Beyond the Data Science Unicorn: A Competency-Stratified Workforce Framework for Resolving Role Ambiguity, Enabling Structured Team Design, and Aligning Education with Practice

Mayank Gupta
Guru Gobind Singh Indraprastha University (GGSIPU), Delhi
Maharaja Agrasen Institute of Technology (MAIT)

[1]*Date of Receiving: 13/07/2025*
Date of Acceptance:04/09/2025
Date of Publication: 26/10/2025

Abstract

Data science has matured from an experimental specialty into a core organizational capability, yet the workforce that performs this work remains poorly defined. Across industry, academia, and government, job titles such as data scientist, data analyst, machine learning engineer, data engineer, analytics translator, and AI specialist are often used inconsistently, even when they refer to overlapping tasks. This ambiguity creates practical problems in hiring, curriculum design, team formation, career progression, and project governance. Recent studies show that employers continue to seek multidisciplinary workers with both technical and soft skills, while research on collaboration demonstrates that data science is rarely an individual activity and instead depends on coordinated work across specialized roles. At the same time, competency frameworks have emerged, but they are not yet universally adopted or consistently mapped to real job families, proficiency levels, and stages of the data lifecycle. This paper argues that role ambiguity is not merely a labeling issue but a structural workforce problem. It reviews current evidence on the fragmentation of data science roles, explains the organizational costs of this fragmentation, and proposes the core design principles of a data science workforce framework. Such a framework should define role families, competency domains, proficiency levels, and workflow responsibilities while remaining adaptable to domain context and technological change. A clearer workforce framework would help employers build balanced teams, help educators align curricula with practice, and help practitioners navigate sustainable career paths in data science.

Keywords: *data science roles; workforce framework; competencies; job ambiguity; data science teams; data value chain*

1. Introduction

The expansion of data-intensive work has increased demand for data professionals, but the field still lacks shared agreement on who does what. Early work on data science roles already warned that organizations were using titles loosely and expecting "unicorn" employees to cover business understanding, data engineering, analytics, modeling, and communication in one person. Saltz and Grady's 2017 study directly framed this as ambiguity in team roles and argued for a workforce framework, while later studies continued to show that the market uses multiple overlapping labels for related functions. More recent research suggests that the field has moved toward specialization, but not yet

[1] *How to cite the article:* Gupta M (October 2025); Beyond the Data Science Unicorn: A Competency-Stratified Workforce Framework for Resolving Role Ambiguity, Enabling Structured Team Design, and Aligning Education with Practice; *International Journal of Law, Management and Social Science*, Vol 9, Issue 4, 47-51

toward full clarity. This combination of growth and conceptual instability is the central problem addressed in this paper.

Role ambiguity matters because data science is not just a technical occupation. It is an organizational system of work that joins data acquisition, engineering, modeling, interpretation, governance, and decision support. When role boundaries are unclear, organizations struggle to recruit the right people, educational institutions struggle to teach the right mix of competencies, and practitioners struggle to understand career expectations. In other words, the ambiguity of data science roles affects both labor supply and labor demand. A workforce framework is therefore needed not only to classify occupations, but to improve how data science work is designed, coordinated, taught, and evaluated.

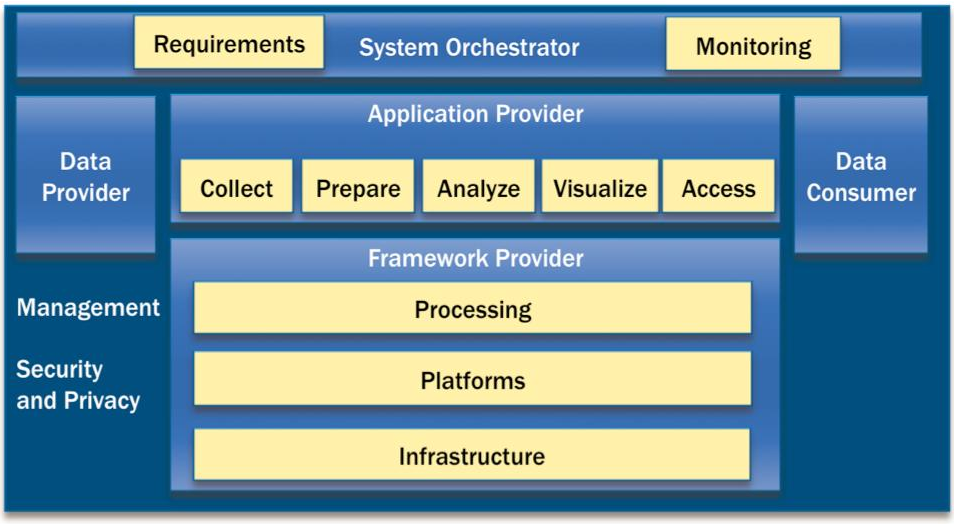

Figure 1. Big data reference architecture

2. Evidence of Ambiguity in Data Science Roles

The literature repeatedly shows that data science job definitions are broad, inconsistent, and highly hybridized. De Mauro et al. classified big data professions systematically and demonstrated that the labor market already contained multiple distinct job families with different skill sets rather than a single unified profession. Costa and Santos likewise showed that the data scientist profile only partially fits existing ICT competence frameworks, which suggests that the role sits across several established professional categories rather than within one. Saltz and Grady later found recurring but inconsistently defined roles, including data scientist, data engineer, data architect, data science programmer, data science researcher, and data analyst. Together, these findings show that fragmentation is not recent; it has been built into the profession from the start.

Job advertisement studies reinforce the same conclusion. Gardiner et al. found that big data postings demanded diverse and sometimes mismatched combinations of technical and managerial capabilities. Coelho da Silveira et al. reported that employers often prioritized a multidisciplinary profile and valued soft skills such as communication, teamwork, and problem solving alongside Python and SQL. Smaldone et al., using U.S. job advertisements, argued that understanding employability criteria is essential because employer expectations are evolving rapidly and can create skills mismatches. Vogt et al. later used more than 5,000 job postings to build a competency data set and explicitly argued that static expert-driven frameworks need more dynamic updating from labor-market evidence. These studies indicate that the problem is not a shortage of titles, but an unstable mapping between titles, tasks, and competencies.

Recent work shows that specialization is increasing, but standardization remains incomplete. Fayyad and Hamutcu proposed three anchor role families, data analyst, data scientist, and data engineer, as a practical way to escape the “unicorn” problem. In contrast, Gunklach et al. analyzed 16,348 job advertisements and identified nine important

roles, arguing that organizations should align skills with the data value chain and contextual needs. This is an important development: the field is moving from vague generalism toward structured specialization, but there is still no universally adopted model. That gap is exactly why a workforce framework is needed.

3. Why Ambiguity Becomes a Team Problem

Role ambiguity would be less harmful if data science were mostly individual work, but it is not. Zhang, Muller, and Wang found that data science workers are highly collaborative and interact with a range of stakeholders and tools across six common workflow stages. Their findings show that data science teams depend on coordination practices, shared artifacts, and communication between technical and nontechnical participants. This means that unclear role boundaries do not stay confined to hiring documents; they directly affect handoffs, accountability, documentation, and project execution.

This team-based reality also exposes the limits of role labels. A data scientist may clean data in one company, build models in another, and act as a client-facing translator in a third. A data engineer may focus on pipelines, governance, or platform architecture depending on the organization. Business users may be passive recipients of dashboards in one setting and active co-creators in another. Gunklach et al. describe data science as a "team sport," while Dorr et al. show that competency adaptation by role and purpose is necessary in fast-changing domains. In practice, then, ambiguity arises because organizations are trying to combine stable labels with unstable local needs.

The consequences are substantial. First, hiring becomes inefficient because job descriptions bundle too many unrelated expectations. Second, team design suffers because leaders cannot distinguish between complementary and redundant capabilities. Third, professional development becomes uneven because employees do not know which competencies belong to their role and which belong to adjacent roles. Fourth, educational programs risk teaching generic data science without showing learners how competencies translate into actual job families. The result is a field in which role names appear clear on paper but remain operationally blurred in practice.

4. The Limits of Existing Competency Models

Several frameworks have attempted to bring order to data science education and workforce planning. Hattingh et al.'s systematic literature review argued that a unified competency model is crucial for building a competitive workforce. Schmitt et al.'s evaluation of the EDISON competency framework described the EDISON project as one of the most complete attempts to define data science knowledge and skills, while also noting disagreement over what should properly count as data science. These studies are valuable because they move the discussion from job titles to competencies. However, they also show that competency models alone do not fully solve the role problem.

The main limitation is that many competency models are broad but not role-explicit. They identify what data science may include, but not always who should own which tasks at different proficiency levels and in different contexts. Vogt et al. directly criticized static descriptions and argued for continuous updating through job advertisement analysis. Zarefard and Marsden's 2024 framework for hiring and training also points toward a more structured competency-based approach, yet the practical challenge remains translating competencies into team architecture. A workforce framework must therefore go beyond a list of skills. It must connect competencies to role families, project workflows, and organizational maturity.

5. What a Data Science Workforce Framework Should Include

A useful workforce framework should begin with a small set of clearly defined role families. The literature supports at least three anchor roles, data analyst, data scientist, and data engineer, with room for additional specializations such as machine learning engineer, analytics translator, data architect, and domain AI specialist where needed. The framework should define each role by primary purpose, typical outputs, decision rights, and dependencies on other

roles. This would reduce the tendency to write job descriptions that ask one person to perform every stage of the data lifecycle.

Second, the framework should organize competencies into domains rather than treating them as a flat list. Across the literature, five domains appear consistently: technical and engineering skills; statistics and analytical reasoning; domain and business understanding; communication and collaboration; and governance, ethics, and professional practice. Coelho da Silveira et al. and Smaldone et al. both show that employers value soft and technical competencies together, not separately. That finding is critical because it means workforce design must treat communication and problem framing as core capabilities, not optional extras.

Third, competencies should be mapped to stages of the data value chain. Gunklach et al. argue that a skill-based perspective should align roles with the demands of the data value chain, and their analysis shows why role design must reflect the progression from data acquisition and processing to analysis, visualization, deployment, and use. A workforce framework should therefore specify where each role contributes most strongly, where responsibilities overlap, and where collaboration is mandatory. This would improve handoffs and reduce confusion over ownership.

Fourth, the framework should include proficiency levels. Entry-level, practitioner, senior, lead, and strategic levels should not merely signal years of experience; they should indicate expanded scope, complexity, leadership, and accountability. The same competency, such as model evaluation or stakeholder communication, looks different at junior and senior levels. Without proficiency tiers, role definitions remain too generic to guide hiring, training, or promotion. Recent work in domain-adapted competencies also supports this layered view by showing that competencies must be adjusted for learner role and purpose.

Finally, a workforce framework should be evidence-based and continuously updated. Job markets, tooling, and AI capabilities are changing too quickly for static role definitions to remain reliable for long. Vogt et al. make a strong case for using job advertisement analysis to monitor competency demand, while Dorr et al. show that rapidly evolving AI domains require adaptable role-specific competency design. A modern framework should therefore combine stable structure with periodic revision based on labor-market evidence and domain change.

6. Conclusion

The ambiguity of data science team roles is not a temporary side effect of a young discipline. It is a persistent organizational problem produced by the intersection of multidisciplinary work, fast-changing tools, inconsistent job labels, and insufficiently role-specific competency models. The literature now provides enough evidence to move beyond general discussion. Data science is collaborative, specialized, and distributed across the data lifecycle, which means organizations need a workforce framework that explicitly links role families, competencies, proficiency levels, and workflow responsibilities.

A well-designed workforce framework would deliver benefits across the ecosystem. Employers could write clearer job descriptions and build more balanced teams. Universities could align curricula with actual role pathways instead of teaching an undifferentiated “data scientist” ideal. Practitioners could understand how to grow from one role into another. Most importantly, organizations could replace the unrealistic search for a single all-purpose expert with a more sustainable model of complementary expertise. The field does not need fewer data science roles. It needs clearer ones.

References

Coelho da Silveira, C., Marcolin, C. B., da Silva, M., & Domingos, J. C. (2020). What is a data scientist? Analysis of core soft and technical competencies in job postings. *Revista Inovação, Projetos e Tecnologias, 8*(1), 25–39. https://doi.org/10.5585/iptec.v8i1.17263

De Mauro, A., Greco, M., Grimaldi, M., & Ritala, P. (2018). Human resources for big data professions: A systematic classification of job roles and required skill sets. *Information Processing & Management, 54*(5), 807–817. https://doi.org/10.1016/j.ipm.2017.05.004

Dorr, D. A., Krussel, A., Hauck, R., Jackson, C., Dalal, A., Bedrick, S., Payne, P. R. O., Bridge2AI-Voice Consortium, & Hersh, W. (2025). Adapting data science competencies by role and purpose: Voice AI. *Frontiers in Digital Health, 7*, Article 1610253. https://doi.org/10.3389/fdgth.2025.1610253

Gottipati, S., Shim, K. J., & Sahoo, S. (2021). Glassdoor job description analytics: Analyzing data science professional roles and skills. In *Proceedings of the 2021 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1329–1336). https://doi.org/10.1109/EDUCON46332.2021.9453931

Gunklach, J., Nadj, M., Michalczyk, S., Jacob, K., Gröger, C., & Mädche, A. (2025). Beyond the unicorn? Job roles in data science. *Business & Information Systems Engineering*. Advance online publication. https://doi.org/10.1007/s12599-025-00954-2

Hattingh, M., Marshall, L., & Seymour, L. F. (2019). Data science competency in organisations: A systematic literature review. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists 2019*. https://doi.org/10.1145/3351108.3351110

Saltz, J. S., & Grady, N. W. (2017). The ambiguity of data science team roles and the need for a data science workforce framework. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)* (pp. 2355–2361). https://doi.org/10.1109/BigData.2017.8258190

Schmitt, K. R. B., Clark, L., Kinnaird, K. M., Wertz, R. E. H., & Sandstede, B. (2023). Evaluation of EDISON's data science competency framework through a comparative literature analysis. *Foundations of Data Science, 5*(2), 177–198. https://doi.org/10.3934/fods.2021031

Smaldone, F., Ippolito, A., Lagger, J., & Pellicano, M. (2022). Employability skills: Profiling data scientists in the digital labour market. *European Management Journal, 40*(5), 671–684. https://doi.org/10.1016/j.emj.2022.05.005

Vogt, J., Voigt, T., Nowak, A., & Pawlowski, J. M. (2023). Development of a job advertisement analysis for assessing data science competencies. *Data Science Journal, 22*, Article 33. https://doi.org/10.5334/dsj-2023-033

Zarefard, M., & Marsden, N. (2024). The essential competencies of data scientists: A framework for hiring and training. In *Human Interface and the Management of Information* (Lecture Notes in Computer Science). https://doi.org/10.1007/978-3-031-60125-5_27

Zhang, A. X., Muller, M., & Wang, D. (2020). How do data science workers collaborate? Roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), Article 22. https://doi.org/10.1145/3392826