Dialogues in
Human Geography

Commentary

Dialogues in Human Geography
3(3) 297–302
© The Author(s) 2013
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2043820613514323
dhg.sagepub.com
SAGE

Big data, little history

Trevor J Barnes
University of British Columbia, Canada

Abstract

The paper makes the argument that what is forgotten in the celebration of big data is history. Big data is presented as if it were disconnected from the past, removed from issues or problems that went before. I argue in this short commentary that the past remains potent for big data and that proponents ignore it at their peril. Rather than being a brand new approach, big data brings a series of problematic assumptions and practices first criticised 40 years ago by opponents of geography's quantitative revolution. Those assumptions, practices and criticisms are reviewed in the paper.

Keywords

big data, causation, computerization, history, models, quantification, quantitative revolution

When I was invited last autumn to participate in the panel session about big data at the 2013 Association of American Geographers annual meeting in Los Angeles, California, I had at best only rudimentary knowledge of the topic. Since then I've been inundated by data about big data. In his brilliant history of statistics, *Taming Chance*, Ian Hacking (1990: 45) speaks of the beginning of an 'avalanche of numbers'. It felt like I had been buried under an avalanche of big data.

First, there was the news of big data's role in Netflix's remake of the BBC's Richard-the-Third-inspired contemporary political thriller, *House of Cards* (Carr 2013). Netflix knew from sifting through its big data that subscribers who watched the original *House of Cards* also watched movies starring Kevin Spacey and films directed by David Fincher. Hence, Kevin Spacey played the lead role, and David Fincher directed. It couldn't go wrong, and it hasn't. In February 2013, it became the number one television series in the world except that it was not on television.[1] Then I heard the all-around American philosopher scientist, James Owen Weatherall, say on a podcast that for every individual in America there are 200 Ancient Alexandria Libraries worth of information that have been collected and stored about them.[2] The Library at Alexandria contained everything worth knowing in the ancient world. So, we now have 200 times what was worth knowing about the ancient world for every living American, and likely a few dead ones too. Finally, I just read that IBM estimated that about 2.5 quintillion bytes of data are generated each day (Yakbuksi, 2013). I had to look that one up. It is ten to the power of eighteen. As I tell my undergraduate students when I use a big number: remember the universe is only 13.8 billion years old; it's a pipsqueak amount in comparison.

Corresponding author:
Trevor J Barnes, University of British Columbia, 1984 W Mall, Vancouver, BC, Canada V6T 1Z2.
Email: trevor.barnes@geog.ubc.ca

'Big' as an adjective, then, doesn't get close to describing the size of the data sets now being analysed and manipulated. Mike Batty (2013) in a recent commentary tells of how his transportation research unit at University College London, UK, through data generated by the swipe-based Oyster travel card, has access to information of about 7 million individual daily journeys taken on London's public transportation system. That amounts to a data set of 15 billion over a 5-year period. It is a far cry, Batty notes, from the comparatively miniscule-sized data sets that made up geographers' early empirical forays into travel studies – initially using 1950s gravity models and later 1970s entropy-maximizing models.

Mike Batty implies, and this is an overriding theme in Chris Anderson's well-known *Wired Magazine* editorial on big data written in 2008, that the bigness of big data has presaged a qualitative shift. According to Anderson, we now live in a 'Petabyte Age', that is, an age of ten to the power of 15, binary 2 to the power of 50, bytes. The important point for Anderson is that the petabyte age is different 'because more is different' (2008). We live in a new world, we do things differently. This is an Apple and Google world that reaches if not to the sky, then at least to the cloud.

In so far as I can contribute to the discussion about big data, it is to caution that things are not as different as they might seem. It is not quite the brand new day that Chris Anderson supposes. In Chris Anderson's account, history drops out. It might be big data, but it is little history. This neglect of history is a typical modernist move. The past is ignored because nothing must constrain or limit what is to come. Only the bright new future matters. 'History is more or less bunk' as the iconic modernist Henry Ford once famously put it.[3] Or more explosively, as Louis Aragon said, 'Nothing is more beautiful than a church and some dynamite' (quoted in Eksteins, 2008: 23). Perhaps for Chris Anderson nothing is more beautiful than a pile of old floppy disks, hard disks and disk arrays and some dynamite. But you can't blow up a cloud.

My starting point in thinking about big data is science studies, a form of enquiry that is always keen to emphasize the centrality of the history, and more recently, the geography of science (for reviews, see Barnes, 2004; Powell, 2007). In the traditional rationalist account of science, knowledge is the product of a disembodied universal logic that has neither history nor geography. In contrast, science studies insist that scientific knowledge is situated, indissolvably connected to its history and geography. For science studies, as with Faulkner, 'the past is never dead. It's not even past' (Faulkner, 1994: 73). Even in a petabyte world, history matters. Big data comes with big history. Let me elaborate.

Big data did not first emerge in 2008 with Chris Anderson's editorial. Nor was it the invention of Google that first came on to the scene in 1998. Nor was it yet another brainchild of Steve Jobs who with two others founded Apple in 1976. One needs to go back even farther, to a pre-Apple world, difficult as that might be to imagine. One needs to reconstruct a history of big data over the *longue durée*. Big data has been made possible because of the particular conjuncture of different elements, each with their own history, coming together at this our present moment.

But precisely because these different elements have a history, the issues, problems and questions that were there in their earlier incarnation can remain even in the new form. Indeed, in some cases they become more exaggerated. I will illustrate by drawing upon some of my work on the history of geography's quantitative revolution that as a movement began in the 1950s. The practices carried out then have at least resonances with the big data paradigm. Certainly from the earliest days of geography's 1950s quantitative revolution, attempts were made to link for then relatively large sets of data to the calculative power of the computer.

For example, the early spatial scientist, William Warntz, used the Newtonian potential model to calculate spatially variable U.S. agricultural supply and demand schedules from census data (Warntz, 1959). It was a Herculean computational task, involving thousands of painstaking individual mathematical operations. He needed help. Consequently, in October 1953, Warntz approached the Princeton astrophysicist John Stewart to ask whether the 'ingenious electrical computer designed and constructed by Thomas B Bissett' that Stewart had used for his own calculations of geographical potential 'be made available to him?'[4] Unfortunately, it couldn't on that occasion. But it wasn't long before Warntz had access to a computer, a machine that came to define his research and career, as well as geography's quantitative revolution.[5]

Or again, soon after Warntz had first made his request to use 'the Bissett', a number of U.S. state universities began acquiring computers of their own including in 1955 the University of Washington, Seattle, which became one of the key 'centres of calculation' within geography's quantitative revolution (the computer's arrival was announced in *The Professional Geographer* by the Chair of Geography, Donald Hudson in 1955). In fall 1956, Waldo Tobler remembers being in the attic of the chemistry building at the University of Washington, Seattle, where the university's IBM 650 had been installed. Because of its high demand, Tobler could use the computer only during the graveyard shift, its crude design permitting in retrospect teeny rather than big data. As Tobler recalls:

> To cover programming on the [IBM] 650 you had to pick up two bytes of information on one rotation of the drum. It had a 2K memory which rotated real fast. And if you were clever, you could pick up two pieces of information in one rotation.[6]

The quantitative revolution in geography did many good things for the discipline, introducing scientific theory, making human geography one of the social sciences, giving the subject intellectual respectability and kudos. But that revolution was also criticised on a number of grounds, and some of those grounds I want to suggest continue to apply to the *über* version of the quantitative revolution that is big data.

The first is that computational techniques and the avalanche of numbers become ends in themselves, disconnected from what is important. That is, techniques and numbers become fetishized, put on a pedestal, prized for what they are rather than for what they do. This was certainly a criticism of geography's quantitative revolution. Even by the early 1970s David Harvey (1972: 6) thought:

> [Geography's] quantitative revolution has run its course and diminishing marginal returns are apparently setting in as . . . [it] serve[s] to tell us less and less about anything of great relevance . . . . There is a clear disparity between the sophisticated theoretical and methodological framework which we are using and our ability to say anything really meaningful about events as they unfold around us.

This goes to the important distinction between data and knowledge. Clearly, big data has information coming out of its ears, but is it generating useful knowledge? Do we now collect data for data's sake? Because it is there. Because we can. Nate Silver (2012: 250), the big data statistician who successfully predicted the November 2012 U.S. election results in each of the 50 states said recently, 'most of the data is just noise, as most of the universe is filled with empty space'. My fear is that big data will increasingly produce noise. But because its output comes in mathematical form, and since this is the hallmark of science ('mathematics is nature's language' as Galileo said), it will be touted as knowledge. And all the while the world is going to hell in a handbasket.

Second, for big data to have purchase, information needs to be converted into numbers, into data. Francis Galton, the Victorian inventor of regression analysis, famously said, 'Whenever you can, count'. That injunction has been raised to a significant power in the current regime of big data. The corollary is that if it can't be counted, it can't be included. Specifically, what is often lost is context, which cannot be put into an equation (Boyd and Crawford, 2012: 671). This was one of the warnings of humanistic geographers who attacked geography's quantifiers during the 1970s (Ley and Samuels, 1978). They argued that geographical context is frequently left out in quantitative studies because it cannot be expressed in numerical form. But its omission can produce distorted, misleading, even tragic outcomes. An extreme example comes from the Vietnam War. The U.S. Secretary of Defense, Robert McNamara, who in many ways practised big data before the term was invented, believed that war could be won by statistical analysis. Context was put to one side. Central was only the numbers: body counts, kill ratios, size of captured armaments, bomb target hits and so on. The war was defined for

McNamara by spread sheets of figures, victory to be achieved by statistical manipulation. The larger historical and geographical context was unnecessary. Despite an earlier meteoric career based precisely on figuring large numbers, McNamara eventually concluded war could not be counted. As he later admitted, context came to overwhelm his efforts.[7] The Vietnam War never computed, the numbers never added up.

Third, in the big data view, the numbers are the story, shorn of the need of any interpretation, shorn of the need of even assigning causation. Chris Anderson, in his 2008 editorial, says, 'with enough data, the numbers speak for themselves'. That is, the data alone determine knowledge without interpretation or causality. Anderson writes, 'Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all' (2008). This is data determinism with a vengeance. But precisely that view was attacked in geography during the late 1970s and early 1980s by Andrew Sayer (1984) from a critical realist perspective. For Sayer, numbers are never innocent, speaking for themselves, but always come marked by prior theorization: they are theory laden. Numbers do not speak for themselves but speak only for the assumptions that they embody. Numbers emerge only from particular social institutions, arrangements and organizations mobilised by power, political agendas and vested interests. In addition, the correlations that Anderson touts are never enough. Correlation does not answer the big question that social science most wants answering: why do things happen? To clarify, the so-called 'why questions', Sayer argued, can't be addressed by enumerating correlation coefficients that record merely contingent relations among events, 'if a, then b'. They do not address why 'a' and 'b' are related. This requires a causal explanatory framework that attends to the powers and liabilities of phenomena and the causal mechanisms that connect them. To suppose that correlation coefficients substitute for causation is to delude ourselves and others.

The last point is that big data as a project is inherently conservative. By utilizing the numbers as they are given big data is stuck with what is rather than what should be. As an issue this was central for critics of geography's quantitative revolution. Harvey's (1972) work already discussed made exactly this point. Arguing that 'the *status* is nothing to *quo* about', Harvey (1973: 95) called for (a 'revolutionary') theory that did not merely describe the world, that is, simply conformed to the data, but changed the world, and along with it the numbers themselves. The geographer who most focussed on this issue, though, devoting 'some of the best years'[8] of his life to it, and developing a sophisticated critique during the 1970s, was Gunnar Olsson (1980). That critique was most clearly delineated in his work on the gravity model, and its use in Sweden as a tool of social engineering (Olsson, 1974). As Olsson (1974: 355) puts it:

> The spatial structure of the present Swedish welfare state has deliberately been built to reflect the structure of existing interaction patterns as these exerted themselves .... [W]hat happened was that a group of academicians – mainly geographers – went into the field of census taking, observed how people interacted over space, translated these observation into the positivistic language of a variant of the gravity model, determined from these fits where the boundaries between service areas actually fell, and then, finally, convinced the political decision makers that these boundaries were efficient boundaries to which the administrative and political areas ought to be adjusted .... [T]he initial purpose of creating a just society became altered to that of finding a set of efficient solutions to a problem of geometric positioning.

The problem, then, was that because past data on spatial interaction was used to determine the locations of Swedish public social service and health facilities, the old sociospatial order was reproduced. Inequalities of the past were revisited on the present. This was not deliberate. Administrators had the best intentions. But the very techniques that they used unintentionally led them to uphold prior inequalities. Nothing changed. That same conservatism is built into big data. That's why Netflix is reshooting *House of Cards*. It is the same old, same old. But it might be better for us all, even Netflix subscribers, if we got something new.

I am not suggesting that big data be scrapped. In any case, it has already become part of our early new millennium reality. There is no turning back the clock. It is here. But we can at least take a critical sensibility, interrogating some of the hyperbole with which it is accompanied. Part of the hype is that it is new. My argument here has been that it is not. What's been around comes around. It has a history. It comes from somewhere. And if we can remember that history we can learn from it. George Santyana said, 'Those who cannot remember the past are condemned to repeat it'. My commentary is a plea to remember the past, and big data's past in particular.

Notes

1 http://www.webpronews.com/house-of-cards-is-the-most-popular-tv-show-in-the-world-right-now-according-to-imdb-2013-02

2. BBC Radio 4, Start the Week, 'Mathematical modelling', first broadcast on February 11, 2013, http://downloads.bbc.co.uk/podcasts/radio4/stw/stw_20130211-1107a.mp3

3. Interview with Henry Ford, *The Chicago Tribune*, May 25, 1916.

4. Warntz to Stewart, October 1, 1953, 'Warntz, W' Box 36, John Q. Stewart Papers, Rare Books and Special Collections, Princeton University.

5. Don Janelle (2000) writes about Warntz's life. Warntz later in the early 1960s used Princeton's IBM 7090 computer to calculate the exact height of 3100 separate nails that he then pounded into a two by three-and-a-half feet block of wood to produce a three-dimensional U.S. potential population map for 1961 (photographs are found in a 2012 lecture that Janelle gave at Harvard: https://cga-download.hmdc.harvard.edu/publish_web/Video/2012_09_05_Don_Janelle_slides.pdf). Warntz later became Director of the Harvard Lab where geographic information system was developed (Chrisman, 2006).

6. Interview by the author with Waldo Tobler, Santa Barbara, California, USA, March, 1998.

7. McNamara's confession is found in Errol Morris's 2003 revealing documentary, *The Fog of War: Eleven Lessons from the Life of Robert S. McNamara.*

8. Interview by the author with Gunnar Olsson, Pittsburgh, Philadelphia, USA, April, 2000.

References

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

Barnes TJ (2004) Placing ideas: *Genius Loci*, heterotopia, and geography's quantitative revolution. *Progress in Human Geography* 29(5): 565–595.

Batty M (2013) Big data, big issues. *Geographical* 85(1): 75.

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

Carr D (2013) Giving viewers what they want. *The New York Times*. 24 February. Available at: http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html?pagewanted=all&_r=0

Chrisman N (2006) *Charting the Unknown: How Computer Mapping at Harvard became GIS*. Redlands, CA: ESRI Press.

Eksteins M (2008) Drowned in Eeu de Vie. 21st February. *London Review of Books* 30(4): 23–24.

Faulkner W (1994) *Requiem for a Nun* (first published 1951). New York: Vintage.

Hacking I (1990) *The Taming of Chance*. Cambridge, UK: Cambridge University Press.

Harvey D (1972) Revolutionary and counter-revolutionary theory in geography and the problem of ghetto formation. *Antipode* 4(2):1–13.

Harvey D (1973) *Social Justice and the City*. London, UK: Edward Arnold.

Hudson D (1955) University of Washington. *The Professional Geographer* 7(4): 28–29.

Janelle D (2000) William Warntz, 1922–1988. In: PH Armstrong and GJ Martin (eds) *Geographers: Biobibliographical Studies* Vol. 19. London and New York: Mansell, pp. 102–118.

Ley D and Samuels M (eds) (1978) *Humanistic Geography*. Chicago, IL: Maroufa Press.

Olsson G (1974) Servitude and inequality in spatial planning: ideology and methodology in conflict. *Antiopde* 6(1): 16–21.

Olsson G (1980) *Birds in Egg/ Eggs in Bird*. London, UK: Pion.

Powell RC (2007) Geographies of science: histories, practices, localities, futures. *Progress in Human Geography* 31(3): 309–329.

Sayer A (1984) *Method in Social Research: A Realist Approach*. London, UK: Hutchinson.

Silver N (2012) *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't*. New York, NY: Penguin.

Warntz W (1959) *The Geography of Price*. Philadelphia, PA: University of Pennsylvania Press.

Yakabuski K (2013) Big Data should inspire humility, not hype. *The Globe and Mail*, 4 March, A11.