Mobility Data Science: Perspectives and Challenges

MOHAMED MOKBEL, Computer Science and Engineering, University of Minnesota, Minneapolis, United States
MAHMOUD SAKR, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium and Ain Shams University, Cairo, Egypt
LI XIONG, Emory University, Atlanta, United States
ANDREAS ZÜFLE, Computer Science, Emory University, Atlanta, United States
JUSSARA ALMEIDA, Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil
TAYLOR ANDERSON, Department of Geography and Geoinformation Science, George Mason University, Fairfax, United States
WALID AREF, Computer Science, Purdue University, West Lafayette, United States
GENNADY ANDRIENKO, KD, Fraunhofer SIT Saint Augustin Branch, Sankt Augustin, Germany
NATALIA ANDRIENKO, KD, Fraunhofer SIT Saint Augustin Branch, Sankt Augustin, Germany
YANG CAO, School of Computing, Kyoto University, Kyoto, Japan
SANJAY CHAWLA, Qatar Computing Research Institute, HBKU, Doha, Qatar
REYNOLD CHENG, Computer Science, HKU, Hong Kong, Hong Kong
PANOS CHRYSANTHIS, Computer Science, University of Pittsburgh, Pittsburgh, United States
XIQI FEI, Department of Geography and Geoinformation Science, George Mason University, Fairfax, United States
GABRIEL GHINITA, Hamad Bin Khalifa University College of Science and Engineering, Doha, Qatar
ANITA GRASER, AIT Austrian Institute of Technology GmbH, Wien, Austria
DIMITRIOS GUNOPULOS, National and Kapodistrian University of Athens, Athens, Greece
CHRISTIAN S. JENSEN, Department of Computer Science, Aalborg University, Aalborg, Denmark
JOON-SEOK KIM, Geospatial Science and Human Security Division, Pacific Northwest National Laboratory, Richland, United States
KYOUNG-SOOK KIM, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology Tokyo Bay Area Center, Koto-ku, Japan
PEER KRÖGER, Kiel University, Kiel, Germany
JOHN KRUMM, Computer Science, University of Southern California, Los Angeles, United States
JOHANNES LAUER, HERE Technologies, Schwalbach am Taunus, Germany
AMR MAGDY, Computer Science and Engineering, University of California Riverside, Riverside, United States
MARIO NASCIMENTO, Khoury College of Computer Sciences, Northeastern University, Vancouver, Canada
SIVA RAVADA, Oracle, Nashua, United States
MATTHIAS RENZ, Computer Science, Kiel University, Kiel, Germany

ACM 2374-0353/2024/06-ART10
https://doi.org/10.1145/3652158

Corrected Version of Record. V.1.1. Published July 18, 2024.

DIMITRIS SACHARIDIS, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium
FLORA SALIM, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
MOHAMED SARWAT, Arizona State University, Tempe, United States
MAXIME SCHOEMANS, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium
CYRUS SHAHABI, Computer Science, University of Southern California, Los Angeles, United States
BETTINA SPECKMANN, Technische Universiteit Eindhoven, Eindhoven, Netherlands
EGEMEN TANIN, University of Melbourne, Australia, Melbourne, Australia
XU TENG, Iowa State University, Ames, United States
YANNIS THEODORIDIS, University of Piraeus, Piraeus, Greece
KRISTIAN TORP, Aalborg Universitet, Aalborg, Denmark
GOCE TRAJCEVSKI, Iowa State University, Ames, United States
MARC VAN KREVELD, Information and Computing Sciences, Utrecht University, the Netherlands, Utrecht, Netherlands
CAROLA WENK, Department of Computer Science, Tulane University, New Orleans, United States
MARTIN WERNER, School of Engineering and Design, Techincal University of Munich, Munchen, Germany
RAYMOND WONG, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong
SONG WU, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium
JIANQIU XU, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
MOUSTAFA YOUSSEF, Computer Science and Engineering, The American University in Cairo, Cairo, Egypt
DEMETRIS ZEINALIPOUR, Department of Computer Science, University of Cyprus, Nicosia, Cyprus
MENGXUAN ZHANG, School of Computing, Australian National University, Canberra, Australia
ESTEBAN ZIMÁNYI, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium

---

This is a community publication. The authors of this article met in Dagstuhl for Seminar #22021 on Mobility Data Science [152]. The first four authors co-organized the Dagstuhl seminar leading to this article and coordinated the creation of this manuscript. All other authors contributed equally to this research. The seminar was held in the week of January 9 - 14, 2022. It had 47 participants specializing in different topics: data management, mobility analysis, geography, privacy, urban computing, systems, simulation, indoors, visualization, information integration, and theory. Due to COVID-19, the seminar took place in hybrid mode, with 8 onsite and 39 remote participants. Despite the challenge of different time zones of the participants, all sessions were attended by at least 37 participants. It was an excellent opportunity to start the discussion about next-decade opportunities and challenges for mobility data science.

Authors' Contact Information: Mohamed Mokbel, Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States; e-mail: mokbel@umn.edu; Mahmoud Sakr, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Bruxelles, Belgium and Ain Shams University, Cairo, Egypt; e-mail: mahmoud.sakr@ulb.be; Li Xiong, Emory University, Atlanta, GA, United States; e-mail: lxiong@emory.edu; Andreas Züfle, Computer Science, Emory University, Atlanta, GA, United States; e-mail: azufle@emory.edu; Jussara Almeida, Computer Science, Federal University of Minas Gerais, Belo Horizonte, Brazil; e-mail: jussara@dcc.ufmg.br; Taylor Anderson, Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, United States; e-mail: tander6@gmu.edu; Walid Aref, Computer Science, Purdue University, West Lafayette, IN, United States; e-mail: aref@cs.purdue.edu;

Corrected Version of Record. V.1.1. Published July 18, 2024.

Mobility data captures the locations of moving objects such as humans, animals, and cars. With the availability of Global Positioning System (GPS)–equipped mobile devices and other inexpensive location-tracking technologies, mobility data is collected ubiquitously. In recent years, the use of mobility data has demonstrated a significant impact in various domains, including traffic management, urban planning, and health sciences. In this article, we present the domain of mobility data science. Towards a unified approach to mobility data science, we present a pipeline having the following components: mobility data collection, cleaning, analysis, management, and privacy. For each of these components, we explain how mobility data science differs from general data science, we survey the current state-of-the-art, and describe open challenges for the research community in the coming years.

Gennady Andrienko, KD, Fraunhofer SIT Saint Augustin Branch, Sankt Augustin, Nordrhein-Westfalen, Germany; e-mail: gennady.andrienko@iais.fraunhofer.de; Natalia Andrienko, KD, Fraunhofer SIT Saint Augustin Branch, Sankt Augustin, Nordrhein-Westfalen, Germany; e-mail: natalia.andrienko@ iais.fraunhofer.de; Yang Cao, School of Computing, Kyoto University, Kyoto, Kyoto, Japan; e-mail: yang@i.kyoto-u.ac.jp; Sanjay Chawla, Qatar Computing Research Institute, HBKU, Doha, Ad-Dawhah, Qatar; e-mail: schawla@qf.org.qa; Reynold Cheng, Computer Science, HKU, Hong Kong, SAR, Hong Kong; e-mail: ckcheng@cs.hku.hk; Panos Chrysanthis, Computer Science, University of Pittsburgh, Pittsburgh, PA, United States; e-mail: panos@cs.pitt.edu; Xiqi Fei, Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, United States; e-mail: xfei@gmu.edu; Gabriel Ghinita, Hamad Bin Khalifa University College of Science and Engineering, Doha, Qatar; e-mail: gabriel.ghinita@umb.edu; Anita Graser, AIT Austrian Institute of Technology GmbH, Wien, Wien, Austria; e-mail: Anita.Graser@ait.ac.at; Dimitrios Gunopulos, National and Kapodistrian University of Athens, Athens, Greece; e-mail: dg@di.uoa.gr; Christian S. Jensen, Department of Computer Science, Aalborg University, Aalborg, Denmark; e-mail: csj@cs.aau.dk; Joon-Seok Kim, Geospatial Science and Human Security Division, Pacific Northwest National Laboratory, Richland, Washington, United States; e-mail: joonseok.kim@pnnl.gov; Kyoung-Sook Kim, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology Tokyo Bay Area Center, Koto-ku, Japan; e-mail: ks.kim@aist.go.jp; Peer Kröger, Kiel University, Kiel, Schleswig-Holstein, Germany; e-mail: pkr@informatik.uni-kiel.de; John Krumm, Computer Science, University of Southern California, Los Angeles, CA, United States; e-mail: jckrumm@outlook.com; Johannes Lauer, HERE Technologies, Schwalbach am Taunus, Hesse, Germany; e-mail: johannes.lauer@here.com; Amr Magdy, Computer Science and Engineering, University of California Riverside, Riverside, CA, United States; e-mail: amr@cs.ucr.edu; Mario Nascimento, Khoury College of Computer Sciences, Northeastern University, Vancouver, Canada; e-mail: m.nascimento@northeastern.edu; Siva Ravada, Oracle, Nashua, NH, United States; e-mail: siva.ravada@oracle.com; Matthias Renz, Computer Science, Kiel University, Kiel, Schleswig-Holstein, Germany; e-mail: mr@informatik.uni-kiel.de; Dimitris Sacharidis, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Bruxelles, Belgium; e-mail: dimitris.sacharidis@ulb.be; Flora Salim, School of Computer Science and Engineering, University of New South Wales, Sydney, New South Wales, Australia; e-mail: flora.salim@unsw.edu.au; Mohamed Sarwat, Arizona State University, Tempe, AZ, United States; e-mail: msarwat@asu.edu; Maxime Schoemans, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Belgium; e-mail: maxime.schoemans@ulb.be; Cyrus Shahabi, Computer Science, University of Southern California, Los Angeles, California, United States; e-mail: shahabi@usc.edu; Bettina Speckmann, Technische Universiteit Eindhoven, Eindhoven, Noord-Brabant, Netherlands; e-mail: b.speckmann@tue.nl; Egemen Tanin, University of Melbourne, Australia, Melbourne, Australia; e-mail: etanin@unimelb.edu.au; Xu Teng, Iowa State University, Ames, IA, United States; e-mail: xteng@esri.com; Yannis Theodoridis, University of Piraeus, Piraeus, Greece; e-mail: ytheod@unipi.gr; Kristian Torp, Aalborg Universitet, Aalborg, Denmark; e-mail: torp@cs.aau.dk; Goce Trajcevski, Iowa State University, Ames, IA, United States; e-mail: gocet25@iastate.edu; Marc van Kreveld, Information and Computing Sciences, Utrecht University, the Netherlands, Utrecht, Netherlands; e-mail: m.j.vankreveld@uu.nl; Carola Wenk, Department of Computer Science, Tulane University, New Orleans, LA, United States; e-mail: cwenk@tulane.edu; Martin Werner, School of Engineering and Design, Techincal University of Munich, Munchen, Germany; email: martin.werner@tum.de; Raymond Wong, The Hong Kong University of Science and Technology, Hong Kong, Hong Kong; e-mail: raywong@cse.ust.hk; Song Wu, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Bruxelles, Belgium; e-mail: song.wu@ulb.be; Jianqiu Xu, College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China; e-mail: jianqiu@nuaa.edu.cn; Moustafa Youssef, Computer Science and Engineering, The American University in Cairo, Cairo, Cairo, Egypt; e-mail: moustafa.youssef@gmail.com; Demetris Zeinalipour, Department of Computer Science, University of Cyprus, Nicosia, Cyprus; email: dzeina@cs.ucy.ac.cy; Mengxuan Zhang, School of Computing, Australian National University, Canberra, ACT, Australia; e-mail: mxzhang@IASTATE.EDU; Esteban Zimányi, Data Science Lab, Université Libre de Bruxelles École polytechnique de Bruxelles, Bruxelles, Bruxelles, Belgium; e-mail: esteban.zimanyi@ulb.be.

CCS Concepts: • **Information systems**; • **Computing methodologies** → **Artificial intelligence**; **Parallel algorithms**; **Machine learning**; • **Applied computing** → **Computers in other domains**; **Transportation**;

Additional Key Words and Phrases: Spatiotemporal data, Geospatial intelligence, GPS data, Mobility Patterns, Environmental impacts, Urban Mobility

ACM Reference Format:
Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, Jussara Almeida, Taylor Anderson, Walid Aref, Gennady Andrienko, Natalia Andrienko, Yang Cao, Sanjay Chawla, Reynold Cheng, Panos Chrysanthis, Xiqi Fei, Gabriel Ghinita, Anita Graser, Dimitrios Gunopulos, Christian S. Jensen, Joon-Seok KIM, Kyoung-Sook Kim, Peer Kröger, John Krumm, Johannes Lauer, Amr Magdy, Mario Nascimento, Siva Ravada, Matthias Renz, Dimitris Sacharidis, Flora Salim, Mohamed Sarwat, Maxime Schoemans, Cyrus Shahabi, Bettina Speckmann, Egemen Tanin, Xu Teng, Yannis Theodoridis, Kristian Torp, Goce Trajcevski, Marc van Kreveld, Carola Wenk, Martin Werner, Raymond Wong, Song WU, Jianqiu Xu, Moustafa Youssef, Demetris Zeinalipour, Mengxuan Zhang, and Esteban Zimányi. 2024. Mobility Data Science: Perspectives and Challenges. *ACM Trans. Spatial Algorithms Syst.* 10, 2, Article 10 (June 2024), 35 pages. https://doi.org/10.1145/3652158

1 INTRODUCTION

The volume of mobility data being collected has been steadily increasing since the advent of affordable personal location-enabled mobile devices. Examples of mobility data continuously generated and collected in huge volumes include (a) individual sporadic locations obtained from mobile app data and location-based social networks; (b) individual pedestrians, biking, or driving trajectories constrained by underlying sidewalks, biking trails, and road networks, respectively; (c) indoor individual or asset tracking data obtained from RFID and Bluetooth devices; (d) athletes' movement data in various sports obtained from wearable devices; (e) public transportation, taxis, ride sharing, and delivery logistics trajectories obtained by location-tracking devices and specially designed app services; (f) aircraft and vessel trajectories moving in an unconstrained environment (i.e., no underlying road network) obtained by air and sea traffic monitoring services; and (g) animal tracking data moving freely in the space obtained from physically tagged and remotely sensed animals. Generally speaking, for each moving object, mobility data is typically available in the form of a sequence of (*location*, *timestamp*) pairs. The *location* attribute could be as simple as a point, represented by either *latitude* and *longitude* coordinates or as relative coordinates with respect to the underlying space. The *location* attribute could also be an area, which can represent the mobility of objects with spatial extents, e.g., flocks or group movement.

The ability to understand and analyze mobility data is crucial for various widely used important sectors and applications. In transportation and traffic management, analyzing traffic data through vehicle mobility helps in predicting accidents [158], traffic congestion [258], and better route planning [51]. In ride sharing and delivery logistics application, analyzing trip mobility data helps in data-driven eco route planning, which results in huge cost and energy savings [96]. In location-based services, analyzing people movements around the city significantly helps in trip planning activities [217], finding popular tourists sites and restaurants [118], and data-driven routing and querying [218]. In indoor navigation, understanding how people move indoors helps in understanding the traffic for various stores inside a mall, which is needed in various market research studies [114]. In urban planning, driving data can significantly help in building highly accurate, reliable, and annotated maps [159] as well as deciding on good locations for various facilities, e.g., restaurants, retail stores, and clinics [206]. In social computing, analyzing how people move in cities and regions helps in understanding the demand for infrastructure and energy

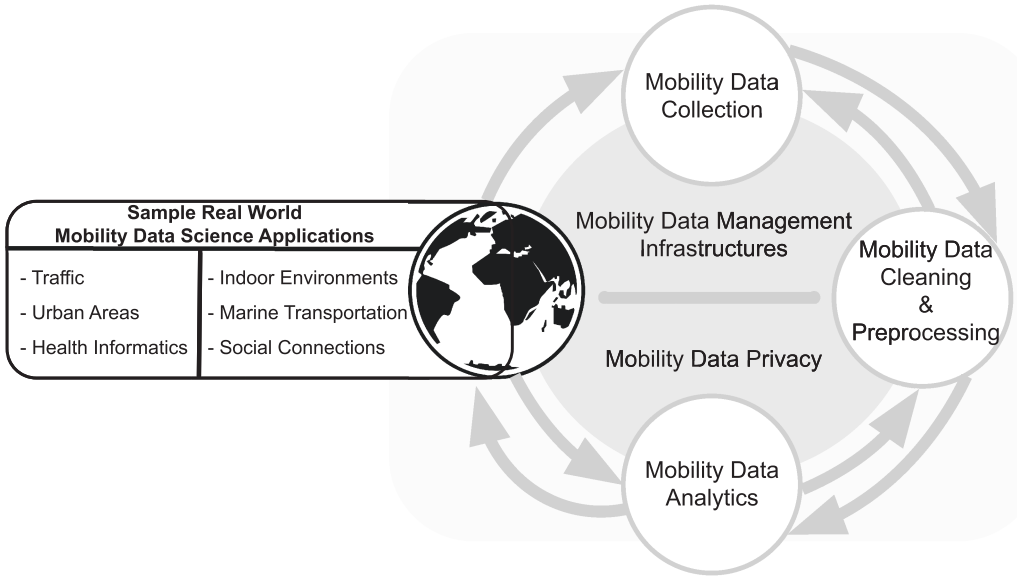

Fig. 1. The mobility data science pipeline.

as a means of reducing inequalities [200]. In disaster response, analyzing crowd movement helps in preparing for natural disasters through rescuing and evacuation efforts [105]. In health informatics, connected wearables can monitor and analyze the movement of elderly people, allowing for timely, and potentially life-saving, interventions [134]. In pandemic prevention, privacy-preserving individual tracking allows for contact tracing, which was deemed to be a cornerstone in limiting pandemic spread [155, 277].

Despite the common goal of acquiring, managing, and generating insights from mobility data, the mobility data science community is largely fragmented, developing solutions in silos. It stems from a range of disciplines with expertise in moving object data storage and management [99], geographic information science [88], spatiotemporal data mining [210], human mobility modelling [27], ubiquitous computing, computational geometry, and more. The sheer volumes of mobility data along with the immense need of mobility data analysis in various applications call for employing a complete Data Science pipeline [190] over mobility data (Figure 1). This includes the whole pipeline of Data Science applications, starting from the data storage and management infrastructure and going through data collection, data cleaning and preprocessing, and data analysis. Unfortunately, this is not straightforward as current Data Science systems, tools, and algorithms are not directly applicable to mobility data. This is mainly due to the fact that these systems, tools, and algorithms are designed in a generic way to support any data type and, hence, they do not lend themselves to the distinguishing characteristics of mobility data. Examples of such characteristics include the spatial and temporal dimensions of the data, the rate of updates, and the privacy requirements. In particular, mobility data is always spatial, in which nearby objects are more related to each other. This is unlike traditional data, in which the concepts of *nearby* and *locality* are not taken into account. Also, similar to time series data, mobility data is temporal, in which one object may have hundreds of updates to its location and all updates are related to each other (e.g., one trajectory). This is again unlike traditional data, in which temporal updates of a single object are not frequent and older updates would be of less importance. Similar to streaming data, mobility data has a high frequency of updates, which is not supported in

typical data science applications. Finally, mobility data is more sensitive to privacy. While privacy preserving in traditional data can be achieved by removing (quasi-)identifier attributes, in mobility data, locations by themselves are considered private information that can reveal not only the users' identities, but also their behavior, lifestyle, medical conditions, and workplaces.

Motivated by ubiquity and sheer volume of mobility data, the importance of mobility applications, and the lack of support from current data science pipelines, this article presents a pipeline for Mobility Data Science. We define *Mobility Data Science* as an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract or extrapolate knowledge and insights from potentially noisy structured and unstructured mobility data, and apply knowledge from mobility data across a broad range of application domains. While currently, the community of developers, practitioners, and researchers dealing with mobility data use off-the-shelf data science techniques and systems to collect, clean, manage, and analyze their mobility data, we firmly believe that this leads to sub-bar performance. We urge this community to build its own mobility data science pipeline to better serve its own purpose. This article makes the case for the need for a mobility data science pipeline along with presenting the challenges that need to be addressed to realize it.

2 MOBILITY DATA COLLECTION

The abundant availability of real data is a cornerstone to any data science application, and mobility data science applications are no exception. However, it is much easier to collect tons of data for data science applications than is the case for mobility data science. In particular, for data science applications, well-established research in anonymizing personal data allows wide data sharing. This is to the extent that governments have released various datasets for public use (e.g., Data.gov). In addition, companies already collect their own inventory data that does not include any personal identifiers; hence, it is suitable to be fed to data science applications. On the other side, data-driven mobility data science research has been in a constant struggle with the need for available mobility data. A main reason is that non-aggregated individual human location data is considered personal identifiable information as it may lead to tracing an individual's identity. For example, it has been shown that only a few spatial locations are sufficient to uniquely identify individuals even among a large population of people [202]. As a result, most datasets are collected in aggregated form, which hinders the deployment of various mobility data science applications. This section discusses current efforts and challenges of mobility data collection.

2.1 Efforts in Mobility Data Collection

Before the wide availability of personal digital devices, human mobility data collection was expensive; therefore, datasets were very sparse. With the advent of personal location-enabled devices, many people's movements have started leaving digital traces that are being collected either by industry as a means of providing location-based services [196] or by governmental entries as a means of data analysis, e.g., traffic-related studies [232]. However, this did not result in a similar explosion of publicly available mobility data, mainly due to privacy and data-sharing concerns.

Current efforts in releasing public non-aggregated mobility data are mainly limited to small datasets and small regions, while removing locations that can lead to one's whereabouts. They mostly include trips obtained from taxis, ride-sharing services, or public transportation. Some of these datasets include detailed trajectory data for the following cities (ordered alphabetically): (1) *Athens* [28]. 500K trajectories collected over 5 days in downtown Athens, Greece; (2) *Beijing 1* [271]. 17+K trajectories with 26 million GPS points over 3 years in Beijing, China; (3) *Beijing 2* [259]. 10+K trajectories with 15 million GPS points over 1 week in Beijing, China; (4) *Rio* [69]. 12+K buses with detailed trajectories of 118+ million GPS points over 30 days in Rio de Janeiro, Brazil;

(5) *Rome* [41]. 320 taxis with detailed trajectories of 21+ million GPS points over 30 days in Rome, Italy; (6) *San Francisco* 1 [179]. 536 taxis with detailed trajectories of 11+ million GPS points over 30 days in San Francisco, California, (7) *San Francisco* 2 [1]. 20+K detailed trajectories with 5+ million GPS points in San Francisco, California; (8) *Shenzhen* [238]. 664 taxis with detailed trajectories of 1.1+ million GPS points over 1 day in Shenzhen, China; (9) *Singapore* [107]. 84K trajectories with 80+ million GPS points over 1 month in Singapore. Other datasets only include the origin and destination of each trajectory. Examples include the following cities: (1) *Austin* [192]. 1.5 million trips for a period of 10 months in Austin, Texas, (2) *Guangdong* [256]. 2.5 million trips over 1 day in Guangdong Province, China; (3) *New York City* [168]. 1.5 million taxi trips over a period of 6 months in New York City, New York; (4) *Porto* [180]. 426K taxi trips over 3 months in Porto, Portugal,

Other than trip and trajectory road network data, there are tons of available biking data across the world, including tens of millions of trips in the Bay Area [139], Boston [138], Chicago [70], Columbus [62], London [135], Los Angeles [38], Madrid [141], Minneapolis [164], New York City [60], Philadelphia [111], Portland [39], and Washington D.C. [44]. There are also available public marine traffic datasets that include detailed vessel trajectories (e.g., [176]), sport datasets for basketball and soccer that include a variety of events that took place in major leagues within one season [174], and indoor data about the behaviors of nearly 30 students in Grade-10 and their teachers collected over four weeks in Australia, with spatial reference (associations to rooms) and highly granular wearable data [83].

However, there are some large-scale aggregated datasets with a coarse granularity that can help in high-level analysis, but not to get insight details of mobility data. Examples of such aggregate data include origin–destination employment statistics in the United States that contain home-to-work commuting flows aggregated to the census tract level [90], cell phone trace datasets capturing the locations of individuals aggregated to their nearest cell tower [235], foot traffic data of check-ins of 35 million anonymized mobile devices in United States aggregated to census block groups [197], and a global database about aggregate indoor occupant behavior, composed of 34 datasets from 15 countries and 39 institutions, collected by occupancy sensors that measure the occupancy count of each space being monitored [72]. An additional source of human mobility data is **location-based social network (LBSN)** data. LBSN data captures both (1) discrete check-ins between users and locations and (2) a social network between users. This dimension of location bridges the gap between the physical world and online social networking services [269]. However, it has been shown that existing LBSN datasets are too small to broadly understand, analyze, and predict human behavior [126].

The lack of available mobility data, combined with the need to stress test various research ideas, has motivated various research groups to either develop their own data simulators or develop publicly available simulators that can also be used by other researchers for benchmark datasets. However, such simulators were mainly designed to test specific aspects of research and not meant to be representative of real mobility data. For example, various simulators were mainly designed to test new index structures for mobility data, query processing algorithms, and system infrastructure scalability for managing spatiotemporal data (e.g., [153]). Within the transportation community, more fine granularity simulators (e.g., [34]) were proposed to study traffic infrastructure, but none of them is meant to provide a comprehensive mobility study.

2.2 Challenges in Mobility Data Collection

This section presents some of the challenges in mobility data collection that the community needs to address towards realizing the pipeline of mobility data science.

Challenge 1. Mobility Data Privacy. In most cases, (human) mobility data is sensitive and considered to be personal identifiable information. This raises major privacy concerns regarding data

sharing. Hence, any attempt to collect fine-granularity detailed trajectory or human mobility data must first address the privacy challenge. Though the general topic of data privacy has been well studied in literature with practical solutions, such solutions are not directly applicable to the case of mobility data. In particular, mobility data gives rise to the **Trajectory-User Linking (TUL)** problem [85]. To protect users' actual locations while preserving meaningful mobility information for various learning tasks, one may wish to generate realistic motions based on real-world mobility datasets [272]. Since privacy is a core problem in mobility data that does not only impact data collection but also impacts all other components of the mobility data science pipeline, we dedicate Section 6 to discussing mobility data privacy in detail.

Challenge 2. Mobility Data Bias. Mobility data collection procedures suffer from all kinds of bias. For example, mobile application data and mobile phone network data are biased against people who do not use smartphones or use prepaid plans. Most traffic counting sensors are installed to count cars but do not count pedestrians, cyclists, wheelchairs, or similar modes of transport. Cells in mobile phone networks vary widely in size. The data traces that are usually collected in cellular networks are cellular themselves. This affects rural areas with larger cells more than urban areas. Volunteered tracking data is biased towards technically savvy people. Sports tracking data is biased towards health-conscious members of the middle and upper classes. It is important to understand, measure, and mitigate data bias in mobility datasets to ensure that actions and policies that are based on mobility data science results are equitable, fair, and include vulnerable populations [205].

Challenge 3. Incentives for Data Sharing. Users need to have good incentives to share their locations. To some degree, users agree to share their locations with commercial entities to get location-based services, ride sharing, cell phone coverage, delivery, and other services. However, it is understood that users would be reluctant to publicly share their mobility traces. Conversely, the biking community have shown a great affinity for sharing their biking trails. A main reason is that, in many places of the world, most of these trails are not really home-to-work commuting, but it is more of an outdoor activity. Hence, sharing biking trails helps fellow bikers in knowing the conditions of biking trails, which is a great incentive for sharing. More incentives need to be offered for drivers to share their mobility traces, even for sporadic trips that do not lead to identifiable locations. Sharing could be for part of the trajectory, where rewards are given back based on the sharing length and resolution. A gamification concept may be exploited to encourage more participants to share.

Challenge 4. Simulated Mobility Data. The dire need for mobility data along with the difficulty of obtaining it made it apparent that simulated synthetic data is immensely needed to enrich and train mobility data science applications. However, the challenge is to go beyond earlier attempts of simulating data for testing very specific techniques to simulating data for the general purpose of having realistic life scenarios. Empowered by modern computational capabilities that make it possible to simulate large populations, the mobility community should work with social scientists to create realistic individual-level human mobility data. Lessons have been learned from the experience of the deep learning community by applying **generative adversarial networks (GANs)** for trajectory generation [262]. However, it is unclear as of yet how to measure the extent to which mobility data is realistic. If synthetic mobility data is too realistic, for example, due to training on real human trajectories, it may invade someone's privacy if, for instance, it shows where members of a given household actually visit. On the flip side, benchmark data that is too disconnected from the real world and does not represent realistic human behavior would not allow generalization to the real world.

3 MOBILITY DATA CLEANING

Until the early 21st century, location data and mobility data available for **geographic information science (GIS)** was mainly collected, curated, standardized [78, 79], and published by authoritative sources such as the **United States Geological Survey (USGS)** [231]. Now, data used for mobility data science is often obtained from sources of **volunteered geographic information (VGI)** [216]. Such data is contributed by millions of individual users (more than 10 million contributors in the case of OpenStreetMap [170]) and is rarely curated. Mobility data collected from such sources is highly uncertain due to physical limitations of sensing devices, due to obsoleteness of observations, and in many cases is simply incorrect due to deliberate misinformation [157]. Consequentially, our ability to unearth valuable knowledge from large sets of mobility data is often impaired by the uncertainty of the data such that geography has been named the "Achilles heel of GIS" [89].

Data cleaning and preprocessing is a milestone to all data science. In fact, it has been reported that data scientists spend more than 80% of their time in data cleaning and preparation [162]. As a result, there are huge efforts in the data science community dedicated to developing various data cleaning algorithms [57] and full-fledged systems [67]. Mobility data is of no exception in terms of its need for data cleaning and preparation procedures. However, for numerous reasons, data cleaning and preparation yields unique challenges. This section discusses current efforts and challenges of mobility data cleaning.

3.1 Efforts in Mobility Data Cleaning

A recent survey [125] and data quality assessment tool [91] have discussed various sorts of errors that negatively impact data quality in spatial and mobile environments. Motivated by the inaccuracy of location tracking devices, several efforts were dedicated to address (a) the spatial inherent inaccuracy of GPS devices and (b) the uncertainty of moving object whereabouts between two known locations, which is a result of low sampling rates due to bandwidth and battery limitations.

As the spatial inaccuracy indicates erroneous GPS coordinates, the efforts to identify and correct such coordinates have focused on either finding and eliminating outliers or map matching all coordinates to an underlying fixed and trusted infrastructure (e.g., road network map). For the case of map matching, existing efforts aim to match/snap all GPS traces to an underlying road network [42, 46]. Proposed techniques vary from as simple as snapping each point to its nearest road to applying Markov Chain to identify the most probable road segment that each point should be snapped to. In the case in which there is no underlying road infrastructure (e.g., marine transportation or animal movement), outlier detection techniques are used to identify and remove erroneous points [224].

Irrespective of the collection method and device settings, there is also indispensable uncertainty in movement data caused by their discreteness. Since time is continuous, the data cannot refer to every possible instant. For any two successive instants, there is a temporal gap in which the whereabouts of the moving objects are unknown. To overcome such location uncertainty, several efforts were dedicated to modeling the uncertainty of mobility data surveyed in [278].

3.2 Challenges in Mobility Data Cleaning

This section delves into some challenges linked to cleaning mobility data that the community needs to tackle.

Challenge 5. Inaccuracy in the Movement Space Infrastructure. A unique challenge in mobility data is that, in many cases, its reference points are the ones that are inaccurate. In particular, mobility data that represent movement on a road network may be more accurate than the road

network itself. Road networks, like any other type of data, suffer from all sorts of inaccuracy and may not even be available in many places [160]. In fact, Microsoft has recently announced that it has found more than 1 million kilometers of roads missing from current maps [148]. This is why there is a whole area of industrial and academic research about map inference, which aims to infer (all or missing parts) of the road network from either satellite images [29] or trajectory data [37]. However, almost all of these techniques focus on making accurate maps in terms of topology. There need to be more efforts to develop map inference algorithms that go beyond inferring the map topology to inferring map metadata (e.g., road speed, traffic lights, number of lanes, and turns), without which mobility data would not be accurate as its road network reference itself is missing important data. A major step towards cleaning mobility data would be to first clean its reference map.

Challenge 6. Filling in Temporal Mobility Gaps. As mentioned earlier, there are lots of efforts dedicated to modeling the uncertainty of moving objects' whereabouts between two consecutive time instances. However, uncertainty poses different challenges to downstream functions and applications, including the need to develop new techniques for indexing, query processing, and data analysis for various uncertainty models. One way to overcome this is to try to infer the actual whereabouts of a moving object between any two time instances with known locations. There are already several efforts to insert artificial points between two consecutive trajectory points, with the promise that these points act as if the trajectory was collected in a very high sampling rate. This process has various names, e.g., *trajectory interpolation* [136, 268], *trajectory completion* [130], *trajectory data cleaning* [261], *trajectory restoration* [124], *trajectory map matching* [42], *trajectory recovery* [243], and *trajectory imputation* [76]. However, the large majority of such work relies on matching the trajectory points on the underlying road network, where the imputation becomes finding the road network's shortest path between two consecutive trajectory points. Unfortunately, this is not applicable to the case in which the road network is unknown, untrusted, or inaccurate. Hence, more recent attempts try to do data-driven trajectory imputation without relying on the underlying road network [76, 80]. However, these techniques are either not scalable to city-scale trajectory datasets or require dense historical data that derives its imputation process. There is an immense need to develop a scalable, accurate, and fine-grained imputation that almost mimics a continuous datastream of trajectory locations.

4 MOBILITY DATA ANALYTICS

Spatial data is special. Unlike non-spatial features, location attributes (e.g., longitude and latitude) rarely exhibit linear or other simple functional relationships to variables of interest. It rarely makes sense to model a variable of interest directly in relation to spatial attributes. Instead, it is distances that matter. According to Tobler's first rule of Geography, "everything is related to everything else, but closer things are more related than things that are far apart" [221]. For mobility data, proximity is further extended with time, i.e., objects that are close in space and time. In addition to this concept of spatiotemporal autocorrelation, what makes mobility data even more challenging to handle is that it is often observed from humans whose behavior can often be irrational and difficult to explain. As Nobel Prize laureate Murray Gell-Mann famously said, "Think how hard physics would be if particles could think" [172]. However, unlike in physics, the "particles" of interest are often humans who can think. Data collection sensors have the capability to capture the spatiotemporal locations of moving objects, but not their behavioral aspects. These difficulties require new paradigms, techniques, and algorithms to analyze and learn from the spatiotemporal data and that can explain and predict the associated behavior. This section discusses current efforts and challenges of mobility data analysis.

4.1 Efforts in Mobility Data Analytics

Mobility data analytics has already gained momentum in research in recent years. Dedicated workshops have existed in major conferences, including the ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial) since 2011 [209], the **Big Mobility Data Analytics (BMDA)** workshop in EDBT since 2018 [177], and the ACM SIGSPATIAL International Workshop on Animal Movement Ecology and Human Mobility (HANIMOB)@SIGSPATIAL since 2021 [171]. Surveys on the status of research exist [20, 198].

Mobility data analytics encompasses various application domains and involves analyzing data from different sources such as urban [265], maritime [61], aviation [59], animal movement [171], and indoor movement [114]. Among these different themes, urban mobility stands out with a fairly large body of research, including green routing [10], traffic anomaly detection [173], hot spot and hot path analysis [166], road traffic prediction [161], and travel time estimation [240]. Trajectories of moving objects have been used as means to create and continuously update the road network [159]. Public transport systems also collect ticketing data in the form of passenger check-ins, sometimes also associated with check-outs. This data has been shown to be very useful to transit planners in understanding passenger demand and movement patterns in daily operations as well as in the strategic long-term planning of the network [227]. Personal mobility of individuals is also a subject of analysis that includes, e.g., activity recognition [50, 175], personalized routing [66], matching with ride-sharing services [19], and crowd-sourcing [178].

While a significant portion of research focuses on understanding and analyzing data through analytics, there are also important efforts dedicated to developing generic analysis tools for spatiotemporal data that are agnostic to the application domain. Efforts regarding generic methods for mobility data analysis include, among many others, trajectory clustering [244], trajectory similarity measures [224], outlier detection [101], transportation mode classification [40], spatiotemporal pattern detection [199], and trajectory completion [121]. However, and despite these many research efforts towards analyzing mobility data, there is a lack of common data analysis tools and systems. The scientific software environment for mobility data analysis is rather fragmented. For example, [117] lists 58 packages in their review of R packages for movement and [92] reviews Python libraries for movement data analysis and visualization.

Recent years have seen a notable increase in research on deep learning for mobility data analysis [137, 250]. This brought an increased adoption of various paradigms and (adapted versions of) architectures used in other areas in whic deep learning has brought improvements in tasks, e.g., clustering/classification [149], prediction [122] and recommendation [30], information propagation [274], etc. For example, **Generative Adversarial Network (GAN)**–based architectures have been used recently to learn representations of trajectories and generate synthetic trajectory techniques [84]. Given the introduction of Transformers [233], transformed-based approaches have also been used for mobility modelling and trajectory prediction [254] given the sequential properties of mobility data. Other deep learning approaches, such as contrastive learning [273], have also been exploited in mobile data settings, along with investigation of the impact/benefits of representation learning [86].

4.2 Challenges in Mobility Data Analysis

This section highlights open problems related to mobility data analysis that need consideration from the community.

Challenge 7. ***Machine Learning (ML)*** *for Mobility Data.* The state-of-the-art **deep learning (DL)** models, such as Transformers [233], were not developed initially for mobility data science in mind. They were derived from **natural language processing (NLP)** and computer vision

domains. The community needs to provide best-case practices for doing ML (and DL) for mobility data.

A major hurdle, and a research opportunity as well, is that existing ML and analytics tools, e.g., TensorFlow and PyTorch, do not support location and mobility as base data types to reason about. Thus, even the basic analysis, such as clustering, classification, and similarity, need to be extended when mobility data is involved. These tasks, as well as higher-level analysis, cannot be totally independent. Instead, common basic building blocks could have an impact on all or some of them. For example, exploring the effectiveness of embedding for mobility data analysis is a basic block that could impact different ML-based analysis tasks. This raises a challenge to build analysis primitives and common building blocks for applications that could shape a framework of ML-based mobility data analysis.

Another major hurdle is the robustness in data-driven mobility models. It is widely known that data-driven models (as in the case of ML and DL) are only as good as the data used to train them. However, given the changes in mobility behaviors, such as during the COVID-19 pandemic and the associated lockdowns, and environmental events and disasters, traditional ML-based, and even recent DL-based, methods are no longer robust. The models' performances deteriorate in unseen events, especially as new behaviors emerge and then persist. Recent effort includes the incorporation of 'contextual awareness' and 'memory' in an enhanced event-aware spatiotemporal network [245] for predicting mobility in multiple modes of transportation, including taxis, cycling, and subways during the unprecedented events such as COVID lockdowns or snowstorms as events emerged and up to 30 days post the event. However, more work needs to be done on modelling and understanding mobility behavior that is robust to changes due to societal events.

Challenge 8. Progressing from Next Location Prediction to Movement Behavior Understanding. Due to the wide availability of aggregated check-in and foot-traffic data, many researchers focus on the problem of location prediction, e.g., [253]. Leveraging predictions such as "User X will visit Coffee Shop A next" or "32±4 users will visit Coffee Shop A in the next hour" has some direct applications. It could be useful for providing information about parking ("parking at location X appears to be a problem today, so consider..."), for battery-charging opportunities, or for providing information about collective transportation status ("Metro station X that you are expected to visit is closed for repairs, so instead..."). One could provide a new transportation schedule and departure time in response to problems at an anticipated future location of a user, just like airlines at times update itineraries in the case of issues. Earlier work has been based on data mining techniques to detect periodic behavior, e.g., [36, 75, 116]. Beyond predicting locations, if we understand the underlying behavior at the individual-, group-, or population-scales that leads to these predictions, we could understand *why* one coffee shop chain has increasing visitor rates (e.g., due to a movement towards organically grown coffee sold by the coffee shop). Through inferring from the data about such behaviors, only then can we take corresponding actions not only to predict locations but also to prescribe actions (e.g., offering more organic coffee) to improve visitor rates. This understanding of (human) behavior will broadly affect applications using mobility data. Traditional spatiotemporal data science allows for predictive analytics to predict the future. In contrast, mobility data science enables prescriptive analytics by understanding the underlying human behavior to devise actions and policies that aim to achieve desirable targets.

An open problem for understanding mobility behavior data is the lack of labels or human annotation to provide insights on the actual observations. There are several other tricks that have been proposed, including cross-domain data fusion as well as developing interpretability mechanisms for ML or DL models. When geographical information is fused with contextual features and social behaviors, not only location prediction can be improved but also insights can be provided

about the underlying visitor behavior [253], even if no human-labelled data are provided about the mobility behaviors.

Therefore, explainability of AI and ML models that have underpinned many of such predictive behavior models remain an open challenge, especially since DL models are black boxes. One such approach for DL-based models is disentangled representation learning, and a recent work [266] shows that the disentanglement of latent spatiotemporal factors can assist the explainability of how the underlying latent factors learned by DL models are correlated. It can also be used for dimensionality reduction and can assist in few-shot learning cases.

Challenge 9. Visual Analytics. Visualization and exploratory analysis of mobility data has long been a hot topic in visual analytics [15]. More recently, the trend turned to combining visualization with modeling and simulation to support decision-making [123]. This kind of research is by necessity application oriented, while much less is done on developing more general ideas and approaches.

One general research problem that has only been slightly touched on in visual analytics but not systematically addressed is human involvement in real-time analysis of big mobility data. Is it possible to define realistic scenarios for involving human intelligence in big data analytics taking into account the cognitive limitations of human analysts with regard to the amount of information that can be perceived, speed of processing, and time required for analytical reasoning and contributing to the analysis process? Also how does one combine computational methods of analysis, such as ML, with human expert knowledge and reasoning? The involvement of human intelligence is limited to thoughtful data preparation, feature selection, parameter setting, and so on. It would be great to find ways to make more direct and effective use of human-possessed concepts and, particularly, knowledge of causal relationships. Hence, a grand research challenge for visual mobility analytics is to develop approaches to understanding and modeling mobility behaviors from low-level movement data, such as trajectories of moving entities.

The following research problem is how to analyze behaviors after they have been extracted from elementary movement data and represented by appropriate data structures. A conceptual framework should be developed to enable defining the types of conceivable patterns of movement behavior. This will provide orientation for developing visualization techniques facilitating visual discovery of behavioral patterns as well as algorithmic methods for detection of specified types of patterns. These techniques and methods should be incorporated into systems and workflows for analyzing the contexts in which various patterns take place and developing models for describing and predicting mobility behaviors depending on the context.

5 MOBILITY DATA MANAGEMENT INFRASTRUCTURE

Classical data management systems have been designed for generic data types, where spatial and temporal data can be supported as new additional types. Yet, the core functionality of the data management engine does not acknowledge the spatial and temporal properties of mobility data. For example, mobility data calls for storing and querying locations of objects that evolve over time. The evolution can be in the location, the extent, and/or the properties of the object. The evolution can happen in discrete steps, e.g., check-ins, or in a continuous form. Thus, it is desired that the data management platform is able to represent the history, the current location, and possibly the near future of the moving object. Another example is classical index structures that are built with the assumption that the read workload is significantly higher than the write workload and, hence, the index structure does not change often. Mobility data exhibits a different workload, in which the write workload (e.g., object location update) is significantly higher than the read workload, which makes all classical index structures simply not applicable to mobility data. A third example

is that simple queries of mobility data, e.g., nearest neighbor search, can be supported by classical data management systems by finding the distance between the user location and all other objects, sorting all objects based on that distance, and getting the closest one. This cumbersome approach is mainly due to the lack of having a specialized nearest-neighbor operator. Should we have one, that operator could seamlessly integrate with the query executor and optimize a data management engine to efficiently support a pretty important query in most data mobility applications. A last example is that classical methods for scaling up data management in distributed environments rely on data distribution, mostly based on the data keys. This does not work well in scaling up mobility data, as it is always desired to distribute mobility data in a way that, spatially and temporally, nearby objects are grouped together in the same cluster or computing node. This section discusses current efforts and challenges of mobility data management.

5.1 Efforts in Mobility Data Management

There has already been extensive research in all layers of mobility data management infrastructure. In terms of data modeling, early models based on the constraint database model aim to support simple moving objects (i.e., points), e.g., [93]. More complex data types (e.g., moving regions) have been supported by later models based on abstract data types, e.g., [100], that are still being used in recent systems, e.g., [276]. More recent efforts have been introduced to capture the semantics of trajectories of moving objects. Other models were also proposed to capture specialized modes of movement, including indoor environments, e.g., [113], network constrained, e.g., [98], fuzzy trajectories, e.g., [225], and detecting periodic moving patterns, e.g., [33, 36, 75, 116]. In terms of indexing, tens of index structures have been proposed to support efficient indexing, storage, and retrieval for spatiotemporal data as either historical data, current locations, or continuously updated locations, e.g., [143, 146, 154, 163]. This forms the infrastructure support for various spatiotemporal query processing techniques for various query operators over moving objects, including spatiotemporal range queries [156], spatiotemporal nearest-neighbor queries, e.g., [11–13, 214, 252], reverse nearest neighbor queries [35], skyline queries [108], and scalable spatial and spatiotemporal joins, e.g., [247, 251].

In terms of academic full-fledged systems, the SECONDO system has been introduced in the early 2000s as a comprehensive testbed for distributed moving object databases covering all aspects of data modeling, indexing, and querying [97]. More recently, MobilityDB, implemented on PostGIS, has been introduced as a scalable system with a wider functionality on moving object databases [228, 276]. In terms of Big Data systems, ST-Hadoop [8], SUMMIT [7] and HadoopTrajectory [22] systems extend the Hadoop system to support spatiotemporal data and trajectories, respectively, while other systems, e.g., [65, 144, 145], extend the Twitter Storm distributed data streaming system to support streamed location data. TrajSpark [264], Dita [207], and TrajMesa [127] extend the Spark system to support various index structures and query operations over trajectory data. SharkDB [242] extends in-memory column-oriented storage engines to support trajectories. In the open-source community and in industry, PostGIS [181] supports very basic trajectory functions and Oracle spatial supports streaming point data to capture real-time mobility [169], whereas Microsoft Azure [25] supports storing trajectory data in Azure table and utilizing Azure Redis for indexing. Distributed-MobilityDB [23] integrates the trajectory data management of MobilityDB with a distributed PostgreSQL database to provide a distributed moving object database.

5.2 Challenges in Mobility Data Management Infrastructure

Though there is already a lot of work in various components of mobility data management infrastructure, there is an apparent lack of integrated systems that offer comprehensive functionality

to end users, encapsulated in full-fledged systems that support mobility data science. Hence, the challenges in this section mainly focus on system building.

Challenge 10. Building Systems with Mobility Data in Mind. Location data has almost always been supported in data systems as an afterthought problem. Many systems, e.g., Postgres, Storm, Spark, and Hadoop, have not been originally designed with location data support in mind. What typically happens is that spatial data types get augmented into tuple-oriented systems to support the location data type. For example, a restaurant tuple that describes various attributes of a restaurant is augmented with the latitude and longitude of the location attribute of the restaurant to support location services. Spatial indexes are provided to speed up the access to these attributes, and some accompanying spatial operators are provided to operate on the location attributes to provide location services, e.g., range or k-nearest-neighbor searches. While this approach works to some extent, systems coming out of this approach end up with sub-par performance for spatial data and, hence, for mobility data. Given the myriad applications that rely on mobility data, it is important that systems are extended with native support for locations and mobility data. Thus, mobility data types and operations should be integrated in the core of these systems and should not be considered as an afterthought problem. This can go through all kinds of systems, starting from database management systems that need to be spatially and temporally aware to support mobility data to scalable big data and NoSQL systems, where injecting spatial and temporal awareness into their core functionality will inherit their scalability to support scalable mobility data science.

Challenge 11. Location Data as First-Class Citizens. Having locations as the core of mobility data calls for treating location data as a first-class citizen in a location data system that at the same time can be extended to support other data types [16]. These location data systems can be presented as Location+X systems, e.g., as in [16], where the data types "X" can be keywords (e.g., to support spatial keywords and tweets), graphs (e.g., to support road-network data), relational data (e.g., to support descriptions of spatial data objects), click streams (e.g., to support check-in data), document data (e.g., to support points of interest and documents that describe them), or annotated trajectories (e.g., location + time + textual annotations), among others. In many location services, more than one data type X may need to be supported, e.g., a graph data type combined with a document or keyword data types, which calls for a multi-model-like data system. This gives rise to an ecosystem where location is at the core with some form of an extensible multi-model data system that supports the multitude of data types "X". However, current multi-model data system technology is lacking in several aspects. First, they do not support data streaming, which is a cornerstone in mobility data due to the online streamed locations of moving objects. Second, we do not want to fall into the trap of adopting existing multi-model technologies that may affect location being a first-class citizen. However, the need for supporting multi-models in one seamlessly integrated location+X system remains a necessity. In addition to supporting location data via a native location+X engine, an ecosystem for mobility data would also include many important utilities to facilitate a broad spectrum of location service applications. From the input data side, to help navigate the vast amounts of available location datasets and discover the right datasets for a given task, a location dataset lake infrastructure and location dataset discovery, cleaning, and integration facilities are needed. From the presentation side, a comprehensive visualization suite is envisioned to support visualizations for combinations of spatial and temporal data analytics on top of location data.

Challenge 12. Streaming, Batch, and Hybrid Workloads. Motivated by the application needs, mobility data management needs to support both batch and real-time data through all system layers, from digesting the data to analyzing and visualizing it. For example, a common requirement is to

visualize the positions of a fleet of vehicles in real time, which only requires access to the most recent positions of the vehicles. Yet, at the same time, there is a need to perform batch analytics on the full trajectory of these vehicles (e.g., to assess whether the trajectories exhibit some unexpected behavior). Generally speaking, the need to have both real-time and historical data has led to the development of the data warehouse domain, where operational databases cover the real-time **Online Transaction Processing (OLTP)** whereas data warehouses cover the historical **Online Analytical Processing (OLAP)**. Since having two different systems for the two kinds of workloads is very costly, a new approach referred to as **Hybrid Transactional and Analytical Processing (HTAP)** has recently been proposed. However, mobility data exhibits significantly different workloads from other data, where streaming data is dominant in terms of objects continuously streaming their new locations. Historical data is not of less importance and is continuously appended. While some efforts have been spent in the direction of write-optimized indexing for location data, e.g., as in [211], more research efforts need to be spent to adopt the concepts behind HTAP systems to support the nature of mobility data.

6 MOBILITY DATA PRIVACY

As we discussed in Challenge 1, mobility data privacy is a core problem in the mobility data science pipeline. Studies have shown that location data could reveal sensitive personal information, such as home and workplace, and religious and sexual inclinations [183]. As localization technology advances and extremely fine-grained location tracking is being enabled, it may even reveal products of interest in the stores we have visited, doctors we saw at a hospital, bookshelves of interest in a library we have visited, artifacts we observed in a museum, and generally anything that might publicize our preferences, beliefs, and habits. A recent survey has shown that 78% of smartphone users among 180 participants believe that apps accessing their location pose privacy threats [47].

While there are many privacy-preserving data collection and data analysis techniques developed for personal data, mobility data introduces unique challenges due to (1) spatiotemporal correlations in the mobility data, which often results in increased privacy cost due to privacy composition for correlated data or downgraded utility for downstream applications; (2) complex location semantics (e.g., corresponding points of interest of locations) and mobility behaviors (e.g., regular vs. one-time visit of a location) that existing privacy definitions may not be able to capture; and (3) diverse and emerging application scenarios, such as contact tracing using mobility data for which existing privacy algorithms designed for aggregate data analytics are not suitable. In this section, we briefly review existing privacy notions and techniques developed for location and mobility data and discuss several open challenges.

6.1 Efforts in Mobility Data Privacy

We categorize existing techniques in mobility data privacy into two main settings corresponding to our data pipeline: (1) local setting (data collection stage) and (2) central setting (data analysis stage). In the local setting, the mobility service provider that collects mobility data is assumed to be untrusted. Hence, each mobile user or entity can apply privacy-preserving mechanisms before the data is collected by the service provider. In the central or global setting, the mobility service provider is assumed to be trusted and collects the raw mobility data. The provider can apply privacy-preserving mechanisms for statistical analysis and share aggregated data, ML models trained from the data, or synthetic data mimicking the original data with untrusted third parties.

Local Setting. In recent years, **local differential privacy (LDP)**, the local variant of differential privacy [63, 94], has become the de facto standard for preserving privacy at the data collection

stage. Users can perturb their raw data using an LDP mechanism before uploading it to an untrusted server. Most existing mechanisms are designed to ensure utility for aggregate queries or analytics (e.g., frequency or density estimation), which requires the aggregation of the perturbed values from a large group of users, whereas the individual perturbed value may not provide much utility. Several works applied existing LDP schemes to location data but the utility is poor [119, 267]. Other works relaxed LDP to personalized LDP [52]. Recent works developed improved LDP mechanisms for location data with better utility [239].

In addition to supporting aggregate data analytics, **location based services (LBSs),** including range queries, spatial crowdsourcing, and the emerging contact tracing for pandemic control, require the precision of the perturbed locations themselves. **Geo-indistinguishability (GeoInd)** [14] relaxes LDP for location data, which requires the locations to be indistinguishable only within a radius and the indistinguishabilty is scaled by their distances, providing a better privacy utility trade-off for LBSs. Later works extended GeoInd to account for temporal correlations between consecutive locations of mobile users [249] and protection of customizable spatiotemporal activities instead of raw locations or trajectories [43]. Other works applied the GeoInd mechanisms and variants for privacy-enhanced spatial crowdsourcing and contact tracing [64, 220]. Besides statistical privacy techniques, **Private Information Retrieval (PIR)** and secure **multiparty computation (MPC)** techniques have also been developed to allow LBS queries such as range queries and contact tracing without revealing individual locations [6, 56, 87, 186] but are generally more computationally expensive and need to be designed for each different query.

Global Setting. Many works have applied **differential privacy (DP)** for computing and publishing aggregate mobility data. Compared with DP algorithms for tabular data, they typically exploit the hierarchical structure of locations and sequential patterns of trajectories to improve utility [2, 49, 150, 184, 204]. Some works also utilized the DP aggregates for task assignment in spatial crowdsourcing [219]. In practice, mobility data providers have started sharing aggregated mobility datasets with DP, especially in response to the pandemic, such as Meta's population density maps and Movement Range maps, Google's COVID-19 Community Mobility Reports, and SafeGraph's Patterns [24]. Other works have applied DP for training ML models using mobility data, for example, for location prediction [5]. Another line of work attempts to generate synthetic trajectories or mobility data based on raw trajectories with formal DP guarantees [103, 241]. From the privacy attack side, recent works demonstrated the possibility of membership inference attacks on aggregate location data and linking attacks, and the defense power of DP against some of these attacks, reinforcing the need for ensuring rigorous privacy even for seemingly anonymous aggregate mobility data and ML models trained from mobility data [115, 182].

6.2 Challenges in Mobility Data Privacy

This section highlights open problems related to mobility data privacy that need consideration from the community.

Challenge 12. Threat Models and Privacy Definitions. The first challenge for mobility data privacy is the need to understand the threat models and adopt or define proper criteria by which to enforce privacy. We need to define first what needs to be protected (i.e., the sensitive information). This may vary for different mobile users and applications. It may be the exact location coordinates of a user at a given time (most existing efforts focus on this). It may also be the association of a user with a sensitive place, co-location of two users (while it's okay for the users to reveal the exact location coordinates), or spatiotemporal activities of a user (e.g., stay at a place, or a trajectory). When defining privacy models and designing subsequent privacy mechanisms, there will almost always be attacks based on side channel information exploitation. While privacy notions such as

DP typically assume the worst case, which also means sacrificed utility, relaxed versions may be needed given specific threat models to enhance the privacy and utility trade-off.

Besides developing rigorous privacy-enhancing mechanisms, it is equally important to understand the privacy risks and the empirical defense power of **privacy-enhancing technology (PET)**. While there has been some work on privacy attacks on aggregate mobility data [182], more work is needed to understand what sensitive information may be revealed and reconstructed from mobility data-based models, e.g., whether membership inference attacks or feature reconstruction attacks [81, 212] can be carried out and potentially build benchmark attacks that can be used to audit the privacy risk of mobility data science systems and privacy mechanisms.

Challenge 13. Privacy and Utility Trade-off and Other Factors. When designing privacy mechanisms for mobility data collection and analysis, it is important to consider the utility of the privacy protected data for the downstream applications. For LBS (as typical in the local setting), the utility needs to be measured by the precision or accuracy of range queries for POI search, or contact detection for contact tracing (instead of how accurate the perturbed location is from the original location for which most algorithms following GeoInd are focused on). Hybrid methods that combine DP and cryptographic techniques may be needed, especially for critical applications such as contact tracing and public health [56]. For aggregate data analytics and ML applications using mobility data (in both local and global settings), the utility need to be measured by the accuracy of the statistics (e.g., frequency or density estimation for which most existing work focuses on), the trained model, or the fidelity of the synthetic data. As a result, the algorithms need to be designed to optimize the corresponding utility and many remain an open challenge. For example, existing methods for DP trajectory synthesization are mainly based on statistical models or low-order Markov models and perform well on some utility metrics [103, 241]. While there are more powerful **generative adversarial network (GAN)**—based models or diffusion models for generating more realistic synthetic trajectories [137, 275], ensuring formal DP for these models would result in deteriorated utility due to the complexity of the models. Designing methods for optimal privacy utility trade-off remains an open challenge.

In addition to the privacy and utility trade-off, privacy-enhancing technology may exacerbate bias in the data or learning algorithms. Mobility data may have inherent bias, as we discussed in Challenge 2. Data analysis algorithms may also have unfair performance for groups that are underrepresented in training data. It has been demonstrated that learning with DP could exacerbate such unfairness, i.e., underrepresented groups suffer from worse privacy/utility trade-offs [21]. Research is needed to understand this impact on mobility data and design privacy algorithms to optimize the privacy utility trade-off while ensuring fairness.

Challenge 14. Explainability and Societal Education. Another important challenge of mobility data privacy is to improve the explainability of privacy definitions and mechanisms and communicate them to the stakeholders, including mobile users (data contributors), mobility service providers, and data analysts. This is a general challenge for privacy-enhancing technology, but more so for mobility data given the complex semantics of location information and diverse applications, as we mentioned. DP-compliant algorithms and location privacy models (such as GeoInd) as described earlier use privacy parameters to control the trade-off between privacy guarantee and the utility of the private outputs. However, there is a significant gap between the theory and practice of DP: we lack principles and guidelines for choosing privacy parameters when collecting or processing mobility data using DP techniques in the real world. While the technology companies have employed DP in releasing the mobility datasets, as we discussed earlier, the choice of the privacy parameter and the associated noise and uncertainty are often

not precisely specified or uniform across companies. This makes it difficult for the downstream applications to quantify the uncertainty of the analysis result.

The parameter ϵ of DP is mathematically defined but not well aligned with the stakeholders' interests. Even for the same ϵ, the privacy guarantees could be different based on the different variants of DP and algorithms at hand. In addition, the ϵ is not always linked to a specific privacy risk for the users (such as "the probability that an attacker can correctly infer my data") or a precise utility level for data analysts (such as "the accuracy of the DP-ML model"). To promote the adoption of mobility data privacy technology such as those based on DP, we should establish principles and design guidelines, and provide tools for explaining DP's protection and limitation from stakeholders' practical interests. For example, we can help data contributors understand the privacy risk (such as membership inference attacks or reconstruction attacks) under different privacy parameters given a concrete DP algorithm. We can also design efficient methods to visualize how data analyzers' utility metrics (such as MSE or model accuracy) may change along with different privacy parameters for specific mobility applications.

7 MOBILITY DATA SCIENCE APPLICATIONS

Mobility data science used to be limited to the domain of transportation. However, recent technological inventions have created an abundance of mobility data, resulting in applications in many other domains of interest for society. Such applications leverage mobility data to understand, explain, and predict where moving entities such as humans, animals, or infectious diseases go, why they go where they go, and where they will go next. This section outlines broad applications of mobility data science to illustrate the recent landscape of mobility data science.

7.1 Traffic

Traffic is a problem of global scale, as recognized by transportation science over a decade ago. Drivers in the United States spend 6.9 billion driving-hours stuck in traffic and waste more than 11 billion liters of fuel per year according to INRIX [112]. Measured per capita, people in Russia and Thailand spend even more time in traffic, whereas Brazil, South Africa, the United Kingdom, and Germany are only slightly behind the United States. Leveraging mobility data science and understanding the underlying behavior of human participants concomitantly with different transportation modes can enable more effective solutions to multiple problems at the heart of improving traffic management. Two main lines of research focus on (1) traffic monitoring at an aggregate level, e.g., to help city administration; and (2) provision of services to road users. Existing work regarding traffic monitoring includes monitoring congestion [128], assessing the safety of roads and intersections [142], traffic prediction [131], evacuation routing [263], and optimizing public transportation schedules [191]. Efforts regarding the services provided to road users include routing queries that balance the traffic across roads [68], helping drivers to find nearest facilities [120], personalized routing [129], eco-routing for minimizing greenhouse emissions [133], and enabling multi-modal trip planning [223]. But there are many open opportunities and challenges in using mobility data to improve traffic conditions. One example is devising accurate models for the dynamic scheduling of public transportation. Another example is the context-aware optimization of traffic signals, e.g., incorporating the impact of additional flux of pedestrians in bus/train stations, to minimize the stop-and-go impacts for vehicles. A challenge of using mobility data science in the transportation domain is monitoring and reduction of emissions. Being able to quantify emissions (e.g., from transportation) is essential to accountability and reduction of emissions. Using data on emissions collected from in-situ sensors but also sensed remotely through earth observation (satellite) data will allow us to better understand the effects of e-mobility, better collective transportation, and infrastructure improvements.

7.2 Urban Areas

In 2018, 55% of the world's population (4.2 billion people) resided in urban areas. This proportion is projected to increase to 68% by 2050 [230]. Urban areas are a focal point for mobility application as they introduce a variety of mobility modalities such as electric vehicles [234] and bicycles and scooters with respective sharing programs [132]. By understanding how, where, and why people move in cities, outer suburban areas, and regional areas, the demand for infrastructure and energy can be better understood [270]. Improving this understanding helps reduce urban inequalities in cities [165] such as access to high-quality food [236] and healthcare [95]. Mobility data also helps improve urban safety by improving crime prediction [82] and helping to recommend safe routes [203].

A specific urban mobility data science supports urban areas through data-driven map construction [3] and updating of existing maps to account for blocked or new road segments [48], which is paramount in autonomous driving applications [140].

The real-time monitoring of urban mobility could result in *situational awareness*, initially a term coined in defense applications, involving *perception* of the environmental states using the surrounding data, *comprehension* of the ingested data to understand the emerging situations, and *projection* of future states and/or events that require predictive analytics. Mobility data provides critical components and insights into situational awareness in cities. When achieved, this applies not only to enabling robust critical infrastructures in cities but also to protecting them from harm, e.g., forest fires, earthquakes, and terrorist attacks. Many researchers use mobility data as input to enable situational awareness in cities as well as in airports [208].

7.3 Health Informatics

The spread of infectious diseases is a highly complex spatiotemporal process that is strongly tied to human mobility [106] and human behavior [74]. Many recent works have used human mobility data for data-driven epidemic forecasting, as surveyed in [195]. A specific example of leveraging mobility data for public health is contact tracing, which refers to the process of tracking persons who may have come into spatial contact with an infected person, and subsequently collecting further information about these contacts [151]. The feature-rich interaction, processing and localization/communication modalities of smartphone devices have brought these to battle on the technological forefront and have curbed the fast spread of pandemics, such as COVID-19. To date, the community has proposed a wide range of contact tracing approaches, including opportunistic [185] and participatory approaches [64] as well as privacy-sensitive [260], decentralized [226], proximity-based (e.g., **Bluetooth Low Energy (BLE)**, sound) [187], and location-based approaches (e.g., Wi-Fi, GPS) [64]. However, a wide range of challenges remain unanswered, including methodologies to improve the penetration and adoption rates, alleviate privacy or expectation skepticism [32], ubiquitous availability on low-end terminals as well as technological/psychological adoption barriers [31], achieving cross-country interoperability with standard formations beyond recommendations, scalability/reliability and accuracy verification of engaged spatial technologies as well as lessons about effectiveness from real large-scale deployments.

Another specific health application for mobility data is health monitoring of older adults. GPS-enabled smartwatch technology can be used to monitor the movement of older-adult users [215]. In particular, if the monitored user is showing early signs of dementia, the user's trajectories could show an abrupt change from the individual's movement history [222]. For instance, a user who normally walks in a park and then goes to a restaurant is found to only stay in the park for a substantial amount of time. Indoor sensors installed in the room can also be used to track whether an an older adult or a patient falls from the bed. Trajectory outlier analysis methods, together with gerontology knowledge, can be very useful for this kind of application.

7.4 Indoor Environments

Indoor mobility data management has been described as a new frontier in data management [114]. However, in addition to data management, large-scale indoor localization data also raises challenges in data collection, data analysis, and data privacy. Indoor data collection is an open research problem due to the non-existence of the indoor equivalent of GPS: a system that can provide the user location in any building worldwide. This is particularly important in applications related to emergency management and infectious disease contact tracing. Systems have been developed over the years to address this problem based on different data sources, including WiFi signal strength and time of arrival [255], cellular signal [194], ultra-wideband [9], ultrasonic [110], magnetic tracking [213], and inertial sensors [102], among others. These novel data sources enable new applications in indoor navigation, contact tracing, indoor analytics, and evacuation management.

Indoor data analytics allows improvement of understanding of indoor behavior, which has multiple benefits and applications, including for crowd management [4], retail and POI recommendation systems [189], and for optimizing energy use and improving sustainability in the long term [200]. For example, by utilizing WiFi logs, Ren et al. [188] find strong correlations between behaviors and user demography (e.g., age, gender, and visitor types), indicating that indoor mobility behavior, in conjunction with online behavior, can be used to predict the underlying demography of the visitors.

Occupancy behaviors are also highly linked with building management systems and controls [45]. By having a more accurate energy use estimation using indoor spatial and mobility data, in addition to historical energy consumption data, the performance of the buildings can be better optimized, towards achieving more sustainable operations [71]. The responsible use of mobility behavior analytics, including indoor and outdoor mobility behaviors, strongly points to the increased capacity for improving sustainable operations of buildings [200], enabling net zero goals to be achieved.

7.5 Marine Transportation

According to UNCTAD, over 80% of the volume of international trade in goods is carried by sea, and the percentage is even higher for most developing countries [229]. Estimates say that global shipping activity emitted 3% of the global emissions worldwide in 2022 [109]. These significant numbers, as well as the availability of large-scale ship trajectory data obtained from the **automatic identification system (AIS)** [18], motivated a lot of research efforts on mobility data analysis for maritime transportation. The stakeholders who seek the benefit of such analyses include the maritime authorities, environment officers, ship owners, port and canal managers, and the transport and logistic sectors.

One major challenge is to ensure safety at sea, which splits down to the technical challenges of identifying positional anomalies [193], locating dark vessels (vessels that switch off their AIS devices) [147], and cleaning location and identity spoofing [73]. Additionally, an essential aspect is the detection of fishing activities to ensure sustainable fishing practices [58]. Since vessels do not have fixed routes in the sea, research has also investigated the density of ship routes [248].

Multi-criteria routing using multiple optimization criteria, including estimated time of arrival, fuel consumption, safety, and comfort, has been increasingly recognized as an important path planning problem [104]. An optimization of ship routes could effectively lead to significant reductions of greenhouse gas emissions and contribute to the actions against anthropogenic global warming. The influence of ocean currents, waves, and wind on the course and speed of ships have been known for centuries. Used optimally, ocean currents lead to more efficient paths between two given ports. Ship route computation approaches that exploit the potentials of wind, wave,

and weather models aiming at minimizing fuel consumption have been addressed by the marine science, maritime engineering, and transportation communities [77].

Since green mobility is currently gaining a huge amount of attention, carbon dioxide emission–aware ship routing is expected to make an enormous impact on the economy, politics, and society and provides very promising opportunities for the spatial and spatiotemporal database and mobility communities. Marine transportation becomes particularly important in the scope of climate change (e.g., the advent of hydrogen/battery/fossil/atom hybrid vessels) as well as digitization for new infrastructure-free localization technologies on-board.

7.6 Social Connections

Location-based social networks (LBSNs) bridge the gap between the physical world and online social networking services [269]. LBSN data capture both human mobility (in the form of check-ins to discrete points of interest) and a social network between individual humans. Combining mobility data and social networks, LBSN data finds many applications. A first application found in the literature was on modeling and describing human mobility patterns (e.g., [55, 167]), analyzing these patterns (e.g.,[54]), and explaining why individual users choose locations and how social ties affect this choice (e.g., [237]). Another application is that of location recommendation, which leverages check-ins of users and their ratings in the user-location network to recommend new locations to users [26]. A closely related application area is location prediction (e.g., [53]), which predicts the future check-ins of users. Another active research field in LBSN analysis is friend recommendation or social link prediction (e.g., [201]), which suggests new friends to users based on similar interests at similar locations while also having similar social connections. Other research topics concerning LBSNs include efficient query processing (e.g., [17]), finding user communities (e.g., [257]), and estimating the social influence of users (e.g., [246]).

This plethora of applications and research shows how mobility data in connection with social network data can be used to understand the social fabric that ties us together. A potential future application is using human mobility data to reinforce this social fabric by recommending social events and meetings to groups of people to help them find new friends, collaborators, sports mates, teachers, mentors, and family members.

8 CONCLUSIONS

This article presented the current state of the mobility data science pipeline in addressing the specific challenges of mobility data. A main question that this article answered is how mobility data science is different from data science. The space and time dimensions in mobility data call for different methods of data acquisition, management, analysis, and privacy preservation that are not addressed by the common data science tools. Accordingly, we surveyed the main problems that are currently being researched, we identified major research questions for the coming years, and described applications that lead to broader impacts of mobility data science. Co-authored by a diversity of academics and industry professionals, this article also conferred a community effort to sketch the boundary of mobility data science as an interdisciplinary field and bring together a dedicated research community around the identified research challenges.

ACKNOWLEDGEMENTS

Mohamed F. Mokbel acknowledges the support of the National Science Foundation under grants nos. IIS-1907855 and IIS-2203553. Mahmoud Sakr acknowledges the support of the EU's Horizon Europe research and innovation program under grant agreement nos. 101070279 (MobiSpaces) and 101093051 (EMERALDS). Li Xiong acknowledges the support of the National Science Foundation under grant nos. CNS-2125530 and CNS-2041952. Andreas Züfle and Taylor Anderson

acknowledge the support of the National Science Foundation under grant no. DEB-2109647. Walid G. Aref acknowledges the support of the National Science Foundation under grant no. IIS-1910216. Gennady and Natalia Andrienko acknowledge the support of the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B), and of the EU in projects *SoBigData++* and *CrexData* (grant agreement no. 101092749). Reynold Cheng acknowledges the support of the Hong Kong Jockey Club Charities Trust (Project No. 260920140), the University of Hong Kong (Project No. 109000579), and the HKU Outstanding Research Student Supervisor Award 2022-23. Panos K. Chrysanthis acknowledges the support of the National Science Foundation under grant no. SES-2017614 and of National Institute of Health under grant no. R01HL159805. Anita Graser acknowledges the support of the EU's Horizon Europe research and innovation program under grant agreement nos. 101070279 (MobiSpaces) and 101093051 (EMERALDS). Matthias Renz acknowledges the support of the German Research Foundation under grant nos. 290391021 and 491008639, the Helmholtz School for Marine Data Science (MarDATA) partially funded by the Helmholtz Association (grant no. HIDSS-0005) and the Federal Ministry for Economic Affairs and Climate Action (BMWi) under grant no. 68GX21002E. Flora Salim acknowledges the support of the Australian Research Council (ARC) Centre of Excellence for Automated Decision-Making and Society (ADM+S) (grant no. CE200100005). Maxime Schoemans acknowledges the support of the Fund for Scientific Research (FNRS) under grant no. 40018132. Yannis Theodoridis acknowledges the support of the EU's Horizon Europe research and innovation program under grant agreement nos. 101070279 (MobiSpaces) and 101093051 (EMERALDS). Song Wu acknowledges the support of the EU's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 955895 (DEDS). Jianqiu Xu acknowledges the support of the National Science Foundation under grant no. U23A20296.

REFERENCES

[1] ACM SIGSPATIAL CUP 2017 n. d. ACM SIGSPATIAL CUP 2017. Retrieved from http://sigspatial2017.sigspatial.org/giscup2017/download.

[2] Gergely Acs and Claude Castelluccia. 2014. A case study: Privacy preserving release of spatio-temporal density in Paris. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1679–1688.

[3] Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, Carola Wenk, Mahmuda Ahmed, Sophia Karagiorgou, Dieter Pfoser, and Carola Wenk. 2015. *Map Construction Algorithms*. Springer.

[4] Tanvir Ahmed, Torben Bach Pedersen, and Hua Lu. 2014. Finding dense locations in indoor tracking data. In *2014 IEEE 15th International Conference on Mobile Data Management*, Vol. 1. IEEE, 189–194.

[5] Ritesh Ahuja, Gabriel Ghinita, and Cyrus Shahabi. 2020. Differentially-private next-location prediction with neural networks. In *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02, 2020*, Angela Bonifati, Yongluan Zhou, Marcos Antonio Vaz Salles, Alexander Böhm, Dan Olteanu, George H. L. Fletcher, Arijit Khan, and Bin Yang (Eds.). OpenProceedings.org, 121–132. https://doi.org/10.5441/002/edbt.2020.12

[6] Wesam Al Amiri, Mohamed Baza, Karim Banawan, Mohamed Mahmoud, Waleed Alasmary, and Kemal Akkaya. 2019. Privacy-preserving smart parking system using blockchain and private information retrieval. In *2019 International Conference on Smart Applications, Communications and Networking (SmartNets)*. IEEE, 1–6.

[7] Louai Alarabi and Mohamed F. Mokbel. 2020. A demonstration of Summit: A scalable data management framework for massive trajectory. In *IEEE International Conference on Mobile Data Management, MDM* (Versailles, France). 226–227.

[8] Louai Alarabi, Mohamed F. Mokbel, and Mashaal Musleh. 2018. ST-Hadoop: A MapReduce framework for spatio-temporal data. *GeoInformatica* 22, 4 (2018), 785–813.

[9] Abdulrahman Alarifi, AbdulMalik Al-Salman, Mansour Alsaleh, Ahmad Alnafessah, Suheer Al-Hadhrami, Mai A. Al-Ammar, and Hend S. Al-Khalifa. 2016. Ultra wideband indoor positioning technologies: Analysis and recent advances. *Sensors* 16, 5 (2016), 707.

[10] Antonio M. R. Almeida, Jose L. A. Leite, Jose A. F. Macedo, and Javam C. Machado. 2017. GPS2GR: Optimized urban green routes based on GPS trajectories. In *Proceedings of the 8th ACM SIGSPATIAL Workshop on GeoStreaming*. 39–48.

[11] Ahmed M. Aly, Walid G. Aref, and Mourad Ouzzani. 2012. Spatial queries with two kNN predicates. *Proc. VLDB Endow.* 5, 11 (2012), 1100–1111. https://doi.org/10.14778/2350229.2350231

[12] Ahmed M. Aly, Walid G. Aref, and Mourad Ouzzani. 2015. Cost estimation of spatial k-nearest-neighbor operators. In *Proceedings of the 18th International Conference on Extending Database Technology, EDBT 2015, Brussels, Belgium, March 23-27, 2015*. OpenProceedings.org, 457–468. https://doi.org/10.5441/002/EDBT.2015.40

[13] Ahmed M. Aly, Walid G. Aref, and Mourad Ouzzani. 2015. Spatial queries with k-nearest-neighbor and relational predicates. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Bellevue, WA, USA, November 3-6, 2015*. ACM, 28:1–28:10. https://doi.org/10.1145/2820783.2820815

[14] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*. 901–914.

[15] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. 2013. *Visual Analytics of Movement*. Springer. https://doi.org/10.1007/978-3-642-37583-5

[16] Walid G. Aref, Yeasir Rayhan, Libin Zhou, and Anas Daghistani. 2022. ILX: Intelligent "location+X" data systems (vision paper). *CoRR* abs/2206.09520 (2022). https://doi.org/10.48550/ARXIV.2206.09520 arXiv:2206.09520

[17] Nikos Armenatzoglou, Stavros Papadopoulos, and Dimitris Papadias. 2013. A general framework for geo-social query processing. *Proc. of the VLDB Endowment* 6, 10 (2013), 913–924.

[18] Alexander Artikis and Dimitris Zissis. 2021. *Guide to Maritime Informatics*. Springer.

[19] Mohammad Asghari, Dingxiong Deng, Cyrus Shahabi, Ugur Demiryurek, and Yaguang Li. 2016. Price-aware real-time ride-sharing at scale: an auction-based approach. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–10.

[20] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. 2018. Spatio-temporal data mining: A survey of problems and methods. *ACM Comput. Surv.* 51, 4 (Aug 2018). https://doi.org/10.1145/3161602

[21] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in Neural Information Processing Systems* 32 (2019).

[22] Mohamed Bakli, Mahmoud Sakr, and Taysir Hassan A. Soliman. 2019. HadoopTrajectory: A Hadoop spatiotemporal data processing extension. *Journal of Geographical Systems* (2019), 1–25.

[23] Mohamed Bakli, Mahmoud Sakr, and Esteban Zimányi. 2020. Distributed spatiotemporal trajectory query processing in SQL. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*. 87–98.

[24] Satchit Balsari, Caroline Buckee, Jennifer Chan, and Andrew Schroeder. 2022. The use of human mobility data in public health emergencies. In *CrisisReady*. https://www.crisisready.io/wp-content/uploads/2022/06/The-Use-of-Human-Mobility-Data-in-Public-Health-Emergencies.pdf

[25] Jie Bao, Ruiyuan Li, Xiuwen Yi, and Yu Zheng. 2016. Managing massive trajectories on the cloud. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPACIAL'16)*. ACM, New York, NY. https://doi.org/10.1145/2996913.2996916

[26] Jie Bao, Yu Zheng, David Wilkie, and Mohamed Mokbel. 2015. Recommendations in location-based social networks: a survey. *GeoInformatica* 19, 3 (2015), 525–565.

[27] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R. James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J. Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74. https://doi.org/10.1016/j.physrep.2018.01.001 Human mobility: Models and applications.

[28] Emmanouil Barmpounakis and Nikolas Geroliminis. 2020. On the new era of urban traffic monitoring with massive drone data: The pNEUMA large-scale field experiment. *Transportation Research Part C: Emerging Technologies* 111 (2020), 50–71.

[29] Favyen Bastani, Songtao He, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Sam Madden, and David J. DeWitt. 2018. RoadTracer: Automatic extraction of road networks from aerial images. In *CVPR*. Salt Lake City, UT, 4720–4728.

[30] Zeynep Batmaz, Ali Yürekli, Alper Bilge, and Cihan Kaleli. 2019. A review on deep learning for recommender systems: Challenges and remedies. *Artif. Intell. Rev.* 52, 1 (2019), 1–37. https://doi.org/10.1007/s10462-018-9654-y

[31] Luca Bedogni, Federico Montori, and Flora Salim. 2022. Location contact tracing: Penetration, privacy, position, and performance. *Digital Government: Research and Practice* 3, 3 (2022), 1–13.

[32] Luca Bedogni, Shakila Khan Rumi, and Flora D. Salim. 2021. Modelling memory for individual re-identification in decentralised mobile contact tracing applications. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–21.

[33] Thomas Behr, Victor Teixeira de Almeida, and Ralf Hartmut Güting. 2006. Representation of periodic moving objects in databases. In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems* (Arlington, VA) *(GIS'06)*. ACM, New York, NY, 43–50. https://doi.org/10.1145/1183471.1183480

[34] Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. 2011. SUMO – Simulation of urban mobility: An overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*.

[35] Rimantas Benetis, Christian S. Jensen, Gytis Karĉiauskas, and Simonas Ŝaltenis. 2006. Nearest and reverse nearest neighbor queries for moving objects. *The VLDB Journal* 15 (2006), 229–249.

[36] Christos Berberidis, Ioannis P. Vlahavas, Walid G. Aref, Mikhail J. Atallah, and Ahmed K. Elmagarmid. 2002. On the discovery of weak periodicities in large time series. In *Proveedings of Principles of Data Mining and Knowledge Discovery, 6th European Conference, PKDD 2002, Helsinki, Finland, August 19-23, 2002 (Lecture Notes in Computer Science, Vol. 2431)*. Springer, Berlin, 51–61. https://doi.org/10.1007/3-540-45681-3_5

[37] James Biagioni and Jakob Eriksson. 2012. Inferring road maps from global positioning system traces: Survey and comparative evaluation. *Transportation Research Record: Journal of the Transportation Research Board* 2291, 1 (2012), 61–71.

[38] Bike Share Metro Data. n.d. Bike Share Metro Data. Retrieved from https://bikeshare.metro.net/about/data/

[39] Bike Town System Data. n.d. Bike Town System Data. Retrieved from https://www.biketownpdx.com/system-data

[40] Filip Biljecki, Hugo Ledoux, and Peter Van Oosterom. 2013. Transportation mode-based segmentation and classification of movement trajectories. *International Journal of Geographical Information Science* 27, 2 (2013), 385–407.

[41] Lorenzo Bracciale, Marco Bonola, Pierpaolo Loreti, Giuseppe Bianchi, Raul Amici, and Antonello Rabuffi. 2014. CRAWDAD dataset roma/taxi (v. 2014-07-17). Retrieved from https://crawdad.org/roma/taxi/20140717

[42] Sotiris Brakatsoulas, Dieter Pfoser, Randall Salas, and Carola Wenk. 2005. On map-matching vehicle tracking data. In *VLDB*. Trondheim, Norway, 853–864.

[43] Yang Cao, Yonghui Xiao, Li Xiong, Liquan Bai, and Masatoshi Yoshikawa. 2019. Protecting spatiotemporal event privacy in continuous location-based services. *IEEE Transactions on Knowledge and Data Engineering* 33, 8 (2019), 3141–3154.

[44] Capital Bike Share System Data. n.d. Capital Bike Share System Data. Retrieved from https://www.capitalbikeshare.com/system-data

[45] Salvatore Carlucci, Marilena De Simone, Steven K. Firth, Mikkel B. Kjærgaard, Romana Markovic, Mohammad Saiedur Rahaman, Masab Khalid Annaqeeb, Silvia Biandrate, Anooshmita Das, Jakub Wladyslaw Dziedzic, et al. 2020. Modeling occupant behavior in buildings. *Building and Environment* 174 (2020), 106768.

[46] Pingfu Chao, Yehong Xu, Wen Hua, and Xiaofang Zhou. 2020. A survey on map-matching algorithms. In *Australasian Database Conference, ADC*. Melbourne, Australia, 121–133.

[47] Kostantinos Chatzikokolakis, Ehab ElSalamouny, Catuscia Palamidessi, Pazii Anna, et al. 2017. Methods for location privacy: A comparative overview. *Foundations and Trends® in Privacy and Security* 1, 4 (2017), 199–257.

[48] Chen Chen, Cewu Lu, Qixing Huang, Qiang Yang, Dimitrios Gunopulos, and Leonidas J. Guibas. 2016. City-scale map creation and updating using GPS collections. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1465–1474. https://doi.org/10.1145/2939672.2939833

[49] Jinchuan Chen and Reynold Cheng. 2006. Efficient evaluation of imprecise location-dependent queries. In *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 586–595.

[50] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (May 2021), 40 pages. https://doi.org/10.1145/3447744

[51] Lisi Chen, Shuo Shang, Christian S. Jensen, Bin Yao, Zhiwei Zhang, and Ling Shao. 2019. Effective and efficient reuse of past travel behavior for route recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 488–498.

[52] Rui Chen, Haoran Li, A. Kai Qin, Shiva Prasad Kasiviswanathan, and Hongxia Jin. 2016. Private spatial data aggregation in the local setting. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*. IEEE, 289–300.

[53] Chen Cheng, Haiqin Yang, Michael R. Lyu, and Irwin King. 2013. Where you like to go next: Successive point-of-interest recommendation.. In *IJCAI*, Vol. 13. 2605–2611.

[54] Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring millions of footprints in location sharing services. *ICWSM* 2011 (2011), 81–88.

[55] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1082–1090.

[56] Hyunghoon Cho, Daphne Ippolito, and Yun William Yu. 2020. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv preprint arXiv:2003.11511* (2020).
[57] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*. 2201–2206.
[58] Buncha Chuaysi and Supaporn Kiattisin. 2020. Fishing vessels behavior identification for combating IUU fishing: Enable traceability at sea. *Wireless Personal Communications* 115 (2020), 2971–2993.
[59] Sai-Ho Chung, Hoi-Lam Ma, Mark Hansen, and Tsan-Ming Choi. 2020. Data science and analytics in aviation.101837.
[60] CitiBike System Data. n.d. CitiBike System Data. Retrieved from https://ride.citibikenyc.com/system-data
[61] Christophe Claramunt, Cyril Ray, L. Salmon, E. Camossi, Melita Hadzagic, A. L. Jousselme, G. Andrienko, N. Andrienko, Y. Theodoridis, and G. Vouros. 2017. Maritime data integration and analysis: recent progress and research challenges. *Advances in Database Technology-EDBT* 2017 (2017), 192–197.
[62] CoGo Bike Share System Data. n.d. CoGo Bike Share System Data. Retrieved from https://www.cogobikeshare.com/system-data
[63] Graham Cormode, Somesh Jha, Tejas Kulkarni, Ninghui Li, Divesh Srivastava, and Tianhao Wang. 2018. Privacy at scale: Local differential privacy in practice. In *Proceedings of the 2018 International Conference on Management of Data*. 1655–1658.
[64] Yanan Da, Ritesh Ahuja, Li Xiong, and Cyrus Shahabi. 2021. REACT: Real-time contact tracing and risk monitoring via privacy-enhanced mobile tracking. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*. IEEE, 2729–2732. https://doi.org/10.1109/ICDE51399.2021.00315
[65] Anas Daghistani, Walid G. Aref, Arif Ghafoor, and Ahmed R. Mahmood. 2021. SWARM: Adaptive load balancing in distributed streaming systems for big spatial data. *ACM Trans. Spatial Algorithms Syst.* 7, 3 (2021), 14:1–14:43. https://doi.org/10.1145/3460013
[66] Jian Dai, Bin Yang, Chenjuan Guo, and Zhiming Ding. 2015. Personalized route recommendation using big trajectory data. In *2015 IEEE 31st International Conference on Data Engineering*. 543–554. https://doi.org/10.1109/ICDE.2015.7113313
[67] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. 2013. NADEEF: A commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 541–552.
[68] Allan M. De Souza, Roberto S. Yokoyama, Guilherme Maia, Antonio Loureiro, and Leandro Villas. 2016. Real-time path planning to prevent traffic jam through an intelligent transportation system. In *2016 IEEE Symposium on Computers and Communication (ISCC)*. IEEE, 726–731.
[69] Daniel Dias and Luis Henrique Maciel Kosmalski Costa. 2018. CRAWDAD dataset coppe-ufrj/RioBuses (v. 2018-03-19). Retrieved from https://crawdad.org/coppe-ufrj/RioBuses/20180319
[70] Divvy Bikes System Data. n.d. Divvy Bikes System Data. Retrieved from https://www.divvybikes.com/system-data
[71] Bing Dong, Yapan Liu, Hannah Fontenot, Mohamed Ouf, Mohamed Osman, Adrian Chong, Shuxu Qin, Flora Salim, Hao Xue, Da Yan, et al. 2021. Occupant behavior modeling methods for resilient building design, operation and policy at urban scale: A review. *Applied Energy* 293 (2021), 116856.
[72] Bing Dong, Yapan Liu, Wei Mu, Zixin Jiang, Pratik Pandey, Tianzhen Hong, Bjarne Olesen, Thomas Lawrence, Zheng O'Neil, et al. 2022. A global building occupant behavior database. *Scientific Data* 9, 1 (2022), 369.
[73] Enrica d'Afflisio, Paolo Braca, and Peter Willett. 2021. Malicious AIS spoofing and abnormal stealth deviations: A comprehensive statistical framework for maritime anomaly detection. *IEEE Trans. Aerospace Electron. Systems* 57, 4 (2021), 2093–2108. https://doi.org/10.1109/TAES.2021.3083466
[74] Justin Elarde, Joon-Seok Kim, Hamdi Kavak, Andreas Züfle, and Taylor Anderson. 2021. Change of human mobility during COVID-19: A United States case study. *PLoS One* 16, 11 (2021), e0259031.
[75] Mohamed G. Elfeky, Walid G. Aref, and Ahmed K. Elmagarmid. 2004. Using convolution to mine obscure periodic patterns in one pass. In *Advances in Database Technology — EDBT 2004, Proceedings of the 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, March 14-18, 2004 (Lecture Notes in Computer Science, Vol. 2992)*. Springer, Berlin, 605–620. https://doi.org/10.1007/978-3-540-24741-8_35
[76] Mohamed M. Elshrif, Keivin Isufaj, and Mohamed F. Mokbel. 2022. Network-less trajectory imputation. In *SIGSPATIAL*. 8:1–8:10.
[77] Ming-Chung Fang and Yu-Hsien Lin. 2015. The optimization of ship weather-routing algorithm based on the composite influence of multi-dynamic elements (II): Optimized routings. *Applied Ocean Research* 50 (2015), 130–140.
[78] Federal Geographic Data Committee. n.d. Geospatial Metadata Standards and Guidelines. https://www.fgdc.gov/metadata/geospatial-metadata-standard.
[79] Robin G. Fegeas, Janette L. Cascio, and Robert A. Lazar. 1992. An overview of FIPS 173, the spatial data transfer standard. *Cartography and Geographic Information Systems* 19, 5 (1992), 278–293.

[80] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. DeepMove: Predicting human mobility with attentional recurrent networks. In *WWW*. 1459–1468.
[81] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1322–1333.
[82] Kaiqun Fu, Zhiqian Chen, and Chang-Tien Lu. 2018. StreetNet: Preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 269–278.
[83] Nan Gao, Max Marschall, Jane Burry, Simon Watkins, and Flora D. Salim. 2022. Understanding occupants' behaviour, engagement, emotion, and comfort indoors with heterogeneous sensors and wearables. *Scientific Data* 9, 1 (2022), 261.
[84] Nan Gao, Hao Xue, Wei Shao, Sichen Zhao, Kyle Kai Qin, Arian Prabowo, Mohammad Saiedur Rahaman, and Flora D. Salim. 2022. Generative adversarial networks for spatio-temporal data: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 2 (2022), 1–25.
[85] Qiang Gao, Fan Zhou, Kunpeng Zhang, Goce Trajcevski, Xucheng Luo, and Fengli Zhang. 2017. Identifying human mobility via trajectory embeddings. In *IJCAI*, Vol. 17. 1689–1695.
[86] Qiang Gao, Fan Zhou, Ting Zhong, Goce Trajcevski, Xin Yang, and Tianrui Li. 2022. Contextual spatio-temporal graph representation learning for reinforced human mobility mining. *Inf. Sci.* 606 (2022), 230–249. https://doi.org/10.1016/j.ins.2022.05.049
[87] Gabriel Ghinita, Panos Kalnis, Ali Khoshgozaran, Cyrus Shahabi, and Kian-Lee Tan. 2008. Private queries in location based services: anonymizers are not necessary. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. 121–132.
[88] Michael F. Goodchild. 2010. Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science* 1 (2010), 3–20. https://doi.org/10.5311/JOSIS.2010.1.2
[89] Michael F. Goodchild. 1998. Uncertainty: The Achilles heel of GIS. *Geo Info Systems* 8, 11 (1998), 50–52.
[90] Matthew Graham, Mark Kutzbach, and Brian McKenzie. 2014. *Design Comparison of LODES and ACS Commuting Data Products*. Technical Report. US Census Bureau, Center for Economic Studies.
[91] A. Graser. 2021. An exploratory data analysis protocol for identifying problems in continuous movement data. *Journal of Location Based Services* 15, 2 (2021), 89–117. https://doi.org/10.1080/17489725.2021.1900612
[92] Anita Graser. 2023. The state of trajectory visualization in notebook environments. *GI_Forum* 1 (2023), 73–91. https://doi.org/10.1553/giscience2022_02_s73
[93] Stéphane Grumbach, Philippe Rigaux, and Luc Segoufin. 1998. The DEDALE system for complex spatial queries. *SIGMOD Rec.* 27, 2 (Jun 1998), 213–224. https://doi.org/10.1145/276305.276324
[94] Xiaolan Gu, Ming Li, Li Xiong, and Yang Cao. 2020. Providing input-discriminative protection for local differential privacy. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 505–516.
[95] Mark F. Guagliardo. 2004. Spatial accessibility of primary care: Concepts, methods and challenges. *International Journal of Health Geographics* 3, 1 (2004), 1–13.
[96] Chenjuan Guo, Bin Yang, Ove Andersen, Christian S. Jensen, and Kristian Torp. 2015. Ecomark 2.0: Empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. *GeoInformatica* 19 (2015), 567–599.
[97] Ralf Hartmut Güting, Thomas Behr, and Christian Düntgen. 2010. SECONDO: A platform for moving objects database research and for publishing and integrating research implementations. *IEEE Data Eng. Bull.* 33 (2010), 56–63.
[98] Ralf Hartmut Güting, Victor Teixeira De Almeida, and Zhiming Ding. 2006. Modeling and querying moving objects in networks. *The VLDB Journal* 15 (2006), 165–190.
[99] Ralf Hartmut Güting and Markus Schneider. 2005. *Moving Objects Databases*. Elsevier.
[100] Ralf Güting, Michael Böhlen, Martin Erwig, Christian Jensen, Nikos Lorentzos, Markus Schneider, and Michalis Vazirgiannis. 2000. A foundation for representing and querying moving objects. *ACM Transactions on Database Systems (TODS)* 25 (03 2000), 1–42. https://doi.org/10.1145/352958.352963
[101] Xiaolin Han, Reynold Cheng, Chenhao Ma, and Tobias Grubenmann. 2022. DeepTEA: Effective and efficient online time-dependent trajectory outlier detection. *Proceedings of the VLDB Endowment* 15, 7 (2022), 1493–1505.
[102] Robert Harle. 2013. A survey of indoor inertial positioning systems for pedestrians. *IEEE Communications Surveys & Tutorials* 15, 3 (2013), 1281–1293.
[103] Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia Procopiuc, and Divesh Srivastava. 2015. DPT: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment* 8, 11 (2015), 1154–1165.
[104] Jörn Hinnenthal and Günther Clauss. 2010. Robust Pareto-optimum routing of ships utilising deterministic and ensemble weather forecasts. *Ships and Offshore Structures* 5, 2 (Aug. 2010), 105–114. https://doi.org/10.1080/17445300903210988

[105] Boyeong Hong, Bartosz J. Bonczak, Arpit Gupta, and Constantine E. Kontokosta. 2021. Measuring inequality in community resilience to natural disasters using large-scale mobility data. *Nature Communications* 12, 1 (2021), 1870.
[106] Xiao Hou, Song Gao, Qin Li, Yuhao Kang, Nan Chen, Kaiping Chen, Jinmeng Rao, Jordan S. Ellenberg, and Jonathan A. Patz. 2021. Intracounty modeling of COVID-19 infection with human mobility: Assessing spatial heterogeneity with business traffic, age, and race. *Proceedings of the National Academy of Sciences* 118, 24 (2021), e2020524118.
[107] Xiaocheng Huang, Yifang Yin, Simon Lim, Guanfeng Wang, Bo Hu, Jagannadan Varadarajan, Shaolin Zheng, Ajay Bulusu, and Roger Zimmermann. 2019. Grab-Posisi: An extensive real-life GPS trajectory dataset in Southeast Asia. In *Proceedings of the ACM SIGSPATIAL International Workshop on Prediction of Human Mobility, PredictGIS 2019*. Chicago, IL, 1–10.
[108] Zhiyong Huang, Hua Lu, Beng Chin Ooi, and Anthony K. H. Tung. 2006. Continuous skyline queries for moving objects. *IEEE Transactions on Knowledge and Data Engineering* 18, 12 (2006), 1645–1658.
[109] IEA (2023). n.d. IEA (2023), CO2 Emissions in 2022, IEA, Paris. License: CC BY 4.0. Retrieved 5 August 2023 from https://www.iea.org/reports/co2-emissions-in-2022
[110] Faheem Ijaz, Hee Kwon Yang, Arbab Waheed Ahmad, and Chankil Lee. 2013. Indoor positioning: A review of indoor ultrasonic positioning systems. In *2013 15th International Conference on Advanced Communications Technology (ICACT)*. IEEE, 1146–1150.
[111] Indego Trip Data. n. d. Indego Trip Data. Retrieved from https://www.rideindego.com/about/data/
[112] INRIX. n.d. 2022 Global Traffic Scorecard. Retrieved 1 November 2023 from https://inrix.com/scorecard/
[113] Christian S. Jensen, Hua Lu, and Bin Yang. 2009. Graph model based indoor tracking. In *2009 10th International Conference on Mobile Data Management: Systems, Services and Middleware*. 122–131. https://doi.org/10.1109/MDM.2009.23
[114] Christian S. Jensen, Hua Lu, and Bin Yang. 2010. Indoor—A new data management frontier. *IEEE Data Eng. Bull.* 33, 2 (2010), 12–17.
[115] Fengmei Jin, Wen Hua, Jiajie Xu, and Xiaofang Zhou. 2019. Moving object linking based on historical trace. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 1058–1069.
[116] Tanvi Jindal, Prasanna Giridhar, Lu-An Tang, Jun Li, and Jiawei Han. 2013. Spatiotemporal periodical pattern mining in traffic data. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp@KDD 2013, Chicago, IL, August 11, 2013*. ACM, 11:1–11:8. https://doi.org/10.1145/2505821.2505837
[117] Rocío Joo, Matthew E. Boone, Thomas A. Clay, Samantha C. Patrick, Susana Clusella-Trullas, and Mathieu Basille. 2020. Navigating through the R packages for movement. *Journal of Animal Ecology* 89, 1 (Jan. 2020), 248–267. https://doi.org/10.1111/1365-2656.13116
[118] Gregor Jossé, Klaus Arthur Schmid, Andreas Züfle, Georgios Skoumas, Matthias Schubert, and Dieter Pfoser. 2015. Tourismo: A user-preference tourist trip search engine. In *Advances in Spatial and Temporal Databases: 14th International Symposium, SSTD 2015, Hong Kong, China, August 26-28, 2015. Proceedings 14*. Springer, 514–519.
[119] Jong Wook Kim and Beakcheol Jang. 2019. Workload-aware indoor positioning data collection via local differential privacy. *IEEE Communications Letters* 23, 8 (2019), 1352–1356.
[120] Mohammad R. Kolahdouzan and Cyrus Shahabi. 2004. Voronoi-based K nearest neighbor search for spatial network databases. In *(e)Proceedings of the 30th International Conference on Very Large Data Bases, VLDB 2004, Toronto, Canada, August 31 – September 3, 2004*, Mario A. Nascimento, M. Tamer Özsu, Donald Kossmann, Renée J. Miller, José A. Blakeley, and K. Bernhard Schiefer (Eds.). Morgan Kaufmann, 840–851. https://doi.org/10.1016/B978-012088469-8.50074-7
[121] John Krumm. 2022. Maximum entropy bridgelets for trajectory completion. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–8.
[122] Pedro Lara-Benítez, Manuel Carranza-García, and José C. Riquelme. 2021. An experimental review on deep learning architectures for time series forecasting. *Int. J. Neural Syst.* 31, 3 (2021), 2130001:1–2130001:28. https://doi.org/10.1142/S0129065721300011
[123] Chunggi Lee, Yeonjun Kim, Seungmin Jin, Dongmin Kim, Ross Maciejewski, David Ebert, and Sungahn Ko. 2020. A visual analytics system for exploring, monitoring, and forecasting road traffic congestion. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (2020), 3133–3146. https://doi.org/10.1109/TVCG.2019.2922597
[124] Bozhao Li, Zhongliang Cai, Mengjun Kang, Shiliang Su, Shanshan Zhang, Lili Jiang, and Yong Ge. 2021. A trajectory restoration algorithm for low-sampling-rate floating car data and complex urban road networks. *International Journal of GIS* 35, 4 (2021), 717–740.
[125] Huan Li, Hua Lu, Christian S. Jensen, Bo Tang, and Muhammad Aamir Cheema. 2023. Spatial data quality in the Internet of Things: Management, exploitation, and prospects. *Comput. Surveys* 55, 3 (2023), 57:1–57:41.
[126] Ming Li, Rene Westerholt, Hongchao Fan, and Alexander Zipf. 2016. Assessing spatiotemporal predictability of LBSN: A case study of three Foursquare datasets. *GeoInformatica* (2016), 1–21.

[127] Ruiyuan Li, Huajun He, Rubin Wang, Sijie Ruan, Yuan Sui, Jie Bao, and Yu Zheng. 2020. TrajMesa: A distributed NoSQL storage engine for big trajectory data. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2002–2005.

[128] Xiaolei Li, Jiawei Han, Jae-Gil Lee, and Hector Gonzalez. 2007. Traffic density-based discovery of hot routes in road networks. In *Proceedings of Advances in Spatial and Temporal Databases: 10th International Symposium, SSTD 2007, Boston, MA, July 16-18, 2007*. Springer, 441–459.

[129] Yang Li, Dimitrios Gunopulos, Cewu Lu, and Leonidas J. Guibas. 2019. Personalized travel time prediction using a small number of probe vehicles. *ACM Trans. Spatial Algorithms Syst.* 5, 1 (2019), 4:1–4:27. https://doi.org/10.1145/3317663

[130] Yang Li, Yangyan Li, Dimitrios Gunopulos, and Leonidas J. Guibas. 2016. Knowledge-based trajectory completion from sparse GPS samples. In *SIGSPATIAL*. 33:1–33:10.

[131] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Conference Track Proceedings of 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018*. OpenReview.net. https://openreview.net/forum?id=SJiHXGWAZ

[132] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. 2015. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 1–10.

[133] Canhong Lin, King Lun Choy, George T. S. Ho, Sai Ho Chung, and H. Y. Lam. 2014. Survey of green vehicle routing problem: Past and future trends. *Expert Systems with Applications* 41, 4 (2014), 1118–1138.

[134] Zongyu Lin, Shiqing Lyu, Hancheng Cao, Fengli Xu, Yuqiong Wei, Hanan Samet, and Yong Li. 2020. HealthWalks: Sensing fine-grained individual health condition via mobility data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–26.

[135] London Cycling Data. n.d. London Cycling Data. Retrieved from https://cycling.data.tfl.gov.uk/

[136] Jed A. Long. 2016. Kinematic interpolation of movement data. *International Journal of Geographical Information Science* 30, 5 (2016), 854–868.

[137] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2023. A survey on deep learning for human mobility. *ACM Comput. Surv.* 55, 2 (2023), 7:1–7:44. https://doi.org/10.1145/3485125

[138] Lyft Bikes Bay Wheels Trip Data. n.d. BlueBikes System Data. Retrieved from https://www.bluebikes.com/system-data

[139] Lyft Bikes Bay Wheels Trip Data. n.d. Lyft Bikes Bay Wheels Trip Data. Retrieved from https://www.lyft.com/bikes/bay-wheels/system-data

[140] Jane Macfarlane and Matei Stroila. 2016. Addressing the uncertainties in autonomous driving. *SIGSPATIAL Special* 8, 2 (2016), 35–40.

[141] Madrid Open Data. n.d. Madrid Open Data. Retrieved from https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)

[142] Hiroya Maeda, Yoshihide Sekimoto, and Toshikazu Seto. 2016. Lightweight road manager: Smartphone-based automatic determination of road damage status by deep neural network. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. 37–45.

[143] Ahmed R. Mahmood, Ahmed M. Aly, Tatiana Kuznetsova, Saleh M. Basalamah, and Walid G. Aref. 2018. Disk-based indexing of recent trajectories. *ACM Trans. Spatial Algorithms Syst.* 4, 3 (2018), 7:1–7:27. https://doi.org/10.1145/3234941

[144] Ahmed R. Mahmood, Ahmed M. Aly, Thamir Qadah, El Kindi Rezig, Anas Daghistani, Amgad Madkour, Ahmed S. Abdelhamid, Mohamed S. Hassan, Walid G. Aref, and Saleh M. Basalamah. 2015. Tornado: A distributed spatio-textual stream processing system. *Proc. VLDB Endow.* 8, 12 (2015), 2020–2023. https://doi.org/10.14778/2824032.2824126

[145] Ahmed R. Mahmood, Anas Daghistani, Ahmed M. Aly, MingJie Tang, Saleh M. Basalamah, Sunil Prabhakar, and Walid G. Aref. 2018. Adaptive processing of spatial-keyword data over a distributed streaming cluster. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL 2018, Seattle, WA, November 06-09, 2018*. ACM, 219–228. https://doi.org/10.1145/3274895.3274932

[146] Ahmed R. Mahmood, Sri Punni, and Walid G. Aref. 2019. Spatio-temporal access methods: A survey (2010–2017). *GeoInformatica* 23, 1 (2019), 1–36.

[147] Fabio Mazzarella, Michele Vespe, Alfredo Alessandrini, Dario Tarchi, Giuseppe Aulicino, and Antonio Vollero. 2017. A novel anomaly detection approach to identify intentional AIS on-off switching. *Expert Systems with Applications* 78 (2017), 110–123. https://doi.org/10.1016/j.eswa.2017.02.011

[148] MicrosoftMissingRoads. n.d. Discover New Roads with Bing Maps. Retrieved from https://blogs.bing.com/maps/2022-12/Bing-Maps-is-bringing-new-roads/

[149] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long. 2018. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access* 6 (2018), 39501–39514. https://doi.org/10.1109/ACCESS.2018.2855437

[150] Darakhshan J. Mir, Sibren Isaacman, Ramón Cáceres, Margaret Martonosi, and Rebecca N. Wright. 2013. DP-WHERE: Differentially private modeling of human mobility. In *2013 IEEE International Conference on Big Data*. IEEE, 580–588.
[151] Mohamed Mokbel, Sofiane Abbar, and Rade Stanojevic. 2020. Contact tracing: Beyond the apps. *SIGSPATIAL Special* 12, 2 (2020), 15–24.
[152] Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, Jussara Almeida, Taylor Anderson, Walid Aref, Gennady Andrienko, Natalia Andrienko, Yang Cao, et al. 2022. Mobility data science (Dagstuhl seminar 22021). In *Dagstuhl Reports*, Vol. 12. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
[153] Mohamed F. Mokbel, Louai Alarabi, Jie Bao, Ahmed Eldawy, Amr Magdy, Mohamed Sarwat, Ethan Waytas, and Steven Yackel. 2013. MNTG: An extensible web-based traffic generator. In *Proceedings of Advances in Spatial and Temporal Databases: 13th International Symposium, SSTD 2013, Munich, Germany, August 21-23, 2013*. Springer, 38–55.
[154] Mohamed F. Mokbel, Thanaa M. Ghanem, and Walid G. Aref. 2003. Spatio-temporal access methods. *IEEE Data Eng. Bull.* 26, 2 (2003), 40–49. http://sites.computer.org/debull/A03june/arefF.ps
[155] Mohamed F. Mokbel, Li Xiong, and Demetrios Zeinalipour-Yazti. 2022. Introduction to the special issue on contact tracing. 2 pages.
[156] Mohamed F. Mokbel, Xiaopeng Xiong, and Walid G. Aref. 2004. SINA: Scalable incremental processing of continuous queries in spatio-temporal databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (Paris, France). 623–634.
[157] Peter Mooney, Marco Minghini, et al. 2017. A review of OpenStreetMap data. *Mapping and the Citizen Sensor* (2017), 37–59.
[158] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. 2019. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 33–42.
[159] Mashaal Musleh, Sofiane Abbar, Rade Stanojevic, and Mohamed Mokbel. 2021. QARTA: An ML-based system for accurate map services. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2273–2282.
[160] Mashaal Musleh and Mohamed F. Mokbel. 2022. RASED: A scalable dashboard for monitoring road network updates in OSM. In *MDM*. 214–221.
[161] Attila M. Nagy and Vilmos Simon. 2018. Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing* 50 (2018), 148–163. https://doi.org/10.1016/j.pmcj.2018.07.004
[162] New York Times. n.d. For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. Retrieved 5 May 2023 from https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html
[163] Long-Van Nguyen-Dinh, Walid G. Aref, and Mohamed F. Mokbel. 2010. Spatio-temporal access methods: Part 2 (2003 - 2010). *IEEE Data Eng. Bull.* 33, 2 (2010), 46–55. http://sites.computer.org/debull/A10june/Aref.pdf
[164] NiceRide System Data. n.d. NiceRide System Data. Retrieved from https://www.niceridemn.com/system-data
[165] Jan Nijman and Yehua Dennis Wei. 2020. Urban inequalities in the 21st century economy. *Applied Geography* 117 (2020), 102188.
[166] Panagiotis Nikitopoulos, Aris-Iakovos Paraskevopoulos, Christos Doulkeridis, Nikos Pelekis, and Yannis Theodoridis. 2018. Hot spot analysis over big trajectory data. In *2018 IEEE International Conference on Big Data (Big Data)*. 761–770. https://doi.org/10.1109/BigData.2018.8622376
[167] Anastasios Noulas, Salvatore Scellato, Cecilia Mascolo, and Massimiliano Pontil. 2011. An empirical study of geographic user activity patterns in Foursquare. *ICWSM* 11 (2011), 70–573.
[168] Kaggle. n.d.. New York City Taxi Trip Duration. Retrieved from https://www.kaggle.com/c/nyc-taxi-trip-duration/data
[169] Oracle. 2022. Stream IoT data to an autonomous database using serverless functions F35431-06. Retrieved from https://docs.oracle.com/en/solutions/iot-streaming-oci/index.html#GUID-FEE82830-EE69-42C8-8068-BF955DD4A025
[170] OSM. n.d. Open Street Map. Retrieved from http://www.openstreetmap.org
[171] Federico Ossi, Fatima Hachem, Francesca Cagnacci, Urška Demšar, and Maria Luisa Damiani. 2022. HaniMob 2021 Workshop Report: The 1st ACM SIGSPATIAL Workshop on Animal Movement Ecology and Human Mobility. *SIGSPATIAL Special* 13, 1-3 (2022), 33–36.
[172] Scott E. Page. 1999. Computational models from A to Z. *Complexity* 5, 1 (1999), 35–41.
[173] Bei Pan, Yu Zheng, David Wilkie, and Cyrus Shahabi. 2013. Crowd sensing of traffic anomalies based on human mobility and social media. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 344–353.
[174] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. 2019. A public data set of spatio-temporal match events in soccer competitions. *Nature Scientific Data* 6, 1 (2019), 236.

[175] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, et al. 2013. Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)* 45, 4 (2013), 1–32.
[176] Kostas Patroumpas, Elias Alevizos, Alexander Artikis, Marios Vodas, Nikos Pelekis, and Yannis Theodoridis. 2017. Online event recognition from moving vessel trajectories. *GeoInformatica* 21 (2017), 389–427.
[177] Nikos Pelekis, Chiara Renso, Yannis Theodoridis, and Karine Zeitouni. 2022. Editor's note. *GeoInformatica* 26, 3 (2022), 449. https://doi.org/10.1007/s10707-022-00468-z
[178] Dieter Pfoser. 2016. *Crowdsourcing Geographic Information Systems*. Springer New York, NY, 1–8. https://doi.org/10.1007/978-1-4899-7993-3_80607-1
[179] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. 2009. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Retrieved from https://crawdad.org/epfl/mobility/20090224
[180] Porto. n.d. Taxi Service Trajectory. Prediction Challenge. ECML PKDD 2015. Retrieved from http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html
[181] PostGIS Project Steering Committee. 2023. PostGIS, spatial and geographic objects for postgreSQL. Retrieved from https://postgis.net
[182] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. Knock knock, who's there? Membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145* (2017).
[183] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. 2017. What does the crowd say about you? Evaluating aggregation-based location privacy. *Proceedings on Privacy Enhancing Technologies* 2017, 4 (2017), 156–176.
[184] Wahbeh Qardaji, Weining Yang, and Ninghui Li. 2013. Differentially private grids for geospatial data. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 757–768.
[185] Sirisha Rambhatla, Sepanta Zeighami, Kameron Shahabi, Cyrus Shahabi, and Yan Liu. 2022. Toward accurate spatiotemporal COVID-19 risk scores using high-resolution real-world mobility data. *ACM Trans. Spatial Algorithms Syst.* 8, 2 (2022), 1–30. https://doi.org/10.1145/3481044
[186] Leonie Reichert, Samuel Brack, and Björn Scheuermann. 2020. Privacy-preserving contact tracing of COVID-19 patients. *Cryptology ePrint Archive* (2020).
[187] Leonie Reichert, Samuel Brack, and Björn Scheuermann. 2021. A survey of automatic contact tracing approaches using Bluetooth low energy. *ACM Transactions on Computing for Healthcare* 2, 2 (2021), 1–33.
[188] Yongli Ren, Martin Tomko, Flora D. Salim, Jeffrey Chan, and Mark Sanderson. 2018. Understanding the predictability of user demographics from cyber-physical-social behaviours in indoor retail spaces. *EPJ Data Science* 7, 1 (2018), 1–21.
[189] Yongli Ren, Martin Tomko, Flora Dilys Salim, Kevin Ong, and Mark Sanderson. 2017. Analyzing web behavior in indoor retail spaces. *Journal of the Association for Information Science and Technology* 68, 1 (2017), 62–76.
[190] El Kindi Rezig, Lei Cao, Michael Stonebraker, Giovanni Simonini, Wenbo Tao, Samuel Madden, Mourad Ouzzani, Nan Tang, and Ahmed K. Elmagarmid. 2019. Data Civilizer 2.0: A holistic framework for data preparation and analytics. *Proc. VLDB Endow.* 12, 12 (Aug 2019), 1954–1957. https://doi.org/10.14778/3352063.3352108
[191] Keven Richly, Ralf Teusner, Alexander Immer, Fabian Windheuser, and Lennard Wolf. 2015. Optimizing routes of public transportation systems by analyzing the data of taxi rides. In *Proceedings of the 1st International ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics*. 70–76.
[192] Ride Austin Dataset. n.d. Ride Austin Dataset. Retrieved from https://data.world/ride-austin/ride-austin-june-6-april-13
[193] Maria Riveiro, Giuliana Pallotta, and Michele Vespe. 2018. Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 5 (2018), e1266.
[194] Hamada Rizk, Marwan Torki, and Moustafa Youssef. 2018. CellinDeep: Robust and accurate cellular-based indoor localization via deep learning. *IEEE Sensors Journal* 19, 6 (2018), 2305–2312.
[195] Alexander Rodríguez, Harshavardhan Kamarthi, Pulak Agarwal, Javen Ho, Mira Patel, Suchet Sapre, and B. Aditya Prakash. 2022. Data-centric epidemic forecasting: A survey. *arXiv preprint arXiv:2207.09370* (2022).
[196] Safegraph. n.d. Safegraph. Places Data Curated for Accurate Geospatial Analytics. Retrieved from https://www.safegraph.com/
[197] SafeGraph Inc. n.d. Weekly Patterns Dataset. Retrieved from https://docs.safegraph.com/docs/weekly-patterns
[198] Mahmoud Sakr, Cyril Ray, and Chiara Renso. 2022. Big mobility data analytics: Recent advances and open problems. *Geoinformatica* 26, 4 (Oct 2022), 541–549. https://doi.org/10.1007/s10707-022-00483-0
[199] Mahmoud Attia Sakr and Ralf Hartmut Güting. 2014. Group spatiotemporal pattern queries. *GeoInformatica* 18 (2014), 699–746.
[200] Flora D. Salim, Bing Dong, Mohamed Ouf, Qi Wang, Ilaria Pigliautile, Xuyuan Kang, Tianzhen Hong, Wenbo Wu, Yapan Liu, Shakila Khan Rumi, Mohammad Saiedur Rahaman, Jingjing An, Hengfang Deng, Wei Shao, Jakub Dziedzic, Fisayo Caleb Sangogboye, Mikkel Baun Kjærgaard, Meng Kong, Claudia Fabiani, Anna Laura Pisello, and Da Yan.

2020. Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. *Building and Environment* 183 (2020).

[201] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. 2011. Exploiting place features in link prediction on location-based social networks. In *ACM SIGKDD*. 1046–1054.

[202] Erik Seglem, Andreas Züfle, Jan Stutzki, Felix Borutta, Evgheniy Faerman, and Matthias Schubert. 2017. On privacy in spatio-temporal data: User identification using microblog data. In *Proceedings of Advances in Spatial and Temporal Databases: 15th International Symposium, SSTD 2017, Arlington, VA, August 21–23, 2017*. Springer, 43–61.

[203] Sumit Shah, Fenye Bao, Chang-Tien Lu, and Ing-Ray Chen. 2011. CROWDSAFE: Crowd sourcing of crime incidents and safe routing on mobile devices. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 521–524.

[204] Sina Shaham, Gabriel Ghinita, Ritesh Ahuja, John Krumm, and Cyrus Shahabi. 2022. HTF: Homogeneous tree framework for differentially-private release of large geospatial datasets with self-tuning structure height. *ACM Transactions on Spatial Algorithms and Systems* (2022).

[205] Sina Shaham, Gabriel Ghinita, and Cyrus Shahabi. 2022. Models and mechanisms for spatial data fairness. *Proc. VLDB Endow.* 16, 2 (2022), 167–179. https://www.vldb.org/pvldb/vol16/p167-ghinita.pdf

[206] Shuo Shang, Bo Yuan, Ke Deng, Kexin Xie, and Xiaofang Zhou. 2011. Finding the most accessible locations: Reverse path nearest neighbor query in road networks. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 181–190.

[207] Zeyuan Shang, Guoliang Li, and Zhifeng Bao. 2018. DITA: Distributed in-memory trajectory analytics. In *Proceedings of the 2018 International Conference on Management of Data*. 725–740.

[208] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jeffrey Chan, Xinhong Hei, Bradley Feest, and Flora D. Salim. 2019. Flight delay prediction using airport situational awareness map. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 432–435.

[209] Ashwin Shashidharan, Varun Chandola, and Ranga Raju Vatsavai. 2021. The 9th ACM SIGSPATIAL International Workshop on Analytics for Big Spatial Data (BigSpatial 2020) November 3, 2020. *SIGSPATIAL Special* 12, 3 (2021), 15–16.

[210] Shashi Shekhar, Zhe Jiang, Reem Y. Ali, Emre Eftelioglu, Xun Tang, Venkata M. V. Gunturi, and Xun Zhou. 2015. Spatiotemporal data mining: A computational perspective. *ISPRS International Journal of Geo-Information* 4, 4 (2015), 2306–2338. https://doi.org/10.3390/ijgi4042306

[211] Jaewoo Shin, Jianguo Wang, and Walid G. Aref. 2021. The LSM RUM-tree: A log structured merge R-tree for update-intensive spatial workloads. In *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*. IEEE, 2285–2290. https://doi.org/10.1109/ICDE51399.2021.00238

[212] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[213] Yuanchao Shu, Cheng Bo, Guobin Shen, Chunshui Zhao, Liqun Li, and Feng Zhao. 2015. Magicol: Indoor localization using pervasive magnetic field and opportunistic WiFi sensing. *IEEE Journal on Selected Areas in Communications* 33, 7 (2015), 1443–1457.

[214] Yasin N. Silva, Walid G. Aref, Per-Åke Larson, Spencer Pearson, and Mohamed H. Ali. 2013. Similarity queries: Their conceptual evaluation, transformations, and processing. *VLDB J.* 22, 3 (2013), 395–420. https://doi.org/10.1007/S00778-012-0296-4

[215] Thanos G. Stavropoulos, Asterios Papastergiou, Lampros Mpaltadoros, Spiros Nikolopoulos, and Ioannis Kompatsiaris. 2020. IoT wearable sensors and devices in elderly care: A literature review. *Sensors* 20, 10 (2020), 2826.

[216] Daniel Sui, Sarah Elwood, and Michael Goodchild. 2012. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Springer Science & Business Media.

[217] Xu Teng, Goce Trajcevski, Joon-Seok Kim, and Andreas Züfle. 2020. Semantically diverse path search. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 69–78.

[218] Xu Teng, Goce Trajcevski, and Andreas Züfle. 2021. Semantically diverse paths with range and origin constraints. In *Proceedings of the 29th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 375–378.

[219] Hien To, Gabriel Ghinita, and Cyrus Shahabi. 2014. A framework for protecting worker location privacy in spatial crowdsourcing. *Proceedings of the VLDB Endowment* 7, 10 (2014), 919–930.

[220] Hien To, Cyrus Shahabi, and Li Xiong. 2018. Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 833–844.

[221] Waldo R. Tobler. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46, sup1 (1970), 234–240.

[222] Magdalena I.. Tolea, John C. Morris, and James E. Galvin. 2016. Trajectory of mobility decline by type of dementia. *Alzheimer Disease and Associated Disorders* 30, 1 (2016), 60.

[223] Dimitrios Tomaras, Vana Kalogeraki, Thomas Liebig, and Dimitrios Gunopulos. 2018. Crowd-based ecofriendly trip planning. In *19th IEEE International Conference on Mobile Data Management, MDM 2018, Aalborg, Denmark, June 25-28, 2018*. IEEE Computer Society, 24–33. https://doi.org/10.1109/MDM.2018.00018
[224] Kevin Toohey and Matt Duckham. 2015. Trajectory similarity measures. *SIGSPATIAL Special* 7, 1 (2015), 43–50.
[225] Goce Trajcevski, Ouri Wolfson, Klaus Hinrichs, and Sam Chamberlain. 2004. Managing uncertainty in moving objects databases. *ACM Trans. Database Syst.* 29, 3 (Sep 2004), 463–507. https://doi.org/10.1145/1016028.1016030
[226] Carmela Troncoso, Mathias Payer, Jean-Pierre Hubaux, Marcel Salathé, James Larus, Edouard Bugnion, Wouter Lueks, Theresa Stadler, Apostolos Pyrgelis, Daniele Antonioli, et al. 2020. Decentralized privacy-preserving proximity tracing. *arXiv preprint arXiv:2005.12273* (2020).
[227] Robert Truong, Olga Gkountouna, Dieter Pfoser, and Andreas Züfle. 2018. Towards a better understanding of public transportation traffic: A case study of the Washington, DC metro. *Urban Science* 2, 3 (2018), 65.
[228] ULB Data Science Lab. n.d. MobilityDB, an open source geospatial trajectory data management and analysis platform. Retrieved from https://mobilitydb.com
[229] United Nations Conference on Trade and Development. n.d. Review of Maritime Transport 2022. Retrieved from https://unctad.org/rmt2022
[230] United Nations Department of Economic and Social Affairs. 2018. Revision of world urbanization prospects. *New York: United Nations Department of Economic and Social Affairs* (2018).
[231] United States Geological Survey. n.d. USGS Science Data Catalog. Retrieved from https://data.usgs.gov/datacatalog/
[232] U.S. Department of Transportation. n.d. Data Inventory. Retrieved from https://www.transportation.gov/data
[233] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).
[234] Mohammad M. Vazifeh, Hongmou Zhang, Paolo Santi, and Carlo Ratti. 2019. Optimizing the deployment of electric vehicle charging stations using pervasive mobility data. *Transportation Research Part A: Policy and Practice* 121 (2019), 75–91.
[235] Ymir Vigfusson, Thorgeir A. Karlsson, Derek Onken, Congzheng Song, Atli F. Einarsson, Nishant Kishore, Rebecca M. Mitchell, Ellen Brooks-Pollock, Gudrun Sigmundsdottir, and Leon Danon. 2021. Cell-phone traces reveal infection-associated behavioral change. *Proceedings of the National Academy of Sciences* 118, 6 (2021), e2005241118.
[236] Renee E. Walker, Christopher R. Keane, and Jessica G. Burke. 2010. Disparities and access to healthy food in the United States: A review of food deserts literature. *Health & Place* 16, 5 (2010), 876–884.
[237] Fengjiao Wang, Guan Wang, and S. Yu Philip. 2014. Why checkins: Exploring user motivation on location based social networks. In *Data Mining Workshop (ICDMW)*. IEEE, 27–34.
[238] Guang Wang, Xiuyuan Chen, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. Experience: Understanding long-term evolving patterns of shared electric vehicle networks. In *Proceedings of the International Conference on Mobile Computing and Networking, MobiCom*. Los Cabos, Mexico, 1–12.
[239] Han Wang, Hanbin Hong, Li Xiong, Zhan Qin, and Yuan Hong. 2022. PrivLBS: Local differential privacy for location-based services with staircase randomized response. In *Proceedings of the ACM Conference on Computer and Communications Security*.
[240] Hongjian Wang, Xianfeng Tang, Yu-Hsuan Kuo, Daniel Kifer, and Zhenhui Li. 2019. A simple baseline for travel time estimation using large-scale trip data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 2 (2019), 1–22.
[241] Haiming Wang, Zhikun Zhang, Tianhao Wang, Shibo He, Michael Backes, Jiming Chen, and Yang Zhang. 2023. PrivTrace: Differentially private trajectory synthesis by adaptive Markov model. In *USENIX Security*.
[242] Haozhou Wang, Kai Zheng, Jiajie Xu, Bolong Zheng, Xiaofang Zhou, and Shazia Sadiq. 2014. SharkDB: An in-memory column-oriented trajectory storage. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 1409–1418.
[243] Jingyuan Wang, Ning Wu, Xinxi Lu, Wayne Xin Zhao, and Kai Feng. 2021. Deep trajectory recovery with fine-grained calibration using Kalman filter. *TKDE* 33, 3 (2021), 921–934.
[244] Sheng Wang, Zhifeng Bao, J. Shane Culpepper, and Gao Cong. 2021. A survey on trajectory data management, analytics, and learning. *ACM Comput. Surv.* 54, 2, Article 39 (Mar 2021), 36 pages. https://doi.org/10.1145/3440207
[245] Zhaonan Wang, Renhe Jiang, Hao Xue, Flora D. Salim, Xuan Song, and Ryosuke Shibasaki. 2022. Event-aware multimodal mobility nowcasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 4228–4236.
[246] Yu-Ting Wen, Po-Ruey Lei, Wen-Chih Peng, and Xiao-Fang Zhou. 2014. Exploring social influence on location-based social networks. In *ICDM*. IEEE, 1043–1048.
[247] Randall T. Whitman, Bryan G. Marsh, Michael B. Park, and Erik G. Hoel. 2019. Distributed spatial and spatio-temporal join on Apache Spark. *ACM Transactions on Spatial Algorithms and Systems (TSAS)* 5, 1 (2019), 1–28.
[248] Lin Wu, Yongjun Xu, Qi Wang, Fei Wang, and Zhiwei Xu. 2017. Mapping global shipping density from AIS data. *The Journal of Navigation* 70, 1 (2017), 67–81. https://doi.org/10.1017/S0373463316000345

[249] Yonghui Xiao and Li Xiong. 2015. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. 1298–1309.
[250] Jiyang Xie, Zeyu Song, Yupeng Li, Yanting Zhang, Hong Yu, Jinnan Zhan, Zhanyu Ma, Yuanyuan Qiao, Jianhua Zhang, and Jun Guo. 2018. A survey on machine learning-based mobile big data analysis: Challenges and applications. *Wirel. Commun. Mob. Comput.* 2018 (2018), 8738613:1–8738613:19. https://doi.org/10.1155/2018/8738613
[251] Xiaopeng Xiong, Mohamed F. Mokbel, Walid G. Aref, Susanne E. Hambrusch, and Sunil Prabhakar. 2004. Scalable spatio-temporal continuous query processing for location-aware services. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 2004), 21-23 June 2004, Santorini Island, Greece*. IEEE Computer Society, 317–326. https://doi.org/10.1109/SSDBM.2004.61
[252] Jianqiu Xu, Ralf Hartmut Güting, and Yunjun Gao. 2018. Continuous k nearest neighbor queries over large multi-attribute trajectories: A systematic approach. *GeoInformatica* 22 (2018), 723–766.
[253] Hao Xue, Flora Salim, Yongli Ren, and Nuria Oliver. 2021. MobTCast: Leveraging auxiliary trajectory forecasting for human mobility prediction. *Advances in Neural Information Processing Systems* 34 (2021), 30380–30391.
[254] Hao Xue and Flora D. Salim. 2021. TERMCast: Temporal relation modeling for effective urban flow forecasting. In *Advances in Knowledge Discovery and Data Mining: 25th Pacific-Asia Conference, PAKDD 2021*. Springer, 741–753.
[255] Chouchang Yang and Huai-Rong Shao. 2015. WiFi-based indoor positioning. *IEEE Communications Magazine* 53, 3 (2015), 150–157.
[256] Yu Yang, Fan Zhang, and Desheng Zhang. 2018. SharedEdge: GPS-free fine-grained travel time estimation in state-level highway systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 48:1–48:26.
[257] Hongzhi Yin, Zhiting Hu, Xiaofang Zhou, Hao Wang, Kai Zheng, Quoc Viet Hung Nguyen, and Shazia Sadiq. 2016. Discovering interpretable geo-social communities for user behavior prediction. In *ICDE*. IEEE, 942–953.
[258] Haitao Yuan and Guoliang Li. 2021. A survey of traffic prediction: From spatio-temporal data to intelligent transportation. *Data Science and Engineering* 6 (2021), 63–85.
[259] Jing Yuan, Yu Zheng, Chengyang Zhang, Wenlei Xie, Xing Xie, Guangzhong Sun, and Yan Huang. 2010. T-drive: Driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 99–108.
[260] Abbas Zaidi, Ritesh Ahuja, and Cyrus Shahabi. 2022. Differentially private occupancy monitoring from WiFi access points. In *23rd IEEE International Conference on Mobile Data Management, MDM 2022, Paphos, Cyprus, June 6-9, 2022*. IEEE, 361–366. https://doi.org/10.1109/MDM55031.2022.00081
[261] Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S. Yu. 2017. Time series data cleaning: From anomaly detection to anomaly repairing. *PVLDB* 10, 10 (2017), 1046–1057.
[262] Liming Zhang, Liang Zhao, and Dieter Pfoser. 2022. Factorized deep generative models for end-to-end trajectory generation with spatiotemporal validity constraints. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–12.
[263] Ping Zhang, Hui Zhang, and Danhuai Guo. 2015. Evacuation shelter and route selection based on multi-objective optimization approach. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management*. 1–5.
[264] Zhigang Zhang, Cheqing Jin, Jiali Mao, Xiaolin Yang, and Aoying Zhou. 2017. TrajSpark: A scalable and efficient in-memory management system for big trajectory data. In *Proceedings of Web and Big Data: First International Joint Conference, APWeb-WAIM 2017, Beijing, China, July 7–9, 2017, Part I*. Springer, 11–26.
[265] Kai Zhao, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1911–1920.
[266] Sichen Zhao, Wei Shao, Jeffrey Chan, and Flora D. Salim. 2022. Measuring disentangled generative spatio-temporal representation. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 522–530.
[267] Xiangguo Zhao, Yanhui Li, Ye Yuan, Xin Bi, and Guoren Wang. 2019. LDpart: Effective location-record data publication via local differential privacy. *IEEE Access* 7 (2019), 31435–31445.
[268] Kai Zheng, Yu Zheng, Xing Xie, and Xiaofang Zhou. 2012. Reducing uncertainty of low-sampling-rate trajectories. In *ICDE*. 1144–1155.
[269] Yu Zheng. 2011. Location-based social networks: Users. In *Computing with Spatial Trajectories*. Springer, 243–276.
[270] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 1–55.
[271] Yu Zheng, Xing Xie, and Wei-Ying Ma. 2010. GeoLife: A collaborative social networking service among user, location and trajectory. *IEEE Data(base) Engineering Bulletin* (June 2010). Retrieved from https://www.microsoft.com/en-us/research/publication/geolife-a-collaborative-social-networking-service-among-user-location-and-trajectory/
[272] Fan Zhou, Xin Liu, Kunpeng Zhang, and Goce Trajcevski. 2021. Toward discriminating and synthesizing motion traces using deep probabilistic generative models. *IEEE Transactions on Neural Networks and Learning Systems* 32, 6 (2021), 2401–2414. https://doi.org/10.1109/TNNLS.2020.3005325

[273] Fan Zhou, Pengyu Wang, Xovee Xu, Wenxin Tai, and Goce Trajcevski. 2022. Contrastive trajectory learning for tour recommendation. *ACM Trans. Intell. Syst. Technol.* 13, 1 (2022), 4:1–4:25. https://doi.org/10.1145/3462331

[274] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. 2022. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Comput. Surv.* 54, 2 (2022), 27:1–27:36. https://doi.org/10.1145/3433000

[275] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James J. Q. Yu. 2023. DiffTraj: Generating GPS trajectory with diffusion probabilistic model. In *NeurIPS*.

[276] Esteban Zimányi, Mahmoud Sakr, and Arthur Lesuisse. 2020. MobilityDB: A mobility database based on PostgreSQL and PostGIS. *ACM Trans. Database Syst.* 45, 4 (Dec 2020). https://doi.org/10.1145/3406534

[277] Andreas Züfle, Taylor Anderson, and Song Gao. 2022. Introduction to the special issue on understanding the spread of COVID-19, Part 1. *ACM Transactions on Spatial Algorithms and Systems* 8, 3 (2022), 1–5.

[278] Andreas Züfle, Goce Trajcevski, Dieter Pfoser, Matthias Renz, Matthew T. Rice, Timothy Leslie, Paul Delamater, and Tobias Emrich. 2017. Handling uncertainty in geo-spatial data. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 1467–1470.

Received 7 August 2023; revised 12 February 2024; accepted 20 February 2024