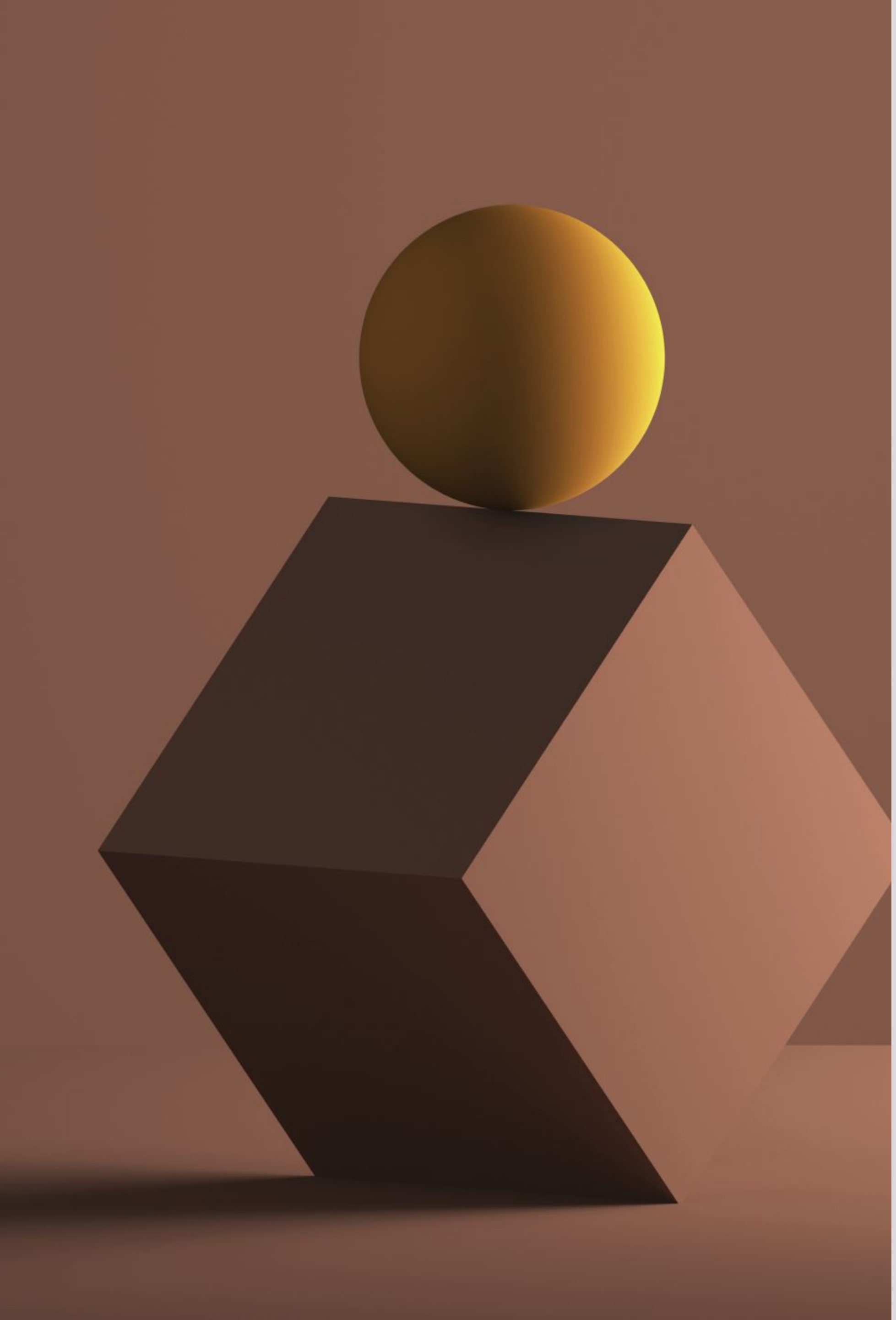


Geometry and Topology in Data Science and Machine Learning

Mikael Vejdemo-Johansson
CUNY College of Staten Island – Mathematics
CUNY Graduate Center – Computer Science / Data Science

Roadmap

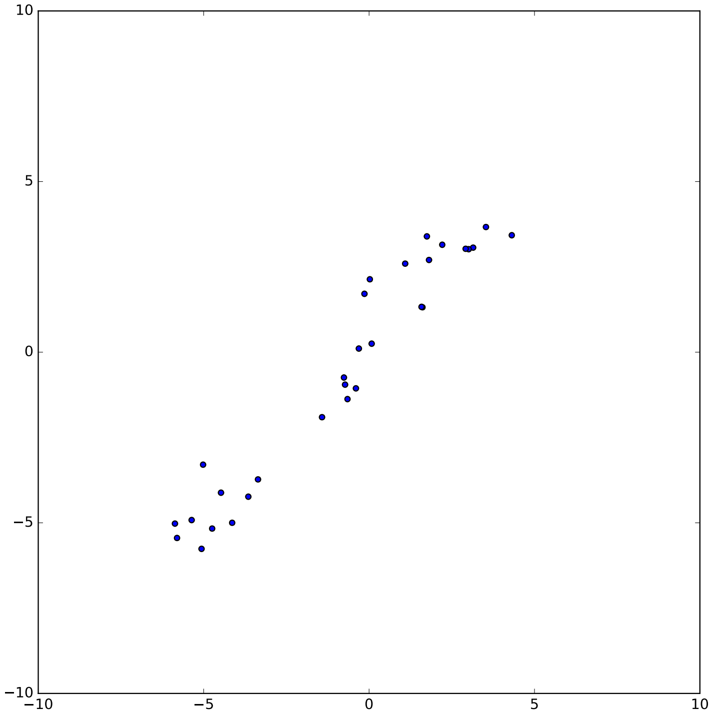
Topological Data Analysis
Use linear algebra to compute homology on data sets measuring their clusters, holes and bubbles.

Geometric Data Analysis
Use manifolds to estimate point cloud data.

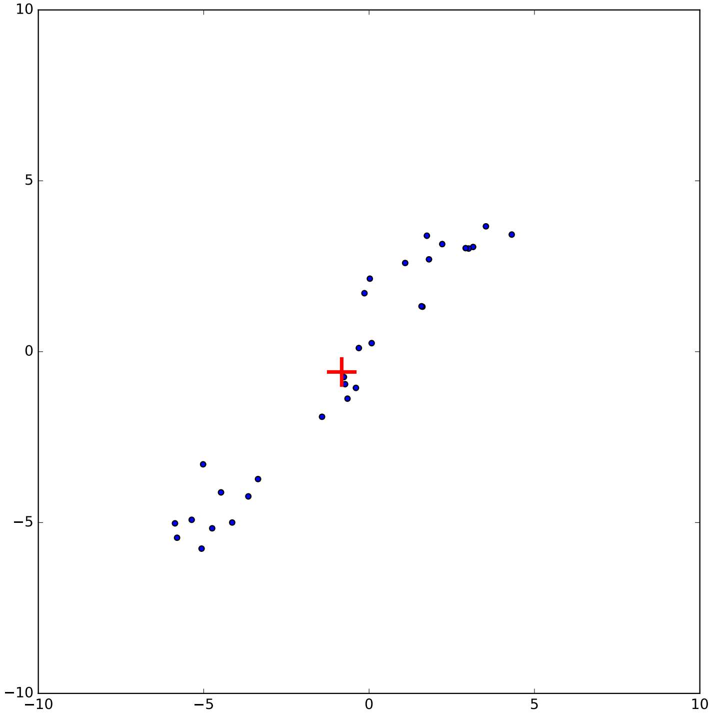
Information Geometry
Use differentiable manifolds to study parametrized distributions.

Algebraic Statistics
Use algebraic geometry to study statistics – with bonus content: use category theory to study statistical models.

Unifying perspective: Data has shape. Shape matters.



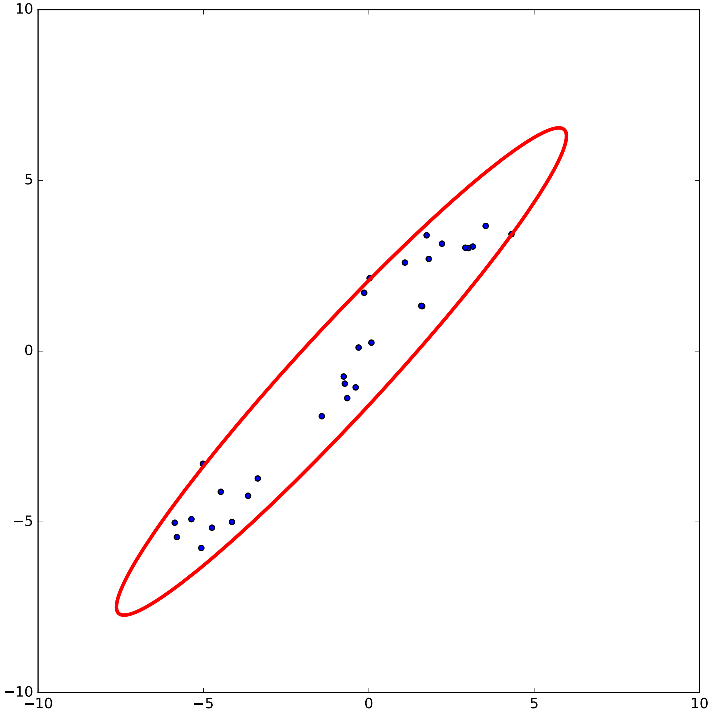
Data has shape



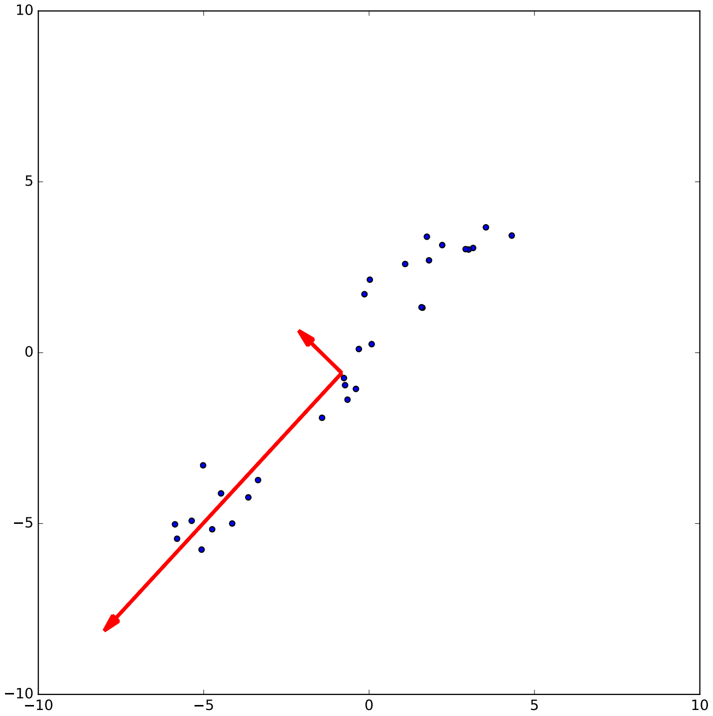
Data has shape

The mean (or centroid) gives us a **location** for the data set

Data has shape



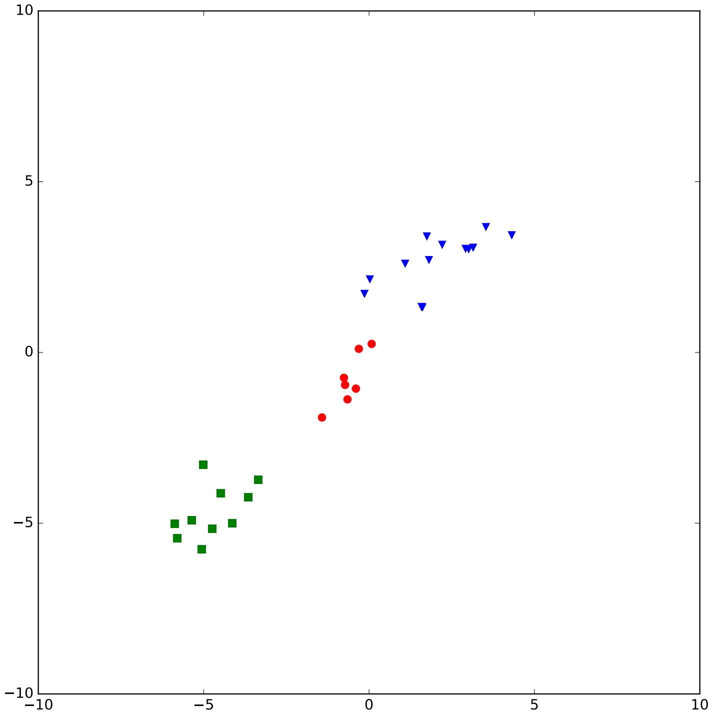
Variance (and covariance) measure how much and in what directions data spreads out.



Data has shape

PCA fits a best matching coordinate system to the data.

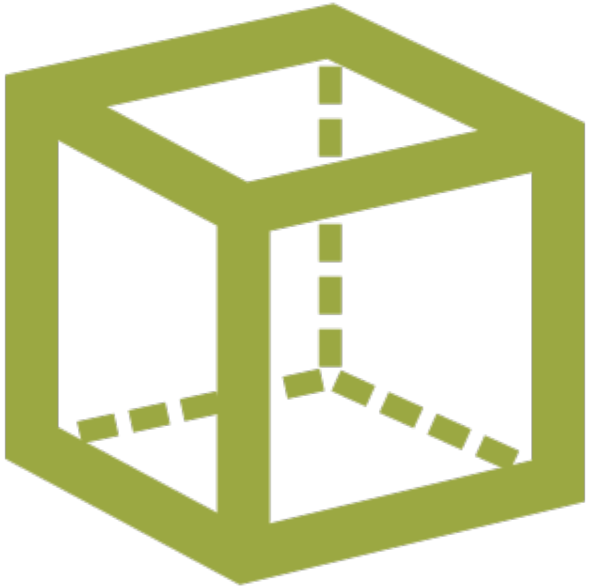
Data has shape



Clustering interprets the data as a collection of discrete unconnected points.

Shape matters

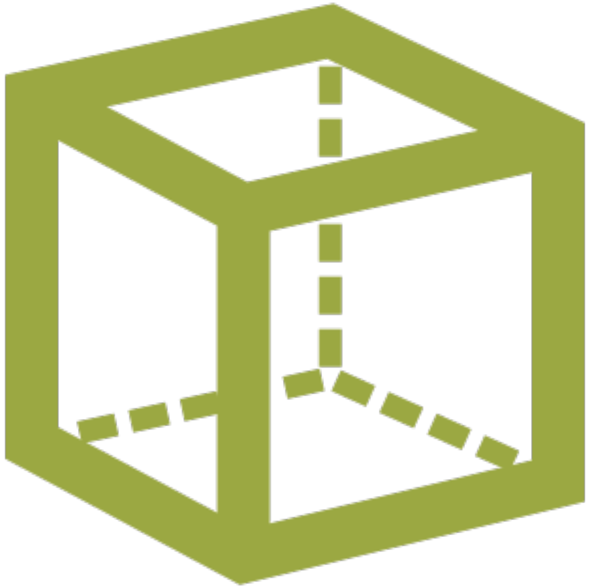
Your choice of data analysis tool imposes assumptions that your data may or may not obey.



Shape matters

Linear Regression assumes data is (close to) an affine hyperplane.

We have diagnostics to discover if this assumption is bad, and plans for when it is.

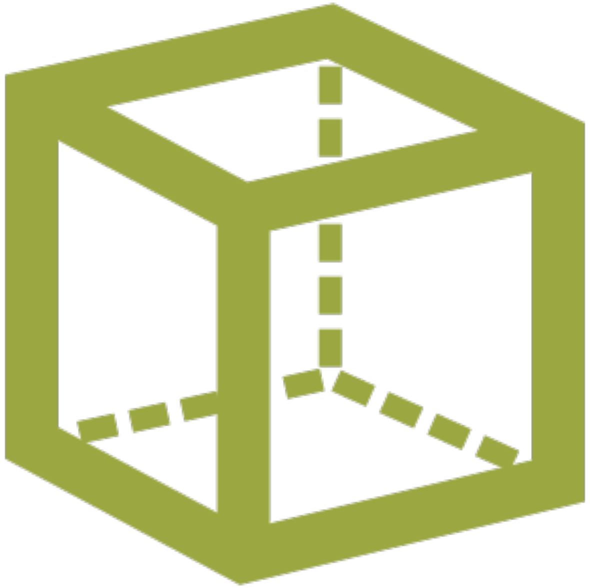


Shape matters

Machine Learning toolkits come with, inter alia, Decision Boundaries that may or may not have desirable properties:

Continuous? Smooth? Connected? Piece-wise linear?

We have approaches to measure the shapes of these Decision Boundaries.



Shape matters

When expanding our toolboxes, we want to control, and preferably limit, the implicit assumptions that our new tools place on our data.

Data has shape.
Shape matters.
How do we measure shape?

This talk describes four approaches:

1. Topological Data Analysis
   Use algebraic topology, especially homology, to study shapes.
2. Geometric Data Analysis
   Use Riemannian geometry, Differential geometry, and manifolds

Data has shape.
Shape matters.
How do we measure shape?

This talk describes four approaches:

3. Information Geometry
   Model families in classical statistics have fruitful differential geometrical properties.
4. Algebraic Statistics
   Classic statistical constructions, model families, etc can have fruitful algebraic geometrical interpretations.

Topological Data Analysis

Represent Data with (Simplicial) Complexes

TDA has several popular approaches for data analysis. All of them build on representing data as a discrete topological space (ie simplicial complex):

- (Persistent) Homology
- (Persistent) Cohomology
- Mapper

Homology uses linear algebra to find *holes* (…or bubbles, or higher-dimensional analogues)

Cohomology is the linear dual of homology, faster to compute, and can find *circular coordinates*.

Mapper constructs a simplicial complex model from data equipped with a *lens function*.

Persistence: Simplicial Complexes from Data

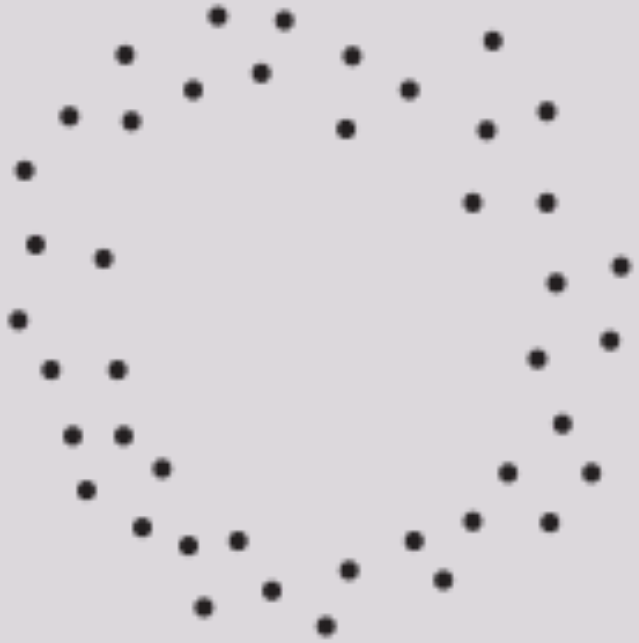
The simplest construction to *topologize* data is the Čech complex at scale ε:

- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots, x_d$ if the intersection of balls with radius ε centered at the vertices is non-empty

Persistence: Simplicial Complexes from Data

The simplest construction to *topologize* data is the Čech complex at scale ε:

- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots,\ x_d$ if the intersection of balls with radius ε centered at the vertices is non-empty



Persistence: Simplicial Complexes from Data

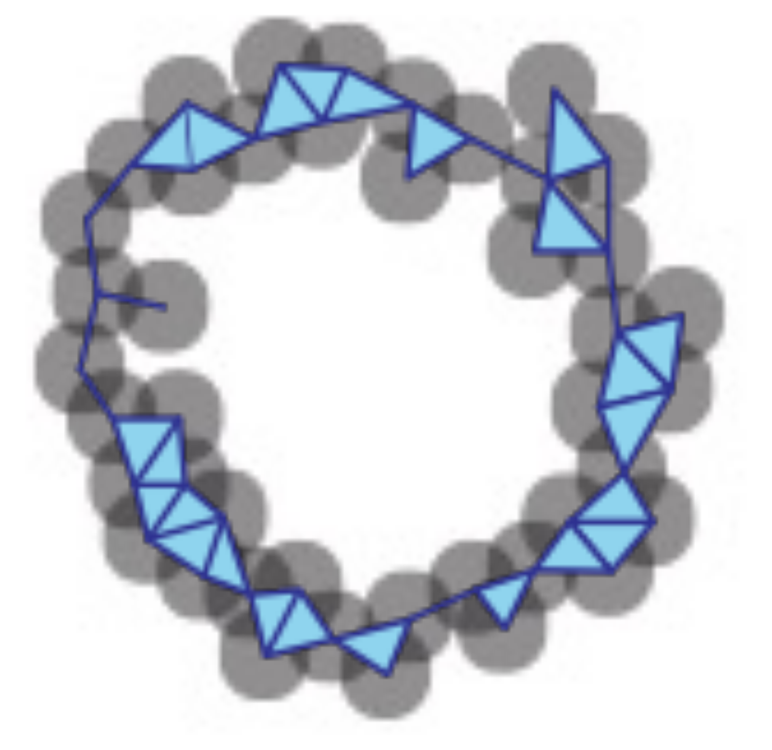
The simplest construction to *topologize* data is the Čech complex at scale ε:

- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots,\ x_d$ if the intersection of balls with radius ε centered at the vertices is non-empty

Persistence: Simplicial Complexes from Data

The simplest construction to *topologize* data is the Čech complex at scale ε:

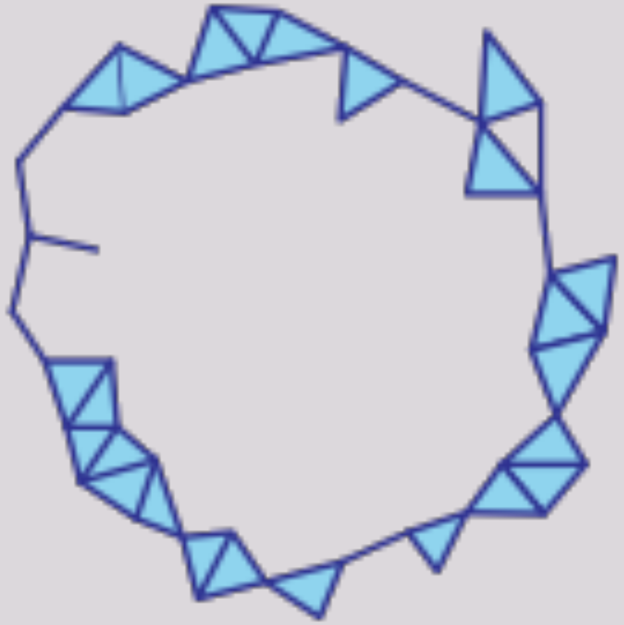
- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots, x_d$ if the intersection of balls with radius ε centered at the vertices is non-empty



Persistence: Simplicial Complexes from Data

The simplest construction to *topologize* data is the Čech complex at scale ε:

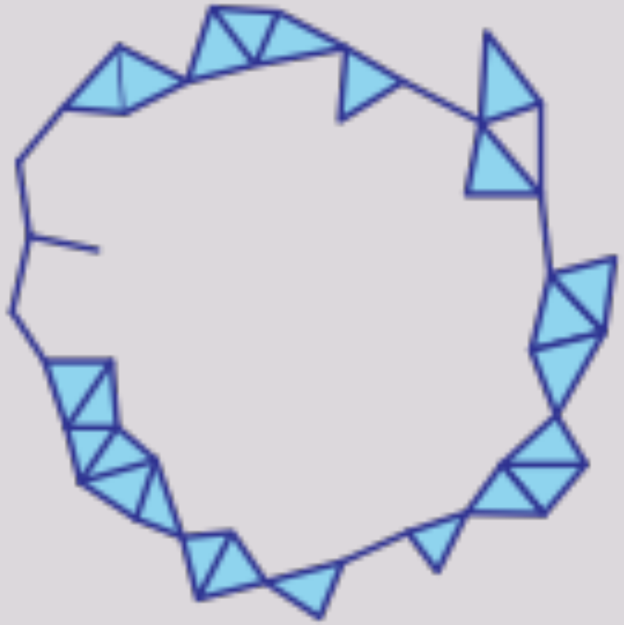
- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots, \ x_d$ if the intersection of balls with radius ε centered at the vertices is non-empty



Persistence: Simplicial Complexes from Data

The most common construction to *topologize* data is the Vietoris-Rips complex at scale ε:

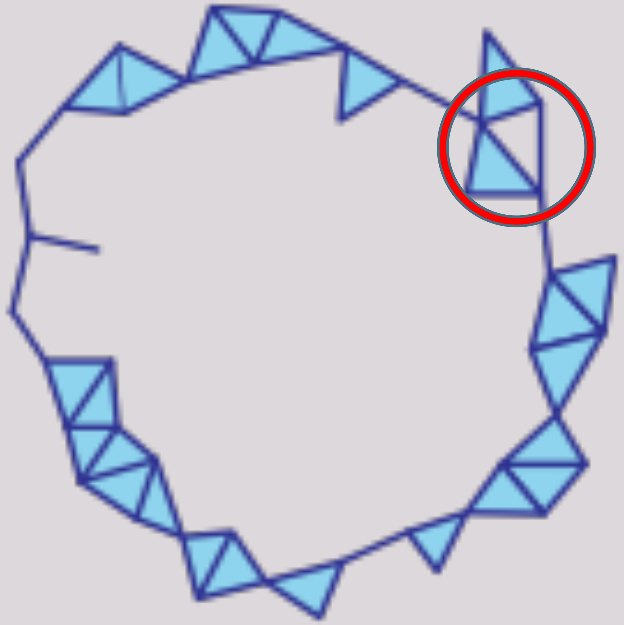
- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots, \ x_d$ if the pair-wise distances between vertices is $<\varepsilon$



Persistence: Simplicial Complexes from Data

The most common construction to *topologize* data is the Vietoris-Rips complex at scale ε:

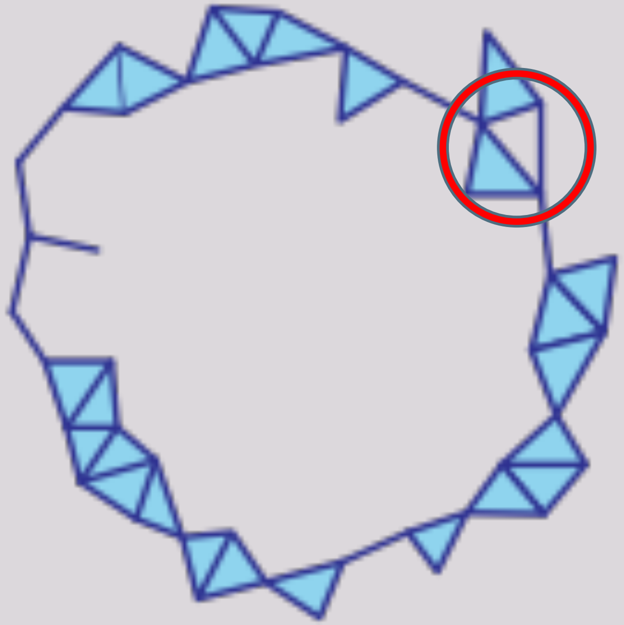
- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots,\ x_d$ if the pair-wise distances between vertices is <ε



Persistence: Simplicial Complexes from Data

The most common construction to *topologize* data is the Vietoris-Rips complex at scale ε:

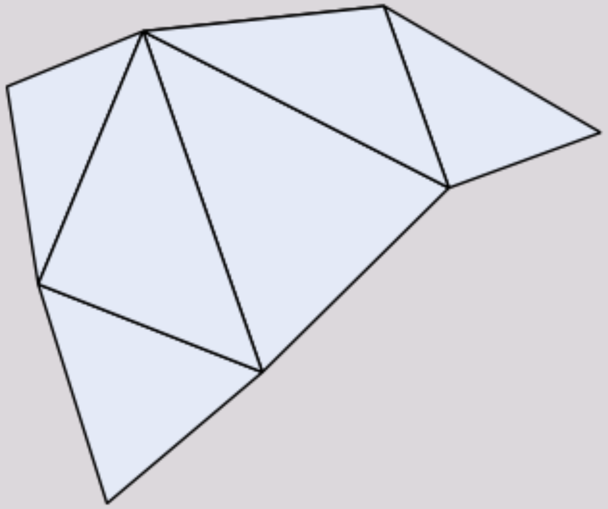
- Vertices are the data points
- Connect vertices $x_0, x_2, \ldots, \ x_d$ if the pair-wise distances between vertices is <ε



The Vietoris-Rips complex is the *clique complex* of the Čech complex: Add triangles (and tetrahedra, and higher simplices) for all cliques in the underlying graph.

Persistence: Homology: Chains

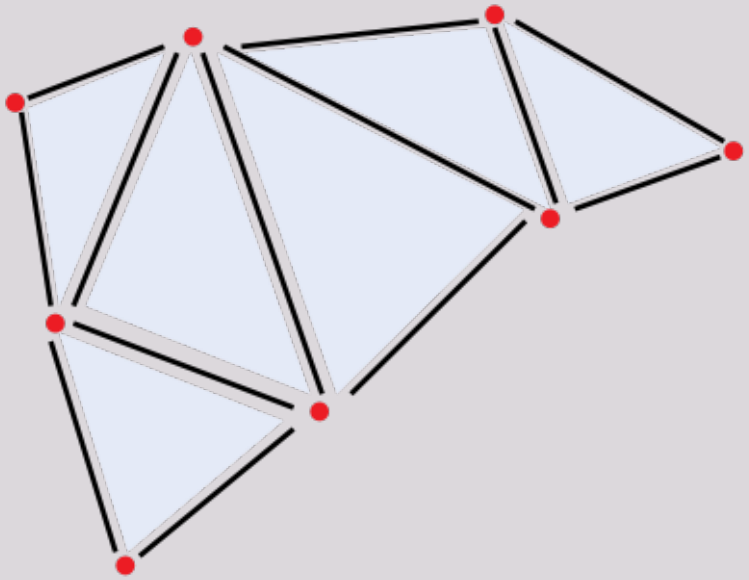
With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.



Persistence: Homology: Chains

With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.

Break up the complex into building blocks.

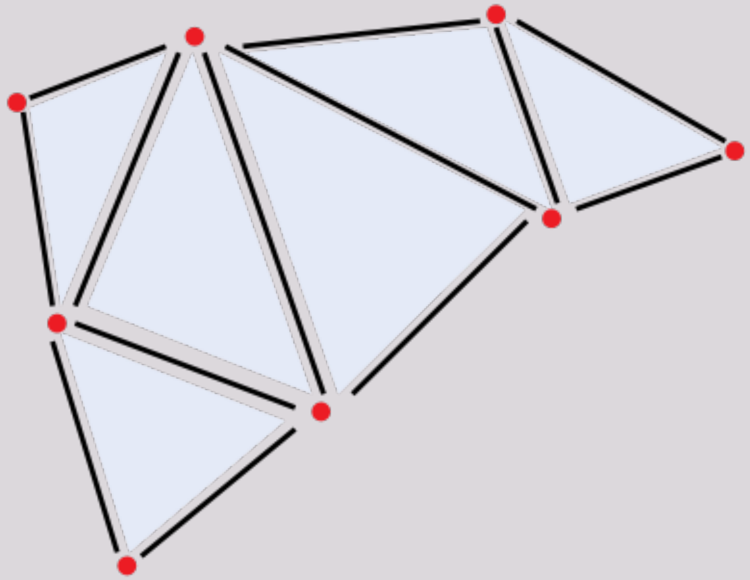


Persistence: Homology: Chains

With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.

Break up the complex into building blocks.

Create vector spaces – one for each dimension – with the simplices as abstract basis set:

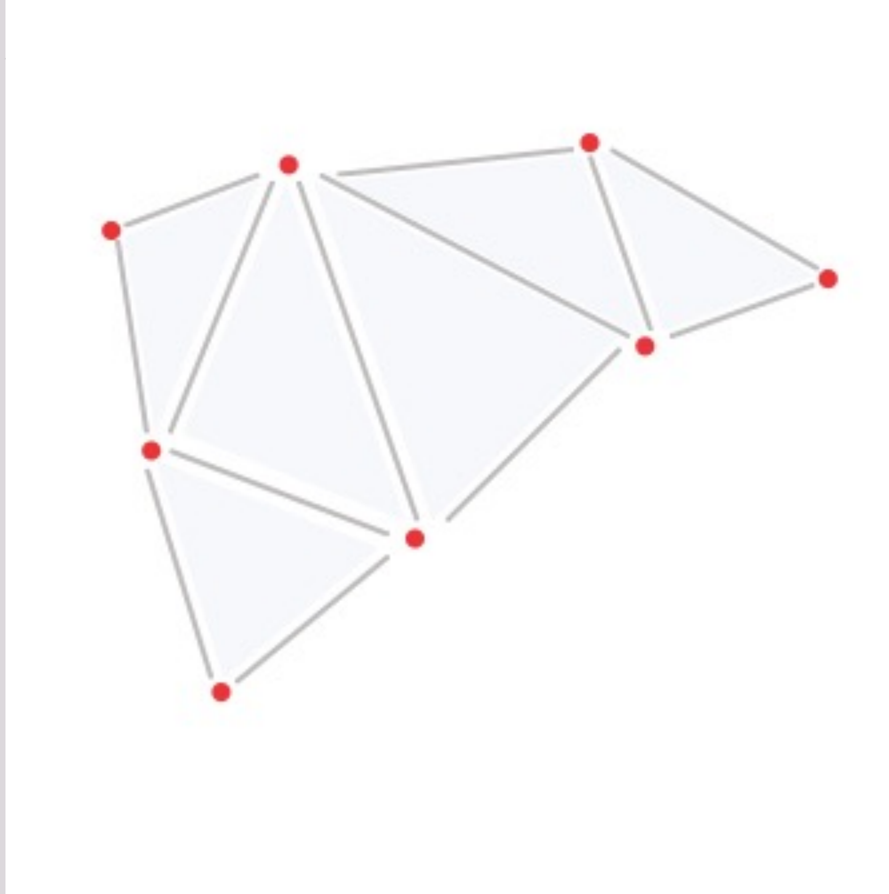


Persistence: Homology: Chains

With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.

Break up the complex into building blocks.

Create vector spaces – one for each dimension – with the s

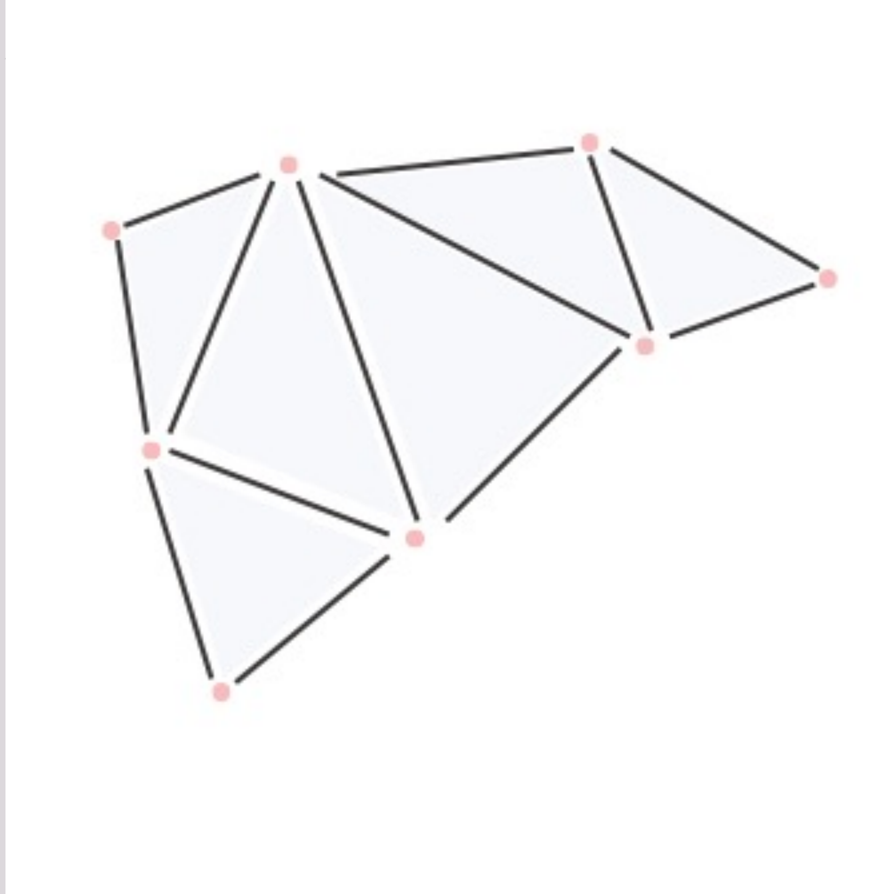


Persistence: Homology: Chains

With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.

Break up the complex into building blocks.

Create vector spaces – one for each dimension – with C_1 the s

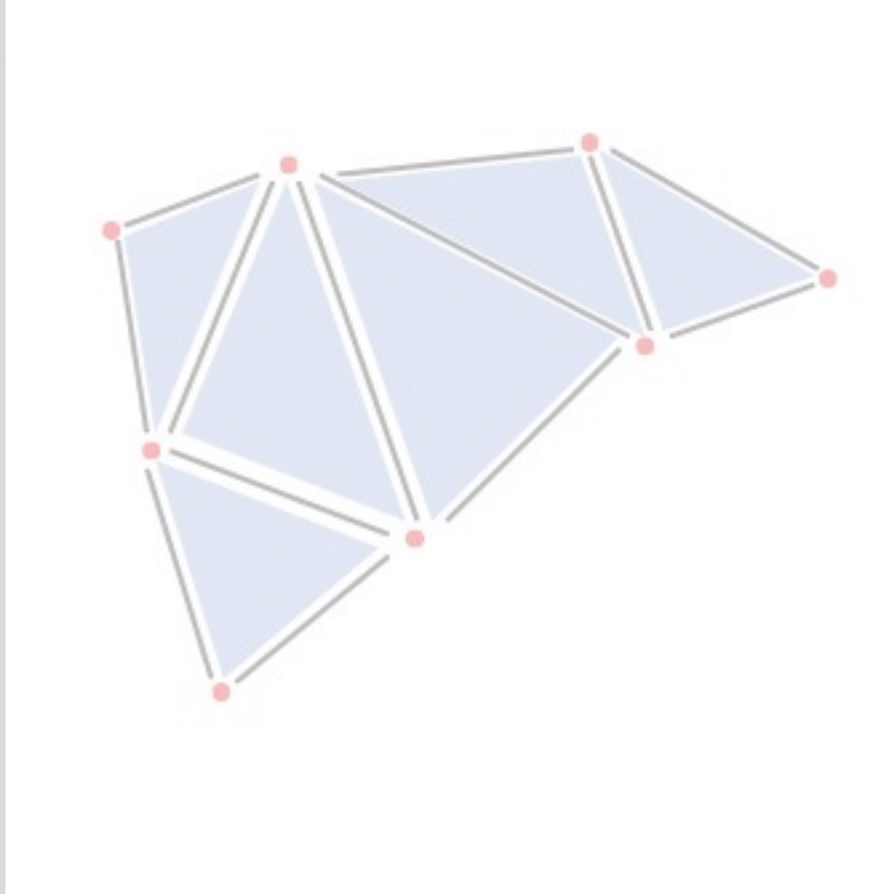


Persistence: Homology: Chains

With a simplicial complex in place, we can compute its *homology* – a vector space measuring *holes* and *bubbles* in the simplicial complex.

Break up the complex into building blocks.

Create vector spaces – one for each dimension – with C_1 C_2 the s

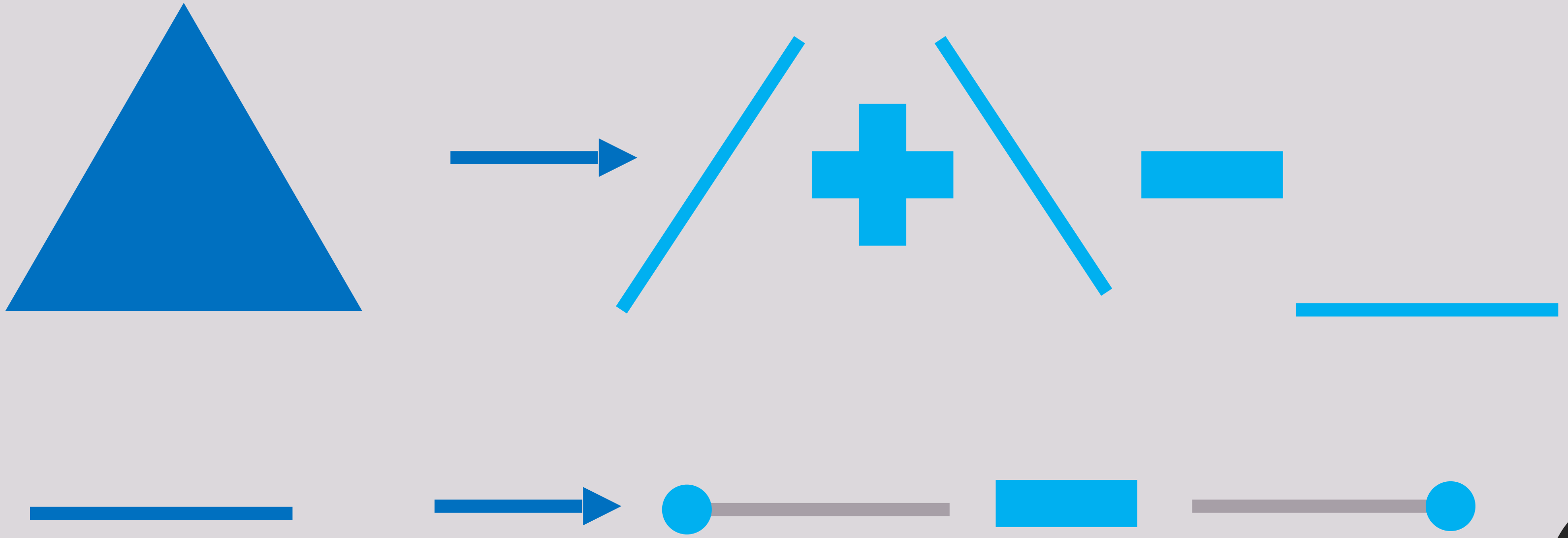


Persistence: Homology: Boundary Map

The boundary of a simplex is built out of simplices 1 dimension lower.

We can define a linear map by sending each simplex to a linear combination of its boundary simplices with alternating signs – then extend linearly.

This defines the *boundary map* ∂.



Persistence: Homology: Boundary Map

In a path of edges, the end-points appear in the boundary with opposite signs, cancelling each other out.

If end-points coincide (ie path forms a cycle), then the boundary is 0.

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if ∂z = 0.

The kernel of ∂ is the *cycle group* Z.

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if ∂z = 0.

The kernel of ∂ is the *cycle group* Z.

Some cycles are not surprising:

the boundary of a boundary is empty – $\partial^2 = 0$.

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if ∂z = 0.

The kernel of ∂ is the *cycle group* Z.

Some cycles are not surprising:

the boundary of a boundary is empty – $\partial^2 = 0$.

A *chain* z is a *boundary* if there is some w such that ∂w = z.

The image of ∂ is the *boundary group* B.

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if ∂z = 0.
The kernel of ∂ is the *cycle group* Z.

Some cycles are not surprising:

the boundary of a boundary is empty - $\partial^2 = 0$.

A *chain* z is a *boundary* if there is some w such that ∂w = z.
The image of ∂ is the *boundary group* B.

We would like to be able to nudge cycles across parts of the shape without changing what cycle we mean (homotopy invariance).

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if $\partial z = 0$.
The kernel of ∂ is the *cycle group* Z.

Some cycles are not surprising:

the boundary of a boundary is empty - $\partial^2 = 0$.

A *chain* z is a *boundary* if there is some w such that $\partial w = z$.
The image of ∂ is the *boundary group* B.

We would like to be able to nudge cycles across parts of the shape without changing what cycle we mean (homotopy invariance).

Boundaries tell us exactly when that is possible.

Persistence: Homology, Cycles and Boundaries

We can use this as a definition:

A *chain* z is a *cycle* if ∂z = 0.
The kernel of ∂ is the *cycle group* Z.

Some cycles are not surprising:

the boundary of a boundary is empty - $\partial^2 = 0$.

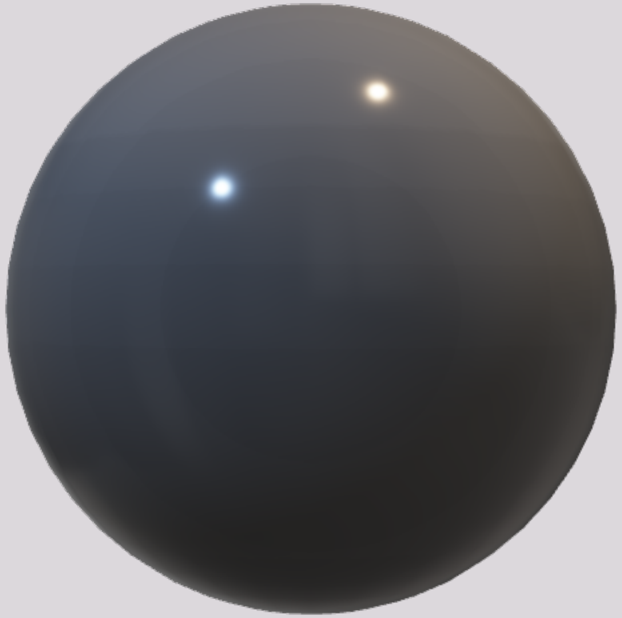
A *chain* z is a *boundary* if there is some w such that ∂w = z.
The image of ∂ is the *boundary group* B.

We would like to be able to nudge cycles across parts of the shape without changing what cycle we mean (homotopy invariance).

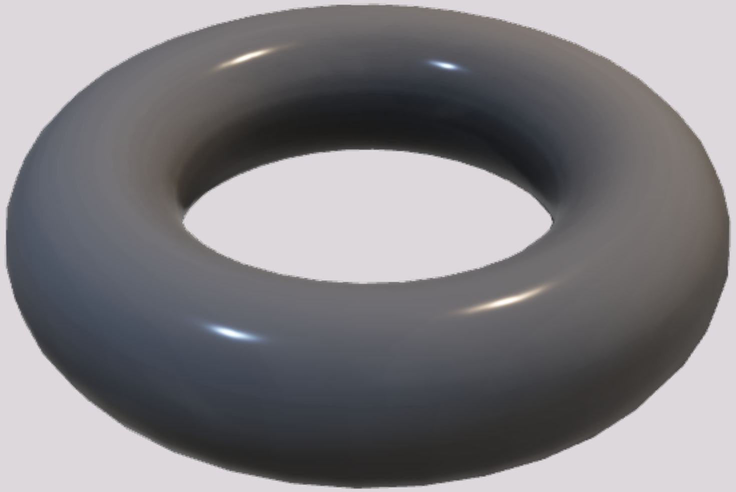
Boundaries tell us exactly when that is possible.

The essential cycles, up to this nudging, are exactly the quotient H = Z / B. We call this the *homology*.

Persistence: Homology Some Examples

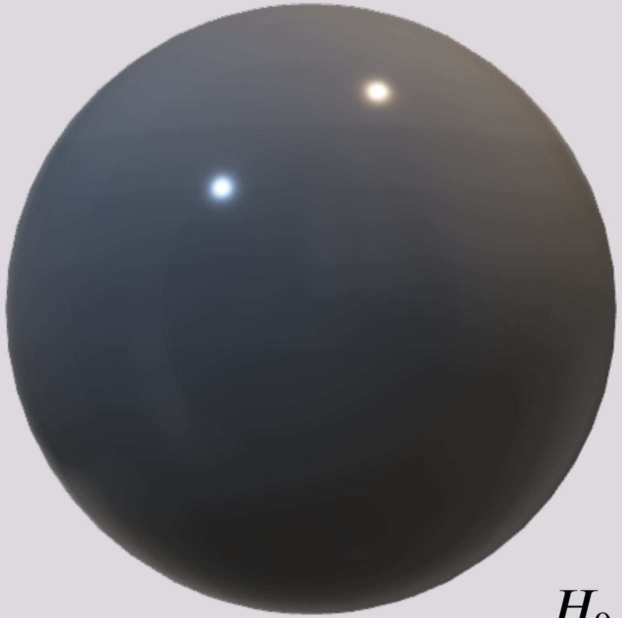


$H_0(sphere, \Bbbk) = \Bbbk^1$
$H_1(sphere, \Bbbk) = \Bbbk^0$
$H_2(sphere, \Bbbk) = \Bbbk^1$



$H_0(torus, \Bbbk) = \Bbbk^1$
$H_1(torus, \Bbbk) = \Bbbk^2$
$H_2(torus, \Bbbk) = \Bbbk^1$

Persistence: Homology Some Examples

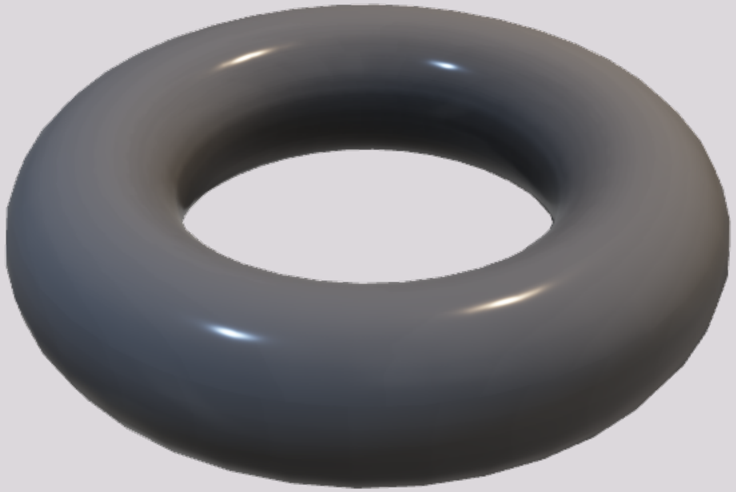


$$H_0(sphere,\ \Bbbk) = \Bbbk^1$$
$$H_1(sphere,\ \Bbbk) = \Bbbk^0$$
$$H_2(sphere,\ \Bbbk) = \Bbbk^1$$

H_0 measures “how many pieces”.

H_1 measures “how incontractible loops”.

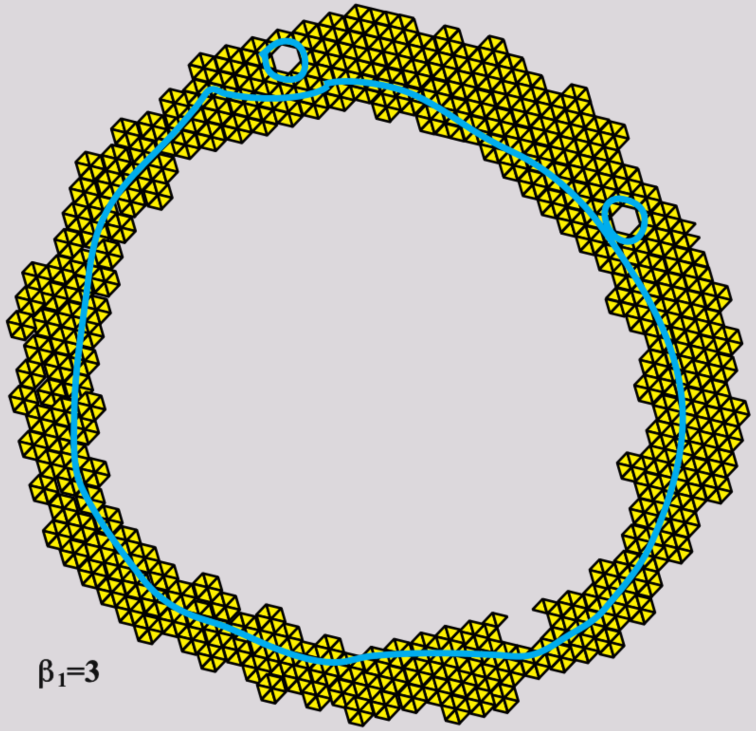
H_2 measures “how many enclosed voids”.



$$H_0(torus,\ \Bbbk) = \Bbbk^1$$
$$H_1(torus,\ \Bbbk) = \Bbbk^2$$
$$H_2(torus,\ \Bbbk) = \Bbbk^1$$

Persistence: What scale do we use?

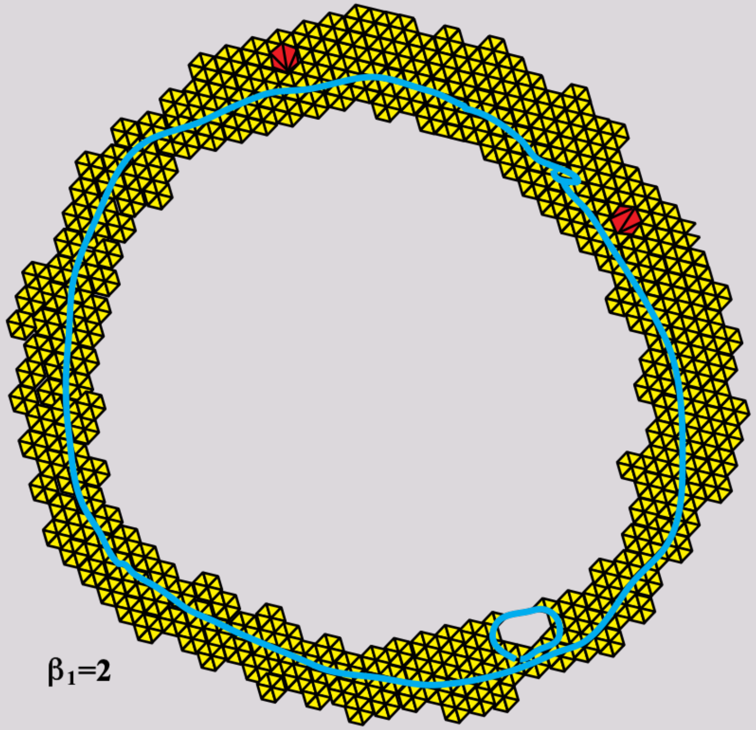
We could try just picking some scale to work at.



Persistence: What scale do we use?

We could try just picking some scale to work at.

But even small changes of scale can have dramatic effects on the detected features.

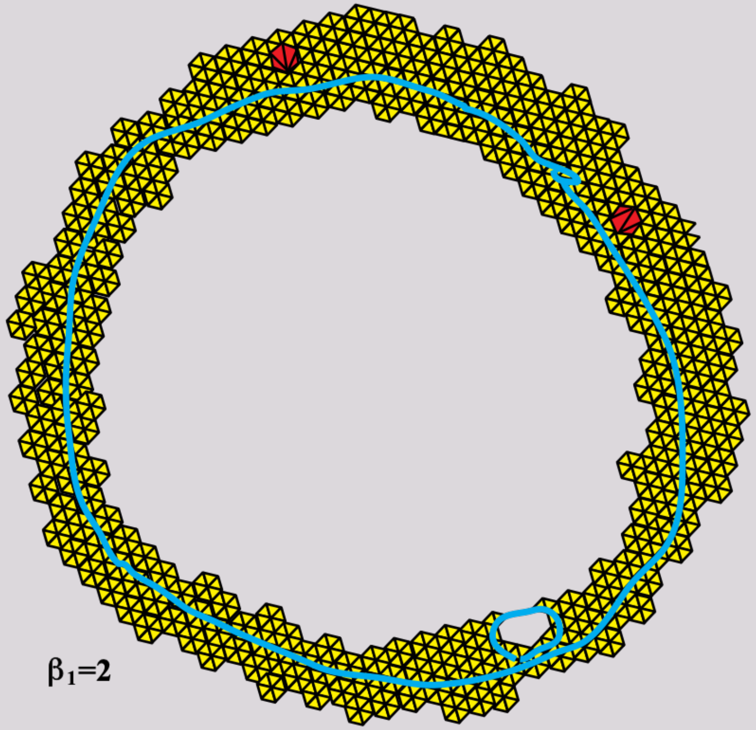


Persistence: What scale do we use?

We could try just picking some scale to work at.

But even small changes of scale can have dramatic effects on the detected features.

The solution: study all scales at once, using functoriality and representation theory.



Persistence: Functoriality

Homology is functorial: a map X → Y induces a linear map H(X) → H(Y) which respects composition.

Persistence: Functoriality

Homology is functorial: a map X → Y induces a linear map H(X) → H(Y) which respects composition.

So from a sequence of possibly interesting scales:

$$\varepsilon_1 < \varepsilon_2 < \varepsilon_3 < \varepsilon_4 < \varepsilon_5$$

Persistence: Functoriality

Homology is functorial: a map X → Y induces a linear map H(X) → H(Y) which respects composition.

So from a sequence of possibly interesting scales:

$$\varepsilon_1 < \varepsilon_2 < \varepsilon_3 < \varepsilon_4 < \varepsilon_5$$

We get a sequence of inclusion maps between simplicial complexes:

$$VR_{\varepsilon_1}(X) \subseteq VR_{\varepsilon_2}(X) \subseteq VR_{\varepsilon_3}(X) \subseteq VR_{\varepsilon_4}(X) \subseteq VR_{\varepsilon_5}(X)$$

Persistence: Functoriality

Homology is functorial: a map X → Y induces a linear map H(X) → H(Y) which respects composition.

So from a sequence of possibly interesting scales:

$$\varepsilon_1 < \varepsilon_2 < \varepsilon_3 < \varepsilon_4 < \varepsilon_5$$

We get a sequence of inclusion maps between simplicial complexes:

$$VR_{\varepsilon_1}(X) \subseteq VR_{\varepsilon_2}(X) \subseteq VR_{\varepsilon_3}(X) \subseteq VR_{\varepsilon_4}(X) \subseteq VR_{\varepsilon_5}(X)$$

By functoriality we get a sequence of linear maps between homology groups:

$$HVR_{\varepsilon_1}(X) \to HVR_{\varepsilon_2}(X) \to HVR_{\varepsilon_3}(X) \to HVR_{\varepsilon_4}(X) \to HVR_{\varepsilon_5}(X)$$

Persistence: Representation Theory

Theorem (Gabriel, 1972)

A diagram of vector spaces

$$V_1 \to V_2 \to V_3 \to V_4 \to V_5$$

Decomposes into a direct sum of component diagrams, each of which is 1-dimensional with identity maps in a connected interval, and 0 elsewhere. ■

Persistence: Representation Theory

Theorem (Gabriel, 1972)

A diagram of vector spaces

$$V_1 \to V_2 \to V_3 \to V_4 \to V_5$$

Decomposes into a direct sum of component diagrams, each of which is 1-dimensional with identity maps in a connected interval, and 0 elsewhere. ■

In other words, we can make a simultaneous basis change for all homology groups at all scales so that the induced topological features get mapped identically between one scale and the next.

The result can be described with the start- and end-indices of the connected interval corresponding to a feature.

Persistent Homology

What do we get out of it?

From a dataset (or *point cloud*) we get a description of the topology of a shape resembling the dataset as a multiset of intervals (b_i, d_i)

These descriptors are *stable*: If the point cloud changes by a bounded amount, the end-points of intervals can change only by that amount. (length 0 intervals vanish, and can be created, at will)

Persistent Homology

3x3 pixel patches in natural images concentrate on a Klein bottle in $\mathbb{R}^9$. This can be used to create compression algorithms, rotation invariant signatures for image textures , or inform a Convolutional Neural Network doing computer vision tasks.

What do people do with it?

Chemical properties of zeolites, induced by pore geometry, can be given persistent homology signatures and used to pre-screen interesting compounds before spending time simulating or synthesizing them.

Persistent Homology

Where can I learn more?

Carlsson, Gunnar. "Topology and data." *Bulletin of the American Mathematical Society* 46.2 (2009): 255-308.

Several books exist by now that focus on different aspects.

Persistent Homology

Where can I learn more?

Carlsson, Gunnar. "Topology and data." *Bulletin of the American Mathematical Society* 46.2 (2009): 255-308.

Several books exist by now that focus on different aspects.

Release date February 2022:

Carlsson, Gunnar and Vejdemo-Johansson, Mikael. “Topological Data Analysis with Applications”. Cambridge University Press (2022).

Introduces all relevant topology and the persistence theory needed, and ends with a sequence of case studies where TDA is applied.

Geometric Data Analysis

Changing Meaning of Geometric Data Analysis

1960s – 1980s

Benzécri et al: Analyse des Données / Analyse des Correspondances.

Data is interpreted as point clouds. PCA and variations (MCA) adapted for categorical data used.

1980s-

Kendall: Shape manifolds, procrustean metrics and complex projective spaces.

Geometric shapes – up to rotation, scaling and translation – form manifolds parametrizing the shapes. Motivates development of statistics without vector space operations – Fréchet Means etc.

2000-

Manifold Learning – fit a *nice* manifold to an observed point cloud. Isomap / Locally-linear embeddings / Laplacian Eigenmaps / t-SNE / UMAP.

Kendall Shape Spaces

Shapes are represented by *k landmark points* in the plane: producing vectors in $\mathbb{R}^{2k} = \mathbb{C}^k$.

Quotient out translation ($\mathbb{C}^{k-1}$), scale and rotation (ie complex scalar multiplication) produces points in $\mathbb{CP}^{k-2} = \Sigma_2^k$. In general, the space of *k* points in *d* dimensions is the shape manifold Σ_d^k.

Kendall Shape Spaces

Shapes are represented by *k landmark points* in the plane: producing vectors in $\mathbb{R}^{2k} = \mathbb{C}^k$.

Quotient out translation ($\mathbb{C}^{k-1}$), scale and rotation (ie complex scalar multiplication) produces points in $\mathbb{CP}^{k-2} = \Sigma_2^k$. In general, the space of *k* points in *d* dimensions is the shape manifold Σ_d^k.

Example

The space of triangles forms a sphere: first two vertices can be fixed to the points $\pm 1 \in \mathbb{C}$. Third vertex is placed in the induced point of $\mathbb{C}$. We augment a point at ∞ representing the case when the first two points coincide. This ends up isometric to the sphere with radius ½.

Kendall Shape Spaces

Shapes are represented by k *landmark points* in the plane: producing vectors in $\mathbb{R}^{2k} = \mathbb{C}^k$.

Quotient out translation ($\mathbb{C}^{k-1}$), scale and rotation (ie complex scalar multiplication) produces points in $\mathbb{CP}^{k-2} = \Sigma^k_2$. In general, the space of k points in d dimensions is the shape manifold Σ^k_d.

Example

The space of triangles forms a sphere: first two vertices can be fixed to the points $\pm 1 \in \mathbb{C}$. Third vertex is placed in the induced point of $\mathbb{C}$. We augment a point at ∞ representing the case when the first two points coincide. This ends up isometric to the sphere with radius ½.

This figure shows the view from the north pole (origin): the equilateral triangle. As we approach the equator, triangles approach sets of collinear points.

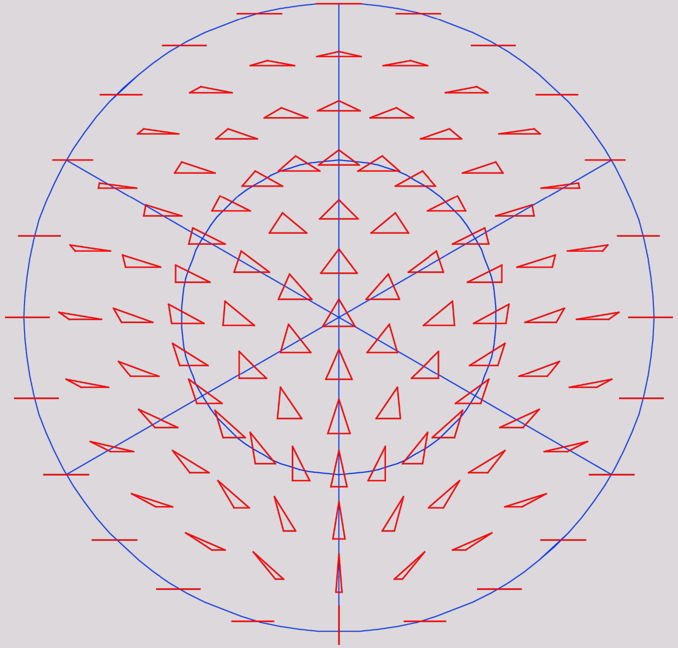


Figure from Klingenberg 2015, https://doi.org/10.3390/sym7020843

Fréchet Means

Statistics without arithmetic: mean can not be defined as $\frac{1}{N}\sum x_i$.

Solution: use the fact that the mean minimizes the aggregated squared distances to the data points.

Fréchet Means

Statistics without arithmetic: mean can not be defined as $\frac{1}{N}\sum x_i$.

Solution: use the fact that the mean minimizes the aggregated squared distances to the data points.

Definition

The *Fréchet variance* is $\mathbb{V}_F(p) = \sum d(p, x_i)^2$.

The *Karcher means* are local minima of $\mathbb{V}_F$.

The *Fréchet mean* is the global minimum of $\mathbb{V}_F$, if it exists.

Fréchet Means

Statistics without arithmetic: mean can not be defined as $\frac{1}{N}\sum x_i$.

Solution: use the fact that the mean minimizes the aggregated squared distances to the data points.

Definition

The *Fréchet variance* is $\mathbb{V}_F(p) = \sum d(p, x_i)^2$.

The *Karcher means* are local minima of $\mathbb{V}_F$.

The *Fréchet mean* is the global minimum of $\mathbb{V}_F$, if it exists.

Arithmetic mean: use Euclidean distance

Median: use square root of Euclidean distance

Geometric mean: use $d(x, y) = |\log x - \log y|$

Harmonic mean: use $d(x, y) = \left|\frac{1}{x} - \frac{1}{y}\right|$

Manifold Learning

When a (Riemannian) manifold is embedded in a Euclidean space, geodesic and ambient distances might be very different.

Figure by Oliver Grisel, CC-BY, Wikipedia

Manifold Learning

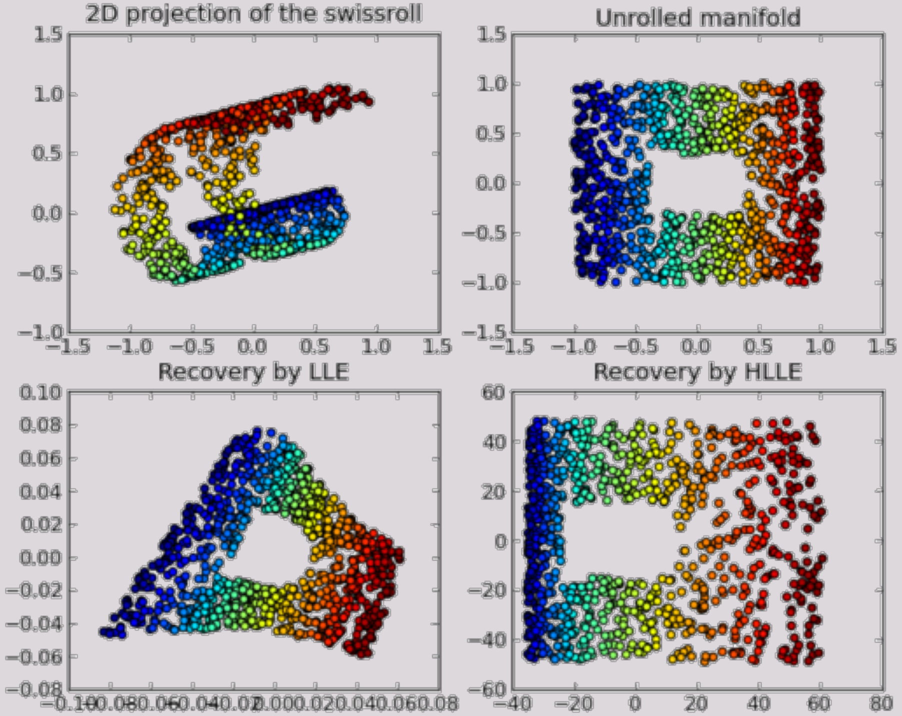


Figure by Oliver Grisel, CC-BY, Wikipedia

When a (Riemannian) manifold is embedded in a Euclidean space, geodesic and ambient distances might be very different.

Example

The Swiss Roll dataset has points sampled on the surface of a rolled-up square. Ambient distance methods (such as PCA) would put points near each other jumping the gap.

Many manifold learning techniques learn a nearest-neighbor graph and use graph distance as a proxy for geodesic distance.

Manifold Learning

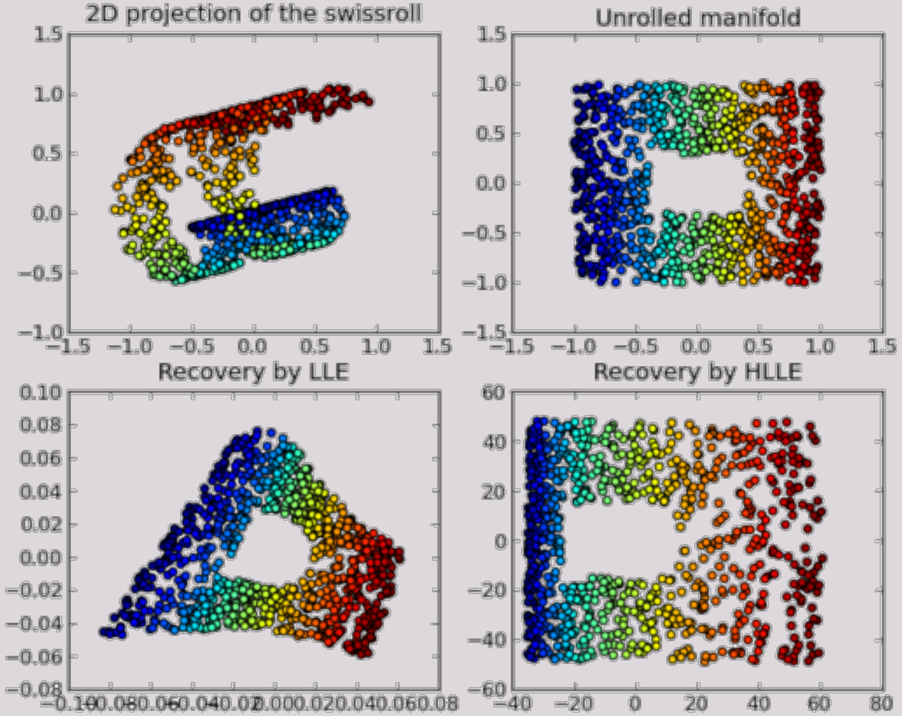


Figure by Oliver Grisel, CC-BY, Wikipedia

When a (Riemannian) manifold is embedded in a Euclidean space, geodesic and ambient distances might be very different.

Example

The Swiss Roll dataset has points sampled on the surface of a rolled-up square. Ambient distance methods (such as PCA) would put points near each other jumping the gap.

Many manifold learning techniques learn a nearest-neighbor graph and use graph distance as a proxy for geodesic distance.

Laplacian Eigenmaps: eigenmaps of graph Laplacian on NN-graph produce coordinates.

Isomap: Multidimensional Scaling (MDS) on weighted NN-graph distances.

Locally-Linear Embeddings: Barycentric coordinates for each point based on its neighbors. Minimize a cost function measuring reconstruction error using eigenvalues.

Geometric Data Analysis Resources

Geomstats – geomstats.github.io
Python package for computing with data on manifolds, for manifold learning, etc.

scikit-learn – scikit-learn.org
Python package for a wide range of machine learning tasks, including manifold learning.

GDAtools
R package for "old school" Geometric Data Analysis: correspondence analysis etc.

KeOps, GeomLoss
PyTorch packages for introducing geometric methods to deep learning.

Information Geometry

Parametrized Distributions

Parametrized Distributions

Many probability distributions live in parametrized families:

$\ell, u \rightarrow Uniform(\ell,\ u)$

$\mu,\ \sigma^2 \rightarrow \mathcal{N}(\mu, \sigma^2)$

$\lambda \rightarrow Exponential(\lambda)$

$p \rightarrow Bernoulli(p)$

$n, p \rightarrow Binomial(n, p)$

Statistical estimation deals with the problem of choosing appropriate parameters θ given observed data and a choice of parametrized family P(x|θ).

Parametrized Distributions

Parameters form Manifolds

Parametrized Distributions

Uniform: the interval $[\ell, u]$

Normal: the half-space $\mathbb{R} \times \mathbb{R}^+$

Exponential: the half-line $\mathbb{R}^+$

Bernoulli: the interval $[0,1]$

Binomial: the stripes $\mathbb{N} \times [0,1]$

Parameters form Manifolds

It turns out that manifolds whose points are probability distributions form a special class of manifolds. We call these *statistical manifolds*.

Information Metric

A *metric* on a parameter manifold *should* measure distinguishability:

$d(p(x|\theta), p(x|\theta+d\theta))$ should measure how different $p(x|\theta)$ is from $p(x|\theta+d\theta)$.

Information Metric

A *metric* on a parameter manifold *should* measure distinguishability:
d(p(x|θ), p(x|θ+dθ)) should measure how different p(x|θ) is from p(x|θ+dθ).

The *Relative Difference*

$$\Delta = \frac{p(x\,|\,\theta + d\theta) - p(x\,|\,\theta)}{p(x\,|\,\theta)} = \frac{\partial \log p(x\,|\,\theta)}{\partial \theta^a} d\theta^a$$

would be an obvious choice to consider.

Information Metric

A *metric* on a parameter manifold *should* measure distinguishability:
d(p(x|θ), p(x|θ+dθ)) should measure how different p(x|θ) is from p(x|θ+dθ).

The *Relative Difference*

$$\Delta = \frac{p(x\,|\,\theta + d\theta) - p(x\,|\,\theta)}{p(x\,|\,\theta)} = \frac{\partial \log p(x\,|\,\theta)}{\partial \theta^a} d\theta^a$$

would be an obvious choice to consider.

However: $\mathbb{E}\left[\Delta\right] = 0$

Information Metric

A *metric* on a parameter manifold *should* measure distinguishability:
d(p(x|θ), p(x|θ+dθ)) should measure how different p(x|θ) is from p(x|θ+dθ).

The *Relative Difference*

$$\Delta = \frac{p(x\,|\,\theta + d\theta) - p(x\,|\,\theta)}{p(x\,|\,\theta)} = \frac{\partial \log p(x\,|\,\theta)}{\partial \theta^a} d\theta^a$$

would be an obvious choice to consider.

However: $\mathbb{E}\left[\Delta\right] = 0$

One thing that works, however, is the variance!
We define

$$d\ell^2 = \mathbb{V}[\Delta] = \int \mathbb{d}x \; p(x\,\big|\,\theta) \frac{\partial \log p(x\,\big|\,\theta)}{\partial \theta^a} \frac{\partial \log p(x\,\big|\,\theta)}{\partial \theta^b} \, \mathbb{d}\theta^a \, \mathbb{d}\theta^b$$

Information Metric

A *metric* on a parameter manifold *should* measure distinguishability:
d(p(x|θ), p(x|θ+dθ)) should measure how different p(x|θ) is from p(x|θ+dθ).

The *Relative Difference*

$$\Delta = \frac{p(x\,|\,\theta + d\theta) - p(x\,|\,\theta)}{p(x\,|\,\theta)} = \frac{\partial \log p(x\,|\,\theta)}{\partial \theta^a} d\theta^a$$

would be an obvious choice to consider.

However: $\mathbb{E}\left[\Delta\right] = 0$

One thing that works, however, is the variance!
We define

$$d\ell^2 = \mathbb{V}\left[\Delta\right] = \int dx\ p(x\,\middle|\,\theta) \frac{\partial \log p(x\,\middle|\,\theta)}{\partial \theta^a} \frac{\partial \log p(x\,\middle|\,\theta)}{\partial \theta^b} d\theta^a d\theta^b$$

The matrix g_{ab} such that $d\ell^2 = g_{ab} d\theta^a d\theta^b$ is the *Fisher Information Matrix*. Interpreting this as a Riemannian Metric Tensor produces the *Fisher Information Metric*.

Chentsov's Theorem

(aka Čencov, Ченцов*)*

Theorem

The Fisher Information Metric is (up to scaling) the *only* Riemannian metric on statistical manifolds that is invariant under Markov mappings.

Chentsov's Theorem

(aka Čencov, Ченцов*)*

Theorem

The Fisher Information Metric is (up to scaling) the *only* Riemannian metric on statistical manifolds that is invariant under Markov mappings.

Here, a Markov mapping can be understood through example: consider a 6-sided die with probabilities $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \theta/3$ and $\mathbb{P}(4) = \mathbb{P}\left(5\right) = \mathbb{P}\left(6\right) = (1 - \theta)/3$. The outcomes low={1,2,3} and high={4,5,6} can be described as a weighted coin with side probabilities θ and 1-θ.

This re-interpretation is an embedding of the statistical manifold of Binomial(n,θ) into the manifold of Multinomial(n;θ_1, θ_2, θ_3, θ_4, θ_5, θ_6).

Example: Multinomial Distributions

$$P(n \mid \theta) = \frac{N!}{n_1! \dots n_m!} \theta_1^{n_1} \dots \theta_m^{n_m}$$

where $\sum n_i = N$ and $\sum \theta_i = 1$ has parameter manifold given by the simplex. The metric tensor has entries

$$g_{ij} = \frac{N}{\theta_i} \delta_{ij} + \frac{N}{\theta_m}$$

where 1≤i,j≤m-1.

So… what is Information Geometry?

Using differential geometry tools to study the Riemannian metric on statistical manifolds.

With the Fisher Information tensor in place, we can find statistical relevance for geodesics, normal projections, parallel transport, covariant derivatives, connections, and curvature.

One first example:
Fisher Information Metric is the curvature of the Kullback-Leibler divergence:

$$KL\left(p:q\right) = \int p(x)\log\frac{p(x)}{q(x)}\,dx$$

What makes me *excited about it?*

Baudot and Bennequin, The Homological Nature of Entropy, MDPI Entropy 2015, 17, 3253-3318.

Using homological algebra tools, a topological space is constructed such that: degree 1 cohomology is one-dimensional, and generated by the Shannon entropy function.

Bradley, Entropy as a Topological Operad Derivation, MDPI Entropy 2021, 23 (9), 1195.

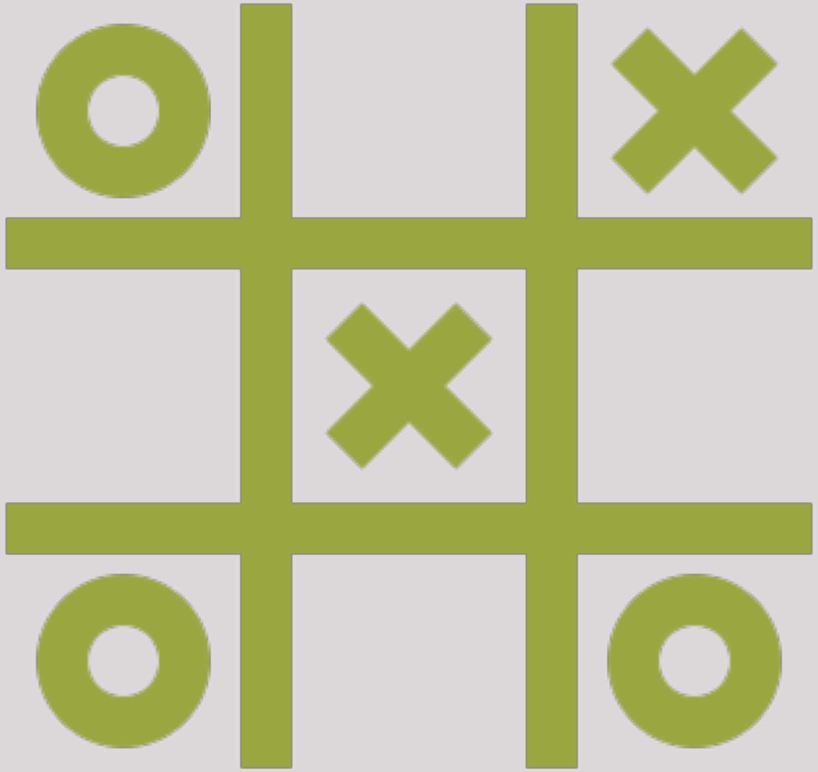
Shannon entropy defines a derivation of the operad of topological simplices, and for every derivation of this operad, at some point it is a constant multiple of Shannon entropy.

Where can I learn more?

This section was *heavily* informed by:
Caticha, The basics of information geometry. AIP Conference Proceedings 1641, 15 (2015).

Canonical reference:
Amari and Nagaoka, Methods of Information Geometry. AMS / Oxford University Press, (2000)

Algebraic Statistics



What is Algebraic Statistics?

The application of algebraic geometry to problems in statistics and probability.

1998 Diaconis and Sturmfels
Conditional inference – random walks on contingency tables correspond to generating sets of toric ideals.

2001 Pistone, Riccomagno and Wynn
Experimental Design using Gröbner Bases

2005 Pachter and Sturmfels
Algebraic Statistics in Computational Biology

2005 Studený
Combinatorics of conditional independence structures

2009 Drton, Sturmfels and Sullivant
Oberwolfach Lecture Notes.

2012 Aoki, Hara and Takemura
Markov Bases

2016 Zwiernik
Tree models using real algebraic geometry

2018 Sullivant
Broad overview of the field.

Example: Markov Chains

A sequence $X_1, X_2, \ldots,\ X_m$ of random variables on the same state space is a Markov Chain if

$$\mathbb{P}\left(X_i = x_i \;\middle|\; X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right) = \mathbb{P}(X_i = x_i \mid X_{i-1} = x_{i-1})$$

or in other words, if the next value only depends on its immediate predecessor.

Example: Markov Chains

A sequence $X_1, X_2, \ldots,\ X_m$ of random variables on the same state space is a Markov Chain if

$$\mathbb{P}\left(X_i = x_i \;\middle|\; X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right) = \mathbb{P}(X_i = x_i \mid X_{i-1} = x_{i-1})$$

or in other words, if the next value only depends on its immediate predecessor.

Let m=3, and the state space $\Sigma = \{0,1\}$. The chain is fully determined by the probabilities of the 8 possible outcome sequences, ie the joint probabilities $p_{ijk} = \mathbb{P}\left(X_1 = i, X_2 = j, X_3 = k\right)$. A full joint probability distribution corresponds to a point $\left(p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}\right) \in \mathbb{R}^8$.

Example: Markov Chains

A sequence $X_1, X_2, \ldots,\ X_m$ of random variables on the same state space is a Markov Chain if

$$\mathbb{P}\left(X_i = x_i \,\middle|\, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}\right) = \mathbb{P}(X_i = x_i \mid X_{i-1} = x_{i-1})$$

or in other words, if the next value only depends on its immediate predecessor.

Let m=3, and the state space $\Sigma = \{0,1\}$. The chain is fully determined by the probabilities of the 8 possible outcome sequences, ie the joint probabilities $p_{ijk} = \mathbb{P}\left(X_1 = i, X_2 = j, X_3 = k\right)$. A full joint probability distribution corresponds to a point $\left(p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}\right) \in \mathbb{R}^8$.

Conditional probabilities for X_3 correspond to ratios $p_{ijk}/(p_{ij0} + p_{ij1})$

Gathering up all these ratios, clearing denominators, and simplifying we can characterize the Markov Chains by:

$$p_{ijk} \geq 0,\ \sum p_{ijk} = 1,\ p_{000}p_{101} - p_{001}p_{100} = 0,\ p_{010}p_{111} - p_{011}p_{110} = 0$$

This defines a semialgebraic set in $\mathbb{R}^8$.

Key Features of Algebraic Statistics

(VERY MANY) STATISTICAL MODELS ARE SEMIALGEBRAIC SETS.

PARAMETRIC STATISTICAL MODELS ARE (OFTEN) POLYNOMIAL FUNCTIONS OF THEIR PARAMETERS.

ESTIMATION AND MODEL FITTING CORRESPONDS TO FINDING POINTS ON VARIETIES OR SEMIALGEBRAIC SETS.

HYPOTHESIS TESTING OF MODEL FIT CORRESPONDS TO CHECKING WHETHER A POINT IS ON A GIVEN VARIETY.

Beyond Algebraic Statistics: Categorical Statistics

Statistics and Probability by creating a category with sufficient structure to enable calculus with string diagrams (ie symmetric monoidal).

- Morphisms are probabilistic functions
- Category contains *copying morphisms* and *deletion morphisms*.

Beyond Algebraic Statistics: Categorical Statistics

Statistics and Probability by creating a category with sufficient structure to enable calculus with string diagrams (ie symmetric monoidal).

- Morphisms are probabilistic functions
- Category contains *copying morphisms* and *deletion morphisms*.

One example: BorelStoch has

- Objects: standard Borel spaces (finite sets, ℕ and [0,1])
- Morphisms: Measurable Markov kernels (generalized Markov transition matrices; a kernel $\kappa : (X, A) \to (Y, B)$ associates to each $x \in X$ a probability measure on (Y, B) so that this association is a measurable map wrt A
  Composition by $(\lambda \circ \kappa)(dz \,|\, x) = \int_Y \lambda\left(dz \,\middle|\, y\right)\kappa(dy \,|\, x)$, ie integrate over all possible intermediary points)
- Monoidal structure by products of measurable spaces.

Categorical Statistics: Theories and Models

(Patterson, 2020)

A *statistical theory* is a small Markov category T with a distinguished *sampling morphism* p.

A *model* of a statistical theory is a functor $T \to Stat$, where Stat is a specific Markov category for modeling statistics.

Categorical Statistics: Theories and Models

(Patterson, 2020)

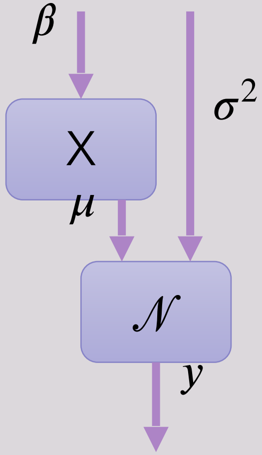
A *statistical theory* is a small Markov category T with a distinguished *sampling morphism* p.

A *model* of a statistical theory is a functor $T \to \mathit{Stat}$, where Stat is a specific Markov category for modeling statistics.

Example:

A linear model with design matrix $X \in \mathbb{R}^{n \times p}$ has sampling distribution $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ with parameters $\beta \in \mathbb{R}^p,\ \sigma^2 \in \mathbb{R}_+$.

A *theory* of a linear model has objects $y,\ \beta,\ \mu,\ \sigma^2$ and morphisms $X : \beta \to \mu$ and $\mathcal{N} : \mu \otimes \sigma^2 \to y$, and sampling morphism:



Categorical Statistics: Theories and Models

(Patterson, 2020)

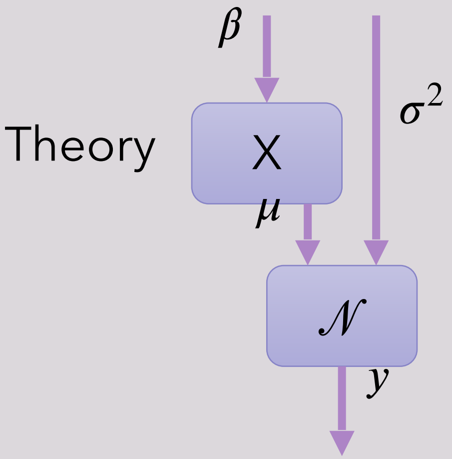
A *statistical theory* is a small Markov category T with a distinguished *sampling morphism* p.

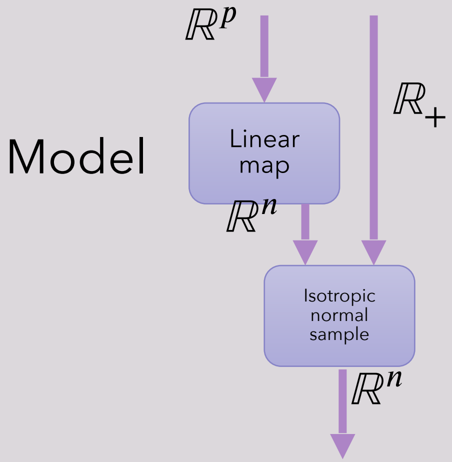
A *model* of a statistical theory is a functor $T \to \mathcal{S}tat$, where Stat is a specific Markov category for modeling statistics.

Example:

A linear model with design matrix $X \in \mathbb{R}^{n \times p}$ has sampling distribution $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ with parameters $\beta \in \mathbb{R}^p,\ \sigma^2 \in \mathbb{R}_+$.

A *theory* of a linear model has objects $y,\ \beta,\ \mu,\ \sigma^2$ and morphisms $X : \beta \to \mu$ and $\mathcal{N} : \mu \otimes \sigma^2 \to y$, and sampling morphism:





Categorical Statistics: Theories and Models

(Patterson, 2020)

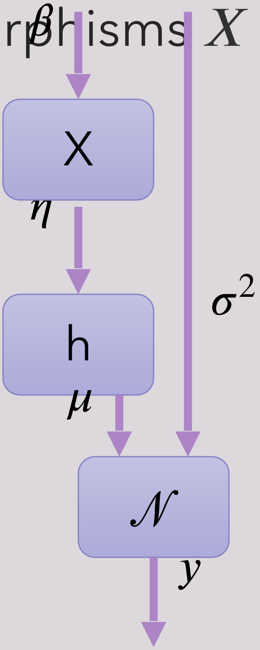
A *statistical theory* is a small Markov category T with a distinguished *sampling morphism* p.

A *model* of a statistical theory is a functor $T \to Stat$, where Stat is a specific Markov category for modeling statistics.

Example:

A *general* linear model has sampling distribution $y \sim \mathcal{N}(h(X\beta), \sigma^2 I_n)$ with h an invertible *link function*.

A *theory* of a general linear model has objects $y,\ \beta,\ \mu,\ \eta,\ \sigma^2$ and morphisms $X : \beta \to \eta, h : \eta \to \mu$ and $\mathcal{N} : \mu \otimes \sigma^2 \to y$:



Categorical Statistics: Theories and Models

(Patterson, 2020)

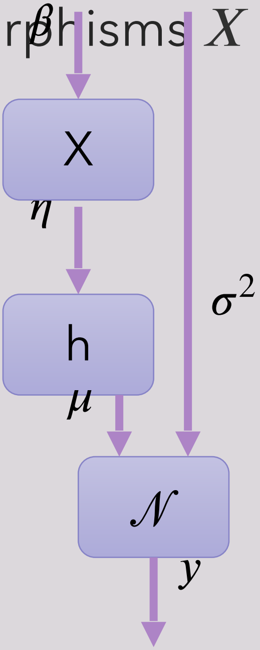
A *statistical theory* is a small Markov category T with a distinguished *sampling morphism* p.

A *model* of a statistical theory is a functor $T \to Stat$, where Stat is a specific Markov category for modeling statistics.

Example:

A *general* linear model has sampling distribution $y \sim \mathcal{N}(h(X\beta), \sigma^2 I_n)$ with h an invertible *link function*.

A *theory* of a general linear model has objects $y,\ \beta,\ \mu,\ \eta,\ \sigma^2$ and morphisms $X : \beta \to \eta$, $h : \eta \to \mu$ and $\mathcal{N} : \mu \otimes \sigma^2 \to y$:



Setting $\eta = \mu$ and choosing $h = Id$ creates a *theory morphism*

$G : GLM \to LM$

which induces a *model migration functor*

$G^* : Mod(LM) \to Mod(GLM)$

Thank you for listening

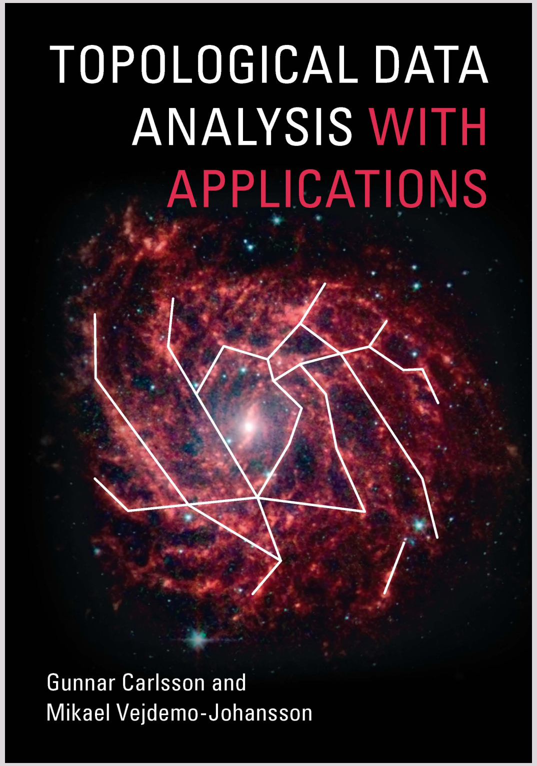
Topological Data Analysis
Use linear algebra to compute homology on data sets measuring their clusters, holes and bubbles.

Geometric Data Analysis
Use manifolds to estimate point cloud data.

Information Geometry
Use differentiable manifolds to study parametrized distributions.

Algebraic Statistics
Use algebraic geometry to study statistics – also: use category theory to study statistical models.

Thank you for listening



TOPOLOGICAL DATA ANALYSIS WITH APPLICATIONS

Gunnar Carlsson and
Mikael Vejdemo-Johansson

www.cambridge.org/9781108838658

20% discount code:
TDAA2021

Topological Data Analysis
Use linear algebra to compute homology on data sets measuring their clusters, holes and bubbles.

Geometric Data Analysis
Use manifolds to estimate point cloud data.

Information Geometry
Use differentiable manifolds to study parametrized distributions.

Algebraic Statistics
Use algebraic geometry to study statistics – also: use category theory to study statistical models.