Topology and Machine Learning

A Global Map of your Data

Anthony Bak

AYASDI

Outline

Part I

- Problems with Big Data

Outline

Part I

- Problems with Big Data
- Topological Summaries

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Part II

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Part II

- Review

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Part II

- Review
- Why Topolgy? (Big ideas with examples)

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Part II

- Review
- Why Topolgy? (Big ideas with examples)
- More Examples

Outline

Part I

- Problems with Big Data
- Topological Summaries
- Examples

Part II

- Review
- Why Topolgy? (Big ideas with examples)
- More Examples

Caveats: I am only talking about the strain of TDA done by Ayasdi

Goals

TDA Review

The Data Problem

How do we extract meaning from **Complex Data**?

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data >500,000 features, complicated interdependencies)

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data $>$500,000 features, complicated interdependencies)
- Or both!

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data $>$500,000 features, complicated interdependencies)
- Or both!

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data >500,000 features, complicated interdependencies)
- Or both!

Problem 1: There isn't a single story happening in your data.

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data $>$500,000 features, complicated interdependencies)
- Or both!

Problem 1: There isn't a single story happening in your data.
Problem 2: Too many hypothesis to check.

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data >500,000 features, complicated interdependencies)
- Or both!

Problem 1: There isn't a single story happening in your data.
Problem 2: Too many hypothesis to check.

TDA will be the tool that summarizes out the irrelevant stories to get at something interesting.

The Data Problem

How do we extract meaning from **Complex Data**?

- Data is complex because it's "Big Data"
- Or has very rich features (eg. Genetic Data $>$500,000 features, complicated interdependencies)
- Or both!

Problem 1: There isn't a single story happening in your data.
Problem 2: Too many hypothesis to check.

TDA will be the tool that summarizes out the irrelevant stories to get at something interesting.

The shape (segmentations, groupings) represent verified hypothesis. You have to decide if they are interesting.

Math World

We start in ”Math World”

Math World

We start in "Math World"

- We'll draw the data as a smooth manifold.

Math World

We start in "Math World"

- We'll draw the data as a smooth manifold.
- Functions that appear are smooth or continuous.

Math World

We start in "Math World"

- We'll draw the data as a smooth manifold.
- Functions that appear are smooth or continuous.

$\Rightarrow$ We will not need either of these assumptions once we're in "Data World".

Math World

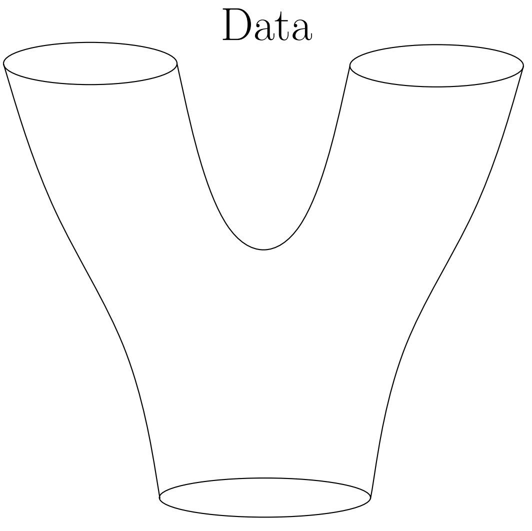
We start in "Math World"

- We'll draw the data as a smooth manifold.
- Functions that appear are smooth or continuous.

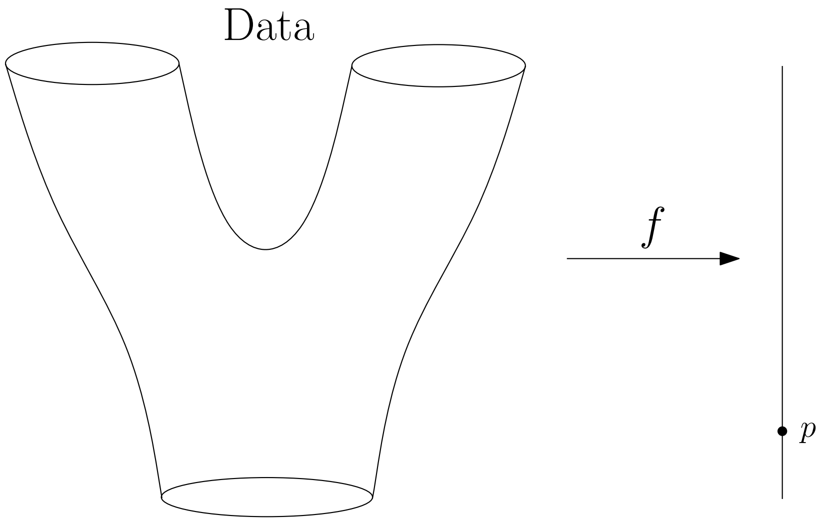
$\Rightarrow$ We will not need either of these assumptions once we're in "Data World".

$\Rightarrow$ Even more importantly, data in the real world is **never** like this.

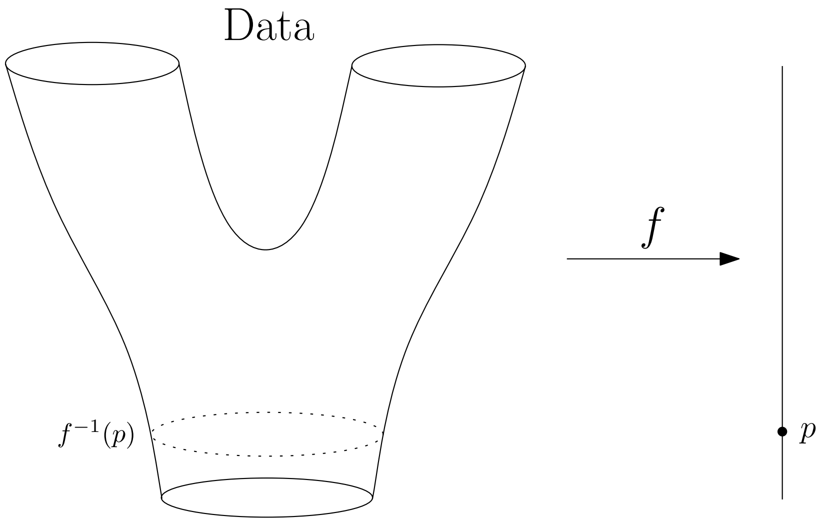
Math World



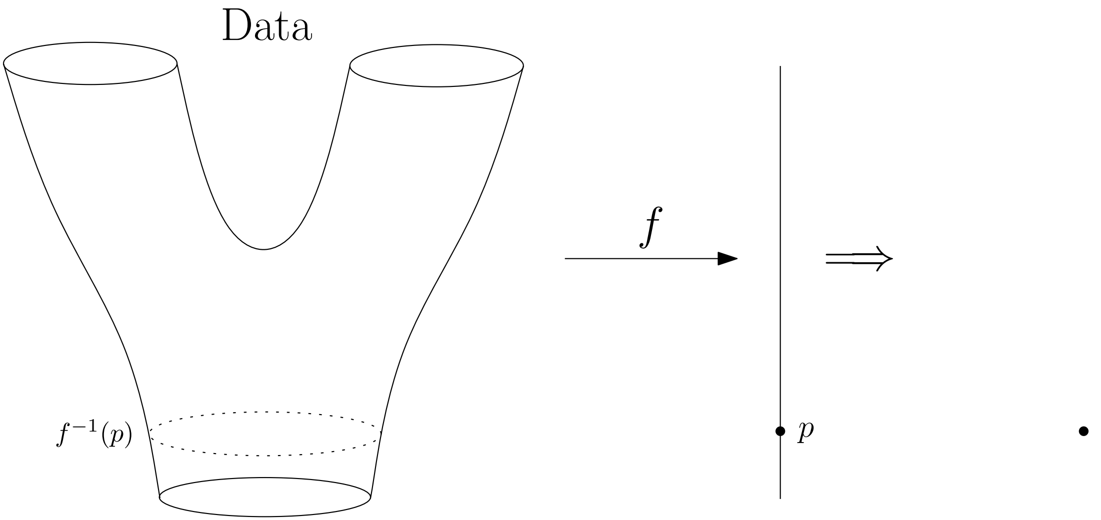
Math World



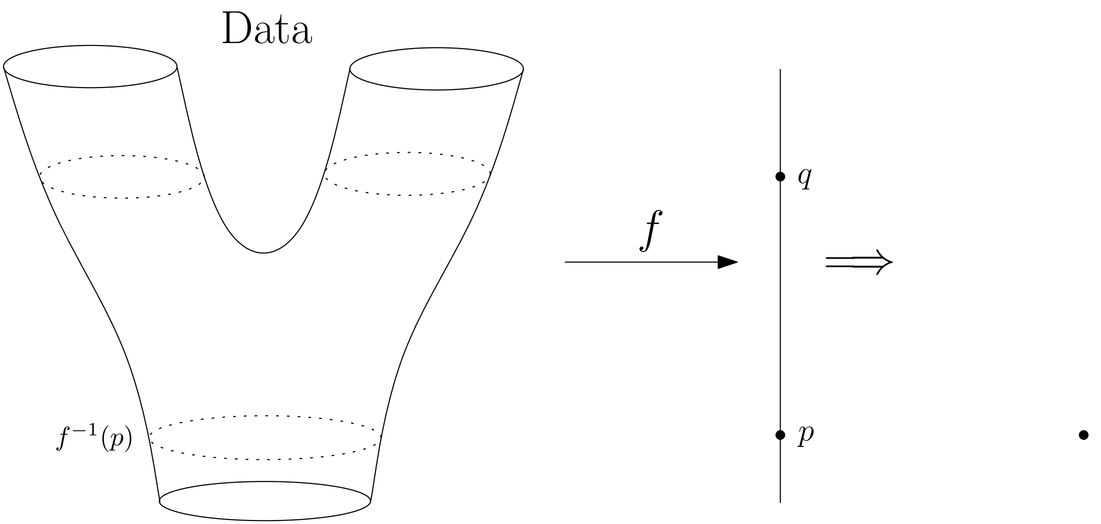
Math World



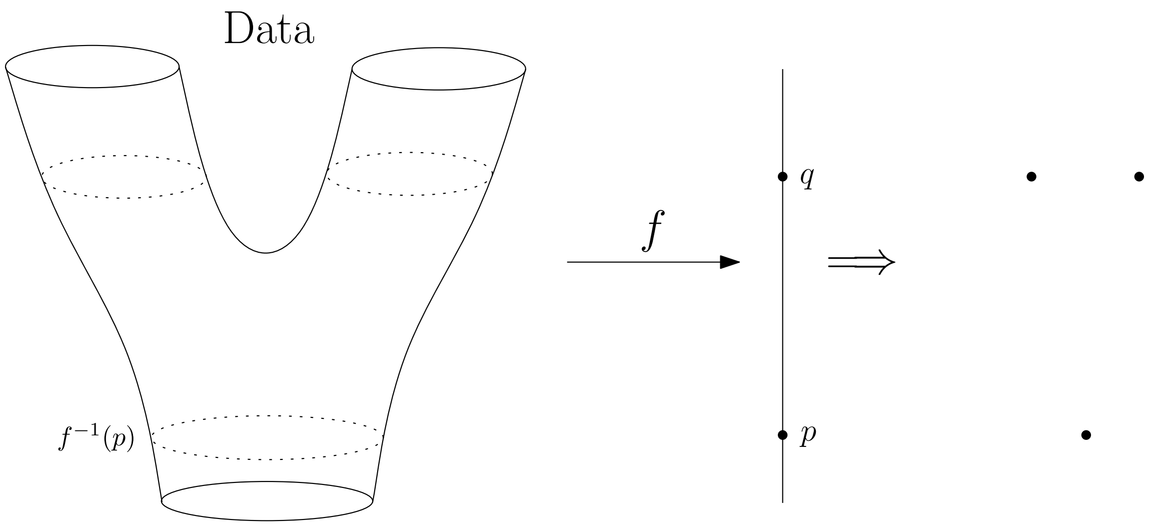
Math World



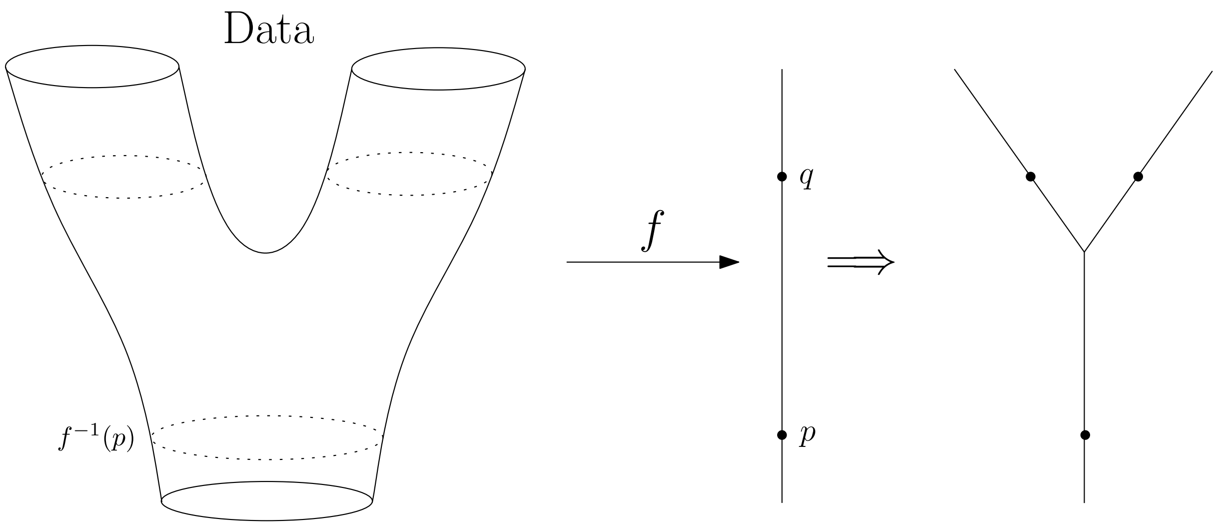
Math World



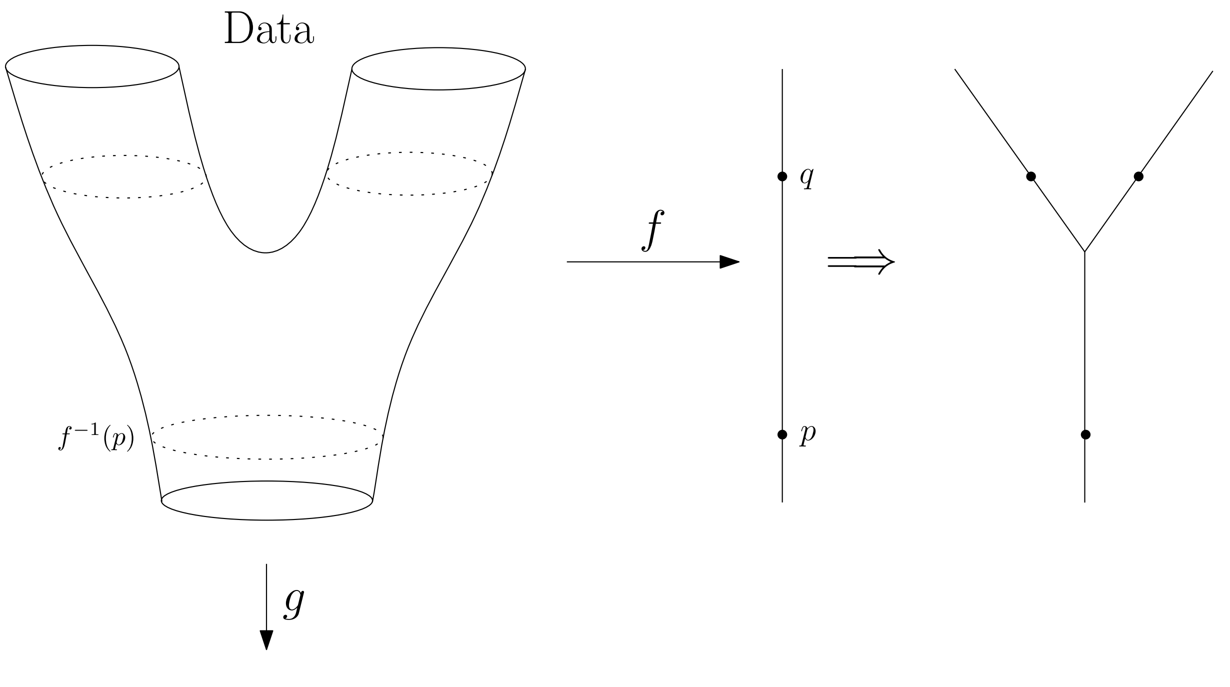
Math World



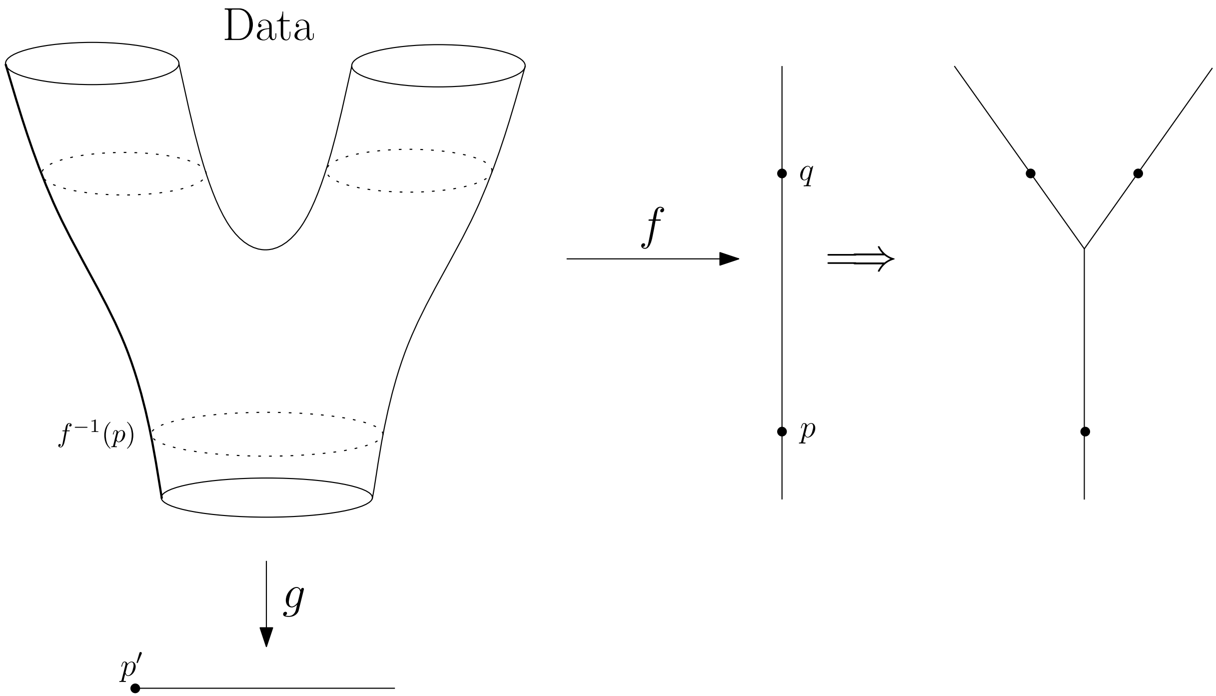
Math World



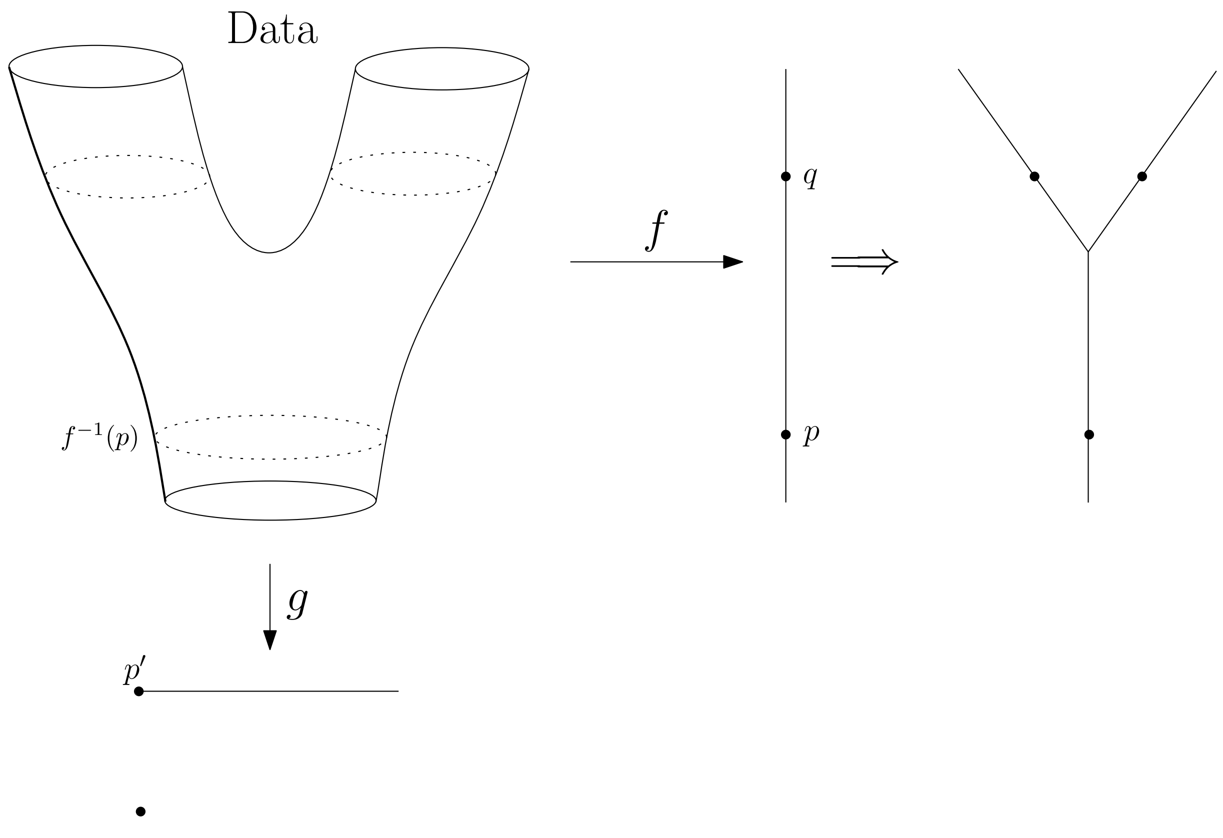
Math World



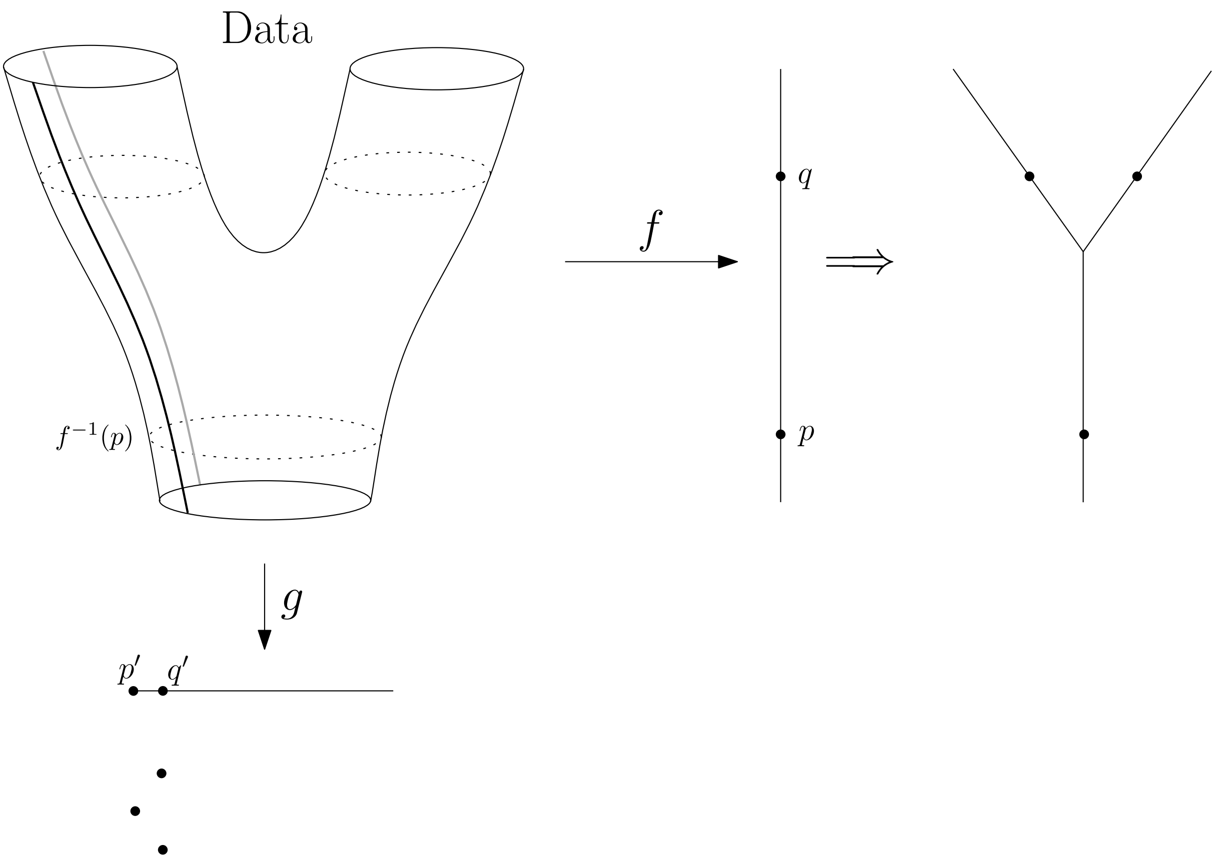
Math World



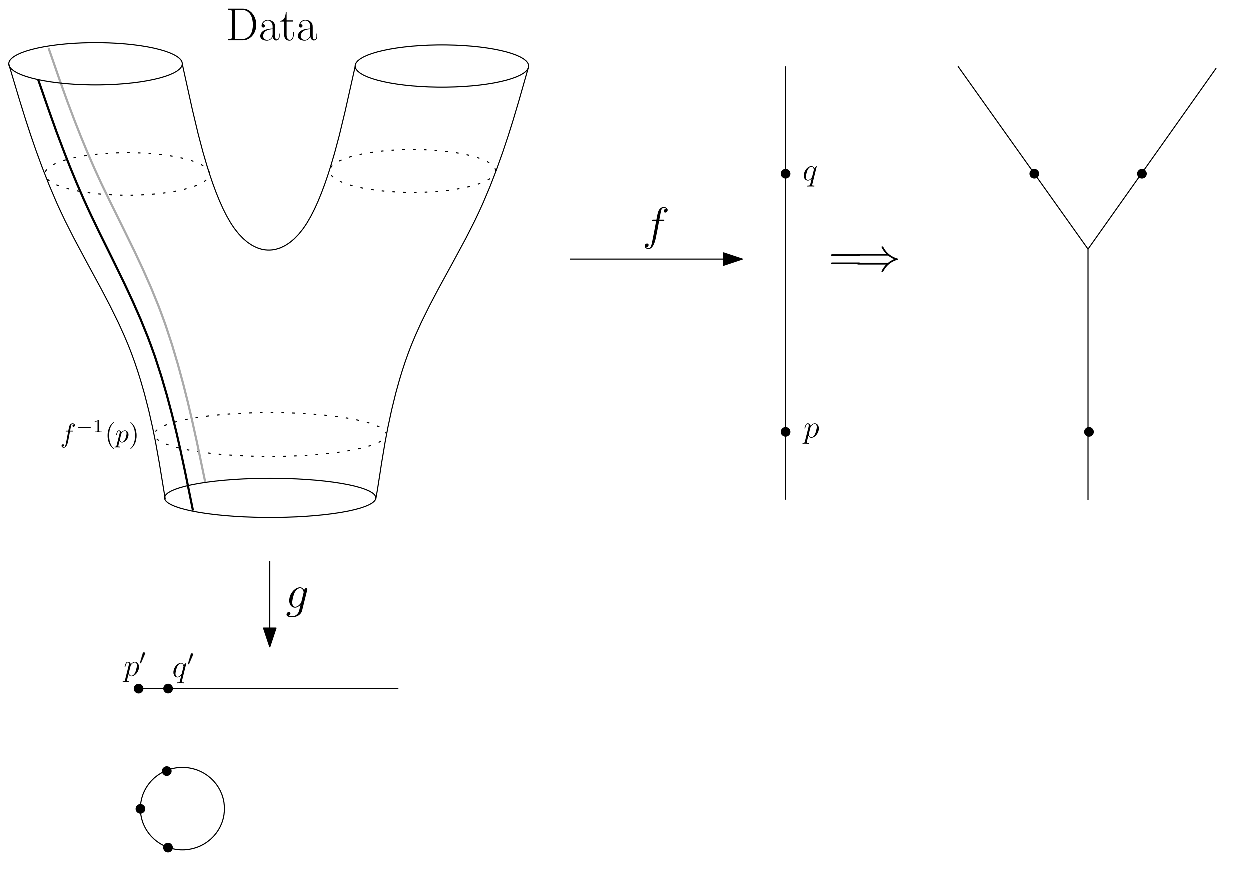
Math World



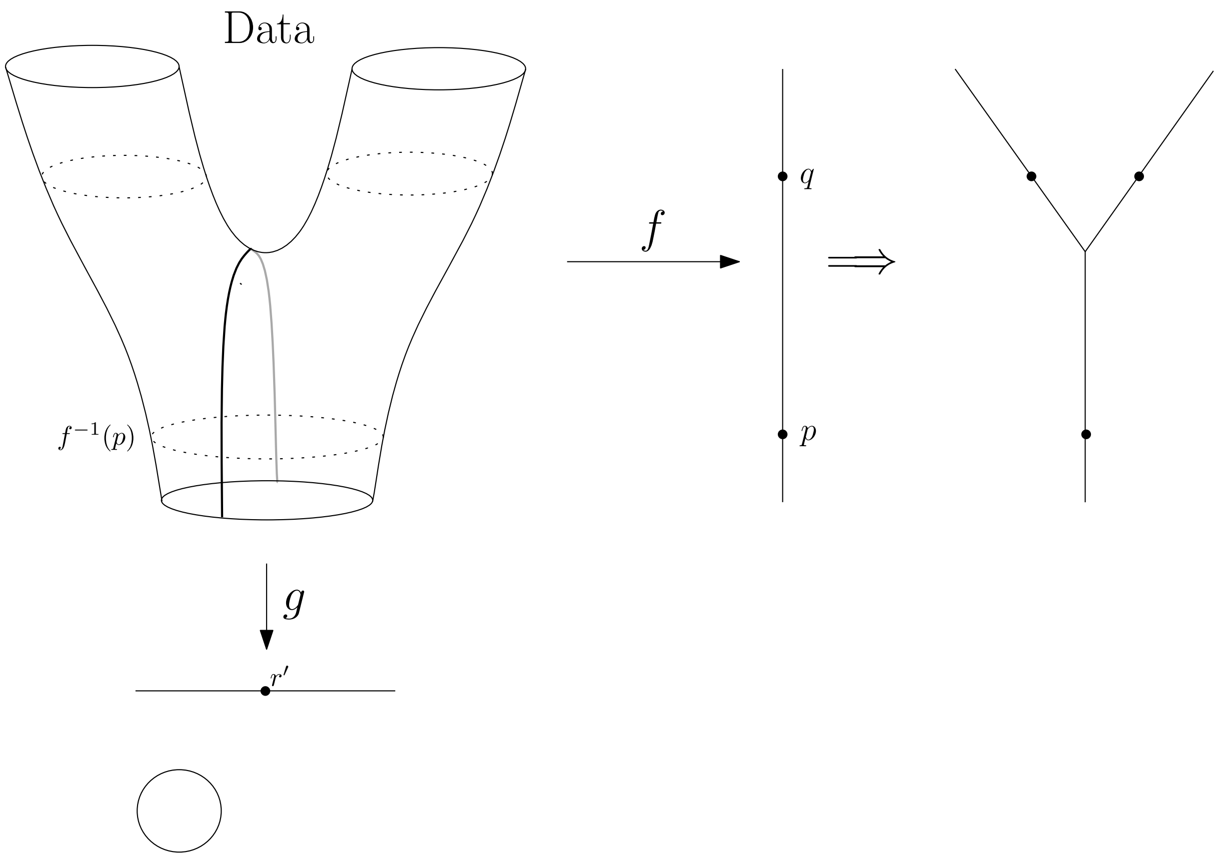
Math World



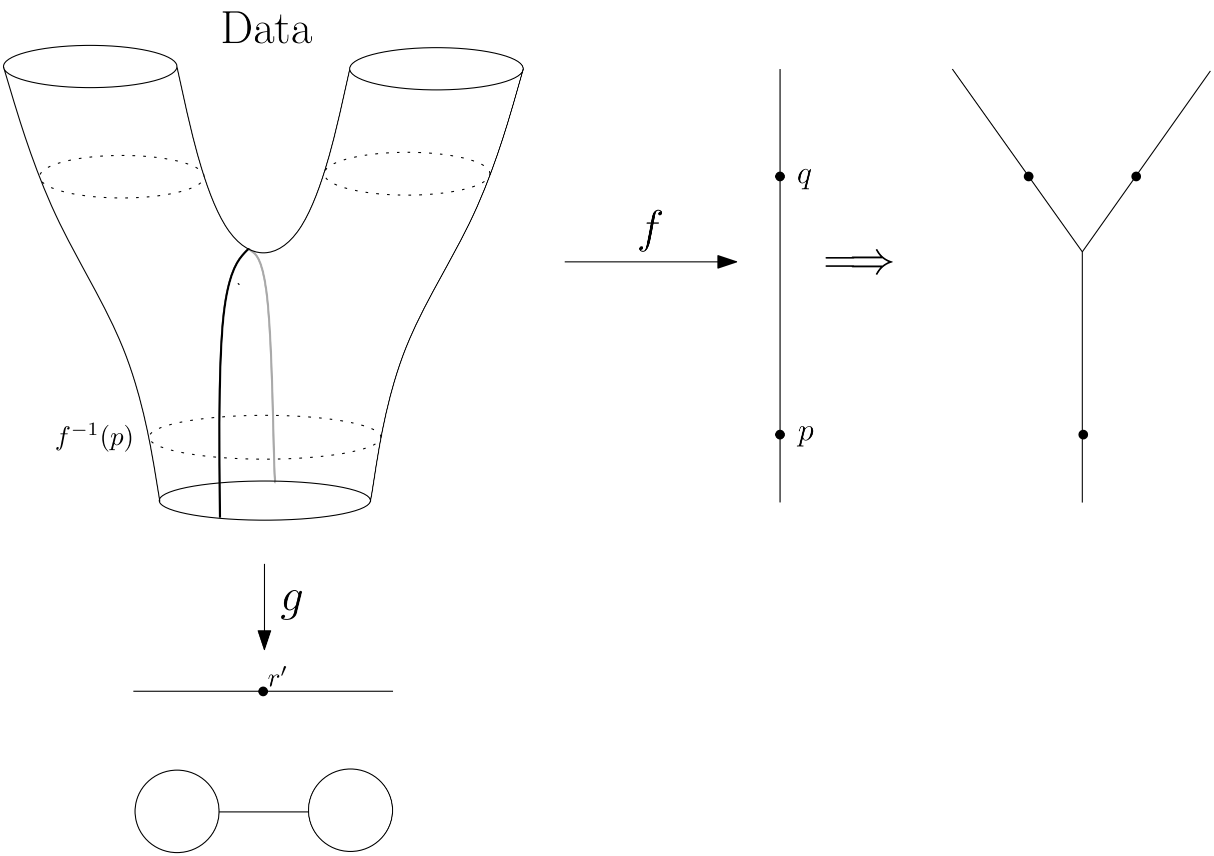
Math World



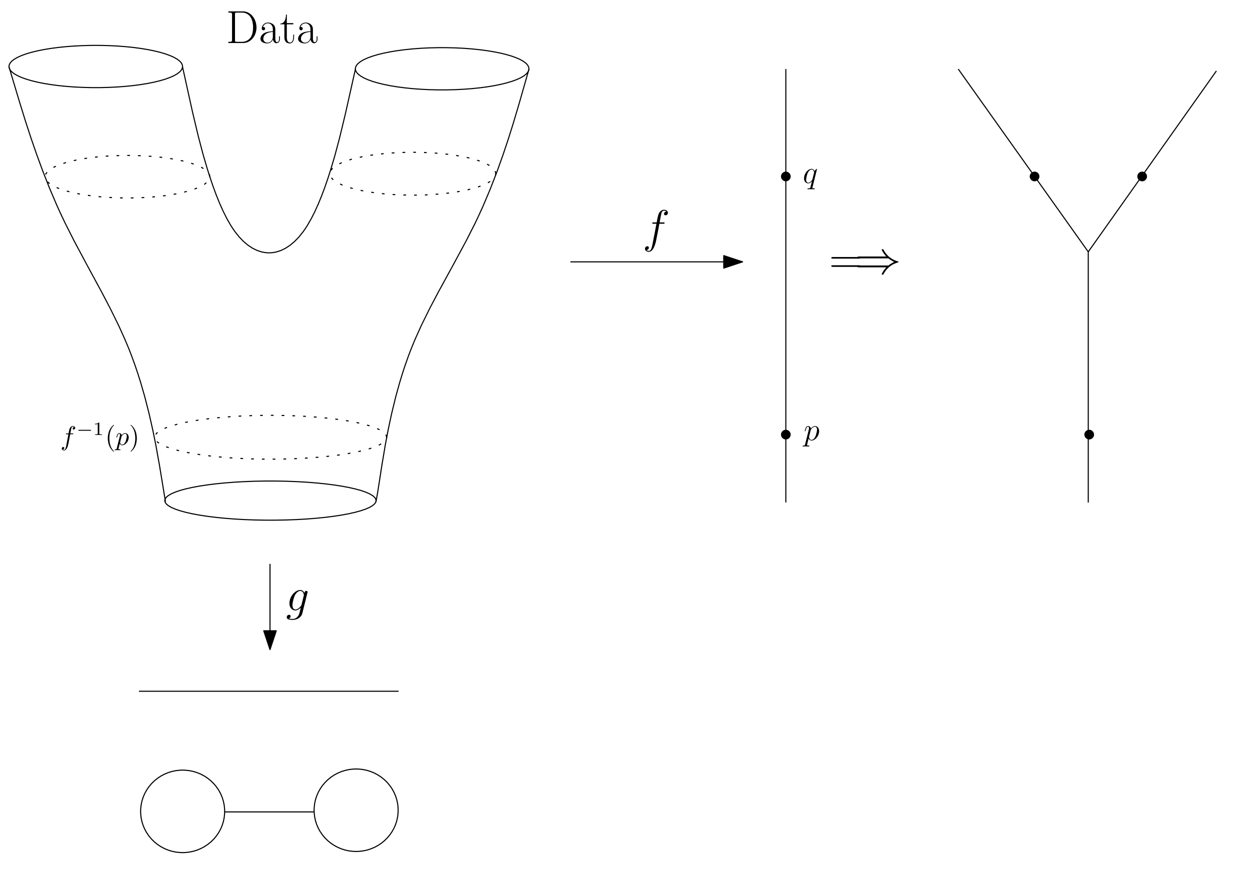
Math World



Math World



Math World

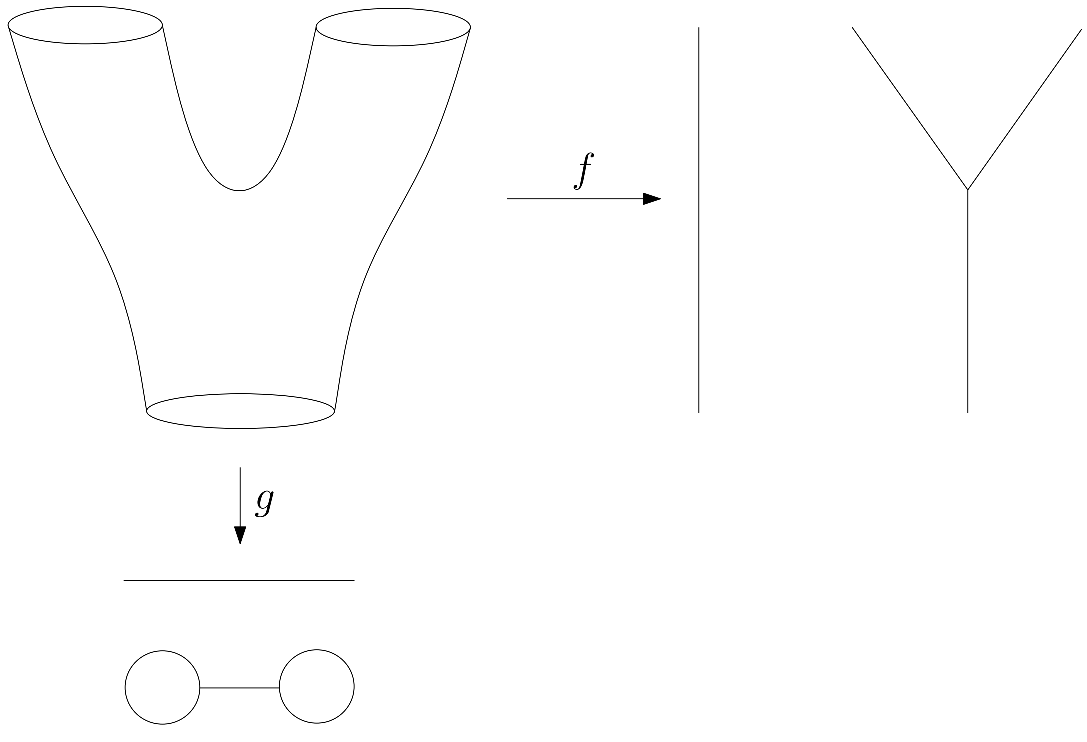


Why is this useful?

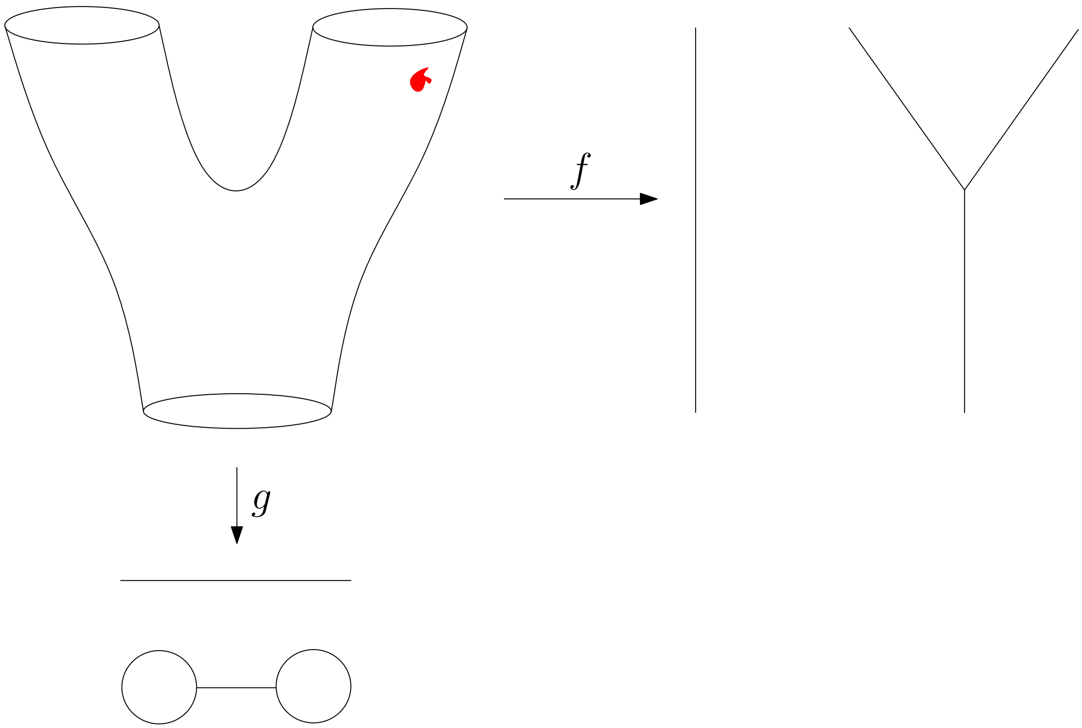
Why is this useful?

$\Rightarrow$ We get "easy" understanding of the localizations of quantities of interest.

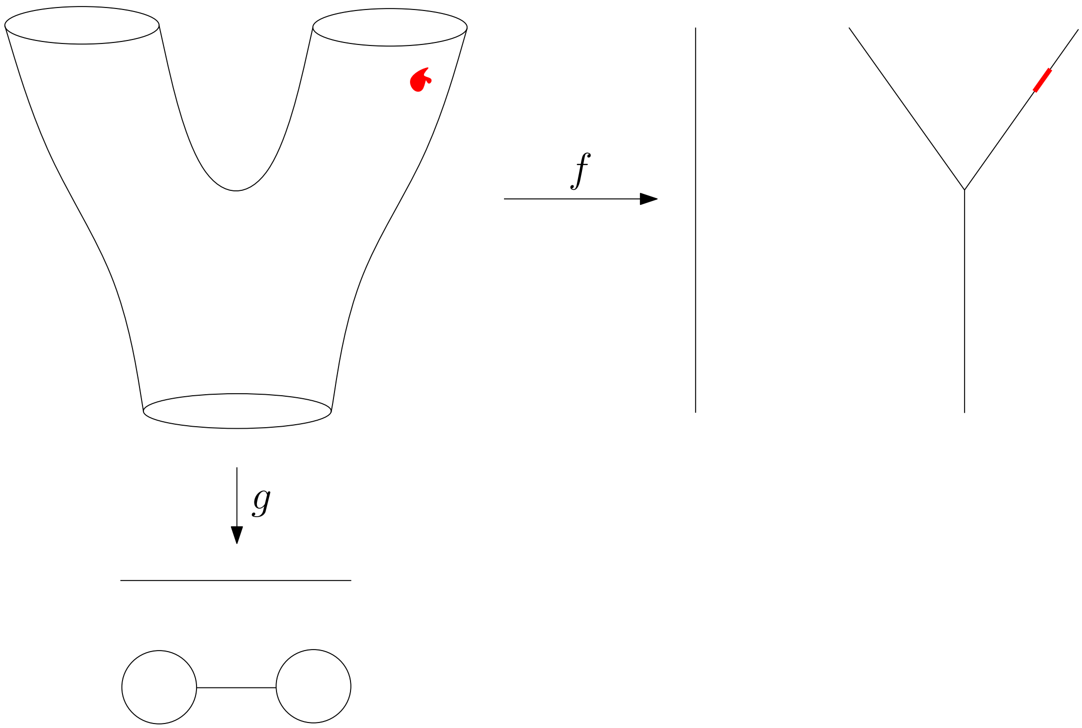
Why is this useful?



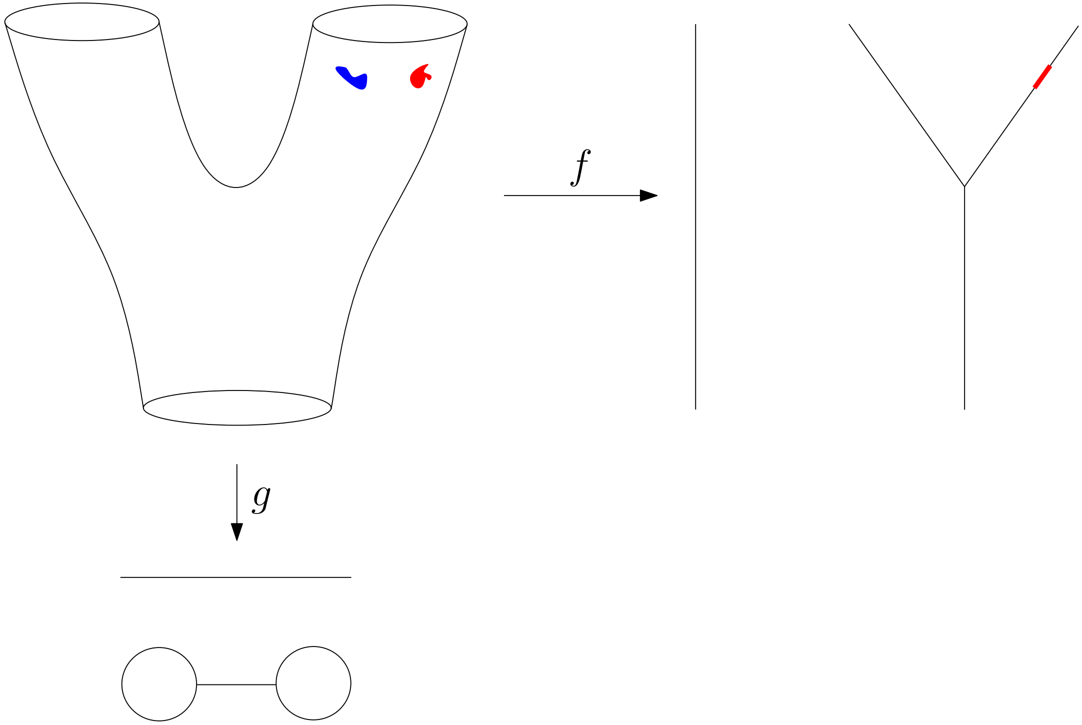
Why is this useful?



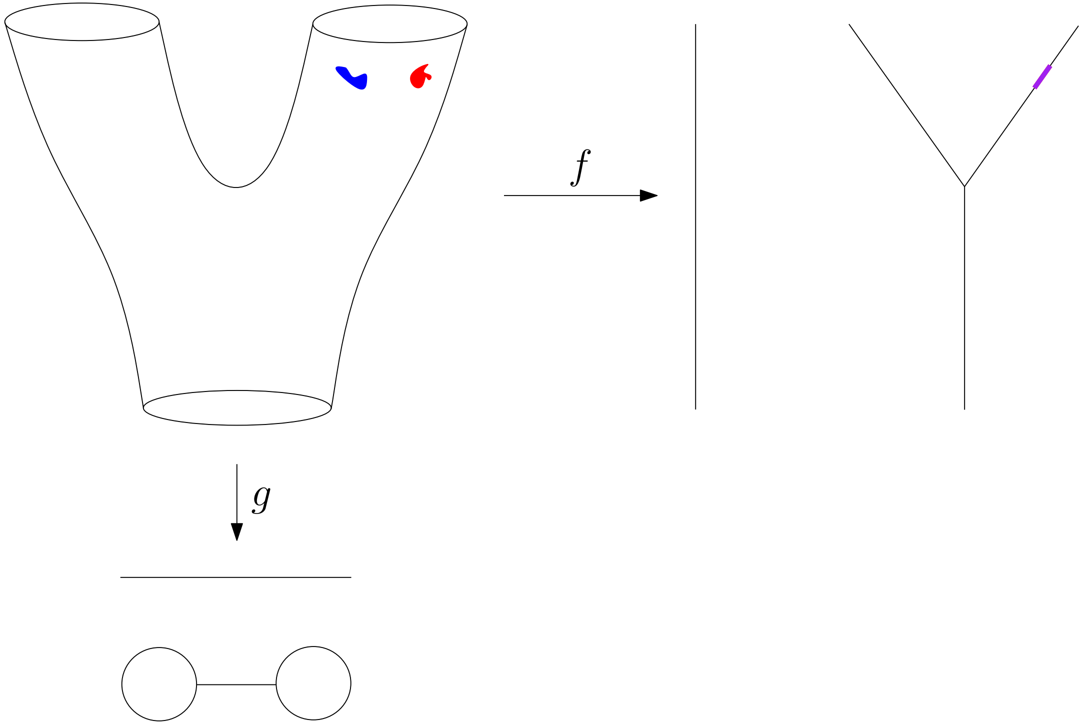
Why is this useful?



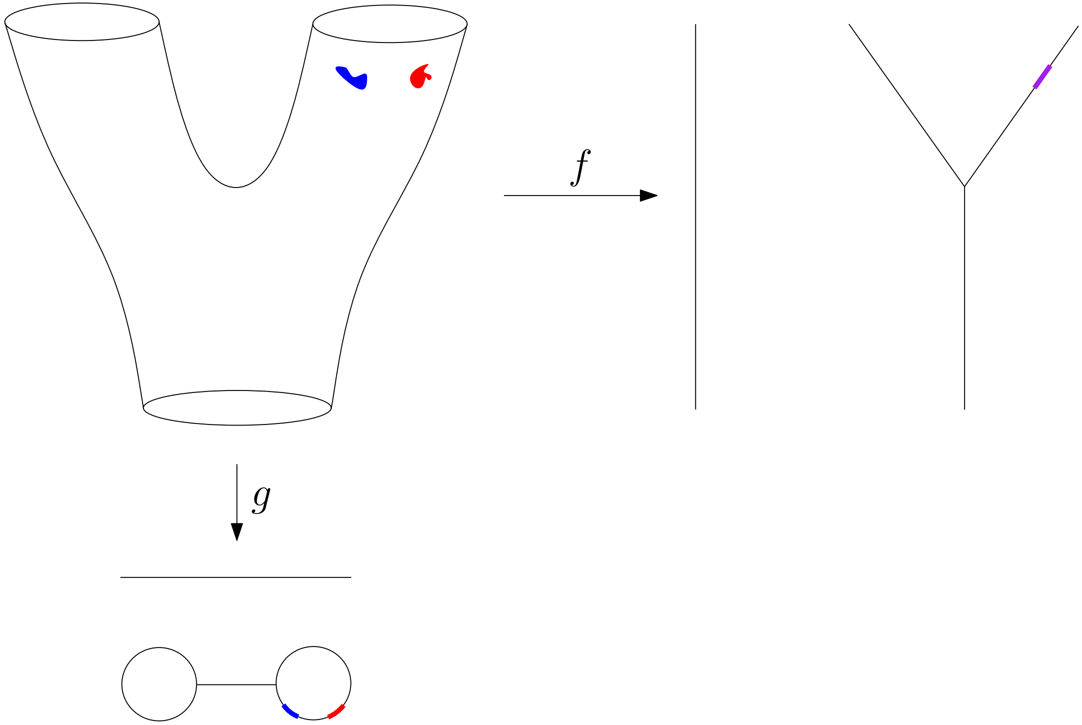
Why is this useful?



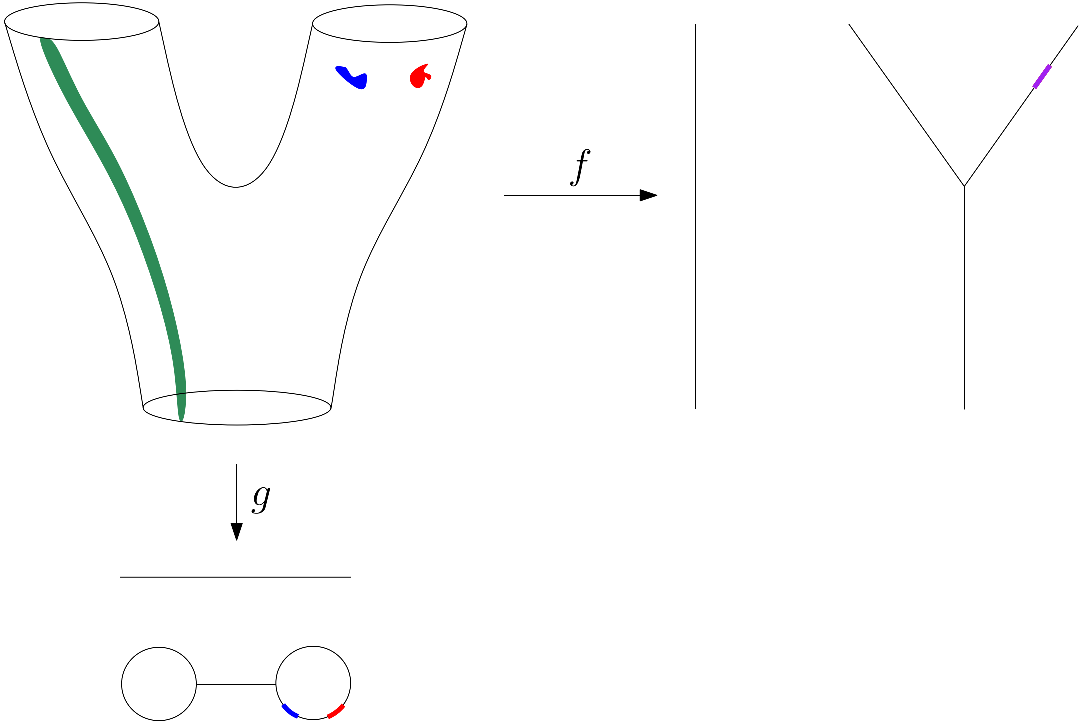
Why is this useful?



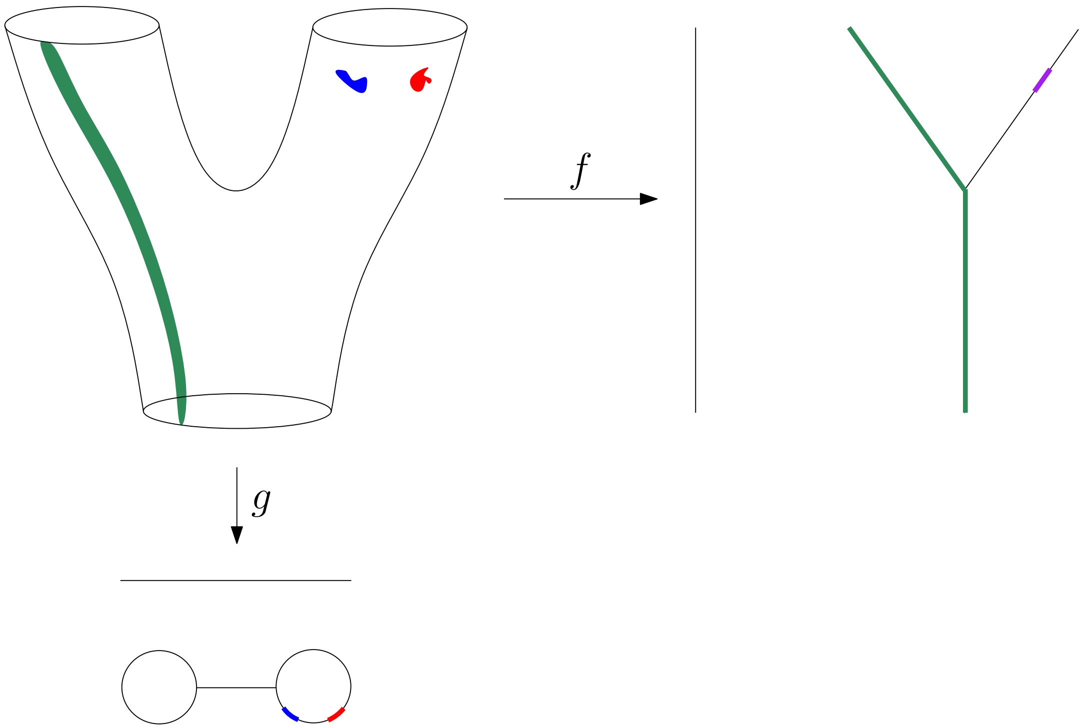
Why is this useful?



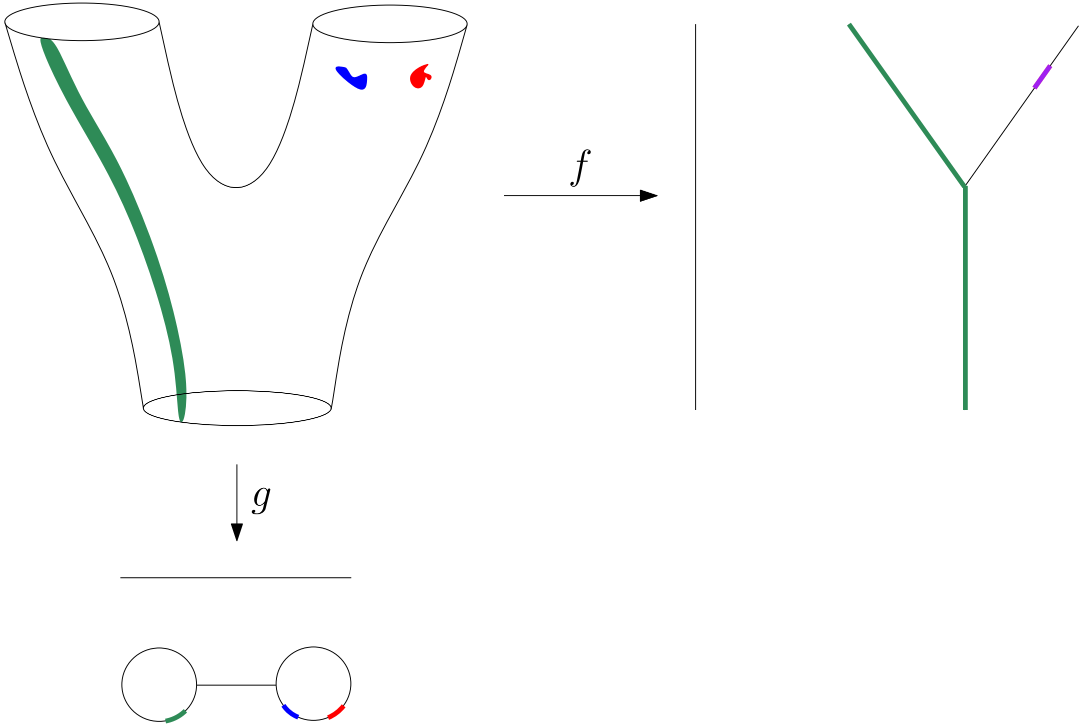
Why is this useful?



Why is this useful?



Why is this useful?



Why is this useful?

Why is this useful?

- Lenses inform us where in the space to look for phenomena.

Why is this useful?

- Lenses inform us where in the space to look for phenomena.
- For easy localizations many different lenses will be informative.

Why is this useful?

- Lenses inform us where in the space to look for phenomena.
- For easy localizations many different lenses will be informative.
- For hard ($=$ geometrically distributed) localizations we have to be more careful.

Why is this useful?

- Lenses inform us where in the space to look for phenomena.
- For easy localizations many different lenses will be informative.
- For hard ($=$ geometrically distributed) localizations we have to be more careful. But even then, we frequently get incremental knowledge even from a poorly chosen lens.

Step 2: Clustering as π_0

We need to adjust the "Math World view to bring the algorithm into "Data World".

Step 2: Clustering as π_0

We need to adjust the "Math World view to bring the algorithm into "Data World".

- We replace points with open sets in the range of the lens.

Step 2: Clustering as π_0

We need to adjust the "Math World view to bring the algorithm into "Data World".

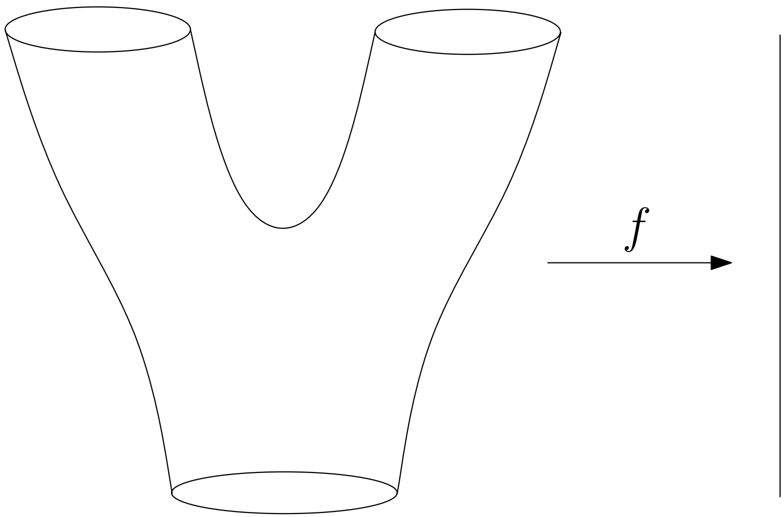
- We replace points with open sets in the range of the lens.
- We replace "connected component of the inverse image" is with "*clusters* in the inverse image".

Step 2: Clustering as π_0

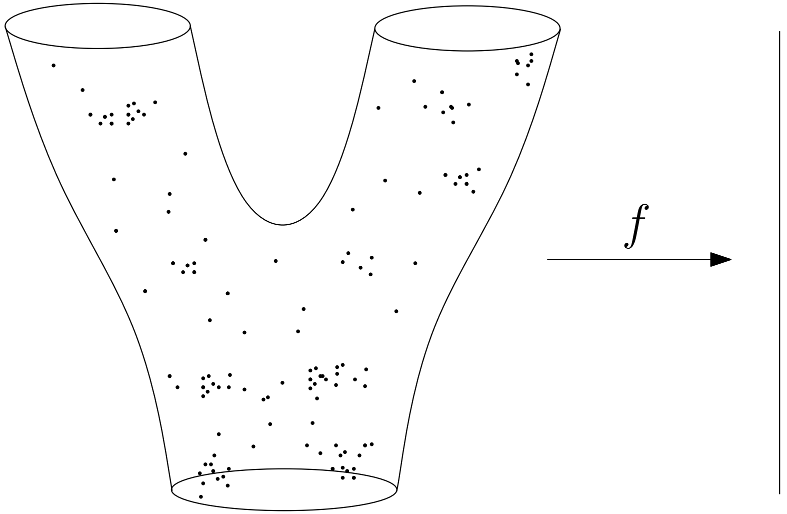
We need to adjust the "Math World view to bring the algorithm into "Data World".

- We replace points with open sets in the range of the lens.
- We replace "connected component of the inverse image" is with "*clusters* in the inverse image".
- We connect clusters (nodes) with an edge if they share points in common.

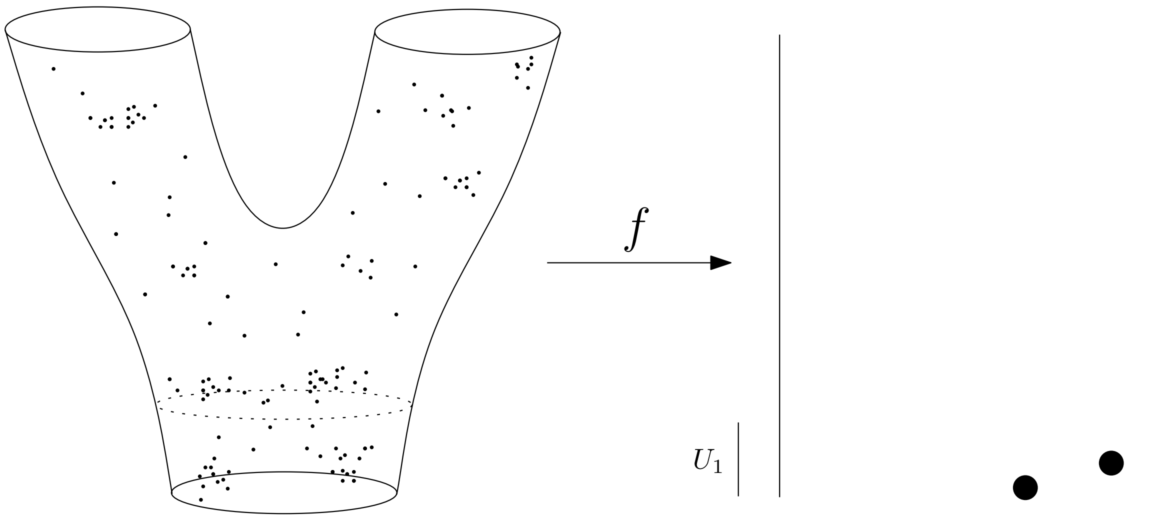
Step 2: Clustering as π_0



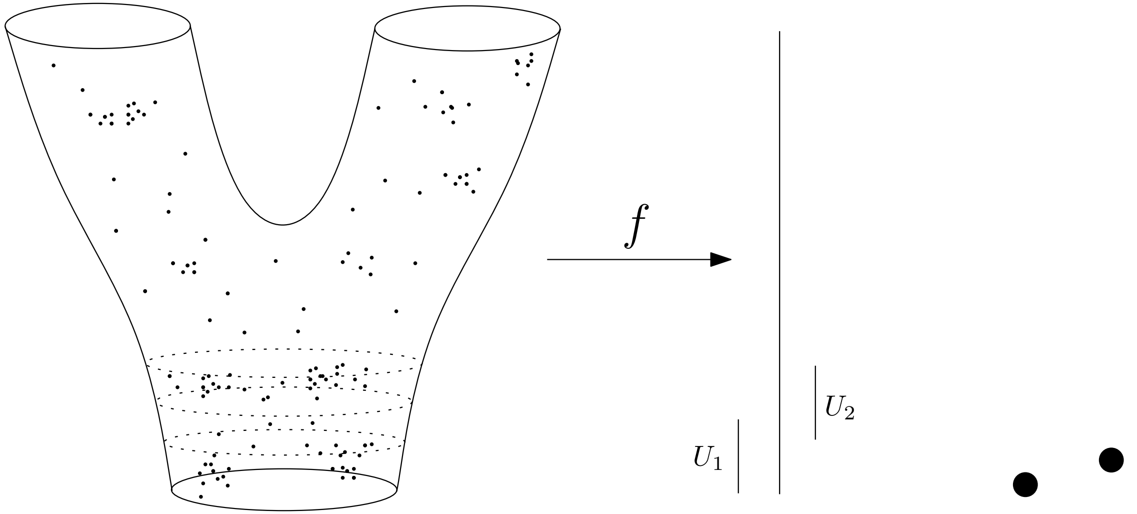
Step 2: Clustering as π_0



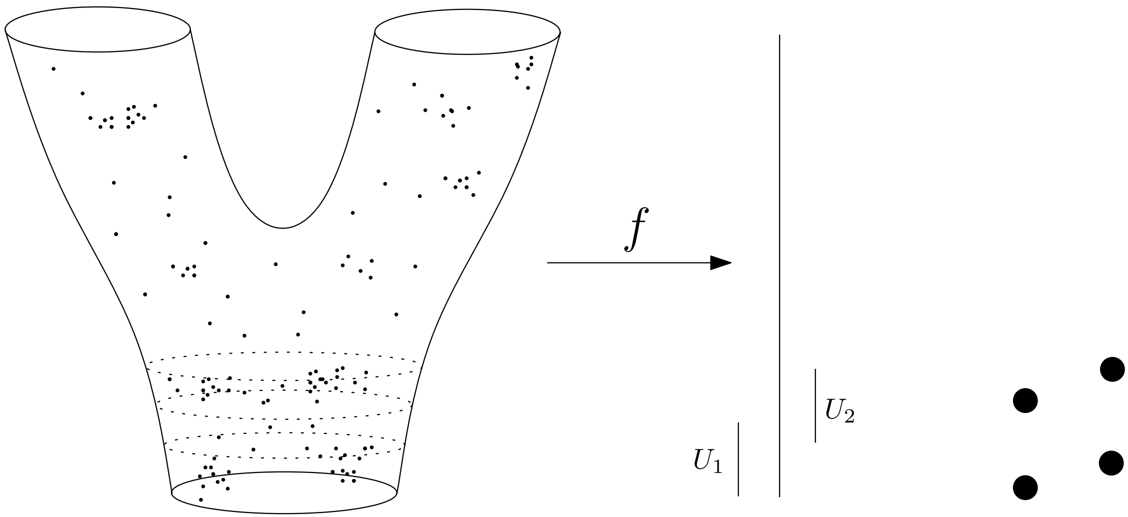
Step 2: Clustering as π_0



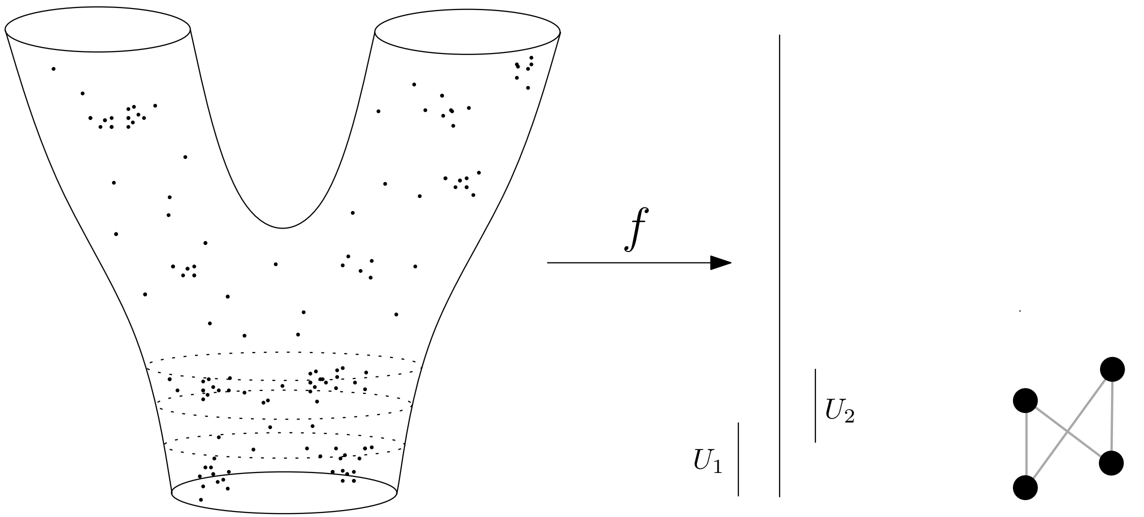
Step 2: Clustering as π_0



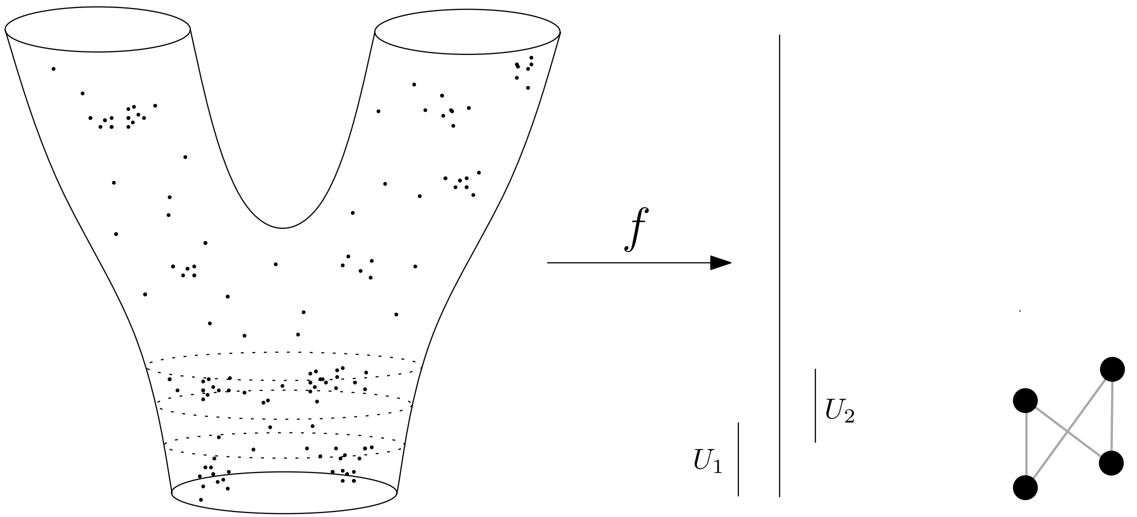
Step 2: Clustering as π_0



Step 2: Clustering as π_0

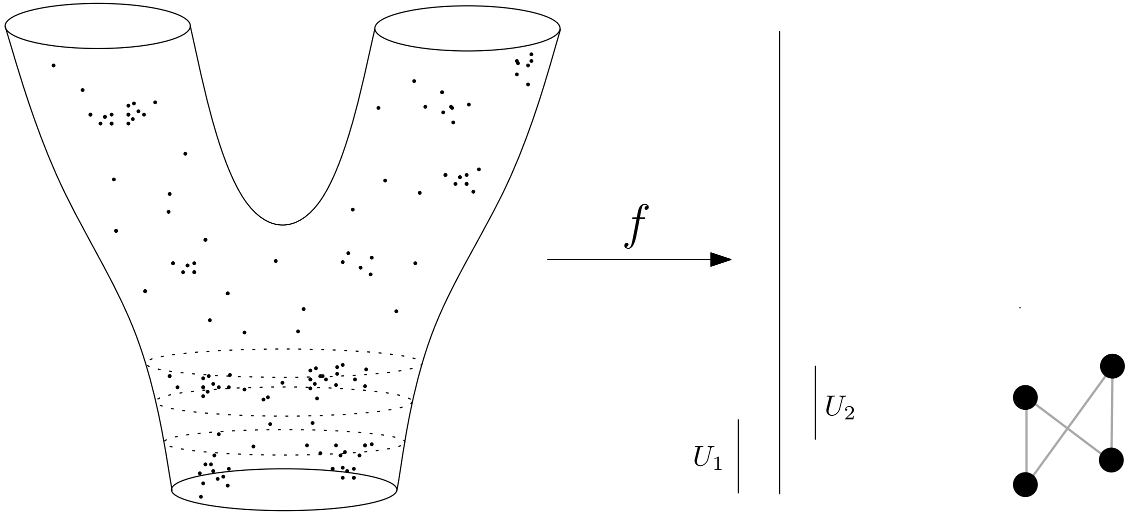


Step 2: Clustering as π_0



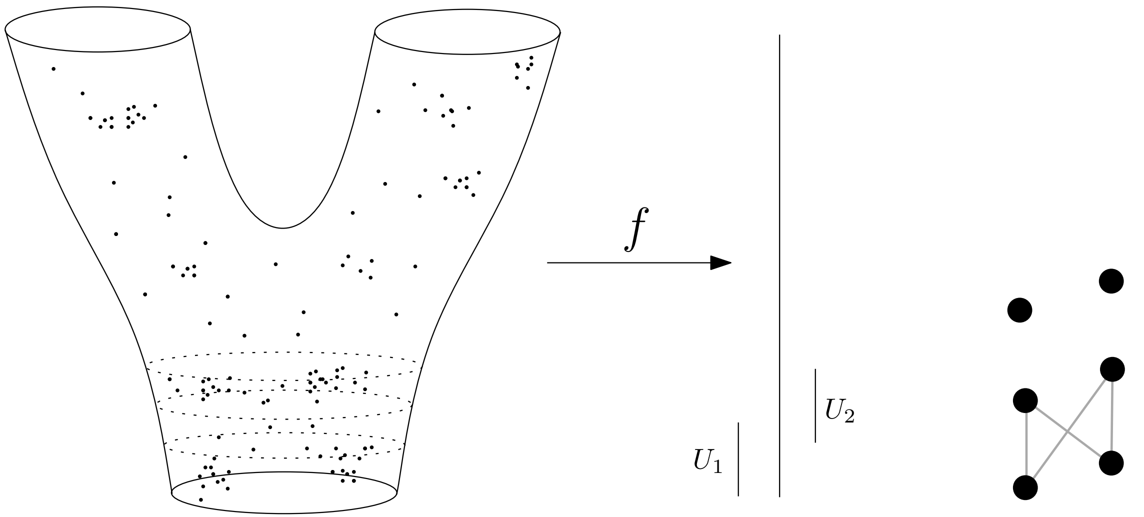
- Nodes are clusters of data points

Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

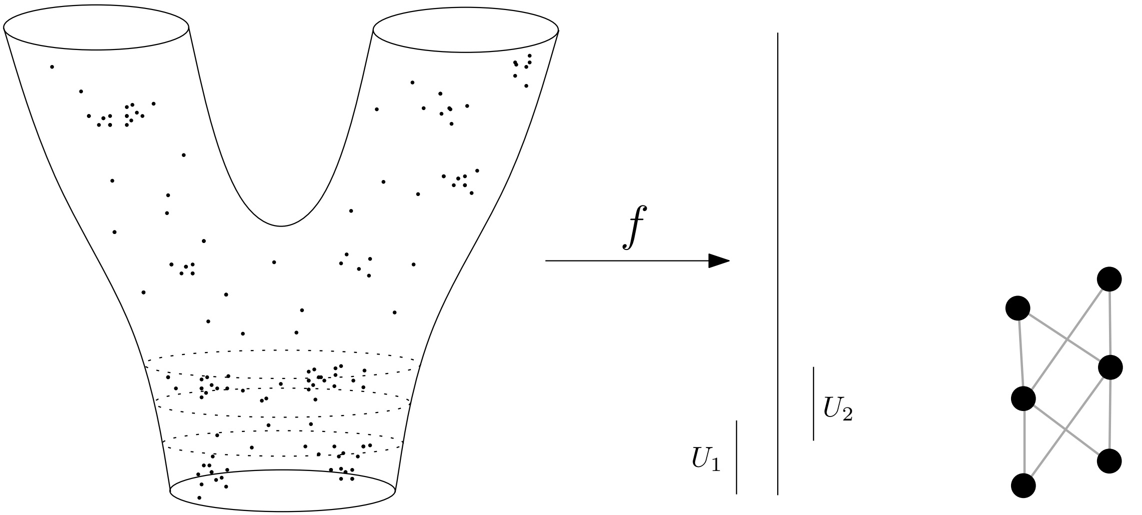
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

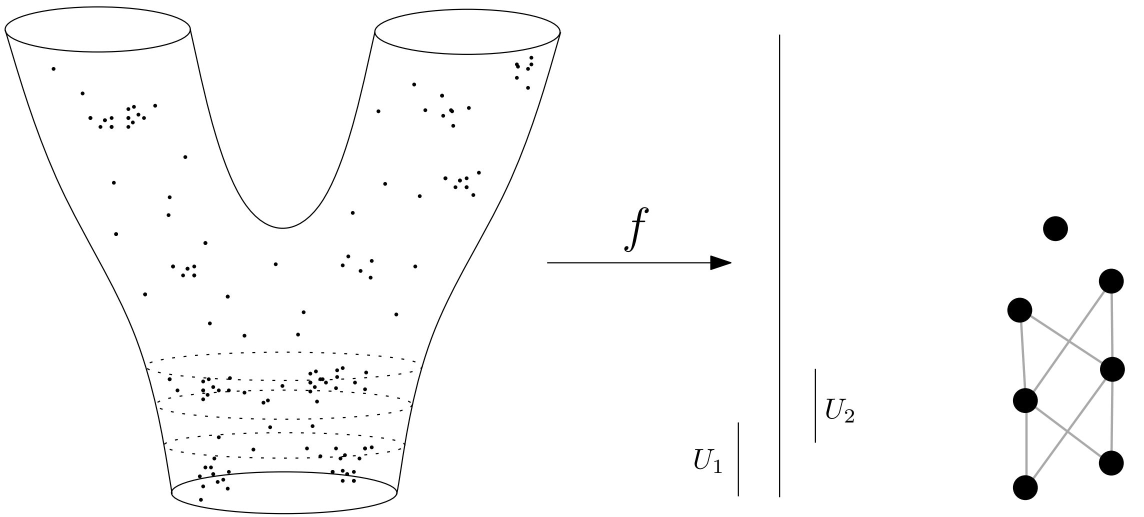
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

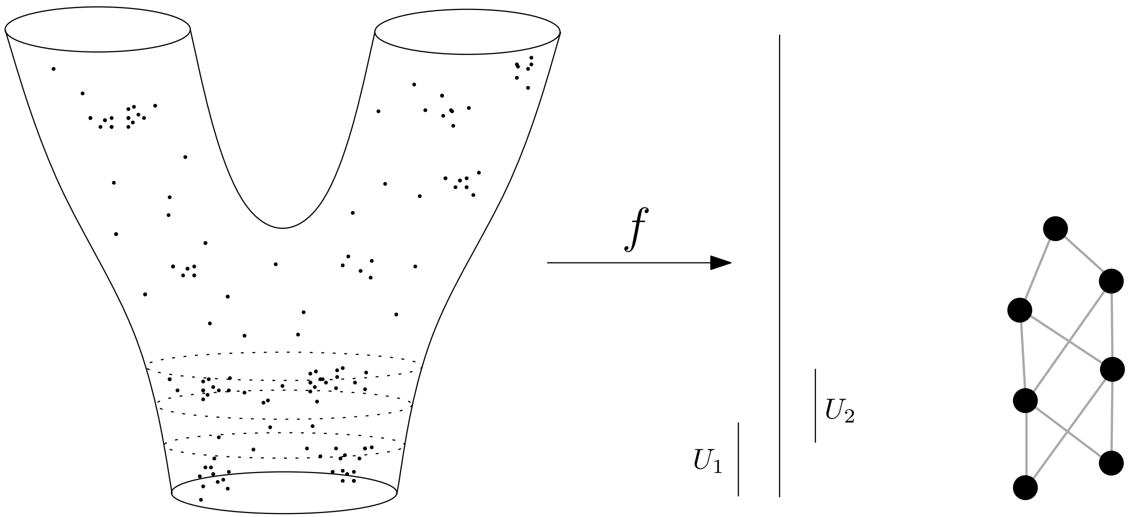
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

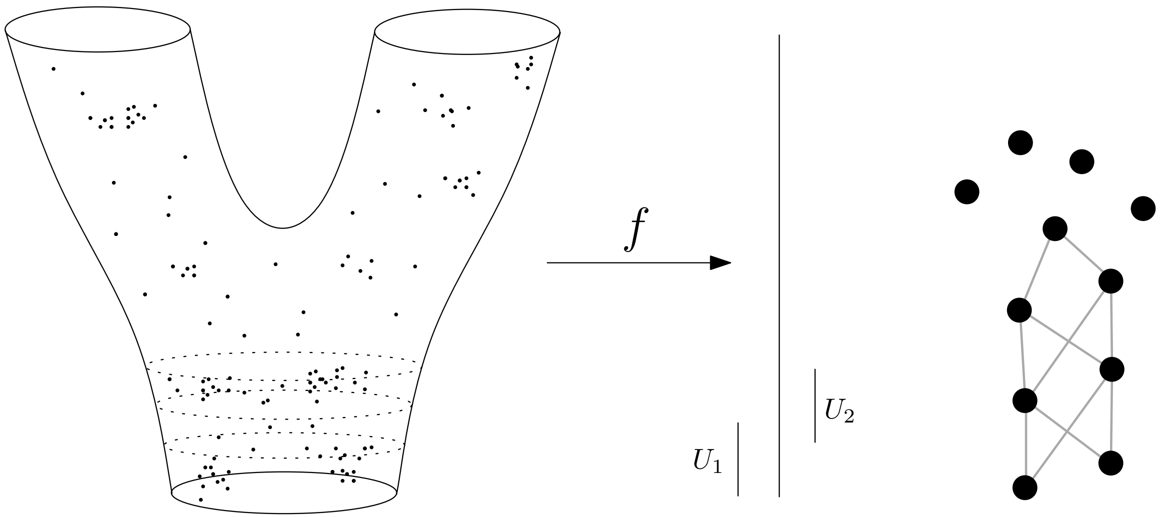
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

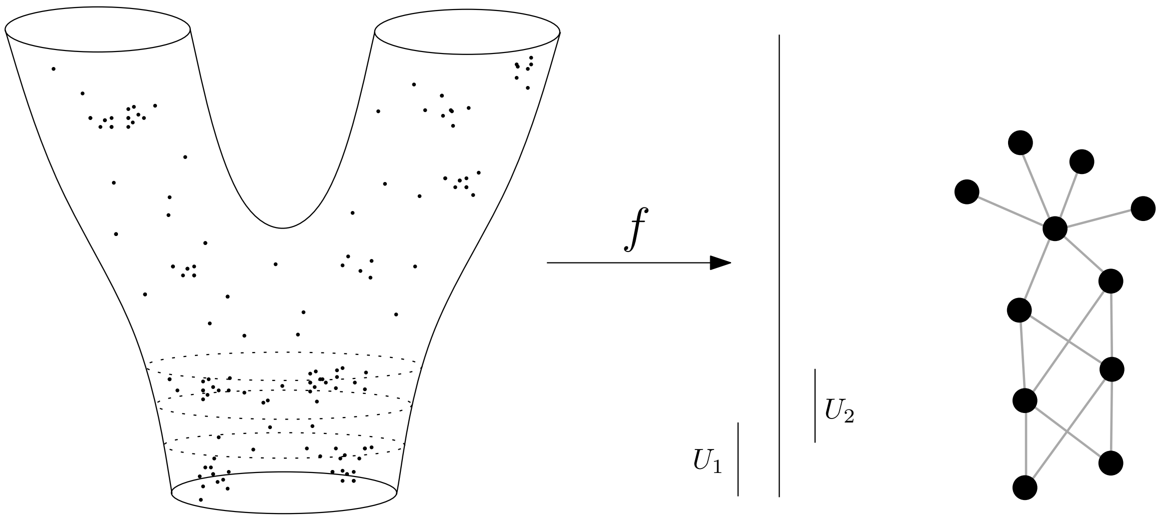
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

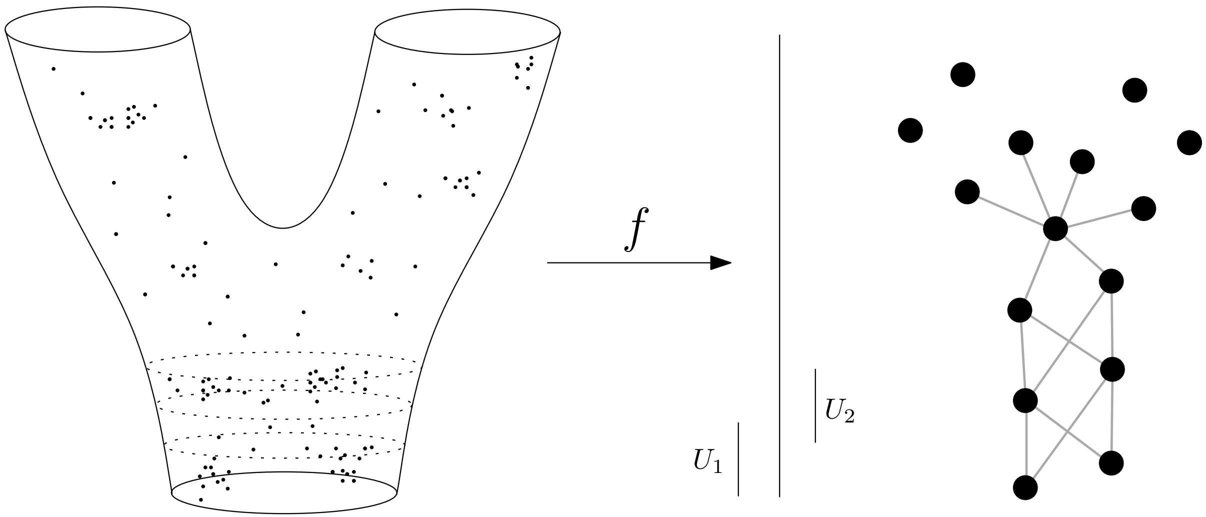
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

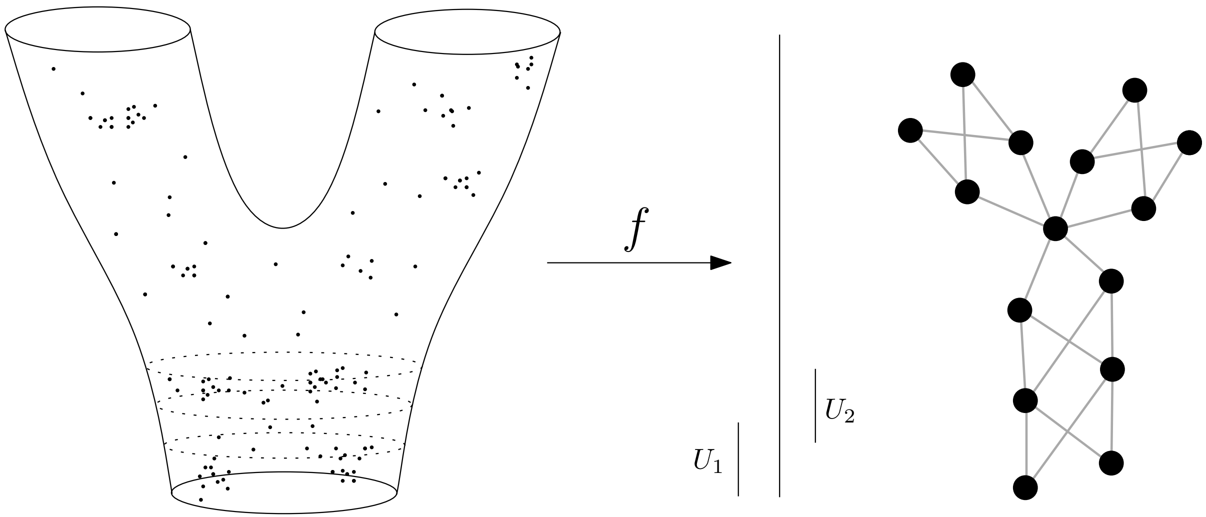
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

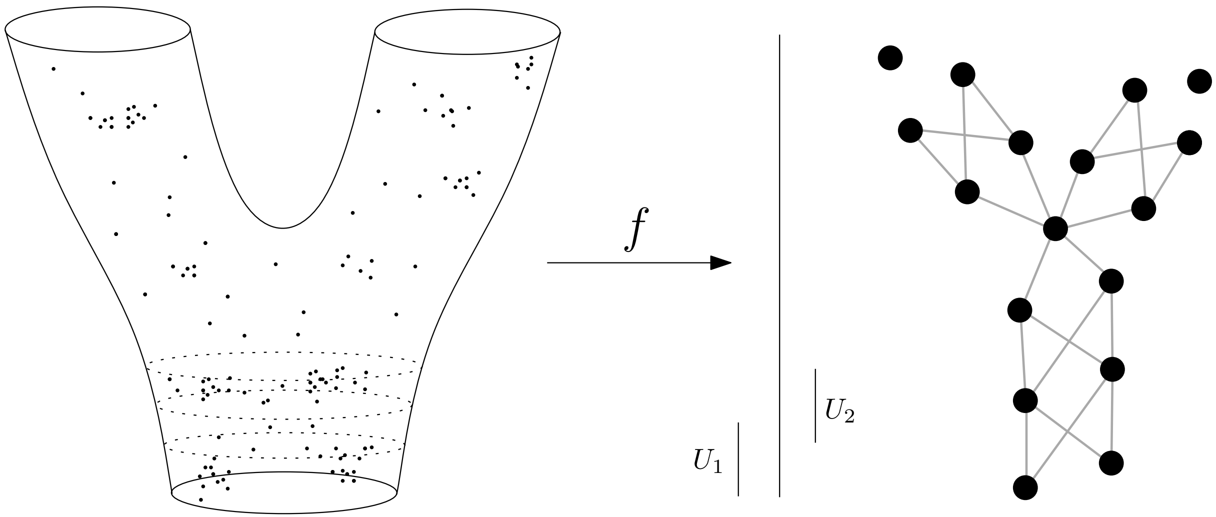
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

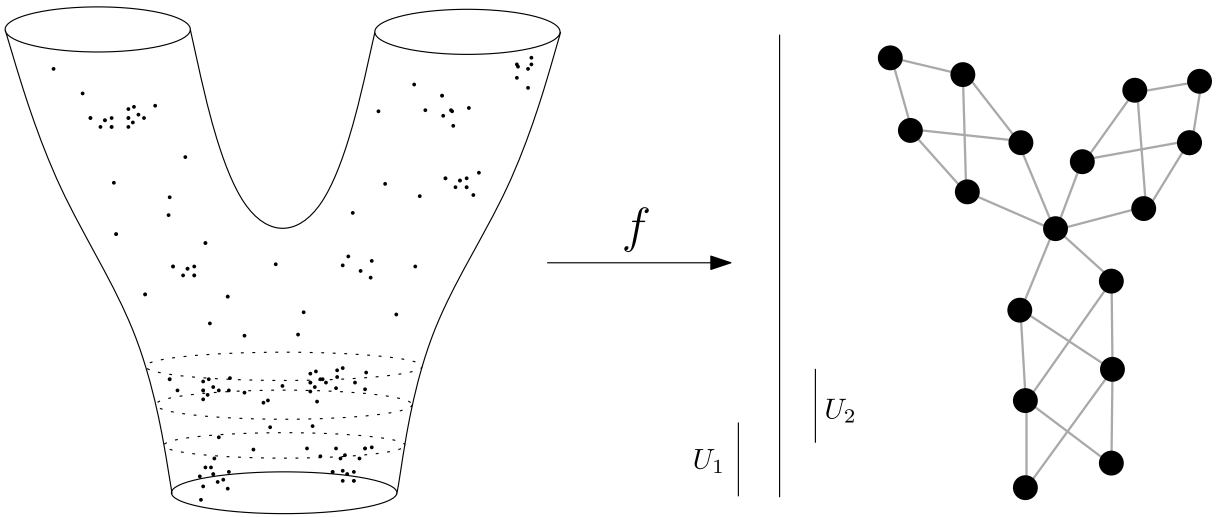
Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

This is also called taking the *nerve* of a covering where the lens+clustering makes the cover.

Lenses: Where do they come from

The technique rests on finding good lenses.

Lenses: Where do they come from

The technique rests on finding good lenses.

$\Rightarrow$ Luckily lots of people have worked on this problem

Lenses: Where do they come from

A Non Exhaustive Table of Lenses

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics
Mean/Max/Min

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics
Mean/Max/Min
Variance

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics
Mean/Max/Min
Variance
n-Moment

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics
Mean/Max/Min
Variance
n-Moment
Density

Lenses: Where do they come from

- Standard data analysis functions

A Non Exhaustive Table of Lenses

Statistics
Mean/Max/Min
Variance
n-Moment
Density
...

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology

A Non Exhaustive Table of Lenses

Statistics	Geometry
Mean/Max/Min	
Variance	
n-Moment	
Density	
...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology

A Non Exhaustive Table of Lenses

Statistics	Geometry
Mean/Max/Min	Centrality
Variance	
n-Moment	
Density	
...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology

A Non Exhaustive Table of Lenses

Statistics	Geometry
Mean/Max/Min	Centrality
Variance	Curvature
n-Moment	
Density	
...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology

A Non Exhaustive Table of Lenses

Statistics	Geometry
Mean/Max/Min	Centrality
Variance	Curvature
n-Moment	Harmonic Cycles
Density	
...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology

A Non Exhaustive Table of Lenses

Statistics	Geometry
Mean/Max/Min	Centrality
Variance	Curvature
n-Moment	Harmonic Cycles
Density	...
...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	
Variance	Curvature	
n-Moment	Harmonic Cycles	
Density	...	
...		

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	
n-Moment	Harmonic Cycles	
Density	...	
...		

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	Autoencoders
n-Moment	Harmonic Cycles	
Density	...	
...		

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	Autoencoders
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE
Density	...	
...		

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	Autoencoders
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE
Density	...	SVM Distance from Hyperplane
...		

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	Autoencoders
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE
Density	...	SVM Distance from Hyperplane
...		Error/Debugging Info

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning
Mean/Max/Min	Centrality	PCA/SVD
Variance	Curvature	Autoencoders
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE
Density	...	SVM Distance from Hyperplane
...		Error/Debugging Info
		...

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics
- Domain Knowledge / Data Modeling

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning	Data Driven
Mean/Max/Min	Centrality	PCA/SVD	
Variance	Curvature	Autoencoders	
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE	
Density	...	SVM Distance from Hyperplane	
...		Error/Debugging Info	
		...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics
- Domain Knowledge / Data Modeling

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning	Data Driven
Mean/Max/Min	Centrality	PCA/SVD	Age
Variance	Curvature	Autoencoders	
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE	
Density	...	SVM Distance from Hyperplane	
...		Error/Debugging Info	
		...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics
- Domain Knowledge / Data Modeling

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning	Data Driven
Mean/Max/Min	Centrality	PCA/SVD	Age
Variance	Curvature	Autoencoders	Dates
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE	
Density	...	SVM Distance from Hyperplane	
...		Error/Debugging Info	
		...	

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics
- Domain Knowledge / Data Modeling

A Non Exhaustive Table of Lenses

Statistics	Geometry	Machine Learning	Data Driven
Mean/Max/Min	Centrality	PCA/SVD	Age
Variance	Curvature	Autoencoders	Dates
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE	...
Density	...	SVM Distance from Hyperplane	
...		Error/Debugging Info	
		...	

Why use TDA?

Basic Example: Higher Fidelity PCA

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.

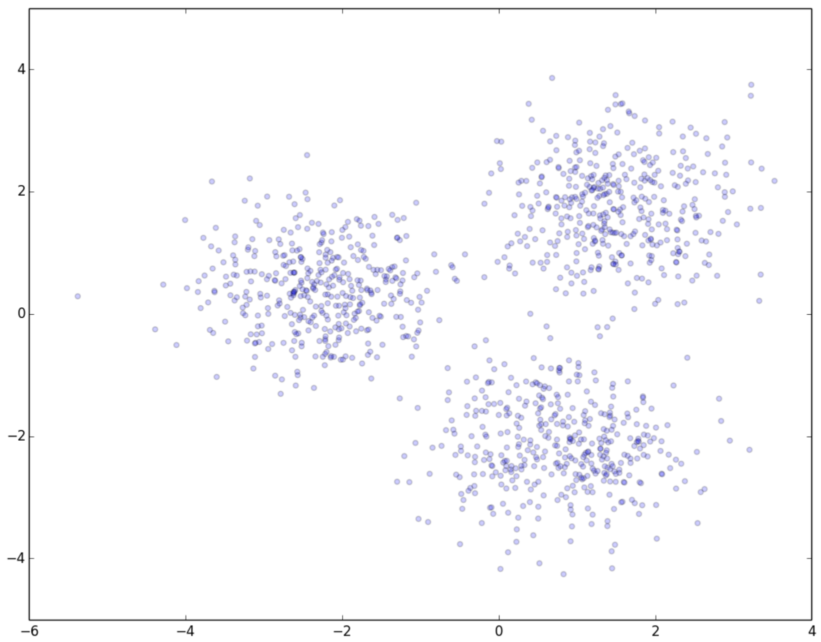
Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

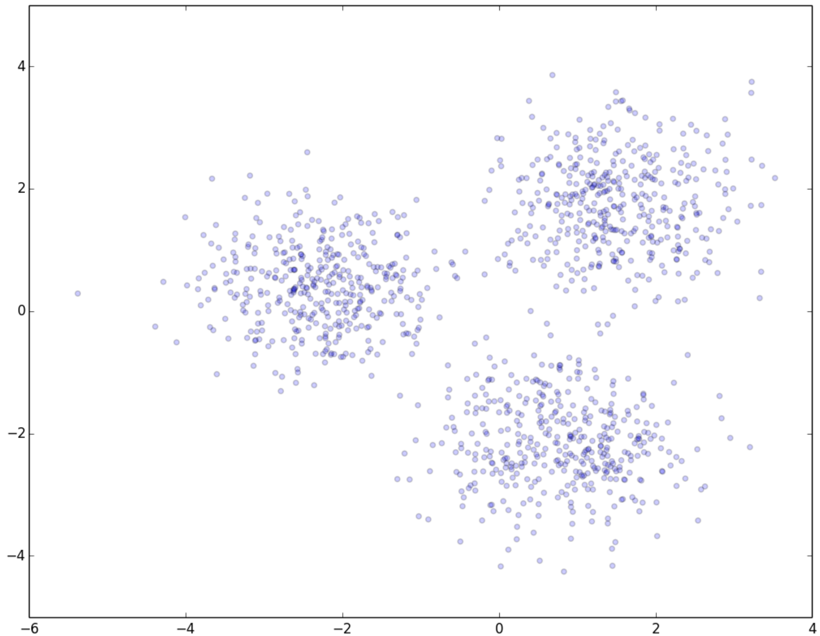
Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Unsupervised Learning: PCA



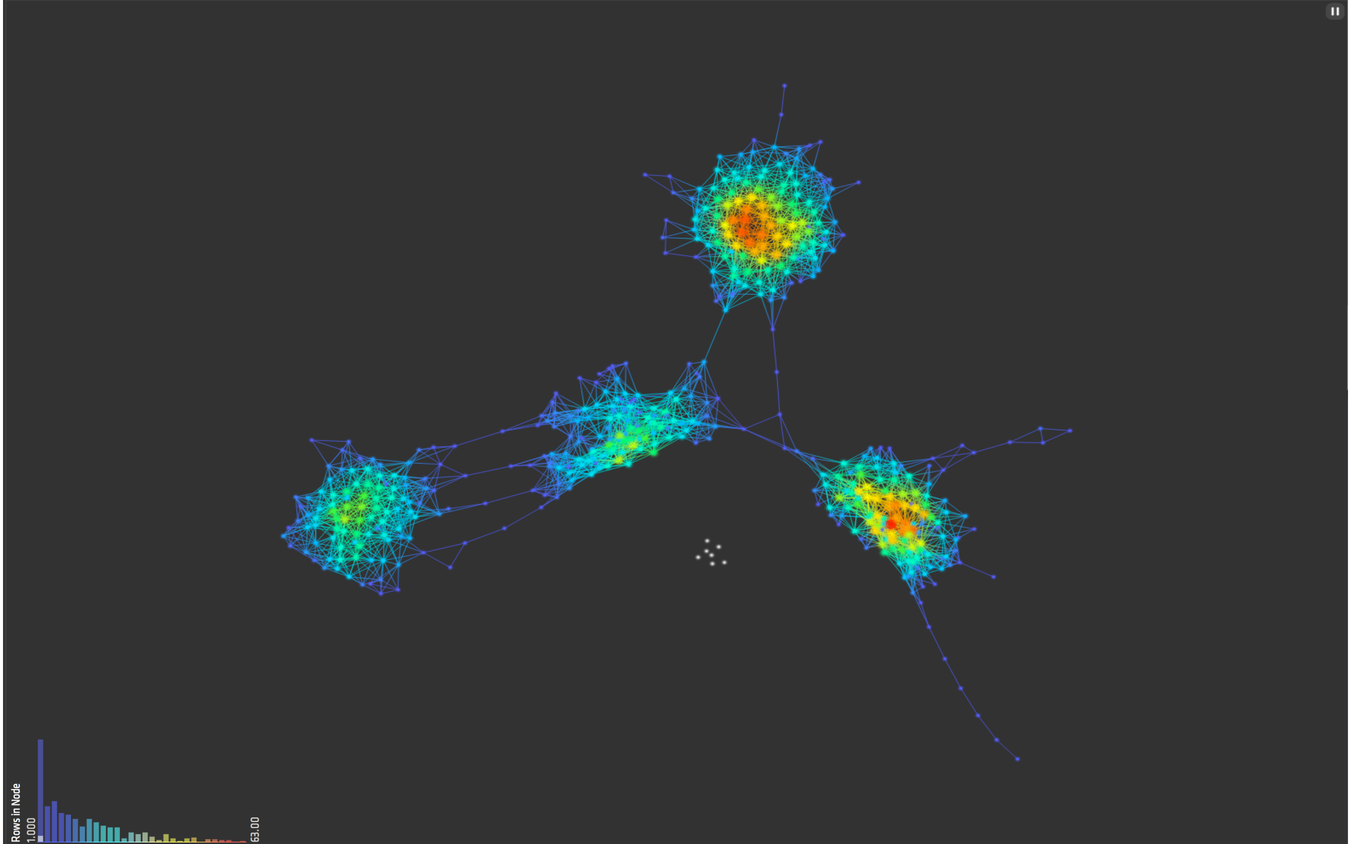
Unsupervised Learning: PCA



PCA captured 98.4% of the variance

Unsupervised Learning: PCA

AYASDI



Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Problems

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Problems

- Clusters are distinct but are projected on top of each other.

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Problems

- Clusters are distinct but are projected on top of each other.
- Points are falsely clustered because of the projection.

Unsupervised Learning: PCA

PCA is roughly speaking orthogonal projection onto the plane that best contains the data.

Advantages:

- Provides unsupervised dimensionality reduction.
- Easy to interpret: Finds the best linear subspace that captures the variance or spread of the data.

Problems

- Clusters are distinct but are projected on top of each other.
- Points are falsely clustered because of the projection.

As a *framework* for data analysis we get **higher fidelity** from existing tools.

Real Examples

Supervised Learning: Model Introspection

We can use TDA to examine what is happening with our machine learning models.

Model Introspection: Outliers

Data: Customer attributes. Service usages, contractual details.

Problem: Customers commit fraud. Find customers with abnormal costs.

Proposed Solution: Create an ensemble of cost outlier models. Use these to flag customers as being fraudulent.

Model Introspection: Outliers

TDA Introspection:

Model Introspection: Outliers

TDA Introspection:

- Create a dataset that contains all non-cost information.

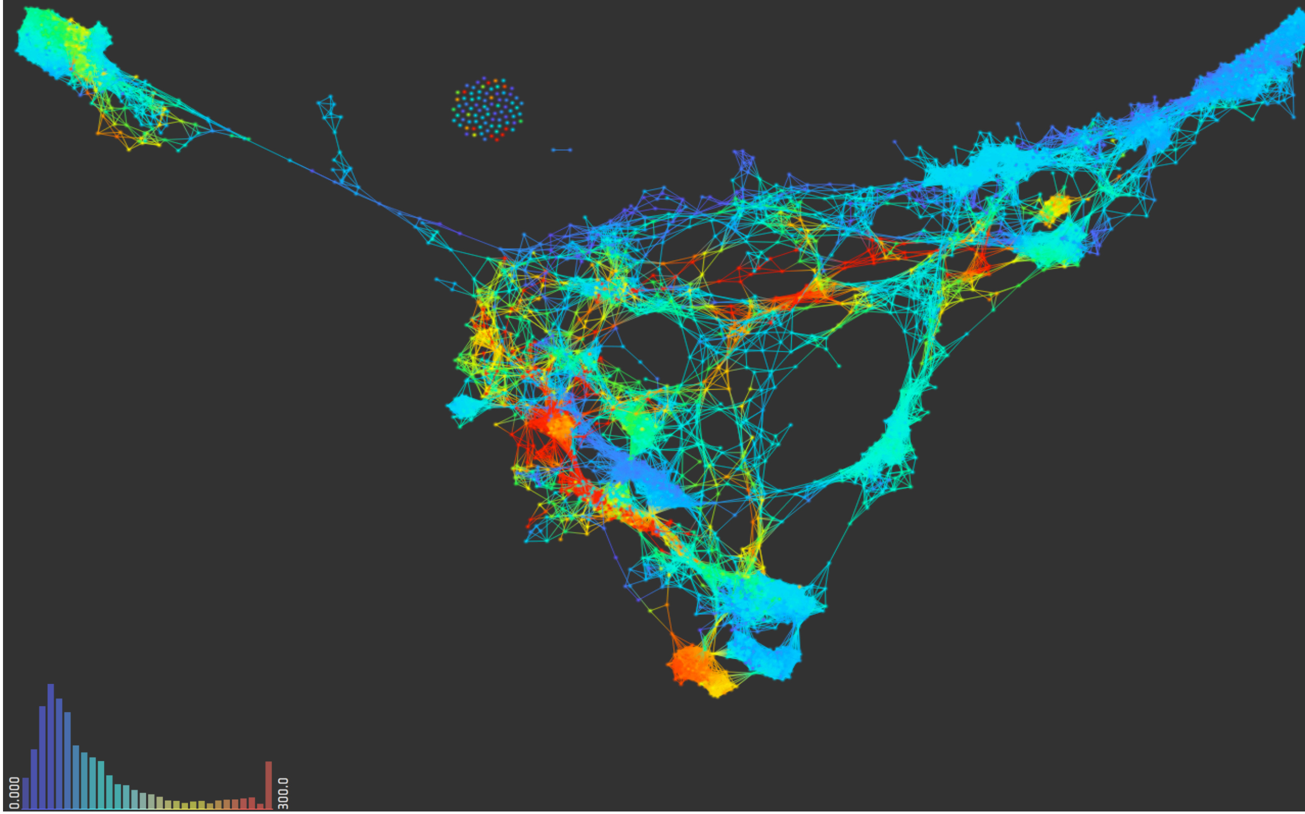
Model Introspection: Outliers

TDA Introspection:

- Create a dataset that contains all non-cost information.
- Color by who is being flagged by the ensemble as being a (high) cost outlier.

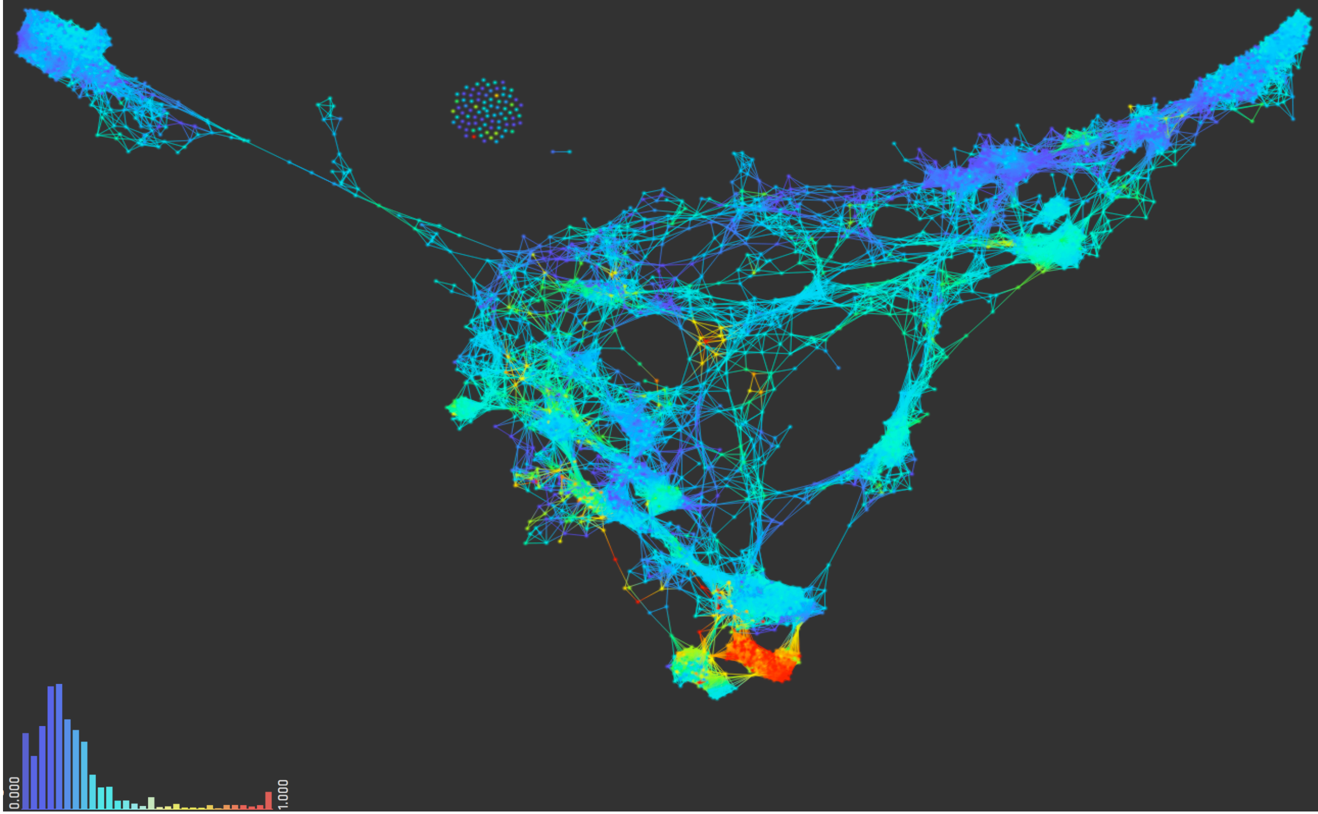
Model Introspection: Customer Cost

AYASDI



Model Introspection: Model Outliers

AYASDI



Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers.

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation
- The client was completely unaware that they we flagging a consistent group of customers. Assumed it was distributed throughout the space.

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation
- The client was completely unaware that they we flagging a consistent group of customers. Assumed it was distributed throughout the space.

Further Investigation:

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation
- The client was completely unaware that they we flagging a consistent group of customers. Assumed it was distributed throughout the space.

Further Investigation:

- Have we found a model free outlier model? No cost information needed?

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation
- The client was completely unaware that they we flagging a consistent group of customers. Assumed it was distributed throughout the space.

Further Investigation:

- Have we found a model free outlier model? No cost information needed?
- More likely: Our models have a systematic bias.

Model Introspection: Outliers

Observation:

- The different (independent?) models are all flagging the same group of customers as cost outliers. Remember that we didn't use cost in the network creation
- The client was completely unaware that they we flagging a consistent group of customers. Assumed it was distributed throughout the space.

Further Investigation:

- Have we found a model free outlier model? No cost information needed?
- More likely: Our models have a systematic bias.

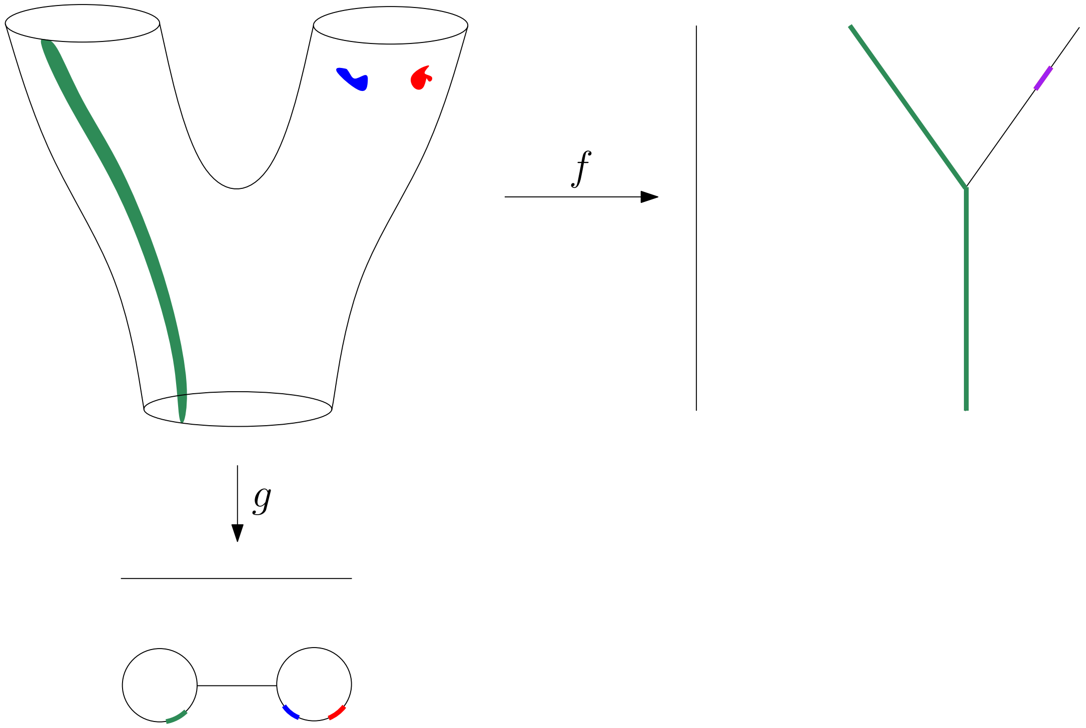
$\Rightarrow$ TDA tells use where to look in our data for problems and questions.

Thank You!

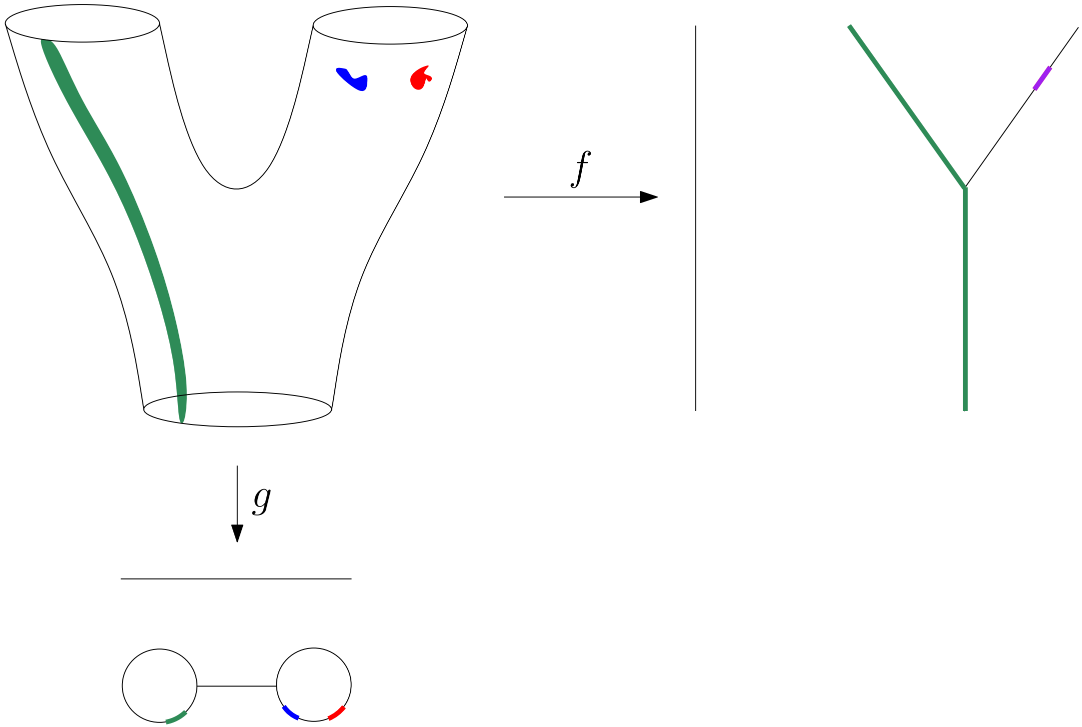
http://www.ayasdi.com/

Part II

1. Review
2. Why Topolgy? (With Examples)
3. More Applications



TDA is a machine for creating geometric/topological summaries.



TDA is a machine for creating geometric/topological summaries.

The shape (segmentations, groupings, features) represent verified hypothesis. You have to decide if they are interesting.

Why Topology?

Topology has three properties that make it well suited for data analysis

Why Topology?

Topology has three properties that make it well suited for data analysis

1. Coordinate Invariance

Why Topology?

Topology has three properties that make it well suited for data analysis

1. Coordinate Invariance
2. Deformation Invariance

Why Topology?

Topology has three properties that make it well suited for data analysis

1. Coordinate Invariance
2. Deformation Invariance
3. Compressed Representation

Why Topology?

Topology has three properties that make it well suited for data analysis

1. Coordinate Invariance
2. Deformation Invariance
3. Compressed Representation

We'll examine them in turn.

1) Coordinate Invariance

Coordinate Invariance

- The topology of shape doesn't depend on the coordinates used to describe the shape.

Coordinate Invariance

- The topology of shape doesn't depend on the coordinates used to describe the shape.
- Many different feature sets can describe the same phenomena

Coordinate Invariance

- The topology of shape doesn't depend on the coordinates used to describe the shape.
- Many different feature sets can describe the same phenomena
- While processing data we frequently alter the coordinates: scaling, rotation, whitening

Coordinate Invariance

- The topology of shape doesn't depend on the coordinates used to describe the shape.
- Many different feature sets can describe the same phenomena
- While processing data we frequently alter the coordinates: scaling, rotation, whitening

$\Rightarrow$ You want to study properties of your data that are invariant under coordinate changes.

Coordinate Invariance: Gene Expression

We want to study a specific biological phenomena via gene expression, such as cancer. We compare the data using:

Coordinate Invariance: Gene Expression

We want to study a specific biological phenomena via gene expression, such as cancer. We compare the data using:

- Samples from different patient populations

Coordinate Invariance: Gene Expression

We want to study a specific biological phenomena via gene expression, such as cancer.
We compare the data using:

- Samples from different patient populations
- Different collections of genes

Coordinate Invariance: Gene Expression

We want to study a specific biological phenomena via gene expression, such as cancer. We compare the data using:

- Samples from different patient populations
- Different collections of genes
- Different underlying technology

Coordinate Invariance: Gene Expression

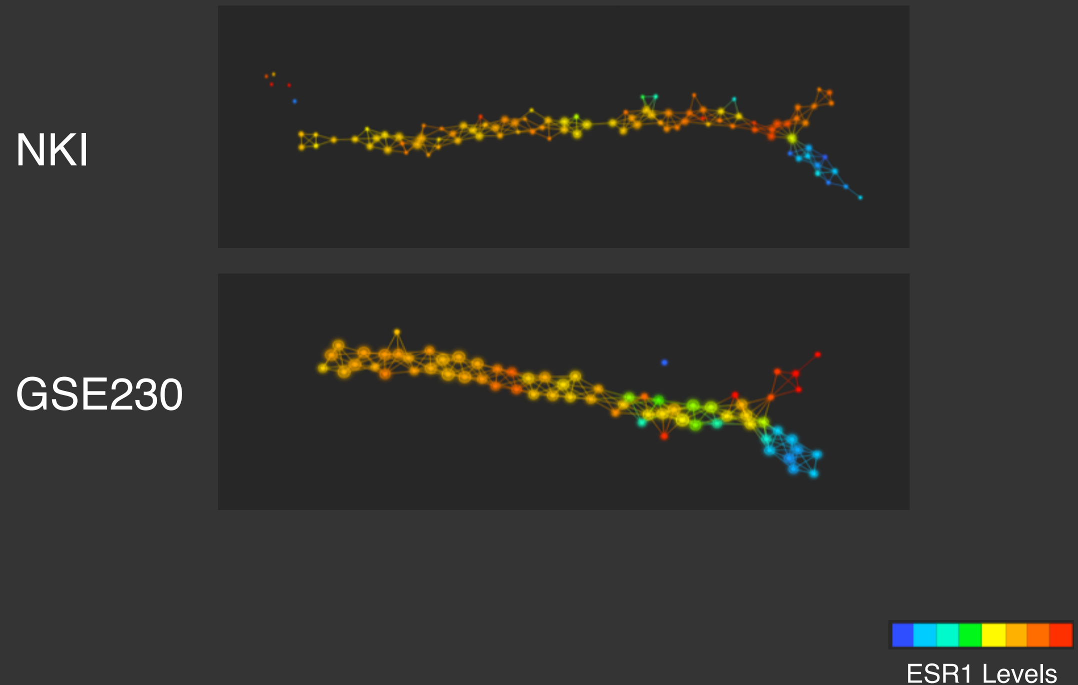
We want to study a specific biological phenomena via gene expression, such as cancer. We compare the data using:

- Samples from different patient populations
- Different collections of genes
- Different underlying technology

$\Rightarrow$ Different coordinates on Cancer

Coordinate Invariance: Gene Expression

AYASDI



2) Deformation Invariance

Deformation Invariance

- Topological features don't change when you stretch and distort the data

Deformation Invariance

- Topological features don't change when you stretch and distort the data

Advantage: Makes problems easier.

Deformation Invariance

- Topological features don't change when you stretch and distort the data

Advantage: Makes problems easier.

- Noise resistance.

Deformation Invariance

- Topological features don't change when you stretch and distort the data

Advantage: Makes problems easier.

- Noise resistance.
- Less preprocessing of the data.

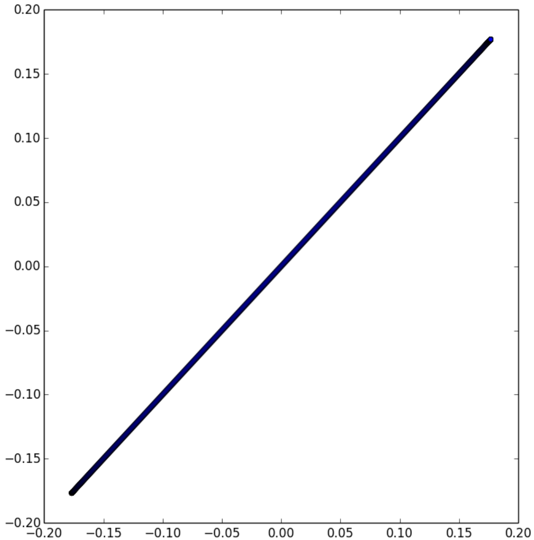
Deformation Invariance

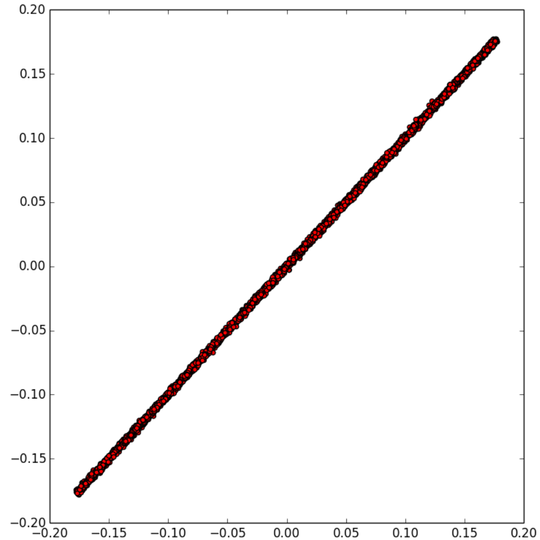
- Topological features don't change when you stretch and distort the data

Advantage: Makes problems easier.

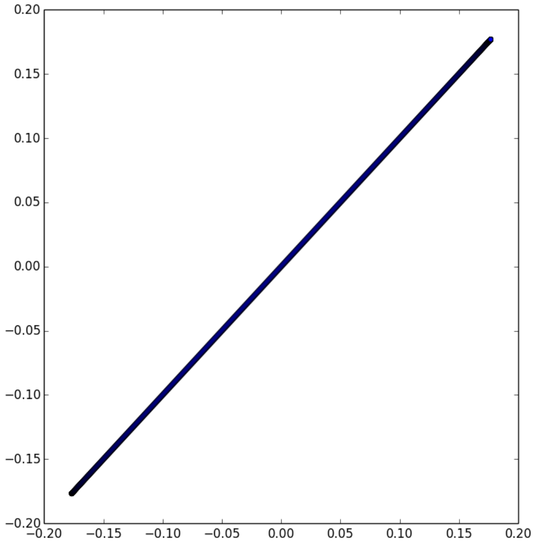
- Noise resistance.
- Less preprocessing of the data.
- Robust (stable) answers.

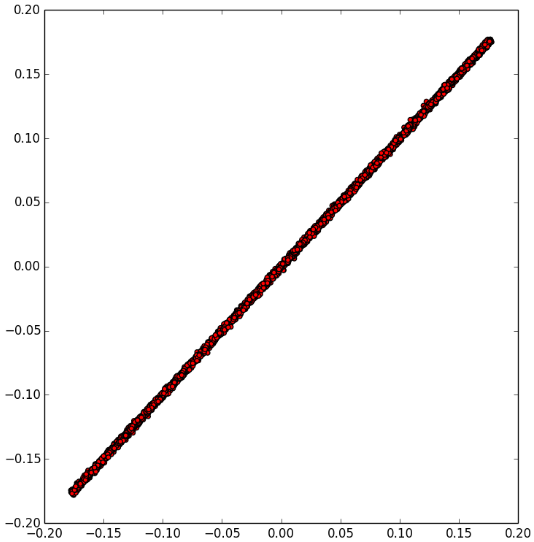
Deformation Invariance: Line and Noisy Line





Deformation Invariance: Line and Noisy Line



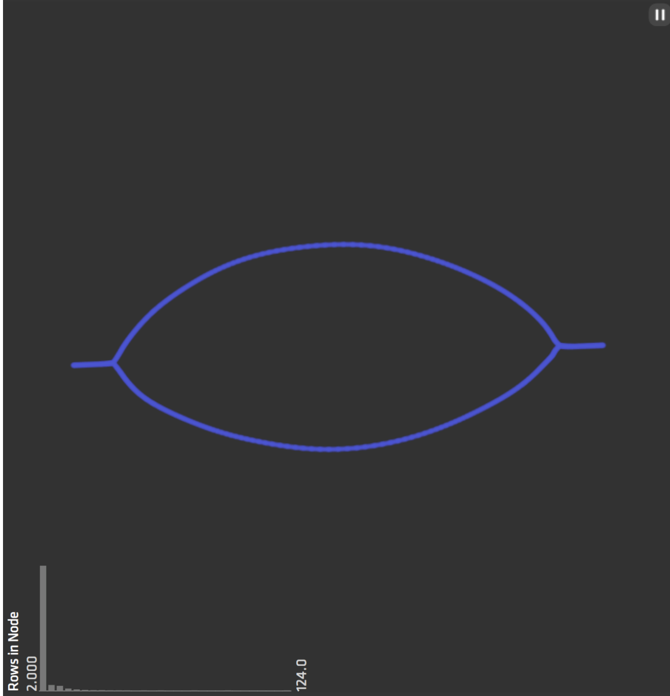


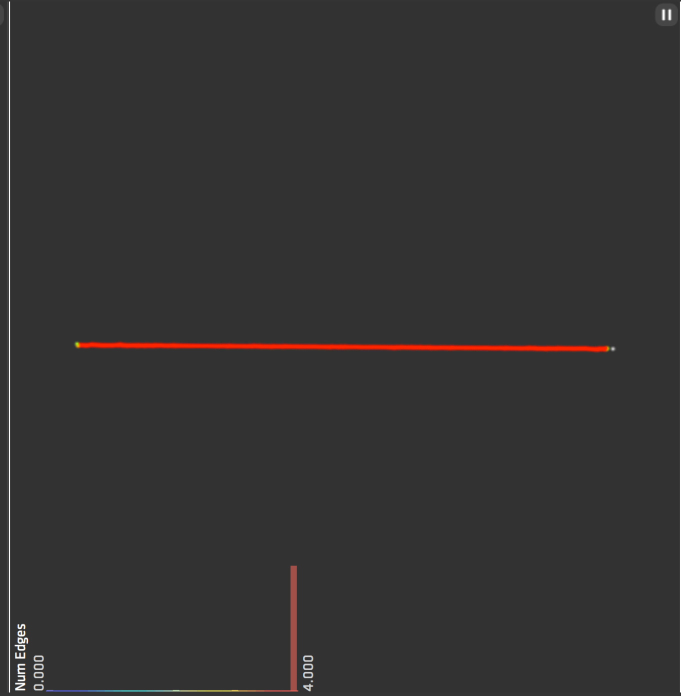
Pearson Correlation: 0.999998 resp 0.9999

Use x-axis coordinate as a lens. Expect that we will get two lines out.

Deformation Invariance: Line is a Circle

AYASDI





Deformation Invariance: Circle and Noisy Line

Some lessons.

- We can be surprised even when we think the solution is obvious. Both examples had almost perfect correlation.

Deformation Invariance: Circle and Noisy Line

Some lessons.

- We can be surprised even when we think the solution is obvious. Both examples had almost perfect correlation.
- We did not think to transform the data to look for structure. TDA saved us.

Deformation Invariance: Circle and Noisy Line

Some lessons.

- We can be surprised even when we think the solution is obvious. Both examples had almost perfect correlation.
- We did not think to transform the data to look for structure. TDA saved us.
- Being insensitive to deformation means we discover unexpected structure.

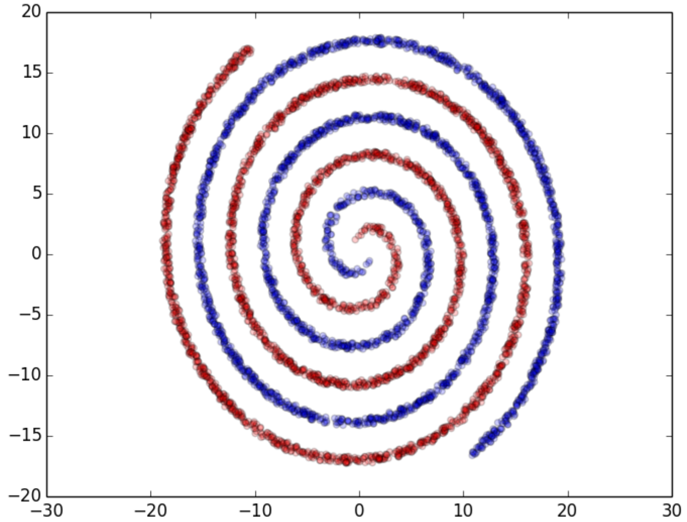
Deformation Invariance: Circle and Noisy Line

Some lessons.

- We can be surprised even when we think the solution is obvious. Both examples had almost perfect correlation.
- We did not think to transform the data to look for structure. TDA saved us.
- Being insensitive to deformation means we discover unexpected structure.
- We **did not** find structure in noise.

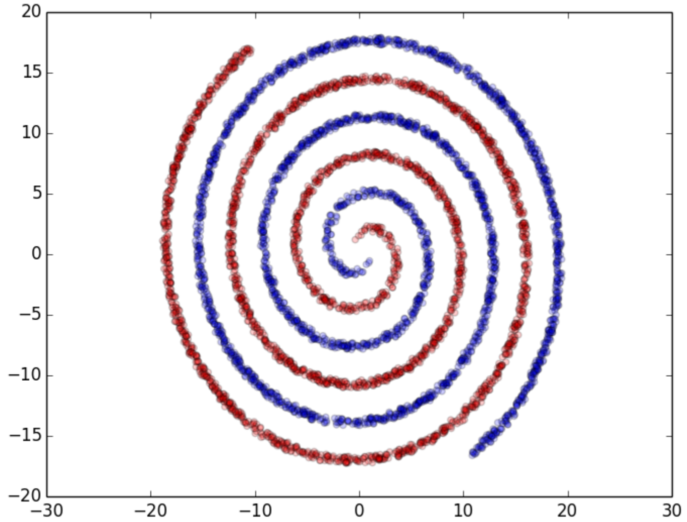
Deformation Invariance: Intertwined Spirals

Separate the two classes.



Deformation Invariance: Intertwined Spirals

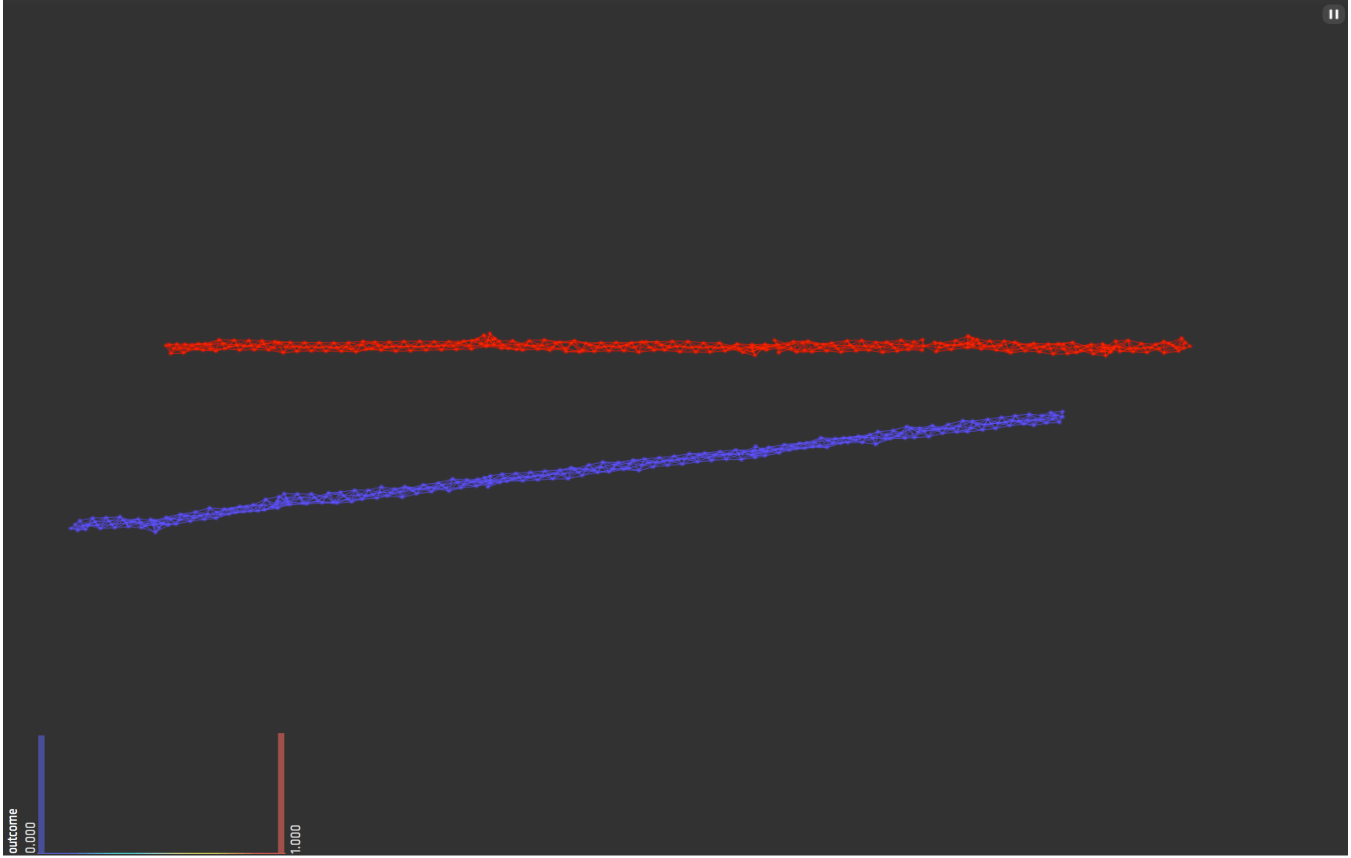
Separate the two classes.



Use x-axis coordinate as a lens.

Deformation Invariance: Intertwined Spirals

AYASDI



Deformation Invariance: Intertwined Spirals

Some lessons.

- Separating the two classes was easy. Take connected components of graph.

Deformation Invariance: Intertwined Spirals

Some lessons.

- Separating the two classes was easy. Take connected components of graph.
- We retained more information than clustering.

Deformation Invariance: Intertwined Spirals

Some lessons.

- Separating the two classes was easy. Take connected components of graph.
- We retained more information than clustering. We remember that we have lines.

Deformation Invariance: Intertwined Spirals

Some lessons.

- Separating the two classes was easy. Take connected components of graph.
- We retained more information than clustering. We remember that we have lines.
- If there was localized structure along the spiral, for example, subclasses of the two major classes, we would find those localizatons on these lines.

3) Compressed Representation

Compressed Representation

- Replace the metric space with a combinatorial summary: a simplicial complex.

Compressed Representation

- Replace the metric space with a combinatorial summary: a simplicial complex.
- The data is easier to manage, search and query while maintaining essential features.

Compressed Representation

- Replace the metric space with a combinatorial summary: a simplicial complex.
- The data is easier to manage, search and query while maintaining essential features.
- Leverage many known algorithms from **Graph Theory** , **Computational Topology** and **Computational Geometry**

Compressed Representation

- Replace the metric space with a combinatorial summary: a simplicial complex.
- The data is easier to manage, search and query while maintaining essential features.
- Leverage many known algorithms from **Graph Theory** , **Computational Topology** and **Computational Geometry**

This is more or less what TDA is about

Analogy: Cartography

To be a good and useful map:

- Come at your problem with as few assumptions as possible but bring tools to measure what's there (metrics & lenses)

Analogy: Cartography

To be a good and useful map:

- Come at your problem with as few assumptions as possible but bring tools to measure what's there (metrics & lenses)
- Measure what you find. Use as few assumptions as possible.

Analogy: Cartography

To be a good and useful map:

- Come at your problem with as few assumptions as possible but bring tools to measure what's there (metrics & lenses)
- Measure what you find. Use as few assumptions as possible.
- Produce a summary relevant to the problem.

Analogy: Cartography

To be a good and useful map:

- Come at your problem with as few assumptions as possible but bring tools to measure what's there (metrics & lenses)
- Measure what you find. Use as few assumptions as possible.
- Produce a summary relevant to the problem.
  - Different problems require different summaries.

Analogy: Cartography

To be a good and useful map:

- Come at your problem with as few assumptions as possible but bring tools to measure what's there (metrics & lenses)
- Measure what you find. Use as few assumptions as possible.
- Produce a summary relevant to the problem.
  - Different problems require different summaries.

Use your map to make decisions! Don't got back and measure from scratch.

$\Rightarrow$TDA is the machine that takes the tools (metrics & lenses) and produces the summary (network)

More Examples

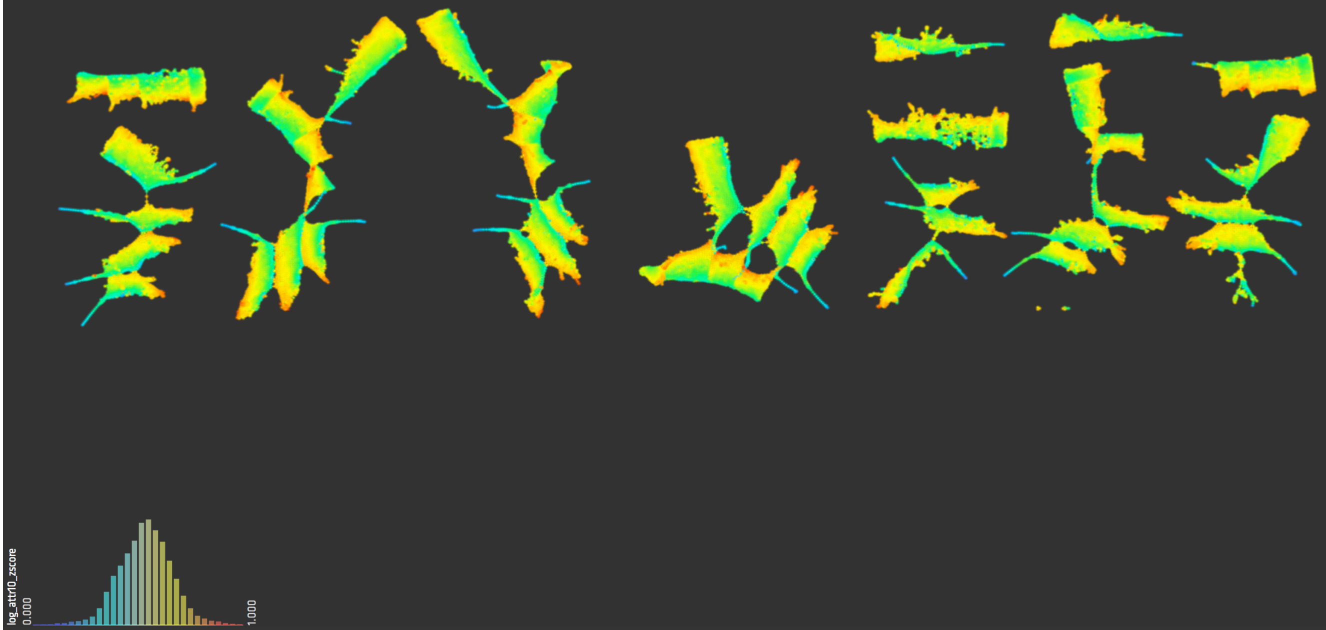
Customer Churn

Data: Customer usage and contractual details for major telco.

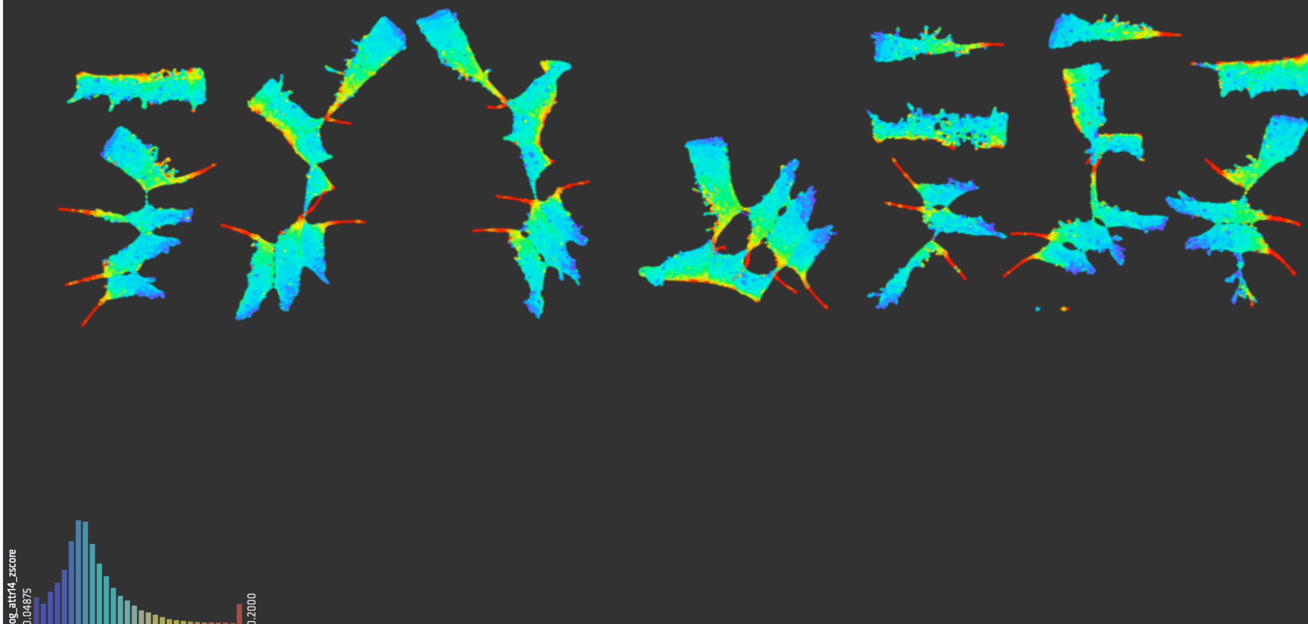
Analysis: A contractual stage data lens was used to split the data into "contractual stage" groups

Customer Churn

AYASDI

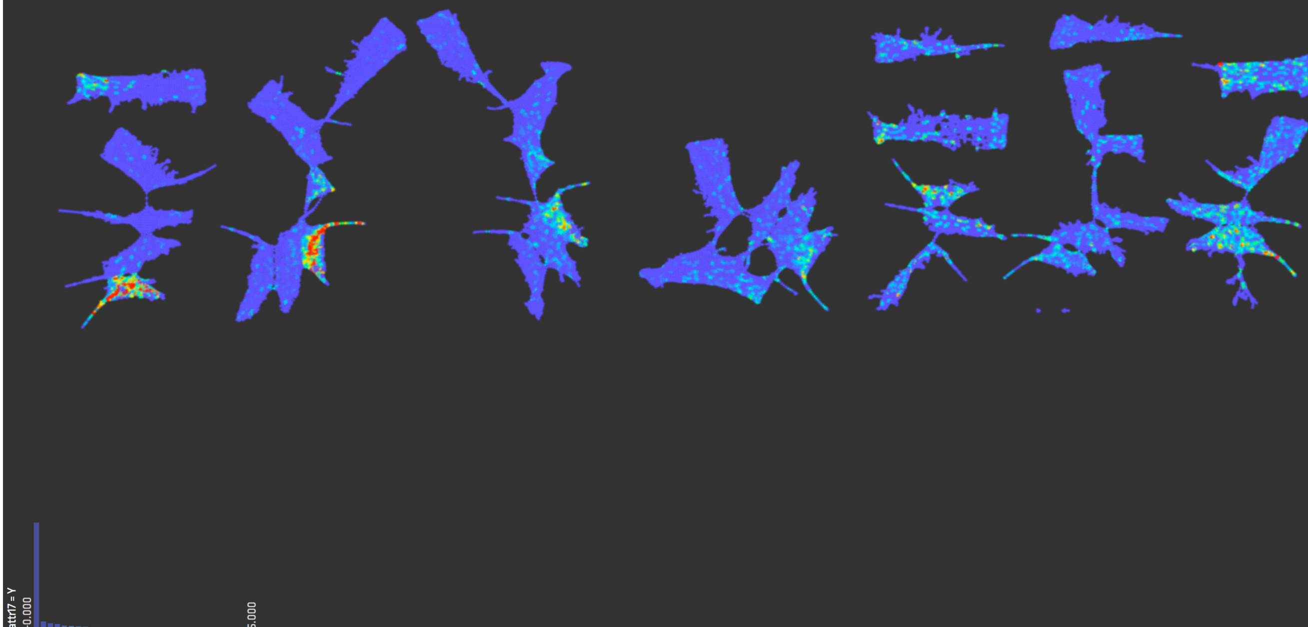


Customer Churn



Customer Churn

AYASDI



Customer Churn

Shape and Meaning

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.
- The shape tells us where to look in the data. We were able to localize churn in certain contract stages.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.
- The shape tells us where to look in the data. We were able to localize churn in certain contract stages.
- Coloring helps us figure out what is going on.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.
- The shape tells us where to look in the data. We were able to localize churn in certain contract stages.
- Coloring helps us figure out what is going on.
- Not shown: The software also allows for statistical queries of subgroups of data points.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.
- The shape tells us where to look in the data. We were able to localize churn in certain contract stages.
- Coloring helps us figure out what is going on.
- Not shown: The software also allows for statistical queries of subgroups of data points.

$\Rightarrow$ We turn our insight into better targeting resulting in fewer lost customers.

Customer Churn

Shape and Meaning

- We see similar shape across all the contract stages.
- We can see natural segmentations the vary over many orders of magnitude (100-50,000 customers) in size.
- Stability gives us confidence in the validity of the results.
- The shape tells us where to look in the data. We were able to localize churn in certain contract stages.
- Coloring helps us figure out what is going on.
- Not shown: The software also allows for statistical queries of subgroups of data points.

$\Rightarrow$ We turn our insight into better targeting resulting in fewer lost customers. This can be automated.

Predictive Maintenance: Industrial Machinery

Setup: We have a large piece of industrial machinery, think turbine, jet engine, locomotive or robot. Built into the machine are sensors measuring physical quantities: pressure, temperature, rpms etc.

Problem: Unscheduled downtime is very expensive.

Question: Can we predict when a part will need to be repaired in the future so we can schedule the downtime appropriately?

Predictive Maintenance: Industrial Machinery

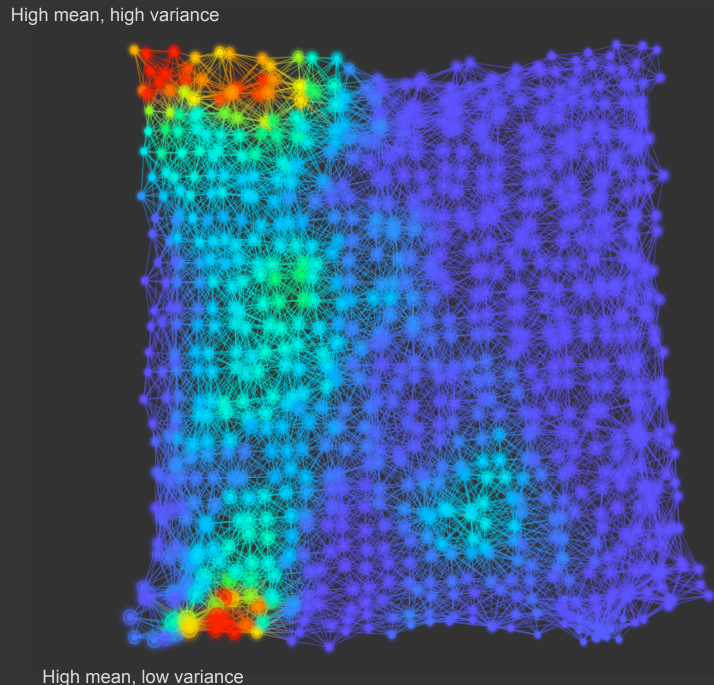
Data Transformation: We want the sensors to be comparable. In this example, z-scoring is sensible

Predictive Maintenance: Industrial Machinery

Data Transformation: We want the sensors to be comparable. In this example, z-scoring is sensible (but there are other sensible choices as well, min/max normalization, logs if sensors vary of several orders of magnitude).

Predictive Maintenance: Industrial Machinery

AYASDI



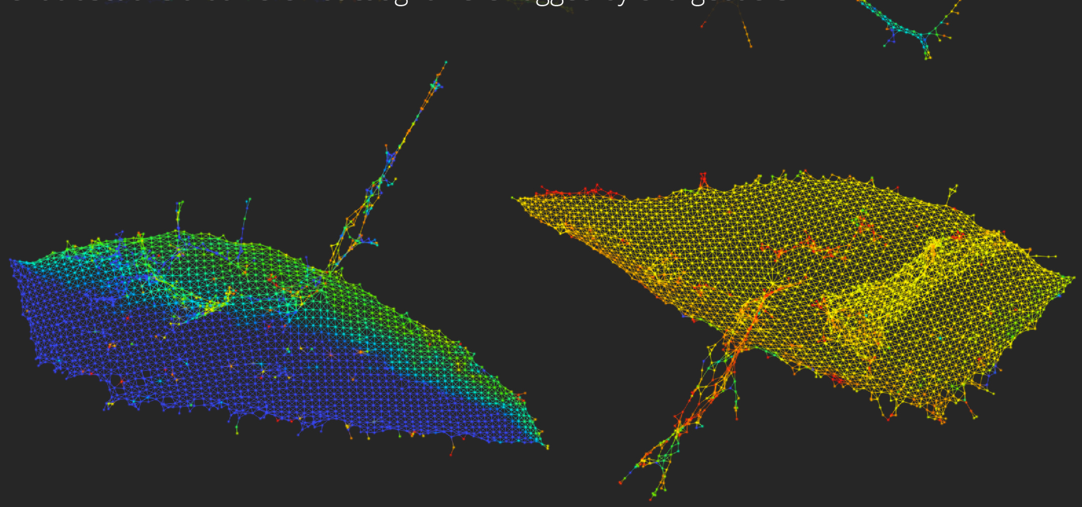
Fraud Detection

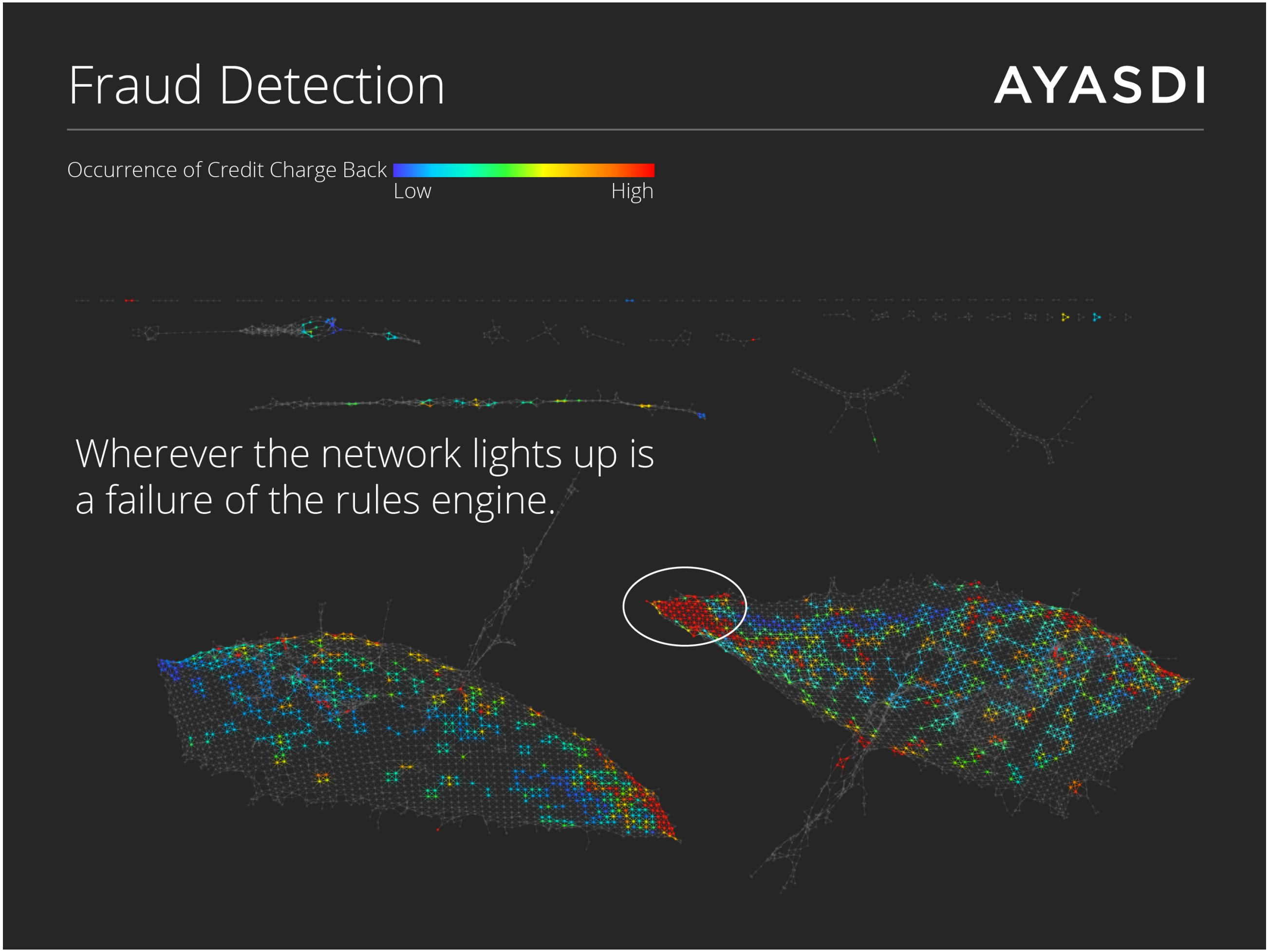
AYASDI

Risk score
Low High

About the data:

600,000+ transactions for a given month
Each transaction has 140 attributes (account, device, timing)
Fraudulent trasctions that were not caught were flagged by chargebacks

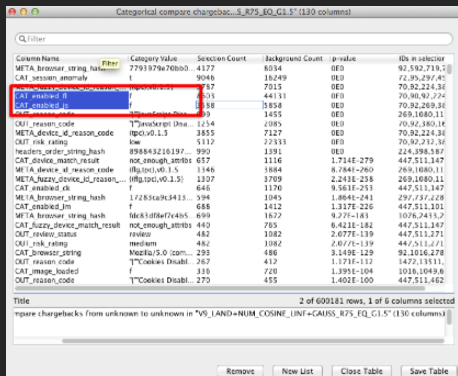



Fraud Detection
AYASDI
Occurrence of Credit Charge Back
Low
High
Wherever the network lights up is
a failure of the rules engine.

Fraud Detection

AYASDI

Occurrence of Credit Charge Back
Low — High



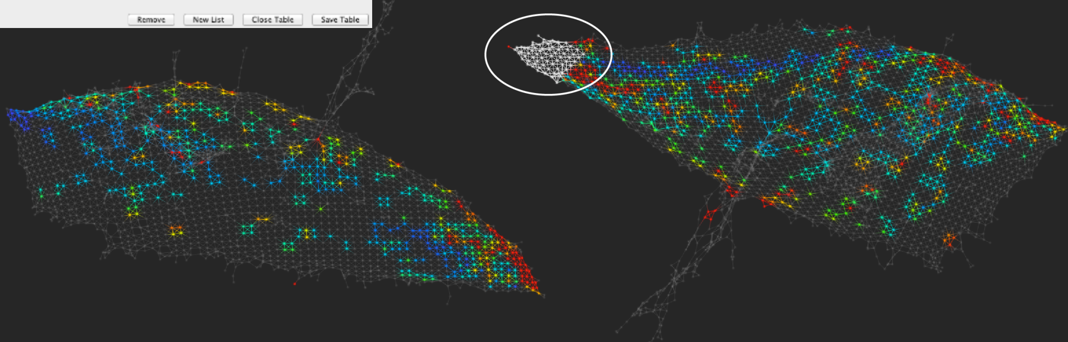
Categorical compare chargebac...S_R75_EQ_G1.5" (130 columns)

Column Name	Category Value	Selection Count	Background Count	p-value	IDs in selection
META_browser_string_hash	7793979e70bb0...	4177	8034	0E0	92,592,719,7
CAT_session_anomaly	t	9046	16249	0E0	72,95,297,45
[illegible]	[illegible]	[illegible]	7015	0E0	70,92,224,38
CAT_enabled_fl	f	[illegible]	44131	0E0	70,90,92,224
CAT_enabled_js	f	[illegible]	5858	0E0	70,92,269,38
[illegible]	[illegible]	[illegible]	1455	0E0	269,1080,11
OUT_reason_code	"]"JavaScript Disa...	1254	2085	0E0	70,92,380,16
META_device_id_reason_code	(tpc),v0.1.5	3855	7127	0E0	70,92,224,38
OUT_risk_rating	low	5112	22333	0E0	70,92,232,38
headers_order_string_hash	898843216197...	990	1391	0E0	224,398,587
CAT_device_match_result	not_enough_attribs	657	1116	1.714E-279	447,511,147
META_device_id_reason_code	(flg,tpc),v0.1.5	1346	3884	8.784E-260	269,1080,11
META_fuzzy_device_id_reason_...	{(flg,tpc),v0.1.5}	1307	3709	2.243E-258	269,1080,11
CAT_enabled_ck	f	646	1170	9.561E-253	447,511,147
META_browser_string_hash	17283ca9c3413...	594	1045	1.864E-241	297,737,228
CAT_enabled_im	f	688	1412	1.317E-226	447,511,101
META_browser_string_hash	fdc83df8ef7c4b5...	699	1672	9.27E-183	1076,2433,2
CAT_fuzzy_device_match_result	not_enough_attribs	440	765	6.421E-182	447,511,147
OUT_review_status	review	482	1082	2.077E-139	447,511,271
OUT_risk_rating	medium	482	1082	2.077E-139	447,511,271
CAT_browser_string	Mozilla/5.0 (com...	293	486	3.149E-129	92,1016,278
OUT_reason_code	"]"Cookies Disabl...	267	412	1.171E-112	1472,13511,
CAT_image_loaded	f	338	720	1.395E-104	1016,1049,6
OUT_reason_code	"]"Cookies Disabl...	270	455	1.402E-100	447,511,462

Title — 2 of 600181 rows, 1 of 6 columns selected

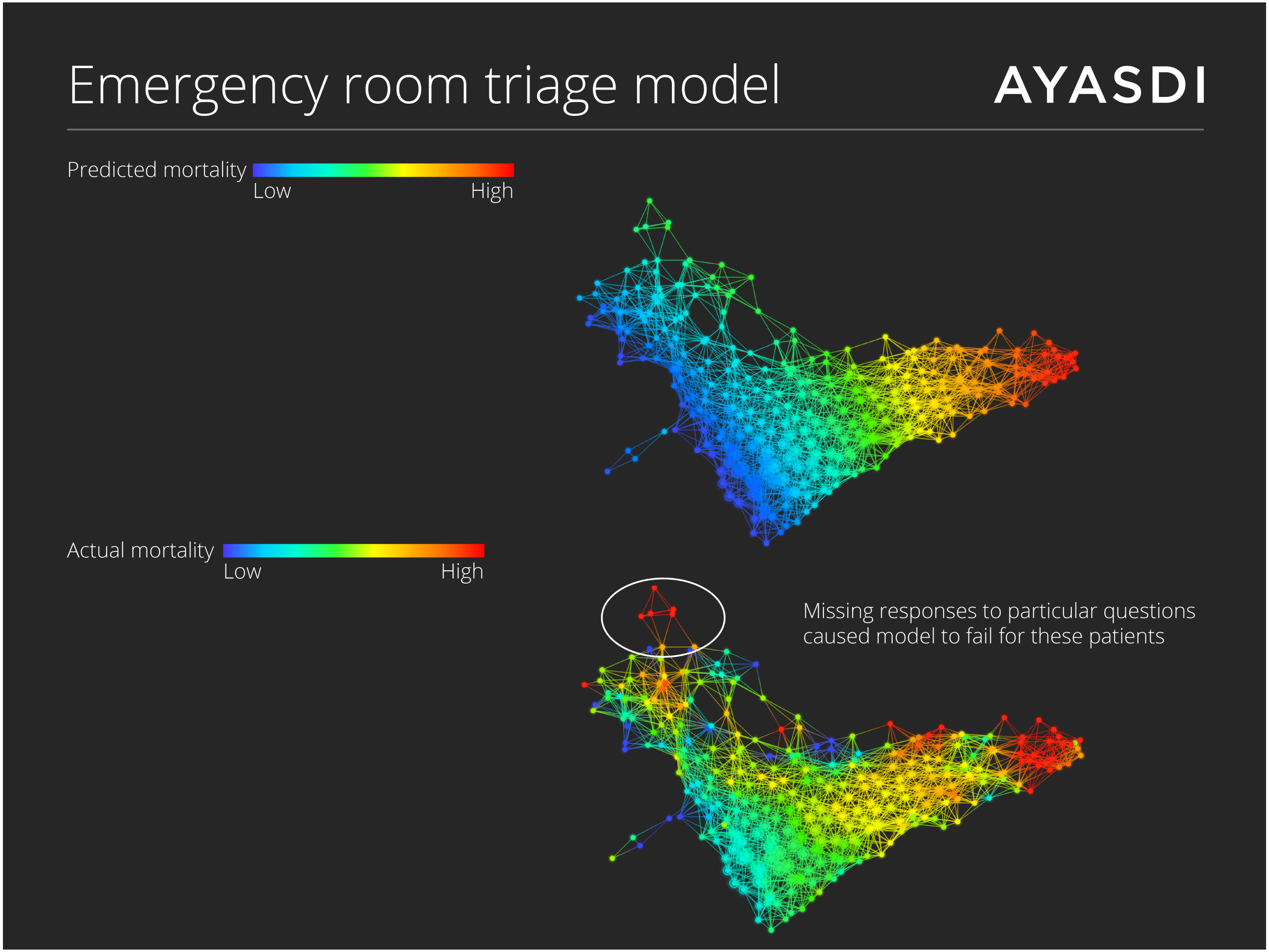
mpare chargebacks from unknown to unknown in "V9_LAND+NUM_COSINE_UNF+GAUSS_R75_EQ_G1.5" (130 columns)

Remove | New List | Close Table | Save Table

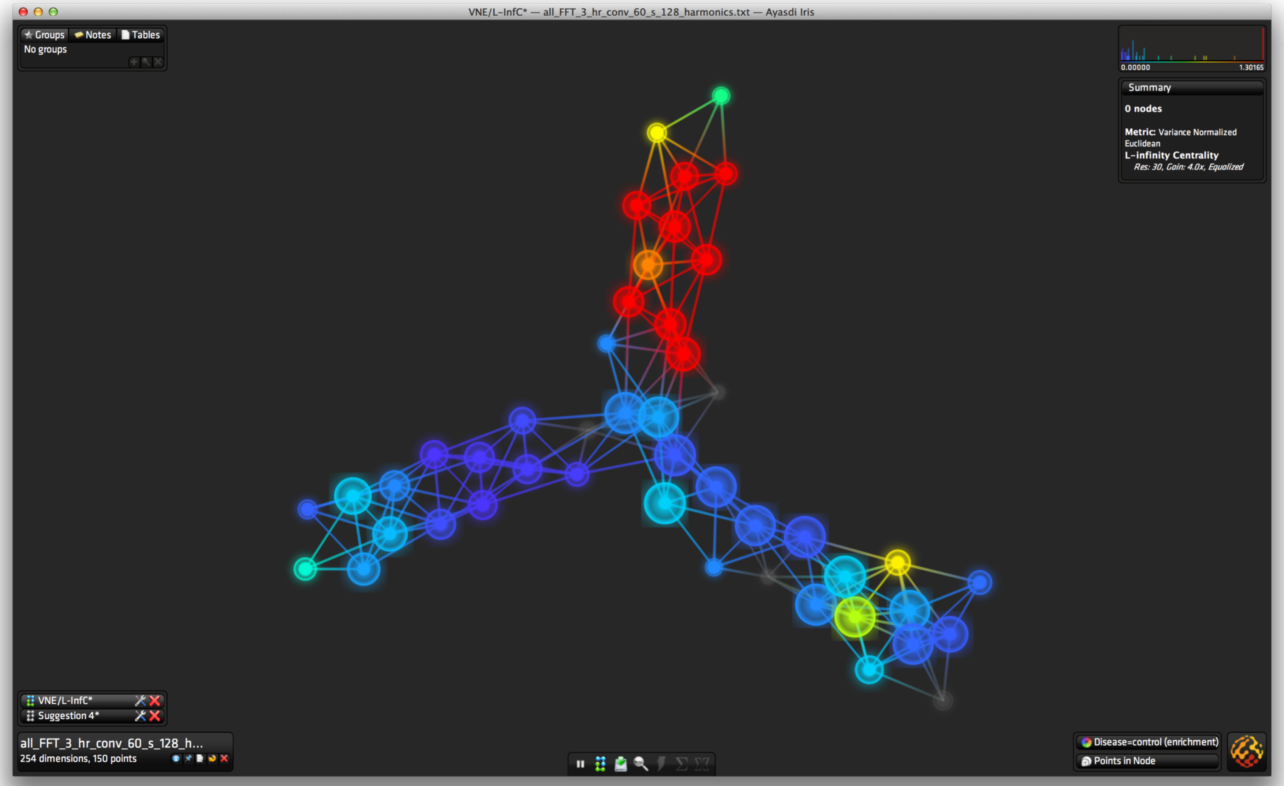


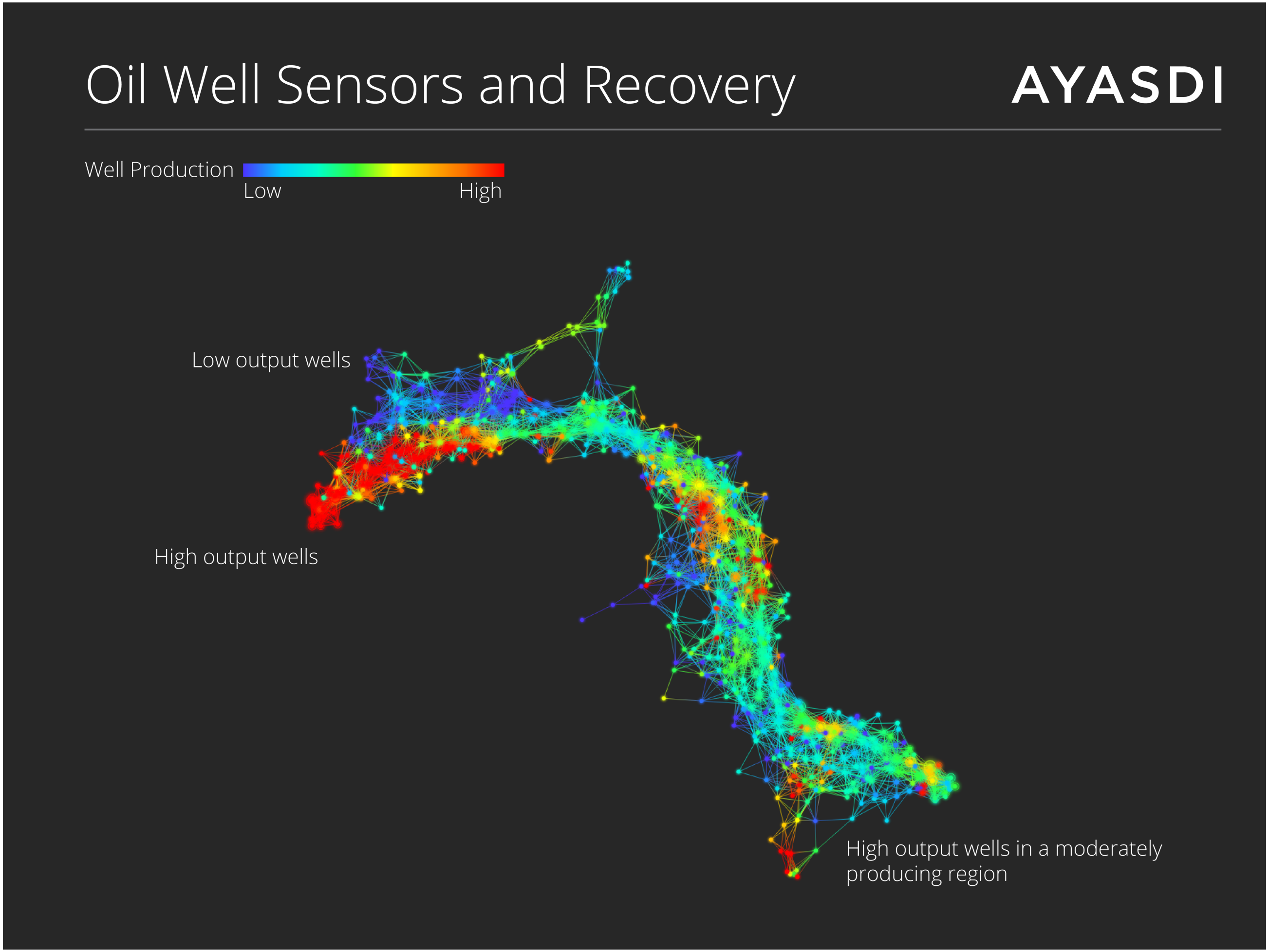
Enriched for the following attributes:

1. Images disabled
2. Javascript disabled
3. Cookies disabled
4. Flash disabled
5. Time spent on page was significantly longer


Emergency room triage model
AYASDI
Predicted mortality
Low
High
Actual mortality
Low
High
Missing responses to particular questions
caused model to fail for these patients

Parkinson's Detection with Mobile Phone



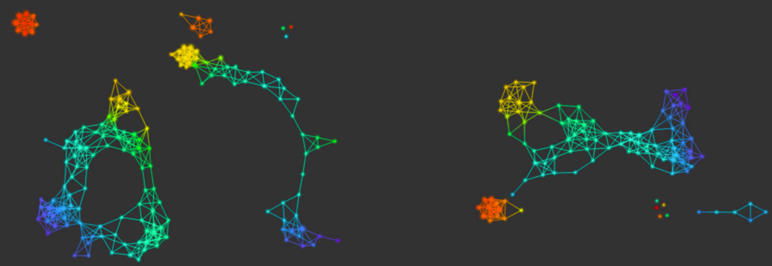


Oil Well Sensors and Recovery
AYASDI
Well Production
Low
High
Low output wells
High output wells
High output wells in a moderately producing region

Analyzing NGS Data with Ayasdi Cure

About the Data

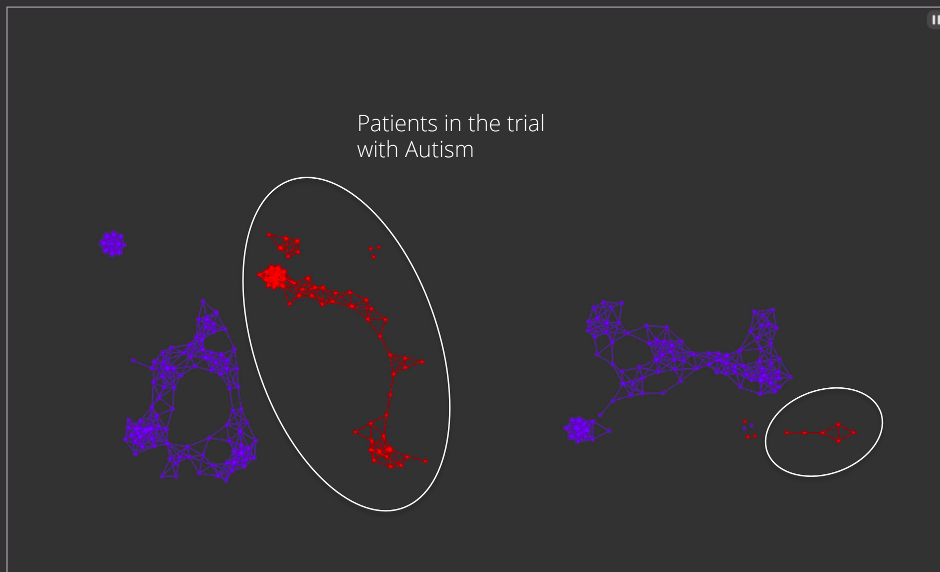
- 164 patients from autism clinical trial
- Some with autism, some without
- Data consists of genotype calls



Goal: Identify genetic drivers of the disease in subpopulations

AYASDI

Analyzing NGS Data with Ayasdi Cure

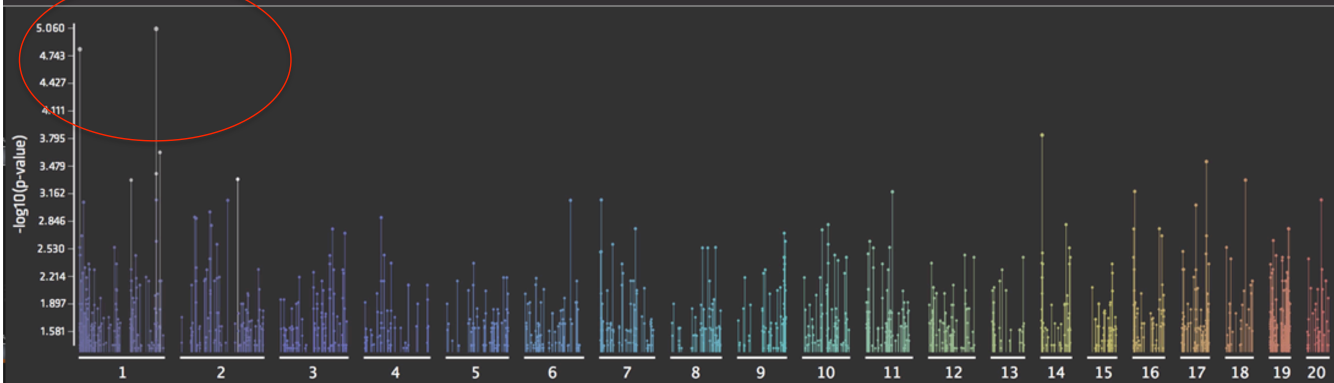


Disease Phenotype for Autism

High — Low

AYASDI

Analyzing NGS Data with Ayasdi Cure



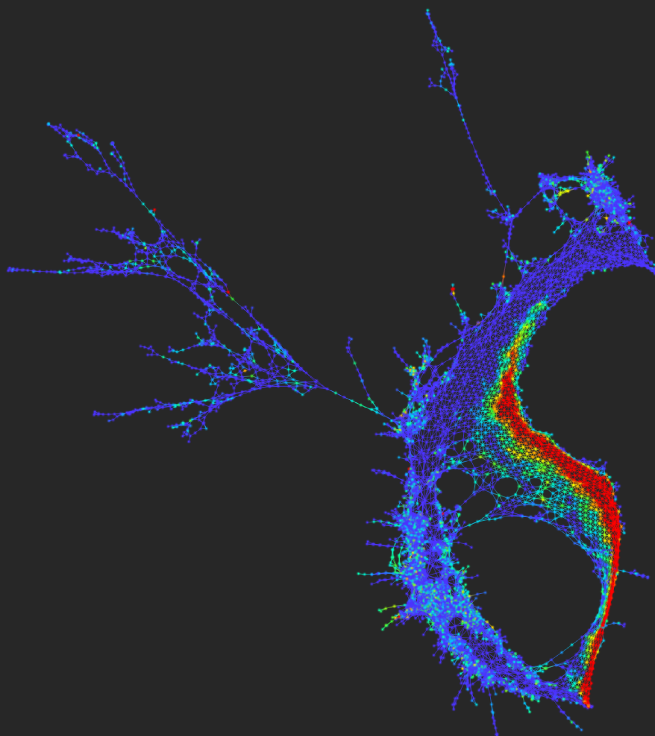
AYASDI

The Wellcome Trust Case Control Consortium

AYASDI

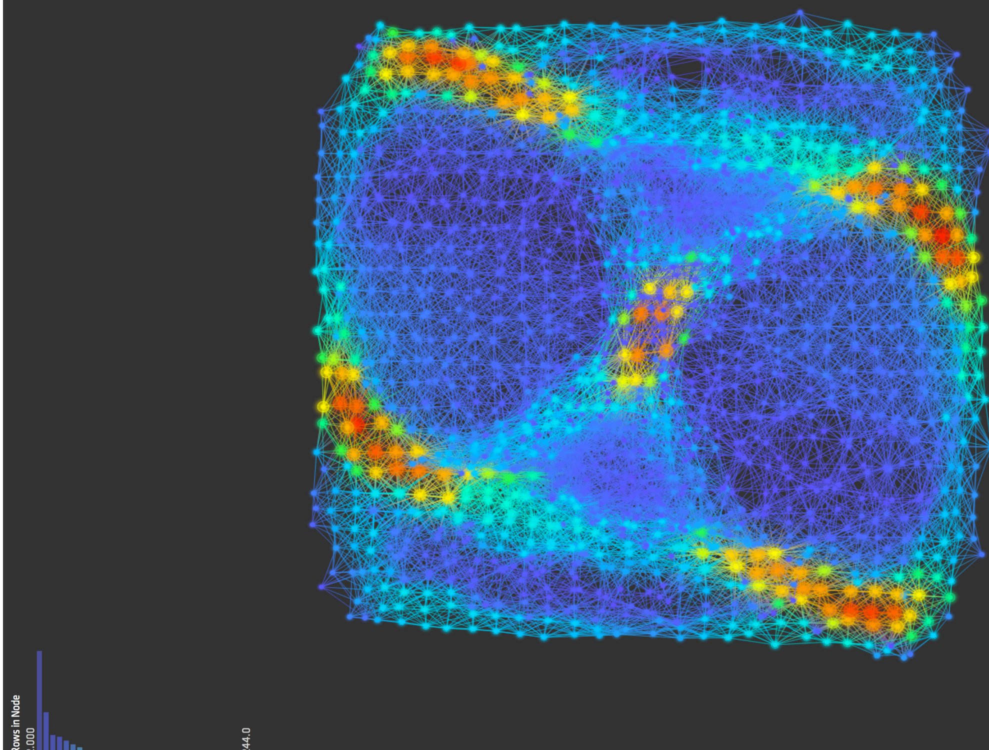
Variants with association to Crohn's Disease

Low High

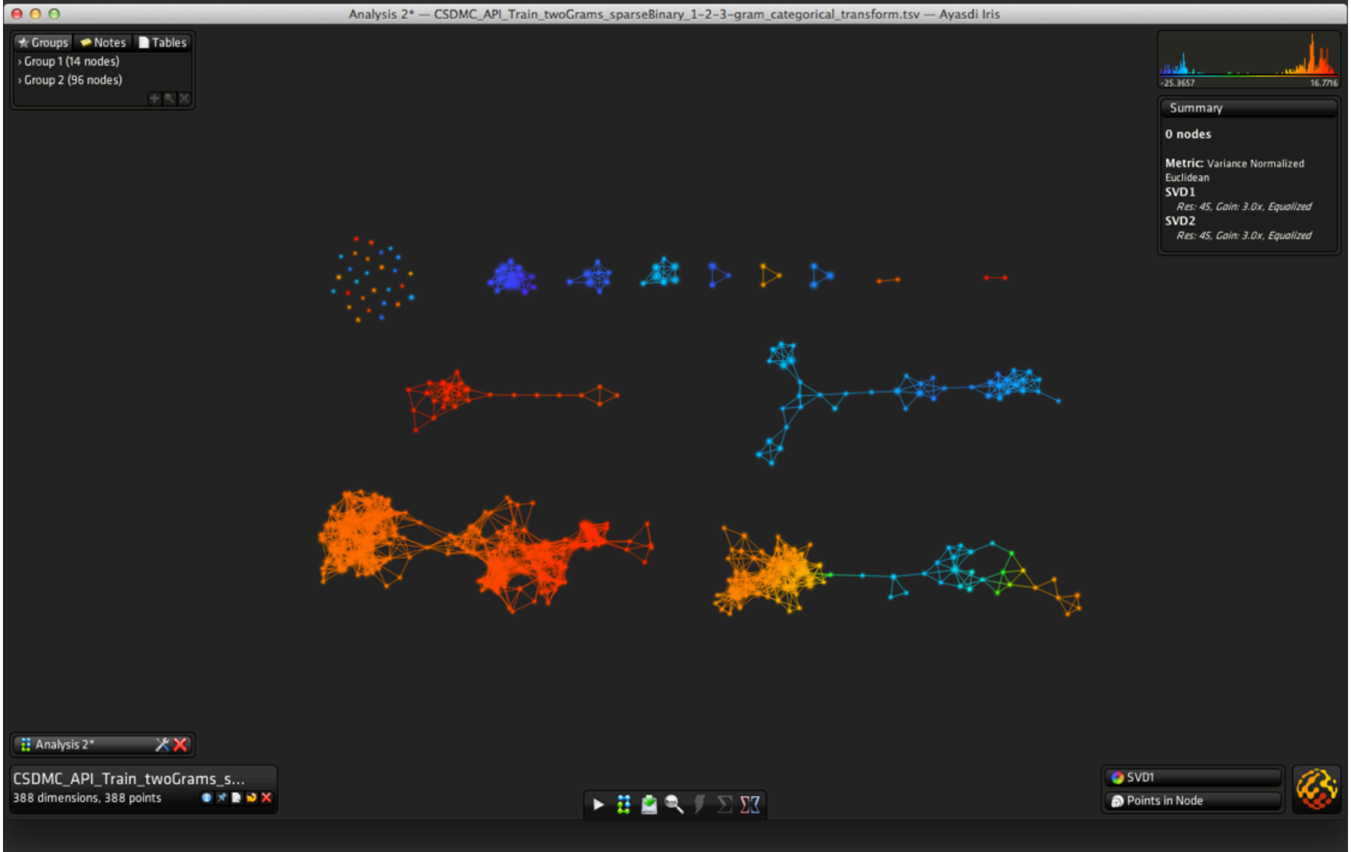


Isomap: Configuration Space of C_8H_{16}

AYASDI

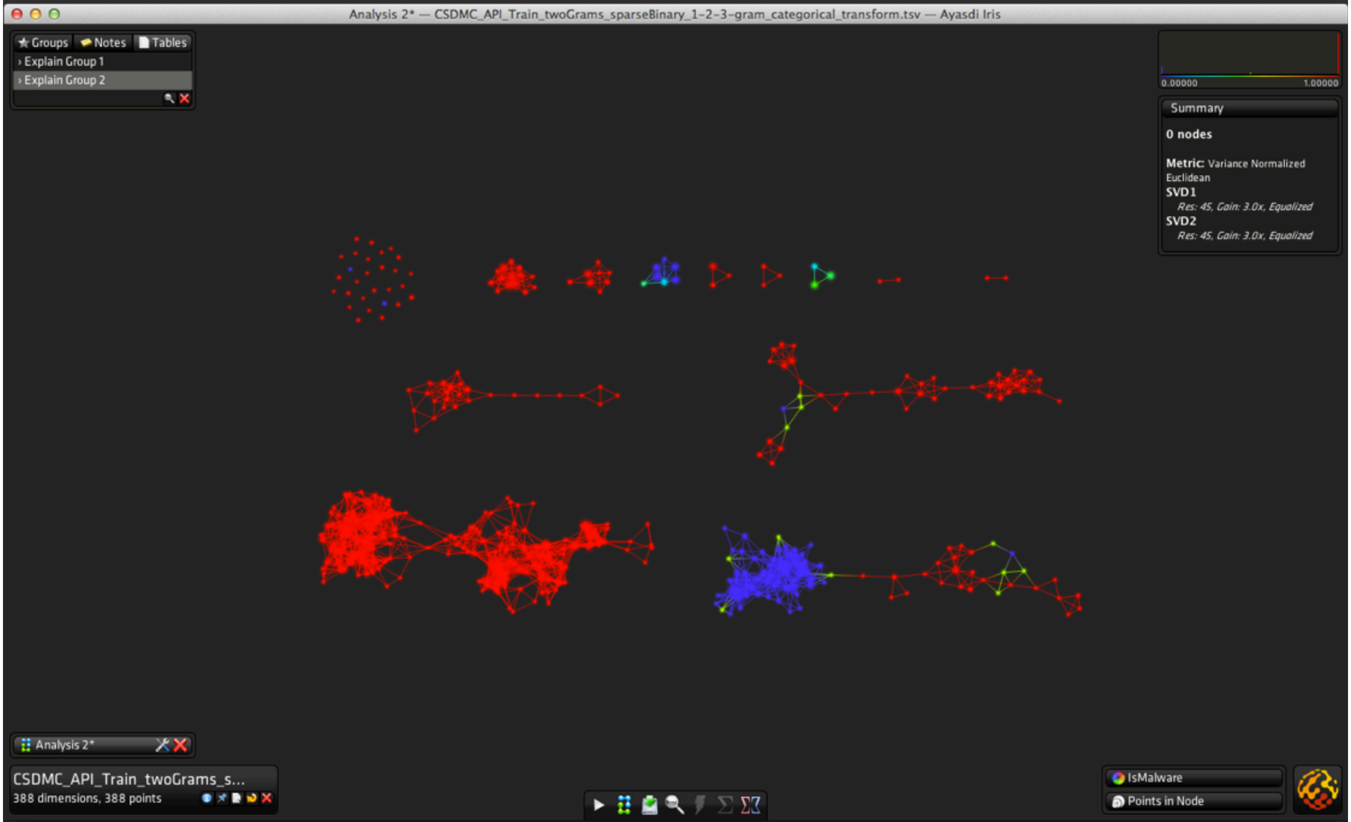


Malware System Calls

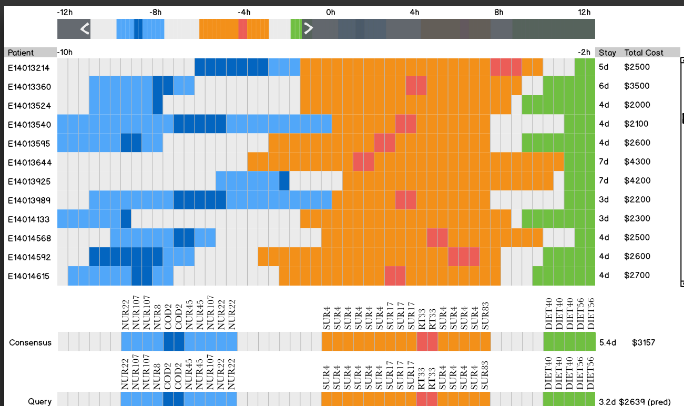


Malware System Calls

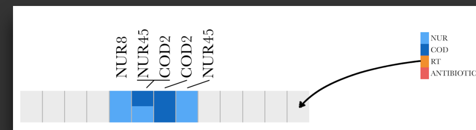
AYASDI



User Experience for Care Paths

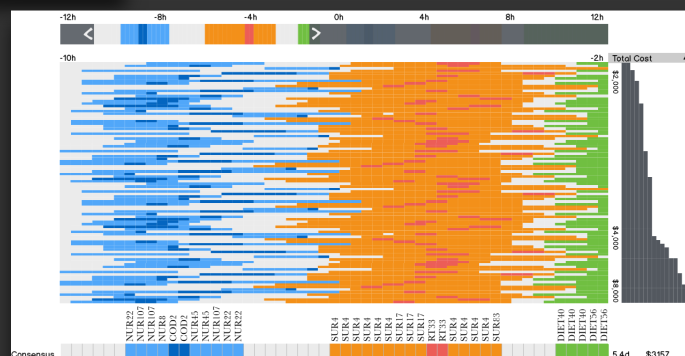


Patient Query

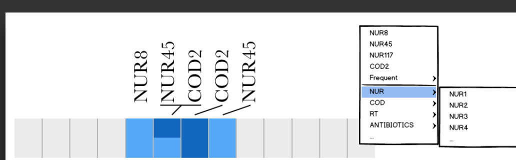


Drag and Drop Interface

Care Path Overview



Patient Query Detail



What's the point of all this?

Data Has Shape

And Shape Has Meaning

Thank You!

http://www.ayasdi.com/