Clustering, Indexing, and Data Structures for High-Dimensional and Categorical Data: Algorithmic Foundations, Modern Advances, and Scalable Analytic Systems

Abstract

This survey provides a comprehensive and critical synthesis of contemporary advances in clustering, indexing, and analytic methodologies for high-dimensional and categorical data. Motivated by the widespread emergence of large, complex datasets in domains such as genomics, healthcare, e-commerce, and network analysis, the paper elucidates the fundamental challenges posed by the "curse of dimensionality," data heterogeneity, and the proliferation of categorical and multimodal variables. The scope encompasses key computational paradigms, including nearest neighbor search, clustering, feature selection, and high-dimensional statistical testing, as well as foundational and emerging indexing structures—from traditional spatial trees to compressed, learned, and hybrid neural indexes.

Key contributions include an in-depth analysis of algorithmic strategies tailored to high-dimensional settings, such as ensemble subspace and consensus spectral clustering, robust tensor decompositions, and adaptive index constructions leveraging machine learning. The survey further evaluates space-efficient storage and hardware-accelerated computation, addressing real-time scalability, dynamic adaptation, and resilience to noisy, adversarial, or streaming data. Comprehensive benchmarking, cluster validation, and open-source ecosystem reviews contextualize methodological innovations within system-level performance and reproducibility frameworks.

Conclusions highlight persisting open problems: balancing statistical rigor and computational efficiency, ensuring robustness and interpretability, integrating ethical and privacy considerations, and advancing standardized benchmarking. The survey delineates future research directions—including federated analytics, neural and retrieval-augmented indexing, and unified analytic platforms—emphasizing that adaptive, accountable, and explainable methodologies are essential to harnessing the potential of high-dimensional data across scientific and societal domains.

ACM Reference Format:
. 2025. Clustering, Indexing, and Data Structures for High-Dimensional and Categorical Data: Algorithmic Foundations, Modern Advances, and Scalable Analytic Systems. In . ACM, New York, NY, USA, 42 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Recent advancements in artificial intelligence (AI) have led to an exponential growth of research output and practical applications across a broad spectrum of fields. This survey aims to provide a comprehensive and accessible synthesis of current developments in the domain, specifically targeting graduate students, researchers, and professionals in computer science, data science, and allied fields seeking an in-depth understanding of contemporary AI paradigms and their foundational techniques.

A unique aspect of our survey is the emphasis on a unified taxonomy that categorizes approaches according to their core methodologies and cross-domain applicability. Notably, this taxonomy explicitly integrates developments from adjacent fields, such as computational neuroscience, statistical physics, and cognitive science, thereby broadening the relevance and impact of AI methodologies beyond the confines of core computer science or AI research alone. We delineate how this taxonomy advances prior surveys by integrating emerging subfields and highlighting underexplored connections between traditionally disparate research areas. The structured mapping provided herein distinguishes our survey and enables practitioners to more readily identify relevant methods for their specific problem domains.

For improved readability and immediate orientation, Table 1 presents a summarized view of the main AI paradigms and their distinguishing features. This table, placed at the outset, reinforces key conceptual takeaways and assists readers in contextualizing subsequent detailed discussions.

Moreover, we identify and detail key research gaps uncovered during our systematic review—including areas such as scalability in complex environments, robustness against adversarial conditions, and ethical implications of large-scale deployment. While recent foundational works (e.g., those appearing in the past two to three years) have addressed certain aspects of these challenges, notable limitations remain. For example, despite advances in model scalability, issues of interpretability and fairness are persistent and often underemphasized relative to performance metrics. Where relevant, we provide a critical comparative discussion of both the strengths and limitations of prevailing approaches, with the intent to guide ongoing and future work toward addressing these open challenges.

The remainder of the paper systematically maps numbered references to their full bibliographic entries, ensuring accurate traceability and improved reader experience.

1.1 Motivation

High-dimensional and categorical data have become pervasive across a broad spectrum of modern analytical domains, driven by rapid advancements in data acquisition, storage, and sensing technologies in fields such as healthcare, genomics, e-commerce, and network analysis [6, 13, 15, 28, 33, 38–40, 49, 52, 53, 60–63, 66,

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Summary of Main AI Paradigms and Their Distinguishing Features

Paradigm	Core Methodology	Applications	Key Limitations
Supervised Learning	Labeled data, loss minimization	Image/audio recognition, NLP	Requires extensive labeled data, limited generalization
Unsupervised Learning	Pattern discovery, clustering	Anomaly detection, data compression	Interpretability, defining useful objectives
Reinforcement Learning	Trial-and-error, rewards	Robotics, game-playing	Sample inefficiency, instability, reproducibility
Transfer Learning	Knowledge reuse, domain adaptation	Cross-lingual learning, medical imaging	Negative transfer risk, domain shift sensitivity
Generative Modeling	Data generation, density estimation	Image synthesis, text generation	Evaluation difficulty, mode collapse

79, 84, 88, 90, 91, 93, 107, 108]. These data types are distinguished not only by an exponential increase in feature dimensionality, but also by the growing prevalence of categorical variables—which are frequently sparse and non-ordinal. This dual trend introduces significant methodological and computational challenges.

One central issue is the so-called "curse of dimensionality," a phenomenon in which distances between data points lose discriminative power as the number of dimensions increases. This undermines the effectiveness of similarity-based techniques, as well as methods for nearest neighbor (NN) search, clustering, and classification [28, 49, 53, 60, 90, 91]. The high-dimensional regime facilitates noise accumulation, where the signal-to-noise ratio degrades, ultimately diminishing the discriminatory capacity of models even as data and computational resources scale [53, 84, 91, 93]. Beyond these statistical hurdles, high dimensionality incurs significant computational overhead in both data storage and algorithmic execution. Classically efficient indexing and search strategies may deteriorate from logarithmic or sublinear time complexity to linear or superlinear, as they struggle with the combinatorial growth of feature configurations [13, 15, 40, 62, 90].

Categorical variables present further complications. Their sparsity and the absence of inherent distance metrics inhibit the straightforward use of standard statistical and machine learning approaches, often necessitating custom distance metrics, specialized encoding techniques, or novel regularization frameworks [6, 38, 39, 52, 88, 93]. In high-stakes applications—such as medical diagnosis or bioinformatics—interpretability is paramount; however, the opacity of many high-dimensional models further constrains their practical adoption [60, 61, 66, 84, 107]. Therefore, the ongoing methodological imperative is to develop algorithms that scale efficiently while delivering robustness, interpretability, and reliability in both statistical inference and practical decision-making.

Ongoing research has yielded notable progress in addressing these obstacles. Innovations include compressed computation, ensemble learning strategies, high-dimensional data structures for efficient indexing, and methods leveraging spectral, consensus, and regularization principles. Collectively, these developments have extended the boundaries of feasible analysis for large and complex datasets [28, 39, 52, 60, 84, 93]. Nevertheless, as the scale, speed, and heterogeneity of contemporary datasets continue to intensify, foundational challenges remain. This necessitates continuous methodological innovation and rigorous evaluation of advancements within the evolving algorithmic landscape.

1.2 Key Concepts and Terminology

Meeting the analytical demands posed by high-dimensional and categorical data necessitates clear definitions of the core computational problems and methodological strategies that underpin modern practice [6, 13, 15, 28, 33, 38–40, 49, 52, 53, 57, 60–63, 66, 79, 84, 88, 90, 91, 93, 107, 108]. Among the fundamental primitives are nearest neighbor (NN) and k-nearest neighbor (kNN) search, which support similarity-based queries essential for clustering, classification, anomaly detection, and recommender systems. In high-dimensional settings, both exact and approximate NN algorithms are central, with ongoing advances in indexing and pruning techniques, as well as metric learning approaches, to safeguard nearest neighbor structures in the face of sparsity and noise.

Other core analytical tasks include:

Similarity and Range Search: These extend NN paradigms to return all objects within a specified distance or similarity threshold from a query. They are pivotal in data mining, information retrieval, and feature-based querying—especially in graph- or spatial-structured data.

Clustering: The process of partitioning data into groups that maximize intra-group similarity. Challenges intensify in high-dimensional contexts, where relevant features are often obscured by spurious or noisy information [61, 66, 84, 93].

Classification: Assigns category labels to data objects, typically in a supervised framework. The abundance of irrelevant or redundant features in high-dimensional spaces impedes both model accuracy and interpretability.

Statistical Testing: In high-dimensional settings, conventional statistical testing must contend with reduced statistical power and inflated type I/II error rates, due to the effects of multiple hypothesis testing and inter-feature dependencies [66].

Indexing: Refers to the construction of data structures—such as k-d trees, ball trees, cover trees, and emerging learned or adaptive indexes—that expedite various types of queries, even as dimensions proliferate [52, 57, 60, 88, 91].

Frequently, high-dimensional and categorical data analysis requires the interplay among these concepts. For instance, graph-based representations exploit both spatial and relational proximity, while spectral and consensus methods adapt clustering and similarity measures to enhance partition quality and retrieval robustness [6, 38, 40, 84]. Categorical data clustering, in particular, integrates specialized encoding schemes, variable selection, and consensus mechanisms to mitigate the effect of noise from less informative dimensions [6, 38, 52, 93]. Thus, the field employs a multifaceted toolbox, extending foundational concepts to address the distinct analytical challenges posed by complex, high-dimensional datasets.

1.3 Scope and Organization

This survey offers a comprehensive synthesis of recent advances in algorithmic, methodological, and system-level approaches for the analysis of high-dimensional and categorical data, with particular emphasis on elucidating the current state of the art and highlighting foundational challenges and opportunities [57, 70, 112]. The review begins with an in-depth analysis of major algorithmic paradigms, including classic and contemporary methods for NN and kNN search, range search, clustering, classification, and statistical testing, each examined through the lens of dimensionality, data heterogeneity, and categorical structure.

Subsequent sections explore indexing methodologies, covering both established data structures and newly emerging approaches such as learned, adaptive, and hybrid indexes, with a focus on computational efficiency, robustness, and adaptability to dynamic data workloads. Special attention is devoted to trends in data compression and representation learning, including advances in compressed computation, symbolic embedding techniques, and spectral models that facilitate scalable and meaningful analytics on massive datasets.

The survey further discusses ensemble and spectral methods, consensus and subspace clustering, and hybrid statistical–machine learning frameworks. Special emphasis is placed on recent techniques such as consensus spectral clustering and self-constrained spectral clustering [57, 112], which have demonstrated strong robustness in the face of high-dimensional noise and uninformative features, and advances in compressed computation that enable direct algorithmic operations on compressed data representations [70]. Each method is critically evaluated for its effectiveness in extracting meaningful structure and mitigating challenges such as dimensionality-induced noise accumulation.

Finally, the survey contextualizes these algorithmic and methodological advances within the broader landscape of practical system integration. It addresses open research questions and emerging trajectories, including dynamic and adaptive computation, interpretable modeling, and resilient, secure indexing strategies for high-dimensional and categorical data analysis. Through a critical engagement with the current literature across these dimensions, this survey aims to provide a foundational orientation for newcomers and a forward-looking roadmap for future research in this rapidly evolving field.

2 Clustering High-Dimensional, Categorical, and Mixed Data

Clustering high-dimensional, categorical, and mixed-type data presents unique challenges due to the nature and complexity of the data involved. This section reviews the key algorithmic paradigms, summarizes their main features, highlights existing research gaps, and introduces frameworks that distinguish this survey from previous works. The taxonomy herein is informed by developments not only from core computer science and artificial intelligence but also from related areas, such as applied statistics and domain-specific data science, ensuring broader relevance and a comprehensive perspective.

Table 2 provides an overview of central paradigms for clustering data with challenging characteristics, and is positioned here to orient readers before deeper technical discussions. Our taxonomy integrates not only traditional algorithm classes but also reflects methodological advances from adjacent fields such as statistical learning, domain engineering, and emerging interdisciplinary applications.

Despite the progress outlined in Table 2, several actionable research gaps persist within this domain, categorized by data characteristics and application context:

For high-dimensional data, the curse of dimensionality impacts both cluster discernibility and algorithmic efficiency. Existing subspace and projected clustering paradigms remain limited in scalability to truly large feature spaces and often require manual or heuristic selection of relevant dimensions. Even methods utilizing adaptive or automatic feature selection are constrained by computational costs as dimensionality grows. There is a need for algorithms that can automatically and adaptively identify informative subspaces in ultra-high-dimensional settings, potentially leveraging recent advances in automatic feature selection and representation learning. However, current approaches often encounter significant challenges in interpretability and efficient search within complex feature spaces.

When clustering categorical data, current algorithms rely on specialized distance metrics and information-theoretic criteria. Nonetheless, they often struggle with rare categories, missing data, and the integration of domain-specific constraints. For example, some techniques are sensitive to noise or skewed distributions, limiting robustness in practical applications. Addressing these issues requires the development of robust categorical similarity measures that can naturally handle noise, rare events, and missing values, possibly with the integration of domain knowledge, while ensuring scalability to large datasets.

Mixed-type data clustering remains an open challenge due to the differing statistical scales of features and the difficulty of balancing continuous and categorical attributes. Most current approaches use simple concatenation techniques or distance-weighting heuristics, which may not perform optimally in heterogeneous settings. These ad-hoc solutions can introduce bias if attribute types are not weighted appropriately, and might obscure intrinsic data structures. Actionable opportunities exist for designing unified, principled objective functions or embeddings that capture the joint structure of mixed features in a more theoretically grounded way.

This survey distinguishes itself from prior reviews by systematically categorizing clustering techniques according to data type compatibility, scalability, and robustness (as summarized in Table 2), introducing a unified framework for evaluating algorithm suitability under mixed real-world constraints, and explicitly integrating foundational insights from interdisciplinary sources. This taxonomy and analytical mapping provide researchers and practitioners with a clearer guide for method selection and highlight novel directions for algorithmic innovation.

In summary, while significant advances have been made across different paradigms of clustering high-dimensional, categorical, and mixed data, substantial gaps remain in terms of scalability, robustness, and holistic support for mixed-type attributes. Continued research—particularly in adaptive dimensionality reduction, domain-aware similarity measures, and unified objective formulations—will be crucial to addressing these open challenges.

Table 2: Summary of Major Clustering Paradigms for High-Dimensional, Categorical, and Mixed Data

Paradigm	Data Type(s) Supported	Strengths	Limitations
Subspace/Projected Clustering	High-dimensional	Discovers meaningful clusters in relevant feature subsets	Sensitive to subspace selection, scalability concerns
Categorical Data Clustering	Categorical	Tailored similarity/distance functions (e.g., k-modes)	May not generalize across domains
Mixed Data Clustering	Mixed (numeric + categorical)	Integrates heterogeneous data types (e.g., k-prototypes)	Balancing influence of each type is challenging
Spectral Methods	Mostly numeric, extensible	Good for non-convex structures, adaptable to high dimensions	Computational cost, tuning parameters
Model-based Clustering	All types (with extensions)	Probabilistic framework, flexible models	Scalability, model selection complexity

2.1 Challenges in Clustering High-Dimensional and Categorical Data

Clustering high-dimensional datasets—encompassing continuous, categorical, or mixed types—entails a suite of formidable statistical and computational challenges. Foremost is the phenomenon of noise accumulation: as dimensionality escalates, the distinction between informative and non-informative features blurs, thereby reducing the reliability of traditional similarity measures. This complication is particularly acute in domains like gene expression analysis and text mining, where only a minority of observed variables substantially contribute to cluster separability. Consequently, uninformative features may give rise to diffuse or spurious clusters, especially under conditions of stochastic or adversarial noise [57].

Categorical attributes further amplify these obstacles due to sparsity and high cardinality, making it difficult to define robust distance or similarity metrics. Such issues undermine both distance-based and model-based clustering algorithms [57], weakening their effectiveness and interpretability in real-world applications.

2.2 Ensemble Subspace and Consensus Spectral Clustering

To alleviate the curse of dimensionality and limitations of single-view clustering, ensemble subspace approaches and consensus spectral clustering have emerged as prominent strategies. These techniques typically employ feature transformation—such as one-hot encoding for categorical variables—followed by procedures like random projection or subspace sampling to generate diverse, information-rich feature subsets [57, 66]. Through subsampling, clusters may be constructed using only the most relevant dimensions, thereby mitigating the influence of noisy or irrelevant variables.

The ensemble process involves aggregating the results from multiple subspace clusterings, often quantified via co-association matrices and consensus functions (e.g., majority voting), to capitalize on the collective insights of partially independent clusterings [4, 55]. Parallel and distributed computation paradigms are frequently leveraged to ensure scalability.

A notable advancement is the incorporation of feature reweighting, with data-driven measures guiding the assignment of greater importance to features or subspaces associated with high signal-to-noise ratios. This renders ensemble clustering methods not only more robust to noise but also adaptive to heterogeneous feature landscapes [29, 57]. Theoretical analyses demonstrate that these methods achieve statistical consistency and minimax-optimal error rates even as the fraction of truly informative features diminishes—a scenario common in omics and text mining tasks [57, 66]. Empirical results corroborate these theoretical gains, with ensemble and consensus spectral approaches often outperforming baseline methods in genomics and unstructured text clustering tasks [57].

Despite their advantages, consensus-based frameworks show reduced efficacy when data exhibits complex feature dependencies (e.g., spatial, temporal, or network structures) or when dealing with genuinely mixed-type attributes, situations where standard one-hot or projection-based strategies fail to capture generative processes [57]. Furthermore, algorithmic complexity—though mitigated through parallelization—can pose practical limitations in very high-dimensional or resource-constrained environments [57].

2.3 Spectral Clustering and Self-Constrained Extensions

To reinforce the survey's core objective of bridging clustering algorithms with efficient indexing and search in high-dimensional and complex data, this subsection synthesizes recent trends in integrating spectral clustering with index-aware or constraint-driven methodologies. Our focus remains on clarifying how advances in clustering can directly enhance large-scale data retrieval and organization, especially for heterogeneous and graph-structured datasets.

Spectral clustering has become a widely adopted method for high-dimensional and categorical datasets, leveraging the global organizational structure encoded within the eigenspaces of similarity or Laplacian matrices [91, 112]. This framework eschews direct modeling of cluster-wise densities, instead utilizing geometric relationships in a transformed, lower-dimensional embedding.

Recent methodological advancements include self-constrained spectral clustering, wherein the canonical objective is augmented with explicit pairwise or label-based constraints. These constraints encode prior knowledge or enforce desired partition properties, implemented through iterative optimization and alternating update rules. This ensures convergence to partitions that honor both intrinsic data similarities and extrinsic supervisory information, as demonstrated in recent work [112]. Bai et al. [112] introduce a theoretical framework and corresponding update strategies, highlighting extensibility and practical integration of supervision into spectral approaches (2023).

Self-constrained extensions are particularly advantageous in semi-supervised contexts and in scenarios requiring alignment with spatial or relational structures—for example, integrating clustering results with spatial databases or graph-indexed data pipelines. This theme resonates with state-of-the-art indexing in graph data, where learning-based similarity search frameworks [91] (Wang et al., 2023) increasingly enable scalable containment queries by embedding structural and label information, thus directly bridging clustering-derived representations with efficient index construction.

Nevertheless, spectral clustering remains sensitive to affinity matrix construction and parameter tuning, necessitating careful preprocessing and validation to ensure reliability [91, 112]. As the field moves forward, synthesizing trends indicate growing convergence between clustering and indexing paradigms—particularly through embedding-based unification—as well as promising directions for future research in robust, application-aligned solutions for high-dimensional and heterogeneous data environments.

2.4 Alternative Clustering Methodologies

As part of this survey's core objective to provide an integrated, up-to-date synthesis of clustering paradigms for high-dimensional and heterogeneous data, this section surveys several foundational and emerging clustering frameworks beyond ensemble and spectral methods. Our aim is to clarify trends, highlight distinguishing innovations, and reinforce connections to state-of-the-art indexing strategies, guiding practitioners in matching methodologies to diverse analytical objectives.

The landscape of clustering methods for high-dimensional and mixed-type data encompasses several paradigms:

Hierarchical Clustering: Hierarchical approaches—both agglomerative and divisive—are valued for their interpretability through dendrograms and the ability to resolve clusters at varying levels of granularity. However, as demonstrated by [82, 88], their scalability in high-dimensional data can be problematic, and performance is closely linked to parameter selection and linkage criteria. Recent work, such as the adaptive frameworks in [19, 99], leverages innovations like local, parameter-free cut-off distances or mass-based merging to automatically generate robust cluster solutions, boost resilience to noise, and mitigate over-merging—advances especially pertinent after 2022.

Bayesian and Model-Based Approaches: Mixture models and their extensions (including mixed membership and tensor-normal mixtures) enable probabilistic cluster assignments and uncertainty quantification. Escalating dimensionality and data heterogeneity have prompted advances in scalable inference, including penalized coordinate descent algorithms that enhance variable selection and estimation accuracy [36], as well as spectral approximations for ultrahigh-dimensional or grouped scenarios [62]. For example, in transcriptomic and microbiome clusters (2025), SCAD penalization outperforms classical LASSO, promoting near-zero false positives [36]. Nonetheless, model-based techniques remain sensitive to compositionality, overparameterization, and complex dependency structures.

Tensor Clustering: For inherently multiway data (such as those in omics or neuroimaging), tensor clustering via tensor normal mixture models (TNMM) offers parsimonious modeling and interpretable clustering, scaling sub-linearly with respect to exponential increases in dimensionality [59]. Penalized approaches incorporating sparsity constraints (e.g., lasso/SCAD) and tailored EM strategies ensure robustness, though implementation depends on appropriate separability assumptions for covariance and careful model selection.

Robust and Hybrid Methods: Integration of diverse clustering objectives, such as density- and partition-based criteria or hybrid probabilistic and distance-based frameworks, allows adaptation to challenging settings (non-globular clusters, heterogeneous features, compositional data). Fully autonomous, parameter-free clustering frameworks [99] and distance-based Bayesian models leveraging both cohesion and repulsion [67] exemplify significant advances in robust, flexible cluster identification. Such methods, several released after 2020, highlight a cross-cutting shift toward less human intervention and greater applicability to real-world, complex datasets.

Deep Clustering Paradigms: Rapid progress in deep learning has transformed clustering, especially for high-dimensional, weakly structured, or task-specific data. Deep clustering methods [73, 111] exploit the simultaneous optimization of representation learning and clustering, demonstrating greater noise resilience, insensitivity to initialization, and improved handling of overlapping or abstract cluster geometries. End-to-end frameworks are now common for modalities like images, text, time series, and graphs, with recent surveys (2022–2024) presenting new taxonomies and benchmarks [73, 111]. Yet, challenges persist—notably, explainability, hyperparameter tuning, and reliable transfer across domains.

In summary, these alternative methodologies each embody unique trade-offs between scalability, robustness, interpretability, and generality. No single approach dominates across all high-dimensional, categorical, or mixed data contexts. Instead, as consistently emphasized in both classic and post-2020 studies [4, 19, 36, 57, 62, 111, 112], the optimal method depends critically on dataset structure, the nature of feature distributions, and the specific analytical or indexing objectives at hand.

Table 3 below synthesizes principal clustering paradigms, contrasting their advantages and limitations, and unifies the section by reflecting the taxonomic logic underpinning this survey. This comparative, trend-driven perspective aims to distinguish our synthesis from prior reviews and provide a practical reference for users facing the clustering/indexing interface in high-dimensional heterogeneous data.

In direct alignment with this review's aims, we next transition to the relationship between clustering and similarity search (indexing), highlighting critical interfaces and open research opportunities for joint indexing-clustering frameworks well-suited for increasingly complex, high-dimensional and heterogeneous data analyses.

2.5 Cluster Validation Metrics and Benchmarking

As a core objective, this survey aims to critically synthesize and clarify the landscape of cluster validation metrics and benchmarking protocols, with a distinct emphasis on high-dimensional, categorical, and mixed-type data. We seek to present not only established practices but also recent trends and emerging challenges, fostering a deeper understanding of the interplay between clustering methods and their evaluation, and positioning our survey as a guide for both comparative assessment and methodological innovation.

Robust evaluation of clustering results in high-dimensional and mixed-type contexts relies upon comprehensive validation and benchmarking metrics. These include:

Table 3: Comparison of Principal Clustering Paradigms for High-Dimensional, Categorical, and Mixed Data

Methodology	Primary Advantages	Key Limitations
Ensemble Subspace/Consensus Spectral	Robustness to noise and irrelevant features; scalable via parallelization	Complexity in affinity aggregation; reduced efficacy for data with intricate dependencies or mixed types
Spectral (Standard/Self-Constrained)	Captures global structure; accommodates constraints/prior knowledge	Sensitive to affinity matrix and parameter selection; scaling may be nontrivial
Hierarchical	Interpretability; flexible resolution	Parameter sensitivity; scalability challenges in high dimensions
Bayesian/Model-Based	Probabilistic inference; uncertainty quantification	Overparameterization; bottlenecks in ultrahigh dimensions
Tensor Clustering	Exploits multiway data; improved parsimony	Requires structured data; complex implementation
Deep Clustering	End-to-end learning; resilience to noise/overlap	Interpretability; hyperparameter tuning; domain transferability
Robust/Hybrid	Adaptive to diverse data; handles irregular shapes	Model selection complexity; computational overhead

External Indices: Metrics such as Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Cohen's Kappa facilitate quantitative comparison against known ground-truth labels, enhancing comparability across algorithms when gold standards are available [4, 6, 8, 11, 17, 21, 31, 39, 42, 43, 50, 56, 62, 66, 73, 84–86, 88, 100–102, 104, 110, 111]. For instance, ARI and NMI are frequently used to provide chance-adjusted and information-theoretic perspectives on agreement, respectively, while Cohen's Kappa quantifies agreement beyond random expectation. It is important to note that, as highlighted in recent studies (e.g., high-dimensional disease classification [6, 8, 88]), external metrics can be influenced by class imbalance and label noise, emphasizing the need for careful metric selection and interpretation [8, 17, 88].

Internal Indices: Metrics such as the Silhouette coefficient, Davies-Bouldin index, Dunn index, accuracy, AUROC, and F1-score provide model-agnostic assessments of cluster cohesion, separation, and overall quality, independent of external references [1, 4, 6, 11, 17, 19, 21, 36, 39, 42, 43, 47, 50, 56, 62, 66, 73, 84, 86, 88, 95, 100–102, 104, 110]. While classical internal indices such as silhouette and Dunn scores are widespread, their practical limitations—including sensitivity to noise, cluster size, and shape—have been highlighted in recent comparative works [1, 73]. These limitations are particularly evident in complex real-world data, motivating the development of alternative measures. For example, Wiroonsri [95] proposes a novel correlation-based index that can reveal multiple local optima when selecting the number of clusters, partially addressing the inadequacy of single-optimum indices in complex scenarios. In high-dimensional genomics, metrics such as F1-score, AUROC, and Cohen's Kappa are critical for performance benchmarking due to pervasive class imbalance and small-sample effects [6, 8, 50, 66, 88].

Multimodality-Based and Modern Indices: Recently, measures such as Dip and Silverman's tests have gained momentum, owing to their robustness and improved discrimination of true clusterability across challenging data scenarios [1, 11, 19, 95, 101, 102, 111]. These multimodality-based metrics are recommended as general-purpose tools for distinguishing clusterable from unclusterable data [1]. However, challenges such as handling outliers, chaining, and very small clusters remain [1], highlighting directions for future research.

Importantly, many classical and even recent metrics exhibit sensitivity to various artifacts (noise, imbalance, parameter selection, etc.), as established in experimental overviews and large-scale benchmarking studies spanning domains such as time-series analysis, genomics, and point cloud indexing [1, 4, 8, 17, 36, 43, 47, 57, 66, 73, 84, 88, 100, 102, 104]. The curse of dimensionality, dominance of uninformative features, instability of benchmarks due to preprocessing bias, and dataset selection aggravate the reproducibility and scalability of performance claims in contemporary settings [17, 57, 73].

To ensure trustworthy and meaningful analysis, current best practices demand the joint and transparent use of both internal and external validation indices, meticulous dataset curation, and open, reproducible benchmarking pipelines [4, 8, 36, 43, 57, 62, 66, 84, 111]. For example, the recent HighDimMixedModels.jl framework (2025) [36] and absolute cluster validity protocols [43] illustrate the growing adoption of such transparency and rigor. The continued development of algorithm-specific and data-type-specific benchmarking frameworks, particularly for high-dimensional, categorical, and mixed data, is now recognized as essential for progress [57, 66, 73].

A further unique aspect of the cluster validation and benchmarking landscape addressed in this survey is the synthesis of recent advances in clustering and indexing, especially for high-dimensional or heterogeneous data types. As evidenced in recent landmark works on deep clustering (2022–2024) [73, 101, 111], similarity search [11, 31, 84, 100, 110], and biomedical applications [4, 36, 43, 57, 62, 66, 104], robust indexing and efficient search algorithms are increasingly intertwined with cluster validation, both in theory and practice. Integrating clustering and indexing is a rising research direction, seeking improved scalability, interpretability, and domain adaptation.

Finally, recent surveys and experimental reports call for the development of standardized reporting frameworks, comprehensive evaluation protocols tailored to specific data modalities, and well-documented repositories for reproducible research [8, 17, 43, 57, 73]. Taxonomically, this review distinguishes itself by explicitly categorizing methods and indices by their applicability to data type, model class, and evaluation purpose, providing a cross-sectional synthesis that is lacking in prior overviews.

Collectively, these methodological innovations and validation frameworks delineate both the considerable progress and enduring open challenges in the clustering of high-dimensional, categorical, and mixed data. Ongoing advances in interpretability, scalability, and rigorous benchmarking—with clearer taxonomies and deeper integration of clustering and indexing—remain essential for the development of effective clustering methodologies and their translation to a broad array of scientific and practical applications.

3 Index Structures and Data Representations

At the heart of this survey is a critical examination of how modern index structures and data representations underpin large-scale, AI-driven clustering and retrieval tasks. This section restates and reinforces the core objective: to synthesize the landscape of indexing and representation methods that enable efficiency, scalability, and adaptability in the face of rapidly expanding and diversifying datasets. Our goal is to clarify the roles these paradigms play in current analytics workflows, especially as the boundaries between clustering and indexing become increasingly intertwined.

We begin by establishing a clear organizational framework, distinguishing the principal classes of index structures—such as tree-based, hash-based, and learned indices—alongside emergent neural and hybrid methods. This taxonomy is designed to guide readers through the breadth of techniques while offering coherence that distinguishes our review from prior surveys.

Transitions between clustering and indexing are made explicit: modern clustering algorithms often depend on scalable indexing schemes for both initialization and dynamic updates, while advanced indexing frequently leverages cluster assignments or structure to optimize access in heterogeneous, high-dimensional spaces. Emphasizing their synergy, we highlight recent trends where clustering and indexing are not independent stages but mutually reinforcing components of end-to-end systems.

Throughout, we underscore key open research directions, particularly the challenge of integrating clustering and indexing in scenarios involving high-dimensional or highly heterogeneous data. This includes the adaptation of data representations to emerging application domains and analytics tasks, which frequently impose novel requirements on scalability and precision.

To orient readers and unify the section, we regularly return to the survey's objectives: to disentangle, organize, and critically synthesize the evolving field of index structures and data representations as they relate to large-scale AI analytics. We focus not merely on enumerating available techniques, but on surfacing cross-cutting themes, recent advances, and future opportunities at the nexus of clustering and indexing for modern data science.

3.1 Traditional Index Structures

Classic spatial and multidimensional index structures—including R-trees, k-d trees, Quadtrees, Grid indexes, Inverted Indexes, and Column Stores—have long been foundational in database systems for managing multi-attribute and spatial queries. R-trees and their variants are optimal for bounding spatial objects and facilitating efficient range and topological searches, while k-d trees and Quadtrees naturally partition multidimensional or spatial data for point queries and region decompositions. Grid and inverted indexes enable rapid filtering and set operations, with inverted indexes excelling particularly in text and categorical data retrieval. Column stores further separate data by attribute, supporting high compression and swift analytical scans. Despite their versatility and widespread adoption, these structures present significant trade-offs: while highly effective for low to moderate dimensionality, scaling to higher dimensions often incurs substantial costs in storage, maintenance, and query performance, particularly as datasets increase in both volume and complexity [25, 26]. Moreover, in high-throughput or real-time environments, the continual maintenance and updating of indexes can amplify these costs, leading to bottlenecks that undermine their intended efficiency.

As shown in Table 4, the effectiveness and limitations of these index structures are tightly coupled to the underlying data characteristics and query requirements.

3.2 Limitations for High-Dimensional and Categorical Data

Despite their general flexibility, traditional indexes typically underperform as dimensionality grows—a phenomenon often described as the "curse of dimensionality." For instance, R-trees experience increased node overlap and size, resulting in excessive I/O during searches. Similarly, k-d trees become imbalanced with high-dimensional inputs, suffering from sharply reduced partitioning efficiency [25, 26]. Beyond numerical dimensions, most classical indexes struggle to integrate categorical and mixed-type attributes alongside spatial or numerical information; supporting such heterogeneous data often necessitates complex, task-specific adaptations that ultimately compromise generality and performance. This persistent set of limitations has catalyzed the search for unified, extensible indexing frameworks that can accommodate both high-dimensional and heterogeneous data types—achieving flexible indexing and querying without incurring prohibitive design or operational complexity [25, 26].

3.3 Modern Memory-Efficient and Compressed Indexes

The explosive growth of data volumes, coupled with physical memory bandwidth limitations, has driven significant advances in compressed and succinct indexing structures. For example, q-gram trees for graph similarity search show the feasibility of highly space-efficient in-memory indexes, requiring as little as 5–15% of the memory needed by conventional methods while providing query acceleration, and scaling to datasets with up to 25 million graphs [23]. These structures combine probabilistic and deterministic methods by using succinct data structures and hybrid encodings, augmented by global and local filters to efficiently localize candidate sets and accelerate search [23].

Probabilistic data structures such as state-of-the-art Bloom filters and Cuckoo filters have further advanced memory efficiency, attaining near-optimal space bounds and constant expected lookup time, with configurable trade-offs among false positive rates, insertion, and deletion [32, 83, 103]. Recent improvements in Cuckoo filters, notably the introduction of signed-offset addressing and overlapping window layouts, have eliminated classic power-of-two bucket size restrictions, significantly reducing space overhead and maximizing achievable load. These advances make windowed Cuckoo filters stand out as the most memory-efficient option for online-insertion-capable filters across practical false positive rates, achieving fast queries and high adaptability for large-scale analytics and scientific data workloads [83].

Trie-based indexes, including the Height-Optimized Trie (HOT) and Adaptive Radix Tree (ART), optimize space by dynamically adjusting node sizes and fan-outs, thereby balancing lookup performance, memory consumption, and update speed—qualities that are critical for real-time in-memory database applications [14, 96]. In particular, HOT leverages dynamic node height adjustment to minimize index height and storage, while ART can further reduce construction and update overhead by integrating techniques such as database cracking, allowing incremental index refinement during query processing [14, 96].

Table 4: Comparison of traditional index structures by usage and limitations

Index Type	Primary Use	Dimensionality Support	Key Limitations
R-tree	Spatial objects, range queries	Low/medium	Degrades with high dimensionality
k-d tree	Point queries, region search	Low/medium	Poor balance in high-dim spaces
Quadtree	2D/3D spatial partitioning	Low	Scalability issues
Grid Index	Numeric filtering	Medium	Inefficient for skewed data
Inverted Index	Text/search, categorical	N/A	Poor for numeric/spatial data
Column Store	Analytical scans	N/A	Write overhead, schema constraints

Suffix-based and run-length encoded indexes excel at handling highly repetitive data, such as web archives or genomic sequences. Approaches leveraging the run-length encoded Burrows–Wheeler Transform (RLBWT) and compressed suffix arrays achieve asymptotically optimal space representation on repetitive inputs [76]. These structures support efficient substring search and factorization directly over compressed representations, as demonstrated by algorithms that can compute LZ77 factorizations or self-indexes using space and working memory as low as 1% of the original dataset, outperforming classical entropy-compressed and pointer-based alternatives by orders of magnitude [32, 76]. However, while static approaches yield optimal space and good performance, supporting dynamic updates in compressed indexes remains challenging in practice and often introduces significant overheads [14, 70, 76].

Despite these major strides, the widespread deployment of memory-efficient indexes still involves critical trade-offs. Maximizing compression may inhibit dynamic operations or slow down update handling. The complex engineering required to design succinct structures that efficiently support range, similarity, and set queries—especially for dynamic scenarios—remains a significant challenge [70, 96]. Consequently, ongoing research seeks to optimize the balance among compression, adaptability, and high query efficiency, aiming to meet the evolving needs for scalable and performant indexing in data-intensive domains.

3.4 Compressed Computation Paradigm

With uncompressed data volumes increasingly surpassing hardware capabilities, a shift towards the "compressed computation" paradigm has emerged as a necessity. Here, compression is no longer merely a storage or transmission optimization but forms the basis for direct in-memory computation. The result is not only minimized storage and I/O but also a fundamentally reduced working set during active processing [70].

Critical advances include data structures and algorithms operating natively upon compressed representations—such as run-length compressed suffix arrays, compressed tries, or space-efficient factorization structures—thus circumventing expensive decompression cycles [14, 23, 32, 70, 76, 83, 96, 103]. For example, the direct transformation from RLBWT to LZ77 factorization enables self-indexing in extremely limited space, as RLBWT-based approaches can exploit highly repetitive data to deliver both asymptotically optimal space and practical memory reductions [76]. These algorithms have demonstrated up to two to three orders of magnitude improvements in working space over previous entropy-compressed and suffix-array based alternatives on datasets such as Wikipedia and genomics repositories, though dynamic maintenance remains a practical performance challenge. Likewise, compressed suffix trees enriched with witness structures provide unified, online computation of Lempel-Ziv factorizations, achieving both linear time and sublinear space efficiency—opening the path for real-time analytics on massive textual data [32]. Succinct dictionaries [103] and improved Cuckoo filters [83] further contribute to this trend by supporting set membership and filtering operations with extremely compact representations, controllable false positive rates, and minimal query overhead, outperforming state-of-the-art alternatives in both space and look-up speed for large-scale applications such as genomics and networking.

Beyond strings and text, compressed computation has extended to more complex structures. For example, the MSQ-Index [23] leverages succinct q-gram trees and hybrid encoding for graph similarity search, requiring an order of magnitude less memory than previous graph editing indexes while retaining fast query response and the ability to scale to massive chemical graph collections. Recent studies on efficient in-memory indexes such as the Adaptive Radix Tree (ART) have also explored dynamic, incremental index construction—so-called database cracking—reducing build overhead and yielding more adaptable structures for streaming or real-time workloads [96].

Despite these advances, significant challenges persist. Indexes that operate on compressed data must mediate among conflicting objectives: compression ratio, query latency, and support for updates. For instance, partially or fully compressed repositories raise open questions for similarity and range search, as classic indexes typically presuppose uncompressed or partially indexed data [70]. The evolution of adaptive and query-aware compressed indexes—capable of dynamically alternating between compressed and uncompressed representations—constitutes a core frontier for ongoing research.

3.5 Learned, Neural, and Adaptive Indexes

This subsection surveys the rapid evolution of learned and adaptive index structures, highlighting their goals, strengths, and emerging limitations. The core objective is to explore how these new indexing paradigms leverage machine learning and system adaptivity to address the growing demands of high-dimensional, large-scale, and heterogeneous data—reaching beyond the bounds of classical and compressed indexes.

Recognizing the limitations of both classical and compressed index approaches for managing high-dimensional, evolving, or heterogeneous datasets, a new generation of learned and adaptive indexes has emerged. Leveraging machine learning, these indexes reinterpret data access as a form of prediction, employing models that estimate the location or probability of a record within the data structure.

Spline-based learned indexes, such as LiLIS, apply error-bounded piecewise linear models to approximate mappings within sorted or spatially partitioned data. These models provide constant-time ($O(1)$) lookup overhead and have been shown, in distributed big data frameworks, to yield dramatic speedups compared to traditional spatial indexes. In particular, LiLIS operates by integrating error-bounded spline-based learned indexes on per-partition data using flexible, spatially-aware partitioning strategies such as R-tree, Quadtree, k-d tree, and grid-based schemes. The mapping of high-dimensional locations into one dimension (e.g., via Z-order curves) enables efficient point, range, k-nearest neighbor, and join queries. Experimental evaluation demonstrates that, for distributed spatial workloads (e.g., on Apache Spark), LiLIS significantly outperforms conventional indexes such as Sedona-R-tree, achieving up to orders-of-magnitude faster query speeds and 1.5–2× faster index build times. For example, on real and large synthetic datasets, LiLIS achieves point query times of 82.59 ms and range query times of 468.64 ms, whereas competing methods are much slower, especially for join and kNN queries. However, these benefits can be sensitive to partitioning choices (R-tree partitioners excel generally, while k-d/Quadtree best support join operations) and to query distribution skew. The reliance on model training also introduces additional computational overhead, particularly for massive datasets, and adapting to more complex queries remains an open concern [25, 60].

Table 18 highlights empirical query time comparisons, illustrating the substantial efficiency advantage of learned spatial indexes in distributed scenarios.

Model-driven indexing further encompasses approaches in which partitioning mechanisms—such as R-tree, k-d tree, and Z-order curves—are tightly coupled with predictive mappings offered by the learned model. This coupling achieves notable reductions in both index construction and query evaluation times, though it necessitates careful attention to partitioner selection, sensitivity to workload skew, and the costs associated with ongoing model retraining and integration into distributed dataflow platforms [25, 60]. While these models offer significant speedups and scalability, notable weaknesses remain regarding guaranteed error bounds in high dimensions, efficient and consistent retraining for evolving datasets, and maintaining predictable performance under adversarial or highly dynamic workloads.

Frameworks for index selection are beginning to adopt online learning paradigms, notably multi-armed bandit and reinforcement learning strategies. For example, recent multi-armed bandit-based approaches remove dependency on DBA intervention and unreliable cost models by directly and consecutively exploring candidate indexes based on observed query performance. These frameworks have demonstrated strong empirical speedups: up to 75% on highly dynamic or shifting workloads and up to 51–59% on hybrid transactional/analytical (HTAP) settings compared to conventional or deep RL-based approaches, while ensuring convergence to (near-)optimal policies [74]. Such methodologies enable continual online adaptation to rapidly fluctuating workload characteristics and transform auto-tuning from a static optimization task into an ongoing learning process. However, they may incur higher exploration costs initially, require robust reward signal estimation, and sometimes struggle with abrupt, unpredictable workload shifts—remaining practical limitations for production deployment.

At a broader system level, frameworks such as annotative indexing aim to subsume both traditional and learned index paradigms within a common, highly modular architecture. Annotative indexes generalize and unify inverted, columnar, and object indexes under a dynamic, transactional model that supports efficient ACID transactions and concurrency for both reads and writes. These designs facilitate expressive query processing over structured, semi-structured (e.g., JSON), heterogeneous, and graph-based data. Annotative indexing is further distinguished by its support for lazy transformation, hybrid neural and sparse retrieval, composability for retrieval augmented generation (RAG), and compatibility with structured and unstructured query modalities. The architecture is capable of scaling to hundreds of concurrent transactional clients and supports advanced operations, including entity lookup, knowledge graph queries, and neural search over both text and other datatypes, all while ensuring modularity and extensibility [26]. Practical challenges include robust transaction isolation under heavy concurrency, efficient garbage collection, full distributed scaling, and support for dense vector and large graph workloads—open issues highlighted for future research and system development.

Nonetheless, these modern indexing strategies introduce their own unresolved challenges. Theoretical concerns include maintaining bounded errors or guarantees in high-dimensional predictive indexing, and robustness of continuous model retraining. Practical issues persist around ensuring safe concurrent access, effective distributed scaling, adversarial resilience and security, garbage collection, and seamless integration into complex, distributed data systems. Emerging research directions span GPU-accelerated retraining for threshold workloads, exploration of hybrid designs combining learned and classical index primitives, and development of transactional guarantees for broad query types over both structured and unstructured data, with a particular emphasis on scaling and extensibility in heterogeneous deployment environments [25, 26, 60].

Synthesizing across index types, the progression from classical and compressed indexes through adaptive, learned, and fully modular frameworks underscores a continual balancing act between efficiency, adaptability, scalability, and practical usability. As catalogued in recent surveys [60], future convergence points are likely to center on hybrid designs—blending strengths of model-driven and classical structures, delivering theoretical guarantees, and supporting broad query capabilities across structured, semi-structured, and unstructured modalities. The overarching contemporary challenge is to enable databases and data platforms to keep pace with the ever-expanding scale and heterogeneity of real-world data, driving ongoing research at the intersection of algorithm design, machine learning, and large-scale system engineering.

4 Similarity, Range Search, and Graph Querying

This section provides an overview of core techniques and challenges in similarity search, range search, and graph querying, which are

Table 5: Query times (ms) for spatial queries using LiLIS and Sedona-RK on large datasets [25]. "Much slower" indicates high latency or infeasibility for the corresponding method.

Method	Point	Range	kNN	Join
LiLIS-K	82.59	468.64	650.2	228581
Sedona-RK	much slower	much slower	790993	much slower

foundational to numerous modern AI and data systems. The primary objective here is to outline the methodological landscape, critically summarize their characteristic strengths and weaknesses, and synthesize developments across classical and contemporary approaches. Explicit connections to the overarching goals of effective, scalable querying for large and complex datasets are highlighted throughout.

4.1 Similarity Search

Similarity search addresses the challenge of identifying data items most similar to a given query under a defined similarity or distance measure. Methods typically fall into two main families: exact and approximate approaches.

Exact methods, such as exhaustive linear scan and classical tree-based structures (e.g., KD-trees), guarantee retrieval accuracy but often suffer from poor scalability in high-dimensional or large datasets. Their computational and storage complexity grows rapidly with data volume and dimension, which limits practicality in modern applications.

Approximate methods, including hashing-based techniques (such as Locality-Sensitive Hashing), product quantization, and graph-based nearest neighbor algorithms have emerged to mitigate these limitations. While they offer significant improvements in speed and resource usage, their accuracy may degrade, and parameter tuning is nontrivial. Notably, performance can fluctuate depending on data distribution, query pattern, and hash function characteristics.

Both exact and approximate categories face challenges in adapting to dynamic datasets, heterogeneous similarity measures, and providing consistent latency. Addressing such practical weaknesses remains a central objective in current research, with ongoing efforts to balance efficiency, accuracy, and adaptability.

4.2 Range Search

Range search involves retrieving all items within a given distance (or similarity threshold) from a query element. Tree-structured indices, like R-trees and ball trees, are widely used to organize and partition search spaces for low- to moderate-dimensional data.

The primary limitation of these classical approaches is their sensitivity to the "curse of dimensionality" and declining pruning power as dimension increases. For high-dimensional data, space-filling curves and partitioning heuristics offer partial remedies but introduce additional complexity and often sacrifice search completeness.

More recent approximate range search mechanisms leverage hashing, learned partitioning, and hybrid indexing. While providing improved scalability, these methods may omit true positives especially when threshold boundaries are near ambiguous regions, highlighting a tradeoff between recall and performance that remains an open research direction.

Enhancing the robustness of range search for diverse data distributions, broader distance functions, and streaming settings is crucial for further progress.

4.3 Graph Querying

Graph querying generalizes pattern or substructure search over graph-structured data, motivated by applications ranging from social networks to knowledge graphs. Techniques encompass subgraph isomorphism, neighborhood matching, and semantic querying frameworks.

Classical algorithms for subgraph matching, although precise, are computationally intractable for large graphs due to inherent NP-completeness. Heuristic and index-based solutions attempt to improve scalability, but often must balance query expressivity, index construction overhead, and update cost.

Graph embeddings and learned representations have more recently enabled efficient approximate querying at the expense of interpretability and semantic fidelity. Key challenges here involve supporting rich queries, coping with rapid data evolution, and ensuring robustness in the presence of noise and incomplete information.

4.4 Synthesis and Outlook

The evolution from exact, index-based approaches to approximate and learned methods reflects an underlying drive to balance efficiency, scalability, and accuracy in diverse querying scenarios. While classical methods provide clear guarantees and conceptual simplicity, state-of-the-art approaches adapt flexibly to modern data regimes at the price of more complex tradeoffs and parameter dependencies.

Future work is expected to further close the gap between structural rigor and practical effectiveness, synthesizing graph, similarity, and range search paradigms into unified, adaptive frameworks. Continued attention to empirical weaknesses and integration of new taxonomies will be critical for advancing both theoretical insight and real-world utility.

4.5 Space-Partitioning Indexes for Query Processing

Space-partitioning indexes play a foundational role in efficient distance, similarity, and range query processing over both spatial and non-spatial datasets. These structures—including grid files, k-d trees, R-trees, spatial hashing, and, more recently, learned and ensemble-based indexes—enable rapid pruning of the search space by hierarchically or adaptively aggregating data into regions with shared characteristics. This results in significant reductions in computational redundancy during query evaluation. Notably, grid-based methods often surpass tree-based counterparts in performance when appropriate partition strategies are adopted, particularly for point, range, and join queries, due to superior data linearization and lower index traversal overheads [54, 94]. The introduction of machine-learned indexing and hybrid approaches has further advanced performance, with learned indexes employing regression models and space-filling curves to efficiently predict object positions and minimize lookup times. These techniques are particularly effective for high-dimensional or irregularly distributed datasets [18, 25, 71, 94].

Classical methods, however, encounter scalability barriers in contexts characterized by large-scale and highly repetitive datasets. Traditional inverted indexes and spatial structures often suffer from inefficiencies in both indexing and memory footprint [21, 44]. Recent breakthroughs have leveraged repetitiveness in data through compressed suffix arrays, run-length compressed structures, and grammar-compressed partial answers, which have substantially reduced storage requirements while supporting efficient document retrieval and counting [22, 38]. For applications involving online or evolving similarity functions—common in active learning and interactive data analysis—adaptive indexing solutions such as OASIS maintain families of locality-sensitive hash (LSH) indexes, dynamically updating them in response to user feedback without costly retraining. This results in heightened responsiveness and improved resource utilization in scenarios where similarity criteria are fluid [25, 48].

Space-partitioning techniques have evolved to address queries over complex multi-attribute datasets, including spatio-textual documents, 3D point clouds with attributes, and genomic sequences. Innovations such as persistent, parallel spatio-textual indexes and compressed attribute-aware spatial indexing facilitate queries across spatial, textual, and temporal dimensions, supporting top-k retrieval and attribute-based filtering with high throughput and efficient updates [18, 21, 44, 47, 70, 97]. At the algorithmic level, secondary partitioning techniques enhance traditional space partitioning by further dividing index cells, enabling duplicate-free and low-latency range and distance queries on spatially extended or non-point objects. For example, recent work [97] partitions each primary cell into secondary partitions defined by the begin and end values of object extents relative to the cell, which reduces both duplication and unnecessary computations in distance-range and join queries and outperforms earlier approaches in empirical comparisons.

The trajectory of research in space-partitioning indexing is determined by the interplay among data distribution, partitioning granularity, compression strategies, and the necessity for adaptation to evolving query patterns and dynamic workloads.

4.6 Efficient Index Management and Scaling

With increasing query volumes and the ever-growing size of datasets, efficient index management and robust scaling techniques have become fundamental for large-scale data retrieval tasks. Avoiding duplicates is particularly important; naive retrieval methods may inadvertently report the same results multiple times, leading not only to redundant computation but also to inefficiency in the presence of overlapping spatial objects or complex join queries. To address this, secondary partitioning techniques have been developed, notably those that subdivide each primary partition into secondary segments based on object boundaries, such as begin and end values in relation to a partition's spatial extent. As demonstrated in [97], this secondary partitioning improves query precision by more accurately localizing candidate object sets, reducing irrelevant verification steps, and providing significant performance improvements over older secondary partitioning approaches and state-of-the-art data-partitioning indexes in practical evaluations.

Scaling these indexing techniques has been further enabled by distributed and parallel index architectures leveraging modern cluster-computing frameworks like Apache Spark and Flink. Recent advancements include the adoption of lightweight learned index structures, which employ regression-based models (such as splines) and space-filling curve mappings to achieve $O(1)$ lookup by predicting object positions inside partitions. For example, the LiLIS system [25] introduces a lightweight, distributed learned index for large-scale spatial data, utilizing custom-trained models for each spatial partition and a range of partitioning strategies (R-tree, Quadtree, KD-tree, grid) tailored to data distribution. LiLIS proves robust to both data size and skewness (although query and partitioner choice can affect performance), supporting efficient point, range, k-nearest neighbor, and join queries. Experimental results indicate that LiLIS achieves 1.5–2 times faster index construction and over an order of magnitude faster query speeds compared to traditional spatial indexes in distributed big data environments, while maintaining full compatibility with Spark APIs. For instance, LiLIS achieves range query times around 472ms versus 521282ms with Sedona's RQ, and join queries an order of magnitude faster as well, across datasets as large as 300M points.

Contemporary data environments, characterized by continuous updates and evolving distributions, require indexing mechanisms with efficient incremental and parallel update capabilities. Persistent spatio-textual indexes, such as those presented in [70, 97], facilitate smooth integration of new data and near-real-time query execution, which is critical for applications demanding frequent refreshes, like event recommendation and geo-tagged search. Moving further, recent trends toward compressed computation and rep-index structures enable computation directly on compressed dataset representations, providing substantial space savings for highly repetitive data, as discussed in [70], although these approaches often present challenges in supporting dynamic updates efficiently.

The main approaches are summarized in Table 6, which presents the trade-offs among scalability, update efficiency, and distinguishing advantages for large-scale spatial query processing.

4.7 Graph Analytics and Advanced Query Structures

The increasing prevalence of graph-structured data in sectors such as bioinformatics, social networks, and software engineering necessitates specialized query mechanisms that extend beyond classical spatial or string indexing paradigms. This subsection aims to clarify the current landscape and motivations for advanced indexing and querying in graph analytics, focusing on similarities and trade-offs among leading techniques.

Table 6: Comparison of Space-Partitioning Index Strategies for Large-Scale Query Processing

Strategy	Scalability	Update Efficiency	Strengths
Tree-Based (e.g., R-tree)	Moderate (suffers in high dimension)	Moderate (requires rebalancing)	General-purpose; established theory
Grid-Based	High (especially with proper partitioning)	High (minimal restructuring)	Fast for point/range queries; low traversal overhead
Learned/Hybrid	Very High (adapts to data, $O(1)$ lookup)	High (can support incremental updates)	Handles skewed, high-dimensional data; efficient memory use
Compressed/Rep-indexes	High (suitable for repetitive data)	Moderate to Low (updates can be complex)	Dramatic space savings for redundant datasets

Among the main requirements is the capacity to efficiently resolve similarity queries—such as those based on edit distance or subgraph containment—which remain computationally demanding. Recent advancements in succinct data structures, including q-gram trees with hybrid encoding, have demonstrated substantial reductions in index memory consumption compared to previous filtering methods, while maintaining or improving filtering effectiveness and query speeds [23]. These compact structures blend global and local filtering strategies (such as degree and label-based refinement), enabling efficient navigation of large candidate spaces for graph similarity search. For example, the approach in [23] uses only 5%–15% of the indexing memory of prior methods, introducing enhanced filtering via degree and label filters, and scales to exceptionally large datasets (up to 25 million graphs). However, such in-memory indexes may face practical challenges when applied to even larger or more dynamic databases, as the construction and maintenance costs can grow with increased data volatility.

In the context of directed acyclic graphs (DAGs), efficient querying is enabled by techniques that exploit order, level, and separator-based decompositions. These methods provide strong worst-case performance guarantees, achieving near-optimal query complexities even under adversarial conditions [58]. Specifically, the Partial Order Multiway Search (POMS) algorithm in [58] uses recursive partitioning to achieve a competitive ratio of $O(\log n)$ compared to the optimal, where n is the number of vertices. This generalizes classical tree search results and shows practical benefits for search models in settings such as debugging and distributed systems. On the other hand, while the theoretical guarantees are robust, practical deployment may be impacted by the computational cost of finding optimal partitions and adapting to evolving DAG structures.

The ongoing convergence of hybrid filtering, succinctness, adaptive partitioning, and strong competitive guarantees reflects broader trends in processing high-dimensional and irregular datasets. Recent work, particularly from 2021 onward [23, 58], highlights both notable progress and emerging trade-offs: highly compact indexes and near-optimal query performance are achievable, but with open challenges regarding adaptability, scalability, and computation overhead for highly dynamic or exceptionally large graphs.

In summary, recent advances in graph analytics offer substantial improvements in query efficiency and memory usage for similarity search and DAG querying. Nonetheless, trade-offs remain between index succinctness, adaptability, and computational overheads. Researchers and practitioners should evaluate these methods based on their specific data scales, update rates, and application requirements.

4.8 Unified Perspectives for kNN, Similarity, and Join Operations

A broad analytic perspective reveals that k-nearest neighbor (kNN), similarity, range search, and join operations can be interpreted as instances of a unified data retrieval paradigm, especially over large, heterogeneous, or multimodal datasets. Recent empirical syntheses highlight the confluence of several methodological directions:

Spatial Partitioning: Organizing the search space hierarchically to prune irrelevant regions. This includes primary and secondary partitioning methods that divide data according to geometric and attribute-based features, shown to accelerate range, kNN, and join queries by reducing unnecessary computations and avoiding duplicate reporting [5, 18, 47, 48, 97].

Machine Learning-Based Index Construction: Leveraging regression models, space-filling curves, and other data-driven techniques to predict locations and enhance lookup speeds. Recent learned index structures combine lightweight models, such as splines or neural nets, with classical spatial partitioners, as in LiLIS and LLM-powered index advisors [25, 45, 71]. These data-driven indexes provide superior throughput and index build efficiency for large and dynamic workloads, with the caveat that model training costs and complex query support remain open research challenges.

Adaptive Query Evaluation: Dynamically tuning the search process to accommodate data characteristics, distributional shifts, and incoming queries. Approaches such as online metric learning and index parameter adaptation allow for real-time or streaming adjustments, as explored in frameworks like OASIS, which incrementally update similarity functions and reuse index structures to maintain performance and reduce overheads [22, 80, 93].

Ensemble and Subspace Techniques: Combining multiple indexing or filtering strategies to mitigate high-dimensional challenges and exploit complementary strengths. Ensemble clustering and consensus methods aggregate results over feature subsets or algorithmic variants, which is especially effective for noisy, high-dimensional, or categorical data [21, 38, 44, 57]. Ensemble schemes, parallelization, and local filtering collectively drive state-of-the-art performance for large and challenging similarity, kNN, and join problems [5, 18, 21, 22, 25, 31, 33, 38, 44, 45, 47, 48, 54, 58, 70, 71, 80, 93, 94, 97].

Robustness to noise and high dimensionality is further achieved through parallelism, compression, and ensemble models. Distributed frameworks have unified formerly distinct operations—such as kNN joins, range queries, and similarity joins—into single-session, high-throughput systems, minimizing I/O overhead and enabling resource-efficient knowledge discovery [21, 25, 57, 70]. The empirical record demonstrates that modern systems such as FML-kNN and LiLIS outperform prior MapReduce-style solutions, supporting scalable, robust retrieval across a variety of analytical tasks and data modalities.

For example, grammar-compressed and LCP-based indexes excel on highly repetitive string collections, outperforming naive approaches, but may introduce compromises in index construction time or incremental update capabilities [22, 33, 38, 44]. Meanwhile, machine-learned index structures and consensus-driven, parallelizable clustering approaches have substantially improved scalability and resilience to noise, albeit with increased algorithmic and training complexities [25, 57]. Distributed learned indexes and data-partition specific strategies enable efficient and accurate large-scale spatial and similarity search, particularly as data and workload heterogeneity increase [25, 71, 97]. As such, the methodological integration of space partitioning, local filtering, parallelization, and ensemble learning now underpins the state-of-the-art across similarity, kNN, and join algorithms in massive data environments.

Looking forward, key research challenges involve:

Supporting nonlinear and complex similarity functions, including those accommodating adaptive metrics [93]; Enabling non-parametric and domain-agnostic retrieval that generalizes robustly across workloads and data types [44, 45, 57]; Developing robust, fine-grained incremental index updates to support real-time and streaming scenarios [48, 54, 80]; Standardizing evaluation protocols for multimodal and streaming data, facilitating benchmark-driven progress [5, 38].

Addressing these challenges will catalyze the continued synthesis and advancement of indexing and querying approaches, fully adapted to the evolving demands of dynamic, large-scale, and heterogeneous data landscapes.

5 Dimensionality, Data Preprocessing, and Visualization

Section Objectives and Audience

This section aims to critically survey methods for data dimensionality reduction, preprocessing, and visualization, in direct relation to the overall objectives of this survey: to provide a unified, comparative, and updated overview of foundational and emerging techniques. It serves readers ranging from graduate students to practitioners in machine learning and data science who seek both breadth and insight into ongoing challenges and developments.

Overview and Scope

Techniques in dimensionality reduction, data preprocessing, and visualization play a pivotal role in preparing data for learning, facilitating model interpretability, and uncovering structural insights. Given the multifaceted objectives of dimensionality reduction—from alleviating the curse of dimensionality to aiding visualization of high-dimensional datasets—it is vital to systematically review and compare these techniques in terms of their tradeoffs, conventions, and evolving interpretations within the machine learning community.

Survey Contributions Compared to Existing Reviews

Dimensionality Reduction: Technical Overview

Dimensionality reduction techniques such as Principal Component Analysis (PCA), t-SNE, and UMAP are widely used to reduce data to tractable or visualizable representations. Linear techniques like PCA emphasize global structure and variance retention, whereas nonlinear approaches (t-SNE, UMAP) focus on preserving local neighborhood relationships. However, the choice of technique directly affects the downstream interpretability and risk of information loss—a contentious aspect in practical applications.

Unresolved weaknesses persist. In high-dimensional settings, nonlinear methods often obscure global relationships and may introduce artificial clusters or fail to preserve important features, a failure mode frequently overlooked in earlier literature. Interpretability tradeoffs are especially acute: while linear projections remain somewhat explainable, nonlinear mappings are opaque, and there is significant debate on methods for reliably interpreting their outputs.

Data Preprocessing and Workflow Considerations

Preprocessing steps such as normalization, feature selection, and imputation are critical not only for model performance but also for the meaningfulness of reductions and visualizations produced. Conventional wisdom—favoring aggressive normalization, for instance—has been challenged by recent work emphasizing the need for context-aware transformations. It remains controversial, for example, whether autoencoding approaches or graph-based techniques require distinct preprocessing pipelines as compared to classical methods. Careful consideration of data type, scale, and distribution is advised to avoid artifacts and misinterpretation in visualization.

Visualization: Interpretability and Unresolved Challenges

Visualization techniques are deeply intertwined with dimensionality reduction and preprocessing choices. The interpretability of low-dimensional scatter plots, for example, is often overstated; misleading cluster structures or loss of class separability represent common failure modes. This survey critically examines how new visualization paradigms challenge or reinforce existing conventions, and where the literature departs on best practices. Readers are directed to Section **??** for a novel taxonomy of interpretability-aware preprocessing approaches that is unique to this survey.

Summary Box: Open Questions and Contentious Areas

Key Unresolved Challenges: When (and for whom) are deep or nonlinear reductions interpretable? What preprocessing standards ensure robust visualization versus introducing artifacts? How should practitioners weigh global versus local structure retention?

Through systematically integrating technical review with these critical perspectives, this section aims to equip readers with not only factual knowledge, but also a nuanced framework for method selection and interpretation, fulfilling the broader objectives of this survey.

Table 7: Distinct perspectives and contributions of this survey vis-à-vis existing reviews on data dimensionality, preprocessing, and visualization.

Aspect	Prior Reviews	This Survey
Taxonomy	Often grouped by algorithm class (e.g., linear/nonlinear)	Introduces a use-case-driven taxonomy distinguishing interpretability, scalability, and intended application
Critical Analysis	Focus on method capabilities	Explicit analysis of unresolved failure modes and interpretability tradeoffs
Change in Conventional Wisdom	Typically omitted or lightly addressed	Highlights where consensus has shifted (e.g., t-SNE vs. UMAP for structure preservation)
Scope	Separate coverage of preprocessing and visualization	Integrated treatment linking preprocessing impact on visualization outcomes

5.1 Data Types and Representational Variety

Modern data science contends with an expanding diversity of data types, including numeric, categorical, temporal, spatial, multimodal, compositional, incomplete, dynamic, and high-variance forms. This variety substantially informs the choice and design of analytical tools by shaping the assumptions underlying algorithmic methods. For example, numeric and continuous variables—ubiquitous across disciplines—facilitate a wide spectrum of quantitative manipulations. In contrast, categorical data, particularly in high-dimensional or sparse contexts as observed in omics or textual datasets, challenge direct statistical analysis and demand well-chosen encoding or embedding methods [72, 75, 99]. Specifically, nominal attributes often require encoding schemes that preserve class informativeness and allow valid correlation or distance-based interpretation [72].

Temporal and sequential datasets further complicate analysis due to the necessity of maintaining order dependencies, affecting similarity computation and clustering methodologies [3, 82]. Spatial data, such as those arising from medical imaging or geographic information systems, impose unique representational requirements that must strike a balance between fidelity, computational efficiency, and the preservation of connectivity or adjacency information [34, 73, 87, 109].

The prevalence of multimodal and compositional data in fields such as systems biology or sensor analytics magnifies these complexities. Compositional data, defined by components representing parts of a whole and summing to a constant, oblige the use of specific transformations—such as log-ratio methods—and purpose-built regression models to ensure inferential validity [62, 102, 104]. Additionlly, the challenges posed by incomplete and dynamic datasets—including non-stationarity, time-varying drift, frequent updates, and deletions—necessitate adaptive preprocessing strategies capable of real-time reaction to evolving data [5, 28, 35, 100, 109, 110]. Data representations must also accommodate the practical realities of high variance and high dimensionality, which drive ongoing innovation in domains such as indexing, compression, and scalable embedding frameworks [5, 62, 87, 100, 109].

5.2 High-Dimensionality Challenges and Solutions

The widespread occurrence of high-dimensional data exacerbates both statistical and computational hurdles, encapsulated by the "curse of dimensionality." As dimensionality increases, the feature space grows exponentially, rendering conventional notions of distance less meaningful and impairing the performance of algorithms reliant on pairwise proximity [3, 82]. The resulting sparsity and noise accumulation compromise statistical power, heighten overfitting risks, and undermine clustering and learning efficacy. Classic distance metrics such as Euclidean and Manhattan distances, and kernel-based approaches, suffer from degraded discrimination in these settings, raising concerns for both exact and approximate k-nearest neighbor searches, high-dimensional clustering, and analyses of large-scale biological data [6, 38, 39, 57, 60].

To address these phenomena, methodologies that scale and adapt to high-dimensionality have emerged:

Feature selection and dimensionality reduction: Linear techniques such as Principal Component Analysis (PCA) and nonlinear methods like t-SNE and UMAP extract salient features and discard noisy or redundant ones [4, 55, 104]. However, recent work highlights that standard dimensionality reduction methods are often vulnerable to scattering noise, which can obscure cluster structures and reduce interpretability. The distance-of-distance (DoD) transformation, for example, has been shown to preprocess neighborhood distances to better separate noise from meaningful clusters in embeddings, significantly improving clustering accuracy, especially in very high-dimensional and low-sample regimes [55]. This approach demonstrates the need for advances specifically targeted at denoising and noise-induced artifact reduction during dimensionality reduction.

Adaptive metric learning: Tools including local Mahalanobis transforms and hierarchical subspace models enable more informative similarity calculations under small sample size relative to feature count ($p \gg n$) [39, 82, 104]. Adaptive metrics, such as in double-weighted k-nearest neighbor frameworks, allow the algorithm to individually weight features by informativeness, thereby mitigating the dominance of irrelevant dimensions and improving classification and clustering outcomes in high-dimensional settings [6]. Clustering performance also depends critically on data normalization and scaling; recent studies introduce scaling approaches based on multidimensional shape complexity to enhance the separation of clusters, albeit with additional computational cost [3].

Ensemble subspace methods: Aggregation over multiple random or systematically chosen low-dimensional projections mitigates overfitting and stabilizes models [57]. Consensus clustering methods, such as those employing co-association matrices and feature reweighting, have shown substantial improvements in accuracy and robustness against noise, particularly for categorical data with only a minority of informative features. Ensemble approaches also benefit high-dimensional regression and classification, as exemplified by ensemble Lasso or trimmed averaging techniques, which outperform traditional methods with complex, less sparse models and provide more stable results across a range of parameter choices [4, 57].

Dynamic and streaming data analyses: Incremental index structures, real-time clustering, and continuous normalization address the demands of evolving datasets [16, 28, 76, 102]. The emergence of learned multi-dimensional indexes introduces machine learning approaches directly into the indexing process, allowing for more adaptive and efficient querying in high-dimensional databases, though challenges in dynamic workloads and precise error bounding persist [60]. Novel data structures for dynamic and streaming updates in geometric and topological spaces have also become central for scalable processing [16, 28, 76].

Despite these advances, many high-dimensionality solutions display sensitivity to specific data distributions and parameterizations. Moreover, practical trade-offs between interpretability, computational cost, and robustness to noise persist as fundamental issues across application domains [22, 50, 102]. Ongoing research also emphasizes the challenge of statistical inference under high-dimensions, such as testing mean vectors or symmetry, particularly with missing data or small sample sizes [22, 50, 102]. The ongoing quest to generalize methods robustly across diverse modalities and to guarantee interpretable, meaningful low-dimensional representations remains central to current research.

5.3 Preprocessing and Normalization

Data preprocessing is foundational to robust and reliable analytics, particularly when analyzing high-dimensional, heterogeneous, or noisy datasets. The main objectives are to mitigate noise and outlier effects, normalize feature scales, and ensure compatibility with downstream models.

Standard normalization methods, such as min-max scaling, z-score standardization, and variance-stabilizing transforms, aim to harmonize feature ranges. However, these approaches may be inadequate when faced with outlier-prone or heavy-tailed distributions, or when compositional constraints are present [3, 35, 73, 95]. For compositional data, specialized transformations like log-contrast or isometric log-ratio are necessary to avoid spurious correlations resulting from constant-sum constraints [50, 104]. Notably, [35] addresses preprocessing challenges for nominal and mixed-type attributes by encoding categorical data numerically, allowing for inclusion in quantitative analyses and downstream clustering, even when classical statistical measures may not be well-defined.

The selection of normalization procedure can significantly influence the interpretability and performance of clustering and classification. Recent work [3] proposes determining feature-wise scaling factors by optimizing shape complexity, emphasizing the importance of balancing intra- and inter-cluster distances before clustering. This approach demonstrates improved clustering performance over traditional normalization, particularly in ambiguous scenarios, though it can require expert intervention. In the context of time-series data, careful consideration of normalization choices is likewise critical, as noted in recent comparative taxonomies [73].

Robust outlier detection and adjustment remain crucial in preprocessing, as these steps have a substantial impact on analytical outcomes. Nevertheless, there is a scarcity of standardized benchmarks for evaluating outlier handling and noise mitigation efficacy, underscoring the importance of transparent and reproducible preprocessing pipelines. Domain-specific customization is frequently needed, especially for normalization and duplicate management [8, 28, 38, 50, 100, 104, 110]. The literature demonstrates that modern k-nearest neighbor (kNN) based methods actively incorporate noise-resilient feature selection and adaptive weighting to achieve improved classification on imbalanced or noisy datasets [8, 38]. Furthermore, methods such as consensus spectral clustering on high-dimensional categorical data illustrate the effectiveness of integrating feature reweighting schemes based on informativeness, resulting in robust performance under both stochastic and adversarial noise, even when only a small fraction of features are informative [57].

In streaming and dynamic data environments, preprocessing must address additional challenges: algorithms should efficiently assimilate new information, adapt to evolving data distributions (concept drift), and handle data deletions or reweighting without necessitating a full retraining of models [5, 16, 28, 76]. These constraints are particularly prominent in real-time analytics and online learning, where the demand for balancing statistical rigor with computational efficiency is ever-present.

5.4 Dimensionality Reduction and Visualization Techniques

Dimensionality reduction and visualization remain fundamental tools for extracting informative low-dimensional representations from complex datasets, elucidating latent structures, and supporting both exploratory and inferential analysis. Principal Component Analysis (PCA) and its advanced variants, such as generalized contrastive PCA (gcPCA) and contrastive PCA (cPCA), are central techniques. Unlike conventional cPCA, which requires tuning a sensitive hyperparameter, gcPCA introduces a normalization strategy that penalizes high-variance dimensions, ensuring robust and interpretable separation of patterns across contrasting datasets without hyperparameter dependence [4, 29]. In gcPCA, the normalization incorporated into the eigendecomposition step makes the method robust to noise and rank deficiency and obviates the need for ad hoc hyperparameter selection. Extensive benchmark analyses show that gcPCA uncovers biologically meaningful axes in data such as hippocampal electrophysiology and single-cell RNA-seq, outperforming both PCA and cPCA in separating relevant patterns [29]. These extensions support accurate distinction of signals across experimental conditions and are particularly advantageous in cases where traditional approaches suffer from interpretability and tuning challenges.

Nonlinear embedding tools, notably t-SNE and UMAP, are widely adopted for visualizing cluster and manifold structures in high-dimensional domains including single-cell genomics, neuroimaging, and image analysis. However, they can be confounded by scattering noise, where random fluctuations obscure cluster boundaries in low-dimensional projections [29]. The distance-of-distance (DoD) transformation addresses this by processing the original distance matrix—by emphasizing differences in local neighborhood distances—to sharpen cluster detection and separate noise as distinct clusters. The DoD transformation, as demonstrated in simulated and real datasets, contracts noise-to-noise distances more strongly than cluster-related distances, especially at higher ambient dimensionality and smaller sample size, thus improving the clarity and recovery of true cluster geometry [55]. DoD's efficacy is quantified via improvements in Adjusted Rand Index (ARI) scores and classification accuracies (e.g., from 84.3% to 91.2% in high-dimensional neural and image data) [55]. However, DoD transformation requires $O(N^2)$ computational resources and careful tuning of neighborhood parameters. Ongoing research aims to optimize computation and parameter selection, as well as broaden application domains.

Supervised dimensionality reduction and feature selection are advanced not only by penalized regression methods, such as Lasso, Elastic Net, and Adaptive Lasso, but also by ensemble subspace approaches. Ensemble subspace methods, like ensemble Lasso or consensus spectral clustering, aggregate results across diverse randomly-selected feature subsets and employ strategies such as trimmed means or majority voting, achieving higher robustness and accuracy—especially in challenging $p \gg n$ or weak signal scenarios. In particular, ensemble approaches such as Ensemble Linear Subspace Analysis (ELSA) and consensus-based spectral clustering have been shown to offer gains in efficiency and minimax error rates under high-dimensional regimes when compared with single-model approaches [4, 57]. These methods remain robust under model complexity, high feature correlation, and adversarial or stochastic noise, attaining strong empirical and theoretical performance in domains such as genomics and text clustering [57]. Especially when only a small fraction of features are informative, ensemble voting or aggregation methods improve resistance to noise accumulation and enhance clustering accuracy, although they entail greater computation that can be mitigated by parallelization [57].

Visualization frameworks have expanded beyond traditional scatterplots to include specialized representations of clusters, graphs, tensors, and multidimensional networks. For example, Flowcube employs interactive matrix-based representations enhanced by direction-based filtering lenses for elucidating complex geographic flows without excessive clutter, enabling richer user-driven pattern discovery in large, spatial datasets [89]. The system models flows as directed triples, (p_i, p_j, m_{ij}), and the Direction-Based Filtering (DBF) lens permits selective, non-distorting interactive exploration of nuanced flow patterns even in massive datasets. While interpretation depends on user interaction and lens configuration, Flowcube systematically outperforms traditional cluttered displays and reveals weak trends often missed by automated filtering [89]. Methods grounded in tensor decomposition and model-based clustering, such as the tensor normal mixture model combined with penalized likelihood and doubly-enhanced EM algorithms, deliver parsimonious and interpretable compression and unsupervised grouping for high-order, structured data with rigorous statistical guarantees [59]. The doubly enhanced expectation-maximization (DEEM) algorithm leverages tensor structure, adaptive penalties, and separability constraints to yield consistent cluster recovery even as tensor dimensions scale exponentially, outperforming vectorized and classical clustering approaches under challenging conditions [59]. Where data are inherently multiway or graph-structured, compact encodings and algorithmic advances further support scalable and interactive exploration [5, 16].

Interpretability, transparency, and reproducibility are increasingly emphasized in the field. Contemporary approaches feature explainable feature allocation, use of cluster validity indices, rigorous benchmarking standards, and accessible software tools, ensuring traceability and practical adoption [16, 29, 62, 66, 89, 104]. Nevertheless, persistent challenges include producing consistent embeddings across runs, mitigating batch effects, and scaling methods to ever-larger and more heterogeneous multimodal datasets.

6 Feature Selection, Classification, and Vector Modeling

6.1 Feature Ranking and Robust Classification

Feature selection and classification in high-dimensional domains—especially when both the number of variables (p) and observations (n) are large and the data presents high variance or outlier contamination—have undergone substantial advances in recent literature. Key challenges include maintaining robustness to outliers, ensuring interpretability, and achieving computational scalability. Traditional classifiers frequently encounter difficulties in settings where class separation is predominantly due to variance differences or where outliers heavily influence the results.

Recent frameworks incorporate rank-based and subsampling strategies to address these issues. Notably, rank-based classification methods rely on transforming pairwise distances between observations into rank information, enabling classification procedures that are resilient to distributional assumptions and offer enhanced robustness against outlier effects [65]. These frameworks follow a systematic approach composed of (i) computing distance matrices among samples, (ii) applying rank transformations to the distances, and (iii) integrating rank-derived features into classifiers such as quadratic discriminant analysis, facilitating class separation in challenging high-dimensional scenarios. The default use of the ℓ_2 distance can be adapted to alternate metrics, improving applicability to data with network or non-Euclidean structure, while multi-class problems are handled by direct extension of the methodology.

Complementing rank-based methods, efficient subsampling strategies tailored for high-dimensional settings have also been developed [20]. For example, recent approaches first utilize a random LASSO-based selection to identify a sparse set of active predictors and then employ leverage-score-informed subsampling to select observations that capture the essential structure of the data. This two-stage framework leads to more accurate variable selection and reduction in predictive error (e.g., mean squared prediction error), while substantially reducing computational demands—particularly in large-scale or $p > n$ contexts. Sensitivity analyses reported in these studies confirm that such procedures are robust across a range of algorithmic parameters.

Empirical evidence, including extensive simulation studies and real-world applications such as sentiment classification and blog post comment prediction, shows that both rank-based classifiers and subsampling-based feature selection frameworks can match or surpass state-of-the-art alternatives. This is especially apparent in settings requiring noise resilience or flexible handling of unconventional feature distributions [20, 65]. Nevertheless, several challenges remain, including algorithmic scalability for extremely large datasets and the principled selection of distance metrics for rank-based methods. These persist as active research directions in high-dimensional robust classification.

6.2 Nonparametric and Subdata Selection Methods

Key Objectives: This subsection surveys recent advances in nonparametric and subdata selection methods, with explicit focus on (i) scalable and robust variable selection when both sample size and dimensionality are large; (ii) methodological trade-offs in dual-stage versus traditional selection pipelines; and (iii) actionable open challenges in high-throughput and omics data analytics. These objectives align directly with the overarching goals of the survey, which are to synthesize state-of-the-art methodologies and offer guidance for practitioners on robust statistical modeling in the era of high-dimensional data.

Nonparametric approaches and advanced subdata selection techniques have become indispensable for analysis in settings where the number of observations (n) and features (p) are both extremely large. The principal aim in such contexts is to achieve robust variable selection and reliable estimation while navigating substantial computational and statistical difficulties. Traditional LASSO-based selection often loses effectiveness with highly correlated predictors or extreme dimensionality ($p \gg n$), motivating strategies that decouple variable screening and subdata extraction for improved performance and interpretability.

A dual-stage methodological framework exemplifies these advances: it first employs a randomized LASSO procedure for initial variable screening, followed by leverage-score-driven sampling to systematically identify the most informative data points for downstream estimation. This combination addresses critical weaknesses of classic LASSO, affording higher estimation accuracy and improved computational efficiency in empirical analyses, as demonstrated in both recent simulation studies and real-world applications [20, 36].

To clarify the relative merits of single- versus dual-stage procedures, Table 8 compares their defining characteristics and main trade-offs.

Immediately following from Table 8, dual-stage procedures deliver systematic gains in both robustness and estimation accuracy compared to classical single-stage approaches. In particular, the integration of randomized LASSO with leverage-score selection yields more informative sampling and sustained predictive performance, especially in challenging high-dimensional settings [20]. Nevertheless, practitioners should note that dual-stage strategies require careful algorithmic tuning and thoughtful parameter selection, as supported by recent sensitivity analyses [20]. These pipelines are particularly advantageous when full-data computation is prohibitive, supporting scalable analysis for massive datasets.

Beyond linear regression settings, methodological advances extend to regression models with high-dimensional compositional covariates. Recent innovations include hierarchical, mixed, and p-value-free false discovery rate (FDR) control schemes, which exploit symmetry in test statistics to enable valid inference under severe correlation or dimensionality. These strategies secure strong FDR control and high asymptotic power, with practical superiority confirmed in synthetic and real omics applications [20].

Penalized likelihood estimation for high-dimensional mixed-effects models has achieved increased accuracy and interpretability with new coordinate descent algorithms using nonconvex penalties such as SCAD. Compared to LASSO, SCAD demonstrates markedly improved variable selection accuracy and reduced estimation bias, particularly in the presence of correlated predictors or group structure, as evidenced across transcriptome, GWAS, and microbiome benchmarks and applications [36]. The release of open-source software (e.g., HighDimMixedModels.jl [36]) supports adoption, though important research questions remain regarding parameter tuning, convergence, and robust uncertainty quantification for non-Gaussian outcomes.

Open Challenges and Methodological Recommendations: Key outstanding challenges span balancing computational cost and information gain in dual-stage subsampling; designing robust metric selection for varying data structures; improving scalability as n and p grow; and accelerating parameter tuning and inference procedures, especially for complex models or non-traditional response types. For practitioners, dual-stage approaches (randomized LASSO plus leverage-based subdata selection) are recommended when predictor correlation or extreme dimensionality undermine standard LASSO, as detailed recommendations and guidelines in [20] attest. For mixed-effects models with group structure, nonconvex penalization (e.g., SCAD) via efficient coordinate descent is preferred over LASSO when precision in variable selection is a primary concern [36].

Summary and Outlook: In summary, nonparametric and subdata selection methods have undergone significant methodological refinement, culminating in dual-stage frameworks that emphasize statistical robustness, computational efficiency, and fidelity to complex data structures. These developments closely support the broader survey goal of bridging scalable inference and practical implementation. Continued advances will focus on tighter integration of subdata selection with robust metric development, innovative tuning schemes, and extending current methodology to address non-Gaussian and multi-view data.

6.3 Statistical Testing in High Dimensions

Section Objective: This subsection surveys key statistical testing methodologies suited to high-dimensional data, clarifies their practical strengths and limitations, and explicitly highlights pivotal open research challenges in benchmarking and cluster validation. The aim is to guide practitioners in method selection while identifying urgent avenues for future work.

Statistical inference in high-dimensional scenarios requires procedures that remain reliable when p is large relative to n, especially under substantial inter-feature dependencies. Traditional mean vector tests, such as Hotelling's T^2, frequently suffer degraded performance in these settings, evident in inflated type I error rates and diminished power. U-statistic-based methods have been developed for both one- and two-sample settings to address these issues, providing test statistics that asymptotically converge to t-distributions as p increases [22, 50]. This construction allows these methods to avoid resampling or complex corrections, ensuring more direct and accurate inference—features demonstrated in areas like neuroimaging or genomics where "large p, small n" prevails. Notably, simulation work confirms that these U-statistic-based approaches maintain type I error control and statistical power even when sample sizes are extremely small and p very large [50].

Table 8: Comparison of Traditional and Dual-Stage Subdata Selection Methods

Method	Variable Selection Stage	Subdata Selection Stage
Standard LASSO	Single-stage (LASSO only); less robust under high correlation or $p \gg n$	No explicit subdata selection; potentially inefficient for large n
Dual-Stage (Random LASSO + Leverage)	Randomized LASSO for variable screening, enhancing robustness to correlation	Leverage-score sampling identifies informative data points, reducing computation and improving accuracy

Practical challenges pertinent to modern applications include handling missing data and computational scalability. For example, tests tailored for data missing at random (MAR) introduce new statistics and establish asymptotic validity as both n and p grow, providing robustness in incomplete observation contexts [102]. Computational efficiency in ultra-high-dimensional settings is advanced by projection-based strategies, which estimate null distributions by leveraging the concentration of measure phenomenon and facilitate valid inference without excessive computational burden [22, 39, 104]. Nevertheless, while random projections can reduce variance and accelerate computing, they may also obscure weak signals or elevate type II error risk, particularly in correlated data scenarios [39].

Recent literature underscores the essential alignment of statistical assumptions with hypothesis structure when applying multiple testing corrections and cluster validity assessment [43, 50, 60]. Benchmarking cluster validity in truly heterogeneous high-dimensional data remains a substantial obstacle. Most currently used metrics either lack generalizability or fail under noisy and structured environments. Notably, the field lacks absolute cluster validity indices that are robust across data modalities and variable noise levels [43]. Benchmarking via simulations persists as the main comparative strategy, yet this route is sensitive to generative assumptions and remains difficult to reliably reproduce across research groups [39, 50].

Ensemble and consensus clustering frameworks have proved increasingly effective in high-dimensional contexts by incorporating dimension reduction, feature reweighting, and model averaging to bolster noise robustness and adversarial resilience. Combining one-hot encoding with random projection and spectral consensus enhances cluster recovery but correspondingly increases computational demands [50, 104]. Despite successes—especially with potential for parallelization—there remains insufficient guidance on optimal parameter selection and theoretical guarantees, particularly for data with extreme heterogeneity.

Methodological Recommendations for Practitioners: For mean testing in high-dimensional and small-sample settings, U-statistic-based tests [50] offer direct, calibration-free alternatives ideal for scenarios where resampling is computationally unsustainable. For incomplete or MAR data, practitioners should employ tests specifically constructed for missingness patterns [102]. When inference speed is critical and data dimensionality extreme, projection-based tests [22, 104] provide strong trade-offs between speed and reliability, though users should interpret findings cautiously if data exhibit heavy correlation or the expected signals are faint [39]. Regarding cluster validation, ensemble approaches and adoption of absolute validity metrics (when available) [43] are advisable, but always with critical attention to the properties and composition of the underlying data.

Example Use Case: Consider neuroimaging data with $p \approx 10,000$ voxel-based features from as few as 20 subjects, coupled with missing entries for some subjects. Here, U-statistic-based t-tests [50] deliver mean difference testing without resampling, while fast projection-based inference [22, 104] offers further computational acceleration. For subsequent cluster validation, practitioners employing modern ensemble clustering and absolute validity indices [43] should remain cautious in interpreting these metrics, especially where real datasets diverge from controlled simulation scenarios; using a mix of metrics and simulation-driven benchmarking is currently best practice.

Summary and Key Takeaways: In summary, high-dimensional statistical testing and validation present unique methodological and computational hurdles (Table 9). Research frontiers include robust p-value calibration in extreme regimes, principled handling of missing data, reproducible benchmarking strategies, and scalable consensus frameworks. Continued progress will require the development of universally reliable validity metrics, practical benchmark suites, and actionable methodological guidelines that harmonize statistical rigor with modern data heterogeneity.

6.4 Vector Space and Distributional Semantic Models

Key Objectives: This subsection aims to clarify the primary objectives and advances in vector space and distributional semantic models, focusing on three central requirements: interpretability, predictive accuracy, and scalability. It further highlights open challenges and provides guidance for future methodological choices.

Semantic modeling in high-dimensional linguistic or biological contexts requires vector representations that balance interpretability, predictive accuracy, and computational tractability. Distributional semantic models, which encode entities as vectors in high-dimensional spaces, have demonstrated strong performance in capturing semantic relationships. Notable approaches such as neural embedding models (e.g., word2vec) and matrix factorization methods (e.g., NMF) achieve strong predictive accuracy; however, their typically dense representations hinder interpretability at the dimension level.

To address this limitation, recent research introduces dimension selection procedures that directly map naturally occurring attributes — such as specific words — onto vector dimensions, thus uniting interpretability with performance. Pakzad et al. [64] exemplify this direction by proposing a method that selects a subset of the most frequent words as the basis dimensions for embedding. The resulting vector spaces are both interpretable and highly accurate; on the ukWaC corpus, selection of $N = 1500$ basis words enabled strong performance. Empirical results on benchmark similarity tasks (MEN, RG-65, SimLex-999, WordSim353) show that reducing the number of basis vectors from 5000 to 1500 comes at a minimal predictive cost of only about 1.5%–2%, while producing notably

Table 9: Summary of Open Problems in High-Dimensional Statistical Testing and Validation, with Promising Current Approaches

Research Challenge	Limitation	Promising Approaches and Reference
Calibration of p-values in ultra-high dimensions	Classical estimators often inflate type I errors for $p \gg n$	U-statistic-based tests with t-calibration [50]
Testing under missing at random (MAR) mechanism	Robustness to missing data requires new test derivations	Asymptotic tests for MAR, leveraging incomplete data [102]
Cluster validity in heterogeneous, noisy datasets	Few absolute cluster metrics work across data types	Emerging absolute validity indices, but generalizability is limited [43]
Benchmarking and reproducibility of new methods	Evaluation depends on simulation design; lacks consensus benchmarks	Simulation-based benchmarking; calls for standardized, diverse datasets [39, 50]
Systematic integration of projection-based tests	Risk of masking weak signals or introducing bias in correlated data	Random projection-based statistics coupled with robust hypothesis designs [22, 39, 104]
Parallel computation for ensemble or consensus testing frameworks	Optimal design and scalability are underexplored in large, real-world settings	Parallelizable ensemble clustering and consensus mechanisms; more theoretical work needed [50]

more interpretable representations. Comparative interpretability assessments further highlight the clear advantage of these word-based vectors over neural embeddings and NMF baselines.

Regarding database indexing and retrieval, construction of vector models leveraging machine learning methods—such as clustering, neural networks, and hybrid systems—enables efficient, adaptive handling of multi-dimensional data. These advances facilitate scalable querying, indexing, and retrieval, thereby supporting large-scale, dynamic datasets [43].

Methodological Recommendations: Practitioners are advised to select metric and technique combinations best suited to their data and task conditions. For interpretable applications, basis selection strategies that map natural attributes (such as specific words) to vector dimensions (as in [64]) are preferable, especially when a slight loss in predictive accuracy is acceptable in exchange for greater transparency. For highly dynamic or large-scale datasets, clustering-validity and adaptive indexing frameworks (see [43]) should be prioritized to maintain both speed and accuracy.

Actionable Open Problems: Progress in this field depends on: (i) establishing principled, data-driven interpretability metrics for high-dimensional vectors, (ii) refining methodological trade-offs between accuracy and interpretability as the number of dimensions is reduced, and (iii) inventing adaptive systems for indexing/retrieval that can respond in real time to data scale and distributional changes. Addressing these issues is critical for advancing transparent, scalable, and effective vector space modeling.

Concluding Key Takeaways: Vector space and distributional semantic models continue to evolve toward greater interpretability without sacrificing accuracy or scalability. Embedding methods rooted in natural attribute selection, together with adaptive indexing systems, represent promising directions for both theoretical innovation and practical implementation.

7 Benchmarking, Evaluation, and Cluster Validation

This section sets forth the essential objectives and persistent challenges inherent to benchmarking, evaluating, and validating clustering methods. We first clarify foundational concepts; next, we present standard methodologies, then conclude by highlighting the most critical open questions, with particular emphasis on high-dimensional and heterogeneous data settings.

The primary objectives of this section are: (i) to define the goals and scope of benchmarking, evaluation, and validation in clustering; (ii) to streamline the methodological landscape and identify actionable recommendations for practitioners; (iii) to summarize the main unresolved problems; and (iv) to propose taxonomical guidance for ongoing and future research.

Benchmarking involves systematic and comparative testing of algorithms in well-designed and controlled scenarios, leveraging standardized datasets and evaluation metrics. Evaluation addresses the assessment of clustering quality—whether by reference to external ground truth or by measuring internal cohort coherence and separation—using instruments such as the adjusted Rand index, silhouette score, and stability or consistency statistics. Cluster validation concerns the reliability and statistical support for detected group structure, most critically in scenarios lacking true labels.

A principal challenge lies in selecting and developing evaluation metrics appropriate to the problem context—especially for high-dimensional or noisy data, where classical measures may falter. Key issues remain in the interpretability of results, sensitivity to initialization schemes, and the ability to generalize evaluation strategies to heterogeneous or complex data types.

Example applications include assessing the stability of clustering results on gene expression profiles via resampling protocols, or benchmarking text clustering algorithms on corpora with differing topic distributions and vocabulary richness. In such cases, practitioners are advised to explicitly define both the evaluation goal (e.g., external vs. internal validity) and the corresponding metrics, adapting technique choices to the expected noisiness, dimensionality, and structure of available data.

Where open problems are summarized in Table 11, we note that dataset diversity is being pursued via community-driven benchmark suites, while metric robustness now frequently leverages ensembles of metrics or consensus strategies. For cluster validation, statistical approaches like bootstrap testing and gap statistics are increasingly adopted to buttress results when labels are unavailable. The need for cohesive frameworks that unify these threads remains a pressing direction for future research.

At the conclusion of each main subsection, we provide concise summary statements of the major research challenges in that area. As a guide to further progress, we propose a stylized taxonomy—integrating validator, benchmarking, and representation-based perspectives—to orient researchers toward holistic, actionable solutions.

The following subsections elaborate on state-of-the-art methods, metrics, and systematizations for benchmarking and validating clustering algorithms. Each is introduced with explicit objectives and concludes with a focused summary of open research questions, supporting methodological clarity and practical utility.

7.1 Cluster Validation and Evaluation Metrics

Robust cluster validation underpins the scientific credibility and reproducibility of unsupervised learning methodologies. Two primary paradigms exist for validating clustering results: internal

Table 10: Summary of Open Problems and Promising Approaches in Vector Space and Distributional Semantic Models

Open Problem	Notable Approach/Direction	Current/Future Recommendations
Evaluating interpretability	Use attribute-based basis dimensions (e.g., frequent words) [64]	Develop rigorous, standardized interpretability metrics for comparing vector spaces
Balancing accuracy and interpretability	Word-selection methods with controlled basis size [64]	Experiment with varying dimension counts to find optimal trade-offs for given tasks
Scalable indexing & retrieval	Integrate clustering and neural-based models [43]	Employ adaptive indexing that re-optimizes as data evolves
Adapting to data shifts	Hybrid indexing and machine learning frameworks	Build systems with continuous monitoring and adaptation to shifting distributions

Table 11: Overview of Key Open Problems in Clustering Benchmarking and Evaluation, with Notes on Promising Approaches

Thematic Area	Open Problem	Description and Promising Directions
Benchmarking	Dataset Diversity	Lack of standardized, diverse benchmarks in high-dimensional and heterogeneous settings; active development of new open repositories and synthetic data generators that reflect domain complexity.
Evaluation	Metric Robustness	Selecting or designing evaluation measures resilient to dimensionality and noise; emerging methods incorporate stability analysis and multi-metric reporting for more robust assessment.
Cluster Validation	Statistical Guarantees	Validating discovered structure without access to ground-truth labels; advances include resampling-based significance testing and tight theoretical criteria for cluster validity.
Cross-cutting	Integration of Methods	Lack of unified frameworks or taxonomies linking benchmarking, evaluation, and validation approaches; growing interest in modular toolkits and meta-evaluation studies to bridge these domains.

(absolute) and external (relative) measures. Internal validation indices—such as the Silhouette coefficient, Dunn index, and Davies-Bouldin score—evaluate clustering quality without recourse to ground truth labels. These methods efficiently quantify cluster compactness and separation, yet they can be influenced by noise, feature scaling, and data dimensionality. Notably, in high-noise or high-dimensional contexts, these metrics often struggle to distinguish true structure, potentially misrepresenting clusterability, particularly when faced with chaining artifacts, small clusters, or overlapping densities [1, 4, 6, 8, 17, 19, 21, 31, 36, 39, 42, 43, 47, 50, 62, 66, 73, 84, 85, 88, 95, 100–102, 104, 111]. Caution is thus advised against relying solely on internal metrics for conclusive assessment [17, 102].

External indices—including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), F1-score, and Cohen's Kappa—compare the computed clustering to reference labels, providing more interpretable and objective benchmarking, especially when a gold standard exists [4, 6, 11, 17, 19, 21, 31, 36, 39, 42, 43, 47, 50, 62, 66, 73, 84, 100–102, 110, 111]. Among these, ARI and NMI have consistently demonstrated robustness and discriminative power, outperforming less nuanced measures such as purity [11, 42]. Nevertheless, these metrics possess their own biases, for example towards certain cluster size distributions and cluster counts—a challenge accentuated in multiclass, imbalanced, or high-dimensional data.

Moreover, metrics such as precision at top-n (P@n) and area under the ROC curve (AUROC), prevalent in related settings like outlier detection, require meticulous adjustment for dataset imbalance and sampling artifacts to avoid misleading results [17].

Recognizing such limitations, recent advancements have introduced more context-sensitive and adaptive validation strategies. These include indices leveraging correlations between within-cluster and centroid distances, which can highlight multiple plausible clustering solutions, aligning more effectively with real, often hierarchical, data structures [42]. Additionally, multimodality tests—including the Dip test and Silverman's test applied to pairwise distances—offer robust, distribution-agnostic assessments of clusterability, generally outperforming classic indices in differentiating between true signal and noise, though challenges remain for specialized structures such as heavy chaining or tiny clusters [50, 102].

Finally, parameter sensitivity and data preprocessing practices—such as normalization, scaling, and duplicate handling—strongly influence metric reliability. Accordingly, adaptive, data-driven feature scaling procedures and robust software implementations are increasingly integral in contemporary clustering analyses [50, 66].

As summarized in Table 12, selection of appropriate validation metrics must account for data characteristics, application context, and the availability of ground truth labels. No single approach suffices across all scenarios; thus, rigorous studies routinely report multiple metrics and qualify their interpretations.

7.2 System-Level and Analytic Metrics

A comprehensive evaluation of clustering and search systems extends beyond standard clustering quality measures, aligning with the overarching survey objective of equipping practitioners to assess and select real-world, scalable solutions. Recent literature emphasizes the need for system-level analytic metrics that capture efficiency, scalability, and robustness—criteria that have become ever more pivotal given the explosive scale and heterogeneity of modern datasets [1, 3, 5, 8, 14, 16, 23, 32, 35, 57, 60, 70, 73, 76, 82, 83, 95, 96, 100, 110, 112].

Latency and Throughput: Real-time search and analytics demand optimized latency and high throughput. Advances in approximate nearest neighbor search (ANNS) leverage improvements in memory layout and vector handling, such as those demonstrated by VSAG and MRQ frameworks that employ software prefetching, cache-aware structures, and adaptive parameter tuning to significantly lower response times and tuning costs for high-dimensional queries [70, 96, 100, 110, 112]. Additionally, compressed and succinct indexing structures [14, 16, 23, 32, 76, 83] bolster performance by enabling rapid lookups while reducing memory burden, and are effective when combined with automation in parameter tuning or hybrid encodings. The computational trade-offs between accuracy and speed are further highlighted in recent comparisons and deployment studies.

Scalability: Solutions must efficiently scale with data volume and dimensional complexity. Modern approaches—including graph-based, quantization-focused, and hybrid indexing—have demonstrated sublinear or near-linear scaling to billions of data points, using techniques such as adaptive partitioning, automaton-based compression, and hierarchical pruning [1, 14, 16, 23, 32, 57, 76]. Recent surveys of learned and dynamic multi-dimensional indexes [60]

Table 12: Common Cluster Validation Metrics: Key Properties and Use Cases

Metric Type	Example Metrics	Requires Ground Truth	Key Strengths / Limitations
Internal (Absolute)	Silhouette, Dunn, Davies-Bouldin	No	Fast; sensitive to noise/dimensionality; may not detect overlap/chaining
External (Relative)	ARI, NMI, F1-score, Cohen's Kappa	Yes	Interpretable w/ labels; can be biased by cluster count/imbalance
Multimodality/Novel	Dip test, Silverman's test	No	Robust to noise; less sensitive to structure; challenges in special cases

consolidate methodological advancements in scalability, supporting dynamic, high-dimensional workloads with mutable layouts and efficient insertion. The combination of these techniques is necessary for maintaining tractable memory and time complexity at scale, a requirement validated by increasingly rigorous benchmarking.

Resource Constraints: In edge and embedded contexts, methods are evaluated under strict resource limitations. Standardized metrics, such as "peak *n* per query" or "memory units per million points" [5, 16, 83, 95], provide fair benchmarks for comparing memory and computation across implementations. Novel designs such as windowed Cuckoo filters [83] and space-efficient voxel data structures [5] demonstrate substantial reductions in memory usage while retaining high lookup or query performance. The principles from compact hashing and succinct automaton-based structures help realize flexible deployments in environments constrained by hardware or energy.

Robustness: As data distributions evolve or are subject to noise, the robustness of clustering and search systems has moved to the foreground. Recent frameworks emphasize resilience through support for streaming and online updates [14, 60, 82] as well as ensemble and consensus mechanisms robust to adversarial noise [57, 112]. The introduction of specialized adaptiveness and self-constrained metrics—often paired with theoretical guarantees of consistency and error bounds—enables more nuanced evaluation of accuracy under evolving or perturbed data.

Systematic benchmarking now commonly mandates reporting a holistic suite of analytic and engineering metrics—spanning timing, space, accuracy, and robustness—across increasingly diverse operational scenarios. This integrated perspective supports well-informed system design and fosters transparency in empirical evaluation, directly contributing to the survey objective of promoting best practices and clearer comparisons between clustering and search solutions.

Transitioning from system-level assessment, the next subsection will discuss how metric selection and evaluation strategies can directly inform the development and choice of subdata and testing methodologies, highlighting their interconnected roles in effective, end-to-end system design.

7.3 Benchmarking Environments and Open-Source Tools

Objectives and Scope: This subsection systematically reviews the landscape of benchmarking environments and open-source tools for clustering, indexing, and similarity search, with explicit focus on facilitating transparency, reproducibility, and meaningful algorithmic comparison in high-dimensional and structured data contexts. We outline the resources, core features, and persistent gaps in current frameworks, clarifying the points of competition and interoperability among benchmarking ecosystems, simulation platforms, and domain-specific libraries.

Transparent and reproducible evaluation is anchored in open, standardized benchmarking ecosystems that encompass both implementations and curated datasets. The proliferation of open-source libraries—spanning Python, R, and, more recently, Julia—has increased access to advanced clustering, indexing, and similarity search techniques [5, 11, 17, 21, 22, 29, 32–34, 36, 39, 50, 51, 60, 76, 81–84, 86, 91, 93, 101, 102, 104, 106–108, 111].

Dataset Repositories: Resources such as the UCI Machine Learning Repository and OpenML furnish benchmarks across diverse data types, annotated with task-specific and preprocessing information. These repositories are central to enabling systematic comparative studies of clustering and similarity search algorithms across domains, as well as to promoting transparency in method evaluation [17, 32, 39, 76, 83, 104].

Simulation Frameworks: Tools for controlled manipulation of data properties—including class separation, noise, and dimensionality—aid the rigorous comparison of algorithmic performance and highlight sensitivity to dataset idiosyncrasies. These environments support experiments in both synthetic and real-world settings, as demonstrated by works such as HighDimMixedModels.jl for penalized high-dimensional mixed-effects modeling [36], and gcPCA, which supports benchmarking of pattern-detection methods on paired high-dimensional datasets [29]. Both approaches underscore the need for careful performance validation under varying complexity, data size, and variability.

Domain-Specific Libraries: Specialized implementations cater to modalities such as time series, trajectories, graphs, and point clouds, offering tailored distance functions and evaluation protocols. For example, scalable and distributed methods for *k*NN joins are vital for big data frameworks [21], while domain-aware tools address the high-dimensional challenges found in compositional, mixed, or structured data scenarios [39, 102, 104]. Libraries advancing voxelization and network embedding techniques play a crucial role in supporting robust analysis of 3D and network datasets [5, 51].

Framework Competition and Comparison: Benchmarking frameworks and storage structures vary in their targeted data modalities, scalability, reproducibility support, and ability to handle emerging challenges. For example, recent surveys of learned multi-dimensional indexes highlight the proliferation of specialized frameworks with trade-offs in mutability, query support, and efficiency, with key differentiators including support for kNN or range queries, storage overhead, update dynamics, and robustness to adversarial or dynamic workloads [60]. The continued development of new index structures for repetitive string or large network data further illustrates the competitive landscape among both classic and learning-augmented approaches [33, 60, 83, 108].

Reproducibility Artifacts: The growing adoption of public code repositories, leaderboards, and open challenge datasets has strengthened verifiability of benchmarking efforts. Increasingly, best practices prioritize publishing full experimental configurations, including random seeds, preprocessing scripts, and evaluation parameters, to advance community-wide interpretability and reproducibility [11, 17, 22, 29, 32, 50, 76, 82]. The work of Campos et al. [17] especially emphasizes rigorous documentation of benchmarking protocols and systematic evaluation for fair comparison, revealing the pitfalls of inconsistent dataset preparation and metric selection.

Despite progress, substantial challenges persist. The field continues to lack universally recognized benchmark suites that reflect the complexity of emerging applications, such as those involving graph, tensor, or mixed-type data [60, 81, 84, 91, 106]. Moreover, robust simulation and evaluation environments addressing data corruption, anonymization, and missingness are still needed, with recent research highlighting the importance of modeling such phenomena explicitly [29, 50, 66, 102]. Continued community investment is essential to curate, annotate, and standardize benchmarks that mirror the intricacies of real-world clustering tasks.

7.4 Visualization for Evaluation and Transparency

The goals of this section are to: (1) survey the roles of visualization in evaluating and communicating clustering and similarity search results, (2) synthesize recent methodological advances in interactive and interpretable visualization, and (3) assess impacts and open challenges, particularly concerning reproducibility and transparency.

Visualization is indispensable for both evaluating and communicating the results of clustering and similarity search, bridging the gap between algorithmic output, expert assessment, and end-user trust. The use of 2D and 3D visualization remains fundamental for exploratory analysis, while interactive dashboards now constitute standard practice for both method development and result dissemination [5, 16, 22, 29, 50, 59, 64, 89, 102, 104].

Modern frameworks commonly integrate dimensionality reduction methods—such as t-SNE, UMAP, and contrastive PCA—not merely as visualization tools but also as preprocessing steps that surface latent structure otherwise masked in high-dimensional data. For instance, recent methods like generalized contrastive PCA (gcPCA) enable the robust extraction of low-dimensional patterns that distinguish between complex high-dimensional biological datasets, offering interpretable axes of variation without the ambiguity of parameter tuning or ad hoc decision-making. gcPCA, for example, performs an eigendecomposition of the normalized difference in covariance matrices between two datasets, eliminating reliance on a manually selected hyperparameter and providing axes that demonstrate biological meaning and distinctiveness in neural and gene expression data [29]. Dimensionality reduction techniques also serve as foundational tools for validating cluster separation, detecting anomalies, and identifying ambiguous or overlapping subpopulations, where they supplement quantitative metrics with intuitive, expert-driven insight [50, 59, 102, 104].

Emerging solutions employ innovative interaction paradigms. For instance, systems that visualize flows or spatial networks, such as Flowcube, enhance the exploration of origin-destination data by introducing interactive spatial filters that reveal nuanced arrangement patterns across large-scale datasets. The Direction-Based Filtering (DBF) lens in Flowcube, for example, systematically isolates flows along specific spatial directions, unveiling subtle concentration, alignment, or dispersal patterns in geographic movement data that might be missed by conventional automated methods. The user-driven, interactive exploration supported by such paradigms enables the discovery of both prominent and subtle structure in complex datasets, as demonstrated through real-world movement analyses in urban environments. Flowcube further formalizes movement as a set of directed pairs with associated magnitudes, supporting detailed analysis of both major and minor movement patterns and outperforming traditional methods through its interactive lensing approach [89].

To ensure transparency and reproducibility, the latest best practices emphasize coupling visualization with systematic, script-driven analytics [29, 64, 102]. Open interfaces, reproducible color mappings, and support for exporting or replaying visualization states significantly strengthen interpretability and facilitate peer verification of findings [5, 64]. For example, approaches that select interpretable feature sets or basis vectors, as in interpretable word embedding models, enhance the transparency of downstream visual analyses by ensuring that each dimension corresponds to an easily understood concept. In recent work, selecting a set of natural words as basis vectors allows word embeddings to be interpreted directly via their component words, improving transparency with minimal performance loss and making validation and presentation more intuitive [64].

Despite these advances, visual analytics continues to confront significant and ongoing limitations. These include challenges with ultra-high-dimensional or massively multi-class data, as well as situations where intrinsic data structure resists intuitive mapping. In high-dimensional tensor and mean estimation tasks, validated through model-based clustering and statistical tests, effective visualization supports interpretation but also highlights bottlenecks in conveying complex relationships or uncertainty [50, 59, 102]. Such issues have motivated research into mixed-modality and interactive visualization methods that can scale alongside increasing analytical and data complexity, as well as the development of advanced frameworks and open-source tools that address the combinatorial and presentational challenges posed by high-resolution or semantically rich datasets [5, 29]. Moreover, the visibility, reproducibility, and interpretability fostered by open implementations and well-documented workflows have direct implications for the ethical deployment of high-dimensional data analysis, including the need for auditing, accountability, and cross-domain policy compliance.

In summary, visualization not only underpins the trustworthy evaluation and communication of clustering and similarity search but also sits at the intersection of methodological innovation, interpretability, and reproducibility, with evolving best practices driving interdisciplinary impact in research and applications.

8 Data Representation, Storage Optimization, and Hardware Acceleration

This section aims to systematically analyze foundational approaches and current advances in data representation, storage optimization, and hardware acceleration, highlighting their interdependencies and practical impact on AI systems. The explicit survey goals for this section are as follows: (1) to compare techniques for representing high-dimensional AI data; (2) to critically evaluate storage efficiency strategies in terms of their practical trade-offs; (3) to examine hardware acceleration mechanisms and their synergy with data structures; and (4) to discuss how these core areas impact reproducibility and policy considerations in AI engineering.

The subsequent discussion transitions from methods for data modeling and compression, through common storage architectures and optimization targets, to hardware designs directly supporting AI workloads. Each major subsection will explicitly state its focus and scope, aiding reader orientation, and will highlight how the discussed techniques may support or challenge reproducibility of AI results, as well as the broader ethical and policy landscape (where relevant).

Our goal is to connect these themes, clarifying how representational, software, and hardware elements interact in modern benchmarks and deployments. Furthermore, we consider the implications of these technologies for the reproducibility and auditability of AI workflows, thus broadening the interdisciplinary synthesis.

Open challenges in this domain include: - Developing scalable metrics for evaluating representation fidelity and downstream task relevance in massive, high-dimensional datasets, and ensuring these metrics are transparent for reproducibility purposes. - Balancing trade-offs between compression ratio, access latency, and energy efficiency in heterogeneous storage systems while remaining mindful of policy-driven requirements for data retention and privacy. - Enabling adaptive hardware-software co-design that remains robust to shifts in model architecture and data modality, and considering the ethical implications of energy consumption and environmental impact.

At the conclusion of this section, we synthesize open research questions—particularly concerning evaluation metrics in high-dimensional settings and the integration of representation, validation, and benchmarking methods—within a unified perspective.

Finally, Table 13 systematically summarizes principal open problems addressed in this section.

8.1 Data Representations for High-Dimensional Analytics

The analytical landscape for high-dimensional and multimodal data demands representations that are both expressive and computationally efficient. Classical strategies have relied on dense, grid-based formats for regular domains; however, in higher dimensions, the storage and computational requirements rapidly become prohibitive, driving the need for more advanced data structures that harness inherent sparsity and structural regularities characteristic of scientific and industrial datasets. Voxel-based encodings, which extend regular grid representations, remain prevalent for 3D spatial data due to their implementation simplicity and direct storage mapping. However, as dimensionality grows or data become increasingly sparse, these encodings exhibit significant memory inefficiency [5].

To mitigate these shortcomings, hierarchical structures have gained prominence. Examples include sparse voxel octrees (SVO), serialized directed acyclic graphs (SVDAG), and a spectrum of dynamic data structures such as OpenVDB, NanoVDB, SPGrid, and DT-Grid. These representations can dramatically reduce memory requirements—often by orders of magnitude—while preserving the capacity for locality-sensitive computations and supporting real-time manipulation [5]. Manifold-based approaches further extend this paradigm by succinctly capturing topological and geometric features, supporting advanced analytical tasks across fields such as computer graphics, computational biology, and scientific simulation.

The effectiveness of these data structures in high-dimensional contexts centers on several critical trade-offs. For instance, static memory layouts such as contiguous arrays deliver high throughput in batch-analytic or streaming scenarios but can be inflexible for adaptive or interactive applications. In contrast, dynamic memory layouts support interactive and adaptive analytics, yet introduce complexity due to sophisticated concurrency controls needed to maintain consistency and performance. Furthermore, hierarchical representations (such as SVO or SVDAG) greatly optimize storage for sparse datasets but at the cost of increased pointer overhead and potentially slower random access, particularly as spatial resolution grows. Dynamic grid solutions like OpenVDB and NanoVDB distinguish themselves by supporting efficient updates and parallelization, but may require substantial engineering effort for integration into existing high-performance workflows [5].

Despite their promise, challenges remain significant. The lack of standardized benchmarks and robust open-source libraries inhibits the objective evaluation of data structures across real-world scenarios. Many current implementations show reduced efficiency when dealing with non-watertight models or datasets that require rich semantic annotations. GPU acceleration, essential for volumetric analytics at scale, is not uniformly supported, which further restricts practical usability. Additionally, a critical obstacle is the absence of mature solutions for efficiently handling extreme resolutions and managing the trade-off between granularity and performance in dynamic or semantically complex contexts [5].

Open Problems and Research Challenges:

Standardized benchmarking and validation present persistent challenges for high-dimensional data analytics. Key unresolved problems include:

- *Establishing widely-used, standardized datasets and performance metrics for rigorously comparing novel data structures and algorithms.*
- *Developing robust and efficient libraries capable of supporting non-watertight and semantically annotated data in dynamic scenarios.*
- *Achieving scalable GPU-ready implementations that can manage large and heterogeneous volumetric datasets at extreme resolutions.*
- *Designing tools that facilitate the annotation and tracking of semantic information within hierarchical or sparse data representations.*

Table 13: Summary of principal open research challenges in data representation, storage optimization, and hardware acceleration.

Area	Open Problem	Key Considerations
Representation	High-dimensional evaluation	Scalability, metric interpretability, task alignment, reproducibility
Storage Optimization	Trade-offs in compression and latency	Efficiency, energy cost, reliability, policy compliance
Hardware Acceleration	Co-design for diverse workloads	Flexibility, future-proofing, integration complexity, ethical/environmental impact

> *Progress in these areas is crucial for enabling reproducible research and for the operational integration of advanced representations into scientific, industrial, and graphics pipelines.*

8.2 Space-Efficient Storage Structures

This subsection surveys recent innovations in storage structures that aim to overcome memory and I/O limitations in high-volume, high-dimensional analytics. The explicit objective of this subsection is to provide a structured synthesis of two main classes of space-efficient storage structures: probabilistic filters and compressed indexes. We clarify the primary scope as follows: (1) to outline the core algorithmic strategies that enable space reduction in these structures, (2) to compare key capabilities such as support for dynamic updates, false positive rates, and adaptability to evolving data, and (3) to discuss how recent developments affect their practical deployment for massive datasets, specifically addressing points of competition or complementarity among the alternatives.

Memory and I/O bottlenecks present fundamental constraints for analytics on high-volume, high-dimensional datasets. Probabilistic summary structures such as Bloom filters, Cuckoo filters, and their many variants provide essential tools by offering efficient, probabilistic set membership queries while greatly reducing memory consumption [14, 23, 32, 70, 76, 83, 96, 103]. Bloom filters are well-established for basic set membership but lack support for deletions and dynamic adaptability. In contrast, Cuckoo filters improve upon classical Bloom filters by enabling deletions and supporting tunable false positive rates without sacrificing speed or flexibility [32, 103]. Recent advances [83] have removed architectural restrictions in Cuckoo filters—such as the classical power-of-two bucket size constraint and inflexible layouts—by introducing signed-offset addressing and overlapping window designs, thereby further reducing memory overhead and supporting higher loads. These contemporary Cuckoo filter variants provide the smallest memory usage among online-insertion-capable filters for practical false positive rates, maintain fast lookups, and offer flexibility, making them particularly competitive for high-throughput, rapidly evolving analytics domains (such as genomics and real-time analytics) [83].

Compressed indexes represent another key advance, employing succinct data structures, run-length encoding, and grammar-based compression to address the space–time tradeoff across a spectrum of workloads [14, 23, 32, 70, 76, 83]. These indexes are especially advantageous for high-redundancy domains—including version-controlled documents, genomic databases, and large-scale logs—by leveraging innovations such as ILCP arrays, grammar-compressed lists, and run-length encoded Burrows–Wheeler Transforms (RLBWT). For example, the work in [76] shows that using RLBWT for factorization cuts working space down to as little as 1% of dataset size, outperforming uncompressed and basic entropy-compression alternatives by orders of magnitude. Meanwhile, compressed suffix trees enable linear-time LZ77/LZ78 computations in sublinear space for repetitive texts [32]. In the context of graph analytics, succinct q-gram tree indexes [23] allow in-memory similarity queries on up to 25 million chemical graphs by compactly representing occurrence patterns, requiring as little as 5–15% of the memory compared with earlier approaches, and yielding faster searches. Height-optimized tries [14] and adaptive radix trees with incremental cracking [96] extend these compression and efficiency techniques to in-memory indexing for fast analytical workloads, providing a competitive advantage in environments with fluctuating query patterns.

Table 14 explicitly compares the distinctive features, strengths, and adaptability of representative probabilistic and compressed storage structures, highlighting their respective areas of competition and complementarity.

Despite their considerable theoretical benefits, several persistent challenges limit the practical deployment of these data structures. Update costs for many compressed or probabilistic structures can be significant, especially when dynamic workloads require frequent re-encoding or adaptation. The inherent risk of false positives in approximate structures restricts their applicability in mission-critical analytics. Additionally, fully supporting real-time, online compression and adaptable indexing in evolving databases remains an early-stage research area. Adversarial or worst-case input distributions also continue to pose unresolved challenges.

For example, while run-length encoded or grammar-compressed indexes deliver exceptional memory savings and enable efficient retrieval for highly repetitive data [32, 76], their construction and update speeds often lag when incremental or dynamic updates are required. Space-optimal Las Vegas dictionaries [103] demonstrate near-information-theoretic efficiency for static set membership queries but are more challenging to extend to dynamic, real-world contexts and more complex query types. There is a growing need for compressed computation algorithms that can efficiently process data directly in compressed form, especially as summarized in [70].

In summary, this subsection establishes that space-efficient storage structures fundamentally trade accuracy, update flexibility, and implementation complexity for dramatic memory savings, enabling scalable analytics on massive datasets. The main open questions at the leading edge of this field include mitigating update costs, lowering query errors, supporting robust dynamic indexing, and generalizing specialized innovations into unified frameworks for broad and realistic real-world deployment.

Table 14: Comparison of Key Features among Probabilistic and Compressed Storage Structures

Feature	Bloom Filter	Cuckoo Filter	Recent Variants*	Use-case Focus
Supports Deletions	No	Yes	Variant-specific (e.g., some variants introduce efficient deletions [83])	
Tunable False Positive Rate	By design	Tunable	Adaptive/Variable, finer-grained with windowed layouts [83]	
Dynamic Resizing	Limited	Possible	Improved via flexible addressing and layouts [83]	
Bucket Structure	Fixed	Power-of-2 (classical) / Relaxed (modern)	More flexible with signed-offset and overlapping windows [83]	
Query/Update Cost	Low queries, static structure	Low queries/updates, supports deletions	Enhanced flexibility, slightly higher overhead in exchange for adaptability [83]	
Use-case Focus	General set membership	High-throughput, frequent updates	Specialized workloads (e.g., genomics, dynamic logs, analytics)	

*Recent variants refer to structures such as windowed Cuckoo filters, succinct q-gram tree indexes, and advanced compressed indexes as described in [14, 23, 83, 96].

8.3 Hardware and Parallelization for Analytic Scalability

Objectives and Scope. This subsection clarifies how modern hardware, parallelization, and distributed systems are leveraged to realize analytic scalability, with a particular focus on large-scale, privacy-sensitive, or latency-critical applications that extend beyond core computer science areas, such as geospatial data management, healthcare analytics, and IoT environments. Our discussion is intended for readers from both technical and interdisciplinary backgrounds seeking a deeper understanding of how underlying hardware and system-level advancements affect scalable analytics design and its broader implications. We aim to address the guiding question: *How do recent advances in computational architectures, privacy mechanisms, and parallel methods coalesce to support scalable analytics, and what are the key challenges and tradeoffs encountered?*

Achieving analytic scalability necessitates leveraging modern hardware architectures, distributed systems, and parallelization paradigms. The advent of SIMD-capable CPUs and massively parallel GPUs has fostered a rich ecosystem of algorithms and data structures optimized for hardware acceleration. For example, fine-grained parallelization of index search—applied to both inverted and compressed indexes—uncovers that memory access patterns, cache locality, and SIMD-friendly encoding formats are as pivotal to query performance as the index design itself [13, 19, 41, 66, 78, 102]. It has been empirically established that leaving postings lists uncompressed can maximize traversal speeds; however, compression schemes such as QMX and Simple-8b attain comparable throughput while halving memory requirements, thereby offering a favorable tradeoff for search engine workloads [102].

To aid navigation and transparency, as we transition from indexing infrastructure to higher-order analytics, it is important to note that index design and parallelization fundamentally shape performance across both initial data access (e.g., retrieval, similarity search) and advanced analytics (e.g., federated clustering, learning systems). This layered view sets the stage for understanding the interplay of hardware constraints with distributed, privacy-aware analytic processes.

These scalability concerns extend to distributed and federated environments, where sheer data volumes and stringent privacy constraints preclude centralization. Distributed range query indices [93], privacy-preserving federated learning [24, 56], and hybrid consensus protocols for secure retrieval [42] increasingly depend on sophisticated, decentralized approaches. Within federated analytics, local differential privacy (LDP) and secure, multi-level storage enable privacy-preserving computation, maintaining sub-second latency across thousands of distributed clients [56]. Recent innovations such as federated pseudo-sample clustering [85] illustrate that communication-efficient and privacy-preserving analytics are feasible through synergy of local summarization, prototype exchange, and robust central aggregation. These methods not only strengthen privacy guarantees in domains like healthcare or connected vehicles, but also broaden the interdisciplinary reach of scalable analytics.

Optimization is further enhanced through adaptive load balancing and streaming quantization, especially in domains like high-velocity recommender systems. Here, rapid index updates, cluster balancing, and repair mechanisms empower complex multi-task learning in the presence of continual data drift [21]. For example, streaming vector quantization architectures deliver immediate and adaptive candidate retrieval at scale, enabling industry-grade recommender deployments and knowledge extraction in fields such as smart metering, resource forecasting, and consumer analytics [13, 21]. The integrated use of real-time streaming index construction with advanced ranking architectures typifies current directions for scalable, high-throughput analytics.

However, extracting optimal performance from hardware and system resources is challenging. Compressed indexes can induce cache bottlenecks; meanwhile, dynamic, parallel query processing—across both document-at-a-time and term-at-a-time paradigms—demands nuanced orchestration for effectiveness and efficiency [13, 102]. A promising avenue is the adoption of learned index structures and adaptive query execution, which dynamically tailor workload strategies to observed hardware characteristics using predictive models [4, 70, 77, 109]. These strategies draw on compressed computation and subspace analysis, which are increasingly relevant for large-scale applications in science, engineering, and biomedicine [4, 70].

Key Takeaways. Synthesizing developments across hardware acceleration, distributed protocols, and privacy-aware federated analytics, recent directions highlight the necessity of co-optimizing data access, encoding, privacy, and adaptability to underlying hardware profiles. These strategies not only sustain sub-second response in distributed and federated scenarios but also provide robustness to data drift, adversarial conditions, and evolving analytic objectives. The interplay between efficient compression, parallelism, decentralized learning, and privacy-by-design principles defines the state of the art in scalable analytic system design and extends their applicability to broader sectors, including but not limited to geoscience, transit, health, and resource management.

Navigating hardware-level design and parallelization remains central for both foundational infrastructure and higher-level analytics. By contextualizing these developments within broader interdisciplinary applications, this survey aims to provide accessible guidance for diverse readers facing real-world scalability demands.

8.4 Adaptive and Online Index Updating

Objectives, Audience, and Scope: This subsection systematically surveys recent advances in adaptive and online index updating, focusing on the underlying algorithmic frameworks, applicability across data types (relational, high-dimensional, and compressed), and key open challenges for robustness and efficiency under dynamic workloads. This overview is intended to serve both core database systems researchers and practitioners, as well as interdisciplinary readers interested in adaptive data management in fields as diverse as bioinformatics, large-scale text analytics, and real-time data science. Guiding questions include: How have online learning and feedback-driven mechanisms transformed index adaptation? What are the practical and theoretical constraints in supporting continuous updates, especially in compressed or high-dimensional contexts? What distinguishes the latest approaches, and where do substantive open problems remain? By introducing terminology as needed and emphasizing practical relevance, this summary aims to be accessible to a broad technical audience.

To maintain agility in ever-changing analytical environments, index structures must accommodate online, dynamic updates and facilitate autonomous tuning. A significant advancement is the deployment of adaptive and self-tuning indexes, often powered by online machine learning and feedback mechanisms rather than static, manually-tuned configurations. For example, frameworks based on multi-armed bandits and online learning algorithms continually explore different structural configurations, enabling index layouts to converge rapidly toward optimal states. Such frameworks achieve both faster adaptation and improved robustness compared with traditional offline or fixed-configuration approaches [16, 74, 76]. These developments yield substantial improvements in real-world systems: Perera et al. [74] show that multi-armed bandit methods, by directly utilizing workload feedback, yield up to 75% speed-up on shifting/ad-hoc analytical workloads and 59% on HTAP (Hybrid Transactional/Analytical Processing) workloads over static indexes, with convergence guarantees and improved stability over deep reinforcement learning.

This adaptability is increasingly important beyond classical relational systems. In high-dimensional nearest neighbor search, adaptive algorithms iteratively refine cluster assignments, metric selection, and index organization based on actual data variability and performance feedback, helping maintain efficiency even in adversarial or non-stationary settings [60]. Recent surveys such as Mamun et al. [60] provide an explicit taxonomy of learned multi-dimensional indexes, distinguishing them by degrees of adaptivity (e.g., immutable vs. mutable), data layout dynamics, and model retraining support. They also synthesize core characteristics—such as supported query types and model classes—with inline comparative tables, aiding researchers and practitioners in navigating this fast-evolving space. Modern systems like OASIS adapt families of locality-sensitive hash indexes in real time as similarity measures evolve, which is crucial for interactive and dynamically-changing analytic tasks. Key ongoing challenges noted include supporting dynamic workloads, guaranteeing precise error bounds, and ensuring effective concurrency.

Recent work on cracking and incremental index construction—for example in Adaptive Radix Trees (ARTs)—reveals that dynamic, workload-based partial indexing can substantially reduce construction overhead without sacrificing query performance [96]. Wu et al. [96] analyze the cost of ART construction and introduce data-driven partitioning algorithms that support efficient incremental index evolution. These progressive strategies embody a broader shift toward continuously workload-aware and progressive adaptation.

Despite these advances, online updating is especially challenging for compressed or succinct data structures, where balancing, merging, and re-encoding may erode or obscure performance gains [14, 32, 76, 96, 103]. For example, Policriti and Prezza [76] demonstrate that while run-length encoded BWT-based and LZ77 indices dramatically reduce working space for static data (sometimes to 1% of the data size, Table 2 in their work), dynamic variants introduce major speed bottlenecks, with performance lagging far behind static models. Yu [103] introduces optimal succinct static set membership structures with fast queries and minimal space, but extending these benefits to dynamic or online settings remains an open research frontier. Similarly, Boissonnat et al. [16] detail automaton-based compression for topological data and highlight how static structures enable significant compression, but state that efficient dynamic and fully online variants are stymied by deep combinatorial barriers.

Across all methods, ensuring resilience to concept drift, adversarial interactions, and catastrophic forgetting is an enduring challenge—particularly as analytic platforms become increasingly autonomous and must handle unpredictable, high-throughput streams across domains. The development of provably efficient and reliable adaptive algorithms for such settings remains a foundational open problem in both theory and large-scale practice.

In summary, this section has clarified the primary aims of adaptive and online index updating: to enable autonomous, robust, and workload-aware index evolution across a wide array of analytic domains. By synthesizing frameworks extending from online learning and adaptive index construction to multi-dimensional and compressed settings, this survey highlights the state of the art as well as the persistent frontiers—ensuring accessibility for readers from computer science and allied fields. The subsequent sections will examine domain-specific applications and expand on open challenges, further contextualizing these adaptive strategies in pursuit of trustworthy and explainable data analytics.

9 Multiway Data, Tensor Methods, and Higher-Order Analytics

This section presents the foundational concepts and emergent methodologies for analyzing multiway (i.e., multi-dimensional) data using tensor-based approaches. Our primary objectives are: (1) to formally characterize the challenges and opportunities inherent in higher-order data analysis, (2) to survey canonical tensor decompositions and algorithmic structures that underpin modern analytics, and (3) to synthesize recent advances into a unifying framework that distinguishes this survey from prior overviews.

This section is intended to be accessible to a broad academic and professional audience, including both researchers and practitioners across data science, engineering, and interdisciplinary domains. Readers with backgrounds in computer science, statistics, applied mathematics, or information sciences will find the exposition approachable, while illustrative examples will facilitate engagement for those less familiar with advanced algebraic concepts.

We seek to address the following guiding questions: How do tensors expand the representational and computational scope relative to matrix-based methods? Which classes of tensor decomposition offer scalable modeling advantages in high-dimensional data settings? What are the primary analytic, computational, and practical considerations when deploying such methods across real-world domains? Finally, how does the structure of this survey differ from previous work, and what new taxonomy or organizational principle do we propose for presenting the state of the field?

To facilitate reader orientation, each major subsection will open with a statement of objectives and thematic focus. For clarity, denser topics are divided into more focused subsections, allowing readers to navigate foundational principles, indexing infrastructures, and higher-order analytics with ease.

The narrative flows from data-centric indexing and representation issues—where we explore the organization and storage of high-dimensional arrays—into algorithmic and theoretical analyses concerning tensor decompositions and their applications. Brief transitional commentary is included when moving from the foundational topic of indexing infrastructure to advanced analytical techniques, ensuring coherence for readers transitioning between practical data handling and abstract algorithmic frameworks.

Throughout, we synthesize technical advances and incorporate illustrative, practical examples that clarify key multidimensional concepts for an interdisciplinary readership. In addition to core technological perspectives relevant to computer science, we lightly expand on broader applications and implications, including scientific computing, bioinformatics, neuroscience, chemometrics, and social network analysis, thereby underscoring the practical significance of tensor methods across diverse research fields.

At the conclusion of this section, we summarize the central takeaways and outline open challenges at the intersection of multidimensional data representation, scalable analytic algorithm design, and interdisciplinary applications. In synthesizing recent work, we explicitly contextualize our proposed taxonomy and organizational framework in relation to existing surveys, highlighting both continuities and novel distinctions in structure and conceptual emphasis. This approach aims to provide a comprehensive yet cohesive perspective on higher-order analytics with tensors, establishing a structural and conceptual framework that underscores the novel contributions of this survey.

9.1 Prevalence and Application Areas

The rapid expansion of high-dimensional, multi-modal data in scientific and engineering disciplines has driven the extensive adoption of tensor-based methods for advanced data modeling and analysis. In contrast to conventional matrix-based techniques, tensor methodologies are specifically designed to preserve and leverage the intrinsic multiway structure characteristic of contemporary datasets. These datasets, arising from domains such as biomedical imaging (for example, functional MRI or hyperspectral imaging), temporal-spatial time series (including climate models and multi-channel EEG), and complex networked systems (such as multi-relational biological interactions or dynamic social networks) [9], often contain interrelations spanning more than two modes. By exploiting this higher-order structure, tensor models enable richer, more expressive representations of data, thereby uncovering multivariate interactions beyond the scope of pairwise (matrix) approaches.

For instance, in imaging science, tensors can simultaneously encode spatial, temporal, and spectral dimensions. Similarly, in network analysis, hypergraph analogs of tensors facilitate the study of multi-entity relationships, significantly advancing the analytical depth achievable in fields such as genomics and chemometrics [9]. This inherent capacity of tensor methods to capture and model complex relationships underscores the imperative for robust analytical frameworks capable of scaling with and adapting to the escalating complexity of contemporary scientific datasets.

9.2 Tensor Decompositions and Higher-Order Methods

At the core of multiway analytics are tensor decomposition techniques, which extend the principles of matrix factorization into higher orders and enable the discovery of latent structures embedded within complex datasets. Among these, the Canonical Polyadic (CP) and Tucker decompositions are foundational. The CP decomposition represents a tensor as a sum of rank-one components, furnishing interpretable multiway analogs to singular vectors, while the Tucker decomposition generalizes principal component analysis (PCA) to encompass multiple modes by extracting interactions through a core tensor and orthonormal factor matrices [9].

Addressing the inherent nonconvexity and computational complexity of these decompositions, recent algorithmic advances employ strategies such as alternating least squares, gradient-based optimization, and stochastic techniques. These methods capitalize on problem-specific structures and incorporate sophisticated initialization procedures, thereby enhancing convergence properties and robustness to noise.

In addition to classical decompositions, contemporary research has expanded the scope of tensor analytics through higher-order statistical techniques, including tensor singular value decomposition (tensor-SVD), multiway PCA, and independent component analysis (ICA). Each of these frameworks brings distinct advantages for source separation and dimensionality reduction in tensor-formatted data [9]. Furthermore, novel mixture modeling and multi-mode regression approaches have been formulated within the tensor paradigm, empowering researchers to construct expressive models tailored to heterogeneous and structured data streams.

A particularly active research area involves tensor completion and recovery, where the objective is to impute missing entries by leveraging low-rank or structured sparsity assumptions. Such methods are critical for real-world scenarios where datasets are often incomplete or partially observed. While a rich variety of algorithms has emerged, all must contend with the significant challenges imposed by the "curse of dimensionality" and the absence of straightforward low-rank characterizations—factors that make the tensor setting fundamentally more complex than the matrix case.

As shown in Table 15, each decomposition method offers unique trade-offs in terms of modeling capabilities and application suitability within multiway data analysis.

Table 15: Comparison of Core Tensor Decomposition Techniques

Decomposition	Core Idea	Advantages / Typical Use Cases
CP Decomposition	Expresses tensor as a sum of rank-one components.	Interpretability, identifies latent factors, applicable in signal processing and topic modeling.
Tucker Decomposition	Generalizes PCA to multiway data, yielding a core tensor and factor matrices.	Captures interactions between modes; flexibility in modeling mode-specific variances; used in compression and feature extraction.
Tensor-SVD	Generalizes SVD to tensors via multi-linear operations.	Enables robust dimensionality reduction and source separation; effective for multi-modal signal processing.

9.3 Complexity and Open Challenges

Despite their substantial potential, tensor methods are accompanied by formidable analytical and computational challenges that stem from fundamental aspects of complexity theory and high-dimensional statistics. Notably, unlike matrices, tensors may not possess best low-rank approximations—a phenomenon posing significant obstacles to the design of optimal decomposition algorithms. Central analytical tasks such as low-rank tensor decomposition and rank determination have been shown to be NP-hard in the general case, establishing fundamental barriers for scalable computation [9]. This hardness sharply delineates the limits of what can be achieved algorithmically, especially in large-scale or high-noise data regimes.

A prominent issue is the disparity between what is statistically or information-theoretically achievable, and what current algorithms can compute efficiently. Even when estimators exist with theoretically optimal statistical guarantees, known algorithms may fail to realize these estimates within practical timeframes due to issues such as nonconvexity and local minima.

Recent research integrates tools from optimization, statistics, and numerical linear algebra to navigate these trade-offs. Methods for tensor decompositions, including CP and Tucker, as well as alternating minimization and gradient-based algorithms, have been developed to address the challenges of nonconvexity and initialization. There are continued efforts to refine our understanding of sample complexity bounds, convergence rates, and error profiles associated with various algorithms in tensor analysis [9]. Nevertheless, empirical performance often lags behind theoretical guarantees, sometimes requiring impractically large datasets or suffering from susceptibility to poor local optima.

Algorithmic frameworks currently used for key tasks such as clustering or indexing on tensor data often remain rigid and do not scale efficiently, limiting their integration into practical analytical pipelines [9]. Progress remains especially urgent on several open problems: constructing algorithms that jointly achieve statistical optimality and computational tractability for high-dimensional and high-order tensors; designing robust initialization and regularization methods tailored to tensor models; and advancing clustering and indexing techniques that exploit the inherent multiway structure of tensor data.

Addressing these open challenges is pivotal to fully realizing the analytical capabilities of tensor and higher-order methods. Success in these areas will have substantial implications for applications in diverse scientific and engineering fields [9].

10 Applications and Deployment Strategies

This section aims to systematically explore the diverse applications and deployment strategies of the methods surveyed in this work. Our primary objective is to identify measurable criteria for successful deployment, distill key deployment patterns, and elucidate how emergent techniques are being adapted across different domains. By synthesizing prevailing approaches and discussing practical considerations, we offer a comprehensive reference for both researchers and practitioners seeking to operationalize these technologies.

Explicit Objectives: In this section, we (1) categorize major application domains by their operational constraints, (2) compare the leading deployment strategies across these domains in terms of latency, privacy, and scalability, and (3) provide a unique taxonomy for broader comparative evaluation. This addresses the need for measurable objectives and structured comparison, as highlighted in the introduction.

We begin by defining the scope of application areas, clarifying the specific characteristics and requirements each domain imposes on deployment design. For clarity, specialized methods—such as *federated learning*[1], *edge inference*[2], and *differential privacy*[3]—are defined with contextual footnotes inline.

We then examine prevailing deployment strategies, highlighting integration workflows and comparing operational trade-offs—such as throughput, resource utilization, and adaptability—in direct relation to common deployment environments. This is summarized in Table 16, which enables at-a-glance digestibility of strategies as observed in recent literature.

To enhance clarity for readers with differing technical backgrounds, brief illustrative examples are provided alongside complex multidimensional concepts. All references are cited within the context of their respective domains for improved traceability.

Transitioning from foundational domains to more specialized applications, we maintain narrative continuity and highlight thematic connections. Our analysis synthesizes and extends previous surveys: whereas earlier works often catalog domains independently, here we propose a framework that categorizes deployment strategies by their underlying operational constraints (such as latency tolerance, data privacy, and scalability), offering a novel lens for comparative evaluation. Throughout, we further distinguish our contribution by including direct comparisons to recent reviews and demarcating our novel taxonomy and commentary.

We conclude by providing extended commentary on how the proposed taxonomy enables future research: by explicating the operational drivers of deployment strategies and mapping them to both foundational and emerging domains, this work lays a structured foundation for comparative studies spanning new areas of application.

Key takeaways: By systematically categorizing application areas and deployment frameworks, and by presenting a comparative summary (Table 16), this section offers measurable practical

[1] Federated learning enables distributed training of models across multiple devices or organizations without centralizing data, thereby preserving privacy.

[2] Edge inference refers to the execution of model predictions on resource-constrained edge devices, improving latency and reducing bandwidth demand.

[3] Differential privacy is a mathematical guarantee for privacy, ensuring that analysis results do not reveal information about individual data points.

Table 16: Comparison of Deployment Strategies Across Application Domains

Application Domain	Common Deployment Strategy	Operational Constraints	Typical Workflow	Notable References (Year)
Healthcare	Federated Learning	High privacy, moderate latency	Distributed training, on-device inference	[49] (2022), [15] (2023)
Autonomous Vehicles	Edge Inference	Ultra-low latency, real-time req.	On-device prediction, periodic cloud sync	[92] (2023), [37] (2022)
Finance	Centralized/Hybrid	Data compliance, high accuracy	Secure aggregation, hybrid cloud-edge deployment	[90] (2022)
Smart Cities	Distributed/Edge	Scalability, heterogeneity	Edge device aggregation, hierarchical coordination	[79] (2023)
E-commerce	Centralized Cloud/Hybrid	High throughput, dynamic scale	Centralized model hosting, elastic resource scaling	[98] (2022)

guidance for real-world adoption in varied operational environments. The synthesized taxonomy establishes a basis for future comparative research, emphasizing both commonalities and unique requirements across diverse deployment scenarios.

10.1 Application Domains and Case Studies

In recent years, state-of-the-art methods for clustering, indexing, and analytics have been deployed across a wide spectrum of scientific and industrial domains. This proliferation attests not only to the versatility of these techniques but also to the complexity inherent in their large-scale application.

In the fields of genomics and transcriptomics, advanced methodologies such as ensemble subspace regression and penalized mixed models have become instrumental. These tools elucidate molecular subtypes and latent structures within high-dimensional sequencing datasets by effectively balancing interpretability, predictive accuracy, and statistical rigor. Notably, ensemble regression techniques confer robust alternatives to classical penalized models, especially where the dimensionality far exceeds available observations, such as in gene expression biomarker discovery. Here, aggregation across random subspaces mitigates tuning sensitivity and overfitting tendencies [16, 36]. High-dimensional mixed-effects frameworks—augmented with sparsity-inducing penalties such as the smoothly clipped absolute deviation (SCAD)—have further advanced feature selection and inference, particularly within compositional microbiome investigations and genome-wide association study (GWAS) designs. Compared to traditional LASSO approaches, these methods offer superior performance amid clustered or highly correlated predictors [103].

Neuroimaging research, dealing with inherently multiway (tensor) data structures, has seen significant uptake of tensor-based clustering models. By exploiting separable covariance structures, these models enable both computational efficiency and scientific interpretability. The tensor normal mixture model, integrating sparsity-enforcing penalties with customized expectation-maximization procedures, exemplifies this paradigm: it delivers state-of-the-art performance on large neuroimaging datasets while providing principled quantification of cluster uncertainty and sensitivity to initialization [67]. Complementary clusterability diagnostics, grounded in multimodality analyses, serve as robust guides for assessing the intrinsic tendency for cluster formation—thereby cautioning against exclusive reliance on traditional, noise-sensitive internal indices [59].

Text analytics and the digital humanities benefit from innovations in indexing and data compression. Methods that leverage the repetitive structure of large textual corpora—such as run-length Burrows-Wheeler Transform (BWT)-based LZ77 factorization and succinct membership data structures—substantially reduce memory consumption and computational demands, thus enabling scalable solutions in digital numismatics, linguistics, and large-scale search applications [35, 66, 100]. In chemical informatics, graph-based indexing strategies (e.g., for PubChem-scale datasets) combine hybrid encoding and succinct filtering to achieve notable space reductions and rapid query operations, even in the presence of millions of diverse molecular graphs [39, 104].

In the financial and social sciences, unsupervised learning approaches such as clustering uncover nuanced subpopulations and latent biases that traditional demographic or regression-based analyses may overlook. Notably, large-scale application of K-means clustering to financial wellbeing surveys has revealed patterns—such as explicit mismatches between subjective and objective financial stability—that challenge prevailing assumptions. These findings highlight both methodological opportunities for more informative clustering objectives and the need for mixed-model frameworks to disentangle complex, overlapping constructs [1].

Methodological advances in environmental analytics, EEG/gene clustering, and chemical informatics have been closely tied to the advent of scalable, distributed, and federated analytics platforms. For example, distributed nearest-neighbor systems utilizing Apache Flink, with domain-specific space-filling curve partitioning and granularity-aware load balancing, have enabled efficient analysis of granular smart meter or environmental sensor data—offering superior wall-clock performance relative to traditional central paradigms [76]. Similarly, approximate nearest neighbor search in high-dimensional chemical or image repositories increasingly employs graph-regularized sparse coding and quantization to reconcile recall, speed, and storage footprint [5, 50, 105].

Emerging domains, such as single-cell transcriptomics and clinical subtyping (e.g., diabetes), have driven the adaptation of techniques like generalized contrastive principal component analysis and mixed-membership modeling. Designed to decouple technical artifacts from biological signals, these frameworks produce interpretable axes of variation and robust unsupervised stratification, supporting research in heterogeneous and high-noise environments [32, 110].

The evolution of large-scale algorithms and data structures is intimately linked with augmented capabilities in massive data and graph indexing. As datasets increasingly exceed main memory capacity, techniques including dynamic polygon nearest-neighbor search, adaptive radix trees, voxelized spatial representations, and automaton-based simplex complex compression are indispensable for real-time analytics within both static and dynamic contexts [60, 75, 96, 102]. Current research in multidimensional learned indexes, database cracking, and compressed or low-footprint computation further underscores a dynamic field, where algorithmic, statistical, and hardware constraints motivate the development of novel theoretical models and practical open-source implementations [4, 14, 44, 58].

10.2 Large-Scale Deployments and Federated Analytics

As we transition from domain-specific method reviews to crosscutting themes, this section aims to explicitly address the core objectives of the survey: to synthesize emerging technical solutions for scalable, trustworthy analytics and to critically examine the operational and methodological barriers to their deployment. Our focus here is to bridge foundational advances in clustering, indexing, and federated learning with the realities of implementing such tools across diverse institutional settings.

Scaling advanced analytics from domain research to operational deployment introduces both computational and institutional challenges. Federated analytics and privacy-preserving clustering are of growing significance for applications in which data are distributed across independent institutions or geographical zones, subject to legal and governance restrictions on access and sharing. In these contexts, the use of open-source libraries and reproducible workflows is not only best practice, but often essential for enabling trustworthy, cross-institutional scientific collaboration [57].

Deployments at scale typically require algorithms for clustering, indexing, and spatial or graph analysis to function efficiently in distributed or parallelized environments. This demands an intricate balancing act between accuracy, processing speed, and memory resource usage. Empirical benchmarking of open-source range query and graph indexing libraries for high-performance computing has highlighted the importance of context-specific profiling—considering build time, query performance, and memory scaling—as well as the limitations of universal, "one-size-fits-all" strategies. Notably, brute-force or hybrid approaches sometimes demonstrate superior performance over more complex alternatives when operational data fall outside nominal parameter regimes [112]. However, these brute-force methods, while robust, can suffer from prohibitive computational cost and poor scalability in high-dimensional or adversarially noisy settings, whereas some sophisticated algorithms may fail to generalize or degrade sharply when input distributions deviate from assumed models [57, 112]. Failure to account for real-world heterogeneity can thus undermine the reliability of both naive and advanced techniques.

Federated learning introduces additional considerations, including statistical heterogeneity, communication overhead, and privacy preservation. While probabilistic model aggregation and distributed subspace consensus mechanisms have been developed to support inference across disparate data sources, these often face drawbacks: for instance, model drift from non-i.i.d. data, increased synchronization latency, and persistent privacy leakage risks if local updates are insufficiently protected [70].

Crucially, the assurance of reproducibility and the broad dissemination of open-source software, workflow templates, and standardized datasets underpin scientific trust, algorithmic benchmarking, and iterative methodological improvement. Practices such as explicit reporting of statistical validation, computational requirements, and parameter sensitivity facilitate fair comparisons and spur innovation across domains [57]. To ensure transparent evaluation, practitioners should routinely disclose not just strengths but also failure cases and domain-specific limitations of deployed algorithms.

In summary, large-scale deployments require both methodological innovation and transparent, reproducible engineering to overcome the inherent drawbacks of even the most advanced technical approaches. This crosscutting perspective sets the stage for our subsequent detailed survey of federated analytics methods.

10.3 Guidelines for Deployment

The objective of this section is to clearly articulate key practical recommendations for reliably deploying analytics solutions on complex, heterogeneous datasets, while highlighting both strengths and notable limitations of current approaches.

Extracting valid scientific and operational insights from complex, heterogeneous datasets requires adhering to principled standards for automation, benchmarking, and statistical validation. Recommendations, along with brief commentary on notable challenges or failure cases, are as follows:

Automation: Streamlining feature preprocessing, model selection, and parameter tuning can enable scalable workflows and help maintain interpretability and domain relevance. However, over-automation may obscure crucial data-specific nuances and introduce risks where model outputs become less transparent or less aligned with evolving domain requirements.

Benchmarking: Comprehensive benchmarking across diverse datasets and operational conditions, leveraging both internal and external evaluation indices, sensitivity analyses, and simulation-based studies, is essential for assessing clusterability, validity, and model robustness [112]. Yet, benchmarking may overlook certain failure cases, such as poor generalization across highly heterogeneous environments or underperformance on rare subpopulations not reflected in standard benchmarks.

Statistical Validation: Rigorous clusterability diagnostics and out-of-sample validation are particularly important in high-noise or high-dimensional environments to avoid spurious discoveries. Nonetheless, in practice, such validation can be challenged by limited access to labeled data, especially in domains with rare ground-truth or highly unbalanced classes, thus impeding the accurate assessment of model performance.

Transparency and Reproducibility: Transparent algorithmic reporting and open-source implementations, alongside publishing benchmarks and reproducible code and workflows, are vital for scientific rigor and collaborative development [57]. These practices, however, can be limited by proprietary data, commercial toolchains, or privacy constraints, which may reduce reproducibility in certain real-world scenarios.

Scalability: Algorithmic efficiency, memory optimization, and distributed computation must be prioritized. Resource-aware methods—including compressed computation and dynamic data structures specifically designed for operating directly on compressed representations [70], as well as federated analytics—are increasingly necessary for managing large-scale, heterogeneous datasets. Developments in methods such as ensemble subspace clustering [57] and consensus spectral clustering have improved scalability by using parallelizable structures and robust aggregation; nonetheless, these approaches can remain computationally demanding compared to simpler models and may encounter efficiency limits imposed by hardware or network constraints in very large-scale deployments.

Interpretability and Ethics: Ensuring model interpretability, transparent feature selection, and fairness auditing is essential, especially in sensitive biomedical, environmental, or social contexts. Despite ongoing efforts, interpretability and fairness auditing can be hampered by model complexity, latent variable effects, and the absence of reliable fairness metrics tailored to all data domains.

These balanced practices collectively support the development of robust, trustworthy, and adaptive analytics pipelines, while underscoring areas where continued research and practical vigilance are necessary to ensure effective real-world analytics in the face of evolving data challenges.

10.4 Comparison of Representative Large-Scale Deployment Strategies

This section aims to explicitly outline the objectives and provide a structured, critical comparison of deployment strategies employed in high-dimensional and distributed analytics. Our goals are to clarify the trade-offs inherent in each approach, spotlighting not only their advantages but also notable drawbacks and representative scenarios where they may falter or fail.

The successful deployment of large-scale clustering and analytics methods thus hinges on careful alignment between methodological strengths and the practical realities posed by domain data, workflow constraints, and institutional context. While the approaches summarized above offer robust solutions within certain bounds, practitioners must remain vigilant regarding scaling ceilings, performance regressions, or failure cases as systems and data evolve. Continuing advancements in open-source dissemination, standardized evaluation, and adaptive algorithmic design will catalyze further innovation and responsible adoption across scientific and industry domains.

11 Crosscutting Themes, Challenges, and Emerging Research Directions

This section aims to explicitly outline the overarching objectives of our analysis: to identify recurring themes, articulate primary challenges, and highlight promising avenues for future research within the surveyed domain. In doing so, we not only underscore the strengths of key approaches but also provide brief, contrasting commentary on their notable drawbacks or documented failure cases, thus ensuring a balanced perspective.

Throughout our synthesis, we observe that while many approaches demonstrate impressive advancements—such as improved scalability and adaptability—certain limitations consistently emerge. For instance, methods that prioritize efficiency may inadvertently compromise robustness or generalizability, often failing in scenarios with high data variability or adversarial conditions. Conversely, models offering high flexibility sometimes incur prohibitive computational costs, impeding their practical deployability.

Additionally, a recurring challenge involves the standardization and interoperability of developed models across diverse application settings. Integration issues and lack of benchmark datasets further compound these obstacles, limiting meaningful performance comparisons and slowing community-wide progress.

By systematically contrasting advantages with corresponding limitations, this section provides deeper clarity on critical research bottlenecks and suggests focal points for forthcoming investigations. All references have been formatted consistently to enhance the professionalism and readability of this synthesis.

11.1 Integration and Adaptivity

The escalating volume, complexity, and heterogeneity of contemporary data have accentuated the necessity for adaptive and integrative systems across indexing, clustering, feature selection, similarity search, and statistical modeling. A pronounced trend has emerged toward the unification of methodologies traditionally addressed in isolation, including but not limited to the joint handling of clustering and feature selection, spatial and graph indexing, learned and annotative indices, and adaptive tensor models [7, 30, 33, 53, 68, 69, 77, 101, 106, 111]. This movement is primarily motivated by empirical limitations observed in “one-size-fits-all” techniques, which become inadequate as data increases in dimensionality, dynamism, or semantic richness.

For instance, hybrid paradigms combining prototype reduction with learned dimensionality compression empower k-NN search to achieve significant gains in speed and accuracy. Recent studies have demonstrated that convex hull selection, stratified sampling, principal component analysis, and auto-encoder-based representations can deliver classification speed-ups of up to 32×, often with minimal or even improved accuracy, but these approaches may be affected by data overlap and class imbalance, requiring flexible representation selection and dynamic parameter tuning [8, 38, 68, 75, 88, 99]. Work on exact kNN methods has further highlighted the effectiveness of hybrid and ensemble approaches, showing, for example, that variants leveraging ensemble strategies achieve robust, domain-adaptable performance, especially for complex and high-dimensional data [8, 38, 88]. Addressing class overlap and imbalance simultaneously, recent kNN models employ composite weighting schemes, enhancing reliability and robustness across imbalanced and noisy datasets without parameter tuning [8].

Similarly, the integration of feature selection and clustering—especially for mixed-type or high-dimensional datasets—exploits joint optimization and ensemble techniques to reinforce cluster robustness and enhance attribute discrimination, even under adverse conditions such as adversarial noise or low signal-to-noise ratios [73, 93]. Advances in deep clustering link feature learning and cluster assignment in unified or iterative frameworks, improving adaptability to diverse data domains, while partitional and hierarchical methods remain essential for balancing accuracy, efficiency, and interpretability [73, 111]. The importance of evaluating clustering effectiveness via multiple internal and external metrics and accommodating varied evaluation methodologies has also been emphasized in recent surveys [73, 101, 111].

Tensor-based modeling constitutes another pivotal frontier in integrative analytics, offering interpretable and scalable substrates

Table 17: Comparison of large-scale deployment strategies for clustering and analytics.

Strategy	Advantages	Constraints / Notable Drawbacks	Application Examples
Distributed parallel analytics (e.g., Apache Flink, Spark)	High scalability; fault tolerance; supports massive input volumes	Requires significant infrastructure setup; can introduce resource contention; may need complex partitioning for optimal performance; failure scenarios include network or node bottlenecks impacting throughput	Smart grid analytics, environmental sensor networks
Federated clustering/learning	Preserves privacy; data remains local; enables cross-institutional collaboration	High communication costs; statistical heterogeneity can degrade model convergence; aggregation becomes complex with divergent local distributions; failure if local sites cannot synchronize updates	Multi-center biomedical studies, cross-jurisdictional finance
Graph-based indexing with hybrid encoding	Space-efficient; supports rapid queries over large, diverse graphs	Computationally expensive index construction; sensitive to parameter tuning; may perform poorly with highly dynamic or evolving graphs; failure can occur when structure does not generalize	Chemical informatics, PubChem-scale search, social network mining
Compressed/learned data structures	Drastically reduced memory footprint; competitive accuracy	Implementation complexity; can be highly sensitive to parameter selection; degradation in accuracy if compression loses key features; may struggle with continual updates or streaming data	Text analytics, high-throughput genomics, image retrieval
Centralized brute-force/hybrid approaches	Simplicity; robust against certain data irregularities; minimal tuning	Does not scale to massive datasets; high per-node resource demand; often fails on data with high dimensionality or volume (memory/compute limits)	Small- to medium-size or irregular dataset scenarios

for multiway data prevalent in scientific and engineering applications. Penalized tensor mixture models and scalable decomposition algorithms have been developed to reconcile statistical consistency in clustering with computational scalability, particularly in high-dimensional scenarios [38, 88, 91]. Furthermore, manifold learning perspectives and nonlinear representation approaches offer additional capabilities for capturing intricate high-dimensional structures and heterogeneous experimental conditions; however, these advancements demand stronger theoretical guarantees and enhanced adaptivity at scale [8, 16, 28].

The convergence of indexing paradigms—most notably through annotative and learned indexing—has yielded robust frameworks for unified, scalable data platforms. Annotative indexes generalize over inverted, columnar, and graph-based strategies, supporting transactional, concurrent, and semi-structured workloads, alongside complex knowledge graph scenarios [25, 26, 60]. Recent work provides a detailed taxonomy for classifying learned indexes across dimensions such as design, mutability, data layout, insertion strategy, and supported query types, and outlines open challenges including precise error bounds, efficient (re-)training, concurrency, and GPU acceleration [60]. Annotative indexing further supports ACID-compliant, transactional ingestion and expressive SQL-like querying over heterogeneous JSON data with high concurrency, showing performance advantages in dynamic, large-scale environments and enabling unified management of textual, structured, and vector/graph data [26]. Lightweight distributed learned indices for big spatial data, such as LiLIS, demonstrate dramatic improvements in query speed and index construction, coupled with robust scalability for diverse spatial workloads within existing big data platforms like Spark [25]. Likewise, proximity graph-based frameworks like UNIFY deliver efficient, range-filtered approximate nearest neighbor search with integrated hybrid filtering and automatic strategy selection, optimizing for high scalability and recall in attribute-constrained searches [53]. Such frameworks have been demonstrated to outperform traditional approaches in both efficiency and flexibility, positioning annotative and learned indexes as central solutions for modern adaptive data management systems [25, 26, 43, 60, 97].

11.2 Machine Learning for Index and Analytic Optimization

This section aims to clarify the concrete research objectives and contributions of machine learning for index and analytic optimization, focusing on measurable advances such as the development of dynamic, workload-aware index structures, unifying frameworks, and analytic guarantees in performance and adaptability. Unlike previous surveys, recent scholarship introduces frameworks that treat index management as an online learning and decision-making problem, with explicit attention to robustness, adaptability to multi-dimensional and hybrid workloads, and error-bounded performance [25, 60]. This refines the objectives of the field from static, manual index construction to automated, adaptive, and theoretically analyzable solutions.

Rather than depending exclusively on hand-crafted heuristics or costly offline tuning, new methodologies leverage workload observation and cost feedback to perpetually adapt index schemas [25, 60]. Noteworthy progress has emerged in resource-efficient recommendation systems utilizing large language models (LLMs) to synthesize features of workloads and infer optimal index strategies with minimal retraining or DBA input. These systems commonly integrate demonstration pools, scalable inference engines, and domain knowledge injection to achieve high recommendation accuracy and low latency, rivaling or exceeding the best conventional index advisors [60, 64]. Key innovations distinguishing these approaches include: formulating index optimization tasks for compatibility with few-shot or in-context learning paradigms; extracting granular workload statistics to inform schema selection; and deploying scalable, aggregation-based mechanisms to ensure robustness and efficiency.

Online learning frameworks inspired by bandit algorithms further extend the paradigm by eliminating dependencies on DBA expertise or traditional query optimizers. These approaches adopt active exploration and exploitation strategies, evaluating alternative indexes based on real-time performance metrics. Notably, they offer theoretical guarantees on convergence to near-optimal configurations under dynamic or evolving workloads, frequently surpassing the adaptability and efficiency of deep reinforcement learning and static analytics methods [60]. Despite these advances, ongoing challenges remain: more sophisticated or hybrid workloads demand greater model expressiveness and faster adaptation, and tight analytic guarantees such as error bounds in higher dimensions are active research topics [60, 64].

For example, recent frameworks such as LiLIS implement lightweight, distributed learned indexes for spatial and multi-dimensional contexts, natively within big data ecosystems like Spark [25]. These systems apply error-bounded spline-based models and space-filling curves, supporting a range of query types efficiently while remaining robust to data partitioning strategies and query skewness. Empirical studies demonstrate that such models can outperform traditional spatial indexes by orders of magnitude in both query speed and index construction time, and retain compatibility with standard analytics APIs [25].

This body of research unifies previously disparate strands—adaptive index learning, analytic performance guarantees, and scalable big data deployment—into a cohesive framework, highlighting both practical advances and remaining open challenges for future investigation [25, 60, 64]. All cited references are current as of 2024 and fully traceable.

Table 18: Comparative Query Latency of Distributed Learned and Traditional Spatial Indexes [25]

Method	Point Query (ms)	Range Query (ms)	kNN Query (ms)	Join Query (ms)
LiLIS-K	82.59	468.64	650.2	228,581
Sedona-RK	(much slower)	(much slower)	790,993	(much slower)

11.3 Transactional and Distributed Perspectives

This section aims to clarify key research objectives by unifying transactional and distributed analytic perspectives across heterogeneous and federated environments. Our focus is to (1) identify actionable integration points between emerging indexing paradigms, transactional guarantees, and adaptive distributed architectures, and (2) articulate measurable directions for optimizing analytic performance and consistency under multimodal, cross-system workloads. Distinct from previous surveys, we emphasize a framework that connects recent advances in learned, annotative, and automaton-based indexing structures, highlighting how these approaches reshape scalability and ACID compliance beyond legacy architectures.

Recent database management and analytic systems must robustly execute distributed queries with strong ACID guarantees, reflecting the necessity for cross-engine and federated access in vector, graph, and hybrid analytic workloads [60, 76, 83]. Modern systems are tasked with seamlessly integrating graph databases, knowledge graphs, and spatial or textual search engines, while upholding high standards for performance and correctness [16, 26, 60]. Our survey differentiates itself by focusing on the convergence of these requirements, and the unifying frameworks now underlying them.

Advances include learned multi-dimensional indexing [60], which leverages machine learning to address adaptive partitioning, workload concurrency, and robustness, and efficient compressed or automaton-based structures for specialized backend integration [16]. Annotative indexing [26] further generalizes this trend, providing a schema-flexible, transactionally robust core for managing both structured and semi-structured data at scale. These frameworks increasingly unify inverted indexes, object stores, and graph paradigms while transparently supporting hybrid queries.

Distributed query execution is made viable by partitioned, dynamic system architectures and innovative data structures, each choice entailing trade-offs among communication cost, consistency, and optimization for privacy-aware or partially accessible data [2, 10, 32, 76, 103]. Federated analytics thus face the challenge of maximizing openness and collaboration while guaranteeing security, transactional integrity, and controllable latency across geographically distributed infrastructures [22, 39, 50].

The rising importance of ACID properties within federated and multimodal systems reflects a crucial trend: transactional guarantees are no longer isolated requirements but are interwoven with scalable, adaptive analytic environments [25, 26, 60]. Notably, systems like LiLIS [25] exemplify this shift by offering distributed learned indexes that combine partition-aware error-bounded models with efficient ACID-compliant spatial data services, demonstrating measurable gains in query throughput and index construction efficiency compared to traditional approaches.

In conclusion, the present section provides an up-to-date synthesis of how transactional and distributed perspectives are unified in current systems. We explicitly characterize recent frameworks that bridge learned, automaton-based, and annotative indexing, articulating concrete research objectives for optimizing scale, adaptability, and correctness across federated multimodal environments. This perspective distinguishes our survey from prior work by connecting system-level guarantees to emergent algorithmic and architectural advances, as documented in the latest literature.

11.4 Robustness and Adversarial Resilience

Robustness and adversarial resilience have become focal research directions as indexing and analytics frameworks extend into sensitive domains such as healthcare, finance, and security. This section aims to clarify: (1) the explicit vulnerabilities exhibited by indexing and analytic models in high-dimensional, graph-structured, and compressed data settings; (2) measurable goals, such as minimizing adversarial query success rates and maximizing index integrity under attack scenarios; and (3) the state-of-the-art in unifying algorithmic frameworks delivering analytic guarantees on robustness.

Recent surveys and studies demonstrate that manipulative adversaries can significantly degrade the effectiveness of systems that employ graph-based, tensor, or compressed representations for indexing and analytics [9, 58, 60, 70]. For instance, [9] details how tensor analytic algorithms—including tensor decomposition and multiway PCA—are especially susceptible to adversarial and stochastic perturbations, and analyzes theoretical sample complexity and error bounds under such challenges. In the context of graph-structured search, [58] presents competitive ratio guarantees for partial order multiway search in directed acyclic graphs (DAGs), providing rigorous bounds ($O(\log n)$ times the optimal) even under adversarial target placement, thus quantifying worst-case search robustness.

With respect to index architectures, [60] surveys learned multi-dimensional indexing schemes, cataloging open research challenges including the precise formulation of error bounds and the quantification of robustness against adversarial attacks (see inline table summaries therein). Notably, they propose taxonomy-driven frameworks that separate index learning from learned model indexing, facilitating systematic benchmarking and ongoing progress toward robust index design. On the compression front, [70] articulates how compressed computation redefines algorithmic trade-offs, bringing efficiency and resilience into focus when conventional, uncompressed algorithms break down due to scale.

A key unification emerging in current literature is the pursuit of analytic guarantees—such as query complexity bounds, error bounds under adversarial noise, and statistical-computational trade-off frontiers—across diverse data modalities. Although no single general-purpose robust analytic framework supersedes all others, progress is apparent in the transfer of worst-case analysis techniques (e.g., competitive ratio, sample complexity) across graphs, tensors, and compressed representations. Typical research objectives now involve explicit maximization of adversarial robustness while limiting system overhead, establishing a balanced triad: privacy, resilience, and efficiency.

Despite these advances, core open challenges persist. Current algorithms providing analytic guarantees on robustness frequently incur high storage, computation, or training costs, and generalizing robust methods across modalities and dynamic workloads remains unresolved [9, 58, 60, 70]. Furthermore, the establishment of universally accepted robustness benchmarks and the consistent traceability of cited results are ongoing priorities for the community.

In summary, this survey extends previous works by explicitly mapping formal robustness objectives to analytic frameworks across studied data types, highlighting measurable goals and drawing on recent results that quantify adversarial resilience. It integrates and cross-compares guarantees from the tensor, graph, and learned index literature, providing a more unified perspective on the analytic frontiers and research priorities in robust and resilient data analytics.

11.5 Online, Adaptive, and Learned Indexing for Dynamic Workloads

Contemporary workloads, characterized by rapid streaming, immense scale, and frequent evolution, elevate the necessity for indexing systems that adapt online and dynamically to shifting data distributions and workload demands. This section aims to articulate explicit research objectives for this domain: (1) minimizing index staleness for real-time and HTAP scenarios, (2) constraining resource usage and overheads during adaptive retraining or modification, and (3) establishing analytic performance guarantees and error bounds for index quality under dynamic and hybrid workloads.

Compared to previous surveys, we synthesize a unified view of recent advances that fuse incremental index maintenance, multi-armed bandit algorithms, and context-sensitive strategies, focusing particularly on measurable adaptability and analytic guarantees [60, 74]. Notably, frameworks such as the bandit-based index tuning in [74] provide provable regret bounds and empirically outperform deep reinforcement learning on shifting and static HTAP/analytical workloads, reporting up to 75% and 59% throughput speed-up, respectively. In the realm of distributed and spatial workloads, lightweight learned indexes such as LiLIS [25] integrate spline-based models and space-partitioning to deliver order-of-magnitude improvements in query and build efficiency on real and synthetic-scale tasks.

The landscape has shifted with the proliferation of hybrid and adaptive approaches: the taxonomy outlined in [60], distinguishing pure learned from hybrid, mutable from immutable, and various adaptation tactics, enables comprehensive comparison and guides the selection of indexing solutions tailored for dynamic environments. These surveys further stress open challenges regarding concurrency, precise error quantification, efficient model retraining, and robustness to query and data distribution variability.

Key approaches and challenges in this domain include the following: Incremental index maintenance methods for online adaptation as new data or query patterns emerge. Bandit-based adaptation mechanisms that balance exploitation with exploration, yielding performance guarantees that can be analytically quantified across workload shifts [74]. Context-sensitive indexing strategies, incorporating workload-aware feature extraction, dynamic retraining, and error-bounded spatial partitioning [25, 60].

Recent advances demonstrate that learned and hybrid adaptive indexing systems can achieve resilient, high-performance outcomes for streaming, spatial, and HTAP workloads, yet several obstacles remain on staleness prevention, resource overhead minimization, and robust generalization across workload types. Studies emphasize the importance of benchmarking, concurrency control, and system integration as open research priorities [60]. For broader perspectives on compressed computation and algorithmic directions, see [70].

11.6 Societal, Fairness, Privacy, and Ethical Issues

In this crosscutting section, our explicit objective is to provide an analytic synthesis and critical comparison of state-of-the-art techniques in adaptive indexing, machine learning, and similarity search from the perspective of their societal, fairness, privacy, and ethical ramifications. We refine our goals by (i) delineating measurable criteria relevant to formal privacy preservation, algorithmic fairness, and reproducibility; (ii) highlighting how recent technical innovations in these areas yield new guarantees or expose novel limitations; and (iii) unifying the discussion to bridge gaps identified in prior surveys, emphasizing connections between technical design choices and real-world impact across sensitive data domains.

Recent integration of advanced analytics and adaptive indexing into workflows involving sensitive, personal, or scientific data (e.g., healthcare, finance, policy) foregrounds ethical, privacy, and fairness issues. While scalable automated decision-making offers efficiency gains, these systems inherently risk propagating societal bias, privacy breaches, and undermining transparency if mitigation mechanisms are insufficient [11, 13, 19, 31, 40, 42, 45, 46, 51, 54, 63, 78, 85, 86, 91, 93, 94, 108].

With respect to privacy guarantees, recent methods increasingly go beyond classical anonymization or reliance on trusted third parties. For example, privacy in vehicular ad hoc networks is now practically enforced via combined local differential privacy (LDP) and distributed ledgers, obviating single points of failure and drastically reducing computation while maintaining robust privacy, as compared to earlier group-leader or centralized approaches [42, 85]. These formal criteria offer measurable standards for privacy that can be tracked during system design and audit. Similarly, immutable distributed ledger solutions not only protect data integrity but also provide audit trails necessary for regulatory compliance.

Research in clustering and high-dimensional analysis highlights the need for fairness and robustness across heterogeneous, noisy feature spaces. Adaptive clustering algorithms that leverage local parameterization or ensemble subspace strategies (as in [4, 19, 57]) demonstrably improve both validity and fairness over traditional fixed-parameter or single-view clustering, especially for datasets with strong feature heterogeneity. Theoretical analytic

Table 19: Comparison of illustrative adaptive learned indexing systems for dynamic and spatial workloads (from [25, 74])

System	Workload Type	Key Technique(s)	Measured Speed-up Over Baseline	Guarantees/Notes
Bandit Index Tuning [74]	Analytical/HTAP (Shifting, Static)	Multi-Armed Bandits	Up to 75% (shifting), 28% (static)	Provable no-regret bounds; no DBA or optimizer required
LiLIS [25]	Distributed Spatial Data	Error-bounded Spline-Based Learned Index	Orders of magnitude; 1.5-2× build speed	O(1) lookup, robust to scale; partitioning method sensitive

guarantees—such as minimax optimal error rates for consensus clustering (cf. [57]) or trimmed ensemble regression (cf. [4])—support rigorously measurable improvements in performance and fairness, yet bring associated challenges of computational cost or increased model complexity that must be transparently documented.

Transparent methodological reporting has also advanced, driven by resource-efficient index recommendation strategies and language-model-powered index advisors that facilitate explainable, reproducible database management [45]. Large language models paired with in-context demonstration pools significantly improve transparency and adaptability compared to legacy heuristics and black-box RL solutions, but hurdles persist for analytic workloads of extreme complexity or nonstandard schema. These advances are best evaluated using open benchmarks, ablation studies, and comprehensive disclosure of synthetic and real-world validation metrics [45, 54], setting measurable reproducibility goals that distinguish newer frameworks from prior surveys.

Ethically, integrating immutable ledgers and anonymous verification promotes system accountability and legal compliance, but ongoing research emphasizes the necessity of harmonizing privacy, auditability, and evolving regulatory obligations in dynamic, data-intensive scenarios. This includes the proactive assessment of algorithmic bias, accessibility, and unintended societal impacts, ensuring that high-stakes decisions in fields like healthcare and public policy adhere to both legal and ethical norms [19, 35, 40, 45].

Relation to Previous Surveys and Unifying Advances: Relative to earlier reviews, this section unifies recent analytic and technical breakthroughs by foregrounding formal guarantees for privacy (notably decentralized LDP and ledger-based methods), ensemble and local-adaptive fairness mechanisms in high-dimensional clustering, and explainability advances in index recommendation and auditability. Measurable goals—such as formal privacy bounds, minimax error rates for robust learning, or documented reproducibility through benchmarks/ablation—are now explicitly synthesized. This perspective bridges the gap between descriptive surveys and framework-oriented comparisons, substantiating claims with up-to-date technical and analytic developments [4, 19, 35, 42, 45, 54, 57, 85, 112].

Traceability and Citation Currency: All cited works referenced in this section, including the latest surveys and analytic contributions (e.g., [4, 19, 35, 42, 45, 54, 57, 85, 112]), are current and fully traceable, supporting transparent scholarly attribution and future review.

In summary, ensuring ethical, privacy-preserving, and fair data practice is inseparable from technical progress in adaptive indexing and analytics. Each innovation must be assessed not only on classic performance and scalability measures, but also on formal, reproducible, and societal criteria—setting a research agenda anchored by transparency, accountability, and inclusivity as defining pillars for future work.

11.7 Emerging Research Directions

Objectives: This section sets forth clearly articulated, measurable research goals that synthesize current and emergent challenges in data analytics and management. Specifically, we propose the following objectives: (a) Quantifying efficiency gains in large-scale indexing and retrieval systems through reproducible benchmarking; (b) Advancing scalable analytic models by establishing formal guarantees on resource usage and predictive performance; (c) Improving model interpretability with transparent reporting mechanisms and standardized evaluation criteria; (d) Ensuring fairness and mitigating bias by developing measurable protocols for demographic parity and outcome equity; (e) Enhancing robustness via adversarial testing suites and reliability metrics; and (f) Fostering adaptability by defining criteria for seamless integration of new modalities or data regimes.

Distinct from previous surveys, our exposition accentuates new opportunities in constructing unified frameworks that systematically bridge advances across indexing, retrieval, and analytic modeling. These integrative approaches move beyond siloed progress, promoting end-to-end analytic guarantees and comprehensive performance auditing. Our analysis not only recapitulates major achievements but also delineates persistent unsolved problems, inviting the community to develop cross-cutting methodologies and empirical baselines that support transparent comparison and replicable advancement.

Unified Indexing Architectures: Neural, Hybrid, Annotative, and Compressed. Neural representations, hybrid model-index combinations, annotative indices, and compressed data structures each enable trade-offs between expressivity, compressibility, and speed. For instance, compressed indexes optimized for repetitive or diverse data types yield significant storage and retrieval efficiencies [25, 60, 64, 70], yet can introduce increased model training complexities and rigidities that impact update flexibilities and query latency [25, 60]. Annotative and interpretable indexing methods [64] offer improved transparency but sometimes incur a modest accuracy penalty or limited adaptability to unseen query types. Open questions include optimal multidimensional trade-offs and guarantees on error bounds, especially for dynamic or adversarial workloads [25, 60].

Retrieval-Augmented Generation (RAG) and Structured LLM Queries. Embedding retrieval engines tightly within large language models, as seen in RAG architectures, advances knowledge-grounded query response [57, 60, 64]. However, RAG and structured LLM queries face challenges including seamless integration across disparate vector, relational, and graph indices, and require new unifying interfaces for knowledge graph management and prompt engineering [60, 64]. These frameworks are often susceptible to inconsistency between retrieved evidence and generated content, and to brittleness in the face of rapid schema evolution or incomplete data.

Unified Statistical–Computational Analytics. Emerging analytic systems increasingly integrate scalable tensor and mixture modeling, ensemble clustering methods such as consensus or self-constrained spectral clustering, and fairness-aware learning, aiming to deliver robust, theoretically-grounded, and practically effective analytics [57, 60, 70, 91, 112]. Key advantages include provable optimality and resilience to noise or partial information; for example, consensus spectral clustering frameworks leverage random projections and feature reweighting to achieve minimax-optimal error rates and robust clustering in high-dimensional, noisy, or low-informative regimes [57, 112]. Parallelizable structures underpin scalability, and neural-based analytics, such as neural similarity search, offer further reductions in computational complexity for specific tasks [91]. However, these unified approaches bring significant computational overhead and increased system complexity, while challenges persist in efficiently handling nonparametric or mixed-type data, maintaining precise error bounds in higher-dimensional settings, and supporting dynamic workloads or compressed computation [57, 60, 70]. The ongoing pursuit of consensus frameworks and minimax-consistent algorithms that generalize across diverse data and task domains is central to future research.

Robust and Scalable Adaptive Systems. Ensuring robustness to adversarial manipulation, distributional shifts, and environmental changes—especially under federated or streaming architectures—remains challenging [58, 60, 70, 112]. Adaptive learning, privacy-preserving analytics, and composable index synthesis are important enablers. However, deployment is complicated by trade-offs among privacy, latency, and update adaptability [60, 112]. While some theoretical models achieve promising constants regarding algorithmic optimality, as demonstrated by near-optimal search algorithms in challenging environments [58], these advances do not always translate into practical solutions that require minimal computational resources in real-world, compressed, or high-volume data scenarios [70]. There is also a continued need to address challenges in dynamic workloads, concurrency, and robust error handling to realize scalable adaptive systems in practice [60].

Crosscutting Challenges and Outlook. Key cross-domain challenges and promising trajectories are summarized as follows:

Achieving efficient, interpretable, and adaptive indexing remains a central concern, particularly in balancing the expressiveness of deep neural indexes with interpretability and practical retraining requirements [25, 60, 64]. For example, emerging learned index structures not only aim for high query throughput and dynamic adaptability in large-scale spatial and multi-dimensional settings but must also ensure tight error bounds, efficient retraining, and interpretability, as demonstrated by recent approaches such as LiLIS in distributed spatial big data processing [25, 60]. Complementary studies highlight that models can yield interpretable representations—e.g., word embeddings with explicit, natural language dimensions—while incurring minimal performance loss [64].

Unifying statistical and algorithmic guarantees is critical for advancing algorithms that ensure rigorous statistical validity and computational feasibility in high-dimensional, heterogeneous, and large-scale data. Recent robust clustering and consensus spectral methods, for instance, have been shown to yield minimax optimal error in challenging noisy regimes, achieving both accuracy and robustness in practical large-scale scenarios [57, 112].

Fairness and robustness are increasing priorities as workloads become more federated and data more diverse. State-of-the-art clustering methods are being designed with built-in resilience to adversarial and noisy regimes, and current approaches stress the importance of fairness-aware analytics that accommodate underrepresented groups and adapt to nonstationary distributions [57, 112].

The development of general-purpose integration frameworks—including unified abstractions, taxonomies, and pipeline construction tools—is underway to support seamless integration and hybridization of clustering, indexing, and analytic techniques [60]. Such frameworks facilitate versatile adaptation to differing data modalities and analytic objectives.

The ongoing integration of adaptivity, efficiency, fairness, and interpretability across disciplines continues to define the trajectory of fundamental research and is expected to underpin future breakthroughs in data-driven intelligence.

12 Synthesis and Conclusion

This survey has systematically examined the landscape of consensus and ensemble methods in the context of artificial intelligence, with particular focus on their taxonomy, operational characteristics, and practical implications. Central objectives included identifying the criteria and metrics used to compare surveyed methods, examining the scaling limits of consensus approaches, and highlighting the breadth of their applicability across domains.

We synthesized taxonomic proposals by critically reviewing their structure in relation to practical deployment. For instance, the modular breakdown of consensus techniques informs stakeholders not only about the algorithmic distinctions but also about their implications regarding scalability, efficiency, and deployability in real-world applications. The analysis highlighted that while certain classes of ensemble methods demonstrate robustness in small to medium-scale scenarios, significant challenges remain for scaling to high-dimensional data and resource-constrained environments. This finding emphasizes the need for research efforts that address load balancing, fault tolerance, and adaptation to distributed architectures.

A critical review of consensus and ensemble methods suggests that, although ensemble averaging and voting-based approaches can enhance prediction reliability, they often encounter bottlenecks in computational and memory usage as scale increases. Techniques that exploit hierarchical consensus or decentralized models partially alleviate these issues but may introduce new complexities in synchronization and result aggregation. Therefore, future research should focus on designing scalable architectures that strike a balance between predictive performance and computational feasibility.

The practical implications of the surveyed methods are extensive. In domains such as medical diagnosis, finance, and autonomous systems, the ability to aggregate diverse models is paramount for robust decision-making. Our synthesis further suggests multi-disciplinary applicability, with emerging use cases in areas like climate modeling and social network analysis.

Looking forward, a key research direction is the integration of consensus mechanisms with explainability and transparency frameworks, enabling not only accurate but also interpretable ensemble predictions. Moreover, foundational questions remain regarding the theoretical limits of consensus under decentralized and adversarial conditions.

In conclusion, this survey provides a consolidated understanding of the field, emphasizing the taxonomy's practical relevance, critical challenges in scaling, and widespread applicability. These insights aim to guide both researchers and practitioners in advancing robust, efficient, and interpretable consensus frameworks for the next generation of intelligent systems.

12.1 Restatement of Objectives

This survey set out to achieve the following objectives: (1) to systematically review the state-of-the-art techniques in clustering, indexing, and analytic methods relevant to the targeted application domains, (2) to provide a comparative analysis illuminating the core advantages and limitations of each approach, and (3) to identify open challenges, future directions, and potential synergies across the examined methodologies.

12.2 Synthesis of Approaches

The reviewed work demonstrates considerable progress in clustering, indexing, and analytic techniques, each addressing specific aspects of scalability, accuracy, and interpretability. Notably, clustering approaches excel in unsupervised structuring and knowledge extraction, while indexing techniques enhance retrieval efficiency for large-scale datasets. Analytic frameworks further enable actionable insights from complex data streams. Despite distinct emphases, an integrative perspective reveals several common challenges, such as handling high-dimensionality, ensuring scalability, and supporting real-time decision-making.

12.3 Proposed Unifying Taxonomy

Based on our analysis, we propose a unifying taxonomy encompassing three interrelated dimensions: 1. Data Characteristics: including dimensionality, heterogeneity, and dynamism. 2. Methodological Axis: spanning clustering granularity, indexing architecture, and analytic depth. 3. Application Targets: ranging from exploration and summarization to predictive analysis.

This taxonomy facilitates systematic comparison and guides both researchers and practitioners in selecting and combining techniques according to specific problem requirements.

12.4 Key Challenges and Future Outlook

Key open challenges and promising future directions emerging from the literature are: 1. Developing adaptive algorithms capable of managing evolving and dynamic data distributions. 2. Bridging the gap between unsupervised clustering and supervised analytic frameworks to enhance interpretability and performance. 3. Designing scalable indexing structures that remain efficient with increasing dataset size and complexity. 4. Integrating privacy-preserving and fairness-aware mechanisms within all methodological stages. 5. Automating parameter selection and hyperparameter tuning to reduce domain expertise barriers. 6. Fostering cross-domain transferability and generalization of developed approaches.

12.5 Concluding Remarks

This survey has provided a comprehensive synthesis of clustering, indexing, and analytic methodologies, clarifying their individual and combined roles in addressing contemporary data challenges. By restating objectives, integrating a taxonomy for conceptual cohesion, and outlining explicit challenges and future avenues, we offer a roadmap for continued advancement and cross-fertilization in this rapidly evolving field.

12.6 Comparative Review and Synthesis

The contemporary landscape of clustering, indexing, and similarity search for high-dimensional and categorical data is characterized by substantial methodological diversity and paradigm shifts. Traditional hard clustering approaches, such as k-means and hierarchical clustering, remain foundational for their simplicity and interpretability. However, these methods encounter significant challenges—including the curse of dimensionality, limited scalability, and sensitivity to noise or parameter selection—when confronted with complex, large-scale, or categorical datasets [76, 77, 85]. In response, modern research has produced a progression of enhanced methodologies: density-based, spectral, consensus, and ensemble clustering techniques, each designed to accommodate heterogeneities in data structure, density, and scale.

Spectral clustering has demonstrated consistently superior performance in high-dimensional contexts, owing to its use of eigenspace transformations that facilitate robust separation and flexible parameterization [82]. Nevertheless, this approach frequently incurs higher computational costs and exhibits increased sensitivity to initialization and hyperparameter configuration [17, 32, 56].

Consensus and ensemble clustering have emerged as pragmatic answers to the instability and ambiguity associated with model selection in high-dimensional or noisy regimes. By aggregating the outputs of multiple clustering executions—employing varying feature projections, subsamples, or foundational algorithms—these strategies capitalize on the "wisdom of the crowd" principle to enhance robustness and accuracy. Theoretical and empirical evidence supports their efficacy in challenging scenarios, such as sub-Gaussian mixtures and mixed-type data [25, 28, 33, 85, 93]. Nonetheless, the computational burden of consensus methods remains a concern, stimulating ongoing research into improving their scalability and refining the minimax optimality of combination rules.

Feature selection and dimensionality reduction are now indispensable for effective clustering and indexing in high-dimensional spaces. Established methods such as Principal Component Analysis (PCA), t-SNE, and UMAP remain prevalent for uncovering manifold structures. Yet, these techniques may yield misleading representations under heavy noise or nonlinearity, exemplified by the "scattering noise" phenomenon. Recent advances, such as the distance-of-distance transformation, address these limitations by disentangling structural signals from noise prior to embedding [26]. Moreover, the adoption of ensemble subspace projections, random feature selection, and regularized tensor decompositions—including tensor PCA and tensor-normal mixture models—expands dimensionality reduction techniques to multiway and highly structured data, thereby bolstering both statistical efficiency and scalability [2, 10, 51, 62].

Indexing methodologies are undergoing transformative change with the advent of massive, high-dimensional, and repetitive or categorical datasets. Classical spatial and metric indexes (e.g., *k*d-tree, R-tree) experience sharp performance degradation in very high dimensions or with heterogeneous attribute types. Consequently, contemporary solutions such as graph-based indexes (HNSW, proximity graphs), neural network-based systems, and compressed/text-indexing structures are increasingly adopted [27, 45, 61, 75, 109]. Annotative indexing innovatively integrates paradigms across inverted indexes, graph databases, and knowledge graphs within unified, scalable frameworks, enabling efficient transactional and concurrent querying, as well as supporting dynamic and semi-structured data formats [26]. This multi-paradigm approach supports efficient retrieval for both structured and unstructured data at scale [87]. In parallel, learned indexes and multi-dimensional neural indexing systems exhibit dynamic adaptability, model-driven querying, and robustness to distributional changes and retrieval-augmented generation workflows [3, 25, 35, 82]. The emergence of lightweight distributed learned indexes further underscores advances in minimizing query latencies and index building costs for spatial data, with approaches like LiLIS and ML-based spatial partitioning demonstrating order-of-magnitude improvements in query speeds and throughput [25, 71].

Substantial advances in similarity and range search have followed the evolution from exact *k*-nearest neighbor (*k*NN) algorithms to approximate methods. Notably, the use of product quantization, residual corrections, and graph traversal heuristics has yielded marked improvements in computational scalability. Recent work on graph-based indices has specifically advanced range search performance, with adaptive queries and early-stopping heuristics providing significant speedups, especially in large and dense datasets [61]. Innovations such as minimization residual quantization (MRQ), range-aware filter and hybrid-search algorithms (e.g., UNIFY, HSIG), and specialized index structures for applications like time series and trajectories now support billion-scale, real-time query workloads with reliable recall and efficient resource use [21, 42, 48, 53, 71–73, 78, 98, 109, 111]. Furthermore, robustness to dynamic workloads and adversarial query patterns is increasingly managed through adaptive algorithms and hybrid or ensemble-based indexing strategies [89, 99, 104]. Unified frameworks, like UNIFY, exemplify this by supporting hybrid pre-, post-, and range-filtered search on high-dimensional attribute-rich datasets while maintaining efficient index maintenance and scalability [53].

Tensor analytics, comprising decomposition models and high-order network embedding, represents a frontier in extracting latent structures from multidimensional data arrays as found in omics, neuroscience, and signal processing. Recent algorithms exploit the interplay between statistical and computational constraints to deliver interpretable and consistent factorizations. These methods overcome difficulties such as the lack of best low-rank approximations or the NP-hardness of optimization tasks, effectively balancing parsimony, scalability, and uncertainty quantification [10, 62].

Hardware-aware and compressed computation paradigms further expand the boundaries of feasible analytics by operating on compressed or in-memory representations, vital for petabyte-scale or streaming datasets [12, 83, 108]. These approaches emphasize CPU/GPU affinity, cache locality, and architecture-specific optimizations, as evident in developments related to index compression, efficient filter structures (e.g., windowed cuckoo filters), and compact data structures for document or sequence analysis [30, 83, 91, 107].

Taken together, the field's synthesis highlights a movement towards hybrid, adaptive, and robust systems. Integrating dimensionality reduction, advanced indexing, and multi-perspective clustering is increasingly recognized as essential for the comprehensive analysis of complex, high-dimensional, and categorical data.

12.7 Ongoing Challenges and Open Problems

Despite significant advances, several important theoretical and practical challenges continue to impede progress in the field:

Scalability and Expressiveness: Achieving scalable solutions for graph indexing and high-order analytics remains difficult, especially for dynamic, streaming, or extremely large datasets (e.g., containing billions of nodes). Current approaches are often limited by high memory requirements, costly maintenance, and inadequate response times [9, 58]. Large-scale graph search and indexing are particularly challenged by computational and storage overheads: for example, succinct indexing methods [23] reduce memory usage but face difficulties scaling beyond tens of millions of graphs, and in tensor-based or high-dimensional representations, computational bottlenecks are created by the NP-hardness and absence of efficient low-rank approximations. Furthermore, the need to process data that frequently evolves, as well as difficulties in partitioning and efficiently querying high-dimensional or partially ordered structures [58], continue to restrict scalable deployment. Key algorithmic advances, such as separator-based search strategies in partially ordered graphs and new succinct data structures, provide promising directions, but further breakthroughs are needed to balance expressiveness and scalability, especially when the underlying data is both massive and rapidly changing.

Robustness: Many existing systems fall short in providing robust handling of adversarial input distributions, substantial noise, and distributional shifts. These limitations constrain the deployment of analytics in real-time or adversarially influenced environments [23]. Techniques such as ensemble subspace clustering and consensus spectral methods [57] have made progress toward mitigating noise and adversarial effects, ensuring that clustering remains consistent and reliable even when only a small fraction of features is informative and the noise level is high. Nonetheless, developing scalable, robust solutions for diverse high-dimensional and noisy settings remains an open area of investigation, especially since many methods still incur high computational costs or require careful feature engineering.

Statistical-Computational Gap: Particularly for high-order and high-dimensional analytics, a persistent gap exists between statistically optimal methods and those attainable by efficient (e.g., polynomial-time) algorithms [9]. For instance, in the context of tensor decompositions and related models, information-theoretic

Table 20: Comparative Overview of Major Methodological Advances in High-Dimensional Analysis

Strategy or Method	Domains of Strength	Primary Advantages	Principal Limitations
Traditional Hard Clustering	Numeric, low-dimension	Simplicity, interpretability, fast convergence	Sensitive to noise, non-scalable, poor in high-dimension
Spectral Clustering	High-dimensional, networks	Robust separation, adaptable parameterization	High computational cost, initialization sensitivity
Consensus/Ensemble Clustering	Heterogeneous, noisy data	Robustness, improved accuracy, model instability handling	Computationally intensive, scaling challenges
Dimensionality Reduction	High-dimensional, manifold	Enhanced visualization, subspace recovery	Potential distortion/noise, manifold discontinuity issues
Graph-based Indexing	Large-scale, high-dimension	Efficient retrieval, adaptability, multi-paradigm support	Memory overhead, maintenance difficulty
Learned/Neural Indexes	Dynamic, large datasets	Model-driven access, adapts to data drift	Training complexity, generalization uncertainty
Approximate Similarity Search	Real-time, billion-scale	Fast query, recall-resource trade-offs	Possibly lower accuracy, adversarial vulnerability
Tensor Analytics	Multimodal, structured data	Latent pattern discovery, scalability, uncertainty quant.	Stat/comp. gap, convergence obstacles, complexity
Hardware-aware Computation	Streaming, petabyte-scale	Efficient memory use, architecture leveraging	Hardware dependency, compression artifacts

optimality is often unachievable by practical algorithms due to non-convexity and complex optimization landscapes. While techniques such as alternating minimization and gradient-based algorithms can sometimes bridge this gap, they frequently depend on sub-optimal initialization and do not guarantee reliable convergence, especially as the structure or dimension of the data increases [9, 60]. Closing this statistical-computational gap by developing both new algorithmic frameworks and improved theoretical understanding is a critical direction for advancing the field.

Statistical Rigor vs. Computational Efficiency: Recent advances often prioritize speed or parallelism at the cost of statistical soundness and interpretability. In sensitive domains such as biomedical analytics, failing to maintain statistical consistency or inferential reliability can undermine the trustworthiness of outcomes and restrict real-world applicability [57, 60]. For example, consensus spectral clustering methods offer higher robustness and accuracy but involve additional computational complexity, and learned multi-dimensional indexes frequently lack precise error analysis or inferential guarantees. There is a continuing need for methods that balance rapid data processing with transparency, consistency, and statistical rigor, particularly as applications demand both efficiency and reliable interpretation of analytic results.

Reproducibility and Benchmarking: The absence of comprehensive and standardized benchmarks, inconsistent data handling practices, and evaluation biases inhibit meaningful comparison of methods. The field critically needs multidimensional benchmarks and open-access repositories to support reproducible research and unbiased progress [57, 60]. Recent surveys highlight the importance of compiling, updating, and systematically classifying datasets and algorithms, as well as introducing standardized evaluation criteria for scalability, accuracy, and robustness [60]. Without these, progress in developing and evaluating new methods remains difficult to quantify or generalize.

Ethical and Societal Considerations: As high-dimensional analysis becomes pervasive in decision-critical domains, issues of fairness, transparency, privacy, and user agency are increasingly urgent. There is intensified demand for algorithmic frameworks that provide guarantees regarding fairness and responsible governance, particularly where analytic outcomes influence individual rights or societal welfare [70, 112]. The shift toward compressed computation [70] and adaptive analytic frameworks introduces new questions of interpretability and user control, especially as these methods become embedded in large-scale, automated systems. Ensuring responsible, ethical, and human-centered development remains a core open problem as analytical tools exert growing influence on sensitive societal and individual outcomes.

12.8 Future Outlook and Roadmap

Looking ahead, several converging trajectories are anticipated in the evolution of analytic systems and data structures for high-dimensional and categorical data:

Scalability remains a central challenge and priority. Future systems are expected to leverage hybrid architectures that blend compressed, hardware-conscious computation with distributed and cloud-native paradigms, making it feasible to manage both static and streaming massive datasets efficiently [12, 83]. Notably, advances such as highly space- and memory-efficient data structures (e.g., improved Cuckoo filters) and near-optimal dynamic algorithms for problems like vertex cover and matching are setting new practical standards in scalability, flexibility, and update efficiency. For instance, the latest Cuckoo filter designs introduce signed-offset addressing and overlapping window layouts, effectively removing traditional architectural overhead and permitting flexible, memory-efficient filtering for diverse large-scale use cases [83]. Similarly, state-of-the-art deterministic dynamic data structures now allow near-optimal updates for vertex cover and matching approximations, fulfilling open theoretical goals and ensuring efficient management in massive graphs [12].

Interpretability will be vital in building trust and driving adoption, especially in critical areas such as medicine, finance, and public policy [57, 60]. Progress here will rely on bridging transparent, explainable outputs from both statistical and machine learning-based models, and on the development and integration of indexes and clustering techniques whose operations are amenable to human scrutiny and reasoning. The development of learned index structures for multi-dimensional data has underscored both the opportunities and the challenges for interpretability, including the need to clarify the operation of machine learning models that are embedded within data management pipelines [60]. Advances in robust clustering for noisy and high-dimensional categorical data have also demonstrated the value of ensemble and consensus formulations, which mitigate classic noise and interpretability issues by aggregating multiple model outputs for more transparent results [57].

Benchmarking and reproducibility form another pillar of future research. There is a growing need to develop comprehensive and standardized benchmarking suites that address clustering, indexing, and similarity search tasks using both synthetic and real-world datasets [57]. Careful benchmarking underpins objective evaluation and fosters reproducible innovation, especially as methods for high-dimensional, categorical, and noisy data become more complex and statistically nuanced. The community is now emphasizing the development of evaluation protocols that are statistically robust and tailored to the unique demands of high-dimensional, mixed, and noisy environments, thereby supporting the rigorous comparison of competing approaches [57, 60].

Ethical and open science integration must be woven into both the algorithmic methodology and practical deployment. Considerations of fairness, privacy, and transparency will ensure that analytic systems serve users equitably and minimize societal risk [70, 112]. As compressed computation and scalable learning become ubiquitous, addressing their impact on society, including open science practices and the mitigation of algorithmic biases, becomes paramount. Adopting algorithmic innovations—such as compressed computation methods and self-constrained learning paradigms—necessitates parallel attention to transparency and openness across the research lifecycle [70, 112].

Research imperatives for the future include expanded investigation into dynamic graph and tensor analytics, compressed and federated computation, robust scalable clustering for mixed-type and noisy data, and interpretable, learning-augmented indexes [57, 60, 70]. Bridging the divide between statistical rigor, computational efficiency, and societal responsibility stands as a defining challenge for the forthcoming era. The literature highlights persistent open questions, such as supporting dynamic workloads, benchmarking with complex data types, guaranteeing model robustness against adversarial settings, and advancing GPU-accelerated multi-dimensional indexing [57, 60].

In summation, there is no singularly dominant approach in the high-dimensional analytic landscape. Instead, progress points toward integrated, adaptive, and accountable systems—where advances in dimensionality reduction, indexing, similarity search, and robust clustering are tightly coupled with rigorous benchmarking, interpretability, and societal stewardship. Achieving a cohesive synthesis among statistical excellence, computational scalability, and ethical responsibility will define the future of high-dimensional data analysis systems.

References

[1] A. Adolfsson, M. Ackerman, and N. C. Brownstein. 2019. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition* 88 (2019), 13–26. doi:10.1016/j.patcog.2018.10.026

[2] P. Afereidoon. 2025. persiansort: an alternative to mergesort inspired by persian rug. *arXiv preprint arXiv:2505.05775 [cs.DS]* (2025). https://arxiv.org/abs/2505.05775

[3] E. J. Aguilar and V. C. Barbosa. 2023. Shape complexity in cluster analysis. *PLoS ONE* 18, 5 (2023), e0286312. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0286312

[4] Md Firoz Ahmed, Sujit Kumar Mitra, and Rajdeep Mitra. 2021. Ensemble Linear Subspace Analysis of High-Dimensional Data: Theory and Applications. *Mathematics* 9, 21 (2021), 2669. https://www.mdpi.com/2227-7390/9/21/2669

[5] M. Aleksandrov, P. J. Prentice, and F. Wereszczuk. 2021. Voxelisation Algorithms and Data Structures: A Review. *Sensors* 21, 24 (2021), 8241. https://www.mdpi.com/1424-8220/21/24/8241

[6] Amjad Ali, Zardad Khan, Hailiang Du, and Saeed Aldahmani. 2025. Double weighted k nearest neighbours for binary classification of high dimensional genomic data. *Scientific Reports* 15 (2025), 12681. doi:10.1038/s41598-025-97505-2

[7] Imran Ali, Maria Balta, and Thanos Papadopoulos. 2023. Social media platforms and social enterprise: Bibliometric analysis and systematic review. *International Journal of Information Management* 69 (2023). doi:10.1016/j.ijinfomgt.2022.102510

[8] A. A. Amer, S. D. Ravana, and R. A. A. Habeeb. 2025. Effective k-nearest neighbor models for data classification enhancement. *Journal of Big Data* 12 (2025). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01137-2

[9] Arnab Auddy, Dong Xia, and Ming Yuan. 2024. Tensor Methods in High Dimensional Data Analysis: Opportunities and Challenges. *arXiv preprint arXiv:2405.18412* (2024). https://arxiv.org/abs/2405.18412

[10] L. P. Barnes, S. Cameron, and B. Howard. 2025. On Unbiased Low-Rank Approximation with Minimum Distortion. *arXiv preprint arXiv:2505.09647 [cs.DS]* (2025). https://arxiv.org/abs/2505.09647

[11] Jean Bertin. 2024. Advancing Similarity Search with GenAI: A Retrieval Augmented Generation Approach. *arXiv preprint arXiv:2501.04006 [cs.IR]* (Dec 2024). https://arxiv.org/abs/2501.04006

[12] Sayan Bhattacharya, Monika Henzinger, and Giuseppe F. Italiano. 2018. Deterministic Fully Dynamic Data Structures for Vertex Cover and Matching. *SIAM J. Comput.* 47, 3 (2018), 859–887. https://dblp.org/rec/journals/siamcomp/BhattacharyaHI18

[13] Xingyan Bin, Jianfei Cui, Wujie Yan, Zhichen Zhao, Xintian Han, Chongyang Yan, Feng Zhang, Xun Zhou, Qi Wu, and Zuotao Liu. 2024. Real-time Indexing for Large-scale Recommendation by Streaming Vector Quantization Retriever. *arXiv preprint arXiv:2501.08695* (2024), 1–20. https://arxiv.org/abs/2501.08695

[14] R. Binna, E. Zangerle, M. Pichl, G. Specht, and V. Leis. 2022. Height Optimized Tries. *ACM Transactions on Database Systems* 47, 1 (2022), 1–46. https://dl.acm.org/doi/10.1145/3506692

[15] Gregory Bint, Anil Maheshwari, Michiel H. M. Smid, and Subhas C. Nandy. 2019. Partial Enclosure Range Searching. *International Journal of Computational Geometry & Applications* 29, 1 (2019), 73–93. https://dblp.org/rec/journals/ijcga/BintMSN19

[16] Jean-Daniel Boissonnat, Karthik C. S., and Sébastien Tavenas. 2017. Building Efficient and Compact Data Structures for Simplicial Complexes. *Algorithmica* 79, 2 (2017), 530–567. doi:10.1007/s00453-017-0373-8

[17] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. 2016. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* 30, 4 (2016), 891–927. doi:10.1007/s10618-015-0444-8

[18] A. Chaves Carniel. 2024. Defining and designing spatial queries: the role of spatial relationships. *Geo-spatial Information Science* 27, 6 (2024), 1868–1892. https://www.tandfonline.com/doi/full/10.1080/10095020.2022.2163924

[19] Luyao Chang, Fan Li, Xinzheng Niu, and Jiahui Zhu. 2022. On an improved clustering algorithm based on node density for WSN routing protocol. *Cluster Computing* 25, 4 (2022), 3005–3017. doi:10.1007/s10586-022-03544-z

[20] Vasilis Chasiotis, Lin Wang, and Dimitris Karlis. 2024. Efficient subsampling for high-dimensional data. *arXiv preprint arXiv:2411.06298* (2024). https://arxiv.org/abs/2411.06298

[21] Georgios Chatzigeorgakidis, Sophia Karagiorgou, Spiros Athanasiou, and Spiros Skiadopoulos. 2018. FML-kNN: scalable machine learning on Big Data using k-nearest neighbor joins. *Journal of Big Data* 5 (2018), 4. doi:10.1186/s40537-018-0115-x

[22] B. Chen, F. Chen, J. Wang, and T. Qiu. 2025. An efficient and distribution-free symmetry test for high-dimensional data based on energy statistics and random projections. *Computational Statistics & Data Analysis* 206 (2025), 108123. https://www.sciencedirect.com/science/article/abs/pii/S016794732400207X

[23] X. Chen, H. Huo, J. S. Vitter, Y. Hu, and Q. Zhu. 2021. MSQ-Index: A Succinct Index for Fast Graph Similarity Search. *IEEE Transactions on Knowledge and Data Engineering* 33, 6 (2021), 2654–2668. doi:10.1109/TKDE.2019.2954527

[24] Yaru Chen, Jie Zhou, and Xinglong Luo. 2024. An improved density peaks clustering based on sparrow search algorithm. *Cluster Computing* 27, 8 (2024), 11017–11037. doi:10.1007/s10586-024-04384-9

[25] Z. Chen, W. Hao, Z. Zeng, Y. Wen, L. Shi, Z.-J. Wang, and Y. Zhao. 2025. LiLIS: Enhancing Big Spatial Data Processing with Lightweight Distributed Learned Index. *arXiv preprint arXiv:2504.18883v3* (2025). https://arxiv.org/abs/2504.18883

[26] C. L. A. Clarke. 2024. Annotative Indexing. *arXiv preprint arXiv:2411.06256* (2024). https://arxiv.org/abs/2411.06256

[27] V. Cohen-Addad, B. Guedj, V. Kanade, and G. Rom. 2021. Online k-means Clustering. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) (Proceedings of Machine Learning Research, Vol. 130)*. 1126–1134. https://arxiv.org/abs/1909.06861

[28] Sarita de Berg and Frank Staals. 2025. Nearest Neighbor Searching in a Dynamic Simple Polygon. *arXiv preprint arXiv:2503.03435* (2025), 22. https://arxiv.org/abs/2503.03435

[29] E. F. de Oliveira, P. Garg, J. Hjerling-Leffler, R. Batista-Brito, and L. Sjulson. 2025. Identifying patterns differing between high-dimensional datasets with generalized contrastive PCA. *PLOS Computational Biology* 21, 2 (2025), e1012747. doi:10.1371/journal.pcbi.1012747

[30] Naveen Donthu, Satish Kumar, Nitesh Pandey, and Prashant Gupta. 2021. Forty years of the International Journal of Information Management: A bibliometric analysis. *International Journal of Information Management* 57 (2021), 102307. doi:10.1016/j.ijinfomgt.2020.102307

[31] Simeon Emanuilov and Aleksandar Dimov. 2024. Billion-scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering. *Cybernetics and Information Technologies* 24, 4 (2024), 45–58. doi:10.2478/cait-2024-0035

[32] Johannes Fischer, Tomohiro I, Dominik Köppl, and Kunihiko Sadakane. 2018. Lempel-Ziv Factorization Powered by Space Efficient Suffix Trees. *Algorithmica* 80, 7 (2018), 2048–2081. doi:10.1007/s00453-017-0354-y

[33] T. Gagie, A. Hartikainen, K. Karhu, J. Kärkkäinen, G. Navarro, S. J. Puglisi, and J. Sirén. 2017. Document retrieval on repetitive string collections. *Information Retrieval Journal* 20 (2017), 273–303. doi:10.1007/s10791-017-9297-7

[34] A. J. Gallego, J. R. Rico-Juan, and J. J. Valero-Mas. 2022. Efficient k-nearest neighbor search based on clustering and adaptive k values. *Pattern Recognition* 122 (2022), 108356. doi:10.1016/j.patcog.2021.108356

[35] Z. Gniazdowski. 2024. New Approach to Clustering Random Attributes. *Zeszyty Naukowe WWSI* 19, 31 (2024), 41–90. doi:10.48550/arXiv.2412.09748

[36] E. Gorstein, R. Aghdam, and C. Solís-Lemus. 2025. HighDimMixedModels.jl: Robust high-dimensional mixed-effects models across omics data. *PLOS Computational Biology* 21, 1 (2025), e1012143. doi:10.1371/journal.pcbi.1012143

[37] Ralf Hartmut Güting, Suvam Kumar Das, Fabio Valdés, and Suprio Ray. 2025. Exact Trajectory Similarity Search With N-tree: An Efficient Metric Index for kNN and Range Queries. *ACM Transactions on Spatial Algorithms and Systems* 11, 1 (2025), 5:1–5:54. doi:10.1145/3716825

[38] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat. 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data* 11, 1, Article 113 (2024). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00973-y

[39] S. W. Harrar and X. Kong. 2022. Recent developments in high-dimensional inference for multivariate data: Parametric, semiparametric and nonparametric approaches. *Journal of Multivariate Analysis* 188 (2022), Article 104855. doi:10.1016/j.jmva.2021.104855

[40] Muhammad Umair Hassan, Xiuyang Zhao, Raheem Sarwar, Naif R. Aljohani, S. M. M. Rahman, K. Muhammad, and M. A. Raza. 2024. SODRet: Instance retrieval using salient object detection for self-service shopping. *Machine Learning with Applications* 15 (2024), 100523. https://www.sciencedirect.com/science/article/pii/S2666827023000762

[41] Majid Hojati, Rob Feick, Steven Roberts, Carson Farmer, and Colin Robertson. 2023. Distributed spatial data sharing: a new model for data ownership and access control. *Journal of Spatial Information Science* 2023, 27 (2023), 1–26. doi:10.5311/JOSIS.2023.27.220

[42] Zainab Iftikhar, Adeel Anjum, Abid Khan, Munam Ali Shah, and Gwanggil Jeon. 2023. Privacy preservation in the internet of vehicles using local differential privacy and IOTA ledger. *Cluster Computing* 26 (2023), 3361–3377. doi:10.1007/s10586-023-04002-0

[43] F. Iglesias, T. Zseby, and A. Zimek. 2020. Absolute Cluster Validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2096–2112. https://ieeexplore.ieee.org/document/8695871

[44] Bharathi B. K. and K. Jaganathan. 2022. The Intrinsic Structure of High-Dimensional Data According to Principal Graphs. *Mathematics* 10, 20 (2022), 3894. https://www.mdpi.com/2227-7390/10/20/3894

[45] Kiyonari Kobayashi, Shusuke Shimbo, and Yuji Matsumoto. 2024. Resource-Efficient Index Advisor Utilizing Large Language Model. *arXiv preprint arXiv:2503.07884* (2024). https://arxiv.org/abs/2503.07884

[46] A. Koudounas, C. Papagiannopoulou, L. Rokach, and S. Papadopoulos. 2020. Gradient-based Learning Methods Extended to Similarity-Based Models for Large-Scale Data. *Journal of Artificial Intelligence Research* 69 (2020), 1209–1247. https://jair.org/index.php/jair/article/view/12192/26600

[47] S. Ladra, M. Rodríguez Luaces, J. R. Parama, and F. Silva-Coira. 2024. Compact and indexed representation for LiDAR point clouds. *International Journal of Geographical Information Science* 27, 4 (2024), 1035–1070. doi:10.1080/10095020.2022.2121664

[48] M. Lawson, W. Gropp, and J. Lofstead. 2021. Exploring Spatial Indexing for Accelerated Feature Retrieval in HPC. *arXiv preprint arXiv:2106.13972* (2021). https://arxiv.org/abs/2106.13972

[49] Kuo-Kai Lee, Wing-Kai Hon, Chung-Shou Liao, Kunihiko Sadakane, and Meng-Tsung Tsai. 2023. Fully Dynamic No-Back-Edge-Traversal Forest via 2D-Range Queries. *International Journal of Computational Geometry & Applications* 33, 1&2 (2023), 43–54. https://dblp.org/rec/journals/ijcga/LeeHLST23

[50] J. Li. 2023. Finite sample t-tests for high-dimensional means. *Journal of Multivariate Analysis* 196 (2023), Article 105183. doi:10.1016/j.jmva.2023.105183

[51] J. Li, B. He, and D. Wang. 2021. A Scalable Random-Walk-Based Network Embedding Algorithm with Local Structural Information. *Journal of Artificial Intelligence Research* 71 (2021), 651–683. https://jair.org/index.php/jair/article/view/12567/26689

[52] Y. Li, R. Zhang, Q. Ma, J. Song, B. Zhang, M. Bai, W. Wang, and Y. Li. 2023. CSD-RkNN: reverse k nearest neighbors queries with category-sensitive distance. *International Journal of Geographical Information Science* 37, 8 (2023), 1709–1730. doi:10.1080/13658816.2023.2249521

[53] Anqi Liang, Pengcheng Zhang, Bin Yao, Zhongpu Chen, Yitong Song, and Guangxu Cheng. 2024. UNIFY: Unified Index for Range Filtered Approximate Nearest Neighbors Search. *arXiv preprint arXiv:2412.02448* (2024). https://arxiv.org/abs/2412.02448

[54] J. Lin and A. Trotman. 2017. The role of index compression in score-at-a-time query evaluation. *Information Retrieval Journal* 20 (2017), 274–314. doi:10.1007/s10791-016-9291-5

[55] J. Liu and M. Vinck. 2022. Improved visualization of high-dimensional data using the distance-of-distance transformation. *PLOS Computational Biology* 18, 12 (2022), e1010764. doi:10.1371/journal.pcbi.1010764

[56] Y. Liu, J. Ding, H. Wang, and Y. Du. 2025. A Clustering Algorithm Based on the Detection of Density Peaks and the Interaction Degree Between Clusters. *Applied Sciences* 15, 7 (2025), 1–19. doi:10.3390/app15073612

[57] S. Loodtoy and V. Yalagandula. 2021. Bibliometric Analysis of International Journal of Information Management. *International Journal of Information Management* (2021). http://repo.lib.jfn.ac.lk/ujrr/bitstream/123456789/4718/2/Bibliometric%20Analysis%20of%20International%20Journal%20of%20Information%20Management.pdf

[58] S. Lu, W. Martens, M. Niewerth, and Y. Tao. 2023. Partial Order Multiway Search. *ACM Transactions on Database Systems* 48, 4 (2023), 1–31. doi:10.1145/3626956

[59] Qing Mai, Xin Zhang, Yuqing Pan, and Kai Deng. 2022. A Doubly Enhanced EM Algorithm for Model-Based Tensor Clustering. *J. Amer. Statist. Assoc.* 117, 540 (2022), 2120–2134. doi:10.1080/01621459.2021.1904959

[60] A.-A. Mamun, H. Wu, Q. He, J. Wang, and W. G. Aref. 2024. A Survey of Learned Indexes for the Multi-dimensional Space. *arXiv preprint arXiv:2403.06456* (2024). https://arxiv.org/abs/2403.06456

[61] Magdalen Dobson Manohar, Taekseung Kim, and Guy E. Blelloch. 2025. Range Retrieval with Graph-Based Indices. *arXiv preprint arXiv:2502.13245* (2025). https://arxiv.org/abs/2502.13245

[62] N. Marco, D. Şentürk, S. Jeste, C. C. DiStefano, A. Dickinson, and D. Telesca. 2024. Flexible regularized estimation in high-dimensional mixed membership models. *Computational Statistics & Data Analysis* 194 (2024), 107931. doi:10.1016/j.csda.2024.107931

[63] Jorge Martinez-Gil. 2022. Evaluation of Code Similarity Search Strategies in Large-Scale Codebases. *Machine Learning with Applications* 10 (2022), 100423. https://www.sciencedirect.com/science/article/pii/S2666827022000868

[64] A. Michalopoulos, D. Tsitsigkos, P. Bouros, N. Mamoulis, and M. Terrovitis. 2025. Efficient Distance Queries on Non-point Data. *ACM Transactions on Spatial Algorithms and Systems* 11, 1 (2025), 1:1–1:37. doi:10.1145/3698194

[65] Xiangbo Mo and Hao Chen. 2024. A new classification framework for high-dimensional data. *arXiv preprint arXiv:2306.15199* (2024). https://arxiv.org/abs/2306.15199

[66] Neda Dousti Mousavi, S. Mostafa Hosseini, and Mahdi Mahmoudi. 2023. Categorical Data Analysis for High-Dimensional Sparse Covariates with Multinomial Responses: An RNA-Seq Cancer Application. *Mathematics* 11, 14 (2023), 3202. https://www.mdpi.com/2227-7390/11/14/3202

[67] Abhinav Natarajan, Maria De Iorio, Andreas Heinecke, Emanuel Mayer, and Simon Glenn. 2024. Cohesion and Repulsion in Bayesian Distance Clustering. *J. Amer. Statist. Assoc.* 119, 546 (2024), 1374–1384. doi:10.1080/01621459.2023.2191821

[68] H. Yepdjio Nkouanga and S. Vajda. 2023. Optimization Strategies for the k-Nearest Neighbor Classifier. *SN Computer Science* 4, 47 (2023). doi:10.1007/s42979-022-01469-3

[69] Daniel Obraczka and Erhard Rahm. 2022. Fast Hubness-Reduced Nearest Neighbor Search for Entity Alignment in Knowledge Graphs. *SN Computer Science* 3, 6 (2022), 501. doi:10.1007/s42979-022-01417-1

[70] A. Pakzad, V. Mehrjou, D. Khosla, and B. Schölkopf. 2021. A Word Selection Method for Producing Interpretable Word Embeddings. *Journal of Artificial Intelligence Research* 71 (2021), 867–900. https://jair.org/index.php/jair/article/download/13353/26748/29105

[71] V. Pandey, A. van Renen, E. T. Zacharatou, A. Kipf, I. Sabek, J. Ding, V. Markl, and A. Kemper. 2023. Enhancing In-Memory Spatial Indexing with Learned Search. *arXiv preprint arXiv:2309.06354* (2023). https://arxiv.org/abs/2309.06354

[72] Y. Pang, X. Zhou, J. Zhang, Q. Sun, and J. Zheng. 2022. Hierarchical electricity time series prediction with cluster analysis and sparse penalty. *Pattern Recognition* 126 (2022), 108599. doi:10.1016/j.patcog.2022.108599

[73] J. Paparrizos, F. Yang, and H. Li. 2024. Bridging the Gap: A Decade Review of Time-Series Clustering Methods. *arXiv preprint arXiv:2412.20582* (2024). https://arxiv.org/abs/2412.20582

[74] R. M. Perera, B. Oetomo, B. I. P. Rubinstein, R. Borovica-Gajic, and M. Roughan. 2023. No DBA? No Regret! Multi-Armed Bandits for Index Tuning of Analytical and HTAP Workloads With Provable Guarantees. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (2023), 12221–12237. doi:10.1109/TKDE.2023.3271664

[75] Nathan Phelps and Adam Metzler. 2024. An exploratory clustering analysis of the 2016 National Financial Well-Being Survey. *PLOS ONE* 19, 9 (2024), e0309260. doi:10.1371/journal.pone.0309260

[76] Alberto Policriti and Nicola Prezza. 2018. LZ77 Computation Based on the Run-Length Encoded BWT. *Algorithmica* 80, 7 (2018), 1986–2011. doi:10.1007/s00453-017-0379-2

[77] Yifan Qiao, Shiyu Ji, Changhai Wang, Jinjin Shao, and Tao Yang. 2023. Privacy-aware document retrieval with two-level inverted indexing. *Information Retrieval Journal* 26 (2023). doi:10.1007/s10791-023-09428-z

[78] A. Rachwał, A. Popławska, and M. Borys. 2023. Determining the Quality of a Dataset in Clustering Terms. *Applied Sciences* 13, 5 (2023), 1–22. doi:10.3390/app13052942

[79] S. Rahul. 2021. Approximate range counting revisited. *Journal of Computational Geometry* 12, 1 (2021), 183–212. https://jocg.org/index.php/jocg/article/view/3153

[80] S. Ray and B. Nickerson. 2022. Temporally relevant parallel top-k spatial keyword search. *Journal of Spatial Information Science* 24 (2022), 1–36. https://josis.org/index.php/josis/article/view/199

[81] R. Reinanda, E. Meij, and M. de Rijke. 2020. Knowledge Graphs: An Information Retrieval Perspective. *Foundations and Trends® in Information Retrieval* 14, 4 (2020), 289–444. https://www.nowpublishers.com/article/Details/INR-063

[82] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. da F. Costa, and F. A. Rodrigues. 2019. Clustering algorithms: A comparative approach. *PLoS ONE* 14, 1 (2019), e0210236. doi:10.1371/journal.pone.0210236

[83] J. E. Schmitz, J. Zentgraf, and S. Rahmann. 2025. Smaller and More Flexible Cuckoo Filters. *arXiv preprint arXiv:2505.05847* (2025). https://arxiv.org/abs/2505.05847

[84] Patrick Schäfer, Jakob Brand, Ulf Leser, Botao Peng, and Themis Palpanas. 2024. Fast and Exact Similarity Search in less than a Blink of an Eye. *arXiv preprint arXiv:2411.17483* (Dec. 2024). https://arxiv.org/abs/2411.17483

[85] S. Song and X. Liang. 2024. Federated Pseudo-Sample Clustering Algorithm: A Label-Personalized Federated Learning Scheme Based on Image Clustering. *Applied Sciences* 14, 6 (2024), 1–18. doi:10.3390/app14062345

[86] Liyang Sun, Yujing Wang, Zejian Wang, Xinyi Wu, Xiangming Dou, Jinji Li, Yicheng Bai, Xuerui Wang, Weinan Zhang, Yong Yu, and Zhenguo Li. 2024. The Disruption Index Measures Displacement Between a Paper and Its Citations. *arXiv preprint arXiv:2504.04677* (2024). https://arxiv.org/abs/2504.04677

[87] B. Tang, H. He, and S. Zhang. 2020. MCENN: A variant of extended nearest neighbor method for pattern recognition. *Pattern Recognition Letters* 133 (2020), 116–122. doi:10.1016/j.patrec.2020.01.015

[88] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide. 2022. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports* 12 (2022). https://www.nature.com/articles/s41598-022-10358-x

[89] Katerina Vrotsou, Georg Fuchs, Natalia Andrienko, and Gennady Andrienko. 2017. An Interactive Approach for Exploration of Flows Through Direction-Based Filtering. *Journal of Geovisualization and Spatial Analysis* 1, 1 (2017), 1–21. doi:10.1007/s41651-017-0001-7

[90] H. Wang and Q. Zeng. 2021. Unit-disk range searching and applications. *Journal of Computational Geometry* 12, 1 (2021), 381–417. https://jocg.org/index.php/jocg/article/view/4015

[91] H. Wang, J. Zhang, Y. Wei, Y. Wang, X. Zhang, and J. Pei. 2023. Neural Similarity Search on Supergraph Containment. *IEEE Transactions on Knowledge and Data Engineering* 35, 11 (2023), 11200–11214. doi:10.1109/TKDE.2023.3279920

[92] Haitao Wang and Wuzhou Zhang. 2019. On Top-k Weighted Sum Aggregate Nearest and Farthest Neighbors in the L1 Plane. *International Journal of Computational Geometry & Applications* 29, 3 (2019), 189–218. doi:10.1142/S0218195919500055

[93] S. Wang, L. Qin, J. X. Yu, R. Jin, and L. Chang. 2020. Continuously Adaptive Similarity Search. *ACM Transactions on Information Systems* 38, 3 (2020), 28:1–28:28. https://dl.acm.org/doi/10.1145/3318464.3380601

[94] H. Wei, P. Li, H. Gao, and C. Wang. 2017. String Similarity Search: A Hash-Based Approach. *IEEE Transactions on Knowledge and Data Engineering* 29, 7 (2017), 1371–1385. doi:10.1109/TKDE.2017.2692024

[95] N. Wiroonsri. 2024. Clustering performance analysis using a new correlation-based cluster validity index. *Pattern Recognition* 145 (2024), 109910. doi:10.1016/j.patcog.2023.109910

[96] G. Wu, J. Zhang, J. Fu, and J. Wang. 2022. A case study for Adaptive Radix Tree index. *Information Systems* 106 (2022), 101920. https://www.sciencedirect.com/science/article/abs/pii/S0306437921001228

[97] Y. Wu, X. Zhou, Y. Zhang, L. Ma, and J. Fan. 2024. Automatic Index Tuning: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 36, 12 (2024), 7657–7676. https://ieeexplore.ieee.org/document/10582533

[98] Jie Xue, Yuan Li, Saladi Rahul, and Ravi Janardan. 2020. Searching for the closest-pair in a query translate. *Journal of Computational Geometry* 11, 2 (2020), 1–33. doi:10.20382/jocg.v11i2a3

[99] J. Yang and C.-T. Lin. 2025. Autonomous clustering by fast find of mass and distance peaks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 1 (2025), 1–14. doi:10.1109/TPAMI.2025.40031325

[100] Mingyu Yang, Wentao Li, and Wei Wang. 2025. Fast High-dimensional Approximate Nearest Neighbor Search with Efficient Index Time and Space. *arXiv preprint arXiv:2411.06158* (2025), 8. https://arxiv.org/abs/2411.06158

[101] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao. 2024. A Rapid Review of Clustering Algorithms. *arXiv preprint arXiv:2401.07389* (2024). https://arxiv.org/abs/2401.07389

[102] Y. Yin. 2021. Test for high-dimensional mean vector under missing observations. *Journal of Multivariate Analysis* 186 (2021), Article 104797. doi:10.1016/j.jmva.2021.104797

[103] Huacheng Yu. 2022. Nearly Optimal Static Las Vegas Succinct Dictionary. *SIAM J. Comput.* 51, 3 (2022), 174–249. doi:10.1137/20M1363649

[104] P. Yuan, C. Jin, and G. Li. 2024. FDR control for linear log-contrast models with high-dimensional compositional covariates. *Computational Statistics Data Analysis* 197 (2024), 107973. https://www.sciencedirect.com/science/article/abs/pii/S0167947324000574

[105] Z. Yuan and C. L. Philip Chen. 2023. Forgetful Forests: Data Structures for Machine Learning on Data Streams with Incremental Computation and Filtering. *Algorithms* 16, 6 (2023), 278. doi:10.3390/algorithms16060278

[106] R. Zanibbi, B. Mansouri, and A. Agarwal. 2025. Mathematical Information Retrieval: Search and Question Answering. *Foundations and Trends® in Information Retrieval* 19, 1–2 (2025), 1–190. https://www.nowpublishers.com/article/Details/INR-095

[107] D. Zhang, Y. Huang, H. Wang, D. Yang, Z. He, and J. Xu. 2021. Continuous Trajectory Similarity Search for Online Outlier Detection. *IEEE Transactions on Knowledge and Data Engineering* 33, 10 (2021), 3405–3419. doi:10.1109/TKDE.2020.3046670

[108] J. Zhang, J. Tang, C. Ma, X. Chen, Y. Liu, and J. Li. 2018. Fast and Flexible Top-k Similarity Search on Large Networks. *ACM Transactions on Information Systems* 36, 2 (2018), 14:1–14:34. doi:10.1145/3086695

[109] Y. Zhang, M. Xiang, and B. Yang. 2017. Graph regularized nonnegative sparse coding using incoherent dictionary for approximate nearest neighbor search. *Pattern Recognition* 70 (Oct. 2017), 75–88. doi:10.1016/j.patcog.2017.05.004

[110] Xiaoyao Zhong, Haotian Li, Jiabao Jin, Mingyu Yang, Deming Chu, Xiangyu Wang, Zhitao Shen, Wei Jia, George Gu, Yi Xie, Xuemin Lin, Heng Tao Shen, Jingkuan Song, and Peng Cheng. 2025. VSAG: An Optimized Search Framework for Graph-based Approximate Nearest Neighbor Search. *arXiv preprint arXiv:2503.17911* (2025), 16. https://arxiv.org/abs/2503.17911

[111] S. Zhou, H. Xu, Z. Zheng, J. Chen, Z. Li, J. Bu, J. Wu, X. Wang, W. Zhu, and M. Ester. 2022. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. *arXiv preprint arXiv:2206.07579* (2022). https://arxiv.org/abs/2206.07579

[112] F. Zhu, Y. Kou, X. Jia, and Y. Zhu. 2023. An Efficient and Robust Semantic Hashing Framework for Similarity Search. *ACM Transactions on Information Systems* 41, 2 (2023), 1–30. doi:10.1145/3570725