Geography

Jennifer Ferreira

Geography

Geography as a discipline is concerned with developing a greater understanding of processes that take place across the planet. While many geographers agree that big data presents opportunities to glean insights into our social and spatial world, and the processes that take place within it, many are also cautious about how it used and the impact it may have on how these worlds are analysed and understood. Given that often big data are either explicitly or implicitly spatially or temporally referenced, this makes it particularly interesting for geographers. Geography, then, becomes part of the big data phenomenon.

As a term that has only relatively recently become commonly used, definitions of big data still vary. Rob Kitchin suggests there are in fact seven characteristics of big data, extending beyond the three V's proffered by Doug Laney (volume, velocity and variety) which are widely cited.

1. Volume: often terabytes and sometimes petabytes of information are being produced.
2. Velocity: often a continuous flow created in near real time.
3. Variety: composed of both structured and unstructured forms.
4. Exhaustivity: striving to capture entire populations.
5. Fine grained: aiming to provide detail
6. Relational: with common fields so data sets can be conjoined.
7. Flexible: so new fields can be added as required, the data can be extended and where necessary exhibit scalability.

This data is produced largely through three forms: directed, generated largely by digital forms of surveillance; automated, generated by inherent automatic functions of digital devices; and volunteered, provided by users for example via interactions on social media or crowdsourcing activities.

The prevalence of spatial data has grown massively in recent years, with the advent of real-time remote sensing and radar imagery, crowdsourcing map platforms such as Open Street Map, and digital trails create by ubiquitous mobile devices. This has meant there is a wealth of data to be analysed about human behaviour, in ways not previously possible.

Large data sets are not a new concept for geography. However, even some of the most widely used large data sets used geography, such as the census, do not constitute big data. While they are large in volume and seek to be exhaustive, and high in resolution, they are very slow to be generated and have little flexibility. The type of big data now being produced is well exemplified by companies such as Facebook which in 2012 alone processed over 2.5 billion pieces of content, 2.7 billion 'likes' and 300 million photo uploads in just one day; or Walmart which generated over 2.5 petabytes of data information every hour in 20102. One of the key issues for using big data is that collecting, storing and analysing these kinds of data is very different from that of traditionally large data sets such as the census. These new forms of data creation are creating new questions about how the world operates, but also about how we analyse and use such data forms.

Governments are increasingly turning to big data sources to consider a variety of issues, for example public transport. A frequent system referred to about the production of big data related to public systems is the use of the Oyster Card in London. Michael Batty discusses the example of public in transport in London (tube, heavy rail and buses) to consider some of the issues with big data sets. With around 45 million journeys every week or around a billion every year the data is endless. He acknowledges that big data is enriching our knowledge of how cities function, particularly with respect to how people move around them. However, it is can be questioned how much this data can actually tell us. Around 85% of all travellers using public transport in London on these forms of transport use the Oyster Card, and so clearly there is an issue about representativeness of the data. Those that don't use the card, tourists, occasional users, and other specialist groups will not be represented. Furthermore because we can't actually trace where an individual begins and ends their journey it only presents a partial view of the travel geographies of those in London. Nevertheless this data set is hugely important for the operation of transport systems in the city.

Disaster response using big data has also received significant media attention in recent years: crisis mapping community after the 2010 Haiti earthquake, or collecting tweets in response to disaster events such as Hurricane Sandy. This has led to many governments and NGO's promoting the use of social media as potentially useful data sources for responding to disasters. While geo-referenced social media provides one lens on the impact of disaster events it should not be relied on as a representative form of data covering all populations involved. Big data in these scenarios presents a particular view of the world based on the data creators, and essentially can mask the complexity and multiplicity of scenarios that actually exist. Taylor Shelton, Ate Poorthuis, Mark Graham and Matthew Zook explore the use of twitter around the time of Hurricane Sandy and they acknowledge that their research did not present any new insights into the geography of twitter, but that it did show how subsets of big data could be used for particular forms of spatial analysis.

Trevor Barnes argues that criticisms of the quantitative revolution in geography are also applicable to the use of big data. First that a focus on the computational techniques and data collected can become disconnected from what is important, i.e. the social phenomena being researched. Second that it may create an environment where quantitative information is deemed superior, and that where phenomena can't be counted they won't be included. Third, that numbers do not speak for themselves - numbers created in data sets (of any size) emerge as a product of particular social constructions even where they are automatically collected by technological devices.

The growth of big data as part of the data revolution presents a number of challenges for geographers. While there has been much hype and speculation over the adoption of big data into societies, changing the ways that businesses operate, government manage places and the way that organizations manage their operations. For some, the benefits are overstated. While it may be assumed that because much technology contains GPS that the use of big data sets is a natural extension of the work geographic information scientists, it should be noted that the emergence of such data sets created by mobile technology has created a large new amount of data, but also data that geographic information scientists have not typically focused their efforts. Therefore work is needed to develop sound methodological frameworks to work with such data sets.

The sheer size of the data sets that are being created, sometimes with millions or billions of observations being created in a variety of forms on a daily basis is a clear challenge. Traditional statistical analysis methods used in geography are designed to work using smaller data sets with much more known about the properties of the data being analysed. Therefore new methodological techniques for data handling and analysis are required to be able to extract useful information for geographical research.

Data without a doubt are a key resource for modern the world, however it is important to remember that data does not exist independently of the systems (and people in them) from which they are produced. Big data sets have their own geographies, they are themselves social constructions formed from variegated socio-economic contexts and therefore will present a vision of the world that is uneven in its representation of populations and their behaviour. Big data, despite attempts to make it exhaustive will always be spatially uneven and biased. Data will always be produced by systems that have been created with influences from different contexts and from groups of people with different interests.

Sandra González-Bailón highlights that while technology has allowed geospatial data to be generated much more quickly than in the past, and if mobilized in an efficient manner people can use these technologies as network of sensors. However, the use of digital tools can produce distorted maps or results if the inputs to the system are systematically biased i.e. those who do not have access to the tools will not be represented. Therefore there are questions about how to extract valid information from the ever growing data deluge. Then there are issues around privacy and confidentiality of the data produced and how it will be used potentially in both the public and private sector.

Michael Goodchild highlights that while a lot of big data is geo-referenced and can contribute to a growing understanding of particular locations, there are issues about quality of data that is produced. Big data set are often comprised of disparate data sources which don’t always have quality controls, or do not have metadata about the provenance of the data. This raises questions about the extent such data can be trusted, or used to make valid conclusions. There is a need for geographers to explore how data can become more rigorous. Michael explains how Twitter streams continue to be seen as important sources of information about social movements, or events, but often little is known about the demographics of those tweeting and so it is impossible to understand the extent to which these tweets represent the wider sentiments of society. Furthermore, only a small percentage of tweets are georeferenced and so the data is skewed toward the data provided by people who opt-in to provide that level of data. Much like many other geographers writing on the topic of big data, the potential for such source s of data to be useful, but questions need to be raised about how it is used, and how the quality is improved.

Mark Graham, has begun to ask questions about the geographies of big data and considered which areas of the world are displayed through big data sets, and what kinds of uneven geographies are produced by them. The geographies of how data is produced is revealing in itself. This is exemplified by examining the content of Wikipedia: every article on Wikipedia was downloaded and place on a map of the world. While this showed a global distribution, particularly for articles in English language, the worlds displayed by those in Persian for example, were much more limited. The key point here was that the representations made available to the world through the use of big data can lead to the omission of other worlds that still exist but may not be visible. These absences or 'data shadows' are also a concern for geographers. It raises questions about what this says about the places they represent. In exploring this phenomenon geographers are seeking to explore the geographies of data authorship in the data deluge, considering why there are differences in the production of information, asking questions about why some people produce large amounts of data while others are excluded.

It is without question that digital technologies have transformed the ways in which we can explore the way the world works; the flood of data now being produced can be used to create more maps of places, more models of behaviour and more views on the world. With companies, governments, and research funding agencies calling for more effort to be put into generating and exploring big data, some geographers have highlighted that in order to deliver significantly valuable insights into societal behaviour then more effort is needed to ensure that the big data collection and analysis is scientifically robust. Big data and particularly data that is geo-referenced has provided a new wealth of opportunities to understand more about people and places, asking new questions, measuring new processes and phenomena in ways not previously possible.

Jennifer Ferreira
Coventry University

See Also: Demographic Data; Disaster Planning, Environment; Smart Cities; Spatial Analytics; Spatial Data; Spatial-Temporal Analytics.

Further Readings

Barnes, Trevor. Big data, little history. *Dialogues in Human Geography*, 3(3): 297-302 (2013).

Batty, Michael. Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3):274-278 (2013).

Goodchild, Michael. The quality of big (geo)data. *Dialogues in Human Geography* 3(3) 280-284 (2013).

Gonzalez –Bailon, Sandra. Big data and the fabric of human geography. *Dialogues in Human Geography*, 3(3): 292-296 (2013).

Laney, Doug. (2001) 3D data management: controlling data volume, velocity, and variety. Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocityand-Variety.pdf [Accessed: 18/11/14].

Li, Linna., Goodchild, Michael., and Xu, Bo. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, 40(2): 61–77 (2013).

Kitchin, Rob. Big data and human geography: opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3): 262-267 (2013).

Kitchin, Rob. The Data Revolution: Big Data Open Data, Data Infrastructures and their Consequences. London: Sage. (2014).

Shelton, Taylor, Poorthuis, Ate, Graham, Mark and Zook, Matthew. (2014) Mapping the data shadows of Hurricane Sandy: uncovering the sociospatial dimensions of ‘big data’ *Geoforum*, 52 (1)167-179.