Symphony AYASDI

# TDA and Machine Learning: Better Together

WHITEPAPER

# Overview

AyasdiAI's award-winning artificial intelligence platform powers the design, development, and deployment of enterprise-scale financial services solutions. Our approach, underlying technology and products have been designed to fight online financial crime in all its forms. This paper details our Topological Data Analysis (TDA) technology, how it interacts with and how it enhances other machine learning technologies.

**Innovative**
The financial services industry is long overdue for an innovation that stays ahead of financial criminals. Our technology can be used sector-wide to support the anti-fraud and financial crime-fighting goals of firms, regulators, customers, partners, and governing bodies.

**Transparent**
One of our core values is the application of transparency from the top down. Our TDA technology is designed to make every potential risk event and behavioral profile understandable to our customers, empowering them to fully understand the scope and scale of the risks they face—even as those risks change over time.

Transparency also means being upfront and driven by results. To that end, we commit to customer-driven success factors at the start of every project and hold ourselves accountable to every metric.

**Trustworthy**
We know that our customers are personally held responsible for financial crimes discovered at their organizations, so we know that failure is not an option. Our track record of crimes discovered, efficiencies achieved, and promises fulfilled has earned us backing by DARPA and the trust of many of the world's largest global banks to protect their businesses and reputations.

Global enterprises increasingly look to their data to make decisions that can affect millions of lives and billions of dollars of revenue, and that's where our TDA technology comes in. It distills business value from large, complex datasets that includes information from activity logs, monitoring sensors, and raw financial data. In any given dataset, the number of potential insights derived from the data is an exponential function of the number of data points. We also treat aggregate data growth as an exponential function over time.

Unfortunately, we cannot train enough data scientists to meet this runaway, double-exponential demand curve. This is driving scientists and mathematicians to examine new approaches, such as TDA, to improve both the quality and speed of their analytics engines.

## Introducing the AyasdiAI Machine Intelligence Platform

This platform is built on the mathematical concept of topology, which studies shape and TDA adapts this discipline to analyze highly complex data. It draws on the philosophy that all data has an underlying shape, and that shape has meaning. AyasdiAI's approach to TDA is embodied in the AyasdiAI Machine Intelligence Platform.

The platform draws together a broad range of machine learning, statistical, and geometric algorithms to create a summary or compressed representation of all the data points in a large data set and thus to rapidly uncover critical patterns and relationships in that data set. By identifying the geometric relationships that exist between data points, TDA offers an extremely simple and efficient way of partitioning data to understand the underlying properties that characterize the segments and sub-segments that lie within that data.

AyasdiAI's platform is the only commercially available implementation of TDA. By marrying TDA and machine learning, AyasdiAI creates an infrastructure that greatly improves deriving meaning from data that is highly resistant to other methods of analysis and interpretation.

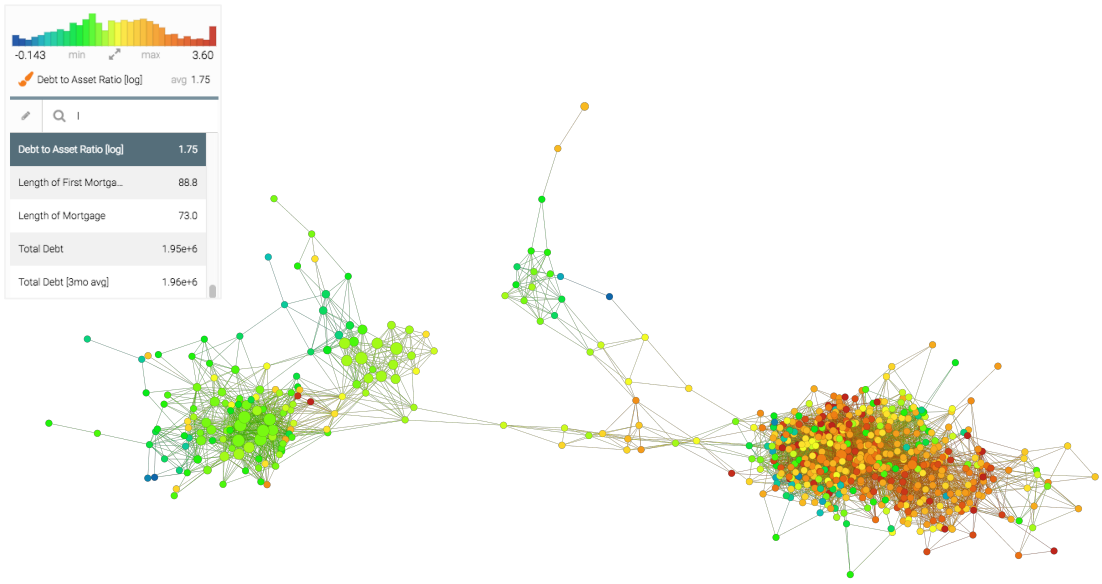



*Figure 1: Creating a Compressed Representation of Data to Uncover Patterns and Subgroups of Interest*

## The Promise of Machine Learning

Machine learning is a class of algorithms that adjust and learn from data that then take or suggest actions. This field helps companies:

1. Effectively segment existing data into meaningful groupings.
2. Identify the key attributes and features that drive segmentation.
3. Find patterns and anomalies in a data set.
4. Precisely classify new data points as they arrive.

There are two classes of machine learning techniques. Unsupervised learning helps discover the hidden structure in data; it utilizes only the content of input data and has no knowledge of the expected output.

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. Innovations in machine learning techniques promise to help drive new revenue streams, forge stronger customer relationships, predict risk, and prevent fraud. However, analyzing complex data using these methods is constrained by certain intrinsic issues as well as a dependency on scarce machine learning expertise.

## Machine Learning – Mind the Gap

### Unsupervised Learning – An Overview

Unsupervised learning algorithms, which have no knowledge of expected results, fall into two categories:

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

2. Dimensionality Reduction: These algorithms are especially useful for reducing 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 divide or segment a dataset into smaller datasets. Different clustering algorithms rely on different techniques to cluster data.

Take the example of a hierarchical clustering algorithm such as single-linkage clustering. With single linkage clustering, each data point starts out as its own cluster. The single-point clusters are combined 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 have been fused into a cluster. The resulting cluster hierarchy can be visualized as a dendrogram (tree diagram) that shows which clusters were fused together to produce new clusters. Knowing the sequence and distance at which cluster fusion took place can help determine the optimal scale for clustering.

In general, there are three key issues with 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 some informed criteria (such as a "Bayesian information score") to make an educated guess at the number of clusters, this is typically 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 such data sets, thereby obscuring the true underlying structure of the data. Moreover, 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 the shape of the clusters. As an example, the k-means family of algorithms implicitly assumes clusters to be 'spherical'. Similarly, model-based algorithms explicitly assume a model of the shape of the cluster. Imagine you find clusters using either of these methods—how might you discover the fact that the clusters are not faithful to the underlying distribution?

## Unsupervised Learning – Dimensionality Reduction

Dimensionality Reduction methods make it easier to visualize data sets that have a large number of data columns. For example, consider credit card transactions that have thousands of attributes that are each represented as a data column. Visualizing these transactions can be extremely difficult given that we mere mortals cannot see more than three dimensions at once.

A good example of a Dimensionality Reduction algorithm is Principal Component Analysis (PCA); other methods include Multi-dimensional Scaling, Isomap, t-Distributed Stochastic Neighbor Embedding, UMAP embedding, and Google's PageRank algorithm.

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

However, there are two issues with Dimensionality Reduction methods:

1. Projection Loss—Dimensionality Reduction methods compress a large number of attributes down to a few. As a result, data points that are well separated in the high-dimensional space might appear as neighbors in the 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 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 of supervised learning algorithms:

1. Regression algorithms
2. Classification algorithms

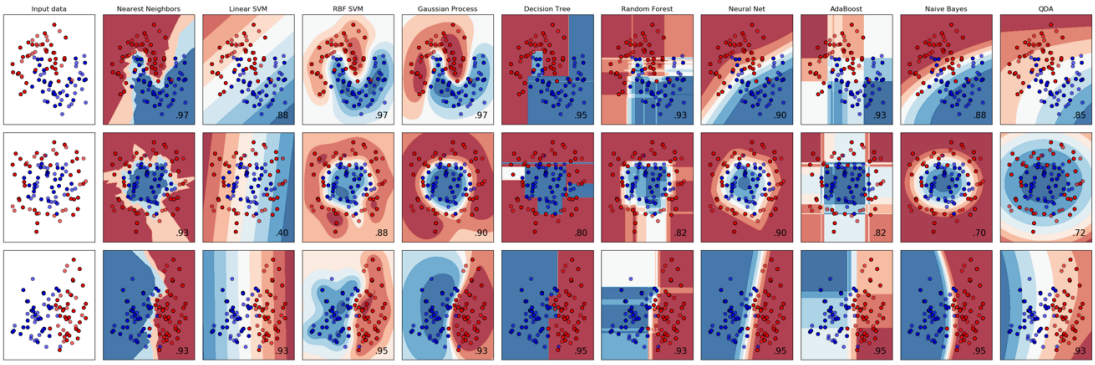
Regressors predict real-valued variables (such as profit margins or stock prices). Classifiers predict discrete variables (such as fraud or customer churn). You may have seen similar examples with Linear and Logistic Regression, Support Vector Machine, and Artificial Neural Networks.

There are two phases in supervised learning:

1. Training—In this phase, the algorithm analyzes a training data set to produce parameters for a function that will infer results when presented with new input data. This phase needs historical data—the empirical evidence known to be correct ("ground truth"), against which to verify the accuracy of the function. By definition, the training set contains associated ground truth information that is used to modify or tweak the function as needed to improve its ability to accurately infer data values.
2. Prediction—In this phase, the function that was produced in the training phase is used to predict the values for new input data points.

Supervised learning algorithms have certain inherent issues that need to be taken into consideration:

1. Assumptions: Choosing an algorithm means making assumptions about the shape of the underlying data. For example, linear regression assumes that 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. The chart below shows decision boundaries of standard machine learning classifiers. Notice that each algorithm has a distinctive shape to its decision boundary. Without first knowing the shape of the underlying data, the chosen algorithm might produce poor results. These approaches rely heavily on a machine learning expert knowing which algorithm to choose. [image: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html]




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. It is unlikely that there is a single shape that fits the entire dataset.
3. **Generalization**: A model may perform well with test data but produce inaccurate results with new data. This is known as a generalization error and it occurs because the model has more parameters than are actually required. Like a cool-looking but poorly fit pair of shoes, the only way to tell they won't work is when you slide your foot in.

### Supervised Learning - What is Missing

Despite its promise, supervised learning has limitations:

1. Successful implementations require experts in machine learning, an increasingly scarce resource to find.
2. It is easy to miss important insights in data by choosing the wrong algorithm or by not trying enough algorithms or hyperparameters.
3. Each class of machine learning algorithms has its own set of intrinsic issues that need to be taken into account.
4. Most supervised machine learning algorithms require large quantities of high-quality labeled data that can be difficult, expensive, or impossible to attain.

The next section details how TDA enhances standard machine learning methods.

## How TDA Improves the Application of Machine Learning Algorithms

All machine learning methods produce functions or maps. For example:

1. Clustering maps each input data point to a cluster.
2. Dimensionality Reduction maps each input data point to a lower dimensional data point.
3. Supervised Learning algorithms map each input data point to a predicted value.

The AyasdiAI platform increases the effectiveness of machine learning by using all of these maps or functions to process input and produce a superior quality output. In addition, not-yet-invented machine learning algorithms can easily be added to the platform infrastructure to further enhance its ability to streamline analysis of large or complex data sets.

Here's how TDA interacts with each of the previously mentioned machine learning approaches.

### Unsupervised Learning - Clustering

Most clustering algorithms rely on global optimization techniques (including HDBSCAN), which are susceptible to noise since they consider all of the data during the optimization. By comparison, 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. The key benefits of this approach include the following:

1. Elimination of the projection loss issue typical of Dimensionality Reduction methods wherein data points that were well-separated 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, data points that were well separated in the original space will typically still be well separated in the output. This enables the easy identification of distinct segments and sub-segments within data that might have been missed using other Dimensionality Reduction methods.
2. Automatic synthesis of 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 Dimensionality Reduction method.

## Supervised Learning - Regression and Classification

TDA augments supervised learning algorithms in the following ways:

1. It eliminates systematic errors. Most supervised learning algorithms are based on global optimization. As such they assume a shape for the underlying data and then try to discover the parameters that best approximate all the data. The result can be mistakes in some data regions. TDA uses the output of these supervised algorithms as an input to discover areas of the underlying data where errors are being made systematically.
2. It's optimized for local data sets. TDA effectively constructs a collection or ensemble of models rather than making global assumptions regarding all the underlying data. Each model is responsible for a different segment of the data. This eliminates the need to create a single model that works well on all of the data, which is a difficult if not impossible mission. This "collection of models" method generates more accurate results and can incorporate any supervised algorithm.

Additionally, TDA utilized on the feature space (that is the transpose of the original data) can be used to substantially improve the convergence capabilities of advanced machine learning techniques such as neural networks.

TDA reduces the possibility of missing critical insights by making machine learning experts less dependent on choosing the right algorithms. Current machine learning techniques are still considered input to find patterns and insights in local data; AyasdiAI's approach enhances any algorithm with which it is paired.

# Summary

While organizations have successfully tackled the challenge of storing and querying vast amounts of data, they continue to lack the necessary tools and techniques for extracting useful information from highly complex data sets. Topology and TDA are well suited for analyzing complex data with potentially millions of attributes. The AyasdiAI 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 software creates highly interactive visual networks that make possible the rapid exploration and understanding of critical patterns and relationships in data.

## About Symphony AyasdiAI

AyasdiAI, a Symphony AI portfolio company, empowers banks and financial institutions with a complete picture of customer, third party and user behavior to discover crime, risk and competitive opportunity through unparalleled, predictive insight. Using a uniquely powerful combination of artificial intelligence and machine learning, AyasdiAI customers dramatically reduce the time to achieve genuine transparency, with full explainability. AyasdiSensa™ leverages unique combinations of topological data analysis, time series and leading analytical innovations to give organizations absolute fidelity for competitive discovery, risk detection and efficiency optimization. Learn more at www.ayasdi.com, LinkedIn, or Twitter.

## About Symphony 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 1,600 skilled technologist 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.

Symphony
AYASDI

555 Twin Dolphin Dr, Suite 370
Redwood City, CA 94065 USA
+1 650.704.3395
sales@ayasdi.com
ayasdi.com | @ayasdi