Spatial Data Quality in the Internet of Things: Management, Exploitation, and Prospects

HUAN LI, Aalborg University, Denmark
HUA LU, Roskilde University, Denmark
CHRISTIAN S. JENSEN, Aalborg University, Denmark
BO TANG, Southern University of Science and Technology, China
MUHAMMAD AAMIR CHEEMA, Monash University, Australia

With the continued deployment of the Internet of Things (IoT), increasing volumes of devices are being deployed that emit massive spatially referenced data. Due in part to the dynamic, decentralized, and heterogeneous architecture of the IoT, the varying and often low quality of spatial IoT data (SID) presents challenges to applications built on top of this data. This survey aims to provide unique insight to practitioners who intend to develop IoT-enabled applications and to researchers who wish to conduct research that relates to data quality in the IoT setting. The survey offers an inventory analysis of major data quality dimensions in SID and covers significant data characteristics and associated quality considerations. The survey summarizes data quality related technologies from both task and technique perspectives. Organizing the technologies from the task perspective, it covers recent progress in SID quality management, encompassing location refinement, uncertainty elimination, outlier removal, fault correction, data integration, and data reduction; and it covers low-quality SID exploitation, encompassing querying, analysis, and decision-making techniques. Finally, the survey covers emerging trends and open issues concerning the quality of SID.

1 INTRODUCTION

The Internet of Things (IoT) interconnects massive numbers of devices to enable functionality such as ubiquitous perception and communication and smart decision-making [152, 185, 246]. IoT plays a pivotal role in many verticals, including in application related to smart cities [87, 152, 182], smart transportation [198], smart buildings [113, 230], smart healthcare [125], and smart energy [9, 207]. With an annual growth rate of 25% in smart, interconnected "things" (e.g., sensors, actuators, wearables, and vehicles) [5], we will witness explosive growth in IoT data collected from the physical world. Market intelligence firm IDC (International Data Corporation) predicts that the volume of data generated by IoT devices will reach 80ZB by 2025 [6]. As a concrete example of an IoT vertical, a smart meter project in Germany produces over 25TB of data per day [9]. As another indication of the growth in data volumes, research by the company Hazelcast [7] reports that the full rollout of 5G networks will increase the number of interconnected mobile devices per square kilometer from the current 4000 to 1 million and will incur high-speed data streams on a vast scale.

In the geographic information and mobile computing communities, IoT data is envisioned as a huge treasure trove since a considerable proportion of IoT devices and the data they generate are spatially referenced [186]. On the one hand, many IoT devices can self-localize through GPS. On the other hand, positioning technologies enabled by the wireless communication infrastructure and wireless and ambient devices have been integrated widely into the IoT infrastructure (called Location of Things [186]) to provide spatial references to other IoT devices. We call such spatially

Authors' addresses: Huan Li, Aalborg University, Aalborg, Denmark, lihuan@cs.aau.dk; Hua Lu, Roskilde University, Roskilde, Denmark, luhua@ruc.dk; Christian S. Jensen, Aalborg University, Aalborg, Denmark, csj@cs.aau.dk; Bo Tang, Southern University of Science and Technology, Shenzhen, China, tangb3@sustech.edu.cn; Muhammad Aamir Cheema, Monash University, Melbourne, Australia, aamir.cheema@monash.edu.

 The definitive Version of Record was published in *ACM Computing Surveys*, https://doi.org/10.1145/3498338.

Table 1. Related Review Papers on IoT, Data Quality (DQ), and Spatial Computing (SC)

Reference	Scope			Concerns
	IoT	DQ	SC	
Li et al. [117]	-	✓	✓	characteristics of big geospatial data and issues in handling them
Goodchild [70, 71]	-	✓	✓	quality and uncertainty issues of big geodata processing
Guptill and Morrison [76]	-	✓	✓	elements of evaluating spatial data quality
Devillers et al. [58]	-	✓	✓	recent achievement and open issues in improving spatial data quality
Li et al. [108]	-	✓	✓	quality assessment tools and uncertainty-aware spatial analytics
Züfle et al. [284, 285]	-	✓	✓	major challenges in handling uncertain geospatial data
Zheng and Su [279]	-	✓	✓	quality and semantics of raw trajectory data
Tsai et al. [208]	✓	◦	-	features of IoT data and data mining techniques for IoT
Siow et al. [187]	✓	◦	-	IoT and big data analytics in creating applications and services
Mohammadi et al. [158]	✓	◦	-	deep learning for IoT big data and stream data analytics
Karkouch et al. [95]	✓	✓	-	IoT factors endangering data quality and IoT outlier detection
Banerjee et al. [24]	✓	✓	-	human-in-the-loop for IoT data quality control
Ann and Wagh [16]	✓	✓	-	IoT data testing layer for data quality assurance
Perez-Castillo et al. [175]	✓	✓	-	IoT data quality in smart, connected product (SCP) environments
Liu et al. [131]	✓	✓	-	IoT data quality dimensions and related measurement methods
Song and Zhang [189]	✓	✓	-	deep learning for validity, completeness, and consistency of IoT data
Shit et al. [186]	✓	-	✓	analysis and taxonomy of IoT-based localization techniques
Javarneh et al. [12]	✓	-	✓	cloud-based big spatial data management frameworks for IoT
Mahdavinejad et al. [152]	✓	◦	◦	machine learning methods for IoT smart cities
Chen et al. [47]	✓	◦	✓	robustness, security, and privacy of IoT-enabled Location-based Services
Li et al. [122]	✓	◦	✓	error sources and mitigation methods of IoT-signal-based localization
Ours	✓	✓	✓	**quality management of SID and exploitation of low-quality SID**

✓ focused, ◦ partially covered, - not mentioned

referenced data from IoT devices *spatial IoT data* (SID). Two important special cases of SID are distinguished: *trajectories*, as time series of location values; and *spatiotemporal IoT data* (STID), general sensory data values with temporal and spatial references.

SID represents frequent observations in potentially large spatial regions, thus offering an exciting foundation for new insights to be utilized in queries, analyses, and decision-making in diverse applications. For example, one study [163] uses spatiotemporal data collected from urban traffic systems to enable dynamic and flexible congestion control. Another study [182] demonstrates how analyses of massive trajectories and STID can contribute to smart city construction.

However, quality issues associated with SID have become an obstacle for IoT-enabled spatial applications [47]. These issues are due to a variety of properties of the IoT, including the following three. First, IoT devices often have limited capabilities or limited resources that cause the generated spatial information to be erroneous, incomplete, or duplicated [122, 123, 191, 267]. Second, the IoT is decentralized and spans potentially massive numbers of devices that continuously collect and emit data. This can lead to excessive, deferred, disordered, or inconsistent spatial and spatiotemporal information [21, 188, 258]. Third, IoT devices are diverse and may use different positioning technologies, having the effect that the generated spatial information may be heterogeneous and may have incompatible formats, resolutions, and semantics [193, 230].

Since SID is a vital resource that drives spatial applications, addressing its quality issues is of high significance—in some cases, it is even essential. Not surprisingly, many recent studies [118, 154, 188, 264, 268] focus on SID quality issues. All such works can be divided into two overall lines of study: some studies aim to control or enhance the quality of SID, while other studies focus on querying, analyses, and decision-making over low-quality SID. These two lines of work, namely **SID quality management** and **exploitation of low-quality SID**, are the focus of this survey.

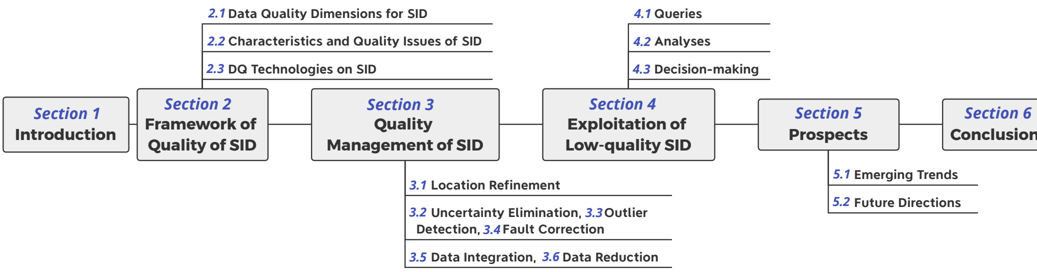

Fig. 1. The survey organization.

Note also the notion of Wireless Sensor Network (WSN) that relates to IoT: a WSN denotes a wirelessly interconnected group of sensors that are separate from the Internet [98]. Within the purview of WSNs, a range of technologies have been invented that are also relevant to IoT. Thus, this survey also covers WSN data quality technologies [18, 63, 106, 197, 248, 251].

Specifically, the survey concerns the intersection of three research areas, namely IoT, data quality, and spatial computing. Table 1 summarizes the scope and technical concerns of the most recent related surveys on IoT, data quality (DQ), and spatial computing (SC). Some existing surveys focus on synergies between two of the three areas, covering topics such as IoT data quality, spatial data quality, and IoT-enabled spatial applications. However, no existing surveys integrate all three research areas in their coverage. Next, some surveys cover spatial computing partially and address selected topics in spatial computing. In particular, Mahdavinejad et al. [152] present machine learning methods related to challenges presented by big IoT data, using smart cities as the main use case. Their study does not focus on DQ technologies. Also, Chen et al. [47] survey solutions for improving the robustness, security, and privacy of the Location-based Services in IoT systems, and Li et al. [122] review IoT-signal-based localization systems, covering localization error sources and mitigation methods. These two studies concern IoT-based localization algorithms—they do not cover a broader range of quality management techniques for SID, and nor do they consider the exploitation of low-quality SID.

In contrast to existing surveys, this survey aims to provide unique insights to practitioners who intend to develop IoT-enabled spatial applications and to researchers who are interested in IoT DQ aspects, from the perspective of quality management and quality-aware data exploitation.

The organization of this survey is illustrated in Fig. 1.

- Section 2 provides an overview of the quality aspects of SID, covering data quality dimensions related to SID, characteristics and quality issues of SID, and DQ technologies relevant to SID.
- Section 3 presents key classical quality management techniques for SID, encompassing the tasks of location refinement, uncertainty elimination, outlier removal, fault correction, data integration, and data reduction.
- Section 4 reviews the most recent works on the exploitation of low-quality SID, encompassing the tasks of querying, analyses, and decision-making.
- Section 5 discusses emerging trends and open issues related to SID quality, identifying research directions that are important in order to enable efficient, effective, and innovative quality-aware SID computing.
- Section 6 concludes the paper.

2 SPATIAL IOT DATA QUALITY FRAMEWORK

Data Quality (DQ) refers to how well data satisfies the purpose of data consumption [95, 175]. In this sense, each data consumer has her/his own DQ criteria for capturing how the data fits her/his task at hand. These criteria are also called *DQ dimensions* [95], and they encompass aspects such as accuracy, completeness, and interpretability. In Section 2.1, we introduce a set of DQ dimensions specific to SID. Given these specific DQ dimensions, we analyze the characteristics of spatial data in the IoT context and identify associated SID quality issues in Section 2.2. Finally, we present DQ technologies for SID from both task and technique perspectives in Section 2.3.

2.1 Data Quality Dimensions for SID

Applications are often associated with particular sets of DQ dimensions that take into account their particular data consumption purposes. DQ dimensions differ across application areas or scenarios even if they have the same name. For example, timeliness is considered as "the most recent time when the data is updated" for a snapshot query processing task [112], and as "the average of the difference between the recording time and current processing time" in a timestamp cleaning task [188]. In this survey, we will investigate the most important data consumption requirements in IoT-enabled spatial applications, and based on this, we analyze and define the major DQ dimensions of spatial data in the IoT context.

SID, including trajectories and STID, is regarded as observations of some real phenomenon or process through IoT facilities, which can be exploited as input to spatial queries, analyses, decision-making, and so on. 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 [95, 113, 131]. IoT deployments generally need to observe a variety of constraints, e.g., cost constraints, and application-level restrictions such as throughput, energy consumption, and privacy policy [95]. From a high-level perspective, quality requirements to SID posed by the consuming IoT-enabled applications span the following aspects.

- SID should be *accurate* and *reliable*. The most basic attribute of SID is location. If there is a deviation in the location, the information it points to is inaccurate and unreliable, which may lead to unsound and untrustworthy query, analysis, and decision-making results [47, 117].
- SID should be *comprehensive* and *informative*. SID serves as the medium to perceive the environment, while IoT digitization results in a certain degree of information loss in the SID. Spatial computing tasks benefit from complete and meaningful SID that preserves critical information on the environment [158, 208].
- SID should be *easy to use*. SID is inherently high-speed, dynamic, and geo-distributed, which makes large-scale exploitation difficult. SID should be ready at hand such that computing with large-scale SID can be realized readily and at a low cost. Moreover, SID is generally collected from heterogeneous devices and therefore differs in format, spatial resolution, and semantics. SID is expected to be simple in format, compatible, and human-readable [175, 187].

In accordance with the above three aspects, we list major DQ dimensions for SID and their meanings in Table 2. As the notion of data quality is open-ended, the DQ dimensions in Table 2 are non-exhaustive. Also, as mentioned above, DQ dimensions with the same name may carry different definitions in different applications. Nevertheless, the DQ dimensions to a large extent reflect the major DQ requirements of IoT-enabled spatial applications.

2.2 Characteristics and Quality Issues of SID

IoT devices continuously monitor variables of interest (e.g., position [186], check-in behavior [203], air quality [130], or electricity consumption [9, 57]) in specific spatial ranges using some form

Table 2. DQ Dimensions Specific to SID

Requirements	DQ Dimension	Meaning
Accurate and Reliable	Precision	The degree to which repeated data values, e.g., measurements, are similar, which can be modeled as the reciprocal of variance.
	Accuracy	The maximum absolute error ϵ such that all data values fall in the interval $[\mu - \epsilon, \mu + \epsilon]$, where μ refers to the true value [95].
	Consistency	The degree to which the available data from different sources match and support each other in a defined spatiotemporal range.
Comprehensive and Informative	Time Sparsity	The maximum time interval between two consecutive data items.
	Space Coverage	The ratio of the area that embraces the location measurements to the area that the IoT system is expected to cover.
	Completeness	The ratio of observed items to the missing ones in a spatiotemporal range.
	Redundancy	The ratio of non-distinct items to all items in a spatiotemporal range.
Easy to use	Latency	The average difference between the time when data is generated and processed.
	Staleness	The difference between the current time and the last time of update.
	Data Volume	The number of data items participating in a computing task.
	Truth Volume	The number of data items having the corresponding true values.
	Resolution	The level of detail of the information that can be provided to a computing task.
	Interpretability	The degree to which the format and meaning of the data items are clear and understandable for a computing task.

of localization. As a result, SID is often associated with specific characteristics. Identifying these characteristics helps find the causes of quality issues. Also, some SID characteristics in turn help address DQ issues. Table 3 summarizes the SID characteristics and their resulting **quality issues**. In particular, some SID characteristics can be regarded as omnipresent in IoT settings (termed 'IoT-omnipresent'), while others are mainly brought about by spatial aspects (termed 'Spatial-specific'). Moreover, a characteristic and its resulting quality issues can relate to the *spatial attribute* or the *thematic attribute* of SID. According to our definition in Section 1, trajectories and STID have spatial attributes, while thematic attributes (i.e., general data values) only exist in STID. As an example, the characteristic *temporally discrete* can be reflected in both spatial and thematic attributes. Also, temporal discreteness tends to yield increased time sparsity, lower completeness, and increased staleness as fewer data points are seen across time.

One notable property of SID is the inherent dependencies among data items in terms of their spatial and temporal aspects. As described in Table 3, the characteristics *spatially autocorrelated* and *spatially anisotropic* characterize spatial dependencies, *Markovian* characterizes temporal dependencies, and *varying smoothly* characterizes both spatial and temporal dependencies. As will be introduced in Section 2.3, techniques for the modeling of spatiotemporal dependencies can help to address DQ issues in SID.

2.3 DQ Technologies on SID

We categorize DQ technologies according to two facets in Fig. 2. From the system architecture perspective, we divide the technologies according to the tasks distributed to different IoT layers (see Section 2.3.1). From the technique perspective, we differentiate among technologies in terms of their data modeling methods, learning paradigms, and computing modes (see Section 2.3.2).

2.3.1 Task Facet. An IoT system adopts a layered approach to organizing its data acquisition, management, and exploitation tasks [16]. To serve spatial applications, an IoT system usually consists of five layers as follows.

Table 3. SID Characteristics and Their Resulting Quality Issues

	Characteristic	**Description**	**Quality Issues**	**Reflected Attr.**	
				Spatial	*Thematic*
IoT-omnipresent	Noisy and erroneous	Device capability limitations and hardware failures cause data uncertainty, noise, and faulty values [122, 191].	low precision, low accuracy, low consistency	✓	✓
	Temporally discrete	The reporting times of data items are not continuous due to the sampling strategy of the IoT devices [121, 281].	low time sparsity, low completeness, high staleness	✓	✓
	Decentralized and heterogeneous	Data stems from IoT devices scattered over the physical space, and these devices' generation mechanisms and data formats vary [197].	low consistency, high latency, low interpretability	✓	✓
	Dynamic	Data is reported continuously and is evolving, and data nodes disconnect irregularly or change strategies [259].	low precision	✓	✓
	Voluminous and duplicated	Devices are connected to the IoT that report data in a high-frequency and repetitive manner [9, 267].	high redundancy, high latency, high data volume	✓	✓
	Isolated and conflicting	Data nodes of different authorities are isolated from each other, and inconsistency is caused by differences in data handling methods at the nodes [170].	low consistency, low interpretability	✓	✓
	Varying smoothly	Physical variables within a spatial or temporal range exhibit smooth variation w.r.t. a particular target [95].	-	✓	✓
	Markovian	A data value is dependent on values generated at previous timestamps [95].	-	✓	✓
Spatial-specific	Unverifiable	Locations are hard to verify due to a limited volume and coverage of true values [71].	low truth volume	✓	
	Hierarchical and multi-scaled	Spatial attributes often exist at different spatial scales [253]. Even symbolic localization results have this issue.	low consistency, low resolution, low interpretability	✓	
	Spatially discrete	Localization results appear only in a fixed set of positions, or the value range of the localization algorithm is non-continuous or non-interpolable [195].	low space coverage	✓	
	Spatially autocorrelated	Data observations in nearby locations tend to resemble each other, instead of being statistically independent [92].	-		✓
	Spatially anisotropic	Spatial dependencies among data values are non-uniform in different directions [92].	-		✓

The **perception layer** manages IoT sensors that collect raw data, which involves the following DQ tasks. 1) *Hardware Reliability Control* combats loss of precision, reading dropping, and fail-dirty [14] by upgrading sensors or sensor components to ones with improved durability, performance-per-watt, and environmental adaptability. Its main goals include precision↑, accuracy↑, and consistency↑ (throughout this paper, we use ↑ to mean lifting and ↓ to mean lowering). 2) *Working Mode Adjustment* improves the devices' capabilities at data acquisition. For example, lifting (or lowering) a sensor's sampling frequency can combat time sparseness (or duplicates). As another example, setting a higher power mode for a wireless hotspot can expand the hotspot's space coverage. 3) *Deployment Planning* formulates the optimal deployment solution that concerns

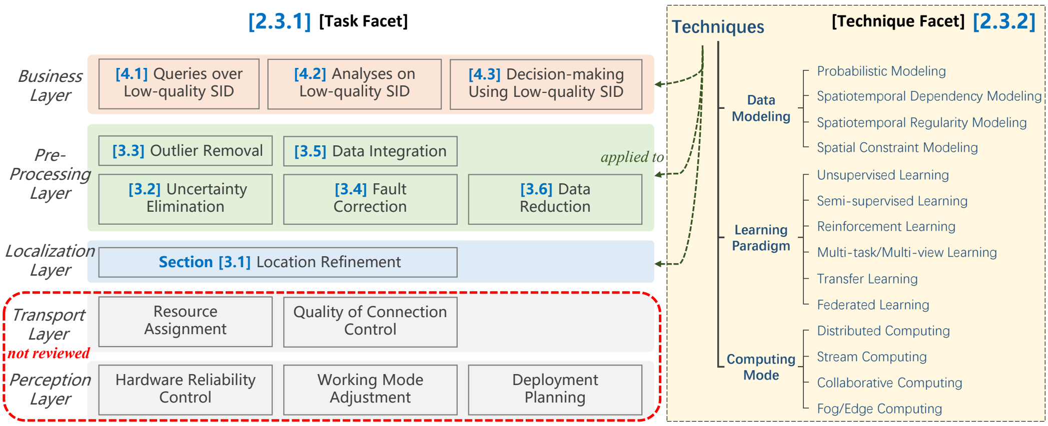

Fig. 2. Task and technique facets of the categorization of DQ technologies (the detailed uses of techniques in task are presented in lower-level categorization diagrams in Sections 3 and 4).

sensor installation, reference point selection, and calibration [66, 156]. This mainly helps achieve precision↑, accuracy↑, space coverage↑, latency↓, and truth volume↑.

The **transport layer** uses communication technology to enable coordination among devices and transmission of data. This layer involves two DQ tasks. 1) *Quality of Connection Control* ensures stable and flexible connectivity and interoperability among roaming devices to address completeness↑, latency↓, and staleness↓ [11]. 2) *Resource Assignment* addresses effective allocation of data, CPU, memory, and storage to IoT nodes [212], which targets latency↓ and staleness↓. Key enabling technologies include the computation offloading [150] and transport SDN [161].

The above technologies optimize mainly the infrastructure. In the remaining part of the survey, we exclude these and focus on data handling for DQ at higher IoT layers.

The **localization layer** estimates object locations that are assigned to data, thus producing spatial data. Here, a key DQ task is the *Location Refinement* (LR)—a process that accompanies or follows the localization process to adjust initial location estimates to reduce system and random errors. Its main goals concern precision↑, accuracy↑, and resolution↑. The concrete techniques are articulated in Section 3.1.

The **pre-processing layer** manages SID, involving the DQ tasks listed in Table 4. These DQ tasks explicitly target improvements of input data quality to serve business applications better.

Unlike the DQ tasks in the pre-processing layer, the DQ tasks in the **business layer** aim to ensure that the data can support the specific needs of diverse spatial applications. Concerning SID quality, these tasks include *Querying over Low-quality SID* (Section 4.1), *Analyses on Low-quality SID* (Section 4.2), and *Decision-making Using Low-quality SID* (Section 4.3). To be detailed in Section 4, different subcategories of these tasks consider different quality issues in their utilized SID. We therefore do not list the specific DQ goals for them here.

2.3.2 *Technique Facet.* We summarize the techniques that address DQ issues from three viewpoints.

From the viewpoint of **data modeling**, the following techniques construct different data representations or models according to the specific characteristics of the data.

- *Probabilistic Modeling* combats uncertainty and noise by introducing probabilistic representations of observations [53] or results [264], this way preserving all possibilities of the target variables. Statistical optimization methods are employed to address dynamic and complex settings [134].

Table 4. DQ Tasks in the Pre-processing Layer

DQ Task	Description	Main DQ Goals
Uncertainty Elimination (Section 3.2)	Uses time series or batch analysis methods to a) reduce uncertain or imprecise measurements and b) impute unknown measurements at unsampled points [257].	precision↑, completeness↑, resolution↑, and time sparsity↓
Outlier Removal (Section 3.3)	Detects and removes items in a data collection that do not conform to their context [10].	precision↑, accuracy↑, and consistency↑
Fault Correction (Section 3.4)	Finds and repairs wrong, conflicting, or missing data values based on comparative analyses within or between data collections [256].	accuracy↑, consistency↑, and completeness↑
Data Integration (Section 3.5)	Obtains a unified data representation by comparing, combining, and fusing data collections from multiple sources [22].	accuracy↑, completeness↑, data volume↑, resolution↑, and interpretability↑
Data Reduction (Section 3.6)	Converts a data collection into a corrected and simplified form based on statistical techniques, by either eliminating invalid and meaningless data or by reconstructing summary or statistical data at different aggregation levels [207].	data volume↓, latency↓, and redundancy↓

- *Spatiotemporal Dependency Modeling* derives spatiotemporal correlations from the inherent characteristics of SID (including varying smoothly [282], Markovian [19, 226], spatially autocorrelated [113], and spatially anisotropic [195] as introduced in Section 2.2). Spatiotemporal dependencies are then incorporated into the handling of noise [226, 282], missing or unknown values [113, 195], errors [19], etc.
- *Spatiotemporal Regularity Modeling* facilitates inference or prediction by discovering and extracting spatial and temporal regularities [110, 225, 226, 267, 273] of large-scale SID collections. Compared with the inherent characteristics of SID, data regularity is often formed by the rules and factors derived from the context, e.g., user preference and semantics of physical entities.
- *Spatial Constraint Modeling* utilizes additional spatial and motion constraints to contend with noisy, incomplete, and faulty SID. Such constraints include, but are not limited to, the topology of road networks [226, 281] and indoor buildings [109], maximum allowed speeds [239, 268, 282], and predefined rules associated with locations and regions [43, 64, 264].

From a **learning paradigm** perspective, techniques choose appropriate schemes or strategies to mitigate low DQ issues in learning. Due to the diversity of related techniques under development, we give only a brief, non-exhaustive review[1].

- *Unsupervised Learning* such as Expectation-Maximization (EM) [82], AutoEncoders (AE) [48, 89, 142, 236], and Generative Adversarial Networks (GAN) [48] can address the scarcity of labels (ground truth data).
- *Semi-supervised Learning* can deal with partial availability of labels (e.g., co-training [46]) and imbalance of labels (e.g., positive-unlabeled (PU) learning methods [40]).
- *Reinforcement Learning*, widely used in sequential decision-making, can deal with the incompleteness [199] and dynamics [104, 192, 218] of trajectories or spatiotemporal sequences.
- *Multi-task Learning* [67, 164, 253, 273] and *Multi-view Learning* [260, 262, 272], which make full use of data for improved overall performance, can contend with scarcity of labels, as well as bias and heterogeneity of data during training.
- *Transfer Learning* [72, 245], borrowing labeled data or knowledge from related domains, can deal with limited data availability and bias of data in a certain domain.
- *Federated Learning* can deal with the scarcity of data across multiple domains [105, 155] and facilitate decentralized model training [141].

[1]In this survey, we do not cover the most basic, heavily used supervised learning as a specific learning paradigm.

From a **computing mode** viewpoint, typical computing paradigms are listed below.

- *Distributed Computing* [162, 231, 248, 252] distributes data and resources among different system components, improving the throughput and overall efficiency of the system (for lower latency and staleness) and reducing single points of failure and system errors (for increased completeness).
- *Stream Computing* [42, 91, 118] processes and forwards data items generated in real-time within a time-limited window and buffer. It is an effective means to enable timely data exploitation.
- *Collaborative Computing* improves the performance of a computing task by coordinating multiple computing nodes [36, 49, 263] and combing their data and intermediate computing results [264, 275]. It helps improve the consistency, completeness, and availability of SID to be exploited for a particular task.
- *Fog/Edge Computing* [118, 161, 267] pushes data and algorithms to nodes that are situated where, or near to where, data is collected, addressing the issues of latency and throughput in systems with large amounts of data. This reduces data volumes and redundancy, as well as latency and staleness of SID.

2.3.3 Connections between Tasks and Techniques. Referring to Fig. 2, different techniques apply to different tasks, and some tasks may involve and assemble multiple techniques. In the following two sections, the literature is organized from the task perspective. When reviewing existing work related to a task, the low-level association between the applicable techniques and the particular task is analyzed and highlighted. E.g., Fig. 3 shows how different techniques are linked to a subcategory of location refinement technologies. Furthermore, Section A.1 in the Supplementary Material shows the associations between the DQ tasks and the DQ techniques from a global viewpoint.

3 QUALITY MANAGEMENT OF SID

This section elaborates on selected technologies that control and improve the quality of SID before they are exploited for business purposes, including location refinement (Section 3.1) in the localization layer and uncertainty elimination (Section 3.2), outlier removal (Section 3.3), fault correction (Section 3.4), data integration (Section 3.5), and data reduction (Section 3.6) in the pre-processing layer.

3.1 Location Refinement (LR)

Given a set $\mathbf{x}$ of measurements from an IoT infrastructure, localization of $\mathbf{x}$ is performed by an algorithm that can be modeled as a function $\mathrm{f} : \mathbf{X} \mapsto \mathbf{Y}$ that maps measurements such as $\mathbf{x} \in \mathbf{X}$ to a location $\mathbf{y} \in \mathbf{Y}$. Due to the inherent non-stationary and noisy nature of IoT measurements (e.g., Wi-Fi signal strengths and RFID readings) [133], the result $\mathbf{y}$ can be imprecise and erroneous. Adopting a probabilistic approach, the objective of LR is to find optimal localization results $\hat{\mathbf{y}} \in \mathbf{Y}$ that maximize the conditional probability $\mathrm{P}(\mathbf{Y} \mid \mathbf{X}, F, C)$, where $F = \{\mathrm{f}_1, \ldots\}$ is a family of functions each corresponding to a localization process and C refers to spatial constraints that can be utilized for refinement. According to the specifics of the input $\mathbf{X}$, we divide LR technologies into three main categories as illustrated in Fig. 3, where dashed arrows indicate DQ techniques that have been used widely in a DQ task or its subcategory.

In an **Ensemble LR** method, $\mathbf{X}$ refers to an individual object's multi-variable measurements at a *single* time point t_i. Here, $\mathbf{X} = \mathbf{X}_i = \{X_i^{(1)}, \ldots, X_i^{(M)}\}$, where $X_i^{(j)}$ $(1 \leq j \leq M)$ is a measured variable at t_i; and the final output $\hat{\mathbf{y}} = \hat{\mathbf{y}}_i$ is a location estimate at time t_i. The variables in $\mathbf{X}$ can be measured by different sensors, including sensors of varying types. Ensemble LR aims to assemble multiple localization results generated from $\mathbf{x} \in \mathbf{X}$ to output a statistically optimal result. Ensemble LR mainly follows the idea of probabilistic modeling. We distinguish between single-source and multi-source ensemble LR.

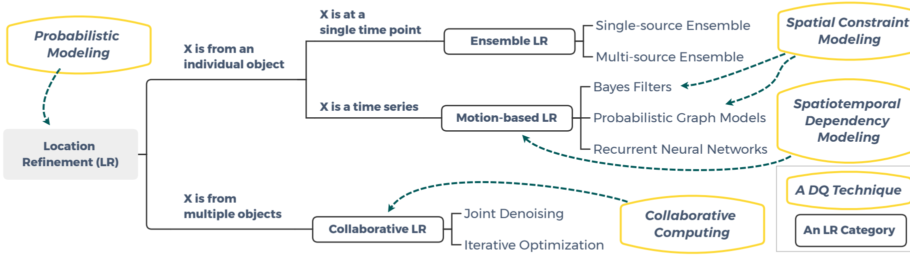

Fig. 3. Categories of location refinement technologies and key DQ techniques.

Single-source ensemble LR aggregates a set of possible localization results $\mathbf{y} = \{y_1, \ldots\}$ produced by a single localization process $\mathrm{f}(\mathbf{x})$. Fang et al. [63] study a weighted k-nearest neighbor (WkNN) method that determines the final location $\hat{\mathbf{y}}$ as the weighted mean of the top-k location estimates from $\mathrm{f}(\mathbf{x})$, i.e., $\hat{\mathbf{y}} = \sum_{j=1}^{k} \omega_j \cdot y_j$. The weight ω_j is modeled as the likelihood $\mathrm{P}(y_j \mid \mathbf{x})$.

In contrast, *multi-source ensemble LR* involves multiple independent localization processes as $F = \{\mathrm{f}_1, \ldots\}$ and fuses their localization results to improve the accuracy of $\hat{\mathbf{y}}$. Here, F can contain different localization algorithms such as lateration/angulation, RSSI (Received Signal Strength Indicators) fingerprinting, and dead reckoning [133]. Each may use a different combination of variables from $\mathbf{X}$ to estimate a location. Chen et al. [45] integrate results of RSSI fingerprinting and dead reckoning that suffer from signal fluctuations and time-growing error propagation, respectively. They use a weighted least squares (WLS) algorithm to combine linearly a fixed number of the highest confidence fingerprinting estimates by minimizing the relative error to the true location. The weight of a fingerprinting estimate is modeled as an exponential function related to the credibility of the dead reckoning. Using a hierarchical procedure, Dai et al. [55] employ a deep neural network (DNN) to generate a candidate reference location set from RSSI measurements, followed by an improved kNN algorithm to interpolate the final result upon the candidate set.

While multi-source ensembles require multi-aspect information from a more complex deployment setting, this also means that better location accuracy is possible than with a single-source ensemble.

In a **Motion-based LR** method, $\mathbf{X}$ refers to an individual object's *sequential* measurements, i.e., $\mathbf{X} = \mathbf{X}_{1:N} = \langle \mathbf{X}_1, \ldots, \mathbf{X}_N \rangle$, where $\mathbf{X}_i$ $(1 \leq i \leq N)$ can be the single-variable or multivariable measurement observed at time t_i. Accordingly, the final output is $\hat{\mathbf{y}} = \langle \hat{y}_1, \ldots, \hat{y}_N \rangle$. As the accuracy and robustness of localization at a single time point are affected adversely by time-varying noise, motion-based LR introduces knowledge of motion dynamics and historical measurements to improve the current localization result over time. Generally, motion-based LR relies on the modeling of spatiotemporal dependencies in localization sequences. Representative techniques for modeling spatiotemporal dependencies include Bayes Filters [18, 69, 202, 229, 247], Probabilistic Graph Models (PGM) [60, 134], and Recurrent Neural Networks (RNN) [80].

Bayes Filters sequentially estimate a dynamic system's state (the target object's current location) from noisy observations by capturing the uncertainty at each time point t_i as a probability distribution $\mathrm{P}(\mathbf{X}_i)$. Yim et al. [247] design an Extended Kalman Filter to linearize the trilateration results modeled with Additive White Gaussian Noise. The correlation between two consecutive estimates is captured as a Kalman filtering process, i.e., $y_{i+1} = \mathbf{A}_i y_i + \mu_i$ $(1 \leq i < N)$, where $\mathbf{A}_i$ is a state transition matrix and μ_i is a system error that follows a Gaussian distribution. Assuming $\mathbf{X}_i$ consists of multi-source sensory data, Giovanelli et al. [69] calculate the velocity based on RSSIs and the distances to Bluetooth hotspots based on Time-of-Flight measurements. The correlation of the velocity and distances is captured by a second-order, linear Kalman Filter to help reduce noise in the localization sequence. While Kalman Filter and its variants assume linear motion and Gaussian measurement noise, Particle Filters (PF) can make use of more sophisticated non-linear and non-Gaussian models. Wu et al. [229] propose an improved PF to evaluate the joint posterior $\mathsf{P}(y_{1:N} \mid z_{1:N}, u_{1:N})$ at time t_N given the RSSIs $z_{1:N}$ and inertial mesurements $u_{1:N}$ from timestamps t_1 to t_N. Based on a sequential Monte Carlo process, they sample a set of particles q whose weight is estimated based on the likelihoods $\mathsf{P}(z_i \mid y_i^{(q)})$ and $\mathsf{P}(y_i^{(q)} \mid y_{i-1}^{(q)}, u_i)$. This way, a particle that fits better with RSSIs and motion dynamics is more likely to be sampled in the next timestamp. PF has also been applied to the refinement of locations using minimalist spatial information from *binary sensor networks* [18]. In this setting, binary values $\{-1, 1\}$ indicate whether a device is approaching or is moving away from an anchored sensor node. Unlike all the above studies, and assuming unknown sensor node locations, Taylor et al. [202] propose a Bayes Filtering framework for location tracking that simultaneously localizes and calibrates the sensor nodes.

PGMs are more suitable for scenarios where object locations are modeled as discrete and piecewise constant states. Liu et al. [134] propose a Hidden Markov Model (S, O, A, B, π) to fuse observations O from smartphone sensors and WLAN signals. In particular, each hidden state $s_j \in S$ corresponds to a grid-based location; the emission probability set $B = \{b_i(s_j) = \mathsf{P}(o_i \mid X_i = s_j)\}$ and the initial state distribution π are estimated by RSSI fingerprinting algorithm; and state transition probabilities in A are calculated and refined using motions derived from smartphone sensor data. Assuming locations can only be at a set of predefined reference points, Dümbgen et al. [60] use a linear-chain Conditional Random Field (CRF) for LR such that the physical connectivity of reference points in a floorplan is captured as links between states at consecutive timestamps. The conditional probability of states $y_{1:N}$ given multi-modal observations $\mathbf{x}_{1:N}$ can be represented by a product of potential functions $\mathsf{P}(y_{1:N} \mid \mathbf{x}_{1:N}) \propto \prod_{i=2}^{N} \phi(y_{i-1}, y_i, \mathbf{x}_i)$. Each such potential function considers the motion between y_{i-1} and y_i as well as the reliability of result y_i given $\mathbf{x}_i \in \mathbf{X}_i$.

RNNs excel at capturing intricate sequential dependencies of observations and results. Hoang et al. [80] study different architectures, such as Multiple-RSSI-In-Single-Location-Out (MISO) and Multiple-RSSIs-In-Multiple-Locations-Out (MIMO), to output an optimal location at a single point or an optimal location sequence. In the case of multiple-location output, sliding window averaging is applied to reduce the accumulated errors. They report that a predicted-location-augmented-MISO LSTM (long short-term memory) achieves the best robustness among different combinations of architecture and RNN models.

Motion-based LR models all require much historical data for training. Also, motion-based LR is difficult to implement in a decentralized computing setting. We compare the three categories of models mentioned above. First, RNNs use more training data than PGMs and far more than Bayes Filters. Second, RNNs often achieve relatively better performance in complex scenes. Third, PGMs can explicitly incorporate mobility knowledge and therefore are suitable for scenarios with known space information.

In a **Collaborative LR** method, $\mathbf{X}$ refers to *multiple* objects' observations at a single time point, i.e., $\mathbf{X} = \mathbf{X}_O = \{\mathbf{X}^{o_1}, \ldots, \mathbf{X}^{o_M}\}$, where $O = \{o_1, \ldots, o_M\}$ is the corresponding object set. In the spirit of collaborative computing, collaborative LR optimizes the results globally as $\{\hat{\mathbf{y}}^{o_1}, \ldots, \hat{\mathbf{y}}^{o_M}\}$. The ideas include *joint denoising* [263, 271] and *iterative optimization* [49, 165].

Joint denoising assumes that any observed location is a combination of the actual location and system noise. Therefore, it separates the system noise that best meets a statistical hypothesis from collective observations to distill the actual locations. Assuming that the errors of a Convolutional Neural Network (CNN) location estimator are Gaussian, Zhang et al. [263] use Gaussian Process Regression to jointly adjust the coordinates of a batch of CNN-estimated locations. To handle

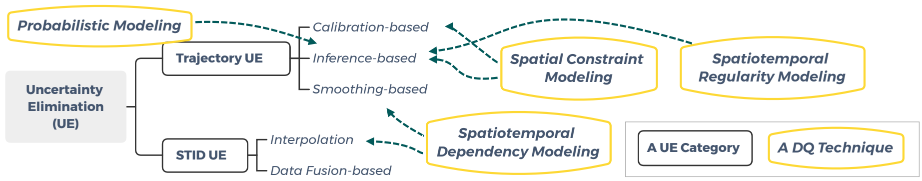

Fig. 4. Categories of uncertainty elimination technologies and key DQ techniques.

non-Gaussian estimation noise, Zhang et al. [271] use Gaussian Mixture-Semidefined Programming to optimize collective results.

Iterative optimization assumes random errors of observed locations and then reduces iteratively the random errors of the observed locations together. Niculescu and Nath [165] propose DV-Hop to optimize the locations of distributed target nodes based on their peer-to-peer hop counts. By sharing hop counts between nodes in the network, they use least squares to derive the medium size of a hop along with the unknown locations of the nodes according to the distribution of anchor nodes (with known locations). Modeling trilateral estimates as particles, Chen and Zou [49] use Particle Swarm Optimization (PSO) to adjust iteratively the particles' locations based on the gains of their location fitness to anchor nodes.

Collaborative LR requires a large number of objects (devices) for data and control coordination, which is a challenge in IoT settings with dynamic changes in connectivity.

Remarks. Most LRs are based on probabilistic modeling. Spatiotemporal dependencies (e.g., Markovian) are utilized widely in motion-based LR, and spatial constraints can be incorporated into Bayes Filters [229] and PGMs [60, 134]. Motion-based LR usually achieves higher accuracy compared to ensemble and collaborative LR that refine results at a single time point. However, motion-based LR often requires a mass of true location values (ground truth) to parameterize the model.

3.2 Uncertainty Elimination (UE)

The uncertain information subjected to UE includes imprecise measurements and unknown values at unmeasured points (see Table 4). Fig. 4 shows UE technologies that target trajectories or STID and indicates DQ techniques that are highly relevant to different categories of technologies.

Trajectory UE can be divided into *calibration-based* [115, 191], *inference-based* [87, 110, 121, 226, 281], and *smoothing-based* [31, 282] approaches.

Calibration-based approaches align noisy and incomplete trajectories with reference points or ranges obtained from maps [191] or extracted from collective trajectory data [115, 191]. Su et al. [191] collect different kinds of stable anchors (e.g., POIs and turning points) and align raw noisy trajectory locations with the anchors for heterogeneous trajectory comparison. Li et al. [115] derive smooth and continuous route skeletons over historical trajectory point clouds and consider the local distributions of points around skeleton points to eliminate deviations. Choosing significant and robust references is a challenge for these approaches that also have to consider updating the references according to environmental changes.

Inference-based approaches exploit structural regularities in collective trajectories to restore a complete path that connects all observed locations of a trajectory. Some studies utilize the topology of road networks [87, 226, 281] or indoor spaces [110] explicitly. Wu et al. [226] recover an optimal route $\hat{R}$ between two location-time records (l_s, t_s) and (l_e, t_e) based on MAP (Maximum a Posteriori) over the posterior $\mathrm{P}(R \mid l_s, t_s, l_e, t_e, \mathcal{T})$, where R is a candidate route and $\mathcal{T}$ contains all historical trajectories. The posterior is decomposed into the product of $\mathrm{P}(\Delta t \mid R, l_s, l_e, t_e, \mathcal{T})$ and

$P(R \mid l_s, l_e, t_e, \mathcal{T})$. The former captures the likelihood of $\Delta t = t_e - t_s$ over the expected time of R, and the latter computes the posterior of a route regardless of Δt based on a Markov Decision Process—an inverse reinforcement learning technique. Generalizing the MAP problem, Li et al. [110] summarize historical trajectories at the level of indoor POIs and model the POI transition probabilities based on an indoor connectivity graph to decode optimal sub-paths in-between. Observing that incomplete trajectories with similar routes often complement each other, Zheng et al. [281] use multiple trajectories to model movements between road network locations and to infer possible paths between consecutively observed locations in a trajectory to improve completeness. Jagadeesh and Srikanthan [87] use a Hidden Markov Model to produce suboptimal inference results in real-time by designing a route choice model to capture the likelihoods of only a small set of path candidates. Without using a topology explicitly, Li et al. [121] extract a network of road junctions and estimate transition probabilities across junctions based on structural regularities learned from massive trajectories. As a result, junctions are used as references to complete a fine-level trajectory. Inference-based approaches require large amounts of data for learning, and their accuracy decreases as an incomplete time interval grows.

Smoothing-based approaches utilize temporal autocorrelation of consecutive data items to mitigate volatility. Moving averages, exponential smoothing, and random walks are typical techniques for time series smoothing [31, 282]. Such approaches are simple to implement, but they do not address the randomness of movements in a specific trajectory.

An important branch of **STID UE** is the *spatiotemporal interpolation* techniques that estimate and insert thematic values at unsampled location-time points that align with spatiotemporally nearby sample points. In this branch, the time-interpolation-primitive[2] and space-interpolation-primitive approaches have been reviewed [116]. Here, we only review approaches that interpolate thematic values in space and time simultaneously. Such approaches can be based on shape functions [116], inverse distance weighting (IDW) [17, 195], and Kriging [113].

Motivated by *Tobler's first law of geography* [204], stating that things close to each other in space-time are more alike than more distant things, Li et al. [116] model a shape function with different time scales to interpolate PM2.5 measures. Appice et al. [17] extract prominent data trends and geographically-aware station interactions to approximate observed data in sensor networks, and they further infer missing data based on IDW. Susanto et al. [195] propose distribution-based distance weighting, where nearby data variations are considered to produce distributions (either Gaussian, Lorentzian, or Laplacian) for weight computation. Li et al. [113] use Kriging to predict PM2.5 distributions such that values at unsampled points can be determined by the values and weights of nearby sample points.

The performance of the interpolation techniques decreases with the expansion of the spatiotemporal range to be covered, and data (with ground truth) needs to be pre-analyzed for selecting an appropriate interpolation model.

Recently, data fusion methods have been considered for reducing measurement uncertainty in STID. Okafor et al. [169] employ feature selection to analyze factors that affect the accuracy of low-cost environmental monitoring sensors and introduce additional environmental features such as temperature and relative humidity for training measurement calibration models. One challenge in such data fusion-based UE approaches is how to find additional relevant and reliable data sources.

Remarks. Calibration-based and inference-based UE approaches both make use of spatial constraints and collective trajectories. The former identifies reference objects while the latter extracts regularities from incomplete trajectories having similar temporal and spatial conditions. Smoothing-based UE is based on temporal dependencies (i.e., varying smoothly and Markovian) of trajectories,

[2] Time series smoothing [31] on the thematic values of STID can be regarded as a time-interpolation-primitive approach.

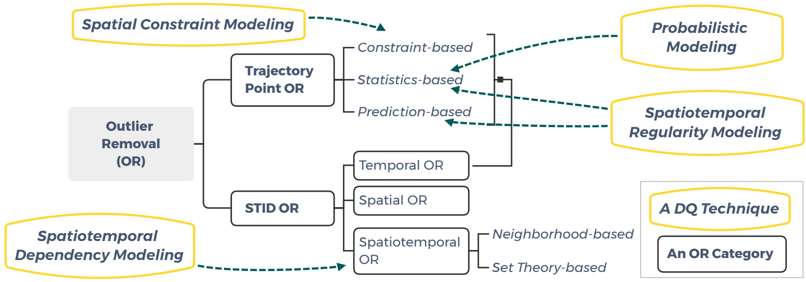

Fig. 5. Categories of outlier removal technologies and key DQ techniques.

which can be integrated easily with stream computing and fog/edge computing techniques to improve efficiency. Interpolation is based on spatiotemporal dependencies characterized as being varying smoothly, spatially autocorrelated, and spatially anisotropic (see Table 3).

3.3 Outlier Removal (OR)

We consider OR technologies for trajectories and STID separately, as indicated in Fig. 5.

Trajectory Point OR aims to remove each location point that is significantly different from its contextual points and does not accord with the expected normal mobility behavior underlying the trajectory. Note that removing point outliers is different from trajectory outlier detection [42, 132, 142, 154] that identifies anomalous trajectories. We consider three subcategories.

Constraint-based OR [239, 282] detects abnormal points that violate mobility constraints based on neighborhood information such as a maximum allowed velocity. Such approaches are simple to implement, but they do not contend well with dynamic and noisy trajectories.

Statistics-based OR identifies anomalous points based on statistical profiling of one trajectory [171] or a collection of trajectories [198]. Patil et al. [171] propose a Z-test-based anomaly detection method using a combination of privacy-insensitive information such as *synchronized Euclidean distance* (SED) [159], velocity, and acceleration. Tang et al. [198] apply Adaptive Density Optimization to a set of vehicle trajectories, in order to find low-density points that are likely to deviate from the roads as revealed by dense location points. Due to the reliance on statistics over historical data, these approaches do not work in scenarios with constraints on the available historical data.

Prediction-based OR [255, 256] identifies a value as an outlier if it differs from the value predicted from historical data. Outliers are then repaired with the predicted values. Zhang et al. [255] study likelihood-based repair over sequential data (e.g., trajectories), in which speed changes are modeled as distributions and a repaired sequence is found based on the maximum likelihood of the distributions. Assuming that some true values are available, Zhang et al. [256] integrate iterative minimum repair with an ARX model (AutoRegressive model with eXogenous inputs). In particular, high confidence repairs generated by ARX in previous iterations guide repairs in subsequent iterations. The key objective of these approaches is to achieve accurate predictions. To achieve that, they rely on trustworthy input data and regularly updated models.

STID OR considers three types of STID outliers, namely *spatial outliers* (outliers w.r.t. their spatial neighbors), *temporal outliers* (outliers w.r.t. their temporal neighbors), and *spatiotemporal outliers* (an item whose thematic attribute value deviates significantly from those of other items in its spatial and temporal neighborhoods). Trajectory point outliers can be regarded as a special case of temporal outliers. Therefore, the three categories of trajectory point OR covered above also apply to temporal outliers. Systematic reviews of temporal OR are available [30, 75].

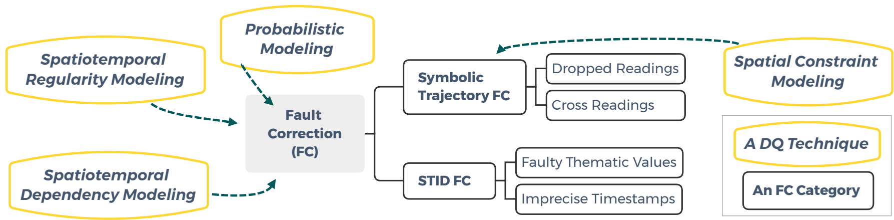

Fig. 6. Categories of fault correction technologies and key DQ techniques.

Aggarwal [10] reviews spatial and then spatiotemporal OR using spatial OR as a fundamental step. Aggarwal [10] also covers the close relationship between temporal OR and spatial OR when the temporal and spatial attributes are contextual attributes (as opposed to thematic attributes) in STID. In this sense, statistics-based and prediction-based approaches used widely in temporal OR also apply to spatial OR. Detecting pure spatial outliers, Zheng et al. [277] utilize both spatial and non-spatial contextual attributes to identify meaningful neighbors. To deal with heterogeneity and different scales of contextual attributes, metric learning is applied to effectively measuring the scores of spatial outliers.

In a classic study of spatiotemporal OR based on neighborhoods, Birant and Kut [29] consider the density of neighborhoods to identify spatial and temporal outliers and then combine the result to provide spatiotemporal outliers. Neighborhood-based approaches can be implemented when data is only partially available. However, the less neighborhood information that is available, the lower the effectiveness. Also, the decoupling of spatial and temporal aspects yields suboptimal results. In a classic set-theoretical study, Albanese et al. [15] utilize the concept of *rough set* to define a spatiotemporal outlier in terms of lower and upper approximations. Compared to neighborhood-based approaches, set theory-based approaches require holistic data and are more suitable for simple data attributes.

Remarks. Probabilistic modeling [171, 239, 255], spatiotemporal dependencies [29, 277] and regularity [255, 256], and spatial constraints [282] have been used widely in OR techniques. Some works [29, 255] follow the unsupervised learning paradigm. Temporal OR including the constraint-based approaches [239, 282] and prediction-based approaches [255, 256] can be implemented in a stream computing fashion.

3.4 Fault Correction (FC)

As illustrated in Fig. 6, we next present FC technologies for symbolic trajectories and STID.

Symbolic Trajectory FC repairs false negatives (FNs) and false positives (FPs) in symbolic trajectories. Unlike trajectories captured as geometric point time series, *symbolic trajectories* are seen in RFID, Infrared, and Bluetooth tracking scenarios where each location of an object is represented as the ID of the sensor that detected that object at that time [146]. In symbolic trajectories, FNs (*dropped readings*) [19, 20, 43, 64, 88] occur when a sensor fails to detect an object, while FPs (*cross readings*) [19, 21, 43, 64] occur when an object is unexpectedly detected by multiple sensors simultaneously (considering that the detection ranges of sensors are disjoint).

In general, symbolic trajectory FC technologies use probabilistic modeling to identify and repair faults. Moreover, these technologies consider spatiotemporal regularities of interactions between sensors and objects [19–21, 43, 64, 88], spatiotemporal dependencies among records in a trajectory [19, 43, 64, 88], and spatial constraints due to the sensor deployment and space structure [19–21, 43, 64].

Jeffery et al. [88] fix dropped readings based on a declarative, adaptive smoothing filter named SMURF, which consists of binomial sampling for per-tag cleaning and π-estimators for multi-tag cleaning. Chen et al. [43] utilize duplicate readings, the prior data distributions and FN rates of readers, and the maximal capacity of zones to capture the likelihood $\mathrm{P}(z_{ij} \mid h_i)$, where $z_{ij} \in \{0, 1\}$ indicates whether reader j reports object o_i and h_i is the zone where object o_i is actually in. Fazzinga et al. [64] embed constraints of direct unreachability, travel time, and latency into the modeling of spatiotemporal dependencies, and they identify the trajectory with the highest conditional probability. Focusing on integrity constraints implied by a sensor deployment, Baba et al. design a distance-aware graph [21] and a probabilistic graph [20] to handle FPs and FNs, respectively. Baba et al. [19] further utilize a multivariate HMM to capture the data uncertainty and correlation between object locations and RFID readings from historical data.

STID FC repairs *faulty thematic values* [98, 178, 184] or *imprecise timestamps* [91, 138, 157, 162, 188]. Pumpichet et al. [178] employ a belief-based approach to identify a group of helpful neighboring sensors based on the consistency of their data streams, estimating replacement values for dirty readings based on the time and distance over the identified group. Kuemper et al. [98] correct faults in IoT data sources. In particular, real-time information-quality vectors are generated for data sources based on cross-validation of heterogeneous sensory information. When these vectors indicate a provisionally unreliable data source, such a source is replaced by an alternate virtual data source that is created based on spatiotemporal analysis and interpolation methods. Providing a centralized data validation method, Sartori et al. [184] measure the Pearson correlation coefficients between the most recent reading sequences of adjacent sensors and find repairs for missing and anomalous readings from a single sensor based on the readings from correlated sensors.

Imprecise timestamps lead to staleness/uncertainty [157, 188] or disorder [91, 138, 162]. To find the optimal result among different combinations of possible timestamp repairs, Song et al. [188] adopt heuristics and linear programming relaxation over the provenance chain of unchanged nodes and the nodes to be repaired. Milani et al. [157] propose a graphical model to capture spatial and temporal dependencies in past update patterns. They also propose a dynamic probabilistic relational model to output repairs for stale cells via Maximum a Posteriori estimation. Mutschler and Philippsen [162] present a distributed and adaptive K-slack[3] for disorder processing on high-speed event streams. Aiming for efficient sliding window aggregate queries over out-of-order streams, Ji et al. [91] extend K-slack by introducing a window-based metric for measuring the aggregation quality. To address the latency of K-slack in heterogeneous networks, Liu et al. [138] propose aggressive and conservative strategies to handle unexpected and prevalent disorders, respectively.

Remarks. Symbolic trajectory FC [19, 43, 88] requires historical data to build models for use when cleaning incoming data. K-slack for disorder resolution [91, 138, 162] can be implemented in a stream and/or distributed computing mode.

3.5 Data Integration (DI)

In Fig. 7, we divide DI technologies for SID into two categories, namely semantic DI and non-semantic DI. The former involves semantic and comprehensible data sources and concerns their integration with raw SID to enrich the interpretability of SID. Without semantic aspects, the latter compares and combines multi-angle spatiotemporal observations to eliminate inconsistencies and to enhance the reliability of the integrated data.

Semantic DI technologies concern trajectories [109, 110, 125, 126, 167, 225, 239] or STID [22, 23, 25, 149, 230].

[3] K-slack buffers the arriving data for K time units for reordering.

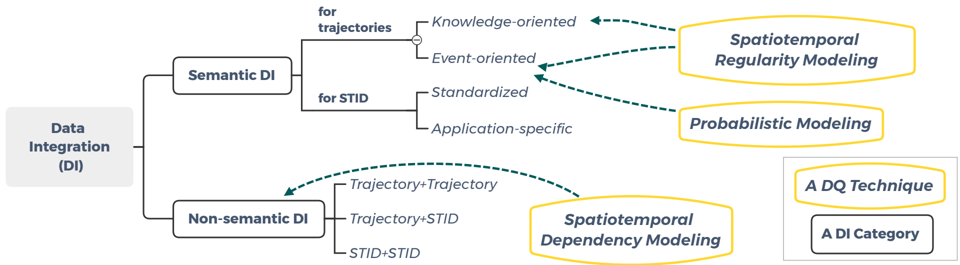

Fig. 7. Categories of data integration technologies and key DQ techniques.

Semantic DI for trajectories aims to annotate raw location traces with concepts or complementary knowledge at particular timestamps or during time intervals, facilitating direct, concise, and explainable exploitation of trajectories. According to the content to be associated with locations, they can be divided into knowledge-oriented technologies [167, 225] and event-oriented technologies [109, 110, 125, 126, 239]. The former annotate a trajectory point or segment with structured tuples [167] or human-readable text/keyword [225]. Nogueira et al. [167] propose an ontology-based framework to enrich GPS traces with Linked Open Data. Wu et al. [225] annotate location records with keywords extracted from geo-referenced social media data using Kernel Density Estimation. The querying of trajectories enhanced with keyword-like events, termed *activity trajectories*, has also been studied [190, 278]. Event-oriented technologies [109, 110, 125, 126, 224, 239] annotate trajectory points or segments with event labels to form sequences of application-specific events. Liao et al. [125] infer activity types and significant places from personal location traces using a hierarchical CRF (conditional random field). They further extend the hierarchical CRF to model the mapping from GPS data to transportation concepts such as destination and transportation mode [126]. Yan et al. [239] use a Hidden Markov Model to annotate trajectories with stops and POI categories on a grid-based map. By analyzing spatiotemporal regularity, Wu and Li [224] facilitate personalized POI category annotation of personal GPS records. Li et al. [110] annotate noisy Wi-Fi positioning data with sequences of semantic mobility triples of the form (time, indoor region, mobility pattern), using density-based partitioning for event detection and weighted estimates of relevant positioning records for region matching. Li et al. [109] further propose a coupled CRF to model indoor spatial constraints as well as probabilistic dependencies among positioning records, regions, and events. As a result, multivariate annotations are decoded with the highest plausibility.

Semantic DI for STID enriches spatial data infrastructures (SDI) with *standardized* [23, 230] or *application-specific* [22, 25, 149] geo-semantic meta information. Wu et al. [230] propose a Semantic-Web-of-Things framework that combines a Semantic Sensor Network (SSN) ontology with other domain-specific semantics extracted from IoT resources based on entity linking. Bajaj et al. [23] categorize existing ontologies required for annotating different aspects (4W1H: What, When, Who, Where, and How) of IoT data acquisition and access. Barnaghi et al. [25] design a lightweight semantic modeling framework to annotate spatial, temporal, and thematic attributes of sensor stream data, using geohashing and clustering to distribute streams to different repositories at different scales. To extract interpretable knowledge from continuous and heterogeneous IoT data streams, Maarala et al. [149] design a mobile reasoner that uses geographical partitioning and brings data processing closer to the data sources. Badidi and Maheswaran [22] design a DI architecture for IoT urban data by combining semantic technologies, edge computing, and cloud computing.

Existing studies assume that the semantics to be integrated is not updated and thus do not address real-world dynamically evolving semantics, which thus remains an open problem.

Non-semantic DI technologies can be divided into three cases: trajectory+trajectory [93, 173, 260], trajectory+STID [261], and STID+STID [51, 283].

Current ubiquitous location systems [14, 186] are constructed with different infrastructures and algorithms, producing trajectories in diverse formats [173], resolutions [260], or ID systems [93]. *Trajectory+trajectory* aims to generate a unified representation for such different trajectories. Peixoto et al. [173] propose the Trajectory Data Description Format (TDDF) to enable the conversion between formats. TDDF can capture statistics to enable efficient data management. To model real-time traffic, Zhang et al. [260] propose a convex multi-view learning method to quantify biases of trajectories and a context-aware tensor decomposition method to calibrate incomplete trajectories at different spatial granularities. To identify the same moving entity that has different IDs in different trajectory datasets, Jin et al. [93] extract trajectory signatures based on four representation strategies (sequential, temporal, spatial, and spatiotemporal) and two quantitative criteria (commonality and unicity) and conduct kNN search over these signatures.

Trajectory+STID attaches spatial or spatiotemporal measurements to points or segments of location traces based on similarities of their spatial or temporal attributes. Zhang et al. [261] propose a DI architecture to analyze real-time mobility patterns based on correlations and divergences in multi-source urban IoT data.

STID+STID fuses multi-source spatiotemporal measurements based on their spatial and temporal commonality. Cheng et al. [51] develop a spatial and temporal nonlocal filter-based fusion model to enhance both the spatial resolution and temporal frequency of remote sensing data. Focusing on how different approaches utilize spatial and temporal dependencies of data, Zhu et al. [283] provide a systematic review of spatiotemporal fusion of multi-source remote sensing data.

In addition to these data pre-processing technologies that integrate multi-source SID to serve business needs, a popular line of research constructs end-to-end models that learn and fuse multi-source data to serve business needs directly. The relevant techniques, such as multi-task learning [164] and multi-view learning [89, 262, 269, 272], are detailed in Section 4.3.

Remarks. Semantic DI for trajectories often exploits spatiotemporal data regularity incurred by geo-semantics (e.g., POI category [125, 239], indoor or road network constraints [109, 110, 126], and personal preferences [225]). To efficiently assign semantics to data at the IoT far end, edge computing [22, 149] and stream computing [25] have been used in semantic DI for STID. Non-semantic DI [93, 261, 283] utilizes mainly the spatiotemporal dependencies in data.

3.6 Data Reduction (DR)

DR aims to improve throughput and computing efficiency in general while minimizing the loss of information as seen from the business level. A categorization in the SID context is shown in Fig. 8. We proceed to cover technologies for trajectory compression and STID reduction in turn.

Trajectory Compression compacts either raw trajectories [33, 103, 128, 136, 137, 144, 159, 160, 275] or network-constrained and map-matched trajectories [41, 78, 97, 118, 119, 177, 243]. Each category can be further divided into online and offline approaches. The related concept of *trajectory simplification* [33, 103, 128, 136, 137, 144, 159, 160] can be regarded as a special form of compression. However, it focuses on eliminating trajectory points and does not consider compression means such as binary encoding. A mainstream technology for trajectory simplification is the error-bounded line simplification algorithms [129].

Raw Trajectory Compression. In the offline setting, all trajectory points are accessible during compression. Cao et al. [33] study trajectory compression based on line simplification. Considering

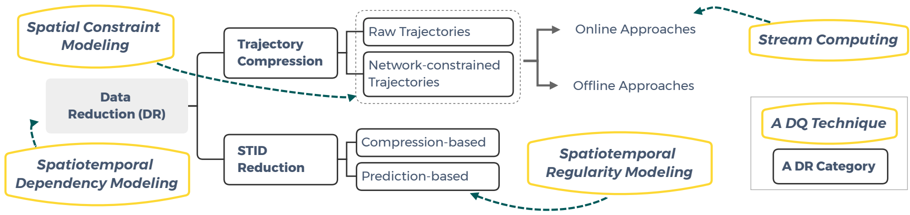

Fig. 8. Categories of data reduction technologies and key DQ techniques.

different approximation distances for line simplification, this study considers the *soundness* concept of whether the answer to a query over approximated trajectories is error-bounded. Moreover, this study considers the *aging of trajectories* in compression, the idea being to allow increasingly coarse approximation as time elapses. Intending to minimize direction-aware distances with a fixed storage budget, Long et al. [144] study direction-preserving trajectory simplification (DPTS) using dynamic programming and binary search. They also design approximate solutions to a dual form of DPTS, i.e., maximizing the span of the minimum covering angular ranges of all line segments. Based on references extracted from collective trajectories, Zhao et al. [275] use greedy algorithms and dynamic programming algorithms to achieve optimal compression among massive combinations of references for resembling a trajectory.

Lange et al. [103] study optimal line simplification for reducing trajectory data in the context of online tracking of trajectories of moving objects in sensor networks. Subsequent online trajectory compression has been formulated as a *Min-Error* problem, where the aim is to minimize the compression error while achieving a compression ratio that satisfies a given threshold; or conversely, as the problem of maximizing the compression ratio while satisfying a given compression error threshold [136, 137]. Assuming a fixed storage budget, Muckell et al. [159] propose the SQUISH method that processes incoming points one by one to achieve a final compression ratio λ that minimizes the Synchronized Euclidean Distance (SED), defined as the sum of Euclidean distances between the same-time positions on two trajectories (the distances between concurrent trajectory positions have been investigated earlier on [33]). The extended SQUISH method [160] allows a user-specified threshold μ for the SED error. Liu et al. [136] propose a Bounded Quadrant System (BQS) that bounds each incoming point by a convex hull in a virtual coordinate system to enable efficient compression error evaluation. They further offer normal, fast, and progressive versions of the BQS algorithm [137] to adjust the storage budget and compress trajectories with different error tolerances subject to *trajectory aging* [33]. Based on the SED error, Lin et al. [128] develop a spatiotemporal cone intersection-based algorithm to check trajectory points in $O(1)$ time. Their simplification allows interpolated data points in its outputs. Recently, Wang et al. [218] adopt reinforcement learning to build online point dropping strategies for different error measures; this also works in offline mode.

Network-constrained Trajectory Compression. In the offline setting, road network constraints are considered globally. Popa et al. [177] discuss the limitations of 2D compression methods for compressing in-network trajectories in road network settings. They propose an extended data model and a network partitioning algorithm to support error-bounded in-network trajectory compression based on line simplification. Han et al. [78] propose a framework that decomposes trajectories into spatial paths and temporal sequences and performs in parallel lossless spatial path compression and lossy, but error-bounded, temporal sequence compression. Yang et al. [243] study the `TED` representation, where a trajectory is represented by a spatial entry path (`E`), distances (`D`) that locations appear in the E, and a time flag sequence (T) to indicate a trajectory's presence at an E edge at a certain time. Koide et al. [97] summarize trajectories as sequences of road edges based on the FM-index (a compressed full-text substring index). Focusing on uncertain trajectories, Li et al. [119] improve TED by considering variations in sample intervals and also generate corresponding referential representations (and binary representations). Other than reducing trajectories on road networks, a *map generalization* process simplifies the geographical data within a map of a certain scale without degrading the readability of information [223].

In an online fashion, Chen et al. [41] calculate the heading of incoming GPS points and compact the data based on heading changes at intersections. Li et al. [118] propose a real-time compression framework, in which referential trajectory representations are built by the selection, deletion, and rewriting operators on edge servers and sent to cloud servers for querying based on a cost-reducing data transmission scheme.

To sum up, dynamically adjusting compression strategies based on data dynamics and reducing data volumes as early as possible on edge devices are directions for trajectory compression to be further strengthened in IoT scenarios.

STID Reduction can be divided into compression-based [9, 57, 106, 201, 207] and prediction-based [34, 197, 248, 267] approaches.

Compression-based approaches can be divided further into lossless and lossy ones. Lossless compression [9, 201] usually works in batch mode and is suitable for applications that demand accuracy. Abuadbba et al. [9] use Gaussian approximation to reduce smart meter readings such that only the margin space between the approximated and actual readings is losslessly compressed. Tate [201] uses Golomb-Rice codes to compress phasor angle data by considering the correlations between the phasor angles of different sensory units. In contrast, lossy compression [57, 106, 207] achieves a higher compression ratio with some precision loss. To deal with multimodal measurement data collected from Wireless Sensor Networks, Li et al. [106] extend the lossy stream compression method *Lightweight Temporal Compression* from 1D to ND by detecting N-ball intersections. Considering data reconstruction based on a reduced volume of transmitted data, de Souza et al. [57] apply Singular Value Decomposition to lossy data compression in smart distribution systems. Tripathi et al. [207] devise an adaptive data reduction algorithm based on compressive sampling and Gaussian Mixture Model-based quality assessment to reduce smart meter data transmission.

Prediction-based approaches [34, 197, 248, 267] are mostly used to reduce the data volume of communication between IoT nodes. Data can be dropped if the error of a predicted value is within an acceptable range. Carvalho et al. [34] deploy a linear regression model at each node and check the prediction consistency between spatially neighboring nodes. Data is transmitted only if inconsistent predictions exist. Instead of using linear regression for multivariate data, Yin et al. [248] use a Kalman Filter to predict future values for univariate readings. Spatial correlation is also utilized to redistribute energy consumption within a cluster of neighbors. Tan and Wu [197] predict reading values both at the source and sink based on a hierarchical Least Mean Square adaptive filter. Sensor nodes are requested only to send readings that deviate from the prediction by an error budget. Zhang et al. [267] combine CNN and LSTM models at edge devices for event prediction, and only the data with events as predicted true is transmitted.

Compression-based approaches fit well in batch processing scenarios, while prediction-based approaches are challenged by the robustness and timeliness of prediction models.

Remarks. DR technologies for trajectories and STID mostly utilize spatiotemporal data dependencies. Online trajectory compression [41, 128, 136, 137, 160] fits well with stream computing. Some prediction-based DR for STID [34, 267] builds machine learning models based on spatiotemporal

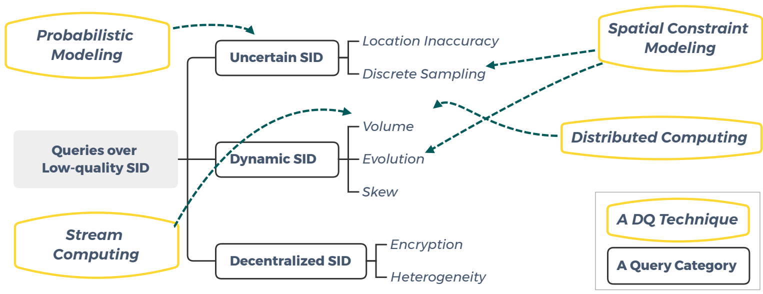

Fig. 9. Categories of queries over low-quality SID and key DQ techniques.

regularity. Edge computing [197, 267] and distributed computing [34, 57, 248] techniques have been explored to reduce data volumes at the IoT edge devices.

4 EXPLOITATION OF LOW-QUALITY SID

This section covers techniques that exploit existing SID of low quality to fulfill various business purposes, including queries (Section 4.1), analyses (Section 4.2), and decision-making (Section 4.3).

4.1 Queries over Low-quality SID

A categorization of queries over low-quality SID is shown in Fig. 9. As three major obstacles to effective and efficient SID query processing, the uncertainty, dynamics, and decentralization of data are discussed in Sections 4.1.1, 4.1.2, and 4.1.3, respectively.

4.1.1 Queries over Uncertain SID. Location uncertainty is a major issue in spatial queries [238], for which probabilistic modeling techniques are exploited widely. In this setting, query processing techniques 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 [53], and a recent survey [284] categorizes the existing queries over uncertain spatial data according to query types. In contrast, we categorize query processing techniques based on the type of location uncertainty they handle in the context of IoT-based localization/tracking, namely the uncertainty caused by inaccuracy of localization algorithms and that caused by the discrete sampling of devices [176].

To handle the **uncertainty caused by location inaccuracy**, an object's location l_i at a single time point t_i is usually described as a probability density function (pdf) $f(l_i, t_i)$, which occurs in continuous and discrete cases:

- *Continuous Case.* A closed-form distribution, satisfying $\int_{l_i \in ur} f(l_i, t_i)\, \mathrm{d}l_i = 1$[4] and $\forall l_i' \notin ur\ (f(l_i', t_i) = 0)$, where ur is a closed uncertainty region that minimally covers all possible object locations;
- *Discrete Case.* A set of instances (samples) s_j with corresponding occurrence probabilities p_j, formally $f(l_i, t_i) = \{(s_1, \mathrm{p}_1), \ldots, (s_N, \mathrm{p}_N)\}$ having $\sum_{j=1}^{N} \mathrm{p}_j = 1$.

Table 5 further differentiates existing studies according to their query types.

[4]A general case is $\int_{l_i \in ur} f(l_i, t_i)\, \mathrm{d}l_i \leq 1$ where $\int_{l_i \in ur} f(l_i, t_i)\, \mathrm{d}l_i < 1$ implements existential uncertainty [56, 213], i.e., an object's overall existence is indicated by a probability value.
[5]Reference [238] finds a subset of O that is covered by O's convex-hull with probability above a given threshold.
[6]It is the ranking version of a probabilistic spatial query that returns m objects with the highest probabilities.

Table 5. Selected Queries over Uncertainty Caused by Positioning Inaccuracy

Query Types	Continuous Case	Discrete Case
NN (Nearest Neighbor) and kNN Queries	[28, 52, 54, 206]	[232]
Range Queries	[200, 220]	[232, 238][5]
Ranking Queries	[56][6]	[84, 254]
Reverse NN Queries	[124]	[27, 39]
Skyline Queries	[211]	[172, 266]
Range Aggregate Queries	[139, 270]	[270]
Contact Similarity Queries and Joins	[26, 213]	[233]

To handle the **uncertainty caused by discrete sampling**, a moving object o's location(s) at unsampled time points is modeled by a distribution that is referenced to o's sampled, known location(s). The distribution can be modeled to infer the location at a single time point or the locations across a time interval. The uncertainty models can also be applied during pre-processing to perform spatial interpolation of original data.

Given an observed location l_c at time t_c and a maximum object speed v_{max}, the possible object locations at a future timestamp $t_f > t_c$ belong to a circular uncertainty region $\bigcirc(l_c, v_{max} \cdot (t_f - t_c))$ centered at l_c with radius $v_{max} \cdot (t_f - t_c)$, following a uniform distribution [146, 240, 241] or a Gaussian distribution [176], sometimes with a distance-decaying effect [112]. Based on the circular region modeling, the possible locations at a time t_x between t_a and t_b can be further reduced to the intersection of the two circular regions $\bigcirc(l_a, v_{max} \cdot (t_x - t_c))$ and $\bigcirc(l_b, v_{max} \cdot (t_b - t_a))$, called a *lens* [176]. Additional space constraints such as indoor topology [112, 146, 240, 241] can be utilized to further reduce the circular region or lens. Different from modeling circular uncertainty region, the future location can be given by linear dead reckoning based on velocity and direction [86, 103]. Thus, given a velocity vector $\vec{v}$, the location at time t_f ($t_f > t_c$) is obtained as $l_f = l_c + \vec{v} \cdot (t_f - t_c)$.

Sometimes, queries require knowing possible locations across a small time interval or the entire duration of a trajectory. To this end, observed locations at multiple timestamps $\langle (l_1, t_1), \ldots, (l_N, t_N) \rangle$ are utilized. The expected location l' at any time $t' \in [t_a, t_b]$ between two consecutively reported locations can be obtained using linear interpolation, and the corresponding uncertainty region is a circular region centered at l' and with a predefined radius [206]. The uncertainty regions across an unsampled time interval combine to form a buffered line segment, or a 3D cylinder in location-time space [90]. In a different approach, the location l' is constrained by an ellipse whose two foci are the reported locations l_a and l_b and whose eccentricity is determined by the maximum speed [176]. In the location-time space, the shape of the ellipse becomes a *bead* (also known as *space-time prism* [77, 100]) as an integrated body of an upward and a downward pointing cone [90, 176], and the bead sequence for a discrete trajectory forms a "necklace" [81, 99, 145, 205, 268]. Speed-constrained beads can be further refined by fusing spatial constraints derived from additional sensory data (e.g., road-side sensor data [257]). Beyond the speed constraint, Markovian dependencies along a trajectory are exploited such that unobserved locations can be instantiated by a stochastic process over observed locations. In the setting of a discrete grid that partitions the space, an object's current grid location can be inferred based on its first-order Markovian record (its last reported location) [62, 166, 265]. Assuming a Gaussian distribution of the current uncertain location, its mean and standard deviation can be inferred from previous uncertain locations that also follow a Gaussian distribution [90]. Considering a complex space structure, uncertain locations are modeled based on a particle filtering process such that particles are resampled to replicate high weight particles and eliminate low weight particles [251].

Table 6. Selected Queries over Uncertainty Caused by Discrete Sampling

Query Type	At a Time Point	Across a Time Interval or the Duration of a Trajectory
NN and kNN Queries	uniform circular [241]; velocity vector [86]	cylinder [206]; particles [251]; first-order Markovian grids [166, 265]
Range Queries	uniform circular [240]	particles [251]; first-order Markovian grids [62, 265]; Markovian Gaussian distributions [90]; combinations of road segments [280]; speed-constrained beads/necklaces [205]; beads with mobility constraints [257]
Similarity Ranked Queries		combination of sample connections [148]
Reverse NN Queries		first-order Markovian grids [61]
Range Aggregate Queries	distance-decaying [112]	combination of sample connections [111]; speed-constrained bead/necklace [145]
Contact Similarity and Alibi Queries	uniform circular [146]	speed-constrained beads/necklaces [99, 268]

From a holistic view, the Cartesian product is used to form all possible trajectories based on observed discrete locations. The probability of each formed trajectory instance is computed as the product of the probabilities of all involved observed locations. In a setting where each observed location is described as a set of location samples, the possible trajectories are generated by connecting two samples at each pair of consecutive timestamps [111, 148]. In a road network, each possible route comes from the combination of possible road segments between each two consecutively observed route locations [280].

Table 6 summarizes different queries and their uncertainty models in the setting of discrete sampling.

Queries over uncertain spatial data have been studied extensively in the last decades, while how to query uncertain SID in a resource-limited and stream setting remains open [284].

4.1.2 Queries over Dynamic SID. The dynamics of SID bring about issues of data volume, data evolution, and data skew in spatial query processing.

To efficiently process **Queries over Massive SID**, distributed computing [50, 153, 231, 250, 252] and stream computing [50, 91, 153] techniques have been proposed.

You et al. [250] implement two systems, namely SpatialSpark based on Apache Spark and ISP-MC based on Apache Impala, to support indexed spatial joins based on point-in-polygon testing and point-to-polyline distance computation. Xie et al. [231] develop an in-memory distributed framework that leverages segment-based partitioning and two-layer indexing of trajectories to enable large-scale similarity search. Mapping and partitioning noisy trajectories based on road networks, Yuan and Li [252] support in-memory distributed similarity search and join by quickly pruning irrelevant partitions and dissimilar trajectories. As an extension of the distributed stream processing platform Apache Storm, Mahmood et al. [153] implement a spatio-textual query processing system with a spatio-textual index that can adapt to the data distribution and query workload. To enable continuous spatio-textual queries over flooding geo-tagged text streams, Chen et al. [50] propose a distributed publish/subscribe system with a workload distribution algorithm that adapts to both space and text properties of the data. These methods focus on the scalability of the query processing, without considering reducing data for high-speed yet low-cost computation.

For **Queries over Evolving SID**, object locations and other information arrive continuously in a streaming fashion. Safe region [13, 38, 74, 107, 168, 180, 237] and incremental evaluation [114, 242, 259] strategies have been applied to reducing the communication and computation overhead.

Qi et al. [180] provide a systematic review of safe region-based techniques for continuous kNN [107, 168] and range [13, 38] queries. We cover representative studies of other types of safe region-based continuous spatial queries as follows. First, to enable efficient processing of subscriptions to incoming events in the proximity of moving users, Guo et al. [74] propose a communication cost model and incremental schemes to construct safe regions for spatial Boolean expression matching over event streams. Second, to continuously find pairs of users whose dynamically changing distance is below a threshold, Xu et al. [237] build safe regions based on predicted locations using non-linear motion patterns. Third, Hidayat et al. [79] devise efficient safe region construction algorithms for both skyline and top-k queries with continuous query location updates.

To enable continuous optimal shortest path queries with dynamic traffic, Yang et al. [242] propose means of quickly finding affected queries and updating their shortest path answers when road conditions change. To continuously provide k alternative paths as a user moves on a path towards the target, Li et al. [114] devise depth-aware algorithms that maintain and exploit previously computed useful information to efficiently update the query result. Assuming a moving object's partial trajectory is being updated, Zhang et al. [259] study continuous trajectory similarity search based on pruning and incremental evaluation. These algorithms are centralized and have not considered the locality of SID in decentralized settings.

Skewed SID generated by mobile users is seen commonly in IoT and cloud computing environments. For **Queries over Skewed SID**, node load-balancing [183] and data partitioning [183, 210, 221] have been adopted.

Ray et al. [183] propose a heterogeneous cluster-based spatial query processing infrastructure that uses declustering to create balanced spatial partitions and dynamic load-balancing to resolve performance heterogeneity and data skew during processing. To support multi-dimensional range and NN queries over skewed data, Wei et al. [221] propose a dynamic and scalable index KR^{+}-index on Cassandra that enhances R-tree with keys constructed as the Hilbert-value of the centroid coordinate of the leaf rectangle. Vo et al. [210] propose a spatial data partitioning framework `SATO` that consists of `S`ampling, `A`nalysis for partitioning strategy, `T`earing for data distribution, and `O`ptimization based on succinct partition statistics. So far, query processing algorithms and data partitioning/indexing strategies have not been considered for decentralized edge devices.

4.1.3 Queries over Decentralized SID. In a distributed architecture, data encryption [73, 94, 249] and heterogeneity [59, 193, 234] pose challenges to query processing.

To enable the outsourcing of range and kNN querying on private spatial data, Yiu et al. [249] propose a spatial transformation scheme that balances efficiency and privacy as well as a cryptographic transformation scheme. Kamel et al. [94] consider updates from data owners to encrypted outsourced data and contribute a dynamic spatial index to support encrypted range query processing in the cloud. Aiming at uncertain data encrypted in decentralized semi-trusted servers, Guo et al. [73] design an authorized ranking method to process kNN queries over ciphertexts.

To enable spatial queries over heterogeneous location data sources, Xu and Güting [234] propose a generic and precise location representation for moving objects referencing a set of defined infrastructures. To query similar asynchronous trajectories generated by multiple sources, Sun et al. [193] select optimally matched points based on spatial and temporal thresholding and use the selected points to measure multi-source trajectory similarity. To integrate heterogeneous data models and workflows (e.g., indexing and query processing) for big and diverse trajectory data, Ding et al. [59] propose a unified data management and analytics platform that provides unified storage and computing engine and an enhanced distributed computing paradigm with flexible APIs. These works collect heterogeneous data and process them in a centralized manner and do not address the querying of heterogeneous data at decentralized nodes with different capabilities.

4.2 Analyses on Low-quality SID

We categorize existing analysis techniques targeting low-quality SID based mainly on quality issues related to uncertainty and dynamics (volume and evolution). Within each category, works are organized according to their analysis tasks. The relevant but special form of *visual analytics* tasks have been covered by Section A.2 in the Supplementary Material.

4.2.1 Analyses of Uncertain SID. To address uncertainty in SID, data analysis techniques often exploit probabilistic modeling [120, 123, 203, 219, 276], spatiotemporal dependencies [132, 140, 214, 244, 276], and spatial constraints [44, 174, 203, 222].

Clustering. Assuming trajectories are captured as sequences of uncertainty regions, Pelekis et al. [174] propose an intuitionistic fuzzy vector representation to compress uncertainty and generate *centroid trajectories* to capture similar movements, upon which they conduct Fuzzy C-Means clustering over the generated centroid trajectories. Considering network-constrained trajectories with positioning errors and low sampling rates, Chen et al. [44] construct an approximate minimum spanning tree of a trajectory to define similarity on candidate segments. In this setting, a graph-based clustering algorithm is proposed that uses representative points to update clusters incrementally.

Anomaly Detection. Li et al. [132] propose an N-gram-based abnormality measurement method to identify missing events in medical devices. They construct hotspots of abnormal events and model transitions between hotspots using finite state machines. To reduce the influence of uncertain tracing data on the abnormality measurement, they devise an iterative algorithm for the recovery of missing records and estimation of transition probabilities.

Frequent Pattern Mining. Considering a high degree of incompleteness and noise in spatiotemporal sequences, Li and Han [123] study techniques for period detection and periodic behavior detection. Using sequence-level and element-level data uncertainty models, Zhao et al. [276] find probabilistically frequent sequential patterns based on a prefix-projection version of the PrefixSpan algorithm. To retrieve sequential stop-by pattern (sequential occurrence regions) from uncertain RFID data, Teng et al. [203] propose a probabilistic model to capture deployment and spatial constraints, find uncertain candidates based on filtering and mapping construction, and output the stop-by patterns by means of an Index 1-itemset algorithm and an event clustering algorithm. To extract sequential stay events from noisy trajectories, Yang et al. [244] design a density function that considers neighborhood movement ability and stay time as well as a trajectory clustering algorithm with dynamic noise tolerance. Assuming multi-instance location uncertainty, Li et al. [120] study probabilistic threshold mining of frequent spatiotemporal sequential patterns based on a dynamic programming method for computing the frequency probability of patterns. Assuming uncertainty is captured as a probability distribution, Wang et al. [219] propose fast co-occurrence pattern mining algorithms based on filter-and-refinement.

Hotspot and Popular Route Discovery. Liu et al. [140] study community detection based on diffusion modeling on noisy trajectories and additional fine-grained markers (e.g., movement velocity and the semantics of locations). To detect high-density crowds from noisy Wi-Fi positioning sequences, Wang et al. [214] simplify and reconstruct sequences based on stay points and Kalman filtering, and propose a spatiotemporal version of the OPTICS algorithm. Wei et al. [222] infer the top-k routes that sequentially pass the given locations within a specified time interval, by aggregating temporally sparse trajectories over a graph constructed for routing.

The above-mentioned proposals are batch-oriented and centralized, and they do not consider real-time and decentralized settings.

4.2.2 Analyses of Dynamic SID. To handle high data volumes in analytics, indexing and pruning [32, 216, 258], distributed computing [83, 194, 228], and stream computing [42, 65, 138] techniques have been proposed. Spatiotemporal dependency modeling and online learning [142, 170, 217, 227] have been utilized to facilitate the analysis of evolving SID.

Clustering. Assuming a decentralized, noisy RFID system, Wu et al. [228] define a Time-Parameterized Edit Distance to form RFID trajectory clusters in a MapReduce framework. Each cluster is a sequence of node-range pairs that describe the co-movement of a group of objects. Given massive trajectories, Hu et al. [83] use coarse-grained Dynamic Time Warping to enable fast similarity computation and further propose a MapReduce-based strategy to slice and cluster trajectories. To enable scalable clustering over map-matched trajectories, Wang et al. [216] propose an edge-based distance (EBD) measure to reduce time complexity, an algorithm extended from Lloyd's algorithm[7] for finding k representative paths, and an indexing framework with a pivot-table and an inverted index to avoid unnecessary distance computations. Wang et al. [217] construct a kNN network to capture changing locations of vehicles, learn low-dimensional vehicle representations by performing dynamic network representation learning on the constructed network, and use K-medoids and Gaussian Mixture Models to cluster vehicles with similar behavior patterns.

Anomaly Detection. Given continuous trajectory streams with changing distributions, Bu et al. [32] monitor anomalous patterns characterized by big spatial deviations within certain time intervals by means of online local cluster construction, pruning strategies, and piecewise metric indexing. By comparing against historically "normal" routes on the fly, Chen et al. [42] identify anomalous sub-trajectories as well as the corresponding parts that indicate the anomalies. To detect anomalies in partial trajectories that have not reached a destination, Wu et al. [227] capture driving behavior and preferences based on a maximum entropy inverse reinforcement learning model. To enable online updates of anomaly scores of trajectories, Liu et al. [142] propose a Gaussian Mixture Variational Sequence AutoEncoder to capture complex sequential information of trajectories and to discover different types of normal routes in a latent space. Mao et al. [154] propose a feature grouping-based algorithm to detect abnormal trajectory fragments on the fly from evolving trajectory streams with skewed distributions.

Frequent Pattern Mining. Sun et al. [194] construct a Probabilistic Suffix Tree to mine significant subsequence patterns from massive uncertain spatiotemporal data using Hadoop. To find spatial co-evolving patterns (groups of sensors that are spatially correlated and co-evolve frequently in their readings) from massive geo-sensory data, Zhang et al. [258] propose a two-stage approach. First, frequent evolutions for individual sensors are detected via a segment-and-group approach. Second, the evolutions are assembled while using spatial pruning enabled by a pattern search tree. Liu et al. [138] propose aggressive and conservative strategies to process sequence pattern mining on out-of-order RFID event streams.

Event Discovery. To discover spatial events from conflicting mobile crowdsourced data, Ouyang et al. [170] propose TSE (Truth finder for Spatial Events) and Personalized TSE models to handle diverse and noisy participant reports in an unsupervised way. Assuming streaming spatial objects, Feng et al. [65] propose a sliding window model to continuously detect *bursty regions* with many spatial objects in a specified spatial and temporal range.

How to migrate the functionality covered above to edge devices to reduce cost and latency are highly relevant future research topics.

4.3 Decision-Making using Low-quality SID

A variety of decision-making tasks based on SID are relevant, such as the prediction of next location(s) [48, 102, 104, 105, 181, 262], traffic volume [141, 199], and spatiotemporal variables [46, 67, 147, 164, 245, 253]; the recommendation of POIs [82, 155, 264, 273] or routes [72]; and the

[7]Lloyd's algorithm [143] is originally for Voronoi-based iterative centroid estimation.

planning of task assignments [192] or site selection [40, 272]. We organize studies according to the DQ issues they address in learning, namely the scarcity of labels, limited data availability and data bias, uncertainty of data, dynamics of data, and heterogeneity and decentralization of data.

Scarcity of Labels. This issue has been addressed in unsupervised learning (such as EM [82], AutoEncoder [48, 236], and GAN (Generative Adversarial Network) [48]), semi-supervised learning (such as co-training [46] and PU learning [40]), and multi-task learning [67, 253, 273]. Considering multiple latent temporal parameters in POI recommendation, Hosseini et al. [82] retrieve multi-aspect temporal similarity maps to reduce user-location matrix sparseness and use the EM algorithm to compensate for incomplete data at each temporal scale. To assess the quality of unlabeled volunteered geographic information (VGI), Xu et al. [236] match VGI and official data to obtain samples to train an AutoEncoder by minimizing reconstructed errors. Chen et al. [48] adopt GANs or Variational AutoEncoders to generate qualified trajectories for self-driving simulation and traffic analyses. To estimate urban air quality at a fine spatial granularity, Chen et al. [46] adopt an ensemble semi-supervised learning method with iterative co-training to counter the limited availability of labeled data. To select new public toilet locations with limited positive labels of regions (i.e., having toilets placed there), Chen et al. [40] identify reachable regions, construct their high-order and semantic representations from multi-source urban data, and adopt PU learning over the representations to identify unlabeled positive regions that should have a toilet. Yuan et al. [253] devise a multi-level multi-task learning framework for predicting lake water quality at multi-scales, in which information among region-specific models are shared to help create models for regions with limited or no training data. Assuming incomplete labels when forecasting the scales of spatial events, Gao et al. [67] propose a multi-task ordinal regression framework that enforces similar feature sparsity patterns for different tasks while preserving the heterogeneity in their scale patterns. Using a Spatio-Temporal Gated Network, Zhao et al. [273] jointly train the POI context prediction and next POI recommendation to fully leverage labeled and unlabeled data.

Limited Availability and Bias of Data. This issue has been addressed in transfer learning [72, 245] and federated learning [105, 155]. To transfer long-period data from other cities for spatiotemporal prediction, Yao et al. [245] train a well-generalized spatial-temporal network based on a meta-learning paradigm. Aiming to learn routing preferences between a pair of identified regions, Guo et al. [72] resolve sparse and skewed trajectories between a region pair by transferring routing preferences from the pairs with dense trajectories. Assuming mobility data is protected locally, Li et al. [105] propose a federated learning framework for location prediction that utilizes self-attention and local-global fusion to achieve personalization. Aiming at privacy-preserving and sparsity-aware location recommendation, Meng et al. [155] propose randomized data obfuscation and region aggregation methods to deal with data sparseness and propose tensor factorization-based spatial similarity to execute predictions at spatial neighbors.

Uncertainty of Data. Probabilistic modeling [181, 264] is used to handle location uncertainty, while reinforcement learning [199] is used to deal with incompleteness. By removing outlier trajectories via clustering, Qiao et al. [181] construct a continuous time Bayesian network to capture correlations among street ID, speed, and direction for predicting the motion of an uncertain moving object. To recommend next individual POIs with uncertain check-ins at collective POIs, Zhang et al. [264] exploit hierarchical category transitions to model users' preference transitions and semantic relatedness of POIs at different granularities. Using incomplete trajectories for traffic volume inference, Tang et al. [199] use deep reinforcement learning to recover vehicle movements and use graph embedding to encode multi-hop traffic propagation between road segments.

Dynamics of Data. Reinforcement learning [104, 192], incremental learning [102], and edge computing [147] techniques have been exploited. To predict a remaining trajectory from an observed partial trajectory, Le et al. [104] use reinforcement learning to model sequential decision-making and employ long-term optimal planning for predictions. Considering emerging crowdsourcing workers in spatial task assignment, Sun et al. [192] propose GRU (Gated Recurrent Unit)-based predictors for tasks and workers and propose adaptive batching strategies based on the Deep Q Network. To predict destinations in data streams, Laha and Putatunda [102] apply a sliding window with the exponentially fading to four incremental learning methods (i.e., multivariate multiple regression, spherical-spherical regression, randomized spherical kNN regression, and their ensemble). For short-term energy prediction over dynamic STID, Luo et al. [147] propose an online edge computing framework that performs acquisition, processing, and deep regression in sensing nodes, routing nodes, and the central server, respectively.

Heterogeneity and Decentralization of Data. Multi-task [164] and multi-view learning [89, 262, 269, 272] techniques have been adopted to integrate multi-source data, while federated learning [141] is used to facilitate decentralized models. Nguyen et al. [164] present a Spatial-temporal Multi-Task Learning algorithm to integrate multiple heterogeneous data sources for within-field crop yield prediction. Zhang et al. [262] utilize context-aware tensor decomposition and iterative multi-view learning to combine cellphone call detail records and transportation data for improving single-view mobility inference. Zhao et al. [272] propose a site selection framework that learns functions of architecture from multi-source urban big data. Zhang et al. [269] extract physical and human semantic features from remote sensing images, POIs, and real-time social media users. They then map them to common subspaces to obtain cross-correlations that enable the recognition of urban functions. Jenkins et al. [89] employ Denoising AutoEncoders and Graph Convolutional Networks to jointly learn region representations from satellite images, POIs, human mobility data, and spatial graph data. Liu et al. [141] propose a federated learning-based GRU network for traffic flow prediction that updates universal learning models through a secure parameter aggregation mechanism rather than by directly sharing raw data across organizations.

Lightweight AI [267] for rapid decision-making close to the data source is a promising direction for the above-mentioned work to move toward more innovative IoT scenarios.

5 PROSPECTS: TRENDS AND FUTURE DIRECTIONS

Based on the review of DQ technologies in Sections 3 and 4, we find that SID quality management is being integrated with different learning techniques[8] and that SID quality related computing is becoming increasingly relevant in dynamic, decentralized, and heterogeneous settings. We proceed to present emerging trends in Section 5.1, and discuss future directions in Section 5.2.

5.1 Emerging Trends

Privacy-preserving Computing. SID, and IoT data in general, may include sensitive data. The use of cloud computing and the decentralized architecture of the IoT combine to yield new requirements for privacy protection and security. Thus, an important direction is to enable effective, privacy-preserving, and secure SID management and analysis [196]. We have seen that SID is often encrypted, obscured, anonymized, or hidden, to address privacy requirements. However, this often comes at the cost of reduced usability of the data from the perspective of applications. In this context, studies have focused on effective queries [73, 94, 249], analyses [179], and decision-making [155] on encrypted or obscured SID. In addition, from the perspective of quality management, how to construct privacy-preserving data representations (e.g., embeddings [89, 142, 217]) or effective cryptographic solutions [47, 73, 94, 249] also call for in-depth research. With the increasing prominence of data protection regulations such as GDPR [1] and CCPA [4], we anticipate much more research dedicated to secure yet effective SID computing.

[8]Please also refer to the connections between techniques and tasks in Supplementary Material Section A.1.

Edge/Fog Computing. The decentralized IoT architecture, where data is created at the edge, also introduces challenges and opportunities related to data processing. In particular, the architecture offers exciting opportunities for edge or fog computing to improve processing efficiency and reduce central, single-point workloads. Market intelligence firm IDC [2] predicts that at least 40% of IoT data will be stored and processed at the edge or close to the edge. To handle quality issues of SID, edge/fog computing has been combined with stream computing [118], blockchain technology [35], transport SDN [161], lightweight AI [267] and system-on-a-chip [151, 274] to increase system scalability, autonomy, and economy.

Reinforcement and Incremental Learning. SID often tracks evolving processes and is updated dynamically. This nature of the data calls for processing models with corresponding capabilities of dynamic and incremental processing. For example, many control and decision-making processes can be abstracted into reinforcement learning models whose parameters can be adjusted incrementally. In the handling of SID quality issues, reinforcement learning has proven effective at addressing data sparseness and incompleteness [104, 192, 199, 218, 226], and reinforcement learning can be expected to find use in a broader range of quality management, data analysis, and decision-making tasks on streaming and dynamically changing SID.

Comprehensive Data Fusion for Improved DQ. Multi-source, multi-modal, and heterogeneous urban IoT data is becoming increasingly available [135]. Research on such data has focused on how to effectively integrate diverse and rich, but also biased, spatiotemporal data sources for better DQ from different technical perspectives. First, multi-task [67, 164, 253, 273] and multi-view learning [260, 272] are being used to extract latent and high-quality features based on correlations in multi-source data. Second, techniques based on transfer learning, federated learning, and pre-trained models [72, 105, 141, 155, 245] are being studied that aim to utilize diverse data to enhance the richness and expressiveness of the training data. Third, representation learning techniques [40, 48, 89, 199, 217] have been proposed that attempt to map heterogeneous and multi-modal data to subspaces to enable joint utilization of their information. Finally, techniques based on data integration [23, 98, 169, 264] aim to exploit extra knowledge or expertise to enhance the quality of SID and the interpretability of models of SID.

5.2 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. We offer several promising directions from this perspective.

Dynamic DQ Modeling. SID collection, processing, and transmission may involve thousands or even millions of heterogeneous and dynamic data nodes, making DQ management potentially very complex. Therefore, effective quality modeling techniques are needed to guide each individual node's data handling and its interaction with other nodes. If capturing decentralized data nodes as vertices and data dependencies between them as edges, representation learning over the constructed dynamic graph [96, 217] holds the potential to enable estimating and predicting quality measures. Furthermore, factors such as external environmental influence, local properties and resource constraints, and the spatiotemporal distribution of nodes can be incorporated into the goal function design of a model to cope with quality modeling in different application contexts.

Secure SID Sharing. Many studies on spatial computing [36, 127, 193, 215, 262] exist that demonstrate the power of integrating multiple data sources. However, IoT data repositories in most enterprises are still in silos, depriving enterprises of valuable insights that can be realized only by mining broader data pools of SID [3]. Constructing such SID data pools calls for trustworthy protocols and data governance mechanisms for secure and reliable data sharing across IoT repositories. Blockchain and federated learning techniques are relevant here—the former authenticate data, and the latter train models globally while safeguarding each enterprise's private data [35, 101].

DQ-aware Task Planning. A variety of quality management services are exploited in IoT settings, including outlier removal, fault correction, compression, and interpolation. From the perspectives of resource optimization, self-adaptivity, and sustainability, it is important to conduct a quantitative cost-benefit analysis of such DQ-related services [209] as a foundation for understanding how they can be applied to optimize DQ locally or globally. To enable fine-grained and reliable cost-benefit analyses, it is relevant to take into account DQ modeling, evolving topology and characteristics of IoT nodes, and the priorities and data dependencies of DQ tasks.

Cross-layer Quality Management. Today's IoT adopts a layered approach that separates DQ tasks with different goals and data scopes logically. In spite of the proliferation and increasing diversity of SID applications, the usage of bottom-layer, general-purpose DQ services (e.g., compression and interpolation) has been rather limited. To enable quality management that is sufficiently general to support diverse applications, an interesting direction is to modularize and containerize services and to organize the resulting modules in a cross-layer fashion [37], e.g., through directed acyclic graph models. To realize this vision, secure and efficient control protocols and interfaces compatible with edge computing and microservices are poised to be key enabling technologies.

Quality Management Middleware for SID. In general, the dynamic nature, heterogeneity, and disorder exhibited by SID represent obstacles to its utilization. To enable ubiquitous quality management of SID and to enable applications to better utilize SID, quality management middleware that fits in the IoT paradigm is highly desirable. Such middleware is expected to integrate the technical directions mentioned above.

We end this section by providing an application perspective. With the continued advances in spatial sensing and autonomous movement, paradigms such as Internet of Vehicles [235], Internet of Flying Robots [85], and Internet of Medical Things [68] are becoming increasingly relevant. However, conflicts and crashes affect peoples' confidence in, and acceptance of, autonomous movement technologies. It is reported [8] that a large fraction of accidents are caused by sensors failing to perceive the environment in a correct and timely manner. Therefore, we believe that DQ technologies for spatial IoT data may play an important role in the context of these paradigms.

6 CONCLUSIONS

In this survey, we focus on the quality-aware utilization of spatial IoT data. First, we analyze the data consumption requirements of SID and define major data quality dimensions. Based on these dimensions, we summarize the significant characteristics of spatial IoT data and identify the associated quality issues related to spatial and thematic attributes. Subsequently, we analyze data quality technologies available for enhancing spatial IoT data and present a taxonomy of these technologies from both task and technique perspectives. Adopting the proposed taxonomy, we extensively review and categorize existing studies on quality management, covering location refinement, uncertainty elimination, outlier removal, fault correction, data integration, and data reduction, and we review studies on low-quality data exploitation, covering querying, analyses, and decision-making. Finally, we provide insight into emerging trends related to data quality in IoT data and discuss the future directions for innovative quality-aware SID utilization.

The survey covers trajectories and spatiotemporal data with general data values separately. Much of work reviewed, while not focused in particular on IoT settings, is applicable to some extent to IoT scenarios. In the coming years, IoT will continue to be a field of continuous development and innovation. Its unique features, such as decentralization, dynamics, and heterogeneity, and the resulting quality issues, will continue to offer opportunities and challenges for the design, development, and deployment of IoT-enabled spatial applications.

ACKNOWLEDGMENTS

We thank the reviewers and the editors for their very insightful, careful, and constructive comments. This work was carried out within the EU MSCA-funded project MALOT (Grant No. 882232) and in collaboration with DIREC, supported by Innovation Fund Denmark. Bo Tang was supported by the NSFC (Grant No. 61802163) and the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001). Muhammad Aamir Cheema was supported by the ARC FT180100140.

REFERENCES

[1] 2016. *General Data Protection Regulation (GDPR).* https://gdpr-info.eu/

[2] 2017. *IDC FutureScape: Worldwide Internet of Things 2018 predictions.* Retrieved Nov 2021 from https://www.idc.com/research/viewtoc.jsp?containerId=US43161517

[3] 2017. *Location-based services for the Internet of Things.* Retrieved Nov 2021 from https://internetofthingsagenda.techtarget.com/blog/IoT-Agenda/Location-based-services-for-the-internet-of-things

[4] 2018. *California Consumer Privacy Act (CCPA).* https://oag.ca.gov/privacy/ccpa

[5] 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

[6] 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/

[7] 2020. *5G and the enablement of IoT edge processing.* Retrieved Nov 2021 from https://hazelcast.com/resources/5g-and-the-enablement-of-iot-edge-processing/

[8] 2021. *The dangers of driverless cars.* Retrieved Nov 2021 from https://www.natlawreview.com/article/dangers-driverless-cars

[9] Alsharif Abuadbba, Ibrahim Khalil, and Xinghuo Yu. 2017. Gaussian approximation-based lossless compression of smart meter readings. *IEEE Transactions on Smart Grid* 9, 5 (2017), 5047–5056.

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

[11] Mamta Agiwal, Navrati Saxena, and Abhishek Roy. 2019. Towards connected living: 5G enabled Internet of Things (IoT). *IETE Technical Review* 36, 2 (2019), 190–202.

[12] 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.

[13] Haidar Al-Khalidi, David Taniar, John Betts, and Sultan Alamri. 2014. Monitoring moving queries inside a safe region. *The Scientific World Journal* 2014 (2014).

[14] Furqan Alam, Rashid Mehmood, Iyad Katib, Nasser N Albogami, and Aiiad Albeshri. 2017. Data fusion and IoT for smart ubiquitous environments: A survey. *IEEE Access* 5 (2017), 9533–9554.

[15] 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.

[16] Fernandes Nicole Ann and Rupali Wagh. 2019. Quality assurance in big data analytics: An IoT perspective. *Telfor Journal* 11, 2 (2019), 114–118.

[17] 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.

[18] Javed Aslam, Zack Butler, Florin Constantin, Valentino Crespi, George Cybenko, and Daniela Rus. 2003. Tracking a moving object with a binary sensor network. In *SenSys*. 150–161.

[19] 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.

[20] Asif Iqbal Baba, Hua Lu, Torben Bach Pedersen, and Xike Xie. 2014. Handling false negatives in indoor RFID data. In *MDM*, Vol. 1. 117–126.

[21] Asif Iqbal Baba, Hua Lu, Xike Xie, and Torben Bach Pedersen. 2013. Spatiotemporal data cleansing for indoor RFID tracking data. In *MDM*, Vol. 1. 187–196.

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

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

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

[25] Payam Barnaghi, Wei Wang, Lijun Dong, and Chonggang Wang. 2013. A linked-data model for semantic sensor streams. In *IEEE GreenCom/iThings/CPScom*. 468–475.

[26] 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.

[27] Thomas Bernecker, Tobias Emrich, Hans-Peter Kriegel, Matthias Renz, Stefan Zankl, and Andreas Züfle. 2011. Efficient probabilistic reverse nearest neighbor query processing on uncertain data. *Proceedings of the VLDB Endowment* 4, 10 (2011), 669–680.

[28] 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.

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

[30] 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.

[31] Robert Goodell Brown. 2004. *Smoothing, forecasting and prediction of discrete time series*. Courier Corporation.

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

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

[34] Carlos Carvalho, Danielo G Gomes, Nazim Agoulmine, and José Neuman De Souza. 2011. Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation. *Sensors* 11, 11 (2011), 10010–10037.

[35] Roberto Casado-Vara, Fernando de la Prieta, Javier Prieto, and Juan M Corchado. 2018. Blockchain framework for IoT data quality via edge computing. In *SenSys BlockSys Workshop*. 19–24.

[36] Dakshak Keerthi Chandra, Pengyang Wang, Jennifer Leopold, and Yanjie Fu. 2019. Collective representation learning on spatiotemporal heterogeneous information networks. In *SIGSPATIAL*. 319–328.

[37] Tung-Chun Chang, Georgios Bouloukakis, Chia-Ying Hsieh, Cheng-Hsin Hsu, and Nalini Venkatasubramanian. 2021. SmartParcels: Cross-layer IoT planning for smart communities. In *IoTDI*. 1–13.

[38] Muhammad Aamir Cheema, Ljiljana Brankovic, Xuemin Lin, Wenjie Zhang, and Wei Wang. 2010. Multi-guarded safe zone: An effective technique to monitor moving circular range queries. In *ICDE*. 189–200.

[39] Muhammad Aamir Cheema, Xuemin Lin, Wei Wang, Wenjie Zhang, and Jian Pei. 2009. Probabilistic reverse nearest neighbor queries on uncertain data. *IEEE Transactions on Knowledge and Data Engineering* 22, 4 (2009), 550–564.

[40] 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).

[41] Chao Chen, Yan Ding, Xuefeng Xie, Shu Zhang, Zhu Wang, and Liang Feng. 2019. TrajCompressor: An online map-matching-based trajectory compression framework leveraging vehicle heading direction and change. *IEEE Transactions on Intelligent Transportation Systems* 21, 5 (2019), 2012–2028.

[42] 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.

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

[44] Jingyu Chen, Ping Chen, Qiuyan Huo, and Xuezhou Xu. 2011. Clustering network-constrained uncertain trajectories. In *FSKD*, Vol. 3. 1657–1662.

[45] 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.

[46] 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.

[47] Liang Chen, Sarang Thombre, Kimmo Järvinen, Elena Simona Lohan, Anette Alén-Savikko, Helena Leppäkoski, M Zahidul H Bhuiyan, Shakila Bu-Pasha, Giorgia Nunzia Ferrara, Salomon Honkala, et al. 2017. Robustness, security and privacy in location-based services for future IoT: A survey. *IEEE Access* 5 (2017), 8956–8977.

[48] 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.

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

[50] 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.

[51] Qing Cheng, Huiqing Liu, Huanfeng Shen, Penghai Wu, and Liangpei Zhang. 2017. A spatial and temporal nonlocal filter-based data fusion method. *IEEE Transactions on Geoscience and Remote Sensing* 55, 8 (2017), 4476–4488.

[52] 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.
[53] 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.
[54] Reynold Cheng, Dmitri V Kalashnikov, and Sunil Prabhakar. 2004. Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge and Data Engineering* 16, 9 (2004), 1112–1127.
[55] Peng Dai, Yuan Yang, Manyi Wang, and Ruqiang Yan. 2019. Combination of DNN and improved KNN for indoor location fingerprinting. *Wireless Communications and Mobile Computing* 2019 (2019).
[56] Xiangyuan Dai, Man Lung Yiu, Nikos Mamoulis, Yufei Tao, and Michail Vaitis. 2005. Probabilistic spatial queries on existentially uncertain data. In *SSTD*. 400–417.
[57] Julio Cesar Stacchini de Souza, Tatiana Mariano Lessa Assis, and Bikash Chandra Pal. 2015. Data compression in smart distribution systems via singular value decomposition. *IEEE Transactions on Smart Grid* 8, 1 (2015), 275–284.
[58] 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.
[59] 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.
[60] 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.
[61] Tobias Emrich, Hans-Peter Kriegel, Nikos Mamoulis, Johannes Niedermayer, Matthias Renz, and Andreas Züfle. 2014. Reverse-nearest neighbor queries on uncertain moving object trajectories. In *DASFAA*. 92–107.
[62] Tobias Emrich, Hans-Peter Kriegel, Nikos Mamoulis, Matthias Renz, and Andreas Zufle. 2012. Querying uncertain spatio-temporal data. In *ICDE*. 354–365.
[63] 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.
[64] 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.
[65] 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.
[66] Massimo Ficco, Christian Esposito, and Aniello Napolitano. 2013. Calibrating indoor positioning systems with low efforts. *IEEE Transactions on Mobile Computing* 13, 4 (2013), 737–751.
[67] Yuyang Gao and Liang Zhao. 2018. Incomplete label multi-task ordinal regression for spatial event scale forecasting. In *AAAI*. 2999–3006.
[68] Arthur Gatouillat, Youakim Badr, Bertrand Massot, and Ervin Sejdić. 2018. Internet of medical things: A review of recent contributions dealing with cyber-physical systems in medicine. *IEEE Internet of Things Journal* 5, 5 (2018), 3810–3822.
[69] Davide Giovanelli, Elisabetta Farella, Daniele Fontanelli, and David Macii. 2018. Bluetooth-based indoor positioning through ToF and RSSI data fusion. In *IPIN*. 1–8.
[70] Michael F Goodchild. 1998. Uncertainty: The Achilles heel of GIS. *Geo Info Systems* 8, 11 (1998), 50–52.
[71] Michael F Goodchild. 2013. The quality of big (geo) data. *Dialogues in Human Geography* 3, 3 (2013), 280–284.
[72] Chenjuan Guo, Bin Yang, Jilin Hu, and Christian S. Jensen. 2018. Learning to route with sparse trajectory sets. In *ICDE*. 1073–1084.
[73] Cheng Guo, Ruhan Zhuang, Chunhua Su, Charles Zhechao Liu, and Kim-Kwang Raymond Choo. 2019. Secure and efficient k nearest neighbor query over encrypted uncertain data in cloud-IoT ecosystem. *IEEE Internet of Things Journal* 6, 6 (2019), 9868–9879.
[74] Long Guo, Dongxiang Zhang, Guoliang Li, Kian-Lee Tan, and Zhifeng Bao. 2015. Location-aware pub/sub system: When continuous moving queries meet dynamic event streams. In *SIGMOD*. 843–857.
[75] 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.
[76] Stephen C Guptill and Joel L Morrison. 2013. *Elements of Spatial Data Quality*. Elsevier.
[77] Torsten Hägerstraand. 1970. What about people in regional Science? *Papers in Regional Science* 24, 1 (1970), 7–24.
[78] 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.
[79] Arif Hidayat, Muhammad Aamir Cheema, Xuemin Lin, Wenjie Zhang, and Ying Zhang. 2021. Continuous monitoring of moving skyline and top-k queries. *The VLDB Journal* (2021).
[80] 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.

[81] Kathleen Hornsby and Max J Egenhofer. 2002. Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence* 36, 1 (2002), 177–194.
[82] 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.
[83] Chunchun Hu, Xionghua Kang, Nianxue Luo, and Qiansheng Zhao. 2015. Parallel clustering of big data of spatio-temporal trajectory. In *ICNC*. 769–774.
[84] Ming Hua, Jian Pei, Wenjie Zhang, and Xuemin Lin. 2008. Ranking queries on uncertain data: A probabilistic threshold approach. In *SIGMOD*. 673–686.
[85] Hailong Huang and Andrey V Savkin. 2018. Towards the internet of flying robots: A survey. *Sensors* 18, 11 (2018), 4038.
[86] 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.
[87] George Rosario Jagadeesh and Thambipillai Srikanthan. 2014. Robust real-time route inference from sparse vehicle position data. In *ITSC*. 296–301.
[88] Shawn R Jeffery, Minos Garofalakis, and Michael J Franklin. 2006. Adaptive cleaning for RFID data streams. *Proceedings of the VLDB Endowment* 6, 163–174.
[89] Porter Jenkins, Ahmad Farag, Suhang Wang, and Zhenhui Li. 2019. Unsupervised representation learning of spatial data via multimodal embedding. In *CIKM*. 1993–2002.
[90] 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.
[91] 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.
[92] Zhe Jiang. 2018. A survey on spatial prediction methods. *IEEE Transactions on Knowledge and Data Engineering* 31, 9 (2018), 1645–1664.
[93] 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).
[94] Ibrahim Kamel, Ayesha M Talha, and Zaher Al Aghbari. 2017. Dynamic spatial index for efficient query processing on the cloud. *Journal of Cloud Computing* 6, 1 (2017), 1–16.
[95] 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.
[96] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. 2020. Representation learning for dynamic graphs: A survey. *Journal of Machine Learning Research* 21, 70 (2020), 1–73.
[97] 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.
[98] Daniel Kuemper, Thorben Iggena, Ralf Toenjes, and Elke Pulvermueller. 2018. Valid.IoT: A framework for sensor data quality analysis and interpolation. In *MMSys*. 294–303.
[99] 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.
[100] Bart Kuijpers, Harvey J Miller, and Walied Othman. 2011. Kinetic space-time prisms. In *SIGSPATIAL/GIS*. 162–170.
[101] Rajesh Kumar, Abdullah Aman Khan, Jay Kumar, A Zakria, Noorbakhsh Amiri Golilarz, Simin Zhang, Yang Ting, Chengyu Zheng, and WenYong Wang. 2021. Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging. *IEEE Sensors Journal* (2021).
[102] 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.
[103] Ralph Lange, Frank Dürr, and Kurt Rothermel. 2011. Efficient real-time trajectory tracking. *The VLDB Journal* 20, 5 (2011), 671–694.
[104] Truc Viet Le, Siyuan Liu, and Hoong Chuin Lau. 2016. Reinforcement learning framework for modeling spatial sequential decisions under uncertainty. In *AAMAS*. 1449–1450.
[105] Anliang Li, Shuang Wang, Wenzhu Li, Shengnan Liu, and Siyuan Zhang. 2020. Predicting human mobility with federated learning. In *SIGSPATIAL*. 441–444.
[106] Bo Li, Omid Sarbishei, Hosein Nourani, and Tristan Glatard. 2018. A multi-dimensional extension of the lightweight temporal compression method. In *IEEE Big Data*. 2918–2923.
[107] Chuanwen Li, Yu Gu, Jianzhong Qi, Ge Yu, Rui Zhang, and Wang Yi. 2014. Processing moving k NN queries using influential neighbor sets. *Proceedings of the VLDB Endowment* 8, 2 (2014), 113–124.
[108] Deren Li, Jingxiong Zhang, and Huayi Wu. 2012. Spatial data quality and beyond. *International Journal of Geographical Information Science* 26, 12 (2012), 2277–2290.

[109] 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.
[110] 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.
[111] Huan Li, Hua Lu, Lidan Shou, Gang Chen, and Ke Chen. 2018. Finding most popular indoor semantic locations using uncertain mobility data. *IEEE Transactions on Knowledge and Data Engineering* 31, 11 (2018), 2108–2123.
[112] Huan Li, Hua Lu, Lidan Shou, Gang Chen, and Ke Chen. 2018. In search of indoor dense regions: An approach using indoor positioning data. *IEEE Transactions on Knowledge and Data Engineering* 30, 8 (2018), 1481–1495.
[113] 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.
[114] Lingxiao Li, Muhammad Aamir Cheema, Mohammed Eunus Ali, Hua Lu, and David Taniar. 2020. Continuously monitoring alternative shortest paths on road networks. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2243–2255.
[115] Lun Li, Xiaohang Chen, Qizhi Liu, and Zhifeng Bao. 2020. A data-driven approach for GPS trajectory data cleaning. In *DASFAA*. 3–19.
[116] Lixin Li, Xingyou Zhang, James B Holt, Jie Tian, and Reinhard Piltner. 2011. Spatiotemporal interpolation methods for air pollution exposure. In *SARA*.
[117] Songnian Li, Suzana Dragicevic, Francesc Antón Castro, Monika Sester, Stephan Winter, Arzu Coltekin, Christopher Pettit, Bin Jiang, James Haworth, Alfred Stein, et al. 2016. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing* 115 (2016), 119–133.
[118] 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.
[119] 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.
[120] 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.
[121] Yang Li, Yangyan Li, Dimitrios Gunopulos, and Leonidas Guibas. 2016. Knowledge-based trajectory completion from sparse GPS samples. In *SIGSPATIAL*. 1–10.
[122] 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).
[123] 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.
[124] 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.
[125] Lin Liao, Dieter Fox, and Henry Kautz. 2007. Extracting places and activities from GPS traces using hierarchical conditional random fields. *The International Journal of Robotics Research* 26, 1 (2007), 119–134.
[126] Lin Liao, Donald J Patterson, Dieter Fox, and Henry Kautz. 2007. Learning and inferring transportation routines. *Artificial Intelligence* 171, 5-6 (2007), 311–331.
[127] Lu Lin, Jianxin Li, Feng Chen, Jieping Ye, and Jinpeng Huai. 2017. Road traffic speed prediction: A probabilistic model fusing multi-source data. *IEEE Transactions on Knowledge and Data Engineering* 30, 7 (2017), 1310–1323.
[128] 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.
[129] 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.
[130] Yijun Lin, Yao-Yi Chiang, Fan Pan, Dimitrios Stripelis, José Luis Ambite, Sandrah P Eckel, and Rima Habre. 2017. Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. In *SIGSPATIAL*. 1–10.
[131] Caihua Liu, Patrick Nitschke, Susan P Williams, and Didar Zowghi. 2020. Data quality and the Internet of Things. *Computing* 102, 2 (2020), 573–599.
[132] 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.
[133] Hui Liu, Houshang Darabi, Pat Banerjee, and Jing Liu. 2007. Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, 6 (2007), 1067–1080.
[134] Jingbin Liu, Ruizhi Chen, Ling Pei, Robert Guinness, and Heidi Kuusniemi. 2012. A hybrid smartphone indoor positioning solution for mobile LBS. *Sensors* 12, 12 (2012), 17208–17233.
[135] Jia Liu, Tianrui Li, Peng Xie, Shengdong Du, Fei Teng, and Xin Yang. 2020. Urban big data fusion based on deep learning: An overview. *Information Fusion* 53 (2020), 123–133.

[136] Jiajun Liu, Kun Zhao, Philipp Sommer, Shuo Shang, Brano Kusy, and Raja Jurdak. 2015. Bounded quadrant system: Error-bounded trajectory compression on the go. In *ICDE*. 987–998.

[137] 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.

[138] Mo Liu, Ming Li, Denis Golovnya, Elke A Rundensteiner, and Kajal Claypool. 2009. Sequence pattern query processing over out-of-order event streams. In *ICDE*. 784–795.

[139] Qiyu Liu, Xiang Lian, and Lei Chen. 2019. Probabilistic maximum range-sum queries on spatial database. In *SIGSPATIAL*. 159–168.

[140] 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.

[141] 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.

[142] Yiding Liu, Kaiqi Zhao, Gao Cong, and Zhifeng Bao. 2020. Online anomalous trajectory detection with deep generative sequence modeling. In *ICDE*. 949–960.

[143] Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28, 2 (1982), 129–137.

[144] 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.

[145] Hua Lu, Chenjuan Guo, Bin Yang, and Christian S. Jensen. 2016. Finding frequently visited indoor POIs using symbolic indoor tracking data. In *EDBT*. 449–460.

[146] Hua Lu, Bin Yang, and Christian S. Jensen. 2011. Spatio-temporal joins on symbolic indoor tracking data. In *ICDE*. 816–827.

[147] 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.

[148] 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.

[149] Altti Ilari Maarala, Xiang Su, and Jukka Riekki. 2016. Semantic reasoning for context-aware Internet of Things applications. *IEEE Internet of Things Journal* 4, 2 (2016), 461–473.

[150] Pavel Mach and Zdenek Becvar. 2017. Mobile edge computing: A survey on architecture and computation offloading. *IEEE Communications Surveys & Tutorials* 19, 3 (2017), 1628–1656.

[151] Michele Magno, Xiaying Wang, Manuel Eggimann, Lukas Cavigelli, and Luca Benini. 2020. InfiniWolf: Energy efficient smart bracelet for edge computing with dual source energy harvesting. In *DATE*. 342–345.

[152] 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.

[153] 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.

[154] Jiali Mao, Jiaye Liu, Cheqing Jin, and Aoying Zhou. 2021. Feature grouping–based trajectory outlier detection over distributed streams. *ACM Transactions on Intelligent Systems and Technology* 12, 2 (2021), 1–23.

[155] Shunmei Meng, Lianyong Qi, Qianmu Li, Wenmin Lin, Xiaolong Xu, and Shaohua Wan. 2019. Privacy-preserving and sparsity-aware location-based prediction method for collaborative recommender systems. *Future Generation Computer Systems* 96 (2019), 324–335.

[156] Weixiao Meng, Ying He, Zhian Deng, and Cheng Li. 2012. Optimized access points deployment for WLAN indoor positioning system. In *WCNC*. 2457–2461.

[157] Mostafa Milani, Zheng Zheng, and Fei Chiang. 2019. CurrentClean: Spatio-temporal cleaning of stale data. In *ICDE*. 172–183.

[158] Mehdi Mohammadi, Ala Al-Fuqaha, Sameh Sorour, and Mohsen Guizani. 2018. Deep learning for IoT big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials* 20, 4 (2018), 2923–2960.

[159] 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.

[160] Jonathan Muckell, Paul W Olsen, Jeong-Hyon Hwang, Catherine T Lawson, and SS Ravi. 2014. Compression of trajectory data: A comprehensive evaluation and new approach. *GeoInformatica* 18, 3 (2014), 435–460.

[161] Raül Muñoz, Ricard Vilalta, Noboru Yoshikane, Ramon Casellas, Ricardo Martínez, Takehiro Tsuritani, and Itsuro Morita. 2018. Integration of IoT, transport SDN, and edge/cloud computing for dynamic distribution of IoT analytics and efficient use of network resources. *IEEE Journal of Lightwave Technology* 36, 7 (2018), 1420–1428.

[162] Christopher Mutschler and Michael Philippsen. 2013. Distributed low-latency out-of-order event processing for high data rate sensor streams. In *IPDPS*. 1133–1144.

[163] Kapileswar Nellore and Gerhard P Hancke. 2016. A survey on urban traffic management system using wireless sensor networks. *Sensors* 16, 2 (2016), 157.

[164] 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.

[165] Dragoş Niculescu and Badri Nath. 2003. DV based positioning in ad hoc networks. *Telecommunication Systems* 22, 1 (2003), 267–280.

[166] Johannes Niedermayer, Andreas Züfle, Tobias Emrich, Matthias Renz, Nikos Mamoulis, Lei Chen, and Hans-Peter Kriegel. 2013. Probabilistic nearest neighbor queries on uncertain moving object trajectories. *Proceedings of the VLDB Endowment* 7, 3 (2013), 205–216.

[167] 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.

[168] Sarana Nutanong, Rui Zhang, Egemen Tanin, and Lars Kulik. 2010. Analysis and evaluation of v*-k nn: An efficient algorithm for moving k nn queries. *The VLDB Journal* 19, 3 (2010), 307–332.

[169] 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.

[170] Robin Wentao Ouyang, Mani Srivastava, Alice Toniolo, and Timothy J Norman. 2015. Truth discovery in crowdsourced detection of spatial events. *IEEE Transactions on Knowledge and Data Engineering* 28, 4 (2015), 1047–1060.

[171] 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.

[172] Jian Pei, Bin Jiang, Xuemin Lin, and Yidong Yuan. 2007. Probabilistic skylines on uncertain data. In *VLDB*. 15–26.

[173] 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.

[174] 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.

[175] Ricardo Perez-Castillo, Ana G Carretero, Moises Rodriguez, Ismael Caballero, Mario Piattini, Alejandro Mate, Sunho Kim, and Dongwoo Lee. 2018. Data quality best practices in IoT environments. In *QUATIC*. 272–275.

[176] Dieter Pfoser and Christian S. Jensen. 1999. Capturing the uncertainty of moving-object representations. In *SSD*. 111–131.

[177] Iulian Sandu Popa, Karine Zeitouni, Vincent Oria, and Ahmed Kharrat. 2015. Spatio-temporal compression of trajectories in road networks. *GeoInformatica* 19, 1 (2015), 117–145.

[178] Sitthapon Pumpichet, Niki Pissinou, Xinyu Jin, and Deng Pan. 2012. Belief-based cleaning in trajectory sensor streams. In *ICC*. 208–212.

[179] Apostolos Pyrgelis, Emiliano De Cristofaro, and Gordon J Ross. 2016. Privacy-friendly mobility analytics using aggregate location data. In *SIGSPATIAL*. 1–10.

[180] 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.

[181] Shaojie Qiao, Changjie Tang, Huidong Jin, Teng Long, Shucheng Dai, Yungchang Ku, and Michael Chau. 2010. PutMode: Prediction of uncertain trajectories in moving objects databases. *Applied Intelligence* 33, 3 (2010), 370–386.

[182] 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.

[183] Suprio Ray, Bogdan Simion, Angela Demke Brown, and Ryan Johnson. 2013. A parallel spatial data analysis infrastructure for the cloud. In *SIGSPATIAL*. 284–293.

[184] Fabio Sartori, Riccardo Melen, and Fabio Giudici. 2019. IoT data validation using spatial and temporal correlations. In *MTSR*. 77–89.

[185] Omer Berat Sezer, Erdogan Dogdu, and Ahmet Murat Ozbayoglu. 2017. Context-aware computing, learning, and big data in Internet of Things: A survey. *IEEE Internet of Things Journal* 5, 1 (2017), 1–27.

[186] 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.

[187] Eugene Siow, Thanassis Tiropanis, and Wendy Hall. 2018. Analytics for the Internet of Things: A survey. *Comput. Surveys* 51, 4 (2018), 1–36.

[188] Shaoxu Song, Ruihong Huang, Yue Cao, and Jianmin Wang. 2021. Cleaning timestamps with temporal constraints. *The VLDB Journal* (2021), 1–22.

[189] Shaoxu Song and Aoqian Zhang. 2020. IoT data quality. In *CIKM*. 3517–3518.

[190] Xiaozhao Song, Jiajie Xu, Rui Zhou, Chengfei Liu, Kai Zheng, Pengpeng Zhao, and Nickolas Falkner. 2020. Collective spatial keyword search on activity trajectories. *GeoInformatica* 24, 1 (2020), 61–84.

[191] Han Su, Kai Zheng, Haozhou Wang, Jiamin Huang, and Xiaofang Zhou. 2013. Calibrating trajectory data for similarity-based analysis. In *SIGMOD*. 833–844.

[192] 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).

[193] Lu Sun, Wei Zhou, Baichen Jiang, and Jian Guan. 2017. A real-time similarity measure model for multi-source trajectories. In *CIIS*. 257–262.

[194] Zi-Yun Sun, Ming-Che Tsai, and Hsiao-Ping Tsai. 2014. Mining uncertain sequence data on Hadoop platform. In *PAKDD*. 204–215.

[195] Ferry Susanto, Paulo De Souza, and Jing He. 2016. Spatiotemporal interpolation for environmental modelling. *Sensors* 16, 8 (2016), 1245.

[196] Romana Talat, Mohammad S Obaidat, Muhammad Muzammal, Ali Hassan Sodhro, Zongwei Luo, and Sandeep Pirbhulal. 2020. A decentralised approach to privacy preserving trajectory mining. *Future Generation Computer Systems* 102 (2020), 382–392.

[197] Liansheng Tan and Mou Wu. 2015. Data reduction in wireless sensor networks: A hierarchical LMS prediction approach. *IEEE Sensors Journal* 16, 6 (2015), 1708–1715.

[198] Luliang Tang, Xue Yang, Zihan Kan, and Qingquan Li. 2015. Lane-level road information mining from vehicle GPS trajectories based on naïve bayesian classification. *ISPRS International Journal of Geo-Information* 4, 4 (2015), 2660–2680.

[199] 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.

[200] Yufei Tao, Xiaokui Xiao, and Reynold Cheng. 2007. Range search on multidimensional uncertain data. *ACM Transactions on Database Systems* 32, 3 (2007), 15–es.

[201] 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.

[202] Christopher Taylor, Ali Rahimi, Jonathan Bachrach, Howard Shrobe, and Anthony Grue. 2006. Simultaneous localization, calibration, and tracking in an ad hoc sensor network. In *IPSN*. 27–33.

[203] 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.

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

[205] Goce Trajcevski, Alok Choudhary, Ouri Wolfson, Li Ye, and Gang Li. 2010. Uncertain range queries for necklaces. In *MDM*. 199–208.

[206] Goce Trajcevski, Roberto Tamassia, Hui Ding, Peter Scheuermann, and Isabel F Cruz. 2009. Continuous probabilistic nearest-neighbor queries for uncertain trajectories. In *EDBT*. 874–885.

[207] Sharda Tripathi and Swades De. 2018. An efficient data characterization and reduction scheme for smart metering infrastructure. *IEEE Transactions on Industrial Informatics* 14, 10 (2018), 4300–4308.

[208] Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T Yang. 2013. Data mining for Internet of Things: A survey. *IEEE Communications Surveys & Tutorials* 16, 1 (2013), 77–97.

[209] M Jeroen Van Der Donckt, Danny Weyns, M Usman Iftikhar, and Ritesh Kumar Singh. 2018. Cost-benefit analysis at runtime for self-adaptive systems applied to an Internet of Things application. In *ENASE*. 478–490.

[210] Hoang Vo, Ablimit Aji, and Fusheng Wang. 2014. SATO: A spatial data partitioning framework for scalable query processing. In *SIGSPATIAL*. 545–548.

[211] Khuong Vu and Rong Zheng. 2013. Efficient algorithms for spatial skyline query with uncertainty. In *SIGSPATIAL*. 412–415.

[212] Jiangtao Wang, Yasha Wang, Daqing Zhang, Feng Wang, Haoyi Xiong, Chao Chen, Qin Lv, and Zhaopeng Qiu. 2018. Multi-task allocation in mobile crowd sensing with individual task quality assurance. *IEEE Transactions on Mobile Computing* 17, 9 (2018), 2101–2113.

[213] Lizhen Wang, Pinping Wu, and Hongmei Chen. 2011. Finding probabilistic prevalent colocations in spatially uncertain data sets. *IEEE Transactions on Knowledge and Data Engineering* 25, 4 (2011), 790–804.

[214] Peixiao Wang, Fei Gao, Yuhui Zhao, Ming Li, and Xinyan Zhu. 2020. Detection of indoor high-density crowds via Wi-Fi tracking data. *Sensors* 20, 18 (2020), 5078.

[215] Pu Wang, Jiyu Lai, Zhiren Huang, Qian Tan, and Tao Lin. 2020. Estimating traffic flow in large road networks based on multi-source traffic data. *IEEE Transactions on Intelligent Transportation Systems* (2020).

[216] 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.

[217] Wei Wang, Feng Xia, Hansong Nie, Zhikui Chen, Zhiguo Gong, Xiangjie Kong, and Wei Wei. 2020. Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems* (2020).
[218] Zheng Wang, Cheng Long, and Gao Cong. 2021. Trajectory simplification with reinforcement learning. In *ICDE*.
[219] Zhanquan Wang, Bowen Lu, Fangli Ying, Man Kong, and Minwei Tang. 2017. Research of mining algorithms for uncertain spatio-temporal co-occurrence pattern. In *KST*. 12–17.
[220] Zhi-Jie Wang, Dong-Hua Wang, Bin Yao, and Minyi Guo. 2014. Probabilistic range query over uncertain moving objects in constrained two-dimensional space. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2014), 866–879.
[221] Ling-Yin Wei, Ya-Ting Hsu, Wen-Chih Peng, and Wang-Chien Lee. 2014. Indexing spatial data in cloud data managements. *Pervasive and Mobile Computing* 15 (2014), 48–61.
[222] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. 2012. Constructing popular routes from uncertain trajectories. In *KDD*. 195–203.
[223] Robert Weibel. 1996. Generalization of spatial data: Principles and selected algorithms. In *Algorithmic Foundations of Geographic Information Systems*. 99–152.
[224] Fei Wu and Zhenhui Li. 2016. Where did you go: Personalized annotation of mobility records. In *CIKM*. 589–598.
[225] Fei Wu, Zhenhui Li, Wang-Chien Lee, Hongjian Wang, and Zhuojie Huang. 2015. Semantic annotation of mobility data using social media. In *WWW*. 1253–1263.
[226] 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.
[227] Hao Wu, Weiwei Sun, and Baihua Zheng. 2017. A fast trajectory outlier detection approach via driving behavior modeling. In *CIKM*. 837–846.
[228] 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.
[229] Zheng Wu, Esrafil Jedari, Roberto Muscedere, and Rashid Rashidzadeh. 2016. Improved particle filter based on WLAN RSSI fingerprinting and smart sensors for indoor localization. *Computer Communications* 83 (2016), 64–71.
[230] Zhenyu Wu, Yuan Xu, Yunong Yang, Chunhong Zhang, Xinning Zhu, and Yang Ji. 2017. Towards a semantic web of things: A hybrid semantic annotation, extraction, and reasoning framework for cyber-physical system. *Sensors* 17, 2 (2017), 403.
[231] Dong Xie, Feifei Li, and Jeff M Phillips. 2017. Distributed trajectory similarity search. *Proceedings of the VLDB Endowment* 10, 11 (2017), 1478–1489.
[232] Xike Xie, Hua Lu, and Torben Bach Pedersen. 2013. Efficient distance-aware query evaluation on indoor moving objects. In *ICDE*. 434–445.
[233] Xike Xie, Hua Lu, and Torben Bach Pedersen. 2014. Distance-aware join for indoor moving objects. *IEEE Transactions on Knowledge and Data Engineering* 27, 2 (2014), 428–442.
[234] Jianqiu Xu and Ralf Hartmut Güting. 2013. A generic data model for moving objects. *GeoInformatica* 17, 1 (2013), 125–172.
[235] Wenchao Xu, Haibo Zhou, Nan Cheng, Feng Lyu, Weisen Shi, Jiayin Chen, and Xuemin Shen. 2017. Internet of vehicles in big data era. *IEEE/CAA Journal of Automatica Sinica* 5, 1 (2017), 19–35.
[236] Yongyang Xu, Zhanlong Chen, Zhong Xie, and Liang Wu. 2017. Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science* 31, 10 (2017), 1929–1951.
[237] Ying Xu, Dongxiang Zhang, Meihui Zhang, Dongsheng Li, Xiaoling Wang, and Heng Tao Shen. 2018. Continuous proximity detection via predictive safe region construction. In *ICDE*. 629–640.
[238] Da Yan, Zhou Zhao, Wilfred Ng, and Steven Liu. 2014. Probabilistic convex hull queries over uncertain data. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2014), 852–865.
[239] 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.
[240] Bin Yang, Hua Lu, and Christian S. Jensen. 2009. Scalable continuous range monitoring of moving objects in symbolic indoor space. In *CIKM*. 671–680.
[241] Bin Yang, Hua Lu, and Christian S. Jensen. 2010. Probabilistic threshold k nearest neighbor queries over moving objects in symbolic indoor space. In *EDBT*. 335–346.
[242] Dingyu Yang, Dongxiang Zhang, Kian-Lee Tan, Jian Cao, and Frédéric Le Mouël. 2014. CANDS: Continuous optimal navigation via distributed stream processing. *Proceedings of the VLDB Endowment* 8, 2 (2014), 137–148.
[243] 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.

[244] Yuqing Yang, Jianghui Cai, Haifeng Yang, Jifu Zhang, and Xujun Zhao. 2020. TAD: A trajectory clustering algorithm based on spatial-temporal density analysis. *Expert Systems with Applications* 139 (2020), 112846.

[245] 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.

[246] Ibrar Yaqoob, Ejaz Ahmed, Ibrahim Abaker Targio Hashem, Abdelmuttlib Ibrahim Abdalla Ahmed, Abdullah Gani, Muhammad Imran, and Mohsen Guizani. 2017. Internet of Things architecture: Recent advances, taxonomy, requirements, and open challenges. *IEEE Wireless Communications* 24, 3 (2017), 10–16.

[247] Jaegeol Yim, Chansik Park, Jaehun Joo, and Seunghwan Jeong. 2008. Extended Kalman filter for wireless LAN based indoor positioning. *Decision Support Systems* 45, 4 (2008), 960–971.

[248] Yihang Yin, Fengzheng Liu, Xiang Zhou, and Quanzhong Li. 2015. An efficient data compression model based on spatial clustering and principal component analysis in wireless sensor networks. *Sensors* 15, 8 (2015), 19443–19465.

[249] 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.

[250] Simin You, Jianting Zhang, and Le Gruenwald. 2015. Large-scale spatial join query processing in cloud. In *ICDE Workshops*. 34–41.

[251] 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.

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

[253] Shuai Yuan, Jiayu Zhou, Pang-Ning Tan, Emi Fergus, Tyler Wagner, and Patricia Soranno. 2017. Multi-level multi-task learning for modeling cross-scale interactions in nested geospatial data. In *ICDM*. 1153–1158.

[254] 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.

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

[256] Aoqian Zhang, Shaoxu Song, Jianmin Wang, and Philip S Yu. 2017. Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1046–1057.

[257] Bing Zhang, Goce Trajcevski, and Liu Liu. 2016. Towards fusing uncertain location data from heterogeneous sources. *GeoInformatica* 20, 2 (2016), 179–212.

[258] 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.

[259] 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).

[260] 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.

[261] 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.

[262] 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.

[263] 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.

[264] 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.

[265] 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.

[266] Wenjie Zhang, Xuemin Lin, Ying Zhang, Muhammad Aamir Cheema, and Qing Zhang. 2012. Stochastic skylines. *ACM Transactions on Database Systems* 37, 2 (2012), 1–34.

[267] 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.

[268] Xichen Zhang, Suprio Ray, Farzaneh Shoeleh, and Rongxing Lu. 2021. Efficient contact similarity query over uncertain trajectories. In *EDBT*. 403–408.

[269] Yatao Zhang, Qingquan Li, Wei Tu, Ke Mai, Yao Yao, and Yiyong Chen. 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. *Computers, Environment and Urban Systems* 78 (2019), 101374.

[270] 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.

[271] Yueyue Zhang, Song Xing, Yaping Zhu, Feng Yan, and Lianfeng Shen. 2017. RSS-based localization in WSNs using Gaussian mixture model via semidefinite relaxation. *IEEE Communications Letters* 21, 6 (2017), 1329–1332.

[272] Guoshuai Zhao, Tianlei Liu, Xueming Qian, Tao Hou, Huan Wang, Xingsong Hou, and Zhetao Li. 2017. Location recommendation for enterprises by multi-source urban big data analysis. *IEEE Transactions on Services Computing* (2017).

[273] 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).

[274] Wenfeng Zhao, Biao Sun, Tong Wu, and Zhi Yang. 2018. On-chip neural data compression based on compressed sensing with sparse sensing matrices. *IEEE Transactions on Biomedical Circuits and Systems* 12, 1 (2018), 242–254.

[275] 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.

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

[277] Guanjie Zheng, Susan L Brantley, Thomas Lauvaux, and Zhenhui Li. 2017. Contextual spatial outlier detection with metric learning. In *KDD*. 2161–2170.

[278] Kai Zheng, Shuo Shang, Nicholas Jing Yuan, and Yi Yang. 2013. Towards efficient search for activity trajectories. In *ICDE*. 230–241.

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

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

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

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

[283] 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.

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

[285] 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.

Table 7. Connections Between DQ Tasks and DQ Techniques

Technique \ Task	LR	UE	OR	FC	DI	DR	Querying	Analyses	Decision-making
Prob. M.	[63, 134]	[110, 226]	[239, 255]	[64, 157]	[109, 126]	[119]	[90, 266]	[203, 276]	[181, 264]
STD M.	[229]	[116]	[29, 277]	[43, 64]	[93, 283]	[136, 159]	[62, 166]	[214, 244]	[102, 262]
STR M.		[121, 281]	[255, 256]	[43, 64]	[126, 225]	[34, 267]		[170]	[141, 273]
SC M.	[60]	[191, 226]	[282]	[20, 21]	[109, 110]	[41, 243]	[237, 268]	[222, 258]	[72]
UL			[29, 255]					[142, 170]	[48, 82]
SSL									[40, 46]
RL						[218]		[227]	[192, 199]
MTL/MVL					[260]				[164, 269]
TL									[72, 245]
FL									[105, 155]
Ds. Com.				[162]		[57, 248]	[231, 252]	[83, 194]	
Str. Com.				[138]	[25]	[41, 128]	[153]	[42, 65]	[102]
Col. Com.	[49, 263]	[115, 281]	[198]			[275]			[264]
E/F Com.					[22, 149]	[197, 267]			[147]

A SUPPLEMENTARY MATERIAL

A.1 Connections between DQ Tasks and Techniques

In Table 7, we use some classic studies to illustrate connections between DQ tasks and DQ techniques. An empty cell does not necessarily mean that a certain technique cannot be used for a certain task. It may simply mean that we do not cover studies that represent this combination.

The full names of abbreviated DQ tasks are listed as follows: LR (Location Refinement), UE (Uncertainty Elimination), OR (Outlier Removal), FC (Fault Correction), DI (Data Integration), and DR (Data Reduction).

The full names of abbreviated DQ techniques are listed as follows: Prob. M. (Probabilistic Modeling), STD M. (Spatiotemporal Dependency Modeling), STR M. (Spatiotemporal Regularity Modeling), SC M. (Spatial Constraint Modeling), UL (Unsupervised Learning), SSL (Semi-supervised Learning), RL (Reinforcement Learning), MTL/MVL (Multi-task Learning/Multi-view Learning), TL (Transfer Learning), FL (Federated Learning), Ds. Com. (Distributed Computing), Str. Com. (Stream Computing), Col. Com. (Collaborative Learning), and E/F Com. (Edge/Fog Computing).

A.2 Visual Analytics on Low-quality SID

We introduce the visual analytic studies for uncertain and dynamic SID, respectively.

Visual Analytics on Uncertain SID. Data uncertainty such as imprecision, sparse sampling, and missing values make visual analytics of trajectories and other spatially referenced data more challenging [286, 287]. Some studies [289, 291, 292] address challenges related to uncertainty in visual analytics. To handle uncertainty in visual analyses of urban mobility patterns over sensor network data, Senaratne et al. [292] construct uncertain markers based on space-time prisms. As conflicts from heterogeneous data impede visual human behavior analytics, Chen et al. [289] propose a semi-automatic pattern and outlier detection approach with a pre-defined set of uncertainty types. Further, to enable visual traceability of faulty IoT data, Lomotey et al. [291] use associative rules and lexical chaining methods to identify (un)linkability between IoT devices for correctness checking in sensor data propagation.

Visual Analytics on Dynamic SID. Visualization tools [288, 290] have also been explored in analyzing large-scale and evolving SID. To ease the analysis of high-dimensional air quality measurements, Kalamaras et al. [290] propose a reactive visual analytics platform that aims to support explainable spatial data analysis. Batista et al. [288] develop a set of visualization tools to enhance the understandability of analyses of data collected from a worldwide climate sensor network.

SUPPLEMENTARY REFERENCES

[286] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, and Stefan Wrobel. 2013. *Visual Analytics of Movement.* Springer Science & Business Media.

[287] Natalia Andrienko and Gennady Andrienko. 2013. Visual analytics of movement: An overview of methods, tools and procedures. *Information visualization* 12, 1 (2013), 3–24.

[288] André FM Batista, Pedro LP Correa, and Giri Palanisamy. 2016. Visual analytics improving data understandability in IoT projects: An overview of the US DOE ARM program data science tools. In *MASS*. 349–354.

[289] Siming Chen, Zuchao Wang, Jie Liang, and Xiaoru Yuan. 2018. Uncertainty-aware visual analytics for exploring human behaviors from heterogeneous spatial temporal data. *Journal of Visual Languages & Computing* 48 (2018), 187–198.

[290] Ilias Kalamaras, Ioannis Xygonakis, Konstantinos Glykos, Sigmund Akselsen, Arne Munch-Ellingsen, Hai Thanh Nguyen, Andreas Jacobsen Lepperod, Kerstin Bach, Konstantinos Votis, and Dimitrios Tzovaras. 2019. Visual analytics for exploring air quality data in an AI-enhanced IoT environment. In *MEDES*. 103–110.

[291] Richard K Lomotey, Joseph C Pry, and Chenshean Chai. 2018. Traceability and visual analytics for the Internet-of-Things (IoT) architecture. *World Wide Web* 21, 1 (2018), 7–32.

[292] Hansi Senaratne, Manuel Mueller, Michael Behrisch, Felipe Lalanne, Javier Bustos-Jiménez, Jörn Schneidewind, Daniel Keim, and Tobias Schreck. 2017. Urban mobility analysis with mobile network data: A visual analytics approach. *IEEE Transactions on Intelligent Transportation Systems* 19, 5 (2017), 1537–1546.