Big data

Rob Kitchin
National University of Ireland

The etymology of "big data" can be traced to the mid-1990s, when the term was first used to refer to the handling and analysis of massive datasets. It is only since 2008, however, that the term has gained traction, becoming a business and industry buzzword. Like many rapidly emerging concepts, big data have been variously defined, but most commentators agree that they differ from what might be termed "small data" with respect to their traits of volume, velocity, and variety. Traditionally, data have been produced in tightly controlled ways using sampling techniques that limit their scope, temporality, and size. Even very large datasets, such as national censuses, have been restricted to generally 30–40 questions, and are carried out once every 10 years in most countries. Advances in computing hardware and software and in networking have, however, enabled a much wider scope for producing, processing, analyzing, and storing massive amounts of diverse data on a continuous basis. Moreover, big data generation strives to be exhaustive, capturing entire populations or systems (n = all); fine-grained in resolution and uniquely indexical in identification; relational in nature, containing common fields that enable the conjoining of different datasets; and flexible, holding the traits of extensionality (new fields can be easily added) and scalability (can expand in size rapidly). Big data thus consist of huge volumes of diverse, fine-grained, interlocking data produced on a dynamic basis. For example, in 2012, Walmart was generating more than 2.5 petabytes (2^{50} bytes) of data relating to more than 1 million customer transactions *every hour*, and Facebook was processing 2.5 billion pieces of content (links, comments, etc.), 2.7 billion "Like" actions, and 300 million photo uploads *per day*. Such big data, its proponents argue, enable new forms of knowledge that produce disruptive innovations with respect to how business is conducted and governance enacted. Given that much big data are georeferenced, they enable new kinds of geographic analysis and insights.

Sources of big data

Big data are produced in three broad ways: through directed, automated, and volunteered systems. Directed systems are controlled by a human operator and include closed-circuit television, spatial video, and LiDAR (light detection and ranging) scans. Automated systems automatically capture data as an inherent function of the technology and include the recording of retail purchases at the point of sale; transactions and interactions across digital networks (e.g., sending emails, Internet banking); the use of digital devices such as mobile phones that record and communicate the history of their own utilization; clickstream data that record navigation through a website or app; measurements from sensors embedded into objects or environments; the scanning of machine-readable objects such as transponders and barcodes; and machine-to-machine interactions across the Internet. Volunteered systems rely on users to gift data through uploads and interactions and include engaging in social

The International Encyclopedia of Geography.
Edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston.
© 2017 John Wiley & Sons, Ltd. Published 2017 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781118786352.wbieg0145

media (e.g., posting comments, observations, photos to social networking sites such as Facebook) and the crowdsourcing of data wherein users generate data and then contribute them to a common platform (e.g., uploading traces supplied by global positioning systems (GPS) to OpenStreetMap).

Analyzing big data

Given their volume, variety, and velocity, big data present significant analytical challenges to traditional methods, which have been designed to extract insights from scarce and static data. The solution has been the development of a new suite of data analytics that are rooted in research around artificial intelligence and expert systems, and new forms of data visualization and visual analytics, both of which rely on high-powered computing. Data analytics seek to produce machine learning that iteratively evolves an understanding of datasets using computer algorithms, automatically recognizing complex patterns and constructing models that explain and predict such patterns and optimize outcomes. Moreover, since different approaches have their strengths and weaknesses, depending on the type of problem and data, an ensemble approach can be employed that builds multiple solutions using a variety of techniques to model and predict the same phenomena. As such, it becomes possible to apply hundreds of different algorithms to a dataset to ensure that the most illuminating insights are produced. Given the enormous volumes and velocity of big data, visualization and mapping have proven a popular way for both making sense of data and communicating that sense. Visualization methods seek to reveal the structure, pattern, and trends of variables and their interconnections. Tens of thousands of data points can be plotted to reveal a structure that is otherwise hidden (e.g., mapping trends across millions of tweets to see how they vary across people and places) or the real-time dynamics of a phenomenon can be monitored using graphic and spatial interfaces (e.g., the flow of traffic across a city).

Pros and cons of big data

There is good reason for the hype surrounding big data. Big data offer the possibility of shifting from data-scarce to data-rich studies of all aspects of the world from narrow to exhaustive samples; static snapshots to dynamic vistas; coarse aggregations to high resolutions; relatively simple models to complex, sophisticated simulations and predictions. Furthermore, big data consist of both qualitative and quantitative data, most of which are spatially and temporally referenced. Big data provide greater breadth, depth, scale, and timeliness, and are inherently longitudinal in nature. They enable researchers to gain greater insights into various systems. For businesses and government, such data hold the promise of increased productivity, competitiveness, efficiency, effectiveness, utility, sustainability, and securitization, and the potential to better manage organizations, leverage value and produce capital, govern people, and create better places.

Big data are not without negative issues, however. For example, most big data are generated by private corporations such as mobile phone operators, app developers, social media providers, financial institutions, retail chains, and surveillance and security firms, none of which are under any obligation to freely share the data they generate. As such, access to such data is at present limited. There are also concerns as to how clean (error- and gap-free), objective (bias-free), and consistent (few discrepancies) the data are, and as to their veracity and the extent to which they accurately (precision) and faithfully (fidelity, reliability) represent what they are meant to. Further, big data raise a number of ethical questions concerning the extent to which they facilitate dataveillance (surveillance through data records), infringe on privacy and other human rights, enable social sorting (provide differential access to services), pose security concerns with regards to identity theft, and enable control creep wherein data generated for one purpose are used for another.

Geography and big data

Geographers have long engaged with massive datasets and nascent big data such as remote-sensing imagery and meteorological records, seeking to map and model environmental and climate change. More recently they have pioneered the analysis of volunteered geographic information (see Sui, Elwood, and Goodchild 2013), such as the mapping of georeferenced and locative social media data (e.g., data from Twitter and Foursquare) (Shelton *et al.* 2014), and begun to model large-scale urban data, such as the constant flow of passengers through a transport system (Batty 2014). They have also been at the forefront of debates concerning the sociospatial implications of big data technologies to urban systems and everyday life. Nonetheless, big data do pose a major challenge to the discipline, namely that, given their volume and velocity, analyzing such data requires a fundamentally different skill set to analyzing traditional forms of data, but as yet few geographers possess such skills.

SEE ALSO: Data quality standards; Metadata; Qualitative data; Quantitative methodologies

References

Batty, Michael. 2014. *The New Science of Cities*. Cambridge, MA: MIT Press.

Shelton, Taylor, Ate Poorthuis, Mark Graham, and Mathew Zook. 2014. “Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of ‘Big Data.’” *Geoforum*, 52: 167–179. DOI:10.1016/j.geoforum.2014.01.006.

Sui, Daniel, Sarah Elwood, and Michael Goodchild, eds. 2013. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Berlin: Springer.

Further reading

Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.