TOPOLOGICAL DATA ANALYSIS FOR CUSTOMER SEGMENTATION AND BEHAVIORAL ANALYSIS IN E-COMMERCE

Keertana Kappuram (PID: A69034324)
kkappuram@ucsd.edu

I. INTRODUCTION & MOTIVATION

The rapid growth of e-commerce has led to the generation of vast volumes of complex, high-dimensional behavioural data from customer interactions, ranging from browsing and clickstream events to purchase histories and product preferences. Extracting meaningful insights from this data is critical for enabling personalized recommendations, segmenting customers effectively, and optimizing marketing strategies. However, traditional data analysis techniques, including clustering algorithms and dimensionality reduction methods like PCA and t-SNE, often fall short in preserving the intrinsic geometry and structure of behavioural datasets, especially in the presence of noise, sparsity, or nonlinear dynamics. In response to these limitations, Topological Data Analysis (TDA) has emerged as a powerful framework for uncovering the shape of data using tools from algebraic topology. Unlike methods that rely purely on distances or projections, TDA focuses on stable geometric and structural features that persist across scales. This report explores the role of TDA, particularly techniques such as Mapper and persistent homology, in capturing hidden patterns, nonlinearities, and multi-scale behaviours in customer interaction data, with the aim of advancing analytics and segmentation capabilities in digital commerce.

II. BACKGROUND

Topological Data Analysis (TDA) is a mathematically grounded approach for studying the intrinsic shape of data. Rooted in algebraic topology, TDA extracts multi-scale features, such as connected components, loops, and voids, that are robust to noise and invariant under transformations. These features help capture the structural essence of complex datasets without imposing strong geometric assumptions.

A core construct in TDA is the simplicial complex, which generalizes graphs into higher dimensions. From a point cloud dataset, one commonly builds a Vietoris–Rips complex: for a fixed radius ε, a k-simplex is included whenever all pairs of $k+1$ points lie within ε of each other. Varying ε gives rise to a filtration, a nested sequence of simplicial complexes that reflects how topological structure evolves with scale. Persistent homology analyses this filtration by tracking the birth and death of features across ε, yielding persistence diagrams or barcodes. These topological signatures can then be converted into vectorized forms such as persistence landscapes, images, or kernel embeddings, enabling downstream statistical and machine learning tasks. In contrast, Mapper is a graph-based algorithm that constructs a compressed topological summary of a dataset using a filter function (e.g., density, projection, recency, or purchase frequency). The data is sliced into overlapping intervals based on this function, clustered locally within each slice, and clusters that share points across overlaps are connected. The resulting graph captures both local and global structures such as branches, cycles, and disconnected regions, offering interpretable visualizations of the dataset's topology.

In e-commerce applications, these TDA tools are particularly valuable. Persistent homology can uncover cyclical user behaviour, temporal transitions in purchasing habits, or structural voids in co-purchase networks. Mapper, on the other hand, allows for meaningful customer segmentation, revealing transition paths between buyer personas and highlighting anomalous or underserved groups. By encoding high-dimensional consumer interactions into stable, interpretable topological features, TDA enables richer behavioural insights than traditional clustering or embedding-based methods, supporting more effective decision-making in personalization, targeting, and churn prevention.

III. LITERATURE SURVEY:

1. Mapper and the Topological Shape of High-Dimensional Data

"Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition" (Singh et al., 2007) [1]

The paper introduces a novel framework called Mapper for the qualitative analysis, simplification, and visualization of high-dimensional data. Unlike traditional manifold learning or dimensionality reduction techniques (e.g., PCA, t-SNE), Mapper uses tools from algebraic topology to construct multiscale, combinatorial representations that preserve local proximity while revealing global structural features like loops, flares, and disconnected components. The core idea is to reduce data into a simplicial complex (e.g., a graph or higher-dimensional analogue) that reflects the topological shape of the data with respect to user-defined functions (called filters). This approach is particularly useful for massive datasets where geometric or metric-based visualizations are ineffective.

Building on this foundation, the authors describe both the theoretical and statistical construction of Mapper. Theoretically, the method is inspired by the Nerve Theorem, where a simplicial complex is built from overlapping covers of a topological space, and a partition of unity enables a continuous map from the space to the nerve. In practice, the statistical implementation begins with a filter function $f: X \to \mathbb{R}$ applied to a point cloud dataset. The range of f is divided into overlapping intervals, each interval selects a subset of the data, and a clustering algorithm (e.g., single-linkage) is applied to these subsets. Each cluster becomes a vertex in the Mapper graph, and edges are drawn between clusters that share data points across overlapping intervals. The construction generalizes to higher dimensions using multiple filter functions, enabling the detection of complex topological features such as loops (β_1) and voids (β_2).

The paper presents several illustrative applications of Mapper that highlight its ability to uncover nontrivial topological structures in noisy, high-dimensional datasets. For instance, it successfully reconstructs the shape of a noisy circle using a distance-based filter, revealing a loop structure that traditional clustering would miss. Similarly, when applied to a sphere in $\mathbb{R}^3$, Mapper detects the spherical topology by leveraging two filter functions and computing higher-dimensional simplices. These examples demonstrate Mapper's capacity to recover known topological features such as connected components, loops, and voids from sampled data, even in the presence of noise and variation in density. Beyond synthetic geometric data, the authors propose Mapper as a general-purpose tool for visualizing and simplifying complex datasets in any domain where structure is difficult to perceive using standard methods. This interpretability and structural sensitivity make Mapper especially promising for e-commerce applications, where customer behaviour data often reside in high-dimensional spaces and exhibit nonlinear patterns. For example, Mapper can be used to construct customer segmentation graphs based on filters such as recency, frequency, and monetary value (RFM), revealing not only clusters of similar buyers but also transition paths between behavioural types and underserved customer groups. Unlike traditional segmentation techniques, Mapper captures both global structure (e.g., disconnected or looping segments) and local density variations, providing a multiscale view of consumer behaviour. This enables more targeted marketing strategies, churn risk identification, and the discovery of emergent patterns in customer lifecycles, critical capabilities in competitive digital marketplaces.

2. Persistent Homology and the Foundations of TDA

"Topology and Data" (Carlsson, 2009) [2]

In "Topology and Data," Gunnar Carlsson is motivated by a fundamental limitation in traditional data analysis: its reliance on geometric or linear techniques that often fail to capture the deeper structure of complex data. The paper argues that in many high-dimensional, noisy, or sparse datasets, especially those not well-behaved in Euclidean terms, important patterns such as loops, voids, or connected components remain hidden. To address this, Carlsson proposes a topological approach that focuses on

the shape and continuity of data, rather than distances or projections. His goal is to develop tools that are robust to noise, insensitive to specific coordinate systems, and capable of extracting global and multi-scale structure. This motivation drives the introduction of topological data analysis (TDA) as a rigorous framework that leverages ideas from algebraic topology, like simplicial complexes and persistent homology, to analyse data through its intrinsic shape, offering a new lens for discovery where conventional methods fall short.

Carlsson advances topological data analysis (TDA) beyond Mapper by emphasizing persistent homology as a mathematically rigorous method to extract and quantify multi-scale topological features from data. The paper discusses how finite metric spaces, often arising as point cloud data, can be transformed into nested families of simplicial complexes (e.g., Vietoris–Rips complexes) indexed by a scale parameter ε. As ε increases, simplices are added based on pairwise distances, resulting in a filtration: a sequence of topological spaces connected via inclusion maps. Persistent homology tracks the birth and death of homological features (e.g., connected components in H_0, loops in H_1, voids in H_2) across the filtration, producing persistence modules and interval decompositions. These are visualized as barcodes or persistence diagrams, which are stable under perturbations of the input data due to foundational stability theorems. This enables TDA to capture the intrinsic shape of data in a coordinate-free, noise-tolerant manner. Carlsson also highlights the algebraic structure underlying persistence modules and the implications of their decomposition in terms of interval modules, establishing persistent homology as not just a visualization tool but a computable and statistically meaningful descriptor of topological structure.

The techniques developed in this work have important implications for e-commerce analytics by enabling robust, structural insight into high-dimensional behavioural data. Persistent homology, in particular, allows one to analyse customer-product interaction data or purchase histories as sampled from an unknown underlying space, where topological features such as connected components, loops, or voids may correspond to latent customer segments, cyclical browsing patterns, or gaps in product offerings. Unlike Mapper, which relies on filter-based decompositions and clustering, persistent homology offers a scale-aware approach that quantifies how long these topological features persist across multiple resolutions, providing a measure of their significance. For example, persistent H_1 features may reveal consistent loops in navigation behaviour or transitions between product categories, while higher-dimensional features (e.g., H_2) could detect complex voids in the purchase co-occurrence space. Importantly, the stability of persistence diagrams under perturbations ensures reliability in noisy, incomplete, or sparsely sampled transactional data. This makes the approach well-suited for applications like product recommendation, customer segmentation, or sales forecasting, where structural robustness and interpretability are critical.

3. Foundations and Practical Implementations of Topological Data Analysis

"An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists" (Chazal & Michel, 2021) [3]

This paper by Chazal and Michel broadens the scope of Topological Data Analysis (TDA) by articulating its integration within modern data science workflows and emphasizing statistical, algorithmic, and software aspects necessary for practitioners. While foundational ideas like persistent homology and Mapper are mentioned, the paper distinguishes itself by focusing on the TDA pipeline: from data preprocessing and metric space construction, to complex selection (e.g., Vietoris–Rips vs. Čech), to vectorization techniques enabling compatibility with classical machine learning models. It provides detailed coverage of topological descriptors such as persistence diagrams, landscapes, silhouettes, and images, and describes how they can be embedded into vector spaces or used with kernel methods. This approach enables TDA to contribute not just as a visualization tool, but as a rigorous and quantifiable feature extraction methodology.

A central theme of the paper is the statistical and algorithmic tractability of TDA tools. The authors explore the stability properties of persistence diagrams under perturbations of the input metric space, providing theoretical guarantees for robustness. They also examine computational challenges, including

the combinatorial explosion of simplicial complexes, and present efficient strategies such as using witness complexes, sparsification techniques, and multiscale approximations to ensure scalability. The paper highlights how open-source software libraries, particularly GUDHI, support real-world experimentation, enabling users to construct filtrations, compute persistent homology, and convert results into numerical features without needing deep mathematical expertise.

The authors illustrate the practical power of TDA through domain-agnostic applications, ranging from trajectory analysis and material science to medical imaging and time series prediction. Unlike earlier works that focused more narrowly on shape analysis or low-dimensional visual insights, this paper emphasizes TDA's role as a feature engineering tool in full-fledged machine learning pipelines. For example, it shows how persistence-based features can be used in conjunction with SVMs, neural networks, or clustering algorithms for interpretable and robust learning. This shift from purely topological intuition to algorithmic integration makes the paper especially valuable for data scientists aiming to harness TDA within scalable, reproducible, and statistically grounded workflows.

4. Statistical Enhancements and Practical Parameterization of Mapper

"Statistical Analysis and Parameter Selection for Mapper" (Carrière, Michel & Oudot, 2018) [4]

This paper presents a grounded and scalable toolkit for applying Topological Data Analysis (TDA) in real-world machine learning pipelines. Rather than reiterating foundational motivations for topology in data science, the authors concentrate on practical integration: translating topological concepts into usable statistical tools within supervised and unsupervised learning frameworks. They emphasize the full processing pipeline from data representation in metric spaces to the extraction of vectorized topological signatures, thereby framing TDA not just as a theoretical approach but as a set of tools for feature engineering and statistical inference.

The technical heart of the paper lies in the systematic development of vectorized topological features. The authors describe algorithms for constructing filtrations such as Vietoris, Rips and alpha complexes and generating persistence diagrams that summarize geometric structure across scales. They then detail how these diagrams can be converted into formats like persistence landscapes and kernel-based embeddings, making them compatible with standard classifiers and regressors. Key computational considerations, such as the use of witness complexes and simplification techniques to reduce complexity, are presented along with references to scalable implementations in libraries like GUDHI and Ripser, which are essential for handling large datasets.

What sets this paper apart for the topic of this project is its focus on vectorization and statistical validation, which enables the deployment of TDA features in predictive models. By framing persistent homology outputs as robust, noise-resistant features that can be fed into machine learning systems, the authors offer a path toward interpretable and scalable applications. In the e-commerce context, this could enable novel customer segmentation strategies, anomaly detection in user behaviour, or representation learning over product interaction graphs. Unlike previous works that emphasized structural summaries or visualization, this paper provides a roadmap for embedding TDA outputs directly into statistical learning frameworks for decision-making at scale.

5. giotto-tda: A Scalable TDA Toolkit for Machine Learning and Exploration

"giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration" (Tauzin et al., 2021) [5]

Tauzin et al. (2021) introduce giotto-tda, a Python library that integrates topological data analysis (TDA) into machine learning workflows through a modular, scikit-learn–compatible interface. The library supports persistent homology and Mapper, along with a wide range of preprocessing tools for diverse data types including time series, graphs, and high-dimensional point clouds. It emphasizes usability and scalability, providing optimized C++ backends, GPU-accelerated components, and

support for hyperparameter tuning, pipeline composition, and cross-validation. By abstracting TDA into reusable components, giotto-tda makes it possible to include topological feature extraction directly within end-to-end machine learning pipelines.

In the context of e-commerce, giotto-tda offers a practical means of extracting structural insights from customer behavior data, which often exhibits high dimensionality and noise. Its Mapper implementation, for instance, enables the construction of topological graphs that summarize customer segmentation structures using domain-relevant filters such as purchase frequency or session duration. The library's real-time visualization capabilities and memory caching allow users to iteratively explore and refine these segmentations, making it particularly suitable for tasks such as identifying behavioral archetypes, detecting anomalous shopping patterns, or revealing latent market segments. Additionally, giotto-tda provides tools for converting persistence diagrams into feature vectors, which can be incorporated into classification or regression models, further extending its applicability to churn prediction, recommendation systems, and marketing optimization.

Overall, giotto-tda bridges the gap between TDA theory and real-world application by offering a production-grade, extensible platform that enables the deployment of topological methods in scalable data science workflows. Its focus on interoperability, visualization, and statistical learning makes it especially useful in domains like e-commerce, where understanding the structure of complex customer interactions is critical for data-driven decision-making.

6. TDA Applications in Time Series and Dynamical Systems

"A Short Survey of Topological Data Analysis in Time Series and Systems Analysis" (Gholizadeh & Zadrozny, 2018) [6]

This paper surveys the use of Topological Data Analysis (TDA) in analysing time series and dynamical systems, with a specific focus on the application of persistent homology to uncover temporal patterns and structural transitions. A central technique explored is delay-coordinate embedding, which transforms a 1D time series into a high-dimensional point cloud by constructing time-lagged vectors from the original signal. This geometric reconstruction captures the underlying phase space of the dynamical system, enabling the application of topological tools such as Vietoris–Rips and witness complexes. Persistent homology is then used to extract features like connected components (H_0) and loops (H_1), which correspond to recurrent system behaviours and periodic structures.

The survey emphasizes topological descriptors derived from time series embeddings, such as persistence diagrams, barcode entropy, and persistent landscapes, and highlights their robustness in detecting transitions between regimes. Technical use cases include identifying fixed points, periodic attractors, and bifurcations in nonlinear systems, phenomena that standard time-series techniques may overlook. Importantly, the paper also reviews the integration of these topological features into machine learning models for tasks such as classification, anomaly detection, and state estimation. Compared to traditional approaches, this methodology offers better resilience to noise and can detect structural changes without relying on strong parametric assumptions.

In the context of e-commerce, these techniques are particularly useful for analysing temporally evolving customer behaviour. For instance, clickstream or session data can be delay-embedded to reveal cycles in browsing or purchase activity. Persistent H_1 features could capture recurring engagement patterns (e.g., weekend spikes, holiday effects), while sudden changes in the topology of embedded time series could indicate churn risk, shifts in interest, or anomalies in engagement. By treating customer behaviour as a dynamic process rather than static snapshots, TDA-based time series analysis offers a principled way to track lifecycle evolution, segment users based on engagement stability, and design adaptive recommendation or retention strategies. This enables businesses to move beyond simple trend detection and toward topologically-informed monitoring and forecasting of customer dynamics.

7. Topological Characterization of Market Crashes in Financial Time Series

"Topological Data Analysis of Financial Time Series: Landscapes of Crashes" (Dłotko, Hess et al., 2019) [7]

In Topological Data Analysis of Financial Time Series: Landscapes of Crashes, Marian Gidea presents a framework for detecting early warning signs of systemic instability in financial markets using persistent homology and persistent landscapes. The core idea is to analyse the evolving correlation structure of financial assets (e.g., stocks in the S&P 500 or Dow Jones) by constructing a sequence of weighted graphs from sliding windows of asset returns. Each graph represents pairwise correlations between assets at a given time.

The author applies Weight Rank Clique Filtration to convert these graphs into topological spaces and then extracts their topological features using persistence diagrams and persistent landscapes. Over time, these landscapes reflect changes in market connectivity, and their L_p norms, real-valued measures of feature intensity, serve as indicators of structural stress. A sharp rise in L_1 or L_2 norms anticipates major crashes, such as the 2008 financial crisis, by several months. This approach offers a model-free, data-driven technique to quantify qualitative shifts in system behaviour. Unlike traditional statistical indicators, the method captures multi-scale topological features such as loops and voids that represent complex interactions among financial entities. These features are stable under noise and minor data perturbations, making them robust for practical forecasting. Persistent landscapes, in particular, provide a smooth, interpretable summary of topological changes, which can be analysed over time for anomaly detection or regime change identification.

Although the focus is on financial markets, the methodology translates well to e-commerce analytics. For example, temporal co-purchase or user-product interaction networks can be treated similarly to financial correlation networks. Applying TDA to such graphs allows for the detection of emerging trends, shifts in customer behaviour, or impending market disruptions (e.g., inventory bottlenecks or viral product cascades). Persistent landscapes and their L_p norms can serve as early indicators of structural changes in the consumer-product graph, enabling timely responses in recommendation systems, marketing strategies, or supply chain planning.

8. Topological Insights for Customer Segmentation and Digital Marketing Strategies

"Topological Data Analysis in Digital Marketing" (Lakshminarayan & Yin, 2020) [8]

This paper proposes a hybrid modelling framework that combines first-order Markov chain modelling with topological data analysis (TDA) to identify and classify patterns in customer clickstream behaviour for digital marketing applications. The central idea is to model each user session as a sequence of page transitions, compute class-specific transition probability matrices (TPMs) for buyer and non-buyer sessions, and generate log-likelihood ratio sequences over the session trajectory. These sequences reflect the evolving belief in a user's likelihood of conversion and form the basis for topological analysis.

To extract shape-based information from these sequences, the authors apply persistent homology to the log-likelihood curves. Specifically, they compute persistence diagrams for the sublevel sets of these curves and transform them into persistent landscapes and silhouettes, which serve as functional summaries encoding the strength and scale of topological features (e.g., local minima in decision confidence). These representations allow classification of new sessions by comparing their topological summaries to class-wise averages, using L^2 distances. A complementary classifier based on raw log-likelihood ratios is also evaluated.

The integration of Markov chains with persistent homology is particularly powerful in capturing both short-term decision dynamics and global browsing structure. Compared to conventional classifiers, the topological approach demonstrates superior precision and recall on longer sessions ($\geq$15 clicks), where complex browsing behaviour is more prevalent. While the methods are applied here to

clickstream data, the framework generalizes to other e-commerce behaviours such as session embeddings or product navigation graphs. It provides a new way to model intent not as a point estimate, but as an evolving topological signature of engagement, enhancing session-based targeting and real-time marketing strategies.

9. Integrating Topological Signatures into Deep Learning Frameworks

"Deep Learning with Topological Signatures" (Hofer et al., 2017) [9]

This paper proposes a novel framework that integrates persistent homology into deep learning pipelines by introducing a differentiable topological signature layer. While persistent homology provides a powerful means of summarizing the multiscale topological structure of data via persistence diagrams, traditional descriptors are not readily compatible with backpropagation. To bridge this gap, the authors define persistence landscapes and persistence images as topological summaries that can be treated as continuous, vector-valued functions. These summaries are stable under perturbations of the input and can be efficiently computed from the persistence diagrams, allowing them to be used as intermediate representations within neural networks.

The key innovation is the construction of a neural network module that accepts persistence diagrams as input and outputs topologically-informed feature vectors, which can then be concatenated with standard features or passed through subsequent layers. The authors demonstrate that this topological layer is fully differentiable, enabling gradient-based optimization and integration into end-to-end training. The framework is tested on both synthetic and real-world datasets, such as orbit classification, graph-based molecule prediction, and shape classification, showing that incorporating topological features improves generalization and robustness, especially in data regimes with inherent geometric or structural complexity. The paper also discusses the interplay between learned and handcrafted topological features, providing empirical evidence that topological priors can enhance learning capacity when fused with neural representations.

For e-commerce, this work introduces a compelling direction: integrating persistent diagrams directly into deep models for behavioural prediction, customer segmentation, or anomaly detection. For example, if Mapper or persistent homology is used to extract features from RFM vectors, session graphs, or co-purchase networks, this approach would allow these summaries to feed into neural recommender systems or churn prediction models. Moreover, the differentiability of the topological layer makes it suitable for online learning and real-time marketing tasks, crucial for adaptive personalization in e-commerce environments. This paper offers a practical and theoretically grounded pathway to scale TDA within modern ML workflows.

10. Statistical TDA with Persistence Landscapes for Functional Analysis

"Statistical Topological Data Analysis using Persistence Landscapes" (Bubenik, 2015) [10]

Bubenik introduces persistence landscapes, a novel topological summary designed to bridge persistent homology with functional analysis and statistical inference. Traditional persistence diagrams, while rich in geometric information, lack structure for direct statistical manipulation due to their representation as multisets in a non-linear space. Persistence landscapes resolve this by transforming each persistence diagram into a sequence of real-valued functions defined over the real line. Each function encodes the prominence of topological features (e.g., birth-death intervals) across scales, producing a functional summary that lives in a separable Banach space, specifically, L_p spaces, where standard operations like mean, variance, and hypothesis testing are well-defined.

The paper establishes strong theoretical foundations for this transformation. Bubenik proves that persistence landscapes are stable under perturbations of input data, making them robust for noisy or high-dimensional datasets. Importantly, these representations are amenable to functional statistics, such as computing confidence bands, performing functional PCA, and using kernel-based methods for classification or regression. The L_p-norm of the landscape, for instance, serves as a scalar summary

capturing the overall “topological energy” of a dataset, analogous to how norms in signal processing represent energy or intensity. The persistence landscape also enables vectorization without arbitrary binning, preserving geometric fidelity while enabling statistical learning methods.

For this project, persistence landscapes provide a scalable and statistically principled mechanism for embedding topological information into modelling pipelines. Whether you're analysing session graphs, purchase trajectories, or Mapper-derived clusters, landscapes allow you to quantify and compare structural complexity across customer segments using L_p norms or landscape distances. Their compatibility with classical machine learning models (e.g., SVMs, random forests) and statistical hypothesis tests makes them ideal for downstream tasks like churn prediction, lifecycle modelling, or A/B testing of customer cohorts. Integrating persistence landscapes with your topological summaries (e.g., from Mapper or persistent homology) ensures interpretability and reproducibility, key priorities in data-driven marketing and personalization strategies.

IV. PROPOSED METHODOLOGY:

1) DATASET DESCRIPTION:

The dataset used in this study is the Online Retail dataset, a publicly available transactional dataset from a UK-based, non-store online retailer. It contains detailed records of all transactions between December 1, 2010 and December 9, 2011, comprising approximately 500,000 entries. The company primarily sells unique, all-occasion gift items, and its customer base includes both individuals and wholesalers.
Each row in the dataset represents a single line item from an invoice and includes the following fields:

- InvoiceNo: A unique 6-digit identifier for each transaction.
- StockCode: A unique 5-digit product identifier.
- Description: Textual description of the purchased item.
- Quantity: The number of units purchased in the transaction.
- InvoiceDate: The date and time the transaction was recorded.
- UnitPrice: The price per unit (in British pounds).
- CustomerID: A unique 5-digit identifier assigned to each customer.
- Country: The country where the customer resides.

The data spans multiple countries, though the majority of purchases are from customers based in the United Kingdom. Other notable countries include France, Germany, and the Netherlands.
Prior to analysis, the data was cleaned by removing:

- Entries with missing CustomerID values.
- Records with negative Quantity (indicating returns).
- Canceled transactions (i.e., InvoiceNo entries beginning with 'C').

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536367	48187	DOORMAT NEW ENGLAND	4	12/01/10 8:34	7.95	13047	United Kingdom
536368	22960	JAM MAKING SET WITH JARS	6	12/01/10 8:34	4.25	13047	United Kingdom
536368	22913	RED COAT RACK PARIS FASHION	3	12/01/10 8:34	4.95	13047	United Kingdom
536368	22912	YELLOW COAT RACK PARIS FASHION	3	12/01/10 8:34	4.95	13047	United Kingdom
536368	22914	BLUE COAT RACK PARIS FASHION	3	12/01/10 8:34	4.95	13047	United Kingdom
536369	21756	BATH BUILDING BLOCK WORD	3	12/01/10 8:35	5.95	13047	United Kingdom
536370	22728	ALARM CLOCK BAKELIKE PINK	24	12/01/10 8:45	3.75	12583	France
536370	22727	ALARM CLOCK BAKELIKE RED	24	12/01/10 8:45	3.75	12583	France
536370	22726	ALARM CLOCK BAKELIKE GREEN	12	12/01/10 8:45	3.75	12583	France
536370	21724	PANDA AND BUNNIES STICKER SHEET	12	12/01/10 8:45	0.85	12583	France
536370	21883	STARS GIFT TAPE	24	12/01/10 8:45	0.65	12583	France
536370	10002	INFLATABLE POLITICAL GLOBE	48	12/01/10 8:45	0.85	12583	France
536370	21791	VINTAGE HEADS AND TAILS CARD GAME	24	12/01/10 8:45	1.25	12583	France
536370	21035	SET/2 RED RETROSPOT TEA TOWELS	18	12/01/10 8:45	2.95	12583	France

Figure 1. Dataset Sample

2) IMPLEMENTATION & EXPERIMENTAL RESULTS:

i. CUSTOMER SEGMENTATION USING MAPPER WITH CLV-PCA LENS

In this implementation, we apply the Mapper algorithm to perform unsupervised segmentation of e-commerce customers using engineered behavioural metrics from the UK Online Retail dataset. The preprocessing pipeline computes Recency, Frequency, and Monetary value (RFM) per customer, followed by a custom Customer Lifetime Value (CLV) formulation:

$$CLV = (Monetary / Frequency) * (1 / (Recency + 1))$$

This definition captures monetary contribution normalized by purchase regularity and penalized by temporal inactivity. Given the heavy-tailed nature of CLV, we apply a logarithmic transformation to obtain log_CLV, which stabilizes variance and improves geometric fidelity for downstream analysis.

For the topological modelling step, the Mapper algorithm (via Kepler Mapper) constructs a 2D lens composed of the first principal component of the standardized RFM variables and the log_CLV value. The data is covered using 12 overlapping hypercubes (with 45% overlap), and local clusters within each bin are identified using DBSCAN (ε=0.4, min_samples = 5). These clusters are assembled into a simplicial complex representing the shape of customer behaviour under the defined lens.

CODE SNIPPET:

```
# Compute RFM and CLV
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']
snapshot = df['InvoiceDate'].max() + pd.Timedelta(days=1)
rfm = df.groupby('CustomerID').agg({
   'InvoiceDate': lambda x: (snapshot - x.max()).days,
   'InvoiceNo': 'nunique',
   'TotalPrice': 'sum'
}).rename(columns={'InvoiceDate': 'Recency', 'InvoiceNo': 'Frequency', 'TotalPrice': 'Monetary'})
rfm['CLV'] = (rfm['Monetary'] / rfm['Frequency']) * (1 / (rfm['Recency'] + 1))
rfm['log_CLV'] = np.log1p(rfm['CLV'])

# Create Mapper lens: PCA projection + log(CLV)
scaler = StandardScaler()
X = scaler.fit_transform(rfm[['Recency', 'Frequency', 'Monetary']])
pca = PCA(n_components=1)
pca_component = pca.fit_transform(X)
lens = np.column_stack([pca_component, rfm['log_CLV'].values.reshape(-1, 1)])

# Construct and visualize the Mapper graph
mapper = km.KeplerMapper(verbose=1)
graph = mapper.map(lens, X, cover=km.Cover(n_cubes=12, perc_overlap=0.45),
clusterer=DBSCAN(eps=0.4, min_samples=5))
mapper.visualize(graph, path_html="clv_mapper_graph.html",
custom_tooltips=np.array(rfm.index.astype(str)))
```

OUTPUT:

CLV features computed.

```
Creating 144 hypercubes.

Created 56 edges and 24 nodes in 0:00:00.295105.
Mapper graph created with 24 nodes.
Wrote visualization to: clv_mapper_graph.html
Saved: clv_mapper_graph.html

Cluster-Level Business Summary:
Cluster 0 — Size: 1298, Avg Recency: 208.6, Avg Frequency: 1.8, Avg Monetary: 346.52
Cluster 1 — Size: 1989, Avg Recency: 127.3, Avg Frequency: 2.4, Avg Monetary: 671.77
Cluster 2 — Size: 1815, Avg Recency: 52.1, Avg Frequency: 3.4, Avg Monetary: 1163.74
Cluster 3 — Size: 1256, Avg Recency: 24.9, Avg Frequency: 4.7, Avg Monetary: 1831.87
Cluster 4 — Size: 665, Avg Recency: 11.5, Avg Frequency: 5.9, Avg Monetary: 2601.23
Cluster 5 — Size: 296, Avg Recency: 4.8, Avg Frequency: 6.4, Avg Monetary: 3048.11
Cluster 6 — Size: 65, Avg Recency: 2.5, Avg Frequency: 6.3, Avg Monetary: 3584.60
Cluster 7 — Size: 41, Avg Recency: 74.5, Avg Frequency: 5.7, Avg Monetary: 780.70
Cluster 8 — Size: 388, Avg Recency: 51.7, Avg Frequency: 5.1, Avg Monetary: 1242.14
Cluster 9 — Size: 965, Avg Recency: 33.7, Avg Frequency: 5.2, Avg Monetary: 1604.05
Cluster 10 — Size: 6, Avg Recency: 14.5, Avg Frequency: 26.5, Avg Monetary: 5952.46
Cluster 11 — Size: 993, Avg Recency: 20.7, Avg Frequency: 5.8, Avg Monetary: 2198.89
Cluster 12 — Size: 10, Avg Recency: 12.2, Avg Frequency: 25.7, Avg Monetary: 6748.76
Cluster 13 — Size: 625, Avg Recency: 9.8, Avg Frequency: 7.1, Avg Monetary: 3011.92
Cluster 14 — Size: 325, Avg Recency: 4.6, Avg Frequency: 8.5, Avg Monetary: 4048.73
Cluster 15 — Size: 65, Avg Recency: 2.5, Avg Frequency: 6.3, Avg Monetary: 3584.60
Cluster 16 — Size: 5, Avg Recency: 1.6, Avg Frequency: 29.8, Avg Monetary: 18491.88
Cluster 17 — Size: 9, Avg Recency: 1.3, Avg Frequency: 23.3, Avg Monetary: 10800.78
Cluster 18 — Size: 5, Avg Recency: 15.2, Avg Frequency: 26.8, Avg Monetary: 6193.25
Cluster 19 — Size: 8, Avg Recency: 13.2, Avg Frequency: 25.9, Avg Monetary: 7339.59
Cluster 20 — Size: 8, Avg Recency: 5.9, Avg Frequency: 24.5, Avg Monetary: 8706.29
Cluster 21 — Size: 20, Avg Recency: 3.0, Avg Frequency: 26.0, Avg Monetary: 12362.70
Cluster 22 — Size: 5, Avg Recency: 1.6, Avg Frequency: 29.8, Avg Monetary: 18491.88
Cluster 23 — Size: 6, Avg Recency: 1.3, Avg Frequency: 24.3, Avg Monetary: 11872.57
```

Figure 2: Cluster-Level Business Summary from Mapper Segmentation (RFM + log_CLV Lens)

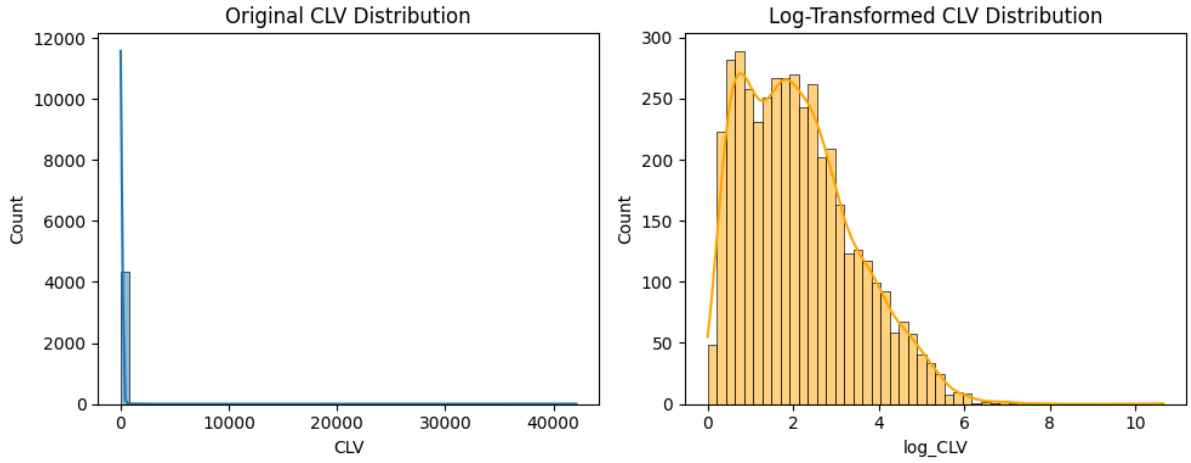

Figure 3: Effect of Log Transformation on CLV Distribution

The resulting Mapper graph, consisting of 24 nodes and 56 edges, offers a topological summary of customer behavior that reflects both local density patterns and smooth behavioral transitions across segments. Each node corresponds to a cluster of customers with similar RFM and log_CLV profiles, while the connections between nodes indicate overlapping behaviors across adjacent bins in the lens space. This structure reveals a nonlinear continuum of customer types, ranging from high-recent, low-frequency purchasers to low-recent, high-value repeat buyers. By examining average RFM statistics for each cluster, we can trace meaningful lifecycle trajectories, for example, identifying customers transitioning from high-frequency purchasing to inactivity or detecting isolated high-value cohorts with unusually high spending and frequent engagement.

The CLV and log_CLV distributions further validate the choice of log transformation. In figure 3, the left panel shows the original distribution of Customer Lifetime Value (CLV), which is highly right-skewed due to a small number of high-value customers. The right panel displays the log-transformed CLV (log_CLV), which produces a more balanced and approximately normal distribution. This

transformation improves numerical stability and enhances the effectiveness of downstream topological and statistical modeling techniques, including Mapper-based segmentation.. This preprocessing step not only enhances the Mapper lens but also ensures that rare but important customer patterns are properly integrated into the graph structure rather than suppressed or isolated by noise.

Crucially, the Mapper output provides a visually interpretable map of the customer landscape, something traditional clustering cannot achieve. Marketers can use this representation to pinpoint strategic clusters for retention, upselling, or re-engagement, while analysts gain a multiscale understanding of behavioral diversity within the customer base. This interpretability and structural insight make Mapper a powerful tool for segmentation-driven decision-making in e-commerce analytics.

ii. VISUALIZING PERSISTENT HOMOLOGY FOR E-COMMERCE CUSTOMER DATA

To further enhance the understanding of customer segmentation beyond Mapper's lens-driven approach, we incorporate Topological Data Analysis using persistent homology. Persistent homology is a mathematically rigorous framework that identifies and quantifies the significance of topological features, such as clusters and cycles, within high-dimensional data. By applying this approach to the e-commerce customer dataset, we can systematically reveal robust patterns and structures that persist across multiple scales, offering deeper insights into the intrinsic organization of customer behaviour.

In our implementation, we preprocess and standardize core behavioural metrics (Recency, Frequency, Monetary value, and log-transformed Customer Lifetime Value) for each customer, and select a manageable random sample for computational efficiency. Using the Vietoris–Rips filtration as implemented in the ripser library, we compute persistent homology up to dimension one, capturing both connected components (H_0) and loops (H_1). The output, visualized as a persistence diagram, summarizes the birth and death of topological features as the neighbourhood scale increases. Features that persist far from the diagonal in the diagram represent meaningful, noise-resistant customer segments or recurring behavioural cycles, while points close to the diagonal are typically considered noise. To provide a geometric baseline for comparison, a 2D PCA projection of the same standardized customer features is visualized alongside. While PCA captures linear variance and reveals coarse clustering structure, it fails to reflect multi-scale topological features such as loops or persistent groupings, which are effectively highlighted by persistent homology.

Due to technical limitations such as constrained RAM and lack of GPU acceleration in the working environment, a representative random sample of 1,000 customers was selected, and a 2D projection (Recency and log_CLV) was used to compute persistent homology efficiently while preserving meaningful behavioural structure.

CODE SNIPPET:

```
from ripser import ripser
from persim import plot_diagrams
from sklearn.preprocessing import StandardScaler

# Select behavioral metrics and scale
X = rfm_sample[['Recency', 'Frequency', 'Monetary', 'log_CLV']]
X_scaled = StandardScaler().fit_transform(X)

# Use Recency and log_CLV for persistent homology
X_tda = X_scaled[:, [0, 3]]  # Axis 0 = Recency, Axis 3 = log_CLV
tda_result = ripser(X_tda, maxdim=1)  # Compute H0 and H1
plot_diagrams(tda_result['dgms'], show=True, title="Persistence Diagram: Customer Segmentation")
```

OUTPUT:

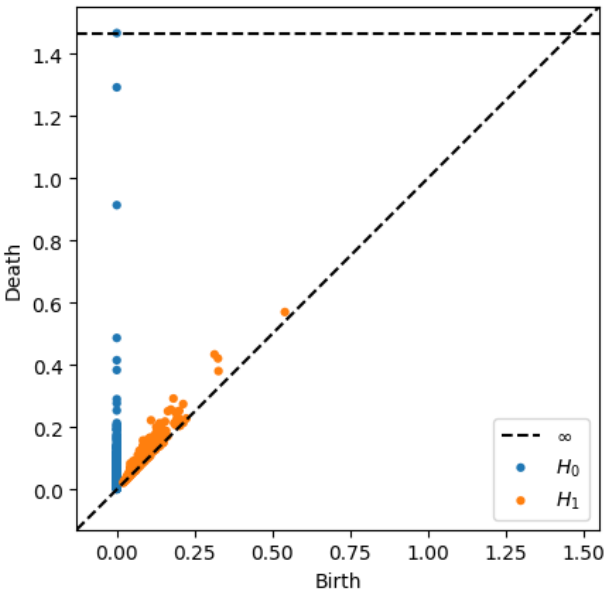

Figure 4: Persistence Diagram of Customer Behavioural Features

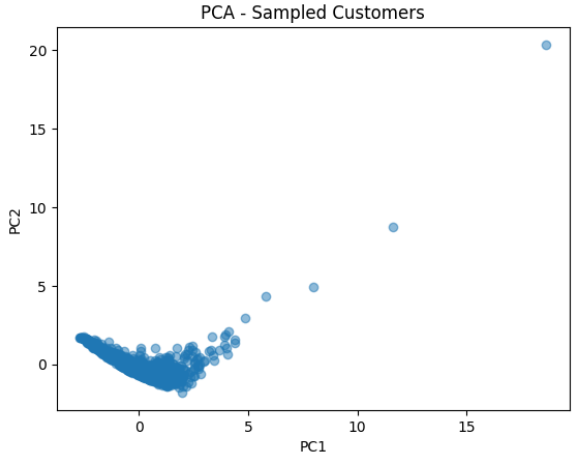

Figure 5: 2D PCA Projection of Standardized Customer Features

The persistence diagram (Figure 4) illustrates the key topological features of the customer data, displaying blue points for H_0 (connected components or clusters) and orange points for H_1 (loops or cycles). Blue points that are far from the diagonal indicate robust customer segments, groups of customers whose behaviour remains distinct across a range of similarity thresholds. Orange points represent cyclic relationships or recurring patterns in customer behaviour, which may correspond to seasonal trends or cyclical engagement. Points close to the diagonal are typically considered noise. For comparison, a 2D PCA scatterplot (Figure 5) of the same sampled data provides a geometric perspective, illustrating the spread and potential clustering of customers, but lacks the multi-scale, topological insights that TDA offers.

This topological perspective is particularly valuable for e-commerce analytics. Persistent homology provides a scale-invariant and noise-robust summary of customer segmentation, complementing Mapper's intuitive graph-based view. By quantifying the "lifetimes" of behavioural patterns, TDA allows analysts to distinguish statistically significant customer groups from spurious ones, and to detect subtle but important structures, such as cyclical buying habits, that traditional clustering or dimensionality reduction may overlook. As a result, persistent homology equips businesses with actionable, interpretable insights for targeted marketing, retention, and lifecycle management.

iii. Feature Importance Analysis Using Classical and Topological Data Analysis

In this section, we focus on understanding which aspects of customer purchasing behaviour are most influential in predicting whether a customer will make a purchase in the following month. To achieve this, we extract two types of features from the transaction history of each customer: classical time series features and topological features derived using Topological Data Analysis (TDA). Classical features include statistical descriptors such as the mean and standard deviation of monthly spending, the minimum and maximum amounts spent, the skewness of spending, the number of months with at least one purchase, the time elapsed since the last purchase, the largest drop in spending between months, and the overall trend in spending as measured by linear regression. These features are well-known in customer analytics and provide interpretable summaries of each individual's purchasing pattern over time.

To complement these classical features, we incorporate TDA features, which are designed to capture complex temporal structures in the data that may not be evident through simple statistics. Specifically, we transform each customer's purchase time series into a point cloud using delay embedding, and then compute persistence diagrams to summarize the birth and death of topological features (such as connected components and loops) in the embedded data. From these diagrams, we extract summary statistics like the number of persistent features and the maximum and total lifespan of these features in both the zero-dimensional (H_0) and one-dimensional (H_1) homology groups. By combining these advanced TDA features with classical features, we create a rich, multidimensional representation of each customer's behavioural history.

We then use a Random Forest classifier to model the relationship between these features and the likelihood of a customer making a purchase in the next month. The model's feature importance scores are visualized in a horizontal bar plot, which reveals which features are most predictive in this context. This analysis provides actionable insights for e-commerce businesses, as it highlights which behavioural signals are most relevant for targeting marketing efforts, designing retention strategies, or personalizing customer experiences. Moreover, by explicitly including TDA-derived features, we test whether more sophisticated mathematical representations of customer behaviour offer predictive value beyond traditional metrics, a valuable consideration for businesses seeking to leverage advanced analytics for competitive advantage.

CODE SNIPPET:

```
def delay_embed(ts, dim=2, tau=1):
    if len(ts) < (dim - 1) * tau + 1:
        return np.empty((0, dim))
    return np.column_stack([ts[i:len(ts)-(dim-1)*tau+i] for i in range(0, dim*tau, tau)])

def tda_summary(ts):
    emb = delay_embed(ts)
    if emb.shape[0] == 0:
        return [0, 0, 0, 0, 0, 0]
    dgms = ripser(emb, maxdim=1)['dgms']
    h0 = dgms[0]
    h1 = dgms[1]
    h0_life = h0[:, 1] - h0[:, 0] if len(h0) else np.array([0])
    h1_life = h1[:, 1] - h1[:, 0] if len(h1) else np.array([0])
    return [len(h0), np.max(h0_life), np.sum(h0_life), len(h1), np.max(h1_life), np.sum(h1_life)]

tda_features = pd.DataFrame([tda_summary(row) for row in pivot.values], index=pivot.index)
```

OUTPUT:

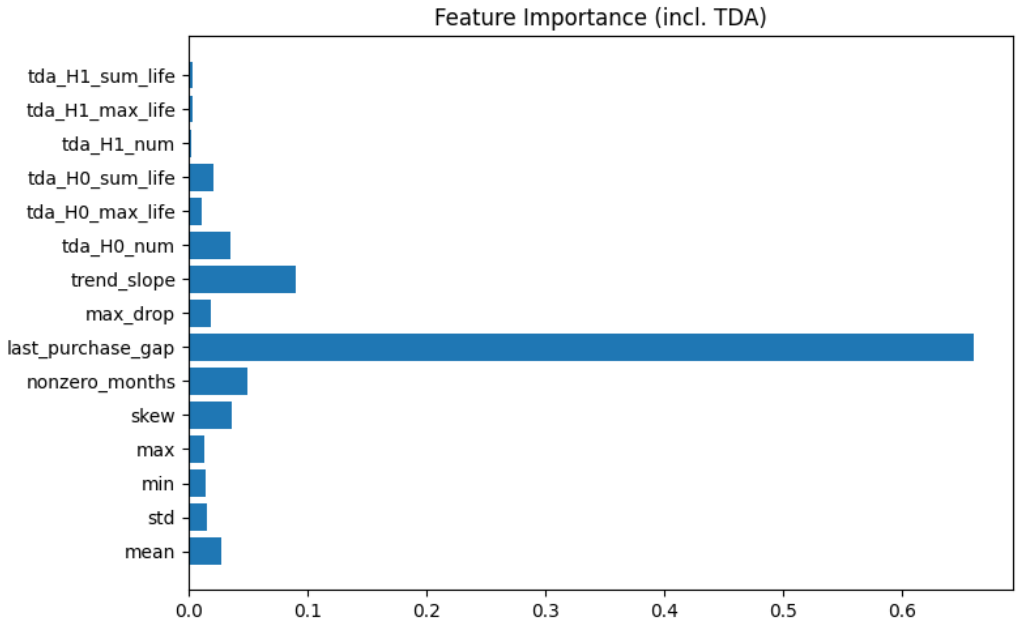

Figure 6: Feature Importance for Next-Month Purchase Prediction Using Classical & Topological Features

The results of the feature importance analysis reveal that classical time series features, particularly those related to recency and purchasing frequency, are the dominant predictors of whether a customer will make a purchase in the following month. The feature importance plot shows that attributes such as the time since last purchase, the number of active (non-zero) months, and the slope of recent spending trends have the highest influence in the Random Forest model's predictions. In contrast, the topological features derived from TDA, including the number and persistence of H_0 and H_1 components, contribute relatively little to the model's predictive power within this dataset. This suggests that, for the customer purchase histories available, traditional statistical descriptors capture most of the relevant behavioural patterns, while the additional complexity offered by TDA does not provide substantial new information. These findings help prioritize which features are most useful for targeting retention strategies or forecasting customer activity in e-commerce applications. This outcome is also consistent with domain knowledge, as last_purchase_gap (a proxy for recency) is a well-established predictor of churn and customer inactivity. However, this does not diminish the potential value of TDA; in more temporally complex or noisier datasets, topological features may capture structural patterns that classical features alone cannot.

V. CONCLUSION

Recent advances have shown that topological approaches can offer unique value in analysing high-dimensional customer data within e-commerce settings. Traditional methods such as clustering and dimensionality reduction often struggle to capture the underlying structure of behavioural datasets, especially in the presence of sparsity or nonlinearity. In contrast, topological methods preserve global structure while remaining robust to noise, offering new perspectives for understanding customer segmentation, lifecycle stages, and complex interaction patterns.

This report presented a series of implementations that applied Topological Data Analysis (TDA) techniques to the UK Online Retail dataset. Using Mapper, we constructed interpretable simplicial graphs that revealed nonlinear customer segments based on Recency, Frequency, Monetary value, and a custom Customer Lifetime Value formulation. Persistent homology was used to extract multi-scale topological features from the same data, allowing us to distinguish meaningful behavioural patterns from noise. We also evaluated the relative predictive power of classical statistical features versus

topological descriptors using a Random Forest classifier. Throughout, the implementation emphasized interpretability and visual analysis, and all results were obtained using efficient code pipelines designed to run within resource-constrained environments.

Our results demonstrate that while classical features remain highly predictive in forecasting near-term customer activity, topological features provide an interpretable and complementary view that captures deeper geometric and temporal structures. These insights open the door to future applications such as anomaly detection, churn modelling in sparse time series, or hybrid systems that fuse topological reasoning with deep learning. As customer data continues to grow in complexity, TDA offers a promising toolkit for extracting actionable, structure-aware insights in e-commerce and beyond.

VI. REFERENCES

[1] G. Singh, F. Mémoli, and G. Carlsson, “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition,” in *Eurographics Symposium on Point-Based Graphics*, 2007.

[2] Gunnar Carlsson. "Topology and Data." Bulletin of the American Mathematical Society, 2009.

[3] F. Chazal and B. Michel, “An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists,” *arXiv preprint arXiv:1710.04019*, revised version published 2021.

[4] M. Carrière, B. Michel, and S. Oudot, “Statistical Analysis and Parameter Selection for Mapper,” *Journal of Machine Learning Research*, vol. 19, no. 58, pp. 1–39, 2018.

[5] G. Tauzin, J. Burella Pérez, A. M. Medina-Mardones, U. Lupo, M. Caorsi, A. Dassatti, L. Tunstall, W. Reise, and K. Hess, “giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration,” *Journal of Machine Learning Research*, vol. 22, no. 39, pp. 1–6, 2021. [arXiv:2004.02551]

[6] S. Gholizadeh and W. Zadrozny, “A Short Survey of Topological Data Analysis in Time Series and Systems Analysis,” in *Proceedings of the 7th International Conference on Complex Networks and Their Applications*, pp. 231–242, 2018.

[7] P. Dłotko, K. Hess, J. P. Smith, M. Massara, and T. Di Matteo, “Topological Data Analysis of Financial Time Series: Landscapes of Crashes,” *Physica A: Statistical Mechanics and its Applications*, vol. 523, pp. 691–700, 2019.

[8] C. Lakshminarayan and M. Yin, “Topological Data Analysis in Digital Marketing,” in *Proceedings of the 2020 IEEE International Conference on Big Data (Big Data)*, pp. 4503–4511, 2020.

[9] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, “Deep Learning with Topological Signatures,” in *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1633–1643, 2017.

[10] P. Bubenik, “Statistical Topological Data Analysis using Persistence Landscapes,” *Journal of Machine Learning Research*, vol. 16, pp. 77–102, 2015.

[11] Dey, T. K., Wang, Y., & Wang, Y. (2022). Computational Topology for Data Analysis. Cambridge University Press.

[12] Edelsbrunner, H., & Harer, J. (2010). Computational Topology: An Introduction.

[13] De Silva, V., & Carlsson, G. (2004). "Topological estimation using witness complexes." Symposium on Point-Based Graphics.

[14] Bendich, P., Marron, J. S., Miller, E., Pieloch, A., & Skwerer, S. (2016). "Persistent homology analysis of brain artery trees." The Annals of Applied Statistics, 10(1), 198–218.

[15] Wasserman, L. (2018). "Topological Data Analysis." Annual Review of Statistics and Its Application, 5, 501–532.

[16] Oudot, S. (2015). Persistence Theory: From Quiver Representations to Data Analysis.

[17] Lum, P. Y., et al. (2013). Extracting insights from the shape of complex data using topology. Scientific Reports, 3.