Learning *from* Quality, Not *for* It: Conditioning Models on Explicit Data Quality

Mattia Sabella
Politecnico di Milano
Milan, Italy
mattia.sabella@polimi.it

ABSTRACT

The rise of Data-Centric AI has established that the quality of the data a model consumes is as decisive as the choice of model itself, refocusing research onto the systematic engineering of data quality as a primary driver of performance. Yet the dominant response to this reasoning has been to treat quality as a precondition: something diagnosed and repaired through preparation pipelines before any learning takes place, or at best inferred implicitly from a model's own training dynamics. This work questions whether explicit, measurable quality information can instead become a first-class entry to the learning process itself, so that a model is conditioned to trust reliable data and discount unreliable data as it learns. We pursue this idea across three connected fronts: designing models that consume quality metadata directly within their optimization; bringing quality-awareness into Federated Learning (FL), where data are decentralized, heterogeneous, and unevenly curated and where no preparation methodology yet exists; and finally clarifying how quality informed learning relates to classical preparation: whether it replaces, complements, or still depends on it. Early results suggest that learning directly from quality is not only feasible but often cheaper and more robust than repairing data first, opening a path toward treating data quality as something a model can reason with rather than a problem to be solved before learning begins.

VLDB Workshop Reference Format:
Mattia Sabella. Learning *from* Quality, Not *for* It: Conditioning Models on Explicit Data Quality. VLDB 2026 Workshop: VLDB Ph.D. Workshop.

1 INTRODUCTION

The diffusion of a data driven culture has made ML pipelines a backbone of organizational and scientific decision making. Within these pipelines, the quality of the data a model consumes is now understood to be as decisive as the choice of model itself. The rise of Data-Centric AI has formalized this intuition, refocusing research effort onto the systematic engineering of data quality, quantity, and reliability as the primary driver of AI system performance [3, 12, 15]. The open question is no longer whether data quality matters, but how systems should act on it. Conventionally, the answer is data preparation: quality is fixed before learning, through a personalized pipeline of cleaning, imputation, augmentation, feature engineering, scaling and standardization that turns raw inputs into ingestable model ready data [1, 7]. This stage is famously the most time consuming and the most consequential in the pipeline, and a large literature is devoted to automating it. Model-centric methods challenge the premise that data must be repaired before reaching the model, allowing training to adapt to imperfect data through mechanisms such as sample reweighting, curriculum learning, and influence based selection. However, these methods usually infer reliability from training dynamics. This work instead asks whether explicit, measurable quality metadata can be fed directly into the learning objective, so that the model conditions its updates on known data reliability. This question opens a new research direction centered on quality informed learning, in which quality is a first class input to optimization, not a property to be repaired beforehand. I use explicit quality metadata to denote measurable descriptors of data or source reliability available independently of the model's training loss, such as feature reliability, missingness, corruption patterns, client level profiles or other externally computed indicators of trustworthiness. The European BETTER[1] project provides a motivating application scenario that can work as a testbed: a decentralized healthcare data space that lets institutions securely share and analyze multi source patient data through FL across national borders, advancing research on rare diseases. FL trains models locally and aggregates only their updates, over data that are simultaneously scarce, statistically heterogeneous (non-IID), and poorly and unevenly prepared across clients. Crucially, almost everything the data-centric community knows about preparing data assumes a centralized view of the dataset that simply does not exist in FL. No data preparation methodology has yet been built specifically for the federated setting. If quality can be handled inside the learning process, then the long standing assumption that data must be cleaned before learning becomes contestable. Is classical data preparation still necessary, or does quality informed learning make it redundant? And if both have value, how should they be combined? Answering this requires holding and comparing three approaches against one another: classical preparation pipelines, explicit conditional quality informed learning, and implicitly robust models. The research problem can therefore be decomposed along three logical dimensions. The first is foundational: data quality is treated as part of the learning process rather than as a mere preprocessing concern, which requires the model architecture and learning objective to become explicitly quality driven. The second is contextual: since the limitations of classical data preparation are most apparent in distributed environments, the quality informed perspective must be studied within FL. The third is comparative:

[1] https://www.better-health-project.eu

Proceedings of the VLDB Endowment. ISSN 2150-8097.

once both classical preparation and quality informed learning are available, their relationship must be clarified to determine whether the latter replaces, complements, or still depends on the former.

This leads to three research questions:

- **RQ1.** How can explicit quality metadata be integrated into a model-centric learning objective to condition the learning process, and what are the effects on predictive performance and robustness to low quality data?
- **RQ2.** How can data preparation pipelines and quality driven mechanisms be adapted and specialized for Federated Learning environments, where data quality varies heterogeneously across clients?
- **RQ3.** Does quality informed learning act as a substitute for, or a complement to, classical data preparation and does this relationship hold consistently across both centralized and federated settings?

The paper is structured as follows: Section 2 reviews related work and isolates motivations and concrete gaps; Section 3 presents the research vision and works conducted, at a high level, together with the ongoing studies; Section 4 lays out the conclusions of the paper.

2 RELATED WORK

Beyond establishing that data quality drives model performance, data-centric AI has produced increasingly structured quality models spanning feature and dataset level dimensions (accuracy, completeness, consistency, uniqueness), profile characteristics (dimensionality, dependency, class imbalance, overlap), and bias characteristics (coverage, disparity, density) [3, 15]. A recurring finding is that a quality issue's impact is not universal but depends jointly on the data profile and the learning task [12]. Two complementary lines act on this quality without ever consuming it as an explicit input to learning. The first lets optimization absorb imperfection through training time mechanisms: sample reweighting [14], easy-to-hard curriculum learning [13], co-teaching between paired networks that exchange small loss samples [4], and influence function reweighting [14]. Each reacts to a quality signal inferred from training dynamics rather than from external metadata. The second line instead repairs data beforehand: here AutoML-style pipeline construction has reached genuine adaptive capability (e.g. CtxPipe, for instance, builds context-aware preparation pipelines from data characteristics [1, 2]) yet its surveyed cost is consistent, namely limited scalability, coverage of only a narrow set of impacting characteristics, and substantial search overhead from exploring task combinations [7]. In Federated Learning, the dominant response to non-IID data operates at the aggregation level: algorithms such as FedAvg, FedProx, FedNova, and SCAFFOLD counter the drift caused by client level system and distribution differences. Benchmarks consistently show that no single method dominates, each wins only under particular kinds and degrees of skew, and their gap narrows or inverts when several skews co-occur [6]. A more recent stream studies noisy-label and quality-aware FL aggregation [5, 9, 16], but data-quality heterogeneity across clients remains severely understudied jointly with non-IID conditions, and surveys taking a data preparation viewpoint reach the same verdict: no preparation techniques specialized for FL have yet been developed [11]. The literature then leaves four gaps:

- **Gap 1**: Model-Centric learning lacks explicit quality-aware conditioning. Existing quality aware methods rely on implicitly inferred quality signals instead of directly incorporating existing quality metadata into the learning objective.
- **Gap 2**: Data Preparation approaches lack scalable, context-adaptive framework capturing a limited set of impacting data characteristics while introducing a non negligible search overhead cost from exploring task combinations (e.g. AutoML).
- **Gap 3**: FL is rarely analyzed under joint quality and distribution heterogeneity.
- **Gap 4**: No data-preparation methodology exists for the federated setting. Clients still clean, normalize, and impute in isolation, with no coordinated, FL-specialized preparation process.

The next section turns the gaps defined into a research vision, mapping each gap to a concrete study.

3 CONTRIBUTIONS

The research follows the three-dimensional decomposition introduced above. The foundational dimension asks how a model can be designed to consume explicit quality metadata and translate it into optimization behaviour; the contextual dimension asks how quality awareness, both inside learning and at the preparation step, can be brought into FL, where no preparation methodology yet exists; and the comparative dimension asks how the resulting quality-informed learning relates to classical preparation and to implicitly noise-robust models, in centralized as well as federated settings. Within this space, I am developing a set of studies that each populate a different region of the map.

3.1 Foundational Dimension

QuAIL: Quality-Aware Inertial Learning for Robust Training with Imperfect Data. Feature level data quality is often documented in practice, whereas instance level quality annotations are rarely available. This observation motivates the core idea of the work: injecting feature-reliability priors directly into the learning process, so that optimization is guided by known feature trustworthiness rather than by the loss signal alone. Concretely, an existing model is augmented with a learnable feature modulation layer whose updates are selectively constrained by a quality dependent proximal regularizer: trustworthy features are allowed to adapt freely, while features known to be unreliable are held back toward a periodically updated anchor, producing controlled adaptation across features of differing quality without any explicit data repair or sample level reweighting. The gate acts only at the input interface, so QuAIL attaches to any differentiable tabular model as a lightweight, $O(D)$ add-on that requires no search. This is the starting point answer to **RQ1** and it fills **Gap 1**. The study targets tabular learning under structured, non-uniform corruption, since uniform robustness is what existing methods already approximate. QuAIL is evaluated on 15 OpenML-CC18 classification benchmarks across six corruption scenarios, Corruption Completely At Random (CCAR) and value dependent Corruption Not At Random (CNAR), each at 10%, 20%, and 30% overall corruption budgets against a representative set of baselines spanning three families: standard models trained on raw, corrupted

data; classical data preparation pipelines applied as preprocessing before the same learner (a search-based reimplementation of Saga++ and a feature aware custom cleaning pipeline, CP); and curriculum learning, which shapes training against noise without consuming explicit quality annotations. To compare fairly across heterogeneous datasets, methods are ranked with a Bradley–Terry Elo model fitted over head-to-head matches for each trial on five classification metrics. Across this benchmark, QuAIL obtains the top Elo rating in almost all metrics and settings, and is the closest method to the clean data reference, at essentially no additional cost beyond unmitigated training. Its advantage over the standard baseline is consistent in every configuration and widens with corruption severity, reaching +110 Elo points at 30% CCAR, so the benefit is best read as a reliability gain rather than a large jump in average accuracy. Against explicit repair, QuAIL outranks both Saga++ and CP while requiring an order of magnitude less computation (roughly 10× and 17× less CPU time, respectively), and under the harder CNAR regime cleaning can actively harm the downstream task—CP's mean F1 collapses below the unmitigated level. Curriculum learning is the strongest competitor, being the only baseline that likewise spends the quality signal on shaping learning rather than on repairing data; QuAIL's edge over it follows from placing the prior at the column granularity where corruption is structured and governance metadata actually exists, rather than collapsing a row's reliable and unreliable features into a single sampling weight. Together these results give a direct answer to **RQ1** and consolidate what was previously only a preliminary signal into a firm finding: explicit quality conditioning matches or surpasses classical data preparation in both predictive performance and computational cost [10].

Ongoing & Future Directions. Overall, QuAIL demonstrates that modeling feature reliability within the optimization process is a promising, effective, and scalable strategy for robust tabular learning under realistic data-quality constraints. It introduces minimal architectural overhead, integrates seamlessly into existing training pipelines, and lowers data preparation cost while improving predictive robustness most visibly in the value-dependent, high-severity regimes where conventional repair strategies fail. The current evaluation derives the quality prior exactly from the injected corruption, so the most immediate next step is to characterize QuAIL's behaviour when this prior is noisy or only partially correct, as real governance metadata inevitably is. Beyond this, future work will extend the approach from a single MLP backbone to more expressive architectures (e.g., Transformer-based tabular encoders) and additional data modalities, generalize it from classification to regression, replace synthetic corruption with datasets carrying real, documented defects, and adapt the mechanism to FL settings (e.g., the BETTER project) to assess whether its gains hold against traditional aggregation mechanisms.

3.2 Contextual Dimension

The foundational dimension develops quality informed learning in a centralized setting; the contextual dimension asks how it must change once that setting is removed. Everything QuAIL demonstrates today, and everything planned for its evolution has to be re-examined under the conditions that define Federated Learning, where data are decentralized, statistically heterogeneous (non-IID), and unevenly curated, and where no centralized view of the dataset ever exists. Does QuAIL adaptation into the federated regime still improve robustness and solve client heterogeneity when training is decentralized and only model updates are aggregated? This shift exposes quality sensitive decisions that have no counterpart in the centralized case. In a Federation where clients differ in both distribution and cleanliness, and where no preparation methodology yet exists, the same metadata can also govern how client contributions are aggregated and how each client's data is prepared. The studies below address these new decisions, starting with the one that most directly drives federated performance: which aggregation strategy the Federation should adopt.

Profiling-Driven Selection of Aggregation Strategies for Federated Data Spaces. Established aggregators such as FedAvg, FedProx, FedNova, and SCAFFOLD each excel only under particular heterogeneity conditions, and the margin between them narrows or even inverts when several kinds of skew co-occur [6]; in practice, however, the choice is fixed a priori and paired with a per-client preparation pipeline tuned in isolation from the rest of the Federation. This study asks whether client level quality and distributional profiles can turn that expertise dependent, quality blind choice into a quality informed one: each client shares only a local profile, never its data, and the coordinator uses these profiles to anticipate, through a learned predictor, the best performing aggregator under the current statistical and quality conditions of the Federation. Conceptually, it is the federated counterpart of context-aware pipeline construction in the centralized case (e.g., CtxPipe [2]) selecting an aggregation strategy from client profiles rather than preparation operators from data characteristics. The work gives a first answer to **RQ2** and fills **Gap 3** by characterizing FL behaviour jointly under data-quality heterogeneity and non-IID skew, and it lays a first building block toward **Gap 4**: a quality-informed decision layer around which an FL-specialized preparation pipeline can later be assembled.

Ongoing & Future Directions. The predictor must be trained on larger samples and validated at scale on unseen datasets and new clinical domains before it can be deployed in a real FL context such as the BETTER project. It should also incorporate more recent aggregation algorithms that may be inherently more robust to heterogeneity, as well as the additional model architectures under study, so that its recommendations generalize beyond the current benchmark.

Collaborative Multi-Agent Data Preparation for FL. The profiling study decides which aggregation strategy to adopt, but leaves the preparation step itself untouched: in current FL each client cleans, normalizes, and imputes in isolation, and this uncoordinated local preprocessing silently amplifies the very distributional divergence that aggregation then has to absorb. The natural continuation is to turn preparation from a fixed local step into a coordinated process, assembling the FL-specialized pipeline anticipated above. The idea is a Collaborative Multi-Agent System embedded in the federated preparation stage: each client runs specialized agents that profile their data locally and exchange only distributional metadata (feature moments, missingness statistics, label prevalence, drift indicators), never raw data, to adapt their local preprocessing policies

jointly. The governing principle is selective harmonization, correcting harmful shifts such as scale inconsistencies or measurement protocol artefacts while preserving beneficial heterogeneity that reflects genuine population differences and strengthens robustness. This reframes federated preparation from a pure model-aggregation problem into a collaborative representation alignment one, providing the most direct answer to **RQ2** and the contribution that fills **Gap 4**, the preparation methodology still missing for the federated setting.

3.3 Comparative Dimension

The comparative dimension addresses **RQ3**. A first, partial answer is already embedded in the first project: in the centralized tabular case, QuAIL has been compared to some generalizable data preparation techniques. The study extracts early preliminary signal that explicit quality conditioning can act as a substitute for preparation rather than merely a complement to it. It remains, however, only a signal: the comparison is so far confined to a single modality, one family of architectures, and the centralized regime, and it does not yet probe the federated case. For this reason, the full comparative analysis is deliberately postponed until the two ingredients that make a fair comparison possible are in place. The first is a mature, validated and extended foundational dimension, the second is an adaptive federated data preparation pipeline that is still under development, which supplies the data preparation side of the comparison with a credible federated counterpart. Only when both sides are equally well established, will the comparative study then quantify, across these settings, when quality informed learning replaces preparation, when the two are complementary and worth combining, and when classical preparation remains indispensable, closing the argument opened by the relative research questions.

3.4 Side Project

Safety-Aware Aggregation for Federated Fine-Tuning of LLMs. One exploratory study tests the same principle in a very different setting: the text modality, via Large Language Model fine-tuning, in dual-use scenarios. Here, quality is interpreted as a safety property (how hazardous a piece of training data is) rather than reliability. In a federated setup, many parties jointly train one shared model without pooling their data, so the server never sees what anyone trains on; the data filtering approach through safeguard screening removing dangerous contributions [8] cannot be applied, and a single party can quietly teach the model harmful knowledge. The response keeps this work's core move: instead of self-declaring, each client runs a shared, standardized safe scorer over its own data and shares only the resulting risk score with the server for merging contributions. The quality-safety signal can steer how updates are combined rather than how the model learns. The developed method, Federated Risk-Aware Projection (RAP-FL), uses the risk scores to estimate the direction in the model's parameter space along which hazardous knowledge is being learned, then projects each contribution away from that direction, erasing the harmful component while keeping the rest, and even recovering the useful knowledge that risky dual-use data still carries rather than discarding whole updates. Experiments showed encouraging results on WMDP benchmarks, an early sign that the principle can be applied to AI Safety challenges.

4 CONCLUSION

So far, this work provides early evidence that models can exploit explicit data-quality information during learning, reducing the need for some forms of preprocessing in select ed settings, and that the same idea can guide how federated clients are combined and prepared. The next steps are to widen this method to more data types and architectures, build the missing preparation pipeline for FL, and then test, on equal terms, whether learning from quality can stand in for, work alongside, or still need classical preparation.

REFERENCES

[1] Alvaro AA Fernandes, Martin Koehler, Nikolaos Konstantinou, Pavel Pankin, Norman W Paton, and Rizos Sakellariou. 2023. Data preparation: A technological perspective and review. *SN Computer Science* 4, 4 (2023), 425.

[2] Haotian Gao, Shaofeng Cai, Tien Tuan Anh Dinh, Zhiyong Huang, and Beng Chin Ooi. 2024. Ctxpipe: Context-aware data preparation pipeline construction for machine learning. *Proceedings of the ACM on Management of Data* 2, 6 (2024), 1–27.

[3] Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. 2023. A survey on dataset quality in machine learning. *Information and Software Technology* 162 (2023), 107268.

[4] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* 31 (2018).

[5] Xuefeng Jiang, Jia Li, Nannan Wu, Zhiyuan Wu, Xujing Li, Sheng Sun, Gang Xu, Yuwei Wang, Qi Li, and Min Liu. 2025. Fnbench: Benchmarking robust federated learning against noisy labels. *arXiv preprint arXiv:2505.06684* (2025).

[6] Zili Lu, Heng Pan, Yueyue Dai, Xueming Si, and Yan Zhang. 2024. Federated learning with non-iid data: A survey. *IEEE Internet of Things Journal* 11, 11 (2024), 19188–19209.

[7] Alhassan Mumuni and Fuseini Mumuni. 2025. Automated data processing and feature engineering for deep learning and big data applications: a survey. *Journal of Information and Intelligence* 3, 2 (2025), 113–153.

[8] Kyle O'Brien, Stephen Casper, Quentin Anthony, Tomek Korbak, Robert Kirk, Xander Davies, Ishan Mishra, Geoffrey Irving, Yarin Gal, and Stella Biderman. 2025. Deep ignorance: Filtering pretraining data builds tamper-resistant safeguards into open-weight llms. *arXiv preprint arXiv:2508.06601* (2025).

[9] Y Peng, C Wang, H Shi, R Ma, H Guan, and Hanbo Yang. 2025. Federated Learning With Client Clustering Selection and Quality-Aware Model Aggregation (2024). *IEEE Internet of Things Journal* (2025).

[10] Mattia Sabella, Alberto Archetti, Pietro Pinoli, Matteo Matteucci, and Cinzia Cappiello. 2026. QuAIL: Quality-Aware Inertial Learning for Robust Training under Data Corruption. *arXiv preprint arXiv:2602.03686* (2026).

[11] Nawraz Saeed, Mohamed Ashour, and Maggie Mashaly. 2025. Comprehensive review of federated learning challenges: A data preparation viewpoint. *Journal of Big Data* 12, 1 (2025), 153.

[12] Camilla Sancricca, Pasquale Castiglione, and Cinzia Cappiello. 2025. Exploring the Influence of Data Characteristics on Machine Learning Outcomes. In *International Conference on Business Process Modeling, Development and Support*. Springer, 227–244.

[13] Zhen Hao Wong, Hansi Yang, Quanming Yao, and Yaqing Wang. 2026. Robust heterogeneous network representation learning by multifaceted curriculum training. *Neural networks: the official journal of the International Neural Network Society* 196 (2026), 108438.

[14] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333* (2024).

[15] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *Comput. Surveys* 57, 5 (2025), 1–42.

[16] Zongxiang Zhang, Gang Chen, Yunjie Xu, Lihua Huang, Chenghong Zhang, and Shuaiyong Xiao. 2024. FedDQA: A novel regularization-based deep learning method for data quality assessment in federated learning. *Decision Support Systems* 180 (2024), 114183.