Bulletin of the Technical Committee on

Data Engineering

December 2014 Vol. 37 No. 4

IEEE Computer Society

Letters

Special Issue on Urban Informatics

Conference and Journal Notices

The TC on Data Engineering

Membership in the TC on Data Engineering is open to all current members of the IEEE Computer Society who are interested in database systems. The TCDE web page is http://tab.computer.org/tcde/index.html.

The Data Engineering Bulletin

The Bulletin of the Technical Committee on Data Engineering is published quarterly and is distributed to all TC members. Its scope includes the design, implementation, modelling, theory and application of database systems and their technology.

Letters, conference information, and news should be sent to the Editor-in-Chief. Papers for each issue are solicited by and should be sent to the Associate Editor responsible for the issue.

Opinions expressed in contributions are those of the authors and do not necessarily reflect the positions of the TC on Data Engineering, the IEEE Computer Society, or the authors' organizations.

The Data Engineering Bulletin web site is at http://tab.computer.org/tcde/bull_about.html.

Letter from the Editor-in-Chief

Delayed Publication

This December, 2014 issue of the Bulletin is, as some of you may notice, being published in July of 2015, after the March and June, 2015 issues have been published. Put simply, the issue is late, and the March and June issues were published in their correct time slots. The formatting of the issue, and the surrounding editorial material, e.g. the inside front cover and copyright notice, are set to the December, 2014 timeframe. Indeed, the only mention of this inverted ording of issues is in this paragraph. Things do not always go as planned. However, I am delighted that the current issue is being published, and I have high confidence that you will enjoy reading about urban informatics, the topic of the issue.

The Current Issue

We all want our cities to be more livable and their citizens to be more informed. Importantly, we in the high tech world are in a position to help. Computers, and more precisely, information technology, can help our cities work better, whether it be in planning urban trips or urban planning, managing data and making it suitably available are key elements in these activities. And it is the handling of data related to our increasingly urban world that is the subject of the current issue.

It is tempting to believe that the technology that is used by commercial enterprises can be exploited to deal with this urban context. And, of course, there is some truth to that belief. But in a large application area such as urban informatics, important issues unique to this application arise. This may be due to legal issues, but as important, it is relevant to providing good and effective governance. And the sheer size and complexity of the urban environment poses its own unique challenge.

David Maier has worked with V. M. Megler and Kristin Tufte to assemble this current issue devoted to an exploration of urban informatics. Any such issue can only sample this very large space. Our editors have brought together papers that not only provide a sample, but also illustrate where technology might be taking us. There is a unique quality to this issue as most of us, being technology centric, are unfamiliar with the application of information technology to urban issues. Not only is this issue of interest for its ability to educate us about urban issues, but even more importantly, it may succeed in engaging us to help further the application of informatics to the problems we face every day in our own urban environments. Thanks to our three editors for bringing us this issue which is on a subject that is important and relevant to us all.

David Lomet
Microsoft Corporation

Letter from the Special Issue Editors

Most data related to people and the built world originates in urban settings. There is increasing demand to capture and exploit this data to support efforts in areas such as Smart Cities, City Science and Intelligent Transportation Systems. *Urban informatics* deals with the collection, organization, dissemination and analysis of urban information used in such applications. However, the dramatic growth in the volume of this urban data creates challenges for existing data-management and analysis techniques. The collected data is also increasingly diverse, with a wide variety of sensor, GIS, imagery and graph data arising in cities. To address these challenges, urban informatics requires development of advanced data-management approaches, analysis methods, and visualization techniques. It also provides an opportunity to confront the "Variety" axis of Big Data head on.

The contributions in this issue cross the spectrum of urban information, from its origin, to archiving and retrieval, to analysis and visualization.

The first two papers deal with sources (and uses) of urban data. The first (by Ilarri et al.) looks at new collection modalities for urban data, specifically collaborative sensing for urban transportation. It classifies different sensor types, and describes their use in parking, traffic-information and trajectory-data settings. It also considers challenges and opportunities in this area, including encouraging user cooperation, data quality and management of trust and privacy. The second (by Sahuguet et al.), on Open Civic Data, focuses on the (re)use of data that has already been collected, often by public agencies, utilities and other civic institutions. It begins with examples of the use of civic data in areas as diverse as identifying mortgage fraud, economic development and crisis management (and also recounts some "horror stories" involving civic data). The authors then turn to the technical challenges of Open Civic Data, including actually opening up the data, making it discoverable, and handling time and space as first-class citizens (a theme that runs through all the papers).

The third paper (by Catlett et al.) concentrates on storing and serving open information that comes from municipal, state and federal sources, in a way that supports easy data discovery and exploration. The authors present Plenario, a platform for making such information available to users who are not expert data scientists, and who are not prepared to spend days or weeks in preparing data. They report on two early use cases for Plenario, the lessons learned from them, and challenges and opportunities for further development.

The fourth paper looks at improving analysis of urban data. The authors want to go beyond batch-oriented queries aimed at a particular pre-determined questions to more interactive and open-ended exploration of urban data sets. They describe a visual query interface that is particularly oriented to discerning patterns in the kinds of spatial and temporal datasets that are ubiquitous in the urban setting. The paper illustrates the approach with taxi-trip records (an example of Open Civic Data), and describes an indexing technique that support working with spatial-temporal data at scale.

We hope this selection of papers helps gives readers a flavor of the range of applications and challenges arising in urban informatics, and perhaps inspires some to take up research in this area.

David Maier, V. M. Megler, Kristin Tufte
Portland State University

Collaborative Sensing for Urban Transportation

Sergio Ilarri[1], Ouri Wolfson[2], Thierry Delot[3],
[1]University of Zaragoza, Spain, silarri@unizar.es
[2]University of Illinois, Chicago, USA, wolfson@cs.uic.edu
[3]University of Valenciennes, France, Thierry.Delot@univ-valenciennes.fr

Abstract

In this paper, we overview the current status of research in the field of collaborative sensing for urban transportation. We first classify the different types of sensors that can be relevant in collaborative urban sensing and then we analyze three different topics: parking spaces, traffic, and trajectories. We discuss issues regarding the sensing and data sharing approaches as well as relevant use cases. Finally, we identify some research challenges and trends.

1 Introduction

A key element for urban informatics [1] is the widespread availability of sensors of different types, which can capture a wide range of environmental conditions and information of the surroundings of the user. Besides traditional wireless sensor networks, where sensors are statically deployed over a fixed area to measure certain values, mobile and urban computing scenarios open up new opportunities for more flexible and dynamic approaches. Collaborative sensing (also called mobile crowdsensing or participatory sensing) [2, 3] implies using sensors in a cooperative way in order to obtain an overall and more complete perspective of the environment. Specifically, the new scenarios support the use of *mobile sensors*. On the one hand, users with their mobile devices have become an important source of sensor data [4]. On the other hand, vehicles can also carry a variety of sensors ("Today's luxury cars have more than 100 sensors per vehicle" [5]).

Collaborative sensing can provide key benefits in urban transportation, contributing to higher travel efficiency, safety, and reduced pollution, through innovative applications that benefit from the data sensed collaboratively. These benefits apply to private transportation as well as public transportation. The collaborative aspect remarks the idea that the role of people is particularly important. For example, the work presented by Campbell et al. adopts a human-centric (or people-centric) view of urban sensing [6] , which emphasizes the significance of the attributes of people, their surroundings, and their interactions with the environment.

In this paper, we survey the state of the art in collaborative sensing for urban transportation. In Section 2 we overview the different types of sensors that could be of relevance. In Section 3 we focus on the problem of sensing and sharing information about parking spaces, as a prime example of a scarce resource in urban transportation (bike stations could be another example). In Section 4 we analyze the problem of traffic estimation from a cooperative sensing perspective. In Section 5 we tackle the management of trajectories. Finally, in Section 6 we summarize some existing challenges. Figure 1 shows an overview of the main topics covered.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

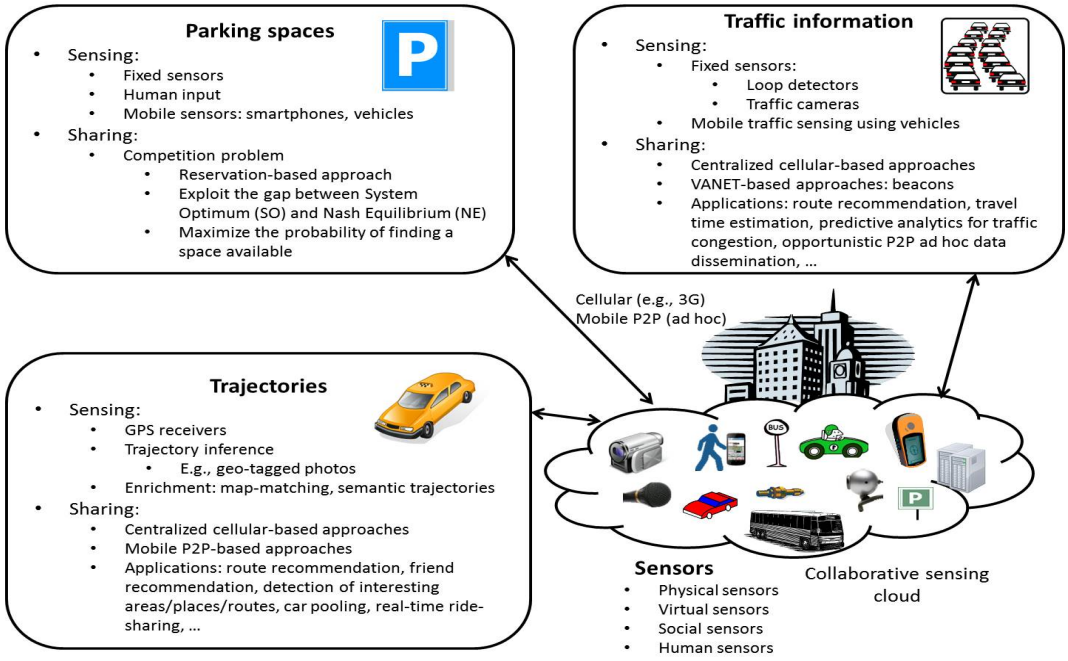


Figure 1: Main topics in collaborative sensing for transportation

2 Types of Sensors

According to the abstraction level of the information they manage, several types of sensors can be distinguished:

- *Physical sensors*. These are traditional *hardware sensors*, which are able to directly capture raw data from the environment. Sensing capabilities can be embedded in existing mobile devices (e.g., smartphones) or as stand-alone measuring devices that communicate their data to other devices or computers. As an example, mobile devices are being equipped with an increasing number of built-in sensors: GPS, microphone, camera, ambient light sensor, accelerometer, gyroscope, compass, proximity sensor, temperature, etc. These mobile devices can be carried by people or be integrated in vehicles (cars, taxis, buses, etc.).

- *Virtual sensors*. These are *software sensors*, which do not directly correspond to physical sensors. Instead, they usually provide higher-level data that are obtained by combining the output of several sensors. A virtual sensor [7] is an abstraction that aggregates data provided by different (and possibly heterogeneous) sensors to compute a virtual measurement (e.g., a virtual position sensor that transparently uses different positioning mechanisms, as needed, to locate the user).

- *Social sensors*. These are sensors that provide data based on information extracted from the social media: social networks (e.g., Facebook, Foursquare, Flickr), blogs, microblogs (e.g., Twitter), etc. For example, Albakour et al. [8] exploit microblogs to detect events in the vicinity, and Sakaki et al. [9] extract data from tweets to infer heavy traffic conditions for drivers. Another prominent effort is the EU FP7 project *SocialSensor* [10].

- *Human sensors*. In the extreme, we can consider humans as sensors, as they themselves can provide interesting measures through explicit cooperation. Humans can use their own senses and provide information explicitly, or manage the sensors that they have available in specific ways to collaborate in the measuring task. As an example, humans can provide *volunteered geographic information (VGI)* [11]. As another

example, the idea of *spatial crowdsourcing* [12, 13, 14] is a hot topic. Drivers can also provide information about parking spaces (see Section 3). Although in some cases social sensors could also be considered as human sensors (e.g., humans can report what they see by sending tweets), we make a distinction by requiring explicit (not indirect or casual) cooperation from people in the case of human sensors.

The previous classification is to some extent similar to the one provided for location sensors in the work by Indulska et al. [15]. According to that work, physical location sensors provide the location of a physical device, virtual location sensors extract information from the virtual space (software applications, operating systems, and networks), and logical location sensors exploit virtual and physical location sensors to infer physical locations. On the other hand, Islam et al. [16] distinguish between *hard sensors* (physical sensors) and *soft sensors* (which actually represent trigger conditions, such as the reception of an email or a new entry in the user's calendar).

3 Parking Spaces

A key element to improve urban transportation is to provide a suitable management of parking spaces. Among the inconveniences of searching for parking, we can highlight that it is usually a time-consuming task, stressful and frustrating for drivers, and it obviously contributes to increasing traffic and pollution. A well-known report by Shoup [17] quantifies the yearly cost of parking in a commercial district next to the campus of the University of California in Los Angeles (Westwood Village): 47000 gallons of gasoline, 730 tons of CO_2 emissions, and 95000 hours of drivers' time. According to a study in France [18], searching for parking represents between 5% and 10% of the traffic in urban areas and it can rise up to 60% in the case of small streets. As a final example, a study performed in the district of Schwabing in Munich (cited in [19]) indicates that about 44% of the traffic is searching for an available parking space. These and many other reports emphasize the importance of promoting an efficient management of parking spaces for sustainable transportation in cities.

For some types of parking, such as parking garages and paid parking lots, an existing infrastructure can accurately monitor in real-time the occupancy of parking spaces and even provide effective reservation and payment methods. However, in the general case, such as in the case of free street parking, a fixed sensor-based infrastructure to detect the occupancy of parking spots (a popular example is SFPark in San Francisco, see `http://sfpark.org/`, which covers 6000 metered spaces and 12250 spaces in parking garages [20]) usually does not exist or it is considered expensive to be deployed at a large scale (e.g., see [21, 22, 23, 24]). Therefore, collaborative parking sensing is a promising avenue to explore.

It should be noted that a variety of methods could be used to capture information about the availability of a parking space. Besides traditional approaches using fixed sensors (e.g., SFPark), smartphone-based techniques are being developed to automatically detect parking and unparking events without user intervention [22, 23, 24, 25, 26]. Other approaches consider the vehicles themselves as mobile sensors that detect surrounding parking spaces [21]. Finally, the driver himself or herself could also provide the information manually (as illustrated in the work by Hoh et al. [27] or by a variety of ephemeral applications such as Google's Open Spot, which is no longer available).

Once the information about parking spaces is collected, the problem is how to effectively share it. Parking spaces represent a scarce resource, and therefore if information about their availability is disseminated freely among the potentially interested vehicles there is a high risk of generating a competition among them to try to get the same parking space. There is a trade-off between competition and information dissemination [28], such that eventually the time to find an available parking space could actually increase as compared to the situation of blind search where no information is provided. Different approaches have been proposed to manage this problem:

- Some approaches have proposed reservation protocols to guarantee the use of the parking space to only a specific vehicle (e.g., the *Centralized Assisted Parking Search* approach [28]). They rely on an infrastruc-

ture for reservation, payment and monitoring of the parking spaces, and so their use is limited to the case of controlled parking facilities. An exception is the reservation protocol for VANETs (Vehicular Ad Hoc Networks) proposed by Delot et al. [29], which uses the term *reservation* in a metaphorical way. What it actually does is to provide the information about an available parking space to a single driver); without the use of an infrastructure, ensuring that no other vehicle can take up the "reserved" space is not possible.

- Other approaches exploit the gap between the System Optimum (SO) and Nash Equilibrium (NE) in matching vehicles and parking slots [30]. SO is more efficient (i.e., it involves minimum total travel-time to slots), but an uncoordinated system of drivers acting selfishly settles into an NE state. One approach to bridge the gap is to price the slots such that an NE matching when considering travel-time and price is the same as an SO matching when considering travel-time alone [31]. Another approach involves payments among drivers looking to park, through a platform, to convert one matching to another [31, 32]. In other words, through payments, selfish drivers are incentivized to behave in a manner that is globally and environmentally responsible. Interestingly, payments raise the issue of truthful location disclosure by a driver, and Vickrey-Clarke-Groves (VCG) mechanisms from economics can be adapted to incentivize drivers to be truthful [32].

- Some approaches try to maximize the probability of finding an available parking space. For example, one alternative is to compute a route that goes through all the available parking spaces [33]. As another example, the Gravity-based Parking Algorithm (GPA) [30] is a gravitational model heuristic where each parking space applies an attraction force on a vehicle searching for a parking space. The magnitude of the force is inversely related to the distance between the space and the vehicle. The force-vectors on the vehicle are added, thus areas with a higher density of parking spaces become more attractive. In the approach proposed by Klappenecker et al. [34], parking lots periodically disseminate certain status parameters (such as their capacity and the number of occupied parking spaces) that vehicles can use to estimate the probability of finding there an available parking space at the time of arrival. Other proposals highlight the interest of guiding drivers towards areas where the probability to find an available parking space is high [35], instead of towards a specific parking space. Zekri et al. [36] present an approach that aggregates information about available parking spaces to extract general knowledge about their spatio-temporal availability.

Although most studies have focused on effectively exploiting availability data about parking spaces to minimize the parking time, the communication overhead should not be overlooked. As an example, a survey concerning the Montreal Region [37] (an area of 5500 km^2) reports 80500 parking events per day during the year 2003. As another example, from the analysis of a dataset of SFPark we conclude that an average of 56.8 records (events) per city block are generated each day in San Francisco; the raw data of each record has a size of 129.7 bytes on average (the names of the blocks have different lengths) which can be reduced (filtering out unnecessary or redundant information) for later processing to an average of 33 bytes of useful data per record. If a high number of messages about parking and unparking events are exchanged among vehicles with no control, this could easily lead to network overload. If we imagine a country or world-wide cloud-based system, the volumes of data in such a system may be particularly significant.

4 Traffic Information

For the collection of traffic information, different methods have been traditionally used. For example, for speed or traffic density estimation static devices and sensors such as single and double loop detectors or traffic surveillance cameras can be exploited. In the last years, rather than relying on sensing approaches based on the use of a fixed infrastructure, which is usually an expensive and non-flexible approach with limited coverage (e.g., see [38]), other alternatives propose the use of vehicles as mobile sensors.

Thus, *probe vehicles* (e.g., taxis and buses) could collect traffic data in a city. For example, they could periodically report their speed and location to a central server (using cellular communication) for later aggregation [38]. This collaborative traffic sensing is a promising approach, but it also faces important challenges: the data are sensed opportunistically as the vehicles move through the city, and therefore the data collected may exhibit significant gaps in time and space, as the spatio-temporal distribution of the vehicles is not uniform (e.g., compressive sensing could be used to deal with missing values [38]). Vehicles can use different types of sensors to obtain the relevant information; for example, VTrack [39] exploits WiFi location samples to estimate the location of the user (with less energy consumption than in a GPS-based solution) and identify travel-time patterns along the road segments. Besides probe vehicles, *floating cars* (such as patrol cars for surveillance) can also play the role of mobile sensors [42] for traffic estimation: the path of these floating vehicles could be controlled and adjusted according to the traffic monitoring requirements.

The wireless bandwidth used for the transmission of traffic information is a resource to minimize [40, 41]. For example, Ayala et al. [41] proposed a *flow-based update policy* where cars use a transmission probability computed based on the number of messages that the server expects to receive from the vehicles in each road segment in order to ensure a desired accuracy for the average speeds computed. They performed experiments based on data generated by a sensor at the end of a road segment on highway 190 in Chicago from 6AM to 8AM on May 22nd, 2007 (Tuesday). In the experiments presented, fewer that 500 messages are required by the proposed policy, whereas the number of messages goes up to about 3000 with an alternative *information cost based policy* (based on the existing trade-off between the communication cost and the data uncertainty) and about 5500 for a *deterministic policy* (a vehicle transmits its velocity to a server if the difference between the broadcasted velocity received from the server and the measured velocity exceeds a certain threshold), considering a data collection period of 300 seconds and a velocity threshold of 1 mph for the information cost based policy and deterministic policy; the benefits of the proposed policy are also shown for other combinations of thresholds.

Whereas traditional mobile traffic sensing approaches using vehicles are centralized and usually use cellular communications, other novel solutions have been proposed in the field of VANETs, which exploit ad hoc communications in a distributed way. In the proposal presented by Sanguesa et al. [43], vehicles in urban areas use information of the beacons received from other neighboring vehicles, as well as features extracted from digital maps, to estimate the density of vehicles in their neighborhood; in the experimental evaluation performed, the authors infer that the optimal time period to estimate the density in vehicular environments is 30 seconds. To offer an overall picture rather than just a local density estimation, some proposals advocate exploiting a complementary fixed-support infrastructure. Thus, the V2X-D architecture [44] combines vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications to estimate the traffic density, based on the number of beacons received by vehicles and fixed support nodes on the roads (Road Support Stations or RSUs) and the features of the road maps. Another recent VANET-based approach for speed estimation is proposed by Yang et al. [45].

It is also interesting to mention that it is possible to combine information provided by (fixed or mobile) traffic sensors with other data sources. For example, several approaches [46, 47, 48] combine data provided by traditional monitoring equipment with social media feeds to obtain a better perception and try to infer the causes of the traffic situation.

The estimation of traffic density and travel times has a range of potential benefits. Providing drivers with estimated travel times would help them to make more informed decision and take the best routes [49], even in multimodal settings involving both public and private transportation. Moreover, traffic density has a significant impact on the performance of dissemination protocols for VANETs, and therefore the knowledge of the traffic density could be exploited to have a more robust data dissemination or routing protocol (e.g., see [50]). Traffic awareness also implies the detection of hazards, such as traffic jams, accidents, or a malfunctioning bus line.

Besides estimating the current situation, once traffic data have been collected, it is also interesting to analyze them to learn something useful that can help to improve traffic-based applications, such as travel recommendation systems, even in the absence of real-time information. Predictive analytics, which implies extracting information from data in order to predict future outcomes and trends, is a hot topic currently. Its potential appli-

cation to the field of transportation, particularly exploiting traffic information, is very promising. For example, Chong et al. [51] present a use case that predicts traffic congestion by using a collaborative analytics system proposed by the authors. Xu et al. [52] describe a time-series approach to prediction for processing moving objects queries, and Min et al. [53] enhance it by exploiting spatial and temporal interactions and considering the road network characteristics (the possible congestion or not of a link influences the traffic flow of others).

As an example application, Waze (`http://www.waze.com`) is a popular system allowing a community of drivers to share information (e.g., traffic information). It captures contextual information (e.g., location, speed) and aggregates the data collected to deliver real-time information about the surroundings of the user (e.g., location and size of traffic jams). Another example is the exchange of multimedia data about traffic [54]; if a 5 second video or voice recording is captured by each vehicle every minute and disseminated, the data volumes generated and exchanged may be significant.

5 Trajectories

Trajectory-data management is another relevant topic in the field of urban transportation. We can identify three basic tasks regarding the collection, processing, and use of trajectories:

- *Capturing trajectory information*. The use of GPS is now widespread and many mobile devices (such as modern smartphones) are equipped with a GPS receiver. With such a positioning mechanism (or using other similar satellite-based techniques, such as the upcoming European *Galileo*), the trajectory of the device can be easily captured as a sequence of location data points. Moreover, less-precise positioning techniques (e.g., network-based positioning [55]) can be applied outdoors if a GPS receiver is not available or if the device is in a covered area with fewer than four satellites in view. Finally, in some cases it could be possible to infer trajectories, such as popular travel paths, from other sources like geo-tagged pictures [56].

- *Enriching trajectories*. Raw location data is sometimes of little use. Therefore, the GPS locations have to be cleaned and corrected. A common preprocessing step, in the case of vehicles, is applying *map-matching* [57], which is based on the assumption that vehicles move along roads and that their movement is smooth. This is not only a cleaning technique that allows the correction of GPS errors, but it also leads to data enrichment, as it allows labeling raw locations with the corresponding streets and roads. This enrichment can be further enhanced, going from *raw trajectories* to *semantic trajectories* [58], which not only encode data points but also higher-level semantic information. For example, semantic annotations could represent the type of movement (walking, running, driving, etc.), the transportation means used (private car, bus, taxi, etc.) [59], or any other feature of interest, such as the type of business visited [60]. For example, trajectory data enriched with the transportation means could be useful for other users to discover new multimodal routes using both public and private transportation. According to Ilarri et al. [61], different mobility dimensions and levels of granularity could be considered, depending on the needs.

- *Sharing trajectory data*. A number of websites have been developed to support users in the task of sharing and searching trajectory data (e.g., ShareMyRoutes and Bikely). These sites allow users to share and find information about interesting touring routes, bike routes, hiking routes, etc. Similarly, there are also many similar applications for smartphones, such as RouteShoot, which captures both a GPS trace and a video of the surroundings of the trajectory. These applications and web sites are based on a centralized architecture, where a server receives the information through the Internet (the mobile devices use cellular communications such as 3G) and can serve trajectory requests from other interested users. Nevertheless, we could also envision advanced applications where the trajectory data are also shared using peer-to-peer (P2P) ad hoc wireless communications in an opportunistic way. As an example, although it is not a pure P2P architecture because it exploits an infrastructure of access points, the *Shared-Trajectory-based Data*

Forwarding Scheme (STDFS) [62] for VANETs is based on the estimation of an encounter graph that is exploited for packet forwarding.

The previous tasks would make trajectory data available in an appropriate form. Then, it is also important to develop techniques that can efficiently and effectively exploit the trajectory information. Trajectory data can be collected and mined to learn and extract interesting information. For example, the T-Drive system [63] mines historical GPS trajectories of taxis (taxi drivers are assumed to be experts in finding the fastest routes) in order to recommend appropriate time-dependent driving directions. Ying et al. [64] advocate the use of a semantic similarity measure between trajectories as the basis of a friend recommender system. Analyzing trajectories to detect interesting areas and popular routes is another key application (e.g., see [65]). The information about popular areas and routes can be exploited in recommendation systems, traffic and service planning, and route finding. Trajectories can also be exploited for car pooling, by matching users with similar profiles based on historical information [66], or even for real-time ride-sharing [67]. Finally, as mentioned above, trajectories provide very useful information for opportunistic data forwarding in delay-tolerant networks (e.g., see [62]). Depending on the specific scenario considered and the purpose of the trajectory-data management, huge volumes of trajectory data may need to be processed; for example, in the study by Giannotti et al. [68] the authors consider a data set of around 17000 cars performing around 200000 travels over a week in the city of Milan (Italy), as well as a dataset of around 40000 cars performing around 1500000 travels over 5 weeks in the region surrounding the city of Pisa (Italy).

6 Some Challenges and Future Trends

Although the wide availability of sensors offers very interesting opportunities for urban sensing in transportation, there are also challenges that need to be tackled, such as:

- *Encouraging cooperation among the users*. A significant research topic is focused on providing incentives for users to cooperate. For example, Yang et al. [69] propose both a platform-centric model (the system itself provides a reward to the participating users) and a user-centric model (based on auctions, the users demand a specific reward in exchange of their sensing services). Another interesting related work is provided by Sugiyama et al. [70], who analyze incentive mechanisms in the context of data acquisition and distributed computing applications. Although these are significant research contributions, we believe that more effort is needed to propose specific incentive mechanisms for practical scenarios and evaluate their real impact on the behavior of the users. The use of *gamification* to engage users should also be further explored.

- *Management of the data quality*. With opportunistic sensing it is difficult to guarantee, or even estimate, a specific degree of quality for the data provided. This could be alleviated through the development of techniques that estimate the expected level of cooperation and the reliability of the different contributors. With this type of information available, it would be possible to quantify the accuracy, completeness, and freshness of the data. Guaranteeing a certain data quality could also require encouraging further cooperation among the users, although best-effort (rather than optimal) approaches should be expected. A number of data quality metrics for sensor feeds have been proposed [71, 72]. Recently, a metric called *quality of contributed service* has been proposed to characterize the quality and timeliness of a sensed quantity obtained through participatory sensing [73].

- *Spatial crowdsourcing*. This is a quite new topic that proposes the engagement of users in performing certain tasks that may require them to explicitly move to specific areas to achieve the intended goals (i.e., capturing data that are linked to that geographic area) [12, 13, 14]. We can envision innovative applications

that could be useful in the context of transportation, like asking vehicles to slightly detour from their route to alleviate congestion, measure some environmental parameters in an area (e.g., the CO_2), help in the task of traffic sensing, verify if a certain parking spot is still available and take a picture that can be provided to an interested car, or even just act as links in a multi-hop ad hoc data communication among vehicles. The idea is very new and should be explored in more depth.

- *Semantic data management*. The use of semantic techniques, such as ontologies [74] and reasoners [75], can provide significant benefits. Among them, we can highlight enabling the interoperability among different devices and enriching sensor data with higher-level information, which could be potentially queried in a flexible and semantic way. This is related to the so-called *Semantic Sensor Web* [76] and to concepts such as *semantic trajectory* [58] and *semantic location granule* [77]. However, how to go from raw sensor data to *smart sensor data* and exploit them effectively is an issue that requires further exploration. Ilarri et al. [61] analyze the state of the art on the semantic management of moving objects (e.g., vehicles) and propose a distributed framework for this purpose.

- *Exploitation of different types of sensors and large-scale data analytics*. How to effectively exploit and integrate all the information that sensors can provide is still an unsolved issue. For example, a geo-tagged picture posted on Facebook and local traffic data captured could be correlated to infer information about the location of the user and the traffic conditions, or a user could tweet complaining about the current service of a certain bus line in the city (which could be used to alert other people about a potential malfunctioning of a public transportation service). There is a need to identify what is relevant and how to benefit from it. The analysis of *big spatio-temporal data* to extract interesting information for urban transportation is also an avenue that deserves further exploration.

- *Different ways to cooperate, unobtrusive cooperation*. There are many different ways in which users can cooperate in the sensing process. Some of them will require explicit user interaction, but others could benefit from existing sensors in an unobtrusive way, without any effort from the user. For example, as commented in Section 3 there are several approaches to automatically detect park and unpark events that can signal the unavailability and availability of a parking space (e.g., [22, 23, 24, 25, 26]). As another example, vehicles could be used for environment monitoring without any direct action by the driver (e.g., [78, 79]). Finally, there are also several approaches for context detection, such as automatic transportation mode detection (e.g., see [59, 80]) and activity recognition (e.g., [81]), which could be exploited in collaborative transportation applications to minimize the involvement required from the user. We can also envision other intermediate possibilities that require intervention by the user but decreased to a certain extent thanks to the use of wearable devices such as the Google Glasses or the Samsung Galaxy Gear.

- *Management of trust and privacy*. Collaborative sensing for transportation can imply the collection and storage of data about users' daily trips. So, ensuring their privacy is a key issue. Similarly, trust management is a fundamental aspect of collaborative sensing in intelligent transportation, particularly if humans actively provide information. For example, malicious users could provide false information to disturb the system or to gain a competitive advantage. A survey of trust management in the transportation context including properties of trust, trust metrics, potential attacks and defenses, is presented by Ma et al. [82].

Acknowledgments

We acknowledge the support of the CICYT project TIN2013-46238-C4-4-R and DGA-FSE. This research was also supported in part by the US Department of Transportation National University Rail Center (NURAIL); and National Science Foundation grants IIS-1213013, CCF-1216096, DGE-0549489, IIP-1315169. Thanks also to the French International Campus on Safety and Intermodality in Transportation, the Nord-Pas-de-Calais Region,

the European Union, the Regional Delegation for Research and Technology, the Ministry of Higher Education and Research and the National Center for Scientific Research.

References

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 38:1–38:55, 2014.

[2] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava, "Participatory sensing," in *World Sensor Web Workshop (WSW)*. ACM, October 2006, pp. 1–5.

[3] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: Current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, November 2011.

[4] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell, "A survey of mobile phone sensing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 140–150, 2010.

[5] B. Fleming, "Sensors – A forecast [automotive electronics]," *IEEE Vehicular Technology Magazine*, vol. 8, no. 3, pp. 4–12, 2013.

[6] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, and R. A. Peterson, "People-centric urban sensing," in *Second Annual International Workshop on Wireless Internet (WICON)*. ACM, August 2006.

[7] S. Kabadayi, A. Pridgen, and C. Julien, "Virtual sensors: Abstracting data from physical sensors," in *International Symposium on on World of Wireless, Mobile and Multimedia Networks (WOWMOM)*. IEEE Computer Society, June 2006, pp. 587–592.

[8] M.-D. Albakour, C. Macdonald, and I. Ounis, "Identifying local events by using microblogs as social sensors," in *Tenth Conference on Open Research Areas in Information Retrieval (OAIR)*. ACM, May 2013, pp. 173–180.

[9] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, "Real-time event extraction for driving information from social sensors," in *International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE Computer Society, May 2012, pp. 221–226.

[10] FP7 Project, "Socialsensor – sensing user generated input for improved media discovery and experience," `http://www.socialsensor.eu/` [Last access: April 21, 2015].

[11] M. F. Goodchild, "Citizens as sensors: The world of volunteered geography," *GeoJournal*, vol. 69, no. 4, pp. 211–221, 2007.

[12] L. Kazemi and C. Shahabi, "GeoCrowd: Enabling query answering with spatial crowdsourcing," in *20th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2012, pp. 189–198.

[13] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," in *Proceedings of the VLDB Endowment (PVLDB)*, vol. 7, no. 10, September 2014, pp. 919–930.

[14] Z. Chen, R. Fu, Z. Zhao, Z. Liu, L. Xia, L. Chen, P. Cheng, C. C. Cao, Y. Tong, and C. J. Zhang, "gMission: A general spatial crowdsourcing platform," in *Proceedings of the VLDB Endowment (PVLDB)*, vol. 7, no. 13, September 2014, pp. 1629–1632.

[15] J. Indulska and P. Sutton, "Location management in pervasive systems," in *Australasian Information Security Workshop Conference on ACSW Frontiers (ACSW Frontiers)*, vol. 21. Australian Computer Society, Inc., January 2003, pp. 143–151.

[16] N. Islam and R. Want, "Smartphones: Past, present, and future," *IEEE Pervasive Computing*, vol. 13, no. 4, pp. 89–92, 2014.

[17] D. Shoup, "Cruising for parking," *Access*, no. 30, pp. 16–22, Spring 2007.

[18] E. Gantelet and A. Lefauconnier, "The time looking for a parking space: Strategies, associated nuisances and stakes of parking management in France," in *Europe Transport Conference (ETC)*, Association for European Transport and Contributors, 2006.

[19] M. Caliskan, D. Graupner, and M. Mauve, "Decentralized discovery of free parking places," in *Third International Workshop on Vehicular Ad Hoc Networks (VANET)*. ACM, September 2006, pp. 30–39.

[20] SFpark, "SFpark pilot project evaluation summary – a summary of the SFMTA's evaluation of the SFpark pilot project," June 2014.

[21] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, and W. Trappe, "ParkNet: Drive-by sensing of road-side parking statistics," in *Eighth International Conference on Mobile Systems, Applications, and Services (MobiSys)*. ACM, June 2010, pp. 123–136.

[22] S. Nawaz, C. Efstratiou, and C. Mascolo, "ParkSense: A smartphone based sensing system for on-street parking," in *19th Annual International Conference on Mobile Computing & Networking (MobiCom)*. ACM, June 2013, pp. 75–86.

[23] K.-C. Lan and W.-Y. Shih, "An intelligent driver location system for smart parking," *Expert Systems with Applications*, vol. 41, no. 5, pp. 2443–2456, 2014.

[24] A. Nandugudi, T. Ki, C. Nuessle, and G. Challen, "PocketParker: Pocketsourcing parking lot availability," in *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*. ACM, September 2014, pp. 963–973.

[25] B. Xu, O. Wolfson, J. Yang, L. Stenneth, P. S. Yu, and P. C. Nelson, "Real-time street parking availability estimation," in *14th International Conference on Mobile Data Management (MDM)*, vol. 1. IEEE Computer Society, June 2013, pp. 16–25.

[26] S. Ma, O. Wolfson, and B. Xu, "UPDetector: Sensing parking/unparking activities using smartphones," in *Seventh SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS)*. ACM, November 2014, pp. 1–10.

[27] B. Hoh, T. Yan, D. Ganesan, K. Tracton, T. Iwuchukwu, and J.-S. Lee, "TruCentive: A game-theoretic incentive platform for trustworthy mobile crowdsourcing parking service," in *15th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE Computer Society, September 2012, pp. 160–166.

[28] E. Kokolaki, M. Karaliopoulos, and I. Stavrakakis, "Opportunistically assisted parking service discovery: Now it helps, now it does not," *Pervasive and Mobile Computing*, vol. 8, no. 2, pp. 210–227, 2012.

[29] T. Delot, S. Ilarri, S. Lecomte, and N. Cenerario, "Sharing with caution: Managing parking spaces in vehicular networks," *Mobile Information Systems*, vol. 9, no. 1, pp. 69–98, 2013.

[30] D. Ayala, O. Wolfson, B. Xu, B. Dasgupta, and J. Lin, "Parking slot assignment games," in *19th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2011, pp. 299–308.

[31] D. Ayala, O. Wolfson, B. Xu, B. DasGupta, and J. Lin, "Pricing of parking for congestion reduction," in *20th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2012, pp. 43–51.

[32] O. Wolfson and J. Lin, "A marketplace for spatio-temporal resources and truthfulness of its users," in *Seventh SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS)*. ACM, November 2014, pp. 1–10.

[33] V. Verroios, V. Efstathiou, and A. Delis, "Reaching available public parking spaces in urban environments using ad-hoc networking," in *12th International Conference on Mobile Data Management (MDM)*. IEEE Computer Society, June 2011, pp. 141–151.

[34] A. Klappenecker, H. Lee, and J. L. Welch, "Finding available parking spaces made easy," *Ad Hoc Networks*, vol. 12, pp. 243–249, 2012.

[35] M. Caliskan, D. Graupner, and M. Mauve, "Decentralized discovery of free parking places," in *Third ACM International Workshop on Vehicular Ad Hoc Networks (VANET)*. ACM, September 2006, pp. 30–39.

[36] D. Zekri, B. Defude, and T. Delot, "Building, sharing and exploiting spatio-temporal aggregates in vehicular networks," *Mobile Information Systems*, vol. 10, no. 3, pp. 259–285, 2014.

[37] C. Morency and M. Trépanier, "Characterizing parking spaces using travel survey data," May 2008, Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), CIRRELT-2008-15.

[38] Z. Li, Y. Zhu, H. Zhu, and M. Li, "Compressive sensing approach to urban traffic sensing," in *31st International Conference on Distributed Computing Systems (ICDCS)*. IEEE Computer Society, June 2011, pp. 889–898.

[39] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson, "VTrack: Accurate, energy-aware road traffic delay estimation using mobile phones," in *Seventh ACM Conference on Embedded Networked Sensor Systems (SenSys)*. ACM, November 2009, pp. 85–98.

[40] M. Tanizaki and O. Wolfson, "Randomization in traffic information sharing systems," in *15th SIGSPATIAL International Symposium on Advances in Geographic Information Systems (GIS)*. ACM, November 2007, pp. 23:1–23:8.

[41] D. Ayala, J. Lin, O. Wolfson, N. Rishe, and M. Tanizaki, "Communication reduction for floating car data-based traffic information systems," in *Second International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOPROCESSING)*, February 2010, pp. 44–51.

[42] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, "Effective urban traffic monitoring by vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 1, pp. 273–286, January 2015.

[43] J. A. Sanguesa, M. Fogue, P. Garrido, F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, "An infrastructureless approach to estimate vehicular density in urban environments," *Sensors*, vol. 13, no. 2, pp. 2399–2418, 2013.

[44] J. Barrachina, J. A. Sanguesa, M. Fogue, P. Garrido, F. J. Martinez, J.-C. Cano, C. T. Calafate, and P. Manzoni, "V2X-d: A vehicular density estimation system that combines V2V and V2I communications," in *IFIP Wireless Days (WD)*, November 2013, pp. 1–6.

[45] J.-Y. Yang, L.-D. Chou, C.-F. Tung, S.-M. Huang, and T.-W. Wang, "Average-speed forecast and adjustment via VANETs," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 9, pp. 4318–4327, 2013.

[46] S. Djahel, R. Doolan, G.-M. Muntean, and J. Murphy, "A communications-oriented perspective on traffic management systems for smart cities: Challenges and innovative approaches," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 125–151, 2015.

[47] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi, "Crowd sensing of traffic anomalies based on human mobility and social media," in *21st SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2013, pp. 344–353.

[48] R. Varriale, S. Ma, and O. Wolfson, "VTIS: A volunteered travelers information system," in *Sixth SIGSPATIAL International Workshop on Computational Transportation Science (IWCTS)*. ACM, November 2013, pp. 13:13–13:18.

[49] H. van Lint, "Reliable travel time prediction for freeways – bridging artificial neural networks and traffic flow theory," Ph.D. dissertation, Delft University of Technology, Delft (Netherlands), May 2004.

[50] S. M. Bilal, C. J. Bernardos, and C. Guerrero, "Position-based routing in vehicular networks: A survey," *Journal of Network and Computer Applications*, vol. 36, no. 2, pp. 685–697, 2013.

[51] C. S. Chong, B. Zoebir, A. Y. S. Tan, W.-C. Tjhi, T. Zhang, K. K. Lee, R. M. Li, W. L. Tung, and F. B.-S. Lee, "Collaborative analytics for predicting expressway-traffic congestion," in *14th Annual International Conference on Electronic Commerce (ICEC)*. ACM, August 2012, pp. 35–38.

[52] B. Xu and O. Wolfson, "Time-series prediction with applications to traffic and moving objects databases," in *Third ACM International Workshop on Data Engineering for Wireless and Mobile Access (MobiDe)*. ACM, September 2003, pp. 56–60.

[53] W. Min and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[54] O. Wolfson and B. Xu, "A new paradigm for querying blobs in vehicular networks," *IEEE MultiMedia*, vol. 21, no. 1, pp. 48–58, 2014.

[55] G. Sun, J. Chen, W. Guo, and K. Liu, "Signal processing techniques in network-aided positioning: A survey of state-of-the-art positioning designs," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 12–23, 2005.

[56] X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang, "Photo2Trip: Generating travel routes from geo-tagged photos for trip planning," in *International Conference on Multimedia (MM)*. ACM, October 2010, pp. 143–152.

[57] H. Yin and O. Wolfson, "A weight-based map matching method in moving objects databases," in *16th International Conference on Scientific and Statistical Database Management (SSDBM)*. IEEE Computer Society, June 2004, pp. 437–438.

[58] C. Parent, S. Spaccapietra, C. Renso, G. Andrienko, N. Andrienko, V. Bogorny, M. L. Damiani, A. Gkoulalas-Divanis, J. Macedo, N. Pelekis, Y. Theodoridis, and Z. Yan, "Semantic trajectories modeling and analysis," *ACM Computing Surveys*, vol. 45, no. 4, pp. 42:1–42:32, 2013.

[59] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *19th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2011, pp. 54–63.

[60] J. Liu, O. Wolfson, and H. Yin, "Extracting semantic location from outdoor positioning systems," in *Seventh International Conference on Mobile Data Management (MDM)*, May 2006, pp. 73–73.

[61] S. Ilarri, D. Stojanovic, and C. Ray, "Semantic management of moving objects: A vision towards smart mobility," *Expert Systems With Applications*, vol. 42, no. 3, pp. 1418–1435, 2015.

[62] F. Xu, S. Guo, J. Jeong, Y. Gu, Q. Cao, M. Liu, and T. He, "Utilizing shared vehicle trajectories for data forwarding in vehicular networks," in *30th International Conference on Computer Communications (INFOCOM)*. IEEE Computer Society, April 2011, pp. 441–445.

[63] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-Drive: Driving directions based on taxi trajectories," in *18th SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS)*. ACM, November 2010, pp. 99–108.

[64] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng, "Mining user similarity from semantic trajectories," in *Second SIGSPATIAL International Workshop on Location Based Social Networks (LBSN)*. ACM, November 2010, pp. 19–26.

[65] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *18th International Conference on World Wide Web (WWW)*. ACM, April 2009, pp. 791–800.

[66] W. He, K. Hwang, and D. Li, "Intelligent carpool routing for urban ridesharing by mining GPS trajectories," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2286–2296, 2014.

[67] S. Ma, Y. Zheng, and O. Wolfson, "T-Share: A large-scale dynamic taxi ridesharing service," in *29th International Conference on Data Engineering (ICDE)*, April 2013, pp. 410–421.

[68] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *The VLDB Journal*, vol. 20, no. 5, pp. 695–719, October 2011.

[69] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *18th Annual International Conference on Mobile Computing and Networking (Mobicom)*. ACM, August 2012, pp. 173–184.

[70] K. Sugiyama, T. Hasegawa, J. Huang, T. Kubo, and J. Walrand, "Motivating smartphone collaboration in data acquisition and distributed computing," *IEEE Transactions on Mobile Computing*, vol. 13, no. 10, pp. 2320–2333, 2014.

[71] L. Ramaswamy, V. Lawson, and S. V. Gogineni, "Towards a quality-centric Big Data architecture for federated sensor services," in *2013 IEEE International Congress on Big Data (BigData Congress)*. IEEE Computer Society, June 2013, pp. 86–93.

[72] Z. Qin, Q. Han, S. Mehrotra, and N. Venkatasubramanian, "Quality-aware sensor data management," in *The Art of Wireless Sensor Networks*, ser. Signals and Communication Technology, H. M. Ammari, Ed. Springer, 2014, pp. 429–464.

[73] C. Tham and T. Luo, "Quality of contributed service and market equilibrium for participatory sensing," *IEEE Transactions on Mobile Computing*, vol. 14, no. 4, pp. 829–842, 2015.

[74] I. Horrocks, "Ontologies and the Semantic Web," *Communications of the ACM*, vol. 51, no. 12, pp. 58–67, 2008.

[75] R. B. Mishra and S. Kumar, "Semantic web reasoners and languages," *Artificial Intelligence Review*, vol. 35, no. 4, pp. 339–368, 2011.

[76] J.-P. Calbimonte, H. Jeung, Ó. Corcho, and K. Aberer, "Enabling query technologies for the Semantic Sensor Web," *Int. Journal on Semantic Web and Information Systems*, vol. 8, no. 1, pp. 43–63, 2012.

[77] J. Bernad, C. Bobed, E. Mena, and S. Ilarri, "A formalization for semantic location granules," *Int. Journal of Geographical Information Science*, vol. 27, no. 6, pp. 1090–1108, 2013.

[78] O. Urra, S. Ilarri, E. Mena, and T. Delot, "Using hitchhiker mobile agents for environment monitoring," in *Seventh International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS)*, ser. Advances in Intelligent and Soft Computing, vol. 55. Springer, March 2009, pp. 557–566.

[79] X. Xu, P. Zhang, and L. Zhang, "Gotcha: A mobile urban sensing system," in *12th ACM Conference on Embedded Network Sensor Systems (SenSys)*. ACM, November 2014, pp. 316–317.

[80] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks*, vol. 6, no. 2, pp. 13:1–13:27, 2010.

[81] N. Győrbíró, Á. Fábián, and G. Hományi, "An activity recognition system for mobile phones," *Mobile Networks and Applications*, vol. 14, no. 1, pp. 82–91, 2008.

[82] S. Ma, O. Wolfson, and J. Lin, "A survey on trust management for intelligent transportation system," in *Fourth SIGSPATIAL International Workshop on Computational Transportation Science (CTS)*. ACM, November 2011, pp. 18–23.

Open Civic Data: Of the People, By the People, For the People

Arnaud Sahuguet arnaud@thegovlab.org
John Krauss john@thegovlab.org
Luis Palacios luis@thegovlab.org
David Sangokoya david@thegovlab.org

The Governance Lab @ NYU
2 MetroTech Center, Brooklyn, NY 11201, USA

Abstract

"Software is eating the world", says Marc Andreessen, with data as its fuel and its by-product. Inspired by the success of various open movements, data is now getting open as well. At the forefront, governments and cities are releasing a trove of civic data with promises of better – data-driven, collaborative and participatory – forms of governance. In this paper, we provide a *definition of Open Civic Data and motivate what makes it special. We present an overview through stories from the field. We look at current technical barriers; future trends and challenges; and hint at how database research can and should contribute.*

1 Introduction

"Software is eating the world", says Marc Andreessen [1]. Data is both its fuel and its by-product. The value of data is a given. In the private sector, fortunes have been made by small and large corporations leveraging big data: Google for search and advertising; Facebook, Twitter and LinkedIn for social media; Amazon for retail. Data has become a competitive advantage and a carefully guarded asset.

The civic sphere – broadly and loosely defined for now – is eager to join the bandwagon. With more than 60% of the world population living in an urban environment by 2050, smarter cities [2] have become a necessity. Complex public problems (climate change, health, transportation, employment, education, etc.) require an effective collaboration between public and private entities. And this kind of collaboration can only happen if the data is made available in order to describe the problem and measure the impact of a solution.

Inspired by the success of various "open movements" – open source for software, open content for creative work, open access for scholarly research, open education for teaching materials – public players have embraced open data as a way to achieve this goal. The White House Executive Order [3] and G8 Open Data principles [4] are just two examples (see Table 1).

"Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike" [5]. According to McKinsey [6], this can translate into $3 to $5 trillion in economic impact. Open data also creates large opportunities for innovation [7] and social impact [8]. But so far, government – at all levels – has been the main steward for open data using its organizing power as a stick rather than a carrot.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

1	Open Data by Default	*"To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable."*
2	Quality and Quantity	
3	Usable by All	
4	Releasing Data for Improved Governance	
5	Releasing Data for Innovation	

Table 1: G8 Open Data Principles (L); White House Executive Order (R).

We define *Open Civic Data* as a subset of open data that either originates from civic sources or is applicable to civic purposes. We also take a very broad definition of *civic* to embrace all things related to the common good or about "doing together what we cannot do alone". In such an open system, "the public becomes part of the data processing system and might process data, enrich data, combine it with other sources, and might even collect their own data" [9].

What makes Open Civic Data different and interesting is the fact that this is data about us that, if handled properly, can have a huge impact on our lives and our immediate environment. Paraphrasing words from a famous address[1], we see Open Civic Data as data *of the people, by the people, for the people*.

Open data as a research topic is still in its infancy, mostly based on anecdotes [10], interviews with practitioners [7], qualitative assessments of practices, descriptive surveys of datasets [11], best practices [12] and decision frameworks [9, 13, 14]. Eckartz et al. [13] propose a decision process that involves answering the questions of ownership, privacy, economic value, data quality and technical aspect. Janssen et al. [9] present an exhaustive list of benefits (political & social, economic, operational & technical) and barriers (institutional, task complexity, use & participation, legislation, information quality, technical) for open data.

Our focus here is on the technology component of open data, which usually goes beyond the narrow technical dimensions mentioned above. Like our fellow researchers, we start from stories from the field to catalog technical barriers and look at future challenges.

The rest of this paper is organized as follows. In Section 2, we present some stories from the field to illustrate examples of Open Civic Data and its applications. In Section 3, we review existing barriers and show how database innovations can address some of them. Section 4 is more speculative and focuses on Open Civic Data 2.0 with related research problems. We finish by presenting our conclusions.

2 Stories from the field: Open Civic Data 1.0

"In God we trust; everyone else, bring data" tweeted[2] Mike Bloomberg in 2010, then mayor of New York City, in the context of his campaign against smoking. And leveraging the city open data portal, the quantitative analyst behind IQuantNY [15] investigates city issues such as noise, most blocked driveways, taxi cab tips, the worst places to swim or the most-money-making fire hydrant.

In this section, we present a few stories from the field organized around (1) quality of life, (2) government-to-government, (3) economic development and (4) crisis management.

2.1 Quality of life

"Governments should concentrate on the three B's: Buses for transit data, Bullets for crime data and Bucks for budget & expenditure data," said Mark Headd, former chief data officer for the city of Philadelphia. Bullets

[1] Abraham Lincoln, Gettysburg, Pennsylvania November 19, 1863.

[2] https://twitter.com/mikebloomberg/status/114493100541489152

and Bucks are sometimes controversial topics, but Buses – transit data – is a clear success story with hundreds of mobile apps released globally, millions of users and millions of minutes – and maybe dollars – saved every day. The Google-initiated GTFS standard, pioneered in Portland, OR and now used by thousands of cities, was instrumental in creating transit maps and mobile applications.

The EU-funded CitySDK project exposes city data like points of interest for tourists via harmonized APIs, with deployments in Helsinki, Amsterdam and Lisbon. The CityByk.es project out of Barcelona aggregates information about hundreds of bike-sharing programs across the world and powers numerous mobile applications.

Through a partnership between the City of San Francisco and Yelp!, food enthusiasts access restaurant hygiene scores from their mobile phone, with a standardization of such scores in the making.

In Brazil, citizens of Recife can find the closest health-unit based based on their location [16].

All of these examples demonstrate the great benefits of smarter access to information as a way to improve directly or indirectly people's lives. The GTFS data standard is particularly interesting. Four components contributed to its success: (a) good timing, with lack of existing standard at the time; (b) ease of publishing of transit data for cities with a simple CSV-based standard; (c) a high-value use-case for people with a problem – transit – they face usually more than once a day; and (d) a ubiquitous delivery channel, Google Maps. Now everyone is looking for "the next GTFS."

2.2 Government-to-Government

Surprisingly, the biggest beneficiaries of government opening its data are actually governments themselves.
The adage that "the left hand doesn't know what the right hand is doing" is often true because of heavily siloed data. Open data helps cities spend less, while generating new revenue and allocating their resources better. In Philadelphia, because data from the Department of Revenue and the Department of Licenses & Inspections is not shared, the city issues building permits and rental licenses to entities who are late with their taxes. This situation amounts to tens of millions in lost revenue. In Baltimore, the city issues permits for occasional events, e.g., street fairs, which sometimes require the presence of additional police forces to guarantee safety. Because the police department does not have access to the calendar of events, assignments are made at the last minute, which translates into overtime and extra cost for the city. Opening the data (a) would let the city know about the late taxes and deny the permits, encouraging tax payment and generating revenue; and (b) would let the police department anticipate the need for extra staff, with likely reduced need for overtime.
In both cases, time and money would be saved.

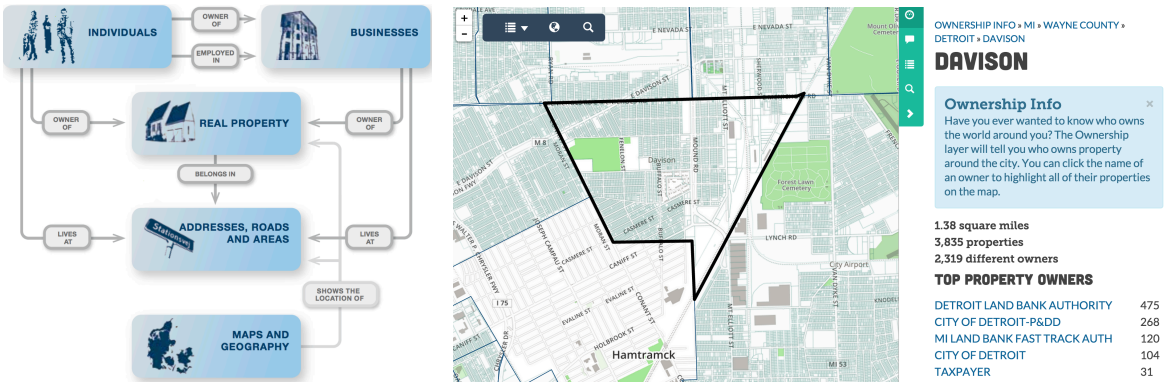
In New York City, leveraging open data, "... the Analytics Team has used data-based targeting to isolate instances of prescription drug abuse, discover cases of mortgage fraud and identify businesses operating with expired licenses [...] worth more than a combined $200 million."

In New Orleans, open data about blighted properties – due to hurricane Katrina in 2005 – lets the city and private partners prioritize which block to renovate and which to demolish.

The fear of being judged, data turf wars, and plain old politics are often barriers to open data in government. But when it opens its data, government highly benefits from it, via cost reduction, better resource allocation and also new sources of revenue. Opening data and seeing others make impactful use of it is a good ice-breaker for future conversations and collaborations.

2.3 Economic Development

Census bureaus worldwide have a long tradition of releasing datasets of key socio-demographics for their population. These datasets are heavily used by the private sector when deciding on investment, in such industries as real-estate, insurance, retail and entertainment. Denmark has identified (a) individuals, (b) businesses, (c)



(a) (b)

Figure 1: (a) What is "Basic Data"? [17] (b)"Why Don't We Own This?" in Detroit.

real estate properties, (d) addresses and (e) maps as "Basic Data" that needs to be open to foster economic development (Figure 1(a)).

The UK's Open Address project is creating a free, fresh and accurate database of all addresses in the UK, to remove the legal and cost barriers of using geographic data in the kingdom.

In Detroit, due to the economic downturn, the city lost a large chunk of its population, with many houses being abandoned. The city had to file for bankruptcy with no clear idea of what it owns. Through crowdsourced efforts, a survey of all properties was conducted and the data made open (Figure 1(b)). This helped to reboot the real-estate market.

In France, the OpenFisca project uses open data from the revenue service to provide an open source micro-simulator of the tax-benefit system. Users and corporations can use it to calculate social benefits and taxes paid. Simulations can also be run at the level of an entire population to estimate the consequences of new tax regulations.

The Danish example shows the importance of opening strategic datasets that build the foundation for more open data. This is not surprising: in the mobile space, location-based applications started to flourish when geo-data was made available via APIs such as Google Maps.

2.4 Crisis Management

Open data also plays a key role before times of crisis (preparedness, resilience planning), during and after (reconstruction).

Ushahidi pioneered the use of crowdsourcing and open data in the context of the disputed 2007 election in Kenya. The open source platform has been widely adapted and used in over 150 countries since.

In 2014 the UN Office for the Coordination of Humanitarian Affairs (OCHA) launched the Humanitarian Data Exchange portal as an open platform for UN agencies, NGOs and government departments to share data on disaster response. The platform has more than 1,300 datasets that can be compared across countries and crises and is currently being used in the Ebola crisis in West Africa.

Before Hurricane Sandy, open datasets about flooding zones and evacuation routes were used to inform residents whether they should evacuate and optimal routes towards shelters and food distribution centers.

In time of crisis, it is crucial to be able to bring "all hands on deck". Open datasets are key for collaboration between sectors and across country boundaries.

2.5 Horror Stories

Open data has also witnessed its share of horror stories.

A recent decision from the Canadian government to cancel the long form census [18] has created a data vacuum that is damaging research, city planning and also the private sector.

Outside of the civic sphere, a seminal example was the release of search logs by AOL for academic research purposes in 2006 [19]. The released data contained some personally identifiable information (PII) on AOL members that made it possible to identify people and reveal their Internet search histories [20]. On the consumer side, advances of Big Data are not sending a reassuring message with the example of Target who can figure out that a teenage girl was pregnant before her father did [21].

Because open civic data is more recent and governments opening their data are extremely risk and publicity averse, we have less examples to choose from. In New York City, taxi data released by the Taxi & Limo Commission was not anonymized properly. As a result, it was used to infer the identity of taxi drivers; in some cases, trips of celebrities were also reconstructed [22].

A few other areas prone to such stories are budget, land and health. Because of its very complicated nature, budget data can be represented in many ways and published data does not always reflect real-time adjustment and modifications. Land data is very tricky to define properly. Health data contains lots of private information.

We should expect more such stories because it is very hard to anticipate how disparate datasets will be combined. Very recently, a curious data enthusiast ran an algorithm combining taxi trip data (time, location) with prayer starting times to guess the religious orientation of some drivers [23].

Open Civic Data success stories are centered around data packaged in actionable intelligence for people, data with strong economic value – either for the government or the private sector – and cases where lives are at stake. There are few horror stories so far because data has been released haphazardly. Also, bigger stories linked to abusive state surveillance have probably eclipsed them. Another fact not to be neglected is the digital divide. By opening the data, we should not create two classes of open data users: the haves – who can get the data and understand it – and the have-nots.

For more examples of open data in action, see Open Data Census, the Open Data Barometer, government using GitHub for data or [24]. Data-Smart City Solutions at Harvard University provide a rich catalog of cases for such categories as civic data, civic engagement, health & human services, infrastructure, public safety, regulation and the responsive city.

3 Technical barriers

There are many barriers to open data. Janssen et al. [9] identify the following categories: institutional, task complexity, use & participation, legislation, information quality and technical. In this section, we look at the technical components of these barriers and relate them to the field of data management.

3.1 What makes civic data special?

Civic data relates to people's behavior and can therefore be very personal, especially with some obvious privacy issues such as personally identifiable information (PII). By its civic nature, the data is big, increasingly real-time and more and more mediated by the "Internet of Everything" [25]). It encompasses data (e.g., crime stats), text (e.g., legislation), audio (e.g., gunshot-detecting microphones) and video (e.g., traffic cameras, CCTV). It also spans across geo-spatial, temporal and social dimensions. Often times, the data starts at the analog level and needs to be digitized (with the cost and quality loss this can imply); or the data comes from legacy systems

Opening	acquisition modeling transformation
Publishing	anonymization quality provenance & versioning
Using	discovery visualization query & processing

Table 2: Barriers when opening the data.

where export is also costly. For accountability and transparency reasons, data needs to be versioned and each change documented.

Civic data is often only one part of the equation and requires integration with other non-civic data sources or data owned by a different governmental level (e.g., city vs state). This integration can prove to be difficult, depending on the nature of the data owner itself. Cities are complicated entities with hundreds of years of civic data, a plethora of agencies with sometimes misaligned incentives and a procurement system that makes it hard to bring in innovative solutions [26].

Last but not least, civic data is of interest to a very large audience with different skills and different needs: from the journalist, the investor, the regulator, the decision maker, the consumer, the citizen or the activist. As data is being requested by any of these diverse parties (e.g., through a Freedom of Information Act (FOIA) request in the US), its structure might be different.

3.2 Opening the data

Opening civic data costs money. Siloed or legacy data requires special handling. Data needs to be kept fresh. Data needs to follow legal requirements. All of this translates into human and software costs. So far, the stick has been more prevalent than the carrot in terms of opening data with more legal mandates (e.g., Obama's Executive Order [3] and FOIA requests) than scientific studies showing the positive impact of open data. If not done properly, opening data can create legal risks for cities. The most relevant datasets (the 3 B's mentioned in Section 2) are often the most controversial from a political point of view at the local level (turf-war between agencies, political strife between candidates) and beyond (cities competing to attract people, business and funding based on their core civic metrics). As a result, the opening of civic data often optimizes for ease, cost, liability and political kudos rather than positive impact.

A key challenge when opening the data is data modeling. The "runtime model" (used inside a database) might not be the same as the "publishing model" (used on an open data portal). There is no agreed-upon standard for civic data: no preferred format, no pre-defined schemas, globally unique identifiers and taxonomies for categories of civic data. Nor is there a preferred way to publish a data catalog. For instance, the city of Boston uses two different schemas for its summer and winter farmers market. New York City, Portland and Chicago each have each their own different ways of publishing their police precinct data. See Barbosa et al. [11] for an in-depth study of civic datasets.

Getting the data itself is often problematic. Civic data often originates from decades-old legacy systems. Water data from the City of Baltimore needs to be exported from an old mainframe. Data is often siloed and requires proper integration before it is published. The first step in the process often consists in agreeing on a unique

identifier to join separate datasets. Lots of open data work in New York City was done around Boro-Block-Lot (BBL[3]) identifiers.

There are some emerging efforts to address these issues, e.g., Sunlight Foundation's OCDIDs, Schema.org GovernmentService and the UK's ESD standards. Database research on information extraction, data integration, schema mapping and ETL languages is very relevant in this context.

3.3 Publishing the data

Access-control and anonymization are critical. Civic data is personal by nature because it describes human behavior. Taxi data from the New York City TLC commission was not sufficiently anonymized (see Section 2). It is also extremely difficult to anticipate how datasets – including external ones – will be combined and guaranteeing that no information will be leaked or inferred is very challenging.

"[...] a city is more than a place in space, it is a drama in time." said P. Geddes in [27]. Civic data keeps evolving. Versioning and provenance of raw datasets can be preserved relatively easily at publishing time using version control systems like GitHub. It is harder when datasets are combined across versions. Also, open data can be reused and modified freely; therefore it becomes very hard to keep track of the various transformations the data went through.

Data quality is also a critical element. In the context of crisis, bad data or stale data can lead to lost lives. The US Department of Commerce imposes the following quality requirements in terms of data: "comprehensive, consistent, confidential, trustworthy and available to all."[4].

Techniques like k-anonymity [28] and differential privacy [29] can of course be applied, but (1) one has to be aware of the issue and (2) one needs the tools to process the data. Practical and theoretical work on data provenance [30, 31, 32] and data quality [33] are also extremely relevant.

3.4 Using the data

Usability of the data is a principle in the G8 Open Data charter [4] and a requirement for innovation.

Publishing the data is just the beginning. First, it needs to be discovered. There is no agreed-upon schema to describe data catalogs (DCAT, VoID, etc.). The best way to have something discoverable on the Web so far is to have it indexed by Google.

Once discovered, data needs to be downloaded. Most data portals provide only bulk access to the datasets, often with no data compression and no way to retrieve only a subset of the data. New York City taxi datasets are notorious for taking forever to download. Datasets often come with no metadata or data dictionary. Datasets often require other datasets to be downloaded. A civic hacker often has to find and fetch the dataset, repeat the process for other datasets it depends on create a relational schema for each dataset and finally load the data into a database management system, before she is able to use the data.

As mentioned before, civic data is geo-spatial, temporal and social, which makes it hard to represent on traditional three dimensional interfaces. Canned solutions often do a poor job.

Innovative approaches like TaxiVis [34] leverage database indexing techniques to provide rich and efficient queries and data visualization over large amounts of multidimensional data. Commercial GIS solutions from ESRI or Google Earth Engine address the very common use case of spatial data.

[3]Borough-Block-Lot (BBL) or parcel numbers identify the location of buildings or properties.

[4]`http://www.commerce.gov/blog/2014/06/19/listening-our-data-customers-open-data-roundtable`

3.5 From the commercial side

Most vendors focus on one-size-fits-all solutions that do not address the specificities of open data. The IBM SmartCities initiative focuses on: "Do more with less," "Bridge silos in information and operations," "Use civic engagement to drive better results" and "Invest in infrastructure for better management". CISCO's emphasis is more on "connecting people, process, data, and things" with an Internet of everything for cities [25].

Unfortunately, the most popular open data solutions such as CKAN, Socrata or GitHub, offer very limited data management capabilities: no versioning (except for GitHub), no schema validation, and no triggers. None of them allows you do joins between datasets [35].

Civic Open Data today has a lot of challenges, whether you are a producer or a consumer. Fortunately, 20+ years of database research provide some answers to most of them, at least on paper. But these answers have to be propagated to data curators, commercial vendors and civic data scientists. A solution might be for these innovations to be unbundled, packaged for easy deployment and integration with existing solutions, including open-sourced ones. A good analogy would be to do for data what Docker is doing for software. A "Docker for Open Civic Data"[5] would package various datasets, their schema, a fitting datastore, some relevant views that represent the most common facets of the data and a ready-to-go user friendly front-end to let people start "playing with the data" right away.

4 Open Civic Data 2.0 & Challenges

A meta-trend we see at the governance level [36] is a push for more data-driven, collaborative and participatory forms of governance. In terms of open data, this means governments pushing the envelope with data, new forms of crowdsourcing for data creation and also the emergence of citizen-science targeted at governance.

In this section we look at where Open Civic Data is going and some of the domain-specific challenges that will emerge.

4.1 Richer structure

First, location and time have to be first-class data citizens, at the storage, query and visualization levels.

As civic data gets more interconnected and more social, we need richer models in terms of structure and semantics. Schema.org and ESD standards in the UK are already using Linked Data.

Civic data is social and connected by nature with annotations, comments and relationships between entities. OpenCorporates compiles a map of relationships between corporations, e.g. holdings, subsidiaries, etc. It also provides unique identifiers and taxonomies to help build knowledge on top. The goal of orgPedia project is to map corporate entities to possible labor, environment or export violations.
The natural structure for such data is a graph database, which presents both storage and query challenges [37, 38, 39].

4.2 From descriptive to predictive, with high quality and concrete metrics

The current wave of Open Civic Data focuses mostly on the descriptive aspect of running a city with reports, dashboards, etc. The next wave will be about predictive analytics.

City managers will anticipate crime before it happens or dispatch emergency response services to reduce response time. Citizens will access information about the best location for parking or the next train with room to sit, e.g. La Tranquilien in Paris.

[5] Docker4Data project

Such applications will require models to be computed via machine learning techniques and also built-in optimization algorithms. Database architectures like LogicBlox [40] combining OLTP, OLAP, machine learning and optimization around a single datastore could provide a very appealing solution.

All of this assumes a very high level of quality for the data. As we mentioned before, quality is often the reason why data is not made available in the first place. Tools to clean the data but also metrics to measure the level of quality of the data are needed.

4.3 Social civic data & civic data science

The next wave of Open Civic Data will be more crowdsourced [41] at the production, selection and consumption levels. Two concrete emerging use cases are population informatics [42] (aka social genome) and precision medicine [43].

More and more data will come from the people directly either via apps like FixMyStreet or Waze, or mediated via wearables or connected sensors [25]. Open data portals need to accommodate individual data contribution and handle issues such as data quality, duplicated data and spam for user contributed content. Research on crowdsourcing [44] and data privacy & anonymization [29, 45] is highly relevant.

People will also want to have their say on which datasets should be made open using legal tools like the Freedom Of Information ACT (FOIA) in the US. Open data portals need to accommodate voting and prioritization to help decision makers pick the next dataset to open.

Tapping into the reservoir of expertise – from citizens, private sector or simply other agencies – implies making Open Civic Data easy to access, process and understand. It also means bringing together the people with data and problems and the people with knowledge and solutions. DataKind, Bayes Impact and Kaggle are examples of such data science marketplaces.

Database management and data science are becoming the two faces of the same coin as recently discussed by Howe et al. [46]. Visualization tools like Tableau or interactive data science frameworks like Trifacta or iPython notebook are good steps towards tools that provide "an immediate connection" to what people look for in the data, as described by Victor [47]. The "Docker-for-data" model mentioned in Section 3 and the PC-AXIS file format are along those lines. The dataset-centric nature of open data makes data immutability a new "inexorable trend" as noted by Helland [48].

Because of a push for more data-driven, collaborative and participatory decision-making and governance at the city level, we should expect Open Civic Data to reflect these changes. This means richer and more social data, contributed directly by people or mediated by sensors. This also means making data more directly accessible and actionable to people in order to tap into their collective intelligence.

We are closing this section by offering a list of 10 "impossible queries" collected through various conversations with people in the field. Answering such queries will require data that does exist, data that exists but cannot be easily joined, query languages that make it easy to express such queries and more.

1. List companies in this geographic area with more than 5 labor infractions?
2. What's the relationship between company X and elected official Y?
3. What are the best default locations for ambulances during the heat season?
4. Where should I send my building inspectors first?
5. What blighted blocks should be demolish and why?
6. What's the impact of releasing (resp. removing) dataset X?

7. Which gas stations are doing price gouging during the current hurricane?
8. What's the impact of car sharing services?
9. Can people affected by flooding afford the surge of insurance premium in areas at risk?
10. Your query here

5 Conclusion

Sixty percent of the world population will live in cities by 2050. Better decision-making at the city level is critical to have a positive impact. And Open Civic Data is key.

Civic data is not just data: it is data about us and how we operate, among ourselves and within our environment. Opening the data is not just publishing the data: it means making the public part of the system and tapping into its wisdom to drive actionable decision and impactful outcomes.

There are lots of barriers, some of them technical; and database research can help. There are also some upcoming challenges. These are opportunities for great research. But whatever good research produces, we have to make sure it is properly packaged and can be consumed and incorporated into the the current work streams and work flows used by people running cities and governments. We also have to make sure we do not create an even bigger digital divide for people and cities.

In the not so distant past, database researchers were forced to pick their paper's "motivating example" based on availability of the data – rather than an application domain they care about or makes sense for their research – or to rely on made-up data. With Open Civic Data, this time is over.
Open Civic Data gives our field a chance to apply years of database research – past, present and future – to solve new challenging problems and to have a concrete impact on society and our environment.

Acknowledgments

The authors would like to thank Michael Flowers, Alberto Lerner and the Data Engineering Bulletin editors for comments on early versions of this paper.

References

[1] M. Andreessen, "Why software is eating the world'," *Wall St. J.*, vol. 20, 2011.
[2] A. M. Townsend, *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. 2013.
[3] B. Obama, "Executive Order-Making open and machine readable the new default for government information," *Whitehouse.gov*, vol. 9, 2013.
[4] U. K. P. o. G. Cabinet Office, "G8 open data charter and technical annex," in *Open Government Partnership Summit*, 18 June 2013.
[5] D. Dietrich, J. Gray, T. McNamara, A. Poikola, P. Pollock, J. Tait, and T. Zijlstra, "Open data handbook." `http://opendatahandbook.org`, 2009.
[6] J. Manyika, *Open data: Unlocking innovation and performance with liquid information*. McKinsey, 2013.
[7] A. Zuiderwijk, N. Helbig, J. R. Gil-García, and M. Janssen, "Special issue on innovation through open data: Guest editors' introduction," *J. Theor. Appl. Electron. Commer. Res.*, vol. 9, no. 2, pp. i–xiii.
[8] A. Howard, "More than economics: The social impact of open data," *Tech Republic*, 31 July 2014.
[9] M. Janssen, Y. Charalabidis, and A. Zuiderwijk, "Benefits, adoption barriers and myths of open data and open government," *Information Systems Management*, vol. 29, no. 4, pp. 258–268, 2012.
[10] J. Gurin, *Open Data Now*. McGraw Hill Education, 2014.

[11] L. Barbosa, K. Pham, C. Silva, M. R. Vieira, and J. Freire, "Structured open urban data: Understanding the landscape," *Big Data*, vol. 2, pp. 144–154, 1 Sept. 2014.

[12] J. Tauberer, "Open government data." https://opengovdata.io/, 2012.

[13] S. M. Eckartz, W. J. Hofman, and A. F. Van Veenstra, "A decision model for data sharing," in *Electronic Government*, Lecture Notes in Computer Science Volume 8653, pp. 253–264, Springer Berlin Heidelberg, 1 Sept. 2014.

[14] A. Sahuguet and D. Sangokoya, "A "calculus" for open data." https://medium.com/p/p-b-d-c-1218ee894400, Feb 2015.

[15] B. Wellington, "I quant NY." http://iquantny.tumblr.com/.

[16] K. dos Santos Brito, M. A. da Silva Costa, V. C. Garcia, and S. R. de Lemos Meira, "Brazilian government open data: Implementation, challenges, and potential opportunities," in *Proceedings of the 15th Annual International Conference on Digital Government Research*, dg.o '14, (New York, NY, USA), pp. 11–16, ACM, 2014.

[17] The Danish Government, "Good basic data for everyone - a driver for growth and efficiency." http://goo.gl/4V6ZPs, 2012.

[18] T. Grant, "Damage from cancelled census as bad as feared, researchers say." The Globe and Mail, Jan. 2015.

[19] M. Arrington, "AOL proudly releases massive amounts of private data," *TechCrunch: http://www. techcrunch. com/2006/08/06/aol-proudly-releasesmassive-amounts-of-user-search-data*, 2006.

[20] D. Kawamoto and E. Mills, "AOL apologizes for release of user search data," *CNET: http://news.cnet.com/AOL-apologizes-for-release-of-user-search-data/2100-1030_3-6102793.html*, 2006.

[21] K. Hill, "How target figured out a teen girl was pregnant before her father did," *Forbes, Inc*, 2012.

[22] C. Gayomali, "NYC taxi data blunder reveals which celebs don't Tip–And who frequents strip clubs," Oct, year = 2014, url =.

[23] A. Berlee, "Using NYC taxi data to identify muslim taxi drivers." http://theiii.org/index.php/997/using-nyc-taxi-data-to-identify-muslim-taxi-drivers/, Jan 2015.

[24] Sunlight Foundation, "The Impacts of Open Data," tech. rep., Sunlight Foundation, 2014.

[25] S. Mitchell, N. Villa, M. Stewart-Weeks, and A. Lange, "The Internet of Everything for Cities," 2013.

[26] J. Bessen, "The Anti-Innovators: How special interests undermine entrepreneurship." http://www.cfr.org/united-states/anti-innovators/p35910, 2015.

[27] P. Geddes, *Cities in Evolution*. Londres, William & Norgate LTD, 1915.

[28] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, (New York, NY, USA), pp. 49–60, ACM, 2005.

[29] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth, "Differential privacy: An economic method for choosing epsilon," in *IEEE 27th Computer Security Foundations Symposium, CSF 2014, Vienna, Austria, 19-22 July, 2014*, pp. 398–410, 2014.

[30] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance," in *Database Theory - ICDT 2001*, Lecture Notes in Computer Science Volume 1973, pp. 316–330, Springer Berlin Heidelberg, 2001.

[31] T. J. Green, G. Karvounarakis, and V. Tannen, "Provenance semirings," in *Proceedings of the Twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '07, (New York, NY, USA), pp. 31–40, ACM, 2007.

[32] R. Ikeda and J. Widom, "Panda: A system for provenance and data," *IEEE Data Eng. Bull.*, vol. 33, no. 3, pp. 42–49, 2010.

[33] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, (Vienna, Austria), pp. 315–326, VLDB Endowment, 2007.

[34] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva, "Visual Exploration Of Big Spatio-temporal Urban Data: A Study Of New York City Taxi Trips," *IEEE Transactions on Visualization and Computer Graphics archive Volume 19 Issue 12*, vol. 19, pp. 2149–2158, Dec. 2013.

[35] M. Headd, "I hate open data portals." http://civic.io/2015/04/01/i-hate-open-data-portals/, Apr 2015.

[36] B. S. S. Noveck, *Wiki Government: How Technology Can Make Government Better, Democracy Stronger, And Citizens More Powerful*. Brookings Institution Press, 2009.

[37] P. T. Wood, "Query languages for graph databases," *SIGMOD Record*, vol. 41, pp. 50–60, Apr. 2012.

[38] J. Seo, J. Park, J. Shin, and M. S. Lam, "Distributed socialite: A datalog-based language for large-scale graph analysis," *Proceedings VLDB Endowment*, vol. 6, pp. 1906–1917, Sept. 2013.

[39] E. Yoneki and A. Roy, "Scale-up graph processing: A storage-centric view," in *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, (New York, NY, USA), pp. 8:1–8:6, ACM, 2013.

[40] LogicBlox, "Datalog for enterprise applications: from industrial applications to research." `http://datalog20.org/slides/aref.pdf`, 16 Mar. 2010.

[41] D. C. Brabham, *Crowdsourcing*. MIT Press, 2013.

[42] H.-C. Kum, A. Krishnamurthy, A. Machanavajjhala, and S. C. Ahalt, "Social genome: Putting big data to work for population informatics," *Computer*, vol. 47, no. 1, pp. 56–63, 2014.

[43] E. Hafen, D. Kossmann, and A. Brand, "Health data cooperatives - citizen empowerment," *Methods Inf. Med.*, vol. 53, pp. 82–86, Feb 2014.

[44] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the World-Wide web," *Commun. ACM*, vol. 54, pp. 86–96, Apr. 2011.

[45] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *SIGMOD Record*, vol. 33, pp. 50–57, Mar. 2004.

[46] B. Howe, M. J. Franklin, J. Freire, J. Frew, T. Kraska, and R. Ramakrishnan, "Should we all be teaching intro to data science instead of intro to databases?," in *Proceedings Of The 2014 ACM SIGMOD International Conference On Management Of Data*, pp. 917–918, ACM, June 2014.

[47] B. Victor, "Inventing on Principle." http://govlabacademy.org/coaching-programs.html, 2013.

[48] P. Helland, "Immutability changes everything," in *7th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2015.

Plenario: An Open Data Discovery and Exploration Platform for Urban Science

Charlie Catlett[a,b] Tanu Malik[a,c] Brett Goldstein[a,b] Jonathan Giuffrida[b] Yetong Shao[a] Alessandro Panella[a] Derek Eder[3] Eric van Zanten[3] Robert Mitchum[1] Severin Thaler[c] Ian Foster[c]
[a] Urban Center for Computation and Data[1]
[b]Harris School of Public Policy[2]
[c]Department of Computer Science[2]
[1]Computation Institute of the University of Chicago and Argonne National Laboratory
[2]University of Chicago
[3]DataMade, LLC

Abstract

The past decade has seen the widespread release of open data concerning city services, conditions, and activities by government bodies and public institutions of all sizes. Hundreds of open data portals now host thousands of datasets of many different types. These new data sources represent enormous potential for improved understanding of urban dynamics and processes—and, ultimately, for more livable, efficient, and prosperous communities. However, those who seek to realize this potential quickly discover that discovering and applying those data relevant to any particular question can be extraordinarily difficult, due to decentralized storage, heterogeneous formats, and poor documentation. In this context, we introduce Plenario, a platform designed to automating time-consuming tasks associated with the discovery, exploration, and application of open city data—and, in so doing, reduce barriers to data use for researchers, policymakers, service providers, journalists, and members of the general public. Key innovations include a geospatial data warehouse that allows data from many sources to be registered into a common spatial and temporal frame; simple and intuitive interfaces that permit rapid discovery and exploration of data subsets pertaining to a particular area and time, regardless of type and source; easy export of such data subsets for further analysis; a user-configurable data ingest framework for automated importing and periodic updating of new datasets into the data warehouse; cloud hosting for elastic scaling and rapid creation of new Plenario instances; and an open source implementation to enable community contributions. We describe here the architecture and implementation of the Plenario platform, discuss lessons learned from its use by several communities, and outline plans for future work.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

1 Plenario Context and Objectives: Open Data Discovery and Exploration

Over the past decade cities worldwide have adopted new policies for releasing municipal and goverment data, resulting in hundreds of open data portals and tens of thousands of datasets [1]. In the spirit of transparency, public access to city information, and collaboration with the community, cities such as Chicago, San Francisco, New York City, Barcelona, and Glasgow have launched online data portals containing datasets on a multitude of topics. Many of these portals include frequently updated data on crime, city contracts, business licenses, food safety inspections, service requests, traffic, energy usage, schools, and other data of importance to residents and researchers. Thus far, this data has been used by software developers to build new applications, by journalists to research stories and watchdog government activities, by researchers from many different disciplines (including sociology, education, economics, and behavioral sciences), and by policymakers to engage the public on new strategies and initiatives.

While this first wave of open data produced undeniable benefits, several issues have thus far prevented the movement from reaching its full potential. Most importantly, "open" does not always mean "usable." Finding relevant data in the many open data portals is largely a manual exercise requiring a high degree of experience and familiarity with portal technologies and their interfaces. Furthermore, most datasets are released in file formats and structures that make integration and analysis time-consuming even for skilled data analysts, and effectively out of reach to the general public. Even the most advanced portals release most datasets as massive spreadsheets or tables, putting the burden on users to extract, visualize, map, or combine data subsets of interest. Further, many of the information technologies and tools used to make data available were designed primarily to support the analysis of individual datasets rather than exploring relationships among datasets. These technical hurdles make asking even simple questions, such as "What datasets are available for the block on which I live?" or "What is the relationship between dataset A and dataset B?" immensely challenging. While data for answering such questions may have been released, in practice that data was often simply dumped without any descriptive metadata, in a wide range of formats, etc.—and is thus only intelligible to those possessing private knowledge.

This problem of combining datasets also exists within city government and has inspired novel solutions in the recent past. One such project, WindyGrid [2], was developed for internal use by the City of Chicago in anticipation of hosting the 2012 NATO Summit. WindyGrid organizes disparate datasets (both internal to the city and from public sources such as social networks) by their space and time coordinates using geospatial database technology. It thus allows city officials to gather multi-dimensional, real-time information about different areas of the city. This system was found to support much more informed and effective deployment and coordination of services, including emergency responses. After the summit, the city continued using WindyGrid, expanding its use by adding tools to analyze and improve city services.

In the same time period, the University of Chicago's Urban Center for Computation and Data (UrbanCCD) [3] organized the Urban Sciences Research Coordination Network (US-RCN) [4] to bring together scientists, policymakers, social service providers, and others to explore the use of open data for social science research. Disciplines represented in US-RCN range from sociology to economics; questions studied range from healthcare to education, crime, and employment. Interaction within this diverse community, along with lessons learned designing and using WindyGrid, revealed that a critical need for many data-enabled inquiries is to be able to easily find and access data about a particular place and for a particular window of time.

The common workflow required for such inquiries, shown in Figure 1(a), relies on the knowledge the investigator has about what datasets are available, and from what sources, as well as familiarity with the portal technologies and their internal search, refinement, and export functions. Additionally, the examination, refinement, and aligning and merging of datasets represents a significant cost in terms of labor and time given the diversity of spatial and temporal resolution and organization of data from multiple sources. The result is that effectively using open data requires both considerable knowledge and expertise in navigating and finding data as well as resources to evaluate and prepare the data.

The Plenario project began with a hypothesis that for open data is to have truly transformative impact, it

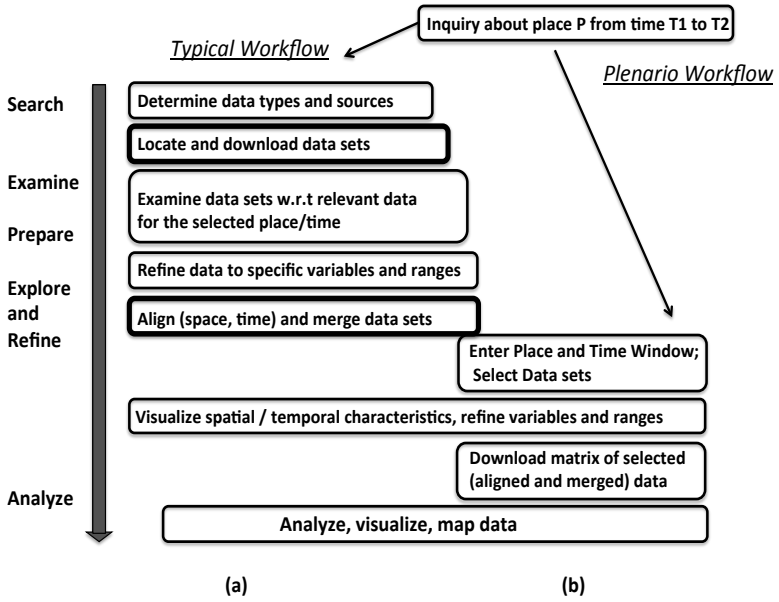


Figure 1: (a) Typical workflow for data discovery and exploration, involving many hours of data preparation and (b) the reformed, interactive space- and time-based exploration workflow using Plenario. Bold lined steps in (a) show steps that are particularly time-consuming, which Plenario eliminates by pre-integrating data from many sources.

must be accessible to non-data scientists, by individuals without a priori familiarity with the growing collection of open data portals (or their user interfaces), and it must be possible to explore the data without first investing weeks or months in data preparation. Plenario is a prototype platform developed to test this hypothesis by bringing many open datasets together, integrating them, and presenting a map-based interface for users to discover datasets relevant to a particular place over a period of time, and to examine the potential for interdependencies among those datasets.

1.1 Plenario: An Overview

Plenario exploits the fact that the vast majority of open data portals are implemented using one of two platforms, each with an API for accessing and downloading data—Socrata Open Data API (SODA) [5] and Comprehensive Knowledge Archive Network (CKAN) [6]. Each platform offers various internal search capabilities, visualization tools, and APIs for external application development. Yet at present there is no federation between these platforms or global search capabilities across portals implementing either of the platform. In part the lack of search capabilities reflects the fact that the data is highly diverse, from text to spreadsheets to shapefiles, and it is not clear how one might search such a collection of sources—keyword? Full text? Based on interactions with the US-RCN community and the experience with WindyGrid in the City of Chicago, Plenario is designed to support place and time inquiries, and consequently uses a map interface with time specifications to implement search. This user interface replaces the first two steps in typical workflows—the "Discover" phase shown in Figure 1.

Beyond the need for search, open data sources are diverse with respect to their spatial and temporal organization, resolution, units of measure, and other factors. Plenario imports datasets and integrates them into a single geospatial database, performing the alignment and merger of the datasets, eliminating the need for the user to do so, as shown in the "Explore and Refine" phase of Figure 1. Moreover, the Plenario workflow does not rely on the knowledge of the user to determine where relevant data might exist. In cases where the use may

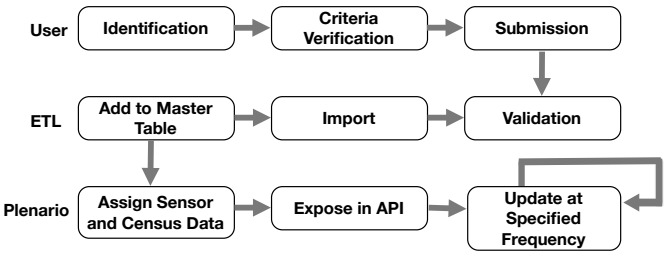


Figure 2: The path from identification of an appropriate dataset to making it available in the Plenario API. This process takes less than 24 hours for all but the largest datasets; we aim to improve this efficiency much further.

be aware of data that is not within Plenario, a request for data import is provided through a web form.

Figure 1(b) illustrates the Plenario workflow, with its new open data discovery capability and automation of several of the most costly steps in the data analysis workflow—notably the "Locate and download" and "Align and Merge" steps in Figure 1(a). Instead of searching for and combing through a multitude of potential open data sources and datasets to find data of interest for a particular location, the user specifies geographical boundaries and instantly receives all of the data available for that location (Figure 3). The labor-intensive work of combining and aligning the various spatial and temporal attributes of datasets is already done as part of the data import functions of Plenario, significantly shortening the path from question to discovery.

Plenario also provides basic data visualization, mapping, and time series creation to give users a snapshot of data before they decide to download it (Figure 4). When datasets are listed in response to a user query, each includes not only basic information and links to provenance and meta data, but a simple time series graph that provides the user with an indication as to the overall signal of the dataset, for instance whether it might provide relevant information for the particular temporal query. datasets can be selected for a map-based view showing spatial density of the data, and the user can modify the aggregation density anywhere from 100 meters to 1 kilometer. Finally, the user can refine the view of a selected dataset by specifying fields and values or ranges of interest. All of these tools enable the user to examine each dataset to determine its relevance to the research questions before exporting the integrated datasets of interest.

Furthermore, the platform helps avoid duplication of effort by providing a space for users to collaborate on cleaning data and exploring datasets. Every dataset only needs to be imported once but can be used by all (or all authorized users if the dataset is not open); similarly, the provenance of how data was cleaned is available for all users.

2 Plenario Architecture and Implementation

The Plenario software platform is open source and available on GitHub [7]. It is implemented using Amazon Web Services commercial cloud services. This facilitates replication in several important ways. First, governments and other organizations can readily create and potentially customize their own instance of Plenario, including both open and internal data. Second, organizations can provide open data for integration into an existing Plenario instance, such as the one operated by the University of Chicago at http://plenar.io. In either case, they can then use Plenario's analysis tools and API to power their own applications. The architecture easily allows data providers to choose which datasets are available to the public and which should remain "closed," available only for internal use for authorized users. The San Francisco Plenario instance, detailed in Section 4.2, is being used to explore functions for aggregating sensitive data prior to presenting to the end user.

Here we describe the Plenario architecture. We describe, in particular, the following features: (a) data import via an automated Extract-Transform-Load (ETL) builder, (b) integration using a geospatial database, and (c) special cases for common datasets such as weather and census data. In brief, an automated extract-transform-

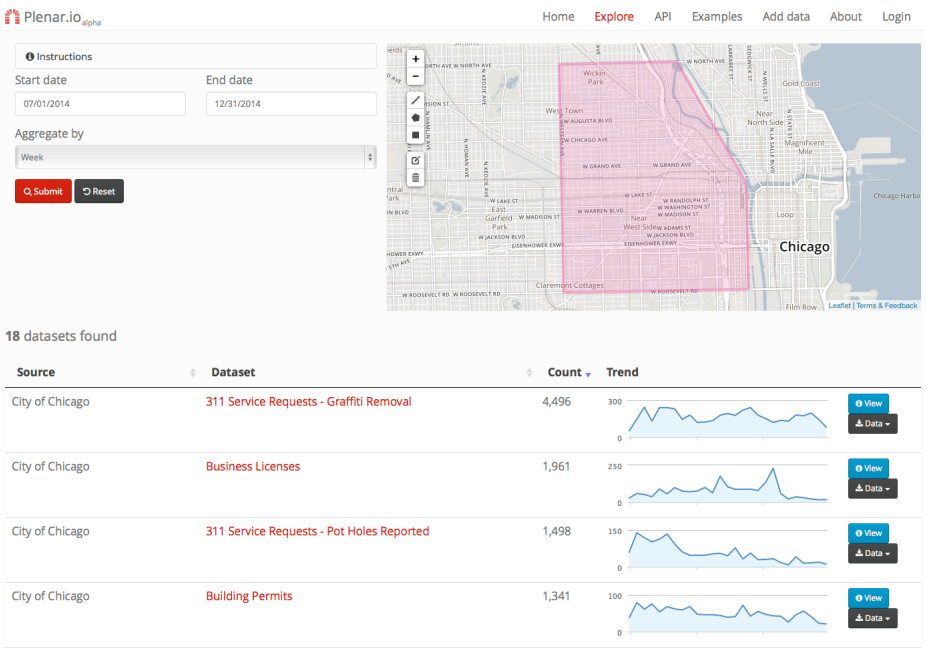


Figure 3: An example search using the Plenario portal. The search panel, at top, specifies the desired time period (the second half of 2014), aggregation (weekly), and spatial extent (the polygon). The results panel, truncated here to the first four of 18 matching data sources, includes not only basic metadata but also time series graphs as an indication of temporal dynamics.

load (ETL) builder imports each dataset and inserts the table into a PostgreSQL database. The ETL builder includes a specification for update frequency, so that Plenario updates the dataset at the same frequency as the source dataset is updated. Besides creating a table for each dataset, every record is represented as a pointer on a single "Master Table," so that all data is indexed on the same spatial and temporal database indices to improve performance (Figure 2). Additionally, the platform joins each observation to a set of commonly used datasets including sensor and place-based data. Finally, the datased is exposed in the API, making it accessible from the Plenario web portal and other clients.

2.1 Data Import: Automated ETL Builder

A dataset can be submitted to Plenario as a URL that points to a publicly available table in CSV format. This approach supports, for example, datasets on a Socrata or CKAN platform, as well as direct links to CSV files. Plenario's automated ETL builder scans the dataset, gets meta-information if available from the source platform, infers field types, and checks for common issues of data integrity. As Plenario currently focuses on spatio-temporal datasets, the user is then asked to identify which fields in the source data contain the required spatial information (location), temporal information (timestamp), and unique identifier, and how frequently the dataset is updated. (Future improvements to Plenario will remove one, two, or all three of these requirements: see discussion in Section 5.) The user can also add information regarding the dataset's provenance, description, and license if these are not automatically populated (as they are with Socrata datasets).

Following a basic check for URL stability and malware, an ETL worker process begins importing a local copy of the dataset as a new table in the PostgreSQL database. After import, Plenario inserts a row into the Master Table for every row in the new dataset, containing the dataset name and dataset-specific identifier (foreign key), the row identifier (primary key), and the spatial and temporal fields. The dataset is then made available via

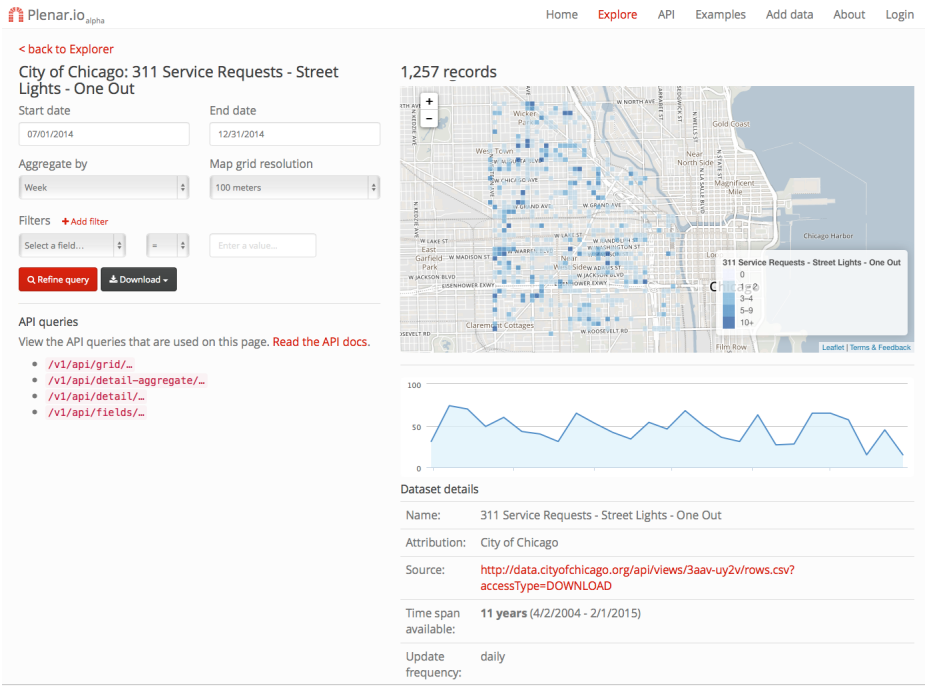


Figure 4: Plenario dataset view. Selected from the first results screen, this view allows the user to view the spatial distribution of a given dataset and provides links to the data and associated metadata. This screen also allows the user to change the temporal and spatial resolution of the query and to refine the dataset by selecting and specifying values or ranges for individual record values.

a RESTful API with endpoints for raw data, data aggregation by time and space, metadata, and weather-specific data (weather is one of several special case base datasets discussed in Section 2.3). Tasks are automatically scheduled to update the dataset according to the refresh frequency of the source dataset, using the unique identifier to avoid re-populating the entire table. Datasets can be imported and updated simultaneously using multiple ETL workers.

2.2 Core Database: Single Spatio-Temporal Index and PostgreSQL Schema

Plenario achieves the workflow optimizations discussed in Section 1 and illustrated in Figure 1 by organizing all records using common spatial and temporal indices in the Master Table (Figure 5). This method has several important implications.

First, data is automatically organized in an intuitive and coherent manner that can be easily searched and accessed by the user. In addition to API access, Plenario includes a portal interface (Figure 3) that allows users to search for datasets by drawing polygons or paths on a map and selecting start and end dates.

Second, data is organized, and can be searched for and accessed, without relying upon user knowledge of the existence of the data or its sources. Any point or polygon, and any time period, can be associated with data from multiple datasets, from multiple government agencies or organizations. This data can then be returned as a result of a search without the user needing to specify the data source. Thus, for example, a query for data points from Midtown Manhattan during June 2013 will return data from the City of New York, New York State, federal government, and numerous local or national organizations and surveys, including sources of which the user is unaware.

The third implication is that data for any arbitrary geography can be readily organized as a time series

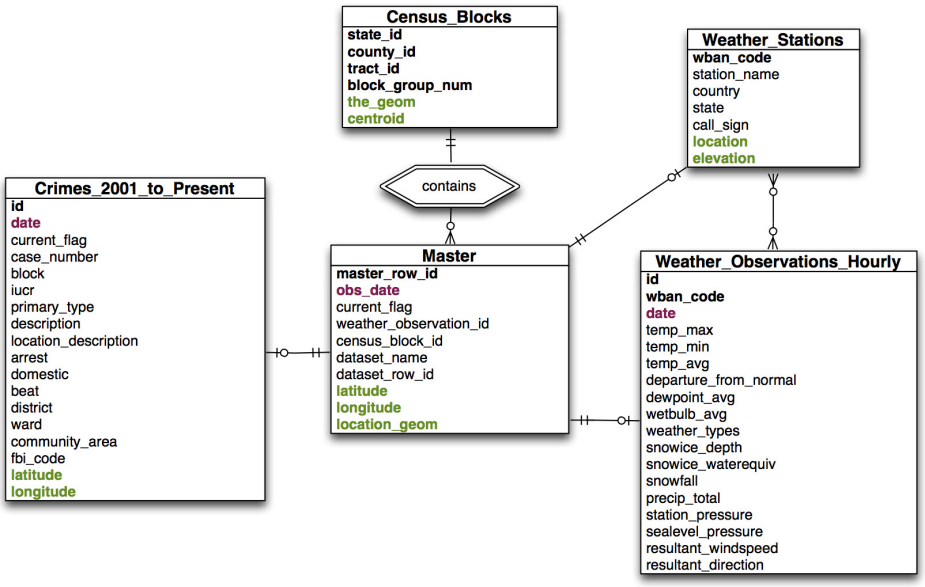


Figure 5: PostgreSQL Schema, showing how a sample dataset (Chicago crimes) feeds into the Master Table, which in turn links to spatial data like Census blocks and sensor data like weather observations through the `census_block_id` and `weather_observation_id` fields. There is one row in the Master Table for every row in a source dataset like Crimes, but a many-to-many relationship exists between the Master Table and Census blocks or weather observations. Note that the `Weather_Observations_Hourly` table (which contains no spatial information) is filtered through the `Weather_Stations` table (which contains no temporal information).

containing counts (or other variables) of observations in each contained dataset. Plenario enables one-click download of such time series matrices—instant snapshots of a geography over time—with any temporal resolution from hours to decades. Plenario thus eliminates the tedious work of data compilation and aggregation along identical temporal and spatial units and allows users to begin simple analyses immediately.

2.3 Special Cases: Commonly Used datasets

Plenario was optimized to support not only particular datasets related to a specific topic but to enable investigations in the context of widely used data sources, such as weather observations (with corresponding station position obtained from shapefiles) and aggregated census information. Once a dataset is imported and inserted into the Master Table, Plenario enriches it with other data relevant to the same geography, including sensor and location-specific data. Below we discuss *sensor data* (time series such as weather) and *local data* (relatively static data about the geography), which are also shown in Figure 5.

Sensor data, which records one or more variables at regular intervals in fixed locations, usually along a network with high coverage (such as weather stations), is important both for tracking environmental variables over time and for enhancing human-generated data (such as noise complaints) with objective recordings from the same time and location (such as noise levels).

Weather data in particular is important to many questions about place, from healthcare studies to traffic or flooding analysis. We thus include NOAA hourly and daily weather weather station data as an integral element of the Master Table scheme. To date, we have loaded into Plenario all U.S. hourly and daily weather station data since 2011. We also assign to every record in the Master Table the identity of the weather station that is closest to it; thus, we can respond to any query requesting weather information about a record by retrieving

the observation from that "nearest" weather station for the time closest to the record's timestamp. The process is efficient because all weather data is stored in only one place, and because the closest weather stations are pre-calculated when a dataset is imported.

This integration of weather station data provides an initial example of how sensor data adds to the open data landscape. Further plans for sensor data include incorporating data from the Array of Things [8] project in Chicago, which will report measures such as air pollution and noise at a much higher resolution, both temporally (every 30–60 seconds) and spatially (sensors will be deployed throughout Chicago, with densities ranging from one per square block to one per square kilometer). This data source will further illustrate the value of sensor data to municipal open datasets, enabling investigations such as the spatial and temporal characteristics of air quality in the context of vehicle flow and weather, or the interrelationships between hyperlocal weather and crime.

Local data refers to data aggregated at a regional (not individual) level, containing variables that are relatively static over time, such as demographic data and local economic data. The Chicago instance of Plenario incorporates a prototypical example of local data, which is data from the United States Census: every row in the Master Table is coded with its FIPS Census block identifier, which allows for easy enhancement with open data from other sources tied to that Census block, Census tract, county, state, etc., all of which can be determined from the Census block identifier.

2.4 Components and AWS Implementation

Plenario is built entirely from open source tools and is released under the MIT license. All code is in a GitHub repository [7], making the platform easy to fork and clone. By default, all source datasets remain under their original license; most are released under the MIT license or are unlicensed.

The data backend of Plenario is built as a PostgreSQL relational database, with PostGIS geospatial extension. SQL Alchemy [9] is used as the object relational mapper with the GeoAlchemy 2 extension. The web application with API was developed using Flask [10], with mapping capabilities provided by Leaflet [11] and Open Street Map [12]. The ETL process uses Celery [13] for logging, and Redis [14] is available for caching support when high loads are anticipated.

We host Plenario on Amazon Web Services (AWS)'s Elastic Cloud Compute (EC2) infrastructure. We currently use four virtual servers within a Virtual Private Cloud (VPC): one web server, one database server, one ETL worker server, and one gateway server. Due to its elastic nature, the EC2 server resources can be upgraded in real time as traffic load and data footprint increase. For larger deployments, the database server can be sharded and replicated for more capacity using the AWS Relational Database Service (RDS). Amazon Machine Images (AMI) can be used to snapshot the various server configurations for easy re-deployability.

For data integrity and redundancy, every raw dataset processed by the Plenario ETL worker is saved as a snapshot and stored on Amazon's Simple Storage Service (S3). This approach allows for data integrity checks, ETL fault tolerance, and history tracking on every dataset Plenario ingests.

Plenario can be used in several ways: a user can fork the GitHub code to develop a separate project, copy the entire project via a machine image on AWS, or feed data into the web portal supported by the University of Chicago at http://plenar.io. Each of these modalities of use has been seen since Plenario's alpha launch in September 2014, including a repurposing of the core API to power the City of San Francisco's Sustainable Systems Framework initiative, as detailed below.

The web portal interface described above is in fact an application that accesses a Plenario instance running on AWS, via the Plenario API. This modular approach enables other front-end frameworks to be built to use the API, ranging from custom mobile and web applications (of which http://plenar.io is an example) to a complex analytics system such as WindyGrid, which uses commercial mapping and user interface software such as ESRI.

3 Evaluating Plenario's Architecture

The key architectural feature of the Plenario system is its spatio-temporal database, which is hosted on the cloud. Users can upload new datasets into the hosted system, and query datasets for discovery and exploration using a RESTful API. The system, given its need to support open-data for all, sets no limits, both in terms of the number of uploads and the size of the upload. The open upload feature raises scalability concerns, especially when performing exploratory spatial querying, which depending upon spatial-extents itself can be I/O-intensive. A complementary, but related concern to scalability is the cost of hosting an open-data service such as Plenario on the cloud. Given the volume of anticipated data, hosting Plenario on high-end instances seems natural but if most uploads are small and queries retrieve small spatial regions then high-end instances do not provide sufficient cost-benefit advantage.

To evaluate our choices, we performed a thorough performance evaluation to evaluate our database choice, and to determine which type of cloud instance provides the best cost-benefit ratio. To conduct the experiments, we developed a benchmark based on available Plenario queries and logs. We evaluated the benchmark workload with open-source relational, NoSQL and array database systems to determined which database system exhibits highest performance for concurrent uploads and query. Finally, we instantiated Plenario's relational database on different cloud-instances to determine the best cost-benefit ratio in terms of transactions per second per dollar spent.

3.1 The Plenario Benchmark

We developed a benchmark specific to Plenario because unlike other geospatial benchmarks [20], Plenario has no separate database instantiation and query generation phase; database tables and queries are determined by an incoming user-workload leading to an openly-writable spatial storage system. To simulate the discovery and exploration phases in Plenario, we simulated a closed loop Markov-chain synthetic workload generator in Python. The Markov-chain consists of two states, `add_data` and `query`, with the query state having five sub-states, `initial_query`, `expand_time`, `narrow_time`, `expand_geography`, and `narrow_geography`. The current state is updated after each query and the system chooses between `query` with probability 0.85 and `add_data` with probability 0.15. Within the query phase, expansion and reduction of spatial attributes occurs with equal probability, and spatial attributes are chosen over time attributes by 0.6 to 0.4. We chose these probabilities based on user query logs and estimated patterns.

As part of a session, users often change spatial and time attributes by either expanding them or narrowing them from an initial specification. Expansion and reduction of the time attribute is simply achieved by changing both the starting and ending dates by an equal amount such that the new date range is a factor k times of that of the old date range. We choose k so as to reflect a daily, weekly, or monthly search pattern. To expand and narrow spatial boundaries such that the new query is again a convex polygon, we follow a simple single parametric method. (A convex polygon is not a requirement of the Plenario API. However logs show that all users to date have indeed specified convex polygons.) For expansion, given an N-sided polygon with coordinates $(x_1, y_1), \ldots, (x_n, y_n)$,

1. Find the smallest rectangle that exactly covers all the vertices of the polygon by finding the maximum and minimum of the coordinates of all the vertices;
2. Fix the center of the rectangle and expand it outwards by an expansion factor of k;
3. Divide the rectangle into four equal sub-rectangles: top-left, top-right, bottom-left and bottom-right;
4. For each vertex of the polygon, identify the sub-rectangle it is located in, and create a new point in the region of the box further from the center than the vertex is;
5. The newly simulated points form the vertices of our new polygon.

The same algorithm works for narrowing the polygon except instead of expanding by a factor k, we narrow by k. The algorithm is guaranteed to produce a convex polygon since all the new points are simulated randomly in four different sub-rectangles, so it's impossible that they lie on a line. Finally, in the `add_data` state, data is generated for a given spatial with a skew ranging from 0.1–0.3 and with the size of the data chosen from a Zipf distribution varying from a few kilobytes to a few gigabytes.

3.2 Evaluation

To evaluate the query and upload workload from the Markov-based generator, we chose three databases: (a) PostGIS, (b) Accumulo, and (c) SciDB. Plenario is currently hosted on PostGIS, which is a traditional RDBMS with spatial index support. RDBMs can encounter scaling problems when dealing with Terabytes of data or thousands of tables. We examine Accumulo and SciDB as alternative database systems that support scaling out onto multiple nodes. Accumulo is a key-value store, and SciDB a multi-dimensional array store.

To correctly ingest two-dimensional geospatial data into the one-dimensional Accumulo key-value store, we create a geohash [18] of each latitude-longitude pair, hence mapping each two-dimensional coordinate into a scalar value. Geohashes have the property that points near to each in space other have geohashes with the same prefix, which improves locality significantly when points are stored by order of geohash. Besides, geohash can obtain good precision in a small scale of bit length. While there are other schemes of linearizing spatial data that are more precise than geohash in terms of maintaining the locality of spatial points, geohash is simple to implement and is used significantly in key-value systems for spatial indexing. In SciDB we consider latitude and longitude as two different dimensions of a grid that is divided into cells. Since SciDB supports only integer dimensions and Plenario's data has spatial coordinates of arbitrary precision, to consider latitude and longitude as dimensions, we perform linear scaling of the spatial coordinates, and round off the resulting values. Using these dimensions, the size of the grid is decided based on the cross-product of the number of latitude and longitude points, respectively, in the data. As not every combination of a latitude and longitude cell may have a corresponding data point in the dataset, the cross-product results in a two-dimensional sparse array in SciDB.

The databases are set up on an AWS T2.Extra Large instance on Ubuntu 14.04. We used Accumulo version 1.7.0 with a single master and tablet server. The accompanying Hadoop version is 2.6. configured for one name node, data node, and secondary node respectively. Hadoop uses Zookeeper v3.4.6 with one server. We used version 14.12 of SciDB with four instances one SciDB master and one worker. With this setup, we use our experiments to answer four important questions regarding Plenario's architecture: (1) Which data model (grid, geohashed key-value, or R-tree-based relational) provides the best performance for retrieving the relevant datasets from the Master table, given a spatial polygon? (2) Given that Plenario accepts CSV data, how do different database systems compare in terms of ingest costs? (3) When different users upload and query datasets concurrently, how do systems compare in terms of the transaction rates that they supported? (4) What is the economic cost of hosting Plenario on different EC2 instances?

Figure 6 compares the query performance on the three different databases. The query workload consists of 4–6 sided spatial polygons. An initial specification of points is chosen randomly. The query may expand or narrow as described in Section 3.1. We measure the average response time of the workload against the same data loaded in each of the database system. In PostGIS, each spatial query uses the spatial index to retrieve the objects; in Accumulo, given a spatial polygon, we calculate the minimum and maximum geohashs across all vertices. We use these geohash values as the range to query to perform a batch scan and determine if the scanned point is in the polygon to get an exact answer. In SciDB there is no index, so a two-dimensional polygon query is approximated to the nearest grid rectangle to obtain all the points from that grid and then each point is checked for containment. As we see, we get best performance from PostGIS for querying but Accumulo is not far off. Since we are only dealing with point data with arbitrary precision, so far geohashing is comparable to R-tree.

We compare the three databases for ingesting data (Table 1). Note that Plenario intends to expand its Chicago instance to thousands of datasets so fast ingest is important. While we have already described a formal ETL

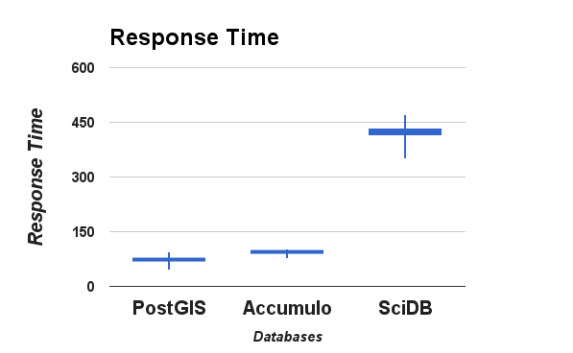


Figure 6: Response Time Comparison

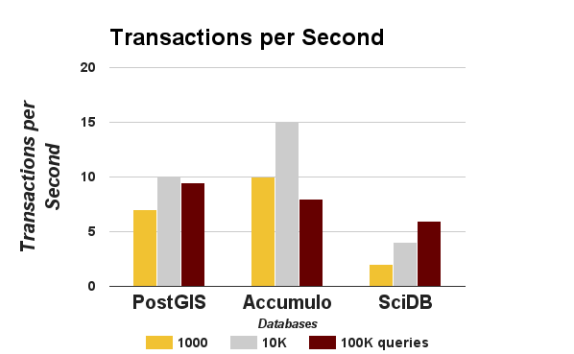


Figure 7: Throughput Comparison

Table 1: Ingest Time (in seconds)

Data Size	PostGIS	Accumulo	SciDB
1	120	65	300
10	800	630	NA
20	1800	1260	NA

Table 2: Cost of hosting Plenario

Machine Size	Transactions per $
Small	9.8
Large	13.6
XL-HM	12.8
2XL-HM	12.3

process for ingesting public datasets, in this experiment we compare in how must time cleaned CSV datasets are ingested. In Accumulo ingest implies using geohashing to first convert data and then using batch mode to ingest the newly formatted data into an Accumulo table. Accumulo loads data at a constant rate into this table, which is approximately at 3700 entries/second or 200 KB/second. In SciDB, the CSV file is first ingested as a single dimensional array and then redimensioned into a two-dimensional grid. In SciDB, while ingesting into single dimensional array is straightforward, re-dimensioning is a memory intensive operation which often fails or is extremely slow due to thrashing for fairly modest size of data such as 8GB. In PostGIS, data is simply ingested into the relational table using bulk copy.

Figure 7 compares the concurrent cost of the querying and update in terms of transactions per second. Given the fast uploads Accumulo is able to achieve higher number of transactions per second, even though it is not running the queries against a spatial index. This is indeed surprising and exciting result, which demonstrates the benefits of using the Accumulo instance especially as thousands of datasets are added to Plenario and the system needs to scale out. In RDBMSs, R-trees (and other spatial indices) are external to the data and must be updated and computed when new features are added. This is computationally intensive and is affecting performance adversely as data size increases.

Finally, in Table 2 we compare the cost of hosting Plenario on different size instances AWS. Given that Chicago Plenario instance is 100GB in size, and most queries are fetching small datasets of a few kibobytes to maximum of 1GB, currently even a large instance is sufficient to host Plenario in terms of the cost benefit ratio. While currently sustainable, this result must be examined in lieu of projected growth and other similar studies [19]

4 Plenario Use Cases and Early Lessons Learned

We have discussed a number of advantages the Plenario project was designed to provide for researchers, government employees, developers, journalists, and citizen users. Here we discuss several specific use cases where Plenario is being used today to support social and economic science research (The Chicago Plenario instance)

and for community engagement on urban sustainability, resilience, and livability goals (The San Francisco Plenario instance).

4.1 Supporting Research: The Chicago Instance

The Chicago instance of Plenario, active at http://plenar.io, has been implemented with data from multiple cities, and with a focus in particular on Chicago, to support research into urban science and computational approaches to public policy, identified through the US-RCN, by supporting rapid data discovery, helping researchers, journalists, and residents identify datasets that interest them regardless of the original source. For example, Goldstein leads a team investigating the interrelationship between local weather and violent crime, which involves all of the base sensor and local data described in Section 2.3 as well as urban crime data, 311 service call data, and other datasets. Without Plenario, this research would begin with an attempt to find data of interest from many independent sources, many of which the user might not know exist. For Chicago these include open data portals operated by the City of Chicago, Cook County, the State of Illinois, the federal government, and data from hundreds of separate government departments such as the Illinois Department of Transportation or the Chicago Department of Public Health. In addition to these sources, relevant data has been imported into http://plenar.io from the National Oceanic and Atmospheric Administration (NOAA) and the U.S. Census Bureau.

As the Chicago instance has grown from dozens of data sets to over 150, we have found that Plenario's automatic data set summary feature has led users to identify a range of previously undetected data quality problems. For example, a single error in the date field is immediately apparent when a summary suggests that Plenario contains data from prehistoric times, or far into the future. We intend to incorporate consistency checks to flag obvious errors of this kind (such as impossible dates), but we note that not all errors are readily flagged by algorithms. Errors and holes in data are inevitable. Thus it will be important both to work with data providers to fix obvious errors and to provide Plenario users with mechanisms to discover and flag errors.

4.2 Enabling Community-Driven Urban Sustainability, Resilience, and Livability: The San Francisco Instance

The Plenario platform has also been deployed experimentally as part of the City of San Francisco's Sustainable Development initiative [15]. This project has motivated important Plenario enhancements, including support for additional data types such as geographical information in the form of ESRI shapefiles. It has also spurred the development of new features to enable the use of Plenario as a community "dashboard," whereby the visual interface is the primary use. (In contrast, the Chicago Instance is mostly used to refine and export data for advanced data analytics.) Several enhancements driven by the San Francisco implementation have already been incorporated into the core Plenario code base. Others will be incorporated after further evaluation in the San Francisco instance.

The San Francisco Plenario instance contains datasets pertaining to a wide variety of sustainability indices, ranging from community structure accessibility to green space, canopy cover, water consumption, and energy use. Having the ability to instantly access datasets of this kind by spatial-temporal queries empowers institutions and communities to assess the status quo and plan future efforts in sustainable development. Thus, for example, the framework is to be used in the sustainable development of the South of Market (SoMa) ecodistrict.

The data needed for the applications that the San Francisco Instance is designed to support are highly heterogeneous in both content and form. For example, quantifying access to green spaces—the vicinity of parkland to residents—requires analysis of geographic information regarding the location and shape of each park, which cannot be treated simply as a point in space. Similarly, a community center is an entity that exists over a certain time span, in contrast to much place-based urban data such as crime or inspections, which are "events" that each occur at a specific instant. To incorporate these and other types of data, Plenario's database schema was extended and more ETL functions were added. Moreover, new types of queries were developed and implemented

efficiently, aimed at questions of the type "What is the average distance for residents of a given area to the closest farmer's market, at any point in time and in a given range?"

In the San Francisco Plenario instance approaches to support a mix of open and sensitive data are being explored. As with the Census data in the Chicago instance, some of the data made available from City of San Francisco is not public and thus must be carefully aggregated to protect privacy. Utilities data is one such kind of available data in which privacy must be protected. One algorithm to protect privacy that is commonly used for utilities datasets is the "15/15 rule," which requires that no aggregation sample may contain less than 15 data points, and any point in any aggregation sample cannot represent more than 15% of the measure for that sample (The "15/15 Rule" was adopted by the California Public Utilities Commission in Decision D.97-10-031.). The methodology being explored in the San Francisco project is to host raw data securely in the Plenario instance, and then to implement the query- and data-specific privacy-preserving aggregations as a function of the particular search, view, and/or data export process.

5 Lessons, Challenges, and Opportunities

With the two large-scale Plenario instances described above, we have identified a number of challenge areas that will be essential to address in order to move Plenario from an alpha platform to a fully supported and sustainable resource. We group these as issues related to data, scaling, and architecture.

5.1 Data Issues

Data is often collected for different purposes and thus in different ways across juridisctions. Even datasets with similar purposes, such as 311 service requests or food safety inspection reports, can rarely be merged across jurisdictions, effectively limiting research to a focus on one particular city rather than incorporating and studying multiple cities at once. These barriers can exist at the metadata level (different variables recorded), in the resolution of the data (spatial and temporal), and even at the level of individual data points and fields (semantics and ontology). For example, a crime classified as "assault" in New York City crime data would be classified as a "battery" in crime data from Chicago, which may mislead a researcher attempting to compare violent crime in the two cities or compile a large dataset of crime in the United States.

We have also encountered the common challenge of poor data quality and documentation. Because all data in Plenario ultimately refers to a source dataset hosted by a municipality, the remedy is limited to either cleaning the data upon insertion into Plenario or providing feedback to the data providers. Data cleaning at insertion would accelerate cleaning in comparison to relying on data providers, but would also require that the platform understand in each case what is "correct." Ultimately this balance might be encoded into the ETL process in a similar fashion to the update frequency. Finally, the lack of unique identifiers on many datasets also means that updating datasets requires a full refresh of the entire dataset, which increases load but more importantly introduces data consistency issues that will impact the applications using the datasets, particularly those aimed at real-time capabilities.

5.2 Scaling Issues

Plenario was designed with consideration regarding scale, given the enormity of the open data landscape and the rapid pace with which open datasets are being released. Nevertheless, as the experiments show the Master Table approach introduces scaling challenges, particularly as the table grows to billions of rows. The team has explored a variety of approaches including partitioning the table along the temporal index, with mixed results. In particular, the number of NOAA's hourly observations for all 2,200+ weather stations since 1997 in the United States was deemed too large to import in its entirety, while maintaining a reliably responsive API. To

work around this limitation, only observations from weather stations within a certain radius of each dataset's bounding box were added.

The sensor data also contributes to scaling challenges. Though the closest weather station to every record is identified upon insertion into the Master Table, the platform executes the join between the Master Table and the weather table at the time of request rather than as part of the insertion process. This has significant impact on query performance but the alternative would exacerbate scaling issues with the Master Table by making it extremely wide. Furthermore, sensor data needs to be spatially smoothed to avoid sharp boundaries in the data such as when two neighboring weather stations record significantly different values for a given variable. To reduce computational load, sensor data is organized spatially using a Voronoi diagram [16] without spatial smoothing.

5.3 Architecture and Data Semantics Issues

Plenario's original purpose as a platform for spatio-temporal data discovery and exploration brings into question what variables count as "space" and "time." For example, should 311 data reflect the location of the caller or the location of the problem reported? How should the location of non-spatial crimes, like fraud or online crimes, be reported? And how should Plenario represent records missing a spatial or temporal value? How, too, could unstructured data be supported in Plenario—especially when the location and datestamp of such data are uncertain?

We have also encountered challenges with respect to how to treat data that lacks resolution in spatial and temporal data. For instance, how do we present city budget data that covers an entire city for the period of one year—and make this data discoverable in typical user searches? Should a query across multiple years return multiple city budgets, ones wholly contained in the temporal arguments, or none at all? How should shapes like parks, streets, and parcel lots be dated? Some of these challenges are being highlighted in the San Francisco Plenario instance, as discussed earlier.

Ultimately these challenges suggest exploration into the optimal approach to support the integration of spatial/temporal data with data that is primarily "entity" based. In some cases, such as with census data, spatial and temporal mapping can be done in concert with data aggregation as is necessary for privacy protection. In other cases, particularly with organizations whose data includes internal private data about businesses and individuals, such mapping is less straightforward. Plenario currently supports questions such as "where were the automobile accidents in mid-town Manhattan during heavy rainstorms in 2014?" but is not organized in order to refine this query to show only those accidents involving cars greater than 10 years old, or male drivers aged 18-24.

Finally, Plenario is currently designed as a portal for open data, which is only a subset of data useful for urban science and research, for policy development, or for many areas of improved urban operations. There are known solutions to challenges to multiple levels of authorization, and it will be important to integrate these into the platform. The San Francisco Plenario instance supports sensitive data by aggregating at the time of query, presenting the aggregated data to the end user. The Chicago Plenario instance uses pre-aggregated census data, eliminating the need to aggregate at query time. While this improves query performance and reduces the sensitivity of the data stored in Plenario, it also requires that the aggregation algorithm is defined a priori, where different aggregation schemes may be more or less optimal for different types of inquiry.

6 Conclusions and a Plenario Roadmap

The Plenario team has begun to develop a 12–18 month roadmap based on input from early users. A rigorous set of performance scaling tests is being developed to explore the architecture issues noted above; the results of these tests may lead us to revisit various design decisions, ranging from the underlying database to the Master Table. Several features requested by researchers are under consideration for this roadmap, including automated

time series analysis to identify correlations between datasets: for instance, identifying subsets of 311 data that are lagged by violent crime in various neighborhoods of a city.

Of particular interest to many place-based investigations is the identification of urban areas that function as units. Traditional boundaries such as neighborhoods or districts often do not reflect the underlying social or economic structure, in part because many such boundaries were drawn generations in the past and/or through political processes. The rapidly expanding variety of data being integrated into Plenario is resulting in increased opportunity to understand what differentiates one neighborhood from another and to use spatial units defined by current data, not solely by a 20^{th} century surveyor's pen. Concurrently, support for place-based research will require more powerful tools for specifying spatial aggregation of data (where Plenario has already provided flexibility in temporal aggregation), necessary to address the Modifiable Area Unit Problem [17]: that is, the fact that the results of spatial analysis are often highly dependent on the spatial units used.

Today's open data landscape largely resembles the Internet of the 1980s when data was shared through anonymous file transfer servers, which were useful only to those with inside knowledge of their locations and contents. The advent of HTTP and web browsers led to today's powerful search and integration capabilities, including those that Plenario uses to import data. An underlying objective of the Plenario project is to contribute to these benefits extending to open data.

The first step toward this vision has been to implement the Plenario platform as a means to reduce or eliminate many of the challenges of working with open data, beginning with discovery, exploration, and integration across many data sources. Addressing these challenges provides increased incentives for governments to release data, reducing the need to develop their own custom data portals and providing the basic tools to start extracting insight and return on investment from their data. By building and encouraging a collaborative open data ecosystem at every stage, from identifying datasets to building third-party tools, Plenario helps push the full potential of this movement closer to realization.

Acknowledgments

We thank Mengyu Zhang for help with experiments, and anonymous reviewers for comments on the paper.
The Plenario project is funded by the John D. and Catherine T. MacArthur Foundation and the National Science Foundation via an NSF Early-Concept Grant for Exploratory Research (EAGER) for software development (award number 1348865), while the interaction capabilities were driven by the Urban Sciences Research Coordination Network, created with an NSF Building Community and Capacity for Data-Intensive Research in the Social, Behavioral, and Economic Sciences and in Education and Human Resources (BCC-SBE/EHR) award.

References

[1] Maksimovic, M.D., Veljkovic, N.Z., and Stoimenov, L.V., "Platforms for open government data," Telecommunications Forum (TELFOR), 2011 19th, vol., no., pp.1234,1237, 22-24 Nov. 2011. doi: 10.1109/TELFOR.2011.6143774

[2] "Chicago's WindyGrid: Taking Situational Awareness to a New Level." `http://datasmart.ash.harvard.edu/news/article/chicagos-windygrid-taking-situational-awareness-to-a-new-level-259` [Accessed July 7, 2015]

[3] The Urban Center for Computation and Data, at the Computation Institute of the University of Chicago and Argonne National laboratory. `http://www.urbanccd.org` [Accessed July 7, 2015]

[4] NSF 1244749, "BCC-SBE: An Urban Sciences Research Coordination Network for Data-Driven Urban Design and Analysis. PI Catlett, C., University of Chicago. 2012-2015.

[5] `http://www.socrata.com/` [Accessed July 7, 2015]

[6] http://ckan.org/ [Accessed July 7, 2015]

[7] https://github.com/UrbanCCD-UChicago/plenario [Accessed July 7, 2015]

[8] Moser, W, "What Chicago's 'Array of Things' Will Actually Do," Chicago Magazine, January 27, 2014. See also http://ArrayofThings.github.io [Accessed July 7, 2015]

[9] http://www.sqlalchemy.org/ [Accessed July 7, 2015]

[10] http://flask.pocoo.org/ [Accessed July 7, 2015]

[11] http://leafletjs.com/ [Accessed July 7, 2015]

[12] http://www.openstreetmap.org/about [Accessed July 7, 2015]

[13] http://www.celeryproject.org/ [Accessed July 7, 2015]

[14] http://redis.io/ [Accessed July 7, 2015]

[15] "The Sustainable Development Program." http://www.sf-planning.org/index.aspx?page=3051 [Accessed July 7, 2015]

[16] Voronoi, G., Nouvelles applications des paramétres continus á la théorie des formes quadratiques. Deuxième mémoiure: recherches sur les parallèloedes primitifs, J. reine angew. Math. 134, 198-287 (1908)

[17] Wong, D., "The modifiable areal unit problem (MAUP)", In Fotheringham, A Stewart; Rogerson, Peter. *The SAGE handbook of spatial analysis*. pp. 105–124 (2009)

[18] "Geohash," https://en.wikipedia.org/wiki/Geohash [Accessed July 7, 2015]

[19] Malik, T., Chard, K., and Foster, I., "Benchmarking cloud-based tagging services". In *In IEEE 30th International Conference of Data Engineering Workshops (ICDEW)*, pp. 231-238 (2014)

[20] Ray, S., Simion, B., and Brown, A. D., "Jackpine: A benchmark to evaluate spatial database performance". In IEEE 27th International Conference on Data Engineering (ICDE), pp. 1139-1150, IEEE. (2011).

Riding from Urban Data to Insight Using New York City Taxis

Juliana Freire[1,2] Cláudio Silva[1,2] Huy Vo[1,2]
Harish Doraiswamy[1] Nivan Ferreira[1] Jorge Poco[1]
{juliana.freire,csilva,hvo,harishd,nivan.ferreira,jpocom}@nyu.edu
[1]Department of Computer Science and Engineering
[2] Center for Urban Science and Progress
New York University

Abstract

About half of humanity lives in urban environments today and that number will grow to 80% by the middle of this century. Cities are thus the loci of resource consumption, of economic activity, and of innovation. Given our increasing ability to collect, transmit, store, and analyze data, there is a great opportunity to better understand cities, and enable them to deliver services efficiently and sustainably while keeping their citizens safe, healthy, prosperous, and well-informed. But making sense of all the data available is hard. Currently, urban data exploration is often limited to confirmatory analyses consisting of batch-oriented queries and the exploration of well-defined questions over specific regions. The lack of interactivity makes this process both time-consuming and cumbersome. This problem is compounded in the presence of big, multivariate spatio-temporal data, which is ubiquitous in urban environments. Another challenge comes from the need to empower social scientists, policy makes and urban residents who lack computer science expertise to leverage these data. In this paper, we give an overview of our recent work on techniques that combine data management and visualization to enable a broad set of users to interactively explore large, spatio-temporal data. We describe a visual query interface that simplifies the process of specifying spatio-temporal queries as well as new indexing technique that enables these queries to be evaluated at interactive rates. We also present a scalable framework that applies computational topology to automatically find interesting *data slices so as to help guide users in the exploratory process.*

1 Introduction

Today, 50% of the world's population lives in cities and the number will be 70% by 2050; North America is already 80% in cities, rising to 90% by 2050 [46]. Cities are thus the loci of economic activity and innovation. At the same time, most cities face huge challenges around transportation, resource consumption, housing affordability, and inadequate or aging infrastructure. Data, along with visualization and analytics capabilities, can help significantly with these challenges.

Our increasing ability to collect, transmit, and store data, coupled with the growing trend towards openness [4, 18, 20, 23, 36, 37, 43], creates a unique opportunity that can benefit government, science, citizens and industry. By integrating and analyzing multiple data sets, city governments can go beyond today's imperfect and

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

often anecdotal understanding of cities to enable better operations and informed planning [16, 20]. Scientists can engage in data-driven science and explore longitudinal processes to understand people's behavior [21]; identify causal relationships across data sets, which can in turn, influence policy decisions [11, 42]; or create models and derive predictions that benefit citizens [15]. Putting urban data in the hands of citizens has the potential to improve governance and participation, and in the hands of entrepreneurs and corporations it will lead to new products and services for governments, firms, and consumers. The challenge now lies in making sense of all the data so that they can be used effectively to answer the right questions.

Urban data is unique in that it captures the behavior of the different components of a city, namely its citizens, existing infrastructure (physical and policies), the environment (e.g., weather), and interactions between these elements [28]. To understand a city and how its components interact, intricate analyses are necessary. These include flexible exploration and visualizations that span different geographical regions and multiple time slices. However, urban data analysis has often been limited to well-defined questions, or confirmatory data analysis [45]. The common practice is for domain experts to formulate hypotheses on the basis of theory and anecdotal experience, then for data scientists with expertise in geographical information systems (GIS) and statistical analysis tools to select relevant data, carry out analyses, and finally, the domain experts can inspect the results to verify whether they disprove or support the hypotheses. In order to answer even simple questions, a large number of plots and tables may have to be generated, each of which is individually programmed and manually composed for analysis. Because data selection is often decoupled from the analysis and visualization, the context switch between the different software components used for these tasks makes it hard to keep spatio-temporal context. Together with the glut of plots derived, this creates a heavy cognitive load for the users, while the batch-oriented analysis pipeline hampers exploration across data sets, which is essential for understanding trends and potential causal mechanisms. Furthermore, the dependency on data specialists distances the domain experts from the data, limiting their opportunity to explore new directions. The trend towards broad-scale data collection, rather than limited collection targeted at specific questions, makes it clear that this process cannot scale, and that tools are needed that foster hypothesis-generating analyses.

One of the shifts we're seeing with observational data is broad-scale collection, rather than limited collection targeted at one hypothesis. So the thought is that the kind of tools that worked for analyzing a "single hypothesis" dataset might not be a good match for the kind of exploratory,

The lack of interactivity, along with the recent explosion in data volume and complexity, make it clear that this process cannot scale.

An important goal of our research is to *enable domain experts to freely explore a large number of urban data sets and interactively analyze the many different facets of these data*. This involves fundamental challenges. First and foremost, we need usable tools, designed for users who do not have computer science training. Second, not only can urban data be large, but often they contain both temporal and spatial components, in addition to multiple variables. In a recent study of open data published by cities in North America, we found that over 50% of the tabular data contained spatial attributes and roughly 48% included time information [4]. Attaining interactivity while exploring spatio-temporal data is difficult. Even though there has been substantial work on spatio-temporal indexing, most techniques aim to speed up batch queries, and are not able to support the query rates that interactive visual analytics applications demand. Another challenge comes from the fact that there are too many data slices to explore, that cover different regions and time ranges, making it hard to identify interesting patterns or events.

In this paper, we give an overview of recent work that combines data management and visualization to support the interactive exploration of large urban data [9, 14]. We use New York City taxi data as a case study to both illustrate general challenges that arise in urban data exploration, and to demonstrate the effectiveness and usefulness of the techniques we have developed. We describe the taxi data in Section 2. In Section 3, we present TaxiVis, a visual analytics tool that allows users to specify complex spatio-temporal queries through a visual interface. We describe the visual language as well as a new indexing strategy that allows queries to be evaluated at interactive rates. TaxiVis also implements a number of visualization and interaction techniques to

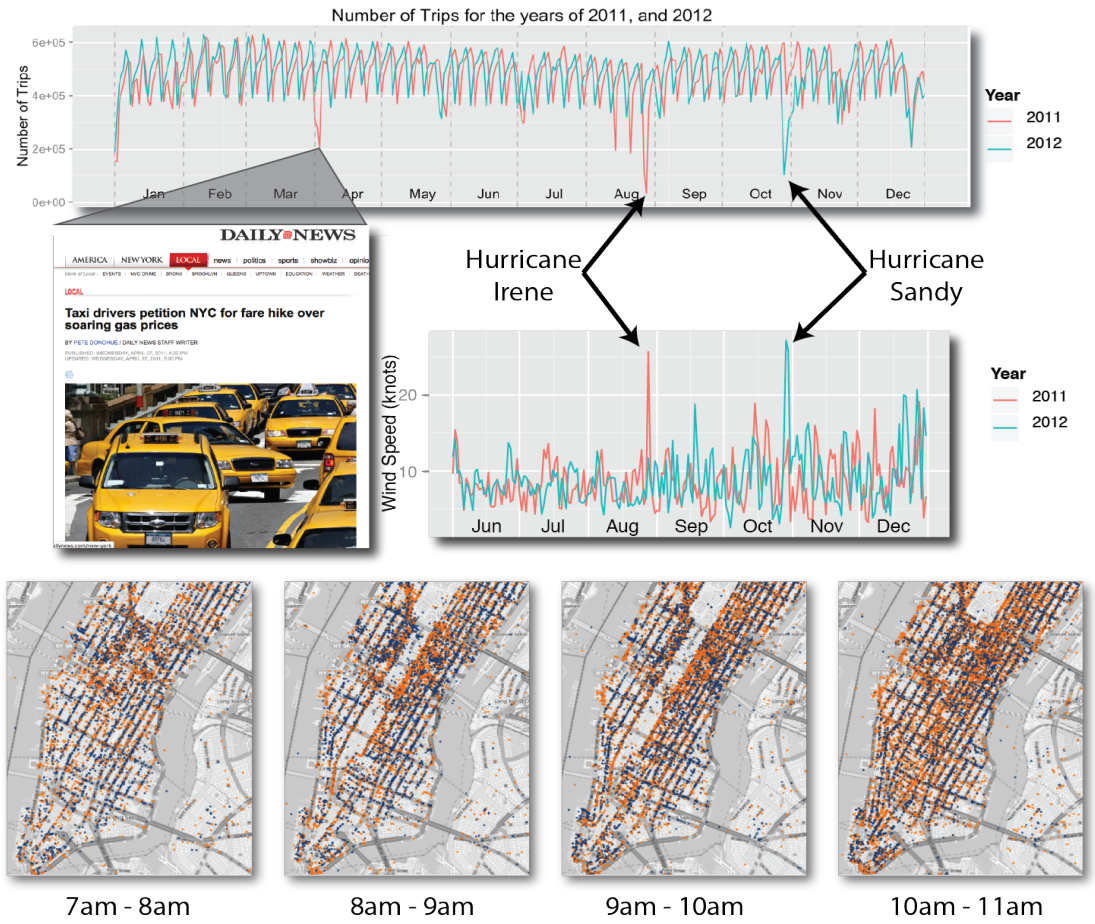


Figure 1: Taxis as sensors of city life. The plot on the top shows how the number of trips varies over 2011 and 2012. While some patterns are regular and appear on both years, some anomalies are clear, e.g., the drops in August 2011 (Hurricane Irene) and in October 2012 (Hurricane Sandy). Another large discrepancy happened in April 2011, when gas prices increased. In the bottom, we show pickups (blue) and dropoffs (orange) in Manhattan on May 1st from 7am to 11am. Notice that from 8-10am, there are virtually no trips along 6th Avenue, indicating the traffic was blocked.

streamline exploration, including, coordinated views and parameter sweeps. The system has been successfully deployed at two New York City agencies: the Taxi & Limousine Commission (TLC) and at the Department of Transportation (DoT), where users have lauded its usability and speed. But even with such a system, looking for interesting patterns across a very large number of spatio-temporal slices can be like looking for a needle in a haystack. To help guide users towards interesting times slices, we designed a new topology-based event detection technique that is both scalable and, unlike existing approaches, can detect events with arbitrary spatial geometry. We describe the technique in Section 4 and show how it can support event queries. We conclude in Section 5, where we discuss open problems and directions for future work.

2 Exploring New York City Taxi Data: Opportunities and Challenges

Taxis are central component of the New York City transportation system. Every day, there are over 500,000 taxi trips transporting about 600,000 people [44]. Through the meters installed in each vehicle, the Taxi & Limousine Commission (TLC) captures detailed information about these trips, which consists of two spatial attributes (pickup and dropoff locations), two temporal attributes (pickup and dropoff times), and additional

attributes including taxi id, distance traveled, fare and tip amount. The large number of vehicles, trips they make, and people they carry, make taxis valuable sensors that can provide unprecedented insight into many different aspects of city life, from economic activity and human behavior to mobility patterns. For example, the transactional attributes present in the taxi data enables the study of the economics of fare structure and optimal fleet size [40, 41].

Consider the plot in Figure 1 (top), which shows how the number of trips per day varies over 2011 and 2012. Note the regularity in the trip distribution over the two years. For example, on Thanksgiving, Christmas and New Year's eve, there is a substantial drop in the number of trips. But the plot also shows some anomalies. There are big drops in August 2011 and October 2012, which correspond to hurricanes Irene and Sandy, respectively. Looking at the data at a finer scale, other interesting patterns emerge. The maps in Figure 1 (bottom) show the density of taxis across Manhattan from 7am to 11am, on May 1st, 2011. From 8am to 10am, taxis disappear along 6th avenue, from Midtown to Downtown; and then, at 10am they reappear. As it turns out, during this period, the streets were closed to traffic for the NYC Five Boro Bike Tour.[1] Other useful information can also be discovered by analyzing the taxi data set, from popular night spots and economically disadvantaged neighborhoods that are underserved by taxis, to mobility patterns across regions at different times and days.

Not surprisingly, exploring these data is challenging due to its size and complexity. There are over 170 million taxi trips in NYC every year. The current approach used by domain experts is to store the data in a general-purpose relational DMBS and perform queries to answer questions posed out of intuition or field observation. The results of these queries are then fed into different software tools such as R, Excel and ArcGIS for further analysis and visualization. This workflow makes the analysis process inefficient. The disconnect between the data selection process and visualization hampers exploration. The context switch between the different software components creates a heavy cognitive load for the users and makes it hard to keep spatio-temporal context. There is also a steep learning curve for users to master these tools. Furthermore, off-the-shelf DMBS are not built for interactivity. For example, common queries over on the taxi data (even in the presence of spatial indexes) range from tens of seconds to minutes (see Section 3). These response times are not acceptable for interactive visual analytics, since users perceive questioning and answering as separate tasks [17].

3 Visually Exploring Spatio-Temporal Data

Visualization and visual analytics systems help people explore and explain data by allowing the creation of both static and interactive visual representations [24, 25, 29, 31, 30, 47]. A basic premise of visualization is that visual information can be processed at a much higher rate than raw numbers and text: as the cliché goes, "A picture is worth a thousand words." Well-designed visualizations substitute perception for cognition, freeing up limited cognitive and memory resources for higher-level problems [35]. There are several visualization tools that give users access to advanced visualization techniques. But the application of visualization technology to large data is non-trivial.

A widely-used method to analyze large data is to take a small subset of the data (often by sub-sampling) and study it with existing (non-scalable) tools. Hypotheses are generated from this sample, which are then tested on the complete data set through confirmatory analysis [17]. This approach has many shortcomings, one of them is the potential bias introduced by the use of small samples. This is further exacerbated by high-dimensional data that comes from multiple sources. Patterns that might be easy to find on the complete data sets might be obscured in small samples.

Working on the whole data set has many advantages but comes at a high computational cost. This creates new challenges for data management systems, since to be effective, visualization tools must be interactive, requiring sub-second response times. In a recent study, Liu and Heer [33] concluded that even relatively short delays in visualization systems can harm user activity, data set coverage, and how many observations and hypotheses

[1] http://www.nycbikemaps.com/spokes/five-boro-bike-tour-sunday-may-1st-2011

are generated. Fekete and Silva [13] argued that although there has been much work on scaling databases for big data, existing technologies do not meet the requirements needed to interactively explore massive or even reasonably sized data sets. Recognizing this limitation, several recent works have started to address the problems of providing efficient support for visualization [49] and interactive queries over large tabular data [2, 5, 26, 27, 32, 34, 48].

In what follows, we present our approach to support interactive exploration of spatio-temporal data, which was implemented in the TaxiVis system.

3.1 The TaxiVis System

TaxiVis was designed to support interactive analysis of NYC taxi data. It implements a visual model that is able to express complex queries and an index structure that enables interactive response times for spatio-temporal queries. We describe these two components below, and discuss how they were combined in the TaxiVis system.

The query model enable users to pose queries over all the dimensions of the data and flexibly explore the attributes associated with the taxi trips. The components of the user interface are shown in Figure 2. Queries can be interactively composed and refined in the Map view (B), as well as generalized by performing parameter sweeps. Query results can be visualized in the data summary view (D) in a variety of ways including time-series plots, histograms, scatterplots and heatmaps. By combining data selection and result visualization in the same environment, TaxiVis allows users to explore multiple data slices while maintaining the spatio-temporal context. The system also implements a number of strategies to render a large number of graphical primitives on a map, as well as the use of adaptive level-of-detail rendering to provide clutter-free visualization of the results (see Figure 3). TaxiVis also implements a number of visualization and interaction techniques to streamline exploration, including, multiple coordinated views and parameter sweeps. The former is illustrated in Figure 3, which shows a comparison of the number of pickups in different neighborhoods on Mondays and Sundays.

Figure 2: TaxiVis user interface components. (A) Time selection widget, (B) Map, (C) Tool bar, and (D) Data summary.

TaxiVis is currently being used at the NYC Department of Transportation (DoT) and the TLC. The feedback we have received from them was very positive. The analysts stated that "The speed at which the tool permits us to work has saved multiple hours of staff time and has dramatically improved the unit's output and capabilities". We should note that while the original motivation to build TaxiVis was to analyze taxi data, we have used the system to explore other spatio-temporal data sets, including: NYC CitiBike, property ownership [22], 311 complaints [1], geo-tagged tweets, and energy consumption.

3.2 Visual Query Model

To address the usability limitations of existing tools, we designed a new visual query model that supports complex spatio-temporal queries over origin-destination data [14]. Users need not be experts in any textual query language: users specify queries visually and they can iteratively refine their queries through direct manipulation of the results. The model is expressive and supports a wide class of queries, including the query classes for spatio-temporal data defined in Peuquet's triad framework [39].

In our model, queries are of the following form: **SELECT** * **FROM** trips **WHERE** <constraints>. The general idea is to have users specify the constraints for this query template through visual operations. There

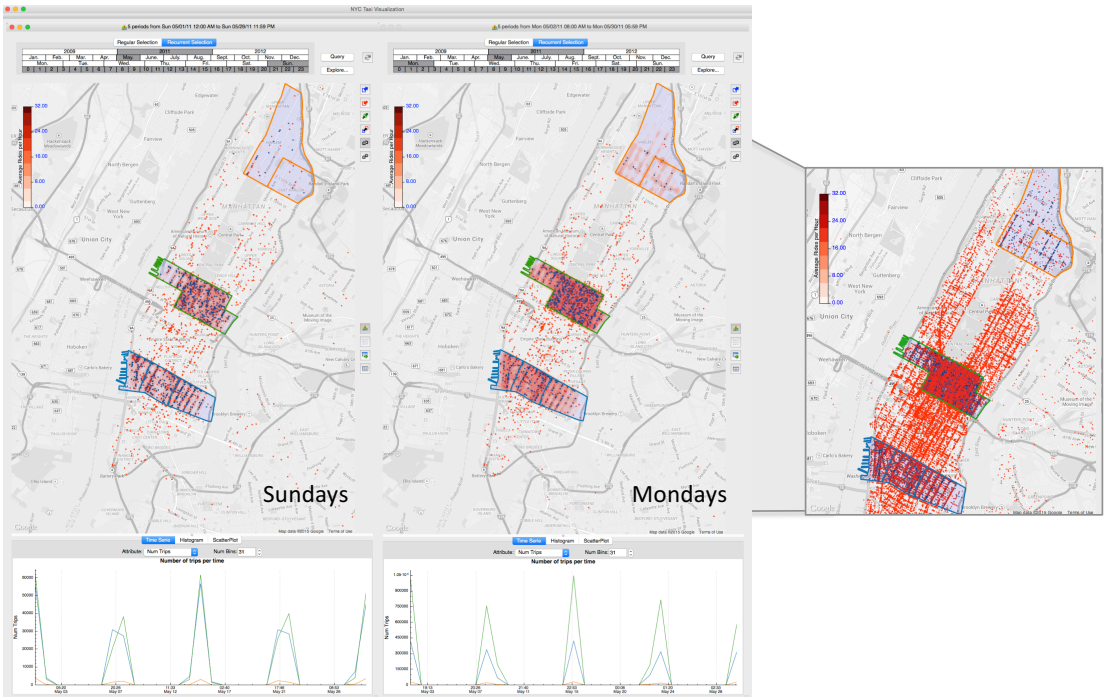


Figure 3: Using multiple views to explore how the number of pickups varies on Sundays and Mondays (in May 2011) for different neighborhoods. While there are many more pickups Midtown (green) throughout the day on Mondays, on Sundays, West, East and Greenwich Village (blue) get busier. This analysis also shows that Harlem (orange) is underserved by taxis. Because queries can return a large number of results, TaxiVis uses adaptive level-of-detail (LOD) rendering to display the results on the map. As shown on the right, without LOD, the map gets cluttered.

are three types of constraints that correspond to the components of the data: *spatial, temporal, and attributes*. Each query is associated with the set of trips contained in its results. Since each trip is uniquely identified by the trip id, *queries can be composed*: users can iteratively refine queries and further explore the results. This has two important implications: it allows the creation of summaries and visualizations while maintaining the spatial and temporal contexts, and enables queries to be applied directly to the derived visualizations. To formalize the process of query composition and properly define query semantics, we use two types of queries: *atomic* queries and *complex* queries, where the latter uses atomic queries as building blocks.

Atomic Queries. An atomic query consists of a set of temporal, attribute and spatial constraints. *Temporal constraints* define intervals that bound the values of the time range of the query. A temporal constraint is specified by an interval $[t_{Min}, t_{Max}]$. A trip satisfies the constraint if $trip.pickup\_time$, $trip.dropoff\_time \in [t_{Min}, t_{Max}]$. It is also possible to have constraints that bound just the pickup or the dropoff time.

An *attribute constraint* can be expressed using equality conditions (for categorical attributes) or interval conditions (for numerical attributes). A trip satisfies an attribute equality constraint associated with a categorical attribute A if for the given value a, $trip.A = a$. If the constraint is associated with a numerical attribute, the trip satisfies the constraint for the interval $[l_A, r_A]$ if $trip.A \in [l_A, r_A]$.

Spatial constraints come in two flavors: single-region and directional constraints. A single-region constraint is defined by a connected spatial region and is associated either with the pickup location (start constraint) or the dropoff location (destination constraint). A trip satisfies the constraint for region r if $trip.pickup\_region \in r$ (for start constraints) or $trip.dropoff\_region \in r$ (for destination constraints). Directional constraints are used to

construct queries about origins and destinations. A directional constraint bounds the regions associated with both pickup and dropoff locations. Given source and destination regions, r_{source} and r_{dest}, respectively, a trip satisfies the constraint if *trip.pickup_location* $\in r_{source}$ and *trip.dropoff_location* $\in r_{dest}$.

We define a function called *result* which takes as input an atomic query and returns the set of all *trip* records that satisfy the query constraints. The *result* function determines how queries are evaluated. Atomic queries are *closed under intersection*, *i.e.*, since the query constraints are *closed under intersection* we can combine atomic queries to construct new ones. We do so by taking two atomic queries Q_1, Q_2, and constructing a third query, Q_3, which is given by the intersection of the corresponding constraints. By definition, we have $result(Q_3) = result(Q_1) \cap result(Q_2)$.

Complex Queries. A complex query is constructed by combining a set of atomic queries through disjunction. We evaluate those queries by extending the *result* function inductively. Note that an atomic query is a special case of a complex query, where the query set has a single element. Given two complex queries, Q_1 and Q_2, $result(Q_1 \cup Q_2) = result(Q_1) \cup result(Q_2)$. In general, given an atomic query Q it is not possible to find an atomic query Q' such that $result(Q') = result(Q)^C$ (the complement of $result(Q)$). However, it is always possible to define a complex query Q' that satisfies this condition. Thus, set theoretic operations can be performed on the result of complex queries to build new complex queries.

Visual Representation. Figure 2 illustrates how atomic and complex queries are represented visually in our system. Temporal constraints are specified using time-selection widgets (A), and attribute constraints are defined in a separate view (see [14] for details). Here, to illustrate the semantics of the query model, we focus on spatial views that are defined on the map view (B). Single-region constraints are defined by polygons and directional constraints are defined by arrows. The transparent color in the interior of the polygons define the type of the constraint: blue means start constraint, red means destination constraint (see Figure 2). The colors on polygon borders and arrows identify distinct queries (there are 3 queries – orange, red, and blue). The orange and red queries are atomic queries, consisting of only atomic temporal and spatial constraints. The blue query Q is a complex query, composed by the union of two atomic queries: a single-region start query Q_1 and a directional query Q_2. In an SQL-like textual notation, Q_1 can be represented as:

SELECT * **FROM** trips
WHERE *trip.pickup_time* $\in [05/01/2011, 05/07/2011]$ **AND** *trip.pickup_location* $\in R_1$

where R_1 denotes the blue region selected in the map. And Q_2:

SELECT * **FROM** trips
WHERE *trip.pickup_time, trip.dropoff_time* $\in [05/01/2011, 05/07/2011]$ **AND** *trip.pickup_location* $\in$ NYCNeighborhood('*Gramercy*') **AND** *trip.dropoff_location* $\in$ NYCRegion('*Times Square*')

where NYCNeighborhood and NYCRegion are functions that given a neighborhood name or region name, respectively, returns the corresponding spatial region.

3.3 Query Evaluation

While easy-to-use, the visual interface leads to an important challenge: a user can issue several large queries and visualizations need to be created for their results at interactive speeds. These queries can be complex. For example, in Figure 3, two queries are represented: the query on the left asks for all pickups in three different neighborhoods on all Sundays in May 2011, and the one on the right explores a different time pattern—all Mondays in May 2011. Users can not only select arbitrary regions, but they can also interactively move polygons around the map, thus generating a series of queries.

To support these queries, we first experimented with traditional database systems, both open source and

Table 1: Summary of experiments with data storage strategies.

	SQLite	PostgreSQL	TaxiVis Storage
Storage space	100GB	200GB	30GB
Index construction time	52h	13h	28m
1k-query	8s	3s	0.2s
100k-query	85s	24s	2s

commercial. In spite of extensions for spatial queries, their query performance is not suitable for interactivity, not to mention the fact that they take a considerable length of time to build the spatial indices. Table 1 shows the performance for SQLite and PostgreSQL with PostGIS, with the former being used for in-memory storage. SQLite took 52 hours just to build the indices for data corresponding to a single year of taxi trips, which as mentioned earlier consists of approximately 170 million trips. Moreover, a single atomic spatio-temporal query could take from seconds to tens of seconds to complete, while complex ones such as those specified by the recurrent time selection widget, can take minutes. Finally, another shortcoming of these database systems is their large memory footprint. In our experiments, SQLite and PostgreSQL used more than 100GB of RAM (in memory setup for SQLite) and 200GB, respectively.

In order to address these issues, we have built a light-weight database variant that allows fast queries on all attributes including spatio-temporal constraints. Our implementation is based on a space-partitioning data structure, the k-d tree [8], that treats each taxi trip as a point in a k-dimensional space. In our implementation, points are only stored in leaves. Our code takes only 30 minutes to build the indices for the full three years of data and uses only 30GB of disk space. At run-time, the whole data structure, including the data points, are mapped to the system virtual memory, therefore, it may operate in-core or out-of-core adaptively, depending on the available resources. In our tests, compared to the database systems mentioned above, the memory usage is considerably smaller, and queries are significantly faster – they can be evaluated within the bounds required by our interactive system. In Table 1, we summarize the results obtained by our our experiments where 1k-query and 100k-query refer to queries returning approximately 1000 and 100,000 trips, respectively.

4 Automatically Finding Interesting Spatio-Temporal Slices

An important challenge in the exploration of large spatio-temporal data is how to identify *interesting* data slices. While TaxiVis simplifies the exploration of the taxi data, it is not practical to exhaustively examine each data slice. Aggregation can help overcome this problem, but it does so at the cost of occluding small or local patterns in the data [3]. For example, events such as the NYC Five Boro Bike tour shown in the maps of Figure 1, affect a relatively small region in the city over a short period of time, and thus may not be visible when the data is aggregated over time or space.

In more recent work, we designed an automated event-detection algorithm based on techniques from computational topology to help *guide* users towards interesting patterns and data slices [9]. The topological representation of large data sets provides an abstract and compact global view that captures different features and leads to enhanced and easier analysis across applications [19, 38]. The advantages of using topology-based techniques are twofold: topological data structures such as contour trees [7] and Reeb graphs [10], which are used to to identify topological features, can be efficiently computed; and unlike existing approaches, topology-based techniques allow for detection of events that can have arbitrary spatial geometry.

Detecting events in the taxi data. Event detection is accomplished in two steps. First, a time-varying scalar function is derived from the input data. Here, we assume that the temporal dimension is represented as a set of discrete time steps. If we consider the taxi data, we can define the time-varying scalar function as the density of

(a) (b)

Figure 4: **(a)** Scalar function for a time step corresponding to 9-10am on May 1, 2011. **(b)** Two of the minima of this function correspond to the the path taken by the NYC Five Boro Bike tour in Lower and Midtown Manhattan.

taxis at each point in NYC over hourly intervals. The domain of this function is represented as a graph. Next, the set of events is computed as topological features. In particular, we consider two types of features: *minimum* and *maximum*. Given a single time step in the taxi density function, a minimum of this function represents a region where the density of taxis is lower than its local neighborhood, implying a relative scarcity of taxis in that region. Such events, when in a busy region, could be due to road blocks. For example, Figure 4(a) shows the scalar function corresponding to a time step when the NYC Five Borough Bike tour from Figure 1 occurred. Two of the minima events are shown in Figure 4(b), which correspond to the path taken by the bike tour in Manhattan. Similarly, a maximum represents a region where the density of taxis is higher than that of its local neighborhood.

Event Index. We implemented a visual interface on top of the TaxiVis infrastructure that allows users to explore events at different time steps. In order to guide users towards potentially interesting events from a possibly large number of events, we designed a rudimentary hash-like indexing scheme that groups similar events across time slices. Let an event E be represented as a pair (R,τ), where R denotes the region associated with an event (a subgraph of the domain graph), and τ is a real number that represents the topological significance of E. The pair above, together with the time of event, is used to identify the appropriate bin, called the *event group*, for E as follows. Given two events $E_1(R_1,\tau_1)$ and $E_2(R_2,\tau_2)$, the *graph distance metric* [6], δ, measures the *geometric similarity* between R_1 and R_2:

$$\delta(E_1,E_2) = 1 - \frac{|R_1 \bigcap R_2|}{\max(|R_1|,|R_2|)},$$

where $R_1 \bigcap R_2$ denotes the maximum common subgraph between R_1 and R_2, and $|R|$ denotes the number of nodes in R. The *topological similarity* between two events is defined as:

$$T(E_1,E_2) = |\tau_1 - \tau_2|$$

Two events E_1 and E_2 are in the same event group of the index if

1. $\delta(E_1,E_2) \leq \varepsilon_\delta$ and $T(E_1,E_2) \leq \varepsilon_\tau$, where, ε_δ and ε_τ are user-defined thresholds.

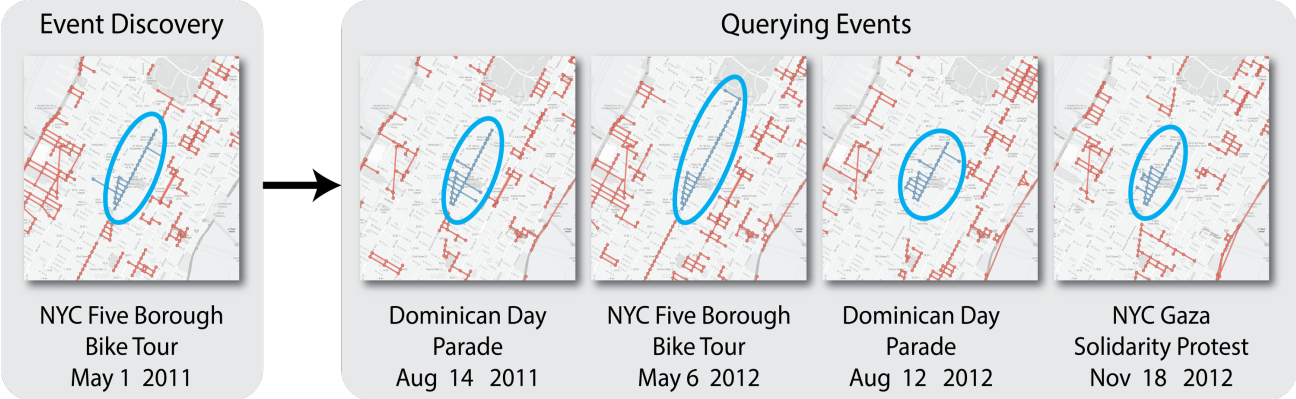


Figure 5: Event-guided exploration of taxi data corresponding to years 2011 and 2012. The user is able to find an event corresponding to the the NYC Five Borough Bike tour that occurred on 1 May 2011 between 8 am and 10 am. Searching for similar events yields the same bike tour that happened in 2012, together with the Dominican day parades for both years. Additionally, we found the Gaza solidarity protest, which was held at the same location.

2. E_1 and E_2 occur within the same time period. We used a time period equal to a month in our implementation.

We define two attributes for each event group – range and density. The *range* of an event group is defined as the amount of time between the first and the last event in that group based on their time steps. Its *density* is defined as the number of events of that group that happen per time unit. It measures the time frequency of the events within the group. These two attributes can be used to identify not only periodic events (hourly, daily, and weekly events), but also events with varying frequency (rare events and trends). Thus by exploring properties of event groups through the visual interface, users can effectively identify events of interest. Each event group $\Sigma = \{E_1, E_2, \ldots, E_k\}$, is also associated with a key (R_Σ, τ_Σ), where

$$R_\Sigma = \bigcap_{i \in [1,k]} R_i \quad \text{and} \quad \tau_\Sigma = \sum_{i=1}^{k} \tau_i / k$$

This allows users to search the data for occurrences of a given pattern. This is accomplished by comparing the similarity between the given pattern and the keys associated with the different event groups. Figure 5 illustrates the result of querying the taxi data for events similar to the NYC Five Boro Bike tour.

5 Conclusions

In this paper, we presented an overview of our recent work on techniques to support exploratory analysis and visualization of spatio-temporal urban data. We described TaxiVis, a visual analytics tool and its two main components: a visual query interface that simplifies the process of specifying spatio-temporal queries, and a indexing technique that enables these queries to be evaluated at interactive rates. We also described a topology-based approach to automatic event detection, and how it can help guide users towards interesting time slices. While this work is a step towards scalable and usable visual analytics for spatio-temporal data, there are many open problems we intend to pursue in future work.

Given the growing trend towards transparency, and the large number of open data sets, there is a great opportunity to leverage these data to better understand cities. But this also creates many challenges. Visualization and

visual analytics systems have been successfully used to aid users obtain insight from data. But to be effective, visualization systems have to be interactive, requiring sub-second response times [33, 13]. Having been designed for batch queries issued through a text-based or terminal interfaces, existing relational database technologies and business intelligence systems used for OLAP analyses are not suitable backends for these tools [50]. New data management techniques are needed to support interactive visualization [13, 12, 49] Another important challenge comes from the sheer number of data sets available: it is impossible to apply conventional database integration and warehousing techniques where the goal is to establish a single mediated schema. Therefore, we need new methods and tools that help users integrate data on the fly, in a task-oriented manner: as users make discoveries and formulate hypothesis, they should be able to bring in new data that allows them to test these hypotheses. As multiple data sets are integrated, we face additional problems, including the need to support interactive queries that span these data sets, and to visually fuse data sets as different components of a visualization.

Acknowledgments: We thank David Maier for his constructive comments on this manuscript, and the New York City TLC and DoT for providing the data used in this paper and feedback on our results. This work was supported in part by a Google Faculty Award, IBM Faculty Awards, the Moore-Sloan Data Science Environment at NYU, the NYU School of Engineering, the NYU Center for Urban Science and Progress, and NSF awards CNS-1229185 and CI-EN 1405927. Silva has been partially funded by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER).

References

[1] NYC 311. `http://www1.nyc.gov/311`.

[2] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data. In *Proc. EuroSys*, pages 29–42, 2013.

[3] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. From movement tracks through events to places: Extracting and characterizing significant places from mobility data. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 161–170. IEEE, 2011.

[4] L. Barbosa, K. Pham, C. Silva, M. Vieira, and J. Freire. Structured open urban data: Understanding the landscape. *Big Data*, 2(3), 2014.

[5] L. Battle, M. Stonebraker, and R. Chang. Dynamic reduction of query result sets for interactive visualizaton. In *Proc. IEEE Big Data*, pages 1–8, 2013.

[6] H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Lett.*, 19(3):255–259, 1998.

[7] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *Comput. Geom. Theory Appl.*, 24(2):75–94, 2003.

[8] M. De Berg, M. Van Kreveld, M. Overmars, and O. Schwarzkopf. *Computational Geometry*. Springer, 1997.

[9] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. Silva. Using topological analysis to support event-guided exploration in urban data. *IEEE TVCG*, 20(12):2634–2643, 2014.

[10] H. Doraiswamy and V. Natarajan. Computing Reeb graphs as a union of contour trees. *IEEE Transactions on Visualization and Computer Graphics*, 19(2):249–262, 2013.

[11] I. Ellen, J. Lacoe, and C. Sharygin. Do foreclosures cause crime? *Journal of Urban Economics*, 74:59–70, 2013.

[12] J.-D. Fekete. Visual analytics infrastructures: From data management to exploration. *IEEE Computer*, 46(7):22–29, 2013.

[13] J.-D. Fekete and C. Silva. Managing data for visual analytics: Opportunities and challenges. *IEEE Data Eng. Bull.*, 35(3):27–36, 2012.

[14] N. Ferreira, J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[15] B. Ferris, K. Watkins, and A. Borning. OneBusAway: Results from providing real-time arrival information for public transit. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1807–1816, New York, USA, 2010. ACM.

[16] A. Feuer. The mayor's geek squad. New York Times, March 24, 2013.

[17] D. Fisher. Incremental, approximate database queries and uncertainty for exploratory visualization. In *Proceedings of the 2011 IEEE Symposium on Large Data Analysis and Visualization*, LDAV '11, pages 73–80. IEEE, 2011.

[18] Freedom of Information Act (FOIA), 2014. http://www.foia.gov.

[19] A. T. Fomenko and T. L. Kunii, editors. *Topological Modeling for Visualization*. Springer Verlag, 1997.

[20] B. Goldstein and L. Dyson. *Beyond Transparency: Open Data and the Future of Civic Innovation*. Code for America Press, San Francisco, USA, 2013.

[21] A. Grossman and A. Sun. MTA swipes show subway trends. http://online.wsj.com, October 19, 2011.

[22] T.-A. Hoang-Vu, V. Been, I. G. Ellen, M. Weselcouch, and J. Freire. Towards understanding real-estate ownership in New York City: Opportunities and challenges. In *Proceedings of the Workshop on Data Science for Macro-Modeling (DSMM)*, 2014. To appear.

[23] J. Höchtl and P. Reichstädter. Linked open data - a means for public sector information management. In *Electronic Government and the Information Systems Perspective*, volume 6866 of *Lecture Notes in Computer Science*, pages 330–343. Springer, Berlin Heidelberg, 2011.

[24] IBM. OpenDX. `http://www.research.ibm.com/dx`.

[25] C. Johnson, H. Pfister, T. Munzner, R. Moorhead, P. Rheingans, and T. Yoo. *NIH-NSF Visualization Research Challenges Report*. IEEE Press, 2006.

[26] U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl. M4: A visualization-oriented time series data aggregation. *PVLDB*, 7(10):797–808, 2014.

[27] N. Kamat, P. Jayachandran, K. Tunga, and A. Nandi. Distributed and interactive cube exploration. In *Proc. ICDE*, pages 472–483, 2014.

[28] B. Katz and J. Bradley. *The Metropolitan Revolution: How Cities and Metros Are Fixing Our Broken Politics and Fragile Economy*. Brookings Focus Book. Brookings Institution Press, 2013.

[29] Kitware. Paraview. `http://www.paraview.org`.

[30] Kitware. The visualization toolkit. `http://www.kitware.com`.

[31] Lawrence Livermore National Laboratory. VisIt: Visualize It in Parallel Visualization Application. `https://wci.llnl.gov/codes/visit` [29 March 2008].

[32] L. Lins, J. Klosowski, and C. Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *IEEE TVCG*, 19(12):2456–2465, Dec 2013.

[33] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2122–2131, Dec 2014.

[34] Z. Liu, B. Jiang, and J. Heer. immens: Real-time visual querying of big data. *Computer Graphics Forum (Proc. EuroVis)*, 32, 2013.

[35] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[36] City of Chicago data portal. `https://data.cityofchicago.org`.

[37] NYC OpenData. `https://nycopendata.socrata.com`.

[38] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny, editors. *Topological Methods in Data Analysis and Visualization*. Springer, 2010.

[39] D. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994.

[40] B. Schaller. Elasticities for taxicab fares and service availability. *Transportation*, 26(3):283–297, 1999.

[41] B. Schaller. A regression model of the number of taxicabs in US cities. *Journal of Public Transportation*, 8(5):63, 2005.

[42] A. E. Schwartz, I. G. Ellen, I. Voicu, and M. H. Schill. The external effects of place-based subsidized housing. *Regional Science and Urban Economics*, 36(6):679 – 707, 2006.

[43] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, H. Wendy, and M. Schraefel. Linked open government data: Lessons from data.gov.uk. *IEEE Intelligent Systems*, 27(3):16–24, 2012.

[44] Taxicab fact book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[45] J. W. Tukey. *Exploratory Data Analysis*. Pearson, 1977.

[46] UN 2012 world urbanization prospects: The 2011 revision highlights. `http://esa.un.org/unup/pdf/WUP2011_Highlights.pdf`, 2012. Page accessed Jan 2014.

[47] C. Upson et al. The application visualization system: A computational environment for scientific visualization. *IEEE Computer Graphics and Applications*, 9(4):30–42, 1989.

[48] H. Wickham. Bin-summarise-smooth: a framework for visualising large data. Technical report, had.co.nz, 2013.

[49] E. Wu, L. Battle, and S. R. Madden. The case for data visualization management systems. *PVLDB*, 7(10):903–906, 2014.

[50] K. Zoumpatianos, S. Idreos, and T. Palpanas. Indexing for interactive exploration of big data series. In *Proc. SIGMOD*, pages 1555–1566, 2014.

32nd IEEE International Conference on

Data Engineering

May 16-20, 2016 · Helsinki, Finland

http://icde2016.fi/

Call for Papers

Conference Organization

General Chairs
Boris Novikov (Saint Petersburg University, Russia)
Eljas Soisalon-Soininen (Aalto University, Finland)

Program Committee Chairs
Mei Hsu (HP Labs, USA)
Alfons Kemper (TU Munich, Germany)
Timos Sellis (RMIT University, Australia)

Program Committee Chair for Industrial and Applications Papers
C. Mohan (IBM Almaden, USA)

Program Committee Chair for Demos
Stefan Manegold (CWI, the Netherlands)

Workshop and Tutorial Chairs
Giovanna Guerrini (University of Genova, Italy)
Georgia Koutrika (HP Labs, USA)

Organizing Committee Chair
Sami El-Mahgary (Aalto University, Finland)

Publicity Chair
Antoni Wolski (AWO Consulting, Finland)

Web Chair
Anna Yarygina (Saint Petersburg University, Russia)

The annual ICDE conference addresses research issues in designing, building, managing, and evaluating advanced data systems and applications. It is a leading forum for researchers, practitioners, developers, and users to explore cutting-edge ideas and to exchange techniques, tools, and experiences. We invite submissions for research papers, industrial and applications papers, demonstrations, tutorials, and workshops.

Topics of Interest

- Cloud Computing and Database-as-a-Service
- Big Data and Data-Warehousing System Architectures
- Data Integration, Metadata Management, and Interoperability
- Modern Hardware and In-Memory Database Architecture and Systems
- Privacy, Security, and Trust
- Query Processing, Indexing, and Optimization
- Social Networks, Social Web, Graph, and Personal Information Management
- Crowdsourcing, Distributed, and P2P Data Management
- Streams and Sensor Networks
- High Performance Transaction Management
- Temporal, Spatial, Mobile, and Multimedia Data
- Strings, Texts, and Keyword Search
- Uncertain and Probabilistic Data
- Visual Data Analytics, Data Mining, and Knowledge Discovery

Important Dates

Submission due:	Oct. 19, 2015
Author's feedback (research papers only):	Dec. 1-4, 2015
Notification:	Dec. 20, 2015
Camera-ready copy due:	Jan. 25, 2016
Tutorial proposals due:	Nov. 15, 2015
Tutorial notification:	Dec. 20, 2015
Workshop proposals due:	Sep. 10, 2015
Workshop notification:	Oct. 10, 2015

Aalto University
School of Science

IEEE computer society

Data Engineering

It's FREE to join!

TCDE

tab.computer.org/tcde/

The Technical Committee on Data Engineering (TCDE) of the IEEE Computer Society is concerned with the role of data in the design, development, management and utilization of information systems.

- Data Management Systems and Modern Hardware/Software Platforms
- Data Models, Data Integration, Semantics and Data Quality
- Spatial, Temporal, Graph, Scientific, Statistical and Multimedia Databases
- Data Mining, Data Warehousing, and OLAP
- Big Data, Streams and Clouds
- Information Management, Distribution, Mobility, and the WWW
- Data Security, Privacy and Trust
- Performance, Experiments, and Analysis of Data Systems

The TCDE sponsors the International Conference on Data Engineering (ICDE). It publishes a quarterly newsletter, the Data Engineering Bulletin. If you are a member of the IEEE Computer Society, you may join the TCDE and receive copies of the Data Engineering Bulletin without cost. There are approximately 1000 members of the TCDE.

Join TCDE via Online or Fax

ONLINE: Follow the instructions on this page:

www.computer.org/portal/web/tandc/joinatc

FAX: Complete your details and fax this form to **+61-7-3365 3248**

Name _____

IEEE Member # _____

Mailing Address _____

_____

Country _____

Email _____

Phone _____

TCDE Mailing List

TCDE will occasionally email announcements, and other opportunities available for members. This mailing list will be used only for this purpose.

Membership Questions?

Xiaofang Zhou
School of Information Technology and Electrical Engineering
The University of Queensland
Brisbane, QLD 4072, Australia
zxf@uq.edu.au

TCDE Chair

Kyu-Young Whang
KAIST
371-1 Koo-Sung Dong, Yoo-Sung Ku
Daejeon 305-701, Korea
kywhang@cs.kaist.ac.kr

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398