International Journal of Communication 8 (2014), 1765–1783 1932–8036/20140005

IJoC Big Data, Big Questions

Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data[1]

JIM THATCHER
Clark University, USA

Amid the continued rise of big data in both the public and private sectors, spatial information has come to play an increasingly prominent role. This article defines big data as both a sociotechnical and epistemic project with regard to spatial information. Through interviews, job shadowing, and a review of current literature, both academic researchers and private companies are shown to approach spatial big data sets in analogous ways. *Digital footprints* and *data fumes*, respectively, describe a process that inscribes certain meaning into quantified spatial information. Social and economic limitations of this data are presented. Finally, the field of geographic information science is presented as a useful guide in dealing with the "hard work of theory" necessary in the big data movement.

Keywords: Big data, critical big data, spatial big data, capital, data fumes

Introduction

In late January 2012, Daniel Rasmus, a writer for *Fast Company*, predicted 2012 to be the "year of Big Data" (Rasmus, 2012). By November of that year, Chris Taylor of *Mashable* had declared Nate Silver and his partner, big data, the "absolute, undoubted winner" of the U.S. presidential election (Taylor, 2012).[2] From marketing (Baker, 2012) to health care (Cerrato, 2012) to the national science funding agencies (National Science Foundation, 2012)**,** big data and its related models and analytical approaches have gained prominence. For many practitioners, the spatial information attached to large sets of data plays an increasingly important role in analysis. By adding "where" to information about who is doing

[1] An earlier, compact version of this article appeared in the *Proceedings of the 26th International Cartographic Conference*.

[2] Silver rose to prominence using aggregated statistics to predict U.S. national elections. His principle website, http://fivethirtyeight.com, was licensed by *The New York Times* throughout the 2012 U.S. presidential election.

Jim Thatcher: jethatcher@gmail.com

Date submitted: 2013–04–10

what, when, and with whom, some authors have gone so far as to proclaim the "end of theory" as the rapid aggregation and analysis of data destroy the need for social explanation, allowing the numbers to "speak for themselves" (Anderson, 2008). As government, private industry, and academic researchers all rush to embrace big data, some scholars have pushed back against this rising tide, calling into question the epistemological, economic, and political commitments big data entails (Batty, 2012; boyd & Crawford, 2012; Burgess & Bruns, 2012). As part of a special section questioning the big data movement, this article examines the commonalities and tacit epistemological commitments inherent in big data approaches in both academic research and private corporations. It argues that, at present, the limits of both what can be known through and done with big data lie in the hands of a relatively small group of individuals and companies that control the data chain. Drawing parallels to discussions in the field of geographic information science (GIS) in the 1990s, the article suggests a critical theoretical approach to understanding the limits and opportunities offered by big data.

This article proceeds in five parts: First, it will briefly define big data in relation to location and spatial information. Recognizing that data has always been big, this article defines big data as both a sociotechnical and epistemic project that involves the rapid combination, aggregation, and analysis of data through an abductive process**.** Building from this definition, it then demonstrates that, despite using different terms, academic researchers and private corporations are approaching spatial big data in analogous ways. Drawing from ethnographic interviews of mobile application designers and developers and a literature review of current big data research, the terms *digital footprints* and *data fumes* are introduced and defined. Third, two problems with this approach are foregrounded for researchers. On the one hand, rather than fully capturing life as researchers hope, end-user interactions within big data are necessarily the result of decisions made by an extremely small group of programmers working for private corporations that have promulgated through the mobile application ecosystem. On the other hand, in accepting that the data gathered through mobile applications reveal meaningful information about the world, researchers are tacitly accepting a commodification and quantification of knowledge. Big data researchers are intrinsically linking the epistemic limits of their own research to information generated through motivations for profit both from end users and application creators. Although science and technology have long been entwined with capitalist profit motives (Marcuse, 1941/1982), this represents a new relation as the very limits of knowledge are set through the data infrastructure of private corporations.

The issues and concerns raised for big data researchers parallel a set of similar debates that occurred within the GIS community in the 1990s. Geographic information systems, predominantly corporately created software packages, were on the rise in terms of popularity and use. A debate around the very name—systems versus science—led to a broader discussion on the relations between science, technology, and modern capitalism. Drawing from this discussion, the fourth section of this article draws three parallels between current concerns over big data and previous discussions within the GIS community. Big data researchers must engage in the "hard work of theory" (Pickles, 1997, p. 370) to actively engage the epistemological and ontological commitments of the big data paradigm. The article concludes with a summation of the arguments presented.

Definitions: Big Data and Spatial Big Data

Definitions of what constitutes big data are as abundant as articles touting its many abilities. In this section, big data is defined in relation to how spatial information functions with big data research. Rather than a singular definition, big data is defined first from a technical perspective and then from a wider social and epistemological perspective. The former situates big data as an ever-shifting target that has less to do with size and more with the ability to rapidly combine, aggregate, and analyze, while the latter situates big data research as a particular view of scientific praxis that privileges abductive reasoning and model building. Finally, the information that constitutes spatial big data sets is defined.

From a technical perspective, data has always been big as what constitutes a large data set has shifted with technological advances (Farmer & Pozdnoukhov, 2012). In the 1980s, a fully automatic tape library capable of storing the entire 1980 U.S. Census database was the definition of Big Data; today's home computers often have 10 times that storage amount (Jacobs, 2009). Rather than defining the technical nature of big data through an ever-increasing quantity measurement, it is more useful to consider it through a combination of "high-volume (increasing amount of data), high-velocity (speed of data in and out), and/or high-variety (range of data types and sources)" (Horvath, 2012, p. 15) that requires new forms of analysis.

In particular, it is not simply the size of a data set that matters, but the ability to access it in a meaningful and timely manner. While consumer-grade computers at the time of writing may be able to store a database several terabytes in size, they often lack the ability to effectively analyze it. For example, Jacobs (2009) notes that PostgreSQL requires more than 24 hours to perform a simple query on over 1 billion rows on a current-generation Mac Pro workstation due to the sorting algorithm used for large data sets. Further, hard computational limits are often encountered. R, another popular analysis software, uses 32-bit integers to index arrays even in nominally 64-bit versions. This means a hard limit of around 2 billion rows is emplaced programmatically even in 64-bit systems with sufficient memory to hold more rows. If the "pathologies of big data are primarily those of analysis" (Jacobs, 2009, p. 4) then, for those wishing to use big data, the decisions made by designers and developers of big data analysis systems place limits on both what is "big" and how it can be accessed. For this reason, Jacobs concludes with a technical definition of big data as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" (Jacobs, 2009, p. 11), so the tape array in the 1980s is just as much big data as the tera- and petabytes of information stored in digital warehouses today. This technical definition incorporates the acceptance of a technocracy that designs and sells the new tools needed to handle big data. The end user, whether she or he is a researcher or a corporate advertiser, predominantly relies upon the big data tools provided by a limited set of vendors, as common techniques and approaches are technically insufficient for new big data purposes. These vendors are removed from the specific data in both time and space, but their design decisions shape the very questions that can be asked of it. This relationship exists in parallel with that experienced by the GIS community in the late 1990s and is discussed in greater detail in Section 5.

Big data, though, often refers to more than the technical and methodological means through which large data sets are accessed. The rapid growth of big data approaches and the acceptance of a big data movement in both private firms and academic research belie a sociotechnical phenomenon that accepts certain epistemological and cultural views on scientific praxis and the nature of reality. In this sense, big data represents the latest iteration of the desire to find efficiency and meaning in quantitative analysis. Lohr (2012) traces this desire for businesses back to Taylorism's "scientific management," although it could just as easily be seen in Max Weber's earlier writing on bureaucracy (Weber, 1946/1973). Regardless of its historical origin, the "belief that large data sets offer a higher form of intelligence and knowledge" (boyd & Crawford, 2012, p. 663) is a mythological one. It requires a certain *belief* in the nature and value of quantification, *faith* in the viability of a model to accurately predict reality. The resulting "computational turn in thought and research" and its resulting epistemological limits are discussed later in this article; here it is important to note that the big data movement requires a belief that life can be captured and modeled by data or even fully transformed into it (Berry, 2011; boyd & Crawford, 2012, p. 665).

The triumph of big data is the acceptance of these myths. With reality accurately captured in ever-larger data sets, scientific praxis transforms into a new paradigm of manipulation and exploration. Hey, Tansley, and Toelle (2009) deem this the "fourth paradigm" beyond the traditional ones of empiricism, analysis, and simulation. Data-intensive science is about "exploiting" large data sets for "discovery and understanding of deeply embedded facts and meanings" (Horvath, 2012, p. 16). Large numbers of potential correlations are equally considered, in contrast to traditional methods driven from a theoretically based hypothesis and a small number of testable variables (Batty, 2012). Science in the big data era is an abductive process in which "hypotheses are developed to account for observed data" (Farmer & Pozdnoukhov 2012, p. 5). The underlying problems that emerge with the acceptance of this praxis are discussed later in this article.

For both researchers and private corporations, spatial big data is predominantly generated through mobile device use, such as smart phones with embedded GPS receivers. Mobile devices and the users who carry them are both active and passive sensors through which the world may be monitored. The big data perspective for spatial information involves "using large-scale mobile data as input to characterize and understand real-life phenomena" (Laurila et al., 2012, p. 1). Most current studies describing themselves as "big data" with a spatial component revolve around two mobile software platforms: Twitter and Foursquare. The data captured and stored from mobile users of these applications is seen as containing "rich information" (Long, Jin, & Joshi, 2012). For example, the Livehoods project, conducted through Carnegie Mellon University, uses Foursquare check-ins[3] and tweets to "discover [a city's] structure" that reflect the "everyday experiences of real people" (Livehoods, 2012).[4] The big data viewpoint posits that check-ins and tweets reveal meaningful information that can be used by researchers to study society and by companies to increase profits. These two distinct goals are both found within the data already generated by end users of certain applications. The next section describes how companies and researchers describe their work through the terms *digital footprints* and *data fumes*. The two terms

[3] For this article, a check-in is defined as an action an end user takes that broadcasts her or his supposed location at a given time and place.

[4] The resulting maps created by the Livehoods project may be found at http://livehoods.org/maps.

reveal that, while the goals may differ, corporations and researchers are both accepting the purpose and value of the spatial information given off by mobile device use.

Spatial Big Data: Data Fumes and Digital Footprints

This section presents two terms used by different groups when referring to spatial big data information: *digital footprints* and *data fumes*. Taken directly from current big data research, digital footprints refer to information that is given off by actions humans are already taking. The information contained in these footprints, which may be generated through purposeful action or passive recording, is believed to offer deep and novel insights into both individual experience and society as a whole. A subset of digital footprints called data fumes refers to information given off through the use of already existing applications. This term emerged from a series of interviews with private industry designers and developers.

Twenty-four semistructured interviews were conducted between April and August 2012. These interviews lasted between 1 and 3 hours and focused broadly on the preconceptions developers held going into project development, how these preconceptions changed during development, and what their applications provided for end users. The interviews were then transcribed and thematically coded. Interviewees spoke repeatedly of "adding value" to end users' generation of data and of using the spatially aware check-in to deepen end users' experiences, and these broad themes became the basis of further analysis. Although limited in number, the interviews ranged from independent, contracted developers to the chief executive officer of a large start-up with several million dollars in funding. The names of interviewees have been obfuscated and, when appropriate, entirely withheld. In addition to the formal interview set, reviews of the current academic and popular literature on big data are used to construct the two concepts and demonstrate their related nature.

Digital Footprints

Much has been written recently on the transformation of everyday citizens into remote sensors capable of recording various physical, social, and economic metrics (Goodchild, 2007; Sui, 2008). As humans move through, record, and share information about the world, they have been called the "best [sensors] of all" ("A Sea of Sensors," 2010). Digital social media data has been used to measure everything from forest fires (Goodchild & Glennon, 2010), to earthquakes (Sakaki et al., 2010), to rioting (Presley, 2011). In this context, Martino et al. (2010) write that citizens are evolving "from urban actors to urban indicators" (p. 1), while Paulos, Honicky, and Hooker (2008) call attention to mobile devices as "personal measurement instruments" (p. 416) that enable new forms of citizen science. Following these observations, Exner, Zeile, and Streich (2011) suggest that "humans should be considered as implicit (passive) and explicit (active) sensors" (p. 1090). Implicit when referring to the more or less continual stream of information passed passively along by a mobile smart phone (GPS signal), explicit when an end user makes use of his or her smart phone to purposely demarcate a bit of information (liking a location in Foursquare or Facebook, uploading a geotagged picture to Flickr).

For researchers of the data sets produced by both implicit and explicit human sensor data, digital footprints involve an additional distinction beyond active and passive. Digital footprints require the researcher to invest a meaning into the data set that goes beyond the intentionality of the person creating the data. This can occur with a passive- or actively- produced data set and occurs when information generated in one context is used to study a different one. The researcher believes that the data can accurately reveal something other than the intended meaning of the end user. For example, Lathia, Quercia, and Crowcroft (2012) used the passive sensor of transit pass use records to measure "community well-being," investing the travel to and from various areas of a city with the ability to determine those areas' community well-being. The remainder of this section provides examples of the types of claims made when researching digital footprints. It is important to note here that this article is not calling into question the validity of the results produced by these researchers, nor the integrity of their methods. Rather, digital footprint is developed as a necessary concept to discuss the epistemological limitations and commitments of spatial big data research.

Researchers of digital footprints often make claims akin to "[d]igital traces of citizen activities in a city contain abundant detail on various aspects of urban dynamics" (Kling & Pozdnoukhov, 2012, p. 482) or "the large amount of geo-spatially tagged data allows one to approach questions of society-level interest in an entirely data-driven manner" (Joseph, Tan, & Carley, 2012, p. 919). The study of check-ins "reveals meaningful spatio-temporal patterns and offers the opportunity to study both user mobility and urban spaces" (Noulas, Scellato, Mascolo, & Pontil, 2012, p. 570) as the data generated by location-based social networks offers unprecedented opportunities for research. It can be used to "understand the relation between urban space, social events, and mobility" (Lathia et al., 2012, p. 93). This type of research will improve the lives of "businesses, urban planners, real-estate developers, researchers and end users" (Cranshaw, Schwartz, Hong, & Sadeh, 2012, p. 59). These and other studies all accept that, through the examination of spatial big data sets, meaning beyond the end user's intent may be found. This belief requires that social life can be accurately captured in big data sets created and maintained for purposes other than research. Further, these data sets can be parsed of meaning through an abductive process of discovery. Research into digital footprints accepts that deep societal and personal meaning is captured and stored in data sets to be revealed through accurate analysis and modeling. Data fumes are the means through which private companies approach similar sets of spatial big data information. Rather than inscribing meaning into the data set, private companies attempt to add value to an end user's life by manipulating existing data and data sets.

***Data Fume*s**

Since 2009, more than $115 million has been invested in location-based start-up companies, making them a major factor in the generation and manipulation of and access to spatial information data sets (Wilson, 2012). Despite repeated announcements of the check-in's "death" (Mitchell, 2012), the vast majority of these start-ups have focused on the *check-in*. As a purposeful, end user–initiated action creates the data point, check-ins are a form of active sensing. "Users *check-in* at *venues* where they are present, effectively reporting their location" (Carbunar & Potharaju, 2012, p. 182) via the application to those who have access to the information—a group that may include friends, the makers of the application, other corporations, and researchers. Check-ins may be distinguished from other, more ubiquitous types of spatial tracking in that they are discrete events: A user checks in at a movie theater and later checks in at a restaurant, but the period between check-ins is not captured. Although other services exist, the dominant systems for checking in at the moment remain Foursquare and Facebook Places. Although different sources cite different numbers, since 2009 it is estimated that Foursquare has seen more than 2.5 billion check-ins (Benner & Robles, 2012), with 20 million coming in March 2012 alone (Long et al., 2012). Meanwhile, Facebook, with over two-hundred million active users, has had upwards of two billion actions tagged with locations in April of 2012 (Long et al., 2012). With hundreds of millions of dollars invested in them and billions of discrete data points being generated, the check-in is clearly alive and well.

Data fumes are attempts to add value to the check-in, to make it more meaningful and to profit from the provision of this meaning. The term itself comes from an interview with the chief executive officer of a popular mobile spatial start-up and former developer of another popular geospatial application:

> In the end, [my applications] all tie to location data, they're all sitting on top of other people's location data and behaviors that we already have when we're going out, like checking-in.
>
> Say Foursquare didn't exist, I'd have to convince everyone in the world to check-in and share their location. The fact that Foursquare exists and aims to get everyone in the world to check-in . . . it means they can concentrate on that, and I can use all the data they're already collecting, the fumes, the exhaust of their data to do really interesting things.
>
> Our intent was never to add any more work to your life, but only to add value based on work you're already doing or other people are doing.

These three quotes capture the essence of mobile spatial big data and its uses by corporations. Mobile spatial start-ups are sitting on top of already existing big data sets, and they are leveraging these sets to do new things, but they are fundamentally not interested in changing behavior. In another quote, the same interviewee (assigned the name John) discussed a desire to not change "base behavior," the check-in, but to let others focus on that. He, and an increasing number of similar start-ups, focuses on the "fumes" or "exhaust" already given off by base behaviors driven by larger, already established corporations. A designer (assigned the name Lon), described building his company's application on Facebook as obvious, because "millions of people use it. It's where the users are." In building off existing frameworks, designers and developers seek to add value, to maximize what John called the "cost versus reward ratio." In this, they accept that the information contained in the fumes has meaning with regard to individuals' lives.

Data fumes refer directly to the data and information that is generated through actions the end user already takes. An application that "live[s] on data fumes," as another interviewee (assigned the name Paul) stated, is one that seeks to access and manipulate this already existing data to "add value" to the end user's experience. Although John and Paul were the only two interviewees to directly mention the term *data fumes*, this type of relation to existing big data sets is increasingly common. Of the mobile spatial applications represented in the interview set, only one was not building its product off of an existing platform. John’s most recent application, for example, uses Foursquare check-in data displayed on a map created by another third-party vendor. Every application relied on mapping information provided by either Google or OpenStreetMap. Every interviewee accepted that the data being generated, the millions of discrete check-ins, represented meaningful information on which to act. The existing behavior offered a rich source on which to add value, to capitalize on the big data sets already being generated, and to profit.

While private companies living off of data fumes run the risk of losing control of and access to their data, researchers face an additional risk: They are necessarily inscribing information created in one context to have meaning for another. From one perspective, this is a banal point; research has long analyzed data beyond the intent of data’s original author. For example, ethnographic work is often analyzed and reinterpreted in ways that diverge from the research subject’s perspective, and economists regularly use substitution to allow one variable to represent another. However, here the inscription of meaning occurs with an unconsidered eye toward the epistemological commitments and limitations with which it comes. big data offers “strong quantitative” and “data-driven” approaches that suggest better, deeper, and new understandings of society than ever available before (Joseph et al., 2012). Context and qualitative meaning are stripped from the data as unnecessary when the “numbers speak for themselves” (Anderson, 2008).

Capital, Private Ownership, and What Can Be Known With Big Data

Data fumes and digital footprints engage identical big data data sets with an identical inscription into and belief in the meaning of the data. Whereas private companies seek to add value to existing behaviors, academic researchers derive meaning through the abductive exploration of the data sets produced by these behaviors. In each case, the data given off is accepted as useful, meaningful, and Big. These beliefs produce two intertwined problems with the very nature of mobile big data research: First, although the data sets may be big in size, their format, composition, and access are regulated by a very small number of individuals. Using information generated by other applications forces a reliance upon the means through which that data is made available. What can be known and what can be done with mobile spatial big data are delimited by the decisions of a small set of programmers and businesses, their choices promulgating throughout the mobile application ecosystem. Second, as these decisions are made by private corporations engaged in profit-seeking activities, when academic researchers accept the resulting data as meaningful, they are fundamentally accepting an epistemological framework of knowledge structured through capitalist imperatives. When social science researchers use big data produced and maintained by private corporations, they tacitly accept the incentives and requirements that shaped the information. Examples from published research and the interview set highlight these problems.

What Data? For Whom? And When?

Much of the allure of digital footprints and data fumes is that they build from data that already exists. It lets start-ups focus on “add[ing] value based on work you’re already doing” (John). For academic researchers, access to new data sets offers tantalizing possibilities of better understanding the world as the spatial information studied helps "gauge diverse social indicators ranging from political views to consumer tastes to public opinions" (Cheng, Caverlee, Lee, & Sui, 2011, p. 81). However, both the form of the information accessed and the ability to access it are held entirely outside the control of the vast majority of individuals. Despite the *size* of the data set, its *quality* and *characteristics* are entirely outside the control of both the mobile start-up and the academic researcher. When researchers study Twitter data, they are accepting not only that the object of study can be found within the 140-character limitation of the tweet but that their ability to access the information and the form it takes will be controlled by the privately owned Twitter Incorporated. When a start-up builds its application using Foursquare check-in data, it is accepting that its users will access information through the venue-based data set that structures the Foursquare check-in. The remainder of this subsection charts the data chain through which those driving the base action that generates data restrict and enable what can be known and what can be done with the resulting fume or footprint.

Academic research into Foursquare check-in data provides a clear example of the danger of current big data approaches. Because Foursquare does not allow a public means of accessing its check-in results, many studies, such as Cheng et al. (2011) and Noulas et al. (2011), use Foursquare data that has been posted on Twitter. It is estimated that only 25% of Foursquare users link their accounts to Twitter (Long et al, 2012). Any check-in that is not published through Twitter is not included in the study. Although Foursquare—and some of its corporate partners—has access to this information and uses it to better target end users for advertisements, it *does not exist* for academic researchers. Further, Twitter only offers public access to 1% of its global stream (Twitter API Documentation, 2012). Using Twitter's current Streaming API,[5] tweets can be monitored for certain terms or for originating from specified locations, but not both conditions (Twitter API Documentation, 2012). To capture Foursquare check-ins, researchers monitor for terms that link to check-in information (Noulas et al., 2011). If the tweets meeting the selected conditions ever exceed 1% of the global Twitter stream, the stream is truncated at 1%, and all information past this threshold is not captured. Although Twitter previously made access to larger percentages of its global data stream available to researchers upon request, it has recently partnered with corporate resellers who monitor, store, and release information for a fee. Some researchers have linked servers to monitor larger segments of the Twitter stream through multiple accounts; however, this work-around continues to function only so long as Twitter does not change its API to prevent it.

For private companies building their applications off of data fumes, the lack of control and guaranteed access to information presents a set of similar concerns. For example, in May 2012, Foursquare announced a change in its public API to prevent applications from accessing the information of users checked into venues other than that of the end user (Thompson, 2012). An end user can still see users checked into the same venue but cannot search other venues. This change was intended to prevent the creation of applications that could be used for stalking other users, like Girls Around Me (Brownlee, 2012), but the change in access to information affected a host of unrelated applications. One interviewee had to abandon a project and begin a new one because there was no way to continue the current

[5] API stands for Application Programming Interface and is a standardized protocol through which applications may communicate with each other.

application. Assisted Serendipity, an application for finding the gender balance in bars that had previously been praised by the CEO of Foursquare, was unable to function and ceased development (Thompson, 2012). Both private companies and academic researchers run clear risks of shifting access to the information they require when relying upon the data given off by interests other than their own. Foursquare changed its API policy due to bad publicity, and Twitter drastically restricted free access to its data when it found a market willing to pay for it. In each case, the decision to change was made by a single corporation, but the repercussions were felt along the entire data chain. What can be known—the limits of research—and what can be done—the limits of capitalization—are inherently controlled by those who control access to the data; their decisions suffuse the data chain, setting the limits of footprints and fumes.

Incentives and Imperatives in Data Created for Profit

In addition to basic concerns over access, reliance upon data given off by other corporations accepts an underlying view of individual and society shaped through the incentives that drive data creation and the imperatives of the corporations that drive it. The big data research paradigm accepts that meaning can be found in ever-increasingly large data sets through an abductive praxis of exploration. In accepting this paradigm, researchers accept the existence of and ascribe meaning to a particular commodified, quantified, and standardized individual that emerges within the tangles of privately owned data. This subsection draws attention to how the processes that drive spatial big data creation may distort or be in direct conflict with the goals of researchers attempting to further societal understanding.

Although the Foursquare and Twitter data upon which much spatial big data research relies is not an exclusive source (see, e.g., Lathia et al., 2012), it represents a burgeoning field of academic research. Researchers have ascribed meaning ranging from societal well-being to neighborhood boundaries to movement patterns of individuals to the data produced from check-in data from location-based social networks (LBSNs). The use of these networks is built around a for-profit model in which end users are incentivized to participate, and the information they provide is commoditized and sold to advertisers (Thatcher, in press). This has a double effect on the commodification of data produced. On one hand, end users may manipulate the system by providing false information to receive rewards. On the other hand, the organization and form of the data produced is structured in such a way as to standardize and quantify location to allow for its commodification and exchange. Each scenario potentially distorts data in unique ways.

It has been well established that the coordinate data provided by services such as Twitter and Foursquare are not nearly as accurate as their decimal points suggest (Xu et al., 2012). For example, a Twitter user may set his or her location to Boston, Massachusetts. Although the user has specified an address to the city level, geocoding these results produces a location specified to within a single meter. While potentially misleading, this issue is well known, and researchers have proposed solutions (Hecht, Hong, Suh, & Chi, 2011). Check-in information, the basis of most LBSNs like Foursquare, avoids this potential distortion by having end users check into already-geocoded venues. To encourage the use of this feature, Foursquare and other services offer rewards for checking into certain venues. The first time a user checks into a restaurant, she or he may be offered a discount on the meal or a free drink from the bar. Similarly, a customer may receive a permanent discount after checking into a location a certain number of times. By rewarding user participation, Foursquare and other services have incentivized location fraud—to "falsely claim to be at a location, to receive undeserved rewards or social status" (Carbunar & Potharaju, p. 1). Incentivizing end users to contribute data so that that data may be sold as a commodity results in distorting end-user behavior to maximize the receipt of rewards.

For Foursquare and other LBSNs, this "fraud" doesn't matter; the data provided is still valuable, and the LBSN is able to turn a profit, so whether an individual user is actually at a venue matters less than that the individual user *claims* to be at said venue. In fact, Foursquare allows and encourages applications like Check In Take Out, which allow users to check into distant restaurants to place take-out orders (http://www.checkintakeout.com). While this has led some technology writers to repeatedly proclaim the death of the check-in (Mitchell, 2012), it can more generally be seen as part of a shift in focus of mobile applications from simply recording location and providing destinations to shaping the consumption patterns of users (Thatcher, in press; Wilson, 2012). As long as the data produced can be exchanged for a profit, the LBSN has no incentive to prevent this separation of physical location from check-in location. Where an end user is physically located is less important than the quantified data representing location.

From the perspective of end users, whether they are actually at a check-in location matters less than the rewards they receive for setting their location to the venue. For Foursquare and its partners, whether the location presented by an end user represents the user's physical location is less important than whether the information can be used to guide consumption patterns in a way that turns a profit for the businesses involved. This framework of data generation is driven by an innate profit motive—end users receive discounts, Foursquare sells data to partners, partners use the data to drive consumption—that is completely divorced from an accurate representation of physical location. Researchers who make use of this data are inherently accepting this capitalist framing for their research. Abductive exploration of big data may reveal patterns, but it reveals patterns of location as a commodity. Movement patterns found within Foursquare information reflect movement first encouraged and then shaped by motives for profit. A behavioral loop is created for both the end users and those who study them: "A person feeds in data, which is collected by an algorithm that then presents the user with choices, thus steering behavior" (Lohr, 2012, para. 19).

Along with the opportunities offered by massive spatial data sets come restrictions on how data can be accessed and on what can be known through it. Spatial big data involves data created and shaped by a motive for profit. Although this does not necessarily matter for a start-up company, it has profound repercussions for academic researchers. These repercussions parallel a series of debates that occurred within the GIS research community throughout the 1990s.

Big Data and GIS: Thoughts Toward a Critical Approach

With about 80% of all data stored by businesses and governments having a spatial component, it has been suggested that GIS scholars hold a "home field advantage" when it comes to the study of big data (Farmer & Pozdnoukhov, 2012). In addition to expertise in the handling and analysis of large, spatially referenced data sets, the debates that have occurred and that continue to occur within the GIS research community speak directly to the concerns raised in the previous sections. Concerns with the very definition of big data as an abductive science, concerns with the corporate control and regulation of access to data, and the epistemological concerns over the effects of commodification on data generation are all tied to similar debates in the field of geographic information science.

Earlier in this article big data was defined as a set of technical and methodological approaches involving the rapid aggregation and analysis of data sets that necessarily accepts a certain epistemological commitment toward the generation of knowledge—namely that of an abductive scientific method. The tension between the two parts of this definition—big data as a methodological approach and big data as an epistemological commitment—echoes an earlier debate between geographic information systems and geographic information science. In the 1990s, GIS was seen as "hardware and software technology in search of applications" (Goodchild, 1992, p. 43) with the limits of knowledge set by its major software vendors. In response, some research began to question the epistemological and ontological assumptions of GIS research (Pickles, 1993; Sheppard, 1993). The so-called technocists, who were engaged in a certain form of empirical research, were "willing to step out of [their] comfort zone" (Wright, Goodchild, & Proctor, 1997, p. 373) and consider postpositivist theoretical considerations. In the same way, those advocating "purely data-driven" (Kling & Pozdnoukhov, 2012, p. 483) approaches would do well to consider what knowledges are being privileged and foreclosed in the acceptance of an abductive science (boyd & Crawford, 2012; Burgess & Bruns, 2012).

The data sets researchers rely upon are increasingly generated through and controlled by privately held corporations, which raises distinct concerns about how what can be known has become regulated by a small set of corporate entities. Although not discussed here, data held by the government comes with a related set of concerns over access, which parallels discussions of a rising technocracy controlling knowledge production in GIS. "This technocracy is hidden in the offices of the vendors that develop the hardware and software and make the technology more generally accessible" (Obermeyer, 1995, p. 78). As GIS technology advanced, the mediation by technology receded from active consideration. Like the telephone, GIS became a "transparent" technology used without conscious consideration of the underlying technical processes (Obermeyer, 1995, p. 81). A similar situation has arisen for big data researchers: APIs standardize the process of access, structuring the data, but they also set the limits of what the data contains. Burgess and Bruns (2012) have demonstrated that the very heuristics of their analysis are shaped by the format of the data available through the API used. Foursquare's recent API change demonstrates that the content of the data, and therefore what can be known through it, is subject to regulation and alteration outside the control of researchers. The opaque process of corporate decision making serves as the hidden technocracy of big data.

Finally, the reliance upon data generated with an explicit motive for profit—both for the end user and the corporation—results in epistemological commitments not dissimilar to concerns raised with regard to the knowledges and approaches privileged by GIS use. For GIS, and now for big data, there is a need to distinguish between "empirical and technical claims about objects, practices, and institutions," (Pickles, 1995, p. 23) the discourses within which these claims, and claims to truth, are made. Big data, like GIS, accepts that a certain quantitative representation of life can stand in for its full meaning (Curry, 1997). Further, in this inscription of meaning, big data must be seen as directly producing new knowledge rather than simply revealing it. The hard work of theory ties big data directly to much longer traditions of social theory and social thought as they have engaged technology. GIS researchers have drawn productively from Benjamin (Kingsbury & Jones, 2009), Heidegger (Pickles, 1995), Foucault (Harley, 1989), and other classical social theory perspectives in deepening an understanding of the relation between the specific technological form and the knowledges produced. These works have, in turn, led to critiques and challenges that have pushed the field further (Crampton, 2003; Kitchin & Dodge, 2007). Theoretical considerations were informed by and, in turn, informed, empirical considerations as researchers stepped out of their comfort zones (Wright et al., 1997, p. 373) and engaged in the "hard work of theory" (Pickles, 1997, p. 370). Rather than let the numbers speak for themselves, big data researchers should draw from these theories to ground big data, its myths, and its practices in wider understandings of the relations between technology, modernity, and capitalism. Stripped of hyperbolic claims to absolute truth, big data must be recontextualized within socioeconomic, cultural, and political bodies of knowledge. This hard work of theory opens new knowledge projects within the realm of big data. For example, if the check-in is viewed as a form of disciplining technology—one that reports location to enmesh it more fully in capitalist exchange—then purposeful location fraud takes on new meaning as a potential form of resistance or protest.

Although there is some debate over the name, the field of GIS approached this task through what has variously been called GIS and society, qualitative GIS, and critical GIS. In most histories of the field, early "abrasive exchanges" between social theorists and technical practitioners are followed by "more thoughtful and considered engagements" (O'Sullivan, 2006, p. 784). Without implying that technical and empirical works are not critical, this article suggests the ongoing development of a critical big data movement. This article and others—such as Batty (2012), boyd and Crawford (2012), and Burgess and Bruns (2012)—have begun to problematize the issues surrounding big data; this process must continue. Further, critical big data has the opportunity to draw from and be inspired by the work done by scholars in the GIS community, enjoying their home field advantage not only in dealing with large, spatially referenced data sets but understanding technology as part and parcel of larger societal trends. Rather than accepting big data as a neutral tool (Heidegger, 1977), critical big data must directly engage it as both science *and* system with its own forms of knowledge production and social imperatives.

Conclusion

This article began by defining big data from both technical and theoretical perspectives. From a technical perspective, although the actual size of data that constitutes "big" will remain ever shifting, big data means the ability to rapidly aggregate and analyze data sets of size far beyond what current technical systems are prepared to handle. From an epistemological perspective, big data research requires a certain investing of meaning into quantified data sets and acceptance of an abductive scientific method where meaning emerges through exploration. For spatial big data sets, private industry and academic researchers rely upon the same type of information: data given off by actions end users already take, called digital footprints and data fumes. Two interrelated concerns were raised with the reliance on this type of big data. First, it places access to data in the hands of a very small group of corporate entities. These for-profit industries control the limits of what can be known through and done with spatial big data. Second, by accepting this data as meaningful, by researching it for insight into society, researchers tacitly accept not only a quantified measuring of life but one that is wholly commodified: End users are financially incentivized to generate data so that the data itself may then be sold. Discussed were the epistemological ramifications of relying upon a commodified representation of location as well as how the accuracy of the information may be distorted through financial incentives. Finally, the concerns raised were charted directly to discussions in the GIS community over the past 20 years. Big data needs a continued critical engagement with the assumptions, limits, and privileges associated with its specific generation of knowledge and research agenda.

This article has problematized a set of issues within big data research and suggested one source of potential guidance in addressing and understanding these issues. Against an ever-burgeoning field of statistical techniques and methodological improvements, it has suggested that academics recontextualize spatial big data by turning to the social theoretical work in the critical GIS tradition. In highlighting and focusing upon the ties between academic research and for-profit business, other aspects of big data have been left unaddressed. End users' agency and experience are unexamined, as are potential state actors' uses of big data for surveillance and control.

Big data research continues to move forward. For example, the abductive power of big data has been highlighted in terms of the National Security Agency's security programs (Crampton, 2013), and the difference between the geolocation of a tweet and the relational experience of its content has been highlighted in terms of the Occupy Wall Street movement (Eckert & Hemsley, 2013). New quantitative methods have been examined at with a critical eye to complementary qualitative ones (DeLyser & Sui, 2013). These and other works embrace both the technical power of spatial big data and theory-based understandings of its limits. As big data's promise and power continue to be touted (Mayer-Schonberger & Cukier, 2013; Siegel, 2013), this work must continue.

References

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Baker, @. (2012, January 5). Can social media sell soap? *The New York Times*. Retrieved from http://www.nytimes.com/2013/01/06/opinion/sunday/can-social-media-sell-soap.html?hp&_r=1&

Batty, M. (2012). Smart cities, big data. *Environment and Planning B, 39,* 191–193.

Benner, J., & Robles, C. (2012). Trending on Foursquare: Examining the location and categories of venues that trend in three cities. In *Proceedings of the Workshop on GIScience in the Big Data Age 2012* (pp. 27–35). Columbus, Ohio.

Berry, D. M. (2011). *The philosophy of software: Code and mediation in the digital age.* London, UK: Palgrave Macmillan.

boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679.

Brownlee, J. (2012, March 30). This creepy app isn't just stalking women without their knowledge, it's a wake-up call about Facebook privacy. *Cult of Mac*. Retrieved from http://www.cultofmac.com/157641/this-creepy-app-isnt-just-stalking-women-without-their-knowledge-its-a-wake-up-call-about-facebook-privacy/

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "big social data" for media and communication research. *M/C* Journal, *15*(5). Retrieved from http://www.journal.media-culture.org.au/index.php/mcjournal/article/view/561

Carbunar, B., & Potharaju, R. (2012). You unlocked the Mt. Everest badge on Foursquare! Countering location fraud in geosocial networks. In *Proceedings of the 2012 IEEE 9th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)* (MASS '12) (pp. 182-190). IEEE Computer Society, Washington, DC.

Cerrato, P. (2012, November 1). Big data analytics: Where's the ROI? *InformationWeek: Healthcare*. Retrieved from http://www.informationweek.com/healthcare/clinical-systems/big-data-analytics-wheres-the-roi/240012701

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. (2011). Exploring millions of footprints in location sharing services. In *Proceedings of the Fifth International AAAI Conference on WSM.* Barcelona, Spain. Retrieved from http://www.aaai.org/ocs/index.php/IC WSM/ICWSM11/paper/view/2783

Crampton, J. (2003). *The political mapping of cyberspace.* Edinburgh, Scotland: Edinburgh University Press.

Crampton, J. (2013). Commentary: Is security sustainable? *Environment and Planning D*, *31,* 571–577.

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The Livehoods Project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on WSM*. Dublin, Ireland.

Curry, M. (1997). The digital individual and the private realm. *Annals of the AAG*, *87,* 681–699.

DeLyser, D., & Sui, D. (2013). Crossing the qualitative-quantitative divide II: Inventive approaches to big data, mobile methods, and rhythmanalysis. *Progress in Human Geography*, *37*(2), 293–305.

Eckert, J., & Hemsley, J. (2013, April 11). *Occupied geographies, relational or otherwise.* Presentation to the American Association of Geographers, Los Angeles, CA.

Exner, J., Zeile, P., & Streich, B. (2011). Urban monitoring laboratory: New benefits and potential for urban planning through the use of urban sensing, geo- and mobile-web. In *Real CORP Proceedings 2011* (pp. 1087–1096). Wien, Austria.

Farmer, C., & Pozdnoukhov, A. (2012). Building streaming GIScience from context, theory, and intelligence. In *Proceedings of the Workshop on GIScience in the Big Data Age 2012* (pp. 5–10). Columbus, Ohio.

Goodchild, M. (1992). Geographical information science. *International Journal of Geographical Information Systems*, *6*, 31–45.

Goodchild, M. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal, 69*(4), 211–221.

Goodchild, M., & Glennon, J. A. (2010). Crowdsourcing geographic information for disaster response: A research frontier. *International Journal of Digital Earth, 3*(3), 231–241.

Harley, J. (1989). Deconstructing the map. *Cartographical*, *26*, 1–20.

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart. In *Proceedings of the ACM CHI Conference 2011* (pp. 237–246). Vancouver, BC.

Heidegger, M. (1977). *The question concerning technology and other essays* (W. Lovitt, Trans.). New York, NY: Harper Perennial.

Hey, T., Tansley, S., & Toelle, K. (Eds.). (2009). *The fourth paradigm: Data-intensive scientific discovery*. Richmond, WA: Microsoft Research.

Horvath, I. (2012). Beyond advanced mechatronics: New design challenges of social-cyber systems. (Draft paper.) *Proceedings of the ACM Workshop on Mechatronic Design*, *Linz 2012*. Linz, Austria

Jacobs, A. (2009). The pathologies of big data. *ACM Queue*, *7*(6), 1–12.

Joseph, K., Tan, C., & Carley, K. (2012). Beyond "local," "categories" and "friends": Clustering Foursquare users with latent "topics." In *Proceedings of ACM Ubicomp '12* (pp. 919–926). Pittsburgh, PA.

Kingsbury, P., & Jones III, J. P. (2009). Walter Benjamin's Dionysian adventures on Google earth. *Geoforum,* 40, 502–513.

Kitchin, R., & Dodge, M. (2007). Rethinking maps. *Progress in Human Geography*, *31*, 331–344.

Kling, F., & Pozdnoukhov, A. (2012). When a city tells a story: Urban topic analysis. In *Proceedings of ACM SIGSPATIAL 2012* (pp. 482–485). Redondo Beach, CA.

Lathia, N., Quercia, D., & Crowcroft, J. (2012). The hidden image of the city: Sensing community well-being from urban mobility. *Pervasive Computing*, *7319*, 91–98.

Laurila, J., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T., Dousse, O., Eberle, J., & Miettinen, M. (2012). The mobile big data challenge. *Nokia Research*. Retrieved from http://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf

Livehoods. (2012). Retrieved from http://livehoods.org

Lohr, S. (2012, December 29). Sure, big data is great. But so is intuition. *The New York Times*. Retrieved from http://www.nytimes.com/2012/12/30/technology/big-data-is-great-but-dont-forget-intuition.html?_r=3&adxnnl=1&partner=rss&emc=rss&adxnnlx=1357590814-L6vdMVi0JnNF0dB5hk1KLg&

Long, X., Jin, L., & Joshi, J. (2012). Exploring trajectory-driven local geographic topics in Foursquare. In *Proceedings of ACM Ubicomp '12* (pp. 927–934). Pittsburgh, PA.

Marcuse, H. (1982 [1941]). Some social implications of modern technology. In A. Arato & E. Gebhardt (Eds.), *The essential Frankfurt School reader* (pp. 138–162). New York, NY: Continuum.

Martino, M., Britter, R., Outram, C., Zacharias, C., Biderman, A., & Ratti, C. (2010). *Senseable city*. Cambridge, MA: MIT Senseable City Lab.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

Mitchell, J. (2012, April 10). Life after death of the check-in. *ReadWrite*. Retrieved from http://readwrite.com/2012/04/10/pronouncing_the_death_of_the_check-in

National Science Foundation. (2012, October 3). *NSF announces interagency progress on administration's big data initiative* (Press release). Retrieved from http://www.nsf.gov/news/news_summ.jsp?cntn_id=125610

Noulas, A., Scellato, S., Mascolo, C., & Pontil, M. (2011). An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the Fifth International AAAI Conference on WSM* (pp. 570–573). Barcelona, Spain.

Obermeyer, N. (1995). The hidden GIS technocracy. *Cartography and Geographic Information Science*, *22*(1), 78–83.

O'Sullivan, D. (2006). Geographical information science: Critical GIS. *Progress in Human Geography*, *30*(6), 783–791.

Paulos, E., Honicky, R. J., & Hooker, B. (2008). Citizen science: Enabling participatory urbanism. In M. Foth (Ed.), *Handbook of research on urban informatics: The practice and promise of the real-time city* (pp. 414–436). Hershey, PA: Information Science Reference.

Pickles, J. (1993). Discourse on method and the history of discipline: Reflections on Jerome Dobson's 1993 "Automated geography." *Professional Geographer*, *45*, 451–455.

Pickles, J. (1995). *Ground truth*. New York, NY: Guilford Press.

Pickles, J. (1997). Tool or science? GIS, technoscience and the theoretical turn. *Annals of the AAG*, *87*, 363–372.

Presley, S. (2011). Mapping out #LondonRiots. *NFPvoice*. Retrieved from http://nfpvoice.com/?p=1348

Rasmus, D. (2012, January 27). Why big data won't make you smart, rich, or pretty. *Fast Company*. Retrieved from http://www.fastcompany.com/1811441/why-big-data-won%E2%80%99t-make-you-smart-rich-or-pretty

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 851–860). Raleigh, NC.

A sea of sensors. (2010, November 4). *The Economist*. Retrieved from http://www.economist.com/node/17388356

Sheppard, E. (1993). Automated geography: What kind of geography for what kind of society? *Professional Geographer*, *45*, 457–460.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken, NJ: John Wiley.

Sui, D. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems, 32*, 1–5.

Taylor, C. (2012, November 7). Triumph of the nerds: Nate Silver wins in 50 states. *Mashable.* Retrieved from http://mashable.com/2012/11/07/nate-silver-wins

Thatcher, J. (in press). Avoiding the ghetto through hope and fear: An analysis of immanent technology using ideal types. *GeoJournal*.

Thompson, C. (2012, May 10). Foursquare alters API to eliminate apps like Girls Around Me. *AboutFoursquare*. Retrieved from http://aboutfoursquare.com/foursquare-api-change-girls-around-me

Twitter API Documentation. (2012). *Streaming API request parameters*. Retrieved from https://dev.twitter.com/docs/streaming-apis/parameters

Weber, M. (1973 [1946]). *From Max Weber* (C. Mills & H. Gerth, Eds.). New York, NY: Oxford University Press.

Wilson, M. (2012). Location-based services, conspicuous mobility, and the location-aware future. *Geoforum*, *43*(6), 1266–1275.

Wright, D., Goodchild, M., & Proctor, D. (1997). Still hoping to turn that theoretical corner. *Annals of the AAG*, *87*(2), 373.

Xu, S., Flexner, S., & Carvalho, V. (2012). Geocoding billions of addresses: Towards a spatial record linkage system with big data. In *Proceedings of the Workshop on GIScience in the Big Data Age 2012* (pp. 17–26). Columbus, Ohio.