Mapper, Manifolds, and More!

Topological Data Analysis and Mapper

Matt Piekenbrock

Data Science and Security Cluster (DSSC) Talk: March 28th, 2018

Topology

- Topology is a branch of mathematics concerned with **spaces** and **maps**
  - Formalizes notions of *proximity* and *continuity*
- What topology is useful for:
  - Feature characterization
    - e.g. suitable for classification
  - Formalize notions of robustness
    - "At what noise level will my estimate be off by X amount?"
  - Learning global representations from *local information*
  - Complexity / Feasibility
    - Proving asymptotic behavior of functions (e.g. Big-O)

Much of this definition is based on Robert Ghrist's Book: Ghrist, Robert W. Elementary applied topology. Seattle: Createspace, 2014.

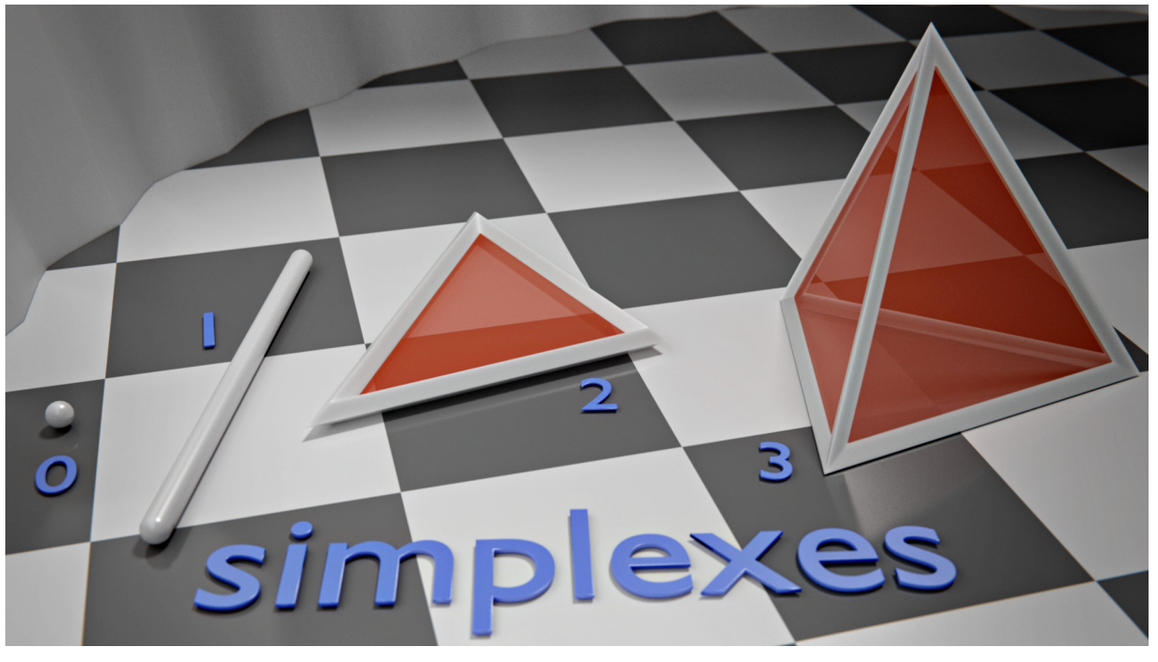
Topological Data Analysis (TDA)

- TDA comprises “a collection of powerful tools that can quantify shape and structure in data in order to answer questions from the data’s domain.” [Munch]
- Is an *emerging* field for data analysis!
- Motivation for TDA:
  - Data is huge, often high dimensional, and complex
  - Traditional techniques have not “kept up”
    - i.e. Rely on *overly-simplistic* assumptions
- Basic Idea:
  - **Data has shape**
  - This shape can be rigorously quantified via topological signatures

Much of this is summarized from: Munch, Elizabeth. "A user’s guide to topological data analysis." Journal of Learning Analytics 4.2 (2017): 47-61.

Whats a Topological Signature?

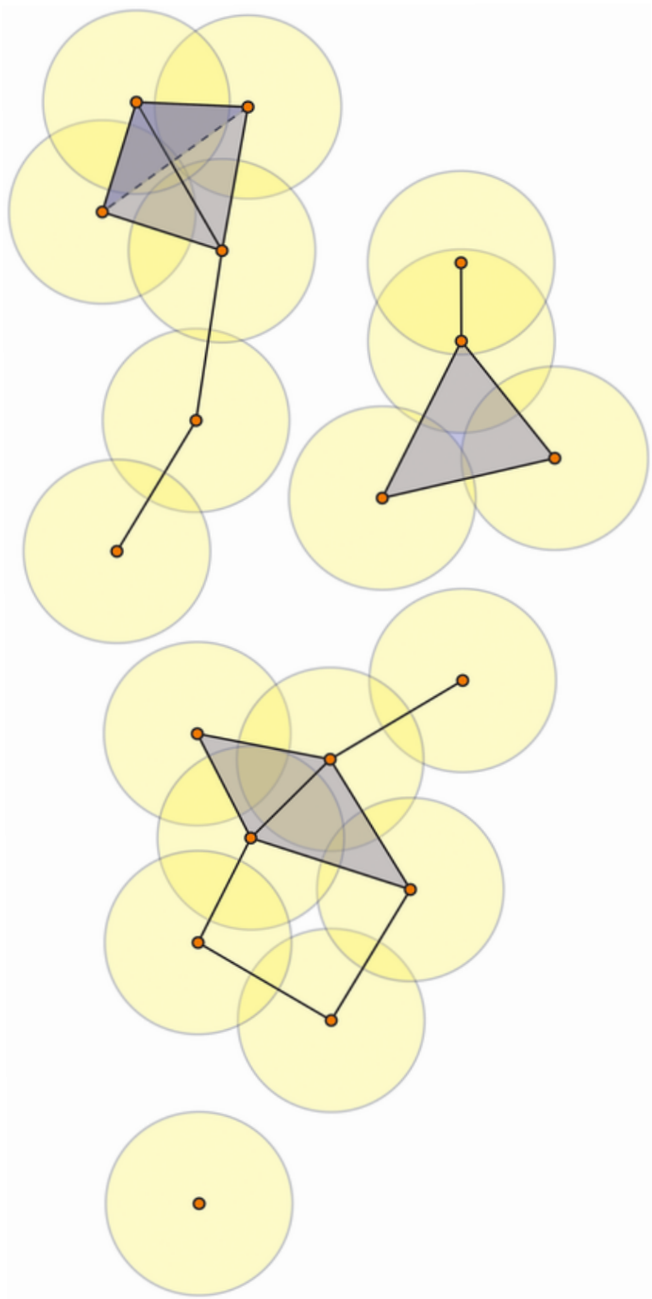
- Informally, a topology on a set is a *description of how elements in the set are spatially related*
  - Can be seen as a formalization of clustering
  - i.e. *the collection of all open sets in a space is called its topology*
- A topological signature is a **simplified representation** of the topology of a given space
- Often, a [discrete] representation used as a topological signature is a **simplicial complex**
  - 0-simplex == vertex
  - 1-simplex == edge
  - 2-simplex == triangle
  - 3-simplex == tetrahedron
  - … *k*-simplex == …



Much of this is summarized from: Munch, Elizabeth. "A user's guide to topological data analysis." Journal of Learning Analytics 4.2 (2017): 47-61.

Example of a Topological Signature

- Perhaps the most common topological signature is the so-called Rips Complex $VR_\epsilon(X)$ formed by forming an simplex between all points which have pairwise distances less than ϵ



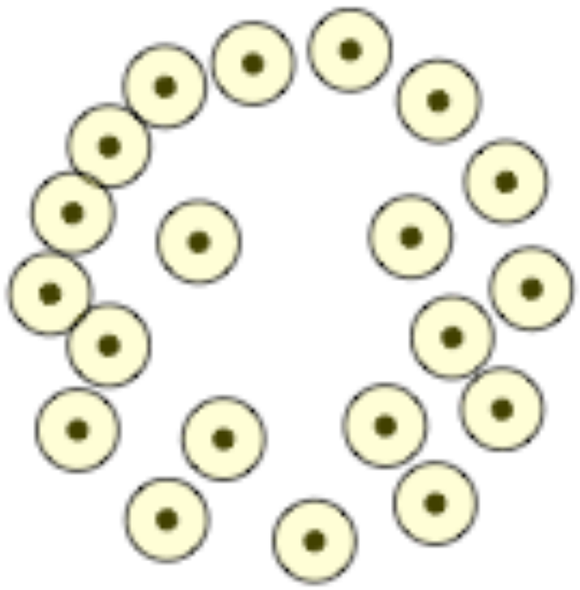
- The simplicial complex formed by non-empty intersections of

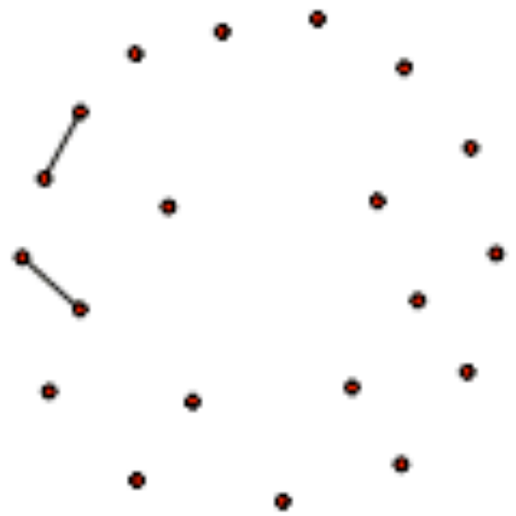
$$B(X_1, \epsilon) \cap B(X_2, \epsilon) \cap \cdots \cap B(X_n, \epsilon)$$

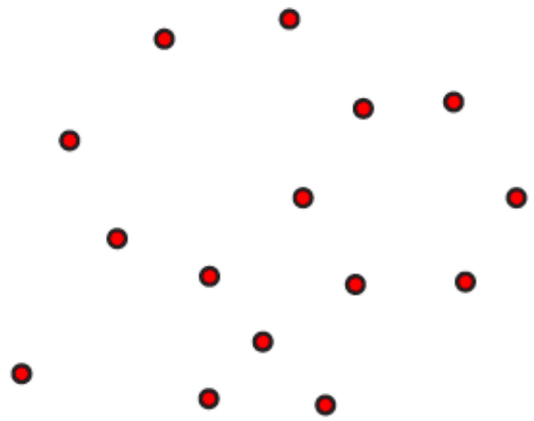
where

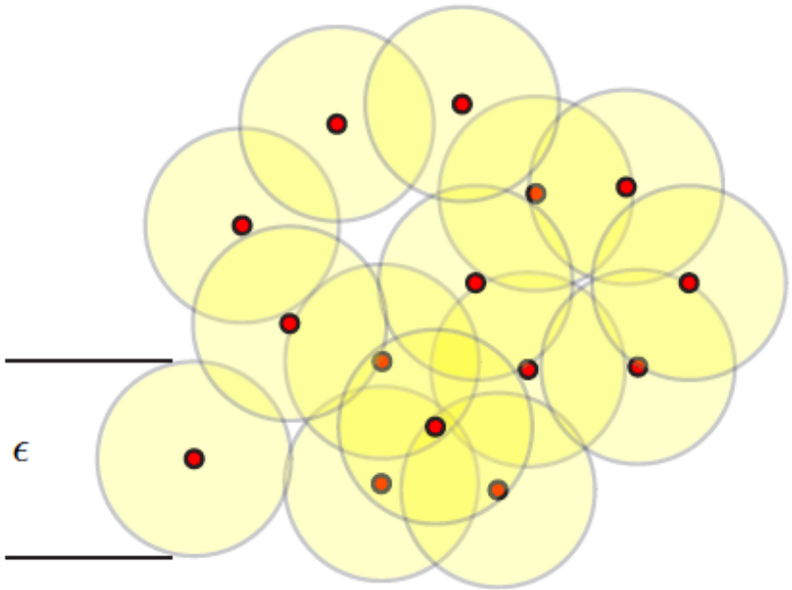
$$B(X_i, \epsilon) = \{x \in X \mid d(X_i, x) \leq \epsilon\}$$

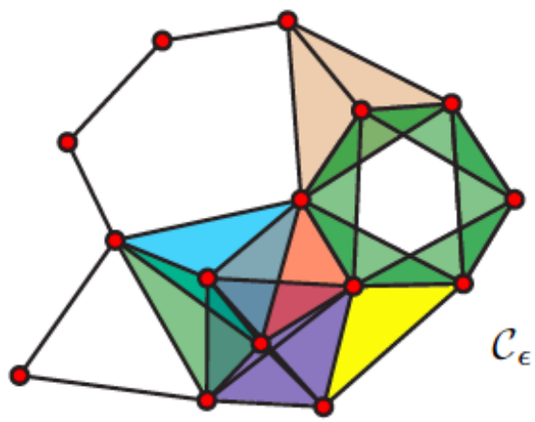
Picture from: https://www.researchgate.net/publication/269876798_Topological_Data_Analysis_of_Biological_Aggregation_Models/figures?lo=1&utm_source=google&utm_medium=organic

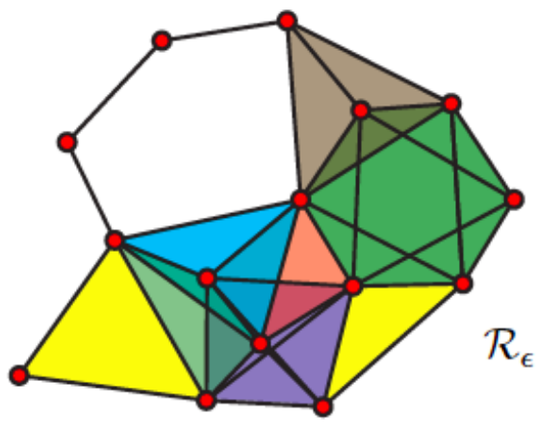







ϵ


$\mathcal{C}_\epsilon$


$\mathcal{R}_\epsilon$

Recall: The Goal of TDA

- Basic Idea:
  - **Data has shape**
  - This shape can be rigorously quantified with topology
  - Quantify shape through topological signatures
  - Such signatures act as **summaries of the data**
- The Goal of TDA:
  - Use tools from topology to make meaningful signatures of the data
  - Topological signatures lead to topological **invariants**, and such invariants ***enable greater understanding of the relationships in—and transformations of—real data***

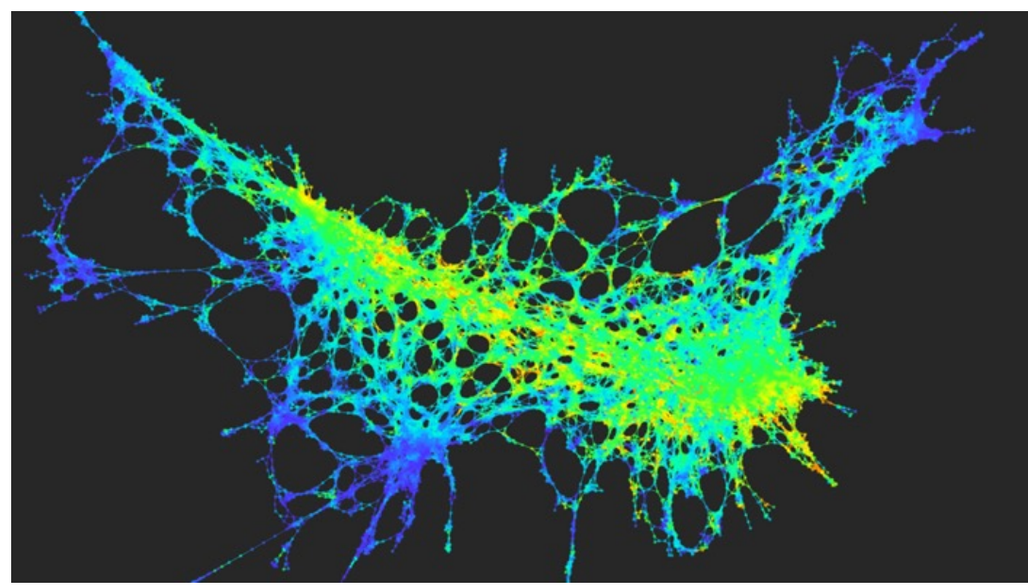
Much of this is gleaned from either:

1. Munch, Elizabeth. "A user's guide to topological data analysis." Journal of Learning Analytics 4.2 (2017): 47-61.
2. Ghrist, Robert W. Elementary applied topology. Seattle: Createspace, 2014.

Mapper: a Topological Signature

- Mapper: *Perhaps* the most used signature in modern TDA *applications*
- Created by Singh, Mémoli, and Carlsson [Mapper]
- Simplest interpretation:
  - Interprets any set of data in $\mathbb{R}^d$ as "point cloud data", turns data into a *simplified topological graph*
  - "Mapper takes as input both a possibly high dimensional dataset and a map defined on the data, and produces a **summary of the data** by using a cover of the codomain of the map."

	V1	V2	V2A	V3	V4	V5	V6	V7	V8	V9	V10	V11
1	6	12	12	1	1	1	1	-2	1	1	2	1
2	6	12	12	2	1	2	3	4	2	2	2	2
3	6	12	12	3	1	3	2	4	2	1	2	2
4	6	12	12	4	1	1	3	4	3	1	2	1
5	6	12	12	5	1	1	1	2	1	1	1	3
6	6	12	12	6	1	2	2	2	4	1	2	1
7	6	12	12	7	1	1	1	1	1	1	2	2
8	6	12	12	8	1	1	1	1	2	2	2	1
9	6	12	12	9	1	1	1	2	2	2	2	2
10	6	12	12	10	1	1	1	2	1	1	1	1
11	6	12	12	11	1	2	3	4	1	1	2	4
12	6	12	12	12	1	1	2	3	2	2	1	2
13	6	12	12	13	1	2	3	3	4	4	2	3
14	6	12	12	14	1	1	1	1	1	1	2	2
15	6	12	12	15	1	2	3	4	4	1	2	3



Picture from: https://www.researchgate.net/publication/269876798_Topological_Data_Analysis_of_Biological_Aggregation_Models/figures?lo=1&utm_source=google&utm_medium=organic

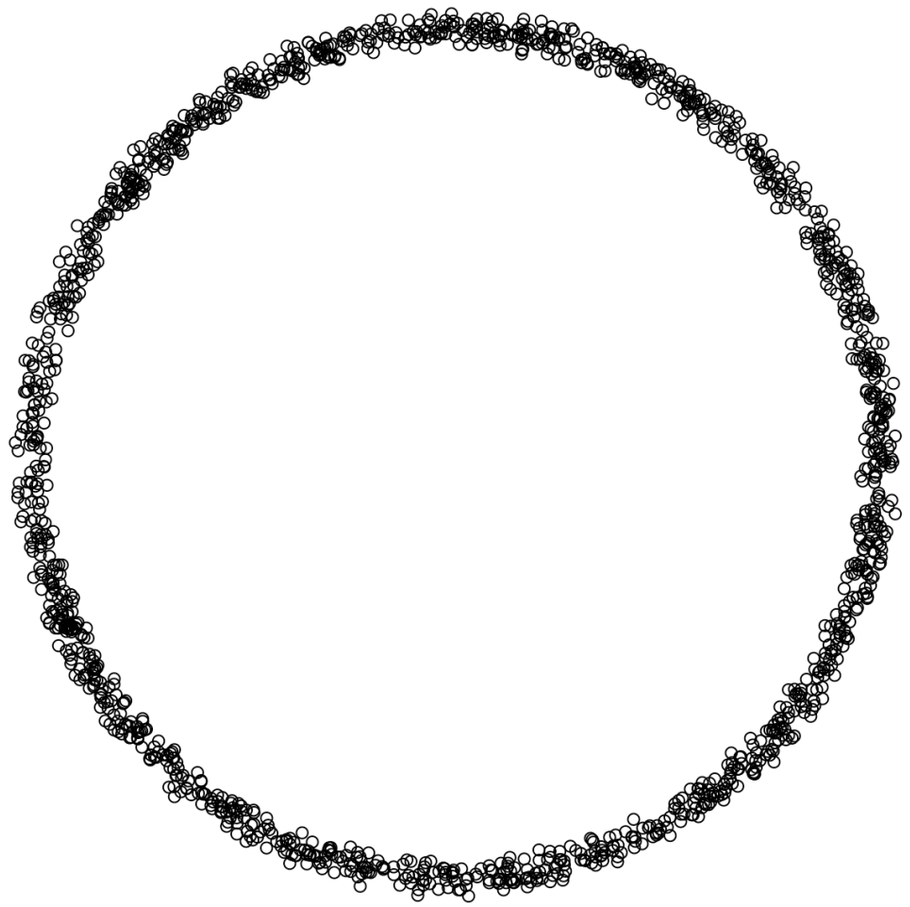
Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

Mapper: Example

Consider the case where your data looks like a circle



$$X = \{x_1, x_2, \ldots, x_n\}$$

Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

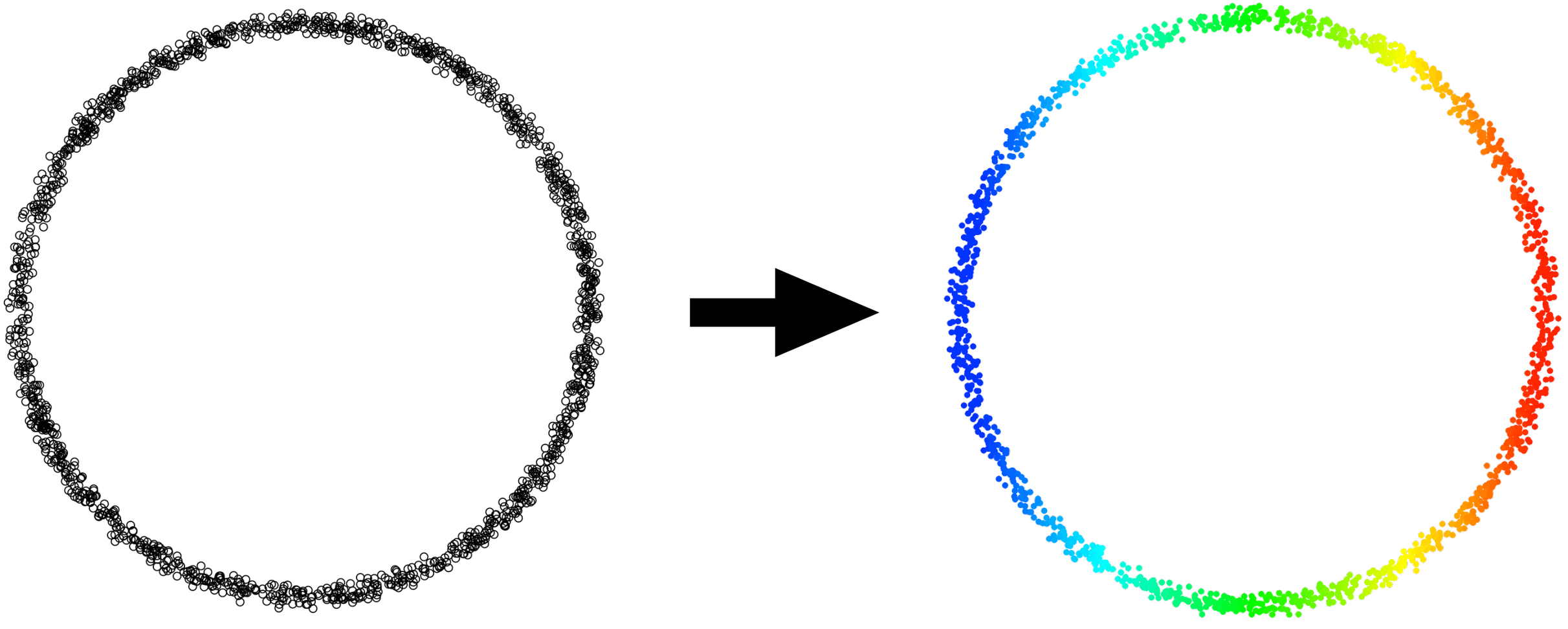
Mapper: Example — Step 1

… and you evaluate a function f

on that circle $f(x) = \|x - p\|_2$

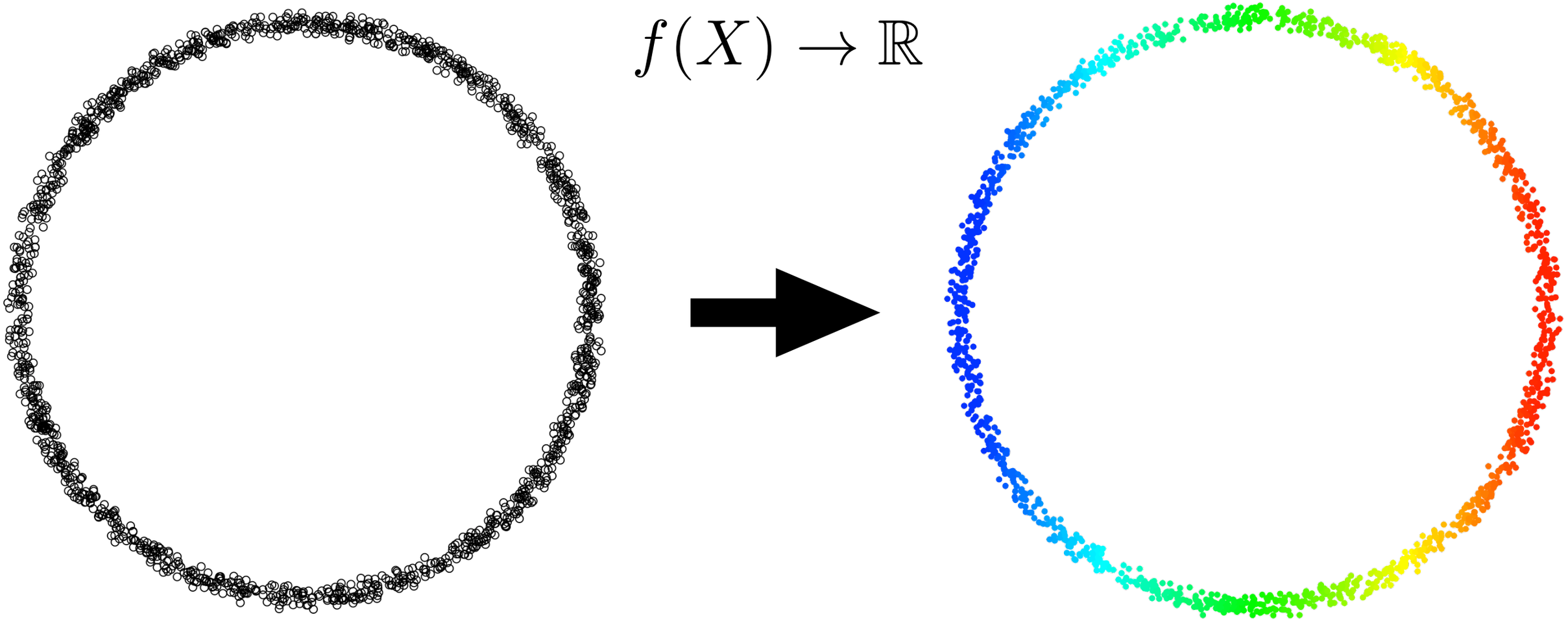
where p is the left-most point in the circle

(blue == low distance, red == high distance)

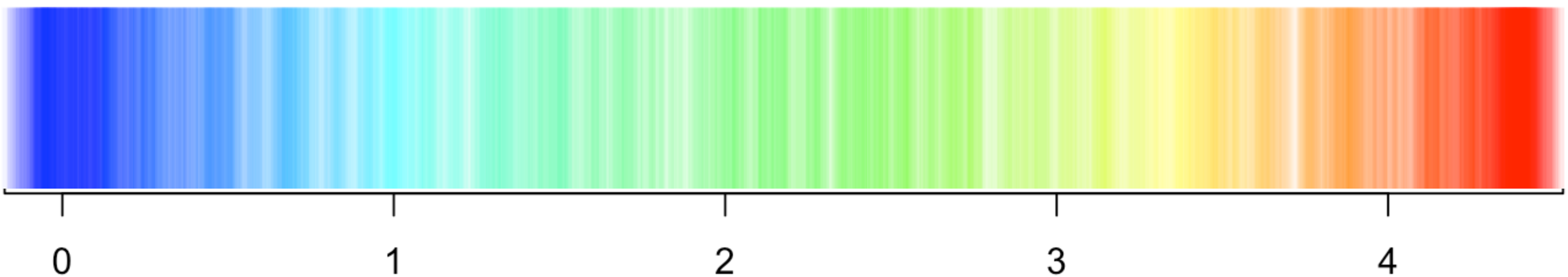


Mapper: Example — Step 1

$$f(X) \to \mathbb{R}$$



$$Z \subset \mathbb{R}$$



Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

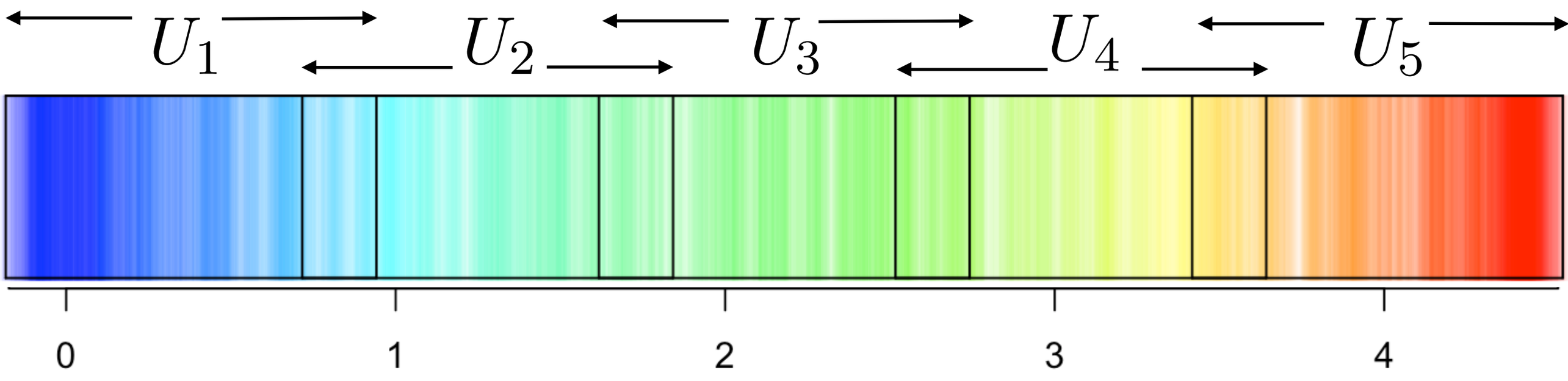
Mapper: Example — Step 2

Define a cover of the filter space Z

Recall the definition of a cover is a collection of sets whose union contains some space as a subset

$$C = \{U_\alpha : a \in A\} \qquad X \subseteq \bigcup_{\alpha \in A} U_\alpha$$

In the case above, *C* covers *X*

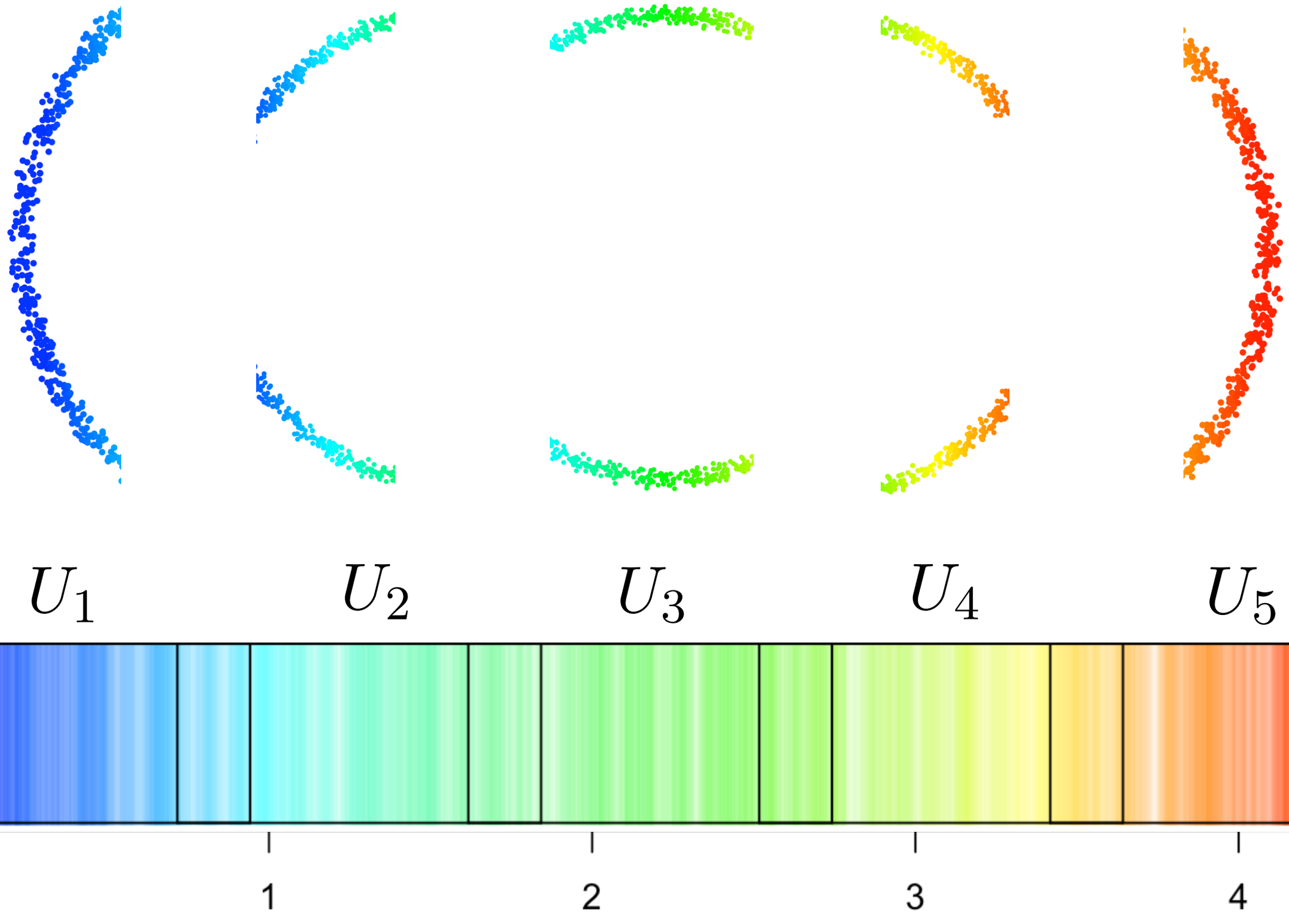


Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

Mapper: Example — Step 3

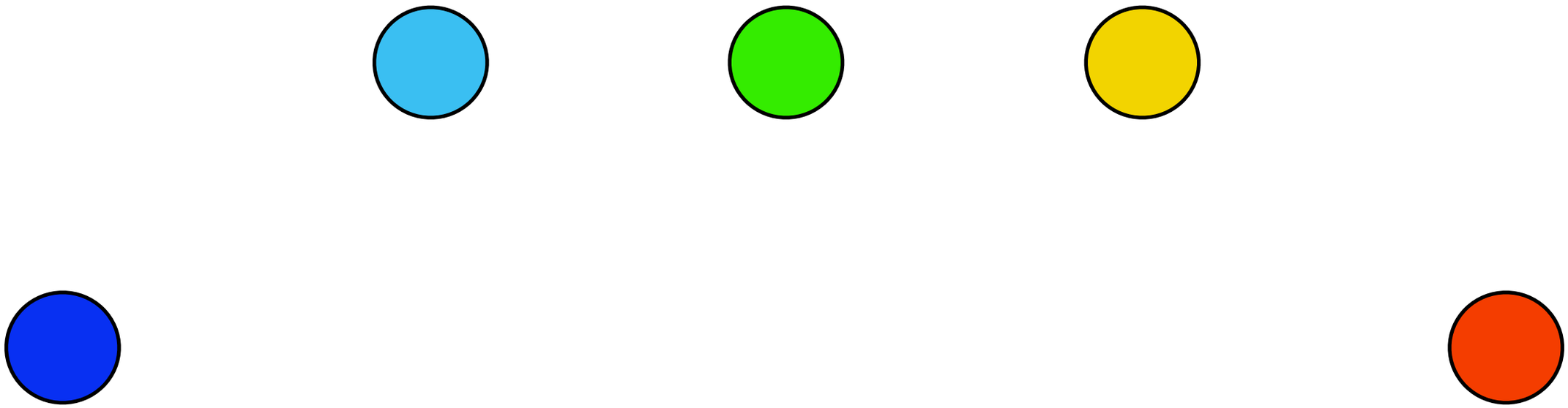


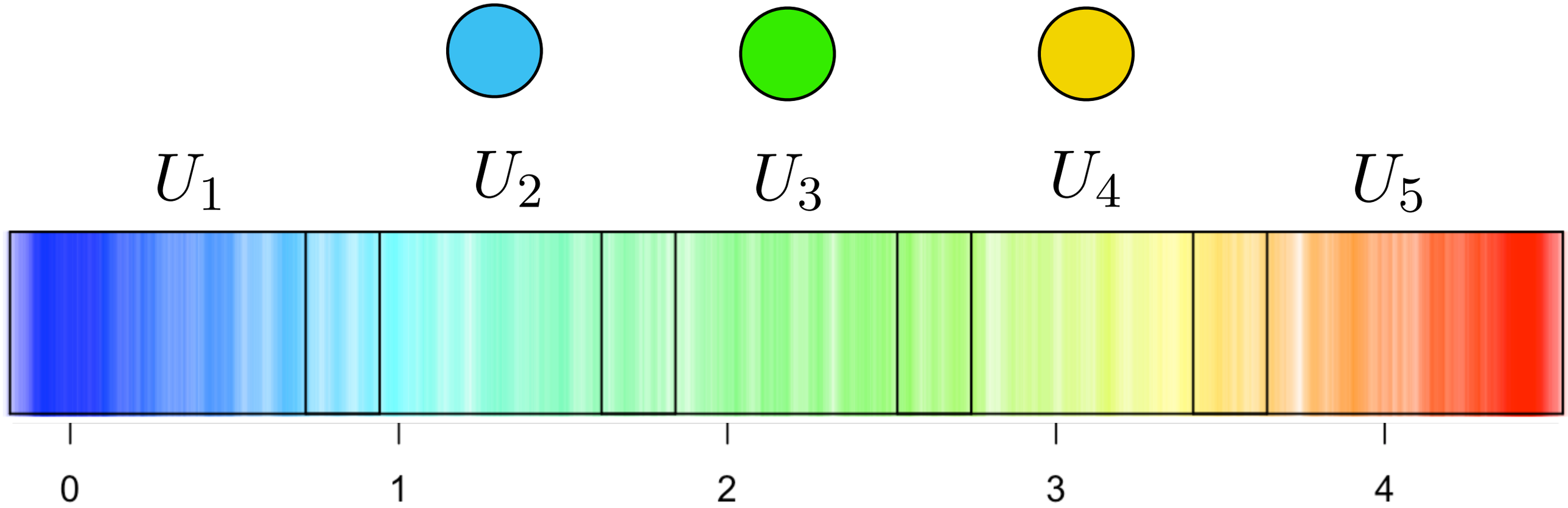
Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

Mapper: Example — Step 4



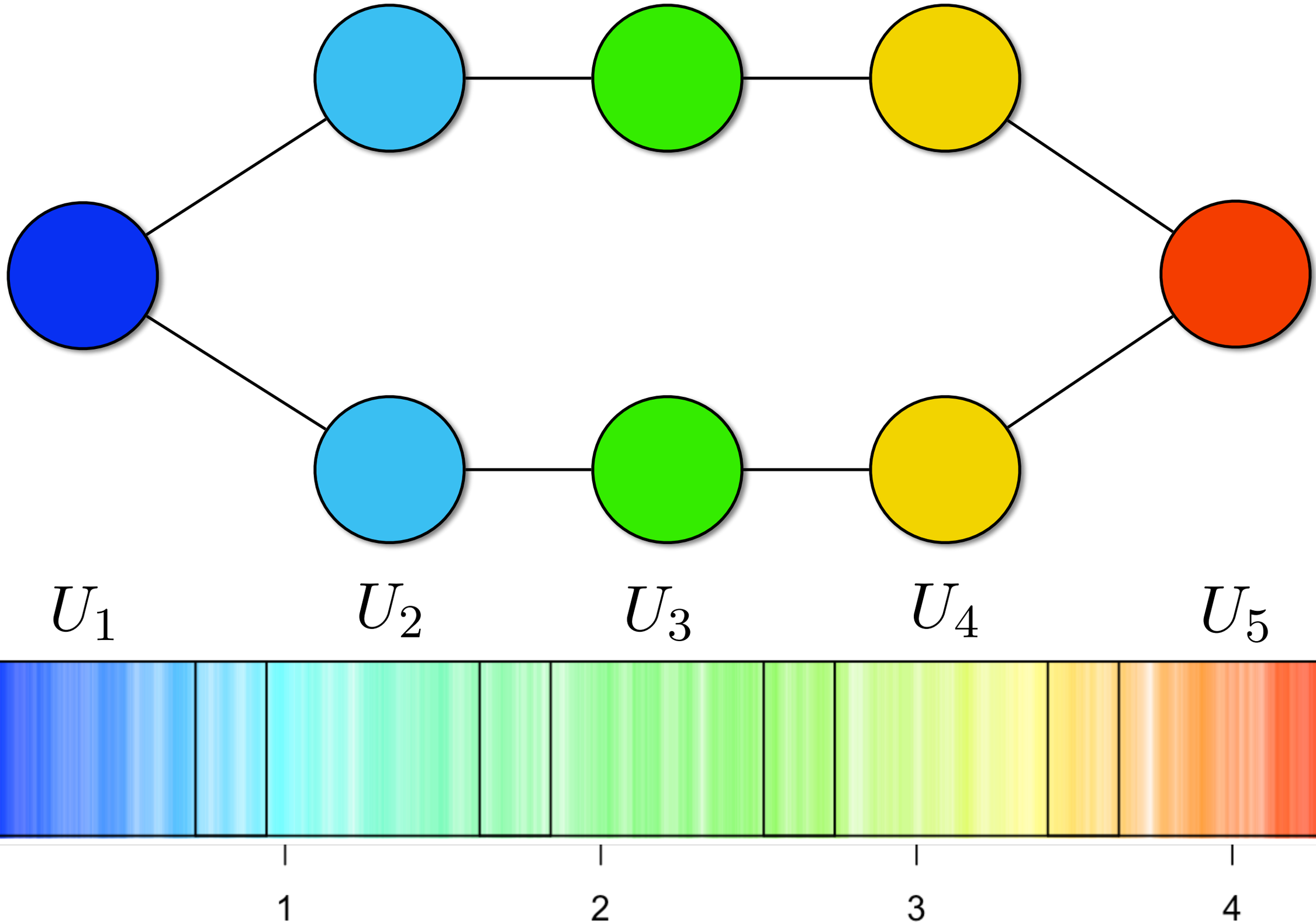


Mapper: Background

1. Define a reference map $f : X \to Z$
2. Construct a covering $\{U_\alpha\}_{\alpha \in A}$ of Z
   - A is called the index set
3. Construct the subsets X_α via $f^{-1}(U_\alpha)$
4. Apply a clustering algorithm $\mathcal{C}$ to the sets X_α
5. Obtain a cover $f^*(\mathcal{U})$ of X_α by considering the path-connected components of X
   - Clusters form "nodes" / 0-simplexes
   - Non-empty intersections form "edges" / 1-simplexes
6. The Mapper construction is the nerve of this cover

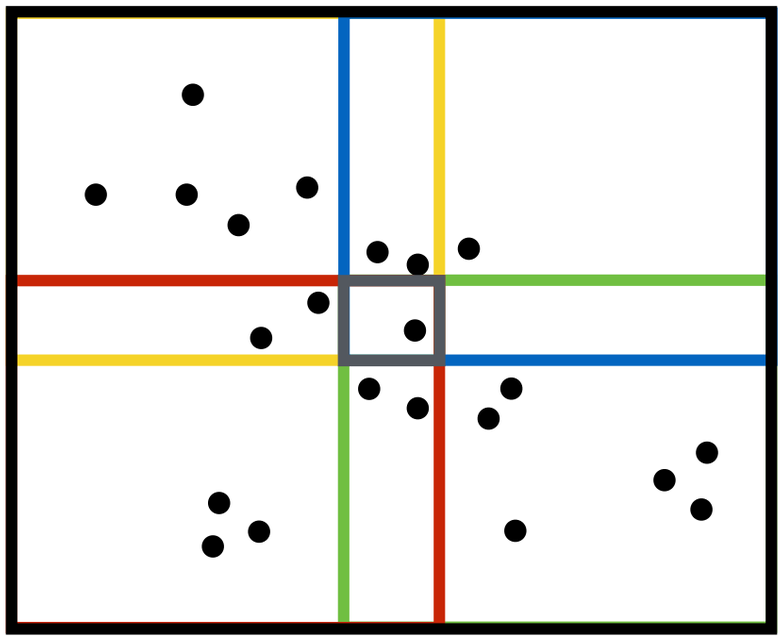
$$M(\mathcal{U}, f) := N(f^*(\mathcal{U})))$$

Mapper: Example — Step 5-6



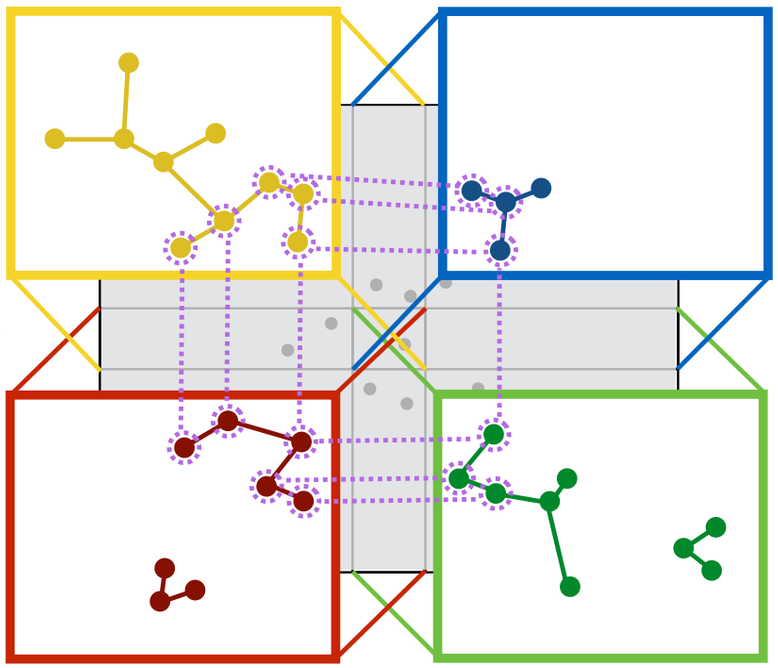
Mapper: In one picture

$f : X \to Z$



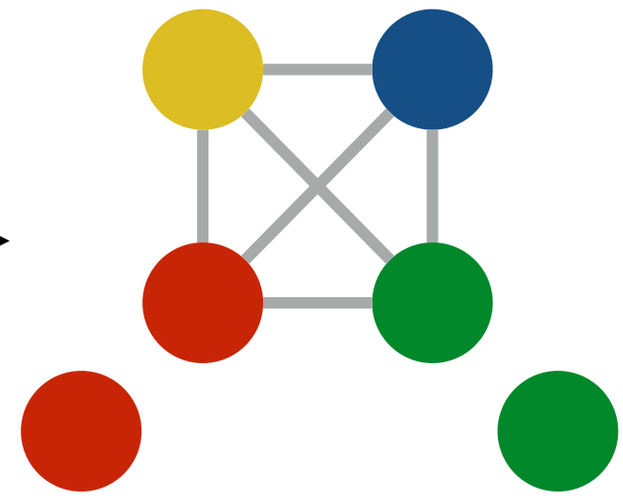
$\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$

$X_\alpha = f^{-1} U_\alpha$



$C(X_\alpha) \to f^*(\mathcal{U})$

$f^*(\mathcal{U}) \to M(\mathcal{U}, f)$



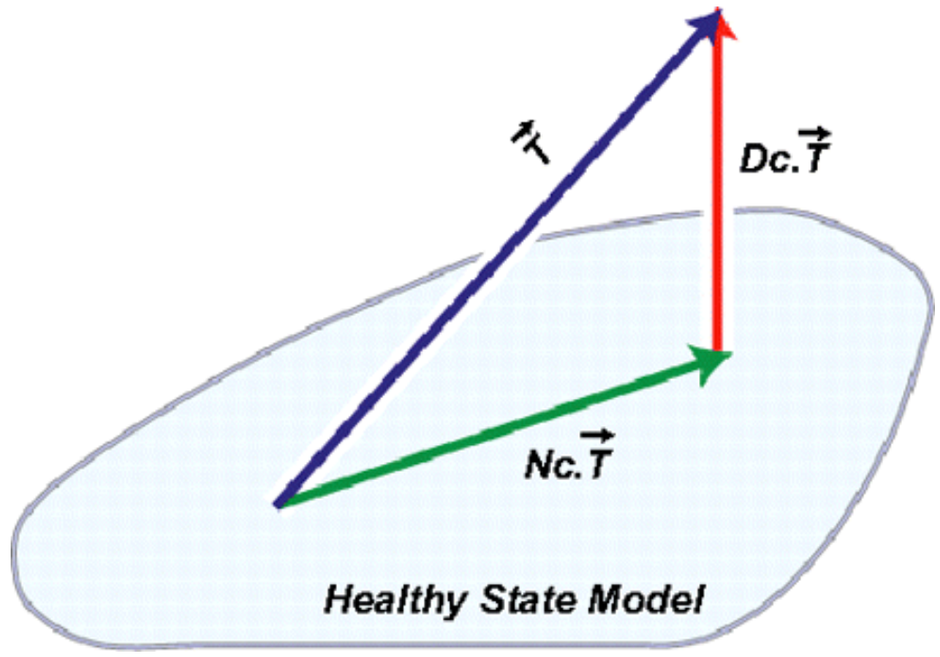
(1-skeleton example)

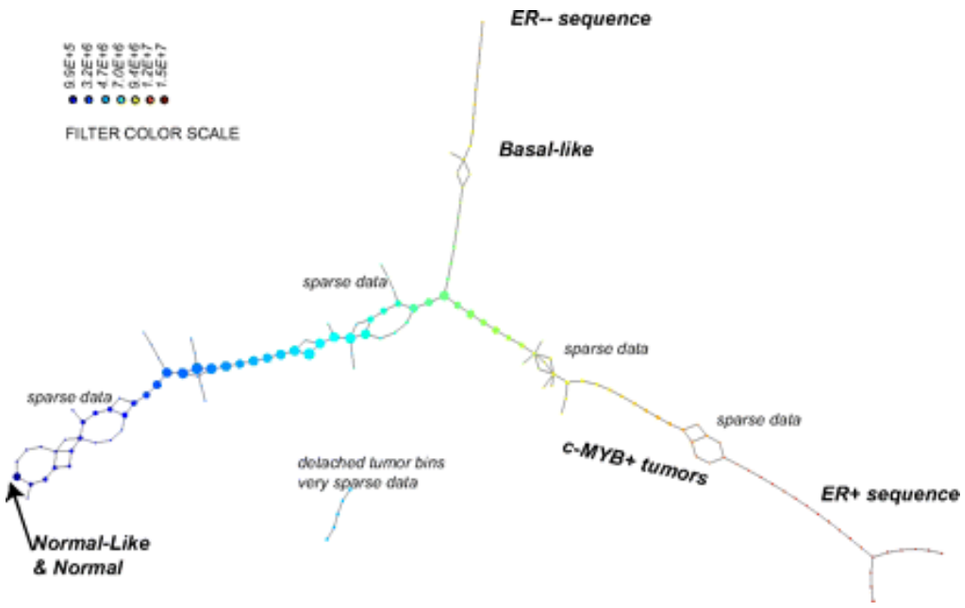
Why Mapper is useful

- Mapper provides a **succinct summary** of **the shape of a data set**
  (expressed via the codomain of the mapping)
- Mappers utility lies in its **generality**:
  - *Any* mapping function can be used
  - Cover may be constructed *arbitrarily*
  - *Any* clustering algorithm may be used
- The resulting graph is often much easier to interpret than, e.g. individual scatter plots of pairwise relationships
- Mapper is often paired with **high-dimensional data**, and is generally used to see the ‘true’ shape or structure of the data
- Mapper is the core algorithm behind the AI Company, Ayasdi Inc.
  - Anti-Money Laundering
  - Detecting Payment Fraud
  - Assessing health risks

How to choose a mapping function?

- Mapper is **highly dependent on the choice of filter function**
- Ideally, a domain-specific map that is well-understood may be appropriate
  - Ex. Biologists created a Healthy State Model (HSM) which encodes tumor cell tissues into orthogonal "disease" and "normal" components
  - Using the disease component allows for disease-specific analysis of their data
- What if we want a general filter function that can be used for many kinds of data sets?

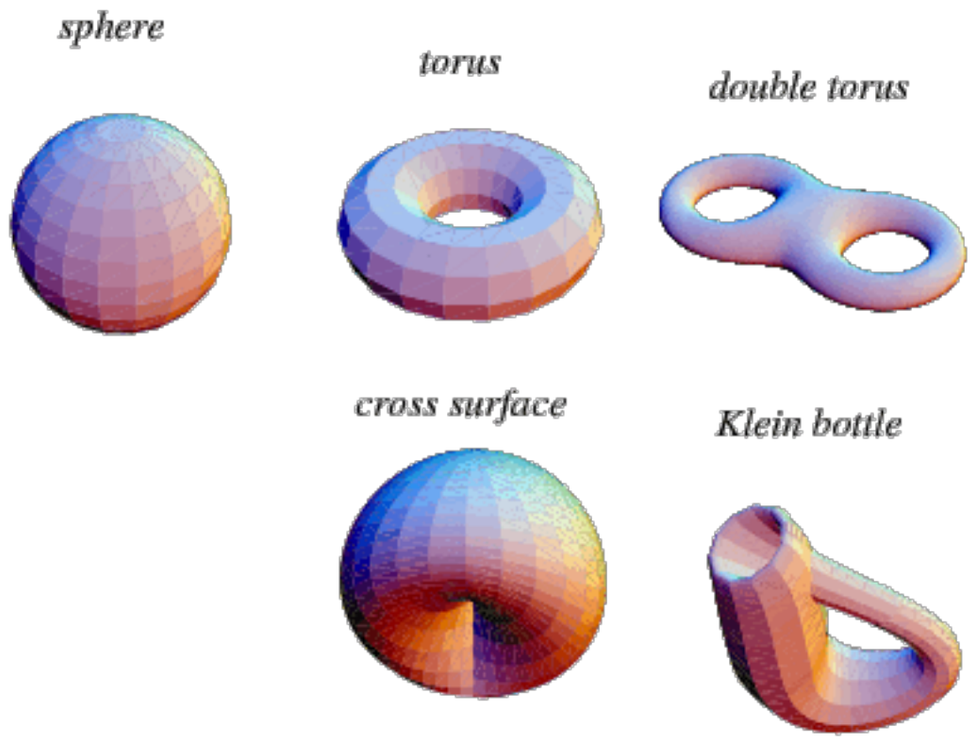




Pictures from: http://www.pnas.org/content/108/17/7265.long

Manifold Learning

- What is a Manifold?
  - A topological space that ***locally resembles Euclidean space***
  - A manifold [generally] is "smooth" if it permits the use of partial differentiation*
  - Ex. of Manifolds: The Earth, a Torus, a swiss roll



- The Manifold Hypothesis
  - The manifold hypothesis is the idea that data tend to lie *on or near a low dimensional manifold*
  - Alt. Def: The dimensionality of data is only **arbitrarily high**; rather, data may exist in some "**intrinsic**" dimensionality

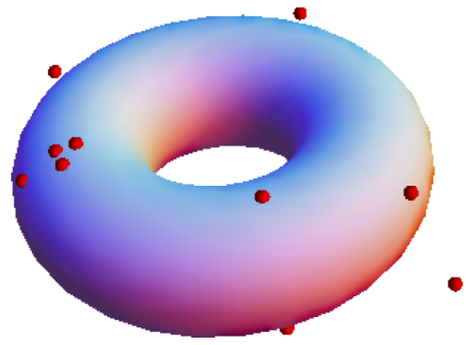
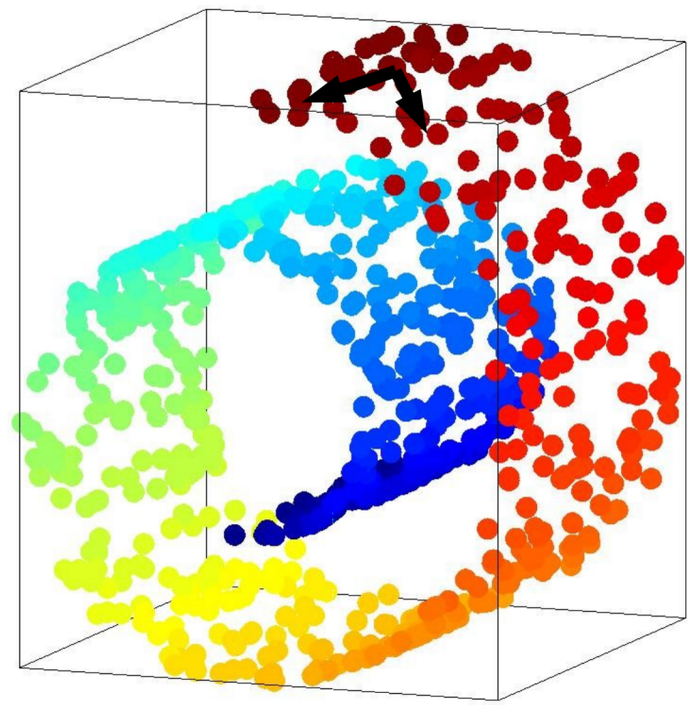


FIGURE 1. Data lying in the vicinity of a two dimensional torus.

*Technically, a Manifold is *smooth* when a Hausdorff space is furnished with an atlas of charts, each of which are infinitely differentiable whenever they intersect
Image from: Fefferman, Charles, Sanjoy Mitter, and Hariharan Narayanan. "Testing the manifold hypothesis." Journal of the American Mathematical Society 29.4 (2016): 983-1049.

IsoMap Example

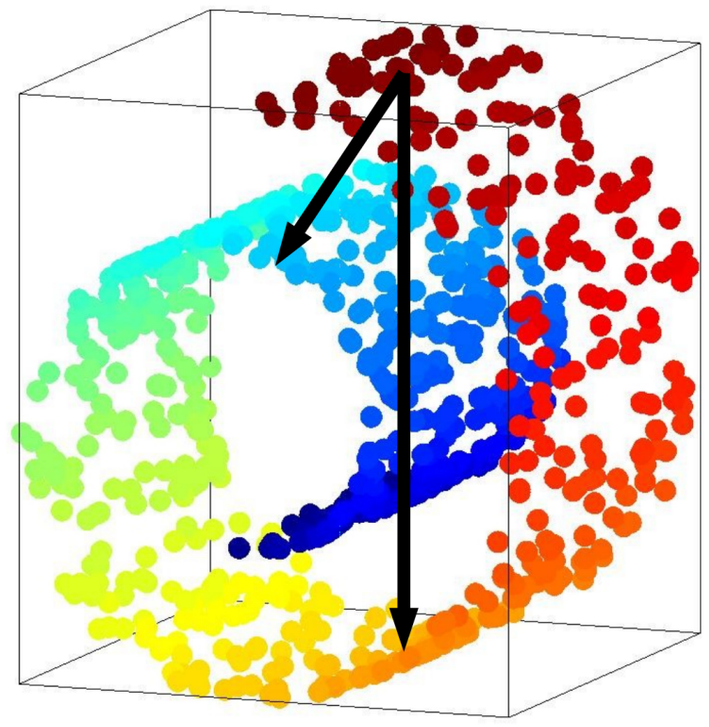
1. Make a neighborhood graph, connected points that are either:
   - Within some distance ϵ to each other
   - Is the k -th nearest neighbor of another



2. Compute Shortest path (Djikstra's)
3. Compute lower-dimensional embedding (MDS)
   - Goal: Given distance matrix $D = (d_{ij})$, compute the set:

$$x_1, \ldots, x_n \in \mathbb{R}^m$$

Such that: $d_{ij} \approx \|x_i - x_j\|_2$

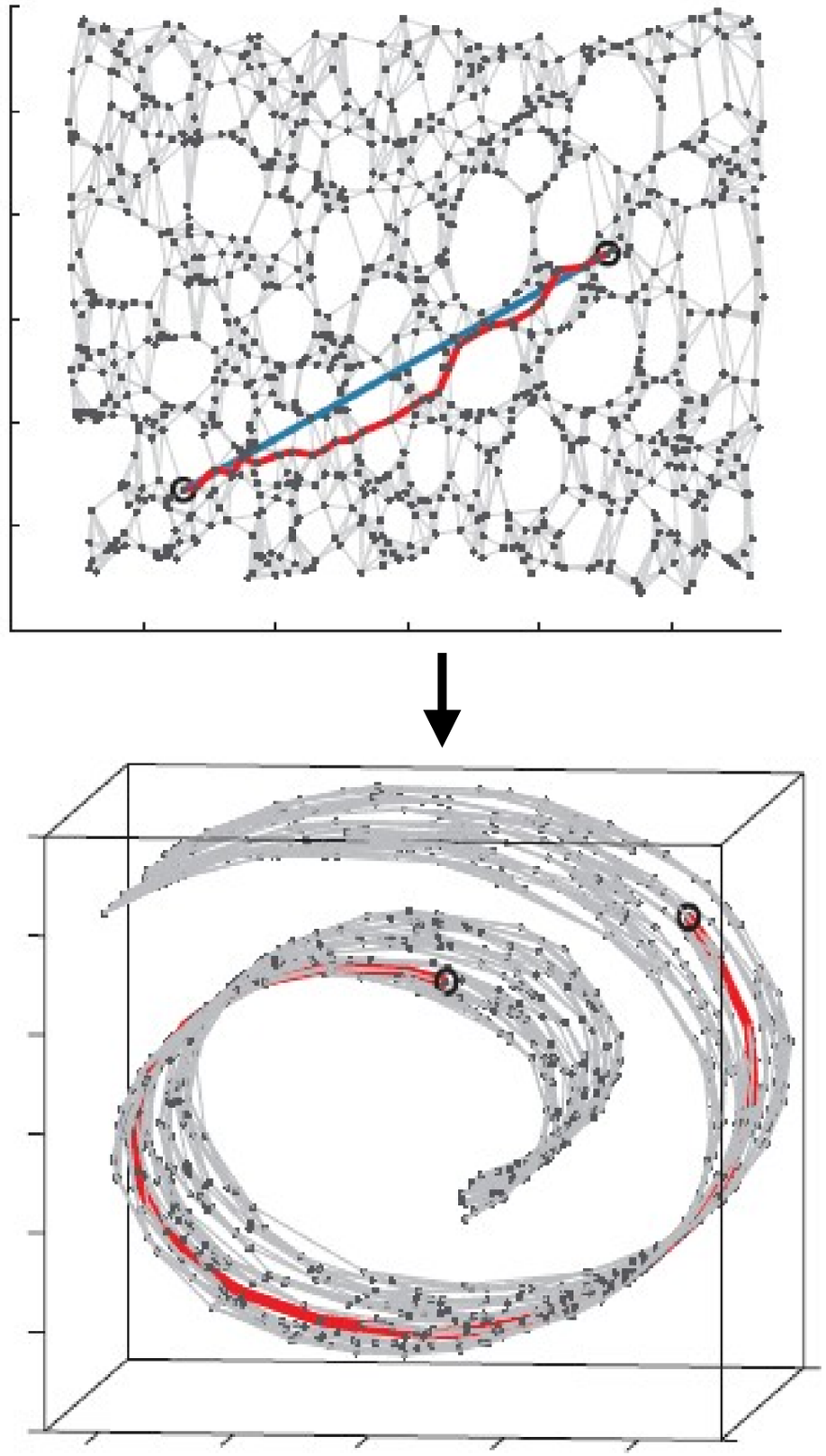


IsoMap Example

1. Make a neighborhood graph, connected points that are either:
   - Within some distance ϵ to each other
   - Is the k -th nearest neighbor of another
2. Compute Shortest path (Djikstra's)
3. Compute lower-dimensional embedding (MDS)
   - Goal: Given distance matrix $D = (d_{ij})$, compute the set:

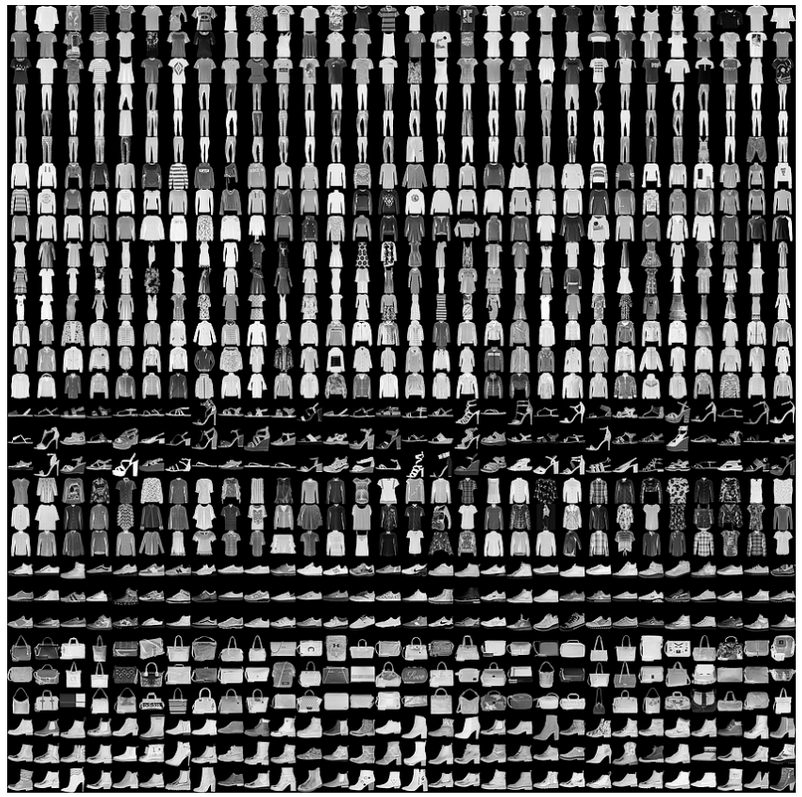
     $$x_1, \ldots, x_n \in \mathbb{R}^m$$

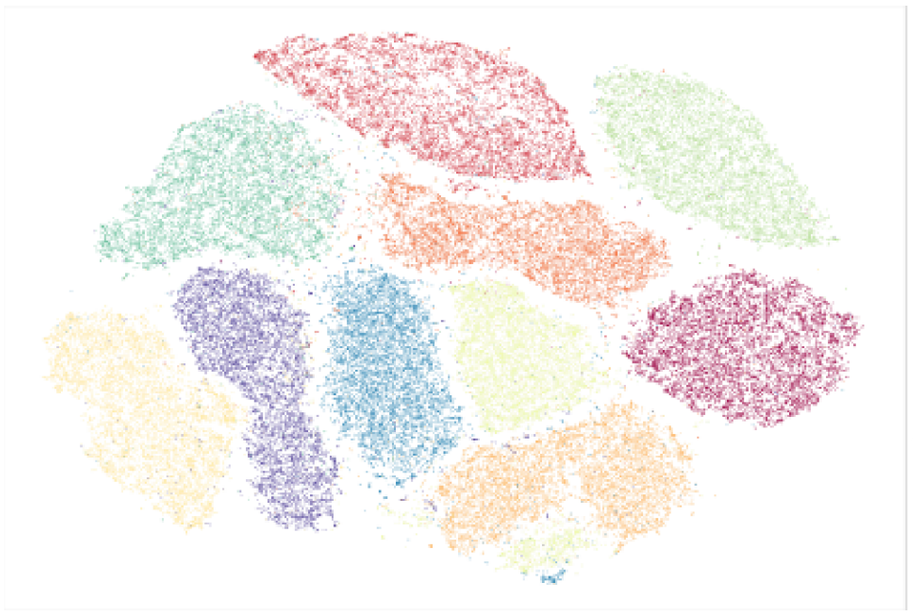
     Such that: $d_{ij} \approx \|x_i - x_j\|_2$



Why care about the Manifold Hypothesis?

- There's a lot of theoretical [3, 4], experimental [2], and empirical [1, 3] evidence supporting the Manifold Hypothesis
- Related: Non-linear Dimensionality Reduction
  - SNE, t-SNE, IsoMap, Locally Linear Embedding
- There's *several ways* to get a different perspective of the manifold; there is (to my knowledge) no "best" manifold approximation technique
- The hypothesis has actually motivated developments in Generative Adversarial Networks [3]





t-SNE applied to MNIST

1. http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/#fn3
2. Carlsson, Gunnar, et al. "On the local behavior of spaces of natural images." International journal of computer vision 76.1 (2008): 1-12.
3. Lui, Kry Yik Chau, et al. "Implicit Manifold Learning on Generative Adversarial Networks." arXiv preprint arXiv:1710.11260 (2017).
4. Fefferman, Charles, Sanjoy Mitter, and Hariharan Narayanan. "Testing the manifold hypothesis." Journal of the American Mathematical Society 29.4 (2016): 983-1049.

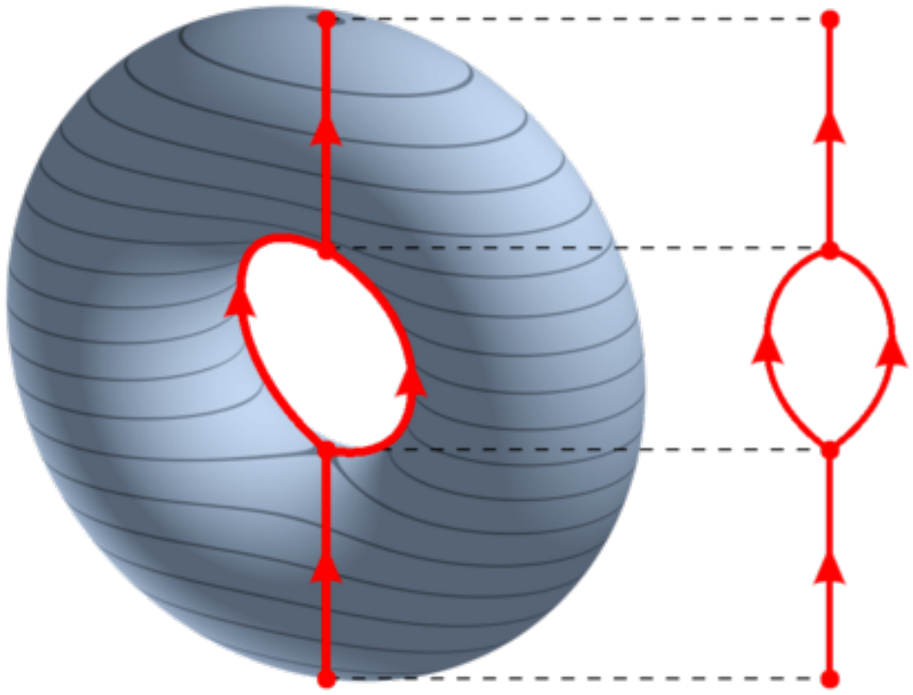
Images from: https://github.com/zalandoresearch/fashion-mnist, https://github.com/lmcinnes/umap

Background: Reeb Graphs

- Level set definition:

$$L_c(f) = \{(x_1, \cdots, x_n) \mid f(x_1, \cdots, x_n) = c\}$$

- A Reeb graph is a mathematical object reflecting the evolution of the level sets of a real-valued function on a manifold.
  - Points are part of the same ‘edge’ if they belong in the same connected component in $f^{-1}(c)$



- Reeb space == multivariate generalization of Reeb graph
- “…compresses the components of the level sets of a multivariate and obtains a summary representation of their relationships”

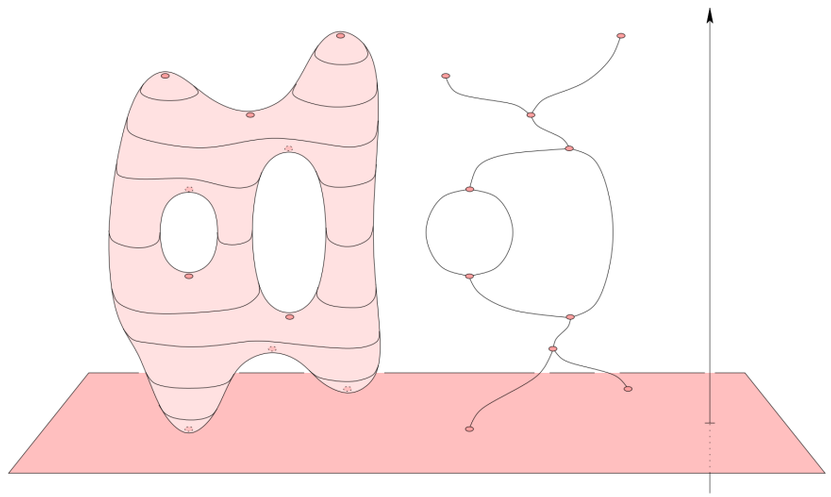
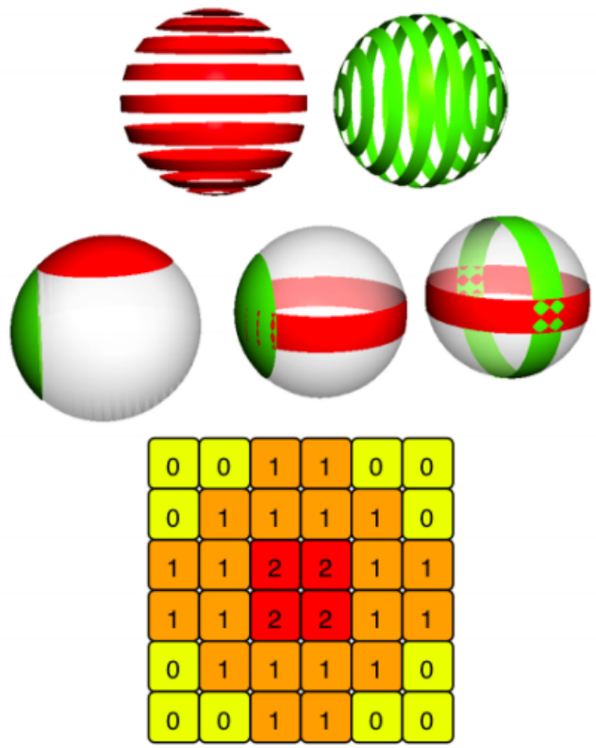


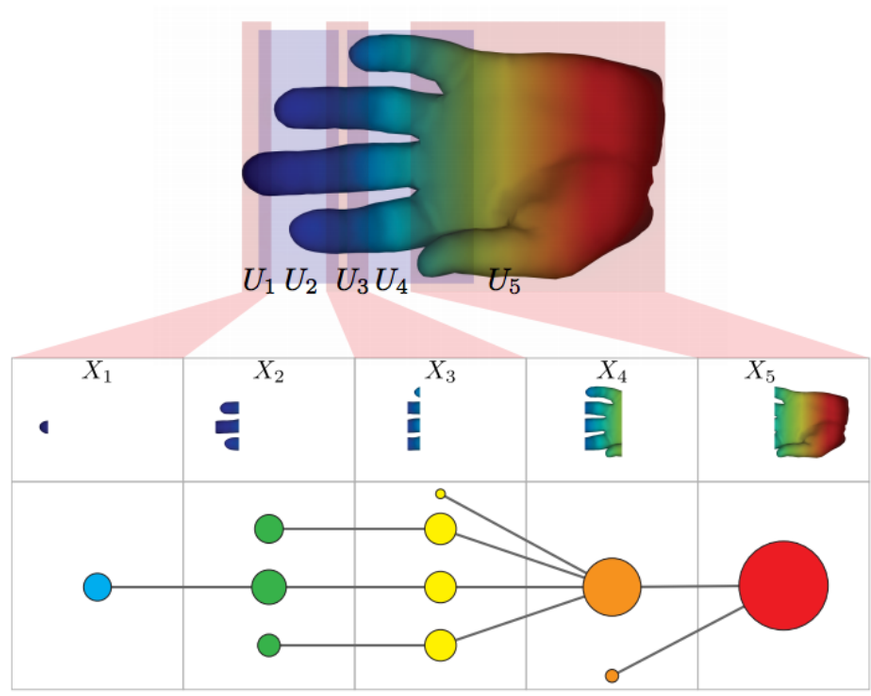
Figure V.13: Level sets of the 2-manifold map to points on the real line and components of the level sets map to points of the Reeb graph.

Top from: https://en.wikipedia.org/wiki/Reeb_graph
Bottom image from: https://www2.cs.duke.edu/courses/fall06/cps296.1/Lectures/sec-V-4.pdf

What does Mapper represent?

- What does mapper *actually* do?
  - "…our construction amounts to a stochastic version of the Reeb graph associated with the filter function. If the covering of R is too coarse, we will be constructing an image of the Reeb graph of the function, while if it is fine enough we will recover the Reeb graph precisely."
  - Munch et. al [1] proved that the categorical representations of the Reeb space and Mapper converge in terms of interleaving distance
- So we may *think* of Mapper as an approximation of the Reeb space





1. Munch, Elizabeth, and Bei Wang. "Convergence between categorical representations of Reeb space and Mapper." arXiv preprint arXiv:1512.04108 (2015).
Top From: Singh, Gurjeet, Facundo Mémoli, and Gunnar E. Carlsson. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." SPBG. 2007.
Bottom from: http://web.cse.ohio-state.edu/~wang.1016/courses/5559/Lecs/mapper-lec5559.pdf

Demo of World Values Survey

LiveSlides web content

To view

Download the add-in.
liveslides.com/download

Start the presentation.

Time Permitting

Persistent Homology

- Recurring theme in applied topological data analysis
- "Despite being both computable and insightful, the homology of a complex associated to a point cloud at a particular ϵ is **insufficient**: it is a mistake to ask which value of ϵ is optimal" [1]
- "The motivation is that, for a parameterized family of spaces (i.e. VR complexes) modeling a point-cloud data set, **qualitative features which persist over a large parameter range have greater statistical significance**"

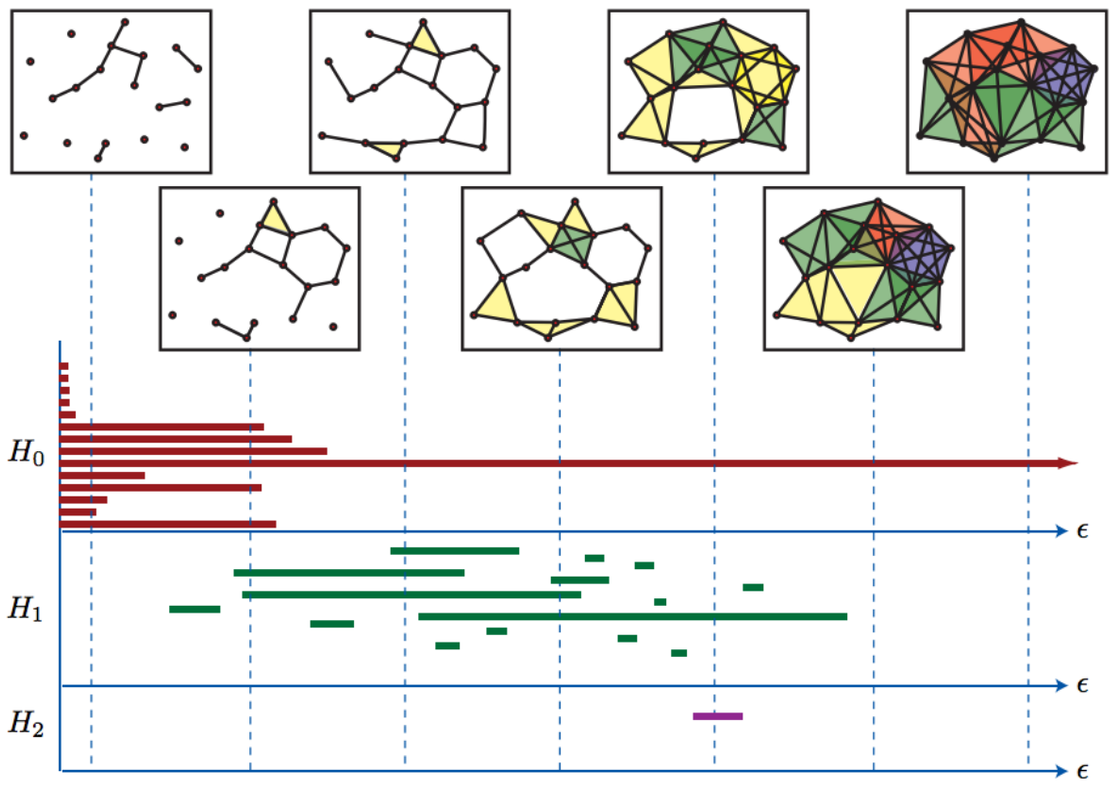
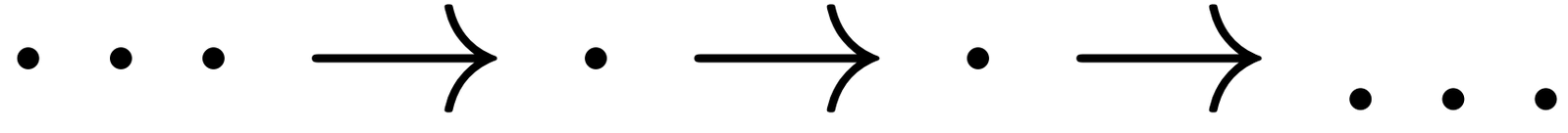


FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

Image from: Ghrist, Robert. "Barcodes: the persistent topology of data." Bulletin of the American Mathematical Society 45.1 (2008): 61-75.

Studying Mapper:
In the context of Persistent Homology

- Is it possible to study Mapper in the context of Persistent Homology?
- “The icon of persistence is a monotone sequence



where arrows connote maps of spaces or chains or the induced homomorphisms on homology." - Ghrist

- So, we just need some kind of monotone sequence between Mapper constructions?
- As it turns out, someone* already has!
  - Published analysis of so-called “MultiScale Mapper” [1]
  - *Happens to be one of the original authors of Mapper

1. Dey, Tamal K., Facundo Mémoli, and Yusu Wang. "Multiscale mapper: Topological summarization via codomain covers." Proceedings of the twenty-seventh annual acm-siam symposium on discrete algorithms. Society for Industrial and Applied Mathematics, 2016.

Multi-scale Mapper

- "The resulting view of the data [produced by Mapper] through a cover of the codomain offers flexibility in analyzing the data. However, *it offers only a view at a fixed scale at which the cover is constructed*."

- *Multiscale mapper:* a "tower" of simplicial complexes, which is a chain of simplicial complexes connected by [induced] simplicial maps

- Nice benefit: if the map is a real-valued PL function, the *exact* persistence diagram *from only the 1-skeleton* (!)

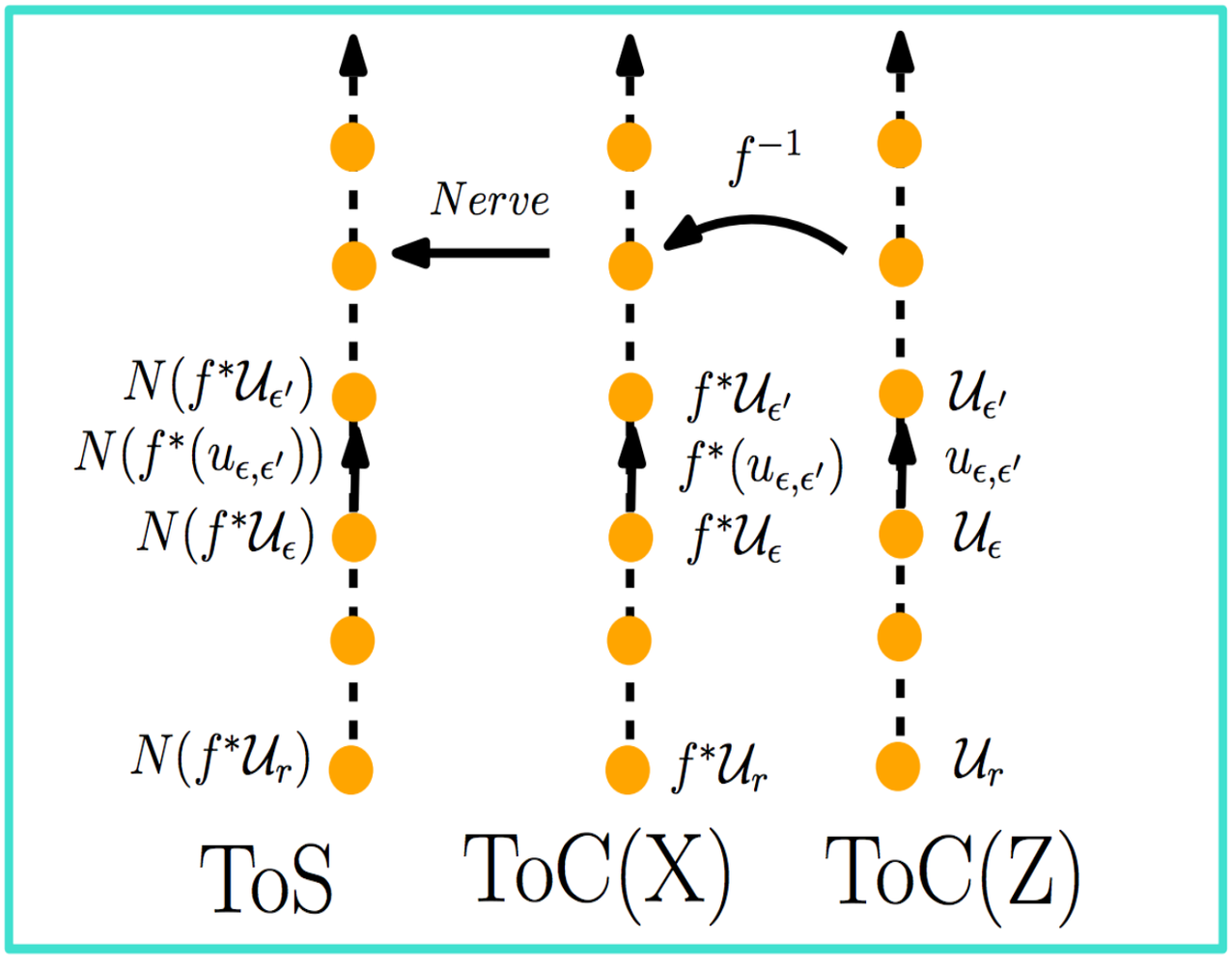


Image from: http://web.cse.ohio-state.edu/~wang.1016/courses/5559/Lecs/mapper-lec5559.pdf

Conclusion

- TDA is an emerging field
- Some see it as a more formal generalization of clustering
  - *How did I get here?* :)
- Mapper is becoming **remarkably popular**!
  - Ayasdi was awarded **$106 million** in funding towards their solution that uses Mapper for TDA [2]
  - Has proven *useful* in several data-analysis domains already!
- Mapper provides useful summary information, e.g. high-dimensional data sets
- ***Can Topology provide a solution to all unsupervised things in ML?***
  - “The reader should not conclude that the subject is quickly or painlessly learned. [The field of topology] is the impetus of future work: hard, slow, and ***fruitful***. “ - Robert Ghrist, Applied Elementary Topology
- Check out my extension to the open-source Mapper if you want to actually use for experiments: https://github.com/peekxc/TDAmapper

Questions?