

# Topological Data Analysis

With a view towards machine learning

Priyangshu Ghosh

at IISER BHOPAL under Dr. Kuntal Roy

*priyangshu@cmi.ac.in*

July 24, 2024

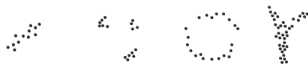
# Outline

- 1 Introduction
  - Simplicial complex
  - Homology groups
- 2 Some background in topology
- 3 Persistent Homology

# Introduction

# Some Motivation

Data sets comes in various shapes and sizes, lets look at some of them below:



- The first data set has a rough shape of a strip or a line. These kinds of data are usually handled using regression models.
- The second set has clumps of points spread on the plain. We use cluster analysis for to deal with such datasets.
- The third set is a type of data set that occurs frequently when one is dealing with time series data representing periodic or recurrent behavior of some kind.
- The fourth data might describe data in which there are one standard or normal mode and three extremal modes.

The latter two do not have dedicated methods for analyzing them, and its not a complete list of data sets which don't have their dedicated methods. And hoping to find a method for every type of data set isn't a great idea.

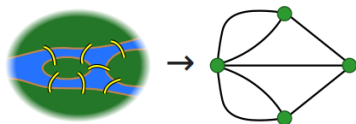
Topological Data Analysis (TDA) is a recent field that emerged from various works in applied (algebraic) topology and computational geometry during the first decade of the century. TDA refers broadly to a set of analytic, computational and statistical methods stemming from the study of the shape and connectivity of data.

Although one can trace back geometric approaches for data analysis quite far in the past, TDA really started as a field with the pioneering works of Edelsbrunner et al. (2002) and Zomorodian and Carlsson (2005) in persistent homology and was popularized in a landmark paper in 2009 Carlsson (2009).

# Some background in Topology

## History fact

Topology originated in 1736, by Euler as he published a paper, solving the Seven Bridges of Königsberg problem. And in 1750, Euler wrote to a friend that he had realized the importance of the edges of a polyhedron. This led to his polyhedron formula:  $V - E + F = 2$



Our primary focus in this introduction, is to understand homology. Homology is a technique of studying topological spaces using algebraic objects. Our primary focus will be on homology groups, which are topological invariants(properties that are invariant under homomorphisms)

# Simplicial complex

## Definition 1.1: Simplex

The  $n$ -simplex,  $\Delta^n$ , is the simplest geometric figure determined by a collection of  $n + 1$  points in Euclidean space  $\mathbb{R}^n$ . Geometrically, it can be thought of as the complete graph on  $n + 1$  vertices, which is solid in  $n$  dimensions.

$$\Delta^n := \{(x_0, \dots, x_n) \in \mathbb{R}^{n+1} \mid \sum x_i = 1, x_i \geq 0 \forall i\}$$



## Definition 1.2: Face

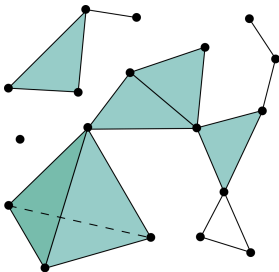
An  $n$ -face of a simplex is a subset of the set of vertices of the simplex with order  $n + 1$ . The faces of an  $n$ -simplex with dimension less than  $n$  are called its proper faces.



## Definition 1.3: Simplicial complex

A simplicial complex  $K$  is a finite set of simplices satisfying the following conditions:

- 1 For all simplices  $A \in K$  with  $\alpha$  a face of  $A$ , we have  $\alpha \in K$ .
- 2 If  $A, B \in K$ , then  $A$  and  $B$  are properly situated.

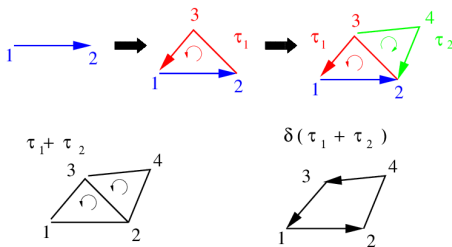


# Homology Groups

## Definition 2.1: Chains

Given a set  $A_1^n, \dots, A_k^n$  of arbitrarily oriented  $n$ -simplices of a complex  $K$  and an abelian group  $G$ , we define an  $n$ -chain  $x$  with coefficients in  $G$  as a formal sum:  $x = g_1 A_1^n + g_2 A_2^n + \dots + g_k A_k^n$ , where  $g_i \in G$ .

Henceforth, we will assume that  $G = \mathbb{Z}$ . The set of  $n$ -chains forms an abelian group over addition: for  $x = \sum_{i=1}^k g_i A_i^n$  and  $y = \sum_{i=1}^k h_i A_i^n$ , we have  $x + y = \sum_{i=1}^k (g_i + h_i) A_i^n$ . We denote the group of  $n$ -chains by  $L_n$ .



## Definition 2.2: Boundary

Let  $A^n$  be an oriented  $n$ -simplex in a complex  $K$ . The boundary of  $A^n$  is defined as the  $(n-1)$ -chain of  $K$  over  $\mathbb{Z}$  given by

$$\delta(A^n) = A_0^{n-1} + A_1^{n-1} + \cdots + A_n^{n-1},$$

where  $A_i^{n-1}$  is an  $(n-1)$ -face of  $A^n$ . If  $n = 0$ , we define  $\delta(\Delta^0) = 0$ .

It is important to note that, since  $A^n$  was oriented, the  $A_i^{n-1}$  have associated orientations as well. We can extend the definition of boundary linearly to all of  $L_n$ : for an  $n$ -chain  $x = \sum_{i=1}^k g_i A_i^n$ , define

$$\delta(x) = \sum_{i=1}^k g_i \delta_n(A_i^n),$$

where  $A_i^n$  are the  $n$ -simplices of  $K$ . Therefore, the boundary operator  $\delta$  is a homomorphism  $\delta : L_n \rightarrow L_{n-1}$ .

## Definition 2.3: Cycle

We call an  $n$ -chain a cycle if its boundary is zero, and denote the set of  $n$ -cycles of  $K$  over  $\mathbb{Z}$  by  $Z_n$ .  $Z_n$  is a subgroup of  $L_n$ , and can also be written as  $Z_n = \ker(\delta)$ .

## Definition 2.4

We say that an  $n$ -cycle  $x$  of a  $k$ -complex  $K$  is homologous to zero if it is the boundary of an  $(n + 1)$ -chain of  $K$ ,  $n = 0, 1, \dots, k - 1$ . A boundary is then any cycle that is homologous to zero. This relation is written  $x \sim 0$ , and the subgroup of  $Z_n$  of boundaries is denoted  $B_n$ . We also write  $B_n = \text{Im}(\delta)$ .

Less formally, a cycle is a member of  $B_n$  if it "bounds" something contained in the complex  $K$ . For example, the chain  $b + c + e$  in Figure 3 is a boundary, but  $a + d + e$  is not. The relation  $x \sim 0$  gives an equivalence relation: for two chains  $x, y$ ,  $(x - y) \sim 0 \Rightarrow x \sim y$ , and we call  $x$  and  $y$  homologous.

Since  $B_n$  is a subgroup of  $Z_n$ , we may form the quotient group  $H_n = Z_n/B_n$ .

## Definition 2.5: Simplicial Homology group

The group  $H_n$  is the  $n$ -dimensional homology group of the complex  $K$  over  $\mathbb{Z}$ .  $H_n$  can also be written as  $\ker(\delta)/\text{Im}(\delta)$ .

Particularly in the lower dimensions, we have an intuitive idea of when two topological spaces are fundamentally “the same”. We have some ways of generalizing and making rigorous this intuition, including the idea of homeomorphism. It would be nice to have some sort of relation between the homology groups of homeomorphic spaces, and, in fact, it turns out that if two topological spaces are homeomorphic, they have isomorphic homology groups.

### Definition 2.6

Given a topological space  $X$ , a singular  $n$ -simplex in  $X$  is a map  $\sigma : \Delta^n \rightarrow X$ , such that  $\sigma$  is continuous.

### Definition 2.7

Let  $C_n(X)$  be the free abelian group with basis the set of singular  $n$ -simplices of  $X$ . Elements of  $C_n(X)$  are called singular  $n$ -chains and are finite formal sums:

$$\sum_i g_i \sigma_i, \text{ where } g_i \in \mathbb{Z}.$$

## Definition 2.8

The boundary map  $\delta_n : C_n(X) \rightarrow C_{n-1}(X)$  is given by:

$$\delta_n(\sigma) = \sum_i (-1)^i \sigma[[v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_n],$$

where  $v_i$  are the 0-simplices of  $\sigma$ , that is, the maps of the vertices of  $\Delta^n : v_i : \Delta^0 \rightarrow X$ .

## Chain Complex

Given topological space  $X$ , we defined the chain complex  $C(X)$  to encode information about  $X$ .  $C(X)$  is the sequence of free abelian groups  $C_0(X), C_1(X), \dots$ , paired with homomorphisms:  $\{\delta_n\}_{n \geq 0}$

$$\dots \xrightarrow{\delta_{n+1}} C_n \xrightarrow{\delta_n} C_{n-1} \xrightarrow{\delta_{n-1}} \dots \xrightarrow{\delta_1} C_0 \xrightarrow{\delta_0} 0$$

## Singular homology group

$$H_n(X) := \ker(\delta_n) / \text{im}(\delta_{n+1})$$

## Informal definition and Intuition

Given topological space  $X$ . The  $k^{\text{th}}$  homology group  $H_k(X)$  describes the number of  $k$  dimensional holes in  $X$ . Intuitively one way to think of a  $k$  dimensional hole is, what dimension of space does one need to fill up the gaps in  $X$ . Soon we will define betti numbers for this.

**Example 1:  $n$  dimensional sphere  $S^n$**

$$H_k(S^n) = \begin{cases} \mathbb{Z} & k = 0, n \\ \{0\} & \text{otherwise} \end{cases}$$







**Example 2: Torus ( $T^2 := (S^1 \times S^1)$ )**

$$H_k(T^2) = \begin{cases} \mathbb{Z} & k = 0, 2 \\ \mathbb{Z} \times \mathbb{Z} & k = 1 \\ \{0\} & \text{otherwise} \end{cases}$$

## Betti numbers

$$\beta_k = \text{rank}(H_k(X))$$

As we discussed earlier  $\beta_k$  denotes number of  $k$  dimensional holes.

$\beta_0=1$ 	$\beta_0=1, \beta_1=1$ 	$\beta_0=1, \beta_1=1$ 	$\beta_0=1, \beta_1=2$ 	$\beta_0=1, \beta_0=1, \beta_2=1$ 	$\beta_0=1, \beta_1=0, \beta_2=1$ 
---	---	---	---	---	--



# Persistent Homology

# Persistence

The concept of persistence is motivated by the practical need to cope with noise in data. This includes defining, recognizing, and possibly eliminating noise. Because of the loose definitions of noise and features, we will focus on a range of scales and try to attain a point of view.

Persistent homology is a measure of the structure of a filtered simplicial complex. It helps us address the major weakness of homology, the instability under small changes.

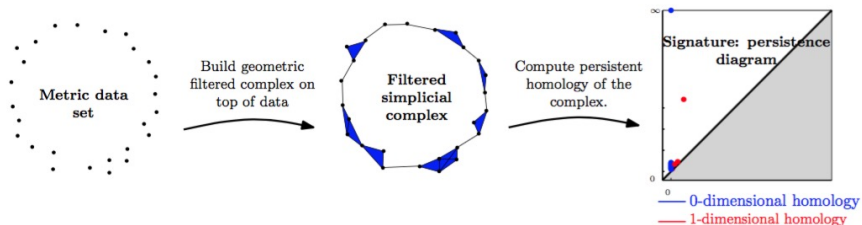
## Filtration

Consider a simplicial complex,  $K$ , and a function  $\varphi : K \rightarrow \mathbb{R}$ . We require that  $\varphi$  be monotonic, by which we mean it is non-decreasing along increasing chains of faces, that is,  $\varphi(\sigma) \leq \varphi(\tau)$  whenever  $\sigma$  is a face of  $\tau$ . Monotonicity implies that the sublevel set,  $K(a) = \varphi^{-1}(-\infty, a]$ , is a subcomplex of  $K$  for every  $a \in \mathbb{R}$ . Letting  $m$  be the number of simplices in  $K$ , we get  $n + 1 \leq m + 1$  different subcomplexes, which we arrange as an increasing sequence:

$$\emptyset = K_0 \subseteq K_1 \subseteq \dots \subseteq K_n = K.$$

This sequence is called a filtered simplicial complex.

# Computing persistence



# Čech and VR Complex

Čech Complex and Vietoris-Rips complex are abstract simplicial complexes that are defined on metric spaces with a distance parameter, over a point cloud.

Let  $S$  be a finite set of points in  $\mathbb{R}^n$ , let  $r \geq 0$  be a real number, then

$$\check{C}ech(S, r) := \{\sigma \subseteq S \mid \bigcap_{x \in \sigma} B_x(r) \neq \emptyset\}$$

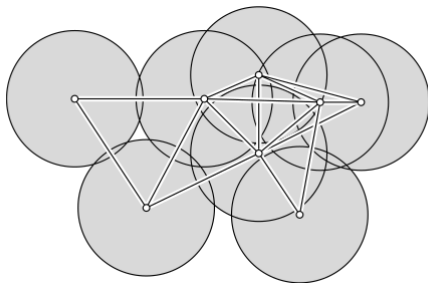
To construct the Čech complex, we need to test whether a collection of disks has a non-empty intersection, which can be difficult or, in some metric spaces, impossible. Hence we introduce Rips complex.

The Rips complex of  $S$  and  $r$ , consists of all abstract simplices in  $2^S$ . whose vertices are at most a distance  $2r$  from one another. In other words, we connect any two vertices at distance at most  $2r$  from each other by an edge, and we add a triangle or higher-dimensional simplex to the complex if all its edges are in the complex.

$$VR(S, r) = \{\sigma \subseteq S \mid \text{diam}(\sigma) \leq 2r\}$$

## Lemma

Let  $S$  be a finite set of points in some Euclidean space, and  $r \geq 0$ , then we have  $\check{\text{Cech}}(S, r) \subseteq \text{VR}(S, r) \subseteq \check{\text{Cech}}(S, \sqrt{2}r)$



**Figure:** Nine points with pairwise intersections among the disks indicated by straight edges connecting their centers. The Čech complex fills all nine of the possible ten triangles. But the VR complex fills all of them.

# Persistence homology group

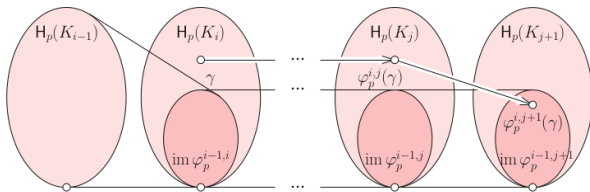
For a filtered complex  $K_0, K_1, \dots, K_n$  we define For  $0 \leq i \leq j \leq n$ , the inclusion  $K_i \hookrightarrow K_j$  induces a homomorphism  $\varphi_{i,j}^p : H_p(K_i) \rightarrow H_p(K_j)$ . The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = H_p(K_0) \rightarrow H_p(K_1) \rightarrow \dots \rightarrow H_p(K_n) = H_p(K),$$

## Definition

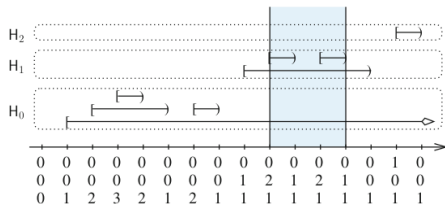
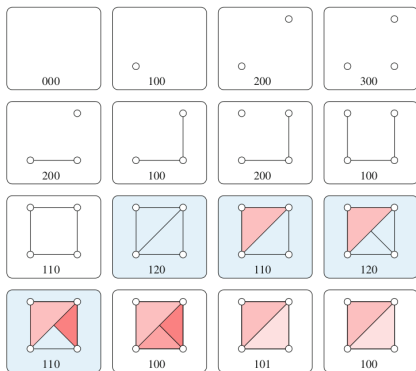
The  $p$ -th persistent homology groups are the images of the homomorphisms induced by inclusion,  $H_{i,j}^p = \text{im } \varphi_{i,j}^p$ , for  $0 \leq i \leq j \leq n$ . The corresponding  $p$ -th persistent Betti numbers are the ranks of these groups,  $\beta_{i,j}^p = \text{rank } H_{i,j}^p$ . Similarly, we define reduced persistent homology groups and reduced persistent Betti numbers. Note that  $H_{i,i}^p = H_p(K_i)$ . The persistent homology groups consist of the homology classes of  $K_i$  that are still alive at  $K_j$  or, more formally,  $H_{i,j}^p = Z_p(K_i) / (B_p(K_j) \cap Z_p(K_i))$ .

We can be more concrete about the classes counted by the persistent homology groups. Letting  $\gamma$  be a class in  $H_p(K_i)$ , we say it is born at  $K_i$  if  $\gamma \notin \text{im } \varphi_{i-1,i}^p$ . Furthermore, if  $\gamma$  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  $K_j$ , that is,  $\varphi_p^{i,j-1}(\alpha) \notin \text{im } \varphi_p^{i-1,j-1}$  but  $\varphi_p^{i,j}(\alpha) \in \text{im } \varphi_p^{i-1,j}$ .



The index persistence of  $\gamma$  is  $j - i + 1$ . In most applications, we have a function that governs the construction of the filtration, and we call the difference between the function values at the birth and the death the persistence of the class.

# Barcodes

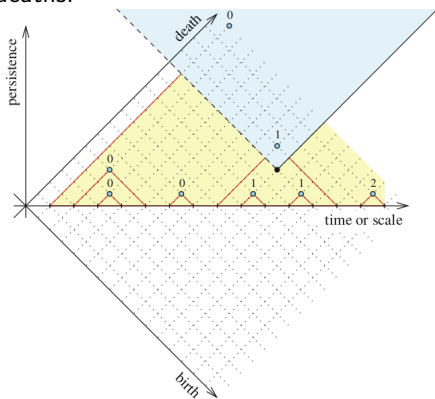


Left: Construction of a tetrahedron, this represents a filtered simplicial complex .  
 The 3 digits represent  $\beta_0, \beta_1, \beta_2$  .  
 Right: The barcode of the filtration.



# Persistence diagrams

The number of bars tends to get large with more complicated shapes and filtrations. Persistence diagrams are convenient ways of representing birth and deaths.



We can read off the betti numbers of the complexes in the filtration using this diagram. The rank of the image of  $H_p(K_i)$  in  $H_p(K_j)$  is the number of classes that are born at or before  $K_i$  and that die entering  $K_{j+1}$  or later.

One of the reasons we care about Persistence is due its stability under perturbations of data. Small changes in the data imply at most small changes in the measured persistence.

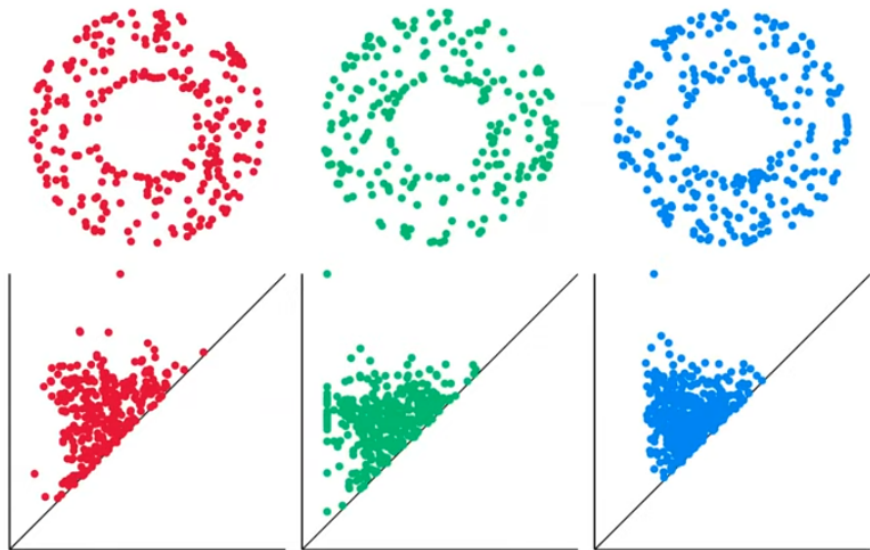
## Bottleneck distance

If  $X$  and  $Y$  are two persistent diagrams, then we define the bottleneck distance

$$W_{\infty}(X, Y) = \inf_{f: X \rightarrow Y} \sup_{x \in X} \|x - f(x)\|_{\infty}$$

where  $f$  is a bijection

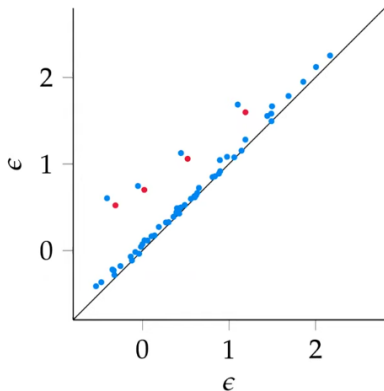
# Bottleneck stability



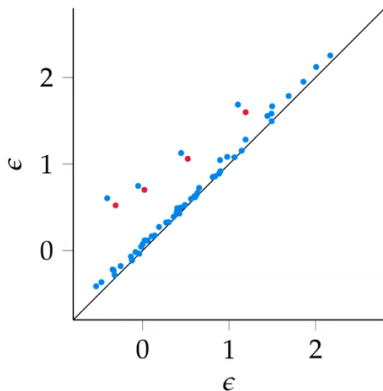
A drawback of the bottleneck distance is that its receptive to noise, as we are only looking at the furthest pair of the corresponding points. Hence we introduce the degree  $q$  Wasserstein distance between two persistence diagrams  $X$  and  $Y$

$$W_q(X, Y) = \left[ \inf_{\eta: X \rightarrow Y} \sum_{x \in X} \|x - \eta(x)\|_{\infty}^q \right]^{\frac{1}{q}}$$

Bottleneck distance

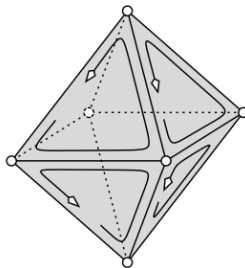
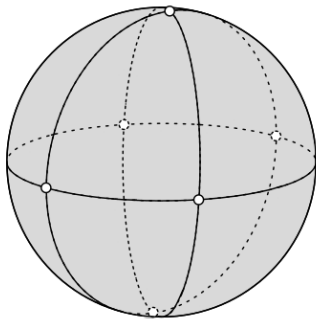


Wasserstein distance



## Triangulations

Sometimes it is useful to have a more structured representation of a space. Triangulation describes the replacement of topological spaces by piecewise linear spaces, i.e. the choice of a homeomorphism in a suitable simplicial complex. Spaces being homeomorphic to a simplicial complex are called triangulable.



## Stability theorem for filtrations

Let  $K$  be a simplicial complex and  $f, g : K \rightarrow \mathbb{R}$  are two monotonic functions. For each dimension  $p$ , the bottleneck distance between the diagrams  $X = \text{Dgm}_p(f)$  and  $Y = \text{Dgm}_p(g)$ ,  $W_\infty(X, Y) \leq \|f - g\|_\infty$

## Tame functions

Let  $X$  be triangulable. A function  $f : X \rightarrow \mathbb{R}$  is tame if it has a finite number of homological critical values, and its homology groups are finite dimensional. We call  $a \in \mathbb{R}$  a homological critical value if  $\nexists \varepsilon > 0$  for which  $f_p^{a-\varepsilon, a+\varepsilon}$  is an isomorphism for each dimension  $p$

We define tame functions to allow the stability theorem for more general functions

# Summary

- Read about simplicial complexes, their homologies, approximating Topological spaces using simplicial complexes.
- Analysing point clouds, by first filtering them then computing their persistent homologies
- Čech and Vietoris Rips Complex
- Persistence barcodes, diagrams, and landscapes.
- Stability of diagrams under perturbations.
- Tried reading about breast cancer detection using tda.
- Scikit-tda for most of my code simulations.

# Bibliography I



Herbert Edelsbrunner. *A Short Course in Computational Geometry and Topology*. Cham, Switzerland: Springer, 2014. URL: <https://www.springer.com/gp/book/9783319059560>.



Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Providence, Rhode Island: American Mathematical Society, 2010. URL: <https://www.ams.org/books/surv/082/>.



Ephraim Robert Love. "Machine Learning with Topological Data Analysis". PhD diss. University of Tennessee, 2021. URL: [https://trace.tennessee.edu/utk\\_graddiss/6170/](https://trace.tennessee.edu/utk_graddiss/6170/).



Wikipedia. *Homology (mathematics)*. [https://en.wikipedia.org/wiki/Homology\\_\(mathematics\)](https://en.wikipedia.org/wiki/Homology_(mathematics)). Accessed: 2024-07-01. 2023.



Wikipedia. *Persistent Homology*. [https://en.wikipedia.org/wiki/Persistent\\_homology](https://en.wikipedia.org/wiki/Persistent_homology). Accessed: 2024-07-01. 2023.



Wikipedia. *Topology*. <https://en.wikipedia.org/wiki/Topology>. Accessed: 2024-07-01. 2023.