[CV-05-019] Big Data Visualization

Abstract

As new information and communication technologies have altered so many aspects of our daily lives over the past decades, they have simultaneously stimulated a shift in the types of data that we collect, produce, and analyze. Together, this changing data landscape is often referred to as "big data." Big data is distinguished from "small data" not only by its high volume but also by the velocity, variety, exhaustivity, resolution, relationality, and flexibility of the datasets. This entry discusses the visualization of big spatial datasets. As many such datasets contain geographic attributes or are situated and produced within geographic space, cartography takes on a pivotal role in big data visualization. Visualization of big data is frequently and effectively used to communicate and present information, but it is in making sense of big data – generating new insights and knowledge – that visualization is becoming an indispensable tool, making cartography vital to understanding geographic big data. Although visualization of big data presents several challenges, human experts can use visualization in general, and cartography in particular, aided by interfaces and software designed for this purpose, to effectively explore and analyze big data.

Keywords: big data, geovisualization, interactive design techniques, visualization

Author & citation

Poorthuis, A. (2018). Big Data Visualization. The Geographic Information Science & Technology Body of Knowledge (3rd Quarter 2018 Edition), John P. Wilson (Ed.).
DOI: 10.22224/gistbok/2018.3.5.

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Computational issues in cartography and visualization. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions

Big data: Datasets that are characterized not only by their high volume but also by their velocity, variety, exhaustivity, resolution, relationality, and flexibility.

Volume: The amount of data necessary to be considered ‘big’ data. Typically, the volume

of big data is measured in terabytes and petabytes, or consisting of millions to billions of observations.

Velocity: The frequency with which a dataset is updated. Typically, big data is produced or updated in real-time or at a fine temporal granularity.

Variety: The diversity of data points available within and between data sets. Big data typically consists of a wide range of structured and unstructured datasets from different sources and provenances.

Exhaustivity: A term that describes the scope of big data. For big data, the data set is typically as wide as possible, focused on entire populations rather than samples.

Resolution: The granularity and detail in big data. Big data is typically as detailed as possible, including being indexical in identifying the objects under study.

Relationality: The extent to which different datasets can be joined together based on common attributes. One of the defining characteristics of big data is its ability to be connected to other datasets.

Flexibility: The ability of a dataset to be easily extended (with additional attributes) and expanded (by adding additional observations).

Data Reduction: A strategy used to reduce the amount of data or summarize relevant parts of a dataset.

Filtering: The subsetting of a dataset based on attributes of the data.

Subsampling: The subsetting of a dataset based on stochastic sampling.

Aggregation: The combination of multiple data points into a higher-level aggregation.

Projection: A data reduction strategy that ‘maps’ data points to either a smaller number of dimensions or narrower data range.

2. Introduction to Big Data

2.1 What is Big Data?

New information and communication technologies have altered many aspects of our daily lives over the past decades, and simultaneously stimulated a palpable shift in the types of data that companies, governments, scientists, and individuals are able to collect, produce, and analyze. These new emerging datasets are often referred to as **big data**. The term ‘big data’ was first coined in the 1990s (Diebold, 2012). While the exact definition of big data remains somewhat fluid, there have been several efforts to define its core characteristics. One of the most commonly used definitions is based on the “three V’s” (Laney, 2001):

1. **Volume.** Big data is massive and is often measured in terabytes and petabytes or consisting of million or billions of observations.
2. **Velocity**. Big data is produced or updated in real-time or at a fine temporal

granularity.
3. **Variety**. Big data consists of a wide range of structured and unstructured datasets from different sources and provenances.

Although the 3V definition is succinct, new and alternative definitions of the concept have also been developed that help to further distinguish big data from “small data.” A useful synthesis of these definitions adds four additional dimensions to the 3V definition (see for an extensive review and (Kitchin, 2013; 2014; Kitchin & McArdle, 2016)):

1. **Exhaustivity**. The scope of big data is as wide as possible, focused on entire populations rather than samples.
2. **Resolution**. Big data is as detailed as possible, including being indexical in identifying the objects under study.
3. **Relationality**. Big data can be connected easily. Different datasets can be joined together based on common attributes.
4. **Flexibility**. Big data can be easily extended (with additional attributes) and expanded (by adding additional observations).

A data source or dataset does not need to exhibit all of the seven characteristics to be considered big data and there is no exact threshold that differentiates small and big data. Instead, it is an accepted notion that there exists a gray transition zone between the two. Further, multiple different forms or ‘species’ of big data may exist at the same time (Kitchin & McArdle, 2016). However, regardless of the semantics, it is clear that many of the datasets that are produced, analyzed, and visualized in the 21st century differ significantly from their 20th century counterparts, prompting a re-evaluation of the role cartography and visualization in this process.

1.2 The Relevance of Big Data for GIS&T

A large portion of big data is geographic in nature and, as such, big data has had a large impact on the geographic disciplines. Spatial big data ranges from mobile phone and traffic data to social media platforms (see **Social Media Analytics**) and credit card transactions, to air quality sensors and satellite imagery – each of which provides not only a data point, but a geographic location associated with that data point. All of these datasets can potentially help us better understand the world around us (see **Citizen Science with GIS&T**) and thus have seen an uptake in spatial research (Arribas-Bel, 2014; Goodchild, 2007; Graham & Shelton, 2013). The increasing prevalence of these types of datasets have spurred an entire new discipline of Data Science and some people working in GIS and related fields have started to relabel themselves as "spatial data scientists," as can be seen in the new Center for Spatial Data Science at the University of Chicago and the Geographic Data Science Lab at University of Liverpool.

More importantly, big data might change how we approach spatial analysis and visualization. While we now have access to unparalleled, large quantities of heterogeneous data about the world around us, it remains a formidable challenge to understand and interact with this data in meaningful ways. As a result, new approaches have been developed to help automate many aspects of data analysis, such as automated machine learning approaches, artificial intelligence and other "unsupervised" computational methods (see **Artificial Intelligence**). While these automated approaches can be useful additions to our toolbox, the human role in spatial data analysis and visualization remains

essential. As Shneiderman (2014) argues, while computer-led data analysis might be effective for well understood topics, the creation of new knowledge and breakthroughs requires human experts who can use and understand visualizations to gain new insights. Visualization is an indispensable tool to make sense of big data, which makes cartography vital to understanding geographic big data.

3. Visualization of Big Data: Visual Communication and Visual Thinking

In the domain of big data visualization, we can make the distinction between roughly two types of visualizations: those that aid in visual thinking and those meant for visual communication (DiBiase, 1990) (see **Cartography & Science** and **Geovisualization** for a more in-depth discussion). Visual communication is best done with a "Map-to-See," a straightforward cartographic representation to be understood in the blink of an eye (Kraak, 1988). On the other hand, visual thinking is often done through more complex cartographic products that may take a while to be fully understood: a "Map-to-Read."

In the context of big data, visual communication has been employed by companies, news desks, and scientists (see **Narrative & Storytelling,** forthcoming) to communicate findings, present narratives, or sometimes simply to impress on the reader the complexity or size of the underlying dataset. A clear example of the latter is the so-called "hairball" visualizations in which complex, large networks are visualized with an equally complex ball of lines (Krzywinski, Birol, Jones, & Marra, 2012). Within cartography, an analogous example is projecting a big dataset consisting of spatial points directly onto a map, resulting in a complex representation with millions of dots. Although many big data sets are indeed visualized to present and communicate – often in beautiful and compelling ways – ultimately the use of big data within this map use mode is not significantly different from that of small or more conventional data sets.

In the "visual thinking" mode, visualization is inextricably linked with big data for the purposes of exploration and analysis, and specifically to make sense of big data and generate new (scientific) knowledge (Fox & Hendler, 2011). Although it comes with its own set of challenges (see next section), visualization allows researchers to explore, analyze, and synthesize datasets that are too large, complex, and heterogeneous to understand by merely looking at the raw data. Visualization as such is an indispensable tool in this process and an important driving force in complex analyses of big data (see **Geovisual Analytics**).

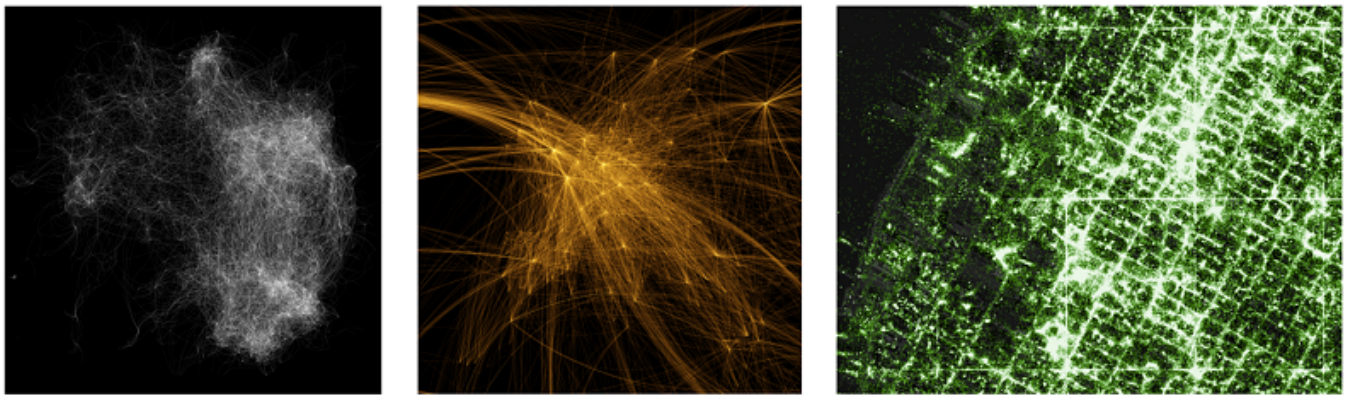


Figure 1: Examples of 'hairball'-type visualizations. From left to right, an example of a

namesake network visualization; a map of global passenger air routes (Josullivan.58 / CC-BY-3.0, https://commons.wikimedia.org/wiki/File:World_airline_routes.png); and a map displaying over 6 billion tweets showcasing Mapbox' mapping platform (Eric Fisher / CC-BY-2.0, https://www.flickr.com/photos/walkingsf/15869589271/in/photostream/).

4. Challenges for Big Data Visualization

4.1 Computational

The most obvious set of challenges with big data visualization are computational in nature. In its simplest form, it can be a challenge for conventional CPU-based mapping software to draw increasingly large amounts of data points (see **Graphics Processing Units**). Large datasets can also complicate even basic functions, such as data storage. For example, the file size of a standard shapefile in a Geographic Information System is limited to 2GB (or roughly 70 million point features) and 255 attributes, and each field is limited to 254 characters. Many big datasets exceed these limits, which warrants new file formats. In addition, the unstructured nature of many big data sets does not necessarily fit in the structured rigidity of conventional relational databases. New database ontologies (such as document-oriented and other NoSQL formats) have been developed to address these issues.

4.2 Visual

Another set of challenges with the visualization of big data lays within the domain of visualization itself. It should be noted here that these issues are not inherently unique to big data. Rather, big data significantly amplifies many pre-existing challenges in cartography and forces us to acknowledge and address them explicitly. The most obvious of these challenges is related to the size of the data. Simply visualizing or plotting such a large number of data points might create confusing visualizations that yield no insights (cf. the hairball visualization discussed above) or visualizations that hide or obscure data, often referred to as overplotting (see (Dang, Wilkinson, & Anand, 2010) for a discussion).

Many spatial big datasets contain precise geographic coordinates for each observation, which poses another, paradoxical challenge: the ease with which these coordinates can be plotted as points on a map may lure us into a potentially narrow or constraining visualization of big data (Crampton et al., 2013). On the flip side, some big data contains less precise, but still spatial, references to vernacular place names, neighborhoods, and spatial regions that might not be easily mapped to the discrete geometry of a polygon.

Of course, the "richness" or heterogeneity of such data presents additional questions. For example, how can the qualitative textual data of social media be effective visualized? This is particularly the case for datasets that have real-time or frequent temporal updates, meaning that the dataset may constantly be in a state of flux. Finally, the unstructured nature big data also means that observations might be inaccurate or less precise. In other words, potential uncertainty within the data might need to be accounted for in the visualization as well (see **Representing Uncertainty**).

4.3 Representation, Ethics, and Privacy

Apart from technical challenges, it is important to be cognizant of a series of ethical challenges for big data visualization. While ethics form an important part of the entire domain of GIS&T (see **Professional & Practical Ethics of GIS&T** and **Cartography & Power**), big data may enlarge or amend those ethical issues. A particularly notable example is the privacy of those whose data are mapped and visualized. Conventional datasets typically aggregate social data to census tracts or other administrative geographies, while many big datasets provide precise coordinate pairs, oftentimes at the level of the individual. Visualizing such data with the same precision may do harm to people. Conversely, coordinate pairs might also be spoofed or altered deliberately, potentially placing people in locations which they have never visited (Zhao & Sui, 2017). There are many additional issues surrounding the visualization of big data (e.g. representation; consent; bias) and it is an essential part of any project to be cognizant of these (see for an overview (boyd & Crawford, 2012; Zook et al., 2017; Zwitter, 2014)).

5. Approaches and Solutions to Big Data Visualization

5.1 Data Reduction

To address some of the challenges above, one important approach to big data visualization is to 'make big data small' (Poorthuis & Zook, 2017; Poorthuis, Zook, Shelton, Graham, & Stephens, 2015) and falls within the domain of **data reduction** or summarization. Visualizations of complex, large datasets do not have to be complex or large themselves. Sarikaya (2017) distinguishes four specific reduction strategies, which are not dissimilar from strategies employed in cartographic generalization (see **Scale & Generalization)**:

- Filtering. Subsetting a dataset based on attributes of the data. For example, only including records relevant to the process under study.
- Subsampling. Subsetting a dataset based on stochastic sampling. For example, by performing a random sample if unnecessary to visualize the entire dataset.
- Aggregation. Combining multiple data points in a higher-level aggregation. This can be a with a bottom-up approach by clustering proximate or similar points (see **Classification & Clustering**, forthcoming**)** or top-down by aggregating individual points to a higher spatial unit (e.g., administrative region) (see **Aggregation of Spatial Entities**).
- Projection. Unstructured or high-dimensional big data can be simplified by 'mapping' data points to either a smaller number of dimensions or narrower data range. In its simplest form, this can be done manually but larger datasets require the use of automated techniques that range from Principal Components Analysis (see **Analyzing Multidimensional Attributes,** forthcoming) to newer machine learning techniques (see **Machine Learning Programming for GIS**, forthcoming).

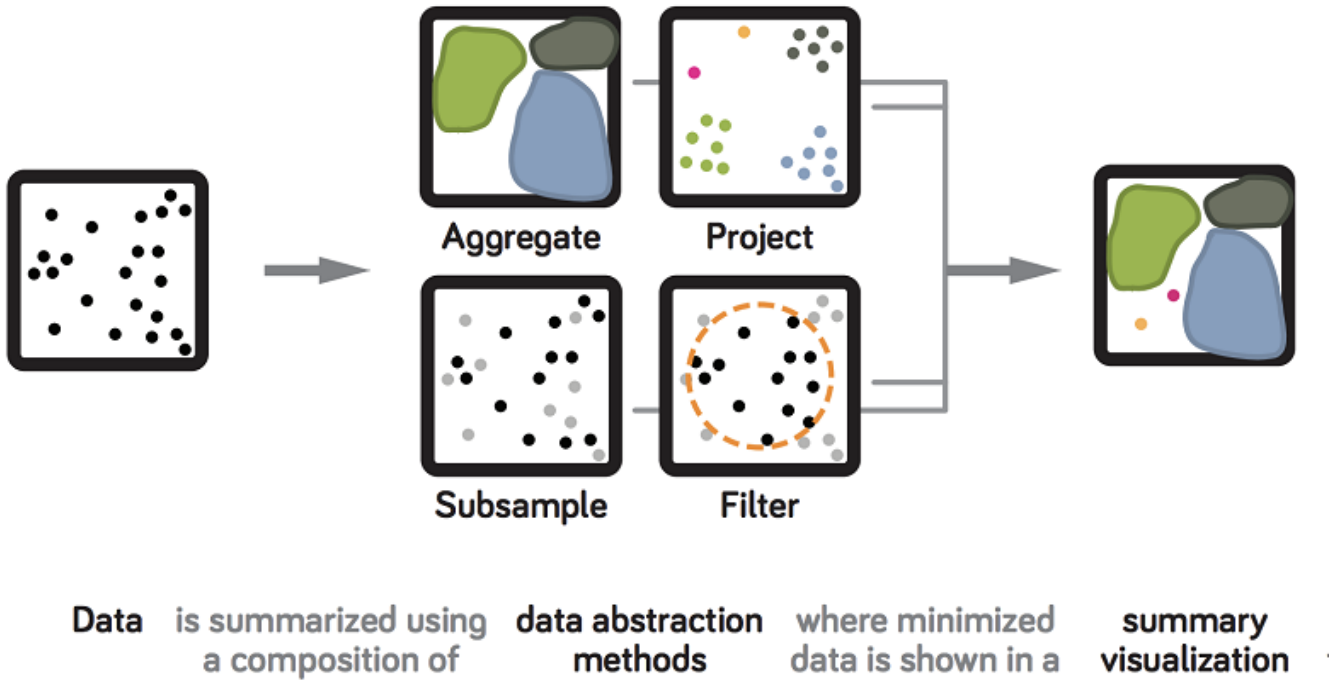


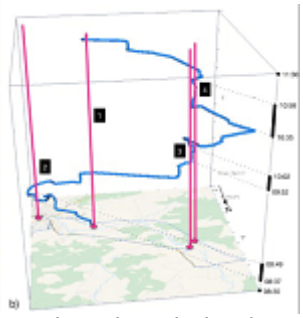
Figure 2: Big data can be made "small" through the use of data reduction strategies that yield summary visualizations. Figure reproduced with permission from Sarikaya (2017).

5.2 Visual Strategies

Data reduction strategies make big data small in order to use relatively conventional, straightforward cartographic techniques. However, depending on the nature of the data and the purpose of the visualization, this is not always an option. Explicitly incorporating big data in cartography, without the simplification from data reduction, is still at the cutting edge of the field, full of new challenges and opportunities (see (Robinson et al., 2017) for an overview). A canon of techniques has yet to crystalize but several examples of strategies can be identified (see Table 1).

Table 1. Examples of Visual Strategies for Big Data Visualization

Technique	Big Data Challenge Addressed	Description	Example
"Heat" or Hot Spot Maps	Large volume of point data	Conversion of point data to smooth surface, often through kernel density estimation. A popular choice due to its inclusion in many mapping software.	 Reproduced with permission (cf. Kumar, Morstatter, & Liu, 2014)
Edge Bundling	Large volume of line data	Big data often contains large amounts of spatial relations (e.g., traffic, movement). Edge bundling visually bundles relations that move in a similar direction.	 Sophie Engle / GPL-3.0 (cf. Holten & Van Wijk, 2009)

Technique	Big Data Challenge Addressed	Description	Example
Three-dimensional Maps	High dimensionality of big data	The third dimension can be used to represent an additional data attribute or temporal dimension (e.g., space-time cube)	 Kraak and Kveladze (2017), CC-BY-4.0
Multiple View Systems	Variety of big data	Multiple view systems provide multiple, often linked or coordinated windows into a dataset.	 Nost, Rosenfeld, Vincent, Moore, & Roth (2017), CC-BY-NC-ND-4.0
Animation	Velocity of big data	Animation can be utilized to include temporal dimension in maps	 Dheeraj Savala / MIT license

The process of big data visualization relies heavily on computationally intensive procedures, which requires us to work in close concert with our computers. To facilitate this process, big data visualization is often done in an exploratory, interactive fashion with interfaces and software that enable the user to quickly perform a series of exploratory analyses through the visualization of different aspects of a dataset (see **Exploratory Spatial Data Analysis** (forthcoming) and **UI/UX Design**). These interfaces can be custom-made for a specific project or tailored to use with big data. An example of such a project is imMens, a browser-based system that allows users to explore millions of multivariate data points in an interactive, real-time environment (Liu, Jiang, & Heer, 2013). To enable this, the system pre-computes visualizations in a way similar to webmap tilesets (see **Web Mapping**) and it performs calculations in parallel to make sure the computer can 'keep up' with the user (see **Parallel Programming and GIS Applications**, forthcoming). More conventional, off-the-shelf software has also been adapted to enable big data visualization. For example, ArcGIS is now using both GPU rendering and parallel processing and popular data science languages (e.g., Python and R) provide authoring environments for interactive visualization and the tight coupling of analysis and visualization (see **Jupyter Notebooks**, forthcoming).

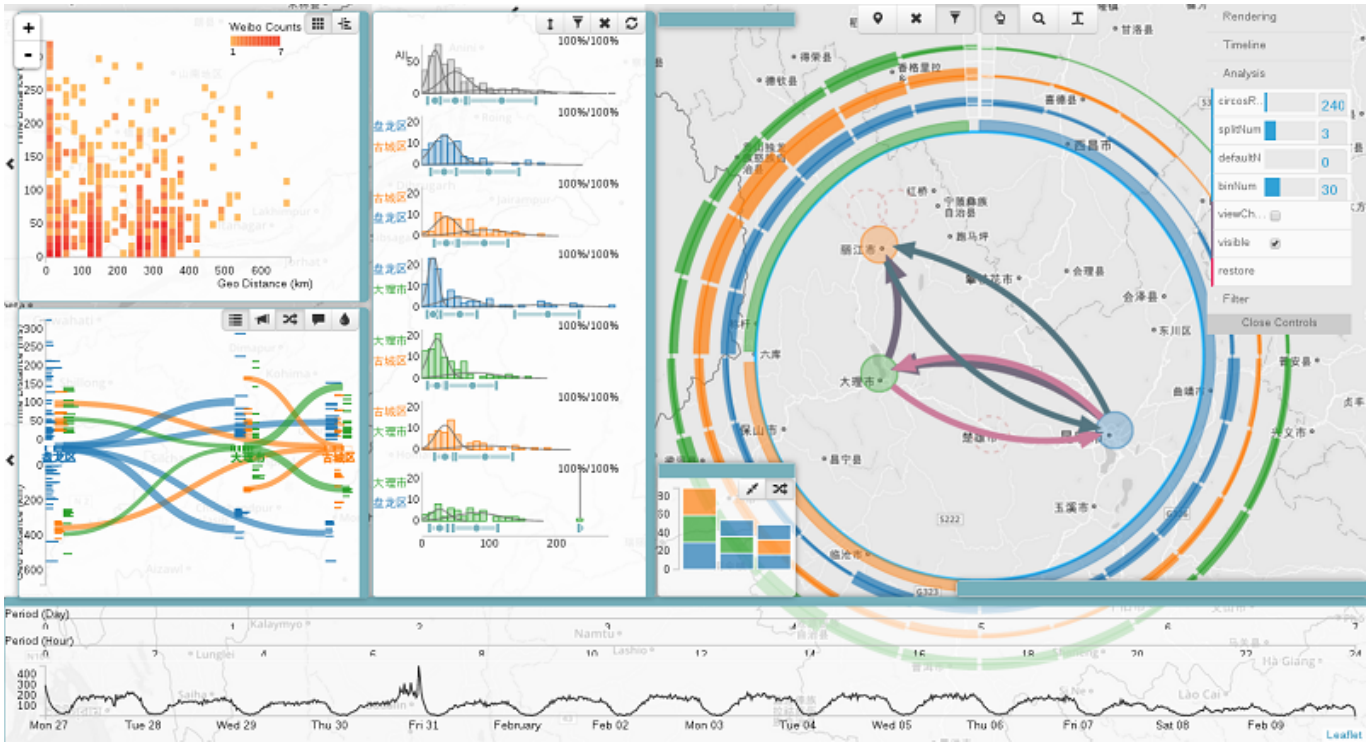


Figure 3: An example of an exploratory, interactive software interface visualizing big data (Chen et al., 2016). It allows the discovery of movement patterns in social media data through both data reduction (e.g., filtering) and visualization strategies (e.g., multiple linked views). Reproduced with permission (http://vis.pku.edu.cn/trajectoryvis/en/weibogeo.html).

It is clear that the smooth interaction between user and computer is crucial to gain insight from big data visualization. Therefore, approaches to big data visualization should not exclusively focus on performance, the computational aspects of processing data or specific visual challenges, but also on effective interface and experience design (see **UI/UX Design** and **Usability Engineering & Evaluation**). In this way, big data visualization necessarily combines the backend (computation) and frontend (visualization) of cartography in a tight coupling, in which both human and computer work together to create new insights from data.

References

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. Applied Geography, 49, 45–53.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679.

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., et al. (2016). Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data. IEEE Transactions on Visualization and Computer Graphics, 22(1), 270–279.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. Cartography and Geographic Information Science 40 (2):130-139.

Dang, T. N., Wilkinson, L., & Anand, A. (2010). Stacking Graphic Elements to Avoid Over-Plotting. IEEE Transactions on Visualization and Computer Graphics, 16(6), 1044–1052.

DiBiase, D. (1990). Visualization in the Earth Sciences. Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences, Pennsylvania State University, 59, 13-18.

Diebold, F. X. (2012). A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version (November 26, 2012). PIER Working Paper No. 13-003.

Fox, P., & Hendler, J. (2011). Changing the Equation on Scientific Data Visualization. Science, 331(6018), 705–708.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), 211-221.

Graham, M. & Shelton, T. (2013). Geography and the future of big data, big data and the future of geography. Dialogues in Human Geography, 3(3), 255–261.

Holten, D. & Van Wijk, J. J. (2009). Force-Directed Edge Bundling for Graph Visualization. Computer Graphics Forum, 28(3), 983–990.

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

Kitchin, R. andMcArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1).

Kitchin, R. M. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 1–12.

Kraak, M-J. (1988). Computer-assisted cartographical three-dimensional imaging techniques (Doctoral Dissertation). Delft University Press, Delft.

Kraak, M-J., & Kveladze, I. (2017). Narrative of the annotated Space–Time Cube–revisiting a historical event. Journal of Maps, 13(1), 56-61.

Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plots—rational approach to visualizing networks. Briefings in Bioinformatics, 13(5), 627–644.

Kumar, S., Morstatter, F., & Liu, H. (2014). Twitter Data Analytics. New York, NY: Springer New York.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety.

META Group Research Note, 6.

Liu, Z., Jiang, B., & Heer, J. (2013). imMens: Real-time Visual Querying of Big Data. Computer Graphics Forum, 32(3), 421–430.

Nost, E., Rosenfeld, H., Vincent, K., Moore, S. A., & Roth, R. E. (2017). HazMatMapper: an online and interactive geographic visualization tool for exploring transnational flows of hazardous waste and environmental justice. Journal of Maps, 13(1), 14–23.

Poorthuis, A., & Zook, M. A. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. Journal of Urban Technology, 36, 1–21.

Poorthuis, A., Zook, M. A., Shelton, T., Graham, M., & Stephens, M. (2015). Using Geotagged Digital Social Data in Geographic Research. In N. Clifford, S. French, M. Cope, & S. Gillespie (Eds.), Key Methods in Geography (3rd ed.).

Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., et al. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. International Journal of Cartography, 18(5), 1–29.

Sarikaya, A. T. (2017). Targeting Designs of Scalable, Exploratory Summary Visualizations (Doctoral Dissertation). The University of Wisconsin - Madison, Madison, WI.

Shneiderman, B. (2014). The Big Picture for Big Data: Visualization. Science, 343(6172), 730–730.

Zhao, B., & Sui, D. Z. (2017). True lies in geospatial big data: detecting location spoofing in social media. Annals of GIS, 23(1), 1–14.

Zook, M. A., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., et al. (2017). Ten simple rules for responsible big data research. PLoS Computational Biology, 13(3), e1005399.

Zwitter, A. (2014). Big Data ethics. Big Data & Society, 1(2), 1–6.