Spatial Data Quality in the IoT Era: Management and Exploitation

Huan Li[†] Bo Tang[‡] Hua Lu[§] Muhammad Aamir Cheema[¶] Christian S. Jensen[†]

[†]Department of Computer Science, Aalborg University, Denmark
[‡]Department of Computer Science and Engineering, Southern University of Science and Technology, China
[§]Department of People and Technology, Roskilde University, Denmark
[¶]Department of Software Systems and Cybersecurity, Monash University, Australia

{lihuan, csj}@cs.aau.dk; tangb3@sustech.edu.cn; luhua@ruc.dk; aamir.cheema@monash.edu

ABSTRACT

Within the rapidly expanding Internet of Things (IoT), growing amounts of spatially referenced data are being generated. Due to the dynamic, decentralized, and heterogeneous nature of the IoT, spatial IoT data (SID) quality has attracted considerable attention in academia and industry. How to invent and use technologies for managing spatial data quality and exploiting low-quality spatial data are key challenges in the IoT. In this tutorial, we highlight the SID consumption requirements in applications and offer an overview of spatial data quality in the IoT setting. In addition, we review pertinent technologies for quality management and low-quality data exploitation, and we identify trends and future directions for quality-aware SID management and utilization. The tutorial aims to not only help researchers and practitioners to better comprehend SID quality challenges and solutions, but also offer insights that may enable innovative research and applications.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; *Sensor networks*; *Geographic information systems.*

KEYWORDS

Internet of Things; geo-sensory data; quality management

1 TUTORIAL OVERVIEW

The Internet of Things (IoT) encompasses numerous devices (e.g., sensors, actuators, wearables, and vehicles) to enable functionality such as ubiquitous perception and decision-making [80]. The IoT enables applications in smart cities [92], transportation [99], healthcare [72], energy [101], etc. An annual growth rate of 25% in IoT devices [1] is evidence of explosive growth in IoT data. Indeed, market intelligence provider IDC predicts that IoT data volume will reach 80 ZB by 2025 [2].

The geographic information and mobile computing communities are finding opportunities from IoT data, as IoT data is often spatially referenced by means of different positioning technologies [94]. In this tutorial, we concentrate on such spatially referenced data from IoT devices, called **spatial IoT data** and abbreviated as SID. Two important special cases of SID are distinguished: *trajectories*, time series of location values; and *spatiotemporal IoT data* (STID), general sensory values with temporal and spatial references. SID includes substantial observations in potentially large spatial regions, thus offering an exciting foundation for new insights that may benefit diverse IoT-enabled applications, including congestion control [99], urban planning [92], air quality monitoring [60], and POI recommendations [41, 128].

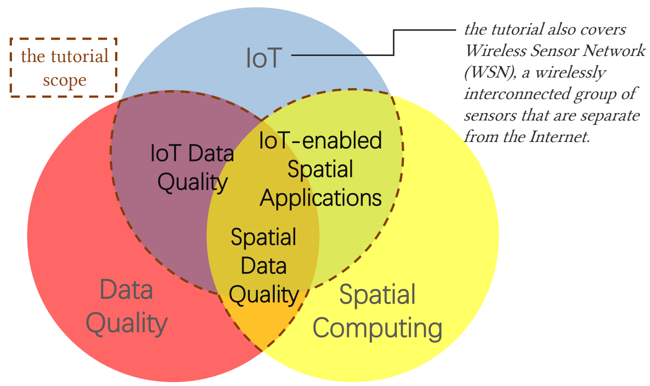

Figure 1: Tutorial scope.

However, applications are often challenged by a variety of SID quality issues, mostly caused by the distinct properties of the IoT [50]. First, IoT devices are often limited by production specifications and resources, resulting in erroneous, incomplete, or duplicated spatial information [67, 97, 130]. Second, the IoT is decentralized and encompasses a wide range of dynamic devices that emit and consume data, which can lead to excessive, deferred, disordered, or inconsistent data [8, 95, 122]. Third, IoT devices use diverse positioning technologies, causing the spatial information generated to be heterogeneous, potentially resulting in incompatible formats, resolutions, and semantics [49, 87, 124].

Since SID is a treasure trove of spatial information that may benefit many spatial applications [5, 47, 80], resolving quality issues is essential. Researchers are continuously making efforts on related topics, including a large number of recent studies [23, 62, 95, 98]. The popularity of the studies on SID quality issues is also evidenced by an increasing emergence of survey papers. However, most of existing survey papers focus on synergies between two of the three related areas, i.e., the IoT, data quality, and spatial computing, covering topics such as IoT data quality [11, 50, 71, 96], spatial data quality [28, 35, 38, 135, 140, 141], and IoT-enabled spatial applications [5, 80, 94]. Although several survey papers [66, 80] on IoT-enabled spatial applications mention quality issues, they do not analyze and summarize DQ technologies.

In contrast to existing survey papers, this tutorial consolidates the IoT, data quality (DQ), and spatial computing. The scope of the tutorial is shown in Figure 1. We organize the related work into two overall lines: 1) **SID quality management**, where the aim is to

The definitive Version of Record was published in SIGMOD'22, https://dl.acm.org/doi/10.1145/3514221.3522568.

control or enhance the quality of SID, and 2) **exploitation of low-quality SID**, where the focus is on querying, analysis, and decision-making over low-quality SID. For both of these lines of research, the goal of the tutorial is to provide unique insights to researchers who are interested in IoT DQ aspects and to practitioners who intend to develop IoT-enabled applications.

The tutorial utilizes the structure of our survey paper [59] in the ACM Computing Surveys. Due to the time limit, the tutorial will be a condensed version of our full survey paper, focusing on comparing the methodologies of representative works. We start by presenting a framework of SID quality aspects, covering the major DQ dimensions, data characteristics, and quality issues, as well as means to address quality issues. We then present key technologies for SID quality management, encompassing location refinement, uncertainty elimination, outlier removal, fault correction, data integration, and data reduction. Furthermore, we cover technologies for low-quality SID exploitation, addressing the tasks of querying, analysis, and decision-making. We end by describing emerging trends and open issues related to SID quality and identify research directions that are important for efficient, effective, and innovative quality-aware SID computing.

2 TUTORIAL OUTLINE

The intended **1.5-hour** tutorial will be tailored for the SIGMOD attendees who are aware of general data management topics, but may not be working on spatial IoT data. We use the first 5 minutes to present the overall background, challenges, and applications of SID, followed by 15 minutes to establish a general picture of SID quality aspects (see Section 2.1). We then cover SID quality management (see Section 2.2) and the exploitation of low-quality SID (see Section 2.3), each for 30 minutes. The last 10 minutes conclude with prospects on SID quality technologies (see Section 2.4).

2.1 SID Quality Framework (20 mins)

DQ Dimensions. DQ reflects how well data satisfies the purpose of data consumption [50]. Therefore, data consumers have their own criteria for assessing the DQ for a task at hand. These criteria, known as *DQ dimensions*, differ across application areas or scenarios. In this tutorial, we cover the most important data consumption requirements in IoT-enabled spatial applications, and based on this, we define and discuss the major DQ dimensions of spatial data in the IoT context.

SID is treated as observations of real phenomena or processes through IoT devices. There is inevitably a difference between the true states of the underlying phenomena or processes and the measurements due to imperfections in the IoT technologies [50, 60]. From a high-level perspective, quality requirements to SID posed by the consuming IoT-enabled applications span three aspects, each with several major DQ dimensions.

- SID should be *accurate* and *reliable*. In this setting, we review the concepts and applications of the DQ dimensions Precision, Accuracy, and Consistency.
- SID should be *comprehensive* and *informative*. Here, we introduce the DQ dimensions Time Sparsity, Space Coverage, Completeness, and Redundancy.
- SID should be *easy to use*. Here, we cover the DQ dimensions Latency, Staleness, Data Volume, Truth Volume, Resolution, and Interpretability.

Table 1: SID Characteristics and Resulting Quality Issues

SID Characteristic	Quality Issues (↓: low; ↑: high)
[omnipresent in IoT setting]	
Noisy and erroneous	↓ precision, ↓ accuracy, ↓ consistency
Temporally discrete	↑ time sparsity, ↓ completeness, ↑ staleness
Decentralized and heterogeneous	↓ consistency, ↑ latency, ↓ interpretability
Dynamic	↓ precision
Voluminous and duplicated	↑ redundancy, ↑ latency, ↑ data volume
Isolated and conflicting	↓ consistency, ↓ interpretability
Varying smoothly	-
Markovian	-
[specific in spatial data domain]	
Unverifiable	↓ truth volume
Hierarchical and multi-scaled	↓ consistency, ↓ resolution, ↓ interpretability
Spatially discrete	↓ space coverage
Spatially autocorrelated	-
Spatially anisotropic	-

Quality Issues. IoT devices continuously monitor variables of interest (e.g., position [94], check-ins [102], or air quality [60]) in specific spatial ranges using some form of positioning. Due to the particular working mechanism of IoT devices and the application need, SID is associated with characteristics. Identifying these characteristics helps find the causes of quality issues and the corresponding solutions. Table 1 presents a brief overview of SID characteristics and their resulting quality issues. A detailed analysis of SID characteristics and quality issues will be offered in the tutorial.

Means to Resolve DQ Issues. Referring to Figure 2, we organize DQ technologies from two perspectives.

Task Perspective. We consider the technologies according to the IoT layers that their tasks concern. The DQ tasks in the **perception** and **transport layers** optimize mainly the infrastructure (cf. our survey [59]). Taking into account the audience, we exclude these and focus on data handling for DQ in higher IoT layers.

- The **localization layer** estimates object locations, thus producing spatial data. A key DQ task is Location Refinement that accompanies or follows the positioning process and adjusts initial location estimates to reduce system and random errors. Its main goals concern ↑ precision, ↑ accuracy, and ↑ resolution.
- The **pre-processing layer** manages SID, encompassing DQ tasks that explicitly target improvements of input data quality. These tasks are 1) Uncertainty Elimination that reduces uncertain or imprecise measurements and imputes unknown measurements at unsampled points, thus addressing ↑ precision, ↑ completeness, ↑ resolution, and ↓ time sparsity; 2) Outlier Removal that detects and removes items in a collection that do not conform to their context, addressing ↑ precision, ↑ accuracy, and ↑ consistency; 3) Fault Correction that finds and repairs wrong, conflicting, or missing data values, addressing ↑ accuracy, ↑ consistency, and ↑ completeness; 4) Data Integration that obtains a unified data representation by comparing, combining, and fusing data sets from multiple sources, thereby addressing ↑ accuracy, ↑ completeness, ↑ data volume, ↑ resolution, and ↑ interpretability; 5) Data Reduction that converts a data set into a corrected and simplified form, addressing ↓ data volume, ↓ latency, and ↓ redundancy.

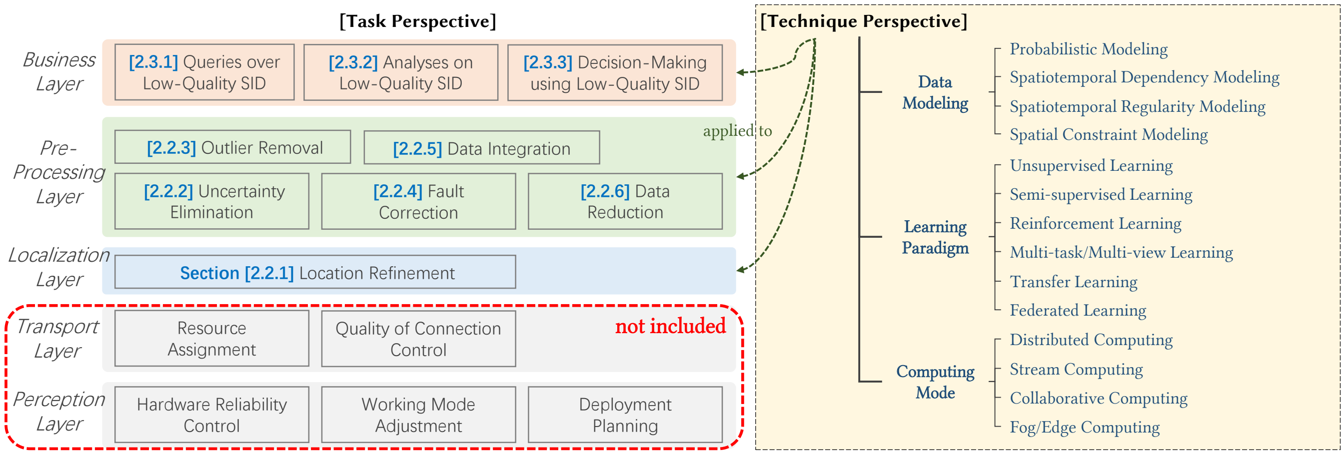

Figure 2: Task and technique perspective of the categorization of data quality technologies.

- The DQ tasks in the **business layer** aim to ensure that the data can support specific needs of diverse spatial applications. Concerning SID quality, these tasks span Querying, Analysis, and Decision-making in the setting of low-quality SID. Different subcategories of these tasks consider different DQ issues. We therefore do not list the specific DQ goals here.

Technique Perspective. We also categorize technologies according to different technical viewpoints.

- From a **data modeling** viewpoint, we categorize techniques into 1) Probabilistic Modeling that combats uncertainty and noise by generating probabilistic representations of observations [27] or results [128] in dynamic and complex settings; 2) Spatiotemporal Dependency Modeling that derives spatiotemporal correlations from the inherent characteristics of SID (including the characteristics of varying smoothly [138], Markovian [8, 108], spatially autocorrelated [60], and spatially anisotropic [7], as listed in Table 1) for handling noise [108, 138], missing or unknown values [7, 60], errors [8], etc.; 3) Spatiotemporal Regularity Modeling that targets the discovery and extraction of spatial and temporal regularities (often formed by external rules and factors derived from the context) [58, 108, 130, 132] from SID collections; and 4) Spatial Constraint Modeling that utilizes additional spatial and motion constraints to contend with noisy, incomplete, and faulty SID [20, 32, 108, 113].
- From a **learning paradigm** viewpoint, techniques choose appropriate schemes or strategies to mitigate low DQ issues in learning: 1) Unsupervised Learning like EM algorithm [41], AutoEncoders [23, 76], and GAN [23] can address the scarcity of labels (ground-truth data); 2) Semi-supervised Learning can address partial availability of labels (e.g., co-training [22]) and imbalanced labels (e.g., positive-unlabeled learning methods [18]); 3) Reinforcement Learning can address the incompleteness [99] and dynamics [98, 106] in sequential decision-making; 4) Multi-task Learning [83, 132] and Multi-view Learning [124, 126] can contend with scarcity of labels and bias/heterogeneity of data in training; 5) Transfer Learning [116], borrowing labeled data or knowledge from related domains, can address limited data availability and biased data; and 6) Federated Learning can address the scarcity of data across multiple domains [55] and facilitate decentralized model training [75].
- From a **computing mode** viewpoint, useful paradigms include: 1) Distributed Computing [111, 119] for improving system throughput and reducing single points of failure; 2) Stream Computing [19, 48, 62] for timely data exploitation; 3) Collaborative Computing for improving consistency, completeness, and availability of SID with multiple computing nodes [24, 127] and their data [128, 133] involved; and 4) Fog/Edge Computing [62, 130] for reducing data volumes, redundancy, latency, and staleness of SID by pushing computing tasks closer to data sources.

2.2 SID Quality Management (30 mins)

2.2.1 Location Refinement (LR). Given a set $\mathbf{x}$ of IoT measurements, a positioning function $\mathrm{f} : \mathbf{X} \mapsto \mathbf{Y}$ maps $\mathbf{x} \in \mathbf{X}$ to a location $\mathbf{y} \in \mathbf{Y}$. Due to the non-stationary and noisy nature of IoT measurements, $\mathbf{y}$ can be imprecise and erroneous. Adopting a probabilistic method, LR aims to find optimal result $\hat{\mathbf{y}} \in \mathbf{Y}$ that maximizes $\mathrm{P}(\mathbf{Y} \mid \mathbf{X}, F, C)$, where $F = \{\mathrm{f}_1, \ldots\}$ is a family of positioning functions and C refers to spatial constraints. Based on the specifics of $\mathbf{X}$, we consider three categories of LR technologies.

Ensemble LR. $\mathbf{X}$ refers to an individual object's multi-variable measurements at a *single* time point t_i, and the output $\hat{\mathbf{y}}$ is a location estimate at t_i. $\mathbf{X}$ may consist of different components that are measured by different sensors, including sensors of varying types. Within Ensemble LR, *single-source methods* [31] aggregate a set of possible results $\mathbf{y} = \{y_1, \ldots\}$ produced by a single process $\mathrm{f}(\mathbf{x})$; *multi-source methods* [21] involve multiple independent processes as $F = \{\mathrm{f}_1, \ldots\}$ and fuse their results for more accurate location $\hat{\mathbf{y}}$.

Motion-based LR. Here, $\mathbf{X}$ refers to an individual object's *sequential* observations where each observation can be single-variable or multivariable. Accordingly, the output $\hat{\mathbf{y}}$ is a location sequence. Motion-based LR introduces knowledge of motion dynamics and historical measurements to improve the positioning results, and this is achieved mainly by capturing spatiotemporal dependencies in observation sequences. Representative techniques include Bayes Filters [34], Probabilistic Graph Models [30], and Recurrent Neural Networks [40].

Collaborative LR. Here, **X** refers to *multiple* objects' observations at a single time. In the spirit of collaborative computing, collaborative LR optimizes all objects' positions altogether. Two subcategories are identified: *joint denoising* [127] assumes system noise and distills the actual locations by eliminating system noise that best meets a statistical hypothesis; *iterative optimization* [24] assumes random errors and iteratively reduces the random errors of a batch of observed locations.

2.2.2 Uncertainty Elimination (UE). We consider both imprecise measurements and unknown values at unmeasured points. We present trajectory UE and STID UE as trajectories and STID are used frequently in applications.

Trajectory UE roughly falls into three categories. *Calibration-based approaches* align noisy and incomplete trajectories with reference points or ranges obtained from maps [97] or extracted from a large set of trajectories [61, 97]. *Inference-based approaches* exploit structural regularities across trajectories to restore complete paths that connect observed locations of a trajectory, using explicit [108, 137] or implicit [65] spatial constraints. *Smoothing-based approaches* utilize temporal autocorrelation of consecutive data points to mitigate volatility [138].

STID UE has often been regarded as a *spatiotemporal interpolation* process, which estimates and inserts thematic values at unsampled location-time points that align with spatiotemporally nearby samples [7]. The interpolation performance degrades with the expansion of the spatiotemporal range covered, and data (with ground-truth labels) needs to be pre-analyzed for model selection. Recently, *data fusion* has been incorporated into reducing measurement uncertainty in STID [85]. One main challenge faced by data fusion is how to find additional relevant and reliable data sources.

2.2.3 Outlier Removal (OR). Probabilistic modeling [86, 113, 121], spatiotemporal dependencies [14] and regularity [121], and spatial constraints [138] have been used widely in OR.

Trajectory Point OR aims to remove location points that are clearly different from their nearby points and do not accord with expected mobility behavior. *Constraint-based methods* [113, 138] detect abnormal points that violate mobility constraints based on neighborhood information. Such methods may not contend well with dynamic and noisy trajectories. *Statistics-based methods* [86] detect anomalous points based on statistical profiling of a single or a set of trajectories. These methods may be restricted by the availability of historical data. *Prediction-based methods* [121] identify a value as an outlier if it differs from a value predicted from historical data. Outliers are then repaired with predicted values. Relying on accurate predictions, these methods entail trustworthy input data and regularly updated models.

STID OR targets *temporal*, *spatial*, or *spatiotemporal* outliers. The last refers to the items whose thematic attribute values deviate clearly from those of other items in their spatial and temporal neighborhoods. Temporal OR has been investigated systematically [15, 36], with trajectory point outliers being a special case. Aggarwal [4] reviews spatial and then spatiotemporal OR using spatial OR as an initial step; this study also reveals the close relationship between temporal OR and spatial OR when regarding temporal and spatial attributes as contextual attributes (as opposed to thematic attributes). Some classic studies specific to spatiotemporal OR are based on neighborhoods [14] or set theory [6]. Compared to neighborhood-based approaches, set theory-based approaches require holistic data and are more suitable for simple data attributes.

2.2.4 Fault Correction (FC). FC technologies are generally based on comparative analyses within or between data collections.

Trajectory FC mainly considers a type of *symbolic trajectory*, a time-ordered sequence of categorical values referring to the detecting sensors or covered regions. Symbolic trajectories are commonly seen in RFID, Infrared, and Bluetooth tracking scenarios, in which false negatives [20, 32, 45] occur when a sensor fails to detect an object, while false positives [8, 20, 32] occur when an object is detected by multiple sensors simultaneously. FC technologies generally use probabilistic modeling to detect and repair faults. In addition, many studies [8, 20, 32, 45] consider spatiotemporal regularities of interactions between sensors and objects, spatiotemporal dependencies among trajectory records, and spatial constraints caused by the sensor deployment and the underlying space.

STID FC repairs *faulty thematic values* [90] or *imprecise timestamps* [48, 95]. These methods rely mostly on modeling spatiotemporal dependencies among neighboring [48, 90] and autocorrelated [95] collections/(sub)sequences.

2.2.5 Data Integration (DI). These technologies are classified based on whether or not semantic aspects are involved or not.

Semantic DI involves semantic and comprehensible data sources and concerns their integration with raw SID to enrich the interpretability of the SID. *Semantic DI for trajectories* aims to annotate raw location traces with concepts/labels [58, 113] or complementary knowledge [84] at particular times or during time intervals, facilitating direct, concise, and explainable utilization of trajectories. These technologies often exploit spatiotemporal regularity incurred by geo-semantics (e.g., POI category [113] and spatial constraints [57, 58]). *Semantic DI for STID* enriches spatial data infrastructures (SDI) with *standardized* [10] or *application-specific* [9] geo-semantic meta-information. Edge computing [9] can be employed to efficiently assign semantics to data at the IoT far end.

Non-semantic DI compares and combines multifaceted spatiotemporal observations to eliminate inconsistencies and to enhance the reliability of the integrated data, relying mainly on pure spatiotemporal data dependencies. *Trajectory+trajectory* techniques target unified representations of trajectories in different formats [87] and scales [124], or using different ID systems [49]. *Trajectory+STID* techniques [125] attach spatial or spatiotemporal measurements to location points or segments based on similarities of their spatial or temporal attributes. Finally, *STID+STID* techniques [139] fuse multi-source spatiotemporal measurements according to their spatial and temporal commonality.

2.2.6 Data Reduction (DR). DR aims to improve throughput and computing efficiency in general while minimizing the loss of information as seen from the business layer.

Trajectory Compression compacts either raw trajectories [17, 54, 69, 73, 77, 82, 133] or network-constrained (map-matched) trajectories [39, 51, 62, 63, 115]. Each category can be divided further into online [54, 62, 69, 73, 73, 82, 106] and offline [17, 39, 51, 63, 77, 115, 133] approaches. The related notation of *trajectory simplification* [17, 54, 69, 73, 77, 82] can be regarded as a special form of compression that focuses on removing trajectory points and does not consider compression techniques such as binary encoding. A mainstream technology for trajectory simplification is error-bounded line simplification [70].

STID Reduction leverages compression [56, 101] or predictions [130]. *Compression-based approaches* can be divided further into lossless compression [101], for applications that demand accuracy, and lossy compression [56] that achieves a higher compression ratio with some precision loss. *Prediction-based approaches* [130] are often used to reduce communication data volume between IoT nodes. Data can be dropped if the prediction error is within an acceptable range. Compression-based approaches fit well in batch processing scenarios, while prediction-based approaches are challenged by the robustness and timeliness of prediction models.

2.3 Exploitation of Low-Quality SID (30 mins)

2.3.1 Queries over Low-Quality SID. The uncertainty, dynamics, and decentralization of data are three major obstacles to effective and efficient SID query processing.

Data Uncertainty is a key issue in spatial querying, and probabilistic modeling techniques are exploited widely to contend with this. In this setting, algorithms estimate upper and lower bounds of query objects based on probability models to enable priority-oriented processing and object pruning. A taxonomy of probabilistic spatial queries is available [27], and a recent survey [140] categorizes queries over uncertain spatial data. In contrast, the tutorial presents query processing techniques based on the type of location uncertainty they handle in the context of IoT-based positioning or tracking: First, to handle *uncertainty caused by location inaccuracy*, an object's location at a single time is usually described as a probability density function (pdf), which occurs in continuous (a closed-form distribution) [12, 13, 26, 68, 100] or discrete (a set of samples with occurrence probabilities) cases [43, 120, 131]. Second, to handle *uncertainty caused by discrete sampling*, a moving object's location(s) at unsampled time points is modeled by a distribution that is referenced to its sampled, known location(s) [3, 89]. The distribution can be modeled to infer the location at a single time point (e.g., uniform circular [114] or velocity vector [44]) or the locations across a time interval (e.g., particles [118], first-order Markovian grids [129], Markovian Gaussian distributions [46], combination of road segments [136], combination of sample connections [79], beads/necklaces [52, 103], etc.).

Data Dynamics bring issues of data volume, data evolution, and data skew to query processing. To efficiently process *queries over massive SID*, distributed computing [25, 81, 111, 119] and stream computing [25, 48, 81] have been employed. For *queries over evolving SID*, object locations and other information arrive in a streaming fashion. Safe regions [91] and incremental evaluation [123] have been proposed to reduce communication and computation overhead. For *queries over skewed SID*, node load-balancing [93] and data partitioning [93, 104] have been studied.

Data Decentralization poses challenges to processing encrypted data [117] and heterogeneous data [29, 112]. To enable outsourcing of queries on private location data, spatial and cryptographic transformation schemes [117] have been invented to balance efficiency and privacy. To enable spatial queries over heterogeneous location data sources, generic location representation [112] and a unified data management platform [29] have been proposed.

2.3.2 Analyses on Low-Quality SID. The tutorial categorizes existing analysis techniques targeting low-quality SID based mainly on quality issues related to uncertainty and dynamics. Within each category, studies are organized according to the tasks they consider.

Uncertainty in SID. To combat data inaccuracy and incompleteness, data analysis techniques often exploit probabilistic modeling [64, 67, 102], spatiotemporal dependencies [72, 74, 134], and spatial constraints [88, 102, 107]. Tasks span clustering [88], anomaly detection [72], frequent-pattern mining [64, 67, 102, 134], and popular-route discovery [107]. The existing techniques are generally batch-oriented and centralized, leaving techniques for real-time and decentralized settings as an open issue.

Dynamics (Volume and Evolution) in SID. To handle high data volumes in analytics, indexing and pruning [16, 105, 122], distributed computing [42, 110], and stream computing [19, 33] techniques have been proposed. Spatiotemporal dependency modeling and online learning [76, 109] have been utilized to facilitate the analysis of evolving SID. Typical applications are discussed, including clustering [105], anomaly detection [16, 19, 76, 109], frequent-pattern mining [122], and event discovery [33]. How to migrate the functionality covered above to edge devices to reduce cost and latency are highly relevant future topics.

2.3.3 Decision-Making using Low-Quality SID. A variety of decision-making tasks leverage SID, such as the prediction of next location(s) [23, 53, 55, 126], traffic volume [75, 99], and spatiotemporal variables [78, 83, 116]; the recommendation of POIs [41, 128]; and the planning of task assignments [98] and site selection [18]. Related studies are organized based on the DQ issues they address during learning as follows.

- **Scarcity of Labels** has been addressed in unsupervised learning [23, 41], semi-supervised learning [18], and multi-task learning [132].
- **Limited Availability and Bias of Data** have been addressed in transfer learning [116] and federated learning [55].
- **Uncertainty of Data** has been handled in probabilistic modeling [128] and reinforcement learning [99].
- **Dynamics of Data** has been explored in reinforcement learning [98], incremental learning [53], and edge computing [78].
- **Heterogeneity and Decentralization of Data** have been studied in multi-task learning [83] and multi-view learning [126] for integrating multi-source data, and in federated learning [75] for constructing decentralized models.

2.4 Trends and Future Directions (10 mins)

2.4.1 Emerging Trends. Having reviewed DQ technologies for a variety of tasks, we observe that SID quality management is being integrated with different learning techniques. Moreover, SID quality related computing is becoming increasingly relevant in dynamic, decentralized, and heterogeneous settings. The tutorial highlights several emerging trends, namely **Privacy-preserving Computing** (effective generation and exploitation of encrypted or obscured SID) [76, 117], **Edge/Fog Computing** (improving efficiency and reducing central, single-point workloads) [62, 130], **Reinforcement and Incremental Learning** (models with corresponding capabilities of dynamic and incremental processing) [98, 99, 106, 108], and **Comprehensive Data Fusion for Improved DQ** (integrating diverse and rich, but also biased, spatiotemporal data sources) [23, 55, 83, 99, 116, 124, 132].

2.4.2 Open Issues and Future Directions. Although many studies consider the quality of SID, no systematic studies exist on how to coordinate DQ technologies in IoT settings. The tutorial covers several promising directions from this perspective.

- **Dynamic DQ Modeling**, which is needed for guiding an individual IoT node's data handling and interactions with other nodes in heterogeneous and dynamic IoT architectures.
- **Secure SID Sharing**, which enables the discovery of valuable insights across IoT data repositories that currently form silos.
- **DQ-aware Task Planning**, which lays the foundation for efficient coordination of multiple DQ-related services.
- **Cross-layer DQ Management**, which aims to make DQ-related services sufficiently general to support diverse applications.
- **Quality Management Middleware for SID**, which serves to integrate the technical directions mentioned above.

3 TUTORIAL INFORMATION

Target Audience. Focusing on quality-aware SID management and utilization, the tutorial targets researchers with interests in DQ in IoT settings and practitioners who aim to develop IoT-enabled applications. The tutorial benefits attendees with different experiences. Beginners in the area will build an overall impression of spatial data quality in the context of the IoT and will learn about the latest achievements in DQ technologies. Experts in related topics will learn techniques and methodologies of particular DQ technologies in-depth and will gain insight into trends and new challenges in quality-aware SID computing.

Excellence. Building on a new ACM Computing Surveys paper [59], the tutorial provides a comprehensive introduction to cutting-edge developments in a good deal of sub-topics on multiple aspects of spatial IoT data quality. The tutorial spotlights the unique challenges of the IoT that are brought to spatial computing, and it expands substantially the techniques and methodology for handling trajectories and spatiotemporal data in IoT settings. By cutting across the IoT, data quality, and spatial computing, the tutorial is different from the KDD'21 tutorial presented by Gupta et al. [37] on DQ for machine learning, the CIKM'20 tutorial by Song and Zhang [96] on IoT data quality, and the ICDE'17 tutorial by Züfle et al. [141] on handling geospatial data uncertainties. Covering topics including data preprocessing tasks as well as querying, analysis, and decision-making, the tutorial inspires a wide spectrum of new research and applications related to spatial IoT data.

4 PRESENTERS

Huan Li [Homepage] is an EU Marie Curie IF Fellow and Assistant Professor with the Department of Computer Science, Aalborg University. He was an Alibaba engineer working on very-large-scale spatial data intelligence platform. He obtained his BSc from Sichuan University and PhD from Zhejiang University. His research interests lie in IoT data management and mobile computing. Most of his research works are published in top-tier database venues.

Bo Tang [Homepage] is an Assistant Professor with the Department of Computer Science and Engineering, SUSTech, China. He is leading the database group there since 2017. He received with BSc from Sichuan University and the PhD degree from Hong Kong Polytechnic University. He won ACM SIGMOD China Rising Star Award in 2021. He was a visiting researcher at CWI Amsterdam and MSRA. His research interests are in the area of data management (e.g., database/big-data systems, query optimization techniques).

Hua Lu [Homepage] is a Professor of Computer Science, Roskilde University, Denmark. He received the BSc and MSc degrees from Peking University, China, and the PhD degree in computer science from National University of Singapore. His research interests include databases, data mining, and data science. He has served as PC cochair or vice-chair for many international conferences. He received the Best Vision Paper Award at SSTD 2019.

Muhammad Aamir Cheema [Homepage] is an ARC Future Fellow and Associate Professor at the Faculty of Information Technology, Monash University, Australia. He obtained his PhD from UNSW Australia in 2011. He is the recipient of 2012 Malcolm Chaikin Prize for Research Excellence in Engineering, 2013 Discovery Early Career Researcher Award, 2014 Dean's Award for Excellence in Research by an Early Career Researcher, 2018 Future Fellowship, 2018 Monash Student Association Teaching Award, and 2019 Young Tall Poppy Science Award. He has also won two CiSRA best research paper of the year awards, and several best paper awards at conferences including ICDE, ICAPS, WISE, and ADC. He is the Associate Editor of IEEE TKDE and DAPD.

Christian S. Jensen [Homepage] is a professor of computer science at Aalborg University. He was recently at Aarhus University for three years and at Google for one year. His research concerns data analytics and management with focus on temporal and spatio-temporal data management. He is a fellow of the ACM and IEEE, and he is a member of the Academia Europaea, the Royal Danish Academy of Sciences and Letters, and the Danish Academy of Technical Sciences. He has received several awards for his research, most recently the 2019 TCDE Impact Award. He was Editor-in-Chief of ACM TODS from 2014 to 2020 and an Editor-in-Chief of The VLDB Journal from 2008 to 2014.

ACKNOWLEDGMENTS

The presenters' work was supported by the EU MSCA program (Grant No. 882232), the Digital Research Centre Denmark, the NSFC (Grant No. 61802163), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), and the ARC (FT180100140).

REFERENCES

[1] 2019. *Growing opportunities in the Internet of Things*. Retrieved Nov 2021 from https://www.mckinsey.com/industries/private-equity-and-principal-investors/our-insights/growing-opportunities-in-the-internet-of-things

[2] 2019. *IDC forecasts connected IoT devices to generate 79.4ZB of data in 2025*. Retrieved Nov 2021 from https://futureiot.tech/idc-forecasts-connected-iot-devices-to-generate-79-4zb-of-data-in-2025/

[3] Pankaj K Agarwal, Leonidas J Guibas, Herbert Edelsbrunner, Jeff Erickson, Michael Isard, Sariel Har-Peled, John Hershberger, Christian Jensen, Lydia Kavraki, Patrice Koehl, et al. 2002. Algorithmic issues in modeling motion. *Comput. Surveys* 34, 4 (2002), 550–572.

[4] Charu C Aggarwal. 2015. Outlier analysis. In *Data Mining*. 237–263.

[5] Isam Mashhour Al Jawarneh, Paolo Bellavista, Antonio Corradi, Luca Foschini, and Rebecca Montanari. 2020. Big spatial data management for the Internet of Things: A survey. *Journal of Network and Systems Management* 28, 4 (2020), 990–1035.

[6] Alessia Albanese, Sankar K Pal, and Alfredo Petrosino. 2012. Rough sets, kernel set, and spatiotemporal outlier detection. *IEEE Transactions on Knowledge and Data Engineering* 26, 1 (2012), 194–207.

[7] Annalisa Appice, Anna Ciampi, Donato Malerba, and Pietro Guccione. 2013. Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. *Journal of Spatial Information Science* 2013, 6 (2013), 119–153.

[8] Asif Iqbal Baba, Manfred Jaeger, Hua Lu, Torben Bach Pedersen, Wei-Shinn Ku, and Xike Xie. 2016. Learning-based cleansing for indoor RFID data. In *SIGMOD*. 925–936.

[9] Elarbi Badidi and Muthucumaru Maheswaran. 2018. Towards a platform for urban data management, integration and processing. In *IoTBDS*. 299–306.

[10] Garvita Bajaj, Rachit Agarwal, Pushpendra Singh, Nikolaos Georgantas, and Valérie Issarny. 2018. 4W1H in IoT semantics. *IEEE Access* 6 (2018), 65488–65506.

[11] Tanvi Banerjee and Amit Sheth. 2017. IoT quality control for data and application needs. *IEEE Intelligent Systems* 32, 2 (2017), 68–73.

[12] Thomas Bernecker, Tobias Emrich, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, and Andreas Züfle. 2011. A novel probabilistic pruning approach to speed up similarity queries in uncertain databases. In *ICDE*. 339–350.

[13] George Beskales, Mohamed A Soliman, and Ihab F Ilyas. 2008. Efficient search for the top-k probable nearest neighbors in uncertain databases. *Proceedings of the VLDB Endowment* 1, 1 (2008), 326–339.

[14] Derya Birant and Alp Kut. 2007. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering* 60, 1 (2007), 208–221.

[15] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. 2021. A review on outlier/anomaly detection in time series data. *Comput. Surveys* 54, 3 (2021), 1–33.

[16] Yingyi Bu, Lei Chen, Ada Wai-Chee Fu, and Dawei Liu. 2009. Efficient anomaly monitoring over moving object trajectory streams. In *KDD*. 159–168.

[17] Hu Cao, Ouri Wolfson, and Goce Trajcevski. 2006. Spatio-temporal data reduction with deterministic error bounds. *The VLDB Journal* 15, 3 (2006), 211–228.

[18] Chaoxiong Chen, Chao Chen, Chaocan Xiang, Songtao Guo, Zhu Wang, and Bin Guo. 2020. ToiletBuilder: A PU learning based model for selecting new public toilet locations. *IEEE Internet of Things Journal* (2020).

[19] Chao Chen, Daqing Zhang, Pablo Samuel Castro, Nan Li, Lin Sun, and Shijian Li. 2011. Real-time detection of anomalous taxi trajectories from GPS traces. In *Mobiquitous*. 63–74.

[20] Haiquan Chen, Wei-Shinn Ku, Haixun Wang, and Min-Te Sun. 2010. Leveraging spatio-temporal redundancy for RFID data cleansing. In *SIGMOD*. 51–62.

[21] Jian Chen, Gang Ou, Ao Peng, Lingxiang Zheng, and Jianghong Shi. 2018. An INS/WiFi indoor localization system based on the weighted least squares. *Sensors* 18, 5 (2018), 1458.

[22] Ling Chen, Yaya Cai, Yifang Ding, Mingqi Lv, Cuili Yuan, and Gencai Chen. 2016. Spatially fine-grained urban air quality estimation using ensemble semi-supervised learning and pruning. In *UbiComp*. 1076–1087.

[23] Xinyu Chen, Jiajie Xu, Rui Zhou, Wei Chen, Junhua Fang, and Chengfei Liu. 2021. TrajVAE: A Variational AutoEncoder model for trajectory generation. *Neurocomputing* 428 (2021), 332–339.

[24] Xiao Chen and Shengnan Zou. 2017. Improved Wi-Fi indoor positioning based on particle swarm optimization. *IEEE Sensors Journal* 17, 21 (2017), 7143–7148.

[25] Zhida Chen, Gao Cong, Zhenjie Zhang, Tom ZJ Fuz, and Lisi Chen. 2017. Distributed publish/subscribe query processing on the spatio-textual data stream. In *ICDE*. 1095–1106.

[26] Reynold Cheng, Jinchuan Chen, Mohamed Mokbel, and Chi-Yin Chow. 2008. Probabilistic verifiers: Evaluating constrained nearest-neighbor queries over uncertain data. In *ICDE*. 973–982.

[27] Reynold Cheng, Tobias Emrich, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, Goce Trajcevski, and Andreas Züfle. 2014. Managing uncertainty in spatial and spatio-temporal data. In *ICDE*. 1302–1305.

[28] Rodolphe Devillers, Alfred Stein, Yvan Bédard, Nicholas Chrisman, Peter Fisher, and Wenzhong Shi. 2010. Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Transactions in GIS* 14, 4 (2010), 387–400.

[29] Xin Ding, Lu Chen, Yunjun Gao, Christian S. Jensen, and Hujun Bao. 2018. UlTraMan: A unified platform for big trajectory data management and analytics. *Proceedings of the VLDB Endowment* 11, 7 (2018), 787–799.

[30] Frederike Dümbgen, Cynthia Oeschger, Mihailo Kolundžija, Adam Scholefield, Emmanuel Girardin, Johan Leuenberger, and Serge Ayer. 2019. Multi-modal probabilistic indoor localization on a smartphone. In *IPIN*. 1–8.

[31] Xuming Fang, Zonghua Jiang, Lei Nan, and Lijun Chen. 2018. Optimal weighted K-nearest neighbour algorithm for wireless sensor network fingerprint localisation in noisy environment. *IET Communications* 12, 10 (2018), 1171–1177.

[32] Bettina Fazzinga, Sergio Flesca, Filippo Furfaro, and Francesco Parisi. 2016. Exploiting integrity constraints for cleaning trajectories of RFID-monitored objects. *ACM Transactions on Database Systems* 41, 4 (2016), 1–52.

[33] Kaiyu Feng, Tao Guo, Gao Cong, Sourav S Bhowmick, and Shuai Ma. 2019. SURGE: Continuous detection of bursty regions over a stream of spatial objects. *IEEE Transactions on Knowledge and Data Engineering* 32, 11 (2019), 2254–2268.

[34] Davide Giovanelli, Elisabetta Farella, Daniele Fontanelli, and David Macii. 2018. Bluetooth-based indoor positioning through ToF and RSSI data fusion. In *IPIN*. 1–8.

[35] Michael F Goodchild. 2013. The quality of big (geo) data. *Dialogues in Human Geography* 3, 3 (2013), 280–284.

[36] Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. 2013. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2013), 2250–2267.

[37] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, et al. 2021. Data quality for machine learning tasks. In *KDD*. 4040–4041.

[38] Stephen C Guptill and Joel L Morrison. 2013. *Elements of Spatial Data Quality*. Elsevier.

[39] Yunheng Han, Weiwei Sun, and Baihua Zheng. 2017. COMPRESS: A comprehensive framework of trajectory compression in road networks. *ACM Transactions on Database Systems* 42, 2 (2017), 1–49.

[40] Minh Tu Hoang, Brosnan Yuen, Xiaodai Dong, Tao Lu, Robert Westendorp, and Kishore Reddy. 2019. Recurrent neural networks for accurate RSSI indoor localization. *IEEE Internet of Things Journal* 6, 6 (2019), 10639–10651.

[41] Saeid Hosseini, Hongzhi Yin, Xiaofang Zhou, Shazia Sadiq, Mohammad Reza Kangavari, and Ngai-Man Cheung. 2019. Leveraging multi-aspect time-related influence in location recommendation. *World Wide Web Journal* 22, 3 (2019), 1001–1028.

[42] Chunchun Hu, Xionghua Kang, Nianxue Luo, and Qiansheng Zhao. 2015. Parallel clustering of big data of spatio-temporal trajectory. In *ICNC*. 769–774.

[43] Ming Hua, Jian Pei, Wenjie Zhang, and Xuemin Lin. 2008. Ranking queries on uncertain data: A probabilistic threshold approach. In *SIGMOD*. 673–686.

[44] Yuan-Ko Huang, Chao-Chun Chen, and Chiang Lee. 2009. Continuous k-nearest neighbor query for moving objects with uncertain velocity. *GeoInformatica* 13, 1 (2009), 1–25.

[45] Shawn R Jeffery, Minos Garofalakis, and Michael J Franklin. 2006. Adaptive cleaning for RFID data streams. *Proceedings of the VLDB Endowment* 6, 163–174.

[46] Hoyoung Jeung, Hua Lu, Saket Sathe, and Man Lung Yiu. 2013. Managing evolving uncertainty in trajectory databases. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2013), 1692–1705.

[47] Hoyoung Jeung, Man Lung Yiu, and Christian S. Jensen. 2011. Trajectory pattern mining. In *Computing with Spatial Trajectories*. 143–177.

[48] Yuanzhen Ji, Hongjin Zhou, Zbigniew Jerzak, Anisoara Nica, Gregor Hackenbroich, and Christof Fetzer. 2015. Quality-driven continuous query execution over out-of-order data streams. In *SIGMOD*. 889–894.

[49] Fengmei Jin, Wen Hua, Thomas Zhou, Jiajie Xu, Matteo Francia, Maria Orowska, and Xiaofang Zhou. 2020. Trajectory-based spatiotemporal entity linking. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[50] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. 2016. Data quality in Internet of Things: A state-of-the-art survey. *Journal of Network and Computer Applications* 73 (2016), 57–81.

[51] Satoshi Koide, Yukihiro Tadokoro, Chuan Xiao, and Yoshiharu Ishikawa. 2018. CiNCT: Compression and retrieval for massive vehicular trajectories via relative movement labeling. In *ICDE*. 1097–1108.

[52] Bart Kuijpers, Rafael Grimson, and Walied Othman. 2011. An analytic solution to the alibi query in the space–time prisms model for moving object data. *International Journal of Geographical Information Science* 25, 2 (2011), 293–322.

[53] Arnab Kumar Laha and Sayan Putatunda. 2018. Real time location prediction with taxi-GPS data streams. *Transportation Research Part C: Emerging Technologies* 92 (2018), 298–322.

[54] Ralph Lange, Frank Dürr, and Kurt Rothermel. 2011. Efficient real-time trajectory tracking. *The VLDB Journal* 20, 5 (2011), 671–694.

[55] Anliang Li, Shuang Wang, Wenzhu Li, Shengnan Liu, and Siyuan Zhang. 2020. Predicting human mobility with federated learning. In *SIGSPATIAL*. 441–444.

[56] Bo Li, Omid Sarbishei, Hosein Nourani, and Tristan Glatard. 2018. A multidimensional extension of the lightweight temporal compression method. In *IEEE Big Data*. 2918–2923.

[57] Huan Li, Hua Lu, Muhammad Aamir Cheema, Lidan Shou, and Gang Chen. 2020. Indoor mobility semantics annotation using coupled conditional Markov networks. In *ICDE*. 1441–1452.
[58] Huan Li, Hua Lu, Gang Chen, Ke Chen, Qinkuang Chen, and Lidan Shou. 2020. Toward translating raw indoor positioning data into mobility semantics. *ACM/IMS Transactions on Data Science* 1, 4 (2020), 1–37.
[59] Huan Li, Hua Lu, Christian S. Jensen, Bo Tang, and Muhammad Aamir Cheema. 2022. Spatial data quality in the Internet of Things: Management, exploitation, and prospects. *Comput. Surveys* 55, 3, Article 57 (2022), 41 pages.
[60] Jiayu Li, Haoran Li, Yehan Ma, Yang Wang, Ahmed A Abokifa, Chenyang Lu, and Pratim Biswas. 2018. Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network. *Building and Environment* 127 (2018), 138–147.
[61] Lun Li, Xiaohang Chen, Qizhi Liu, and Zhifeng Bao. 2020. A data-driven approach for GPS trajectory data cleaning. In *DASFAA*. 3–19.
[62] Tianyi Li, Lu Chen, Christian S. Jensen, and Torben Bach Pedersen. 2021. TRACE: Real-time compression of streaming trajectories in road networks. *Proceedings of the VLDB Endowment* 14, 7 (2021), 1175–1187.
[63] Tianyi Li, Ruikai Huang, Lu Chen, Christian S. Jensen, and Torben Bach Pedersen. 2020. Compression of uncertain trajectories in road networks. *Proceedings of the VLDB Endowment* 13, 7 (2020), 1050–1063.
[64] Yuxuan Li, James Bailey, Lars Kulik, and Jian Pei. 2013. Mining probabilistic frequent spatio-temporal sequential patterns with gap constraints from uncertain databases. In *ICDM*. 448–457.
[65] Yang Li, Yangyan Li, Dimitrios Gunopulos, and Leonidas Guibas. 2016. Knowledge-based trajectory completion from sparse GPS samples. In *SIGSPATIAL*. 1–10.
[66] You Li, Yuan Zhuang, Xin Hu, Zhouzheng Gao, Jia Hu, Long Chen, Zhe He, Ling Pei, Kejie Chen, Maosong Wang, et al. 2020. Location-Enabled IoT (LE-IoT): A survey of positioning techniques, error sources, and mitigation. *arXiv preprint arXiv:2004.03738* (2020).
[67] Zhenhui Li and Jiawei Han. 2014. Mining periodicity from dynamic and incomplete spatiotemporal data. In *Data Mining and Knowledge Discovery for Big Data*. 41–81.
[68] Xiang Lian and Lei Chen. 2009. Efficient processing of probabilistic reverse nearest neighbor queries over uncertain data. *The VLDB Journal* 18, 3 (2009), 787–808.
[69] Xuelian Lin, Jiahao Jiang, Shuai Ma, Yimeng Zuo, and Chunming Hu. 2019. One-pass trajectory simplification using the synchronous Euclidean distance. *The VLDB Journal* 28, 6 (2019), 897–921.
[70] Xuelian Lin, Shuai Ma, Jiahao Jiang, Yanchen Hou, and Tianyu Wo. 2021. Error bounded line simplification algorithms for trajectory compression: An experimental evaluation. *ACM Transactions on Database Systems* 46, 3 (2021), 1–44.
[71] Caihua Liu, Patrick Nitschke, Susan P Williams, and Didar Zowghi. 2020. Data quality and the Internet of Things. *Computing* 102, 2 (2020), 573–599.
[72] Chuanren Liu, Hui Xiong, Yong Ge, Wei Geng, and Matt Perkins. 2012. A stochastic model for context-aware anomaly detection in indoor location traces. In *ICDE*. 449–458.
[73] Jiajun Liu, Kun Zhao, Philipp Sommer, Shuo Shang, Brano Kusy, Jae-Gil Lee, and Raja Jurdak. 2016. A novel framework for online amnesic trajectory compression in resource-constrained environments. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 2827–2841.
[74] Siyuan Liu and Shuhui Wang. 2016. Trajectory community discovery and recommendation by multi-source diffusion modeling. *IEEE Transactions on Knowledge and Data Engineering* 29, 4 (2016), 898–911.
[75] Yi Liu, JQ James, Jiawen Kang, Dusit Niyato, and Shuyu Zhang. 2020. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet of Things Journal* 7, 8 (2020), 7751–7763.
[76] Yiding Liu, Kaiqi Zhao, Gao Cong, and Zhifeng Bao. 2020. Online anomalous trajectory detection with deep generative sequence modeling. In *ICDE*. 949–960.
[77] Cheng Long, Raymond Chi-Wing Wong, and HV Jagadish. 2014. Trajectory simplification: On minimizing the direction-based error. *Proceedings of the VLDB Endowment* 8, 1 (2014), 49–60.
[78] Haidong Luo, Hongming Cai, Han Yu, Yan Sun, Zhuming Bi, and Lihong Jiang. 2019. A short-term energy prediction system based on edge computing for smart city. *Future Generation Computer Systems* 101 (2019), 444–457.
[79] Chunyang Ma, Hua Lu, Lidan Shou, and Gang Chen. 2012. KSQ: Top-k similarity query on uncertain trajectories. *IEEE Transactions on Knowledge and Data Engineering* 25, 9 (2012), 2049–2062.
[80] Mohammad Saeid Mahdavinejad, Mohammadreza Rezvan, Mohammadamin Barekatain, Peyman Adibi, Payam Barnaghi, and Amit P Sheth. 2018. Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks* 4, 3 (2018), 161–175.
[81] 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 Basalamah. 2015. Tornado: A distributed spatio-textual stream processing system. *Proceedings of the VLDB Endowment* 8, 12 (2015), 2020–2023.
[82] Jonathan Muckell, Jeong-Hyon Hwang, Vikram Patil, Catherine T Lawson, Fan Ping, and SS Ravi. 2011. SQUISH: An online approach for GPS trajectory compression. In *COM.Geo*. 1–8.
[83] Long H Nguyen, Jiazhen Zhu, Zhe Lin, Hanxiang Du, Zhou Yang, Wenxuan Guo, and Fang Jin. 2019. Spatial-temporal multi-task learning for within-field cotton yield prediction. In *PAKDD*. 343–354.
[84] Tales P Nogueira, Reinaldo B Braga, Carina T de Oliveira, and Hervé Martin. 2018. FrameSTEP: A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications* 92 (2018), 533–545.
[85] Nwamaka U Okafor, Yahia Alghorani, and Declan T Delaney. 2020. Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. *ICT Express* 6, 3 (2020), 220–228.
[86] Vikram Patil, Priyanka Singh, Shivam Parikh, and Pradeep K Atrey. 2018. Geosclean: Secure cleaning of GPS trajectory data using anomaly detection. In *MIPR*. 166–169.
[87] Douglas Alves Peixoto, Xiaofang Zhou, Nguyen Quoc Viet Hung, Dan He, and Bela Stantic. 2018. A system for spatial-temporal trajectory data integration and representation. In *DASFAA*. 807–812.
[88] Nikos Pelekis, Ioannis Kopanakis, Evangelos E Kotsifakos, Elias Frentzos, and Yannis Theodoridis. 2011. Clustering uncertain trajectories. *Knowledge and Information Systems* 28, 1 (2011), 117–147.
[89] Dieter Pfoser and Christian S. Jensen. 1999. Capturing the uncertainty of moving-object representations. In *SSD*. 111–131.
[90] Sitthapon Pumpichet, Niki Pissinou, Xinyu Jin, and Deng Pan. 2012. Belief-based cleaning in trajectory sensor streams. In *ICC*. 208–212.
[91] Jianzhong Qi, Rui Zhang, Christian S. Jensen, Kotagiri Ramamohanarao, and Jiayuan He. 2018. Continuous spatial query processing: A survey of safe region based techniques. *Comput. Surveys* 51, 3 (2018), 1–39.
[92] M Mazhar Rathore, Awais Ahmad, Anand Paul, and Seungmin Rho. 2016. Urban planning and building smart cities based on the Internet of Things using big data analytics. *Computer Networks* 101 (2016), 63–80.
[93] Suprio Ray, Bogdan Simion, Angela Demke Brown, and Ryan Johnson. 2013. A parallel spatial data analysis infrastructure for the cloud. In *SIGSPATIAL*. 284–293.
[94] Rathin Chandra Shit, Suraj Sharma, Deepak Puthal, and Albert Y Zomaya. 2018. Location of Things (LoT): A review and taxonomy of sensors localization in IoT infrastructure. *IEEE Communications Surveys & Tutorials* 20, 3 (2018), 2028–2061.
[95] Shaoxu Song, Ruihong Huang, Yue Cao, and Jianmin Wang. 2021. Cleaning timestamps with temporal constraints. *The VLDB Journal* (2021), 1–22.
[96] Shaoxu Song and Aoqian Zhang. 2020. IoT data quality. In *CIKM*. 3517–3518.
[97] Han Su, Kai Zheng, Haozhou Wang, Jiamin Huang, and Xiaofang Zhou. 2013. Calibrating trajectory data for similarity-based analysis. In *SIGMOD*. 833–844.
[98] Lijun Sun, Xiaojie Yu, Jiachen Guo, Yang Yan, and Xu Yu. 2021. Deep reinforcement learning for task assignment in spatial crowdsourcing and sensing. *IEEE Sensors Journal* (2021).
[99] Xianfeng Tang, Boqing Gong, Yanwei Yu, Huaxiu Yao, Yandong Li, Haiyong Xie, and Xiaoyu Wang. 2019. Joint modeling of dense and incomplete trajectories for citywide traffic volume inference. In *WWW*. 1806–1817.
[100] Yufei Tao, Xiaokui Xiao, and Reynold Cheng. 2007. Range search on multidimensional uncertain data. *ACM Transactions on Database Systems* 32, 3 (2007), 15–es.
[101] Joseph Euzebe Tate. 2015. Preprocessing and Golomb-Rice encoding for lossless compression of phasor angle data. *IEEE Transactions on Smart Grid* 7, 2 (2015), 718–729.
[102] Shan-Yun Teng, Wei-Shinn Ku, and Kun-Ta Chuang. 2017. Toward mining stop-by behaviors in indoor space. *ACM Transactions on Spatial Algorithms and Systems* 3, 2 (2017), 1–38.
[103] Goce Trajcevski, Alok Choudhary, Ouri Wolfson, Li Ye, and Gang Li. 2010. Uncertain range queries for necklaces. In *MDM*. 199–208.
[104] Hoang Vo, Ablimit Aji, and Fusheng Wang. 2014. SATO: A spatial data partitioning framework for scalable query processing. In *SIGSPATIAL*. 545–548.
[105] Sheng Wang, Zhifeng Bao, J Shane Culpepper, Timos Sellis, and Xiaolin Qin. 2019. Fast large-scale trajectory clustering. *Proceedings of the VLDB Endowment* 13, 1 (2019), 29–42.
[106] Zheng Wang, Cheng Long, and Gao Cong. 2021. Trajectory simplification with reinforcement learning. In *ICDE*.
[107] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. 2012. Constructing popular routes from uncertain trajectories. In *KDD*. 195–203.
[108] Hao Wu, Jiangyun Mao, Weiwei Sun, Baihua Zheng, Hanyuan Zhang, Ziyang Chen, and Wei Wang. 2016. Probabilistic robust route recovery with spatio-temporal dynamics. In *KDD*. 1915–1924.
[109] Hao Wu, Weiwei Sun, and Baihua Zheng. 2017. A fast trajectory outlier detection approach via driving behavior modeling. In *CIKM*. 837–846.
[110] Yanbo Wu, Hong Shen, and Quan Z Sheng. 2014. A cloud-friendly RFID trajectory clustering algorithm in uncertain environments. *IEEE Transactions on Parallel and Distributed Systems* 26, 8 (2014), 2075–2088.

[111] Dong Xie, Feifei Li, and Jeff M Phillips. 2017. Distributed trajectory similarity search. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1478–1489.

[112] Jianqiu Xu and Ralf Hartmut Güting. 2013. A generic data model for moving objects. *GeoInformatica* 17, 1 (2013), 125–172.

[113] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. 2013. Semantic trajectories: Mobility data computation and annotation. *ACM Transactions on Intelligent Systems and Technology* 4, 3 (2013), 1–38.

[114] Bin Yang, Hua Lu, and Christian S. Jensen. 2009. Scalable continuous range monitoring of moving objects in symbolic indoor space. In *CIKM*. 671–680.

[115] Xiaochun Yang, Bin Wang, Kai Yang, Chengfei Liu, and Baihua Zheng. 2017. A novel representation and compression for queries on trajectories in road networks. *IEEE Transactions on Knowledge and Data Engineering* 30, 4 (2017), 613–629.

[116] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. 2019. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *WWW*. 2181–2191.

[117] Man Lung Yiu, Gabriel Ghinita, Christian S. Jensen, and Panos Kalnis. 2010. Enabling search services on outsourced private spatial data. *The VLDB Journal* 19, 3 (2010), 363–384.

[118] Jiao Yu, Wei-Shinn Ku, Min-Te Sun, and Hua Lu. 2013. An RFID and particle filter-based indoor spatial query evaluation system. In *EDBT*. 263–274.

[119] Haitao Yuan and Guoliang Li. 2019. Distributed in-memory trajectory similarity search and join on road network. In *ICDE*. 1262–1273.

[120] Liming Zhan, Ying Zhang, Wenjie Zhang, and Xuemin Lin. 2015. Finding top k most influential spatial facilities over uncertain objects. *IEEE Transactions on Knowledge and Data Engineering* 27, 12 (2015), 3289–3303.

[121] Aoqian Zhang, Shaoxu Song, and Jianmin Wang. 2016. Sequential data cleaning: A statistical approach. In *SIGMOD*. 909–924.

[122] Chao Zhang, Yu Zheng, Xiuli Ma, and Jiawei Han. 2015. Assembler: Efficient discovery of spatial co-evolving patterns in massive geo-sensory data. In *KDD*. 1415–1424.

[123] Dongxiang Zhang, Zhihao Chang, Sai Wu, Ye Yuan, Kian-Lee Tan, and Gang Chen. 2020. Continuous trajectory similarity search for online outlier detection. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[124] Desheng Zhang, Tian He, and Fan Zhang. 2019. National-scale traffic model calibration in real time with multi-source incomplete data. *ACM Transactions on Cyber-Physical Systems* 3, 2 (2019), 1–26.

[125] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *MobiCom*. 201–212.

[126] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. coMobile: Real-time human mobility modeling at urban scale using multi-view learning. In *SIGSPATIAL*. 1–10.

[127] Guolong Zhang, Ping Wang, Haibing Chen, and Lan Zhang. 2019. Wireless indoor localization using convolutional neural network and Gaussian process regression. *Sensors* 19, 11 (2019), 2508.

[128] Lu Zhang, Zhu Sun, Jie Zhang, Horst Kloeden, and Felix Klanner. 2020. Modeling hierarchical category transition for next POI recommendation with uncertain check-ins. *Information Sciences* 515 (2020), 169–190.

[129] Meihui Zhang, Su Chen, Christian S. Jensen, Beng Chin Ooi, and Zhenjie Zhang. 2009. Effectively indexing uncertain moving objects for predictive queries. *Proceedings of the VLDB Endowment* 2, 1 (2009), 1198–1209.

[130] Xiaopu Zhang, Jun Lin, Zubin Chen, Feng Sun, Xi Zhu, and Gengfa Fang. 2018. An efficient neural-network-based microseismic monitoring platform for hydraulic fracture on an edge computing architecture. *Sensors* 18, 6 (2018), 1828.

[131] Ying Zhang, Xuemin Lin, Yufei Tao, Wenjie Zhang, and Haixun Wang. 2011. Efficient computation of range aggregates against uncertain location-based queries. *IEEE Transactions on Knowledge and Data Engineering* 24, 7 (2011), 1244–1258.

[132] Pengpeng Zhao, Anjing Luo, Yanchi Liu, Fuzhen Zhuang, Jiajie Xu, Zhixu Li, Victor S Sheng, and Xiaofang Zhou. 2020. Where to go next: A spatio-temporal gated network for next POI recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[133] Yan Zhao, Shuo Shang, Yu Wang, Bolong Zheng, Quoc Viet Hung Nguyen, and Kai Zheng. 2018. REST: A reference-based framework for spatio-temporal trajectory compression. In *KDD*. 2797–2806.

[134] Zhou Zhao, Da Yan, and Wilfred Ng. 2012. Mining probabilistically frequent sequential patterns in uncertain databases. In *EDBT*. 74–85.

[135] Kai Zheng and Han Su. 2015. Go beyond raw trajectory data: Quality and semantics. *IEEE Data Engineering Bulletin* 38, 2 (2015), 27–34.

[136] Kai Zheng, Goce Trajcevski, Xiaofang Zhou, and Peter Scheuermann. 2011. Probabilistic range queries for uncertain trajectories on road networks. In *EDBT*. 283–294.

[137] Kai Zheng, Yu Zheng, Xing Xie, and Xiaofang Zhou. 2012. Reducing uncertainty of low-sampling-rate trajectories. In *ICDE*. 1144–1155.

[138] Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology* 6, 3 (2015), 1–41.

[139] Xiaolin Zhu, Fangyi Cai, Jiaqi Tian, and Trecia Kay-Ann Williams. 2018. Spatiotemporal fusion of multisource remote sensing data: Literature survey, taxonomy, principles, applications, and future directions. *Remote Sensing* 10, 4 (2018), 527.

[140] Andreas Züfle. 2020. Uncertain spatial data management: An overview. *arXiv preprint arXiv:2009.01121* (2020).

[141] 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 *ICDE*. 1467–1470.