

Data Science: Foundations, Methodologies, Applications, and Emerging Challenges — A Comprehensive Review

Dushyant Kaushik

Assistant Professor, MERI College of Engineering & Technology

Abstract

Data science has grown from a loosely defined intersection of statistics and computer science into one of the most consequential disciplines shaping contemporary research, industry, and public policy. This paper presents a structured review of the field, tracing its conceptual origins, examining the methodological frameworks that guide applied projects, surveying its principal application domains, and critically assessing the ethical and workforce-related challenges that accompany its rapid expansion. The review synthesises evidence from peer-reviewed literature, industry research, and professional bodies to argue that data science is best understood not as a single technique but as a socio-technical practice whose success depends as much on organisational process, governance, and human judgement as on algorithmic sophistication. Particular attention is paid to the persistent gap between technical capability and reliable real-world deployment, to algorithmic bias and privacy as central ethical concerns, and to the way generative artificial intelligence is reshaping the skills and roles the discipline demands. The paper concludes that the next phase of data science will be defined less by raw analytical power and more by the discipline's capacity to institutionalise trustworthy, transparent, and reproducible practice.

Keywords: *data science, big data, machine learning, CRISP-DM, data ethics, algorithmic bias, data governance, generative artificial intelligence, workforce skills*

1. Introduction

Over the past two decades, the term "data science" has moved from a niche label used by a handful of statisticians to a mainstream discipline that underpins decision-making across healthcare, finance, retail, manufacturing, and government. This expansion has been driven by three simultaneous developments: an explosion in the volume and variety of digitally recorded information, dramatic increases in affordable computing power, and the maturation of machine learning algorithms capable of extracting patterns from that information at scale. Yet, despite its prominence, data science remains a discipline whose boundaries are contested. Some scholars treat it as a rebranding of applied statistics; others regard it as a genuinely new transdisciplinary field with its own theoretical commitments, professional norms, and modes of knowledge production.

This review paper does not attempt to resolve that definitional debate outright, but instead takes the position, informed by the literature surveyed below, that data science is most usefully understood as an applied, problem-driven practice that draws on statistics, computer science, and domain expertise to convert raw data into decisions and knowledge. The purpose of this review is fourfold: first, to trace the historical and conceptual evolution of the field; second, to examine the methodological frameworks that practitioners use to structure data science projects; third, to survey representative applications across major industry sectors; and fourth, to critically evaluate the ethical, governance, and workforce challenges that will shape the discipline's next phase of development. The review is organised so that each of these four aims corresponds to a dedicated section, followed by a discussion that draws the threads together and a concluding summary.

2. Historical Evolution and Conceptual Foundations

2.1 From Statistics to a Distinct Discipline

The lineage of data science is usually traced to the statistical community rather than to computer science alone. As early as 1962, the statistician John Tukey argued that his own central interest had shifted from formal inference toward what he called data analysis, an empirical activity concerned with extracting meaning from observations rather than testing

pre-specified hypotheses (Priestley & McGrath, 2019). The term itself is generally attributed to the Danish computer scientist Peter Naur, who used "data science" repeatedly in his 1974 Concise Survey of Computer Methods to describe the systematic handling and interpretation of data. Decades later, the statistician William S. Cleveland proposed an explicit action plan for expanding the technical scope of statistics under the data science label, a proposal that is widely regarded as the moment the term began to acquire its modern connotation (Priestley & McGrath, 2019).

Priestley and McGrath (2019) offer one of the more theoretically grounded accounts of this evolution, framing data science as an instance of what Gibbons and colleagues termed "Mode 2" knowledge production: a form of inquiry that is organised around real-world application rather than disciplinary purity, that draws on collaborators from both academia and industry, and that is subject to a broader and more socially accountable form of quality control than traditional peer review alone. Their analysis of graduate programmes in the United States found that a majority of master's-level data science programmes, and a substantial share of doctoral programmes, are housed outside conventional academic departments, which is itself evidence of the discipline's transdisciplinary and applied character. Other historical accounts converge on a similar timeline, generally locating the term's popularisation in the early 2000s and its rapid institutionalisation during the big data boom of the 2010s, when distributed computing frameworks made it economically feasible to store and analyse data at previously unimaginable scale (Dataversity, 2025; Press, 2013).

2.2 Defining Data Science and Big Data

Despite widespread use of the term, no single definition of data science commands universal agreement. Martínez, Viles, and Olaizola (2021) survey a range of formulations, noting that some authors define data science as the intersection of computer science, statistics, and domain expertise, while others emphasise the combination of mathematical knowledge, programming skill, and business acumen. After reviewing this landscape, Martínez et al. (2021) settle on a definition that treats data science as a multidisciplinary field situated between computer science, mathematics, and statistics, oriented toward using scientific methods to extract knowledge and value from large volumes of structured and unstructured data. This review adopts a closely related view: data science is the applied practice of turning data into reliable, actionable knowledge through a combination of statistical reasoning, computational technique, and contextual judgement.

Closely related to data science is the concept of big data, which refers not to a technique but to a class of datasets whose characteristics exceed the capacity of conventional data-processing tools. The most widely cited characterisation is Doug Laney's "three Vs" framework, introduced in a 2001 industry analysis, which described big data along the dimensions of volume (the sheer scale of data generated), velocity (the speed at which it is produced and must be processed), and variety (the heterogeneity of data types and structures) (Nexla, 2025; TechTarget, n.d.). Later commentators have extended this framework to include veracity, the reliability and accuracy of the data, and value, the practical benefit that can be extracted from it, producing a "five Vs" model that is now common in both the academic and industry literature (Martínez et al., 2021; Nexla, 2025). The distinction matters conceptually: big data describes the raw material, while data science describes the disciplined process of converting that material into insight.

3. Methodological Frameworks for Data Science Projects

3.1 CRISP-DM and Its Legacy

The single most influential process model in the field remains the Cross-Industry Standard Process for Data Mining (CRISP-DM), developed in the mid-1990s by a consortium that included SPSS, Daimler-Benz, and NCR. CRISP-DM decomposes a project into six iterative phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Data Science PM, 2024; KDnuggets, 2024). Successive KDnuggets community polls, beginning in 2002 and continuing through 2014, consistently found CRISP-DM to be the most widely used methodology among analytics and data mining practitioners, a dominance that persists in practitioner surveys today even though the framework itself has not been formally revised since its creation (Martínez et al., 2021).

CRISP-DM's endurance rests on its simplicity and its explicit acknowledgement that data science work is rarely linear: the model's cyclical structure allows teams to move backward from, for instance, the modelling phase to data preparation as new issues are discovered (Flatiron School, 2026). At the same time, a critical review by Martínez et al. (2021) argues

that CRISP-DM's central limitation is its silence on team management: the framework specifies what technical activities should occur but says little about how roles, responsibilities, and communication among data scientists, engineers, and business stakeholders should be organised. Their comparative evaluation of nineteen published data science methodologies found that only a handful qualify as genuinely "integral," in the sense of addressing project management, team management, and data or information management simultaneously; most, including CRISP-DM itself, are strong on project structure but weak on team coordination or knowledge retention.

3.2 Contemporary and Alternative Methodologies

Several later frameworks were designed explicitly to remedy CRISP-DM's gaps. Microsoft's Team Data Science Process, for example, defines four distinct project roles and standardises project artefacts such as a living project charter and per-experiment model reports, improving reproducibility and stakeholder communication at the cost of a dependency on a specific vendor's tooling (Martínez et al., 2021). Other proposals, including agile-inspired data science lifecycles and IBM's Foundational Methodology for Data Science, extend or reorganise the CRISP-DM phases without resolving the underlying tension between the exploratory, trial-and-error nature of analytical work and the desire for predictable, auditable project timelines. A recurring finding across this literature is sobering: surveys cited by Martínez et al. (2021) report that a large majority of data science professionals do not follow any explicit methodology at all, even though most agree that doing so would improve project outcomes, and separate industry estimates suggest that the majority of data science projects never reach production deployment. This gap between methodological availability and methodological adoption is arguably as significant a challenge to the discipline's maturity as any purely technical limitation.

4. Applications Across Domains

4.1 Healthcare

Healthcare is among the most extensively studied application domains for data science and machine learning. Bhatt and Kumar (2024) survey the use of machine learning algorithms in health data science and find applications spanning disease diagnosis, extraction of relevant clinical features from electronic health records, and prediction of disease progression from heterogeneous patient data. Supervised learning techniques, particularly neural networks, have achieved notably high accuracy in diagnostic imaging tasks, while unsupervised methods are used to segment patient populations in ways that reveal clinically meaningful subgroups that would otherwise remain hidden in complex records. Reviews of the field also emphasise that value in healthcare analytics is not purely diagnostic: machine learning is increasingly applied to operational problems such as hospital resource optimisation and financial risk assessment, where predictive models support decisions about staffing, scheduling, and cost containment alongside clinical decision-making (ResearchGate, 2024).

These clinical gains are not unconditional. Several surveys note that the same properties that make healthcare data valuable, namely its sensitivity, heterogeneity, and regulatory protection, also make it difficult to work with responsibly. Data scarcity, privacy restrictions, and the risk that models trained on non-representative populations will underperform for under-served groups are recurring themes in this literature, motivating growing interest in privacy-preserving techniques such as federated learning, in which models are trained across decentralised datasets without the underlying patient records ever leaving their originating institution.

4.2 Finance

In financial services, data science techniques are used across the full spectrum of institutional activity, from consumer-facing credit decisions to systemic risk monitoring. Reviews of the field highlight fraud detection, algorithmic trading, and credit risk modelling as core use cases, with support vector machines and random forest algorithms particularly prominent in credit scoring applications, and deep learning architectures increasingly used to model time-series financial data (ResearchGate, 2024). Real-time or near-real-time processing is especially important in this sector: because financial markets and fraud patterns evolve quickly, the velocity dimension of big data is often as operationally important as its volume.

A growing strand of research also examines the intersection of finance and healthcare analytics, applying machine learning to model long-term financial sustainability and life expectancy trends within healthcare systems, which illustrates how the same underlying analytical toolkit is increasingly reused across what were once separately administered sectors.

4.3 Other Sectors and Cross-Cutting Applications

Beyond healthcare and finance, data science has become embedded in retail, where it supports dynamic pricing, personalised recommendation, and supply chain optimisation; in manufacturing, where predictive maintenance and failure prediction reduce downtime; and in transportation and logistics, where it underpins route optimisation, congestion modelling, and the perception systems used in autonomous vehicles (Martínez et al., 2021; ResearchGate, 2024). A common thread across these sectors is that the underlying algorithmic techniques, principally supervised and unsupervised machine learning, are broadly transferable, while the value of any given application depends heavily on domain-specific data quality, regulatory context, and organisational readiness to act on the resulting insights. This reinforces a point made repeatedly in the methodological literature: technical sophistication alone rarely determines project success; organisational fit and stakeholder alignment matter just as much.

5. Ethical, Legal, and Societal Challenges

5.1 Algorithmic Bias and Fairness

As data-driven systems have moved from experimental settings into decisions that materially affect people's lives, algorithmic bias has become the most widely discussed ethical concern in the field. A systematic analysis published in the International Journal of Current Science Research and Review documents cases in which hiring algorithms disproportionately disadvantaged particular demographic groups and facial recognition systems trained on skewed datasets exhibited markedly different accuracy rates across racial and gender groups (IJCSRR, 2025). The same review situates these failures within a broader ethical landscape that also includes transparency deficits and accountability gaps, and it surveys mitigation strategies such as fairness-aware machine learning, adversarial debiasing, and explainable artificial intelligence techniques designed to make model reasoning legible to affected stakeholders.

High-profile real-world incidents, including well-documented cases of biased automated recruitment tools and skewed image-classification systems, have been used across this literature to argue that bias is not a hypothetical risk confined to laboratory benchmarks but a demonstrated failure mode with concrete social consequences (IJCSRR, 2025). Existing ethical frameworks, including the ACM Code of Ethics and the IEEE's Ethically Aligned Design principles, provide normative guidance, but the literature is candid that translating high-level principles into specific, auditable engineering practice remains an unresolved challenge for most organisations.

5.2 Privacy, Governance, and Regulation

Privacy concerns run alongside, and often intersect with, concerns about bias. As organisations collect ever larger and more granular datasets, regulatory regimes such as the European Union's General Data Protection Regulation and the California Consumer Privacy Act have introduced formal obligations around data minimisation, consent, and the right to explanation, obligations that data science teams must now treat as engineering constraints rather than optional legal add-ons (IJCSRR, 2025). Techniques such as differential privacy and other privacy-preserving transformations are increasingly discussed not as academic curiosities but as practical tools for reconciling analytical value with individual privacy protection, particularly in sensitive domains such as healthcare and public-sector data. A recurring theme across this literature is the recommendation that organisations adopt structured governance mechanisms, including bias detection audits, data provenance tracking, and human review checkpoints, rather than relying on the good intentions of individual data scientists.

6. Workforce, Skills, and Emerging Trends

6.1 The Data Science Skills Gap

Even as demand for data-driven decision-making has grown, the supply of professionals equipped to meet that demand has consistently lagged. Priestley and McGrath (2019) describe this as a "talent gap": traditional university structures were, for a period, ill-equipped to train analytical talent at the pace industry required, prompting the rapid proliferation of dedicated data science degree programmes outside conventional academic departments. More recent workforce research indicates that this gap has not closed so much as changed shape. IBM (2026) reports that a large share of employers expect the core skills required of their workforce to change substantially by the end of the decade, and that organisational leaders increasingly expect formal reskilling and upskilling to become a permanent feature of workforce planning rather than a one-off response to a single technology shift.

Industry surveys converge on a related point: many organisations have adopted data and artificial intelligence tools faster than they have built the internal expertise to use them effectively. Gloat (2026) reports that a large majority of organisations now use artificial intelligence in at least one business function, yet only a small fraction consider themselves to have achieved full analytical or artificial intelligence maturity, a mismatch that the same source attributes less to a shortage of tools than to a shortage of people who can deploy those tools with judgement.

6.2 Generative AI, Automation, and the Discipline's Trajectory

The emergence of large-scale generative artificial intelligence has introduced a further inflection point. The World Economic Forum (2025) notes that nearly half of surveyed business leaders still anticipate significant skills gaps in critical technical roles even as automation reduces demand for some traditional data-handling tasks, creating what has been described as a dual challenge of simultaneous workforce redundancy in some functions and acute talent scarcity in others. Analyses of job-posting data show a sharp rise in explicit demand for generative artificial intelligence skills across both technical and non-technical roles, suggesting that the boundary between "data scientist" as a specialised job title and broader data literacy as a general workplace competency is becoming increasingly blurred.

Within data science specifically, generative and foundation models are being explored as tools for synthetic data generation, capable of producing realistic but privacy-preserving training data in domains such as medical imaging where real data is scarce or sensitive, as well as for automating routine elements of the data preparation and exploratory analysis phases historically performed manually by analysts. Industry forecasts, including projections that a meaningful share of routine analytical decisions will be made autonomously through agentic artificial intelligence systems within the next several years (USDSI, n.d.), suggest that the data scientist's role is shifting away from manual model-building toward the supervision, validation, and ethical governance of increasingly automated analytical pipelines. This shift reinforces, rather than undermines, the importance of the methodological and ethical foundations discussed earlier in this review: as more of the technical pipeline is automated, the judgement required to frame the right business problem, validate model outputs, and guard against bias becomes, if anything, more consequential.

7. Discussion: Open Problems and Research Directions

Drawing the preceding sections together, three cross-cutting observations emerge. First, the gap between technical capability and reliable deployment remains the discipline's most persistent weakness. The methodological literature reviewed in Section 3 shows that most data science projects still proceed without an explicit process methodology, and that even well-established frameworks such as CRISP-DM address technical workflow far more thoroughly than team coordination or organisational knowledge retention. Future work would benefit from empirically testing integrated methodologies, of the kind proposed by Martínez et al. (2021), across a wider range of industries and project scales, rather than continuing to propose new frameworks without comparative evaluation against existing ones.

Second, the ethical challenges surveyed in Section 5, particularly algorithmic bias and privacy, are not peripheral concerns to be addressed after a model is built; they are, in the framing increasingly adopted by both academic and industry sources, engineering requirements that should shape data collection, model selection, and validation from the outset. The continuing gap between high-level ethical principles and concrete, auditable practice suggests that future

research should prioritise practical tooling, such as standardised fairness audits and provenance-tracking infrastructure, over further elaboration of abstract ethical frameworks.

Third, the workforce and technological trends discussed in Section 6 indicate that the skills profile of a data scientist is shifting away from manual coding and model-building toward problem framing, model validation, and governance oversight, as generative and agentic artificial intelligence systems absorb an increasing share of routine analytical work. This has direct implications for data science education and professional training, which will need to place correspondingly greater weight on critical evaluation, domain expertise, and ethical reasoning rather than treating these as secondary to technical proficiency.

8. Conclusion

Data science has evolved from a loosely defined extension of statistics into a distinct, application-driven discipline that now shapes decisions across nearly every major economic sector. This review has traced that evolution from its statistical and computational origins through the methodological frameworks that structure applied projects, the domains in which the discipline delivers demonstrable value, and the ethical and workforce challenges that accompany its continued growth. The evidence surveyed here supports a central conclusion: the discipline's future development will be determined less by further gains in raw algorithmic performance and more by its capacity to institutionalise trustworthy process, rigorous governance, and a workforce equipped to exercise judgement over increasingly automated analytical systems. Organisations and researchers that treat data science as a purely technical pursuit risk repeating the well-documented pattern of technically sound projects that fail to reach production or to earn stakeholder trust; those that treat it as an integrated technical, organisational, and ethical practice are better positioned to realise its long-term value.

References

- Bhatt, V., & Kumar, M. (2024). Applications of machine learning algorithms in health data science (HDS) for next research directions: A survey report. *Journal of Advances in Information Research and Applications (JoAIRA)*. <https://journals.stmjournals.com/joaira/article=2024/view=131474/>
- Capco. (2026). *Workforce of the future in the age of AI*. Capco Intelligence. <https://www.capco.com/intelligence/capco-intelligence/workforce-of-the-future-in-the-age-of-ai>
- Data Science PM. (2024). *What is CRISP-DM?* <https://www.datascience-pm.com/crisp-dm-2/>
- Dataversity. (2025). *A brief history of data science*. <https://www.dataversity.net/articles/brief-history-data-science/>
- Flatiron School. (2026). *Intro to data science: Understanding CRISP-DM*. <https://flatironschool.com/blog/intro-to-data-science-understanding-crisp-dm>
- Gloat. (2026). *AI skills demand in the U.S. job market (2026)*. <https://gloat.com/blog/ai-skills-demand/>
- IBM. (2026). *AI skills gap*. <https://www.ibm.com/think/insights/ai-skills-gap>
- International Journal of Current Science Research and Review [IJCSRR]. (2025). Ethical challenges in data science: Navigating the complex landscape of responsibility and fairness. *International Journal of Current Science Research and Review*, 8. <https://ijcsrr.org/wp-content/uploads/2025/02/09-0703-2025.pdf>
- KDnuggets. (2024). *Data science lifecycle process*. <https://www.kdnuggets.com/data-science-lifecycle-process>
- Martínez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research* (preprint). arXiv:2106.07287. <https://arxiv.org/abs/2106.07287>
- Nexla. (2025). *Big data* [Glossary entry]. <https://nexla.com/glossary/big-data/>
- Press, G. (2013, May 28). A very short history of data science. *Forbes*. <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/>

Priestley, J., & McGrath, R. J. (2019). The evolution of data science: A new mode of knowledge production. *International Journal of Knowledge Management*, 15(2), 1–19. <https://doi.org/10.4018/IJKM.2019040106>

ResearchGate. (2024). *Applications of machine learning in healthcare, finance, agriculture, retail, manufacturing, energy, and transportation: A review*. <https://www.researchgate.net/publication/385207893>

TechTarget. (n.d.). *What are the 3 V's of big data?* <https://www.techtarget.com/whatis/definition/3Vs>

United States Data Science Institute [USDSI]. (n.d.). *Next era of data science: Skills, trends, and opportunities*. <https://www.usdsi.org/data-science-insights/next-era-of-data-science-skills-trends-and-opportunities>

World Economic Forum. (2025, October). *How we can balance AI overcapacity and talent shortages*. <https://www.weforum.org/stories/2025/10/ai-s-new-dual-workforce-challenge-balancing-overcapacity-and-talent-shortages/>