B

Big Geo-Data

Song Gao
Department of Geography, University of California, Santa Barbara, CA, USA

Synonyms

Big georeferenced data; Big geospatial data; Geospatial big data; Spatial big data

Definition/Introduction

Big geo-data is an extension to the concept of big data with emphasis on the geospatial component and under the context of geography or geosciences. It is used to describe the phenomenon that large volumes of georeferenced data (including structured, semi-structured, and unstructured data) about various aspects of the Earth environment and society are captured by millions of environmental and human sensors in a variety of formats such as remote sensing imageries, crowdsourced maps, geotagged videos and photos, transportation smart card transactions, mobile phone data, location-based social media content, and GPS trajectories. Big geo-data is "big" not only because it involves a huge volume of georeferenced data but also because of the high velocity of generation streams, high dimensionality, high variety of data forms, the veracity (uncertainty) of data, and the complex interlinkages with (small) datasets that cover multiple perspectives, topics, and spatiotemporal scales. It poses grand research challenges during the life cycle of large-scale georeferenced data collection, access, storage, management, analysis, modeling, and visualization.

Theoretical Aspects

Geography has a long-standing tradition on the duality of research methodologies: the *law-seeking* approach and the *descriptive or explanatory* approach. With the increasing popularity of data-driven approaches in geography, a variety of statistical methods and machine learning methods have been applied in geospatial knowledge discovery and modeling for predictions. Miller and Goodchild (2015) discussed the major challenges (i.e., populations not samples, messy not clean data, and correlations not causality) and the role of theory in the data-driven geographic knowledge discovery and spatial modeling, with addressing the tensions between idiographic versus nomothetic knowledge in geography. Big geo-data is leading to new approaches to research methodologies in capturing complex spatiotemporal dynamics of the Earth and the society directly at multiple spatial and temporal scales instead of just snapshots. The data streams play a driving-force role in data-driven methods rather

L.A. Schintler, C.L. McNeely (eds.), *Encyclopedia of Big Data*,
DOI 10.1007/978-3-319-32001-4_492-1

than a test or calibration role behind the theory or models in conventional geographic analyses.

While data-driven science and predictive analytics evolve in geographic and provide new insights, sometimes it is still very challenging for humans to interpret the meanings of machine learning or analytical results or relate findings to underlying theory. To solve this problem, Janowicz et al. (2015) proposed a semantic cube to illustrate the need for semantic technologies and domain ontologies to address the role of diversity, synthesis, and definiteness in big data researches.

Social and Human Aspects

The emergence of big geo-data brings new opportunities for researchers to understand our socio-economic and human environments. In the journal of *Dialogues in Human Geography* (volume 3, Issue 3, November 2013), several human geographers and GIScience researchers discussed a series of theoretical and practical challenges and risks to geographic scholarship and raised a number of epistemological, methodological, and ethical questions related to the studies of big data in geography. With the advancements in location-awareness technology, information and communication technology, and mobile sensing technology, researchers employed emerging big geo-data for investigating the geographical perspective of human dynamics research within such contexts in the special issue on *Human Dynamics in the Mobile and Big Data Era* on the *International Journal of Geographical Information Science* (Shaw et al. 2016). By synthesizing multi-sources of big data, those researches can uncover interesting human behavioral patterns that are difficult or impossible to uncover with the traditional datasets. However, challenges still exist in the scarcity of demographics and cross-validation or getting the identity of individual behaviors rather than aggregated patterns. Moreover, the location-privacy concerns and discussions arise in both academic world and the society. There exist social tensions among big data accessibility and privacy protection.

Technical Aspects

Cloud computing technologies and their distributed deployment models offer scalable computing paradigms to enable big geo-data processing for scientific researches and applications. In the geospatial research world, cloud computing has attracted increasing attention as a way of solving data-intensive, computing-intensive, and access-intensive geospatial problems and challenges, such as supporting climate analytics, land-use and land-cover change analysis, and dust storm forecasting (Yang et al. 2017). Geocomputation facilitates fundamental geographical science studies by synthesizing high-performance computing capabilities with spatial analysis operations, with providing a promising solution to aforementioned geospatial research challenges.

There are a variety of big data analytics platforms and parallelized database systems emerging in the new era. They can be classified into two categories: (1) the massively parallel processing data warehousing systems like *Teradata* which are designed for holding large-scale structured data and support standard SQL queries and (2) the distributed file storage systems and cluster-computing framework like *Apache Hadoop* and *Apache Spark*. The advantages of Hadoop-based systems mainly lie in their high flexibility, scalability, low cost, and reliability for managing and efficiently processing a large volume of structured and unstructured datasets, as well as providing job schedules for balancing data, resources, and task loads. A MapReduce computation paradigm on Hadoop takes the advantages of divide-and-conquer strategy and improves the processing efficiency. However, big geo-data has its complexity on the spatial and temporal components and requires new analytical framework and functionalities compared with nonspatial big data. Gao et al. (2017) built a scalable Hadoop-based geoprocessing platform (GPHadoop) and ran big geo-data analytical functions to solve crowdsourced gazetteers harvesting problems. Recently, more efforts have been made in connecting traditional GIS analysis research community to the cloud computing research community for the next frontier of big geo-data analytics.

In one special issue on *big data* at the journal of *Annals of GIS* (volume 20, Issue 4, 2014), researchers further discussed several key technologies (e.g., cloud computing, high-performance geocomputation cyberinfrastructures) for dealing with quantitative and qualitative dynamics of big geo-data. Advanced spatiotemporal big data mining and geoprocessing methods should be developed by optimizing the elastic storage, balanced scheduling, and parallel computing resources in high-performance geocomputation cyberinfrastructures.

Conclusion

With the advancements in location-awareness technology and mobile distributed sensor networks, large-scale high-resolution spatiotemporal datasets about the Earth and the society become available for geographic research. The research on big geo-data involves interdisciplinary collaborative efforts. There are at least three research areas that require further work: (1) the systematic integration of various big geo-data sources in geospatial knowledge discovery and spatial modeling, (2) the development of advanced spatial analysis functions and models, and (3) the advancement of quality assurance issues on big geo-data. Finally, there will still be ongoing comparisons between data-driven and theory-driven research methodologies in geography.

Further Readings

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. (2017). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems, 61*, 172–186.

Janowicz, K., van Harmelen, F., Hendler, J., & Hitzler, P. (2015). Why the data train needs semantic rails. *AI Magazine,* Association for the Advancement of Artificial Intelligence (AAAI), pp. 5–14.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *Geo Journal, 80*(4), 449–461.

Shaw, S. L., Tsou, M. H., & Ye, X. (2016). Editorial: Human dynamics in the mobile and big data era. *International Journal of Geographical Information Science, 30*(9), 1687–1693.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big data and cloud computing: Innovation opportunities and challenges. *International Journal of Digital Earth, 10*(1), 13–53.