# Chapter 31
# Cultivating Urban Big Data


**Ningchuan Xiao and Harvey J. Miller**

**Abstract** Urban big data often contain spatial and temporal elements that have increasingly become an integral part of various applications and projects such as smart mobility, smart city, and other digitally enhanced urban infrastructure. It is critical to develop an open and collaborative environment so that these data can be used by a wide range of users. This chapter first discusses some characteristics and sources of urban big data. Three hypothetical user stories are described to highlight the potential of these data. After describing the internal data structure of these data and techniques that can be used to retrieve the data, we discuss the difficulty in making the data useful for the general public and elaborate on a self-organizing agile approach to developing an urban big data infrastructure.

## 31.1 Introduction

Big data are one of the most popular topics of the past decade (Marr 2015). The concept of big data has evolved beyond the original context as a buzz word into the reality of daily life and has shown tangible values for businesses, governments, research communities, and the general public (Kim et al. 2014; Günther et al. 2017). Informally, big data refer to the vast amount of data that are generated, collected, or distributed at a high frequency or speed. More formal definitions of big data vary widely in the literature (Mergel et al. 2016), and researchers have generally agreed that big data all share certain characteristics, including volume, variety, veracity, velocity, and value (Chen and Zhang 2014).

Urban areas are a significant playground where multiple players are engaged in the generation, storage, and applications of big data (Kitchin 2014). For much of the urban population, big data have become an integral part of their daily lives. Many technological, economic, and demographic factors have contributed to this rapid

N. Xiao (✉) · H. J. Miller
Center for Urban and Regional Analysis, The Ohio State University, Columbus, USA
e-mail: xiao.37@osu.edu

Department of Geography, The Ohio State University, Columbus, USA

© The Author(s) 2021

W. Shi et al. (eds.), *Urban Informatics*, The Urban Book Series,
https://doi.org/10.1007/978-981-15-8983-6_31

growth. Various sensor technologies used in domains such as environmental monitoring and shared transportation means are the data sources that provide continuous feeds (Cuff et al. 2008). These sensors have been connected through a network that forms what is dubbed the Internet of things or IoT (Atzori et al. 2010). In an urban area, the IoT plays an especially important role in everyday life because the so-called things in the IoT include both physical objects such as GPS devices and environmental sensors, and also people who are equipped with sensors that can provide information about the location and surrounding area of the person. In many cities around the world, public transportation systems have increasingly applied GPS to allow more accurate and accessible transit to their residents. For example, many public transit agencies instrument their vehicles with GPS receivers and share these data publicly to support real-time bus tracking and arrival applications. In the meantime, passengers of these transportation systems use new ticketing methods such as smart cards to pay the transit fare, which also allows the transportation authorities to record and track their movements. In addition, citizens in urban areas have become a special kind of sensor (Goodchild 2007). These "sensors" have multiple ways of generating data. For example, they may provide spatial and temporal data using technology developed by commercial companies, as in the case of Google Traffic, in exchange for services (Heipke 2010), or they collect data about gas prices or traffic and exchange them with companies such as GasBuddy or Waze for rewards or other types of membership benefits (Boulos et al. 2011). Telecommunication companies have established vast databases that contain user identities and spatiotemporal activities. Cell phones have been mostly replaced by smartphones where the original function of making phone calls has been reduced to merely one of a huge number of uses relying on the network provided by the telecommunication companies, where many of the other functions are enabled to track the user's location.

Urban big data generated through sensor technology have all the characteristics of big data in general, but more critically they have their own features. First, urban big data involve a wide range of users from the general public to those in private services. It is important to recognize that these groups of people are active in multiple roles in the entire ecosystem of urban big data, including the phases of data generation, maintenance, storage, and usage. The users of the data, for example, also contribute to the generation of the very data they are using, as in the case of GasBuddy[1] where members report gas prices at different stations and also use the information provided by the Web service. Second, urban big data always have a geographic footprint as the data must relate to an urban extent. This is different from other big data sources (e.g. Web search and tweets without geotags) where the geographic dimension is not salient. Along with the spatial dimension, urban big data also have an important and sensitive temporal dimension as many applications depend on the time stamp of the data (e.g., real-time bus information is important for users to schedule activities around bus operations). Third, urban big data as a whole are often ill-structured because many data sources often do not coordinate their data generation and collection efforts. Data tend to exist in a loosely managed environment where a particular

[1] www.gasbuddy.com.

data set may not be connected to other data sets and may not be known to other groups of people.

The purpose of this chapter is two twofold: We provide an overview of urban big data and discuss the technical aspects how data can be made useful for various purposes. We specifically focus on the part of big data within the urban context as described above. The remainder of this chapter starts with a discussion of data sources. We then discuss the elements of the data, followed by several hypothetical user stories. On the technical aspects of urban big data, we discuss several data-collecting techniques and then extend the discussion into the needs and requirements for developing an urban big data infrastructure.

## 31.2 Sources of Urban Big Data

Urban big data come from a wide range of sources, and it may not be straightforward to categorize these sources. For example, in a study of the characteristics of 26 data sets (Kitchin and McArdle 2016), seven types were used to categorize the data sets, including mobile communication, Web sites, social media/crowdsourcing, cameras/lasers, transactions of process-generated data, and administrative. Not all these data have the urban context. Here, we group big data sources by the type of data providers, which can be from private or public sectors. In addition, we also recognize the types of data that are generated voluntarily. Each data set can be open to the public to use or may be protected so that only authorized users can access it. The distinction between open and protected data is important, especially for the urban context, as many data sources may have limited uses because they are difficult to share among potential users of the data. Table 31.1 lists a number of example

**Table 31.1** Example sources of urban big data

| Provider | Open | Protected |
|---|---|---|
| Private | Bike sharing | Bike sharing<br>Mobile phone calls<br>Surveillance camera and CCTV<br>Health data |
| Public | Real-time bus operation<br>Census data<br>LiDAR and remote sensing<br>Traffic cameras and CCTV<br>Air pollution sensors | Public transit usage<br>Individual survey<br>Public health data |
| Volunteers | Social media<br>Community sensor network | Social media<br>Health data on mobile devices |

data sets for each category. The purpose of listing these examples is to give a brief overview of possible and practical data sources. We note that these are merely a small sample as different cities in different counties will certainly have more sources.

The private sector generates a huge amount of data on a daily basis. We only list a few examples that are more related to the urban context. Popular bike-sharing companies, for example, provide both open and protected data. The open slice of the data may include the number and locations of bike stations, and available bikes and docks at each location, while the protected part results from tracking the movement of each individual bike along with information about customers. Some companies (e.g. Waze) may choose to release an aggregated version of their individual data in the form of averages over space and time as the open part, while protecting the actual individual data. It is obvious that private companies have been collecting such data sets as phone calls, surveillance, and individual health information. These data are highly protected due to privacy laws and even the need to maintain good relationships with the public (Chap. 32).

Urban big data from sources in the public sector cover a variety of domains such as demography, transportation, environment, and public health. These data are not necessarily open to the general public due to privacy concerns. For example, while many municipal services provide public transit data (e.g. bus operations), individual usage of bus data that can be obtained through the records of bus passes is often protected. The duality also applies to census data, where the aggregated version of the demographic, housing, and economic data is open to the general public, but individual surveys are tightly guarded.

The third type of data source includes individuals or groups who volunteer their own data for various uses. These providers generate their own data as they are themselves sensors (Goodchild 2007; Chaps. 28 and 29), which is different from the other two provider types where data are passively collected. A significant source in this category is the social media data. Tweets, for example, can be harvested using different licensing policies granted by Twitter. While the users generate the data, they do not necessarily own their own data, and not all social media data are open to the public. Other important kinds of volunteered data are those generated by the general public using various sensors. One of the prominent examples is the use of affordable air quality sensors (Kumar et al. 2015), and the users of these sensors can share their data to form community sensor networks (Yi et al. 2015). Though the quality of such data may be questionable (Lewis and Edwards 2016), they have been used for mapping[2] or other analysis.[3]

[2] www.purpleair.com/map?#1/25/-30.

[3] www.citylab.com/environment/2018/07/cheap-sensors-are-democratizing-air-quality-data/563990/.

## 31.3 User Stories

Let us consider three user stories of urban big data. These stories are hypothetical, but they do represent some of the examples we have encountered in our previous applications. They are not limited just to the data but extend to the entire ecosystem of urban big data that includes, in addition to data, the software systems as deployed in a hardware or network setting. We assume the existence of the data, and we aim to demonstrate how such data can be used in meaningful ways to address real-life problems. These stories are based on examples from experiences in the USA, but we believe it is possible to find relevant examples in other countries. We note that we use the term user story instead of use case for a specific reason, as use cases are a software engineering term that requires more formal description of the system. However, in this chapter, as will be discussed later, the specific requirements of the data usages will be difficult to define, and we argue that an agile method is more suitable. More discussion about the agile method will be presented later in this chapter.

The first user story involves a resident, Jon, in an urban area. Jon plans to invite a few of his friends to a party over the weekend. He has a few requirements for the party venue. His friends like biking, and he wants to use the bike-sharing system so that his friends can rent bikes for some fun riding. The party location needs to have sufficient available bikes and be close enough to the trails. Not all of his friends have cars, so Jon must consider a place that can be accessed by public transit or only by biking. He also desires the place to be close to some respectable restaurants for a happy hour after the ride. There is no existing app that will help Jon plan the event. But Jon is data savvy and can use the openly available data and mapping tools to put together some candidate locations. He can also use historical data to tell roughly what will happen in the weekend. He then shares what he has found with his friends before he finalizes the party venue.

The second user story involves a group of individuals who are interested in the city's development direction. They are busy with their own daily work, and it is hard for them to find a good time to have face-to-face meetings. Most of their activities rely on the use of online communication tools. Recently, the county planning authority posted a statement that gives the overall environment of the county a low rating. But the group does not feel this rating fairly represents the progress the county has made over the past few years and would like to give the overall environment another look. Two group members, Rachie and Lieta, are especially critical of the county's rating. Rachie is interested in air quality, and he is able to collect official air quality data and unofficial, open-source data for the past year. These are daily average data. Leita works on water quality, and she acquires some environmental measures for the gauges in the major streams and lakes within the county. These are again daily averages. They make the data sets available on the group Web site where the members can see the maps and the dynamics of each of the environmental factors. In the discussion board, the group members eventually conclude that it is incorrect and unfair to use a single rating to represent the overall environment quality, and they will present their findings in a hearing.

A third user story involves, again, a group of citizens who are dissatisfied by the congressional redistricting plan put forward by the state commission. They believe the plan is biased toward a political party, even though the commission has clearly stated their anti-gerrymandering stance. The group collected population data at the census block level and voters' data to support their arguments that while the official plan has the overall population evenly divided into the congressional districts, the voters of one of the political parties are strongly concentrated in one district and diluted in others, which gives the other party the edge in the majority of districts. The group also wants to further their argument by establishing that there are multiple alternative plans that can be considered to be equally good. While there are software packages that can be used to generate different kinds of alternative aggregations, they also need to use different demographic and other social and economic data at various spatial resolutions. More importantly, the group uses the alternatives generated by the software and then each group member will start to modify those plans manually to create their own plans. The group members will then share their plans on an online platform that allows them to compare and even synthesize new plans.

Clearly, these user stories involve more than just data. For example, software tools and Web-based applications are essential, and developing those tools is a great challenge. However, it is also clear that data are the cornerstone of the entire ecosystem.

## 31.4 Elements of Urban Big Data

Urban big data exhibit different forms due to the standard chosen to suit the preferred application. For example, a public transit agency may tend to release data using the popular standard called the General Transit Feed Specification (GTFS, discussed later in this chapter). However, we can decompose the data into its smallest items where each can be formulated as a space–time–attribute (STA) tuple of three elements $d = (x, t, a)$, where $x$ is the location or a representation of location of the data item, $t$ is the time stamp to indicate when the observation of the data item occurs or is released, and $a$ is a set of attributes that are associated with the data item.

The above encoding strategy is similar to that of a geo-atom (Goodchild et al. 2007). Here, we separate location and time and relax the way location and attributes can be represented. Location can be explicitly recorded using either a set of coordinates or a set of indicators such as identification numbers that can be used to uniquely refer to locations (see examples below). The attributes associated with the location and time together are a set that is considered as one item in the tuple. This can be done by formatting an attribute as an object formed by a pair of the name of the attribute and the actual value. For example, an attribute of a specific PM2.5 measure can be formed as {PM2.5: 65}. Multiple attributes can be put together in the same manner as {PM2.5: 65, Ozone: 35}, a format commonly used in many data encoding strategies such as JavaScript Object Notation (JSON) that is supported in many programming languages. Putting everything together, an example of ((−83, 40), Mon Jul 01 2019

23:52:00 GMT + 0800 (CST), {PM2.5: 65, Ozone: 35}) encodes two air quality measures at a location in Columbus, OH on Monday, July 1, 2019 at 11:52 PM. Another example is (101.1, 2010, {total: 1200}), indicating a total (population) of 1200 for census tract 101.1 in the year of 2010.

An STA tuple can be viewed as a special kind of observation that occurs at a certain time and location. The big data for an urban area is a set $d$ for all available locations and time periods in the area for the kinds of attributes that can observed or collected. This data model can be used to represent different spatial and temporal phenomena. For example, air quality of an urban area can be represented by a sequence of measures at a number of air quality stations, where each station is marked by its coordinates. Air quality as a geographic phenomenon is a field where observations are possible at any point in space. However, as far as data are concerned, we often resort to discrete data points to represent the phenomenon. For areal data, locations can be represented by the identification numbers or other indicators. For example, different demographic data can be collected for census tracts for multiple years, where each tract is represented by an identification number. The actual geometry (shape and its corresponding coordinates) may not be crucial for the data collection purpose as each tract can be uniquely identified and referred to geographically through another data set containing the coordinates. Similar examples can be found for phenomena on linear features such as water quality measures along a stream, where discrete locations are used for observations.

An interesting case is social media data, which occur in huge volume and at high speed. Such data can still be captured using the STA tuple of three elements, where each social media event (such as a tweet, a Facebook post, and a weichat post) always has the time, location (though it may not be shared), and attribute (the content as in text or a mixture of multiple formats). Another example in the same manner is the vast volume of Web pages. While the location of a Web page may not seem to be essential, each Web page can be assigned a location since each will ultimately be either hosted by a Web site that has a physical and meaningful geographic location or created by a person at some location.

## 31.5 Data-Collecting and Processing Techniques

Urban big data can be obtained using various methods. Many data providers typically offer an application program interface (API) that allows users to collect the data through Internet connections. The APIs may have different constraints in terms of how data can be collected. In general, data providers have full control of how their data can be collected. For example, Twitter uses layers of data-streaming policies, where the free and public license only provides a tiny portion of the tweets, and the way those small numbers of tweets are sampled is not clear to users (Morstatter et al. 2013). Some other data providers, on the other hand, make their data more open. For example, many public transit systems use a particular data protocol to make their schedule and real-time vehicle positions available. In this section, we show how

to stream urban big data using two examples. We focus on open data here, though similar techniques can be applied to more restricted data sources.

The first example is the public transit system. A commonly used format for public transit data (schedules and updates) is the General Transit Feed Specification or GTFS (Harrelson 2006). Since its invention in 2005, GTFS has become the standard for publishing public transit data by agencies such as TriMet in Portland, OR, and BART in San Francisco, CA, to bring data to the general public (McHugh 2013). GTFS data have also been incorporated into Google Maps, where users can find real-time transit information on a common platform. The actual data structure of GTFS consists of multiple text files in comma-separated values (CSV) format. Google also provides a Python package called `google.transit`,[4] where the `gtfs_realtime_pb2` module can be used to help extract information from GTFS without having to directly handle the text files.

The transit agency in Columbus, OH, Central Ohio Transit Authority (COTA), uses GTFS to publish the bus schedule and real-time information for bus trips and its vehicle positions. To retrieve data for vehicle positions, we first use the following four lines of code to import the necessary Python modules and request to open an online GTFS database. In the fourth line, the file called `VehiclePositions.pb` is not the database itself, but a Google Protocol Buffer that describes the structure of the data and the necessary encoding/decoding methods of the data.

```
>> from google.transit import gtfs_realtime_pb2
>>> import requests
>>> import datetime
>>> response = requests.get('http://realtime.cota.com/\
        TMGTFSRealTimeWeb Service/Vehicle/VehiclePositions.pb')
```

Now, we can establish the feed from the actual database and read the actual data using the following code:

```
>>> feed = gtfs_realtime_pb2.FeedMessage()
>>> feed.ParseFromString(response.read())
>>> print(len(feed.entity)) 182
```

There were 182 buses at the time of running the code, among which the first bus can be examined using the following code:

```
>>> bus = feed.entity[0]
>>> bus
id: "1001"
vehicle {
  trip {
    trip_id: "665028"
    start_date: "20190722"
    route_id: "001"
  }
  position {
    latitude: 39.944339752197266
```

[4] https://developers.google.com/transit/gtfs-realtime/examples/python-sample.

```
    longitude: -82.86833953857422
    bearing: 270.0
    speed: 7.93974322732538e-06
  }
  timestamp: 1563818766
  vehicle {
    id: "11001"
    label: "1001"
  }
}
>>>
d = datetime.datetime.fromtimestamp(bus.vehicle.timestamp)
>>>d.strftime("%h %d, %Y, %H:%M:%S")
'Jul 22, 2019, 14:06:06'
```

Along with the position of the vehicle, the data also include the trip ID on which the vehicle is currently running and the vehicle ID, and it will be straightforward to use an STA tuple to encode this information. The default timestamp uses the epoch time, and the last two lines of code show how to convert it into calendar date and time.

We can run the same code after a few seconds, and below is the result. The following example was obtained exactly 20 s after the previous result and the position has also changed, while the bus was running on the same trip.

```
id: "1001"
vehicle {
  trip {
    trip_id: "665028"
    start_date: "20190722"
    route_id: "001"
  }
  position {
    latitude: 39.94470977783203
    longitude: -82.87486267089844
    bearing: 270.0
    speed: 8.457552212348673e-06
  }
  timestamp: 1563818786
  vehicle {
    id: "11001"
    label: "1001"
  }
}
```

While the vehicle position feed provides real-time data about bus location, detailed information about bus stops must be obtained from another real-time feed. The following example uses a similar procedure to retrieve real-time stop information.

```
>>> response = requests.get('http://realtime.cota.com/\
…   TMGTFSRealTimeWebService/\
…   TripUpdate/TripUpdates.pb')
>>> feed = gtfs_realtime_pb2.FeedMessage()
>>> feed.ParseFromString(response.content)
```

Below we explore some information about the first trip. The following example reveals the information about the trip and the vehicle that was currently operating on this trip. This corresponds to the bus information from our previous example.

```
>>> feed.entity[0].trip_update.trip
trip_id: "665028"
start_date: "20190722"
route_id: "001"
>>> feed.entity[0].trip_update.vehicle
id: "11001"
label: "1001"
>>> len(feed.entity[0].trip_update.stop_time_update)
74
```

There are 74 stops made on this trip so far, and we look at the first two stops:

```
>>> feed.entity[0].trip_update.stop_time_update[0]
stop_sequence: 9
arrival {
  time: 1563818515
}
departure {
  time: 1563818515
}
stop_id: "LIVNOEW"
>>> ft.entity[0].trip_update.stop_time_update[1]
stop_sequence: 10
arrival {
  time: 1563818711
}
departure {
  time: 1563818711
}
stop_id: "LIVCOUNW"
```

Based on the difference in departure times between the two stops, the data show that the bus arrived at the second stop (coded "LIVCOUNW") after 156 s (3.3 min). Each stop has its unique code, and COTA maintains a master file for all the stops,[5] where each stop is associated with a set of attributes that include the address and coordinates.

With the above examples, it is clear that at a specific time and location, each bus is associated with certain attributes such as the trip information and speed, which can be encoded as an STA tuple. The same can be said about stops that are made by the busses. We can then write a program that automatically requests the real-time data for bus positions and stop updates at a desirable time interval (every second, for example). The information retrieved can then be recorded in a database where each record is an STA tuple $(x, t, a)$. For the buses, for example, each record contains fields such as latitude, longitude, timestamp, vehicle ID, trip ID, bearing, along with any other information that is deemed to be useful. For each stop, we can do the same by

[5] https://github.com/joeshaw/cota-bus/blob/master/cota-gtfs/stops.txt.

recording fields such as the coordinates, arrival and departure times, trip ID, vehicle ID, and stop ID. The accuracy of the database is partly dependent on the time interval of data collection. A one-minute time interval may be sufficient for the purpose of information visualization and some analysis, and a smaller interval will be needed if we aim to provide real-time service to the general public for tasks such as trip planning that require higher accuracy.

The Environmental Protection Agency (EPA) of the USA maintains a network of air quality sensors across the country. EPA also provides an API to allow users to access air quality data.[6] This API provides a Web service based on a software architecture called REST (Richardson and Ruby 2008) that supports the use of a URL to query a database in order to retrieve data. For example, the following URL specifies the time frame, geography boundaries, and environment variable, along with other necessary parameters. The last parameter must be replaced by an actual API key that can be applied from the Web site.

```
https://airnowapi.org/aq/data/?
    parameters=pm25&
    bbox=-83.368244,39.586371,-82.269611,40.344184&
    startDate=2019-05-19T03&endDate=2019-05-19T04&
    DataType=B&format=application/json&verbose=1&
    API_KEY=XXXX
```

This request will return the following data formatted in JSON. It shows that during the two-hour time frame specified, there are two PM2.5 sensors at two locations, and their data (e.g., locations, values, air quality index values) are provided. Again, we can write a program that automatically and repeatedly retrieves information like the above as STA tuples and store them into a database.

```
[
  {
    "Latitude": 40.11109, "Longitude": -83.065376,
    "UTC": "2019-05-19T03:00",
    "Parameter": "PM2.5",
    "Unit": "UG/M3", "Value": 14.8, "AQI": 57, "Category": 2,
    "SiteName": "Columbus NR - Smoky Row",
    "AgencyName": "Ohio EPA-DAPC",
    "FullAQSCode": "390490038", "IntlAQSCode":
"840390490038"
  },
  {
    "Latitude": 40.0845, "Longitude": -82.81552,
    "UTC": "2019-05-19T03:00",
    "Parameter": "PM2.5",
    "Unit": "UG/M3", "Value": 12.2, "AQI": 51, "Category": 2,
    "SiteName": "New Albany",
    "AgencyName": "Ohio EPA-DAPC",
    "FullAQSCode": "390490029", "IntlAQSCode":
"840390490029"
  },
```

[6] https://docs.airnowapi.org.

```
  {
    "Latitude": 40.11109, "Longitude": -83.065376,
    "UTC": "2019-05-19T04:00",
    "Parameter": "PM2.5",
    "Unit": "UG/M3", "Value": 14.7, "AQI": 56, "Category": 2,
    "SiteName": "Columbus NR - Smoky Row",
    "AgencyName": "Ohio EPA-DAPC",
    "FullAQSCode": "390490038", "IntlAQSCode":
"840390490038"
  },
  {
    "Latitude": 40.0845, "Longitude": -82.81552,
    "UTC": "2019-05-19T04:00",
    "Parameter": "PM2.5",
    "Unit": "UG/M3", "Value": 12.1, "AQI": 51, "Category": 2,
    "SiteName": "New Albany",
    "AgencyName": "Ohio EPA-DAPC",
    "FullAQSCode": "390490029", "IntlAQSCode":
"840390490029"
  }
]
```


The raw data collected in the above examples are merely STA tuples of the form (*x, t, a*) and must be processed to support purposes such as analyzing urban traffic status or mapping density of air pollution. In a bigger context, this is an area of data mining of big data (Vatsavai et al. 2012). In our example of using the GTFS feeds, two kinds of real-time raw data are acquired: vehicle positions and stop updates. Among all the GTFS text files, the file called `stop_times.txt` is used to store the bus schedule for all routes, containing detailed arrival and departure time as scheduled for each stop on each trip. By comparing the real-time trip updates of the actual arrival and departure time of each trip with the scheduled times, it is possible to compute the delay of each bus and conduct further analysis of how the delays propagate along the trip (Park et al. 2019). It is also possible to visualize the discrepancy in places that can be reached by the scheduled and actual buses (Fig. 31.1).

The above data collection examples show the general procedure of harvesting urban big data and the considerations of storing them in spatiotemporal databases. There are of course many other sources for urban big data that are designed for different purposes (e.g. Twitter data). Though these data sets differ in technical details such as data format and APIs, it can be argued that STA tuples can be used to capture most (if not all) of these data sets. To this extent, from a data perspective alone, it suffices to say that the data are "out there" for users to use. The real and more difficult challenge is how to make these data accessible to all.

## 31.6 Toward Urban Big Data Infrastructure

Urban big data as described above have the necessary elements to support the user stories described in the previous section of this paper. These data sets are also relatively straightforward to obtain. However, it should also be clear that the ecosystem

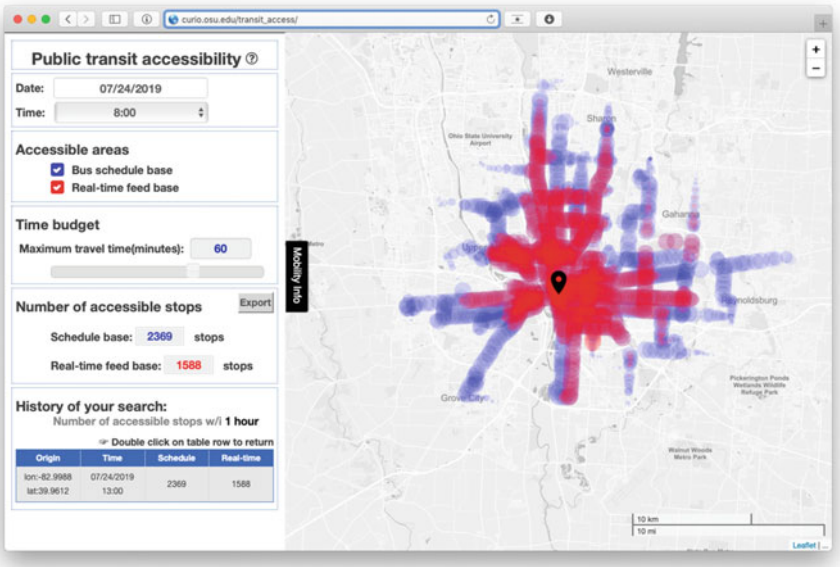


**Fig. 31.1** Visualizing the difference between the scheduled stops (blue) and those that were actually reached (red) in a one-hour time frame from a given location (black pin icon). *Source* http://curio.osu.edu/transit_access/

of urban big data does not always suit regular users from the general public, who are often not trained to be as data savvy as the experts who generate the data. The difficulty these regular users may face can be as simple as where to find the data and as complicated as how to use them. These are the major limitations that make it difficult for the data to be accessible to a wide audience.

To address these problems, we advocate the idea of urban big data infrastructure under the spirit of data for all. The concept of infrastructure refers to the ubiquitous availability of resources such as electricity where a person, who does not need to be an electricity expert, can use it by simply plugging in. We would ponder if it is possible for a regular user to find a desired spatiotemporal data set by specifying it instead of by carrying out a process of searching and coding. For example, is it possible to ask a virtual assistant (e.g. Apple's Siri) on a smartphone to find the spatiotemporal data set by giving a description of the data? In the remainder of this section, we review some methods that may shed light in the future development of such an infrastructure.

There are a few existing methods that can be used to address *some* of the issues mentioned above. A geoportal (Tait 2005), for example, is designed as a gateway to serve geospatial data on a Web-based platform. More specifically, a geoportal can be used to allow users to do the following tasks:

- Discover geospatial data based on a catalog of the data maintained in the geoportal.
- Provide useful information about how to use each of the geospatial data sets.

- View and map the data sets discovered.
- Automatically harvest (collect) online data sources and store them in the geoportal for further uses.
- Provide data using various data query techniques such as REST, GeoRSS, and KML.

The implementation of a geoportal requires work on the server side and is suitable as a solution to data needs at the enterprise level. Ideally, by logging into a geoportal, a user can find relevant data sets and explore the properties of those data through mapping, tabulating, or simply describing the data. However, these geoportals are usually developed for data experts to use instead for the regular users, who may not have the necessary skill sets in understanding the portal and navigating the numerous data sets served. It is also difficult to expect users to develop their own geoportals or to develop data sets within existing portals. In this sense, the ultimate users (the general public in our case) are entirely at the mercy of the data experts or data enterprises.

Another approach is spatial data infrastructure (SDI). The term often involves technologies for data collection and retrieval, along with metadata, as well as policies that promote access to spatial data. For this reason, SDIs are not technological solutions to data problems but more of a social and political response to the data needs that emerge from communities at different scales. In an ideal situation, implementing an SDI requires the efforts of government agencies, the private sector, representatives of the general public, and even members of academia. In the past, SDIs have been effective in consolidating traditional data sets such as the cadastre, national base maps, large-scale topographic maps, and remotely sensed images. While it is well recognized that the success of SDIs is critically dependent on how the users, citizens, and institutions are engaged, their involvements have been a significant challenge (Erik de Man 2006; Elwood 2008). It should be noted that a major portion of the SDI literature is focused on the technological aspects, especially taking a GIS-centered perspective (Maguire and Longley 2005; Steiniger and Hunter 2012; Evangelidis et al. 2014; Helmi, Farhan and Nasr 2018). Through such a technological perspective, unfortunately, the concept of SDI tends to be reduced to merely a form of GIS or geoportal.

We argue that it is necessary to develop an urban big data infrastructure in order to address the issues discussed above and to fulfill the goals of using the data as mentioned in the user stories. The technical aspects of such an infrastructure, though still challenging, can be relatively straightforward, as much of the effort has already focused on how to utilize the technology in getting the data and making the data accessible. For example, the development of geoportals has already demonstrated that various data can be incorporated in commonly used formats and standards for users to discover and use. Many geospatial database management systems (e.g. GeoServer and Esri's geoportal) can be used to harvest data from different sources. More importantly, these systems typically also support data discovery. For example, Catalogue Services[7] is a specification standard proposed by the Open Geospatial Consortium

[7] https://www.opengeospatial.org/standards/cat.

(OGC) and has been supported by major software systems such as GeoServer[8] and Esri's geoportal.[9]

The fundamental challenge of developing urban big data infrastructures goes beyond the technological domain: It is the often ill-defined relationship among data, data providers, data users, and software developers and vendors that makes it difficult for such an infrastructure to be effective, as shown in the case of SDIs. From an engineering perspective, this challenge is due to the changing requirements as new user stories emerge whenever new data sources or new technology become available. There is no silver bullet that will solve all the problems. Instead, it is important to understand that a fully functional urban big data infrastructure (or SDIs at a lesser level of difficulty) takes time and must wait for collaborations to emerge.

We envision an agile process (Stellman and Greene 2014) where all parties involved in the use and production of urban big data will constantly engage with each other and revise any previous understandings about the data, even though the understandings may be preliminary and sometimes trivial at the early stages of development. A top-down approach to developing the infrastructure is bound to fail since such an approach is typically dependent on well-defined requirements, as shown repeatedly in the history and literature of software engineering (Sommerville 2016). The strong social and human aspects of urban big data infrastructure make it natural to consider an agile approach that stresses how the development process should actively engage with the system (data) users (Stellman and Greene, 2014). A typical agile development process starts from user stories that roughly but meaningfully describe the fundamental requirements of a system but often do not specify the details of how the system should be run and built. In order for the project to advance, the end user or client must constantly be involved in the process and provide feedbacks so that the requirements can become increasingly clear. Lack of user involvement will cause adverse consequences to both the team and the project (Hoda et al. 2011). User involvement in turn helps the developers understand the direction of the project and enables them to work together with the users, toward the end product.

Among the many agile methods, self-organizing agile methods are a promising recent development that have gained much recognition (Hoda et al. 2012) and can be especially suitable for the development of urban big data infrastructures. Researchers have studied the potential of such an approach from different perspectives, including organizational theory that focuses on how organizations may learn from past experience (Morgan 1998) and complex adaptive systems that show how feedback among individuals can help the system evolve (Lansing 2003). In addition to the customer/user, a regular agile team includes a product owner who maintains a close relationship with the customer and plays the role of a stakeholder, a coordinator (scrum master) who operates the daily routines of the team and keeps the team together, and team members who are dedicated to work on various parts of the project with a strong leadership from the coordinator and product owner. In the case of a self-organizing agile method, a team may still have those roles among team members,

[8] https://docs.geoserver.org/latest/en/user/services/csw/index.html.

[9] https://www.esri.com/en-us/arcgis/products/geoportal-server/overview.

but is a more autonomous group where the role of each member may change. A strong point of such an approach is that decisions about the project are made not by the product owner but more spontaneously from the collaborations among all team members, and more importantly with the customer (Hoda et al. 2011).

The key aspect of a self-organizing agile process is the collaborative leaders who play the most critical role. In the agile literature, these are team members who act as mentors and coordinators. Mentors are not bosses because they do not make decisions; instead, they are coaches who provide guidance and support the team's confidence. Coordinators are essential too because they work directly with users in order for the development to be on the right track as the users require.

Self-organizing agile methods are promising, and it should be noted that the development of an urban big data infrastructure will not emerge just because there are demands from users and data experts. Strong bonds between them are important, and leadership is required. We do not imagine that an infrastructure can be developed over just a few projects where big data are involved. Instead, given the fact that SDIs are still far from being functional despite the efforts of the past three decades (Erik de Man 2006; Grus et al. 2010), it is reasonable to believe that a fully functional urban big data infrastructure will also take a long time to materialize. However, with strong and collaborative leadership formed through the bond between the user (demand) and the developers (skills), it is possible to evolve the infrastructure through multiple projects where data and knowledge derived from the use of data will accumulate. An open and collaborative environment will be especially useful at the urban scale where similar tasks may repeat in different urban areas and therefore good practices can be adopted and improved through time.

## 31.7 Concluding Remarks

Urban big data have exhibited potential in helping us to better understand the city and make better and informed decisions. Such data have a wide range of sources, and the technology to retrieve the data is relatively straightforward. However, the social and human aspects have made the use of the data by the general public a real challenge. Cultivating urban big data requires long-term planning and sustainable collaboration between many parties. It is not reasonable to expect silver bullet solutions.

Technology aside, data have become the cornerstone of an ecosystem that is sustained by a chain of users, developers, companies, analysts, and investors. The roles of each player in this ecosystem are not the same as in the old economy. For example, while users are still using the services provided by companies such as Google and Facebook, they also contribute to data collection through using the Internet (e.g. conducting searches or posting on social media). To some extent, this era of urban big data is also an era where users act as products. Schneier (2015) describes the relationship between the (private) data provider and users as a feudalist system where the data "lords" have full and firm control on the properties (data) that are similar to the land in a feudal system, and the users receive benefits from the data

"lords" through payment or other types of contribution (their own data, for example), similar to peasants in a feudal system who must trade their labor in order to have access to land and services. We do not believe such a feudalist world in the data domain is healthy for data to be used to its optimal extent. Through collaboration and policy, we can develop an open (though not necessarily free) urban big data infrastructure that will enable the data to be used by their true constituents: the general public.

## References

Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Networks 54(15):2787–2805

Boulos MNK, Resch B, Crowley DN, Breslin JG, Sohn G, Burtner R, Pike WA, Jezierski E, Chuang KYS (2011) Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int J Health Geogr 10(1):67

Chen C, Zhang C (2014) Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Inf Sci 275:314–347

Cuff D, Hansen M, Kang J (2008) Urban sensing: out of the woods. Commun ACM 51(3):24–33

Elwood S (2008) Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing. Int J Geogr Inf Sci 22(1):71–90

Erik de Man WH (2006) Understanding SDI; complexity and institutionalization. Int J Geogr Inf Sci 20(3):329–343

Evangelidis K, Ntouros K, Makridis S, Papatheodorou C (2014) Geospatial services in the cloud. Comput Geosci 69:116–122

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Goodchild MF, Yuan M, Cova TJ (2007) Towards a general theory of geographic representation in GIS. Int J Geogr Inf Sci 21(3):239–260

Grus L, Crompvoets J, Bregt A (2010) Spatial data infrastructures as complex adaptive systems. Int J Geogr Inf Sci 24(3):439–463

Günther WA, Mehrizi MHR, Huysman M, Feldberg F (2017) Debating Big Data: a literature review on realizing value from Big Data. J Strategic Inf Syst 26(3):191–209

Harrelson C (2006) Happy trails with google transit. Google Official Blog (https://googleblog.blogspot.com/2006/09/happy-trails-with-google-transit.html). Accessed: July 25 2019

Heipke C (2010) Crowdsourcing geospatial data. ISPRS J Photogrammetry Remote Sens 65(6):550–557

Helmi AM, Farhan MS, Nasr MM (2018) A framework for integrating geospatial information systems and hybrid cloud computing. Comput Electrical Eng 67:145–158

Hoda R, Noble J, Marshall S (2011) The impact of inadequate customer collaboration on self-organizing agile teams. Inf Softw Technol 53(5):521–534

Hoda R, Noble J, Marshall S (2012) Self-organizing roles on agile software development teams. IEEE Trans Softw Eng 39(3):422–444

Kim GH, Trimi S, Chung JH (2014) Big-data applications in the government sector. Commun ACM 57(3):78–85

Kitchin R (2014) The real-time city? Big Data and Smart Urbanism. GeoJournal 79(1):1–14

Kitchin R, McArdle G (2016) What makes Big Data, Big Data? exploring the ontological characteristics of 26 datasets. Big Data and Society 3(1):1–10

Kumar P, Morawska L, Martani C, Biskos G, Neophytou M, Di Sabatino S, Bell M, Norford L, Britter R (2015) The rise of low-cost sensing for managing air pollution in cities. Environ Int 75:199–205

Lansing JS (2003) Complex adaptive systems. Ann Rev Anthropol 32(1):183–204

Lewis A, Edwards P (2016) Validate personal air-pollution sensors. Nature News 535(7610):29–31

Maguire DJ, Longley PA (2005) The emergence of geoportals and their role in spatial data infrastructures. Comput Environ Urban Syst 29(1):3–14

Marr B (2015) Big Data: using SMART Big Data, analytics and metrics to make better decisions and improve performance. Wiley, Hoboken

McHugh B (2013) Pioneering open data standards: THE GTFS story. In: Goldstein B (ed) Beyond transparency: Open Data and the future of civic innovation. Code for America Press, San Francisco, pp 125–135

Mergel I, Rethemeyer RK, Isett K (2016) Big Data in public affairs. Public Adm Rev 76(6):928–937

Morgan G (1998) Images of organization. SAGE, Thousand Oaks

Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. In: Proceedings of the seventh international AAAI conference on weblogs and social media. Association for the Advancement of Artificial Intelligence, pp 400–408

Park Y, Mount J, Liu L, Xiao N, Miller HJ (2019) Assessing public transit performance using real-time data: spatiotemporal patterns of bus operation delays in Columbus, Ohio, USA. Int J Geogr Inf Sci 34(2):367–392. https://doi.org/10.1080/13658816.2019.1608997

Richardson L, Ruby S (2008) RESTful Web services. O'Reilly Media, Sebastopol

Schneier B (2015) Data and Goliath: the hidden battles to collect your data and control your world. WW Norton, New York

Sommerville I (2016) Software Engineering. Pearson, London

Steiniger S, Hunter AJ (2012) Free and open source GIS software for building a spatial data infrastructure. In: Bocher E, Neteler M (eds) Geospatial free and open source software in the 21st century. Springer, Dordrecht, pp 247–261

Stellman A, Greene J (2014) Learning agile: understanding scrum, XP, lean, and kanban. O'Reilly Media, Sebastopol

Tait MG (2005) Implementing geoportals: applications of distributed GIS. Comput Environ Urban Syst 29(1):33–47

Vatsavai RR, Ganguly A, Chandola V, Stefanidis A, Klasky S, Shekhar S (2012) Spatiotemporal data mining in the era of big spatial data: algorithms and applications. In: Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data. ACM. Washington, pp 1–10

Yi W, Lo K, Mak T, Leung K, Leung Y, Meng M (2015) A survey of wireless sensor network based air pollution monitoring systems. Sensors 15(12):31392–31427

**Ningchuan Xiao** is Professor of Geography and Associate Director of the Center for Urban and Regional Analysis (CURA) at The Ohio State University. He is interested in spatial data science and technology.

**Harvey J. Miller** is the Bob and Mary Reusche Chair in Geographic Information Science, Professor of Geography, and Director of the Center for Urban and Regional Analysis (CURA) at The Ohio State University. His research interests include mobility analytics, sustainable transportation, and time geography.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.