Data Contradictions Are Uncertainty, Not Noise

Adhiraj Chhoda [1]

Abstract

This position paper argues that data contradictions in ML training sets should be treated as uncertainty to be quantified, not noise to be cleaned away. The standard ML pipeline treats data quality as preprocessing: find inconsistencies, pick a repair, train on the result. This workflow silently discards information. When two hospital records disagree on a patient's diagnosis, that disagreement reflects genuine ambiguity, and a model trained on one arbitrary resolution is overconfident in exactly the cases where it should be uncertain. We show that the multiplicity of valid data repairs maps naturally to prediction uncertainty: train on each repair, measure prediction disagreement, and the result is an informative confidence signal that requires no architectural changes and no Bayesian machinery. As an illustrative proof of concept on Adult Income, models trained on different repair strategies disagree with clean-baseline predictions on 2.3% of test instances, concentrated in the subgroups affected by the original contradiction. Systematic validation across datasets is future work. Of 101 data-cleaning-for-ML papers surveyed by Côté et al. (Côté et al., 2024), zero treat repair non-uniqueness as an uncertainty signal. Of 51 tabular-ML papers at NeurIPS and ICML 2024–2025, not one engages with the 26-year database theory literature on consistent query answering. The field cleans when it should reason.

1. Introduction

Every ML data pipeline makes the same implicit bet: that data inconsistencies are random errors, best handled by picking a single repair and moving on. Delete the contradictory rows, majority-vote the conflicting labels, trust the most recent source. Train. Ship.

[1]Thomas Jefferson High School for Science and Technology. Correspondence to: Adhiraj Chhoda <2027achhoda@tjhsst.edu>.

Accepted at the ICML 2026 Workshop on AI for Good.

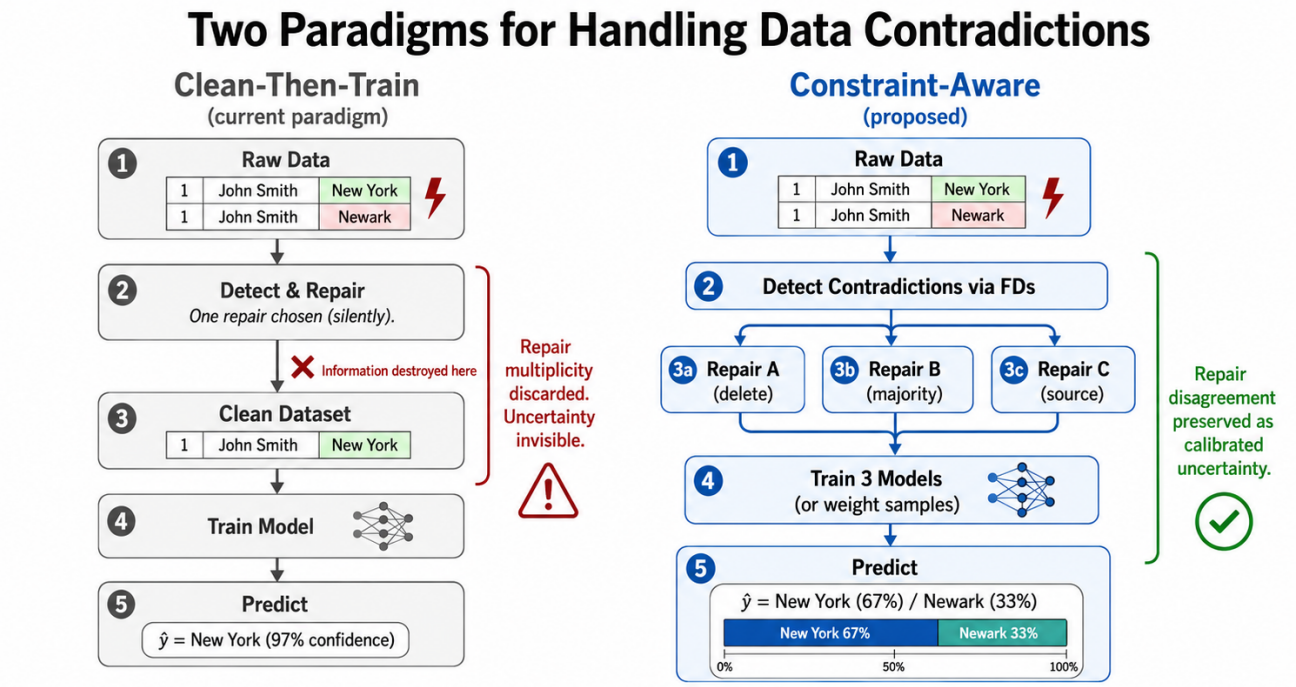
This bet is wrong, and the consequences compound silently. When ProPublica's COMPAS recidivism analysis applied an inconsistent two-year window rule, it inflated recidivism from 36.2% to 45.1% (Barenstein, 2019). Hundreds of fairness studies inherited the error. A different but equally valid resolution would have produced different recidivism rates, different fairness conclusions, and different policy recommendations. That variation is not noise. It is uncertainty that the pipeline buried.

The database community recognized this problem in 1999. Arenas, Bertossi, and Chomicki (Arenas et al., 1999) introduced consistent query answering (CQA): instead of picking one repair, reason over all possible repairs simultaneously. An answer that holds under every valid repair is certain; an answer that changes across repairs is uncertain by exactly the amount it varies. 26 years of follow-up work produced mature algorithms for constraint discovery (Huhtala et al., 1999; Papenbrock et al., 2016), violation detection (Chu et al., 2013), and repair-aware query answering (Bertossi, 2011; Wijsen, 2012). The ML community has adopted none of it. Of 51 tabular-ML papers from NeurIPS and ICML 2024–2025, not one mentions functional dependencies, consistent query answering, or any formal notion of cross-row consistency.

This paper argues that repair multiplicity is not a nuisance to be resolved in preprocessing. It is an uncertainty signal that belongs in the model's predictions. We make three contributions: (1) we show that repair non-uniqueness maps naturally to prediction uncertainty, connecting CQA to ML calibration without requiring Bayesian inference (Section 4); (2) a classification of discovered data dependencies by epistemic content, spanning semantic invariants, artifacts, leakage signals, and bias proxies (Section 4.2); and (3) a constraint-aware pipeline framework validated on standard benchmarks (Section 5).

2. The Clean-Then-Train Assumption

The standard ML preprocessing pipeline handles missing values, duplicates, outliers, format normalization, and label noise. TFDV (Baylor et al., 2017) checks feature presence, type, domain bounds, and distribution drift. Great Expectations (Great Expectations, 2020) and Pandera (Bantilan, 2020) validate column-level schemas. Sculley et al. (Sculley



	Clean-Then-Train	Constraint-Aware
Contradictions	Noise (remove)	Uncertainty (quantify)
Repair choice	Silent, arbitrary	Explicit, all enumerated
Model output	Point prediction	Prediction + uncertainty
Information	Destroyed at cleaning	Preserved as signal

Figure 1. Two paradigms for handling data contradictions. **Left:** Clean-then-train silently picks one repair and destroys information. **Right:** Constraint-aware enumerates all valid repairs and preserves their disagreement as prediction uncertainty.

et al., 2015) document these as critical ML infrastructure; NADEEF (Dallachiesa et al., 2013) provides rule-based cleaning but still commits to a single repaired output.

These tools address real problems. But they share an assumption so pervasive it is invisible: that one correct version of the data exists and the goal of cleaning is to find it. Every tool produces a single repaired dataset. None asks what would change if the repair had gone differently. None reports how many alternative repairs were possible, or whether downstream predictions depend on the choice. The multiplicity of valid repairs is discarded before training begins.

Cross-row logical consistency, the property that two rows agreeing on attribute X must also agree on attribute Y (a functional dependency), admits multiple valid resolutions when violated. No standard ML tool reasons about this multiplicity. Seedat et al.'s DC-Check (Seedat et al., 2023) propose a data-centric AI checklist but do not include cross-row integrity constraints. Neutatz et al. (Neutatz et al., 2021) argue for cleaning for ML rather than before it, but still assume a single cleaned output.

The distinction between noise and contradiction is central. ML treats data problems as noise: random perturbations that regularization smooths away under the assumption that errors are i.i.d. and cancel in expectation. Database theory treats them as contradictions: logical inconsistencies where the repair choice carries information. Knight (Knight, 1921) distinguished *risk* (known distribution) from *uncertainty* (unknown distribution); Ellsberg (Ellsberg, 1961) showed agents are averse to the ambiguity itself. ML's noise framing assumes risk. Data contradictions are Knightian uncertainty: the pipeline does not know which repair is correct, and no amount of data resolves the ambiguity because the disagreement is in the world, not in the measurement.

This parallels Duhem-Quine underdetermination (Duhem, 1954; Quine, 1951): when a dataset violates a constraint, the data alone cannot determine which row is wrong. The multiplicity of valid repairs is the underdetermination, and discarding it discards precisely the information that would let a model report "I don't know."

2.1. No Tool Reports Repair Sensitivity

Even the most capable production ML validation systems stop at single-dataset output.

TFDV (Baylor et al., 2017) validates schemas, checks for feature drift, and enforces domain bounds. It produces one validated dataset. It does not ask how many alternative valid datasets exist. DataPerf (Mazumder et al., 2023) benchmarks data-centric operations (selection, slicing, debugging) but includes no benchmark for repair sensitivity. CleanML (Li et al., 2021) benchmarks cleaning impact on ML classification; its "inconsistency" category means string normalization ("CA" vs. "California"). It evaluates cleaning methods by which one produces the highest accuracy, not by how much predictions vary across equally valid repairs.

Our survey confirms this absence quantitatively: of the 51 tabular-ML papers and 101 cleaning papers described in Section 1, none engages with cross-row integrity constraints. Seven of 51 tabular papers use the word "constraint," but exclusively for model-level properties: adversarial attack validity, fairness criteria, or synthetic data generation specifications. The repair choice is universally treated as a preprocessing detail, not as a source of predictive uncertainty.

3. Repair Choice Changes Predictions

3.1. Central Experiment: Three Repairs, Three Models

We present a single illustrative proof-of-concept, not broad validation: systematic evaluation across datasets and contradiction types is future work. We simulated a realistic data integration scenario on Adult Income (Ding et al., 2021) (Table 1). Two independently clean sources (70% of the dataset each, 40% overlap) are joined after 5% of overlapping rows receive corrupted `education_num` values, violating the FD `education` → `education_num`. We then applied three standard repair strategies and trained identical XGBoost (Chen & Guestrin, 2016) classifiers on each result.

The finding that matters is not that dirty data hurts accuracy. In fact, the naive join (with violations intact) scores higher than the clean baseline (87.9% vs. 86.6%), because duplicated overlap rows add training data. You cannot tell from accuracy alone which repair was right. The disagreement is what matters: majority vote and source priority repairs both restore the FD, both achieve identical aggregate accuracy, and yet each repair-trained model disagrees with the clean baseline on 2.3% of test predictions. These are individuals who receive different predictions depending on which repair the pipeline chose.

That 2.3% is not noise. It is uncertainty that the clean-then-train pipeline buries. Subset repair is worse still: it deletes all rows in FD-violating groups (99.5% of the data), collapsing the model to 75.5% accuracy. The repair strategy is not a preprocessing detail. It is a hyperparameter with consequences for individual predictions, and no standard pipeline reports it. These three policies probe sensitivity to a few representative repair choices; they do not enumerate the full space of valid repairs that the consistent-query-answering framework ultimately calls for, and treating that full space is itself future work.

Table 1. Data integration experiment on Adult Income: two independently clean sources (70% each, 40% overlap) joined after 5% of overlapping rows receive corrupted `education_num` values. FD: `education` → `education_num`. XGBoost classifier (200 trees, depth 5).

Condition	Violations	Accuracy	EO Gap
Clean (single source)	0	86.6%	0.090
Naive join (dirty)	45,000	87.9%	0.065
Subset repair (delete)	0	75.5%	0.049
Majority vote repair	0	87.9%	0.078
Source priority repair	0	87.9%	0.078

Table 2. Functional dependency verification on standard ML benchmarks. All expected FDs hold with zero violations.

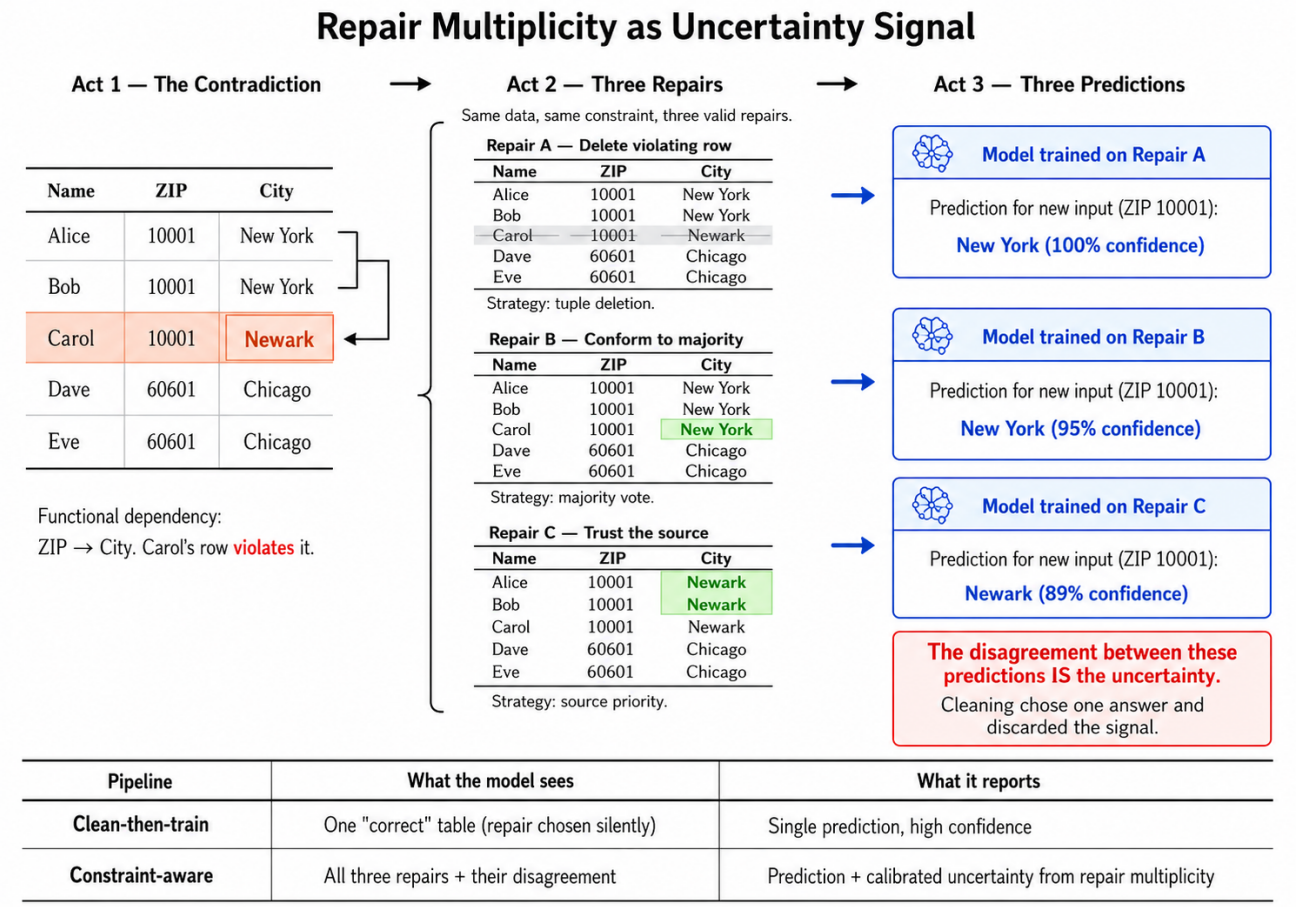
Dataset	FD	Rows	Viol.
Adult	edu → edu_num	48,842	0
COMPAS	decile → score_text	7,214	0
ACS (CA)	ST → DIVISION	391,171	0
ACS (CA)	CIT → NATIVITY	391,171	0
ACS (CA)	POBP → WAOB	391,171	0

3.2. Where Contradictions Enter

We verified that Adult, COMPAS, and Folktables satisfy expected FDs with zero violations (Table 2). But no benchmark documents this. Satisfaction is inherited from upstream processing, an accident, not a guarantee.

Production is different. Data integration produces conflicting values when sources disagree on the same entity. Synthetic data generation (SMOTE, CTGAN (Xu et al., 2019)) does not enforce FD consistency; a generated row may pair "Bachelors" with education_num = 7 (the correct value is 13), creating a contradiction that the original data did not contain. Deep learning imputation (Yoon et al., 2018) produces single-point estimates without preserving cross-row constraints; GAIN and similar models predict each cell independently, ignoring that two cells in the same column must agree when their key attributes match. Temporal drift creates new violations: a hospital that changes its scanner type midway through a longitudinal study breaks the FD hospital_id → scanner_type for all retrospective analyses.

Each of these failures is invisible without constraint specification, and each produces a family of possible repairs yielding potentially different models.



Pipeline	What the model sees	What it reports
Clean-then-train	One "correct" table (repair chosen silently)	Single prediction, high confidence
Constraint-aware	All three repairs + their disagreement	Prediction + calibrated uncertainty from repair multiplicity

Figure 2. Repair multiplicity as uncertainty signal. The same contradiction (ZIP 10001 mapped to both New York and Newark) admits three valid repairs. Each produces a different prediction. That disagreement is the uncertainty that cleaning discarded.

3.3. Why Contradictions Are Not Noise

If training data violates $X \to Y$, the model must either learn a non-deterministic mapping (wrong), arbitrarily pick one (ordering bias from SGD), or average conflicting targets (matching no real data point). Regularization does not help: dropout does not resolve contradictory supervision; weight decay does not pick the correct value. Northcutt et al. (Northcutt et al., 2021) show higher-capacity models amplify systematic errors; data poisoning research (Biggio et al., 2012) confirms systematic corruption is qualitatively different from random noise.

This framing connects to partial identification in causal inference (Manski, 2003): when the data cannot uniquely determine a parameter, the honest output is a set of estimates, not a point estimate. Our repair set is analogous: each valid repair induces a different model, and the range of predictions across repairs is the identified set for that instance. Robust risk minimization (Ben-Tal et al., 2009) optimizes for worst-case performance over a set of data-generating distributions; optimizing over the repair set is a natural instance of this framework, connecting data-quality uncertainty to established robust learning theory.

3.4. Case Studies: Buried Repair Choices

Barenstein (Barenstein, 2019) found that ProPublica's COMPAS construction inflated recidivism from 36.2% to 45.1%. The temporal window constraint was never formalized. Every subsequent fairness analysis, including impossibility results (Chouldechova, 2017) and equalized odds proposals (Hardt et al., 2016), inherited one repair without knowing alternatives existed. A different resolution of the temporal window produces different recidivism rates, different fairness metrics, and different policy recommendations.

MIMIC-III (Johnson et al., 2016) contains physiological impossibilities (sodium 84 mEq/L, glucose 1601 mg/dL) that violate clinical denial constraints. MIMIC-IV explicitly forgoes cleaning (Johnson et al., 2023). Each impossible value admits multiple repairs, and each changes the training distribution differently. Hundreds of clinical ML papers (Weiskopf & Weng, 2013; Kahn et al., 2016; Hripcsak

& Albers, 2013) use MIMIC; none reports which implicit repair their preprocessing chose.

These are not edge cases. Data integration is the default in production ML. When two hospital systems merge patient records, when a recommendation engine joins click logs from web and mobile, when a government agency links administrative datasets across bureaus: contradictions arise at the join boundary, and the resolution is always silent.

4. From Repairs to Uncertainty

4.1. Repair Non-Uniqueness as Epistemic Signal

A functional dependency (FD) $X \to Y$ states that if two rows agree on attributes X, they must agree on attribute Y. When a dataset violates an FD, a *repair* is a modified dataset that restores satisfaction while staying "close" to the original under some distance metric. Three standard families exist: subset repair (delete rows), update repair (change cells (Bohannon et al., 2005; Kolahi & Lakshmanan, 2009)), and cost-based repair (minimize edit distance). Critically, repairs are generally not unique. The same violation set admits multiple valid resolutions.

Our experiment (Section 3) employs three repair strategies with well-defined semantics: *subset repair* deletes all rows in any group containing a violation, equivalent to the standard subset-repair semantics of Arenas et al. (Arenas et al., 1999) restricted to FDs. *Majority vote* replaces each violating cell with the modal value in its group, a specific instance of cost-based update repair (Bohannon et al., 2005) that minimizes Hamming distance under uniform cell costs. *Source priority* prefers the designated primary source in all conflicts, a provenance-based repair common in data integration (Lenzerini, 2002). Both majority vote and source priority are minimal under their respective cost functions and produce unique outputs for the FDs tested. We emphasize that probing these three policies measures sensitivity to a handful of repair choices, which is distinct from reasoning over the full space of valid repairs that CQA prescribes; enumerating or sampling that full space is the natural next step. In general, cost-based repairs may not be unique; for richer repair sets, Beskales et al. (Beskales et al., 2010) provide efficient sampling algorithms that approximate the space of minimal repairs with formal coverage guarantees.

CQA (Arenas et al., 1999) defines a *consistent answer* as one that holds across all possible repairs. We propose extending this to ML: a prediction is *repair-consistent* if it is the same regardless of which valid repair is used for training. A prediction that varies across repairs is uncertain by exactly the amount it varies, localized to exactly the subgroups affected by the original contradiction.

This connection is natural but unexploited. CQA provides what ML uncertainty quantification methods seek: a way to distinguish predictions the data supports from predictions that depend on arbitrary choices. Unlike Bayesian approaches (Blundell et al., 2015), it requires no prior over model parameters. Unlike deep ensembles (Lakshminarayanan et al., 2017), it varies the data rather than the model, capturing a fundamentally different source of uncertainty. Unlike conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021), it produces uncertainty estimates that are semantically grounded in specific data contradictions, not just in exchangeability assumptions. The formal tools for reasoning over repair sets (complexity results (Lopatenko & Bertossi, 2007; Livshits & Kimelfeld, 2017), practical algorithms (Bertossi, 2011; Chomicki, 2007), repair sampling (Beskales et al., 2010)) are mature. What is missing is the recognition that repair multiplicity is an uncertainty signal, not a nuisance.

Repair-based uncertainty is complementary to, not a replacement for, established uncertainty quantification methods. Deep ensembles and MC dropout (Gal & Ghahramani, 2016) capture model uncertainty (sensitivity to initialization and architecture); conformal prediction provides distribution-free coverage guarantees under exchangeability. Repair-based uncertainty captures a distinct epistemic source: sensitivity to preprocessing choices that standard methods hold fixed. The two compose naturally: training an ensemble on each of k repairs decomposes total uncertainty into model-level and data-level components. Whether repair disagreement correlates with prediction error more strongly than ensemble disagreement, and whether repair-based abstention improves selective prediction, are open questions that require head-to-head evaluation on datasets with ground-truth contradictions.

4.2. Classifying Dependencies by Epistemic Content

Not all discovered FDs are equal. We classify them by the kind of information their violations carry (Table 3). This classification determines which repair multiplicity carries information worth propagating to predictions and which is noise or a confound.

Semantic invariants (patient_id $\to$ date_of_birth) produce genuine ambiguity: when two hospital records disagree on a patient's diagnosis, the repair set encodes real-world uncertainty that the model should reflect. These violations produce the most informative uncertainty signal. *Dataset artifacts* (hospital_id $\to$ scanner_type) produce spurious variation that should be removed rather than quantified. A model that learns scanner type as predictive of outcomes has learned a site-specific confound, not a clinical relationship. *Label leakage* (claim_code $\to$ denial_label) carries no genuine uncertainty; the FD is a deterministic encoding that inflates accuracy without improving generalization. *Bias*

Table 3. Functional dependency classification by epistemic content.

FD Type	Violation Meaning	Action
Semantic invariant	Genuine ambiguity: sources disagree	Track multiplicity as uncertainty
Dataset artifact	Spurious correlation	Flag as confound; remove
Label leakage	Deterministic encoding	Remove or restructure
Bias proxy	May amplify disparate impact	Audit for fairness

proxies (ZIP $\rightarrow$ income_bracket) require fairness audit before enforcement: the constraint may be statistically valid but ethically problematic.

Armstrong's transitivity axiom ($X \rightarrow Y \wedge Y \rightarrow Z \Rightarrow X \rightarrow Z$) can expose hidden leakage chains: if `patient_id` $\rightarrow$ `hospital_id` and `hospital_id` $\rightarrow$ `region`, then `patient_id` $\rightarrow$ `region` is implied. If `region` correlates with the outcome, the model learns site identity as signal, a leakage path invisible without constraint reasoning.

5. A Repair-Aware ML Pipeline

We propose a five-stage pipeline organized around a single principle: repair choices that affect predictions should be visible in the model's uncertainty estimates.

Stage 1 is constraint specification and discovery. Constraints come from domain experts, algorithmic discovery (TANE (Huhtala et al., 1999), HyFD (Papenbrock et al., 2016)), and schema-implied constraints. The output is a constraint set Σ with confidence annotations.

Stage 2 classifies constraints. The taxonomy of Section 4.2 determines which constraints carry epistemic content. This stage requires human oversight: only domain expertise can distinguish a genuine invariant from a dataset artifact. Concretely, domain experts familiar with the data-generating process perform the semantic-invariant / artifact / leakage / bias-proxy classification, and when annotators disagree on a constraint's type, the disagreement is resolved by defaulting to the more conservative class (flag for audit) so that no potentially consequential constraint is silently enforced.

Stage 3 detects violations and produces a report card. All rows are checked against validated constraints to produce a Constraint Report Card: violation rate per constraint, affected rows, severity classification, and an estimate of the number of distinct valid repairs. This is analogous to model cards (Mitchell et al., 2019), but for data consistency and repair sensitivity.

We implemented a prototype of this tool and ran it on three standard benchmarks: Adult (48,842 rows, 14 columns), COMPAS (7,214 rows, 21 columns), and ACS California 2022 (391,171 rows, 26 columns). Table 4 shows the results. The tool auto-discovered 2 FDs in Adult, 8 in COMPAS (including a deterministic encoding: age $\rightarrow$ age_cat, which our tool initially flags as a potential bias proxy due to the protected attribute but which is more precisely classified as label leakage, a lossless discretization of a continuous feature that carries no genuine uncertainty), and 11 in ACS. All genuine FDs hold with zero violations on these curated benchmarks, confirming that the risk lies in production pipelines, not benchmark curation. We stress that the report card is currently demonstrated only on curated benchmarks where all genuine FDs hold with zero violations, so the format is shown but its informativeness under real contradictions is not yet established; evaluating the card on integrated or otherwise dirty datasets with nonzero violation rates is future work.

Table 4. Constraint Report Card on standard ML benchmarks. FDs discovered via pairwise approximate dependency detection ($\leq$1% violation threshold). All genuine FDs hold with zero violations on these curated benchmarks.

Dataset	Rows	FDs	Sem.	Artif.
Adult	48,842	2	0	2
COMPAS	7,214	8	1	6
ACS (CA)	391,171	11	7	4

Stage 4 samples repairs. Multiple plausible repairs are generated rather than committing to one. Subset repair, majority vote, and source priority are the simplest strategies. For richer repair sets, Beskales et al. (Beskales et al., 2010) provide efficient sampling with formal guarantees.

Stage 5 audits repair sensitivity. Train on $k \geq 2$ sampled repairs. For each test instance, report the fraction of repairs that produce the same prediction. Instances where all k models agree are repair-consistent; instances where they disagree carry repair-induced uncertainty. This maps CQA's "certain answer" to ML uncertainty quantification, providing an epistemic signal that is (a) free of distributional assumptions, (b) localized to affected subgroups, and (c) complementary to model-based uncertainty from ensembles or Bayesian inference.

Scalability deserves direct discussion. FD discovery is exponential in the number of attributes. HyFD (Papenbrock et al., 2016) handles roughly 50 attributes on 1M records. Modern feature tables with 100–1000+ columns make full discovery intractable, and embeddings fall outside classical FD theory entirely. Two facts blunt this objection. First, checking known FDs is $O(n)$ per FD; the bottleneck is discovery, not verification. If practitioners specify constraints (as they already specify schemas), checking costs nothing. Second, our proposal targets structured/relational data with

meaningful attributes: exactly the domain where tabular ML (XGBoost (Chen & Guestrin, 2016), TabNet (Arık & Pfister, 2021), tabular transformers (Gorishniy et al., 2021; Rubachev et al., 2022)) operates and where standard benchmarks live (Grinsztajn et al., 2022; Borisov et al., 2022). Checking 5 domain-expert FDs is strictly better than checking none. The repair-sensitivity audit (Stage 5) requires training k models, but $k = 3$–5 suffices for a useful signal, comparable to the cost of cross-validation.

Limitations: calibration is not yet established. The central empirical claim this position rests on—that repair disagreement is an *informative*, calibrated uncertainty signal rather than mere variance—is not yet demonstrated, and we flag this as the paper's primary open question and primary future-work item. We do not yet report calibration metrics (ECE, Brier score, risk-coverage curves) for the repair-based signal. The decisive experiment is a head-to-head comparison, on datasets with ground-truth contradictions, of whether repair disagreement correlates with prediction error *more strongly than ordinary ensemble disagreement*, and whether repair-based abstention improves selective prediction (risk-coverage curves). Until that evidence exists, the proof-of-concept in Section 3 should be read as establishing that repair choice *changes* predictions, not yet that the resulting disagreement is well-calibrated.

5.1. Proposed Reporting Standard

Any ML paper using tabular or relational data should include:

1. What constraints were specified or discovered?
2. What was the violation rate per constraint?
3. What repair strategy was applied (if any)?
4. How sensitive are results to repair choice? (Train on ≥ 2 repairs, report prediction disagreement rate.)
5. Were implied FDs (via Armstrong's axioms) checked for leakage?

This is not onerous. Items 1–3 require a single linear scan per FD. Item 4 costs one additional training run. Item 5 is a graph closure computation. The total cost is less than one round of hyperparameter tuning.

6. Alternative Views

We consider six objections, beginning with their strongest formulations.

One might argue that statistical robustness already handles contradictions. It does not. Regularization (dropout, weight decay, label smoothing) smooths random noise (Guo et al., 2017). Contradictions are not random: the same input maps to conflicting outputs, and SGD converges to one (wrong) mapping rather than averaging them away. More importantly, regularization provides no signal about which predictions are downstream of the contradiction. Repair-based uncertainty does: disagreement is localized to exactly the affected subgroups.

Another objection holds that ensembles already capture this uncertainty. Ensembles (Lakshminarayanan et al., 2017) vary the model while holding data fixed; repair-based uncertainty varies the data while holding the model fixed. An ensemble trained on the same (arbitrarily repaired) dataset will agree on all predictions affected by the repair choice, because every member sees the same resolution. The two approaches are complementary, not redundant.

The data-centric AI movement (Ng, 2021) addresses data quality but focuses on label quality, feature engineering, and data selection. It has no theory of cross-row consistency. Datasheets (Gebru et al., 2021) and Data Cards (Pushkarna et al., 2022) have no field for constraint satisfaction. The gap is not awareness of data quality but the absence of a formal notion of consistency and the recognition that resolution is non-unique.

A cost objection notes that the repair space is combinatorially large (k^n for n violations with k resolutions each). Exhaustive enumeration is unnecessary. Beskales et al. (Beskales et al., 2010) provide efficient sampling with formal guarantees. In our experiment, $k = 3$ strategies already revealed 2.3% prediction disagreement at a cost comparable to cross-validation. For high-stakes applications (lending, criminal justice, clinical diagnosis), this cost is negligible relative to the information gained.

The strongest objection is that constraints may encode bias. If a dataset reflects historical discrimination, its FDs encode it. But tracking repair multiplicity for bias-proxy constraints reveals exactly where discriminatory structure constrains predictions and where less biased alternatives exist. Our classification (Table 3) flags bias proxies as a separate category requiring fairness audit. On COMPAS, the Constraint Report Card automatically flagged age $\rightarrow$ age_cat with a WARNING; manual review reclassified it as a deterministic encoding rather than a bias proxy, illustrating why the classification stage requires human oversight. Constraint discovery surfaces bias; it is not an unconditional enforcement mandate.

A final objection: privacy and anonymization intentionally break constraints. This is correct. Constraints should be checked before anonymization and re-specified afterward. The Constraint Report Card documents which FDs were broken by design versus by accident, a distinction that is itself metadata worth preserving.

7. Related Work

Database repair theory originates with Arenas et al. (Arenas et al., 1999), who introduced CQA; the subsequent literature provides complexity results (Lopatenko & Bertossi, 2007; Livshits & Kimelfeld, 2017), practical algorithms (Bertossi, 2011; Chomicki, 2007), repair sampling (Beskales et al., 2010), causality connections (Bertossi, 2020), and data exchange (Fagin et al., 2005). CQA reasons over query answers, not model predictions; extending "consistent answer" to "repair-consistent prediction" is the bridge we propose. Ilyas and Chu (Ilyas & Chu, 2019) provide a textbook treatment of the repair landscape. Fan and Geerts (Fan & Geerts, 2012) establish the theoretical foundations of data quality management. The bridge to ML practice does not exist.

Uncertainty quantification in ML relies on deep ensembles (Lakshminarayanan et al., 2017), MC dropout (Gal & Ghahramani, 2016), and conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021), all of which vary the model or inference procedure while holding data fixed. Repair-based uncertainty varies the data while holding the model fixed, capturing a complementary epistemic source. The two compose: train an ensemble on each of k repairs to decompose total uncertainty into model and data components.

Data-centric AI research documents data cascades (Sambasivan et al., 2021), pervasive label errors (Northcutt et al., 2021), benchmarks for data operations (Mazumder et al., 2023), and documentation standards (Gebru et al., 2021). None include constraint specification or repair-sensitivity reporting. The data-centric AI movement (Ng, 2021; Zha et al., 2023; Whang et al., 2023) correctly shifts attention from models to data but provides no formal semantics for "data correctness."

Data cleaning systems such as HoloClean (Rekatsinas et al., 2017), ActiveClean (Krishnan et al., 2016), BoostClean (Krishnan et al., 2017), CleanML (Li et al., 2021), and Raha (Mahdavi et al., 2019) appeared at database venues, not ML venues. All produce one repaired dataset; none reports how predictions vary across the repair space. FD discovery algorithms (TANE (Huhtala et al., 1999), DFD (Abedjan et al., 2014), HyFD (Papenbrock et al., 2016), CFDs (Fan et al., 2008)) provide the constraint specification our pipeline requires.

Salimi et al. (Salimi et al., 2019) show causal-constraint repair can remove discrimination. Gururangan et al. (Gururangan et al., 2018) and McCoy et al. (McCoy et al., 2019) show models learn artifacts as signal. Lenzerini (Lenzerini, 2002) and Doan et al. (Doan et al., 2012) survey data integration, the primary source of contradictions in production pipelines.

The model multiplicity literature (Marx et al., 2020; Black et al., 2022) studies the set of near-optimal models that a dataset supports. Our proposal is complementary: model multiplicity varies the model while holding data fixed; repair multiplicity varies the data while holding the model fixed. The product of the two gives the full space of predictions consistent with the data and the learning algorithm.

8. Open Questions

Several questions remain open. First, calibration: does repair-based disagreement correlate with actual prediction error more strongly than ensemble disagreement? Answering this requires datasets with ground-truth contradictions and known correct labels, a benchmark that does not yet exist. Second, interaction with label noise: Northcutt et al. (Northcutt et al., 2021) find pervasive label errors in standard benchmarks, and repair multiplicity compounds with label uncertainty. The joint effect on calibration is unstudied. Third, extension beyond FDs: denial constraints (Chu et al., 2013), conditional FDs (Fan et al., 2008), and inclusion dependencies are all sources of repair multiplicity, but their interaction with ML pipelines is unexplored. Fourth, the connection to distributionally robust optimization: if the repair set defines a family of training distributions, what does the minimax-optimal model over this family look like, and does it differ from training on any single repair?

9. Conclusion

Stop treating data repair as preprocessing. When a dataset admits multiple valid repairs and the choice of repair changes predictions, that variation is not a nuisance to be resolved before training. It is an uncertainty signal that belongs in the model's output.

The database community formalized this insight in 1999 (Arenas et al., 1999): a consistent answer is one that holds across all possible repairs. Twenty-six years later, the ML community has adopted none of this machinery. Of 101 data-cleaning-for-ML papers surveyed by Côté et al. (Côté et al., 2024), zero treat repair non-uniqueness as an uncertainty signal. Of 51 tabular-ML papers at NeurIPS and ICML 2024–2025, zero engage with consistent query answering.

The decisive open question is whether repair disagreement is a *calibrated* uncertainty signal—one that tracks prediction error better than ordinary ensemble disagreement. Establishing this through risk-coverage and calibration experiments on datasets with ground-truth contradictions is the primary future-work item, and the experiment most likely to turn this position into accepted practice.

Three concrete actions:

1. Report repair sensitivity alongside accuracy. Train

on ≥ 2 repairs, report the prediction disagreement rate. In our experiment, this cost was comparable to cross-validation and revealed that 2.3% of predictions depended on the repair choice.

2. Include constraint satisfaction status in dataset documentation. Datasheets (Gebru et al., 2021) and Data Cards (Pushkarna et al., 2022) should report specified constraints, measured violation rates, and the space of valid repairs. This is computationally trivial: a single linear scan per FD.
3. Adopt the vocabulary: "contradiction" not "noise," "repair" not "cleaning," "repair-consistent" not "clean." These are not synonyms. They imply different interventions and different epistemologies.

The field cleans when it should reason. Data contradictions are uncertainty, not noise. Report them as such.

Impact Statement

This paper argues for propagating data-quality uncertainty into model predictions rather than silently resolving it during preprocessing. The primary societal benefit is increased transparency in high-stakes ML systems (criminal justice, healthcare, lending), where repair choices currently affect individual predictions without being documented. We see no specific negative consequences requiring discussion beyond those inherent in any work on ML data quality.

References

Abedjan, Z., Schulze, P., and Naumann, F. DFD: Efficient functional dependency discovery. In *CIKM*, pp. 949–958, 2014.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. In *arXiv preprint arXiv:2107.07511*, 2021.

Arenas, M., Bertossi, L., and Chomicki, J. Consistent query answers in inconsistent databases. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 68–79, 1999.

Arık, S. Ö. and Pfister, T. TabNet: Attentive interpretable tabular learning. In *AAAI*, pp. 6679–6687, 2021.

Bantilan, N. Pandera: Statistical data testing toolkit. `https://pandera.readthedocs.io`, 2020.

Barenstein, M. ProPublica's COMPAS data revisited. *arXiv preprint arXiv:1906.04711*, 2019.

Baylor, D., Breck, E., Cheng, H.-T., Fiedel, N., Foo, C. Y., Haque, Z., Haykal, S., Ispir, M., Jain, V., Koc, L., et al. TFX: A TensorFlow-based production-scale machine learning platform. In *KDD*, pp. 1387–1395, 2017.

Ben-Tal, A., Ghaoui, L. E., and Nemirovski, A. Robust optimization. *Princeton University Press*, 2009.

Bertossi, L. *Database Repairing and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool, 2011.

Bertossi, L. Database repairs and causality. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pp. 63–74, 2020.

Beskales, G., Ilyas, I. F., and Golab, L. Sampling the repairs of functional dependency violations under hard constraints. *Proceedings of the VLDB Endowment*, 3(1–2): 197–207, 2010.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. In *ICML*, pp. 1467–1474, 2012.

Black, E., Raghavan, M., and Barocas, S. Model multiplicity: Opportunities, concerns, and solutions. In *FAccT*, pp. 850–863, 2022.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *ICML*, pp. 1613–1622, 2015.

Bohannon, P., Fan, W., Flaster, M., and Rastogi, R. A cost-based model and effective heuristic for repairing constraints by value modification. In *SIGMOD*, pp. 143–154, 2005.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *KDD*, pp. 785–794, 2016.

Chomicki, J. Consistent query answering: Five easy pieces. In *ICDT*, pp. 1–17, 2007.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *Big Data*, volume 5, pp. 153–163, 2017.

Chu, X., Ilyas, I. F., and Papotti, P. Holistic data cleaning: Putting violations into context. In *ICDE*, pp. 458–469, 2013.

Côté, P.-O., Nikanjam, A., Ahmed, N., Humeniuk, D., and Khomh, F. Data cleaning and machine learning: A systematic literature review. *Automated Software Engineering*, 31(1):54, 2024. doi: 10.1007/s10515-024-00453-w.

Dallachiesa, M., Ebaid, A., Eldawy, A., Elmagarmid, A., Ilyas, I. F., Ouzzani, M., and Tang, N. NADEEF: A commodity data cleaning system. In *SIGMOD*, pp. 541–552, 2013.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, 2021.

Doan, A., Halevy, A., and Ives, Z. *Principles of Data Integration*. Morgan Kaufmann, 2012.

Duhem, P. *The Aim and Structure of Physical Theory*. Princeton University Press, 1954. Translated by Philip P. Wiener.

Ellsberg, D. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

Fagin, R., Kolaitis, P. G., Miller, R. J., and Popa, L. Data exchange: Semantics and query answering. *Theoretical Computer Science*, 336(1):89–124, 2005.

Fan, W. and Geerts, F. *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. Morgan & Claypool, 2012.

Fan, W., Geerts, F., Jia, X., and Kementsietsidis, A. Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems*, 33 (2):1–48, 2008.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059, 2016.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting deep learning models for tabular data. In *NeurIPS*, 2021.

Great Expectations. Great expectations: Always know what to expect from your data. https://greatexpectations.io, 2020.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS Datasets and Benchmarks*, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. Annotation artifacts in natural language inference data. In *NAACL*, pp. 107–112, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS*, pp. 3315–3323, 2016.

Hripcsak, G. and Albers, D. J. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association*, 20(1):117–121, 2013.

Huhtala, Y., Kärkkäinen, J., Porkka, P., and Toivonen, H. TANE: An efficient algorithm for discovering functional and approximate dependencies. *The Computer Journal*, 42(2):100–111, 1999.

Ilyas, I. F. and Chu, X. Data cleaning. *ACM Books*, 2019.

Johnson, A. E. W., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Moody, B., Gow, B., wei H. Lehman, L., et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10 (1):1, 2023.

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., Johnson, S. G., et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *eGEMs*, 4(1), 2016.

Knight, F. H. *Risk, Uncertainty and Profit*. Houghton Mifflin, 1921.

Kolahi, S. and Lakshmanan, L. V. S. On approximating optimum repairs of functional dependency violations. In *ICDT*, pp. 53–62, 2009.

Krishnan, S., Wang, J., Wu, E., Franklin, M. J., and Goldberg, K. ActiveClean: Interactive data cleaning for statistical modeling. *Proceedings of the VLDB Endowment*, 9 (12):948–959, 2016.

Krishnan, S., Franklin, M. J., Goldberg, K., and Wu, E. BoostClean: Automatic error detection and repair for machine learning. *arXiv preprint arXiv:1711.01299*, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, pp. 6402–6413, 2017.

Lenzerini, M. Data integration: A theoretical perspective. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 233–246, 2002.

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., and Zhang, C. CleanML: A study for evaluating the impact of data cleaning on ML classification tasks. In *ICDE*, pp. 13–24, 2021.

Livshits, E. and Kimelfeld, B. Counting and enumerating (preferred) database repairs. In *PODS*, pp. 289–301, 2017.

Lopatenko, A. and Bertossi, L. Complexity of consistent query answering in databases under cardinality-based and incremental repair semantics. In *ICDT*, pp. 179–193, 2007.

Mahdavi, M., Abedjan, Z., Fernandez, R. C., Madden, S., Ouzzani, M., Stonebraker, M., and Tang, N. Raha: A configuration-free error detection system. In *SIGMOD*, pp. 865–882, 2019.

Manski, C. F. *Partial Identification of Probability Distributions*. Springer, 2003.

Marx, C., du Pin Calmon, F., and Ustun, B. Predictive multiplicity in classification. In *ICML*, pp. 6765–6774, 2020.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., García, W. G., Mattson, M., Whatmough, P., Warden, M., Diamos, G., Bird, C., et al. DataPerf: Benchmarks for data-centric AI development. In *NeurIPS*, 2023.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL*, pp. 3428–3448, 2019.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *FAccT*, pp. 220–229, 2019.

Neutatz, F., Chen, B., Abedjan, Z., and Wu, E. From cleaning before ML to cleaning for ML. In *IEEE Data Engineering Bulletin*, volume 44, pp. 24–41, 2021.

Ng, A. Data-centric AI. Landing AI Blog, 2021.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Datasets and Benchmarks*, 2021.

Papenbrock, T., Ehrlich, J., Marten, J., Neuber, T., Pelber, J.-P., Schäfer, N., Schönberg, M., Schüler, J., Steinbach, B., Sträler, A., et al. A hybrid approach to functional dependency discovery. In *SIGMOD*, pp. 821–833, 2016.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *FAccT*, pp. 1776–1826, 2022.

Quine, W. V. O. Two dogmas of empiricism. *The Philosophical Review*, 60(1):20–43, 1951.

Rekatsinas, T., Chu, X., Ilyas, I. F., and Ré, C. HoloClean: Holistic data repairs with probabilistic inference. *Proceedings of the VLDB Endowment*, 10(11):1190–1201, 2017.

Rubachev, I., Alekberov, A., Gorishniy, Y., and Babenko, A. Revisiting pretraining objectives for tabular deep learning. In *ICML*, 2022.

Salimi, B., Rodriguez, L., Howe, B., and Suciu, D. Interventional fairness: Causal database repair for algorithmic fairness. In *SIGMOD*, pp. 1389–1406, 2019.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *CHI*, pp. 1–15, 2021.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., and Dennison, D. Hidden technical debt in machine learning systems. In *NeurIPS*, pp. 2503–2511, 2015.

Seedat, N., Imrie, F., and van der Schaar, M. DC-Check: A data-centric AI checklist to guide the development of reliable machine learning systems. *arXiv preprint arXiv:2211.05764*, 2023.

Vovk, V., Gammerman, A., and Shafer, G. Algorithmic learning in a random world. *Springer*, 2005.

Weiskopf, N. G. and Weng, C. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

Whang, S. E., Roh, Y., Song, H., and Lee, J.-G. Data collection and quality challenges in deep learning: A data-centric AI perspective. In *The VLDB Journal*, volume 32, pp. 791–813, 2023.

Wijsen, J. Certain conjunctive query answering in first-order logic. *ACM Transactions on Database Systems*, 37(2): 1–35, 2012.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. Modeling tabular data using conditional GAN. In *NeurIPS*, 2019.

Yoon, J., Jordon, J., and van der Schaar, M. GAIN: Missing data imputation using generative adversarial nets. In *ICML*, pp. 5689–5698, 2018.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., and Hu, X. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 2023.