Reliable Context

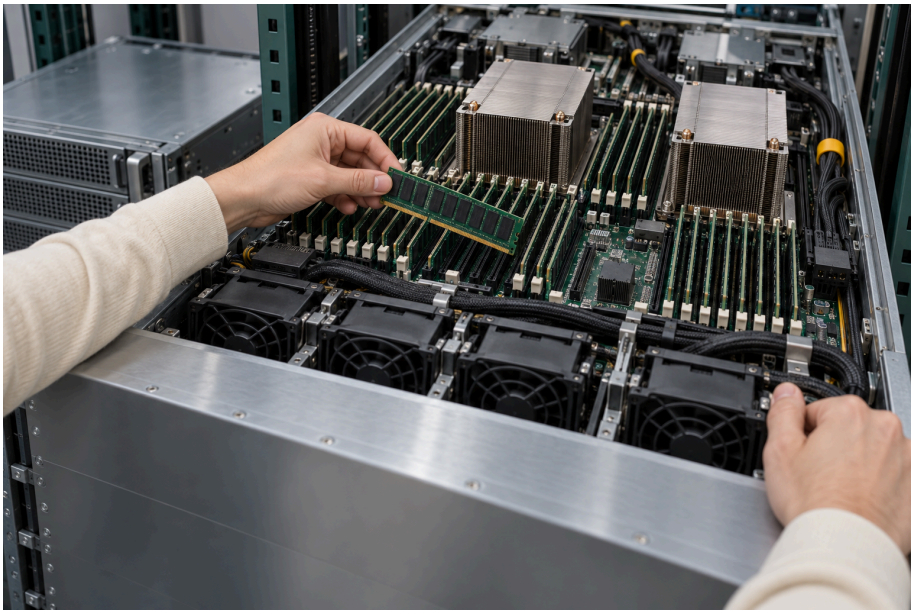
RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

How to Test Truth-Discovery Algorithms Before Choosing

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-09-22 · truth discovery algorithms · truth discovery algorithm comparison · source reliability estimation · multi-source data fusion · truth discovery benchmarks · empirical bayes · average proximity · knowledge fusion · provenance

Original article: https://reliablecontext.com/articles/test-truth-discovery-algorithms



In this article

1. Executive Summary
2. Introduction and Background
3. Formal Property Tests
4. Voting and Unweighted Aggregation
5. Iterative Source-Weighting Operators
6. Proximity Weighting and Empirical Bayes
7. Feature Comparison
8. Performance and Benchmarks
9. Data Analysis and Evidence
10. Implementation Acceptance Protocol
11. Implications and Future Directions
12. Frequently Asked Questions (FAQs)
13. Conclusion

Executive Summary

Choosing a truth-discovery algorithm is not a leaderboard exercise. A defensible acceptance test asks three different questions: does the operator behave coherently on small counterexamples, does it beat a transparent baseline on representative data, and are its statistical assumptions credible for the deployment? Singleton and Booth formalize the first question with a truth-discovery network whose primary output is an ordering of sources and facts (link.springer.com). Their results also prevent careless wish lists: no operator can jointly satisfy Coherence, Symmetry, Monotonicity and Per-object Independence in their model (link.springer.com). A passed axiom is therefore a behavioral guarantee, not proof of factual accuracy, and a failed axiom may be an explicit tradeoff rather than an automatic rejection.

The minimum baseline ladder should begin with **majority voting or an unweighted mean**, then add a competence-weighted mean, then average-proximity weighting, and only then consider empirical-Bayes post-processing. Meir et al. define the real-valued unweighted baseline as the arithmetic mean (arxiv.org) and show that, under a particular Gaussian model, their average-proximity estimate is the unique maximum-likelihood solution with constant regularization (ojs.aaai.org). That is a scoped theorem, not a universal ranking. Ben Shabat et al. require an unbiased base estimator, and their known-competence dominance result applies when there are **more than 3 questions** (arxiv.org). Their real-data results were mixed when truth variation was high relative to noise (arxiv.org).

Empirical evidence reinforces the case for a ladder rather than a winner. A **12-algorithm** comparison found no method consistently best on precision (arxiv.org). One benchmark contained **877 sources and 33,235 claims** (arxiv.org), while a **100 million-claim** configuration exceeded the study's memory capacity (arxiv.org). Those results justify reporting accuracy, uncertainty, convergence, wall time, memory and sensitivity together.

The recommendation is a preregistered stop rule: reject a complex method unless it passes the declared counterexamples, improves the relevant loss with uncertainty over splits or resamples, remains stable under added irrelevant components and sparse participation, and justifies every assumption. Preserve the input,

parameters, output and evidence for each run. W3C PROV provides an interchange model for heterogeneous provenance (www.w3.org), while Reliable Context's relevant observable guarantees are cited values, explicit unknowns and reproducible past states. Neither provenance nor axioms substitute for ground-truth evaluation, but both make the decision inspectable.

Introduction and Background

Truth discovery is the joint problem of estimating **claim trustworthiness** and **source reliability** when sources disagree. The official KDD topic description captures this coupled objective directly (www.kdd.org). This truth discovery algorithm comparison treats multi-source data fusion under conflicting data as an acceptance problem: how to evaluate behavior, evidence and assumptions before choosing. A minimal input contains objects, candidate facts about each object, sources, and directed source-to-fact claims. In the Singleton and Booth model, every fact belongs to exactly one object (link.springer.com). An operator then returns fact and source rankings. Other models instead return probabilities, labels or real-valued estimates.

These output types are not interchangeable. **Categorical single-truth** tasks can use accuracy or error rate, while **multi-truth** tasks require precision and recall because several candidates may be valid (vldb.org). **Continuous** tasks need a numerical distance such as mean absolute error (MAE) (hanj.cs.illinois.edu). A valid comparison must therefore fix the estimand, output type and loss before running code.

Accuracy alone underdetermines behavior. Two operators can tie on a finite test set while reacting differently to a unanimous claim, an unclaimed fact, a duplicate source, an unrelated graph component or a low-participation source. Public datasets with complete ground truth are also difficult to obtain (arxiv.org). The acceptance suite below makes those hidden choices observable before test-set performance decides anything.

Formal Property Tests

Capabilities

A property test is a small graph transformation with a declared expected relation between outputs. It should be run against exact scores where available and against the induced ranking in every case. The core tests are:

- **Coherence:** increasing the credibility of supporting facts should not perversely lower a source, and support from more credible sources should strengthen a fact. This encodes the reciprocal intuition that sources reporting believable facts should rank as trustworthy (link.springer.com).
- **Symmetry:** renaming sources, facts or objects without changing graph structure should merely rename the output. Rankings should not depend on node names (link.springer.com).
- **Unanimity:** a fact claimed by every source should be at least as highly ranked as any other candidate (link.springer.com).
- **Groundedness:** an unclaimed fact should rank no higher than a claimed fact (link.springer.com).
- **Monotonicity:** adding support for a fact should strictly improve it relative to facts it already weakly outranked (link.springer.com).

- **Per-component Independence:** adding or changing a disconnected component should not alter rankings inside the original component. Relevance is defined through membership in the same connected component (link.springer.com).

Counterexamples and scope

Four minimal fixtures expose implementation behavior without a large benchmark:

- **Unanimity fixture:** `s1 -> a, s2 -> a, s3 -> a`, with `b` supported only by `s1`. Require `a >= b`.
- **Groundedness fixture:** `s1 -> a`, while candidate `b` has no incoming claim. Require `a >= b`.
- **Irrelevant-component fixture:** rank `a` and `b`, then append disconnected `s9 -> z`. Require the original ordering to remain unchanged.
- **Added-support fixture:** begin with `a >= b`, then add a fresh source claiming `a`. Require a strict improvement of `a` relative to `b` where Monotonicity is declared.

These tests are scoped. The axiomatic model assumes good-faith independent sources that do not copy or collude (link.springer.com), and its finite fact domains effectively restrict it to categorical values (link.springer.com). The test harness must add separate fixtures for copying, correlation, continuous claims and multiple truths rather than pretending the axioms cover them.

Voting and Unweighted Aggregation

Capabilities

For categorical claims, majority voting counts support. For real values, unweighted aggregation returns the arithmetic mean, as defined earlier. These are not naive throwaways. They are **identifiable baselines** with no learned source weights, low compute cost and easily audited failure cases. Every proposed operator should report incremental improvement over the relevant one.

The formal attraction is also a warning. Per-object Independence, Symmetry and Monotonicity together force within-object rankings toward simple vote counts (aamas.csc.liv.ac.uk). Demanding all three while also expecting cross-object source learning is incoherent. The stronger impossibility result shows that no operator jointly satisfies Coherence, Symmetry, Monotonicity and Per-object Independence in the stated framework (link.springer.com).

Adoption

Majority Voting is implemented alongside Sums, Average.Log, Investment, PooledInvestment and TruthFinder in the open TruthDiscovery documentation (truthdiscovery.readthedocs.io). A separate QCRI repository lists **12 methods** used in a comparative study, including TruthFinder, Cosine, 2-Estimates, 3-Estimates, Depen and maximum likelihood estimation (github.com). These implementation lists demonstrate availability, not benchmark superiority.

Strengths and limitations

- **Strength:** outputs are explainable as counts or means.
- **Strength:** computational cost is low enough to function as a smoke test. A streaming Bayesian study reported that its fast variant matched majority voting's computational speed (proceedings.neurips.cc).
- **Limitation:** copying breaks the independence behind vote counts. Straightforward voting falls short when sources copy (www.vldb.org).
- **Limitation:** honest-majority assumptions must be explicit. Some worker-filtering methods assume honest workers are the majority (www.jmlr.org).

Voting remains the first rung because a complex method that cannot beat it under the chosen loss has not justified its complexity.

Iterative Source-Weighting Operators

Capabilities

Iterative operators alternate between source scores and fact scores. **Sums** resembles graph centrality: sources gain weight from supported facts, and facts gain weight from supporting sources. **TruthFinder** estimates the chance that at least one supporting claim is correct from source precision (hanj.cs.illinois.edu). **Investment** has each source distribute credit uniformly among its claims (hanj.cs.illinois.edu). Their shared attraction is cross-object learning: evidence on one object can affect how much influence a source receives elsewhere.

Adoption

The practical ecosystem is broad but heterogeneous. The TruthDiscovery package documents six implemented methods (truthdiscovery.readthedocs.io), while QCRI's study repository covers dependence-aware and likelihood-based variants (github.com). A comparison must pin repository revision, initialization, stopping rule and numerical tolerance because the algorithm label alone does not specify a run.

Strengths and limitations

Singleton and Booth prove that Sums satisfies Coherence, Symmetry, Unanimity and Groundedness in their ordinal framework (link.springer.com). They also show failures of Per-object Independence, Strong Independence, Per-component Independence and Monotonicity (link.springer.com). The Per-component failure is concrete: global normalization can drive scores in an unrelated component to zero (link.springer.com).

UnboundedSums removes that normalization and satisfies Coherence, Symmetry, Unanimity, Groundedness and Per-component Independence (link.springer.com). Its numerical scores grow without bound, creating a computational concern even if rankings stabilize (link.springer.com). Its Monotonicity is a conjecture, not an established property. This correction matters: PCI here means **Per-component Independence**, not “Property of Consistent Improvement.”

Proximity Weighting and Empirical Bayes

Capabilities

Meir et al.'s **average-proximity** method estimates worker competence from average similarity to other workers. Its informal Anna Karenina principle is that expected average similarity rises roughly linearly with competence (arxiv.org). The procedure computes pairwise proximity, converts it to estimated fault or competence, then gives more weight to lower-fault workers (arxiv.org). Under independent Gaussian errors with known variances, the mean-squared-error-optimal unbiased baseline is an inverse-variance weighted mean (www.itl.nist.gov).

Ben Shabat et al.'s **empirical-Bayes estimator** is different. It is a second stage applied across the vector of object estimates, shrinking a base truth-discovery output under a real-valued statistical model. The official paper page states that, under its conditions, this second step can reduce expected error (proceedings.mlr.press).

Adoption

The official average-proximity paper reports improvement over unweighted aggregation and other truth-discovery methods in its tested settings (ojs.aaai.org). The empirical-Bayes paper reports an expected-error reduction under its stated conditions (proceedings.mlr.press). Both are author-reported results, not independent reproductions.

Strengths and limitations

- **Average-proximity strength:** it is a simple competence estimator with an explicit Gaussian likelihood interpretation (ojs.aaai.org).
- **Average-proximity limitation:** the detailed result treats pairwise distances as conditionally independent given worker types (arxiv.org). Coordinated similar reports can inflate proximity.
- **Empirical-Bayes strength:** the theory supplies an explicit expected-risk comparison with the base estimator, rather than relying only on a benchmark table (proceedings.mlr.press).
- **Empirical-Bayes limitation:** the general result assumes the base truth estimate is unbiased (arxiv.org). Systematic source bias can invalidate that premise.
- **Scope limitation:** the paper analyzes real-valued estimation; categorical acceptance requires a different model and loss.

Feature Comparison

Table 1 compares what can actually be asserted about the principal options. “Unknown” means the cited literature does not establish the property in the stated scope, not that the property fails.

Option	Output and added complexity	Property evidence	Required assumptions	Appropriate acceptance role
Majority or unweighted mean	Count or arithmetic mean, minimal	Symmetry is expected from a label-free implementation; other properties depend on the exact output definition	Equal source influence; copying not corrected	Mandatory first baseline
Sums	Iterative source and fact rankings	Pass: Coherence, Symmetry, Unanimity, Groundedness. Fail: Monotonicity and three independence variants (link.springer.com)	Categorical, single-object membership, independent good-faith sources	Property-test reference implementation
UnboundedSums	Iterative unnormalized scores	Pass: Coherence, Symmetry, Unanimity, Groundedness, Per-component Independence. Monotonicity unknown (link.springer.com)	Same formal scope; numerical growth must be managed	Sensitivity check for normalization effects
Average proximity	Pairwise distances plus competence conversion	Statistical consistency and likelihood results under paper-specific models	Suitable distance, independent bounded or Gaussian noise, limited coordination	Transparent learned-weight baseline
Empirical-Bayes second stage	Shrinkage after a base real-valued estimator	Expected-risk improvement under theorem conditions	More than 3 questions in known-competence case; unbiased base in general case	Optional post-processing rung
Reliable Context	System-of-record assurance layer, not a published truth-discovery operator	Observable guarantees: canonical entities, cited values, explicit unknowns, replayable state	Customer-defined permitted sources, routing, clearance and retention	Preserve defensible selected values and their evidence, not replace algorithm evaluation

The matrix separates three kinds of evidence that are often conflated. A formal pass describes behavior inside a model. A statistical theorem describes expected loss under assumptions. A system-of-record guarantee describes what a reader can inspect after a value is selected. None implies the other two.

Table 2 is the deployment acceptance matrix. It determines which tests must be present before performance is compared.

Data regime	Required baseline	Required property or stress test	Primary metric	Stop condition
Categorical, single truth	Majority vote	Unanimity, Groundedness, Symmetry, added-support	Accuracy or log loss	No gain over voting across uncertainty interval
Categorical, multiple truths	Per-candidate vote threshold	Candidate correlation and threshold sensitivity	Precision, recall, F1	Gains vanish across thresholds
Real-valued	Mean, median, inverse-variance mean	Scale, outlier and variance-estimation sensitivity	MAE and mean squared error	No improvement over best transparent estimator

Data regime	Required baseline	Required property or stress test	Primary metric	Stop condition
Copied or coordinated sources	Deduplicated or cluster-weighted vote	Copy-rate and correlation sweep	Loss by dependence level	Ranking reverses under plausible copying
Sparse participation	Global vote or mean	Minimum claims per source and missingness sweep	Loss by coverage bin	Low-coverage sources cause unstable weights
Dynamic stream	Rolling unweighted estimate	Arrival-order, drift and stale-source tests	Instantaneous and final accuracy	Material path dependence without declared model

This matrix blocks invalid cross-regime conclusions. A Markov random field can represent candidate correlations in a multi-truth model (eudl.eu), while continuous estimates use numerical distances. The same score cannot fairly rank both without redefining the task.

Performance and Benchmarks

Benchmark design

A benchmark should vary one causal factor at a time and also include mixed regimes. Waguih and Berti-Equille controlled the number of claims, the number and distribution of distinct values, and the distribution of true-positive claims (arxiv.org). They applied the same convergence test to iterative algorithms, using cosine similarity between successive source-trustworthiness vectors (arxiv.org). Equal convergence rules are necessary for a fair runtime comparison.

The benchmark grid should sweep:

- **Source competence:** constant, bimodal and continuous spectra.
- **Coverage:** dense to highly sparse participation. Crowdsourced datasets are often sparse and imbalanced (www.ijcai.org).
- **Dependence:** independent, clustered copying and changing dependence. Dynamic copying can stop or switch over time (www.vldb.org).
- **Truth multiplicity:** one valid fact versus correlated multiple truths.
- **Noise family:** bounded, Gaussian, heavy-tailed, biased and adversarially coordinated.
- **Scale:** sources, objects, claims and candidate values, each varied separately.

Crowdsourcing benchmarks also differ in annotation density. One recognizing-textual-entailment corpus contains **800 tasks, 164 workers and 8,000 labels** (proceedings.neurips.cc); its source study gathered **10 annotations for each of 800 sentence pairs** (aclanthology.org). Density must be reported, not hidden behind the dataset name.

Reporting protocol

Report point loss, uncertainty, failure rate, iterations, wall time and peak memory. NIST requires measurement results to carry quantitative uncertainty statements (www.nist.gov). A bootstrap is a transparent option: NIST describes repeated resampling with replacement, typically **500 to 1,000 subsamples** (itl.nist.gov). Use object-level resampling when sources answer overlapping objects, and cluster resampling when dependence is the risk of interest.

The interval should be interpreted as uncertainty about an estimate, not decorative error bars. NIST describes measurement uncertainty as incomplete knowledge about the measurand (www.nist.gov), and notes that confidence intervals for a mean narrow as sample size increases (itl.nist.gov). Report the unit and resampling scheme so that narrowing is not mistaken for independence.

Runtime must include failures. MLCommons treats failure to reach a target inside the budget as infinite completion time (mlcommons.org) and standardizes benchmark hardware to make runtime comparable (mlcommons.org). A truth-discovery report need not copy that hardware, but it should copy the discipline: publish hardware, software, stopping rule and timeout treatment.

Data Analysis and Evidence

The quantitative literature does not support a universal winner. In the 2014 comparison, none of the tested methods consistently outperformed the others on precision (arxiv.org). Under an exponential spectrum of source trustworthiness, none of the tested methods was reliable (arxiv.org). This is a negative result worth retaining in an acceptance plan.

Table 3 converts published measurements into a baseline ladder. Reported values remain tied to their original designs.

Rung	Published evidence	Added assumptions	Decision use
Majority or mean	**438 ms** and **528 ms** in two 2014 synthetic scale configurations (arxiv.org)	Equal source influence	Establish error and compute floor
Inverse-variance mean	Mean-squared-error-optimal unbiased estimator under independent Gaussian errors with known variances (www.itl.nist.gov)	Known, reliable variance estimates	Test whether competence weighting alone explains gains
Average proximity	Author-reported improvement over unweighted aggregation and other methods (ojs.aaai.org)	Informative pairwise distance and independence model	Challenge more complicated learned weights
Empirical Bayes	Author-reported synthetic risk reductions ranged from **1% to 90%** across tested algorithms (arxiv.org)	Unbiased base estimate and shrinkage model	Add only after base estimator qualifies
Dependence-aware or multi-truth	Evaluate precision and recall, plus loss by copying or correlation level	Explicit dependence or candidate-correlation model	Use only when data diagnostics show the regime

The table's essential column is “added assumptions.” An estimated weighted-mean variance can understate the true variance, especially in small samples (www.itl.nist.gov). Likewise, significant input correlations cannot be ignored in an uncertainty evaluation (www.bipm.org). A smaller point error with a fragile variance estimate is not decisive.

Worked Gaussian example (Hypothetical Example)

Suppose three independent sources report **10**, **12** and **20**, with known error variances **1**, **4** and **25**. The unweighted mean is **14.00**. Inverse-variance weights are `1`, `0.25` and `0.04`, so the weighted estimate is `(10 + 3 + 0.8) / 1.29 = 10.70`. This follows the Gaussian common-mean result for the optimal unbiased estimator (www.itl.nist.gov).

Average proximity cannot be fairly demonstrated from one object because competence is learned across questions. An acceptance test would generate or retain many objects, estimate competences on training objects, and apply the resulting weights to held-out objects. Empirical Bayes comes later still: it shrinks a vector of base estimates across more than **3 questions** in the known-competence setting (arxiv.org). Reporting an empirical-Bayes “answer” for this single object would misstate the method.

Published dataset sizes show why regime labels matter. The Book Authors data include **1,263 books, 24,364 listings and 877 bookstores** (www.microsoft.com). A biographical continuous benchmark contains **1,372,066 claims** about **9,924 date entities**, with **1,912 verified dates** (hanj.cs.illinois.edu). The empirical-Bayes evaluation used data shapes including **208 workers and 25 questions**, **50 and 300**, and **10 and 200** (arxiv.org). No one shape establishes robustness across the others.

Annotation count can change the baseline itself. Snow et al. collected **7,000 affect labels** in one experiment (aclanthology.org) and reported that averaging **4 non-expert annotations** per example matched one expert's average agreement (aclanthology.org). This does not generalize automatically, but it is a useful reminder to compare aggregation against the cost of collecting another label.

Implementation Acceptance Protocol

Preregister the test

Before observing comparative results, write a one-page protocol:

- **Estimand:** one label, a set of valid labels, a probability, or a real value.
- **Required properties:** name each axiom, its scope and whether failure is fatal.
- **Expected failures:** state where copying, bias, sparsity or multiple truths should break the model.
- **Baseline ladder:** majority or mean, robust unweighted estimator, competence weighting, then optional post-processing.
- **Splits:** hold out objects, time periods or source clusters without leaking source-quality labels.
- **Metrics:** loss, uncertainty, calibration where probabilistic, wall time, peak memory and convergence failures.

- **Stop rule:** reject added complexity if its improvement interval includes zero, if a required counterexample fails, or if results reverse under a plausible perturbation.
- **Artifacts:** fix code revision, environment, seeds, inputs, parameters and complete outputs.

Sparse participation deserves its own stop test. In one city-population experiment, removing low-participation sources discarded **10,491 claims**, or **20.74%** of all claims (www.vldb.org). Filtering may simplify estimation while changing the target population. Report both the retained-source result and the all-source result.

Preserve evidence and reproduction paths

An acceptance result is only defensible if another team can reconstruct it. ACM SIGMOD's reproducibility package asks for configuration, initialization, scripts, workload and measurement protocol (reproducibility.sigmodconf.hosting.acm.org). It also recognizes that different hardware need not produce identical performance numbers if the overall behavior and conclusions reproduce (reproducibility.sigmodconf.hosting.acm.org). NIST similarly describes reproducibility as obtaining similar analytic results with identical methods (www.nist.gov). ACM SIGSIM's artifact criteria require documentation, consistency, completeness, executability, and verification and validation evidence (sigsim.acm.org); its Results Reproduced badge requires a team other than the authors to obtain the main results (sigsim.acm.org).

Provenance should connect every reported metric to a run, parameter set, dataset version and output. W3C PROV defines provenance as a record of the entities, activities, people and institutions involved in producing data (www.w3.org). Its bundle construct can also represent provenance about a provenance record (www.w3.org).

Reliable Context belongs at this evidence layer, not as an undocumented algorithm claim. Its published contract says every value has a cited, verbatim, dated source and that past state is reproducible. Those observable properties can retain the chosen value, competing evidence and decision history. They do not establish that a particular truth-discovery operator is accurate.

Implications and Future Directions

The acceptance suite changes algorithm selection from “which score is highest?” to “which guarantees and assumptions are justified?” Four research directions follow.

- **Dependence-aware competence:** conditional independence can prevent a model from representing correlations between workers' labels (proceedings.mlr.press). Future tests should sweep correlation structures rather than add a single “copied” flag.
- **Dynamic evaluation:** streams require both instantaneous and final accuracy (pmc.ncbi.nlm.nih.gov). Arrival order, delayed truth and source drift should be first-class variables.
- **Calibrated outputs:** a predicted probability near **0.8** should correspond to about **80%** positives over comparable cases (scikit-learn.org). Yet Brier loss alone does not prove better calibration because discrimination and uncertainty also affect it (scikit-learn.org). Reliability plots and proper scores should accompany ranking metrics.

- **Truth versus consensus:** proximity and iterative reinforcement can reward agreement even when agreement is wrong. External labels, temporal verification or retained primary evidence remain necessary for factual validation.

No-method outcomes are valid. If every operator fails a required property, or no complex method beats the baseline with uncertainty accounted for, the correct decision is to retain the simpler estimator and document the unresolved regime. Negative results narrow the conditions under which a system can be trusted.

Frequently Asked Questions (FAQs)

What is the first test for a truth-discovery algorithm?

Run majority voting for categorical data or the unweighted mean for real values, then execute Unanimity, Groundedness, Symmetry and irrelevant-component counterexamples. Only after those results are recorded should a learned-weight method enter the comparison.

Does satisfying an axiom mean the output is true?

No. An axiom states how outputs should change when inputs change. It does not compare the output with external ground truth. Conversely, failing an axiom is not automatically fatal when the application deliberately rejects that normative requirement.

Which truth-discovery evaluation metrics should be used?

Use accuracy or log loss for single categorical truths, precision and recall for multiple truths, and MAE plus mean squared error for real values. Add calibration for probabilistic outputs, uncertainty intervals, runtime, memory and convergence failures.

When is empirical-Bayes post-processing appropriate?

Use it only for real-valued outputs after checking the assumptions detailed above. The official paper presents empirical Bayes as a second stage of a truth-discovery algorithm, not a replacement for task definition or baseline validation (proceedings.mlr.press).

How should sparse sources be handled?

Do not silently remove them. Compare inclusion, filtering and partial-pooling policies, then report loss by source-coverage bin and the number of claims removed. Crowdsourced datasets are often sparse and imbalanced (www.ijcai.org).

Conclusion

A truth-discovery algorithm deserves use only after passing a layered acceptance test. First, define the data regime and loss. Second, declare the normative properties that matter and execute small counterexamples. Third, climb a transparent baseline ladder from voting or an unweighted mean through competence weighting,

adding proximity estimation or empirical-Bayes post-processing only when their assumptions fit. Fourth, quantify uncertainty, compute, convergence and sensitivity rather than reporting accuracy alone.

The literature supplies no universal champion. It instead supplies sharper questions: which independence demand conflicts with cross-object learning, whether normalization makes unrelated components interact, whether source competence is identifiable under the observed participation pattern, and whether shrinkage improves expected error under a credible model. The strongest decision is often conditional: one method for categorical single-truth data, another for continuous measurements, and no complex method when copying or sparse participation invalidates the assumptions.

Finally, preserve the evidence behind the choice. Inputs, source claims, code revision, parameters, outputs and uncertainty calculations should form a reproducible chain. That record lets a reviewer distinguish a theorem from an author-reported experiment, a behavioral guarantee from factual accuracy, and a measured gain from an artifact of the benchmark. The result is not a crowned algorithm. It is a defensible decision rule for when added complexity has earned deployment.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://link.springer.com/article/10.1007/s10458-022-09569-3
2. https://arxiv.org/html/1905.00629v4
3. https://ojs.aaai.org/index.php/AAAI/article/view/25750
4. https://arxiv.org/html/2206.04816v1
5. https://arxiv.org/html/1409.6428
6. https://www.w3.org/TR/prov-dm
7. https://www.kdd.org/kdd2016/topics/view/data-reliability-and-truthfulness
8. https://vldb.org/pvldb/vol5/p550_bozhao_vldb2012.pdf
9. https://hanj.cs.illinois.edu/pdf/qdb12_bzhao.pdf
10. https://aamas.csc.liv.ac.uk/Proceedings/aamas2020/pdfs/p2011.pdf
11. https://truthdiscovery.readthedocs.io/en/latest/algorithms.html
12. https://github.com/qcri/DAFNA-EA
13. https://proceedings.neurips.cc/paper_files/paper/2019/file/54ee290e80589a2a1225c338a71839f5-Paper.pdf
14. https://www.vldb.org/pvldb/vol2/vldb09-335.pdf
15. https://www.jmlr.org/papers/v18/15-650.html
16. https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/consmean.htm
17. https://proceedings.mlr.press/v180/ben-shabat22a.html
18. https://eudl.eu/doi/10.1007/978-3-030-92638-0_2
19. https://www.ijcai.org/proceedings/2023/333
20. https://aclanthology.org/D08-1027.pdf
21. https://www.nist.gov/director/nist-information-quality-standards

22. https://itl.nist.gov/div898/handbook/eda/section3/bootplot.htm
23. https://www.nist.gov/itl/sed/topic-areas/measurement-uncertainty
24. https://itl.nist.gov/div898/handbook/eda/section3/eda352.htm
25. https://mlcommons.org/benchmarks/algorithms
26. https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6?download=true&t=1659082531978&version=1.10
27. https://www.microsoft.com/en-us/research/wp-content/uploads/2011/03/fr743-yin.pdf
28. https://www.vldb.org/pvldb/vol8/p425-li.pdf
29. https://reproducibility.sigmodconf.hosting.acm.org
30. https://sigsim.acm.org/conf/pads/2024/blog/artifact-evaluation
31. https://proceedings.mlr.press/v97/li19i/li19i.pdf
32. https://pmc.ncbi.nlm.nih.gov/articles/PMC12948928
33. https://scikit-learn.org/1.8/modules/calibration.html

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.