Applied Science, Engineering and Management Bulletin [ASEMB]
ISSN: 3049-3005
Vol 3 Issue 4 (Oct-Dec 2026) |Pg: 81-89

Big Data Analytics for Predictive Decision-Making in Large-Scale Information Systems

Dr Anupama Jha
Assistant Professor, VSIT, VIPS-TC, GGSIP University, Delhi
anupama.jha@vips.edu

Abstract

Large-scale information systems generate heterogeneous, high-velocity data that can support decisions extending beyond descriptive reporting toward forecasts of future events and risks. Yet the conversion of data abundance into dependable decisions is not automatic: it depends on data integration, measurement quality, scalable computation, predictive validation, organizational interpretation, and governance. This paper examines how big data analytics can enable predictive decision-making in large-scale information systems through a structured conceptual review of established research and institutional frameworks. It synthesizes literature on business intelligence and analytics, predictive modelling in information systems, big data architectures, machine learning operations, and responsible data use. The paper adopts a secondary-research design; it does not report original survey, interview, or experimental data. A synthesis framework is developed around five connected capabilities: data foundation, analytical modelling, decision translation, operational feedback, and governance. The review distinguishes prediction from explanation and prescription, and argues that predictive performance should be evaluated using temporally valid tests, calibration, operational utility, and subgroup-level diagnostics rather than accuracy alone. It further identifies recurring barriers, including fragmented data, concept drift, opaque models, weak accountability, and the organizational gap between model output and action. The paper proposes an integrated architecture and implementation lifecycle for embedding predictions into decision workflows while retaining human oversight. The central conclusion is that sustainable value arises not from model sophistication in isolation but from alignment among technical scalability, decision context, institutional capability, and trustworthy governance.

Keywords: *big data analytics; predictive analytics; decision-making; large-scale information systems; machine learning; data governance; decision intelligence*

1. Introduction

Organizations increasingly coordinate operations through interconnected information systems that capture transactions, machine-generated events, user interactions, geospatial observations, documents, and external data feeds. Enterprise resource planning platforms, customer relationship management systems, digital payment networks, industrial internet-of-things environments, health information systems, and public-sector platforms can produce datasets whose scale and heterogeneity challenge conventional database and reporting practices. Big data is therefore not simply a synonym for a large database. Its practical significance arises from the combination of volume, velocity, variety, and associated concerns such as veracity and value (Chang & Grady, 2019; Tosi et al., 2024).

The analytical opportunity is to use these data to anticipate events before they occur. Predictive analytics applies statistical learning, data mining, and machine learning to estimate unknown outcomes from observed patterns. Examples include forecasting equipment failure, estimating customer churn, detecting payment fraud, anticipating demand, predicting service congestion, and identifying patients who may require additional care. Such estimates can change the timing and allocation of organizational resources. In contrast with retrospective dashboards, predictive systems can support early intervention, scenario planning, and risk-sensitive prioritization.

However, a forecast is not itself a decision. A model may estimate a high probability of default, failure, or churn, but an organization must still determine whether intervention is justified, which action is feasible, what its costs are, and who is accountable for the outcome. Shmueli and Koppius (2011) emphasize that explanatory statistical modelling and predictive modelling have different objectives and evaluation logics. An explanatory model may illuminate relationships without being the best forecasting instrument; conversely, a highly

predictive model may not establish causality. Confusing these aims can produce analytically impressive but operationally weak systems.

This distinction is especially important in large-scale information systems. The system must move data from distributed sources into reliable analytical environments, execute models within acceptable latency and cost, deliver results to decision-makers, and learn from subsequent outcomes. The technical architecture must therefore be connected to organizational processes and governance. Chen et al. (2012) describe the evolution of business intelligence and analytics from structured enterprise reporting toward web and mobile data and more advanced analytics. The progression illustrates that analytical capability depends on changes in both data and organizational practice.

This paper investigates the relationship between big data analytics and predictive decision-making in large-scale information systems. It develops an integrated conceptual account of the technical, methodological, and managerial conditions required for effective use. Its contribution is a synthesis that links architecture, predictive validity, decision utility, feedback, and responsible governance. The paper is structured as follows: literature is reviewed, objectives and gaps are stated, the secondary research methodology is explained, and a thematic analysis is followed by findings and implications.

2. Review of Literature

2.1 Big data and the evolution of analytics

Big data research has developed around the challenge of extracting value from data that exceed the capabilities of traditional processing approaches. The NIST Big Data Interoperability Framework treats big data as an architectural and interoperability problem as much as a data-volume problem; it emphasizes roles, interfaces, management fabrics, and security and privacy considerations (Chang & Grady, 2019). This perspective is useful because it discourages a narrow focus on storage capacity. A large-scale information system must coordinate ingestion, processing, storage, access, and governance across heterogeneous components.

Chen et al. (2012) position business intelligence and analytics as an evolving field, distinguishing traditional BI&A based on structured data and relational systems from approaches that incorporate web data and more advanced analytics. Their framework provides a historical account of how data sources and methods broadened. Gandomi and Haider (2015) similarly examine big data concepts, technologies, and analytics methods, emphasizing that value extraction requires methods suited to the characteristics of the data. Together, these works suggest that technological scale matters only when analytical methods and business questions are aligned.

More recent reviews show that the field remains methodologically diverse. Jamarani et al. (2024), in a systematic review of 109 studies published between 2014 and 2023, map the use of predictive analytics across application domains and methods. Tosi et al. (2024) review fifteen years of big data research and identify persistent challenges alongside emerging directions. These reviews indicate substantial progress in application and infrastructure, but also show that implementation difficulties—data quality, integration, scalability, and interpretability—remain consequential.

2.2 Predictive analytics and information systems

Predictive analytics estimates the likelihood or magnitude of future or unknown outcomes. Typical approaches include regression, classification, time-series forecasting, clustering for segmentation, anomaly detection, ensemble learning, and deep learning. The choice of method depends on the prediction target, data structure, decision horizon, cost of error, and operational constraints. Hastie et al. (2009) provide a foundational treatment of statistical learning methods and the trade-offs among model flexibility, interpretability, and generalization.

A key contribution to information systems research is Shmueli and Koppius (2011), who argue that predictive analytics has been underused relative to explanatory statistical modelling. They identify roles for prediction in theory generation, measurement development, comparison of theories, model improvement, relevance assessment, and evaluation of predictability. Their central methodological warning is that explanatory power does not automatically imply predictive power. Models should be tested on data not used for fitting, and their performance should be assessed in relation to the intended prediction task.

Provost and Fawcett (2013) make a related point from the perspective of data science: analytical value depends on understanding the business problem, the costs and benefits of decisions, and the operational setting in which predictions will be used. A model that improves a generic metric may fail to improve organizational outcomes if the decision threshold is poorly chosen or if the intervention is ineffective. Predictive analytics is therefore best understood as a socio-technical capability rather than a standalone algorithm.

2.3 Architectures and scalable processing

Large-scale analytics typically combines distributed storage, parallel processing, streaming or batch ingestion, data integration, and analytical services. Distributed architectures make it possible to process workloads that would be impractical on a single machine, but they introduce complexity in orchestration, consistency, security, metadata management, and cost control. NIST's reference-architecture work emphasizes interoperability and cross-cutting management and security functions rather than prescribing a single vendor stack (Chang & Grady, 2019).

A data lake or lakehouse can retain raw and semi-structured data for flexible analysis, while curated warehouses and feature stores support consistent reporting and model inputs. Batch processing is appropriate for periodic planning and retrospective analysis; stream processing is useful when event latency affects the value of a decision, such as fraud detection or industrial monitoring. A hybrid architecture is often preferable because not every decision requires real-time computation. The system should allocate latency, reliability, and cost according to the decision's time sensitivity.

Architecture also shapes analytical reproducibility. Data lineage, schema versioning, access controls, feature definitions, and model registries help ensure that a prediction can be traced to the data and code that generated it. Sculley et al. (2015) describe how machine-learning systems accumulate hidden technical debt through data dependencies, feedback loops, configuration complexity, and changing external conditions. Their analysis implies that deployment and maintenance are core parts of model quality, not secondary engineering concerns.

2.4 From prediction to decision support

Decision-making requires more than ranking cases by estimated risk. A prediction becomes actionable when it is connected to a policy, resource constraint, and outcome objective. For example, a churn score has limited value unless the organization knows which retention actions are available, their costs, and whether different customer groups respond to them. A predictive maintenance score must be linked to maintenance windows, spare-parts availability, safety requirements, and the expected cost of downtime. Prescriptive analytics extends prediction by evaluating actions under constraints; recent reviews note that scalable prescriptive analytics remains less mature than predictive methods (Wissuchek & Zschech, 2025).

This transition can be framed through expected utility. If a model estimates an event probability, the decision-maker compares the expected loss of acting and not acting, accounting for intervention cost and consequences. Thresholds should therefore reflect the decision context rather than defaulting to 0.5. Calibration is important because a predicted probability should correspond to observed event frequency over suitable groups and time periods. Ranking metrics such as area under the ROC curve may be useful, but they do not by themselves determine whether a decision policy creates value.

Human oversight is particularly important where decisions affect access to credit, employment, health services, public benefits, or other high-impact outcomes. Models can reproduce historical patterns that reflect unequal access, measurement bias, or past institutional decisions. Barocas and Selbst (2016) explain how data mining can produce disparate impacts even when protected attributes are not explicitly used. Responsible predictive systems consequently require assessment of data provenance, proxy variables, error distributions, recourse, transparency, and appeal mechanisms.

2.5 Synthesis of prior research

The principal perspectives and their relevance to the present study are compared in Table 1.

Table 1. Synthesis of major literature streams

Research stream	Representative contribution	Implication for this paper
BI&A evolution	Chen et al. (2012) map the evolution of business intelligence and analytics.	Analytics capability evolves with data sources, methods, and organizational use.
Big data concepts	Gandomi and Haider (2015); Chang and Grady (2019).	Scale requires architecture, interoperability, and governance.
Predictive validity	Shmueli and Koppius (2011); Hastie et al. (2009).	Prediction must be evaluated out of sample and separately from causal explanation.
Operational data science	Provost and Fawcett (2013); Sculley et al. (2015).	Business context, deployment, and technical debt affect realized value.
Recent synthesis	Jamarani et al. (2024); Tosi et al. (2024); Wissuchek and Zschech (2025).	Applications are expanding, while scalable integration and prescriptive translation remain challenges.

3. Research Methodology

3.1 Research design and approach

The study uses a qualitative, conceptual literature-review design based on secondary sources. It synthesizes peer-reviewed journal articles, academic books, and institutional publications relevant to big data analytics, predictive modelling, information systems, and responsible data use. A conceptual review is appropriate because the research question concerns relationships among technical architecture, analytical methods, and decision processes rather than the measurement of a single intervention in a particular organization.

3.2 Data sources and selection

The evidence base was developed from established academic publications and institutional sources identifiable through bibliographic records and publisher or institutional pages. Priority was given to foundational works that define the field, methodological publications on prediction, peer-reviewed reviews published in recent years, and authoritative standards-oriented material. Sources were selected for relevance to at least one of the following themes: big data characteristics and architecture; business intelligence and analytics; predictive model design and evaluation; deployment and monitoring; decision utility; and governance, privacy, or fairness.

The paper uses secondary data only. No primary data were collected. There were no surveys, interviews, experiments, organizational records, or participant observations. Consequently, no sample of human respondents, response rate, inferential statistics, or empirical effect size is reported. This distinction is important because the findings are interpretive propositions derived from literature synthesis, not newly observed outcomes.

3.3 Analytical procedure

The literature was organized thematically. First, recurring concepts were extracted: data characteristics, architecture, modelling, validation, decision translation, operational monitoring, and governance. Second, contributions were compared to identify areas of convergence and tension—for example, the distinction between explanatory and predictive objectives, and the difference between model performance and decision value. Third, the themes were arranged into a lifecycle framework that follows data from source systems through model deployment and decision feedback. Finally, implications were derived for system design and future empirical research.

4. Data Analysis and Discussion

4.1 An integrated architecture for predictive decision-making

The synthesis indicates that a predictive information system should be designed as a connected pipeline rather than a sequence of isolated tools. Data sources include internal transactional systems, sensors, logs, external datasets, and user-generated content. Ingestion and integration services standardize formats, resolve identities, manage event time, and attach metadata. A governed storage and processing layer supports both historical

analysis and low-latency event processing. Curated features then feed model training and inference services. Predictions are exposed through operational applications, dashboards, alerts, or automated workflows, while outcomes and human overrides return to the analytical environment.

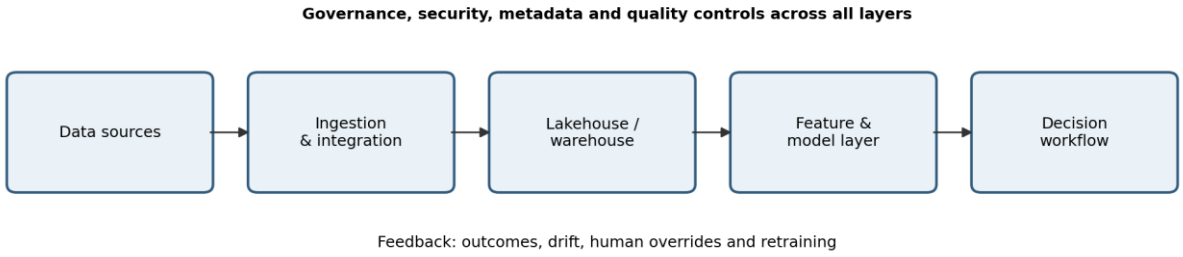

Figure 1. Reference architecture for big data-enabled predictive decision-making

Figure 1 illustrates the integrated architecture proposed in this paper. It is an author-developed conceptual diagram informed by the NIST interoperability framework and literature on BI&A and machine-learning systems (Chang & Grady, 2019; Chen et al., 2012; Sculley et al., 2015). Governance is represented as a cross-cutting capability because access control, privacy, data quality, metadata, and auditability cannot be confined to a final approval step. The feedback path is equally important: outcomes, overrides, and drift signals should inform monitoring and retraining.

The architecture should not imply that every workload belongs in one centralized platform. Federated or domain-oriented architectures may be more suitable where data residency, ownership, latency, or institutional boundaries constrain centralization. What matters is that interfaces, definitions, and controls are sufficiently consistent to support trustworthy reuse. NIST's emphasis on interoperability is relevant because large-scale systems frequently span multiple technologies and organizational owners (Chang & Grady, 2019).

4.2 Data quality and feature readiness

The predictive value of a dataset depends on whether its variables represent the target phenomenon at the time a decision is made. Missing values, duplicates, inconsistent identifiers, delayed records, selection bias, and changing definitions can distort model inputs. Data volume does not compensate for systematic measurement error. For instance, a model predicting service demand from incomplete transaction records may understate demand in channels with weaker digital capture. Similarly, a maintenance model trained on sensor readings collected only from functioning equipment may fail to learn patterns preceding catastrophic failures.

Feature engineering should therefore be governed as a reproducible process. Definitions need owners, business meaning, permitted use, refresh frequency, and lineage. Leakage must be prevented: a feature that contains information recorded after the prediction time can create artificially strong test performance while being unavailable in actual deployment. Temporal systems should preserve event time and processing time, since late-arriving events can change the interpretation of a prediction. These practices operationalize the veracity and validity concerns identified in big data frameworks (Chang & Grady, 2019).

4.3 Model selection and predictive evaluation

A further distinction is between discrimination and calibration. Discrimination measures whether higher-risk cases tend to receive higher scores; calibration assesses whether predicted probabilities correspond to observed frequencies. Decision-makers need both when scores are interpreted as probabilities or used in expected-cost calculations. Shmueli and Koppius (2011) caution against assuming that explanatory fit demonstrates predictive performance. Accordingly, validation should test out-of-sample performance and compare models against meaningful baselines.

4.4 Decision translation and organizational use

The decision threshold should be selected in relation to the costs and benefits of available actions. Where interventions are limited, ranking may be more useful than binary classification: a team may act on the highest-

risk cases until its capacity is exhausted. Where the consequences of errors are asymmetric, threshold selection should reflect those consequences. Where interventions change outcomes, historical labels may not represent what would happen under a new policy. This creates a need to distinguish prediction from causal effect estimation and, where feasible, to evaluate interventions using appropriate experimental or quasi-experimental designs.

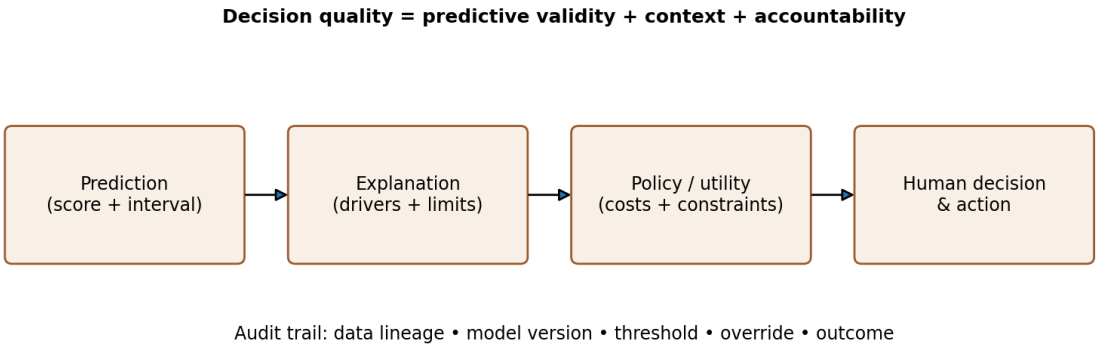

Figure 2. Translating model output into an accountable decision

Figure 2 presents the decision-translation logic developed in this review. The diagram is an author-developed synthesis drawing on the distinction between predictive and explanatory modelling (Shmueli & Koppius, 2011), the business-problem orientation of data science (Provost & Fawcett, 2013), and research on algorithmic disparate impact (Barocas & Selbst, 2016). An estimate becomes operationally meaningful only when accompanied by an explanation of its limits, a policy or utility rule, and an accountable decision process.

4.5 Operational monitoring and model lifecycle

A mature lifecycle includes versioned datasets, feature definitions, code, model artifacts, thresholds, and deployment configurations. It also includes rollback procedures, incident response, documentation, and retraining criteria. Retraining should not be automatic merely because a statistical monitor signals change: teams must determine whether the change is meaningful, whether labels are reliable, and whether the new model improves the relevant decision objective. Sculley et al. (2015) show why unmanaged dependencies and feedback loops can create hidden technical debt. The implication is that operational ownership must be assigned before launch.

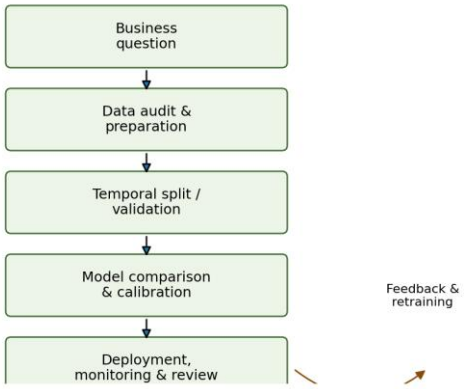

Figure 3. Predictive analytics lifecycle with validation and feedback

Figure 3 depicts the iterative lifecycle proposed in this paper. It is an author-developed process model informed by statistical learning principles (Hastie et al., 2009), predictive validation in information systems (Shmueli &

Koppius, 2011), and the operational risks of deployed machine-learning systems (Sculley et al., 2015). Its key feature is that monitoring and review are integral to the lifecycle, not an optional final phase.

4.6 Governance, privacy, fairness, and explainability

Fairness cannot be reduced to one universal metric. Different parity criteria can conflict when base rates differ, and the appropriate assessment depends on the decision, legal context, and harm model. Barocas and Selbst (2016) demonstrate that seemingly neutral data mining can create disparate impacts through proxy variables and historical patterns. Accordingly, fairness review should consider data collection, target definitions, label quality, error rates, access to recourse, and the consequences of interventions. Model explanations can support scrutiny, but explanations should not be presented as causal proof or as a substitute for contestability.

4.7 Comparative implications across application domains

Table 2. Illustrative predictive decision contexts

Application	Prediction target	Decision supported	Critical design concern
Financial services	Fraud, default, transaction anomalies	Investigate, authorize, price, or manage exposure	Rare events, asymmetric error costs, fairness and explainability
Manufacturing	Equipment failure, yield, quality deviation	Schedule maintenance, adjust process, allocate spares	Sensor reliability, temporal drift, safety and downtime costs
Healthcare	Deterioration risk, readmission, demand	Prioritize clinical review and resource planning	Clinical validation, representativeness, privacy and human oversight
Retail and digital services	Demand, churn, recommendation response	Stock, retain customers, personalize services	Feedback loops, changing preferences, consent and intervention effects
Public administration	Service demand, fraud risk, case duration	Allocate inspection or service capacity	Due process, transparency, data minimization and contestability

4.8 Proposed implementation framework

The integrated framework can be operationalized through six stages. First, define the decision and its value proposition before selecting data or algorithms. Second, assess data availability, provenance, quality, representativeness, and permitted use. Third, establish a scalable architecture with clear interfaces, metadata, security, and reproducibility. Fourth, build and validate candidate models using baselines, temporal or entity-aware splits, calibration, and subgroup diagnostics. Fifth, embed predictions in a workflow with decision thresholds, escalation rules, human review, and an audit trail. Sixth, monitor outcomes and system health, investigate drift, and revise the model or decision policy when evidence warrants.

Table 3. Implementation stages for predictive decision-making

Stage	Core questions	Expected control or output
1. Decision framing	What action will a prediction change? What are error costs?	Decision specification and success criteria
2. Data readiness	Are data valid, representative, timely, lawful, and traceable?	Data-quality profile, lineage, access approval
3. Platform design	Can the system process required scale and latency reliably?	Governed data pipeline and reproducible feature layer
4. Model validation	Does the model generalize and calibrate under realistic testing?	Validation report, baseline comparison, subgroup analysis
5. Workflow integration	Who acts, at what threshold, and with what oversight?	Decision policy, user interface, escalation and audit logs
6. Monitoring	How are drift, errors, incidents, and outcomes detected?	Monitoring dashboard, review cadence, retraining/rollback rules

A practical implementation should begin with a bounded use case where the decision is frequent enough to evaluate, the outcome can be observed, and the intervention is feasible. A pilot should compare the proposed system with the existing decision process, not merely with a weak statistical baseline. Where possible, controlled rollout or phased evaluation can estimate whether the analytics intervention improves outcomes. Documentation should record assumptions, limitations, intended use, prohibited uses, and the responsible owner.

5. Conclusion

Big data analytics can strengthen predictive decision-making in large-scale information systems by enabling organizations to identify patterns, estimate future events, and intervene earlier. Nevertheless, data scale and algorithmic sophistication are not sufficient conditions for value. Reliable outcomes depend on the full chain from data generation and integration to model validation, decision translation, operational feedback, and governance. This paper has synthesized relevant research and proposed an integrated framework that treats these components as mutually dependent.

Three conclusions are particularly important. First, prediction must be evaluated as prediction: models should be tested on unseen and temporally appropriate data, compared with relevant baselines, and assessed for calibration and subgroup performance. Second, a predictive score should be linked to a defined decision and an explicit account of costs, constraints, and responsibility. Third, systems must be monitored after deployment because drift, feedback loops, technical debt, and changes in organizational context can undermine initial performance.

For practitioners, the proposed framework suggests beginning with a decision problem rather than a technology purchase, establishing data governance early, and assigning operational ownership before deployment. For researchers, the framework points to empirical questions about the relative contribution of data quality, architecture, model design, human oversight, and governance to decision outcomes. Future studies could compare implementation approaches across sectors, examine the relationship between calibration and realized utility, and evaluate how monitoring and recourse mechanisms affect trust and error correction.

The study is limited by its conceptual, secondary-source design and selective literature coverage. It does not estimate the causal impact of big data analytics on organizational performance. The proposed framework should therefore be tested through case studies, longitudinal field research, and controlled evaluations in specific domains. Subject to that limitation, the central argument is that predictive decision-making is best understood as a governed organizational capability: the sustainable advantage comes from connecting scalable information infrastructure to valid predictions and accountable action.

References

1. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.15779/Z38BG31
2. Chang, W., & Grady, N. (2019). NIST big data interoperability framework: Volume 1, definitions (NIST Special Publication 1500-1r2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1500-1r2
3. Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188. https://doi.org/10.2307/41703503
4. Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: The new science of winning. Harvard Business School Press.
5. Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
6. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
7. Jamarani, A., Haddadi, S., Sarvizadeh, R., Haghi Kashani, M., Akbari, M., & Moradi, S. (2024). Big data and predictive analytics: A systematic review of applications. Artificial Intelligence Review, 57, Article 176. https://doi.org/10.1007/s10462-024-10811-5
8. Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. SAGE Publications.
9. Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

10. Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.
11. Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS Quarterly, 35(3), 553–572. https://doi.org/10.2307/23042796
12. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems 28 (pp. 2503–2511).
13. Tosi, D., Kokaj, R., & Roccetti, M. (2024). 15 years of big data: A systematic literature review. Journal of Big Data, 11, Article 73. https://doi.org/10.1186/s40537-024-00914-9
14. Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. International Journal of Production Economics, 165, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031
15. Wissuchek, C., & Zschech, P. (2025). Prescriptive analytics systems revised: A systematic literature review from an information systems perspective. Information Systems and e-Business Management, 23, 279–353. https://doi.org/10.1007/s10257-024-00688-w