackers: Heroes of the Computer Revolution* (Anchor Press, Garden City, NY).

Light J S, 2003 *From Warfare to Welfare: Defense Intellectuals and Urban Problems in Cold War America* (Johns Hopkins University Press, Baltimore, MD).

Manson S M, 2001, "Simplifying complexity: a review of complexity theory" *Geoforum* **32**(3) 405-414.

Manson S M, O'Sullivan D, 2006, "Complexity theory in the study of space and place" *Environment and Planning A* **38**(4) 677-692.

Mayer-Schönberger V and Cukier K, 2013 *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Houghton Mifflin Harcourt, Boston, MA).

McChesney R W, 2013 *Digital Disconnect: How Capitalism is Turning the Internet Against Democracy* (The New Press, New York).

Miller H J, Goodchild M F, 2014, "Data-driven geography" *GeoJournal*, http://link.springer.com/10.1007/s10708-014-9602-6 (accessed June 2, 2015).

Miller H J, Han J eds, 2009 *Geographic Data Mining and Knowledge Discovery* 2nd ed (CRC Press, Boca Raton, FL).

Mitchell M, 2008 *Complexity: A Guided Tour* (Oxford University Press, New York).

Morozov E, 2013 *To Save Everything, Click Here: The Folly of Technological Solutionism* (Public Affairs, New York).

Openshaw S, 1994, "Two explanatory space-time-attribute pattern analysers relevant to GIS", in *Spatial Analysis and GIS* Eds A S Fotheringham, P Rogerson (Taylor & Francis, London), pp 83-104.

O'Sullivan D, 2004, "Complexity science and human geography" *Transactions of the Institute of British Geographers* **29**(3) 282-295.

O'Sullivan D, 2005, "Geographical information science: time changes everything" *Progress in Human Geography* **29**(6) 749-756.

O'Sullivan D and Perry G L W, 2013 *Spatial Simulation: Exploring Pattern and Process* (Wiley-Blackwell, Chichester, England).

Pariser E, 2011 *The Filter Bubble: What the Internet is Hiding from You* (Penguin Press, New York).

Pigliucci M, 2009, "The end of theory in science?" *EMBO reports* **10**(6) 534.

Richardson K A, Cilliers P and Lissack M R, 2001, "Complexity science: A 'gray' science for the 'stuff in between'" *Emergence* **3**(2) 6-18.

Sayer A, 1992 *Method in Social Science: A Realist Approach* 2nd edition (Routledge, London).

Shrader C R, 2006 *History of Operations Research in the United States Army* (Office of the Deputy Under Secretary of the Army for Operations Research, U.S. Army, Washington,

D.C), Available at http://www.history.army.mil/catalog/browse/title.html#h (accessed June 2, 2015).

Taylor A R, 2012, "The Square Kilometre Array" *Proceedings of the International Astronomical Union* **8**(Symposium S291) 337-341.

Thrift N J, 1999, "The place of complexity" *Theory, Culture and Society* **16**(3) 31-69.

Turner F, 2008 *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism* (University Of Chicago Press, Chicago).

Waldrop M, 1992 *Complexity: The Emerging Science at the Edge of Order and Chaos* (Simon and Schuster, New York).

Wallace D and Wallace R, 2001 *A Plague on Your Houses: How New York Was Burned Down and National Public Health Crumbled* (Verso, London).

Weinberger D, 2011 *Too Big to Know: Rethinking Knowledge Now that the Facts Aren't the Facts, Experts are Everywhere, and the Smartest Person in the Room is the Room* (Basic Books, New York).

Weisberg M, 2013 *Simulation and Similarity: Using Models to Understand the World* (Oxford University Press, Oxford).

Wilson M W, 2012, "Location-based services, conspicuous mobility, and the location-aware future" *Geoforum* **43**(6) 1266-1275.

Winner L, 1986 *The Whale and the Reactor: A Search for Limits in an Age of High Technology* (University of Chicago Press, Chicago, IL).

Winsberg E, 2010 *Science in the Age of Computer Simulation* (University Of Chicago Press, Chicago, IL).

Wyly E, 2009, "Strategic Positivism" *The Professional Geographer* **61**(3) 310-322.

Wyly E, 2014a, “Automated (post)positivism” *Urban Geography* **35**(5) 669-690.

Wyly E, 2014b, “The new quantitative revolution” *Dialogues in Human Geography* **4**(1) 26-38.

Zuckerman E, 2014, “The Internet’s Original Sin” *The Atlantic*. Available at http://www.theatlantic.com/technology/archive/2014/08/advertising-is-the-internets-original-sin/376041/4/ (accessed June 2, 2015).