

**O'REILLY**<sup>®</sup>  
Strata Data Conference  
PRESENTED WITH **CLOUDERA**

# Assumed Risk vs Actual Risk: Behavior-based Risk Modeling

Viridiana Lourdes, PhD  
Data Scientist, AyasdiAI

# Agenda

1. Problem: Money laundering.
2. Risk modeling: assumed vs actual risk.
3. Approach: TDA Segmentation.

# Money Laundering

The laundering of dirty money occurs when the perpetrators steer the ill-gotten cash through legitimate businesses or financial institutions to legitimize the money.

Running dirty money through the wash allows the criminals to spend that money without fear of reprisal.



# Money Laundering

Between \$500 billion and \$1.5 trillion cash is laundered internationally per year.

If a financial institution processes funds from criminal activity, the institution could be drawn into active complicity with criminals and become part of the criminal network itself. Even if it is unintentional.

Money Laundering rewards corruption and crime, it damages the integrity of the entire society.

# Anti-Money Laundering (AML)

Procedures, laws and regulations intended to prevent criminals from Money Laundering.



In case of robbery, extortion or fraud, money laundering investigation is frequently the only way to locate the stolen funds and restore them to the victims.

# Anti-Money Laundering (AML)

Criminals are using more sophisticated means to remain undetected, AML actions need to be at the same level.

In the last five years, there has been an explosion of companies with proposals on how to address regulatory requirements using technology.

# AML process

-   
Transactions
-   
Client Profiles  
(CDD, KYC, etc.)
-   
Sanctions/  
PEP/Watch  
Lists



Risk breakdown based on assumed risk, profiles captured during onboarding (Country, Line of business, products, ...).

Transaction Monitoring System



Event Creation with filtering. Some priority/ranking. High rate of false positive.

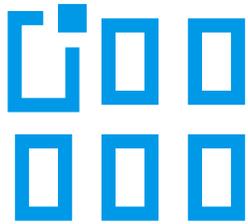


Alert investigation are lengthy and expensive because of limited context.

# Agenda

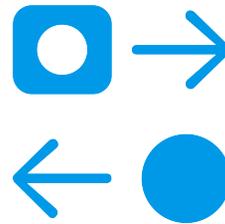
1. Problem: Money laundering.
2. Risk modeling: assumed vs actual risk.
3. Approach: TDA Segmentation.

# Assumed Risk

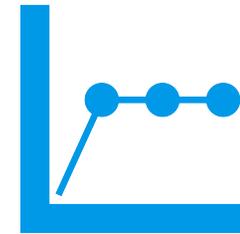


## Standard KYC data

- Customer
- Products & Services
- Geographies

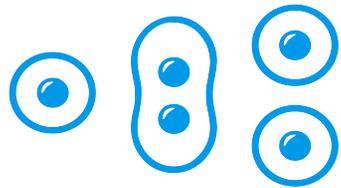


## Risk scoring

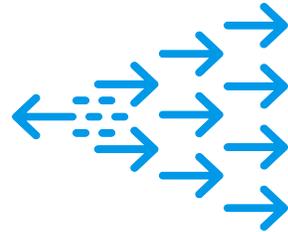


## Relatively static in nature

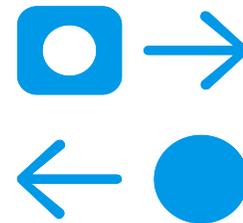
# Actual Risk



Based on behavior



Augmented by changes  
to that behavior and/or  
environment over time



Dynamic in  
nature

# AML process

-  Transactions
-  Client Profiles (CDD, KYC, etc.)
-  Sanctions/ PEP/Watch Lists



Risk breakdown based on assumed risk, profiles captured during onboarding (Country, Line of business, products, ...).

Transaction Monitoring System



Event Creation with filtering. Some priority/ranking. High rate of false positive.

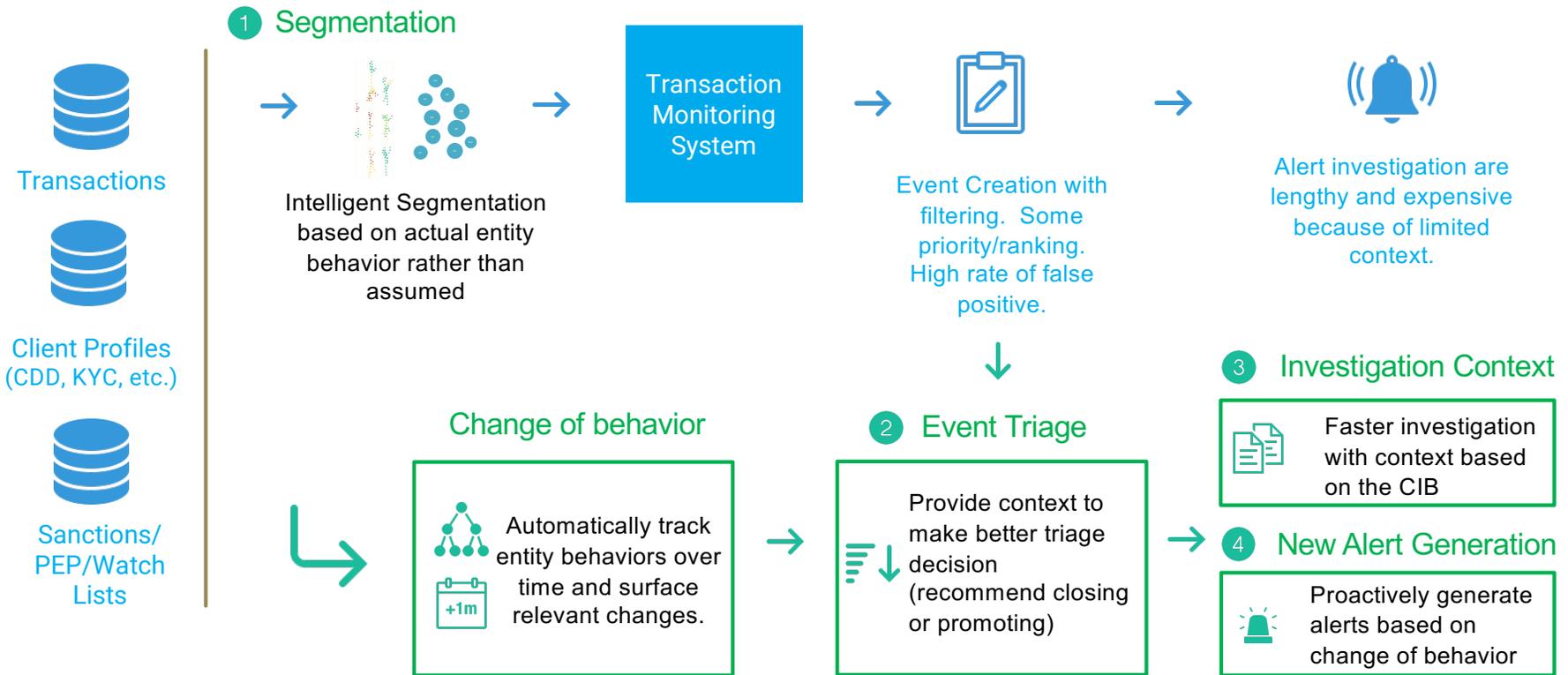


Alert investigation are lengthy and expensive because of limited context.

# Agenda

1. Problem: Money laundering.
2. Risk modeling: assumed vs actual risk.
3. Approach: Segmentation.

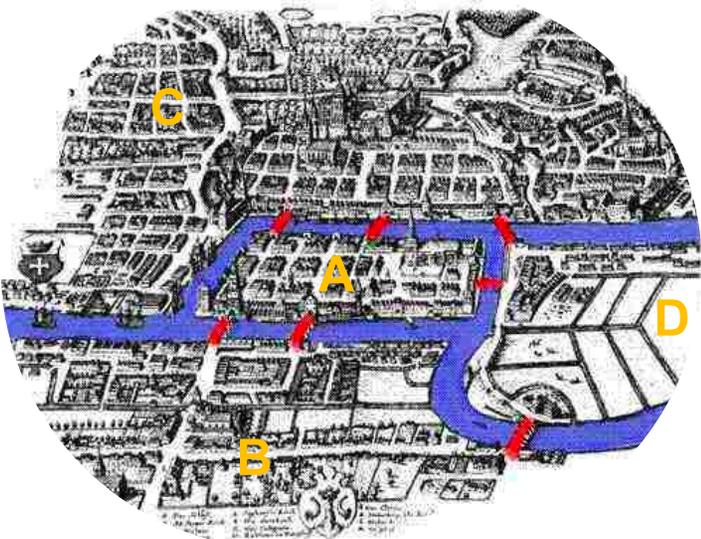
# AML process with Segmentation



# TDA Segments

- The challenge facing enterprises today is not data size, but data complexity.
- We are able to define meaningful segments using Topological Data Analysis (TDA).
- TDA is the use of **topology** to data analysis.

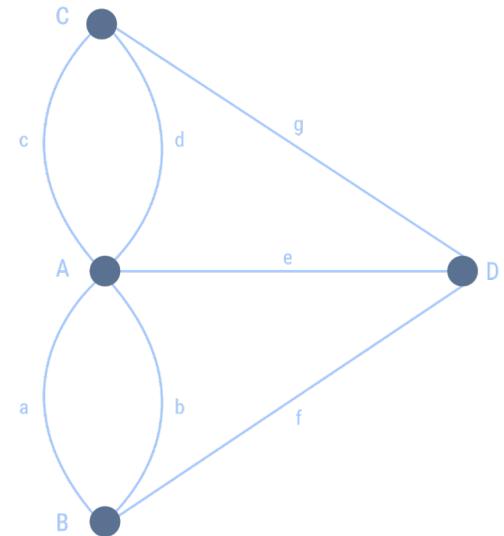
# Topology



City of Königsberg in Prussia set on both sides of Pregel river

Challenge: design a walk through the city that would cross each of those bridges once and only once.

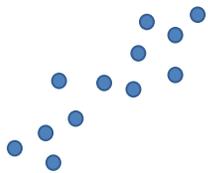
Euler's thinking: the only important feature of a route is the sequence of bridges crossed. Replace each land mass with a node and each bridge with an edge.



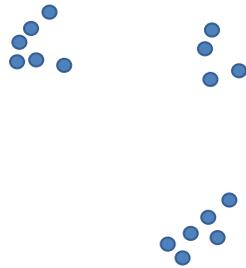
Topology studies the properties of spaces that are preserved under stretching and bending (not tearing or gluing).

# Topological Data Analysis

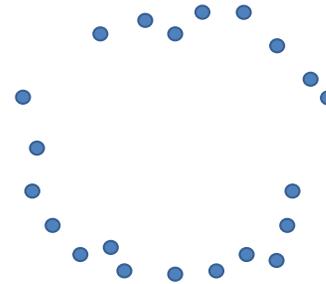
- TDA is the approach that uses the “shape” of the data to extract information on complex datasets to create segments.



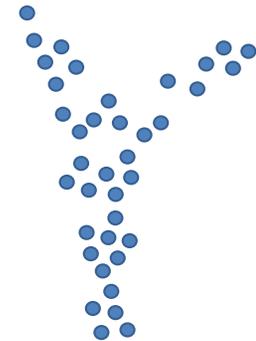
Line



Clusters



Loop

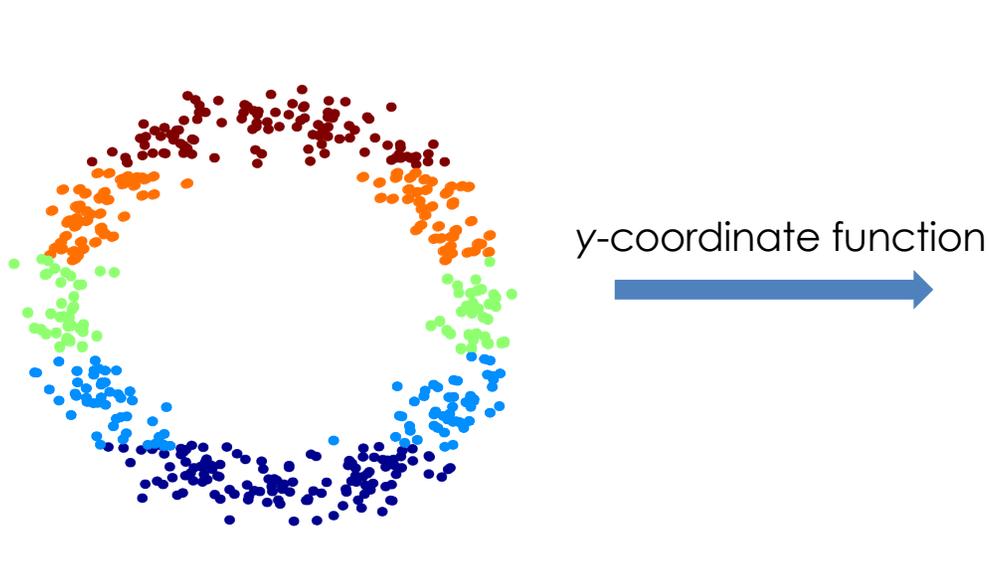


Flares

# Topological Data Analysis

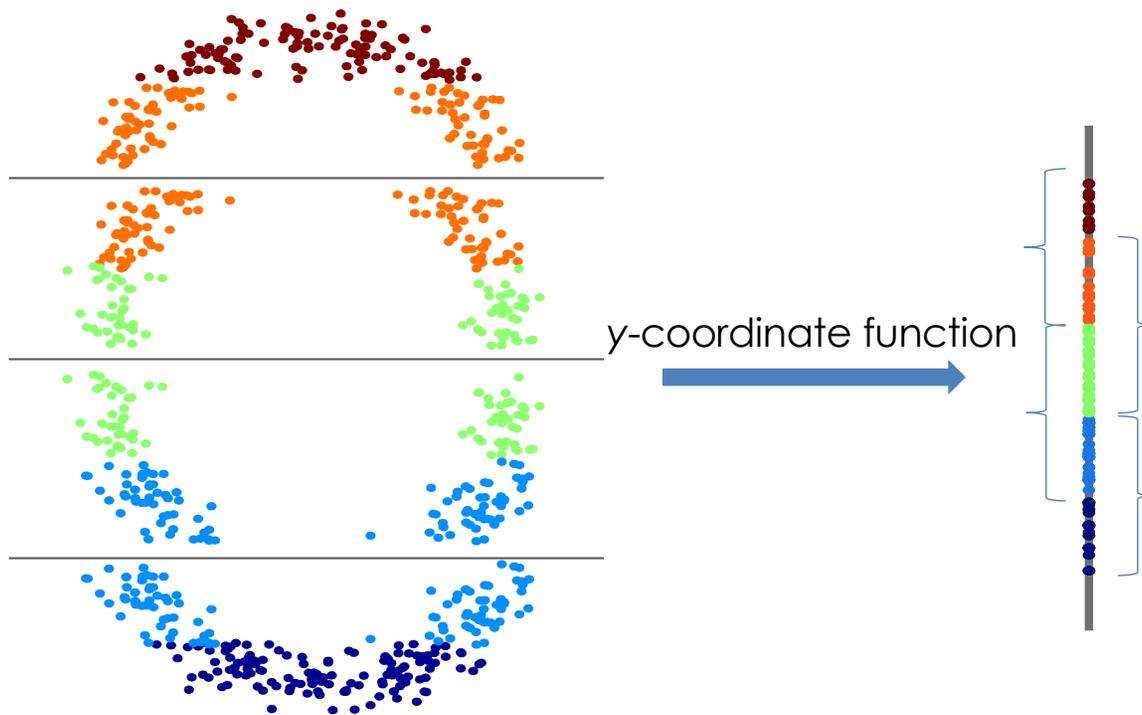
- TDA is the approach that uses the “shape” of the data to extract information on complex datasets to create segments.
- The core idea behind TDA is the Mapper algorithm.
- The Mapper is a method created by *Gurjeet Singh, Facundo Memoli and Gunnar Carlsson* and published in 2007.
- We used AyasiAI’s approach of TDA, which offers a simple way of interrogating data to understand the underlying properties that characterize the segments and sub-segments that lie within data.

# Creating Topological Networks



- TDA applies a function (lens) to the data set
- In this example, data points are mapped to their y-coordinate value

# Creating Topological Networks

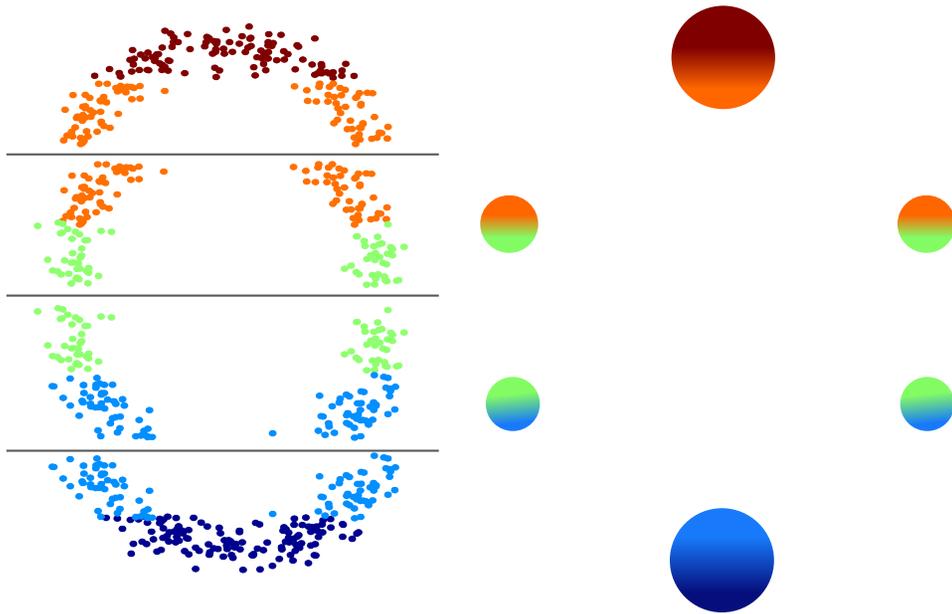


The algorithm subdivides the image of the function into overlapping bins of data points

Points within bins have similar function values

Because of the overlap, data points can fall into multiple bins

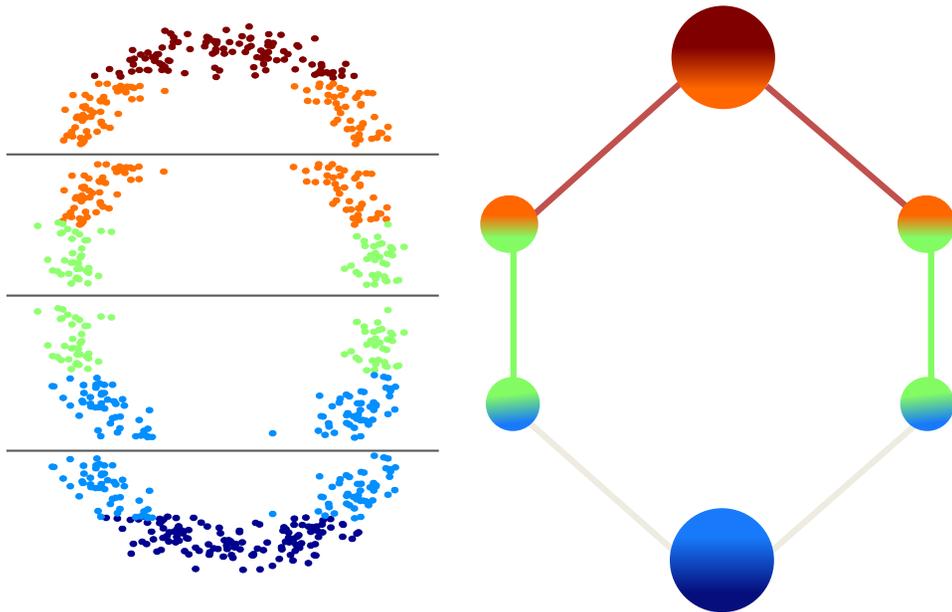
# Creating Topological Networks



The algorithm clusters each of these sets of data points independently using a measure of similarity on the data points

A node represents a set of data points that are similar with respect to the measure of similarity

# Creating Topological Networks

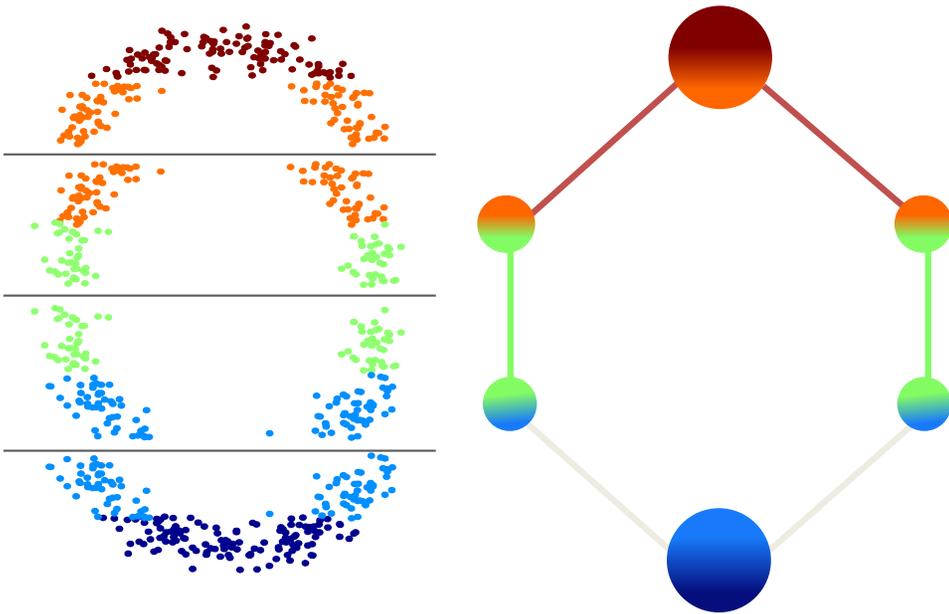


Nodes with data points in common are connected by edges to create a network

As the data was divided into overlapping data sets, a data point can be in multiple nodes

The network captures the underlying shape and behavior of the data

# Creating Topological Networks

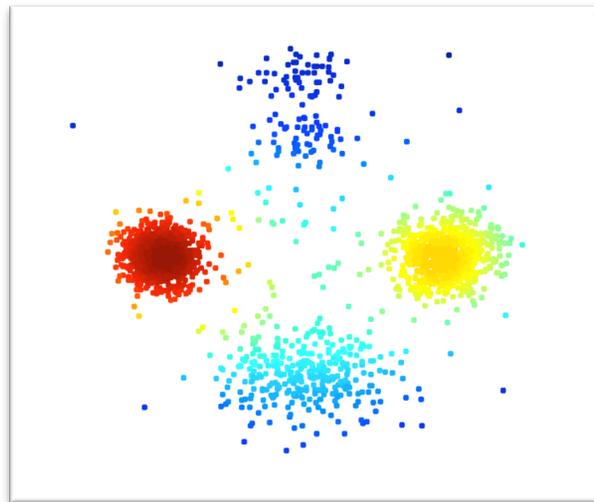


1. Apply a function (lens) to a data set
2. Create a visual network of nodes connected by edges using a measure of similarity.

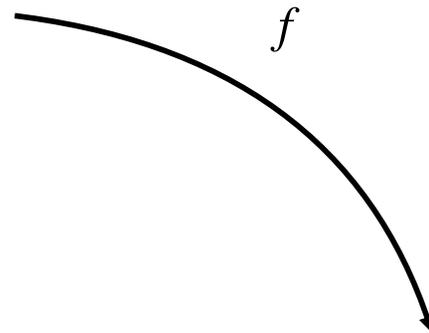
Result:

A compressed summary of the data.

# Creating Topological Networks



$d$  : metric on data



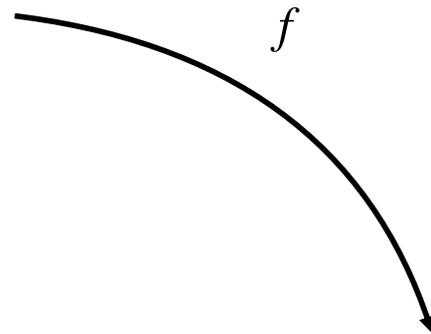
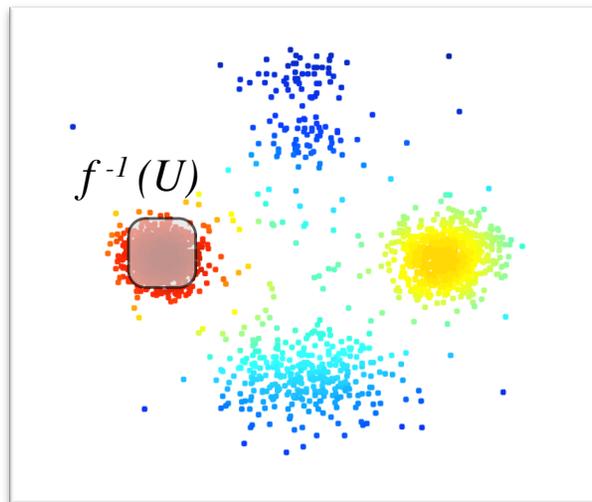
$f$  is a function from the data to some other space (e.g. the real line)

In this example,  $f$  is a density estimator at each point

Data points are colored by a density estimator function

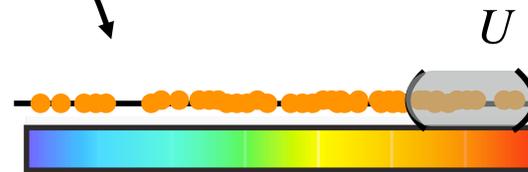


# Creating Topological Networks

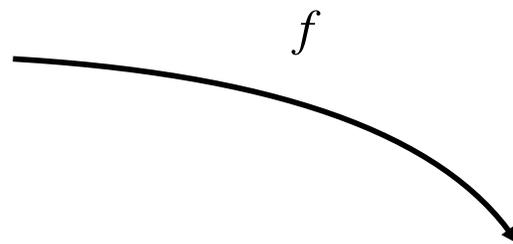
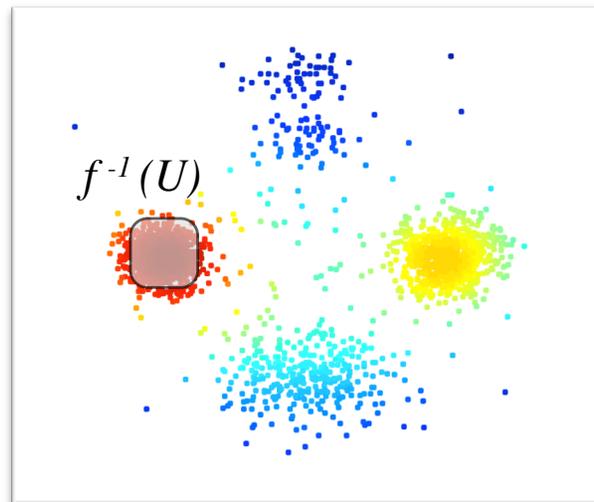


$U$  defines a set of similar points in the image of  $f$

$f^{-1}(U)$  is a set of data points that are similar in the image of  $f$

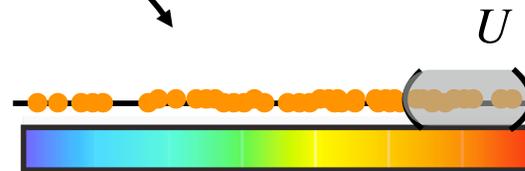


# Creating Topological Networks

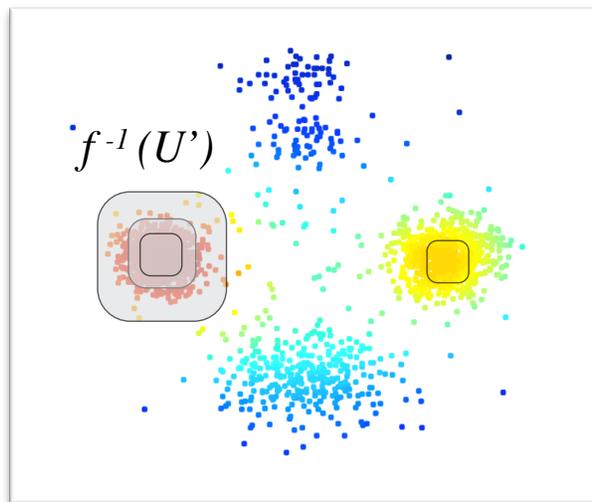


Using the metric, perform clustering to determine the sets of similar points in  $f^{-1}(U)$

Represent each set of points similar in both function and metric as node

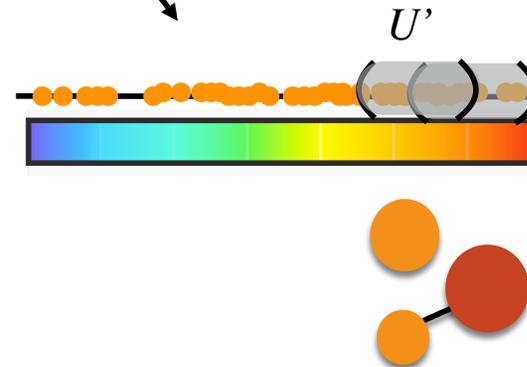


# Creating Topological Networks

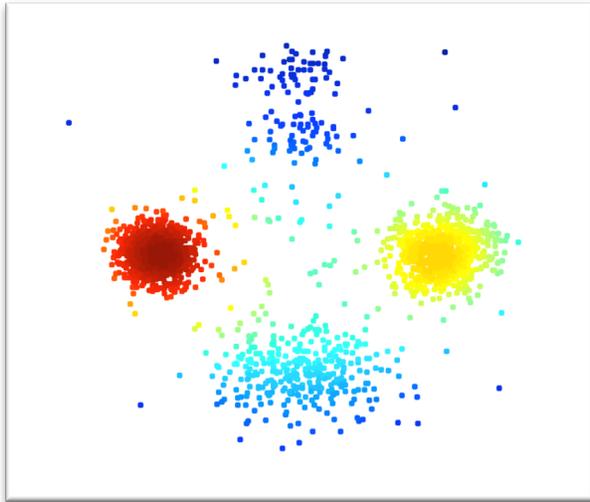


Repeat process with a different set of similar points in the image of the function

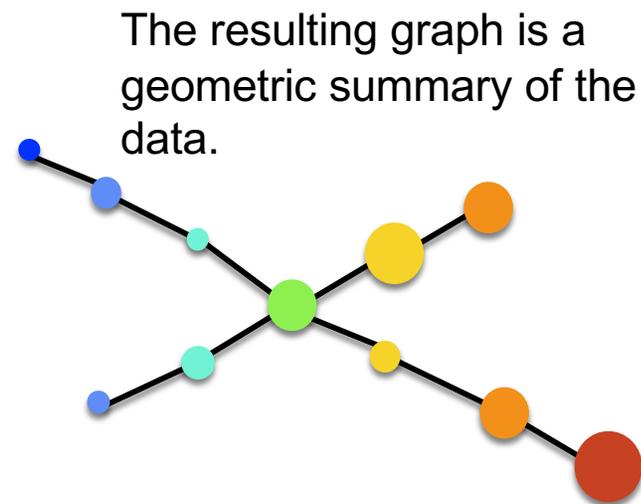
Edges between nodes indicate overlapping points. They capture the continuous nature of the data when viewed through the function



# Creating Topological Networks

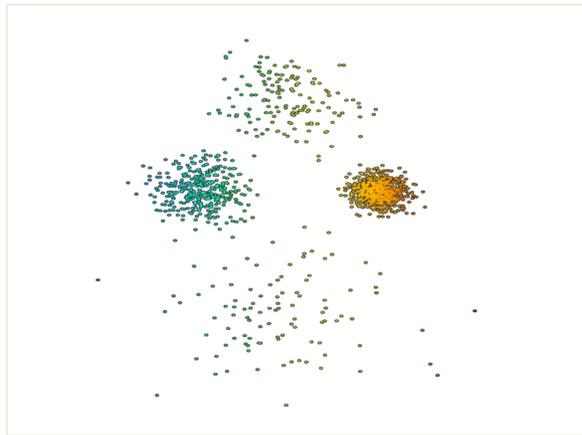


Nodes represent a set of points similar in both function and metric



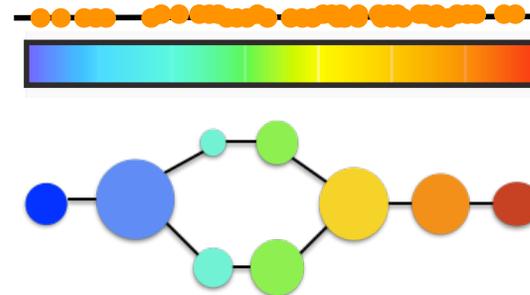
Edges between nodes indicate overlapping points.

# Creating Topological Networks



$f$

Different functions produce different summaries of the data. In this example,  $f$  is now the projection of each point on the x-axis

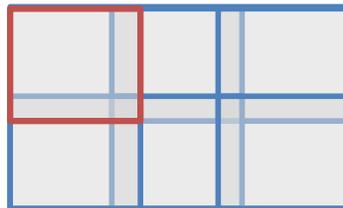


# TDA Mapper Overview

- 1 Use the **Lens** to perform dimensionality reduction on the data

E.g. PCA, MDS, Neighborhood Lens, Entropy etc...

- 2 Use the **resolution** and **gain** to create an open cover (overlapping sections) on that low dimensional space



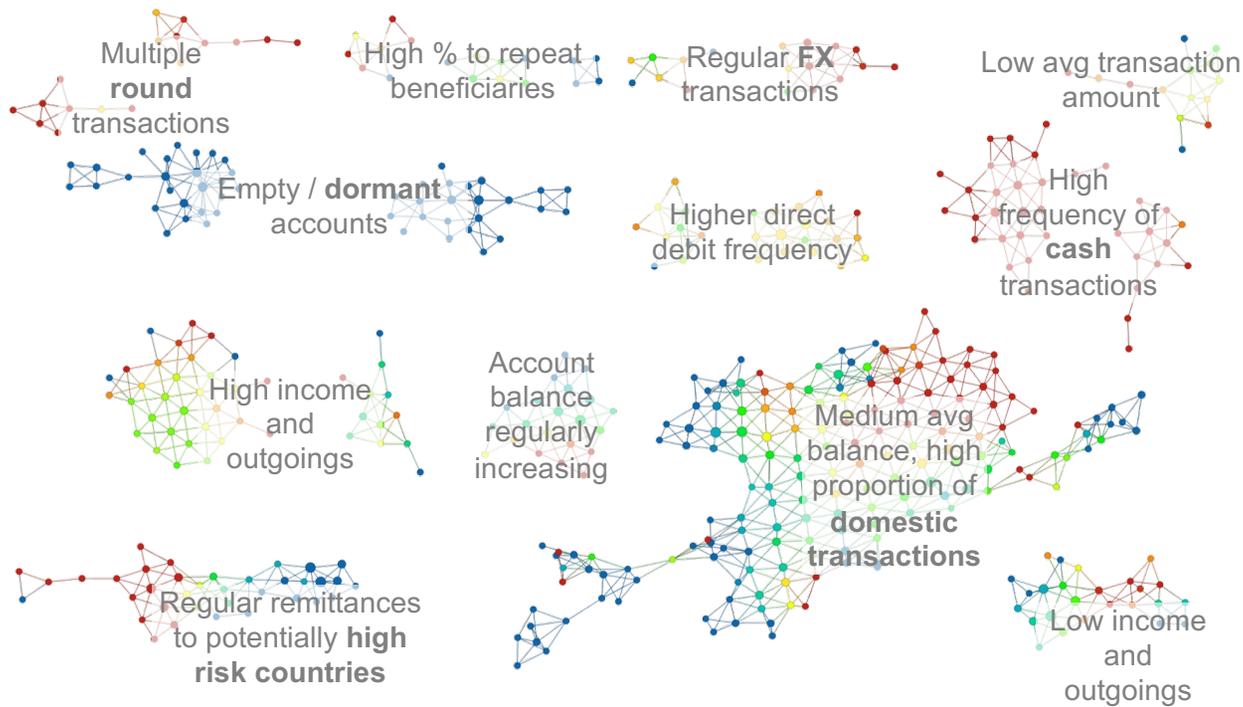
- 3 Use the **metric** (the measure of similarity) to cluster in the high dimensional space within each low dimensional section

E.g. haversine distance, Euclidean distance, Hamming distance etc..

- 4 Create a **network of similarity** - the clusters become **nodes**, and any shared points add in an **edge**



# Assumed vs Actual Risk



Network of customers based on similarity of transactional behaviours

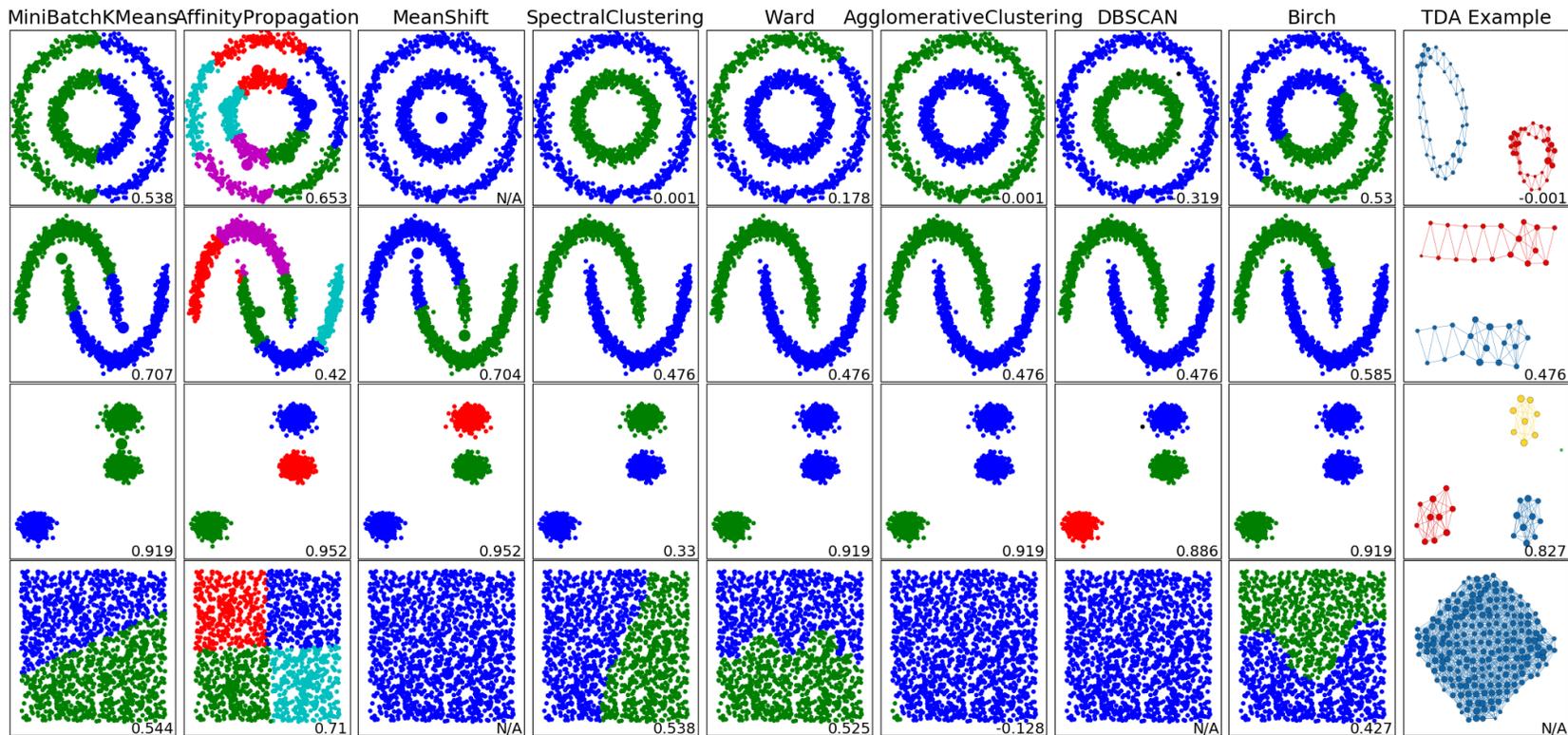
Node: Group of similar customers

Connection: Links two similar groups

# Segments using TDA

- No need to specify the number of segments in advance.
- Represents continuous and cyclic phenomena much better than any form of clustering.
- No assumptions on the shape of the data.
- Label and unlabeled transaction information.
- TDA segments can be used with other models to improve performance.

# TDA Benefits - Segmentation



# Key Takeaways

1. Segmentation is the foundational element to improve **actual risk** modeling.
2. Segmentation can be integrated with existing systems to enhance the performance and operational efficiency of AML.
3. Data has shape and shape has meaning.

# References

1. Gurjeet Singh, Facundo Memoli and Gunnar Carlsson (2007). *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*. Eurographics Symposium on Point Based Graphics, European Association for Computer Graphics.
2. Gunnar Carlson (2009). *Topology and Data*. Amer. Math. Soc. 46

AyasdiAI was founded in 2008 by Gurjeet Sigh, Gunnar Carlsson and Harlan Sexton.

# Thank you!

[viridiana.lourdes@ayasdi.com](mailto:viridiana.lourdes@ayasdi.com)

**O'REILLY**<sup>®</sup>  
Strata Data Conference

PRESENTED WITH **CLOUDERA**

[strataconf.com](http://strataconf.com)  
[#StrataData](https://twitter.com/StrataData)

# Rate today's session

**Cyberconflict: A new era of war, sabotage, and fear**

[See passes & pricing](#)

David Sanger (The New York Times)  
9:55am-10:10am Wednesday, March 27, 2019  
Location: Ballroom  
Secondary topics: Security and Privacy

[Add to Your Schedule](#)  
[Add Comment or Question](#)

**Rate This Session**

We're living in a new era of constant sabotage, misinformation, and fear, in which everyone is a target, and you're often the collateral damage in a growing conflict among states. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. Moving from the White House Situation Room to the dens of Chinese, Russian, North Korean, and Iranian hackers to the boardrooms of Silicon Valley, David reveals a world coming face-to-face with the perils of technological revolution—a conflict that the United States helped start when it began using cyberweapons against Iranian nuclear plants and North Korean missile launches. But now we find ourselves in a conflict we're uncertain how to control, as our adversaries exploit vulnerabilities in our hyperconnected nation and we struggle to figure out how to deter these complex, short-of-war attacks.

**David Sanger**  
The New York Times

David E. Sanger is the national security correspondent for the *New York Times* as well as a national security and political contributor for CNN and a frequent guest on *CBS This Morning*, *Face the Nation*, and many PBS shows.



Session page on conference website

✓ Attending [Notes](#) [Remove](#)

**Cyberconflict: A new era of war, sabotage, and fear**

9:55 AM - 10:10 AM, Wed, Mar 27, 2019

**Speakers**

 David Sanger  
National Security Correspondent  
The New York Times

📍 Ballroom

*Keynotes*

David Sanger explains how the rise of cyberweapons has transformed geopolitics like nothing since the invention of the atomic bomb. From crippling infrastructure to sowing discord and doubt, cyber is now the weapon of choice for democracies, dictators, and terrorists.

[SESSION EVALUATION](#)

O'Reilly Events App