Topological Data Analysis and Machine Learning

Hamilton-GCA

Introduction

John Harer

Duke University

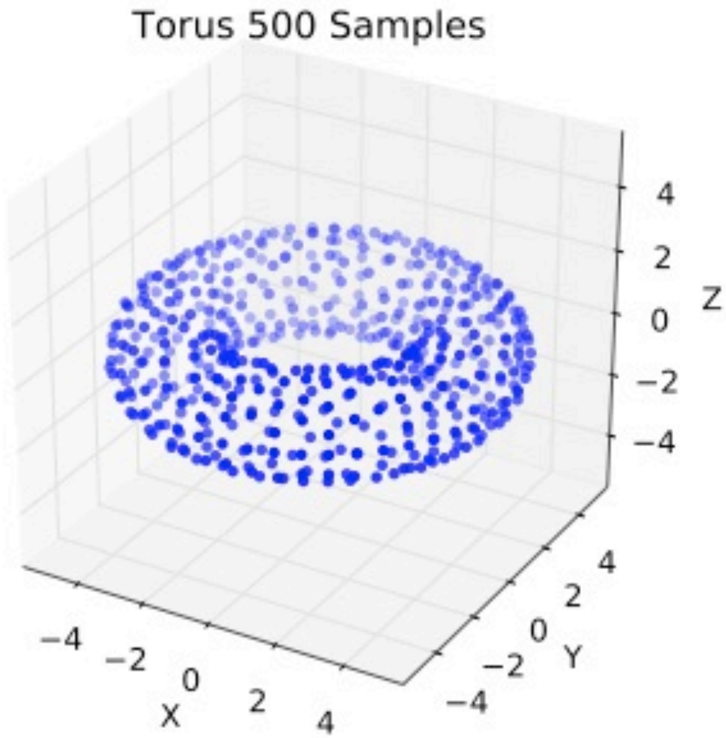
Geometric Data Analytics, Inc.

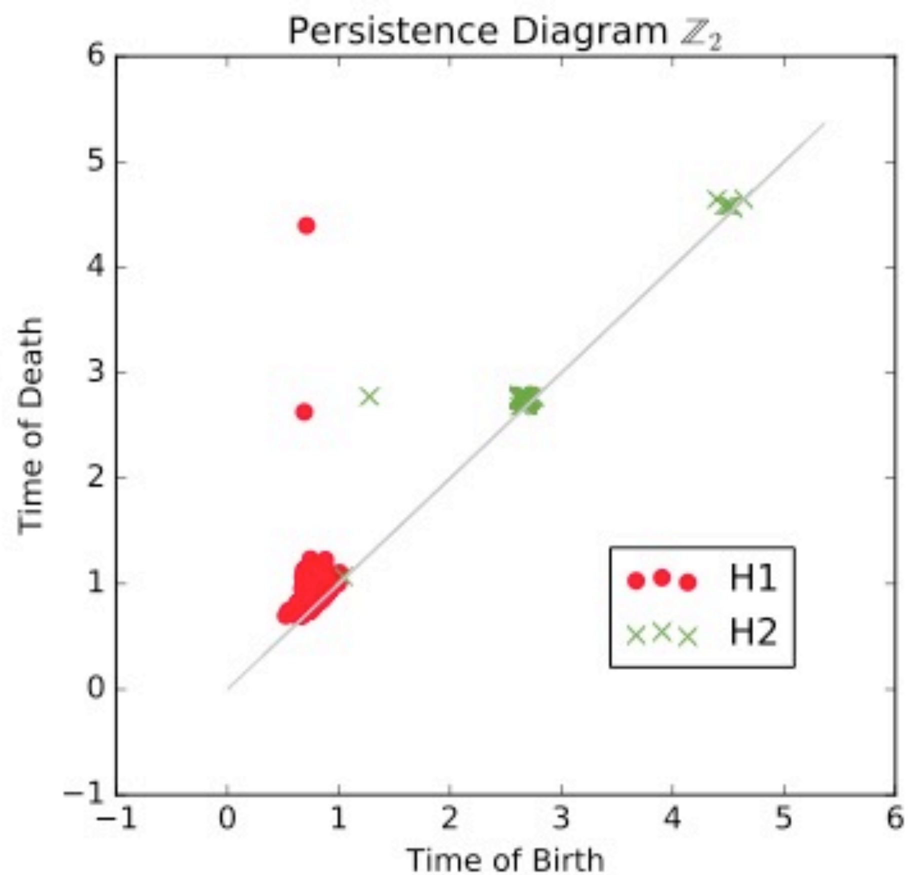
SOFTWARE FROM CHRIS TRALIE: http://github.com/ctralie/TDALabs

Topics

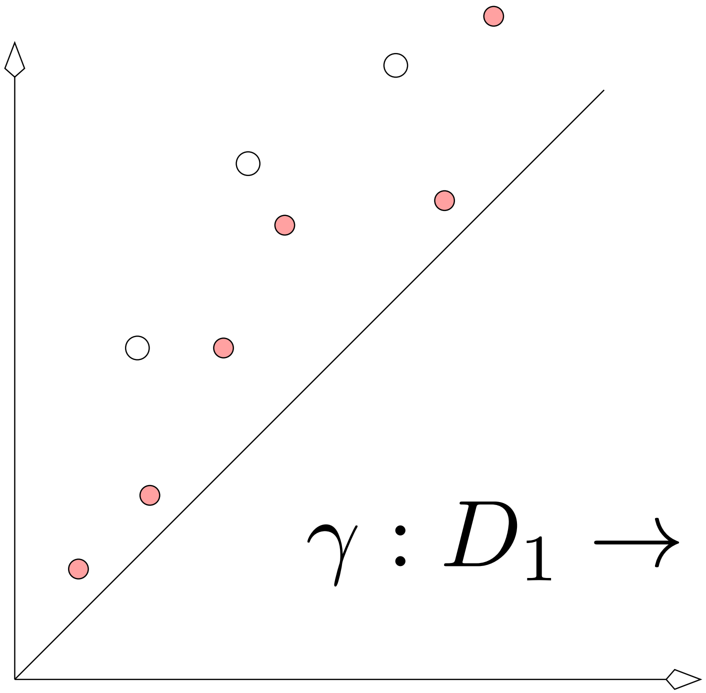
0) Overview
1) Persistence for Functions
2) Simplicial Complexes, Filtrations
3) Homology, Persistent Homology, Matrix Reduction
4) Alpha, Cech and Rips complexes
5) Metrics on Persistence Diagrams, Stability
6) Statistics on Persistence Diagrams

1- Persistence and Persistence Diagrams

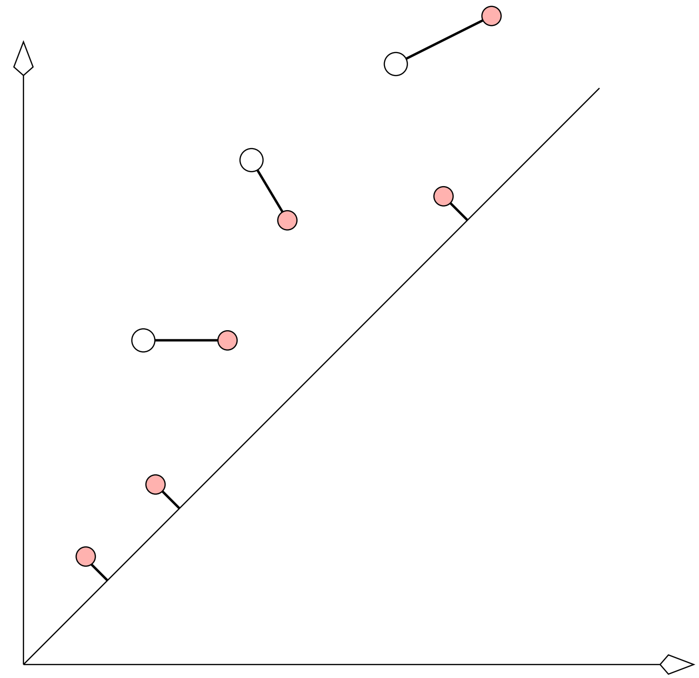




2- Stability and Metrics



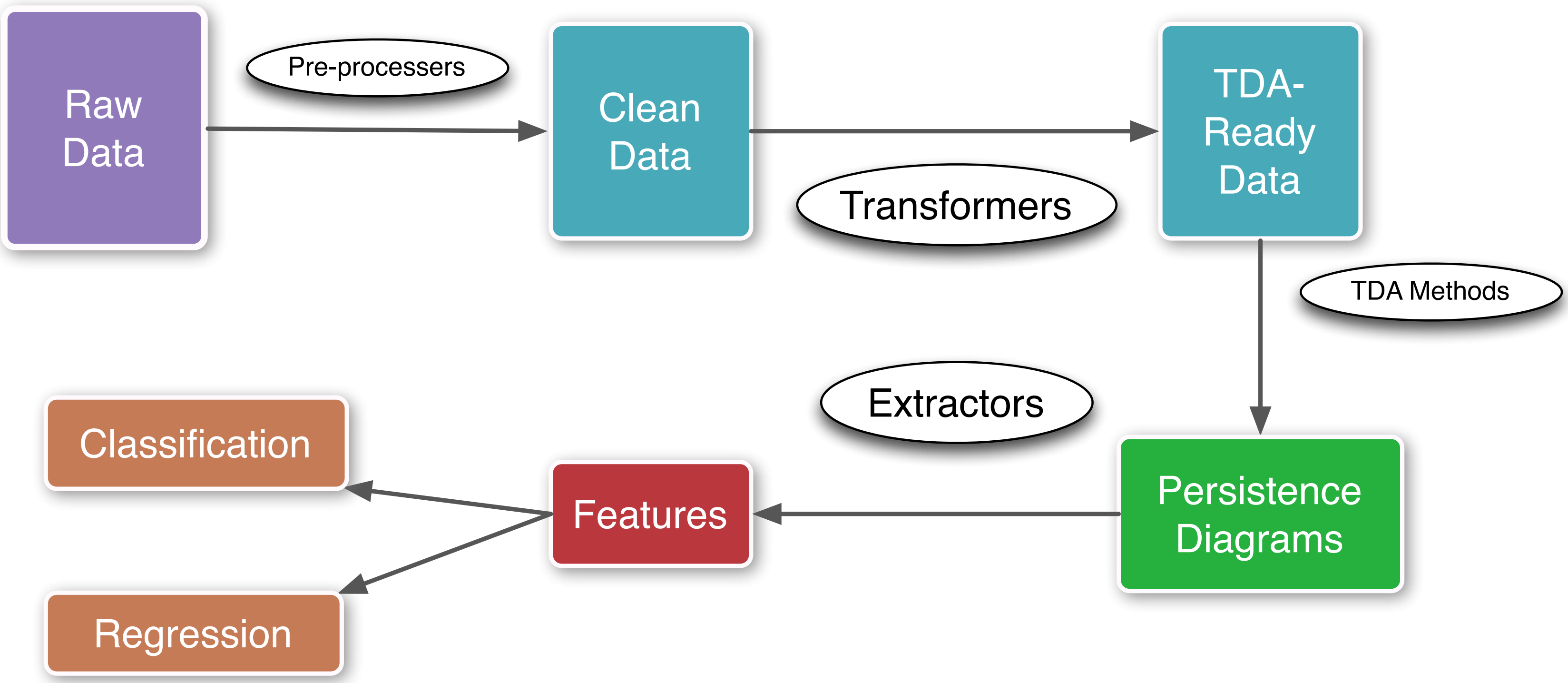
$$\gamma : D_1 \to D_2$$



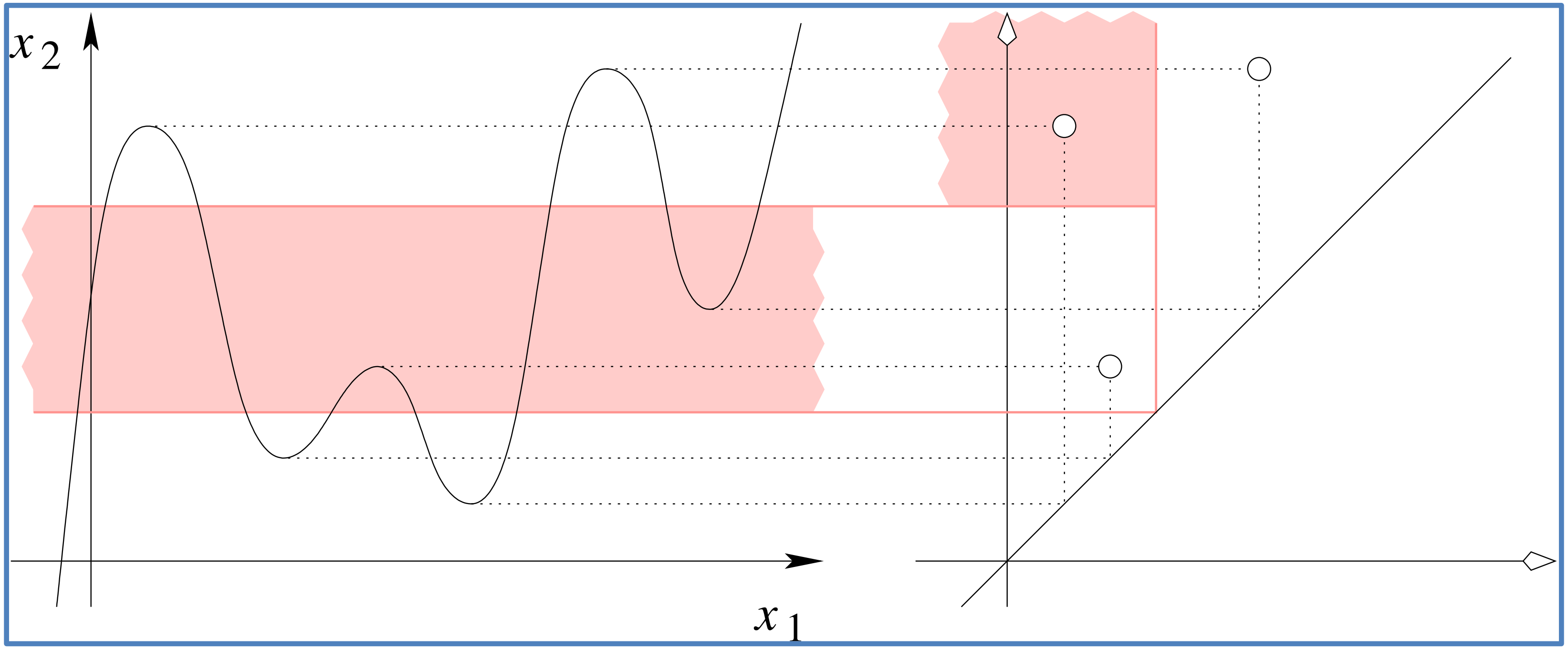
Matching

$$W_p(D_1, D_2) = min_\gamma(\sum_{x \in D_1} ||x - \gamma(x)||_\infty^p)^{1/p}$$

3 - TDA + Machine Learning



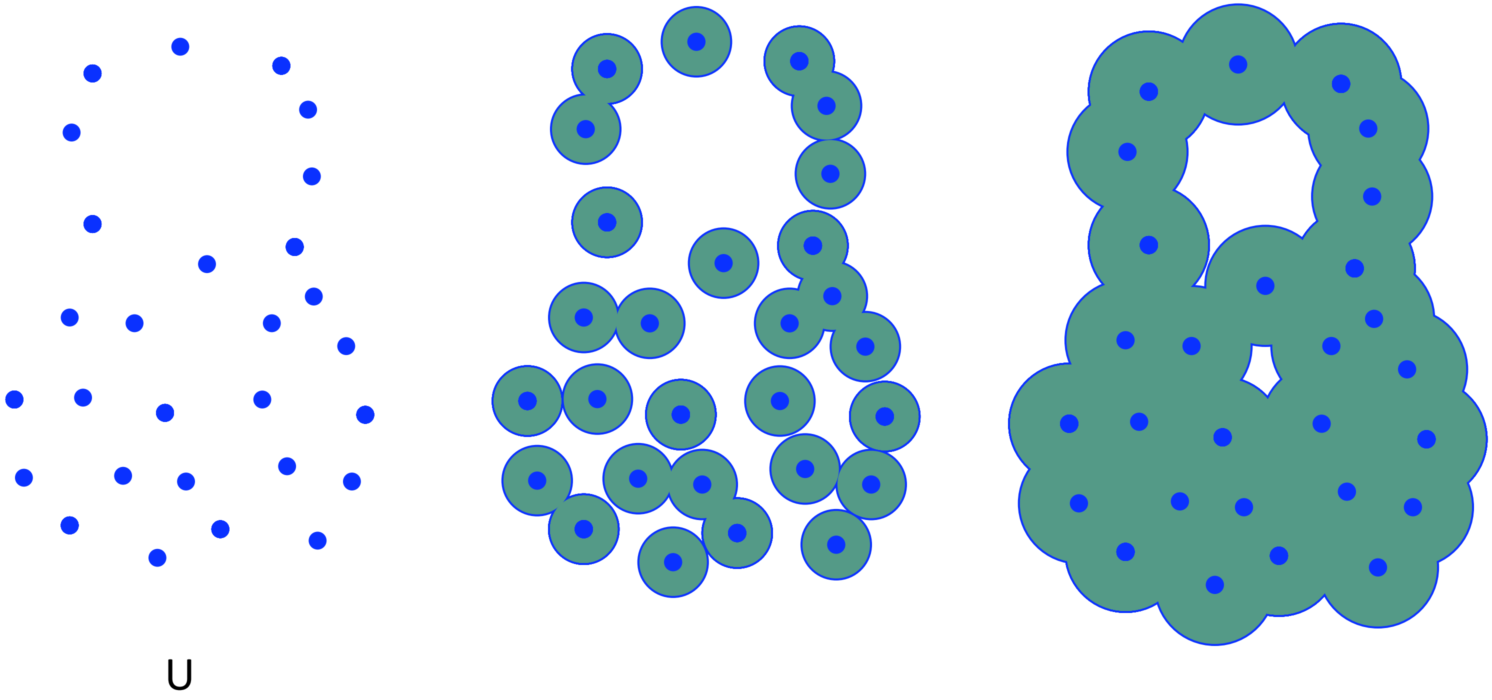
Shape of a Function
Persistence



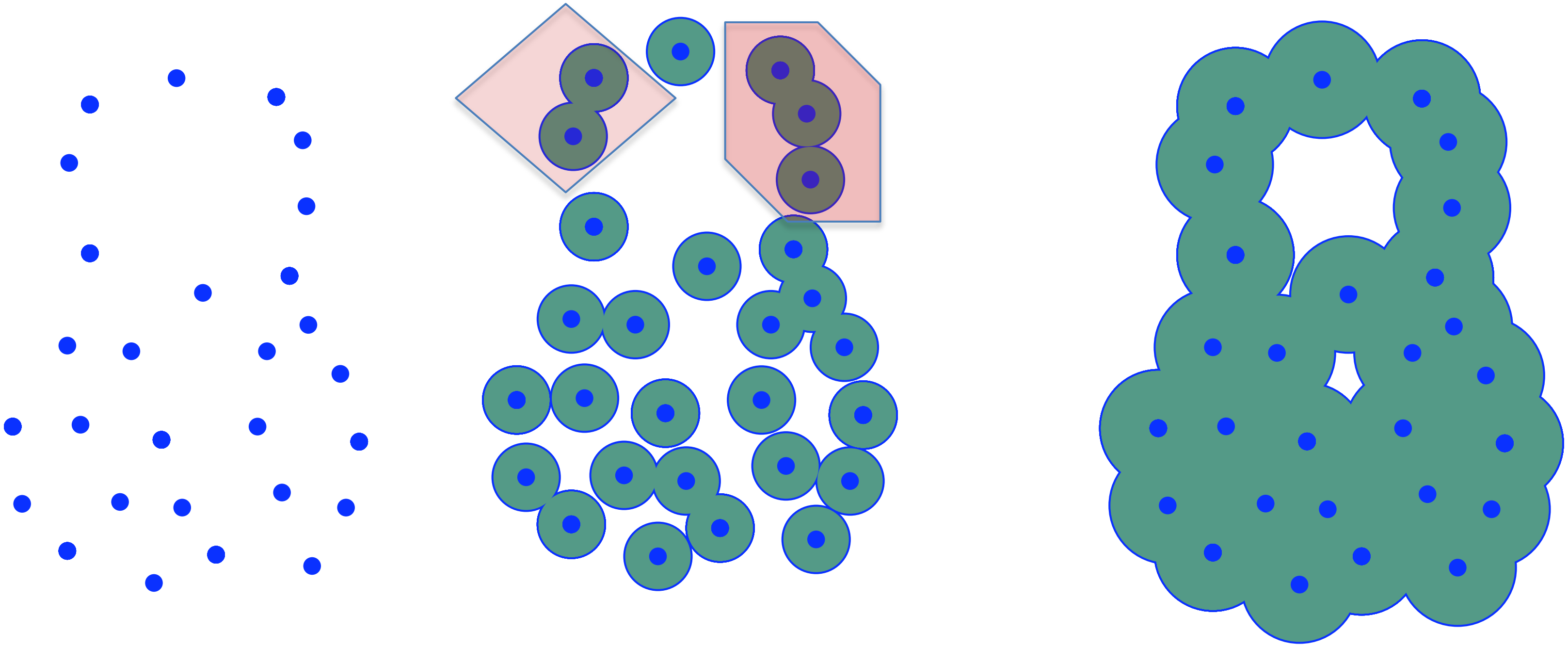
Pair Mins and Maxs in a careful way.
Sublevelset filtration.

Shape of a Set of Points (e.g. Sample)

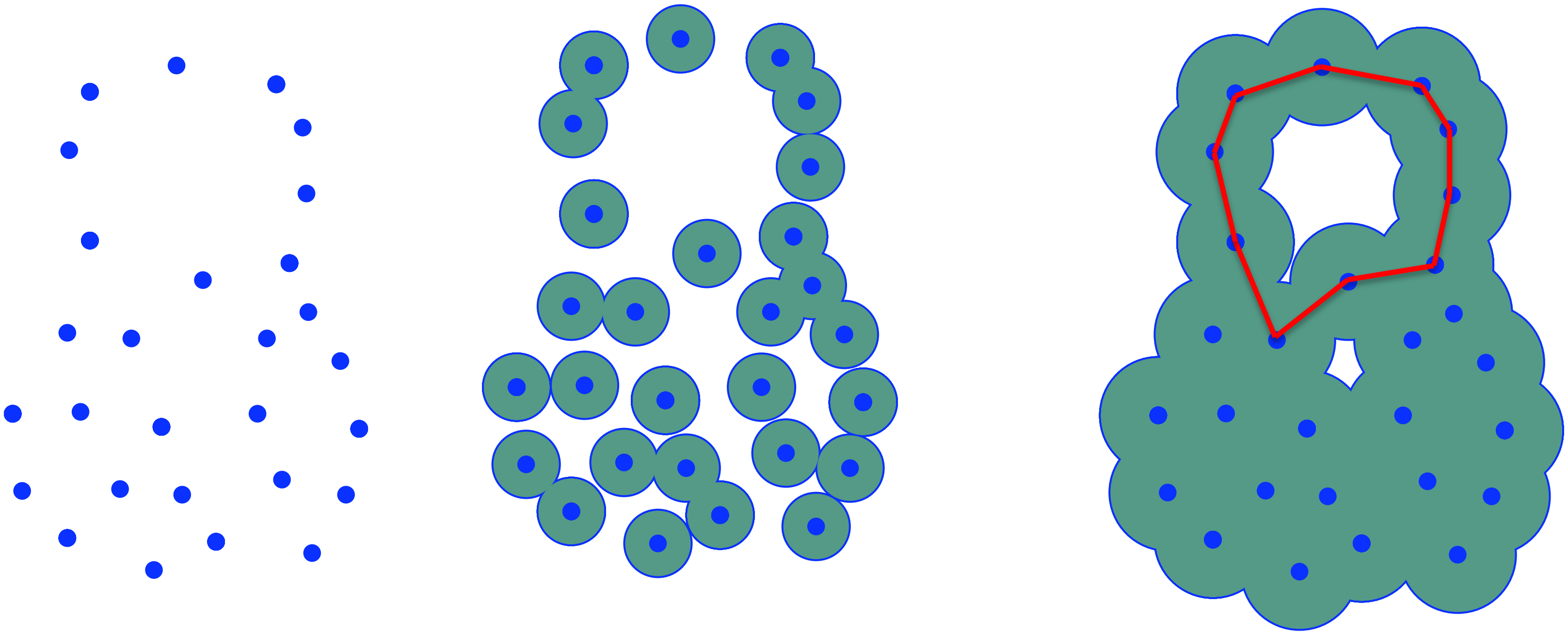
Track Components and 1-Cycles as balls grow around U



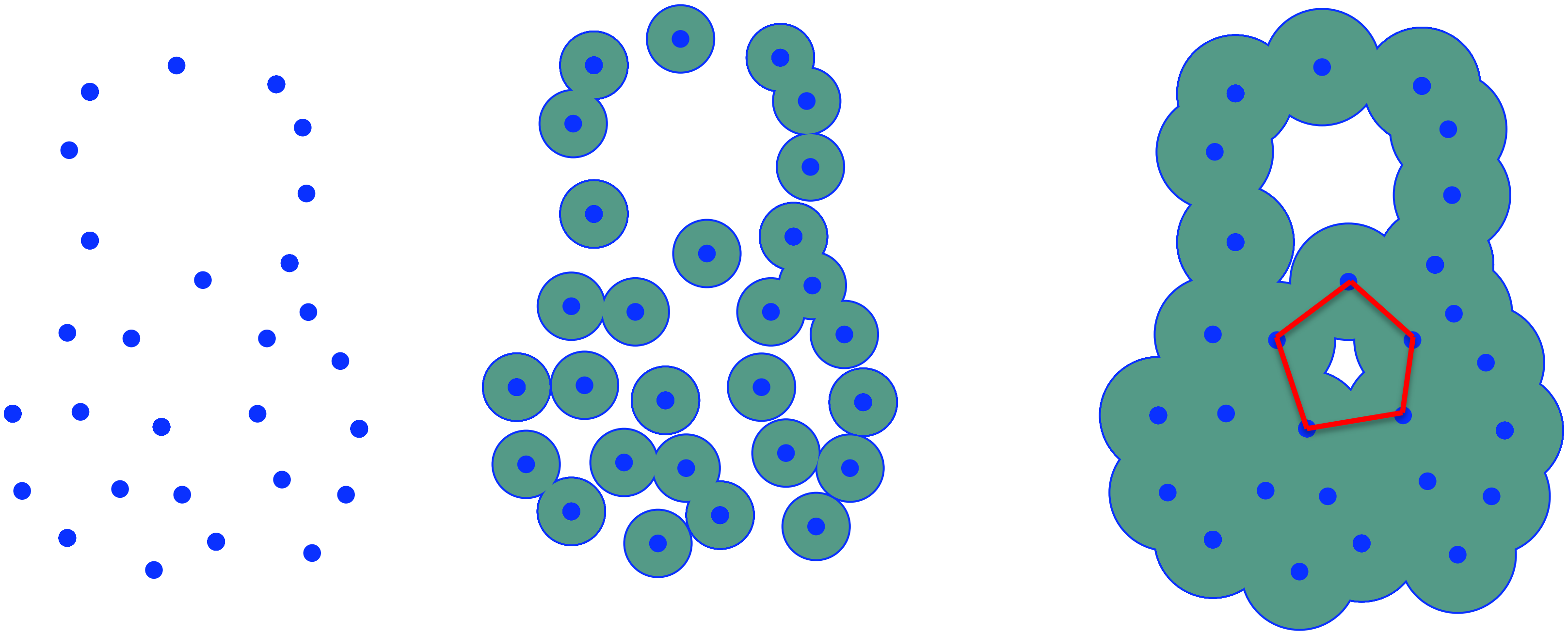
Track Components and 1-Cycles



Track Components and 1-Cycles

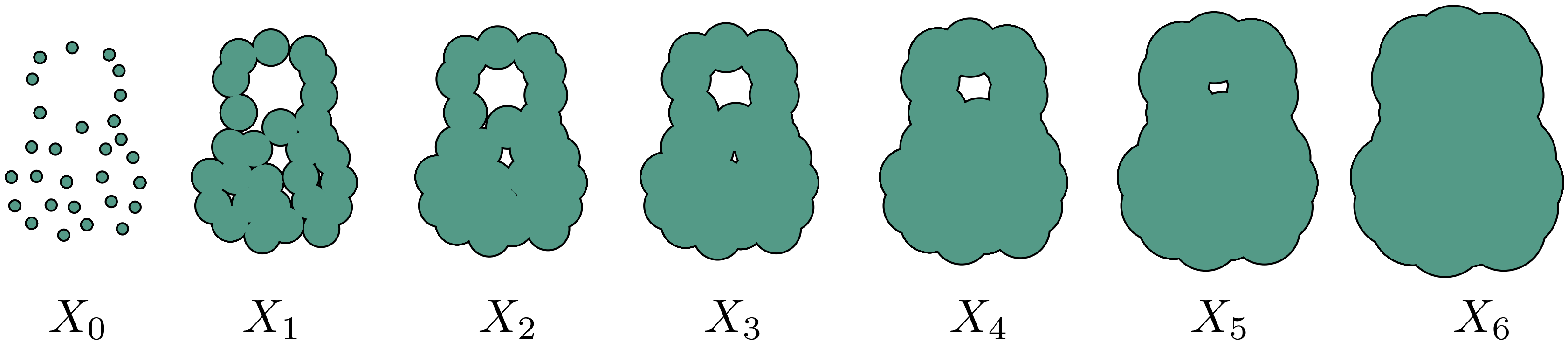


Track Components and 1-Cycles

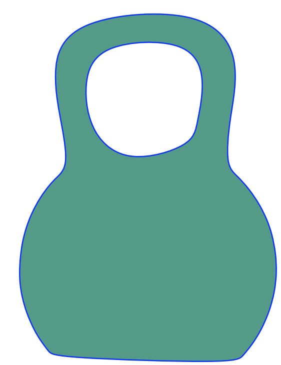


Filtration

$$X_0 \subset X_1 \subset \cdots \subset X_n$$

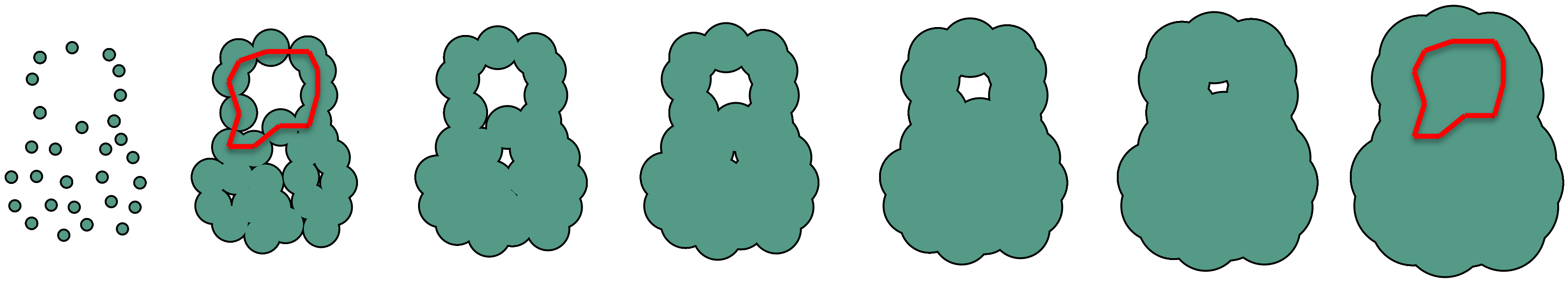


Shape Never Occurs at a Fixed Scale
But homology class (cycle) persists
for some period in the filtration.



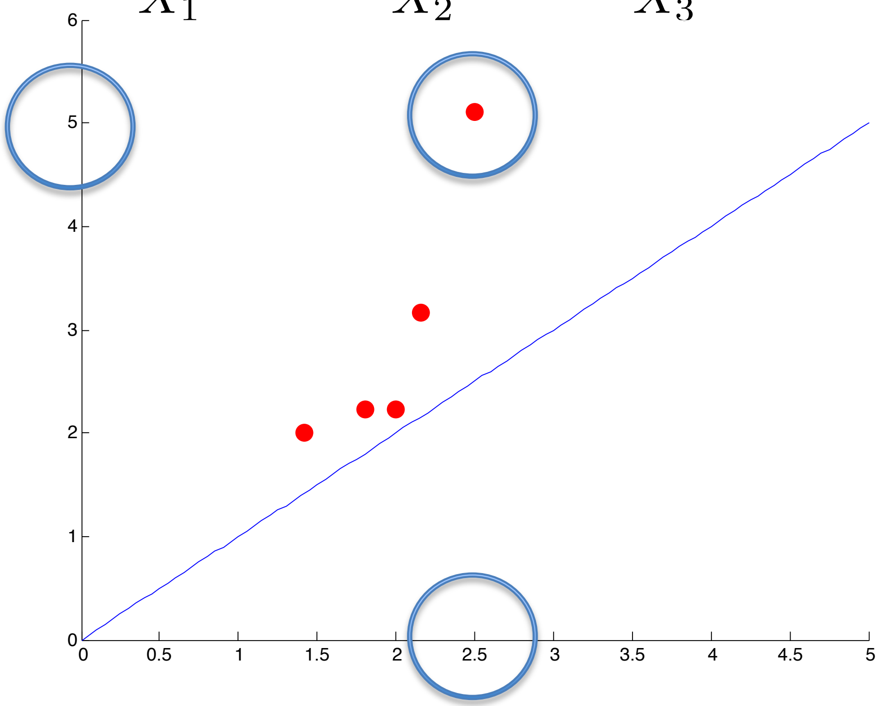
Persistence Diagram

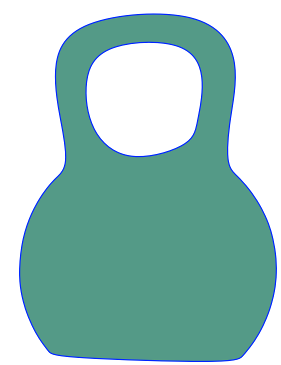
$$X_0 \subset X_1 \subset \cdots \subset X_n$$



X_0 X_1 X_2 X_3 X_4 X_5 X_6

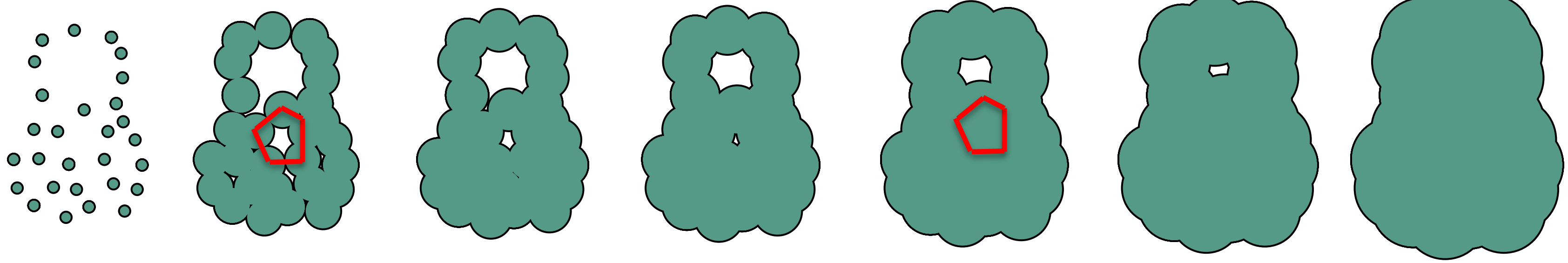
1D





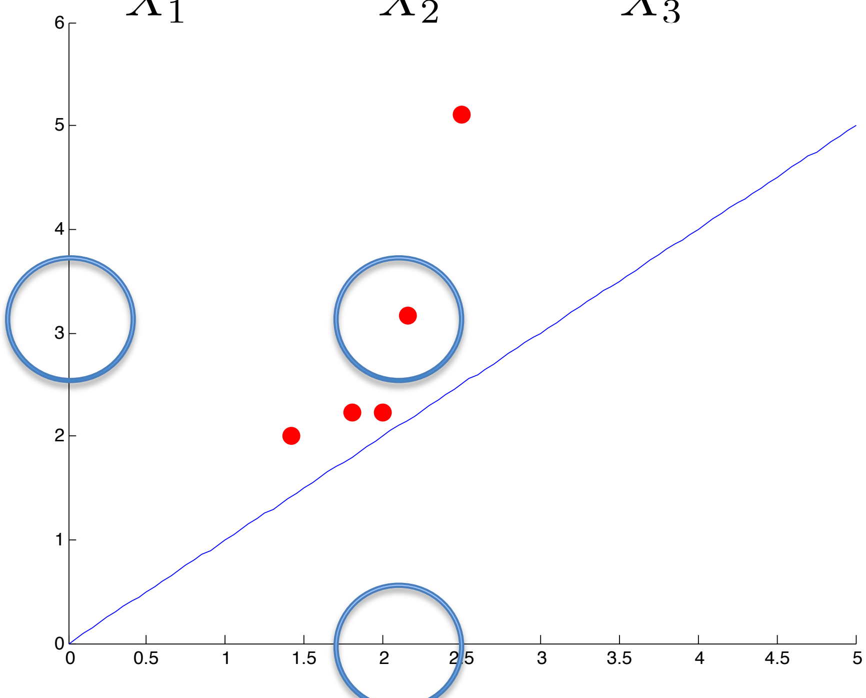
Persistence Diagram

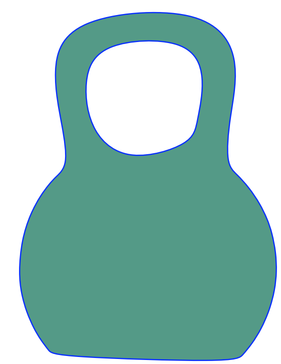
$$X_0 \subset X_1 \subset \cdots \subset X_n$$



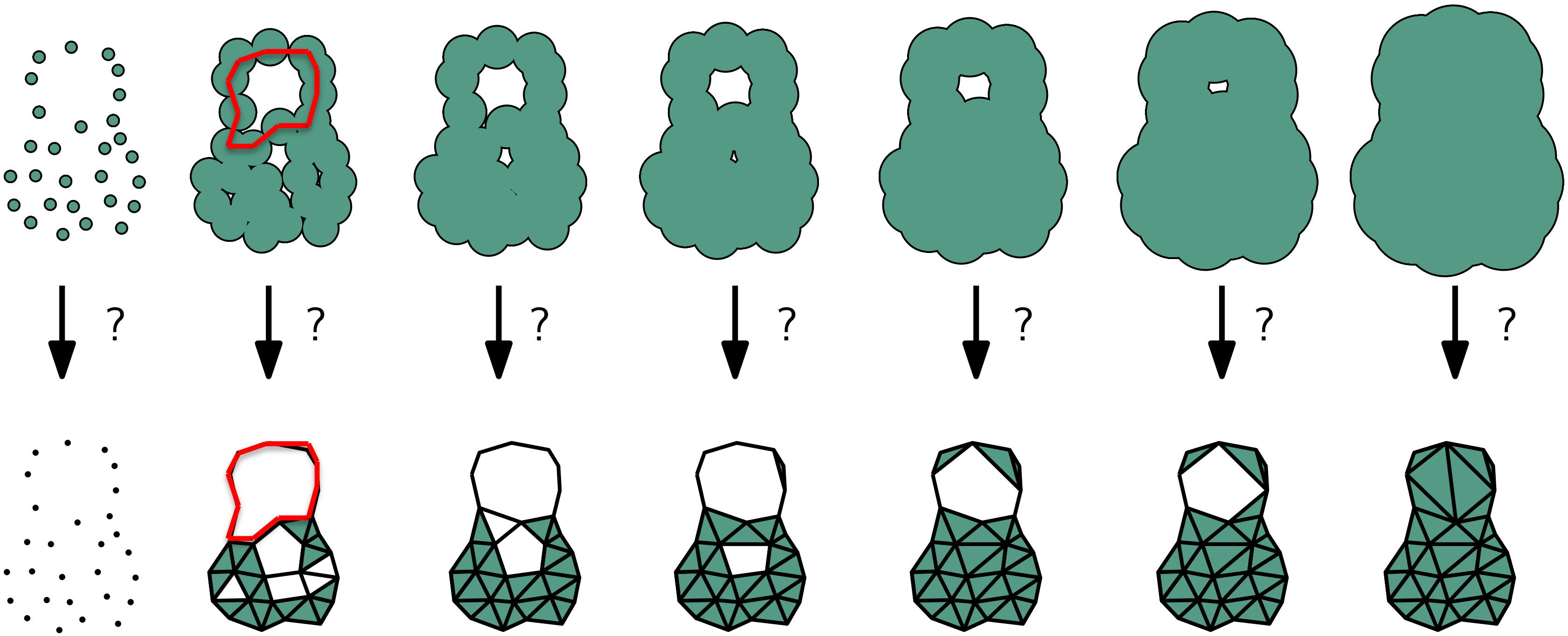
X_0 X_1 X_2 X_3 X_4 X_5 X_6

1D





To Compute: Cech Complex



Complexes

Cěch complex

$$< u_0, \cdots, u_k > \subset C_l(X)$$

iff

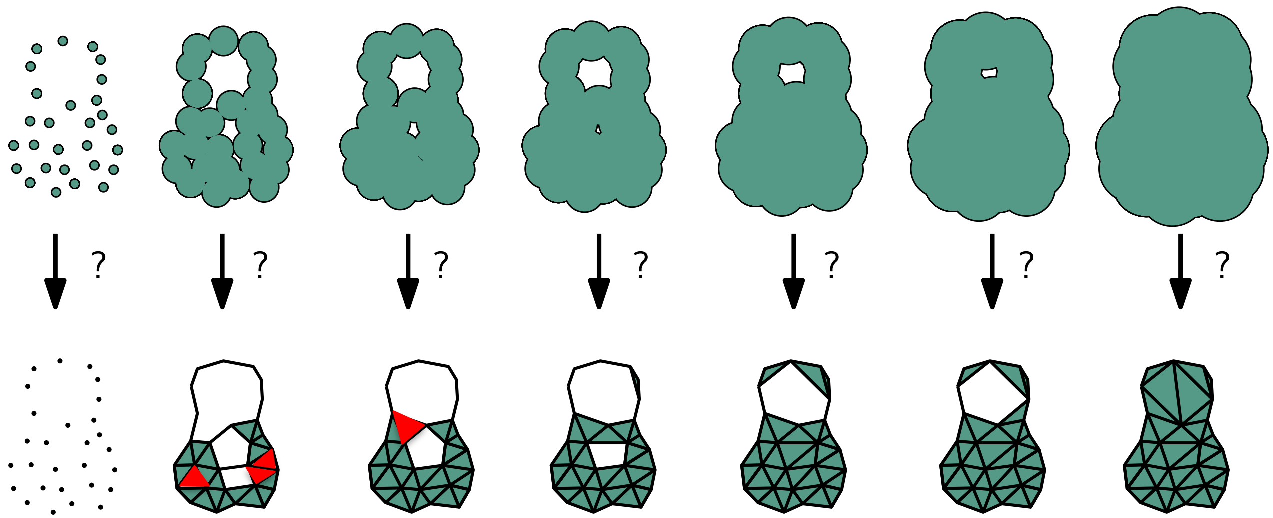
$$B_l(u_0) \cap \cdots \cap B_l(u_k) \neq \emptyset$$

Rips complex

$$< u_0, \cdots, u_k > \subset R_l(X)$$

iff

$$B_l(u_i) \cap B_l(u_j) \neq \emptyset \; \forall i, j$$



$$C_l(X) \subset R_l(X) \subset C_{\sqrt{2}l}(X)$$

$\mathbb{Z}/2$ Homology of a Simplicial Complex

$X =$ simplicity complex

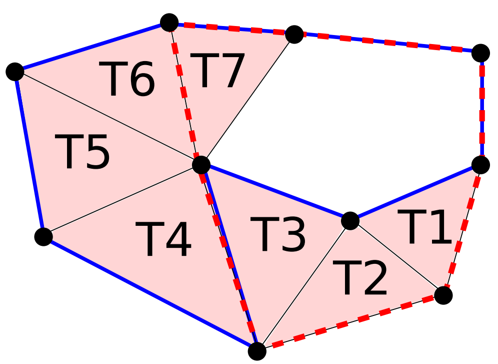
A p-chain is a formal sum $\sigma_1 + \cdots + \sigma_n$
where each σ_i is a p simplex

$C_p(X) =$ abelian group of all p-chains

$$\sigma = [v_0, v_1, \ldots, v_n]$$

$$\partial([v_0, v_1, \ldots, v_n]) = \sum_{k=1}^{n} [v_0, \ldots, v_{k-1}, v_{k+1}, \ldots, v_n]$$

$\partial^2 = 0$



C_0 rank 10

C_1 rank 17

C_2 rank 7

H_0 rank 1

H_1 rank 1

Group of Cycles

$Z_p(K) = ker(\partial_p)$

Group of Boundaries

$B_p(K) = im(\partial_{p+1})$

$$B_p(K) \subset Z_p(K)$$

$$H_p(K) = Z_p(K)/B_p(K)$$

Persistent Homology for Simplicial Complexes

Edelsbrunner
Letscher
Zomorodian

$$X_0 \subset X_1 \subset \cdots \subset X_n = X$$ Filtration

$X = \text{simplicity complex}$ $\quad X_i = \text{sub-complex}$

$$H_p(X_0) \to H_p(X_1) \to \cdots \to H_p(X_n)$$ Induced on Homology

$f_i^j : H_p(X_i) \to H_p(X_j)$ $\quad$ Field, usually $\mathbb{Z}/2$, coefficients

α born at k if

$\alpha \in H_p(X_k)$

$\alpha \notin \mathrm{im} f_{k-1}^k$

α dies at l if

$f_k^{l-1}(\alpha) \notin \mathrm{im} f_{k-1}^{l-1}$

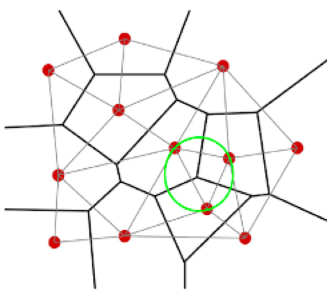
$f_k^l(\alpha) \in \mathrm{im} f_{k-1}^l$

Complexes for Point Clouds

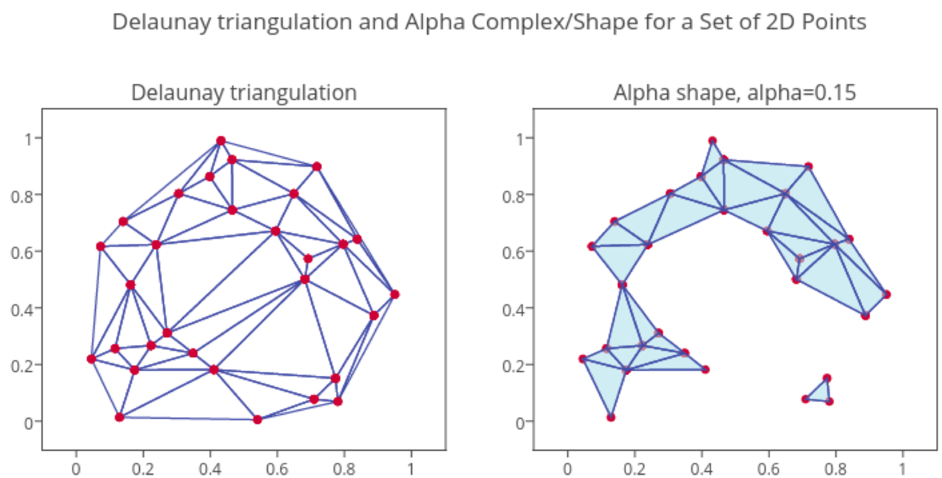
Voronoi, Delaunay and Alpha Complexes

$$U \subset X \subset \mathbb{R}^d$$

$$V(U) = \{x \in \mathbb{R}^d | d(x, u_i) = d(x, u_j)$$
$$\text{for all } u_i, u_j \in U \text{ and } d(x, v) > d(x, u_i) \text{ when } v \notin U\}$$



Delaunay Triangulation
Is the dual



$$R_u(r) = B_u(r) \cap V_u$$

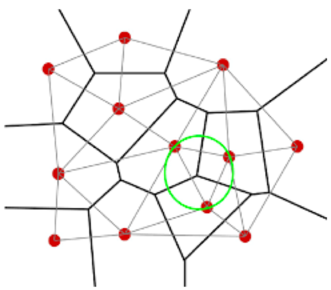
$$Alpha(r) = \{\sigma \subset U | \cap_u R_u(r) \text{ not empty}\}$$

Complexes for Point Clouds

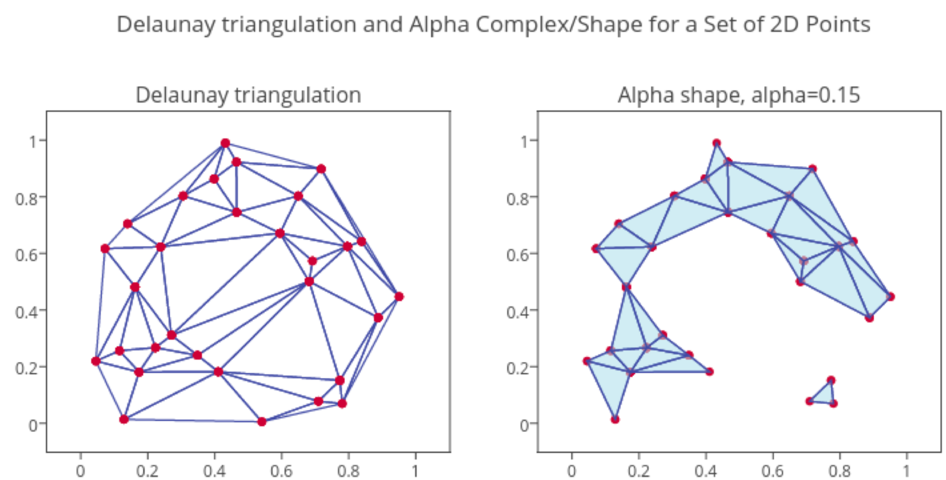
Voronoi, Delaunay and Alpha Complexes

$$U \subset X \subset \mathbb{R}^d$$

$$V(U) = \{x \in \mathbb{R}^d | d(x, u_i) = d(x, u_j) \text{ for all } u_i, u_j \in U \text{ and } d(x, v) > d(x, u_i) \text{ when } v \notin U\}$$



Delaunay Triangulation
Is the dual

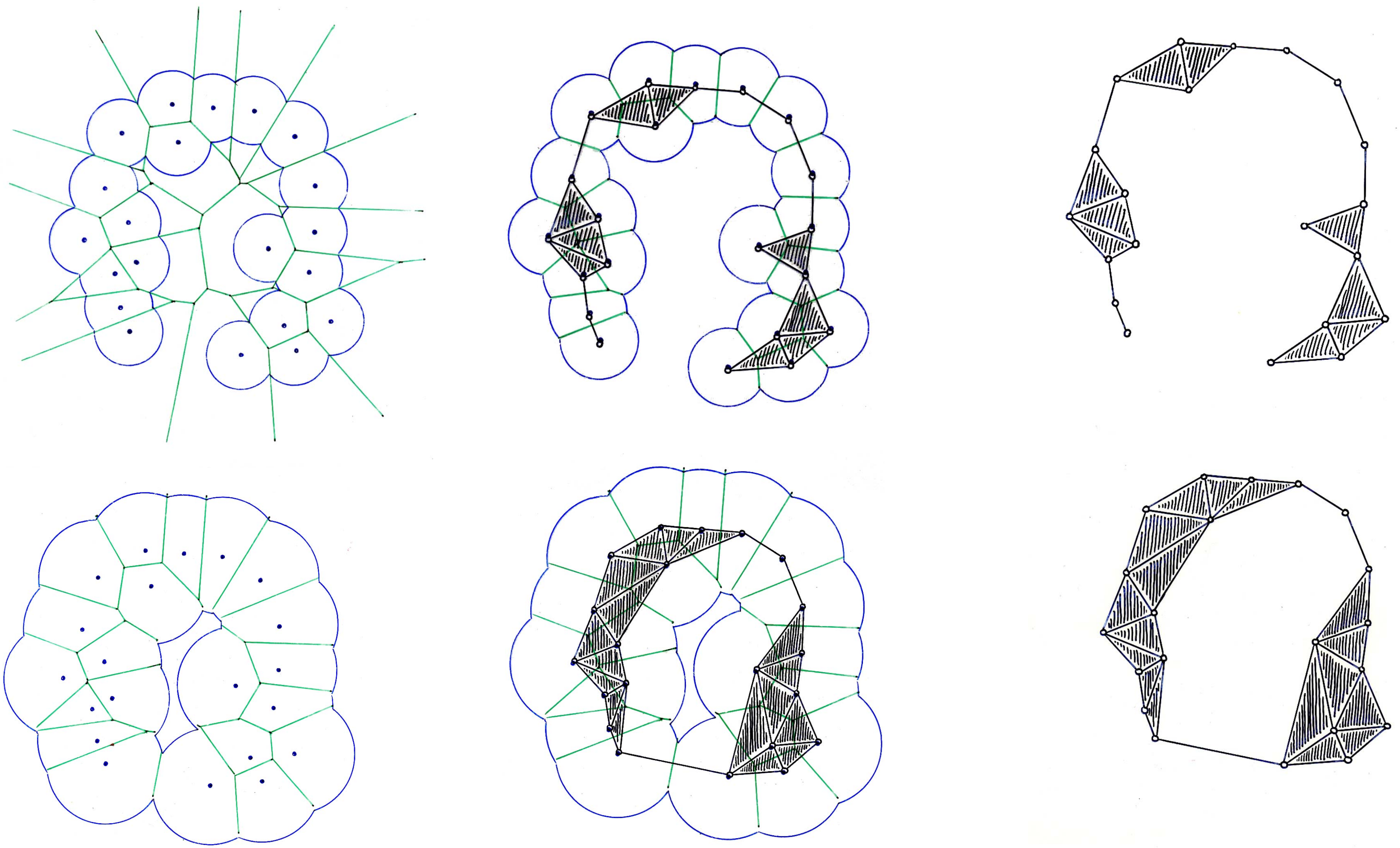


$$R_u(r) = B_u(r) \cap V_u$$

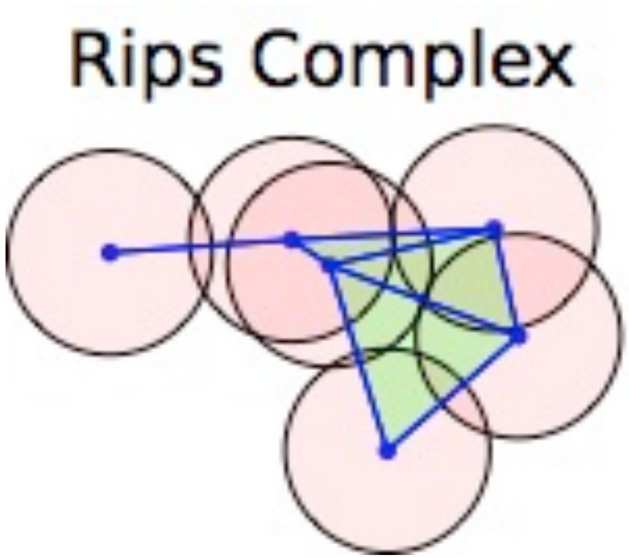
$$Alpha(r) = \{\sigma \subset U | \cap_u R_u(r) \text{ not empty}\}$$

Complexes for Point Clouds

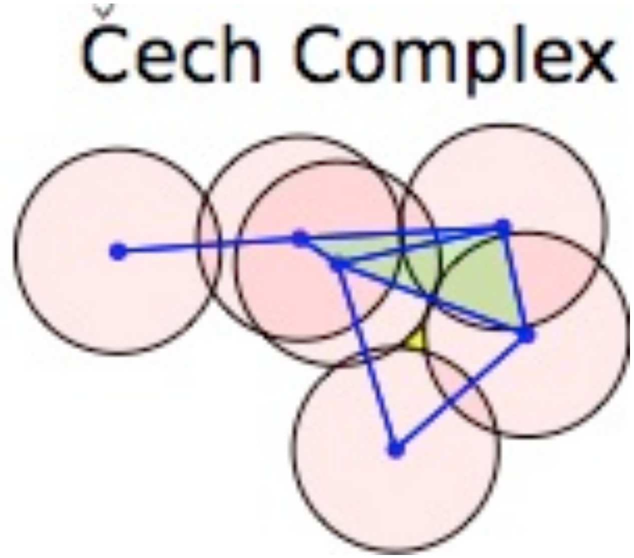
Alpha Complexes



Complexes for Point Clouds

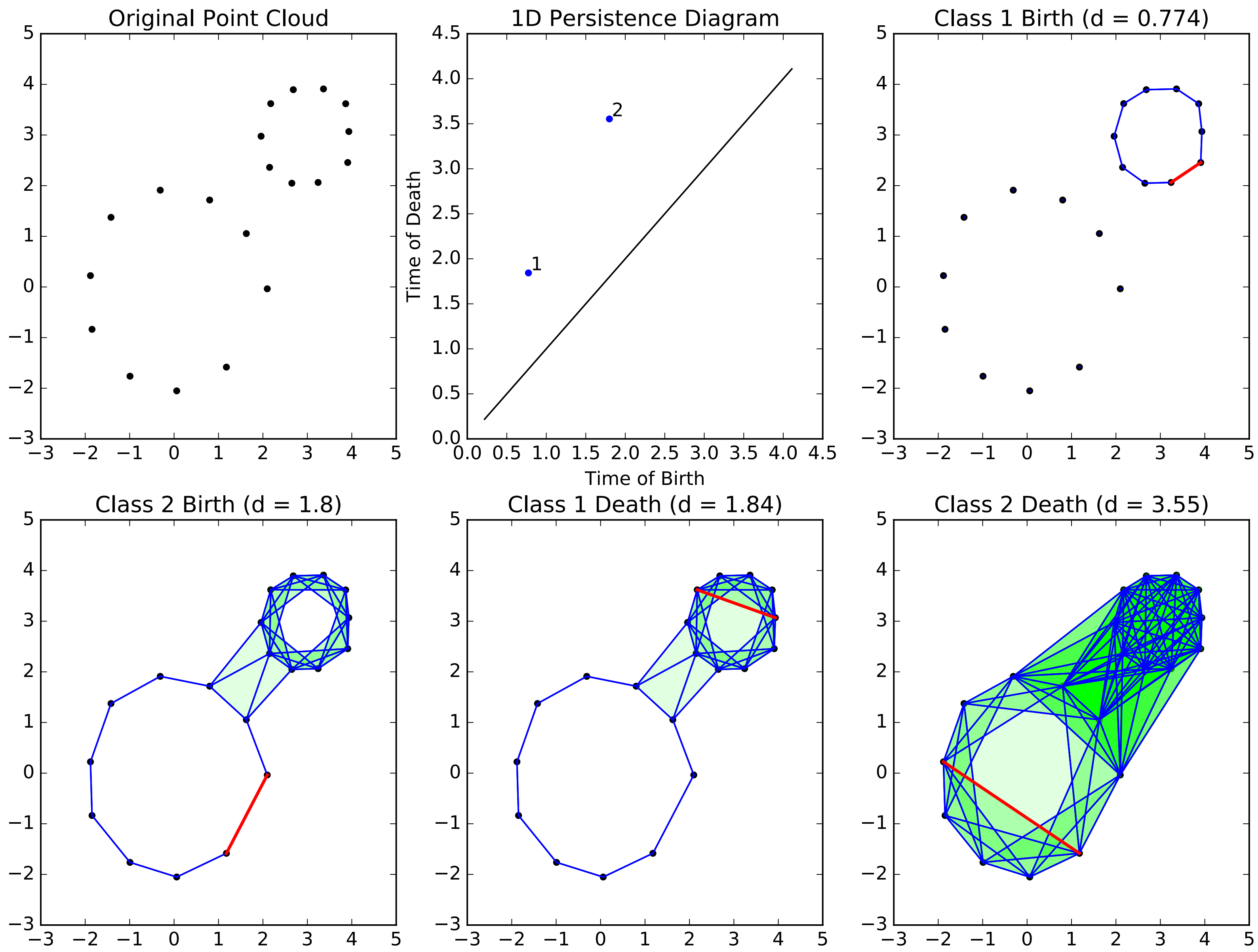


$$< u_0, \cdots, u_k > \subset R_l(X)$$
iff
$$B_l(u_i) \cap B_l(u_j) \neq \emptyset \ \forall i, j$$



$$< u_0, \cdots, u_k > \subset C_l(X)$$
iff
$$B_l(u_0) \cap \cdots \cap B_l(u_k) \neq \emptyset$$

Persistent Homology for Simplicial Complexes



To Calculate Persistent Homology

$$X_0 \subset X_1 \subset \cdots X_n = X$$

$$X_i = X_{i-1} \cup \sigma$$

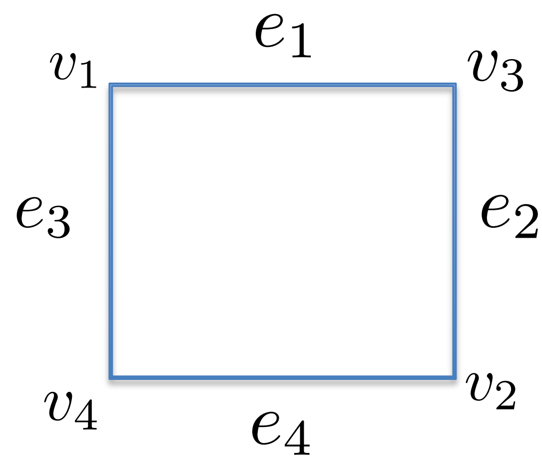
$\sigma = k$ cell

Birth: H_k increases rank by 1

Death: H_{k-1} decreases rank by 1

Matrix Reduction Algorithm

Example: Square



$$\begin{array}{c|cccc} & e_1 & e_2 & e_3 & e_4 \\ \hline v_1 & 1 & 0 & 1 & 0 \\ v_2 & 0 & 1 & 0 & 1 \\ v_3 & 1 & 1 & 0 & 0 \\ v_4 & 0 & 0 & 1 & 1 \end{array}$$

$e_1 + e_2$

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$$H_0 = \mathbb{Z}/2$$

$$H_1 = \mathbb{Z}/2$$

$e_3 + e_4$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$e_1 + e_2 + e_3 + e_4$

$$\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Pairs

(v_2, e_2)

(v_3, e_1)

(v_4, e_3)

Matrix Reduction Algorithm

$C_p \to C_{p-1}$

Columns correspond to p simplices $\sigma_1, \ldots \sigma_n$

Rows correspond to p-1 simplices $\tau_1, \ldots \tau_m$

Start with the Incidence Matrix $M_{i,j} = 1$ iff $\tau_i < \sigma_j$

Find lowest 1 in column 1 - Create pair

Column j - find lowest 1

- If same as lowest 1 of earlier column, add that column to column j
- Continue until new lowest 1 found, or column is 0

Pairs are (row, column) indices of lowest 1s

p-cycles correspond to all-0 columns

Sublevel Set Filtration

X metric space
$f : X \to \mathbb{R}$

Sublevel set $L_f(t) = \{x \in X | f(x) \leq t\}$

Tame

Finite number of homological critical values $t_1 < \cdots < t_n$
$H_k(f^{-1}((-\infty, a]$ finite dimensional for all a, k.

Morse functions on manifolds are tame.

Take $s_0 < t_1 < s_1 < \cdots < t_n < s_n$

Set $X_i = L_f(s_i)$

$X_0 \subset X_1 \subset \cdots X_n = X$

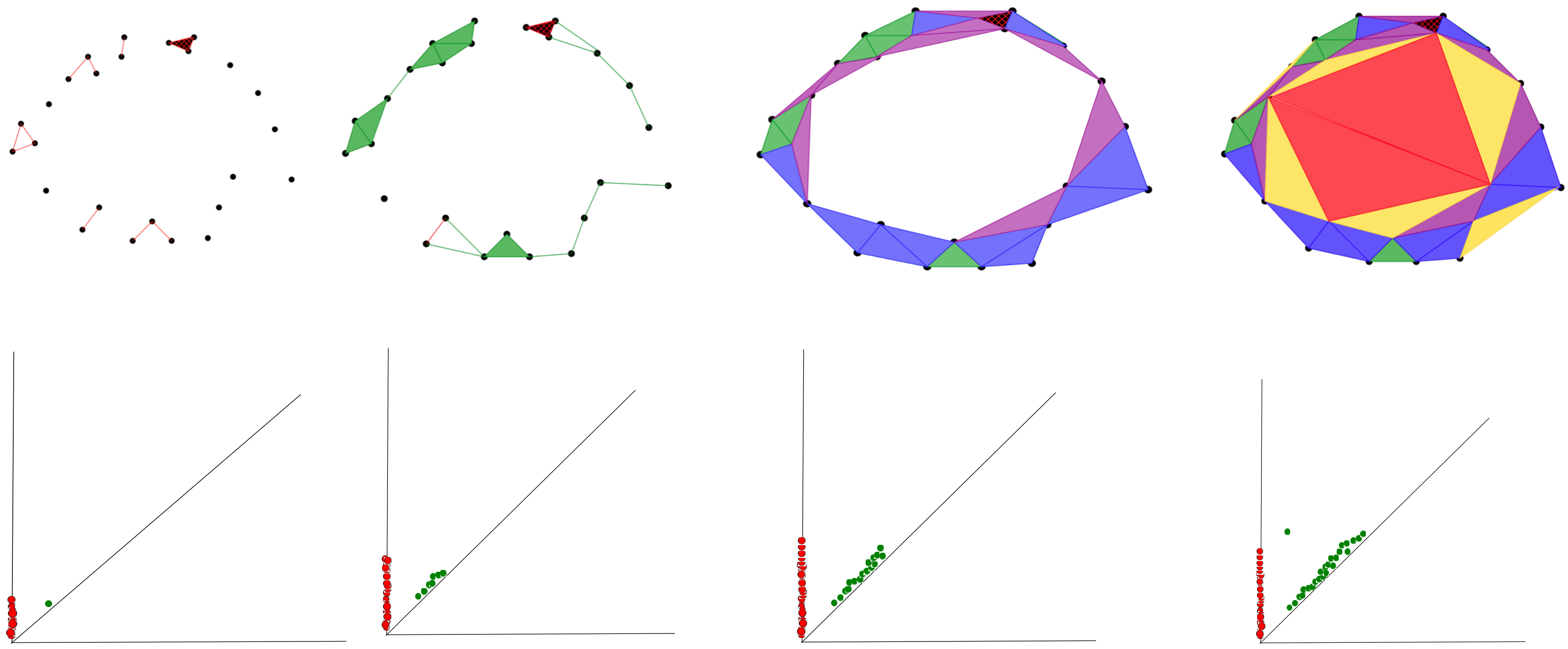
Sublevel Set Filtration

Key example:

$U \subset \mathbb{R}^D$ a point cloud
$d_U : \mathbb{R}^D \to \mathbb{R}$
distance to closest point of U

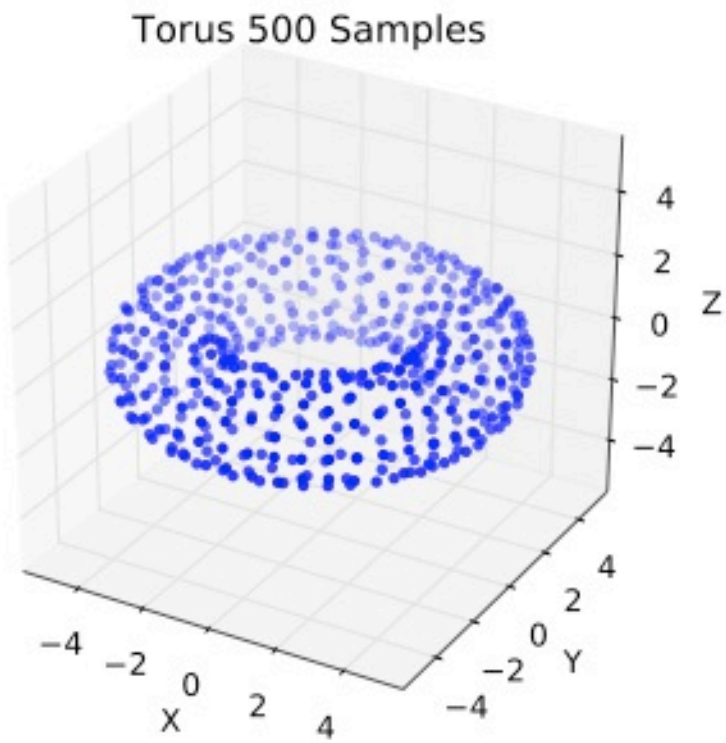
$L_{d_U}(t) = \cup_{u \in U} B_t(u)$

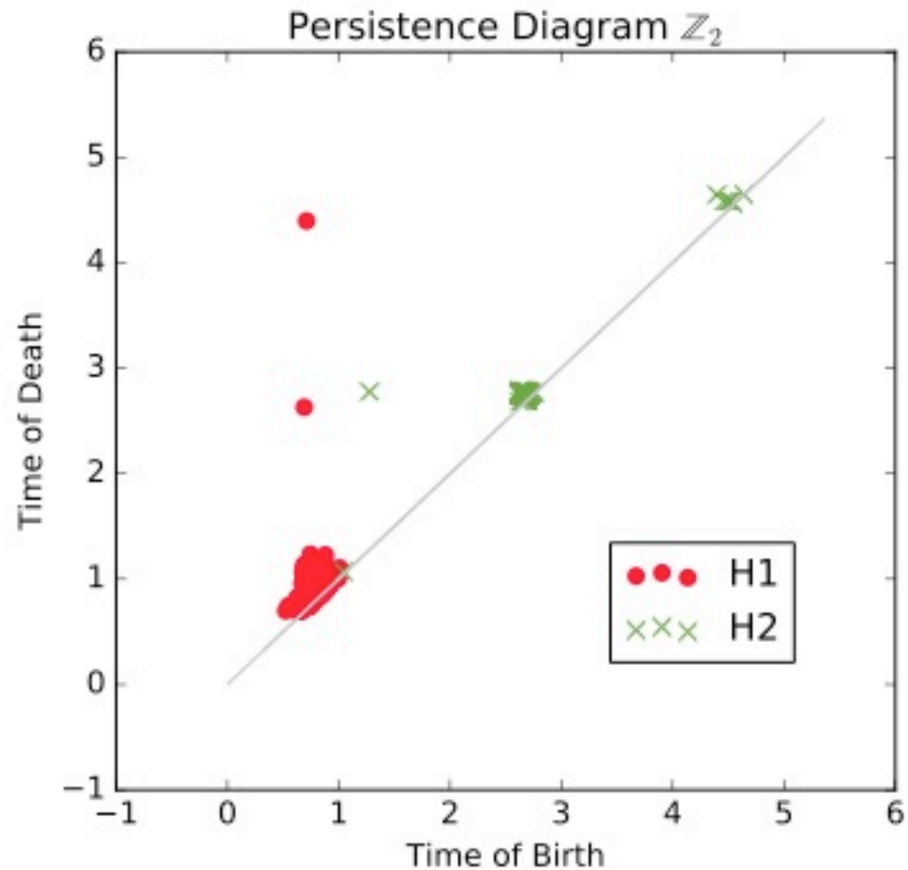
Persistence and Persistence Diagrams



Red = 0D
Green = 1D

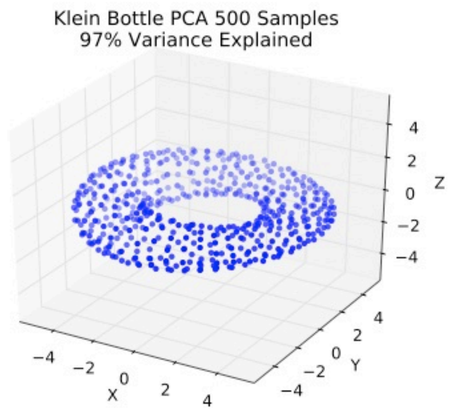
Persistence and Persistence Diagrams

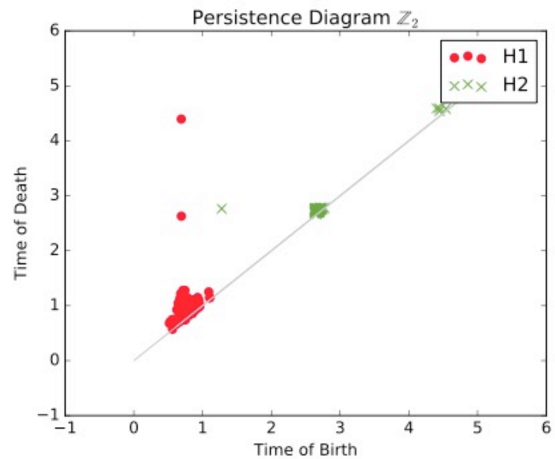


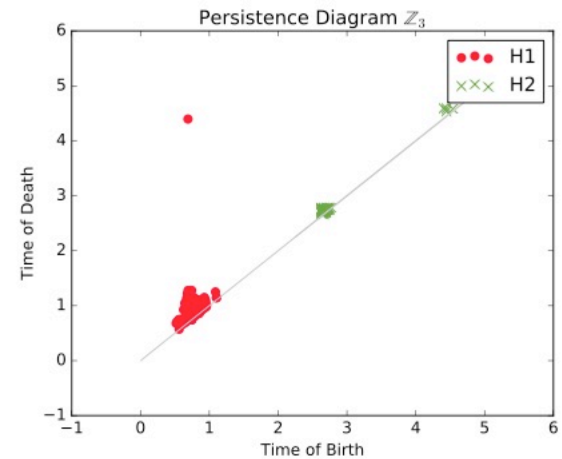


Persistence and Persistence Diagrams

Coefficients Matter



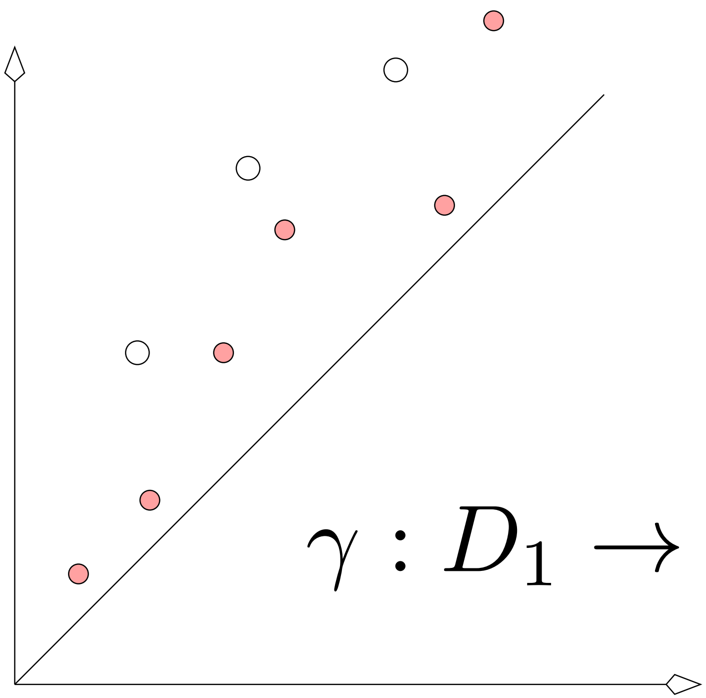




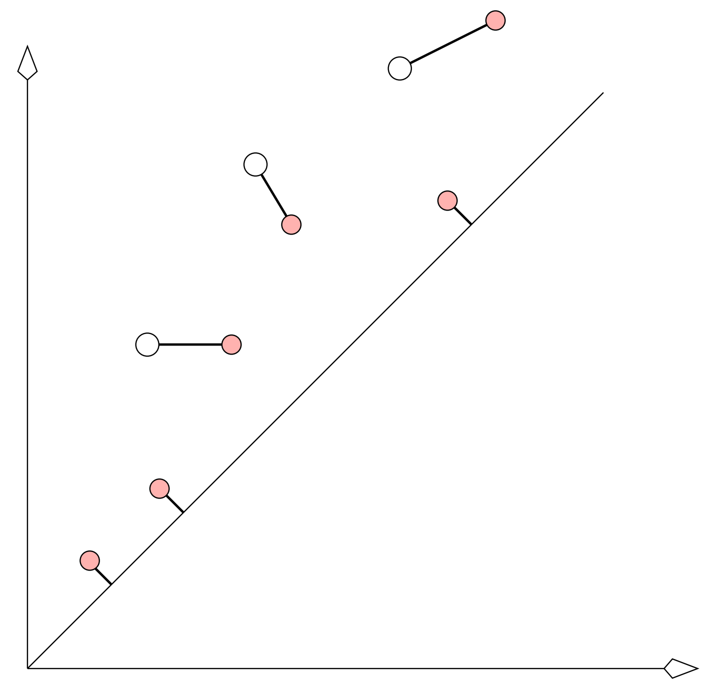
$$H_*(K;\mathbb{Z}/2) = H_*(T^2;\mathbb{Z}/2)$$

$$H_*(K;\mathbb{Z}/3) = H_*(S^1;\mathbb{Z}/3)$$

Wasserstein Distance



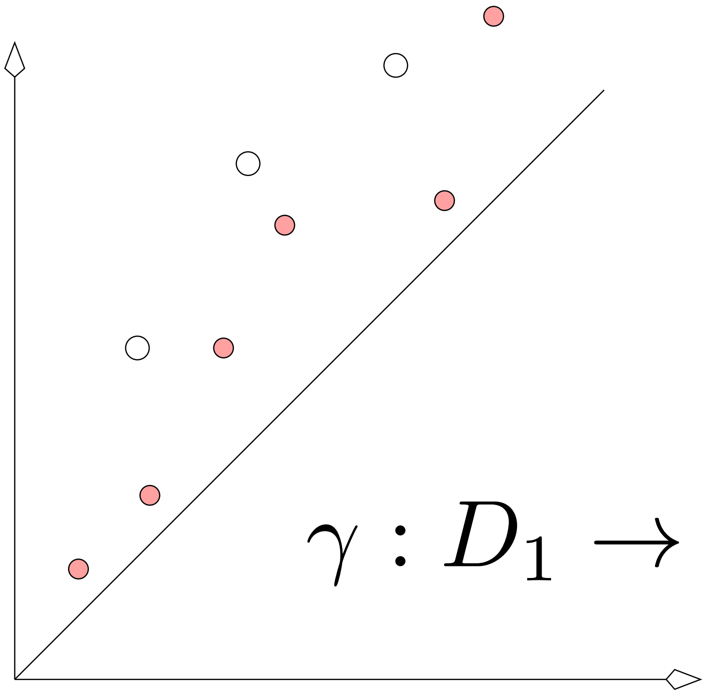
$$\gamma : D_1 \to D_2$$



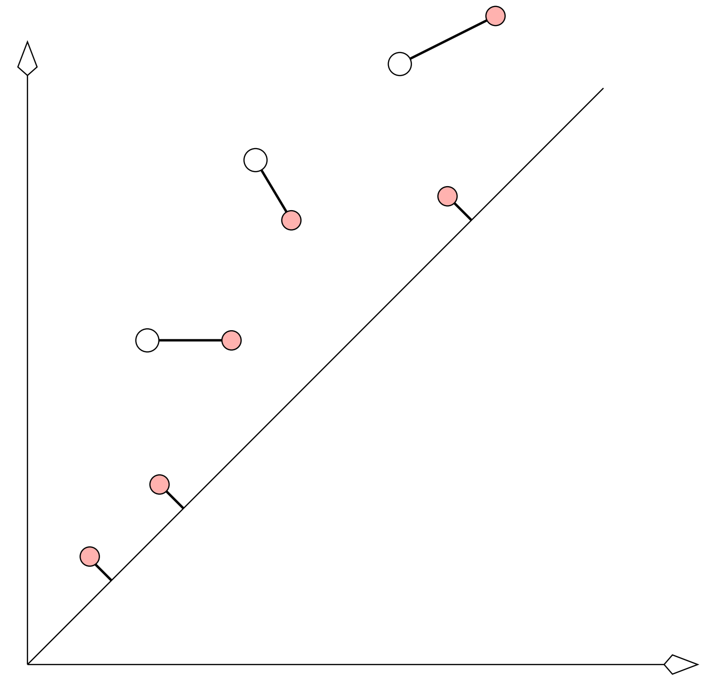
Matching

$$W_p(D_1, D_2) = min_\gamma (\sum_{x \in D_1} ||x - \gamma(x)||_\infty^p)^{1/p}$$

Bottleneck Distance
L_p Distance



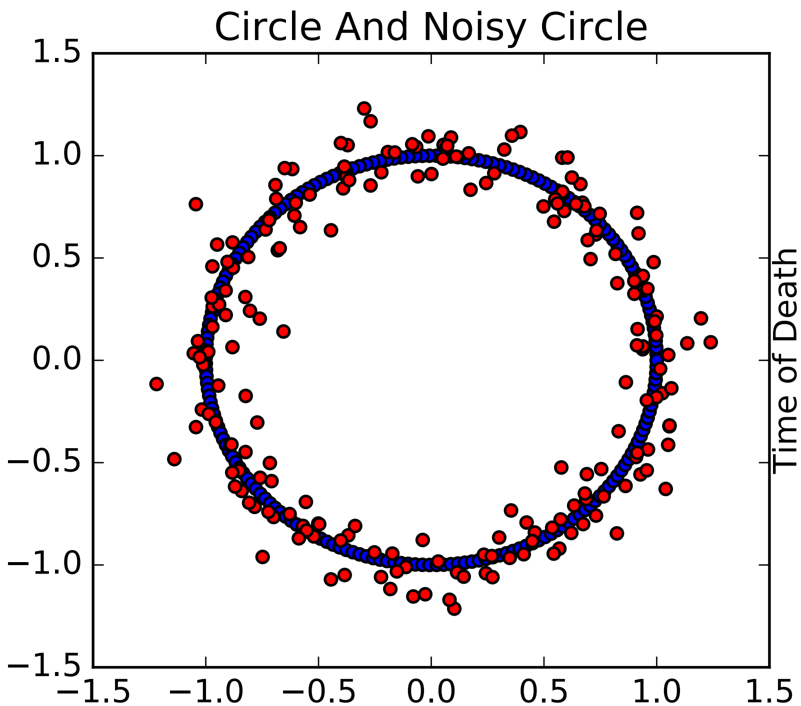
$$\gamma : D_1 \rightarrow D_2$$

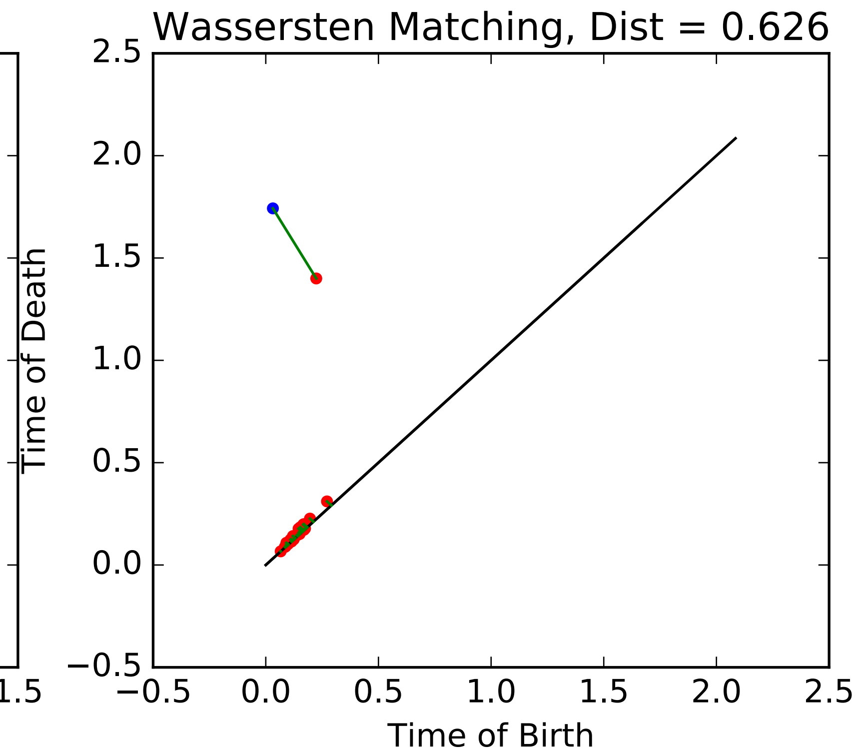


$$B(D_1, D_2) = \min_{\gamma}\big(\max_{x \in D_1} ||x - \gamma(x)||_{\infty} \big)$$

$$L_p(D_1, D_2) = \min_{\gamma}\Big(\sum_{x \in D_1} ||x - \gamma(x)||^p \Big)^{\frac{1}{p}}$$

Wasserstein Distance



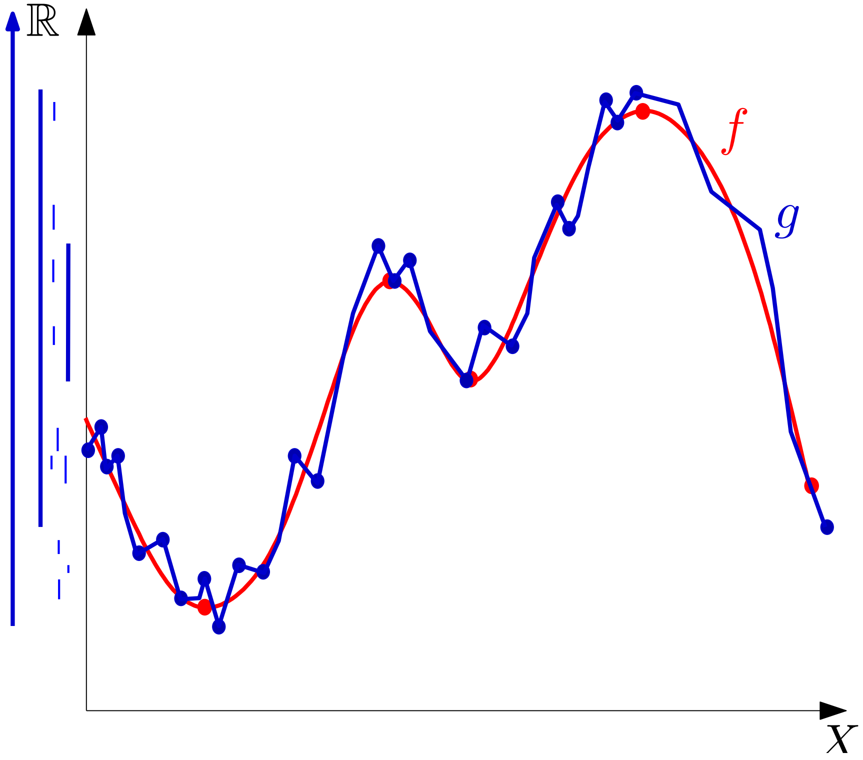


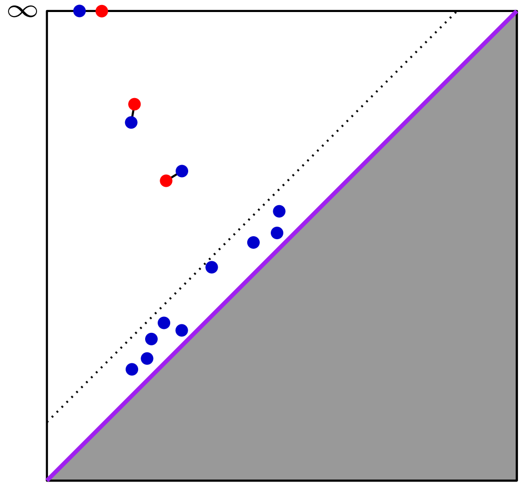
Both cases using sublevel sets of distance
from object to compute peersistence

Stability

Theorem: For any pfd functions $f, g : X \to \mathbb{R}$,

$$\mathrm{d}_b(\mathrm{Dg}\, f, \mathrm{Dg}\, g) \leq \|f - g\|_\infty$$





Stability

Cohen-Steiner
Edelsbrunner
H.

Cohen-Steiner
Edelsbrunner
H.
Mleyko

Early Stability Theorems

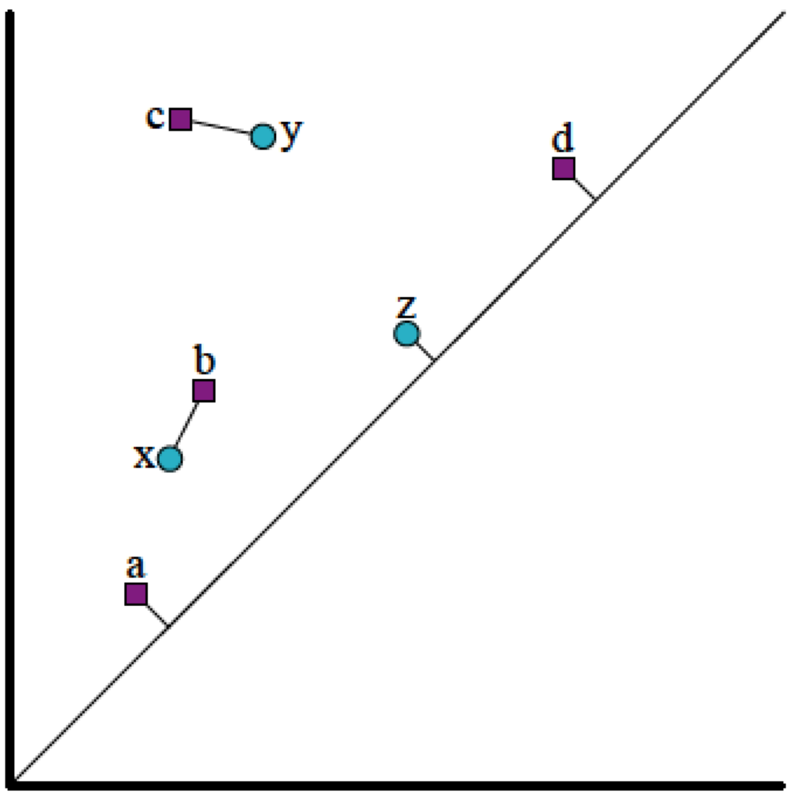
$f, g : X \to \mathbb{R}$ $\quad$ X metric space

Tame $\quad$ $d_B(P(f), P(g)) \leq ||f - g||_\infty$

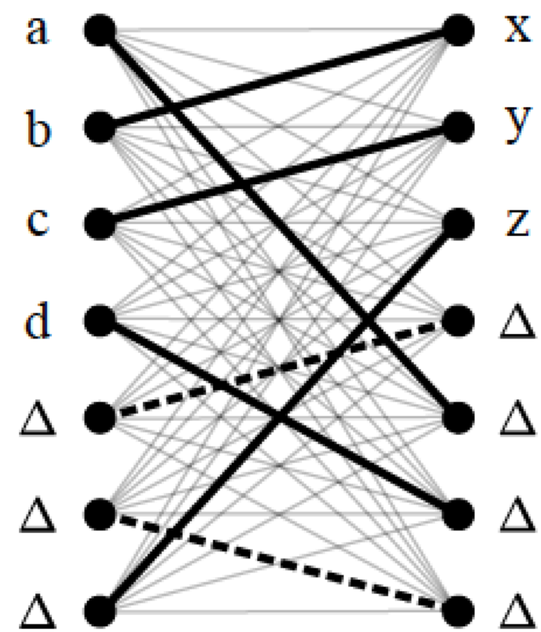
Tame, Lipschitz $\quad$ $W_p(f, g) \leq C||f - g||_\infty^{1-k/p}$

C, k constants that depend only on
X and Lipschitz constants of f, g

Bottleneck and Wasserstein Algorithms



(a)



(b)

Bottleneck and Wasserstein Algorithms

Have two sets U, V and cost function

$$c : U \times V \to \mathbb{R}_{\geq 0}$$

X, Y

Persistence Diagrams

$U = X \cup Y_0$

$V = Y \cup X_0$

$G = (U \cup V, U \times V)$ Matching of G is a subset of $U \times V$ of vertex disjoint edges

$G(\epsilon)$ = subgraph where all edges have cost $\leq \epsilon$

Bottleneck distance is smallest $\epsilon \geq 0$ s.t. $G(\epsilon)$ has a perfect matching

$Wasserstein$

$W_q^q =$ total cost of a matching of G with cost function c^q

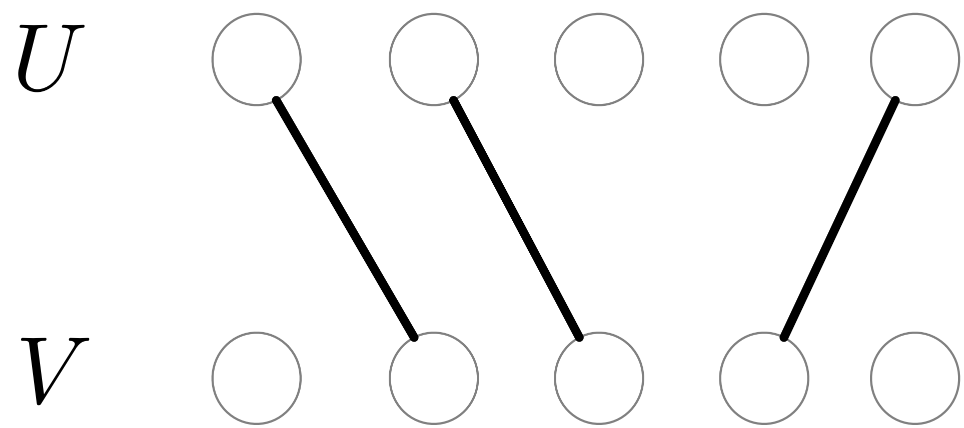
Bottleneck and Wasserstein Algorithms

Algorithm Description to find Bottleneck:

Find the smallest ϵ s.t. $G(\epsilon)$ has a match by binary search on ϵ

For each ϵ run the following:

- Take a partial matching M between U and V



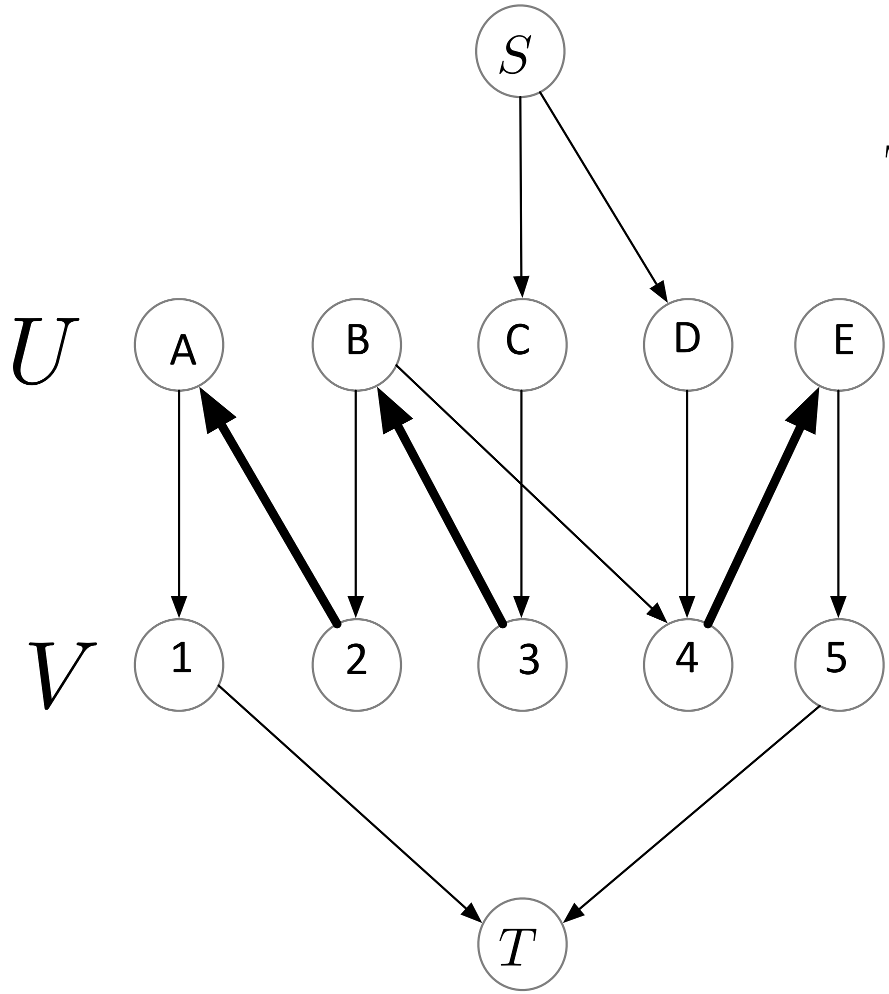
Bottleneck and Wasserstein Algorithms

Edges in M go up

Edges in $G(\epsilon) - M$ go down

S connected to unmatched vertices of U

T connected to unmatched vertices of V



Augmenting Path P from S to T

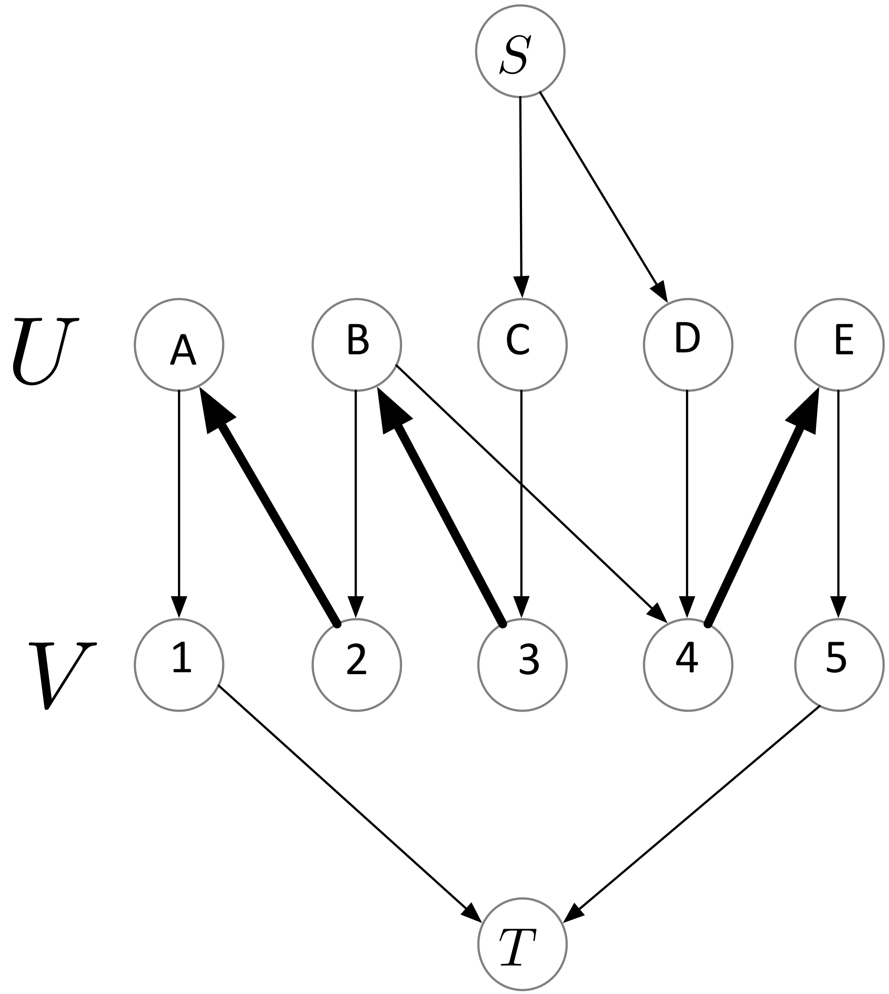
e.g. S, C, 3, B, 4, E, 5, T

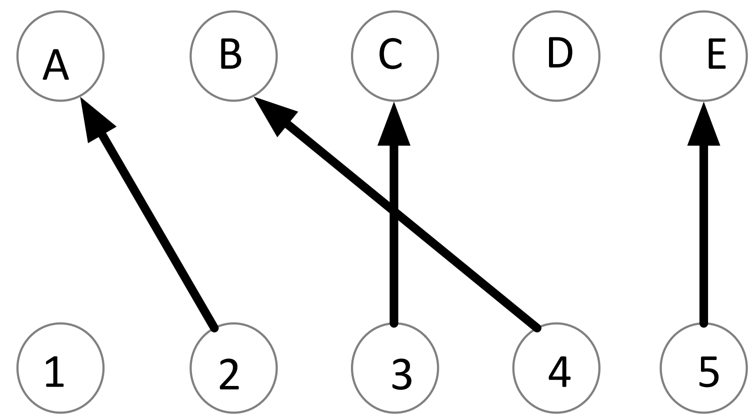
Has 2k+1 edges, k down, k-1 up between U and V

Switch the edges in P

Bottleneck and Wasserstein Algorithms

e.g. S, C, 3, B, 4, E, 5, T





Terminates at largest partial matching

Interleaving Distance

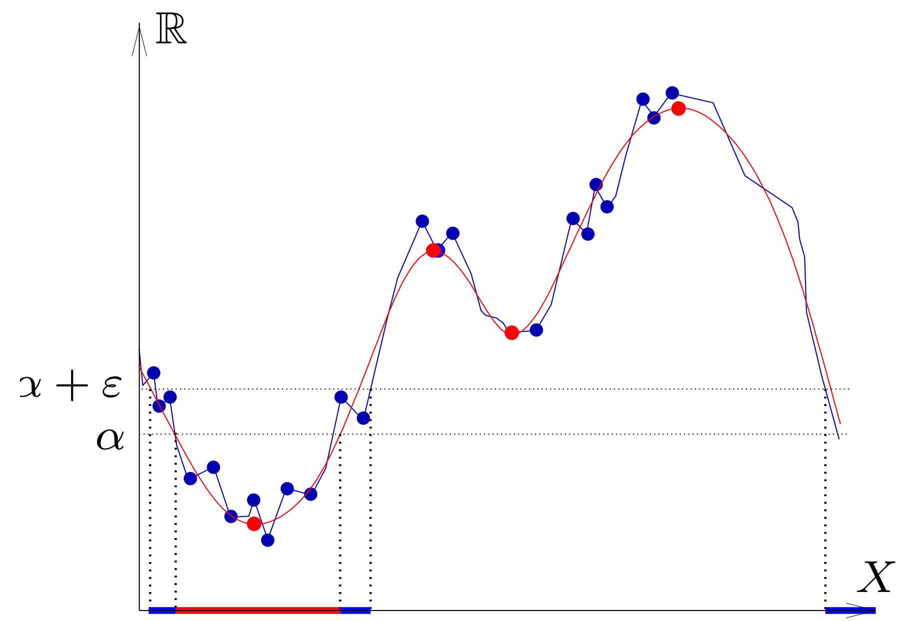
Let $f, g : X \to \mathbb{R}$ be pfd, and let $\varepsilon = \|f - g\|_\infty$.

$$F_t := f^{-1}((-\infty, t])$$

$$G_t := g^{-1}((-\infty, t])$$

- Key observation: $\{F_t\}_t$ and $\{G_t\}_t$ are ε-**interleaved** w.r.t. inclusion:

$$\cdots \subseteq \color{blue}{F_0} \subseteq \color{red}{G_\varepsilon} \subseteq \color{blue}{F_{2\varepsilon}} \subseteq \cdots \subseteq \color{blue}{F_{2n\varepsilon}} \subseteq \color{red}{G_{(2n+1)\varepsilon}} \subseteq \cdots$$



Slide thanks to Steve Oudut

Interleaving Distance

$$[\varepsilon] : M \longmapsto M[\varepsilon] \qquad \text{s.t.} \qquad M[\varepsilon](t) = M(t+\varepsilon)$$

ε-**interleaving:** morphisms $\phi : M \Rightarrow N[\varepsilon]$ and $\psi : N \Rightarrow M[\varepsilon]$ s.t.

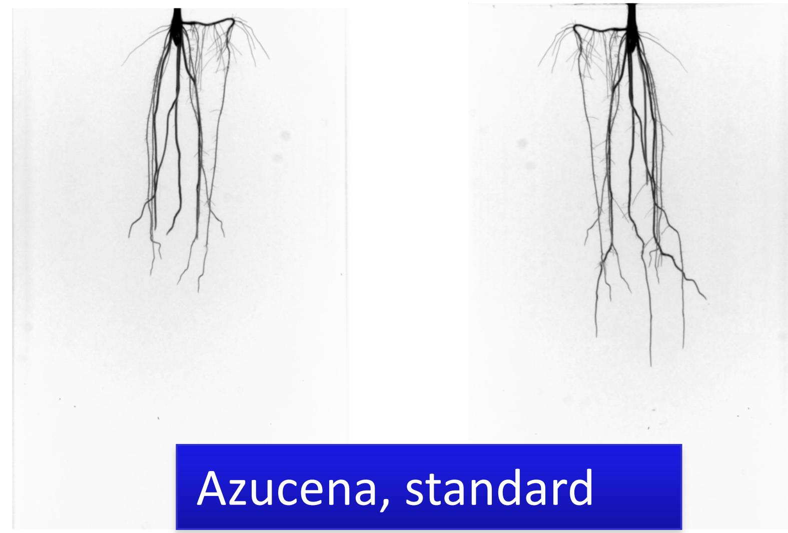
$$\begin{tikzcd}
{\color{blue}M} \arrow[rr, Rightarrow, "\xi"] \arrow[drr, Rightarrow, "\phi"'] & & {\color{blue}M[\varepsilon]} \arrow[rr, Rightarrow, "\xi[\varepsilon]"] \arrow[drr, Rightarrow, "\phi[\varepsilon]"'] & & {\color{blue}M[2\varepsilon]} \\
{\color{red}N} \arrow[rr, Rightarrow] \arrow[urr, Rightarrow, "\psi"] & & {\color{red}N[\varepsilon]} \arrow[rr, Rightarrow] \arrow[urr, Rightarrow, "\psi[\varepsilon]"] & & {\color{red}N[2\varepsilon]}
\end{tikzcd}$$

- **interleaving distance:** $\mathrm{d}_i(M,N) := \inf\{\varepsilon \mid M, N\ \varepsilon\text{-interleaved}\}$

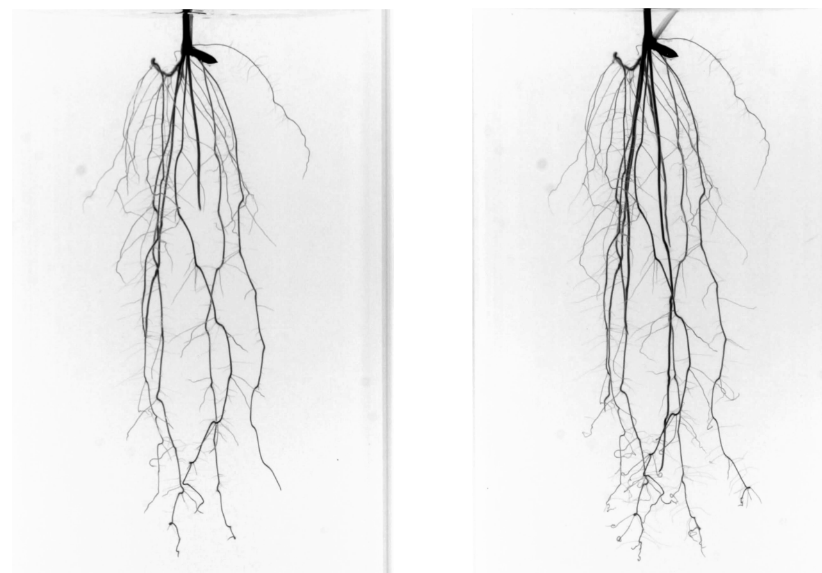
Slide thanks to Steve Oudut

Do Statistics on/with Persistence Diagrams?

- Define means and variances of PDs
- Talk about mean shape and variance in shape of different objects



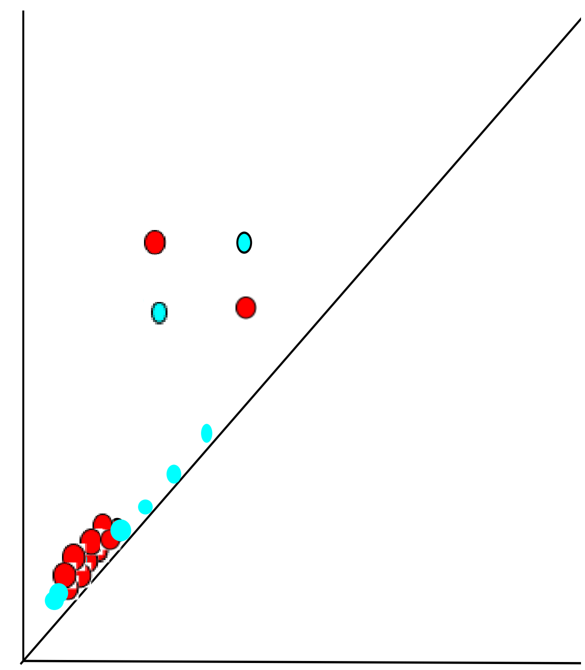
Azucena, standard



Azucena, -Nitrogen

Do Statistics on/with Persistence Diagrams?

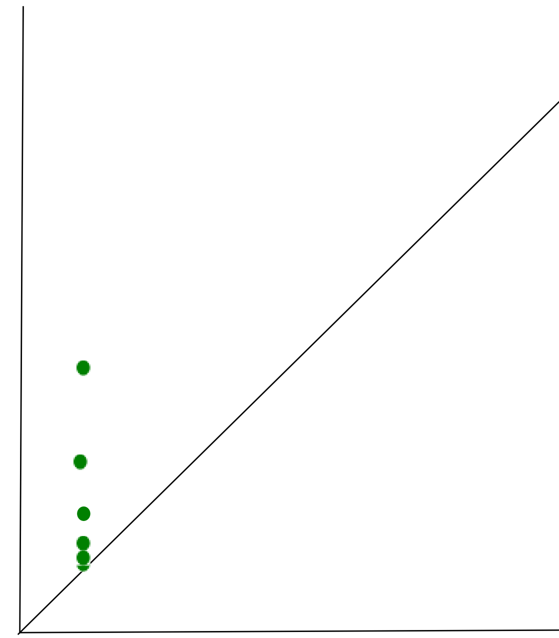
- Can’t naturally add diagrams
- Suggests Frechet Mean
- Need Polish space in which to work



Frechet
Mean

Is the Space of Diagrams Complete ?

NO



$$p_n = (1, 1 + 1/2^n)$$

$$d_n = d_{n-1} \cup p_n$$

d_n Cauchy – match later points to diagonal

$d_n \rightarrow d$ diagram with infinite number of points but total persistence that's finite.

Space of Persistence Diagrams

Yuriy Mileyko
Sayan Mukherjee
H.

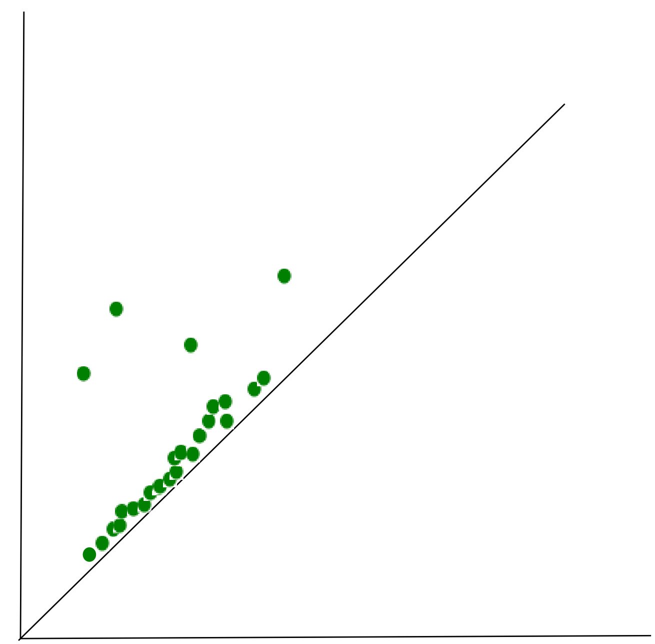
$$Q = (x, y), \quad pers(Q) = y - x$$

$$pers(d) = \sum_{Q \in d} pers(Q)^p$$

allow d to be infinite (but countable)

$$D_p = \{d : pers(d) < \infty\}$$

with metric W_p



$\boxed{D_p}$ is Polish – Complete and Separable

Frechet Function

$$F(d) = \int_{D_p} W_p(d,e)\,d\mu(e)$$

μ compact support

$$\mathrm{FVar}(\mu) = \inf_{d \in D_p} F(d)$$

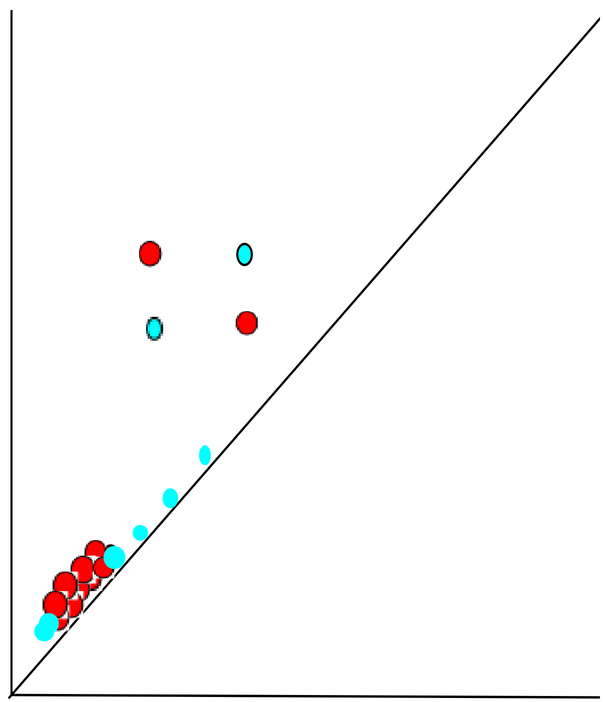
$$E(\mu) = \{d \in D_p : F(d) = \mathrm{FVAR}(\mu)\}$$

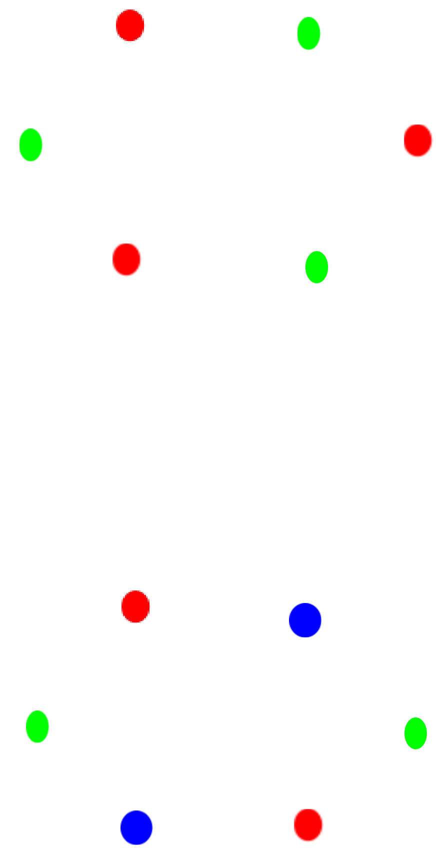
$$\boxed{\textbf{Theorem: } E(\mu) \neq \phi.}$$

But it's not necessarily a unique diagram

Non- Uniqueness of the Frechet Mean

For two diagrams, points of the Mean should be the barycenter of the edge between matched points.





Non- Uniqueness of the Frechet Mean

Ways to fix this:

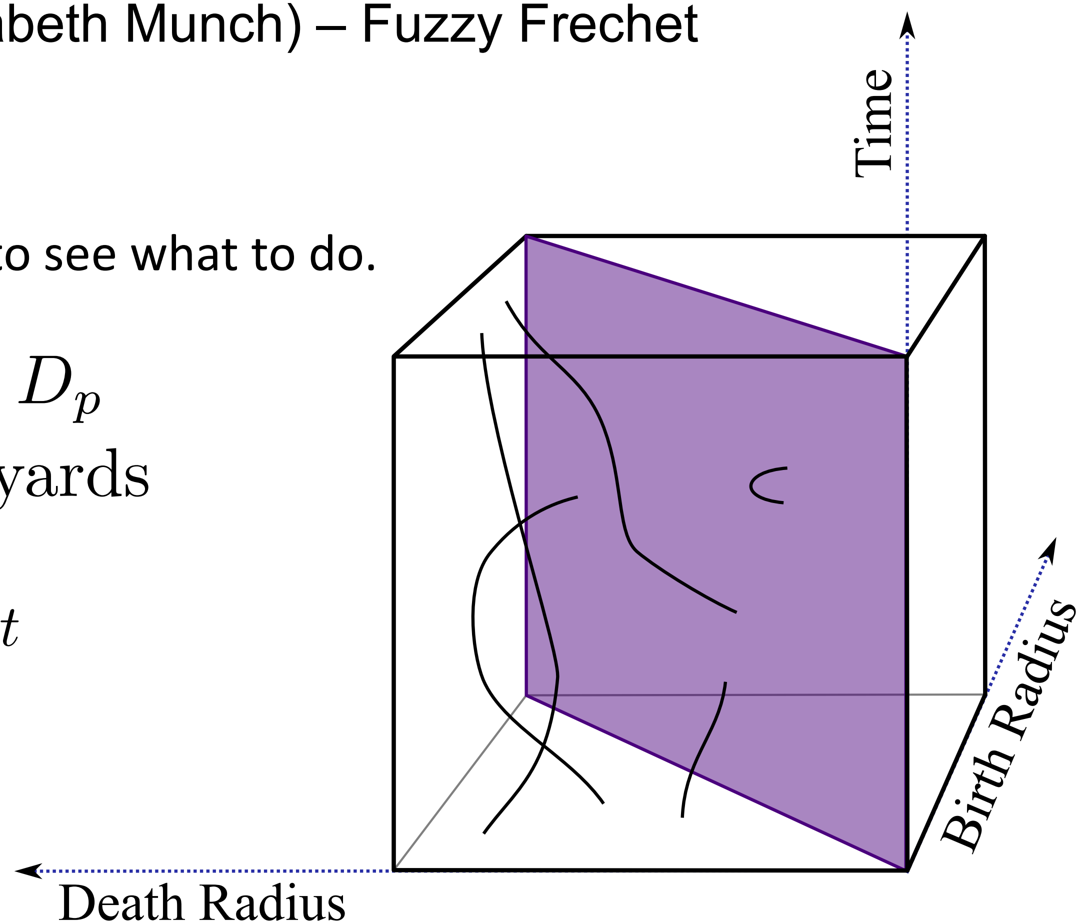
1. Drop the requirement that the mean should be realized as diagram (Peter Bubenik) or
2. Redefine the metric (Elizabeth Munch) – Fuzzy Frechet Mean

Study the time varying case to see what to do.

$$\begin{aligned} P(D_p) &= \text{Path Space of } D_p \\ &= \text{Space of Vineyards} \end{aligned}$$

$$V_p(v,w) = \int_0^1 W_p(v(t), w(t)dt$$

$(P(D_p), V_p)$ is Polish



Frechet Means for Time Varying Data