

Adaptive Data Quality Scoring Using Machine Learning for Large-Scale Enterprise Data Pipelines: A Systematic Review

Sai Mohan Dasari

Senior Data Engineer

Abstract

Enterprise data pipelines increasingly ingest heterogeneous, high-velocity data, and the quality of that data directly determines the reliability of downstream analytics and machine learning systems. This review synthesises the literature on data quality assessment and on the use of machine learning to automate and scale that assessment. It examines foundational multidimensional quality frameworks, data profiling and rule-based error detection, supervised and unsupervised approaches to anomaly and duplicate detection, the theory of probabilistic record linkage, and production validation systems for continuously evolving pipelines. The synthesis traces a transition from static, rule-based checking towards learned scoring functions that adapt to schema evolution, concept drift, and distributional shift. Holistic, inference-based repair methods and constraint-learning validators substantially reduce the manual specification burden of earlier rule-based approaches, while ensemble methods such as random forests and gradient-boosted trees provide effective classifiers for error and duplicate detection over mixed-type data. Three problems persist across the literature: the scarcity of labelled ground truth for supervised scoring, the entanglement of data quality with the broader technical debt of production machine learning systems, and the difficulty of keeping scores stable under non-stationary data distributions. The review argues that adaptive quality scoring is best structured as three coordinated layers—dimension-specific statistical profiling, learned anomaly and duplicate detection models, and drift-aware recalibration—and offers guidance on integrating quality scoring with automated pipeline orchestration.

Keywords: data quality, machine learning, enterprise data pipelines, data profiling, entity resolution, anomaly detection, concept drift, data validation, automated data cleaning, record linkage

1. Introduction

Automated pipelines are increasingly used by large enterprises to transfer data from operational systems to warehouses, feature stores and machine learning platforms. Data volumes have grown while the data itself has become more complex and varied, making it progressively harder to maintain data quality rules by hand (Cai & Zhu, 2015). Poor data quality has long been known to impose substantial operational costs, including inaccurate reporting, flawed decision making, and degraded model performance (Redman, 1998). In machine learning settings, quality issues arising upstream propagate silently into trained models, forming part of the broader technical debt that accumulates in production systems (Sculley et al., 2015). This review examines the convergence of two longstanding research themes: data quality assessment, and the use of machine learning to automate, scale, and adapt that assessment within enterprise-scale pipelines.

The review is structured as a systematic synthesis rather than a narrative review. Section 2 describes the review protocol. Section 3 presents the foundational multidimensional definition of data quality. Section 4 addresses data profiling and rule-based error detection. Section 5 examines machine learning techniques for quality scoring, including supervised classification, entity resolution, and anomaly detection. Section 6 surveys production validation and repair systems suited to pipeline-scale operation. Section 7 considers adaptivity under concept drift. Section 8 offers a comparative synthesis, Section 9 sets out open challenges, and Section 10 concludes. Figure 1 gives an overview of the architecture around which the review is organised, connecting data ingestion, profiling, learned scoring, and drift-aware recalibration into a closed-loop process.

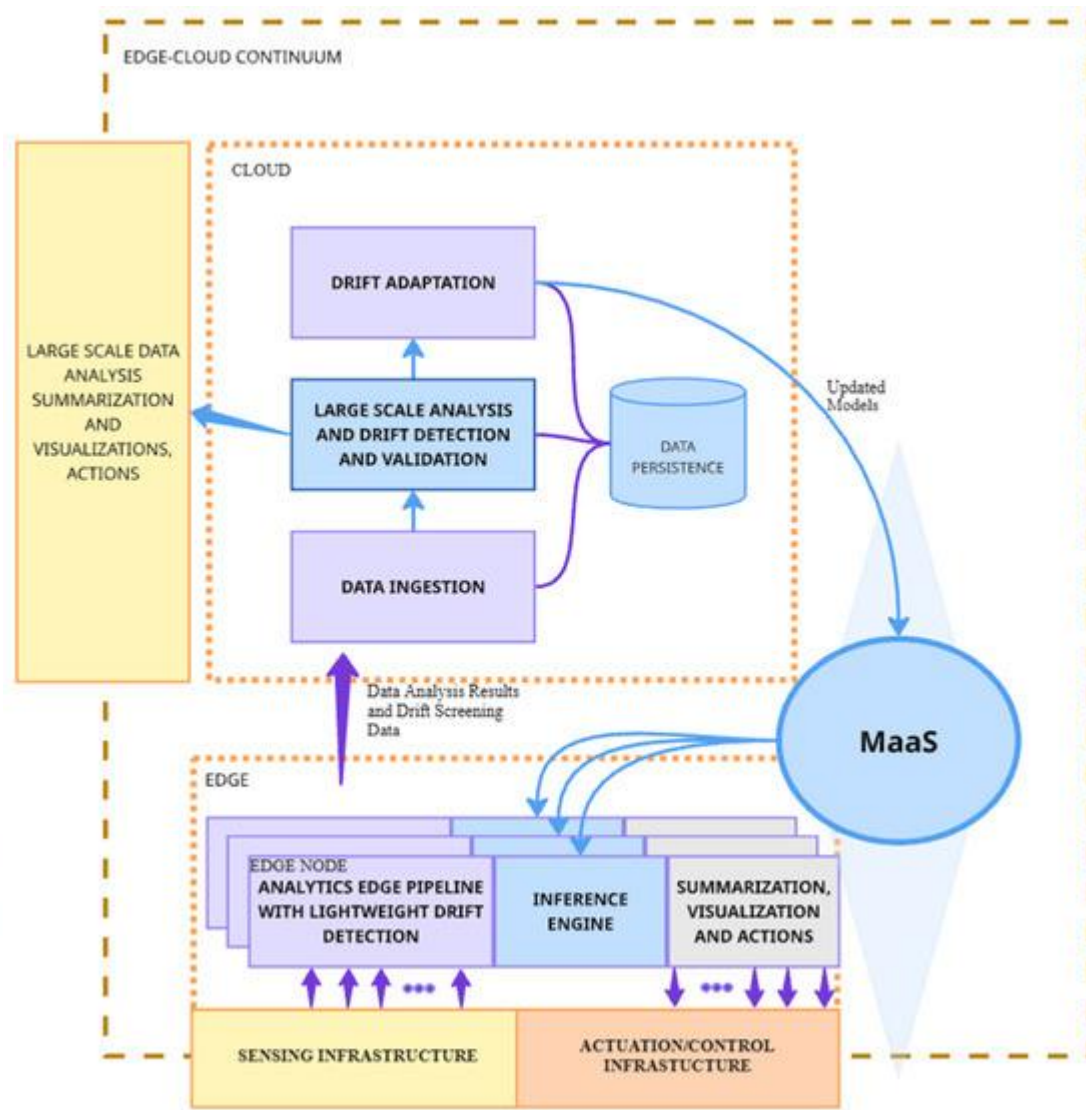


Figure 1

High-Level Adaptive Data Quality Scoring Pipeline, from Ingestion Through Composite Scoring and Drift-Aware Recalibration (MDPI, 2020)

2. Review Methodology

The reviewed corpus comprises twenty-five peer-reviewed and archival sources spanning database systems, data mining, statistics, and applied machine learning venues, including the Proceedings of the VLDB Endowment, ACM Computing Surveys, IEEE Transactions on Knowledge and Data Engineering, SIGKDD, SIGMOD, NeurIPS, and MLSys. The sources were chosen because they directly addressed three themes: formal definitions and dimensions of data quality; techniques for error and duplicate detection, and anomaly detection algorithms; and system-level approaches to data validation and repair for automated pipelines. Findings were presented thematically instead

of chronologically, focusing on the methodological contribution and the problem formulation for each source so that complementary contributions could be synthesized across the venues in a coherent manner.

3. Foundations of Data Quality Assessment

Early research on data quality evolved from a singular notion of “accuracy” towards a multidimensional conception grounded in the needs of data consumers (Wang & Strong, 1996). In this view, dimensions such as accuracy, completeness, consistency, timeliness, and believability are treated as distinct constructs rather than as facets of a single underlying dimension. This was followed by a number of studies which provided a more methodical way of measuring and

evaluating quality in this manner, and which began to distinguish between quantitative measures, which were applied to structured or semi-structured data, and qualitative measures, which were judgmental (Batini et al., 2009). A parallel line of work developed practical measurement instruments, arguing that quality dimensions must be translated into concrete, repeatable metrics usable within an organisation (Pipino et al., 2002). Table 1 summarises the dimensions common to these foundational frameworks.

Table 1

Core Data Quality Dimensions Synthesized from Foundational Frameworks

Dimension	Definition	Primary Source
Accuracy	Degree to which recorded values correctly represent the real-world state they describe	(Pipino et al., 2002; Wang & Strong, 1996)
Completeness	Extent to which all required attribute values are present in a data set	(Batini et al., 2009; Wang & Strong, 1996)
Consistency	Absence of contradictory representations of the same fact across records or systems	(Batini et al., 2009; Pipino et al., 2002)
Timeliness	Extent to which data remains sufficiently current for its intended use	(Pipino et al., 2002; Wang & Strong, 1996)
Believability	Extent to which data is regarded as credible by its consumers	(Wang & Strong, 1996)

Dimension	Definition	Primary Source
Uniqueness	Absence of unintended duplicate representations of the same real-world entity	(Elmagarmid et al., 2007; Getoor & Machanavajjhala, 2012)

Note. Dimension definitions synthesized from the foundational frameworks cited in the Primary Source column.

A composite quality score for a record or data set may be formed as a weighted sum of dimension-specific sub-scores. Given normalized dimension scores $D_1, D_2, \dots, D_n$ on the unit interval $[0, 1]$, with non-negative weights w_i that sum to 1, the composite score for record r is defined as $DQS(r)$ in Equation 1 (Pipino et al., 2002).

$$DQS(r) = \sum_{i=1}^n w_i \cdot D_i(r) \quad (1)$$

This weighted-sum formulation, originally specified by hand, remains the basic form followed by most of the automated scoring systems described in Section 5: the dimensions D_i are replaced by machine-learned functions, and the weighting scheme w_i remains a mechanism by which domain and use-case priorities are added to an otherwise automated pipeline (Polyzotis et al., 2019). The individual dimensions of Table 1 are aggregated into the composite score given in Equation 1, as shown in Figure 2.

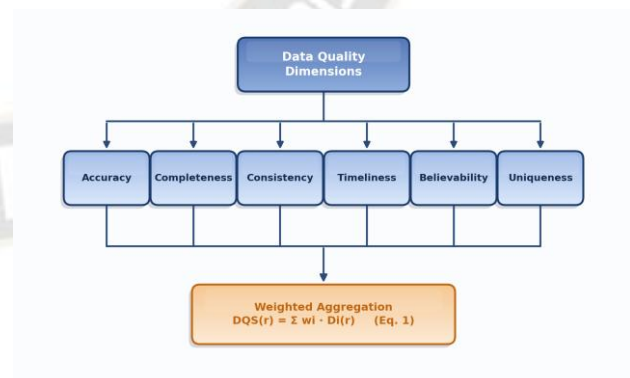


Figure 2

Core Quality Dimensions Aggregated into the Composite Data Quality Score of Equation 1

4. Data Profiling and Rule-Based Error Detection

Computing a quality score first requires characterising a data set's structural and statistical properties through profiling, which includes techniques such as estimating column cardinality and discovering functional dependencies (Abedjan et al., 2015). Profiling outputs are used to derive candidate integrity constraints, such as denial constraints or conditional functional dependencies, which then serve as rule-based detectors for records that violate them. A comprehensive examination of the error detection literature concluded that, after decades of development, no single technique dominates across all error classes, and that combining complementary detectors is generally more effective than any individual detector (Abedjan et al., 2016).

This observation motivated systems that treat error detection as a learnable ensemble rather than a fixed rule set, directly anticipating the machine learning-based scoring approaches discussed later in this review.

The wider data cleaning literature frames the cleaning pipeline as error detection, error repair, and human-in-the-loop quality control, identifying scalability to enterprise-sized data sets and reduction of manual effort as the key open problems (Chu et al., 2016). In keeping with this perspective, interactive cleaning systems have developed sampling-based approaches that tie the effort of data cleaning directly to the quality improvement of the records needed to support a statistical model that is relevant to a given use case, as opposed to considering cleaning as a generic pre-processing step (Krishnan et al., 2016). For enterprise pipelines, where the volume, velocity, and variety of data creates a need for quality assessment at ingestion, rather than at a batch post-processing phase, large-scale quality frameworks have been proposed specifically for these contexts (Taleb et al., 2015).

5. Machine Learning Approaches to Quality Scoring

5.1 Supervised Classification for Error and Anomaly Scoring

Ensemble tree methods are a prominent supervised approach to data quality scoring, owing to their robustness to heterogeneous feature types and their capacity to capture non-linear interactions among profiling features without extensive manual feature engineering. Random forests build a forest of decorrelated decision trees using bootstrap samples of the training data and average the votes of the various trees, a process that has lower variance than a single

decision tree, but is still tractable at scale (Breiman, 2001). Gradient-boosted tree ensembles (Chen & Guestrin, 2016) further follow this concept by fitting successive trees to the residual error of previously built trees using a regularized objective, and have been demonstrated to work well with sparse, large-scale feature matrices with good generalization, which makes fitting trees to profile statistics-derived features for quality-scoring an appealing application. The regularized learning objective minimized by such boosted ensembles can be expressed as Equation 2, where l is a differentiable loss function that compares the predicted quality label $\hat{y}_j$ with the true quality label y_j , f_k is the k -th regression tree among K regression trees, and Ω is a tree complexity penalty that controls overfitting (Chen & Guestrin, 2016).

$$L = \sum_j l(\hat{y}_j, y_j) + \sum_k \Omega(f_k) \quad (2)$$

The label y_j is generally binary or ordinal, and it represents the outcome of a historical audit or failure signal in the downstream application, and the outputted classifier score is used as an approximation of the probability that a record meets the quality criteria for the downstream use case, after calibration. The sequential residual-fitting process of Equation 2 is shown in Figure 3.

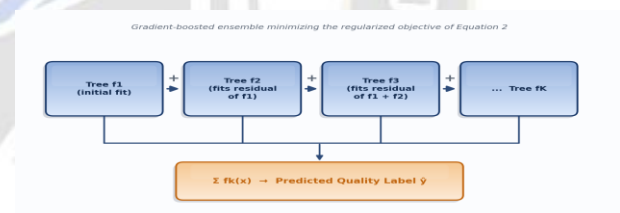


Figure 3

Sequential Residual Fitting in a Gradient-Boosted Tree Ensemble Applied to Quality-Scoring Features

5.2 Entity Resolution and Probabilistic Record Linkage

Many enterprise data quality issues arise not from individual erroneous attribute values, but from the presence of multiple records representing the same real-world entity across different systems, which is the subject of the entity resolution literature. The basic probabilistic theory of record linkage frames the question of whether two records refer to the same entity as a hypothesis test based on an observed agreement pattern (Fellegi & Sunter, 1969). To compare corresponding fields of two records, an agreement vector γ is used, and the likelihood ratio R is the ratio of the probability of observing the pattern under the matching hypothesis M

to the probability of observing the pattern under the non-matching hypothesis U as defined in Equation 3.

$$R(\gamma) = P(\gamma | M) / P(\gamma | U)(3)$$

Pairs with a likelihood ratio above an upper threshold are classified as matches, those below a lower threshold as non-matches and those within the intermediate range are deferred for manual review; this is a three-way decision rule that is the conceptual foundation of contemporary learned matchers (Fellegi & Sunter, 1969) as shown in Figure 4. Later surveys introduced supervised and semi-supervised classifiers, blocking strategies with lower quadratic cost of pairwise comparison, and similarity functions tailored to special types of attributes like names and addresses (Elmagarmid et al., 2007). A thorough treatment of data matching techniques was also provided, formalizing indexing and blocking techniques which enable entity resolution to be tractable on enterprise-scale record volumes, as well as classification and clustering techniques to identify the candidate pairs (Christen, 2012). Later research has also explored entity resolution as a decision problem, considering it to be online and restricted in budget, and proposing strategies for the online evaluation of oracles that would minimize residual uncertainty upon manual verification of the queried sets of entities (Firmani et al., 2016); and comprehensive surveys of the field have highlighted the need to extend the resolution process to streaming and continuously updated sources, where the number of entities is often vast, instead of a static batch collection (Getoor & Machanavajjhala, 2012). Typically, the quality of a resolution outcome, and consequently of any duplicate-detection module in an adaptive scoring mechanism, can be assessed by its precision, recall and harmonic mean in the following formula:

$$F1 = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)}(4)$$

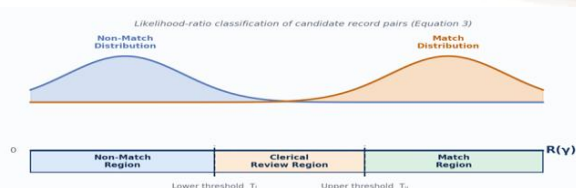


Figure 4

Three-Way Decision Regions Produced by Thresholding the Fellegi-Sunter Likelihood Ratio of Equation 3

5.3 Statistical and Learning-Based Anomaly Detection

In addition to the duplicate detection, quality scoring should identify any individual outlier values or records that may not be duplicated but fall outside of the statistical range. A foundational survey of anomaly detection organised the field into supervised, semi-supervised, and unsupervised paradigms, and distinguished point anomalies from contextual and collective anomalies; the latter two are particularly important in time-indexed enterprise settings such as transaction or sensor streams (Chandola et al., 2009). A machine-learning approach to sensor data quality assessment was developed based on this taxonomy, showing that classifiers based on features describing signal variance, rate of change, and cross-sensor correlation were more successful in detecting faulty sensor readings than fixed-threshold rules, especially when the notion of a “valid reading” is variable, depending on the sensor's operating conditions (Rahman et al., 2014). This result is notable for the present review because it demonstrates, within a narrow application domain, the central principle of adaptive quality scoring: a learned model can replace a hand-tuned threshold while retaining sensitivity to context that a fixed threshold cannot capture.

6. Adaptive Validation and Repair Systems for Production Pipelines

The approaches outlined in Section 5 are integrated into system-level frameworks designed to run continuously within production machine learning pipelines, rather than as one-off analyses. A platform for end to end production scale machine learning included an explicit data validation step to compare data batches coming from the data source to the data schema learned from the data and expected feature statistics, automatically disqualifying data batches that are skewed from the baseline distribution before they are used for model training (Baylor et al., 2017). A similar proposal for data validation for machine learning is to use schema and statistics as versioned, first-class artifacts that change with the data, and raise “data is weird” alerts when new data contains anomalies to the existing schema, instead of having to engineer all possible data failure modes ahead of time (Polyzotis et al., 2019). A large-scale verification framework also suggested declarative constraint specification over data quality dimensions, proposed computing constraint-satisfaction metrics incrementally based on single-pass, distributed algorithms that are suitable for data sets that cannot fit in

the memory of a single machine (Schelter et al., 2018), and further argued that many of the data quality constraints can be automatically suggested based on a profile of historical data, thereby reducing the need to manually write constraints.

A complementary line of work is aimed not solely at detecting quality violations, but also at automatically fixing them. The joint distribution of clean values, integrity constraints and external reference values was modeled with a holistic repair framework, which was based on probabilistic graphical inference, to enable repairs that take correlations across all values of the data into account, not only repairs that resolve violations column by column (Rekatsinas et al., 2017). This reflects a general pattern across the systems reviewed: rather than validating each field independently, they score and repair quality holistically by performing inference over structured, interdependent data. Table 2 compares the rule-based and learning-based error-detection paradigms introduced in Sections 4 through 6.

Table 2

Rule-Based Versus Learning-Based Error Detection Paradigms

Characteristic	Rule-Based Detection	Learning-Based Detection
Constraint origin	Manually authored integrity constraints and functional dependencies	Learned from profiling statistics and, where available, labeled examples
Adaptation to new patterns	Requires manual rule revision	Can be retrained or recalibrated as data evolves
Handling of ambiguous cases	Binary violation flag	Probabilistic or ranked score amenable to thresholding

Characteristic	Rule-Based Detection	Learning-Based Detection
Representative systems	Denial-constraint and conditional functional dependency detectors (Abedjan et al., 2016)	TFX data validation (Baylor et al., 2017; Polyzotis et al., 2019); Deequ (Schelter et al., 2018); HoloClean (Rekatsinas et al., 2017); ActiveClean (Krishnan et al., 2016)
Principal limitation	Coverage bounded by rules an engineer anticipates	Requires representative training signal and periodic recalibration

Note. Synthesized from the systems and surveys cited in Sections 4 through 6.

7. Adaptivity Under Concept Drift

The statistical properties of enterprise data streams rarely remain stationary: schemas evolve, upstream business processes change, and the relationships among attributes shift over time.

This phenomenon is known as concept drift, and it can cause a quality score to degrade as the underlying data-generating process evolves. A thorough study for concept drift adaptation identified sudden, gradual, incremental, and recurring drift types and discussed the mechanisms used for monitoring concept drift through error rates of classifiers over sliding windows, distributions of features, and explicit change-point statistics over sliding windows (Gama et al., 2014). In the context of data quality scoring, such a literature suggests that a static threshold or a model trained with historical data is unlikely to be valid in perpetuity; rather, the data quality scoring system would need to be periodically recalibrated. A representative adaptive update rule (as in exponentially weighted moving-average formulations studied in this literature) updates an expected baseline statistic μ_t at time t using the newly

observed batch statistic x_t , and a smoothing parameter α , as in Equation 5.

$$\mu_t = \alpha \cdot x_t + (1 - \alpha) \cdot \mu_{t-1} \quad (5)$$

When the deviation of a new batch from μ_t exceeds a tolerance distance determined from the data, it can be treated as a candidate quality violation, and the tolerance itself can be re-estimated periodically as drift accumulates. This mechanism extends the fixed-schema comparison performed by production validation systems (Baylor et al., 2017) and is the intersection of the concept-drift literature and the production data-validation literature, as summarized by the closed-loop recalibration process in Figure 5.

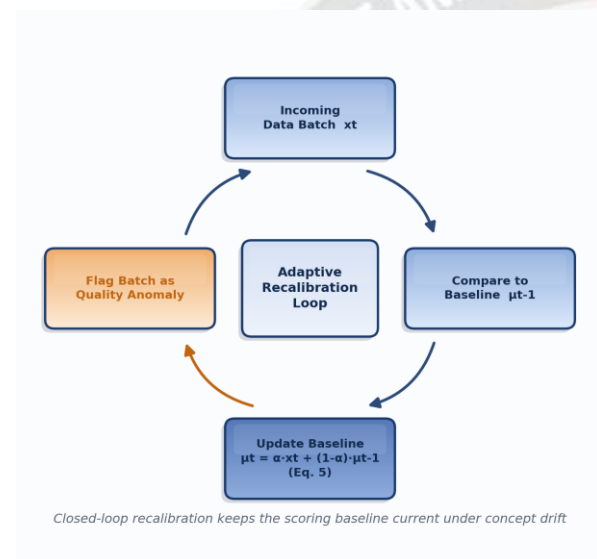


Figure 5

Closed-Loop Recalibration of the Scoring Baseline Under Concept Drift, Following Equation 5

8. Comparative Synthesis

Taken together, the reviewed contributions depict the development of quality assessment across three partially overlapping generations. The first generation, represented here by multidimensional quality frameworks and constraint discovery through profiling, established what should be measured and introduced the concepts—accuracy, completeness, and consistency—that later automated systems inherited. The second generation applied statistical learning to individual sub-problems, such as duplicate detection using probabilistic record linkage (Fellegi & Sunter, 1969) and anomaly detection using supervised and unsupervised classifiers (Chandola et al., 2009) instead of fixed thresholds, by learning the decision boundaries. The third generation integrates these learned components into production-

grade, run-time, continuously operating pipeline infrastructure, providing schema-aware validation and incremental constraint checking, and, where feasible, automated repair (Schelter et al., 2018). Table 3 compares four representative third-generation, production-oriented systems along dimensions relevant to enterprise deployment.

Table 3

Comparison of Representative Production Data Quality Systems

System	Core Mechanism	Primary Focus	Source
TFX Data Validation	Schema and statistics comparison against a learned baseline	Detecting distributional anomalies in ML training data	(Baylor et al., 2017; Polyzotis et al., 2019)
Deequ	Declarative constraints verified via incrementally, distributed computation	Scalable constraint checking over large batch data sets	(Schelter et al., 2018)
HoloClean	Probabilistic graphical inference over constraints and reference data	Holistic automated repair of detected violations	(Rekatsinas et al., 2017)
ActiveClean	Sampling prioritized by estimated impact on a downstream model	Efficient human-in-the-loop cleaning for model training	(Krishnan et al., 2016)

Note. Comparison synthesized qualitatively from the cited system descriptions; no numeric benchmark values are reported in common across the sources.

9. Challenges and Open Issues

Several problems recur across the literature reviewed. First, supervised methods of error and duplicate detection require expensive labeled ground truth data that is impractical to acquire at an enterprise scale, which is why the sampling-prioritization and oracle-querying methods described in Sections 4 and 5 (Firmani et al., 2016) are attractive. Second, quality scoring cannot be separated from the larger system of machine learning in which it is employed; the technical-debt view of production machine learning emphasizes that data dependencies and feedback loops extend beyond the scope of any single quality score, so a locally valid score does not guarantee the absence of hidden system-level dependencies between features (Sculley et al., 2015). Third, the incremental distributed verification that is needed by enterprise-scale volumes limits the choice of statistical methods that can be applied: for example, single-pass, mergeable computations are more practical than algorithms requiring multiple passes over the entire data set (Schelter et al., 2018). Fourth, adaptivity creates a tension: a scoring system that adapts too readily may absorb a genuine decline in quality as if it were a benign distributional shift, whereas one that adapts too conservatively may generate alert fatigue by flagging every legitimate change in the data (Gama et al., 2014).

10. Conclusion and Future Directions

This review has synthesised research at the intersection of classical data quality assessment and the machine learning-based automation of enterprise-scale data pipelines. Multidimensional quality frameworks continue to provide the conceptual basis for validating automated scores, while ensemble classification, probabilistic record linkage, and anomaly detection supply the algorithmic components from which learned scores are constructed. Production systems increasingly combine these components into schema-aware, incrementally computed and, in some cases, self-repairing validation layers, with the concept drift literature providing the theoretical foundation for their recalibration. The reviewed work points to three priorities for future research: coupling quality scoring more tightly to the specific downstream model it serves; reducing the labelling burden of supervised detectors through sampling and active learning; and establishing

drift-aware recalibration as a standard component of enterprise data pipelines.

References

1. Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, 9(12), 993–1004. <https://doi.org/10.14778/2994509.2994518>
2. Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *The VLDB Journal*, 24(4), 557–581. <https://doi.org/10.1007/s00778-015-0389-y>
3. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), Article 16. <https://doi.org/10.1145/1541880.1541883>
4. Baylor, D., Breck, E., Cheng, H.-T., Fiedel, N., Foo, C. Y., Haque, Z., Haykal, S., Ispir, M., Jain, V., Koc, L., Koo, C. Y., Lew, L., Mewald, C., Modi, A. N., Polyzotis, N., Ramesh, S., Roy, S., Whang, S. E., Wicke, M., ... Zinkevich, M. (2017). TFX: A TensorFlow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1387–1395). Association for Computing Machinery. <https://doi.org/10.1145/3097983.3098021>
5. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
6. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, Article 2. <https://doi.org/10.5334/dsj-2015-002>
7. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15. <https://doi.org/10.1145/1541880.1541882>
8. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>

9. Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer. <https://doi.org/10.1007/978-3-642-31164-2>
10. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data* (pp. 2201–2206). Association for Computing Machinery. <https://doi.org/10.1145/2882903.2912574>
11. Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 1–16. <https://doi.org/10.1109/TKDE.2007.250581>
12. Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183–1210. <https://doi.org/10.1080/01621459.1969.10501049>
13. Firmani, D., Saha, B., & Srivastava, D. (2016). Online entity resolution using an oracle. *Proceedings of the VLDB Endowment*, 9(5), 384–395. <https://doi.org/10.14778/2876473.2876474>
14. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44. <https://doi.org/10.1145/2523813>
15. Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12), 2018–2019. <https://doi.org/10.14778/2367502.2367564>
16. Krishnan, S., Wang, J., Wu, E., Franklin, M. J., & Goldberg, K. (2016). ActiveClean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9(12), 948–959. <https://doi.org/10.14778/2994509.2994514>
17. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211–218. <https://doi.org/10.1145/505248.506010>
18. Polyzotis, N., Zinkevich, M., Roy, S., Breck, E., & Whang, S. (2019). Data validation for machine learning. In *Proceedings of Machine Learning and Systems 1* (pp. 334–347).
19. Rahman, A., Smith, D. V., & Timms, G. (2014). A novel machine learning approach toward quality assessment of sensor data. *IEEE Sensors Journal*, 14(4), 1035–1047. <https://doi.org/10.1109/JSEN.2013.2291855>
20. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79–82. <https://doi.org/10.1145/269012.269025>
21. Rekatsinas, T., Chu, X., Ilyas, I. F., & Ré, C. (2017). HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11), 1190–1201. <https://doi.org/10.14778/3137628.3137631>
22. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, 11(12), 1781–1794. <https://doi.org/10.14778/3229863.3229867>
23. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28* (pp. 2503–2511).
24. Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. In *Proceedings of the 2015 IEEE International Congress on Big Data* (pp. 191–198). IEEE. <https://doi.org/10.1109/BigDataCongress.2015.35>
25. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–34. <https://doi.org/10.1080/07421222.1996.11518099>