WHITEPAPER

Topological Data Analysis and Machine Learning: Better Together

SymphonyAI

Overview

SymphonyAI Group's award-winning Ayasdi AI platform powers the design, development, and deployment of enterprise-scale, intelligent applications to deliver extraordinary business value. This paper details the underlying topological data analysis (TDA) technology and how it interacts with and enhances other unsupervised and supervised machine learning technologies and approaches.

As a technology, TDA can distill business value from large, complex datasets. Global enterprises increasingly look to data to make decisions that can affect millions of lives and billions of dollars of revenue. That can be a challenge given the explosion of data – log, sensor, social, health, and financial. Aggregate data growth is an exponential function with time. Unfortunately, we cannot train enough data scientists to meet this runaway curve. This gap is driving scientists and mathematicians to examine approaches such as TDA to improve the quality and the speed of analytics.

Topology, TDA, and the Ayasdi AI Platform

Topology is a mathematical discipline that studies shape. TDA is the adaptation of this discipline to analyze highly complex data. It draws on the philosophy that all data has an underlying shape, and that shape has meaning.

The Ayasdi AI platform draws together TDA and a broad range of machine learning, statistical, and geometric algorithms to create a visual representation of all the data points in a large data set. This visual representation helps to rapidly uncover critical patterns and relationships. By identifying the geometric relationships that exist between data points, TDA offers a straightforward and efficient way of partitioning data to understand the underlying properties that characterize the segments and sub-segments within that data.

Figure 1: TDA creates a compressed representation of data to uncover patterns and subgroups of interest

As the only commercially available TDA implementation, the Ayasdi AI platform provides organizations with tools that greatly enhance their ability to process and draw meaning from data that is highly resistant to other methods of analysis and interpretation.

The Ayasdi AI platform supports application technologies that facilitate the design, development, and deployment of intelligent AI applications. These applications serve use cases in vertical markets such as healthcare, financial services, and the public sector.

The Ayasdi AI platform supports application technologies that facilitate the design, development, and deployment of intelligent AI applications.

The Promise of Machine Learning

Machine learning is a class of algorithms that adjust and learn from data and then take or suggest actions. Machine learning allows companies to segment existing data into meaningful groupings, identify key segmentation attributes and features, find patterns and anomalies, and precisely classify new data points as they arrive.

There are two classes of machine learning techniques – supervised and unsupervised. Supervised learning constructs predictive models by using a training data set to create a function that can accurately infer outputs when presented with new input data. Unsupervised learning helps discover the hidden structure in data; it utilizes only the content of input data and does not know the expected output. Both classes help to drive new revenue streams, forge stronger customer relationships, predict risk, improve medical outcomes, and prevent fraud. However, analyzing complex data using only these methods is constrained by intrinsic issues and dependency on scarce machine learning expertise.

Machine Learning – Mind the Gap

Unsupervised learning algorithms, which do not

know expected results, fall into two categories:

Clustering algorithms discover the underlying sub-segments within data by grouping sets of data points so that those in the same group (called a cluster) are more similar to each other than to those in other clusters.

Dimensionality reduction algorithms reduce the number of properties or attributes (represented by data columns) required for describing each data point while retaining the inherent structure of the data.

Unsupervised Learning – Clustering

Clustering methods segment a dataset into smaller datasets. Different clustering algorithms rely on various techniques to cluster data. In a hierarchical clustering algorithm such as single-linkage clustering, each data point starts as its own cluster. The single-point clusters combine into larger clusters of points by sequentially fusing data point pairs that are the most similar to each other. The process continues until all data points are fused into a cluster. The resulting cluster hierarchy can be visualized as a dendrogram (tree diagram) that shows which clusters were fused to produce new clusters. Knowing the sequence and distance at which cluster fusion took place can help determine the optimal scale for clustering. Three issues limit the effectiveness of clustering algorithms:

1. **The number of clusters –** Some clustering algorithms require that the number of clusters be determined in advance. While a machine learning expert might use informed criteria (such as a "Bayesian information score") to make an educated guess at the number of clusters, typically this is an arbitrary choice that can greatly impact conclusions drawn from the data.

2. **Continuous data sets and multi-modality -** Clustering methods work well when data sets decompose cleanly into distinct groups that are well separated. However, many data sets are continuous and exhibit progressions rather than sharp divisions. Clustering methods can create spurious divisions in

The Ayasdi AI platform supports application technologies that facilitate the design, development, and deployment of intelligent AI applications.

such data sets, thereby obscuring the real underlying structure of the data. Most algorithms implicitly assume that each cluster that we are looking for is equally dense. This is rarely the case in real datasets.

3. Shape assumptions – Many algorithms have implicit or explicit constraints on cluster shape. For example, k-means algorithms assume clusters are spherical. Model-based algorithms assume a model of the shape of the cluster.

Unsupervised Learning – Dimensionality Reduction

Dimensionality reduction methods make it easier to visualize data sets with a large number of data columns. For example, credit card transactions have thousands of attributes, each represented as a data column, and so visualizing these transactions can be challenging. Principal component analysis is a good example of a dimensionality reduction algorithm. Other methods include multi-dimensional scaling, Isomap, T-distributed stochastic neighbor embedding, UMAP embedding, and Google's Pagerank.

Dimensionality reduction methods are powerful since they can reduce the number of dimensions required to describe data while revealing some inherent structure in that data. However, two issues hamper dimensionality reduction methods:

1. Projection loss – Dimensionality reduction methods compress a large number of attributes down to a few. As a result, well-separated data points in the high-dimensional space might appear as neighbors in a projection. Distinct clusters might overlap. This increases the chances of missing out on important insights.

2. Inconsistent results - Distinct dimensionality reduction algorithms produce dissimilar projections because they encode different assumptions. None of the results are wrong; they are simply unlike one another because different algorithms accentuate

Dimensionality reduction methods are powerful since they can reduce the number of dimensions required to describe data while revealing some inherent structure in that data.

different aspects of the data. Relying on a single algorithm might result in missed critical insights.

Supervised Learning - Regression and Classification

Supervised learning algorithms are used for producing predictive models. There are two types: **regression and classification**. Regressors predict real-valued variables such as profit margins or stock prices. Classifiers predict discrete variables such as fraud or customer churn. Examples include linear and logistic regression, support vector machine, and artificial neural networks. There are two phases of supervised learning:

1. **Training -** The algorithm analyzes a training data set using historical data to produce parameters for a function that will infer results when presented with new input data. The historical data provides empirical, correct evidence ("ground truth") against which to verify the function accuracy. By definition, the training set contains data used to modify or tweak the function as needed to improve its accuracy.

2. **Prediction -** the function produced in the training phase is used to predict the values for new input data points.

Supervised learning algorithms face three inherent issues:

1. **Data Shape Assumptions -** The choice of algorithm entails an assumption about the shape of the underlying data. For example, linear regression assumes the data is planar (though possibly higher dimensional) and tries to find the best plane that fits the data. If the actual shape of the underlying data is not planar, then the analysis will produce incorrect results.

2. **Global optimization -** All supervised learning algorithms try to find parameters for a function that best approximate all of the data. However, data is rarely homogeneous, so it is unlikely a single shape fits the entire dataset.

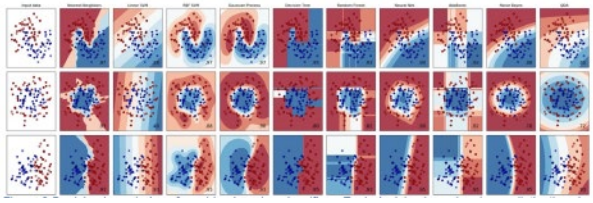


Figure 2 Decision boundaries of machine learning classifiers: Each decision boundary has a distinctive shape. Without knowing the shape of the underlying data, the algorithm might produce poor results. https://scikit-learn.org/stable/auto_examples/classification//plot_classifier_comparison.html

3. Generalization - A model may perform well with test data but produce inaccurate results with new data. A generalization error—also known as overfitting—occurs because the model has more parameters than required.

As a result, despite its promise, supervised learning faces limitations. Successful implementations require scarce machine learning experts. It is easy to miss important insights in data by choosing the wrong algorithm or by not trying enough algorithms or hyperparameters. The required large quantities of high-quality labeled training data can be difficult, expensive, or impossible to attain. Adding TDA to existing machine learning methods can overcome these limitations.

How TDA Improves Machine Learning Algorithms

All machine learning methods produce functions or maps. For example, **clustering** maps each input data point to a cluster. **Dimensionality reduction** maps each input data point to a lower-dimensional data point. **Supervised learning** maps each input data point to a predicted value.

TDA uses all of these functions to dramatically increase the effectiveness of machine learning simultaneously to process input and thereby produce superior quality output. Additionally, TDA can add future machine learning algorithms to further improve its analysis of large, complex data sets.

TDA uses all of these functions to dramatically increase the effectiveness of machine learning simultaneously to process input and thereby produce superior quality output.

Unsupervised Learning - Clustering

TDA uses clustering as an integral step in building a network representation of data. Rather than trying to find disjoint groups, TDA applies clustering to small portions of data. It combines these "partial clusters" into a network representation that shows the similarity between the data points. TDA is more appropriate for constructing a connected representation of continuous data sets or data with heterogeneous densities.

Most clustering algorithms rely on global optimization techniques (including HDBSCAN), susceptible to noise since they consider all data during optimization. TDA splits data into multiple independent parts using lens functions and runs clustering algorithms within each part of the data independently. The multiple local optimizations executed by TDA dramatically reduce the effect of noise on the final results.

Unsupervised Learning - Dimensionality Reduction

TDA supports the automatic execution and synthesis of dimensionality reduction algorithms. However, TDA eliminates the projection loss issue typical of dimensionality reduction where well-separated data points in higher dimensions end up overlapping in a lower-dimensional projection. This is achieved by clustering the data in the original high dimensional space. As a result, well-separated data points in the original space will typically still be well separated in the output. It is therefore easy to identify distinct segments and sub-segments within data that might have been missed. TDA also automatically synthesizes the results of different dimensionality reduction algorithms into a single output. This eliminates the need to know or guess the correct sets of assumptions for a particular dimensionality reduction method.

Supervised Learning - Regression and Classification

TDA augments supervised learning algorithms by eliminating systematic errors and optimizing for local data sets. Most supervised learning algorithms are based on global optimization that assumes a shape for the underlying data and tries to discover parameters that best approximate that data. This can lead to mistakes in some data regions. TDA uses this output as an input to discover where errors are being made. Similarly, TDA constructs a collection of models responsible for a different segment of data, rather than making global assumptions about it. This eliminates the need to create a single model

TDA augments supervised learning algorithms by eliminating systematic errors and optimizing for local data sets.

that works well on all of the data—a difficult, if not impossible, mission. This method generates more accurate results and can incorporate any supervised algorithm. Finally, TDA used on the feature space (that is, the transpose of the original data) can substantially improve the convergence capabilities of advanced machine learning techniques such as neural networks.

TDA reduces the possibility of missing critical insights by reducing the dependency on data scientists choosing the right algorithms. It uses current machine learning techniques to find subtle patterns and insights in local data. In general, TDA enhances any algorithm with which it is paired.

Creating Topological Networks with TDA

TDA identifies data points related to each other and pieces them together to build a global, compressed summary in the form of a network. This network can be executed programmatically or visualized for further investigation.

In a visual network, TDA consistently applies a function (call it f) to the data while using a measure of similarity to generate a compressed representation of the data. The resulting visualization consists of nodes wherein each node represents data points with similar function values clustered together. Two simple examples illustrate this. The first steps through the general methodology and the second demonstrates how TDA enhances machine learning.

Example 1: In Figure 3, a circle in the xy-plane represents a data set. A function f maps each point in the data set to its y-coordinate value, as shown on the right. TDA then subdivides the function image into overlapping sets of nearby values. The points are divided into four overlapping groups that have similar y-coordinate values (Figure 4).

Next, TDA clusters each group of data points independently using a measure of similarity. In this example, similarity is defined using standard Euclidean (straight line) distance. Each cluster is shown as a node. A node represents a set of

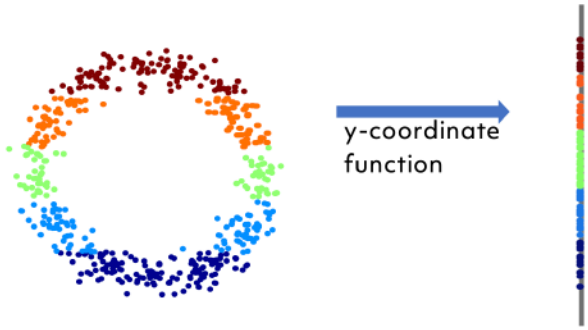


Figure 3: Using a function to map data points in the shape of a circle to their y-coordinate values

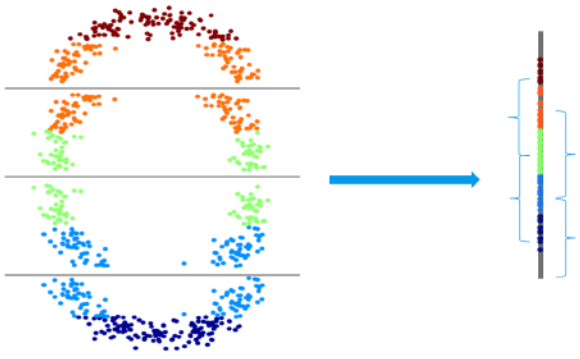


Figure 4: Dividing data points into overlapping sets with similar y-coordinate values

data points that have a measure of similarity (Euclidean distance) and the function value (y-coordinate) in common. The size of each node reflects the number of data points. In Figure 5, the top node represents both red and orange data points. The second set of data points from the top in the original circular pattern contains two distinct regions that produce two separate nodes in the topological image at right.

Nodes with data points in common are connected by edges in the network. Since the data set was divided into overlapping sets, a data point can be represented in multiple nodes. In Figure 6, the orange data points on the left are represented in the top red node and the orange node on the left. These nodes are connected by an edge because they contain data points in common. The resulting network is a compressed representation of the original data set that retains its fundamental circular shape. The network is much simpler to visualize and work with, yet it captures the essential behavior of the data.

Example 2: In the second example, a data set is sampled in the two-dimensional Euclidean plane from four Gaussian distributions. In Figure 7, the data points are colored by the values of the density estimator function. TDA divides the data set into overlapping groups with similar function values—in this case, density estimations. Each data subset is clustered to create nodes that represent data points with similar function values.

The resulting network in Figure 7 captures both the overall structure of the data and its fine-grained behavior. The four flares in the network correspond to the four regions of varying densities. The flares connect to each other because these contain common data points with varying degrees of density. Standard machine learning techniques would have identified the four regions but lost the continuous transitions between them. TDA captures both the differences and similarities.

Complex data holds useful information that could go undetected when using standard machine learning and statistical techniques. By contrast, the Ayasdi AI platform begins by understanding data on a small scale. It then stitches together these pieces of information to

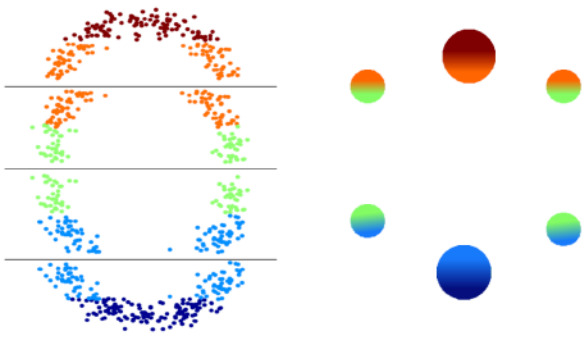


Figure 5: Nodes Represent Clusters of Data Points with Similar Function Values and Measures of Similarity

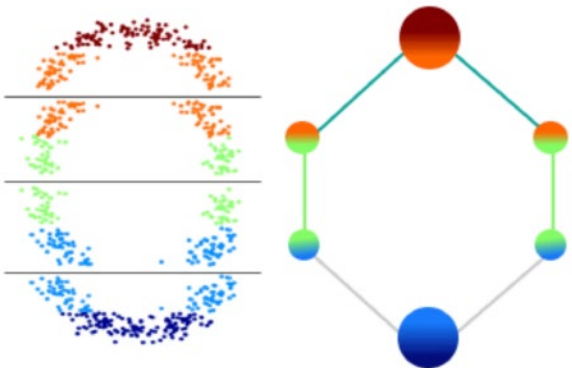


Figure 6: Nodes with data points in common are connected by edges to form a network

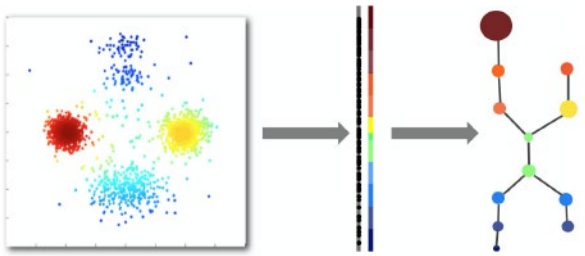


Figure 7: TDA Enhances Machine Learning by Capturing the Overall Structure and Fine-Grained Behavior in a Data Set

create a topological summary, or compressed representation, of the entire data set. This is another instance of TDA's ability to create networks that display subtle insights in the data while also showing the global behavior of the data.

TDA can incorporate virtually any machine learning, statistical, or geometric technique and apply them as a function f to display as a visual network a compressed facsimile of a data set. Principal component analysis, autoencoders, random forests, and density estimators are some of the functions that TDA can use to derive insights from large, complex data sets.

TDA can incorporate virtually any machine learning, statistical, or geometric technique and apply them as a function f to display as a visual network a compressed facsimile of a data set.

How TDA Makes Complex Data More Intelligible

The Ayasdi AI platform uses features that support the discovery of otherwise hidden information in data sets. These include segmentation, feature discovery, classification, model creation, model validation, and anomaly detection. Each is discussed below.

Segmentation and Clustering

Data segmentation groups data points that are more similar to each other. Typical segmentation approaches involve either a data scientist manually generating and testing hypotheses or the use of clustering. Manual testing can be a huge undertaking, even when dealing with small data sets. Typically, a domain expert chooses a logical data attribute to create segments. This approach may have limited utility because it does not consider key information. For example, segmenting customers by the amount of money spent might seem appropriate, but it ignores the impact of key factors such as demographics.

By comparison, standard clustering methods for segmentation produce better results. However, these methods still suffer from the limitations described earlier: a need to know the number of clusters before applying the algorithm, the unsuitability for tackling continuous data sets, and the assumption of spherical shape or uniform

densities.

An example illustrates this: a financial institution segments its clients by investment behavior under specific market conditions and then precisely targets them at the right time with tailored recommendations. Such an approach relies on macro trends to explain client behavior, yet it can miss subtle trends tied, for example, to specific regional events. An event might result in a particular group of clients trading in a specific class of products that diverges from the general trend but is never identified because of the segmentation approach applied. Such subtle trends are more likely if the number of clients exhibiting a particular behavior is small compared to the total number of clients.

In contrast, TDA will discover that while these regional investors are similar to the majority of clients in the data, they are more similar to each other than to the majority. This subtle signal is captured in TDA output as a flare in the visual network and encoded in an application that alerts the bank's sales force. Sales can then prioritize and target these clients with tailored recommendations, rather than those for the majority of the clients.

Anomaly Detection

The model validation description above relies on the availability of predicted outcomes and accurate training data. There are cases in which this information is not readily available. An alternative approach, anomaly detection, does not require the existence of predicted outcome and ground truth information. The Ayasdi AI platform supports anomaly detection by using a transaction data set that does not require accurate training data from current models. Regions of the resulting network topology that represent low-density points or points far away from the central core of the data set are "anomalous," with less in common than data points in network nodes.

Sales can then prioritize and target these clients with tailored recommendations, rather than those for the majority of the clients.

Feature Discovery

Understanding the underlying features or attributes of the data that drive segmentation can be invaluable when pinpointing the factors that impact business outcomes. TDA helps feature discovery by automatically producing a list of the attributes (data columns) that drive segmentation, ranked in order of statistical significance.

For example, to understand the reasons for customer churn, spotting the root causes is significantly more critical than prediction as it often brings systemic issues to the surface. This involves identifying the attributes of departing customers by creating clusters and node groups that can be used to capture underlying features.

Recommendation Engines

Recommendation engines precisely target customers with sales efforts for products and services purchased by other customers with similar profiles. The Ayasdi AI platform is an ideal foundation for a recommendation engine. It enables creating precise sub-segments of a customer base by correlating and analyzing a wide range of data, including demographics, buying behavior, market, CRM, and social media information. Figure 8 shows this information distilled into a topology for additional analysis.

Model Creation

Supervised learning models can predict future actions or behavior. TDA supports multiple different algorithms to a data set when creating a topological model. The result is a group or collection of models, referred to as piece-wise models or an ensemble of models, that best represent all of the data. This ensemble of models tends to provide a far more accurate representation of the data since each is optimized for one or more different segments of the data. Creating a topological model uses standard supervised learning methods like linear regression applied to network node groups to predict the placement of newly arrived data points accurately.

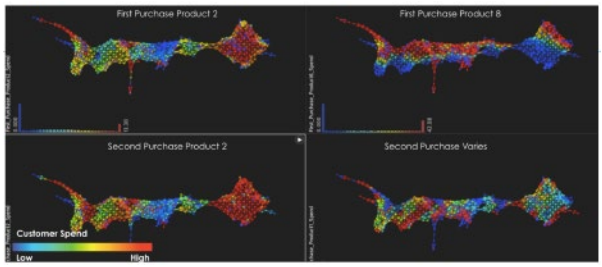


Figure 8: Analyzing Returning Customers by Buying Patterns and Spend

Model Validation

Most organizations rely on many automated models to help with a variety of predictive and investigative business tasks. One of the primary steps involved in validation or auditing exercises is the discovery of systematic errors or biases in a model. Typically, models created by supervised learning algorithms produce systematic errors due to incorrect assumptions about the shape of the underlying data. TDA uncovers these errors in models.

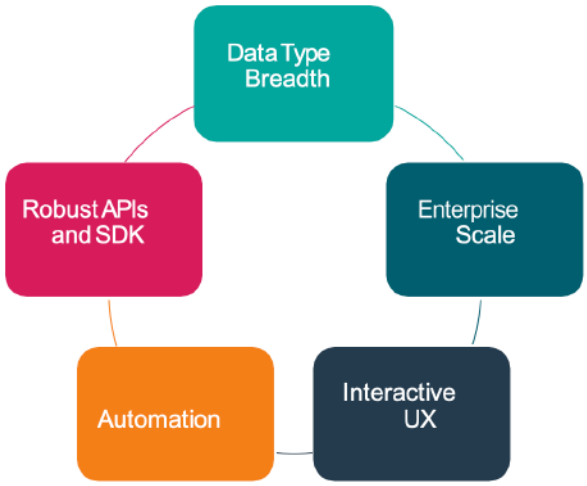
Consider the process of validating models used to detect fraud in credit card transactions. Identifying issues in these models using TDA involves using known fraudulent transactions to create a topological network. This network highlights nodes that do not conform to predicted outcomes or actual results to identify questionable transactions that were not discovered by the existing fraud detection model.

The Ayasdi AI Platform vs. Open Source

The AyasdiAI platform provides unique machine learning and TDA capabilities to solve some of the world's most challenging problems. It represents more than $50 million invested in the forefront of innovation over the past decade. A data scientist may claim to perform the same functions and capabilities but will always fall short in a real-world setting. The AyasdiAI platform is superior to open source in the following ways:

Data Type Breadth: The Platform can handle both text and numeric data with sparsity: integrating large amounts of heterogeneous and diverse data sets into common data and modeling environment at enterprise scale.

Enterprise Scale: The Ayasdi AI platform optimally analyzes large datasets, millions of rows, and hundreds of thousands of columns. Open source TDA cannot handle large datasets with consistency.



Interactive UX: The Ayasdi AI platform's Workbench interactive UI allows for unparalleled rapid discovery. Users can interact with data, quickly understand subpopulations, and compare them for insights. All open-source code bases lack this rapid discovery interface.

Automation: AyasdiAI's Platform has a suite of auto topological model generation algorithms that intelligently evaluates a large space of metrics and filters to suggest the most optimum models. These automation techniques are beneficial for both supervised and unsupervised scenarios.

Robust APIs and SDK: The Ayasdi AI platform's APIs and SDK give users access to a large, extensively tested code base for fast prototyping and easy integration into applications.

Tested and Proven

Our years of experience working with the US government and financial and health organizations have proven our data security infrastructure and scale readiness time and time again. The AyasdiAI platform has been tested in real-world applications to deliver capabilities including high availability, user management, data security, team collaboration, and support for multi-tenancy up to 1000 users. Open source can never provide these capabilities. Additionally, the Ayasdi AI platform provides the needed support for enterprise-grade operations:

A head-to-head comparison between the Ayasdi AI platform and open-source data science tools always yields the same conclusion: open source does not stack up. It will always fall short in many areas crucial for enterprise-scale data science.

Summary

While organizations have successfully tackled the challenge of storing and querying vast amounts of data, they continue to lack the tools and techniques for extracting useful and predictive information from highly complex data sets. Topology and TDA are well suited for analyzing complex data with potentially millions

of attributes.

The Ayasdi AI platform uses TDA to bring together a broad range of machine learning, statistical, and geometric algorithms to create compressed representations of data. This advanced analytics platform creates highly interactive visual networks that enable rapid exploration and understanding of critical patterns and relationships in data. The Ayasdi AI platform's use of TDA augments current machine-learning techniques by ameliorating issues and reducing the dependency on scarce human expertise. The platform provides superior tested and proven enterprise-scale capabilities to solve a broad range of use cases. Innovative organizations are using TDA and the Ayasdi AI platform to:

1. Precisely segment their data
2. Identify the underlying features that drive segmentation
3. Create more effective predictive models and tailor product recommendations
4. Develop, validate and improve models
5. Detect subtle anomalies in data sets

SymphonyAI Group

The SymphonyAI Group is the fastest growing and most successful group of B2B AI companies, backed by a $1 billion commitment to build advanced AI and machine learning applications that transform the enterprise. Symphony AI is a unique operating group of over 1,900 skilled technologists and data scientists, successful and proven entrepreneurs, and accomplished professionals, under the leadership of one of Silicon Valley's most successful serial entrepreneurs, Dr. Romesh Wadhwani.

For more information, visit
www.symphonyai.com/governmentsolutions

SymphonyAI