Dialogues in Human Geography

Commentary

Dialogues in Human Geography
3(3) 274–279

DOI: 10.1177/2043820613513390
dhg.sagepub.com
SAGE

Big data, smart cities and city planning

Michael Batty
University College London, UK

Abstract

I define big data with respect to its size but pay particular attention to the fact that the data I am referring to is urban data, that is, data for cities that are invariably tagged to space and time. I argue that this sort of data are largely being streamed from sensors, and this represents a sea change in the kinds of data that we have about what happens where and when in cities. I describe how the growth of big data is shifting the emphasis from longer term strategic planning to short-term thinking about how cities function and can be managed, although with the possibility that over much longer periods of time, this kind of big data will become a source for information about every time horizon. By way of conclusion, I illustrate the need for new theory and analysis with respect to 6 months of smart travel card data of individual trips on Greater London's public transport systems.

Keywords

big data, managing disruptions, new theory, real-time streaming, shorter time horizons, smart cities

There are many definitions of 'big data' but one of the best I have heard, and I do not know who to attribute it to, is 'any data that cannot fit into an Excel spreadsheet'. This immediately gives one some idea of size, for such spreadsheets now have a dimension of about one million rows and a much lesser number of columns; this definition also suggests that big data must be defined in relation to the standard tools that enable it to be processed to some purpose (Reades, 2013). This implies that big data is not a new concept but exists in every era where the tools for data processing are always being stretched by increasing size. There are some wonderful stories about how big data sets have driven the development of new hardware, software and mathematical methods throughout the history of computing but the present focus does appear to herald a rather different kind of response to the way we might use such data for a better understanding of the world around us. It is always risky to argue that 'this time is different' but our current obsession with big data is driven not just by questions of size but by the actual ways in which more and more data are being collected.

Much if not most of what we now call big data is produced automatically, routinely, and by various forms of sensors. This has been the case since the first industrial technologies were developed using continual monitoring of routine tasks and activities in analogue form and then electrical technologies gave a dramatic boost to such sensing through

Corresponding author:
Michael Batty, Centre for Advanced Spatial Analysis, University College London, Gower Street, London, WC1E 6BT, UK.
Email: m.batty@ucl.ac.uk

various machines including telephones and computers. But it was digital miniaturisation that has really changed the game in that, for the first time, we are now seeing computers being embedded into every conceivable type of object, including ourselves, and it is this that is generating unprecedented quantities of data. In fact, what has really spurred the rise of big data is the collection of data pertaining to activities that humans are intimately involved with. With 7 billion people on the planet, who access about 1.2 billion personal computers, there is now (mid-2013) more than this number of smart phones, some 1.5 billion, growing at around 30% annually. The scale of data being generated by these devices is daunting. In fact, sensor technologies have become ubiquitous with almost plug and play like qualities, thus enabling anyone to monitor and collect data from objects with motion that can be sensed by these devices. In our domain of the city, for example, fast automated travel data now records demand and supply, the performance of the various devices, the costs involved, usage of fuel, energy and so on. The data we have for public transit in London where some 8 million trips a day are made on tubes, heavy rail and buses are taken from the smart card that 85% of all passengers use. This yields about 45 million journeys a week, 180 million a month, a billion or so every half year and so on. The data set will be endless until the technology is changed and even when this occurs, the data being generated will at least in principle still remain a continuing stream. This is the kind of data that are of concern here.

In the contemporary jargon of computing, big data is linked to the general notion that computation since its inception 60 years ago has followed a path of decentralisation in terms of its hardware and software, and now its data. As computers have become ever smaller, they have become spatially all pervasive, located anywhere and everywhere but in parallel. This has been accompanied by decentralised forms of communication where hardware is able to interact at the most basic level. Computers and communications have converged during this process of global spreading, and this has led to the decomposition of the many different tasks that define computation. Twenty or more years ago, the client–server architecture came to define how we interacted with computers and data and software moved 'offshore', in the sense that computers began to be used to access resources in remote locations. This became writ large with the spread of the World Wide Web and in the last decade, the notion of computation through the web – where resources can be accessed interactively – has come to define major ways of working with computers. Into this mix has come the idea that a computer is simply a device whose processing power enables a user to simply interact with data, software and other users where the majority of computation, now defined as both storage of data and its processing, is located somewhere out there in the 'cloud'.

Into this mix has come big data, twisting the logic one step further in that computers are for the first time being used en masse to sense change and collect data, while being located remotely and passing their data to other places, all of which are remote from the ultimate users whose task is to work with the data. This indeed is a very new prospect in that it is both a new step in the miniaturisation and fragmentation of computation. To provide some sense of how recent this move to big data has been, we can use big data to explore 'the rise of big data' through Google Trends, which enables us to mine the Google search engine to generate the popularity of different trends (Choi and Varian, 2009). The keyed terms – cloud computing, Web 2.0 and big data – reveal that over the last 9 years, it is big data that has had by far the most rapid growth in interest. These trends are shown in Figure 1(a), where it is quite clear how Web 2.0, cloud computing and big data have generated successive waves of interest, with big data being the current buzz word. It is tempting to speculate that this wave, like the others shown in the figure, will pass and something else relating to the digitisation of the world will take its place. Alongside, in Figure 1(b), I show the popular interest in big data in the form of the banners that you encounter from the EMC^2 company, which are currently on display in the rail station at terminal 5 in London's Heathrow airport. This should be sufficient to convince that big data has entered the popular imagination.

This is a particularly interesting demonstration of the rise of big data largely because we are using big

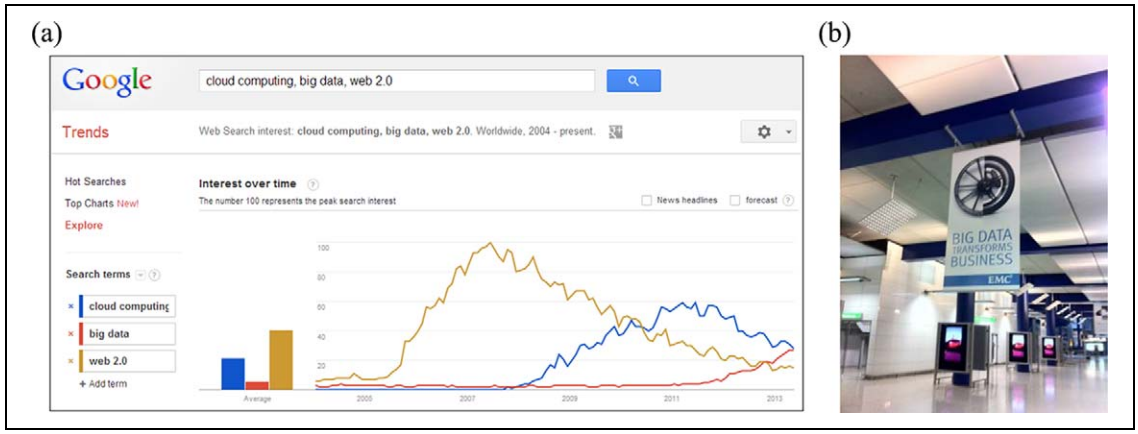

Figure 1. (a) The rise of Web 2.0, clouding computing and big data from 2004 to 2013; (b) big data transforms business: the EMC^2 company advertisements in the Heathrow airport terminal 5 rail station.

data to actually track the same. Google Trends enables one to use the entire resources of the Google search engine to assess the post-2004 popularity (or otherwise) of such key words, clichés, catch phrases and so on (Preis et al., 2012). The search engine now processes enough data to be able to yield quite focussed trends in word usage on a scale (in terms of queries) and at a frequency (of access) which is simply unprecedented. The Google search engine processes something like 30 billion searches a week; hence, the search for the term big data illustrated above has been carried out on a truly 'big' data set where the search volumes are available by the week.

The smart city and urban theory

The perspective here is manifestly spatial and urban in that the data sets pertain to large cities. My interest in them is primarily for developing a new understanding of how cities function, albeit on much shorter time horizons than has traditionally been the focus in urban geography. This, however, immediately generates a concern for how these data can be used to derive rather new theories of how cities function in that the focus is on much shorter term issues than hitherto, and much more on movement and mobility than on the location of land use and the long-term functioning of the city system. This is city planning in a new guise – that is, thinking of cities as being plannable in some sense over minutes, hours and days, rather than years, decades or generations.

The rise of big data is not solely associated with sensing technologies, for there are many big data sets that are generated by human responses; nevertheless, an increasing share of big data is produced automatically and routinely from sensors. In time, it is likely that big data will become associated entirely with routinely sensed data, especially as traditional data sets tend to be increasingly complemented by routine sensing, as well as crowdsourcing (where individuals enter their own data). There is, however, a coincidence between what are now being called smart cities and big data, with smartness in cities pertaining primarily to the ways in which sensors can generate new data streams in real time with precise geo-positioning, and how the data bases that are subsequently generated can be integrated so that value can be added. This last possibility is a somewhat pious hope in that unless there are common keys, which invariably there are not, data sets are usually impossible to link. Nevertheless, these kinds of data which so far deal with movement and transport, some energy and utility flows, and in time may well extend into spatial financial market data – housing markets and point of sales data pertaining to other kinds of consumption – are the stuff of smart cities (Batty et al., 2012). Of course, it is often pointed out that cities only become smart when people are smart^2, and this is sine qua non of our argument here.

Smart cities can also be synonymous with intelligent cities, information cities, virtual cities, amongst many other nomenclatures, but here our usage pertains rather narrowly to data and theory that brings much more immediacy to our urban understanding. In the history of urban studies and planning, most theory and applications have focussed on the long term – on what happens in cities measured over months and years, certainly not over minutes and hours. In fact, transport has always been concerned with peak daily flows, but these are assumed to pertain to a much longer period and thus most of our theory and planning has been focussed on what happens to cities over planning horizons that relate to years – the short term being 5 years and the long term 20 or 50 years. Of course, many groups are concerned with how cities function more routinely over the day but this has not been considered part of urban planning except in terms of urban operations research for emergencies and related services. Much of the routine management of cities has been accomplished in ad hoc ways, not necessarily without any data or science but certainly without the kind of comprehensive theory and modelling that characterises the longer term.

Smart cities belie a shift in this emphasis to a deeper understanding of how urban systems function in the short term. The notion of disruption is all important as big data and the various tools that are being built as part of complexity theory, particularly those pertaining to networks, are being quickly fashioned to deal with how one can respond and plan for very short-term crises. These can range from those that beset the transport system to issues pertaining to the housing market and the provision of social and other services, all of which have been handled in the past in pretty ad hoc ways. But to really get a grip on these issues new theory is needed. As West (2013) so vividly argues, 'big data requires big theory': data without theory is meaningless despite the argument that has been made by commentators such as Anderson (2008) who argues that we can now forget theory in the age of big data. All one needs to look for in big data, so the argument goes, are more and more correlations. In fact as data get bigger, the number of spurious correlations increases exponentially, as Taleb (2013) has indicated. In terms of cities and their functioning, the search for such correlations would be something of a diversion, for what we need to look for in big data can only ever be discovered through the lens of theory.

The other issue is that big data – data streamed in real time at the resolution of seconds – becomes data that pertains to every kind of time horizon if it collected for long enough. We will be able to see the sort of changes in big data that were once only sampled or collected every 10 years in complete population censuses. So, we do not stand at threshold of a time when our attention span is necessarily becoming shorter but at a threshold where our attention spans are being enriched over many time horizons. Thus, new theory should address these diverse horizons. The challenge of course is that big data will push the world ever further into short termism and already there is evidence of this from recent reactions to global crises. This must be resisted.

Planning the smart city

Big data is certainly enriching our experiences of how cities function, and it is offering many new opportunities for social interaction and more informed decision-making with respect to our knowledge of how best to interact in cities. Whether these spontaneous developments will be to our collective advantage or otherwise is yet to be seen for there is undoubtedly a dark side to these developments, quite obviously in questions of privacy and confidentiality. It is worth illustrating some of the potentials of these developments as well as some of the problems and to close the argument, an example from the biggest data set that our group (Centre for Advanced Spatial Analysis at University College London, UK) are working with is worth presenting. We have 1 billion or so records of all those who have tapped 'in' and 'out' of the public transport systems deploying the smart 'Oyster' card for paying for travel, which includes buses, tube trains and overground heavy rail in Greater London. The time period for the data is over 6 months during 2011–12. About 85% of travellers use the card and immediately we have a problem of comprehensiveness in that those who do not use it are likely to be specialist groups – tourists, those who are

occasional users, those who cannot afford the actual card and so on. The data set is remarkable in that we know where people enter the system and leave it, apart from about 10% of users who do not tap out due to open barriers. The data set is thus further reduced in its comprehensiveness. Because tap-ins and tap-outs cannot be associated with origins and destinations of trips, we cannot easily use this data in our standard traffic models without some very clever detective work on associating this travel trip data to locations of home and work and other land use activities. This is possible by good estimation but requires us to augment and synthesise the data with other independent data sets, and thus there is always error.

Such data are potentially extremely useful for figuring out disruptions on the system. We do need, however, to generate some clever cognitive analyses of how people make their way through the various transport systems, just as we need to assign travellers to different lines to ensure that we can measure the correct number of travellers on each line. As the tube, for example, is extremely complicated geometrically, we need to figure out how people navigate it. The state of the art in what we know about navigation in complex environments is still fairly primitive. Many assumptions have to be made and we have no data on what different users of the system have actually learned about their routes. New users of the system will behave differently from seasoned users and this introduces further error. We can see disruption in the data by determining the times at which travellers enter and exit the system, but to really predict disruption on individual lines and in stations, we need to match this demand data to the supply of vehicles and trains that comprise the system.

Fortunately, Transport for London (TfL) have very detailed data on buses and trains that give us precise geo-positioning, times and delays with respect to timetables. In principle, these can be matched against the travellers using the Oyster card. These demand and supply data sets, however, are entirely incompatible because there is no way of knowing which passenger in the system gets onto which train or bus, so cumulative delays for individual passengers cannot be assigned. Again, it is possible to make assumptions about passengers and their temporal positioning in the system, but to my knowledge no one has attempted this kind of synthesis. Currently what we are able to do with the Oyster data is assign it to lines and then to close stations and lines and figure out where passengers might divert to. There are many issues in this kind of analysis but in principle, predictions can be made to give some sense of disruption. So far, most measures of disruptions are done to the network system without loading the passenger volumes, and so far, simple network analyses are all that is available for figuring out delays. In short, matching the demand data to the network is possible and is being attempted, but matching it with supply data is almost impossible. Diversion behaviour of travellers is also tricky as people can walk between stations and bus stops, and there is considerable analysis needed to indicate how people might change mode of travel from one network to another – either for making a straightforward trip or a disrupted trip. These are massive challenges that will require new theories about how people behave in such situations at a very fine spatial scale. Big data provides the context for the study of this kind of short-term behaviour, but we are at the beginning of such explorations, and the many pitfalls that we have indicated here are likely to preoccupy us for some time to come (Reades, 2013).

Notes

1. Jon Reades of Kings College London is responsible for the data mining and analysis of the Oyster card data sets referenced in this paper (Reades, 2013).
2. The adjective smart is a peculiarly North American usage where it is used much more widely in conversation than in English. It pertains to the earlier smart growth movement in cities in the United States, where it has been used for 20 years or so as a synonym for planned or contained sprawl.

References

Anderson C (2008) The end of theory: will the data deluge make the scientific method obsolete? *Wired Magazine*, July 16, http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 30 July 2013).

Batty M, Axhausen K, Fosca G, Pozdnoukhov A, Bazzani A, Wachowicz M, et al. (2012) Smart cities of the future. *European Physical Journal Special Topics* 214: 481–518.

Choi H and Varian H (2009) Predicting the present with *Google Trends*, http://www.google.com/googleblogs/pdfs/google_predicting_the_present.pdf, (accessed 31 May 2013).

Preis T, Moat HS, Stanley HE and Bishop SR (2012) Quantifying the advantage of looking forward. *Scientific Reports* 2: 350.

Reades J (2013) Big data's little secrets: part 1. *Placetique: People, Data, Place*, http://www.reades.com/2013/05/31/big-data-little-secret/ (accessed 31 May 2013).

Taleb NN (2013) Beware the big errors of 'big data'. *Wired Blog*, February 2, http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/ (accessed 31 May 2013).

West GF (2013) Big data needs a big theory to go with it. *Scientific American*, May 15, http://www.scientificamerican.com/article.cfm?id=big-data-needs-big-theory (accessed 31 May 2013).