Shape and Meaning

An Introduction to Topological Data Analysis

Anthony Bak

AYASDI

Goals

For this talk I want to:

Goals

For this talk I want to:

- Show you how TDA provides a framework for many machine learning/data analysis techniques

Goals

For this talk I want to:

- Show you how TDA provides a framework for many machine learning/data analysis techniques
- Demonstrate how Ayasdi provides insights into the data.

Goals

For this talk I want to:

- Show you how TDA provides a framework for many machine learning/data analysis techniques
- Demonstrate how Ayasdi provides insights into the data.

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

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 Problem in both cases is that 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!

The Problem in both cases is that there isn't a single story happening in your data.

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

Data Has Shape

And Shape Has Meaning

Data Has Shape

And Shape Has Meaning

$\Rightarrow$ In this talk I will focus on how we extract meaning.

But first... What is shape?

But first... What is shape?

- Shape is the global realization of local constraints.

But first... What is shape?

- Shape is the global realization of local constraints.
- For a given problem determined by a choice of
  - Features, columns or properties to measure.
  - A metric on the columns.

But first... What is shape?

- Shape is the global realization of local constraints.
- For a given problem determined by a choice of
  - Features, columns or properties to measure.
  - A metric on the columns.
- But not necessarily so. There are more relaxed definitions of shape and we can use those too.

But first... What is shape?

- Shape is the global realization of local constraints.
- For a given problem determined by a choice of
  - Features, columns or properties to measure.
  - A metric on the columns.
- But not necessarily so. There are more relaxed definitions of shape and we can use those too.

The goal of TDA is to understand (for us, summarize) the shape with no preconceived model of what it should be.

Math World

To show you how we extract insight from shape we start in "Math World"

Math World

To show you how we extract insight from shape we start in "Math World"

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

Math World

To show you how we extract insight from shape we start in "Math World"

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

Math World

To show you how we extract insight from shape 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

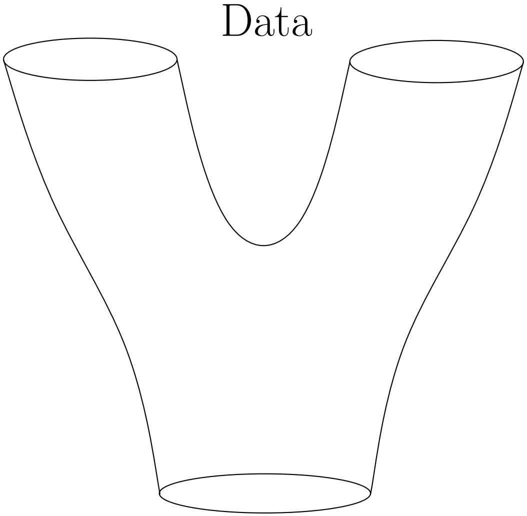
To show you how we extract insight from shape we start in "Math World"

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

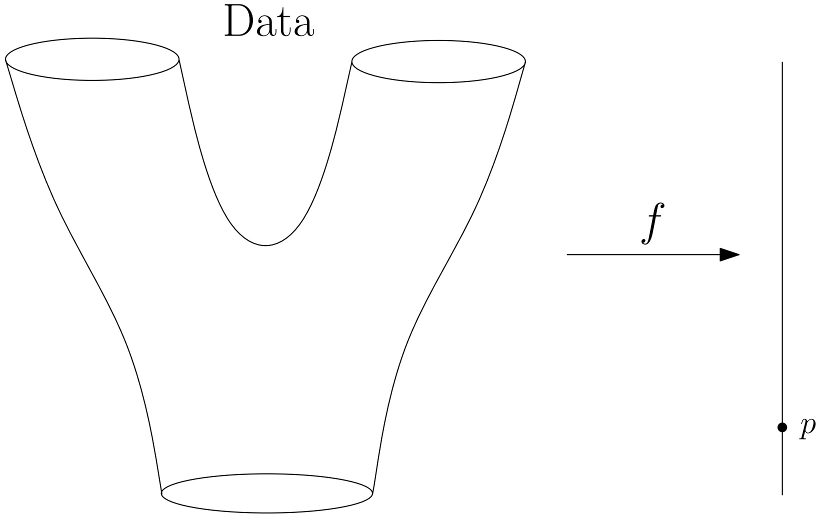
$\Rightarrow$ We will not need either of these assumptions once we're in "Data World".

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

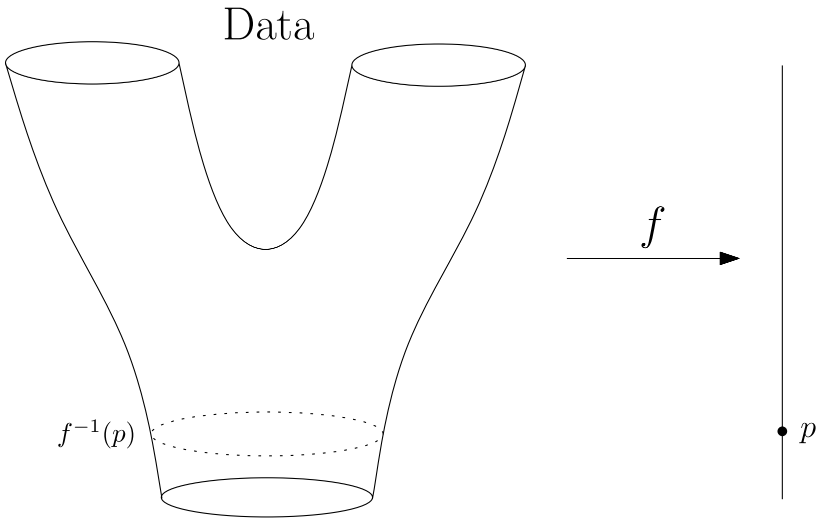
Math World



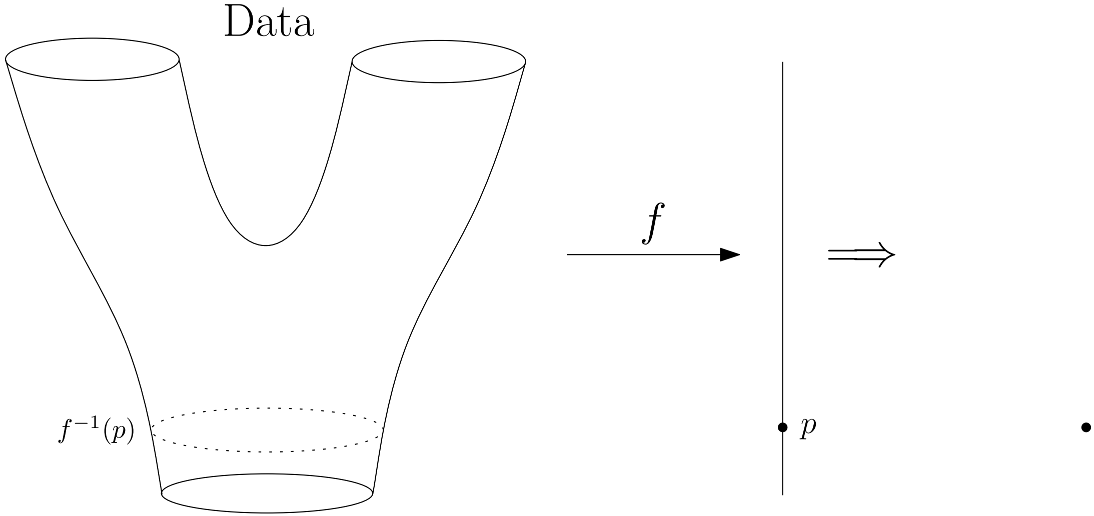
Math World



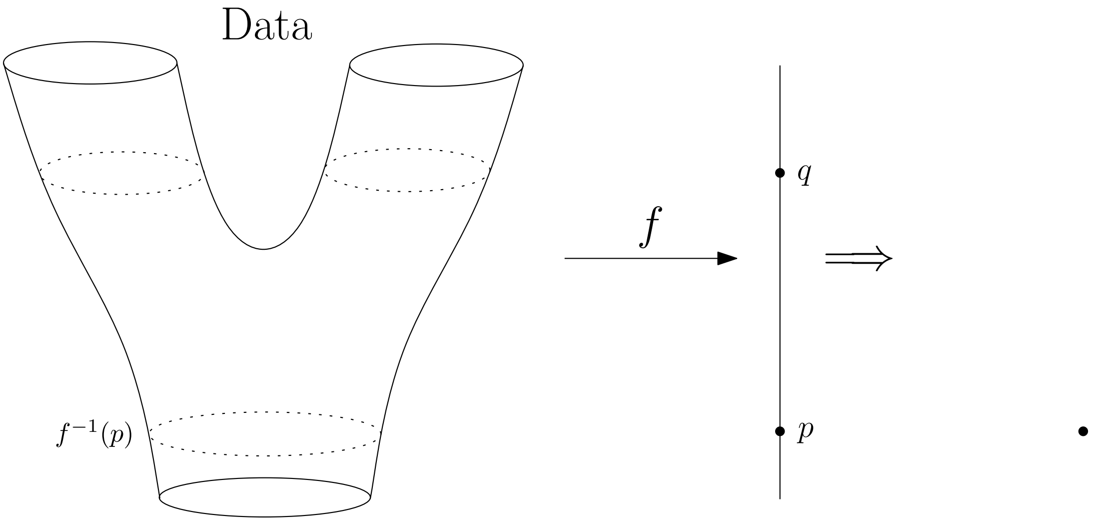
Math World



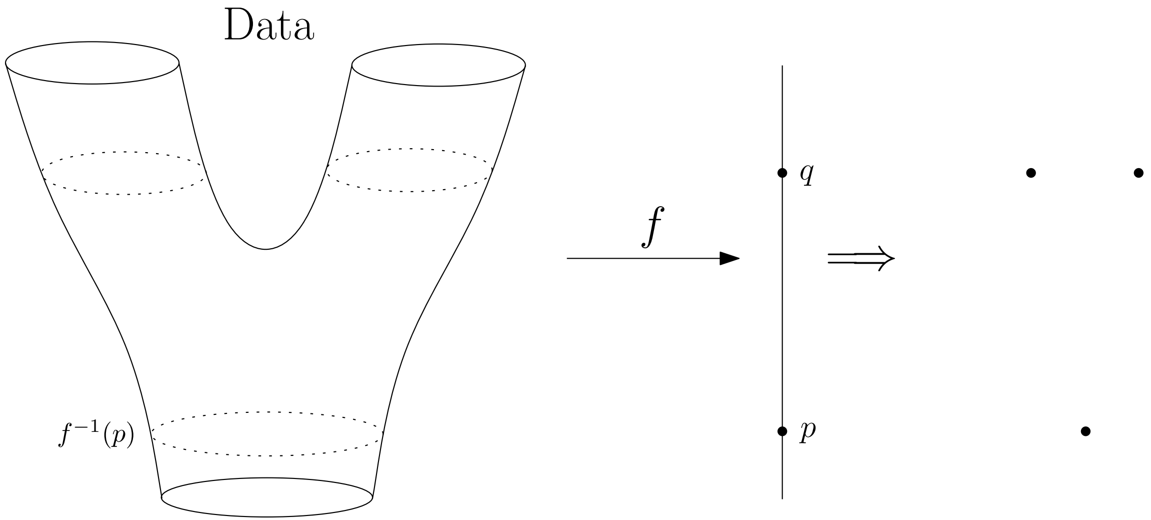
Math World



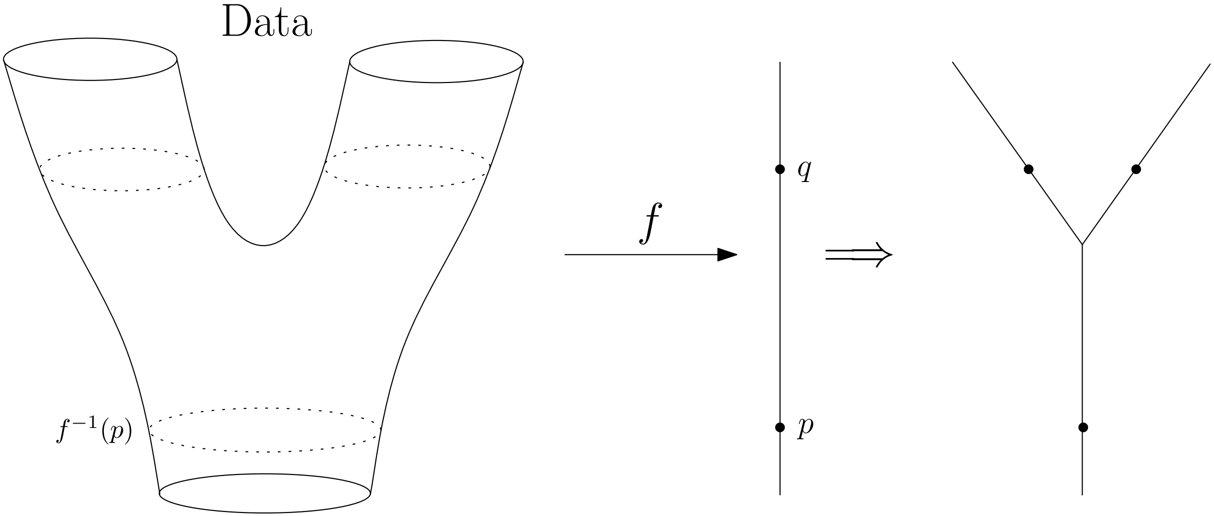
Math World



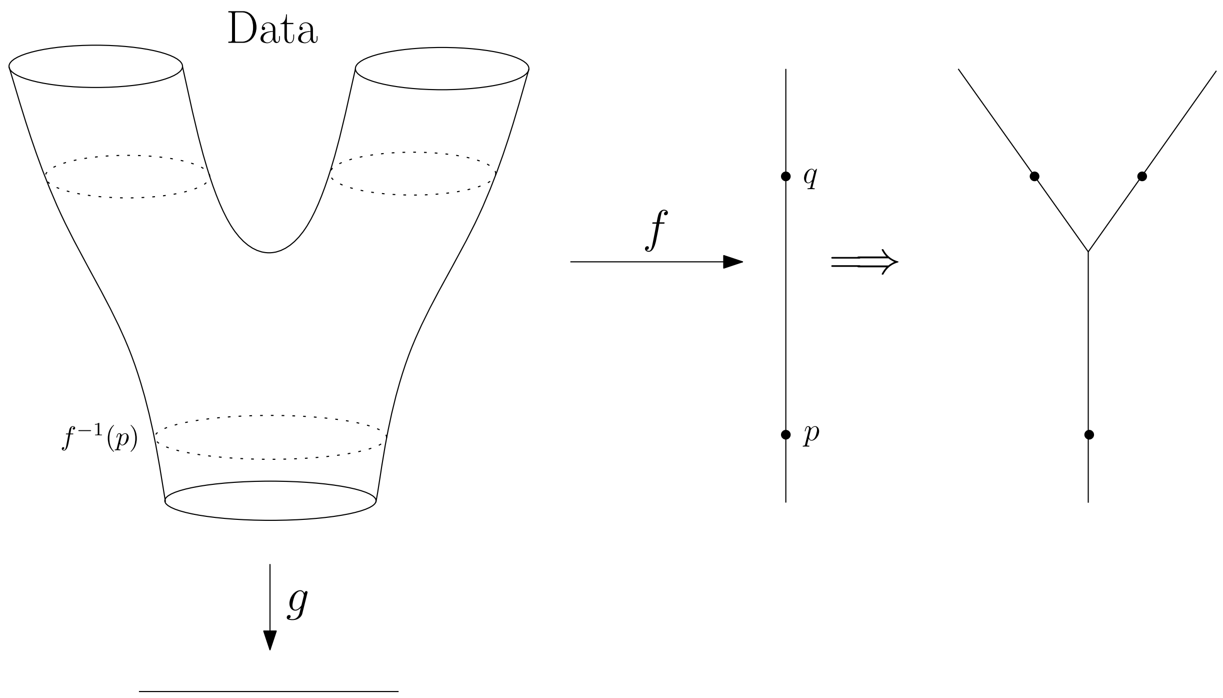
Math World



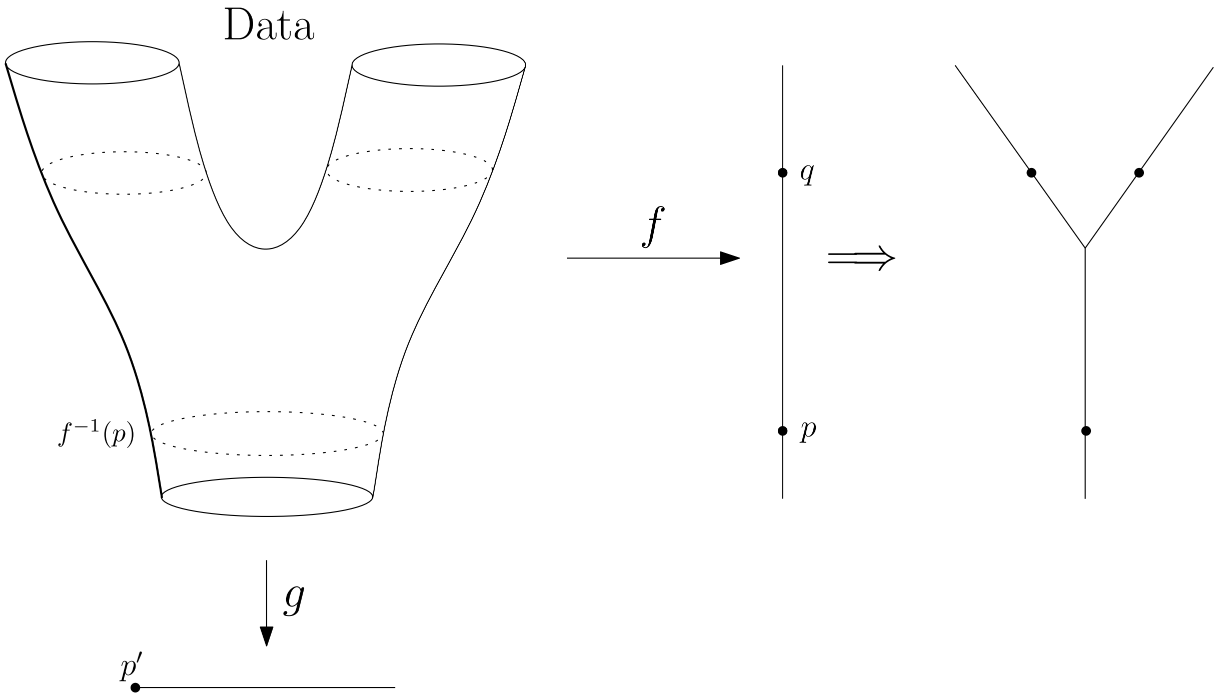
Math World



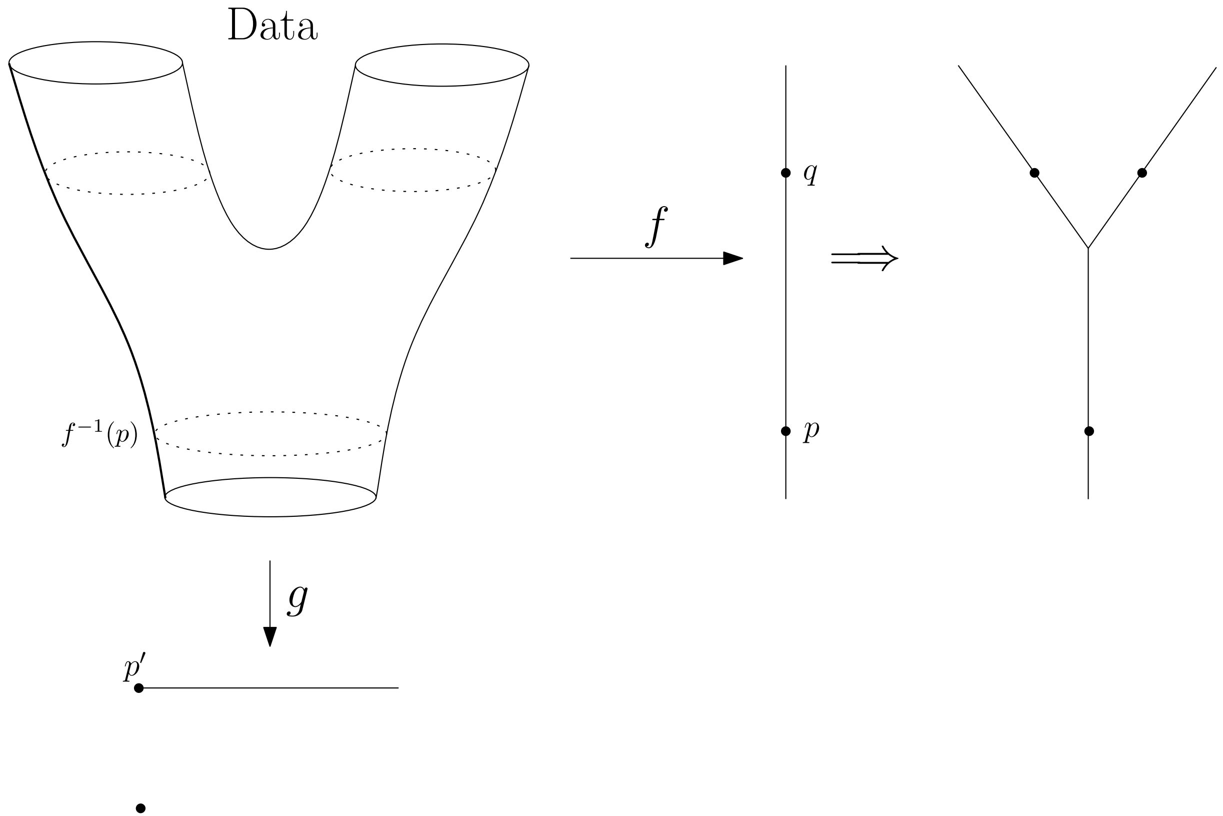
Math World



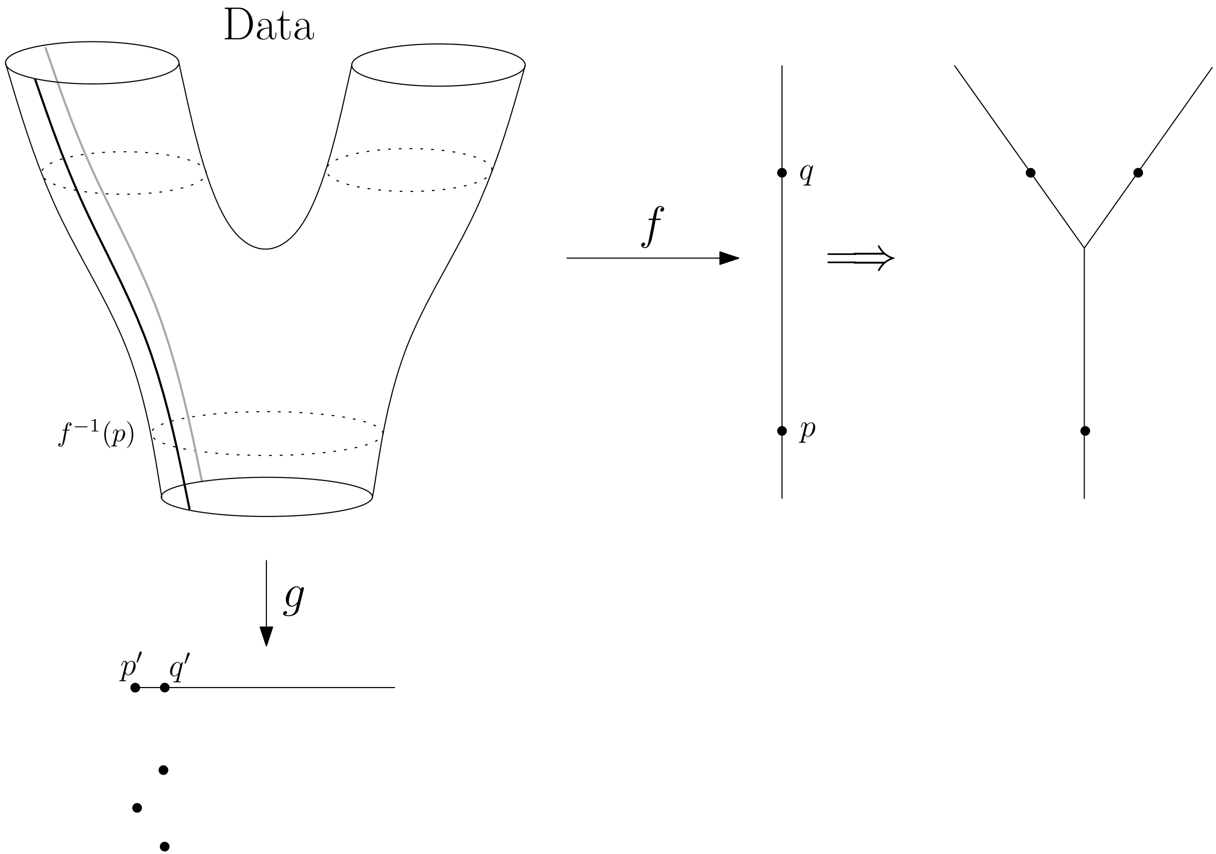
Math World



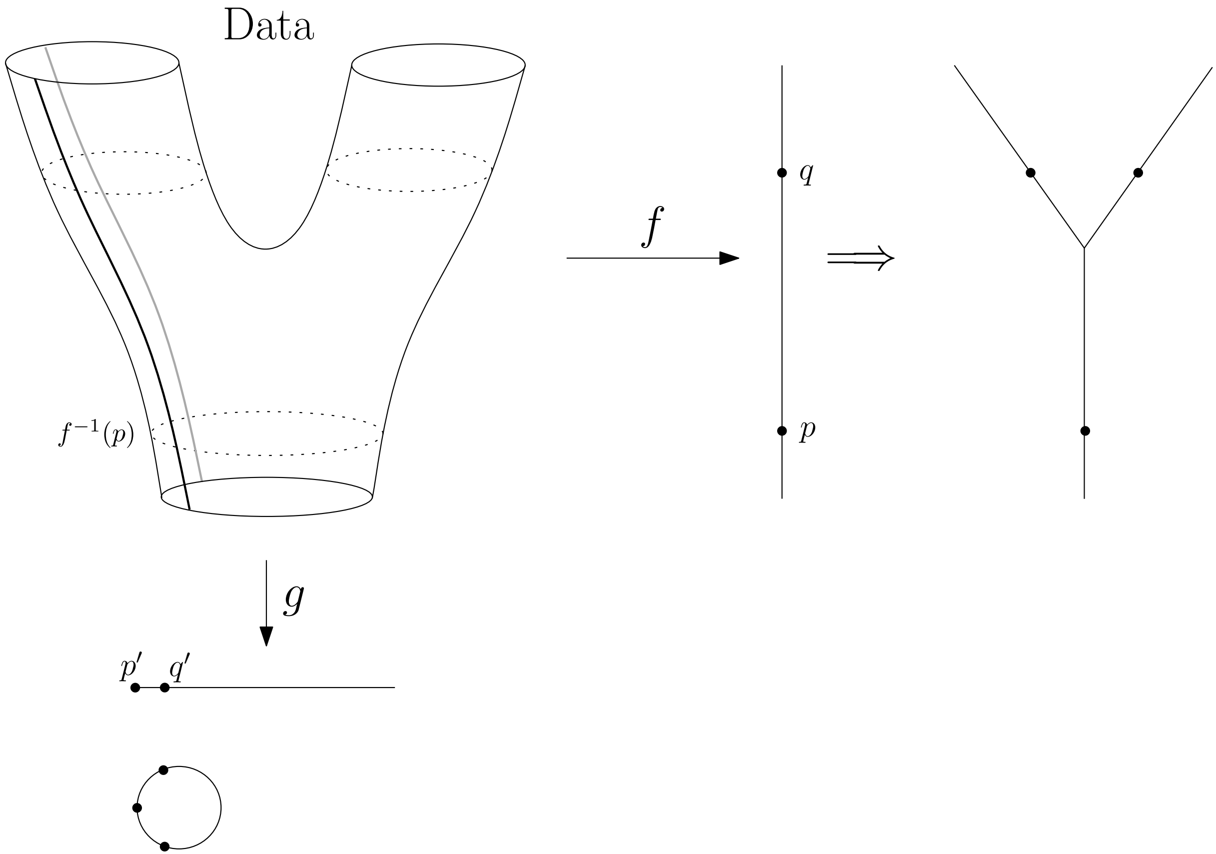
Math World



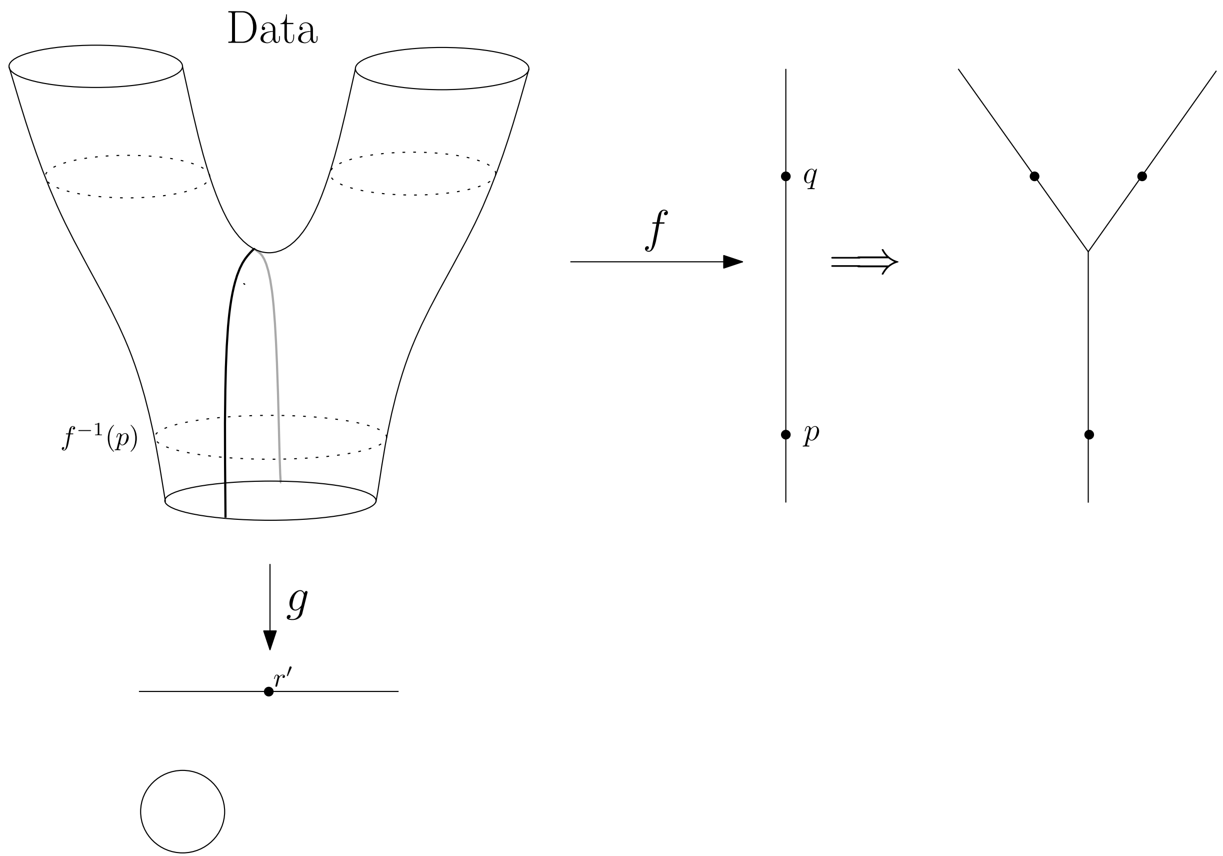
Math World



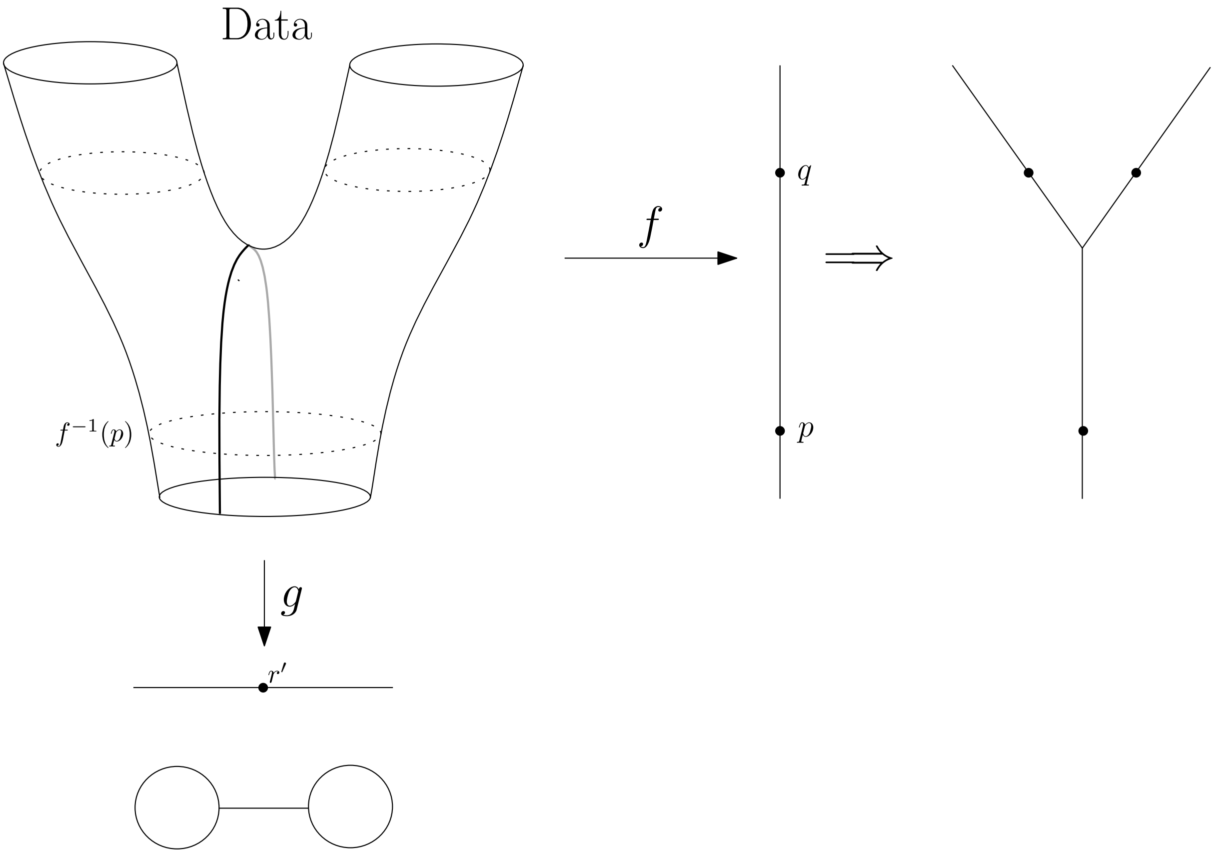
Math World



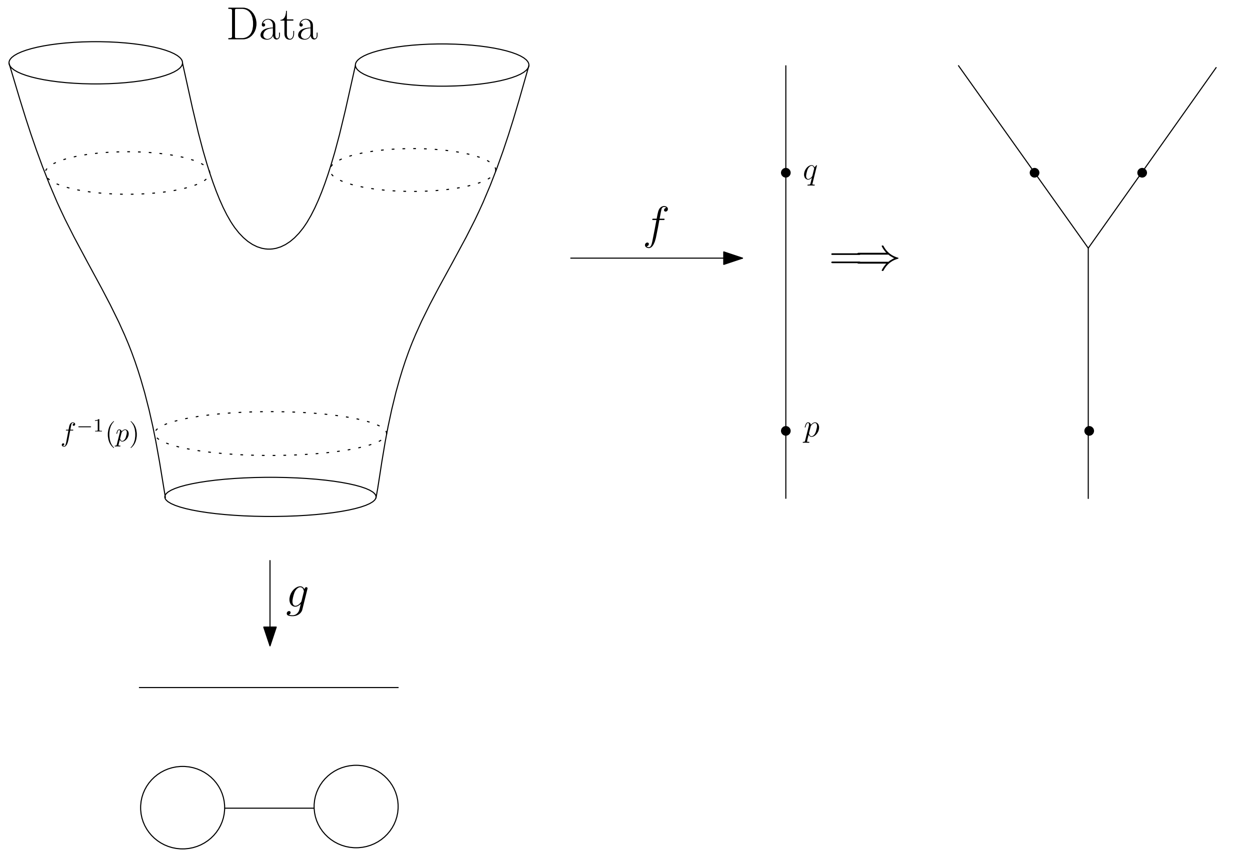
Math World



Math World

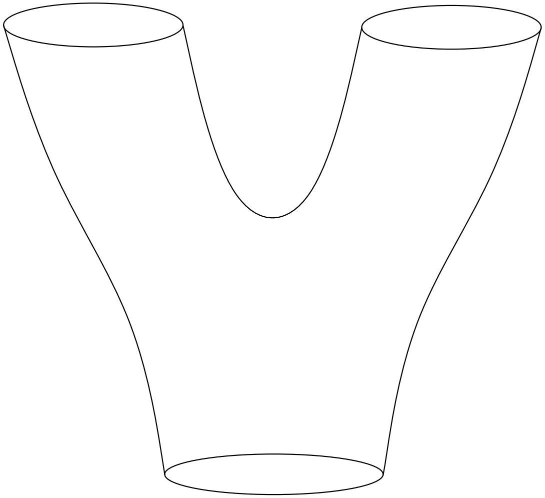


Math World



Exercise:

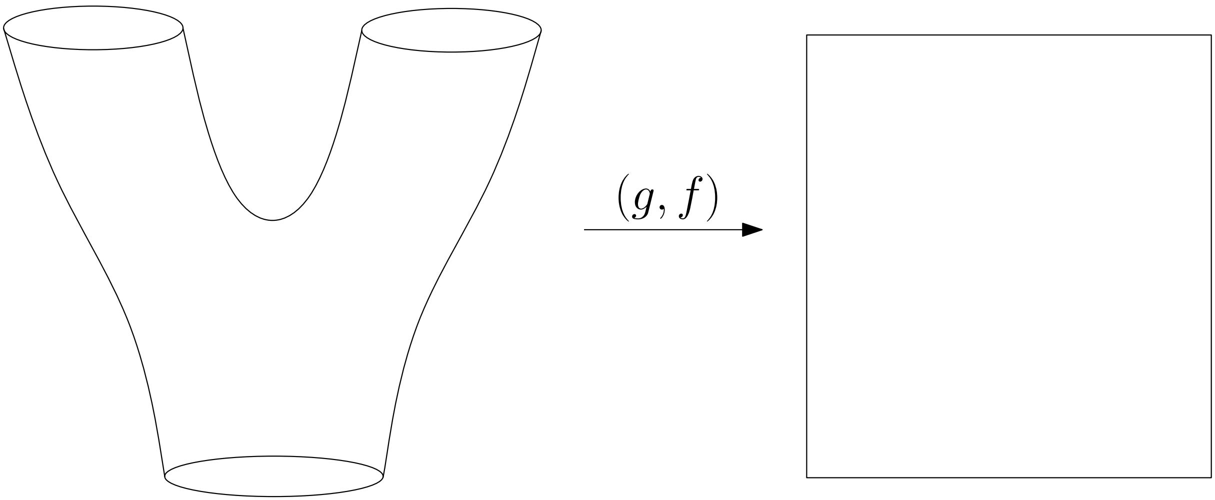
What is the summary if we use both lenses, *g* and *f* at the same time?



$$\xrightarrow{(g, f)}$$

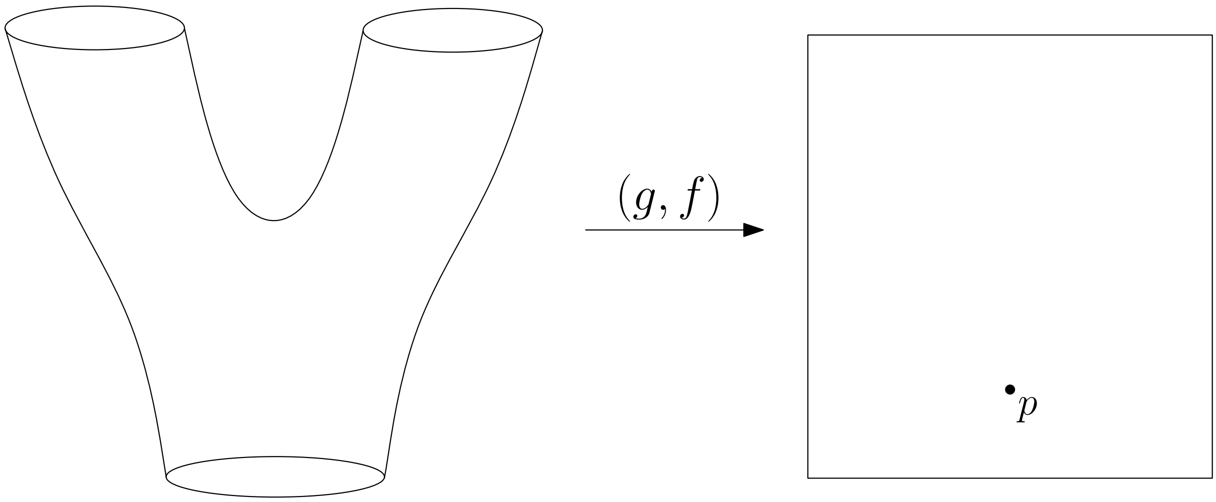
Exercise:

What is the summary if we use both lenses, *g* and *f* at the same time?



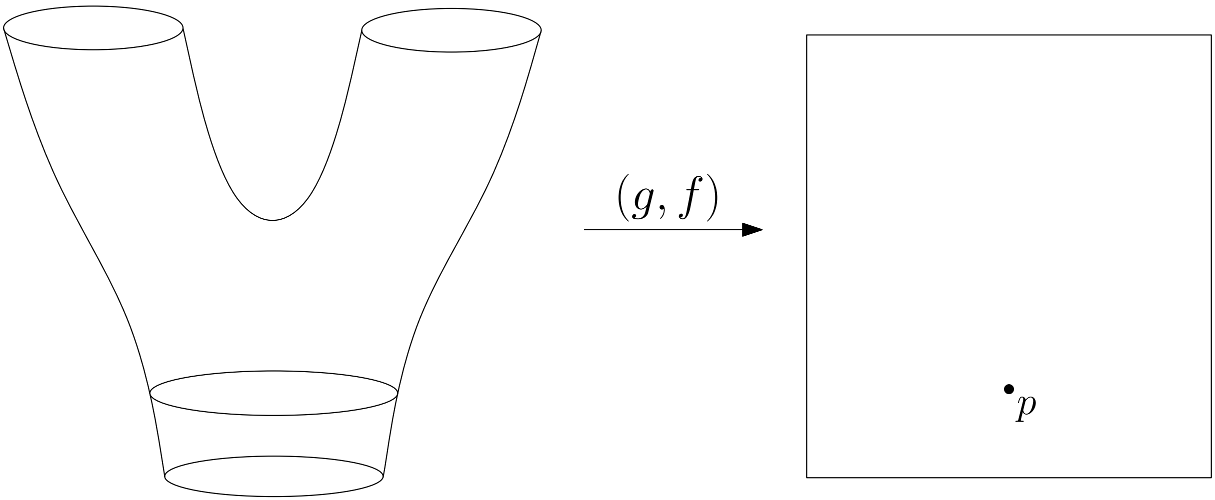
Exercise:

What is the summary if we use both lenses, *g* and *f* at the same time?



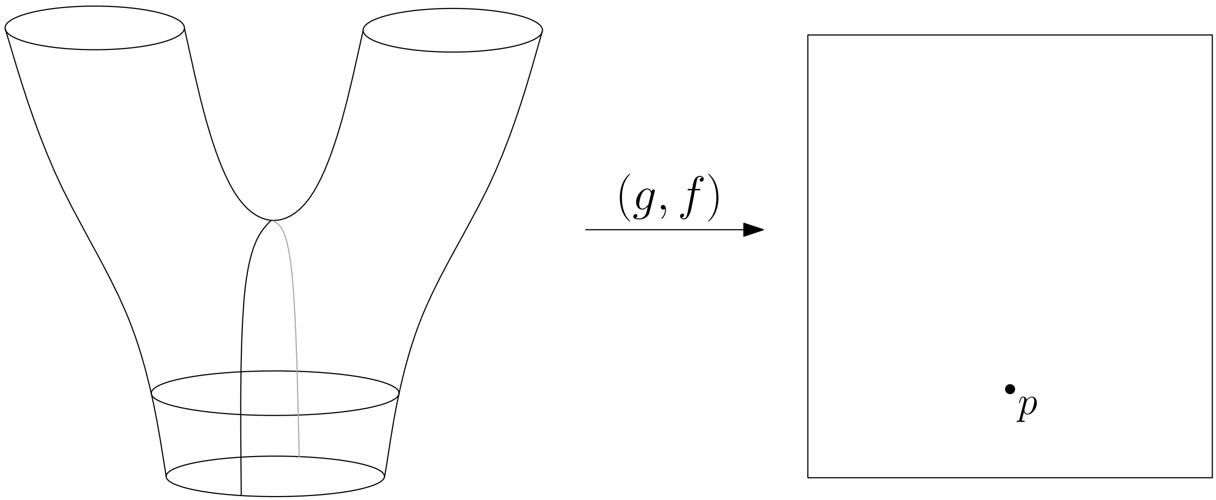
Exercise:

What is the summary if we use both lenses, g and f at the same time?



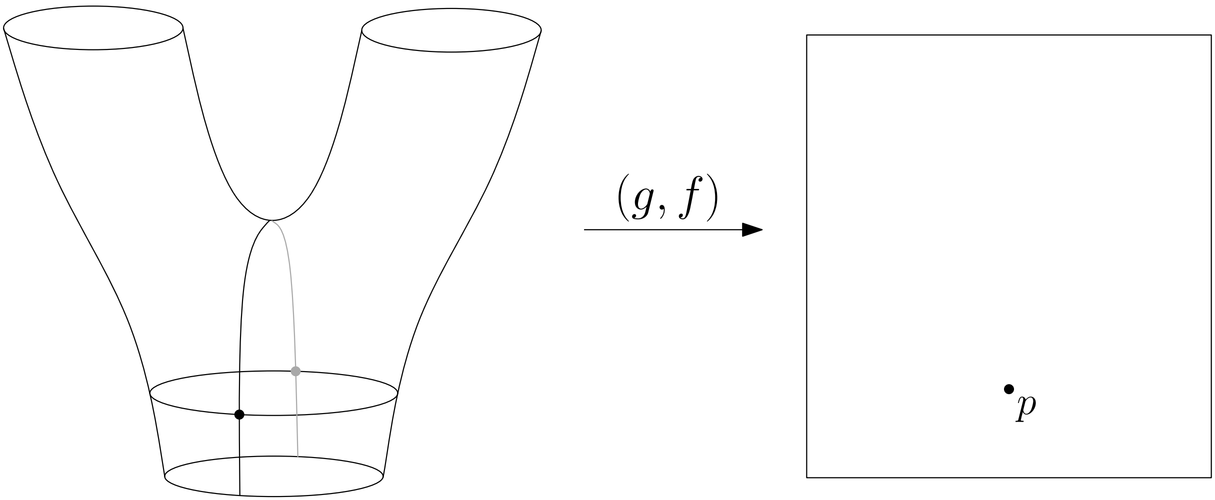
Exercise:

What is the summary if we use both lenses, *g* and *f* at the same time?



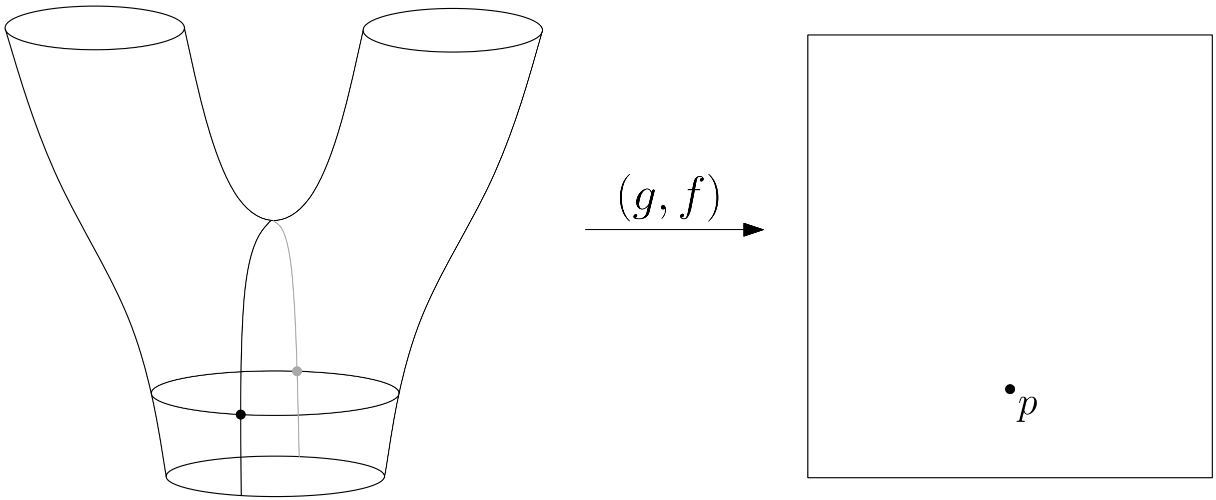
Exercise:

What is the summary if we use both lenses, *g* and *f* at the same time?



Exercise:

What is the summary if we use both lenses, g and f at the same time?



$\Longrightarrow$ We recover the original space

What did the exercise tell us?

What did the exercise tell us?

- With a rich enough set of functions (lenses) we can recover the original space

What did the exercise tell us?

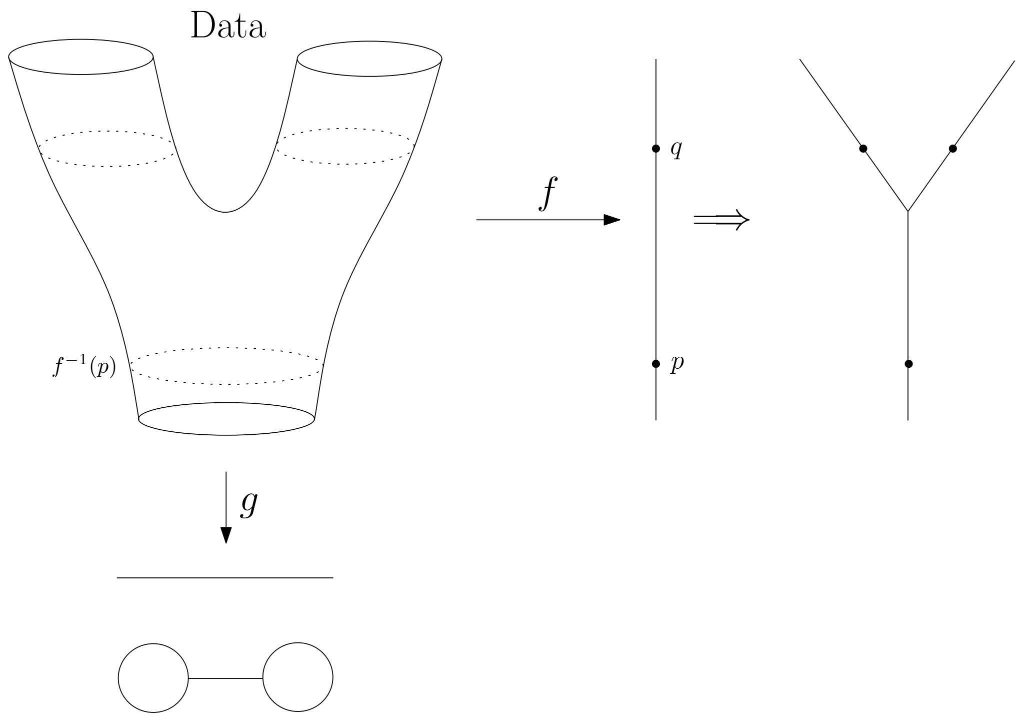
- With a rich enough set of functions (lenses) we can recover the original space
- Of course this leaves us no better off then where we started.

What did the exercise tell us?

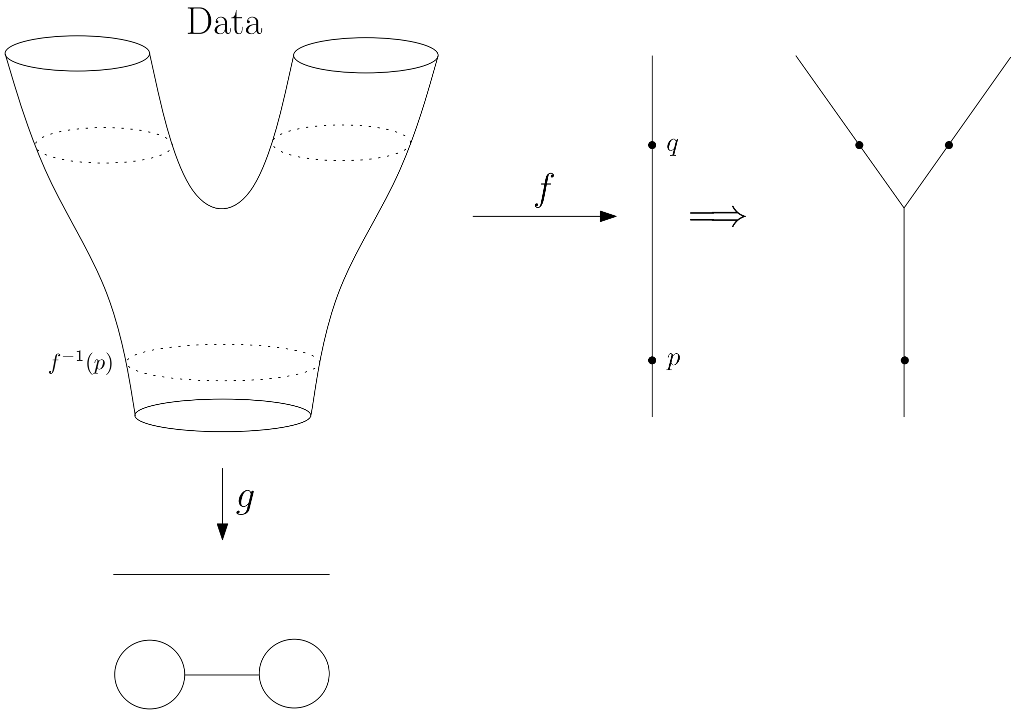
- With a rich enough set of functions (lenses) we can recover the original space
- Of course this leaves us no better off then where we started.

$\Rightarrow$ Instead we select a set of functions to *tune in* to the signal we want.

This is what Ayasdi does:



This is what Ayasdi does:



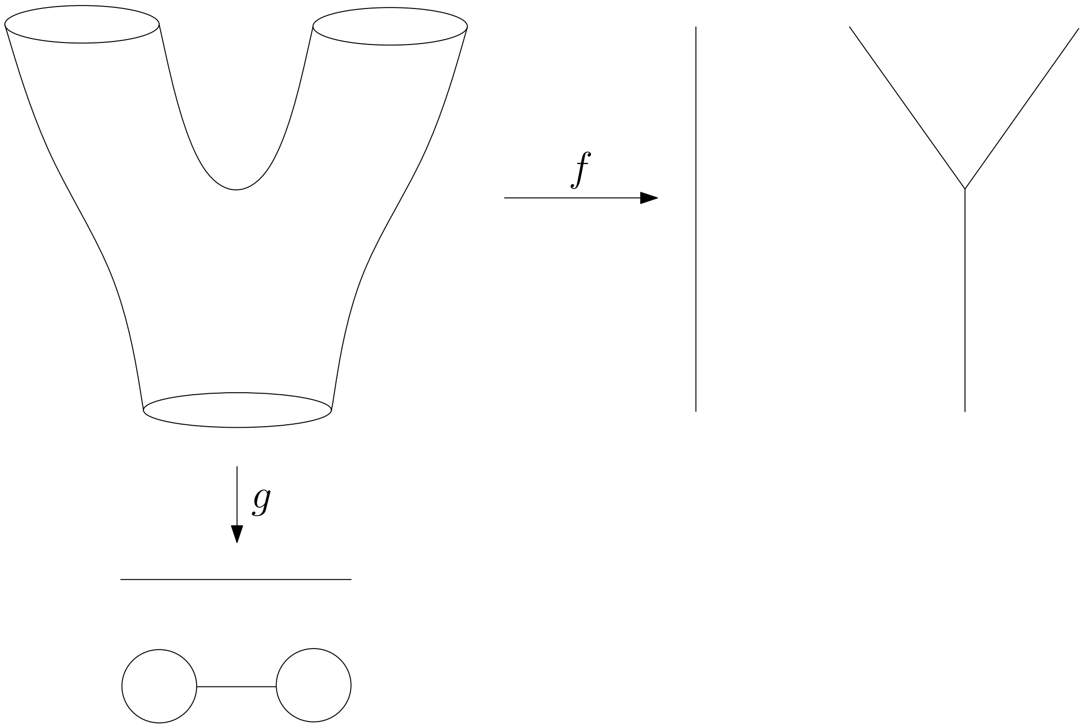
Modulo some details....

Why is this useful?

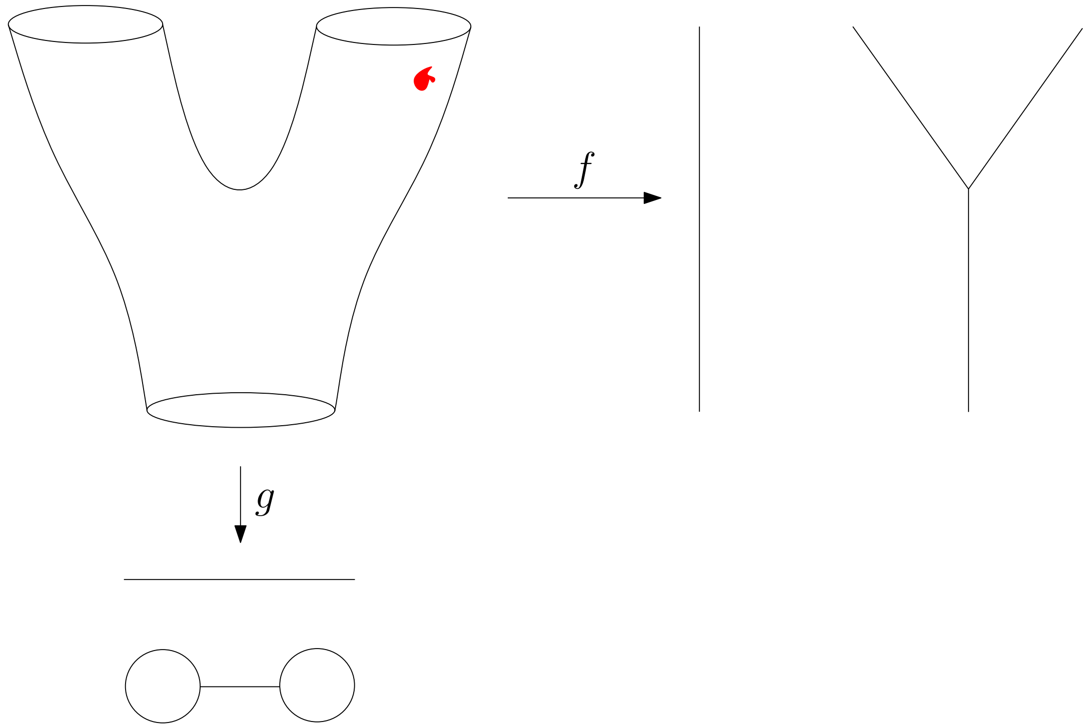
Why is this useful?

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

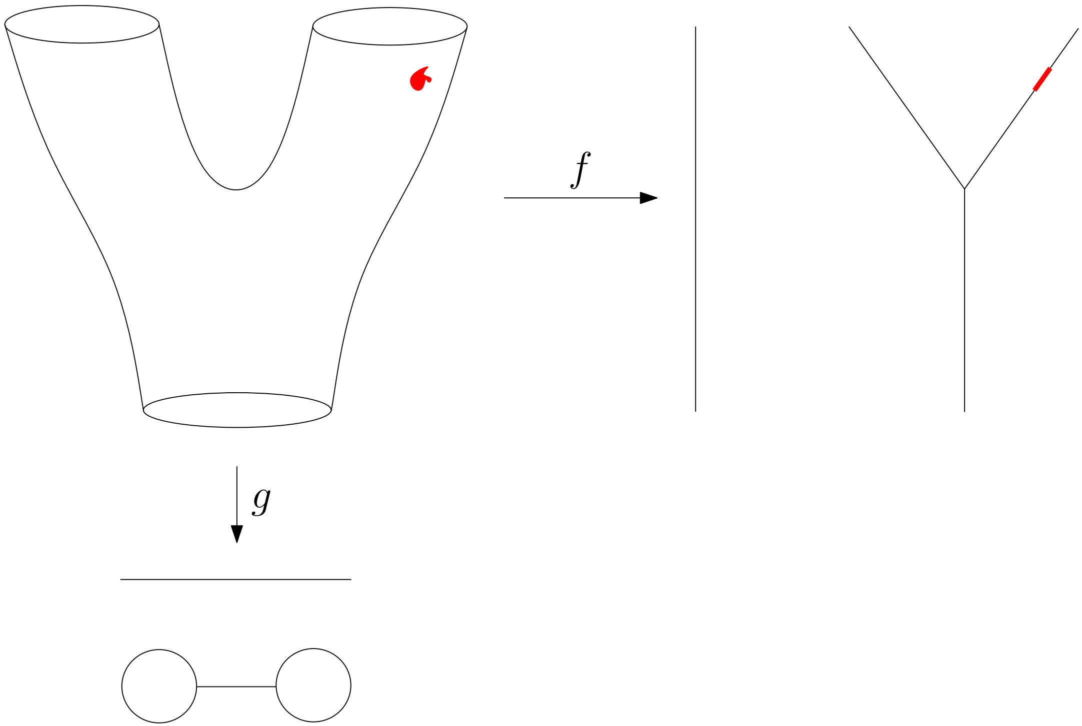
Why is this useful?



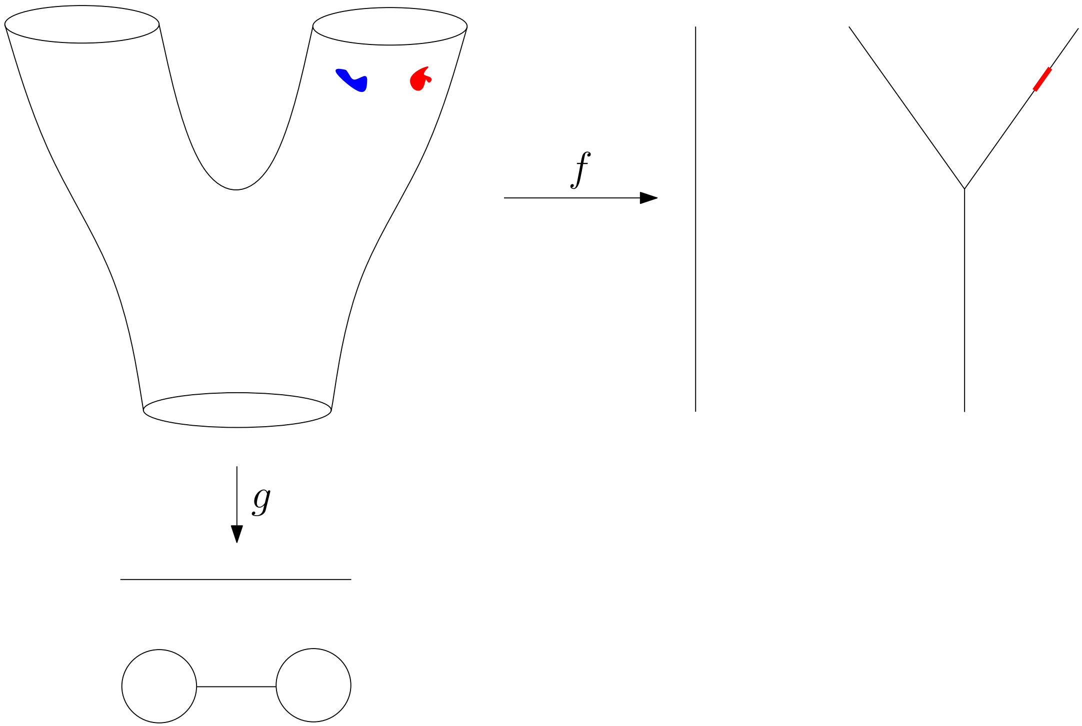
Why is this useful?



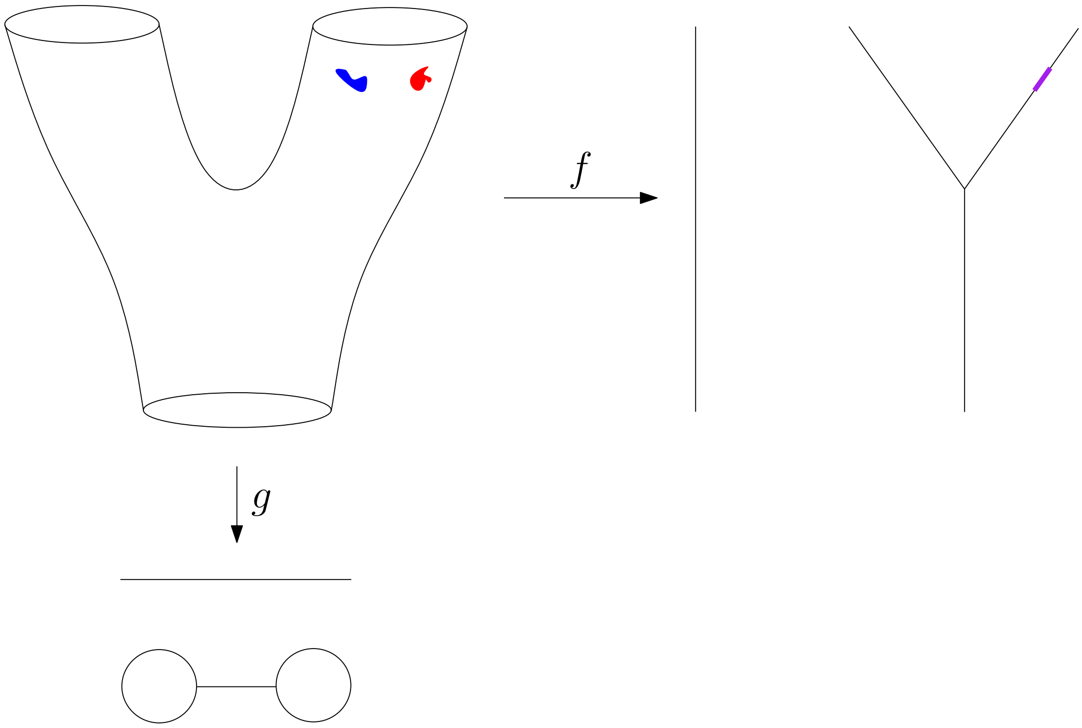
Why is this useful?



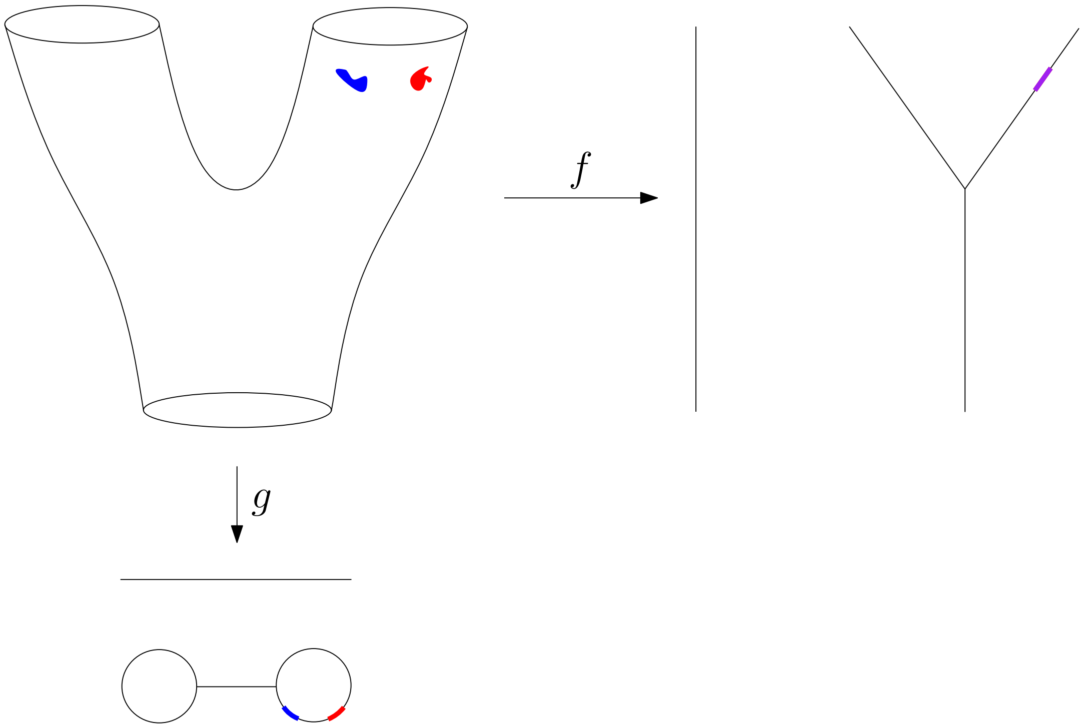
Why is this useful?



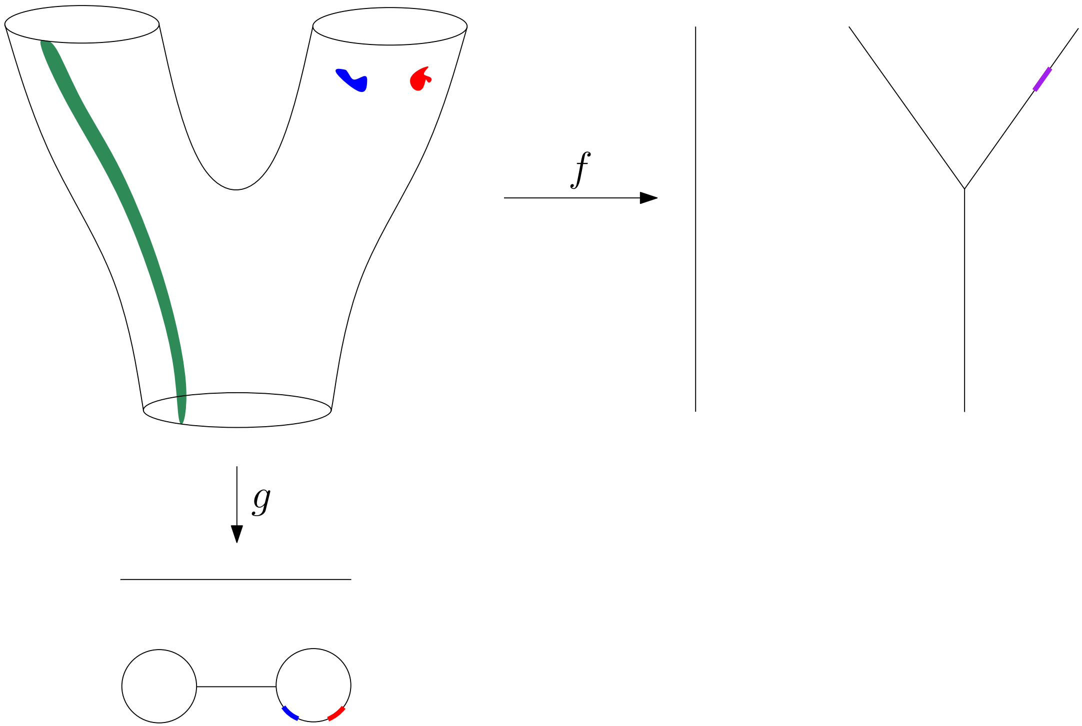
Why is this useful?



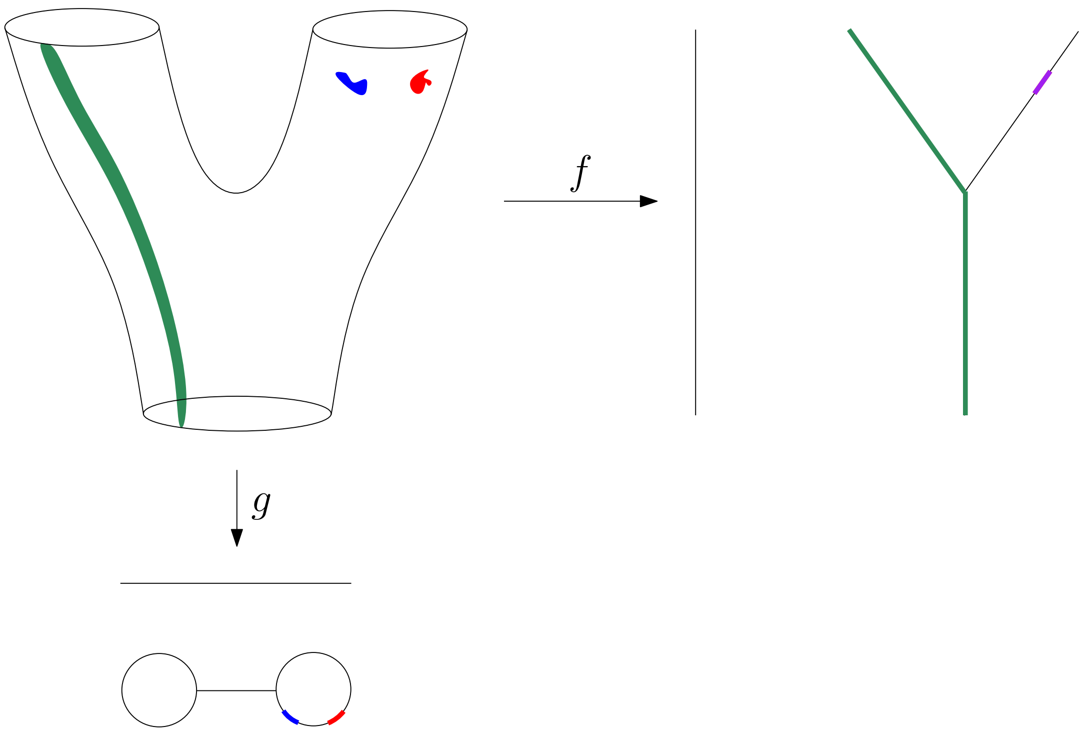
Why is this useful?



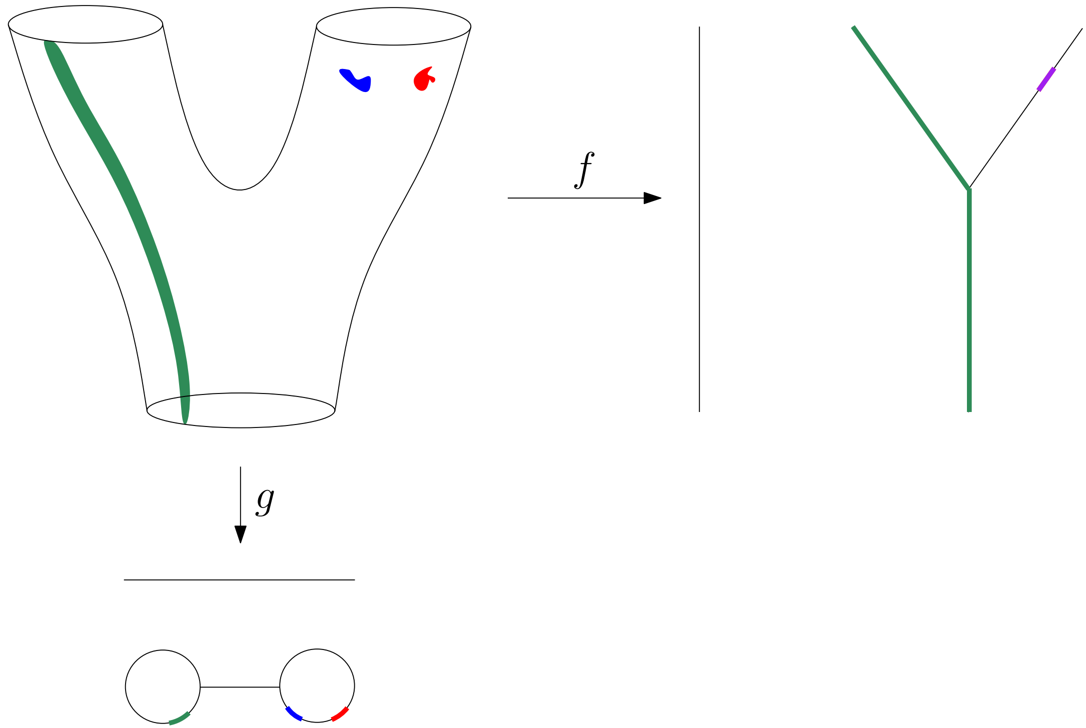
Why is this useful?



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.

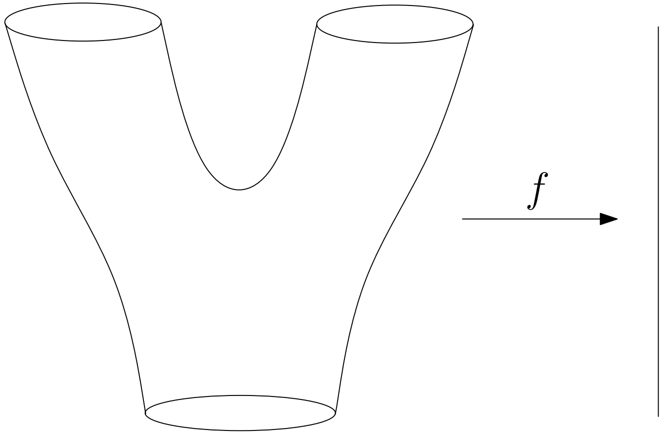
Why is this useful?

- 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.

Modulo Details....

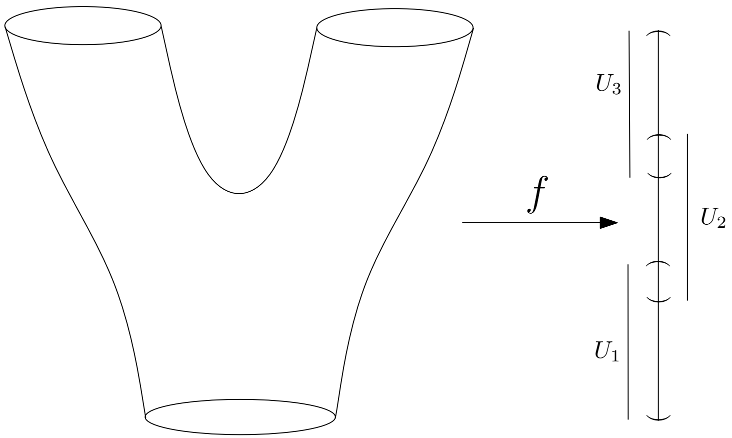
We want to move from this mathematical model to a data driven setup.

Step 1



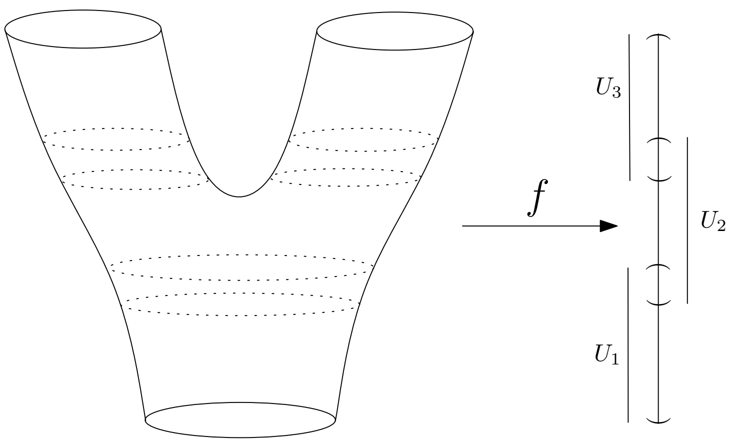
Step 1

- Replace points in the range with an open covering of the range.



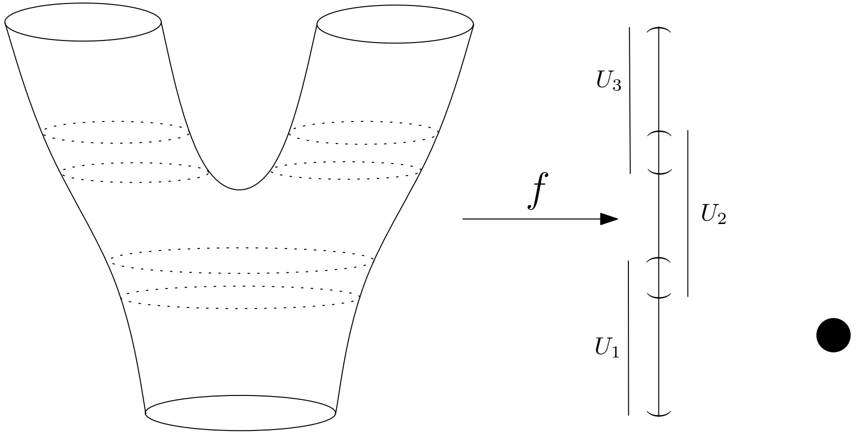
Step 1

- Replace points in the range with an open covering of the range.



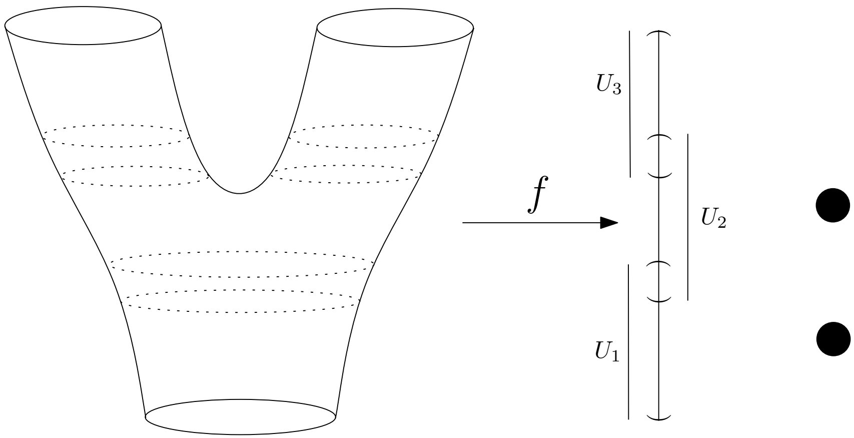
Step 1

- Replace points in the range with an open covering of the range.



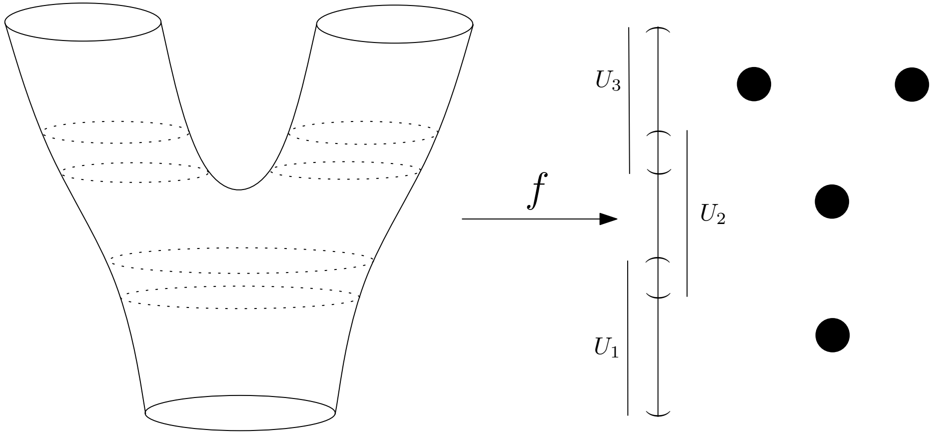
Step 1

- Replace points in the range with an open covering of the range.



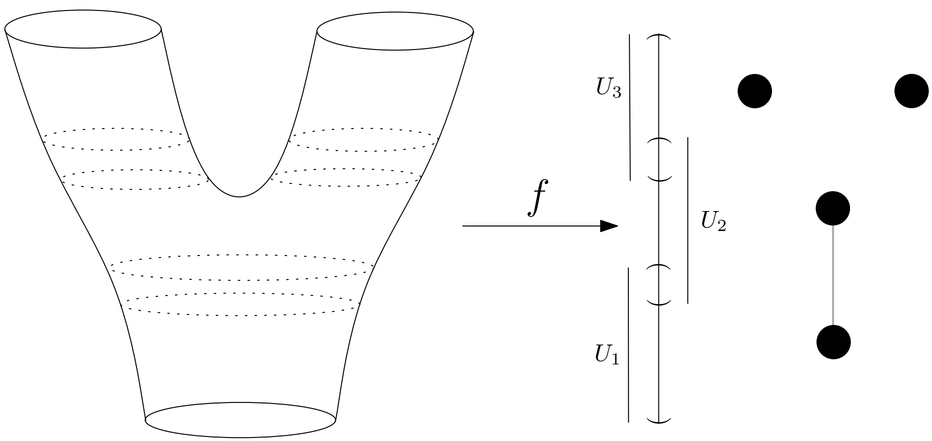
Step 1

- Replace points in the range with an open covering of the range.



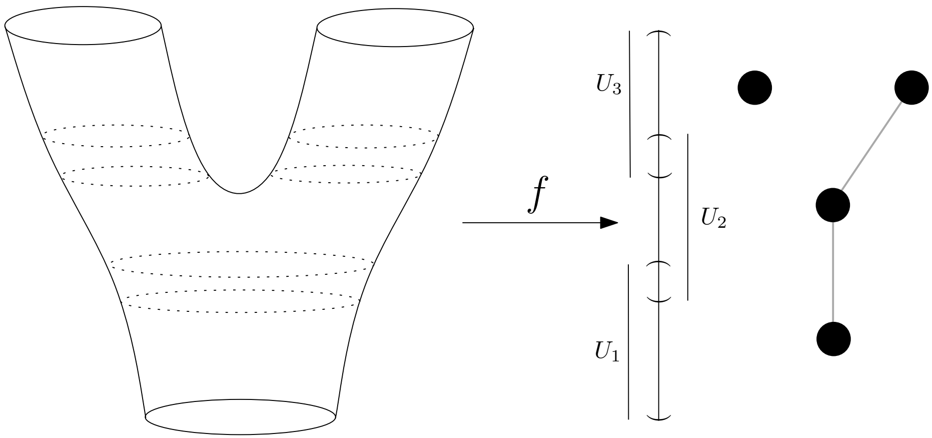
Step 1

- Replace points in the range with an open covering of the range.
- Connect nodes when their corresponding sets intersect.



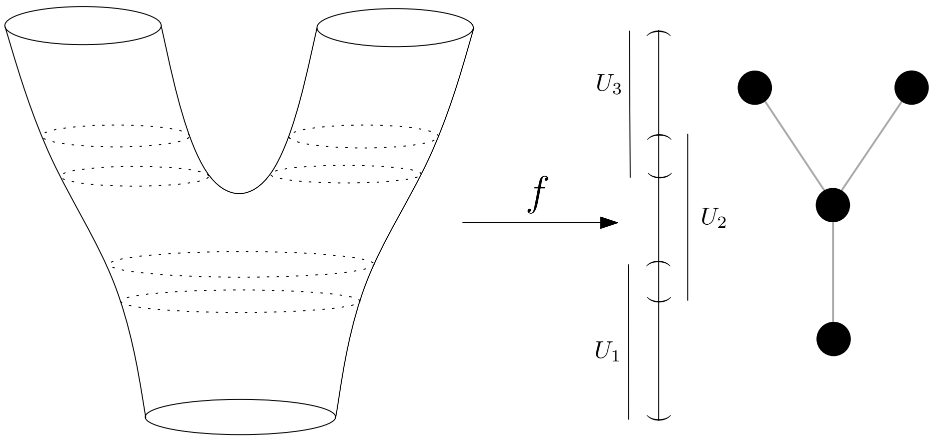
Step 1

- Replace points in the range with an open covering of the range.
- Connect nodes when their corresponding sets intersect.



Step 1

- Replace points in the range with an open covering of the range.
- Connect nodes when their corresponding sets intersect.



$\Rightarrow$ The output is now a graph.

New Parameters

We’ve introduced new parameters into the construction:

New Parameters

We've introduced new parameters into the construction:

- The **resolution** is the number of open sets in the range.

New Parameters

We’ve introduced new parameters into the construction:

- The **resolution** is the number of open sets in the range.
- The **gain** is the amount of overlap of these intervals.

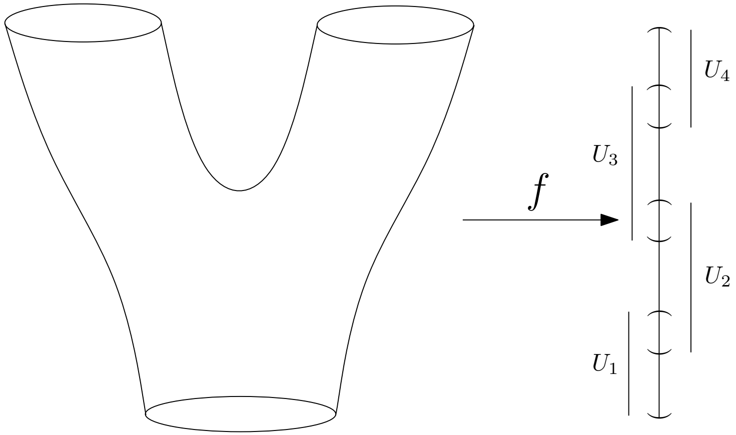
New Parameters

We've introduced new parameters into the construction:

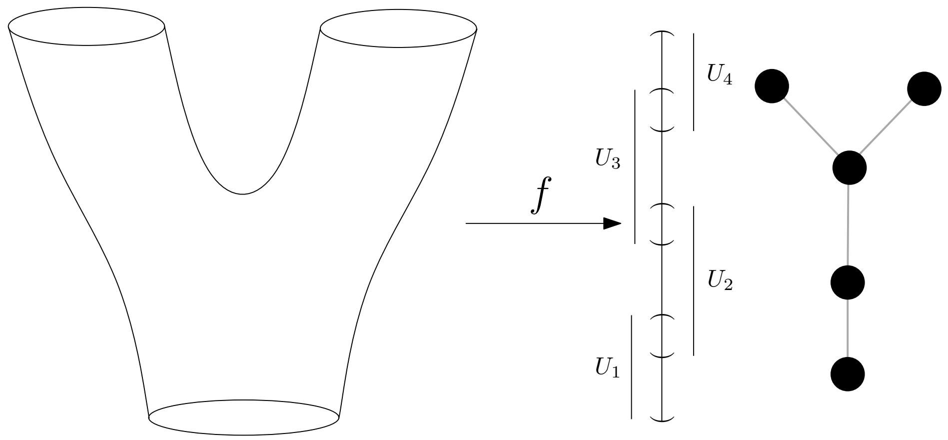
- The **resolution** is the number of open sets in the range.
- The **gain** is the amount of overlap of these intervals.

Roughly speaking, the resolution controls the number of nodes in the output and the 'size' of feature you can pick out, while the gain controls the number of edges and the 'tightness' of the graph.

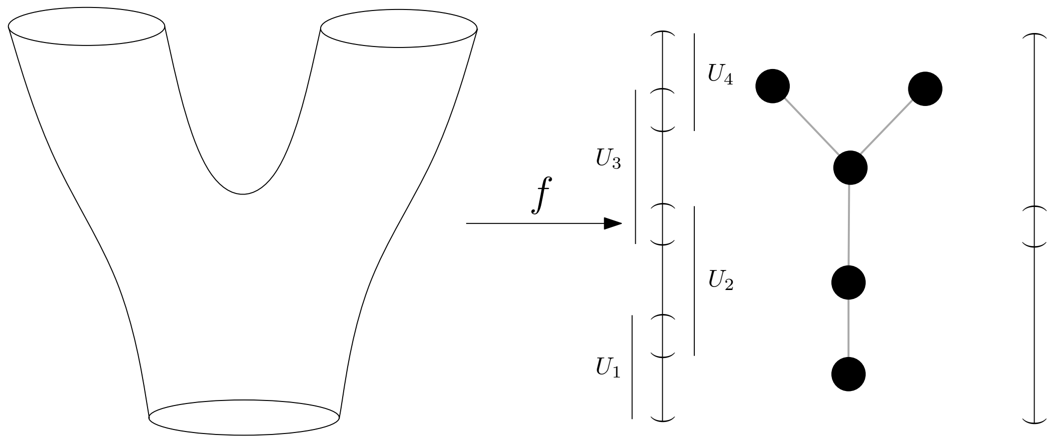
Resolution: A closer look



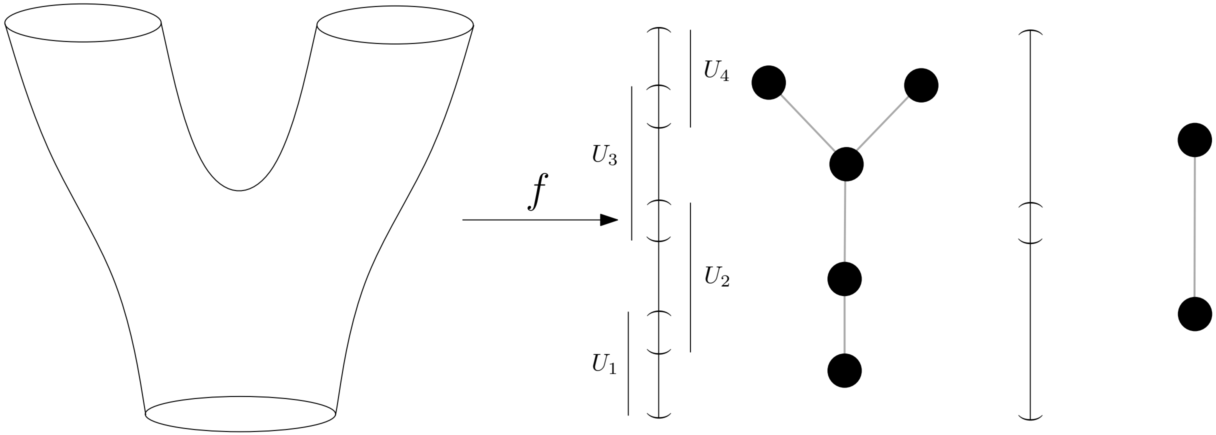
Resolution: A closer look



Resolution: A closer look



Resolution: A closer look



Step 2: Clustering as π_0

We need to make a final adjustment to the algorithm to bring it into data world.

Step 2: Clustering as π_0

We need to make a final adjustment to the algorithm to bring it into data world.

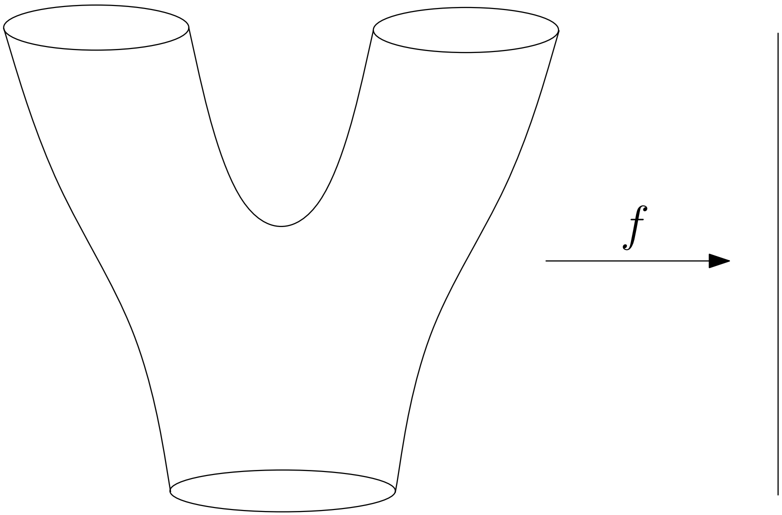
- We replace "connected component of the inverse image" is with "*clusters* in the inverse image".

Step 2: Clustering as π_0

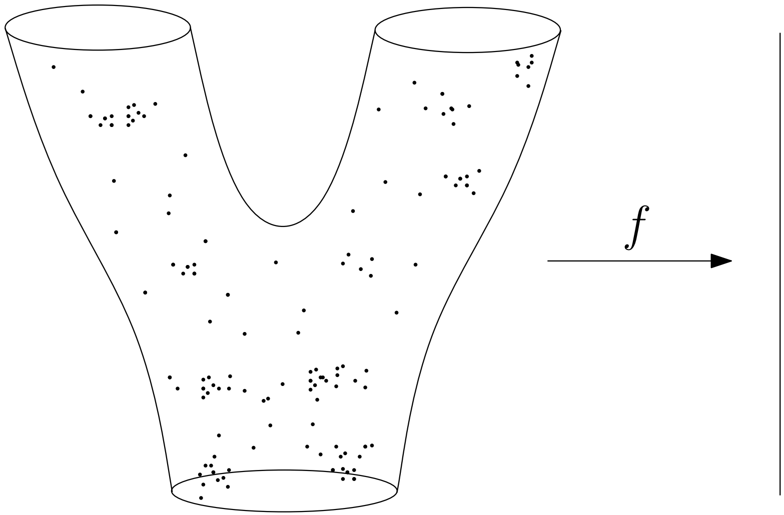
We need to make a final adjustment to the algorithm to bring it into data world.

- 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.

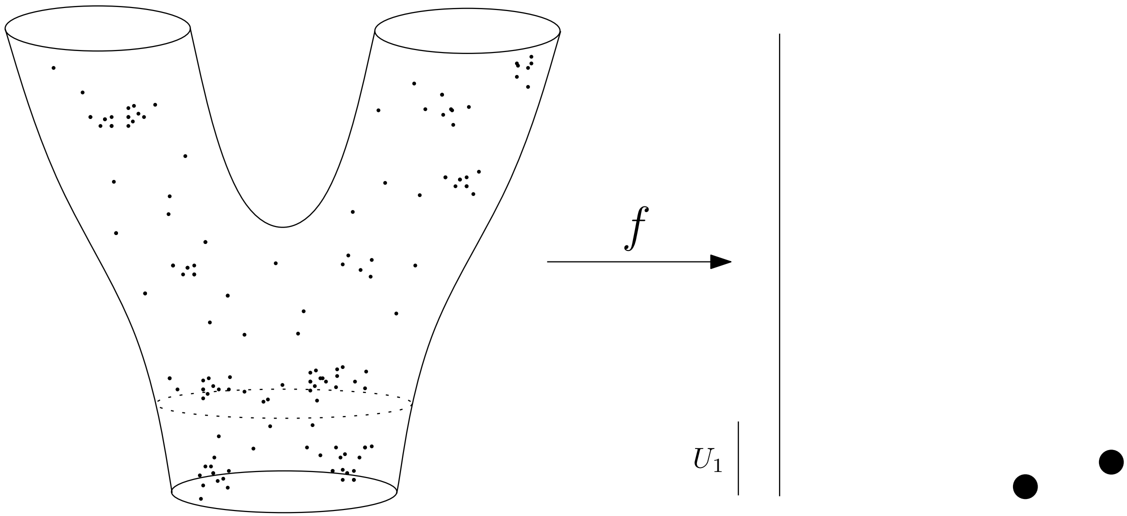
Step 2: Clustering as π_0



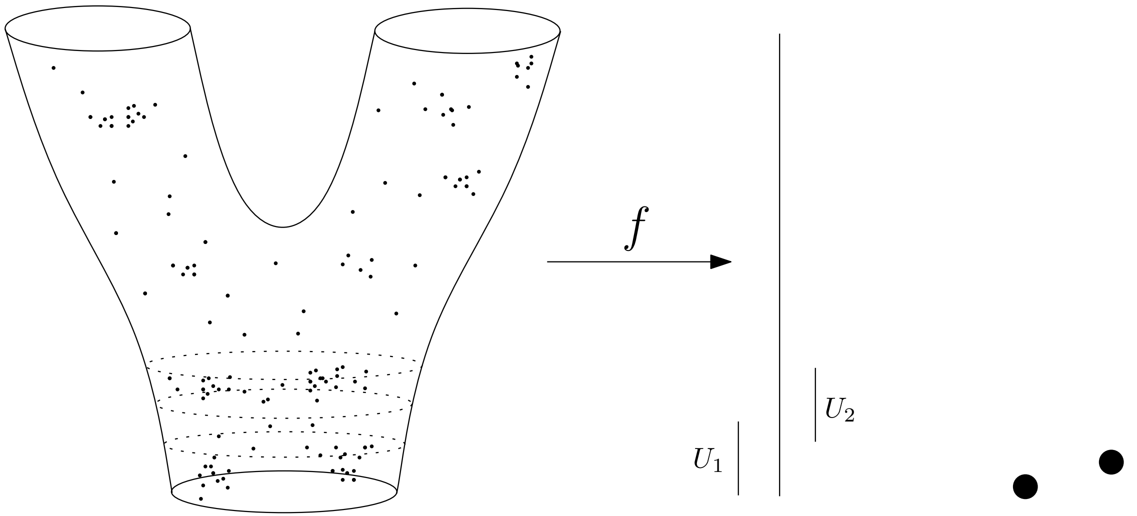
Step 2: Clustering as π_0



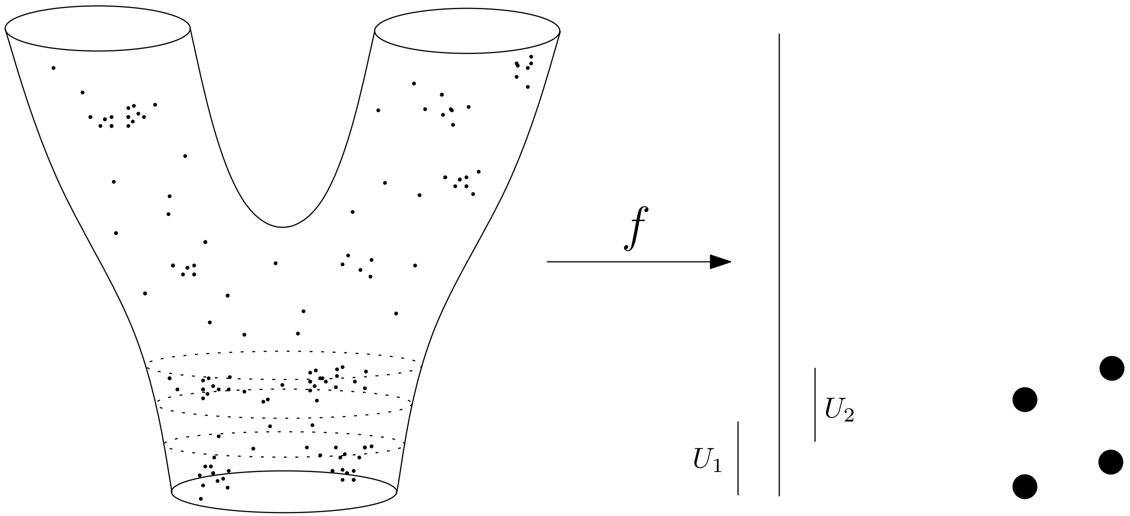
Step 2: Clustering as π_0



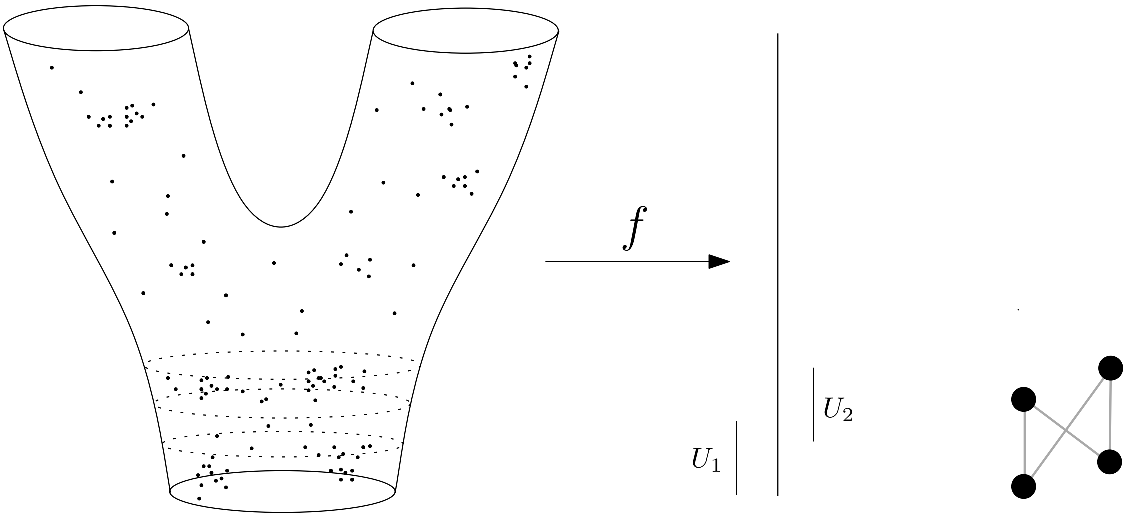
Step 2: Clustering as π_0



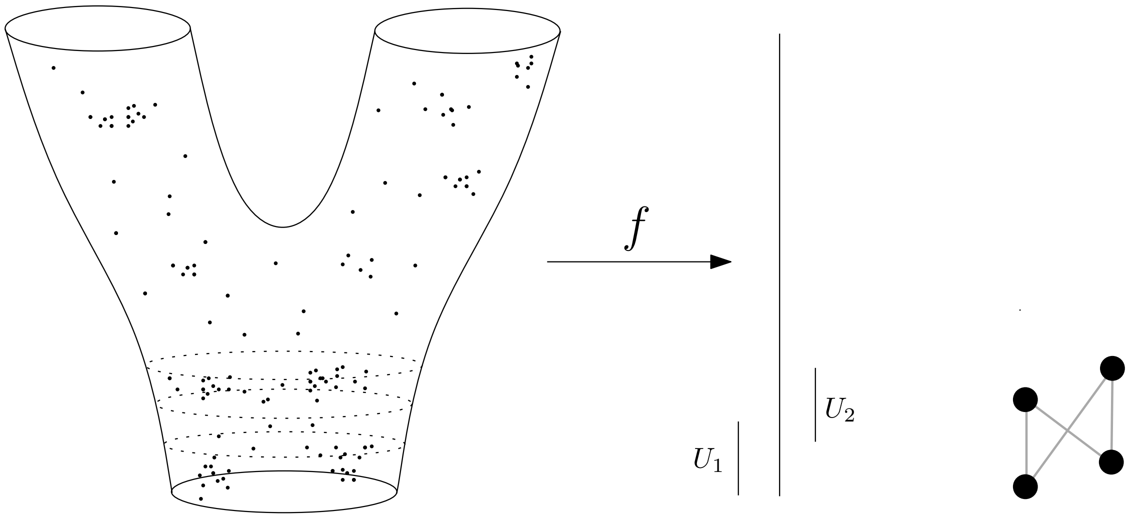
Step 2: Clustering as π_0



Step 2: Clustering as π_0

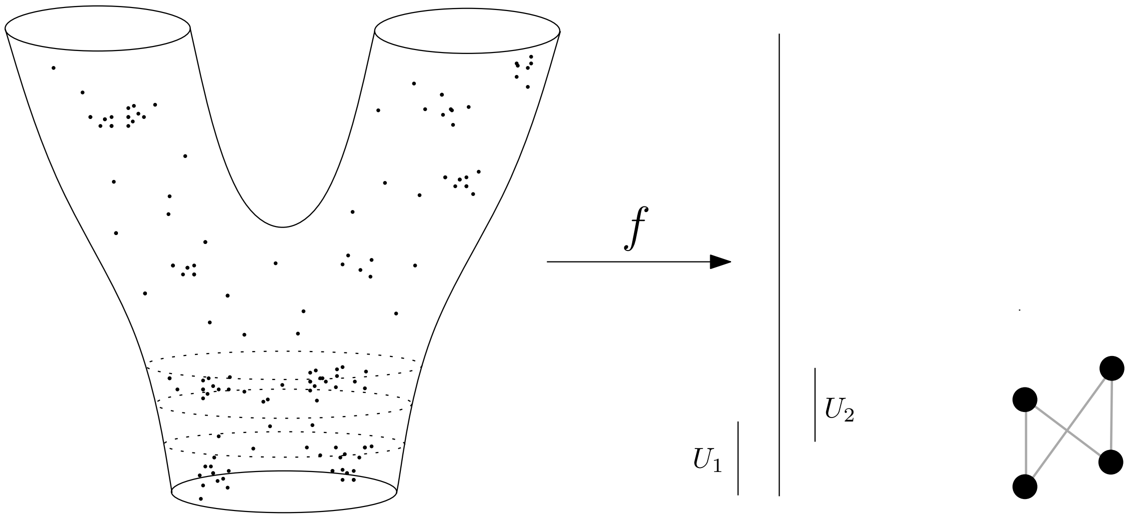


Step 2: Clustering as π_0



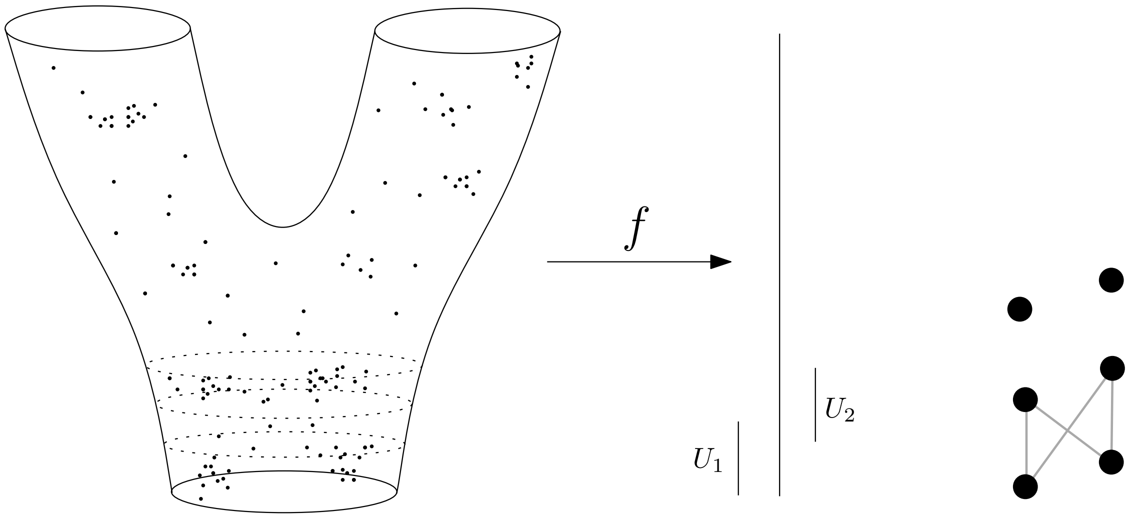
- Nodes are clusters of data points

Step 2: Clustering as π_0



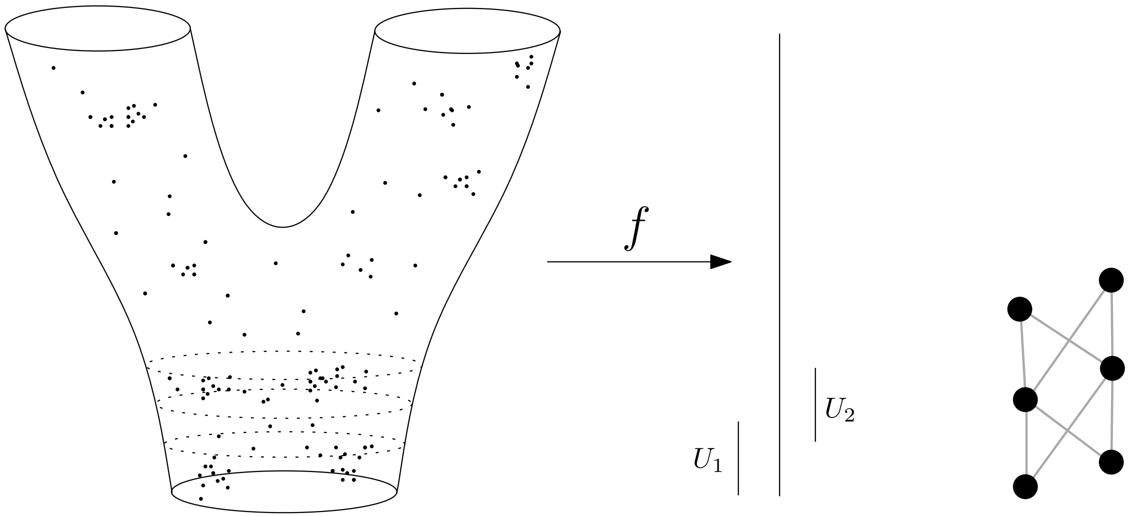
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



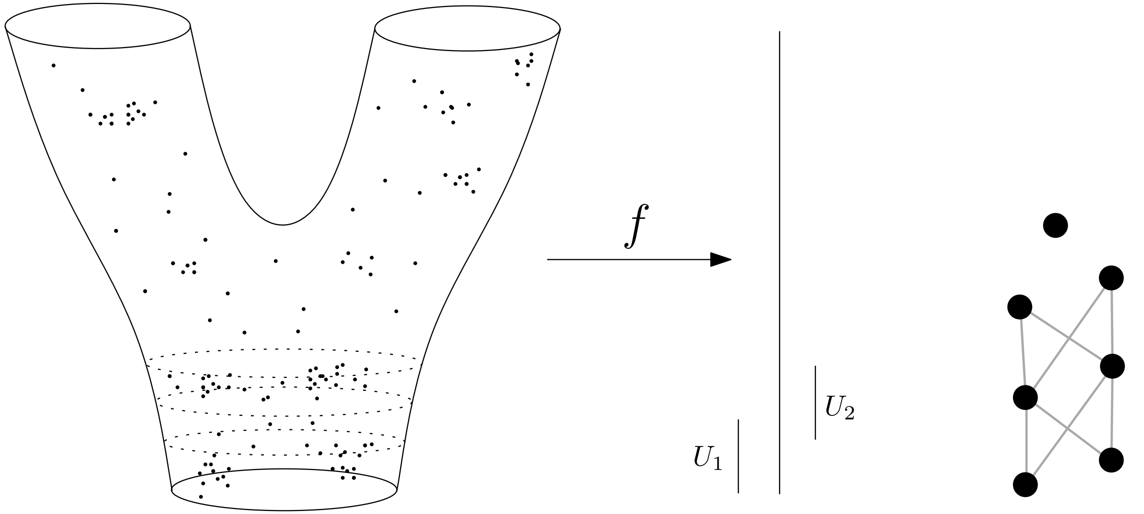
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



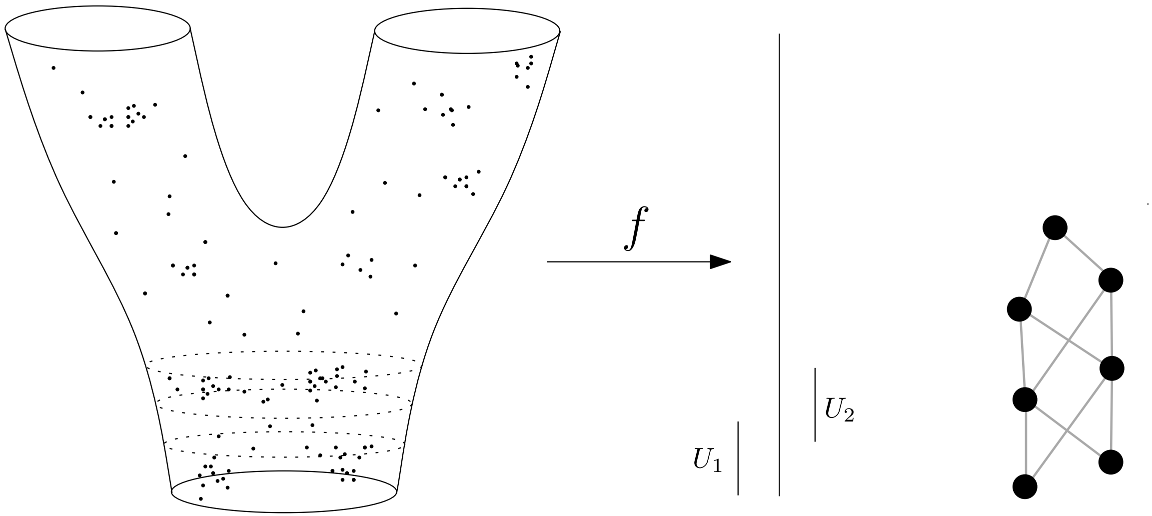
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



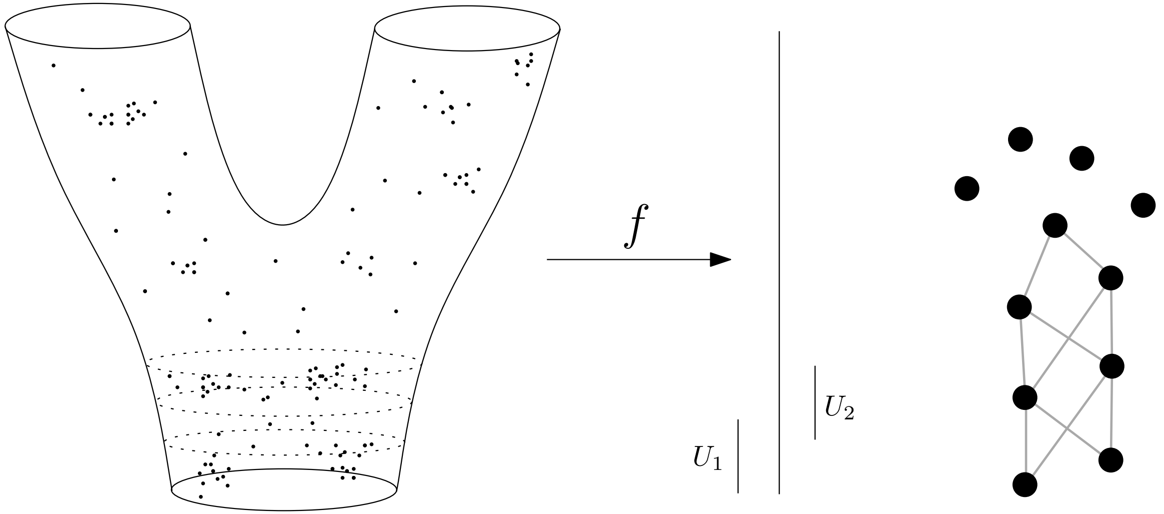
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



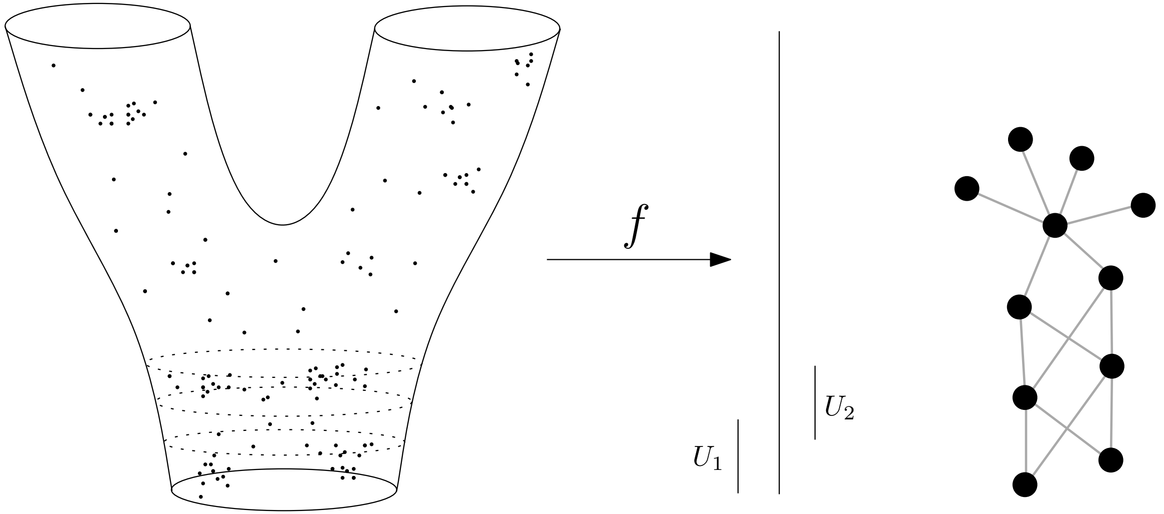
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



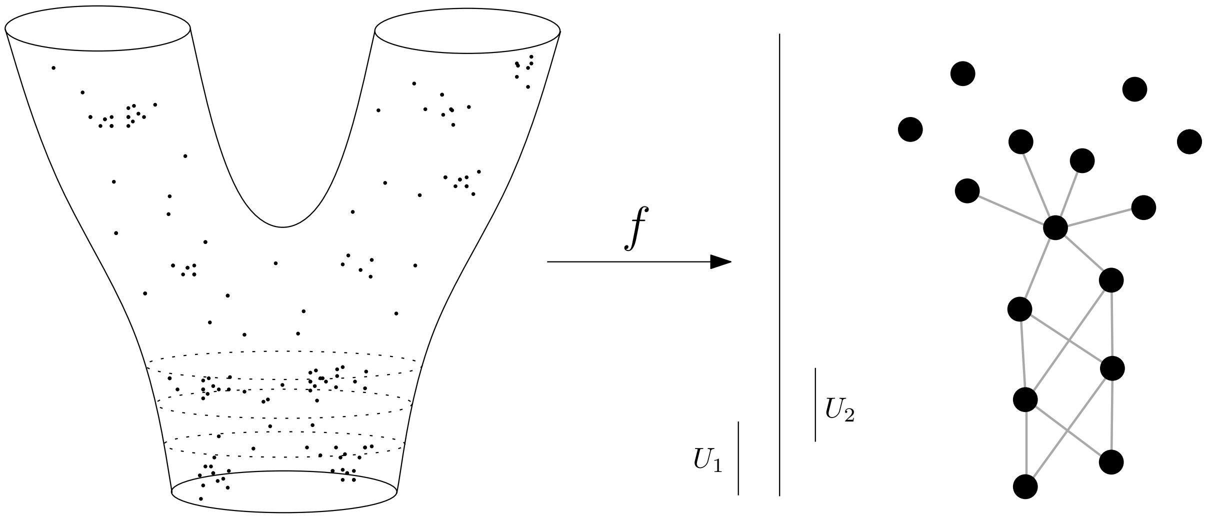
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



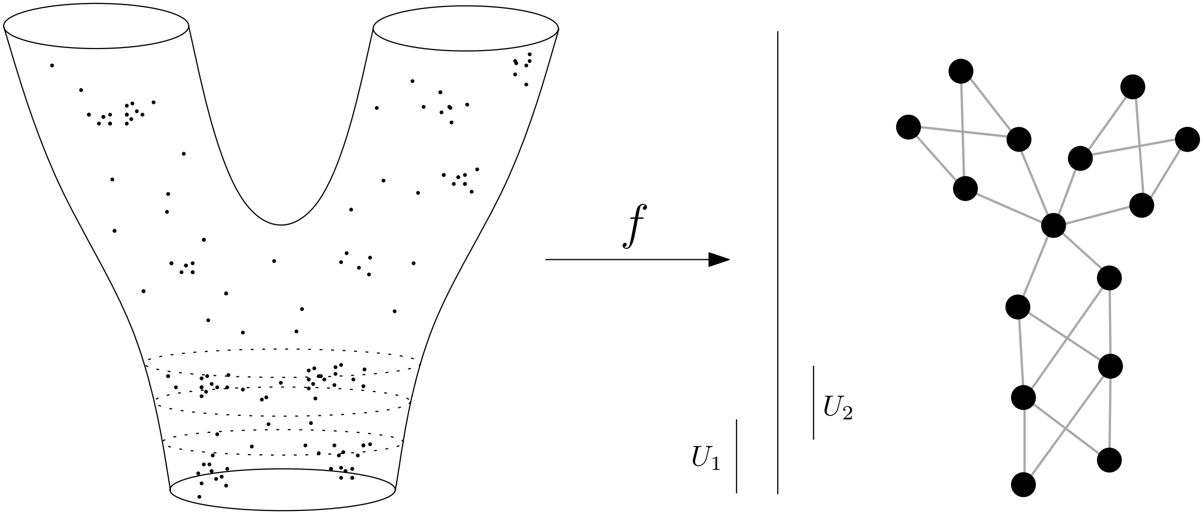
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



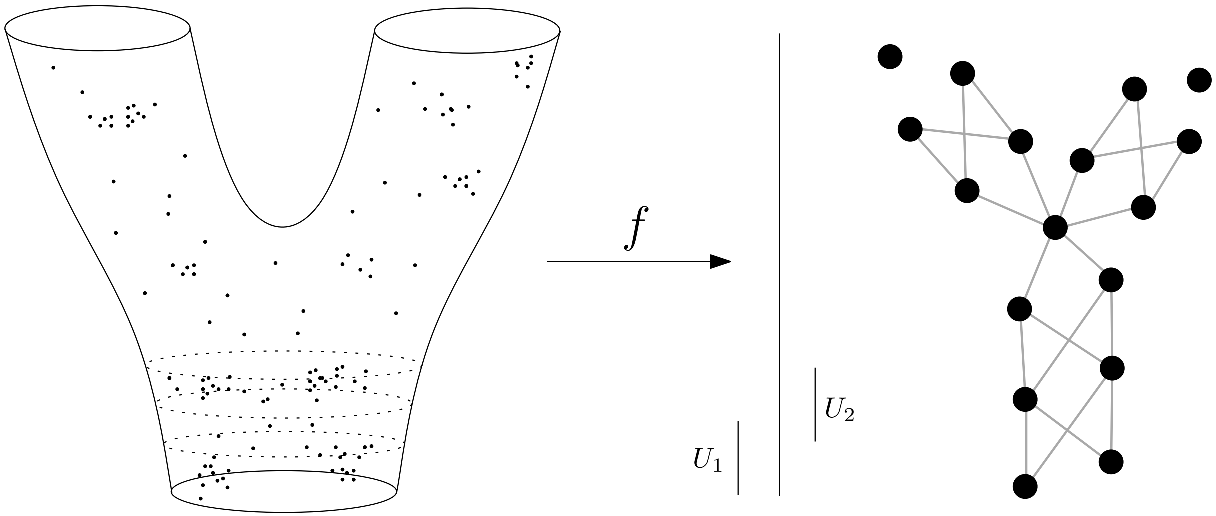
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



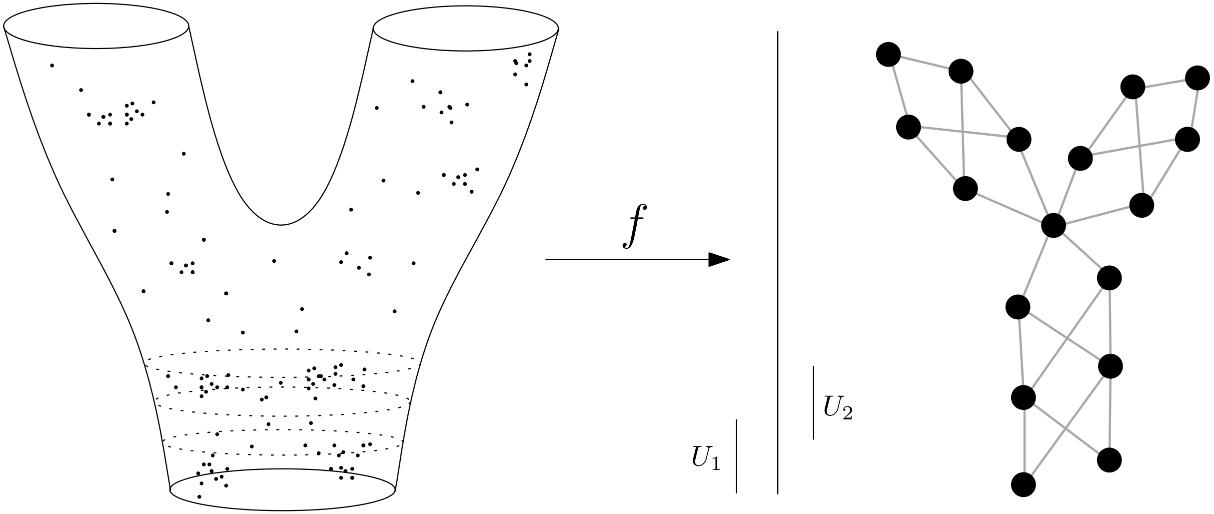
- Nodes are clusters of data points
- Edges represent shared points between the clusters

Step 2: Clustering as π_0



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

Step 2: Clustering as π_0



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

That’s It

That's It

Ok not quite...

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

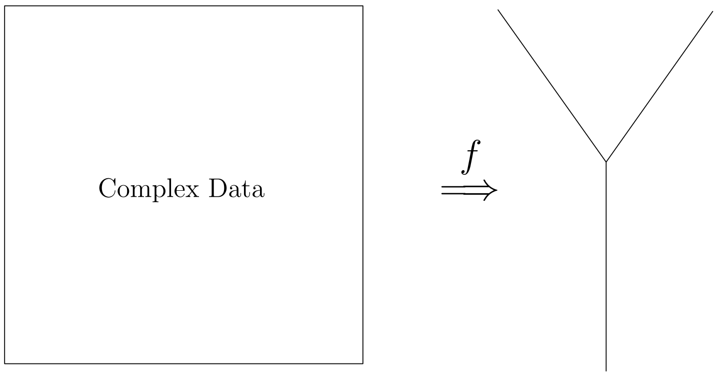
A Non Exhaustive Table of Lenses

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	
		...	

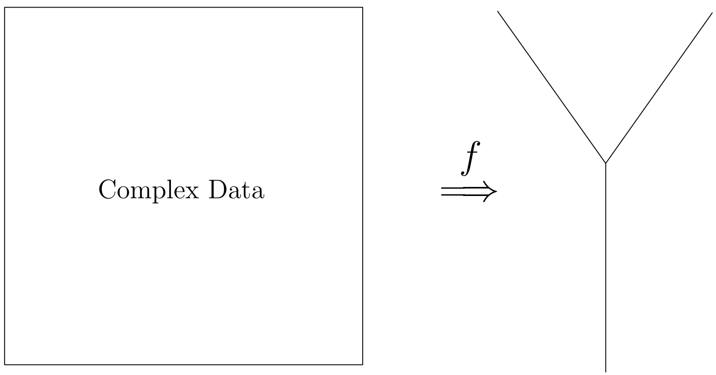
Interperability and Meaning

But what about insight? meaning?

Interperability and Meaning

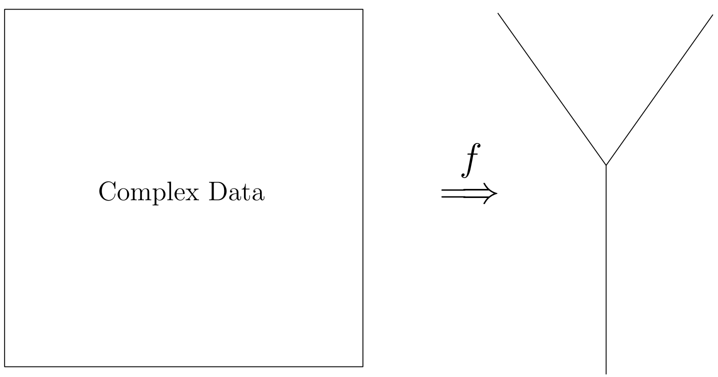


Interperability and Meaning



f is gaussian density

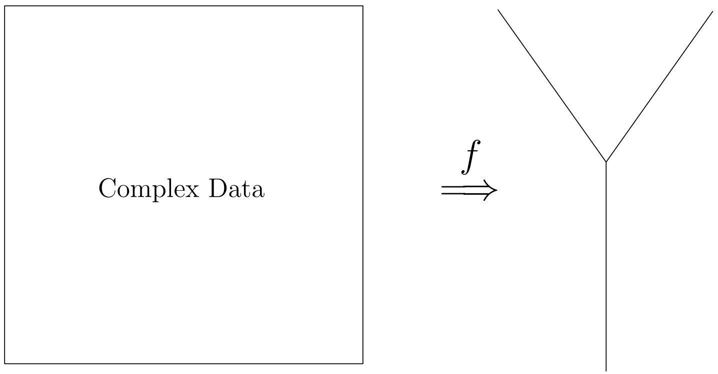
Interperability and Meaning



f is gaussian density

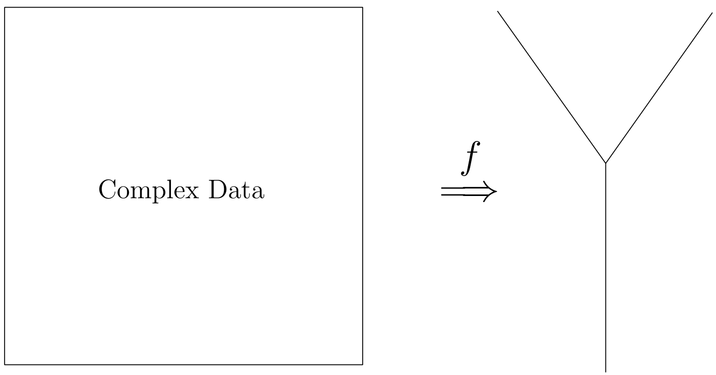
$\Rightarrow$ The data is bi-modal.

Interperability and Meaning



f is centrality

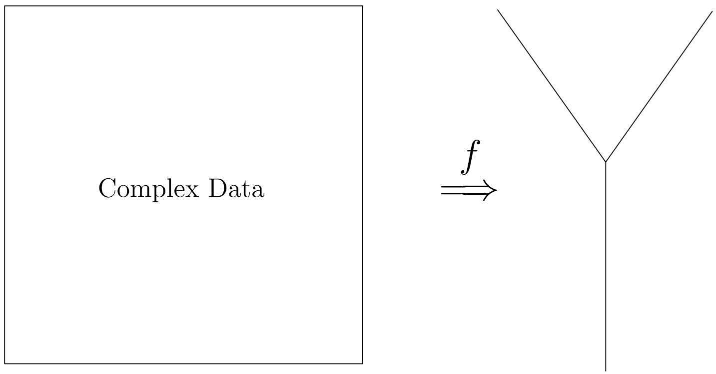
Interperability and Meaning



f is centrality

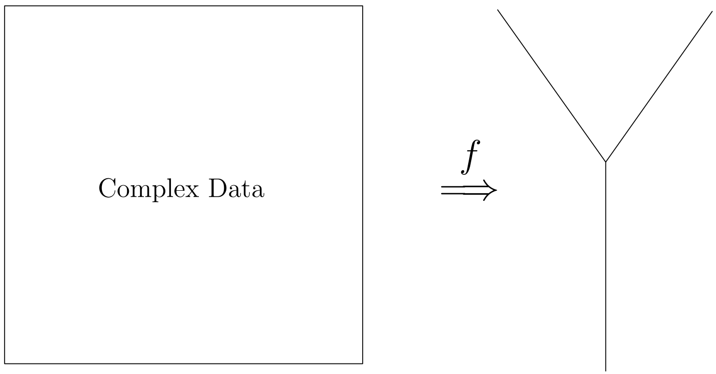
$\Rightarrow$ The data has two ways of being abnormal.

Interperability and Meaning



f is mean

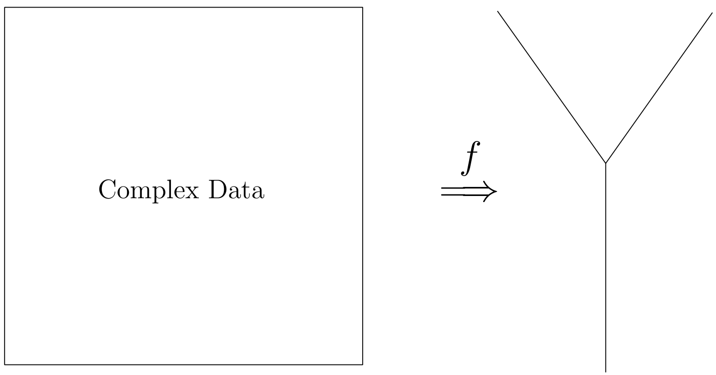
Interperability and Meaning



f is mean

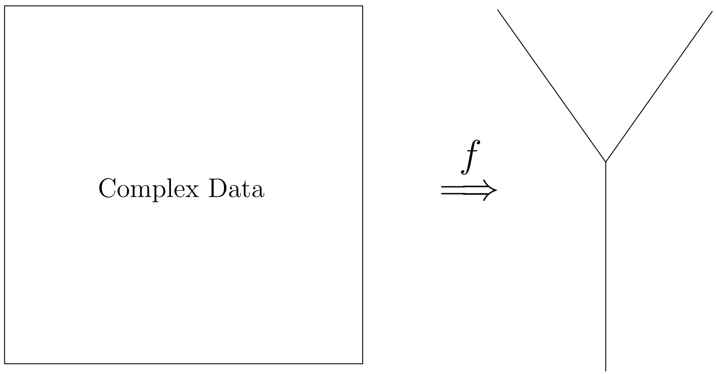
$\Rightarrow$ Two groups of high mean data.

Interperability and Meaning



f is error

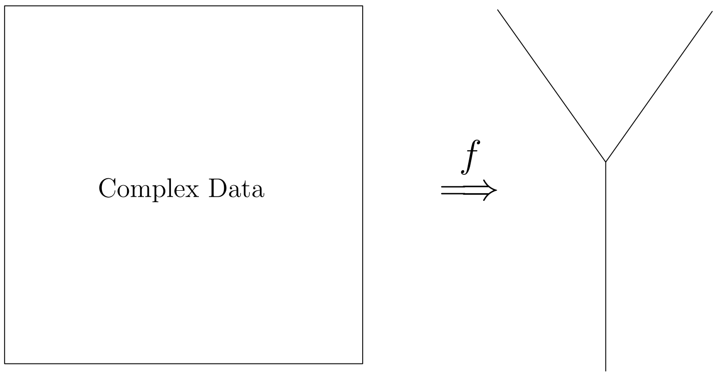
Interperability and Meaning



f is error

$\Rightarrow$ Two types of error.

Interperability and Meaning



f is error

$\Rightarrow$ Two types of error.

The units on the lens give interperability/meaning to the topological summary.

Interperability and Meaning

Another way to think about lenses is as a kind of ’geometric query’.

Examples

Interperability and Meaning

Another way to think about lenses is as a kind of ’geometric query’.

Examples

1. Heart disease study

Interperability and Meaning

Another way to think about lenses is as a kind of ’geometric query’.

Examples

1. Heart disease study
   - Stratification by age without making arbitrary cutoffs.

Interperability and Meaning

Another way to think about lenses is as a kind of ’geometric query’.

Examples

1. Heart disease study
   - Stratification by age without making arbitrary cutoffs.

2. Heavy machinery

Interperability and Meaning

Another way to think about lenses is as a kind of ’geometric query’.

Examples

1. Heart disease study
   - Stratification by age without making arbitrary cutoffs.
2. Heavy machinery
   - Use mean a variance as a lens to find what operating regimes lead to failure of mechanical components.

Some generalizations and extensions

Some generalizations and extensions

Metrics

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Some generalizations and extensions

Metrics

- We don't need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don’t need to be continuous - just "sensible".

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don’t need to be continuous - just "sensible".
- We can use multiple lenses at the same time.

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don’t need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don’t need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.
- In fact, can work with "open covers" of the space (Here taken to mean overlapping partitions). Don’t need a lens at all.

Some generalizations and extensions

Metrics

- We don’t need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don’t need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.
- In fact, can work with "open covers" of the space (Here taken to mean overlapping partitions). Don’t need a lens at all.

Data

Some generalizations and extensions

Metrics

- We don't need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don't need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.
- In fact, can work with "open covers" of the space (Here taken to mean overlapping partitions). Don't need a lens at all.

Data

- Input space can be anything with a topology. Typically we work with row/column numeric/categorical data but, for example, graphs are ok.

Some generalizations and extensions

Metrics

- We don't need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don't need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.
- In fact, can work with "open covers" of the space (Here taken to mean overlapping partitions). Don't need a lens at all.

Data

- Input space can be anything with a topology. Typically we work with row/column numeric/categorical data but, for example, graphs are ok.

Output

Some generalizations and extensions

Metrics

- We don't need a metric, just a notion of similarity - or perhaps a clustering mechanism.

Lenses

- Lenses don't need to be continuous - just "sensible".
- We can use multiple lenses at the same time.
- Lenses can map to space other then $\mathbb{R}$.
- In fact, can work with "open covers" of the space (Here taken to mean overlapping partitions). Don't need a lens at all.

Data

- Input space can be anything with a topology. Typically we work with row/column numeric/categorical data but, for example, graphs are ok.

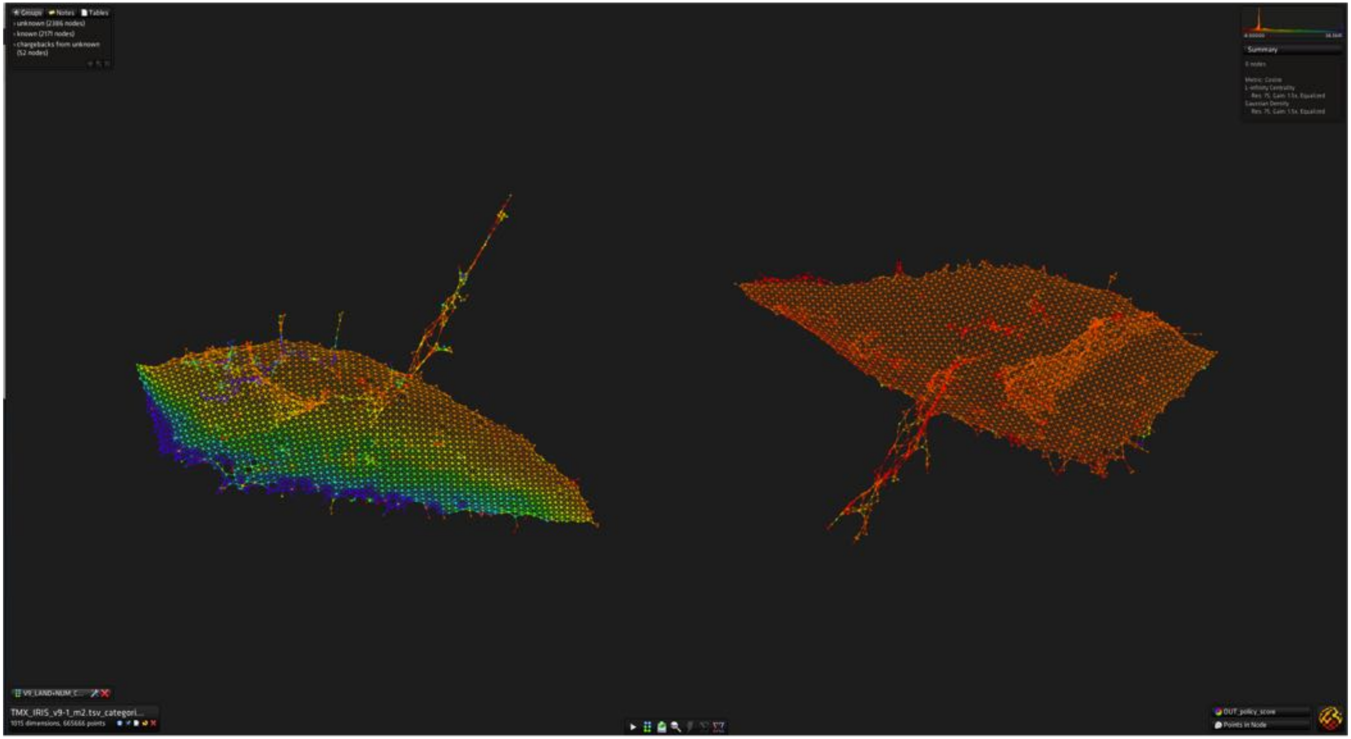
Output

- The output of the algorithm isn't just a graph but is an abstract simplicial complex (swept under the rug in this presentation).

Demo

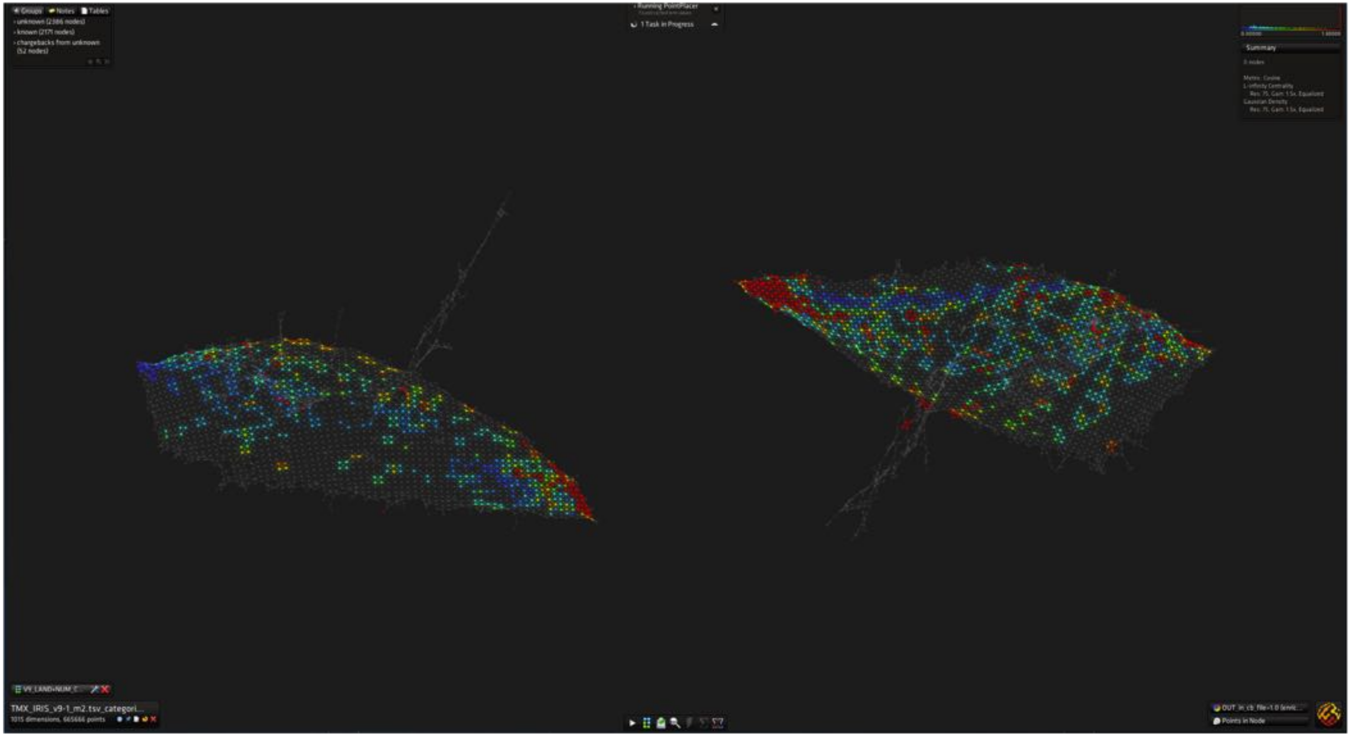
Online Fraud

Fraud Score



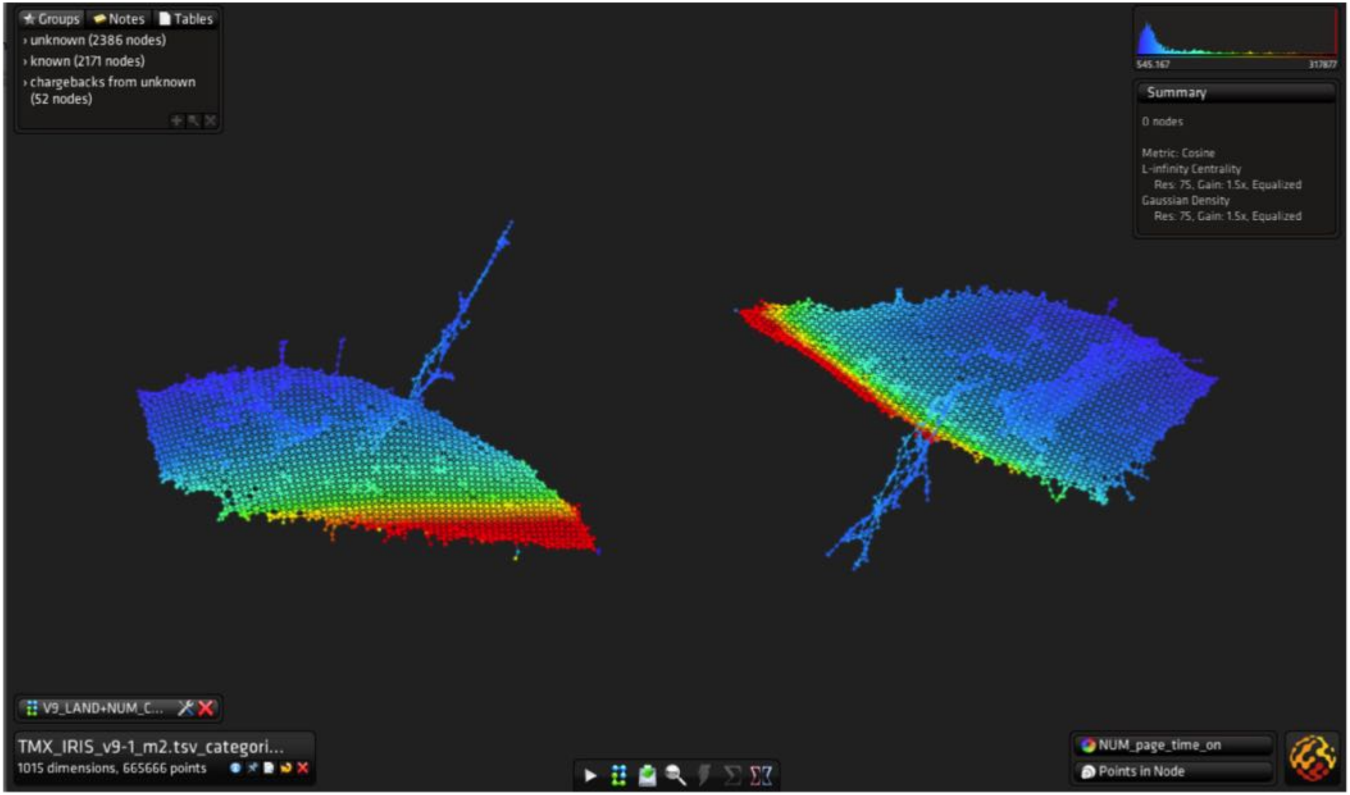
Online Fraud

Charge Back (Ground Truth)



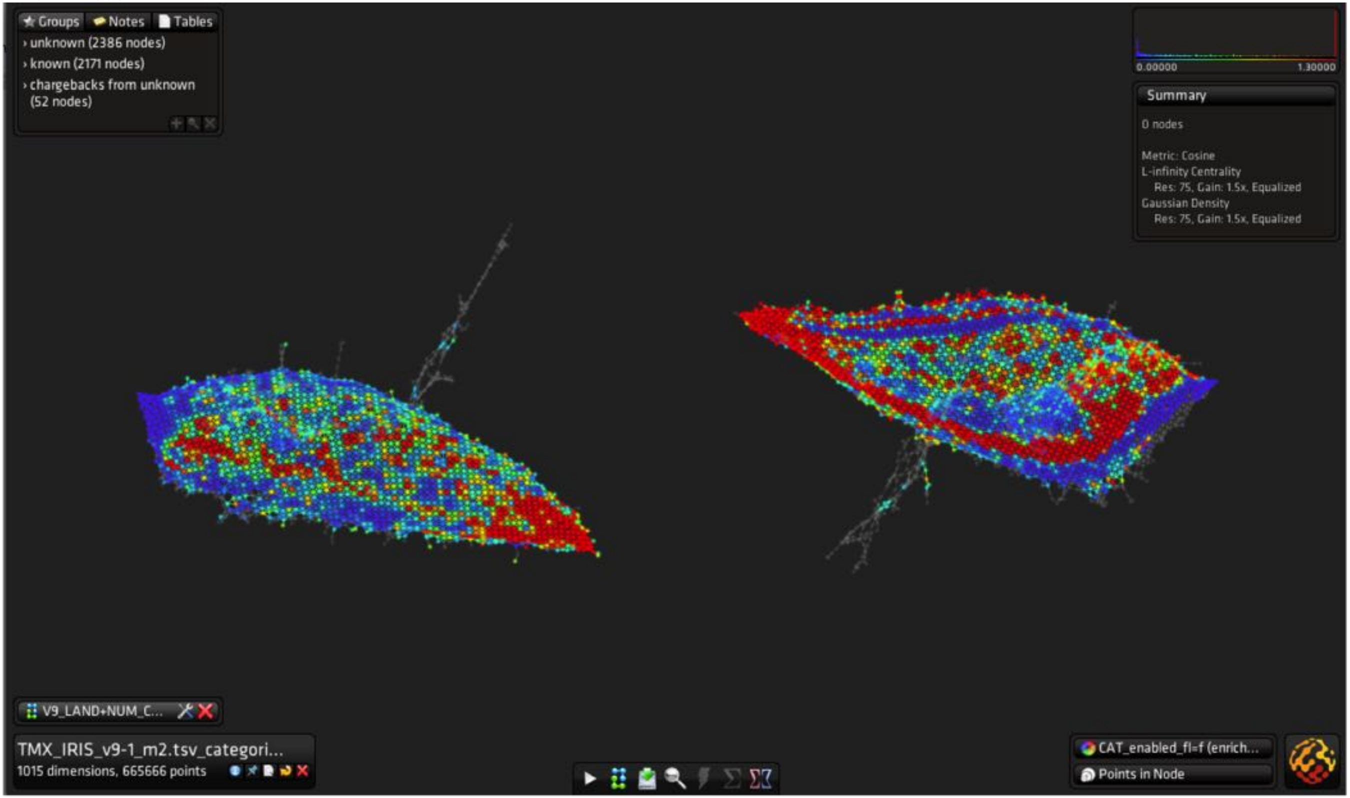
Online Fraud

Time On Page



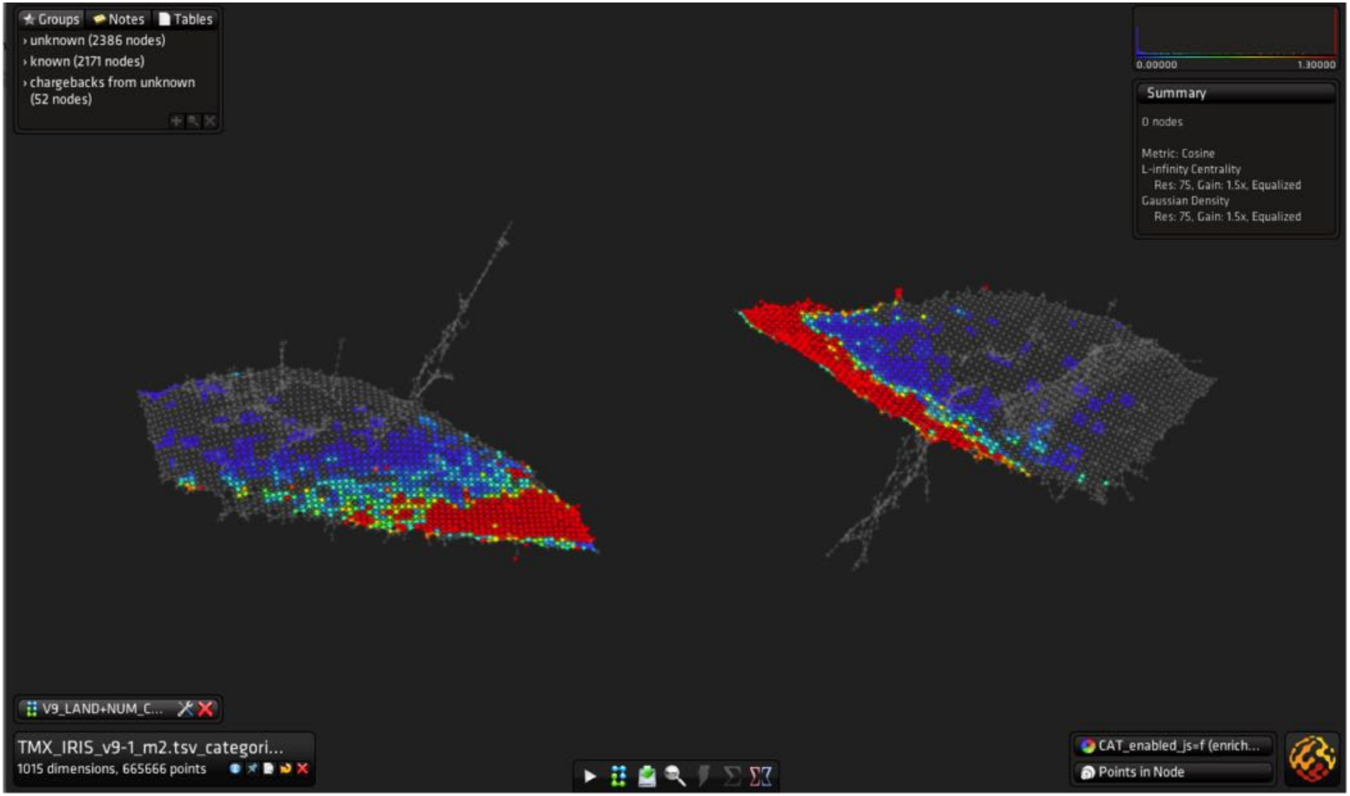
Online Fraud

No Flash



Online Fraud

No Javascript



Parkinson's Detection with Mobile Phone