1: Questioning the City through Urban Analytics

Learning Objectives

By the end of this chapter students will understand the following:

- The majority of future population growth will be concentrated in urban areas.
- The planning and management of population change creates a range of challenges for cities.
- New forms of technology are increasingly embedded into city systems and are providing a wealth of new data.
- Urban analytics represents a methodological toolkit for studying and managing data-rich cities.

Human Settlements and Urbanization

Looking down onto the surface of the earth from space reveals a patchwork of urbanization. By 2015 the United Nations (UN) estimated that global population reached 7.3 billion people, and projected that this will have grown to 9.7 billion by 2050 (United Nations 2015). Not only has the world's population grown, but the distribution of people has changed over time, and at some point in 2009 we reached a significant tipping point where more than half of the global population lived within urban as opposed to rural areas (United Nations 2009). This transition to urbanized living is driven by complex economic, technological, cultural, and geopolitical forces. Global urbanization has fundamentally changed how we (as a species) interact with one another and our natural environment, creating a range of challenges for the planning and management of urban areas.

By night, the scale and extent of urban areas can be rendered visible through satellite imagery that records the use of electric lighting (see Figure 1.1). However, the social, economic, and environmental impacts of these brightly lit places extend far beyond these borders. From space, those dark areas between city lights seem empty and undifferentiated. One can imagine these dark spaces to be a patchwork of agricultural and natural areas that are disconnected from the networks of lights. However, much of the globe, even these dark places, are linked together through a network of ecological and economic exchanges that fuel and support urbanization. Much of this network is material – a vast communications infrastructure that, while concentrated in cities, spans the entire globe.

What has fueled this growth? There is not one story – the forces that drive urbanization in the more developed parts of the world are different from the forces driving urbanization in the global South. Generically, residents of cities are attracted by the advantages of proximity to sources of employment, infrastructure, and cultural assets, or perhaps as a result of improved provision of healthcare or sanitation. The net result of this growth of cities is that a large share of the world's population is connected to the economic and technological infrastructure that emanates from them.

The phenomenon of urbanization is not new, nor are those challenges emerging from these processes. However, a critical difference between urbanization of the twenty-first century and the waves of urbanization that have occurred in the past is that there are entirely new ways to understand these processes. The same information technologies that connect cities in a global network can also be used within them to manage the provision of services, and to mitigate the environmental impact of their metabolism within and beyond their borders. New ways of knowing and managing cities are occurring because of advances in those instruments that can monitor activities within or attributes of urban environments (see Chapter 2). Enhancements to communications infrastructure are enabling the data generated by such devices to be utilized by services in real time, and for devices to communicate with one another, potentially making automated decisions based upon derived information.

Figure 1.1 At night, urbanization is rendered visible on the earth's surface through a satellite originally designed to detect cloud coverage

Source: Data courtesy Marc Imhoff of NASA GSFC and Christopher Elvidge of NOAA NGDC. Image by Craig Mayhew and Robert Simmon, NASA GSFC

The macro shape, structure, and function of urban areas emerge through an incredibly complex set of human interactions; cities are dynamic systems that evolve from the bottom up and over time (Batty 2013). Currently the most urbanized regions of the world include North America, Europe, Latin America, and the Caribbean (see Table 1.1); however, by 2050, it is estimated that 37 percent of new projected urban population growth will be attributable to just three countries in Africa and Asia, namely, China, India, and Nigeria (United Nations 2015).

Table 1.1 Regions of the world vary significantly by their levels of urbanization (United Nations 2015)

Region	Urban	Rural	Total	Percentage urban
Africa	455,345	682,885	1,138,229	40.0
Asia	2,064,211	2,278,044	4,342,255	47.5
Europe	545,382	197,431	742,813	73.4
Latin America and the Caribbean	495,857	127,565	623,422	79.5
North America	291,860	66,376	358,236	81.5
Oceania	27,473	11,356	38,829	70.8
World	3,880,128	3,363,656	7,243,784	53.6

Note: Population is in thousands.

Source: United Nations, World Urbanization Prospects, 2015

Note: Population is in thousands.
Source: United Nations, World Urbanization Prospects, 2015

However, the global narrative of increasing urbanization does not hold everywhere. While some cities expand rapidly, others are shrinking in response to macroeconomic trends, like the decline of manufacturing in many developed countries, environmental disasters (Box 1.1), and/or political instability. Urban areas require careful management, planning and investment if the negative societal impacts that can be associated with rapid or long-term change are to be avoided or mitigated. For cities with declining population, sustaining infrastructure (e.g., roads or schools) that was designed to service much larger populations can strain municipal finances and make it difficult to provide basic services like education and safety. Conversely, in a place with rapid population growth, how might public transit systems be reconfigured to meet this extra demand given land use constraints, such as increasing the density of people's homes located at the site of a potential train stop, that might prevent the expansion of the transit network?

Governing cities is a complex and political process. While the new forms of data about cities do not ease the political burden, they can feed into decision-making processes and help to manage existing infrastructure more efficiently. Decisions about cities are complex, involving difficult trade-offs. However, there are a wealth of often disparate empirical data, a suite of methods or tools for translating this into information, and a mechanism for the communication of findings to stakeholders (Longley et al. 2015b). This book provides a background to this process.

Box 1.1: Change in Urban Areas – Hurricane Katrina and Information Technology

Hurricane Katrina struck the Gulf Coast of the United States in August 2005, and had devastating impacts on the population, infrastructure, and economy of the region. The city of New Orleans, Louisiana was acutely affected after a number of levees failed, leaving 80 percent of the city flooded under 15 to 20 feet (4.5 to 6 m) of water (Figure 1.2).

Figure 1.2 Searching for survivors in New Orleans

Source: US Coast Guard photo by Petty Officer 2nd Class NyxoLyno Cangemi (Wikipedia)

The total costs of Hurricane Katrina were catastrophic, with the National Oceanic and Atmospheric Administration (NOAA) estimating that there were 1,353 direct fatalities, 275,000 homes damaged or destroyed, along with financial costs in excess of $100 billion (Johnson 2006). With such significant impact to housing infrastructure and property, this resulted in a huge number of people being displaced into new locations. Figure 1.3 is not a direct proxy of these movements, but illustrates change at a state level after this event.

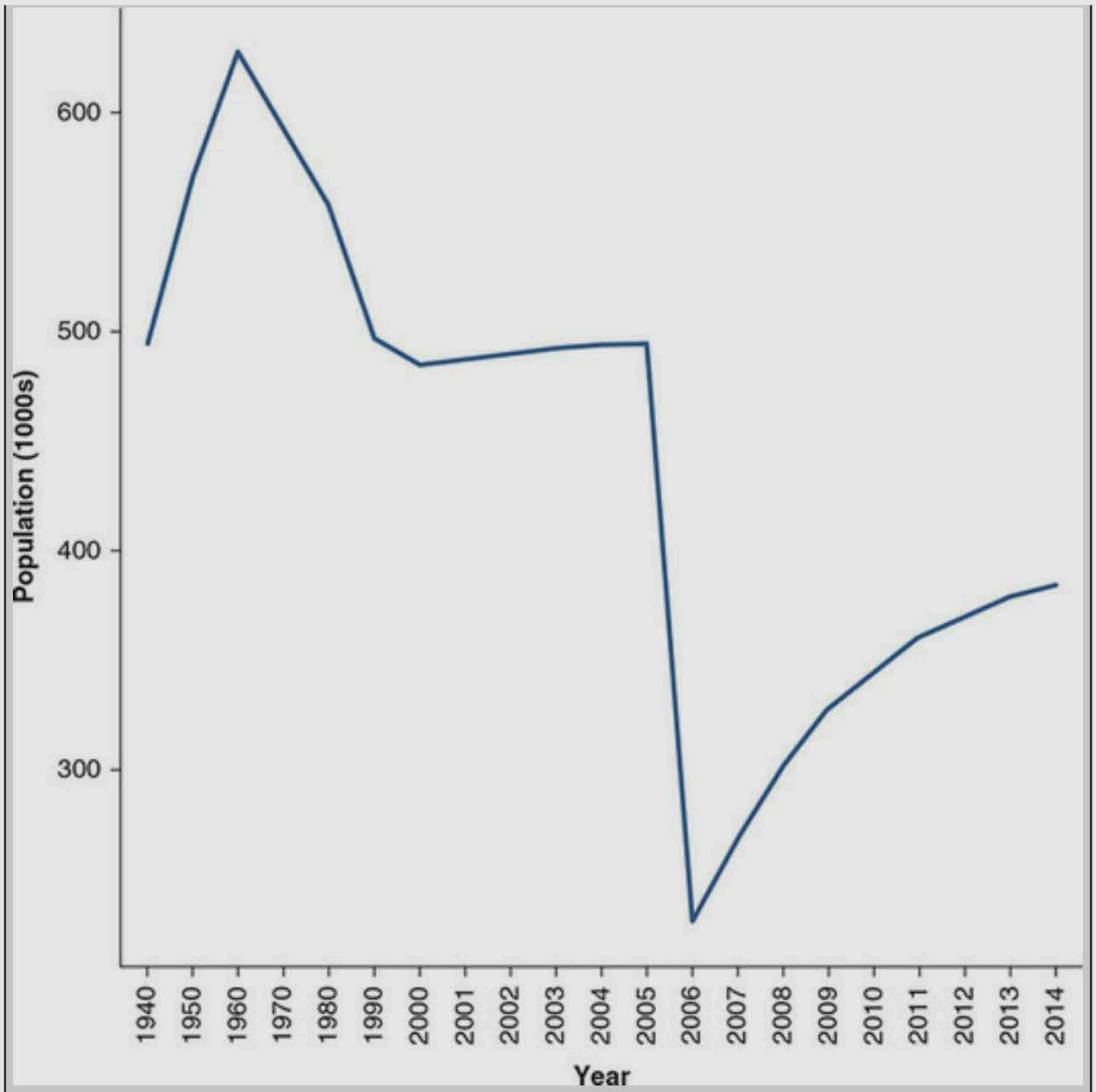
Figure 1.3 States where migrants housed within Louisiana during 2005 had moved to as of 2006

Source: Authors' own; data – US Census Bureau, American Community Survey 2006; State to State Migration Flows 2004–2015

The impact of Hurricane Katrina on New Orleans illustrates the dynamic nature of urban areas, and indeed, over a decade on from the event, population levels have not recovered fully – although, even prior to the hurricane, the population had also been in decline from a peak in the 1970s (Figure 1.4).

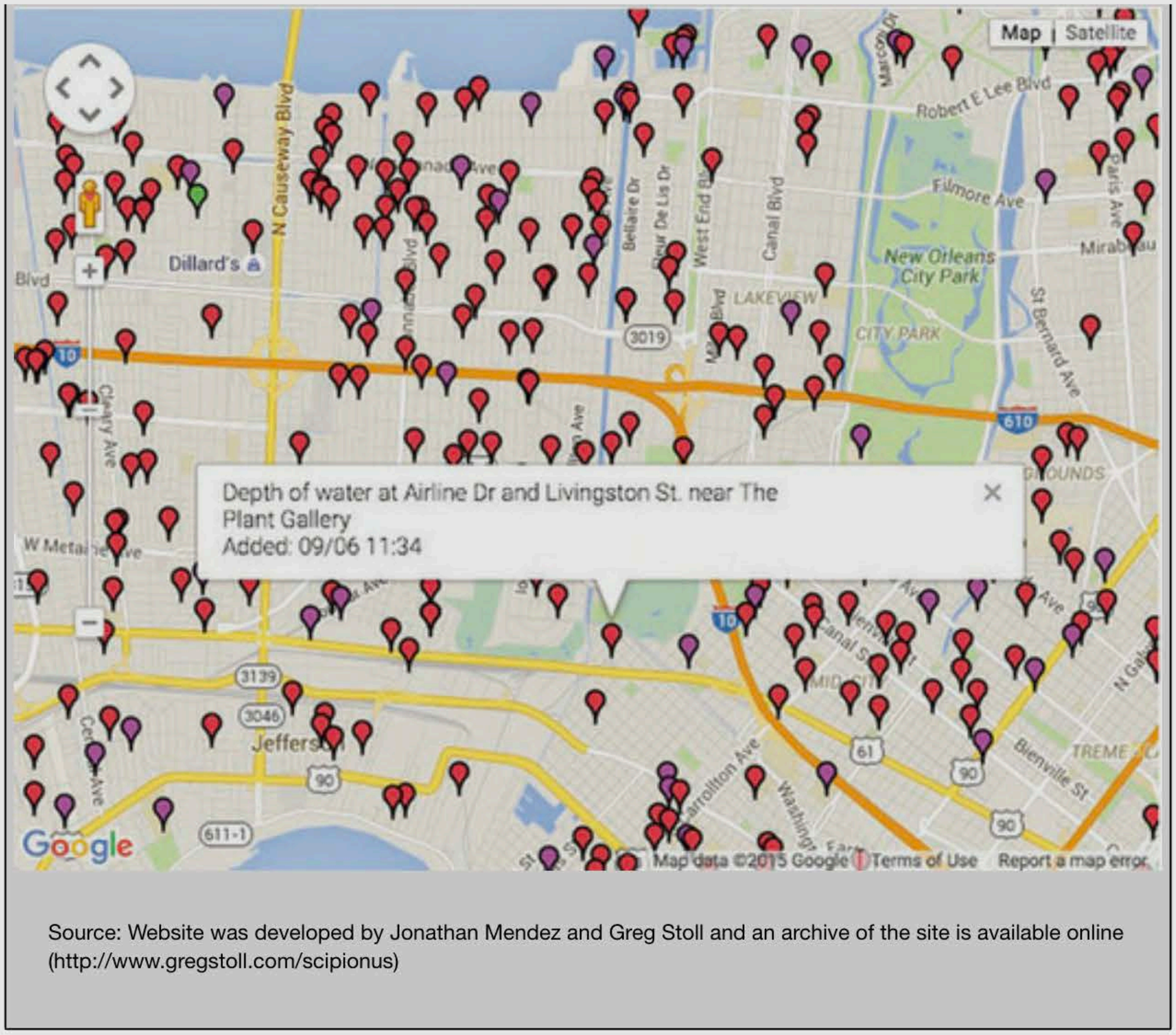
In the aftermath of Hurricane Katrina, the then nascent technologies from Google, including its Maps and Earth platforms, were used by both official bodies and the public to collect or disseminate spatial data related to the event. This included the loading of high-resolution aerial imagery into Google Earth by NOAA, and additionally an early grassroots example of a map "mashup" where survivors posted details of their locations or other useful information (Figure 1.5). The relationship between technology platforms, urban areas, and their inhabitants is an enduring feature of contemporary urban living, and a key facet of urban analytics relates to the use of these new spatial data infrastructures for decision making.

Figure 1.4 Population change in New Orleans between 1950 and 2014



Source: Authors' own; data – US Census Bureau, Population of the Largest 100 Urban Places 1940–2016

Figure 1.5 An early map "mashup" created in the aftermath of Hurricane Katrina



Source: Website was developed by Jonathan Mendez and Greg Stoll and an archive of the site is available online (http://www.gregstoll.com/scipionus)

Urban Data Systems

New York City (NYC) is situated on the north-eastern seaboard of the United States of America, and is the country's most populated city, with the 2010 Census recording over 8 million residents. As with many large metropolitan areas across the globe, technology and data are increasingly embedded into the fabric of the urban area making up NYC.

Figure 1.6 shows a photograph of a fairly typical street scene occurring within Times Square, which is located within the Mid-Town district on the island of Manhattan. People can be seen intermingling with traffic, and theaters, cinemas, and shops line the pedestrianized area, with residential and office towers dominating the skyline. However, beyond the obvious visible features, various technologies infiltrate multiple aspects of urban living and mobility.

Way beyond the visible range of the optics used on the camera that took this photograph, and at approximately 20,000 km above the surface of the earth, a constellation of Global Positioning System (GPS) satellites orbit. Back in Times Square, and among the crowd of pedestrians, a family of tourists has become disorientated; fortunately, however, they have a GPS-enabled cell phone which combines information returned from the satellites with data from a technology company in California to calculate their location. This is then plotted on a map, and routing functionality assists them to navigate to their intended destination. The maps and route calculation

are provided over a wireless Internet connection which is also being used by hundreds of other people within this area. Many of these people have security settings on their cell phones that enable the sharing of locational information collected by a range of technology companies. One provider of free public maps collects this data, pooling the locations, times, and speeds of users within this area, and uses these attributes in a model that predicts traffic congestion. Based on this real-time information, the taxi cab just pulling out of the photograph has rerouted to avoid additional traffic that will cause delay to its current fare.

On top of a number of lampposts are security cameras streaming live video over the Internet. A shop owner, watching the video, identifies an instance of petty theft and the NYC Police are dispatched. The perpetrator is arrested with the location, time, and offense recorded. These attributes are later integrated into extracts placed on the NYC Open Data store. Researchers at a university then use these data to develop a statistical model that estimates where and when crime is most likely to occur, which later feeds into a revised policing strategy report.

Many of the people in the photograph are holding shopping bags, having made purchases from the stores running along the side of Times Square, and typically have paid for these goods using credit cards. After the transactions are processed, records of these purchases are added to the flow from millions of other transactions by consumers from across the United States, with this data being stored by the credit card company. The updated records feed dynamically into the consumer segmentation models which are then provided as a service to third-party marketing agencies.

Collectively all of these small interactions can have a profound impact on the state of a city at any given moment in time. Consider that taxi, using a routing application to avoid traffic. If 10 percent of drivers are using such an application in an area, a significant portion of traffic will be redirected from one part of the city to another. However, such routing applications optimize a single person's route, but when lots of people are using them they have an unobservable impact on the fabric of a city. The wealth of data generated within or used by urban areas is not restricted to these brief examples; however, they aim to be illustrative of the ways in which data flows through urban systems and may be repurposed. These small technology-mediated interactions have a growing role in the shape of urban life. Increasingly cities are directly supporting digital infrastructure as they would traditional "hard" infrastructure like roads. In NYC, for example, public pay telephones are being replaced with LinkNYC (www.link.nyc), which will provide a digital communications infrastructure, including support for navigation, Wi-Fi access, and device charging (Figure 1.7).

Figure 1.6 Times Square, NYC: technology is integral to how cities can be managed and studied

Source: Photograph by Aurelien Guichard CC BY-SA (Flickr)

Figure 1.7 LinkNYC Station – reimagination of the telephone booth

Source: Seth Spielman

Box 1.2: Urban Big Data

When talking about cities it is difficult to escape the term Big Data, which has become commonplace in the description of a particular aspect of the evolving data economy and its links to servicing infrastructure. There are numerous definitions of Big Data (Kitchin and McArdle 2016); however, many have typically made reference to three main “Vs” which include: volume, velocity, and variety, although these are increasingly supplemented by other characteristics (see Table 1.2). Not all data related to urban areas are Big Data, and the majority of data featured within this book are typically large in size, but do not fit a formal definition of Big Data.

Table 1.2 The characteristics of Big Data

Characteristic	Description
Volume	Huge in *volume*, consisting of terabytes or petabytes of data
Velocity	High in *velocity*, being created in or near real time
Variety	Diverse in *variety* in type, being structured and unstructured in nature, and often temporally and spatially referenced
Exhaustive	*Exhaustive* in scope, striving to capture entire populations or systems (n = all), or at least much larger sample sizes than would be employed in traditional, small data studies
Resolution	Fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely indexical in identification
Relational	Containing common fields that enable *relational* joins to different datasets
Flexible	Structured in a *flexible* way that enables new fields to be easily added
Scalable	Fully *scalable* and can be expanded rapidly

Source: Adapted from Kitchin (2014: 68)

Source: Adapted from Kitchin (2014: 68)

Multiple factors have accelerated the availability of data within an urban setting, and this creates a demand for professionals trained in analytics and urban planning to effectively use such new resources. Until recently, access to urban data was often difficult; however, a trend toward more open licensing of data by cities enables use of municipal data without financial cost. The emergence of central repositories that consolidate data for a city facilitate discovery and download of these assets. An example of such a project is outlined in Profile 1.1.

Profile 1.1: Charlie Catlett

Figure 1.8 Charlie Catlett, Director, Urban Center for Computation and Data, University of Chicago

Source: Supplied by Charlie Catlett, image copyright © Eileen Moloney

Describe how you got interested in urban data science and your labs research

Over the longer term my research interests have concerned various aspects of the Internet, distributed and high-performance computing; however, about five years ago I became particularly interested in the rapid growth of cities, and in particular those huge investments happening and will happen over the next few decades in new urban infrastructure. So my interest started by asking if we have the right design tools and understanding of cities to develop infrastructure over the next 50 years that will be sustainable, and will correct some of our past mistakes. Through conversations with the administration of numerous cities over the course of a few years, it became clear to me that having a measurement system like the Array of Things (arrayofthings.github.io) would be useful to the science community to enable them to work with residents and city organizations or governments to address neighborhood challenges.

In general, within our center we have three areas of research related to cities' urban data science. The first concerns the development of tailored applications and tools that aim to make Open Data more accessible to the science community and to the general public. The second area is to develop new capabilities that use open and other data to better understand neighborhoods and cities, particularly in terms of their economic resilience, public safety, education, and health. This was motivated through early discussion with the City of Chicago Mayors Office who had interest in examining neighborhood trends over time; and specifically to identify where interventions may be necessary before a neighborhood reaches a crisis point. Finally, we have interests in embedding systems within cities to support research concerning the development of key capabilities in both measurement and actuation.

How can intelligent infrastructure help cities?

Intelligent infrastructure can really help by reducing the time frame within which you can make a decision, or examine how the result of that decision manifest. For many decisions within cities, such as the design of a new highway, park, or installation of a new rapid bus transit, their design and implementation may unfold over several years, and their impact on a city potentially over several decades. For such applications, historical data and simulation of potential outcomes are important; however, there are other areas that consider decisions for much higher temporal resolutions, for example,

traffic safety. Within these examples, embedded systems become important, and especially so when coupled with capability to make decisions in real time.

What is the Array of Things?

This started off as a connected sensor network project focused on the urban environment; however, it has evolved into a project about embedded capabilities to support research over multiple areas. Rather than try to develop a set of particular capabilities like controlling street lights or an application that tells you about air quality, we instead have focused on a general platform that allows for innovation and the rapid deployment of new hardware and software within the city. We also do this while making data available in forms and through mechanisms that will encourage people to innovate in terms of analytics and application development; but, and very importantly, in a way that preserves the privacy of individuals.

How can you reassure citizens about the collection and use of urban sensor data?

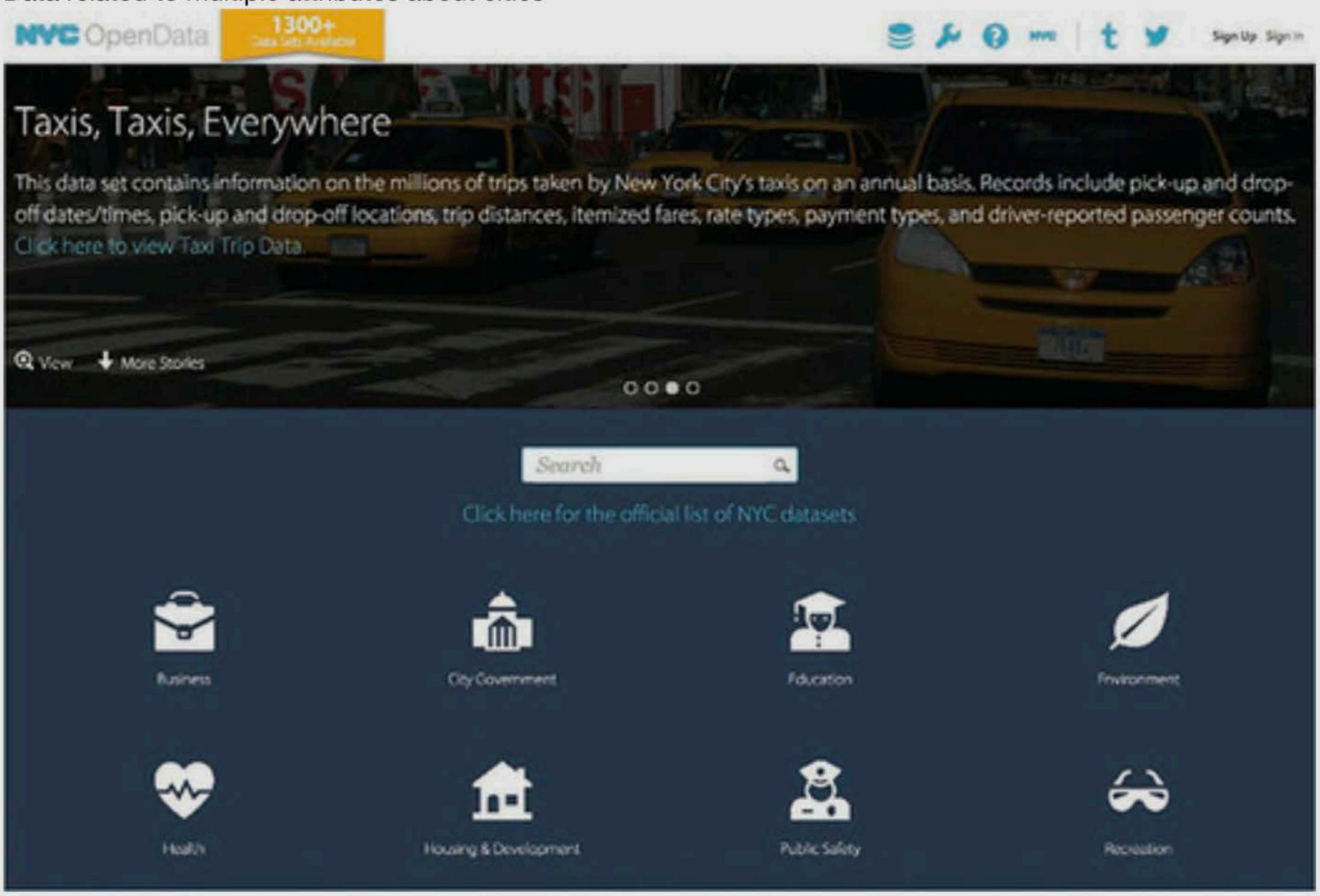
One of the things we have done was to develop a set of clear messages about the Array of Things project and take these out into public meetings where we have a chance to interact with people who are interested and can spend time answering questions. The bottom line for most people is that they want to know what you are doing with the data and how you are ensuring privacy. However, what we have learnt within Chicago is that clarity of communication and transparency, although really important, aren't entirely sufficient. As such, the third area that we have developed as part of our policies is structured and transparent accountability, that we ensure through an independent external privacy committee whose role is explicitly embedded within our governance policies. Where we wish to make changes to our underlying technology (e.g., a new sensor), we are required to make a report to the committee about the scientific justification for a change, the privacy implications that will come about because of them, and what actions we will take to continue to preserve privacy. They may have feedback or recommendations which are given to our governance committee, with these transactions made public.

These developments have required engagement from data owners such as government departments, who invest time in preparing data in formats that make them suitable for wider use. Given the cost of municipal data collection and curation, not all cities are able to offer such resources. In addition to city-to-city variation within countries, practices of municipal data dissemination are variable between countries. For example, within the United States there is a history of data collected by public agencies being placed into the public domain; for example, the decennial census of the population. However, this is not the case everywhere; within the UK, for example, prior to 2001, census data were only made available through a restricted set of commercial resellers, despite the collection of the data being publicly funded.

The arguments for making data open have both an ethical and a pragmatic dimension. Ethically, if the public have funded the collection of a particular dataset (e.g., through taxes), then it could be argued that this should also be made available publicly, with the caveat of appropriate disclosure controls or security constraints. Secondly, the release of Open Data has the potential to generate large economic or societal benefit. In 2013, the global consulting firm McKinsey estimated that worldwide there were approximately $3 trillion worth of value contained within Open Data (Manyika et al. 2013). The availability of Open Data has created new business opportunities, with organizations reselling value-added versions of the data, or integrating the data into new products or services. Transparency in the collection and dissemination of urban data is a good thing. When cities expose to citizens (and visitors) the kinds of information they collect, people are more informed about how their activities are monitored. However, simply making data available does not lead to its effective use. For the value identified by McKinsey to be materialized, creative and innovative use of data are necessary. Understanding how to effectively use urban data is a core goal of this book.

As the volumes of Open Data have grown, a number of Open Data platforms have emerged that provide search and discovery of resources. The two most prevalent systems implemented include CKAN (ckan.org) and Socrata (www.socrata.com), and platforms based on these systems have been developed for a variety of national extents (see Table 1.3), or within the context of specific regions or cities (see Figure 1.9).

Figure 1.9 NYC OpenData. Many cities across the world now have datastores that provide access to Open Data related to multiple attributes about cities



Source: https://nycopendata.socrata.com/. Reprinted with permission

Table 1.3 Some examples of national Open Data portals

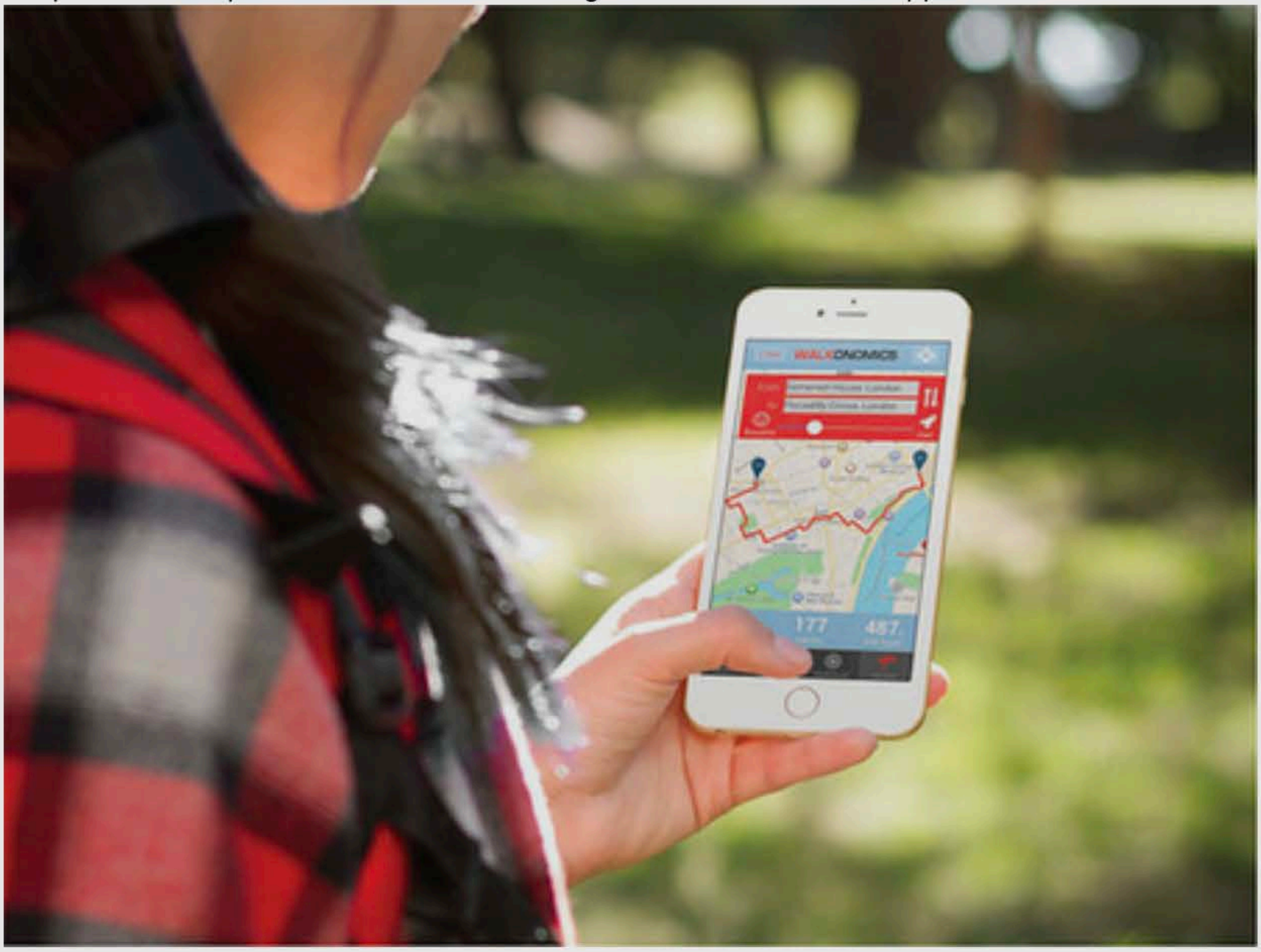
Country	Web address	Platform
UK	data.gov.uk	CKAN
Japan	www.data.go.jp	CKAN
Kenya	www.opendata.go.ke	Socrata
United States	data.gov	CKAN
India	data.gov.in	Open Government Platform

Practicing Urban Analytics

The widespread availability of urban data is a new phenomenon, and creates potential for a new field of inquiry called "urban analytics." Urban analytics is the practice of using new forms of data in combination with computational approaches to gain insight into urban processes. Increasing data availability allows us to ask new and often complex questions about cities, their economy, how they relate to the local and global environment, and much more. For example, something as mundane as adjusting the timing of traffic signals in response to

fluxes in vehicle volume derived though sensor networks could dramatically reduce carbon emissions in a city. Providing automated alerts to mobile devices if air quality were to fall bellow an acceptable threshold could improve public health and awareness of pollution problems. In a more traditional sense, these new data can also be integrated to improve strategic planning systems, by providing a richer evidence base upon which to make more optimal decisions given constraints. Finally, data can empower citizens to make more informed decisions about its use, mobility, and governance of urban areas (Figure 1.10).

Figure 1.10 Walkonomics use a variety of data about cities to grade streets by a series of measures that impact the pedestrian experience. These can be integrated into their mobile application that enables routing



Source: Image available at data.gov.uk. Walkonomics app, www.walkonomics.com/

Data in itself does not offer insight, and methods are required to both generate and communicate findings. Urban analytics provides the tools, technologies, and processes for the pursuit of this new data-intensive science of cities. However, even when we ask the right questions with an appropriate method, we can potentially get incorrect answers. With most new data we may trade the breadth and geographic scale of the attributes being measured for incompleteness or uncertainty. The overarching objective of this book is to introduce the main techniques of urban analytics, their underlying science, and provide consideration of their appropriate use.

Questions

1. Where are the largest cities in the world and how can these be measured? Think about the different ways in which we might measure “large” and explore what data are available to derive this information.

2. Explore changes in city population over time for a country of your choice. Describe the patterns with reference to short-term (e.g., environmental disaster, political instability, etc.) and long-term stimulus (e.g., economic decline etc.).
3. Think about a city near you or the one in which you live. Write a 2,500-word essay on the different ways in which data are embedded into the urban environment, making reference to how ethical considerations in the use of the data can be balanced against social, environmental, or economic benefits.
4. Compare and contrast two city datastores, making reference to the breadth, depth, and scale of data that they hold. How might these datastores be improved?

Supplementary Reading

Batty, M. (2013). *The new science of cities*. Cambridge, MA: MIT Press.

This book provides a window onto a lifetime of experience in urban and regional modeling, and makes the case for a new science of cities. It is an advanced text, but very readable, even by those who are new to this area.

LeGates, R. T., & Stout, F. (2015). *The city reader*. Abingdon: Routledge.

One of the best anthologies of academic writing on cities, planning, and urban studies more generally, mixing a range of classic and contemporary writings from key thinkers. The contents are mainly theoretical with many of the writings also providing rich historical accounts.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). *Geographic information science and systems* (4th edition). Hoboken, NJ: Wiley.

This book is a classic reference text for GISc and provides excellent background materials to this chapter and the rest of the book more generally.

2: Sensing the City

Learning Objectives

By the end of this chapter students will understand the following:

- There are many different types of data collected within urban contexts.
- Data can record activity, attributes and dynamics over a range of spatio-temporal scales.
- Data are generated organically, as a byproduct of our daily lives, or through purposeful data collection processes.
- Sensors and social media complement traditional sources and are generating an increasing amount of new data about urban areas.

Thinking about Cities through Data

The composition, dynamics, and complexity of cities can be understood at multiple scales. At one scale a city can be conceptualized as a single entity, for example, we might want to say something about Los Angeles or Sao Paolo. However, we might also be interested in talking about events within cities, a car accident at a particular intersection, the redevelopment of a specific land parcel.

At a high level of abstraction cities have been thought about as objects existing within spatially bounded extents. However, such definitions do not represent either the functional or relational aspects of cities. Cities possess both enormous centripetal and centrifugal force that attract people, resources, capital, and ideas, but also propel these same assets outwards. For example, in the nineteenth century, massive textile mills in cities of the north-eastern United States processed raw cotton (picked by slaves in the south-eastern United States) into finished products which were then shipped to cities in Europe (and elsewhere). Such connectivity binds individual locations into networks (or systems) of cities (Berry 1964).

However, systems of cities are seldom governed directly – there is no mayor or governor who has authority over a network of cities' trade. For example, the Mayor of Detroit, an economically depressed city in the United States, cannot reverse the global macroeconomic trends that led to that city's decline. Even within the Detroit metropolitan area there are many municipalities, each with their own governments; the city of Detroit cannot control actions of even these nearby neighbors. This kind of intra-metropolitan competition is sometimes called regional fragmentation. While it is important to recognize the interconnectedness of cities, in urban analytics we are usually focused on dynamics *within* cities, not *between* them. Most new forms of government-generated data about cities exist within those bounded extents that are common to formal legal definitions of a city.

Urban analytics becomes especially powerful when we zoom-in and examine cities at a finer level of abstraction, as they are atomic in nature; composed of myrid physical and natural objects, each with different degrees of interaction. These intra-city relationships are highlighted through data – for example, the relationship between pedestrian traffic within a shopping precinct and store sales revenue.

For decades much work within urban analytics could only look at cities as a whole or at best through very course aggregates such as census enumeration areas or postal zones, as computers were simply not capable of doing meaningful analysis for very disaggregate data. Consider the challenge of counting everyone who uses the London Underground or every tree in NYC. Ten years ago this data would have been very difficult to collect and store; now it is available in almost real time. One can refer to this data as being of "high spatio-temporal

resolution" because it is generated at high temporal frequency and has substantial geographic detail. It is also increasingly possible to turn such data into actionable information, which is a recurrent theme of this book.

The combination of data with analytic technologies has been described as a macroscope. Whereas a microscope might allow us to see things that are too small to be observed with the naked eye, a macroscope lets us focus on things that are too big to be observed directly. For example, we might be able to observe a small sample of pedestrians with our naked eye, but cannot without the assistance of data observe the ebb and flow of all pedestrians within an entire shopping mall over the course of a day.

Data Types

In the study of cities there is an important distinction between data that can be considered as *organic* versus that which is *purposeful* or *designed*, with the principal difference between them being the intentionality of collection.

Organic data are the byproduct of some process like communicating with friends, buying something online or in a store, riding the bus or train. On the other hand, purposeful data are collected through a carefully designed, statistically robust collection program. The archetypal forms of designed data are the national demographic and economic surveys that report on things like the unemployment rate, or count the entire population (such as a Census). Designed data provides critical insights into cities; for example, purposefully created statistical surveys might be the only way to produce an accurate estimate of household income for some neighborhoods because designed data attempts to represent the entire population of interest, not just those who shopped at a particular store or rode the bus. The collection of purposeful data are usually overseen by statisticians with a specialization in the design of surveys that are representative of a population.

Organic data are usually available at very high frequency. For example, we might be able to count the number of tweets within an area in near real time, or, based on automated ticketing, the number of people on a mass transit system over the course of a day. Organic data arises as a byproduct of some transactional process of daily life within cities. By contrast, designed data are collected at enormous expense. Consider the "simple" task of counting the population in a country. How would you do it? People are dynamic and move around. It might be possible to mail a questionnaire to every address in the country, but what if people do not respond? In practice, national surveys often involve creating a register – a list of all known home addresses. Each address is contacted by sending a questionnaire via a variety of mechanisms including mail, phone, and the Internet, and then dispatching people called enumerators to those households that do not respond to the survey. Designed surveys are enormously expensive: in the United States the 2010 Census cost about $42 per person counted!

Another common type of designed data, at least in the United States, is a political poll. Considering political polls is useful, even though they are not purely urban, because they highlight an important aspect of designed data. Political polls are usually reported with a margin of error. This margin of error reflects that when a poll is conducted not everyone within a country is contacted, and some of the people who are contacted do not respond. The fact that political polls are generated using a sample of the population creates uncertainty in how close each poll's findings are to the "truth". This is exacerbated by non-response. Imagine if supporters of a certain political candidate did not trust the media and thus refused to respond to pollsters. These kinds of systematic problems require consideration when deriving insight from designed data. However, one of the strengths of designed data are that methods exist for calculating the amount of uncertainty injected into estimates due to sampling and non-sampling errors. This uncertainty is captured in a "margin of error." The margin of error reflects a range of values within which the true value is expected to lie.

The differences between organic and purposeful data are important because they have different use cases. Organic data are more useful for measuring what we might call “fast” urban dynamics, whereas purposeful data are better used to target “slow” dynamics. Fast dynamics record how phenomena quickly change in time and space over high temporal resolution. Whereas “slow” dynamics might be taken to mean change over months, years, or even decades. For example, daily traffic patterns might be considered an example of fast dynamics, while gradual changes in average household income, or the percentage of the population who speak Spanish as their primary language, are examples of slow dynamics.

Data recording fast dynamics are messier in that they usually do not come with estimates of uncertainty (i.e., margin of error): we might know that 15 percent of tweets within a city are in Portuguese, but we have no way of knowing how representative that is of a population (is the city 15 percent Portugese speakers?). Conversely, from a designed survey we get a picture of the entire city, which might inform us that 7 percent of the population plus or minus 4 percent speak Portuguese as their primary language.

Real-Time Data

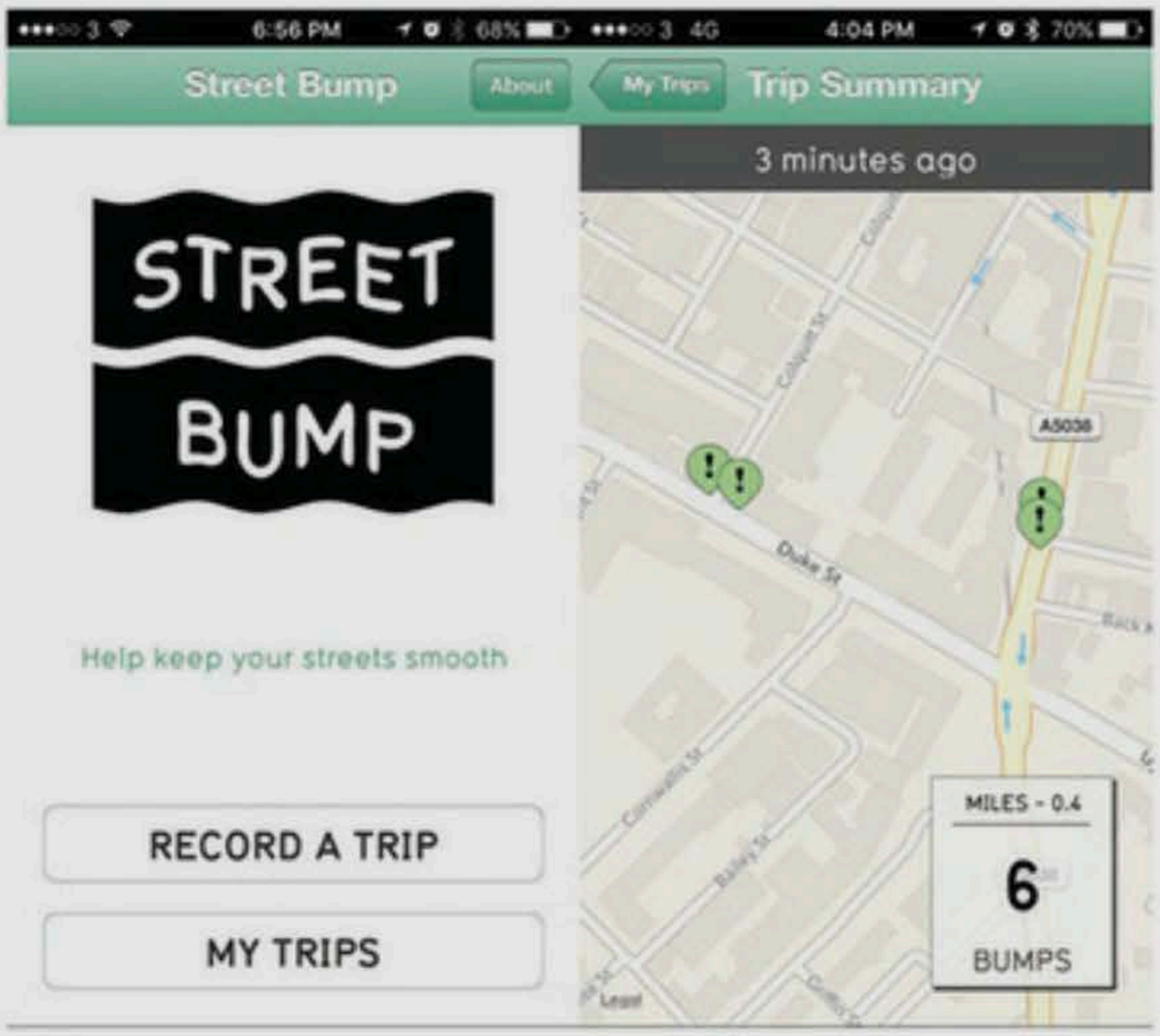
Increasingly, traditional sources of data within cities are complemented by an array of new forms of real-time data generated by both sensors embedded within the fabric of urban areas and through the volunteering of geographic information by people flowing through these spaces. Such data has a tendency to be of higher spatial and temporal granularity; however, much of it is also less representative and comprehensive than traditional sources. These aspects of data create a range of challenges for their use and interpretation (see Box 2.1 and Profile 2.1).

Box 2.1: Real-Time Pothole Detection

The City of Boston Mayor’s Office of New Urban Mechanics released an iPhone mobile app in 2012 that uses the built-in accelerometer of the phone to detect sudden depressions in the road that may be caused by driving over potholes. Each depression event is geolocated using GPS, thus providing a useful example of crowdsourced geographic information (Figure 2.1A and Figure 2.1B). Innovatively, this app sends the recorded depression events to a server, thus separating the sensing of potential potholes by multiple individuals from their later assessment, identification, or rejection. Through analysis of the data generated by the application, the City of Boston discovered that one of the main causes of recorded bumps was as a result of manhole covers being sunk into the road; as a result, over 1,000 of the worst manhole covers were repaired in association with the utility companies (Mechanics 2015).

Figure 2.1 Potholes detected using mobile phone accelerometers

A Potholes in a road



B The Street Bump mobile app showing a number of bumps recorded along two street sections

However, uncertainty and bias may be inherent in the data generated. Depending on variability in dashboard materials where the phone is placed, suspension stiffness, and positioning of the phone, readings may differ between vehicles or users. Furthermore, because the final application was only released for the Apple platform, this restricts use to a group of users able to afford one of these devices and a further subset of these potential users who would be willing to install the app and volunteer the information. Only data collected by these individuals will be included in any analysis, and, as such, is only representative of those areas of the city in which they drive. Despite such caveats, through appropriately designed

and calibrated detection algorithms the rates of false positive results have been shown to be less than 10 percent (Mechanics 2015), and as noted by the developers of previous versions of the app, these were available as Android devices, so for future releases the app could be made cross-platform compatible (Carrera et al. 2013).

Profile 2.1: Rob Kitchin

Figure 2.2 Professor Rob Kitchin, National Institute of Regional and Spatial Analysis at Maynooth University

Source: Rob Kitchin

Briefly describe your research interests

My most consistent areas of research concern mapping (its practices and theories) and networked urbanism (and how digital infrastructure, software, and urban Big Data are reshaping city life). Beyond those I just seek to answer whatever questions I think seem interesting or important. For example, when the financial crisis hit Ireland, I swapped my attention to researching issues of planning and housing.

How do you think Big Data will or has changed society?

I think there's little doubt that Big Data are changing how business is conducted, governance enacted, and science undertaken. If we consider cities, there is now a deluge of real-time, exhaustive, fine-grained data being generated by digital cameras, sensors, transponders, meters, actuators, GPS, and transduction loops that monitor various phenomena and send data to an array of control and management systems such as city operating systems, centralized control rooms, intelligent transport systems, logistics management systems, and smart energy grids. This is changing the operational governance of cities, but also how we model and plan them. And with respect to academia it is changing how questions are asked and answered, leading to the formation of urban informatics and urban science.

Should a commercial Big Data analyst be concerned with ethics?

Yes, I think everyone who is undertaking analysis should be concerned with ethics and ethical practice. On the one hand, there are various laws relating to privacy and data protection and security that have to be complied with. On the other, analysts have a duty of care to their fellow citizens not to expose them to harm through their analysis. What constitutes harm is sometimes difficult to define, and harms can occur directly or indirectly, nonetheless I think every analyst should consider how their work might or will be used and to practice their work responsibly. They should also consider the reputation damage they or their company might experience if they cross what some have termed “the creepy line.” There is little doubt that many Big Data ventures are crossing this line, used to predicatively profile, socially sort, behaviorally nudge, and regulate, control, and govern individuals and the various systems and infrastructures with which they interact.

What is the most important lesson that you would instill on newcomers to urban data science?

Try and start with a question and then find the data to try and answer that, rather than starting with the data and trying to find something useful to do with them. Alternatively, use guided theory to undertake exploratory data analysis that might reveal some potentially interesting hypotheses. Do not fall into the trap of believing that the age of Big Data means the “end of theory” or that the data can inherently speak for themselves!

Passive Data Generation

An increasingly pervasive source of dynamic information about urban areas are derived from data generated by passive technologies supplying a variety of real-time measures. These are wide ranging in scope, but typically involve different sensor technologies including: carbon dioxide (or other gases), temperature, humidity, noise, and light. Such environmental sensors can be deployed at a variety of scales, ranging from city-wide implementations down to the level of individual buildings. Such persistent and Internet-connected monitoring can enable a variety of innovative applications when sensors communicate. For example, air quality within a building may be monitored, and when this reaches some undesirable level, windows could be opened either automatically or occupants given advice that such an action might be necessary. At the city scale, emissions monitoring might enable geographically targeted warnings to be sent to those with respiratory problems through a connected mobile app if pollution levels were likely to adversely impact their breathing.

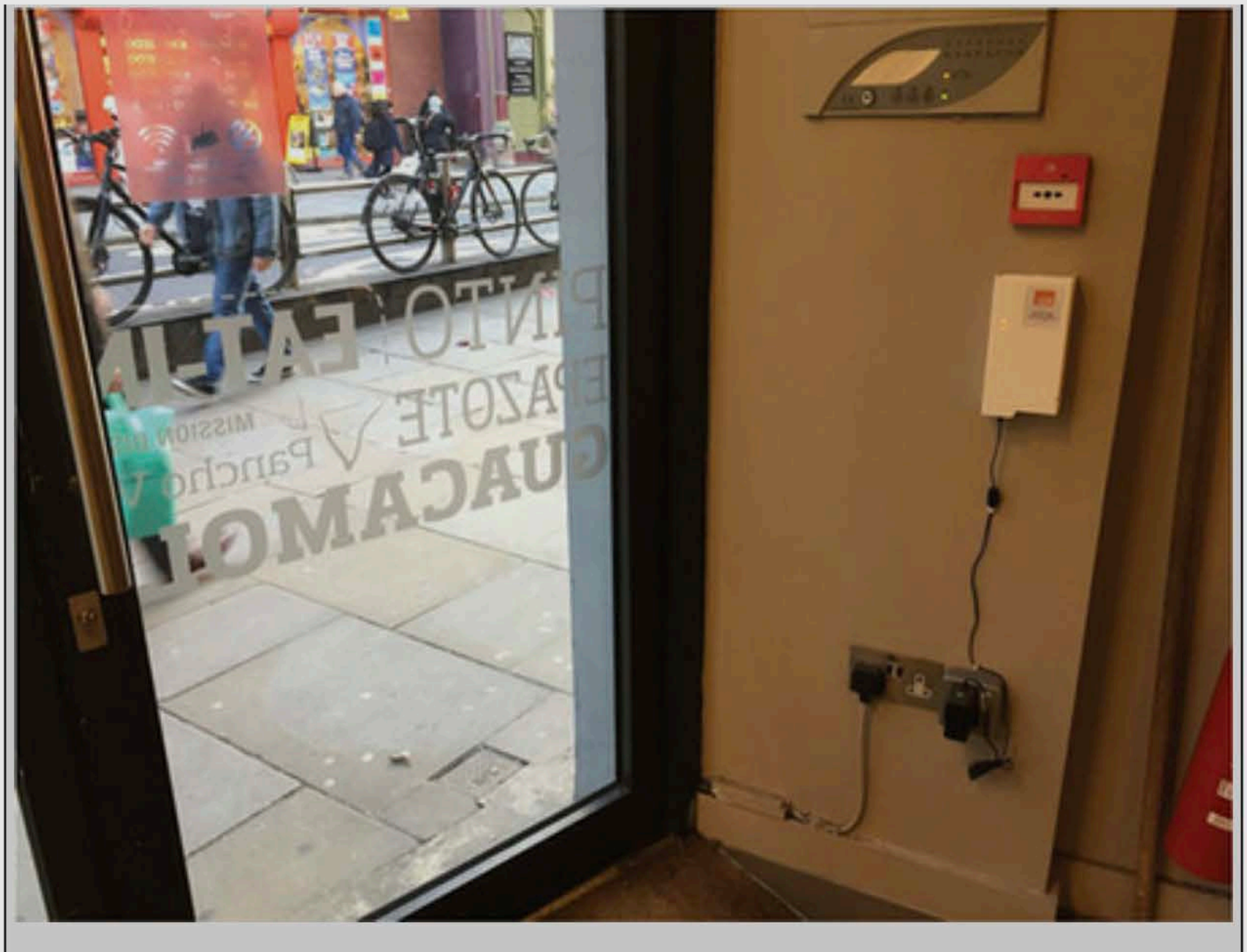
Another class of passive data generation in cities relates to a range of emerging technologies linked with cellular phones that are enabling pedestrian or vehicular volumes to be estimated. Sensors can be installed that detect the presence of a phone through a variety of wireless technologies including Bluetooth, Wi-Fi and the GSM network (see Box 2.2).

Box 2.2: SmartStreetSensor

The SmartStreetSensor Project was developed as a partnership between the Local Data Company (www.localdatacompany.com) and the ESRC Consumer Data Research Centre (CDRC; www.cdrc.ac.uk) to provide a platform for the analysis and interpretation of a live feed of estimated footfall across high streets within Great Britain. The network comprises 1,000 sensors situated over 81 towns and cities across Great Britain that were identified to offer a wide geographical spread, but were also representative in terms of different shopper demographics and retail area characteristics (Figure 2.3).

The sensors detect passing Wi-Fi-enabled devices which are used as a proxy for people; however, these devices do not track individuals through space and time given privacy constraints and legislative guidance. Field research has been very important to the process of developing count estimates, with traditional manual counting surveys used in the calibration of the recorded data streams. This process is necessary to account for variability in the number of people using Wi-Fi-enabled devices across Great Britain. The calibration process also ensures that the sensors only feed data back from the immediate area in front of a shop.

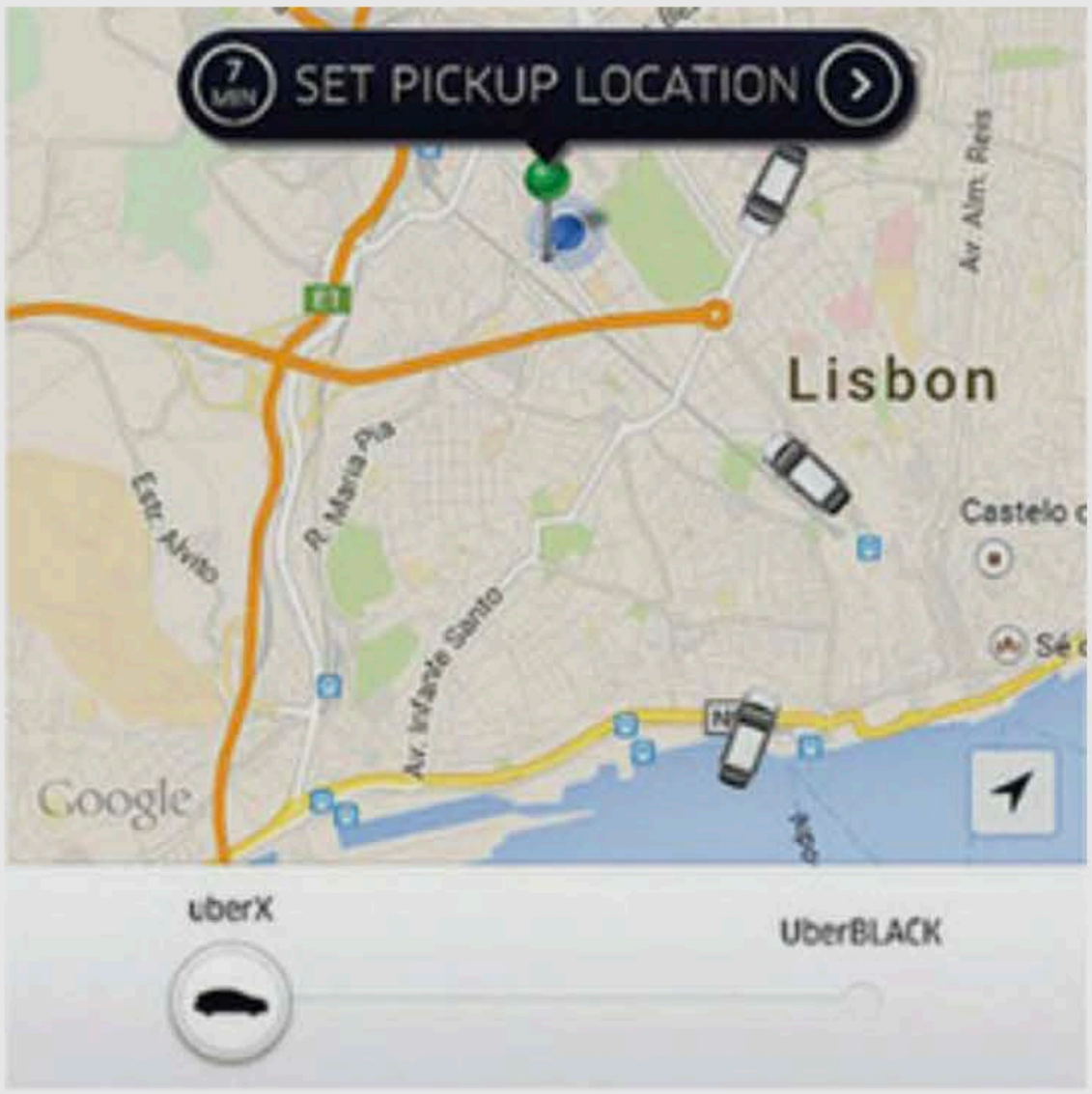
Figure 2.3 An example of a SmartStreetSensor device within a shop window in London



Source: Local Data Company

A further source of passive data are generated through selected apps which, when running on cell phones, can provide streamed location data either in the background or as integral to the function of the app, and will typically use built-in GPS and Wi-Fi functionality of the cell phone to derive location (Figure 2.4). Such data are often collected by the app owner and enable a variety of additional products or services to be created. For example, Google uses pooled location data to measure the speed of drivers along road lengths, which then feeds into estimated congestion measures that appear as a traffic layer within Google's maps product. More recently, Google has also produced temporal profiles for many commonly visited locations such as retailers, thus providing guidance on when destinations may be quieter.

Figure 2.4 The Uber application connects riders with drivers using paired mobile device locations and estimates a wait time for a trip



Source: Photography by Gustavo da Cunha Pimenta CC BY-SA 2.0 (Flickr)

In addition to cell phones, GPS receivers are also attached to other objects that move through urban areas such as taxis or buses (Figure 2.5). In New York City (NYC), this includes the iconic Yellow Taxi Cabs which, through a Freedom of Information Law request in 2014 by Chris Whong (chriswhong.com), had all trip and fare information released by NYC's Taxi and Limousine Commission. Such data allowed the GPS tracks of the cab routes to be mapped, and has enabled a number of interesting studies exploring how trip sharing could mitigate the social and environmental costs that are otherwise embedded within these transportation systems (Santi et al. 2014) (see Figure 2.5B).

Figure 2.5 Data generated by NYC cabs

A NYC cabs are equipped with GPS receivers enabling their locations to be tracked, with route information provided to riders on electronic displays

B A screenshot of HubCab (www.hubcab.org), showing taxi flows and potential taxi-sharing benefits between two locations in Manhattan

A long-established technology for real-time surveillance is closed circuit television (CCTV), which can form a dense network of video coverage for urban areas (see Figure 2.6). These devices are used for a variety of purposes ranging from security monitoring related to policing activities, to deter or detect store thefts, and for a variety of traffic management applications. For many instances of CCTV, operators manage and monitor the video feeds manually; however, by coupling the live feeds to computer vision and streaming analytic techniques for pattern recognition, this has enabled a variety of new applications within urban areas. For example, automatic number plate recognition can detect uninsured vehicles or enable charging for certain road bridges and tunnels. However, emerging technology can also use CCTV to track objects through urban areas including cars and, more recently, at the level of individual pedestrians (Chu et al. 2012).

Figure 2.6 A crowdsourced map of CCTV cameras in Prague, Czech Republic

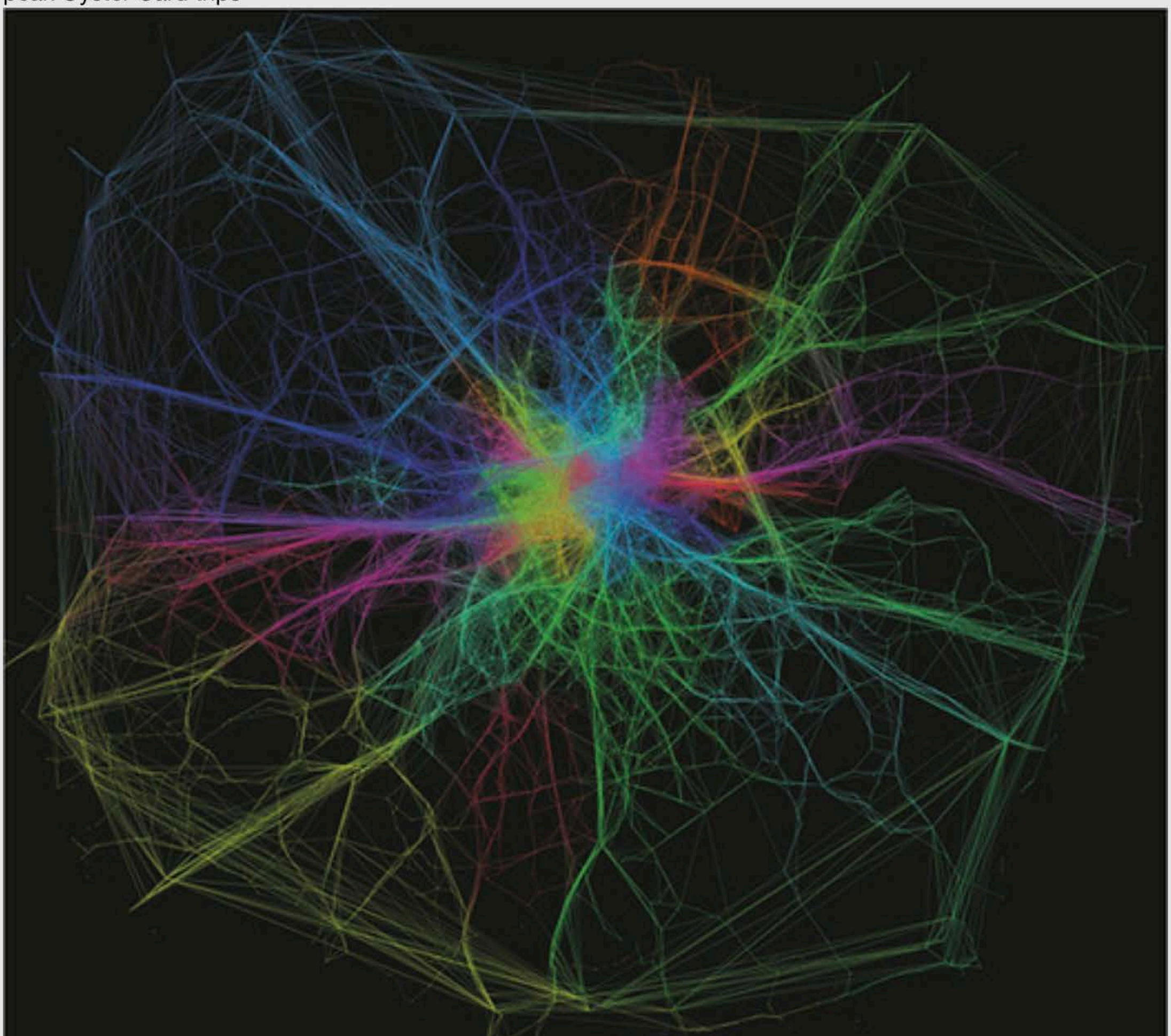
Source: Website http://mapakamer.cz/mapakamer/. Licensed under the Open Data Commons License and CC-BY-SA License

Active Monitoring Technologies

Our movements within urban areas can also be detected through a number of monitoring technologies embedded within these environments that require active use. These mostly relate to transport through the use of smart card ticketing systems or shared bike scheme docking stations. A common monitoring technology that is integral within many public transport systems involves Near Field Communication (NFC) and enables two devices within close proximity to communicate through radio waves. An overlapping technology to NFC includes Bluetooth Low Energy (BLE) which has an extended range and has been used as the base technology for a number of new protocols including the proprietary Apple iBeacon and the open Google Eddystone (see Box 2.3). An example of NFC is implemented within an urban setting as part of the London Underground network, with users swiping/hovering to enter and exit stations, and the trips being charged accordingly. Each year there are around 1.3 billion trips within this network, generating huge volumes of spatio-temporal data about how populations move within the capital. Figure 2.7 shows a subset of the data covering every Oyster Card trans-

action over a three-month period during the summer of 2012, illustrating connections between every London station and its most popular destination.

Figure 2.7 Flows between the most popular destinations by origin, derived from a large dataset of morning peak Oyster Card trips



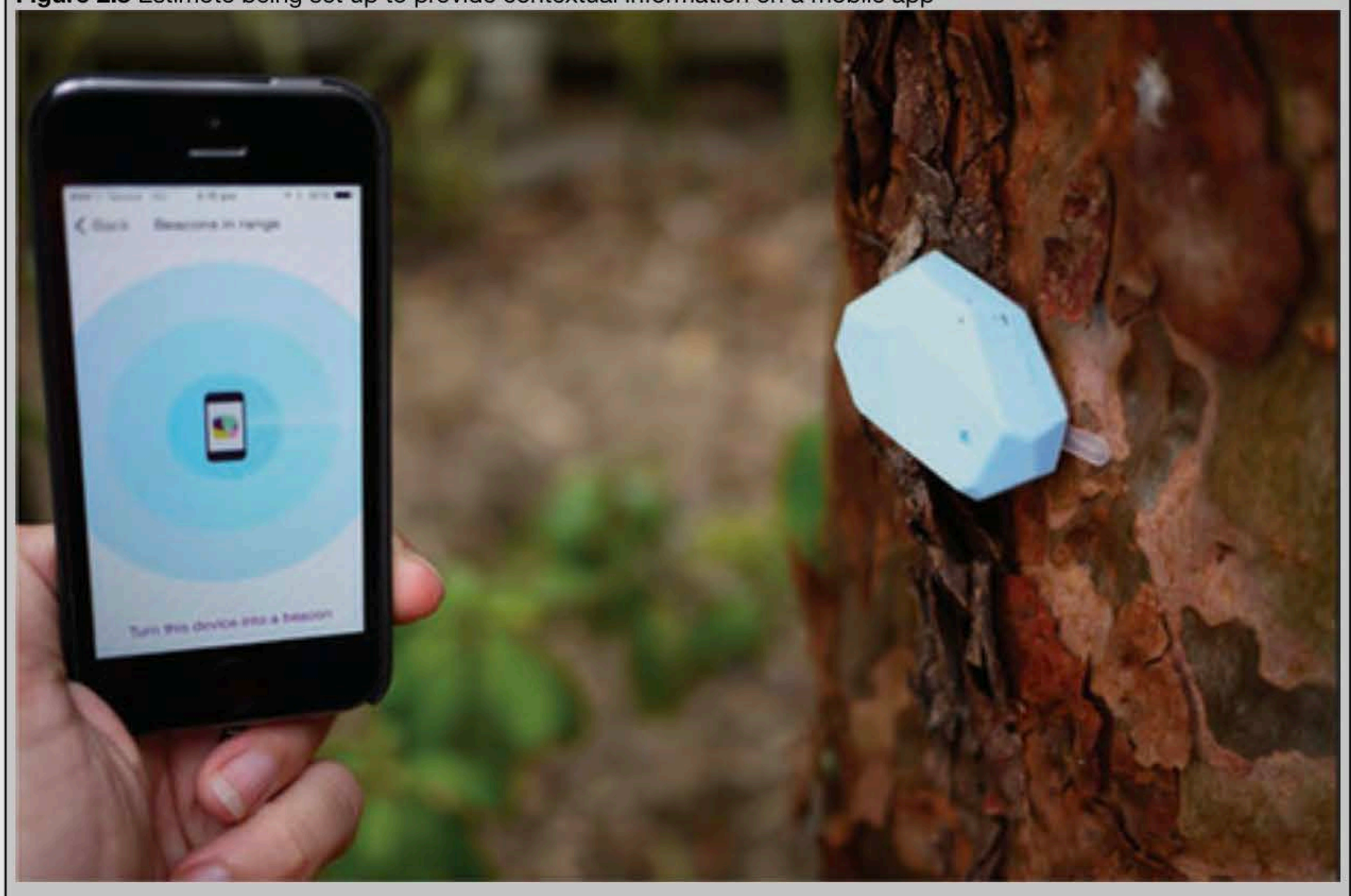
Source: Ed Manley, Centre for Advanced Spatial Analysis, UCL. Reproduced with permission

Box 2.3: Proximity Marketing

A variety of BLE devices are enabling retailers to generate locational information about customer activities within stores where GPS technology (which requires a sky line of sight) would not function. An example of such sensors include the Estimote beacons and stickers (Figure 2.8). By placing these devices onto objects or at specific locations within a store, customers proximal to the sensors can be provided with additional information through a mobile app. As these technologies become more prevalent, they have the potential to transform the way in which visual merchandising is managed

within retail environments, as much richer information can be derived about shopper behavior, including dwell times on particular products.

Figure 2.8 Estimote being set up to provide contextual information on a mobile app



Source: Photograph by Jona Nalder CC BY 2.0 (Flickr)

Thus far in this chapter a range of sensor technologies have been discussed that generate data about how people move within cities both indoors and outdoors, by foot, public transport, or car. However, many large cities also have shared bike schemes where users can check out bikes and are charged for the duration of their use, which is completed by reattaching the bike to a docking station. These schemes generate data as bikes are tracked when checked in and out of stations on the network. Such data has been used to explore how cyclists are using cities at different times of the day or week. Figure 2.9 maps data from the NYC bike-sharing scheme to illustrate differences in use during the day, and between weekdays and weekends.

Figure 2.9 NYC bike-sharing scheme data illustrates differences in use during the day, and between weekdays and weekends

A Weekday interpeak (10 a.m. to 4 p.m.)

B Weekday rush-hour peaks (7–10 a.m. and 4–7 p.m. starts)

C Weekday nights (7 p.m. to 7 a.m.)

D Weekends

Source: Oliver O'Brien, University College London; oobrien.com

Social Media

Social media platforms are numerous (e.g., Twitter, Foursquare, Facebook, etc.), but share one commonality in that the content is mostly user generated. There are a wide array of these platforms that primarily but not exclusively relate to one or many of the following functions: photography sharing, communications, social networks (personal or business), reviews, blogging, music and video. The variety of data generated and stored within these platforms has provided an extensive resource upon which a large body of urban research has been conducted during the past 10 years. Much of the data contains records and relationships that are also georeferenced, often with high temporal granularity. Innovative work within this area concerns analysis where the data are not simply reported in terms of prevalence, but instead utilized to infer some spatial process or structure that otherwise may not be measurable through traditional sources. For example, Figure 2.10 presents a map for the extent of Greater London, UK where Google Translate (translate.google.co.uk) was used to extract the spoken language of georeferenced tweet content.

Figure 2.10 The linguistic Twitter geography of Greater London

Source: James Cheshire and Ed Manley, University College London. Reproduced with permission

Data from National Statistical Agencies

Traditionally a census is a complete count of the population within a country or city, however, conducting such surveys are expensive, but have long been seen as an essential tool for governance. Alternatively, within some countries government agencies keep population registers, which track each citizen through a unique reference, storing information about their age, education, gender, as well as where they live, work, and the structure of their household (married, single, etc.). Such registers are common in a number of North European countries and are amazing sources of data about a population, but they are often closely controlled. In those countries that do not keep such population registers, to develop an understanding of the population one must periodically ask questions of the residents in an area through Census or other large sample survey.

A census is an effort to collect details on every person in a country, typically using a questionnaire delivered through a range of channels including mail, telephone, face to face, and online. Within many countries, the completion of a census is obligatory. For example, within the United States, the Constitution requires that the federal government performs a census every 10 years (a decennial census). Article 1 of the US Constitution requires a complete counting of the population for the allocation of seats to the House of Representatives. However, even the first census in 1790 went beyond this minimum mandate. That census asked each household six questions, collecting basic information on age, gender, race, and slave status of the people living there. Setting aside the contentious issue of counting slaves, questions regarding age and race are not strictly necessary under the constitutional mandate. In the first and each subsequent US census, there has been a tension between collecting data that informs governance and minimizing the amount of data collected (and thus the burden on those responding to the survey).

Counting the entire population for a country is not easy, and thousands (or even tens of thousands) of people are hired by the US Census Bureau to conduct the decennial census. The cost of the endeavor is also related to the number of questions asked. Longer questionnaires and certain types of questions are also less likely to be answered, so it is in a government's interest to optimize questionnaire length and composition. In planning a census, a great deal of care is taken about these attributes, with many countries doing extensive testing or trials to monitor impacts upon response rates and data quality.

One way to balance the cost and burden of long questionnaires on respondents is to talk to a subset of the population. By asking questions of a *sample* of the population, national statistical agencies can collect more detailed attributes about the population without bothering everyone. These surveys tend to be longer and more in depth than census enumeration of a total population.

Urban Data

Urban data can exist at a range of scales and record a brevity of different social, natural, and built environment characteristics. These are generated through multiple different processes, some of which are planned or purposeful (e.g., surveys, censuses, etc.), while others are organic and emerge as a result of systems that operate within cities (e.g., transit etc.). All data possesses a degree of error, uncertainty, or bias and can be considered as socially constructed; as such, it is important that these attributes are given careful consideration as part of the analytic process.

Questions

1. What is the difference between organic and purposeful data generation? Illustrate this discussion with examples, and consider the advantages and disadvantages of these two approaches.
2. How might data be considered as socially constructed, and to what extent might issues manifest because of this in the process of urban analytics?
3. Compare and contrast the use of social media versus survey data to investigate an urban issue of your choice.

Supplementary Reading

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

An excellent critical treatment of the contemporary spatial data economy, drawing together a very rich body of information. Of particular note should be Chapter 10, which provides a very important introduction to ethical issues and considerations.

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

This book has a very positive endorsement to the utility of Big Data, but should be read in conjunction with Kitchin (2014) for a more balanced view!

Townsend, A. M. (2013). *Smart cities: Big data, civic hackers, and the quest for a new utopia*. New York: WW Norton.

This book introduces smart cities as a mechanism to rethink governance through data and infrastructure that enable a more direct connection to populations. There is an interesting discussion around grassroots approaches to smart cities over those dominated by large corporations.