Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Source selection in data fusion: evaluating marginal evidence

> 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 · 2026-10-15 · source selection in data fusion · data fusion · truth discovery · source reliability · conflicting data sources · correlated sources · evaluation metrics · data provenance · source selection

Original article: https://reliablecontext.com/articles/when-more-sources-make-data-fusion-worse



In this article

1. Executive Summary
2. Introduction and Background
3. What source selection optimizes
4. A preregistered source-subset experiment
5. Dependence, reliability and negative marginal gains
6. A provenance ledger and reversal analysis
7. Analysis of key segments
8. Data Analysis and Evidence
9. Implications and Future Directions
10. Conclusion

Executive Summary

Source selection in data fusion should measure what another source changes in the inferred result, rather than equate additional observations with additional independent evidence. The literature supports evaluating both inclusion and exclusion. In Dong et al.'s 100-book illustration, voting fell from **93 correct author lists at 583 sources to 78 at all 894 sources**. (www.vldb.org) Conversely, a confidence-aware study found that removing small sources reduced coverage and increased mean absolute error. (cse.buffalo.edu) Neither establishes a universally optimal subset, consistent with the assumptions-dependent landscape described by Li et al.'s survey. (www.kdd.org)

This report proposes a **held-out source-subset experiment**, not a newly executed benchmark. Compare all sources, repeated random subsets, development-set accuracy rankings and the proposed selector using identical objects, labels, fusion rules and resource boundaries. Final test labels must not influence ranking or subset-size selection. (scikit-learn.org) Related objects require group-aware splitting. (scikit-learn.org) Publish the plan before inspecting the final comparisons, separating exploratory development from confirmation. (www.cos.io) Each comparison should retain both the source set and its object-level answer changes.

The quantitative core separates **correct-answer yield on a fixed verified population**, **conditional accuracy among answered verified objects**, and **coverage**. This follows the coverage-accuracy distinction in selective prediction. (www.jmlr.org) A source can expand coverage while reversing previously correct answers. Dependence analysis must distinguish common true answers from evidence of copying, because agreement alone does not establish dependence. (lunadong.com) Objects without an independently verified answer remain unverifiable, while truth-existence research also warns that the correct answer may be absent from every candidate source. (pmc.ncbi.nlm.nih.gov)

Acquisition expenditure, integration labor and compute should remain separately measured unless a decision-maker explicitly supplies a conversion into common utility. State measurement units. (www.w3.org) Pair before-and-after object outcomes when constructing bootstrap intervals, and declare the variability those intervals represent. (docs.scipy.org) Preserve source snapshots, observation windows and superseded claims

so another evaluator can reconstruct the result. Provenance supplies derivation records; it does not establish correctness by itself. (www.w3.org) The resulting acquisition decision is local to the tested sources, objects, time period and costs, with broader generalization requiring further evidence. (nips.cc)

Introduction and Background

Source selection in data fusion asks which sources should contribute to a combined answer, under specified quality and resource constraints. Truth discovery adds the problem of estimating source reliability when trustworthy answers are not already known. The survey by Li et al. (2015 issue) organizes methods by their assumptions rather than identifying a universally superior algorithm. The Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD) archive labels the issue December 2015; this report follows that verified issue date. (www.kdd.org) (kdd.org)

The unit of analysis must be explicit. Here, an **object** is an entity-attribute-time question; a **claim** is a source's proposed answer; a **source set** is the collection admitted to fusion; and the **fusion output** is an answer or an explicit abstention. These are operational definitions for the proposed experiment. For set-valued attributes, evaluation must distinguish exact-set correctness from partially correct answers, a distinction addressed by Lin and Chen's (2018) domain-aware multi-truth research. (researchportal.hkust.edu.hk)

An **independently verified value** is an evaluation label established through a process separated from the selector's training and tuning. Independence is a study requirement, not a property conferred by calling a label “gold.” The protocol therefore records label provenance, adjudication and access restrictions, and prohibits using final test labels to rank sources. Official evaluation guidance similarly excludes test data from model choices. (scikit-learn.org)

The decision is whether acquiring and integrating an additional source improves the defensible result enough to justify its burden. That requires separating answer coverage from answer quality. Selective-classification research formalizes a coverage-accuracy tradeoff; its relevance here is methodological, rather than evidence that any particular fusion selector works. (www.jmlr.org)

This report combines published evidence with a proposed evaluation protocol. It reports no new source-subset experiment, no universal optimal subset and no measured performance for private organizational data. Generalization claims should match the scope of the actual experiments, as research-reporting guidance requires. (nips.cc)

What source selection optimizes

A selector and a fusion rule solve different problems. The selector chooses admissible evidence; the fusion rule resolves that evidence into outputs. Changing both simultaneously makes attribution difficult. The proposed evaluation first fixes the fusion rule and varies the source set, then separately measures interactions with alternative fusion rules. The literature supplies multiple objectives: Lin et al. (2016) formulate expected true-value contribution and cost-constrained selection, while Shao et al. (2017) minimize expected fusion error under solicitation constraints. (arxiv.org) (research.ibm.com)

Dong, Saha and Srivastava's **MARGINALISM** maximizes gain minus cost in common units and uses a greedy randomized adaptive search procedure (GRASP) heuristic. It assumes single truths, mapped schemas, linked entities and independent sources. (www.vldb.org)

For a new experiment, a **marginal evidence contribution** is the change produced by admitting a specified source set relative to a specified baseline. It is conditional on the objects, fusion rule, time window and other sources already present. Domain-aware research challenges treating source reliability as a universal scalar across subject domains, supporting the need to describe the population in which contribution is measured. (researchportal.hkust.edu.hk)

The proposed decision record retains a vector of outcomes rather than assuming that accuracy, staff time and currency share an objective unit. This is consistent with the World Wide Web Consortium's Data Quality Vocabulary, or W3C DQV, which asks that measurement units be specified and provides a framework rather than a complete definition of quality. (www.w3.org)

Predefine the following quantities before optimizing:

- **Eligible population:** the objects the source acquisition is intended to serve.
- **Verified population:** the fixed subset with usable independent labels.
- **Answer policy:** which outputs count as answers, abstentions or unresolved conflicts.
- **Quality endpoints:** correct values, incorrect values and exact-set or component-level scores.
- **Coverage endpoints:** answered objects and newly observed objects, measured separately.
- **Resource endpoints:** acquisition fees, integration labor, computation and recurring maintenance.
- **Decision constraints:** a budget, a minimum quality requirement or an explicitly elicited utility.

These are proposed measurement choices, not a claim that one literature objective covers every organization. Gebru et al.'s 2018 Datasheets for Datasets proposal provides a complementary basis for recording collection context, fees and access restrictions. Such documentation prevents a nominally inexpensive source from being treated as costless when access or maintenance imposes additional work. (www.microsoft.com)

A source should therefore be evaluated for the questions it changes. A broad source may add only redundant claims; a narrow source may answer rare questions that the existing collection misses. Confidence-aware truth discovery explicitly studies the uncertainty associated with sources that supply few claims, making source size an inadequate substitute for demonstrated reliability. (cse.buffalo.edu)

A preregistered source-subset experiment

Preregistration fixes the intended comparisons and analysis before examining their results. The Center for Open Science defines it as specifying a research plan in advance and distinguishes exploratory model development from confirmation on previously unanalyzed data. The proposed protocol applies that separation to source ranking, dependence estimation and subset-size selection. (www.cos.io)

Partition objects into development, validation and final test sets. Estimate source-quality parameters and tune selection budgets using the first two; evaluate the locked method on the third. If related records share an entity, document family or recurring measurement process, split at that group level. Scikit-learn's group cross-

validation guidance explicitly prevents the same group from appearing in both training and test sets. (scikit-learn.org)

Identical held-out objects and labels are essential. Every method receives the same eligible test population and the same verification mask. A method that does not answer an object remains in the comparison through an abstention, rather than dropping the object from its denominator. Official evaluation guidance warns that different evaluation splits invalidate straightforward fold-level comparisons. (scikit-learn.org)

Table 1 defines the comparison ladder and distinguishes selection methods from record-keeping infrastructure. It includes the publisher's stated record guarantees because retaining evidence is relevant to reproducibility; that row supplies no source-selection or accuracy result.

Comparand or record option	Controlled evaluation role	Evidence required before interpreting a gain
All sources	Fixed fusion rule over every permitted candidate source.	The complete candidate inventory, output coverage and verified errors.
Random subsets	Repeated subsets at matched source counts and separately matched budgets.	Sampling design, seeds, repetitions and the distribution of results.
Accuracy-ranked subsets	Ranking estimated exclusively from development objects.	Training-label availability, uncertainty and missing-source handling.
Proposed selection method	Locked selector evaluated at the same budgets and test objects.	Parameter estimates, selection costs, chosen sets and paired output changes.
Reliable Context, record-keeping option	First-party guarantee of a cited, verbatim, dated source behind every value.	These observable record properties do not establish a selection method's marginal accuracy; no comparative benchmark is supplied here.

The method rows are experimental comparands, not vendor recommendations. The record row addresses the distinct requirement of preserving evidence. A researcher can implement that requirement using suitable infrastructure, but cannot count documentation quality as an observed gain in fused-value accuracy. W3C provenance distinguishes entities, their production and their derivations, providing vocabulary for the retained evidence. (www.w3.org)

The preregistered execution plan should include:

- **Candidate freeze:** identify source versions and permitted uses before scoring.
- **Object freeze:** publish inclusion rules and the verification mask.
- **Split freeze:** retain group assignments and time boundaries.
- **Parameter freeze:** estimate reliability and dependence without final labels.
- **Budget grid:** compare matched counts and matched measured costs.
- **Randomization plan:** specify subset sampling and source-order repetitions.
- **Stopping rule:** define the endpoint before inspecting favorable curves.
- **Release plan:** identify which artifacts can be shared and which require controlled access.

These choices operationalize separation between exploration and confirmation; they should be recorded even when some sources are private. Preregistration does not itself make a method correct, but it makes deviations from the planned comparison visible. (www.cos.io)

An accuracy ranking estimated from the final labels may be retained as an explicitly labeled oracle analysis. It cannot represent a deployable baseline. The same restriction applies to choosing the best subset size after looking at the final test curve. Those choices use test data for model selection, the leakage that the official guidance warns against. (scikit-learn.org)

Dependence, reliability and negative marginal gains

Correlated sources need not provide independent confirmations. Conversely, agreement does not prove copying: independent sources can report the same true value. Dong, Berti-Equille and Srivastava's dependence report makes this distinction explicit. The proposed experiment therefore records overlapping claims separately from any inference about the process that produced that overlap. (lunadong.com)

Dependence is also not necessarily a whole-source property. The same report permits a source to reuse some values while adding or revising others independently. Excluding that entire source can remove both redundant evidence and useful contributions. Li et al.'s survey considers correlations more general than copying, so a binary duplicate-source flag is an incomplete dependence model. (lunadong.com) (www.kdd.org)

The ablation design should distinguish the selector's assumptions from the fusion rule's assumptions. A gain that appears only when both are given the same inaccurate quality estimates may reflect a shared estimation artifact. A gain that survives changed estimators, budgets and source orders is more informative, although its scope remains bounded by the tested collection. Research-reporting guidance requires that such generalization limits remain explicit. (nips.cc)

Use the following ablations:

- **Dependence disabled:** compare a model that treats sources independently with the selected dependence-aware alternative.
- **Overlap-only control:** select for distinct object coverage without quality ranking.
- **Quality-estimate perturbation:** vary reliability within stated estimation uncertainty.
- **Domain-specific quality:** compare global estimates with estimates for prespecified domains.
- **Sparse-source retention:** test whether removing small sources loses distinctive answers.
- **Group representative control:** retain representatives from inferred dependent groups.
- **Fusion-rule interaction:** repeat the locked subset comparison under another declared rule.
- **Temporal holdout:** estimate parameters before the period used for evaluation.

These are proposed tests, motivated by distinct literature findings. Confidence-aware work emphasizes uncertainty in sparse-source reliability; domain-aware multi-truth work allows reliability to vary by domain; and Critical Source Selection explicitly considers both dependence and how often a source contributes, termed its speak rate. (cse.buffalo.edu) (researchportal.hkust.edu.hk) (hub.hku.hk)

A negative marginal gain is informative only after accounting for the relevant denominator. Additional observations may reverse previously correct answers, introduce wrong answers on previously unanswered objects, or move confidence above an answer threshold. Those mechanisms require separate output traces. Selective prediction's coverage-accuracy distinction supplies the conceptual reason to preserve threshold behavior rather than interpreting answer count as quality. (www.jmlr.org)

The converse also matters. In a confidence-aware truth-discovery experiment, removing small sources reduced coverage and increased mean absolute error. That result prevents an equally unsupported rule that fewer sources are inherently better. A source-selection claim must survive comparison with both inclusion and exclusion controls, rather than relying on the direction suggested by one motivating example. (cse.buffalo.edu)

A provenance ledger and reversal analysis

Data provenance in data fusion should connect the published score to the exact inputs and decisions that generated it. W3C's provenance data model (PROV-DM) represents revision as derivation from an earlier entity. For this protocol, each source snapshot, normalized claim set, selected subset and evaluation output should remain addressable as a distinct experimental artifact. Retaining a derivation is a reproducibility measure, not independent proof that its answer is true. (www.w3.org)

Time requires more than a download timestamp. PROV-DM's generation event concerns production of an entity, while DataCite distinguishes collection information from temporal coverage and defines a Valid date for the period during which a resource is accurate. The proposed ledger therefore records observation time separately from the time a claim concerns, whenever that information is available. (www.w3.org) (datacite-metadata-schema.readthedocs.io)

Table 2 gives the proposed minimum ledger. The entries are fields to collect, not observations from a completed experiment. The source-context fields adapt questions from Datasheets for Datasets, including collection periods, fees and maintenance. (www.microsoft.com)

Ledger level	Fields to preserve	Purpose in the comparison
Source identity and permission	Publisher, source identifier, evidence type, permitted use, access restrictions and acquisition terms.	Establish the eligible source universe and identify unavailable evidence.
Source overlap and time	Objects covered, overlapping claims, dependence-group hypothesis, observation window, snapshot identifier and claim-validity period.	Distinguish additional observations from independent evidence and temporal disagreement.
Measured integration burden	License expenditure, analyst hours, mapping and entity-linking work, compute, recurring maintenance, and measurement boundary.	Compare actual resource use without converting unlike units implicitly.
Object-level change	Prior claim, new claim, before/after output, selected sources, verification status, reference label, reversal type and adjudication notes.	Explain each net gain or loss on the fixed test set.
Reproducibility package	Input versions, parameters, random seeds, execution environment, output hashes and access instructions.	Enable the result to be recomputed from the retained artifacts.

The ledger joins two questions: where a value came from, and why the reported evaluation changed. The findable, accessible, interoperable and reusable (FAIR) principles call for detailed provenance and for metadata to remain accessible when the underlying data are no longer available. A restricted source can therefore require controlled access while still having a documented identifier, version and availability status. (www.nature.com)

For each changed object, distinguish wrong-to-correct, correct-to-wrong, abstain-to-correct and correct-to-abstain transitions. Also retain wrong-to-abstain and abstain-to-wrong changes because they affect error and coverage even when the correct-answer count does not change. This transition accounting is the proposed analysis unit; paired resampling should preserve correspondence between the before and after outcomes. (docs.scipy.org)

Audit preservation can be evaluated independently of fusion performance. Reliable Context states that any past record state can be reproduced; that is an observable record guarantee, not evidence that any particular subset improves correctness. A study using a record system should test replay against retained inputs and outputs rather than infer an unreported internal selection mechanism.

The release package should distinguish artifacts merely being accessible from results actually being reproduced. The Association for Computing Machinery, or ACM, defines reproducibility around an independent group obtaining the result with the authors' artifacts; its artifact-availability designation does not require formal artifact evaluation. A repository link alone is therefore insufficient evidence of successful reproduction. (www.acm.org)

Analysis of key segments

The primary endpoints should be defined before any segmentation. Let N denote all eligible held-out objects and V the fixed independently verified subset, with size n. For source set S, let C be verified correct answers, W verified wrong answers and B verified abstentions, so C + W + B = n. These are proposed definitions, designed to preserve the distinction between answering and answering correctly found in selective prediction. (www.jmlr.org)

Report **verified correct-answer yield** as C/n, **verified answer coverage** as (C + W)/n, and **conditional accuracy** as C/(C + W). Report overall answer coverage over N separately. If no verified objects are answered, conditional accuracy is undefined, not zero. W3C DQV's requirement to specify measurement units reinforces the need to name these denominators explicitly. (www.w3.org)

The marginal fixed-population change is $\Delta Y = (C_{new} - C_{old})/n$. It may be described as a change in verified accuracy only if that denominator is made explicit. Coverage change is $\Delta Q = ((C_{new} + W_{new}) - (C_{old} + W_{old}))/n$. Because both methods are evaluated on identical objects, a gain cannot be manufactured simply by excluding difficult unanswered cases from one method's denominator. (scikit-learn.org)

Hypothetical Example: on a fixed set of 100 verified objects, a baseline answers 80, of which 72 are correct. Adding a source produces 90 answers, of which 70 are correct. Coverage rises from 80% to 90%, while verified correct-answer yield falls from 72% to 70%. Conditional accuracy falls from 90% to approximately 77.8%. These are illustrative arithmetic values, not observations from a publication or product.

Objects without reference labels require a different status. **Unverifiable** means the evaluation cannot establish correctness; **wrong** requires a supported contradiction with the reference answer. Zhi et al. (2015) study cases where true answers are absent from the candidate set and distinguish silent, falsely spoken and truly spoken behavior. Neither source silence nor candidate consensus supplies the missing independent verification. (pmc.ncbi.nlm.nih.gov)

Prespecify segment reports for:

- **Overlap:** newly covered objects versus objects already covered by the baseline.
- **Reference status:** verified, unresolved and outside the labeling process.
- **Source density:** sparse versus heavily observed objects.
- **Domain:** subject areas defined before inspecting performance.
- **Time:** stable claims versus claims spanning an update boundary.
- **Answer structure:** scalar values, exact sets and partially correct sets.
- **Evidence access:** public evidence, private evidence and blocked acquisition attempts.

Interpret each slice with its sample size and verification rate. Domain-aware multi-truth research supports distinguishing subject-specific reliability and partially correct sets. It does not establish that any proposed segmentation will improve a selector; that remains an empirical question for the registered experiment. (researchportal.hkust.edu.hk)

A verified subset may itself be selective. Informative-missingness research studies observation processes that depend on unobserved values. Its application here is a methodological warning: easily verified objects may differ from the unresolved population. Report verification rates and sensitivity analyses rather than extrapolating measured accuracy to every unverified answer. This is not a measured fusion result from that matrix-completion study. (jmlr.org)

If the fusion system reports probabilities, evaluate calibration separately from rank or accuracy. Guo et al. (2017) define calibration through correctness probabilities, and official scikit-learn guidance cautions that lower Brier loss, a squared probability-error score, need not mean better calibration because other components contribute. A confidence threshold should be locked before evaluating the final accuracy-coverage curve. (proceedings.mlr.press) (scikit-learn.org)

Data Analysis and Evidence

Dong et al.'s paper carries a 2012 copyright and a VLDB 2013 invitation. Its 100-book illustration used title-page checks and added sources in decreasing accuracy, estimated on that same sample. (www.vldb.org)

Table 3 separates published endpoints from their evaluation scope. Values are the authors' reported results unless explicitly identified as arithmetic.

Study or experiment	Reported quantitative evidence	Interpretive boundary
Dong et al., source-selection illustration	Voting reached 93 correct author lists at 583 sources and ended at 78 with all 894; accuracy-weighted ACCU ended at 80. All 100 reference books were covered by 548 sources. (www.vldb.org)	This is an in-sample illustrative curve, not a held-out estimate of the best subset. The voting decrease is 15 percentage points on the fixed sample, calculated from the endpoints. (www.vldb.org)
Li et al. (2012), dependent-group control	Keeping one randomly selected source per identified dependent group increased dominant-value precision from 0.908 to 0.923 for Stock and from 0.864 to 0.927 for Flight. (www.vldb.org)	Real collections, but reference construction differs: Stock uses agreement among reference sites, while Flight uses airline data. (www.vldb.org)
Li et al. (2014), confidence-aware sparse-source experiment	Removing sources with fewer than 1% as many claims as the largest source left 88.07% coverage in the City Population experiment. Coverage fell and mean absolute error increased as small sources were removed. (cse.buffalo.edu)	Dataset-specific threshold experiment, not a general instruction to retain every source.
Lin et al., scalability evaluation	Synthetic source collections ranged from one million to ten million. (arxiv.org)	Their real Book evaluation reused AbeBooks, so it is not independent cross-domain confirmation of that benchmark. (arxiv.org)
Critical Source Selection and influence-graph selection	The former used two real social-sensing traces; the latter used both simulations and real Twitter data. (hub.hku.hk) (research.ibm.com)	Different source-generation processes and objectives prevent pooling these as a common effect size.

The voting decline is study-specific; the sparse-source experiment demonstrates the opposite risk of exclusion. Later reuse of a dataset does not create an independent replication. These boundaries are central to interpreting the table. (www.vldb.org) (cse.buffalo.edu)

Object overlap also differs from attribute overlap. Li et al. report average data-item redundancy of **66% for Stock and 32% for Flight**. A source-count curve without overlap information can therefore compare very different amounts of additional evidence. Publish the overlap definition alongside source counts rather than assigning one generic “redundancy rate” to the collection. (www.vldb.org)

For the proposed experiment, produce paired plots of correct-answer yield, conditional accuracy and coverage against source count. Repeat them against cumulative measured acquisition expenditure, analyst hours and compute, each on its own axis. No cost curve is estimated here because the cited benchmark endpoints do not supply the proposed study's integration costs. DQV supports stating units rather than silently equating them. (www.w3.org)

Bootstrap intervals should resample the same test objects for both methods and recompute their difference. SciPy's paired bootstrap preserves correspondence by using the same resampled indices across arrays. Where observations are clustered, the protocol should resample the independent groups instead; ordinary paired resampling does not itself correct dependence between objects. State which variability the interval represents. (docs.scipy.org) (nips.cc)

For a prespecified binary correct-or-not endpoint, McNemar's test is a possible paired comparison; the National Institute of Standards and Technology (NIST) describes binary paired outcomes and assumes independence between pairs. It is not a complete analysis of coverage, unverified labels or correlated repeated observations. Report discordant transitions and uncertainty rather than relying on a significance label alone. (www.itl.nist.gov)

Finally, repeated searches across subsets, budgets and fusion rules create multiple-comparison questions. Dror et al.'s (2017) work provides a framework for multiple algorithm comparisons, while Conference on Neural Information Processing Systems (NeurIPS) reporting guidance asks authors to identify the variability behind error bars. Separate source-order sensitivity, random-subset variability and object-sampling uncertainty in the reported results. (aclanthology.org) (nips.cc)

Implications and Future Directions

A favorable source set is a conditional finding. It depends on the available sources, observation period, label construction and decision objective. Koh et al.'s WILDS benchmark (2021) demonstrates distribution-shift problems across naturally varying datasets, but it is not a source-selection benchmark. Its relevance is to motivate a second-period or second-domain evaluation rather than to supply a measured penalty for fusion transfer. (proceedings.mlr.press)

Three transfer tests are particularly useful. First, lock the selector and evaluate a later observation window. Second, transfer the trained policy to another population without retuning on its test labels. Third, change cost assumptions while preserving the measured quality outcomes. The first two test generalization; the third tests the stability of a decision under locally chosen resource priorities. The scope of any claim should remain tied to these actual tests. (nips.cc)

Private evidence creates an additional limit: another researcher may be unable to reconstruct the candidate universe or reference labels. Preserve identifiers, permissions, observation windows, selection outputs and controlled-access instructions. FAIR's separation between retained metadata and available underlying data offers a useful documentation principle; it does not remove licensing or confidentiality restrictions. (www.nature.com)

Source drift should trigger measurement rather than an assumption of permanence. The datasheet proposal asks who maintains a dataset and how updates are communicated. A fusion study can extend that record with scheduled reevaluation, version differences and changes in coverage, calibration and error transitions. That extension is a proposed maintenance practice, not a validated universal update frequency. (www.microsoft.com)

Remaining research questions include how to value evidence when integration cost is unknown, how to evaluate truth absent from all candidate claims, and how to assess performance where verification is nonrandom. Truth-existence research and informative-missingness models address parts of these questions under their own assumptions; neither supplies a general answer for all private and public source collections. (pmc.ncbi.nlm.nih.gov) (jmlr.org)

A defensible acquisition decision should present a frontier of quality and resource outcomes, together with reversals and uncertainty. Where a monetary benefit is explicitly elicited, a scalar utility can supplement that frontier. Where it is not, report the tradeoff in its measured units. A documented choice is more informative than a subset labeled “optimal” without a stated objective. (www.w3.org)

Conclusion

Source selection in data fusion should be evaluated through marginal evidence contribution on a fixed population. The literature provides a concrete example in which additional sources reduced correct fused answers, and a separate example in which excluding small sources damaged coverage and error. Neither result supports a universal preference for larger or smaller collections. (www.vldb.org) (cse.buffalo.edu)

The proposed protocol compares all sources, random subsets, accuracy-ranked subsets and the selection method using identical held-out objects and labels. It separates coverage, correct-answer yield and conditional accuracy, preserves unverified outputs as unverified, and records acquisition and integration resources in explicit units. Source-quality estimates, source dependence and answer thresholds belong inside the evaluation design rather than outside its accounting. (scikit-learn.org) (www.w3.org)

The actionable result is a bounded decision: which additional source set changed which answers, how many verified reversals it produced, what coverage it added, and what resources it consumed. Paired uncertainty, prespecified sensitivity analyses and a versioned evidence ledger make that decision inspectable. Provenance supports reconstruction of the result; independent verification supports the correctness assessment. (docs.scipy.org) (www.w3.org)

A selection method earns a stronger claim when its gains persist across declared budgets, source orders and later populations. Until those tests exist, its reported optimum remains local to the evaluated conditions. That boundary is part of the finding and should travel with every recommendation to acquire, retain or exclude another source. (nips.cc)

Source links

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

1. https://www.vldb.org/pvldb/vol6/p37-dong.pdf
2. https://cse.buffalo.edu/~jing/doc/vldb15_CATD.pdf
3. https://www.kdd.org/exploration_files/Article1_17_2.pdf
4. https://scikit-learn.org/stable/common_pitfalls.html
5. https://scikit-learn.org/stable/modules/cross_validation.html
6. https://www.cos.io/initiatives/prereg
7. https://www.jmlr.org/papers/v11/el-yaniv10a.html
8. https://lunadong.com/publication/indep_techReport.pdf
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC4688015
10. https://www.w3.org/TR/vocab-dqv

11. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bootstrap.html
12. https://www.w3.org/TR/prov-dm
13. https://nips.cc/public/guides/PaperChecklist
14. https://kdd.org/explorations/view/december-2015-volume-17-issue-2
15. https://researchportal.hkust.edu.hk/en/publications/domain-aware-multi-truth-discovery-from-conflicting-sources
16. https://arxiv.org/pdf/1610.09506
17. https://research.ibm.com/publications/optimizing-source-selection-in-social-sensing-in-the-presence-of-influence-graphs
18. https://www.microsoft.com/en-us/research/wp-content/uploads/2019/01/1803.09010.pdf
19. https://hub.hku.hk/handle/10722/308892
20. https://datacite-metadata-schema.readthedocs.io/en/4.6/appendices/appendix-1/dateType
21. https://www.nature.com/articles/sdata201618
22. https://www.acm.org/publications/policies/artifact-review-and-badging-current
23. https://jmlr.org/papers/volume23/20-748/20-748.pdf
24. https://proceedings.mlr.press/v70/guo17a.html
25. https://scikit-learn.org/stable/modules/calibration.html
26. https://www.vldb.org/pvldb/vol6/p97-li.pdf
27. https://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/mcnemar.htm
28. https://aclanthology.org/Q17-1033
29. https://proceedings.mlr.press/v139/koh21a.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.