Andrienko, Gennady; Andrienko, Natalia; Boldrini, Chiara; Caldarelli, Guido; Cintia, Paolo; Cresci, Stefano; Facchini, Angelo; Giannotti, Fosca; Gionis, Aristides; Guidotti, Riccardo; Mathioudakis, Michael; Muntean, Cristina Ioana; Pappalardo, Luca; Pedreschi, Dino; Pournaras, Evangelos; Pratesi, Francesca; Tesconi, Maurizio; Trasarti, Roberto

(So) Big Data and the transformation of the city

Published in:
International Journal of Data Science and Analytics

DOI:
10.1007/s41060-020-00207-3

Published: 01/05/2021

Document Version
Publisher's PDF, also known as Version of record

International Journal of Data Science and Analytics (2021) 11:311–340
https://doi.org/10.1007/s41060-020-00207-3

REGULAR PAPER

(So) Big Data and the transformation of the city

Gennady Andrienko[1,2] · Natalia Andrienko[1,2] · Chiara Boldrini[3] · Guido Caldarelli[5,6] · Paolo Cintia[10] · Stefano Cresci[3] · Angelo Facchini[5,7] · Fosca Giannotti[4] · Aristides Gionis[8,12] · Riccardo Guidotti[10] · Michael Mathioudakis[9] · Cristina Ioana Muntean[4] · Luca Pappalardo[4] · Dino Pedreschi[10] · Evangelos Pournaras[11] · Francesca Pratesi[10] · Maurizio Tesconi[3] · Roberto Trasarti[4]

Received: 5 August 2019 / Accepted: 28 February 2020 / Published online: 31 March 2020

Abstract
The exponential increase in the availability of large-scale mobility data has fueled the vision of smart cities that will transform our lives. The truth is that we have just scratched the surface of the research challenges that should be tackled in order to make this vision a reality. Consequently, there is an increasing interest among different research communities (ranging from civil engineering to computer science) and industrial stakeholders in building knowledge discovery pipelines over such data sources. At the same time, this widespread data availability also raises privacy issues that must be considered by both industrial and academic stakeholders. In this paper, we provide a wide perspective on the role that big data have in reshaping cities. The paper covers the main aspects of urban data analytics, focusing on privacy issues, algorithms, applications and services, and georeferenced data from social media. In discussing these aspects, we leverage, as concrete examples and case studies of urban data science tools, the results obtained in the "City of Citizens" thematic area of the Horizon 2020 SoBigData initiative, which includes a virtual research environment with mobility datasets and urban analytics methods developed by several institutions around Europe. We conclude the paper outlining the main research challenges that urban data science has yet to address in order to help make the smart city vision a reality.

Keywords Big data · Urban data science · SoBigData · Mobility datasets

1 Introduction

The digital revolution witnessed in the last decade offers tremendous opportunities to improve people's quality of life and to transform the way they interact with each other and experience their environment. Living in cities is a big part of modern society. Today, 55% of the world's population lives in urban areas, a proportion that is expected to increase to 68% by 2050.[1] Nowadays, an unprecedented amount of data is collected about human life in cities and other environments. Furthermore, the amount of gathered data is expected to increase in volume, variety, and granularity, in the coming years. The massive amount of available data is the result of monitoring and recording a wide range of human activities by a multitude of sensors and devices. Examples of data associated with the smart city initiative include mobility traces of citizens collected by mobile phones [110], vehicle trajectories collected by GPS devices [76], geolocated content uploaded by citizens to social media platforms [107], social relationships data recorded through mobile phone networks and online social networks [42], passenger trajectories collected by travel cards and other transportation devices [117], transportation data collected by vehicle sharing services [23] (bikes, scooters, cars, etc.), traffic volumes gathered by road sensors, video and photograph streams produced by cameras deployed in different parts of the city, spatiotemporal pollution levels, electricity and water grid data [108], satellite images, credit card transaction data, shopping records [54], crime and other safety-related open data [20], and much more.

[1] https://population.un.org/wup/.

This work was supported by the European Commission through the Horizon 2020 European Project "SoBigData Research Infrastructure—Big Data and Social Mining Ecosystem" (Grant Agreement 654024).

✉ Aristides Gionis
argioni@kth.se; aristides.gionis@aalto.fi

Extended author information available on the last page of the article

Distilling value from the available data to support the smart city vision is a major challenge that requires multidisciplinary research in engineering, city planning, operations research, statistics, economics, computer science, e-governance, policymaking, sociology, and more. The objective is to use the data in order to improve the quality of life of citizens, and at the same time optimize resources, reduce costs, and increase sustainability. Thus, big data analytics can be leveraged to provide new smart services to citizens or to help government and policymakers in making decisions supported by data. This is the goal of urban data science. Many steps forward have been made in the last few years, and the aim of this paper is to provide the reader with a systematic overview of the main cutting-edge topics regarding big data algorithms, methods, data issues, and tools that are shaping the transition to the digital urban environment. In order to illustrate these topics with concrete examples, we leverage the different research activities conducted within the "City of Citizens" exploratory of the SoBigData European project.[2] The project SoBigData [50] covers a wide range of topics in the area of data science applied to human social life, ranging from migration studies and demography 2.0 to sports analytics and analysis of the polarization in the political discourse. SoBigData features an entire thematic area (the "City of Citizens" exploratory) devoted to the topic of urban data science, whose results we discuss in the paper in order to illustrate its potential for improving the individual and collective well-being of people living in cities.

We begin our overview with Sect. 2, where we introduce algorithmic tools that address critical challenges in urban data science and that can be leveraged and combined to provide innovative urban services. First, we discuss how to model information extracted from location-based social networks in order to detect the rhythms of urban activities throughout the day. Then, we discuss the problem of location detection, which entails inferring the semantic (e.g., work or home) of the physical locations in which users roam. Another challenging problem considered in this section is that of trajectory generation, i.e., how to design algorithms able to generate a population of agents whose mobility patterns are indistinguishable from those of real individuals. Moving from descriptive to predictive approaches, we discuss two different prediction problems, namely predicting individual movement (next location) and predicting movement agenda (locations a user will visit during a given day). These types of predictive problems can be used to plan events and infrastructures, for individual gains as well as for public good. In Sect. 3, we present several methods of visual analytics for geolocated social media data, with an emphasis on photograph sharing and micro-blogging platforms. We accompany our discussion on visual analytics with several examples showcasing our methods for the tasks of detecting events and annotating them with explanatory tags for real-world data taken from locations around the world. Next, we move to services and applications (Sect. 4). First, we discuss how to leverage big data analysis to provide sightseeing recommendations to tourists. Then, we present an overview of how car sharing services can be improved by exploiting data analytics techniques. We next discuss the use of big data analytics for studying the link between human mobility, socioeconomic development, urban sustainability, and net negative cities. In Sect. 5, we overview the main software platforms developed and made available within SoBigData. These platforms (M-Atlas and EPOS) are fully fledged software solutions ready to be used and deployed in real systems. The important topic of data gathering with respect to privacy issues is discussed in Sect. 6. We emphasize the crucial issue of personal data privacy, and we present a distributed crowdsensing approach for data collection, which can offer a means for the users to control their data contribution and data privacy. We conclude our presentation in Sect. 7 by discussing a number of future challenges.

2 Algorithms for urban data analytics

In this section, we discuss algorithmic tools that are not focused on a specific urban application, but that can be leveraged and combined to provide advanced services to the cities of the future. Specifically, we discuss algorithms for extracting information about urban activities (i.e., how people engage with the different areas of a city) from location-based social networks like Foursquare (Sect. 2.1), for extracting the most important locations (e.g., home, workplace, etc.) of a user (Sect. 2.2), for simulating realistic human mobility patterns (Sect. 2.3), and for predicting people movements in a city (Sect. 2.4).

2.1 Characterizing urban activities

Location-based social networks (LBSNs) are online social platforms that allow their users to share their whereabouts with their friends and the public [107]. For example, Foursquare enables its users to generate "check-ins," i.e., digital notifications that inform their friends of their whereabouts. Each check-in contains information that reveals who (which user) spends time where (at what location), when (what time of day, what day of the week), and doing what (according to the kind of venue: shopping at a grocery store, dining at a restaurant, and so on).

By analyzing large amounts of activity traces from location-based social networks, we can obtain a fine-grained description of how citizens experience their cities and, in particular, a description that indicates what activity takes

[2] http://www.sobigdata.eu.

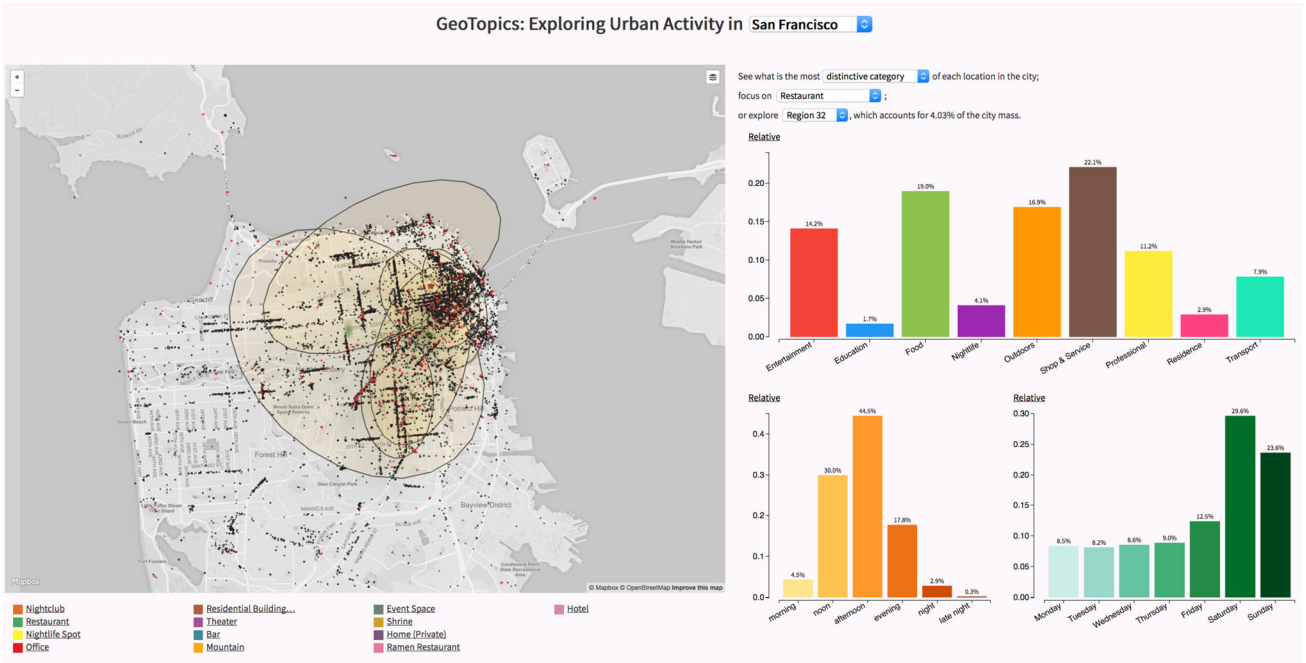


Fig. 1 A screenshot of the GeoTopics system. Users of the system can explore how urban activity is decomposed into different local activity "topics"

place at different locations of the city. For example, urban activity traces might reveal that citizens visit one region mainly for shopping in the morning, while another for dining in the evening. This information is presently collected by municipalities through costly, typically small-scale surveys involving direct interviews with people. Leveraging LBSN allows us to scale up and automatize data collection in a very efficient way, at the same time enabling real-time now-casting of urban activities. Furthermore, once such an urban activity description is available, one can ask more elaborate questions. For example, one might ask what features distinguish one region from another—some regions might be different in terms of the type of venues they host and others in terms of the visitors they attract. As another example, one might ask which regions are similar across cities.

One way to obtain such a description is to use location-based social network activity traces to build a *GeoTopics* model [34]. Inspired by topic models for text documents [19], a *GeoTopics* model aims to describe urban activity in terms of a number of geographic "topics"—where each topic is defined as a probability distribution that describes a particular type of activity. Specifically, each topic models activity that takes place in the vicinity of a city region and is characterized by a certain distribution of activities, users who participate in them, and times when they do that. GeoTopics models are straightforward to train from urban activity datasets, such as Foursquare activity datasets.

GeoTopics models have at least two characteristics that make them desirable in many scenarios of urban analysis. First, they are easily interpretable, as each topic corresponds to the activity of well-defined features. Second, they are probabilistic, and so, they allow for the numerical quantification of various aspects (e.g., derive a probability-based similarity of two predefined regions in terms of activity they host). Making use of GeoTopics models trained on Foursquare activity data, we built[3] a synonymous interactive system [33], which allows users to explore and understand urban activity in tens of large cities around the world. The output of GeoTopics for the city of San Francisco is provided in Fig. 1.

2.2 Personal location detection

One of the key tasks in mobility data analysis (and a necessary preprocessing step for many applications) is detecting the locations of users. The objective is to identify the users' personal location, i.e., the areas where users perform their activities, based on the analysis of the locations (essentially, GPS points) that they have stopped, herein called stop observations. Examples of locations are home, workplace, supermarket, gym, fuel station, etc. More precisely, given a set of users GPS stop observations, i.e., coordinates in which the users have stopped, the *location detection problem* consists in grouping together the observations corresponding

[3] https://mmathioudakis.github.io/geotopics/.

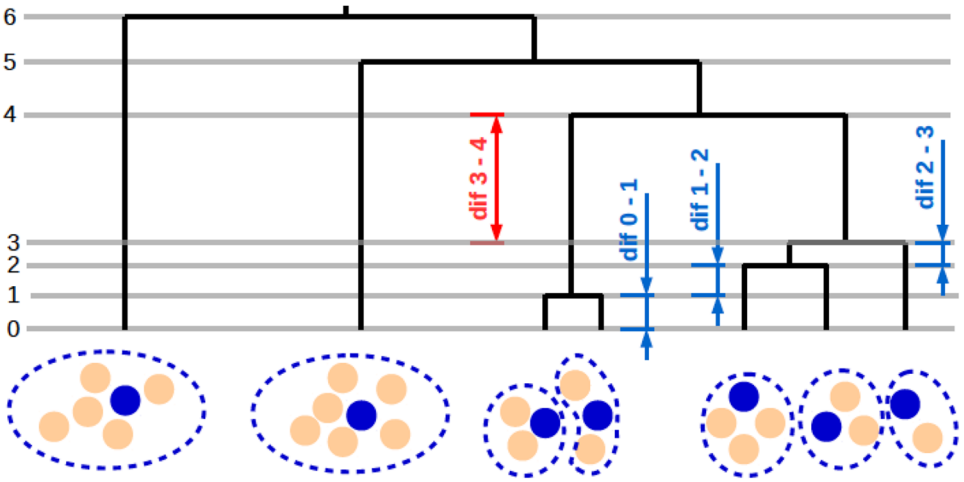


Fig. 2 TOSCA main steps. Orange points are the stop observations. Blue dotted circles correspond to X-means clusters and the blue points to their medoids, which are then processed by single linkage. On the resulting dendrogram, we highlight the differences among distances

to the same location. Correctly discovering such personal locations is an important problem with a wide range of applications. In the literature, this problem is typically addressed using a grid partitioning of the studied area or generic clustering algorithms like DBSCAN [43] or OPTICS [10].

However, this type of clustering methods shows various drawbacks. First, some of them are focused on specific optimization criteria, such as maximizing compactness or density connectivity, which does not always correspond exactly to the notion of locations, and therefore, the results, though optimal with respect to its own criteria, are not good locations. Second, in some cases, the algorithms need parameters that are not easy to guess (e.g., the size of the cell for the grid partitioning and the radius and minimum points for DBSCAN) and that should be tuned ad hoc for the data of each user analyzed. Indeed, in most cases an experienced analyst or some expensive self-tuning procedure might be needed to select accurately the parameters. On the other hand, in most cases such parameters are fixed for all users, while each individual might show specific features that require a treatment different from the others.

TOSCA (two-step clustering algorithm) [52] overcomes these drawbacks. TOSCA is a robust, efficient, statistically well-founded, and parameter-free algorithm explicitly shaped for personal location detection. The two steps of TOSCA are realized by combining two clustering methods and a statistical analysis approach. TOSCA enables in this way to produce high-quality clusters with a low computational cost. The idea behind TOSCA comes from the need to detect the locations of the users in an efficient way without sacrificing the clustering quality and, most importantly, without any tuning phase for the parameters.

Extensive experimentation showed that center-based clustering methods tend to correctly identify subgroups of observations that should belong to the same location. The side effect of such constraints is that the result usually splits real locations into several pieces that are connected with each other in a relatively loose way. On the other hand, single-linkage and density-based clustering methods are very good at spotting such loose connections, with the drawback of not distinguishing well those loose connections that are actually boundaries with other clusters. By exploiting these observations, the two main steps of TOSCA work as follows (see Fig. 2): (1) extract (sub-)clusters and corresponding medoids through center-based methods. X-means [83] algorithm was selected through empirical evaluations; and (2) cluster the medoids through a single-linkage hierarchical algorithm [106]. Stop the iterative cluster aggregation (or, equivalently, cut the dendrogram resulting from a complete run of the algorithm) through a statistically determined threshold on the increase in the distance between the clusters to be merged at each iteration. The cut criteria considered in TOSCA come from the outlier detection theory [16]. The distribution of the difference of the distances in the dendrogram returned by single linkage experimentally shows a sudden spike indicating the change in trend in the aggregation of the clusters.

It has been shown how, in contrast to algorithms commonly used in the literature, TOSCA automatically detects a good distance threshold for the clusters produced, thus adapting the clustering to the individual mobility behavior of each user in the data [52]. Therefore, it is perfectly suitable as *autofocus* clustering algorithm for analyzing individual mobility data. TOSCA evaluation against a large set of competitors over data generated from a null model and a mobility-like model shown that both in the mobility-like model and in the real case study TOSCA performs better than the general-purpose algorithms producing the desirable clustering for personal mobility data mining (see Fig. 3).

2.3 Simulating realistic mobility

The goal of generative algorithms of human mobility is to create a population of agents whose mobility patterns are

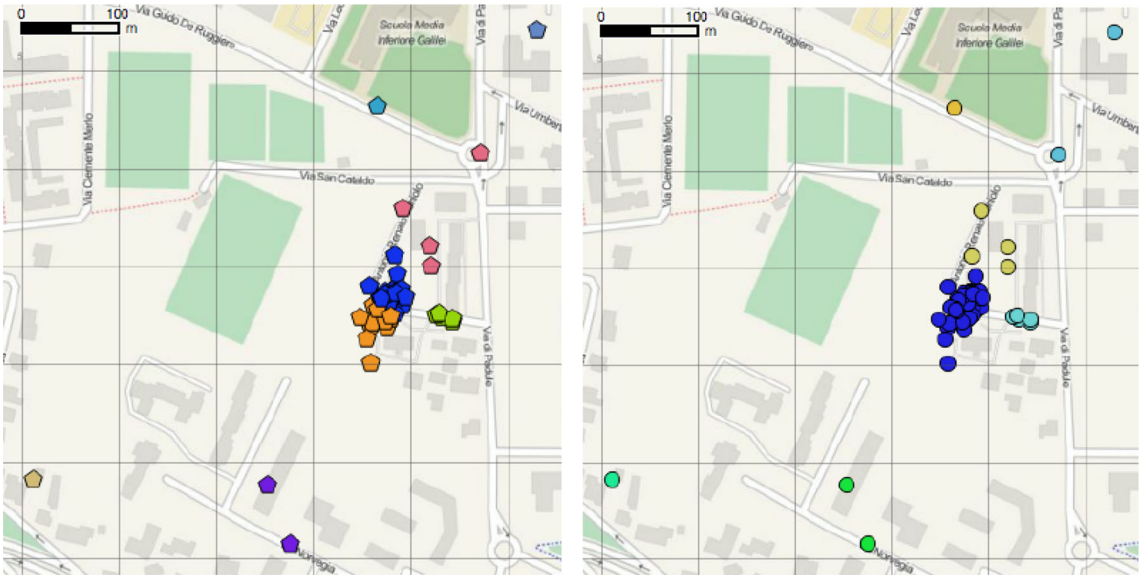


Fig. 3 Personal locations detected with X-means (left) and TOSCA (right). Different colors denote the different clusters (personal locations) to which each point is assigned

statistically indistinguishable from those of real individuals [48]. Typically each generative algorithm focuses on just a few properties of human mobility. A class of algorithms aims to realistically represent spatial properties such as the trip distance distribution [49] or the visitation frequency to a set of preferred locations [80]. Another class of algorithms focuses on accurately representing the time-varying behavior and the routine schedules of individuals [64,100]. However, the major challenge for generative algorithms lies in the creation of realistic temporal patterns and in the description of both routine and out-of-routine mobility patterns.

DITRAS (DIary-based TRAjectory Simulator) [79] is a framework to simulate the spatiotemporal patterns of human mobility (Fig. 4). DITRAS separates the temporal characteristics of human mobility from the spatial ones by operating in two steps. In the first step, DITRAS uses a diary generator to generate a mobility diary, i.e., an algorithm that captures the temporal patterns of human mobility by specifying the arrival time and the time spent in each location visited by the individual. A diary generator is an algorithm that generates a mobility diary for an individual given a diary length. Pappalardo and Simini [79] propose a data-driven algorithm called Mobility Diary Learner (MDL) which is able to infer from real mobility data a Markovian diary generator (MD) which captures the propensity of individuals to follow quasi-periodic daily schedules as well as to modify their mobility habits.

In the second step, DITRAS transforms the mobility diary into a trajectory by using a mechanism for the exploration of

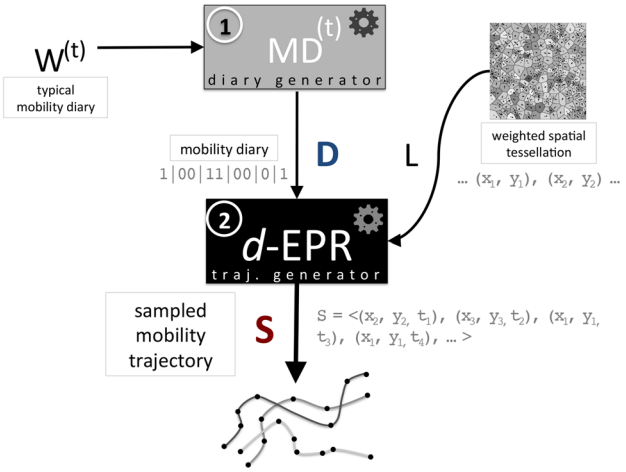


Fig. 4 Outline of the DITRAS framework. DITRAS combines two probabilistic models: a diary generator and trajectory generator. The diary generator uses a typical diary to produce a mobility diary D. The mobility diary D is the input of the trajectory generator together with a weighted spatial tessellation of the territory L. From D and L, the trajectory generator produces a sampled mobility trajectory S

locations on the mobility space. Pappalardo and Simini [79] suggest to use the d-EPR trajectory generator [80,81], which embeds mechanisms to explore new locations and return to already visited locations. The exploration phase takes into account both the distance between locations and their relevance on the mobility space, though taking into account the underlying urban structure and the distribution of population density.

The combination through DITRAS of the d-EPR trajectory generator and the MD diary generator constitutes the generative algorithm d-EPR$_{\text{MD}}$. A comparison with nationwide mobile phone data, region-wide GPS vehicular data and synthetic trajectories produced by other generative algorithms showed that d-EPR$_{\text{MD}}$ simulates the spatiotemporal properties of human mobility in a realistic manner, typically reproducing the mobility patterns of real individuals better than the other considered algorithms.

The DITRAS modeling framework goes toward a comprehensive approach that combines network science and data mining to improve the realism of human mobility models. Recently, to foster the combination of generative algorithms through the DITRAS framework as well as the comparison among existing generative algorithms, the Python library `scikit-mobility`[4]. has been developed [82].

2.4 Predicting mobility

A *prediction* (or forecast) is a statement about the way things will happen in the future, often, but not always, based on experience or knowledge. Although error-free prediction about the future is in most cases impossible, prediction is necessary to allow plans to be made about possible developments. Human predictability can be used to plan events and infrastructures, both for the public good and for private gain. Predictability is a vast research field, tackled with a number of approaches and for a number of different reasons. Mobility data mining is a field of research in which prediction is a fundamental task widely studied in the literature.

Powered by the increasing diffusion of location-based services, predicting the future locations of a mobile user is a flourishing research area. Knowledge of the position of mobile users can support applications that require access to this information in order to operate efficiently. Examples of such services are traffic management, navigational services, mobile phone control, etc. Many location-based services are based on the current or future locations of a user. By using the knowledge about locations, it is possible to fetch relevant information, such as nearby points of interest and available services. Moreover, predicting future positions can inform a driver about services like restaurants, banks, and shops that are present in future locations, or traffic problems that may occur along her route. The strong interest in this kind of applications led to the study of several approaches in the literature addressing the location prediction problem. Some of them base the prediction on single users' movement history, while others extract common behaviors from the histories of all the users in the system.

Indeed, the approaches proposed in the literature for location and trajectory prediction can be classified according to the prediction strategy used. The majority of the studies addressing the location prediction problem propose methods that base the prediction only on the movement history of the object itself. These approaches use an *individual* strategy for the prediction of user future positions. Other approaches in this category adopt time series analysis [32,102] to forecast user behavior in different locations. In this case, it is necessary to define the set of interesting locations to be considered in the analysis (see Sect. 2.2). The main problem of approaches implementing the individual strategy is that they fail in predicting future locations of non-systematic users. In these cases, applying a *collective* strategy could improve the prediction. Prediction approaches belonging to this category first extract mobility behavior for each user considering only the user's movement history, like in the individual strategy, and then, they merge all the individual models for the construction of the predictor [74].

2.4.1 Individual movement prediction

MyWay [112] is a system to forecast the exact future positions of a user, while she is moving. In line with the Personal Data Analytics paradigm [51], MyWay predictors exploit the *individual* systematic behaviors of a single user, the individual systematic behaviors of all the users in the system (called *collective* behavior), and an *hybrid* combination of them. To predict the future positions of a user, MyWay first uses her systematic behaviors and, if they are not sufficient, it exploits the systematic behavior of the crowd. This idea is based on the conviction that typically any user systematically visits a small set of locations and regularly moves between them by choosing the best movements learned by the daily experience [49].

MyWay requires that each individual computes an abstract representation of her systematic behavior: the *individual mobility profile* that captures the paths that are regularly followed by the user, called *routines* [111]. Thus, each routine is a representative trajectory (obtained through clustering methods applied to raw trajectory data) and the collection of the routines of each person constitutes her mobility profile. (An example of mobility profile composed of two routines is provided in Fig. 5).

MyWay predictions leverage the individual mobility profiles in the following way. The *individual* strategy predicts the future positions using only the routines part of the user's individual mobility profile. The *collective* strategy considers the routines of all users exploiting the possibility that a user could follow a path that is atypical for her, but systematic for another user. The *hybrid* strategy uses the collective strategy when the individual one fails. The MyWay framework with the three strategies is illustrated in Fig. 6, where, for each prediction flow, a different color line is used: individual history, the individual profile, and the individual predictor (red

[4] https://github.com/scikit-mobility/scikit-mobility

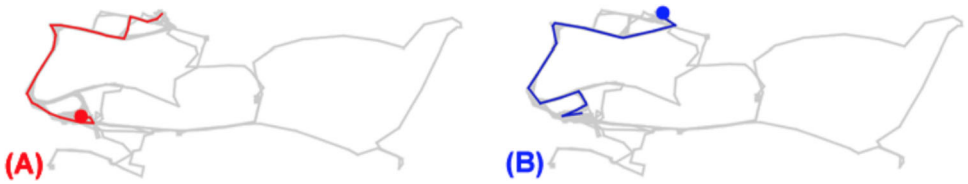


Fig. 5 Individual mobility profile formed by two routines A and B

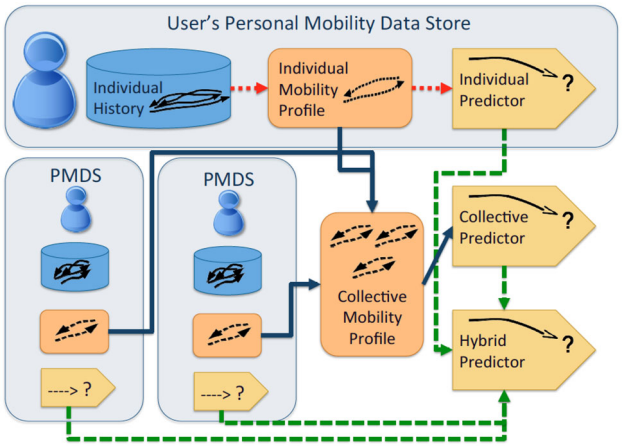


Fig. 6 MyWay prediction strategy system

lines) reside on the user local device, while the collective predictor (blue lines) is outside and therefore handled by a third party that orchestrates the users' information as well as the hybrid predictor (green lines). This third party, usually called coordinator, has the responsibility for the storage and management of the users' profiles. In the case of the hybrid strategy, the coordinator stores all the mobility profiles of the users (which are compact representations of their mobility) and receives the query for the prediction only in the case the individual predictor of a specific user fails.

As theorized by Guidotti et al. [51], MyWay exploits the possibility to use two levels of knowledge (individual and collective), obtaining advantages from the previous strategies. Moreover, in line with TOSCA (Sect. 2.2), MyWay does not apply any a priori spatial discretization. In fact, most of the works proposed so far in the literature apply a spatial discretization such a fixed grid on the space [78] or a territory tessellation obtained by clustering spatial points [63]. The spatial discretization often affects the precision of the prediction that instead of returning spatiotemporal points, it returns regions with higher granularity.

The prediction strategy that uses only individual mobility profiles is comparable with a prediction strategy based on raw movement data. There are some important advantages: (i) it dramatically minimizes the quantity of information required since a mobility profile is a concise representation of the information of the user; (ii) it can help to reduce the privacy risks: The mobility profile represents a systematic behavior, i.e., paths that are regularly followed by the user, but does not reveal all the details of her past spatiotemporal positions. An evaluation of MyWay on large real-world trajectory data show that the best prediction strategy is the hybrid one. Furthermore, a study of how the participation of the user affects the overall performances through an analysis of the prediction rate and the accuracy varying the percentage of users sharing their profiles shows how a greater sharing of routines enables better performances. Furthermore, the prediction rate dramatically increases at each step, while the accuracy slightly decreases. This happens because a larger number of trajectories become predictable allowing more errors, but the overall performances clearly improve.

2.4.2 Individual agenda prediction

The combination of procedures like TOSCA [52] and MyWay [112] can be used to further improve the prediction of human mobility. RAMA (Routinary Actions Mobility Agenda) is an approach that extracts the user's personal mobility model and uses it to reproduce the user's personal mobility agenda representing the predicted positions where the user accomplishes her activities during the whole day [53]. RAMA is completely unsupervised and adaptive to different users and mobility scenarios. In particular, RAMA combines the personal locations extracted by TOSCA with the routines and the technique of prediction of MyWay in the individual mobility network defined by Rinzivillo et al. [100].

RAMA first learns the *personal mobility model* by observing the personal mobility history and then exploits the model learned to reproduce future *personal mobility agendas*. In Figure 7, there is an example of mobility history (a) and personal mobility model (b). In summary, RAMA uses the personal mobility model together with the probabilities of staying in a location or of moving from a location to another one along a certain routine to predict the whole user mobility agenda at predefined time intervals. RAMA has shown to be very flexible and able to adapt to various city contexts obtaining comparable results over Rome, London, Boston, and Beijing. The agenda reproduction obtains a good performance in comparison with naive approaches from the state of the art.

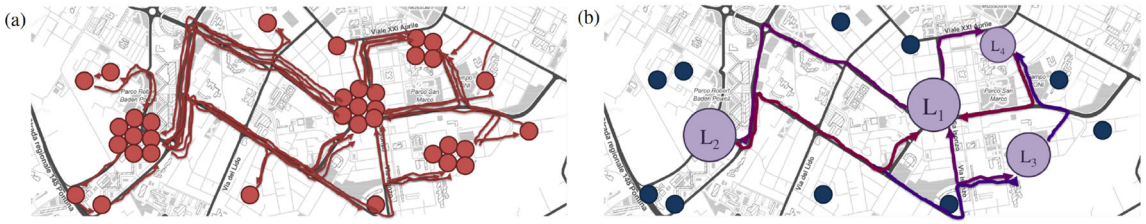


Fig. 7 Example of mobility history (left) and RAMA personal mobility model (right)

3 Visual analytics for urban data

Visual analytics is a research discipline that aims to support synergistic human–computer analytical workflows by combining computational analysis techniques with interactive visual interfaces supporting human interpretation, judgment, and reasoning [66]. Micro-blogging platforms, such as Twitter, and platforms for sharing photographs and videos, such as Flickr and Instagram, allow the users for the annotation of their content with geographic coordinates. The high popularity of these services in conjunction with the widespread proliferation of devices capable of providing location information has led to great and constantly increasing volumes of location- and time-referenced data produced by myriads of users [38]. By analyzing these data, it is possible to extract interesting new information about various places and events as well as about people's interests, mobility behaviors, and lifestyles. For these reasons, the analysis of social media data is currently a popular topic in visual analytics.

In the following, we provide the reader with a brief literature overview of the published literature, with an emphasis on works published within SoBigData describing visual analytics approaches to extract different kinds of information from georeferenced social media data. Most of the works do not focus on extracting a single type of information, but deal with several types.

3.1 Analysis of georeferenced photograph data

Photographs published on Flickr, Panoramio, and other photograph sharing platforms are supplied with metadata, which include the dates and times of the shots and may also include titles and text tags indicating the content of the photographs. For many photographs, the metadata include the coordinates of the locations where the photographs had been taken. Collections of metadata records including geographic coordinates have been analyzed in multiple ways according to the possible analysis foci (space and place or people) and respective tasks [3]. The photograph datasets have been considered from two distinct perspectives: as spatial events (independent points in space and time) and as trajectories of people (i.e., the photograph owners). We discuss these two perspectives below.

Event-based analysis of photographs In analyzing the data as spatial events, spatial density-based clustering has been used for identifying popular places, which have attracted the attention of the photograph owners. Visualization of the times when the photographs had been taken in these places has revealed different seasonal patterns of the visited places. To study the spatial distribution of the photographs over a territory and compare the temporal patterns of visiting different parts of it, the territory is divided into compartments, e.g., by a regular [3] or irregular [5,62] grid, and the photograph-taking events are aggregated by these compartments and by time intervals. The resulting time series of the event counts are visualized on a map [4] or on a time graph [5,62], which is linked to a map display through interactive techniques, including synchronous highlighting, selection, and filtering of corresponding visual objects. By analyzing the time series using either mostly interactive [62] or computationally supported [5] techniques, the researchers detected places with interesting temporal patterns of visits, such as periodic peaks at particular times of the year, very high irregularly occurring peaks, and significant increase in place popularity starting from a particular time. To support a better understanding of these patterns, the visualization is supported with tools for extracting frequently occurring words and word combinations from the titles of the photographs that had been taken in the places and times of the peaks or sudden increases in attendance. In most cases, the extracted words refer to various public events (festivals, open-air shows, and concerts), but also to interesting natural phenomena, such as cherry tree blossoming or abundant snowfalls. A different approach to identify public events and other happenings attracting people's attention is by using spatiotemporal clustering of the photograph-taking events [6, section 6.2.3], where the authors find occurrences of multiple photographs taken closely in space and time, i.e., spatiotemporal clusters. For the clusters, frequently occurring words and word combinations are extracted and investigated using a text cloud display linked to a map, as shown in Figure 8.

An example of an in-depth analysis of the time series resulting from the presence of distinct photographers in regions of Switzerland is presented in the book of Andrienko et al. [6, Sections 7.2.1-7.2.5]. The analysis includes, among other techniques, visually supported clustering of the time

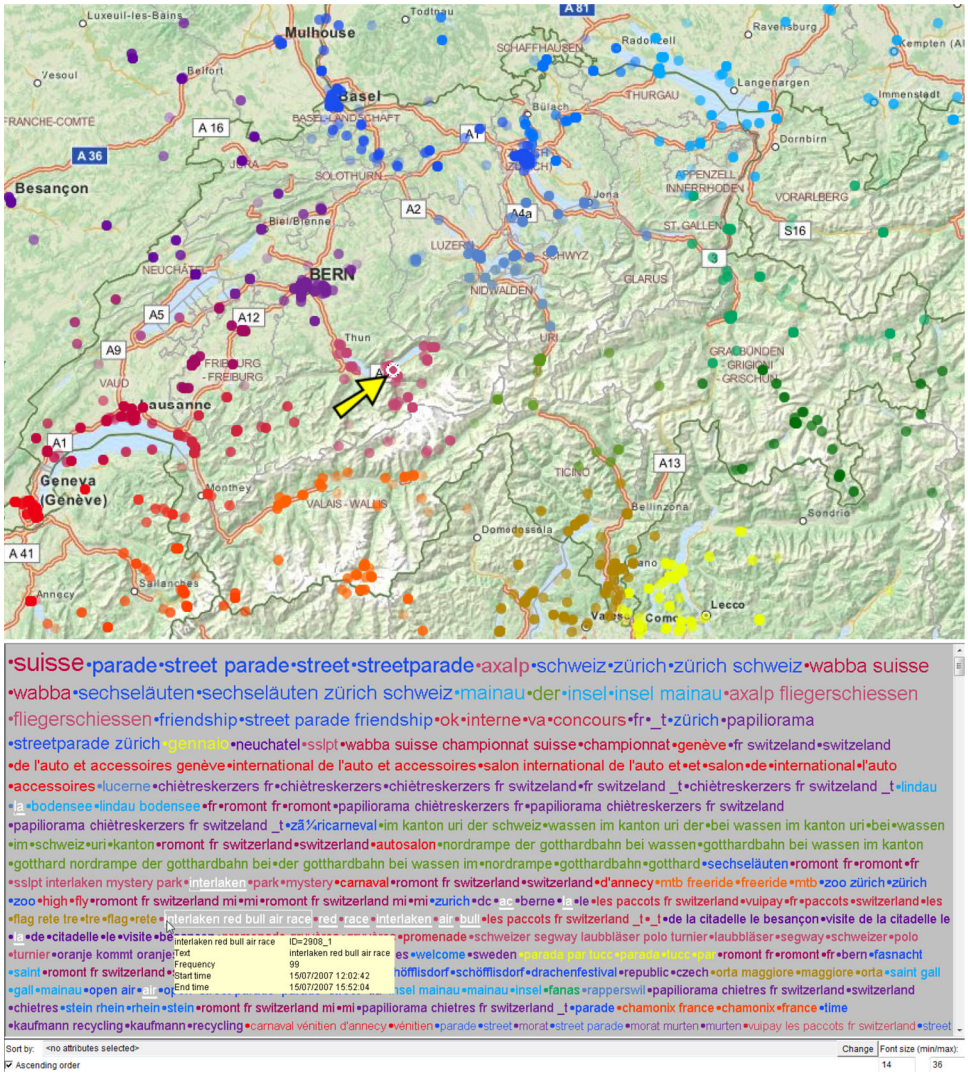


Fig. 8 Top: the frequent occurrences of words and combinations in the photograph titles within spatiotemporal clusters of Flickr photographs are represented on a map by point symbols colored according to the spatial positions of the clusters. Bottom: the words and combinations are represented in a text cloud display, the font sizes being proportional to the frequencies, and the colors corresponding to the spatial locations, as in the map. One of the word combinations ("Interlaken red bull air race") is selected in the text cloud view by mouse pointing; the corresponding point is highlighted on the map (marked with an arrow)

series and interactive generation of models for predicting the number of photographers who are expected to visit the regions in the future at different times of the year. The time series can also be viewed from a different perspective: as a sequence of spatial distributions of the photographers' presence at different time intervals. To study the temporal patterns of the occurrence of similar and dissimilar spatial distribution patterns, the distributions are clustered by similarity, summarized by the resulting clusters, and compared using multiple map displays and special interactive operations supporting comparisons [6, Section 8.1.1]. The temporal distribution of the clusters is visually represented on temporal displays. The example demonstrates how the analysis reveals an interaction between temporal periodicity and temporal trends in the sequence of the spatial distributions of the presence of Flickr photographers over the territory of Switzerland.

Trajectory-based analysis of photographs Trajectories of people can be constructed from georeferenced photograph data by arranging the records of each individual photographer in a chronological sequence—the same idea applies to any kind of georeferenced data that include identifiers of individuals, in particular, to data from YouTube, Twitter, Instagram, and other social media. Trajectories of individuals can be aggregated into flows between compartments of a territory division and visualized on flow maps to enable studying of mass movement patterns [3]. The aggregation of trajectories into flows can be done by time intervals for studying seasonal differences between the mass movement patterns [62]. A set of trajectories can also be analyzed for discovering frequent sequences of place visits [6, Section 7.3.4]. The extracted frequent sequences can be explored using a text cloud display combined with an interactive map and a space–time cube. By analyzing people's trajectories, one can also detect meetings of two or more individuals, including repeated meetings of the same pairs or groups of individuals, and joint trips of two or more photographers [3]; however, performing such analysis may be unethical, as it may compromise the privacy of the individuals.

This overview gives an idea about the diversity of possible approaches to analyzing georeferenced photograph data and the kinds of information and knowledge that can be extracted from such data. The same range of approaches is also applicable to georeferenced micro-blogging data, such as data from Twitter. The types of information that can be extracted from the two different sources of data are the same, but the interpretation may be different. Thus, people mostly take photographs when they encounter interesting places, objects, or events; besides, not all taken photographs but only the best or the most interesting ones may be published. It should also be taken into account that photographs are rarely taken in low-light conditions, and that there are situations and places in which taking photographs is prohibited. Therefore, the photograph data cannot be considered representative of people's presence and movements over a territory and of people's everyday activities. Figure 8 shows that photograph data may reflect people's leisure activities and touristic travels. However, it would be wrong to assume that this is always the case. The possible relation of the published photographs to the author's leisure time, travels, or professional activities can be judged from the temporal frequency and regularity of the photographs and from their spatial distribution.

3.2 Analysis of georeferenced micro-blog data

Posting micro-blog messages from mobile devices may occur more frequently and spontaneously and in a wider range of places and situations than taking and publishing photographs. Besides, there is no time gap between producing and publishing a message, while photograph authors may not publish their photographs immediately after taking them, but may do this after some (often quite long) time. Therefore, unlike photographs, micro-blog data are suitable for real-time analyses, which may discover information about currently happening events, in particular, abnormal and disastrous events, such as earthquakes, floods, storms, and even terrorist attacks [7,12,14,35]. All these activities require the processing of the message text.

One first and crucial issue in processing textual messages is related to the scarcity of explicit and native geolocations. While multimedia posts are often complemented with geotags, the same only occurs in 1% to 4% of text-only micro-blog posts [12,36]. To mitigate this issue and to enable geospatial textual analyses of micro-blogs, a wide array of *geoparsing* techniques have been proposed [15]. The high-level goal of geoparsing is that of enriching any given piece of text with the geographic coordinates of places and locations mentioned within the text itself. In this way, even if the original message did not explicitly carry geotags, it can still be placed on a map by leveraging the geotags discovered during the geoparsing operation. Traditional approaches to geoparsing involved natural language processing of the text

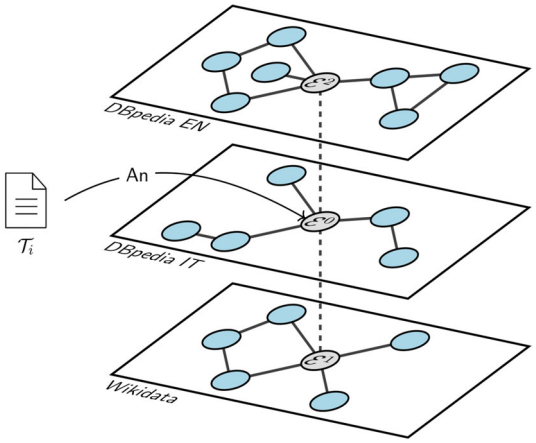


Fig. 9 GSP geoparsing technique links a textual document $\mathcal{T}_i$ to an entity within a knowledge base, via a semantic annotation process. Then, it exploits the Linked Open Data knowledge graph to find additional candidate entities from which to extract geographic information about the input text

to identify geographic named entities that were subsequently matched against gazetteers (i.e., databases containing associations between toponyms and their geographic coordinates) such as Geonames or OpenStreetMap [77]. More recent and more accurate approaches developed within SoBigData are instead based on a combination of machine learning applied to the rich and structured information contained in Linked Open Data. In particular, the geosemantic-parsing (GSP) technique [15] first applies semantic annotation to highlight relevant portions (i.e., tokens) of the input text and to link them to pertinent entities in one or more knowledge bases, such as DBpedia and Wikidata. Then, the algorithm traverses the knowledge graph in order to find a set of candidate entities from which to extract geographic information, as shown in Figure 9. In a subsequent step, all candidates are evaluated and possibly pruned. Finally, all entities that have not been pruned contribute to determining the geographic coordinates of the text, by majority voting [15]. The adoption of geoparsing techniques, such as GSP, has been shown to increase the number of geolocated messages from less than 5% up to 50% [12] across several benchmark datasets.[5]

Given a set of geolocated messages, further analyses can be applied to extract meaningful knowledge on interesting events. One of the possible approaches is prefiltering of the messages for selecting only those that contain analysis-relevant keywords, such as terms denoting extreme weather conditions [7,12,14]. Another approach is extracting significant terms, i.e., such words that do not occur frequently in micro-blog messages in general or in the times (seasons) or places where they have occurred [25,35]. Each occurrence

[5] http://data.d4science.org/ctlg/ResourceCatalogue/geo-annotated_tweets_eng-ita.

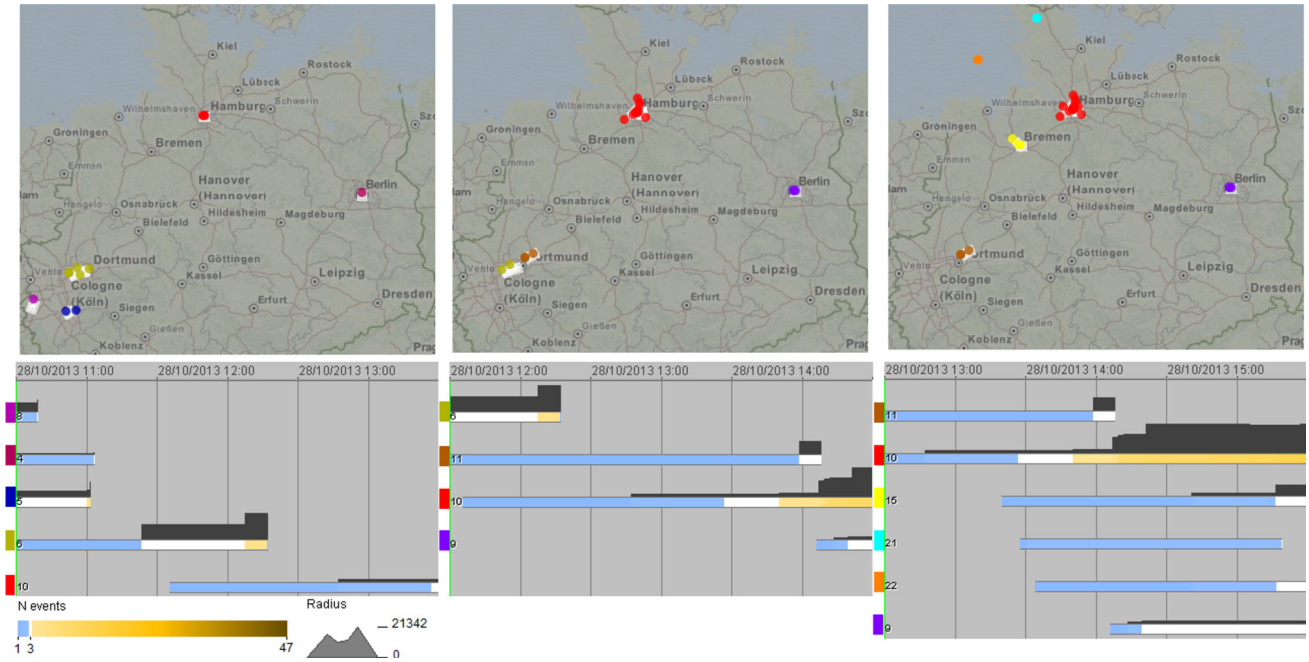


Fig. 10 Emergence and evolution of spatiotemporal clusters of georeferenced tweets related to a hurricane on October 28, 2013

of a significant term is treated as a separate spatial event. A spatiotemporal concentrations (clusters) of events with the same term may indicate that something is happening in this place and time, and the term gives an idea of what may be happening. The significant terms from such spatiotemporal clusters are shown on a map display using the text cloud technique, with the font size being proportional to the number of the term occurrences. The map is constantly updated in real time as new messages appear [14]. By means of an interactive tool called Content Lens, the user can select a particular area and explore in more detail the term occurrences in this area. To increase the relevance of the information that is shown to the user, various user-constructed filters can be applied to the data [25]. In the other approach [7], the message texts are only used for the selection of potentially relevant messages and not used in further analysis. The work focuses on real-time detection of spatiotemporal clusters of relevant events, taking into account only the event locations and times, but not the texts, and on tracing the cluster evolution (growing, shrinking, moving, merging, and splitting) over time (Figure 10). The individual events making the clusters and their message texts can be accessed on demand. Yet another approach focuses on identifying geographically relevant areas in the aftermath of an event (e.g., for detecting those areas mostly stricken by a disaster) [13]. This time, micro-blog texts are analyzed in order to detect mentions of known locations. Then, the mentioned time series of each location are computed and compared to reference values or baselines. Relevant locations are those for which, at a given point in time, the related time series are significantly greater than the references. Finally, each location is graphically highlighted by drawing choropleth maps—that is, by using different shades of colors for different locations in a geographic map, depending on the relevance of the locations [12,14].

An example of an off-line investigation of micro-blog posts related to a disastrous event (an epidemic) is presented by Andrienko et al. [6, Section 6.3.2]. Although it uses synthetically generated data (which, however, are generated by a model learned from real data), it shows the principal possibility of using micro-blog data for identifying the origin and possible cause of an epidemic, the ways of disease propagation, the spatial spread, and the evolution over time. Other studies [12,14,77] are instead based on real-world Twitter data related to disastrous earthquakes and floods. Figure 11 shows the results of such studies that allowed to promptly identify mostly stricken areas in order to guide first responders.

However, detecting and investigating disastrous or abnormal happenings are not the only possible use case for micro-blog data. Georeferenced micro-blog posts, at least those from active bloggers, may to some extent be considered as representative of the people's daily lives and used for studying people's behaviors. Thus, an analysis of a collection of tweets posted by residents of the Seattle area (USA) revealed interesting patterns of collective and individual behaviors [109]. For this analysis, the tweets were classified according to their topics, such as family, work, education, food, and sports, based on the occurrences of topic-specific keywords. (For example, the topic "family" is

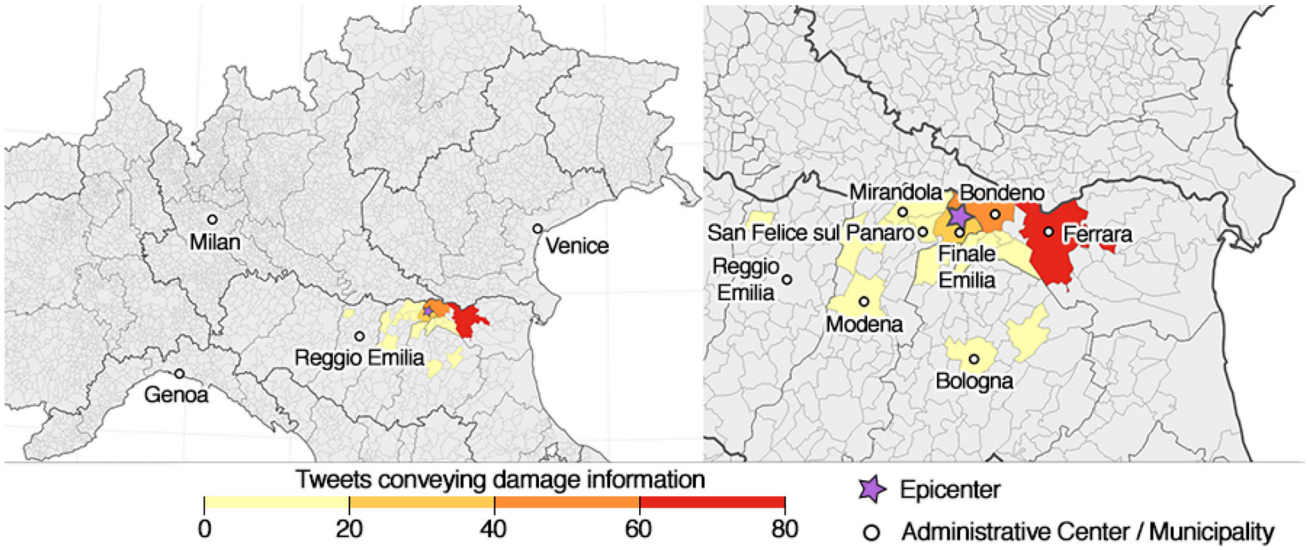


Fig. 11 Real-time geospatial analyses of social media allow to promptly obtain accurate maps of stricken locations in the aftermath of mass emergencies

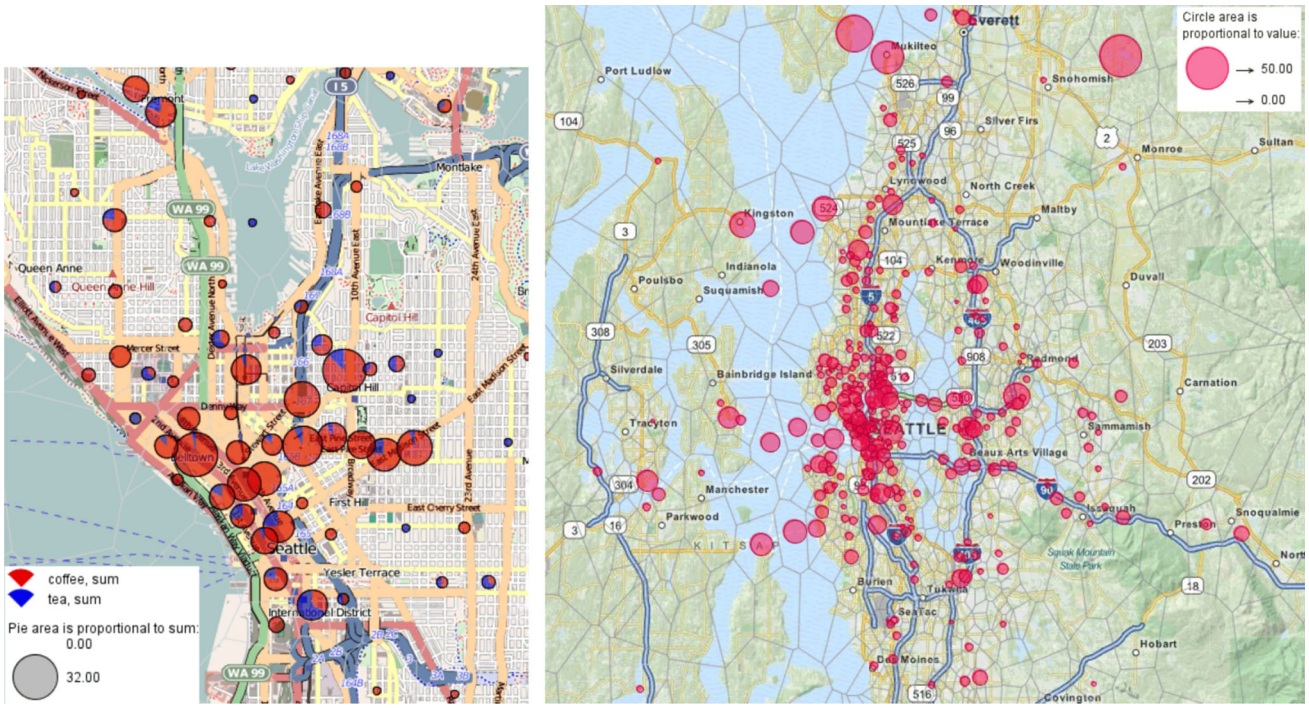


Fig. 12 Left: the spatial distribution of the tweet topics "coffee" and "tea" in the central area of Seattle. Right: the spatial distribution of the topic "transportation"

associated with the terms denoting family members: mother, mom, father, daddy, and so on.) The researchers explored how much the tweet topics are related to the locations from which the tweets were posted and to the times when this happened. For this purpose, they aggregated the tweets by the topics, areas in space, and time intervals and visually explored the results using maps and time histograms. It was found that there are areas where particular topics prevail, which may be related to the kinds of objects or facilities located in the areas (e.g., a university or a stadium) or to the characteristics of the population (e.g., an international district; see Figure 12, left panel). The researchers also looked at the spatial distributions of the different topics and found that some of them are correlated with the distribution of certain kinds of objects or facilities. Thus, the topic "transportation" occurs along the main transportation corridors (Figure

12, right panel). Regarding the temporal distributions of the tweet topics, the researchers found several very interesting patterns of when certain topics occupy the peoples' minds. Thus, "food" occurs more frequently during lunch and dinner times, "coffee" during/after breakfast and over the forenoon, "transportation" during working day rush hours, and "sports" and "alcohol" in the evenings and over the weekend.

Although the study shows that the contents of some micro-blog posts are related to the places the authors visit and/or the activities they perform, these data in general contain a large proportion of noise, which includes texts with unidentifiable topics and texts with topics that are not relevant to the places of message posting. (Thus, a person may tweet about work while being at home or about food while traveling in public transport.) In fact, the proportion of noise outweighs the proportion of potentially relevant data. Therefore, it makes sense to analyze the topic distribution in space and time at the level of a large population of micro-bloggers, to have a sufficiently large amount of potentially relevant data and to be able to use valid statistical summaries. At the level of individuals, the message texts can hardly be indicative of the individuals' activities or purposes for visiting different places.

In analyzing mobility behaviors of individuals, it is reasonable to look not at the message texts, but at the temporal patterns of visiting different places [8]. Significant (repeatedly visited) personal places are extracted from the collection of posts of each individual by spatial clustering of the post locations. Place semantics (i.e., the meanings, purposes for visiting, or activities performed in the places) can be determined based on the times over the weekly cycle when the individuals were present in the places. Thus, a place where a person is present in the evenings and nights of all days can be identified as the person's home place. However, separate consideration of the data of each individual is unfeasible and harmful for personal privacy. Andrienko et al. [8] proposes a privacy-respecting approach, in which data of a large number of Twitter users are analyzed all together using a combination of computational techniques and visualizations presenting the data and analysis result in aggregated form. After extracting personal places and identifying their meanings in this manner, the original georeferenced data are transformed into trajectories in an abstract semantic space. The semantically abstracted data can be further analyzed without the risk of re-identifying people based on the specific places they attend. The paper presents an example of analyzing mobility behaviors of Twitter users in the area of San Diego, California, USA.

To summarize, georeferenced data from social media can be analyzed as spatial events (i.e., independent points in space and time) and as trajectories of people. To analyze such data, visual analytics proposes a number of approaches combining computational techniques (clustering, aggregation, statistical summarization, pattern detection, etc.) with interactive visualizations. By means of these approaches, it is possible to extract interesting information and gain new knowledge about places, events, and people's interests, behaviors, and habits. Metadata of the photographs published through photograph sharing services can reveal people's interests in tourist attractions, public events, and other happenings, or natural phenomena and patterns of touristic behavior. Georeferenced micro-blog posts can be analyzed in real time for early detection of abnormal or disastrous events. It may also be useful to analyze the evolution of such events by looking at the spatiotemporal distribution of the event-related posts. Besides the information concerning unusual happenings, micro-blog data may be a source of knowledge about everyday mobility and activities of people. As both the popularity of social media and the interest in analyzing social media data are growing, we can expect the appearance of new analysis methods and new use cases for information that can be extracted by these methods.

4 Shaping the urban landscape: applications and services

In order to showcase the potential of applying data science techniques to improve the quality of life of people engaging with the city, as well as to make cities more sustainable, in this section we discuss three challenging applications domains of urban data science: tourism (Sect. 4.1), smart transportation (Sect. 4.2), and sustainable urban development (Sect. 4.3).

4.1 Sightseeing tour recommendations

When visiting a new city, it is difficult for tourists to decide which attractions match their preferences better [39]. Likewise, it is also difficult for city stakeholders and city administrators to advertize the right attractions to the right people. The TRIPBUILDER service tries to fill this gap.

TRIPBUILDER, an application available in the SoBigData infrastructure, offers a service that can help tourists organize sightseeing tours within a city, given a set of temporal constraints and preferences, e.g., building a 6-h sightseeing tour where museums and churches are preferred to local architecture or entertainment. TRIPBUILDER is an unsupervised system for building personalized sightseeing tours. Given the target destination, the time available for the visit, and the user's profile, the system recommends a time-budgeted tour that maximizes user's interests and takes into account both the time needed to enjoy the attractions and the time to move from one point of interest to the next one. Moreover, the knowledge base feeding TRIPBUILDER recommendation model is entirely and automatically extracted from publicly available Web services, namely Wikipedia,

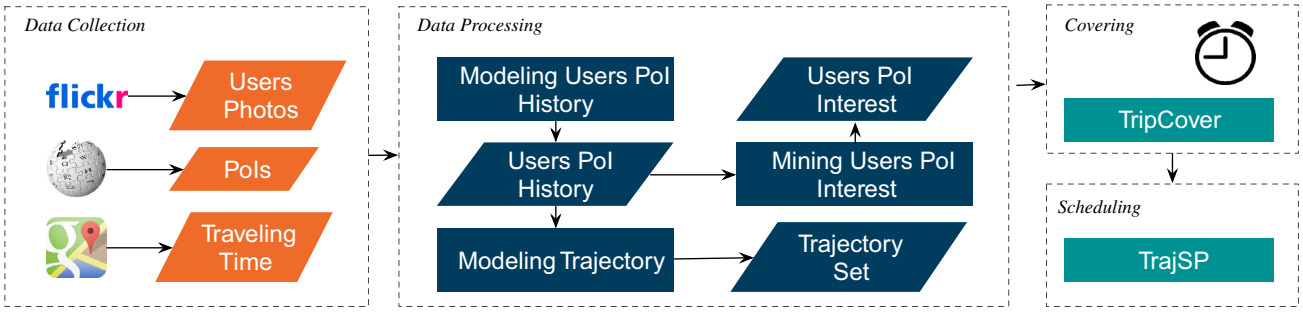


Fig. 13 Overview of the unsupervised process used to build the TRIPBUILDER knowledge base and service

Flickr and Google maps. Each photograph comes with useful information such as tags, comments, and likes from Flickr social network, number of views, information about the user, time stamp, GPS coordinates of the place where the photograph was taken. This allows us to roughly reconstruct the movements of users and their interests by analyzing the time-ordered sequence of their photographs.

The problem of planning the visit to the city is a two-step process. First, given the profile of the user and the amount of time available for the visit, the TRIPCOVER problem is addressed: choosing the set of itineraries across the points of interest that best fits user interest and respects the given time constraint. Second, the selected itineraries are joined in a sightseeing itinerary by means of a heuristic algorithm addressing the trajectory scheduling problem (TRAJSP), a particular instance of traveling salesman problem (TSP). The formalization of TRIPCOVER as an instance of the generalized maximum coverage (GMC) problem can be found in an earlier work [26,28], while a subsequent study [27] demonstrates the capabilities of the TRIPBUILDER application.[6]

Figure 13 depicts an overview of the TRIPBUILDER architecture. The component related to "data collection" retrieves relevant data from Flickr, Wikipedia, and Google Maps. The second component called "data processing" extracts the knowledge used to devise relevant points of interest and model users' visiting behaviors from data provided by the "data collection" component. Given a budget B, the third component "covering" deals with the exploitation of the models and the knowledge base to compute the solution to the TRIPCOVER(B) problem. The result is a set of trajectories in the chosen city on the basis of user interests and time budget that are finally scheduled on the user agenda by the fourth component "scheduling."

The TRIPBUILDER knowledge base, generated in an unsupervised way, covered initially three Italian cities: Pisa, Florence, and Rome, important from a sightseeing point of view and which guarantee variety and diversity in terms of size and richness of public user-generated content available for download: Rome, Florence, and Pisa. Obviously, the methodology can be applied to other cities as well.

Table 1 Performance of TRIPBUILDER (TB) on the Pisa dataset by varying the parameter α and the baselines (Tpop, Tppro) according to various metrics

	Days	Recall-P	Recall-C	S_u^{pro}	S^{vt}
Pisa					
Tpop	1/2	0.480	0.755	0.298	14443
	1	0.833	0.990	0.609	28984
Tppro	1/2	0.560	0.803	0.391	14535
	1	0.797	0.962	0.618	28272
TB,(0)	1/2	0.712	0.910	0.391	16086
	1	0.822	0.988	0.601	28968
TB,(0.5)	1/2	0.725	0.904	0.565	16027
	1	0.863	0.984	0.709	29452
TB,(1)	1/2	0.721	0.898	0.570	15931
	1	0.871	0.984	0.715	29510

The column *Days* denotes the itinerary length

The effectiveness of TRIPBUILDER is assessed by: (i) selecting a set of trajectories of interest for a given user (TripCover) and (ii) scheduling that set on the user agenda (TrajSP). The performance is compared to those obtained by competitive baselines: one (Tpop) that considers the trajectory popularity and one (Tppro) that relies on a normalized user/POI similarity score, using evaluation metrics that consider the actual behavior of test users as mined from Flickr, as explained in detail by Brilhante et al. [28].

The results on the Pisa dataset are reported in Table 1. (Similar results have been obtained for Florence and Rome.) The TRIPBUILDER approach aims at maximizing the user's total profit/interest over the PoIs fitting her budget. In terms of Personal Profit Score S_u^{pro}, a measure of the relevance with respect to the user preferences, the solution improves over the baselines with up to 91% in Pisa (up to 173% in Florence and 130% in Rome, not shown in the table). In addition, it builds trips that increase Visiting Time Score S^{vt} (i.e., the actual time spent enjoying attractions and not traveling to reach one)

[6] http://tripbuilder.isti.cnr.it/.

up to 25 min in Pisa (about 4 h in Florence and approximately 11 h in Rome). Therefore, it suggests itineraries that better match user preferences and involve lower intra-POI movement time than the baselines. In terms of PoIs and categories recall (Recall-P, for PoIs, and Recall-C, for Categories, in Table 1), all algorithms get at least 75% of the relevant PoIs and 96% of the categories for Pisa. Looking at PoIs recall, on the other hand, TRIPBUILDER gets better results than the baselines: 87% compared to 83% of Tpop and 79% of Tppro for the one-day time budget. The proposed solution outperforms the baselines in terms of all the metrics adopted for assessment, by suggesting itineraries that better match user preferences. Such itineraries present higher visiting time and, consequently, lower intra-point of interest movement time than the baselines, meaning users spend the budget in actually visiting points of interest rather than in transit.

4.2 Data-driven urban transportation: the car sharing case

Soon after their invention, cars have boosted people's personal mobility, but at the price of environmental pollution, city congestion, and huge public health issues (such as air pollution-related diseases or the stress associated with traffic and long commutes). An increasing awareness by policy makers and citizens alike has brought traditional automobile transportation at a turning point, and at the center of this personal mobility revolution are the concepts of data-driven smart transportation, sharing economic, and green vehicles: Shared vehicles with small carbon footprints whose usage is optimized by data-fueled approaches may be the solution for the mobility of the future. In this context, car sharing is emerging as one of the most promising examples of Mobility as a Service [104]. The members of a car sharing system can pick up a shared vehicle of the car sharing fleet when they need it. Different operators may implement different pickup/drop-off policies [22]. Here, we focus on free-floating car sharing—such as Car2go, DriveNow, Enjoy—whose customers can pick up and drop off vehicles anywhere within a predefined service area.

Urban data science is at its heart the science of detecting and putting to good use the many signals that stratify into an urban landscape. Car sharing remains a *weak signal*, though: the fraction of people using car sharing for their daily trips is rapidly increasing, but it is still in the order of single-digit percentage points in the best cases [73]. The standard approach for studying car sharing is still relying on surveys and direct interviews with car sharing members [103,104]. In many cases, car sharing is not even included in households travel diaries periodically collected by city councils. However, the digital upgrade of cities thanks to the cyber-physical convergence of urban infrastructure and ICT means that we can now know exactly when and where cars are available, and we can observe shared vehicle flows *as they happen* in the city. This knowledge opens up a new avenue of research that goes in the direction of the new science of cities [17] and urban computing [116]: using data and electronic devices to extract knowledge and to improve urban solutions.

In this section, we showcase how urban computing ideas can be applied to the car sharing domain. To this aim, we rely on a dataset comprising pickup and drop-off times of vehicles in 10 European cities for one of the major free-floating car sharing operators [23]. For nine of these cities, data have been collected between May 17, 2015 and June 30, 2015. For the remaining one, data cover the period from March 11, 2016, to May 12, 2016. The data have been collected every 1 min using the available public API.

These data can be explored in a variety of directions. As an example, in [23] they are analyzed in relation to geo-referenced socio-demographic and urban fabric indicators, coming from official institutes for statistics and Foursquare, respectively. The outcome of this analysis is that, while a single explanatory pattern does not emerge across the cities, they share indeed several similarities. In general, the car sharing demand is positively associated with high educational attainment and negatively correlated with commuting outside of the municipality area. These findings confirm the conclusion of the most recent socio-demographic surveys about car sharing services, but at a much finer spatial granularity and without relying on expensive and time-consuming interviews/questionnaires. With regard to the urban fabric indicators, the only activity category that seems to have a statistically significant effect on car sharing demand is that of nightlife-related activities, suggesting that leisure is the most typical trip purpose.

In the rest of the section, we will focus our attention on the problem of vehicle relocation. It is a well-known issue in car sharing that there are often empty stations in an area of high demand and at the same time stations with several cars in areas of low demand. The solution to this unbalance problem is to move cars from one area to the other one, but the *redistribution* is costly for the car sharing operator (personnel costs plus the costs of these "rides without customers"); thus, it has to be optimized as much as possible [21]. To this aim, being able to characterize the demand is crucial. We leverage the above datasets in order to illustrate how to predict future demand at specific geographic locations.

Features: From our dataset, we extract the following features for prediction: number of events $e_{(i,d,t)}$ observed in cell i at time t of day d, the time of the day (corresponding to bin t), the day of the week (Sunday, Monday, etc.), whether the day is a weekday or not, and the average number of events $\hat{e}_{(i,d,t)}$ observed at bin t of day d in the neighboring cells (we consider 2-hop neighbors only).

Methods: We use the first 80% of the days in the dataset for training, and we predict the remaining 20%. We set the

time window T to 1 h, implying that we want to forecast pickups and drop-offs happening in a 1-h time frame. We only consider cells that have more than 30 events during the observation period. Then, we run the prediction algorithms and we measure the prediction error in terms of root mean squared error (RMSE).

We now define a set of relevant prediction techniques to be evaluated on the datasets at hand. In the following, we use the general term *event* to denote either pickups or drop-offs. HA and HM are two simple prediction functions returning the average and the median, respectively, number of events observed in the same time window across different days. As car sharing typically exhibits marked differences between weekdays and weekends [22], we also test a version of the algorithms (denoted as HA+ and HM+) that distinguish between working days and weekends. ARIMA is the standard autoregressive integrated moving average technique, popular among time series forecasting methods. Then, we also consider random forest (RF) and a neural network (NN) composed of a single-layer perceptron with as many neurons in the input layer as the features described above and one hidden layer. For completeness, we also test the custom algorithm proposed by Weikl and Bogenberger [114] (WEIKL), whose rationale is to represent each timeslot of each day through a vector, whose components are the number of events at each cell during the timeslot.

Results show that for most cities, the error is small, with forecasts off, on average, by less than one drop-off/pickup for the vast majority of cells. However, there are a few cells for which the prediction error is high. These cells are typically near the airport, and both the high volume of traffic observed at the airport and the bustier nature of arrivals and departures there may explain this variability.

The fact that future car sharing requests can be predicted quite accurately using state-of-the-art prediction algorithms is decisively good news for car sharing: It shows that vehicle redistribution could generally be performed very efficiently and this is crucial for improving the reliability of the service (i.e., maximizing the chances that users find shared cars when and where they need them) and, as a consequence, customer satisfaction regarding urban transportation.

4.3 Urban sustainability and net negative cities

As discussed earlier, the twenty-first century is the century of the city. Since 2007, more than half of the global population lives in cities and this figure is expected to grow up to 60% by 2050. In developing countries, which have experienced the most growth during the last two decades, urbanization is growing fast and is leading to the formation of huge urban agglomerations (UN, 2014). At the pinnacle of urbanization, megacities, i.e., urban agglomerations with more than 10 million people, are a perfect example of the urban growth that our society is experiencing over the last decades. In a global scale, the number of megacities was 7 at the beginning of 1960, grew up to 27 in 2010, and in 2020 the number is expected to be over 37 [69]. Due to their size and complexity, megacities tend to concentrate and amplify drawbacks of urbanization-like inequalities (e.g., slum formation and unequal distribution of income), environmental pollution, greenhouse gas emissions, and unequal use of resources. On the other hand, they can provide a test bed for developing best practices and good examples of sustainability solutions from which many can learn. Understanding the drivers of energy and material flows in megacities is of paramount importance for addressing topics, such as global environmental stress, efficiency in resource use, and resource competition. To this aim, urban metabolism studies have become crucial for understanding urban sustainability [45,70] and for identifying the actions needed to address urban sustainability worldwide.

The structural and functional organization of urban systems is a classic example of a multi-scale system of systems [101], in which new connections are established and new behaviors emerge. On a smaller scale, since their first emergence about 10,000 years ago, cities have always played an important role in concentrating goods, minds, and social relationships. It can be argued that cities are an emergent phenomenon made of people who build social relationships on a larger scale, and that the development and growth of urban systems are directly related to the richness and the quality of relationships. The concentration of minds represents a formidable engine for technological, cultural, and social innovation, and cities are places where new behaviors, cultures, economies, and technologies emerge. Thus, cities are not only growing in size, but also in complexity, with more and more layers of interaction between their inhabitants and various actors as a consequence of the shifts from flows of energy to flows of information.

Each city, each local economic system, produces services, goods, and cultures, playing a complex role in the general dynamics of global sustainability, which cannot be described simply by a numerical value. The same level of energy has a different meaning in Mumbai or Detroit, so a new global geography, equipped with physical, thermodynamic, and economic indicators, is needed in order to consider the quality of the energy flows crossing urban boundaries, and to indicate and reinforce those flows, thereby contributing to the development of the city rather than to its growth.

The concept of *urban metabolism* provides a means of understanding the sustainable development of cities by drawing an analogy with the metabolic processes of organisms. The parallels are strong: *"Cities transform raw materials, fuel, and water into the built environment, human biomass and waste"* [41]. In practice, the study of urban metabolism (in urban ecology) requires quantification of the inputs, out-

puts, energy storage, water, nutrients, materials, and wastes. Indeed, urban metabolism is a suitable approach for the quantification of raw materials and energy supply [87]. This methodology is defined as "*the sum total of the technical and socioeconomic processes that occur in cities, resulting in growth, production of energy, and elimination of waste*" [68]. In other words, urban metabolism is a metaphorical framework that can be used to evaluate the interactions (i.e., flows) between natural and urban ecosystems. In order to assess the sustainability of a city, these interactions should be quantified with appropriate measurement methods that apply a holistic approach in order to account for all the interactions occurring in a system [87].

While urban metabolism has to be considered a mature framework [67], its influence on sustainable urban development is still restricted due to a number of limitations [105]. These include: (a) lack of data at the city scale; (b) strong requirements in data and resources; (c) lack of follow-up and evaluation of the evolution of a city's urban metabolism; and (d) difficulties in identifying cause-and-effect relationships for the metabolic flows.

As a response to these limitations and to the growing digitization of cities worldwide, the concept of *smart urban metabolism* has been suggested by Shahrokni et al. [105]. The implementation of the smart urban metabolism concept in the case of Stockholm Royal Seaport [60] demonstrated its potential to improve data quality, with regard to both resolution and frequency, and to reduce the number of assumptions and simplifications required when using statistical data. Thus, Internet of Things (IoT), real-time heterogeneous data sources, and real-time analytics can act as the foundation to study the flow of materials and energy in urban areas in new ways. For instance, by integrating information and communication technology (ICT) and smart city technologies, the smart urban metabolism model can provide real-time feedback on energy and material flows, from the level of the household to that of the urban district and the city. Despite the high potential, it should be noted that smart urban metabolism is a real-time, data-dependent approach with a number of challenges that must be overcome to unleash its potential. While open datasets relevant to urban metabolism may exist in some circumstances, much of the real-time data or big data needed is contained in silos owned by public or private utilities. Gaining and securing long-term access to such data is thus an essential but challenging task.

Within the framework of smart urban metabolism, big data methods play a fundamental role in improving urban sustainability, especially in large regions and urban areas, where millions of meters and sensors can be installed for the implementation of a real-time monitoring of energy and water flows, as well as providing citizens and policy/decision-makers with a real-time picture of air quality, traffic, and public transportation. Impact on energy and water infrastructures is also relevant: Analysis of real-time data flows is detrimental for the full deployment of renewable energy sources and micro-grids and for the new emerging market of peer-to-peer electricity [44]. Furthermore, an immediate impact of real-time data monitoring of cities can be found in supporting the realization of net negative electric cities, i.e., an electricity fueled city with a negative carbon balance [71,72,108].

5 SoBigData software suites

In this section, we overview the main software platforms developed and made available within SoBigData. While the works described so far are accompanied mostly by standalone packages or demonstrators, the platforms discussed in this section are fully fledged software solutions ready to be used and deployed in real systems.

5.1 The M-Atlas tool

M-Atlas is developed with the main objective of providing a tool to express the analytical power of massive collections of trajectory and positional data in unveiling the complexity of human mobility. M-Atlas is a mobility querying and data mining system centered on the concept of spatiotemporal data. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechanisms for mining trajectory patterns and models that, in turn, can be stored and queried. The knowledge discovery process is based on these kinds of data and helps the user to answer his/her questions of mobility analysts. M-Atlas is equipped with a querying and mining language that makes this analytical process possible and providing the mechanisms to master the complexity of transforming raw GPS tracks into mobility knowledge. M-Atlas is centered on the concept of a trajectory, but is able to handle other kinds of data such as positional data (e.g., Call Data Records), and the mobility knowledge discovery process can be specified by M-Atlas queries that realize all the steps of the knowledge discovery process.

In Fig. 14, an example of the interaction with the system is shown. A Web-based interface is developed and integrated into SoBigData to be used for teaching purposes as well as the downloadable version which is suitable for researchers and industry who may want to use it on their local server and data.

5.2 Self-regulating sharing economies: the EPOS system

Citizens' participation in bottom-up sharing economies of smart cities can contribute to several sustainability goals of the United Nations [61]. Three application scenarios are

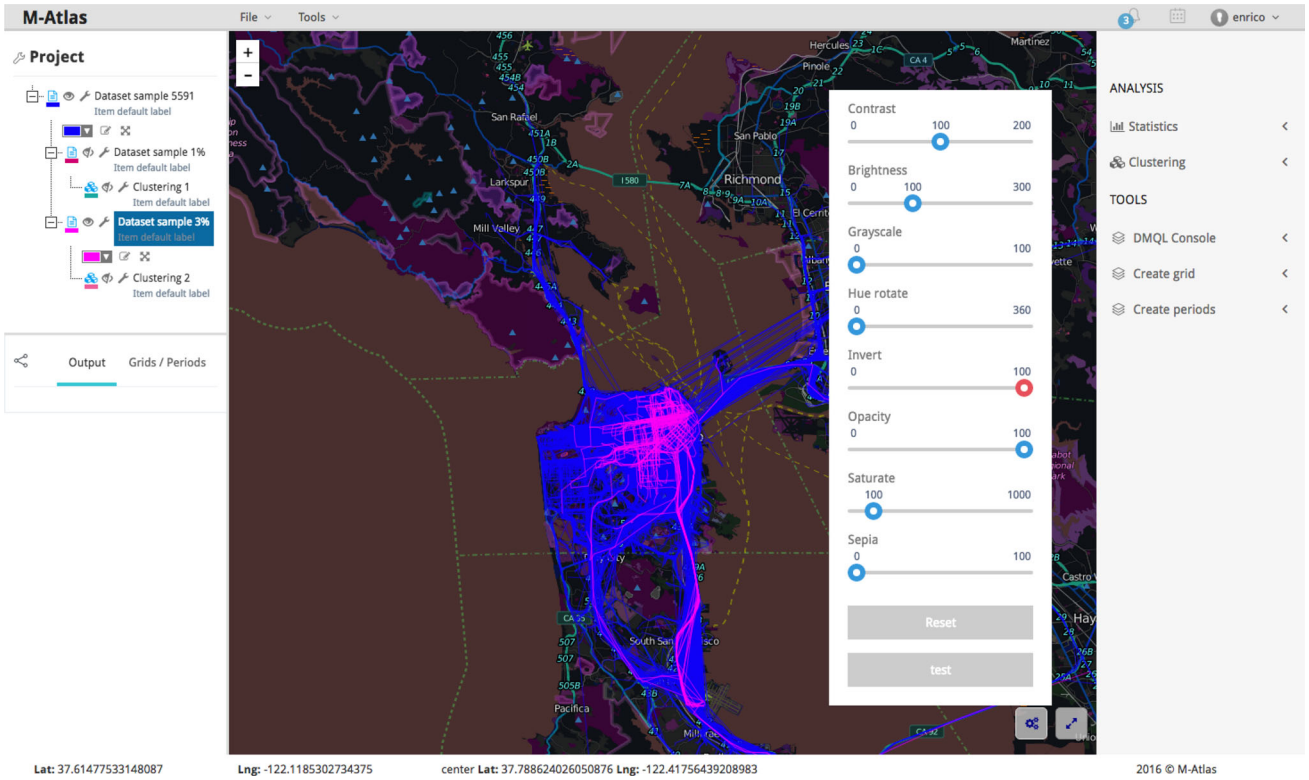


Fig. 14 Web-based interface of M-Atlas. On the left: a tree structure with the analytical steps already computed are shown; in the center, the geographic layers selected on the tree are shown; on the right, the panel where data analytical tools are listed

studied: (i) residential energy management, (ii) charging management of electric vehicles to improve the smart grid reliability, and (iii) managing the utilization of shared bikes to improve the load balancing of bike sharing stations [96,98].

Sharing economies in the aforementioned application scenarios face a foundational challenge of aligning *individual* (citizens) and *collective* (system/city) objectives. On the one hand, citizens make autonomous choices of how they consume or produce resources; for instance, when to turn on their laundry machine, the power level of the heating and cooling system, when to plug in an electrical vehicle to charge or even at which bike sharing station a citizen picks up and leaves a bike. These citizens' decisions all together have a tremendous collective system-level impact; for instance, consuming power at high peak hours can cause blackouts, high power generation costs, and inefficient penetration of renewable energy resources. Similarly, bike sharing stations can become overloaded or underloaded, which increases the operational costs due to manual bike relocations by their system operators. Formally, when agents have a number of discrete options to choose from and coordination is required to minimize a nonlinear cost function such as balancing or matching resource consumption/production, the computational problem of finding the optimal choice for each agent is a combinatorial problem known to be NP-hard [96].

EPOS,[7] the *Economic Planning and Optimized Selections* [95,96], is a fully decentralized learning system to address such challenging computational problems for self-regulating sharing economies. In EPOS, a software agent runs in citizens' personal devices and generates a number of *possible plans* that represent the operational flexibility of the citizen in terms of resource scheduling/allocation. In practice, a plan is a sequence of real values. For instance, determining a time window instead of a certain desired time to turn on a home appliance is a way to generate possible energy consumption plans. Similarly determining several stations from which a citizen is willing to move to pick up or leave a bike is also possible allocation plans of bike sharing stations. These plans are generated under the full authority of the citizens to preserve their autonomy. Plans can be alternative or citizens may have preferences over them, i.e., certain plans may cause higher inconvenience and discomfort than others, e.g., a bike sharing station at far proximity or using a home appliance too late at night. Agents need to make a choice that satisfies the global system-wide objective as well as the citizens' preferences.

EPOS performs coordinated decision-making by self-organizing the communication network of agents in a tree topology over which aggregation of information and

[7] http://epos-net.org.

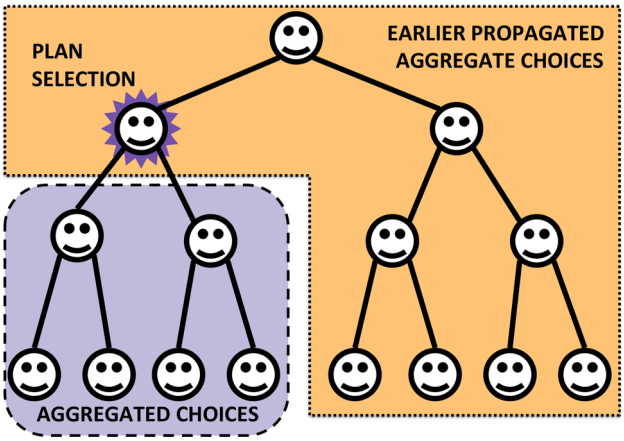
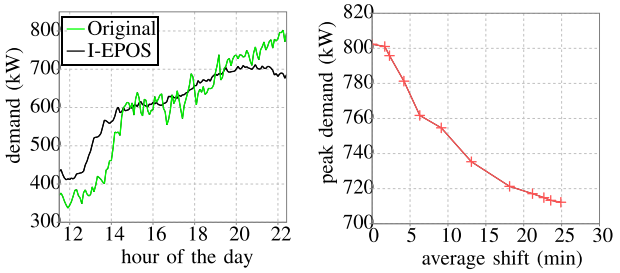


Fig. 15 Decentralized learning and coordination concept of EPOS. An agent makes a plan choice by taking into account (i) the aggregate agent choices in the tree branch underneath and (ii) the aggregate agent choices of the previous learning iteration which is the ones to improve



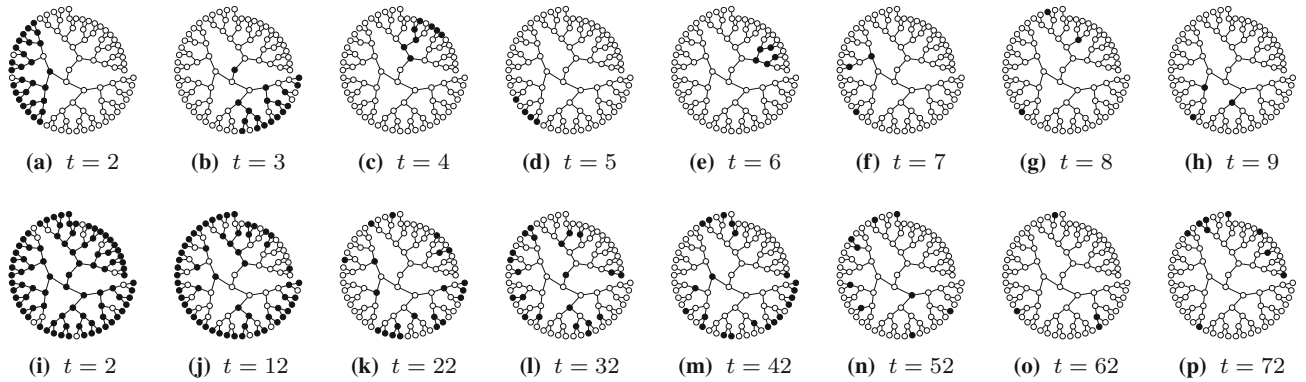
(a) Power load uniformity with and without EPOS **(b)** Power peak shaving capacity vs. citizens' discomfort (load shifting)

Fig. 17 Load balancing of power demand with EPOS [96]

decision-making can be efficiently performed [96]. Learning is performed by consecutive bottom-up and top-down learning iterations during which agents interact in a peer-to-peer fashion. The algorithm resembles backpropagation learning, but in the context of remote agents communicating via a network. Each learning iteration results in a combination of selected plans among the agents. These plans are aggregated (summed up) to evaluate a global objective, for instance the minimization of the variance that is an indicator of load balance in energy consumption and bike sharing stations. A next learning iteration generates a new combination of selected plans such that the variance decreases. Coordination is performed during the process of plan selection by taking into account (i) the aggregate agent choices in the tree branch underneath the agent that selects its plan and (ii) the aggregate agent choices of the previous learning iteration which is the ones to improve. Figure 15 illustrates the concept.

The EPOS algorithm has some striking properties: Its improvement over the learning iterations is fast (3-10 iterations) and monotonous. In terms of optimality, the top 3% of the solutions are found in systems with over a million possible solutions. The high performance of EPOS is also confirmed by comparison with the related work as the ones shown in Fig. 16. EPOS is released as an open-source community software artifact to encourage further research and adoption of decentralized self-management systems for bottom-up sharing economies.

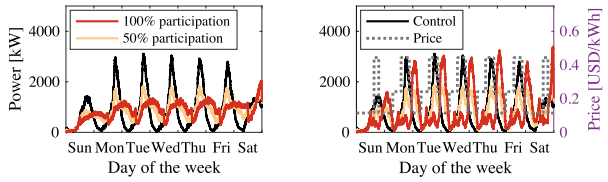
Moreover, EPOS provides the option to the agents to bias the algorithm toward their preferred plans. On the contrary, system operators and utility companies can (monetary) incentivize citizens to sacrifice some of their comfort so that EPOS finds a better solution that contributes to the public good. Socio-technical trade-offs between (i) cost reduction, e.g., variance, (ii) discomfort, e.g., load shifting of energy consumption [91], and (iii) fairness, e.g., dispersion of discomfort among the agents [92] are measured and regulated



(a) $t = 2$ **(b)** $t = 3$ **(c)** $t = 4$ **(d)** $t = 5$ **(e)** $t = 6$ **(f)** $t = 7$ **(g)** $t = 8$ **(h)** $t = 9$

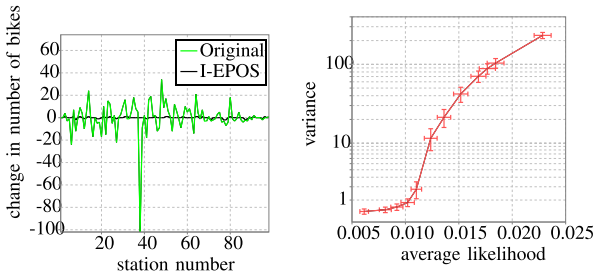
(i) $t = 2$ **(j)** $t = 12$ **(k)** $t = 22$ **(l)** $t = 32$ **(m)** $t = 42$ **(n)** $t = 52$ **(o)** $t = 62$ **(p)** $t = 72$

Fig. 16 Qualitative comparison of the learning process in (a)–(h) EPOS [96] vs. (i)–(p) COHDA [57,58]. The network snapshots indicate the stability and convergence of the learning process, where t is the number of learning iterations. Agents in black change their plan selection to improve the global solution, while agents in white remain to the same plan selection. EPOS relies on exclusively aggregate information exchange, and its convergence speed is significantly faster than the one of COHDA although the latter uses full information exchange



(a) Minimizing power oscillations (standard deviation) **(b)** Minimizing power costs (price incentive signal)

Fig. 18 Decentralized smart grid optimization via coordination of charging electric vehicles [98]



(a) Utilization of bike sharing stations with and without EPOS **(b)** Load-balancing capacity vs. citizens' discomfort (non-preferred bike sharing stations)

Fig. 19 Load balancing of bike sharing stations with EPOS [96]

via the reconfigurable parameters of EPOS. Figures 17, 18 and 19 illustrate the performance of EPOS in three management application domains: residential energy, charging of electrical vehicles, and bike sharing.

6 Privacy-aware data gathering and management

As discussed in "Introduction," while the widespread availability of mobility data opens up new and exciting opportunities to design the smart cities of the future and to improve the quality of urban life, the downside is that such availability may put the privacy of people at risk. In this section, we discuss two approaches for mitigating this problem: mining data (Sect. 6.1) and collecting data (Sect. 6.2) in a privacy-preserving way.

6.1 Privacy-preserving data mining

One of the main reasons for users to share their location is to take advantage of location-based services (LBS). Examples are the trip planning services, where users ask for tips on the nearest point of interest (POI) with respect to their actual location. The typical network structure for location-based service infrastructure relies on wireless technologies, such as WiFi, 3G, or GSM, and has a simple client–network architecture: Users upload their location and they get back the service they have asked for. There are many examples of this kind of application: Consider, for instance, trip planners, which are useful for finding the closest points of interest (hotels, restaurants, touristic places, etc.) [39].

Existing platforms that provide location-based services include location-based social networks, such as Foursquare. Other popular services include route planners that are aware of traffic conditions, e.g., Google maps or the most common navigation assistants. These systems monitor the road network state and suggest to the users the best paths for reaching the requested destination.

In order to provide better services, location-based service platforms collect and mine user data, which are used to develop recommendation systems. Collecting user data raises privacy concerns, and the problem of developing privacy-preserving models has been intensively studied in the literature [47]; however, it is still an open problem. Many users do not perceive the privacy threat, because they are sharing their location only on few occasions, or simply because they are not aware of the privacy implications and considerations. However, the increasing attention drawn to privacy issues has made the field of mobility-related big data research even more challenging.

Dealing with individual and collective views of personal data handling is still a process not taken into account. Actually, the state-of-the-art works focus mainly on the role of who has to handle and analyze a huge amount of personal data. According to Clifton et al. [37], a definitive understanding of what is meant by privacy is still missing. However, privacy-preserving data mining is a research field rich in activities and studies. Bonchi et al. [24] report that in the most commonly used approaches of privacy-preserving data mining there are different levels of privacy, as shown in Fig. 20, from individuals to corporations, to more complex levels of privacy-preserving methods for spatiotemporal data mining. One of the commonly used techniques for privacy-preserving data mining is data perturbation, which relies on either adding noise to the original data or randomizing it. Data perturbation techniques were initially used for statistical disclosure control [1] and later on for privacy-preserving data mining [2]. When considering spatiotemporal data mining, the research on privacy issues is still evolving. As an example, to address privacy considerations in spatiotemporal data, and in particular mobility trajectories, Hoh et al. [59] propose the *path confusion algorithm* for perturbing object trajectories. The idea is that if the proximity of two non-intersecting paths falls below the threshold called perturbation radius, these paths are crossed and their ids are interchanged after the intersection. The objective is that an adversary cannot identify whether these two paths were intersecting in the original dataset or

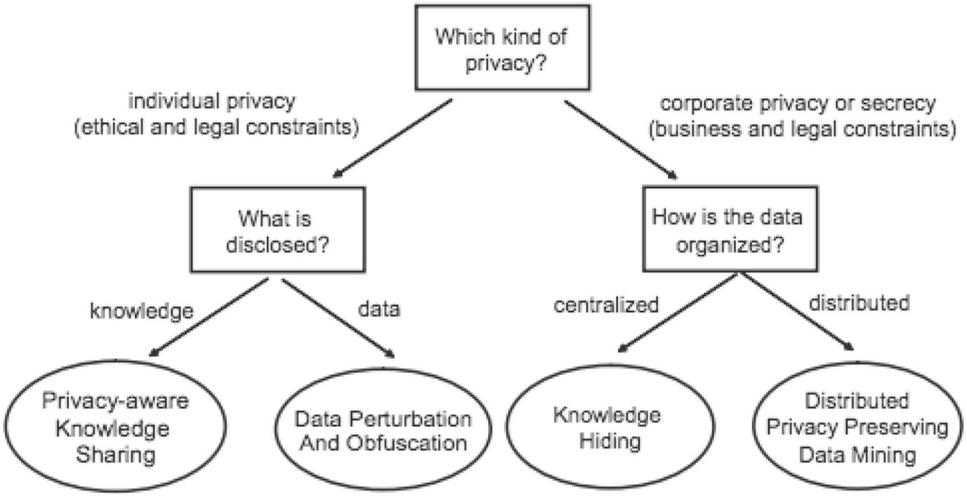


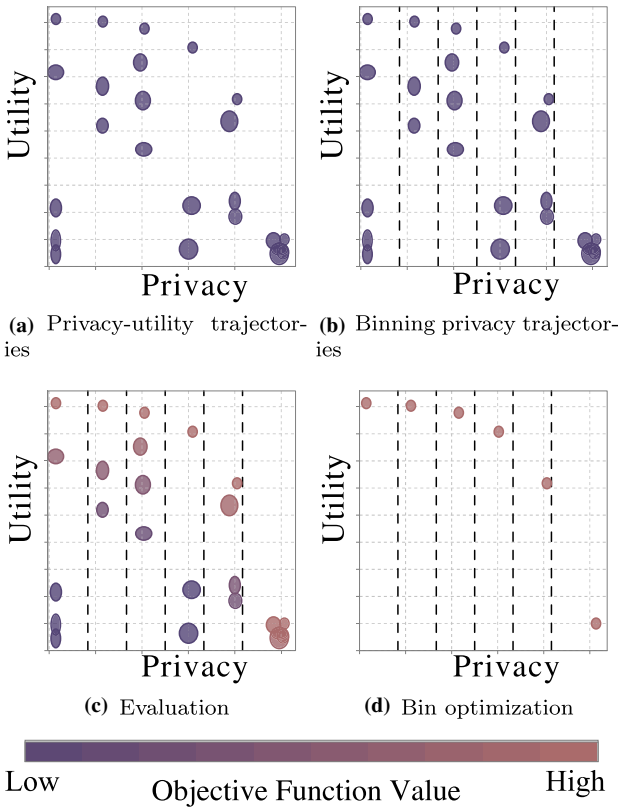
Fig. 20 A classification of different approaches to privacy-preserving data mining

not, since path confusion is only applied to non-intersecting paths.

A new point of view in this field is the so-called new deal on data [86], where Alex Pentland outlines an open information market: Users have the right to possess their data with full control over it, and they may choose to sell their data to some companies, getting in return services or revenue. While there still is not a protocol or a model that implement this new deal on data, we believe that in the near future users will pay more attention to what private information they are disclosing and what is the payback for them.

From a technical perspective [11], data sharing can be modeled as an optimization problem to regulate privacy–utility trade-offs under information self-determination [46]. Such trade-offs are, for instance, the obfuscation of citizens' location vs. the prediction accuracy of traffic congestion. The parameters of differential privacy mechanisms can be computed to satisfy several Pareto efficient privacy–utility values as shown in Fig. 21. These parameters can be set universally or autonomously by citizens via, for instance, user-friendly privacy settings [9]. The latter finding is proved both empirically and theoretically and it has significant implications on the design of (monetary) incentivization schemes of service providers to acquire citizens' data using more socially responsible practices.

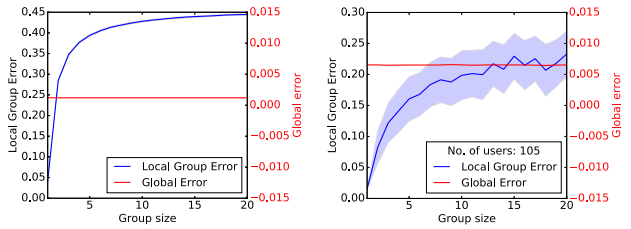
A different approach, which is based on privacy-by-design methodology, can be found in earlier studies [84,85,99]. Here, the framework PRUDEnce is presented, providing an approach that, before applying any privacy-preserving transformation, allows looking at the effective risk there is in the data, as well as the service or purpose for which the data are queried, instead of relying only on theoretical results in terms of privacy. The proposed approach is validated using different data formats underlying many services, defined on real mobility data.



(a) Privacy-utility trajectories **(b)** Binning privacy trajectories **(c)** Evaluation **(d)** Bin optimization

Fig. 21 Optimization of privacy–utility trade-offs [11]: Ellipses denote privacy–utility values for differential privacy settings evaluated according to privacy and utility metrics. Privacy settings are filtered out to form a Pareto efficient trajectory of privacy–utility values

From a social perspective [18], citizens' grouping can be used as a masking mechanism to lower the information revealed to third-party service providers. Grouping can be performed according to semantic criteria (for instance, the proximity of privacy preferences) or according to physical criteria (e.g., geographic location). It is shown that when data



(a) Smart Grid data **(b)** Transportation data

Fig. 22 Average local group error (privacy) and global error (quality of service—aggregation accuracy) in crowdsensing via citizens' grouping [18]. Increasing the group sizes improves privacy, while quality of service remains constant

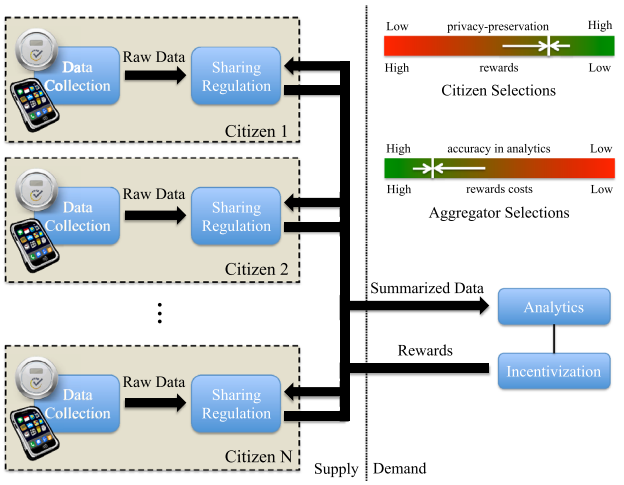


Fig. 23 Modeling data sharing as a supply–demand system [93]. Citizens self-regulate the data they share via data summarization or obfuscation techniques [65]. In contrast, data consumers can incentivize citizens via (monetary) rewards to share more data to improve the quality of service, i.e., the accuracy of data analytics. Therefore, data sharing is an equilibrium and a result of trade-offs: (i) privacy versus rewards for data suppliers and (ii) quality of service versus reward costs for data consumers. The model has been empirically evaluated with real-world data from smart grid pilot projects and smartphone sensor data

are aggregated at a group level before being shared to third-party service providers, the privacy of citizens increases, while service providers preserve the same level of quality of service (Fig. 22).

From an economic perspective [93], data sharing regulatory systems designed to manage trade-offs between privacy and utility can be modeled as supply–demand systems running socially responsible data markets as shown in Fig. 23. Such systems make citizens more aware of their privacy and the value of their data, monetize citizens' data and ultimately incentivize participatory crowdsensing campaigns for the public good.

6.2 Challenges and opportunities in data gathering: distributed crowdsensing

An approach to gathering data at large scale and in aggregate form, while addressing some privacy considerations, is *distributed crowdsensing*. In this approach, citizen data are required at an aggregate level to run a data-intensive service, for instance, computing the total load of a power grid to monitor power peaks that may cause catastrophic blackouts [56] or computing the average speed of vehicles as an indicator of traffic congestion [113]. A critical system design choice is how aggregation is performed, which party performs the aggregation, and what the implications of the aggregation design are for the citizens.

On the one hand, collecting individual citizen data to perform a centralized aggregation at the site of the service provider requires the reveal of personal data and opens up opportunities for discriminatory data analytics [31,115]. For instance, utility companies can perform energy disaggregation to infer with high accuracy the lifestyle and residential activities of citizens [55]. Similarly, the vehicle speed and locations can reveal infractions of the traffic laws and sensitive mobility patterns [18]. Moreover, centralized computations are not scalable and can be single points of failure.

On the other hand, crowdsourcing the aggregation process to citizens by using distributed algorithms for information dissemination and collective computations on the citizens' interconnected devices, e.g., smartphones and wearables, provides privacy-by-design, scalable, trusted, and accountable computations. In this approach, the aggregation of the information is turned into public good by citizens and for citizens.

Performing real-time distributed computations of aggregation functions over a network of citizens' interconnected devices is challenging. More specifically, different aggregation functions, e.g., summation, average, maximum, etc., usually require different algorithms for the same input data [30,40]. The feasibility to provide on-time accurate estimates of the aggregation functions is hindered when input sensor data are continuously updated, i.e., processing a stream of sensor data. Moreover, challenges such as double counting of data, join and drop of devices can deteriorate the performance further. An alternative paradigm for cost-effective data gathering and management in large-scale decentralized networks is required for a distributed approach to crowdsensing.

The *Dynamic Intelligent Aggregation Service* (DIAS)[8] has been introduced by Pournaras et al. [89,90,94] to empower this alternative socially responsible crowdsensing paradigm. A schematic view of the DIAS architecture is shown in Fig. 24. The same DIAS network system can compute almost

[8] http://dias-net.org.

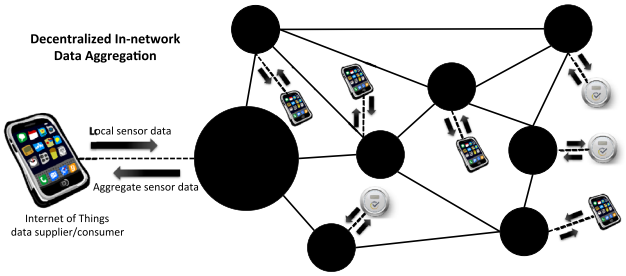


Fig. 24 DIAS crowdsensing approach [94]. Citizens' devices are interconnected to a large-scale decentralized network on which the data aggregation process is crowdsourced. Each device acts as both a *data supplier* and *data consumer*, and therefore, data analytics are turned into a public good run by citizens and for citizens

any aggregation function even under rapid changes in the input sensor data. Citizens do not need to share their exact personal data, but rather representative data profiles, which are used to obtain accurate estimates of the true aggregates. Although sensor data can rapidly change, e.g., the power consumption records of a smart meter, DIAS aggregates, instead, the low, medium, and high profiles of power consumption that are more stable over time and do not change frequently. Therefore, a real-time distributed computation is made feasible [89,90,94].

Moreover, when citizens disconnect from the DIAS network, which may happen due to system failures or when pausing data sharing, self-corrective operations are performed to adjust the values of the aggregation functions to the latest available data and online devices [89,90]. Data management is performed over a distributed memory system of probabilistic data structures, i.e., Bloom filters [29], which label how sensor data should be counted, e.g., counting new values or replacing outdated ones.

The DIAS architecture has been extensively evaluated with real-world data from smart grid pilot projects. Figure 25 illustrates some indicative measurements. A fully working prototype and deployment of DIAS are available for the Euler supercomputer infrastructure of ETH Zurich as well as in servers managed by local communities. Figure 26 illustrates a visualization of such a deployment. DIAS is connected with several front-end systems, such as GDELT [75,97] and the Smart Agora platform [88].

7 Conclusion and future directions

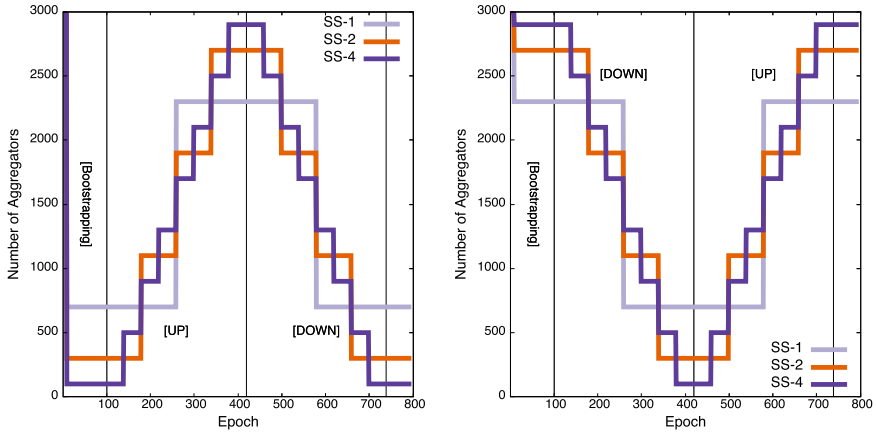
In this paper, we have discussed a wide range of topics related to urban data science, including data collection and management, personal data privacy, distributed crowdsensing data collection, visual analytics for geolocated social media data, modeling, generation, location detection, prediction, and recommendation (see Table 2 for a summary). We also presented resources developed within the SoBigData project, which can be used by researchers, practitioners, or stakeholders. In particular, we presented the M-Atlas tool, a platform developed for querying and mining spatiotemporal mobility data, and EPOS, a fully decentralized system designed to address challenges on managing self-regulating sharing economies.

The presented research, and other related work in the area of smart cities, has made advances on different aspects of data analytics, modeling, and learning from large volumes of data so as to support applications that have the potential to improve individual and collective well-being. Yet, several challenges remain to be addressed and many steps to be taken so as to move forward. In the following, we discuss the most important ones.

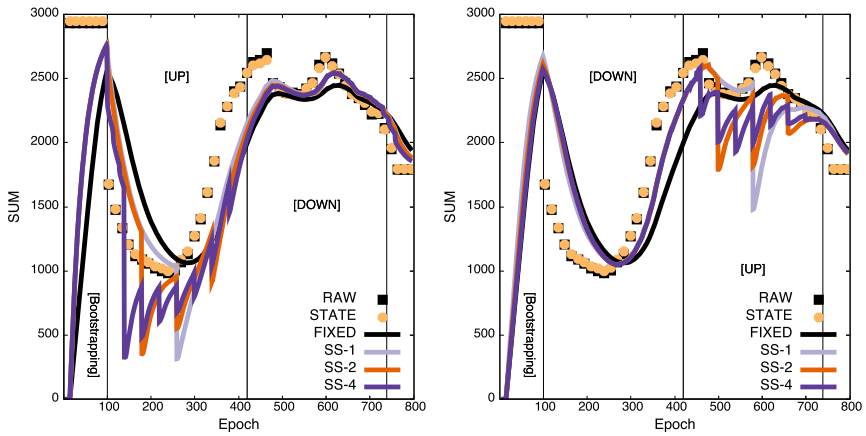
Lack of benchmarks First, on a practical level, research in the area lacks well-established benchmark datasets and well-identified research problems so as to make it easier to quantify progress and to allow researchers to push the state of the art on problem settings with practical validity and high potential for impact. The "City of Citizens" exploratory of the SoBigData project, with a strong focus on developing resources that can be used widely, and cataloguing datasets and methods, is one step toward this direction. However, more consolidation is necessary, also with the participation of researchers outside the SoBigData project.

Trade-off between privacy preservation and social good Second, we discussed in some detail the topic of privacy of personal data. This is a crucial consideration that characterizes one of the most important dilemmas of big data analysis for social good, namely identifying the right balance between offering social benefit by collecting and processing large-scale data of individuals, with protecting the privacy rights of the individuals, as well as avoiding bias and discrimination. In fact, this is an issue of larger scope that goes beyond smart city applications, but obviously it should be also addressed within this context. The steps forward should combine increasing awareness so that citizens become aware of the privacy considerations and demand their rights to be respected, with improving the education of scientists to conduct ethical research, but also with developing the computational solutions required to harness value from data while protecting the privacy of individuals.

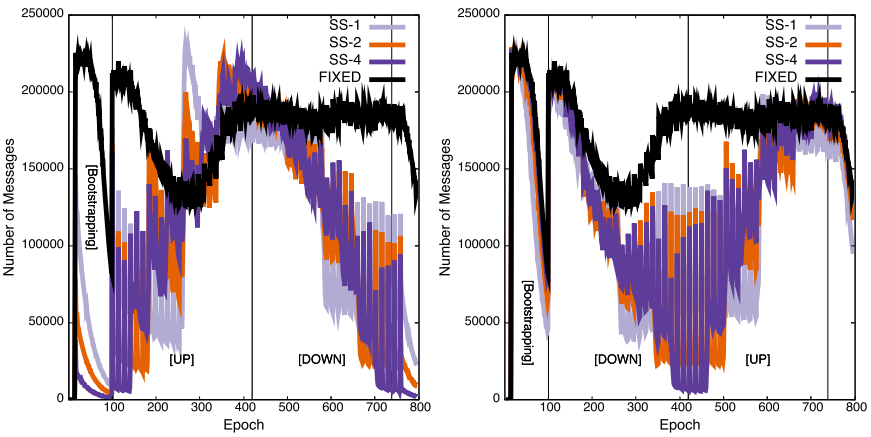
Multi-modal data In all research tasks that we discussed in this paper, we have assumed a single data modality, for example, trajectories, or geolocated social media data or transportation links. In reality, data come in many different modalities, and combining those can be used to extract richer representations and build more accurate models. Developing methods for combining different data modalities is an important research challenge. One should note again the conflict between aggregating more data sources and personal pri-



(a) Scaling up-down and down-up the number of citizens aggregating sensor data.

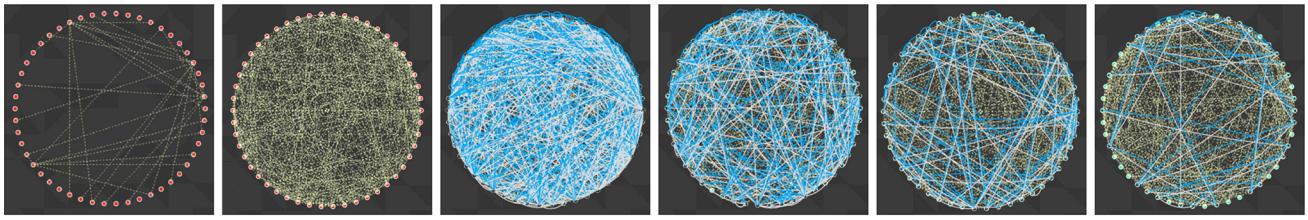


(b) Sum estimation: DIAS estimates accurately the actual (RAW) and approximated (STATE) aggregate sensor values.



(c) Automated adjustments in the communication are performed based on the number of participating citizens.

Fig. 25 Decentralized crowdsensing performance under different scaling scenarios (SS) of citizens' participation [94]



(a) Epoch 14 **(b)** Epoch 15 **(c)** Epoch 18 **(d)** Epoch 20 **(e)** Epoch 21 **(f)** Epoch 23

Fig. 26 Visualization of a DIAS deployment of 50 nodes at the Euler supercomputer infrastructure of ETH Zurich. Epochs indicate the time progress in the emulation of the system operations. The dashed yellow lines indicate the peer-to-peer connections established by a gossiping communication protocol used for distributed node discovery. White and blue solid lines indicate the exchange of sensor data for the local computation of aggregation functions. Nodes are colored red at the very beginning, indicating an inaccurate estimation of the aggregation functions. As more exchanges of sensor data are performed, the nodes turn to green, indicating a maximal accurate estimation of the aggregation functions. DIAS eliminates the communication cost as accuracy increases and devices aggregate acquire the available sensor data in the network [94]

Table 2 Summary of the main topics covered by the paper

Section title	Main findings	Reference literature
Algorithms for urban data analytics	Algorithmic tools addressing critical challenges in urban data science: (i) how to model information extracted from location-based social networks, (ii) TOSCA, RAMA—location detection, (iii) DITRAS—simulation of realistic mobility, (iv) MyWay—individual movement prediction	[10,34,43,51,52,79,100,107,112]
Visual analytics for urban data	Visual analytics for geolocated social media data: photograph sharing and micro-blogging platforms	[3–5,5,6,38,62]
Shaping urban landscape	Use of big data analytics for (i) recommendation to tourists (TRIPBUILDER), (ii) improving shared mobility, (iii) studying the link between human mobility, socioeconomic development, urban sustainability, and net negative cities	[17,21–23,27,39,44,45,60,70,101,114]
SoBigData software suites	Fully fledged platforms: (i) the M-Atlas tool for mining spatiotemporal data, (ii) EPOS for self-regulating sharing economies	[57,95,96]
Privacy-aware data gathering and protection	New deal on data: (i) managing mobility data (ii) anonymization, (iii) PRUDEnce framework, (iv) DIAS	[11,18,37,39,47,59,86,94]

vacy, so this research challenge is as much about learning with complex and heterogeneous data as it is about privacy-preserving data mining.

Incentives to user participation We presented a number of participatory approaches, where citizens are given the opportunity to contribute their data and in return to harnessing gains via access to applications and services. In many cases, however, citizens are reluctant to contribute data or use technologically innovative applications, either because of privacy concerns or because these applications are not useful enough or simply because they are inconvenient to use. As a final challenge, we pose the task of consolidating research in computational methods and data analysis with psychology, gamification, mechanism design, and smart computer–human interaction, so as to increase the participation of citizens in those services and applications by providing meaningful incentives, but also by designing services and applications that are easy to use and transparent.

Acknowledgements Open access funding provided by Aalto University.

Availability of data and materials The data and methods that support the findings described in this paper can be found in the SoBigData catalogue at https://sobigdata.d4science.org/catalogue-sobigdata.

Compliance with ethical standards

Conflict of interest statement On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Adam, N.R., Worthmann, J.C.: Security-control methods for statistical databases: a comparative study. ACM Comput. Surv. (CSUR) **21**(4), 515–556 (1989)
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. SIGMOD Rec. **29**(2), 439–450 (2000)
3. Andrienko, G., Andrienko, N., Bak, P., Kisilevich, S., Keim, D.: Analysis of community-contributed space- and time-referenced data (example of flickr and panoramio photos). In: 2009 IEEE Symposium on Visual Analytics Science and Technology, pp. 213–214 (2009a)
4. Andrienko, G., Andrienko, N., Bak, P., Kisilevich, S., Keim, D.: Analysis of community-contributed space- and time-referenced data by example of panoramio photos (2009b)
5. Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., Poelitz, C.: Identifying place histories from activity traces with an eye to parameter impact. IEEE Trans. Vis. Comput. Graph. **18**(5), 675–688 (2012)
6. Andrienko, G., Andrienko, N., Bak, P., Keim, D., Wrobel, S.: Visual Analytics of Movement. Springer, Berlin (2013)
7. Andrienko, N., Andrienko, G., Fuchs, G., Rinzivillo, S., Betz, H.D.: Detection, tracking, and visualization of spatial event clusters for real time monitoring. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10 (2015)
8. Andrienko, N., Andrienko, G., Fuchs, G., Jankowski, P.: Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. Inf. Vis. **15**(2), 117–153 (2016)
9. Angulo, J., Fischer-Hübner, S., Wästlund, E., Pulls, T.: Towards usable privacy policy display and management. Inf. Manag. Comput. Secur. **20**(1), 4–17 (2012)
10. Ankerst, M., Breunig, M.M., Kriegel, H.P., Sander, J.: Optics: ordering points to identify the clustering structure. SIGMOD Rec. **28**(2), 49–60 (1999)
11. Asikis, T., Pournaras, E.: Optimization of privacy-utility trade-offs under informational self-determination. arXiv preprint arXiv:1710.03186 (2017)
12. Avvenuti, M., Cresci, S., Del Vigna, F., Tesconi, M.: Impromptu crisis mapping to prioritize emergency response. Computer **49**(5), 28–37 (2016a)
13. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., Tesconi, M.: Predictability or early warning: using social media in modern emergency response. IEEE Internet Comput. **20**(6), 4–6 (2016b)
14. Avvenuti, M., Cresci, S., Del Vigna, F., Fagni, T., Tesconi, M.: CrisMap: a big data crisis mapping system based on damage detection and geoparsing. Inf. Syst. Front. **20**(5), 993–1011 (2018a)
15. Avvenuti, M., Cresci, S., Nizzoli, L., Tesconi, M.: GSP (Geo-Semantic-Parsing): geoparsing and geotagging with machine learning on top of linked data. In: European Semantic Web Conference, Springer, pp. 17–32 (2018b)
16. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, Hoboken (1974)
17. Batty, M.: The New Science of Cities. MIT Press, Cambridge (2013)
18. Bennati, S., Pournaras, E.: Privacy-enhancing aggregation of internet of things data via sensors grouping. Sustain. Cities Soc. **39**, 387–400 (2018)
19. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J Mach Learn Res **3**, 993–1022 (2003)
20. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th International Conference on Multimodal Interaction, ACM, pp. 427–434 (2014)
21. Boldrini, C., Bruno, R.: Stackable vs autonomous cars for shared mobility systems: A preliminary performance evaluation. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 232–237 (2017)
22. Boldrini, C., Bruno, R., Conti, M.: Characterising demand and usage patterns in a large station-based car sharing system. In: The 2nd IEEE INFOCOM Workshop on Smart Cities and Urban Computing, IEEE, pp. 1–6 (2016)
23. Boldrini, C., Bruno, R., Laarabi, M.H.: Weak signals in the mobility landscape: car sharing in ten European cities. EPJ Data Sci. **8**(1), 7 (2019)
24. Bonchi, F., Saygin, Y., Verykios, V.S., Atzori, M., Gkoulalas-Divanis, A., Kaya, S.V., Savaş, E.: Privacy in spatiotemporal data mining. In: Mobility, Data Mining and Privacy, Springer, pp. 297–333 (2008)
25. Bosch, H., Thom, D., Heimerl, F., Puettmann, E., Koch, S., Krueger, R., Woerner, M., Ertl, T.: Scatterblogs2: real-time monitoring of microblog messages through user-guided filtering. IEEE Trans. Vis. Comput. Graph. **19**(12), 2022–2031 (2013)
26. Brilhante, I., Macedo, J.A., Nardini, F.M., Perego, R., Renso, C.: Where shall we go today?: Planning touristic tours with tripbuilder. In: Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management, ACM, New York, NY, USA, CIKM 2013, pp. 757–762 (2013)
27. Brilhante, I., Macedo, J.A., Nardini, F.M., Perego, R., Renso, C.: Tripbuilder: A tool for recommending sightseeing tours. In: Rijke, M., Kenter, T., Vries, A.P., Zhai, C.X., Jong, F., Radinsky, K., Hofmann, K. (eds.) Advances in Information Retrieval, Lecture Notes in Computer Science, vol. 8416, Springer International Publishing, pp. 771–774 (2014)
28. Brilhante, I.R., de Macêdo, J.A.F., Nardini, F.M., Perego, R., Renso, C.: On planning sightseeing tours with tripbuilder. Inf. Process. Manag. **51**(2), 1–15 (2015)
29. Broder, A., Mitzenmacher, M.: Network applications of bloom filters: a survey. Internet Math. **1**(4), 485–509 (2004)
30. Can, Z., Demirbas, M.: A survey on in-network querying and tracking services for wireless sensor networks. Ad Hoc Netw. **11**(1), 596–610 (2013)
31. Carmichael, L., Stalla-Bourdillon, S., Staab, S.: Data mining and automated discrimination: a mixed legal/technical perspective. IEEE Intell. Syst. **31**(6), 51–55 (2016)
32. Ceci, M., Appice, A., Malerba, D.: Time-slice density estimation for semantic-based tourist destination suggestion. In: ECAI, pp. 1107–1108 (2010)
33. Çelikten, E., Falher, G.L., Mathioudakis, M.: "What Is the City but the People?": Exploring Urban Activity Using Social Web Traces. In: Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016, Companion Volume, pp. 167–170 (2016)
34. Çelikten, E., Falher, G.L., Mathioudakis, M.: Modeling urban behavior by mining geotagged social data. IEEE Trans. Big Data **3**(2), 220–233 (2017)
35. Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S., Ertl, T.: Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. In: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 143–152 (2012)

36. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM international conference on Information and knowledge management, ACM, pp. 759–768 (2010)
37. Clifton, C., Kantarcioglu, M., Vaidya, J.: Defining privacy for data mining. In: National Science Foundation Workshop on Next Generation Data Mining, Citeseer, vol. 1, p. 1 (2002)
38. Cresci, S.: Harnessing the social sensing revolution: challenges and opportunities. PhD dissertation, University of Pisa (2018)
39. Cresci, S., D'Errico, A., Gazzé, D., Duca, A.L., Marchetti, A., Tesconi, M.: Towards a DBpedia of tourism: The case of TourPedia. In: International Semantic Web Conference (Posters & Demos), pp. 129–132 (2014)
40. Dagar, M., Mahajan, S.: Data aggregation in wireless sensor network: a survey. Int. J. Inf. Comput. Technol. **3**(3), 167–174 (2013)
41. Decker, E.H., Elliott, S., Smith, F.A., Blake, D.R., Rowland, F.S.: Energy and material flow through the urban ecosystem. Annu. Rev. Energy Environ. **25**, 685–740 (2000)
42. Dunbar, R.I., Arnaboldi, V., Conti, M., Passarella, A.: The structure of online social networks mirrors those in the offline world. Social Netw. **43**, 39–47 (2015)
43. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, vol. 96, pp. 226–231 (1996)
44. Facchini, A.: Distributed energy resources: planning for the future. Nat. Energy **2**, 17129 (2017)
45. Facchini, A., Kennedy, C., Stewart, I., Mele, R.: The energy metabolism of megacities. Appl. Energy **186**, 86–95 (2017)
46. Fialová, E.: Data portability and informational self-determination. Masaryk UJL Technol. **8**, 45 (2014)
47. Giannotti, F., Pedreschi, D.: Mobility, Data Mining and Privacy Geographic Knowledge Discovery. Springer, Berlin (2008)
48. Giannotti, F., Pappalardo, L., Pedreschi, D., Wang, D.: A complexity science perspective on human mobility. In: Mobility Data: Modeling, Management, and Understanding, Cambridge University Press, pp. 297–314 (2013)
49. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. Nature **453**(7196), 779 (2008)
50. Grossi, V., Rapisarda, B., Giannotti, F., Pedreschi, D.: Data science at sobigdata: the european research infrastructure for social mining and big data analytics. Int. J. Data Sci. Anal. **6**(3), 205–216 (2018)
51. Guidotti, R.: Personal Data Analytics: Capturing Human Behavior to Improve Self-Awareness and Personal Services through Individual and Collective Knowledge. PhD thesis, University of Pisa (2017)
52. Guidotti, R., Trasarti, R., Nanni, M.: TOSCA: two-steps clustering algorithm for personal locations detection. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, p. 38 (2015)
53. Guidotti, R., Trasarti, R., Nanni, M., Giannotti, F., Pedreschi, D.: There's a path for everyone: A data-driven personal model reproducing mobility agendas. In: Data Science and Advanced Analytics (DSAA), 2017 IEEE International Conference on, IEEE, pp. 303–312 (2017)
54. Guidotti, R., Gabrielli, L., Monreale, A., Pedreschi, D., Giannotti, F.: Discovering temporal regularities in retail customers' shopping behavior. EPJ Data Sci. **7**(1), 6 (2018)
55. Hattori, M., Hirano, T., Matsuda, N., Shimizu, R., Wang, Y.: Privacy-utility tradeoff for applications using energy disaggregation of smart-meter data. In: Australasian Conference on Information Security and Privacy, Springer, pp. 214–234 (2017)
56. Helbing, D.: Globally networked risks and how to respond. Nature **497**(7447), 51 (2013)
57. Hinrichs, C., Sonnenschein, M.: A distributed combinatorial optimisation heuristic for the scheduling of energy resources represented by self-interested agents. IJBIC **10**(2), 69–78 (2017)
58. Hinrichs, C., Lehnhoff, S., Sonnenschein, M.: A decentralized heuristic for multiple-choice combinatorial optimization problems. In: Operations Research Proceedings 2012, Springer, pp. 297–302 (2014)
59. Hoh, B., Gruteser, M.: Protecting location privacy through path confusion. In: First International Conference on Security and Privacy for Emerging Areas in Communications Networks, 2005. SecureComm 2005. IEEE, pp. 194–205 (2005)
60. Hossein, S., Louise, Å., David, L., Anders, N., Nils, B.: Implementing smart urban metabolism in the Stockholm royal seaport: smart city SRS. J. Ind. Ecol. **19**(5), 917–929 (2015)
61. Ibrahim, M., El-Zaart, A., Adams, C.: Smart sustainable cities: A new perspective on transformation, roadmap, and framework concepts. In: The Fifth International Conference on Smart Cities, Systems, Devices and Technologies (includes URBAN COMPUTING 2016), IARIA, pp. 8–14 (2016)
62. Jankowski, P., Andrienko, N., Andrienko, G., Kisilevich, S.: Discovering landmark preferences and movement patterns from photo postings. Trans. GIS **14**(6), 833–852 (2010)
63. Jeung, H., Liu, Q., Shen, H.T., Zhou, X.: A hybrid prediction model for moving objects. In: IEEE 24th International Conference on Data Engineering, 2008. ICDE 2008. IEEE, pp. 70–79 (2008)
64. Jiang, S., Ferreira, J., González, M.C.: Clustering daily patterns of human activities in the city. Data Min. Knowl. Discov. **25**(3), 478–510 (2012)
65. Kandappu, T., Misra, A., Cheng, S.F., Tandriansyah, R., Lau, H.C.: Obfuscation at-source: privacy in context-aware mobile crowd-sourcing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **2**(1), 16 (2018)
66. Keim, E.D., Kohlhammer, J., Ellis, G.: Mastering the information age: solving problems with visual analytics, eurographics association (2010)
67. Kennedy, C., Hoornweg, D.: Mainstreaming urban metabolism. J. Ind. Ecol. **16**(6), 780–782 (2012)
68. Kennedy, C., Cuddihy, J., Engel-Yan, J.: The changing metabolism of cities. J. Ind. Ecol. **11**(2), 43–59 (2007)
69. Kennedy, C., Stewart, I., Ibrahim, N., Facchini, A., Mele, R.: Developing a multi-layered indicator set for urban metabolism studies in megacities. Ecol. Indic. **47**, 7–15 (2014)
70. Kennedy, C., Stewart, I., Facchini, A., Cersosimo, I., Mele, R., Chen, B., Uda, M., Kansal, A., Chiu, A., Kim, K.G., Dubeux, C., La Rovere, E., Cunha, B., Pincetl, S., Keirstead, J., Barles, S., Pusaka, S., Gunawan, J., Adegbile, M., Nazariha, M., Hoque, S., Marcotullio, P., Otharán, F., Genena, T., Ibrahim, N., Farooqui, R., Cervantes, G., Sahin, A.: Energy and material flows of megacities. Proc. Natl. Acad. Sci. USA **112**(19), 5985–5990 (2015)
71. Kennedy, C., Stewart, I.D., Facchini, A., Mele, R.: The role of utilities in developing low carbon, electric megacities. Energy Policy **106**, 122–128 (2017)
72. Kennedy, C., Stewart, I.D., Westphal, M.I., Facchini, A., Mele, R.: Keeping global climate change within 1.5C through net negative electric cities. Curr. Opin. Environ. Sustain. **30**, 18–25 (2018)
73. Kortum, K.: Driving smart: Carsharing mode splits and trip frequencies. In: Transportation Research Board 93rd Annual Meeting, 14-4009 (2014)
74. Krumm, J., Horvitz, E.: Predestination: inferring destinations from partial trajectories. In: International Conference on Ubiquitous Computing, Springer, pp. 243–260 (2006)
75. Leetaru, K., Schrodt, P.A.: Gdelt: Global data on events, location, and tone, 1979–2012. In: ISA Annual Convention, Citeseer, vol. 2, pp. 1–49 (2013)
76. Méneroux, Y., Le Guilcher, A., Saint Pierre, G., Hamed, M.G., Mustière, S., Orfila, O.: Traffic signal detection from in-vehicle

GPS speed profiles using functional data analysis and machine learning. Int. J. Data Sci. Anal. 1–19 (2019)
77. Middleton, S.E., Middleton, L., Modafferi, S.: Real-time crisis mapping of natural disasters using social media. IEEE Intell. Syst. **29**(2), 9–17 (2013)
78. Morzy, M.: Prediction of moving object location based on frequent trajectories. In: International Symposium on Computer and Information Sciences, Springer, pp. 583–592 (2006)
79. Pappalardo, L., Simini, F.: Data-driven generation of spatio-temporal routines in human mobility. Data Min. Knowl. Discov. **32**(3), 787–829 (2018)
80. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.L.: Returners and explorers dichotomy in human mobility. Nat. Commun. **6**, 8166 (2015)
81. Pappalardo, L., Rinzivillo, S., Simini, F.: Human mobility modelling: Exploration and preferential return meet the gravity model. Procedia Computer Science 83:934–939. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016)/The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016)/Affiliated Workshops (2016)
82. Pappalardo, L., Simini, F., Barlacchi, G., Pellungrini, R.: scikit-mobility: a Python library for the analysis, generation and risk assessment of mobility data. arXiv preprint arXiv:1907.07062 (2019)
83. Pelleg, D., Moore, A.W., et al.: X-means: extending k-means with efficient estimation of the number of clusters. In: Icml, vol. 1, pp. 727–734 (2000)
84. Pellungrini, R., Pappalardo, L., Pratesi, F., Monreale, A.: Fast estimation of privacy risk in human mobility data. In: Tonetta, S., Schoitsch, E., Bitsch, F. (eds.) Computer Safety, Reliability, and Security, pp. 415–426. Springer, Cham (2017)
85. Pellungrini, R., Pappalardo, L., Pratesi, F., Monreale, A.: A data mining approach to assess privacy risk in human mobility data. ACM Trans. Intell. Syst. Technol. **9**(3), 31 (2018)
86. Pentland, A.: Reality mining of mobile communications: toward a new deal on data. Glob. Inf. Technol. Rep. **2008–2009**, 1981 (2009)
87. Pincetl, S., Bunje, P., Holmes, T.: An expanded urban metabolism method: toward a systems approach for assessing urban energy processes and causes. Landsc. Urban Plan. **107**(3), 193–202 (2012)
88. Pournaras, E.: Proof of witness presence: blockchain consensus for augmented democracy in smart cities. arXiv preprint arXiv:1907.00498 (2019)
89. Pournaras, E., Nikolić, J.: On-demand self-adaptive data analytics in large-scale decentralized networks. In: 2017 IEEE 16th International Symposium on Network Computing and Applications (NCA), IEEE, pp. 1–10 (2017a)
90. Pournaras, E., Nikolić, J.: Self-corrective dynamic networks via decentralized reverse computations. In: 2017 IEEE International Conference on Autonomic Computing (ICAC), IEEE, pp. 11–20 (2017b)
91. Pournaras, E., Vasirani, M., Kooij, R.E., Aberer, K.: Decentralized planning of energy demand for the management of robustness and discomfort. IEEE Trans. Ind. Inform. **10**(4), 2280–2289 (2014a)
92. Pournaras, E., Vasirani, M., Kooij, R.E., Aberer, K.: Measuring and controlling unfairness in decentralized planning of energy demand. In: 2014 IEEE International on Energy Conference (ENERGYCON), IEEE, pp. 1255–1262 (2014b)
93. Pournaras, E., Nikolic, J., Velásquez, P., Trovati, M., Bessis, N., Helbing, D.: Self-regulatory information sharing in participatory social sensing. EPJ Data Sci. **5**(1), 14 (2016)
94. Pournaras, E., Nikolic, J., Omerzel, A., Helbing, D.: Engineering democratization in internet of things data analytics. In: 2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), IEEE, pp. 994–1003 (2017a)
95. Pournaras, E., Yao, M., Helbing, D.: Self-regulating supply-demand systems. Future Gener. Comput. Syst. **76**, 73–91 (2017b)
96. Pournaras, E., Pilgerstorfer, P., Asikis, T.: Decentralized collective learning for self-managed sharing economies. ACM Trans. Auton. Adapt. Syst. **13**(2), 1–33 (2018)
97. Pournaras, E., Gaere, E., Kunz, R., Ghulam, A.N.: Democratizing data analytics: crowd-sourcing decentralized collective measurements. In: 2019 IEEE 4th International Workshops on Foundations and Applications of Self* Systems (FAS* W), IEEE, pp. 265–266 (2019a)
98. Pournaras, E., Jung, S., Yadhunathan, S., Zhang, H., Fang, X.: Socio-technical smart grid optimization via decentralized charge control of electric vehicles. Appl. Soft Comput. **82**, 105573 (2019b)
99. Pratesi, F., Monreale, A., Trasarti, R., Giannotti, F., Pedreschi, D., Yanagihara, T.: PRUDEnce: a system for assessing privacy risk vs utility in data sharing ecosystems. Trans. Data Priv. **11**(2), 139–167 (2018)
100. Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., Giannotti, F.: The purpose of motion: Learning activities from individual mobility networks. In: 2014 International Conference on Data Science and Advanced Analytics (DSAA), IEEE, pp. 312–318 (2014)
101. Scala, A., D'Agostino, G.: Networks of networks: the last frontier of complexity. Springer, Berlin (2014)
102. Scellato, S., Musolesi, M., Mascolo, C., Latora, V., Campbell, A.T.: Nextplace: a spatio-temporal prediction framework for pervasive systems. In: International Conference on Pervasive Computing, Springer, pp. 152–169 (2011)
103. Schwieger, B., Victorero-Solares, P., Brook, D.: Global carsharing operators report 2015. Technical Report, Team Red (2015)
104. Shaheen, S., Cohen, A.: Mobility and the Sharing Economy: Impacts Synopsis–Spring 2015. Technical Report, Transportation Sustainability Research Center, University of California, Berkeley (2015)
105. Shahrokni, H., Lazarevic, D., Brandt, N.: Smart urban metabolism: towards a real-time understanding of the energy and material flows of a city and its citizens. J. Urban Technol. **22**(1), 65–86 (2015)
106. Sibson, R.: Slink: an optimally efficient algorithm for the single-link cluster method. Comput. J. **16**(1), 30–34 (1973)
107. Spinsanti, L., Berlingerio, M., Pappalardo, L.: Mobility and geo-social networks. In: Mobility Data: Modeling, Management, and Understanding, Cambridge University Press, pp. 315–333 (2013)
108. Stewart, I.D., Kennedy, C.A., Facchini, A., Mele, R.: The electric city as a solution to sustainable urban development. J. Urban Technol. **25**(1), 3–20 (2018)
109. Thom, D., Jankowski, P., Fuchs, G., Ertl, T., Bosch, H., Andrienko, N., Andrienko, G.: Thematic patterns in georeferenced tweets through space-time visual analytics. Comput. Sci. Eng. **15**, 72–82 (2013)
110. Tosi, D.: Cell phone big data to compute mobility scenarios for future smart cities. Int. J. Data Sci. Anal. **4**(4), 265–284 (2017)
111. Trasarti, R., Pinelli, F., Nanni, M., Giannotti, F.: Mining mobility user profiles for car pooling. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 1190–1198 (2011)
112. Trasarti, R., Guidotti, R., Monreale, A., Giannotti, F.: MyWay: location prediction via mobility profiling. Inf. Syst. **64**, 350–367 (2017)
113. Wan, J., Liu, J., Shao, Z., Vasilakos, A.V., Imran, M., Zhou, K.: Mobile crowd sensing for traffic prediction in internet of vehicles. Sensors **16**(1), 88 (2016)

114. Weikl, S., Bogenberger, K.: Relocation strategies and algorithms for free-floating car sharing systems. Intell. Transp. Syst. Mag. IEEE **5**(4), 100–111 (2013)
115. Winter, J.: Algorithmic discrimination: big data analytics and the future of the internet. In: The Future Internet, Springer, pp. 125–140 (2015)
116. Zheng, Y., Capra, L., Wolfson, O., Yang, H.: Urban computing: concepts, methodologies, and applications. ACM Trans. Intell. Syst. Technol. **5**(3), 38 (2014)
117. Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G.: Variability in regularity: mining temporal mobility patterns in london, singapore and beijing using smart-card data. PLoS One **11**(2), e0149222 (2016)

Affiliations

Gennady Andrienko[1,2] · Natalia Andrienko[1,2] · Chiara Boldrini[3] · Guido Caldarelli[5,6] · Paolo Cintia[10] · Stefano Cresci[3] · Angelo Facchini[5,7] · Fosca Giannotti[4] · Aristides Gionis[8,12] · Riccardo Guidotti[10] · Michael Mathioudakis[9] · Cristina Ioana Muntean[4] · Luca Pappalardo[4] · Dino Pedreschi[10] · Evangelos Pournaras[11] · Francesca Pratesi[10] · Maurizio Tesconi[3] · Roberto Trasarti[4]

Gennady Andrienko
gennady.andrienko@iais.fraunhofer.de

Natalia Andrienko
natalia.andrienko@iais.fraunhofer.de

Chiara Boldrini
c.boldrini@iit.cnr.it

Guido Caldarelli
guido.caldarelli@imtlucca.it

Paolo Cintia
cintia@di.unipi.it

Stefano Cresci
s.cresci@iit.cnr.it

Angelo Facchini
angelo.facchini@imtlucca.it

Fosca Giannotti
fosca.giannotti@isti.cnr.it

Riccardo Guidotti
guidotti@di.unipi.it

Michael Mathioudakis
michael.mathioudakis@helsinki.fi

Cristina Ioana Muntean
cristina.muntean@isti.cnr.it

Luca Pappalardo
luca.pappalardo@isti.cnr.it

Dino Pedreschi
pedreschi@di.unipi.it

Evangelos Pournaras
e.pournaras@leeds.ac.uk

Francesca Pratesi
pratesi@di.unipi.it

Maurizio Tesconi
m.tesconi@iit.cnr.it

Roberto Trasarti
roberto.trasarti@isti.cnr.it

[1] Fraunhofer Institute IAIS, Sankt Augustin, Germany

[2] City University London, London, UK

[3] CNR-IIT, Pisa, Italy

[4] CNR-ISTI, Pisa, Italy

[5] IMT Alti Studi Lucca, Lucca, Italy

[6] ECLT, CNR-ISC Dip. Fisica, Università "Sapienza" Piazzale A. Moro 2, Rome, Italy

[7] CNR-ISC, Rome, Italy

[8] KTH Royal Institute of Technology, Stockholm, Sweden

[9] University of Helsinki, Helsinki, Finland

[10] University of Pisa, Pisa, Italy

[11] School of Computing, University of Leeds, Leeds, UK

[12] Aalto university, Espoo, Finland