STM JOURNALS

Recent Trends in
Mathematics

Volume 1, Issue 2, 2024
July—December
DOI (Journal): 10.37591/RTM

https://journals.stmjournals.com/rtm

Review Case Study RTM

Topology and Geometry in Data Science: Persistent Homology and Beyond

Harinath Shukla*

Abstract

In recent years, the interplay between topology, geometry, and data science has gained substantial momentum, offering powerful frameworks to analyze and interpret complex datasets. Traditional statistical and machine learning methods often rely on linear or metric- based assumptions, which may fail to capture the intrinsic structure of high-dimensional or nonlinear data. In contrast, topological and geometric methods provide shape-oriented, scale- invariant tools that focus on the continuity, connectivity, and global organization of data. One of the most impactful developments in this area is persistent homology, a central concept in topological data analysis (TDA). Persistent homology enables the detection and quantification of topological features—such as connected components, cycles, and voids—across multiple spatial resolutions. These features are represented in persistence diagrams or barcodes, which provide stable, noise-resistant summaries of the underlying data topology. As a result, persistent homology has proven effective in a wide range of applications, including image analysis, time-series modeling, biological data interpretation, and material science. This review explores the theoretical underpinnings of persistent homology, its computational aspects, and practical implementation using simplicial complexes and filtrations. Moreover, the article discusses several real-world applications, showcasing the versatility and interpretability of topological summaries in diverse scientific domains. Beyond persistent homology, the field continues to evolve with emerging approaches such as multiparameter persistence, topological machine learning, and sheaf-theoretic frameworks, which further enrich the analytical capacity of TDA. These advanced techniques aim to address challenges such as data heterogeneity, scalability, and integration with other computational paradigms. Overall, this review highlights how the fusion of topology, geometry, and data science opens new avenues for understanding complex data landscapes. By bridging rigorous mathematical theory with practical algorithms, these methods offer a promising direction for future research in mathematical data science and computational topology.

Keywords: Topological data analysis, persistent homology, computational topology, geometric data analysis, multi-scale data representation, topological machine learning

INTRODUCTION

In the era of big data, the ability to understand and interpret the underlying structure of complex datasets has become increasingly important across a wide range of disciplines, including biology, medicine, finance, engineering, and social sciences. Traditional statistical and machine learning techniques often rely on assumptions such as linearity, independence, or low-dimensionality, which may not hold true in practice. As data becomes more intricate, high-dimensional, and unstructured, there is a growing need for mathematical tools that can capture the intrinsic geometry and topology of data beyond conventional metrics and Euclidean assumptions [1]. Topology and geometry offer powerful frameworks for

***Author for Correspondence**
Harinath Shukla
E-mail: harinath14shukla@gmail.com

Research Fellow, Department of Mathematics, Shri Ramswaroop Memorial University, Lucknow, Uttar Pradesh, India

Received Date: May 21, 2025
Accepted Date: August 04, 2025
Published Date: August 20, 2025

Citation: Harinath Shukla. Topology and Geometry in Data Science: Persistent Homology and Beyond. Recent Trends in Mathematics. 2024; 1(2): 21–27p.

addressing this need. Topology, in particular, is concerned with the properties of spaces that remain invariant under continuous deformations such as stretching or bending, while geometry focuses on measurements like angles, distances, and curvature. Together, these fields provide a language for describing shapes, patterns, and connectivity in data that are often invisible to purely statistical approaches [2]. The advent of computational topology has enabled the translation of these abstract mathematical ideas into practical tools for data analysis, giving rise to what is now known as Topological Data Analysis (TDA).

One of the central techniques in TDA is persistent homology, which allows us to detect and quantify topological features such as connected components, holes, and voids at multiple spatial scales.

Rather than examining a dataset at a single resolution, persistent homology tracks how these features appear and disappear across a filtration—a nested sequence of simplicial complexes built from the data. This multi-scale approach makes persistent homology particularly robust to noise and capable of capturing meaningful structures that might otherwise go unnoticed [3].

The visual representations of persistent homology, such as persistence diagrams and barcodes, serve as compact yet informative summaries of the topological features in a dataset. These summaries are stable under small perturbations in the data, making them well-suited for real-world applications where data is often noisy or incomplete. Persistent homology has been successfully applied to a wide array of domains, including shape recognition, sensor network coverage, neuroscience, genomics, time-series analysis, and material science, among others [4].

While persistent homology has garnered significant attention, it represents only a subset of the broader movement toward integrating topology and geometry into data science. Recent advances include multi-parameter persistence, topological machine learning, and the use of sheaf theory for modeling complex relationships in data. These emerging approaches aim to extend the capabilities of persistent homology, address its limitations, and provide new ways of understanding data through a topological lens [4].

This review article aims to provide a comprehensive overview of the role of topology and geometry in data science, with a primary focus on persistent homology and its theoretical underpinnings, computational techniques, and practical applications. We also explore the current trends and future directions that extend beyond persistent homology, highlighting how this growing field continues to reshape our approach to data analysis and interpretation [5].

FOUNDATIONS OF TOPOLOGY AND GEOMETRY IN DATA SCIENCE

The fields of topology and geometry provide essential mathematical tools for understanding the shape, structure, and continuity of data. These concepts help uncover features that are often hidden in high-dimensional or unstructured datasets. In this section, we introduce the foundational concepts that support topological data analysis (TDA), focusing on topological spaces, the transformation of data into combinatorial structures, and the algebraic tools used to analyze these structures [6].

Basic Topological Concepts

Topology is the mathematical study of spatial properties preserved under continuous transformations such as stretching, compressing, or twisting, without tearing or gluing. Key topological properties include connectedness, which refers to whether a space remains whole; compactness, which involves boundedness and closedness; and continuity, which governs how functions behave under deformation. Homotopy and homeomorphism are central ideas used to classify spaces. These concepts allow us to group data structures based on their intrinsic shape rather than specific geometric measurements like distances or angles [7].

From Data to Topological Spaces

Raw data in the form of point clouds or samples often lie in a high-dimensional space. To apply topological techniques, the data must be translated into a structure that reflects its shape. This is typically done by constructing simplicial complexes—combinatorial structures made up of vertices, edges, triangles, and higher-order simplices that encode the proximity and interaction among data points. Common methods include the Vietoris–Rips complex and the Čech complex, which approximate the data's underlying topology based on a distance threshold [8].

Homology and Betti Numbers

Homology is a tool from algebraic topology that assigns a sequence of algebraic objects (homology groups) to a topological space, revealing the presence of holes or voids in different dimensions.

These features are quantified by Betti numbers, where:

- β_0 counts the number of connected components,
- β_1 counts independent loops or cycles,
- β_2 counts enclosed voids or cavities, and so on.

Homology is powerful because it provides a way to compare topological features across spaces using algebraic invariants (Table 1).

PERSISTENT HOMOLOGY: THEORY AND COMPUTATION

Persistent homology has emerged as a fundamental tool in topological data analysis, providing a rigorous method for studying how topological features in data evolve across multiple spatial or scale parameters. This section delves into the core theoretical constructs and computational strategies that enable persistent homology to extract meaningful structural insights from complex datasets [9].

Filtrations and Persistence Modules

Filtrations are the starting point of persistent homology. They represent sequences of nested topological spaces built from data points by gradually increasing a scale parameter. Common filtrations include Vietoris–Rips, Čech, and Alpha complexes. Each space in the filtration captures different levels of proximity or connectivity among data points. As the scale increases, new topological features emerge or vanish. These evolving structures are encoded in persistence modules, which describe how homology groups change across the filtration. They allow for a systematic study of features that persist over multiple scales, distinguishing signal from noise [10].

Persistence Diagrams and Barcodes

Once homology groups have been computed across a filtration, the results are visualized using persistence diagrams and barcodes, both of which summarize the birth and death of topological features. A persistence diagram plots each feature as a point in a 2D plane where the x-coordinate indicates its appearance and the y-coordinate its disappearance.

Alternatively, barcodes represent these intervals as horizontal lines, offering an intuitive timeline of topological persistence. These representations are stable under small perturbations in the data, meaning that small changes do not significantly alter the results—an essential property for applications in noisy, real-world datasets [11].

Table 1. Summary of topological concepts in data science.

Concept	Description
Connectedness	Determines whether a space is in one piece or consists of disjoint parts
Compactness	Indicates boundedness and completeness in topological terms
Simplicial Complex	A combinatorial object made from simplices used to approximate data shape
Homology Groups	Algebraic structures that identify topological features like holes
Betti Numbers	Integers representing the number of features per dimension (β_0, β_1, β_2...)

Table 2. Comparison of libraries.

Library	Language	Strengths	Use Case
GUDHI	C++/Python	Flexible, supports multiple complex types	General-purpose TDA workflows
Ripser	C++	Fastest for Vietoris–Rips computation	Large-scale point cloud analysis
Dionysus	C++/Python	User-friendly, educational applications	Learning and prototyping

Algorithms and Software

Efficient computation of persistent homology requires algorithmic techniques capable of handling large and high-dimensional data. Central to this is matrix reduction, which simplifies boundary matrices to compute homology classes efficiently. Additional techniques like discrete Morse theory help reduce computational complexity. Several robust software libraries have been developed to implement these algorithms. Among the most widely used are GUDHI, Ripser, and Dionysus, each offering unique strengths in performance, scalability, and ease of use. Table 2 below provides a brief comparison of these libraries:

These tools are integral to translating the mathematical theory of persistent homology into practical applications across diverse fields such as biology, physics, finance, and computer vision.

APPLICATIONS OF PERSISTENT HOMOLOGY IN DATA SCIENCE

Persistent homology has proven to be a versatile and powerful tool in various domains of data science. By capturing topological features across multiple scales, it provides insights into the shape, structure, and connectivity of data that are difficult to obtain using classical techniques. This section explores key application areas where persistent homology has made significant contributions.

Shape Recognition and Computer Vision

In computer vision, recognizing and distinguishing shapes under varying conditions is a fundamental challenge. Persistent homology provides a method for extracting shape features that remain stable despite noise, rotation, scaling, or deformation. By analyzing point clouds derived from image data, it captures topological patterns such as loops and voids that correspond to structural features of objects. These topological descriptors are used as inputs to classification algorithms, enhancing object recognition, surface reconstruction, and 3D image segmentation. They have proven especially useful in tasks involving complex and irregular shapes.

Biological and Medical Data

Biological systems are inherently complex and structured across multiple scales. Persistent homology enables the extraction of structural and functional information from such data, offering a new way to understand biological forms and functions. For example, in neuroscience, it helps reveal the topological organization of brain networks. In cancer research, it distinguishes malignant from benign tissue by analyzing spatial patterns in histological images. In structural biology, it reveals folding patterns in proteins and DNA. These insights support diagnosis, classification, and discovery in medical and bioinformatics research.

Sensor Networks and Robotics

Topological data analysis provides a framework for understanding coverage, connectivity, and navigation in spatial environments, which are critical in robotics and sensor network applications. Persistent homology can identify coverage holes in wireless sensor networks, ensuring effective deployment. In robotics, it helps in mapping unknown environments and guiding autonomous exploration by identifying the topological features of physical spaces. This enhances robustness in localization and path planning, particularly in unstructured or dynamic environments where traditional geometric mapping techniques may fail [12].

Time-Series and Dynamical Systems

Time-series data often arise from dynamic processes and exhibit repeating patterns or chaotic behavior. Persistent homology can be applied to time-delay embeddings of such data to uncover recurrent topological features like loops and toroidal structures, which correspond to cycles or attractors in the underlying system. (Table 3) This allows for the classification of dynamical regimes, anomaly detection, and prediction of future states. It has applications in climatology, finance, engineering systems, and any domain where understanding temporal evolution is key.

BEYOND PERSISTENT HOMOLOGY: EMERGING DIRECTIONS

While persistent homology has established itself as a cornerstone of topological data analysis, recent developments have pushed the boundaries of this framework to address increasingly complex data structures and analysis goals. These emerging directions explore multi-faceted perspectives on data, improve theoretical expressiveness, and enable broader integration with machine learning. The following subsections introduce key advancements that go beyond classical persistent homology, revealing how topological and geometric approaches continue to evolve in response to the demands of modern data science.

Multiparameter Persistence

Multiparameter persistence generalizes classical persistent homology by allowing filtrations indexed over multiple variables instead of a single scale parameter. This extension enables simultaneous tracking of topological changes across several data features or metrics, offering a more nuanced view of complex datasets. However, the mathematical structure becomes significantly richer, leading to greater computational complexity and a lack of straightforward visualization tools like barcodes. Despite these challenges, recent progress in algebraic and computational techniques has begun to make multiparameter persistence more accessible for practical applications.

Sheaf-Theoretic Approaches

Sheaf theory offers a powerful and abstract framework for modeling how local data patches relate to global structures. In data science, sheaves help track contextual or overlapping information across datasets, especially when data is spatially distributed or hierarchically organized. By associating algebraic data to regions and studying their interrelationships, sheaf-theoretic tools can capture data consistency, constraints, and coherence. This approach has shown promise in applications like sensor fusion, signal processing, and dynamic systems, where integrating local observations into a coherent global picture is crucial.

Geometric Deep Learning and Topological Features

As deep learning continues to advance, researchers are increasingly integrating topological features—such as those derived from persistent homology—into neural networks to enhance their learning capabilities. These hybrid models aim to improve interpretability, generalization, and robustness by incorporating shape, structure, and connectivity into the learning process. Additionally, geometric deep learning methods, which generalize neural networks to operate on non-Euclidean domains like graphs and manifolds, naturally align with the goals of topological data analysis, creating promising synergies for complex pattern recognition tasks.

Table 3. Summary of persistent homology applications.

Application Domain	Role of Persistent Homology
Shape Recognition & Vision	Extracts robust shape features for classification and segmentation
Biological & Medical Data	Reveals topological structures in neural, protein, and tissue data for diagnosis
Sensor Networks & Robotics	Detects coverage gaps and navigational features for mapping and deployment
Time-Series & Dynamical Systems	Identifies recurring patterns and attractors to understand dynamic behavior

Mapper and Other TDA Tools

The Mapper algorithm is a widely used TDA tool that provides a visual and simplified representation of the data's underlying shape. It constructs a graph-based summary by clustering data within overlapping regions of a filtered space and connecting clusters based on shared points. Mapper is particularly useful for exploratory data analysis, enabling the detection of patterns, anomalies, and stratifications that might be hidden in raw data. Other complementary tools, such as Reeb graphs and merge trees, further enrich the topological toolkit available for data analysis.

CHALLENGES AND FUTURE PERSPECTIVES

Despite the growing success of topological and geometric methods in data science, several key challenges remain. One major limitation lies in the scalability of current algorithms, especially when dealing with large-scale, high-dimensional datasets. Computing persistent homology becomes computationally expensive as data size and complexity increase. Another challenge is interpretability, while persistence diagrams and barcodes summarize topological features, translating these summaries into meaningful domain-specific insights often requires expert knowledge.

Additionally, integrating topological tools with probabilistic models, statistical frameworks, and machine learning algorithms remains an active area of research. Future advancements are expected to focus on enhancing computational efficiency, improving theoretical foundations, and fostering interdisciplinary collaborations. This will support the development of more robust, explainable, and scalable topological approaches in real-world data science applications.

CONCLUSION

The intersection of topology, geometry, and data science represents a transformative shift in how we analyze and interpret complex datasets. Tools like persistent homology have proven especially valuable for uncovering multi-scale topological features such as connectivity, cycles, and voids— characteristics that often elude traditional statistical techniques. By offering robustness to noise and flexibility across various domains, persistent homology has established itself as a cornerstone of Topological Data Analysis (TDA).

However, the journey does not end there. Emerging methods—such as multiparameter persistence, sheaf theory, and topological machine learning—are pushing the boundaries of what topology can achieve in data science. Continued advancements in theoretical foundations, algorithmic development, and real-world applications will ensure that this interdisciplinary frontier remains rich, evolving, and impactful for years to come.

REFERENCES

1. Carlsson G. Topology and data. Bull Am Math Soc. 2009;46(2):255–308.
2. Edelsbrunner H, Harer J. Computational topology: an introduction. Providence (RI): American Mathematical Society; 2010.
3. Ghrist R. Barcodes: the persistent topology of data. Bull Am Math Soc. 2008;45(1):61–75.
4. Oudot SF. Persistence theory: from quiver representations to data analysis. Providence (RI): American Mathematical Society; 2015.
5. Chazal F, Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. Front Artif Intell. 2017;4:667963.
6. Zomorodian A. Topological data analysis. Adv Appl Comput Topol. 2020;70:1–39.
7. Munch E. A user's guide to topological data analysis. J Learn Anal. 2017;4(2):47–61.
8. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. Sci Rep. 2013;3:1236.
9. Otter N, Porter MA, Tillmann U, Grindrod P, Harrington HA. A roadmap for the computation of persistent homology. EPJ Data Sci. 2017;6:17.

10. Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S. Persistent homology analysis of brain artery trees. Ann Appl Stat. 2016;10(1):198–218.
11. Curry J, Mukherjee S, Turner K. How many directions determine a shape and other sufficiency results for two topological transforms. Found Comput Math. 2018;18(5):1167–1201.
12. Robinson AM. Topological signal processing. IEEE Signal Process Mag. 2014;31(4):113–27.

