Rethinking the Evaluation of Data Science and Data Management Research

Dimas Cassimiro Nascimento[1], **Daliton da Silva**[1]

[1]Federal University of Agreste of Pernambuco (UFAPE)
Laboratory of Applied Data Intelligence (LIDA)
Garanhuns – PE – Brazil

{dimas.cassimiro,daliton.silva}@ufape.edu.br

Abstract. *The evaluation paradigm in Data Science and Data Management has become predominantly focused on methodological novelty and quantitative metrics, constraining the diversity of valuable contributions. This article presents a framework for recognizing and evaluating broader research outputs, including theoretical foundations, replication studies, reflective analyses, and visionary perspectives. We contribute a critical analysis of evaluation biases, a classification of underrecognized contribution types, a visual mapping of contribution relationships, a formal multi-criteria evaluation model, and practical implementation guidelines. This framework addresses critical gaps in current assessment practices and provides guidance for fostering more inclusive research ecosystems.*

1. Introduction

Scientific publishing in Data Science and Data Management operates within a well-established paradigm where research contributions are predominantly evaluated based on methodological innovation and empirical performance improvements. This ”syntactically correct” article template [Abadi et al. 2020] has undoubtedly propelled technical advancements but has simultaneously narrowed the scope of recognized scholarly contributions. As observed in [Wagstaff 2012], the machine learning community had become overly focused on incremental improvements on benchmark datasets, often losing sight of whether research actually matters for real-world problems. The authors of [Lipton and Steinhardt 2019] identified persistent troubling trends in ML scholarship, including an overemphasis on mathematical sophistication without conceptual clarity and a failure to distinguish genuine insight from speculative interpretation.

Together, these critiques highlight how the field's evaluation practices have systematically privileged technical novelty over foundational understanding, replication, and reflection, precisely the types of contributions that our framework seeks to revalue. These concerns are particularly relevant to modern AI, where large-scale datasets, computational infrastructure, and costly experimental validation can make methodological and benchmark-driven evaluation especially dominant. Such requirements may also disproportionately favor institutions with greater computational and financial resources. The current evaluation ecosystem systematically undervalues crucial forms of knowledge production, such as theoretical conceptualizations that provide foundational clarity, rigorous

replications that ensure result reliability, visionary thinking that charts future directions, and reflective analyses that critically examine research practices.

Data Science and Data Management face unique epistemological challenges that demand diverse contribution types. The interdisciplinary nature of these fields, spanning computer science, statistics, domain expertise, and organizational studies, necessitates multiple forms of scholarly communication [Provost and Fawcett 2013]. Furthermore, the rapid evolution of data technologies and the critical importance of data governance, ethics, and management practices require both reflective assessment and forward-looking vision. This article presents a structured framework for valuing and evaluating a broader diversity of article types, building on seminal reflections from database research [Abadi et al. 2020] and computer science evaluation practices [Prechelt 2021]. Our contributions include: i) A critical analysis of current evaluation biases in Data Science and Management venues, identifying five systematic limitations that constrain epistemological diversity; ii) A classification of contribution archetypes with specific relevance to data-centric research and their relationships; iii) A visual mapping of contribution relationships and complementarities across theoretical-practical and consolidation-innovation dimensions; iv) A formal multi-criteria evaluation model with tailored characteristic weightings and acceptance thresholds adaptable to diverse publication contexts and article types; and v) Practical implementation guidelines for conferences and journals, including track structures, reviewer training protocols, and editorial policies to foster more inclusive research assessment practices.

2. Related Work

The discourse on research evaluation methodologies spans multiple disciplines and has evolved significantly over recent decades. In computer science and information systems, several influential reports and studies have questioned narrow evaluation criteria and advocated for broader recognition of contribution types. The Beckman Report on Database Research [Abadi et al. 2014] highlighted the field's increasing focus on incremental improvements at the expense of foundational and visionary work, arguing that overemphasis on immediate practical impact and quantitative metrics could undermine long-term innovation. Similarly, the Seattle Report [Abadi et al. 2020] emphasized the importance of theoretical contributions and replication studies for building cumulative knowledge in database systems.

In Data Science specifically, [Provost and Fawcett 2013] identified the tension between methodological innovation and practical application, suggesting that both dimensions require recognition. The interdisciplinary nature of Data Science creates particular challenges for evaluation, as contributions may span technical, statistical, and domain-specific knowledge [Dhar 2013]. The reproducibility crisis affecting many scientific domains has prompted increased attention to replication studies in computational fields [Bausell 2021], with several initiatives, such as ML Reproducibility Challenge, emerging to formally recognize replication work. However, these contributions remain marginalized in mainstream Data Science venues [Olszewski et al. 2023]. In a seminal contribution, Ioannidis [Ioannidis 2014] examined systemic interventions to increase the proportion of true research findings, advocating for practices such as large-scale collaboration, replication culture, study registration, data sharing, and more stringent statistical thresholds, many of which resonate directly with the underrecognized contribution types we

seek to promote.

Research on peer review systems has documented systematic biases against novel and interdisciplinary work [Lee et al. 2013], and studies of conference reviewing [McLeish and Strang 2016] have identified limitations in how reviewers assess contributions that don't fit traditional templates. A comprehensive systematic review by Dwan et al. [Dwan et al. 2013] provided empirical evidence of study publication bias and outcome reporting bias, demonstrating that statistically significant results are more likely to be published and fully reported, distorting the available evidence base for decision-making. These findings underscore the need for more nuanced evaluation frameworks.

The authors of [Larivière et al. 2015] showed that scientific publishing has become increasingly concentrated among a small number of major publishers. Although the study does not directly address publication types, this growing consolidation raises concerns about its potential impact on the diversity of research outputs, including the balance between theoretical, experimental, and system-building contributions. More recently, Hernandez and Colom [Hernandez and Colom 2025] conducted a systematic survey of reproducibility policies in scientific computing journals, revealing significant variability in policy enforcement and identifying critical gaps in how journals support computational reproducibility.

Our work builds upon these foundations by developing a comprehensive framework specifically tailored to Data Science and Management, providing not just critique but a structured classification framework, evaluation model, and implementation guidelines to facilitate concrete changes in research assessment practices.

3. Current Evaluation Paradigm: Strengths and Limitations

The prevailing evaluation paradigm in Data Science and Management follows a well-defined template that emphasizes methodological innovation, technical implementation, empirical evaluation, and demonstrable performance gains over existing approaches. Figure 1 illustrates this linear progression from novelty to improvement. While this template has served important functions in maintaining technical rigor and driving measurable progress, it exhibits several significant limitations when applied as a universal standard.

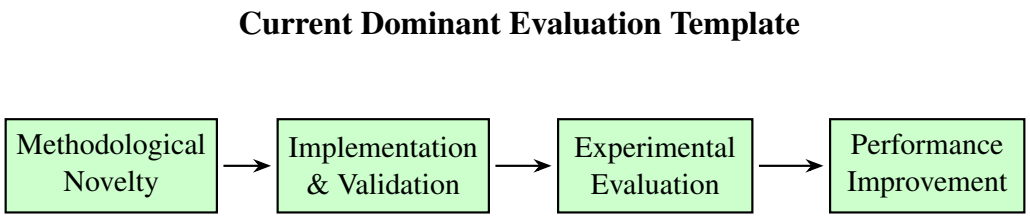


Figure 1. The linear evaluation template predominant in Data Science and Management venues.

3.1. Strengths of the Current Paradigm

The strengths of the current approach include providing clear, relatively objective criteria for assessing research contributions, which benefits authors, reviewers, and editors by establishing shared expectations. The focus on empirical validation helps filter unsubstantiated claims and promotes accountability, while quantitative performance comparisons

facilitate direct assessment of technical advancements and provide benchmarks for future work.

3.2. Weaknesses of the Current Paradigm

Despite these strengths, the current paradigm's limitations have become increasingly apparent. Theoretical marginalization occurs as contributions that develop formal foundations, conceptual frameworks, or mathematical insights without immediate empirical application struggle to meet evaluation criteria [Dhar 2013], impeding development of the theoretical underpinnings necessary for Data Science's maturation as a discipline. Reproducibility challenges emerge from the scarcity of replication studies in mainstream venues, which undermines result verification and knowledge consolidation [Bausell 2021] because researchers have limited incentives to conduct these crucial validation exercises without formal recognition. Visionary constraint appears as speculative or forward-looking contributions that lack immediate empirical support face significant publication barriers, limiting the field's ability to anticipate future directions and engage in long-term planning. A practical-reflective gap exists as insights from real-world data deployments, including lessons from failures and organizational challenges, receive limited recognition [Provost and Fawcett 2013], separating academic research from practical data management concerns. Methodological monoculture may result as overemphasis on a narrow range of contribution types can lead to methodological homogeneity and reduced innovation diversity [Lee et al. 2013].

Beyond these well-documented limitations, the prevailing emphasis on "syntactically correct" articles focused on empirical validation creates additional structural weaknesses that impede scientific progress. First, the very complexity of modern empirical evaluations often renders published studies exceedingly difficult to replicate, even when researchers are motivated to do so. As machine learning techniques become increasingly sophisticated and datasets grow larger, the barrier to independent verification rises correspondingly. Recent investigations have revealed that researchers who rushed to apply machine learning methods without comprehensive understanding of their techniques and limitations produced results that were rarely reproducible [Olszewski et al. 2023]. A survey of machine-learning-based scientific studies found methodological pitfalls such as data leakage affecting hundreds of papers across multiple disciplines, highlighting widespread reproducibility problems in ML-driven research [Kapoor and Narayanan 2023]. When articles achieve acceptance based primarily on impressive empirical results rather than transparent, reproducible methodologies, the scientific record accumulates findings that cannot be reliably verified, undermining the very foundation of cumulative knowledge building.

Second, the paradigm's demand for mature, polished contributions before publication creates significant barriers to incremental scientific progress. Articles are typically only accepted when they present complete, well-packaged solutions with thorough experimental validation, which means that promising but preliminary ideas rarely enter the scholarly conversation. This maturity requirement has several deleterious consequences. It discourages the publication of early-stage work that could benefit from community feedback and collaboration, stifling the kind of open scientific dialogue that accelerates discovery. It also makes it difficult for other researchers to build upon or extend existing work, as the full complexity of mature implementations can be overwhelming to unpack

and modify. When foundational papers present finished systems rather than extensible frameworks, subsequent researchers face steep learning curves simply to understand, let alone improve upon, published methods. This creates what might be termed an innovation bottleneck: the field advances through occasional leaps represented by mature, publishable contributions rather than through continuous, incremental improvements that could be collectively driven by the broader research community.

Third, the emphasis on polished empirical validations often obscures the methodological transparency necessary for genuine scientific progress. Detailed documentation of failed approaches, parameter sensitivity analyses, and implementation pitfalls are routinely omitted from published articles in favor of clean narratives of success. Yet this information is precisely what other researchers need to build upon previous work effectively. Without transparency about what didn't work, under what conditions methods fail, and how sensitive results are to implementation details, subsequent researchers are condemned to repeat the same mistakes, wasting valuable time and resources. This lack of transparency compounds the reproducibility crisis while simultaneously hindering the cumulative development of knowledge that defines mature scientific disciplines.

Evaluation Bias	**Impact on Research Ecosystem**	**Field-Specific Consequences in Data Science**
Empirical Supremacy	Undervalues theoretical and conceptual contributions	Limited development of formal foundations for data-centric disciplines
Novelty Priority	Discourages replication, refinement, and negative results	Reproducibility crisis and insufficient attention to method robustness
Quantitative Focus	Neglects qualitative insights and practical wisdom	Gap between academic research and organizational data challenges
Immediate Impact Emphasis	Constrains long-term and visionary thinking	Limited anticipation of ethical, societal, and technological shifts
Methodological Conformity	Reduces diversity of approaches and perspectives	Homogenization of data science practices and tools

Table 1. Evaluation biases and their consequences in Data Science and Management research.

Table 1 summarizes key evaluation biases and their specific consequences for Data Science and Management. These biases collectively constrain the field's epistemological diversity and long-term health by systematically shaping what types of research are produced and recognized.

4. Research Contribution Archetypes for Data Science and Management

To address the limitations of current evaluation practices, we identify eight contribution archetypes that capture important forms of scholarly work in Data Science and Data Management without claiming to exhaust all possible contribution types. Figure 2 illustrates

selected relationships among these archetypes, showing how different forms of knowledge production may interact and reinforce one another. The proposed classification is extensible, and the archetypes may overlap when a paper performs multiple epistemic functions.

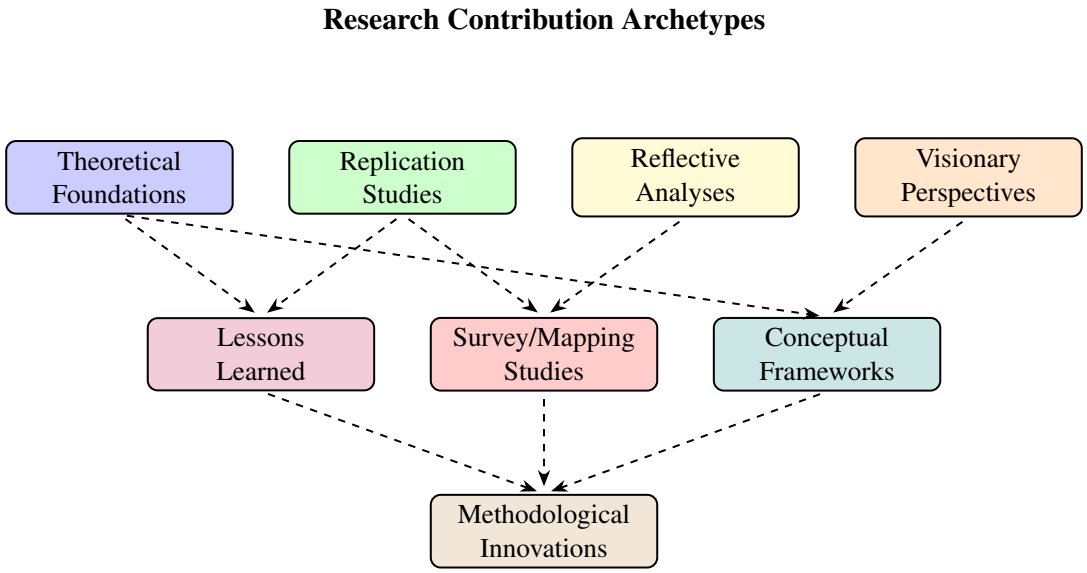


Figure 2. Conceptual relationships among the proposed research contribution archetypes. Dashed arrows represent selected knowledge flows and mutual reinforcement.

Figure 2 illustrates selected functional relationships among the proposed contribution archetypes, highlighting how different forms of scholarly work may interact and reinforce one another. Each contribution type occupies a distinct epistemic role, yet its value becomes clearer when understood in relation to the others. The arrows in the figure do not indicate hierarchy, but knowledge flow and mutual reinforcement across stages of theoretical development, empirical validation, synthesis, and methodological advancement.

Theoretical Foundations provide the formal and conceptual substrate upon which other contributions build. By establishing principles, models, and formal structures, they create the conceptual space later operationalized by Conceptual Frameworks and tested or refined through Replication Studies. Framework articles translate foundational abstractions into integrative structures that connect domains, enabling broader interpretability and cross-context application. In turn, Replication Studies stabilize the knowledge base by validating, challenging, or delimiting previously established theoretical or methodological claims. This interaction between theory construction and empirical verification forms a structural feedback loop essential for epistemic robustness.

Reflective Analyses operate at a meta-level, interrogating assumptions, trajectories, and normative implications of both theoretical and applied work. Their relationship with Survey and Mapping Studies is complementary. While surveys consolidate what has been produced (e.g., [Castro and Aguiar 2023]), reflection critically examines why and how that production occurred. Surveys organize and structure the accumulated literature, identifying patterns and gaps that often motivate both new theoretical inquiries and

methodological refinements. Reflection, by contrast, problematizes implicit commitments and methodological conventions, frequently influencing future visionary or foundational directions.

Visionary Perspectives function as anticipatory mechanisms within the ecosystem. They project possible futures, identify emerging paradigms, and articulate research agendas that may not yet be empirically mature. These forward-looking contributions often stimulate new Conceptual Frameworks and inspire subsequent Methodological Innovations. Their interaction with foundational theory is dialectical: visionary ideas may extend existing theory or challenge its assumptions, thereby initiating cycles of theoretical reformulation and empirical testing.

Lessons Learned contributions anchor the ecosystem in practical reality. By synthesizing insights from real-world deployments, they provide grounded feedback to both methodological and theoretical work. Their relationship with Replication Studies is particularly important, as both address robustness and contextual variability, albeit from different starting points. Practical experiences frequently expose boundary conditions or contextual constraints that motivate methodological adaptation and theoretical refinement.

Methodological Innovations occupy a central integrative position in the structure. They frequently draw upon theoretical principles, are evaluated through replication efforts, are contextualized through lessons learned, and are systematized through surveys. Although often perceived as the dominant form of contribution in computational fields, methodological advances are structurally dependent on the stability provided by replication, the coherence supplied by conceptual frameworks, and the strategic direction suggested by visionary and reflective work.

Collectively, these relationships show that scientific progress does not follow a single linear path: different contribution types can initiate, consolidate, challenge, or redirect one another. Foundational theory informs frameworks and methods; replication and surveys consolidate and validate; reflection critiques and recalibrates; visionary perspectives expand the horizon of inquiry; lessons learned reconnect research to practice; and methodological innovation translates conceptual advances into operational capability. The strength of the field depends not on the predominance of any single contribution type, but on the dynamic equilibrium among them. Recognizing these relationships makes explicit the functional complementarity that sustains cumulative, reliable, and societally relevant advancement in Data Science and Data Management.

5. Visual Mapping of Contribution Relationships

Understanding how different contribution types relate to and complement each other is crucial for appreciating the full research ecosystem. Figure 3 presents a comprehensive mapping of these relationships across two key dimensions: theoretical-practical orientation and consolidation-innovation focus. This mapping reveals several important patterns in how different contribution types complement each other.

The theoretical-practical spectrum shows contributions ranging from highly theoretical (e.g., foundational principles) to strongly practical (e.g., lessons learned), with many types occupying intermediate positions that bridge this divide. The innovation-consolidation balance distinguishes between contributions that primarily drive innovation

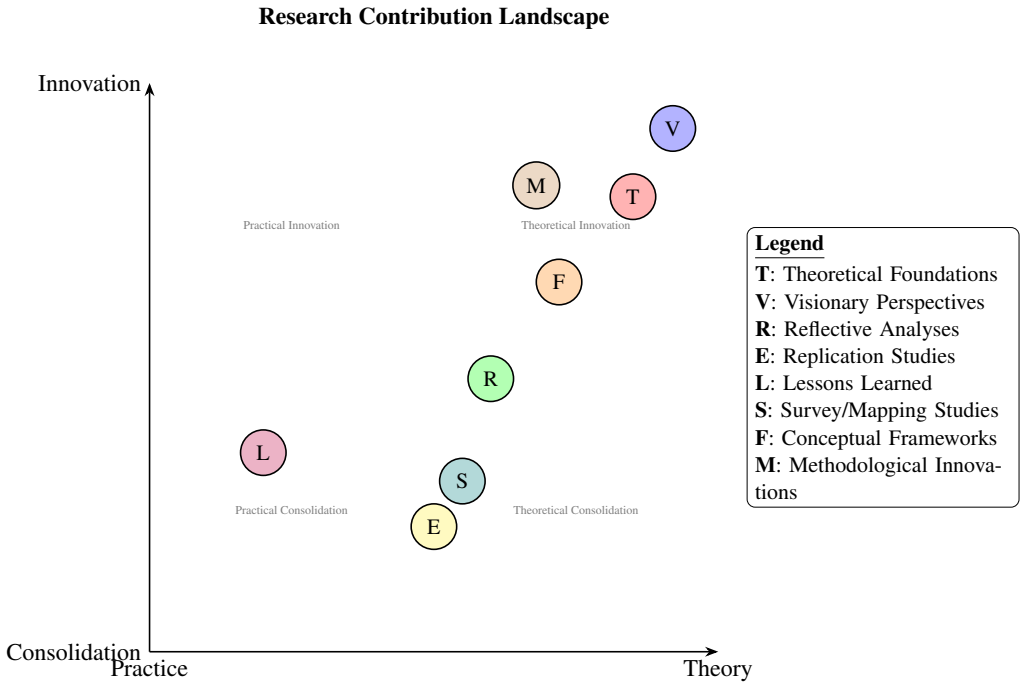


Figure 3. Visual mapping of research contributions across theory–practice and consolidation–innovation dimensions. Each contribution type is positioned according to its dominant epistemic orientation, revealing structural complementarities and distributional patterns within the research ecosystem.

(e.g., visionary articles) and those that focus on knowledge consolidation (e.g., replication studies). Complementarity clusters appear as certain contribution types naturally cluster together, suggesting natural synergies; for example, theoretical, framework, and reflective articles often inform and strengthen each other. Coverage gaps can be revealed through this mapping, showing areas where certain types of contributions are underrepresented and suggesting opportunities for future work. This visual representation helps researchers, reviewers, and editors appreciate the diverse roles different contributions play in advancing the field.

6. Formal Decision Model

To operationalize the evaluation of diverse contribution types, we propose a formal multi-criteria model that can be adapted by conferences and journals. Let $T \in \mathcal{T}$ denote an article type and let $C = \{1, \ldots, 6\}$ be the set of evaluation characteristics. For each characteristic $i \in C$, let $s_i \in [0, 1]$ denote the score assigned to criterion i, w_i^T the weight associated with criterion i for article type T, θ_i^T the minimum individual threshold required for that criterion and type, and Θ_T the global acceptance threshold defined for type T. The weights are assumed to satisfy the following normalization constraints:

$$\sum_{i=1}^{6} w_i^T = 1 \quad \text{and} \quad w_i^T \geq 0. \tag{1}$$

Given a score vector $s = (s_1, \ldots, s_6)$, the aggregated score for an article of type

T is defined as: $S_T(s) = \sum_{i=1}^{6} w_i^T s_i$. Acceptance is therefore determined through a hybrid rule that combines non-compensatory constraints on critical characteristics with compensatory aggregation across all criteria. Let $C_T^* \subseteq C$ denote the subset of critical characteristics for article type T. The decision function is defined as:

$$\text{Accept}(T, s) = \begin{cases} 1, & \text{if } \left(s_i \geq \theta_i^T \ \ \forall i \in C_T^*\right) \wedge \left(S_T(s) \geq \Theta_T\right), \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Therefore, an article is accepted if and only if it simultaneously satisfies all minimum thresholds for its critical characteristics and achieves a sufficient aggregated weighted score. This hybrid structure ensures that essential quality dimensions are not compensable by excellence in unrelated criteria.

Table 2 defines core evaluation characteristics and provides illustrative weightings for different contribution types. These values are not empirically estimated or intended as universal prescriptions; they encode the relative importance assigned to each characteristic under the proposed evaluation logic. In practice, venues may calibrate these parameters according to their scope and community priorities through expert elicitation, reviewer calibration, or retrospective analysis of past decisions.

Article Type	**Orig.**	**Rig.**	**Sig.**	**Clar.**	**Comp.**	**Cont.**
Theoretical Foundations	0.20	0.30	0.15	0.10	0.15	0.10
Visionary Perspectives	0.30	0.15	0.25	0.15	0.05	0.10
Replication Studies	0.10	0.40	0.10	0.10	0.20	0.10
Reflective Analyses	0.15	0.15	0.20	0.15	0.15	0.20

Table 2. Evaluation characteristics and suggested weightings for different article types. Orig. = Originality, Rig. = Rigor, Sig. = Significance, Clar. = Clarity, Comp. = Completeness, Cont. = Contextualization.

Different article types require tailored evaluation protocols that emphasize their distinctive qualities. Table 2 summarizes the critical characteristics for each contribution type, while Table 3 proposes specific minimum thresholds for these characteristics and global acceptance scores.

For theoretical articles, critical characteristics include rigor, originality, and completeness, with minimum thresholds of $\theta_{\text{rigor}} = 0.90$, $\theta_{\text{originality}} = 0.80$, and $\theta_{\text{completeness}} = 0.80$, and a global threshold of $\Theta_T = 0.80$. Evaluation should primarily assess the soundness and internal consistency of formal development, the originality of conceptual or mathematical contributions, and the completeness with which foundational implications and boundary conditions are articulated. The relatively high individual thresholds reflect the non-compensatory importance of epistemic robustness in foundational work, even though the global requirement allows balanced performance across additional dimensions.

Visionary articles prioritize originality, significance, and clarity, with thresholds of $\theta_{\text{originality}} = 0.90$, $\theta_{\text{significance}} = 0.70$, $\theta_{\text{clarity}} = 0.70$, and a global threshold of $\Theta_T = 0.85$. Assessment should emphasize the novelty of the proposed direction, the plausibility and potential long-term impact of the vision articulated, and the clarity with which future

challenges and opportunities are framed. The higher global threshold reflects the expectation that visionary contributions, while less dependent on formal completeness, must nevertheless present a coherent and compelling strategic orientation for the field.

Replication studies require rigor, completeness, and contextualization, with thresholds of $\theta_{\text{rigor}} = 0.90$, $\theta_{\text{completeness}} = 0.80$, $\theta_{\text{contextualization}} = 0.70$, and a global threshold of $\Theta_T = 0.80$. Evaluation should focus on methodological transparency, fidelity to original protocols when appropriate, careful documentation of deviations, and explicit analysis of contextual factors influencing reproducibility. The threshold structure underscores that replication quality depends not on novelty but on technical precision, documentation completeness, and interpretative clarity regarding boundary conditions.

Reflective articles require significance, clarity, and contextualization, with thresholds of $\theta_{\text{significance}} = 0.80$, $\theta_{\text{clarity}} = 0.75$, $\theta_{\text{contextualization}} = 0.90$, and a global threshold of $\Theta_T = 0.85$. Assessment should consider the depth of critical engagement, the relevance of the issues examined to ongoing developments in Data Science and Data Management, and the strength of contextual grounding in historical, methodological, or societal perspectives. The elevated contextualization requirement reflects the expectation that reflective analyses must be firmly anchored in the evolution and current state of the field in order to generate meaningful and actionable reflections and conclusions.

These thresholds are illustrative rather than empirically validated and should be calibrated by each venue together with the corresponding weights. The global threshold Θ_T represents the minimum weighted sum required for acceptance, while critical characteristic thresholds ensure that articles meet essential quality standards for their contribution type.

Article Type	**Characteristic**	**Threshold** (θ_i)	**Global Threshold** (Θ_T)
Theoretical Foundations	Rigor	0.90	0.80
	Originality	0.80	
	Completeness	0.80	
Visionary Perspectives	Originality	0.90	0.85
	Significance	0.70	
	Clarity	0.70	
Replication Studies	Rigor	0.90	0.80
	Completeness	0.80	
	Contextualization	0.70	
Reflective Analyses	Significance	0.80	0.85
	Clarity	0.75	
	Contextualization	0.90	

Table 3. Suggested minimum thresholds for critical characteristics and global acceptance scores by contribution type. The global threshold Θ_T represents the minimum weighted sum required for acceptance, while θ_i indicates the minimum score required for each critical characteristic.

Implementation Remarks. The proposed model implies several implementation considerations. Hybrid contributions may declare one primary and one or more secondary

types. Evaluation should be anchored in the primary contribution, while secondary types may inform the assessment without requiring the paper to satisfy the complete criteria of several independent article categories. It is also important to note that venue customization allows program committees to adjust weights and thresholds to reflect specific venue priorities and scope. In turn, reviewer calibration through training and guidelines may help reviewers consistently apply characteristic scores across different submission types. Finally, publishing transparent evaluation criteria helps authors understand venue expectations and prepare their submissions accordingly.

Implementation Guidelines for Venues. Successfully implementing this expanded evaluation framework requires systematic changes to publication processes and community practices. Implementation requires contribution-specific review guidelines, representative examples, reviewer calibration, and explicit guidance against applying inappropriate standards across contribution types. Also, reviewer guidelines should explicitly address common evaluation pitfalls, such as applying methodological standards to theoretical work or demanding immediate practical impact from visionary contributions.

7. Conclusions and Future Directions

Our proposed framework addresses critical limitations in current evaluation practices by providing a structured approach to recognizing and evaluating a broader spectrum of research outputs in Data Science and Data Management. We have contributed a critical analysis of current evaluation biases identifying five systematic limitations that constrain epistemological diversity, a classification of eight contribution archetypes and their functions within the research ecosystem, a visual mapping of contribution relationships across theoretical-practical and consolidation-innovation dimensions, a formal multi-criteria evaluation model with tailored characteristic weightings and acceptance thresholds, and practical implementation guidelines for conferences and journals seeking to foster more inclusive research assessment practices. Ultimately, by valuing diverse forms of scholarly contributions, the field can develop more comprehensive understanding, more robust methodologies, and more meaningful impact on both theory and practice.

Future research directions include empirical validation of the framework across different contexts, field-specific adaptation to particular Data Science subfields, development of automated assessment support tools, longitudinal impact studies tracking how different contribution types influence subsequent research, and cross-field comparison with related disciplines.

References

Abadi, D., Ailamaki, A., Andersen, D., Bailis, P., Balazinska, M., Bernstein, P., Boncz, P., Chaudhuri, S., Cheung, A., Doan, A., Dong, L., Franklin, M. J., Freire, J., Halevy, A., Hellerstein, J. M., Idreos, S., Kossmann, D., Kraska, T., Krishnamurthy, S., Markl, V., Melnik, S., Milo, T., Mohan, C., Neumann, T., Ooi, B. C., Ozcan, F., Patel, J., Pavlo, A., Popa, R., Ramakrishnan, R., Ré, C., Stonebraker, M., and Suciu, D. (2020). The seattle report on database research. *ACM SIGMOD Record*, 48(4):44–53.

Abadi, D. J., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P. A., Carey, M. J., Chaudhuri, S., Dean, J., Doan, A., Franklin, M. J., Gehrke, J., Haas, L. M., Halevy, A. Y., Hellerstein, J. M., Ioannidis, Y. E., Jagadish, H. V., Kossmann, D., Madden,

S., Mehrotra, S., Milo, T., Naughton, J. F., Ramakrishnan, R., Markl, V., Olston, C., Ooi, B. C., Ré, C., Suciu, D., Stonebraker, M., Walter, T., and Widom, J. (2014). The beckman report on database research. *ACM SIGMOD Record*, 43(3):61–70.

Bausell, R. B. (2021). *The Problem with Science: The Reproducibility Crisis and What to Do About It*. Oxford University Press.

Castro, J. P. C. and Aguiar, C. D. (2023). Big data architectures for fair-compliant repositories: A systematic review. In *Proceedings of the Brazilian Symposium on Databases (SBBD)*, pages 76–88. SBC.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Dwan, K., Gamble, C., Williamson, P. R., and Kirkham, J. J. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PLOS ONE*, 8(7):e66844.

Hernandez, J. A. and Colom, M. (2025). Reproducible research policies and software/data management in scientific computing journals: A survey, discussion, and perspectives. *Frontiers in Computer Science*, 6:1491823.

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, 11(10):e1001747.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804.

Larivière, V., Haustein, S., and Mongeon, P. (2015). The oligopoly of academic publishers in the digital era. *PLOS ONE*, 10(6):e0127502.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

Lipton, Z. C. and Steinhardt, J. (2019). Troubling trends in machine learning scholarship. *Queue*, 17(1):45–77.

McLeish, T. and Strang, V. (2016). Evaluating interdisciplinary research: The elephant in the peer-reviewers' room. *Palgrave Communications*, 2:16055.

Olszewski, D., Lu, A., Stillman, C., Warren, K., Kitroser, C., Pascual, A., Ukirde, D., Butler, K. R. B., and Traynor, P. (2023). "Get in Researchers; We're Measuring Reproducibility": A reproducibility study of machine learning papers in tier 1 security conferences. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3433–3459.

Prechelt, L. (2021). On implicit assumptions underlying software engineering research. In *Evaluation and Assessment in Software Engineering (EASE 2021)*, pages 336–339. ACM.

Provost, F. and Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59.

Wagstaff, K. L. (2012). Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning*, pages 529–536.