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, 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

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

1 Introduction

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, 79, 84, 89, 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, 91, 92]. 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, 92, 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, 91].

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, 89, 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, 89, 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, 89, 92].

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. Each 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

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

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 [92, 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 [112].

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 [112]. Nevertheless, spectral clustering remains sensitive to affinity matrix construction and parameter tuning, necessitating careful preprocessing and validation to ensure reliability [92, 112].

2.4 Alternative Clustering Methodologies

The rich landscape of clustering for high-dimensional and mixed-type data extends beyond ensemble and spectral paradigms. Alternative methods include:

- **Hierarchical Clustering**: Both agglomerative and divisive strategies offer interpretability via dendrograms and flexible cluster resolution, though they may struggle to scale efficiently or maintain robustness in high-dimensional settings [3, 4, 19, 36, 37, 47, 57, 62, 72, 75, 82, 88, 89, 92, 99, 104, 111, 112].
- **Bayesian and Model-Based Approaches**: Mixture models, mixed membership, and tensor-normal mixtures provide probabilistic inference and meaningful uncertainty quantification. Advances in penalization and scalable inference address overparameterization and bottlenecks but challenges persist in extreme dimensionality [3, 57, 59, 67–69, 73, 85, 89, 93, 108, 111, 112].
- **Tensor Clustering**: Capitalizes on multiway data structures (e.g., in omics or imaging), affording improved model parsimony and interpretability. When paired with penalized or ensemble strategies, tensor clustering effectively reduces false positives and accounts for correlated predictors [57, 59, 67, 93].
- **Robust and Hybrid Methods**: These combine, for example, density- and partition-based criteria or incorporate deep learning representations, permitting flexible adaptation to irregular, compositional, or heterogeneous feature structures [3, 4, 36, 57, 72, 73, 75, 82, 88, 89, 92, 93, 99, 108, 111, 112].
- **Deep Clustering Paradigms**: These approaches jointly optimize representation learning and clustering within neural frameworks, delivering resilience to noise, initialization sensitivity, and overlapping structure. Such modularity enables end-to-end, domain-adaptive clustering, particularly effective for high-dimensional images, text, and graph data [3, 57, 73, 82, 108]. However, issues related to interpretability, hyperparameter selection, and domain generalization remain open [3, 57, 73].

Crucially, no single methodology demonstrates universal superiority; optimal selection is invariably tailored to dataset characteristics and analytical objectives [4, 19, 57, 62, 111, 112].

For a concise overview, structured comparison of main clustering paradigm features is provided in Table 1.

2.5 Cluster Validation Metrics and Benchmarking

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

Table 1: 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, 87, 89, 100–102, 104, 110, 111].
- **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, 87, 89, 95, 100–102, 104, 110].
- **Multimodality-Based and Modern Indices**: Recently, measures such as Dip and Silverman's tests have been advocated due to their sensitivity in uncovering true cluster structure and their ability to cope with noise, irregular cluster shapes, and the detection of multiple optimal cluster counts [1, 11, 19, 95, 101, 102, 111].

Despite notable advances, recent studies highlight that many classical metrics are sensitive to noise, class imbalance, and manipulation, underscoring the importance of multimodal validation and robust benchmarking protocols [4, 8, 11, 19, 39, 42, 66, 73, 87, 89, 100, 104].

Best practice now mandates joint use of internal and external metrics, meticulous dataset curation with transparency, and open, reproducible benchmarking pipelines to facilitate method development and comparison [4, 8, 36, 43, 57, 62, 66, 84, 111].

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 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

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 2, 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 demonstrate the feasibility of in-memory, space-efficient indexes, achieving storage reductions of 85–95% relative to conventional structures—while maintaining query performance parity. Such indexes synthesize probabilistic and deterministic techniques via hybrid encodings and succinct filters to effectively localize candidate sets for rapid searches, even at scales involving tens of millions of objects [76].

Probabilistic data structures like advanced Bloom filters and Cuckoo filters extend these innovations, achieving near-optimal

Table 2: 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

space usage and constant expected lookup times with explicit trade-offs between false positive rates, insertion, and deletion capabilities [32, 103]. Recent developments in dynamic address management for Cuckoo filters, such as signed-offsets and overlapping windows, have removed classic constraints, establishing new benchmarks for space efficiency and making these structures well-suited for large-scale analytics and scientific workloads [103].

In trie-based indexes, methods such as the Height-Optimized Trie (HOT) and Adaptive Radix Tree (ART) exploit node compression and dynamic fan-out, achieving a notable balance of lookup speed, memory usage, and update efficiency—attributes particularly valuable for real-time, in-memory database systems [23, 83, 96]. Suffix-based and run-length encoded indexes are tailored for repetitive data scenarios, as seen in web archives or genomics, by constructing compact representations that efficiently support factorization and substring queries. These approaches, particularly when leveraging compressed suffix arrays or run-length Burrows–Wheeler Transform (RLBWT), can achieve asymptotically optimal space on repetitive data [14]:

- High compression ratios, reducing storage requirements by orders of magnitude
- Efficient substring search and factorization capability
- Potential for optimal theoretical bounds on repetitive inputs

Nonetheless, supporting dynamic updates in compressed indexes can introduce notable overhead in practice [14, 70].

The broader adoption of memory-efficient indexes is not without trade-offs. Striving for optimal compression may inhibit support for dynamic operations or slow update pathways. The intricate engineering required to support range, similarity, and set queries over succinct structures remains a key challenge [70, 96]. As a result, contemporary research continues to seek an optimal balance among compression, adaptability, and query efficiency.

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 has enabled self-indexing in extremely limited space, paving the way for on-the-fly analytics in genomics, textual, and scientific archives [14]. Concurrently, compressed suffix trees embedded with witness structures provide unified, online computation of classic text factorizations, achieving linear time and sublinear space while supporting practical query efficiency [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

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 furnish O(1) lookup overhead and integrate seamlessly with distributed big data frameworks [60]. Model-driven indexing tightly couples partitioning strategies (e.g., R-tree, k-d tree, Z-order curves) with the index's predictive mapping, enabling dramatic reductions in build and query times. LiLIS, for example, achieves orders-of-magnitude speedup relative to classical distributed indexes, though it introduces additional overhead for model training and is sensitive to partitioning choice and query distribution [60].

New frameworks for index selection leverage multi-armed bandit and reinforcement learning algorithms to replace static, manually curated designs with dynamic, workload-adaptive index architectures [74]. These strategies base their optimization directly on observed workload performance, eschewing reliance on traditional

cost models and autonomously adapting to rapidly-changing patterns. They deliver significant speedups over both conventional rule-based and deep reinforcement learning techniques in dynamic analytic and hybrid workloads [74]. Importantly, this advancements transform auto-tuning from a fixed optimization task into a continual, online learning process that is robust to evolving data and workloads.

At a broader system level, architecture frameworks such as annotative indexing aspire to subsume the complete spectrum of traditional and learned indexes. Annotative indexes unify inverted, columnar, and object indexes within modular, transactional, and extensible designs, supporting expressive queries over semi-structured, heterogeneous, and graph- or vector-based data [26]. Their architectural flexibility enables features such as:

- Native transactional update support and concurrency control
- Lazy transformation and hybrid search, encompassing neural and sparse retrieval modalities
- Compatibility with structured, unstructured, and knowledge graph queries
- Composability for retrieval-augmented generation and next-generation analytics

Nevertheless, these modern approaches present their own unresolved challenges, spanning theoretical concerns—such as reliability of error bounds with high-dimensional predictions and robustness of model training—to practical issues including concurrent access, distributed scaling, adversarial resilience, and seamless system integration [25, 26, 60, 74]. Key open directions cover GPU-accelerated retraining, hybridization with classical index primitives, and the development of indexes supporting transactional guarantees for both structured and unstructured queries [25, 26].

This progression from classical structures through to compressed, adaptive, and learned indexing reflects the dynamic interplay between algorithmic innovation and practical system engineering. The principal contemporary challenges now focus on reconciling scalability, adaptability, and efficiency across ever-diversifying and accelerating data scales and modalities—a goal that continues to drive index structure research at all levels.

4 Similarity, Range Search, and Graph Querying

4.1 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. These approaches outperform earlier duplicate-avoidance schemes by leveraging finer-granularity partitioning logic [97].

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.

4.2 Efficient Index Management and Scaling

With the rise in query volumes and dataset sizes, managing and scaling indexes becomes paramount for large-scale data retrieval tasks. Avoiding duplicate results is critical; naive solutions risk both multiple reporting and redundant computation, particularly in operations involving overlapping spatial objects or intricate join queries. Secondary partitioning addresses these challenges by subdividing primary partitions according to object boundaries, thereby precisely localizing candidate sets for queries and minimizing unnecessary verifications [97].

Scalability is further reinforced through distributed and parallel index architectures, which leverage cluster-computing environments such as Apache Spark and Flink. Lightweight, learned index structures—often employing spline-based regression or space-filling curve mappings—enable $O(1)$ lookups. Each spatial partition is equipped with a custom-learned index, mitigating the curse of dimensionality and adapting to data skew [18, 25]. These strategies

markedly reduce both index construction times and query latencies while maintaining compatibility with big data frameworks, thereby facilitating real-time analytics at unprecedented scales.

Moreover, the contemporary focus on dynamic data environments has spurred development of indexing mechanisms capable of incremental and parallel updates. Persistent spatio-textual indexes, for example, support rapid integration of new data, facilitating prompt querying and supporting frequent updates, which are crucial in applications such as event recommendation systems and geo-tagged information retrieval [70, 97]. Despite these advancements, challenges remain, including the efficient construction of indexes for massive, evolving datasets and maintaining low-overhead rebalancing as data distributions shift.

The main approaches are structured in Table 3, illustrating the trade-offs between scalability, update efficiency, and their core strengths in large-scale settings.

4.3 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. Among these requirements, the capacity to efficiently resolve similarity queries—such as those based on edit distance or subgraph containment—is essential, yet 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.

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]. Such approaches generalize classical tree search techniques by partitioning the graph to minimize the upper bound on search effort, thus facilitating practical querying on large and intricately structured networks.

Key advances in graph querying are driven by the convergence of hybrid filtering, succinctness, adaptive partitioning, and strong competitive guarantees—reflecting broader trends in processing high-dimensional and irregular datasets.

4.4 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.
- **Machine Learning-Based Index Construction:** Leveraging regression models, space-filling curves, and other data-driven techniques to predict locations and enhance lookup speeds.
- **Adaptive Query Evaluation:** Dynamically tuning the search process in response to data characteristics, distributional shifts, and query patterns.
- **Ensemble and Subspace Techniques:** Combining multiple indexing or filtering strategies to mitigate high-dimensional challenges and exploit complementary strengths [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 [57, 70]. The success of these methodologies, however, remains contingent upon the underlying dataset characteristics (e.g., repetitiveness, attribute richness), query workload complexity, and computational limitations.

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, 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]. 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
- Enabling nonparametric and domain-agnostic retrieval
- Developing robust, fine-grained incremental index updates
- Standardizing evaluation protocols for multimodal and streaming data

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

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

Table 3: 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

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, 88, 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, 88, 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].
- **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].
- **Ensemble subspace methods:** Aggregation over multiple random or systematically chosen low-dimensional projections mitigates overfitting and stabilizes models [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].

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]. 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, especially when encountering high-dimensional, heterogeneous, or noisy datasets. Key objectives include mitigating noise and outlier effects, normalizing feature scales, and ensuring compatibility with downstream models.

Standard normalization approaches—such as min-max scaling, z-score standardization, and variance-stabilizing transforms—seek to harmonize feature ranges, but can falter when confronted with outlier-prone or heavy-tailed distributions, or when subject to compositional constraints [3, 35, 73, 95]. For compositional data, transformations like the log-contrast or isometric log-ratio are essential to avoid spurious correlations introduced by constant-sum constraints [50, 104].

Robust outlier detection and adjustment are critical, as preprocessing steps substantially influence analytical outcomes. However, the paucity of standardized benchmarks for evaluating outlier detection highlights the necessity for transparent and reproducible preprocessing pipelines. Domain-specific customization is frequently required, particularly for normalization and duplicate management [8, 28, 38, 50, 100, 104, 110]. Feature weighting and selection methods, which may integrate prior knowledge or leverage data-driven informativeness, are increasingly integral to highlight relevant variables and suppress noise. Such methods underpin the high performance of gene expression classifiers and cluster analysis workflows in complex biological data [57].

In streaming and dynamic data environments, preprocessing faces fresh constraints: algorithms must assimilate new information efficiently, adapt to evolving distributions (concept drift), and process deletions or reweighting without necessitating full model retraining [5, 16, 28, 76, 102]. These challenges are especially pronounced in real-time analytics and online learning scenarios, where the interplay of statistical rigor and computational efficiency must be dynamically managed.

5.4 Dimensionality Reduction and Visualization Techniques

Dimensionality reduction and visualization are vital for elucidating latent structures, ensuring interpretability, and supporting exploratory analysis within complex datasets. Principal Component Analysis (PCA) and its extensions, such as guided contrastive PCA (gcPCA) and contrastive PCA (cPCA), serve as core tools, with contemporary variants emphasizing robustness, contrast extraction, and the integration of domain-specific penalties [4, 29].

Nonlinear embedding techniques, including t-SNE and UMAP, are essential for visualizing clusters and manifold structures, particularly in single-cell genomics, neuroimaging, and image analysis. However, these techniques are sometimes susceptible to scattering noise, wherein random fluctuations obscure clear cluster identity in low-dimensional projections [29]. To address such challenges, advanced approaches like the distance-of-distance (DoD) transformation preprocess distance matrices to clarify cluster structures under noisy, high-dimensional conditions, demonstrably improving the fidelity of classification and visualization tasks [29].

Penalized regression methods—including the Lasso, Elastic Net, and Adaptive Lasso—play a central role in supervised dimensionality reduction and feature selection, especially under $p \gg n$ regimes [55]. Ensemble subspace methodologies further enhance resilience by aggregating results over diverse feature subsets, offering robust prediction even in correlated or weak-signal settings [57].

Visualization itself has moved beyond traditional scatterplots to embrace representations of clusters, graphs, tensors, and multidimensional mappings. Methods such as Flowcube for geographic flows and tensor decomposition-based clustering support both reduction and interactive exploration, thereby facilitating new insights at scale [5, 16, 22, 29, 50, 59, 90, 102, 104]. Where data possess inherent multiway structure, such as in tensor-valued datasets, model-based strategies like tensor normal mixture models utilize penalized likelihood to compress and cluster high-order data, promoting tractable and interpretable analyses [104].

Interpretability, transparency, and reproducibility are increasingly foregrounded in dimensionality reduction and visualization research. Progress is visible in the adoption of explainable feature allocation methods, the use of cluster validity indices, the development of reproducible benchmarking protocols, and the proliferation of interactive analytic tools [16, 29, 62, 66, 90, 104]. Nevertheless, substantial obstacles endure, particularly in delivering consistent low-dimensional embeddings, managing batch effects, and achieving scalability for large, multimodal data environments.

6 Feature Selection, Classification, and Vector Modeling

6.1 Feature Ranking and Robust Classification

Feature selection and classification in high-dimensional domains—especially in contexts characterized by a large number of features (p) and the presence of noisy, high-variance data—have experienced substantial methodological evolution. Central challenges include achieving robustness to outliers, ensuring interpretability, and maintaining computational scalability. Traditional classifiers often struggle when between-class distinctions are predominantly driven by differences in variance rather than mean shifts, or under substantial outlier contamination. To address these issues, innovative rank-based classification frameworks have emerged. These methods exploit rank information derived from pairwise distances among observations, enabling classification that is resilient against strict parametric modeling assumptions and highly robust to outlier effects [20].

Such rank-based algorithms typically involve the following steps:

- Computation of distance matrices between sample observations.
- Application of rank transformations to these distances.
- Integration with classifiers (e.g., quadratic discriminant analysis) to capitalize on variance-driven class separation.

Empirical evidence from both simulated and real-world data demonstrates that these frameworks frequently match or surpass the performance of state-of-the-art classifiers, particularly in scenarios where sensitivity to noise and adaptability to unconventional feature distributions are essential [65]. However, current challenges remain regarding algorithmic scalability and the dependence on the selection of appropriate distance metrics, motivating ongoing research in this area.

6.2 Nonparametric and Subdata Selection Methods

Nonparametric approaches and advanced subdata selection techniques are indispensable for analysis in high-throughput settings, where both the number of observations (n) and features (p) can be extremely large. The well-known limitations of standard LASSO-based variable selection—in particular, its diminished efficacy under strong predictor correlation or when $p \gg n$—have spurred the development of dual-stage frameworks. A representative procedure entails performing random Lasso for initial variable screening, followed by leverage-score-based sampling to identify the most influential data points for subsequent estimation. This yields demonstrable improvements in both statistical efficiency and computational resource usage [36].

To clarify the comparative benefits of dual-stage selection relative to conventional algorithms, the following overview synthesizes key aspects:

As shown in Table 4, dual-stage procedures systematically enhance robustness and estimation accuracy compared to single-stage or unstructured approaches—especially in highly correlated or computationally constrained scenarios.

Methodological progress also extends to regression with high-dimensional compositional covariates, spurring the development of hierarchical, mixed, and p-value-free false discovery rate (FDR)

Table 4: Comparison of Traditional and Dual-Stage Subdata Selection Methods

Method	Variable Selection Stage	Subdata Selection Stage
Standard LASSO	Single-stage (Lasso only)	No explicit subdata selection
Dual-Stage (Random Lasso + Leverage)	Randomized Lasso for variable screening	Leverage-score sampling for influential data points

control schemes. The latter leverage symmetry properties of test statistics under null hypotheses to facilitate valid inference even when conventional significance testing fails owing to elevated dimensionality or complex correlation patterns. Theoretical foundations ensure strict FDR control and asymptotically optimal power as sample sizes scale, with practical benefits confirmed both in simulation and applied omics research [20].

Further, penalized likelihood estimation for high-dimensional mixed-effects models has benefited from the introduction of coordinate descent algorithms with nonconvex penalties (e.g., smoothly clipped absolute deviation, SCAD). These approaches consistently deliver improved variable selection and greater estimation accuracy relative to LASSO, particularly where predictors are correlated or data exhibit group structures. Despite the availability of open-source implementations, outstanding challenges remain, such as guaranteeing algorithmic convergence in non-Gaussian settings and accelerating the tuning process [57].

6.3 Statistical Testing in High Dimensions

Statistical inference in high-dimensional environments requires robust procedures that remain effective when p is large relative to n and dependencies among features are significant. Classical mean vector testing methods—including Hotelling's T^2 statistic—deteriorate in reliability as dimensionality increases, resulting in inflated type I error rates and poor statistical power. To overcome these limitations, U-statistic-based techniques have been introduced for one- and two-sample testing paradigms, providing test statistics that converge to t-distributions as p becomes large relative to fixed n [22]. These tests obviate the need for resampling or complex adjustments, delivering direct and reliable inference for applications such as neuroimaging and genomics where "large p, small n" is prevalent.

Advances addressing missing data and random projections have likewise propelled high-dimensional inference forward. New test statistics accommodate data missing at random, fortified by asymptotic guarantees as both n and p scale up [102]. Random projection-based approaches draw upon the concentration of measure to efficiently approximate null distributions, thereby preserving computational feasibility and statistical validity even in ultra-high dimensions [39, 104].

Multiple comparison corrections and cluster validity metrics have also undergone rigorous evaluation, with contemporary studies emphasizing that proper alignment of statistical assumptions and hypothesis structuring is critical for balancing power and error control [43, 50, 60]. Extensive simulation studies enable practitioners to benchmark and interpret these methods for practical, high-dimensional scenarios.

The integration of ensemble and consensus clustering frameworks further enriches the analytic repertoire. These approaches combine dimension reduction, feature reweighting, and robust consensus techniques, permitting structured analysis even in the presence of noise and uninformative features. In particular, high-dimensional clustering is enhanced via the union of one-hot encoding, random projections, and spectral consensus mechanisms, augmenting robustness against stochastic and adversarial perturbations [50, 104]. Although computationally intensive, these frameworks lend themselves to parallelization and are supported by theoretical optimality, cementing their value for accurate structure detection in extremely large categorical or heterogeneous datasets.

6.4 Vector Space and Distributional Semantic Models

Semantic modeling in high-dimensional linguistic or biological contexts demands vector representations that attain an effective balance among interpretability, predictive accuracy, and computational tractability. Distributional semantic models—which encode entities as vectors in high-dimensional spaces—have been instrumental in capturing semantic relationships. Leading approaches such as neural embedding models (e.g., word2vec) and matrix factorization methods (e.g., NMF) provide high predictive accuracy; however, their dense representations typically lack dimension-wise interpretability.

Recent innovations address this shortcoming by proposing dimension selection procedures that directly map naturally occurring attributes (such as specific words) onto dimensions, enabling both interpretability and high accuracy. Empirical results from large-scale text analyses demonstrate that with judicious dimensionality selection, it is possible to retain competitive performance on semantic tasks (e.g., similarity judgments) while providing semantically meaningful vector axes. These transparent representations are advantageous for downstream interpretability, standing in contrast to black-box neural embeddings and NMF-derived alternatives [64].

In the area of database indexing and retrieval, vector model construction leveraging machine learning techniques—including clustering, neural networks, and hybrid systems—underpins efficient and adaptive multi-dimensional index structures. These advances facilitate scalable querying, indexing, and retrieval, accommodating the demands of large, continuously evolving datasets [43].

Collectively, contemporary developments underline that strategic construction and allocation of feature dimensions—tailoring model complexity to both data structure and interpretive priorities—are essential to scalable, informative, and interpretable analysis within high-dimensional and distributional semantic frameworks.

7 Benchmarking, Evaluation, and Cluster Validation

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 (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, 86, 89, 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 5, 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

Beyond clustering quality per se, practical deployments—especially at scale—demand evaluation of system-level characteristics that directly affect efficiency and usability. Key considerations include query efficiency, memory consumption, and index construction time, all of which become critical as datasets scale to millions or billions of points [1, 3, 5, 8, 14, 16, 23, 32, 35, 57, 60, 70, 73, 76, 82, 83, 95, 96, 100, 110, 112].

- **Latency and Throughput:** Central in operational and real-time analytics, where approximate nearest neighbor search (ANNS) techniques employ optimized memory layouts, vector quantization, and adaptive parameter tuning to balance speed and accuracy [1, 70, 95, 96, 112]. For example, memory-efficient graph indices can yield substantial reductions in query response times by minimizing cache misses while sustaining high precision.
- **Scalability:** Measured via memory footprint and algorithmic complexity with respect to data size and feature dimensionality. Hybrid indexing, adaptive partitioning, and hierarchical pruning are common strategies benchmarked for sublinear or near-linear scaling, often on community benchmarks with up to billions of vectors [1, 14, 16, 57].
- **Resource Constraints:** Particularly relevant in edge devices or embedded settings, memory and computational limitations prompt the use of per-query or per-million-point metrics for fair comparison, such as "peak n per query" or "memory units per million points" [5, 16, 83, 95].
- **Robustness:** Adaptive clustering and search frameworks are increasingly evaluated for resilience against data drift, adversarial perturbations, or input noise. Real-time systems support online or streaming updates, and maintain performance even as data distributions evolve, assessed with both standard accuracy measures and specialized metrics for adaptiveness [14, 60, 82].

Systematic benchmarking now requires comprehensive reporting of both analytic and engineering criteria, including timing, space, and accuracy under a broad spectrum of operational settings.

7.3 Benchmarking Environments and Open-Source Tools

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, 87, 92, 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 [17, 32, 39, 76, 83, 104].
- **Simulation Frameworks:** Tools that enable controlled manipulation of data properties (e.g., class separation, noise, dimensionality) foster rigorous algorithmic comparison and highlight sensitivity to dataset idiosyncrasies [29, 36, 60, 87].

Table 5: 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

- **Domain-Specific Libraries:** Implementations tailored for modalities like time series, trajectories, and point clouds offer specialized distance functions and evaluation procedures, addressing unique analytical challenges found in these domains [21, 39, 102, 104].
- **Reproducibility Artifacts:** Public web repositories, leaderboards, and challenge datasets facilitate verifiable benchmarking, with increasing emphasis on publishing full experimental configurations—random seeds, preprocessing scripts, and evaluation parameters—to support community-wide interpretability and reproducibility [11, 17, 22, 29, 32, 50, 76, 82].

Despite progress, substantial challenges persist. The landscape lacks universally recognized benchmark suites that capture the complexity inherent in emerging domains (e.g., graph, tensor, or mixed-type data) [60, 81, 84, 92, 106]. Issues of data corruption, anonymization, and missingness demand sophisticated simulation and evaluation environments that can model these phenomena explicitly [29, 50, 66, 102]. Continued community investment is required to curate, annotate, and standardize benchmarks that reflect real-world clustering complexities.

7.4 Visualization for Evaluation 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, 90, 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 [22, 50, 104]. Visualization is thus invaluable for qualitative validation of cluster separation, anomaly detection, and the identification of ambiguous or overlapping subpopulations, supplementing quantitative metrics with intuitive, expert-driven insight [50, 59, 102, 104].

Emerging solutions employ innovative interaction paradigms. For example, systems that visualize flows or spatial networks (e.g., Flowcube) leverage spatial filters to reveal patterns in large-scale, complex datasets, unveiling insight beyond the reach of automated scores [90].

To ensure transparency and reproducibility, current best practices emphasize the coupling of visualization with systematic, script-driven analytics [29, 64, 102]. Open interfaces, reproducible color mappings, and capabilities for exporting or replaying visualization states strengthen the interpretability and peer verifiability of findings [5, 64].

However, visual analytics also confront significant limitations, particularly with ultra-high-dimensional or massively multi-class data, or where intrinsic data structure resists intuitive mapping. These bottlenecks spur ongoing research into mixed-modality and interactive visualization methods capable of scaling alongside analytical and data complexity.

8 Data Representation, Storage Optimization, and Hardware Acceleration

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.

Notably, the success of these data structures in high-dimensional contexts hinges on the careful management of trade-offs among memory efficiency, update cost, and query speed. The following factors delineate this balance:

- **Static memory layouts** (e.g., contiguous arrays) excel in batch-analytic or streaming scenarios, offering superior throughput.

- **Dynamic memory layouts** support interactive and adaptive analytics but demand sophisticated concurrency controls to ensure consistency and performance.

Despite their promise, several challenges persist: the paucity of standardized benchmarks and mature libraries hampers widespread adoption; existing implementations often underperform with non-watertight models or when semantic annotations are required. These limitations highlight the ongoing need for robust, semantically-aware, and GPU-ready data formats [5].

8.2 Space-Efficient Storage Structures

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]. Cuckoo filters, in particular, improve upon classical Bloom filters by enabling deletions and supporting tunable false positive rates without sacrificing speed or flexibility [32, 103]. Recent advances have lifted previous architectural constraints—including the necessity for power-of-two bucket counts—and pioneered windowed or overlapping layouts to further reduce overhead and balance load, making Cuckoo filters well-suited for applications such as genomics and real-time analytics [32].

Compressed indexes represent another key advance, leveraging succinct data structures, run-length, and grammar-based compression to optimize the space–time tradeoff for a range of workloads. In domains with high data redundancy—such as version-controlled documents, genomic sequences, and large-scale log collections—modern indexes employ innovations like ILCP arrays and grammar-compressed document lists, enabling sublinear or compressed-space retrieval, counting, and ranking [14, 23, 83]. In graph analytics, succinct q-gram tree indexes dramatically reduce memory overhead for subgraph and similarity queries by compactly encoding occurrence patterns, thereby scaling to millions of complex objects [96].

This interplay of features among widely used space-efficient storage structures is summarized in Table 6, which outlines their distinctive capabilities and recent improvements.

Despite their compelling theoretical benefits, these advanced structures are accompanied by significant practical challenges:

- **Update costs** can be substantial, with many compressed and probabilistic structures struggling to support dynamic workloads without wholesale re-encoding.
- **False positives and query errors** inherent in approximate structures limit their use in mission-critical analytics.
- **Dynamic and online compression strategies**, essential for real-time and evolving databases, remain an early area of exploration.
- **Handling adversarial or worst-case distributions** effectively is an unresolved issue.

8.3 Hardware and Parallelization for Analytic Scalability

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].

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 [86] illustrate that communication-efficient and privacy-preserving analytics are feasible through the synergy of local summarization, prototype exchange, and robust central aggregation.

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]. 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].

8.4 Adaptive and Online Index Updating

To maintain agility in ever-changing analytical environments, index structures must accommodate online, dynamic updates and facilitate autonomous tuning. A significant advancement in this direction is the deployment of adaptive and self-tuning indexes, often powered by online machine learning and feedback mechanisms rather than static, manually-tuned configurations. Frameworks based on multi-armed bandits and online learning algorithms allow for the continual exploration and exploitation of possible structural

Table 6: Salient Features of Probabilistic and Compressed Storage Structures

Feature	Bloom Filter	Cuckoo Filter	Recent Variants	Use-case Focus
Supports Deletions	No	Yes	Variant-specific	
Tunable False Positive Rate	By design	Tunable	Adaptive/Variable	
Dynamic Resizing	Limited	Possible	Variant-specific	
Bucket Structure	Fixed	Power-of-2 (classical) / Relaxed (modern)	Flexible	
Use-case Focus	General Set Membership	High-throughput, Frequent Updates	Application-specific	

configurations, achieving rapid convergence toward optimal indexing layouts and demonstrating faster adaptation and improved robustness compared to traditional approaches [16, 74, 76]. These developments translate into substantial performance improvements, particularly in hybrid transactional-analytical (HTAP) systems and in responding to dynamic query workloads.

Such adaptability is not exclusive to relational systems. In domains like high-dimensional nearest neighbor search, adaptive algorithms iteratively refine cluster assignments, metric selection, and index organization based on observed data variability and feedback, thus maintaining performance even in adversarial or rapidly evolving settings [60]. Systems like OASIS can maintain families of locality-sensitive hash indices that adapt in real time as underlying similarity measures evolve—capabilities vital for interactive, non-stationary analytic workflows [60]. Research into cracking and incremental construction methods (such as those used for Adaptive Radix Trees) reveals that dynamic, workload-driven partial indexing can yield significant construction-time gains without deteriorating query performance [96].

Nonetheless, efficient online updating remains difficult for compressed or succinct data structures, where balancing, merging, and re-encoding may obscure or counteract performance benefits [14, 32, 76, 96, 103]. Furthermore, ensuring resilience to concept drift, adversarial interactions, and catastrophic forgetting is an unresolved challenge, especially as analytic platforms become more autonomous and must contend with unpredictable, high-throughput streams. It is thus imperative to develop adaptive algorithms that are provably efficient and reliable under continuous workload evolution.

This section has integrated recent advances and critically examined the complex interplay among data representation, compression, hardware optimization, and adaptive indexing. Together, these advances are converging to address the multifaceted demands of modern, large-scale analytics. The subsequent sections focus on domain-specific applications of these principles and outline open challenges in the pursuit of trustworthy and explainable analytics.

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

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

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 synthesizes methods from optimization, convex geometry, and random matrix theory to navigate these trade-offs. Efforts are ongoing to sharpen our understanding of sample complexity bounds, convergence rates, and the error profiles of different algorithms. Despite these advances, existing approaches often display suboptimal empirical performance, either requiring prohibitive data quantities or exhibiting susceptibility to poor local optima.

The structure of current algorithms for tasks such as clustering or indexing on tensor data tends to be rigid and insufficiently scalable, hindering deployment in practical analytics pipelines [9]. Several key open problems remain at the forefront of the field:

- Designing algorithms that reconcile statistical optimality with computational tractability for high-dimensional and high-order tensors.
- Developing robust initialization and regularization techniques suited to the unique challenges of tensor models.
- Extending clustering and indexing methodologies that natively operate on, and exploit, multiway tensor structures.

Addressing these open challenges is pivotal to fully realizing the analytical power of tensor and higher-order methods, with substantial implications for their application across varied scientific and engineering domains.

10 Applications and Deployment Strategies

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].

Table 7: 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.

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

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].

Federated learning introduces additional considerations, including statistical heterogeneity, communication overhead, and privacy preservation. Increasingly, these challenges are addressed through probabilistic model aggregation, distributed subspace consensus, or other federated cluster analysis mechanisms for inference across disparate data sources [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].

10.3 Guidelines for Deployment

Extracting valid scientific and operational insights from complex, heterogeneous datasets requires the implementation of analytics solutions that adhere to principled standards for automation, benchmarking, and statistical validation. Key recommendations include:

- **Automation:** Streamline feature preprocessing, model selection, and parameter tuning to support scalable workflows, while ensuring outputs remain interpretable and relevant to domain needs.
- **Benchmarking:** Conduct comprehensive benchmarking across diverse datasets and operational conditions. Leverage both internal and external evaluation indices, sensitivity analyses, and simulation-based studies to ascertain clusterability, validity, and model robustness [112].
- **Statistical Validation:** Employ rigorous clusterability diagnostics and out-of-sample validation, particularly in high-noise or high-dimensional environments, to mitigate risks of spurious discoveries.

- **Transparency and Reproducibility:** Promote transparent algorithmic reporting and open-source implementations. Publish benchmarks and share reproducible code and workflows to ensure scientific rigor and foster collaborative development [57].
- **Scalability:** Prioritize algorithmic efficiency, memory optimization, and distributed computation. Utilize resource-aware methods—including compressed computation, dynamic data structures, and federated analytics—to manage large-scale, heterogeneous datasets [70].
- **Interpretability and Ethics:** Uphold model interpretability, transparent feature selection, and fairness auditing, especially in sensitive biomedical, environmental, or social contexts. Ensure that analytics outputs are free from unintended bias and actionable in real-world deployments.

These practices collectively support robust, trustworthy, and adaptive analytics pipelines that are responsive to the evolving landscape of large-scale and heterogeneous data.

10.4 Comparison of Representative Large-Scale Deployment Strategies

To facilitate a structured overview of deployment options for high-dimensional and distributed analytics, Table 8 summarizes key characteristics of several representative strategies. This comparison highlights typical advantages, constraints, and application scenarios relevant to practitioners.

The successful deployment of large-scale clustering and analytics methods thus hinges on careful alignment between methodological strengths and the practical realities of domain data, workflow constraints, and institutional contexts. Continuing advancements in open-source dissemination, standardized evaluation, and adaptive algorithmic design will further catalyze innovation and responsible adoption across the sciences and industry.

11 Crosscutting Themes, Challenges, and Emerging Research Directions

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. Nonetheless, these frameworks are susceptible to challenges such as data overlap and class imbalance, necessitating the incorporation of flexible representation selection and dynamic parameter tuning mechanisms [68, 75, 99]. 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].

Tensor-based modeling constitutes another pivotal frontier in integrative analytics, offering interpretable and scalable substrates 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, 89, 92]. 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 indexing—has yielded a robust framework 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]. Given the contemporary requirement for end-to-end transactional semantics with heterogeneous data models, annotative and learned index frameworks—achieving ACID compliance, high concurrency, and the seamless integration of dense, sparse, and graph features—are poised to establish new standards in adaptive data management [43, 60, 97].

11.2 Machine Learning for Index and Analytic Optimization

Machine learning has assumed a central role in the optimization of index structures within database management and analytics platforms. Rather than depending exclusively on hand-crafted heuristics or costly offline tuning, modern methodologies treat index management as a learning or decision-making procedure, leveraging workload observation and cost feedback to perpetually adapt [25, 60]. Noteworthy progress has been made through resource-efficient recommendation systems utilizing large language models (LLMs), which synthesize workload characteristics and deduce ideal index schemas with minimal retraining or manual intervention. These systems integrate demonstration pools, scalable inference engines, and domain knowledge injection, attaining recommendation quality and latency that rivals or exceeds traditional index advisors [60, 64]. Key advances include:

- Modeling the index recommendation task for compatibility with few-shot or in-context learning,
- Extracting granular statistics from diverse workloads, and
- Deploying scalable aggregation mechanisms for robustness.

In addition to LLM-driven approaches, online learning frameworks—inspired by bandit algorithms—eliminate dependencies on DBA expertise or traditional query optimizers by utilizing active exploration and exploitation of index alternatives based on real-time performance measurement. These methods guarantee convergence to near-optimal performance, even in rapidly evolving or ad hoc

Table 8: Comparison of large-scale deployment strategies for clustering and analytics.

Strategy	Advantages	Constraints	Application Examples
Distributed parallel analytics (e.g., Apache Flink, Spark)	High scalability; fault tolerance; supports massive input volumes	Requires infrastructure setup; may require custom partitioning for optimal performance	Smart grid analytics, environmental sensor networks
Federated clustering/learning	Privacy-preserving; data remains local; enables cross-institutional models	Communication cost; statistical heterogeneity; model aggregation complexity	Multi-center biomedical studies, cross-jurisdictional finance
Graph-based indexing with hybrid encoding	Space-efficient; supports rapid queries over large, diverse graphs	Index construction cost; application-dependent parameter tuning	Chemical informatics, PubChem-scale search, social network mining
Compressed/learned data structures	Drastically reduced memory footprint; competitive accuracy	Potentially complex implementation; sensitivity to parameter selection	Text analytics, high-throughput genomics, image retrieval
Centralized brute-force/hybrid approaches	Simplicity; robust to data irregularities; minimal tuning	Poor scaling for massive inputs; high resource demands per node	Small- to medium-size or irregular dataset scenarios

workloads, and frequently surpass the efficiency of both deep reinforcement learning models and static analytics methods [60]. Nonetheless, more sophisticated or hybrid analytic workloads continue to pose challenges related to model expressiveness and rapid adaptation [60, 64].

11.3 Transactional and Distributed Perspectives

Recent developments in database management and analytical infrastructures have been characterized by the deepening intertwining of transactional, distributed, and computational paradigms. The capability to robustly execute distributed queries with strong ACID guarantees has become increasingly pivotal given the emergence of vector, graph, and hybrid analytics that necessitate cross-engine and federated access [60, 76, 83]. State-of-the-art systems are required to efficiently manage and query across heterogeneous backends, which often entails seamless integration among graph databases, knowledge graphs, and spatial or textual search engines—all while upholding high standards of performance and correctness [16, 26, 60].

Innovative indexing structures and partitioned system architectures enable distributed query execution and dynamic partitioning, though they introduce new trade-offs involving communication costs, consistency maintenance, and optimization under the constraints of partial or privacy-preserving data access [2, 10, 32, 76, 103]. Federated analytics, in particular, must balance the ideals of openness and collaboration with the complexities of security, transactional integrity, and latency management in geographically dispersive environments [22, 39, 50]. The rising emphasis on ACID properties in open, federated, and multimodal systems reflects a growing awareness of the imperative to integrate transactional guarantees within scalable, adaptive analytic environments [25, 26, 60].

11.4 Robustness and Adversarial Resilience

The expansion of indexing and analytics frameworks into sensitive domains—including healthcare, finance, and security—has rendered adversarial and randomized query resilience an essential system property. In high-dimensional, graph-structured, and compressed data environments, deliberate manipulations can degrade system performance and expose critical patterns, particularly when underlying indexes rely on learned or compressed representations [58, 60]. Evaluations have shown that graph-based and tensor analytic models are vulnerable to perturbations, underscoring the need for robust and regularized representations capable of withstanding worst-case and stochastic adversarial behaviors without compromising retrieval or inference quality [9, 58, 70].

For example, privacy-preserving document retrieval has adopted cryptographically fortified indexing, randomized responses, and obfuscation strategies to counteract leakage and inference-based attacks, typically at the cost of increased storage or computational burden [60, 70]. Likewise, resilient similarity and clustering algorithms for graph and high-dimensional data integrate adversarial feedback, robust scoring metrics, and continual adaptation. However, computational efficiency and universal applicability remain open challenges in scaling robust data analytics [58, 60]. Achieving equilibrium between privacy guarantees, adversarial robustness, and system efficiency is a persisting core problem for the community [70].

11.5 Online, Adaptive, and Learned Indexing for Dynamic Workloads

Contemporary workloads, characterized by rapid streaming, immense scale, and frequent evolution, have elevated the importance of online, adaptive indexing systems that can dynamically adjust to shifting data distributions and workload requirements. Emerging learned and hybrid index solutions continuously evolve in response to new data patterns and real-time feedback, outperforming static or heuristically managed systems on throughput and accuracy, especially for streaming and hybrid transactional/analytical processing (HTAP) datasets [25, 60, 70, 74]. Central techniques include:

- Incremental index maintenance,
- Bandit-based adaptation mechanisms, and
- Context-sensitive indexing strategies.

Recent findings indicate that such frameworks can consistently surpass their fixed counterparts, but obstacles persist regarding staleness prevention, resource overhead management, and robust generalization across diverse query and workload types. The integration of multi-armed bandit strategies, adaptive feedback loops, and sophisticated feature extraction are identified as vital constructs for future universally adaptive indexing systems [60, 70].

11.6 Societal, Fairness, Privacy, and Ethical Issues

The integration of advanced analytics, machine learning, and adaptive indexing within domains involving sensitive, personal, or scientific data has brought ethical, fairness, and privacy considerations to the forefront. High-throughput automated decision-making provides scale and efficiency, but also raises significant risks related to bias propagation, privacy violations, and transparency loss if not properly mitigated [11, 13, 19, 31, 40, 42, 45, 46, 51, 54, 63, 78, 86, 87, 92–94, 108].

Recent research emphasizes that modern data systems must:

- Satisfy formal privacy criteria, including differential privacy and immutable distributed ledgers,
- Incorporate explainability, auditability, and reproducibility by design,
- Rigorously validate index recommendations and cluster attributions, and

- Systematically evaluate societal implications of data accessibility, reusability, and potential bias

particularly for high-stakes applications in healthcare, finance, and public policy [19, 35, 40, 45].

Transparent methodological disclosure, provenance-aware index construction, and open/reproducible analytics are being advanced to strengthen accountability, while parallel research seeks to harmonize privacy, auditability, and regulatory compliance in data-intensive environments [4, 21, 28, 29, 39, 42, 44, 54, 57, 60, 66, 67, 70, 100, 104, 110, 112].

11.7 Emerging Research Directions

Several converging research trajectories are poised to shape the future landscape of data analytics and management:

- **Neural, Hybrid, Annotative, and Compressed Indexes:** Integrating neural representations with classic, hybrid, and annotative indexing enables responsive and semantically rich retrieval at scale. Advances in compressed indexing, optimized for repetitive, multimodal, or spatio-textual datasets, deliver efficiency gains but highlight ongoing challenges in balancing compression, latency, and update flexibility [25, 60, 64].
- **Retrieval-Augmented Generation (RAG) and Structured LLM Queries:** The emergence of tightly coupled retrieval engines and generative language models facilitates user queries that are directly grounded in curated, structured knowledge sources. Structured querying of LLMs further catalyzes the need for unified interfaces across vector, relational, and graph indices, spanning knowledge graph management and prompt engineering [57, 60, 64].
- **Unified Statistical–Computational Analytics:** Progress in scalable tensor modeling, high-dimensional mixture models, ensemble clustering, and fairness-aware learning is converging to underpin robust, scalable, and ethical analytical systems, increasingly uniting the goals of information-theoretic optimality with computational feasibility [57, 60, 70, 92, 112].
- **Robust and Scalable Adaptive Systems:** Achieving robust performance—resilient to adversarial threats and distribution shifts—while ensuring scalable operation across federated and streaming environments remains an open grand challenge, with adaptive learning, privacy-preserving techniques, and general-purpose index synthesis serving as crucial enablers [58, 60, 70, 112].

Across these crosscutting domains, the integration of adaptivity, efficiency, fairness, and interpretability continues to drive both fundamental scientific inquiry and transformative technological advancements.

12 Synthesis and Conclusion

12.1 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, 86]. 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. 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, 86, 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., kd-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. This multi-paradigm approach supports efficient retrieval for both structured and unstructured data at scale [88]. 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, 35, 82].

Substantial advances in similarity and range search have followed the evolution from exact k-nearest neighbor (kNN) algorithms to approximate methods. Notably, the use of product quantization, residual corrections, and graph traversal heuristics has yielded marked improvements in computational scalability. 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 [90, 99, 104].

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, 92, 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.

12.2 Ongoing Challenges and Open Problems

Despite notable progress, several theoretical and practical challenges remain unresolved:

- **Scalability and Expressiveness:** The ability to scale graph indexing and high-order analytics to dynamic, streaming, or exceptionally large datasets (e.g., billions of nodes) is still constrained by memory consumption, maintenance costs, and responsiveness [9, 58].
- **Robustness:** Existing systems frequently lack adequate resilience to adversarial input distributions, noise, and distribution shifts, which compromises applicability in real-time or adversarial environments [23].
- **Statistical-Computational Gap:** Particularly for high-order analytics, optimal statistical solutions may not be computationally attainable. Polynomial-time algorithms often underperform in the high-dimensional regime, with suboptimal initialization and uncertain convergence properties [60].
- **Statistical Rigor vs. Computational Efficiency:** Many recent innovations prioritize speed or parallellism at the expense of statistical consistency, transparency, and inferential reliability. In domains such as biomedical analytics, this trade-off can compromise trustworthiness and real-world utility [57, 60].
- **Reproducibility and Benchmarking:** The scarcity of comprehensive benchmarks, inconsistent data practices, and evaluation bias impede method comparison and progress. The field requires the establishment of multidimensional benchmarks and open repositories for reproducible research [57].
- **Ethical and Societal Considerations:** High-dimensional analysis raises urgent issues of fairness, transparency, privacy, and user agency, with growing demand for algorithmic solutions that offer provable fairness and responsible governance, especially in decision-critical domains [70, 112].

12.3 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:** Next-generation systems will necessitate hybrid architectures that combine compressed and hardware-aware computation with distributed/cloud-native designs, enabling efficient management of massive datasets in both static and streaming formats [12, 83].
- **Interpretability:** Unifying statistical and learning-based models with transparent, explainable outputs will underpin trust and adoption in sensitive domains such as medicine, finance, and policy [57, 60].
- **Benchmarking and Reproducibility:** The systematic development of diverse, standardized benchmarking suites—for clustering, indexing, and similarity search, across synthetic and real-world scenarios—will be essential to ensure objective evaluation and reproducible innovation [57].
- **Ethical and Open Science Integration:** Societal considerations including fairness, privacy, and open science must be integrated into both methodology and implementation, ensuring that algorithms equitably serve users and minimize risks [70, 112].
- **Research Imperatives:** Ongoing research should concentrate on dynamic graph and tensor analytics, compressed and federated computation, scalable clustering for mixed types, and interpretable, learning-augmented indexes. Bridging statistical rigor with computational efficiency and societal responsibility constitutes a principal goal for the forthcoming era.

In summation, there is no singularly dominant approach in the high-dimensional analytic landscape. Rather, the momentum is toward integrated, adaptive, and accountable systems—wherein advances in dimensionality reduction, indexing, similarity search, and robust clustering are 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.

Table 9: 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

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] Nijaguna Gollara Siddappa and Thippeswamy Kampalappa. 2020. Imbalance Data Classification Using Local Mahalanobis Distance Learning Based on Nearest Neighbor. *SN Computer Science* 1, 76 (2020). doi:10.1007/s42979-020-0085-x

[86] 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

[87] 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

[88] 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

[89] 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

[90] 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

[91] 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

[92] 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

[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