Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery

Piyushimita (Vonu) Thakuriah, University of Glasgow
Nebiyou Tilahun, University of Illinois at Chicago
Moira Zellner, University of Illinois at Chicago

Please cite as:
Thakuriah, P., N. Tilahun and M. Zellner (2015). Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In Proc. of NSF Workshop on Big Data and Urban Informatics, pp. 4-32.

Forthcoming as:
Thakuriah, P., N. Tilahun and M. Zellner (estimated Feb 2016). Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. In *Seeing Cities through Big Data: Research Methods and Applications in Urban Informatics*, to be published by Springer, Consisting of papers presented in an NSF-funded Workshop on Big Data and Urban Informatics.

Abstract

Big Data is the term being used to describe a wide spectrum of observational or "naturally-occurring" data generated through transactional, operational, planning and social activities that are not specifically designed for research. Due to the structure and access conditions associated with such data, research and analysis using such data becomes significantly complicated. New sources of Big Data are rapidly emerging as a result of technological, institutional, social, and business innovations. The objective of this background paper is to describe emerging sources of Big Data, their use in urban research, and the challenges that arise with their use. To a certain extent, Big Data in the urban context has become narrowly associated with sensor (e.g., Internet of Things) or socially generated (e.g., social media or citizen science) data. However, there are many other sources of observational data that are meaningful to different groups of urban researchers and user communities. Examples include privately held transactions data, confidential administrative micro-data, data from arts and humanities collections, and hybrid data consisting of synthetic or linked data.

The emerging area of Urban Informatics focuses on the exploration and understanding of urban systems by leveraging novel sources of data. The major potential of Urban Informatics research and applications is in four areas: (1) improved strategies for dynamic urban resource management, (2) theoretical insights and knowledge discovery of urban patterns and processes, (3) strategies for urban engagement and civic participation, and (4) innovations in urban management, and planning and policy analysis. Urban Informatics utilizes urban Big Data in innovative ways by retrofitting or repurposing existing urban models and simulations that are underpinned by a wide range of theoretical traditions, as well as through data-driven modeling approaches that are largely theory agnostic, although these divergent research approaches are starting to converge in some ways. The paper surveys the kinds of urban problems being considered by going from a data-poor environment to a data-rich world and ways in which such enquiry have the potential to enhance our understanding, not only of urban systems and processes overall, but also contextual peculiarities and local experiences. The paper concludes by commenting on challenges that are likely to arise in varying degrees when using Big Data for Urban Informatics: technological, methodological, theoretical/epistemological, and the emerging political economy of Big Data.

Keywords: Big Data, Urban Informatics, Knowledge Discovery, dynamic resource management, user-generated content

1. Introduction

Urban and regional analysis involve the use of a wide range of approaches to understand and manage complex sectors, such as transportation, environment, health, housing, the built environment, and the economy. The goals of urban research are many, and include theoretical understanding of infrastructural, physical and socioeconomic systems; developing approaches to improve urban operations and management; long-range plan making, and impact assessments of urban policy.

Globally, more people live in urban areas than in rural areas, with 54% of the world's population estimated to be residing in urban areas in 2014 (United Nations, 2014), levying unprecedented demand for resources and leading to significant concerns for urban management. Decision-makers face a myriad of questions as a result, including: What strategies are needed to operate cities effectively and efficiently? How can we evaluate potential consequences of complex social policy change? What makes the economy resilient and strong and how do we develop shockproof cities? How do different cities recover from man-made or natural disasters? What are the technological, social and policy mechanisms needed to develop interventions for healthy and sustainable behavior? What strategies are needed for lifelong learning, civic engagement, and community participation, adaptation and innovation? How can we generate hypothesis about the historical evolution of social exclusion and the role of agents, policies and practices?

The Big Data tsunami has hit the urban research disciplines just like many other disciplines. It has also stimulated the interest of practitioners and decision-makers seeking solutions for governance, planning and operations of multiple urban sectors. The objective of this background paper is to survey the use of Big Data in the urban context across different academic and professional communities, with a particular focus on Urban Informatics. Urban Informatics is the exploration and understanding of urban systems for dynamic resource management, knowledge discovery and understanding of urban patterns and dynamics, urban engagement and civic participation, and urban planning and policy analysis. Urban Informatics research approaches involve both a theory-driven as well as an empirical data-driven perspective centered on emerging Big Data sources. New sources of such data are arising as a result of technological, institutional, social and business innovations, dramatically increasing possibilities for urban researchers. Equally importantly, new ways of accessing existing sources of data, or innovations in the linkage of data belonging to different owners and domains are leading to new connected data systems. We identify major research questions that may be possible to investigate with the data, as well as existing questions that can be revisited with improved data, in an attempt to identify broad themes for the use of Big Data in Urban Informatics.

While the main research agenda is about better understanding and knowledge discovery of urban systems, there are equally important questions relating to technical challenges in managing the data and in addressing methodological and measurement questions that arise. The use of Big Data in Urban Informatics pose significant epistemological challenges regarding the overall modes of research inquiry, and about institutions and the overall political economy regarding the access and use.

This chapter is organized as follows: in Section 2, we review background information and different types of Big Data being used for urban research. This is followed in Section 3 by a discussion of research approaches and applications in Urban Informatics that involve the use of Big Data. Challenges that arise with the use of such data are discussed next in Section 4 and conclusions are drawn in Section 5.

2. Big Data: Complexities and Types

For many, Big Data is just a buzzword and to a certain extent, the ambiguity in its meaning reflects the different ways in which it is used in different disciplines and user communities. The ambiguity is further perpetuated by the multiple concepts that have become associated with the topic. However, the vagueness and well-worn clichés surrounding the subject have overshadowed potentially strong benefits in well-considered cases of use.

Based on a review of 1,437 conference papers and articles that contained the full term "Big Data" in either the title or within the author-provided keywords, De Mauro et al. (2014) arrived at four groups of definitions of Big Data. These definitions focus on: (1) the characteristics of Big Data (massive, rapid, complex, unstructured and so on), with the 3-Vs - Volume, Variety and Velocity - referring to the pure amount of information and challenges it poses (Laney, 2001) being a particularly over-hyped

example; (2) the technological needs behind the processing of large amounts of data (e.g., as needing serious computing power, or, scalable architecture for efficient storage, manipulation, and analysis); (3) as Big Data being associated with crossing of some sort of threshold (e.g., exceeding the processing capacity of conventional database systems); and (4) as highlighting the impact of Big Data advancement on society (e.g., shifts in the way we analyze information that transform how we understand and organize society).

Moreover, the term Big Data has also come to be associated with not just the data itself, but with curiosity and goal-driven approaches to extract information out of the data (Davenport and Patil, 2012), with a focus on the automation of the entire scientific process, from data capture to processing to modeling (Pietsch, 2013). This is partly an outcome of the close association between Big Data and data science, which emphasizes data-driven modeling, hypothesis generation and data description in a visually appealing manner. These are elements of what has become known as the Fourth Paradigm of scientific discovery (Gray, 2007 as given in Hey et al., 2009), which focuses on exploratory, data-intensive research, in contrast to earlier research paradigms focusing on describing, theory-building and computationally simulating observed phenomenon.

Quantitative urban research has historically relied on data from censuses, surveys, and specialized sensor systems. While these sources of data will continue to play a vital role in urban analysis, declining response rates to traditional surveys, and increasing costs of administering the decennial census and maintaining and replacing sensor systems have led to significant challenges to having high-quality data for urban research, planning and operations. These challenges have led to increasing interest in looking at alternative ways of supplementing the urban data infrastructure.

For our purposes, Big Data refers to structured and unstructured data generated naturally as a part transactional, operational, planning and social activities, or the linkage of such data to purposefully designed data. The use of such data gives rise to technological and methodological challenges and complexities regarding the scientific paradigm and political economy supporting inquiry. Established and emerging sources of urban Big Data are summarized in Table 1: sensor systems, user-generated content, administrative data (open and confidential micro-data), private sector transactions data, data from arts and humanities collections, and hybrid data sources, including linked data and synthetic data. While there are many ways to organize Big Data for urban research and applications, the grouping here is primarily informed by the user community typically associated with each type of data; other factors such as methods of generation, and issues of ownership and access are also considered. The grouping is not mutually exclusive; for example, sensor systems might be owned by public agencies for administrative and operational purposes as well as by private companies to assist with transactions.

Table 1: Types of Urban Big Data and Illustrative User Communities

Urban Big Data	Examples	Illustrative User Communities
Sensor systems (infrastructure-based or moving object sensors)	Environmental, water, transportation, building management sensor systems; connected systems; Internet of Things	Public and private urban operations and management organizations, independent ICT developers, researchers in the engineering sciences
User-Generated Content ("social" or "human" sensors)	Participatory sensing systems, citizen science projects, social media, web use, GPS, online social networks and other socially generated data	Private businesses, customer/client-focused public organizations, independent developers, researchers in data sciences and urban social sciences
Administrative (governmental) data (open and confidential micro-data)	Open administrative data on transactions, taxes and revenue, payments and registrations; confidential person-level micro-data on employment, health, welfare payments, education records	Open data: innovators, civic hackers, researchers Confidential data: government data agencies, urban social scientists involved in economic and social policy research, public health and medical researchers
Private Sector Data (customer and transactions records)	Customer transactions data from store cards and business records; fleet management systems; customer profile data from application forms; usage data from utilities and financial institutions; product purchases and terms of service agreements	Private businesses, public agencies, independent developers, researchers in data sciences and urban social sciences
Arts and Humanities Data	Repositories of text, images, sound recordings, linguistic data, film, art and material culture, and digital objects, and other media	Urban design community, historical, art, architecture and digital humanities organizations, community organizations, data scientists and developers, private organizations
Hybrid data (linked and synthetic data)	Linked data including survey-sensor, census-administrative records	Urban planning and social policy community, government data organizations, private businesses and consultants

2.2.1 Sensor Systems: Infrastructure and Moving Object Sensors and Internet of Things

Sensors in urban infrastructure (transportation, health, energy, water, waste, weather systems, structural health monitoring systems, environmental management, buildings and so on) result in vast amounts of data on urban systems. Novel patterns of demand and usage patterns can be extracted from these data. The sensors detect activity and changes in a wide variety of urban phenomena involving

inanimate objects (infrastructure, building structure), physical aspects of urban areas (land cover, water, tree cover, and atmospheric conditions), movement (of cars, people, animals), and activity (use patterns, locations).

As noted earlier, sensor systems might be government or privately owned, with very different access and data governance conditions, and some have been operational for a long time. Typical user communities are public and private urban operations management organizations, independent ICT developers, and researchers in the engineering sciences. However, sensor data, if linked to other sources and archived over long periods of time, can be used by urban social scientists studying long-term social, economic and environmental changes, and their effects on neighborhoods and communities. The emerging smart cities community has become increasingly involved with urban sensor systems, particularly with their integration and performance enhancement through ICT solutions. Many urban sensor systems are now likely to be wirelessly connected, mobile, and significantly more embedded and distributed. Examples from a vast range of operational and planned applications include cooperative or connected vehicle systems, Vehicle-to-Grid systems, Smart Grid systems, and a wide range of indoor and outdoor assistive technologies for seniors and persons with disabilities. The diverse field of remote sensing has been undergoing rapid developments as well, with massive amounts of high-resolution temporal and spatial data being collected more rapidly than ever before with sensors that are mounted on satellites, planes, and lately, drones.

Potentially the "next phase" in this ever-transforming technology landscape is the creation of tiny, intelligent devices that are embedded into everyday objects such as houses, cars, furniture, and clothes, and which can "listen in" and produce recommendations and interventions as needed. The concept of the "Internet of Things" (IoT) attributed to Ashton (1999) is still primarily a vision at this stage, although there are many individual IoT technologies and systems that are operational, although a future with Machine-to-Machine (M2M) communications is envisioned by some, where "billions to trillions of everyday objects and the surrounding environment are connected and managed through a range of devices, communication networks, and cloud-based servers" (Wu, 2011). Needless to say, the number and variety of data streams available to study cities will greatly increase.

2.2.2 User-Generated Content: Social and Human Sensing Systems

Transformative changes have taken place in the last decade regarding ways in which citizens are being involved in co-creating information, and much has been written about crowd-sourcing, Volunteered Geographic Information, and, generally, User-Generated Content (UGC). Citizens, through the use of sensors or social media, and other socially generated information resulting from their participation in social, economic or civic activities, are going from being passive subjects of survey and research studies to being active generators of information. Citizen-based approaches can be categorized as contributory, collaborative, or co-created (Bonney et al., 2009). UGC can generally occur: (1) proactively when users voluntarily generate data on ideas, solve problems, and report on events, disruptions or activities that are of social and civic interest, or (2) retroactively, when analysts process secondary sources of user-submitted data published through the web, social media and other tools (Thakuriah and Geers, 2013).

UGC can be proactively generated through idea generation, feedback and problem solving. Developments in Information and Communications Technology (ICT) have expanded the range and diversity of ways in which citizens provide input into urban planning and design sourcing, vote on and share ideas about urban projects, and provide feedback regarding plans and proposals with the potential to affect life in cities. These range from specialized focus groups where citizens provide input to "hackathons" where individuals passionate about ICT and cities get together to generate solutions to civic problems using data. Citizens also solve problems; for example, through human computation (described further in Section 3.4) to assess livability or the quality of urban spaces where objective metrics from sensors and machines are not accurate. These activities produce large volumes

of structured and unstructured data that can be analyzed to obtain insights into preferences, behaviors and so on.

There has been an explosive growth in the wealth of data proactively generated through different sensing systems. Depending on the level of decision-making needed on the part of users generating information, proactive sensing modes can be disaggregated into participatory (active) and opportunistic (passive) sensing modes. In participatory sensing, users voluntarily opt into systems that are specifically designed to collect information of interest (e.g., through apps which capture information on quality of local social, retail and commercial services, or websites that consolidate information for local ride-sharing), and actively report or upload information on objects of interest. In opportunistic sensing, users enable their wearable or in-vehicle location-aware devices to automatically track and passively transmit their physical sensations, or activities and movements (e.g., real-time automotive tracking applications which measure vehicle movements yielding data on speeds, congestion, incidents and the like, as well as biometric sensors, life loggers and a wide variety of other devices for personal informatics relating to health and well-being). The result of these sensing programs are streams of content including text, images, video, sound, GPS trajectories, physiological signals and others, which are available to researchers at varying levels of granularity depending on, among other factors, the need to protect personally identifiable information.

In terms of retroactive UCG, massive volumes of content are also created every second of every day as a result of users providing information online about their lives and their experiences. The key difference from the proactive modes is that users are not voluntarily opting into specific systems for the purpose of sharing information on particular topics and issues. There are many different types of retroactive UGC that could be used for urban research including Internet search terms, customer ratings, web usage data, and trends data. Data from social networks, micro-blogs or social media streams have generated a lot of interest among researchers, with the dominant services at the current time being online social networks such as Twitter, Facebook and LinkedIn, and Foursquare, the latter being a location-based social network. Additionally, there are general question-and-answer databases from which data relevant to urban researchers could be retrieved, as well as a wide variety of multimedia online social sharing platforms such as YouTube and Flickr, and user-created online content sites such Wikipedia, TripAdvisor and Yelp. Such naturally occurring UGC provide a rich source of secondary data on the social fabric of cities, albeit through the lens of their user communities, raising questions regarding bias and lack of generalizability. However, provided appropriate information retrieval and analytics techniques are used, such data can allow detection and monitoring of events and patterns of interest, as well as the ability to identify concerns, emotions and preferences among citizens, particularly in response to news, urban operation disruptions and policy changes, for real-time understanding of urban dynamics.

2.2.3 Administrative Data

Governments collect micro-data on citizens as a part of their everyday business or operational processes on registration, transactions and record keeping which typically occur during the delivery of a service. Tax and revenue agencies record data on citizens and taxes paid revenues generated, licenses issued and real estate or vehicle transactions made. Employment and benefits agencies collect information on income, earnings and disability or retirement benefits. Administrative micro-data in particular contain a wealth of information that is relevant to urban policy evaluation. The advantages often cited regarding the use of administrative data in research include being relatively cheap and potentially less intrusive and yet comprehensive (Gowans et al., 2015), as well as having larger sample sizes, and fewer problems with attrition, non-response, and measurement error compared to traditional survey data sources (Card et al., n.d.).

One particular development with administrative data is the increasing availability of administrative and other governmental data through "Open Data" initiatives. These initiatives have largely been driven by open government strategies, generally thought of as reflecting transparent government, collaborative government, and innovative government, with some degree of confusion and ideological tensions about what these terms mean in practice (Shkabatur, 2013; Pyrozhenko, 2011). Open data initiatives are based on the idea that governmental data should be accessible for everyone to use and to republish without copyright or other restrictions in order to create a knowledgeable, engaged, creative citizenry, while also bringing about accountability and transparency. Open Data initiatives have the potential to lead to innovations (Thorhildur et al., 2013) and to address the needs of the disadvantaged (Gurstein, 2011).

National and local governments around the world have now supported open data policies. This has led to a proliferation of open data portals where government agencies upload administrative data that are aggregated or anonymized by removing personal identifiers, and is license-free and in non-proprietary formats. Although they present many opportunities, open data initiatives can face challenges due to a number of reasons including closed government culture in some localities, privacy legislation, limitations in data quality that prohibit publication, and limited user-friendliness (Huijboom and van den Broek, 2011).

Many valuable uses of administrative data require access to personally identifiable information, typically micro-data at the level of individual persons, which are usually strictly protected by data protection laws or other governance mechanisms. Personally-identifiable information are those that can be used on its own or together with other information to identify a specific individual, and the benefits of accessing and sharing identifiable administrative data for research purposes have to be balanced against the requirements for data security to ensure the protection of individuals' personal information. Confidential administrative micro-data are of great interest to urban social scientists involved in economic and social policy research, as well as to public health and medical researchers.

There are several activities currently that are likely to be of interest to urban researchers. The UK Economic and Social Research Council recently funded four large centers on administrative data research, including running data services to support confidential administrative data linkage, in a manner similar to that offered in other countries such as Denmark, Finland, Norway and Sweden. In the US, the Longitudinal Employment Household Dynamics (LEHD) program of the Census Bureau is an example of an ambitious nationwide program combining federal, state and Census Bureau data on employers and employees from unemployment insurance records, data on employment and wages, additional administrative data and data from censuses and surveys (Abowd et al., 2005), to create detailed estimates of workforce and labor market dynamics.

Administrative data in some cases can be linked both longitudinally for the same person over time and between registers of different types, e.g. linking employment data of parents to children's test scores, or linking medical records to person's historical location data and other environmental data. The latter, for example, could potentially allow research to investigate questions relating to epigenetics and disease heritability (following Aguilera et al., 2010). Such linkages are also likely to allow in-depth exploration of spatial and temporal variations in health and social exclusion.

2.2.4 Private Sector Transaction Data

Like government agencies, businesses collect data as a part of their everyday transactions with customers. They also develop detailed customer profiles from different sources. Such privately held data may be contrasted with the aforementioned privately owned sensor systems data as those that continuously track customer activity and use patterns. In a report titled "New Data for Understanding the Human Condition: International Perspectives" (OECD Global Science Forum, 2013), customer transactions was identified as a major data category, within which the following were noted as useful

data sources: store cards such as supermarket loyalty cards, customer accounts on utilities, financial institutions, and other customer records such as product purchases and service agreements.

Companies have historically used such data to improve business process management, market forecasts and to improve customer relations. Many of these data sources can provide key insights into challenges facing cities and have been increasingly of interest to urban researchers. For example, utility companies have records on energy consumption and transactions, which can help to understand variations in energy demand and impact for sustainable development policy, while also understanding implications for fuel poverty where households spend more than some acceptable threshold to maintain adequate heating (NAREC, 2013).

2.2.5 Arts and Humanities Collections and Historical Urban Data

There are vast arts and humanities collections that depict life in the city that include text, image, sound recording, and linguistic collections, as well as media repositories such as film, art, material culture, and digital objects. These highly unstructured sources of data allow the representation of the ocular, acoustic and other patterns and transformations in cities to be mapped and visualized, to potentially shed light on social, cultural and built environment patterns in cities. For example, a recent project seeks to digitize a treasure trove of everyday objects such as "advertisements, handbills, pamphlets, menus, invitations, medals, pins, buttons, badges, three-dimensional souvenirs and printed textiles, such as ribbons and sashes" to provide "visual and material insight into New Yorkers' engagement with the social, creative, civic, political, and physical dynamics of the city, from the Colonial era to the present day" (Museum of the City of New York, 2014), which will have detailed metadata making it searchable through geographic querying.

Inferring knowledge from such data involves digitization, exploratory media analysis, text and cultural landscape mapping, 3-D mapping, electronic literary analysis, and advanced visualization techniques. With online publishing and virtual archives, content creators and users have the potential to interact with source materials to create new findings, while also facilitating civic engagement, community building and information sharing. Recent focus has been on humanities to foster civic engagement; for example, the American Academy of Arts and Sciences (2013), while making a case for federal funding for the public humanities, emphasized the need to encourage "civic vigor" and to prepare citizens to be "voters, jurors, and consumers". There is potential for this line of work in improving the well-being of cities by going beyond civic engagement, for example, to lifelong learning (Hoadley and Bell, 1996; CERI/OECD, 1992). Stakeholders engaged in this area are typically organizations involved in cultural heritage and digital culture, such as museums, galleries, memory institutions, libraries, archives and institutions of learning. Typical user communities for this type of data are history, urban design, art and architecture, and digital humanities organizations, as well as community and civic organizations, data scientists, and private organizations. The use of such data in quantitative urban modeling opens up a whole new direction of urban research.

2.2.6 Hybrid Data and Linked Data Systems

Data combinations can occur in two ways: combination through study design to collect structured and unstructured data during the same data collection effort (e.g., obtaining GPS data from social survey participants, so that detailed movement data are collected from the persons for whom survey responses are available), or through a combination of different data sources brought together data by data linkage or multi-sensor data fusion under the overall banner of what has recently been called "broad data" (Hendler, 2014).

There are now several examples where data streams have been linked by design, an example of which is household travel surveys and activity diaries have been administered using both questionnaire-based survey instrument and a GPS element. One of many examples is the 2007/2008 Travel Tracker data collection by the Chicago Metropolitan Agency for Planning (CMAP), which included travel

diaries collected via computer assisted telephone interviews (CATI) and GPS data collected from a subset of participants over 7 days. Recent efforts have expanded the number of sensing devices used and the types of contextual data collected during the survey period. For example, the Integrated Multimedia City Data (iMCD) (Urban Big Data Center, n.d.), which is being administered at the time of writing this paper, involves a questionnaire-based survey covering travel, ICT use, education and literacy, civic and community engagement, and sustainable behavior of a random sample of households in Glasgow, UK. Respondents undertake a sensing survey using GPS and life logging sensors leading to location and mobility data and rapid still images of the world as the survey respondent sees it. In the survey background is a significant Information Retrieval effort from numerous social media and multimedia web sources, as well as retrieval of information from transport, weather, crime-monitoring CCTV and other urban sectors. Alongside these data streams are Very High Resolution satellite data and LiDAR allowing digital surface modeling creating 3D urban representations.

The census is the backbone for many types of urban analysis; however, its escalating costs has been noted to be unsustainable, with the cost of the 2010 US Census being almost $94 per housing unit, representing a 34% increase in the cost per housing unit over Census 2000 costs, which in turn represents a 76% increase over the costs of the 1990 Census (Reist and Ciango, 2013). There was an estimated net undercount of 2.07% for Blacks, 1.54% for Hispanics, and 4.88% for American Indians and Alaska Natives, while non-Hispanic whites had a net over-count of 0.83 percent (Williams, 2012). Vitrano and Chapin (2014) estimated that without significant intervention, the 2020 Census would cost about $151 per household. This has led the US Census Bureau to actively consider innovative solutions designed to reduce costs while maintaining a high quality census in 2020. Some of the strategies being considered include leveraging the Internet and new methods of communications to improve self-response by driving respondents to the Internet and taking advantage of Internet response processes. Another census hybridization step being considered is the use of administrative records to reduce or eliminate some interviews of households that do not respond to the census and related field contacts.

Similar concerns in the UK led to the Beyond 2011 program where different approaches to produce population statistics were considered. The program recommended several potential approaches such as the use of an online survey for the decennial census and a census using existing government data and annual compulsory surveys (Office for National Statistics, 2015). The ONS Big Data project is also evaluating through a series of pilot projects the possibility of using web scraping of Internet price data for the Consumer Price Index (CPI) and the Retail Price Index (RPI) and Twitter data to infer student movement, which is a population that has been historically been difficult to capture through traditional surveys (Naylor et al., 2015). Other Big Data sources being studied as a part of the pilots are smart meter data to identify household size/structure and the likelihood of occupancy during the day, and mobile phone positioning data to infer travel patterns of workers.

Another situation is where data on survey respondents are linked to routine administrative records; one approach involved the use of an informed consent process where respondents who agree to participate in a survey are explicitly asked if the information they provide can be linked to their administrative records. One example of this approach is the UK Biobank Survey (Lightfoot and Dibben, 2013). Having survey responses linked to administrative data enables important urban policy questions to be evaluated; the key issue here is that participants understand and agree to such linkage.

From urban operations point of view, connected systems allow a degree of sophistication and efficiency not possible with data from individual data systems. This was touched upon briefly in Section 2.2.1; clearly weather-responsive traffic management systems (Thakuriah and Tilahun, 2013) and emergency response systems (Salasznyk et al., 2006) require extensive integration of very different streams of data, often in real time. This can be computationally challenging, but also perhaps

equally challenging to get data owners to share information. These types of linked data are likely to accrue a diverse user community including urban planning and operations management researchers, as well as the economic and social policy community, in addition to public and private data government data organizations.

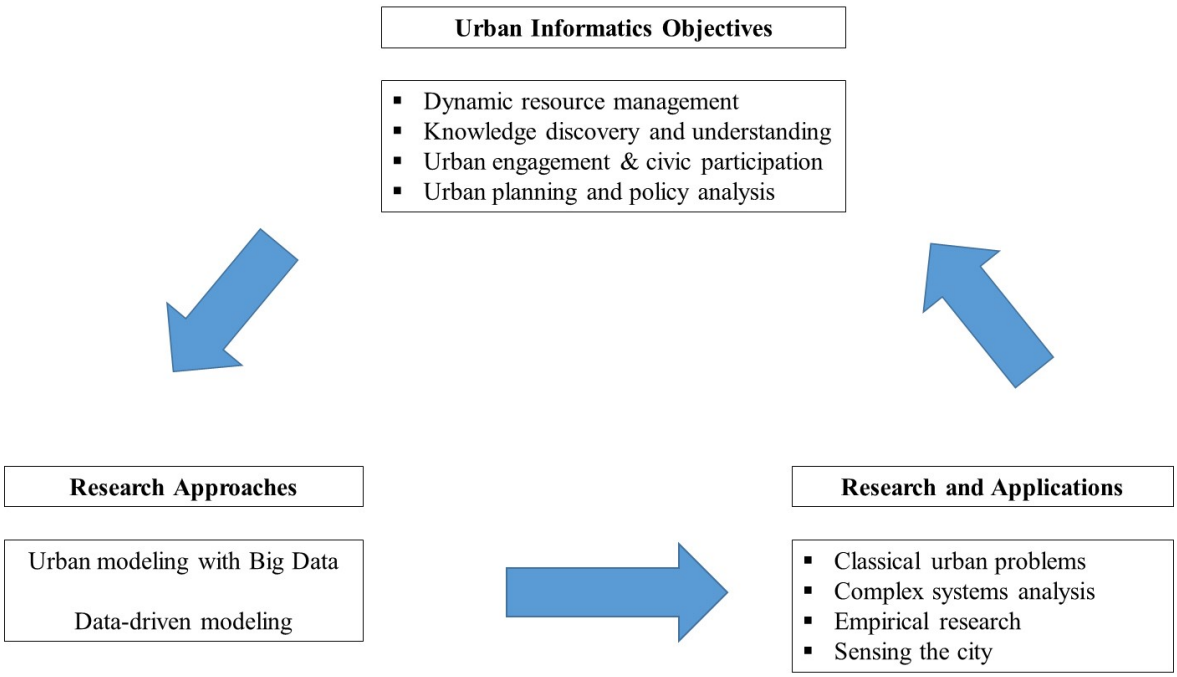
3. Urban Informatics

Overall, developments with urban Big Data have opened up several opportunities for urban analysis. Building on previous definitions (Foth et al., 2011; Bays and Callanan, 2012; Batty, 2013; Zheng, et al., 2014), we view Urban Informatics as the exploration and understanding of urban patterns and processes, and it involves analyzing, visualizing, understanding, and interpreting structured and unstructured urban Big Data for four primary objectives:

1) Dynamic resource management: developing strategies for managing scarce urban resources effectively and efficiently and often making decisions in real-time regarding competitive use of resources;
2) Knowledge discovery and understanding: discovering patterns in, and relationships among urban processes, and developing explanations for such trends;
3) Urban engagement and civic participation: developing practices, technologies and other processes needed for an informed citizenry and for their effective involvement in social and civic life of cities;
4) Urban planning and policy analysis: developing robust approaches for urban planning, service delivery, policy evaluation and reform, and also for the infrastructure and urban design decisions.

The overall framework used here, in terms of the objectives, research approach and applications, and their interdependencies, is shown in Figure 1.

Figure 1: Relationships among Urban Informatics objectives, research approaches and applications



3.1 Research Approaches in Urban Informatics

The analysis of urban systems is theoretically underpinned by myriad economic, social, behavioral, biological and physical principles, allowing the simulation of complex interactions, movements, transactions, trading, diffusion and other urban dynamics and diffusion patterns. While some urban models aim to improve long-range economic and infrastructural planning and program evaluation, others attempt to generate empirical understanding of urban dynamics and verification of theoretical urban concepts, and to provide input into shorter-term operations and management of urban sectors. However, Big Data has become closely associated with data-driven science and modeling, which is typically an empirical approach without the social, psychological, economic and regional planning theory which frame urban research. Data-driven modeling brings novel new methodological approaches particularly in using some of the highly unstructured and voluminous types of Big Data, and a bottom-up approach to understanding urban systems, particularly for improved dynamic resource management, knowledge discovery and citizen engagement.

The research approaches utilized in Urban Informatics are:

1) **Urban modeling and analysis with Big Data:** The use of Big Data within existing urban modeling and simulation frameworks, and in practical empirical approaches grounded in theoretical urban research paradigms, by: (a) reconfiguring/restructuring emerging Big Data through specialized data preparation techniques so that it meets the input requirements of existing urban modeling approaches; or (b) retrofitting or repurposing existing methods through the integration of data-driven approaches (e.g., machine learning, data mining) in the overall analysis scheme, so that urban models are able to use emerging forms of data.

2) **Data-driven models towards "bottom-up" sensing of the city:** Empirical data-driven methods that are derived primarily from the data science and statistical learning communities which focus on retrieval and extraction of information from unstructured or very voluminous streams of data that are not easily accessible to non-specialists, and their pattern detection, knowledge discovery, empirical explanation and hypothesis generation regarding urban phenomena, events and trends.

3.2 Urban Informatics Applications with Big Data

We organize the discussion on Urban Informatics applications using Big Data through urban models and data-driven models in the following ways: (1) reconsidering classical urban problems with new forms of data, (2) use of Big Data for complex systems analysis, (3) applications to address complex urban challenges through empirical research, and (4) through methods to collaboratively sense the city. The applications in turn, help to fine-tune the objectives of Urban Informatics for more comprehensive to knowledge discovery, urban planning and operations.

3.2.1 Reconsidering Classical Urban Problems with Big Data

Classical approaches to urban systems analysis include mathematical models of human spatial interaction to measure flows of travelers and services between pairs of points in urban areas (Wilson, 1971; Erlander, 1980; Sen and Srivastava, 1995), models of urban development, and study of urban structure, and the interaction between transportation and land-use systems (Burgess, 1925; Alonso, 1960; Lowry, 1964; Fujita and Ogawa, 1982; Fujita, 1988). Other areas are transportation network dynamics and travel mode choice analysis (Beckman, McGuire and Winston, 1956; Sheffi, 1985; Ben Akiva and Lerman, 1985), models of housing dynamics and residential location theory (Ellis, 1967; Muth, 1969; Beckman, 1973; Richardson, 1977); and models of regional and local economies, labor markets and industry location and agglomeration (Marshall, 1920; Isard, 1956; Krugman, 1991; Fujita et al., 1999). These methods are typically equation-based and draw from operations research and

statistical techniques. These models and their numerous variants have contributed to a voluminous and diverse literature on which several decades of planning, policy and operational decisions have been based.

These approaches typically use data from traditional sources such as the census or surveys, and to a lesser degree aggregated administrative or sensor data. Using many other urban Big Data sources would require significant modifications to such models "to see around the corners", perhaps new model development or integration with data science approaches, in addition to specialist curation and processing of the data itself. However, there are now an increasing number of examples where emerging forms of data have been used within the framework of these classical urban models. Recent examples are in the areas of travel demand models, e.g., the use of GPS data to estimate flows of travelers from travel origins to destinations traditionally achieved using census journey-to-work data (Zhang et al., 2010), and the use of detailed freeway and arterial street sensor data along with the synthetic LEHD and other data to measure job accessibility (Levinson et al., 2010). Other examples include studies of labor market dynamics using administrative data (e.g., Bijwaard et al., 2011), use of social media data to measure labor market flows and indexes of job loss, job search, and job posting (Antenucci et al., 2014) and the use of online housing searches to study housing market dynamics in terms of area definition, submarket geography and search pressure locations (Rae, 2014).

3.2 Complex Systems Analysis

Large-scale urban modeling practice also use complex systems approaches utilizing Agent-based Models (ABM) and myriad forms of specialized survey, administrative, synthetic and other data sources, to study outcomes that are emergent from individual agent action in interaction with other agents and the environment while also incorporating agent heterogeneity. Well-known early implementations of ABMs include Schelling's segregation model (Shelling, 1971) and Conway's Game of Life (Conway, 1970). ABMs have found widespread application in diverse areas of urban research. Examples include urban and environmental planning (e.g. Zellner et al., 2009; Zellner and Reeves, 2012), transportation (e.g. Tilahun and Levinson, 2013; Zellner et al., forthcoming), environmental studies (e.g. Evans and Hugh, 2004), large-scale agent based micro-simulation models such as ILUTE (Salvini and Miller, 2005), and integrated land, transportation and environment modeling system such as MATSim (Balmer et al., 2009), which provides agent-based mobility simulations. Related developments in computational network perspectives to study a variety of phenomenon have also entered modeling practice, including studies of community structure (Girvan and Newman, 2002) and susceptibility of power grids to failure (Kinney et al., 2005).

ABMs have recently used unstructured sources of data, one example of which is the use of GPS trajectories to obtain a better understanding of human mobility patterns within an ABM framework (Jia et al., 2012). Some researchers have also focused on the use of social network data (Kowald and Axhausen, 2015 gives examples for the case of transportation planning), while others have utilized social networks to examine peer effects, and processes to exchange opinions, preferences and to share experiences, as well as to see how individual's participation in social networks lead to outcomes of interest (e.g., Christakis and Fowler, 2007, demonstrated the spread of obesity via social relationships in a social network while Tilahun et al., 2011 examined the role of social networks in location choice). The use of online social networks in ABMs has been an interesting development in this respect allowing the flexibility of ABMs to incorporate detailed representation and analysis of the effects of social networks that underlie complex decision problems. One example of this nascent literature is the use of Twitter data within an ABM framework to model diffusion of crisis information (Rand et al., 2015).

3.3 Empirical Urban Research

A vast body of empirical work embedded in the urban disciplines is among the most active consumers of urban data, for better understanding, hypothesis testing and inference regarding urban phenomenon. Among these, one vast research area with requirements for specialized data sources, models and tools is that of environmental sustainability and issues relating to clean air, non-renewable energy dependence and climate change. While difficult to generalize, a recent OECD report identified gaps in quantitative urban and regional modeling tools to systematically assess the impacts of urban systems on climate change and sustainability (OECD, 2011). Significant developments in sensor technology have led to smart commodities ranging from household appliances to smart buildings leading to cost-efficiencies and energy savings, for the design of Vehicle-to-Grid (V2G) systems (Kempton et al., 2005), personal carbon trading (Bottrill, 2006) and vehicular cap-and-trade systems (Lundquist, 2011), with data-analytic research around technology adoption, impacts on behaviors and consumption patterns and so on.

Urban models that detect disparities relating to social justice and distributional aspects of transportation, housing, land-use, environment and public health are other consumers of such data. These approaches provide an empirical understanding of the social inclusion and livability aspects of cities, and operational decisions and policy strategies needed to address disparities. This line of work has focused on social exclusion and connections to work and social services (Kain and Persky, 1969; Wilson, 1987; Krivo and Petersen, 1996; Thakuriah et al., 2013), issues of importance to an aging society (Federal Interagency Forum on Aging-Related Statistics, 2010), health and aging in place (Black, 2008; Thakuriah et al., 2011) and needs of persons with disabilities (Reinhardt et al., 2011). Administrative data has played a significant role in this type of research leading to knowledge discovery about urban processes as well as in evaluation of governmental actions such as welfare reform and post-recession austerity measures. Linked and longitudinal administrative data can support understanding of complex aspects of social justice and changes in urban outcomes over time. For example, Evan et al. (2010) highlighted the importance of using longitudinal administrative data to understand the long-term interplay of multiple events associated with substance abuse over time, while Bottoms et al. (2009) discuss the role that longitudinal police crime records can play in studying repeat victimization of crime.

New ICT-based solutions to track and monitor activities allow urban quality and well-being to be assessed at more fine-grained levels. Personalized data generated by assistive technologies and ambient assisted living situations (Abascal et al., 2008) and other ICT applications can be used to assess contributory factors to urban quality of life as well as to design solutions supporting urban wellness objectives for seniors and persons with disability (e.g., hybrid qualitative-GPS data enabled as described by Huang et al., 2012 to understand barriers to accessing food by midlife and older adults with mobility disability). Mobile heath and awareness technologies (Consolvo et al., 2006) particularly those embedded within serious medical pervasive gaming environments (e.g., DiabetesCity—Collect Your Data, Knoll, 2008) and numerous mobile, wearable and other sensor-based physical health recommender systems, one example of which is Lin et al. (2011), open up possibilities for urban researchers to tap into a wealth of data to understand overall built environment and activity-based conditions fostering health and well-being.

3.4 Approaches to Collaboratively Sense the City

The discussion above shows that urban information generation and strategies to analyze the data increasingly involve ICT solutions and the active participation of users. Strategies such as focus groups, SWOT, Strategic Approach, Future Workshops and other approaches have been extensively used in the past as a part of urban participatory practice to generate ideas and even to generate solutions to problems. However, advances in ICT solutions have led to the emergence of new models of citizens input into problem solving, plan and design sourcing, voting on projects, and sharing of

ideas on projects. Examples range from civic hackers analyzing data from Open Data portals to generate ideas about fixing urban problems to using serious games and participatory simulations for the ideation process (Poplin, 2014; Zellner et al., 2012).

As noted earlier, citizens may also engage by generating content through human computation, or by performing tasks that are natural for humans but difficult for machines to automatically carry out (von Ahn et al., 2008). Human computation approaches provide structured way for citizens to engage in play, to provide input and to interact with, and learn about the urban environment. For example, citizens may be able to judge different proposed urban design, or they may be used to assess the quality of urban spaces where objective metrics from data derived through machine vision algorithms are not accurate. Celino et al. (2012) gives an example of this called UrbanMatch, a location-based Game with a Purpose (GAWP), which is aimed at exploiting the experience that players have of the urban environment to make judgments towards correctly linking points of interests in the city with most representative photos retrieved from the Internet. There are multiple variants of human computation including social annotations (where users tag or annotate photos or real-world objects), information extraction (e.g., where users are asked to recognize objects in photos), and others.

By "sensing" the city and its different behavioral and use patterns, data-driven models have stimulated research into a broad range of social issues relevant to understanding cities, including building participatory sensing systems for urban engagement, location-based social networks, active travel and health and wellness applications, and mobility and traffic analytics. Other objectives include dynamic resource management of urban assets and infrastructure, assisted living and social inclusion in mobility, and community and crisis informatics. For example, one of the major cited benefits of social media analysis has been the ability to instantaneously and organically sense sentiments, opinions and moods to an extent not previously possible, and ways in which these diffuse over space and time, thereby enabling the policy community to monitor public opinion, and predict social trends. A part of this trend is being stimulated by major governmental agencies which are increasingly realizing the power of social media in understanding where needs are, and how the public are reacting to major policy changes and political events and people's political preferences (Golbeck and Hansen, 2013).

A data-driven focus is also being seen in learning analytics (e.g., Picciano, 2012), location-based social networks (Zheng and Xie, 2011), recommender systems based on collaborative filtering for travel information (Ludwig et al., 2009) and approaches to detect disruptions from social media (Sasaki et al., 2012). Presumably if these information streams are collected over time and linked to other socio-demographic data, it would be possible to examine variations in the outcomes currently measured by the socially generated data to capture urban dynamics to a greater degree.

Overall, Big Data is being increasingly utilized for a range of Urban Informatics research and applications. By using existing urban models with new forms of data, or through data-driven modeling, urban processes and behaviors can be studied in a timely manner and contextual peculiarities of urban processes and local experiences can examined in greater detail. Yet significant challenges arise in their use, which are addressed next.

4. Challenges in Using Big Data for Urban Informatics

The challenges associated with the use of Big Data for Urban Informatics are: (1) technological, (2) methodological, (3) theoretical and epistemological, and (4) due to political economy that arise from accessing and using the data. These challenges are given in Table 2 along with the characteristics of the challenges and examples of the complexities involved with different types of Big Data.

Table 2: Challenges in using Big Data for Urban Informatics and illustrative topics

Challenges	Characteristics	Challenges by type of data
Technological	Urban information management challenges: 1) Information generation and capture 2) Management 3) Processing 4) Archiving, curation and storage 5) Dissemination and discovery	Information management challenges likely to very high with real-time, high-volume sensor and UGC data which require specific IT infrastructure development and information management solutions
Methodological	1) Data Preparation Challenges a) Information retrieval and extraction b) Data linkage/information integration c) Data cleaning, anonymization and quality assessment 2) Urban Analysis Challenges a) Developing methods for data-rich urban modeling and data-driven modeling b) Ascertaining uncertainty, biases and error propagation	Data preparation challenges likely to be very high with unstructured or semi-structured sensor, UCG and arts and humanities data, and data from real-time private-sector and administrative transactional systems All types of observational Big Data pose significant methodological challenges in deriving generalizable knowledge requiring specialist knowledge to assess and address measurement issues and error structures
Theoretical and epistemological	1) Understanding metrics, definitions, concepts and changing ideologies and methods to understanding "urban" 2) Determining validity of approaches and limits to knowledge 3) Deriving visions of future cities and the links to sustainability and social justice	All types of observational Big Data pose limitations in deriving theoretical insights and in hypothesis generation without adequate cross-fertilization of knowledge between the data sciences and the urban disciplines, but the challenges are greater with certain forms of UGC and sensor data which yield high-value descriptions but are less amenable to explanations and explorations of causality
Political economy	1) Data entrepreneurship, innovation networks and power structures 2) Value propositions and economic issues 3) Data access, governance framework and provenance 4) Data confidentiality, security and trust management 5) Responsible innovation and emergent ethics	Data confidentiality and power structures pose significant challenges to use of administrative data in open government and program evaluation, while access to private sector transactions data, and privately-controlled sensor and UGC are potentially susceptible to changing innovation and profitability motivations; challenges to ethics and responsible innovation are significantly high for certain sensor-based (e.g., IOT) applications

4.1 Technological Challenges

Technological challenges arise due to the need to generate, capture, manage, process, disseminate and discover urban information. The challenges to managing large volumes of structured and unstructured information have been extensively documented elsewhere. Some of the major

information management challenges are those relating to building a data infrastructure, cloud stores and multi-cloud architectures, as well as resource discovery mechanisms, and language and execution environments. Other considerations include hardware, software, well-defined Application Programming Interfaces (API) needed to capture, integrate, organize, search and query and analyze the data. Of equal importance are scalability, fault-tolerance, and efficiency, and platforms for scalable execution. Various Big Data solutions have emerged in the market such as Hadoop, MapReduce and other solutions, some of which are open source.

One of the biggest challenges with using Big Data for Urban Informatics is not that the data are necessarily huge as in the case of financial, genomics, high-energy physics or other data (although this may change with the incoming deluge of the connected vehicle and the IoT world). Rather, it is that urban Big Data tends to be fragmented, messy and sometimes unstructured. Particularly for data linkage, when one goes beyond structured, rectangular databases to streaming data through APIs leading to text, image and other unstructured data formats, the diversity and fragmentation can pose significant problems.

Data privacy also becomes all-important with many sources of Big Data, whether they are administrative micro-data or user-generated image or GPS data, and is often a major roadblock to data acquisition for research, particularly for research that requires potentially personally identifiable data. There are many approaches to data privacy, and these range from technological encryption and anonymization solutions to design, access and rights management solutions. A vast range of Privacy Enhancing Technologies (PETs) (Beresford et al., 2003, Grusteser et al., 2003) are relevant to urban Big Data that focuses on anonymization of GPS data, images and so on. In the case of administrative micro-data, many approaches to ensure confidentiality are used, including de-identified data, simulated micro-data (called synthetic data) that is constructed to mimic some features of the actual data using micro-simulation methods (Beckman et al., 1996; Harland et al., 2012) and utilization of Trusted Third Party (TTP) mechanisms to minimize the risks of the disclosure of an individual's identity or loss of the data (Gowans et al., 2012).

One major capability needed to progress from data-poor to data-rich urban models is that data should be archived over time, enabling storage of very high-resolution and longitudinal spatio-temporal data. The linkage to other socio-economic, land-use and other longitudinal data opens up additional avenues for in-depth exploration of changes in urban structure and dynamics. Although this was previously a challenge, decrease in storage costs and increase in linkage capacity has made this possible.

Another important determinant in data access is having access to high-quality resource discovery tools for urban researchers to find and understand data, ontologies for knowledge representation, and data governance framework that includes harmonization of standards, key terms and operational aspects. Given the vast and dispersed sources of urban Big Data, resource discovery mechanisms to explore and understand data are critical. This includes metadata or data about the data, containing basic to advanced information describing the data and the management rights to it, including archiving and preservation, in a consistent, standardized manner so that it is understandable and usable by others. Other issues are data lifecycle management (the strategic and operational principles underpinning long-term publication and archiving), access to necessary para-data, (i.e., data about the processes used to collect data), and social annotations (i.e., social bookmarking that allows users to annotate and share metadata about various information sources). These issues not only have significant technical requirements in terms of integrating urban data from different sources; they also have legal (e.g., licensing, terms of service, non-disclosure), ethical (e.g., regarding lack of informed consent in some cases, or use by secondary organizations which did not seek consent), and research culture implications (e.g., establishing a culture of reanalysis of evidence, reproduction and verification of

results, minimizing duplication of effort, and building on the work of others, as in Thanos et al., 2015).

The above technology issues are not likely to be directly relevant to urban researchers in many cases. However, methodological aspects of Big Data such as information retrieval, linkage and curation or the political economy of Big Data including data access, governance and privacy and trust management requirements may have direct implications for, and limit, urban researchers if appropriate technology solutions capable of handling these IT requirements cannot be found.

4.2 Methodological Challenges

We consider two types of methodological challenges: data preparation methods (such as cleaning, retrieving, linking, and other actions needed to prepare data for the end-user) and empirical urban analysis methods (data analytics for knowledge discovery and empirical applications). Sensor and co-created data require special processing and analysis methods to manage very large volumes of unstructured data, from which to retrieve and extract information. With traditional sources of urban data, the specific aspects of the workflow from data collection/generation to analysis are clearly demarcated among professionals from different backgrounds (e.g., data collection is typically done by census takers or surveyors who create a clean data file along with the necessary data documentation, which is then used by urban researchers for further analysis). In contrast to this model, in the case of certain forms of unstructured data (e.g., social media data such as Twitter), the analytics of the data (e.g., using machine learning for topic detection and classification algorithms) happens alongside with, or as a part of, information retrieval or the "gathering" of information from the raw data streams. Thus the "data gathering" and the "data analytics" aspects of the workflow are much more tightly coupled, requiring new skills to be learned by urban researchers wishing to use such data or to have close collaboration with data scientists who have this type of skills.

Observational Big Data involves having to address several methodological challenges for inference. Administrative data, for example, may pose challenges due to causality, endogeneity, and other issues that can bias inference. Socially generated data obtained from participatory sensing and crowd-sourcing are likely to be non-representative in the sense that participants probably do not resemble random samples of the population. Those who are easiest to recruit may also have strong opinions about what the data should show and can provide biased information. Social media users are typically not representative of the overall population since they are more likely to be younger and more digitally savvy (Mislove et al., 2011), and they are also more likely to be concentrated in certain areas or generate differing content depending on where they live (Ghosh and Guha, 2013), although these patterns may change over time as the technology becomes more widely used.

In addition, technology changes rapidly and there would always be the issue of the first adopters with specific, non-representative demographics and use patterns. Aside from this, there is dominance by frequent issues and lack of data generation by passive users, and the proliferation of fake accounts which does not add real or true representation of moods, opinions and needs, and are sometime maliciously created to swell sentiments in one direction or the other. Other challenges include lack of independence or herding effects, which is the effect of prior crowd decisions on subsequent actions. Samples may need to be routinely re-weighted, again on the fly, with the weights depending on the purpose of the analysis. Recent work by Dekel and Shamir (2009), Raykar et al., (2010), and Wauthier and Jordan (2011) consider issues on sampling and sampling bias in crowd-sourcing or citizen science while others have considered sampling issues relating to social networks (Gjoka et al., 2010) and social media (Culotta, 2014). However, this work is in its infancy, and further developments are necessary in order to use highly unstructured forms of data for urban inference.

Using Big Data for Urban Informatics require methods for information retrieval, information extraction, GIS technologies, and multidisciplinary modeling and simulation methods from urban

research as well as the data sciences (e.g., machine learning and tools used to analyze text, image and other unstructured sources of data). Methods of visualization and approaches to understanding uncertainty, error propagation and biases in naturally occurring forms of data are essential in order to use and interpret Big Data for urban policy and planning.

4.3 Theoretical and Epistemological Challenges

The theoretical and epistemological challenges pertain to the potential for insights and hypothesis generation about urban dynamics and processes, as well as validity of the approaches used, and the limits to knowledge discovery about urban systems derived from a data focus. As noted earlier, Big Data for Urban Informatics has two distinct roots: quantitative urban research and data science. Although the walls surrounding what may be considered as "urban models and simulations" are pervious, these are typically analytical, simulation-based or empirical approaches that are derived from diverse conceptual approaches (e.g., queuing theory, multi-agent systems) and involve strong traditions of using specialist data to calibrate. These models support the understanding of urban structure, forecasting of urban resources, simulation of alternative investment scenarios, strategies for engagement of different communities, and evaluation of planning and policy, as well as efficient operations of transportation, environmental and other systems, using principles derived from theory. Such models are now using Big Data in varying degrees.

At the same time, exploratory data-driven research is largely devoid of theoretical considerations but is necessary to fully utilize emerging data sources to better discover and explore interesting aspects of various urban phenomena. Social data streams and the methods that are rapidly building around them to extract, analyze and interpret information are active research areas, as are analytics around data-driven geography that may be emerging in response to the wealth of geo-referenced data flowing from sensors and people in the environment (e.g., Miller and Goodchild, 2014). The timely discovery and continuous detection of interesting urban patterns possible with Big Data and the adoption of innovative data-driven urban management are an important step forward and serves useful operational purposes. The knowledge discovery aspects of data-driven models are important to attract the attention of citizens and decision-makers on urban problems and to stimulate new hypotheses about urban phenomena, which could potentially be rigorously tested using inferential urban models.

The limitation of the data-driven research stream is that there is less of a focus on the "why" or "how" of urban processes and on complex cause-and-effect type relationships. In general, data-driven methods have been the subject of interesting debates regarding the scope, limitations and possibility of such approaches to provide solutions to complex problems beyond pattern detection, associations, and correlations. The current focus on data-driven science and the advocacy for it have in some cases led to rather extreme proclamations to the effect that the data deluge means the "end of theory" and that it will render the scientific process of hypothesizing, modeling, testing, and determining causation obsolete (Anderson, 2008). Quantitative empirical research has always been a mainstay for many urban researchers but there is inevitably some conceptual underpinning or theoretical framing which drive such research. Long before Big Data and data science became options to derive knowledge, the well-known statistician, Tukey (1980), noted in an article titled "We Need Both Exploratory and Confirmatory" that "to try to replace one by the other is madness", while also noting that "ideas come from previous exploration more often than from lightning strikes".

A part of the debate is being stimulated by the fact that data-driven models have tended to focus on the use of emerging sources of sensor or socially co-created data, and is closely connected to the data science community. At the time of writing this paper, entering "GPS data" into the Association for Computing Machinery (ACM) Digital Library, a major computer science paper repository returns about 11,750 papers, while entering the same term in IEEE XPlore Digital Library, another such

source, returns another 6,727 papers; these numbers are in fact higher than the counts obtained when the first author was writing her book "Transportation and Information: Trends in Technology and Policy" (Thakuriah and Geers, 2013), indicating not just a voluminous literature on these topics in the data sciences but one that continues to grow very fast. Such data sources have become most closely associated with the term Big Data in the urban context, to the exclusion of administrative data, hybrids of social survey and sensing data, humanities repositories and other novel data sources, which play an important role in substantive, theoretically-informed urban inquiry, beyond detection, correlations and association.

Sensor and ICT-based UGC has also become closely associated with smart cities, or the use of ICT-based intelligence as a development strategy mostly championed and driven by large technology companies for efficient and cost-effective city management, service delivery and economic development in cities. There are numerous other definitions of smart cities, as noted by Hollands (2008). The smart cities movement has been noted to have several limitations, including having "a one-size fits all, top-down strategic approach to sustainability, citizen well-being and economic development" (Haque, 2012) and for being "largely ignorant of this (existing and new) science, of urbanism in general, and of how computers have been used to think about cities since their deployment in the mid-20th century" (Batty, 2013), a point also made by others such as Townsend (2013). It needs to be pointed out that the scope of smart cities has expanded over time to include optimal delivery of public services to citizens and on processes for citizen engagement and civic participation, as encapsulated by the idea of "future cities".

Urban Big Data is also now being strongly associated with Open Data; Open Data is now being increasingly linked to smart cities, along with efforts to grow data entrepreneurship involving independent developers and civic hackers to stimulate innovations and promote social change. Nevertheless, at least in the European Union, the European Innovation Partnership (EIP) on Smart Cities and Communities has received some 370 commitments to fund and develop smart solutions in the areas of energy, ICT and transport. These commitments involve more than 3,000 partners from across Europe towards creating "a huge potential for making our cities more attractive, and create business opportunities" (European Commission, 2015). It is little wonder that the term Big Data for cities is being referred to in some circles almost exclusively in the context of smart cities, to the exclusion of urban research contributions, including a long-standing urban operations literature using sensor and at least some types of user-generated data.

Notwithstanding these tensions, some of the benefit of using sensor and socially generated forms of Big Data is in identifying contextual particularities and local experiences that are very often smoothed over by the systems-oriented view of quantitative urban research; the latter often emphasizes generalizability, sometimes masking elements of complex urban challenges. Such "local" focus lends the hope that observations of unique local features from data will stimulate interest in exploring previously unknown hypothesis of urban processes and that the unique local urban problems identified potentially lends itself to context-dependent urban policy and plan-making. The "new geographies of theorizing the urban" (Robinson, 2014, Roy, 2009) is oriented to skepticism regarding authoritative and universalizing claims to knowledge about urban experiences and is committed to giving attention to contextual particularities and local experiences within places (Brenner and Schmid, 2015). Although epistemological links between data-driven urban modeling and critical urban theory is virtually non-existent at the current time, and may never be explicitly articulated, novel sources of Big Data have the potential to allow the capture of data on social, behavioral and economic aspects of urban phenomena that have either not been previously measured or have been measured at resolutions that are too aggregated to be meaningful. However, such localized observations are far from being a substitute for qualitative social science research, as noted by Smith (2014), who advocates a continued need for ethnographic approaches and qualitative methods and cautions against the continued separation of method from methodology and discipline.

Further, causality assessment is challenging with many forms of Big Data and it does not lend itself easily to the derivation of counterfactuals and to forming an etiologic basis for complex urban processes. Instead of moving from using urban models to a completely different data-driven era, as noted earlier, the focus may be to shift to using administrative, sensing or socially generated urban Big Data as input into estimating and testing traditional models. Urban Big Data analysis would also benefit from being linked to behavioral models needed to build alternative scenarios to understand the effects of unobserved assumptions and factors, or to derive explanations for parts of the urban environment not measured by data. The linked data hybrids suggested previously potentially offers a way to address these limitations.

4.4 Challenges due to the Political Economy of Big Data

The political economy of Big Data arises due to the agendas and actions of the institutions, stakeholders and processes involved with the data. Many of the challenges facing urban researchers in using Big Data stem from complexities with data access, data confidentiality and security, and responsible innovation and emergent ethics. Access and use conditions are in turn affected by new types of data entrepreneurship and innovation networks, which makes access easier in some cases through advocacy for Open Data or makes it more difficult through conditions imposed as a result of commercialization and are generally underpinned by power structures and value propositions arising from Big Data.

The economic, legal and procedural issues that relate to data access and governance are non-trivial and despite the current rhetoric around the open data movement, vast collections of data that are useful for urban analysis are locked away in a mix of legacy and siloed systems owned and operated by individual agencies and private organizations, with their own internal data systems, metadata, semantics and so on. Retrieving information from social media and other online content databases, and the analytics of the resulting retroactive UGC either in real-time or from historical archives have mushroomed into a significant specialized data industry, but the data availability itself is dictated by the terms of service agreements required by the private companies which own the system or which provide access, giving rise to a new political economy of Big Data. User access is provided in some cases using an API, but often there are limits on how much data can be accessed at any one time by the researcher and the linkage of a company's data to other data. Others may mandate user access under highly confidential and secure access conditions requiring users to navigate a complex legal landscape of data confidentiality, and special end-user licensing and terms of service and non-disclosure agreements. Data users may also be subject to potentially changing company policy regarding data access and use. There are also specific restrictions on use including data storage in some cases, requiring analytics in real-time.

Undoubtedly, a part of the difficulty in access stems from data confidentiality and the need to manage trust with citizens, clients and the like. Privacy, trust and security are concepts that are essential to societal interactions. Privacy is a fundamental human right and strategies to address privacy involve privacy-enhancing technology, the legal framework for data protection, as well as consumer awareness of the privacy implications of their activities (Thakuriah and Geers, 2013), especially as users leave a digital exhaust with their everyday activities. However, privacy is also not a static, immutable constant. People are likely to trade off some privacy protection in return for utility gained from information, benefits received, or risks minimized (Cottrill and Thakuriah, 2015). Aside from technological solutions to maintain privacy, a process of user engagement is necessary to raise consumer awareness, in addition to having the legal and ethics processes in place in order to be able to offer reassurance about confidential use of data. Further, many private data owners may not release data due to being able to reserve competitive advantage through data analytics. However, lack of knowledge about the fact-moving legal landscape with regards to data confidentiality, copyright

violations and other unintended consequences of releasing data are central elements of the political economy of Big Data.

The social arguments for and against Big Data, connected systems and IoT are similar to other technology waves that have been previously witnessed, and these considerations also generate multiple avenues for research. Increasingly pervasive sensing and connectivity associated with IoT, and the emphasis on large-scale highly coupled systems that favor removing human input and intervention has been seen to increase exposure to hacking and major system crashes (BCS, 2013). Aside from security, the risks for privacy are greatly enhanced as the digital trail left by human activities may be masked under layers of connected systems. Even those systems that explicitly utilize privacy by design are potentially susceptible to various vulnerabilities and unanticipated consequences since the technological landscape is changing very rapidly and the full implications cannot be thought through in their entirety. This has prompted the idea of "responsible innovation", which "seeks to promote creativity and opportunities for science and innovation that are socially desirable and undertaken in the public interest" and which makes clear that "innovation can raise questions and dilemmas, is often ambiguous in terms of purposes and motivations and unpredictable in terms of impacts, beneficial or otherwise. Responsible Innovation creates spaces and processes to explore these aspects of innovation in an open, inclusive and timely way" (Engineering and Physical Sciences Research Council, n.d.).

Against this backdrop of complex data protection and governance challenges, and the lure of a mix of objectives such as creating value and generating profit as well as public good, a significant mix of private, public, non-profit and informal infomediaries, ranging from very large organizations to independent developers that are leveraging urban Big Data have emerged. Using a mixed-methods approach, Thakuriah et al. (2015) identified four major groups of organizations within this dynamic and diverse sector: general-purpose ICT companies, urban information service providers, open and civic data infomediaries, and independent and open source developer infomediaries. The political economy implication of these developments is that publicly available data may become private as value is added to such data, and the publicly-funded data infrastructure, due to its complexity and technical demands, are increasingly managed by private companies that in turn, potentially restricts access and use.

5. Conclusions

In this paper, we discussed the major sources of urban Big Data and their benefits and shortcomings, and ways in which they are enriching Urban Informatics research. The use of Big Data in urban research is not a distinct phase of a technology but rather a continuous process of seeking novel sources of information to address concerns emerging from high cost or design or operational limitations. Although Big Data has often been used quite narrowly to include sensor or socially generated data, there are many other forms that are meaningful to different types of urban researchers and user communities, and we include administrative data and other data sources to capture these lines of scholarship. But even more importantly, it is necessary to bring together (through data linkage or otherwise), data that have existed in fragmented ways in different domains, for a holistic approach to urban analysis.

We note that both theory-driven as well as data-driven approaches are important for Urban Informatics but that retrofitting urban models to reflect developments in a data-rich world is a major requirement for comprehensive understanding of urban processes. Urban Informatics in our view is the study of urban patterns using novel sources of urban Big Data that is undertaken from both a theory-driven empirical perspective as well as a data-driven perspective for dynamic resource management, knowledge discovery and understanding, urban engagement and civic participation, and urban planning and policy. The research approaches utilized to progress these objectives are a mix of enriched urban models underpinned by theoretical principles and retrofitted to accommodate emerging forms of data, or data-driven modeling that are largely theory-agnostic and emerge bottom-

up from the data. The resulting Urban Informatics research applications have focused on revisiting classical urban problems using urban modeling frameworks but with new forms of data; evaluation of behavioral and structural interactions within enriched complex systems approach; empirical research on sustainable, socially-just and engaged cities; and applications to engage and collaboratively sense cities.

The use of Big Data pose considerable challenge for Urban Informatics research, This includes technology-related challenges putting requirements for special information management approaches, methodological challenges to retrieve, curate and draw knowledge from the data; theoretical or epistemological challenges to frame modes of inquiry to derive knowledge and understand the limits of Urban Informatics research; and finally, an issue that is likely to play an increasingly critical role for urban research – the emerging political economy of urban Big Data, arising from complexities associated with data governance and ownership, privacy and information security, and new modes of data entrepreneurship and power structures emerging from the economic and political value of data. From the perspective of urban analysts, the use of sensor data, socially generated data, and certain forms of arts and humanities and private sector data may pose significant technical and methodological challenges. With other sources such as administrative micro-data, the data access challenges and issues relating to political economy and data confidentiality might be non-trivial. Issues such as sustainability of the data infrastructure, dealing with data quality, and having access to the skills and knowledge to make inferences, apply to all forms of naturally occurring data.

While many types of urban Big Data such as administrative data and specific sensor systems have been used for a long time, there are many novelties as well, such as new, connected sensor systems, and socially generated or hybrid, linked data that result in data in new formats or structure. There is a need for a wide variety of skills due to the tight coupling of preparing unstructured data and data analysis, but also due to the wide variety of technological, methodological and political economy issues involved. Additionally, data and analytics are only one part of the data-focused approach to urban operations, planning and policy-making; having the mechanisms to interpret the results and to highlight the value derived, is critical for adoption of data-driven strategies by decision-making, and for its eventual impact on society.

It is therefore an opportune time for an interdisciplinary research community to have a discussion on the range of issues relating to the objectives of Urban Informatics, the research approaches used, the research applications that are emerging, and finally, the many challenges involved in using Big Data for Urban Informatics.

References

[1] Abascal, J., B. Bonail, Á. Marco, R. Casas, and J. L. Sevillano (2008). AmbienNet: an intelligent environment to support people with disabilities and elderly people. In Proc. 10th International ACM SIGACCESS Conference on Computers and Accessibility (Assets ’08), pages 293–294.

[2] Abowd, J. M., B. E. Stephens, L. Vilhuber, F. Andersson, K. L. McKinney, M. Roemer and S. Woodcock (2005). The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In Producer Dynamics: New Evidence from Micro Data. Published 2009 by University of Chicago Press, 149 - 230. Accessed from http://www.nber.org/chapters/c0485.pdf on March 1, 2014.

[3] Aguilera, O., A. F. Fernández , A. Muñoz and M. F. Fraga (2010). Epigenetics and environment: a complex relationship. Journal of Applied Physiology, Vol. 109(1), 243-251.

[4] Alonso, W. (1960). A Theory of the Urban Land Market. Papers in Regional Science, Vol. 6(1), pp. 149-157.

[5] American Academy of the Arts and Sciences (2013). The Heart of the Matter. Accessed from http://www.amacad.org on April 1, 2015.

[6] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine, June 23, 2008. Accessed from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory on Feb 10, 2012.

[7] Antenucci, D., M. Cafarella, M. C. Levenstein, C. Ré, M. D. Shapiro (2014). Using Social Media to Measure Labor Market Flows. Report of the University of Michigan node of the NSF-Census Research Network (NCRN) supported by the National Science Foundation under Grant No. SES 1131500.

[8] Auto-id Labs. http://www.autoidlabs.org.

[9] Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., Nagel, K., & Axhausen, K. (2009). MATSim-T: Architecture and simulation times. Multi-agent systems for traffic and transportation engineering, 57-78.

[10] Batty, M. (2013). Urban Informatics and Big Data: A Report to the ESRC Cities Expert Group. Accessed from http://www.smartcitiesappg.com/wp-content/uploads/2014/10/Urban-Informatics-and-Big-Data.pdf on December 15, 2014.

[11] Bays, J. and L. Callanan (2012). 'Urban informatics' can help cities run more efficiently. McKinsey on Society. Access from http://mckinseyonsociety.com/emerging-trends-in-urban-informatics/ on July 1, 2014.

[12] BCS, the Chartered Institute for IT (2013). The Societal Impact of the Internet of Things. Accessed from www.bcs.org/upload/pdf/societal-impact-report-feb13.pdf on April 10, 2015.

[13] Beckman, M.J., C. B. McGuire and C.B. Winston (1956). Studies in the Economics of Transportation. Yale University Press, Connecticut.

[14] Beckman, R., K. A. Baggerly, M. D. McKay (1996). Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice. Vol 30(6), 415–429.

[15] Beckmann, M.J. (1973). Equilibrium models of residential location. Regional and Urban Economics, Vol. 3, 361-368.

[16] Ben-Akiva, M. and S. R. Lerman (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press.

[17] Beresford, A. R. and F. Stajano (2003). Location privacy in pervasive computing. Pervasive Computing, IEEE, Vol. 2(1), pp. 46–55.

[18] Black, K. (2008). Health and Aging-in-Place: Implications for Community Practice. Journal of Community Practice, 16: 1, 79-95.

[19] Bonney, R., H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk, and C. C. Wilderman (2009). Public participation in scientific research: Defining the field and assessing its potential for informal science education. Technical report, Center for Advancement of Informal Science Education.

[20] Bottoms, A. E. and A. Costello (2009). Crime prevention and the understanding of repeat victimization: a longitudinal study. In Urban Crime Prevention, Surveillance, and Restorative Justice: Effects of Social Technologies, ed. Knepper, P., J. Doak, J. Shapland, CRC Press, pp. 23-54.

[21] Bottrill, C. (2006). Understanding DTQs and PCAs. Technical report, Environmental Change. Institute/UKERC, October.

[22] Burgess, E. W. (1925). The Growth of the City: an Introduction to a Research Project. In R. E. Park, E. W. Burgess and R. D. Mackenzie (eds). The City. University of Chicago Press, 47-62.

[23] Card, D., R. Chetty, M. Feldstein and E. Saez (n.d.) Expanding Access to Administrative Data for Research in the United States. NSF-SBE 2020 White Paper. Accessed from http://www.nsf.gov/sbe/sbe_2020/all.cfm on April 10, 2015.

[24] Celino , I., S. Contessa, E. Della Valle, T. Krüger, M. Corubolo and S. Fumeo. (2012). UrbanMatch – linking and improving Smart Cities Data. LDOW2012, Lyon, France.

[25] Consolvo, S., K. Everitt, I. Smith, and J. A. Landay (2006). Design requirements for technologies that encourage physical activity. In Proc. SIGCHI Conference on Human Factors in computing systems (CHI '06), 457–466.

[26] Conway, John. "The game of life." Scientific American 223.4 (1970): 4.

[27] Cottrill, C. D. and P. Thakuriah (2015). Location privacy preferences: A survey-based analysis of consumer awareness, trade-off and decision-making. In Transportation Research Part C: Emerging Technologies, Vol. 56, July 2015, pp. 132-148.
[28] Culotta, A. (2014). Reducing Sampling Bias in Social Media Data for County Health Inference. JSM Proceedings, 2014.
[29] Davenport, T. H. and D. J. Patil (2012). Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review, October, 70-76.
[30] Dekel, O., and O. Shamir (2009). Vox Populi: Collecting high-quality labels from a crowd. Pp. 377-386 in Proceedings of the 22nd Annual Conference on Learning Theory (COLT). Available at http://www.cs.mcgill.ca/~colt2009/proceedings.html.
[31] Drake, J. S., J. L. Schofer, A. May, A. D. May (1965). Chicago Area Expressway Surveillance Project, and Expressway Surveillance Project (Ill.). A Statistical Analysis of Speed-density Hypotheses: A Summary. Report (Expressway Surveillance Project (Ill.)). Expressway Surveillance Project.
[32] Ellis, R.H. (1967). Modelling of household location: A statistical approach. Highway Research Record, No. 207, 42-51.
[33] Engineering and Physical Sciences Research Council (n.d) Framework for Responsible Innovation. https://www.epsrc.ac.uk/research/framework/ on April 10, 2015.
[34] Erlander, S. (1980). Optimal Spatial Interaction and the Gravity Model. Lecture Notes in Economics and Mathematical Systems. Lecture Notes in Economics and Mathematical Systems. Volume 173. Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.
[35] European Commission (2015). Digital Agenda for Europe: A Europe 2020 Initiative: European Innovation Partnership (EIP) on Smart Cities and Communities. Accessed from https://ec.europa.eu/digital-agenda/en/smart-cities on August 1, 2015.
[36] Evans, E., C.E. Grella, D. A. Murphy and Y-I. Hser (2010). Using Administrative Data for Longitudinal Substance Abuse Research. The Journal of Behavioral Health Services and Research, Vol. 37(2), pp. 252–271.
[37] Evans, Tom P., and Hugh Kelley. "Multi-scale analysis of a household level agent-based model of landcover change." Journal of Environmental Management 72.1 (2004): 57-72.
[38] Federal Interagency Forum on Aging-Related Statistics (2010). Older American 2010: Key Indicators of Well-Being. Accessed from http://www.agingstats.gov/agingstatsdotnet/Main_Site/Data/2010_Documents/Docs/OA_2010.pdf July 31, 2010.
[39] Foth, M., J. H. Choi, and C. Satchell (2011). Urban Informatics. Proc. of the ACM 2011 conference on Computer supported cooperative work (CSCW '11). ACM, New York, NY, USA, 1-8.
[40] Fujita, M. (1988). A monopolistic competition model of spatial agglomeration: Differentiated product approach. Regional Science and Urban Economics, Vol.18(1), 87-124.
[41] Fujita, M. and H. Ogawa (1982). Multiple equilibria and structural transition of non-monocentric urban configurations. Regional Science and Urban Economics, Vol. 12(2), 161-196.
[42] Fujita, M., P. Krugman and A. J. Venables (1999). The spatial economy: Cities, regions, and international trade. MIT Press, Cambridge, MA.
[43] Ghosh, D. and Rajarshi Guha (2013). What are we tweeting about obesity? Mapping tweets with topic modeling and geographic information system. Cartography and Geographic Information Science. Vol 40(2), 90–102.
[44] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12), 7821-7826.
[45] Gjoka, M., M. Kurant, C. T. Butts and A. Markopoulou (2010). Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. Proc. IEEE 2010 INFOCOM 2010.
[46] Golbeck, J., and D. Hansen. (2013). A method for computing political preference among Twitter followers. Social Networks, Vol. 36, 177-184.
[47] Gowans, H., Elliot, M., Dibben, C. and Lightfoot, D. (2012) Accessing and sharing administrative data and the need for data security, Administrative Data Liaison Service.

[48] Gruteser, M. and D. Grunwald (2003). Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In Proc. 1st international conference on Mobile systems, applications and services, MobiSys '03, pages 31–42.

[49] Gurstein, M. (2011). Open data: Empowering the empowered or effective data use for everyone? First Monday, Vol 16, No 2 - 7 February. Accessed from http://firstmonday.org/ojs/index.php/fm/article/view/3316/2764 on July 1, 2013.

[50] Haque, U. (2012). Surely there's a smarter approach to smart cities? Wired Magazine. April 17, 2012. Accessed from http://www.wired.co.uk/news/archive/2012-04/17/potential-of-smarter-cities-beyond-ibm-and-cisco on April 10, 2012.

[51] Harland, K., A. Heppenstall, D. Smith and M. Birkin (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. Journal of Artificial Societies and Social Simulation, Vol. 15 (1) 1. Accessed from http://jasss.soc.surrey.ac.uk/15/1/1.html on April 19, 2015.

[52] Huang, D. L., D. E. Rosenberg, S. D. Simonovich, and B. Belza (2012). Food Access Patterns and Barriers among Midlife and Older Adults with Mobility Disabilities. In Journal of Aging Research, Vol 2012, pp. 1-8.

[53] Hendler, J. (2014). Data Integration for Heterogeneous Datasets. , BD205, Vol 2(4).

[54] Hey, T., S. Tansley, and K. Tolle (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond, WA: Microsoft Research.

[55] Hoadley, C. M. and P. Bell (1996). Web for Your Head: The Design of Digital Resources to Enhance Lifelong Learning. D-Lib Magazine, September. Accessed from http://www.dlib.org/dlib/september96/kie/09hoadley.html on Jan 15, 2015.

[56] Hollands, R. G. (2008). Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? City: analysis of urban trends, culture, theory, policy, action. Vol. 12(3), 303-320.

[57] Huijboom, N. and T van den Broek. Open data: an international comparison of strategies. European Journal of ePractice, Vol 12, March/April 2011.

[58] Isard, W. (1956) Location and space-economy. MA: MIT Press, Cambridge.

[59] Jia, T., Jiang, B., Carling, K., Bolin, M., Ban, Y.F. (2012). An empirical study on human mobility and its agent-based modeling. In Journal of Statistical Mechanics: Theory and Experiment, P11024

[60] Kain, J. F. and J. J. Persky (1969). Alternatives to the "Gilded Ghetto". In The Public Interest, Winter, 77-91.

[61] Kempton, W and S. E. Letendre (1997). Electric vehicles as a new power source for electric utilities. In Transportation Research Part D: Transport and Environment, Vol. 2(3), pp. 157–175.

[62] Kinney, R., Crucitti, P., Albert, R., & Latora, V. (2005). Modeling cascading failures in the North American power grid. The European Physical Journal B-Condensed Matter and Complex Systems, 46(1), 101-107.

[63] Kowald, M. and K.W. Axhausen (2015) Social Networks and Travel Behaviour, Ashgate, Burlington.

[64] Kinsley, S (n.d.) A political economy of Twitter data? Conducting research with proprietary data is neither easy nor free. Accessed from http://blogs.lse.ac.uk/impactofsocialsciences/2014/12/30/a-political-economy-of-twitter-data/ on April 1, 2015.

[65] Knoll, M. (2008). Diabetes City: How urban game design strategies can help diabetics. In eHealth'08, pp. 200–204.

[66] Krivo, L. J. and R. D. Peterson (1996). Extremely disadvantaged neighborhoods and urban crime. In Social Forces, Vol.75, 619-648.

[67] Krugman, Paul (1991). Geography and Trade. Cambridge, MA: MIT Press.

[68] Lee, D. B., Jr. (1973). Requiem for Large-Scale Models. Journal of the American Institute of Planners, Vol. 39, 163-78.

[69] Levinson, D., B. Marion and M. Iacono (2010). Access to Destinations, Phase 3: Measuring Accessibility by Automobile. Accessed from http://www.cts.umn.edu/Research/ProjectDetail.html?id=2009012

[70] Lightfoot, D. and Dibben, C. (2013) Approaches to linking administrative records to studies and surveys - a review, Administrative Data Liaison Service, University of St Andrews. Accessed from https://www.google.co.uk/search?q=Approaches+to+linking+administrative+records+to+studies+and+surveys+-+a+review&ie=utf-8&oe=utf-8&gws_rd=cr&ei=rfxFVZWeM4KUaumTgdAC# on March 15, 2015.

[71] Lin, Y., J. Jessurun, B de Vries, and H. Timmermans (2011). Motivate: Towards context-aware recommendation mobile system for healthy living. In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on, pages 250–253.

[72] Liu, B. (2007). Web Data Mining: Data-Centric Systems and Applications. Springer.

[73] Lowry, I.S. (1964). A Model of Metropolis. The Rand Corporation, Santa Monica.

[74] Ludwig, Y., B. Zenker and J. Schrader (2009). Recommendation of personalized routes with public transport connections. In D. Tavangarian, T. Kirste, D. Timmermann, U. Lucke, and D. Versick, editors, Intelligent Interactive Assistance and Mobile Multimedia Computing, Vol. 53 of Communications in Computer and Information Science, 97–107, Springer.

[75] Lundquist, D. (2011). Pollution credit trading in Vehicular Ad Hoc Networks. http://connected.vehicle.challenge.gov/submissions/2926-pollution-credit-trading-in-vehicular-ad-hocnetworks,

[76] Marshall, Alfred (1920). Principles of Economics. London, UK: MacMillan and Co.

[77] Miller, H. and M. F. Goodchild (2014). Data-driven geography. GeoJournal, October, 1-13.

[78] Mislove, A., S. Lehmann, Y. Ahn, J. Onnela, and J. N. Rosenquist. (2011). Understanding the demographics of twitter users. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11), Barcelona, Spain.

[79] Moor, J. H. (2005). Why we need better ethics for emerging technologies. Vol. 7, 111-119.

[80] Museum of the City of New York (2014). Museum of the City of New York receives grant from the National Endowment for the Humanities to process, catalog, digitize and rehouse the Ephemera Collections. Accessed from http://www.mcny.org/sites/default/files/Press_Release_NEH_Grant_FINAL.pdf on April 5, 2015.

[81] Muth, R. F. (1969). Cities and Housing: The Spatial Pattern of Urban Residential Land Use. University of Chicago Press.

[82] NAREC Distributed Energy (2013). ERDF Social Housing Energy Management Project - Final Project Report. UK National Renewable Energy Centre. Accessed from https://ore.catapult.org.uk/documents/10619/127231/Social%20Housing%20final%20report/6ca05e01-49cc-43ca-a78c-27fe0e2dd239 on April 1, 2015.

[83] Naylor, J., N. Swier, S. Williams, K. Gask and R. Breton (2015). ONS Big Data Project – Progress report: Qtr 4 October to Dec 2014. ONS Big Data Project Qtr 4 Report. Accessed from http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/the-ons-big-data-project/index.html on April 15, 2015.

[84] OECD Global Science Forum (2013). New Data for Understanding the Human Condition: International Perspectives. Report on Data and Research Infrastructure for the Social Sciences. Accessed from http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf on April 1, 2015.

[85] Office for National Statistics (2015). Beyond 2011 research strategy and plan – 2015 to 2017. Accessed from http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/research-strategy-and-plan---2015-2017.pdf on March 1, 2015.

[86] Organisation for Economic Co-operation and Development (OECD) Global Science Forum (2011). Effective Modelling of Urban Systems to Address the Challenges of Climate Change and Sustainability. October. Accessed from www.oecd.org/sti/sci-tech/49352636.pdf on April 13, 2013.

[87] Ortega, F., J. Gonzalez-Barahona and Robles, G. (2008). On the inequality of contributions to Wikipedia. HICSS '08 Proc. 41st Annual Hawaii International Conference on System Sciences., 304.

[88] Picciano, A. G. (2012). The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks, Vol. 16(3), 9-20.
[89] Pietsch, W. (2013) Big Data – The New Science of Complexity. In 6th Munich-Sydney-Tilburg Conference on Models and Decisions (Munich; 10-12 April 2013). Accessed from http://philsci-archive.pitt.edu/9944/ on April 1, 2015.
[90] Poplin, A. (2014). Digital serious game for urban planning: B3—Design your Marketplace! Environment and Planning B: Planning and Design, Vol 41(3), 493 – 511.
[91] Pyrozhenko, V. (2011). Implementing Open Government: Exploring the Ideological Links between Open Government and the Free and Open Source Software Movement. Prepared for 11th Annual Public Management Meeting.
[92] Quinn, A. J. and B. B. Bederson (2011). Human Computation: A Survey and Taxonomy of a Growing Field. In Proc. Annual Conference on Human factors in computing systems (CHI '11), 1403–1412.
[93] Rae, A. (2014). Online Housing Search and the Geography of Submarkets. Housing Studies, Vol 30(3), pp. 453-472.
[94] Rand, W., Herrmann, J., Schein, B. and Vodopivec, N. (2015). An Agent-Based Model of Urgent Diffusion in Social Media. In Journal of Artificial Societies and Social Simulation, Vol 18(2), 1.
[95] Raykar, V.C., S. Yu, L.H. Zhao, G.H. Valadez, C. Florin, L. Bogoni, and L. Moy (2010). Learning from crowds. Journal of Machine Learning Research 11, 1297-1322.
[96] Reinhardt, J., J. Miller, G. Stucki, C. Sykes and D. Gray (2011). Measuring Impact of Environmental Factors on Human Functioning and Disability: A Review of Various Scientific Approaches. Disability and Rehabilitation, Vol 33(23-24), 2151-2165.
[97] Reist, B. H. and A. Ciango (2013) Innovations in Census Taking for the United States in 2020. Proc. 59th ISI World Statistics Congress, Hong Kong. Accessed from http://www.statistics.gov.hk/wsc/STS017-P2-S.pdf on March 15, 2015.
[98] Richardson, H. W. (1977). A Generalization of Residential Location Theory. Regional Science and Urban Economics, Vol. 7, 251-66.
[99] Robinson, J. (2014). New Geographies of Theorizing the Urban: Putting Comparison to Work for Global Urban Studies. In The Routledge Handbook on Cities of the Global South, edited by S. Parnell and S. Oldfield, New York: Routledge, pp. 57–70.
[100] Roy, A. (2009). The 21st Century Metropolis: New Geographies of Theory. In Regional Studies, Vol. 43(6), pp. 819–830.
[101] Salasznyk, P. P., E. E. Lee, G. F. List and W. A. Wallace (2006). A systems view of data integration for emergency response. International Journal of Emergency Management, Vol 3(4), 313 – 331.
[102] Salvini, Paul, and Eric J. Miller. "ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems." Networks and Spatial Economics 5.2 (2005): 217-234.
[103] Sasaki, K., S. Nagano, K. Ueno and K. Cho (2012). Feasibility Study on Detection of Transportation Information Exploiting Twitter as a Sensor. Sixth International AAAI Conference on Weblogs and Social Media. Workshop on When the City Meets the Citizen, AAAI Technical Report WS-12-0
[104] Schelling, Thomas C. "Dynamic models of segregation†." Journal of mathematical sociology 1.2 (1971): 143-186.
[105] Sen, A. and T.E. Smith (1995). Gravity Models of Spatial Interaction Behavior. Advances in Spatial and Network Economics Series, Springer-Verlag, Berlin Heidelberg NY.
[106] Sheffi, Y. (1985). Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. Prentice-Hall, Inc.
[107] Shkabatur, J. (2013) Transparency With(out) Accountability: Open Government in the United States (March 25, 2012). Yale Law & Policy Review, Vol. 31, No.
[108] Smith, R. J. (2014). Missed miracles and mystical connections: qualitative research and digital social sciences and big data. In Big Data? : Qualitative Approaches to Digital Research, edited by Hand, M. and S. Hillyard. Edward Group Publishing Limited, pp. 181-204. Retrieved from http://www.eblib.com on August 1, 2015

[109] Tang, K. P., J. Lin, J. I. Hong, D. P. Siewiorek, and N. Sadeh (2010). Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In Proc. 12th ACM International Conference on Ubiquitous computing (Ubicomp '10), 85–94.
[110] Thakuriah, P., and G. Geers (2013) Transportation and Information: Trends in Technology and Policy. Springer: New York.
[111] Thakuriah, P., and N. Tilahun (2013) Incorporating weather information into real-time speed estimates: comparison of alternative models. Journal of Transportation Engineering, Vol. 139(4), 379-389.
[112] Thakuriah, P., L. Dirks and Y. Mallon-Keita (forthcoming). Emerging Urban Digital Infomediaries and Civic Hacking in an Era of Big Data and Open Data Initiatives. Proc. of 2015 NSF Big Data and Urban Informatics Workshop.
[113] Thakuriah, P., Persky, J., Soot, S., and Sriraj, P. (2013) Costs and benefits of employment transportation for low-wage workers: An assessment of job access public transportation services. Evaluation and Program Planning, 37, pp. 31-42.
[114] Thakuriah, P., and Mallon-Keita, Y. (2014) An analysis of household transportation spending during the 2007-2009 US economic recession. In: Transportation Research Board 93rd Annual Meeting, Washington D.C., USA, 12-16 Jan 2014.
[115] Thakuriah, P., Soot, S., Cottrill, C., Tilahun, N., Blaise, T., and Vassilakis, W. (2011) Integrated and continuing transportation services for seniors: case studies of new freedom program. Transportation Research Record, 2265, pp. 161-169
[116] Thanos, C. and A. Rauber (2015). Scientific Data Sharing and Re-Use. ERCIM News, Number 100, Jan., 13.
[117] Thorhildur, J. Avital, M., and N. BjÃ¸rn-Andersen (2013). The Generative Mechanisms Of Open Government Data. ECIS 2013 Proceedings. Paper 179.
[118] Tilahun, Nebiyou, and David Levinson. "An Agent-Based Model of Origin Destination Estimation (ABODE)." Journal of Transport and Land Use 6.1 (2013): 73-88.
[119] Townsend, A. (2013). Smart Cities: Big Data, Civic Hackers and the Quest for a New Utopia. W. W. Norton and Co, New York.
[120] Tukey, J. W. (1980). We Need Both Exploratory and Confirmatory. The American Statistician. Vol. 34(1), 23-25.
[121] Urban Big Data Center n.d. Integrated Multimedia City Data (iMCD). Accessed from http://ubdc.ac.uk/our-research/research-projects/methods-research/integrated-multimedia-city-data-imcd/ on January 15, 2015.
[122] U.S. Census Bureau (2012). Press release, May 22, 2012, available at https://www.census.gov/2010census/news/releases/operations/cb12-95.html.
[123] United Nations, Department of Economic and Social Affairs, Population Division (2014). World Urbanization Prospects: The 2014 Revision, Highlights (ST/ESA/SER.A/352).
[124] Vitrano, F. A. and M. M. Chapin (2010). Possible 2020 Census Designs and the Use of Administrative Records: What is the impact on cost and quality?" Suitland, Maryland: U.S. Census Bureau, available at https://fcsm.sites.usa.gov/files/2014/05/Chapin_2012FCSM_III-A.pdf.
[125] Vitrano, F. A. and M. M. Chapin (2014). Possible 2020 Census Designs and the Use of Administrative Records: What is the impact on cost and quality? Accessed from https://fcsm.sites.usa.gov/files/2014/05/Chapin_2012FCSM_III-A.pdf on March 1, 2015.
[126] von Ahn, L., B Maurer, C. McMillen, D. Abraham, and M. Blum. (2008) reCAPTCHA: Human-based character recognition via web security measures. Science, 321(5895), 1465–1468.
[127] von Ahn, L., M. Blum, N. J. Hopper and J. Langford. (2003). CAPTCHA: Using hard AI problems for Security. Technical, Report 136.
[128] Wauthier, F.L., and M.I. Jordan (2011). Bayesian bias mitigation for crowdsourcing. Pp. 1800-1808 in Proceedings of the Conference on Neural Information Processing System, Number 24. Available at http://machinelearning.wustl.edu/mlpapers/papers/NIPS2011_1021.
[129] Wegener, M. (1994) Operational Urban Models State of the Art, Journal of the American Planning Asociation, Vol. 60(1), 17-29.

[130] Weil, R., J Wootton, and A Garca-Ortiz (1998). Traffic incident detection: Sensors and algorithms. Mathematical and Computer Modelling, Vol. 27(911):257–291.
[131] Williams, J. D. (2012). The 2010 Decennial Census: Background and Issues. Congressional Research Service R40551. Accessed from http://www.fas.org/sgp/crs/misc/R40551.pdf on March 1, 2015.
[132] Wilson, A. G. (1971). A family of spatial interaction models, and associated developments. Environment and Planning, Vol. 3(1) 1 – 32.
[133] Wilson, A. G. (2013). A modeller's utopia: Combinatorial evolution. Commentary. Environment and Planning A, Vol. 45, 1262 – 1265.
[134] Wilson, W. J. (1987). The Truly Disadvantaged: The Inner City, the Underclass and Public Policy Chicago, University of Chicago Press.
[135] Wu, G., S. Talwar, K. Johnsson, N. Himayat, and K. D. Johnson (2011). M2m: From mobile to embedded internet. Communications Magazine, IEEE, 49(4):36–43, April.
[136] Zellner, M.L., Lyons, L.; Hoch, C. J.; Weizeorick, J.; Kunda, C.; Milz, D. (2012). "Modeling, Learning and Planning Together: An Application of Participatory Agent-Based Modeling to Environmental Planning." URISA Journal, GIS in Spatial Planning Issue, 24(1): 77-92.
[137] Zellner, M.; Massey, D.; Shiftan, Y.; Levine, J.; Arquero, M. (forthcoming) "Overcoming the last-mile problem with transportation and land-use improvements: An agent-based approach" International Journal of Transportation, Special Issue on Agent-Based Modeling in Transportation Planning and Operations.
[138] Zellner, M. L.; Page, S. E.; Rand, W.; Brown, D. G.; Robinson, D. T.; Nassauer, J.; Low, B. (2009). "The Emergence of Zoning Games in Exurban Jurisdictions." Land Use Policy 26 (2009): 356-367.
[139] Zellner, M.L., Reeves, H. W. (2012). "Examining the contradiction in 'sustainable urban growth': An example of groundwater sustainability." Journal of Environmental Planning and Management, 55(5): 545-562.
[140] Zhang, X., S. Qin, B. Dong and B. Ran (2010). Daily OD matrix estimation using cellular probe data. Proc. 9th Annual Meeting Transportation Research Board.
[141] Zheng, Y. and X. Xie (2011). Location-Based Social Networks: Locations. In Y. Zheng and X. Zhou, eds, Computing with Spatial Trajectories, 277–308, Springer New York.