10.20378/irb-112076

Data Quality Processing in Data Streaming Environments: A Literature Review

ABOUBAKR BENABBAS and DANIELA NICKLAS*, University of Bamberg, Germany

The increasing reliance on real-time analytics and sensor-driven systems has elevated the importance of maintaining high data quality in streaming environments. This literature review provides an overview of data quality (DQ) concepts, representations, and processing techniques tailored to continuous data streams. It traces the evolution of data quality from early database systems to modern big data and streaming contexts, emphasizing intrinsic, representational, and contextual quality dimensions such as accuracy, completeness, consistency, and timeliness. The paper reviews major DQ models, metrics, and standards, highlighting methods for assessing and improving quality in sensor-based and high-velocity data systems. Furthermore, it examines state-of-the-art data cleaning, fault detection, and anomaly management approaches, identifying their limitations in flexibility and generalizability. As a literature review, it synthesizes key foundational and recent contributions rather than providing an exhaustive systematic survey. The literature was analyzed through targeted review of seminal and recent works focusing on peer-reviewed contributions to DQ models, metrics, and processing techniques. Finally, the study discusses emerging trends such as adaptive, pattern-based, and AI-driven quality processing toward building accessible, real-time, and domain-independent frameworks for quality-aware data stream management.

1 Introduction

In an era dominated by real-time analytics, the continuous generation of data from sensors, Internet of Things (IoT) devices, and online platforms has become a cornerstone of modern information systems. Applications ranging from smart cities and environmental monitoring to financial trading and autonomous vehicles increasingly depend on the timely and accurate processing of streaming data. However, the reliability and usefulness of these applications are critically dependent on the quality of the underlying data. No matter how advanced the analytical methods or machine learning models employed, poor data quality can compromise decision-making, degrade system performance, and erode trust in data-driven processes [3, 20].

Traditional data quality frameworks were designed primarily for static or batch-oriented data environments, where quality assessment and cleaning could occur offline. In contrast, data streaming environments present unique challenges: data arrives continuously, often at high velocity, and cannot be easily revisited once processed. Ensuring data quality in such dynamic contexts requires adaptive, automated, and context-aware mechanisms capable of evaluating and improving data in motion. Addressing issues such as missing values, outliers, duplication, and inconsistency in real time is essential for maintaining reliable and actionable data streams [12, 18].

This paper provides a literature review of data quality processing in streaming environments. It revisits the historical origins of data quality, tracing its evolution from early database systems to contemporary real-time data ecosystems. It then explores how data quality can be represented and modeled, introducing formal frameworks, standards, and contextualization strategies that enable quality-aware data management. The paper proceeds to discuss the fundamental dimensions and metrics of data quality—such as accuracy, completeness, consistency, and timeliness—and how these can be quantified and applied to data streams. Building on this foundation, it reviews state-of-the-art methods for data quality processing, including fault detection, cleaning, and transformation techniques, with a focus on their applicability to sensor-based systems. Finally, the paper discusses current limitations, identifies research gaps, and highlights emerging trends

Authors' Contact Information: Aboubakr Benabbas, benabbas.aboubakr@uni-bamberg.de; Daniela Nicklas, daniela.nicklas@uni-bamberg.de, University of Bamberg, Bamberg, Bavaria, Germany.

such as adaptive and AI-driven quality processing frameworks designed to enhance usability and domain independence.

Through this structured review, the paper aims to provide researchers and practitioners with a holistic understanding of data quality challenges and solutions in streaming environments, fostering the development of more reliable, transparent, and intelligent real-time data processing systems.

2 Scope and Purpose of This Literature Review

This paper presents a focused literature review of data quality concepts, representations, and processing techniques in the context of continuous data streams and sensor-driven systems. Unlike a systematic survey, the purpose of this review is not to exhaustively cover all existing approaches or provide a formal classification of the field. Instead, it synthesizes key foundational and recent contributions that inform ongoing work on quality-aware stream processing. The review consolidates relevant background that could not be included in an accompanying research article due to space limitations and aims to support researchers and practitioners seeking an integrated understanding of data quality principles applicable to streaming environments. This review focuses on peer-reviewed research that explicitly addresses data quality concepts, metrics, or processing techniques in the context of continuous data streams or sensor-generated data. Seminal works predating modern streaming systems were included when they established terminology or introduced concepts still referenced in current approaches. Publications were identified through targeted searches in leading digital libraries (e.g., ACM, IEEE, Springer, and Elsevier), supplemented by backward citation tracing. We prioritized works that contributed generalizable models, widely cited frameworks, or domain-independent techniques, while papers addressing narrow, application-specific implementations without broader methodological implications were excluded.

3 Data Quality Origins

The concept of data quality in computer science emerged during the formative years of digital systems in the mid-20th century, when early computing environments required reliable mechanisms to ensure that data captured, stored, and transmitted could be trusted. At that time, the main emphasis was on detecting and preventing errors arising from manual data entry, physical media imperfections, or hardware malfunctions. Techniques such as error detection and correction [17] were developed to safeguard the integrity of information flowing through computational pipelines, laying the groundwork for more formal treatments of data quality.

A major shift occurred with the introduction of relational database systems in the 1970s, led by the foundational work of Edgar F. Codd [5]. The relational model not only transformed data storage and querying but also fostered new principles for organizing and maintaining structured datasets. Normalization procedures, in particular, became essential for reducing redundancy, enforcing logical constraints, and ensuring that datasets reflected coherent representations of real-world entities. As relational technologies matured, the need to integrate data across organizational units grew, especially with the rise of data warehousing in the 1980s. These developments drew attention to broader quality considerations, including dimensions such as consistency, completeness, and timeliness [13].

The onset of the big data era and the widespread adoption of machine learning in the 21st century further expanded the scope of data quality practices. The increased volume, variety, and velocity of data demanded more sophisticated capabilities—ranging from automated profiling and large-scale anomaly detection to advanced governance and monitoring frameworks [15]. These tools enabled organizations to address quality challenges in real-time settings and across diverse data modalities. Modern approaches now combine traditional quality dimensions with requirements related to

transparency, traceability, and ethical compliance, particularly as data-driven applications become more tightly integrated with operational and decision-making processes [6].

4 Data Quality Representation

Data Quality (DQ) representation refers to the principles and mechanisms used to articulate, organize, and convey information about the quality properties of data in a clear and structured manner. Rather than relying on implicit cues or scattered documentation, DQ representation brings together dimensions, metrics, and metadata into an integrated framework that enables systematic assessment, monitoring, and communication of data quality across different stakeholders and systems. The following points highlight central components involved in representing data quality.

(1) **DQ Models and Standards:**
Standardized frameworks such as ISO/IEC 25012 [6] provide widely accepted guidelines for describing data quality characteristics. These models define and categorize quality attributes and outline relationships between them, offering a common language that supports consistent interpretation and interoperability across heterogeneous data management environments.

(2) **Dimensions of Data Quality:**
A fundamental aspect of representation is specifying which quality dimensions are relevant for a particular dataset or application. Commonly emphasized dimensions include accuracy, completeness, timeliness, and consistency. Each dimension highlights a distinct facet of quality and is tied to concrete indicators or metrics that help quantify the degree to which the dataset satisfies corresponding requirements.

(3) **Quality-Aware Query Processing:**
Quality information can be embedded directly into query execution mechanisms. Systems may, for example, rank or filter records based on their associated quality scores, enabling queries to prioritize more reliable or up-to-date data. In simple cases, low-quality tuples can be excluded using threshold-based rules (e.g., `accuracy < 0.8`), while more sophisticated frameworks integrate quality metadata into query optimization strategies.

(4) **Context Integration:**
Data quality cannot be meaningfully represented without considering the context in which the data is used. Requirements vary widely across domains: a timeliness constraint that is essential for financial trading may be far more relaxed in healthcare reporting or long-term archival systems. Effective representation therefore integrates contextual parameters that clarify how quality should be interpreted relative to specific operational or analytical needs.

(5) **Adaptive Quality Representation:**
Because data usage conditions can change, some systems adopt adaptive mechanisms for representing quality. For example, *Adaptive Timeliness* may enforce tighter freshness requirements during high-risk operational periods, while *Adaptive Completeness* may allow missing fields in exploratory analytics but demand full coverage in regulatory reporting. These adaptive strategies enable flexible and context-aware quality assessment.

(6) **DQ Quantification:**
Quantitative metrics serve as measurable indicators of the quality dimensions. Accuracy can be computed through error rates, completeness through the proportion of populated fields, and timeliness through the age of data relative to validity windows. Representation frameworks often formalize these metrics so that quality evaluations can be performed consistently across datasets and over time.

(7) **DQ Metadata:**
Metadata is central to representing data quality because it captures descriptive information about datasets beyond their raw values. This may include lineage traces, reliability assessments of data sources, refresh schedules, or other provenance-related details. Such metadata helps analysts and systems understand both the origins and the reliability of the data they consume.

(8) **DQ Labeling:**
In environments such as streaming systems or sensor networks, quality indicators may be attached directly to individual data elements. Attributes such as accuracy, timeliness, or confidence scores can be encoded as annotations, enabling downstream applications to interpret quality on a per-record basis and make more informed processing decisions.

(9) **Error Propagation Estimation:**
Some approaches represent data quality by modeling how errors propagate through transformations, aggregations, or query operations. These models estimate how processes may introduce uncertainty, degrade accuracy, or obscure provenance, helping users understand how transformations affect the reliability of final outputs.

(10) **DQ Visualization:**
Representing data quality often involves generating dashboards, summaries, or structured reports that illustrate the current quality state of datasets. Such visualizations may track trends in completeness, highlight anomalies in timeliness, or show distributions of quality scores. These outputs support informed decision-making and help organizations identify areas requiring quality improvements.

Data quality representation is especially important in data streams and Data Stream Management Systems (DSMSs), where information arrives continuously and must be evaluated under real-time constraints. These systems rely on dynamic and context-aware representations of quality to handle high-volume, high-velocity data effectively. Clear and actionable quality representation enables organizations to detect problems promptly, observe quality evolution over time, communicate expectations to stakeholders, and align quality-related practices with broader organizational objectives.

5 Data Quality Dimensions

Data quality (DQ) is commonly understood as a multifaceted concept rather than a single property of data. It is described through a collection of *dimensions*—distinct evaluative categories that capture different aspects of data suitability. These dimensions are often grouped into intrinsic, representational, and contextual categories [20]. Intrinsic dimensions describe characteristics inherent to the data itself, such as accuracy or believability [14]. Representational dimensions capture qualities related to how data is presented and interpreted, including clarity and interpretability [19]. Contextual dimensions reflect the relationship between data and its intended use, covering properties such as completeness, timeliness, or relevance [3].

Different applications may assess DQ dimensions in various ways. Some dimensions are evaluated using objective, rule-based criteria, while others are graded along continuous scales. Wang and Strong [20] distinguish between *absolute* dimensions—those that represent inherent properties of the data—and *relative* dimensions, which depend on specific use cases or operational requirements. In real-world settings, the boundary between these categories is often fluid. For example, an application might relax accuracy expectations in situations where approximate values suffice for analysis, effectively converting an absolute requirement into a relative one [7].

The survey by Batini et al. [3] remains influential for its systematic categorization of DQ dimensions and its synthesis of definitions used across the literature. Their work emphasizes four core dimensions—accuracy, completeness, consistency, and timeliness—which form the backbone of most DQ frameworks. Although the exact descriptions differ across studies, these dimensions capture the essential qualities that determine data reliability and usefulness. With the emergence of large-scale, heterogeneous, and real-time data environments, additional dimensions such as provenance, availability, and security have become increasingly relevant [10]. When dealing with data produced by sensors, emphasis typically falls on intrinsic characteristics captured by the foundational four dimensions. The following paragraphs provide consolidated definitions of these dimensions as reported across the literature.

Accuracy: Accuracy captures how closely a recorded value reflects the true state of the observed phenomenon. It encompasses both measurement precision and the degree to which data corresponds to ground truth. Definitions in the literature [2, 16, 20] converge on the idea that accuracy is fundamental to generating dependable analyses and inferences.

Table 1. Core data quality dimensions and representative metrics for streaming data (adapted from Batini and Scannapieco [3] and related work).

Dimension	Streaming-oriented definition	Representative metrics
Accuracy	Degree to which a data value correctly describes the real-world phenomenon it represents, taking sensor noise and calibration into account.	Error deviation from reference, root mean squared error (RMSE), proportion of validated observations.
Completeness	Extent to which expected data points or attributes are present in a stream, considering packet loss, communication failures, or offline sensors.	Missing-value ratio, record availability ratio, attribute coverage.
Consistency	Conformance of data to physical, logical, and semantic constraints within and across streams in real time.	Number of rule violations, unit-conversion conflicts, cross-stream agreement ratio.
Timeliness	Degree to which data arrives within its validity window for a given real-time application.	Data age distribution, lateness ratio, freshness indicator.
Traceability	Ability to track the origin of observations and the sequence of transformations applied along the processing pipeline.	Percentage of tuples with complete provenance, lineage depth, provenance-completeness index.
Duplication	Extent to which redundant or repeated tuples appear due to retransmissions, buffering, or unstable connections.	Duplicate-tuple ratio, deduplication effectiveness, unique-record ratio.
Volatility	Frequency and magnitude with which measurements change over time, indicating stability or instability of a stream.	Update rate, temporal variance index, change-frequency ratio.

Completeness: Completeness measures the presence of required information within a dataset. It is frequently quantified as the proportion of available (non-null) data relative to what is expected [4, 11, 19]. Two common forms are:

- *Horizontal Completeness*: The degree to which full records or data points exist.
- *Vertical Completeness*: The degree to which attributes within a record are populated.

Both forms influence how effectively data can support downstream decision-making.

Consistency: Consistency evaluates whether data values obey the semantic and structural rules defined for the dataset [19]. It may involve:

- *Intra-relation constraints*: Rules governing attribute values within a single dataset.
- *Inter-relation constraints*: Requirements describing how data must align across multiple related datasets.

High consistency improves reliability and minimizes anomalies caused by integration or transformation errors.

Timeliness: Timeliness concerns whether data remains sufficiently current for its intended use. The term overlaps with related concepts such as *currency* and *volatility*, though definitions vary across sources. Some works define currency as the moment of data acquisition [4, 11], while Wang and Strong [20] associate volatility with how quickly data becomes outdated. Redman [16] suggests that data is timely if its validity is preserved relative to the passage of time. Timeliness is especially important in operational or real-time contexts.

Frequency: In some frameworks, frequency refers to how often data elements change or how frequently new observations arrive. This concept is related to volatility but emphasizes update patterns rather than staleness.

As emphasized by Wang and Strong [20] and Batini et al. [3], data quality must be understood as a multidimensional construct. Each dimension isolates a particular perspective for evaluating data, and different dimensions may require different metrics. Some dimensions can be assessed using a single measure, while others require several indicators to capture their nuances comprehensively. This multidimensional framework helps organizations develop targeted strategies for improving data quality and ensures that the data they rely upon is fit for diverse analytical, operational, and strategic purposes.

6 Data Quality Metrics

Data quality (DQ) metrics serve as the operational mechanisms used to quantify the various DQ dimensions. Each metric formalizes how strongly a dimension is expressed within a dataset or stream, typically by defining a measurable scale or function that captures the extent to which the corresponding DQ property is satisfied. As shown in Table 1, single dimensions can be assessed through multiple metrics, enabling more granular or specialized evaluations. Batini et al. [3] provide an extensive catalog of such measures, while Geisler et al. [8] group them into three main categories: content-based, query-based, and application-based metrics.

Content-based metrics focus on intrinsic characteristics of the data itself. These measures evaluate properties such as accuracy, completeness, or consistency directly from the values present in the dataset, without considering how the data is queried or used. Because they capture fundamental aspects of data integrity, they are often employed as baseline indicators of quality. In contrast, query-based metrics assess quality relative to specific information needs. Their evaluation depends

on the interaction between user queries and the dataset, emphasizing attributes like precision, recall, and the relevance of retrieved records. Application-based metrics are highly contextualized and determine whether the data supports the particular requirements of a designated application or operational workflow. Such metrics incorporate task-specific thresholds, relevance criteria, or performance expectations, complementing more general-purpose assessments.

Given our focus on integrating quality-aware processing into sensor data streams at the raw-data level, our study emphasizes content-based metrics. These metrics naturally align with the intrinsic properties of sensor observations. Query-based and application-based metrics, which depend heavily on user intent or task-specific constraints, fall outside the scope of this work. By augmenting streaming data with intrinsic quality information, we aim to enable real-time reasoning about the reliability and suitability of incoming sensor measurements.

DQ dimensions and their associated metrics thus create an abstraction layer that improves reusability and modularity. They allow users and systems to reason about quality in a structured way, apply different methods to evaluate the same dimension, and incorporate user-specific requirements—such as application-defined validity intervals for *Timeliness*—without altering the underlying data-processing logic.

7 Data Quality Processing

Teh et al. [18] conducted a systematic literature review following a structured methodology to identify and synthesize research on sensor-related data errors. Their study provides a comprehensive overview of the various types of errors that can occur in sensor outputs and categorizes existing approaches according to the error types they target, how these errors are quantified, and whether the techniques aim at correcting faulty values or improving overall data quality. This taxonomy helps clarify the landscape of sensor data quality management and highlights the methodological diversity present in the field.

In the area of data cleaning, Gill and Lee [9] introduced a hybrid technique that combines declarative rules with statistical modeling to perform distributed cleaning on environmental data streams. Their approach exploits the complementary strengths of structured rule-based logic and data-driven statistical inference, enabling effective handling of large, continuously generated datasets. This approach is valuable for real-time applications where errors must be addressed without interrupting system operation.

Outlier detection constitutes a major portion of the work on sensor fault identification. Ayadi et al. [1] provide an extensive survey of outlier detection methods designed for Wireless Sensor Networks (WSNs), comparing their operating principles, assumptions, and applicability across different scenarios. They also offer a decision-making framework to assist practitioners in choosing the most suitable approach for their specific context. However, despite their sophistication, many of these methods are tightly coupled to the characteristics of the target application, making them difficult to generalize or transfer to new settings without substantial adaptation.

Although techniques for cleaning and fault detection can be highly effective within the domains for which they are designed, their strong specialization limits cross-domain applicability. This specialization often requires significant domain knowledge to implement successfully, creating barriers for practitioners who lack deep expertise in data quality. As a result, many existing solutions remain challenging to adopt outside the specific environments for which they were originally developed.

To address the need for more flexible processing pipelines, Jeffery et al. [12] proposed the Extensible Sensor Stream Processing (ESP) framework, which offers a programmable architecture for performing real-time cleaning of sensor data as it arrives. The framework attempts to move beyond narrowly targeted methods by providing an extensible structure for composing and deploying

cleaning functions. Nevertheless, ESP still faces limitations in the breadth of available techniques and does not fully support the diversity of data quality requirements encountered across different domains.

Data quality processing more broadly includes a range of procedures and algorithms designed to enhance the reliability, accuracy, and usability of data. These activities typically involve detecting anomalies, correcting errors, transforming data into standardized forms, and validating values against predefined rules or expectations. A robust processing framework must be capable of handling recurring issues including missing entries, noise, redundant values, and structural inconsistencies.

Recent progress in machine learning and artificial intelligence has expanded the capabilities of data quality improvement. Deep learning models have shown considerable promise for adaptive anomaly detection by learning patterns in data streams and automatically identifying deviations. Federated learning approaches, which enable model training across distributed data sources without sharing raw data, have also emerged as tools for privacy-preserving quality processing in distributed environments.

Despite these advancements, a persistent challenge lies in making data quality processing accessible to non-expert users. Many existing tools require substantial technical expertise, limiting broader adoption. Developing intuitive interfaces, automating routine tasks, and integrating visual guidance for selecting or configuring quality-improvement methods could significantly enhance usability and foster wider deployment across diverse industry sectors.

8 Discussion on Data Quality Processing

DQ processing approaches often lack flexibility and are tailored to specific use cases, making them unsuitable for broader applications. Additionally, for users who are not experts in data quality, integrating quality assessment into data processing pipelines remains a significant challenge.

A more versatile data quality solution, built upon Data Quality Patterns, addresses these challenges by narrowing the focus of data quality tasks while ensuring simplicity and adaptability. Model-based approaches that utilize *Data Quality Patterns* and semantics to define sensors and their various properties offer a more accessible and comprehensive pathway to quality-aware data processing.

9 Conclusion

The growing reliance on real-time analytics, sensor networks, and Internet of Things (IoT) infrastructures has intensified the need for reliable and high-quality data in streaming environments. This literature review has examined the conceptual foundations, dimensions, metrics, and processing techniques that collectively define data quality (DQ) in such dynamic contexts. Beginning with the historical evolution of DQ practices—from early database systems to modern big data and stream processing architectures—the paper has illustrated how traditional quality concepts are being reinterpreted to suit the velocity, volume, and variability of streaming data.

A detailed review of DQ representation and measurement methods revealed that accuracy, completeness, consistency, and timeliness remain the most fundamental dimensions, yet their evaluation in continuous data streams requires adaptive and context-aware approaches. The discussion of DQ processing techniques highlighted advances in fault detection, data cleaning, and anomaly management, particularly in sensor-driven systems. However, most existing solutions remain domain-specific, requiring expert intervention and lacking the flexibility needed for cross-domain applicability.

Future progress in data quality processing will depend on developing scalable, autonomous, and interpretable solutions capable of operating across diverse streaming environments. Promising directions include the integration of AI-driven methods for automatic quality assessment, pattern-based frameworks for reusable quality strategies, and semantic models that enhance interoperability.

Equally important is the design of user-friendly interfaces and tools that empower non-experts to monitor and manage data quality in real time.

In summary, ensuring high data quality in streaming environments is not merely a technical challenge but a prerequisite for trustworthy, data-driven decision-making. By consolidating existing knowledge and identifying open research gaps, this literature review provides a foundation for future work toward adaptive, transparent, and intelligent data quality management in real-time systems.

References

[1] Aya Ayadi, Oussama Ghorbel, Abdulfattah M. Obeid, and Mohamed Abid. 2017. Outlier detection approaches for wireless sensor networks: A survey. *Computer Networks* 129 (2017), 319–333.

[2] Donald P. Ballou and Harold L. Pazer. 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science* 31 (1985), 150–162. https://api.semanticscholar.org/CorpusID:62126224

[3] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–52.

[4] Matthew Bovee, Rajendra P Srivastava, and Brenda Mak. 2003. A conceptual framework and belief-function approach to assessing overall information quality. *International journal of intelligent systems* 18, 1 (2003), 51–74.

[5] Edgar F Codd. 1970. A relational model of data for large shared data banks. *Commun. ACM* 13, 6 (1970), 377–387.

[6] International Organization for Standardization. 2008. ISO/IEC 25012: Data Quality Model. https://www.iso.org/standard/35736.html.

[7] Mouzhi Ge and Markus Helfert. 2004. A framework for measuring data quality. In *Proceedings of the 2004 International Conference on Information Quality (ICIQ)*. 105–116.

[8] Sandra Geisler, Christoph Quix, Sven Weber, and Matthias Jarke. 2016. Ontology-based data quality management for data streams. *Journal of Data and Information Quality (JDIQ)* 7, 4 (2016), 1–34.

[9] Saul Gill and Brian Lee. 2015. A Framework for Distributed Cleaning of Data Streams. *Procedia Computer Science* 52 (2015), 1186–1191. doi:10.1016/j.procs.2015.05.156 The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015).

[10] Alejandra U Gonzalez, Joaquín Ordieres-Meré, and Clara AC Guevara. 2015. Data quality assessment: The case of energy performance certificates. *Energy and Buildings* 93 (2015), 167–174.

[11] Matthias Jarke, Maurizio Lenzerini, Yannis Vassiliou, and Panos Vassiliadis. 2002. *Fundamentals of data warehouses*. Springer Science & Business Media.

[12] Shawn R. Jeffery, Gustavo Alonso, Michael J. Franklin, Wei Hong, and Jennifer Widom. 2006. Declarative Support for Sensor Data Cleaning. In *Proceedings of the 4th International Conference on Pervasive Computing* (Dublin, Ireland) *(PERVASIVE'06)*. Springer-Verlag, Berlin, Heidelberg, 83–100. doi:10.1007/11748625_6

[13] Ralph Kimball and Margy Ross. 1996. *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. John Wiley & Sons.

[14] Leo L Pipino, Yang W Lee, and Richard Y Wang. 2002. Data quality assessment. *Commun. ACM* 45, 4 (2002), 211–218.

[15] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc.

[16] Thomas C Redman. 1997. *Data quality for the information age*. Artech House, Inc.

[17] Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.

[18] Hui Yie Teh, Andreas W Kempa-Liehr, and Kevin I-Kai Wang. 2020. Sensor data quality: A systematic review. *Journal of Big Data* 7, 1 (2020), 1–49.

[19] Yair Wand and Richard Y Wang. 1996. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* 39, 11 (1996), 86–95.

[20] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33. doi:10.1080/07421222.1996.11518099