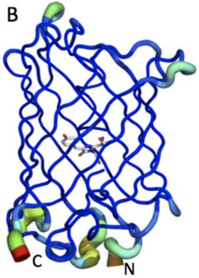
An Introduction to Topological Data Analysis

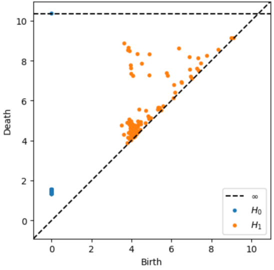
October 14, 2024

The Big Picture



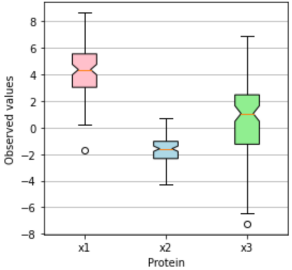
Raw Data

Point clouds
Networks
X-ray CT scans



Topological Summary

Mapper graphs
Persistence Diagrams
Euler Characteristic



Analysis

Statistics
Machine Learning
Prediction

Shape is Data, and Data has Shape

Main goal of TDA: Provide summaries of the shape of data that are

quantifiable

- Rigorously describe qualitative properties

comparable

- Establish a distance metric to compare any two summaries

robust

- A small change in the data set should result in a small change in the summary

concise

- Summaries should simplify the data

Shape is Data, and Data has Shape

Main goal of TDA: Provide summaries of the shape of data that are

quantifiable

- Rigorously describe qualitative properties

comparable

- Establish a distance metric to compare any two summaries

robust

- A small change in the data set should result in a small change in the summary

concise

- Summaries should simplify the data

Shape is Data, and Data has Shape

Main goal of TDA: Provide summaries of the shape of data that are

quantifiable

- Rigorously describe qualitative properties

comparable

- Establish a distance metric to compare any two summaries

robust

- A small change in the data set should result in a small change in the summary

concise

- Summaries should simplify the data

Shape is Data, and Data has Shape

Main goal of TDA: Provide summaries of the shape of data that are

quantifiable

- Rigorously describe qualitative properties

comparable

- Establish a distance metric to compare any two summaries

robust

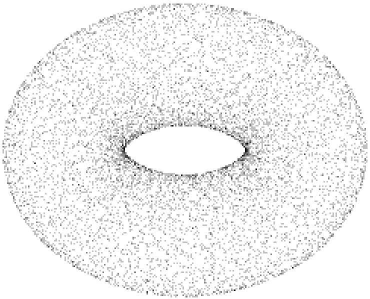
- A small change in the data set should result in a small change in the summary

concise

- Summaries should simplify the data

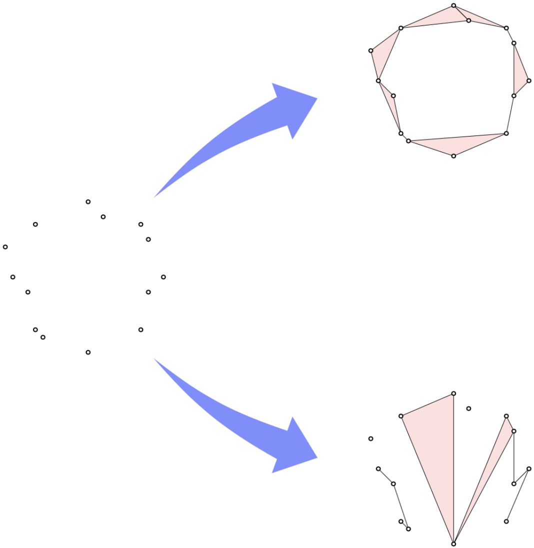
What is Topological Data Analysis?

Topological data analysis (TDA) uses techniques from topology to analyze the underlying structure of data.

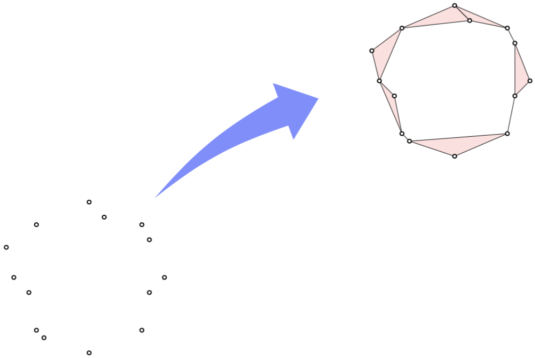


Persistent Homology

Simplicial Complex



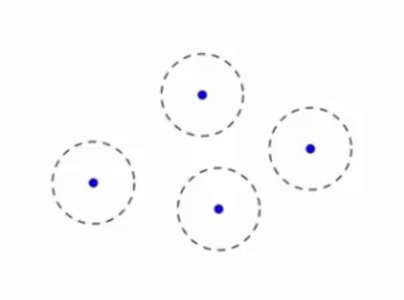
Simplicial Complex



Čech Complex

We need a way to construct a simplicial complex from point cloud data.

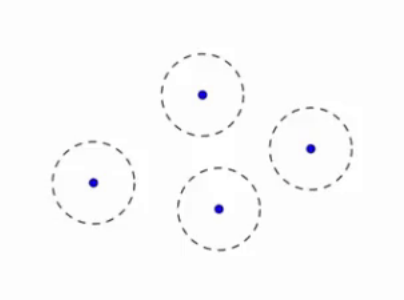
- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Čech complexes. At time r, the given Čech complex is denoted $C(r)$.



Čech Complex

We need a way to construct a simplicial complex from point cloud data.

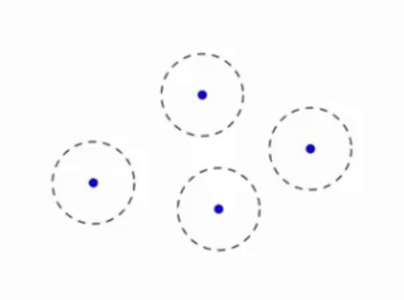
- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Čech complexes. At time r, the given Čech complex is denoted $C(r)$.



Čech Complex

We need a way to construct a simplicial complex from point cloud data.

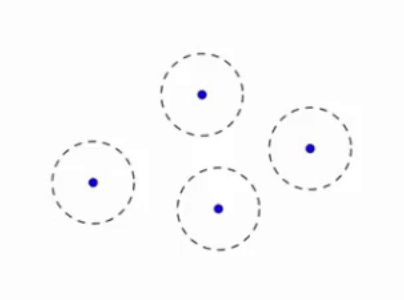
- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Čech complexes. At time r, the given Čech complex is denoted $C(r)$.



Čech Complex

We need a way to construct a simplicial complex from point cloud data.

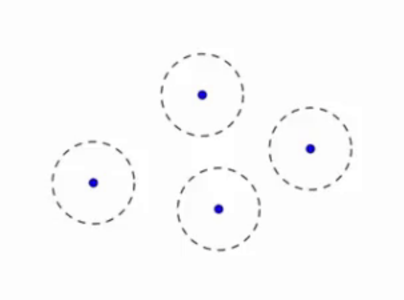
- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Čech complexes. At time r, the given Čech complex is denoted $C(r)$.



Čech Complex

We need a way to construct a simplicial complex from point cloud data.

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Čech complexes. At time r, the given Čech complex is denoted $C(r)$.



Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

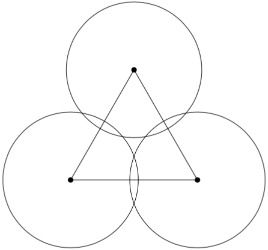
Vietoris-Rips Complex

There are many ways to create a simplicial complex from point cloud data

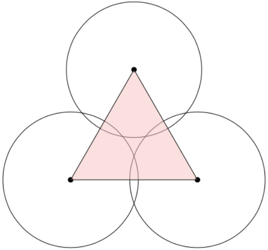
- Draw a small ball around each point.
- Expand the balls according to time r.
- When n balls pairwise intersect, draw an $n-1$-simplex between their vertices.
- As r increases, the simplicial complex changes.
- The resulting complexes are called Vietoris-Rips complexes.
- Vietoris-Rips complexes can be computed more efficiently than Čech complexes
- At time r, the given Vietoris-Rips complex is denoted $VR(r)$.

The Nerve Theorem

How do we know that the union of balls and the corresponding simplicial complex have the same topological structure?



Čech complex



Vietoris-Rips complex

The Čech complex captures the hole but the Vietoris-Rips complex does not!

The Nerve Theorem

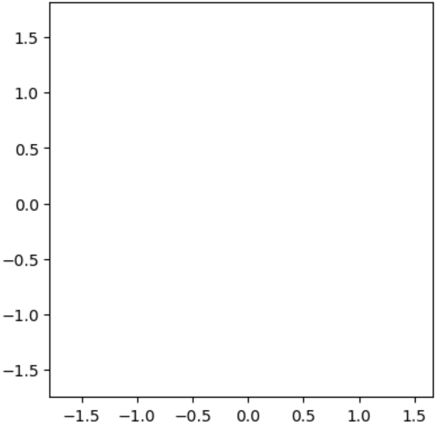
Let X be a set of point cloud data. Let $X(r)$ be the union of balls with radius r around the points of X, and let $C(r)$ be the corresponding Cech complex. Then $X(r)$ and $C(r)$ are homotopy equivalent.

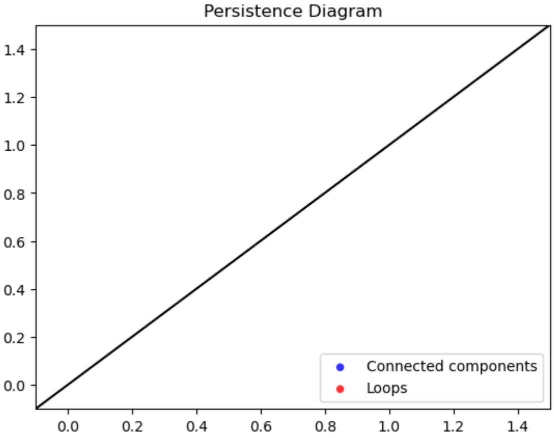
As we saw on the previous slide, the Nerve Theorem is not true for the Vietoris-Rips complex. However, we have the nice inclusion

$$C(r) \subset VR(r) \subset C(2r)$$

So if $C(r)$ and $C(2r)$ are good approximations of the structure of the point cloud data, then $VR(r)$ is as well.

Persistent Homology





Persistence Diagram

Persistent homology groups

Start with a filtration of simplicial complexes $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$. Let $f_i \colon K_i \to K_{i+1}$ be the inclusion map. From this filtration, we obtain a sequence of homomorphisms $f_p^{i,j} \colon H_p(K_i) \to H_p(K_j)$.

Definition

The p-th persistent homology groups are the images of the homomorphisms induced by the inclusion $H_p^{i,j} = \operatorname{im} f_p^{i,j}$ for $0 \leq i \leq j \leq n$. The corresponding p-th persistent Betti numbers are $\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}$.

Persistence Diagram

Persistent homology groups

Start with a filtration of simplicial complexes $\emptyset = K_0 \subseteq K_1 \subseteq \ldots \subseteq K_n = K$. Let $f_i \colon K_i \to K_{i+1}$ be the inclusion map. From this filtration, we obtain a sequence of homomorphisms $f_p^{i,j} \colon H_p(K_i) \to H_p(K_j)$.

Definition

The p-th persistent homology groups are the images of the homomorphisms induced by the inclusion $H_p^{i,j} = \operatorname{im} f_p^{i,j}$ for $0 \leq i \leq j \leq n$. The corresponding p-th persistent Betti numbers are $\beta_p^{i,j} = \operatorname{rank} H_p^{i,j}$.

Birth and Death

Birth and Death

We say that a homology class $\gamma \in H_p(K_i)$ is born at K_i if

$$\gamma \notin H_p^{i-1,i} = \text{im}\, f_p^{i-1,i}$$

If γ is born at K_i, then it dies entering K_j if it merges with an older class as we go from K_{j-1} to K_j.

The Elder Rule

If γ is born at K_i and dies entering K_j, then the index persistence of γ is $j - i$. If γ is born at K_i and never dies, then the persistence index is infinity.

Birth and Death

Birth and Death

We say that a homology class $\gamma \in H_p(K_i)$ is born at K_i if

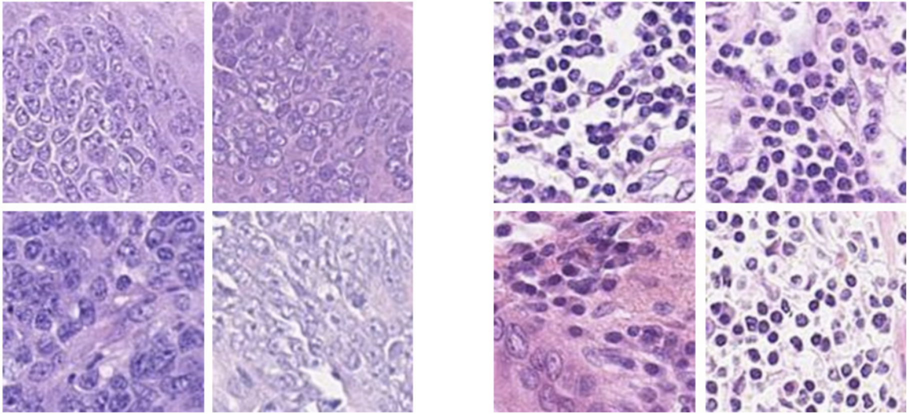
$$\gamma \notin H_p^{i-1,i} = \operatorname{im} f_p^{i-1,i}$$

If γ is born at K_i, then it dies entering K_j if it merges with an older class as we go from K_{j-1} to K_j.

The Elder Rule

If γ is born at K_i and dies entering K_j, then the index persistence of γ is $j - i$. If γ is born at K_i and never dies, then the persistence index is infinity.

Application: Cancer



Tumor (left) and non-tumor (right) patches of colorectal tissue.

Application: Cancer

- Take images of tissue and divide them into patches.
- These images are processed into a greyscale image.
- Each pixel has a intensity value between 0 and 255. Let $B(n)$ be the union of all pixels with intensity $\leq n$. To form a filtration, let $K_i = B(i)$ for $0 \leq i \leq 255$.
- Nuclei in tumor regions lie much closer to each other than in non-tumor regions, so the homology of tumor regions does not change much compared to the homology of non-tumor regions.

This method was able to identify tumor tissue that expert analysis missed!

Application: Cancer

- Take images of tissue and divide them into patches.
- These images are processed into a greyscale image.
- Each pixel has a intensity value between 0 and 255. Let $B(n)$ be the union of all pixels with intensity $\leq n$. To form a filtration, let $K_i = B(i)$ for $0 \leq i \leq 255$.
- Nuclei in tumor regions lie much closer to each other than in non-tumor regions, so the homology of tumor regions does not change much compared to the homology of non-tumor regions.

This method was able to identify tumor tissue that expert analysis missed!

Application: Cancer

- Take images of tissue and divide them into patches.
- These images are processed into a greyscale image.
- Each pixel has a intensity value between 0 and 255. Let $B(n)$ be the union of all pixels with intensity $\leq n$. To form a filtration, let $K_i = B(i)$ for $0 \leq i \leq 255$.
- Nuclei in tumor regions lie much closer to each other than in non-tumor regions, so the homology of tumor regions does not change much compared to the homology of non-tumor regions.

This method was able to identify tumor tissue that expert analysis missed!

Application: Cancer

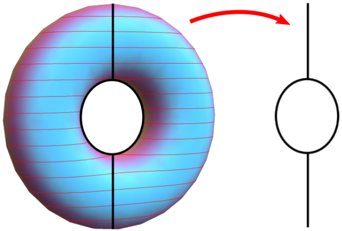
- Take images of tissue and divide them into patches.
- These images are processed into a greyscale image.
- Each pixel has a intensity value between 0 and 255. Let $B(n)$ be the union of all pixels with intensity $\leq n$. To form a filtration, let $K_i = B(i)$ for $0 \leq i \leq 255$.
- Nuclei in tumor regions lie much closer to each other than in non-tumor regions, so the homology of tumor regions does not change much compared to the homology of non-tumor regions.

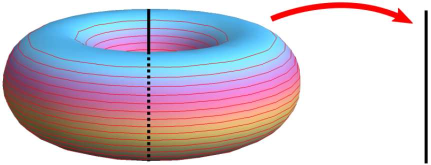
This method was able to identify tumor tissue that expert analysis missed!

Reeb spaces

The Reeb Graph

Let S be a topological space, and let $f\colon S \to \mathbb{R}$ be a continuous function. The *Reeb graph* of f, denoted $\mathrm{Reeb}(f)$, is the space $S/\sim$, where $s_1 \sim s_2$ if and only if $f(s_1) = f(s_2)$ and s_1 and s_2 are in the same connected component of $f^{-1}(f(s_1)) = f^{-1}(f(s_2))$.

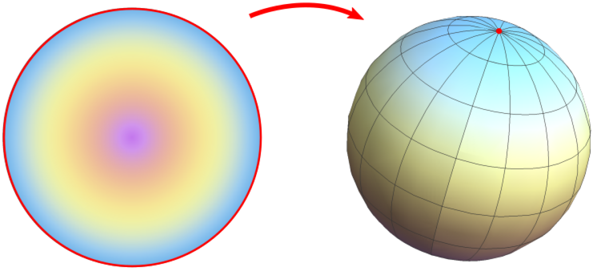




The Reeb graph depends both on the space X and the function f.

The Reeb space of a function

Let S and X be semi-algebraic sets, and let $f\colon S \to X$ be a semi-algebraic map continuous. The *Reeb space* of f, denoted $\mathrm{Reeb}(f)$, is the space $S/\sim$, where $s_1 \sim s_2$ if $f(s_1) = f(s_2)$ and s_1 and s_2 are in the same connected component of $f^{-1}(f(s_1)) = f^{-1}(f(s_2))$.



Letting $f\colon \mathbf{D}^2 \to \mathbf{S}^2$ be the map shown above, the resulting Reeb space is S^2.

Topological complexity

Dey et al. (2017)

If $f : X \to Y$ is a proper map and X is connected, then $\beta_1(\mathrm{Reeb}(f)) \leq \beta_1(X)$.

The previous example shows that this theorem does not generalize to $\beta(\mathrm{Reeb}(f))$. However, $\beta(\mathrm{Reeb}(f))$ can be bounded in terms of the complexity of the map f.

Theorem (Basu et al. (2018))

Let $S \subset \mathrm{R}^n$ be a bounded $\mathcal{P}$-closed semi-algebraic set, and $f = (f_1, \ldots, f_m) : S \to \mathrm{R}^m$ be a polynomial map. Suppose that $s = \mathrm{card}(\mathcal{P})$ and the maximum of the degrees of the polynomials in $\mathcal{P}$ and $f_1, \ldots, f_m$ is bounded by d. Then,

$$\beta(\mathrm{Reeb}(f)) \leq (sd)^{(n+m)^{O(1)}}.$$

Topological complexity

Dey et al. (2017)

If $f\colon X \to Y$ is a proper map and X is connected, then $\beta_1(\mathrm{Reeb}(f)) \leq \beta_1(X)$.

The previous example shows that this theorem does not generalize to $\beta(\mathrm{Reeb}(f))$. However, $\beta(\mathrm{Reeb}(f))$ can be bounded in terms of the complexity of the map f.

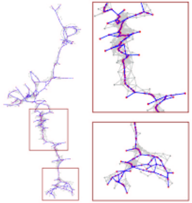
Theorem (Basu et al. (2018))

Let $S \subset \mathrm{R}^n$ be a bounded $\mathcal{P}$-closed semi-algebraic set, and $f = (f_1, \ldots, f_m) : S \to \mathrm{R}^m$ be a polynomial map. Suppose that $s = \mathrm{card}(\mathcal{P})$ and the maximum of the degrees of the polynomials in $\mathcal{P}$ and $f_1, \ldots, f_m$ is bounded by d. Then,

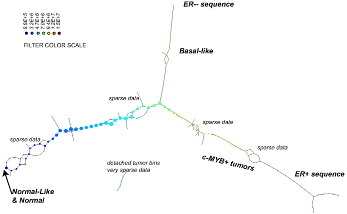
$$\beta(\mathrm{Reeb}(f)) \leq (sd)^{(n+m)^{O(1)}}.$$

Applications of Reeb Graphs

- Ge et al. (2012) use the Reeb graph to obtain a skeleton graph of a metric graph.

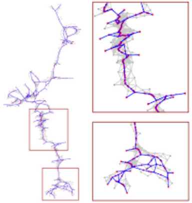


- Nicolau et al. (2011) used Mapper, a discrete approximation of the Reeb graph, to analyze breast cancer tumor expression.

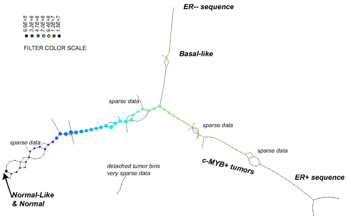


Applications of Reeb Graphs

- Ge et al. (2012) use the Reeb graph to obtain a skeleton graph of a metric graph.



- Nicolau et al. (2011) used Mapper, a discrete approximation of the Reeb graph, to analyze breast cancer tumor expression.



Mapper

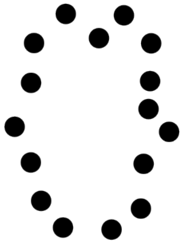
The Mapper Algorithm Captures the Shape of Data

The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.

Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

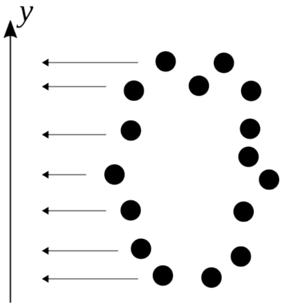
The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

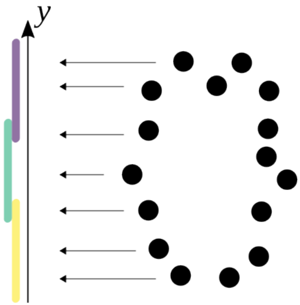
The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

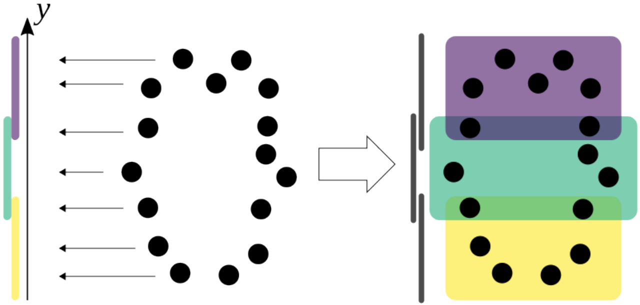
The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

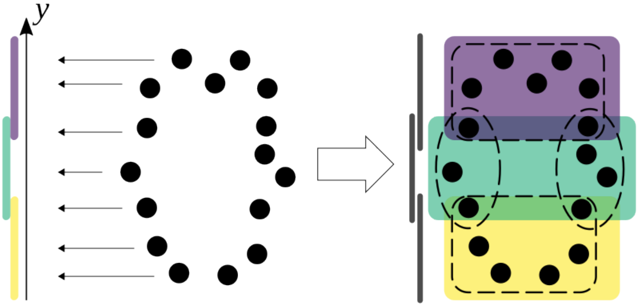
The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

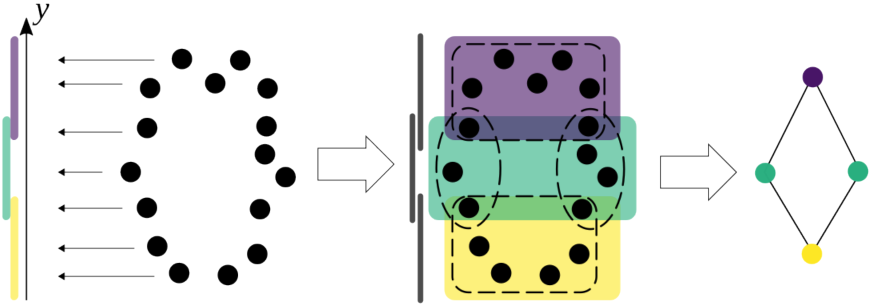
The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

The Mapper Algorithm Captures the Shape of Data

The Mapper algorithm produces a one-dimensional abstract graph that reflects the underlying structure of the input data.



Singh et al. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. *PBG@Eurographics*.

Basu, S., Cox, N., and Percival, S. (2018). On the Reeb spaces of definable maps. *arXiv e-prints*, page arXiv:1804.00605.

Dey, T. K., Memoli, F., and Wang, Y. (2017). Topological Analysis of Nerves, Reeb Spaces, Mappers, and Multiscale Mappers. *ArXiv e-prints*.

Ge, X., Safa, I., Belkin, M., and Wang, Y. (2012). Data skeletonization via reeb graphs. *Adv. Neural Inform. Process. Syst.*, 24.

Nicolau, M., Levine, A. J., and 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*, 108(17):7265–7270.

Singh, G., Memoli, F., and Carlsson, G. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In Botsch, M., Pajarola, R., Chen, B., and Zwicker, M., editors, *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association.