Symphony AYASDI

TDA Networks for Next Generation Detection

WHITEPAPER

Creating Topological Networks with TDA by using Symphony AyasdiAI's Machine Intelligence Platform

TDA as a discipline identifies data points that are related to each other. It then pieces these regions of data together to build a global, compressed summary of the data in the form of a network. This network can be executed programmatically, or it can be visualized for further investigation.

In the case of a visual network, the AyasdiAI Machine Intelligence Platform 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 visual network consists of nodes wherein each node represents data points having similar function values that have been clustered together based on a measure of that similarity.

Consider two simple examples to illustrate how the AyasdiAI Machine Intelligence Platform creates networks. The first example steps through the general methodology and the second example demonstrates how the AyasdiAI Platform enhances machine learning.

For the first example, consider a data set that is represented by a circle in the xy-plane as depicted in Figure 1. Apply a function f that maps each point in the data set to its y-coordinate value, as shown on the right side of Figure 1.

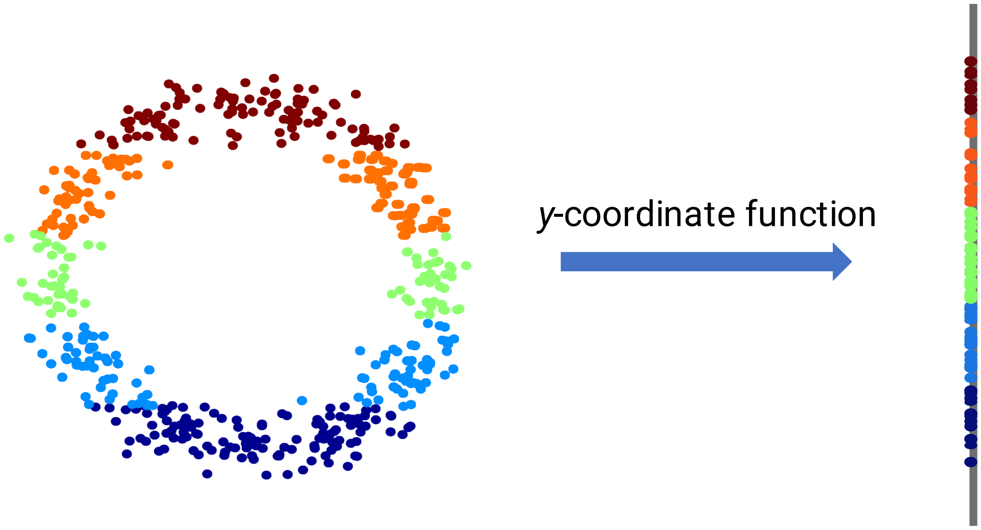

Figure 1: Using a Function to Map Data Points in the Shape of a Circle to their y-Coordinate Values

The AyasdiAI Machine Intelligence Platform then subdivides the image of the function into overlapping sets of nearby values. In this example, the points are divided into four overlapping groups that have similar y-coordinate values as illustrated in Figure 2.

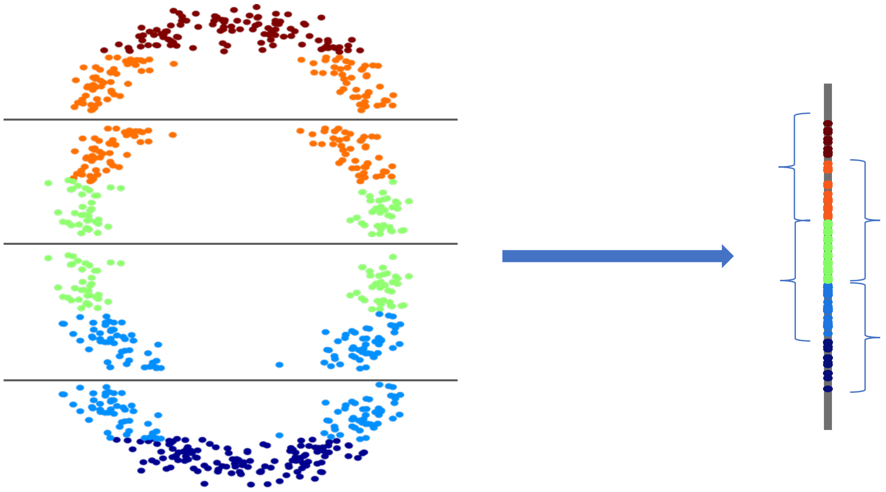

Figure 2: Dividing Data Points into Overlapping Sets with Similar y-Coordinate Values

Next, the AyasdiAI Platform clusters each group of data points independently using a measure of similarity. In this example, similarity is defined using the standard Euclidean (straight line) distance. Each cluster is represented as a node. A node represents a set of 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 within. Notice that as shown in Figure 3 the top node represents both red and orange data points while the second set of data points from the top in the original circular pattern contains two distinct regions of data points that produce two separate nodes in the topological image at right in Figure 3.

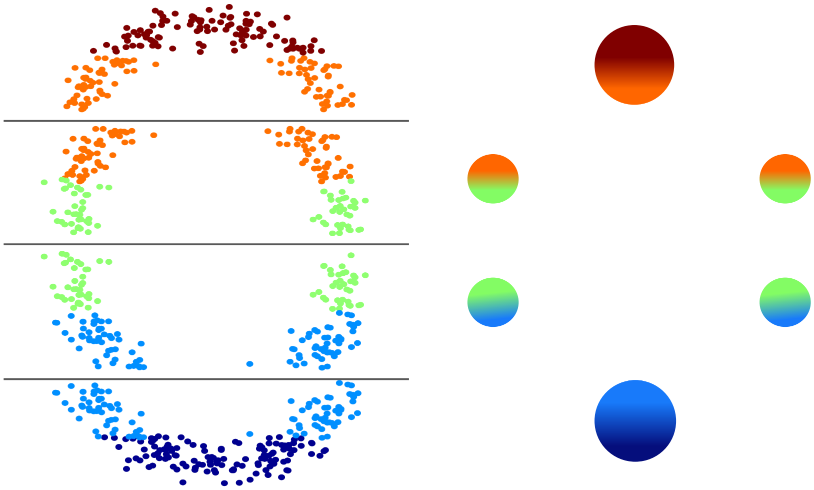

Figure 3: Nodes Represent Clusters of Data Points with Similar Function Values and Measures of Similarity

Nodes with data points in common are connected by edges in the Ayasdi-generated network. Since the data set was divided into overlapping sets, a data point can be represented in multiple nodes. In this example, as can be seen in Figure 4, the orange data points on the left are represented in both the top red node as well as in the orange node on the left. These nodes are connected by an edge because they contain data points in common.

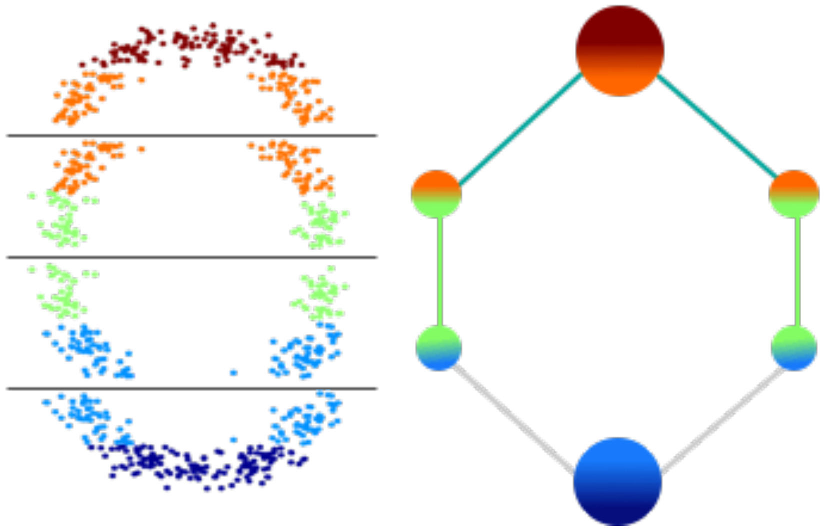

Figure 4: Nodes with Data Points in Common are Connected by Edges to Form a Network

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 than the original data, yet it captures the essential behavior of the data.

Let's turn now to a second example, in which a data set is sampled in the two-dimensional Euclidean plane from four Gaussian distributions. In Figure 5, we color the data points by the values of the density estimator function. The AyasdiAI Machine Intelligence Platform then divides the data set into overlapping groups with similar function values (in this case, density estimations). Each subset of the data is clustered to create nodes that represent data points with similar function values.

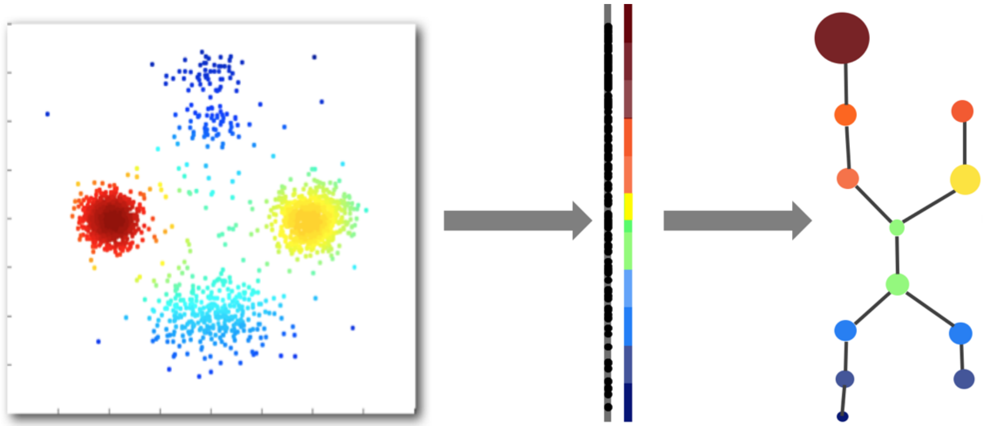

Figure 5: TDA Enhances Machine Learning by Capturing the Overall Structure and Fine-Grained Behavior in a Data Set

The resulting network depicted in Figure 5 captures both the overall structure of the data as well as its fine-grained behavior. The four flares in the network correspond to the four regions of varying densities. The flares in the network connect to each other because these contain common data points having varying degrees of density.

Standard machine learning techniques would have identified the four regions, but they would have lost the continuous transitions between them. The AyasdiAI Machine Intelligence Platform captures both the differences and the similarities in the data.

Complex data holds useful information that could go undetected when using standard machine learning and statistical techniques. By contrast the AyasdiAI Machine Intelligence Platform begins by understanding data at a small scale. It then stitches together these pieces of information to create a topological summary or compressed representation of the entire data set. This is another instance of the platform's ability to create networks that display subtle insights in the data while also showing the global behavior of the data.

The AyasdiAI Machine Learning Platform can incorporate virtually any machine learning, statistical, or geometric technique (or combination thereof) and apply these techniques as a function *f* to display as a visual network of compressed facsimile of a data set. Principal component analysis, autoencoders, random forests, and density estimators are just some examples of functions that the AyasdiAI Platform can use to derive insights from large, complex data sets.

Exploring and Using the AyasdiAI Machine Intelligence Platform to Understand your Data

As discussed above, the AyasdiAI Machine Intelligence Platform uses TDA to create a visual representation of data in the form of a topological network. A network is comprised of the following:

- Nodes that represent collections of similar data points.
- Edges that connect nodes that share data points in common.

TDA helps automatically discover these networks that reveal the underlying structure of a data set. The networks produced by TDA are simple, yet extremely powerful representations of the data and the information encoded in them feed discrete business applications – including fraud detection, money laundering, clinical variation, credit risk or program performance.

How the AyasdiAI Machine Intelligence Platform Uses TDA to Make Complex Data More Intelligible

The AyasdiAI Machine Intelligence Platform has a number of features that support 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

As stated above, data segmentation involves grouping data points that are more similar to each other by some metric than to the remainder of the data. The most common segmentation approaches involve either a data scientist manually generating and testing hypotheses or the use of clustering algorithms.

Manual testing of hypotheses can be a huge undertaking even when dealing with small data sets. This approach typically consists of a domain expert choosing a logical attribute of the data in order to create segments. This approach may have limited utility because key information is not taken into account. Consider, for example, segmenting customers by the amount of money spent; this might seem like a good idea, 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 including: 1. A need to know the number of clusters in advance of applying the algorithm; 2. The unsuitability for tackling continuous data sets; and 3. The assumption of spherical shape or uniform densities.

Figure 6: An intelligent segmentation process delivers far more granular and uniform groups, resulting in higher thresholds and fewer false positives. In addition, these granular groups catch false negatives.

As a more detailed example of clustering, consider a financial institution that segments its clients by their 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-buying behavior, yet it can miss subtle trends hidden in the data that are tied, for example, to specific regional events. Such an event might result in a particular group of clients trading in a specific class of products that diverges from the general trend and is never seen because of the segmentation approach applied. The invisibility of such subtle trends is more likely if the number of clients exhibiting a particular behavior is small compared to the total number of clients in the data set. Complex human behavior such as financial markets typically has numerous instances of overlapping trends both large and small; teasing out the “small” is challenging with traditional methodologies.

The AyasdiAI Machine Intelligence Platform would, on the other hand, 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 would be captured in the AyasdiAI Machine Intelligence Platform output as a flare in the visual network and its presence encoded in an application designed to alert the bank’s sales force of this unique subpopulation. Moreover, the application would direct the sales force to prioritize and target these highly responsive clients with tailored recommendations ahead of those that are more similar to the majority of the clients in the data set.

Anomaly Detection

The Model Validation description above relies on the availability of predicted outcome and ground truth information for a data set. There are, however, cases in which this information is not readily available. The AyasdiAI Platform offers an alternative approach, anomaly detection, that does not require the existence of predicted outcome and ground truth information. The workflow within the AyasdiAI Platform for anomaly detection is as follows:

1. Construct a data set of transactions. Ground truth or other information from current models is not required.
2. Segment the data set based on all data columns to generate a topological network.
3. Explore onscreen regions of the network that represent low density points or points far away from the central core of the data set. This “anomalous” data has less in common with data points in network nodes, which is the reason it was not displayed as part of a node.

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. AyasdiAI’s software helps with feature discovery by automatically producing a list of the attributes (data columns) that drive segmentation, ranked in order of statistical significance.

Take the example of using the AyasdiAI Platform to understand the reasons for customer churn. While the ability to predict churn is useful, spotting the root causes for churn is significantly more important as it often brings systemic issues to the surface.
Identifying the attributes of departing customers with the AyasdiAI Platform involves the following steps:

1. Construct a data set with columns – each column is a customer attribute of interest. Optionally create an outcome data column by which data can be segmented. In this example, an output data column tracks whether a customer departed or remained.
2. Segment the data set using all data columns. For the present example, the output data column that tracks customer churn serves as an additional data lens through which data is viewed.
3. Create clusters of data points or node groups that form a visual network by using the outcome data column for tracking churn.
4. Use the “Explain” operation in the AyasdiAI Platform to get a tabular listing of the underlying features or attributes of the node groups that represent departing customers. This listing will be arranged in statistical order of importance.
5. Encode this information into the workflow of the customer care teams responsible for intercepting potential churners.

Recommendation Engines

Recommendation engines are designed to help organizations generate more revenue by precisely targeting customers with sales efforts for products and services purchased in the past by other customers with similar profiles. The AyasdiAI Machine Intelligence Platform serves as an ideal foundation for a recommendation engine. The steps to build a recommendation application with the AyasdiAI Platform is as follows:

1. Create precise sub-segments of a customer base by correlating and analyzing a wide range of client-related data including demographics, buying behavior, market, CRM, and social media information. Figure 6 presents an example of using the AyasdiAI Platform to distill this information into a topology diagram that can be subjected to further analysis.

2. Assign all newly arriving customer data points to a specific node or group of nodes (sub-segments) in the topology diagram created in the previous step.
3. Look up the buying behavior of the other customers that are represented in the sub-segments to which the new customers were assigned.
4. Develop tailored recommendations based on what similar customers (those in the same sub-segments identified in the previous step) have bought in the past.
5. Use the information encoded in the network to feed the existing CRM application or to build a client facing application.

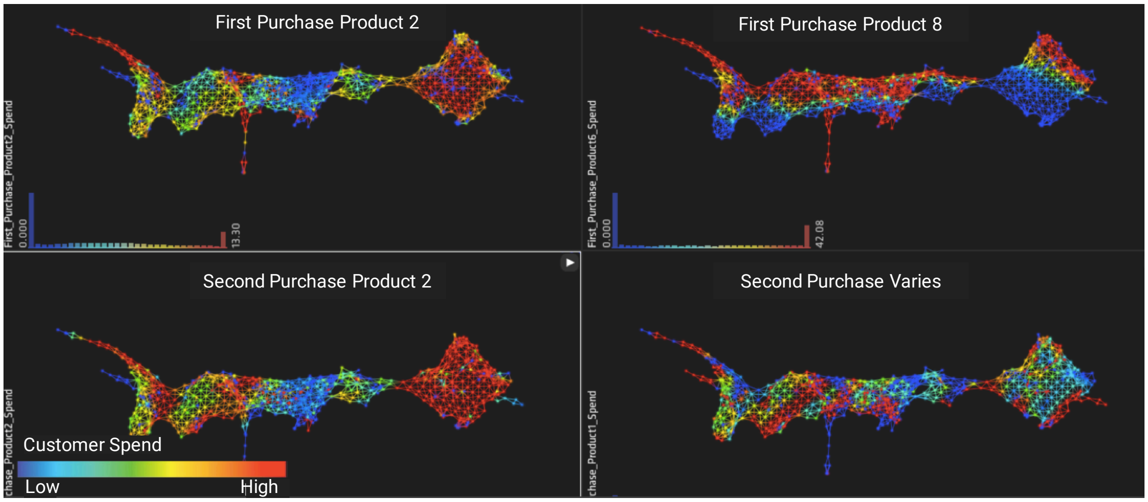

Figure 7: Analyzing Returning Customers by Buying Patterns and Spend

Model Creation

Supervised learning methods are typically employed to create models that can predict future actions or behavior. The AyasdiAI Machine Intelligence Platform supports the application of multiple different algorithms to a data set when creating a topological model. The result is a group or collection of models, referred to as piecewise models or an ensemble of models, that together 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.

A sample workflow for using the AyasdiAI Platform to create a topological model contains the following steps:

1. Construct a data set with columns (attributes) of interest as well as an outcome data column.
2. Segment the data set without using the outcome data column.
3. Create node groups within the topological network.
4. Create a simple, distinct model for each node group using standard supervised learning methods like linear regression.
5. Use the model associated with each node group to accurately predict the placement of newly arriving data points within this node group.

Model Validation

Most organizations rely on a plethora of automated models to help with fraud detection, compliance, regulatory risk management, network security, and client relationship management. These models range from simple rule-based systems to those that are the result of supervised learning algorithms. 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 as a result of incorrect assumptions about the shape of the underlying data. The AyasdiAI Platform uses TDA to uncover these errors in models.

Consider the process of validating models used to detect fraud in credit card transactions. Identifying issues in these models using the AyasdiAI Platform involves the following steps:

1. Construct a data set where each data point is a transaction. Create and populate two additional data columns within this data set as follows:
    a. The predicted outcome from the model for each transaction.
    b. The actual ground truth – was each transaction fraudulent or not?
2. Segment the transaction data using all columns except those that track predicted outcomes and ground truth. The result will be a topological network.
3. Use the color scheme capability of the AyasdiAI Platform to create two separate continuums of hues whose color values are controlled by the predicted outcome and ground truth data columns respectively. Apply each of these color schemes to the topological network created earlier to highlight nodes and/or data points that are outliers, i.e., that do not conform to the predicted outcome or the ground truth values.
4. Focus on the subgroups of transactions in the topological network where the model made mistakes.
5. Use the “Explain” operation in the AyasdiAI Platform to retrieve a list of the data columns (features or attributes) associated with these subgroups of transactions that were not handled properly by the model. This approach can identify the attributes of questionable transactions that were not discovered by the existing fraud detection model.

Summary

AyasdiAI’s use of TDA augments current machine-learning techniques by ameliorating some of their intrinsic issues and reducing the dependency on increasingly scarce human expertise. Innovative companies are using TDA and the AyasdiAI Machine Intelligence 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

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.

Symphony AYASDI

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