Data Quality Awareness: A Journey from Traditional Data Management to Data Science Systems

Sijie Dong
Université Paris Cité
F-75006 Paris, France
sijie.dong@etu.u-paris.fr

Soror Sahri
Université Paris Cité
F-75006 Paris, France
soror.sahri@u-paris.fr

Themis Palpanas
Université Paris Cité
F-75006 Paris, France
themis@mi.parisdescartes.fr

ABSTRACT

In this paper, we present a comprehensive review of the evolution of data quality awareness from traditional data management systems to modern data-driven ML systems, which are integral to data science. We synthesize the existing literature, highlighting the quality challenges and techniques that have evolved from traditional data management to data science, including Big Data and ML fields. As data science systems support a wide range of activities, our focus in this paper lies specifically in the analytics aspect driven by ML. We use the cause-and-effect connection between the quality challenges of ML and those of Big Data to allow a more thorough understanding of emerging DQ challenges and the related quality awareness techniques in data science systems. To the best of our knowledge, our paper is the first to provide a review of DQ awareness spanning traditional and emergent data science systems.

1 INTRODUCTION

The rapid evolution of data science (DS) systems has fundamentally shifted how data is processed and analyzed. Driven by the convergence of Big Data—distinguished by its scale, speed, and complexity—and Machine Learning (ML), these systems operate as end-to-end pipelines where models are learned from data to automate decision-making. These systems, which handle massive datasets and complex models, rely on data quality (DQ) to ensure reliable performance. In this context, DQ awareness—which encompasses defining quality goals via dimensions and metrics, systematically assessing their state to understand their impact, and dynamically adapting processes to improve data quality—becomes critical.

While recent research admits the value of traditional DQ frameworks, directly applying them to DS systems is difficult. This requires moving from traditional quality dimensions, like accuracy and completeness, to the new quality criteria for modern DS systems.

Motivated by the growing importance of DQ awareness in modern DS systems, this paper traces the transition across three evolving paradigms: from *Traditional Data Management* (schema-driven systems designed for structured, transactional processing across both centralized and traditional distributed environments), to *Big Data* environments (distributed ecosystems engineered for extreme Volume, Velocity, and Variety), and ultimately to *DS Systems* (analytical architectures that integrate Big Data infrastructure with ML pipelines for predictive decision-making). This journey highlights how increasing data scale, complexity, and usage progressively reshape DQ concerns, where DQ awareness perspectives are determined by the dominant quality concerns of each system. In traditional data management systems, the singular focus on meeting the needs of immediate users often oversimplified the complexity of DQ issues. In a Big Data context, DQ issues are multiplying due to (i) very large datasets that go beyond the scale assumed by traditional DQ methods, (ii) different types of data that existing quality dimensions and evaluation methods cannot handle, and (iii) real-time, evolving data that may alter data characteristics and derived insights. These DQ issues result in two main challenge areas, as identified by Saha and Srivastava [101]: (i) discovering data quality semantics and performing data repairing in Big Data environments; and (ii) managing the trade-off between accuracy and efficiency across different computing models. This emphasizes the increased complexity of big DQ dimensions, leading to new DQ techniques supported by Big Data platforms.

The ability to extract value from Big Data largely relies on data analytics, where ML plays a central role alongside business intelligence, data visualization, and statistical analysis. In this survey paper, we focus on ML, which is considered as the core of the Big Data revolution [19, 92]. This synergy between Big Data and ML allows DS to transform Big Data into insights, decisions, and predictions. However, the complex configurations of the datasets used in DS systems and the diverse backgrounds of ML practitioners introduce new DQ issues, such as the propagation of errors and biases across ML pipelines, and pose challenges in managing their impact on model training, evaluation, and downstream decisions.

Despite the rich literature on DQ in both Big Data [18, 19], and ML [95], and studies on their integration [77, 92], there is still a gap in exploring the interplay between DQ issues and the unique challenges arising from linking ML and Big Data within data science systems. Unlike prior surveys that treat Big Data and ML quality in isolation, the explicit goal of our study is to map the compounding cause-and-effect relationships at their intersection. To this end,

this paper makes the following contributions: (1) We take readers on a journey through the evolution of DQ awareness from traditional systems to data science. We focus on ML pipelines as a key component of DS systems and emphasize the critical heritage from Big Data and traditional data management, including quality awareness techniques, for navigating emerging DQ challenges in ML pipelines. (2) We highlight emerging DQ challenges and new opportunities, based on the cause-and-effect intersection between Big Data and ML challenges, as depicted in Figure 2. (3) We provide an extension of existing classifications of DQ dimensions and techniques for ML pipelines (Table 1), linking them to specific pipeline stages. (4) We identify key research gaps and future opportunities, including the emerging dual role of LLMs in data quality.

The structure of this survey mirrors the cumulative rather than substitutive evolution of DQ challenges. We organize the remaining sections to highlight this "inheritance" relationship: Section 3 discusses traditional systems, where DQ mainly concerns syntactic correctness and transactional reliability; Section 4 shows how Big Data inherits these requirements while adding challenges from volume, variety, and velocity; and Section 5 examines how ML pipelines incorporate both the syntactic strictures of traditional data and the issues of Big Data characteristics, while adding novel challenges in semantic fitness.

2 FOUNDATIONAL CONCEPTS OF DQ

This section introduces the main systems and DQ concepts used throughout the paper. By traditional data management systems, we refer to schema-driven, transactional environments—including relational databases and early distributed databases—designed to manage structured data with predefined schemas and ACID guarantees [6]. By Data Science (DS) systems, we refer to analytical architectures that couple Big Data infrastructure with ML pipelines to derive predictive insights from large, heterogeneous datasets [19, 92].

In this paper, we define *Data Quality Awareness* [6, 19] as the ability of a data system to explicitly define quality goals via dimensions and metrics, systematically monitoring these states (Assessment), and dynamically adapting processes to mitigate issues (Improvement).

Data quality refers to the extent to which data is suitable for a specific task, emphasizing its actual utility and relevance to the context. Data quality is often measured by its "fitness for use" in supporting operations, decision-making, and planning [97]. To evaluate and ensure DQ, two main components are used: DQ Dimensions and DQ Metrics. Formally, a DQ Dimension is defined as a quality attribute d belonging to a set of dimensions D, where each $d \in D$ represents a specific aspect of data "fitness for use" to be evaluated [8, 122]. Rather than acting merely as descriptive labels, dimensions serve as the conceptual variables that structure and categorize data quality requirements, such as intrinsic accuracy, contextual relevance, or timeliness [6, 122]. To operationalize these qualitative constructs, DQ Metrics provide specific measuring procedures. As established in the literature [6, 90], a metric is a measurement function $m_d : X \rightarrow V$ associated with a dimension d, where X denotes a dataset (or a subset of data) and V represents the metric value domain (e.g., numerical scores, boolean values, categorical levels, or vectors) indicating the degree to which the dimension is satisfied. This mathematical operationalization enables objective, reproducible, and comparable evaluations of data quality across datasets, as well as continuous quality monitoring.

While dimensions and metrics provide the quantitative basis, realizing *DQ Awareness* requires a structured execution framework. Continuous improvement and assessment efforts are then essential to improve source quality and meet or exceed user expectations [6, 18]. *Assessment* is the systematic evaluation of data against quality dimensions to identify anomalies, errors, and their causes; and *improvement* uses these findings to enhance data through data-driven updates and process-driven changes [111].

3 DQ AWARENESS IN TRADITIONAL DATA MANAGEMENT

Traditional data management systems use various techniques to support DQ awareness, primarily from the perspective of data producers or sources, focusing on modeling and measuring data quality at its origin. However, addressing quality challenges also requires considering data consumers by modeling user requirements and ensuring data meets their expectations. In this paper, perspectives distinguish whether DQ is primarily evaluated from the viewpoint of the data itself, its operational use, or the broader actors affected by the system, depending on the dominant quality concerns of each system.

3.1 Quality Awareness from Data Perspective

The data perspective for quality awareness focuses on data characterization. *Dimensions* are used to describe various data properties [109], commonly categorized as intrinsic, accessibility, contextual, and representational quality [123]. To understand data characteristics and assess quality, data profiling techniques are applied at different levels: (i) attribute/tuple level (e.g. missing values, domain violations); (ii) single relation (e.g., business rule violations); (iii) multiple relations (e.g., referential integrity violations); and (iv) multiple sources (e.g., inconsistent duplicates). Profiling includes defining new quality rules (e.g., Functional Dependencies) and identifying quality issues (e.g., inconsistent data) [1, 22, 60, 85, 89].

In [11], profiling involves associating quality contracts with data sources, where a set of contracts forms a quality

profile. These profiles are used to negotiate quality requirements with data source wrappers, ensuring the query framework selects sources based on quality characteristics. Conditional DQ profiling, as proposed in [129], associates attribute quality to conditions in user queries. Although guided by user-defined conditions, the technique remains data-oriented because it primarily evaluates and characterizes data quality properties.

3.2 Quality Awareness from User perspective

Among techniques, query processing is the most important component in traditional systems, embedding quality measures directly into their processes. Quality-aware query processing techniques ensure that data meet user requirements and preferences, mainly through *query language extensions* and *adaptive query processing*.

3.2.1 Query language extensions They integrate DQ considerations into query processing by allowing the expression of quality metrics and constraints in a simple, declarative manner. In [129], an SQL extension was proposed to model user preferences through hierarchical prioritization. Additionally, the DQ-Aware Query System (DQAQS) framework was introduced to improve user satisfaction by considering these preferences in query results.

In [11], the quality-extended query language XQual was introduced for selecting dynamic sources using a negotiation strategy. It extends SQL with a *Qwith* operator to specify quality constraints via contracts and profiles. This approach was later applied to skyline queries with graph-based nearest neighbor search [12], and more recently to enforce DQ thresholds in ML training data [26].

3.2.2 Adaptive query processing It allows for dynamically handling DQ issues and provides more reliable query results. It involves creating multiple query execution plans that can be switched based on the quality of the data. In [84, 86], a distributed query planning algorithm discards low-quality sources, ordering plans by completeness. System P [99] uses a completeness-driven approach where peers rank local plans by potential result size and prune based on budget thresholds, balancing completeness and cost. Similarly, [129] incorporates planning and optimization to assess each plan's utility based on expected data quality, ensuring efficient handling of diverse data sources.

4 DQ AWARENESS IN BIG DATA

As data-centric technologies advanced beyond the traditional data management paradigms discussed in Section 3, the scope of data quality expanded. Big Data ecosystems do not abandon traditional quality perspectives; rather, they inherit foundational dimensions and adapt them to new operational scales. Following, we present the dimensions related to new DQ challenges and the impact of Big Data characteristics (BDCs) on quality dimensions.

4.1 Big Data Quality Dimensions

Traditional quality dimensions (e.g., accuracy, completeness, consistency, etc.) are not sufficient to assess Big Data [7]. In the context of Big Data, the meaning and calculation methods of these traditional dimensions undergo significant changes. This section presents common quality dimensions relevant to Big Data. We classify them according to the existing literature [18, 19, 41, 45], into source-specific and user perspective dimensions. This classification extends the data and user perspectives introduced in Section 3 into the context of Big Data to address its inherent complexity.

4.1.1 Source perspective dimensions The UNECE (United Nations Economic Commission for Europe) classification identified three main types of data sources: process-mediated, machine-generated, and human-sourced [45]. Big DQ dimensions from the source perspective are categorized accordingly:

- *Process-mediated* data sources, typically relational databases with structured data (e.g., customer records), face quality issues such as incorrect values, duplicates, and incompleteness. Related dimensions include consistency, accuracy, and freshness.
- *Machine-generated* sources, using sensors and machines to record real-world events, produce well-structured data. Quality issues often stem from measurement environments (e.g., machine noise, environmental effects). Key dimensions are accuracy, completeness, consistency, trustworthiness, and freshness.
- *Human-sourced* information sources, such as social networks, store human experiences like photos, audio, and videos. In addition to common quality dimensions, ambiguity in short text presents a unique challenge.

4.1.2 User perspective dimensions To better understand Big Data applications, quality dimensions are defined from the user perspective, based on two main assessment approaches. The *effective assessment* focuses on evaluating dimensions that influence user interaction with data. Key dimensions include reliability, availability, usability, relevance, and presentation quality [18], reflecting ease of access, usefulness, trustworthiness, alignment with expectations, and overall user satisfaction. The *context-dependent assessment*, highlighted in [4], emphasizes adapting dimensions based on the specific context of data assessment, including the data source, type, and intended application. In [79], the adaptive quality model adds user requirements related to execution time and performance, introducing the confidence dimension to address accuracy and trustworthiness. The importance of context is further demonstrated in specific domains [42, 105]. Social media platforms prioritize timeliness and accuracy for sentiment analysis,

while online news platforms value credibility to ensure trustworthy reporting [37, 38, 115].

4.2 Impact of Big Data Characteristics on DQ

Previous work characterizes big DQ by (i) exploring the link between Big Data traits and quality dimensions, (ii) identifying specific dimensions, and (iii) analyzing their evaluation. However, a gap remains in connecting BDCs to quality dimensions and understanding their impact on Big Data applications. This stems from the distinction between data quality and the value of insights derived from it, despite their correlation [2]. Moreover, DQ affects the whole Big Data pipeline, including acquisition, analysis, and interpretation [100]. As a result, assessing the overall quality of Big Data for a specific application remains a major challenge. To address this gap, prior work investigated the correlation between BDCs and DQ in specific domains. In the financial sector, for example, [121] found that data variety, due to multiple sources, has more impact on dimensions like accuracy, consistency, security, timeliness, and completeness. Velocity was also linked to timeliness, as faster data generation and processing support timely use.

[47] studied the effect of data volume on DQ dimensions influencing the effectiveness and adoption of Big Data analytics in business contexts. The study focused on less common dimensions: data diagnosticity (value of insights), accessibility (ease of access), security (risks in aggregation and analysis), and task complexity (data processing difficulty). Using the theory of valence from economics and psychology, the authors analyzed the positive and negative roles of these dimensions to guide business data practices.

The findings show that contextual dimensions, particularly timeliness and accessibility, are most closely linked to BDCs in user applications. However, existing studies have limitations. Dimensions beyond security and accessibility may also impact Big Data analytics, yet the influence of variety and velocity on these dimensions remains unclear. Most research focuses on the financial sector, limiting broader applicability. This narrow scope, along with limited user validation, hinders generalization across diverse Big Data contexts. Large volumes amplify security risks, and frameworks like Hadoop add further challenges in distributed environments [13]. Despite these limitations, the findings are useful for understanding the relationship between BDCs and DQ dimensions in user applications. Based on this, we summarize the impact of BDCs on DQ dimensions in Figure 1. The diagram illustrates the causal links we identified: BDCs (left) act as root causes impacting DQ dimensions (right). For instance, an increase in 'Variety' (a BDC) has a documented negative impact on 'Accuracy' and 'Consistency' (DQ dimensions) because integrating heterogeneous sources often leads to conflicts. Conversely, 'Velocity' (a BDC) positively correlates with 'Timeliness' (a DQ dimension), enabling real-time analytics.

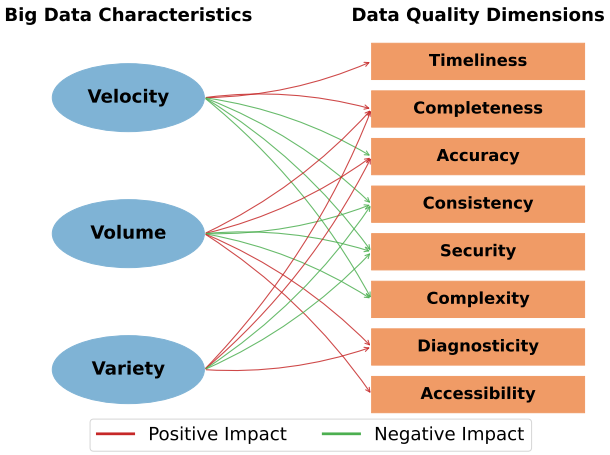

Figure 1. The impact of the BDCs on DQ dimensions.

4.3 Quality Awareness Techniques for Big Data

This section presents quality awareness techniques in the Big Data context, aligned with the V's characteristics, focusing on sampling, parallel processing, and incremental techniques.

4.3.1 Techniques for Managing Data Volume

Parallel Computing. The sheer volume of Big Data makes traditional, single-node DQ assessment computationally infeasible. Therefore, new quality awareness techniques emerged that rely on distributed frameworks such as Apache Hadoop and Apache Spark. These frameworks enable DQ tasks, such as large-scale profiling and assessment, to be parallelized. This parallel processing capability is the foundation for modern, large-scale DQ libraries. Several studies assess performance and scalability using these frameworks. Their findings show that Spark significantly improves computational efficiency for very large data volumes [19, 25, 103]. For instance, Deequ [103], built on Apache Spark, is designed specifically to automate data quality verification on terabyte-scale datasets.

Sampling and Sketch Techniques. *Sampling* reduces the time required for DQ assessment by approximating results. Commonly used methods include simple random, systematic, stratified, cluster, and reservoir sampling [14, 55, 75, 108, 119]. These techniques help determine sample sizes and select effective subsets to evaluate quality metadata. For instance, [19] shows that sampling improves accuracy under time constraints, while [75] demonstrates that systematic sampling better supports accuracy and completeness, and simple random sampling is better for timeliness. [113] introduces bootstrap sampling to profile datasets and select suitable metrics.The subsequent work in [114] applies the Bag of Little Bootstrap (BLB) to improve efficiency in processing unstructured data.

Sampling heterogeneous data requires preparation to ensure it is suitable for quality assessment. [114] proposed

preparing unstructured data (e.g., text, images, videos) using techniques like text mining and feature extraction to identify relevant information. For remote sensing images, [124] introduced a multi-level non-uniform spatial sampling method. Sampling is often integrated into Big Data frameworks, with approaches like block-based sampling using MapReduce [54].

Sketch Techniques, compared to sampling techniques, also greatly reduce the size of an input dataset. Unlike sampling, they maintain these properties more reliably [28]. They are fast, parallelizable, and offer high approximation accuracy [27, 32]. Sketches are commonly used for approximate analytical queries in systems like Pig, Hive, and Spark SQL, and are especially useful for real-time data stream processing [72].

4.3.2 Techniques for Managing Data Variety

Schema Alignment and Entity Resolution. To handle the high variety of Big Data—such as integrating data from IoT devices, social media, and enterprise databases—systems must resolve structural and semantic inconsistencies [34, 67]. Schema alignment techniques utilize metadata to match and unify the distinct structures of heterogeneous sources [52]. Additionally, entity resolution and transformation techniques perform data mapping, normalization, and deduplication to ensure accuracy and consistency across different formats [16]. When integrating overlapping but conflicting data, truth discovery and source dependence evaluation algorithms are applied to identify the correct values, embedding these conflict-resolution mechanisms into modern data-science pipelines [33, 34, 39].

4.3.3 Techniques for Managing Data Velocity

Incremental and Continuous Profiling Algorithms. Profiling methods should efficiently process data growth, without reprocessing entire datasets, and quality metrics should be updated continuously. To support this, [1] proposed incremental and continuous profiling: the former updates metrics based on periodic changes, while the latter processes data as it arrives. Metrics can be refreshed on-demand, periodically, or event-driven [19, 102]. Interactive profiling, including online profiling [85], improves user satisfaction by incorporating quality preferences and displaying intermediate results. An interface [131] further supports user-driven data cleaning, error detection, and transformation.

5 QUALITY AWARENESS IN ML PIPELINES

ML is a data-driven approach in which models are learned from data rather than defined through explicit rules or logic. As a result, it is particularly sensitive to data quality issues (e.g., bias, drift), making ML pipelines the primary focus of DQ awareness in modern DS systems. To effectively manage these sensitivities, we must first establish a comprehensive framework of DQ dimensions tailored to the ML pipeline.

5.1 ML-based quality dimensions

DQ dimensions tailored to ML pipelines must cover the entire lifecycle, from data preparation to model training and monitoring. To structure our analysis, we synthesize the existing literature [31, 82, 87, 93, 95, 138] into five main categories of quality dimensions. These categories consolidate and extend previous frameworks to explicitly address the unique challenges of ML systems.

This classification adopts a socio-technical lens [92, 95], as DQ issues in ML pipelines result from interactions between technical components and human actors, the *data*, *model*, and *process* dimensions cover the core technical components and their operational lifecycle. Externally, the *use-case* and *stakeholder* dimensions form two mutually exclusive pillars: the former defines the objective operational boundaries of the application, while the latter corresponds to the socio-ethical aspects of the human actors involved.

Table 1 presents this classification as a roadmap for our discussion. It maps these five dimension categories (columns) to the specific ML pipeline stages (rows) where they are most critical, and links them to the corresponding quality-aware techniques that will be detailed in Section 5.2. We now review these five categories in detail.

5.1.1 Data-based Dimensions The data considered here is used as input for each ML component. [112] distinguishes between training data, testing data, and serving data. The training data refers to the dataset used to train the model during development, the testing data is used for model evaluation, while the serving data refers to the dataset used during deployment for real-time predictions.

In addition to conventional dimensions, ML introduces specific dimensions that are critical to ensuring the integrity and fairness of the model's performance. Train/Test Independence is a critical data quality dimension that ensures strict separation between datasets. Violations lead to data leakage, where test-set information biases the training process, resulting in artificially inflated performance and poor generalization in production.

5.1.2 Model-based Dimensions Model-based dimensions assess the quality of the trained model itself. These include *performance* (e.g., accuracy, precision) [62, 81], *robustness* (the ability to handle adversarial or noisy inputs) [9], *scalability* (performance as data/user load increases) [9, 29], and *model complexity* (relates to explainability) [5]. These dimensions help ensure that the model operates effectively across various scenarios and environments. In addition to these conventional metrics, *Fairness* and *Explainability* are increasingly important for building

models that are not only technically sound but also responsible and trustworthy. *Fairness* ensures that the model's decisions are unbiased and do not systematically disadvantage certain groups. Addressing fairness allows for preventing discriminatory outcomes and promoting ethical AI use. *Explainability* refers to the model's ability to provide clear and understandable reasons for its decisions, which helps build user trust and promote transparency.

5.1.3 Process-based Dimensions Effective process management [88, 116] ensures that the system remains reliable, efficient, and secure throughout its lifecycle. Key dimensions, as highlighted by [82] under the system facet, include recoverability, portability, efficiency, transparency, traceability, cost, accessibility, ease of manipulation, and security. These dimensions are important for ensuring the robustness, adaptability, and security of machine learning systems, particularly as they scale or evolve.

5.1.4 Use Case and Context-based Dimensions Modeling quality requirements in ML pipelines based on specific use cases and application contexts is presented in [110, 120] to emphasize the importance of tailoring quality dimensions to the particular needs of ML applications. For instance, healthcare applications require stringent data privacy measures and high model accuracy. However, if the data used in these applications lacks contextual relevance, such as using a generic dataset for specialized medical diagnoses, this can lead to significant performance issues and may compromise patient safety. Addressing these issues helps ensure that ML models are effectively aligned with their intended applications, taking into account dimensions such as *value*, *contextual relevance*, and *use case specificity*.

5.1.5 Stakeholders-based Dimensions Data quality challenges related to stakeholders in ML pipelines are mainly presented in [92, 95]. The important role of ML practitioners, including data engineers, data scientists, and domain experts, should be emphasized, as stakeholder concerns extend beyond operational users to actors affected by model decisions and their ethical, legal, and societal implications, to ensure both effectiveness and ethical responsibility in ML pipelines. This allows the data to align with specific use case requirements, mitigating the risk of using data in ways that may be ethically misaligned with the original intent of data curators. *Ethical alignment* is then a key quality dimension allowing data usage to adhere to the ethical standards set by the data curators and comply with legal and social norms. *Transparency* is another key dimension to build trust and facilitate collaboration across diverse users, including developers, business analysts, and policymakers.

5.2 Quality Awareness Techniques for ML Pipelines

In order to align with the ML lifecycle, we categorize the algorithmic and statistical techniques based on the core pipeline tasks they fulfill. Each subsection introduces a task category, while the corresponding techniques and methods used to address it are discussed within the subsection.

5.2.1 Data Validation, Cleaning, and Imputation. *Data Validation* techniques verify that data conforms to certain standards [51] and detect anomalies before they impact model performance. One fundamental technique is *descriptive statistics*, which involves calculating metrics such as mean, median, variance, and standard deviation to summarize data distribution characteristics [57]. Recent advancements in this area include robust statistical methods that improve the detection of outliers and anomalies in large-scale datasets [107]. Schema validation ensures that the data adheres to a predefined schema, checking for consistency in data types, ranges, and formats [3], and recent schema evolution techniques allow adaptation to changes in data formats over time [9]. *Anomaly detection* techniques like Isolation Forests and One-Class SVM [104], identify data points that deviate from the norm. Deep learning-based approaches, such as autoencoders and LSTMs, enhance detection in high-dimensional data [20]. Automated validation tools are increasingly used in large-scale ML systems. Deequ [103], for example, is a Spark-based tool that supports user-defined constraints and offers an API for incremental validation. It also applies ML to automate tasks like outlier detection. Data Linter [58] suggests cleaning actions by analyzing schemas and distributions. Google's TFX [94] includes schema checks and anomaly detection throughout training and deployment.

Data cleaning techniques address a wide array of challenges identified during validation, such as noise, outliers, and duplicate records. While the broader field of data cleaning [69, 73, 98] covers a vast range of topics, including integrity constraint violations, semantic inconsistencies, typographical errors, unit mismatches, and value normalization, our focus here is on techniques most critical for ML accuracy. Cleaning is most effective when focused on improving model accuracy and making training more robust to noise. Indeed, data noise is considered adversarial when it contains malicious poisoning. In such cases, cleaning involves data sanitization to remove malicious inputs, along with outlier detection [17] like Isolation Forests and deduplication [80, 127] to eliminate anomalies and duplicate records, strengthening model reliability. *Data imputation* enhances dataset completeness and consistency by estimating missing values. Techniques range from traditional methods like MICE [118] and iterative k-NN [134], to deep learning approaches leveraging GANs [130] and latent-variable autoencoders [78] for synthetic generation. Recently, adaptive frameworks have emerged to automate model and hyperparameter selection, further optimizing imputation quality [61].

Table 1. Classification of Quality Dimensions and Techniques Across ML Pipeline Stages

ML Stages	Techniques	Quality Dimensions in the ML Pipeline				
		Data-based Dimensions	**Model-based Dimensions**	**Process-based Dimensions**	**Use Case/ Context-based Dimensions**	**Stakeholder-based Dimensions**
Data Preparation	*Sampling*	Representativeness, Balancedness			Contextual Relevance	
	Data Validation	Correctness, Completeness				
	Data Cleaning	Correctness, Completeness				
	Data Imputation	Correctness, Completeness, Intra-Consistency				
	Labeling	Correctness, Absence of Bias				
Model Training	*Feature Engineering*	Representativeness, Train/Test Independence, Balancedness	Performance, Model Complexity		Use Case Specificity	
	Fairness and Bias Detection	Absence of Bias	Fairness			Ethical Alignment
	Data Augmentation	Balancedness	Fairness		Use Case Specificity	
Model Validation & Monitoring	*Data Drift Detection*	Currentness	Robustness	Real-time Performance		
	Monitoring Model		Performance, Scalability, Trust	Reliability, Documentation Quality, Auditability, Reproducibility	Trust	Transparency

5.2.2 Data Augmentation. Data augmentation techniques allow for improving the robustness of models, particularly for small or imbalanced datasets. They expand training data to increase their diversity, mitigate overfitting, and maintain model performance. For imbalanced datasets, techniques such as SMOTE [21] and Mixup [133] are commonly used. Deep learning-based techniques like Variational Autoencoders (VAEs) [68], and GAN-based data augmentation [48] models such as StyleGAN [66] have further advanced the quality and diversity of synthetic data. However, GANs are limited in generating data that is significantly different from the original. AutoAugment [30] addresses this by automating augmentation strategies and improving the balance between realistic and diverse data.

5.2.3 Data Labeling. Supervised learning relies on high-quality labeled data. Advanced techniques and tools have been developed to automate labeling and validation, improving efficiency while ensuring consistency and accuracy. Snorkel [96] uses weak supervision to generate training data from heuristic rules and domain knowledge. Prodigy [40], developed by Explosion AI, applies active learning to present the most informative samples to annotators, streamlining the labeling process.

5.2.4 Fairness and Bias Detection. Addressing fairness involves detecting and mitigating biases in datasets, algorithms, and model outputs. Fairness metrics assess model predictions across several demographic groups, and different stages of ML pipelines (before/in/post processing) [87]. Common metrics include individual fairness (treating similar individuals similarly) and group fairness (across or within groups) [35], with key metrics such as disparate impact [44], equal opportunity [53], and demographic parity [35].

Bias mitigation techniques like reweighting [64] and adversarial debiasing [132] adjust the model's learning process. Reweighting compensates for imbalances via sample weights, while adversarial debiasing introduces an adversarial component to promote fairer predictions. Tools like IBM's Fairness 360 [59] and Google's What-If [49] support bias testing and mitigation, helping practitioners address issues from data drift or evolving societal norms.

5.2.5 ML monitoring. After deployment, quality efforts shift to monitoring serving data to ensure it remains contextually aligned with training data. Various techniques address training-serving skew, with some studies [93] detecting skew in predefined variables. However, selecting the right variables and thresholds remains challenging. One solution is to base these thresholds on the expected distribution of relevant features (e.g., the usage frequency of specific attributes or demographics), allowing for more targeted monitoring.

Automated systems [36, 103] enhance monitoring by flagging anomalies, data integrity issues, or drops in metrics like accuracy, precision, and recall. When combined with performance-tracking tools such as control charts from statistical process control [83], these systems effectively track model performance over time.

5.2.6 Data drift detection and adaptation. In production environments, data distributions may shift over time, a phenomenon known as drift, which can impact model accuracy. Detecting and adapting to drift is therefore critical. Statistical methods such as the Kolmogorov-Smirnov Test [43] are commonly used to compare distributions and

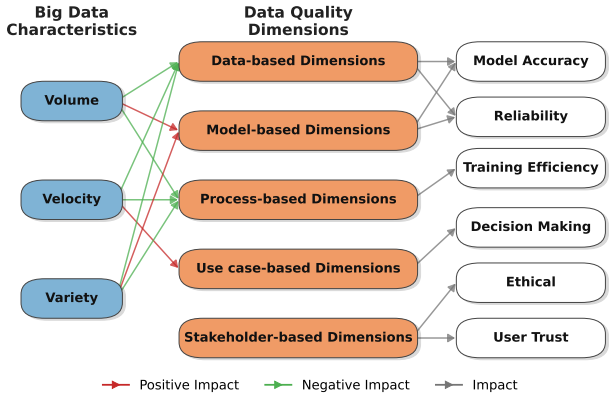

Figure 2. DQ impact in ML Pipelines. The arrows illustrate a left-to-right cascading effect: Big Data characteristics (left) impact specific DQ dimensions (center), which subsequently affect downstream ML (right).

detect significant changes. More advanced methods include Maximum Mean Discrepancy (MMD) [24, 50], a kernel-based test for subtle shifts, and the Classifier Two-Sample Test (C2ST) [76], which trains a classifier to distinguish between datasets. MMD-D [74] enhances MMD for more focused drift detection.

To adapt to drift, reweighting [15] adjusts sample weights during training, and incremental learning [91] updates models continuously with new data. Adversarial training [46] trains models using adversarial examples to make them more robust to distribution changes, enhancing the model's ability to generalize across different data distributions. Another important technique is active learning [106], which reduces labeling effort.

5.3 Impact of data quality in ML pipelines

Data quality challenges in ML pipelines [77, 125] are not isolated phenomena but rather a cumulative stratification of issues inherited from traditional data management, amplified by BDCs, and complicated by machine learning requirements. Primarily, ML systems inherit the impacts of traditional data quality, where dimensions such as correctness, completeness, and consistency remain the bedrock of model utility. Poor data quality at this syntactic level—such as missing values or typographical errors—directly impairs the learning process, as models trained on flawed data inevitably replicate these errors in their predictions. For instance, strictly enforcing schema constraints and referential integrity, while traditional, is a prerequisite for preventing feature extraction failures in ML pipelines. Without ensuring these intrinsic quality dimensions, the downstream model reliability is fundamentally compromised regardless of the algorithmic sophistication.

These foundational issues are amplified by the impacts of BDCs. The transition to high-volume, high-velocity, and high-variety data introduces complexity that traditional cleaning cannot fully address. As illustrated in our conceptual model in Figure 2, BDCs act as causes that propagate effects throughout the pipeline: 'Variety' stemming from heterogeneous sources (e.g., social media combined with sensors) often degrades 'Consistency' and 'Reliability', introducing noise that ML models may mistake for signal. Similarly, 'Velocity' necessitates 'Currentness', where outdated data can lead to immediate model degradation in dynamic environments like fraud detection. Figure 2 further depicts how these distorted dimensions eventually impair 'Stakeholder-based Dimensions' such as User Trust, demonstrating a causal chain where BDCs degrade the final decision-making confidence.

Beyond inheritance and amplification, ML pipelines face unique semantic impacts that do not exist in traditional or Big Data systems. Here, data quality defects are defined not by syntactic errors, but by distributional properties such as bias, fairness, and concept drift. A dataset can be accurate and complete (in a traditional sense) yet functionally toxic to an ML model if it contains historical biases or violates train/test independence (data leakage). These unique dimensions directly affect the ethical alignment and generalizability of the model. For example, a lack of representativeness in training data does not break the database query, but it causes the ML model to fail catastrophically on underrepresented minority groups, creating severe fairness issues in high-stakes decision-making.

The interplay of these three layers is best observed in critical application scenarios. In financial trading, a system must ensure transaction accuracy (Traditional), handle the velocity of market ticks (Big Data), and detect subtle distribution shifts to prevent model drift (ML-Unique). Despite the availability of techniques like sampling and incremental profiling to manage Big Data volume, a significant gap remains: current Big Data quality tools often lack the semantic awareness required to detect nuanced ML-specific issues like fairness violations and explainability deficits. Consequently, future DQ awareness must evolve from purely ensuring data cleanliness to validating the semantic fitness and ethical implications of data within the specific context of ML tasks.

6 NEW OPPORTUNITIES OF DQ FOR DATA SCIENCE SYSTEMS

The intersection of Big Data and machine learning within data science systems presents unique challenges in data quality, which opens up several opportunities for advancing quality awareness techniques.

6.1 Enhancing Quality Awareness in DS Systems

Challenges with Multimodal Data: Current DQ systems struggle with the diverse characteristics of multimodal data, such as discrepancies between discrete text and continuous

image data [71]. Synchronization issues in time-dependent data formats like video and audio further complicate accurate model performance. Future research should develop unified quality metrics and cross-modal consistency checks tailored for multimodal environments [128]. Advanced data fusion techniques, such as multi-modal GANs, could address data gaps and reduce redundancy, ensuring robust data integrity across different modalities [135].

Adaptive Quality Monitoring: Dynamic environments where data continuously evolves, such as those involving real-time data streams, demand adaptive quality monitoring frameworks. Integrating ML models with dynamic profiling techniques could facilitate real-time adaptation to emerging data patterns like drift, enhancing the responsiveness of DQ validation systems [94, 103].

6.2 Addressing AI Ethics

An important research direction within the intersection of fairness and data cleaning is addressing how existing quality awareness techniques affect fairness in ML pipelines. Specifically, we need to examine whether existing data cleaning techniques can be effectively adapted to handle fairness constraints, or if new techniques should be developed. One promising opportunity lies in extending current cross-validation and model optimization techniques to include fairness, rather than only prioritizing accuracy. Techniques like Shapley values [65] have been proposed to identify data points that negatively impact fairness, suggesting a path for fairness-enhanced cleaning. The Big Data challenges further amplify fairness issues in ML pipelines, emphasizing the need for scalable, fairness-aware solutions that balance accuracy and ethical considerations. In addition, enhancing transparency and interpretability in fairness-aware cleaning is also a key research direction to leverage trust and accountability in data science systems [56, 126]. Developing techniques that clarify how these fairness-aware cleaning techniques impact model fairness will help reveal features affecting outcomes.

6.3 The Dual-Role of LLMs in Data Quality

LLMs occupy a unique dual role in the DQ landscape of modern DS systems: they simultaneously act as *objects* of DQ assessment—requiring carefully curated and unbiased training corpora—and as emerging *tools* for performing DQ tasks on other datasets. The following two subsections examine each role in turn.

6.3.1 Data Quality for LLMs LLMs have become integral to data science systems, offering the ability to extract insights and make data-driven decisions at scale. However, one of the main challenges associated with LLMs is the issue of *hallucination* [10, 63]. This issue often arises from harmful errors in the original training data, where biases or inaccuracies lead to hallucinations, and can be exacerbated when LLMs are trained on outputs generated by other LLMs, which creates a feedback loop that further degrades output quality [117].

Addressing these challenges requires ensuring the quality and authenticity of the data used in both training and evaluation [23, 136]. Conventional methods, such as verifying data sources, using human-in-the-loop validation, and ensuring transparency in data provenance, help mitigate hallucination risks. However, the scale and complexity of LLMs require new approaches to effectively manage and improve data quality as these models continue to expand [63, 70]. Additionally, new data quality metrics may be required to capture the specific challenges introduced by LLMs, including mechanisms to verify generated content and manage internal biases, ensuring accuracy and reliability as they scale.

6.3.2 Using LLMs for Data Quality In addition to being subjects of DQ assessment, LLMs are rapidly emerging as powerful tools for DQ assessment and improvement. Their ability to understand natural language and context enables new approaches to data cleaning. For instance, recent studies explore using LLM-based agents for data science tasks, including automated data preprocessing and cleaning. LLMs can be used to identify semantic anomalies (e.g., a "Sony" value in a "Brand" column misspelled as "Soni"), suggest repairs for functional dependency violations, and even automate data quality assessment (DQA) tasks that previously required human-in-the-loop validation. This 'LLM4DATA' paradigm represents a significant new research frontier [137].

7 CONCLUSIONS

This paper traces the evolution of Data Quality (DQ) awareness. We conclude with three key takeaways. First, DQ challenges are cumulative: each new system inherits the concerns of its predecessors—syntactic correctness from traditional data management, scalability from Big Data—while adding new layers, culminating in semantic and ethical dimensions (fairness, bias, drift, label quality) that are unique to ML/DS systems. Second, DQ awareness itself has shifted from a *data-at-rest* concern (traditional systems), to a *data-in-motion* concern (Big Data), and finally to a *data-for-decision-making* concern (ML pipelines), demanding holistic, end-to-end approaches rather than isolated preprocessing steps. Third, the frontier of DQ is increasingly semantic: future challenges center less on technical cleanliness and more on fitness-for-use, fairness, and the reliability of LLM-derived outputs.

Consequently, the path forward requires holistic, end-to-end quality frameworks—adaptive enough to manage this growing complexity—that prioritize not just the data's form, but a broader awareness of the influence of data

quality across modern data ecosystems, particularly in the era of Large Language Models.

Acknowledgements We would like to thank Dr. Divesh Srivastava for his several comments that helped improve our paper.

References

[1] Z. Abedjan, L. Golab, and F. Naumann. Data profiling: A tutorial. In *Proc. SIGMOD*, pages 1747–1751, 2017.

[2] S. Abiteboul et al. The elephant in the room: getting value from big data. In *WebDB*, pages 1–5, 2015.

[3] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases*, volume 8. Addison-Wesley Reading, 1995.

[4] D. Ardagna et al. Context-aware data quality assessment for big data. *Future Gener. Comput. Syst.*, 89:548–562, 2018.

[5] A. B. Arrieta et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion*, 58:82–115, 2020.

[6] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41:1–52, 2009.

[7] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi. From data quality to big data quality. *Journal of Database Management (JDM)*, 26(1):60–82, 2015.

[8] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. 01 2006.

[9] D. Baylor et al. Tfx: A tensorflow-based production-scale machine learning platform. In *Proc. KDD*, pages 1387–1395, 2017.

[10] E. M. Bender et al. On the dangers of stochastic parrots: Can language models be too big? In *Proc. FAccT*, pages 610–623, 2021.

[11] L. Berti-Equille. Quality-adaptive query processing over distributed sources. In *Proc. ICIQ*, 2004.

[12] L. Berti-Équille. Contributions to quality-aware online query processing. *IEEE Data Eng. Bull.*, 29(2):32–42, 2006.

[13] E. Bertino and E. Ferrari. Big data security and privacy. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, volume 31, pages 425–439. 2018.

[14] P. J. Bickel and D. A. Freedman. Asymptotic Normality and the Bootstrap in Stratified Sampling. *The Annals of Statistics*, 12(2):470 – 482, 1984.

[15] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10, 2009.

[16] J. Bleiholder and F. Naumann. Data fusion. *ACM computing surveys (CSUR)*, 41(1):1–41, 2009.

[17] A. Boukerche, L. Zheng, and O. Alfandi. Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, 53(3):1–37, 2020.

[18] L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2015.

[19] C. Cappiello, W. Samá, and M. Vitali. Quality awareness for a successful big data exploitation. In *Proc. IDEAS*, pages 37–44, 2018.

[20] S. Chauhan and L. Vig. Anomaly detection in ecg time signals via deep long short-term memory networks. In *Proc. IEEE DSAA*, pages 1–7, 2015.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[22] F. Chiang and R. J. Miller. Discovering data quality rules. *Proceedings of the VLDB Endowment*, 1(1):1166–1177, 2008.

[23] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[24] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015.

[25] Cisneros-Cabrera et al. Experimenting with big data computing for scaling data quality-aware query processing. *Expert Systems with Applications*, 178:114858, 2021.

[26] U. Comignani, N. Novelli, and L. Berti-Équille. Data quality checking for machine learning with mesqual. In *EDBT 2020*, pages 591–594, 2020.

[27] G. Cormode. Sketch techniques for approximate query processing. 2010.

[28] G. Cormode. Data sketching. *Commun. ACM*, 60(9):48–55, 2017.

[29] D. Crankshaw et al. The missing piece in complex analytics: Low latency, scalable model management and serving with velox. *arXiv preprint arXiv:1409.3809*, 2014.

[30] E. D. Cubuk et al. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019.

[31] G. d'Aloisio, A. D. Marco, and G. Stilo. Modeling quality and machine learning pipelines through extended feature models, 2022.

[32] A. Dobra, M. N. Garofalakis, J. Gehrke, and R. Rastogi. Sketch-based multi-query processing over data streams. In *Data Stream Management*, 2004.

[33] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[34] X. L. Dong and D. Srivastava. Big data integration. In *2013 IEEE 29th international conference on data engineering (ICDE)*, pages 1245–1248. IEEE, 2013.

[35] C. Dwork et al. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[36] L. Ehrlinger and W. Wöß. Automated data quality monitoring. In *ICIQ*, 2017.

[37] I. El Alaoui and Y. Gahi. The impact of big data quality on sentiment analysis approaches. *Procedia Computer Science*, 160:803–810, 2019.

[38] I. El Alaoui, Y. Gahi, and R. Messoussi. Big data quality metrics for sentiment analysis approaches. In *Proceedings of the 2019 International Conference on Big Data Engineering*, pages 36–43, 2019.

[39] A. Esmaelizadeh, J. Rorseth, A. Yu, P. Godfrey, and e. a. Golab. On integrating the data-science and machine-learning pipelines for responsible ai. In *Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI*, pages 50–53, 2024.

[40] Explosion AI, 2018.

[41] H. Fadlallah, R. Kilany, H. Dhayne, R. El Haddad, R. Haque, Y. Taher, and A. Jaber. Context-aware big data quality assessment: A scoping review. *J. Data and Information Quality*, 15(3), 2023.

[42] H. Fadlallah, R. Kilany, H. Dhayne, and et al. Context-aware big data quality assessment: a scoping review. *ACM Journal of Data and Information Quality*, 15(3):1–33, 2023.

[43] G. Fasano and A. Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, pages 155–170, 1987.

[44] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the KDD*, pages 259–268, 2015.

[45] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini. On the meaningfulness of "big data quality" (invited paper). *Data Science and Engineering*, 1(1):6–20, 2016.

[46] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*,

17(59):1–35, 2016.

[47] M. Ghasemaghaei. The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage. *International Journal of Information Management*, 50:395–404, 2020.

[48] I. J. Goodfellow et al. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[49] Google. The What-If Tool, 2018.

[50] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1), 2012.

[51] N. Gupta, H. Patel, S. Afzal, N. Panwar, R. S. Mittal, S. Guttula, and et al. Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets, 2021.

[52] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: The teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16, 2006.

[53] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[54] Y. He, J. Z. Huang, H. Long, Q. Wang, and C. Wei. I-sampling: A new block-based sampling method for large-scale dataset. In *2017 IEEE International Congress on Big Data (BigData Congress)*, pages 360–367. IEEE, 2017.

[55] R. H. Henderson and T. Sundaresan. Cluster sampling to assess immunization coverage: a review of experience with a simplified sampling method. *Bulletin of the World Health Organization*, 60(2):253 – 260, 1982.

[56] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 2019.

[57] P. J. Huber and E. M. Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

[58] N. Hynes, D. Sculley, and M. Terry. The data linter: Lightweight, automated sanity checking for ml data sets. In *NIPS MLSys Workshop*, volume 1, 2017.

[59] IBM. AIF360: IBM Fairness 360, 2018.

[60] I. F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. *Foundations and Trends in Databases*, 5(4):281–393, 2015.

[61] D. Jarrett, B. Cebere, T. Liu, A. Curth, and M. van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *ICML*, volume 162, pages 9916–9937, 2022.

[62] V. Jayawardene, S. Sadiq, and M. Indulska. An analysis of data quality dimensions. 2015.

[63] Z. Ji, N. Lee, R. Frieske, T. Yu, and et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[64] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[65] B. Karlavs, D. Dao, M. Interlandi, S. Schelter, W. Wu, and C. Zhang. Data debugging with shapley importance over machine learning pipelines. In *Proc. ICLR*, 2024.

[66] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4401–4410, 2019.

[67] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, and e. a. Mahmoud Ali. Big data: survey, technologies, opportunities, and challenges. *The scientific world journal*, 2014(1):712826, 2014.

[68] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[69] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. Activeclean: Interactive data cleaning for statistical modeling. *PVLDB*, 9(12):948–959, 2016.

[70] P. Lewis, E. Perez, A. Piktus, and e. a. Petroni. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[71] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, 2017.

[72] K. Li and G. Li. Approximate query processing: What is new and where to go? *Data Science and Engineering*, 3:379–397, 2018.

[73] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang. Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In *ICDE 2021*, pages 13–24. IEEE, 2021.

[74] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *ICML*, pages 6316–6326. PMLR, 2020.

[75] H. Liu, Z. Sang, and S. Karali. Approximate quality assessment with sampling approaches. In *CSCI 2019*.

[76] D. Lopez-Paz and M. Oquab. Revisiting classifier two-sample tests. *ICLR*, 2017.

[77] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz. Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797, 2017.

[78] P.-A. Mattei and J. Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proc. ICML*, pages 4413–4423, 2019.

[79] J. Merino, I. Caballero, B. Rivas, M. A. Serrano, and M. Piattini. A data quality in use model for big data. *Future Gener. Comput. Syst.*, 63:123–130, 2016.

[80] D. T. Meyer and W. J. Bolosky. A study of practical deduplication. *ACM Transactions on Storage (ToS)*, 7(4):1–20, 2012.

[81] S. Mohammed, L. Budach, et al. The effects of data quality on machine learning performance on tabular data. *Inf. Syst.*, 132, 2025.

[82] S. Mohammed, L. Ehrlinger, H. Harmouch, F. Naumann, and D. Srivastava. The five facets of data quality assessment. *ACM SIGMOD Record*, 54(2):18–27, 2025.

[83] D. C. Montgomery. *Introduction to statistical quality control*. John wiley & sons, 2019.

[84] F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer, 2002.

[85] F. Naumann. Data profiling revisited. *SIGMOD Rec.*, 42(4):40–49, 2013.

[86] F. Naumann et al. Quality-driven integration of heterogenous information systems. In *PVLDB '99*, pages 447–458, 1999.

[87] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu. From cleaning before ml to cleaning for ml. *IEEE Data Eng. Bull.*, 44(1):24–41, 2021.

[88] M. H. Ofner, B. Otto, and H. Österle. Integrating a data quality perspective into business process management. *Business Process Management Journal*, 18(6):1036–1067, 2012.

[89] P. Oliveira, F. Rodrigues, and P. Henriques. Data profiling versus data quality problems. *Actes du 2nd Atelier Qualité des données et des connaissances en conjonction avec EGC 2006*, page 9, 2006.

[90] L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[91] R. Polikar, L. Upda, S. S. Upda, and V. Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, 31(4):497–508, 2001.

[92] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data management challenges in production machine learning. In *Proc. SIGMOD*, pages 1723–1726, 2017.

[93] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data lifecycle challenges in production machine learning: a survey.

ACM SIGMOD Record, 47(2):17–28, 2018.

[94] N. Polyzotis, M. Zinkevich, S. Roy, E. Breck, and S. Whang. Data validation for machine learning. In *Proceedings of Machine Learning and Systems*, volume 1, pages 334–347, 2019.

[95] M. Priestley, F. O'donnell, and E. Simperl. A survey of data quality requirements that matter in ml development pipelines. jun 2023.

[96] A. Ratner, S. H. Bach, H. Ehrenberg, and e. a. Fries. Snorkel: Rapid training data creation with weak supervision. In *PVLDB*, volume 11, page 269, 2017.

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

[98] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. Holoclean: Holistic data repairs with probabilistic inference. *arXiv preprint arXiv:1702.00820*, 2017.

[99] A. Roth. Completeness-driven query answering in peer data management systems. In *Proc. of the VLDB 2007 PhD Workshop*. Citeseer, 2007.

[100] S. W. Sadiq, T. Dasu, X. L. Dong, J. Freire, I. F. Ilyas, S. Link, R. J. Miller, F. Naumann, X. Zhou, and D. Srivastava. Data quality: The role of empiricism. *SIGMOD Rec.*, 46(4):35–43, 2017.

[101] B. Saha and D. Srivastava. Data quality: The other face of big data. *2014 IEEE 30th International Conference on Data Engineering*, pages 1294–1297, 2014.

[102] S. Schelter et al. Differential data quality verification on partitioned data. In *ICDE 2019*, pages 1940–1945.

[103] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger. Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12):1781–1794, 2018.

[104] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[105] F. Serra, V. Peralta, A. Marotta, and P. Marcel. Use of context in data quality management: a systematic literature review. *ACM Journal of Data and Information Quality*, 2022.

[106] B. Settles. Active learning literature survey. 2009.

[107] Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[108] H. Shomorony and A. S. Avestimehr. Sampling large data on graphs. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 933–936. IEEE, 2014.

[109] F. Sidi, P. H. S. Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*, pages 300–304. IEEE, 2012.

[110] J. Siebert, L. Joeckel, J. Heidrich, K. Nakamichi, K. Ohashi, and I. N. et al. Towards guidelines for assessing qualities of machine learning systems. In *Communications in Computer and Information Science*, pages 17–31. Springer International Publishing, 2020.

[111] S. Song, F. Gao, R. Huang, and C. Wang. Data dependencies extended for variety and veracity: A family tree. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4717–4736, 2020.

[112] S. Studer, T. B. Bui, C. Drescher, and et al. Towards crisp-ml(q): A machine learning process model with quality assurance methodology, 2021.

[113] I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui. Big data quality: A quality dimensions evaluation. In *UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld 2016*. IEEE, 2016.

[114] I. Taleb, M. A. Serhani, and R. Dssouli. Big data quality assessment model for unstructured data. In *2018 International Conference on Innovations in Information Technology (IIT)*, pages 69–74, 2018.

[115] F. A. A. Tarmizi, P. X. Tan, K. Y. Sharif, and E. Kamioka. Online news veracity assessment using emotional weight. In *Proceedings of the 2019 2nd International Conference on Information Science and Systems*, pages 60–64, 2019.

[116] A. H. Ter Hofstede, A. Koschmider, A. Marrella, et al. Process-data quality: the true frontier of process mining. *ACM Journal of Data and Information Quality*, 15(3):1–21, 2023.

[117] S. Tonmoy, S. Zaman, V. Jain, A. Rani, and e. a. Rawte. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

[118] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, and e. a. Hastie. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[119] J. S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11:37–57, 1985.

[120] S. Wagner, A. Goeb, L. Heinemann, M. Kläs, C. Lampasona, and et al. Operationalised product quality models and assessment: The quamoco approach. *Information and Software Technology*, 62:101–123, jun 2015.

[121] A. Wahyudi, A. Farhani, and M. Janssen. Relating big data and data quality in financial service organizations. In *IFIP International Conference on e-Business, e-Services, and e-Society*, 2018.

[122] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[123] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.*, 12(4):5–33, 1996.

[124] Z. Wang et al. A multi-level non-uniform spatial sampling method for accuracy assessment of remote sensing image classification results. *Applied Sciences*, (16), 2020.

[125] S. E. Whang and J.-G. Lee. Data collection and quality challenges for deep learning. *Proc. VLDB Endow.*, 13(12), 2020.

[126] X. Wu, H. Zhao, Y. Zhu, et al. Usable xai: 10 strategies towards exploiting explainability in the llm era. *arXiv preprint arXiv:2403.08946*, 2024.

[127] W. Xia, H. Jiang, D. Feng, et al. A comprehensive study of the past, present, and future of data deduplication. *Proc. of the IEEE*, 104(9):1681–1710, 2016.

[128] J. Ye, J. Guo, Y. Xiang, et al. Noise-robust cross-modal interactive learning with text2image mask for multi-modal neural machine translation. In *Proc. COLING*, pages 5098–5108, 2022.

[129] N. K. Yeganeh et al. A framework for data quality aware query systems. *Information Systems*, 46:24–44, 2014.

[130] J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR, 2018.

[131] B. M. v. Zernichow and D. Roman. Usability of visual data profiling in data cleaning and transformation. In *OTM Confederated International Conferences*, pages 480–496. Springer, 2017.

[132] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proc. of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[133] H. Zhang et al. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

[134] S. Zhang. Nearest neighbor selection for iteratively knn imputation. *J. Syst. Softw.*, 85(11), 2012.

[135] Y. Zhang, J. Shen, Z. Zhang, and C. Wang. Partial modal conditioned gans for multi-modal multi-label learning with arbitrary modal-missing. In *DASFAA 2021*, pages 413–428. Springer, 2021.

[136] A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, and G. Huang. Expel: Llm agents are experiential learners. In *Proc. AAAI*, volume 38, pages 19632–19642, 2024.

[137] X. Zhou et al. A survey of llm× data. *arXiv preprint arXiv:2505.18458*, 2025.

[138] Y. Zhou, F. Tu, K. Sha, J. Ding, and H. Chen. A survey on data quality dimensions and tools for machine learning, 2024.