Topological data analysis: Mapper, Persistence and Applications.

Paweł Dłotko, Dioscuri Centre in TDA, IMPAN.

Vincent Rouvreau, Inria Saclay.

9 January 2021.

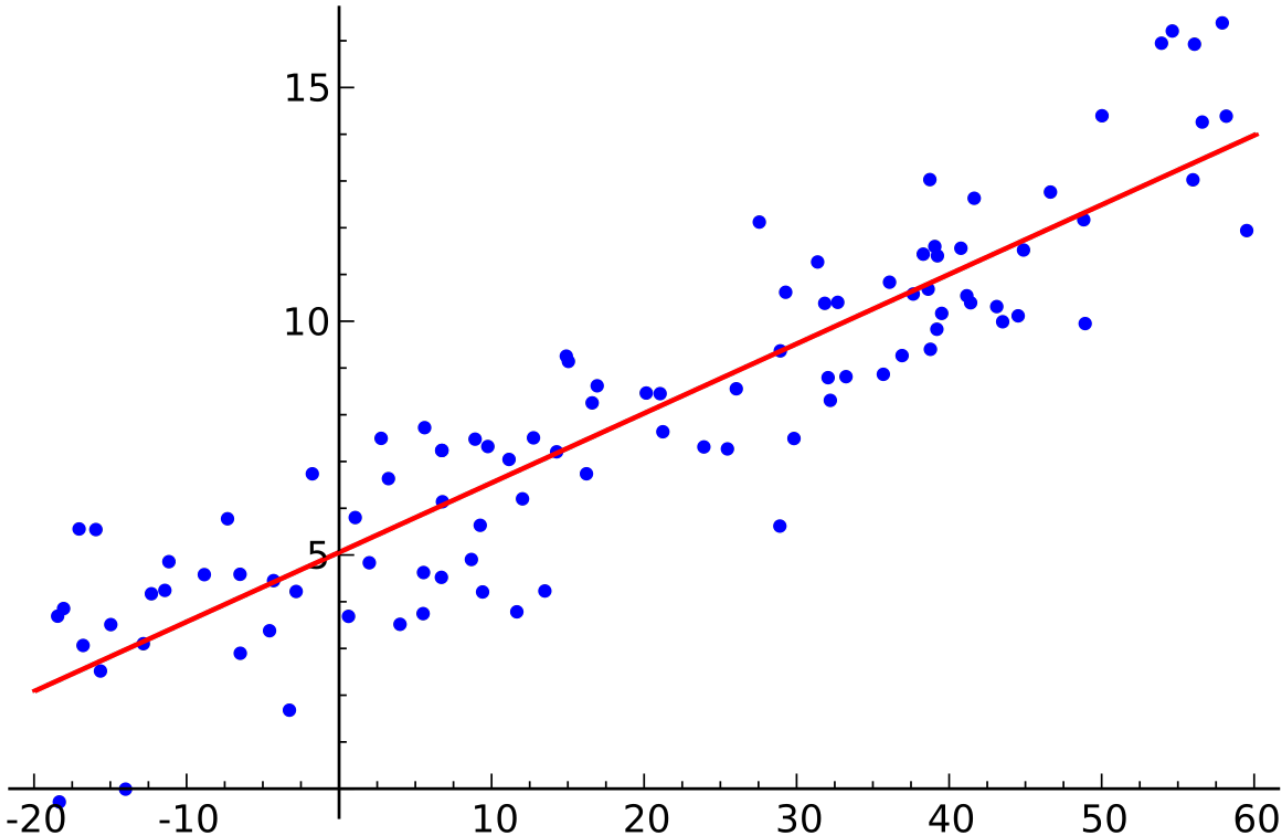
Topological Data Analysis

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

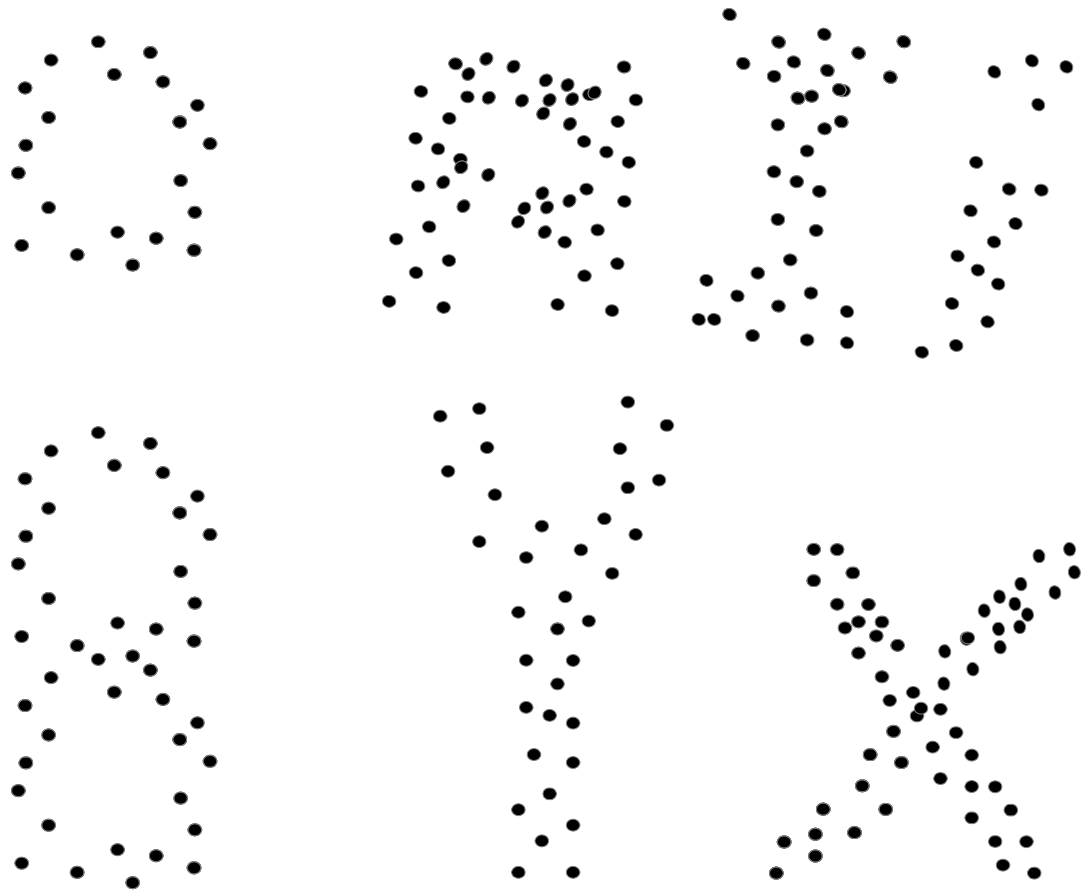
The credo.

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

We all know this story.

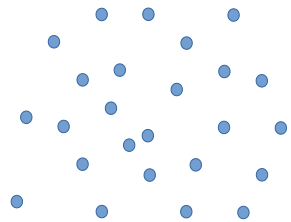


Trap of models.



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

The pipeline.



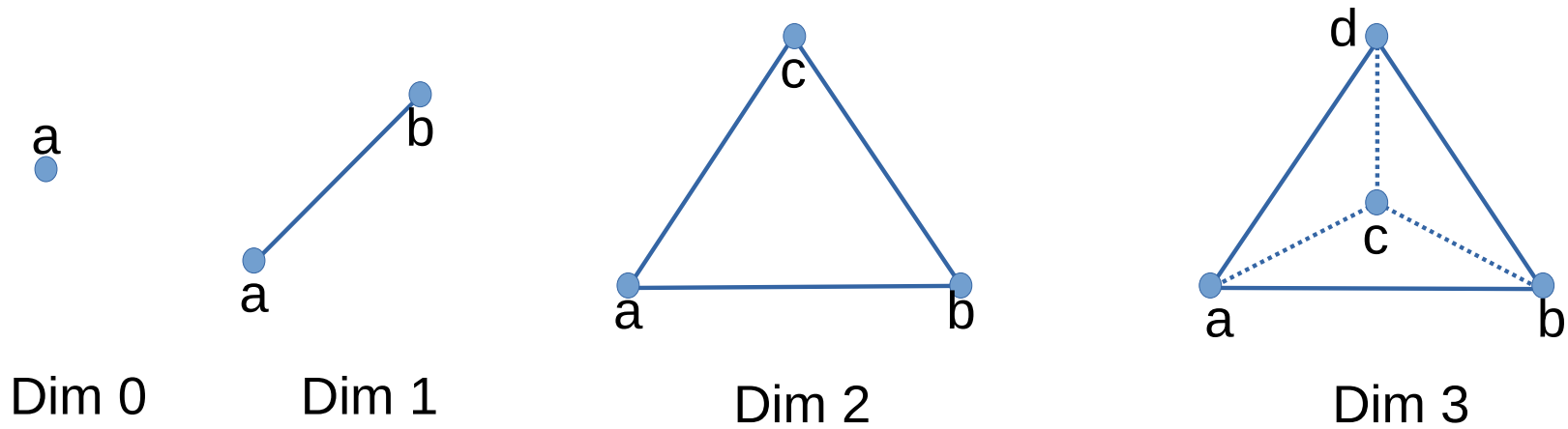
Point cloud

Topological descriptor

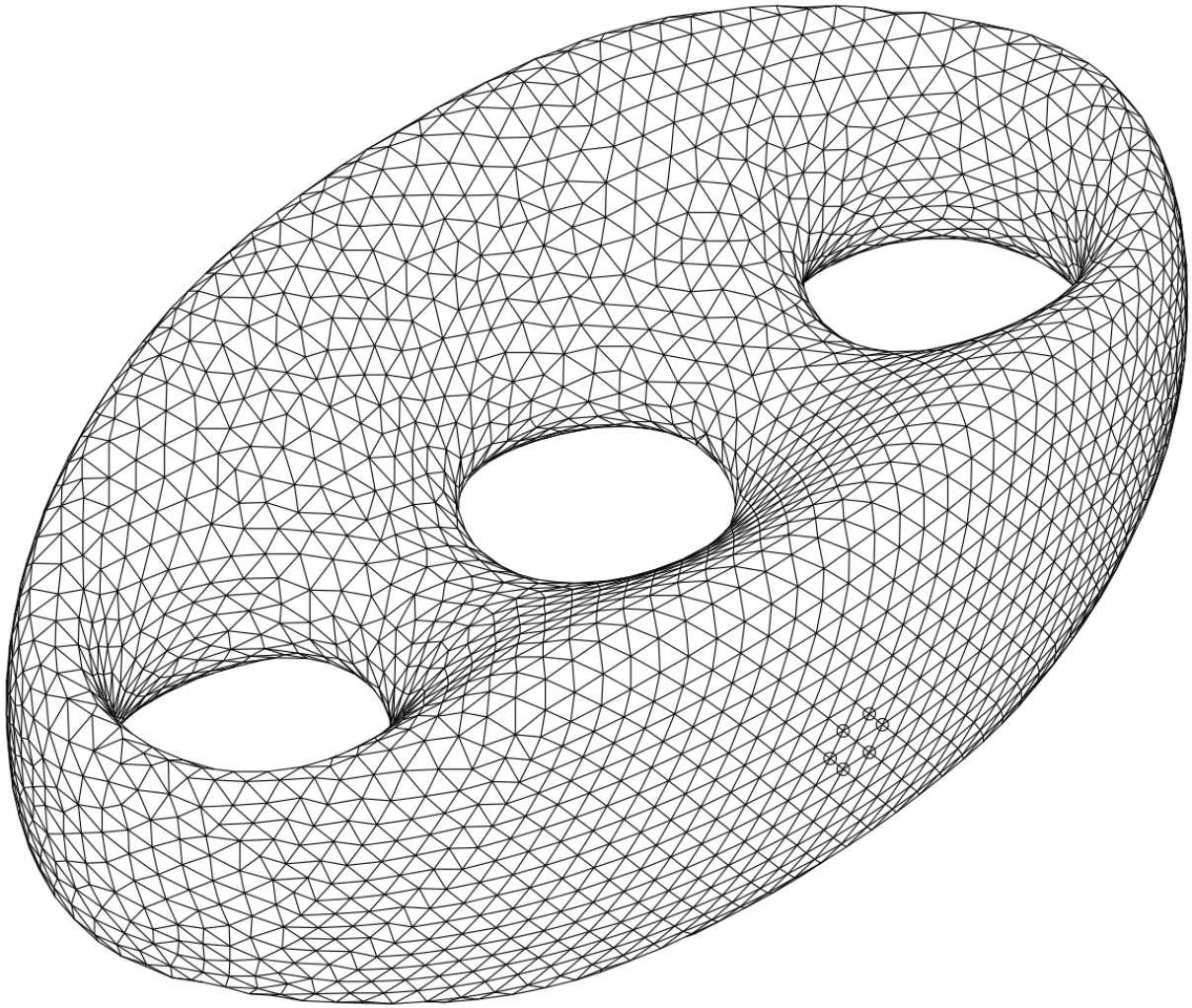
Inference

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.



Sample simplicial complexes



Source: Wikipedia.

Let the data tell you the story.

Topological data analysis:

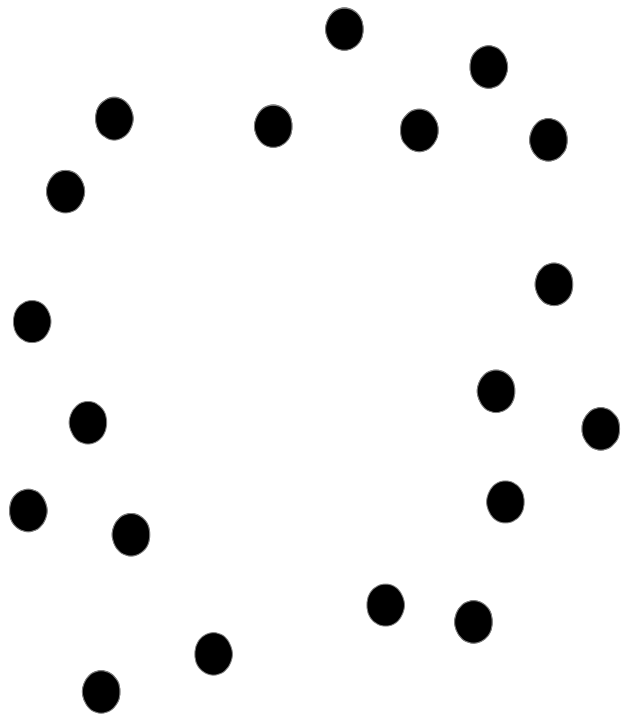
- Persistent homology – point-cloud based homology.
- Accurate network models to examine landscapes of data,

! Stable.

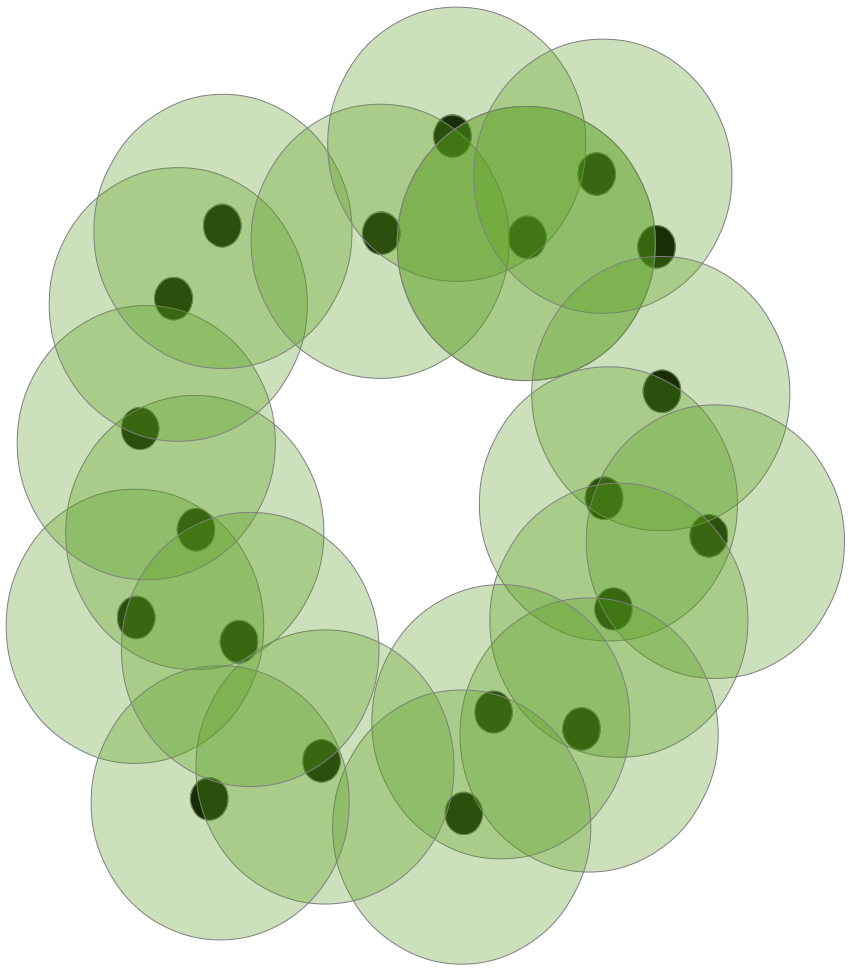
!! No black boxes.

!!! We do not enforce *any* models of data.

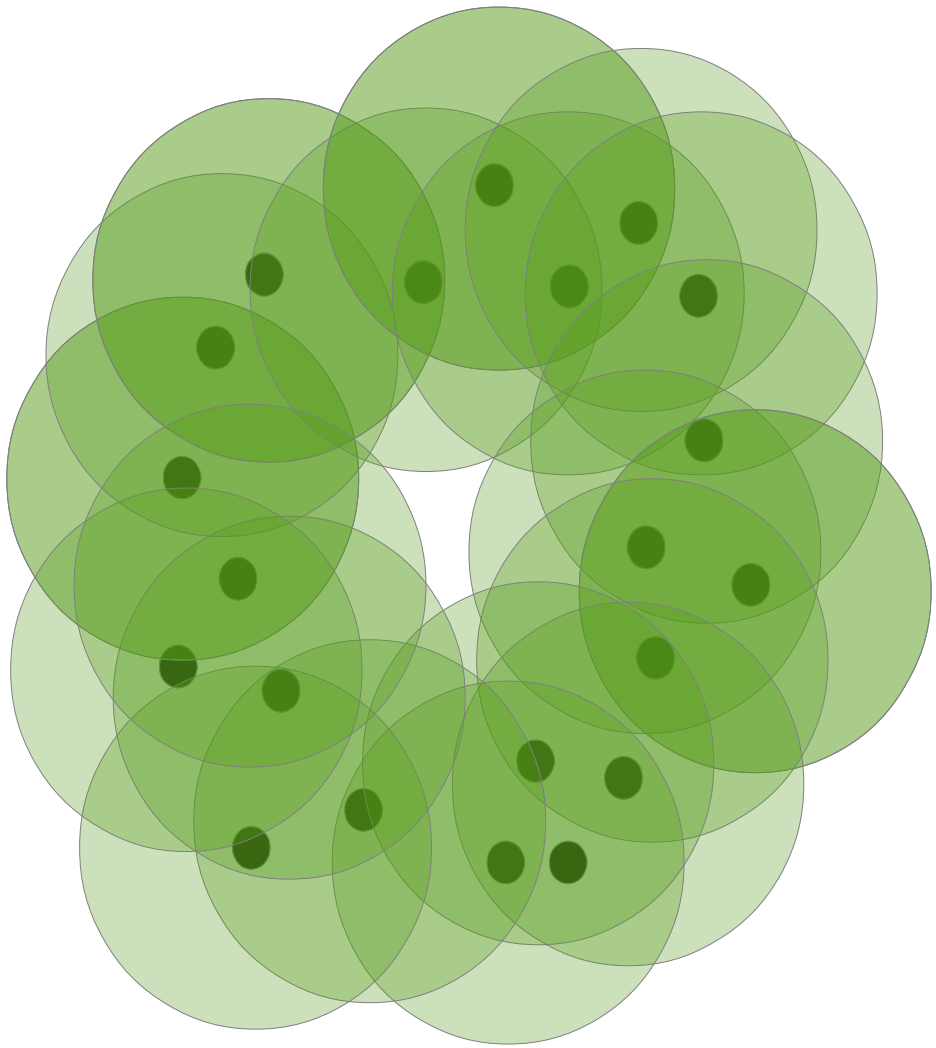
What do you see?



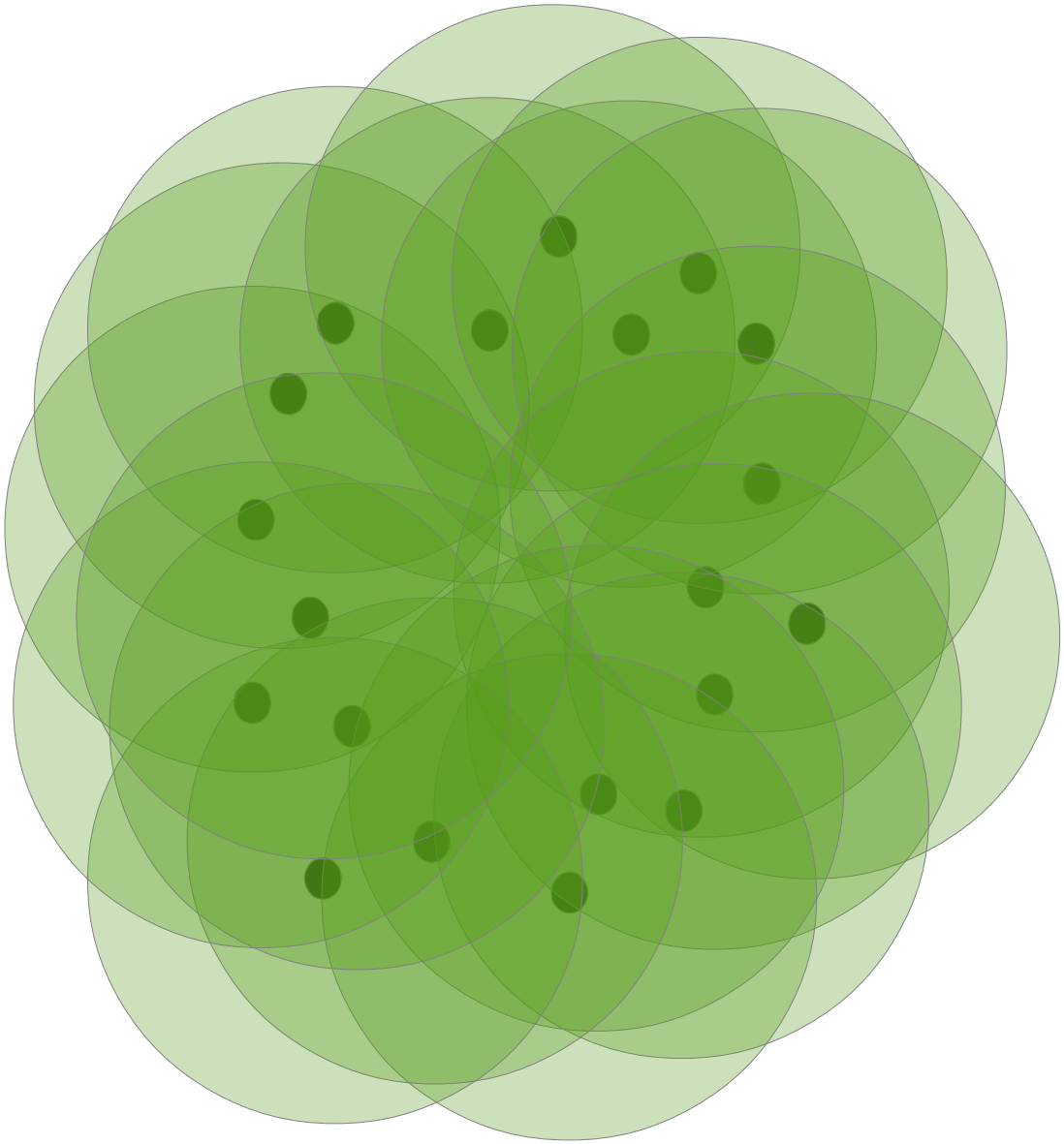
What do you see?



What do you see?



What do you see?



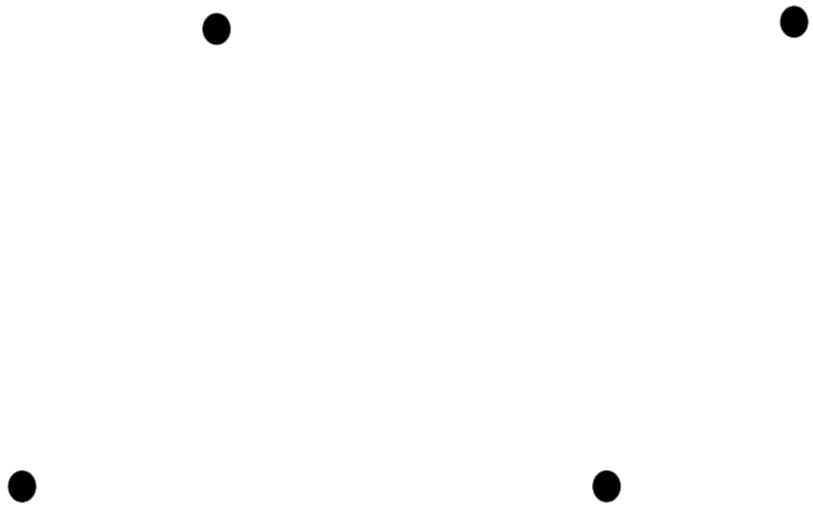
What do you see?

- We may say that we see a circle,
- But we really see 19 points...
- ...that may be sampled from a probability distribution supported at a circle.
- Persistent homology is a tool to make this observation precise.
- To do so, we need to construct a *filtered complex* of the point cloud.
- A filtered complex is a nested sequence of subcomplexes - a way of building a model by adding a sequence of simplices in a number of steps.

Simplicial complexes built from point clouds

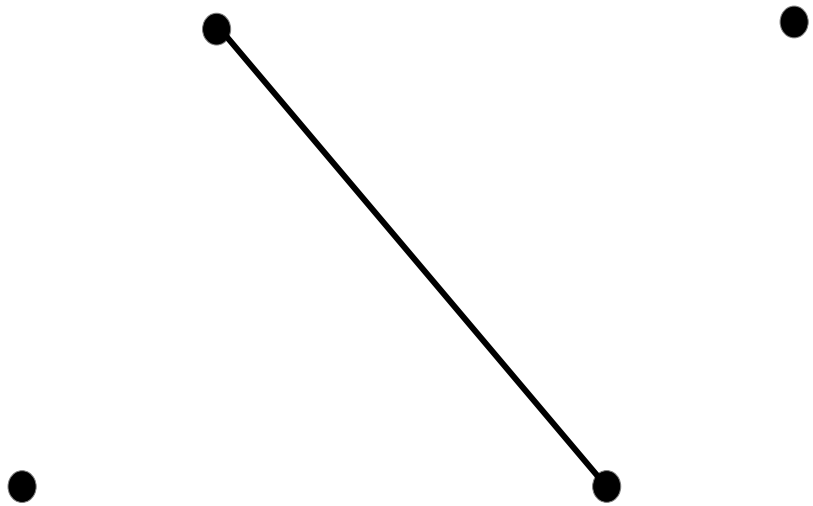
- P finite point cloud with a metric d.
- Rips complex at level ϵ consists of simplices supported in $p_0, \ldots, p_n$ if $B(p_i, \frac{\epsilon}{2}) \cap B(p_j, \frac{\epsilon}{2}) \neq \emptyset$ for every $i, j \in \{0, \ldots, n\}$.
- Čech complex at level ϵ consists of simplices supported in $p_0, \ldots, p_n$ iff $\bigcap_{i=0}^{n} B(p_i, \frac{\epsilon}{2}) \neq \emptyset$.

Filtration of Rips complex



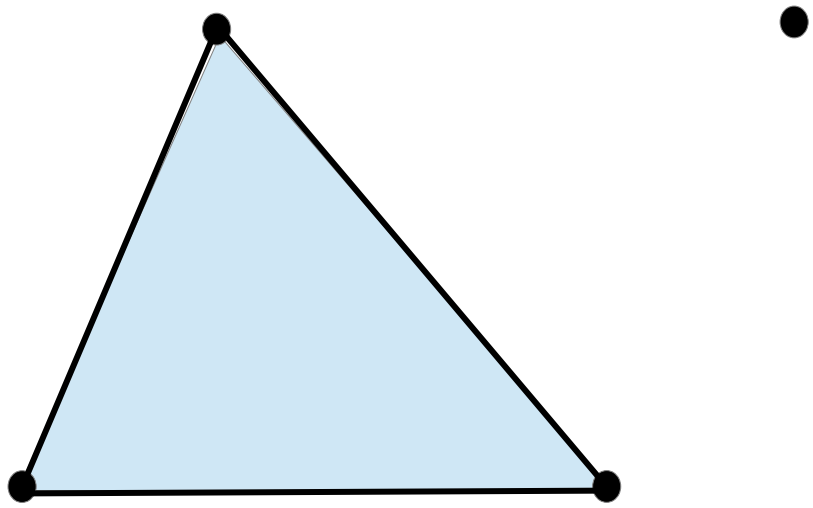
4 vertices

Filtration of Rips complex



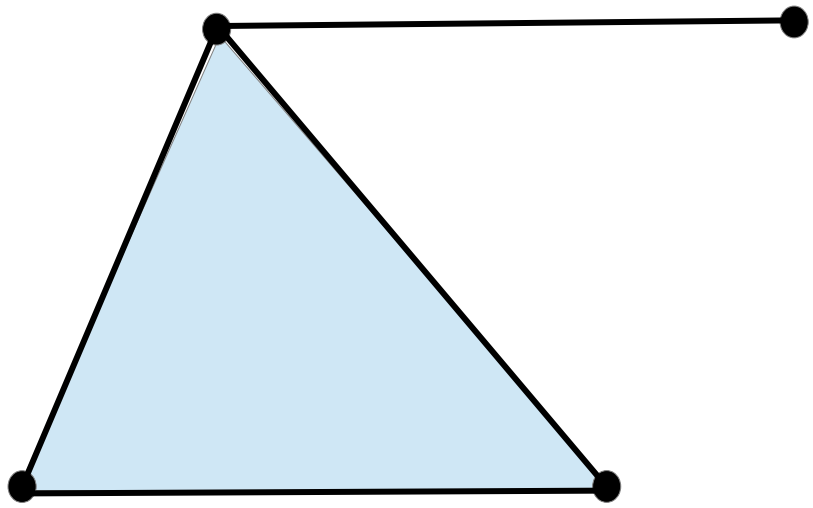
4 vertices, 1 edge

Filtration of Rips complex



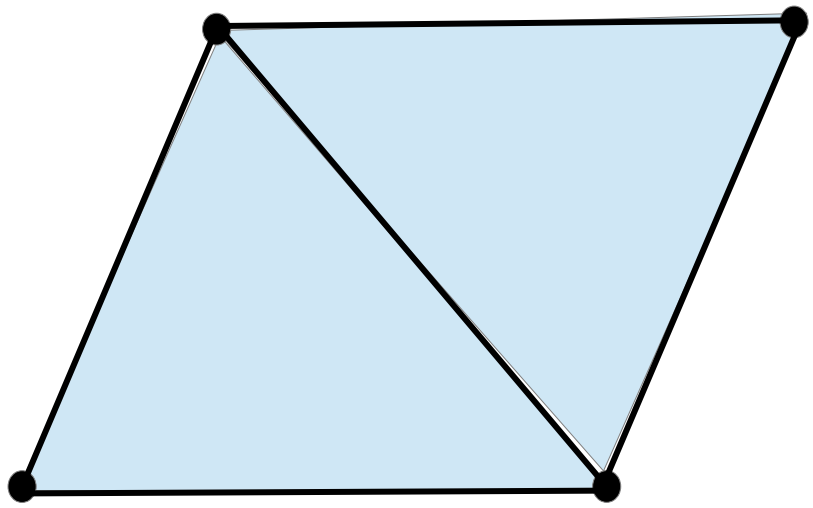
4 vertices, 3 edges, 1 triangle

Filtration of Rips complex



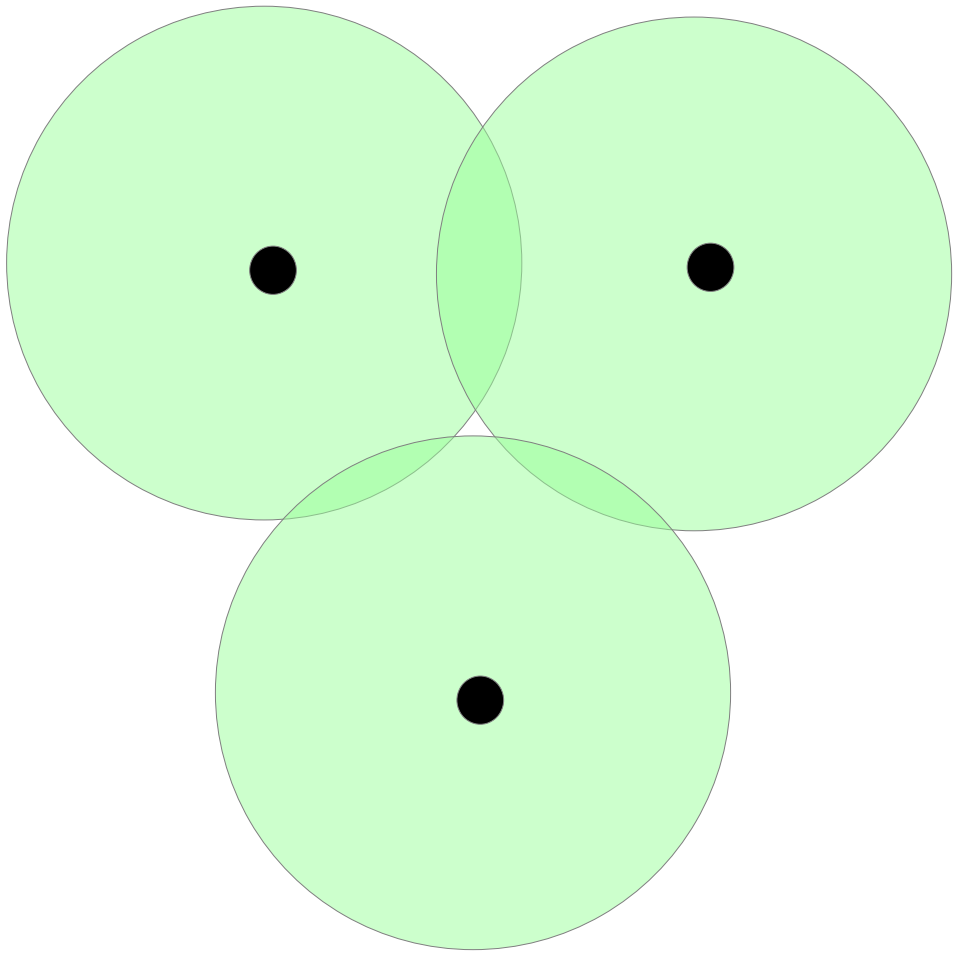
4 vertices, 4 edges, 1 triangle

Filtration of Rips complex

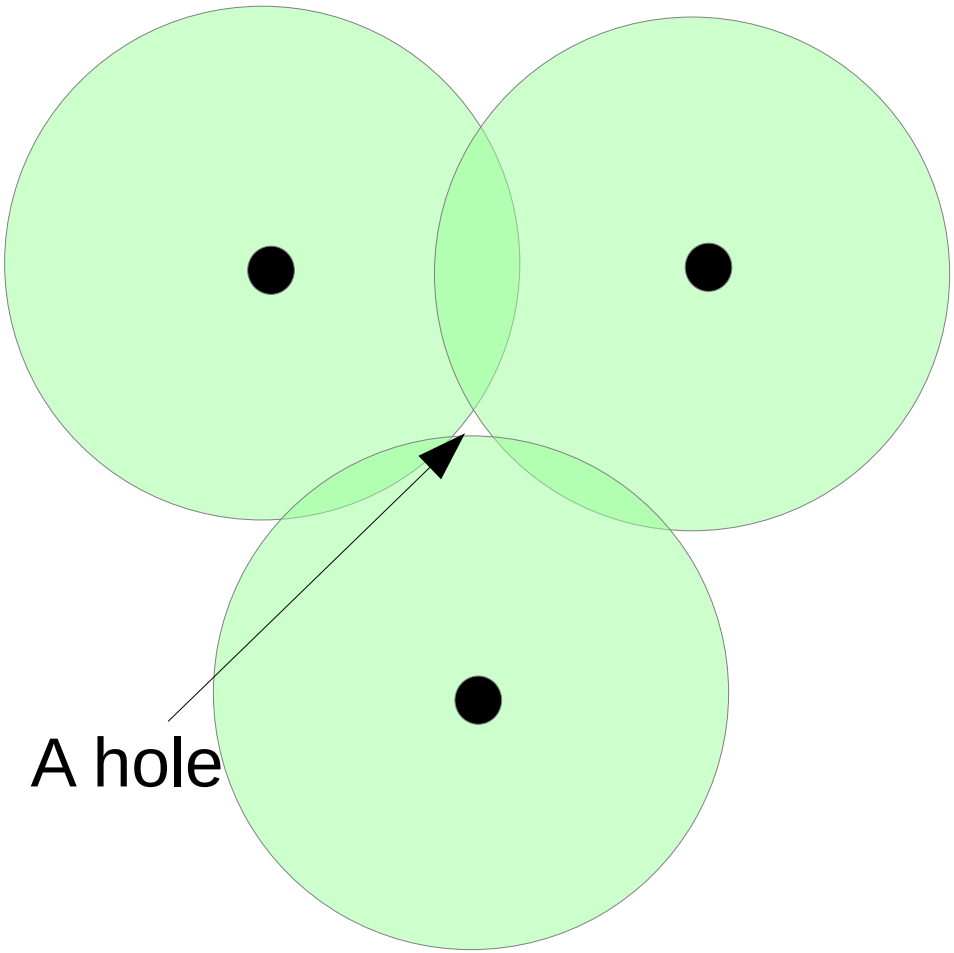


4 vertices, 5 edges, 2 triangles

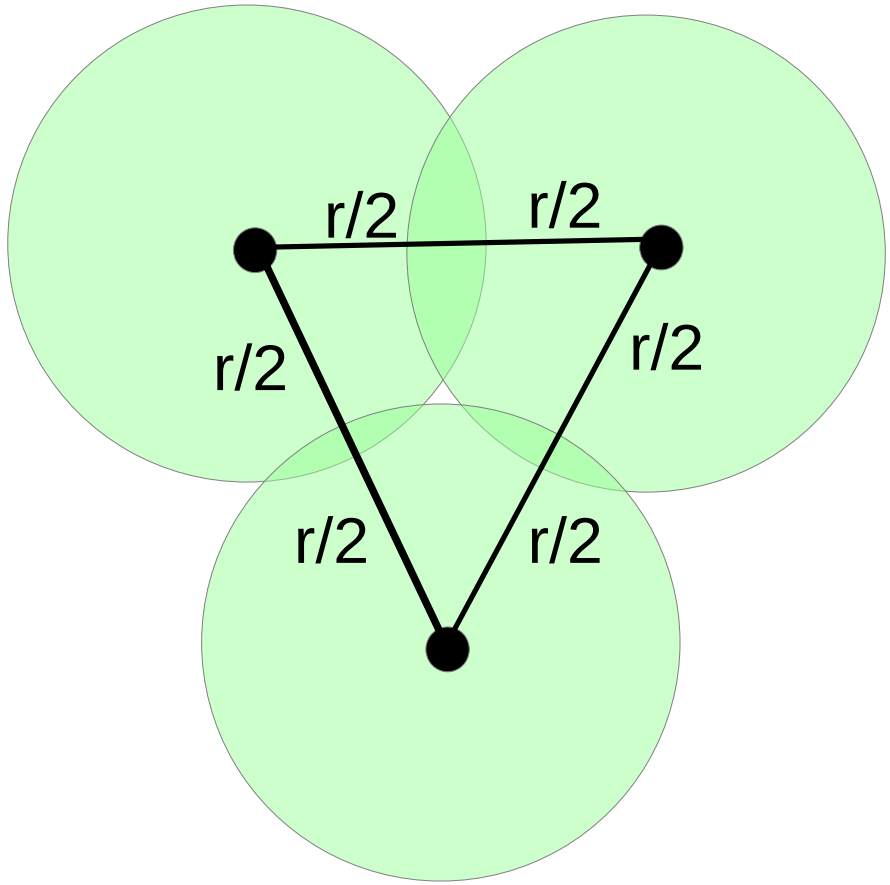
Rips vs Čech



Rips vs Čech



Rips vs Čech

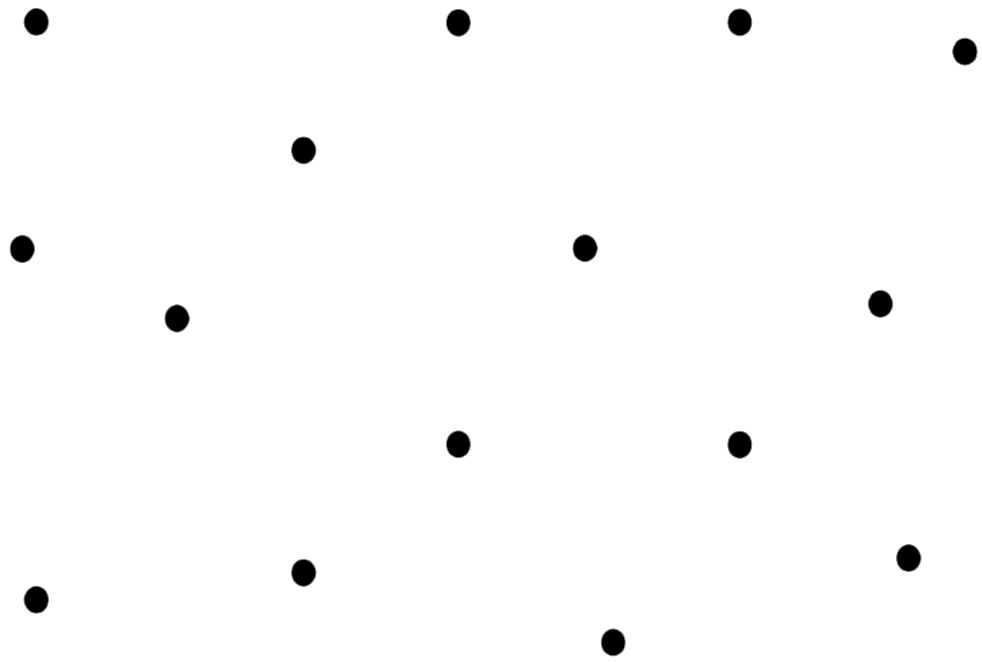


In this case Rips complex is a triangle with boundary, a Čech complex is a boundary of triangle

Čech complex is topologically accurate

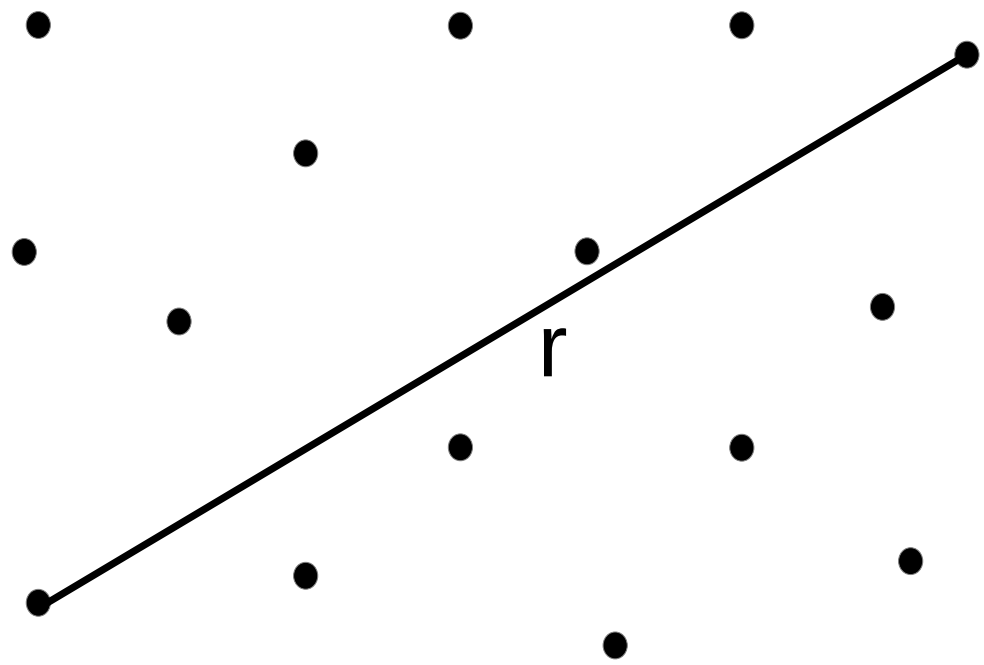
- $\bigcup_{p \in P} B(p, \frac{\epsilon}{2})$ is topologically equivalent to the Čeach complex based on those balls.
- Meaning, there exist a continuous deformation from one into another.
- No tearing, no gluing.

Rips and Čech complexes can grow large



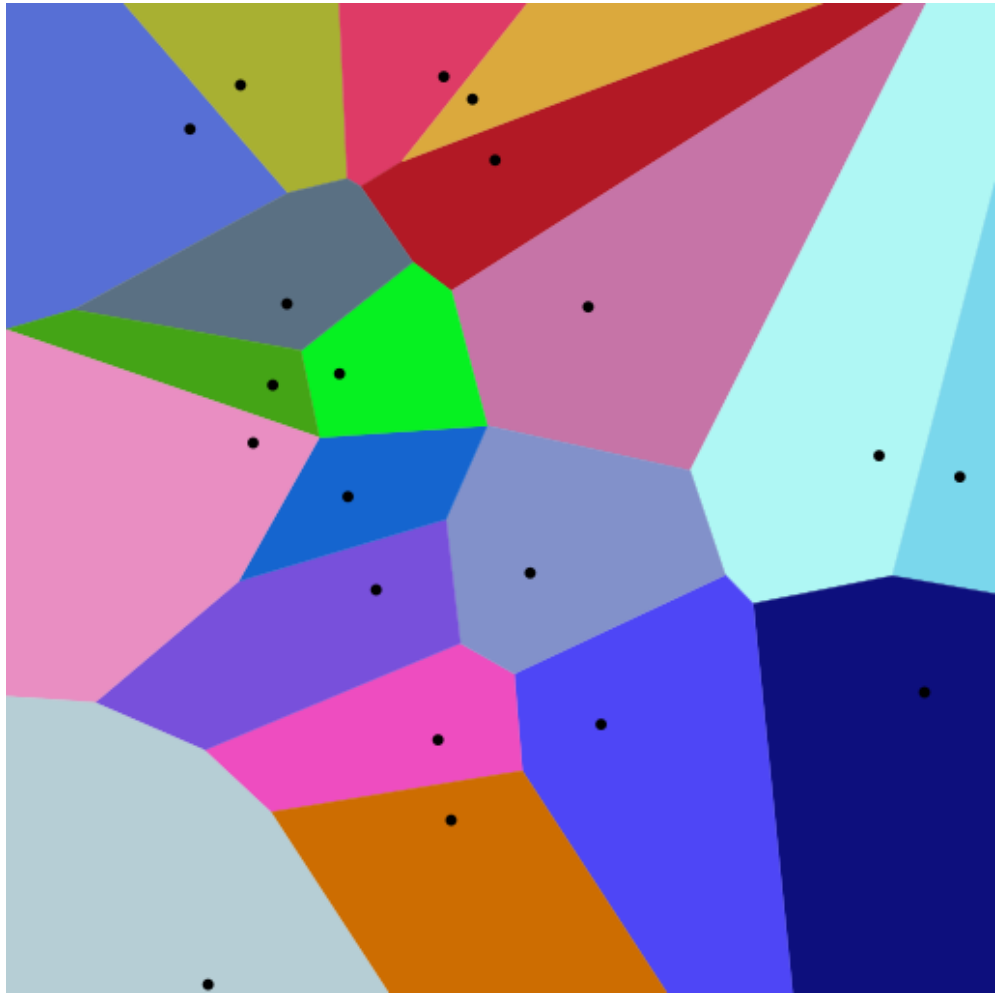
If all points get connected by edges in the complex, we witness so called *combinatorial explosion*. You will encounter it when using Rips complexes.

Rips and Čech 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.

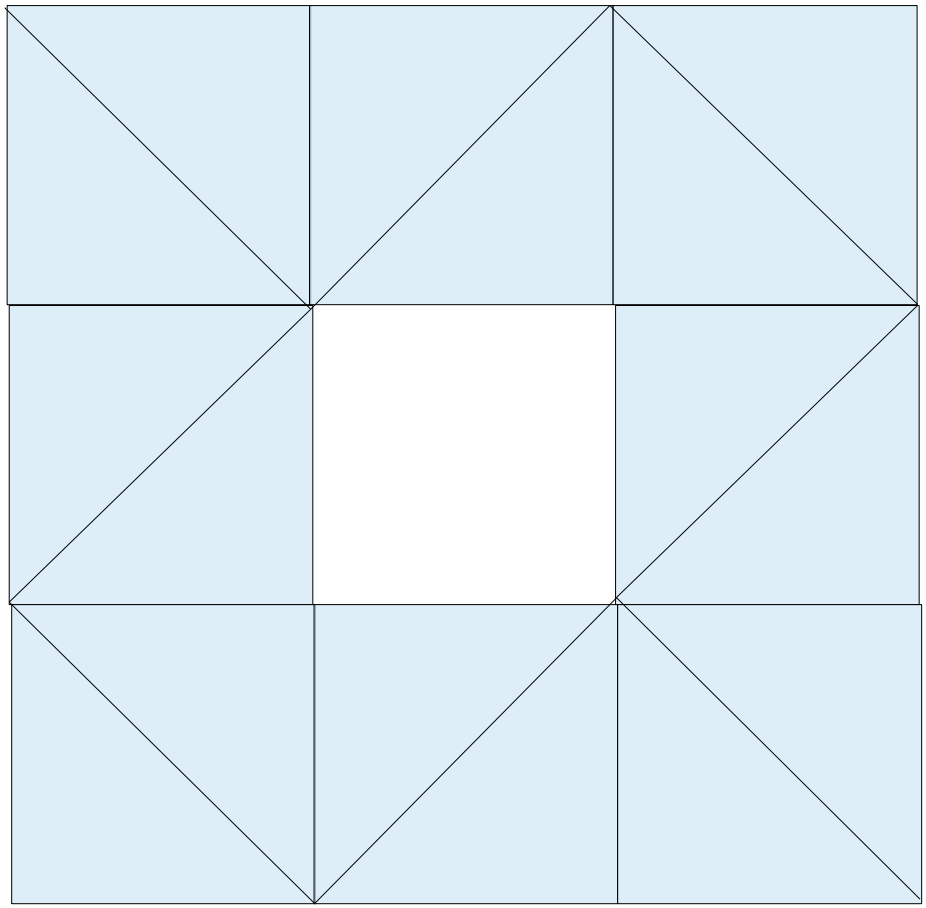
Alpha complexes



Intersecting $B(x, r)$, for $x \in X$ with Voronoi cells of X allows to built much smaller complexes that preserves homotopy type of $\bigcup_{x \in X} B(x, r)$.

From complexes to parameter dependent homology

Homology

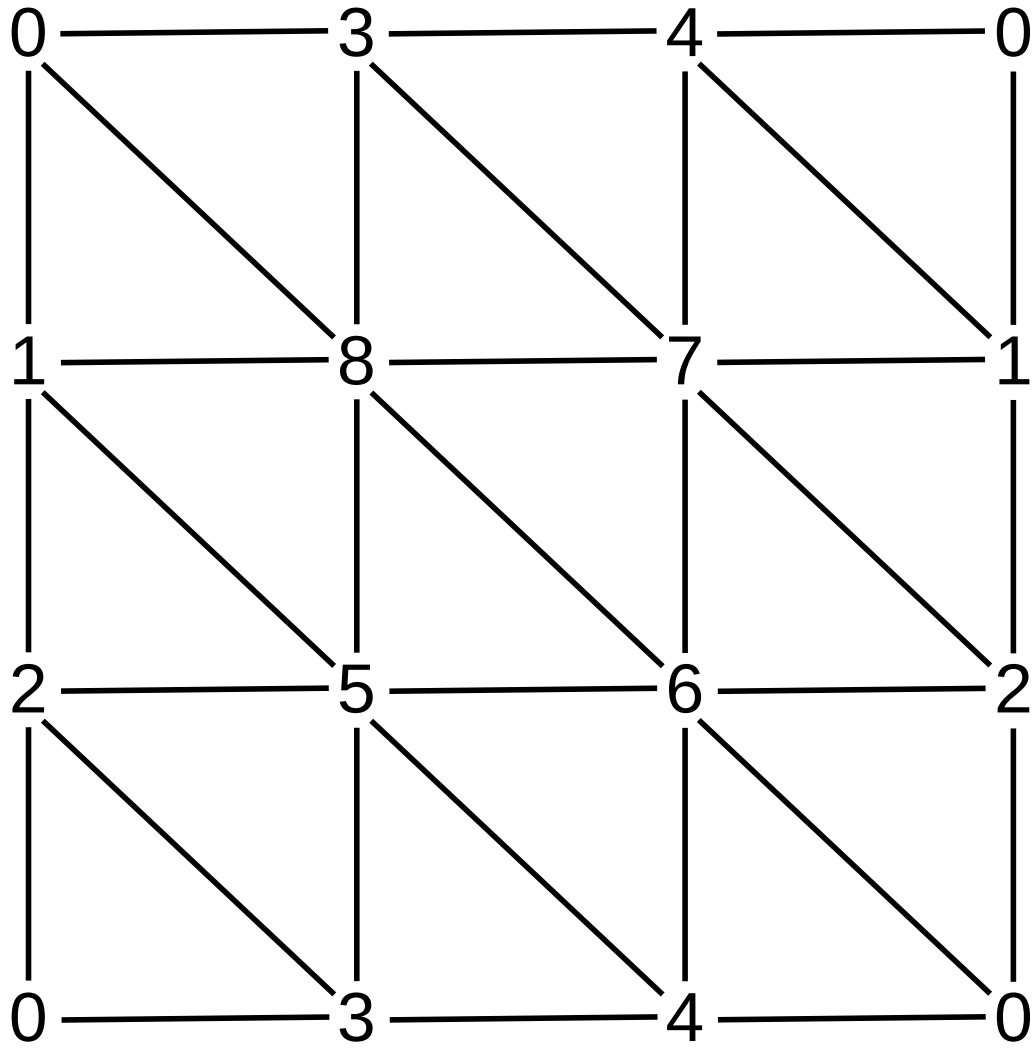


One connected component, one hole in dimension 1.

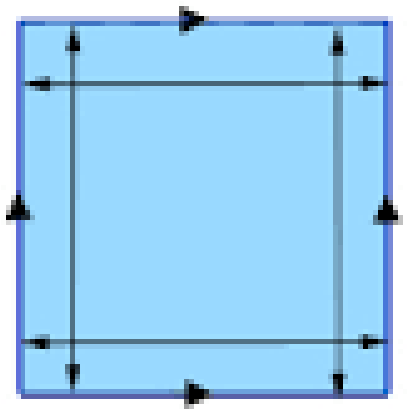
Practical exercise 1.

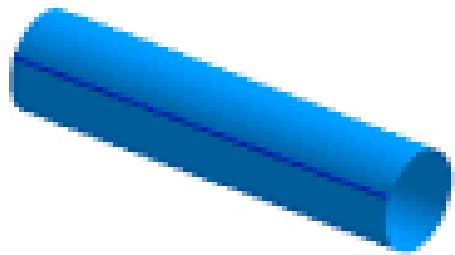
- Please go to `https://dioscuri-tda.org/Paris_TDA_Tutorial_2021.html`,
- Download *exercises in Persistent homology*,
- Open intro_to_homology.ipynb and play with triangulation of a torus.
- What are the homology groups of this triangulation?

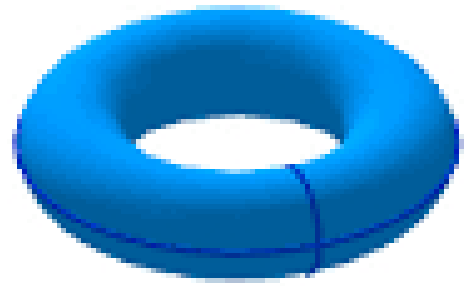
Triangulation of a torus.



Triangulation of a torus.







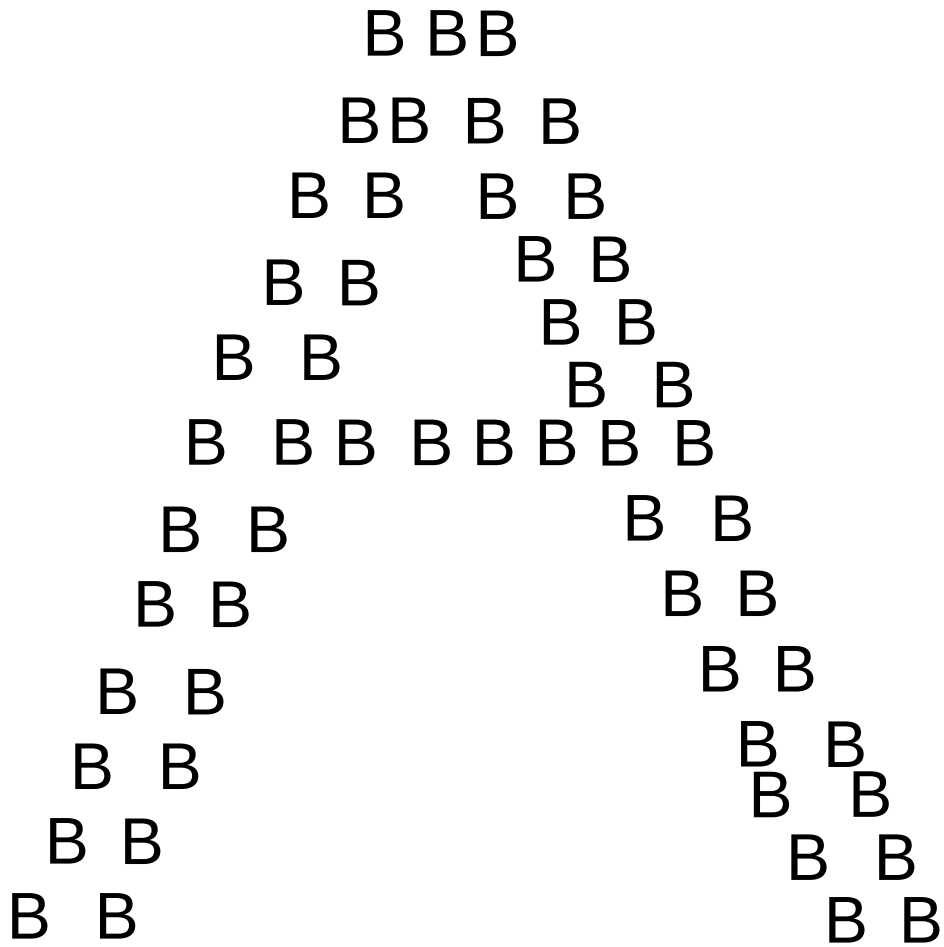
Persistent homology, under the hood

- Let us order simplices according to the minimal ϵ for which they appear (filtration).
- Algorithm to compute (persistent) homology is a version of Gaussian elimination.
- If we run it for a prefix of filtration, we will get homology of the complex composed by simplices in that prefix (a subcomplex of the final complex).
- Analyzing the structure of zero and non-zero columns in the reduced matrix allows to find generators that are created and which became trivial as we move along the filtration.

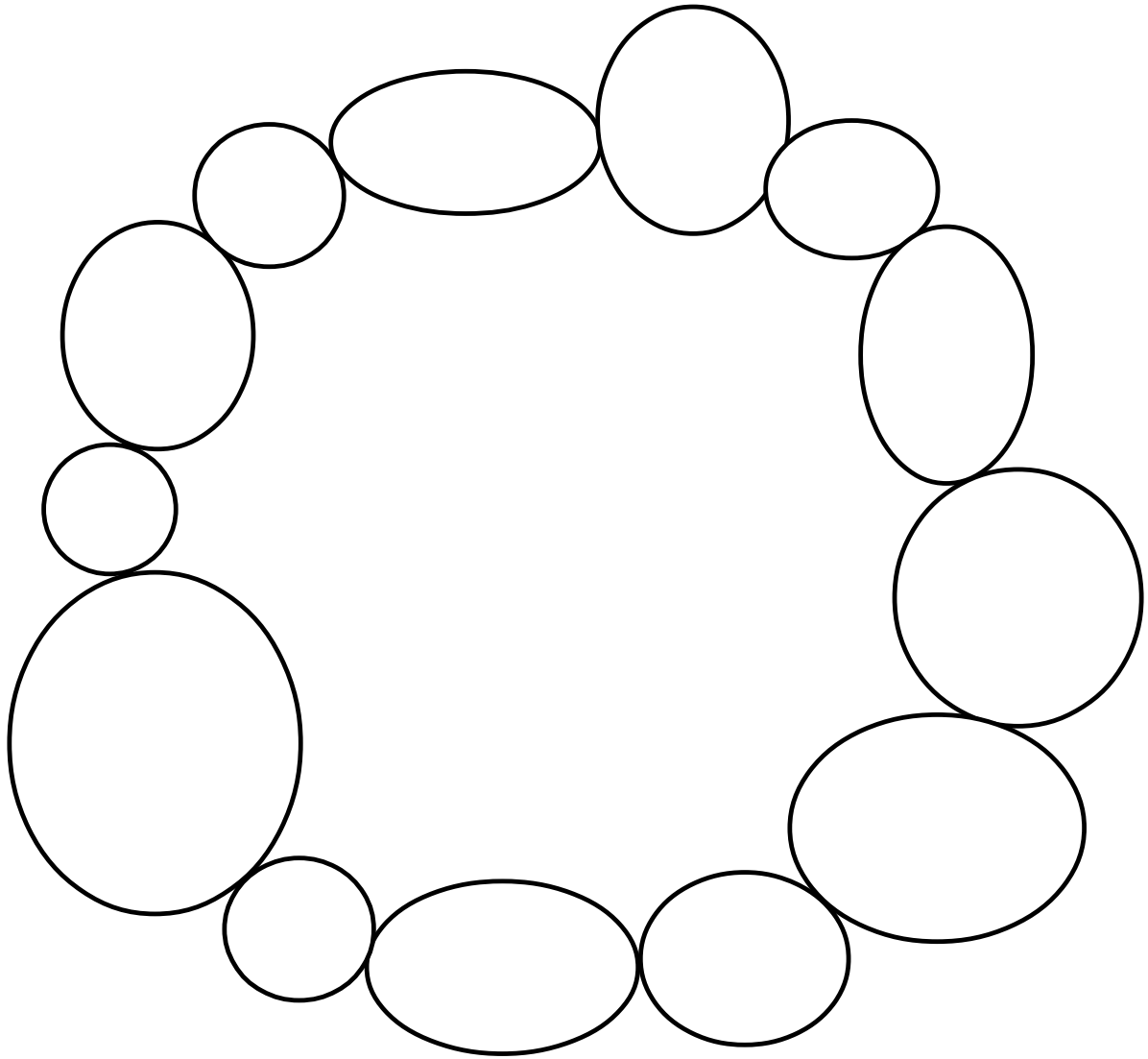
Invariance.

- Persistent homology is a rigorous way of quantifying closed *shapes*,
- ... like connected components, cycles, voids and more.
- No matter if they are embedded in two or million dimensional space,
- No matter if they are rotated, stretched or transformed in any other way.
-
-

Lots of **B**, or a single **A**?



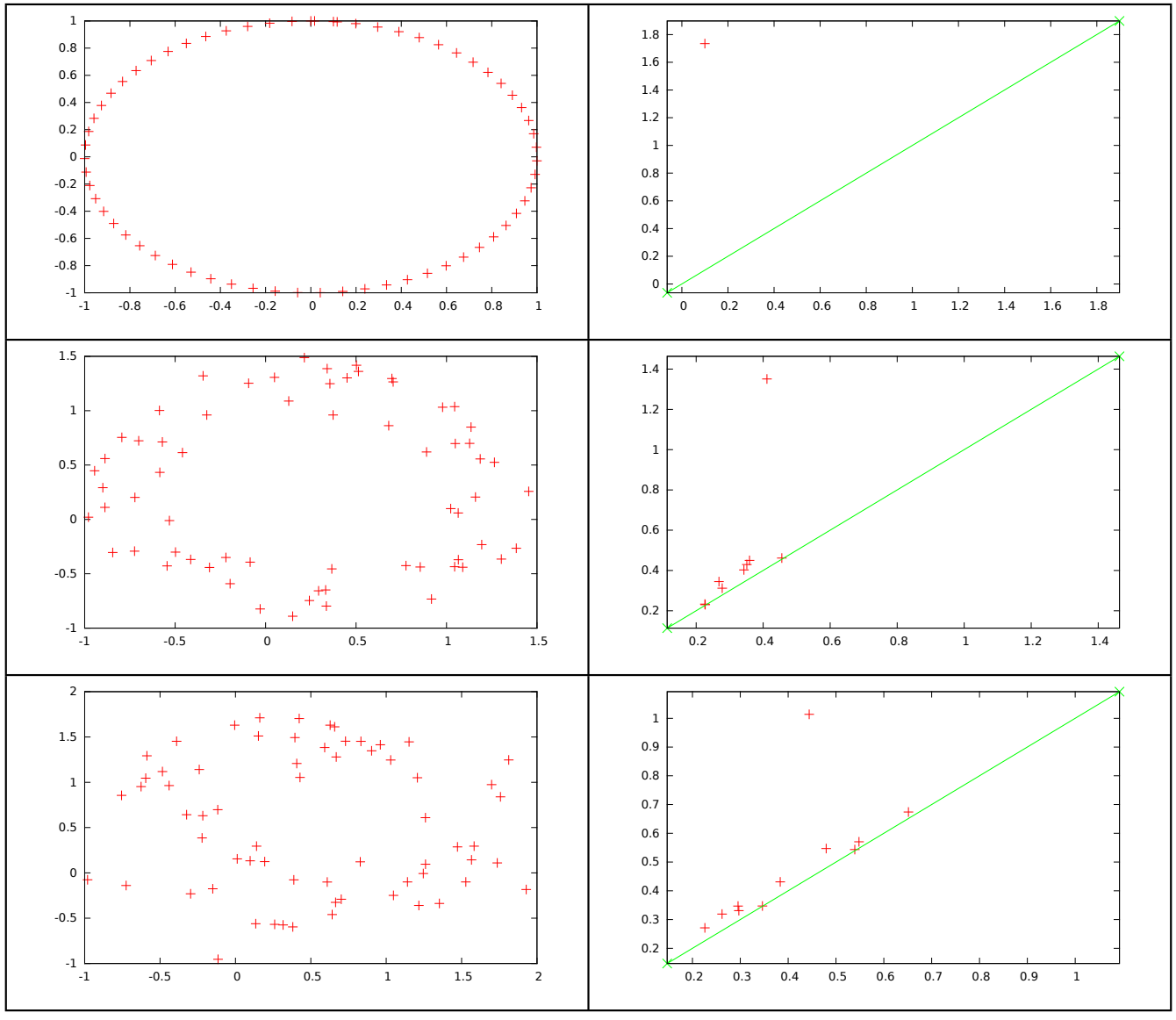
Lots of small circles, or a large one?



Multiscale.

- Persistent homology is a rigorous way of quantifying closed *shapes*,
- ... like connected components, cycles, voids and more.
- No matter if they are embedded in two or million dimensional space,
- No matter if they are rotated, stretched or transformed in any other way.
- Multi-scale,
-

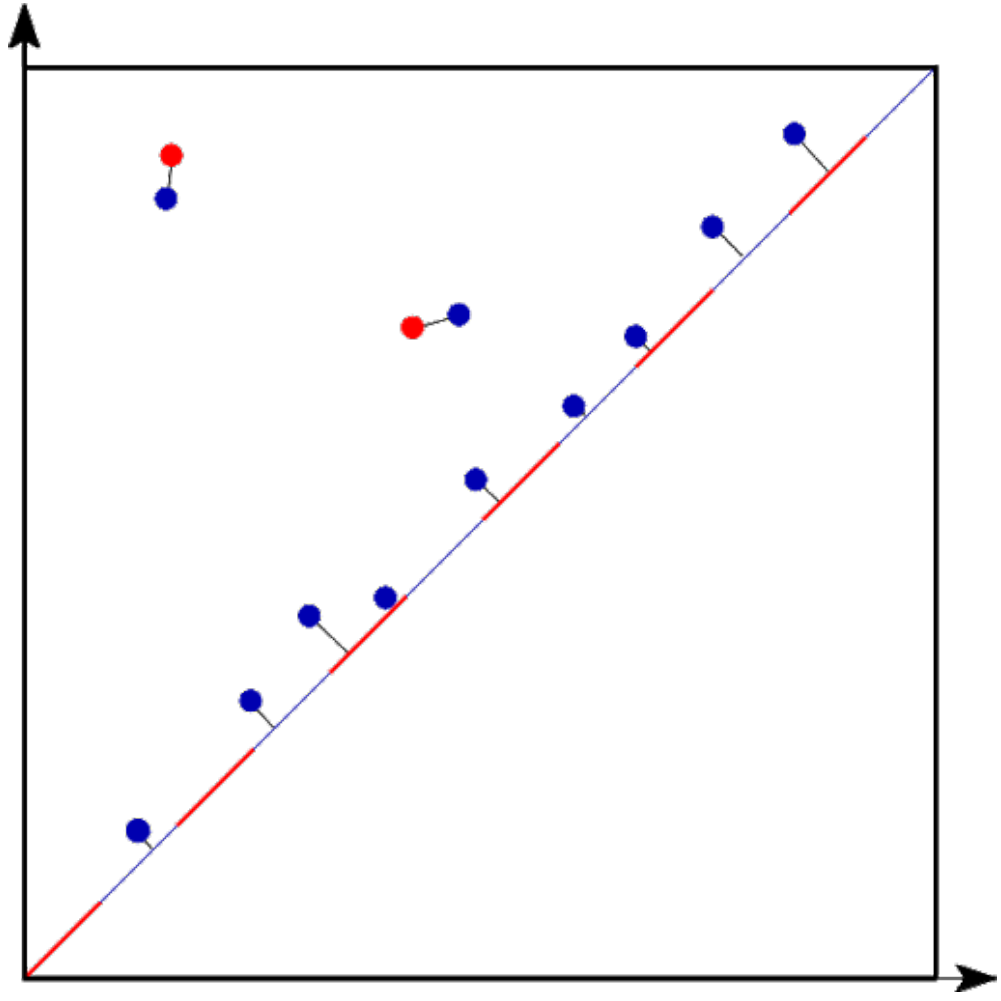
Robustness.


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8
1.5
1
0.5
0
-0.5
-1
-1 -0.5 0 0.5 1 1.5
1.4
1.2
1
0.8
0.6
0.4
0.2
0.2 0.4 0.6 0.8 1 1.2 1.4
2
1.5
1
0.5
0
-0.5
-1
-1 -0.5 0 0.5 1 1.5 2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Robustness.

- Persistent homology is a rigorous way of quantifying closed *shapes*,
- ... like connected components, cycles, voids and more.
- No matter if they are embedded in two or million dimensional space,
- No matter if they are rotated, stretched or transformed in any other way.
- Multi-scale,
- Robust.

Distances between diagrams.



Optimal matchings between points of two persistence diagrams allow us to define standard distances between them – bottleneck (length of longest edge in matching) and p-Wasserstein (sum of lengths of matching lines to the power q) to the power $\frac{1}{q}$.

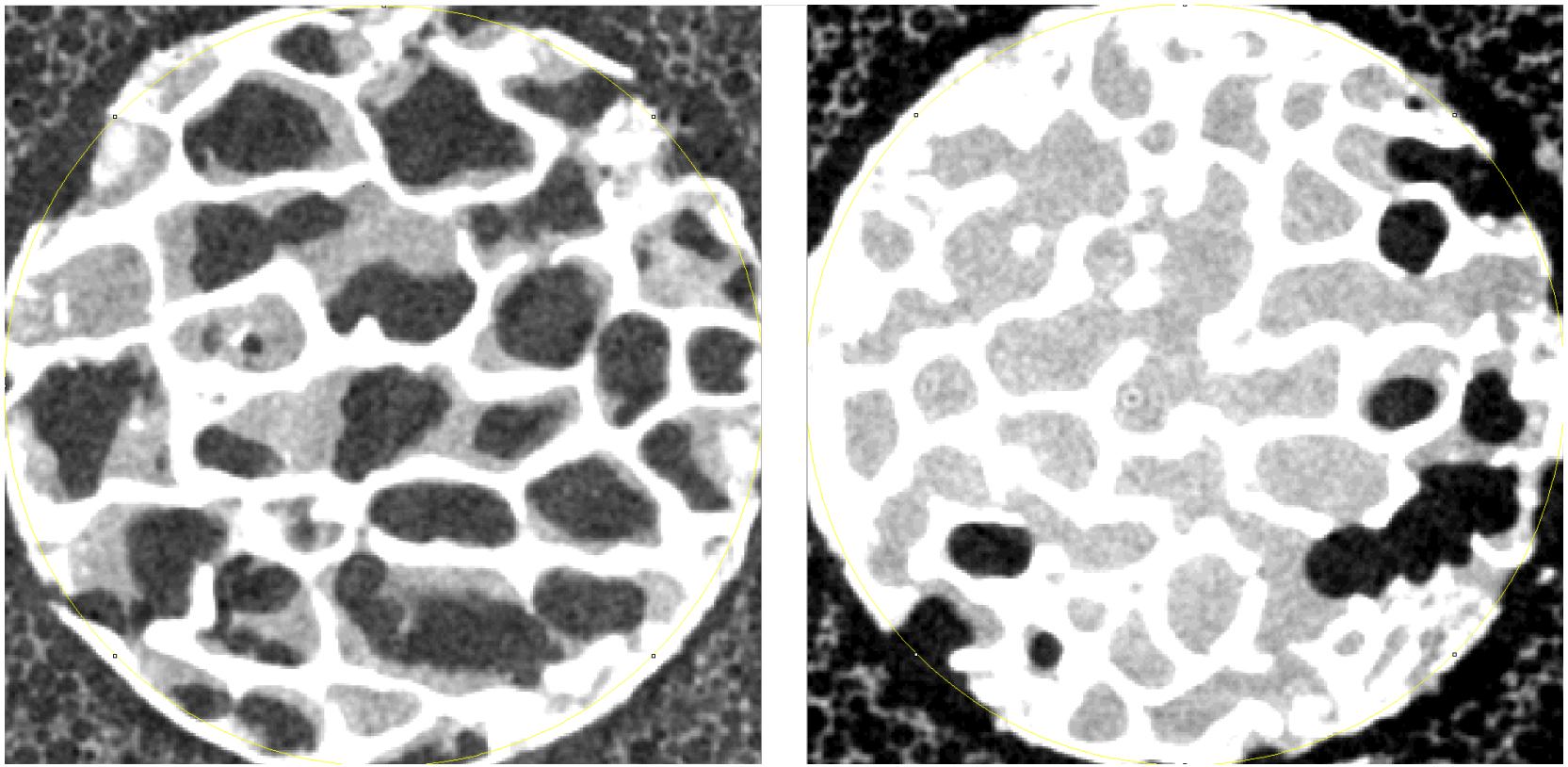
Practical exercise 2.

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

Not only point clouds....

- If you work with:
  - Pixel / voxel / cubical data,
  - Time series,
  - Correlation and similarity measures,
  - ...
- you may still use similar ideas and track connected components and holes emerging and disappearing.

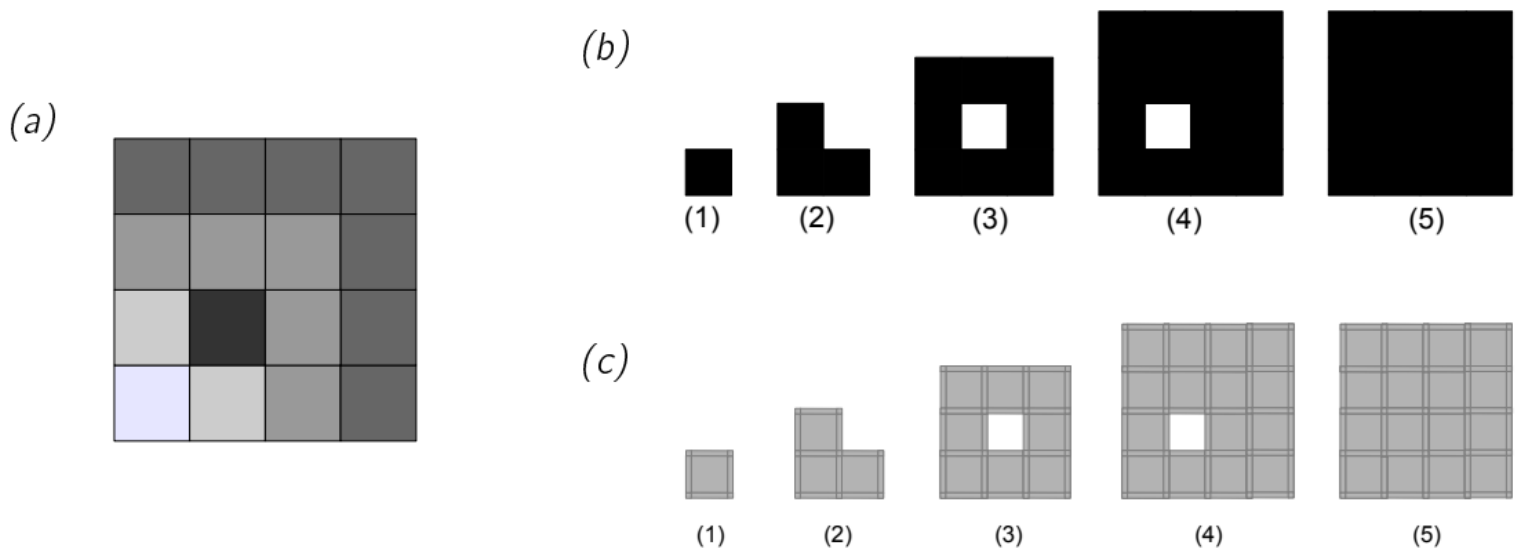
Apply to digital images.



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

What is a cubical persistence?

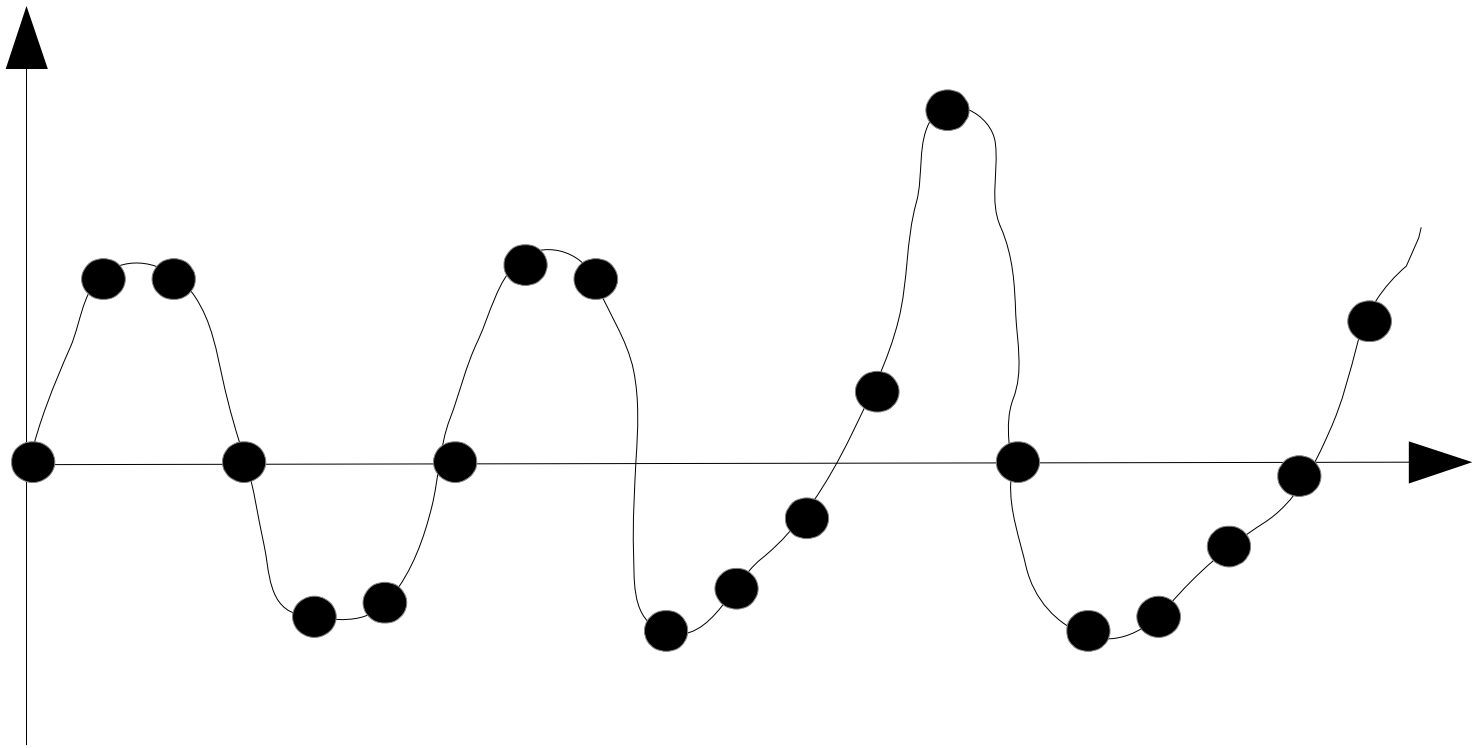
- Sub-level sets of function.
- Cubes enter from lower to highest function/filtration value.
- We track changes in homology of sub-level sets.



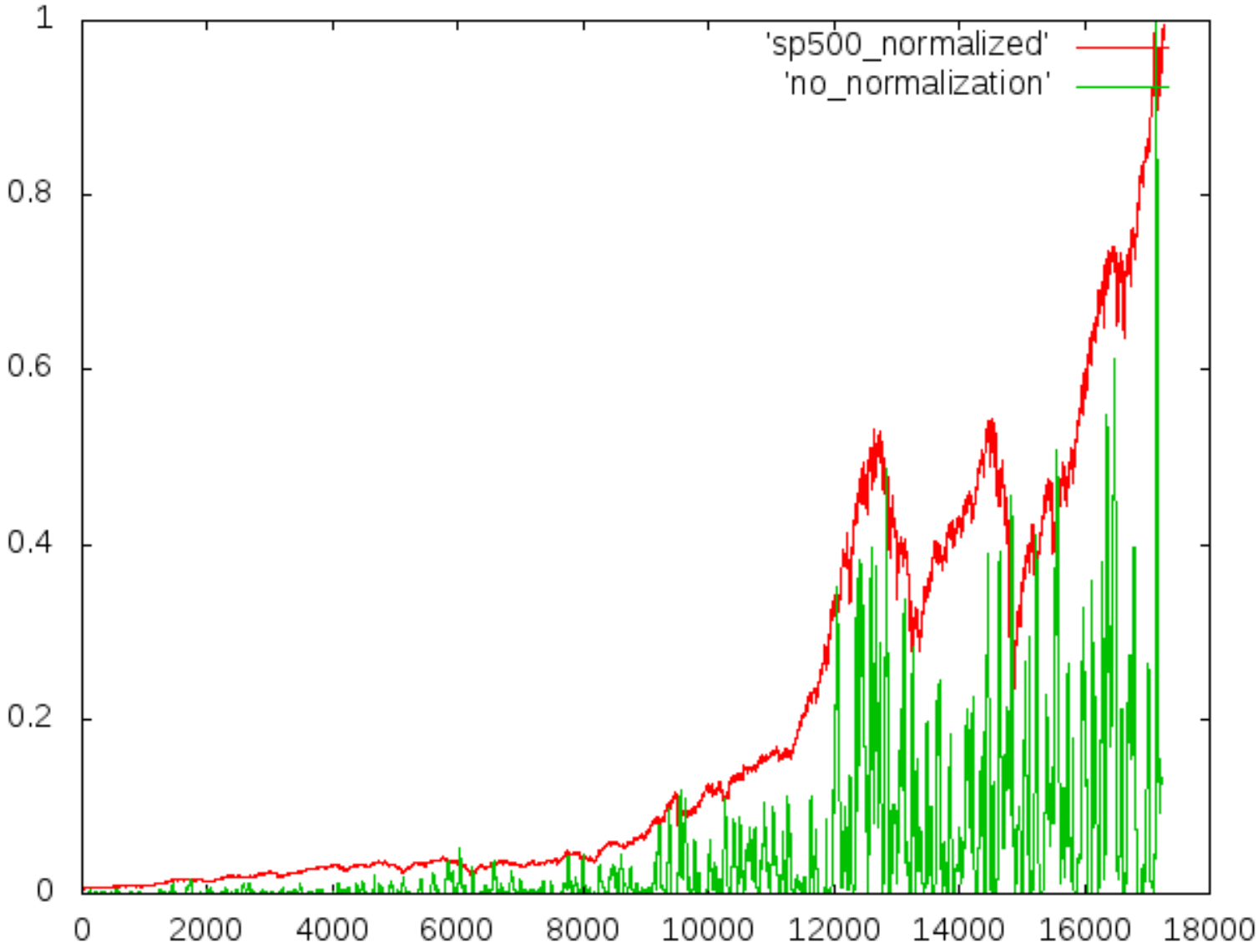
Practical exercise 3.

- Digital images are partially-constant discretization of functions.
- Let us go back to our exercises.
- Open Distance_from_circle.ipynb,
- In this exercise we will construct a cubical approximation of a function $f : [-2, 2]^2 \to \mathbb{R}$. $f(x, y)$ is a distance from (x, y) to a unit circle $x^2 + y^2 = 1$.
- Let us visualize it as an image, and let us compute persistent homology of the corresponding cubical complex.

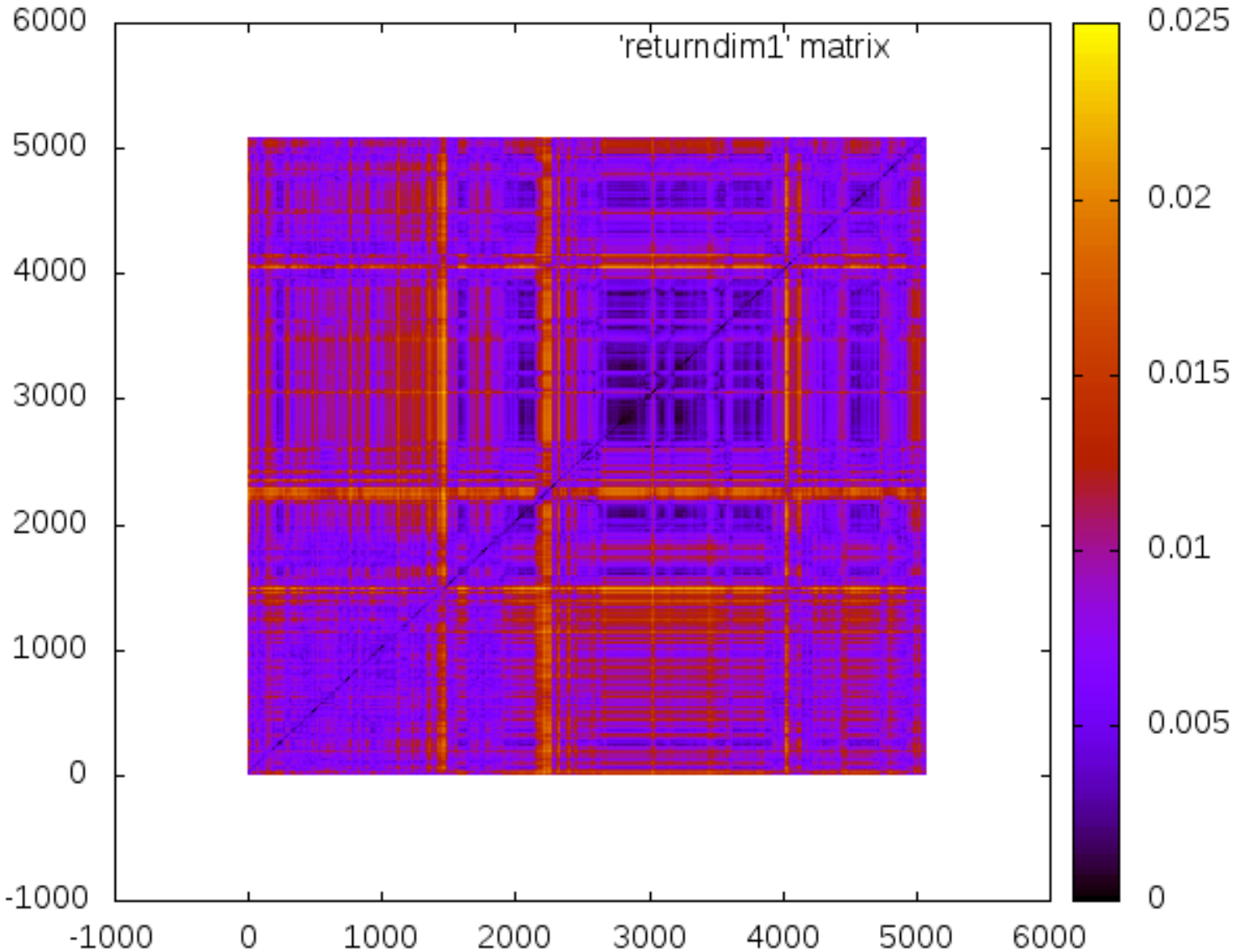
Persistence for time series analysis.



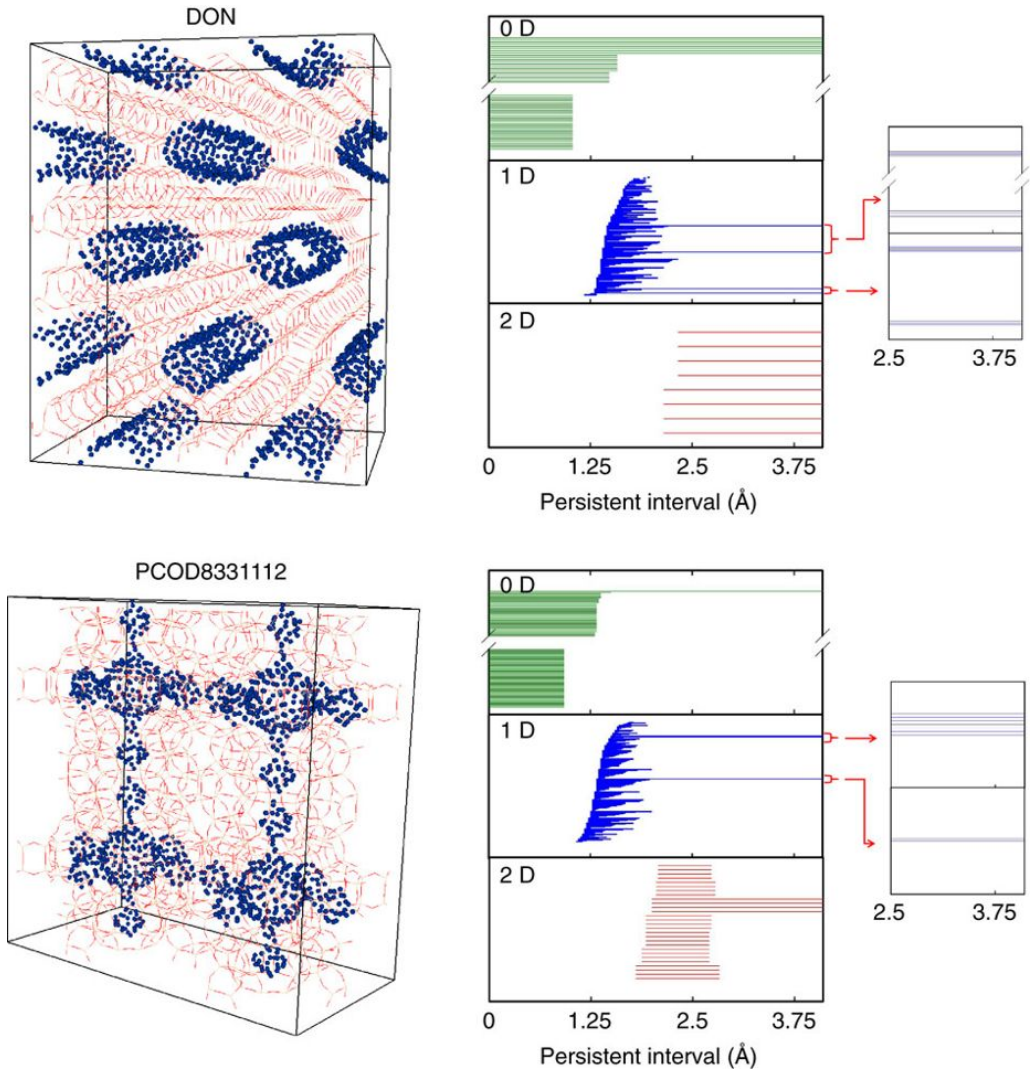
S&P-500 and crashes.



Persistence of correlations or similarity measures.



Persistence–based descriptors of nanoporous materials.



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

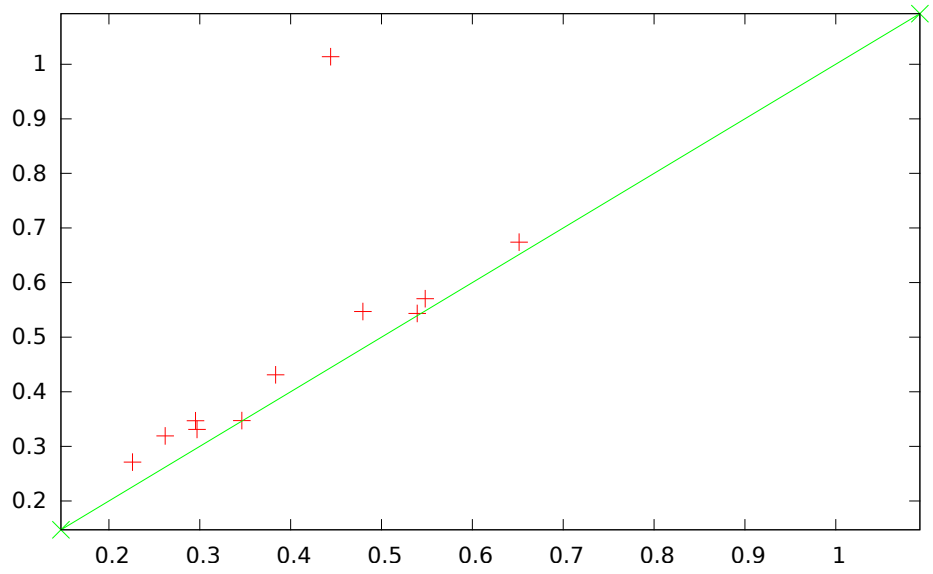
And more...

- 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...

Persistent homology.

- We have robust,
- multi scale,
- coordinate–free,
- compressed,
- tool to detect connected components, cycles, voids and their generalizations.
- It can be interfaced in various ways with standard stat. and ML tools.

Persistent homology, the output.



- Muti set of points in $\mathbb{R}^2$.
- Variable size, not ideal representation to interface with ML/AI and statistics $\rightarrow$ persistence representations, embeddings, ...
- We need to embed persistence diagrams into a Hilbert space (vectorize them).
- That makes topological/statistical inference - hypothesis testing, confidence intervals,... possible.

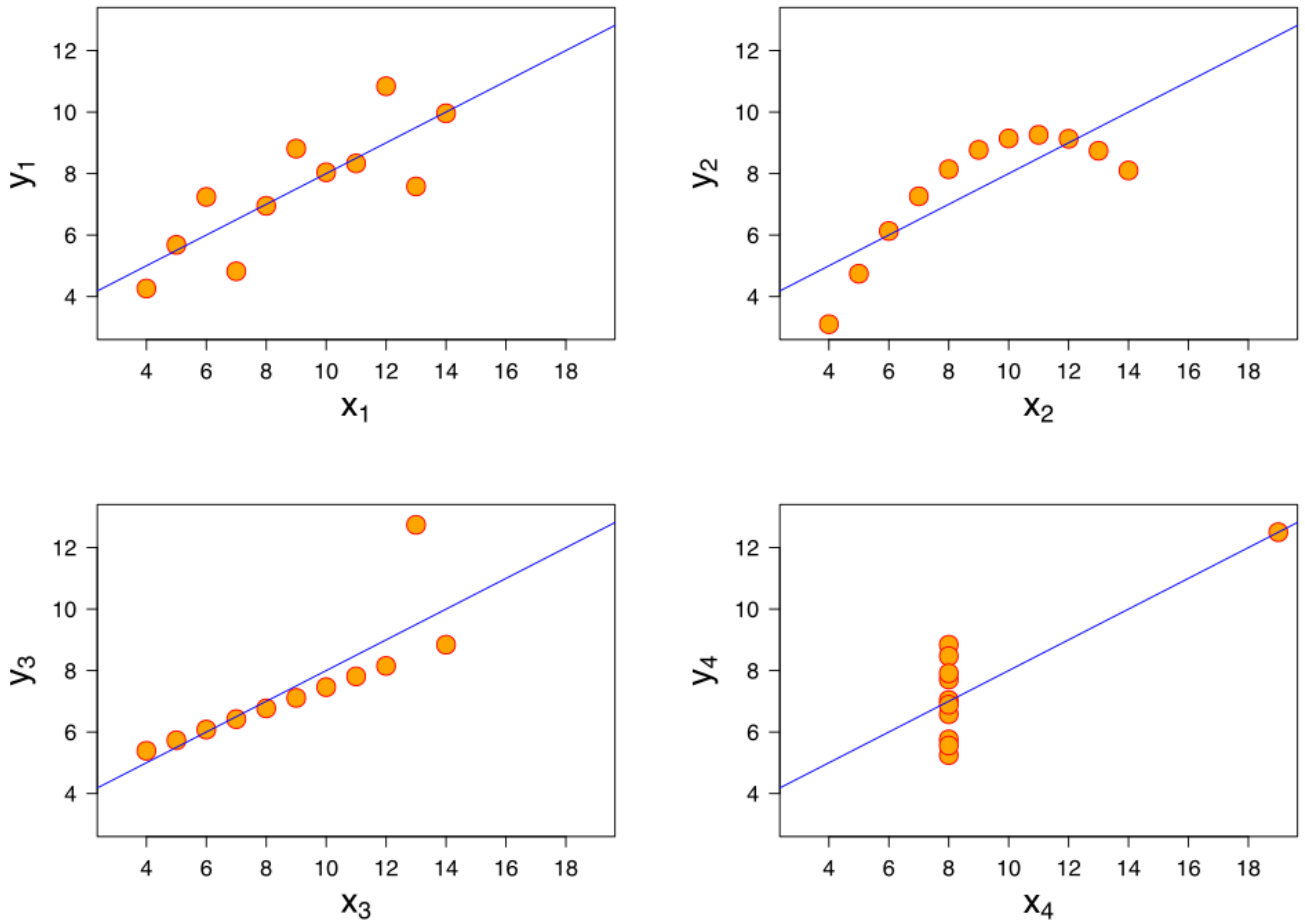
Practical exercise 4.

- This is the last exercise in persistent homology.
- Open classification_of_distributions.ipynb,
- In this exercise we will consider point clouds sampled from two dimensional normal distributions with different averages and covariance matrices.
- We will use persistence homology as their signatures,
- And attempt to classify them using machine learning tools based on those signatures.

Topology and statistics. Together.

- Statistics provide a vast collection of tools to summarize properties of point clouds.
- However, there are numerous examples (line Anscombe's quartet and Datasaurus dataset presented below) of point clouds with the same descriptive statistics, but very different shape.
- This is why, in statistics, we should always *visualize* considered dataset.
- This is however not possible to visualize high dimensional data.
- This is where tools from topology came into rescue – topological tools we discuss in this tutorial allow us to quantify if two datasets may, or may not, have similar shape.

Anscombe's Quartet.



Same statistics, different shapes

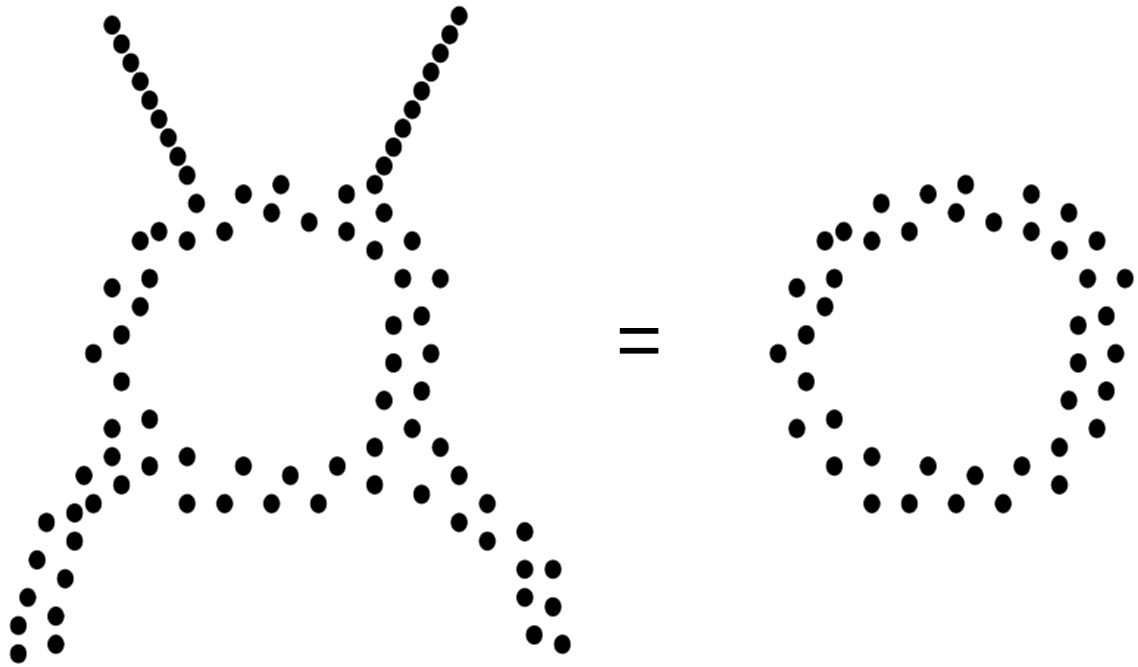
Datasaurus Dataset.

Same statistics, different shapes

Homology and persistent homology, biased collection of resources.

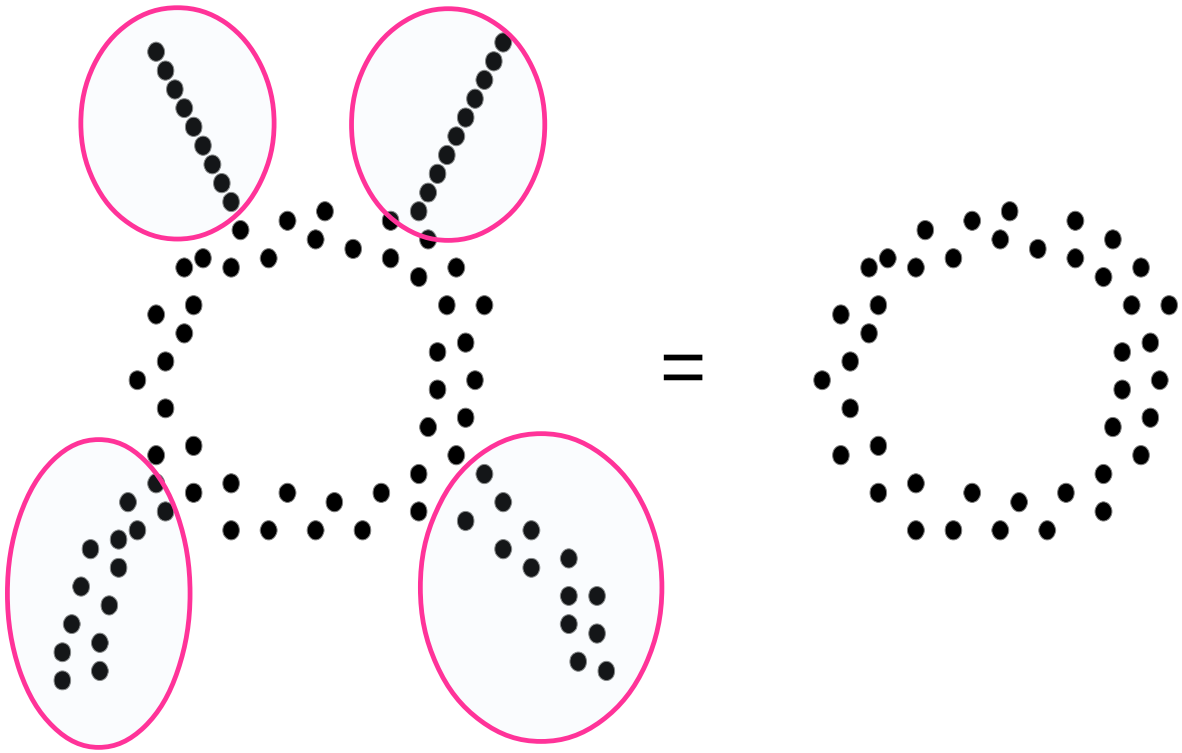
- Edelsbrunner and Harer, Computational Topology, An Introduction, AMS.
- Kaczynski, Mischaikow, Mrozek, Computational Topology, Springer 2003.
- Dłotko, Applied and Computational Topology, Tutorial
- Multiple youtube videos.

Persistence is nice, but, what about flares?



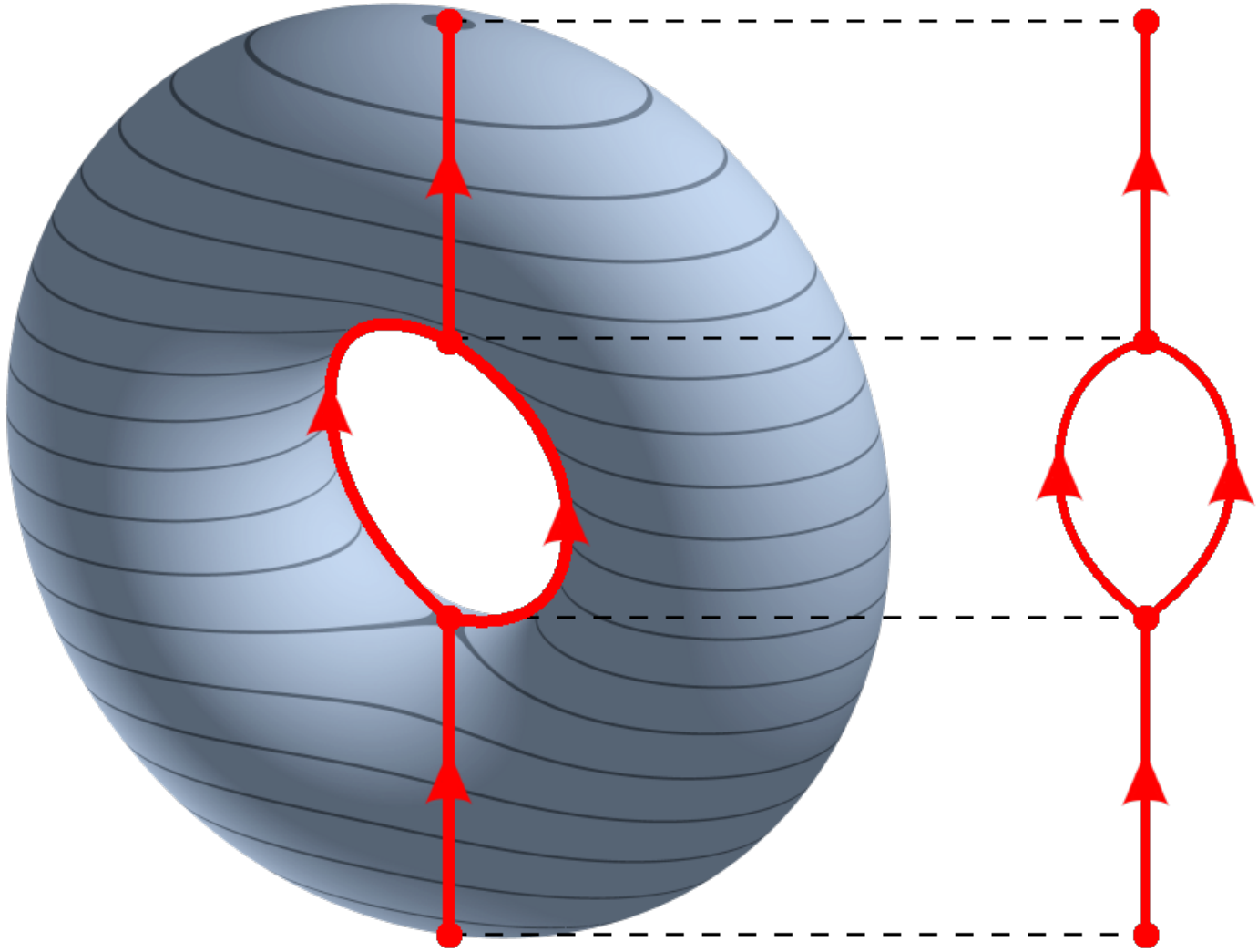
Persistence homology of those two point clouds will be very similar, as they both have one connected component and one hole.

But, what about flares?



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

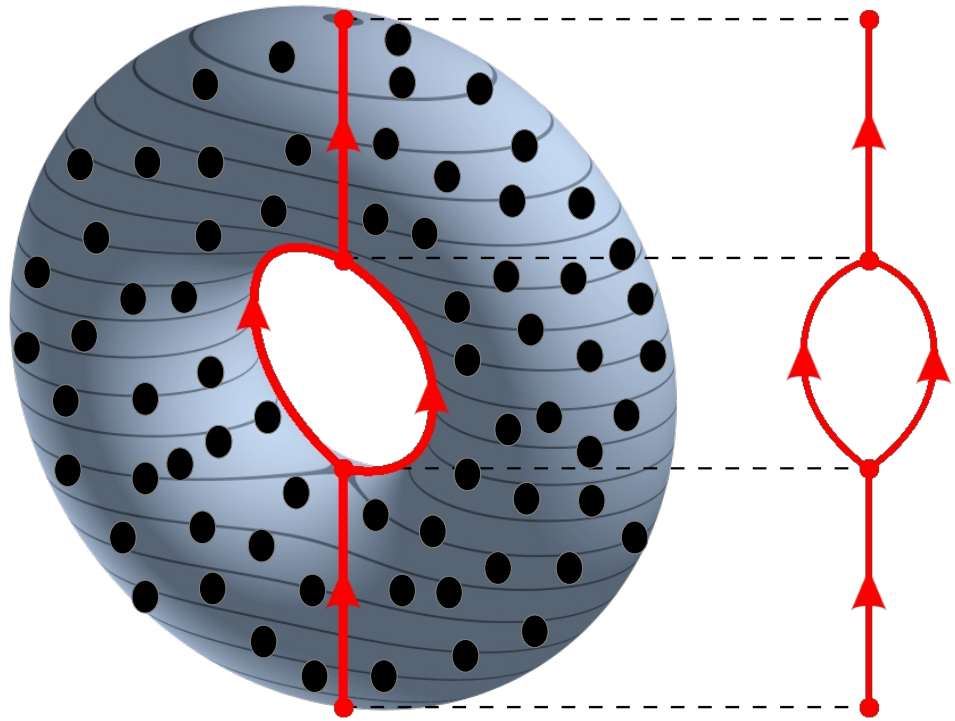
Reeb graph.



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.

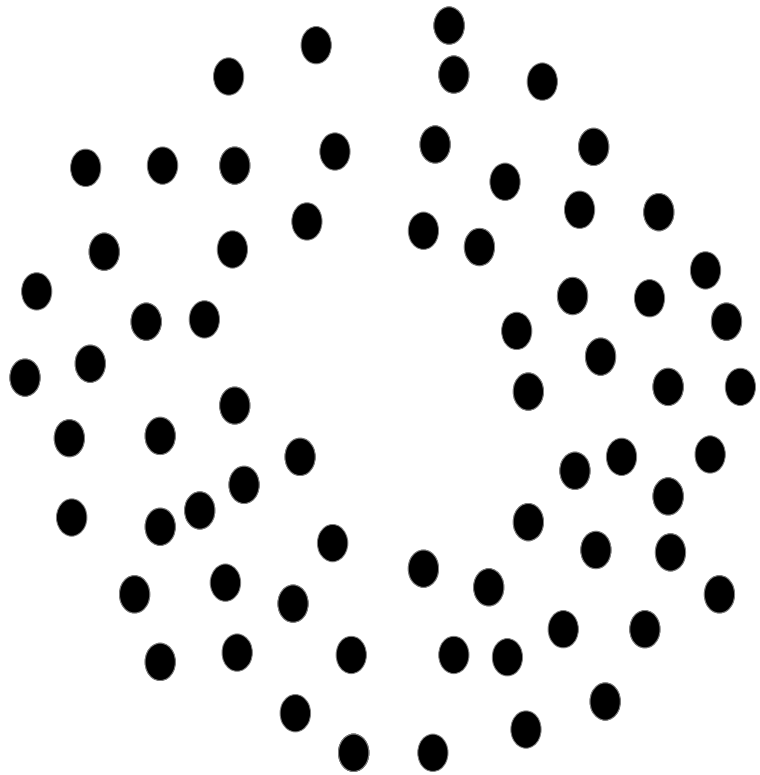


Conventional mapper graph is an attempt to define Reeb graph for discrete point cloud instead of a manifold.

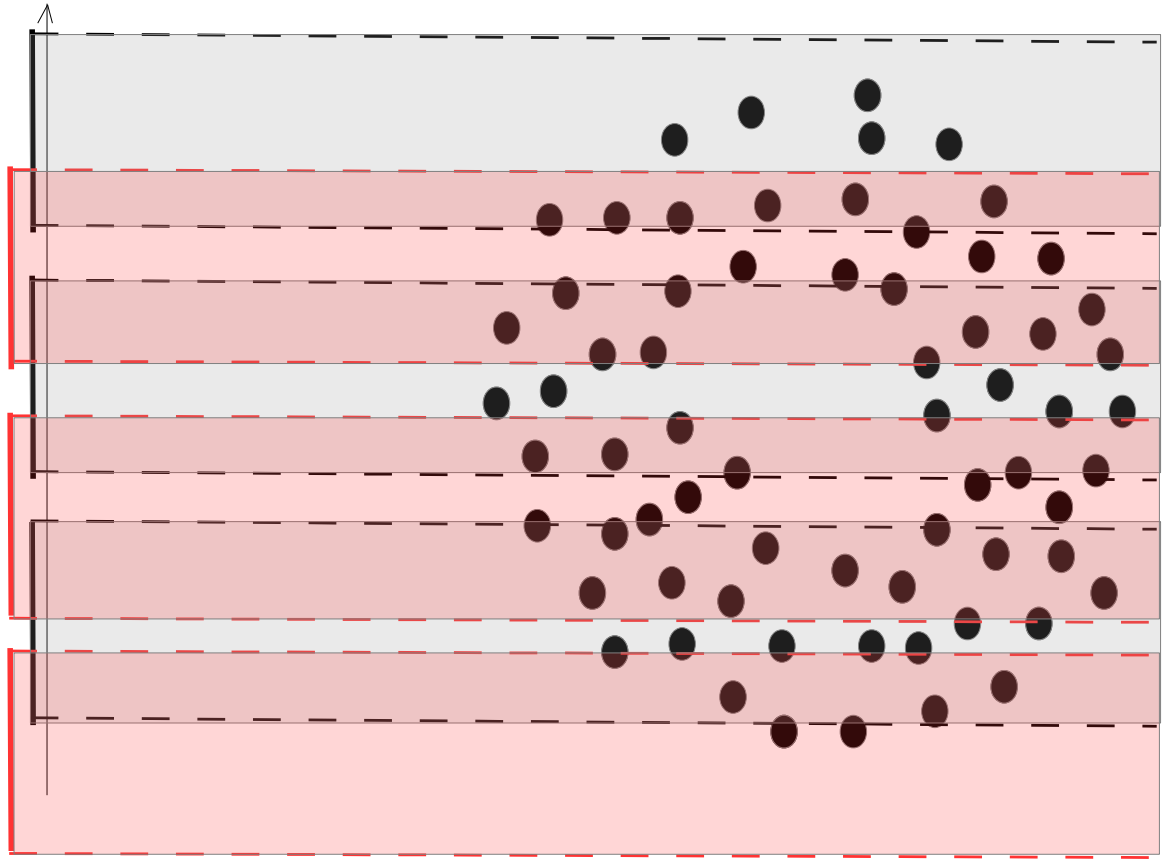
Mapper algorithm, idea.

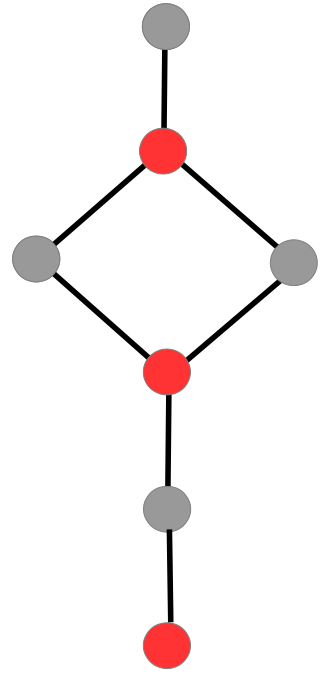
- Input: finite collection of points sampled from M, $f : M \to \mathbb{R}$.
- We define a relation x R y iff:
  - $f(x)$ is close to $f(y)$,
  - x and y belong to the same cluster ...

Conventional Mapper algorithm.



Conventional Mapper algorithm.





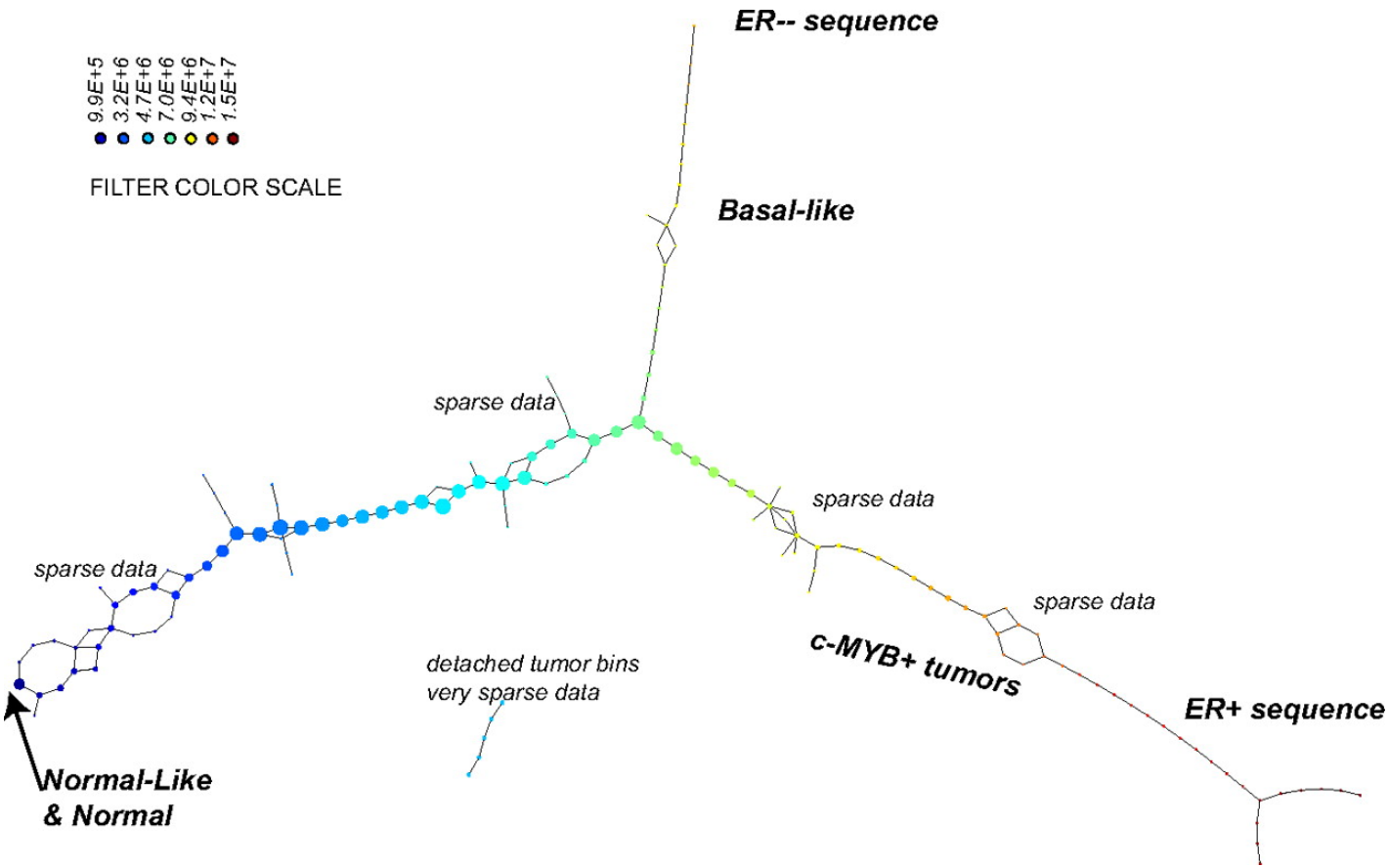
Mapper algorithm, formally.

- Input: finite collection of points sampled from M, $f : M \to \mathbb{R}$.
- Cover of the range of f with overlapping boxes.
- Clustering algorithm
- We define a relation x 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 a value of an objective function.
- 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/Paris_TDA_Tutorial_2021.html`, download exercises in Standard Mapper.
- Let us start from something simple – open standard_mapper_concentric_circles.ipynb
- 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.ipynb
- 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!

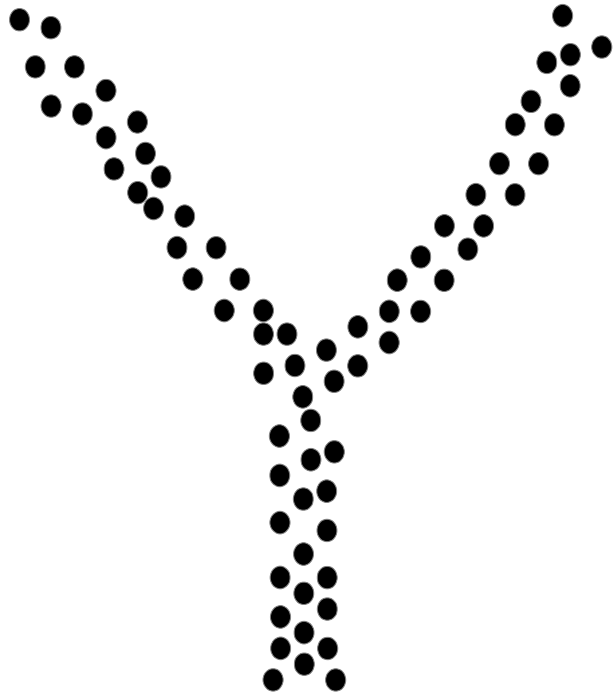
Practical exercise 3.

- This is the last exercise on conventional Mapper.
- Please open standard_mapper_two_dimensional_projection_of_digits.ipynb
- Here we want to understand the structure of the space of hand written digits.
- We will use umap-learn library to project each 8×8 digit into $\mathbb{R}^2$ – note that this time lens function have range in $\mathbb{R}^2$, which is not a typical scenario.
- Once again, we will be able to choose from a number of clustering methods.
- The obtained mapper graph will be colored by the label (which digits we consider).

Ball Mapper algorithm.

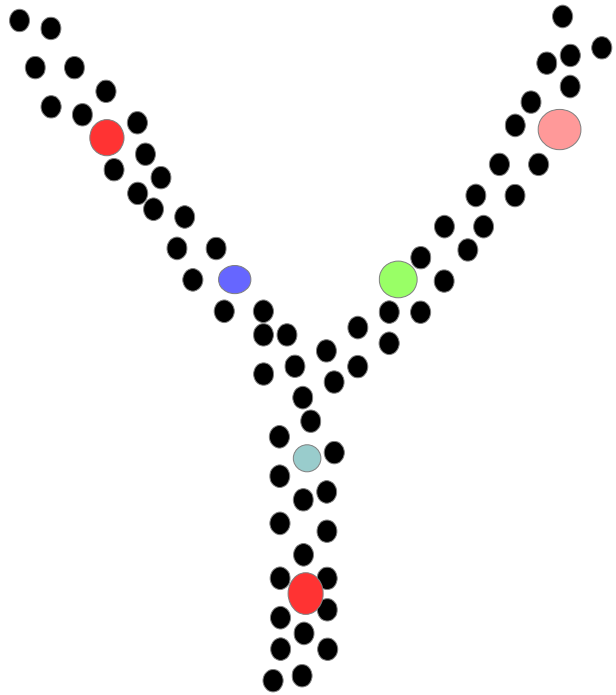
- As a last part of our schedule, we will play with Ball Mapper algorithm.
- Thanks to Davide Gurnari, it is now available both in R and in Python - you are free to pick up the language to proceed!
- As you might have noticed, it is not always trivial to choose the *lens function* as well as *clustering algorithm* in standard Mapper construction.
- The idea of Ball Mapper is intuitively explained in the following slides.

Ball Mapper algorithm.



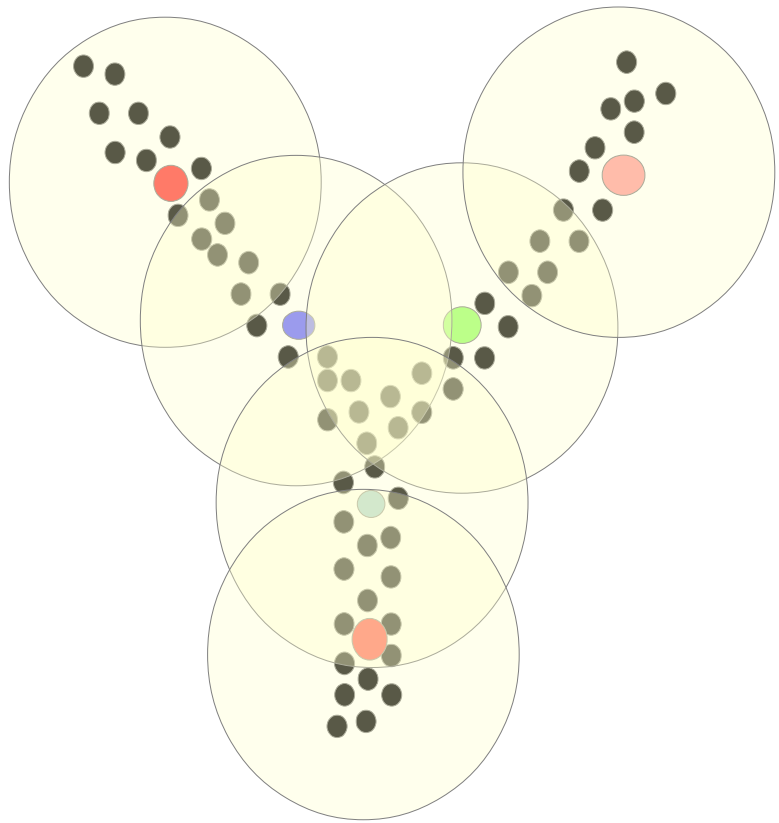
Take a point cloud X

Ball Mapper algorithm.



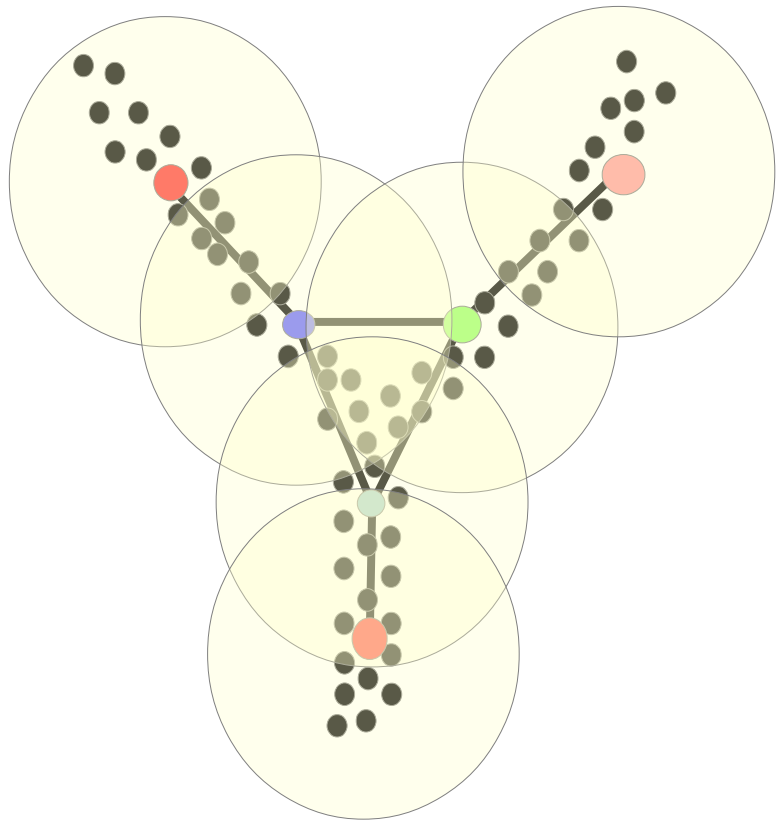
Given $\epsilon > 0$, select subset of points $N \subset X$ such that for every $x \in X$ there exist $n \in N$ such that $d(x, n) \leq \epsilon$ (we call N an ϵ-net)

Ball Mapper algorithm.



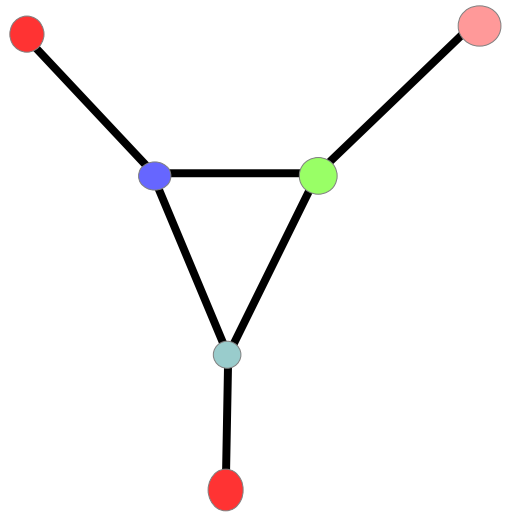
Consequently $X \subset \bigcup_{n \in N} B(n, \epsilon)$, i.e. $\{B(n, \epsilon), n \in N\}$ cover X.

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 install R (I recommend RStudio) and open basic_circle.R or use analogous Jupyter Notebook.
- 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 Boston_property.R
- 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.

Practical exercise 3.

- Our last example is based on UK Census data. We will try to understand phenomena behind Brexit referendum in 2016.
- Please open Brexit_example.R
- The dataset contain the 2011 census data with coloration coming from 2016 Brexit referendum and the results of 2017 elections.
- For more detailed political interpretation please visit `https://arxiv.org/abs/1909.03490`

Thank you for your time.

Dioscuri Centre in Topological Data Analysis
@Facebook

DIOSCURI
CENTRE IN TOPOLOGICAL DATA ANALYSIS

Jointly sponsored by

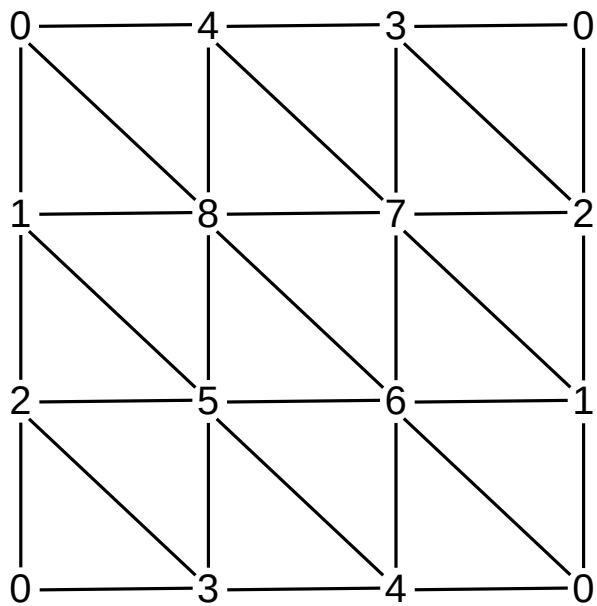
Paweł Dłotko
pdlotko @ impan.pl
pdlotko @ gmail
pawel_dlotko @ skype

Practical exercise, continuous development.

- This is an additional part of a tutorial.
- We will add here some additional exercises aiming to modify some stuff that we did already.
- They will require some modification of the notebooks that we have used.

Practical exercise, projective plane.

- We will start by playing again with intro_to_homology.ipynb, but this time we will play with more abstract space – a *projective plane*.
- Here is the triangulation.
- Please adjust the triangulation and compute the homology groups.



Practical exercise, figure eight.

- In the persistence_simple_point_cloud.ipynb we have computed persistent homology of a point cloud sampled from a circle with and without noise.
- Please modify it so that this time we sample points from a figure 8 point cloud.

Practical exercise, Betti numbers percolation.

- In this exercise we will generate a number of random cubical complexes of a size $N \times N$.
- Each 2 dimensional cube will be set to 1 with probability p and 0 otherwise.
- Please check how the Betti numbers evolves as p varies between 0 and 1. Does this phenomena look something like a phase transition?

Practical exercise, classification.

- Download MNIST dataset, consider only digits 0 and 1.
- Compute their cubical persistent homology.
- Run classification.
- Check if it also works well for other pair digits, for instance 1 and 2.
- Why this is the case?

Practical exercise, exploration with mappers.

- Visit UC-Irvine database `https://archive.ics.uci.edu/ml/index.php`
- Pick any numerical dataset, ideally with no missing values.
- Run Mapper and Ball Mapper algorithms on it. Use labels as coloring.
- Explore the results and try to draw conclusions.

Some solutions.

- Please note that the solutions to some of the questions are available at `https://dioscuri-tda.org/Paris_TDA_Tutorial_2021.html` and download solutions to extra exercises.
- Please however make an attempt to solve it by yourself before moving to it!