Introduction to TDA.

Paweł Dłotko, Dioscuri Centre in TDA, IMPAN.

University of Bremen

19 December 2022.

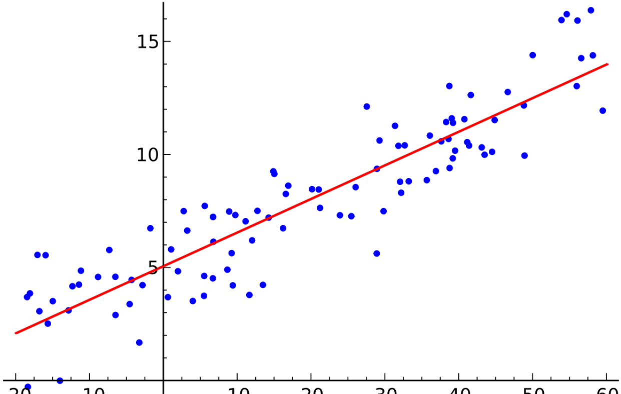
Topological Data Analysis

- Persistent homology,
- Conventional mapper,
- Ball mapper,
- On a very intuitive level,
- With a number of practical examples.

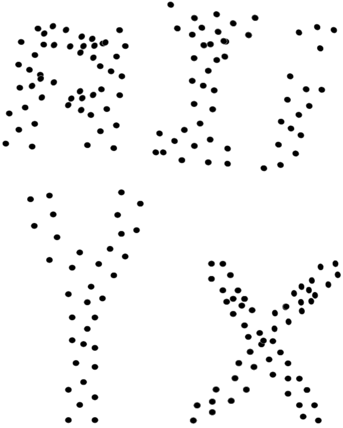
The credo

Data have shape,
shape has meaning,
meaning brings value.

We all know this story

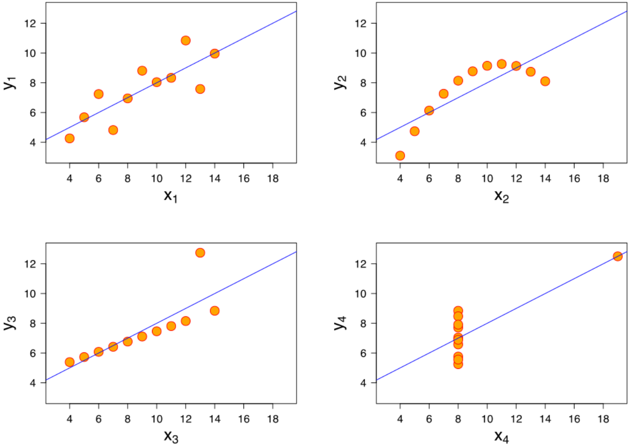


Trap of models



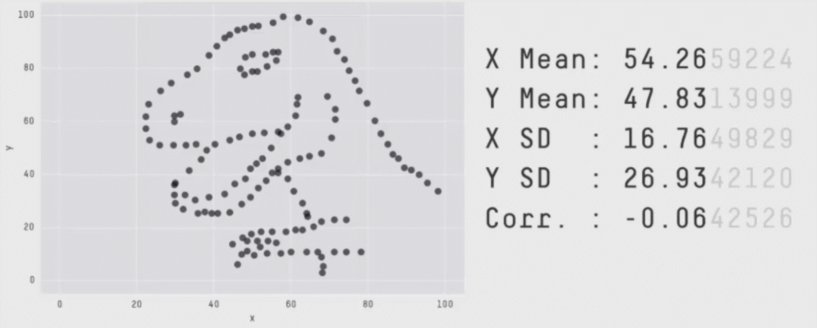
It is not possible to adjust an algebraic model to any possible shape of the data – over-fitting.

Summary statistics do not suffice, always visualize!



Anscombe's Quartet; Same statistics, different shapes

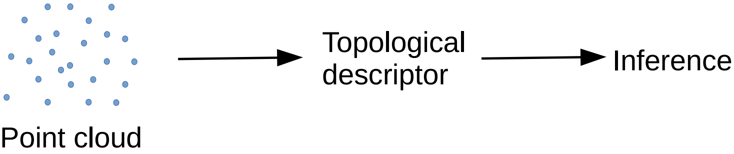
Datasaurus Dataset



Same statistics, different shapes

Shape of data may be important...
But, how to see in high dimensions?

The pipeline



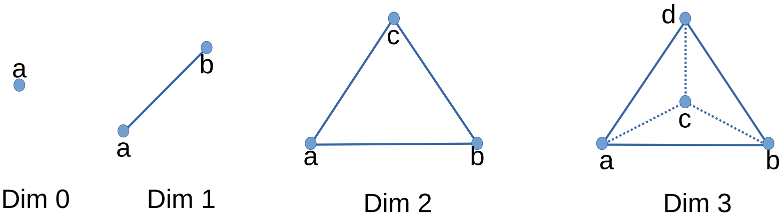
Point cloud

Topological descriptor

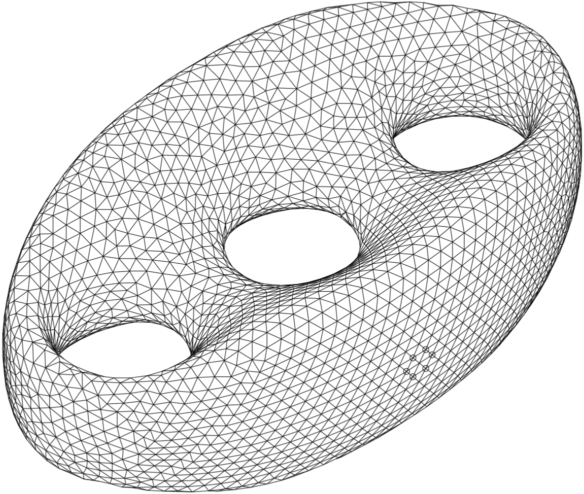
Inference

Simplicial complexes

- $\mathcal{K}$ is an abstract simplicial complex iff for every $A \in \mathcal{K}$ and $B \subset A$, $B \in \mathcal{K}$.
- Each abstract simplicial complex has its geometrical realization built from simplices.
- In this case, simplices consist of points in a general position.

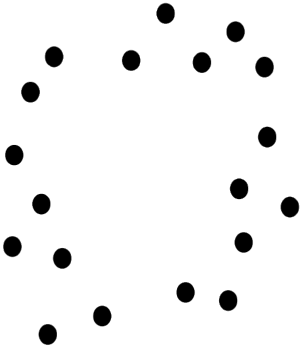


Primary use: FEM



Source: Wikipedia.

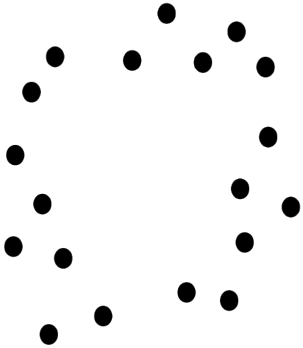
What do you see?



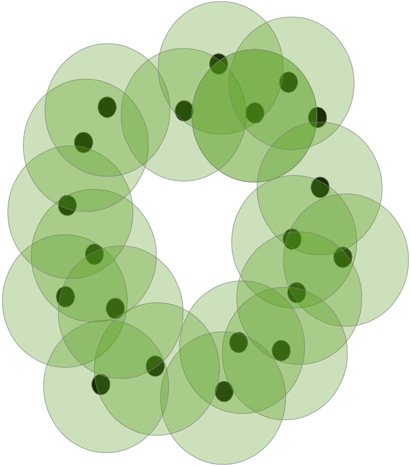
What do you see?

- A circle?
- 19 points...?
- 19 points sampled from a circle?
- Persistent homology is a tool to make this observation bit more precise.
- Filtration – multiscale model of the data.

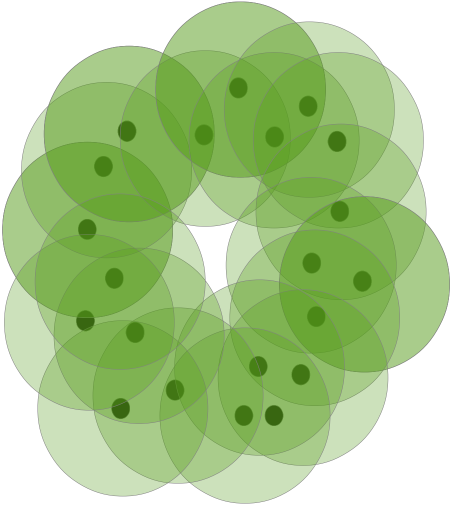
What do you see?



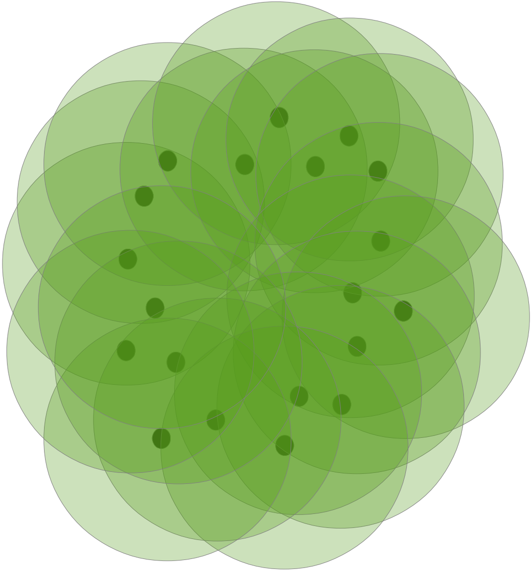
What do you see?



What do you see?



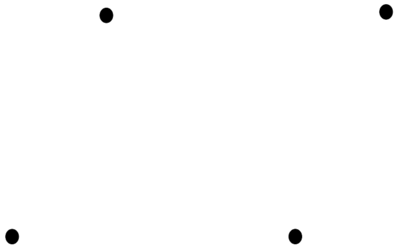
What do you see?



Simplicial complexes from point clouds

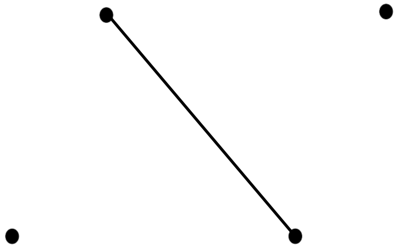
- $P = \{p_1, \ldots, p_n\}$, a finite point cloud with a metric d.
- We need a finite, combinatorial representation of the union of balls.
- Rips complex at level ϵ consists of simplices supported in $p_0, \ldots, p_n$ if $d(p_i, p_j) \leq \epsilon$ for every $i, j \in \{0, \ldots, n\}$.

Filtration of Rips complex



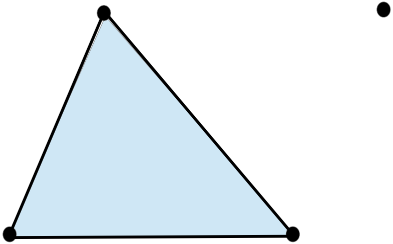
4 vertices

Filtration of Rips complex



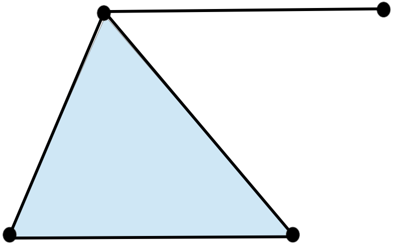
4 vertices, 1 edge

Filtration of Rips complex



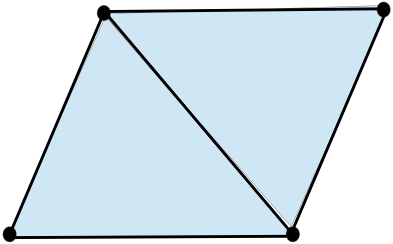
4 vertices, 3 edges, 1 triangle

Filtration of Rips complex



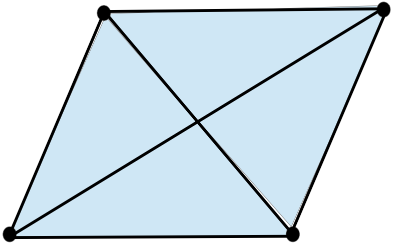
4 vertices, 4 edges, 1 triangle

Filtration of Rips complex



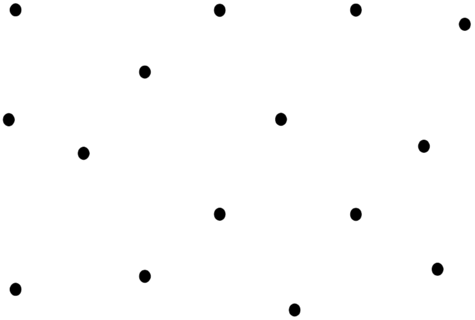
4 vertices, 5 edges, 2 triangles

Filtration of Rips complex



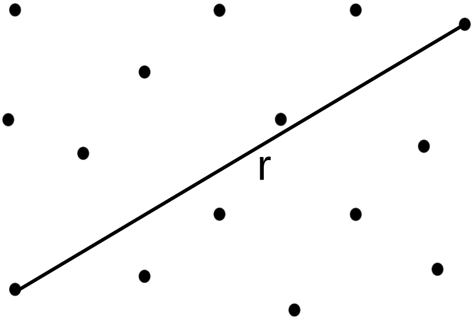
4 vertices, 6 edges, 4 triangles, 1 tetrahedra

Rips complex can grow large



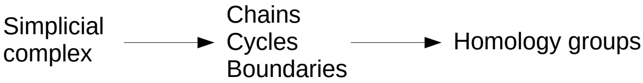
If all points get connected by edges in the complex, we witness so called *combinatorial explosion*. You will encounter it when using Rips complexes.

Rips complexes can grow large



For N points, $\binom{N}{1}$ vertices, $\binom{N}{2}$ edge, $\binom{N}{3}$ triangles, ...
This is why we always limit the level (ϵ) and the maximal dimension of simplices in the complex.

From complexes to parameter dependent homology

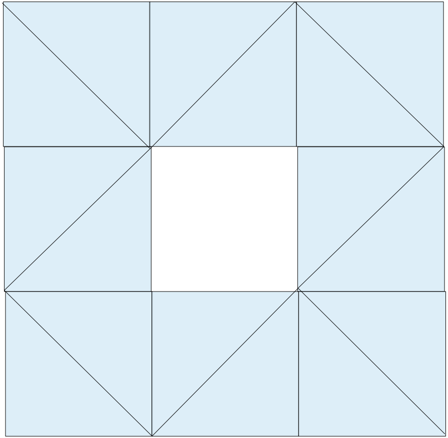


Simplicial complex

Chains
Cycles
Boundaries

Homology groups

Homology

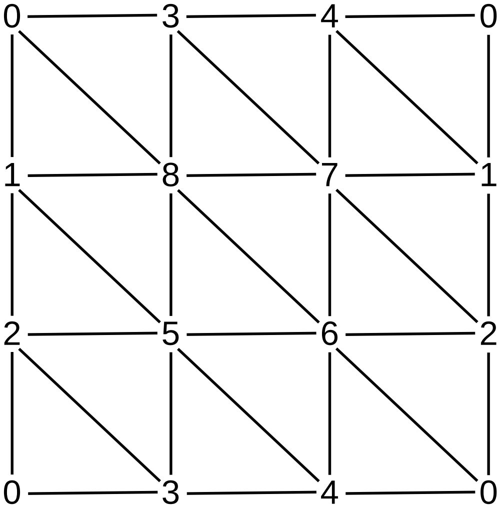


One connected component, one hole in dimension 1.

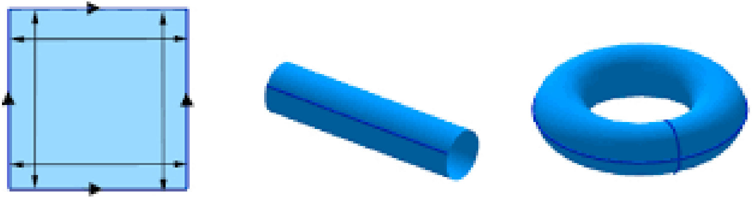
Persistence

1. Suppose we track homology for each radius
2. We obtain Persistent homology, an invariant that is...
3. Multiscale
4. Robust
5. Equipped with metric
6. Aplicable to variety of complexes,
7. Time series analysis,
8. Similarity measures and more...

Example; triangulation of a torus



Triangulation of a torus



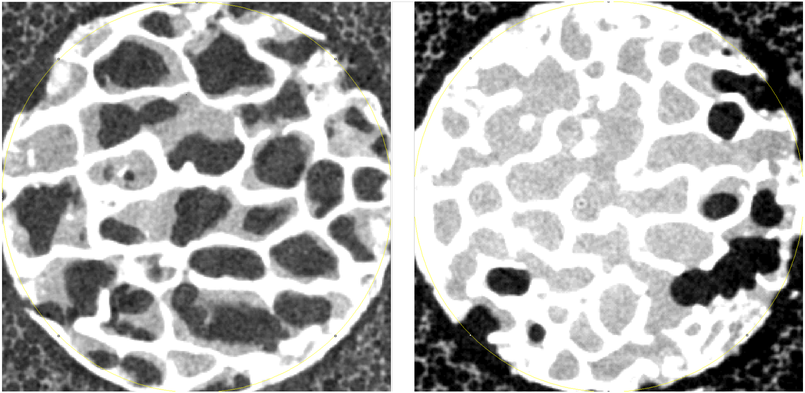
Practical exercise 1

- Please go to `dioscuri-tda.org/Bedlewo_TDA_Tutorial_2021.html`,
- Download *exercises in Persistent homology*,
- Open intro_to_homology and play with triangulation of a torus.
- What are the homology groups of this triangulation?

Practical exercise 2

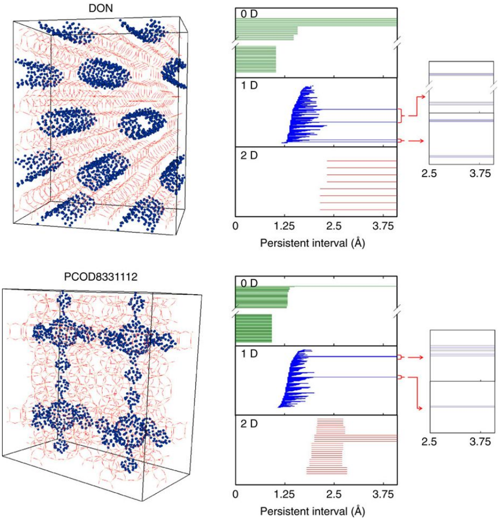
- Let us go back to our jupyter-notebooks exercises.
- Open persistence_simple_point_cloud,
- Compute persistent homology of a point cloud sampled from a circle (without and with a considerable amount of noise).

Apply to digital images



Left – osteoporotic, right – normal bone (vertebrae).
Not only density, but mostly structure is responsible for osteoporotic fractures.

Persistence–based descriptors of nanoporous materials



Lee, Bathel, Dłotko, Mossavit, Smit, Hess, Quantifying similarity of pore-geometry in nanoporous materials, Nature Communications, 15396

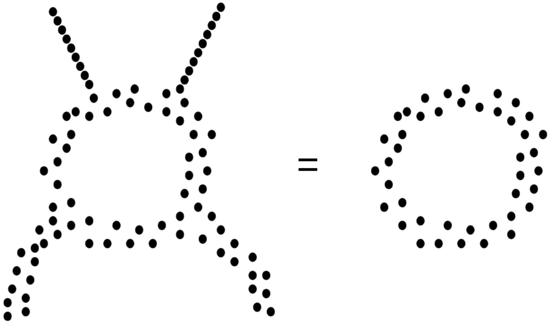
And more...

- We do not have time to cover all this ground.
- But, there are numerous resources for further work:
  - `https://arxiv.org/abs/1807.08607`
  - `https://www.maths.ed.ac.uk/~v1ranick/papers/edelcomp.pdf`
  - `https://gudhi.inria.fr/tutorials/`
  - and many more...

Homology and persistent homology, biased collection of resources

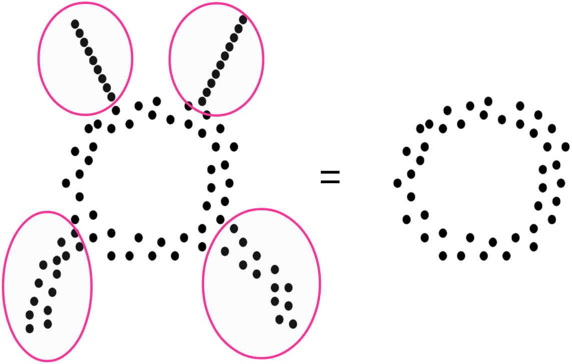
- Edelsbrunner and Harer, Computational Topology, An Introduction, AMS.
- Kaczynski, Mischaikow, Mrozek, Computational Topology, Springer 2003.
- Dłotko, Applied and Computational Topology, Tutorial
- Multiple youtube videos.

Can we have something more visual?



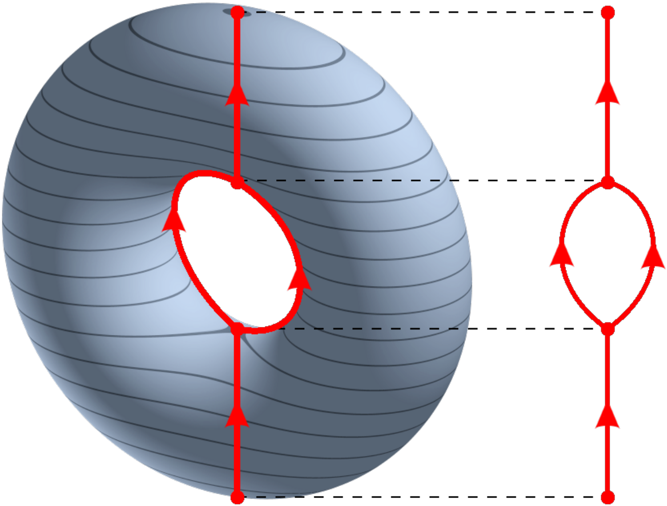
Persistence homology of those two point clouds will be very similar, as they both have one connected component and one hole.

To see flares?



But, oftentimes the information in the *flares* may be important (it may for instance carry information about anomalies).

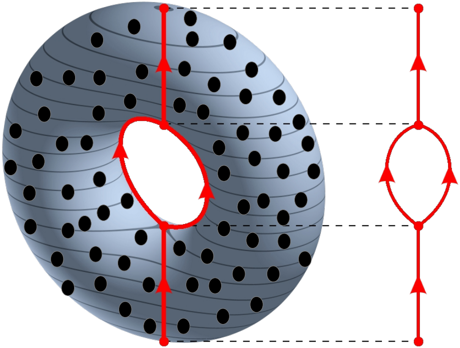
Reeb graph



Reeb graph, formally

- Input: M, $f : M \to \mathbb{R}$.
- We define an equivalence relation x R y iff:
  - $f(x) = f(y)$,
  - x and y belong to the same connected component of $f^{-1}(x)$.
- $M/_{R}$.

Conventional Mapper algorithm

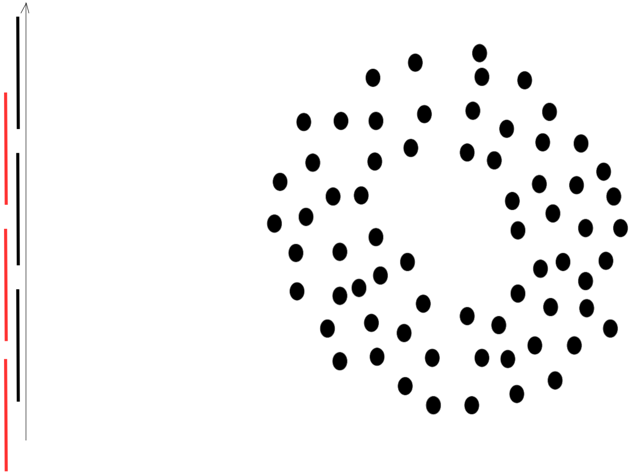


Conventional mapper graph is an attempt to define Reeb graph for discrete point cloud instead of a manifold.

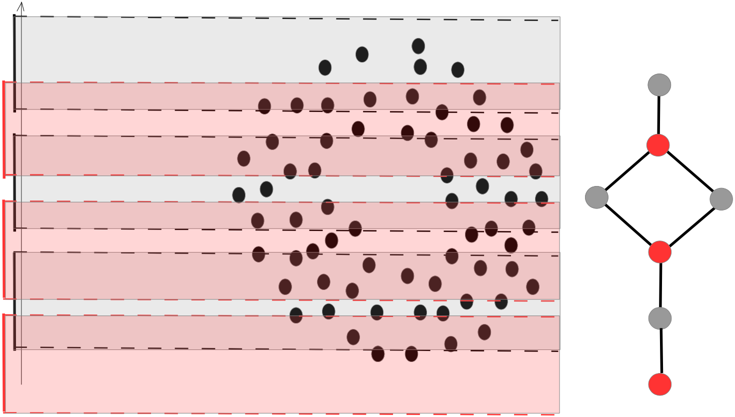
Mapper algorithm, idea

- Input: finite collection of points sampled from M, $f : M \to \mathbb{R}$.
- We define a relation x R y iff:
  - $f(x)$ is close to $f(y)$,
  - x and y belong to the same cluster ...

Conventional Mapper algorithm



Conventional Mapper algorithm



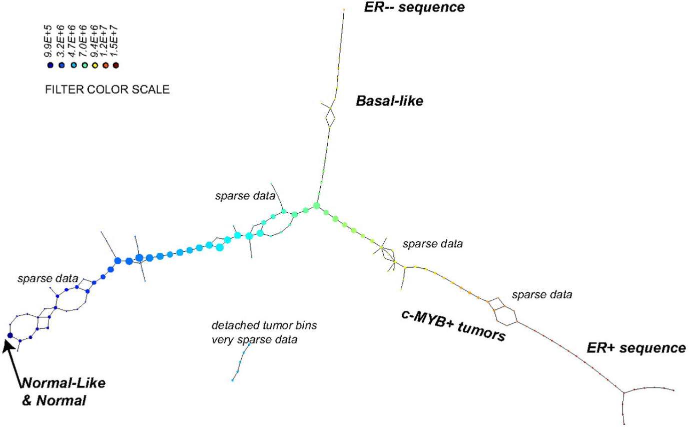
Mapper algorithm, formally

- Input: finite collection of points sampled from M, $f : M \to \mathbb{R}$.
- Cover of the range of f with overlapping boxes.
- Fix a clustering algorithm
- We define a relation $x \text{ R } y$ iff:
    - $f(x)$ and $f(y)$ belong to the same element I of a cover of the range of f,
    - x and y belong to the same cluster in $f^{-1}(I)$.
- Vertices of Mapper graph correspond to the clusters,
- An edge is placed between two vertices if the corresponding clusters have nonempty intersection.

Mapper algorithm, coloring

- Vertices of the Mapper graph may be colored by an average value of an objective function on points covered by clusters.
- Fix a point cloud X and an objective function $f : X \to \mathbb{R}$.
- Each vertex of the Mapper graph correspond a subset (cluster) of points from X.
- Typically the value of the vertex will be an average value of f on the corresponding cluster.

Mapper is the most well know tool of TDA



Nicolau, Levine, Carlsson, *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*, PNAS 2011.

Practical exercise 1

- Let us play with Mapper algorithm!
- Go to `https://dioscuri-tda.org/Bedlewo_TDA_Tutorial_2021.html`, download exercises in Standard Mapper.
- Let us start from something simple – open standard_mapper_concentric_circles
- In this exercise we will generate two concentric circles in a plane.
- We will use projection to the y coordinate as a lens function,
- And a DBSCAN with certain parameters as a clustering algorithm.
- What is the Mapper graph we obtain?

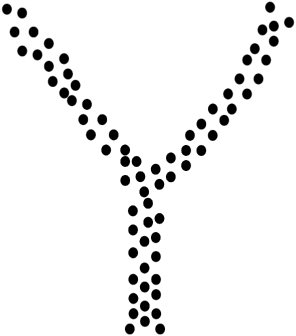
Practical exercise 2

- Let us play with something more advanced, let us consider standard Boston property dataset.
- Please open standard_mapper_boston_dataset
- It contains 13 variables, we want to understand its relation to prices of properties in Boston area (in '1970).
- Here we will use t-distributed stochastic neighbor embedding as a filtering function.
- We will be able to experiment with numerous clustering methods as well.
- Obtained mapper graphs will be colored by the average price of a property in a given cluster.
- This is not the last time we see Boston Property Dataset!

Ball Mapper algorithm

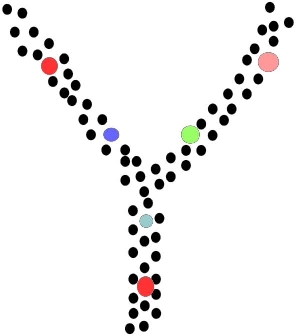
- As a last part of our schedule, we will play with Ball Mapper algorithm.
- As you might have noticed, it is not always trivial to choose the *lens function* as well as *clustering algorithm* in standard Mapper construction.
- The idea of Ball Mapper is intuitively explained in the following slides.

Ball Mapper algorithm



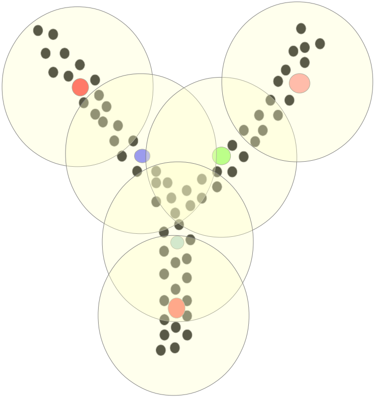
Take a point cloud X

Ball Mapper algorithm



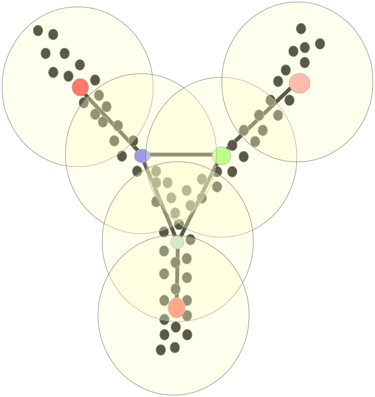
Given $\epsilon > 0$, select subset of points $N \subset X$ such that for every $x \in X$ there exist $n \in N$ such that $d(x, n) \leq \epsilon$ (we call N an ϵ-net)

Ball Mapper algorithm



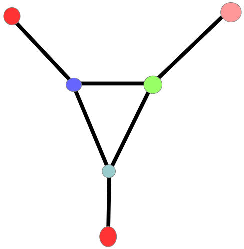
Consequently $X \subset \bigcup_{n \in N} B(n, \epsilon)$, i.e. $\{B(n, \epsilon), n \in N\}$ cover X.

Ball Mapper algorithm



Take one dimensional nerve of that cover (an abstract graph whose vertices correspond to $B(n, \epsilon)$, and edges to nonempty intersections of balls)

Ball Mapper algorithm



This way we obtain a Ball Mapper graph of X with radius ϵ. Vertices of the graph can be colored analogously to those of standard Mapper graph.

Practical exercise 1

- Please open example_basic_circle.
- In this proof-of-concept example we will generate a collection of points sampled from a unit circle $x^2 + y^2 = 1$.
- And built a Ball Mapper graph based on it.
- Do we see what we expect to see?

Practical exercise 2

- In our second example we will re-visit already known Boston Property Dataset.
- Please open example_Boston_property
- This time we will use Ball Mapper to examine the structure of the 13 dimensional point cloud, and the distribution of the explanatory variable (price of properties) on the top of it.
- We will use tools from the Ball Mapper implementations to recognize which coordinates makes most statistical differences between the regions of the graph.

Some solutions

- Please note that the solutions to some of the questions are available at `https://dioscuri-tda.org/bedlewo_2021_tutorial/extra.zip` and download solutions to extra exercises.
- Please however make an attempt to solve it by yourself before moving to it!

Thank you for your time!

Dioscuri Centre in Topological Data Analysis
@Facebook

DIOSCURI
CENTRE IN TOPOLOGICAL DATA ANALYSIS

MAX PLANCK GESELLSCHAFT

NATIONAL SCIENCE CENTRE POLAND

Jointly sponsored by

Ministry of Science and Higher Education
Republic of Poland

Federal Ministry of Education and Research

Paweł Dłotko
pdlotko @ impan.pl
pdlotko @ gmail
pawel_dlotko @ skype