Big Data and Regional Science: Opportunities, Challenges, and Directions for Future Research

Laurie A. Schintler, Schar School of Policy and Government, George Mason University, USA, lschintl@gmu.edu

Manfred M. Fischer, Vienna University of Economics and Business, Austria, manfred.fischer@wu.ac.at

Abstract. Recent technological, social, and economic trends and transformations are contributing to the production of what is usually referred to as Big Data. Big Data, which is typically defined by four dimensions -- Volume, Velocity, Veracity, and Variety -- changes the methods and tactics for using, analyzing, and interpreting data, requiring new approaches for data provenance, data processing, data analysis and modeling, and knowledge representation. The use and analysis of Big Data involves several distinct stages from 'data acquisition and recording' over 'information extraction' and 'data integration' to 'data modeling and analysis' and 'interpretation', each of which introduces challenges that need to be addressed. There also are cross-cutting challenges, which are common challenges that underlie many, sometimes all, of the stages of the data analysis pipeline. These relate to 'heterogeneity', 'uncertainty', 'scale', 'timeliness', 'privacy' and 'human interaction'. Using the Big Data analysis pipeline as a guiding framework, this paper examines the challenges arising in the use of Big Data in regional science. The paper concludes with some suggestions for future activities to realize the possibilities and potential for Big Data in regional science.

Key Words: Spatial Big Data, data analysis pipeline, methodological and technical challenges, cross-cutting challenges, regional science

JEL Classification: C18, C45, C55, C82, R23

1 Introduction

Over the past two decades, we have seen a paradigm shift in the way information and data is generated and handled. This shift is driven by several factors: (i) the significant improvements in storage capacity and computing power to process very large data sets; (ii) the rapid increase in remote sensors generating new streams of digital data from telescopes, traffic monitors and video cameras monitoring the environment; (iii) the introduction of the Internet of Things, implying that even simple components and devices can communicate over the internet; (iv) the mobile revolution with the advent of mobile and smartphones enabling to receive and send information anytime and everywhere; (v) the emergence of e-commerce channels and social media platforms; and (vi) crowd-sourcing platforms for volunteered geographic information (VGI), a type of user-generated content with a geospatial component. These changes together have resulted in what is generally called Big Data.

The analysis of Big Data involves multiple distinct stages from 'data acquisition and recording' over 'information extraction' and 'data integration' to 'data modeling and analysis' and 'interpretation', each of which introduces challenges that need to be addressed. In this contribution we briefly discuss these challenges from the perspective of regional science. We begin with some definition of the notion of Big Data and its potential for regional science. Using the Big Data analysis pipeline as a guiding framework, we then discuss the challenges arising in the use of Big Data. The paper closes with some remarks on future activities to realize the potential of Big Data in regional science.

2 Big Data and Opportunities for Regional Science

2.1 What is Big Data?

The term Big Data has been widely used for any sort of data flow that is larger than usual. Big Data, however, is not just larger data, nor a question of sampling in large data flows. The crucial point of Big Data is that it changes the way to approach data analysis, requiring new processing models and knowledge representations. The challenges associated with handling and analyzing Big Data are due to four of its basic characteristics.

Volume: Big Data is about size – massive volumes of data beyond the capability of traditional approaches of data analytics. Much of the Big Data is geographic in nature containing explicit or implicit spatial information. Terabyte archives for remotely sensed imagery data, vast volumes of real-time sensor observations and location-based media data, and VGI data are examples where new innovative procedures for handling and analyzing massive volumes of spatial data had or still have to be developed.

Velocity: Big Data is generated in a very rapid pace. Traffic data in mobile communication networks and streaming video data are prime examples. The velocity of Big Data is also relevant in the Internet of Things where an up to date picture of information and a near real-time response are prerequisites.

Variety: Big Data is highly heterogeneous in nature. Conventional data analysis could – to a great extent – rely on data structured in tables and databases with entries of pre-defined types. Big Data, in contrast, is characterized by unordered and unstructured formats. The data can be, for example, map data, imagery data, geotagged text data, structured and unstructured data, raster and vector data. All these different types of data call for more efficient models, structures and data management.

Veracity: The quality of Big Data is uncertain, as the data may come from unknown or ever-changing sources. Much of the geospatial Big Data are from unverified sources, raising issues on quality assessment of such data. Analysis of Big Data will certainly require a new mindset, but also new tools and processes to handle the veracity of Big Data, i.e., avoiding noise and abnormalities in the data.

Big Data are complex in a variety of ways. They are voluminous, noisy, heterogeneous, multi-source and collected over a range of temporal and spatial scales. Spatial data may come from earth observations, social media, mobile phone calls, and unmanned aerial vehicles. Sensor technology is also being embedded in cars and containers, adding to the abundance of data. Moreover, the deployment of the Internet of Things will produce large amounts of text-like communication between devices.

Through the whole spectrum of society and business, vast volumes of data are collected on our physical and human-made environment, including building structures, nightlights, land use cover, meteorological conditions, water quality, and so on. Large-scale simulations based on this data (e.g., global climate modeling) provide an additional layer of data in Geographical Information Systems (GISs). The world wide web, and complex ecosystems of online e-commerce websites and infomediaries (e.g., job markets, dating websites, recommendation services), repositories of digitized documents, open data portals, social media platforms, and other websites it encompasses, give us a rich and unfolding picture of the interests, preferences, needs, and activities of individuals, organizations, and firms in regions and cities all over the world. Web 2.0 or the interactive web and related social media platforms, ‘apps’, and discussion fora, in particular, have created a new generation of sensors, namely humans (or citizens) as sensors. Mobile devices including smart phones and location acquisition technologies such as global position systems are producing realms of spatial trajectory data, that capture detailed information on human, material, and information, and animal movements.

Emerging technologies, such as computational intelligence, block chain, nanotechnology, cloud robotics, and so on, are contributing to even newer sources of Big Spatial Data. Such cutting-edge technological innovations are also advancing our capacity to store, process, and glean insight and intelligence from Big Data. The Internet of Things, which comprises a large and growing assemblage of interconnected devices, is actively monitoring and intelligently processing everything from the contents of our refrigerators, for example, to the second-to-second operational characteristics of large-scale infrastructure. Cyber-physical systems, which integrate computing, networking, and physical technologies in a complex and adaptive fashion, are a burgeoning source of Big Data. For example, automated vehicles collect vast amounts of real-time data about traffic conditions and other aspects of the surrounding environment,

information that is instantaneously fed back to the cloud for processing to optimize vehicular routing and performance. Indeed, machine-generated data – i.e., raw data produced and processed by machines – is a rapidly expanding source of data. In fact, machine-generated data could soon make up 50 percent of all of the data in the world (Gantz and Reinsel, 2012).

Just like a-spatial Big Data, geospatial Big Data (or Big Spatial Data) contains disparate formats, structures, semantics, granularity, and so on. However, space and time dimensions of the data add further heterogeneity. To this point, spatial data comprises varying spatial and temporal scales, levels of resolution, and extents of coverage, and with different spatial referencing systems (Fischer, Scholten and Unwin, 1996). Citizen sensing, crowd-sourced and other forms of user-generated data tends to have a high degree of spatial and temporal resolution – i.e., information that is often summarized down to latitude and longitude coordinates, and seconds of the day – and coverage that extends over the entire globe. Other types of spatial data, such as those collected from official organizations are more aggregated and limited in geographic scope. The heterogeneity of Big Data also stems from the particular characteristics of the data acquisition devices themselves. Regarding Big Spatial Data, sensors are either positioned on moving objects or static, continually monitoring the changing environment in an area or at a particular location (Li et al., 2016). Thus, spatial objects are classified geometrically as line, point, or area (Fischer and Wang, 2011).

Spatial Big Data is fraught with heterogeneity, but also with noise, incompleteness, redundancy, uncertainty, and other undesirable features. For example, sensors that monitor the environment produce repetitive coverage, since multiple images must be collected in a short amount of time to achieve appropriate and adequate spatial coverage. Mobile trace data tends towards noise and incompleteness, given that location positioning technologies are currently unable to produce proper signals in specific environments. Crowd-sourced geographic information data often contain duplicate records stemming from human error and technological glitches. Moreover, user-generated data is notoriously biased towards demographic characteristics, preferences, interests, and activity patterns of their users. The digital divide is a further source of bias and gaps in Big Data (Schintler, 2017). Given that regions have different demographic, economic, cultural, and technological profiles, the type and extent of bias vary from place to place.

2.2 Opportunities for Regional Science

Big Data can provide fresh insight into old phenomena, and a better understanding of new phenomena (Arribas-Bel, 2014). For instance, we are now able to examine the complex interplay between cyber socialization and spatial interaction. ‘Apps’ that enable users to share their consumption patterns with friends facilitate the study of the leisure class and so-called Veblen consumer (McLaughlin, Reid and Moore, 2014). Big Data can also be used to construct new notions of time such as ‘social time’ as opposed to solar or standard time (Ahas et al., 2015). Or it can be used to derive novel conceptions of space, for example, bottom-up derived characterizations of regions, as opposed to top-down defined administrative boundaries. Because spatial data tend to cover large areas, and much of it is spatially and temporally fine-grained, it

allows us to move from static to dynamic, aggregate to disaggregate, and local to global. Thus, it is now possible to gain a more robust and refined understanding of spatial and spatiotemporal statistical artifacts, such as spatial and temporal dependence, non-stationarity/heterogeneity phenomena and the modifiable areal unit problem. It also allows us to do more detailed and disaggregated transportation and spatial interaction modeling (Li et al., 2014; Fischer and Wang, 2011). Big Data also enables bottom-up, self-organizing simulation and modeling.

One concern about Big Data is that it signals an end to theory (Anderson, 2008). To this point, some see Big Data as a computational philosophy in research and practice, in which automated algorithmic processes eclipse domain expertise (Graham and Shelton, 2013). In other words, it is viewed as an approach where 'the numbers speak for themselves' (Thatcher, 2014) or where raw data replaces modeling altogether. Hence, Big Data research has been criticized for its strong reliance on supporting inductive reasoning (Li et al., 2016). On the other hand, Big Data can play a pivotal role in bridging ideographic (description-seeking) and nomothetic (law-seeking) research activities (Miller and Goodchild, 2015), and help in moving from relatively simple hypothesizes to more complex postulates and theories (Kitchin, 2014). Moreover, Big Data enables us to revisit and recalibrate old theories where limitations of traditional data constrain our ability for operationalization and testing, and to uncover complex, universal laws, and principles from micro-observations. In fact, the efforts of Zipf, Stewart, and Warntz to bring social physics to geography through large-scale numerical analyses in the 1950s led to geographic potential theory and Zipf's law, both of which are now fundamental theoretical principles in the field of regional science (Barnes and Wilson, 2014). We need to explore further the epistemological implications of Big Data and emerging data-driven methods (Kitchin, 2014).

On a more practical level, Big Data can help to support various aspects of urban planning and management, especially in a smart city context (Batty et al., 2012). Participatory sensing data provides a lower cost mechanism than traditional surveys for collecting information, and as mentioned, it allows us to study an entire population, rather than just a sample. It can also be used to support planning efforts in (and research on) regions where administrative data is lacking, incomplete, or untrustworthy. Further, Big Data can be used to track and measure phenomena where erroneous or incomplete data prohibits the development of useful and comprehensive indicators. Further, crowd-sourced data enables 'nowcasting' – or on-the-fly, near real-time forecasting of economic and other kinds of activity (Glaeser et al., 2017).

3 Challenges

In this section we briefly discuss the challenges that need to be addressed in the five distinctive stages of the data analysis pipeline that leads from 'data acquisition and recording' over 'information extraction and cleaning' and 'data integration, aggregation and representation' to 'query processing, modeling and analysis' and 'visualization and interpretation'.

3.1 Data acquisition and recording

Big Data is first acquired from some generating source (or sources) and then transmitted to storage and recorded for future use. However, given the size and speed of Big Data, it is often not possible to transfer and store all the data. Moreover, the raw data frequently contain information that may be of no interest to the user – i.e., attributes, features, geographies, and time periods that are irrelevant for the intended use of the data. Thus, the data may be compressed and filtered before it goes into storage. One challenge is that Big Data can be filtered and condensed in magnitudes of order. This problem is more severe in the case of Big Spatial Data given the additional degrees of freedom related to space and time. At the same time, one can improve the efficiency of data compression by exploiting spatial and temporal dependencies in the data (Yang and Chen, 2017).

Another challenge is to define filter compression methods in such a way that they do not discard useful information. Lossless compression techniques can preserve all the original raw data, but they fail to optimize data reduction. But, while lossy compression is effective in reducing the volume of Big Data, it comes at the cost of information loss. Information loss is especially problematic in the case of data produced by multiple sensors of different types. For Big Spatial Data, in particular, information on spatial relations and generalization can be lost in the compression. In such cases, we can employ dimensionality reduction techniques, such as clustering methods, to reduce the size the of Big Data such that there is minimal information loss. However, such processes are computationally intensive. Use of clustering algorithms explicitly designed for spatial (and spatiotemporal) data – e.g., Spatio-Temporal Density-based Spatial Clustering of Applications with Noise (ST-DBSCAN) algorithm– can help in managing this problem (Li et al., 2016). But there is also a need for more research to investigate how lossy compression can be applied such that the integrity of scientific results based on the data is preserved. Ultimately, we need to develop filtering and compression techniques that are flexible and adaptive, and tailored to the particular features of the Big Data.

Once data is acquired, transmitted and stored, information about it should be recorded and catalogued for future use, but this poses some challenges. Metadata is used to record information about the data, for example, sample size, sampling strategy, scale, availability, age, ownership, and price (if relevant) (Getis, 1999). However, creating metadata for Big Data is complicated and often impractical. One challenge is that Big Data tends to change hands frequently, where it gets repurposed, repackaged, and reprocessed at each stop (Schintler and Chen, 2017). Thus, details of the data often get lost as it travels from one person or organization to another. Moreover, attributes are sometimes hidden, as is often the case with proprietary or personally-sensitive Big Data (Getis, 1999). In crowd-sourced or user-generated data, information on the granularity of data in space and time and related details are often missing, making full and proper documentation of such data difficult (Li et al., 2016). Another challenge is to automatically generate the right metadata to describe what data to record and how it is recorded and measured. This task remains an ongoing challenge.

3.2 Information extraction and cleaning

Frequently raw data collected will not be in a format ready for analysis. For example, we must convert unstructured data in the form of text to structured data before it is suitable for using traditional modeling and analysis tool. In the case of geospatial data, it requires geocoding before using it in a GIS. Thus, we expect an information extraction process that pulls out the required information from the underlying sources and expresses it in a standard form appropriate for the intended analysis. Doing this correctly is a continuing technical challenge. Information extraction is often application-dependent, as in the case of images and videos. In regional science, we need to be able to extract information on the location of features, and the spatial context of these objects from the data. While some sources of Big Data contain explicit geographic references – e.g., latitude and longitude coordinates – many others do not. For example, in social media data geographic information is embedded in the feeds, often across multiple rather than a single entity, and the information is in poorly-defined formats. We need more research to develop efficient methods for extracting geographic information from such kinds of data sources.

Big Data involves extensive pre-processing and cleaning to remove imperfections and other undesirable features in the data. Existing work on data cleaning assumes well-recognized constraints on valid data or well-understood error models. However, Big Data often comes from unverified sources with low or unknown precision (Li et al., 2016). Further, given the transitory nature of Big Data, information on its quality may be known only to those who have produced or processed the data along the way. While there are international standards and procedures for assessing the quality of spatial data (and small data), best practices and guidelines for Big Spatial Data are lacking (Goodchild, 2013). We need procedures for attempting to ensure the quality of Big Spatial Data (Batty et al., 2012), and also to better understand how the quality and veracity of such data varies by source, type, and region (Schintler and Chen, 2017). Lastly, we need to explore how to exploit redundancy in Big Spatial Data for detecting errors and inconsistencies in the data (Goodchild, 2013).

3.3 Data integration, aggregation, and representation

Data analysis is considerably more challenging than merely locating, identifying, understanding, and recording data. It is often necessary to merge and aggregate data to make it more meaningful, computationally tractable, and compatible with other sources – e.g., administrative records. In a smart city context, there is a need to be able to integrate real-time streaming data with data from traditional cross-sectional sources, such as administrative records, for modeling of real-time problems that relate to longer-term planning objectives (Batty et al., 2012).

In general, the value of data increases, when linked with other data. Hence, data integration can act as a useful means to create value. But integration of Big Data collected from different sources is difficult due to the diversity of data types and formats, semantics, ownership, organizational structures and levels of resolution, and so on. It is even more complicated in the case of Big Spatial Data, given the varying spatial/temporal scales, levels of granularity and coverage the

data comprises (Fischer, Scholten and Unwin, 1996). To reduce the size and dimensionality of networked Big Data, such as data describing interrelated socioeconomic and transportation systems, and flows in cities and regions, network analytic methods and software solutions are badly needed (Batty et al., 2012). Data aggregation in spatial and temporal data is fraught by the modifiable areal unit and the modifiable temporal unit problems. Spatial data magnifies these issues, given that there are countless ways to parse and aggregate the data spatially and temporally. Accordingly, we need to explore rigorously how different spatial and temporal aggregations in Big Data affect patterns of association, and outcomes based on modeling and analysis of the data.

3.4 Query Processing, Modeling, and Analysis

Methods for querying and mining Big Data are fundamentally different from traditional statistical tools for small samples. Query processing intends to extract meaningful sets of observations from the raw or pre-processed data. For Big Spatial Data, spatial indexing methods, which use simple rules – such as 'find all features located in a particular region' or 'find all objects that contain a given query point' – are used for querying data. However, query processing is computationally intensive because of the polynomial complexity of the geometric operations required to pull data. Moreover, in multidimensional data, there are additional spatial relationships, which further impede the efficiency of query processing (Wang et al., 2015).

Recent research on spatial query processing of real-time streaming Big Data focuses on designing indexing methods, which segment the search space into tiles, such that search time focuses on a single tile at a time. However, an ongoing challenge is how to organize the tiles in such a way that the search process is efficient. Hilbert space-filling curves may help in addressing this concern (Li et al., 2016). When querying Big Spatial Data, we also need to ensure adequate extracting and appropriate samples from the data, as failure to do so increases the probability of erroneous conclusions. This is a challenge with immense spatial data as there are many possible realizations that can be drawn from a single source (Getis, 1999). In sum, more research is needed to develop techniques for querying Big Spatial Data to improve query speed and accuracy, while extracting appropriate and representative samples at the same time.

The key for extracting value out from data, is characteristically to build an appropriate model of the interesting aspects in data, and use that model to analyze key values, detect anomalies, patterns, relationships and trends, make predictions, and carry out other analyses. Because of this, machine learning and statistical models have received increasing attention. Conventional parametric statistical techniques are not well-suited for Big Data. Use of such methods requires that certain assumptions hold, in particular that the observations are normally and independently distributed. Big Data often violates such assumptions. In Big Spatial Data, assumptions of spatial and temporal independence of observations and non-stationarity are rarely satisfied, given the high degree of spatial and temporal dependence in the data. Moreover, problems associated with attributes, areal framework and area/attribute interaction, which are present in spatial regression modeling, are magnified with more massive spatial data sets (Getis, 1999). There are

also computational challenges when trying to apply conventional methods to large spatial data sets.

Another problem relates to the spatial weight matrix used in spatial econometric models to describe the arrangement of observational units in space. Given that computational complexity increases exponentially as the number of locations increases linearly, such models tend to suffer from the curse of dimensionality (Li et al., 2016). This problem compromises the computational efficiency of maximum likelihood estimation (a problem not arising in the case of Bayesian model estimation) in spatial autoregressive modeling (Smirnov and Anselin, 2001). For spatial network data – e.g., georeferenced social media data, where a spatial weight matrix represents relationships between origin-destination locations – this problem is even more extreme (Zhou et al., 2017). While we can apply sampling strategies to reduce the size and dimensionality of the spatial weight matrix, this approach can lead to underestimation of spatial autocorrelation (Zhou et al., 2017). 'Divide-and-conquer' methods, which iteratively reduce complex problems into subtasks, until the solution of subproblems is scalable, may be better suited for dealing with dimensionality in spatial regression modeling (Smirnov and Anselin, 2001). Further, Big Data tools can be used to create spatial weights from huge spatial data sets to manage computing resources efficiently (Li et al., 2014). More research is needed to develop methods and tools for reducing the dimensionality of spatial weight matrices for large spatial data sets, and for improving the efficiency of maximum likelihood estimation in spatial regressions involving Big Data.

Machine learning, based on well-grounded statistical models and algorithms such as reinforcement learning, support vector machines and Bayesian networks, is an important means to read out value from data. The problem of Big Data objects in machine learning is generally solved through parallelization of algorithms accomplished either by data parallelism or task parallelism. Machine learning methods can capture non-linearity, heterogeneity, noise, and other complexities in spatial and temporal data (Fischer, 2015). Feedforward neural networks, in particular, may be used for non-parametric statistical inference, as they do not require a priori specification of a specific functional form (Fischer, 2015). However, one drawback of such network models is that they generally cannot scale to large data sets, given the network structure of their underlying architectures. Further, if the statistical properties of phenomena under consideration evolve over space or time, feedforward neural network models have to be retrained to account for this (Li et al., 2016). Deep neural learning, an emerging paradigm in the era of Big Data, has enormous potential for Big Data in regional science. However, while neural networks, in general, have been applied extensively in regional science (Fischer and Gopal, 1994), deep learning has yet to catch on in the field. One issue is that deep learning is still very much a black box, with multiple layers of hidden and uninterpretable parameters. We need to better understand the inner workings of deep learning to make it a more meaningful approach. Recent efforts to develop theories of deep learning may be useful in this regard (Walchover, 2017). We also need to explore how to use deep learning (and machine learning, in general) for causal inference, and to examine further how to use machine learning techniques in combination with econometric/statistical methods to adequately address the array of challenges related to modeling of Big Data (Varian, 2014).

Other challenges relate to the validation of models based on Big Data. In particular, we must be able to assess the performance of different Big Data algorithms from a standard base, and to that point, we require standardized benchmarks and metrics (Li et al., 2016). Often analyses based on Big Data do not match 'ground truth' (Chen and Schintler, 2014). Accordingly, further research is needed to enrich the results of Big Data modeling with more reliable sources of data – e.g., survey data, administrative records.

3.5 Visualization and Interpretation

Visualization and interaction technologies may give users a gateway into their data. They can help in identifying patterns and outliers, which can reveal ways in which the data could be better partitioned for further computational analysis. Systems with a rich palette of visualization tools become essential in conveying to the users the results of the queries in a way that is best understood in the particular domain. Ultimately, display of Big Data appears to be useful only if succinctly and correctly summarizing the underlying information. Related to this, with a few clicks the user should be able to drill down into each piece of data that she sees, to learn to know its provenance, which is a key to understanding the data.

For smart cities, the challenge is to design visualization tools that enable policy and decision makers, city planners, and the community-at-large to visually explore and analyze the data for better decision making (Li et al., 2016). Dashboards and geoportals have great utility in this context (Batty et al., 2012). More research is needed to develop visualization tools that can efficiently deal with all of the dimensions of Big Data, including quality and veracity of the data. Ideally, the design of visualization should be informed by capabilities and constraints in human information processing, perception, and cognition (Li et al., 2016).

Interpretation is at the center of data analysis. Regardless of the size of the data, it is subject to limitations and bias. Without these biases and limitations being understood and outlined, misinterpretation tends to be the rule rather than an exception. One issue is that Big Data used for research is often done for purposes that deviate from the original intents of data collection, which can ultimately contribute to slanted perspectives and insights based on the data (Thatcher, 2014). Analyses based on bottom-up and other kinds of spatially and temporally refined data are prone to ecological fallacy, including the modifiable areal unit and the modifiable time unit problems. Moreover, while data mining techniques may enable us to understand patterns of association, such methods do not convey information on causation or explain the 'why and how' (Li et al., 2016). Correlations discovered in the data may, in fact, be spurious. At the same time, correlation does imply causation in some instances. With more data and techniques like Bayesian networks, we can rule out scenarios where causation is unlikely and hone in on where and why it may be present. It is critical to consider context when designing algorithms, and when interpreting results based on application of the algorithms. Failure to do so can lead to inaccurate and misleading conclusions, as happened in the case of Google Flu (Lazer et al., 2014). To avoid specious findings, we must frame and contextualize data and information in appropriate theory. Ontologies, which use a shared vocabulary to characterize the types, properties, and

interrelationships of concepts representing knowledge in a particular domain, can help in organizing the data and placing it in proper context. We also need to view data through a broader social and epistemological lens to be able accessing it in a meaningful way (Thatcher, 2014).

It is rarely enough to provide just the results. Instead, one must provide also supplementary information that explains how results obtained were derived, and based on which inputs and assumptions. This task is critical for reproducibility efforts. It is also crucial for ensuring that analyses based on Big Data are useful to end users – e.g., decision makers, researchers, policy makers, and citizens. Further, the validity of conclusions about the world drawn from Big Data is often just as much a function of the integrity of the algorithms used to process the data as the raw data itself. Thus, 'algorithmic transparency' is also imperative (Kwan, 2016).

4 Cross-Cutting Challenges

Cross-Cutting challenges are common challenges that underlie many, sometimes all, of the stages of the data analysis pipeline. These include 'heterogeneity', 'uncertainty', 'scale', 'timeliness', 'privacy' and 'human interaction'.

4.1 Heterogeneity

When humans consume information, a great deal of heterogeneity is comfortably tolerated or even desired. However, machine algorithms expect homogeneous data, and cannot easily understand nuances. Consequently, one must structure Big Data carefully as a first step in or before data analysis. To do this efficiently, we need to express differences in data structures and semantics to be shown in forms that are computer understandable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolution.

Big Data is difficult to work with, using relational database management systems and desktop statistics and visualization software. NoSQL (not only structured query languages) database management systems, instead, provide support for clouding architectures and the facility to generate patterns and trends without the need for additional infrastructure. Sometimes it is just not possible or practical to combine Big Data with varying spatial and temporal scales, hierarchies, and levels of resolution to make it compatible for analysis. In such cases, multilevel and hierarchical modeling may be appropriate. Ultimately, to deal with heterogeneity of Big Spatial Data, we need flexible, adaptive, and tailored approaches at each stage in the data pipeline. Indeed, increasing diversity of data means the need for diverse solutions.

4.2 Uncertainty

Uncertainty is present in all stages of the pipeline. Sources of uncertainty stem from the black box nature of algorithms used for analyzing Big Data, imperfections in the data (e.g., sampling

bias, incompleteness, redundancy, etc.), model selection issues, the computing environment itself, e.g., in managing resources in the cloud, but also from misinterpretations of the data and results obtained. Careful data cleaning and scrubbing is a necessary first step in minimizing uncertainty, but even after doing this some imperfections in the data are likely to remain.

If errors are present in the raw data, they can propagate to all stages in the Big Data pipeline. Recent work on managing probabilistic data and modeling suggests one way to make progress. For example, interval analysis allows one to model the uncertainty of the input variables (e.g., from sensor observations) and the corresponding uncertainty of the functions based on the variables (Li et al., 2016). Functional analysis methods (e.g., wavelets, homotopy continuation) are also useful for modeling uncertainty. Moreover, precision analysis can be used to evaluate the veracity of Big Data from the perspective of data quality, while simultaneously ensuring that the utility of the data is preserved.

To address the issue of gaps and other sources of uncertainty in Big Data, we should also consider exploiting the models of the very phenomena that the underlying data corresponds to (Batty et al., 2012). About econometric modeling, there are countless viable model approaches, methods and techniques that can be adopted in the era of Big Data. But we need systematic, comprehensive, and efficient procedures for selecting and formulating robust models (Doornik and Hendricks, 2015). The Bayesian Model Averaging approach represents a more formal Bayesian solution to the issue of model uncertainty in the context of spatial econometric models (LeSage and Fischer, 2008).

4.3 Scale

Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. A dramatic shift is underway to move towards cloud computing, which aggregates multiple disparate workloads with varying performance goals across large numbers of processors to manage computational efficiency. The level of sharing resources on expensive and large clusters requires new ways of determining how to run and execute data processing jobs and to deal with system failures. This task requires us to rethink how to design, build and operate data processing components to support activities at each stage in the data pipeline.

Partitioning techniques, divide-and-conquer, and incremental and distributed processing (including cloud computing) can help to manage the computational complexity of large data sets. Ensemble analysis, which strategically integrates multiple algorithms, can enable us to model an entire data set rather than a subsample of the data. Spatial or ‘place-based’ ensemble methods can be applied to deal with the nuances of data. However, use of ensemble methods poses some challenges, including ensuring that there is consistency between the algorithms and defining the relative weights of the algorithms. Moreover, many partitioning techniques are not yet optimized for geometric computation (Wang et al., 2015). Another approach for managing scalability issues in Big Spatial Data is to exploit complex properties of such data, e.g., fractal patterns. Indeed, data produced via bottom-up mechanisms such as crowd-sourced data tend to exhibit fractal structure and related properties, which lends itself to such tactics (Batty et al., 2012; Li et al.,

2016). In sum, we need more research to develop scalable computational and analytical processing methods and tools, especially for Big Spatial Data.

4.4 Timeliness

The design of a system that effectively deals with size is likely to result into a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of velocity in the context of Big Data. Rather, there is an acquisition rate challenge and a timeliness challenge. There are many situations in which we require the results of the analysis immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed, potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real-time. Instead, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination.

For spatial algorithms, in particular, we cannot wait until all the data are known (Li et al., 2016). A significant requirement for data-intensive spatial applications is fast query response which requires a scalable architecture that can query spatial data on large-scale data. However, speed must not come at the sacrifice of the validity and trustworthiness of the data and results based on the data (Li et al., 2016). For useful large-scale, real-time analysis of Big Data, most if not all of the processes should be automated. In other words, we need the ability to intelligently process, analyze, visualize, and interpret Big Data on the fly. We also need automated procedures for assessing the quality of Big Data (Goodchild, 2013). While techniques like complex event processing and online analytical processing are useful for managing multiple, fast-moving data streams, they are not yet able to adequately support geospatial features and computations in an efficient manner (Lee and Kang, 2015).

4.5 Privacy

Ethics and privacy are another major concern and one that increased with Big Data. This problem is present in all stages of the pipeline. Indeed, triangulation techniques can be applied to multiple sources of data to paint complete pictures of human activity (Graham and Shelton, 2015). But within just a single source of data, information about the location of individuals and their actions at those sites at particular points in time can be inferred either directly or indirectly (Schintler and Chen, 2017), especially in spatial data that has high degree of resolution (Fischer, Scholten and Unwin, 1996). However, even coarse data sets may provide little anonymity (De Montjoye et al., 2013). As more data are being made available to the public – e.g., through geoportals and open data repositories – privacy issues are becoming even more pronounced. Indeed, many online services and platforms today require us to share private information, but beyond record-level access control, we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

Accordingly, we need to develop strategies and policies for ensuring that individuals are 'in full control of their contributed data/profiles, how the data are acquired/managed, analyzed and used, when, and or how long' (Batty et al., 2012). Further, it is important to rethink security for information sharing in Big Data use cases.

Research ethics in the era of Big Data is another concern. To this point, in many research institutions, institutional review boards (IRBs) are in place to enforce the legal and ethical use of data in research. However, for certain kinds of Big Data, such as data that is publicly available or contributed voluntarily by individuals, IRB approval is not required. We can use synthetic data and de-anonymization algorithms to mask the identity of individuals and additional sensitive information in Big Data. It is also important to point out that privacy is not just an issue with the raw data. It also comes into play with the algorithms used to process the data. Algorithmic transparency is critical for ensuring that algorithms and processes do not contribute to undesirable societal outcomes, such as discriminatory policies and practices. Ultimately, managing privacy efficiently is a technical as well as a sociological issue, that has to be addressed jointly from both perspectives to realize the promise of Big Data. We need to carefully consider the privacy implications of Big Data, including open data, and design our IT systems according to well-defined principles to ensure that the personal privacy of individuals is protected in the use of Big Data.

4.6 Human Collaboration

Ideally, analytics for Big Data will be designed to have a human in the loop. Humans are needed to understand the context, adequately frame analyses using Big Data, and position models in appropriate theoretical and empirical contexts. The new field of visual analytics is attempting to do this, at least concerning the modeling and analysis stage in the pipeline. A popular new method of harnessing human ingenuity to solve problems is through crowd-sourcing. However, in this context, we are relying on information provided by unvetted strangers. While most such errors will be detected and corrected by others in the crowd, we need technologies to facilitate this. Related to this, we need the very users who are producing the data in the first place to be part of the process of creating metadata.

In a smart city context, community participation and engagement are critical for ensuring the creation of reliable, timely and trustworthy information about collective phenomena (Batty et al., 2012). Another issue relates to bots – or automated software agents – in crowd-sourced fora and social media. In fact, their presence and influence in such data are far from trivial (Varol et al., 2017). We need methods for efficiently and effectively identifying and removing records in data that relate to bots, as they may not necessarily reflect human behavior and intents (Schintler, 2017). Lastly, hybrid forecasting, which integrates human and machine learning processes, is an emerging paradigm that deserves further attention. Such systems have the potential to address the shortcomings of human forecasting (e.g., cognitive biases) while exploiting the advantages of machine-generated approaches (hybridforecasting.com).

5 Closing remarks

In recent years, geotagged data are generated at a dramatic pace. It is straightforward to acknowledge that the more data we have, the more insight we can obtain from it. But we have also to point out that the volume, the updating velocity and the variety of data are too big, too fast and also too diverse for existing regional science methods and spatial analysis tools. Regional scientists may be not willing to wait for weeks to process terabyte-scale geotagged data streams.

Fortunately, however, several Big Data processing programming models and frameworks, such as MapReduce and Hadoop, have been designed as useful environments that provide parallel processing of large-scale data in a timely, failure-free, scalable and load balance manner. This will diminish the efforts of redesigning spatial analysis tools, adapting them towards online analytical processing and querying/reporting dataware housing tools, and implementing them on top of the distributed systems.

Spatial data mining and knowledge discovery represent an important direction in the development of new generation of spatial analysis tools appropriate in a Big Data environment. The challenges described in this chapter, however, have to be addressed before the potential of Big Data can be realized in regional science and beyond. The challenges include not only the issues of scale and timeliness, but also heterogeneity, error-handling and visualization at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains. It would be not cost-effective to address them in the context of one domain only. We should, instead, support and encourage fundamental interdisciplinary research towards addressing these technical challenges to achieve the promised benefits of Big Data in academia.

References

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M. and Zook, M., 2015. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, *29*(11), pp.2017-2039.

Anderson, C., 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, *16*(7), pp.16-07.

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

Barnes, T.J. and Wilson, M.W., 2014. Big data, social physics, and spatial analysis: The early years. *Big Data & Society*, *1*(1), p.2053951714535365.

Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G. and Portugali, Y., 2012. Smart cities of the future. *The European Physical Journal Special Topics*, *214*(1), pp.481-518.

Chen, Z. and Schintler, L.A., 2015. Sensitivity of location-sharing services data: evidence from American travel pattern. *Transportation*, *42*(4), pp.669-682.

De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M. and Blondel, V.D., 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, p.1376.

Doornik, J.A. and Hendry, D.F., 2015. Statistical model selection with "Big Data". *Cogent Economics & Finance*, *3*(1), p.1045216.

Fischer, M.M., 2015. Neural networks. A class of flexible non-linear models for regression and classification. *Handbook of Research Methods and Applications in Economic Geography; Karlsson, C., Andersson, M., Norman, T., Eds*, pp.172-192.

Fischer, M.M. and Wang, J., 2011. *Spatial data analysis: models, methods and techniques*. Springer Science & Business Media.

Fischer, M.M., Scholten H., and Unwin, D. (eds.), 1996. Spatial analytical perspectives on GIS. Taylor & Francis, Basingstoke.

Gantz, J. and Reinsel, D., 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView: IDC Analyze the future*, *2007*(2012), pp.1-16.

Getis, A., 1999. Some thoughts on the impact of large data sets on regional science. *The Annals of Regional Science*, *33*(2), pp.145-150.

Glaeser, E.L., Kim, H. and Luca, M., 2017. *Nowcasting the Local Economy: Using Yelp Data to Measure Economic Activity* (No. w24010). National Bureau of Economic Research.

Goodchild, M.F., 2013. The quality of big (geo) data. *Dialogues Human Geogr*. 3 (3), 280–284.

Gopal, S. and Fischer, M.M., 1996. Learning in Single Hidden-Layer Feedforward Network Models: Backpropagation in a Spatial Interaction Modeling Context. *Geographical Analysis*, *28*(1), pp.38-55.

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

Kitchin, R., 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), p.2053951714528481.

Kwan, M.P., 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, *106*(2), pp.274-282.

Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), pp.1203-1205.

Lee, J.G. and Kang, M., 2015. Geospatial big data: challenges and opportunities. *Big Data Research*, *2*(2), pp.74-81.

LeSage, J.P. and Fischer, M.M., 2008. Spatial growth regressions: model specification, estimation and interpretation. *Spatial Economic Analysis*, *3*(3), pp.275-304.

Li, S., Dragicevic, S., Castro, F.A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A. and Cheng, T., 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, pp.119-133.

Li, X., Li, W., Anselin, L., Rey, S. and Koschinsky, J., 2014, November. A MapReduce algorithm to create contiguity weights for spatial analysis of big data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (pp. 50-53). ACM.

McLaughlin, R. B., Reid, N., & Moore, M. S. (2014). The ubiquity of good taste: A spatial analysis of the craft brewing industry in the United States. In *The geography of beer* (pp. 131-154). Springer Netherlands.

Miller, H.J. and Goodchild, M.F., 2015. Data-driven geography. *GeoJournal*, *80*(4), pp.449-461.
Schintler, L.A., 2017. The constantly shifting face of the digital divide: Implications for Big Data, urban informatics and regional science. In *Big Data for Regional Science*. Routledge.
Schintler, L.A. and Chen, Z. eds., 2017. *Big Data for Regional Science*. Routledge.
Smirnov, O. and Anselin, L., 2001. Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics & Data Analysis*, *35*(3), pp.301-319.
Thatcher, J., 2014. Big data, big questions| Living on fumes: Digital footprints, data fumes, and the limitations of spatial big data. *International Journal of Communication*, *8*, p.19.
Varian, H.R., 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), pp.3-28.
Varol, O., Ferrara, E., Davis, C.A., Menczer, F. and Flammini, A., 2017. Online human-bot interactions: Detection, estimation, and characterization. *arXiv preprint arXiv:1703.03107*.
Walchover, N., 2017. New theory cracks open the black box of deep learning. *Quanta Magazine*.
Wang, F., Aji, A. and Vo, H., 2015. High performance spatial queries for spatial big data: from medical imaging to GIS. *Sigspatial Special*, *6*(3), pp.11-18.
Zhou, J., Tu, Y., Chen, Y. and Wang, H., 2017. Estimating spatial autocorrelation with sampled network data. *Journal of Business & Economic Statistics*, *35*(1), pp.130-138.