Data-Driven Geography

Harvey J. Miller
Department of Geography
The Ohio State University
miller.81@osu.edu

Michael F. Goodchild
Department of Geography
University of California, Santa Barbara
good@geog.ucsb.edu

Abstract:

The context for geographic research has shifted from a data-scarce to a data-rich environment, in which the most fundamental changes are not just the volume of data, but the variety and the velocity at which we can capture georeferenced data; trends often associated with the concept of Big Data. A data-driven geography may be emerging in response to the wealth of georeferenced data flowing from sensors and people in the environment. Although this may seem revolutionary, in fact it may be better described as evolutionary. Some of the issues raised by data-driven geography have in fact been longstanding issues in geographic research, namely, large data volumes, dealing with populations and messy data, and tensions between idiographic versus nomothetic knowledge. The belief that spatial context matters is a major theme in geographic thought and a major motivation behind approaches such as time geography, disaggregate spatial statistics and GIScience. There is potential to use Big Data to inform both geographic knowledge discovery and spatial modeling. However, there are challenges, such as how to formalize geographic knowledge to clean data and to ignore spurious patterns, and how to build data-driven models that are both true and understandable.

1. Introduction

A great deal of attention is being paid to the potential impact of data-driven methods on the sciences. The ease of collecting, storing, and processing digital data may be leading to what some are calling the fourth paradigm of science, following the millennia-old traditional of empirical science describing natural phenomena, the centuries-old tradition of theoretical science using models and generalization, and the decades-old traditional of computational science simulating complex systems. Instead of looking through telescopes and microscopes, researchers are increasingly interrogating the world through large-scale, complex instruments and systems that relay observations to large databases to be processed and stored as information and knowledge in computers (Hey, Tansley, and Tolle 2009).

This fundamental change in the nature of the data available to researchers is leading to what some call *Big Data*. Big Data refer to data that outstrip our capabilities to analyze. This has three dimensions, the so-called "three Vs": i) *volume* – the amount of data that can be collected and stored; ii) *velocity* – the speed at which data can be captured; and iii) *variety* – encompassing both structured (organized and stored in tables and relations) and unstructured (text, imagery) data (Dumbill 2012). Some of these data are generated from massive simulations of complex systems such as cities (e.g., TRANSIMs; see Cetin 2002), but a large portion of the flood is from sensors and software that digitize and store a broad spectrum of social, economic, political, and environmental patterns and processes (Graham and Shelton 2013; Kitchin 2014). Sources of geographically (and often temporally) referenced data include location-aware technologies such as the Global Positioning System and mobile phones; in-situ sensors carried by individuals in phones, attached to vehicles, and embedded in infrastructure; remote sensors carried by airborne and satellite platforms; radiofrequency identification (RFID) tags attached to objects; and georeferenced social media (Miller 2007, 2010; Sui and Goodchild 2011; Townsend 2013).

Yet despite the enthusiasm over Big Data and data-driven methods, the role it can play in scholarly research, and specifically research in geography may not be immediately

apparent. Are theory and explanation archaic when we can measure and describe so much, so quickly? Does data velocity really matter in research, with its traditions of careful reflection? Can the obvious problems associated with variety – lack of quality control, lack of rigorous sampling design – be overcome? Can we make valid generalizations from ongoing, serendipitous (instead of carefully designed and instrumented) data collection? In short, can Big Data and data-driven methods lead to significant discoveries in geographic research? Or will the research community continue to rely on what for the purposes of this paper we will term Scarce Data: the products of public-sector statistical programs that have long provided the major input to research in quantitative human geography?

Our purpose in this paper is to explore the implications of these tensions – theory-driven versus data-driven research, prediction versus discovery, law-seeking versus description-seeking – for research in geography. We anticipate that geography will provide a distinct context for several reasons: the specific issues associated with location, the integration of the social and the environmental, and the existence within the discipline of traditions with very different approaches to research. Moreover, although data-driven geography may seem revolutionary, in fact it may be better described as evolutionary since its challenges have long been themes in the history of geographic thought and the development of geographical techniques.

The next section of this paper discusses the concepts of Big Data and data-driven geography, addressing the question of what is special about the new flood of georeferenced data. Section 3 of this paper discusses major challenges facing data-driven geography; these include dealing with populations (not samples), messy (not clean) data, and correlations (not causality). The fourth section discusses the role of theory in data-driven geography. Section 5 identifies ways to incorporate Big Data into geographic research. Section 6 concludes this paper with a summary and some cautions on the broader impacts of data-driven geography on society.

2. Big Data and Data-Driven Geography

Humanity's current ability to acquire, process, share, and analyze huge quantities of data is without precedent in human history. It has led to the coining of such terms as the "exaflood" and the metaphor of "drinking from a firehose" (Sui, Goodchild and Elwood 2013; Waldrop 1990). It is also led to the suggestion that we are entering a new, fourth phase of science that will be driven not so much by careful observation by individuals, or theory development, or computational simulation, as by this new abundance of digital data (Hey, Tansley, and Tolle 2009).

It is worth recognizing immediately, however, that the firehose metaphor has a comparatively long history in geography, and that the discipline is by no means new to an abundance of voluminous data. The Landsat program of satellite-based remote sensing began in the early 1970s by acquiring data at rates that were well in excess of the analytic capacities of the computational systems of the time; subsequent improvements in sensor resolution and the proliferation of military and civilian satellites have meant that four decades later data volumes continue to challenge even the most powerful computational systems.

Volume is clearly not the only characteristic that distinguishes today's data supply from that of previous eras. Today, data are being collected from many sources, including social media, crowdsourcing, ground-based sensor networks, and surveillance cameras, and our ability to integrate such data and draw inferences has expanded along with the volume of the supply. The phrase Big Data implies a world in which predictions are made by mining data for patterns and correlations among these new sources, and some very compelling instances of surprisingly accurate predictions have surfaced in the past few years with respect to the results of the Eurovision song contest (O'Leary 2012), the stock market (Preis, Moat, and Stanley 2013), and the flu (Butler 2008). The theme of Big Data is often associated not only with volume but with variety, reflecting these multiple sources, and velocity, given the speed with which such data can now be analyzed to make predictions in close-to-real time.

Ubiquitous, ongoing data flows are a big deal because they allow us to capture spatio-temporal dynamics directly (rather than inferring them from snapshots) and at multiple scales. The data are collected on an ongoing basis, meaning that both mundane and unplanned events can be captured. To borrow Nassim Taleb's metaphor for probable and inconsequential versus improbable but consequential events (Taleb 2007): we do not need to sort the white swans from the black swans before collecting data: we can measure all swans and then figure out later which are white or black. White swans may also combine in surprising ways to form black-swan events.

Big Data is leading to new approaches to research methodology. Fotheringham (1998) defines geocomputation as quantitative spatial analysis where the computer plays a pivotal role. The use of the computer drives the form of the analysis rather than just being a convenient vehicle: analysts design geocomputational techniques with the computer in mind. Similarly, data play a pivotal role in data-driven methods. From this perspective data are not just a convenient way to calibrate, validate, and test but rather the driving force behind the analysis. Consequently, analysts design data-driven techniques with data in mind - and not just large volumes of data, but a wider spectrum of data flowing at higher speeds from the world. In this sense we may indeed be entering a fourth scientific paradigm where scientific methods are configured to satisfy data rather than data configured to satisfy methods.

3. Data-Driven Geography: Challenges

In *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Mayer-Schonberger and Cukier (2013) identify three main challenges of Big Data in science: i) populations, not samples; ii) messy, not clean data, and; iii) correlations, not causality. We discuss these three challenges for geographic research in the following subsections.

3.1. Populations, not samples

Back when analysis was largely performed by hand rather than by machines, dealing with large volumes of data was impractical. Instead, researchers developed methods for

collecting representative samples and for generalizing from them to inferences about the population from which they were drawn. Random sampling was thus a strategy for dealing with information overload in an earlier era. In statistical programs such as the U.S. Census of Population it was also a means for controlling costs.

Random sampling works well, but it is fragile: it works only as long as the sampling is representative. A sampling rate of one in six (the rate previously used by the US Bureau of the Census for its more elaborate Long Form) may be adequate for some purposes, but becomes increasingly problematic when analysis focuses on comparatively rare subcategories. Random sampling also requires a process for enumerating and selecting from the population (a sampling frame), which is problematic if enumeration is incomplete. Sample data also has a lack of extensibility for secondary uses. Because randomness is so critical, one must carefully plan for sampling, and it may be difficult to re-analyze the data for purposes other than those for which it was collected (Mayer-Schonberger and Cukier 2013).

In contrast, many of the new data sources consist of populations, not samples: the ease of collecting, storing, and processing digital data means that instead of dealing with a small representation of the population we can work with the entire population and thus escape one of the constraints of the past. But one problem with populations is that they are often self-selected rather than sampled: for example, all people who signed up for Facebook, all people who carry smartphones, or all cars than happened to travel within the City of London between 8am-11:00am on 2 September 2013. Geolocated tweets are an attractive source of information on current trends (e.g., Tsou et al. 2013), but only a small fraction of tweets are accurately geolocated using GPS. Since we do not know the demographic characteristics of any of these groups, it is impossible to generalize from them to any larger populations from which they might have been drawn.

Yet geographers have long had to contend with the issues associated with samples and their parent populations. Consider, for example, an analysis of the relationship between people over 65 years old and people registered as Republicans, the case studied by

Openshaw and Taylor in their seminal article on the modifiable areal unit problem (Openshaw and Taylor 1979). The 99 counties of Iowa (their source of data) are all of the counties that exist in Iowa. They are not therefore a random sample of Iowa counties, or even a representative sample of counties of the US, so the methods of inferential statistics that assume random and independent sampling are not applicable. In remote sensing it is common to analyze all of the pixels in a given scene; again, these are not a random sample of any larger population.

However, the cases discussed above are where we can be assured that the entire population of interest is included: we are interested in all of the land cover in a scene, or all of the people over 65 and Republicans in Iowa. This is often not true with many new sources of data. A challenge is how to identify the niches to which monitored population data can be applied with reasonable generality. This inverts the classic sampling problem where we identify a question and collect data to answer that question. Instead, we collect the data and determine what questions we can answer.

Another issue concerns what people are volunteering when they volunteer geographic and other information (Goodchild 2007). Social media such as Facebook may have high penetration rates with respect to population, but do not necessarily have high penetration rates into peoples' lives. Checking in at an orchestra concert or lecture provides a noble image that a person would like to promote, while checking in at a bar at 10am is an image that a person may be less keen to share. In the classic sociology text *The Presentation of Self in Everyday Life*, Erving Goffman uses theater as a metaphor and distinguishes between stage and backstage behaviors, with stage behaviors being consistent with the role people wish to play in public life and backstage behaviors being private actions that people wish to keep private (Goffman 1959). While there are certainly cases of over-sharing behavior (especially among celebrities) we cannot be assured that the information people volunteer is an accurate depiction of their complete lives or just of the lives they wish to present to the social sphere. Several geographic questions follow from these observations. What is the geography of stage versus backstage realms in a city or region?

Does this distribution vary by age, gender, socioeconomic status, or culture? What do these imply for what we can know about human spatial behavior?

In addition to selective volunteering of information about their lives, there also may be selection biases in the information people volunteer about environments. OpenStreetMap (OSM) is often identified as a successful crowdsourced mapping project: many cities of the world have been mapped by people on a voluntary basis to a remarkable degree of accuracy. However, some regions get mapped quicker than others, such as tourist locations, recreation areas, and affluent neighborhoods, while locations of less interest to those who participate in OSM (such as poorer neighborhoods) receive less attention (Haklay 2010). While biases exist in official, administrative maps (e.g., governments in developing nations often do not map informal settlements such as favelas), the biases in crowdsourced maps are likely to be more subtle. Similarly, the rise of civic hacking where citizens generate data, maps, and tools to solve social problems tends to focus on the problems that citizens with laptops, fast internet connections, technical skills, and available time consider to be important (Townsend 2013).

3.2 Messy, not clean

The new data sources are often messy, consisting of data that are unstructured, collected with no quality control, and frequently accompanied by no documentation or metadata. There are at least two ways of dealing with such messiness. On the one hand, we can restrict our use of the data to tasks that do not attempt to generalize or to make assumptions about quality. Messy data can be useful in what one might term the softer areas of science: initial exploration of study areas, or the generation of hypotheses. Ethnography, qualitative research, and investigations of Grounded Theory (Glaser and Strauss 1967) often focus on using interviews, text, and other sources to reveal what was otherwise not known or recognized, and in such contexts the kinds of rigorous sampling and documentation associated with Scarce Data are largely unnecessary. We discuss this option in greater detail later in the paper.

On the other hand, we can attempt to clean and verify the data, removing as much as possible of the messiness, for use in traditional scientific knowledge construction. Goodchild and Li (2012) discuss this approach in the context of crowdsourced geographic information. They note that traditional production of geographic information has relied on multiple sources, and on the expertise of cartographers and domain scientists to assemble an integrated picture of the landscape. For example, terrain information may be compiled from photogrammetry, point measurements of elevation, and historic sources; as a result of this process of synthesis the published result may well be more accurate than any of the original sources.

Goodchild and Li (2012) argue that that traditional process of synthesis, which is largely hidden from popular view and not apparent in the final result, will become explicit and of critical importance in the new world of Big Data. They identify three strategies for cleaning and verifying messy data: i) the crowd solution; ii) the social solution; and iii) the knowledge solution. The *crowd solution* is based on Linus' Law, named in honor of the developer of Linux, Linus Torvalds: "Given enough eyeballs, all bugs are shallow" (Raymond 2001). In other words, the more people who can access and review your code, the greater the accuracy of the final product. Geographic facts that can be synthesized from multiple original reports are likely to be more accurate than single reports. This is of course the strategy used by Wikipedia and its analogs: open contributions and open editing are evidently capable of producing reasonably accurate results when assisted by various automated editing procedures.

In the geographic case, however, several issues arise that limit the success of the crowd solution. Reports of events at some location may be difficult to compare if the means used to specify location (place names, street address, GPS) are uncertain, and if the means used to describe the event is ambiguous. Geographic facts may be obscure, such as the names of mountains in remote parts of the world, and the crowd may therefore have little interest or ability to edit errors.

Goodchild and Li (2012) describe the *social solution* as implementing a hierarchical structure of volunteer moderators and gatekeepers. Individuals are nominated to roles in the hierarchy based on their track record of activity and the accuracy of their contributions. Volunteered facts that appear questionable or contestable are referred up the hierarchy, to be accepted, queried, or rejected as appropriate. Schemes such as this have been implemented by many projects, including Open Street Map and Wikipedia. Their major disadvantage is speed: since humans are involved, the solution is best suited to applications where time is not critical.

The third, the *knowledge solution*, asks how one might know if a purported fact is false, or likely to be false. Spelling errors and mistakes of syntax are simple indicators which all of us use to triage malicious email. In the geographic case, one can ask whether a purported fact is consistent with what is already known about the geographic world, in terms both of facts and theories. Moreover such checks of consistency can potentially be automated, allowing triage to occur in close to real time; this approach has been implemented, although on a somewhat unstructured basis, by companies that daily receive thousands of volunteered corrections to their geographic databases.

A purported fact can deviate from established geographic knowledge in either syntax or semantics, or both. Syntax refers to the rules by which the world is constructed, while semantics refers to the meaning of those facts. Syntactical knowledge is often easier to check than semantic knowledge. For example, Figure 1 illustrates an example of syntactical geographic knowledge. We know from engineering specifications that an on-ramp can only intersect a freeway at a small angle (typically 30 degrees or less). If a road-network database appears to have on-ramp intersections of greater than 30 degrees we know that the data are likely to be wrong; in the case of Figure 1, many of the apparent intersections of the light-blue segments are more likely to be overpasses or underpasses. Such errors have been termed errors of *logical consistency* in the literature of geographic information science (e.g., Guptill and Morrison 1995).

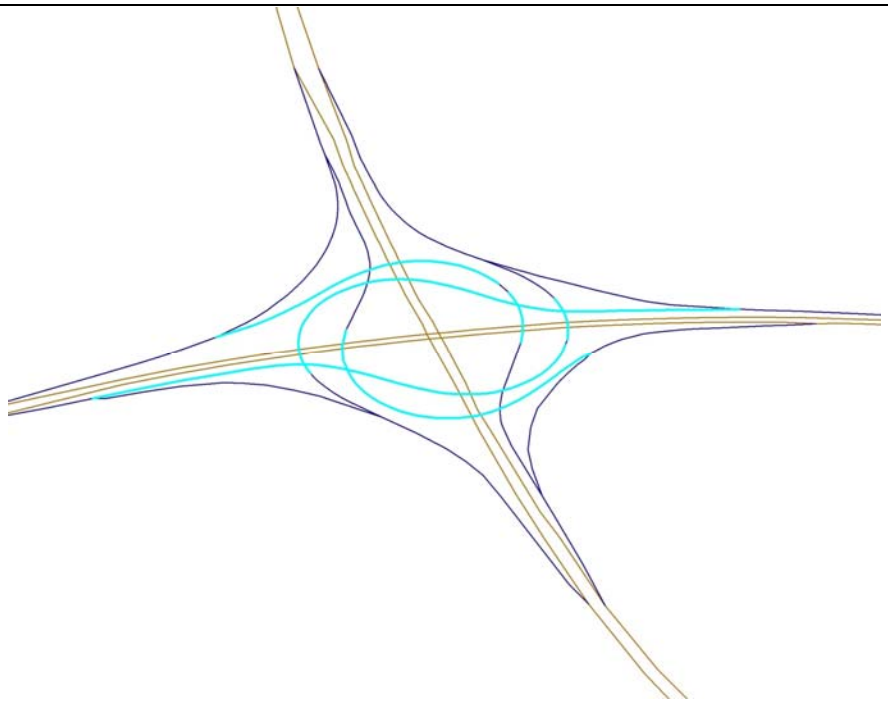

Figure 1. Syntactical geographic knowledge: Highway on-ramp feature geometry.

In contrast, Figure 2 illustrates semantic geographic knowledge: a photograph of a lake that has been linked to the Google Earth map of The Ohio State University campus. However, this photograph seems to be located incorrectly: we recognize the scene as *Mirror Lake*, a campus icon to the southeast of the purported location indicated on the map. The purported location must be wrong, but can we be sure? Perhaps the university moved *Mirror Lake* to make way for a new Geography building? Or perhaps *Mirror Lake* was so popular that the university created a mirror *Mirror Lake* to handle the overflow? We cannot immediately and with complete confidence dismiss this empirical fact without additional investigation since it does not violate any known rules by which the world is constructed: there is nothing preventing *Mirror Lake* from being moved or mirrored. Of course, there are some semantic facts that can be dismissed confidently as absurd – one would not expect to see a lake scene on the top of Mt. Everest or in the Sahara Desert. Nevertheless, there is no firm line between clearly absurd and non-absurd semantic facts – e.g., one would not expect to see Venice or New York City in the Mojave Desert, but Las Vegas certainly exists.

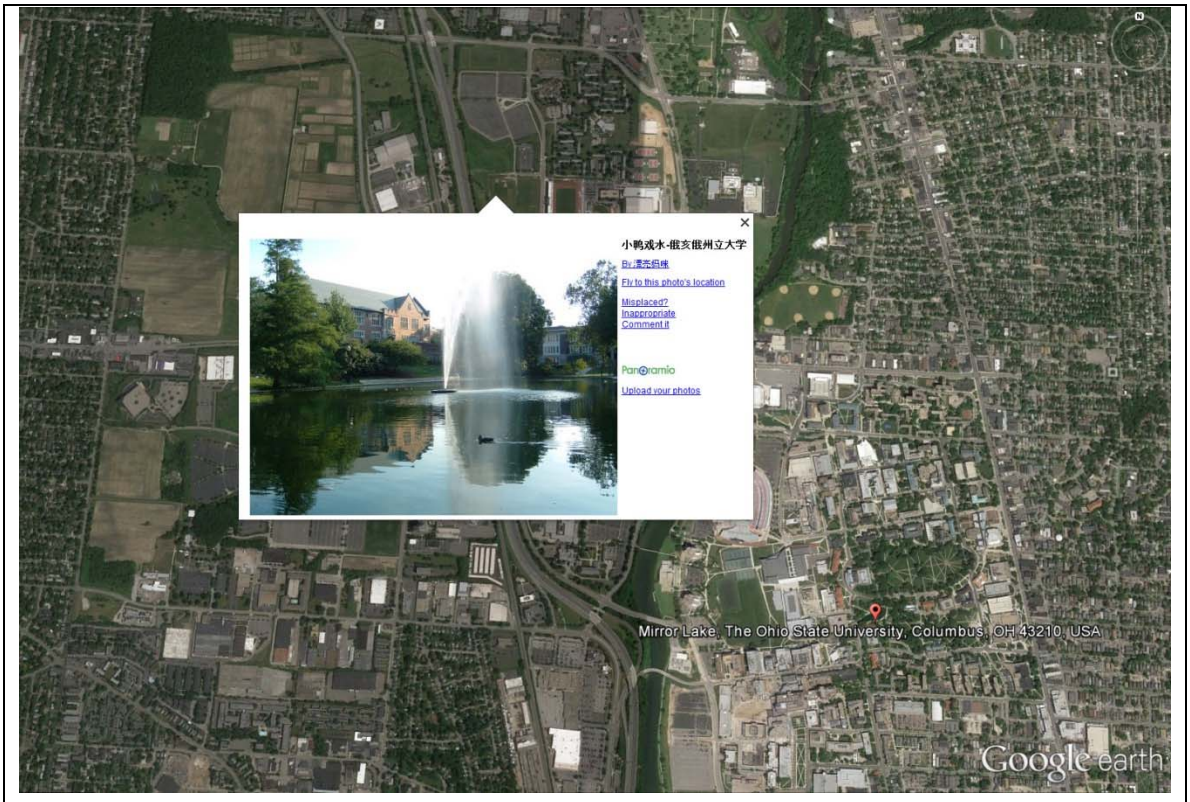

Figure 2: Semantic geographic knowledge: Where is *Mirror Lake*? (Google Earth; last accessed 24 September 2013 10:00am EDT)

A major task for the knowledge solution is formalizing knowledge to support automated triage of asserted facts and automated data fusion. Knowledge can be derived empirically or as predictions from theories, models, and simulations. In the latter case, we may be looking for data at variance with predictions as part of the knowledge discovery and construction processes.

There are at least two major challenges to formalizing geographic knowledge. First, geographic concepts such as neighborhood, region, the Midwest, and developing nations can be vague, fluid, and contested. A second challenge is the development of explicit, formal, and computable representations of geographic knowledge. Much geographic knowledge is buried in formal theories, models, and equations that must be solved or processed, or in informal language that must be interpreted. In contrast, knowledge-

discovery techniques require explicit representations such as rules, hierarchies, and concept networks that can be accessed directly without processing (Miller 2010).

3.3. Correlations, not causality

Traditionally, scholarly research concerns itself with knowing *why* something occurs. Correlations alone are not sufficient, because the existence of correlation does not imply that change in either variable causes change in the other. In the correlation explored by Openshaw and Taylor cited earlier (Openshaw and Taylor 1979), the existence of a correlation between the number of registered Republicans in a county and the number of people aged 65 and over does not imply that either one has a causal effect on the other. Over the years, science has adopted pejorative phrases to describe research that searches for correlations without concern for causality or explanation: "curve-fitting" comes to mind. Nevertheless correlations may be useful for prediction, especially if one is willing to assume that an observed correlation can be generalized beyond the specific circumstances in which it is observed.

But while they may be sufficient, explanation and causality are not necessary conditions for scientific research: much research, especially in such areas as spatial analysis, is concerned with advancing method, whether its eventual use is for explanation or for prediction. The literature of geographic information science is full of tools that have been designed not for finding explanations but for more mundane activities such as detecting patterns, or massaging data for visualization. Such tools are clearly valuable in an era of data-driven science, where questions of "why" may not be as important. In the next section we extend this argument by taking up the broader question of the role of theory in data-driven geography.

4. Theory in data-driven geography

In a widely discussed article published in *Wired* magazine, Anderson called for the end of science as we know it, claiming that the data deluge is making the scientific method obsolete (Anderson 2008). Using physics and biology as examples, he argued that as science has advanced it has become apparent that theories and models are caricatures of a

deeper underlying reality that cannot be easily explained. However, explanation is not required for continuing progress: as Anderson states "Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

Duncan Watts makes a similar argument about theory in the social sciences, stating that unprecedented volumes of social data have the potential to revolutionize our understanding of society, but this understanding will not be in the form of general laws of social science or cause-and-effect social relationships. Although Watts suggests the limitations of theory in the era of data-driven science, he does not call for the end of theory but rather for a more modest type of theory that would include general propositions (such as what interventions work for particular social problems) or how more obvious social facts fit together to generate less obvious outcomes. Watts links this approach to calls by sociologist Robert Merton in the mid-twentieth century for *middle-range theories*: theories that address identifiable social phenomena instead of abstract entities such as the entire social system (Watts 2011). Middle-range theories are empirically grounded: they are based in observations, and serve to derive hypotheses that can be investigated. However, they are not endpoints: rather, they are temporary stepping-stones to general conceptual schemes that can encompass multiple middle-range theories (Merton 1967).

Data-driven science seems to entail a shift away from the general and towards the specific – away from attempts to find universal laws than encompass all places and times and towards deeper descriptions of what is happening at particular places and times. There are clearly some benefits to this change: as Batty (2012) points out, urban science and planning in the era of Scarce Data focused on radical and massive changes to cities over the long term, with little concern for small spaces and local movements. Data-driven urban science and planning can rectify some of the consequent urban ills by allowing greater focus on the local and routine. However, over longer time spans and wider spatial domains the local and routine merges into the long-term; a fundamental scientific challenge is how local and short-term Big Data can inform our understanding of

processes over longer temporal and spatial horizons; in short, the problem of generalization.

Geography has long experience with partnerships – and tensions - between *nomothetic* (law-seeking) and *idiographic* (description-seeking) knowledge (Cresswell 2013). Table 1 provides a summary. The early history of geography in the time of Strabo (64/63 BCE – 24 CE) and Ptolemy (90-168 CE) involved both generalizations about the Earth and intimate descriptions of specific places and regions; these were two sides of the same coin. Bernhardus Varenius (1622 – 1650) conceptualized geography as consisting of general (scientific) and special (regional) knowledge, although he considered the latter to be subsidiary to the former (Warntz 1989, Goodchild et al. 1999). Alexander von Humboldt (1769 – 1859) and Carl Ritter (1779 – 1859), often regarded as the founders of modern geography, tried to derive general laws through careful measurement of geographic phenomena at particular locations and times. In more recent times, the historic balance between nomothetic and idiographic geographic knowledge has become more unstable. The early 20th century witnessed the dominance of nomothetic geography in the guise of the environmental determinism in the early 1900s, followed by a backlash against its abuses and the subsequent rise of idiographic geography in the form of areal differentiation: Richard Hartshorne famously declared in *The Nature of Geography* that the only law in geography is that all areas are unique (Hartshorne 1939). The dominance of idiographic geography and the concurrent crisis in American academic geography (in particular, the closing of Harvard's geography program in 1948; Smith 1992) led to the Quantitative Revolution of the 1950s and 1960s, with geographers such as Fred Schaefer, William Bunge, Peter Haggett, and Edward Ullman asserting that geography should be a law-seeking science that answers the question "why?" rather than building a collection of facts describing what is happening in particular regions. Physical geographers have – perhaps wisely – disengaged themselves from these debates, but the tension between nomothetic and idiographic approaches persists in human geography (see Cresswell 2013; Schuurman 2000; Sui 2004; Sui and DeLyser 2012, 2013).

However, attempts to reconcile nomothetic and idiographic knowledge did not die with Humboldt and Ritter. Approaches such as time geography seek to capture context and history and recognize the roles of both agency and structure in human behavior (Cresswell 2013). In spatial analysis, the trend towards local statistics, exemplified by Geographically Weighted Regression (Fotheringham, Brundson, and Charlton 2002) and Local Indicators of Spatial Association (Anselin 1995), represents a compromise in which the general principles of nomothetic geography are allowed to express themselves differently across geographic space. Goodchild (2004) has characterized GIS as combining the nomothetic, in its software and algorithms, with the idiographic in its databases.

Path to geographic knowledge	**Advocates**
Nomothetic ↔ Idiographic	Strabo Ptolemy
Nomothetic → Idiographic	Varenius
Nomothetic ← Idiographic	Humboldt Ritter
~~Nomothetic~~ Idiographic	Hartshorne
Nomothetic ~~Idiographic~~	Schaefer
Nomothetic ↔ Idiographic	Hägerstrand (Time geography) Fotheringham/Anselin (local spatial statistics) Tomlinson/Goodchild (GIScience)
Table 1: A brief history of partnerships and tensions between nomothetic (law-	

seeking) and idiographic (description-seeking) knowledge in geographic thought.

In a sense, the paths to geographic knowledge engendered by data-intensive approaches such as time geography, disaggregate spatial statistics and GIScience are a return to the early foundation of geography where neither law-seeking nor description-seeking were privileged. Geographic generalizations and laws are possible but space matters: spatial dependency and spatial heterogeneity create local context that shapes physical and human processes as they evolve on the surface of the Earth. Geographers have believed this for a long time, but this belief is also supported by recent breakthroughs in complex systems theory, which suggests that patterns of local interactions lead to emergent behaviors that cannot be understood in isolation at either the local or global levels. Understanding the interactions among agents within an environment is the scientific glue that binds the local with the global (Flake 1998).

In short, data-driven geography is not necessarily a radical break with the geographic tradition: geography has a longstanding belief in the value of idiographic knowledge by itself as well as its role in constructing nomothetic knowledge. Although this belief has been tenuous and contested at times, data-driven geography may provide the paths between idiographic and nomothetic knowledge that geographers have been seeking for two millennia. However, while complexity theory supports this belief, it also suggests that this knowledge may have inherent limitations: emergent behavior is by definition surprising.

5. Approaches to Data-Driven Geography

If we accept the premise – at least until proven otherwise – that Big Data and data-driven science harmonize with longstanding themes and beliefs in geography, the question that follows is: How can data-driven approaches fit into geographic research? Data-driven approaches can support both geographic knowledge discovery and spatial modeling. However, there are some challenges and cautions that must be recognized.

5.1. Data-driven geographic knowledge discovery

Geographic knowledge discovery refers to the initial stage of the scientific process where the investigator forms his or her conceptual view of the system, develops hypotheses to be tested, and performs groundwork to support the knowledge-construction process. Geographic data facilitates this crucial phase of the scientific process by supporting activities such as study-site selection and reconnaissance, ethnography, experimental design, and logistics.

Perhaps the most transformative impact of data-driven science on geographic knowledge discovery will be through data exploration and hypothesis generation. Similar to a telescope or microscope, systems for capturing, storing, and processing massive amounts of data can allow investigators to augment their perceptions of reality and see things that would otherwise be hidden or too faint to perceive. From this perspective, data-driven science is not necessarily a radically new approach, but rather a way to enhance inference for the longstanding processes of exploration and hypothesis generation prior to knowledge construction through analysis, modeling, and verification (Miller 2010).

Data-driven knowledge discovery has a philosophical foundation: *abductive reasoning*, a form of inference articulated by astronomer and mathematician C. S. Peirce (1894-1914). Abductive reasoning starts with data describing something and ends with a hypothesis that explains the data. It is a weaker form of inference relative to deductive or inductive reasoning: deductive reasoning shows that X *must* be true, inductive reasoning shows that X *is* true, while abductive reasoning shows only that X *may* be true. Nevertheless, abductive reasoning is critically important in science, particularly in the initial discovery stage that precedes the use of deductive or inductive approaches to knowledge construction (Miller 2010).

Abductive reasoning requires four capabilities: i) the ability to posit new fragments of theory; ii) a massive set of knowledge to draw from, ranging from common sense to domain expertise; iii) a means of searching through this knowledge collection for connections between data patterns and possible explanations, and; iv) complex problem-

solving strategies such as analogy, approximation, and guesses. Humans have proven to be more successful than machines in performing these complex tasks, suggesting that data-driven knowledge discovery should try to leverage these human capabilities through methods such as geovisualization rather than try to automate the discovery process. Gahegan (2009) envisions a human-centered process where geovisualization serves as the central framework for creating chains of inference among abductive, inductive, and deductive approaches in science, allowing more interactions and synergy among these approaches to geographic knowledge building.

One of the problems with Big Data is the size and complexity of the information space implied by a massive multivariate database. A good data-exploration system should generate all of the interesting patterns in a database, but *only* the interesting ones to avoid overwhelming the analyst. Two ways to manage the large number of potential patterns are *background knowledge* and *interestingness measures*. Background knowledge guides the search for patterns by representing accepted knowledge about the system to focus the search for novel patterns. In contrast, we can use interestingness measures *a posteriori* to filter spurious patterns by rating each pattern based on dimensions such as simplicity, certainty, utility, and novelty. Patterns with ratings below a user-specified threshold are discarded or ignored (Miller 2010). Both of these approaches require formalization of geographic knowledge, a challenge discussed earlier in this paper.

5.2. Data-driven modeling

Traditional approaches to modeling are deductive: the scientist develops (or modifies or borrows) a theory and derives a formal representation that can be manipulated to generate predictions about the real world that can be tested with data. Theory-free modeling, on the other hand, builds models based on induction from data rather than through deduction from theory.

The field of economics has flirted with data-driven modeling in the form of *general-to-specific* modeling (Miller 2010). In this strategy, the researcher starts with the most

complex model possible and reduces it to a more elegant one based on data, founded on the belief that, given enough data, only the true specification will survive a sufficiently stringent battery of statistical tests designed to pare variables from the model. This contrasts with the traditional specific-to-general strategy where one starts with a spare model based on theory and conservatively builds a more complex model (Hoover and Perez 1999). However, this approach is controversial, with some arguing that given the enormous number of potential models one would have to be very lucky to encompass the true model within the initial, complex model. Therefore, predictive performance is the only relevant criterion; explanation is irrelevant (Hand 1999).

Geography has also witnessed attempts at theory-free modeling, also not without controversy. Stan Openshaw is a particularly strong advocate for using the power of computers to build models from data: examples include the Geographical Analysis Machine (GAM) for spatial clustering of point data, and automated systems for spatial interaction modeling. GAM uses a technique that generates local clusters or "hot spots" without requiring a priori theory or knowledge about the underlying statistical distribution. GAM searches for clusters by systematically expanding circular search from locations within a lattice. The system saves circles with observed counts greater than expected and then systematically varies the radii and lattice resolution to begin the search again. The researcher does not need to hypothesize or have any prior expectations regarding the spatial distribution of the phenomenon: the system searches, in a brute-force manner, all possible (or reasonable, at least) spatial resolutions and neighborhoods (Charlton 2008; Openshaw et al. 1987).

GAM is arguably an exploratory technique, while Openshaw's automated system for exploring a universe of possible spatial interaction models leaps more into the traditional realm of deductive modeling. The automated system uses genetic programming to breed spatial interaction models from basic elements such as the model variables (e.g., origin inflow and destination outflow totals, travel cost, intervening opportunities), functional forms (e.g., square root, exponential), parameterizations, and binary operators (add,

subtract, multiply and divide) using goodness of fit as a criterion (Diplock 1998; Openshaw 1988).

One challenge in theory-free modeling is that it takes away a powerful mechanism for improving the effectiveness of a search for an explanatory model – namely, theory. Theory tells us where to look for explanation, and (perhaps more importantly) where not to look. In the specific case of spatial interaction modeling, for example, the need for models to be dimensionally consistent can limit the options, though the possibility of dimensional analysis (Gibbings, 2011) was not employed in Openshaw's work. The information space implied by a universe of potential models can be enormous even in a limited domain such as spatial interaction. Powerful computers and clever search techniques can certainly improve our chances (Gahegan 2000). But as the volume, variety, and velocity of data increase, the size of the information spaces for possible models also increases, leading to a type of arms race with perhaps no clear winner.

A second challenge in data-driven modeling is that the data drive the form of the model, meaning there is no guarantee that the same model will result from a different data set. Even given the same data set, many different models could be generated that fit the data, meaning that slight alterations in the goodness-of-fit criterion used to drive model selection can produce very different models (Fotheringham 1998). This is essentially the problem of statistical overfitting, a well-known problem with inductive techniques such as artificial neural networks and machine learning. However, despite methods and strategies to avoid overfitting, it appears to be endemic: some estimate that three-quarters of the published scientific papers in machine learning are flawed due to overfitting (The Economist 19 October 2013).

A third challenge in theory-free modeling is the complexity of resulting models. Traditional model building in science uses parsimony as a guiding principle: the best model is the one that explains the most with the least. This is sometimes referred to as "Occam's Razor": given two models with equal validity, the simpler model is better. Model interpretation is an informal but key test: the model builder must be able to

explain what the model results say about reality. Models derived computationally from data and fine-tuned based on feedback from predictions can generate reliable predictions from processes that are too complex for the human brain (Townsend 2013; Weinberger 2011). For example, Openshaw's automated system for breeding spatial interaction models has been known to generate very complex, non-intuitive models (Fotheringham 1998), many of which are also dimensionally inconsistent. Figure 3 illustrates some of the spatial interaction models generated by Openshaw's automated system; as can be seen, they defy easy comprehension.

$$T_{ij} = \frac{\left(\frac{\left(\arctan V_{ij}^{-0.8}\right) + 1}{\text{cotan}\left(\text{lgamma}\,(O_i)^{-0.4}\right)} + 0.19\ln(O_i D_j)\right)}{\text{lgamma}\left(\cos(O_i D_j) + 0.04\right) - \text{cotan}(\cosh A_{ij} - 0.9)}$$

$$T_{ij} = \left(\arctan\left(\tanh(V_{ij})^{-1.05}\right) + 1\right)\text{atan}(\text{lgamma}\,(O_i)^{-0.82}) + \frac{\sinh\left(\tanh(O_i D_j) - 0.6\right)}{\text{lgamma}\left(\cos(O_i D_j) + 0.4\right)} - \text{cotan}\left(\tan(A_{ij})^{2.3}\right)$$

$$T_{ij} = \frac{\left(\text{atan}\left(V_{ij}^{-1.4}\right)\tanh(\text{lgamma}(O_i)^{-0.9}) + \tanh(D_j)\right)}{\cos\left(5.1\tanh(O_i D_j)\right)}$$

Figure 3: Three of the spatial interaction models generated by Openshaw's automated modeling system (Openshaw 1988)

The knowledge from data-driven models can be complex and non-compressible: the data are the explanation. But if the explanation is not understandable, do we really have an explanation? Perhaps the nature of explanation is evolving. Perhaps computers are fundamental in data-driven science not only for discovering but also for representing

complex patterns that are beyond human comprehension. Perhaps this is a temporary stopgap until we achieve convergence between human and machine intelligence as some predict (Kurzweil 1999). While we cannot hope to resolve this question (or its philosophical implications) within this paper, we can add a cautionary note from Nate Silver: telling stories about data instead of reality is dangerous and can lead to mistaking noise for signal (Silver 2012).

A final challenge in data-driven spatial modeling is de-skilling: a loss of modeling and analysis skills. While allocating mundane tasks to computers frees humans to perform sophisticated activities, there are times when mundane skills become crucial. For example, there are documented cases of airline pilots, due to a lack of manual flying experience, reacted badly in emergencies when the autopilot shuts off (Carr 2013). Although rarely life-threatening, one could make a similar argument about automatic model building: if a data-driven modeling process generates anomalous results, will the analyst be able to determine if they are artifacts or genuine? With Openshaw's automated spatial interaction modeling system, the analyst may become less skilled at spatial interaction modeling and more skilled at combinatorial optimization techniques. While these skills are valuable and may allow the analyst to reach greater scientific heights, they are another level removed from the empirical system being modeled. However, the more anomalous the results, the deeper the thinking required.

A solution to de-skilling is to force the skill: require it as part of education and certification, or design software that encourages or requires analysts to maintain some basic skills. However, this is a difficult case to make compared to the hypnotic call of sophisticated methods with user-friendly interfaces (Carr 2013) Re-reading Jerry Dobson's prescient essay on automated geography thirty years later (Dobson 1983), one is impressed by the number of the activities in geography that used to be painstaking but are now push-button. Geographers of a certain age may recall courses in basic and production cartography without much nostalgia. What skills that we consider essential today will be considered the pen, ink, and lettering kits of tomorrow? What will we lose?

6. Conclusion

The context for geographic research has shifted from a data-scarce to a data-rich environment, in which the most fundamental changes are not the volume of data, but the variety and the velocity at which we can capture georeferenced data. A data-driven geography may be emerging in response to the wealth of georeferenced data flowing from sensors and people in the environment. Some of the issues raised by data-driven geography have in fact been longstanding issues in geographic research, namely, large data volumes, dealing with populations and messy data, and tensions between idiographic versus nomothetic knowledge. However, the belief that spatial context matters is a major theme in geographic thought and a major motivation behind approaches such as time geography, disaggregate spatial statistics, and GIScience. There is potential to use Big Data to inform both geographic knowledge discovery and spatial modeling. However, there are challenges, such as how to formalize geographic knowledge to clean data and to ignore spurious patterns, and how to build data-driven models that are both true and understandable.

Cautionary notes need to be sounded about the impact of data-driven geography on broader society (see Mayer-Schonberger and Cukier 2013). We must be cognizant about where this research is occurring – in the open light of scholarly research where peer review and reproducibility is possible, or behind the closed doors of private-sector companies and government agencies, as proprietary products without peer review and without full reproducibility. Privacy is a vital concern, not only as a human right but also as a potential source of backlash that will shut down data-driven research. We must be careful to avoid pre-crimes and pre-punishments (Zedner 2010): categorizing and reacting to people and places based on potentials derived from correlations rather than actual behavior. Finally, we must avoid a data dictatorship: data-driven research should support, not replace, decision-making by intelligent and skeptical humans. Some of the other papers in this special issue explore these challenges in depth.

Literature Cited

Anderson, C. (2008) "The end of theory: The data deluge makes the scientific method obsolete," *Wired*, 16.07.

Anselin, L. (1995) "Local indicators of spatial association: LISA," *Geographical Analysis*,27, 2, 93–115.

Batty, M. (2012) "Smart cities, Big Data," *Environment and Planning B*, 39, 191–193.

Butler, D. (2008) "Web data predict flu." *Nature*, 456, 287–288.

Carr, N. (2013) "The great forgetting," *The Atlantic*, November 2013, 77-81.

Cetin, N., Nagel, K., Raney, B. and Voellmy, A. (2002) "Large-scale multi-agent transportation simulations," *Computer Physics Communications*, 147, 559–564.

Charlton, M. (2008) "Geographical Analysis Machine (GAM)," in K. Kemp (ed.) *Encyclopedia of Geographic Information Science*, Sage, 179–180.

Cresswell, T. (2013). *Geographic Thought: A Critical Introduction* Wiley-Blackwell.

Diplock, G. (1998) "Building new spatial interaction models by using genetic programming and a supercomputer," *Environment and Planning A*, 30, 1893-1904

Dobson, J. E. (1983) "Automated geography," *The Professional Geographer*, 35, 135-143.

Dumbill, E. (2012) "What is big data? An introduction to the big data landscape," http://strata.oreilly.com/2012/01/what-is-big-data.html; last accessed 17 April 2014.

The Economist (19 October 2013) "Trouble at the lab," 26-30.

Flake G. W. (1998) *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation*, Cambridge, MA: MIT Press.

Fotheringham, A. S. (1998) "Trends in quantitative methods II: Stressing the computational," *Progress in Human Geography*, 22, 283–292

Fotheringham, A. S., Brunsdon, C., Charlton, M. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*," Chichester, UK: Wiley.

Gahegan, M. (2000) "On the application of inductive machine learning tools to geographical analysis," *Geographical Analysis,* 32, 113-139

Gahegan, M. (2009) "Visual exploration and explanation in geography: Analysis with light," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery*, second edition, Taylor and Francis, 291-324.

Gibbings, J. C. (2011) *Dimensional Analysis*, New York: Springer.

Glaser, B. G., Strauss, A. L. (1967) *The Discovery of Grounded Theory*, Chicago: Aldine.

Goffman, E. (1959) *The Presentation of Self in Everyday Life*, Anchor Books.

Goodchild, M. F. (2004) "GIScience, geography, form, and process," *Annals of the Association of American Geographers*, 94, 4, 709–714.

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

Goodchild, M. F., Egenhofer, M. J., Kemp, K. K., Mark, D. M., Sheppard E. (1999) "Introduction to the Varenius project," *International Journal of Geographical Information Science*, 13, 8, 731–745.

Goodchild, M. F., Li, L. (2012) Assuring the quality of volunteered geographic information. *Spatial Statistics* 1: 110–120. DOI: 10.1016/j.spasta.2012.03.002

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, 255-261

Guptill, S. C., Morrison, J. L., editors (1995) *Elements of Spatial Data Quality*, Oxford: Elsevier.

Haklay, M. (2010) "How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets," *Environment and Planning B: Planning and Design*, 37, 682-703

Hand, D. J. (1999) "Discussion contribution on 'Data mining reconsidered: Encompassing and the general-to-specific approach to specification search' by Hoover and Perez," *Econometrics Journal*, 2, 241–243

Hartshorne, R. (1939) *The Nature of Geography: A Critical Survey of Current Thought in the Light of the Past*, Association of American Geographers.

Hey, T., Tansley S., Tolle, K. (eds.) (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*.

Hoover, K. D., Perez, S. J. (1999) "Data mining reconsidered: Encompassing and the general-to-specific approach to specification search," *Econometrics Journal*, 2, 167–191.

Kitchin, R. (2014) "Big data and human geography: Opportunities, challenges and risks," *Dialogues in Human Geography*, 3, 262-267.

Kurzweil, R. (1999) *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*, Vintage.

Mayer-Schonberger, V., Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*

Merton, R. K. (1967) "On sociological theories of the middle range," in R. K. Merton, *On Theoretical Sociology*, New York: The Free Press, 39–72.

Miller, H. J. (2007) "Place-based versus people-based geographic information science," *Geography Compass*, 1, 503-535.

Miller, H. J. (2010) "The data avalanche is here. Shouldn't we be digging?" *Journal of Regional Science*, 50, 181–201.

O'Leary, M. (2012) "Eurovision statistics: post-semifinal update," *Cold Hard Facts* (May 23). Available: http://mewo2.com/nerdery/2012/05/23/eurovision-statistics-post-semifinal-update/ (accessed October 25, 2013).

Openshaw, S. (1988) "Building an automated modeling system to explore a universe of spatial interaction models," *Geographical Analysis*, 20, 31-46.

Openshaw, S., Charlton, M., Wymer, C., Craft, A. (1987) "A Mark I geographical analysis machine for the automated analysis of point data sets," *International Journal of Geographical Information Systems*, 1, 335–358

Openshaw, S., Taylor, P.J. (1979) "A million or so correlation coefficients: three experiments on the modifiable areal unit problem," in N. Wrigley, editor, *Statistical Methods in the Social Sciences*, London: Pion, 127-144.

Preis, T., Moat, H. S., Stanley, H. E. (2013) "Quantifying trading behavior in financial markets using *Google Trends*," *Scientific Reports*, 3, No. 1684. doi:10.1038/srep01684

Raymond, E. S. (2001) *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, Sebastopol, CA: O'Reilly Media.

Schuurman, N. (2000) "Trouble in the heartland: GIS and its critics in the 1990s," *Progress in Human Geography*, 24, 569-589.

Silver, N. (2012) *The Signal and the Noise: Why Most Predictions Fail – But Some Don't*

Smith, N. (1992) "History and philosophy of geography: Real wars, theory wars," *Progress in Human Geography*, 16, 2, 257–271.

Sui, D. (2004) "GIS, cartography, and the "Third Culture": Geographic imaginations in the computer age," *Professional Geographer*, 56, 62-72.

Sui, D. and DeLyser, D. (2012) "Crossing the qualitative-quantitative chasm I: Hybrid geographies, the spatial turn, and volunteered geographic information (VGI)," *Progress in Human Geography*, 36, 111-124.

Sui, D. and DeLyser, D. (2013) "Crossing the qualitative-quantitative chasm II: Inventive approaches to big data, mobile methods, and rhythmanalysis," *Progress in Human Geography*, 37, 293-305.

Sui, D. and Goodchild, M. F. (2011) "The convergence of GIS and social media: Challenges for GIScience," *International Journal of Geographical Information Science*, 25, 1737-1748.

Sui, D., Goodchild, M. F. and Elwood, S. (2013) "Volunteered geographic information, the exaflood, and the growing digital divide,' in D. Sui, S. Elwood, and M. F. Goodchild (eds.) *Crowdsourcing Geographic Knowledge*, Springer, 1-12.

Taleb, N. N. (2007) *The Black Swan: The Impact of the Highly Improbable*, Random House

Townsend, A. (2013) *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*, Norton.

Tsou, M.H., Yang, J.A., Lusher, D., Han, S., Spitzberg, B., Gawron, J.M., Gupta, D., An, L. (2013) "Mapping social activities and concepts with social media (Twitter) and web search engines (Yahoo and Bing): a case study in 2012 US Presidential Election." *Cartography and Geographic Information Science*, 40, 4, 337–348.

Waldrop, M. M. (1990) "Learning to drink from a fire hose," *Science*, 248, 674-675.

Warntz, W. (1989) **"**Newton, the Newtonians, and the Geographia Generalis Varenii," *Annals of the Association of American Geographers*, 79, 2, 165–191

Watts, D. J. (2011) *Everything is Obvious – Once You Know the Answer*, Crown Business.

Weinberger, D. (2011) "The machine that would predict the future," *Scientific American*, November 15, 2011. http:// www.scientificamerican.com/ article.cfm?id = the-machine-that-would-predict.

Zedner, L. (2010) “Pre-crime and pre-punishment: a health warning,” *Criminal Justice Matters*, 81, 24-25.