Topological Data Analysis

Mikael Vejdemo-Johansson

School of Computer Science
University of St Andrews
Scotland

22 November 2011

Outline

1. The Shape of Data
2. Topological Data Analysis
   - Persistent Homology
   - Signal Analysis
   - Robotics
   - Biomedicine
   - Natural images analysis

Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

What is data?

Data is a collection of observations.

`http://en.wikipedia.org/wiki/Data_analysis` gives three commonly used categories:

Quantitative	Measured by some (real) number.
Categorical	Assigned to one of several possible categories.
Qualitative	Measured by presence or absence of some characteristic.

A *datum* will be some collection of such observations. There are interesting metrics for all, which allows us to define:

A *data set* is a finite metric space.

Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Tasks of data analysis

Summarize Provide a description that is, preferably, smaller than the dataset.

Model Provide a description that allows for predictions of the behaviour of the source of the data.

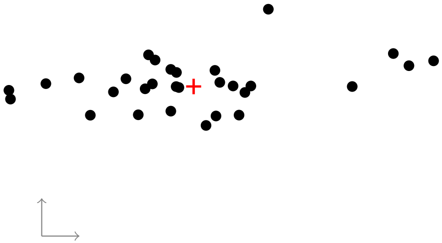
Highlight Provide emphasis on certain interesting properties of the data.

Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Fundamental data analysis techniques

Mean (centroid) tells us *where* the data is located.

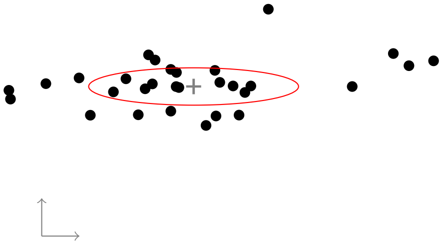


Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Fundamental data analysis techniques

Standard deviation tells us how spread out the data is.

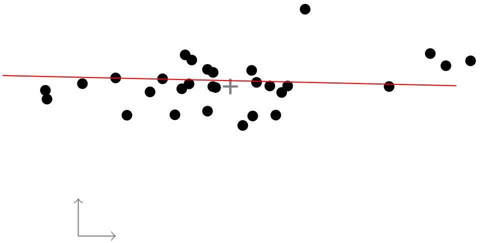


Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Fundamental data analysis techniques

Regression analyses fit the data to an easy to analyze model.

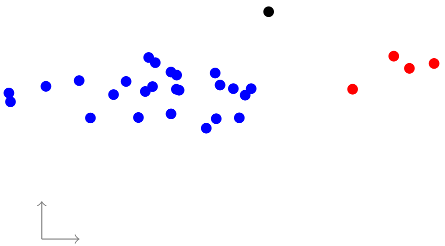


Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Fundamental data analysis techniques

Cluster analysis divides the data into its connected components.

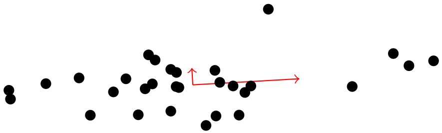


Shape of data

Fundamentally, data analysis is the task of describing the shape of data:

Fundamental data analysis techniques

Principal Component Analysis (and other dimension reduction techniques) give a new coordinate frame that more faithfully represent the data.



Outline

1. The Shape of Data

2. Topological Data Analysis
   - Persistent Homology
   - Signal Analysis
   - Robotics
   - Biomedicine
   - Natural images analysis

Wherefore topology?

Issues with classical data analysis

Unreliable metric The metrics in use may well not be accurate measures of dissimilarity as distances grow.

Ill motivated metric The metrics in use may be arbitrarily chosen, not well anchored as distances grow.

Noisy data Data may be very noisy.

High-dimensional data Data may be very high-dimensional, and thus slow to process.

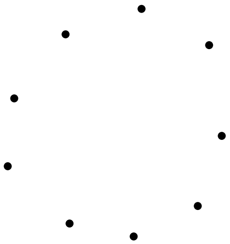
Topology only depends on a notion of *nearness*. Produces dimension-agnostic qualitative features.

Discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

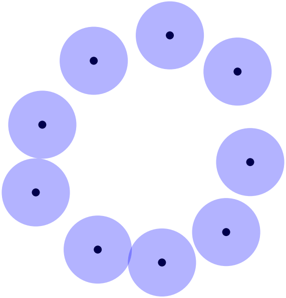


Discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

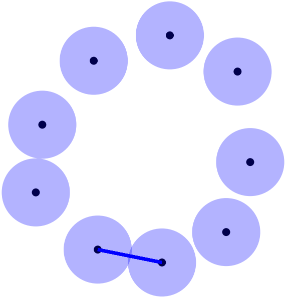


Discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

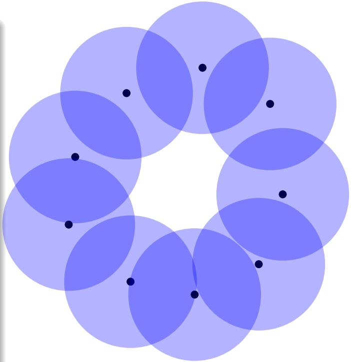


Discrete made continuous

Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.

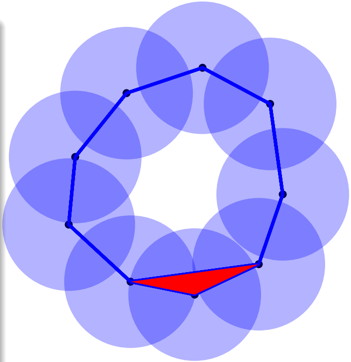


Discrete made continuous

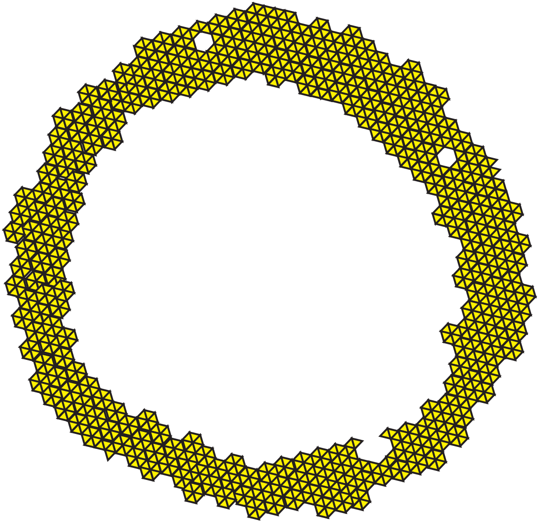
Definition

The Vietoris-Rips complex is an abstract simplicial complex $VR_\epsilon(X)$ for $\epsilon \in \mathbb{R}_+$ and X a finite metric space:

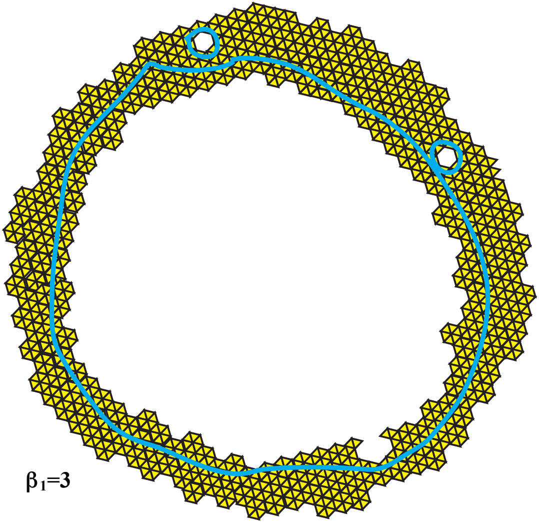
- Contains one vertex for each element in X.
- Contains a simplex $(x_0, \ldots, x_k)$ exactly when $d(x_i, d_j) < \epsilon$ for all $i, j \in [k]$.



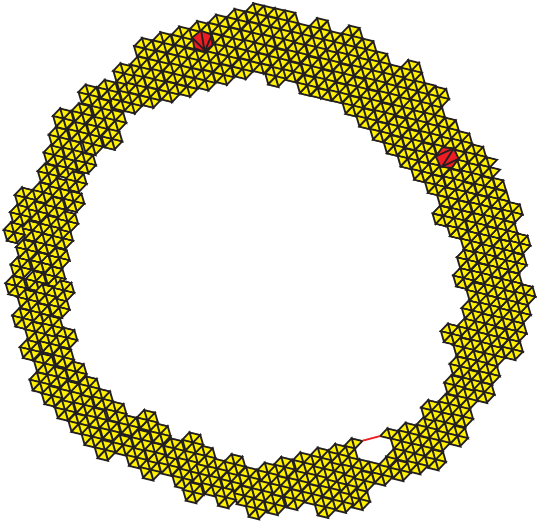
Functoriality and persistent homology



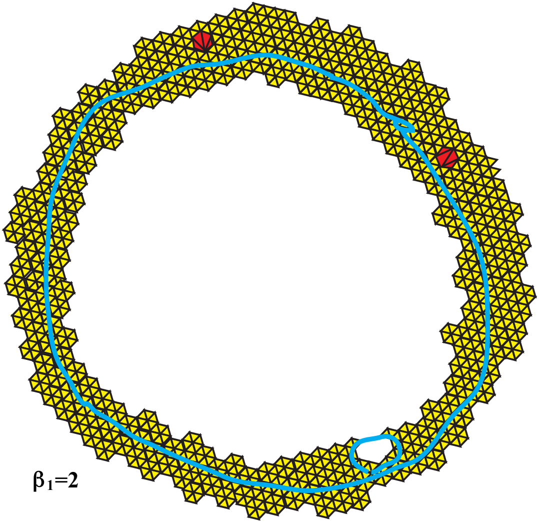
Functoriality and persistent homology



Functoriality and persistent homology



Functoriality and persistent homology



Functoriality and persistent homology

Homology is a *functor*: if $f : X \to Y$ is a function, there is an induced function $H_*(f) : H_*X \to H_*Y$.

If $\varepsilon < \varepsilon'$, then $VR_\varepsilon(X) \subseteq VR_{\varepsilon'}(X)$.

Definition

The *persistent homology space* $H_n^{\varepsilon,\varepsilon'}(VR_*(X))$ is the subspace of the vector space $H_n(VR_{\varepsilon'}(X))$ consisting of the image of the induced map $H_n(VR_\varepsilon(X)) \to H_n(VR_{\varepsilon'}(X))$.

Barcodes

We can summarize, pictorially, the collection of persistent homology spaces as a *barcode*.

Definition

The *persistence barcode* for a filtered complex $VR_*(X)$ is a collection of pairs (s, t).
(s, t) is in the barcode if there is a basis element of $H_n^{s,t}(VR_*(X))$ not in $H_n^{s-\varepsilon,t}(VR_*(X))$ nor in $H_n^{s,t+\varepsilon}(VR_*(X))$.
s and t can take values in positive reals, and ∞.

The dendrogram of single linkage clustering is (almost) exactly the barcode for H_0.

Delay embedding quality

Delay Embedding

$$\{a_x\} \mapsto \{(a_x, a_{x+\varepsilon}, \ldots, a_{x+(d-1)\varepsilon}\}$$

Converts a 1-dimensional signal into a d-dimensional signal.

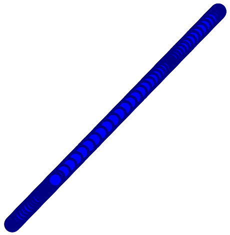
Problem

Choose appropriate parameters ε, d.

Topology helps for *periodic* signals: closed curves are embeddings of S^1, recognizable by Betti numbers.

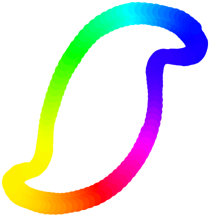
Detecting good delay embeddings (de Silva—Skraba—VJ)

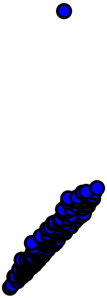
Clarinet middle E tone in $\mathbb{R}^2$:



Detecting good delay embeddings (de Silva—Skraba—VJ)

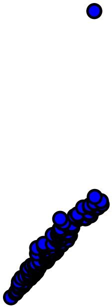
Clarinet middle E tone in $\mathbb{R}^2$:





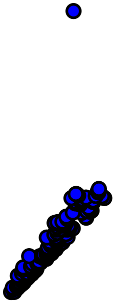
Detecting good delay embeddings (de Silva—Skraba—VJ)

Clarinet middle E tone in $\mathbb{R}^2$:



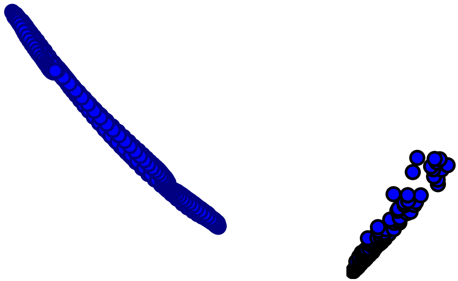
Detecting good delay embeddings (de Silva—Skraba—VJ)

Clarinet middle E tone in $\mathbb{R}^2$:



Detecting good delay embeddings (de Silva—Skraba—VJ)

Clarinet middle E tone in $\mathbb{R}^2$:



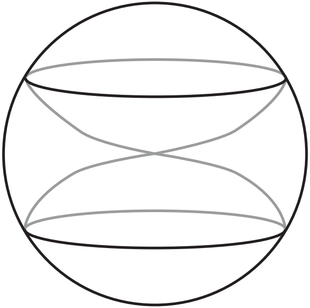
Cyclo-octane configurations

Working to replicate an analysis in *Topology of cyclo-octane energy landscape* (Martin, Thompson, Coutsias, Watson; J. Chem. Phys. **132**, 234115 (2010)), we consider cyclo-octane as a linkage. Martin-Thompson-Coutsias-Watson established the topology of this to be a sphere and a Klein bottle, fused along two circles. They also computed the Betti numbers to be $\beta_0 = 1$, $\beta_1 = 1$, and $\beta_2 = 2$.

Requiring rest-state distances between atoms, and rest-state planar angles for carbon-carbon bonds, the resulting linkage has only rotational joints at each carbon atom.

We sampled 34k points from the resulting variety, and are using this as a test-case for a systematic method for the computation.

Worked example: Cyclo-octane configurations



- Sphere
- Klein bottle
- Intersect in disjoint, unlinked pair of circles.

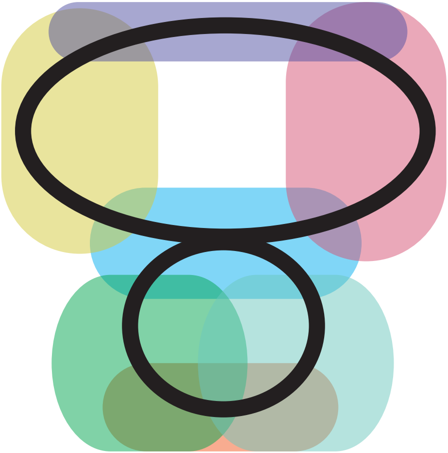
A topological analysis method

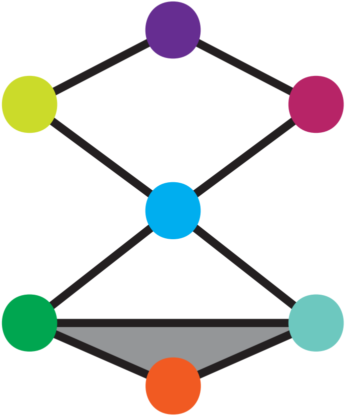
In a recent PhD thesis at Stanford [Singh, '08], a topological method for data analysis was introduced.

Fundamental topological result: Nerve lemma

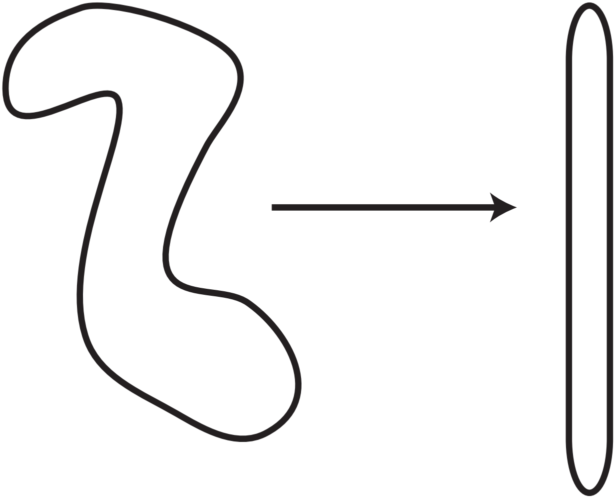
Suppose a space X is subdivided $X = \bigcup_i X_i$ into *contractible* (read simple) components.
Then X is equivalent to the *nerve* of the covering.

The nerve of a covering

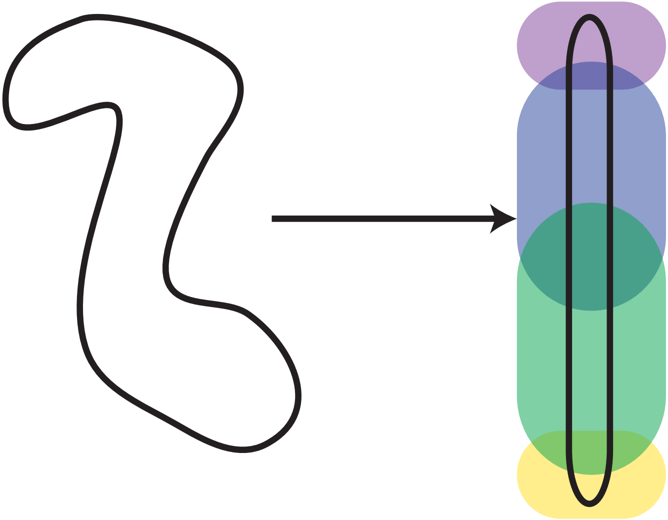




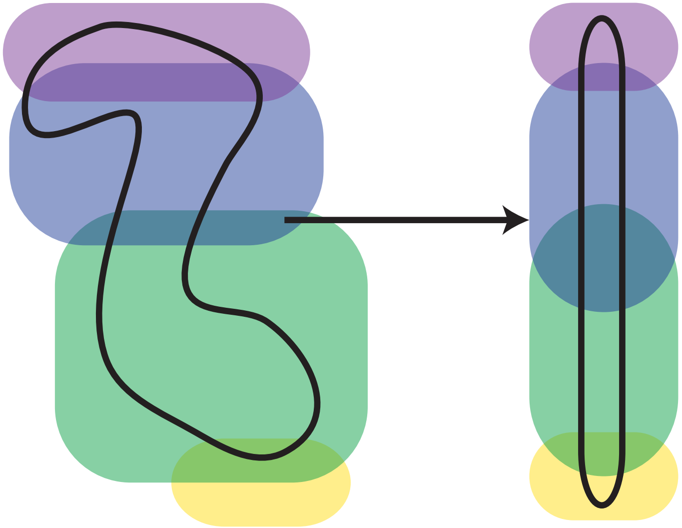
Topological application



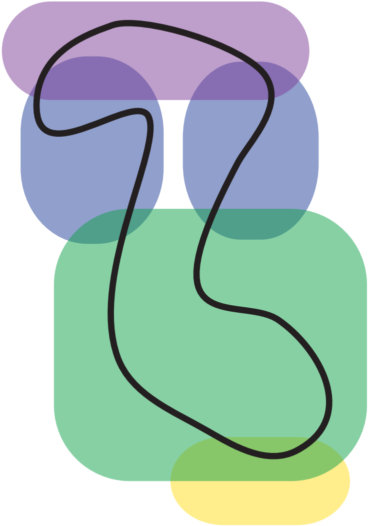
Topological application



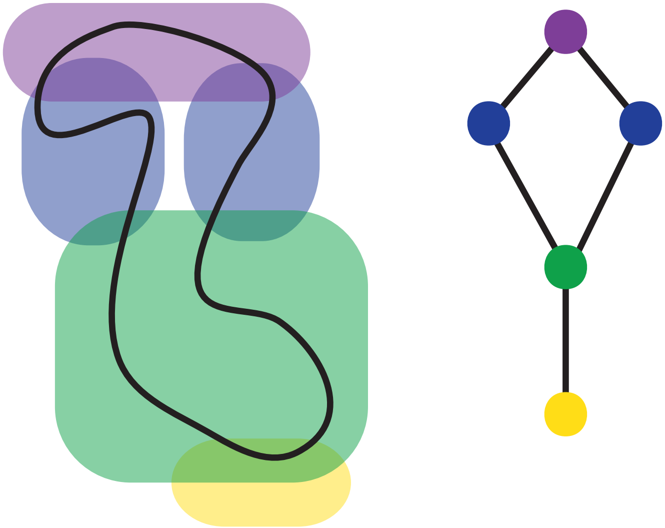
Topological application



Topological application



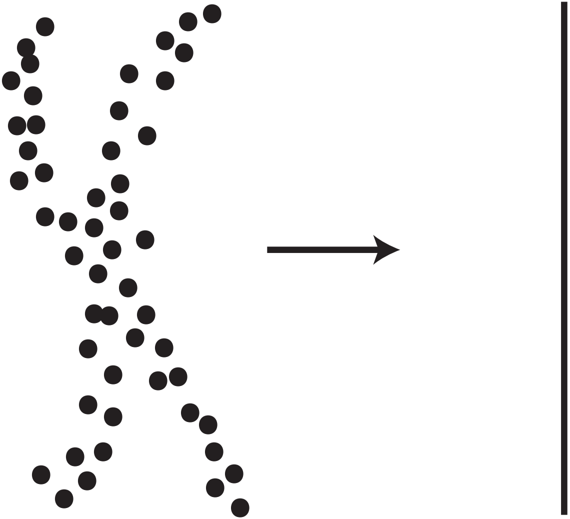
Topological application



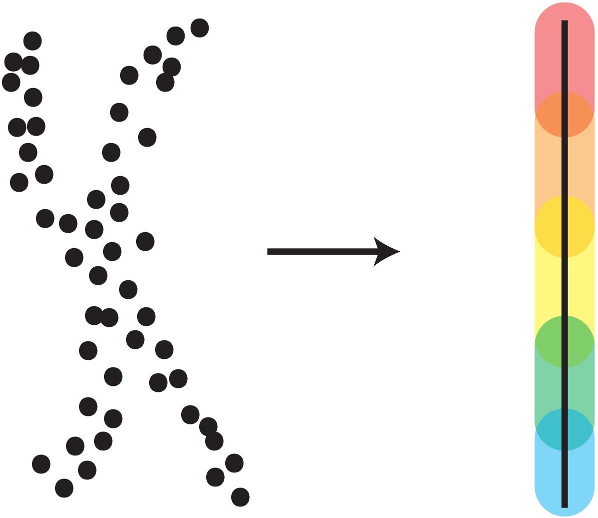
Translate topology to statistics

Continuous function	Measurement function on datapoints
Covering of target space	Covering of datapoints
Preimages	Preimages
Connected components	Clusters
Nerve complex	Mapper diagram

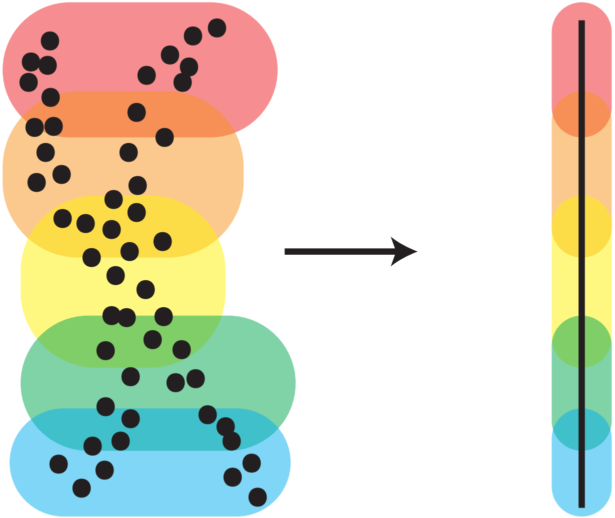
Mapper algorithm



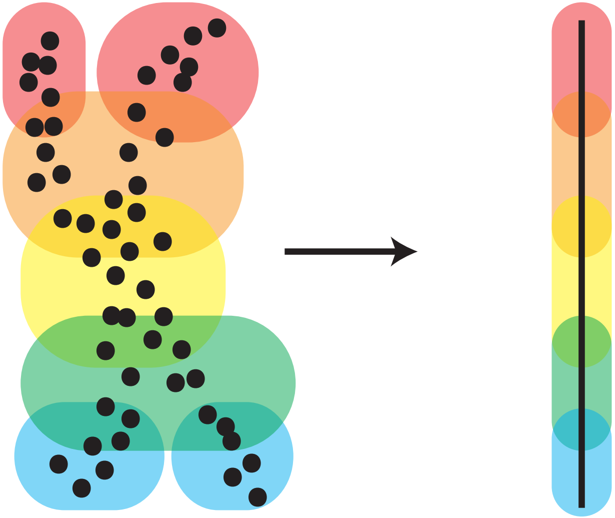
Mapper algorithm



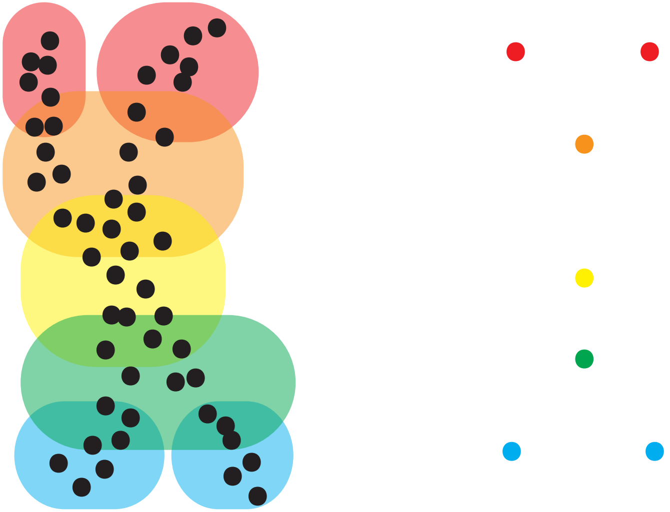
Mapper algorithm



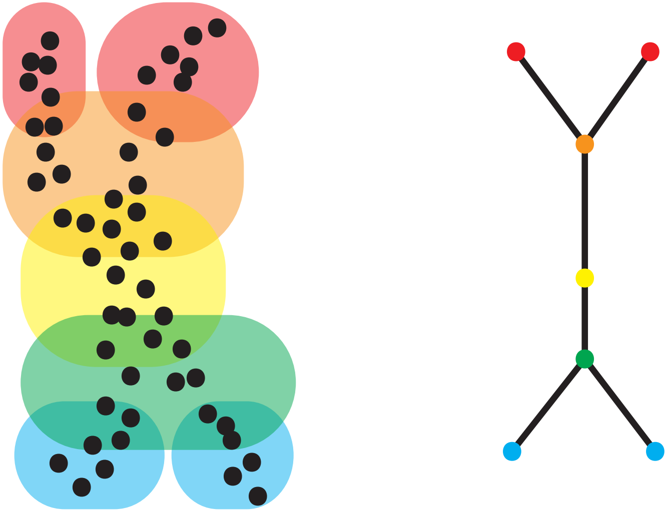
Mapper algorithm



Mapper algorithm



Mapper algorithm



Mapper algorithm

Implementation

This method is provided in a software package currently marketed by Ayasdi.
Startup company founded by Gurjeet Singh (original thesis on Mapper) and Gunnar Carlsson (thesis advisor).

Ayasdi
Mapping worlds of data

Cancer data

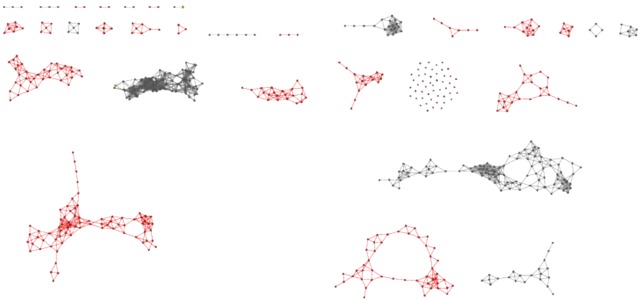
Carlsson – Nicolau and the team at Ayasdi studied physiological data from around 170 breast cancer patients.

Mapper plot structured as a core with flares extending.

One flare consisted exclusively of survivors (0% mortality). Cluster analyses and PCA techniques dispersed this group among high mortality patients.

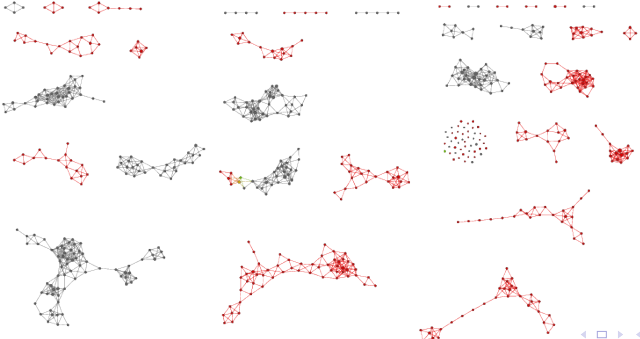
Political data

Carlsson – Lum – Sandberg – V-J and the team at Ayasdi have studied vote data from the US congress.



Political data

Carlsson – Lum – Sandberg – V-J and the team at Ayasdi have studied vote data from the US congress.



Example: Spaces of natural images

Lee-Mumford-Pedersen investigated whether a statistically significant difference exists between natural and random images.

Natural images form a "subspace" of all images. Dimension of ambient space e.g. $640 \times 480 = 307\,200$.

This space of natural images should have:

- high dimension: there are many different images.
- high codimension: random images look nothing like natural ones.

Natural 3x3 patches

Instead of studying entire images, we consider the distribution of 3×3 pixel patches.

Most of these will be approximately constant in natural images. Allowing these drowns out structure.

Lee-Mumford-Pedersen chose $8\,500\,000$ patches with high contrast from a collection of black-and-white images used in cognition research. Each 3×3-patch is considered a vector in $\mathbb{R}^9$.

Normalised brightness: $\mathbb{R}^9 \to \mathbb{R}^8$. Normalised contrast: $\mathbb{R}^8 \to S^7$.

Subsequent topological analysis by Carlsson–de Silva–Ishkanov–Zomorodian.

Pixel patches in S^7

The resulting patches are dense in S^7 – so we consider high-density regions.

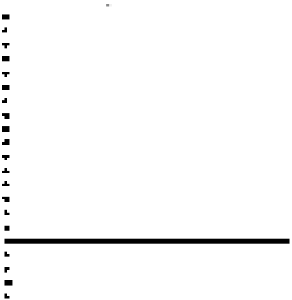
Pick out 25% densest points. We can pick a parametrised method to measure density:

Definition

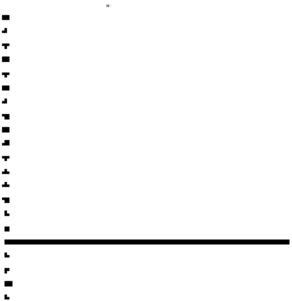
k-codensity $\delta_k(x)$ of a point x is the distance to its kth nearest neighbour. k-density $d_k(x)$ is $1/\delta_k(x)$.

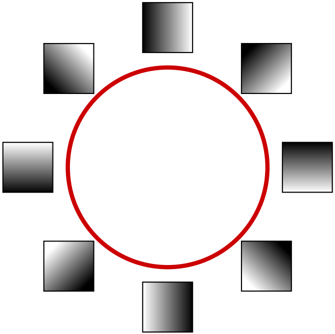
High k yields a smoothly changing density measure capturing global properties. Low k yields a wilder density measure capturing local properties. k acts as a kind of focus control.

300-density

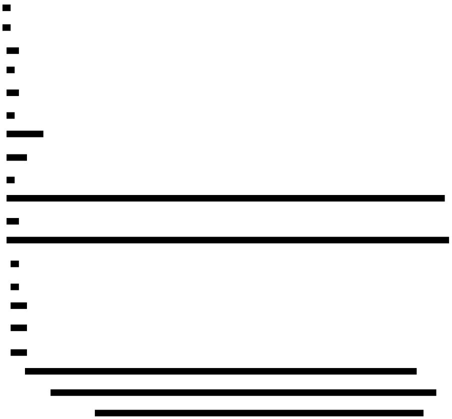


300-density

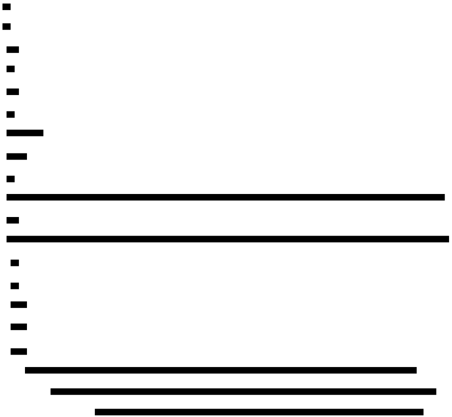


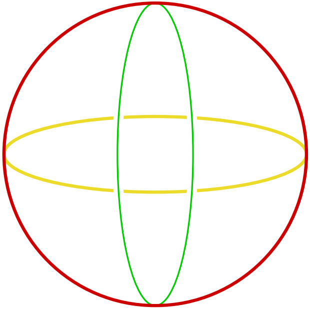


15-density

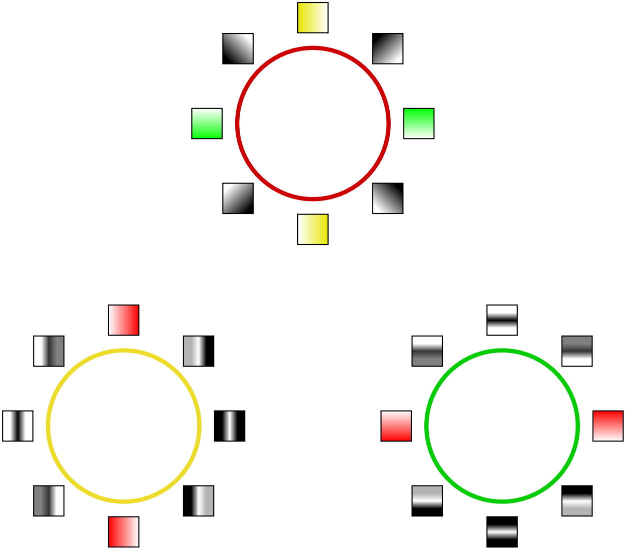


15-density





Three circles

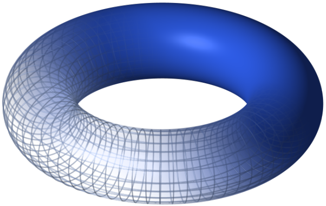


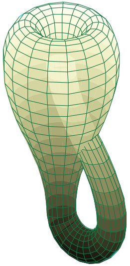
Identifying the subspace of natural pixel patches

Raising the cut-off bar yields, with coefficients in $\mathbb{F}_2$

$$\beta_0 = 1 \qquad \beta_1 = 2 \qquad \beta_2 = 1$$

Assuming the shape is a surface, this corresponds to one of



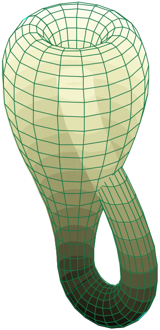


Identifying the subspace of natural pixel patches

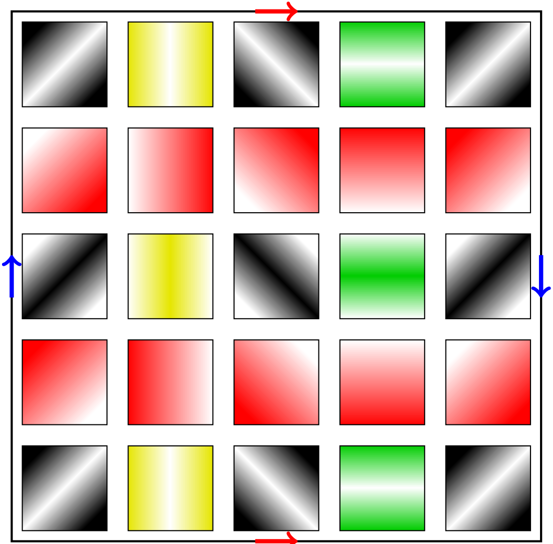
Raising the cut-off bar yields, with coefficients in $\mathbb{F}_3$

$$\beta_0 = 1 \qquad \beta_1 = 1$$

Thus, the relevant shape is:



Klein bottle of pixel patches



Applications of this analysis

Image compression

A 3×3-cluster may be described using 4 values:

- Position of its projection onto the Klein bottle
- Original brightness
- Original contrast

Texture analysis

Textures yield distributions of occuring patches on the Klein bottle. Rotating the texture corresponds to translating the distribution. [J Perea]