Topology and Data

Gunnar Carlsson [1]
Department of Mathematics
Stanford University
http://comptop.stanford.edu/

June 27, 2008

[1] Research supported in part by DARPA and NSF

Introduction / Motivations

- General area of *geometric data analysis* attempts to give insight into data by imposing a geometry on it

Introduction

- General area of *geometric data analysis* attempts to give insight into data by imposing a geometry on it
- Sometimes very natural (physics), sometimes less so (genomics)

  基因组学

Introduction

- General area of *geometric data analysis* attempts to give insight into data by imposing a geometry on it
- Sometimes very natural (physics), sometimes less so (genomics)
- Value of geometry is that it allows us to organize and view data more effectively, for better understanding

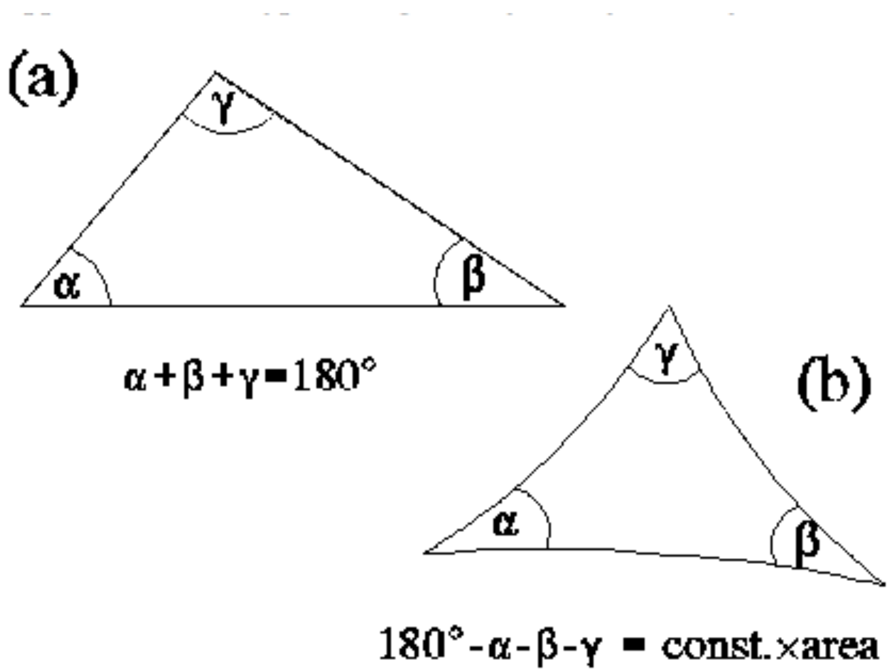
Introduction

- General area of *geometric data analysis* attempts to give insight into data by imposing a geometry on it
- Sometimes very natural (physics), sometimes less so (genomics)
- Value of geometry is that it allows us to organize and view data more effectively, for better understanding
- Can obtain an idea of a reasonable layout or overview of the data

Introduction

- General area of *geometric data analysis* attempts to give insight into data by imposing a geometry on it
- Sometimes very natural (physics), sometimes less so (genomics)
- Value of geometry is that it allows us to organize and view data more effectively, for better understanding
- Can obtain an idea of a reasonable layout or overview of the data
- Sometimes all that is required is a qualitative overview

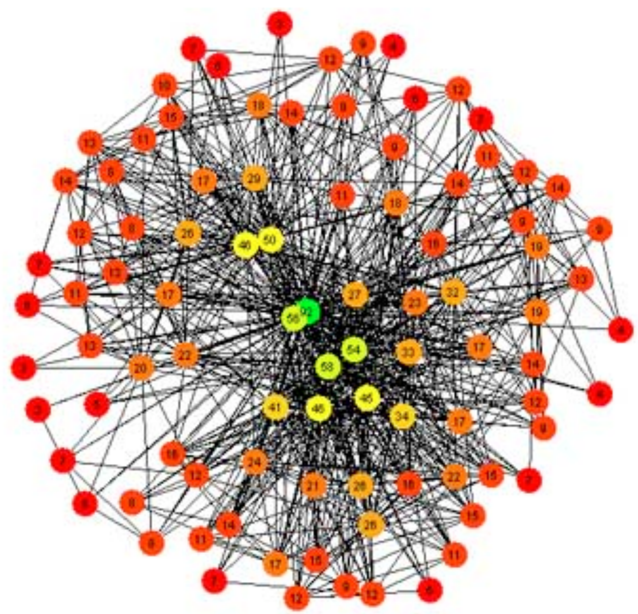
Methods for Imposing a Geometry



hyperbolic metric
(non-Euclidean)

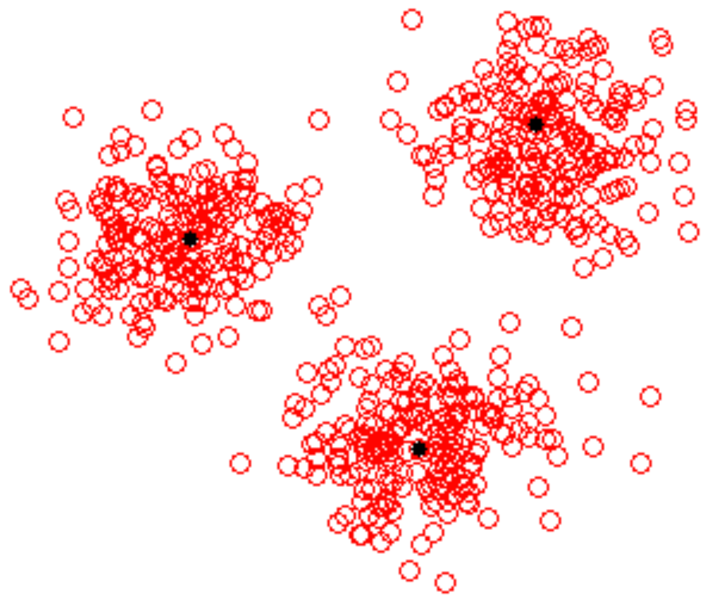
Define a metric

Methods for Imposing a Geometry



Define a graph or network structure

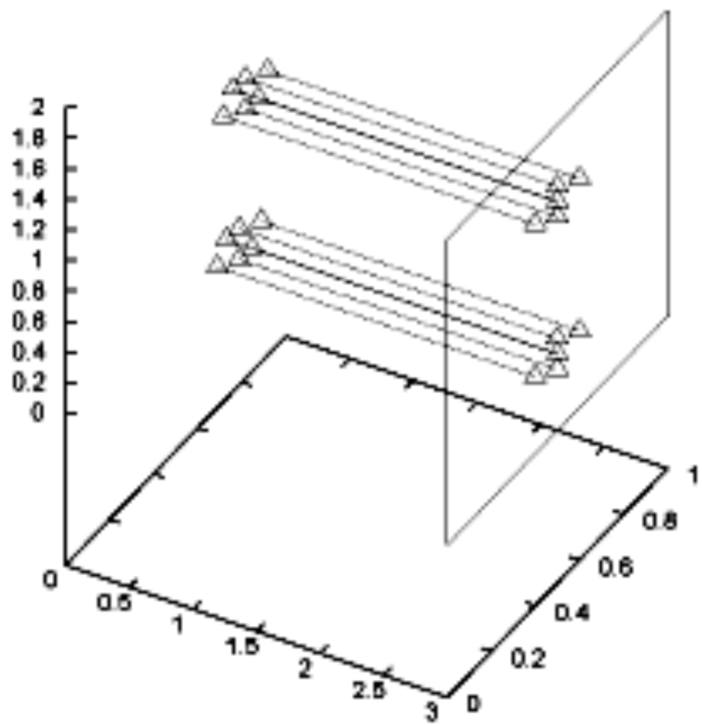
Methods for Imposing a Geometry



Cluster the data

聚类

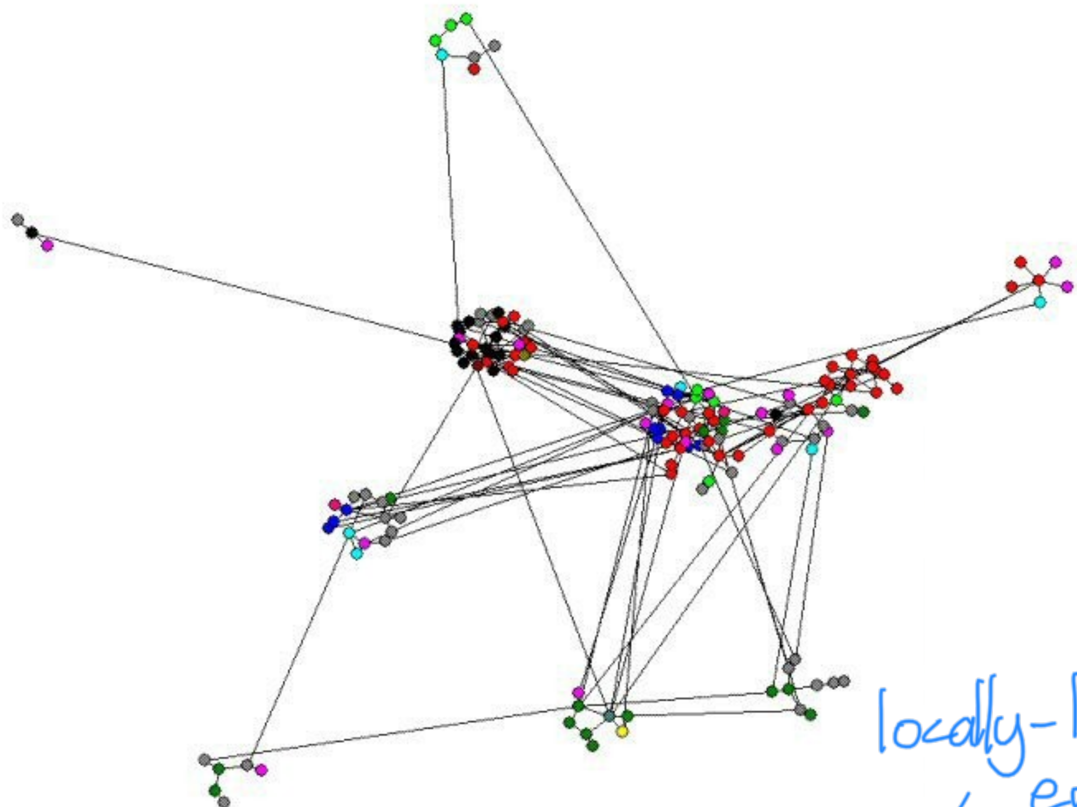
Methods for Summarizing or Visualizing a Geometry



dimensionality reduction

Linear projections

Methods for Summarizing or Visualizing a Geometry



locally-linear embedding

Multidimensional scaling, ISOMAP, LLE

nonlinear dimensionality reduction

Methods for Summarizing or Visualizing a Geometry

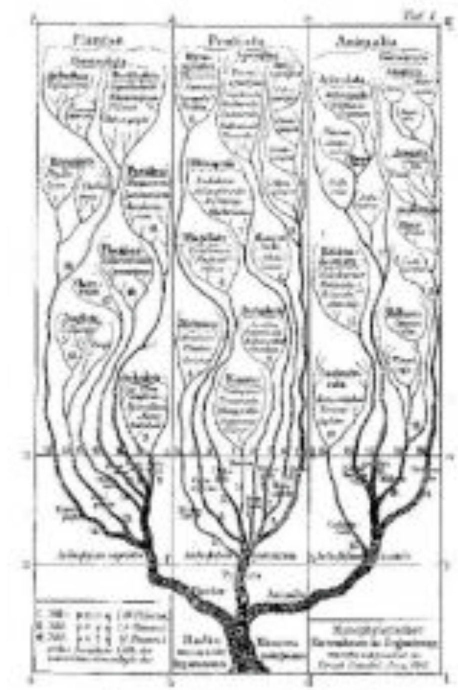


Figure 1: Haeckel's tree with 3 branches

Project to a tree

Properties of Data Geometries

We Don't Trust Large Distances

Properties of Data Geometries

We Don't Trust Large Distances

- In physics, distances have strong theoretical backing, and should be viewed as reliable

We Don't Trust Large Distances

- In physics, distances have strong theoretical backing, and should be viewed as reliable
- In biology or social sciences, distances are constructed using a notion of similarity, but have no theoretical backing (e.g. Jukes-Cantor distance between sequences)

models of DNA evolution

evolutionary distance (in terms of the expected number of changes) between two sequences:

$$d = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

p: portion of sites that differ between the two sequences

Properties of Data Geometries

We Don't Trust Large Distances

- In physics, distances have strong theoretical backing, and should be viewed as reliable
- In biology or social sciences, distances are constructed using a notion of similarity, but have no theoretical backing (e.g. Jukes-Cantor distance between sequences)
- Means that small distances still represent similarity, but comparison of long distances makes little sense

Properties of Data Geometries

We Only Trust Small Distances a Bit

Properties of Data Geometries

We Only Trust Small Distances a Bit

Properties of Data Geometries

We Only Trust Small Distances a Bit

- Both pairs are regarded as similar, but the strength of the similarity as encoded by the distance may not be so significant

Properties of Data Geometries

We Only Trust Small Distances a Bit

- Both pairs are regarded as similar, but the strength of the similarity as encoded by the distance may not be so significant
- Similarity more like a 0/1-valued quantity than $\mathbb{R}$-valued

Properties of Data Geometries

Connections are Noisy

Properties of Data Geometries

Connections are Noisy

- Distance measurements are noisy, as are the connections in many graph models

Properties of Data Geometries

Connections are Noisy

- Distance measurements are noisy, as are the connections in many graph models
- Requires stochastic geometric methods for study

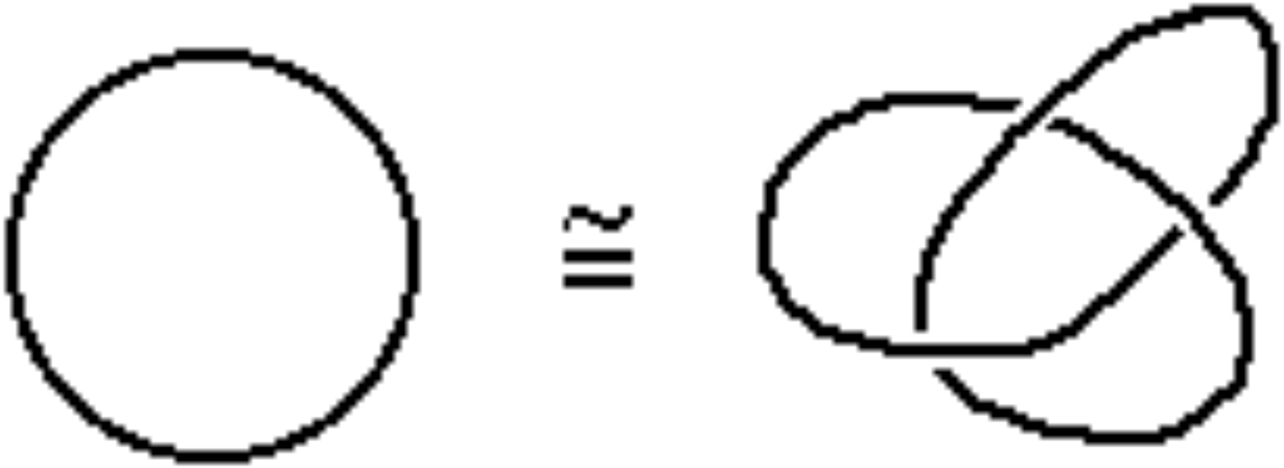
Properties of Data Geometries

Connections are Noisy

- Distance measurements are noisy, as are the connections in many graph models
- Requires stochastic geometric methods for study
- Methods of Coifman et al and others relevant here

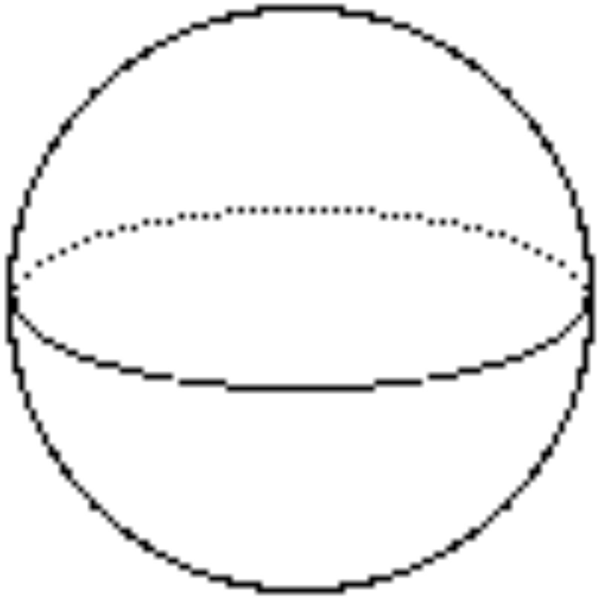
Geometric diffusions as a tool for harmonic analysis and structure definition of data
(Proceedings of the National Academy of Sciences) 2005

Topology

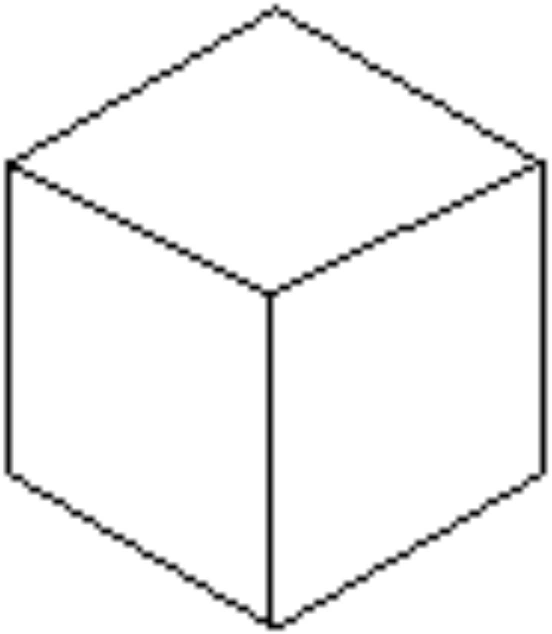


Homeomorphic

Topology



$\cong$



Homeomorphic

Topology

- To see that these pairs are "same" requires distortion of distances, i.e. stretching and shrinking

Topology

- To see that these pairs are "same" requires distortion of distances, i.e. stretching and shrinking
- We do not permit "tearing", i.e. distorting distances in a discontinuous way

Topology

- To see that these pairs are "same" requires distortion of distances, i.e. stretching and shrinking
- We do not permit "tearing", i.e. distorting distances in a discontinuous way
- How to make this precise?

Topology

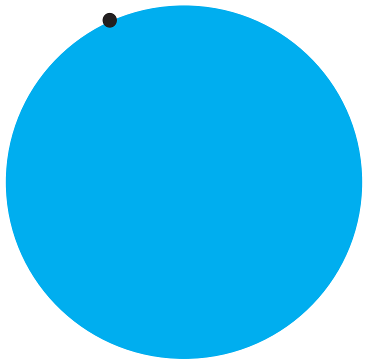
- One would like to say that all non-zero distances in a metric space are the same

Topology

- One would like to say that all non-zero distances in a metric space are the same
- But, $d(x, y) = 0$ means $x = y$

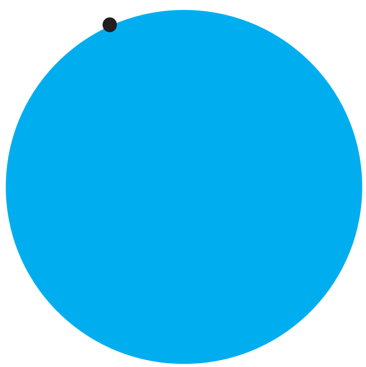
Topology

- One would like to say that all non-zero distances in a metric space are the same
- But, $d(x, y) = 0$ means $x = y$
- Idea: consider instead distances from points to subsets. Can be zero.



Topology

- One would like to say that all non-zero distances in a metric space are the same
- But, $d(x, y) = 0$ means $x = y$
- Idea: consider instead distances from points to subsets. Can be zero.



This accomplishes the intuitive idea of permitting arbitrary rescalings of distances while leaving "infinite nearness" intact.

closure

Topology

- Topology is the idealized form of what we want in dealing with data, namely permitting arbitrary rescalings which vary over the space

Topology

- Topology is the idealized form of what we want in dealing with data, namely permitting arbitrary rescalings which vary over the space
- Now must make versions of topological methods which are "less idealized"

Topology

- Topology is the idealized form of what we want in dealing with data, namely permitting arbitrary rescalings which vary over the space
- Now must make versions of topological methods which are "less idealized"
- Means in particular finding ways of tracking or summarizing behavior as metrics are deformed or other parameters are changed

Topology

- Topology is the idealized form of what we want in dealing with data, namely permitting arbitrary rescalings which vary over the space
- Now must make versions of topological methods which are "less idealized"
- Means in particular finding ways of tracking or summarizing behavior as metrics are deformed or other parameters are changed
- Ultimately means building in noise and uncertainty. This is in the future - "statistical topology".

Outline

1. Homology as signature for shape identification

Outline

1. Homology as signature for shape identification
2. Image processing example

Outline

1. Homology as signature for shape identification
2. Image processing example
3. Topological "imaging" of data

Outline

1. Homology as signature for shape identification
2. Image processing example
3. Topological “imaging” of data
4. Signatures for significance of structural invariants

Persistent Homology

- Homology: crudest measure of topological properties

Persistent Homology

- Homology: crudest measure of topological properties
- For every space X and dimension k, constructs a vector space $H_k(X)$ whose dimension (the k-th *Betti number* β_k) is a mathematically precise version of the intuitive notion of counting "k-dimensional holes"

Persistent Homology

- Homology: crudest measure of topological properties
- For every space X and dimension k, constructs a vector space $H_k(X)$ whose dimension (the k-th *Betti number* β_k) is a mathematically precise version of the intuitive notion of counting "k-dimensional holes"
- Computed using linear algebraic methods, basically Smith normal form

Gaussian elimination

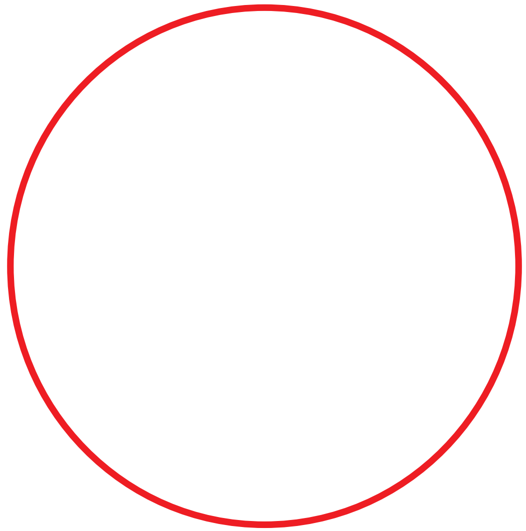
Persistent Homology

- Homology: crudest measure of topological properties
- For every space X and dimension k, constructs a vector space $H_k(X)$ whose dimension (the k-th *Betti number* β_k) is a mathematically precise version of the intuitive notion of counting "k-dimensional holes"
- Computed using linear algebraic methods, basically Smith normal form
- β_0 is a count of the number of connected components

Persistent Homology

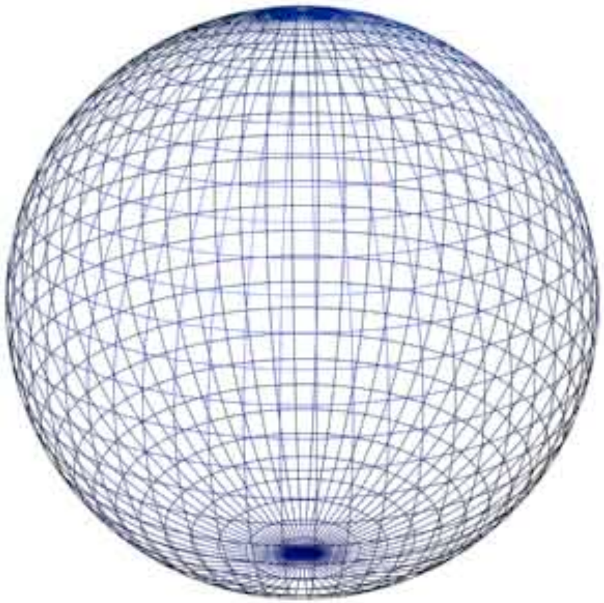
- Homology: crudest measure of topological properties
- For every space X and dimension k, constructs a vector space $H_k(X)$ whose dimension (the k-th *Betti number* β_k) is a mathematically precise version of the intuitive notion of counting "k-dimensional holes"
- Computed using linear algebraic methods, basically Smith normal form
- β_0 is a count of the number of connected components
- β_i's form a signature which encodes topological information about the shape

Persistent Homology



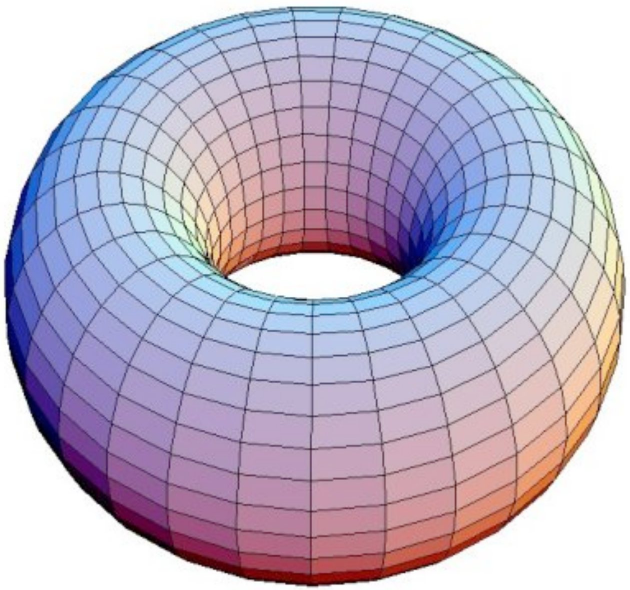
$\beta_0 = 1$, $\beta_1 = 1$, and $\beta_i = 0$ for $i \geq 2$

Persistent Homology



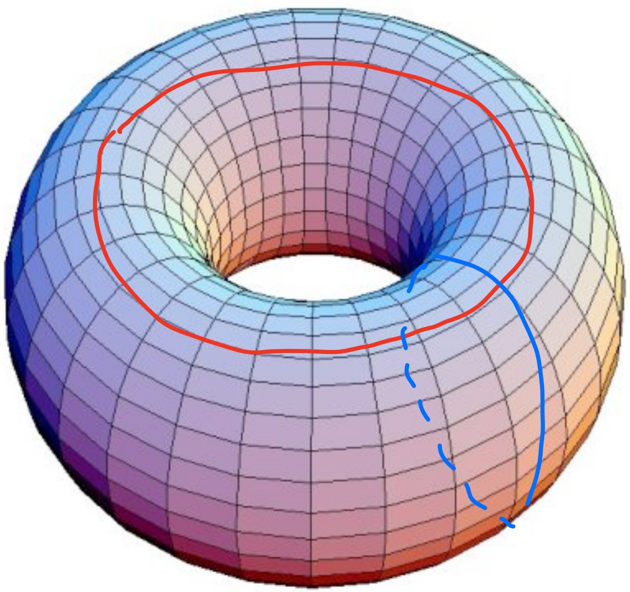
$\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_k = 0$ for $k \geq 3$

Persistent Homology



$\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_k = 0$ for $k \geq 3$

Persistent Homology



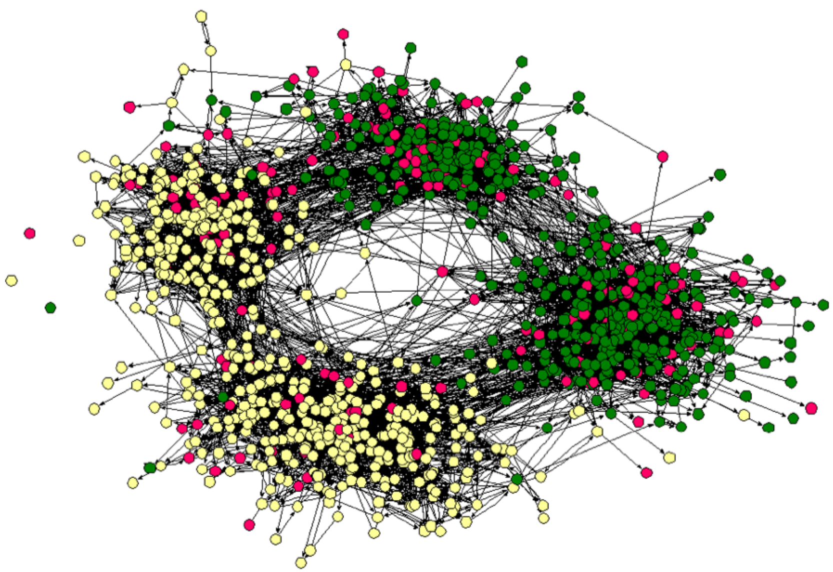
$\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_k = 0$ for $k \geq 3$

Persistent Homology

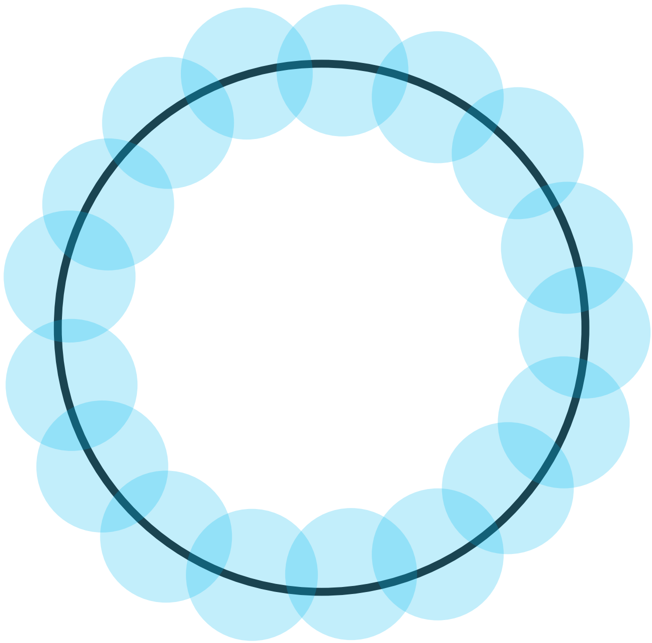
modeling

Question: For a point cloud X , can one infer the Betti numbers of the space $\mathbb{X}$ from which it is sampled?

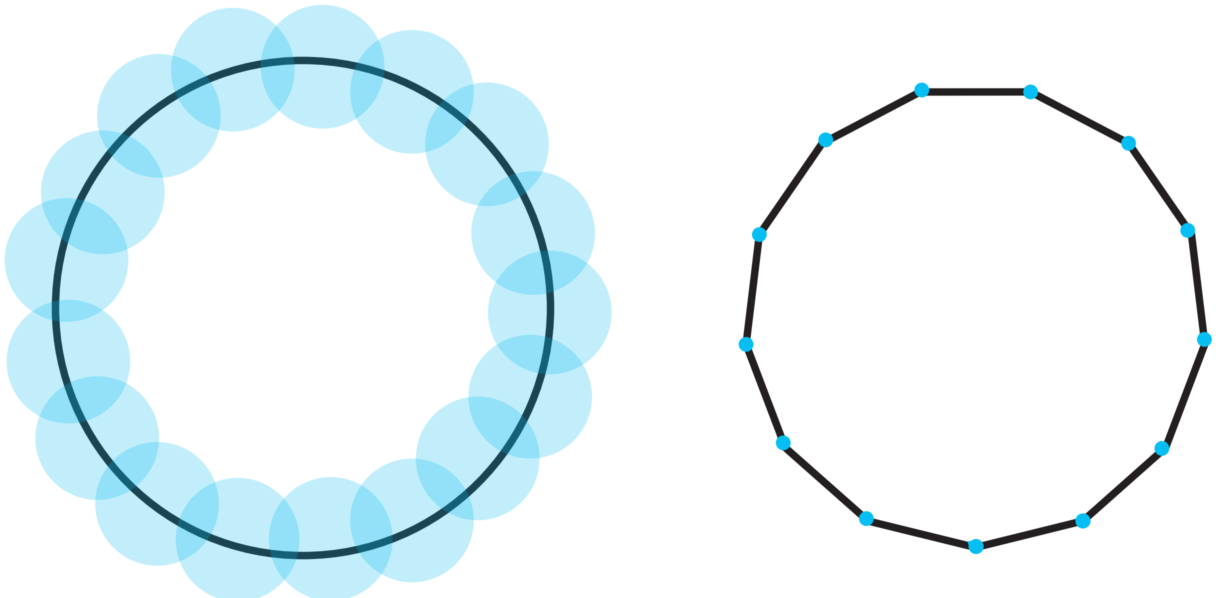
reality



Persistent Homology - Čech Complex

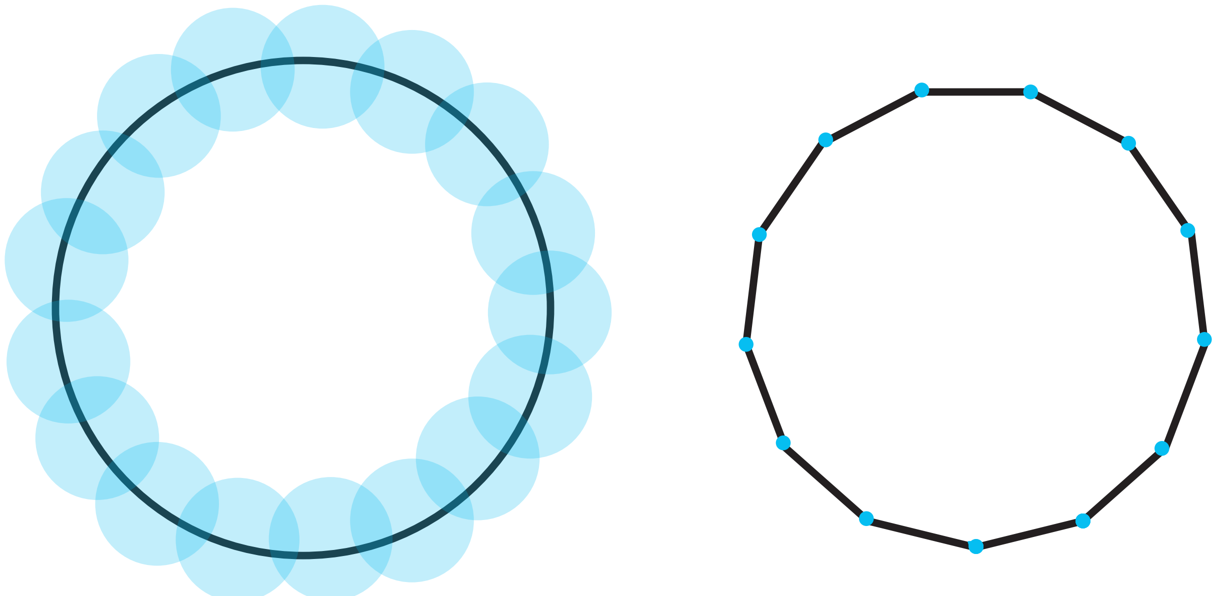


Persistent Homology - Čech Complex



$\check{C}(X, \epsilon)$ - involves a choice of a parameter ϵ (radius of the balls)

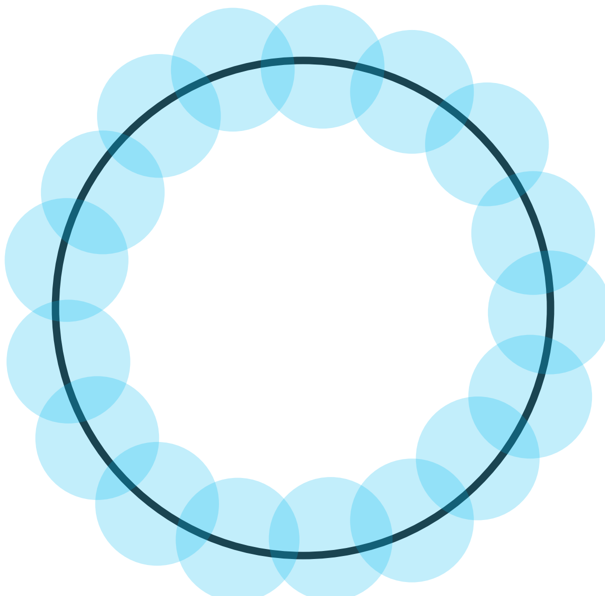
Persistent Homology - Čech Complex

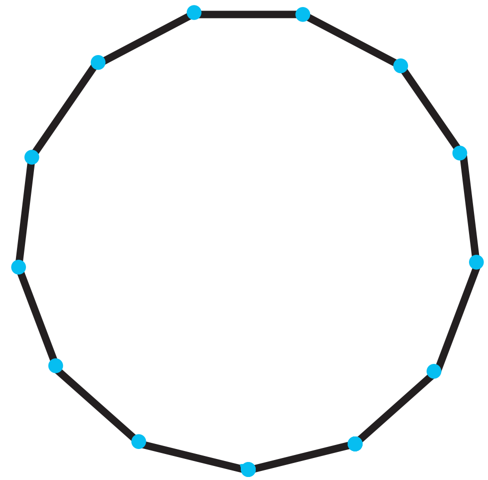


$\check{C}(X, \epsilon)$ - involves a choice of a parameter ϵ (radius of the balls)

Points are connected if balls of radius ϵ around them overlap

Persistent Homology - Čech Complex



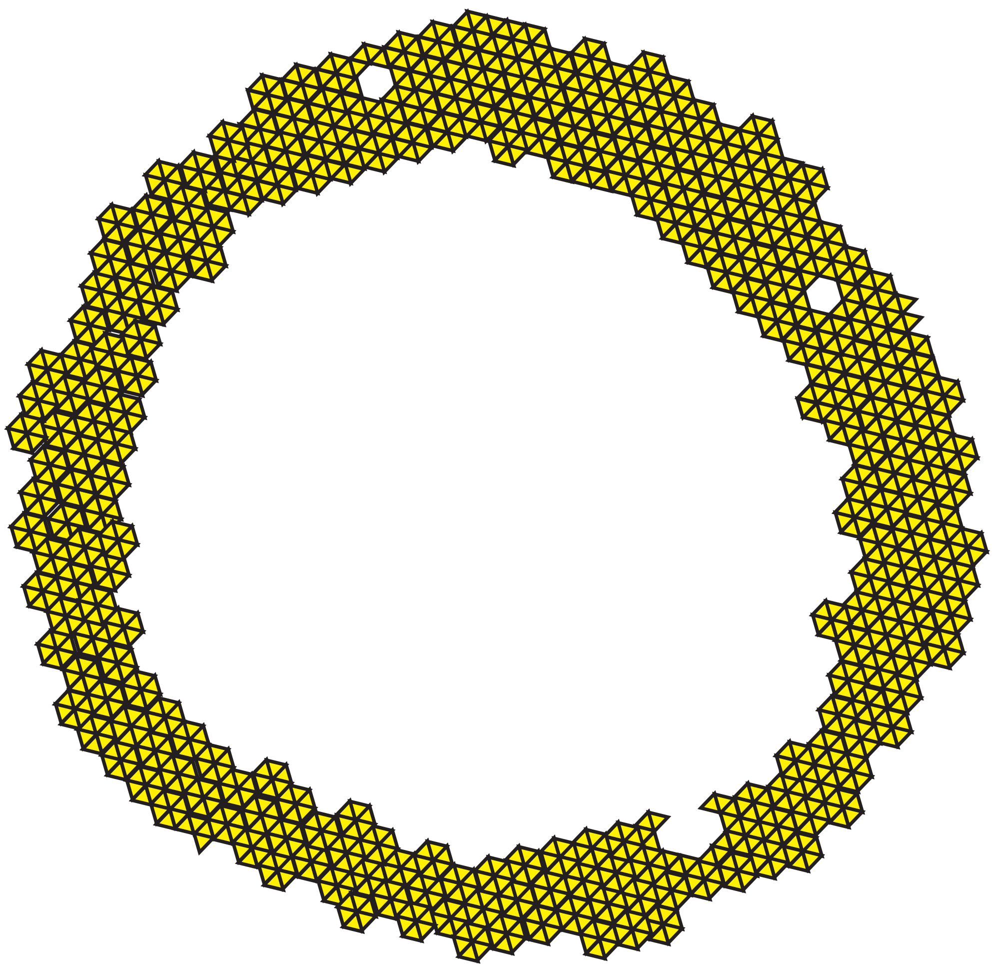


$\check{C}(X, \epsilon)$ - involves a choice of a parameter ϵ (radius of the balls)

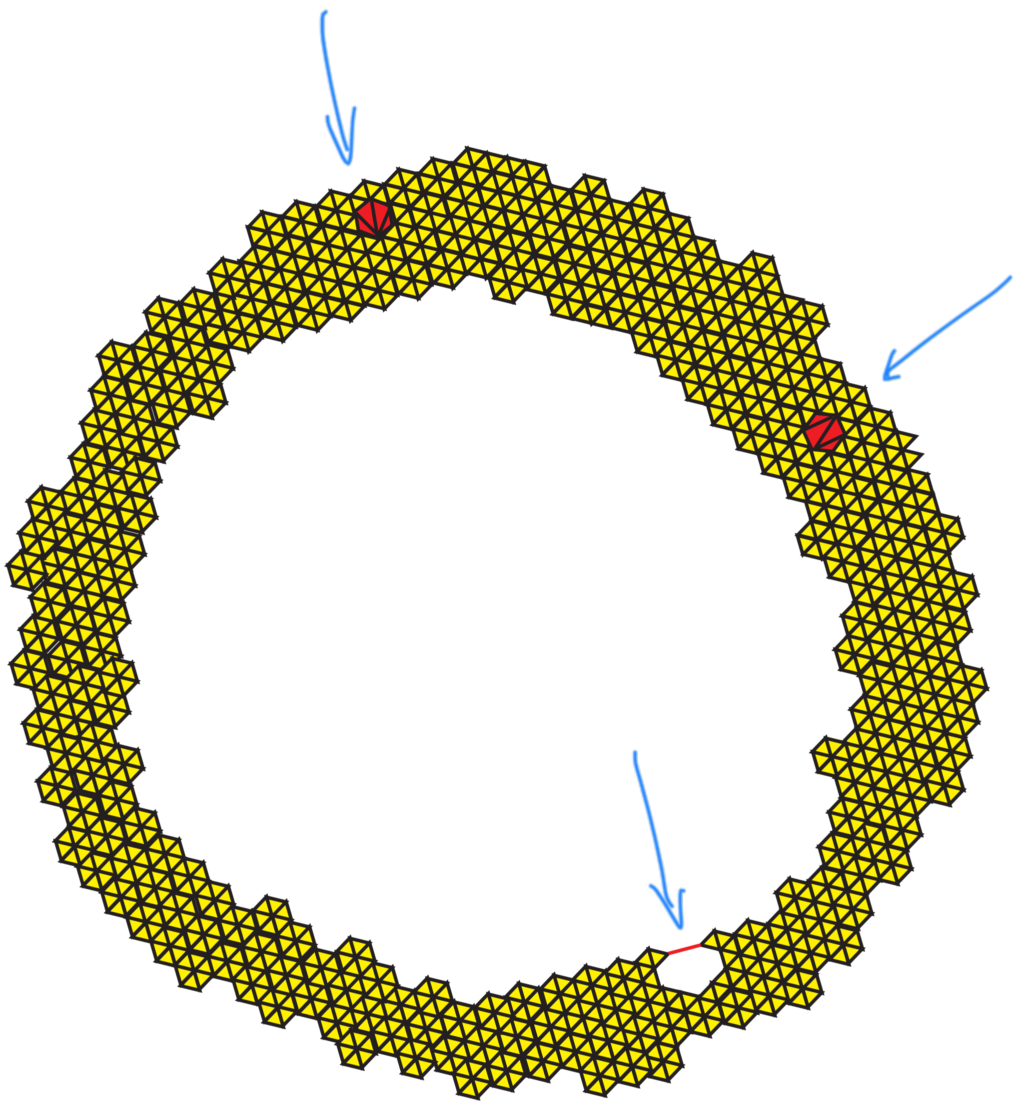
Points are connected if balls of radius ϵ around them overlap

Complex grows with ϵ

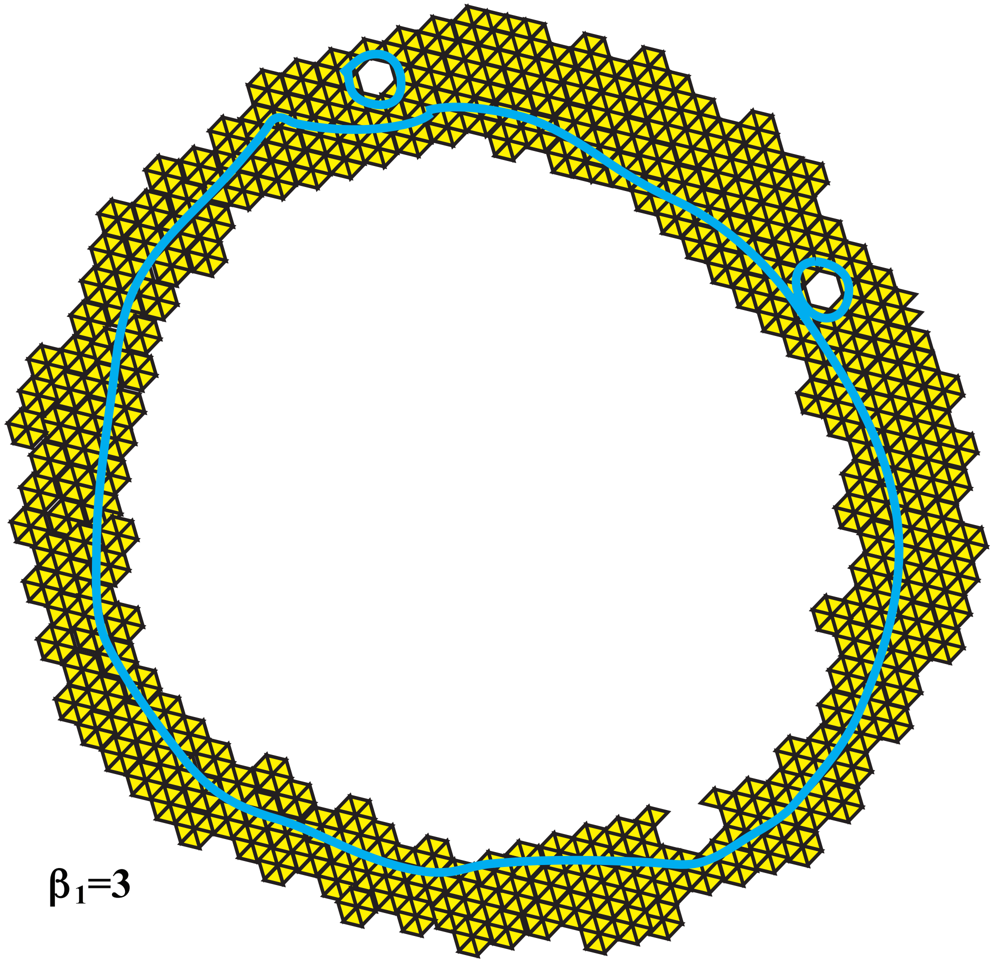
Persistent Homology



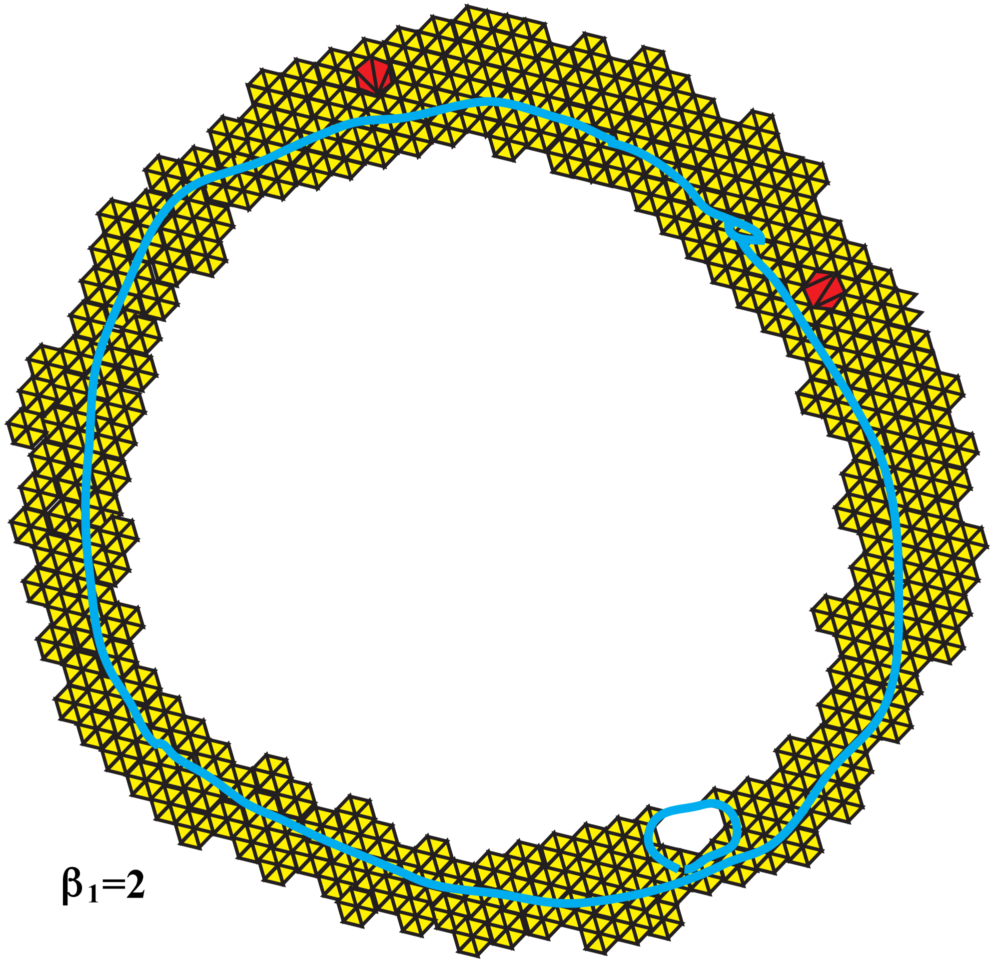
Persistent Homology



Persistent Homology



Persistent Homology



Persistent Homology

- Obtain a diagram of vector spaces

$$\cdots \to H_i(\check{C}(X, \epsilon_1)) \to H_i(\check{C}(X, \epsilon_2)) \to H_i(\check{C}(X, \epsilon_3)) \to \cdots$$

when $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3$ etc.

Persistent Homology

- Obtain a diagram of vector spaces

$$\cdots \to H_i(\check{C}(X, \epsilon_1)) \to H_i(\check{C}(X, \epsilon_2)) \to H_i(\check{C}(X, \epsilon_3)) \to \cdots$$

  when $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3$ etc.
- Called persistence vector spaces

Persistent Homology

- Obtain a diagram of vector spaces

$$\cdots \to H_i(\check{C}(X, \epsilon_1)) \to H_i(\check{C}(X, \epsilon_2)) \to H_i(\check{C}(X, \epsilon_3)) \to \cdots$$

  when $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3$ etc.
- Called persistence vector spaces
- Such diagrams can be classified by *bar codes*

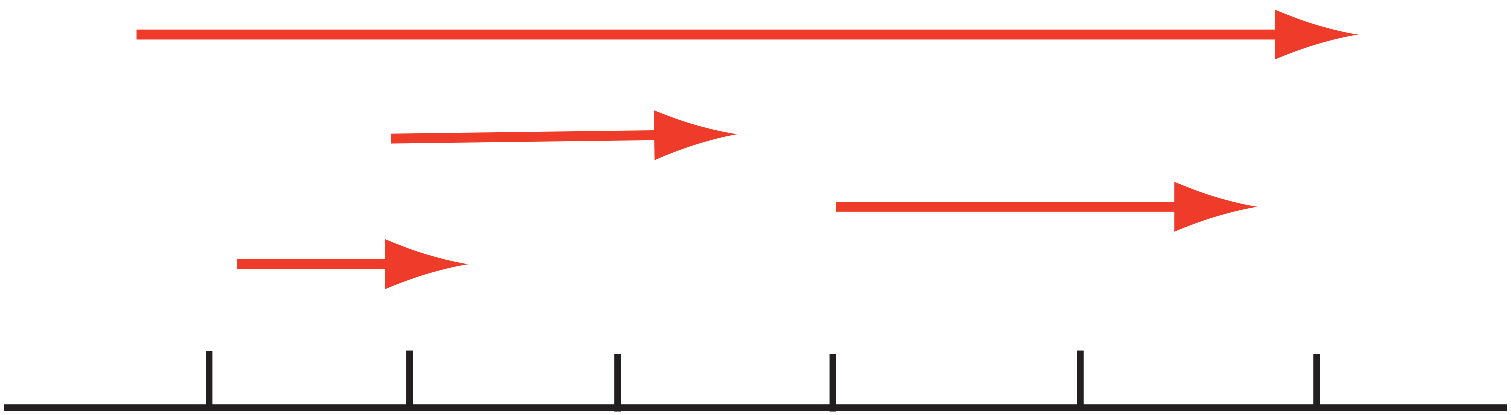
Persistent Homology

- Obtain a diagram of vector spaces

$$\cdots \to H_i(\check{C}(X,\epsilon_1)) \to H_i(\check{C}(X,\epsilon_2)) \to H_i(\check{C}(X,\epsilon_3)) \to \cdots$$

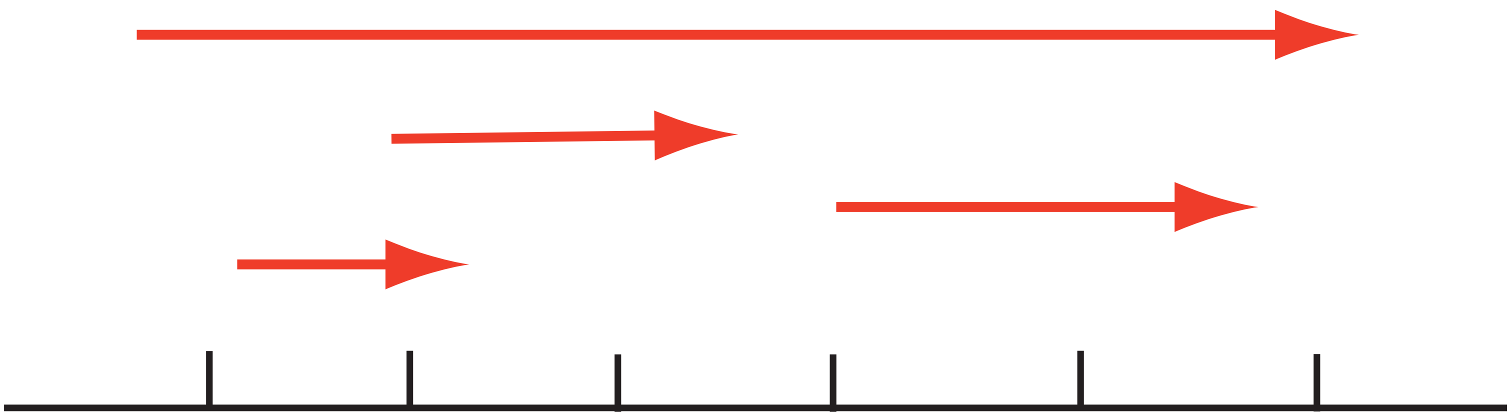
  when $\epsilon_1 \leq \epsilon_2 \leq \epsilon_3$ etc.
- Called persistence vector spaces
- Such diagrams can be classified by *bar codes*
- Analogue of dimension for ordinary vector spaces

Persistent Homology - Bar Codes



A segment indicates a basis element "born" at the left hand endpoint and which dies at the right hand endpoint

Persistent Homology - Bar Codes



A segment indicates a basis element “born” at the left hand endpoint and which dies at the right hand endpoint

Geometrically, means a loop which begins to exist (i.e. becomes closed) at the left hand point and is filled in at the right hand endpoint.

Persistent Homology - Bar Codes

Interpretation:

Persistent Homology - Bar Codes

Interpretation:

Long segments correspond to “honest” geometric features in the point cloud

Persistent Homology - Bar Codes

Interpretation:

Long segments correspond to “honest” geometric features in the point cloud

Short segments correspond to “noise”

Persistent Homology - Bar Codes

Interpretation:

Long segments correspond to “honest” geometric features in the point cloud

Short segments correspond to “noise”

Look at an example.

Example: Natural Image Statistics

- Joint with V. de Silva, T. Ishkanov, A. Zomorodian

Example: Natural Image Statistics

- Joint with V. de Silva, T. Ishkanov, A. Zomorodian
- An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel

Example: Natural Image Statistics

- Joint with V. de Silva, T. Ishkanov, A. Zomorodian
- An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel 像素
- Each pixel has a "gray scale" value, can be thought of as a real number (in reality, takes one of 255 values)

Example: Natural Image Statistics

- Joint with V. de Silva, T. Ishkanov, A. Zomorodian
- An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel
- Each pixel has a "gray scale" value, can be thought of as a real number (in reality, takes one of 255 values)
- Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional space, call it *pixel space*, $\mathcal{P}$

Example: Natural Image Statistics

D. Mumford: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?

Example: Natural Image Statistics

D. Mumford: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?

(Lee, Mumford, Pedersen): Useful to study *local* structure of images statistically

Example: Natural Image Statistics

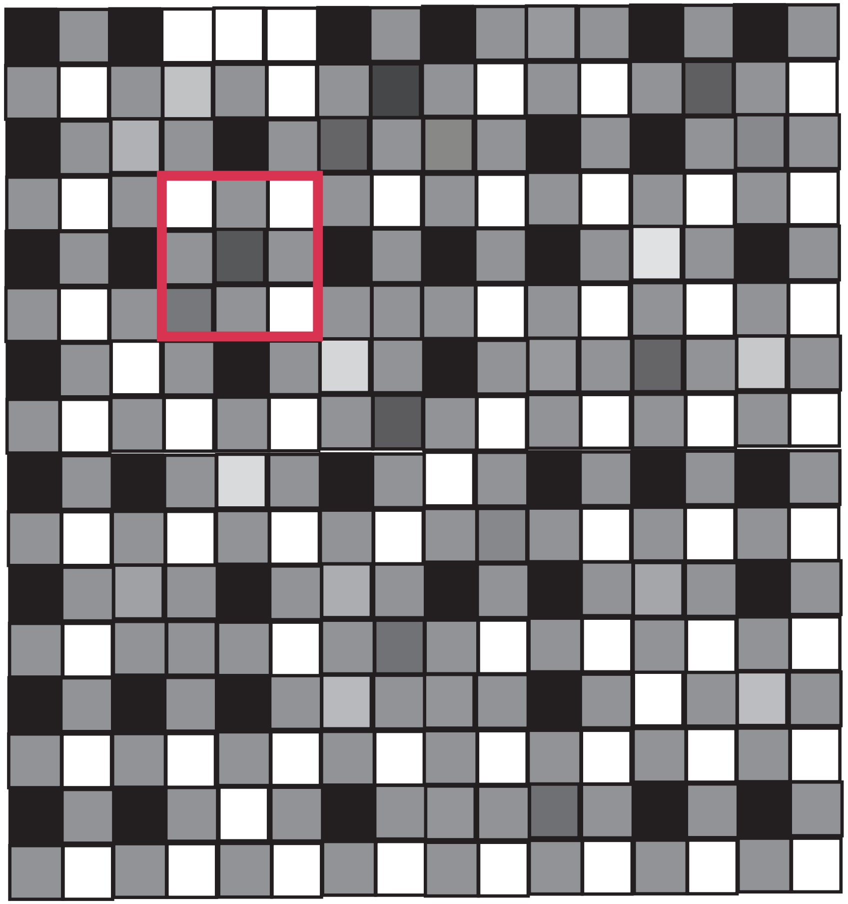
D. Mumford: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?

(Lee, Mumford, Pedersen): Useful to study *local* structure of images statistically

The nonlinear statistics of high-contrast patches in natural images

(International Journal of Computer Vision 2003)

Example: Natural Image Statistics



3×3 patches in images

Example: Natural Image Statistics

Observations:

Example: Natural Image Statistics

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$

Example: Natural Image Statistics

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$
2. Most patches will be nearly constant, or *low contrast*, because of the presence of regions of solid shading in most images

Example: Natural Image Statistics

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$
2. Most patches will be nearly constant, or *low contrast*, because of the presence of regions of solid shading in most images
3. Low contrast will dominate statistics, not interesting

Example: Natural Image Statistics

- Lee-Mumford-Pedersen [LMP] study only high contrast patches

Example: Natural Image Statistics

- Lee-Mumford-Pedersen [LMP] study only high contrast patches
- Collect c:a 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf

Example: Natural Image Statistics

- Lee-Mumford-Pedersen [LMP] study only high contrast patches
- Collect c:a 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$

Example: Natural Image Statistics

- Lee-Mumford-Pedersen [LMP] study only high contrast patches
- Collect c:a 4.5×10^6 high contrast patches from a collection of images obtained by van Hateren and van der Schaaf
- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$
- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$

Example: Natural Image Statistics

- Normalize contrast by dividing by the norm, so obtain patches with norm $= 1$

Example: Natural Image Statistics

- Normalize contrast by dividing by the norm, so obtain patches with norm $= 1$
- Means that data now lies on a 7-D ellipsoid, $\cong S^7$

Example: Natural Image Statistics

Result: Point cloud data $\mathcal{M}$ lying on a sphere in $\mathbb{R}^8$

Example: Natural Image Statistics

Result: Point cloud data $\mathcal{M}$ lying on a sphere in $\mathbb{R}^8$

We wish to analyze it with persistent homology to understand it qualitatively

Example: Natural Image Statistics

First Observation: The points fill out S^7 in the sense that every point in S^7 is “close” to a point in $\mathcal{M}$

Example: Natural Image Statistics

First Observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathcal{M}$

However, density of points varies a great deal from region to region

Example: Natural Image Statistics

First Observation: The points fill out S^7 in the sense that every point in S^7 is "close" to a point in $\mathcal{M}$

However, density of points varies a great deal from region to region

How to analyze?

Example: Natural Image Statistics

Threshholding $\mathcal{M}$

Example: Natural Image Statistics

Threshholding $\mathcal{M}$

Define $\mathcal{M}[T] \subseteq \mathcal{M}$ by

$$\mathcal{M}[T] = \{x | x \text{ is in } T\text{-th percentile of densest points}\}$$

Example: Natural Image Statistics

Threshholding $\mathcal{M}$

Define $\mathcal{M}[T] \subseteq \mathcal{M}$ by

$$\mathcal{M}[T] = \{x | x \text{ is in } T\text{-th percentile of densest points}\}$$

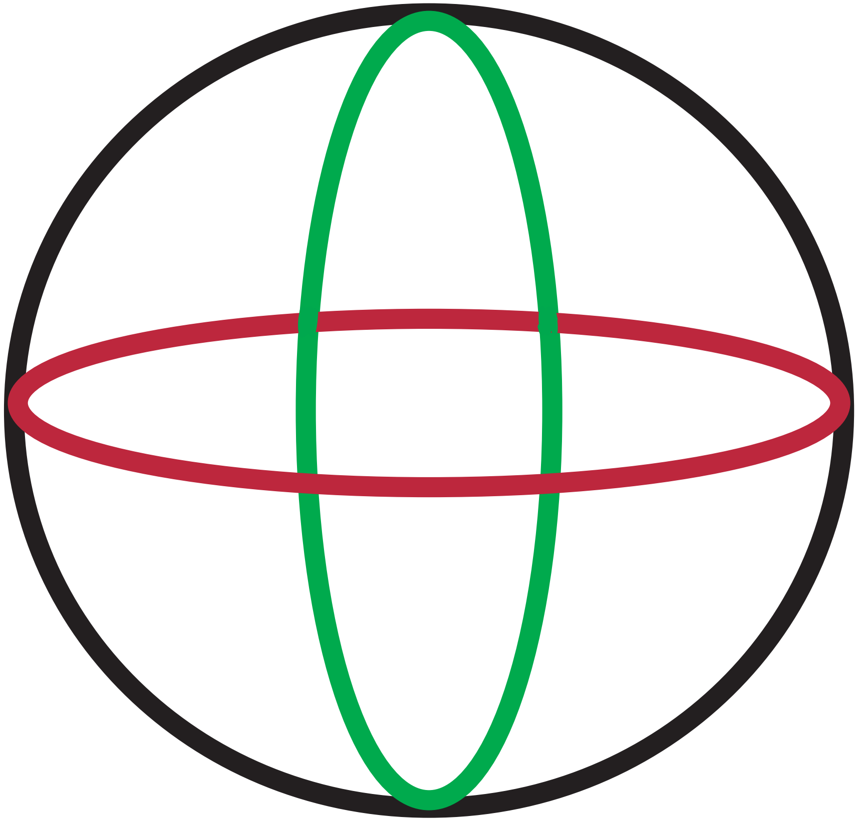
What is the persistent homology of these $\mathcal{M}[T]$'s?

Example: Natural Image Statistics

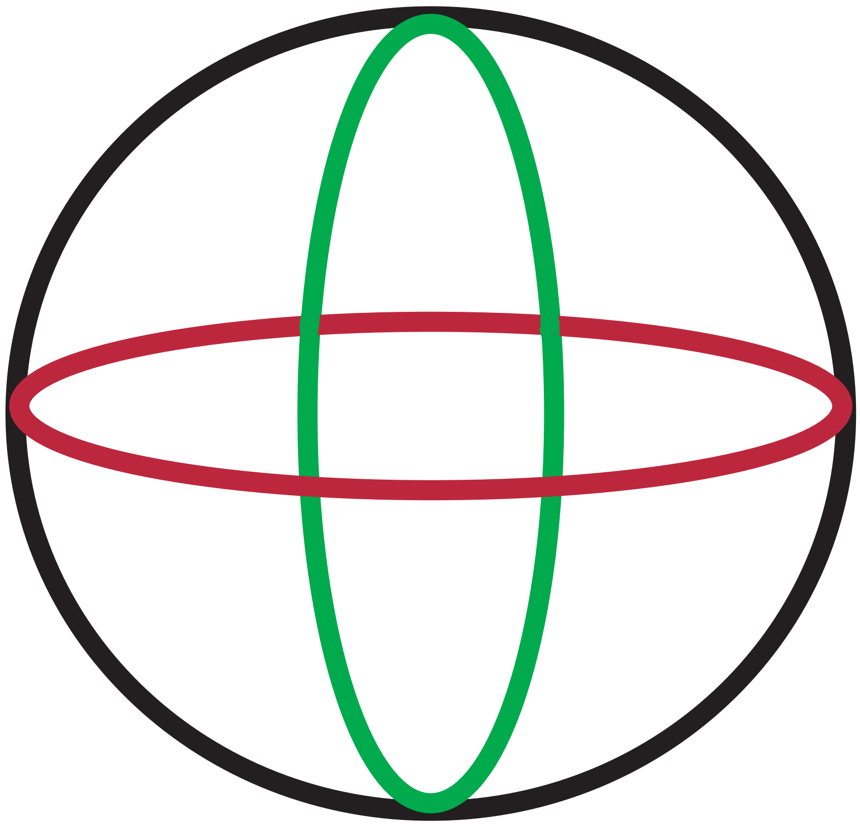
$$5 \times 10^4 \text{ points, } T = 25$$

One-dimensional barcode, suggests $\beta_1 = 5$

Example: Natural Image Statistics

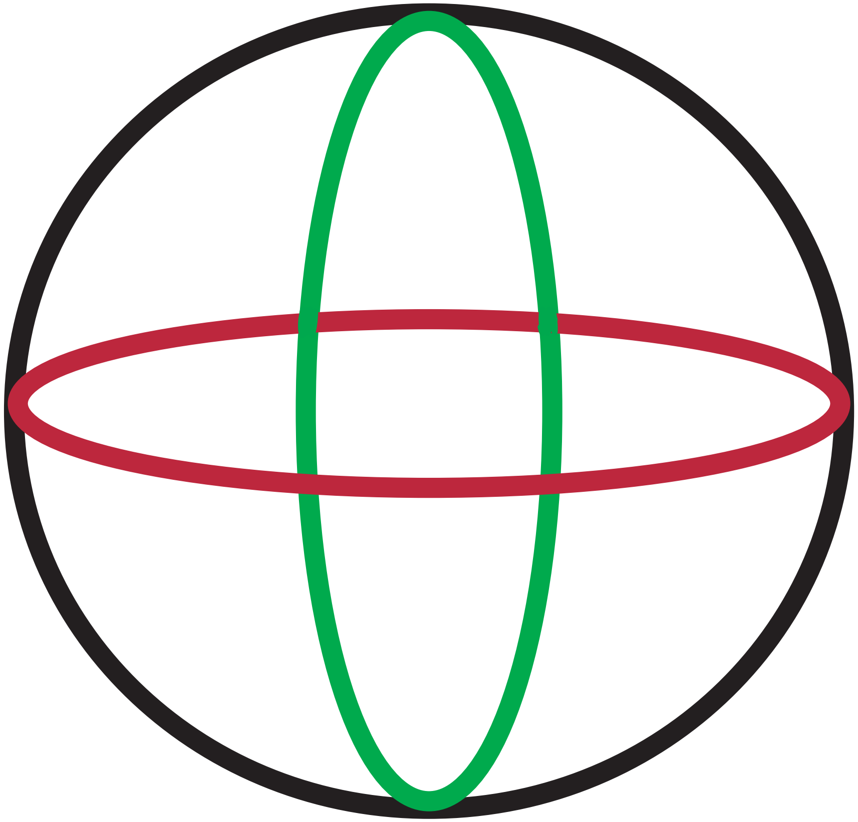


Example: Natural Image Statistics



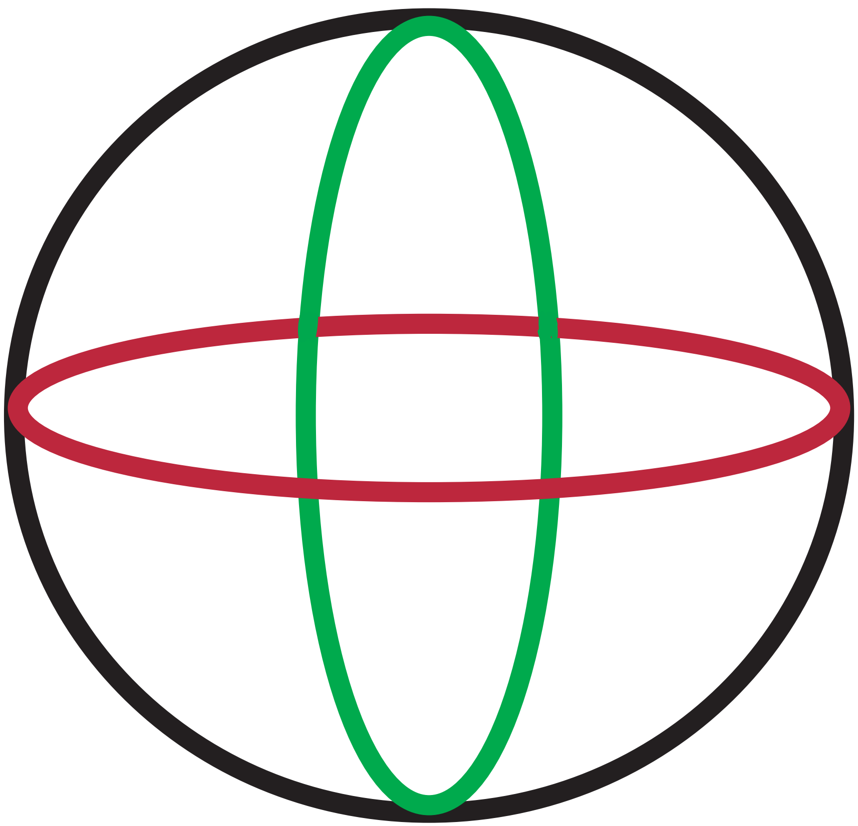
THREE CIRCLE MODEL

Three Circle Model



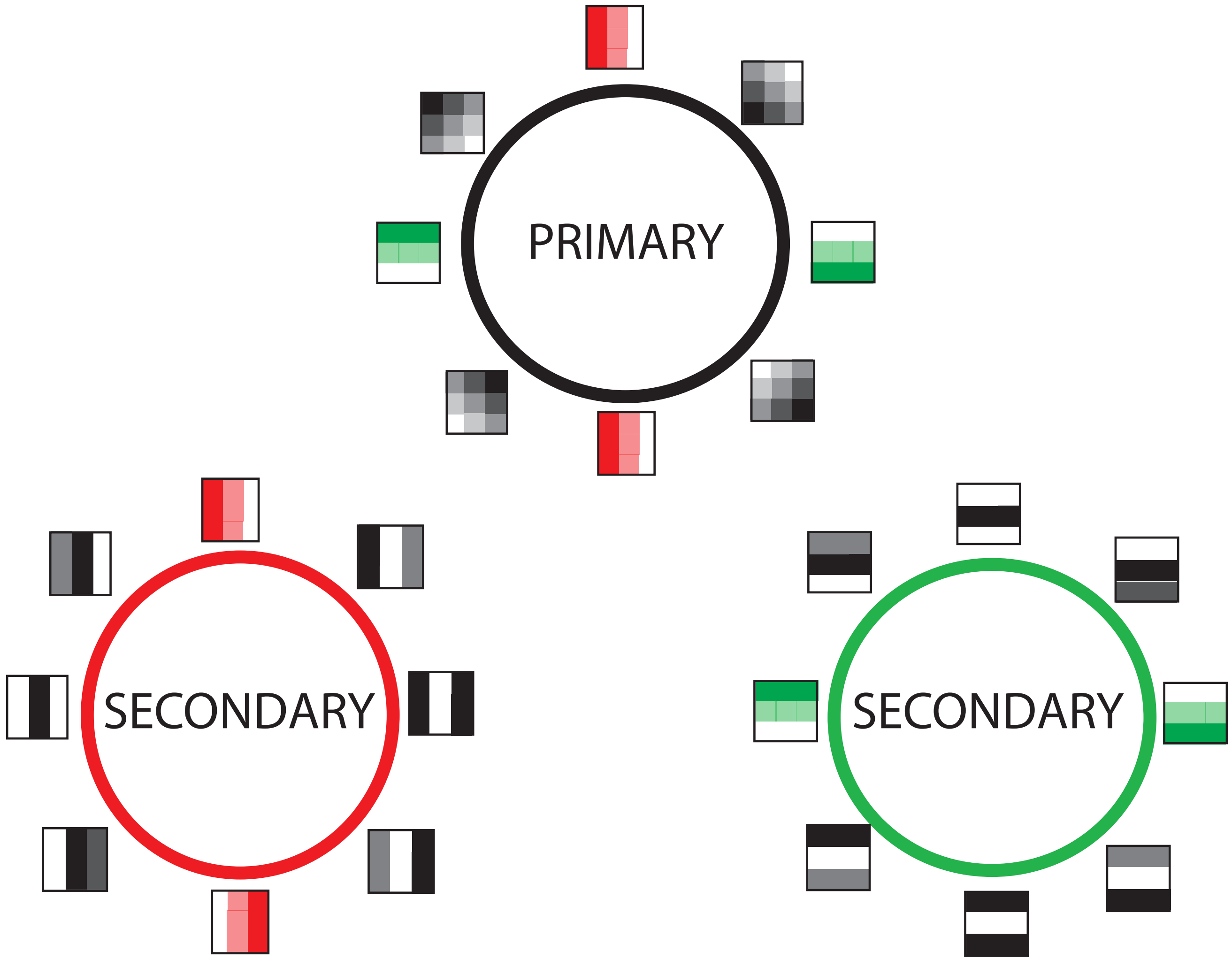
Red and green circles do not touch, each touches black circle

Example: Natural Image Statistics



Does the data fit with this model?

Example: Natural Image Statistics

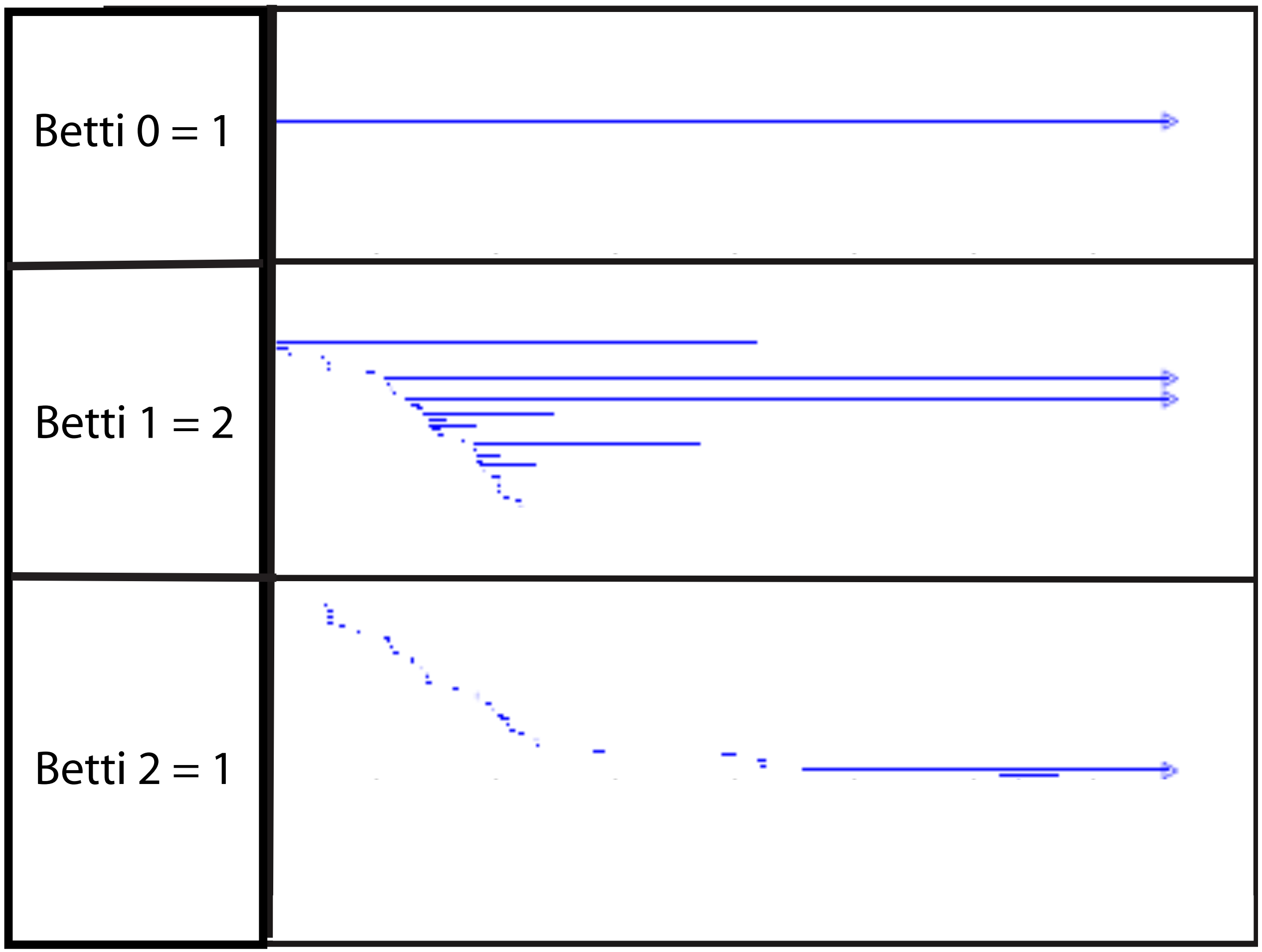


Example: Natural Image Statistics

IS THERE A TWO DIMENSIONAL SURFACE IN WHICH THIS PICTURE FITS?

Example: Natural Image Statistics

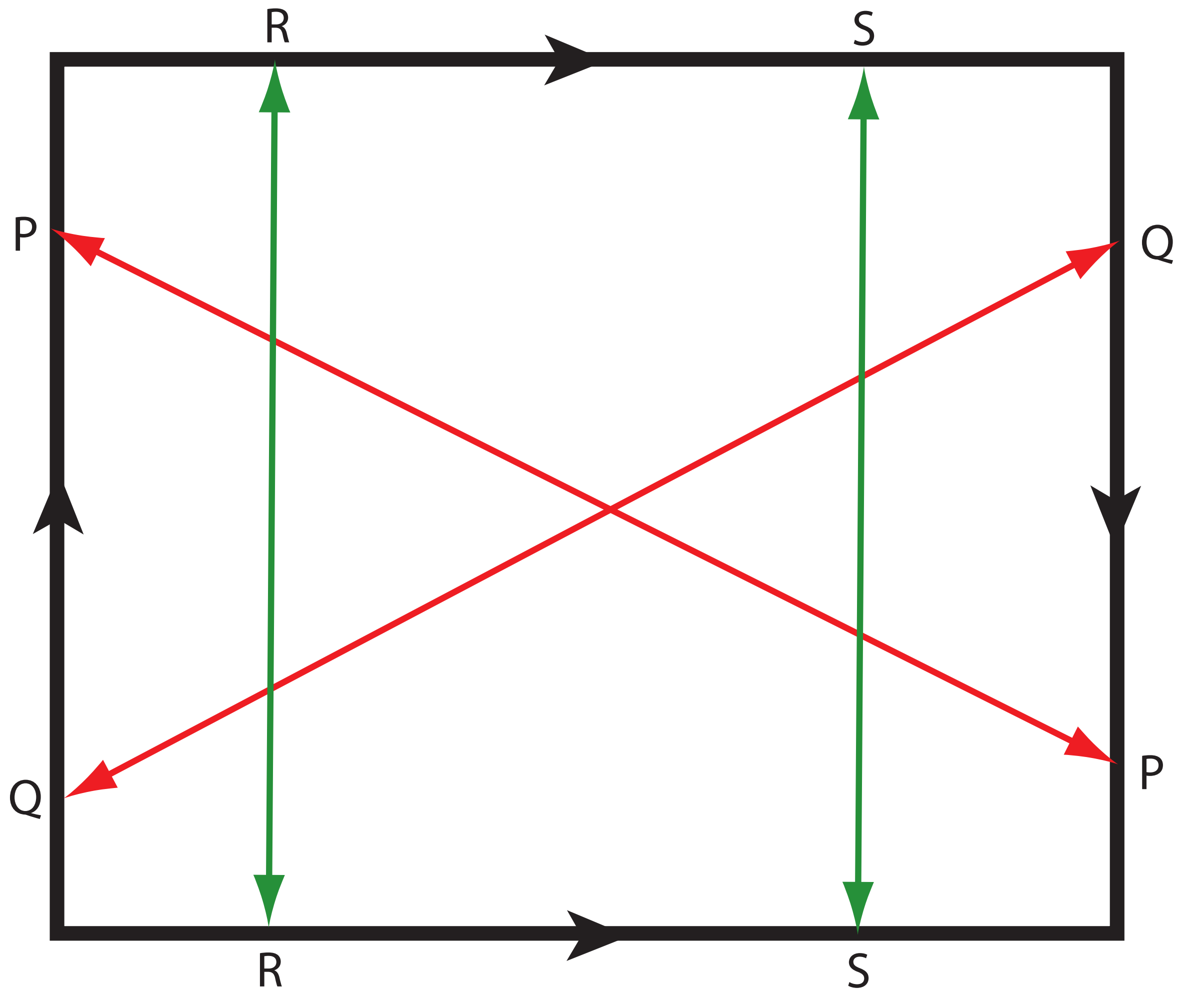
$$4.5 \times 10^6 \text{ points}, \; T = 10$$



Example: Natural Image Statistics

$\mathcal{K}$ - KLEIN BOTTLE

Example: Natural Image Statistics

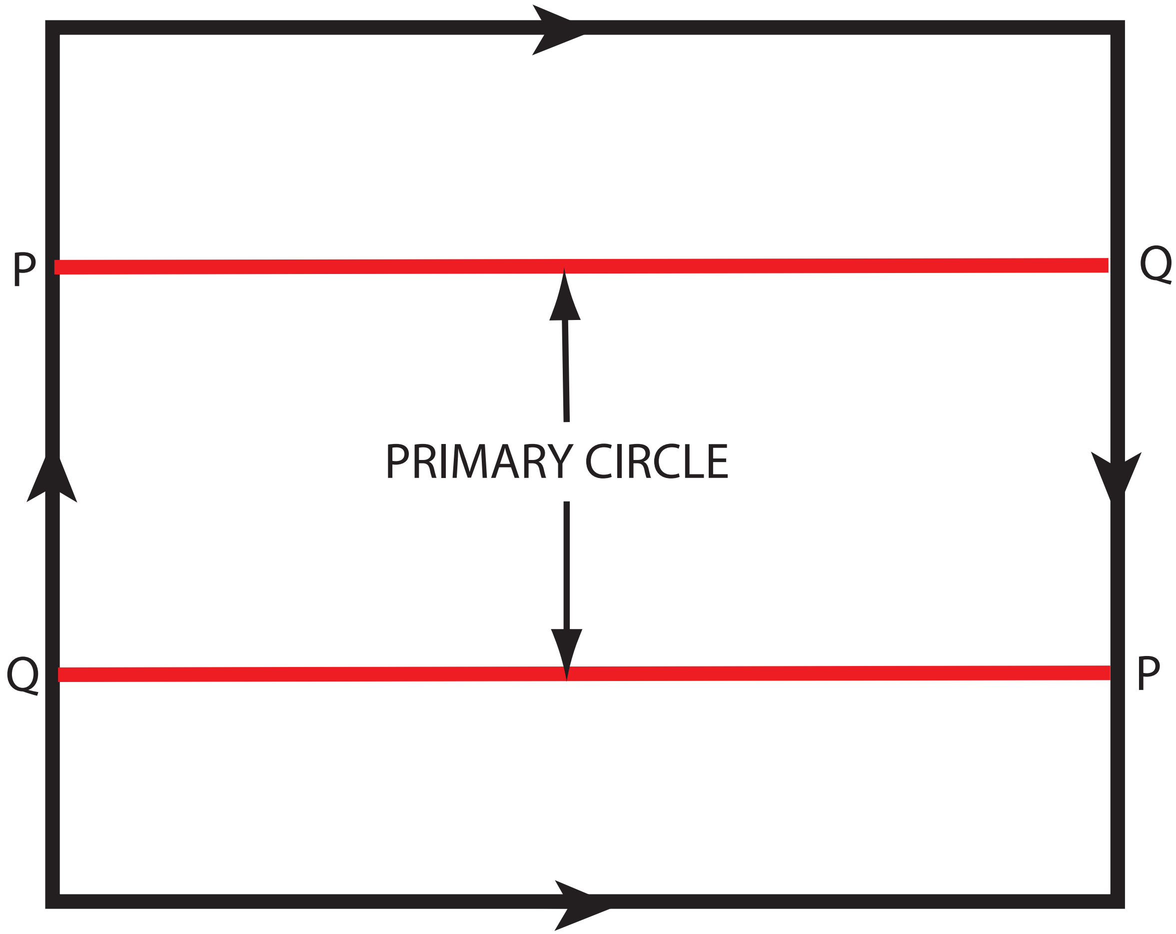


Identification Space Model

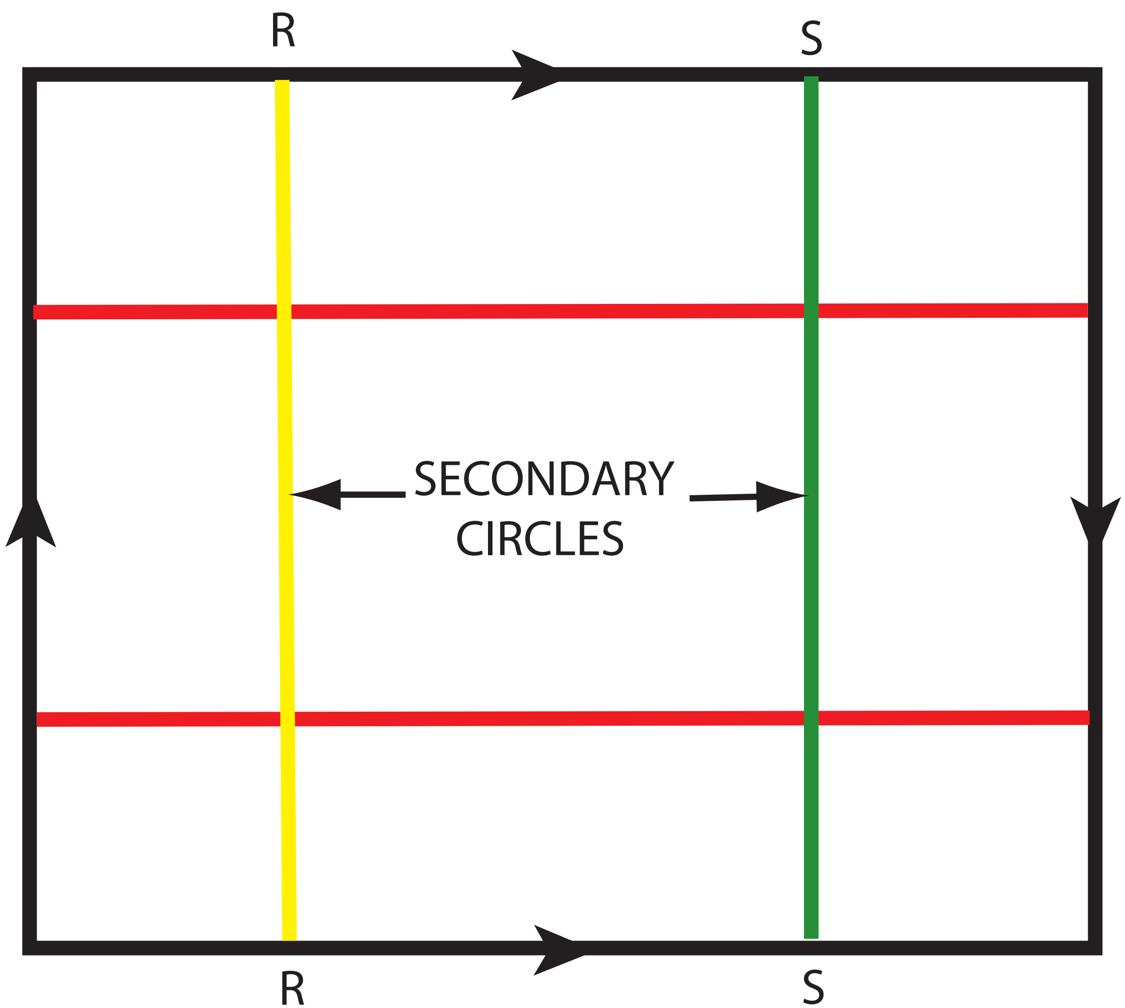
Example: Natural Image Statistics

Three circles fit naturally inside $\mathcal{K}$?

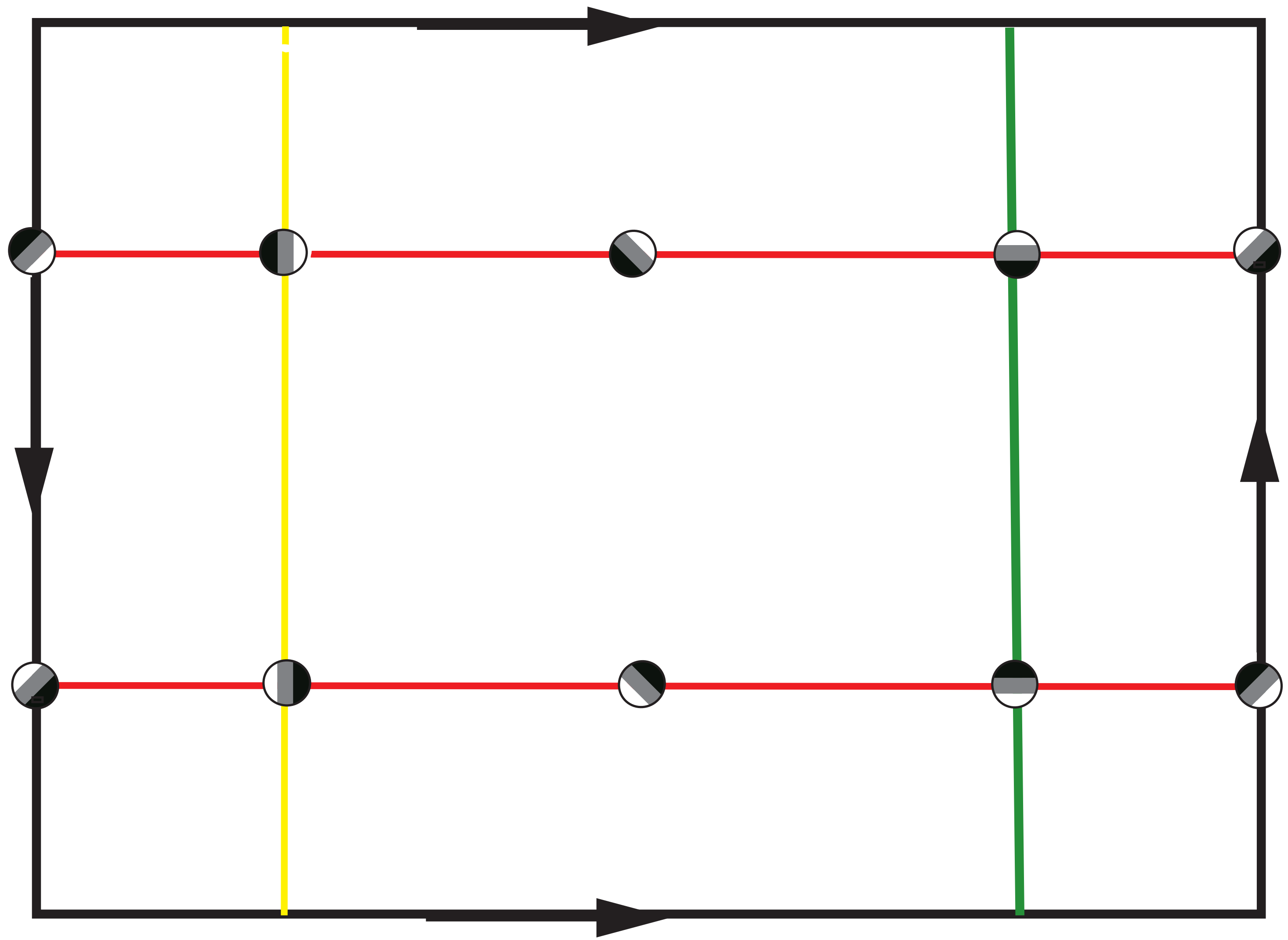
Example: Natural Image Statistics



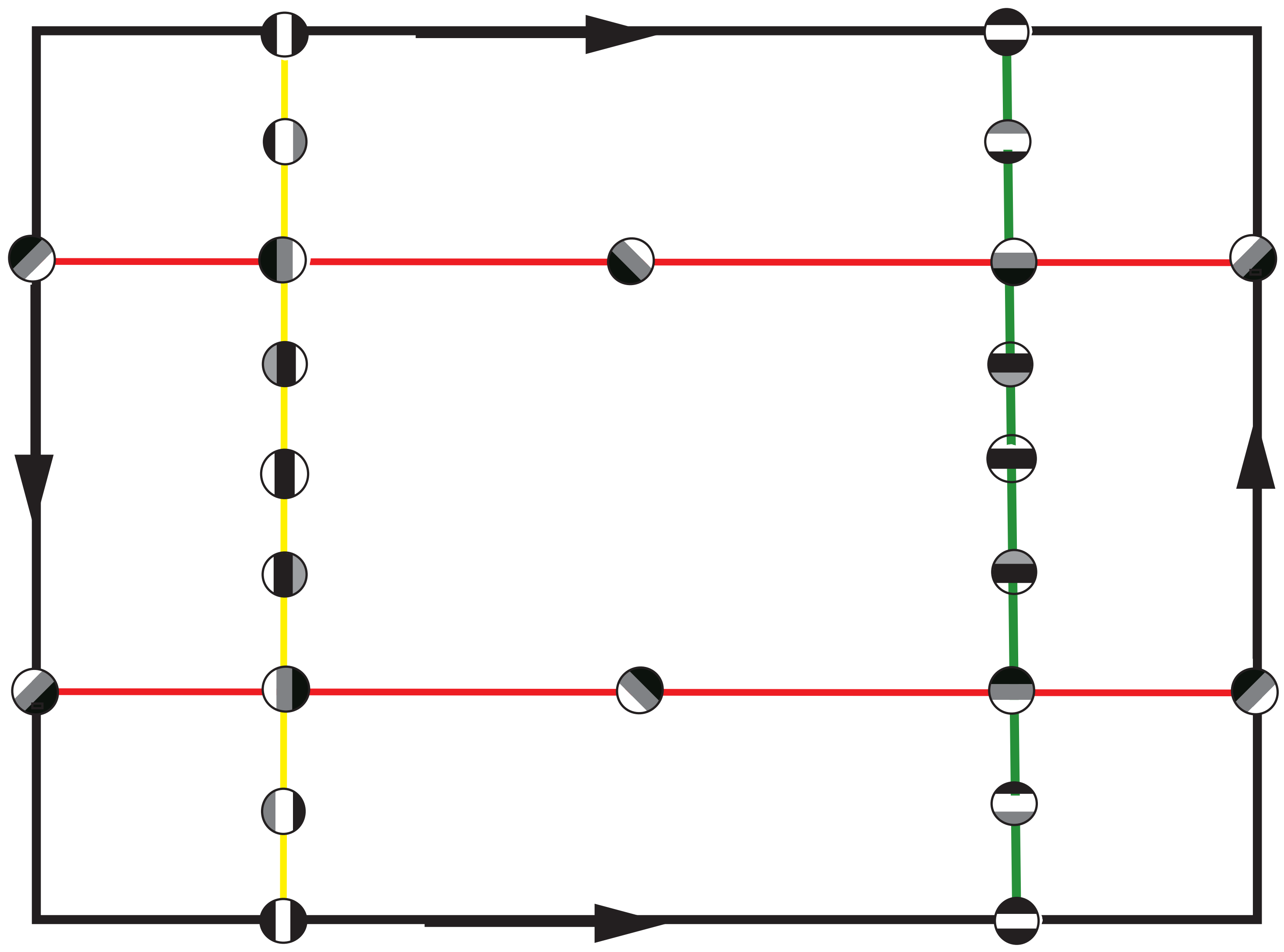
Example: Natural Image Statistics



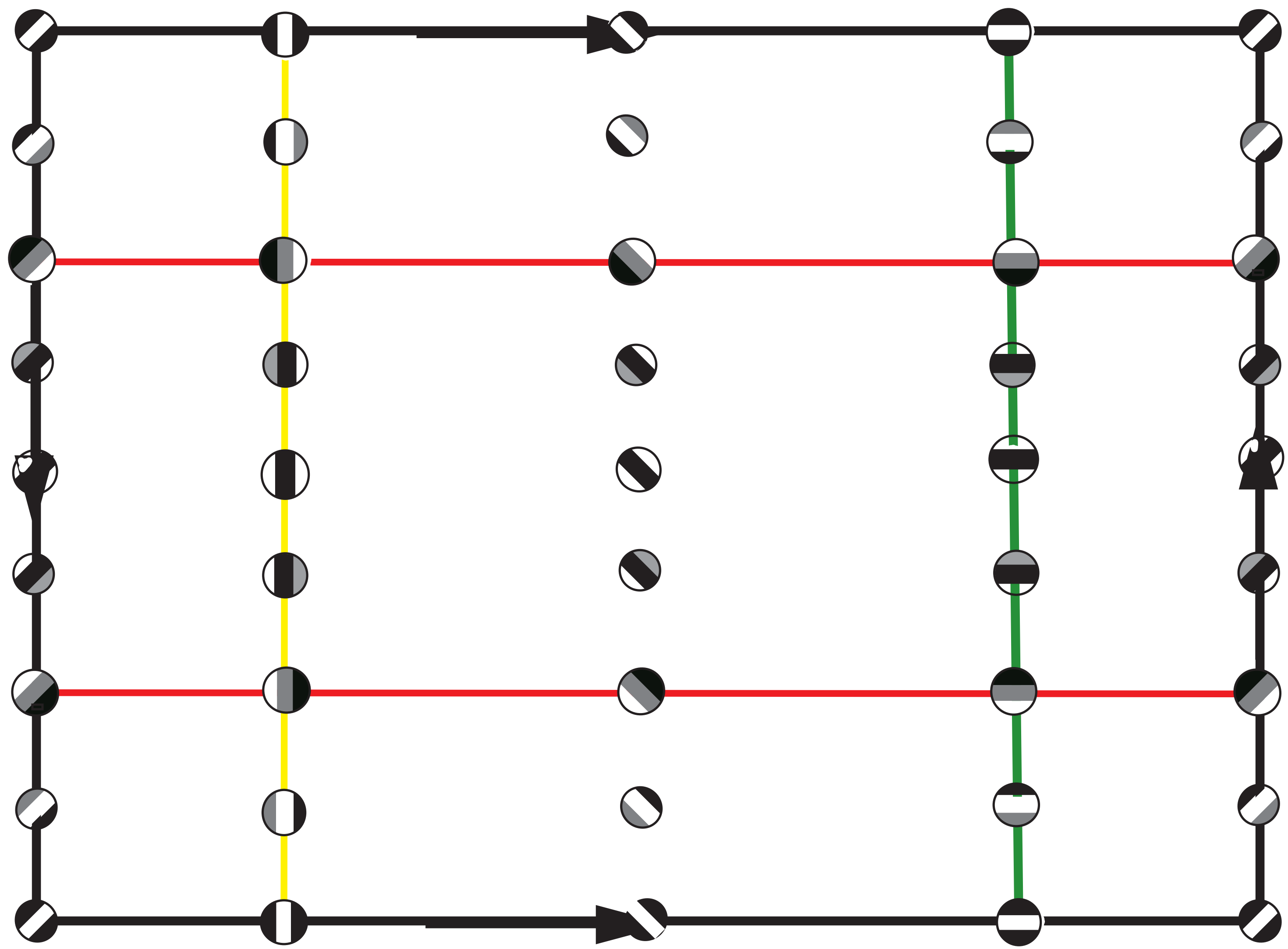
Example: Natural Image Statistics



Example: Natural Image Statistics



Example: Natural Image Statistics



Natural Image Statistics

Klein bottle makes sense in quadratic polynomials in two variables, as polynomials which can be written as

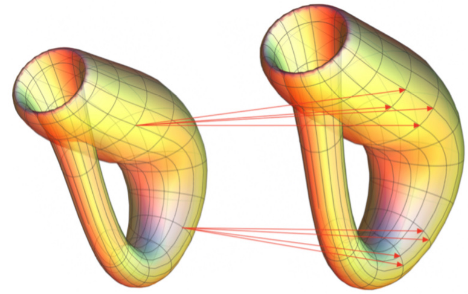
$$f = q(\lambda(x))$$

where

1. q is single variable quadratic
2. λ is a linear functional
3. $\int_D f = 0$
4. $\int_D f^2 = 1$

Carlsson, Ishkhanov, De Silva, Zomorodian
On the local behavior of spaces of natural images.
International Journal of Computer Vision, 2008

A decade later, Love, Filippenko, Maroulas, and Carlsson have made the Klein bottle as a topological input for designing convolutional layers in neural networks that learn image data. Moreover, they have incorporated the tangent bundle of a Klein bottle into TCNNs for learning video data. Both learnings achieved higher accuracies with smaller training sets.

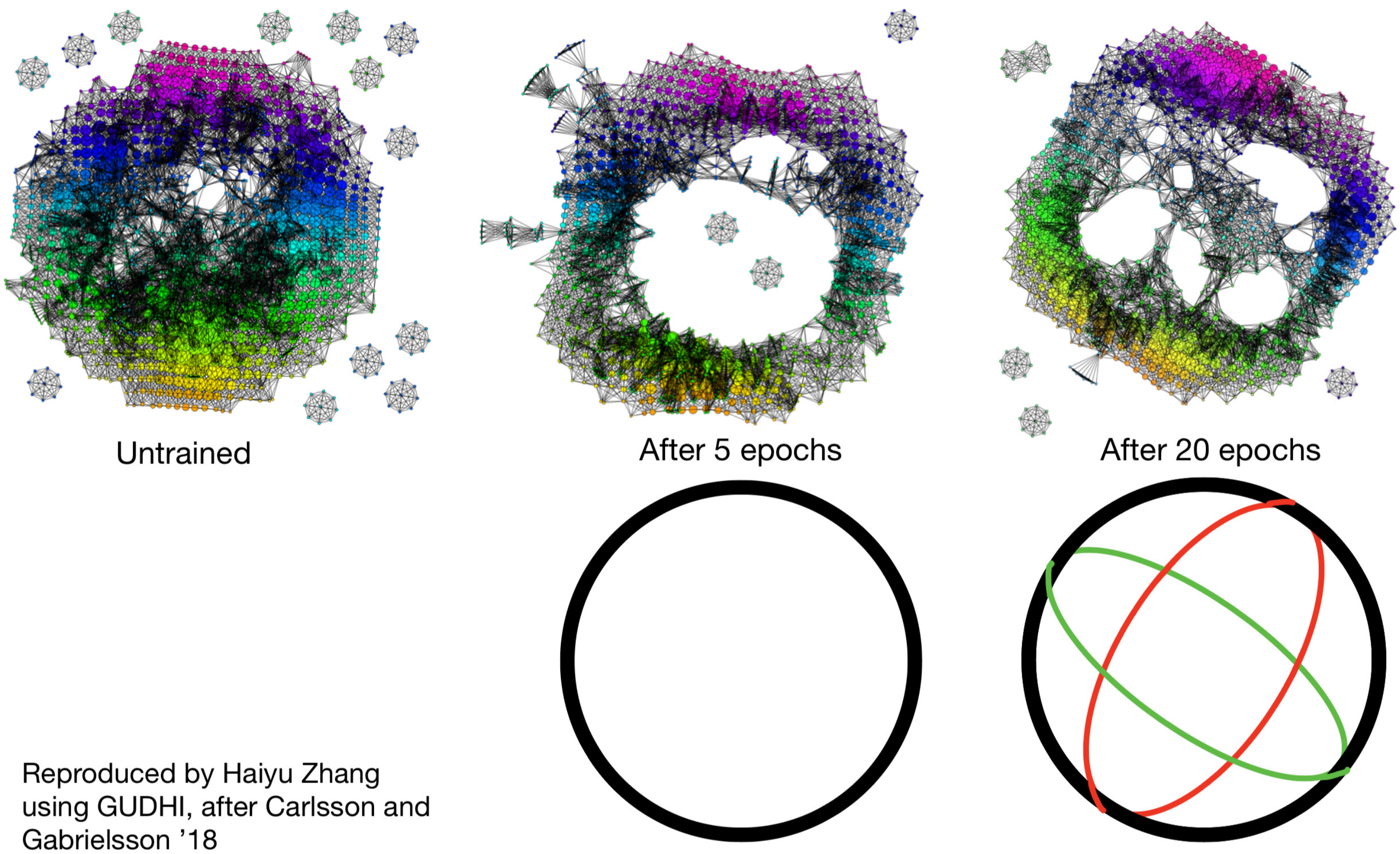


Ephy R. Love et al., Topological convolutional layers for deep learning, ***Journal of Machine Learning Research****, 2023.*

Gunnar Carlsson and Rickard Brüel Gabrielsson, Topological approaches to deep learning, ***Topological Data Analysis: The Abel Symposium****, 2018.*

From topological data analysis to topological deep learning

Topology of convolutional neural networks: Emergence of cycles during a training process



Reproduced by Haiyu Zhang using GUDHI, after Carlsson and Gabrielsson '18

Mapper

Algebraic topology can produce signatures which can help in mapping out a data set.

Mapper

Algebraic topology can produce signatures which can help in mapping out a data set.

Can one obtain flexible topological mapping methods, with combinatorial simplicial complex images?

Mapper

Algebraic topology can produce signatures which can help in mapping out a data set.

Can one obtain flexible topological mapping methods, with combinatorial simplicial complex images?

Yes, joint work with G. Singh and F. Memoli.

Mapper - Mayer-Vietoris Blowup

X a space, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a covering of X.

Mapper - Mayer-Vietoris Blowup

X a space, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a covering of X.

Δ is the simplex with vertex set A

Mapper - Mayer-Vietoris Blowup

X a space, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a covering of X.

Δ is the simplex with vertex set A

$\emptyset \neq S \subseteq A, X(S) = \bigcap_{s \in S} U_s$ and $\Delta[S] =$ face spanned by S

Mapper - Mayer-Vietoris Blowup

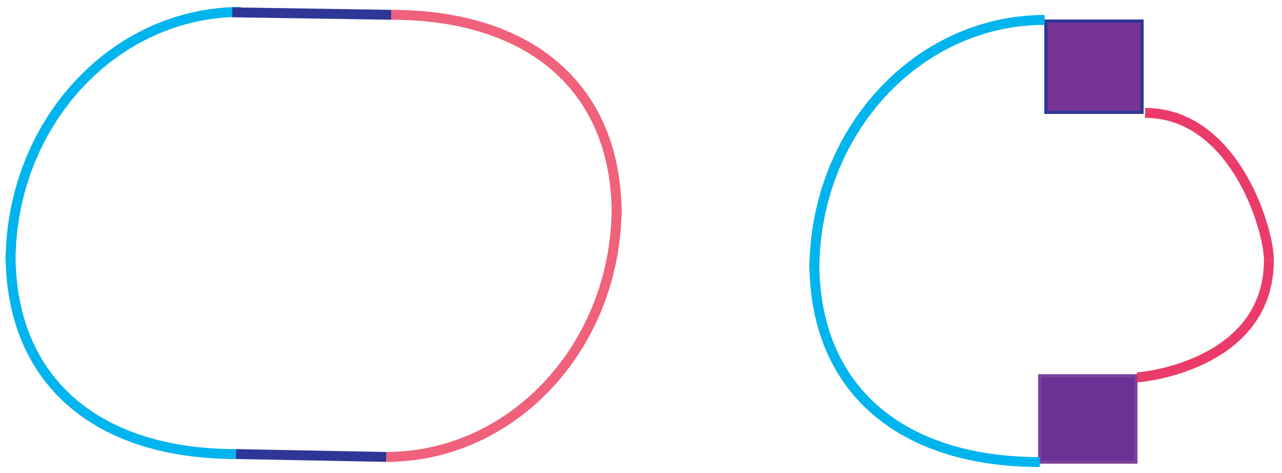
X a space, $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a covering of X.

Δ is the simplex with vertex set A

$\emptyset \neq S \subseteq A, X(S) = \bigcap_{s \in S} U_s$ and $\Delta[S] =$ face spanned by S

Let $X^{\mathcal{U}} \subseteq X \times \Delta$, $X^{\mathcal{U}} = \bigcup_S X(S) \times \Delta[S]$

Mapper - Mayer-Vietoris Blowup



Mapper - Mayer-Vietoris Blowup

Exists a map $\pi_X : X^{\mathcal{U}} \rightarrow X$, which is a homotopy equivalence with mild hypotheses

Mapper - Mayer-Vietoris Blowup

Exists a map $\pi_X : X^{\mathcal{U}} \to X$, which is a homotopy equivalence with mild hypotheses

$N(\mathcal{U}) = \bigcup_{\{S | X(S) \neq \emptyset\}} \Delta[S]$

Mapper - Mayer-Vietoris Blowup

Exists a map $\pi_X : X^{\mathcal{U}} \to X$, which is a homotopy equivalence with mild hypotheses

$N(\mathcal{U}) = \bigcup_{\{S | X(S) \neq \emptyset\}} \Delta[S]$

Exists a second map $\pi_\Delta : X^{\mathcal{U}} \to N(\mathcal{U})$

Mapper - Mayer-Vietoris Blowup

Exists a map $\pi_X : X^{\mathcal{U}} \to X$, which is a homotopy equivalence with mild hypotheses

$N(\mathcal{U}) = \bigcup_{\{S | X(S) \neq \emptyset\}} \Delta[S]$

Exists a second map $\pi_\Delta : X^{\mathcal{U}} \to N(\mathcal{U})$

π_Δ is equivalence if all $X(S)$'s are empty or contractible

Mapper - Mayer-Vietoris Blowup

Intermediate construction $\mathcal{M}(X, \mathcal{U})$

Mapper - Mayer-Vietoris Blowup

Intermediate construction $\mathcal{M}(X, \mathcal{U})$

$$\mathcal{M}(X, \mathcal{U}) = \coprod_{S} \pi_0(X(S)) \times \Delta[S] / \simeq$$

Mapper - Mayer-Vietoris Blowup

Intermediate construction $\mathcal{M}(X,\mathcal{U})$

$$\mathcal{M}(X,\mathcal{U}) = \coprod_{S} \pi_0(X(S)) \times \Delta[S] / \simeq$$

$$\pi_0(X(S)) \times \Delta[S] \xleftarrow{\phi} \pi_0(X(T)) \times \Delta[S] \xrightarrow{\psi} \pi_0(X(T)) \times \Delta[T]$$

Mapper - Mayer-Vietoris Blowup

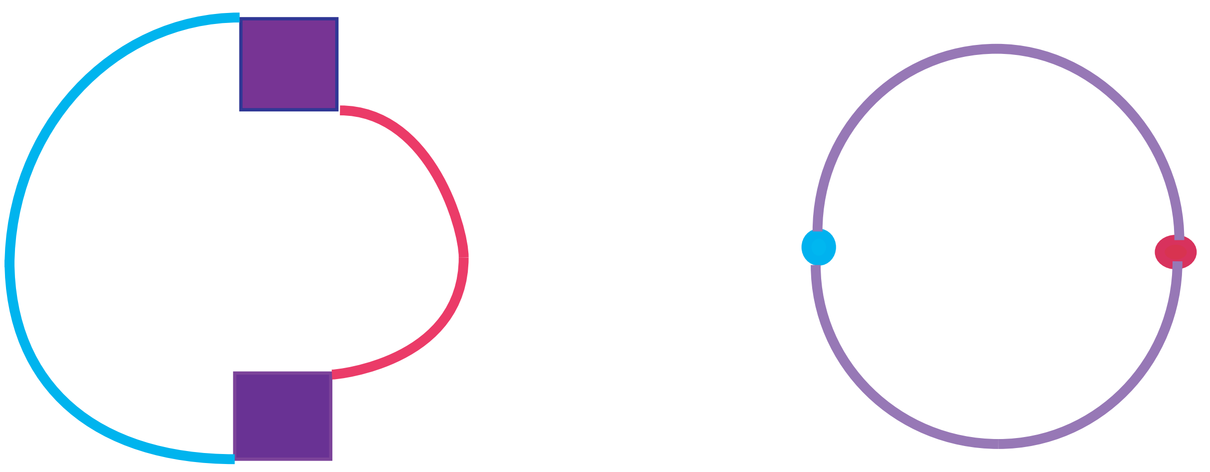
Intermediate construction $\mathcal{M}(X,\mathcal{U})$

$$\mathcal{M}(X,\mathcal{U}) = \coprod_S \pi_0(X(S)) \times \Delta[S]/\simeq$$

$$\pi_0(X(S)) \times \Delta[S] \xleftarrow{\phi} \pi_0(X(T)) \times \Delta[S] \xrightarrow{\psi} \pi_0(X(T)) \times \Delta[T]$$

$\phi(x,\zeta) \simeq \psi(x,\zeta)$

Mapper - Mayer-Vietoris Blowup



Mapper - Statistical Version

Now given point cloud data set $\mathbb{X}$, and a covering $\mathcal{U}$.

Mapper - Statistical Version

Now given point cloud data set $\mathbb{X}$, and a covering $\mathcal{U}$.

Build simplicial complex same way, but π_0 operation replaced by single linkage clustering with fixed error parameter ε.

Mapper - Statistical Version

Now given point cloud data set $\mathbb{X}$, and a covering $\mathcal{U}$.

Build simplicial complex same way, but π_0 operation replaced by single linkage clustering with fixed error parameter ε.

Critical that clustering operation be functorial.

Mapper - Statistical Version

Now given point cloud data set $\mathbb{X}$, and a covering $\mathcal{U}$.

Build simplicial complex same way, but π_0 operation replaced by single linkage clustering with fixed error parameter ε.

Critical that clustering operation be functorial.

Partition of unity subordinate to $\mathcal{U}$ gives map from $\mathbb{X}$ to $\mathcal{M}(\mathbb{X},\mathcal{U})$.

Mapper - Statistical Version

How to choose coverings?

Mapper - Statistical Version

How to choose coverings?

Given a reference map (or filter) $f : \mathbb{X} \to Z$, where Z is a metric space, and a covering $\mathcal{U}$ of Z, can consider the covering $\{f^{-1}U_\alpha\}_{\alpha \in A}$ of $\mathbb{X}$. Typical choices of Z - $\mathbb{R}$, $\mathbb{R}^2$, S^1.

Mapper - Statistical Version

How to choose coverings?

Given a reference map (or filter) $f : \mathbb{X} \to Z$, where Z is a metric space, and a covering $\mathcal{U}$ of Z, can consider the covering $\{f^{-1}U_\alpha\}_{\alpha \in A}$ of $\mathbb{X}$. Typical choices of Z - $\mathbb{R}$, $\mathbb{R}^2$, S^1.

Construction gives an image complex of the data set which can reflect interesting properties of $\mathbb{X}$.

Mapper - Statistical Version

Typical one dimensional filters:

- Density estimators

Mapper - Statistical Version

Typical one dimensional filters:

- Density estimators
- "Eccentricity" : $\sum_{x' \in \mathbb{X}} d(x, x')^2$

Mapper - Statistical Version

Typical one dimensional filters:

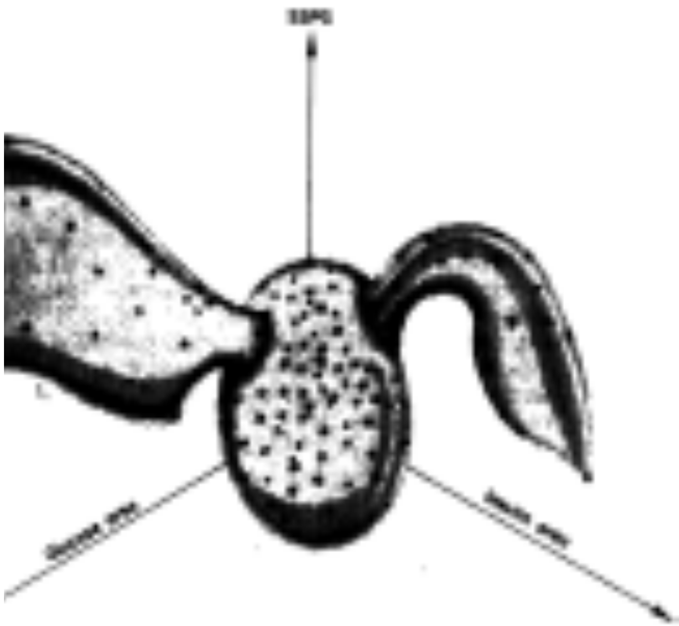
- Density estimators
- "Eccentricity" : $\sum_{x' \in \mathbb{X}} d(x, x')^2$
- Eigenfunctions of graph Laplacian for Vietoris-Rips graph

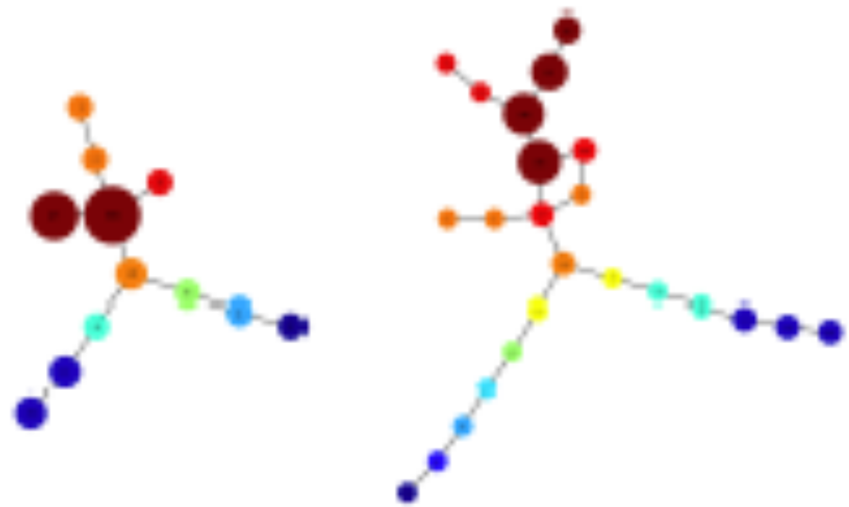
Mapper - Statistical Version

Typical one dimensional filters:

- Density estimators
- "Eccentricity" : $\sum_{x' \in \mathbb{X}} d(x, x')^2$
- Eigenfunctions of graph Laplacian for Vietoris-Rips graph
- User defined, data dependent filter functions

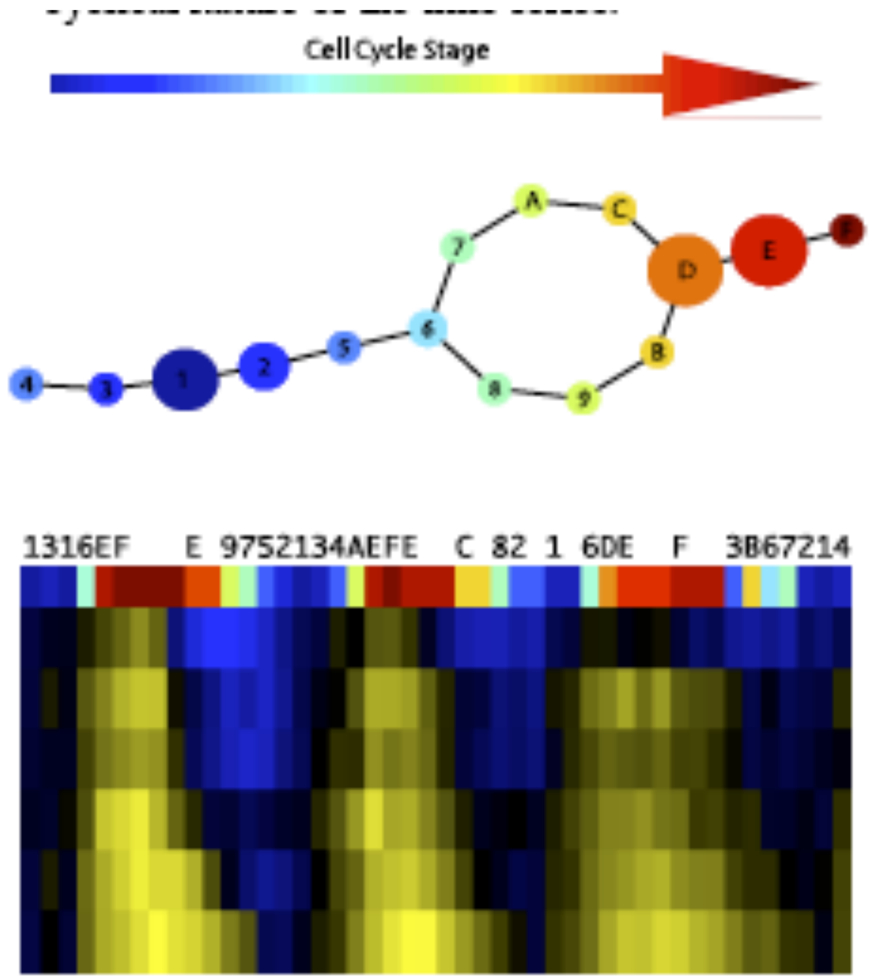
Mapper - Statistical Version





Miller-Reaven Diabetes Study, 1976

Mapper - Statistical Version



Cell Cycle Microarray Data

Joint with M. Nicolau, Nagarajan, G. Singh

Mapper - Scale Space

How to choose the parameter ε in the single linkage clustering?

Mapper - Scale Space

How to choose the parameter ε in the single linkage clustering?

Can one allow ε to vary with α?

Mapper - Scale Space

How to choose the parameter ε in the single linkage clustering?

Can one allow ε to vary with α?

Important question: too many parameter choices makes tool unusable, and choosing one ε for the entire space is too restrictive.

Mapper - Scale Space

Construct a new space with reference map to Z.

Mapper - Scale Space

Construct a new space with reference map to Z.

For each α, we construct the zero dimensional persistence diagram for $f^{-1}U_\alpha$.

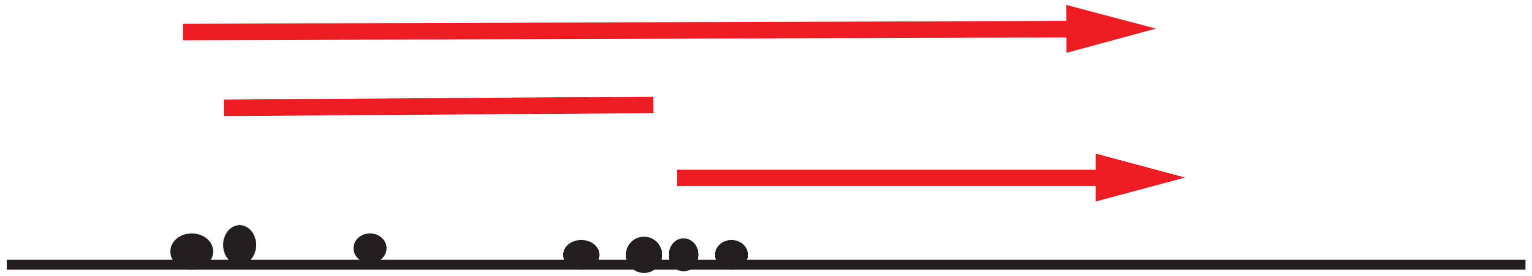
Mapper - Scale Space

Construct a new space with reference map to Z.

For each α, we construct the zero dimensional persistence diagram for $f^{-1}U_\alpha$.

Consider the set of all endpoints of intervals in the persistence diagram. Provides a decomposition of the real line in which ε is varying into intervals. Call these intervals S-intervals.

Mapper - Scale Space



Mapper - Scale Space

- Vertex set of $SS(X, \mathcal{U})$ consists of a pair (α, I), where $\alpha \in A$ and I is an S-interval for the zero dimensional persistence diagram for $f^{-1}(U_\alpha)$.

Mapper - Scale Space

- Vertex set of $SS(X, \mathcal{U})$ consists of a pair (α, I), where $\alpha \in A$ and I is an S-interval for the zero dimensional persistence diagram for $f^{-1}(U_\alpha)$.

- We connect (α, I) and (β, J) with an edge if (a) $U_\alpha \cap U_\beta \neq \emptyset$ and (b) $I \cap J \neq \emptyset$.

Mapper - Scale Space

- Vertex set of $SS(X, \mathcal{U})$ consists of a pair (α, I), where $\alpha \in A$ and I is an S-interval for the zero dimensional persistence diagram for $f^{-1}(U_\alpha)$.
- We connect (α, I) and (β, J) with an edge if (a) $U_\alpha \cap U_\beta \neq \emptyset$ and (b) $I \cap J \neq \emptyset$.
- $SS(X)$ is equipped with a reference map $\pi : SS(X, \mathcal{U}) \to N\mathcal{U}$ given on vertices by $(\alpha, I) \to \alpha$

Mapper - Scale Space

A varying choice of scale is now determined by a *section* of π, i.e a map

$$\sigma : N\mathcal{U} \longrightarrow SS(X, \mathcal{U})$$

so that $\pi\sigma = id_{N\mathcal{U}}$.

Mapper - Scale Space

A varying choice of scale is now determined by a *section* of π, i.e a map

$$\sigma : N\mathcal{U} \longrightarrow SS(X, \mathcal{U})$$

so that $\pi\sigma = id_{N\mathcal{U}}$.

Sections can be given an weighting depending on the length of I for the vertices and depending on the length of $I \cap J$ for the edges.

Mapper - Scale Space

A varying choice of scale is now determined by a *section* of π, i.e a map

$$\sigma : N\mathcal{U} \longrightarrow SS(X, \mathcal{U})$$

so that $\pi\sigma = id_{N\mathcal{U}}$.

Sections can be given an weighting depending on the length of I for the vertices and depending on the length of $I \cap J$ for the edges.

Finding the high weight sections in the case of 1-D filters is computationally tractable.

Variants on Persistence: Zig-Zags

Bootstrap - B. Efron

- Studies statistics of measures of central tendency across different samples within a data set

Variants on Persistence: Zig-Zags

Bootstrap - B. Efron

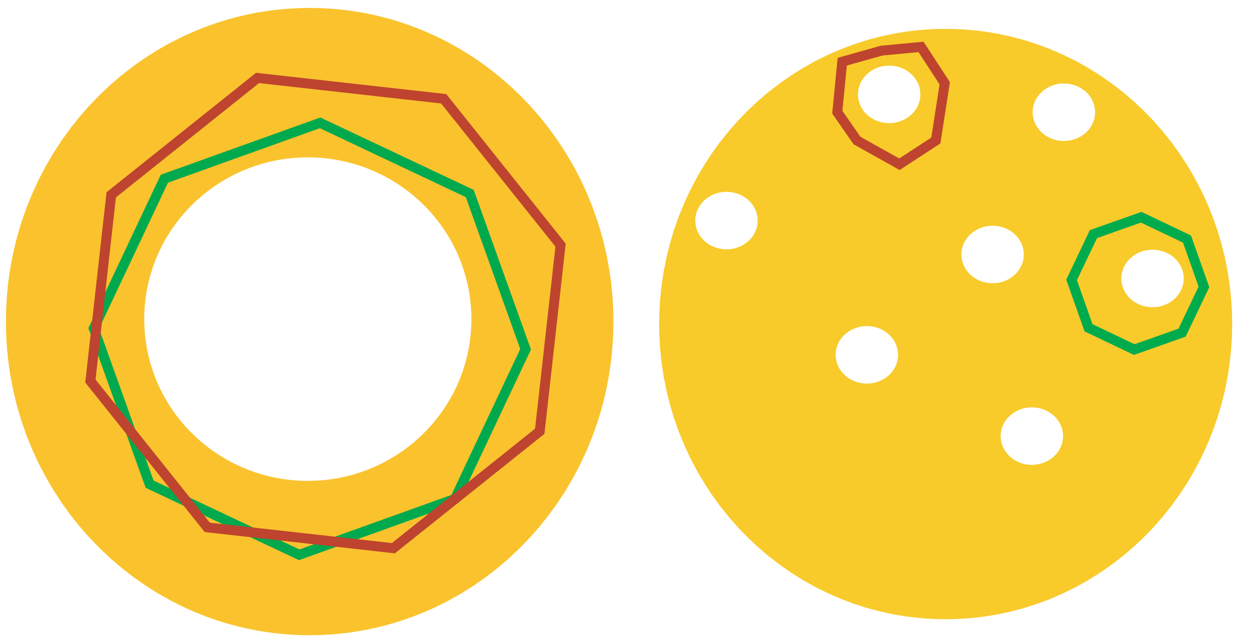
- Studies statistics of measures of central tendency across different samples within a data set
- Can give assessment of reliability of conclusions to be drawn from the statistics of the data set

Variants on Persistence: Zig-Zags

Bootstrap - B. Efron

- Studies statistics of measures of central tendency across different samples within a data set
- Can give assessment of reliability of conclusions to be drawn from the statistics of the data set
- How can one adapt the technique to apply to qualitative information, such as presence of loops or decompositions into clusters?

Variants on Persistence: Zig-Zags



How to distinguish?

Variants on Persistence: Zig-Zags

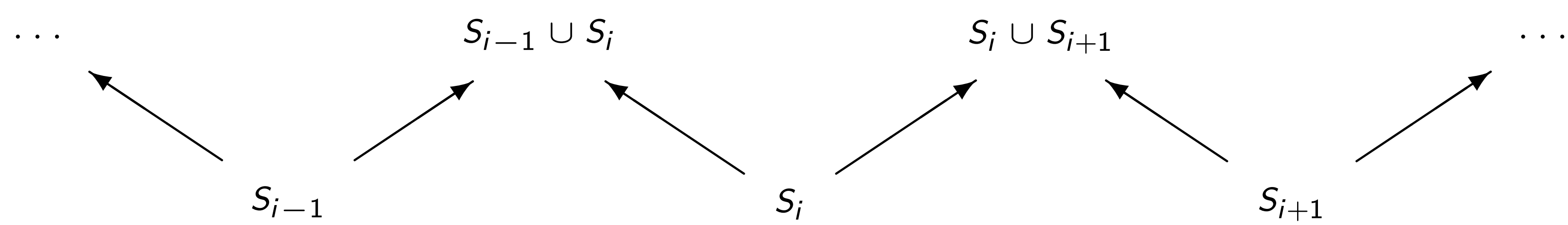
- Family of samples $S_1, S_2, \ldots, S_k$ from point cloud data $\mathbb{X}$

Variants on Persistence: Zig-Zags

- Family of samples $S_1, S_2, \ldots, S_k$ from point cloud data $\mathbb{X}$
- Construct new samples $S_i \cup S_{i+1}$

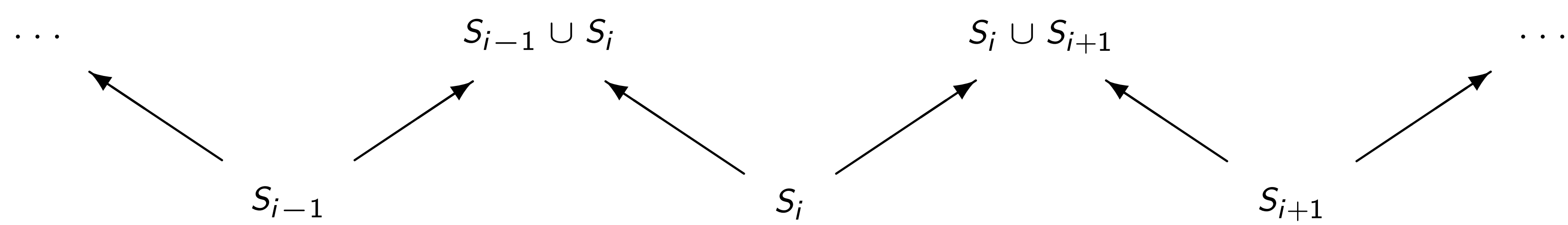
Variants on Persistence: Zig-Zags

- Family of samples $S_1, S_2, \ldots, S_k$ from point cloud data $\mathbb{X}$
- Construct new samples $S_i \cup S_{i+1}$
- Fit together into a diagram



Variants on Persistence: Zig-Zags

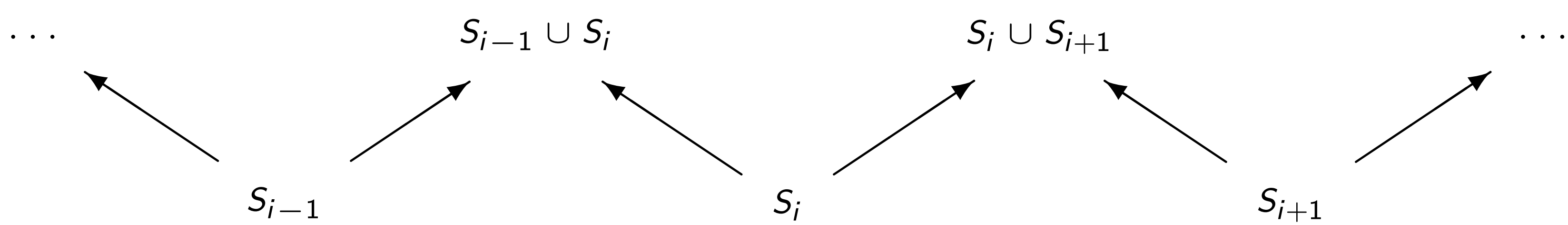
- Family of samples $S_1, S_2, \ldots, S_k$ from point cloud data $\mathbb{X}$
- Construct new samples $S_i \cup S_{i+1}$
- Fit together into a diagram



- Apply H_k to VR-complexes on each of these, get a diagram of vector spaces of same shape

Variants on Persistence: Zig-Zags

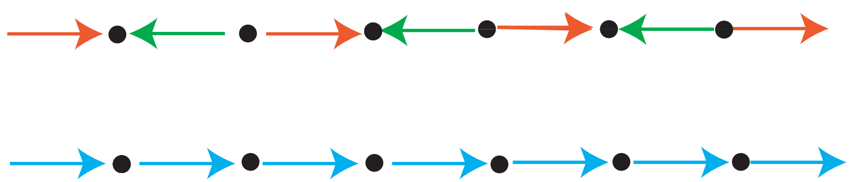
- Family of samples $S_1, S_2, \ldots, S_k$ from point cloud data $\mathbb{X}$
- Construct new samples $S_i \cup S_{i+1}$
- Fit together into a diagram



- Apply H_k to VR-complexes on each of these, get a diagram of vector spaces of same shape
- If a family of homology classes "matches up" under induced maps, then they are stable across samples

Variants on Persistence: Zig-Zags

To carry out analysis, one needs a classification of diagrams of vector spaces of shape of upper row. Second row is shape for ordinary persistence.



Variants on Persistence: Zig-Zags

Classification exists, due to P. Gabriel

Variants on Persistence: Zig-Zags

Classification exists, due to P. Gabriel

Every diagram of this shape has a decomposition into a direct sum of cyclic diagrams, i.e. diagrams which consist of either a one-dimensional or a zero dimensional vector space.

Variants on Persistence: Zig-Zags

Classification exists, due to P. Gabriel

Every diagram of this shape has a decomposition into a direct sum of cyclic diagrams, i.e. diagrams which consist of either a one-dimensional or a zero dimensional vector space.

Can therefore parametrize isomorphism classes by barcodes, just as in the case of ordinary persistence.

Variants on Persistence: Zig-Zags

Classification exists, due to P. Gabriel

Every diagram of this shape has a decomposition into a direct sum of cyclic diagrams, i.e. diagrams which consist of either a one-dimensional or a zero dimensional vector space.

Can therefore parametrize isomorphism classes by barcodes, just as in the case of ordinary persistence.

Long intervals correspond to elements stable across samples, others are artifacts.

Variants on Persistence: Zig-Zags

Results have value in other situations:

Variants on Persistence: Zig-Zags

Results have value in other situations:

- Analysis of time varying data

Variants on Persistence: Zig-Zags

Results have value in other situations:

- Analysis of time varying data
- Analysis of behavior of data under varying choice of density estimators

Variants on Persistence: Zig-Zags

Results have value in other situations:

- Analysis of time varying data
- Analysis of behavior of data under varying choice of density estimators
- Analysis of behavior of witness complexes under varying choices of landmarks

Variants on Persistence: Zig-Zags

Results have value in other situations:

- Analysis of time varying data
- Analysis of behavior of data under varying choice of density estimators
- Analysis of behavior of witness complexes under varying choices of landmarks

This analysis is relevant and interesting even in zero dimensional case, i.e. clustering.