# What is the shape of data?

## A brief introduction to topological data analysis

Mauricio Montes

Auburn University

October 10, 2023

# What is TDA?

TDA stands for Topological Data Analysis.

# What is TDA?

TDA stands for Topological Data Analysis.

- Topology is the study of geometric properties preserved under continuous deformations.

# What is TDA?

TDA stands for Topological Data Analysis.

- Topology is the study of geometric properties preserved under continuous deformations.
- This means that we can stretch, bend, and deform objects, but we cannot tear or glue them.

# What is TDA?

TDA stands for Topological Data Analysis.

- Topology is the study of geometric properties preserved under continuous deformations.
- This means that we can stretch, bend, and deform objects, but we cannot tear or glue them.
- Data analysis is the process of studying and modeling data to extract useful information.

# What is TDA?

TDA stands for Topological Data Analysis.

- Topology is the study of geometric properties preserved under continuous deformations.
- This means that we can stretch, bend, and deform objects, but we cannot tear or glue them.
- Data analysis is the process of studying and modeling data to extract useful information.
- Typically, we use statistics and machine learning to analyze data. We can improve these methods by using tools from topology.

# Developing our intuition

# Developing our intuition

- One of the core ideas of TDA is that our data is sampled from a manifold.

# Developing our intuition

- One of the core ideas of TDA is that our data is sampled from a manifold.
- We can think of manifolds as "things that locally look like" $\mathbb{R}^n$. Most objects in the real world are manifolds.

# Developing our intuition

- One of the core ideas of TDA is that our data is sampled from a manifold.
- We can think of manifolds as "things that locally look like" $\mathbb{R}^n$. Most objects in the real world are manifolds.
- If we take enough samples, we can approximate the shape of the manifold our data is living in. This allows us to answer topological questions about our data.

# The TDA Pipeline

- The input is a finite set of points with some notion of similarity between them. This could be distance, correlation, or some other metric.

# The TDA Pipeline

- The input is a finite set of points with some notion of similarity between them. This could be distance, correlation, or some other metric.
- We construct a simplicial complex from the data. This is a combinatorial object that encodes the shape of the data.

# The TDA Pipeline

- The input is a finite set of points with some notion of similarity between them. This could be distance, correlation, or some other metric.
- We construct a simplicial complex from the data. This is a combinatorial object that encodes the shape of the data.
- We compute topological invariants of the simplicial complex. This would be something like the Euler characteristic, homology, or persistent homology of the complex.

# The TDA Pipeline

- The input is a finite set of points with some notion of similarity between them. This could be distance, correlation, or some other metric.
- We construct a simplicial complex from the data. This is a combinatorial object that encodes the shape of the data.
- We compute topological invariants of the simplicial complex. This would be something like the Euler characteristic, homology, or persistent homology of the complex.
- This new topological data provides us with a new descriptor of our original dataset. This can then be used in conjunction with other methods to improve our analysis.

# Simplicial Complexes

- The problem with data is that it is discrete, but manifolds are continuous.

## Simplicial Complexes

- The problem with data is that it is discrete, but manifolds are continuous.
- We can remedy this by building a simplicial complex from our data to approximate the manifold.

# Simplicial Complexes

- The problem with data is that it is discrete, but manifolds are continuous.
- We can remedy this by building a simplicial complex from our data to approximate the manifold.
- A $k$-simplex is the convex hull of $k+1$ linearly independent points in $\mathbb{R}^n$.

# Simplicial Complexes

- The problem with data is that it is discrete, but manifolds are continuous.
- We can remedy this by building a simplicial complex from our data to approximate the manifold.
- A $k$-simplex is the convex hull of $k+1$ linearly independent points in $\mathbb{R}^n$.
- A simplicial complex $K$ is a collection of simplices that satisfies two conditions:

1. If $\sigma \in K$, then every face of $\sigma$ is also in $K$.
2. If $\sigma_1, \sigma_2 \in K$, then $\sigma_1 \cap \sigma_2$ is either empty or a face of both $\sigma_1$ and $\sigma_2$.

# Topological Preliminaries

- A topological space is a pair $(X, \tau)$ consisting of a set $X$ with a collection of subsets $\tau$, called the topology on $X$, such that:
  1. $\emptyset, X \in \tau$
  2. If $U_i \in \tau$ for $i \in I$, then $\bigcup_{i \in I} U_i \in \tau$
  3. If $U_1, U_2 \in \tau$, then $U_1 \cap U_2 \in \tau$
- The Euclidean topology on $\mathbb{R}^n$ is the collection of all open subsets of $\mathbb{R}^n$. We define an open set as a set $B_r(x) = \{y \in \mathbb{R}^n : d(x, y) < r\}$. Where $d$ is the Euclidean metric.

# Topological Preliminaries

- We say that a topological space is Hausdorff if for any two points $x, y \in X$ there exist disjoint open sets $U, V \in \tau$ such that $x \in U$ and $y \in V$.

# Topological Preliminaries

- We say that a topological space is Hausdorff if for any two points $x, y \in X$ there exist disjoint open sets $U, V \in \tau$ such that $x \in U$ and $y \in V$.
- A basis for a topological space $(X, \tau)$ is a collection $\mathcal{B}$ of open sets such that every open set in $\tau$ can be written as a union of sets in $\mathcal{B}$.
- A topological space is second countable if it has a countable basis.

# Topological Preliminaries

- We say that a topological space is Hausdorff if for any two points $x, y \in X$ there exist disjoint open sets $U, V \in \tau$ such that $x \in U$ and $y \in V$.
- A basis for a topological space $(X, \tau)$ is a collection $\mathcal{B}$ of open sets such that every open set in $\tau$ can be written as a union of sets in $\mathcal{B}$.
- A topological space is second countable if it has a countable basis.
- In the Euclidean topology, the topology is actually generated by the collection of open balls. These are a basis for the Euclidean topology.

# Topological Preliminaries

- The fundamental map between topological spaces is the Homeomorphism.

# Topological Preliminaries

- The fundamental map between topological spaces is the Homeomorphism.
- A homeomorphism is a continuous map $f : X \to Y$ such that there exists a continuous map $g : Y \to X$ such that $g \circ f = id_X$ and $f \circ g = id_Y$. (Bicontinuous, Bijective)

# Topological Preliminaries

- The fundamental map between topological spaces is the Homeomorphism.
- A homeomorphism is a continuous map $f : X \to Y$ such that there exists a continuous map $g : Y \to X$ such that $g \circ f = id_X$ and $f \circ g = id_Y$. (Bicontinuous, Bijective)
- If there exists a homeomorphism between two topological spaces, we say that they are homeomorphic.

# Topological Preliminaries

- We can finally define a manifold.

# Topological Preliminaries

- We can finally define a manifold.
- A topological manifold is a second countable Hausdorff space that is locally homeomorphic to $\mathbb{R}^n$.

# Topological Preliminaries

- We can finally define a manifold.
- A topological manifold is a second countable Hausdorff space that is locally homeomorphic to $\mathbb{R}^n$.
- The last condition means that for every point $x \in X$, there exists an open set $U \in \tau$ such that $x \in U$ and $U$ is homeomorphic to an open set $\hat{U} \subseteq \mathbb{R}^n$.
- To make this a smooth manifold, we require that the homeomorphism is actually a diffeomorphism. (Smooth map with smooth inverse)