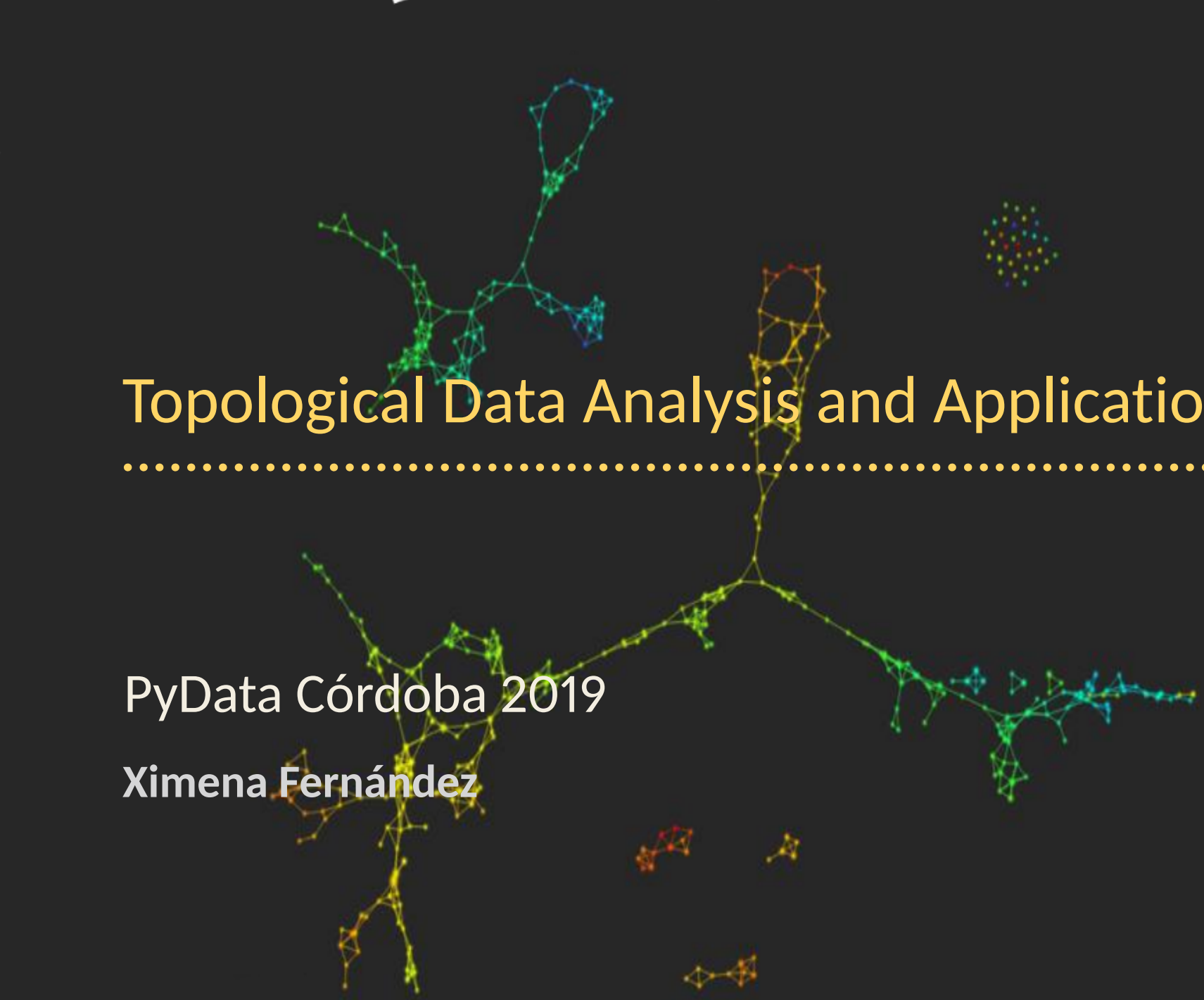
Topological Data Analysis and Applications



PyData Córdoba 2019

Ximena Fernández

Outline

Topological Data Analysis

Topo… what?

Topological Data Analysis

Topo... what?

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a point cloud, that is, a sample of a space X (where X is a metric space, a manifold, a topological space). Our goal is to recover *(the topology of)* X.

Topological Data Analysis

Topo... what?

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a point cloud, that is, a sample of a space X (where X is a metric space, a manifold, a topological space). Our goal is to recover *(the topology of)* X.

Topological Data Analysis

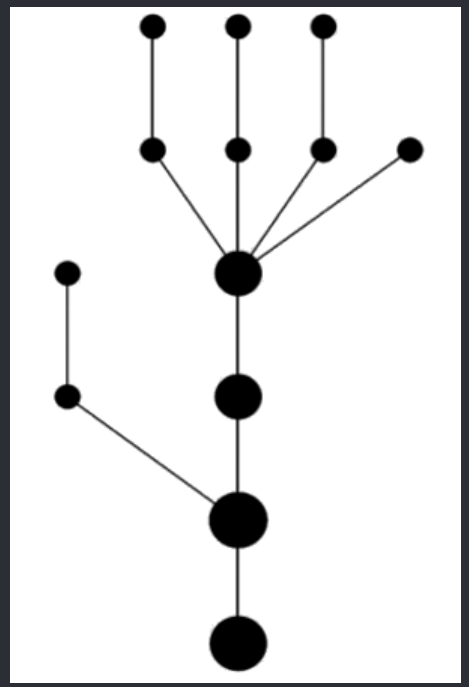
Topo... what?

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a point cloud, that is, a sample of a space X (where X is a metric space, a manifold, a topological space). Our goal is to recover *(the topology of)* X.

Topological Data Analysis

Topo... what?

Let $S = \{x_1, x_2, \cdots, x_n\}$ be a point cloud, that is, a sample of a space X (where X is a metric space, a manifold, a topological space). Our goal is to recover *(the topology of)* X.

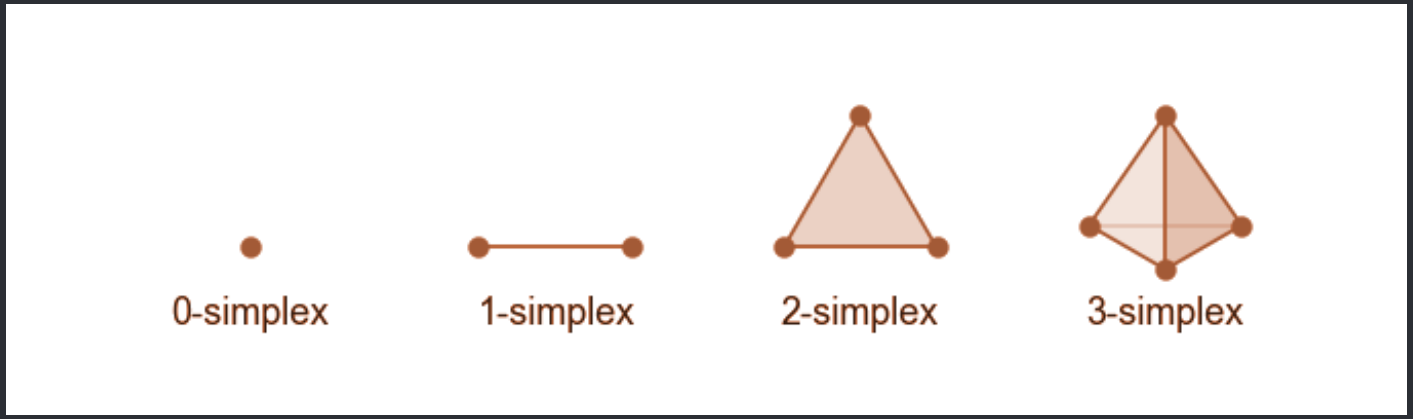


Reconstruction method

Simplicial complexes

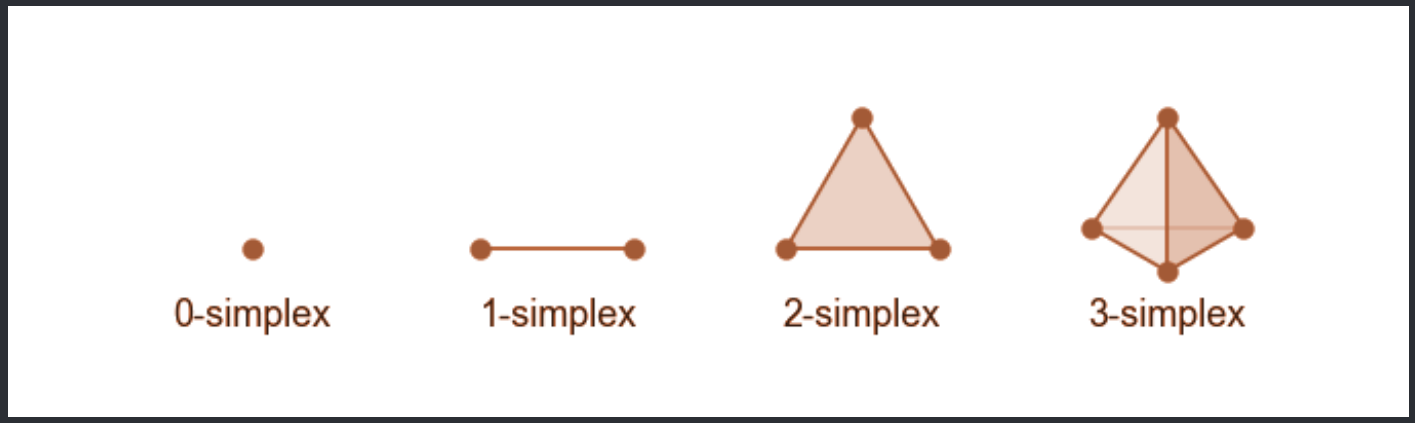
Reconstruction method

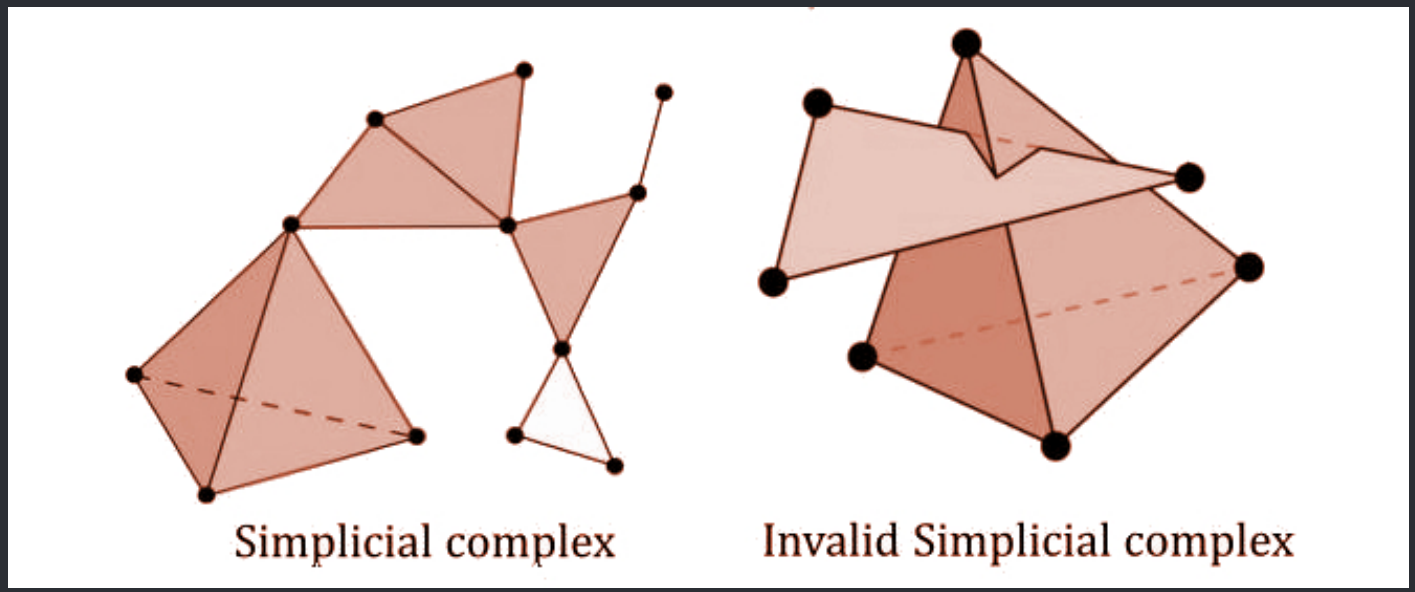
Simplicial complexes



Reconstruction method

Simplicial complexes





Reconstruction method

Reconstruction method

Definition

Let X be a topological space, and let $\mathcal{U} = \{U_1, U_2, \cdots, U_m\}$ be a *cover* of X. The nerve of $\mathcal{U}$ is the simplicial complex $N(\mathcal{U})$ with:

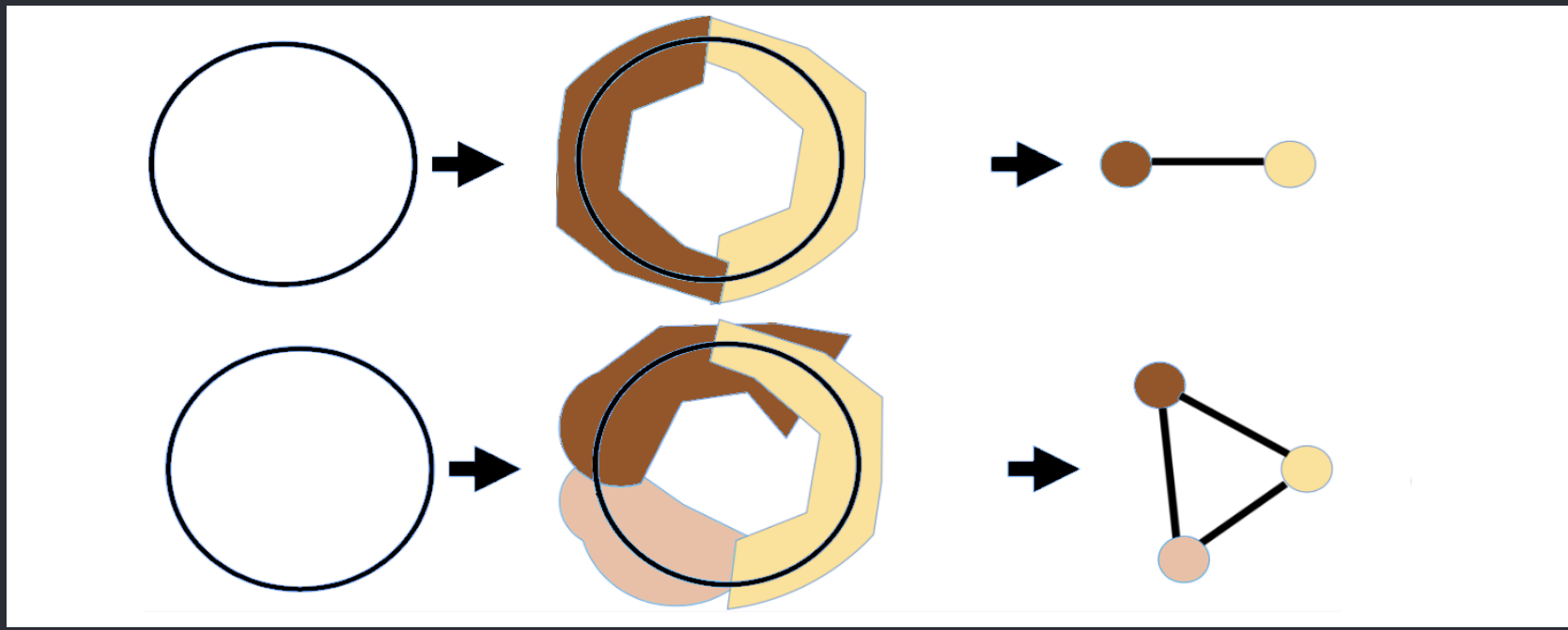
- one **vertex** for each element in the cover
- one **simplex** for each intersection of elements in the cover

Reconstruction method

Definition

Let X be a topological space, and let $\mathcal{U} = \{U_1, U_2, \cdots, U_m\}$ be a *cover* of X. The nerve of $\mathcal{U}$ is the simplicial complex $N(\mathcal{U})$ with:

- one **vertex** for each element in the cover
- one **simplex** for each intersection of elements in the cover



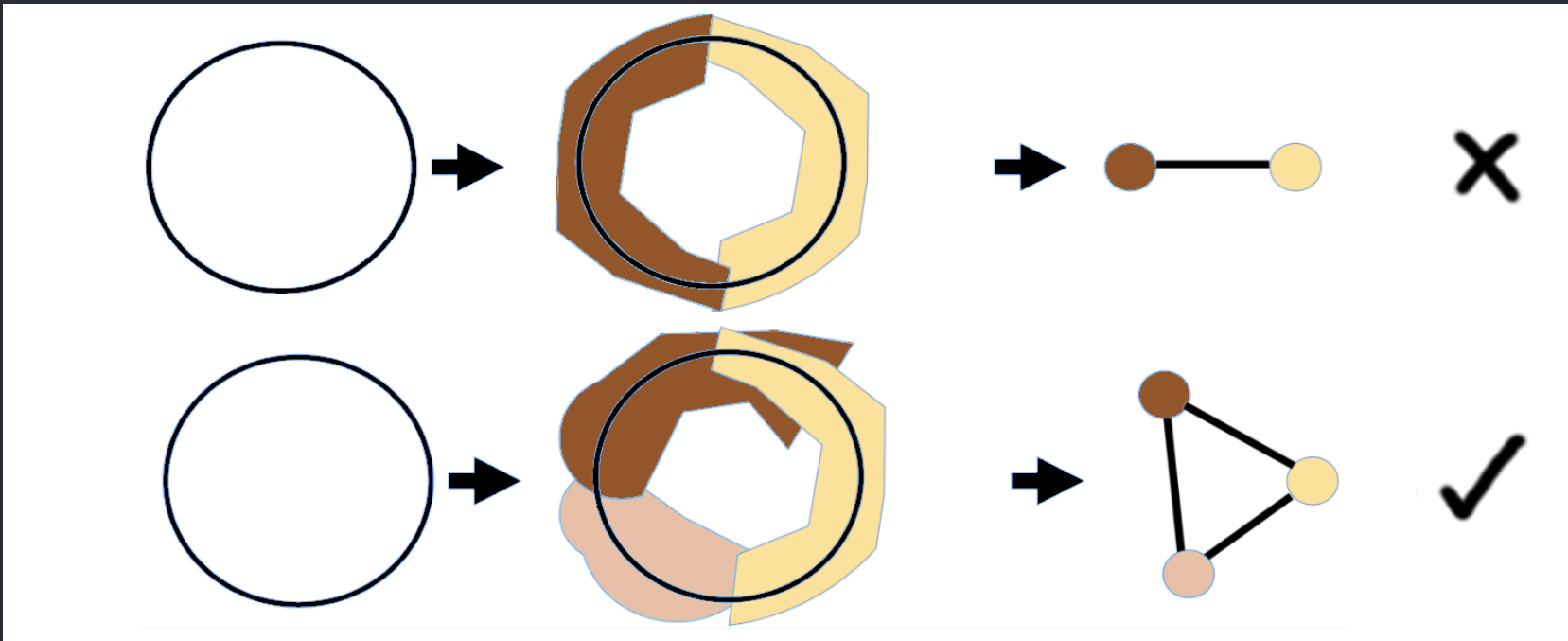
Reconstruction method

Nerve Theorem

Theorem

Let X be a topological space, and let $\mathcal{U} = \{U_1, U_2, \cdots, U_m\}$ *be a* ***good cover*** *of X by open sets.*

Then X is ***homotopically equivalent*** *to* $N(\mathcal{U})$.

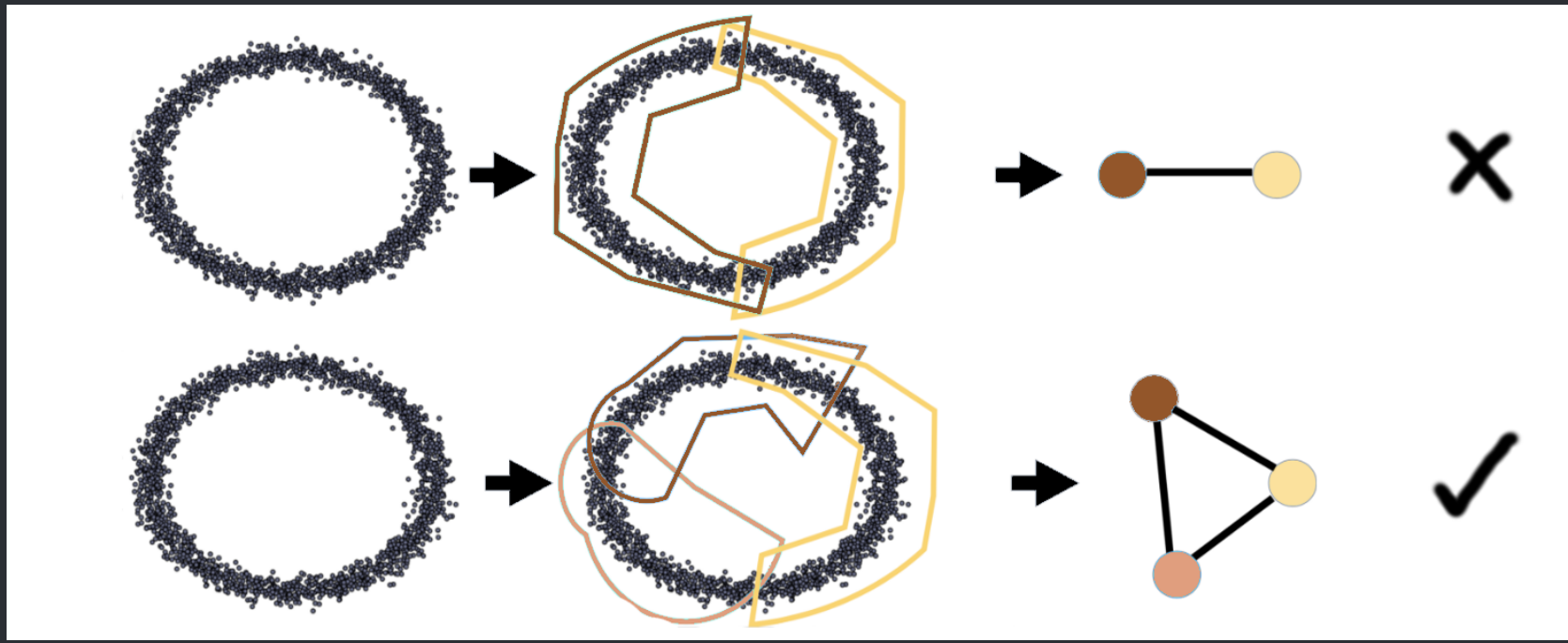


Reconstruction method

Nerve Theorem

Theorem

Let X be a topological space, and let $\mathcal{U} = \{U_1, U_2, \cdots, U_m\}$ be a ***good cover*** *of X by open sets.*
Then X is ***homotopically equivalent*** *to $N(\mathcal{U})$.*



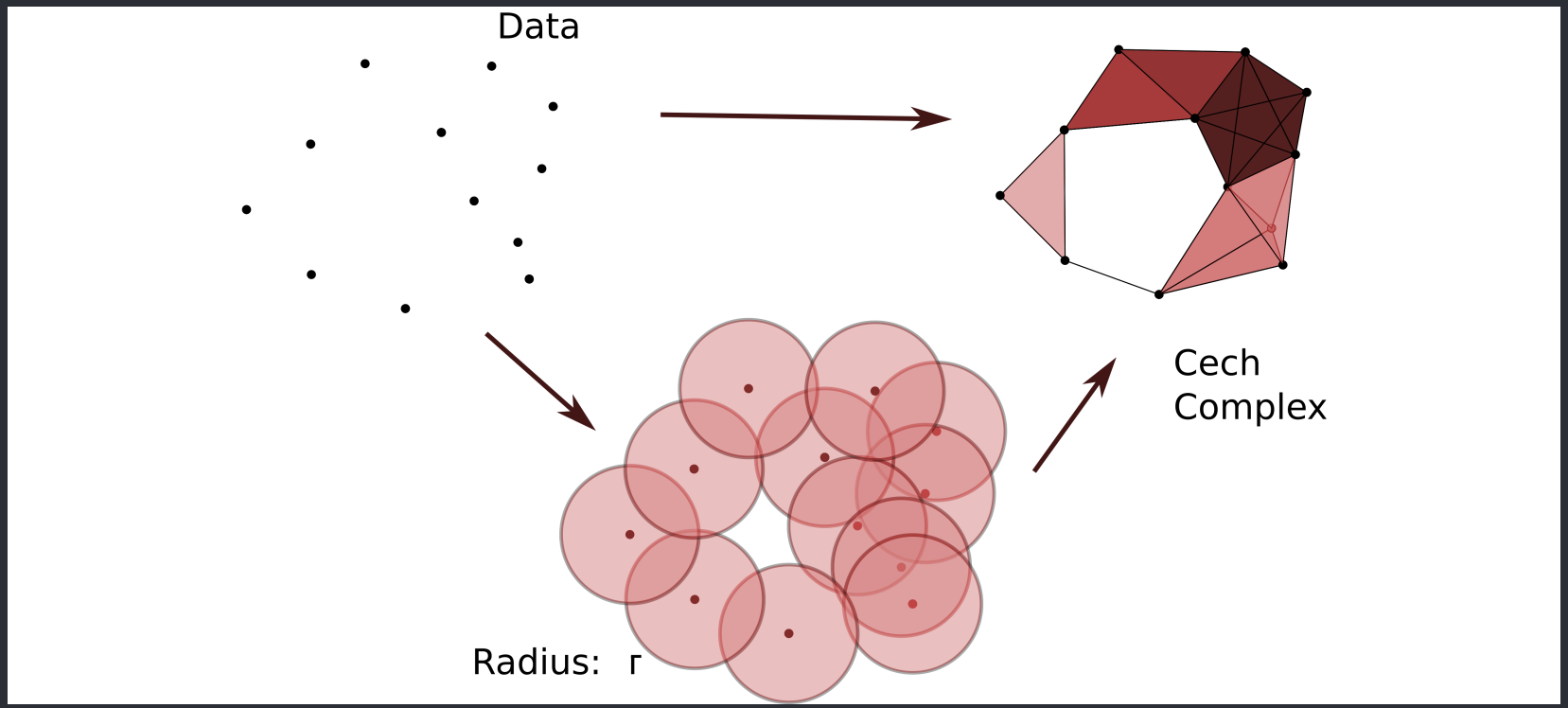
Reconstruction method

How can I select a (good) cover in a point cloud?

Reconstruction method

How can I select a (good) cover in a point cloud?

Let $S = \{x_1, x_2, \cdots, x_n\}$ a point cloud. Suppose you have a metric on S. Define the cover $\mathscr{U}$ as the collection of balls $B(x_1, r), B(x_2, r), \cdots, B(x_n, r)$ with center in the points $x_1, x_2, \cdots x_n$ of S and radius $r > 0$.



Reconstruction method

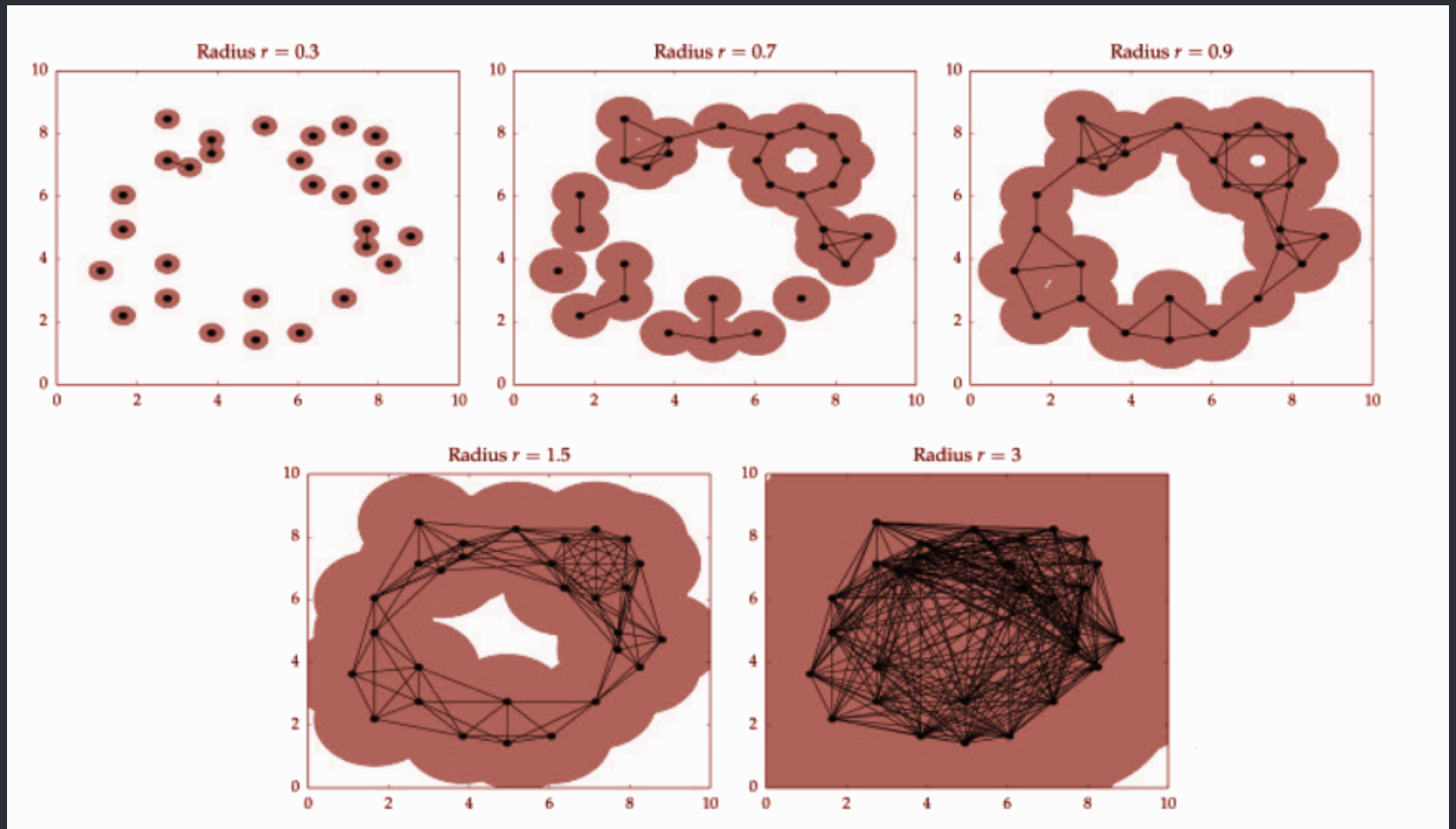
Which radius should I choose?

Reconstruction method

Which radius should I choose?

Problem

Different choices of r give rise to different topologies.



Persistent Homology

All radii

Persistent Homology

All radii

Solution

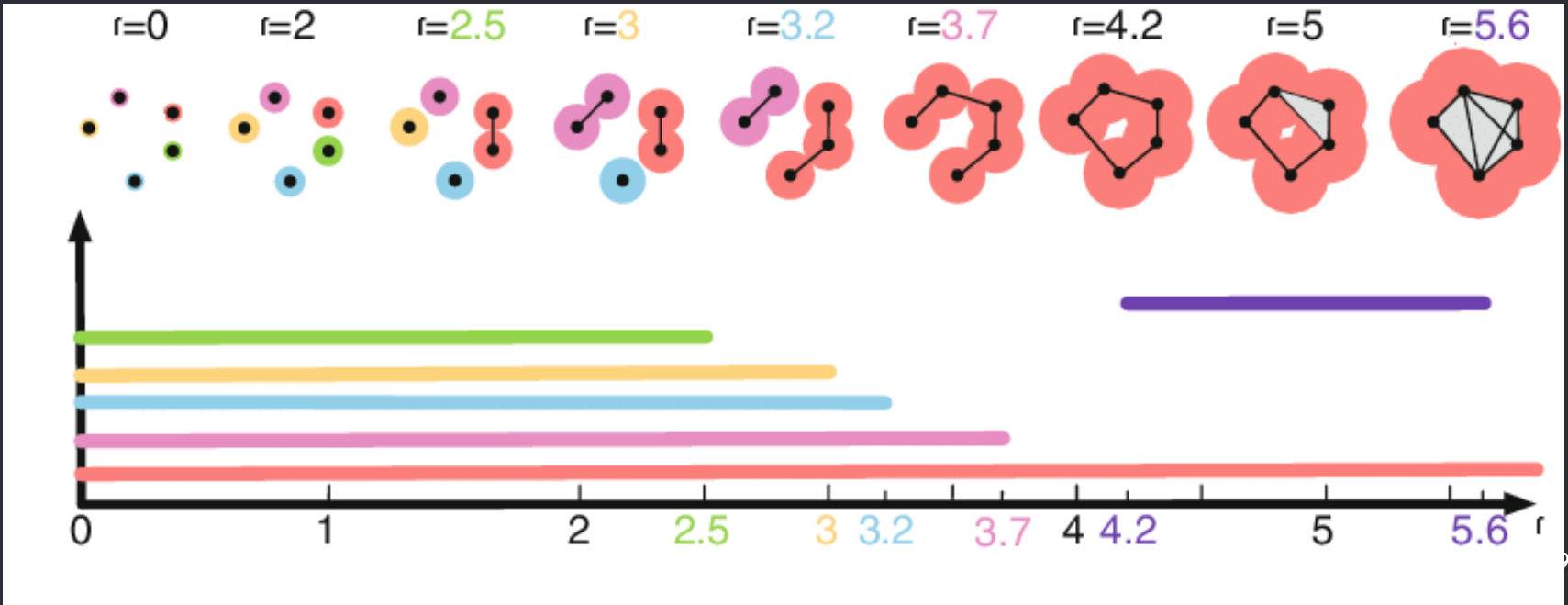
Study the *geometry* of the data with a *multiscale resolution*.

Persistent Homology

All radii

Solution

Study the *geometry* of the data with a *multiscale resolution*. Specifically, compute the *evolution* of the homology of the nerve of a cover *with the radius* of the balls.



Dynamical systems

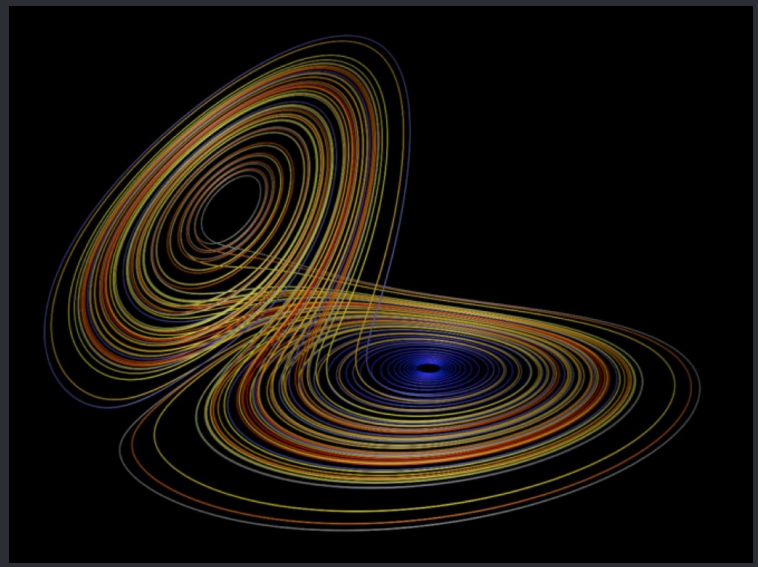
ft. Persistent Homology

- Maletic, S., Zhao, Y., Rajkovic, M. (2016). *Persistent topological features of dynamical systems*. Chaos, 26 5, 053105.

Dynamical systems

ft. Persistent Homology

- Maletic, S., Zhao, Y., Rajkovic, M. (2016). *Persistent topological features of dynamical systems*. Chaos, 26 5, 053105.



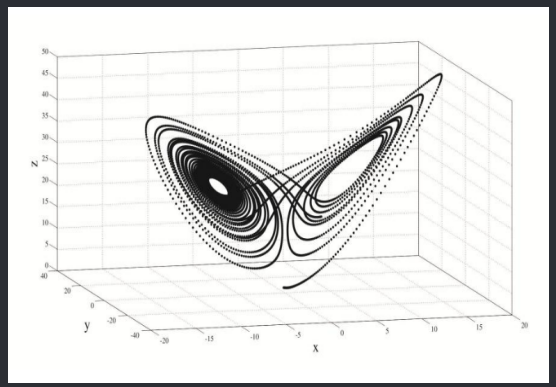
Classify the dynamical systems by identifying geometrical invariants of the *state space*. Concretely, perform a discrete simulation of the dynamical system and compute the persistent homology of this point cloud.

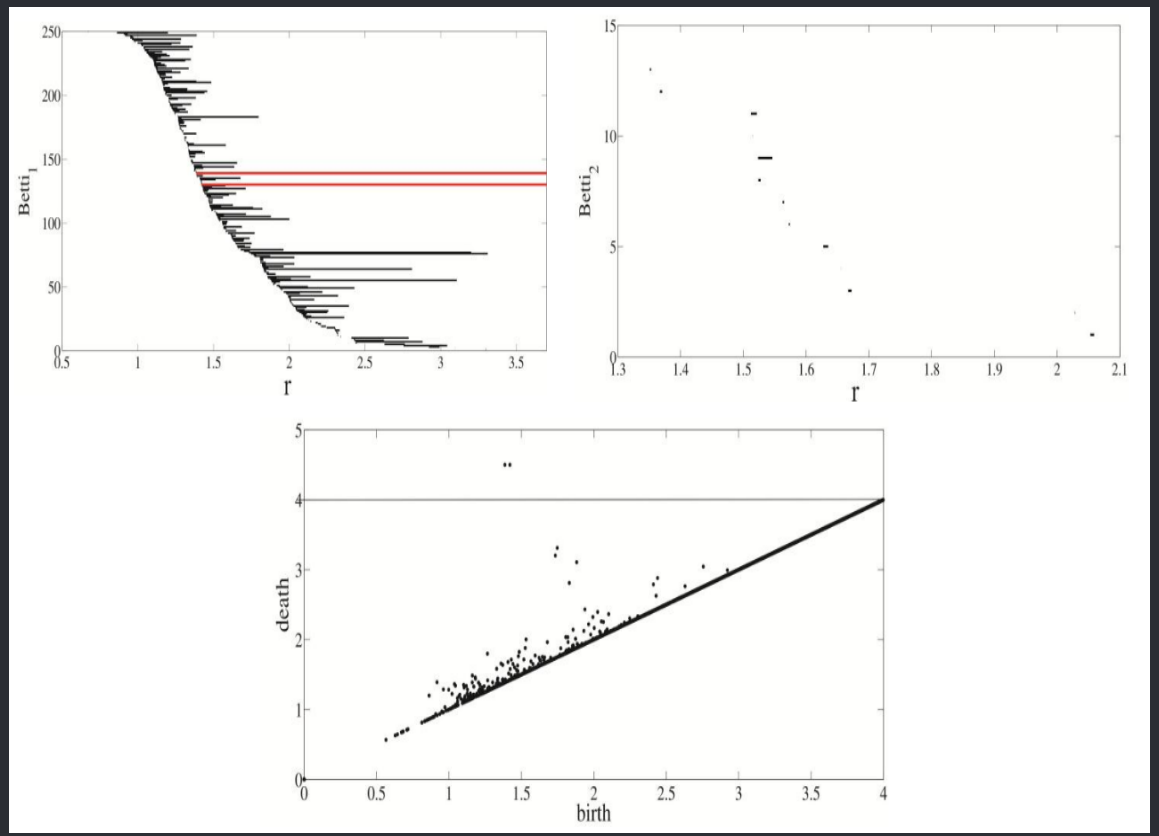
Dynamical systems

ft. Persistent Homology

Lorenz

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = rx - y - xz$$
$$\dot{z} = xy - bz$$



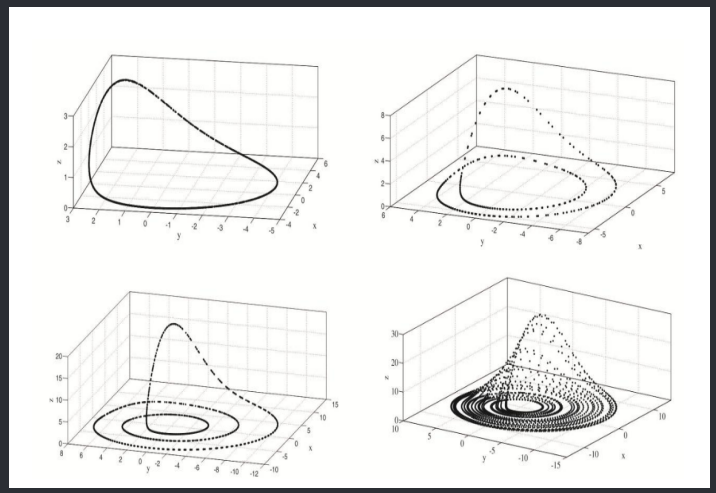


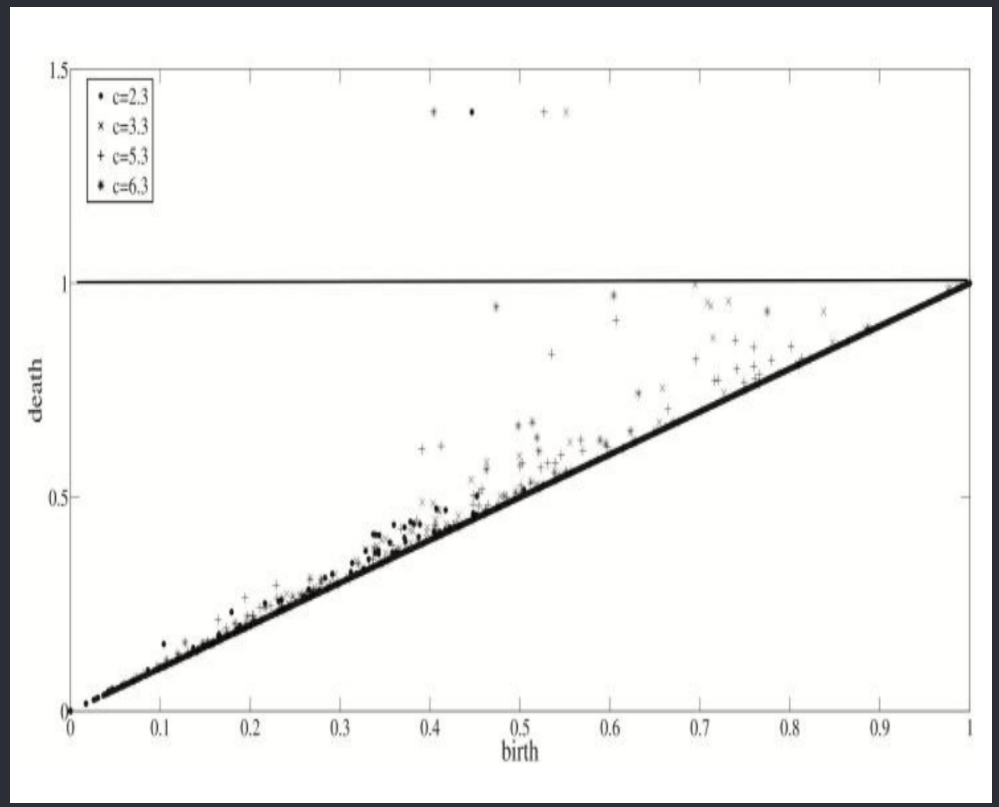
Dynamical systems

ft. Persistent Homology

Rösler

$$\dot{x} = -\sigma(y + z)$$
$$\dot{y} = x + ay$$
$$\dot{z} = b + xz + cz$$





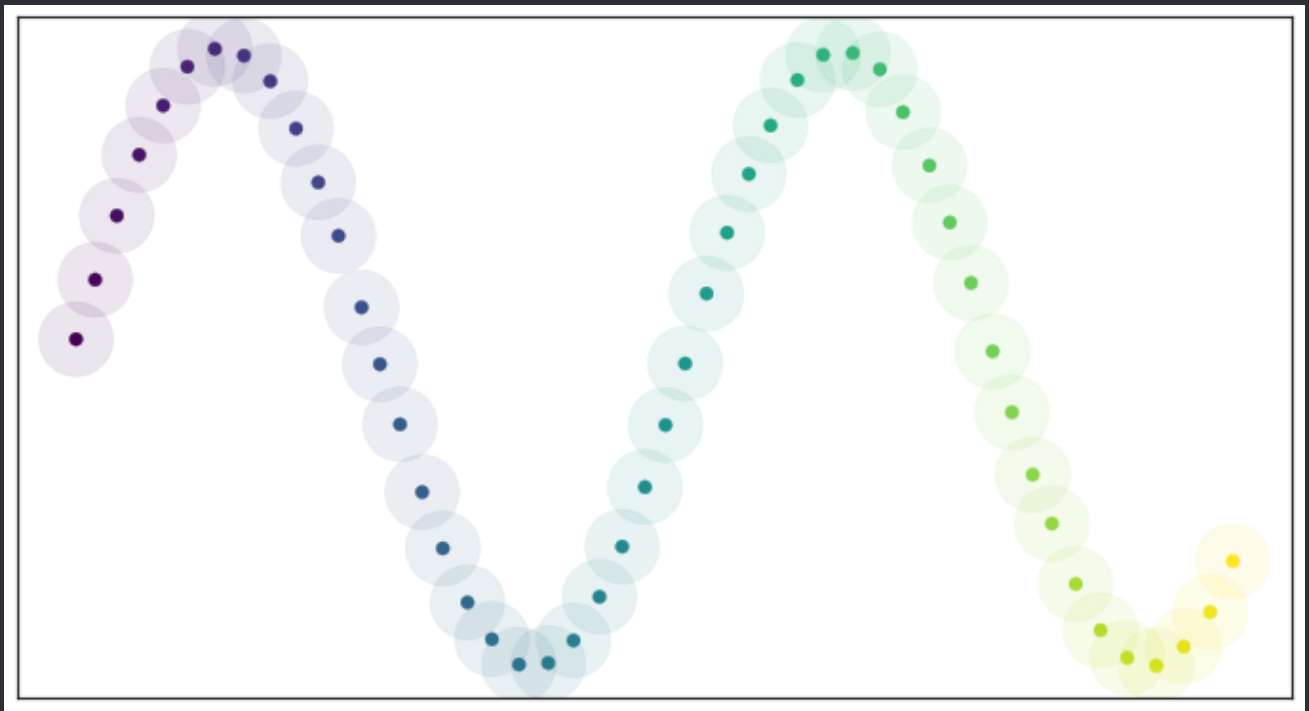
Uniform Manifold Approximation and Projection (UMAP)

Smart metric

Uniform Manifold Approximation and Projection (UMAP)

Smart metric

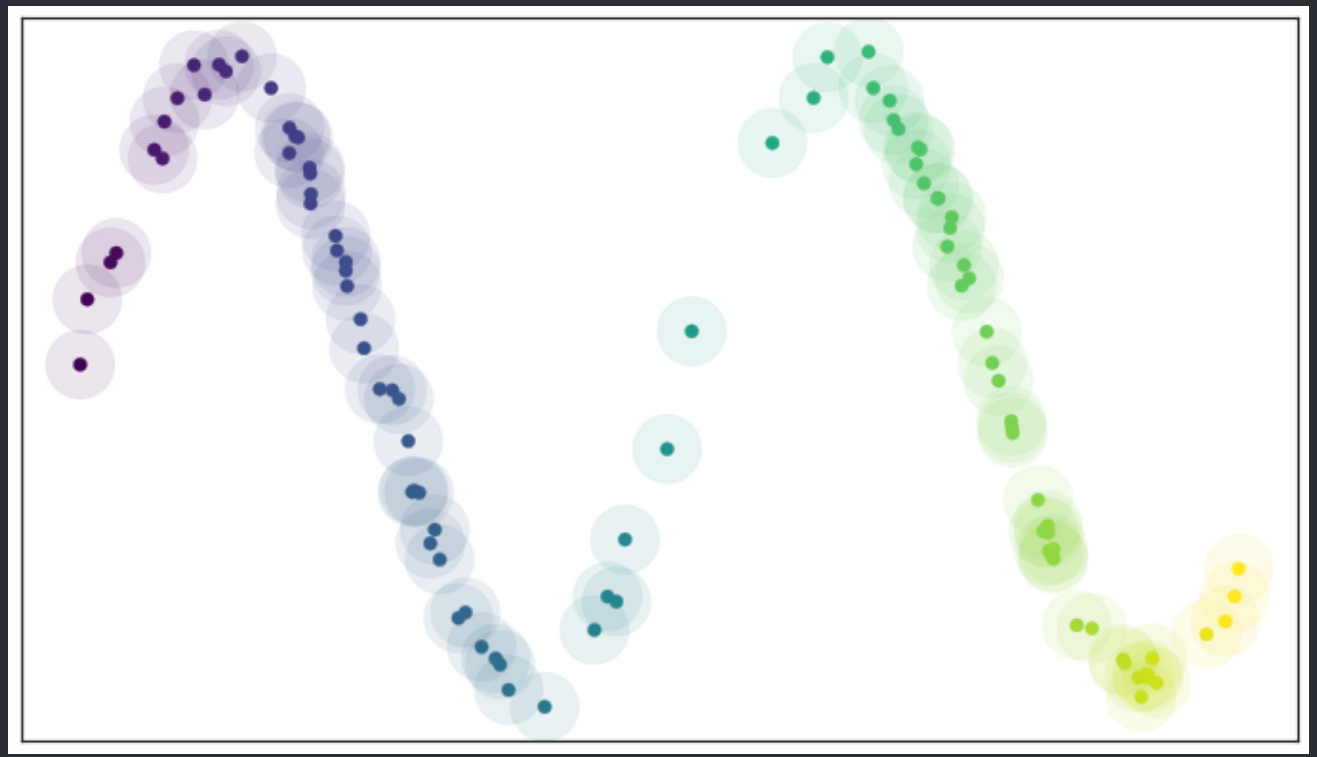
Expectation



Uniform Manifold Approximation and Projection (UMAP)

Smart metric

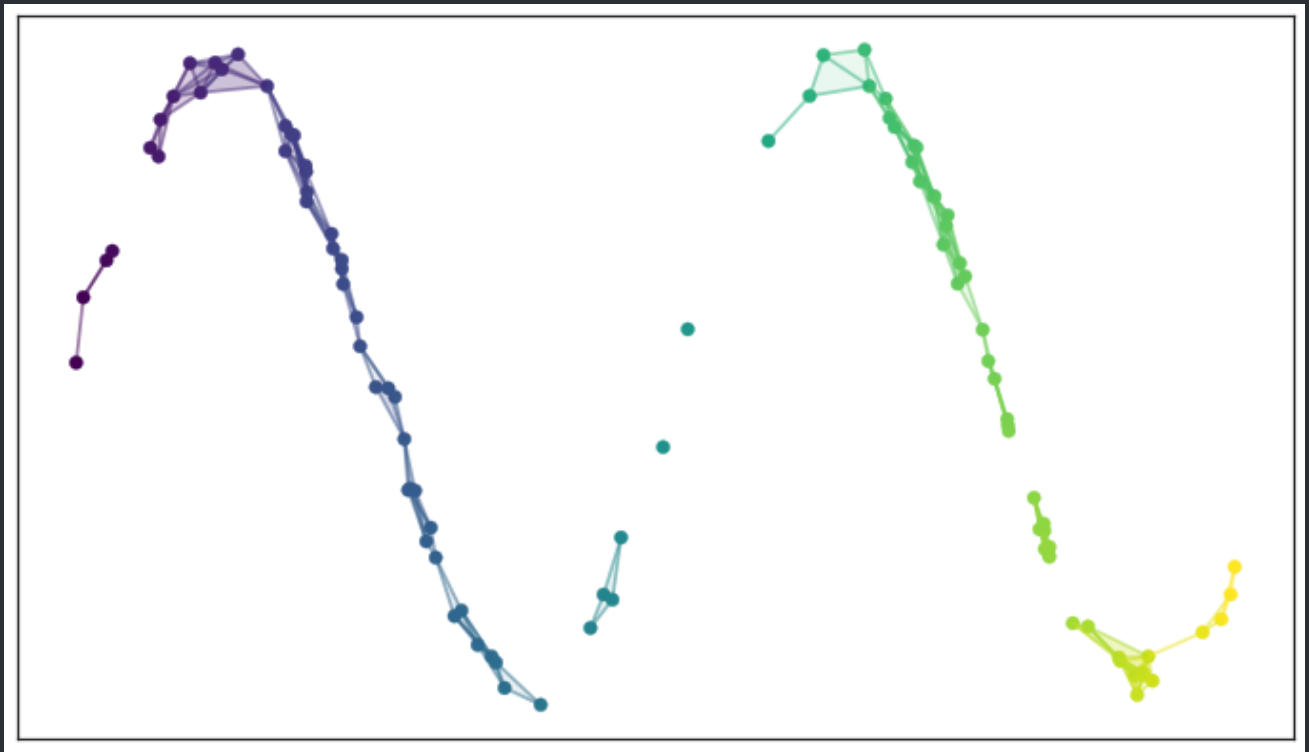
Reality



Uniform Manifold Approximation and Projection (UMAP)

Smart metric

Reality



UMAP

Smart metric

Problem

- The data is not uniformly distributed :(

UMAP

Smart metric

Problem

- The data is not uniformly distributed :(

Solution

UMAP

Smart metric

Problem

- The data is not uniformly distributed :(

Solution

We will **assume** that:

- The data is uniformly distributed on Riemannian manifold :)

UMAP

Smart metric

Problem

- The data is not uniformly distributed :(

Solution

We will **assume** that:

- The data is uniformly distributed on Riemannian manifold :)
- The Riemannian metric is locally constant (or can be approximated as such)
- The manifold is locally connected

UMAP

Smart metric

Let S be a point cloud in $\mathbb{R}^m$.

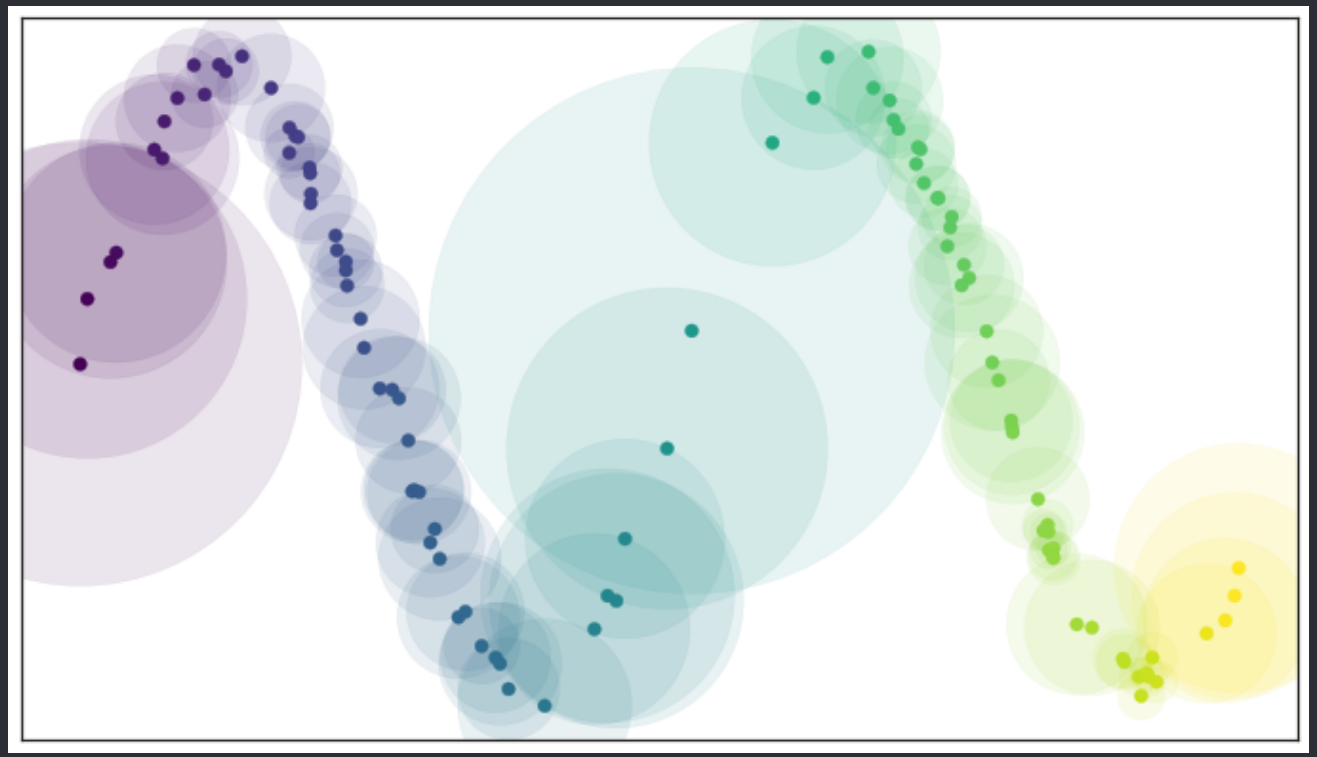
Define (an approximation of) a local notion of distance such that: a unit ball about each point in S stretches to the k-th nearest neighbor of the point (with k a parameter).

UMAP

Smart metric

Let S be a point cloud in $\mathbb{R}^m$.

Define (an approximation of) a local notion of distance such that: a unit ball about each point in S stretches to the k-th nearest neighbor of the point (with k a parameter).

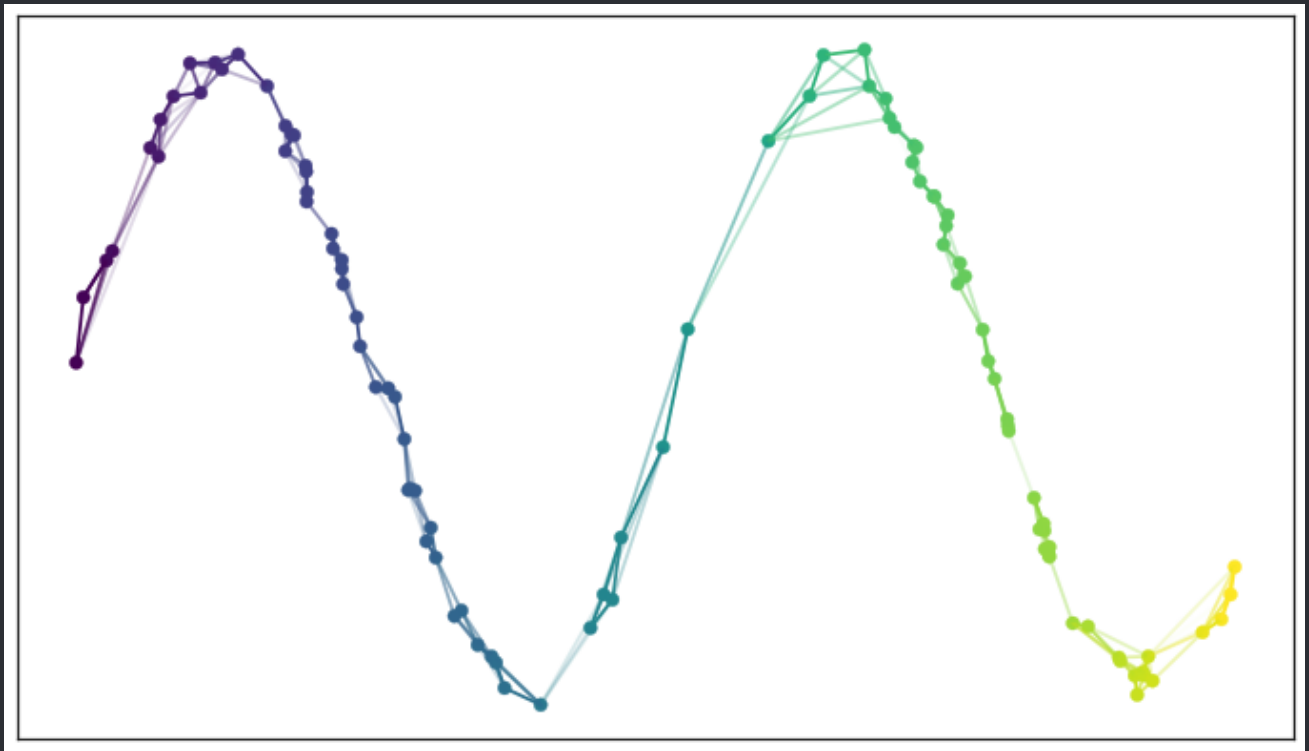


UMAP

Smart metric

Make a compatibilization of this locally sense of distance.

Compute the nerve of that covering by unit balls.

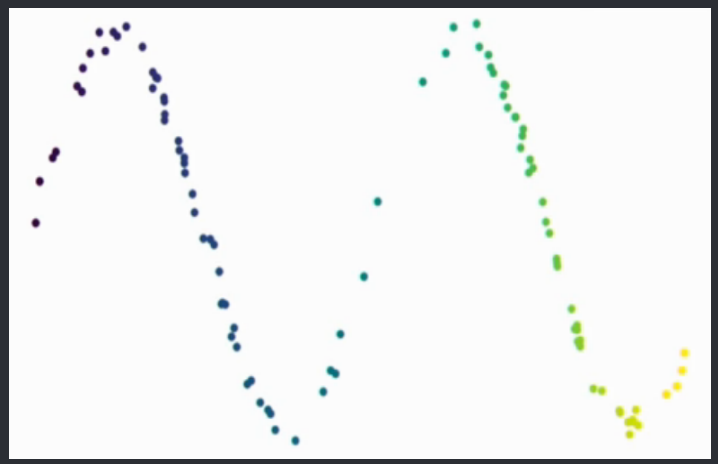


UMAP

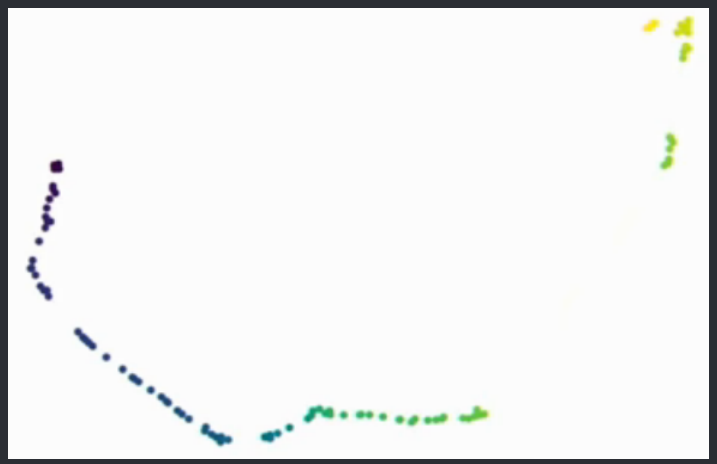
Projection

High dimensional representation → **Low dimensional representation**

Riemmanian manifold with varying metric



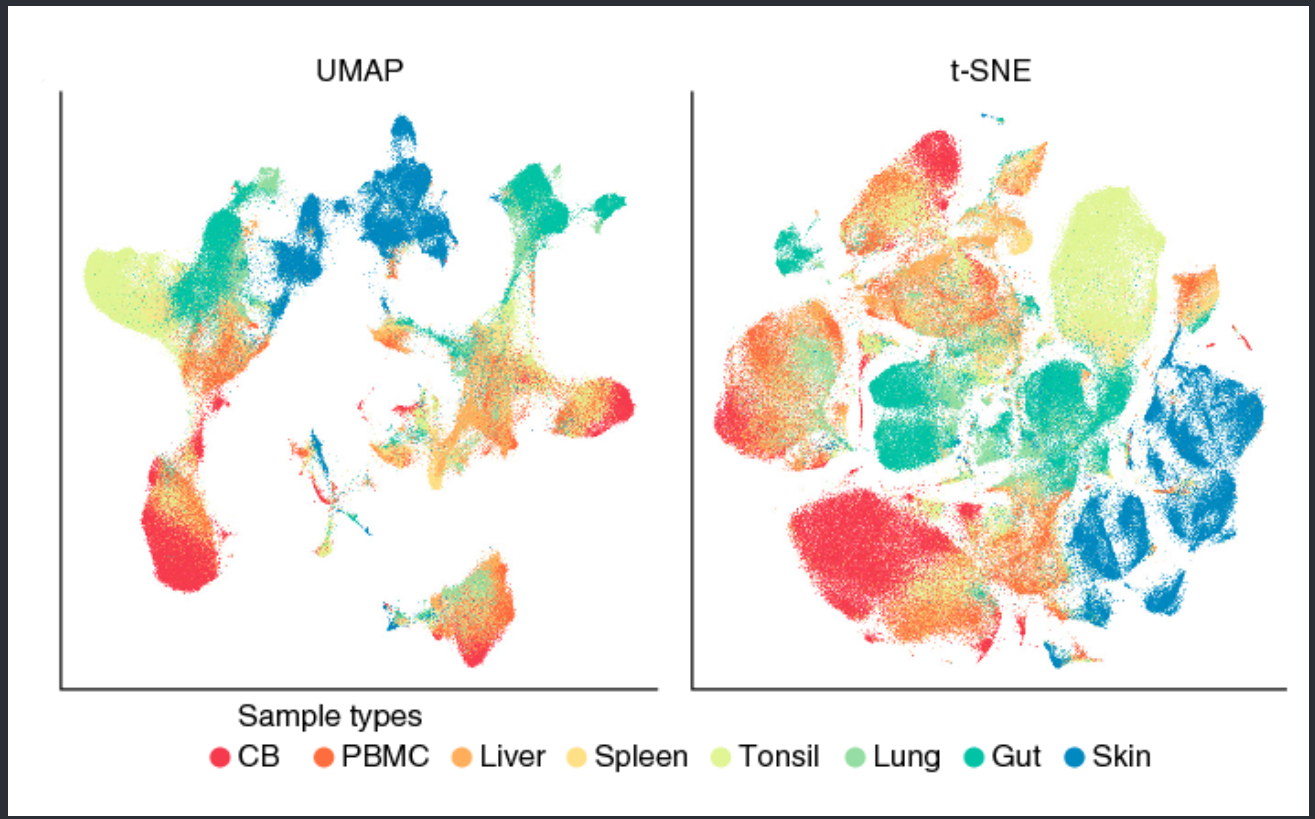
Low dimensional euclidean space with the euclidean metric



Single-cell RNA-sequences

ft. UMAP

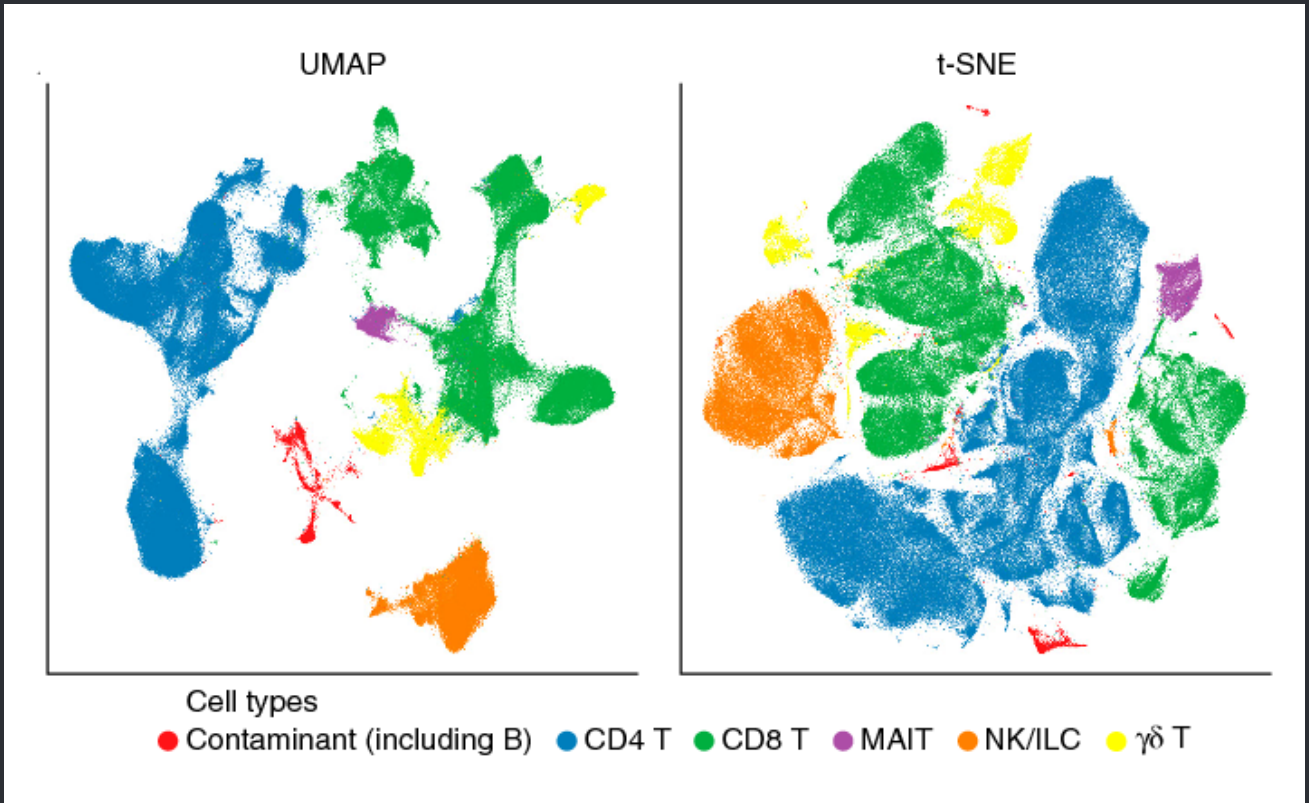
- Becht, E., McInnes, L., Healy, J., Dutertre, C., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W. (2018). *Dimensionality reduction for visualizing single-cell data using UMAP*. Nature Biotechnology, 37, 38-44.



Single-cell RNA-sequences

ft. UMAP

- Becht, E., McInnes, L., Healy, J., Dutertre, C., Kwok, I.W., Ng, L.G., Ginhoux, F., Newell, E.W. (2018). *Dimensionality reduction for visualizing single-cell data using UMAP*. Nature Biotechnology, 37, 38-44.



Mapper

Smart cover

Mapper

Smart cover

- Let X be a topological space.

Mapper

Smart cover

- Let X be a topological space.
- Let $f : X \to [a, b] \subset \mathbb{R}$ be a continuous function, also called filter.

Mapper

Smart cover

- Let X be a topological space.
- Let $f : X \to [a, b] \subset \mathbb{R}$ be a continuous function, also called filter.
- Let $\mathscr{I} = \{I_1, I_2, \cdots, I_k\}$ be a cover of $[a, b]$ by overlapping open intervals.

Mapper

Smart cover

- Let X be a topological space.
- Let $f : X \to [a, b] \subset \mathbb{R}$ be a continuous function, also called filter.
- Let $\mathscr{I} = \{I_1, I_2, \cdots, I_k\}$ be a cover of $[a, b]$ by overlapping open intervals.
- Define $\mathscr{U}^{\pi_0}$ as the connected components of the preimage $f^{-1}(I_j)$ of each interval I_j of $\mathscr{I}$.

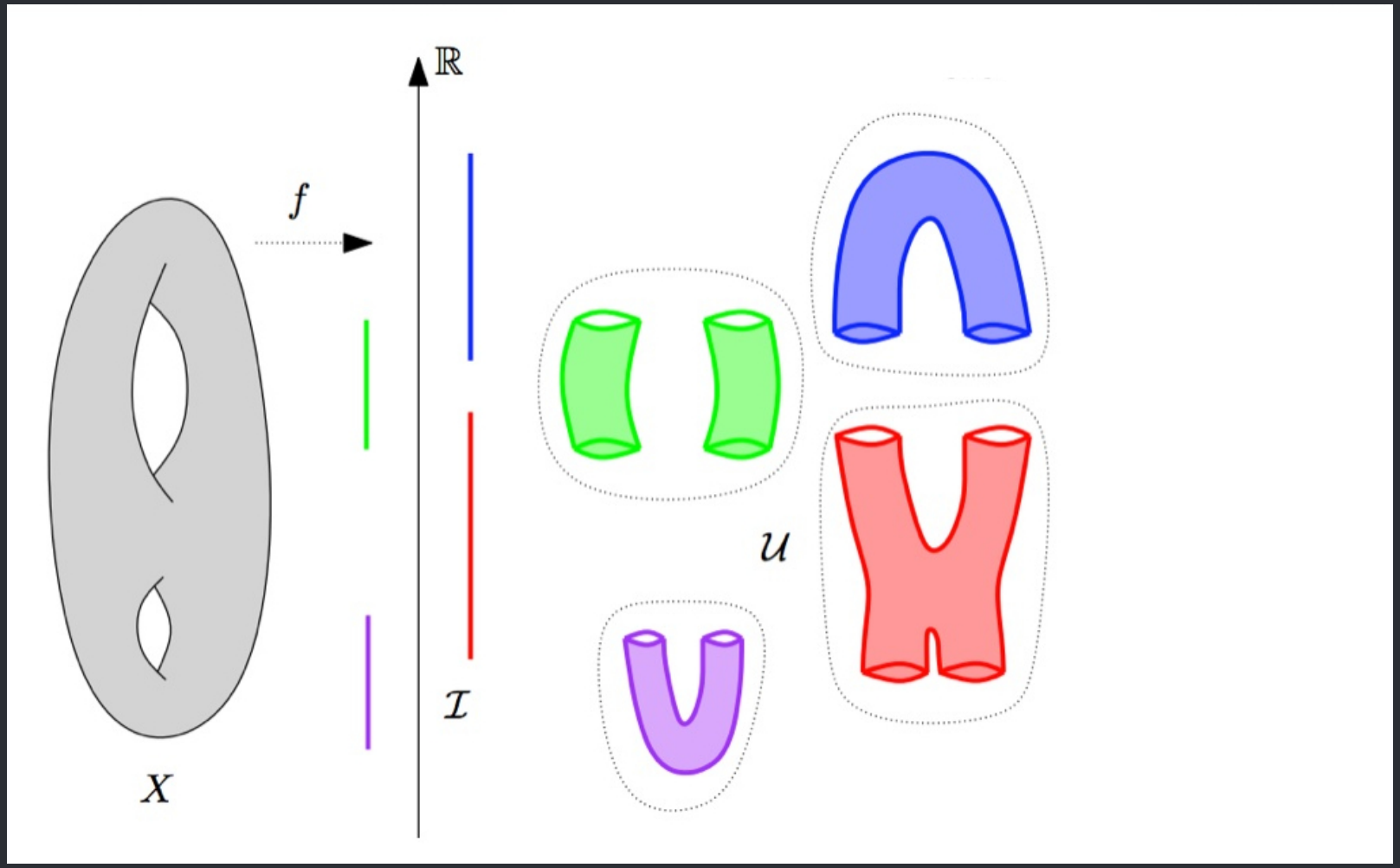
Mapper

Smart cover

- Let X be a topological space.
- Let $f : X \to [a, b] \subset \mathbb{R}$ be a continuous function, also called filter.
- Let $\mathscr{I} = \{I_1, I_2, \cdots, I_k\}$ be a cover of $[a, b]$ by overlapping open intervals.
- Define $\mathscr{U}^{\pi_0}$ as the connected components of the preimage $f^{-1}(I_j)$ of each interval I_j of $\mathscr{I}$.
- $N(\mathscr{U}^{\pi_0})$ *recovers* the topology of X.

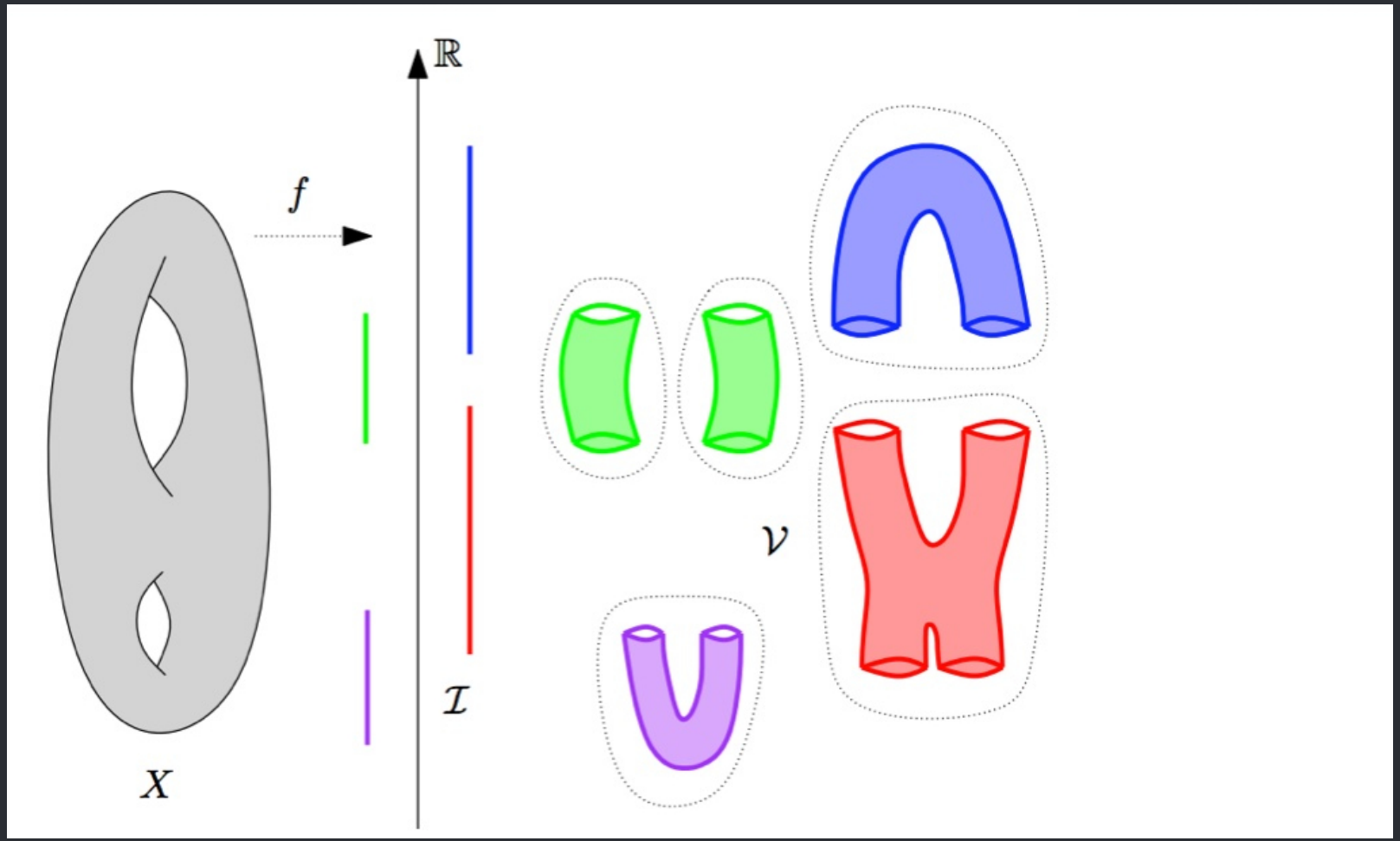
Mapper

Smart cover



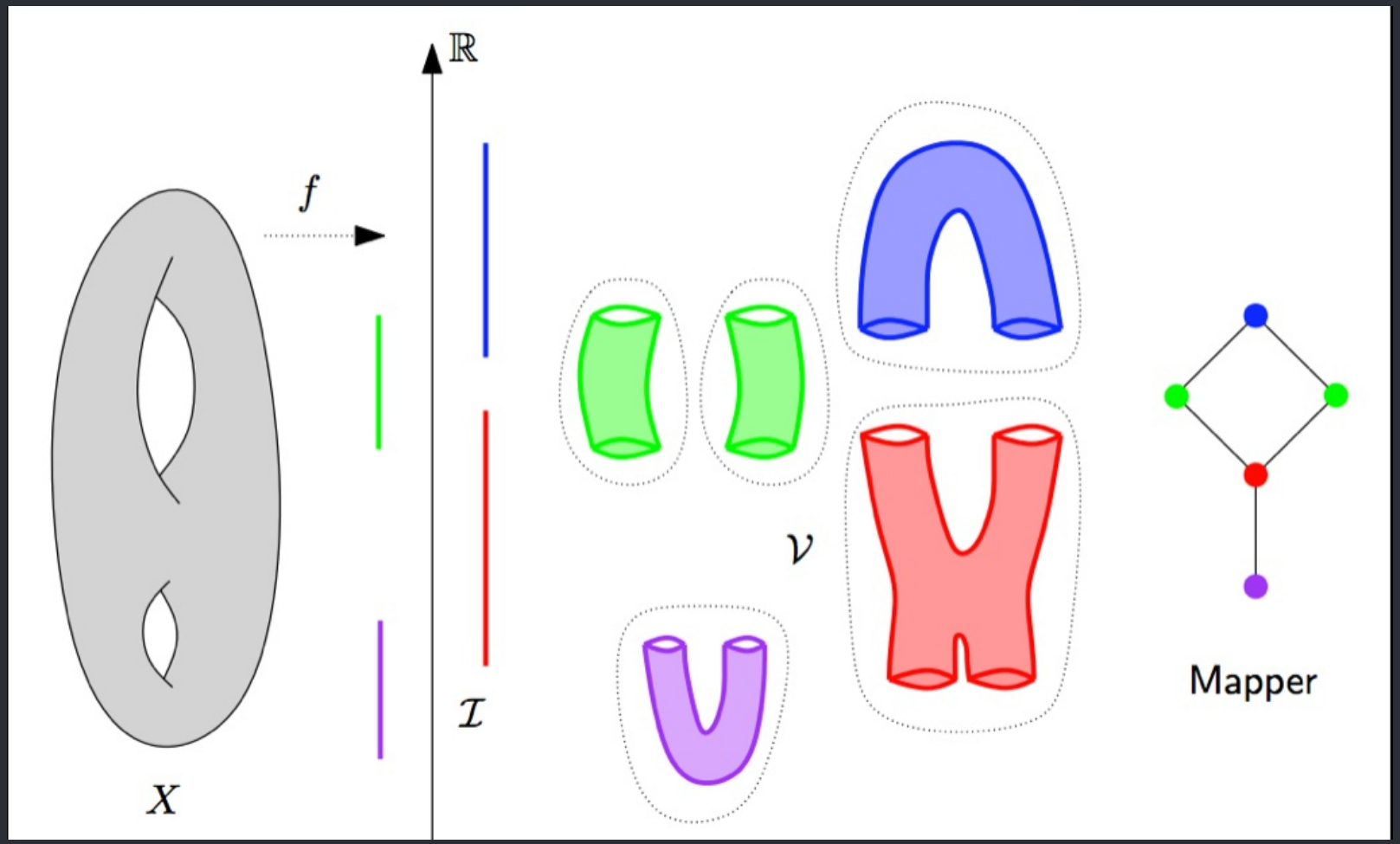
Mapper

Smart cover



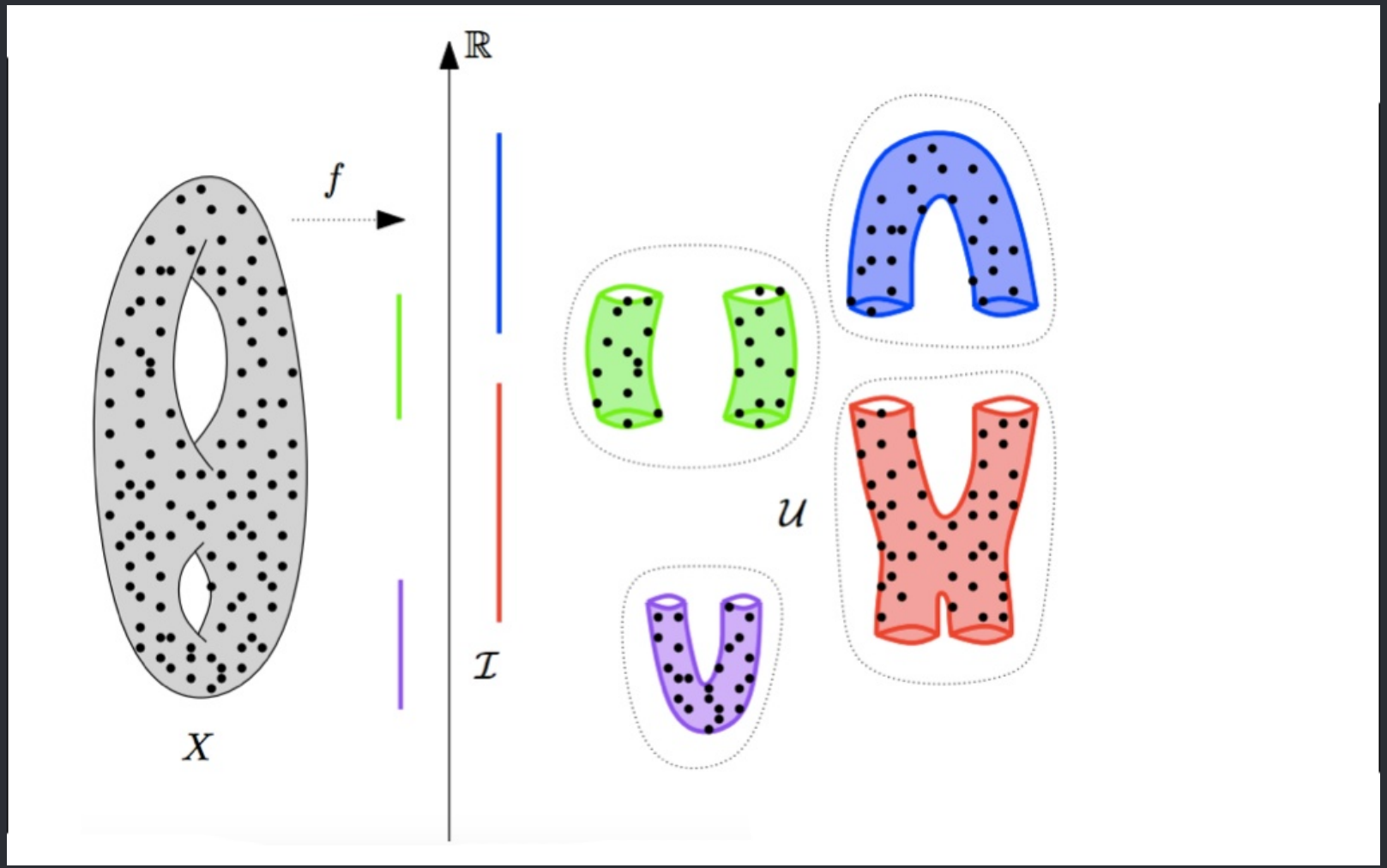
Mapper

Smart cover



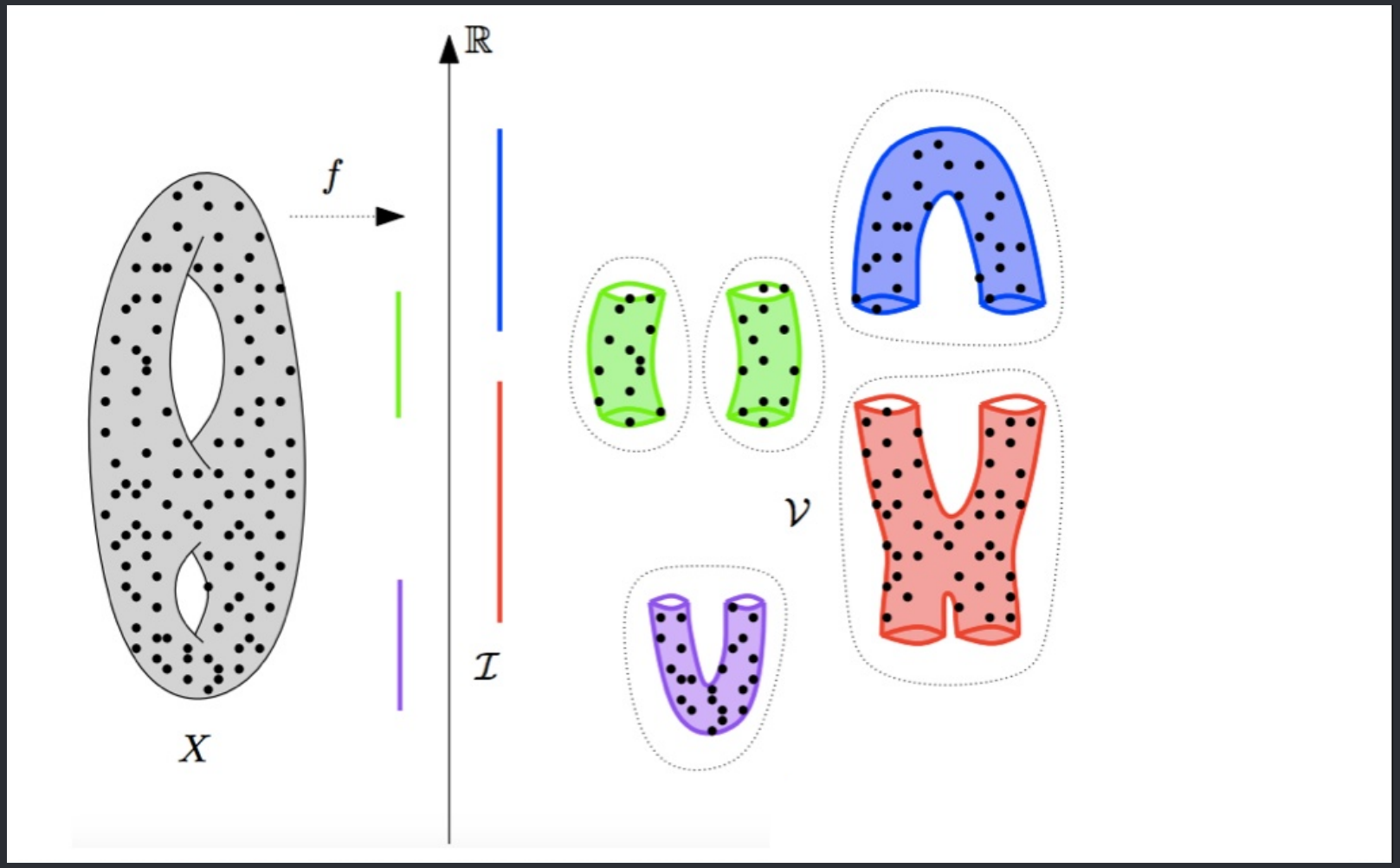
Mapper

Smart cover



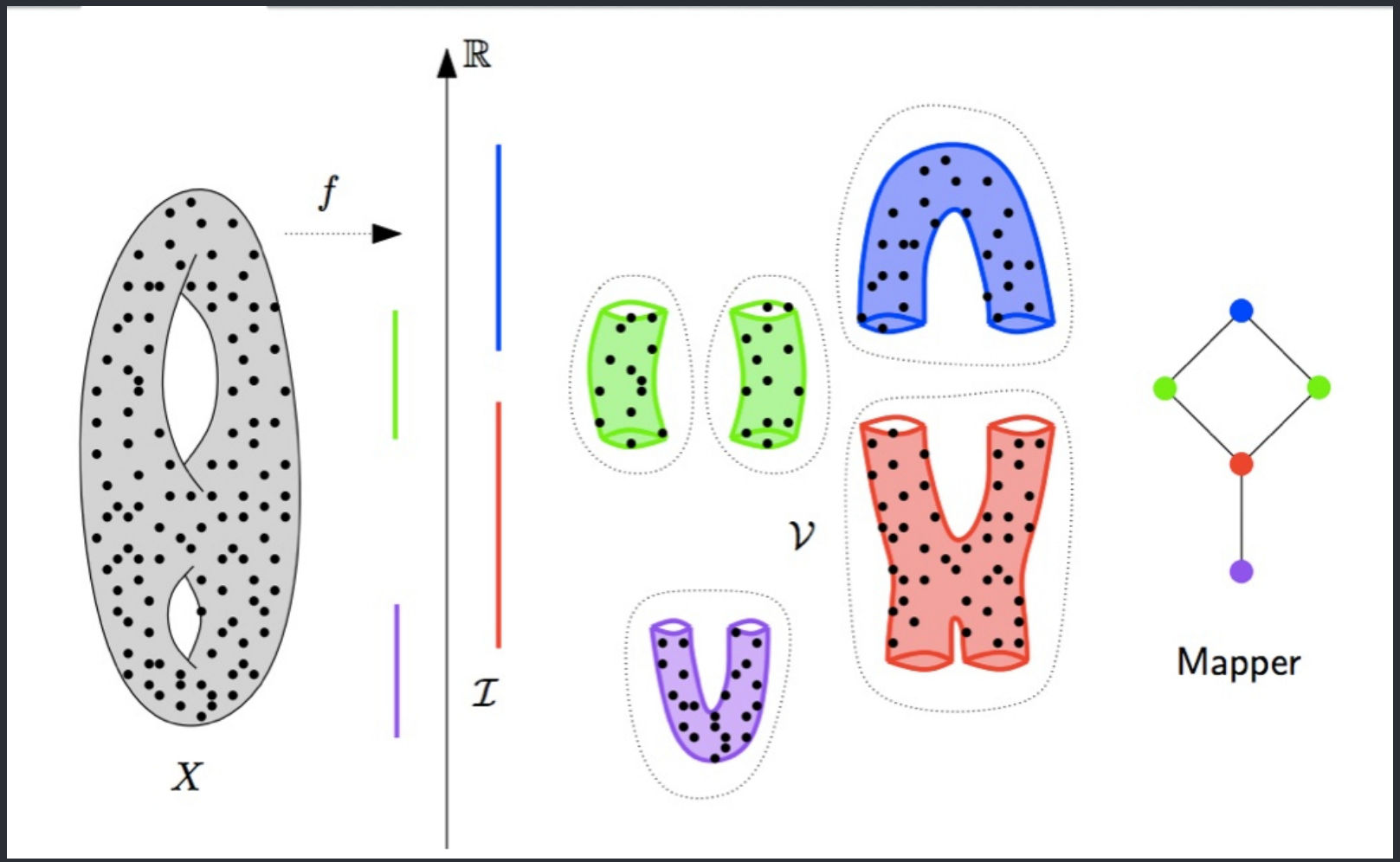
Mapper

Smart cover



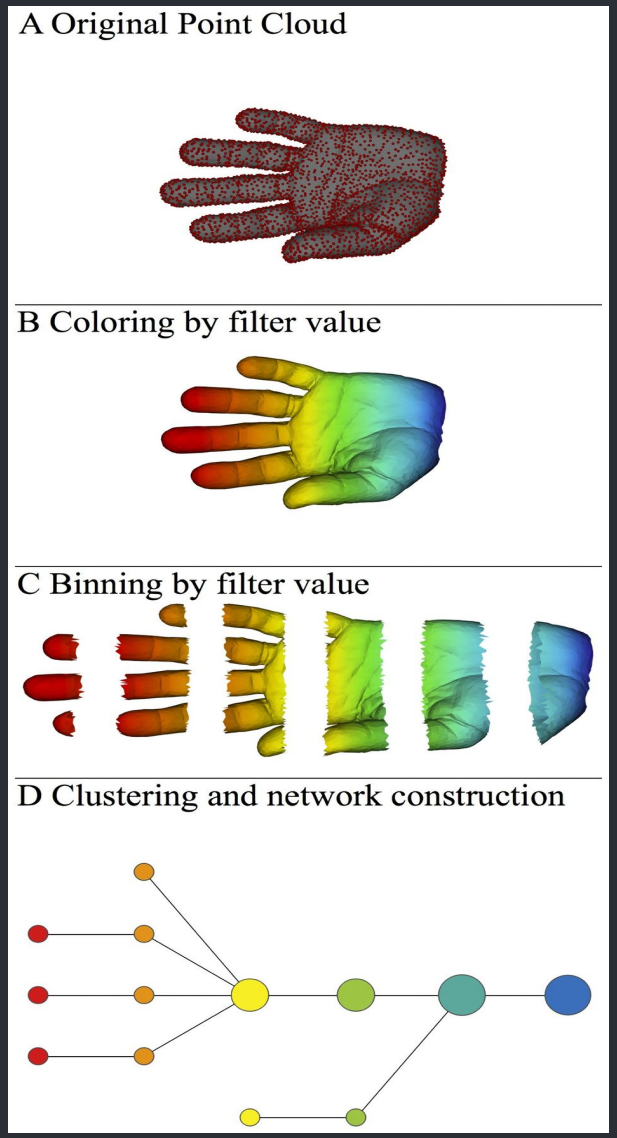
Mapper

Smart cover



Mapper

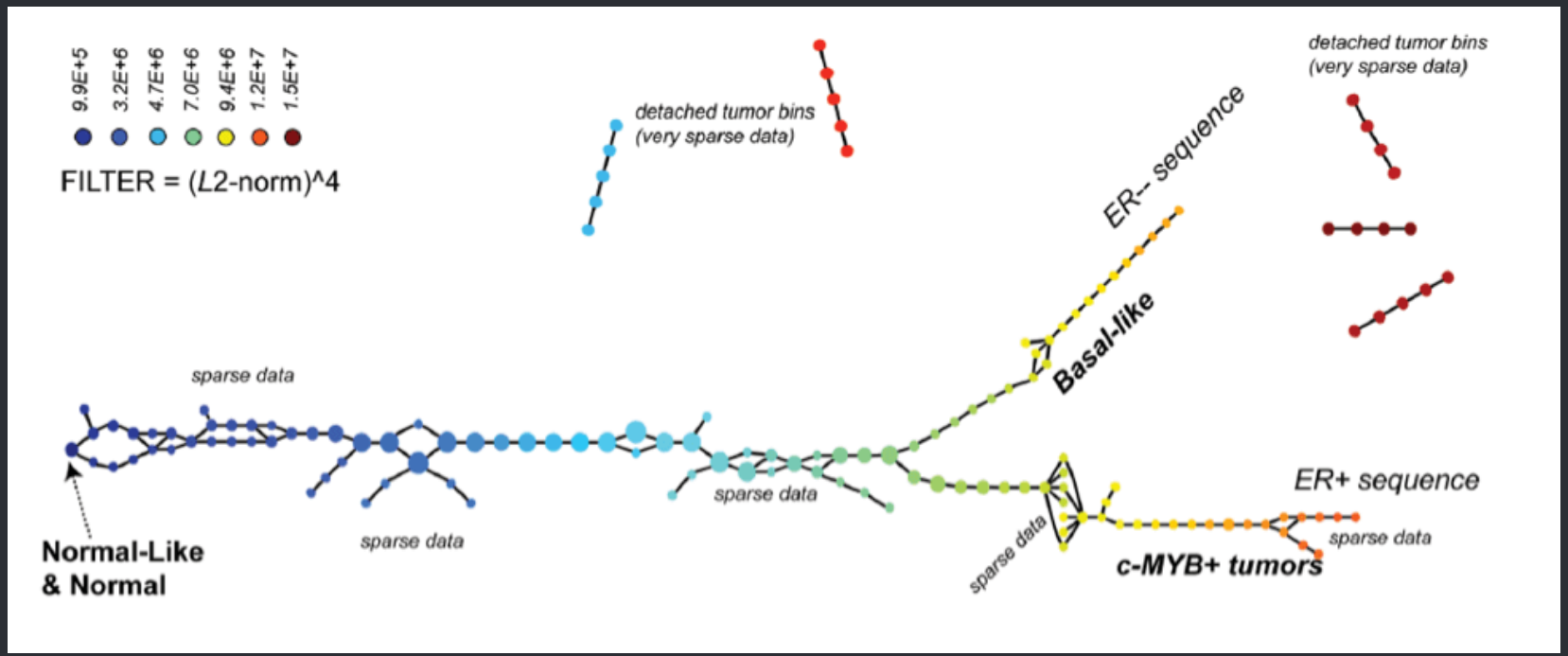
promise made is a debt unpaid



Cancer detection

ft. Mapper

- Nicolau, M., Levine, A.J., Carlsson, G. (2011). *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences of the United States of America, 108 17, 7265-70.



Outline

Python Libraries

1. *Scikit-TDA*
   - *Installation:* `pip install sktda`
   - *Documentation: scikit-tda.org*
   - *Contribution: github.com/scikit-tda*

   1.1 *Persistent Homology*
      - *Installation:* `pip install ripser`
      - *Documentation: https://ripser.scikit-tda.org*

   1.2 *Kepler Mapper*
      - *Installation:* `pip install kmapper`
      - *Documentation: kepler-mapper.scikit-tda.org*

2. *UMAP*
   - *Installation:* `pip install umap-learn`
   - *Documentation: umap-learn.readthedocs.io*
   - *Contribution: github.com/lmcinnes/umap*

Code examples

Public repository:

`https://github.com/ximenafernandez/PyData2019TDA`

Outline

Bibliography

Persistent Homology, *Mapper*, *UMAP* and *general theory*.

[1] Carlsson, G. (2009) *Topology and data.* Bulletin of the American Mathematical Society, 46(2):255–308.

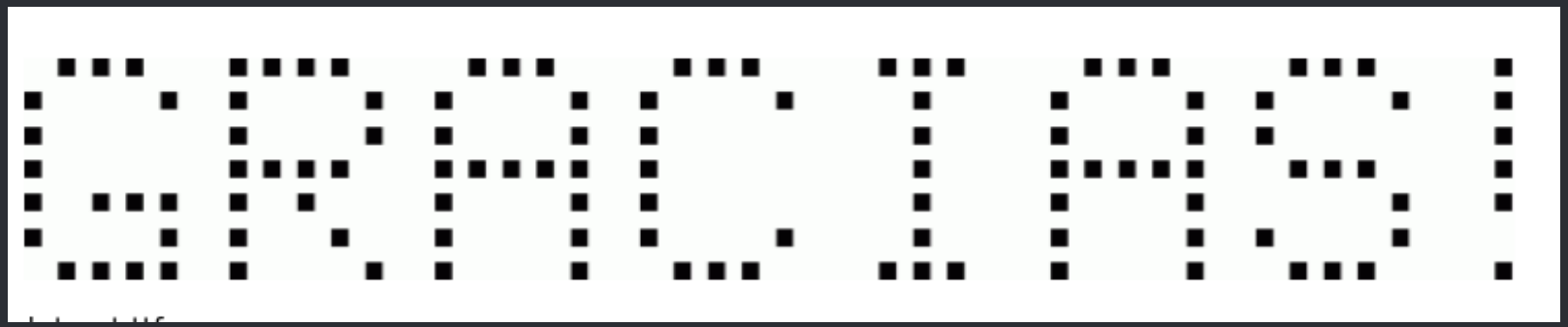
[2] Edelsbrunner, H., Letscher, D., Zomorodian, A. (2002) *Topological persistence and simplification.* Discrete Computational Geometry 28, pages 511–533.

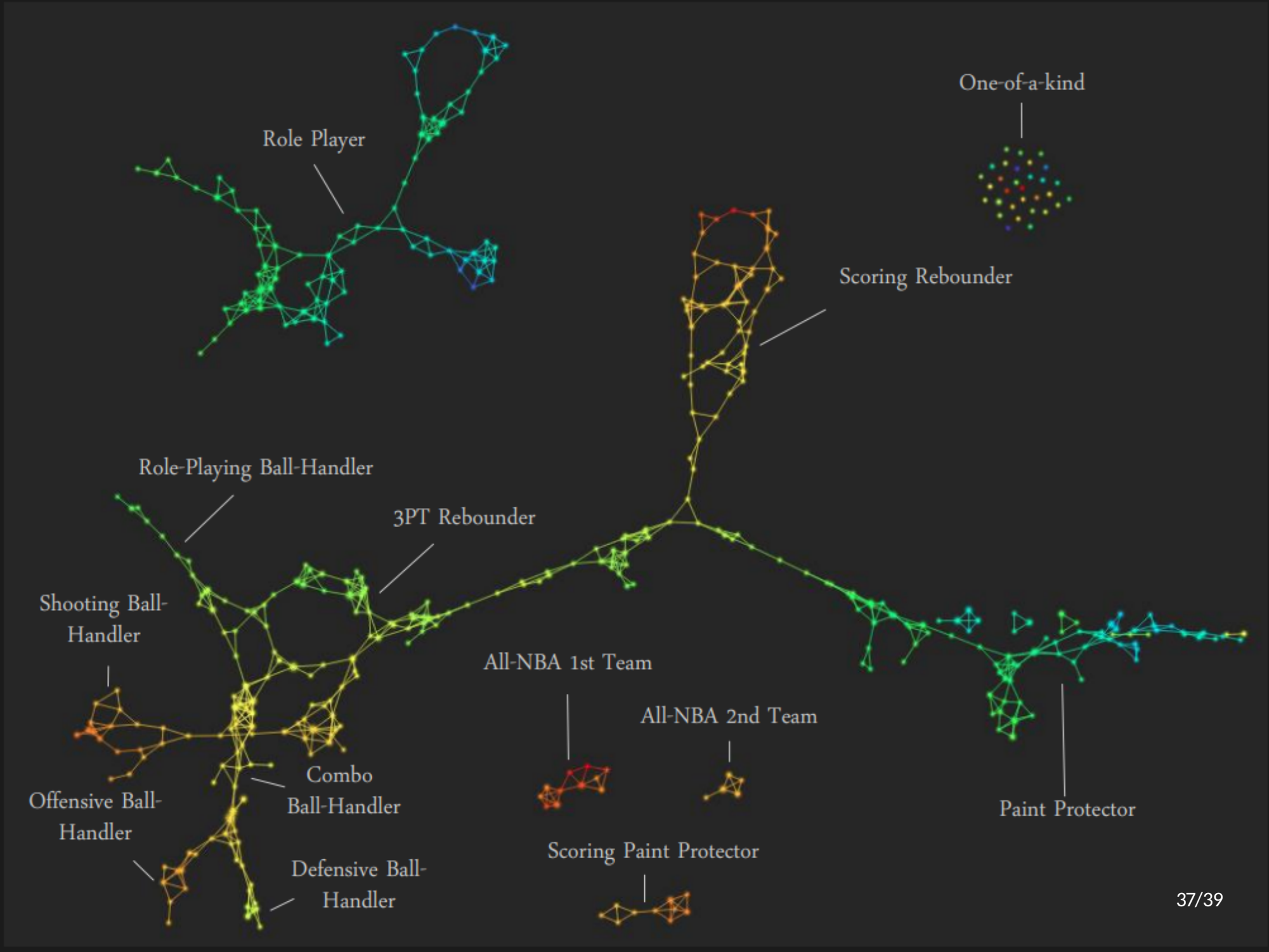
[3] Fernandez, X. Minian, E. (2018). *The Cylinder of a Relation and Generalized Versions of the Nerve Theorem.* Discrete Computational Geometry.

[4] McInnes, L., Healy, J. (2018) *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, ArXiv e-prints 1802.03426.

[5] Singh G., Mémoli F., Carlsson G. (2007) *Topological methods for the analysis of high dimensional data sets and 3d object recognition.* InSPBG, pages 91–100.

[6] Zomorodian, A. (2001) *Computing and Comprehending Topology: Persistence and Hierarchical Morse Complexes.* PhD thesis, University of Illinois at Urbana-Champaign.




One-of-a-kind
Role Player
Scoring Rebounder
Role-Playing Ball-Handler
3PT Rebounder
Shooting Ball-Handler
All-NBA 1st Team
All-NBA 2nd Team
Combo Ball-Handler
Offensive Ball-Handler
Paint Protector
Scoring Paint Protector
Defensive Ball-Handler

- Lum, P.Y., Singh, G., Lehman, A.R., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G.E. (2013). *Extracting insights from the shape of complex data using topology.* Scientific reports.

DATA: rates (per minute played) of rebounds, assists, turnovers, steals, blockedshots, personal fouls, and points scored.

- **Offensive Ball-Handler.** This guy handles the ball and specializes in points, free throws and shots attempted, but is below average in steals and blocks. Examples include Jason Terry and Tony Parker.
- **Defensive Ball-Handler.** This is a defense-minded player who handles the ball and specializes in assists and steals, but is only so-so when it comes to points, free throws and shots. See also: Mike Conley and Kyle Lowry.
- **Combo Ball-Handler.** These players are adept at both offense and defense but don't stand out in either category. Examples include Jameer Nelson and John Wall.
- **Shooting Ball-Handler.** Someone with a knack for scoring, characterized by above-average field goal attempts and points. Stephen Curry and Manu Ginobili are examples. Role-Playing Ball-Handler. These guys play fewer minutes and don't have as big a statistical impact on the game. Hello, Arron Afflalo and Rudy Fernandez.

- **3-Point Rebounder.** Such a player is a ball-handler and big man above average in rebounds and three-pointers, both attempted and made, compared to ball-handlers. Luol Deng and Chase Budinger fit the bill.
- **Scoring Rebounder.** He grabs the ball frequently and demands attention when on offense. Dirk Nowitzki and LaMarcus Aldridge play this position.
- **Paint Protector.** A big man like Marcus Camby and Tyson Chandler known for blocking shots and getting rebounds, but also for racking up more fouls than points. Scoring Paint Protector. These players stand out on offense and defense, scoring, rebounding and blocking shots at a very high rate. Examples include Kevin Love and Blake Griffin.
- **NBA 1st-Team.** This is a select group of players so far above average in every statistical category that the software simply groups them together regardless of their height or weight. Kevin Durant and LeBron James fall in this category.
- **NBA 2nd-Team.** Not quite as good, but still really, really good. Rudy Gay and Caron Butler are examples.
- **Role Player.** Slightly less skilled than the 2nd-team guys, and they don't play many minutes. Guys like Shane Battier and Ronnie Brewer fall under this position.
- **One-of-a-Kind.** These guys are so good they are off the charts — literally. The software could not connect them to any other player. Derrick Rose and Dwight Howard are examples, but you already knew that.