Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities

Michael R. Evans
Dev Oliver
KwangSoo Yang
Xun Zhou
Shashi Shekhar

1 Introduction

Recent years have seen the emergence of many new and valuable spatial datasets. Examples include trajectories of cell-phones and Global Positioning System (GPS) devices, vehicle engine measurements, global climate models (GCM) simulation data, volunteered geographic information (VGI), geosocial media, tweets, etc.

The value of these datasets is already evident. For example, while monitoring tweets, the American Red Cross learned of a tornado touchdown in Texas before news reports [27]. Google has been able to estimate flu activity from search terms [23]. Everyday citizens around the world shape pop culture globally via crowd-sourced talent identification (e.g., Justin Bieber and Psy's breakthrough via YouTube).

However, these location-aware datasets are of a volume, variety, and velocity that exceed the capability of current CyberGIS technologies. We refer to these datasets as Spatial Big Data (SBD).

1.1 Defining Spatial Big Data

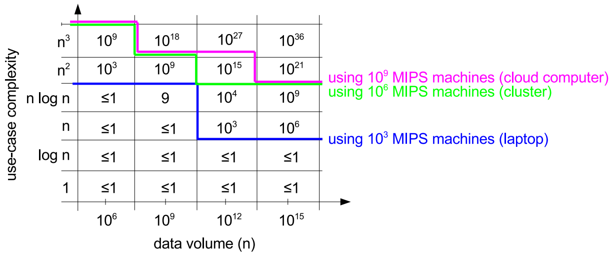
Whether spatial data is defined as "Big" depends on the context. Spatial big data cannot be defined without reference to value proposition (use-case) and user experience, elements which in turn depend on the computational platform, use-case, and dataset at hand. User experience may be unsatisfactory due to computational reasons that often stem from workloads exceeding the capacity of the platform (Table 1). For example, users may experience unacceptable response times, which may be caused by high data volume during

Computer Science Department, University of Minnesota, Minneapolis, MN

Table 1 Unsatisfactory user experience due to computational reasons

Challenging user experiences on a given platform	Data Attribute	Use-Case (Value Proposition)
Unacceptable response time	Volume	● Correlation, Optimization ● Mapping current check-ins ● Kriging crowd-sourced temperature data
Frequent data loss, system failures	Velocity	● Real-time monitoring of moving objects ● Real-time map of all smart phones ● Real-time map of tweets related to disasters
Large human effort to accomplish task	Variety	● Fusion of multiple data sources (e.g., Google time-lapse like video for history or projected future of Earth) ● Map of post-disaster situation on the ground

correlation or optimization. Users may also experience frequent data loss due to high data velocity relative to the data ingest capacity of the computational platform or they may find themselves expending large amounts of effort to pre-process or post-process SBD due to its high variety.



use-case complexity	10^6	10^9	10^{12}	10^{15}
n^3	10^9	10^{18}	10^{27}	10^{36}
n^2	10^3	10^9	10^{15}	10^{21}
n log n	≤1	9	10^4	10^9
n	≤1	≤1	10^3	10^6
log n	≤1	≤1	≤1	≤1
1	≤1	≤1	≤1	≤1

Fig. 1 Relationship between data volume and use-case complexity. (Best in color)

Figure 1 further illustrates the interplay between data volume and use case complexity. For example, while cloud computers that can process 10^9 million instructions per second (MIPS) may be able to handle data volumes of 10^{12} for use-cases of n^2 complexity (e.g., Kriging crowd-sourced temperature data), clusters capable of processing 10^6 MIPS would not be able to handle these use-cases. In this scenario, therefore, the same 10^{12} data volume (for n^2 complexity use-cases) represents spatial big data for the cluster, but not for the cloud computer.

Table 2 Example use-cases with data-volume sensitive complexities

Use-Case	Complexity
Mapping current check-ins	$O(n)$
Delineate river basins of a given elevation grid map	$O(nlogn)$
Kriging crowd-sourced temperature data	$O(n^2)$
Spatial auto-regression parameter estimation	$O(n^3)$

1.2 SBD vs. Spatial Data

Table 3 contrasts traditional spatial data with SBD, assuming simple use-cases requiring a single scan of data on a pre-cloud computing platform (e.g., $1 - 10^4$ CPUs, each with 10^3 MIPS). Traditional spatial data includes point-data (e.g., crime reports), linear-data (e.g., road centerline), raster (e.g., satellite imagery), and graphs (e.g., traditional roadmaps). A typical use of such data might be mapping presidential voter preferences across the country during an election year. Examples of SBD include check-ins [1], GPS-tracks from smart phones, Unmanned aerial vehicle (UAV)/Wide area motion imagery (WAMI) [31] video, temporally detailed roadmaps, Waze, Open Street Map, etc. While traditional spatial data might be used to map the presidential election voter preferences across the country, use-cases for SBD involve showing (near) real-time maps of tweets related to disasters and showing (near) real-time maps of traffic using Waze [61] user-generated content. The volume of traditional spatial data is limited to about 10^6 crime reports/year or gigabytes of roadmaps. SBD's volume is significantly higher. GPS traces can reach 10^{14} items per year and temporally detailed maps can reach 10^{13} items per year, given constant minute-resolution measurements.

Table 3 Traditional Spatial Data vs. Spatial Big Data (for simple use-cases requiring a scan of the dataset on pre-cloud computing platforms)

	Traditional Spatial Data	Emerging Spatial Big Data
Simple Use Cases	• Map of 2012 presidential election voter preferences	• Show (near) real-time map of tweets related to disasters • Show (near) real-time map of traffic using Waze user-generated content
Examples	• Point-data, e.g., crime reports • Linear-data, e.g., road centerline • Raster, e.g., satellite imagery • Graph, e.g., traditional roadmaps	• Check-ins • Gps-tracks from smart phones • UAV/WAMI Video • Temporally detailed roadmaps, Waze, Open Street Map
Volume	• 10^6 crime reports/year • Gigabytes of roadmaps • Decadal census	• 10^{14} GPS traces (10^9 smart phones at 10^6 readings/year with 10% of users allowing traces) • Temporally detailed maps can reach 10^{13} items per year given constant minute-resolution measurements
Variety	• Eulerian frame of reference • Raster, vector, graph	• Lagrangian frame of reference • Temporal graph • Moving objects • Spatial time-series
Velocity	• Limited velocity in traditional spatial data (e.g., new census/decade)	• High velocity (e.g., Show near real-time map of 400 million tweets/day related to disasters)

1.3 SBD vs. Big Data

Table 4 summarizes the differences between big data and SBD, assuming a pre-cloud computing platform with $1 - 10^4$ central processing units (CPUs), each with 10^3 MIPS. Examples of Big Data include Google search terms, clicks on web-page elements, Facebook posts, etc. In contrast, SBD examples include check-ins, geo-located tweets and posts, GPS tracks from all sensors, climate observations and projections, Open Street Map, Waze, etc. Datatypes for big data entail text keywords and web logs whereas SBD datatypes include global climate models (GCM) projecting UAV video, GPS traces, and temporally-detailed roadmaps. Big data explores questions such as: What are (previously unknown) side-effects of FDA-approved medicines? On the other hand, SBD raises questions such as: What are hotspots of spring-break vacationers today? What are critical places with many smart phone users in the last hour? Are there any hotspots of recent disaster-related tweets? Where? Are there traffic congestion areas reported by Waze? Representative computational paradigms for big data are Hadoop, Round-robin data partition, and Hashing (key-value store), e.g., Big Table. By contrast, SBD's computational paradigms entail Spatial Hadoop [5] / GIS on Hadoop [43] and declustering or data division.

Table 4 Big Data vs. Spatial Big Data

	Big Data	Spatial Big Data
Examples	• Google search terms • Clicks on web-page elements • Facebook posts	• Check-ins, Geo-located tweets and posts • GPS tracks from all sensors • Climate observations and projections • Open Street Map, Waze
Data Types	• Text keywords • Web logs	• GCM projecting UAV video • GPS traces • temporally-detailed roadmaps
Questions	• What are (previously unknown) side-effects of FDA-approved medicines?	• What are hotspots of spring-break vacationers today? • What are critical places with many smart phone users in the last hour? • Are there any hotspots of recent disaster-related tweets? Where? • Are there traffic congestion areas reported by Waze?
Representative Computational Paradigms	• Hadoop • Round-robin data partition • Hashing (key-value store), e.g., Big Table	• Spatial Hadoop / GIS on Hadoop • Declustering • Space partitioning is not efficient for nearest neighbors

1.4 Relationship with CyberGIS

CyberGIS seeks to synthesize cyberinfrastucture, GIS, and spatial analysis. The implementation of domain decomposition and task scheduling for parallel and distributed computing are often tailored to exploit the spatial characteristics of specific types of spatial data and analysis methods [60]. Spatial big data represents the next frontier of datasets that CyberGIS needs to be tailored towards. By expanding cyber-infrastructure and hence CyberGIS, we can harness the power of these massive spatial datasets to transform society.

Table 5 compares pre-cloud CyberGIS with Big CyberGIS, the confluence of CyberGIS and spatial big data. In terms of cyberinfrastructure, pre-cloud CyberGIS includes Message Passing Interface (MPI) [16], Open Multi-Processing (Open MP) [16], and Unix clusters with 10^2 to 10^4 nodes. Users typically need to learn detailed programming and performance tuning. Failure of a node during a long simulation usually requires restart. In contrast, Big CyberGIS includes Hadoop, Hadoop Distributed File System (HDFS) [12], and MapReduce [18] with 10^6 nodes. Big CyberGIS features simpler programming paradigms with graceful recovery and minor slow downs. The file system goals of traditional CyberGIS are geared towards performance whereas Big CyberGIS places fault tolerance and usability before performance.

Pre-cloud CyberGIS included traditional tools, e.g., ArcGIS or Postgres/PostGIS, whereas Big CyberGIS includes emerging tools, e.g., GIS on Hadoop or Spatial Hadoop. In terms of spatial analysis, CyberGIS handles R libraries or Arc/Geostatistics whereas CyberGIS facilitates MapR, Ricardo, Mahout, etc. (i.e., scalable spatial data analytics libraries for spatial big data).

1.5 Related Work, Contributions, Scope, and Outline

Related Work: Previous work has focused on specific aspects of spatial big data such as algorithms [59], benchmarking [50], or specific use-cases [51]. However, they do not consider value proposition and user experience. They also lack a broad overview of the challenges and opportunities available when spatial big data is enabled via next-generation CyberGIS.

Contributions: The main contributions of this chapter are to 1) define spatial big data in terms of value proposition and user experience, which depends on the computational platform, use-case, and dataset at hand and 2) provide an overview of the current efforts, challenges and opportunities available when spatial big data is enabled via next-generation CyberGIS. We approach the discussion from both an analytics and infrastructure perspective. From an analytics perspective, we expound on current accomplishments, e.g., GIS on Hadoop, and four novel opportunities that SBD provides, i.e., estimating spatial neighbor relationships, supporting place-based ensemble models, simplifying spatial models, and on-line spatio-temporal data analyt-

Table 5 Pre-cloud Computing CyberGIS vs. Big CyberGIS

	Pre-cloud Computing CyberGIS	Big CyberGIS
Cyberinfrastructure	• MPI, Open MP • Unix clusters with 10^2 10^4 nodes • User needs to learn detailed programming and performance tuning • Failure of a node during a long simulation requires restart	• Hadoop, HDFS, MapReduce • 10^6 nodes • Simpler programming paradigms • Graceful recovery and minor slow downs
File system goals	Geared towards performance	• Fault tolerance and Usability before performance • Economic cost
GIS	ArcGIS, Postgres/PostGIS	GIS on Hadoop, Spatial Hadoop
Spatial Analysis	R libraries, Arc/Geostatistics	Hadoop port of R, e.g., MapR
Use Cases	Map of 2012 presidential election voter preferences	• Show (near) real-time map of tweets related to disasters • Show (near) real-time map of traffic using Waze user-generated content

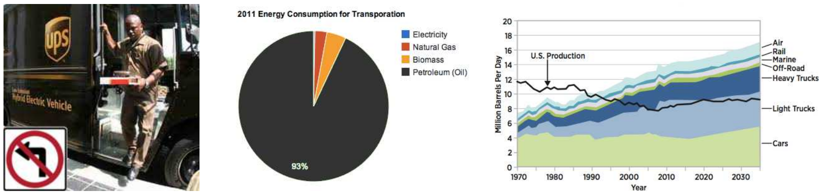
ics. From an infrastructure perspective, we discuss current accomplishments (e.g., Spatial Hadoop), and three new opportunities, i.e., computing spatial auto-correlation of check-ins, parallelizing range queries on polygonal maps, and parallelizing spatial auto-regression (SAR) and spatial graph algorithms.
Scope: This chapter focusses on spatial big data. Detailed discussion of Big Data and Spatial data are outside the scope of the present research, even though they are contrasted briefly with spatial big data. Pre-cloud computing platforms [47, 48] and their spatial use-cases will not be discussed. SBD may also have an impact on scientific knowledge production methodologies, which is currently not explored.
Outline: The rest of this chapter is organized as follows: Section 2 presents compelling societal applications of spatial big data. Section 3 lists and illustrates common types of SBD. Section 4 discusses novel opportunities in spatial big data analytics. Section 5 gives an overview of SBD infrastructure and Section 6 concludes the chapter.

2 Societal Applications of Spatial Big Data

We believe that harnessing Spatial Big Data (SBD) will enable a number of transformative societal applications. This section illustrates societal applications in the context of understanding climate change, next-generation routing services, and emergency and disaster response.

Climate Change: Climate change has been identified by the United Nations as one of the 15 most significant challenges facing human society [2]. Due to the limitations of existing physics-based models, there are still considerable uncertainties regarding the social and environmental impact of climate change. As an alternative solution, data driven approaches hold significant potential for application in environmental sciences due to the availability of tremendous amounts of climate and ecosystem data from satellite and ground-based sensors, observational records for atmospheric, oceanic, and terrestrial processes, and physics-based climate model simulations. For example, with the historical observation data available, climate scientists may discover previously unidentified changes in precipitation over the past century and build models to predict future trends of global climate.

Next-Generation Routing Services: A 2011 McKinsey Global Institute report estimated savings of "about $600 billion annually by 2020" in terms of fuel and time saved [37] by helping vehicles avoid congestion and reduce idling at red lights or left turns. Preliminary evidence for the transformative potential includes the experience of UPS, which saves millions of gallons of fuel by simply avoiding left turns (Figure 2(a)) and associated engine-idling when selecting routes [33]. Immense savings in fuel-cost and greenhouse gas (GHG) emissions are possible in the future if other fleet owners and consumers avoided left-turns and other hot spots of idling, low fuel-efficiency, and congestion. 'Eco-routing' may help identify routes which reduce fuel consumption and GHG emissions, as compared to traditional routing services reducing distance traveled or travel-time. Eco-routing has the potential to significantly reduce US consumption of petroleum, the dominant source of energy for transportation (Figure 2(b)). It may even reduce the gap between domestic petroleum consumption and production (Figure 2(c)), helping bring the nation closer to the goal of energy independence [57].



(a) UPS avoids left-turns to save fuel [33].

(b) Petroleum is dominant energy source for US Transportation [58].

(c) Gap between US petroleum consumption and production is large and growing [9, 17].

Fig. 2 Eco-routing supports sustainability and energy independence. (Best in color)

SBD for next-generation routing services includes GPS trace data, engine measurements, and temporally-detailed roadmaps. For these SBD, a key hurdle is the dataset size. For example, GPS traces can reach 10^{13} items per year given constant minute-resolution measurements for all 100 million US vehicles, engine-measurement data may have 10^{14} data-items per year for measurements of 10 engine variables, once a minute, over the 100 million US vehicles in existence, and temporally-detailed roadmaps may exceed 10^{13} items per year for the 100 million road-segments in the US when associated with per-minute values for speed or travel-time.

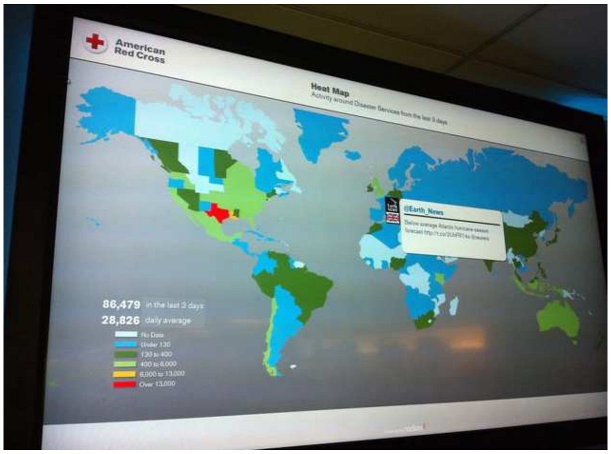
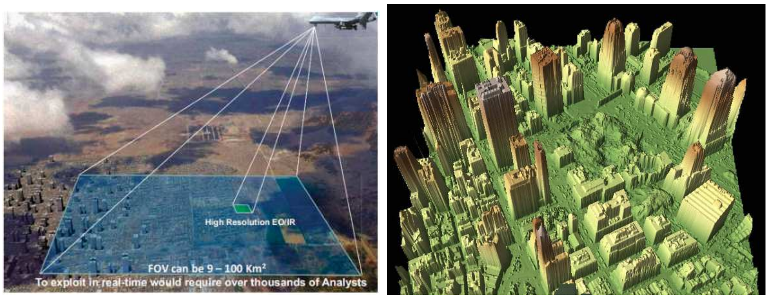


Fig. 3 The Red Cross' new social media center leveraging social media for disaster monitoring [27]. (Best in color)

Emergency and Disaster Response Disaster response agencies are leveraging geo-social media and Volunteered Geographic Information (VGI) such as tweets, check-ins, Waze, and traffic reports. Figure 3 shows an example where the Red Cross has leveraged tweets for disaster monitoring. Indeed, even before cable news outlets began reporting the tornadoes that rippled through Texas in November 2013, a map of the state began blinking red on a screen in the Red Cross' new social media monitoring center [27], alerting weather watchers that something was happening in the hard-hit area.

3 Types of Spatial Big Data

Spatial data are discrete representations of continuous phenomena over the surface of our changing planet. Discretization of continuous space is necessitated by the nature of cyber representation. There are three basic models to represent spatial data: raster (grid), vector, and network. Satellite images are examples of raster data. Vector data consists of points, lines, polygons and their aggregate (or multi-) counterparts. Graphs consisting of spatial networks are an important data type used to represent transportation networks.


High Resolution EO/IR
FOV can be 9 – 100 Km²
To exploit in real-time would require over thousands of Analysts

(a) (b)

Fig. 4 (a) Wide-area persistent surveillance. FOV: Field of view. (Photo courtesy of the Defense Advanced Research Projects Agency.) EO: Electro-optical. [31] (b) LIDAR images of ground zero rendered Sept. 27, 2001 by the U.S. Army Joint Precision Strike Demonstration from data collected by NOAA flights. Thanks to NOAA/U.S. Army JPSD. (Best in color)

Raster data, such as geo-images (Google Earth), are frequently used for remote sensing and land classification. New spatial big raster datasets are emerging from a number of sources.

GCM data: Global Climate Models (GCM) simulation data are generated via computer simulations of atmospheric and oceanic systems based on various models of their physical, chemical, and biological mechanisms. Typically, GCM data contains the major attributes (temperature, precipitation, sea level pressure, etc) across all or part of the Earth's surface in a long span of time (e.g., a hundred years). These datasets can be used to validate the hypothesis of the climate system by comparing their outputs with actual observations; they can also be used to project the future trend of global climate. However, these data have unprecedented volume, velocity and variety that exceed the capability of current cyberGIS tools. For example, a GCM dataset of daily temperature, precipitation, and sea level pressure at 0.5 by 0.5 degree spatial resolution with 20 vertical layers for the next 100 years may reach 100 Terabytes (TB). In addition, GCM data vary across a wide range of spatial and temporal scales, and attributes. Managing and analyzing these data to improve the understanding of climate change is challenging with the limited capabilities of existing cyberGIS infrastructure.

Unmanned aerial vehicle (UAV) Data: Wide area motion imagery (WAMI) sensors are increasingly being used for persistent surveillance of large areas, including densely populated urban areas. The wide-area video coverage and 24/7 persistence of these sensor systems allow for new and interesting patterns to be found via temporal aggregation of information. However, there are several challenges associated with using UAVs in gathering and managing raster datasets. First, UAV has a small footprint due to the relatively low flying height; therefore, it captures a large amount of images in a very short period of time to achieve the spatial coverage for many applications. Storing a rapidly increasing number of digital images poses a significant challenge.

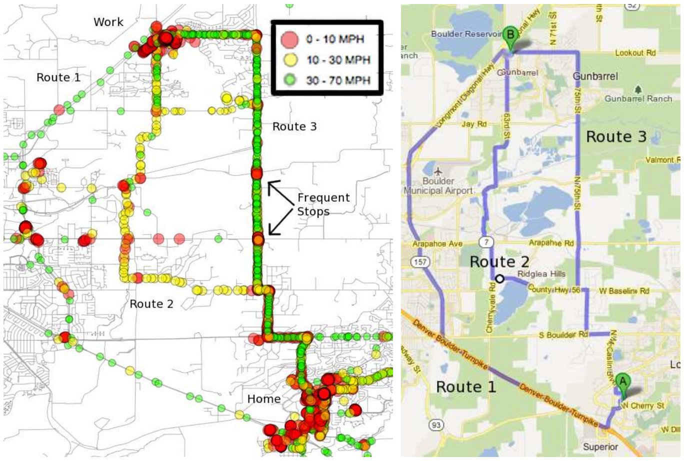
Image processing becomes time consuming and costly because it is hard to rectify and mosaic UAV photography for large areas. The large quantity of data far exceeds the capacity of the available pool of human analysts [42]. Developing automated, efficient, and accurate techniques to handle these spatial big data is essential.

LiDAR: Lidar (Light Detection and Ranging or Laser Imaging Detection and Ranging) data is generated by timing laser pulses from an aerial position (plane or satellite) over a selected area to produce a surface mapping [41].. Lidar data are very rich to analyze surface or extract features. However, these data sets contain irrelevant data for spatial analysis and sometimes miss critical information. These large volumes of data from multiple sources pose a big challenge on management, analysis, and timely accessibility. Particularly, Lidar points and their attributes have tremendous sizes making it difficult to categorize these datasets for end-users. Data integration from multiple spatial sources is another challenge due to the massive amounts of Lidar datasets. Therefore, Spatial Big Data and its management is an essential issue for Lidar remote sensing

Vector data over space is a framework to formalize specific relationships among a set of objects. Traditionally vector data consists of points, lines and polygons; and with the rise of Spatial Big Data, corresponding datasets have arisen from a variety of sources.

VGI Data: Volunteered geographic information (VGI) brings a new notion of infrastructure to collect, synthesize, verify, and redistribute geographic data through geo-location technology, mobile devices, and geo-databases. These geographic data are provided, modified, and shared by everyday citizens using interactive online services (e.g., OpenStreetMap, Wikimapia, GoogleMap, GoogleEarth, Microsofts Virtual Earth, Flickr, etc). Recent explosive growth in user-generated geographic information in the form of tweets, check-ins, Waze, and traffic reports requires bigger storage models to handle large scale spatial datasets.

GPS Trace Data: An example of emerging Spatial Big Data, GPS trajectories, are becoming available for a larger collection of vehicles due to rapid proliferation of cell-phones, in-vehicle navigation devices, and other GPS data-logging devices [21] such as those distributed by insurance companies [62]. Such GPS traces allow indirect estimation of fuel efficiency and GHG emissions via estimation of vehicle-speed, idling, and congestion. They also make it possible to provide personalized route suggestions to users to reduce fuel consumption and GHG emissions. For example, Figure 5 shows 3 months of GPS trace data from a commuter with each point representing a GPS record taken at 1 minute intervals, 24 hours a day, 7 days a week. As can be seen, 3 alternative commute routes are identified between home and work from this dataset. These routes may be compared for engine idling which are represented by darker (red) circles. Assuming the availability of a model to estimate fuel consumption from speed profiles, one may even rank alternative routes for fuel efficiency. In recent years, makers of consumer GPS products [21, 56] are evaluating the potential of this approach. Again, a key hurdle is the dataset size, which can reach 10^{13} items per year given constant minute-resolution measurements for all 100 million US vehicles.



(a) GPS Trace Data. Color indicates speed.

(b) Routes 1, 2, & 3 [25].

Fig. 5 A commuter's GPS tracks over three months reveal preferred routes. (Best in color)

Finally, **Graph** data, is commonly used to represent transportation networks or road maps for routing queries. While the network structure of the graph may not be changing, the amount of information about the network is rising drastically. New temporally-detailed road maps give minute by minute speed information, along with elevation and engine measurements to allow for more sophisticated querying of road networks.

Spatio-Temporal Engine Measurement Data: Many modern fleet vehicles include rich instrumentation such as GPS receivers, sensors to periodically measure sub-system properties [29,30,35,38,54,55], and auxiliary computing, storage and communication devices to log and transfer accumulated datasets. Engine measurement datasets may be used to study the impact of the environment (e.g., elevation changes, weather), vehicles (e.g., weight, engine size, energy-source), traffic management systems (e.g., traffic light timing policies), and driver behaviors (e.g., gentle acceleration or braking) on fuel savings and GHG emissions. These datasets may include a time-series of attributes such as vehicle location, fuel levels, vehicle speed, odometer values, engine speed in revolutions per minute (RPM), engine load, emissions of greenhouse gases (e.g., CO2 and NOX), etc. Fuel efficiency can be estimated from fuel levels and distance traveled as well as engine idling from engine RPM. These attributes may be compared with geographic contexts such as elevation changes and traffic signal patterns to improve understanding of fuel efficiency and GHG emission. For example, Figure 6 shows heavy truck fuel consumption as a function of elevation from a recent study at Oak Ridge National Laboratory [15]. Notice how fuel consumption changes drastically with elevation slope changes. Fleet owners have studied such datasets to fine-tune routes to reduce unnecessary idling [7,8]. It is tantalizing to explore the potential of such datasets to help consumers gain similar fuel savings and GHG emission reduction. However, these datasets can grow big. For example, measurements of 10 engine variables, once a minute, over the 100 million US vehicles in existence [19,53], may have 10^{14} data-items per year.

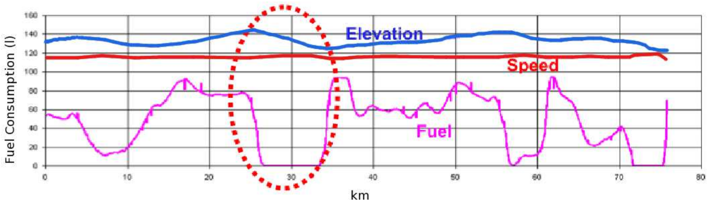
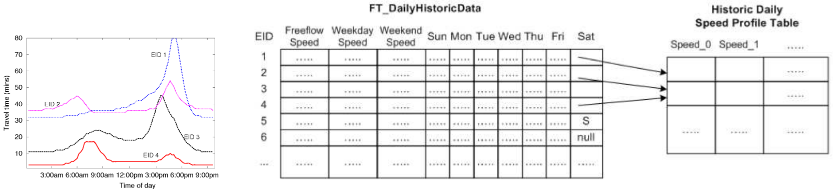


Fig. 6 Engine measurement data improve understanding of fuel consumption [15]. (Best in color)

Historical Speed Profiles: Traditionally, digital road maps consisted of center lines and topologies of the road networks [22,49]. These maps are used by navigation devices and web applications such as Google Maps [25] to suggest routes to users. New datasets from NAVTEQ [40] and other companies use probe vehicles and highway sensors (e.g., loop detectors) to compile travel time information across road segments for all times of the day and week at fine temporal resolutions (seconds or minutes). This data is applied to a profile model, and patterns in the road speeds are identified throughout the day. The profiles have data for every five minutes, which can then be applied to the road segment, building up an accurate picture of speeds based on historical data. Such temporally-detailed (TD) roadmaps contain much more speed information than traditional roadmaps. While traditional roadmaps have only one scalar value of speed for a given road segment (e.g., EID 1), TD roadmaps may potentially list speed/travel time for a road segment (e.g., EID 1) for thousands of time points (Figure 7(a)) in a typical week. This allows a commuter to compare alternate start-times in addition to alternative routes. It may even allow comparison of (start-time, route) combinations to select distinct preferred routes and distinct start-times since route ranking may differ across rush hour and non-rush hour and in general across different start times. However, TD roadmaps are big and their size may exceed 10^{13} items per year for the 100 million road-segments in the US when associated with per-minute values for speed or travel-time. Thus, industry is using speed-profiles, a lossy compression based on the idea of a typical day of a week, as illustrated in Figure 7(b), where each (road-segment, day of the week) pair is associated with a time-series of speed values for each hour of the day.



(a) Travel time along four road segments over a day.

(b) Schema for Daily Historic Speed Data.

Fig. 7 Spatial Big Data on Historical Speed Profiles. (Best in color)

4 Spatial Big Data Analytics

The rise of spatial big data has motivated innovative reasearch in SBD analytics. We summarize recent achievements in this area. We also identify four research areas where opportunity is especially ripe to advance the field.

4.1 Current Accomplishments

Pattern mining and statistical analysis of spatial data are computationally intensive. To improve the scalability of CyberGIS systems on big data, research has focused on parallel and cloud computing solutions for spatial data analytics and mining tasks. For example, recent research has attempted to implement Kriging (i.e., spatial interpolation) under the parallel computing paradigm [26] and over heterogeneous computer architectures and systems by utilizing graphics processing units (GPU) and central processing units (CPU) [52]. Hotspot analysis has also drawn research attention. A recent MapReduce approach to computing $g_i^*(d)$, a statistic for local hotspots, uses an innovative application-level load balancing mechanism [32]. Pang et al. designed a general purpose graphics processing unit (GPGPU)-based approach to compute the likelihood ratio test (LRT), which is a state-of-the-art method for identifying hotspots or anomalies from spatially referenced data [43]. In addition, parallel solutions have been investigated to accelerate existing spatiotemporal pattern discovery algorithms. For example, efforts have been made to scale up the interesting sub-path discovery problem [64] using parallel computing on GPU/GPGPU [45]. Cloud-based spatial analytical tools have also been developed. For example, ESRI released the GIS tool on Hadoop [43] to enable geometry application programming interfaces (APIs) running with MapReduce.

4.2 Areas of Opportunity

There may be many new analytics opportunities provided by SBD. This section describes four examples, i.e., estimating spatial neighbor relationships, supporting place-based ensemble models, simplifying spatial models, and online spatio-temporal data analytics.

4.2.1 Estimating Spatial Neighbor Relationships

The data inputs of spatial data mining (SDM) are complex because they include extended objects such as points, lines, and polygons in vector representation and field data in regular or irregular tessellation such as raster data [11]. During data input, relationships among spatial objects are often implicit (e.g., overlap, intersect, etc.) and are often captured by models or techniques that incorporate spatial information into the SDM process. One such technique is to model the spatial relationship among locations in a spatial framework via a contiguity matrix which may represent a neighborhood relationship defined using adjacency or Euclidean distances. These neighborhood or W matrices are used in many SDM tasks such as spatial outlier detection, co-location pattern discovery, spatial classification and regression modeling, spatial clustering, and spatial hotspot analysis [46].

Table 6 Spatial Auto-Regression and the W-matrix

NAME	MODEL
Classical Linear Regression	$y = x\beta + \epsilon$
Spatial Auto-Regression	$y = \rho W y + x\beta + \epsilon$

The W matrix poses a significant burden to end users due to the fact that W is quadratic in the number of locations, and reliable estimation of W needs a very large number of data samples. In spatial classification and regression modeling, for example, the logistic spatial autoregressive model (SAR) includes the neighborhood relationship contiguity matrix. Table 6 shows a comparison of the classical linear regression model and the spatial auto-regression model where the spatial dependencies of the error term, or the dependent variable, are directly modeled in the regression equation.

SBD Opportunity 1: Post-Markov Assumption. SBD may be large enough to provide a reliable estimate of W. This may ultimately relieve user burden and may improve model accuracy. Traditional assumptions might not have to be made such as limited interaction length (e.g., the Markov assumption), spatially invariant neighbor relationships (e.g., the eight-neighborhood contiguity matrix), and tele-connections derived from short-distance relationships.

4.2.2 Supporting Place-based Ensemble Models

Spatial heterogeneity (or nonstationarity) is an important concept in SDM that is rarely modeled. An important feature of spatial data sets is the variability of observed processes over space. Spatial heterogeneity refers to the inherent variation in measurements of relationships over space. In other words, no two places on Earth are identical. The influence of spatial context on spatial relationships can be seen in the variation of human behavior over space (e.g., differing cultures). Different jurisdictions tend to produce different laws (e.g., speed limit differences between Minnesota and Wisconsin). The term spatial heterogeneity is often used interchangeably with spatial nonstationarity, which is defined as the change in the parameters of a statistical model or change in the ranking of candidate models over space [10].

Traditional astro-physics-based models have been place-independent for the most part with the notable exception of geographically weighted regression (GWR) [13, 20]. The regression equation for GWR, shown by Eq. 1, has the same structure as standard linear regression, with the exception that the parameters are spatially varying, where $\beta(s)$ and $\epsilon(s)$ represent the spatially varying parameters and the errors, respectively. GWR provides an ensemble of linear regression models, one per place of interest.

$$y = X\beta(s) + \epsilon(s) \tag{1}$$

Opportunity 2: SBD may support a place-based ensemble of models beyond GWR. Examples include place-based ensembles of decision trees for land-cover classification and place-based ensembles of spatial auto-regression models. The computational challenge stems from the fact that naive approaches may run a learning algorithm for each place. Reducing the computation cost by exploiting spatial auto-correlation is an interesting possibility that will need to be explored further.

4.2.3 Simplifying Spatial Models

Spatial models are usually computationally more expensive than traditional models. For example, spatial auto-regression requires more computing power due to the fact that W is quadratic in the number of locations (Table 6). Geographically weighted regression has the same limitation as opposed to classical linear regression, also due to the inclusion of the W matrix (Eq. 1). Colocation pattern mining, which finds the subsets of features frequently located together is more computationally expensive that traditional association rule mining [4] and confidence estimation adds more costs (e.g., Markov chain Monte Carlo simulations).

Opportunity 3: The bigger the SBD, the simpler the spatial models. SBD creates an opportunity to simplify spatial models in traditional SDM. It may be the case that some of the complexity from SDM is due to the paucity of data at individual places, which in turn forces one to leverage data at nearby places via spatial autocorrelation and spatial joins. SBD may provide a large volume of data at each place which may allow algorithmic techniques such as place-based divide and conquer. Consequently, it may only be necessary to build one model per place using local data and simpler models. There are, however, a few challenges that must be considered when comparing place-based ensembles of simpler models with current spatial models. For one, it is unclear when bigger data leads to simpler models. In addition, the definition of SBD from an analytics perspective is also unclear (e.g., ratio of samples to number of parameters).

4.2.4 On-line Spatio-Temporal Data Analytics

A fundamental limitation of SDM is off-line batch processing where spatial models are usually not learned in real time (e.g., spatial auto-regression, colocation pattern mining, and hotspot detection). However, SBD includes streaming data such as event reports and sensor measurements. Furthermore, the use cases for SBD include monitoring and surveillance which requires on-line algorithms. Examples of such applications include 1) the timely detection of outbreak of disease, crime, unrest and adverse events, 2) the displacement or spread of a hotspot to neighboring geographies, and 3) abrupt or rapid change detection in land cover, forest-fire, etc. for quick response.

Opportunity 4: On-line Spatio-Temporal Data Analytics. Models that are purely local may leverage time-series data analytics models but regional and global models are more challenging. For spatial interactions (e.g., colocations and tele-connections) with time-lags, SBD may provide opportunities for precisely computing them in an on-line manner. If precise on-line computation is not possible, SBD might be useful in providing on-line approximations.

5 Spatial Big Data Infrastructure

Innovative research in SBD infrastructure has been motivated by the rise of spatial big data. We summarize recent infrastructure accomplishments and identify three areas of opportunity.

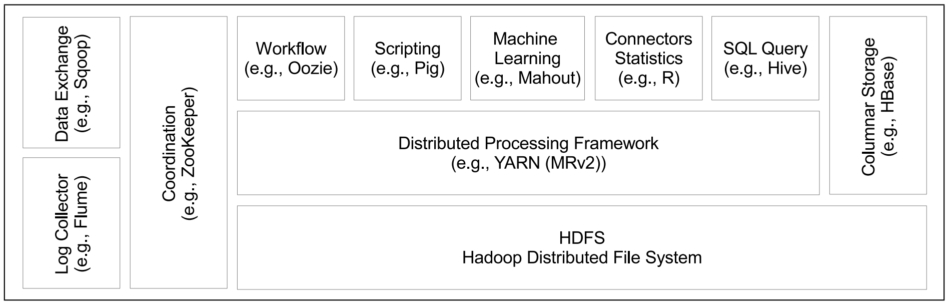

Data Exchange (e.g., Sqoop)
Log Collector (e.g., Flume)
Coordination (e.g., ZooKeeper)
Workflow (e.g., Oozie)
Scripting (e.g., Pig)
Machine Learning (e.g., Mahout)
Connectors Statistics (e.g., R)
SQL Query (e.g., Hive)
Columnar Storage (e.g., HBase)
Distributed Processing Framework (e.g., YARN (MRv2))
HDFS
Hadoop Distributed File System

Fig. 8 Intel Distribution for Apache Hadoop software components [28]

5.1 Current Accomplishments

Research on alternatives to MapReduce are being explored to address some of the emerging challenges that spatial big data raises (e.g., the need to iterate multiple times). Initial efforts in this vein include Pregel [36], Spark [3], GraphLab [34], PowerGraph [24], HaLoop [14], PrIter [63], and CIEL [39], which focus on large-scale, fault-tolerant graph or iterative computing. Research on providing spatial indexes (e.g., R-trees, distributed partitioned R-trees) is also underway. Spatial indexes help to improve the I/O cost of queries retrieving a small part of the data file. Representative efforts include 1) SpatialHadoop [5], which is a MapReduce extension to Apache Hadoop designed specially to work with spatial data by providing specialized spatial data types, spatial indexes, and spatial operations and 2) Hadoop GIS, a high performance spatial data warehousing system over MapReduce [6]. Research on parallel R-tree construction on a GPU is also ongoing [44].

Spatial and CyberGIS intiatives exist at various layers of Hadoop. Figure 8 shows the Intel distribution for Apache Hadoop software components [28]. The figure shows many components running on top of the HDFS for distributed processing (MapReduce), workflow (Oozie), scripting (Pig), machine learning (Mahout), sql queries (Hive), and column store (HBase). At the Hadooop Distributed File System (HDFS) level, SpatialHadoop [5] and Hadoop GIS [6] have added spatial indexes. At the scripting layer (e.g., Pig), SpatialHadoop has added OGIS data types and operators. GIS on Hadoop [43] has also added OGIS data types and operators at the SQL query level (e.g., Hive).

5.2 Areas of Opportunity

There may be many new infrastruture opportunities provided by SBD. We provide three examples in increasing order of difficulty, as shown in Table 7.

Table 7 Example infrastructure opportunities and their relative difficulty

Infrastructure Opportunity	Difficulty
Compute spatial auto-correlation of check-ins	Simple
Parallelize range queries on polygonal maps	Harder
Parallelize spatial auto-regression (SAR) and spatial graph algorithms	Hardest

5.2.1 Compute Spatial Auto-correlation of Check-Ins

Computing spatial auto-correlation of check-ins, multiscale-multigranular images, etc. represents the lowest hanging fruit in terms of infrastructure opportunity. These tasks are relatively easier to parallelize due to the lack of dependency across iterations. Hence, a quality measure for each unit of work may be calculated relatively independently across mapper tasks. Future research should identify a set of spatial problems that are easy to parallelize on MapReduce platforms. Additionally, investigating whether the query-compiler can generate MapReduce parallel code may prove to be an interesting research direction.

5.2.2 Parallelize range queries on polygonal maps

Parallelizing range queries on polygonal maps is a relatively harder task due to the non-uniform distribution of data over space and the wider-ranging sizes of polygons. These result in load imbalance across mapper tasks in simple implementations of the MapReduce platform. A core task in parallelizing the GIS-range-query problem is declustering the spatial data. The goal of declustering is to partition the data so that each partition imposes approximately the same load for any range query. Intuitively, the polygons close to each other should be scattered among different mapper tasks such that for each range-query, every mapper task has an equal amount of work. Polygonal maps can be declustered at the polygonal or sub-polygonal level. Optimal declustering of extended spatial data like polygons is difficult to achieve due to the non-uniform distribution and variable sizes of polygons. In addition, the load imposed by a polygon for each range query depends on the size and location of the query. Since the location of the query is not known *apriori*, it is hard to develop a declustering strategy that is optimal for all range queries.

MPI or OpenMP may provide alternate approaches to parallelizing range queries on polygonal maps. Because their memory-to-memory communication contrasts the disk-based communication of MapReduce, additional work is needed to assess the tradeoff between MPI or OpenMP-based approaches and MapReduce.

5.2.3 Parallelize spatial auto-regression and spatial graph algorithms

Perhaps parallelizing spatial auto-regression (SAR) and other iterative spatial problems such as spatial graph algorithms (e.g., breadth-first search and shortest path) is even more challenging. The reason is that these algorithms use previous information for the next iteration. Although processing one iteration is parallelizable, the synchronization overhead across iterations for cloud environments is too enormous to maintain speedups. Approaches such as Spark [3], which has a cheaper "Reduce" step, will need to be evaluated with iterative GIS workloads. Future work also should include non-iterative algorithms or different parallel programming models. Finally, recent small diameter social graph processing tools, e.g., Pregel [36], need to be evaluated for larger diameter spatial networks, e.g., roadmaps.

6 Conclusion

Recent years have seen the emergence of many new and valuable spatial datasets such as trajectories of cell-phones and GPS devices, vehicle engine measurements, global climate models (GCM) simulation data, volunteered geographic information (VGI), geo-social media, tweets, etc.

However, these emerging and valuable location-aware datasets, which we refer to as Spatial Big Data (SBD), are of a volume, variety, and velocity that exceed the capability of CyberGIS technologies.

This chapter defined spatial big data in terms of value proposition (use-case) and user experience, which depends on the computational platform, use-case, and dataset at hand. User experience may be unsatisfactory due to computational reasons that often stem from workloads exceeding the capacity of the platform. For example, users may experience unacceptable response times, which may be caused by high data volume during correlation or optimization. Users may also experience frequent data loss due to high data velocity relative to the data ingest capacity of the computational platform or they may find themselves expending large amounts of effort to pre-process or post-process SBD due to its high variety.

This chapter also provided an overview of the current efforts, challenges, and opportunities available when spatial big data is enabled via next-generation CyberGIS. From an analytics perspective, we expounded on current accomplishments, e.g., GIS on Hadoop, and four novel opportunities that SBD provides, i.e., estimating spatial neighbor relationships, supporting place-based ensemble models, simplifying spatial models, and on-line spatio-temporal data analytics. From an infrastructure perspective, we discussed current accomplishments (e.g., Spatial Hadoop), and three new opportunities, i.e., computing spatial auto-correlation of check-ins, parallelizing range queries on polygonal maps, and parallelizing spatial auto-regression (SAR) and spatial graph algorithms.

We believe that leveraging spatial big data via CyberGIS will enable a number of transformative societal applications. Next-generation routing services and the leveraging of geo-social media to track disease outbreaks are just the beginning.

References

1. `https://www.facebook.com/about/location`. Facebook Check-in.
2. `http://www.millennium-project.org/millennium/challenges.html`. The Millennium Project, Global Challenges for Humanity.
3. `http://spark.incubator.apache.org/`. Apache Spark.
4. R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499, 1994.
5. Mohamed Mokbel Ahmed Eldawy. Spatial Hadoop, October 9, 2013.
6. Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel Saltz. Hadoop gis: a high performance spatial data warehousing system over mapreduce. *Proceedings of the VLDB Endowment*, 6(11):1009–1020, 2013.
7. American Transportation Research Institute (ATRI). Atri and fhwa release bottleneck analysis of 100 freight significant highway locations. `http://goo.gl/C0NuD`, 2010.
8. American Transportation Research Institute (ATRI). Fpm congestion monitoring at 250 freight significant highway location: Final results of the 2010 performance assessment. `http://goo.gl/3cAjr`, 2010.
9. Austin Brown. Transportation Energy Futures: Addressing Key Gaps and Providing Tools for Decision Makers. Technical report, National Renewable Energy Laboratory, 2011.
10. T.C. Bailey and A.C. Gatrell. *Interactive spatial data analysis*, volume 413. Longman Scientific & Technical Essex, 1995.
11. P. Bolstad. *GIS Fundamentals: A first text on geographic information systems*. Eider Pr, 2005.
12. D. Borthakur. The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11:21, 2007.
13. C. Brunsdon, A.S. Fotheringham, and M.E. Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4):281–298, 1996.
14. Yingyi Bu, Bill Howe, Magdalena Balazinska, and Michael D Ernst. Haloop: Efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, 3(1-2):285–296, 2010.
15. G. Capps, O. Franzese, B. Knee, MB Lascurain, and P. Otaduy. Class-8 heavy truck duty cycle project final report. *ORNL/TM-2008/122*, 2008.
16. Barbara Chapman, Gabriele Jost, and Ruud Van Der Pas. *Using OpenMP: portable shared memory parallel programming*, volume 10. The MIT Press, 2008.
17. Davis, S.C. and Diegel, S.W. and Boundy, R.G. Transportation energy data book: Edition 28. Technical report, Oak Ridge National Laboratory, 2010.
18. J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.
19. Federal Highway Administration. Highway Statistics. *HM-63, HM-64*, 2008.
20. A.S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons Inc, 2002.

21. Garmin. http://www.garmin.com/us/.
22. Betsy George and Shashi Shekhar. Road maps, digital. In *Encyclopedia of GIS*, pages 967–972. Springer, 2008.
23. Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.
24. Joseph E Gonzalez, Yucheng Low, Haijie Gu, Danny Bickson, and Carlos Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 17–30, 2012.
25. Google Maps. http://maps.google.com.
26. Qingfeng Guan, Phaedon C Kyriakidis, and Michael F Goodchild. A parallel computing approach to fast geostatistical areal interpolation. *International Journal of Geographical Information Science*, 25(8):1241–1267, 2011.
27. InformationWeek. Red Cross Unveils Social Media Monitoring Operation. http://www.informationweek.com/government/information-management/red-cross-unveils-social-media-monitorin/232602219, 2012.
28. Intel. Intel Distribution for Apache Hadoop Software. http://hadoop.intel.com/pdfs/IntelDistributionProductBrief.pdf, October 9, 2013.
29. H. Kargupta, J. Gama, and W. Fan. The next generation of transportation systems, greenhouse emissions, and data mining. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1209–1212. ACM, 2010.
30. H. Kargupta, V. Puttagunta, M. Klein, and K. Sarkar. On-board vehicle data stream monitoring using minefleet and fast resource constrained monitoring of correlation matrices. *New Generation Computing*, 25(1):5–32, 2006. Springer.
31. G. Levchuk, A. Bobick, and E. Jones. Activity and function recognition for moving and static objects in urban environments from wide-area persistent surveillance inputs. In *Proceedings of SPIE*, volume 7704, page 77040P, 2010.
32. Yan Liu, Kaichao Wu, Shaowen Wang, Yanli Zhao, and Qian Huang. A mapreduce approach to g i*(d) spatial statistic. In *Proceedings of the ACM SIGSPATIAL International Workshop on High Performance and Distributed Geographic Information Systems*, pages 11–18. ACM, 2010.
33. Joel Lovell. Left-hand-turn elimination. http://goo.gl/3bkPb, December 9, 2007. New York Times.
34. Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Carlos Guestrin, and Joseph M Hellerstein. Graphlab: A new framework for parallel machine learning. *arXiv preprint arXiv:1006.4990*, 2010.
35. Lynx GIS. http://www.lynxgis.com/.
36. G. Malewicz, M.H. Austern, A.J.C. Bik, J.C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 international conference on Management of data*, pages 135–146. ACM, 2010.
37. J. Manyika et al. Big data: The next frontier for innovation, competition and productivity. *McKinsey Global Institute, May*, 2011.
38. MasterNaut. Green Solutions. http://www.masternaut.co.uk/carbon-calculator/.
39. Derek G Murray, Malte Schwarzkopf, Christopher Smowton, Steven Smith, Anil Madhavapeddy, and Steven Hand. Ciel: a universal execution engine for distributed dataflow computing. In *Proceedings of the 8th USENIX conference on Networked systems design and implementation*, page 9, 2011.
40. NAVTEQ. www.navteq.com.
41. New York Times. Mapping Ancient Civilization, in a Matter of Days. http://www.nytimes.com/2010/05/11/science/11maya.html, 2010.
42. New York Times. Military Is Awash in Data From Drones. http://www.nytimes.com/2010/01/11/business/11drone.html?pagewanted=all, 2010.

43. Linsey Xiaolin Pang, Sanjay Chawla, Bernhard Scholz, and Georgina Wilcox. A scalable approach for lrt computation in gpgpu environments. In *Web Technologies and Applications*, pages 595–608. Springer, 2013.
44. Sushil K. Prasad, Shashi Shekhar, Xi He, Satish Puri, Michael McDermott, Xun Zhou, and Michael Evans. Gpgpu-based data structures and algorithms for geospatial computation a summary of results and future roadmap. position paper. *Proceedings of rhe All Hands Meeting of the NSF CyberGIS project. Seattle*, 2013.
45. Sushil K Prasad, Shashi Shekhar, Michael McDermott, Xun Zhou, Michael Evans, and Satish Puri. Gpgpu-accelerated interesting interval discovery and other computations on geospatial datasets–a summary of results. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial-2013)*. ACM, 2013.
46. S. Shekhar, M.R. Evans, J.M. Kang, and P. Mohan. Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):193–214, 2011.
47. S. Shekhar, S. Ravada, D. Chubb, and G. Turner. Declustering and load-balancing methods for parallelizing geographic information systems. *Knowledge and Data Engineering, IEEE Transactions on*, 10(4):632–655, 1998.
48. S. Shekhar, S. Ravada, V. Kumar, D. Chubb, and G. Turner. Parallelizing a gis on a shared address space architecture. *Computer*, 29(12):42–48, 1996.
49. S. Shekhar and H. Xiong. *Encyclopedia of GIS*. Springer Publishing Company, Incorporated, 2007.
50. Shashi Shekhar, Michael R Evans, Viswanath Gunturi, KwangSoo Yang, and Daniel Cintra Cugler. Benchmarking spatial big data. In *Specifying Big Data Benchmarks*, pages 81–93. Springer, 2014.
51. Shashi Shekhar, Viswanath Gunturi, Michael R Evans, and KwangSoo Yang. Spatial big-data challenges intersecting mobility and cloud computing. In *Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pages 1–6. ACM, 2012.
52. Xuan Shi and Fei Ye. Kriging interpolation over heterogeneous computer architectures and systems. *GIScience & Remote Sensing*, (ahead-of-print):1–16, 2013.
53. Daniel. Sperling and D. Gordon. *Two billion cars*. Oxford University Press, 2009.
54. TeleNav. `http://www.telenav.com/`.
55. TeloGIS. `http://www.telogis.com/`.
56. TomTom. TomTom GPS Navigation. `http://www.tomtom.com/`, 2011.
57. US Congress. Energy independence and security act of 2007. *Public Law*, (110-140), 2007. `http://goo.gl/6Kspz`.
58. U.S. Energy Information Adminstration. Monthly Energy Review June 2011. `http://www.eia.gov/totalenergy/data/monthly/`.
59. Ranga Raju Vatsavai, Auroop Ganguly, Varun Chandola, Anthony Stefanidis, Scott Klasky, and Shashi Shekhar. Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–10. ACM, 2012.
60. Shaowen Wang. A cybergis framework for the synthesis of cyberinfrastructure, gis, and spatial analysis. *Annals of the Association of American Geographers*, 100(3):535–557, 2010.
61. Waze Mobile. `http://www.waze.com/`.
62. Wikipedia. Usage-based insurance — wikipedia, the free encyclopedia. `http://goo.gl/NqJE5`, 2011. [Online; accessed 15-December-2011].
63. Yanfeng Zhang, Qixin Gao, Lixin Gao, and Cuirong Wang. Priter: a distributed framework for prioritized iterative computations. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, page 13. ACM, 2011.
64. Xun Zhou, Shashi Shekhar, Pradeep Mohan, Stefan Liess, and Peter K Snyder. Discovering interesting sub-paths in spatiotemporal datasets: A summary of results. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 44–53. ACM, 2011.