ZUCMAP

An Introduction to Persistent Homology

Siddharth Setlur

ETH Zürich

April 4, 2022

ZUCMAP

zucmap.ethz.ch

Every Monday – 12:15 HG G3

Next Week – Tim De Ryck

Physics Informed Neural Networks

Goal: Introduce persistent homology in an intuitive manner & highlight its power by discussing a recent application.

Outline

1) Motivation

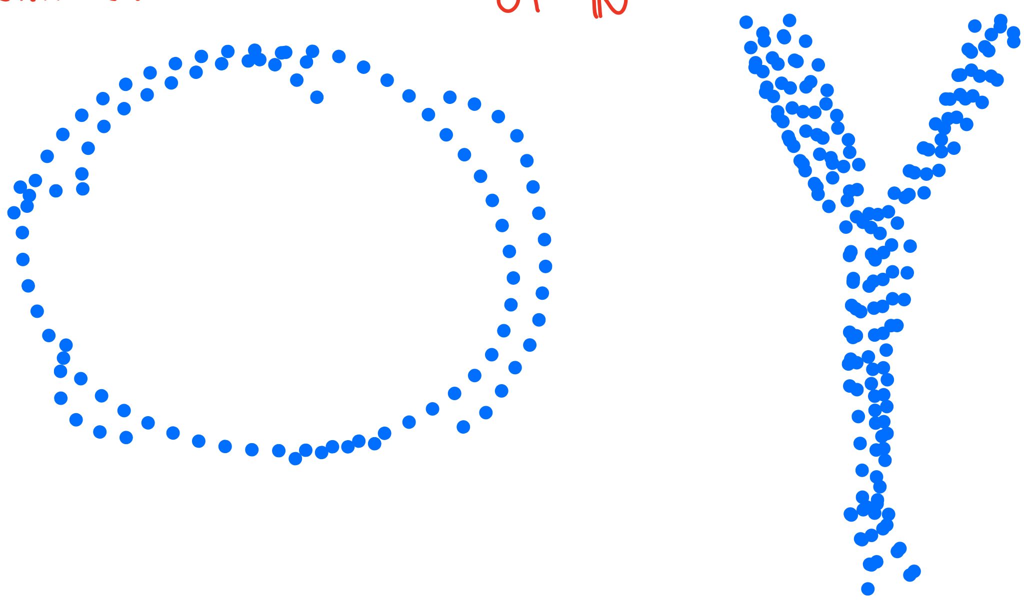
2) Introducing Persistent Homology with an example

3) A glance at a recent application

4) The theory underpinning persistent homology

Data has shape:

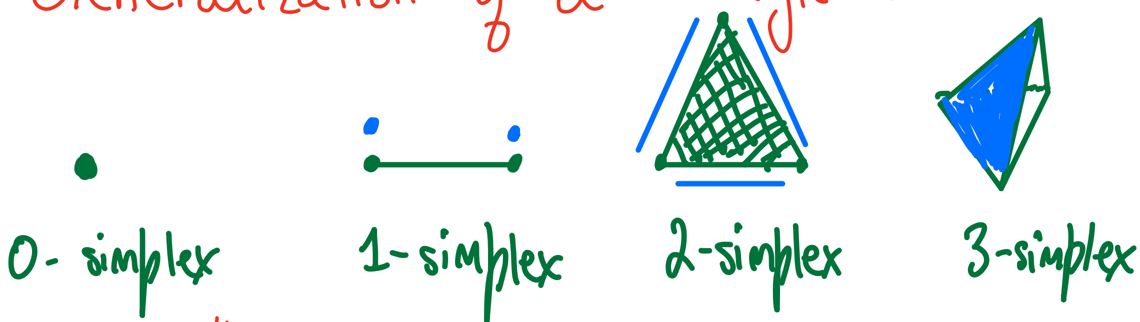
Point cloud data - finite discrete subset of $\mathbb{R}^n$



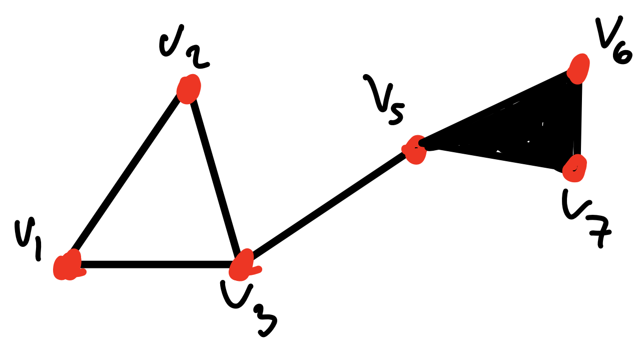
In low dimensions we can "see" this, but data is usually high dimensional, for example patient data to study diseases or in material sciences.

Simplices & Simplicial Complexes

"Generalization" of a triangle to $n \geq 0$ dims

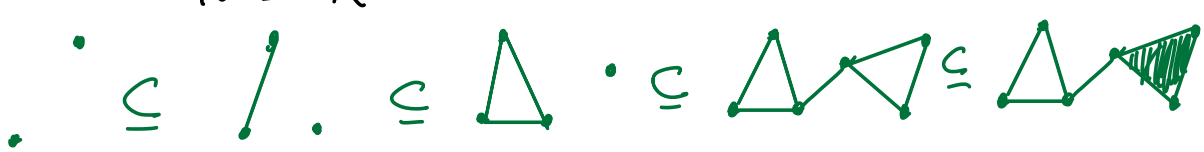


"Sticking" simplices together forms simplicial complexes



A filtration of a simplicial complex K is a sequence of subcomplexes

$$K^0 \subseteq K^1 \subseteq K^2 \subseteq \cdots \subseteq K^n = K$$



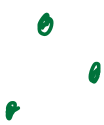
Simplicial Homology

1) Assigns an algebraic invariant to simplicial complexes

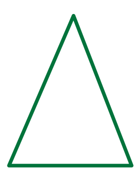
$$\begin{aligned} H_n : \text{SimpComp} &\longrightarrow \text{Ab} \\ K &\longmapsto H_n(K) \end{aligned}, \quad n \in \mathbb{N}_{\geq 0}$$

We work over $\mathbb{Z}_2$, so $H_n(K)$ is a vector space.

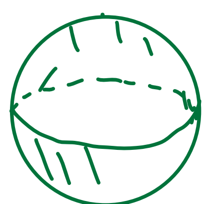
Visually: $H_0 = \#$ of connected comp.
$H_1 = \#$ of "loops"
$H_2 = \#$ of "voids" (think S^2/sphere).



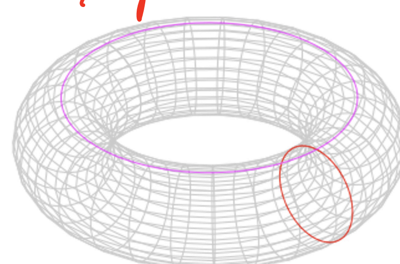
$H_0 = \mathbb{Z}_2^3$
$H_n = 0, n>0$



$H_0 = \mathbb{Z}_2$
$H_1 = \mathbb{Z}_2$
$H_n = 0 \; \forall n>1$



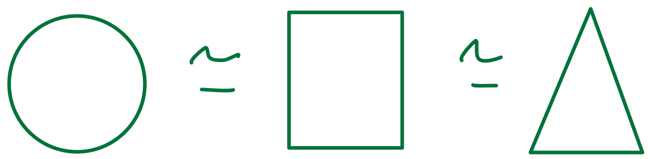
$H_0 = \mathbb{Z}_2$
$H_1 = 0$
$H_2 = \mathbb{Z}_2$
$H_n = 0 \; \forall n>2$



$H_0 = \mathbb{Z}_2$, $H_1 = \mathbb{Z}_2^2$
$H_2 = \mathbb{Z}_2$, $H_n = 0 \; \forall n>2$

1) Homology is homotopy invariant

→ Shape of data



2) Homology is a functor

$$X \xrightarrow{f} Y \rightsquigarrow H_n(X) \xrightarrow{f_*} H_n(Y)$$

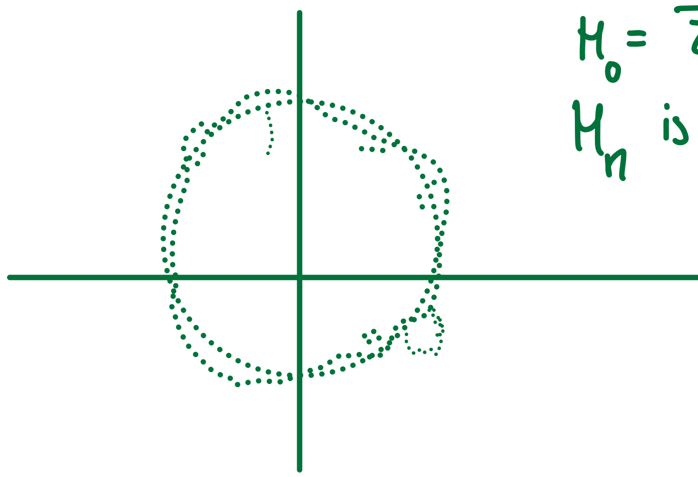
→ Critical in constructing barcodes

3) Simplicial Homology is computable

→ Basically row reduction

Going back to the original problem – the topology of point cloud data.

At first glance, applying simplicial homology to point cloud data yields no interesting results. Consider

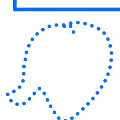


$H_0 = \mathbb{Z}_2^{(\#\text{ of points})}$

H_n is trivial $\forall n > 0$.

But visually we see a circle, so we expect $H_1 = \mathbb{Z}_2$!

$$\boxed{\text{Data}} \longrightarrow \boxed{\text{Filtered Simplicial Complex}} \xrightarrow{\text{Homology}} \boxed{\text{Persistence Modules}} \xrightarrow{\text{Struc. Thm}} \text{Barcode}$$

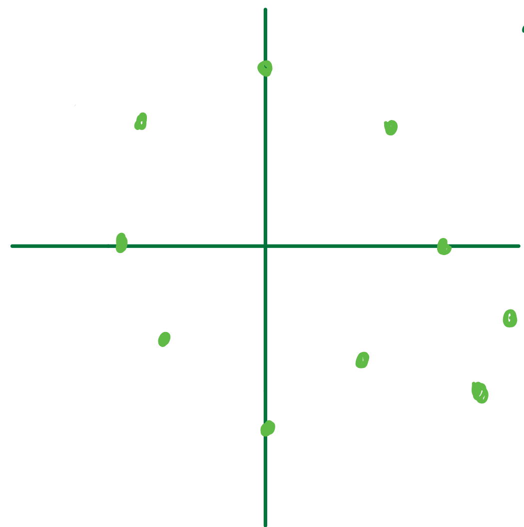


$$K^0 \subseteq K^1 \subseteq \cdots \subseteq K^n$$

$$H_i(K^0) \to H_i(K^1) \to \cdots \to H_i(K^n)$$

From Point Cloud Data to Simplicial Complexes

Consider a simple example with 10 points:
Let X denote the data set.

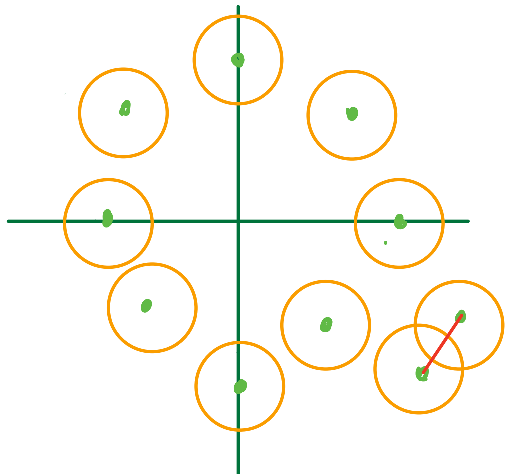


$$\left.\right\} \begin{array}{l} H_0 = (\mathbb{Z}/_2)^{10} \\ H_n = 0 \ \forall n > 0. \end{array}$$

Let $\varepsilon > 0$, then we define a simplicial complex $VR(X, \varepsilon)$ as follows:

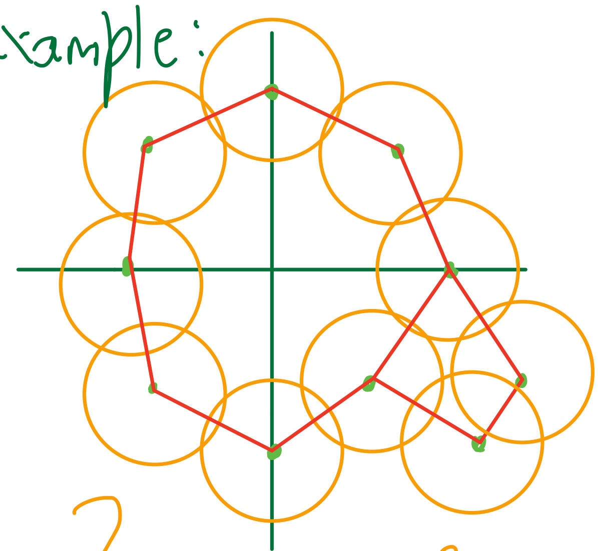
1. The vertices are simply the points from X
2. Let σ denote the simplex spanned by $\{x_0, \dots, x_n\} \in X$. Then $\sigma \in VR(X, \varepsilon)$ iff $d(x_i, x_j) < \varepsilon \ \forall i, j$.

In the case of our example:



ε_1

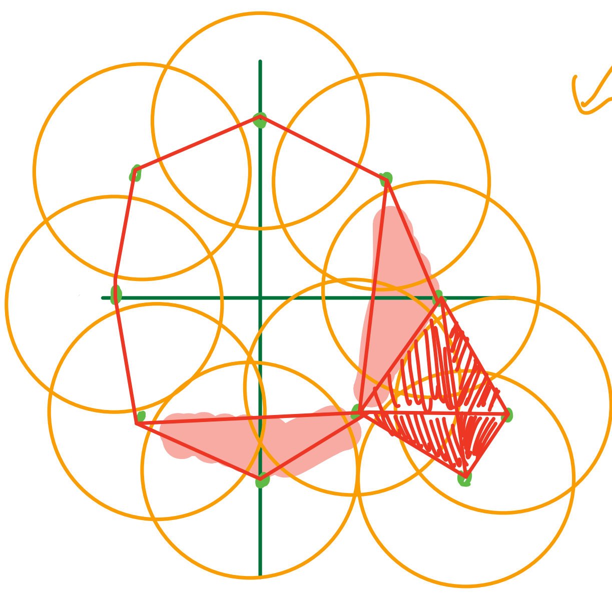
$$H_0(X, \varepsilon_1) = (\mathbb{Z}/_2)^9$$
$$H_n(X, \varepsilon_1) = 0 \quad \forall n>0$$



ε_2

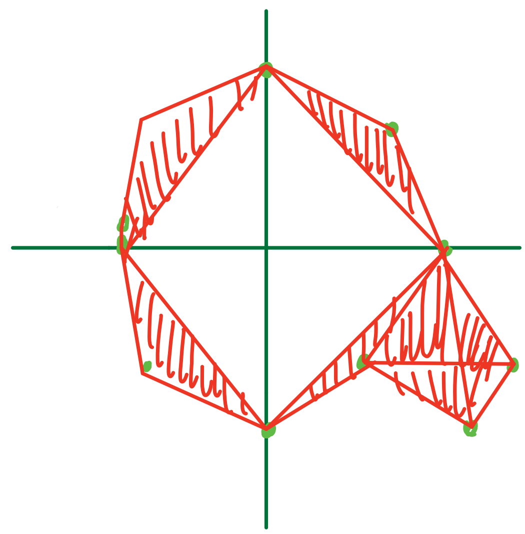
$$H_0(X, \varepsilon_2) = \mathbb{Z}/_2$$
$$H_1(X, \varepsilon_2) = (\mathbb{Z}/_2)^2$$

Noise!



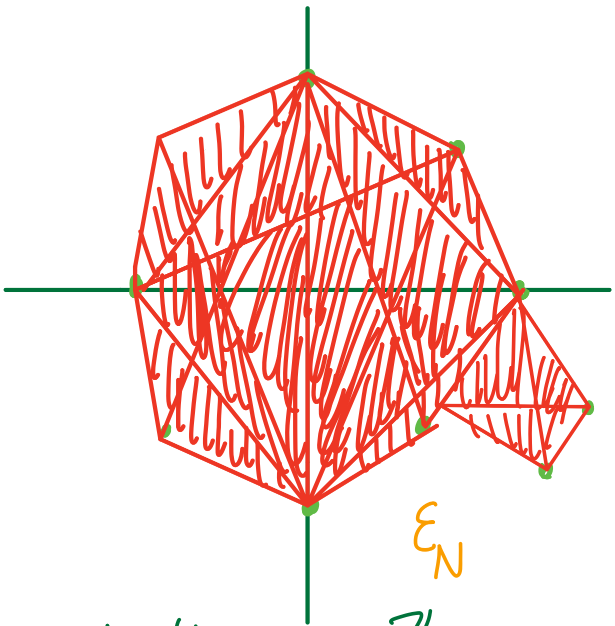
ε_3

$$H_0(X, \varepsilon_3) = \mathbb{Z}/_2$$
$$H_1(X, \varepsilon_3) = \mathbb{Z}/_2$$



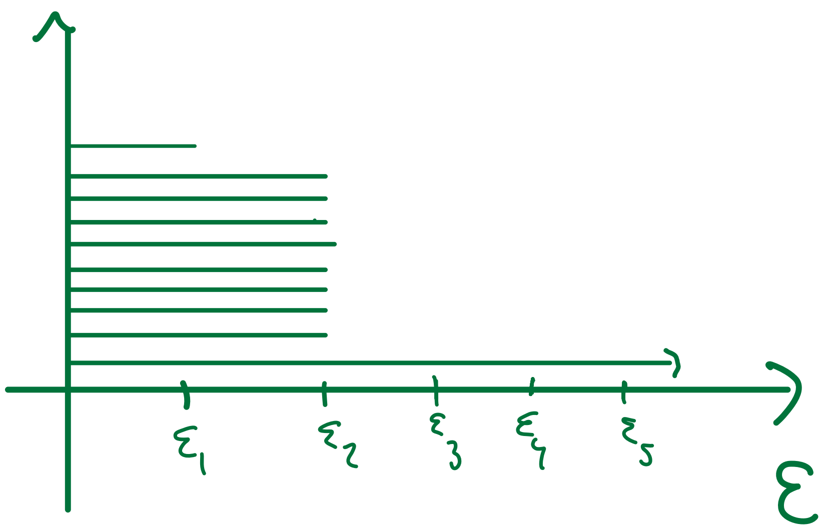
ε_4

$$H_0(X, \varepsilon_4) = \mathbb{Z}/_2$$
$$H_1(X, \varepsilon_4) = \mathbb{Z}/_2$$

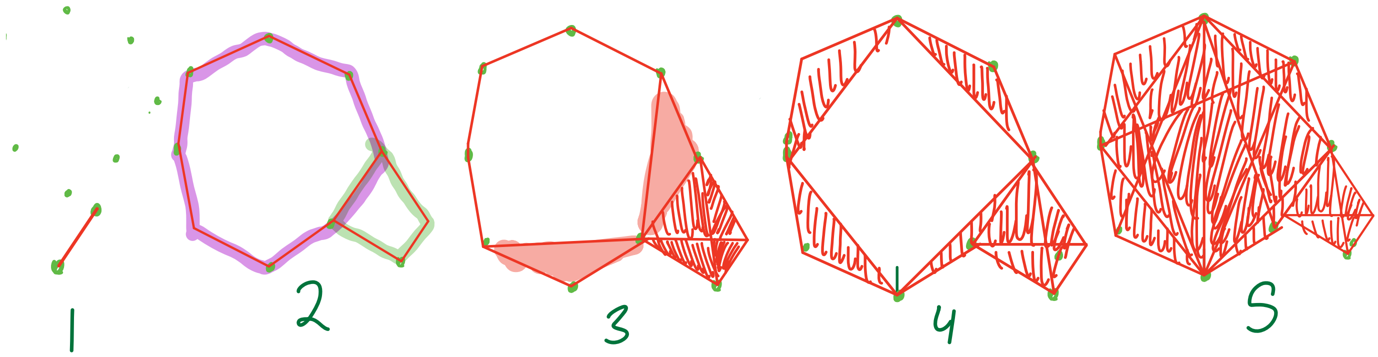


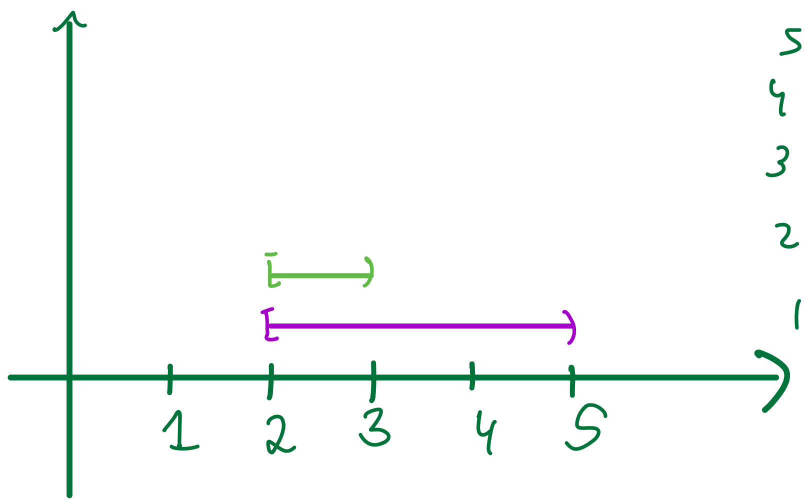
$H_0(X, \varepsilon_N) = \mathbb{Z}_2$

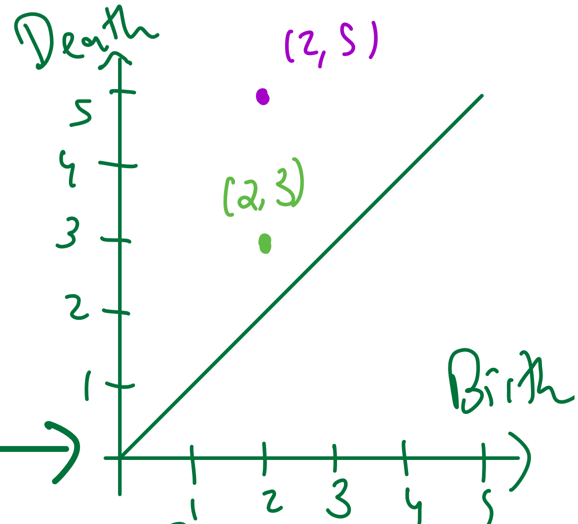
$H_1(X, \varepsilon_N) = 0$



Barcode for H_0







Barcode for H_1

Persistence Diagram

The short intervals can be interpreted as noise.
For example the cycle that is born at ε_1 and dies immidiately after at ε_2, is caused by the two noisy outliers
On the persistence Diagram, values close to the diagonal are noise.

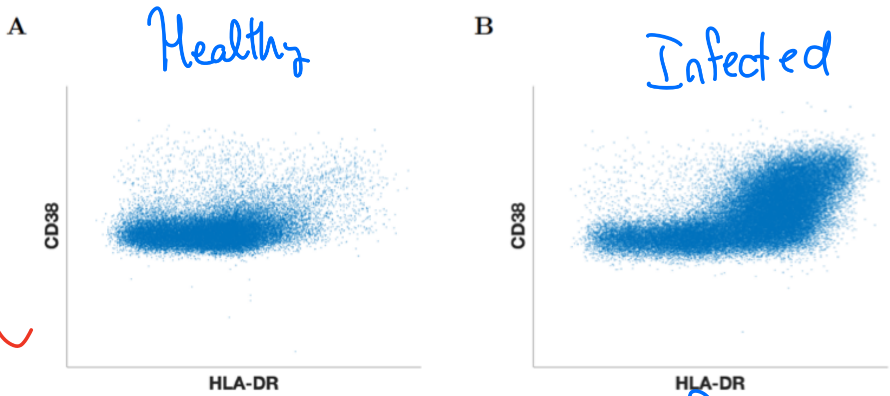
An Application

One of the first results, when I looked on arxiv 2 days ago

Determining clinically relevant features in cytometry data using persistent homology

Soham Mukherjee[1], Darren Wethington[2,4], Tamal K. Dey[1], Jayajit Das[2,3,4,5*]

Transformed Scatter Plot - earlier paper



Data set - Protien expression data of a certain kind of cells - focusing on 3 protiens in particular.

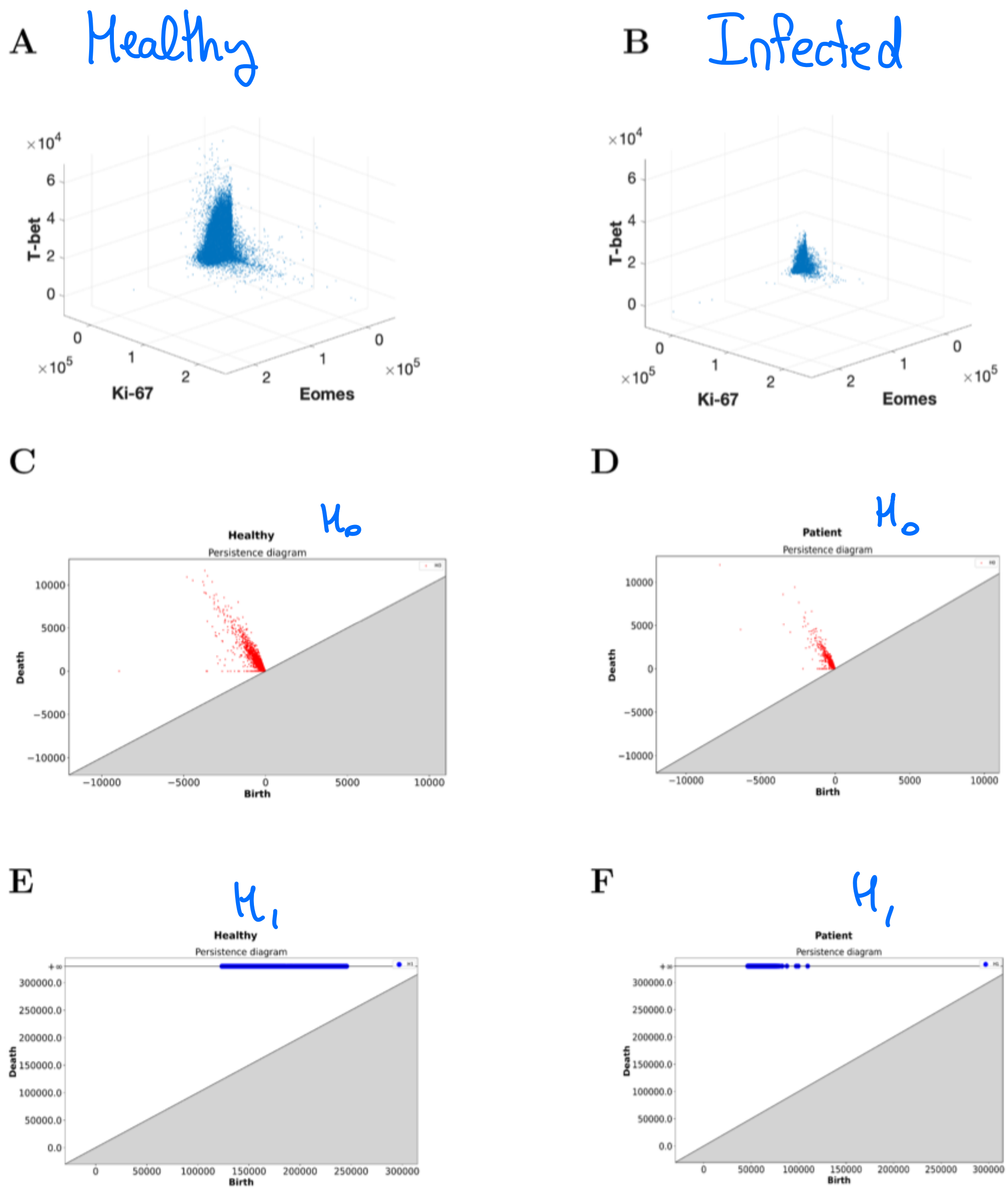
Protien expression from each donor → Persistence Diagram for each donor

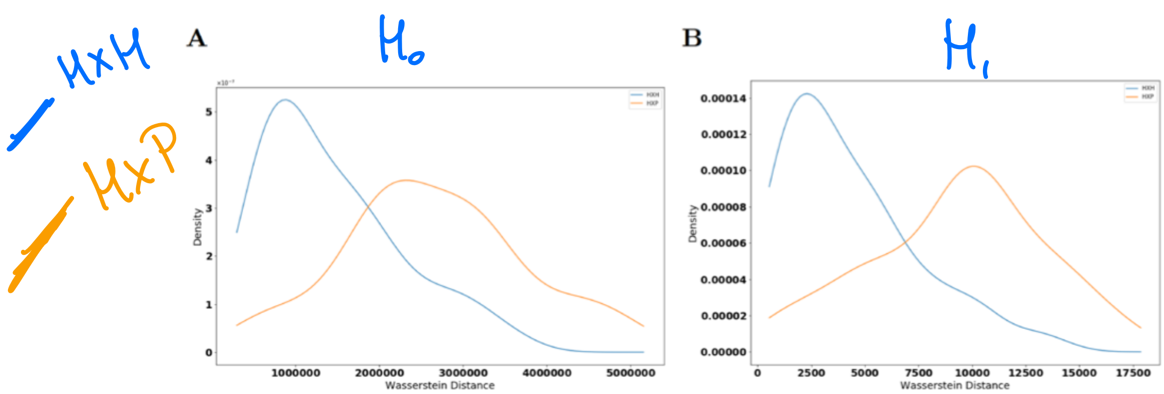
→ Randomly select pairs of either

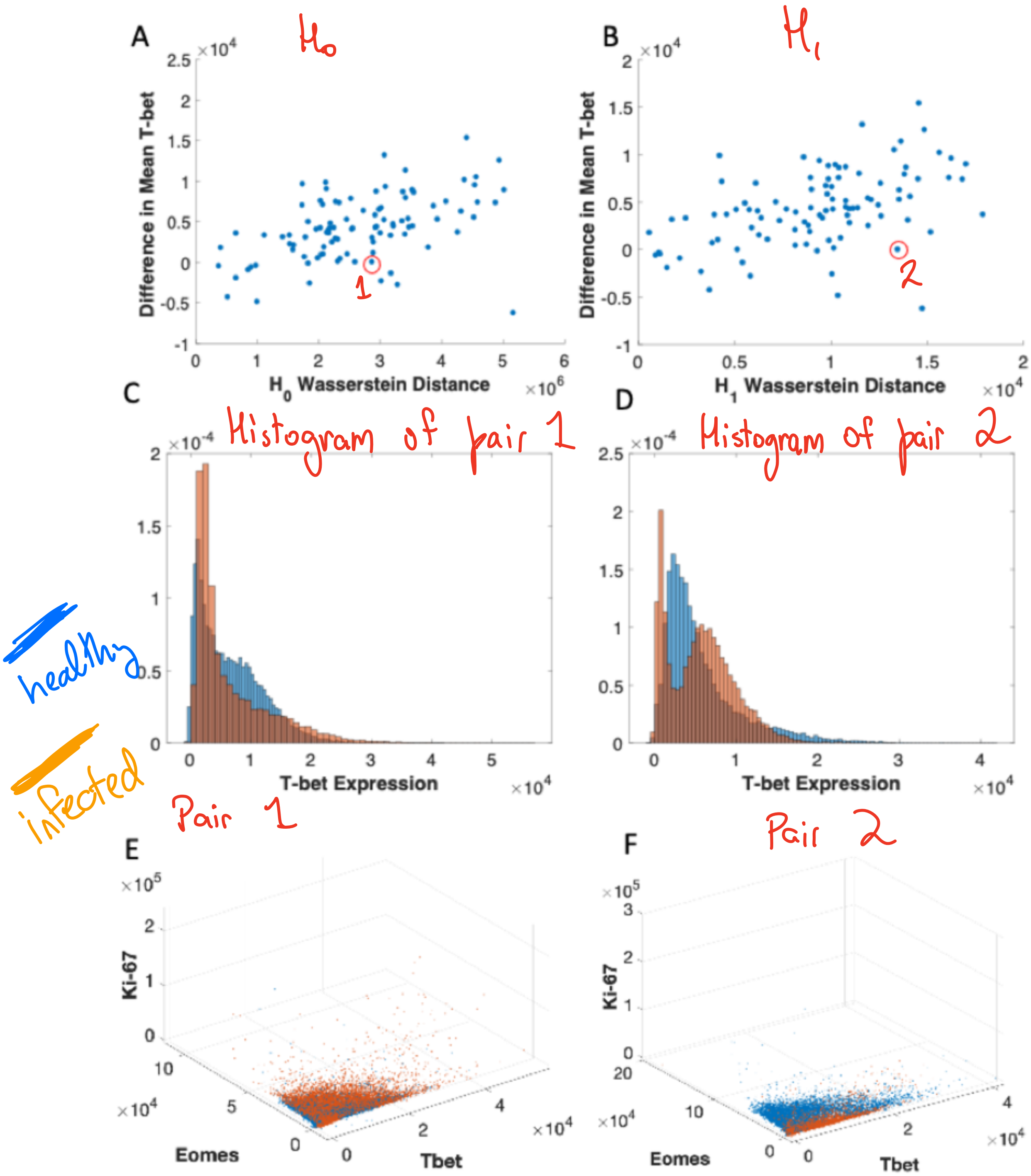
- Healthy - Healthy - H x H
- Healthy - Infected - H x P

→ Compute the distances of the persistence diagrams for each pair

→ Is the distribution of distances for H x H different than the distribution for H x P?

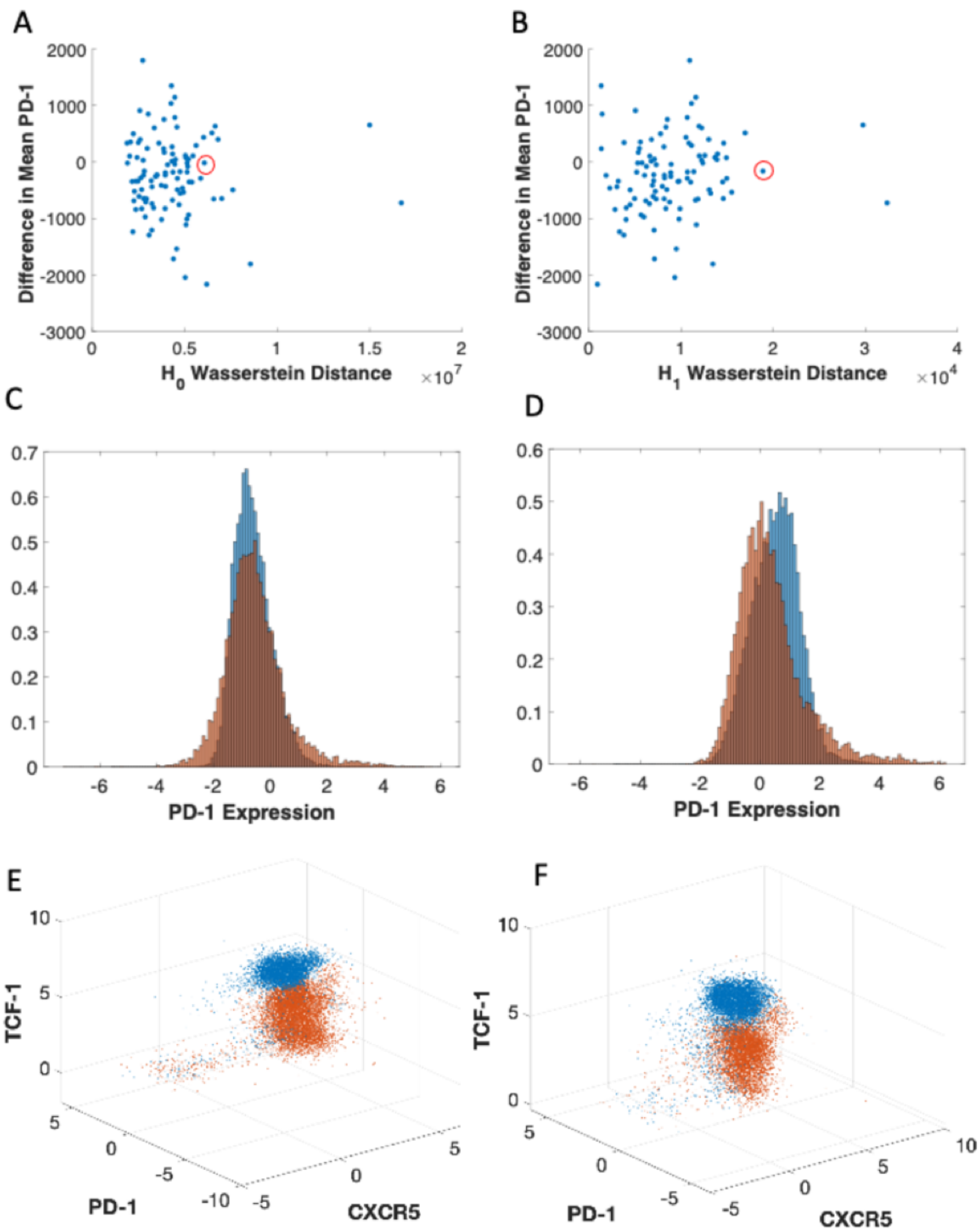

A
Healthy
B
Infected
T-bet
Ki-67
Eomes
C
Healthy
Persistence diagram
H_0
D
Patient
Persistence diagram
H_0
Death
Birth
E
Healthy
Persistence diagram
H_1
F
Patient
Persistence diagram
H_1
Death
Birth


HxH
HxP
A
H_0
Density
Wasserstein Distance
B
H_1
Density
Wasserstein Distance



Persistent Homology reveals new information!

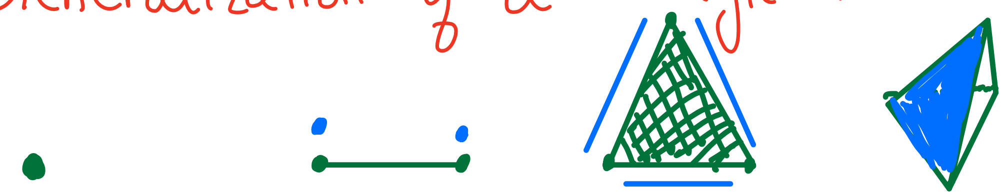
The same analysis but for B-cells



Def (n-simplex): Let $\{v_0, \dots, v_n\}$ be affinely independent points in $\mathbb{R}^n$. The n-simplex spanned by $\{v_0, \dots, v_n\}$ is the convex hull of these points. Points of the form:

$$x = \sum_{i=0}^{n} \lambda_i v_i, \quad \sum \lambda_i = 1, \quad \lambda_i \in \mathbb{R}_{\geq 0}$$

"Generalization" of a triangle to $n \geq 0$ dims



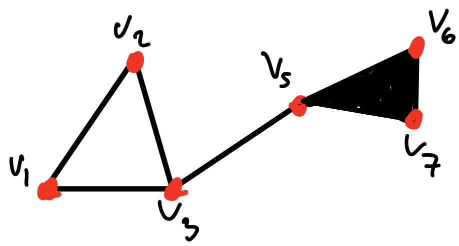
0-simplex 1-simplex 2-simplex 3-simplex

"Sticking" simplices together forms simplicial complexes

Def (Simplicial Complex): A simplicial complex is a set K of vertices & a collection S of subsets of K s.t.:

1) $\{v\} \in S \ \forall v \in K$ } all vertices are included.

2) $\sigma \in S$ and $\tau \subseteq \sigma \Rightarrow \tau \in S$ } faces of simplices in the complex are in the complex

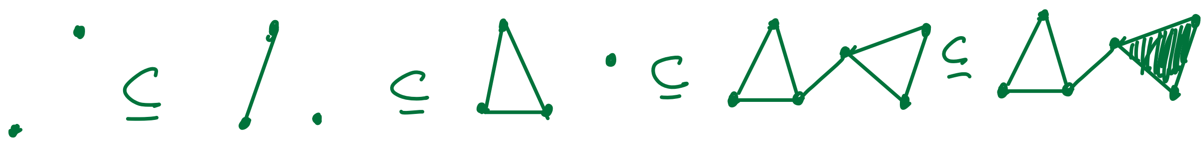


$$S = \Big\{ \{v_i\}_{i=1}^{7}, \underbrace{\{v_1, v_2\}, \{v_2, v_3\}, \dots, \{v_6, v_7\}}_{\text{edges}}, \ \text{triangle}\{\{v_5, v_6, v_7\}\} \Big\}.$$

Remark: One usually also picks an orientation for each simplex in a simplicial complex ($[v_5, v_6, v_7]$ for ex.), but for our purposes the orientation is irrelevant.

Def (Subcomplex, Filtration): A subcomplex L of a simplicial complex K is a subset $L \subseteq K$ which is also a simplicial complex.
A filtration of K is a sequence of complexes:

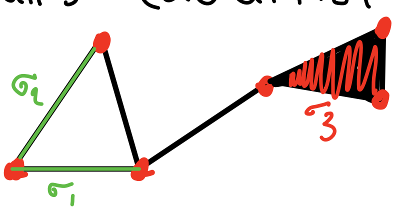
$$K^0 \subseteq K^1 \subseteq K^2 \subseteq \dots \subseteq K^n = K$$



Def (Chain group): Let K be a simplicial complex. Then an n-chain is a linear combination of n-simplices:

$$c = \sum_i c_i \sigma_i \;, \quad \underbrace{c_i \in \mathbb{Z}_2}_{\text{choice of coefficients!}}, \quad \sigma_i \in K \text{ an } n\text{-simplex}$$

$C_n(K)$ is the free abelian group generated by n-chains (addition is pointwise).



$\sigma_1 + \sigma_2 \in C_1(K)$

$\sigma_3 \in C_2(K)$.

Remark: For coefficients in $\mathbb{Z}_2$ (or any field), $C_n(K)$ is a vector space, since it is a free module (basis n-simplices in K) over a field $\mathbb{Z}_2$.

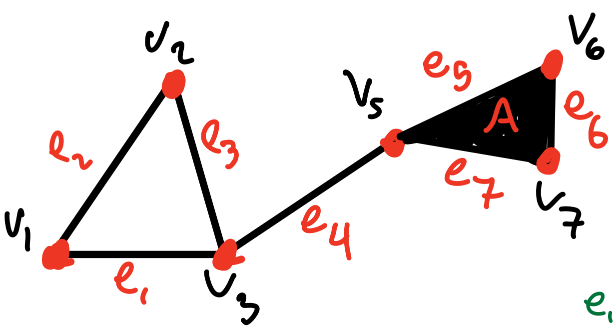
Def (Boundary homomorphism): Define a homomorphism $\partial_n : C_n \longrightarrow C_{n-1}$ by defining it on n-simplices $\sigma \in K$ and extending linearly:

$$\partial_n \sigma = \sum_{i=0}^{n} (-1)^i \underbrace{[v_0, \dots, \hat{v}_i, \dots, v_n]}_{\text{(n-1)-simplex with } v_i \text{ removed}}$$

$$\overset{(\text{over } \mathbb{Z}_2)}{=} \sum_{i=0}^{n} [v_0, \dots, \hat{v}_i, \dots, v_n]$$

Easy to verify that $\partial_n \partial_{n+1} = 0$

$$\underbrace{\cdots \longrightarrow C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots}_{(C_\bullet, \partial) \text{ is a chain complex}}$$



$C_2 = \langle A \rangle$, $C_1 = \langle e_1, \dots, e_7 \rangle$, $C_0 = \langle v_0, \dots, v_7 \rangle$, $C_n = 0 \ \forall n > 2$

$C_1 \longrightarrow C_0$

$$\partial_2 = \begin{array}{c} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{array} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \partial_1 = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{array} \begin{array}{c} \begin{array}{ccccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array}, \quad \partial_0 = 0$$

Boundary homomorphisms are linear maps!

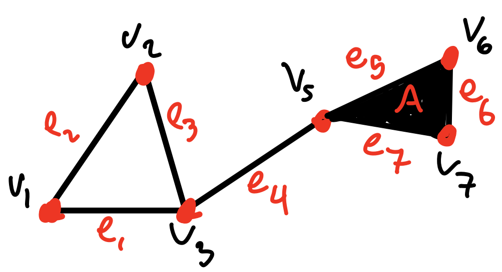
Def (Cycles, Boundaries, Homology): Let:

$$\underbrace{Z_n(K) = \ker \partial_n}_{\text{Cycles}} \quad \text{and} \quad B_n(K) = \underbrace{\operatorname{im} \partial_{n+1}}_{\text{Boundaries}}$$

Since $\partial_n \partial_{n+1} = 0$ we have $B_n \subseteq Z_n$ so

$$\underbrace{H_n(K) = \frac{\ker \partial_n}{\operatorname{im} \partial_{n+1}}}_{\text{Homology}} = Z_n / B_n \text{ is well defined}$$

Homology – over $\mathbb{Z}_2$ a quotient vector space.



$$\partial_1 = \begin{array}{c|ccccccc} & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 & e_7 \\ \hline v_1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ v_2 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ v_3 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ v_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ v_5 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ v_6 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ v_7 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}, \quad \partial_2 = \begin{array}{c|c} e_1 & 0 \\ e_2 & 0 \\ e_3 & 0 \\ e_4 & 0 \\ e_5 & 1 \\ e_6 & 1 \\ e_7 & 1 \end{array}$$

Check: $\ker \partial_1 = \langle e_1 + e_2 + e_3,\ e_5 + e_6 + e_7 \rangle$, $\operatorname{im} \partial_2 = \langle e_5 + e_6 + e_7 \rangle$

$\Rightarrow H_1 = \langle e_1 + e_2 + e_3 \rangle \simeq \mathbb{Z}_2$

Visually: $H_0 = \#$ of connected comp.

$H_1 = \#$ of "loops"

$H_2 = \#$ of "voids" (think S^2/sphere).

→ Computing simplicial homology amounts to computing kernels & images of matrices — easy!

→ Row reduction on ∂_n yields $Z_n = \ker \partial_n$ and $B_{n-1} = \operatorname{im} \partial_{n-1}$

Let $X \subseteq \mathbb{R}^n$ be a finite, discrete subset, then:

① $VR(X, \varepsilon_1) \subseteq VR(X, \varepsilon_2) \quad \forall \varepsilon_1 \leq \varepsilon_2$

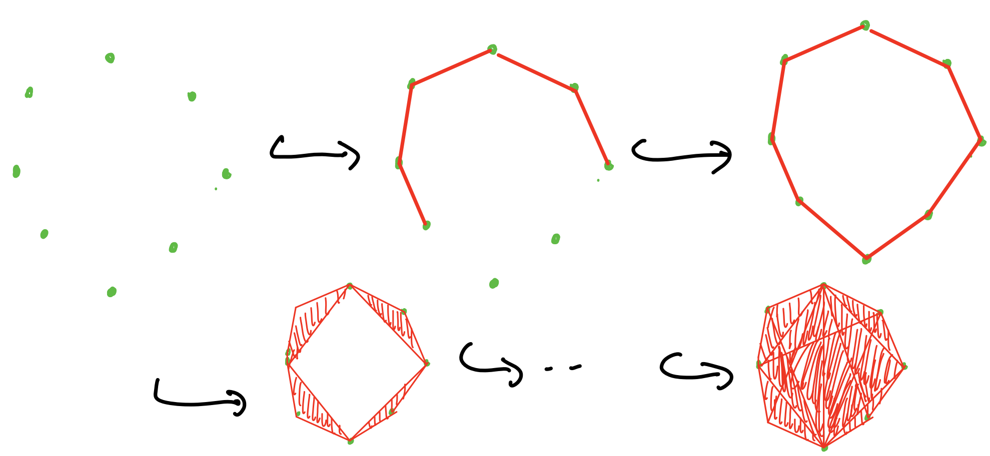
② $VR(X, \varepsilon) = VR(X, \varepsilon_N)$ for some $\varepsilon_N > 0$

Thus $VR(X, \varepsilon)$ defines a filtration of simplicial complex.

The filtration is finite. New simplices are added only at finitely many times.

So given finite point cloud data $X \subseteq \mathbb{R}^n$ we have a filtered simplicial complex

$$K^0 \subseteq K^1 \subseteq K^2 \subseteq \cdots \subseteq K^n$$



Applying H_k we get

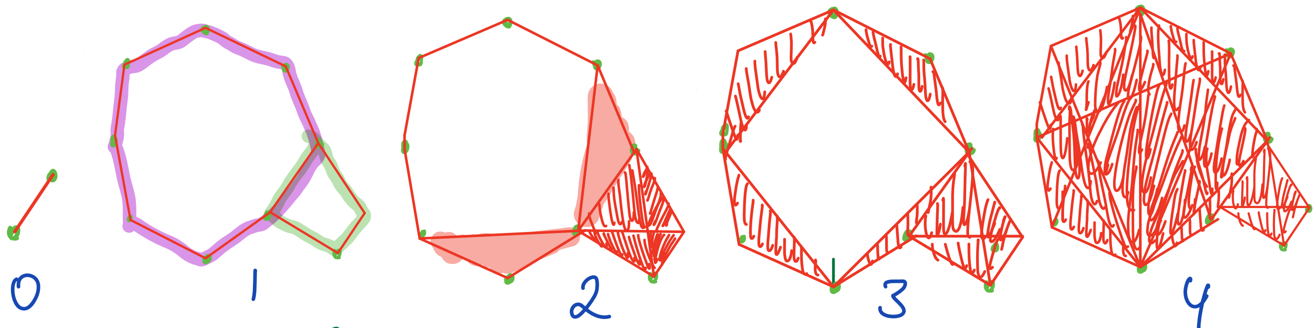
$$H_k(K^0) \longrightarrow H_k(K^1) \longrightarrow \cdots \longrightarrow H_k(K^n)$$

Notation: $H_k(K^i) = H_k^i$, $Z_k(K^i) = Z_k^i$, $B_k(K^i) = B_k^i$

Def (Persistent homology): The p^{th} persistent, $\underbrace{k^{th}\text{ homology}}_{\text{homology dim.}}$ group of K^i is:

$$H_k^{p,i} = \frac{\underbrace{Z_k^i}_{k\text{-cycles of } K^i = Z_k(K^i)}}{\underbrace{\underbrace{B_k^{i+p}}_{k\text{-boundaries of } K^{i+p}} \cap Z_k^i}_{k\text{-cycles of } K^i \text{ that are boundaries in } K^{i+p}}}$$

$H_k^{p,i}$: which cycles in K^i survive in K^{i+p} – no longer persist.



$\notin H_1^{1,2}$

$\in H_1^{1,2}$, $\in H_1^{1,3}$ since the loop still persists.

Thus calculating persistence groups boils down to finding a common basis across

$$H_k(K^0) \to H_k(K^1) \to \cdots \longrightarrow H_k(K^n).$$

Def (Persistence Modules): Let R be a ring, then a persistence Module M is a family of R-modules M^i together with homomorphism $\varphi^i : M^i \longrightarrow M^{i+1}$ $\forall i$

A persistence module is of finite type if:

$\exists N$ s.t. $\varphi^n : M^n \to M^{n+1}$ is an iso for $n \geq N$.

Clearly $H_k(K^0) \to H_k(K^1) \longrightarrow \cdots \longrightarrow H_k(K^n)$ is a persistence Module of finite type.

Our task is now to somehow classify pers. modules of finite type

Algebra Interlude

Polynomial ring w/ standard grading: Let F be a field and $F[t]$ be its polynomial ring, then

Stand. grading: $F[t] = \bigoplus_{i \in \mathbb{N}} F_i$, $F_i = F t^i$

Clearly $\underbrace{\alpha t^n}_{F_n} \underbrace{\beta t^m}_{F_m} = \alpha\beta t^{n+m} \in F_{n+m}$

Important for us: Let M be a fin. gen graded $F[t]$-module, then

$$M \cong \bigoplus_{i=1}^{n} \Sigma^{\alpha_i} F[t] \oplus \bigoplus_{j=1}^{m} \Sigma^{\gamma_j} F[t] / t^{n_j}$$

where $\Sigma^{\alpha} F[t]$ denotes an α-shift upward in the gradatation, $\alpha_i, \gamma_j, n_j \in \mathbb{Z}$

This is completely analagous to the non graded case: Let M be a fin. gen. R-module over a PID R, then

$$M \cong R^n \oplus \bigoplus_{j=1}^{m} R / d_j R$$

In particular if $R = \mathbb{Z}$, then we have the familiar theorem for fin. gen. abelian groups

$$G \cong \mathbb{Z}^n \oplus \bigoplus_{j=1}^{m} \mathbb{Z} / n_j \mathbb{Z}$$

Correspondence

Theorem: Let $\mathcal{M} = (M^i, \varphi^i)$ be a persistence module of finite type, then $\bigoplus H_k(K^i$

$$\alpha(\mathcal{M}) = \bigoplus_{i=0}^{n} M^i \quad \text{endowed with}$$

$$t \cdot (m^0, m^1, m^2, \dots, m^n) = (0, \varphi^0(m^0), \varphi^1(m^1), \varphi^2(m^2), \dots, \varphi^{n-1}(m^{n-1}))$$

$$t^2 (m^0, m^1, \dots, m^n) = (0, 0, \varphi^1\varphi^0(m^0), \varphi^2\varphi^1(m^1), \dots, \varphi^{n-1}\varphi^{n-2}(m^{n-2}))$$

$$\vdots$$

is a finitely generated, non-neg. graded $R[t]$-Mod.

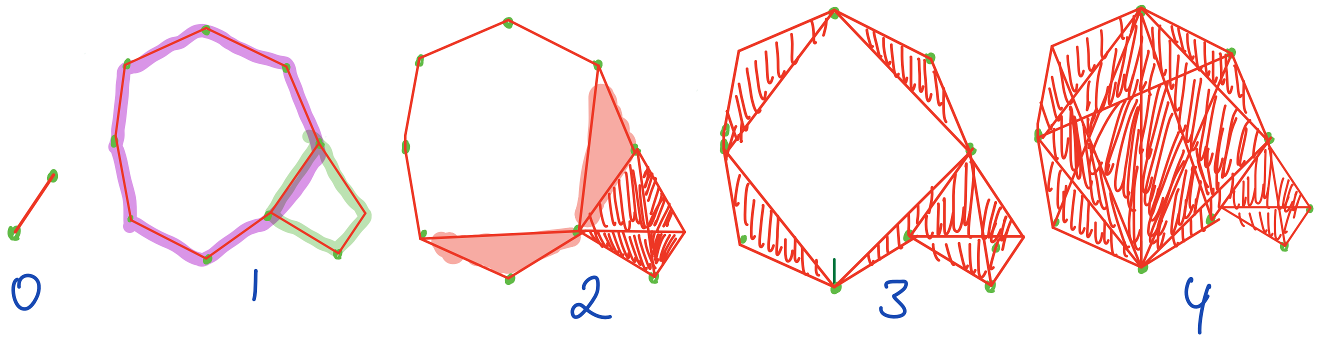
α establishes a equivalence of categories:

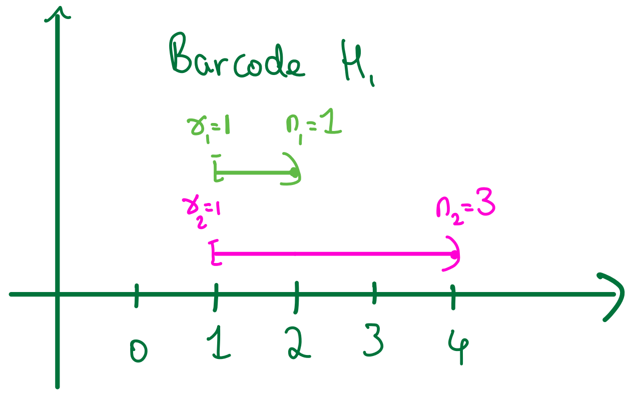
$$\mathrm{FinPers}_R \longleftrightarrow \mathrm{FinGrad}_{R[t]}$$

The Structure Theorem then delivers a classification, but only when R is a field. In the case of $\mathbb{Z}$ for example, $\mathbb{Z}[t]$ is <u>not</u> a PID and we have no structure theorem for fin. gen. modules over general rings.

However if we work over a field F (like $\mathbb{Z}_2$):

$$\alpha(\mathcal{M}) \cong \bigoplus_{i=1}^{n} \Sigma^{\alpha_i} F[t] \oplus \bigoplus_{j=1}^{n} \Sigma^{\gamma_j} F[t] / t^{n_j}$$





γ_i = birth times

$\gamma_i + n_j$ = death time.

$$\Sigma^1 F[t] / t$$

$$\Sigma^1 F[t] / t^3$$

Applying t in $F[t]$ $\rightsquigarrow$ Applying $f^i : H_1(K^1) \to H_1(K^2)$ to $M = (H_1(K^i), f^i)$

$\Sigma^1 F[t] / t$ $\rightsquigarrow$ dies in K^2.

applying t once takes me to 0 (mod t)

$\mathcal{M}$ | $\alpha(\mathcal{M})$

$f^1 : H_1(K^1) \to H_1(K^2)$ $\Longleftrightarrow$ $\rightsquigarrow$ $\Sigma^{\alpha_i} F[t] / t$

$f^1(\ ) = 0$ $\Longleftrightarrow$ Same as applying t in $\alpha(\mathcal{M})$, but this goes to $0 \pmod{t}$

dies in K^2

In particular we can read off the barcode

$$M \cong \underbrace{\bigoplus_{i=1}^{n} \Sigma^{\alpha_i} F[t]}_{\text{n intervals of the form } [\alpha_i, \infty)} \oplus \underbrace{\bigoplus_{j=1}^{m} \Sigma^{\gamma_j} F[t] / t^{n_j}}_{\text{m intervals of the form } [\gamma_j, \gamma_j + n_j)}$$

G69