Understanding TDA:
What we do, how we do it, why we do it.

WHITE PAPER

Introduction

Our company has built an AI platform that powers the design, development, and deployment of enterprise-scale, intelligent applications. Our approach, our underlying technology, and our products are expressly crafted to deliver against an enterprise's requirements in this area – with the goal of delivering extraordinary business value.

This paper details our perspective on intelligent applications, what constitutes an intelligent application, and the importance of enterprise scale. The paper continues with our underlying technology and the architecture that supports that technology.

It closes with a review of the architecture of our solution, the platform components and a brief discussion of the underlying technology.

The Elements and Architecture of Enterprise Scale

For an organization considering how to deploy artificial intelligence capabilities across their enterprise, the five pillars represent the starting point. To scale that effort, however, requires additional considerations. Primary among them is the concept of enterprise scale.

To deliver applications into the enterprise, they need to meet the IT requirements of the enterprise.

Enterprise scale goes beyond compute. Math is far more difficult to scale than servers and the result with many approaches is that the machine learning algorithms must be "downgraded" or simplified to run against larger and larger datasets. TDA has invested significant resources to deliver enterprise-class workload management to ensure its math scales with the size of the challenge.

Enterprise scale also means that your AI strategy needs to integrate into your existing IT infrastructure – this means Active Directory for User Management, leveraging existing data lake infrastructure and respecting the security and compliance guidelines for operational systems.

Enterprise scale also means defining exceptional user experiences. For most analytic and "pseudo-AI" solutions the user interface is Excel, PDF or PowerPoint. This limits the impact these insights can have on an organization - not only in terms how broadly it can be consumed, but also in the fact that it represents a point in time. Enterprise scale demands that AI go to market in the form of well-designed applications that put the user experience first, ensuring these powerful technologies are adopted broadly.

Finally, enterprise scale means applications that facilitate collaboration and accelerate the creation of workflows between line of business, data science and IT. This framework for application development should provide powerful technologies in the right application package, in the right places, solving the right problems – automatically.

These are all elements that TDA focuses on what it means when it talks about enterprise scale. They come from our experience in dozens of enterprise-grade deployments – not just pilots or POCs.

Our Machine Intelligence Platform

Our platform is able to ingest data from a number of sources, but ultimately it will convert that data into something that can reside in a Hadoop. If the data is already in Hadoop, TDA will interact with it there.

Once the data is in Hadoop, TDA's technology can interact with it on a number of different levels, from our Workbench product to a specific application.

Our intelligence platform combines Topological Data Analysis, Machine Learning, Statistical, and Geometric Algorithms to automatically find patterns that elude other technologies. Furthermore, our approach provides exceptional explainability and justification for its selections. The combination of algorithms and the way in which we are leveraging them enable our platform to support a range of algorithmic selection from fully automated (unsupervised) to semi-supervised and ultimately supervised analysis.

Up Stack

There are three elements that take advantage of the REST API. The first is the Python SDK, the second is TDA Workbench and the third is TDA Envision.

The Python SDK provides programmatic access to the technology layer for the purpose of building intelligent applications. Developers with Python and Data Science backgrounds can construct applications using this approach.

TDA Workbench is a data science application that enables users to create topological models of their data and visually explore the relationships inherent in those models. Again, that analysis can be done in an unsupervised approach or in a supervised approach. Elements of interest can be explained and justified in simple terms – a key capability noted earlier in this document.

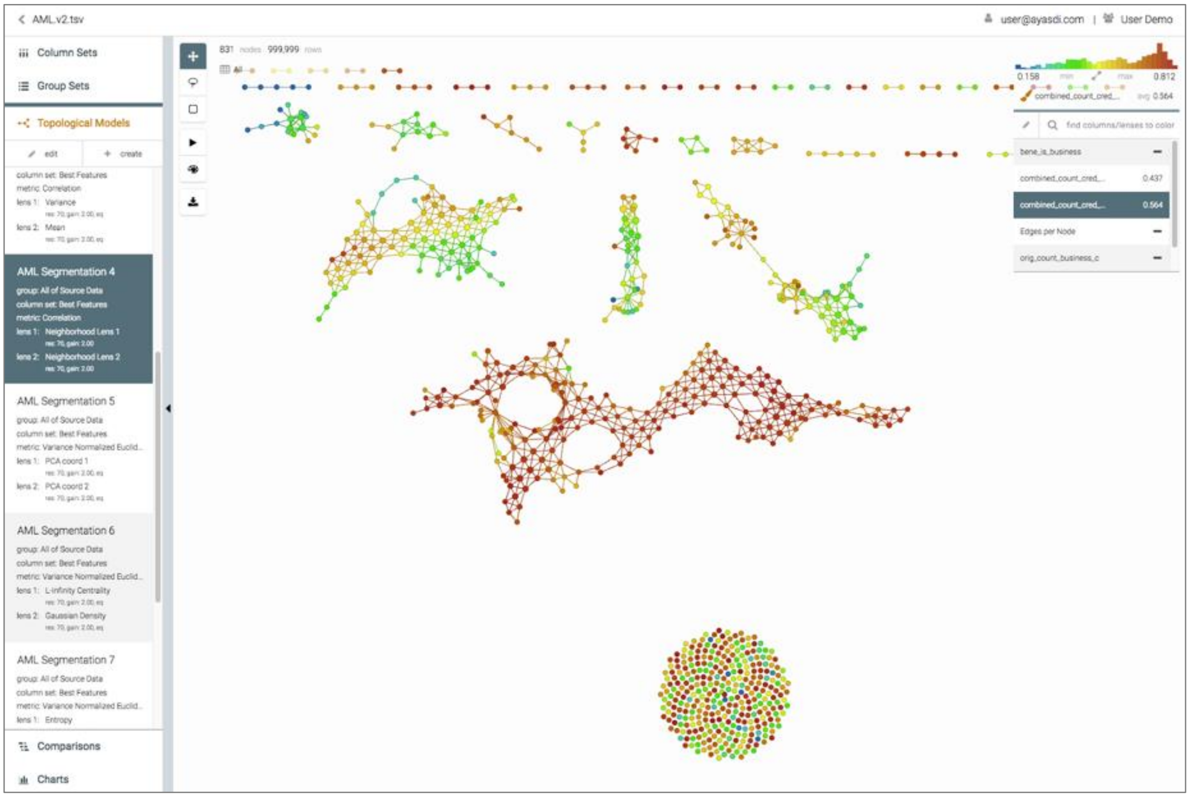

Figure 1: TDA Workbench

While powerful, Workbench offers a subset of the overall capabilities exposed by the REST API and the Python SDK. Workbench is targeted at data science teams and represents an interim step towards the creation of application.

To further facilitate application development, TDA has built an application development framework called Envision. TDA Envision democratizes and accelerates the creation of intelligent applications through a collaborative, workflow-oriented approach that is designed to include business users throughout the process, ensuring the requirements of subject-matter experts are met. Analysts, data scientists, and IT groups can leverage our powerful AI platform to build business-specific intelligent applications using a pre-built library of UI components. The new web-based framework features a simple user interface that is easy to learn and navigate enabling business users and data analysts to work in lock-step together.

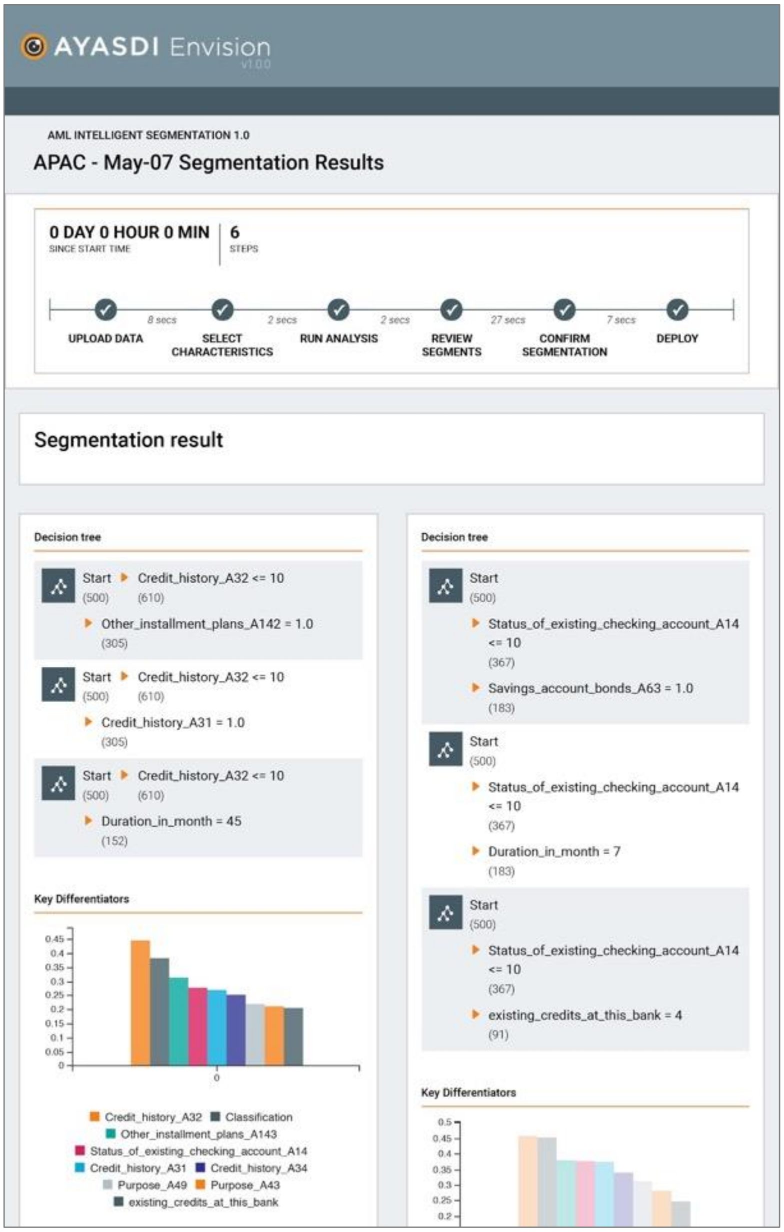

Figure 2: TDA Envision

Applications built using the Python SDK are consumed by a broader portion of the organization from analysts to subject matter experts and business leaders.

The Analytical Engine of TDA

The engine that powers our platform is called Topological Data Analysis (TDA). TDA is considered one of the most significant technological advancements ever funded by DARPA and is the source of a broad range of awards and recognition. TDA has the only enterprise implementation of this powerful technology.

TDA is based on topology, the mathematical subdiscipline that studies the notion of shape. TDA refers to the adaptation of this discipline to analyzing highly complex data. It draws on the philosophy that all data has an underlying shape and that shape has meaning. This knowledge is widely understood for certain shapes – but data does not restrict itself to certain shapes. For example, look at the following:

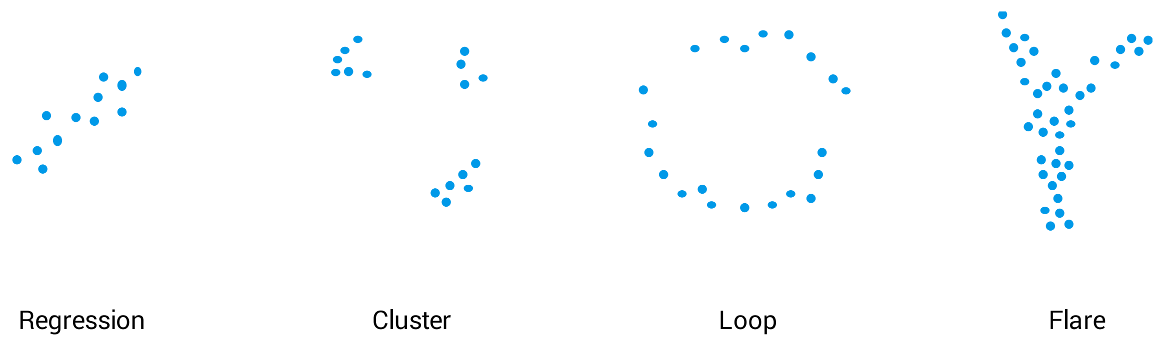

Figure 3: Four examples of the shape of data.

The shape on the far left is well known to generations of mathematicians, statisticians, and data types. This represents a regression line, obtained by fitting a straight line to the data points using a natural measure of fit. A straight line is certainly a shape, and in the above example, we find that a straight line fits the given data quite well. That piece of information is extremely important for several reasons. One is that it gives us the qualitative information that the y-variable varies directly with the x-variable (i.e. that y increases as x increases). Another is that it permits us to predict with reasonable accuracy one of the variables if we know the value of the other variable. The idea is that the shape of a line is a useful organizing principle for the data set, which permits us to extract useful information from it.

Similarly, the second image is that of three distinct groups. It is easy to see that no straight line faithfully represents this data.

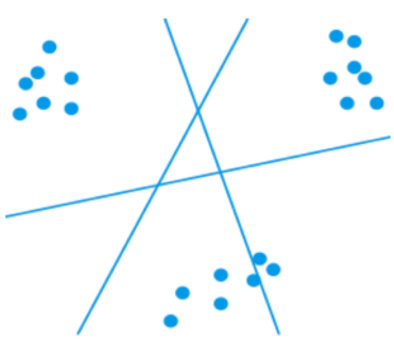

Figure 4: Data that forms distinct clusters.

The reason is that this data set breaks into a set of three tightly concentrated clusters. One might not initially think of this as having anything to do with shape, but after a moment's reflection, we realize that the most fundamental aspect of any shape is the number of connected pieces it breaks into. So, in this case, we see that the shape of this data set is of fundamental importance, and that its shape is that of a set of three discrete points.

While clearly distinct from regression, clustering is also well understood from a mathematical perspective. It is the subject of many books, and there are journals devoted to it.

But what about the third shape?

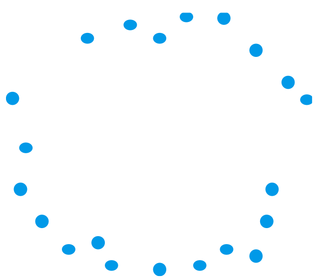

Figure 5: Data that forms a loop.

Notice that this shape also does not fit along a line, and does not break into clusters, but rather has a “loopy” behavior. This kind of structure is often associated with periodic or recurrent behavior in the data set. In both preceding examples, if you were to approximate them with a regression line you would be missing the vast majority of the information encoded within the shape. The actual shape is “hidden” to regression. The circle is “hidden” to clustering algorithms.

Yet we find this data shape everywhere from economic cycles to disease states. Unlike regression or clustering, loops are not well understood or modeled in data science.

One final example.

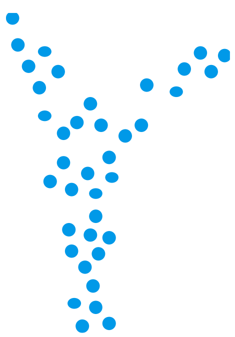

Figure 6: Data that forms the shape of a “Y”.

The shape is in this case that of a capital letter “Y”.

This is another kind of shape, which also occurs frequently in data. Note that it has a central core and three "flares" extending from it. This might represent a situation where the core represents the most frequently occurring behaviors, and the tips of the flares represent the extreme behaviors in the data. It is clearly distinct from the three other shapes we have already discussed.

One example of this kind of structure would occur if we considered a data set gathered from measurements taken from an airliner, during the entire course of a flight. The center "core" is when the plane is cruising at 35K feet, in non-turbulent conditions. The upper left flare may represent thrust associated with takeoff. The bottom center might represent deceleration on landing and the upper right flare may represent flying at 35,000 feet in turbulent conditions.

Again, a line, loop or clusters wouldn't describe this particular shape.

Topology describes shape and does so very well. If we only had simple shapes, this would not be of great importance.

But data doesn't have just four shapes. Complex data has billions if not more shapes – all of which encode information.

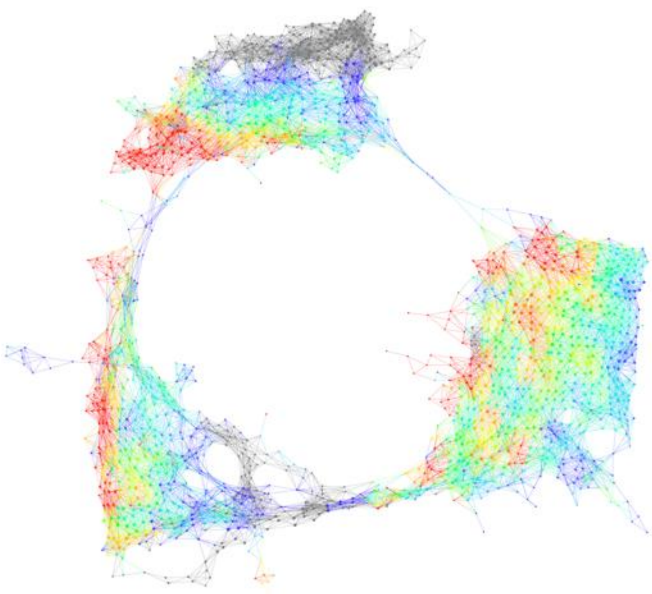

Figure 7: Complex data has billions if not more shapes.

The immense variety possible among shapes suggests that we should not attempt to enumerate all the possible shapes with a individual mathematical approaches (e.g. regression), and create templates for each, but rather find a flexible way of representing all shapes.

That is one of the key problems topological data analysis deals with – it determines the shape of the data and the information encoded therein.

As a framework, TDA uses dozens of machine learning algorithms to determine the best shape of the data. In doing so it uncovers subtle patterns missed by other approaches that rely on a single algorithmic dimension. Furthermore, it accelerates the process of determining what in the data is important and what is not and can do so in an unsupervised fashion – meaning users often find answers to question they didn't even know to ask. Topological data analysis extends and improves the performance of any machine learning algorithm creating a measure of future proofing that does not exists in other solutions. Finally, topological data analysis provides justification for its findings and recommendations.

Summary

Enterprises that adopt machine intelligence will outperform enterprises that don't. Intelligence in business applications is not a fad, it is an inflection point. Intelligence is not reserved for Google, Facebook, Amazon and Microsoft, it is available to any enterprise with the sophistication and drive to transform their business.

Intelligence will define the winners and loser over the coming decade – perhaps less.

As a company, TDA has pioneered the development of these intelligent systems. The five pillars outlined at the start of this paper: Discover, Predict, Justify, Act and Learn represent all the necessary components of an intelligent system. Choosing one, two or even three components will deliver but a fraction of what the entire system would deliver.

A commitment to establishing an intelligence driven enterprise requires a commitment to every principle - however, the payoff for those enterprises making that investment will be immense.

About