Academic Journal of Emerging Technologies
ISSN Print: 3104-4417
ISSN Online: 3104-4425
CD ISSN: 3104-4433
https://doi.org/10.63313/AJET.9063

Feature Selection and Noise Control in High-Dimensional Data: Tourism vs. Healthcare

Shengyu Gu
School of Geography and Tourism, Huizhou University, Huizhou, China
Email: miller@hzu.edu.cn

How to cite this paper: Gu, S. (2026). Feature Selection and Noise Control in High-Dimensional Data: Tourism vs. Healthcare. Academic Journal of Emerging Technologies, 3(2), 11–30. ISSN Print: 3104-4417; ISSN Online: 3104-4425.
https://doi.org/10.63313/AJET.9063
Published: 2026-07-29

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Abstract

This study proposes a unified governance-oriented framework for managing high-dimensional multi-source data through the integration of feature selection and noise control, with the explicit goal of improving decision efficiency rather than purely technical accuracy. Drawing on comparative evidence from smart tourism and healthcare systems, the research demonstrates that the main challenge of high-dimensional data lies not in domain-specific content but in shared governance structures, including multi-source integration, systematic noise generation, and decision pressure under uncertainty. Feature selection is conceptualized as an institutional information purification mechanism that filters variables according to relevance, interpretability, and actionability. Noise control is framed as a governance pipeline involving profiling, harmonization, cleaning, validation, and continuous monitoring. Together, these processes transform complex and heterogeneous datasets into reliable and governable decision inputs. The study shifts evaluation criteria from medical or predictive accuracy toward governance-oriented indicators such as latency, stability, interpretability, robustness, and actionability. Tourism is positioned as the methodological origin and practical field of governance-oriented data purification, while healthcare serves as a comparative application context. The findings establish feature selection and noise control as a transferable management science methodology for high-dimensional data governance and institutional decision-making.

Keywords

High-dimensional data; feature selection; noise control; multi-source governance; decision efficiency; smart tourism; precision healthcare; data purification; decision usability

1. Introduction: Data Purification as Decision Efficiency

The rapid expansion of digital data infrastructures has transformed tourism management and healthcare administration into data-intensive decision environments. However, the increasing availability of multi-source and high-dimensional data does not automatically translate into better decisions.

Instead, organizations face growing difficulties in extracting reliable and actionable information from complex datasets characterized by redundancy, noise, and semantic inconsistency. This study positions feature selection and noise control not as purely algorithmic procedures, but as a form of governance-oriented data engineering designed to enhance managerial decision efficiency rather than model accuracy alone. High-dimensional data environments are widely recognized as problematic for decision-making due to the "curse of dimensionality," in which the number of variables increases faster than the capacity of analytical and organizational systems to interpret them (Bellman, 1961). In tourism, smart destination systems integrate heterogeneous data sources such as mobile trajectories, booking transactions, online reviews, sensor networks, and regulatory indicators. These data streams generate high-dimensional representations of tourist behavior and destination performance, but they also introduce substantial noise, duplication, and bias that can delay or distort policy responses (Adamış & Pınarbaşı, 2022). Similarly, in precision healthcare, multi-omics profiles, medical imaging, and electronic health records (EHRs) produce extremely high-dimensional datasets that challenge institutional risk stratification and resource planning processes (Raghupathi & Raghupathi, 2014). Across both domains, the core difficulty is not data scarcity but data overload and contamination. Common challenges include dimensional explosion, redundant features, measurement errors, semantic inconsistency across platforms, missing values, and selection bias. These problems undermine the credibility and usability of analytical outputs and directly affect decision reliability. As Batini and Scannapieco (2016) argue, data quality and noise are not merely technical issues but organizational governance concerns because they influence accountability, transparency, and trust in decision systems. Feature selection and noise control address these challenges by acting as mechanisms of information purification. Feature selection reduces dimensionality by identifying variables that are relevant, non-redundant, and interpretable for decision purposes (Guyon & Elisseeff, 2003). Noise control mitigates the impact of errors, inconsistencies, and distortions that arise from measurement processes and data integration pipelines (Herland, Khoshgoftaar, & Wald, 2014). Together, these processes transform raw data into structured decision inputs that are aligned with institutional objectives. In smart tourism governance, the consequences of insufficient data purification are evident. Unfiltered social media reviews may contain manipulation or bot-generated content; GPS traces may include positional drift; and booking data may follow different classification standards across platforms. These sources generate noise that weakens demand forecasting and congestion management. In healthcare systems, batch effects in genomic data, coding differences in EHRs, and incomplete clinical records introduce uncertainty into patient risk grouping and population-level planning (Raghupathi & Raghupathi, 2014). In both cases, decision delays and misclassification can lead to inefficient

resource allocation and reduced institutional legitimacy. The shared challenge across tourism and healthcare suggests that high-dimensional noise is not domain-specific but structural. Chen, Mao, and Liu (2014) emphasize that big data analytics must be evaluated by their capacity to improve decision effectiveness, not merely predictive performance. From this perspective, feature selection and noise control should be understood as governance tools that enable institutions to manage complexity and support timely action. This study advances three core research questions: RQ1: Can the noise structures of high-dimensional data in different domains be unified and conceptualized as a governance problem rather than a purely technical one? RQ2: How does feature selection function as an information purification mechanism that enhances decision efficiency across tourism and healthcare systems? RQ3: How can the benefits of dimensionality reduction and noise control be evaluated using governance-oriented indicators rather than domain-specific accuracy metrics? By framing feature selection and noise control as instruments of decision governance, this research departs from conventional algorithm-centered approaches. It proposes that the value of data purification lies in its contribution to decision efficiency, defined as the ability to produce reliable, interpretable, stable, and actionable classifications within institutional workflows. Tourism and healthcare are treated as comparative cases that reveal the same underlying logic: high-dimensional data must be purified before it can serve governance objectives. In doing so, the study situates feature selection and noise control within management science and data governance rather than biomedical or purely computational research. Healthcare appears as one application context of high-dimensional decision systems, while smart tourism represents another. The central contribution is a unified conceptualization of data purification as a governance-oriented process that links dimensionality reduction and noise management directly to institutional decision performance.

2. Conceptual Framework: Governance-Oriented Feature Selection

Classification in tourism management and healthcare analytics must be understood within a broader theoretical framework that integrates complex systems theory, bounded rationality, heuristic decision-making, and managerial decision support systems (DSS). Together, these four pillars explain why classification in real-world environments is fundamentally a decision problem rather than a purely computational task.

2.1. Complex Systems Theory

High-dimensional data environments require more than computational optimization; they demand institutional mechanisms that regulate how information is filtered, validated, and translated into decisions. This study proposes a

governance-oriented conceptual framework in which feature selection and noise control are understood as managerial instruments for data purification rather than as isolated analytical techniques. The framework integrates four core modules: (1) high-dimensional multi-source data governance, (2) feature selection as information purification, (3) noise control as reliability and validity assurance, and (4) decision efficiency as a governance outcome. First, high-dimensional multi-source data governance refers to the institutional capacity to manage excessive and heterogeneous variables generated by digital systems. Governance theory emphasizes that data are not neutral artifacts but organizational resources that must be curated through standards, rules, and accountability structures (Floridi et al., 2018). In tourism systems, hundreds of indicators derived from mobility tracking, booking platforms, and social media exceed the cognitive and operational capacity of destination managers. In healthcare systems, tens of thousands of genomic and clinical variables overwhelm institutional decision processes. Wang and Strong (1996) argue that without explicit governance mechanisms, data richness leads to poor decision quality due to inconsistency and lack of contextual meaning. Second, feature selection as information purification functions as an institutional filtering mechanism that transforms raw data into policy-relevant knowledge. Rather than maximizing predictive performance, governance-oriented feature selection prioritizes relevance, interpretability, and actionability. Liu and Motoda (2012) conceptualize feature selection as a knowledge discovery process that reduces complexity while preserving decision significance. In tourism governance, this means retaining indicators directly linked to congestion control or destination investment strategies. This data purification approach is highly critical when managing the multi-source, heterogeneous records of localized cultural assets, such as filtering platform and behavioral noise to optimize the digital exhibition and sustainable tourism mapping of Huidong Intangible Cultural Heritage. In healthcare administration, it implies focusing on clinically actionable predictors rather than statistically marginal biomarkers. Feature selection therefore acts as a rule-based gatekeeper between data abundance and managerial sense-making. Third, noise control as reliability and validity assurance addresses the distortions introduced by measurement error, semantic inconsistency, and integration bias. Noise is not merely a statistical inconvenience but a governance risk because it undermines institutional trust in data-driven decisions. Zhu and Wu (2004) demonstrate that noisy data systematically degrade classification outcomes and increase misclassification probability. From a governance perspective, noise control includes harmonization standards, anomaly detection, and missing-value management that stabilize decision inputs. In tourism, this involves filtering fake reviews and correcting GPS drift; in healthcare, it requires batch-effect correction and coding normalization across clinical systems. These processes ensure that decisions are grounded in consistent and credible evidence. Fourth, decision efficiency as a

governance outcome reframes evaluation criteria away from algorithmic accuracy toward institutional usability. Decision efficiency is defined by four dimensions: speed, stability, interpretability, and actionability. March and Simon (1993) highlight that organizations operate under bounded rationality and therefore value decision procedures that reduce uncertainty and time costs. Shmueli and Koppius (2011) further argue that predictive analytics should be judged by their contribution to managerial action rather than statistical fit alone. Feature selection and noise control enhance decision efficiency by simplifying models, stabilizing outputs, and enabling accountable interventions. Together, these four modules form a unified governance-oriented logic of data purification. High dimensionality creates decision overload; redundancy introduces instability; noise generates misclassification risk; and unfiltered data undermine institutional accountability. Feature selection and noise control respond to these challenges by operationalizing governance principles within analytical pipelines. As summarized in Table 1, the same conceptual constructs appear in both tourism and healthcare systems, confirming their structural equivalence as governance problems rather than domain-specific anomalies. The table demonstrates how high dimensionality leads to decision latency, redundancy to instability, and noise to misclassification risk, while feature selection and noise control generate interpretability and reliability. The final outcome-decision efficiency-represents the core managerial value of data purification. This conceptual framework supports the central argument that feature selection and noise control are not merely technical preprocessing steps but institutional mechanisms embedded in decision governance. Tourism and healthcare serve as comparative cases illustrating how different sectors confront the same structural challenge: transforming high-dimensional, noisy data into usable and accountable decisions. By shifting the analytical focus from algorithms to governance outcomes, this study establishes data purification as a core component of management science and digital governance.

Table 1. Governance-Oriented Conceptual Mapping of Data Purification

Construct	Definition (Governance Lens)	Tourism Example	Healthcare Example	Managerial Outcome
High dimensionality	Excess variables exceed decision capacity	Hundreds of mobility + booking indicators	Tens of thousands of omics markers	Decision latency
Redundancy	Overlapping indicators inflate models	Duplicate demand proxies	Correlated biomarkers	Instability
Noise	Measurement and semantic errors distort signal	Fake reviews, GPS drift	Batch effects, coding inconsistency	Misclassification risk
Feature selection	Institutional filtering rule-set	Keep policy-relevant predictors	Keep clinically actionable predictors	Interpretability
Noise control	Quality assurance and harmonization	De-duplication, normalization, anomaly control	Batch correction, missingness handling	Reliability
Decision efficiency	Faster, stable, accountable decisions	Crowding policy triggers	Risk cohort planning	Governance usability

3. Data Environments & Noise Taxonomy

High-dimensional data environments in tourism and healthcare are often treated as

technical systems in which noise is assumed to be random error. This study adopts a different perspective: noise is not accidental, but a structural byproduct of institutional settings, platform architectures, and governance regimes. In digitally mediated systems, data are produced through sensors, platforms, administrative rules, and strategic human behavior. As a result, noise reflects how organizations define, collect, and integrate information rather than simple statistical disturbance. In smart tourism systems, data originate from booking platforms, GPS devices, social media, and policy monitoring tools. These infrastructures embed commercial incentives and technical constraints that systematically shape data quality. Platform recommendation and promotion mechanisms bias observed demand patterns, while bot-generated or strategically manipulated reviews distort sentiment signals (Luca & Zervas, 2016). This high-dimensional challenge is particularly evident in the digital management of cultural heritage, such as Huidong Intangible Cultural Heritage platforms. In these environments, management systems must apply rigorous feature selection and noise control to filter out redundant behavioral data and unstructured multimedia noise, thereby capturing core tourist engagement metrics for sustainable tourism planning. GPS-based mobility data suffer from positional drift and device heterogeneity, leading to measurement noise and artificial spikes in visitor counts (Fuller, 2009). Moreover, different tourism platforms apply incompatible definitions of "visit," "stay," or "overnight," creating semantic noise that prevents cross-platform comparability (Bowker & Star, 2000). Healthcare data environments exhibit parallel but domain-specific noise structures. Multi-omics datasets are affected by batch effects, where measurements vary systematically across laboratories and instruments rather than biological conditions (Leek et al., 2010). Clinical data are shaped by coding standards, reimbursement rules, and institutional routines, leading to semantic inconsistencies in diagnosis and procedure categories. Missing data mechanisms further complicate decision-making. Rubin (1976) distinguishes between missing at random (MAR) and missing not at random (MNAR), both of which introduce bias into risk stratification models. Selection bias also emerges when datasets represent only certain hospitals or demographic groups, echoing the well-known problem of non-representative samples in policy analysis (Heckman, 2013). Across both tourism and healthcare, noise is therefore embedded in governance structures rather than isolated sensors. Behavioral noise arises from strategic user actions. Tourists manipulate booking systems and review platforms for discounts or visibility; patients follow irregular care pathways influenced by access, cost, and institutional constraints. These behaviors produce data patterns that deviate from idealized assumptions of randomness and independence. Temporal noise further destabilizes systems. Seasonality, holidays, epidemics, and policy shocks create regime shifts in data distributions that invalidate static models (Hamilton, 1989). Integration noise represents a particularly critical governance risk. When multiple

platforms and agencies merge datasets, identifier mismatches, duplicate records, and inconsistent schemas generate conflicts that compromise accountability. Classic record linkage research demonstrates that errors in merging records across systems propagate directly into classification and resource allocation decisions (Fellegi & Sunter, 1969). In tourism governance, mismatched platform IDs prevent accurate monitoring of visitor flows. In healthcare administration, incorrect patient matching undermines safety and resource planning. These observations support the central claim that noise should be conceptualized as a governance artifact rather than random fluctuation. Data infrastructures are socio-technical systems shaped by institutional rules, incentives, and coordination failures. Noise reflects tensions between platforms, organizations, and users. Without explicit governance-oriented noise control mechanisms, feature selection becomes unreliable and decision efficiency deteriorates. Table 2 synthesizes this argument by presenting a cross-domain taxonomy of noise types in multi-source high-dimensional data. The taxonomy highlights that the same categories of noise appear in tourism and healthcare, though manifested through different mechanisms. Measurement noise generates false alarms; semantic noise creates non-comparability; behavioral noise produces biased decisions; sampling bias leads to inequity; temporal noise causes instability; and integration noise results in governance failure. Understanding noise as institutional rather than random shifts the analytical focus from technical correction to governance design. Noise control becomes a matter of harmonization standards, auditing rules, and accountability frameworks rather than merely filtering algorithms. In this sense, feature selection and noise management are inseparable from data governance. They determine which signals are legitimate for decision-making and which are excluded as unreliable. This taxonomy provides the foundation for governance-oriented feature selection by clarifying what types of distortions must be addressed before dimensionality reduction can improve decision efficiency. Tourism and healthcare serve as comparative cases demonstrating that different sectors encounter structurally similar noise regimes. The implication is that effective noise control requires institutional coordination and policy design, not only computational sophistication.

Table 2. Noise Taxonomy in Multi-Source High-Dimensional Data

Noise Type	Mechanism	Typical Symptoms	Tourism Data Manifestation	Healthcare Data Manifestation	Governance Risk
Measurement noise	Sensor / platform error	Drift, spikes	GPS drift, visitor count errors	Instrument error, imaging artifacts	False alarms
Semantic noise	Inconsistent definitions	Incompatible fields	"Visit / stay / overnight" discrepancies	Diagnosis coding differences	Non-comparability
Behavioral noise	Strategic user behavior	Manipulation	Fake reviews, booking gaming	Irregular care pathways	Biased decisions
Sampling bias	Non-representative data	Skewed distributions	Only online tourists covered	Single-center or subgroup data	Inequity
Temporal noise	Seasonality and shocks	Regime shifts	Holidays, crises, mega-events	Epidemic waves, policy changes	Instability
Integration noise	Merge conflicts	Duplicates, mismatches	Cross-platform ID failures	Multi-system patient ID mapping errors	Governance failure

4. Feature Selection Mechanisms

Feature selection in high-dimensional environments is commonly treated as a technical preprocessing step aimed at improving model accuracy. This study adopts a different stance: feature selection is a decision mechanism that governs which pieces of information are allowed to influence institutional action. Its primary function is not computational efficiency alone, but the purification of information so that decisions become interpretable, stable, and actionable. In this sense, feature selection operates at the intersection of analytics and governance, transforming data abundance into managerial relevance. Rather than focusing on algorithmic implementation, this section conceptualizes feature selection through four complementary mechanisms: filter-based selection, wrapper-based selection, embedded selection, and domain-governed selection. Each mechanism plays a distinct role in reducing dimensionality, controlling noise, and supporting decision efficiency in tourism and healthcare systems.

Filter mechanisms evaluate features independently of any predictive model, relying on statistical relevance, mutual information, or variance thresholds to remove irrelevant or redundant variables. Their governance value lies in speed and scalability, enabling institutions to quickly reduce dimensional explosion before deeper analysis. In tourism governance, filter-based selection is used to identify a small set of indicators that most strongly correlate with visitor flows, congestion levels, or seasonal demand. Variables with negligible variance or weak association with policy outcomes can be discarded without affecting strategic decisions. In healthcare administration, filters are applied to eliminate biomarkers or clinical variables with minimal contribution to risk stratification, thereby reducing batch-induced noise and simplifying institutional dashboards. From a decision perspective, filter mechanisms support early-stage simplification, allowing managers to work with condensed information rather than overwhelming data matrices. However, their governance limitation is that they may overlook interaction effects that are meaningful for policy interventions. As Hall (1999) notes, relevance screening improves efficiency but does not capture the structural dependencies among variables that characterize complex systems.

Wrapper mechanisms evaluate subsets of features according to their contribution to a target decision objective, such as congestion alerts or risk group classification. Instead of asking whether a variable is relevant in isolation, wrappers ask whether combinations of variables improve decision outcomes. In tourism systems, wrapper-based selection is applied to identify the optimal set of indicators for early-warning models of overcrowding or crisis response. The focus is not on statistical elegance but on assembling the feature combination that best supports operational triggers. In healthcare governance, wrappers search for the most effective combination of patient attributes for risk grouping or resource prioritization. The governance strength of wrapper mechanisms lies in their performance orientation. They align feature selection directly with institutional

objectives. However, this comes at the cost of computational complexity and reduced transparency. Kohavi and John (1997) emphasize that wrapper methods are powerful but resource-intensive, which limits their feasibility in real-time governance environments. Moreover, their dependence on specific decision objectives may reduce generalizability across institutional contexts.

Embedded mechanisms integrate feature selection into the structure of analytical models themselves, using regularization or tree-based importance measures to identify influential variables. Their key contribution is the balance between stability and interpretability. In tourism management, embedded selection allows policymakers to extract the most influential predictors of demand or satisfaction directly from forecasting models. Tree-based approaches reveal which behavioral or environmental factors drive congestion or destination choice. In healthcare administration, regularization-based selection identifies stable clinical variables that consistently contribute to patient stratification across datasets. From a governance standpoint, embedded mechanisms produce feature sets that are institutionally communicable. They provide justifications for why certain variables are retained and others excluded, supporting accountability and auditability. Friedman, Hastie, and Tibshirani (2010) argue that embedded selection stabilizes decision rules by penalizing complexity and discouraging overfitting, which aligns with governance requirements for robustness. The limitation of embedded mechanisms is their dependency on specific model structures. Different analytical frameworks yield different feature importance rankings, which can create interpretive conflicts across departments or agencies.

The most distinctive contribution of this study is the introduction of domain-governed selection, which extends beyond technical criteria to include institutional constraints such as actionability, legitimacy, and policy relevance. This mechanism recognizes that not all statistically significant features are suitable for governance. In tourism governance, domain-governed selection retains only indicators that can trigger concrete policy actions, such as crowd control measures, pricing adjustments, or infrastructure deployment. Variables that cannot be translated into operational responses are excluded even if they improve predictive accuracy. In healthcare administration, only clinically actionable predictors that support resource allocation or treatment prioritization are preserved. This mechanism reflects the management science perspective that decision systems must satisfy organizational accountability. As Suchman (1995) argues, legitimacy is a fundamental requirement of institutional action. Feature selection must therefore align with regulatory frameworks, professional standards, and public trust. Domain-governed selection ensures that analytical outputs are not only accurate but also governable. The governance caveat is that this constraint may reduce statistical performance by excluding complex but informative variables. However, the trade-off favors interpretability and institutional usability over marginal gains in

prediction.

Across these four mechanisms, feature selection emerges as a layered decision process rather than a single algorithmic step. Filter mechanisms enable rapid reduction of dimensionality; wrapper mechanisms align features with objectives; embedded mechanisms stabilize interpretability; and domain-governed mechanisms ensure legitimacy and actionability. Together, they form a governance-oriented architecture of information purification. Table 3 summarizes these mechanisms and illustrates how they operate in tourism and healthcare systems. The table highlights that feature selection serves managerial functions such as policy triggering, risk grouping, and resource planning. It also shows that governance caveats must be considered alongside technical strengths. By reframing feature selection as a decision mechanism, this study shifts attention from computational performance to institutional value. The core question becomes not "Which algorithm selects the best features?" but "Which features enable accountable, timely, and stable decisions?" Tourism and healthcare provide parallel illustrations of this logic. Both domains require that dimensionality reduction be aligned with governance objectives rather than purely statistical criteria.

Table 3. Feature Selection as a Decision Mechanism

Category	Selection Logic	Tourism Use-Case	Healthcare Use-Case	Strength	Governance Caveat
Filter	Relevance screening	Rapidly identify key drivers of visitor flows	Initial screening of high-impact biomarkers	Fast, scalable	May miss interactions
Wrapper	Goal-driven search	Optimal indicators for congestion warning	Optimal features for risk stratification	Performance-oriented	Computational cost
Embedded	Model-structural	Tree models extract demand drivers	Regularization selects stable predictors	Stable + interpretable	Model dependency
Governance-constrained	Actionability + legitimacy	Retain only policy-triggering indicators	Retain only resource-allocation predictors	Accountability	May reduce accuracy

5. Noise Control Pipelines & Data Harmonization

Noise control in high-dimensional data environments cannot be reduced to a single preprocessing step. This study conceptualizes noise control as a governance pipeline, in which multiple stages of institutional coordination transform heterogeneous and unreliable data into decision-ready information. Rather than treating noise as random disturbance, the pipeline approach recognizes noise as a structural outcome of cross-platform data production, incompatible standards, and organizational fragmentation. Tourism and healthcare systems share this condition because both rely on multi-source data infrastructures that require continuous harmonization rather than one-time correction. A governance-oriented noise control pipeline typically includes the following stages: profiling, harmonization, cleaning, validation, and monitoring. Each stage fulfills a distinct institutional function and produces artifacts that support accountability and decision reliability.

Profiling is the diagnostic stage in which institutions identify the types and

distributions of noise embedded in datasets. It involves descriptive analysis of outliers, inconsistencies, missingness patterns, and distributional anomalies. In tourism systems, profiling reveals abnormal review frequencies, suspicious rating clusters, or unexpected spikes in GPS-derived visitor counts. In healthcare systems, profiling detects batch effects across laboratories or systematic differences between data collection sites. From a governance perspective, profiling generates data quality reports that document known distortions and uncertainties. These reports serve as institutional evidence that data limitations are recognized and addressed rather than ignored. Pipino, Lee, and Wang (2002) argue that explicit data quality assessment is essential for organizational trust in information systems.

Harmonization addresses semantic noise by aligning definitions, units, and schemas across platforms. Tourism governance requires a unified interpretation of concepts such as "visit," "stay," and "overnight," which are often defined differently across booking systems and mobility platforms. Healthcare governance faces similar problems with diagnosis codes, laboratory measurements, and administrative fields. Harmonization produces data dictionaries and metadata standards that formalize shared meanings. Bowker and Star (2000) demonstrate that classification systems shape organizational action; without harmonized semantics, cross-system integration leads to non-comparable indicators and governance failure. Harmonization therefore acts as a normative mechanism that stabilizes data interpretation.

Cleaning focuses on technical reduction of noise through normalization, anomaly detection, deduplication, and missing-value handling. In tourism data, GPS trajectories may be smoothed and duplicate records removed. In healthcare data, batch correction and imputation procedures address systematic and random distortions. However, cleaning is not purely technical. It encodes institutional decisions about what counts as an anomaly or an acceptable value. Rahm and Do (2000) emphasize that data cleaning reflects organizational assumptions and priorities. The output of this stage is a cleaned dataset that is suitable for feature selection and decision modeling.

Validation introduces institutional checks to ensure that noise control improves decision reliability rather than introducing new biases. In tourism governance, cleaned indicators are compared with manually sampled surveys or observational audits. In healthcare systems, processed datasets are benchmarked against reference cohorts or gold-standard registries. Validation produces an audit trail, documenting how data were transformed and verified. This trail supports accountability and allows decision-makers to trace policy actions back to data preparation steps. According to Kahn et al. (2015), transparent validation procedures are critical for maintaining confidence in analytical outputs used for governance.

Noise control is not static. Temporal shifts, seasonal effects, and policy changes

continually introduce new distortions. Monitoring establishes dashboards and alerts that track data drift and emerging anomalies. In tourism, holiday periods and crisis events require real-time monitoring of indicator stability. In healthcare, shifts in testing protocols or population composition necessitate ongoing surveillance. Monitoring outputs a governance dashboard that integrates quality indicators with operational metrics. This reinforces the view that noise control is part of continuous institutional management rather than a one-off technical fix.

The key contribution of this pipeline framework is to demonstrate that tourism and healthcare systems are structurally isomorphic in their noise control needs. Both domains must govern cross-system semantics, align measurement standards, and institutionalize accountability for data transformations. The difference lies in content, not in governance logic. This reinforces the argument that noise control is fundamentally a problem of cross-system coordination rather than domain expertise. The same pipeline architecture-profiling, harmonization, cleaning, validation, and monitoring-applies regardless of whether the data describe tourists or patients. Feature selection and decision efficiency depend on these governance processes, not on specialized medical or tourism knowledge. Table 4 summarizes the noise control pipeline and illustrates how each stage functions in tourism and healthcare contexts while producing governance artifacts that support institutional decision-making.

Table 4. Noise Control Pipeline for Multi-Source Governance

Stage	Objective	Tourism Example	Healthcare Example	Output Artifact (Governance)
Profiling	Discover noise patterns	Abnormal review distributions	Batch difference detection	Data quality report
Harmonization	Align semantics	Unified definition of "visit"	Unified coding and fields	Data dictionary
Cleaning	Reduce random/system noise	GPS smoothing, deduplication	Batch correction, imputation	Cleaned dataset
Validation	Ensure decision reliability	Manual sample comparison	Benchmark cohort comparison	Audit trail
Monitoring	Continuous governance	Holiday drift surveillance	Data drift monitoring	Governance dashboard

6. Decision Efficiency Evaluation

Conventional evaluation of high-dimensional data models relies heavily on domain-specific accuracy metrics, such as diagnostic precision in healthcare or forecasting error in tourism demand. This study shifts the evaluation lens from technical correctness to governance performance, arguing that the value of feature selection and noise control lies in their ability to improve institutional decision efficiency. Decision efficiency is defined as the capacity of a data pipeline to generate timely, stable, interpretable, robust, and actionable outputs that support managerial and policy interventions. This section operationalizes decision efficiency through five governance-oriented dimensions: decision latency, stability, interpretability, robustness, and actionability. These criteria emphasize organizational usability rather than algorithmic optimization.

Decision latency measures the time required for data to be transformed into an

actionable decision. In high-dimensional systems, excessive features and unmanaged noise increase processing time and slow organizational response. Feature selection reduces dimensionality, while noise control pipelines eliminate reprocessing caused by inconsistencies or anomalies. In tourism governance, reduced latency enables faster congestion alerts and crowd-control responses during peak seasons. In healthcare administration, shorter latency allows rapid generation of patient risk cohorts for resource planning. From a governance perspective, latency is a critical indicator of institutional responsiveness. Davenport and Harris (2007) note that competitive and policy advantages increasingly depend on how quickly analytics can be translated into decisions rather than on marginal accuracy gains.

Stability refers to the consistency of classification or grouping results when small perturbations occur in the data. High-dimensional noise amplifies volatility, leading to frequent reclassification and policy oscillation. Feature selection improves stability by removing redundant and weakly informative variables, while noise control reduces random fluctuations. In tourism systems, stable clustering of visitor segments supports long-term destination strategies rather than short-term reactive measures. In healthcare systems, consistent risk stratification prevents frequent shifts in resource allocation plans. Breiman (1996) argues that unstable models undermine trust and interpretability, making stability a governance requirement rather than a technical luxury.

Interpretability evaluates whether decision drivers can be clearly communicated to stakeholders. Governance requires explanations that justify why certain actions are taken. Feature selection enhances interpretability by limiting decision inputs to a manageable set of variables, while noise control ensures that these variables are credible. In tourism management, interpretability means identifying which factors-such as weather, accessibility, or events-drive congestion policies. In healthcare governance, it involves explaining which indicators justify prioritizing certain patient groups. Lipton (2018) emphasizes that interpretability is essential for accountability and institutional legitimacy, especially when decisions affect public welfare.

Robustness measures the ability of decision systems to maintain performance when confronted with noise, missing values, or integration errors. High-dimensional datasets are particularly vulnerable to such disruptions. Noise control pipelines enhance robustness through harmonization, anomaly detection, and imputation, while feature selection limits sensitivity to unstable variables. In tourism, robustness ensures that fake reviews or GPS drift do not trigger false policy alarms. In healthcare, it ensures that batch effects or incomplete records do not distort population risk assessments. Huber and Ronchetti (2009) argue that robustness is fundamental for decision systems operating under uncertainty and imperfect information.

Actionability is the defining criterion of governance-oriented evaluation. It measures whether analytical outputs can be directly mapped onto institutional actions. A classification result has value only if it triggers policies, resource deployment, or operational adjustments. In tourism governance, actionable outputs include crowd-limiting measures, pricing adjustments, or transportation reallocation. In healthcare administration, they include bed assignment, workforce planning, or preventive intervention programs. Actionability connects data purification to institutional authority, ensuring that analytics are embedded in decision rules rather than remaining abstract insights (Mintzberg, Raisinghani, & Théorêt, 1976).

These five metrics together redefine performance as decision efficiency, not predictive accuracy. Feature selection and noise control improve decision efficiency by shortening latency, increasing stability, enhancing interpretability, strengthening robustness, and ensuring actionability. Tourism and healthcare systems exhibit structural equivalence in this respect: both must transform complex data into governable decisions under uncertainty. Table 5 summarizes how these metrics apply across domains and how they can be operationalized through institutional measurement approaches such as pipeline logs, resampling tests, and audit artifacts. This evaluation framework demonstrates that data purification should be judged by its contribution to governance usability rather than technical perfection.

Table 5. Decision Efficiency Metrics After Feature Selection + Noise Control

Metric	Definition	Tourism Decision Example	Healthcare Governance Example	Measurement Approach
Latency	Time from data ingestion to decision output	Congestion warning response time	Risk cohort generation time	Pipeline timing logs
Stability	Consistency under perturbation	Repeatability of visitor clustering	Consistency of risk stratification	Bootstrap / resampling
Interpretability	Clarity of decision drivers	Key factors behind crowd control	Clear basis for resource allocation	Feature rationale cards
Robustness	Performance under noise and missing data	Resistance to fake reviews	Resistance to batch effects / missingness	Stress tests
Actionability	Ability to trigger concrete actions	Flow limitation and scheduling	Bed and staff allocation	Decision rule mapping

7. Cross-Domain Comparison & Transferability

The comparison between tourism and healthcare reveals a fundamental distinction between data content and governance structure. While tourism systems process mobility traces, bookings, and user-generated content, and healthcare systems handle clinical records, imaging, and molecular profiles, both domains confront the same institutional challenges: integrating multi-source data, managing systematic noise, reducing dimensionality, and supporting high-stakes decisions under uncertainty. This section argues that the shared governance logic, rather than sectoral specificity, determines the effectiveness of feature selection and noise control. From a structural perspective, both tourism and healthcare operate as

complex decision systems characterized by heterogeneous inputs and time-sensitive outcomes. In tourism governance, destination managers must classify visitor segments and detect congestion risks in real time. In healthcare administration, institutions stratify patient risk and allocate limited resources. Despite different subject matter, both rely on data pipelines that must filter irrelevant variables, control distortions, and generate interpretable groupings. Donabedian (1988) emphasizes that organizational performance depends on how information is transformed into action through structured processes. Feature selection and noise control represent such processes, ensuring that data are suitable for institutional decision-making. The cross-domain similarity becomes particularly evident when examining noise generation mechanisms. Tourism platforms and healthcare information systems both embed institutional incentives and technical constraints that produce systematic distortions. Platform bias, inconsistent definitions, and integration conflicts arise in tourism due to commercial platforms and fragmented policy regimes. In healthcare, batch effects, coding differences, and missing data mechanisms stem from institutional heterogeneity and regulatory frameworks. Power and Heavin (2017) argues that data quality issues are inseparable from governance arrangements because they reflect organizational priorities and accountability structures. This insight supports the claim that noise control is not domain-specific but governance-specific. Feature selection also exhibits strong transferability across domains. In tourism, dimensionality reduction simplifies visitor classification and enables timely policy responses. In healthcare, it stabilizes patient stratification and improves resource planning. In both cases, feature selection reduces cognitive and operational overload. Simon's (1960) theory of bounded rationality suggests that organizations must simplify information environments to make decisions feasible. Feature selection operationalizes this principle by translating high-dimensional complexity into manageable decision inputs. Another dimension of transferability lies in harmonization and semantic alignment. Tourism governance requires standardizing concepts such as "visit" or "stay" across platforms, while healthcare governance standardizes diagnoses and clinical fields. These harmonization processes are structurally identical in function: they ensure interoperability and comparability across systems. Borgman (2017) highlights that data infrastructures achieve value only when supported by shared conventions and standards. This reinforces the view that feature selection and noise control are embedded in institutional coordination rather than technical domains. Dynamic environments further strengthen the case for transferability. Tourism demand fluctuates with seasons and crises; healthcare data shift with epidemics and policy changes. Both require drift monitoring and adaptive noise control to maintain decision reliability. Adaptive governance frameworks emphasize the need for continuous feedback and adjustment in complex systems (Janssen & van der Voort, 2016). Feature selection and noise control pipelines thus function as adaptive

governance instruments, not static preprocessing tools. The central implication is that data purification constitutes a transferable methodology of governance. It can be applied to any smart governance system that faces high-dimensional, noisy, and heterogeneous data. Tourism and healthcare serve as empirical illustrations of this broader logic. Their differences in content mask deep similarities in decision structures: both must reduce complexity, ensure reliability, and support accountable action. Table 6 summarizes this transferability by mapping core data purification functions across tourism, healthcare, and smart governance systems. Each function-feature selection, noise control, harmonization, drift monitoring, and decision efficiency-appears consistently across domains. The rationale for transfer is not algorithmic similarity but governance necessity: complex systems require institutional mechanisms that regulate which data count as legitimate inputs for decision-making. This cross-domain synthesis confirms that feature selection and noise control should be understood as a data governance methodology rather than a sector-specific analytical technique. Their primary value lies in enabling decision efficiency, institutional trust, and policy usability across domains. Healthcare is therefore not a defining research boundary but one application context of a general governance logic rooted in management science and digital governance.

Table 6. Transferability Matrix of Data Purification Logic

Function	Tourism	Healthcare	Smart Governance Systems	Transfer Rationale
Feature selection	✔	✔	✔	Reduces decision complexity
Noise control	✔	✔	✔	Ensures reliability and legitimacy
Harmonization	✔	✔	✔	Semantic interoperability
Drift monitoring	✔	✔	✔	Dynamic environments
Decision efficiency	✔	✔	✔	Governance outcome

8. Implications: Data Governance for Management Science

This study advances important theoretical, practical, and methodological implications for management science by reframing feature selection and noise control as governance-oriented decision mechanisms rather than technical preprocessing steps.

From a theoretical perspective, this research integrates dimensionality reduction and noise control into the core of management decision theory. Traditional management science has emphasized bounded rationality, information processing limits, and decision structures, but it has rarely examined how high-dimensional data are institutionally filtered before entering decision processes. By conceptualizing feature selection and noise control as information purification mechanisms, this study extends decision support and governance theory into the

domain of data engineering. This perspective shifts attention from algorithmic performance to decision usability. Feature selection becomes a formalized institutional rule set that defines which variables are legitimate inputs to policy and managerial action. Noise control becomes a reliability and validity assurance process embedded in organizational accountability. Together, they operationalize bounded rationality in digital governance environments by reducing complexity while preserving decision relevance. This contributes to management science by establishing data purification as a structural component of decision systems, linking analytics directly to governance outcomes such as stability, interpretability, and institutional trust.

For practice, the findings suggest that smart tourism and healthcare management require a shared governance checklist for high-dimensional data environments. Despite sectoral differences, both domains face the same institutional risks: semantic inconsistency, uncontrolled noise, and opaque feature selection. Effective governance therefore depends on coordinated mechanisms including: Standardization: shared definitions and metadata frameworks for core indicators; Auditing: routine validation of data transformations and feature selection rules; Data lineage: documentation of how raw data evolve into decision variables; Responsibility boundaries: clear institutional ownership of data quality and decision consequences. These elements transform data pipelines into accountable governance infrastructures rather than opaque analytical systems. For tourism authorities, this means ensuring that crowd-management indicators are derived from harmonized and auditable sources. For healthcare administrators, it means guaranteeing that patient stratification variables are traceable, consistent, and aligned with institutional policies. The implication is that data purification must be managed as an organizational function, not delegated solely to technical teams.

Methodologically, the study proposes the principle of actionable feature governance. Not all statistically significant variables should be retained; only those that support legitimate, interpretable, and executable actions should enter decision systems. This introduces a governance constraint into feature selection: variables must satisfy not only relevance but also institutional actionability and legitimacy. Actionable feature governance requires that each retained feature can be mapped onto a management response, such as congestion control, resource allocation, or risk mitigation. This principle aligns analytical design with policy logic and prevents decision systems from becoming technically sophisticated but institutionally unusable. It also establishes a bridge between data science and management science, ensuring that dimensionality reduction serves decision structures rather than purely computational goals. Overall, the implications underscore that feature selection and noise control are not peripheral technical choices but central components of data governance. They shape how organizations see problems, define priorities, and justify actions. Tourism and healthcare illustrate this shared logic: both require

institutionalized data purification to achieve decision efficiency and governance legitimacy.

9. Conclusion

This study reconceptualizes feature selection and noise control in high-dimensional data environments as a governance-oriented methodology for management decision-making. Rather than treating dimensionality reduction and denoising as technical optimization steps, the research positions them as institutional mechanisms that purify information, stabilize classifications, and enhance decision efficiency. Through a comparative analysis of smart tourism and healthcare systems, the study demonstrates that differences lie primarily in data content, while similarities emerge in governance structure. Both domains confront multi-source integration, systematic noise generation, redundancy, and decision pressure. Feature selection and noise control function as transferable instruments that reduce complexity, ensure reliability, and enable accountable action across domains. The core contribution lies in establishing information purification (feature selection + noise control) as a central construct in management science. This construct links high-dimensional data processing with bounded rationality, institutional legitimacy, and governance usability. Decision efficiency is redefined not as predictive accuracy but as the capacity to produce timely, stable, interpretable, robust, and actionable outputs. In terms of research identity, this work affirms that the primary focus is on high-dimensional multi-source data governance and decision efficiency, not on medical analytics per se. Healthcare serves as a comparative application context, while tourism represents the methodological origin and practical laboratory of governance-oriented data purification. The findings confirm that feature selection and noise control constitute a generalizable management science methodology applicable to smart governance systems beyond any single sector. By embedding data purification within institutional decision logic, this study contributes to a broader understanding of how digital transformation reshapes governance, accountability, and managerial action in complex systems.

Acknowledgements

This work was supported by the project "Technical Services for the Operation of Huidong Intangible Cultural Heritage Exhibition Hall" (Project No. 18/35900259094).

References

[1] Adamış, E., & Pınarbaşı, F. (2022). Unfolding visual characteristics of social media communication: reflections of smart tourism destinations. Journal of Hospitality and Tourism Technology, 13(1), 34-61.

[2] Batini, C., & Scannapieco, M. (2016). Data and information quality. Cham, Switzerland: Springer International Publishing, 63.

[3] Bellman, R. (1961, January). A mathematical formulation of variational processes of adaptive type. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics (Vol. 4, pp. 37-49). University of California Press.

[4] Borgman, C. L. (2017). Big data, little data, no data: Scholarship in the networked world. MIT press.

[5] Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences. MIT press.

[6] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The annals of statistics, 24(6), 2350-2383.

[7] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile networks and applications, 19(2), 171-209.

[8] Davenport, T. H., & Harris, J. G. (2007). Competing on analytics: the new science of Winning. Harvard business review press, Language, 15(217), 24.

[9] Donabedian, A. (1988). The quality of care: how can it be assessed?. Jama, 260(12), 1743-1748.

[10] Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American statistical association, 64(328), 1183-1210.

[11] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People-An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and machines, 28(4), 689-707.

[12] Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33, 1-22.

[13] Fuller, W. A. (2009). Measurement error models. John Wiley & Sons.

[14] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[15] Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

[16] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica: Journal of the econometric society, 357-384.

[17] Heckman, J. (2013). Sample selection bias as a specification error. Applied Econometrics, 31(3), 129-137.

[18] Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. Journal of Big data, 1(1), 2.

[19] Huber, P. J., & Ronchetti, E. M. (2009). Robust statistics. 2nd john wiley & sons. Hoboken, NJ, 2.

[20] Janssen, M., & Van Der Voort, H. (2016). Adaptive governance: Towards a stable, accountable and responsive government. Government Information Quarterly, 33(1), 1-5.

[21] Kahn, M. G., Brown, J. S., Chun, A. T., Davidson, B. N., Meeker, D., Ryan, P. B., ... & Zozus, M. N. (2015). Transparent reporting of data quality in distributed data networks. Egems, 3(1), 1052.

[22] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

[23] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics, 11(10), 733-739.

[24] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue, 16(3), 31-57.

[25] Liu, H., & Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer science & business media.

[26] Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. Management science, 62(12), 3412-3427.

[27] March, J. G., & Simon, H. A. (1993). Organizations. John wiley & sons.

[28] Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of" unstructured" decision processes. Administrative science quarterly, 246-275.

[29] Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. Communications of the ACM, 45(4), 211-218.

[30] Power, D. J., & Heavin, C. (2017). Decision support, analytics, and business intelligence. Business Expert Press.

[31] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1), 3.

[32] Rahm, E., & Do Hong, H. (2000). Data cleaning: Problems and current approaches.

[33] Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

[34] Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. MIS quarterly, 553-572.

[35] Simon, H. A. (1960). The new science of management decision.

[36] Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. Academy of management review, 20(3), 571-610.

[37] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.

[38] Zhu, X., & Wu, X. (2004). Class noise vs. attribute noise: A quantitative study. Artificial intelligence review, 22(3), 177-210.