Introduction to Topological Data Analysis

Priyavrat Deshpande

Chennai Mathematical Institute

Data has shape.

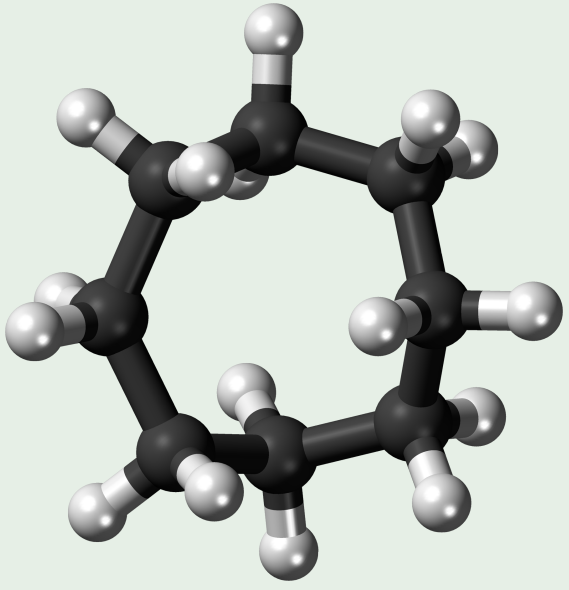
Molecular Conformation Space

Definition

Given an isomer it is the space of all possible configurations of constituting atoms in 3-space up to rigid motions.

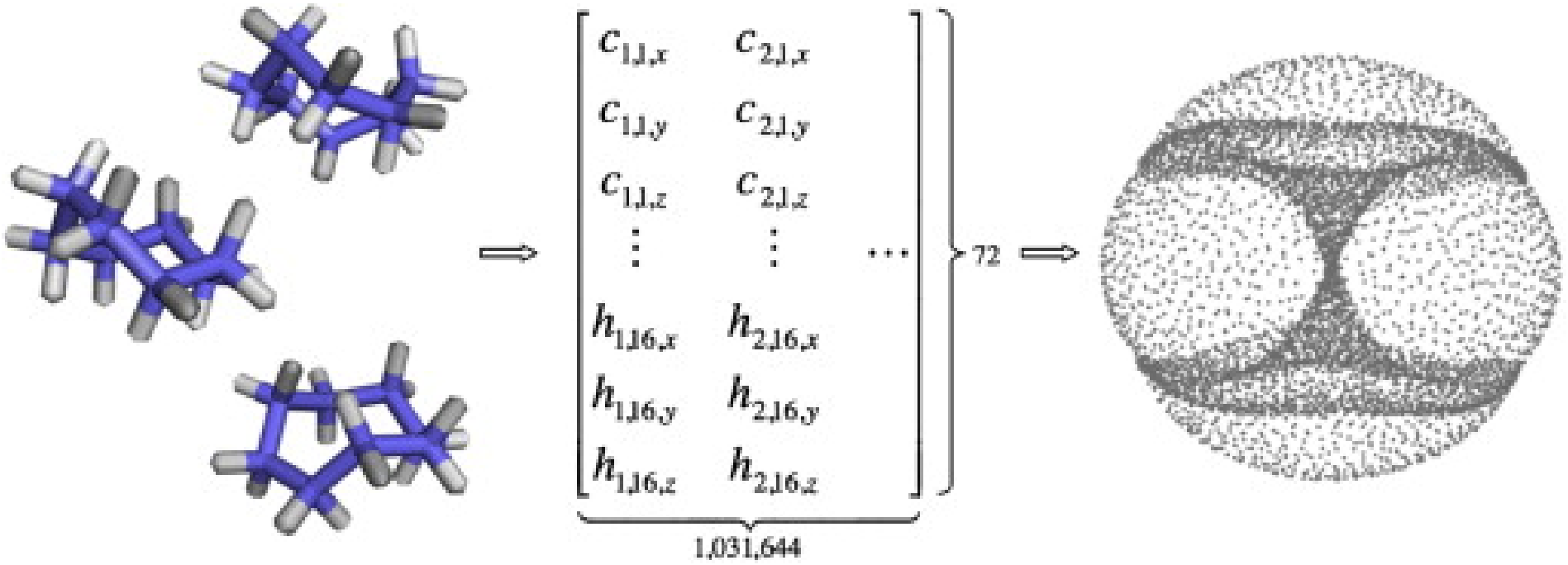
Example (Cyclo-octane molecule)

Chemical formula C_8H_{16}. Each of the 8 carbon atoms have 3 degrees of freedom; the conformation space is 24-dimensional.



The space description?

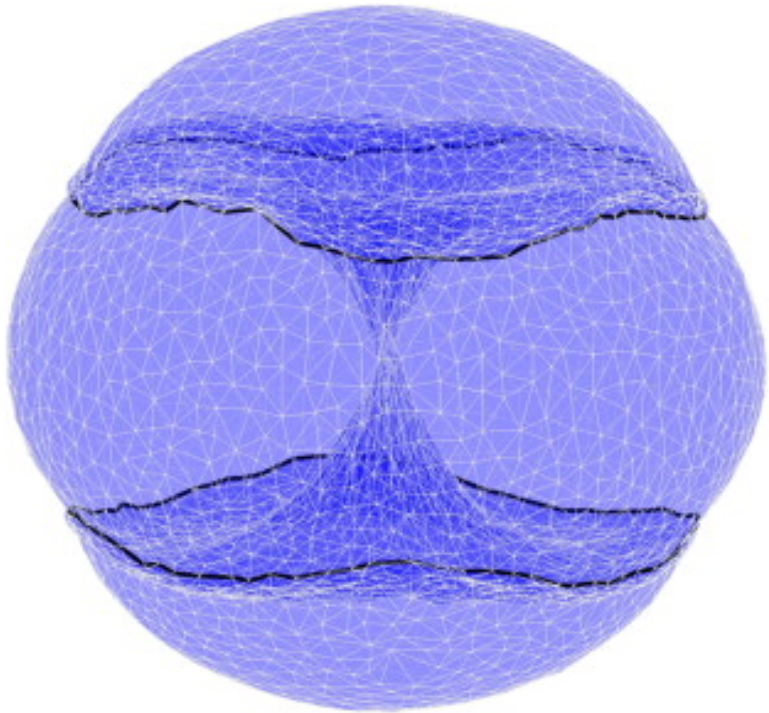
The Cyclo-Octane dataset contains 1 million$+$ points in 24-D. However, the actual configuration space is 2-dimensional.



Source

Non-manifold surface reconstruction from high-dimensional point cloud data by S. Martin and JP Watson in Comp. Geo. 2011.

Triangulation of the surface



Source

Non-manifold surface reconstruction from high-dimensional point cloud data by S. Martin and JP Watson in Comp. Geo. 2011.

Flocking, Swarming etc.

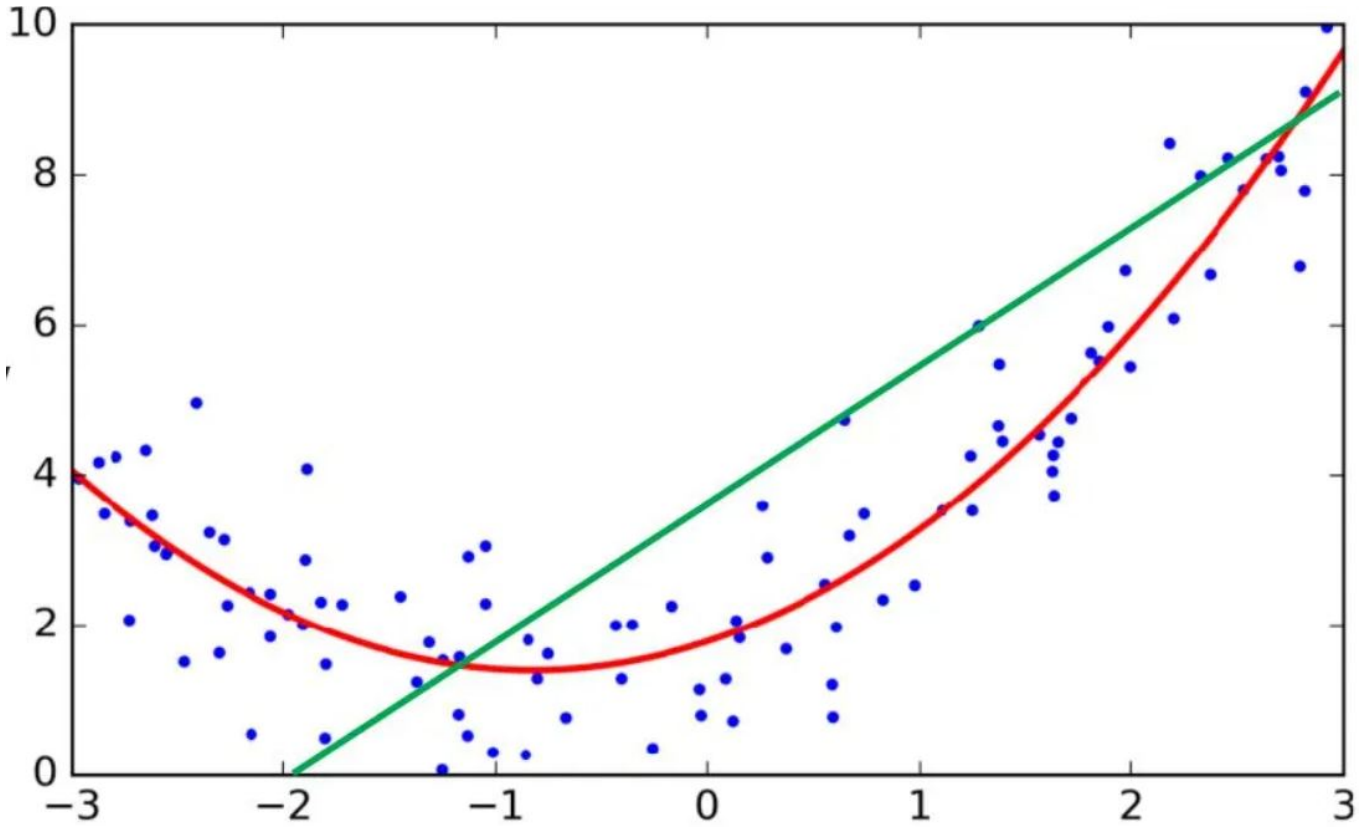
Murmuration

It is the phenomenon that results when thousands of starlings fly in swooping, intricately coordinated patterns through the sky.

Shape has meaning.

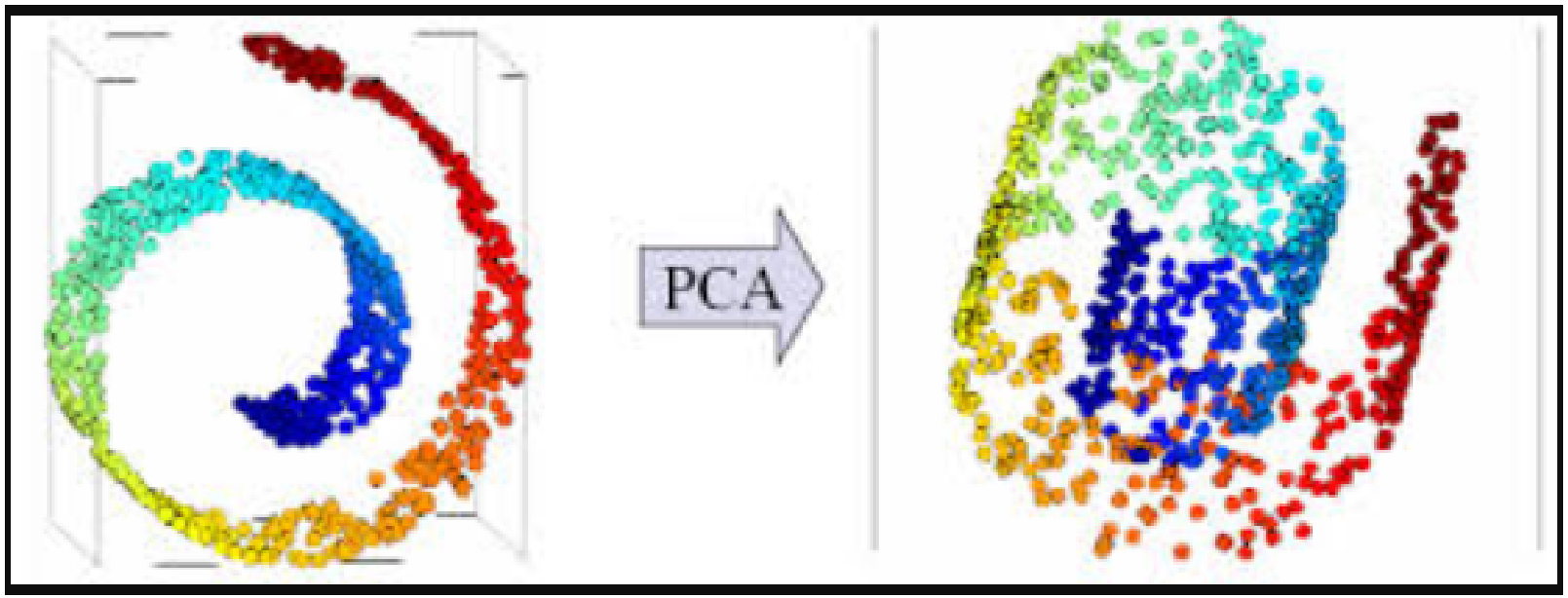
Shape encodes structure & meaning

Fitting a linear regression makes sense if data is along a hyperplane.



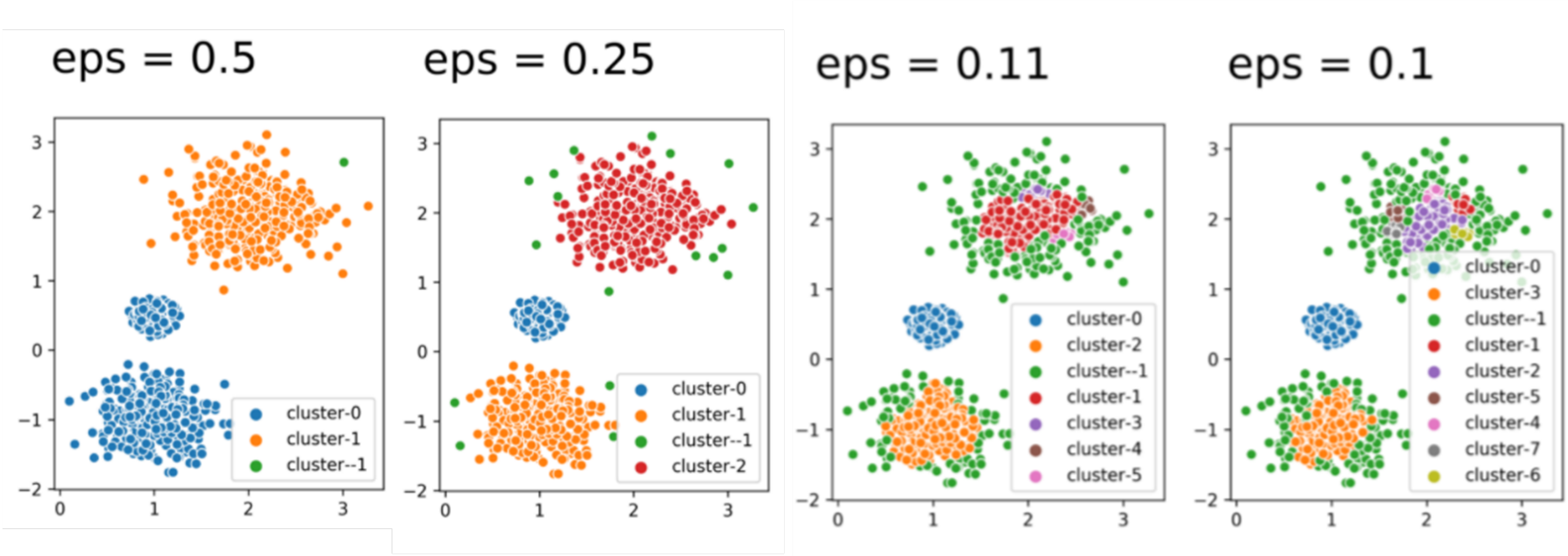
Effective application requires understanding the *shape* of the data.

Complex data



Linear methods fail when data has non-linear structure.

Simple shape?



Clustering algorithms are sensitive to choice of parameters which might be hard to determine in high-dimensional data.

TDA addresses the *shape* aspect of the data analysis.

Course information

Schedule	Mondays and Wednesdays at 2 pm.
Venue	Zoom
Instructor	Priyavrat
Office hours	?
Prerequisites	Metric spaces, linear algebra, Python, R
Texts	No specific textbook. Computational topology by Edelsbrunner and Harer. Elementary applied topology by R. Ghrist. Papers/ Notes posted on Moodle
Evaluation	HW (40%), project (40%), presentations (20%)

Outline

A brief intro

- A large, actively growing interdisciplinary field.
- Uses ideas from algebraic topology to create robust tools for studying high-dimensional data sets.
- Avoids shape-dependent cottage industry.
- TDA "*describes*" the shape of the data that is invisible to classical techniques.

TDA helps...

- Gain insight into the organization of data at a global level
- Analyze data in a manner which is independent of choice of embedding and co-ordinates.
- Capture complex relationships among features in the data.
- Translate ideas of topological invariance to the study of data.

The TDA pipeline

1. Input: Point cloud equipped with a pairwise distance matrix or a similarity metric.
2. Build: A *continuous shape* is constructed on top of the data.
3. Extract: Topological/ geometric information is quantified.
4. Analyze: New families of features and data descriptors are discovered; expressed through visualization.

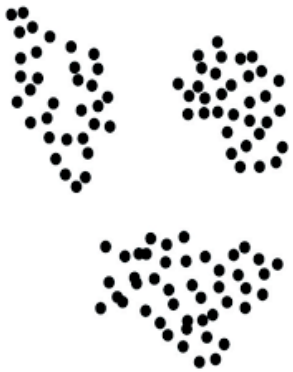
TDA: in detail

The aim of TDA

Apply topology to develop tools for studying

qualitative features of data.

- Data: A finite metric space embedded in a Riemmanian manifold (say, $\mathbb{R}^n$).
- qualitative features: coarse-scale, global geometric features.



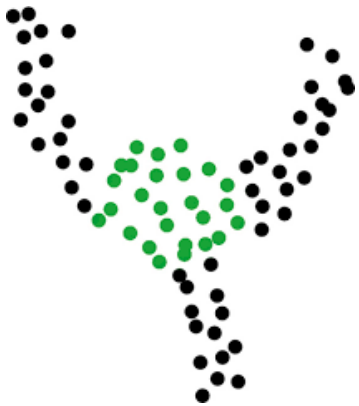


Image source: Ayasdi blog

In detail

To develop

- Formal definitions of such features.
- Computational tools for detection.
- Tools for visualization.
- Methodology for quantifying the statistical significance.
- Tools suitable for high-dimensional PCD.

The key assumption

Insight into the *shape* of data yields insight into the data itself.

The TDA successes

A list of few applications -

- Image analysis, character/ object recognition, medical imaging etc.
- Genetic analysis, evolutionary biology.
- Medicines (discovery of new form of breast cancer and diabetes)
- Sports analytics.
- Sensor networks and coverage.
- Cosmology, galaxy distribution etc.
- Customer analysis.
- Time series classification, anomaly detection

More attention needed in -

- Financial and insurance sector.
- Developmental economics, (Geo)politics etc.
- Business analytics.

Outline

What is Topology?

A branch concerned with properties of *geometric* objects that are invariant under continuous deformations.

What is Algebraic Topology?

Aim

- To translate topological structures in algebraic language.
- To understand properties that remain invariant under deformations.

Example

- Homology groups
- Cohomology algebra
- Characteristic classes
- Homotopy groups
- Many more...

What is homology?

Presence of holes is a property that remains invariant under continuous maps.

In homology we ...

- Formalize the notion of holes.
- Calculate the numbers of holes of different types.
- Study various implications of presence of holes.

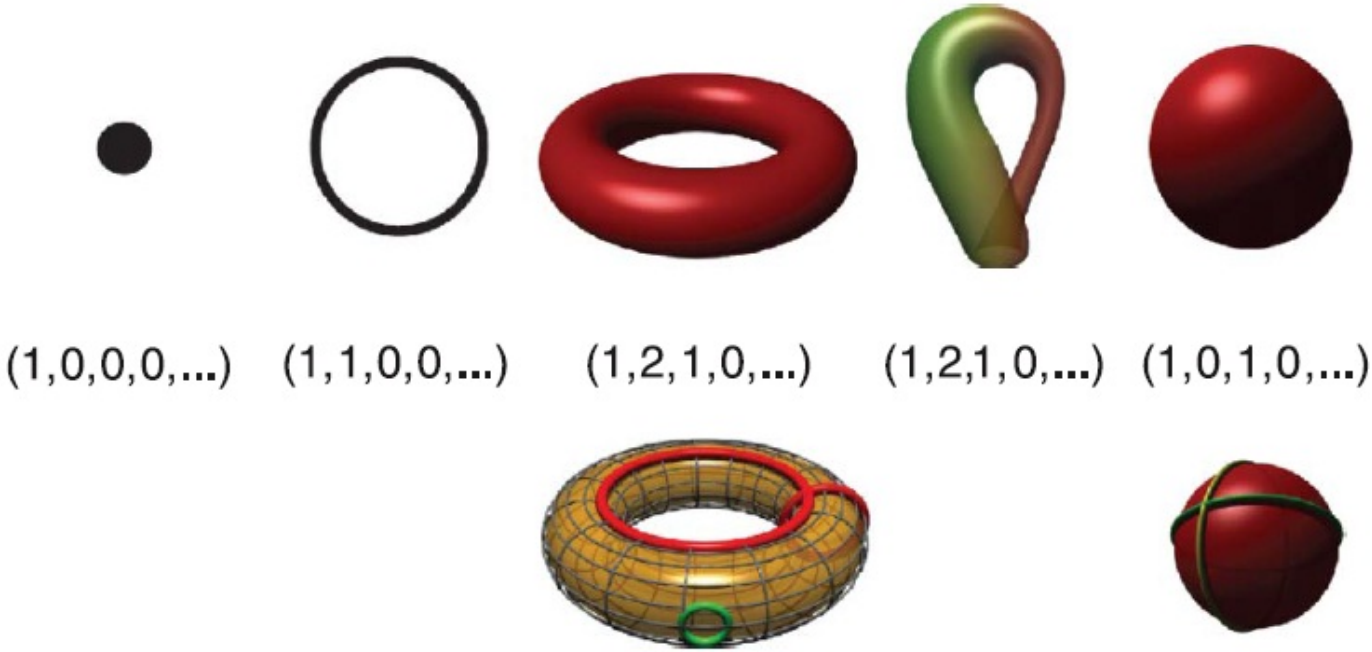
Betti numbers

A numerical invariant that computes the number of i-dimensional holes in a space X for every $i \geq 0$.

Types of holes

- 0-dimensional holes: Number of connected components of the space.
- 1-dimensional hole: Tunnels you can see through. (medu vadai)
- 2-dimensional hole: Hollow space in an object. (balloon)

Examples of mod-2 Betti numbers.



In general, a hole is a "nice" copy of an i-sphere sitting inside.

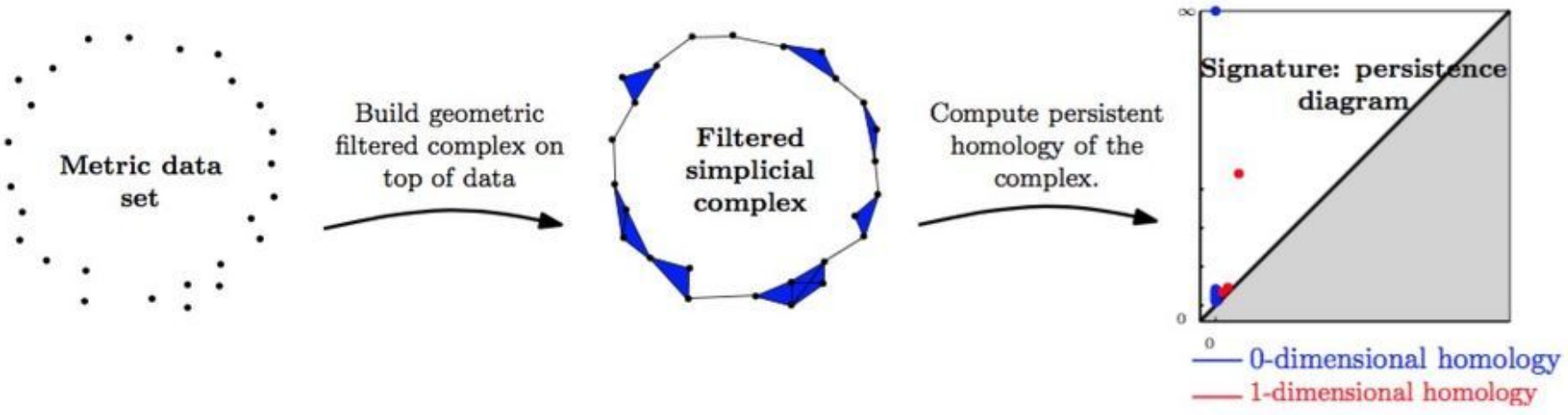
Back to data

Persistent Homology - answers

How can we use the hole-detection formalism of AT to develop robust computational methods to extract qualitative features of data?

- It produces simple descriptors of qualitative features called *bar codes*.
- Starts by building spaces (filtered simplicial complex) on top of PCD.
- Meta theorem - "*every reasonable PCD is sampled from a nice manifold*".
- The idea is that bar codes capture homological features that *persist*.
- Bar codes are not sensitive to noise.

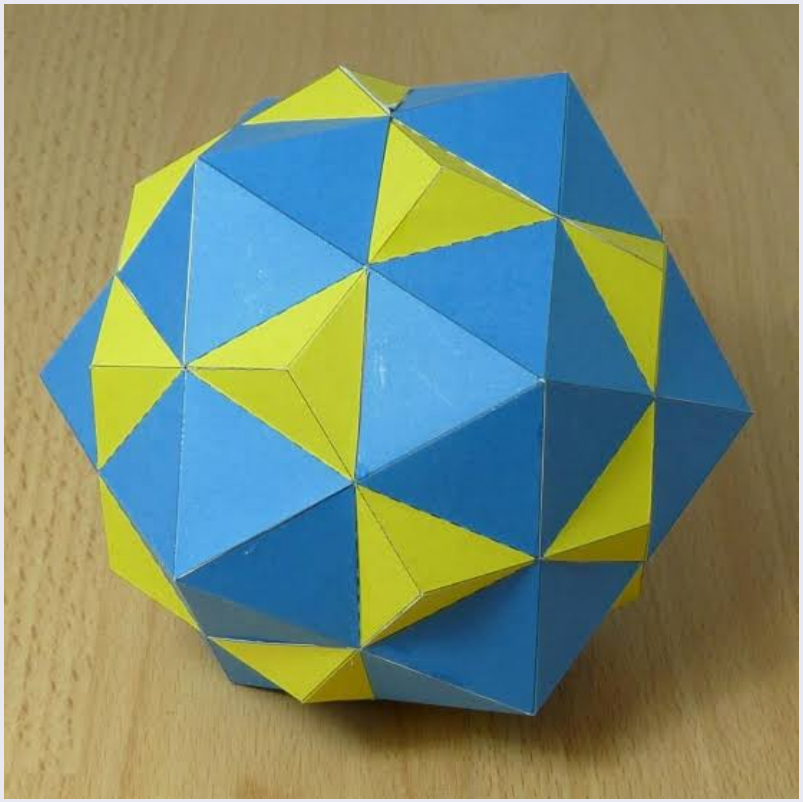
The PH pipeline



How do we model shapes?

One needs a suitable mathematical model to construct and manipulate shapes in all dimensions.

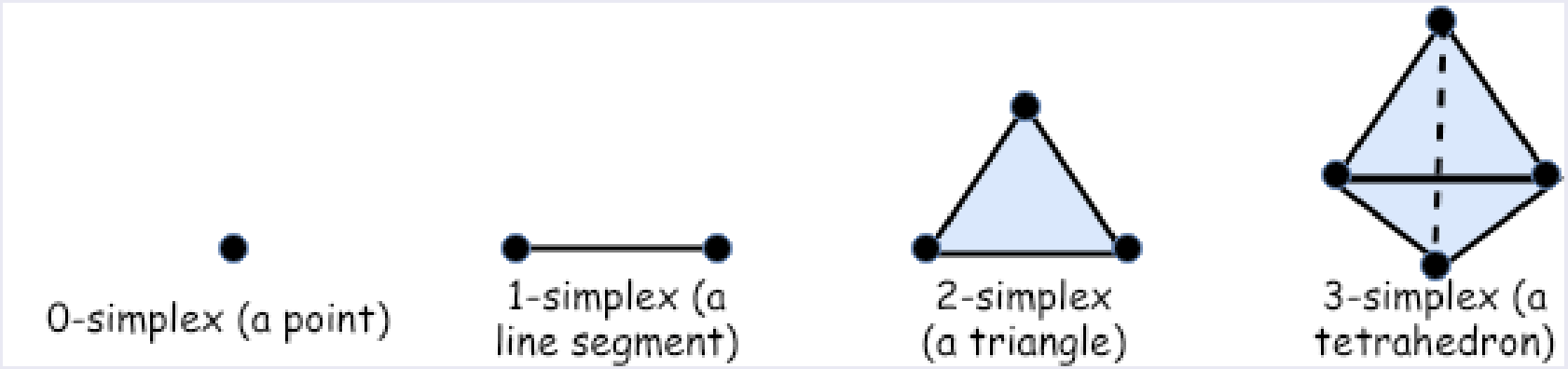
Modular origami



Building blocks

Simplices

- 0-simplex: single point.
- 1-simplex: line segment.
- 2-simplex: filled triangle.
- 3-simplex: filled tetrahedron.



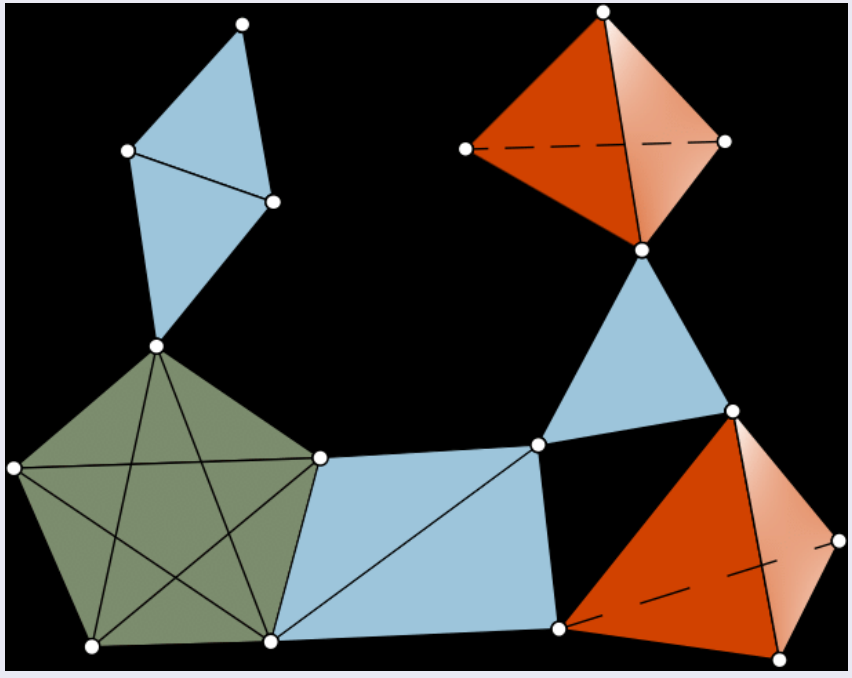
Simplicial complexes

Definition (Simplicial complex K)

It is (finite) union of simplices such that

i for all $\sigma \in K$ all the faces of σ are also in K;

ii the intersection of any two simplices is either empty or a common face.



Simplicial chains

- Let $\{\sigma_1, \ldots, \sigma_p\}$ be the set of k-simplices of K.
- A simplicial k-chain is a linear combination $c := \sum_{i=1}^{p} \epsilon_i \sigma_i$ where $\sigma_i \in \mathbb{F}_2$.
- The set of all simplicial chains, $C_k(K)$, form a vector space over $\mathbb{F}_2$.
- We have $\dim C_k(K) = p$.

Definition (The boundary operator)

The linear map $\partial_k : C_k(K) \to C_{k-1}(K)$ is given by

$$\partial_k([v_{i_0} \ldots, v_{i_k}]) = \sum_j [v_{i_0}, \ldots, \widehat{v_{i_j}}, \ldots, v_{i_k}]$$

Important Property

$$\partial_{k+1} \circ \partial_k = 0.$$

The homology

Definition (The simplicial chain complex)

$$0 \to C_n(K) \xrightarrow{\partial_n} C_{n-1}(K) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1(K) \xrightarrow{\partial_1} C_0(K) \to 0.$$

Definition

The simplicial homology groups

$$H_k(K, \mathbb{F}_2) := \frac{\ker \partial_k}{\operatorname{im} \partial_{k+1}}.$$

Definition (Betti numbers)

For $1 \leq i \leq n$

$$\beta_i(K) := \dim H_i(K, \mathbb{F}_2).$$

Examples

Space	Betti numbers
Point	$\beta_0 = 1$ and $\beta_i = 0$ for $i \geq 1$.
$\mathbb{R}^n$, $n \geq 1$	$\beta_0 = 1$ and $\beta_i = 0$ for $i \geq 1$.
n-sphere, $n \geq 1$	$\beta_0 = \beta_n = 1$ and $\beta_i = 0$ otherwise.
$(S^1)^n$	$\beta_i = \binom{n}{i}$, $1 \leq i \leq n$ and 0 otherwise.
Genus g surface	$\beta_0 = \beta_2 = 1, \beta_1 = 2g$ and $\beta_i = 0$ otherwise.

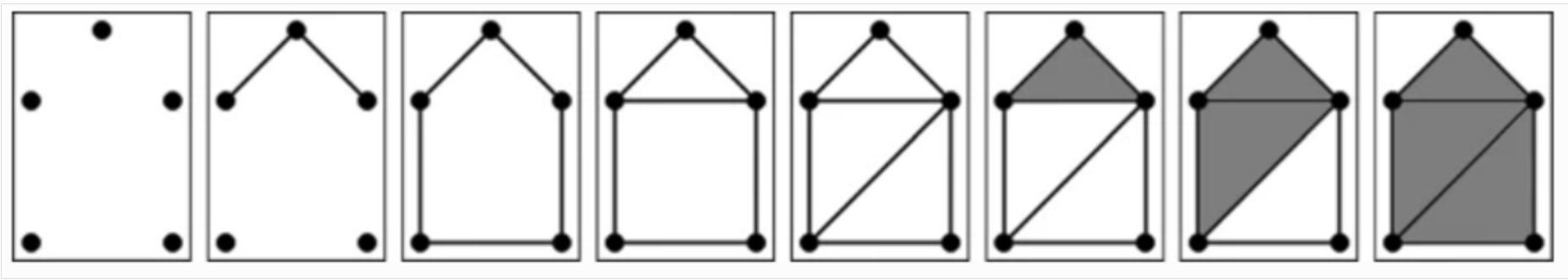
Topological invariance

- The simplicial homology is a functor from simplicial complexes to abelian groups/ vector spaces.
- A homeomorphism of simplicial complexes induces an isomorphism of homology groups in all dimensions.

Outline

Persistent Homology of Simplicial Complexes

Given a filtration $\{C_n\}$ of a simplicial complex C, persistent homology(PH) can be used to study how the homology modules evolve.

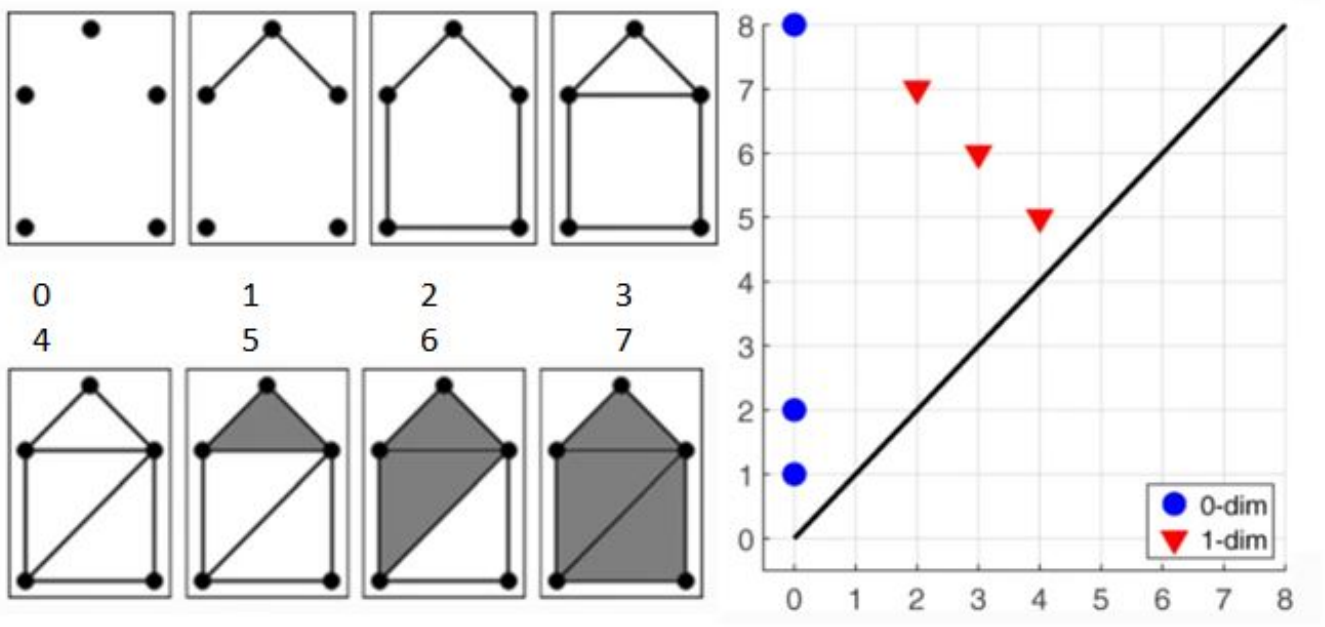


Definition

For $i < j$ the $(i, j)^{th}$ **persistent homology** of C at dimension p is denoted by $H_p^{i \to j}(C)$ is the image of the map $i : \ H_p(C_i) \longrightarrow H_p(C_j)$.

Persistent Homology of Simplicial Complexes

The dimension of persistent homology vector spaces is encoded in a **persistence diagram**; it is a multiset of points in $\mathbb{R}^2$. Each tuple corresponds to the 'birth' and 'death' of a Betti number.



Data to space: the idea

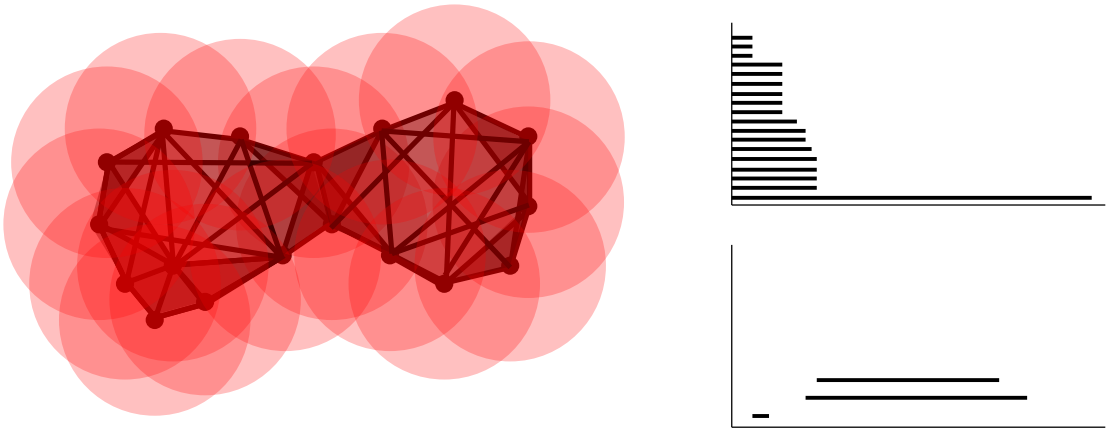
The most common type of data is a point cloud — a set of vectors $\{x_1, \ldots, x_N\}$ in $\mathbb{R}^d$.

A PCD is a 0-dimensional simplicial complex, no interesting topology

The idea

Construct a family of simplicial complexes following the example of hierarchical clustering.

Data to space: an example



Vietoris-Rips complex

Definition

Let $X \subset \mathbb{R}^N$ be a finite point cloud and $\epsilon > 0$. The Vietoris-Rips complex, $VR_\epsilon(X)$, has as k-simplices those $(k+1)$-subsets $\{x_{i_0}, \ldots, x_{i_k}\}$ of X for which

$$d(x_{i_j}, x_{i_l}) \leq \epsilon.$$

- In general $VR_\epsilon(X)$ does not embed in $\mathbb{R}^N$.
- If $\epsilon < \epsilon'$ then $VR_\epsilon(X) \subseteq VR_{\epsilon'}(X)$.
- It is a clique complex, i.e., completely determined by its 1-skeleton.

The Cech complex

Point cloud data

A point cloud $\mathbb{X}$ is a set of vectors $\{x_1, \ldots, x_N\}$ in $\mathbb{R}^d$.

Definition (Cech Complex)

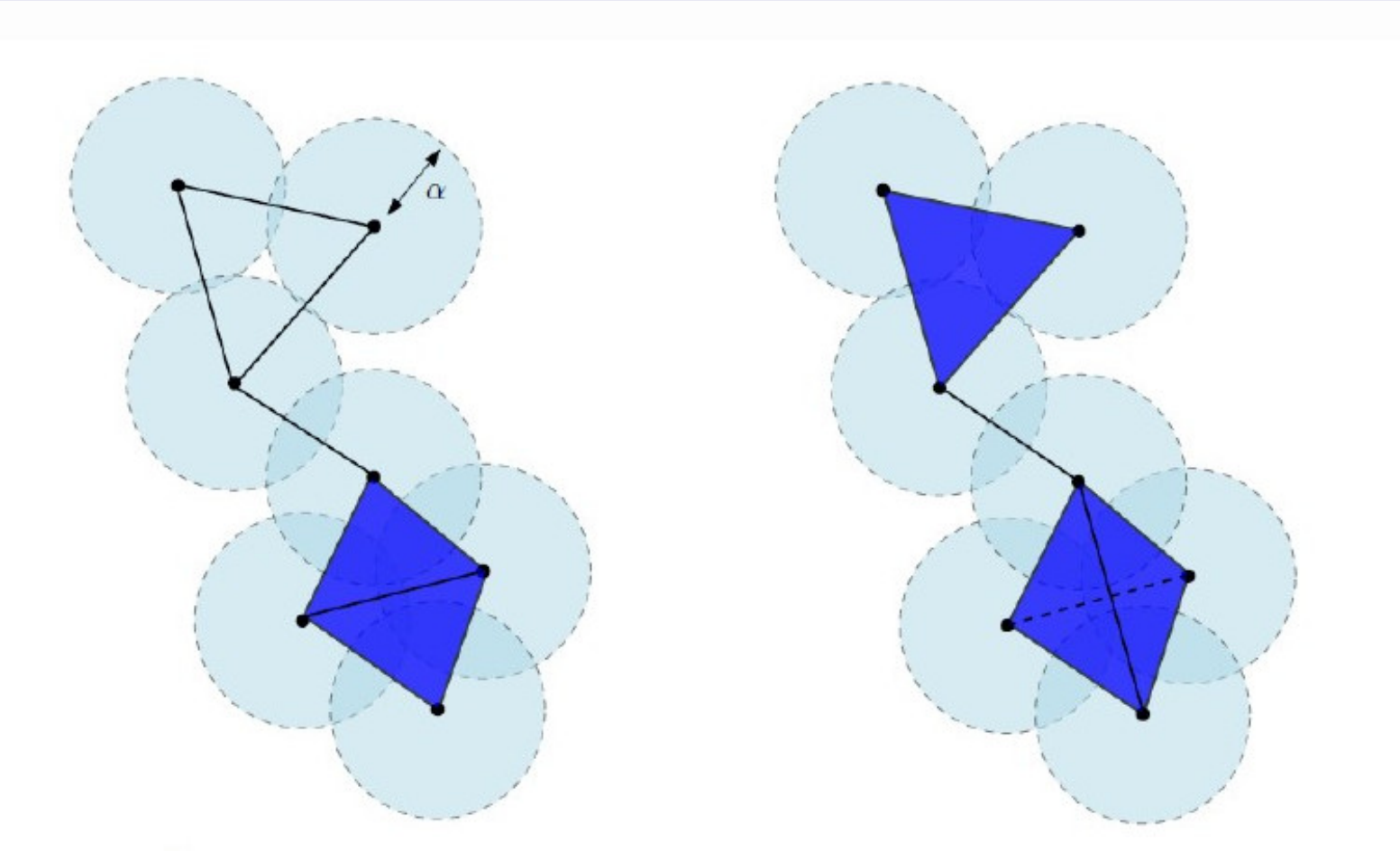
Given a PCD $\mathbb{X}$ and $r > 0$ the Cech complex $\mathrm{Ch}_r(\mathbb{X})$ is an (abstract) simplicial complex whose simplices are those subsets $\sigma \in \mathbb{X}$ such that

$$\bigcap_{x \in \sigma} B(x; r) \neq \emptyset.$$

The Cech complex

- If $r < r'$ then $\mathrm{Ch}_r(\mathbb{X}) \subseteq \mathrm{Ch}_{r'}(\mathbb{X})$.
- For a radius r, a subset σ of $\mathbb{X}$ a simplex if and only if the corresponding set of points is contained in a ball of radius r.
- Checking whether a set of points is contained in a ball of given radius is a well studied problem in computational geometry.

An example



Left hand side we have Ch_α and on the right hand side we have $\mathrm{V}_{2\alpha}$.

The relationship

Proposition

$$\mathrm{Ch}_r(\mathbb{X}) \subseteq \mathrm{V}_{2r}(\mathbb{X}) \subseteq \mathrm{Ch}_{2r}(\mathbb{X}).$$

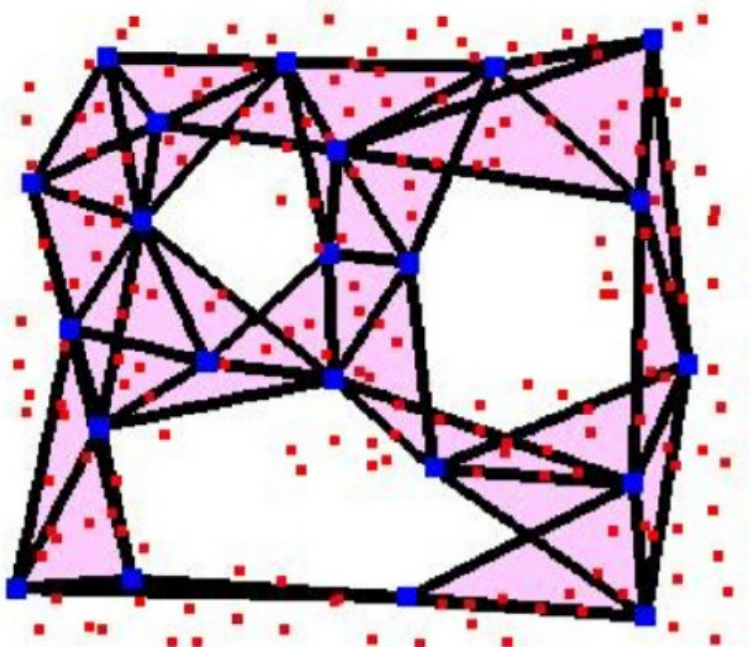
Proof

- Let $\sigma \in \mathrm{Ch}_r$. Then there is a ball of radius r that contains σ. The diameter of such a ball is at most $2r$, hence $\sigma \in \mathrm{V}_{2r}$.
- Let $\tau \in \mathrm{V}_{2r}$. Then $\mathrm{diam}(\tau) \leq r$. Hence there is a ball of radius $2r$ containing τ. Hence $\tau \in \mathrm{Ch}_{2r}$.

The witness complex

Definition

Given a PCD $\mathbb{X}$ and a chosen subset L of landmark points the witness complex, $W(\mathbb{X}, L)$ is defined as follows: the vertices of W are the points in L. For each $x \in \mathbb{X}$, we find two points $l_1, l_2 \in L$ that are closest to x, and add the edge $\{l_1, l_2\}$. A higher simplex is added if and only if all its edges are present.



The witness complex

How to choose L?

- Randomly.
- Pick l_1 at random. Then choose l_2 such that $d(l_1, l_2)$ is maximum. Choose l_3 that maximizes $\min\{d(l_1, l_3), d(l_2, l_3)\}$ etc.
- Choose from denser regions.

The scale parameter

For a scale parameter $r > 0$ the witness complex $W_r(\mathbb{X}, L)$ has vertex set L and $\{l_1, l_2\}$ is an edge if there exists $x \in X$ such that

$$d(x, l_1), d(x, l_2) \leq \text{const.} + r.$$

Other complexes

1. The Alpha complex.
2. The flow complex.
3. The Delaunay triangulation.

Persistent homology of PCD

An increasing sequence $\epsilon_{i_1} < \cdots < \epsilon_{i_n}$ induces a filtration

$$\emptyset \subset VR_{\epsilon_{i_1}}(X) \subseteq \cdots \subseteq VR_{\epsilon_{i_n}}(X).$$

For every $p \geq 0$ we have:

$$H_p(VR_1(X)) \xrightarrow{f_p^{0,1}} H_p(VR_2(X)) \xrightarrow{f_p^{0,2}} \cdots \xrightarrow{f_p^{n-1,n}} H_p(VR_n(X)).$$

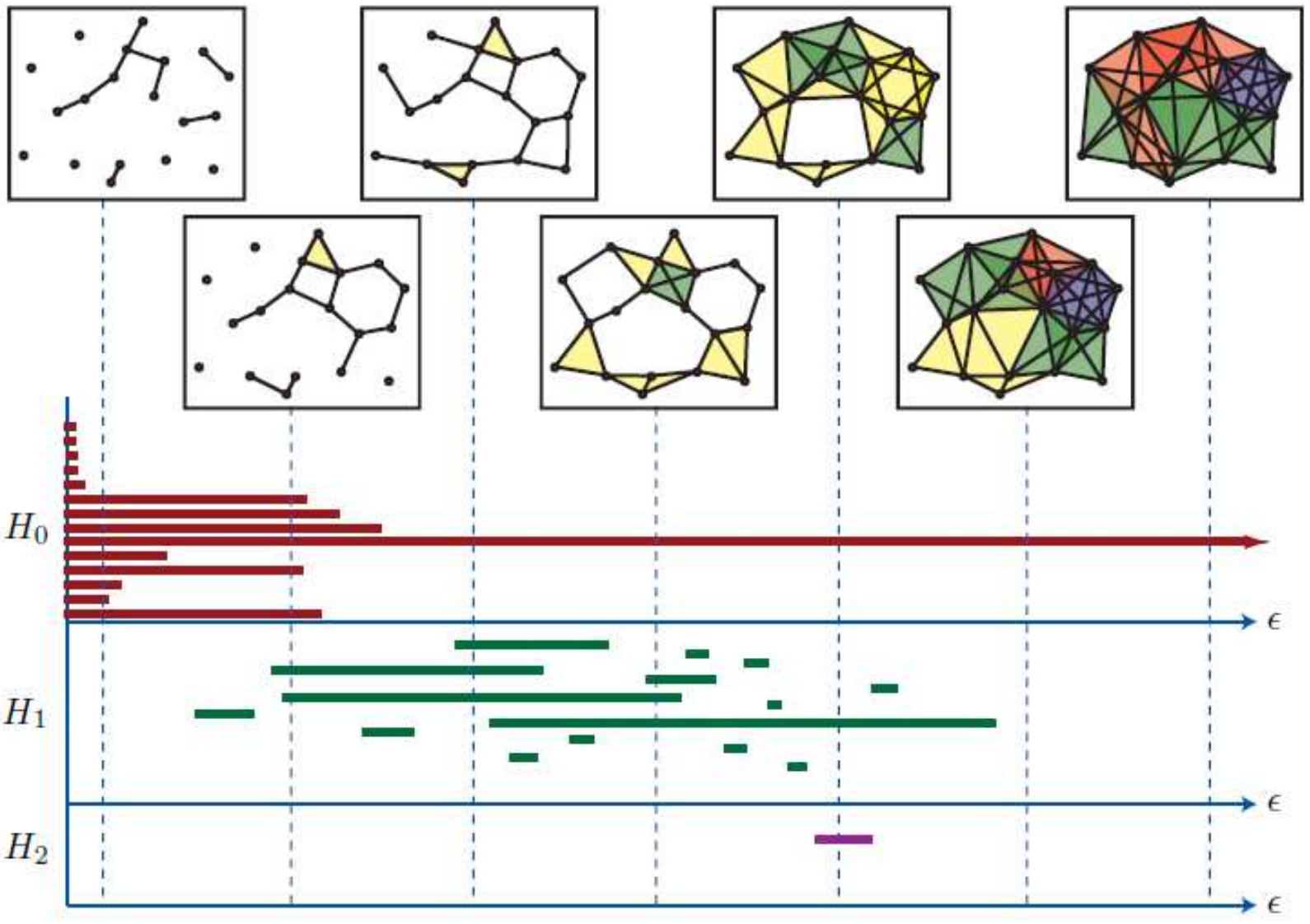
In general for $i < j$

$$f_p^{i,j} : H_p(VR_i(X)) \to H_p(VR_j(X)).$$

Definition

p-th persistent homology group: $\mathcal{H}_p^{i,j} := \mathrm{Im}(f_p^{i,j})$.

Example



Barcodes and diagrams

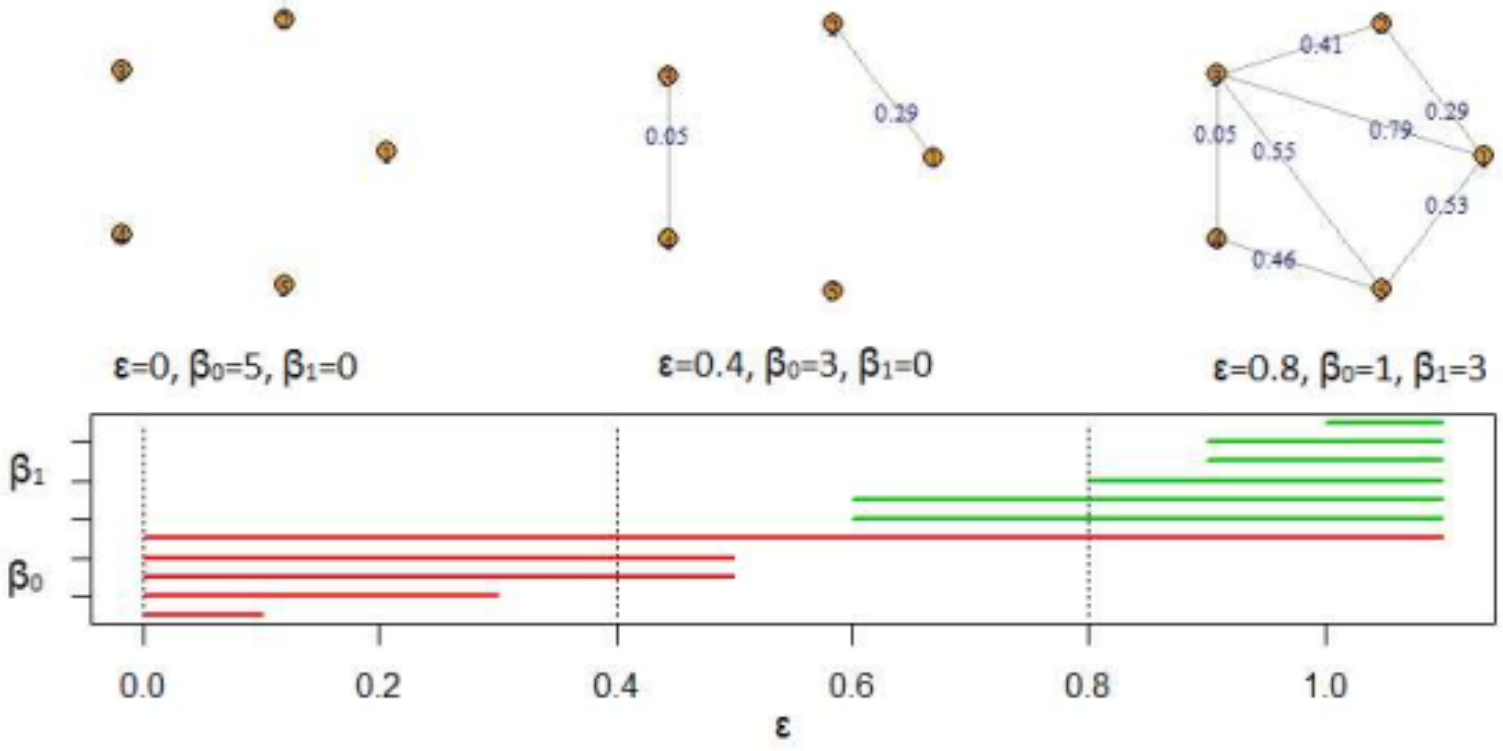
Definition (Persistence barcodes)

For every $p \geq 0$ we draw a graph whose vertical axis corresponds to all possible p-homology generators and the horizontal axis is the time parameter.

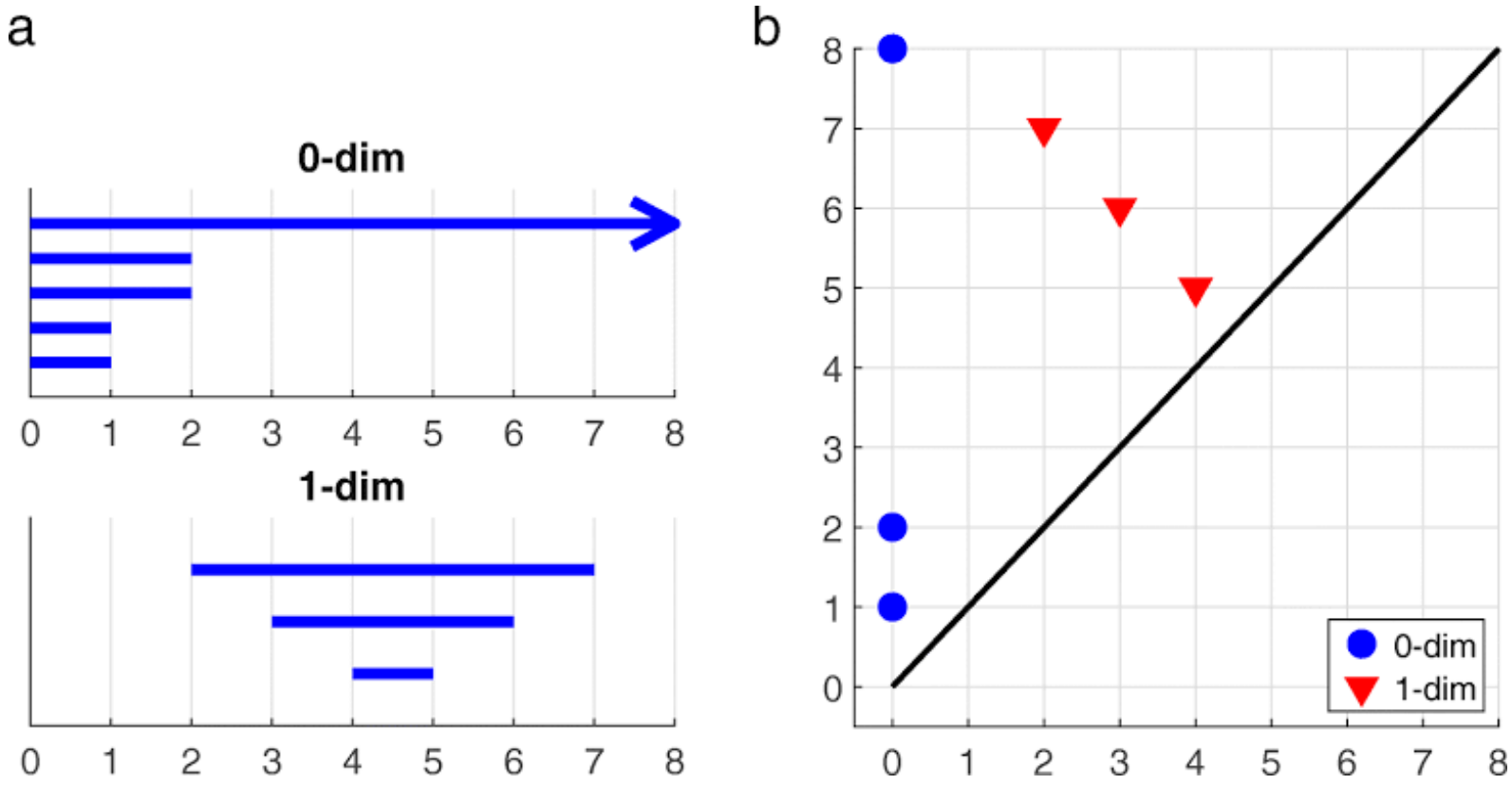
Definition (Persistence diagram)

The p-persistence diagram is a 2-d coordinate system where x is the birth coordinate and y is the death coordinate. For every p-homology class there is a point (i, j) representing its birth and death time.

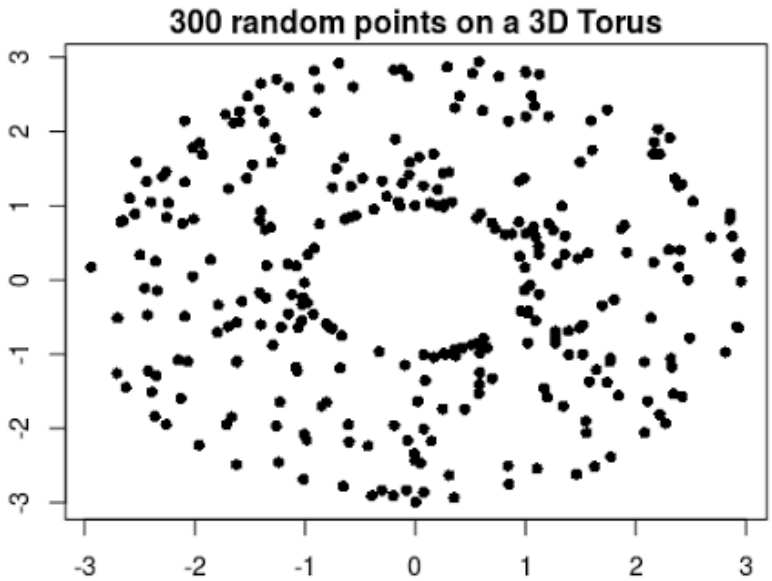
Example

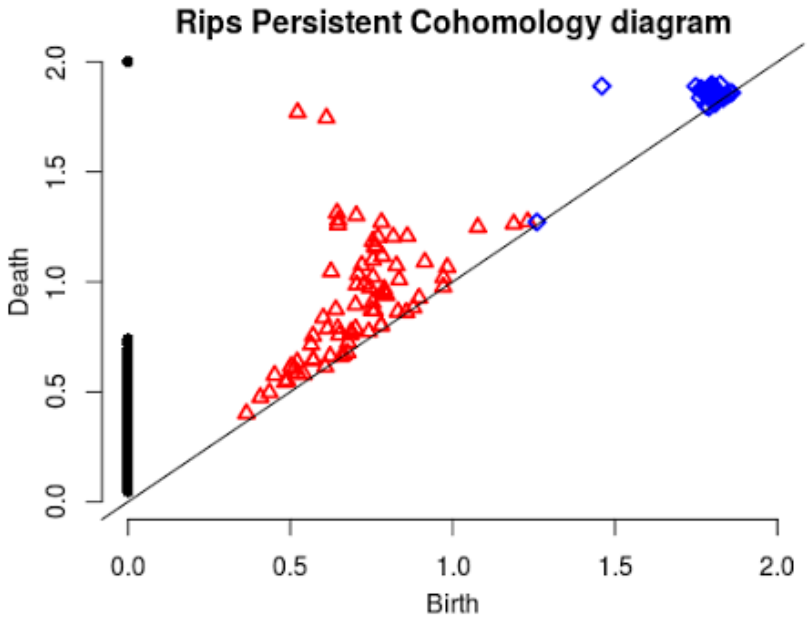

0.05
0.29
0.41
0.05
0.55
0.79
0.29
0.53
0.46
ε=0, β0=5, β1=0
ε=0.4, β0=3, β1=0
ε=0.8, β0=1, β1=3
β1
β0
0.0
0.2
0.4
0.6
0.8
1.0
ε

Example


a
0-dim
0 1 2 3 4 5 6 7 8
1-dim
0 1 2 3 4 5 6 7 8
b
8 7 6 5 4 3 2 1 0
0 1 2 3 4 5 6 7 8
0-dim
1-dim

Persistence diagrams





How to interpret?

1. High persistence implies existence of **robust** features.
2. Spurious topological features are short-lived, i.e., **noise**.
3. The *summary description* is always 2-dimensional.
4. Persistent diagrams are a similarity metirc.
5. β_0: number of connected components (clusters?).
6. β_1: number of cycles (periodic features?).
7. β_2: number of hollow spaces (?).

Outline

R TDA

- Install `TDA` library in R.

PH using Rips complex

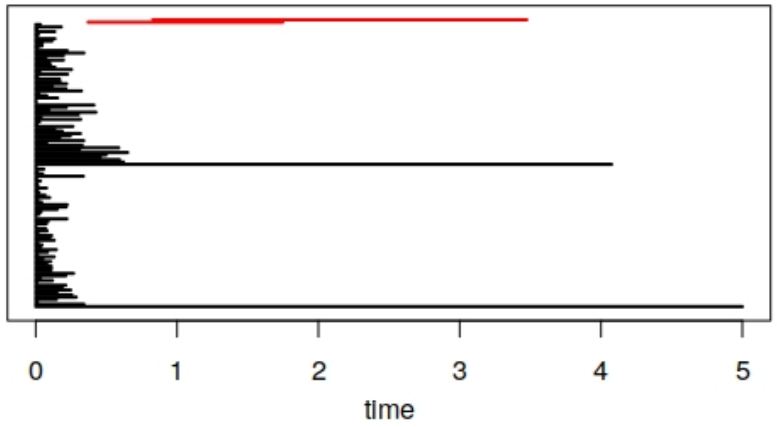
ripsDiag(X, maxdimension, maxscale, dist = "euclidean", library = "GUDHI", location = FALSE, printProgress = FALSE)

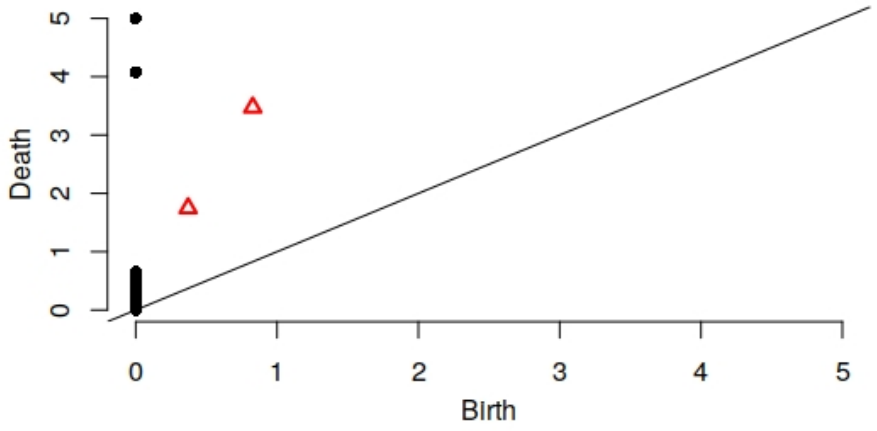
- X is and $n \times d$ matrix of coordinates; $n =$ number of points and $d =$ dimension of the ambient space.
- *maxdimension*: max dimension of the topological feature.
- *maxscale*: maximum value of the filtration.

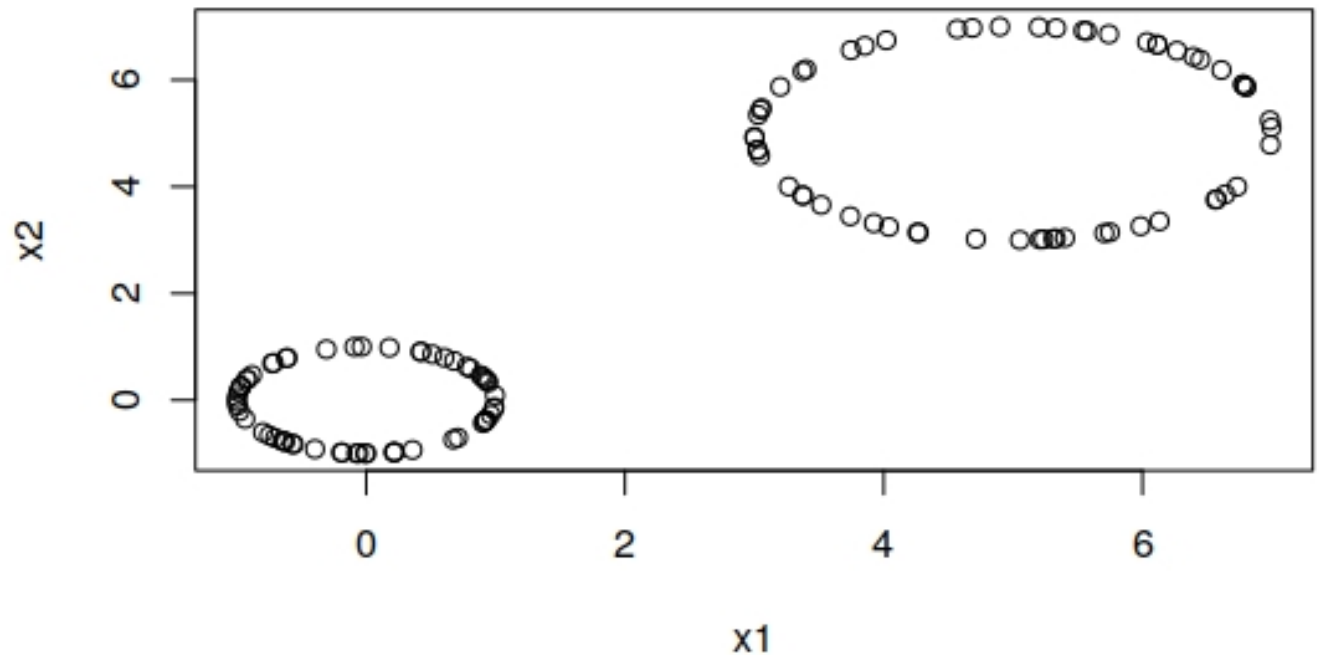
Plot the diagram

plot(x, diagLim = NULL, dimension = NULL, col = NULL, rotated = FALSE, barcode = FALSE, band = NULL, lab.line = 2.2, colorBand = "pink", colorBorder = NA, add = FALSE)

Examples





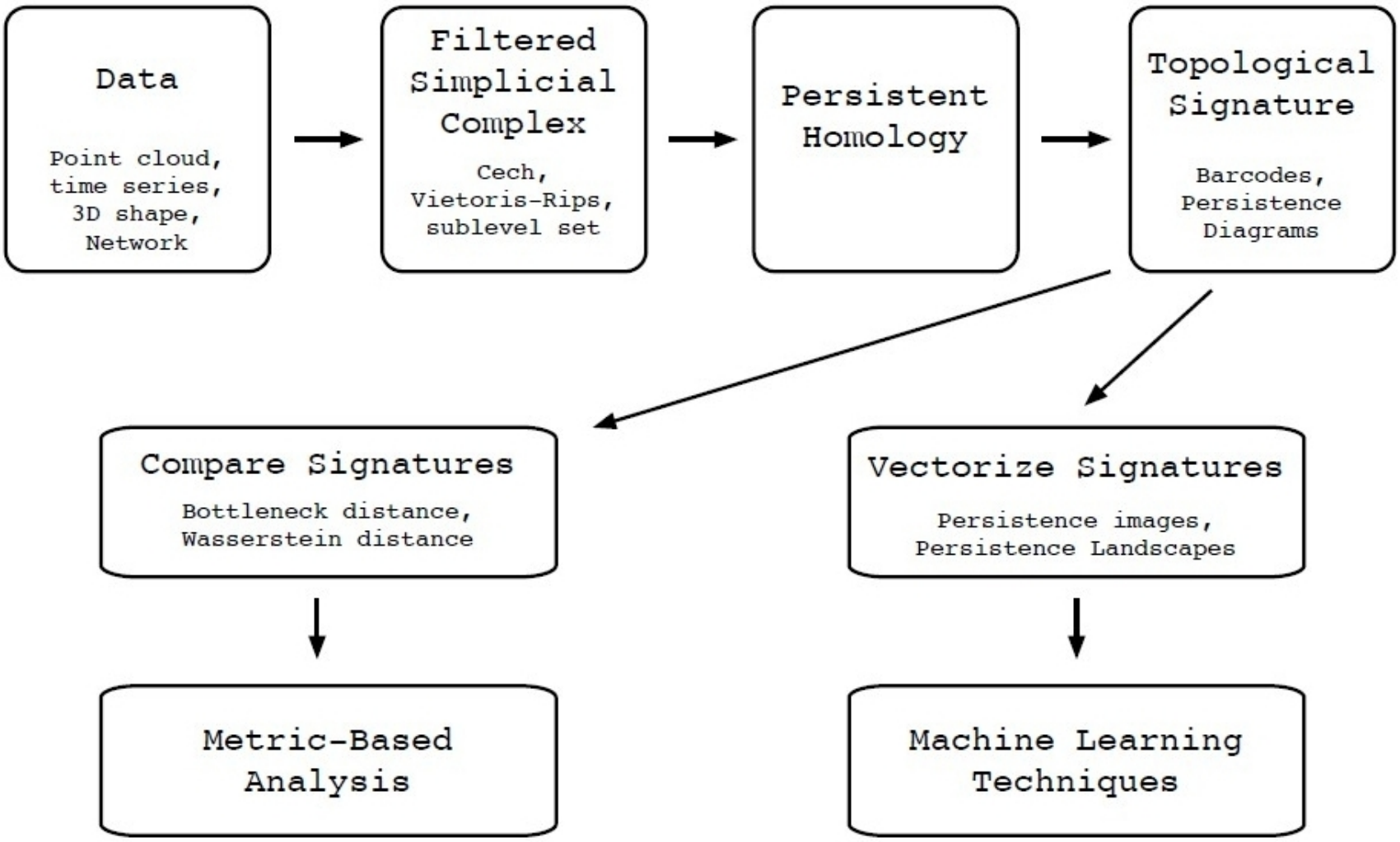

x2
0
2
4
6
0
2
4
6
x1

Properties of a barcode

1. Stable w.r.t. perturbation in the data.
2. Sensitive to "*small/big*" holes.
3. Possible to track holes, record size/scale of the feature.
4. Not sensitive to outliers.
5. Distinguishes between significant and insignificant features.
6. Computable in practice.
7. Provides a flexible framework (i.e., clusters/flares etc.).

Outline

The TDA pipeline



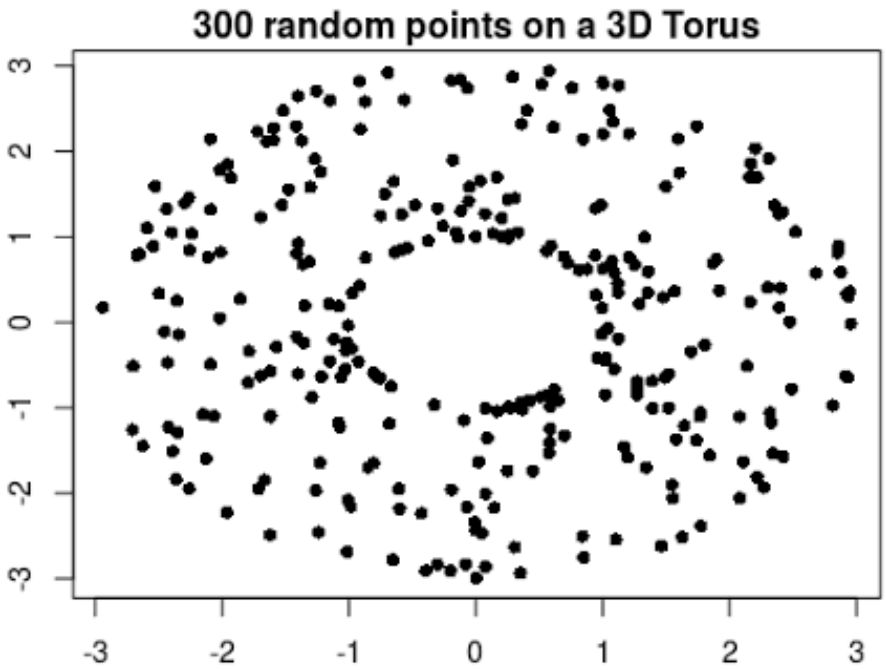
Definitions

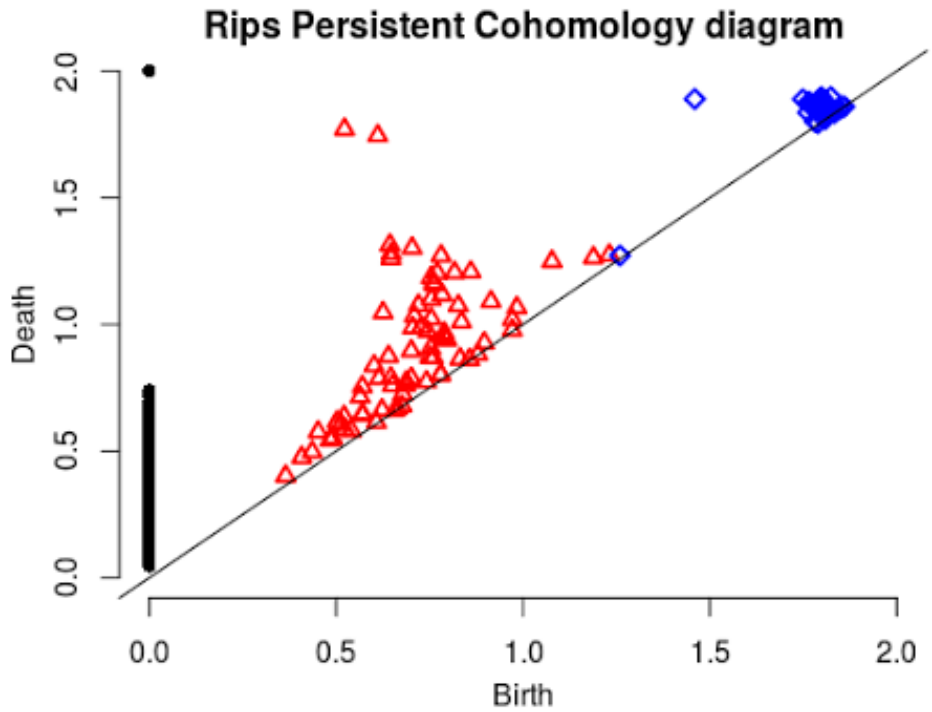
Definition (Persistence diagram)

It is a (finite) multiset X in $\mathbb{R}^2$ containing *birth-death* pairs $x_i := (b_i, d_i)$.

1. Each pair x_i is called a cycle (of appropriate dimension).
2. The lifetime of a cycle x_i is called persistence; $\mathrm{pers}(x_i) = d_i - b_i$.
3. Persistence diagram is a type of topological summary.
4. The space of all PDs supports various metrics that differentiate topological features.

Persistence diagrmas





The bottleneck distance

Definition

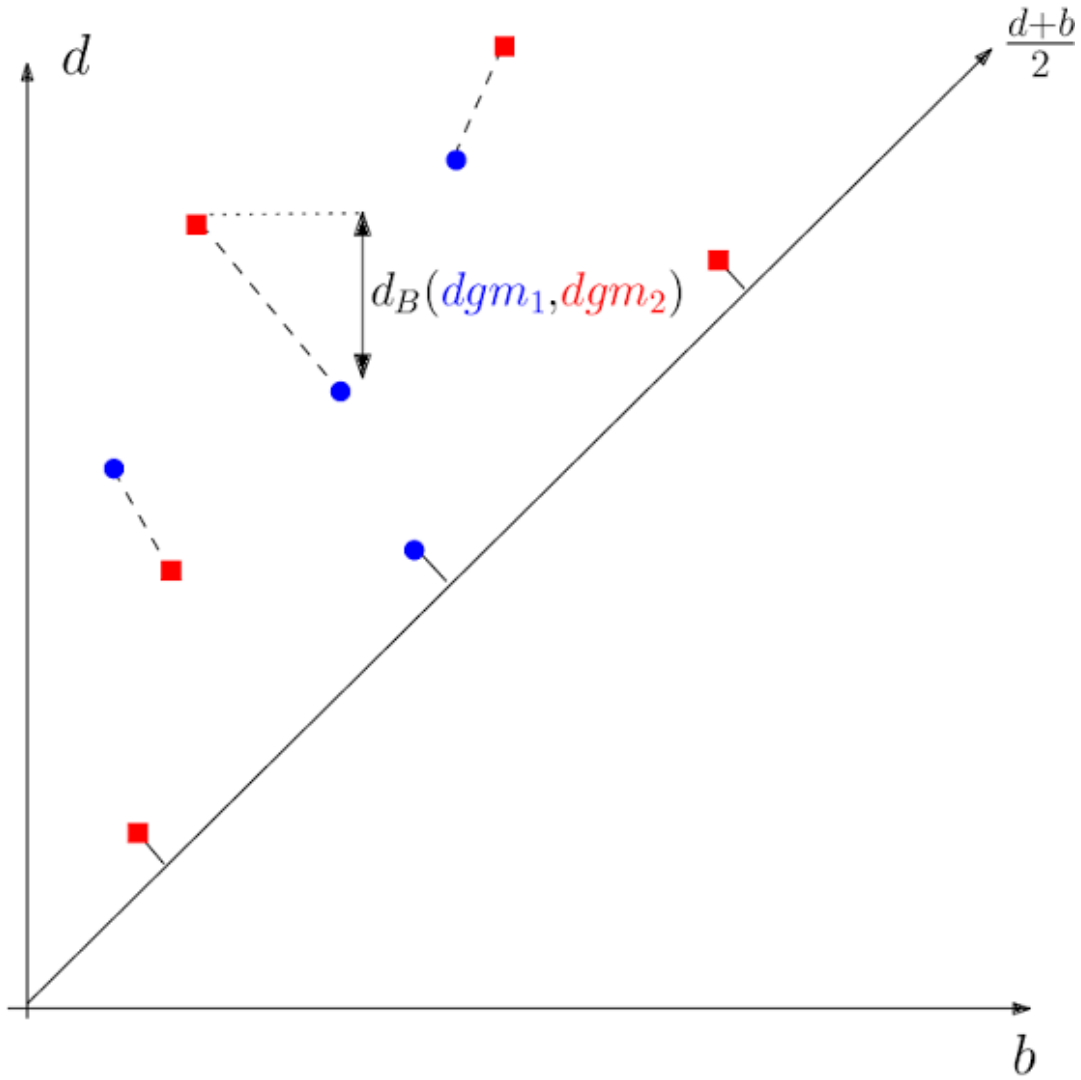
For two PDs X, Y the bottleneck distance (∞-Wasserstein metric) is defined as

$$d_B(X,Y) := \inf_{\gamma} \sup_{x \in X} ||x - \gamma(x)||_\infty,$$

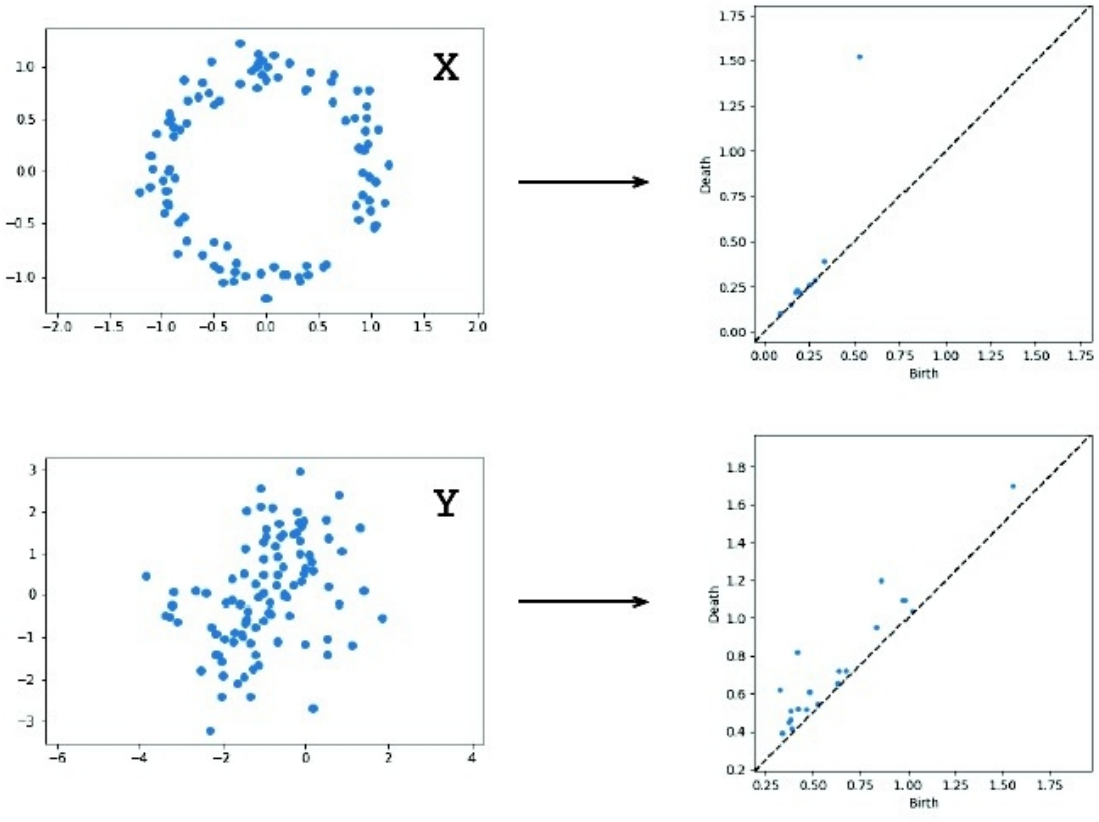
where γ runs over all the matchings (bijections) from X to Y.

1. The space of PDs with d_B is a metric space.
2. There are similar distance functions.
3. Proves stability of PH operation.
4. PD is not a vector.

Optimal transport



Persistence diagrams



The stability theorem

Theorem

Denote by Let $\mathbb{X}_1, \mathbb{X}_2$ be two PCDs and denote by $D_p(\mathbb{X})$ the persistence diagram corresponding p-persistence homology. Then

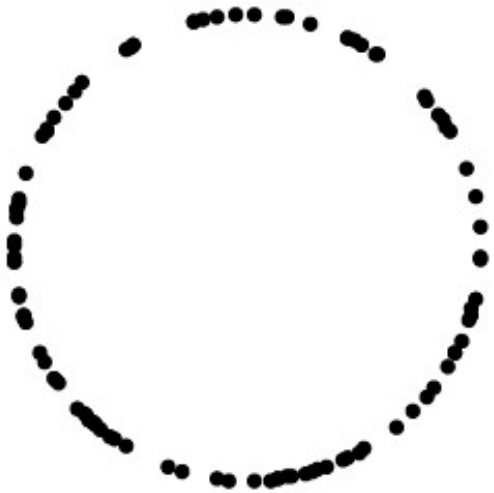
$$d_B(D_p(\mathbb{X}_1), D_p(\mathbb{X}_2)) \leq d_H(\mathbb{X}_1, \mathbb{X}_2),$$

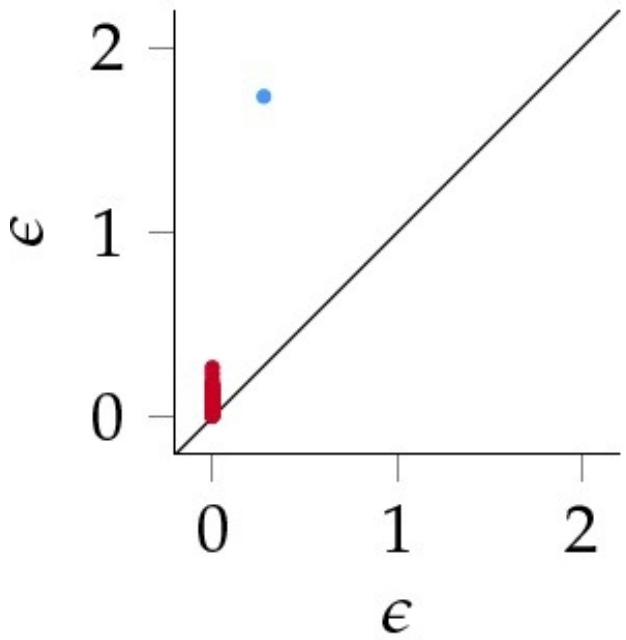
where $d_H(\ ,\ )$ is the Hausdorff distance between the sets.

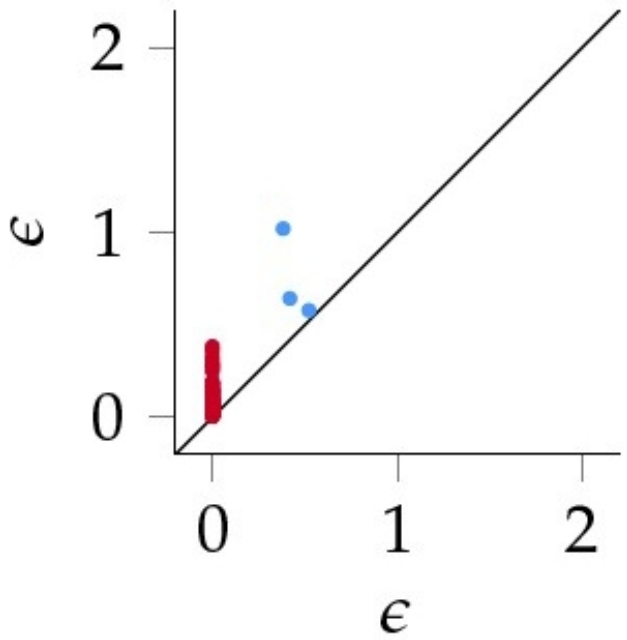
Intuitive meaning

The persistent homology doesn't change under mild perturbation of the data.

An example







Statistics on PDs

Question

Given a PD $D_p(\mathbb{X})(=: X)$. Does X behave like a random variable of the data?

PDs equipped with the bottleneck distance is not suitable for applying staistics.

Statistics and machine learning

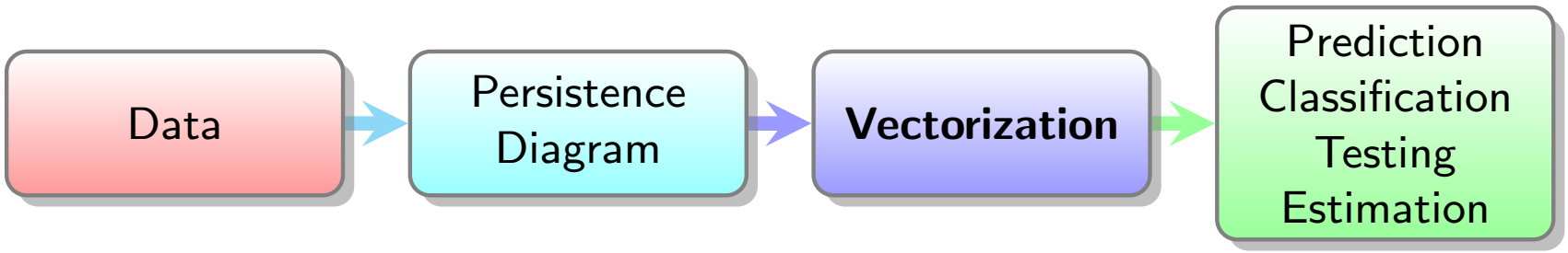
Stats and ML tasks

- clustering
- hypothesis testing
- calculating averages
- variance
- classification

Expectations from vectorization

- no information loss
- stable w.r.t. perturburations
- continuous as well as discrete versions.

Topology to ML



Recently a lot of methods have been discovered that convert topological features into vectors that can be used for statistical analysis as well as input to ML algorithms.

Persistence landscapes

First described in

P. Bubenik, Statistical topological data analysis using persistence landscapes. J. Machine Learning Research, (2015).

- They quantify 'covered' topological features.
- The idea is to 'peel off' layers iteratively.
- A landscape can be sampled at regular intervals to obtain a fixed-size feature vector.
- There is a built-in hierarchy.
- No information is lost.
- Recently it has been used a neural net layer.

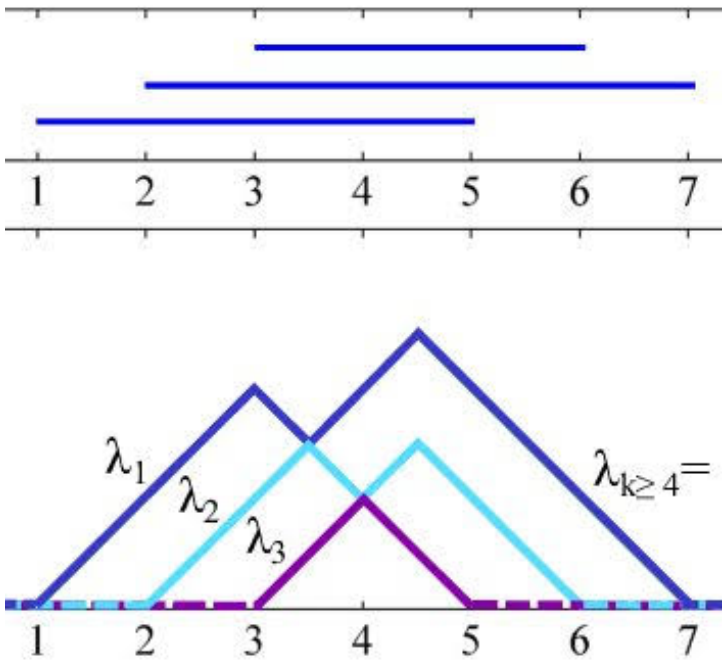
Persistence Landscape

Given an interval $J := [b, d]$ consider the real valued function

$$f_J(t) := \begin{cases} 0 & \text{if } t \notin J, \\ t - b & \text{if } b \leq t \leq \frac{b+d}{2} \\ d - t & \text{if } \frac{b+d}{2} \leq t \leq d \end{cases}$$

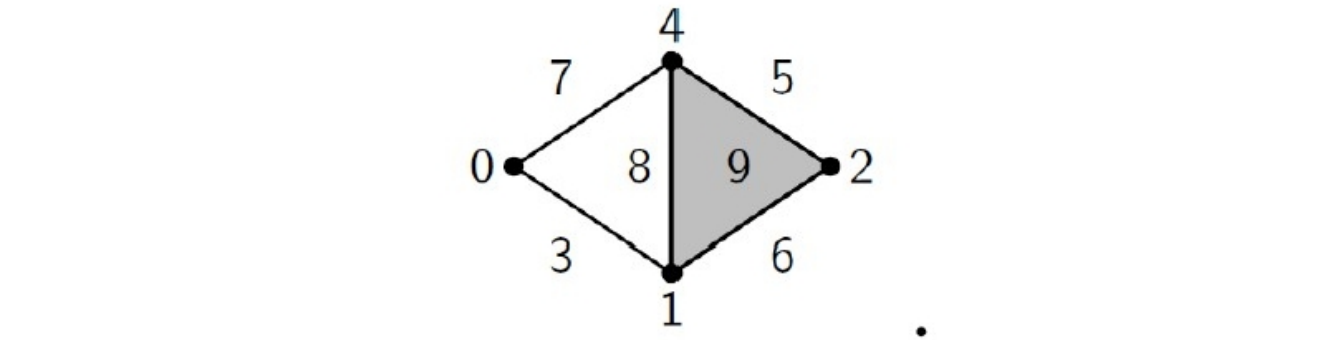
Given a collection of intervals J_i in a barcode B, we get a sequence λ_k of functions, for $k \in \mathbb{N}$:

$$\lambda_k(x) := k \max\{f_{J_i}(x)\}.$$



1. The sequence $\{\lambda_k\} \in L^p(\mathbb{N} \times \mathbb{R})$, a Banach space.
2. The norm measures *how much homology* there is (quantifies long and many barcodes).
3. The distance compares shapes of point clouds.

Consider the following simplicial complex



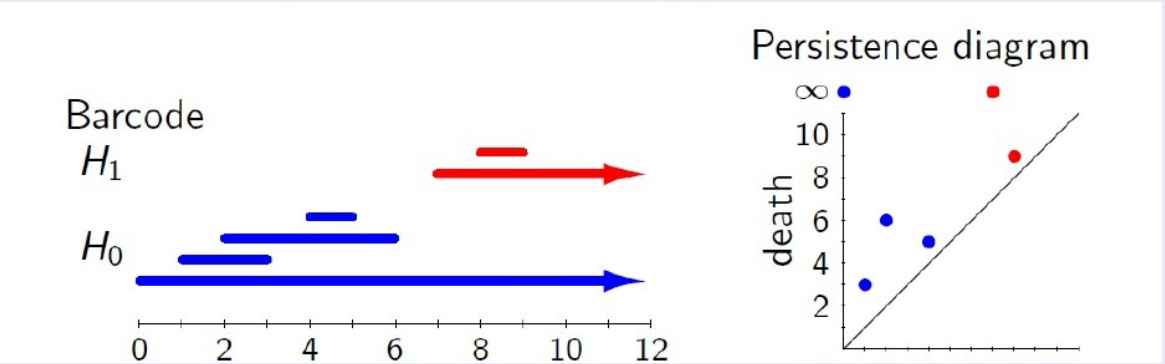
Time	0	1	2	3	4	5	6	7	8	9
Betti number	β_0	β_0	β_0	β_0	β_0	β_0	β_0	β_1	β_1	β_1
effect	$+$	$+$	$+$	$-$	$+$	$-$	$-$	$+$	$+$	$-$

Birth–Death pairs for H_0: $(0,\infty)$, $(1,3)$, $(2,6)$, $(4,5)$

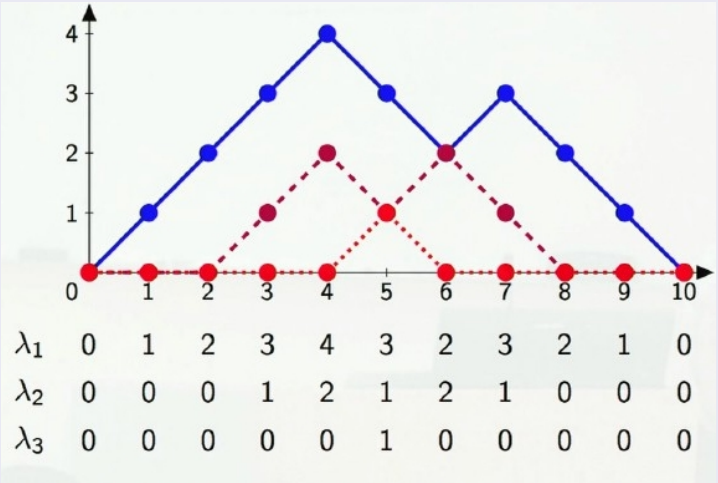
Birth–Death pairs for H_1: $(7,\infty)$, $(8,9)$

Example

The barcode



The 0-landscape



Persistence Entropy

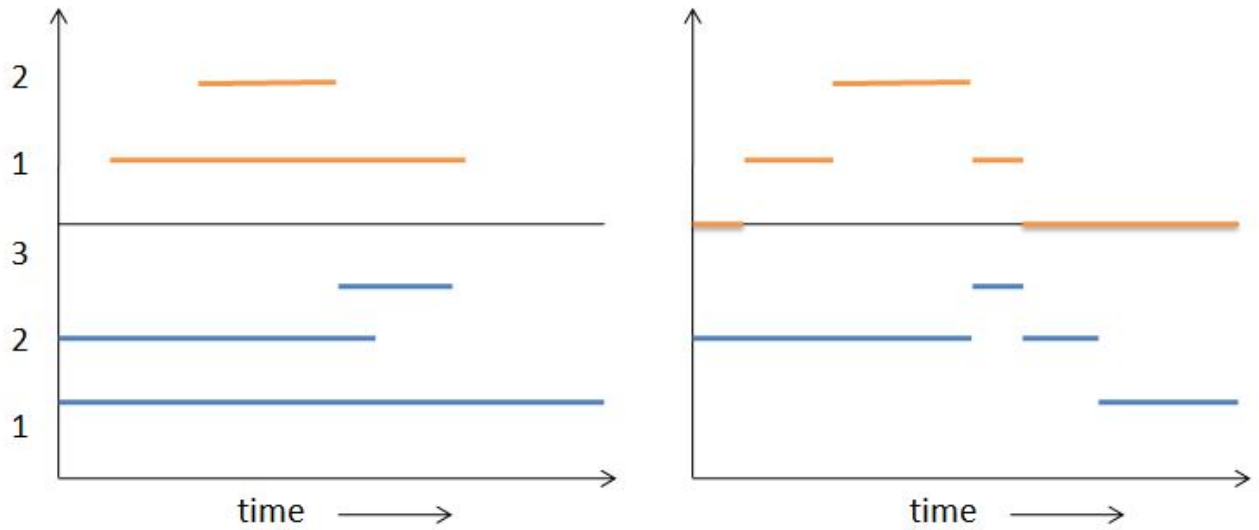
Let the persistence diagram be represented by $D = \{(b_j, d_j)\}_{j \in I}$, where I is the set of all points. The length of each bar is $l_i = d_i - b_i$. Let $L = \sum_i l_i$ denote the total length. Persistent Entropy: The persistent entropy of the barcode is the Shannon entropy of the lengths of the bars.

$$PE(D) = \frac{1}{L} \sum l_i \log(\frac{l_i}{L}) \tag{1}$$

This gives a measure of how similar the length of the barcodes are with the maximum entropy of persistent diagram achieved when all bars are equal.

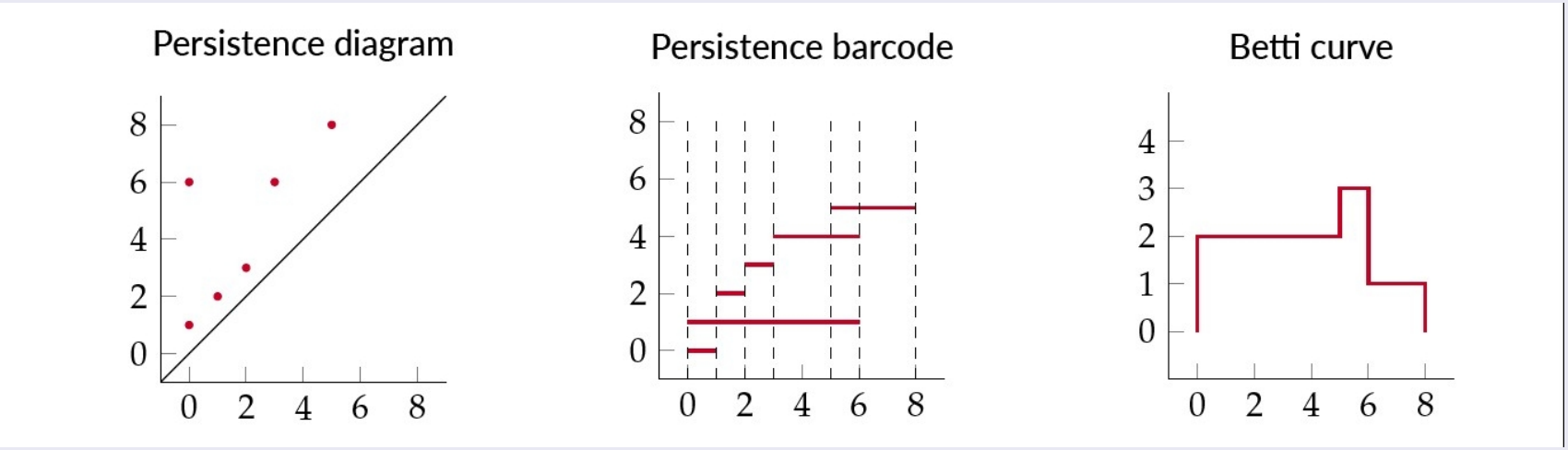
The Betti curve

The Betti curve is a real valued function defined on the set of parameter values. At each point, its value is the number of bars that contain this point. The L^p norm of these curves are considered.



(Left) Persistent Barcode; (Right) Betti Curve

The Betti curve

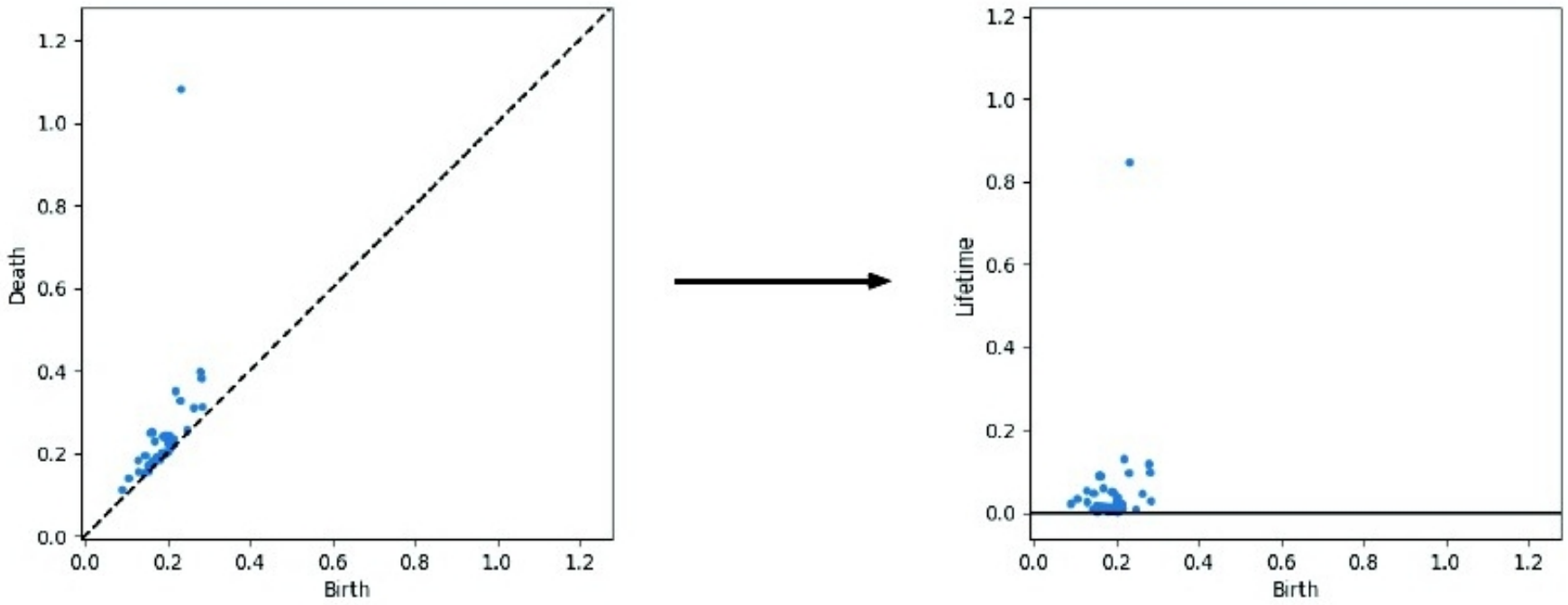


- Easy to calculate.
- Simple representation: a piecewise linear function.

Persistence images

Step 1: The lifetime representation

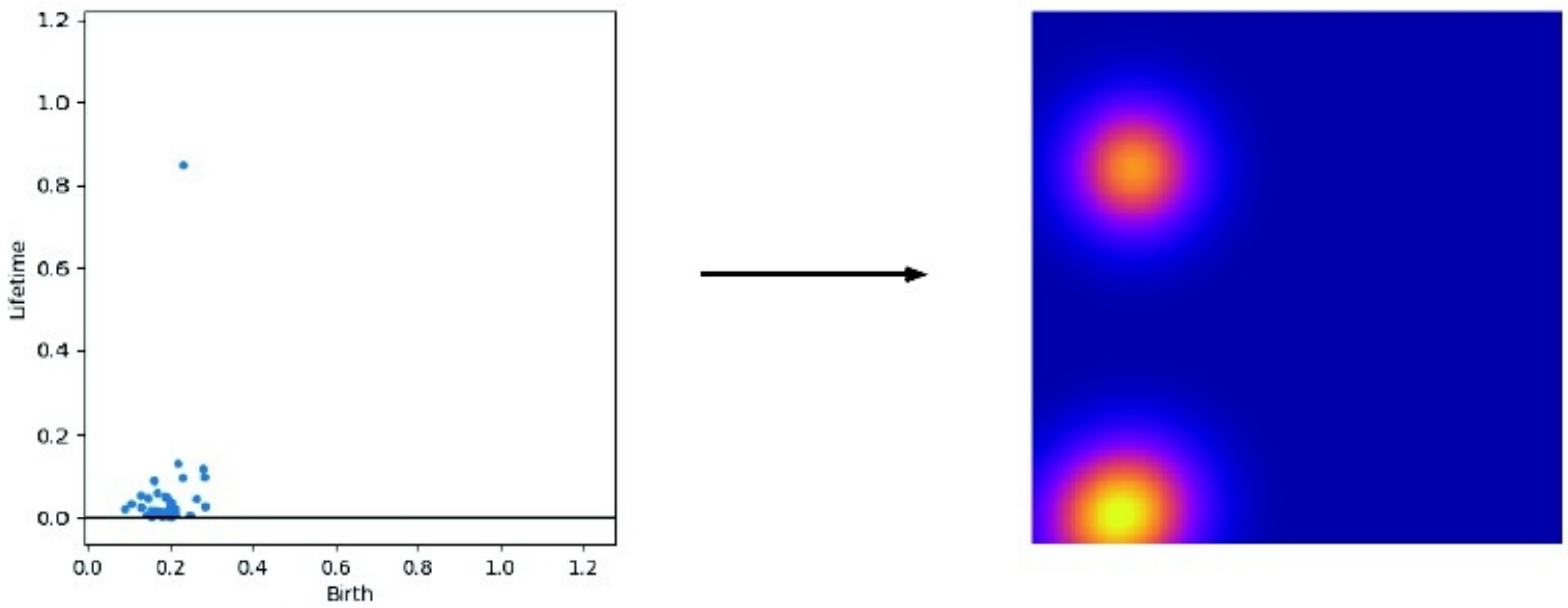
Change the coordinates from $(b, d) \mapsto (b, d - b)$.



Persistence images

Step 2: Heat map

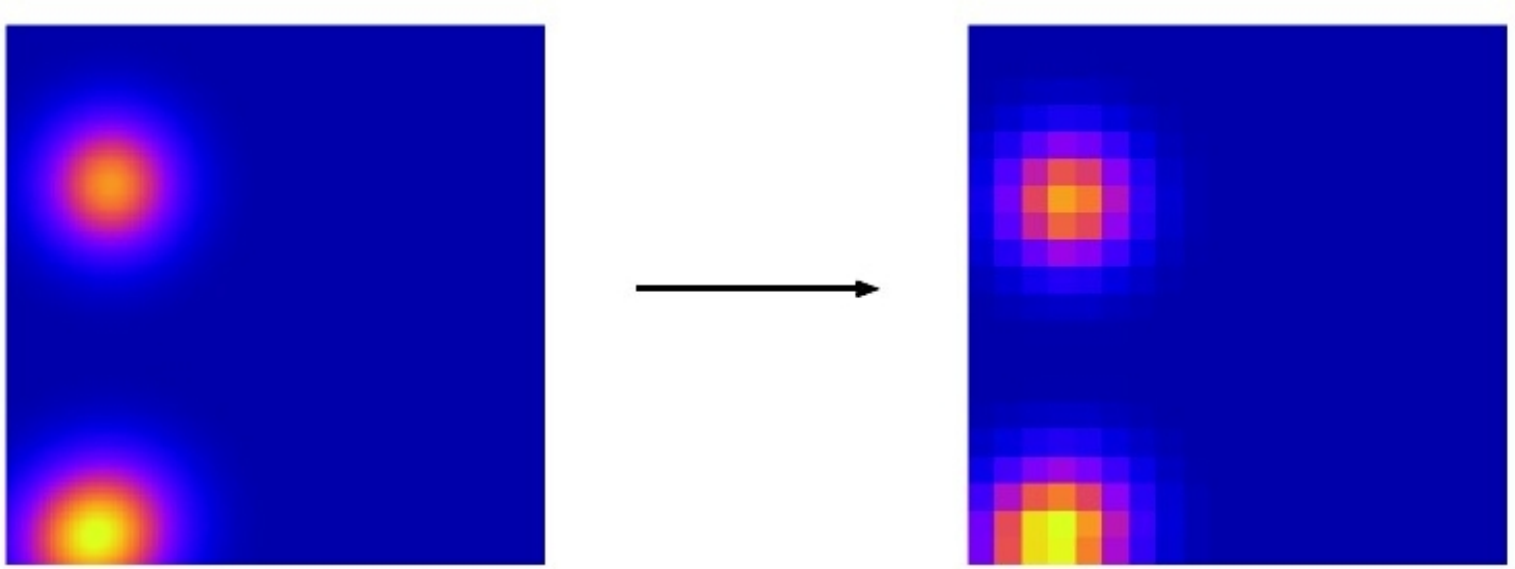
1. Each cycle in the PD is the center of a symmetric Gaussian.
2. Sum the Gaussians to get a real-valued function.
3. Multiply by a weight function, say $w(x, y) = y$.



Persistence images

Step 3: vectorization

Pixelate the image: slice the domain into a grid, then take the average of the function over each square.



Other approaches

- Wasserstein amplitude of order p is the L_p norm of the vector of point distances to the diagonal.
- A vector obtained by rearranging the entries of the distance matrix between points in a PD.
- A vector obtained by superimposing a grid over PD and counting the number of points in each bin.
- First produce a surface from a PD by taking sum of a positive Gaussian centered at each point together with negative Gaussian centered on its reflection below the diagonal.

Stats for non-vector data

- Let Ω be a data set from which certain finite data points are obtained.
- To calculate statistical summaries, the set Ω is desired to have structures of addition, scalar multiplication and even inner product.
- The space of PDs is not an inner product space.
- If we can define a 'nice' map

$$\phi : \Omega \to \mathcal{H}$$

where $\mathcal{H}$ is a Hilbert space then we can calculate summaries and ML models from the inner product

$$< \phi(x_i), \phi(x_j) > .$$

The kernel method: basics

Definition

Let Ω be a set, a function $k : \Omega \times \Omega \to \mathbb{R}$ is called a positive definite kernel if:

- $k(x, y) = k(y, x)$,
- for any $x_1, \ldots, x_n \in \Omega$, the matrix (called the Gram matrix) $[k(x_i, x_j)]$ is positive semi-definite.

Example

Let $\Omega = \mathbb{R}^n$:

- Linear kernel: $< x, y >$.
- Polynomial kernel: $(< x, y > +c)^n$.
- Gaussian kernel: $e^{-\frac{||x-y||^2}{2\sigma^2}}$.

The kernel method

Theorem (Reproducing kernel Hilbert space)

A p.s-d. kernel uniquely defines a Hilbert space $\mathcal{H}$ satisfying

- *for any $x \in \Omega$ the function $k(\cdot, x) : \Omega \to \mathbb{R}$ is an element of $\mathcal{H}$,*
- *the span of $\{k(\cdot, x) : x \in \Omega\}$ is dense,*
- *for $x \in \Omega$ and $f \in \mathcal{H}$, $< f, k(\cdot, x) >= f(x)$.*

Given a data set Ω and a kernel k use the Gram matrix to construct the corresponding RKHS. If k has additional differentiable properties then the RKHS embeds in the space of signed Radon measures.
Conclusion: One can talk about probability distributions on Ω.

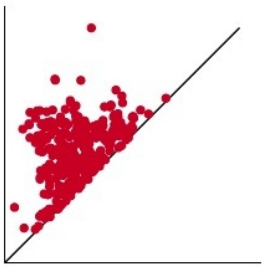
Example

Stable multi-scale kernel of Reininghaus et al.

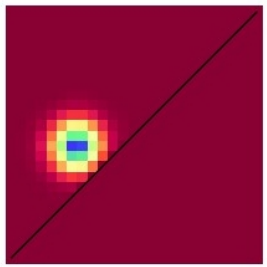
For two PDs B, D we have

$$k(B, D) := \frac{1}{8\pi\sigma} \sum_{p \in B, q \in D} \exp\left(\frac{||p-q||^2}{8\sigma}\right) - \exp\left(\frac{||p-\overline{q}||^2}{8\sigma}\right).$$

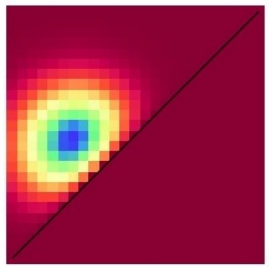
Gaussians of standard deviation σ are placed over every point of B and a -ve Gaussian of σ over the mirror image of the point across the diagonal. The output of this operation is a real-valued function on $\mathbb{R}^2$.



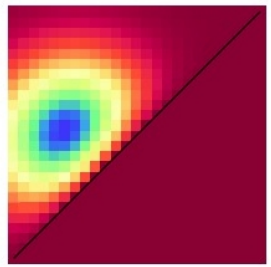
Persistence diagram



$\sigma = 0.1$



$\sigma = 0.5$



$\sigma = 1.0$

Other kernels

- Kernel based on sliced Wasserstein distance by by Carriere et al. (PMRL 2017)
- Kernel embeddings method by Kusano et al. (JMLR 2018)
- Kernel based on Riemannian geometry by Le et al. (ANIPS 2018)
- Kernels on Betti curves by Rieck et al. (arXiv 1907.13496)

Outline

The study of natural images

In the article

G. Carlsson, T. Ishkhanov, V. de Silva, A. Zomorodian, On the Local Behavior of Spaces of Natural Images. Int. J. Comput. Vision 76(1) (2008) 1–12.

Data: high-contrast 3×3 image patches from 4167 outdoor digital photographs; (after suitable normalization) it was spread all over S^7.

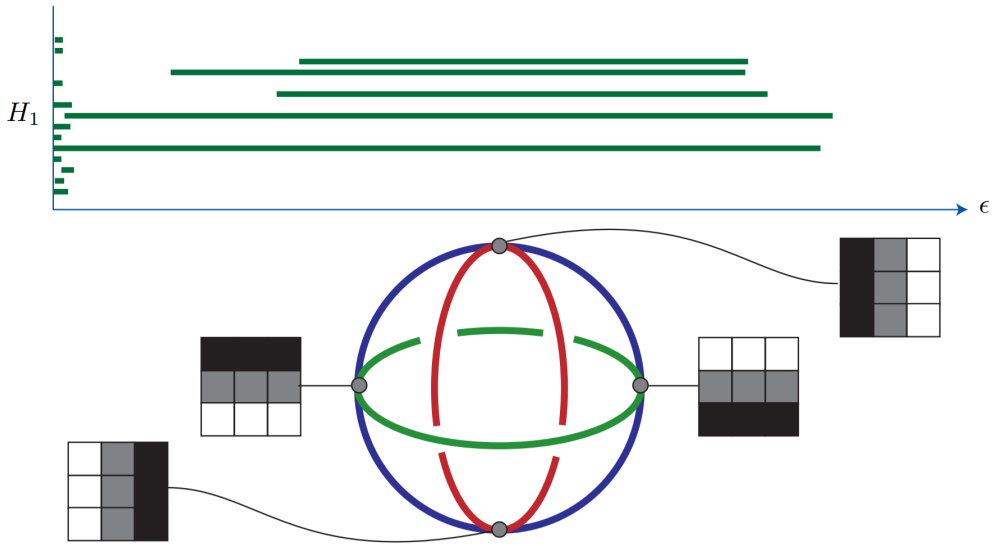
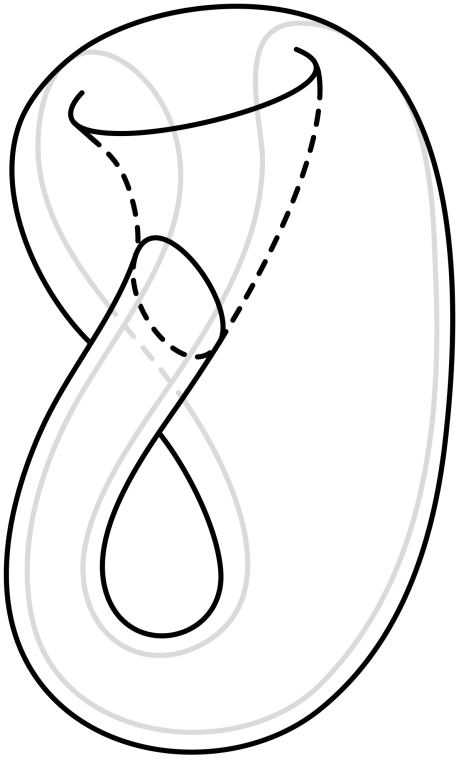
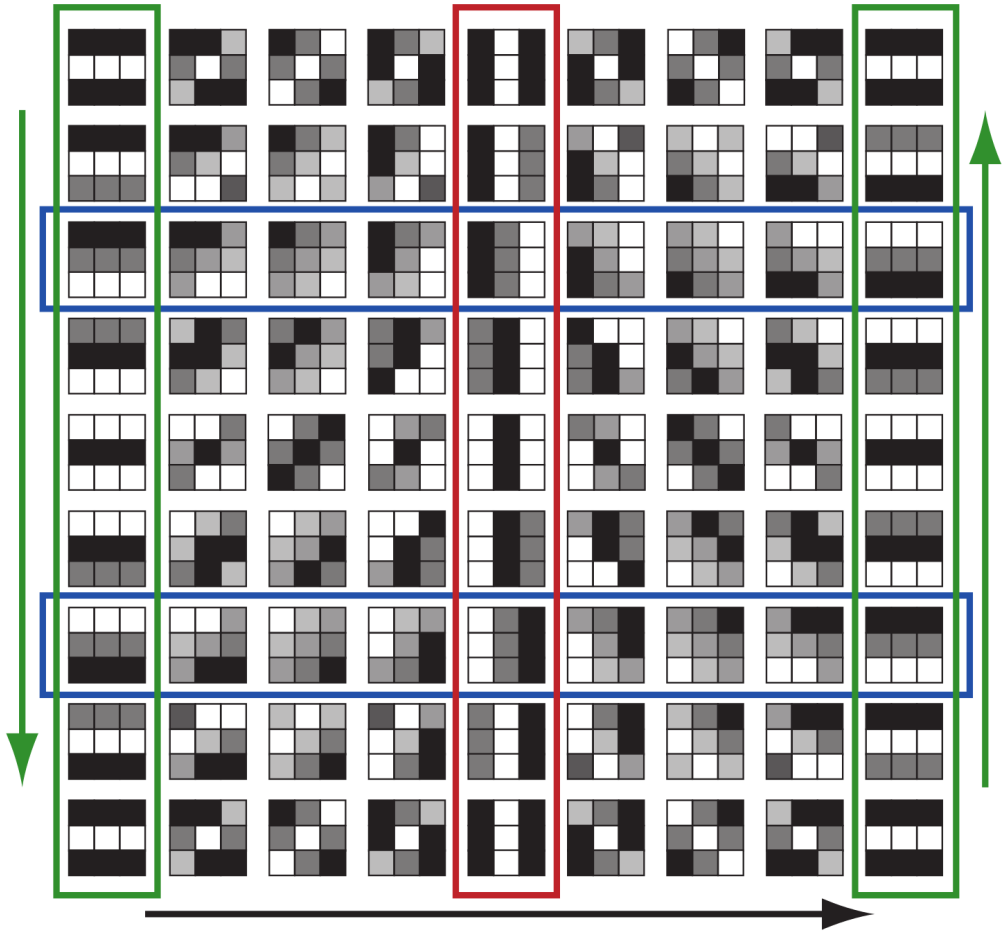


Image credit: Ghrist

The Klein bottle

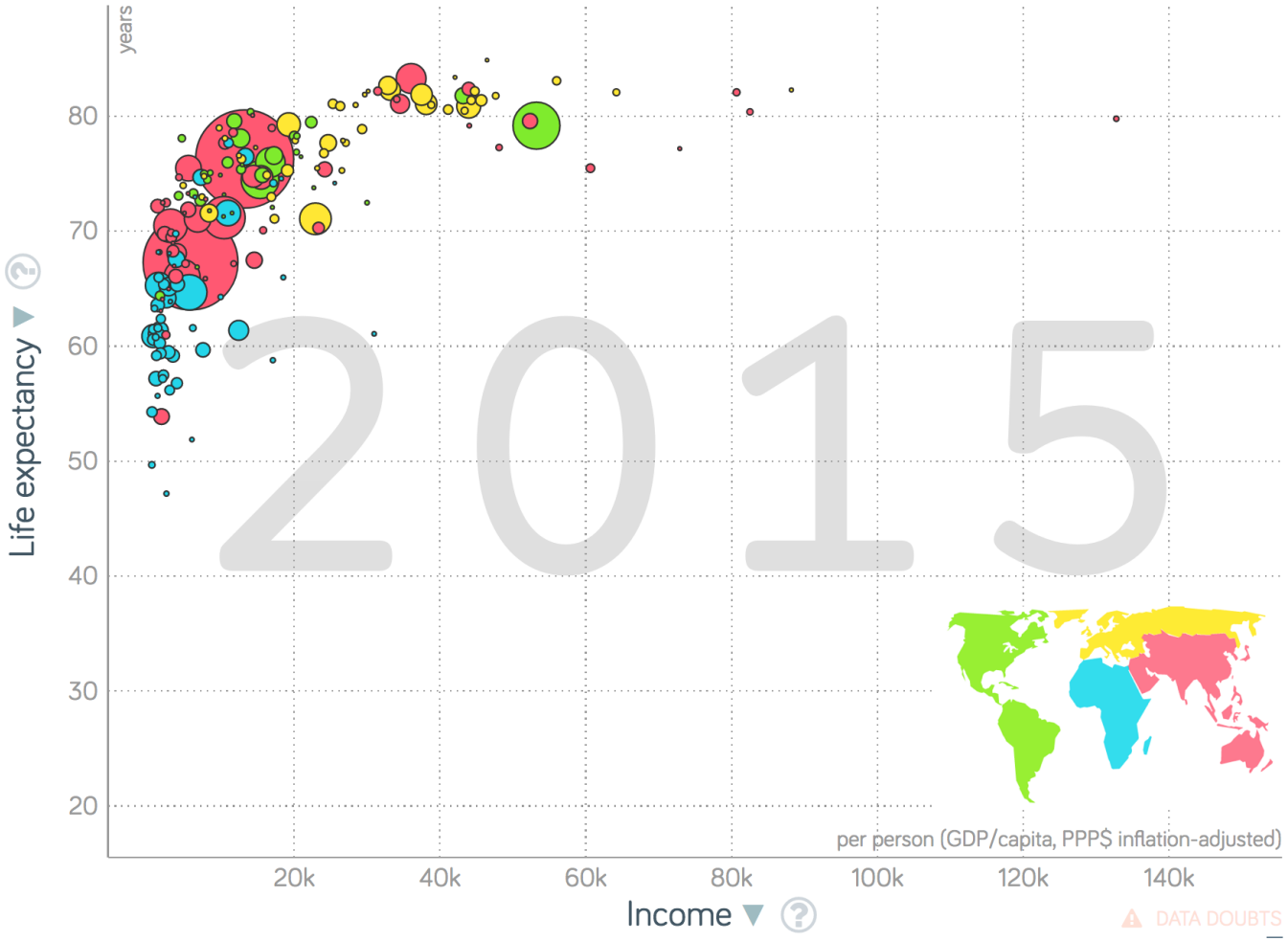
The data revealed a dense region which looks like





Developmental economics

Paper of Banmann and Ziegelmeir on health-wealth index.



Local development cycles

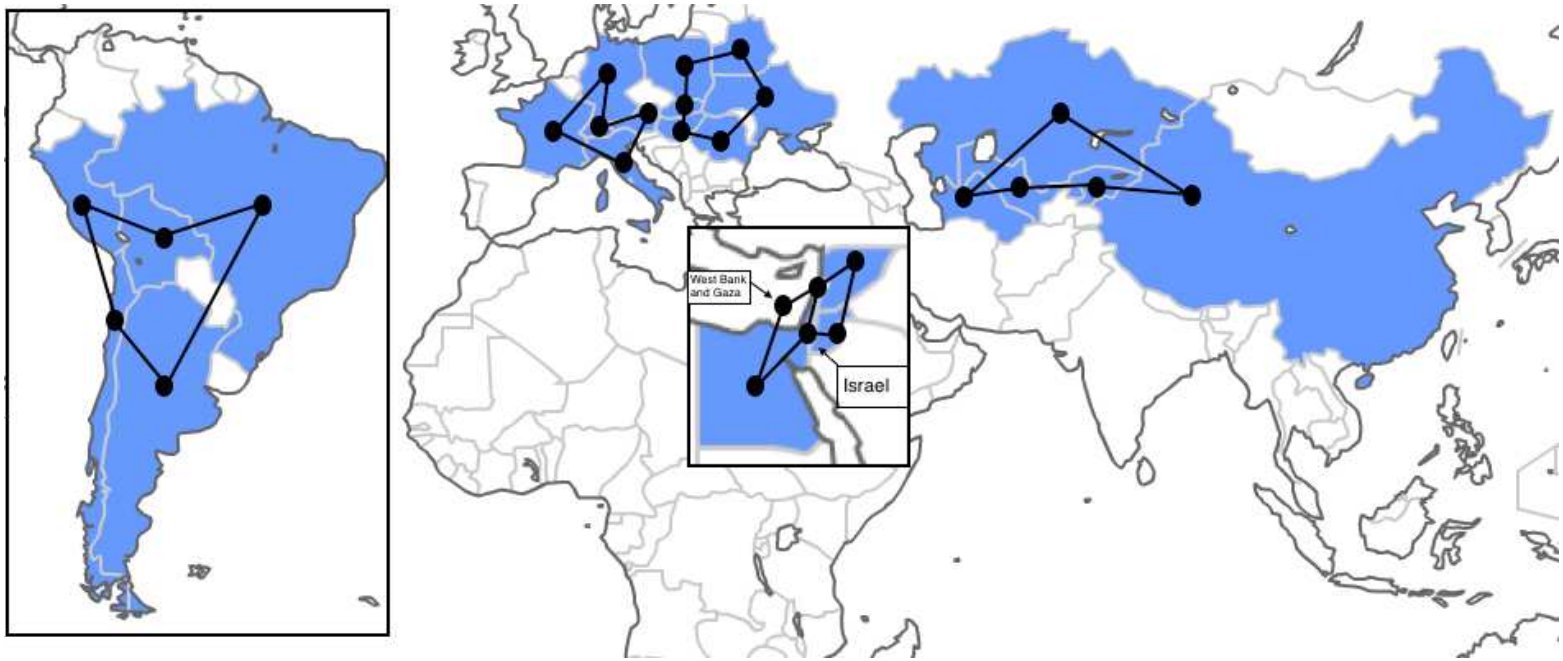


Fig. 6 Map of six cycles in the country border graph with distance d_I, where $I = \{\text{GDP}, \text{LifeExp}\}$, as the edge weight. Software-provided generators for each cycle are shown, and the involved countries are shaded.

Landscape of crashes

The article

M. Gidea, Y. Katz, Topological data analysis of financial time series: landscapes of crashes. Physica A: Statistical Mechanics and its Applications, 491 (2018), 820-834.

1. Consider a d-TS $\{x_n^k\}, 1 \leq k \leq d$. For each time instance t_n we have a point in $\mathbb{R}^d$.
2. Use sliding window technique to construct a PCD of w points $X_n := \{x(t_n), \ldots, x(t_{n+w-1})\}$.
3. TDA is applied on these (X_n) PCDs to study time-varying topological properties of multidimensional time series.

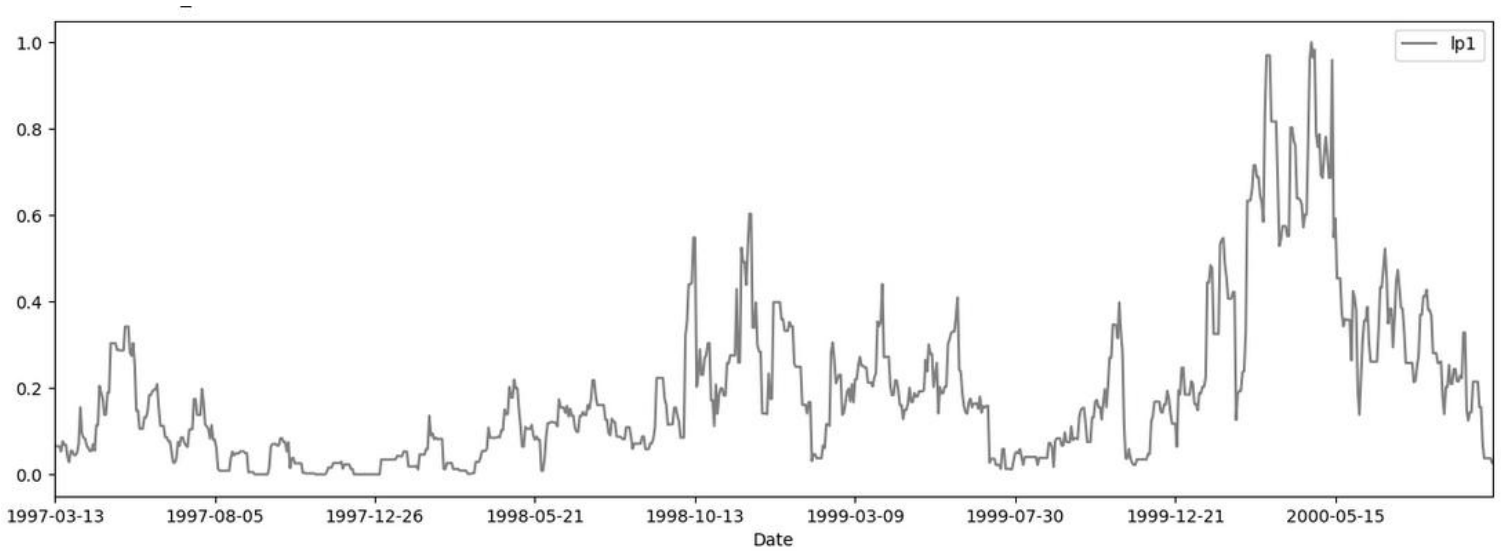
TDA on Time Series

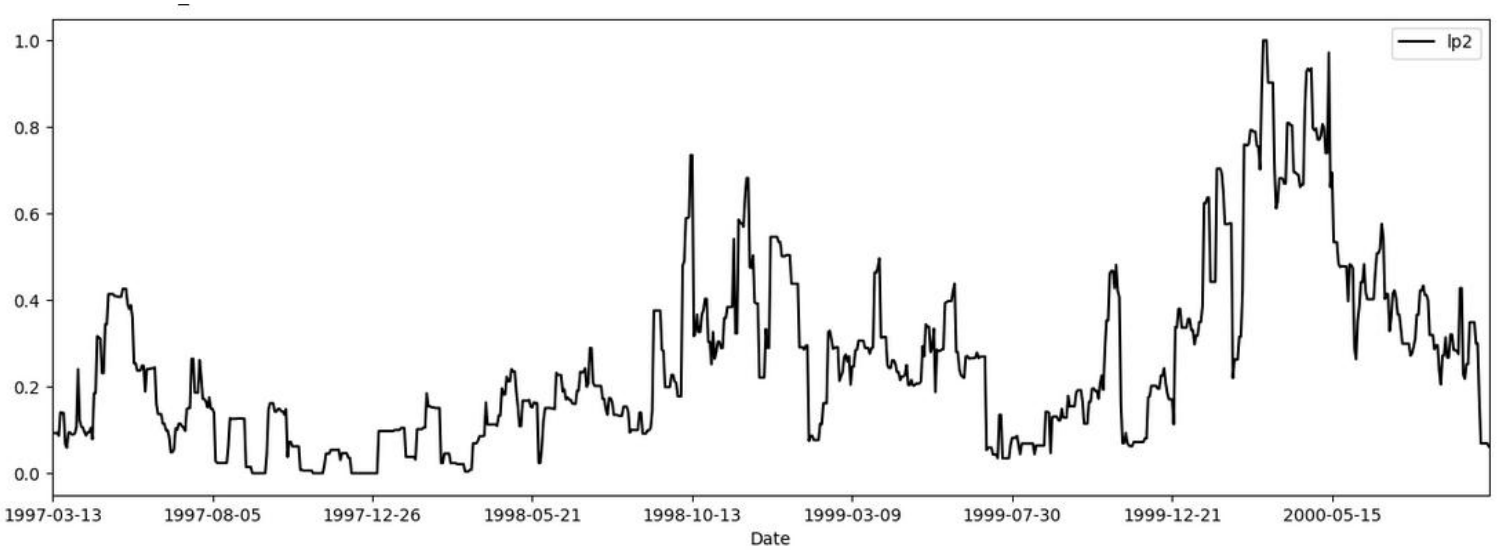
For each point cloud X_n:

1. Compute the persistent homology.
2. Compute the persistence landscapes.
3. Study the L^p norms for $p = 1, 2$.

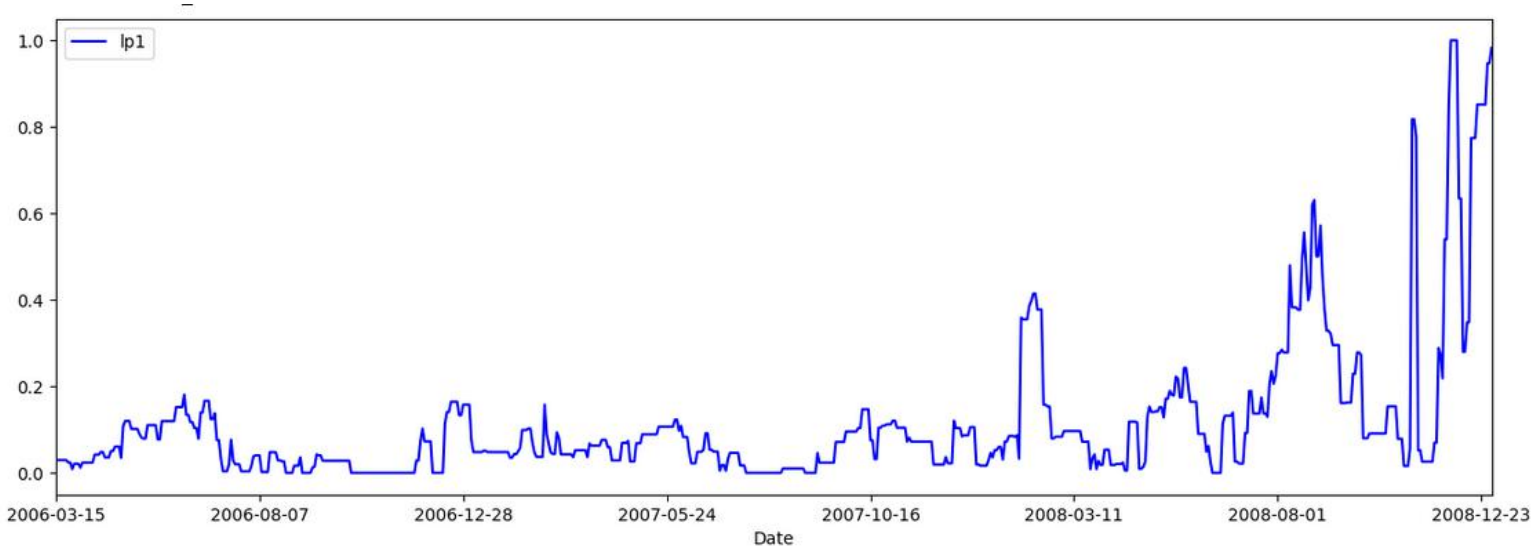
The authors study how the L^p norms behave around market crashes. Conclusion is that the norms are sensitive to transitions in the state of a system from regular to heated.

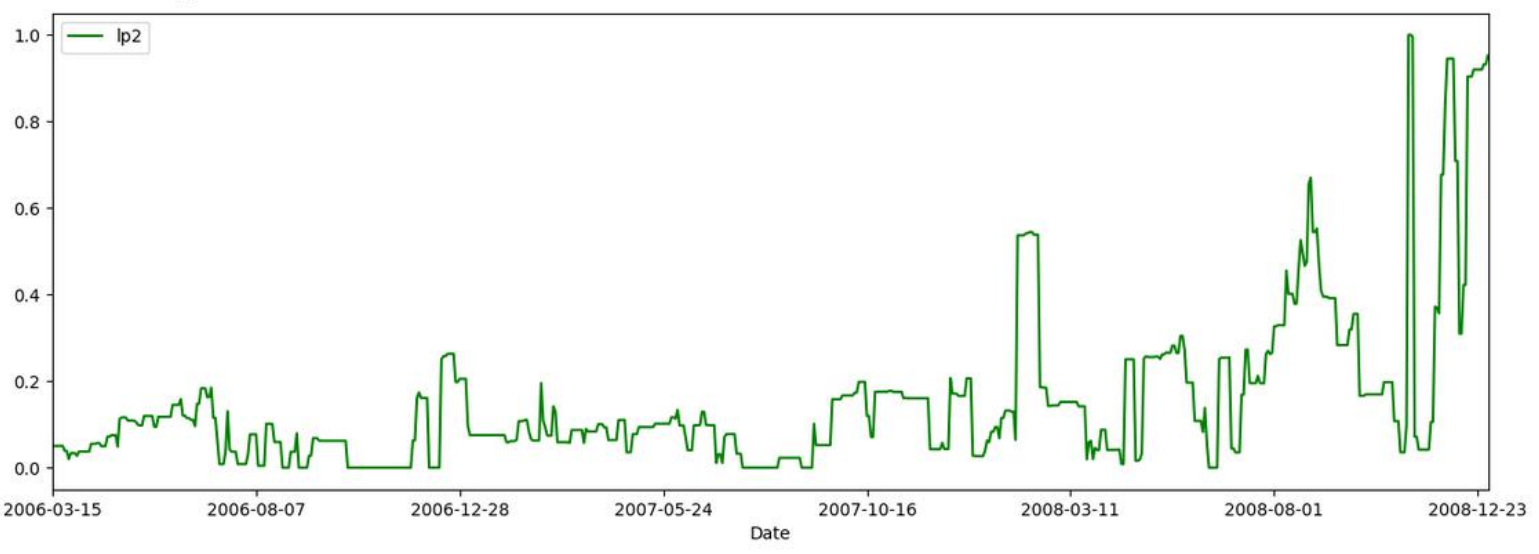
Dotcom crash





Collapse of Lehman Brothers





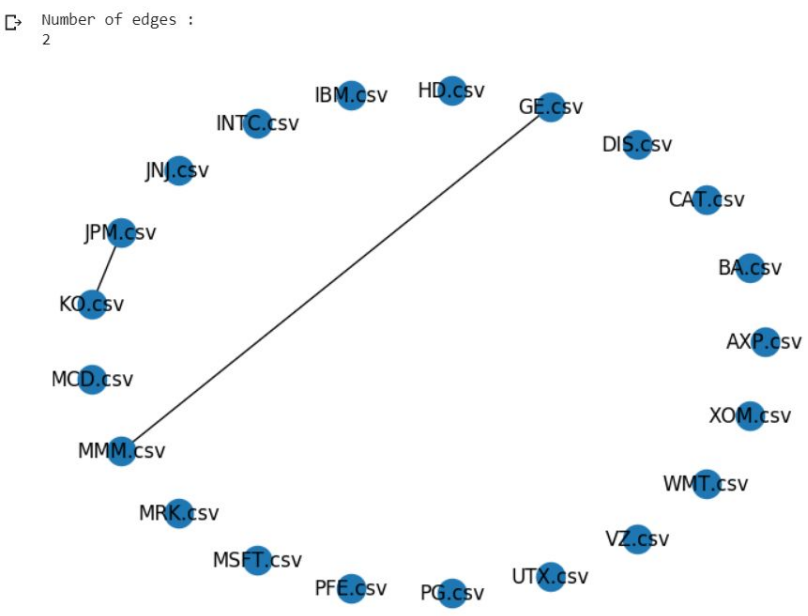
Financial network

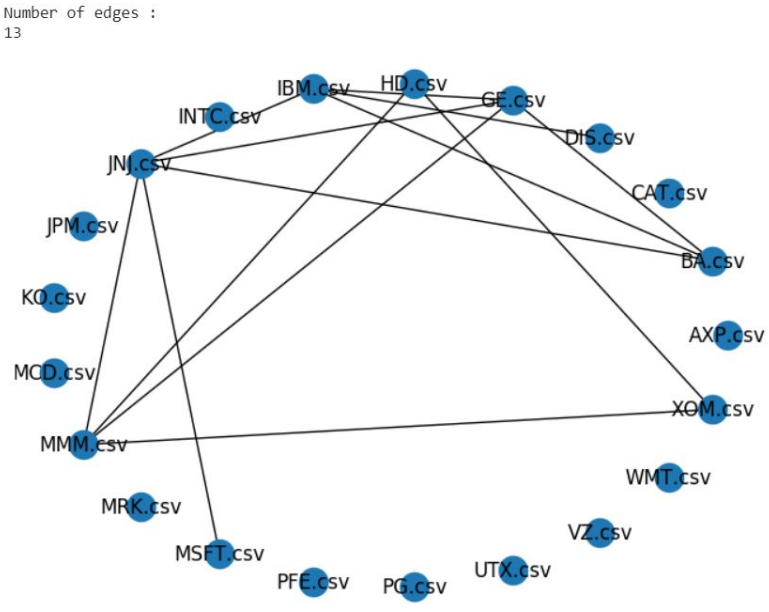
The Article

Gidea M. (2017) Topological Data Analysis of Critical Transitions in Financial Networks. In: 3rd International Winter School and Conference on Network Science. NetSci-X 2017. Springer Proceedings in Complexity. Springer,

The author first creates a time-evolving weighted network based on market data. A filtered simplicial complex is built on top of this. The topology of these complexes is then used to predict early signs of transition.

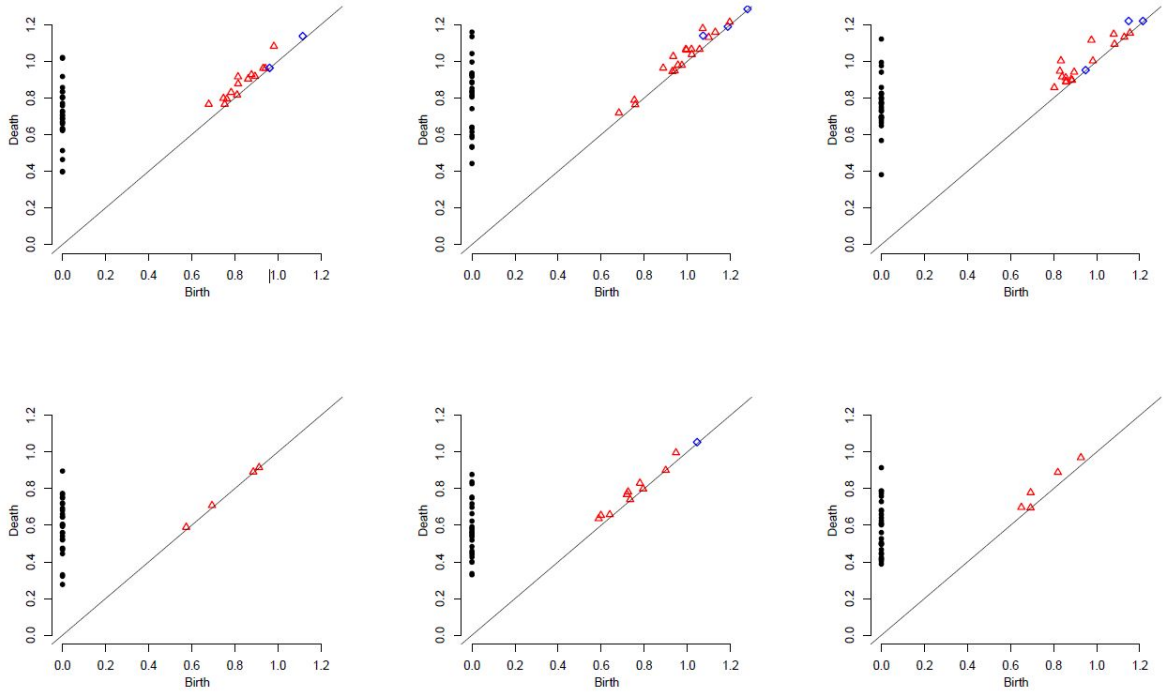
The network





- Nodes are companies.
- This is a complete graph.
- The weight of edge (i, j) is computed using Pearson correlation coefficient.
- Minimum weight 0: perfect correlation; maximum 2 anti-correlation.

The persistence diagrams



Conclusion

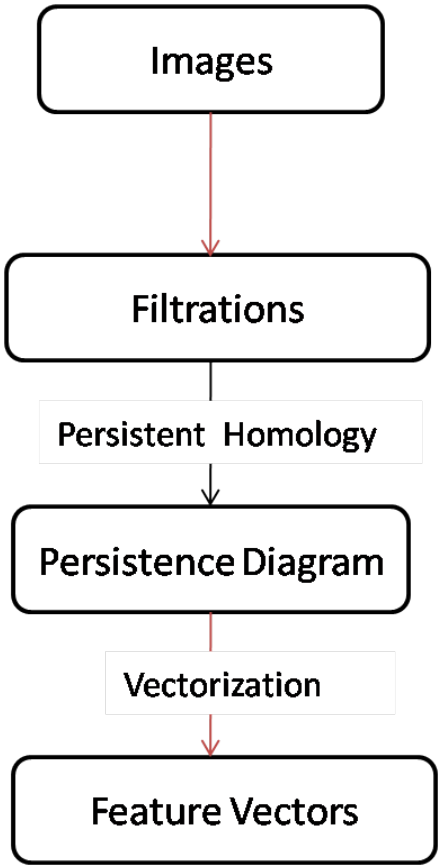
Before a potential crash there are many short lived cycles.

Image Classification using PH

In image classification problems, we can often identify distinct shape features that characterize images in each class. By using tools from TDA, we can classify images based on these properties.

The aim of this project is to develop a 'topological pipeline' to extract various topological descriptors from images which the serve as a basis for classification.

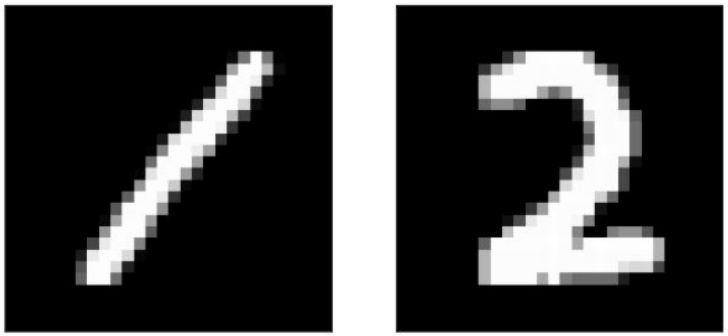
Topological Pipeline: Outline



Two key steps in the pipeline are:

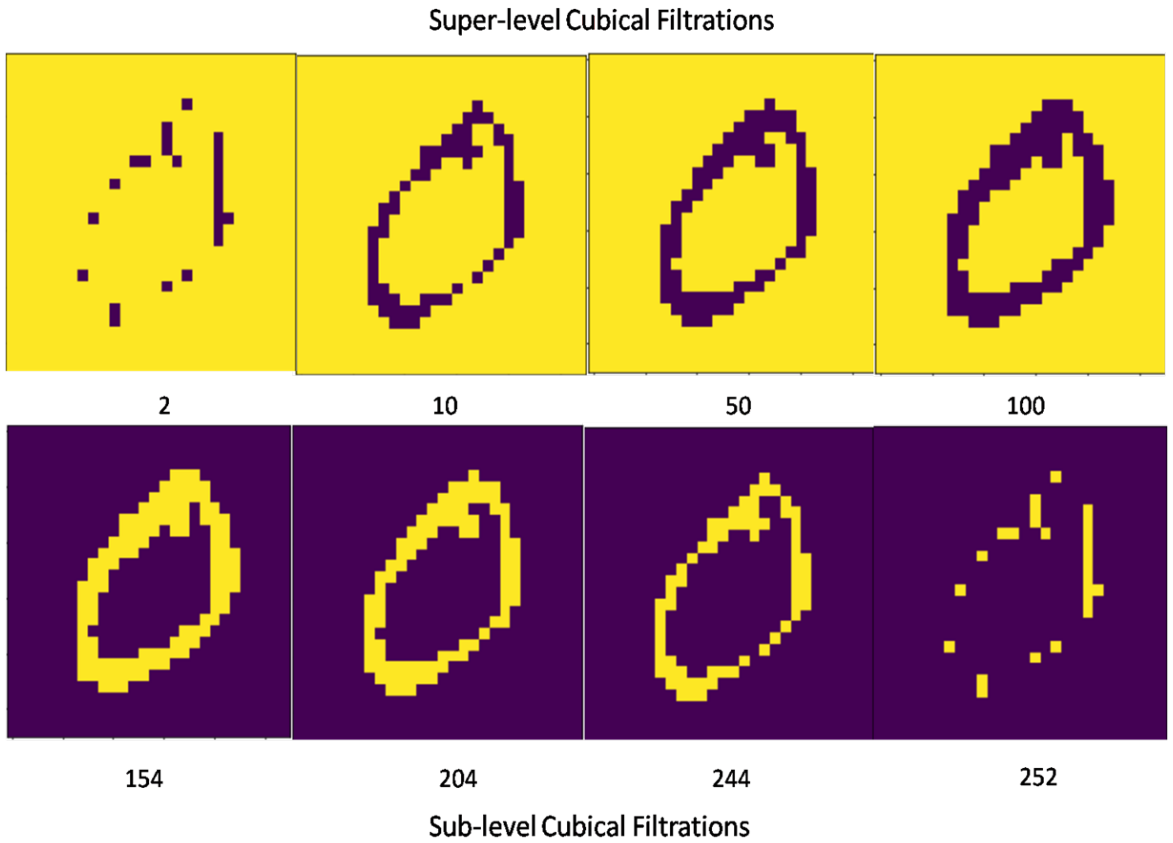
- Constructing filtrations from the image.
- 'Vectorization' of the persistent diagram.

The MNIST dataset of hand written digits will be used to expand on these steps.
This consists of 70000 grayscale images of dimension 28×28 pixels.



Topological Pipeline : Filtrations

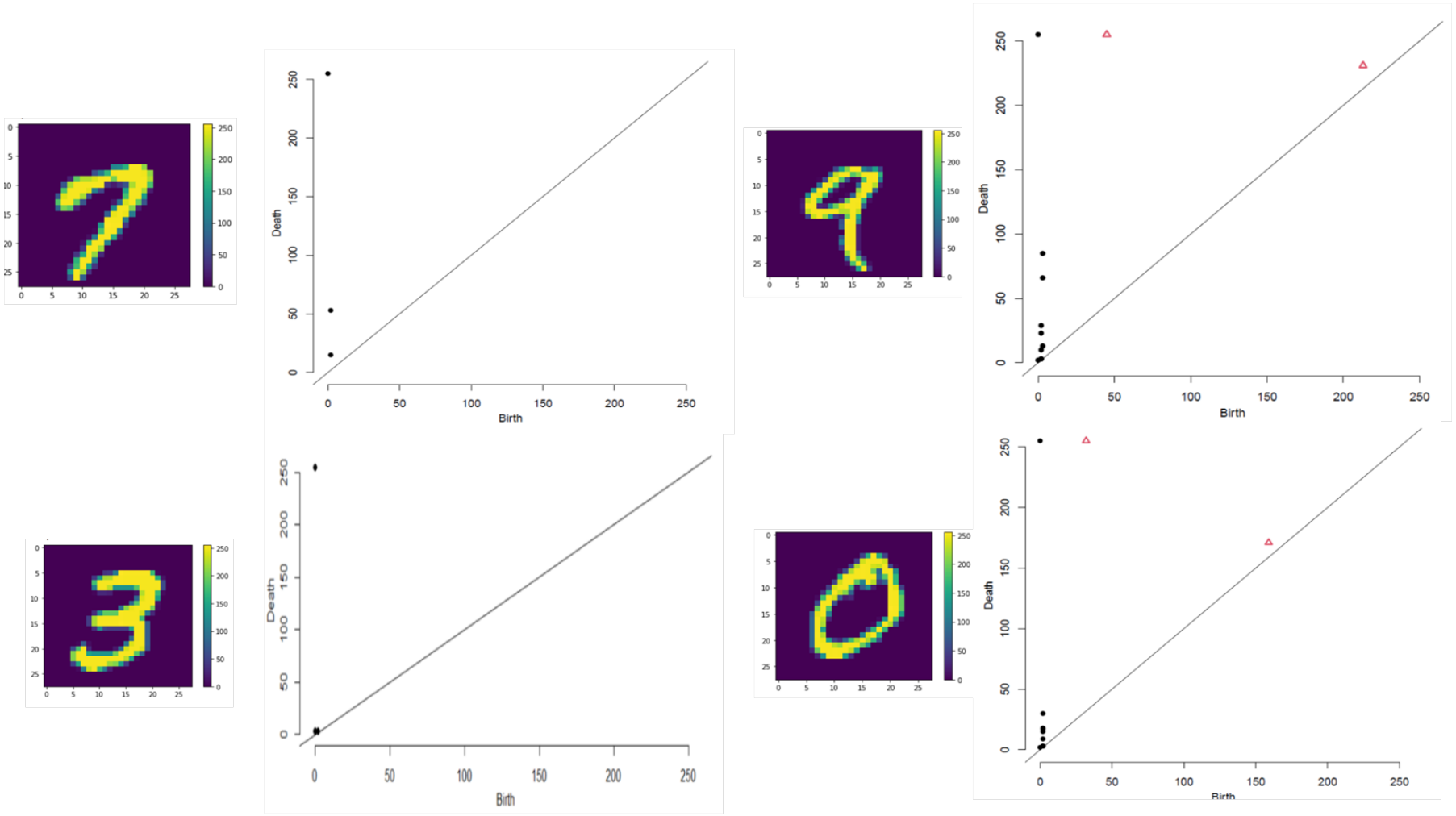
Grayscale images have a natural pixel structure and can be represented by a pixel map on a rectangular grid. As a result, the sublevel and superlevel sets of this map can be interpreted as a cubical filtration.



In the image above, cubical complexes are colored purple.

Topological Pipeline : Filtrations

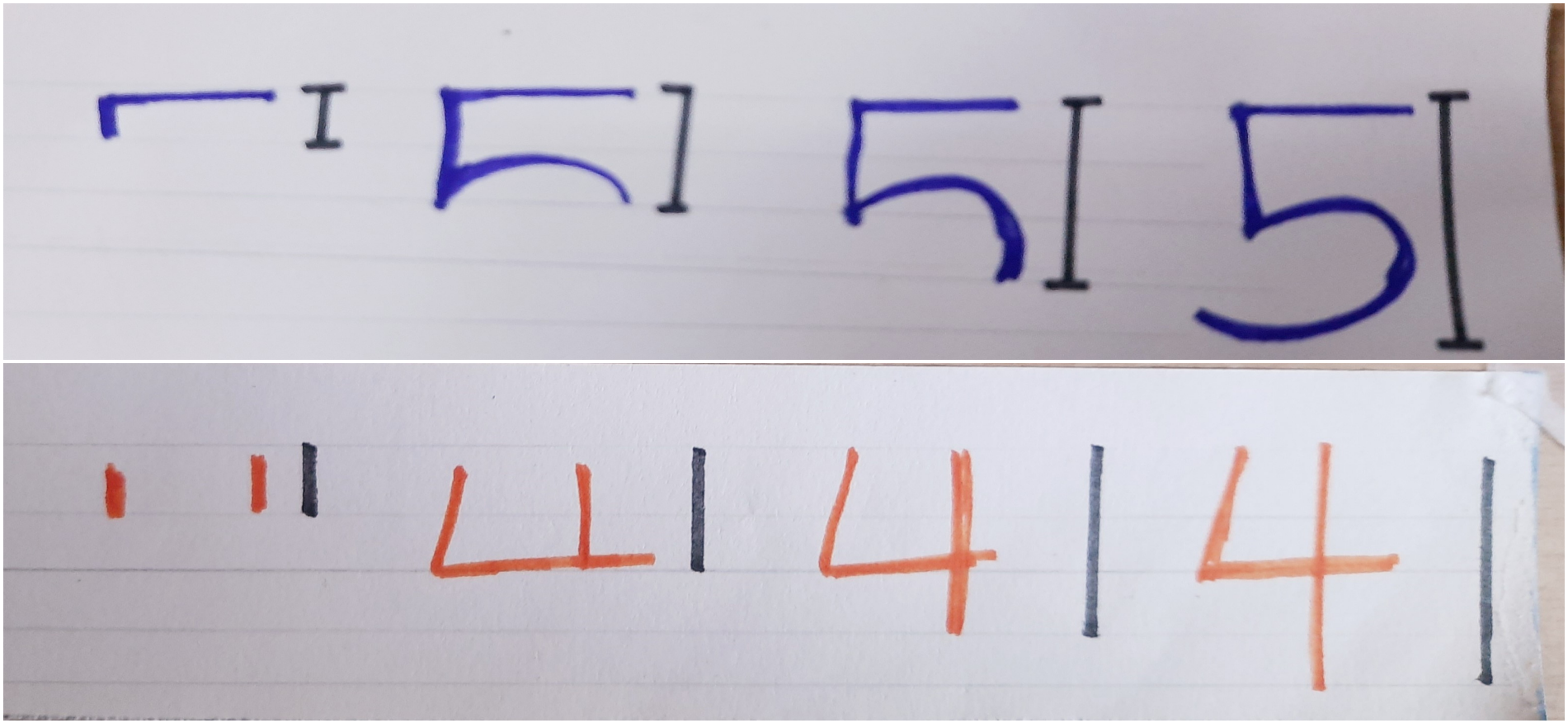
Grayscale filtration alone does not help in distinguishing between digits in the same homotopy class.



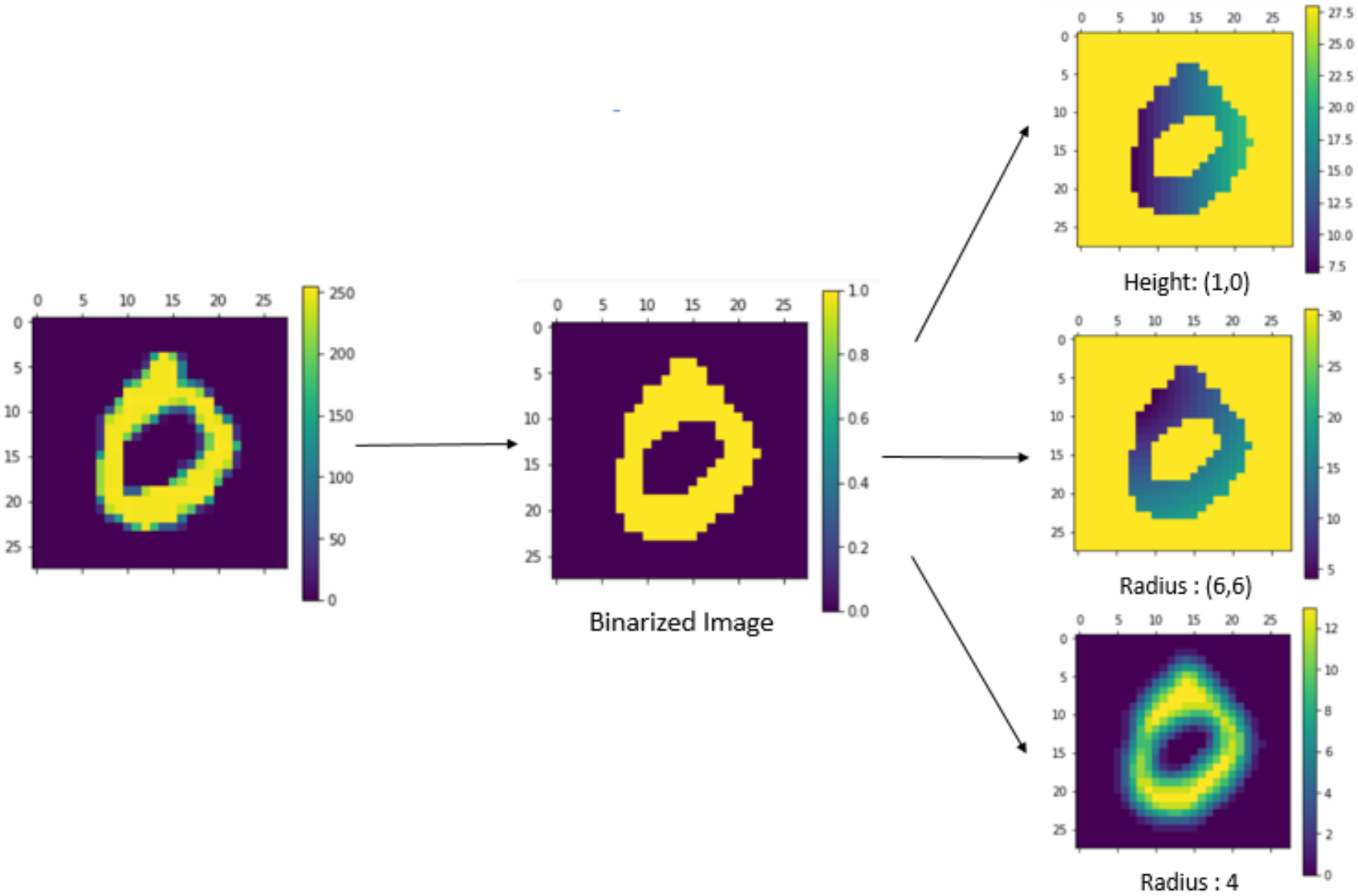
Persistence diagrams corresponding to super-level set cubical complexes.

Topological Pipeline : Filtrations

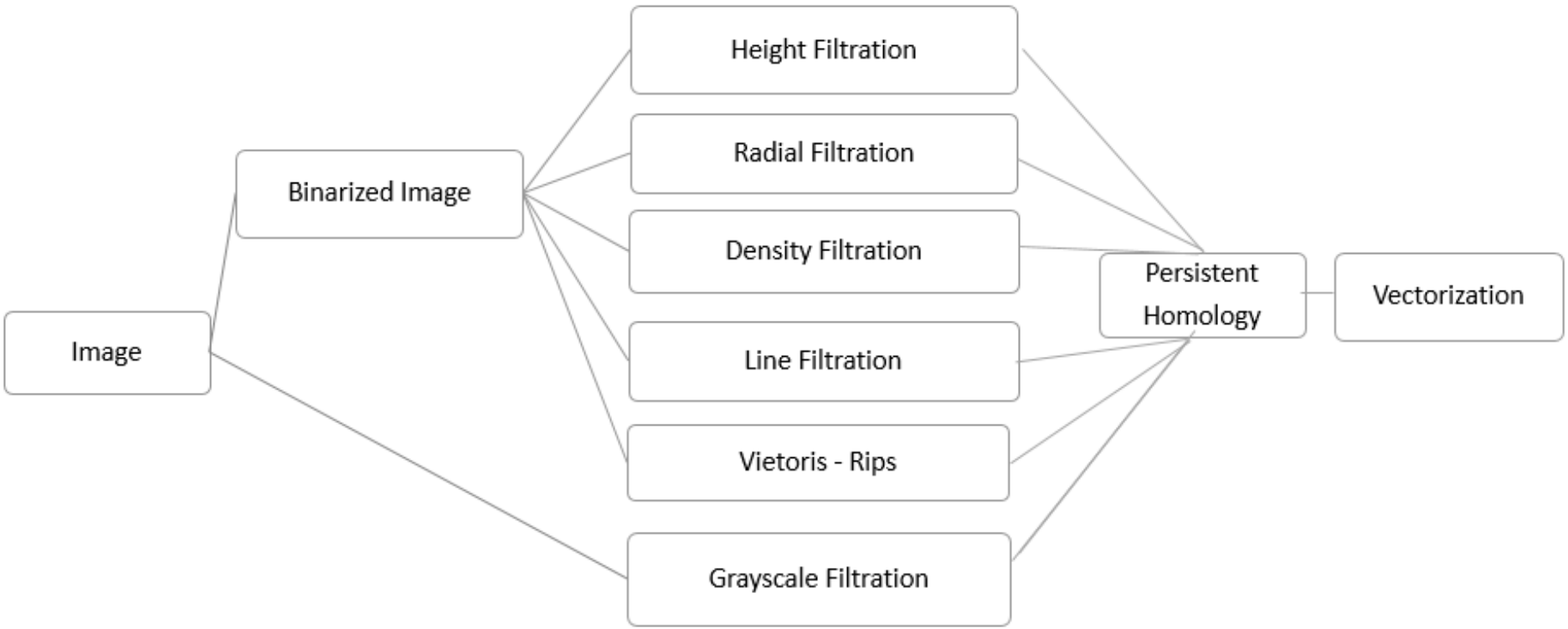
Other filtrations which take into account how the digits are built help distinguish between digits of the same homotopy type.



Topological Pipeline: Filtrations

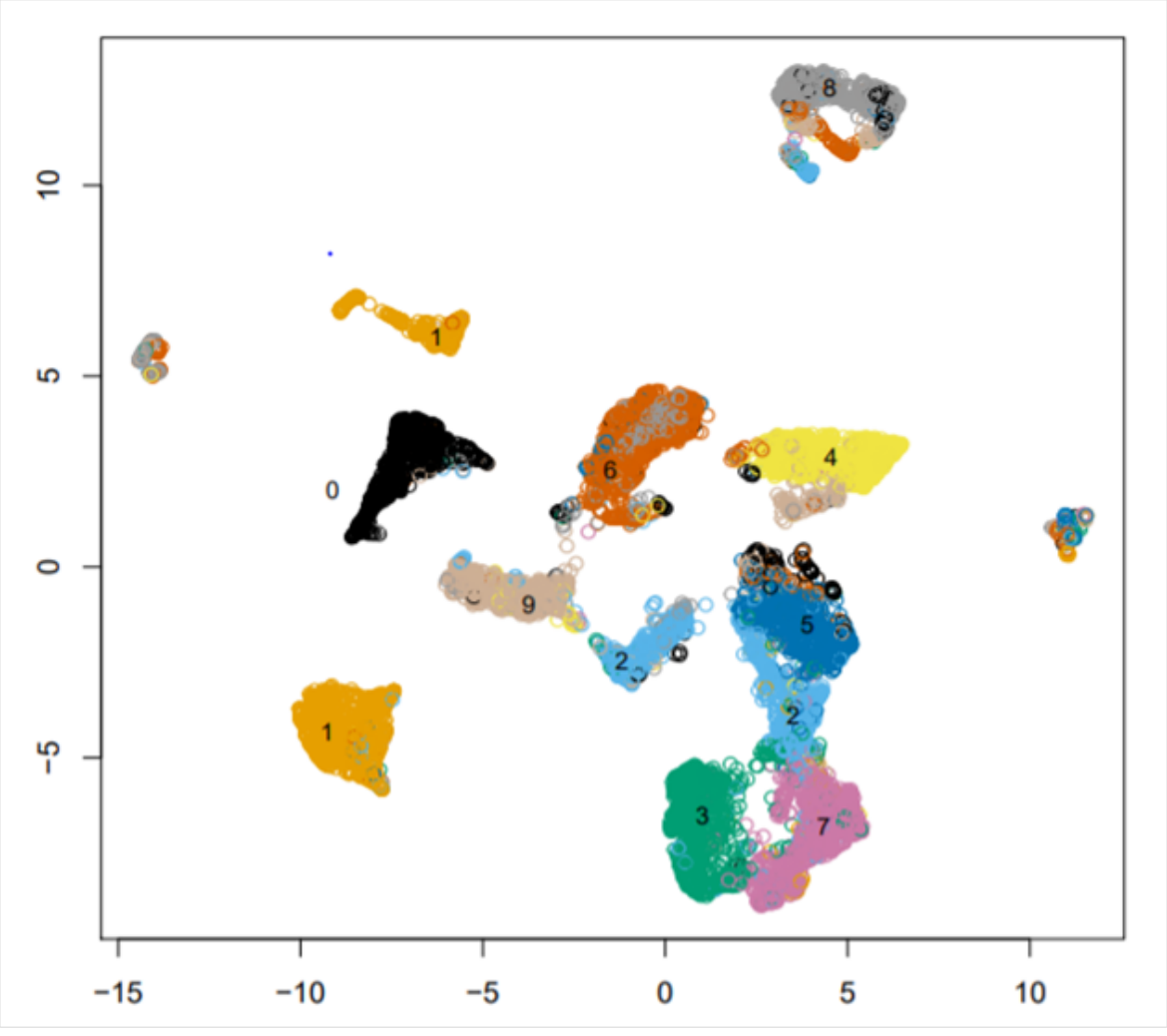


Topological Pipeline



Visualizing the Pipeline Data

2 dimensional projection of 52 dimensional pipeline data (persistent entropy vectorisation) using UMAP.



Visualizing the Pipeline Data

The two clusters each of digits '1' and '2' represent different ways in which these digit are written.

Show the interactive Bokeh plot.

Classification of MNIST

Classification using Random Forest Classifier: Number of trees = 1000

Dimension	Description	Accuracy
703	Reference classifier	96.3%
52	MNIST Pipeline with persistent entropy	96.48%
202	MNIST Pipeline using all 4 vectorization	97.16%

For reference classifier, all the pixel values except those that are 0 for all images were considered as a vector.

S.No	Binarisation	Dimension	Filtrations	Vectorisation	Accuracy
1	0.2	50	Height, Radial, Density, Line	Persistent Landscape	94.92
2	0.4	52	Height, Radial, Density, Line, V-R	Entropy	96.15
3	0.3	52	Height, Radial, Density, Line, V-R	Entropy	96.21
4	0.2	52	Height, Radial, Density, Line, V-R	Entropy	96.48
5	0.2	202	Height, Radial, Density, Line, V-R	All Vectorisations	97.16

Outline

Things I've missed

1. The manifold hypothesis.
2. The sliding window embedding.
3. UMAP and other non-linear dimensionality reduction techniques.
4. Other invariants as topological signatures.
5. Persistent homotopy groups.

Thank You