# VRDI TDA Breakout Session: Applications of Topological Data Analysis

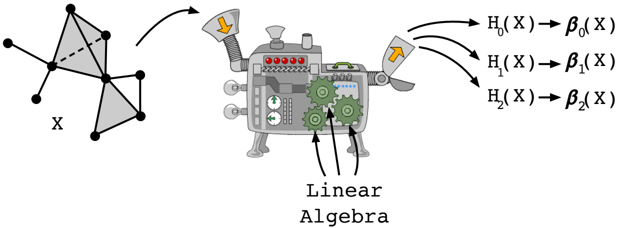
Moon Duchin, Tom Needham, Thomas Weighill

Voting Rights Data Institute
June 26, 2019

# Quick Review of TDA

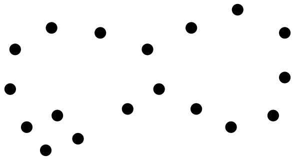
Topology studies geometrical objects (called spaces) up to a loose notion of equivalence.

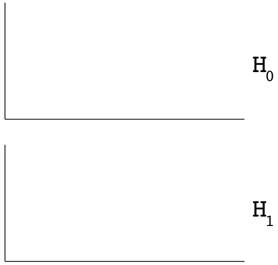
Algebraic topology distinguishes spaces by computing invariants; e.g., Betti numbers $\beta_k(X)$ count $k$-dimensional holes in a space $X$.



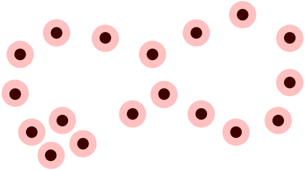

Topological Data Analysis explores the shape of a dataset (e.g. a point cloud in $\mathbb{R}^d$) by computing invariants across a family of spaces generated from the dataset.

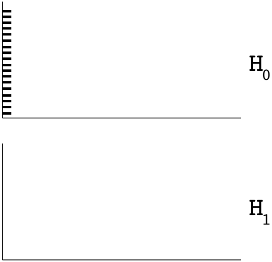
# Persistent Homology - An Example



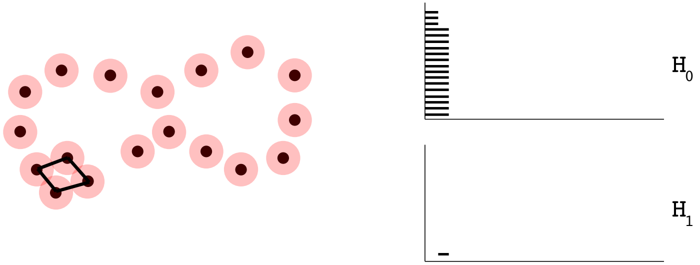


# Persistent Homology - An Example

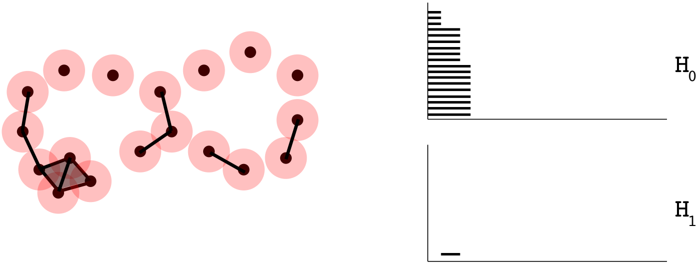




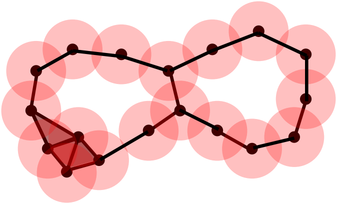
# Persistent Homology - An Example

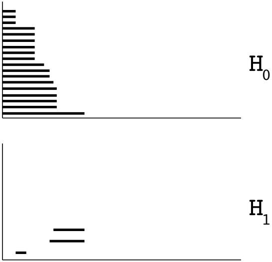


# Persistent Homology - An Example

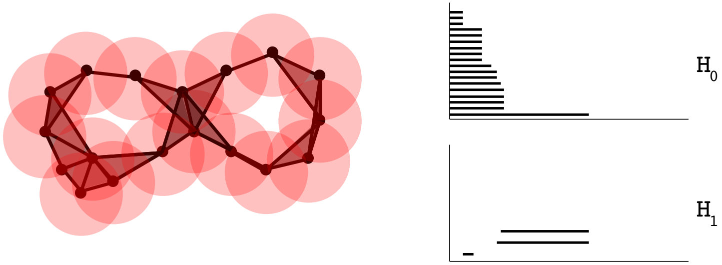


# Persistent Homology - An Example

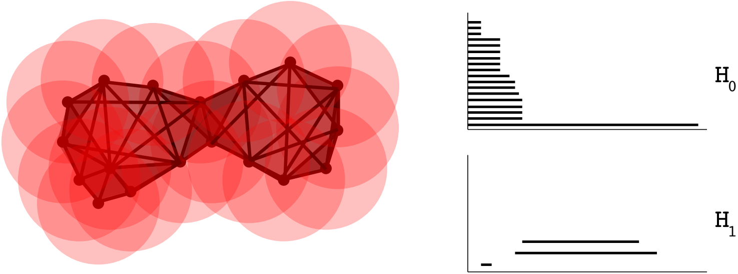




# Persistent Homology - An Example

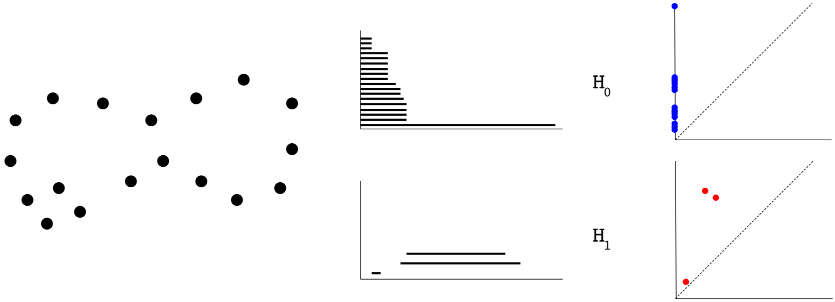


# Persistent Homology - An Example



# Terminology

We can record the "birth time" and "death time" of each topological feature to get a barcode or a persistence diagram.




Dataset X

Barcodes for X

Persistence Diagrams for X

# Distance Between Diagrams

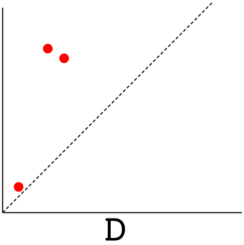
Important piece of the story: comparing persistence diagrams.

Each diagram $D$ is a set[1] of points

$$D = \{(b_i, d_i)\}_{i=1}^{N}$$

with each $b_i < d_i$. Each point in $D$ represents a topological feature of a dataset.

Let $\mathcal{D}$ denote the set of all diagrams.



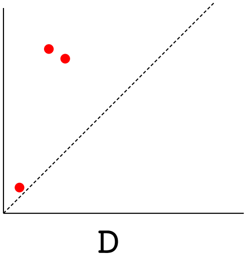

[1] Actually it's a multiset, but let's ignore that...

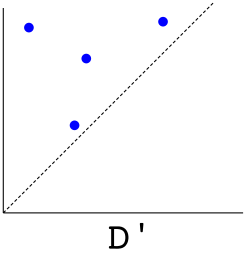
# Metric on Diagrams

We wish to define a metric on $\mathcal{D}$.

This is a function $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ satisfying:

- (Positivity) $d(D, D') = 0 \Leftrightarrow D = D'$
- (Symmetry) $d(D, D') = d(D', D)$
- (Triangle Inequality) $d(D, D'') \leq d(D, D') + d(D', D'')$.




How "close"?

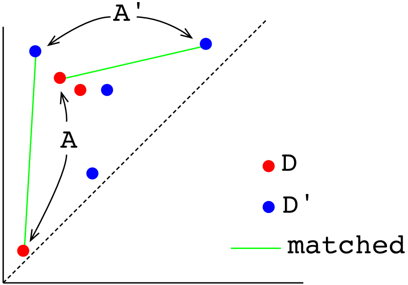
# Idea: Match Features

Given diagrams $D$, $D'$, we want to match points as best as possible.

Problem: diagrams might have a different number of points! So we can't expect a perfect matching.

Let $\phi : A \to A'$ be a bijection, where $A \subset D$ and $A' \subset D'$.

This is called a partial matching.



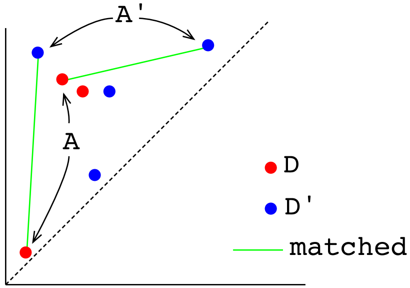
# Idea: Match Features

Pick a matching cost $c_m$ (TBD).

For each $p = (b, d) \in A$, we evaluate $c_m(p, \phi(p))$.

Pick a cost for unmatched points $c_u$ (TBD).

For each $p \in D \setminus A$ and each $p' \in D' \setminus A'$, we evaluate $c_u(p)$ and $c_u(p')$.



# Idea: Match Features

For a partial matching $\phi$, we evaluate the total cost:

$$\mathrm{TC}(\phi) = \max\left\{\max_{p\in A} c_m(p,\phi(p)), \max_{p\notin A} c_u(p), \max_{p'\notin A'} c_u(p')\right\}.$$

Define a metric as min over all partial matchings

$$d(D,D') = \min_{\phi} \mathrm{TC}(\phi).$$

For theoretical reasons, choose costs for $p = (b,d)$ and $p' = (b',d')$ as

$$c_m(p,p') = \max\{|b'-b|, |d'-d|\}, \quad c_u(p) = \frac{d-b}{2}.$$

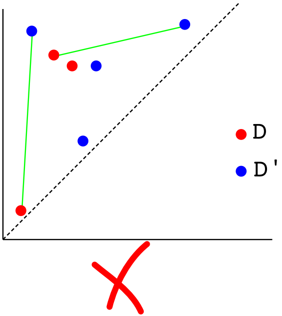
# Bottleneck Distance

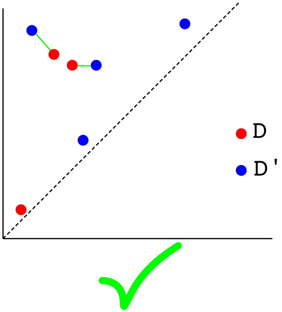
The bottleneck distance between persistence diagrams $D$ and $D'$ is

$$d_b(D, D') = \min_{\phi} \max \left\{ \max_{p \in A} c_m(p, \phi(p)), \max_{p \notin A} c_u(p), \max_{p' \notin A'} c_u(p') \right\}$$

with min over partial matchings and

$$c_m(p, p') = \max\{|b' - b|, |d' - d|\}, \quad c_u(p) = \frac{d - b}{2}.$$



# Applications

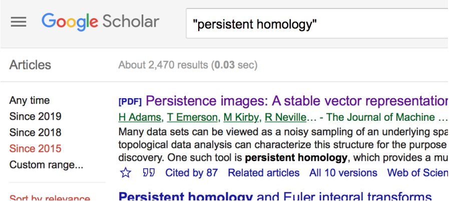
## Question

So what is this actually good for?

Let's look at some interesting examples of TDA “in action".

## Disclaimer

This is very far from an exhaustive list!

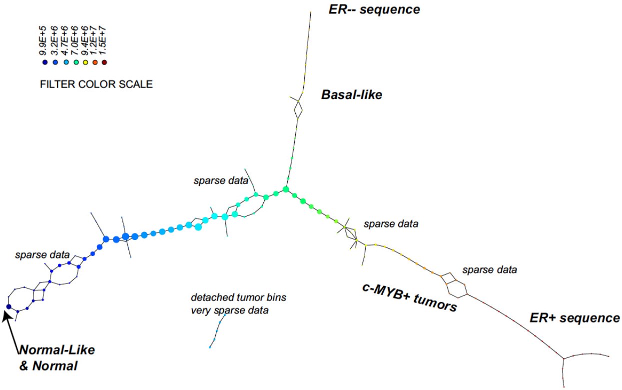



The applications are skewed toward my own interests and things which I think might be relevant to the districting problem.

# Biomedicine

*Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival*

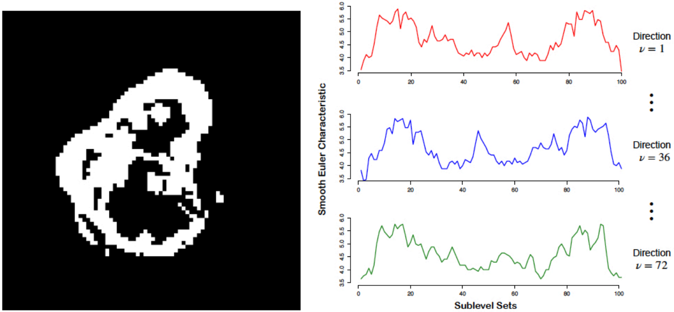
Nicolau, Levine, Carlsson, 2011.



# Biomedicine

*Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform*

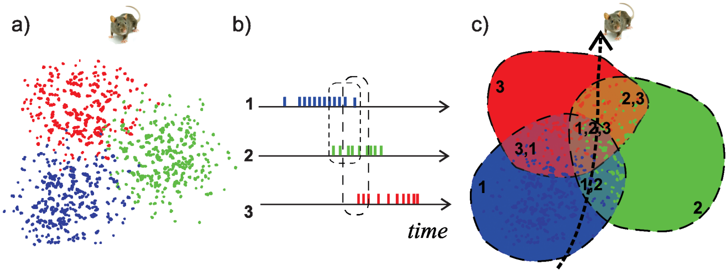
Crawford, Monod, Chen, Mukherjee, Rabadán, 2019



# Cognition

*A Topological Paradigm for Hippocampal Spatial Map Formation Using Persistent Homology*

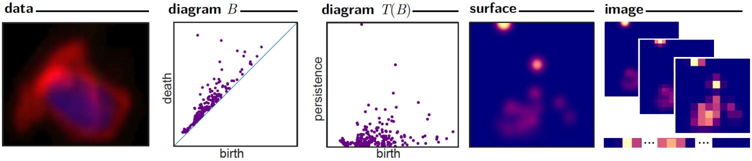
Dabaghian, Mémoli, Frank, Carlsson, 2012



# Machine Learning

## Extracting Feature Vectors with TDA

- Persistence Landscapes, Bubenik, 2015
- Persistence Images, Adams et. al. 2016




## TDA Layers in Neural Networks

- Brüel-Gabrielsson et. al. 2019
- Carrieère et. al. 2019

## Studying the Structure of Neural Networks

- Guss, Salakhutdinov, 2018
- Rieck et. al. 2019

# Shape Analysis

*Persistent Homology Transform for Modeling Shapes and Surfaces*

Turner, Mukherjee, Boyer, 2013

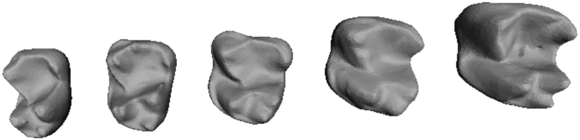


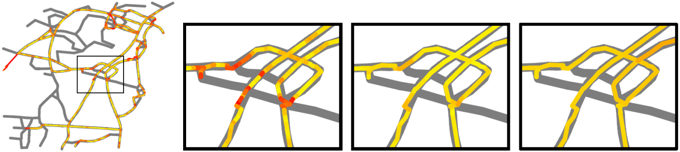
Figure 1: Images of the meshes of five teeth. A common problem in morphology is to measure distances between these five teeth.

Continued in:
Curry, Mukherjee, Turner, 2018
Ghrist, Levanger, Mai, 2018

# Shape Analysis

*Local Persistent Homology Based Distance Between Maps*

Ahmed, Fasy, Wenk, 2014



(a) Local Homology (b) LH Detailed View (c) Hausdorff (d) Fréchet

# Shape Analysis

*Gromov-Hausdorff Stable Signatures for Shapes using Persistence*

Chazal, Cohen-Steiner, Guibas, Mémoli and Oudot, 2009

Theorem

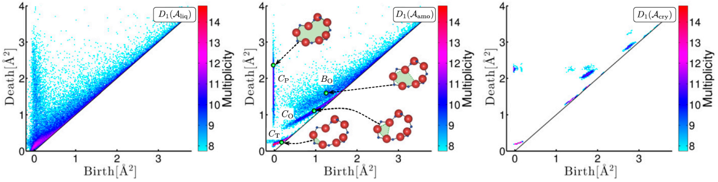
Let $(X, d_X)$ and $(Y, d_Y)$ be finite metric spaces. Let $D_k(X)$ and $D_k(Y)$ be the persistence diagrams for the $k$-dimensional persistent homology of their Vietoris-Rips complexes. Then

$$d_b(D_k(X), D_k(Y)) \leq d_{GH}(X, Y),$$

where $d_{GH}$ denotes Gromov-Hausdorff distance.

# Other Cool Examples

- Materials Science: Hiraoka, et. al. 2015




- Human Activity Recognition: Venkataraman et. al. 2016
- Biobiotic Insect Networks: Dirafzoon, et. al. 2017