Topological Adventures in Machine Learning

Applied Machine Learning Days
28 January 2020

EPFL

Blue Brain Project


shape
deformation
connectivity
continuous
path
cavity
invariants
geometry
connected
donut
simplex
classification
open
topology
proximity
mug
equivalence
homology
algebra
closed
continuity
complex

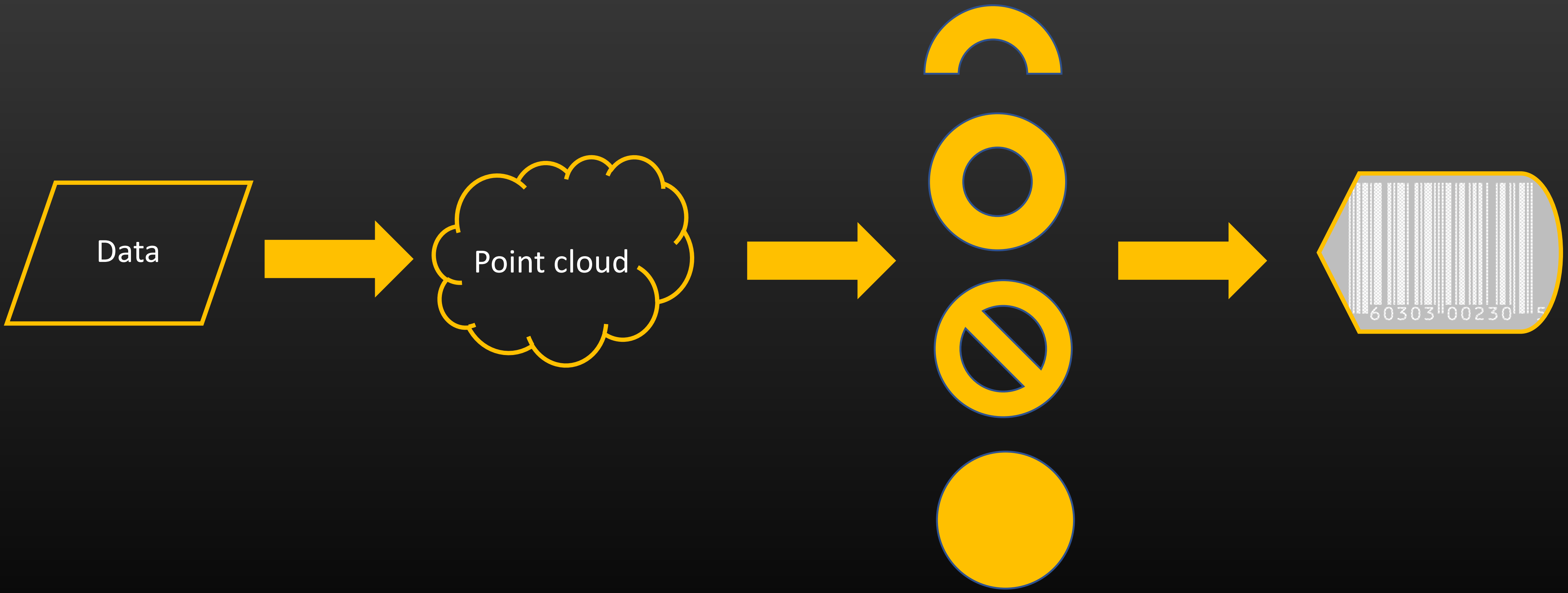

DATA ANALYSIS
PSYCHIATRIC
HELP 5¢
TOPOLOGIST
THE DOCTOR
IS IN

Topological Data Analysis (TDA)

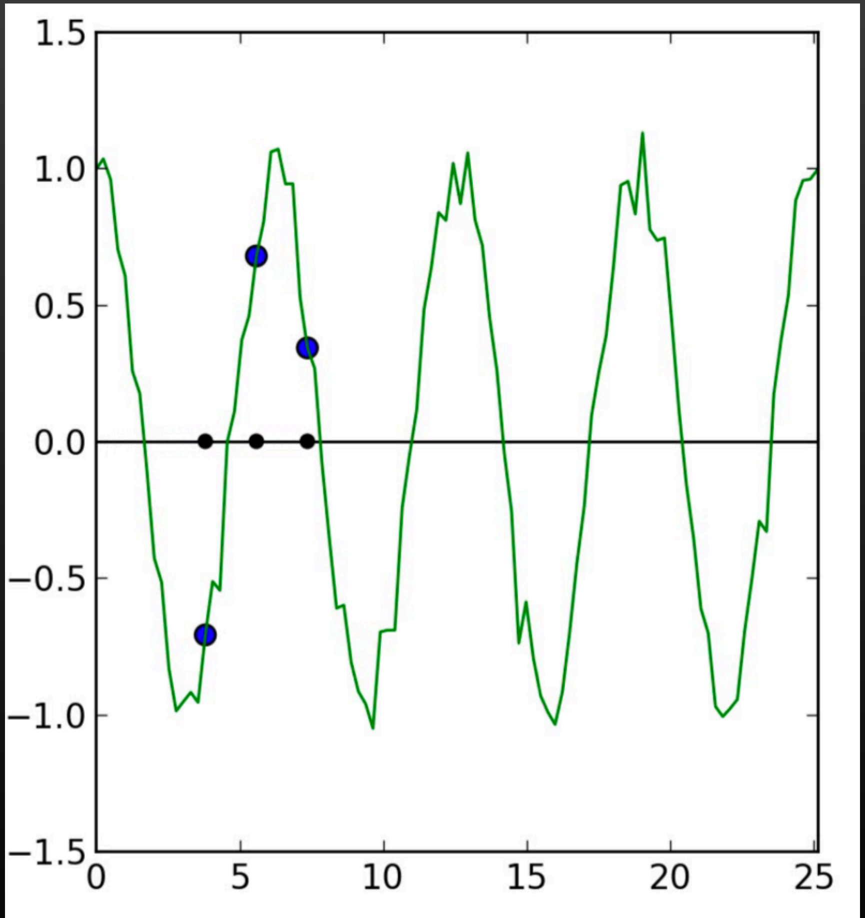
Guiding philosophy of TDA

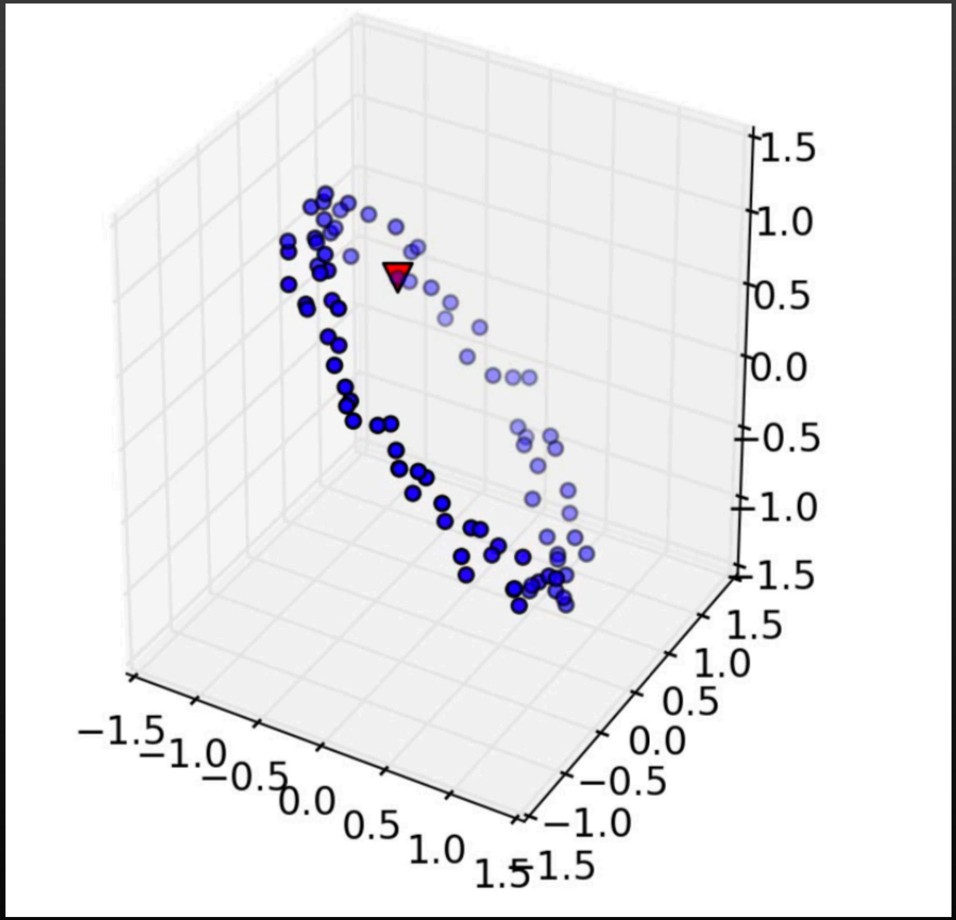
The shape of a data set, encoded by a topological signature, should reveal important relations among the data points with the help of machine learning.

The usual TDA workflow



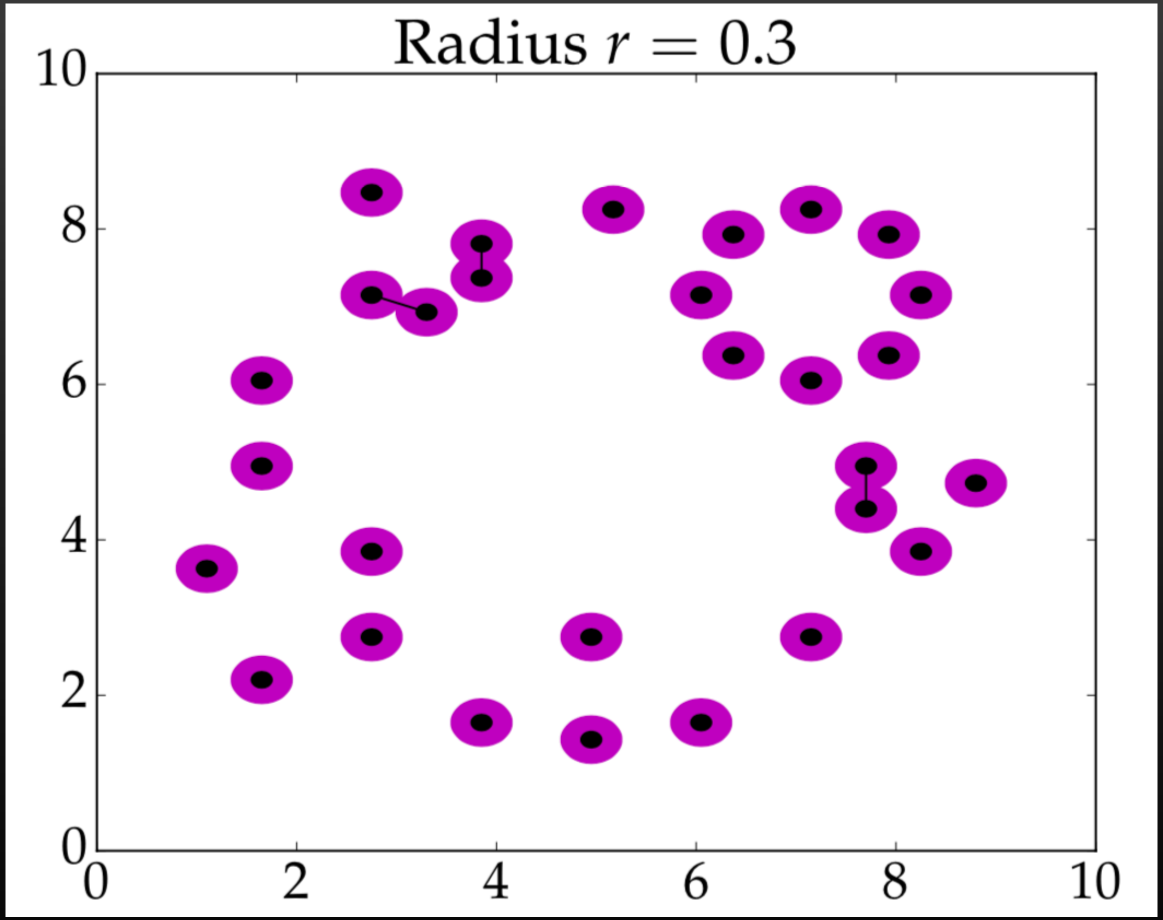
Step 1: Data to Point Cloud





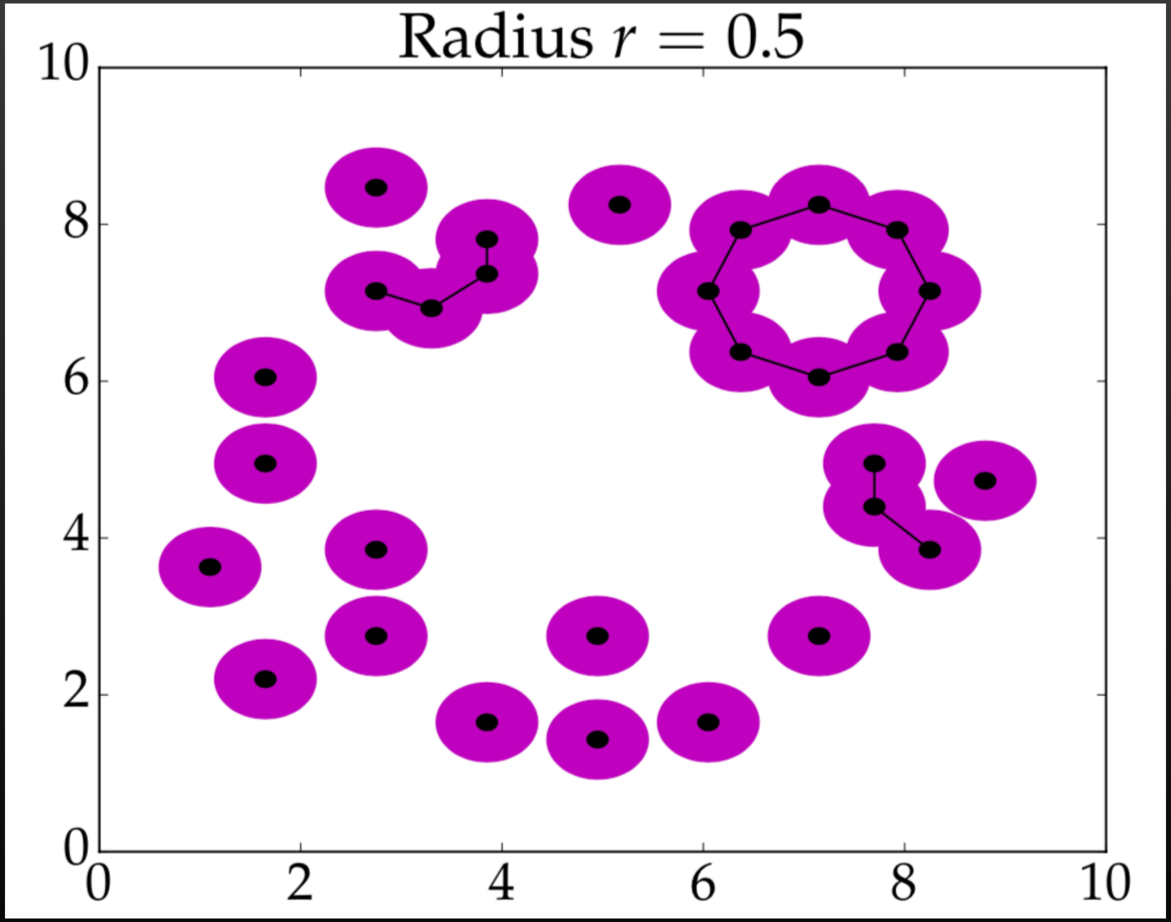
L. Munch, 2019.

Step 2: Point cloud to nested complexes



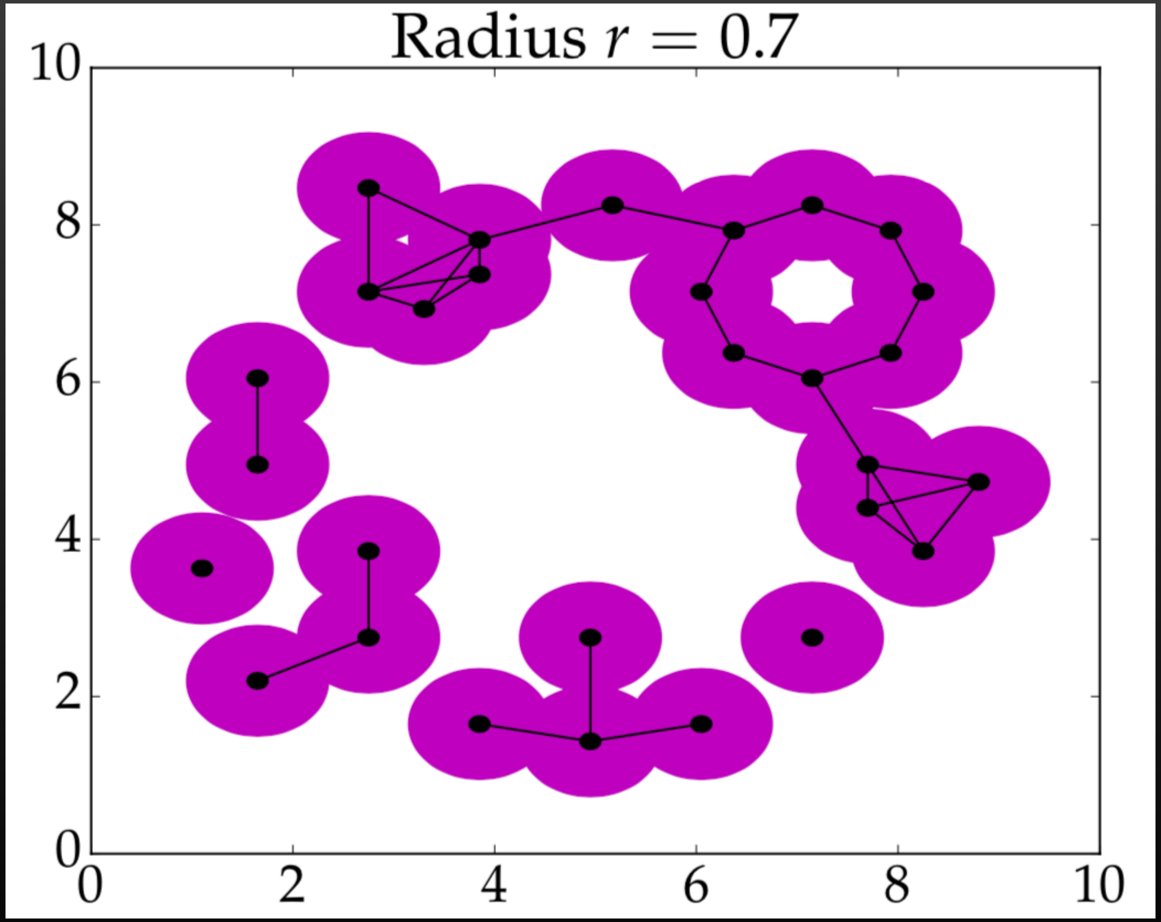
L. Munch, 2019.

Step 2: Point cloud to nested complexes



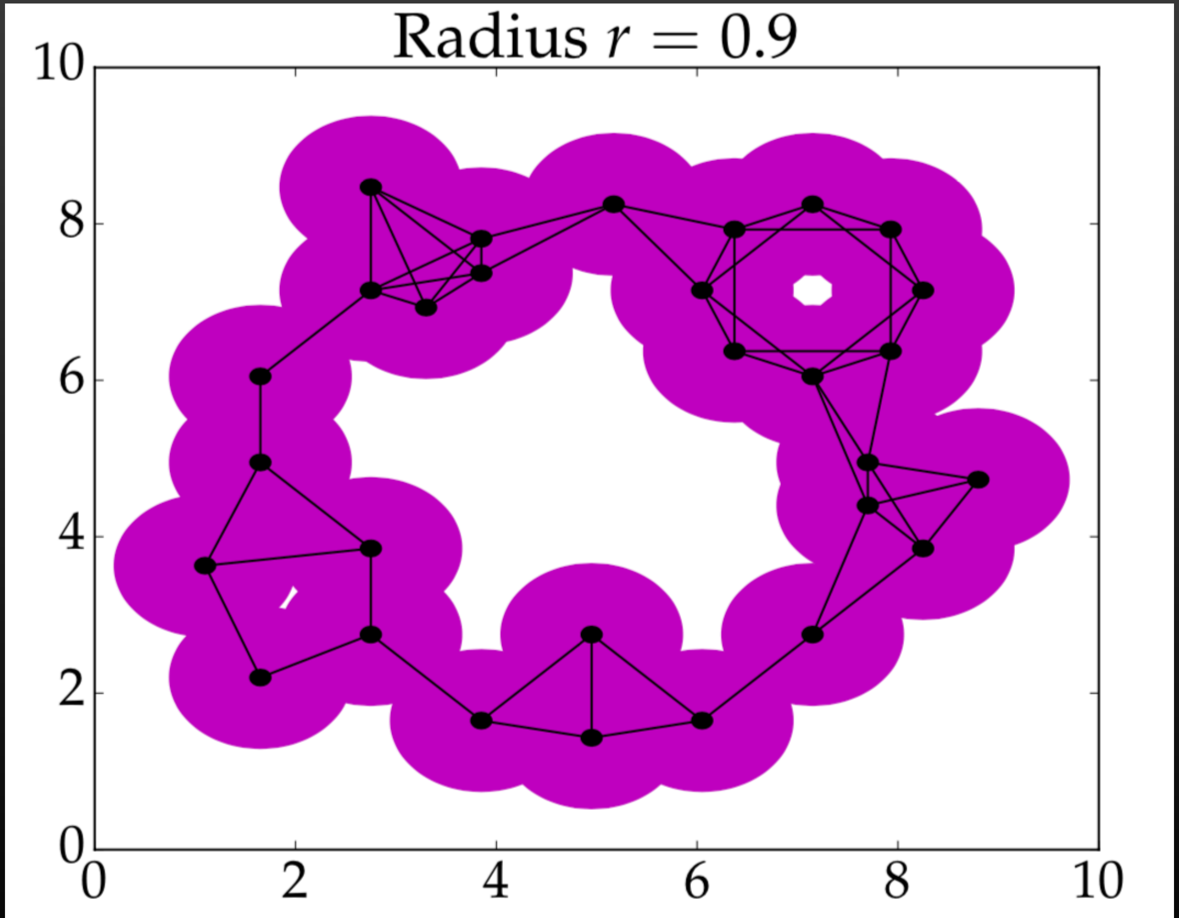
L. Munch, 2019.

Step 2: Point cloud to nested complexes



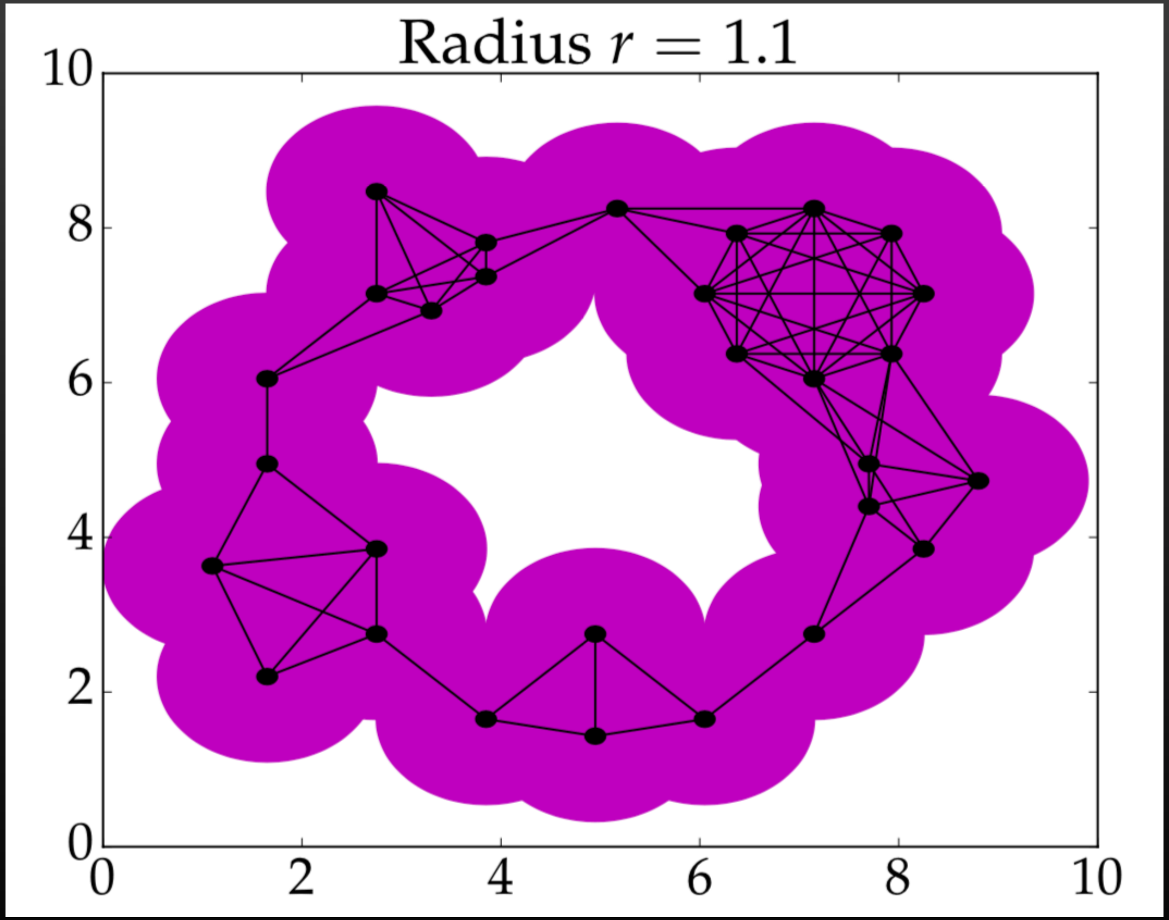
L. Munch, 2019.

Step 2: Point cloud to nested complexes



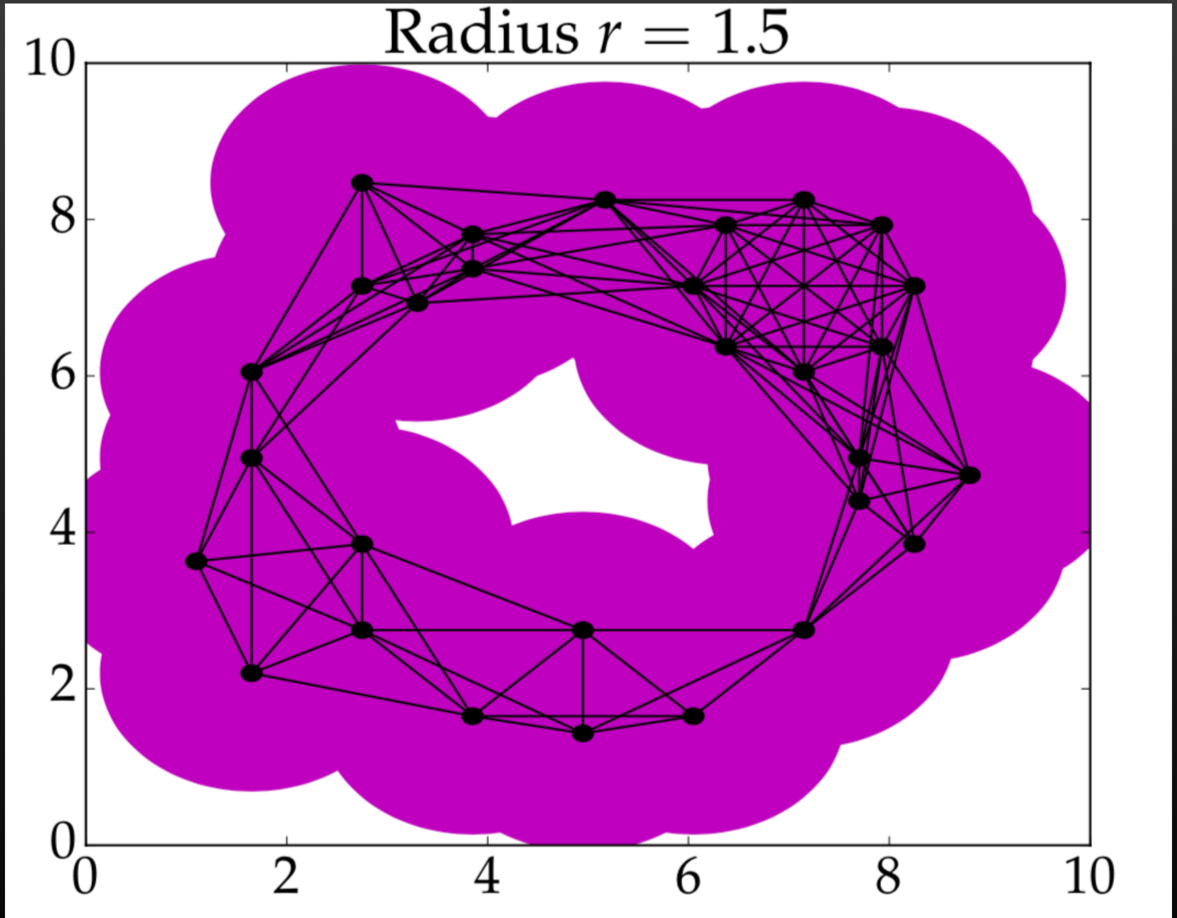
L. Munch, 2019.

Step 2: Point cloud to nested complexes



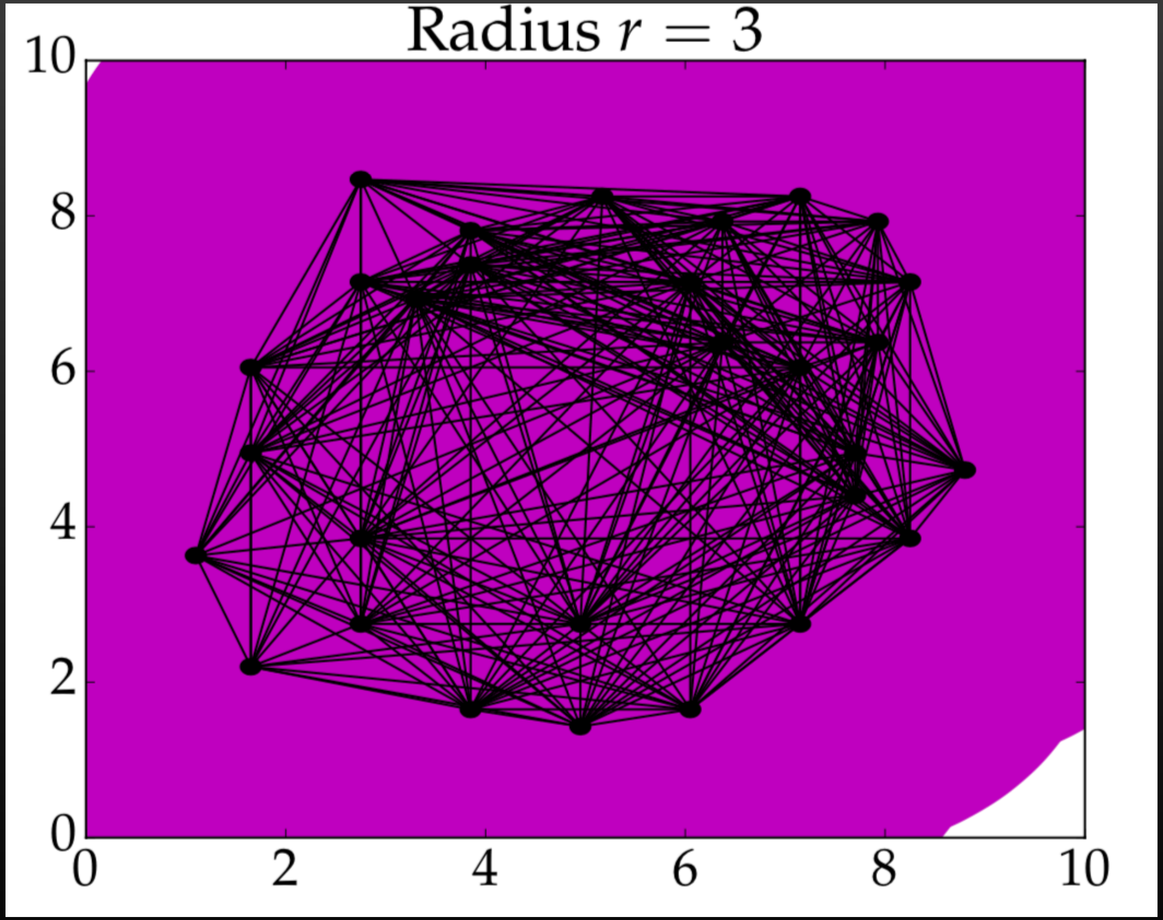
L. Munch, 2019.

Step 2: Point cloud to nested complexes



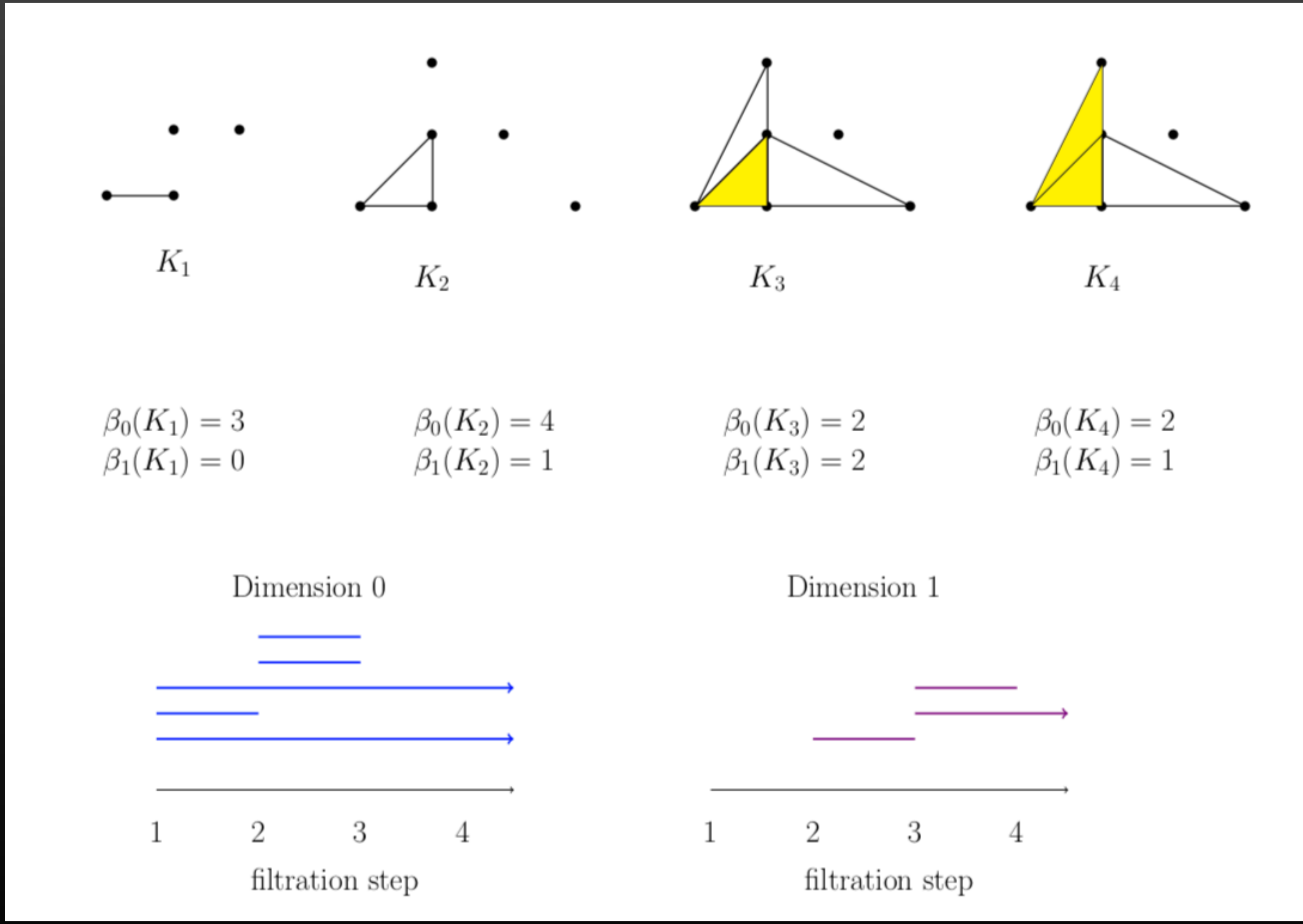
L. Munch, 2019.

Step 2: Point cloud to nested complexes



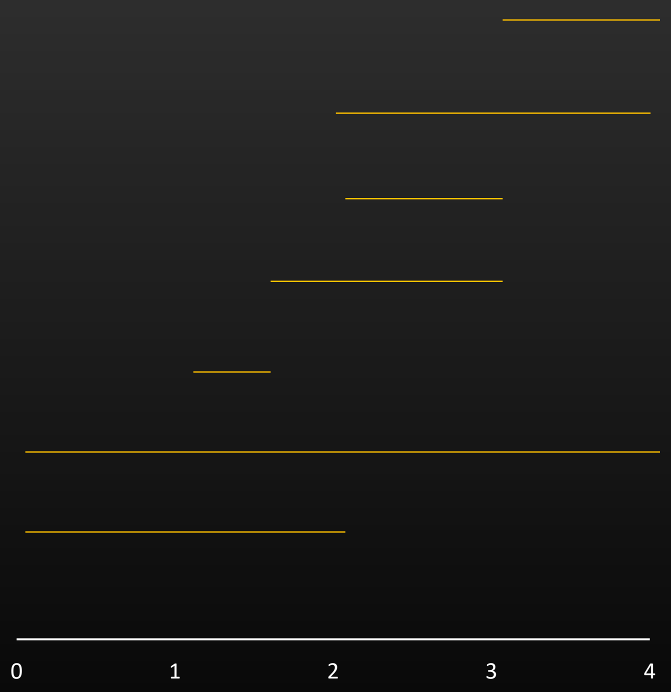
L. Munch, 2019.

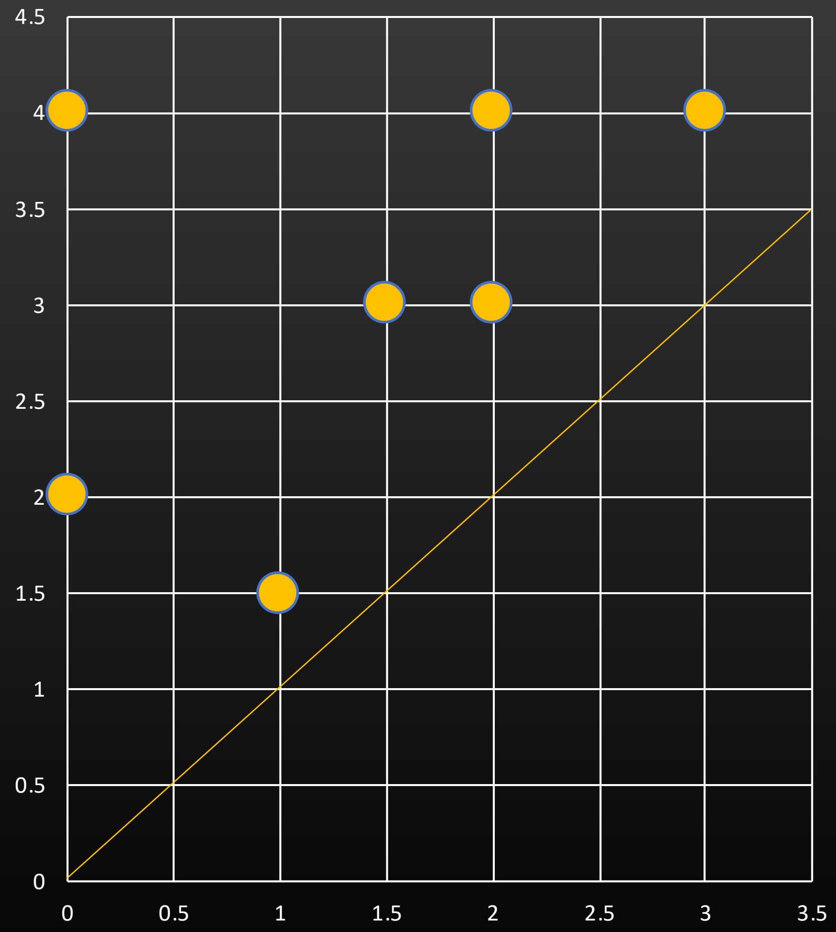
Step 3: Nested complexes to barcode



Otter et al., arXiv, 2016.

Barcodes vs persistence diagrams





Stability

- The set of barcodes/persistence diagrams can be equipped with a variety of earthmover-type distances: the Wasserstein distances of L_p-type and the bottleneck distance of L_∞ -type.

- Most reasonable known instantiations of the TDA pipeline are Lipschitz continuous with respect to Hausdorff distance on point clouds and bottleneck distance on persistence diagrams.

Practicalities

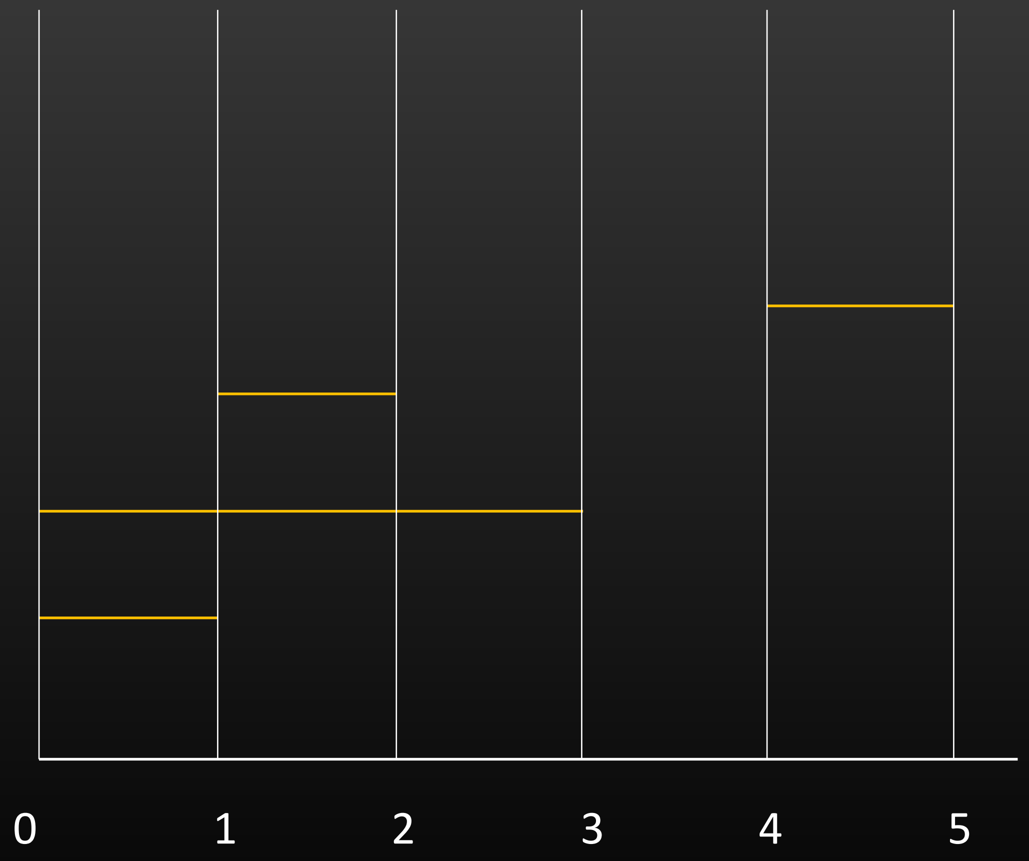
- There are extensive libraries of software, mostly open source, for TDA computations (e.g., GUDHI, Ripser, Flagser, Giotto,…).
- There exist “inverse analysis” tools for interpreting results of TDA computations (e.g., work of Hiraoka et al.).

From TDA to ML

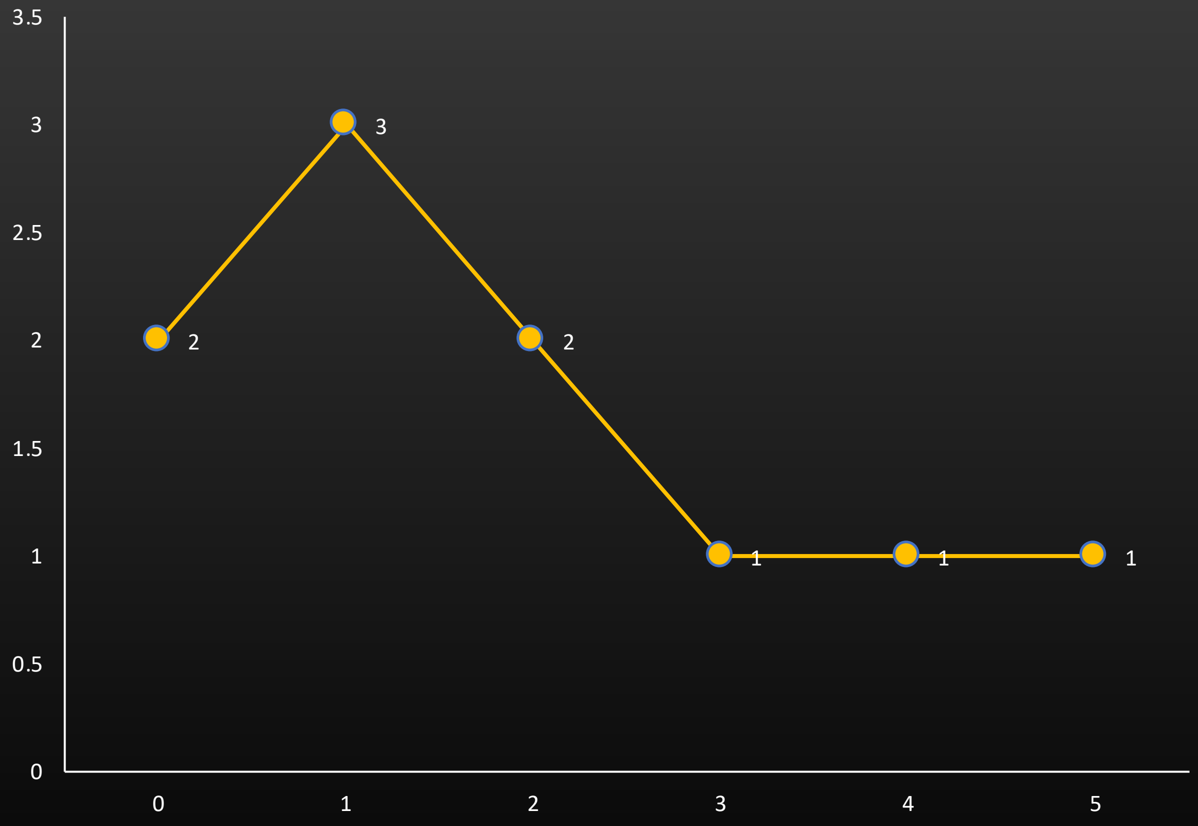
Strategies for featurization

- Problem: Cannot compute statistics in the space of barcodes or the space of persistence diagrams.
- Solution:
  - Define a Lipschitz-continuous mapping from the space of barcodes/persistence diagrams to a vector space $\mathcal{V}$ equipped with an inner product.
  - Compute statistics in $\mathcal{V}$!
  - [Leygonie-Oudot-Tillmann, 2019] New differentiable approach, enabling the use of gradient descent.

Betti curves

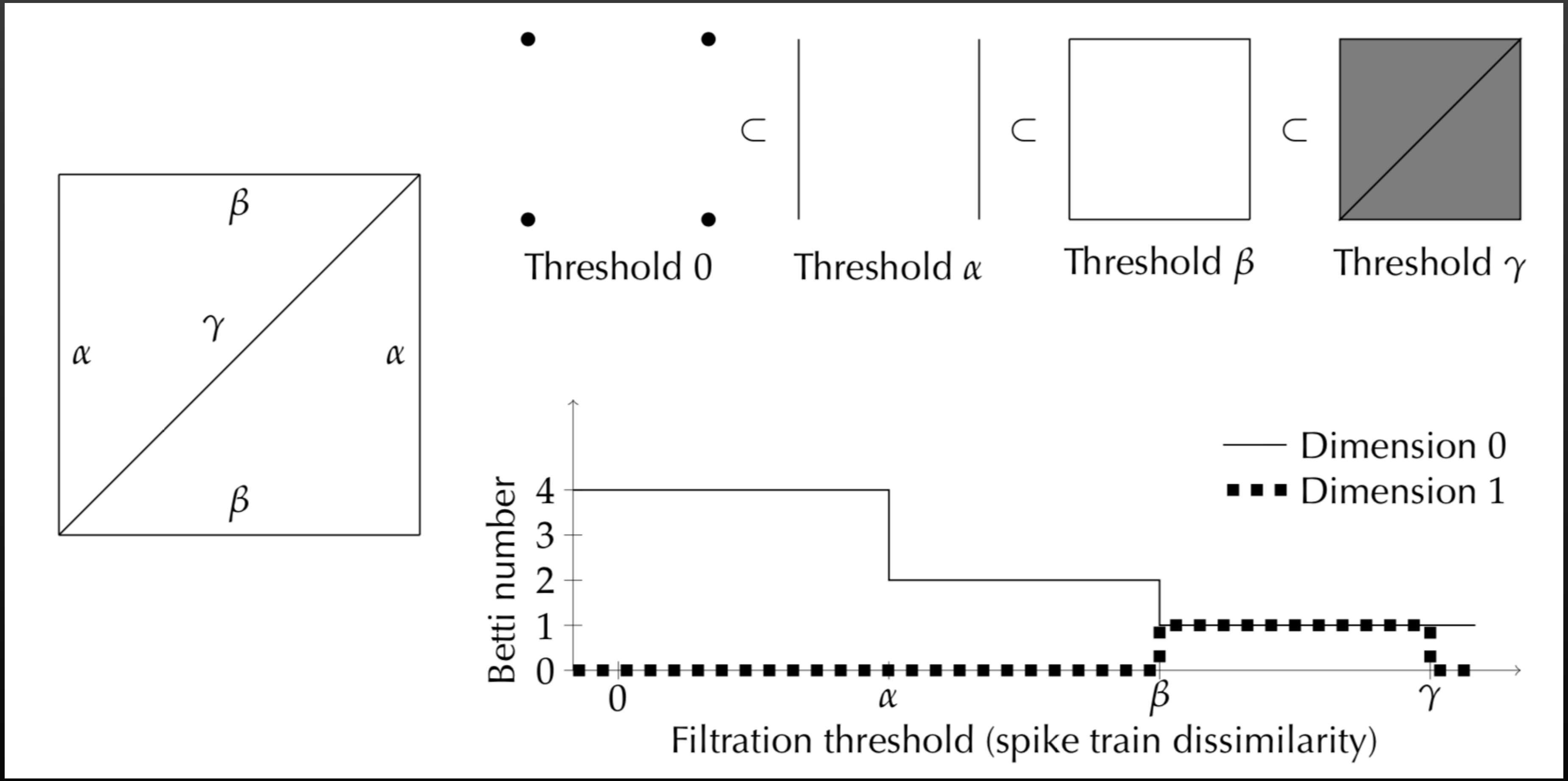


Bar code for cavities of dimension k



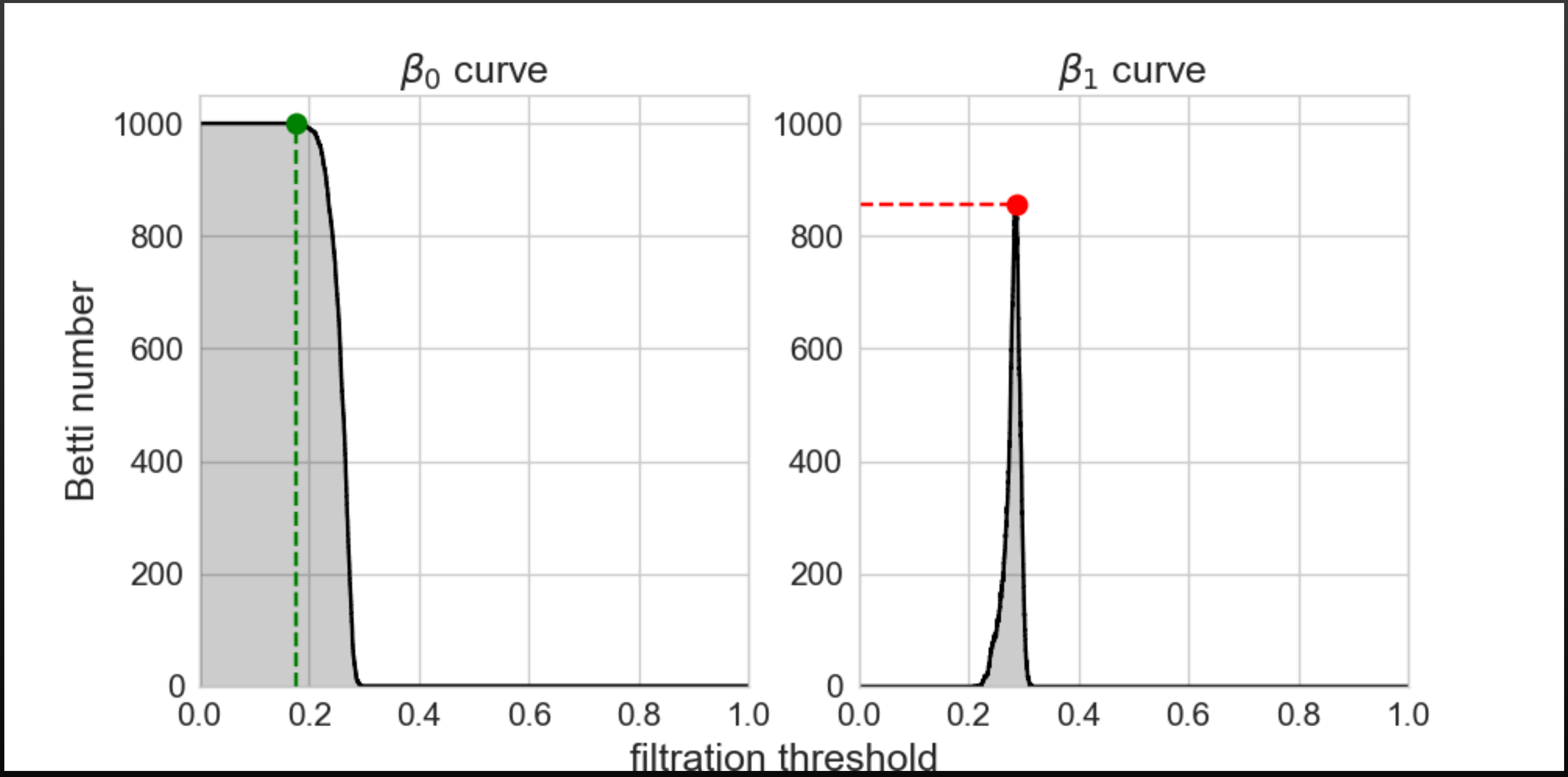
Betti_k curve

Nested complex to Betti curve



Bardin, et al., Network Neuroscience, 2019.

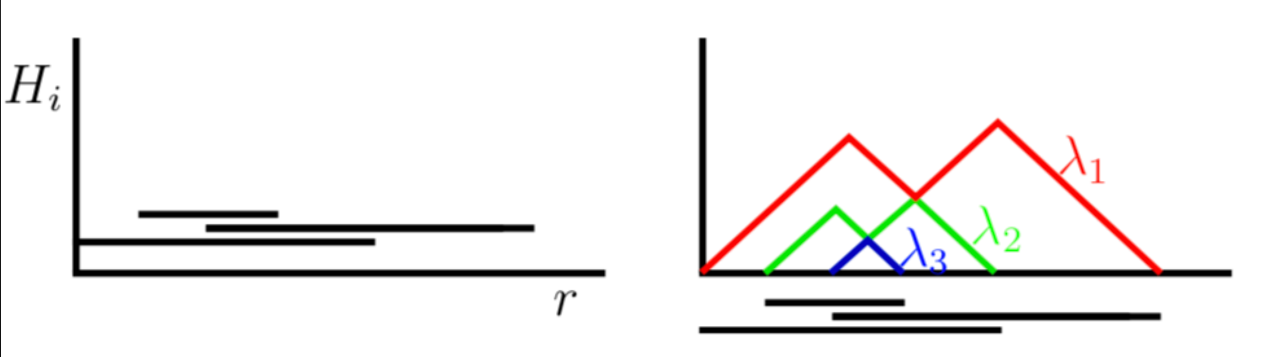
Extracting numerical features



Bardin, et al., Network Neuroscience, 2019.

Persistence landscapes

- Barcodes also give rise to *persistence landscapes*.



$$\lambda = \{\lambda_k : \mathbb{R} \to \mathbb{R} \cup \{\infty\} \mid k \in \mathbb{N}\}$$

- The *L2-landscape distance* between barcodes B and B' with associated landscapes **λ** and **λ**':

$$\Lambda(B, B') = \|\lambda - \lambda'\|_2 = \sum_{k=1}^{\infty} \left(\int |\lambda_k(t) - \lambda'_k(t)|^2 dt \right)^{\frac{1}{2}}$$

Bubenik, J Mach Learn Res (2015)
Dlotko & Bubenik, J Symbolic Comp (2017)

Persistence curves

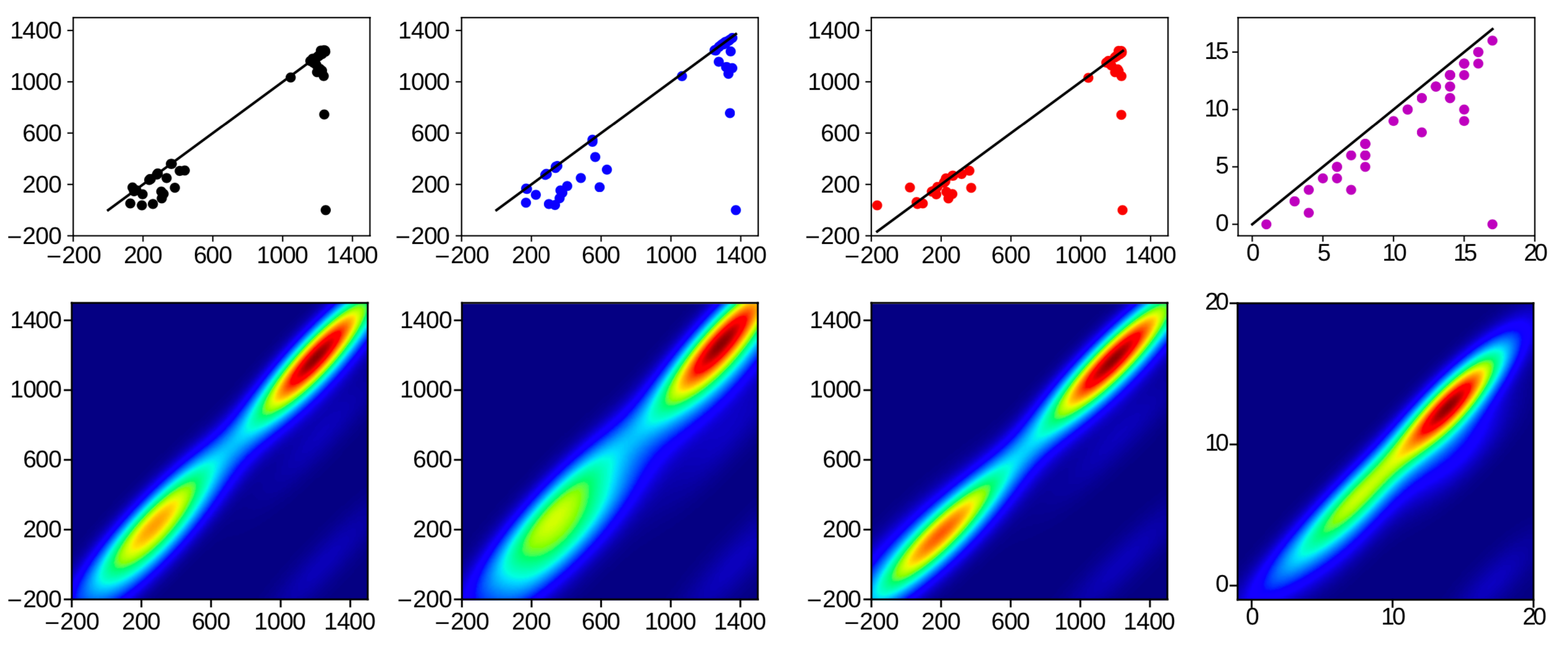
Name	Notation	$\psi(b,d,t)$	T
Betti	$\boldsymbol{\beta(D)}$	1	sum
Midlife	$\mathbf{ml}(D)$	$(b+d)/2$	sum
Life	$\boldsymbol{\ell}(D)$	$d-b$	sum
Multiplicative Life	$\mathbf{mul}(D)$	d/b	sum
Life Entropy [2]	$\mathbf{le}(D)$	$-\frac{d-b}{\sum(d-b)}\log\frac{d-b}{\sum(d-b)}$	sum
Midlife Entropy	$\mathbf{mle}(D)$	$-\frac{d+b}{\sum(d+b)}\log\frac{d+b}{\sum(d+b)}$	sum
Mult. Life Entropy	$\mathbf{mule}(D)$	$-\frac{d/b}{\sum(d/b)}\log\frac{d/b}{\sum(d/b)}$	sum
k-th Landscape [5]	$\boldsymbol{\lambda_k}(D)$	$\min\{t-b,d-t\}$	$\max_k$

Simultaneous generalization of Betti curves and persistence landscapes

Chung and Lawson, arXiv, 2019

Persistence images

- Smooth the PD: replace each point by a Gaussian kernel, then sum
- Discretize



Kanari, et al., Neuroinformatics, 2018.

ML methods applied to featurized TDA

- Decision tree
- Random forest
- Support Vector Machine
- CNN
- Graph CNN

Examples

- Topological characterization of neuron morphologies
- Automated classification of dynamic regimes in networks of neurons
- High-throughput screening of nanoporous materials

Thank you!