COVENANT AEGIS REVIEW OF ADVANCED COMPUTING VOL. 4 · NO. 2 · 2026 · PP. 4–22

Operationalising Data Science: Reproducible Pipelines, MLOps and Reliable Decisions

P. Vasquez[1], F. Adeyemi[2], H. Said[1], E. Rossi[3]

[1]*Covenant Aegis Technology Division, Data Engineering Laboratory*

[2]*Department of Statistics, Partner Research Network*

[3]*Center for Reproducible Computational Science*

Received 15 Jan 2026 · Revised 2 Mar 2026 · Accepted 11 Mar 2026 · REVIEW ARTICLE · DOI 10.48-CA/advcomp.2026.0421 ·
Corresponding author: p.vasquez@covenantaegis.com

ABSTRACT

Background. Data science delivers value only when models reach reliable, monitored production, yet most projects fail on process and data rather than algorithms. **Objective.** To review the data-science lifecycle, the principal failure modes—leakage, distribution shift, technical debt and irreproducibility—and the MLOps practices that mitigate them. **Methods.** Narrative synthesis of process models and the large-scale data/ML-systems literature. **Results.** Effort is dominated by data preparation; leakage is a pervasive cause of irreproducible ML claims across scientific fields; and ML systems accrue hidden technical debt that raises long-term cost. Disciplined versioning, automated pipelines, continuous validation and monitoring are what separate durable analytics from fragile prototypes. **Conclusion.** Reproducibility, data quality and operational engineering—not model novelty—determine impact; MLOps turns prototypes into reliable decision systems.

Keywords: data science; MLOps; reproducibility; data leakage; distribution shift; technical debt; data engineering

1. Introduction

Organisations increasingly base decisions on statistical and machine-learning models, yet a large share of projects never reach production or degrade once deployed. The gap is rarely algorithmic; it is one of process, data quality and engineering [1].

2. The Lifecycle

Process models such as CRISP-DM frame the lifecycle as business understanding, data understanding, preparation, modelling, evaluation and deployment, iterated as understanding improves [2]. Most effort is spent on data preparation (Figure 1); rigorous problem framing and honest evaluation against business-relevant metrics determine whether modelling pays off.

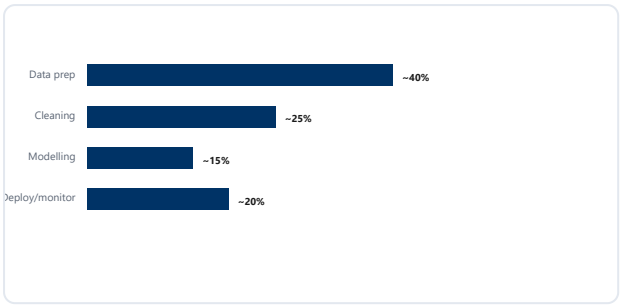


Figure 1. Approximate distribution of practitioner effort across the lifecycle; data preparation and cleaning dominate, consistent with repeated practitioner surveys [1,2]. Illustrative proportions.

3. Scalable Data Processing

Modern analytics rests on distributed data systems. The MapReduce model made large-scale batch processing tractable [3], while in-memory engines such as Apache Spark accelerated iterative and interactive workloads through resilient distributed datasets [4]. Reliable feature pipelines, lineage and data-quality controls underpin everything downstream.

4. Common Failure Modes

4.1 Data Leakage

Leakage—information from outside the training context contaminating the model—produces optimistic estimates that collapse in production and is a pervasive cause of irreproducible ML claims across fields [5].

4.2 Distribution Shift and Technical Debt

Deployed models degrade as data drift from training conditions, requiring monitoring and retraining. Sculley et al. catalogued the hidden technical debt of ML systems—entanglement, undeclared consumers, pipeline jungles and configuration sprawl—that accrues silently and raises long-term cost [6].

Table 1. Failure modes, symptom and primary mitigation.

Failure	Symptom	Mitigation
Leakage	Great test, poor prod	Strict temporal splits
Distribution shift	Slow accuracy decay	Drift monitoring, retrain
Technical debt	Fragile pipelines	Modularity, contracts
Irreproducibility	Cannot rebuild result	Version data/code/env

5. MLOps

MLOps applies software-engineering rigour to the ML lifecycle: version control for data, code and models; automated, reproducible pipelines; continuous integration, delivery and training; and production monitoring of data quality, drift and performance [7]. Reproducibility—reconstructing a result from recorded data, code and environment—is the foundation of trust and iteration [8].

6. The Covenant Aegis Methodology

We standardise experiment tracking, dataset and model versioning, automated validation gates, and live monitoring with drift and performance alarms. Every model carries documentation of its data, assumptions and intended use, and promotion to production requires passing predefined quality and fairness checks.

7. Discussion and Conclusion

The decisive factors in data science are reproducibility, data quality and operational engineering, not model novelty. The figures and table here are illustrative syntheses of practitioner experience rather than measured benchmarks. MLOps discipline turns analytics from one-off prototypes into reliable, monitored decision systems.

References

1. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. *Harv Bus Rev*. 2012;90(10):70–76.
2. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. *PAKDD*. 2000.
3. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM*. 2008;51(1):107–113.
4. Zaharia M, et al. Apache Spark: a unified engine for big data processing. *Commun ACM*. 2016;59(11):56–65.
5. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804.
6. Sculley D, et al. Hidden technical debt in machine learning systems. *NeurIPS*. 2015.
7. Kreuzberger D, Kühl N, Hirschl S. Machine learning operations (MLOps): overview, definition, and architecture. *IEEE Access*. 2023;11:31866–31879.
8. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. National Academies Press; 2019.

9. Data Quality and Feature Engineering

Model performance is bounded by data quality. Common defects—missingness, inconsistent units, duplicate records, label noise and selection bias—propagate into models and evaluation. Disciplined practice enforces data contracts, schema validation and automated quality checks at ingestion, and treats feature engineering as versioned, tested code. Feature stores provide consistent feature definitions across training and serving, eliminating training-serving skew, a frequent cause of silent production failure [A1].

Table A1. Data-quality dimensions and representative checks.

Dimension	Check	Risk if ignored
Completeness	Null/coverage rates	Biased estimates
Consistency	Schema, units	Silent errors
Validity	Range, domain	Outlier leakage
Timeliness	Freshness SLA	Stale decisions

10. Experimentation and Causal Caution

Predictive correlation does not establish causation; acting on a model can change the system it was trained on. Where decisions intervene, A/B testing and, where feasible, causal inference are required to estimate true effect [A2]. Simpson's paradox, confounding and feedback loops routinely mislead naive analyses. Randomised experimentation remains the gold standard for measuring the impact of a deployed model on business or clinical outcomes.

11. The Reproducibility Crisis in ML

A growing body of work documents non-reproducible machine-learning results across scientific fields, with data leakage the most common cause [A3]. Leakage takes many forms—temporal leakage, duplicate records spanning train and test, target leakage from features computed using the outcome—and consistently

inflates reported performance. Reproducibility requires versioning data, code and environment, fixing seeds, and recording the full pipeline. Figure A1 shows the characteristic gap between leaked and corrected evaluation.

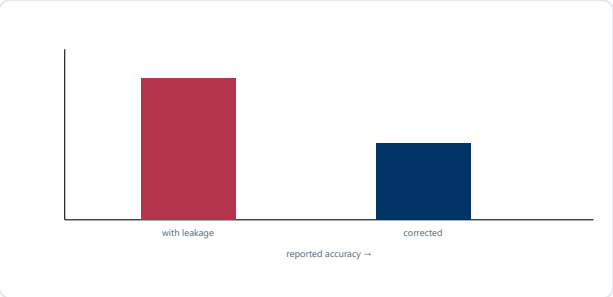


Figure A1. Schematic: leakage inflates reported accuracy; corrected evaluation reveals the true, lower performance [A3]. Conceptual.

12. MLOps Architecture

MLOps applies CI/CD/CT (continuous integration, delivery and training) to the ML lifecycle, with versioned data and models, automated pipelines, a model registry and staged promotion [A4]. Levels of MLOps maturity range from manual scripts to fully automated, monitored pipelines. The architecture must support rollback, reproducible rebuilds and lineage from a prediction back to the data and code that produced it.

13. Monitoring and Drift

Deployed models degrade through covariate shift, label shift and concept drift. Monitoring tracks input distributions (e.g., population-stability index), prediction distributions and, when labels arrive, live performance. Statistical tests and control charts trigger alerts; predefined policies govern recalibration and retraining. Without monitoring, degradation is silent until it causes business or safety harm [A5].

14. Governance, Documentation and Ethics

Responsible data science documents datasets (datasheets) and models (model cards), records intended use and known limitations, and assesses fairness and privacy [A6,A7]. Governance frameworks such as the NIST AI Risk Management Framework structure these obligations across the lifecycle [A8]. Documentation and auditability are prerequisites for trust, regulatory compliance and safe iteration.

15. Future Directions

Priorities include automated data-quality and leakage detection, standardised lineage and reproducibility tooling, causal and uncertainty-aware decision systems, and governance integrated into pipelines by default. Durable impact comes from engineering discipline and reproducibility, not from chasing marginal model-accuracy gains.

Extended References

9. Sculley D, et al. Hidden technical debt in machine learning systems. *NeurIPS*. 2015.
10. Kohavi R, Tang D, Xu Y. *Trustworthy Online Controlled Experiments*. Cambridge University Press; 2020.
11. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in ML-based science. *Patterns*. 2023;4(9):100804.

12. Kreuzberger D, Kühl N, Hirschl S. MLOps: overview, definition, and architecture. *IEEE Access*. 2023;11:31866–31879.
13. Klaise J, et al. Monitoring and explainability of models in production. *arXiv*:2007.06299. 2020.
14. Gebru T, et al. Datasheets for datasets. *Commun ACM*. 2021;64(12):86–92.
15. Mitchell M, et al. Model cards for model reporting. *FAccT*. 2019:220–229.
16. NIST. AI Risk Management Framework (AI RMF 1.0). 2023.

Author contributions: PV led the review; FA contributed statistical rigour; HS and ER reviewed systems and reproducibility. All authors revised. **Funding:** Covenant Aegis Technology Division. **Competing interests:** Authors are affiliated with Covenant Aegis.

© 2026 Covenant Aegis Technology Division. This review synthesises publicly available peer-reviewed literature; figures are illustrative. For professional and educational use.