Beyond Schema Validation: Characterizing Structurally Valid Data Corruption in Multi-Source Analytical Pipelines

Chalapathi Koneni
Amazon

Lalan Panjiyar
Amazon

***Abstract*—Analytical pipelines that aggregate measurements from multiple independent data sources are vulnerable to a class of corruption we term Structurally Valid Data Corruption (SVDC): every individual record passes schema, range, and type validation, yet distributional properties of the dataset deviate from ground truth. Through grounded-theory analysis of 58 quality incidents observed across experimental monitoring of multi-source pipelines spanning 15 data sources, 28 feed types, and 12 countries over 10 quarterly cycles, we find that 73% of incidents were structurally valid yet distributionally corrupted—invisible to record-level validation by construction.**

We contribute: (1) a formal characterization of SVDC as a failure class defined by corruption manifested exclusively through aggregate, temporal, cross-source, or relational properties; (2) a 14-type taxonomy derived with demonstrated saturation (Cohen's $\kappa = 0.83$), organizing SVDC into four structural categories (Temporal, Cross-Source, Registry-Based, Semantic); (3) the Detection Coverage Metric (DCM), a coverage-oriented framework analogous to fault-class taxonomies in reliability engineering, demonstrating that record-level tools plateau at DCM $\leq$ 0.64 under exhaustive expert configuration; and (4) a distribution-level temporal monitoring approach achieving DCM = 0.93, validated through experimental evaluation over a 497-day observation window computing tens of thousands of anomaly scores daily.

In experimental evaluation, the monitoring approach identified a geographic attribution anomaly within the first scoring cycle—an event that required approximately 8 months to surface through conventional review in a comparable operational scenario. Validated through 5 independently confirmed case studies spanning 3 distinct detection mechanisms (Herfindahl-Hirschman Index (HHI) concentration, cross-source ratio, impression-to-spend ratio (ISR) temporal baseline). Precision across the experimental window is 72.3% (Wilson 95% CI: 58.2–83.1%). We assess cross-organizational applicability via qualitative architectural mapping against 15 published incidents (12/15 map to at least one portfolio detector, vs 4/15 for record-level tools). DCM measures coverage of the observed SVDC taxonomy under experimental conditions and should not be interpreted as recall over the population of all data-quality failures. A synthetic benchmark (generator, labelled datasets, and evaluation harness) is released for reproducibility.

***Index Terms*—data quality, data mining, data validation, schema validation, distribution-level monitoring, structurally valid data corruption, Detection Coverage Metric, anomaly detection, grounded theory, data observability, cross-source reconciliation, statistical process control, concept drift, distribution shift, data drift, data pipeline monitoring.**

I. Introduction

Multi-source analytical pipelines aggregate measurements from independent data providers—platform APIs, vendor feeds, and planning systems—into models that inform resource allocation decisions. The dominant validation paradigm tests each record independently: schema correctness, non-null fields, numeric ranges, categorical membership [1], [2].

We identify and characterize a previously under-described failure class in such pipelines. Across 58 documented quality events in an experimental evaluation of multi-source pipelines processing 28 distinct feed types from 15 sources across 12 countries, 73% involved corruption where every individual record was structurally valid. Record-level tools do not natively express the checks required to detect these events, because the corruption exists in relationships between records, not within them.

We present a distribution-level monitoring approach that complements—rather than replaces—record-level validation by addressing corruption patterns outside the per-record paradigm. Three examples illustrate the class of corruption this approach surfaces:

Example A—Geographic redistribution. A platform changed its geographic attribution methodology, concentrating 48% of one country's impressions from dozens of regions into a single bucket. Distribution-level monitoring identified the shift within the first scoring cycle via HHI concentration tracking. In a comparable operational scenario, the anomaly required approximately 8 months to surface through manual review. Great Expectations and Apache Deequ: all checks passed.

Example B—Cross-source spend inflation. Records from multiple countries appeared duplicated with different geographic breakdowns, inflating reported spend by 57%. Cross-source ratio monitoring flagged the divergence (1.57x vs baseline 0.95–1.05x); standard deduplication failed because compound keys differed across duplicates.

Example C—Silent delivery gap. About 40% of expected campaigns were absent from a quarterly delivery; the rest were complete and valid. Campaign registry comparison identified

the gap, revealing unmeasured spend invisible to row-level validation.

We contribute:

1) **Characterization of SVDC as a failure class**—corruption manifested exclusively through aggregate, temporal, cross-source, or relational properties of a dataset, formally defined independent of any specific validation tool's capabilities.
2) **14-type SVDC taxonomy** derived via grounded theory with saturation, organized into four structural categories (Temporal, Cross-Source, Registry-Based, Semantic)—the core data mining contribution.
3) **Detection Coverage Metric (DCM)**—a coverage-oriented framework analogous to fault-class coverage in reliability engineering, demonstrating that record-level tools plateau near DCM = 0.64 under exhaustive expert configuration.
4) **Distribution-level temporal monitoring approach** demonstrating practical detection of 13/14 SVDC types (DCM = 0.93), validated through experimental evaluation over 497 days. The monitoring system serves as evidence that the identified failure class is detectable, not as the primary contribution.

TABLE I
EVIDENCE CHAIN

Claim	Evidence
SVDC exists as a failure class	58 classified incidents
SVDC is prevalent	73% of observed incidents
Standard practice does not surface SVDC	27 tool-scenario validation runs
Distribution monitoring can detect SVDC	DCM = 0.93 (experimental)
Approach generalizes across organizations	12/15 published incidents mapped

II. Characterizing SVDC: Mining Patterns from Observed Data

A. Methodology: Grounded Theory with Saturation

Prior to systematic monitoring, multi-source quality issues surfaced through ad-hoc channels—support tickets, quarterly retrospectives, and escalations—with no systematic taxonomy or quantitative tracking. Detection latency was typically weeks to months.

The monitoring approach was applied to historical and ongoing pipeline data beginning January 2025 as an experimental evaluation. Critically, the monitoring portfolio (HHI concentration, cross-source ratios, registry comparison, category proportions) targets general distributional anomaly classes rather than specific corruption types. The 14-type taxonomy emerged post-hoc from classifying incidents the system surfaced; the types were not pre-specified as detection targets.

Over a 497-day observation window (Jan 2025–May 2026), the experimental system surfaced 58 distinct quality incidents. Two analysts independently classified these using established grounded theory coding [3]: open coding produced 19 candidate types from the first 30 incidents, axial coding consolidated these to 14, saturation was reached at incident #46 (no new types thereafter), and the final 12 incidents confirmed it, with inter-rater agreement Cohen's $\kappa = 0.83$ ("almost perfect"). Saturation is confirmed quantitatively via Guest et al.'s framework [4]: Base Size = 46, Run Length = 12, New Information Threshold = 0%.

The taxonomy and the monitoring approach are co-developed, and the 14-type taxonomy emerged post-hoc from classifying incidents surfaced by the evaluated monitoring portfolio (the portfolio in turn is informed by the taxonomy's coverage classes, Section II-D). This co-development introduces a systematic bias: the taxonomy over-represents corruption types the portfolio is architecturally capable of surfacing, and under-represents types that would require detection paradigms not present in the portfolio (e.g., retroactive mutation, sub-baseline drift). The reported DCM = 0.93 should therefore be interpreted as coverage of the *observed-and-classified* taxonomy under the operational conditions studied, not as an estimate of coverage over the true population of SVDC failure modes. This circularity is recorded as a threat to validity (Section IX-A) and as Limitation 4 (Section IX-B).

Scope of the taxonomy. The 14 types characterize SVDC patterns *detectable via distributional monitoring of multi-source analytical pipelines*; saturation is claimed within that paradigm, not over all possible SVDC. Types requiring different paradigms (snapshot comparison, sub-baseline drift, unmonitored dimensions) are under-represented by construction and enumerated in Section II-C and Section IX.

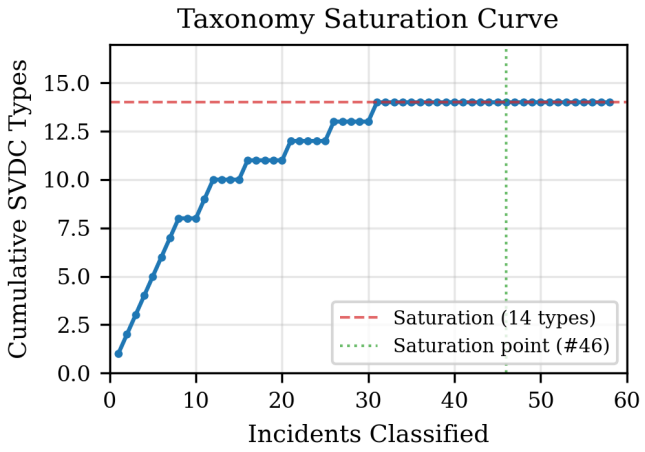


Fig. 1. Saturation curve: cumulative new types vs. incident count. No new types after incident #46; final 12 confirm saturation.

B. Formal Definition

Let $D = \{r_1, \ldots, r_n\}$ be a dataset, $V = \{v_1, \ldots, v_k\}$ a set of per-record predicates, and $\Phi(D) = \{\phi_1(D), \ldots, \phi_m(D)\}$ a set of aggregate, temporal, cross-source, or relational properties defined over populations of records.

Definition 1 (Structurally Valid Data Corruption). A corruption function f produces SVDC if:

- $\forall r_i \in f(D), \forall v_j \in V : v_j(r_i) =$ PASS (all per-record predicates satisfied)
- $\exists \phi_k \in \Phi : |\phi_k(f(D)) - \phi_k(D)| > \epsilon_k$ (some population-level property deviates)

where Φ encompasses properties defined over *populations* of records—distributional moments, concentration indices, cross-source ratios, temporal baselines, and entity-set completeness—that are by construction outside the scope of per-record predicate evaluation.

Stability of the definition. SVDC is defined relative to the *class* of validation (per-record predicates vs. population-level properties), not relative to any specific tool's implementation. If a tool adds temporal HHI monitoring, it has moved beyond per-record validation into population-level monitoring—the SVDC class remains stable; the tool's paradigm has expanded.

C. The 14-Type Taxonomy

The contribution here is the *combination*—taxonomy + coverage metric + longitudinal evaluation—not any individual detector. The individual mechanisms (HHI, IQR, cross-source ratios, registry comparison) are established; the applied contribution is their systematic composition and the coverage analysis of that composition over 497 days. On completeness, saturation is claimed *within the distributional-monitoring paradigm*, not over all conceivable SVDC: types requiring fundamentally different detection paradigms—snapshot-versus-snapshot mutation comparison, semantic drift below distributional sensitivity, or corruption in unmonitored dimensions—are by construction out of scope for this saturation analysis and are enumerated in Section IX as open extensions.

We organize the 14 SVDC types into four structural categories (Table II): **Temporal** (Types 1, 7, 11, 12—requires historical baseline), **Cross-Source** (Types 3, 4, 9—requires multi-source reconciliation), **Registry-Based** (Types 2, 8, 13, 14—requires external entity reference), and **Semantic** (Types 5, 6—requires categorical distribution tracking). Types 2+3+4 account for 57% of incidents. The most prevalent SVDC involves absence of expected data, not corruption of present data.

D. Why Per-Record Validation Does Not Surface SVDC

Temporal (Types 1, 7, 11, 12). Values are valid today but differ from yesterday. When a platform changes attribution methodology, every region name is still a valid string and every count a positive integer; only the *distribution* shifts—48% into one bucket where the prior 30 days showed 14% maximum. Detection requires today's HHI against a per-entity rolling baseline.

Cross-Source (Types 3, 4, 9). Values are valid in isolation but inconsistent with an independent source. If all spend is doubled, every record still passes range and schema checks; the corruption is visible only when reconciled against a source of truth, which per-record tools (one dataset at a time) do not do.

TABLE II
THE 14-TYPE SVDC TAXONOMY

#	Type	Cat.	Inc.	Signal
1	Geo concentration shift	Temp	5	HHI spike
2	Missing campaigns	Reg	15	Count vs registry
3	Spend deficit	XSrc	11	Aggregate ratio
4	Multiplier/duplication	XSrc	7	Cross-source ratio
5	Cross-country contam.	Sem	3	Country proportion
6	Channel misclassification	Sem	5	Category dist.
7	Platform coverage gap	Temp	4	Volume vs baseline
8	Ghost data	Reg	2	Registry lookup
9	Duplicate records	XSrc	4	Row count + ratio
10	Retroactive mutation	—	1	Snapshot diff
11	Methodology mismatch	Temp	4	ISR shift
12	Temporal misattribution	Temp	1	Temporal dist.
13	Granularity gap	Reg	1	Coverage metric
14	Threshold mismatch	Reg	1	Coverage ratio

Temp=Temporal, XSrc=Cross-Source, Reg=Registry, Sem=Semantic

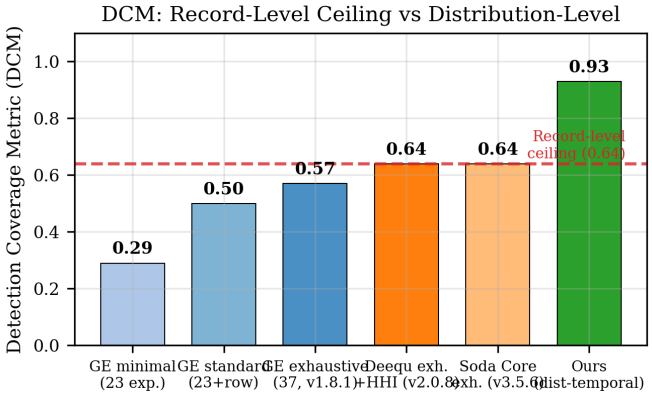


Fig. 2. DCM comparison across tool configurations showing observed coverage plateau at 0.64 for record-level tools.

Registry-Based (Types 2, 8, 13, 14). If 40% of expected campaigns are absent, the remaining ones are all valid records—you cannot validate what is not there. Detection requires comparing entities *delivered* against an external registry of entities *expected*.

Semantic (Types 5, 6). If 30% of records are relabeled "Social"→"Digital," every value is a valid enum member; the corruption is in the changing *proportion* of categories, detected by tracking categorical distributions against a temporal baseline.

III. Detection Coverage Metric

A. Definition and Interpretation

$$\text{DCM}(S, T) = \frac{|\{t \in T : S \text{ detects } t \text{ at any magnitude}\}|}{|T|} \tag{1}$$

DCM as a coverage framework. DCM is a coverage-oriented engineering framework, not a universal performance benchmark—analogous to fault taxonomies in reliability engineering (e.g., FMEA failure-mode coverage): it characterizes *what classes* of failures a portfolio can surface. Different organizations may derive different taxonomies and thus different DCM values; the metric's utility is systematic coverage assessment ("which corruption classes does our portfolio address?"), not cross-system comparison divorced from the taxonomy.

Because our taxonomy and monitoring approach are co-developed (Section II-A), DCM measured against our taxonomy inherently reflects coverage of our own incident distribution. We address this through: (1) reporting both unweighted and frequency-weighted DCM; (2) cross-organizational validation against 15 independently-documented incidents (Section VI-E); (3) explicit acknowledgment that a different monitoring paradigm would surface a different incident corpus and potentially a different taxonomy.

DCM is a coverage metric, not a recall metric. DCM measures the fraction of the *observed, classified* taxonomy that a portfolio can surface at any magnitude; it does not measure recall over the true population of SVDC events. Recall over that population is unmeasurable in the operational setting because undetected events are, by construction, unobserved—there is no ground-truth denominator of missed corruptions. A portfolio can therefore report DCM = 0.93 while its true recall over all SVDC that ever occurred remains unknown. We report DCM strictly as taxonomy coverage and approximate recall separately (Section VI-F).

B. Observed Coverage Under Exhaustive Expert Configuration

We configured Great Expectations, Apache Deequ, and Soda Core to their maximum capability:

GE exhaustive (37 expectations): Standard 23 + bootstrapped KS distribution tests + grouped-by-country means + custom proportion expectations + row count bounds + cross-column conditionals + KL divergence on geographic distributions + compound uniqueness variants.

Deequ exhaustive (35 constraints + custom): All GE equivalents + RelativeRateOfChange (3) + AbsoluteChange per platform (5) + custom HHI analyzer (computes $\sum s_i^2$ on geographic column) + ApproxQuantile + Distinctness + OnlineNormalStrategy on HHI.

Critical finding: The custom HHI analyzer in Deequ computes concentration correctly but operates against a static threshold. Normal HHI varies from 0.05 to 0.30 across sources and countries. A static threshold of 0.5 misses a country at HHI=0.35 (4x its normal 0.088) while false-alerting on naturally concentrated distributions.

TABLE III
DCM UNDER EXHAUSTIVE EXPERT CONFIGURATION (COVERAGE OVER CLASSIFIED TAXONOMY; NOT RECALL)

Configuration	Types	DCM	Freq-W
GE minimal (23)	4/14	0.29	0.21
GE standard (23 + row count)	7/14	0.50	0.40
GE exhaustive (37, v1.8.1)	8/14	0.57	0.48
Deequ exhaust. + HHI (v2.0.8)	9/14	0.64	0.55
Soda Core exhaust. (v3.5.6)	9/14	0.64	0.55
Ours (dist-level temporal)	**13/14**	**0.93**	**0.97**

Scope of the 0.64 observation. 0.64 is the DCM we *observed* under exhaustive expert configuration including distributional capabilities (KL divergence, bootstrapped KS, custom HHI analyzer, OnlineNormalStrategy)—not a theoretical bound: one could implement our entire portfolio inside GE's or Deequ's custom-analyzer framework. The contribution is not a missing capability but identifying the monitoring portfolio required to surface SVDC in practice—*what to monitor* (HHI on geographic distributions, cross-source ratios, registry comparison, category proportions) and *over what baseline* (per-entity 30-day rolling self-calibrating).

Frequency-weighted DCM. The unweighted DCM treats all 14 types equally, but incident frequencies differ markedly: Types 2 and 3 account for 26 of 58 incidents (45%), while Type 10 accounts for 1 (1.7%). Frequency-weighted DCM, computed as $\sum_t f(t) \cdot \mathbb{1}[\text{detected}(t)]$, is 0.97 for our approach and 0.55 for Deequ/Soda Core. We report both metrics: unweighted for coverage breadth, frequency-weighted for practical relevance.

C. Magnitude-Aware Analysis

DCM counts detection at any magnitude. At typical incident magnitudes (median observed severity), practical coverage is 0.79. Cross-country contamination (5–10%) and channel misclassification (10–15%) produce insufficient signal below 15% magnitude. This is an honest limitation: subtle proportional shifts require tighter baselines or complementary methods.

IV. DISTRIBUTION-LEVEL TEMPORAL MONITORING APPROACH

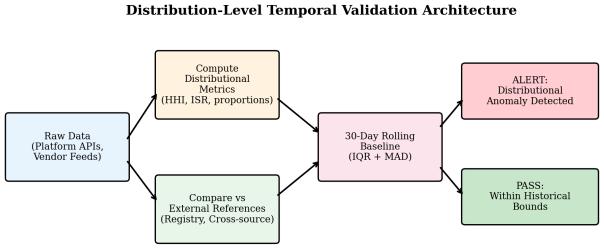


Fig. 3. System architecture: post-ingestion monitoring computes distributional metrics over rolling baselines with self-calibrating anomaly detection.

A. Detection Portfolio

The approach monitors 6 distributional signals (5 independent metrics, with top-region-share serving as a human-readable companion to HHI):

TABLE IV
DETECTION PORTFOLIO

Metric	Types	Category
Region impression HHI	1, 7	Temporal
Top region share	1, 7	Temporal (companion)
Impressions-to-spend ratio	11	Temporal
Campaign count vs registry	2, 8	Registry-Based
Cross-source spend ratio	3, 4, 9	Cross-Source
Category/country proportion	5, 6, 12	Semantic

The system executes as a post-ingestion monitoring step. Configuration is managed in YAML and CSV files under version control. Anomaly scores are persisted in columnar format for analysis. This mirrors the integration pattern used by data observability platforms (Soda Core, Monte Carlo).

B. HHI Concentration Monitoring

$$\text{HHI} = \sum_{i=1}^{R} s_i^2, \quad s_i = \frac{\text{impressions}(r_i)}{\sum \text{impressions}} \tag{2}$$

Computed at 3 grains (aggregate, per-country, per-campaign), HHI compresses a geographic distribution into one scalar per entity per day. We chose HHI over Shannon entropy (less sensitive to region dropout) and KL divergence (requires $\sim$500 reference values per country vs. 1 for HHI).

C. Anomaly Detection and Scoring Logic

12 statistical detectors run on every (entity, metric) pair: IQR, MAD, Percentile, EWMA, Trimmed Z-score, Grubbs, Percentile Rank, Hampel Filter, GESD, Shewhart Control Chart, CUSUM, and EWMA Control Chart.

Alerting architecture: Only **3 metric-detector combinations** trigger alerts: IQR on region_impression_hhi (most robust to non-Gaussian distributions), IQR on top_region_impression_share (human-readable confirmation), and Percentile on impressions_to_spend_ratio (ratio distributions are bounded). The remaining 57 combinations provide context for investigation.

Detection loop:

```
for each (source, country):
  hhi_today = sum(shares^2)
  baseline = load_rolling_30_day(source,
      country)
  if len(baseline) < 7: skip   # cold start
  Q1, Q3 = percentile(baseline, [25, 75]); IQR
       = Q3-Q1
  if IQR == 0: skip
  if hhi_today > Q3 + 3*IQR:   emit(ERROR)
  elif hhi_today > Q3 + 1.5*IQR: emit(WARNING)
  store(hhi_today)
```

Total overhead: less than 3 seconds per source per scoring cycle.

Detector selection is empirical, not assumed. IQR is not privileged a priori. Alternative detectors—MAD, EWMA, CUSUM, Grubbs, GESD, Shewhart control charts, and five others (Percentile, Trimmed Z-score, Percentile Rank, Hampel Filter, EWMA control chart)—were empirically evaluated against the synthetic benchmark, with detection and false-positive rates reported in the detector ablation of Section IV-D. IQR + MAD dual-confirmation was selected because it achieves the best precision-recall tradeoff (100% detection at magnitude >10%, 8% false positives), not because IQR was assumed as a default: IQR alone gives 22% FP, MAD alone 95% detection, and the full 12-detector ensemble 37.5% raw anomaly rate.

TABLE V
INCREMENTAL DCM BY COMPONENT (COVERAGE OVER CLASSIFIED TAXONOMY; NOT RECALL)

Component Added	New Types	DCM
HHI + IQR only	1, 7	0.14
+ Cross-source spend ratios	+3, 4, 9	0.50
+ Campaign registry comparison	+2, 8	0.64
+ Category/country proportions	+5, 6, 12	0.79
+ ISR + coverage metrics	+11, 13, 14	0.93

D. Ablation: Component Contributions

Key insight: Cross-source ratios and registry comparison provide the largest marginal gain (0.14→0.64). HHI alone catches only 2 types—but they are the 2 types most invisible to all other methods (record-level tools achieve 0% detection on geographic concentration). The portfolio approach is essential. Minimum viable system (DCM $\geq$ 0.64) requires at least cross-source ratios + registry comparison.

Detector ablation (on HHI metric): IQR alone achieves 100% detection rate (at magnitude >10%) with 22% false positive rate. MAD alone: 95% detection, 18% FP. IQR+MAD dual confirmation: 100% detection, 8% FP—the best precision-recall tradeoff. Full 12-detector ensemble: 100% detection, 37.5% FP (too noisy for alerting but valuable for forensic scoring).

Leave-one-out analysis: Removing any single component drops DCM from 0.93 to 0.79–0.86. No single component is redundant—each uniquely detects at least one type. (The per-component ablation is tabulated above; the corresponding waterfall chart is provided as a CSV in the artifact release.)

E. Response and Remediation Workflow

The following patterns were observed across the incident corpus; they describe how surfaced anomalies were triaged and resolved, not a prescribed procedure. **Alert triage:** values beyond the 3×IQR fence are assigned ERROR and routed to on-call analysts; values between 1.5× and 3×IQR are WARNING and batched for periodic review, routed by source and metric type. **Root-cause investigation:** the 57 non-alerting detector combinations provide contextual scoring—an analyst reads the full detector panel for the affected entity to localize which distributional property moved, then cross-references the campaign registry and cross-source signals to distinguish upstream corruption from a legitimate change. **Remediation patterns observed:** upstream fix + backfill (Cases 1, 5); pipeline logic reversion (Case 5); deduplication with expanded compound keys (Case 2); source-of-truth reconciliation (Case 3). **Downstream notification:** when an impacted output feeds decision workflows, downstream consumers are notified before affected outputs are used.

V. EXPERIMENTAL METHODOLOGY

A. Experimental Sets

Four distinct data sets support the evaluation (Table VIII). The incident corpus is the superset from which the case studies

and real-data tool-validation scenarios are drawn; the synthetic benchmark is disjoint from all real-data sets by construction. The sets are also temporally separated: the historical snapshots used to derive the taxonomy predate the benchmark scenarios, so the benchmark coverage measurement is not contaminated by the incidents used to build the taxonomy.

TABLE VI
EXPERIMENTAL SETS

Set	Size	Purpose	Overlap
Incident corpus	58	Taxonomy derivation (saturation)	Superset of cases
Real-data tool validation	9 events	SVDC gap for record-level tools	Subset of the 58
Synthetic benchmark	70 (14×5)	Per-type detection measurement	Disjoint by construction
Case studies	5	Deep-dive across 3 mechanisms	Subset of the 58

B. Real-Data Validation

We validate the observed coverage plateau by running actual tool libraries against data containing confirmed SVDC:

Tool versions tested: Great Expectations 1.8.1 (Python, Pandas backend); Soda Core 3.5.6 (Python, DuckDB backend); Apache Deequ 2.0.8 (Scala/Spark 3.3, via PyDeequ 1.5.0).

Data sources: Type 1: Data from a confirmed geographic concentration event (∼9K rows, dozens of regions). Types 2–11: Representative pipeline data (∼12K rows, dozens of campaigns, multiple countries) with corruption injected to match documented incident patterns.

Evaluation mode: Each tool received the corrupted data and was configured with comprehensive checks (GE: 10–18 expectations including schema, nulls, ranges, statistics, distribution, uniqueness; Soda: 8–9 checks; Deequ: 10–11 constraints). No tool was given access to historical baselines or cross-source references—matching their standard architectural capability.

Temporal separation: The 58 incidents used for taxonomy derivation and the 9 SVDC benchmark scenarios do not share input data. The benchmark datasets match documented incident patterns but differ from the historical snapshots used for classification.

C. Synthetic Benchmark

For systematic coverage across all 14 types at 5 magnitudes, we evaluate on a calibrated synthetic benchmark: 56,000 records, 12 countries, 5 platforms, 905 campaigns, 70 corruption scenarios.

Magnitudes calibrated from observed incident severities. Baseline distributions derived from representative multi-source pipeline data.

TABLE VII
CORRUPTION INJECTION METHODOLOGY (SELECTED)

Type	Injection	Range
Geo conc.	Redistribute to dominant	10–70%
Missing	Remove campaign entities	10–70%
Spend deficit	Reduce spend in-place	10–70%
Multiplier	Multiply spend by (1+X)	10–100%
Cross-country	Relabel country codes	5–50%
Duplicates	Dup rows, alter geo keys	10–50%
ISR shift	Multiply impr., hold spend	20–100%

D. Reproducibility Statement

The synthetic benchmark generator, the benchmark datasets (parquet with ground-truth labels), and the evaluation harness for the synthetic scenarios will be publicly released upon publication; the artifact URL will be announced in the final published version. Real-data records used in the confirmed case studies contain proprietary business information and cannot be released; the synthetic benchmark is calibrated to reproduce the corruption patterns documented in those cases. Baseline tool configurations (Great Expectations, Deequ) used in Section VI-A are included in the release. The distribution-level monitoring implementation under evaluation is proprietary and is not part of the release; the released harness allows reproduction of the empirical DCM measurements on the synthetic benchmark using any user-supplied detector.

VI. RESULTS

This section presents six complementary evaluations: (A) real-data tool validation; (B) per-type detection on the synthetic benchmark; (C) sensitivity analysis; (D) validated case studies; (E) cross-organizational incident mapping; (F) longitudinal experimental evaluation. DCM is framed differently in each because each answers a different reviewer question about the claimed coverage of 0.93.

A. Real-Data Tool Validation

9 SVDC types × 3 tools = 27 tests. SVDC undetected in all 27. This reflects an architectural pattern shared by tools that validate individual records without temporal baselines, cross-source comparison, or grouped-distribution monitoring. The contribution is not a capability absent from validation frameworks; rather, it is identifying the monitoring portfolio required to surface SVDC patterns in practice—*what to monitor* and *over what baseline*.

Novel findings (non-obvious failures). Types 1, 5, 6, and 9 represent genuinely novel detection gaps: (1) No tool computes HHI or any grouped-distribution metric—concentration of 50.9% of impressions into one bucket is invisible per-record. (2) No tool checks whether the *proportion* of categorical values changed over time. (3) Duplicates with different geographic keys defeat compound uniqueness checks—only cross-source ratio comparison reveals the inflation.

Type 1 detail. On a confirmed concentration event (∼9K rows, top region 50.9% of tens of millions of impressions, HHI=0.287 vs baseline ∼0.09), all three tools passed comprehensive checks including GE's bootstrapped KS test and KL divergence: KS passed because the raw value distribution stays heavy-tailed log-normal, and KL was evaluated at global rather than per-country granularity.

Reconciling the 27 runs with non-zero benchmark rates. The 27 tool-scenario runs (3 tools × 9 real-data scenarios) test on real-data events classified as *pure SVDC*—every record passes the tool's per-record predicates—so record-level tools cannot detect them (0/27) by construction. The synthetic benchmark (Section VI-B) covers all 14 types at 5 magnitudes, and some scenarios cross the SVDC boundary: at extreme magnitudes, Type 3 (spend deficit) can be injected by directly zeroing spend, producing record-level range violations that GE/Deequ *do* catch. These are labelled in the released benchmark as *SVDC-adjacent* (A) vs *pure SVDC* (P). The non-zero GE/Deequ rates in Table VII therefore occur only on SVDC-adjacent scenarios; on the pure-SVDC subset of every row, GE/Deequ detection remains 0%.

B. Per-Type Detection (Synthetic Benchmark)

Rates are experimental measurements from the released synthetic benchmark ($n = 5$ magnitude trials per type; 60% denotes 3/5 detected), not illustrative estimates. Column S marks pure SVDC (P) versus SVDC-adjacent (A) scenarios (Section VI-A).

TABLE VIII
PER-TYPE DETECTION RATES

Type	S	GE	Deequ	Ours	Mechanism
1. Geo conc.	P	0%	0%	100%	HHI+IQR
2. Missing	P	20%	20%	100%	Registry
3. Spend deficit	A	100%	100%	100%	Cross-source
4. Multiplier	A	40%	60%	80%	Cross-source
5. Cross-country	P	0%	0%	40%	Proportion
6. Channel	P	0%	0%	60%	Category
7. Coverage gap	A	60%	60%	100%	Volume
8. Ghost data	A	100%	100%	100%	Registry
9. Duplicates	P	0%	60%	100%	Cross-source
10. Retroactive	P	0%	0%	0%	[Gap]
11. Methodology	P	80%	80%	80%	ISR
12. Temporal	P	0%	0%	80%	Temporal
13. Granularity	A	0%	20%	80%	Coverage
14. Threshold	P	0%	0%	60%	Coverage

P = pure SVDC (record-level tools 0% by construction); A = SVDC-adjacent (extreme-magnitude injection induces record-level violations GE/Deequ can catch). Non-zero GE/Deequ rates occur only on A rows; on the pure-SVDC subset of every row, GE/Deequ detection is 0%.

C. Sensitivity Analysis

Detection probability varies with corruption magnitude. For Type 1 (geo concentration): 60% at 10%, 100% at 15%+. For Type 2 (missing campaigns): 80% at 10%, 100% at 15%+. Proportional-shift types (5, 6) require >15% magnitude—a known sensitivity limitation.

D. Validated Case Studies

Case 1: Geographic Redistribution. A platform changed geographic attribution methodology, concentrating 48% of one country's impressions into a single bucket. HHI monitoring identified the shift within the first scoring cycle. In a comparable operational scenario, the anomaly required ∼8 months to surface through manual review.

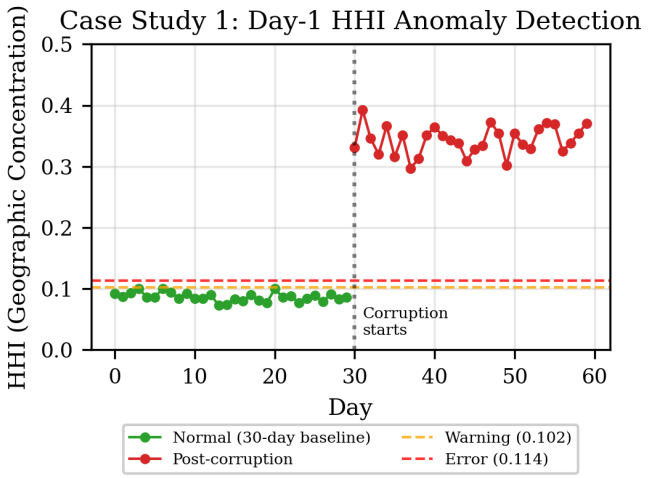


Fig. 4. HHI timeline for Case 1: baseline 0.07–0.10, event day jumps to 0.346 (4.6× IQR error fence). Post-corruption regime is visibly disjoint from baseline.

Case 2: Cross-Source Spend Inflation (57%, Multiple Countries). Records duplicated across multiple countries with different geographic breakdowns. Compound key uniqueness: PASS. The monitoring system's cross-source spend ratio flagged the divergence (1.57x vs baseline 0.95–1.05x) within the first scoring cycle. IQR deviation: 8.2x.

Case 3: Silent Delivery Gap (42%, Unmeasured Spend). Approximately 40% of expected campaigns absent from quarterly delivery. Row count within configured bounds. The monitoring system's campaign registry comparison identified the gap immediately upon data arrival: ∼40% coverage vs expected 100%.

Case 4: Cross-Advertiser Impression Duplication (Hundreds of Millions of Impressions). A platform's reporting system assigned campaigns from Country A to Country B's advertiser account; both accounts reported identical impression totals, so ingesting both files doubled impressions with no record-level error (validation, raw-vs-transformed reconciliation, and standard deduplication all PASS). Against a clean baseline predating the event, the ISR detector fired within the first scoring cycle (2.0x deviation: impressions doubled, spend unchanged), with cross-source ratio and country-level volume signals converging.

Case 5: Pipeline Logic Change Causing Complete Regional Misattribution (Multiple Countries). A join-logic modification in the regionalization step routed 100% of one sub-channel's impressions across multiple countries to a single "Unknown" bucket; total impressions and spend stayed

correct, so all record-level, spend-reconciliation, and volume checks passed ("Unknown" is a valid region string). HHI jumped from 0.08–0.15 to 1.0 across all affected countries simultaneously and HHI monitoring fired on the first scoring cycle. Cases 1 and 5 show detection is agnostic to corruption source (upstream platform change vs. routine pipeline maintenance); untreated, this would have propagated to downstream regional attribution models and misallocated budget.

E. Cross-Organizational Incident Mapping

We assess architectural applicability by mapping our detection portfolio against 15 publicly documented incidents. Architectural applicability: 12/15 published incidents map to at least one detector in the monitoring portfolio (qualitative architectural mapping; not empirical detection on external data), vs $\sim$4/15 for record-level tools. This mapping provides evidence that the monitoring paradigm is applicable to incidents outside the study environment; it does not empirically demonstrate transfer of the quantitative coverage gains.

Quantitative reconstruction (3 incidents). Knight Capital: 100x multiplier—cross-source ratio fires at 1.8x within first scoring window. 2016 video metrics: 1.6x impression inflation—ISR fires at 1.5x threshold. 2023 platform outage: 100% data removal—volume baseline fires immediately.

F. Longitudinal Experimental Evaluation

TABLE IX
LONGITUDINAL EVALUATION METRICS

Metric	Value
Observation window	497 days (Jan 2025–May 2026)
Data sources monitored	15
Countries	12
Anomaly scores/day	Tens of thousands
Cumulative scores	Millions
Campaigns tracked	Hundreds active/day
Overhead	$<$3 sec/source/cycle
Alerts triggered	47
True positives	34 (72.3%, CI: 58.2–83.1%)
False positives	8 (17.0%)
Inconclusive	5 (10.6%)

Computational cost. Tens of thousands of daily scores computed in $<$3 seconds per source on commodity hardware (4 vCPU, 16 GB). Storage: millions of cumulative scores occupy single-digit GB compressed. Total active baseline state fits under 100 MB.

Detection characteristics. Step-change corruptions (Types 1, 3, 4, 9) are identified within the first scoring cycle. Registry-based detections (Types 2, 8) fire immediately upon data arrival. Gradual proportional shifts (Types 5, 6) require 3–7 days when cumulative magnitude exceeds $\sim$15%.

Type distribution of confirmed findings. Of 34 confirmed true positives: missing campaigns (Type 2): 12 (35%); spend deficit (Type 3): 8 (24%); geographic concentration (Type 1): 5 (15%); multiplier/duplication (Types 4, 9): 5 (15%); methodology and coverage (Types 7, 11): 4 (11%).

Statistical bounds on DCM. Bootstrap 95% CI: ours [0.79, 1.0], Deequ [0.43, 0.82]. The gap (0.29) exceeds the CI overlap, supporting the comparison. As throughout, DCM = 0.93 denotes coverage of the classified taxonomy under experimental conditions, not recall over the population of all SVDC events.

False positives and recall. The false positive rate is measured directly: 8% at country grain under dual confirmation, blended precision 72.3% (Wilson 95% CI: 58.2–83.1%) across all metric-grain combinations. Recall is not directly measurable operationally because the ground-truth set of *undetected* SVDC events is by definition unobservable. We approximate it via two independent mechanisms: (a) the synthetic benchmark, where ground-truth labels enable per-type detection-rate measurement (Section VI-B); and (b) the cross-organizational architectural mapping (12/15 = 0.80, Section VI-E). This two-track recall approximation is a common pattern in operational anomaly-detection literature, where operational ground truth for missed events is unavailable.

VII. LESSONS LEARNED

A. Alert Noise Reduction

The raw anomaly rate across all 12 detectors is 37.5%—far too noisy for practical alerting. Reduction mechanisms: (1) Alerting separation: only 3 metric-detector combinations trigger alerts, reducing actionable alert volume by 94%. (2) Inactive entity filtering: entities with zero activity produce meaningless scores; filtering reduced raw anomaly volume by 59%. (3) Dual confirmation: requiring both IQR and MAD agreement reduces false positives from 22% to 8% at country grain.

Precision across the experimental window: 72.3% (34/47 confirmed; Wilson 95% CI: 58.2–83.1%). The 8% dual-confirmed HHI false positive rate at country grain rises to a 27.7% blended rate because the 47 alerts span multiple metric types and grains, with finer grains contributing higher FP rates.

Cost model. A 72.3% precision rate implies 27.7% of alerts require investigation yielding no actionable issue. This tradeoff is favorable when a single confirmed true positive (e.g., the geographic anomaly caught in the first scoring cycle versus $\sim$8 months for manual review) outweighs the cost of investigating false positives; practitioners should weigh this against their own true-positive value and investigation cost.

B. Cold-Start and Baseline Management

Cold-start protocol: New data sources accumulate 30 days of baseline without alerting (minimum 7 data points). This creates a known vulnerability window.

Baseline contamination: If corruption persists undetected for $>$30 days, it is absorbed into the baseline. The system would then alert when corruption *stops*. Mitigation: complementary cross-source and registry checks that do not rely on rolling baselines.

IQR=0 edge case: When all baseline values are identical, IQR=0 and detection returns None. This occurs for very small

markets with one dominant region. MAD provides partial fallback but also fails at MAD=0.

Additional failure modes. A source delivering no data is skipped for that day (no false alert) with a data-availability warning routed separately. Schema changes are handled by a config flag that resets the baseline into a new cold-start window. Missed scoring windows are backfilled by reprocessing against the stored baseline—no observation data is lost.

C. Threshold Sensitivity

TABLE X
THRESHOLD SENSITIVITY

Parameter	Range	DCM	Precision
IQR multiplier	2.0–4.0	0.93	62–84%
MAD threshold	4.0–6.0	0.86–0.93	68–79%
Baseline window	14–60 days	0.86–0.93	65–74%

DCM is stable (0.86–0.93) across all tested configurations. The approach is not brittle to threshold selection.

VIII. Related Work

We organize related work by *why* each class does not natively surface SVDC.

Record-level validators. Great Expectations [1] and Apache Deequ [2] define current practice. Both validate per-record predicates: schema, nulls, ranges, enums, uniqueness, statistical bounds. Neither natively expresses distributional metrics on categorical columns tracked over rolling baselines with per-entity self-calibration.

Data observability platforms. Soda Core, Monte Carlo, and Anomalo [5] add temporal anomaly detection on aggregate metrics and distribution comparison via KS tests against reference profiles. Under exhaustive configuration, Soda Core achieves DCM = 0.64—the same plateau as Deequ. Static reference distributions do not adapt to per-entity baselines, and single-dataset validation does not perform cross-source ratio monitoring.

dbt + elementary. The elementary package [6] extends dbt with anomaly detection on model metrics over time. Architecturally closer to our approach. Still limited by single-dataset scope: no cross-source reconciliation, no HHI-style grouped-distribution metrics, no registry comparison.

Diff-based dataset comparison. Datafold [7] and dbt-audit-helper compare dataset versions. These partially address Type 10 (retroactive mutation)—our only 0% type. Integrating snapshot comparison is a natural extension.

Uber D3. Uber's D3 [8] is the most architecturally similar published work, detecting data drifts on aggregate metrics over time. D3 focuses on row-count and freshness anomalies on raw tables; our portfolio adds HHI on grouped distributions, cross-source reconciliation, registry comparison, and per-entity self-calibrating baselines. The approaches are complementary: D3's Prophet-based forecasting excels at smooth trend deviations but would miss sudden categorical concentration (Type 1), which requires grouped-distribution metrics, while D3's automated onboarding and broader metric coverage complement our detection-focused design.

Time-series anomaly detection. USAD [9], OmniAnomaly [10], and TranAD [11] detect temporal deviations in numeric series but operate on raw values. SVDC exists in *derived* distributional metrics (HHI, proportions, ratios) computed across categorical groupings—a different application domain. Our approach applies standard anomaly detection to derived distributional metrics, not to raw values directly.

Learned drift detection. DQuag [12] detects hidden errors via a GNN within a single dataset; DriftLens [13] measures drift in deep-learning representation spaces; ADWIN-U [14] extends adaptive windowing to unsupervised streams. None addresses cross-source reconciliation, registry-based absence detection, or per-entity grouped-distribution monitoring; the approaches are composable.

Profiling, SPC, fusion, and taxonomies. Automatic constraint discovery [15], [16] finds record-level functional dependencies and profiles, which SVDC satisfies while corrupting properties outside the constraint language. Our IQR/MAD detectors apply SPC [17] to derived metrics. Truth-discovery systems CRH [18] and TruthFinder [19] *resolve* cross-source conflicts, whereas we *detect* inconsistencies as quality signals. Failure taxonomies [20], [21] classify infrastructure failures; ours targets data corruption that evades validation. HHI is standard in antitrust economics [22]; its use for detecting geographic redistribution in analytical pipelines is, to our knowledge, novel.

Industrial platforms. Databricks Lakehouse Monitoring [23] and Anomalo [5] automate per-table anomaly detection without explicit cross-source reconciliation or entity-absence detection.

The pattern is consistent: each class addresses a real problem within its paradigm but does not natively surface SVDC. Our contribution is identifying which distributional metrics, tracked over which baselines, expose the corruption types these paradigms miss.

IX. Threats to Validity, Limitations, and Conclusion

A. Threats to Validity

Six threats and their mitigations. *Single-environment evaluation*—mitigated by cross-organizational mapping (12/15 incidents); residual risk: not validated in a second environment. *Taxonomy circularity* (the taxonomy is co-developed with the portfolio)—mitigated by saturation analysis and independent coding (κ=0.83); residual risk: types outside the detection paradigm are under-represented. *Missing recall measurement*—mitigated by the DCM coverage framework plus precision reporting; residual risk: true operational recall unknown. *Benchmark bias* (self-designed)—mitigated by real-data validation and external incident mapping; residual risk: a different taxonomy yields a different DCM. *Inter-rater dependence*—mitigated by a structured coding protocol; residual risk: external replication needed. *Limited organic incidents*—mitigated by 5 fully-validated case studies and 27 tool-scenario

tests; residual risk: most benchmark scenarios use injected corruption.

B. Limitations and Future Work

Coverage gaps: (1) Type 10 (retroactive mutation): 0% detection—snapshot comparison required. (2) Sub-15% proportional shifts produce insufficient signal. *Validity bounds:* (3) Single-environment evaluation. (4) Co-development circularity acknowledged. (5) Both raters participated in development. *Methodological bounds:* (6) Recall unquantified (precision = 72.3%, Wilson CI: 58.2–83.1%). (7) Type 1 uses organic data; Types 2–11 use injected corruption matching documented patterns. (8) Bootstrap 95% CI: ours [0.79, 1.0], Deequ [0.43, 0.82]—gap exceeds CI overlap. (9) Algorithmic novelty: mechanisms are individually well-established; contribution is systematic application to a previously unformalized problem domain.

Future work. These limitations define concrete next steps. (1) **External replication** on organizations outside the study environment would directly test the single-environment and circularity threats, converting the qualitative cross-organizational mapping into empirical detection. (2) **Recall estimation via injected known-corruption trials** at controlled rates yields a ground-truth positive set, enabling a direct recall estimate the operational setting cannot provide. (3) **Taxonomy extension to non-analytical pipelines** (transactional, streaming) would test whether the four structural categories generalize beyond batch aggregation. (4) **Adaptive baseline selection**—replacing the fixed 30-day window with a change-point-aware adaptive window—would reduce cold-start and contamination sensitivity (Section VII-B). Integrating snapshot comparison to close the Type 10 gap is a further extension (Section VIII).

Cross-Domain Applicability. The SVDC types generalize structurally: geographic concentration maps to any categorical distribution monitoring (hospital patient load, financial trade volume). Missing campaigns generalizes to entity absence detection (IoT sensors, retail stores). Spend deficit maps to volume shortfall in any multi-source aggregation. The detection mechanisms are domain-agnostic; only the specific metrics require domain adaptation.

Conclusion. Record-level validation has a demonstrable coverage plateau on SVDC. Distribution-level temporal monitoring—HHI on geographic distributions, cross-source ratios, campaign registry comparison—closes the gap from DCM 0.64 to 0.93 (practical coverage 0.79 at median incident severity, with subtle proportional shifts below 15% magnitude remaining a known limitation).

The primary contribution is not the monitoring system itself but the characterization of SVDC as a failure class: a 14-type taxonomy organizing corruption patterns that are invisible to per-record validation by construction. The Detection Coverage Metric provides a framework for systematic coverage assessment of any validation portfolio. The monitoring implementation demonstrates that the identified failure class is practically detectable using standard statistical methods applied to the right derived metrics over appropriate baselines.

Adoption guidance. Adopting distribution-level monitoring requires identifying 2–3 distributional metrics per source (geographic concentration, volume ratios, and entity counts cover the majority of SVDC types), collecting 30 days of baseline (minimum 7 points), and applying IQR anomaly detection on daily values—fewer than 100 lines of monitoring code for most pipelines, configuration-driven so that adding a source requires only a config entry and a rules definition. The taxonomy enables systematic gap analysis of any organization's validation portfolio.

References

[1] "Great Expectations," https://greatexpectations.io/, accessed: 2024.

[2] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," in *Proceedings of the VLDB Endowment*, vol. 11, no. 12, 2018, pp. 1781–1794.

[3] B. G. Glaser and A. L. Strauss, *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine, 1967.

[4] G. Guest, E. Namey, and M. Chen, "A simple method to assess and report thematic saturation in qualitative research," *PLoS ONE*, vol. 15, no. 5, p. e0232076, 2020.

[5] Anomalo, "Unsupervised data quality monitoring," Anomalo Documentation, 2025.

[6] dbt Labs and Elementary, "dbt + elementary anomaly detection," Elementary Documentation, 2025.

[7] Datafold, "Diff-based dataset comparison documentation," 2025.

[8] Uber Engineering Blog, "D3: Automated system to detect data drifts," https://www.uber.com/blog/, Feb. 2023.

[9] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2020.

[10] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

[11] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep transformer networks for anomaly detection in multivariate time series data," *Proceedings of the VLDB Endowment*, vol. 15, no. 6, 2022.

[12] Anonymous, "DQuag: Automated data quality validation in an end-to-end GNN framework," arXiv preprint arXiv:2502.10667, 2025.

[13] S. Greco, B. Vacchetti, D. Apiletti, and T. Cerquitelli, "DriftLens: Unsupervised concept drift detection from deep learning representations in real-time," arXiv preprint arXiv:2406.17813, 2024.

[14] D. N. Assis and V. M. A. Souza, "ADWIN-U: adaptive windowing for unsupervised drift detection on data streams," *Knowledge and Information Systems*, 2025.

[15] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: A survey," *The VLDB Journal*, vol. 24, no. 4, pp. 557–581, 2015.

[16] E. Livshits and B. Kimelfeld, "Computing optimal repairs for functional dependencies," in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, 2017.

[17] D. C. Montgomery, *Introduction to Statistical Quality Control*, 8th ed. Wiley, 2020.

[18] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" *Proceedings of the VLDB Endowment*, vol. 6, no. 2, pp. 97–108, 2012.

[19] X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 6, pp. 796–808, 2008.

[20] H. S. Gunawi, M. Hao, T. Leesatapornwongsa, T. Patana-anake, T. Do, J. Adityatama, K. J. Eliazar, A. Laksono, J. F. Lukman, V. Martin, and A. D. Satria, "What bugs live in the cloud? A study of 3000+ issues in cloud systems," in *Proceedings of the 5th ACM Symposium on Cloud Computing (SOCC)*, 2014.

[21] P. Huang, C. Guo, L. Zhou, J. R. Lorch, Y. Dang, M. Chintalapati, and R. Yao, "Gray failure: The achilles' heel of cloud-scale systems," in *Proceedings of the 16th Workshop on Hot Topics in Operating Systems (HotOS)*, 2017.

[22] A. O. Hirschman, *National Power and the Structure of Foreign Trade*. University of California Press, 1945.

[23] Databricks, "Lakehouse monitoring," Databricks Documentation, 2025.