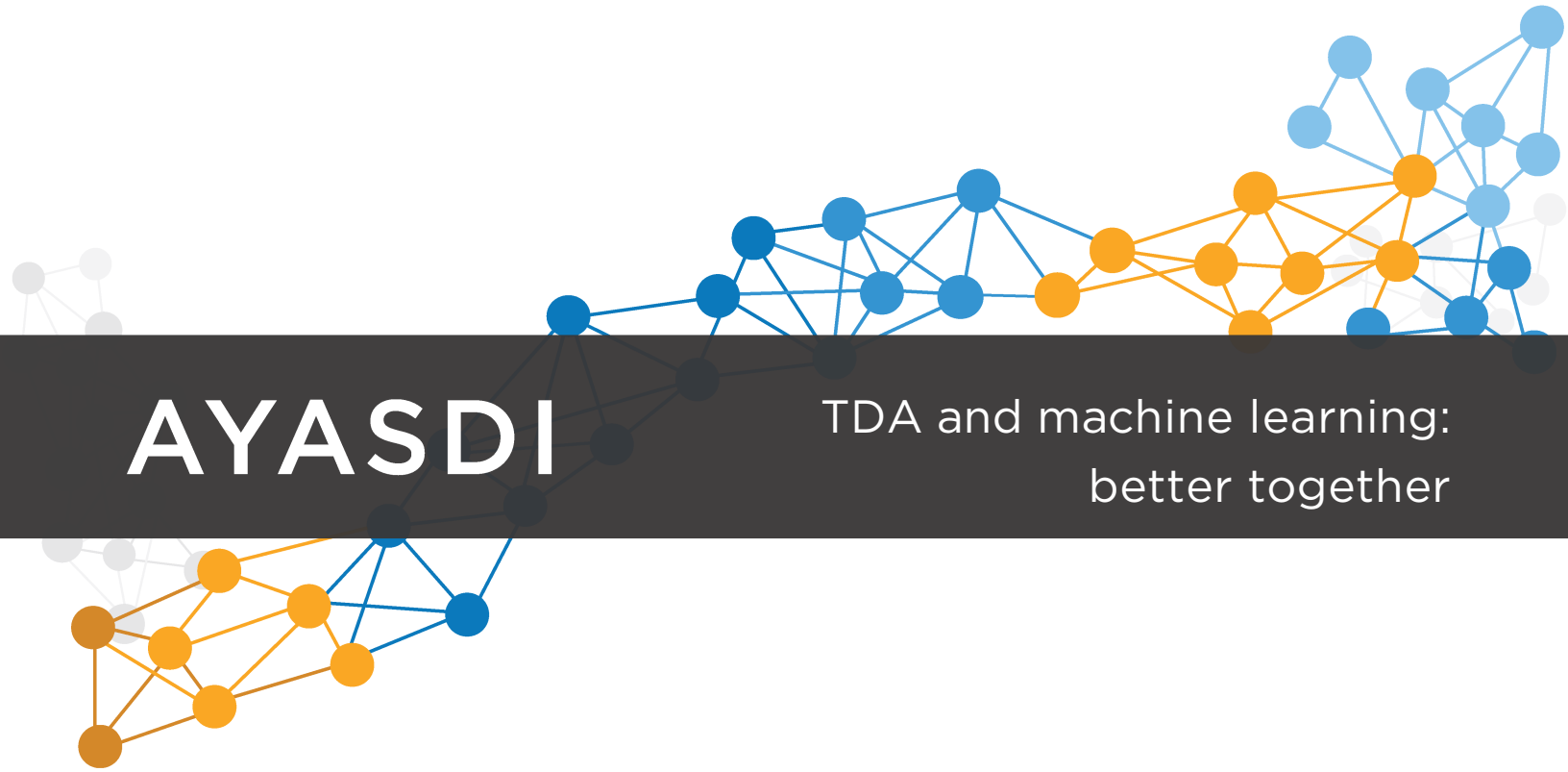


AYASDI

TDA and machine learning: better together

The big data analytics opportunity and dilemma

As artificial intelligence operating on big data becomes a reality, organizations will radically transform. A new generation of computing systems will augment your human capital with machine intelligence, driving speed, efficiency and precision into your business that has never before been possible. You will leverage your vast data assets to deploy whole new classes of intelligent, predictive applications that allow you to transform your relationships with your clients, capitalize on rapidly evolving market conditions, mitigate risk, block fraud and discover new cures for disease.

Machine intelligence is fundamentally about tapping into the latent insights in the massive volumes of unique data that you already have to better inform and empower your enterprise. It automates and accelerates the process of discovering critical patterns and insights from your big and complex data. It leverages commodity high-performance computing to generate mathematical models that you can deploy to drive a whole new generation of intelligent processes and applications across your enterprise.

With machine intelligence, small teams of analysts and business people can solve major analytical and operational challenges in days. Because machine intelligence is literally thousands of times more efficient than manual analytic processes, your organization can get much more done, with far fewer resources. Think of businesses like Google, Facebook or Amazon – like them you can scale with compute operating on big data instead of by adding people.

The ultimate goal is to become an automated, data and compute-driven enterprise – machine intelligence is helping organizations like yours become radically more efficient by leveraging raw computing power and predictive algorithms tapping your own unique data to drive your business.

The collection and analysis of data from transactions, sensors, and biometrics continues to grow at a prodigious rate, taxing the analytic capabilities of even the most sophisticated organizations. The quantity of possible insights in a given dataset is an exponential function of the number of data points. On top of this, aggregate data growth is an exponential function with time. Unfortunately, the world cannot train enough data scientists to meet this runaway, double-exponential demand curve for analytics.

This dynamic is driving computer scientists and mathematicians alike to examine new approaches to improve both the quality and the speed of their analytics platforms. Today's hypothesis-driven analytics and manual machine learning algorithms and statistical tools will not suffice. High-performance computers and algorithms can examine big and complex data far faster and seek insights more comprehensively than any human is capable of. As such, we need to find exponential improvements in analysis and modeling techniques to meet the growing demand.

> "Citi's unmatched multinational business footprint creates a complex data analytics landscape. Ayasdi's big data technology simplifies and accelerates the analysis of thousands of discrete variables and delivers insights that enable Citi to tailor services to specific client needs, operate more efficiently, and mitigate risk."
>
> Deborah Hopkins,
> Chief Innovation Officer
>
> citi®

Introducing topology and Topological Data Analysis

Topology is a mathematical discipline that studies shape. Topological Data Analysis (TDA) refers to the adaptation of this discipline to analyzing big and highly complex data. It draws on the insight that all data has an underlying shape and that shape has meaning.

Ayasdi's approach is to deliver an enterprise software platform that layers TDA on top of a broad range of machine learning, statistical, and geometric algorithms. The analysis creates a summary or compressed representation of all of the data points to help rapidly uncover critical patterns and relationships in the data. By identifying the geometric relationships that exist between data points, Ayasdi offers an extremely intuitive way of interrogating your data to understand the underlying properties that characterize the segments and sub-segments that lie within your data set.

Once you have completed your analysis for a specific business problem, your developers can use the Ayasdi platform create intelligent applications that let your business people develop, validate and maintain new models on their own, and you can use the Ayasdi's API to deploy operational applications that deploy your models in production.

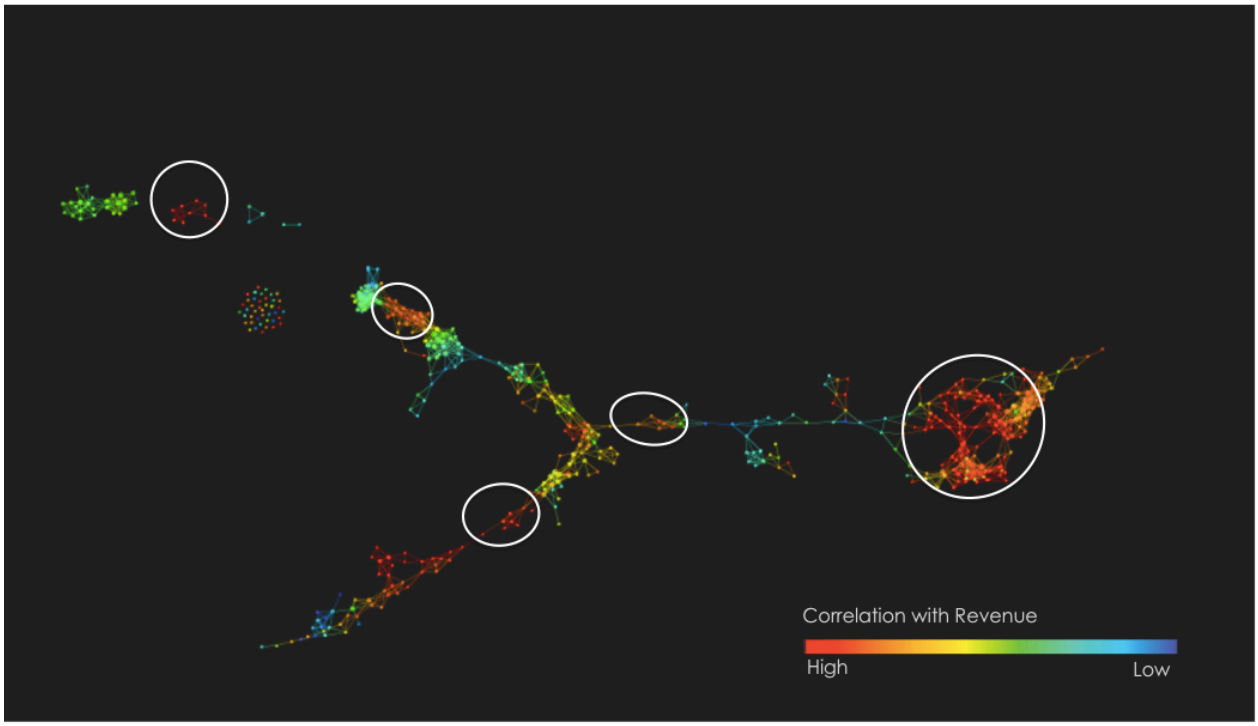


Figure 1: TDA creates a compressed representation of data to uncover patterns and subgroups of interest

The promise of machine learning

Machine learning is a class of algorithms that adjust and learn from data to take or suggest actions in the future. It promises to help companies achieve the following goals:

- Effectively segment existing data
- Identify the key attributes and features that drive segmentation
- Find patterns and anomalies in the data
- Precisely classify new data points as they arrive

There are two broad classes of machine learning techniques – unsupervised and supervised. Unsupervised learning helps with discovering the hidden structure in data. Supervised learning helps with the construction of predictive models. Innovations in these techniques promise to help drive new revenue streams, forge stronger customer relationships, predict risk, and prevent fraud. However, analyzing big and complex data using these methods alone is constrained by certain intrinsic issues as well as a dependency on scarce data science expertise.

Machine learning – what you need to watch out for

We at Ayasdi are big fans of machine learning. The Ayasdi platform includes more than 30 different machine learning, statistical and geometric algorithms, and you can add your own on top of what we provide.

But while machine learning is revolutionary, it is important that you understand and compensate for its inherent limitations as you deploy advanced analytics.

There are two classes of unsupervised learning algorithms:

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 (data columns) required for describing each data point while retaining the inherent structure of the data.

UNSUPERVISED LEARNING - CLUSTERING

You can use clustering methods to segment a dataset into smaller datasets. Different clustering algorithms draw on different techniques to cluster data. Take the example of a hierarchical clustering algorithm such as single-linkage clustering. In the beginning, each data point is its own cluster. The clusters are combined into larger clusters by sequentially fusing pairs that are the most similar to each other by some measure. The process continues until all the data points are fused into one large cluster. The clustering hierarchy can be visualized as a dendrogram (a 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 two key issues with clustering algorithms:

1. The number of clusters – many clustering algorithms require that the number of clusters to be returned must be determined in advance of applying the algorithm to the data. 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, typically this is an arbitrary choice that will greatly impact your conclusions.

2. Continuous data sets - 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 real structure and leading to anomalous results.

UNSUPERVISED LEARNING – DIMENSIONALITY REDUCTION

Dimensionality reduction methods make it easier to interrogate very wide data sets (data sets that have a large number of columns, also known as highly dimensional data.) For example, consider credit card transactions that have thousands of attributes that are represented as data columns. Visualizing these transactions can be extremely difficult given that humans cannot see more than three dimensions at a time. There are many more examples – your customer data is likely to be highly dimensional, and your own personal medical data (think about your genome) is massively dimensional.

Principal Component Analysis (PCA) is a good example of a dimensionality reduction algorithm. Other examples of dimensionality reduction methods include multidimensional scaling, isomap, t-distributed stochastic neighbor embedding and even Google's PageRank algorithm.

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

However, there are two issues with dimensionality reduction methods:

1. Projection loss – by definition, dimensionality reduction methods compress a large number of attributes down to a few. As a result, data points might appear to be in clusters that they are not actually a part of in the full data set. Clusters that should be distinct may overlap in reduced dimensional space. Projection loss increases the chances of missing out on subtle insights inherent in your data.

2. Inconsistent results - Different dimensionality reduction algorithms produce different projections because they encode different assumptions. None of the results are wrong; they are just different as they accentuate different aspects of your data. Depending on a single algorithm can 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

Regression algorithms predict real-valued variables such as profit margins or stock prices. Classifiers predict discrete variables, which is useful for applications such as fraud or customer churn. Examples of regression algorithms include linear and logistic regression, support vector machine, and artificial neural networks.

There are at least two phases in supervised learning:

1. Training - In this phase, the algorithm analyzes a training data set to produce parameters for a function that is assumed to represent the data. This phase needs historical data for which the results are known (the "ground truth").

2. Prediction - In this phase, the function that was produced in the training phase is used to predict the values for new data points.

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

1. Assumptions - The choice of algorithm entails an assumption about the structure of your underlying data. For example, linear regression assumes that your data is planar (possibly higher dimensional) and tries to find the best plane that fits your data. If the actual structure of your underlying data is not planar, then linear regression analysis will produce incorrect results. With supervised learning, there is typically a heavy reliance on data scientists knowing which algorithm to choose. And if you choose the wrong algorithm, you can end up with the wrong answer.

2. Global optimization - All supervised algorithms try to find parameters for a function that best approximates all of your data. However, data is rarely homogeneous. It is unlikely that there is a single structure that fits your entire dataset. A better approach would be to break your data into subsets, and use an appropriate algorithm that best fits each subset. But few analysts do this, because using typical machine learning tools is too complex, slow and difficult to attempt.

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 was built with more variables than were actually required. This issue is also known as overfitting. A better approach would be to have the system help the analyst avoid overfitting errors, but this is impossible with manual approaches to machine learning.

So now we understand the main limitations of machine learning

1. Successful implementations of big data analytics using machine learning require data scientists, an increasingly scarce resource to find, and the more analyses you want to undertake, the more data scientists you need. The problem is that scaling with scarce people just doesn't work.

2. It is easy to miss important insights your data by choosing the wrong algorithm, or even worse to come up with invalid results.

3. Because the selection and implementation of machine learning algorithms is manual, it would be prohibitively expensive and time consuming to exhaustively evaluate all alternatives. So this is never done, leading to non-optimal results.

4. Because it is so costly and time consuming for each machine learning analysis, data scientists are unable to create collections of localized models that more accurately fit your data, again leading to non-optimal results.

5. And finally, because existing machine learning tools are fundamentally designed around manual interaction with a data scientist, they are not taking advantage of the opportunity for automating repetitive tasks or scaling using compute instead of adding more people.

The next section details how TDA enhances standard machine learning methods and improves both the quality of life and productivity of your data scientists.

How TDA improves machine learning algorithms

TDA makes machine learning algorithms dramatically more effective, and your analytic processes vastly faster and more scalable. Plus, it supports going beyond analysis and deploying intelligent applications and operational systems based on your data.

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

1. Clustering maps an input data point to a cluster.
2. Dimensionality reduction maps an input data point to a lower dimensional data point.
3. Supervised learning algorithms map an input data point to a predicted value.

TDA uses the maps / functions from machine learning, statistical and geometric algorithms as its input to produce superior results.

UNSUPERVISED LEARNING - CLUSTERING

TDA uses clustering as an integral step in building a network representation of your data. As opposed to trying to find disjoint groups, TDA applies clustering to small portions of your data. It then combines these "partial clusters" into a network representation that gives an overview of the similarity between your data points. This makes TDA appropriate for constructing a connected representation of both discrete and continuous data sets as well as data with heterogeneous densities.

UNSUPERVISED LEARNING - DIMENSIONALITY REDUCTION

TDA supports the automatic execution and synthesis of many dimensionality reduction algorithms. The key benefits include the following:

1. TDA eliminates the projection loss issue typical of dimensionality reduction methods wherein data points that are well-separated in higher dimensions end up overlapping in lower dimensional projections. TDA achieves this by clustering your data in its original high dimensional space. As a result, data points that were well separated in the original space will typically remain well separated in the TDA output. This enables the easy identification of distinct segments and sub-segments within data that would have been missed using standard dimensionality reduction methods.

2. TDA is able to synthesize the results of many 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, and automates the process of applying different dimensionality reduction techniques to subsets of your data.

SUPERVISED LEARNING - REGRESSION AND CLASSIFICATION

TDA augments supervised learning algorithms in the following ways:

1. Eliminates systematic errors - Most supervised learning algorithms are based on global optimization. These algorithms try to assume a structure for the underlying data and then find the parameters that best approximate all of the data, thereby making mistakes in some regions. TDA uses the output of these supervised algorithms as an input to discover areas of the underlying data where such errors are being made systematically.

2. Optimized for local data sets - As opposed to making global assumptions regarding all of your underlying data, TDA can construct a collection or ensemble of models. Each model is responsible for a different segment of your data. This eliminates the need to create a single model that works well on all of your data. A collection of models is usually much more accurate. This approach works for any supervised algorithm.

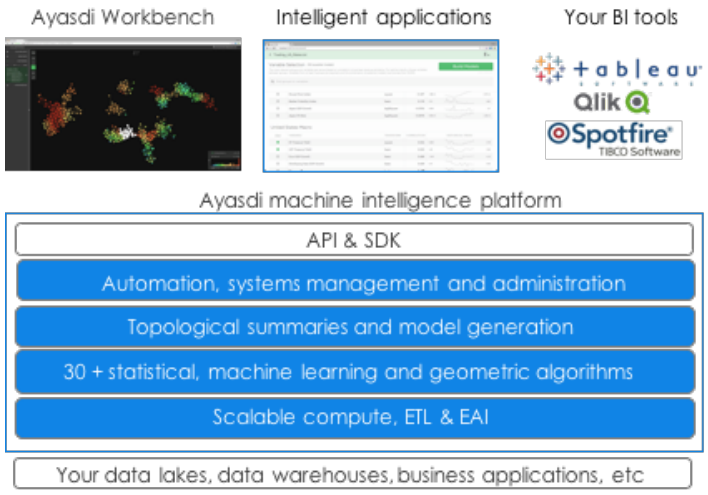
TDA reduces the possibility of missing critical insights by reducing the dependency on data scientists choosing the right algorithms. It uses multiple machine learning techniques as input to find subtle patterns and insights in local data.

In general, TDA enhances any machine learning statistical or geometric algorithm that it is paired with, and it enables the use of multiple algorithms simultaneously, each of which is optimized for those localized regions within your data to which it is applied.

The Ayasdi machine intelligence platform

Ayasdi is an enterprise scale machine intelligence platform that implements and automates TDA plus a wide range of machine learning, statistical and geometric algorithms to help you gain competitive advantage from your big and complex data, without requiring large teams of data scientists to write queries or code algorithms. It supports large numbers of business analysts, data scientists, end-users, developers and operational systems across your organization, simultaneously creating, validating, using and deploying sophisticated analyses and mathematical models.

Ayasdi rides on top of information already resident in your business applications, data warehouses, data lakes or other big data infrastructure and automatically applies many algorithms to your data, dramatically speeding the discovery and model development process.



Ayasdi layers topological data analysis on top of machine learning, statistical and geometric algorithms to extract critical intelligence from your data that was previously hidden or overlooked by conventional analytical approaches.

Ayasdi Workbench is a graphical modeling environment that creates a compressed visual summary of all of your data so your analysts can rapidly uncover the relationships, clusters, progressions, anomalies, and cycles in your data, and explain the underlying reasons for these patterns.

Ayasdi automatically creates and validates mathematical models based on your data, and allows these models to be deployed into your production systems. Your developers can leverage scalable APIs, web services and robust scripting capabilities to deploy intelligent applications at scale.

The Ayasdi machine intelligence platform can be deployed in public or private cloud infrastructures, and leverages inexpensive, scalable Intel-based computing platforms and Hadoop infrastructure.

Using Ayasdi and TDA to analyze your big and complex data

TDA first identifies data points that are related to each other. It then pieces these regions of your data together to build a global, compressed summary of your data in the form of a network. Ayasdi uses a function (f) on the data and a measure of similarity to generate compressed representations of your data. The resulting network, which you can visualize and manipulate in Ayasdi Workbench, consists of nodes that represent data points with similar function values and that form a cluster based on a measure of similarity.

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

First take a data set that is represented by a circle in the xy-plane. We will then use a function f that maps each point in the data set to its y-coordinate value.

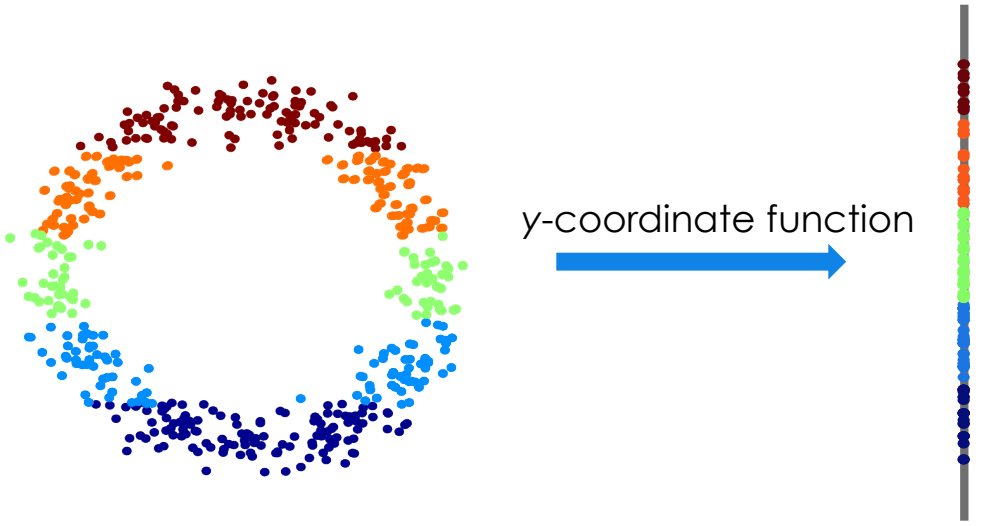


Figure 2: Using a function to map data points in the shape of a circle to their y-coordinate values

Ayasdi 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.

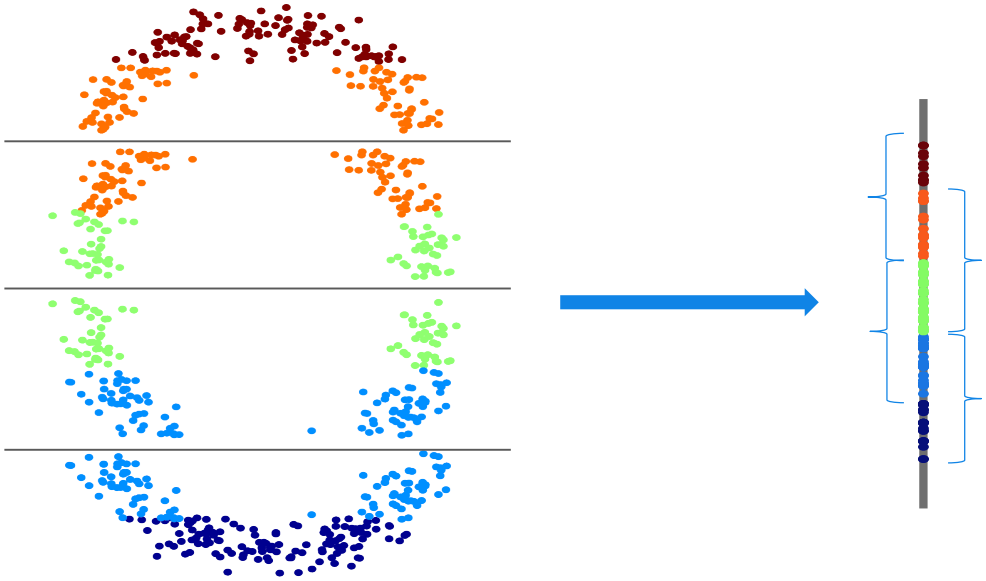


Figure 3: Dividing data points into overlapping sets with similar y-coordinate values

Next, Ayasdi clusters each group of data points independently using a measure of similarity. In this example, similarity is defined using standard Euclidean distance. Each cluster is represented as a node. A node represents a set of data points that are similar with respect to the measure of similarity (Euclidean distance) and the function value

(y-coordinate). The size of the node reflects the number of data points within. Notice that the top node represents both red and orange data points. The second set from the top containing two distinct regions of data points produces two separate nodes.

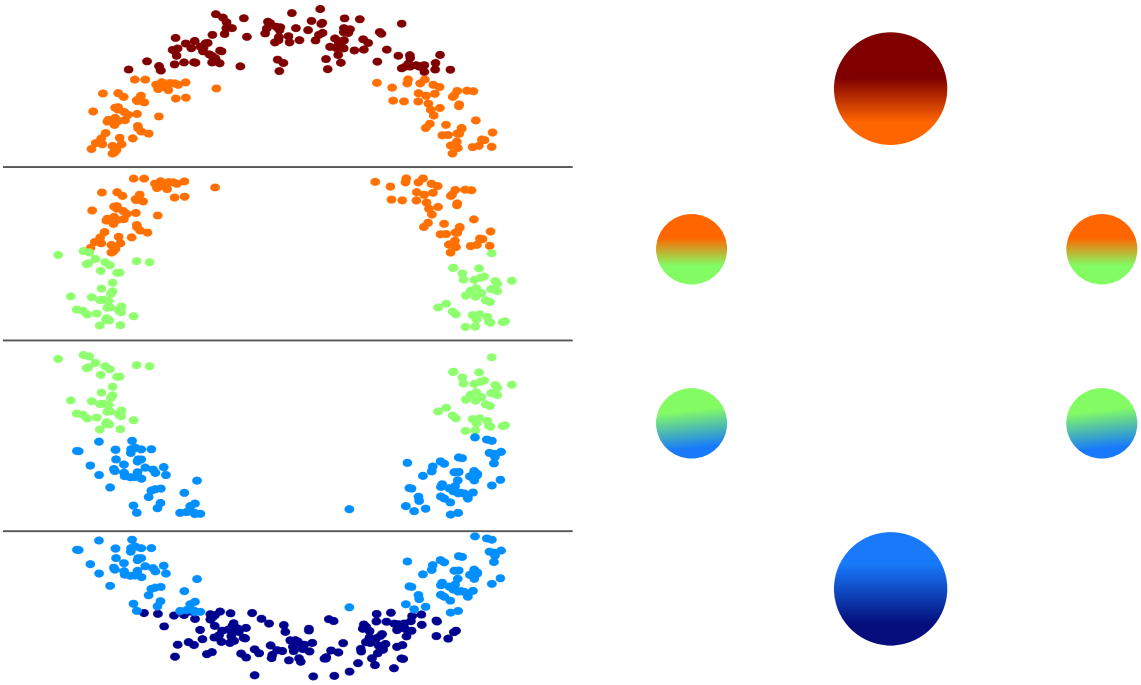


Figure 4: Nodes represent clusters of data points with similar function values and measures of similarity

Nodes with data points in common are connected by edges, which are automatically visualized in Ayasdi Workbench. Since the data set was divided into overlapping sets, a data point can be represented in multiple nodes. In this example, the orange data points on the left are represented in both the top red node as well as the orange node on the left in figure 5 below. These nodes are connected by an edge because they contain data points in common.

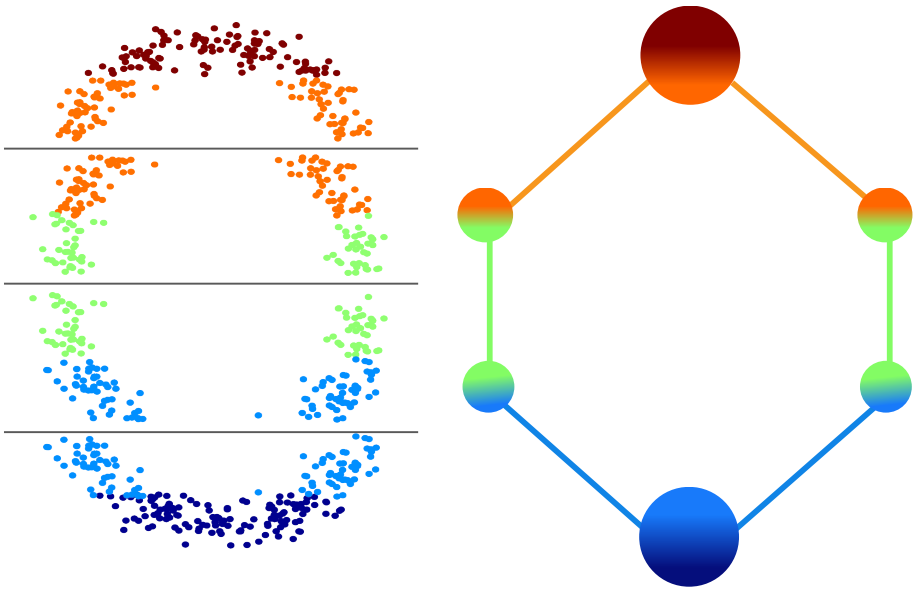


Figure 5: Nodes with data points in common are connected by edges to form a network

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 original data.

For our second example, we will examine a data set that is sampled in the two-dimensional Euclidean plane from four Gaussian distributions. In figure 6, we color the data points by the values of the density estimator function. Ayasdi then divides the data set into overlapping sets 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 and measures of similarity.

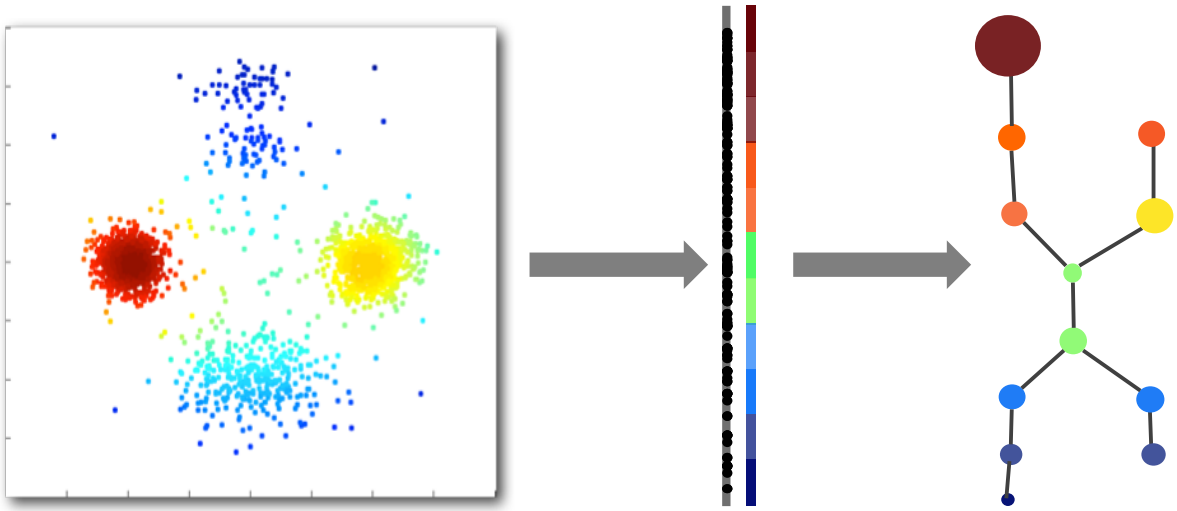


Figure 6: TDA enhances machine learning by capturing the overall structure and fine-grained behavior

In figure 6, you can see how the resulting network 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 are connected to each other because of the data points that are common to these regions with varying degrees of density.

Standard machine learning techniques would have identified the four regions but they would have lost the continuous transitions between them. Ayasdi and TDA capture both the differences and the similarities in the data.

Your big and complex data holds useful information that often goes undetected when using standard machine learning and statistical techniques. The Ayasdi platform begins by understanding data at a small scale. It then automatically stitches together these pieces of information to create a topological summary, which is a compressed representation of the entire data set. The resulting networks are visualized for you to explore in Ayasdi workbench, and make it easy to surface the subtle insights in the data while also representing the global behavior of your data.

Ayasdi can draw on the power of the function f to incorporate virtually any machine learning, statistical, or geometric technique into the creation of compressed representations of your data visualized as networks. Principal component analysis, autoencoders, random forests, and density estimators are some examples of functions f that the Ayasdi platform uses to derive insights from your data. In this way, Ayasdi uses TDA a framework for advanced analytics.

Using Ayasdi Workbench to understand and explore your data

The Ayasdi platform leverages extensive automation and TDA to create a representation of your data in the form of a network. Ayasdi Workbench visualizes this network for you and lets you interact with it. A network comprises of the following:

- Nodes that represent collections of similar data points in your original data set
- Edges that connect nodes that share data points

Ayasdi uses TDA to help automatically discover these networks, which reveal the underlying structure of your data set.

The networks produced by the Ayasdi platform and TDA are simple, yet extremely powerful representations of your data. The following section outlines various techniques for exploring, understanding, and using the insights uncovered from your data using Ayasdi Workbench.

EXPLORING YOUR DATA

1. Visualize - Ayasdi Workbench presents the output of TDA as an interactive visual network that is a compressed representation of your data. Tugging and pulling at the nodes changes the orientation of the network. Changing the appearance of a network on the screen by moving nodes around or changing their colors does not impact the insights it represents. In fact, visualizing a network can help with selecting regions of interest and creating node groups that can be inspected further using the “Explain” and “Export” operations.

2. Color – The color of the nodes and edges of a network can be changed to allow for the quick exploration of your data. When coloring a network by a particular data column, Ayasdi computes the mean value of the specified data column for all the data points within each node individually. It then maps all these values to a color palette and renders them in Ayasdi Workbench.

3. Find - Within Ayasdi Workbench, specific conditions can be applied to data columns that evaluate to either true or false. For example, to find all your customers whose "Net Worth" (the data column) is greater than $500,000, the software creates a color scheme that highlights nodes that contain data points that meet this condition.

4. Contrast - Ayasdi Workbench also lets you explore the differences between two specified color schemes.

UNDERSTANDING AND MAKING USE OF INSIGHTS INTO YOUR DATA

1. Explain - Ayasdi includes an automated "Explain" operation to help you find the data columns or attributes that differentiate your node groups. It automatically runs a wide array of tests (e.g., KS tests, T-test, P-value, hypergeometric enrichment) and returns a list of data columns that are ordered by their statistical power in differentiating the specified node groups.

2. Resolution - Ayasdi Workbench provides a multi-resolution view that supports the discovery of subtle, otherwise hard-to-find signals in your data.

3. Export - Ayasdi Workbench also allows for the export of data points along with the associated list of points that belong to a node for use in downstream operations (e.g., to view using BI tools).

How to use Ayasdi and TDA to make your big and complex data useful

SEGMENTATION

Data segmentation involves grouping data points that are more similar to each other in comparison with the remainder of the data. The most common approaches involve either a data scientist manually generating and testing hypotheses or the use of clustering algorithms.

The manual testing of hypotheses can be a huge undertaking, even when dealing with relatively small data sets. Typically, a domain expert starts by choosing a logical attribute of the data to create segments. However, while segmenting your customers by their spending might seem like a good idea, it ignores the impact of other factors that might be even more important such as demographics or psychographics.

In comparison, standard clustering methods for segmentation produce better results. However, these methods still suffer from the issues described earlier such as the need to know the number of clusters in advance of applying the algorithm and unsuitability for tackling continuous data sets.

For instance, a financial institution can profit from segmenting their clients by their investment behavior under specific market conditions and then precisely targeting them with tailored recommendations, at the right time. Typically, there are macro-trends in client buying behavior. However, there can also be subtle trends hidden in the data that are driven by specific regional events. For instance, such an event might result in a particular group of

clients trading in a specific class of products. In addition, their response is more likely on a continuum, with some clients responding more significantly to the event than others. If the number of clients in this particular region is small compared to the total number of clients in the data, given that they also respond to the same broader market conditions as the rest of the investors, this subtle regional trend will likely be ignored by standard clustering methods.

The Ayasdi platform, on the other hand, would discover that while these regional investors are similar to the majority of the clients in the data, they are much more similar to each other. This subtle signal would be captured in the Ayasdi platform and rendered in Ayasdi Workbench as a flare in the network, thereby informing the bank's sales force of this subpopulation's presence. Moreover, the flare would capture the continuum of responsiveness within this regional group, enabling the sale force to prioritize and target highly responsive clients with tailored recommendations, ahead of those that are more similar to the majority of the clients in the data.

FEATURE DISCOVERY

Understanding the underlying features or attributes of the data that drive segmentation can be invaluable when it comes to pinpointing the factors that impact business outcomes. The Ayasdi platform 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 Ayasdi to understand the reasons for customer churn. While the ability to predict churn is useful, being able to get to the root causes for churn is significantly more important as it often brings systemic issues to the surface.

Identifying the features that result in customer churn with Ayasdi involves the following steps:

1. Construct a data set with data columns (in this case, customer attributes) 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 churned or not.

2. Segment the data set using all the data columns. In this case, 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 network, in this case, using the outcome data column for tracking churn.

4. Use the "Explain" operation to get a tabular listing of the underlying features or attributes of the node groups that represent customers that have churned, in statistical order of importance.

CLASSIFICATION

Recommendation engines are designed to help organizations drive more revenue by precisely targeting customers with products and services that other customers with similar profiles purchased in the past. The Ayasdi platform can serve as an ideal foundation for a recommendation engine. Using Ayasdi to deliver tailored recommendations involves the following steps:

1. Create precise sub-segments of a customer base by correlating and analyzing a wide range of client-related data including demographic, buying behavior, market, CRM, and social media information (Figure 7).

2. Assign all newly arriving customer data points to a specific node or group of nodes (sub-segments).

3. Look up the buying behavior of the other customers that are represented in these sub-segments.

4. Present tailored recommendations based on what similar customers have bought to the sales force or directly to customers via a dashboard or through targeted alerts.

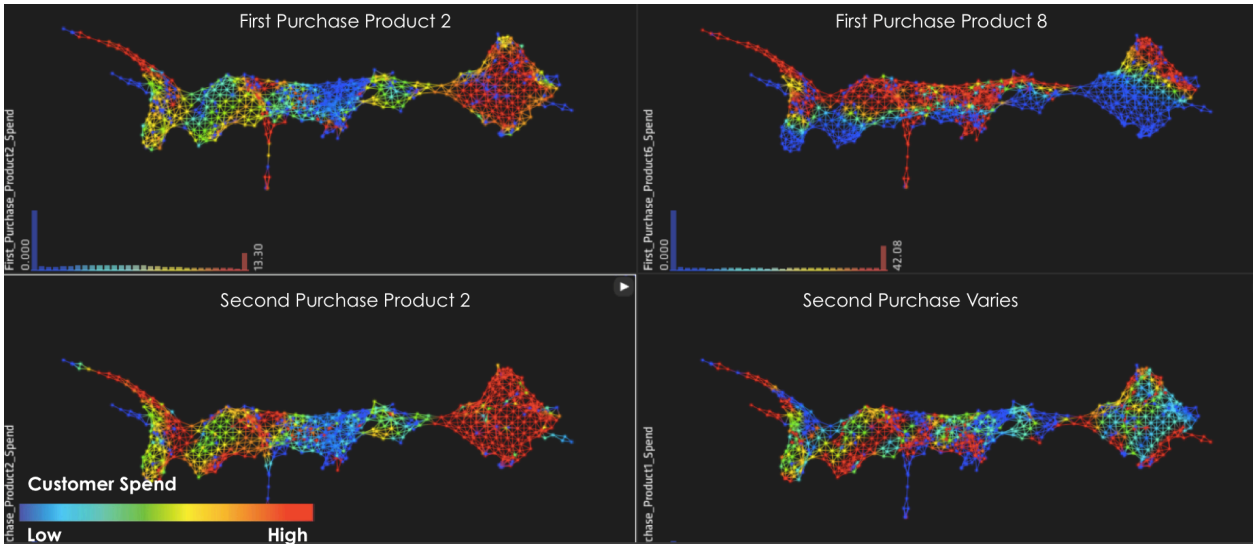


Figure 7: Analyzing returning customers by buying patterns and spend

MODEL CREATION

Supervised learning methods are typically employed to create models that can predict future actions or behavior. Ayasdi leverages TDA to automate the creation of collections of models, referred to as piecewise or ensemble models, that best represent all of your data. These local models tend to be far more accurate than a single global model as they are each optimized for different segments of your data.

An example workflow for creating a model with Ayasdi involves the following steps:

1. Construct a data set with data columns (attributes) of interest as well as an outcome data column.

2. Segment your data set without using the outcome data column and create node groups within the network.

3. Create a simple, distinct model for each node group using supervised learning methods like linear regression that are built into the Ayasdi platform.

4. Use the model associated with each node group to accurately predict the placement of newly arriving data points within these node groups.

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 results of supervised learning algorithms. One of the primary steps involved in validation or auditing exercises is the discovery of systematic errors or biases in the model. Typically, models created by supervised learning algorithms produce systemic errors as a result of incorrect assumptions about the structure of your underlying data. The Ayasdi platform uses TDA to help you uncover these errors in your models.

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

1. Construct a data set where each data point is a transaction. Create two additional data columns:
    a. The predicted outcomes from the model
    b. The actual ground truth - were the transactions fraudulent or not?
2. Segment the transaction data using all but the columns that track predicted outcomes and the ground truth.
3. Color the network by both the model estimation and the ground truth and focus in on the subgroups of transactions in the network where the model made mistakes.
4. Use the "Explain" operation to get a list of the data columns (features) associated with these subgroups to identify fraud that previously went undetected.

ANOMALY DETECTION

The traditional approach to detecting new patterns of fraud can be manually intensive. Manual investigations often result in the creation of new fraud rules that then need to be incorporated into the fraud detection models.

The Ayasdi platform uses TDA to help automatically detect patterns of fraud in data. Detecting anomalies using Ayasdi Workbench involves the following steps:

1. Construct a data set of transactions. Unlike model validation exercises, anomaly detection does not require knowledge of the ground truth or other information from current models

2. Segment the data set based on all data columns

3. Explore regions of the network that represent low density points or points far away from the central core of the data set.

TDA and Ayasdi don't stop with analytics

While the prior section of this document has been focused on what the analyst / data scientist can do in Ayasdi Workbench, the graphical network and data exploration tool that is part of the Ayasdi platform, it's important to remember that machine intelligence does not stop with analytics and insight.

Ayasdi doesn't just help you gain insights; it also automatically creates statistically proven mathematical models based on those insights. If your business is dependent on creating and deploying models, such as risk models in financial services, or population health models in healthcare, once you have performed your initial analysis, your application developers can create intelligent applications that allow your business people to create, validate and manage new models entirely on their own. Global banks have been able to reduce the time it takes to create new risk models from 5,000 hours using machine learning tools to as little as 15 minutes using intelligent applications built on the Ayasdi platform with this approach.

You can also deploy the models you create in operational applications at scale, so your business can start to run based on algorithms that leverage your own data, and can scale using compute instead of by adding more people. Any repetitive process that currently relies on manual human decision making is a candidate for replacing with an intelligent application based on machine intelligence. Anti-money laundering is a classic example of this in financial services, as is clinical variation in healthcare. Assessing and rating prospective clients, and deciding what products to offer to which clients is a cross-industry example of a process that can be automated and improved with machine intelligence.

SUMMARY

There is a tremendous opportunity for you to tap into the massive amounts of client, product, and market-related data at your disposal to uncover previously hidden insights, to create predictive models, and ultimately to automate your business with intelligent applications. It is difficult to understate how transformational this is – in effect your firm will start to behave like Google, Facebook and/or Amazon, where algorithms and raw computing power fed by your unique data will drive your business. Leading organizations worldwide are finding Ayasdi machine intelligence to be as much as 1000x more effective than their current systems and processes for advanced analytics.

You can leverage Ayasdi to support the industrialization of analytics within your firm, and to scale without having to hire armies of data scientists. You can repurpose employees currently dedicated to repetitive manual processes. And many firms are further deploying internal centers of excellence, designed specifically to propagate machine intelligence and drive innovation across the enterprise.

Ayasdi's machine intelligence platform combines innovations in scalable computing, automation, machine learning and topological data analysis to help your firm find previously unknown insights in massive volumes of data with thousands of variables. It leverages the shape of your data to surface subtle relationships, often hard to uncover using conventional analytical tools.

Using Ayasdi's platform, organizations worldwide are building and deploying intelligent applications at scale to become truly data driven businesses.

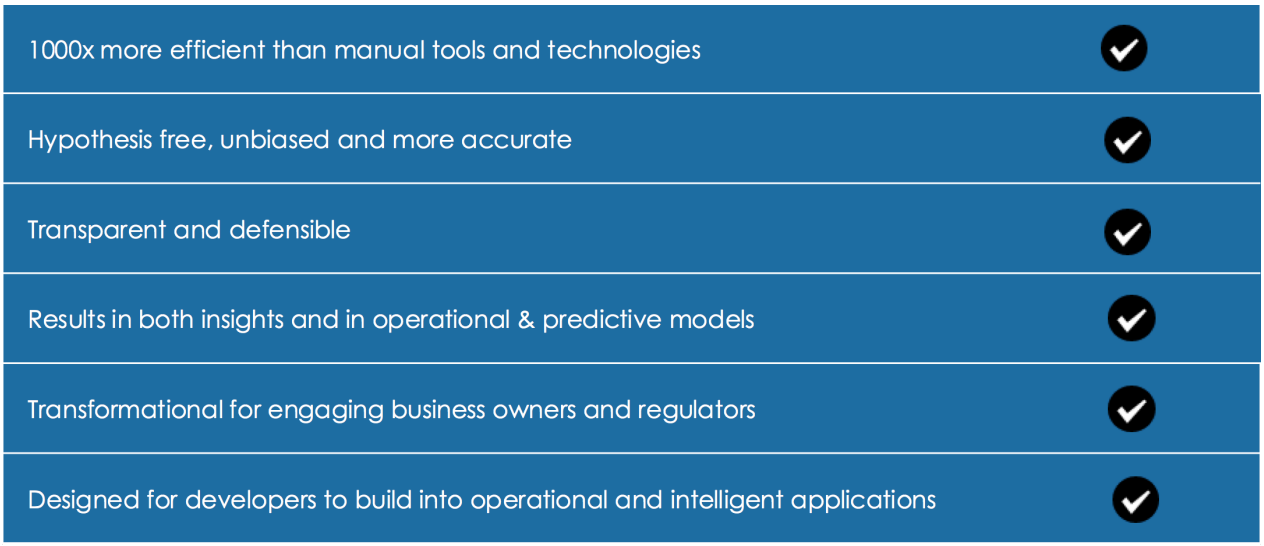


Figure 8: Why organizations are finding Ayasdi to be so transformative

ABOUT AYASDI

Ayasdi is on a mission to help our customers gain transformative advantage through their big and complex data. Our revolutionary machine intelligence platform leverages automation, machine learning and topological data analysis to simplify the extraction of knowledge from even the largest and most complex data sets confronting organizations today and to facilitate the deployment of intelligent applications across the enterprise. Developed by Stanford computational mathematicians, Ayasdi's unique approach to machine intelligence leverages breakthrough mathematics, highly automated software and scalable compute to revolutionize the process of converting big data into business impact. We are excited to count many of the Fortune 500, plus leading governments and research institutions as our clients and partners.

AYASDI

CONTACT US

Ayasdi, Inc.
4400 Bohannon Drive
Suite #200
Menlo Park, CA 94025

sales@ayasdi.com
visit ayasdi.com

@ayasdi