IT Security

Topological Data Analysis

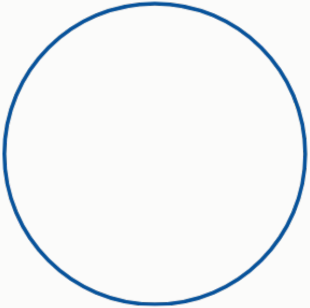
Christian Bungartz
bungartz@cs.uni-bonn.de
University of Bonn | Institute of Computer Science 4

UNIVERSITÄT BONN

About me

	Christian Bungartz
2015 – 2019	B.Sc. in Computer Science in Bonn
2019 – 2022	M.Sc. in Computer Science in Bonn
since 2022	Researcher at the University of Bonn
Interests	Anomaly Detection, Topological Data Analysis, Machine Learning, AI Security, …

Motivation



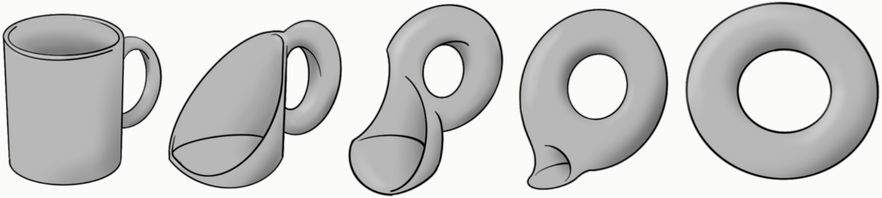
Motivation

Motivation

Geometric properties are those that are defined by **shape**, **size**, and **position**. Topological properties are those that are defined by **connectivity** and **continuity**.

Foundations of Topology

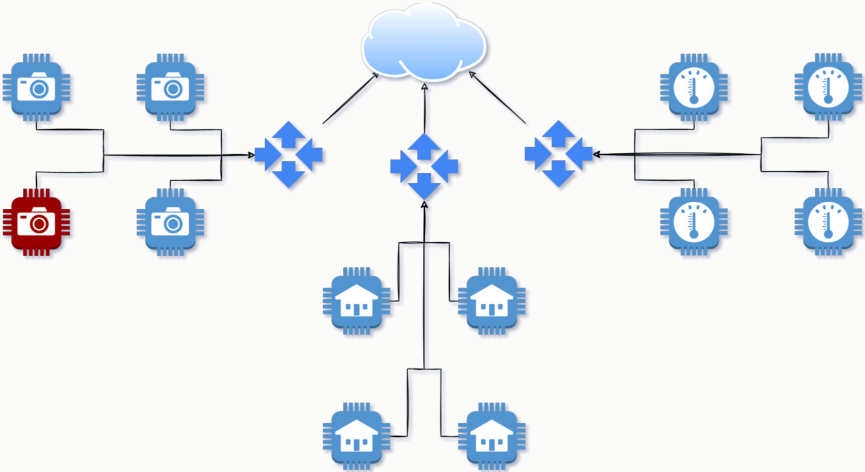
What is Topology?



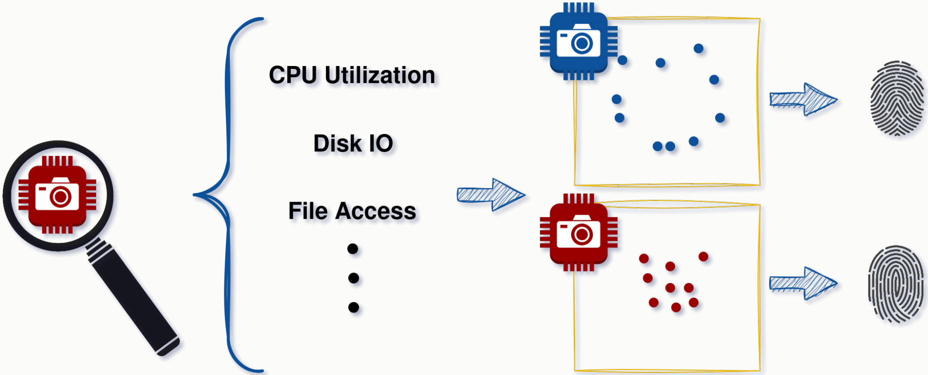
Topology is the study of properties that are preserved under **continuous deformations**, such as stretching, twisting, crumpling and bending. That is all operations without closing or opening holes, tearing or gluing, or passing through itself. In other words: Topology studies properties that are **invariant as long as the connectivity of the space does not change**.

UNIVERSITÄT BONN

Case Study - Introduction



Case Study - Introduction



Foundations of Topology

Essential Terminology

UNIVERSITÄT BONN

Essential Terminology

Metric Space

A metric space is a set X together with a metric $d : X \times X \to \mathbb{R}$ such that

- $d(x, y) \geq 0$
- $d(x, y) = 0 \iff x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leq d(x, y) + d(y, z)$

Invariant

An invariant is a quantity that remains unchanged by a certain classes of transformations. For example, the area of a square is an invariant under rotation.

Foundations of Topology

Topological Spaces

Topological Space

A **topological space** is a set X together with a collection of subsets T that tells us how to group elements together in a meaningful way. T can be thought of as "neighborhoods" within X

The pair (X, T) is called a topological space with T being the **topology** on X.

In practice we typically work with **metric spaces**, a special case of topological spaces. While topological spaces give a broad way to talk about closeness, **metric spaces** provide a concrete way by introducing distances between points.

Topological Space

In a metric space, the concept of closeness is defined by the distance between points. For each point $x \in X$ and any $\epsilon > 0$, we can define the neighborhood of x as the set of points within distance ϵ of x. This neighborhood is called the ϵ**-ball** around x, denoted $B_\epsilon(x)$:

$$B_\epsilon(x) = y \in X \mid d(x, y) < \epsilon$$

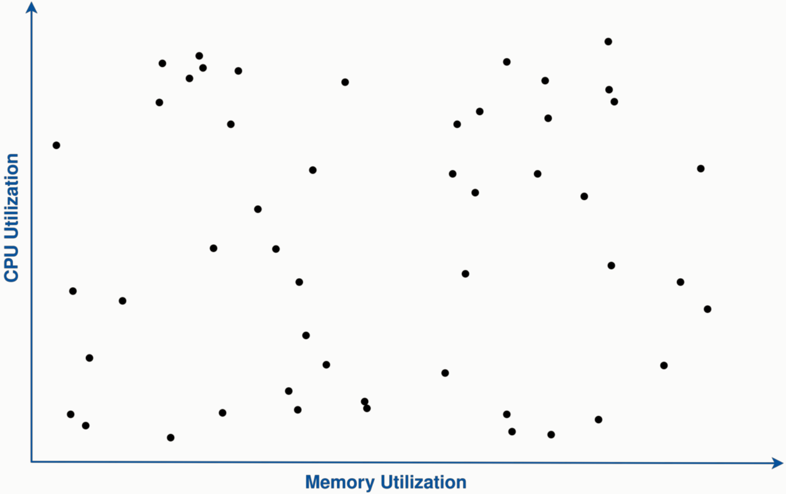
A subset $U \subseteq X$ is considered to be part of T iff

$$\forall x \in U : \exists \epsilon > 0 : B_\epsilon(x) \subseteq U$$

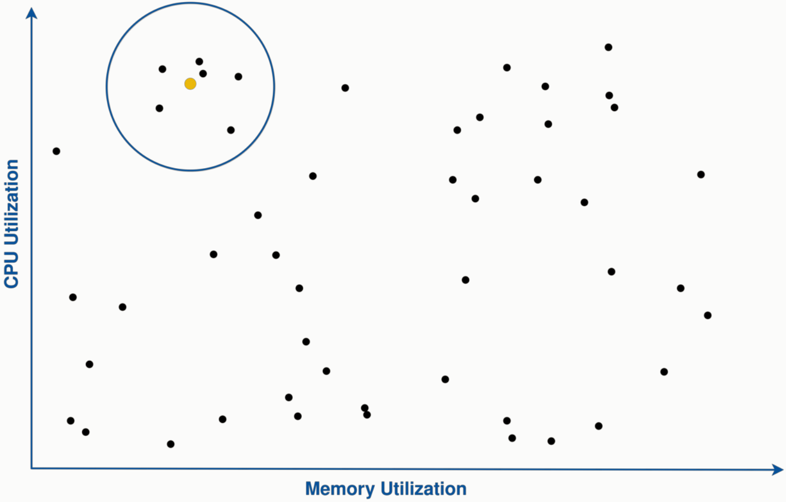
This means that points in U are ”locally near” each other in terms of distance.
Note that a metric space is always a topological space, but the converse is not true.

UNIVERSITÄT BONN

Example Data - Euclidean Space



Example Data - Euclidean Space



Topological Data Analysis

Topological Data Analysis

Topological Data Analysis (TDA) sits at the intersection of topology and data science. It allows us to study high-dimensional data (e.g. from point clouds, networks, etc.) in a quantitative way by examining the shape of the data (e.g. connected components, holes, etc.). In particular, TDA allows us to study **topological invariants**, which allows for the examination of higher-order interactions in the data.

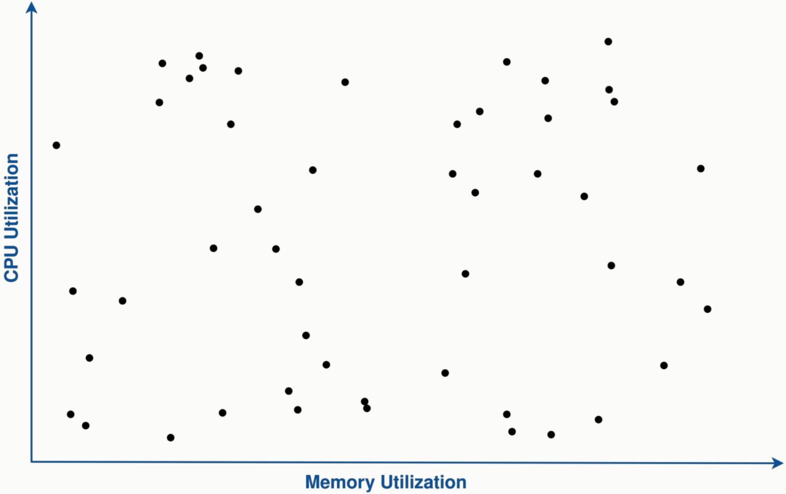
The pipeline of TDA

- **Data**
  - Derived from the real world (e.g. point cloud, network, image, …)
- **Topological Objects**
  - Constructed from the data, typically a simplicial complex (more about this later)
- **Homology Groups**
  - Computed from the topological objects (more about this later)

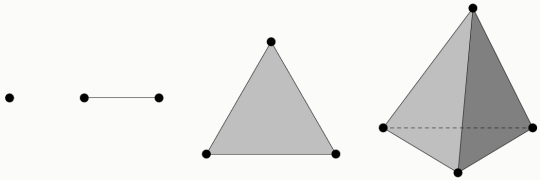
Topological Data Analysis

Simplicial Complexes

Simplicial Complexes



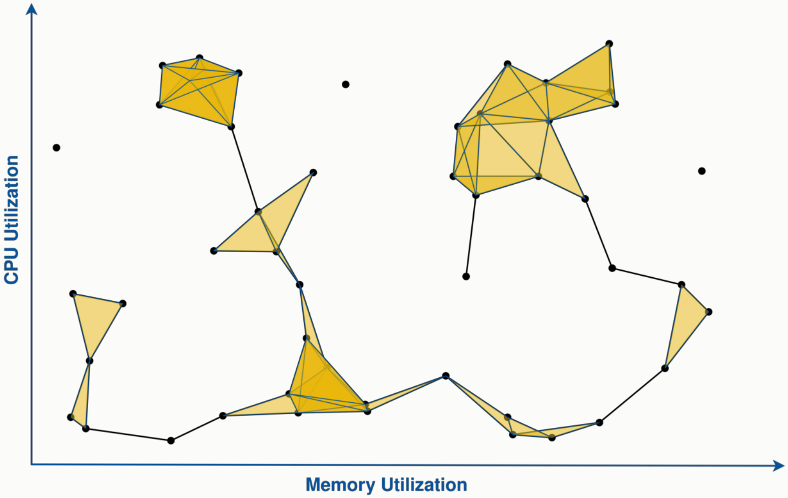
Simplicial Complexes



A simplicial complex K is a finite collection of **simplices** defining a **triangulation** of a topological space. In particular:

- Every face of each simplex in K is also in K
- The intersection of any two simplices in K is either empty or a face of both

Simplicial Complexes



Abstract Simplicial Complex

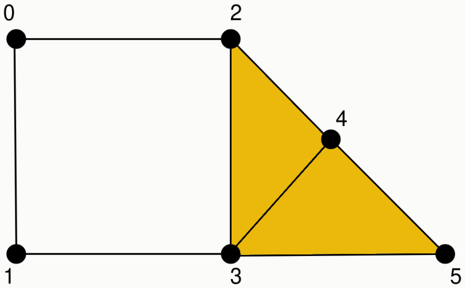
- An **abstract simplicial complex** is the general and combinatorial way of thinking about simplicial complexes.
- It consists of a set of vertices and a collection of subsets of these vertices (the **simplices**), such that if a set is in the collection, all of its subsets are as well.
- Simplices in this context are not geometric objects but rather sets. For example, a 2-simplex is represented by a triple $\{a, b, c\}$, indicating a triangle.

UNIVERSITÄT BONN

Geometric Realization

- The process of converting an abstract simplicial complex into a geometric one is known as **geometric realization**.
- Here, each simplex $\{v_0, v_1, \ldots, v_k\}$ is associated with the convex hull of points representing these vertices in a geometric space.
- This process helps visualize and understand the topological properties of the abstract complex by providing a tangible geometric structure.

Simplical Complex - Example



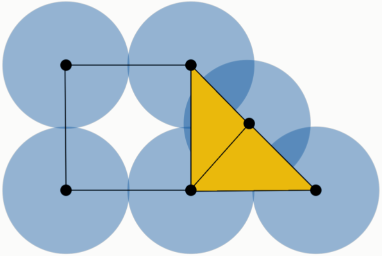
Abstract Simplicial Complex K:

$$\{\emptyset, \{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \{5\},$$
$$\{0,1\}, \{0,2\}, \{1,3\}, \{2,3\}, \{2,4\},$$
$$\{3,4\}, \{3,5\}, \{4,5\}, \{2,3,4\}, \{3,4,5\}\}$$

Vietoris Rips Complex

In practice we typically work with metric spaces. A common way to construct a simplicial complex from a metric space is the **Vietoris Rips Complex**. The Vietoris Rips Complex for a metric space X and a radius ϵ is given by

$$\mathrm{VR}_\epsilon(X) = \{\sigma \subseteq X \mid \forall x, y \in \sigma : d(x, y) \leq \epsilon\}$$



Topological Data Analysis

Persistent Homology

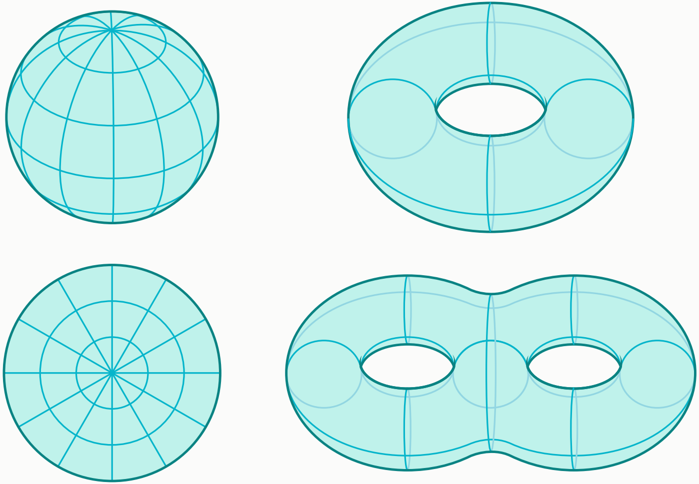
Homology

Homology enables the association of a sequence of algebraic structures (e.g. vector spaces) to a topological space.

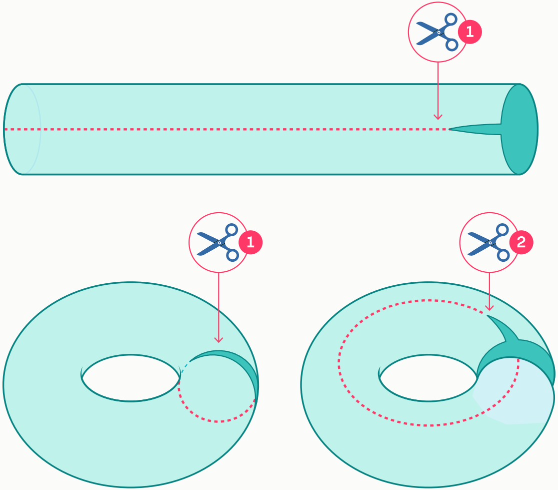
- These structure are algebraic invariants of the topological space (i.e. depend only on the topological properties of the space)
- Measures the number of connected components, 1-dimensional holes, 2-dimensional voids and so on
- Comparing these invariants allows us to compare the respective spaces

Definition: Homology
Given a space X homology is a sequence of abelian groups $H_0(X), H_1(X), H_2(X), \ldots$ (the homology groups), where $H_d X$ gives an measure of the number of "d-dimensional holes" in X.

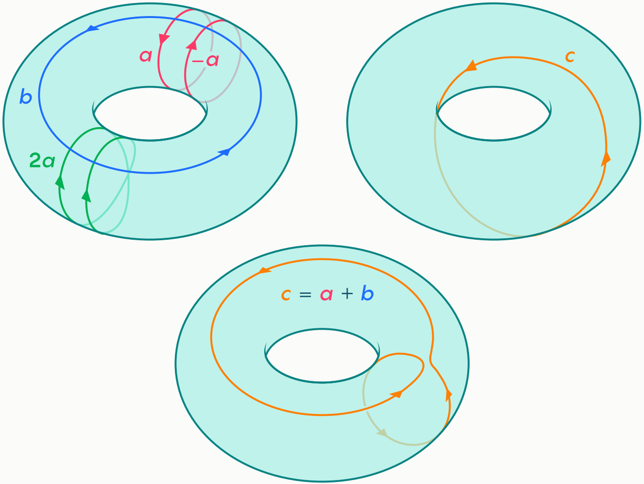
Homology



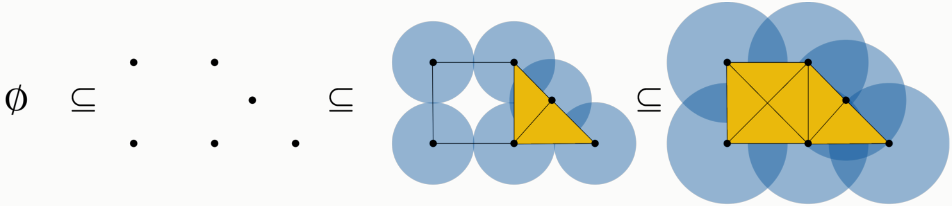
Homology



Homology



Filtration



A filtration is a nested sequence of simplicial complexes $K_0 \subseteq K_1 \subseteq K_2 \subseteq \cdots \subseteq K_n$ such that K_{i-1} is a subset of K_i. For Vietoris Rips Complexes the filtration is given by the increase of the radius ϵ from 0 to ∞.

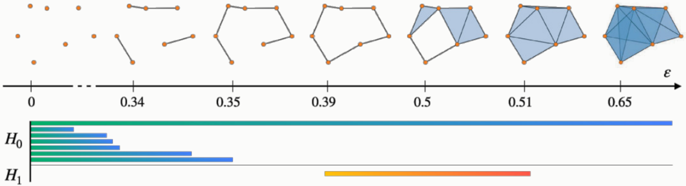
Persistent Homology

Persistent homology allows us to study the evolution of homology groups over a filtration. In other words it allows us to keep track of the i-dimensional holes and their **persistence**. At any given filtration threshold ϵ_i persistent homology tells us:

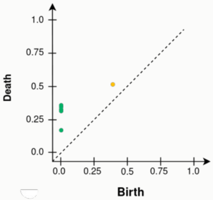
- which holes were not present at ϵ_{i-1}
- which holes die at scale ϵ_i
- which holes persist

The persistence of a feature can be seen as a measure of its importance. The longer a feature persists, the more important it is. We use **barcodes** and **persistence diagrams** to visualize the persistence of features.

Understanding Barcodes and Persistence Diagrams



H_0	H_1
(0,∞)	(0.39,0.51)
(0,0.18)	
(0,0.31)	
(0,0.33)	
(0,0.34)	
(0,0.35)	
(0,0.35)	



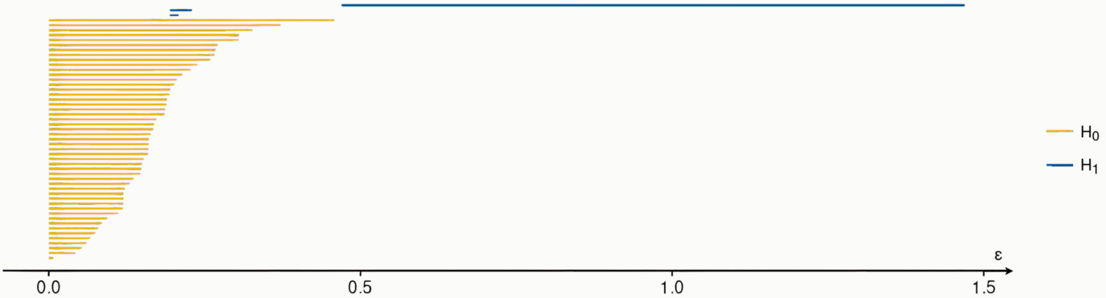
[1]

[1] https://www.nature.com/articles/s41598-021-84486-1

TDA in Practice

Topological Pipeline

- **Data:** Behavior of multiple homogenous devices each presented by a point cloud
- **Topological Objects:** Vietoris Rips Complexes for each point cloud based on euclidean distance, where a point cloud represents device behavior over a time period *t*

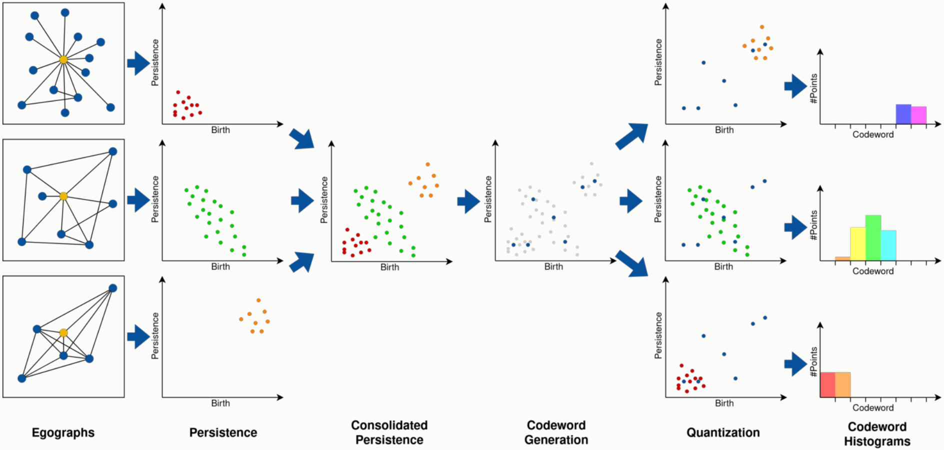


Feature Extraction - Defining Quantities

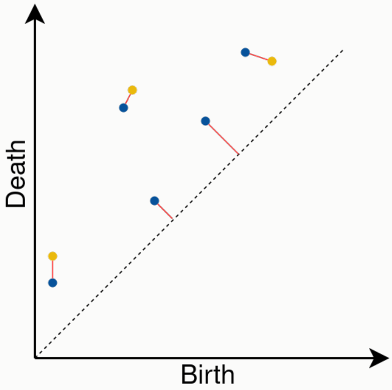
- How do we interpret a barcode?
- We look at the number of features per dimension, the persistence of the features, relevant outliers, ...
    - → Let’s put all of this into a vector.

We represent a barcode by a vector of **defining quantities** for each dimension. Typically this includes the persistence’ mean, median, standard deviation, sum, minimum, the persistence of the three longest living features, the birth time of the longest living feature, and the number of features.

Feature Extraction - Persistence Code-books



Persistence Distance



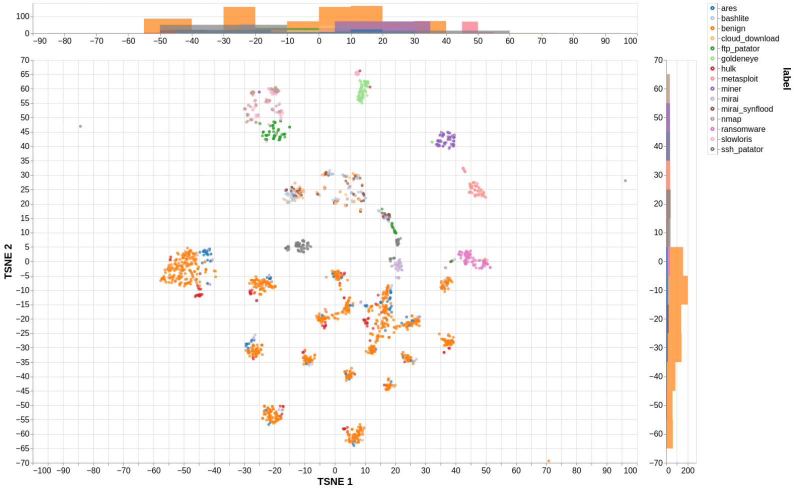
The most common distance metrics are the **Bottleneck distance** and the **Wasserstein distance**. Given a perfect alignment η of two persistence diagrams $\mathcal{X}$ and $\mathcal{Y}$ the Bottleneck distance is given by

$$w_\infty(\mathcal{X}, \mathcal{Y}) = \inf_\eta \sup_{x \in \mathcal{X}} \|x - \eta(x)\|_\infty$$

and the Wasserstein distance — depending on the L^p-norm — by

$$w_q^p(\mathcal{X}, \mathcal{Y}) = \inf_\eta \Big(\sum_{x \in \mathcal{X}} \|x - \eta(x)\|_p^q\Big)^{\frac{1}{q}}$$

Embedding based on the Wasserstein Distance



UNIVERSITÄT BONN

Conclusion

- TDA allows us to study high-dimensional data in a quantitative way by examining the shape of the data.
- In particular, TDA allows us to study topological invariants, which allows for the examination of higher-order interactions in the data.
- In practice we typically utilize the persistent homology of a simplicial complex representing the data.
- Persistent homology allows us to study the evolution of homology groups over a filtration and is typically visualized using barcodes and persistence diagrams.
- The most straightforward way to utilize TDA in cybersecurity is to extract features from these representations

Lightning Survey

Lightning Surveys

Christian Bungartz

University of Bonn | Institute of Computer Science 4

bungartz@cs.uni-bonn.de