An introduction to Topological Data Analysis

Yue Liu

Duke University

MISM Journal Club
2026/11/23

yue.liu@duke.edu
https://liuyue002.github.io/



0: Outline

I Motivation

II A non-technical primer

Munch (2017). A user's guide to topological data analysis.

Journal of Learning Analytics, 4(2), 47–61. doi: 10.18608/jla.2017.42.6

III Demonstration on zebrafish patterns

Liu & Volkening (2026). Multi-objective Bayesian inference in an agent-based model of zebrafish patterns via topological data analysis. arXiv:2605.18685

IV Mathematical foundations

Topaz et al (2015). Topological Data Analysis of Biological Aggregation Models.

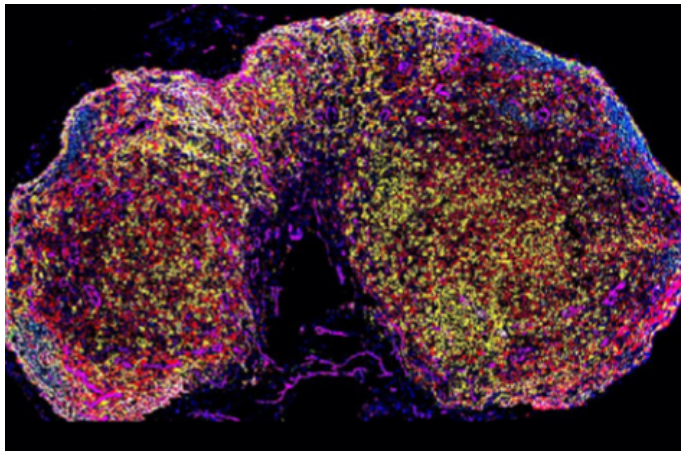
PLOS ONE e0126383. doi: journal.pone.0126383

V Beyond Vietoris-Rips and persistent homology

I: What is the shape of a point-cloud?

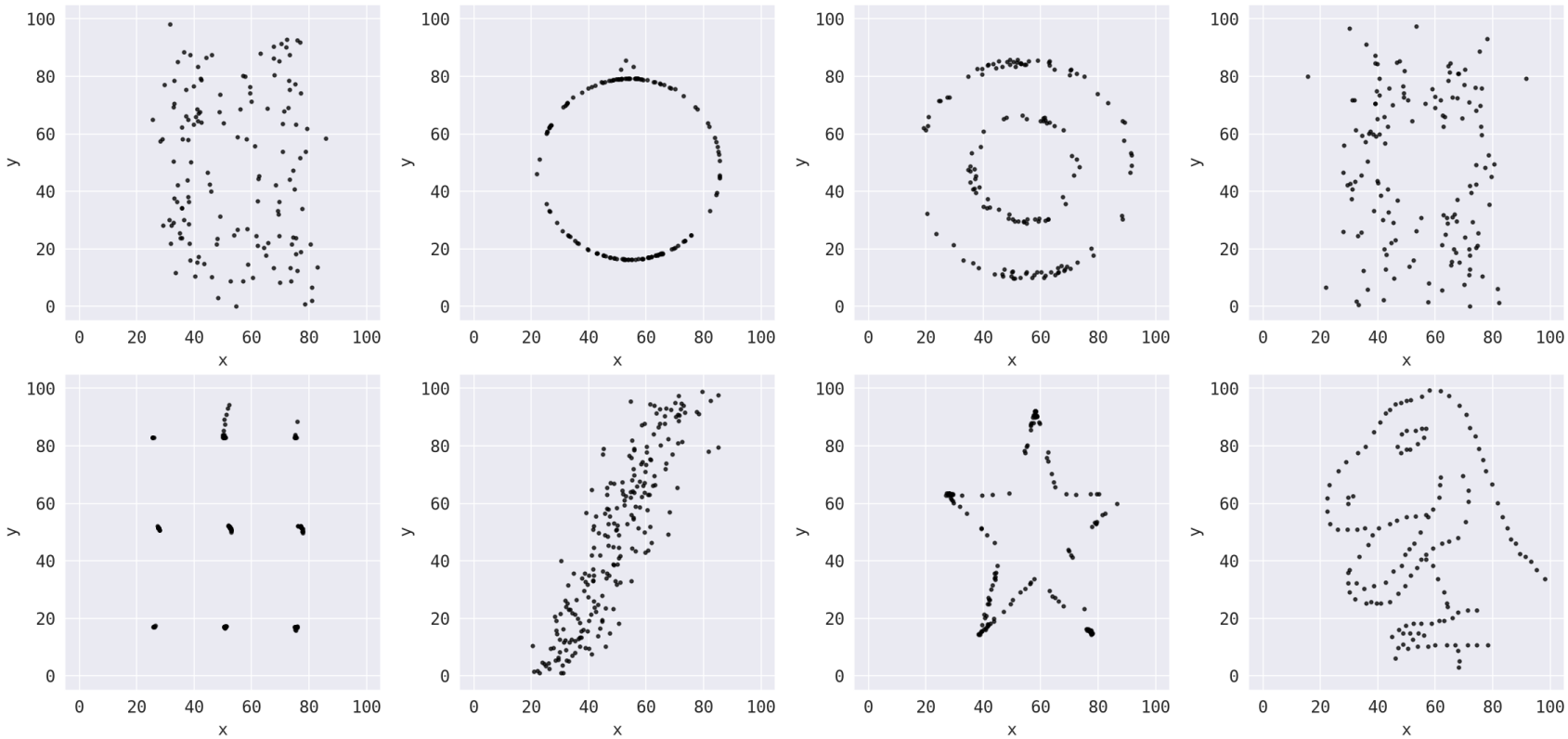
And how to mathematically describe it?





Swan Murmuration: Pete Lau, Yorkshire Wildlife Trust; CODEX cell imaging: Hickey et al 2024

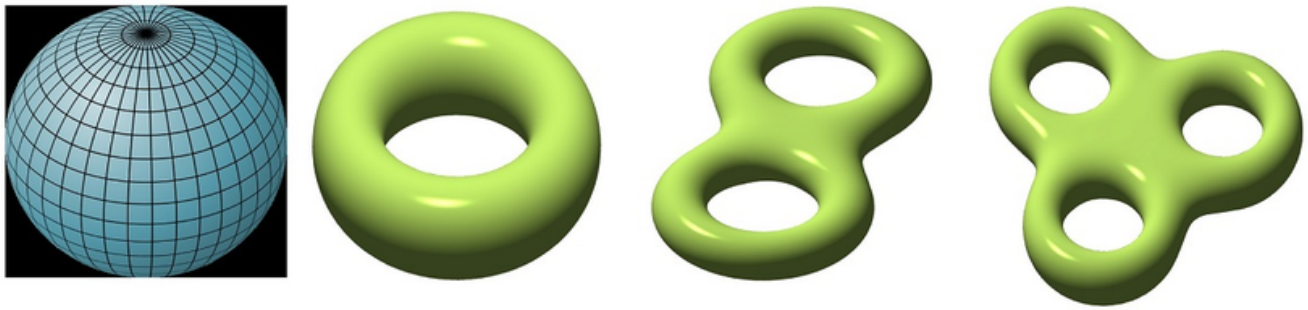
I: Simple summary statistics do not capture topological information



The "Datasaurus Dozen" datasets (Cairo 2016): all have the same $E[X], E[Y], Var[X], Var[Y], Cov(X, Y)$, linear regression fit

I: What summary stats missed: Topological information

Example of topological information: connected components, loops, holes



These shapes has 0-3 holes, respectively

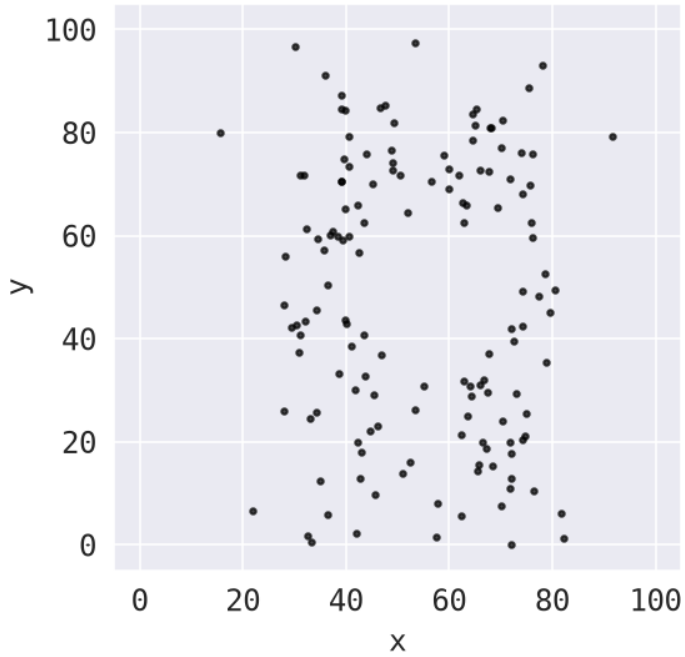
I: Topological info is not a complete description, either

Topology discard certain geometric/statistical information that might be important



Ceramic models by Crane & Segerman, 2015

I: How do we make "connected components" and "holes" precise?



Meso-scale: 2 connected components and 2 holes ?
Micro-scale: many connected components and holes
Macro-scale: 1 connected component, no holes
To "zoom" across scale and make things precise, we need **persistent homology**

I: Outline

I Motivation

II A non-technical primer

Munch (2017). A user's guide to topological data analysis.

Journal of Learning Analytics, 4(2), 47–61. doi: 10.18608/jla.2017.42.6

III Demonstration on zebrafish patterns

Liu & Volkening (2026). Multi-objective Bayesian inference in an agent-based model of zebrafish patterns via topological data analysis. arXiv:2605.18685

IV Mathematical foundations

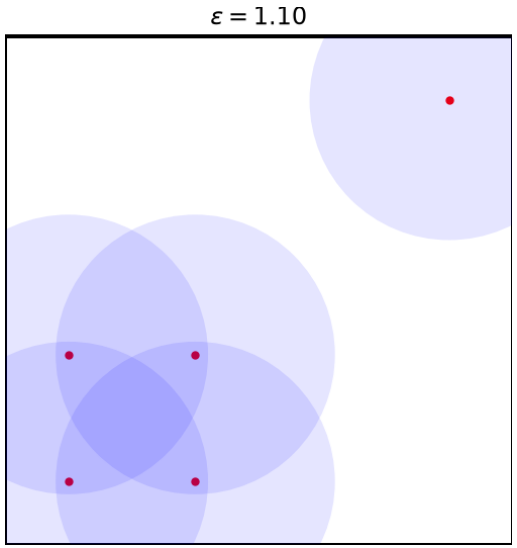
Topaz et al (2015). Topological Data Analysis of Biological Aggregation Models.

PLOS ONE e0126383. doi: journal.pone.0126383

V Beyond Vietoris-Rips and persistent homology

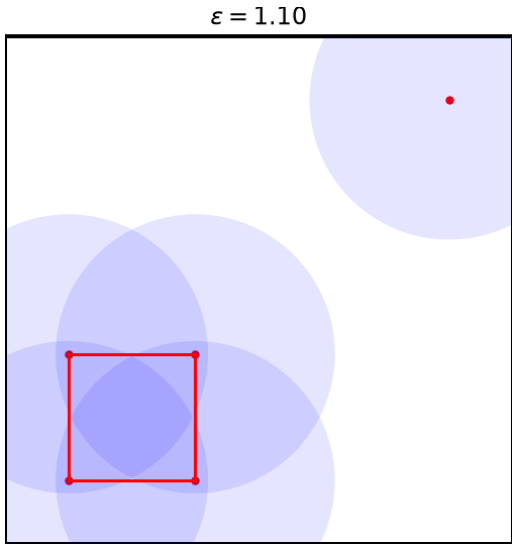
II: Vietoris-Rips Filtration

Step 1: Given a point cloud dataset $\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^n$ and for a given $\varepsilon > 0$, consider ε-balls around each point



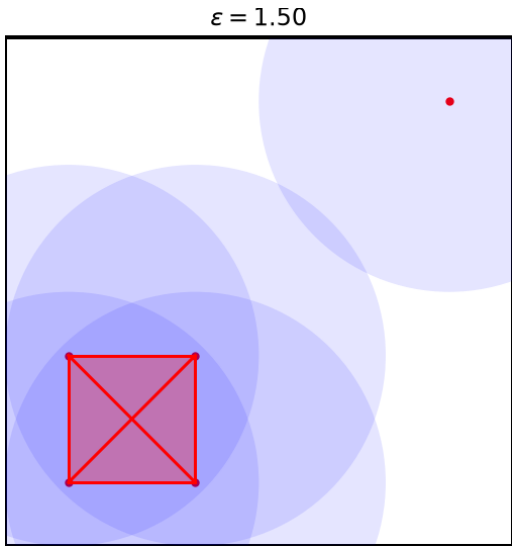
II: Vietoris-Rips Filtration

Step 1: We add a an edge joining $\{x_i, x_j\}$ if $d(x_i, x_j) < \varepsilon$, i.e, if their respective ε-ball includes the other point



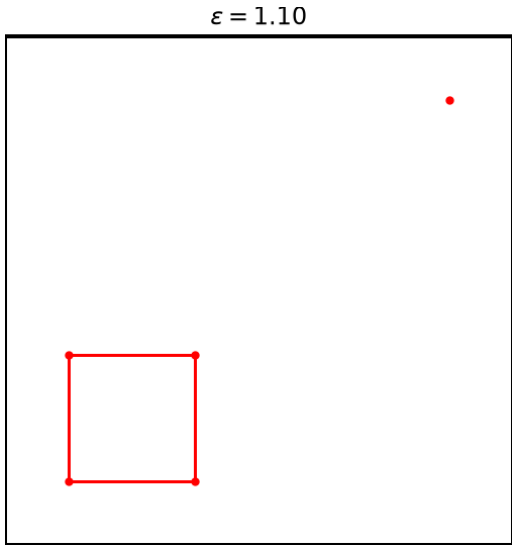
II: Vietoris-Rips Filtration

Step 2: If there is a subset of points that are pairwise joined, we add a filled simplex:
3 points $\rightarrow$ 2D simplex=triangle,
4 points $\rightarrow$ 3D simplex=tetrahedron, etc



II: Vietoris-Rips Filtration

Step 3: Now we view the resulting object as an embedding in $\mathbb{R}^N$ (N=#points, not $\mathbb{R}^n$ the space of the original data!) and count the topological features of the resulting object



$$\beta_0 = \text{“betty-0”} \quad = \#\text{connected components} \quad = 2 \qquad (1)$$

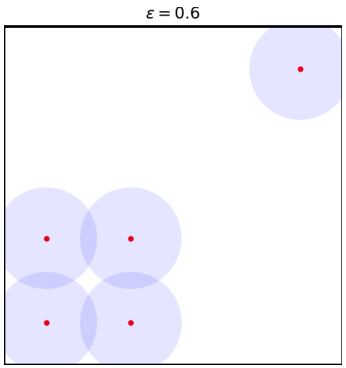
$$\beta_1 = \text{“betty-1”} \quad = \#\text{loops} \quad = 1 \qquad (2)$$

$$\beta_2 = \text{“betty-2”} \quad = \#\text{voids} \quad = 0 \qquad (3)$$

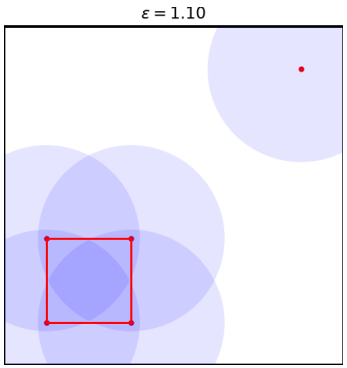
β_k = higher dimensional analogues=0

II: Persistent homology

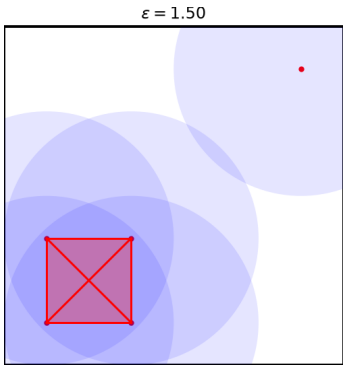
Key to robust characterization of the dataset: vary the scale parameter ε. The most significant features should **persist** for a wide range of ε



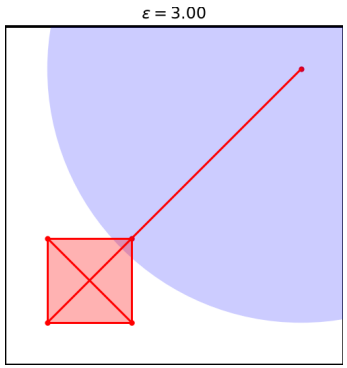
(a)
$\beta_0 = 5, \beta_1 = 0$



(b)
$\beta_0 = 2, \beta_1 = 1$



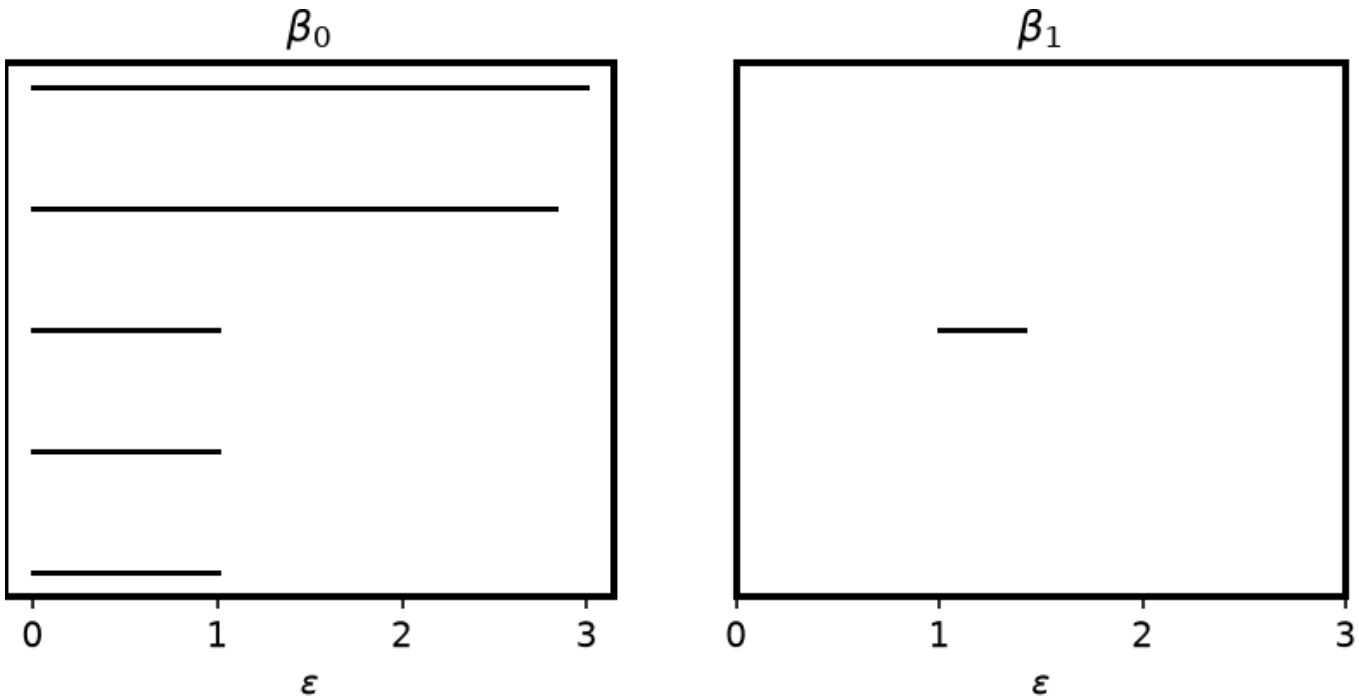
(c)
$\beta_0 = 2, \beta_1 = 0$



(d)
$\beta_0 = 1, \beta_1 = 0$

II: Barcodes

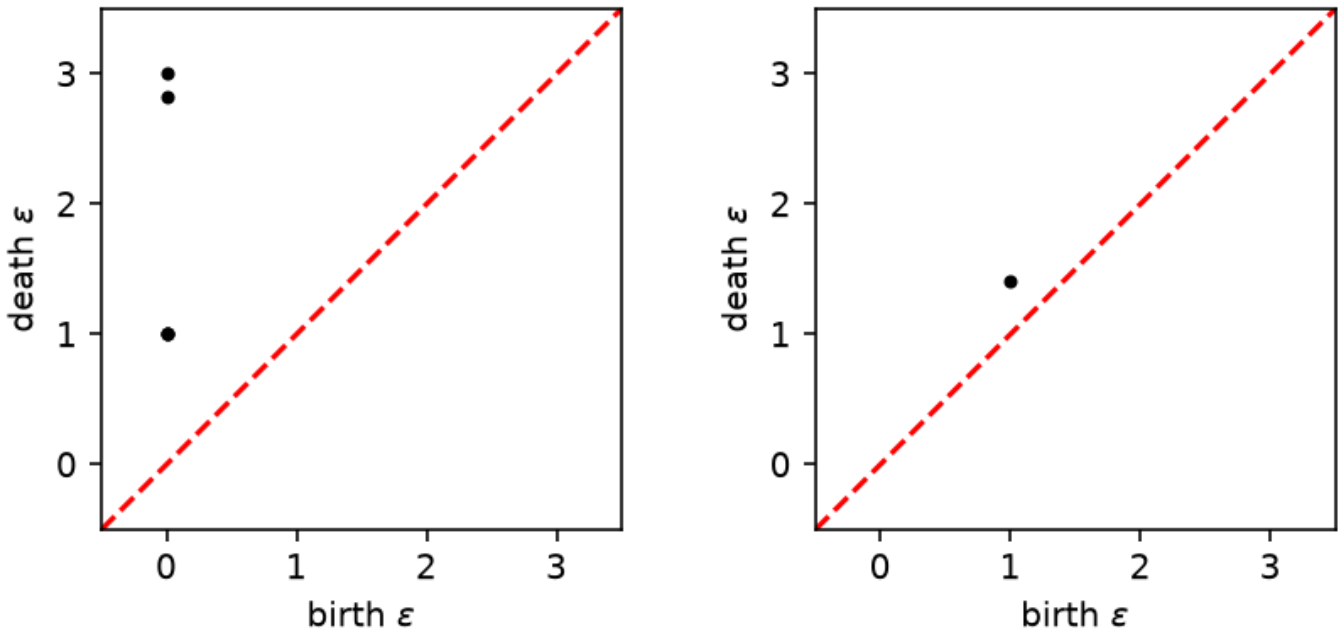
We can plot the persistence of each topological feature in a barcode graph



Left/right end point of a bar represents the value of ε where the feature comes into/out of existence, referred to as the "birth time" and "death time"

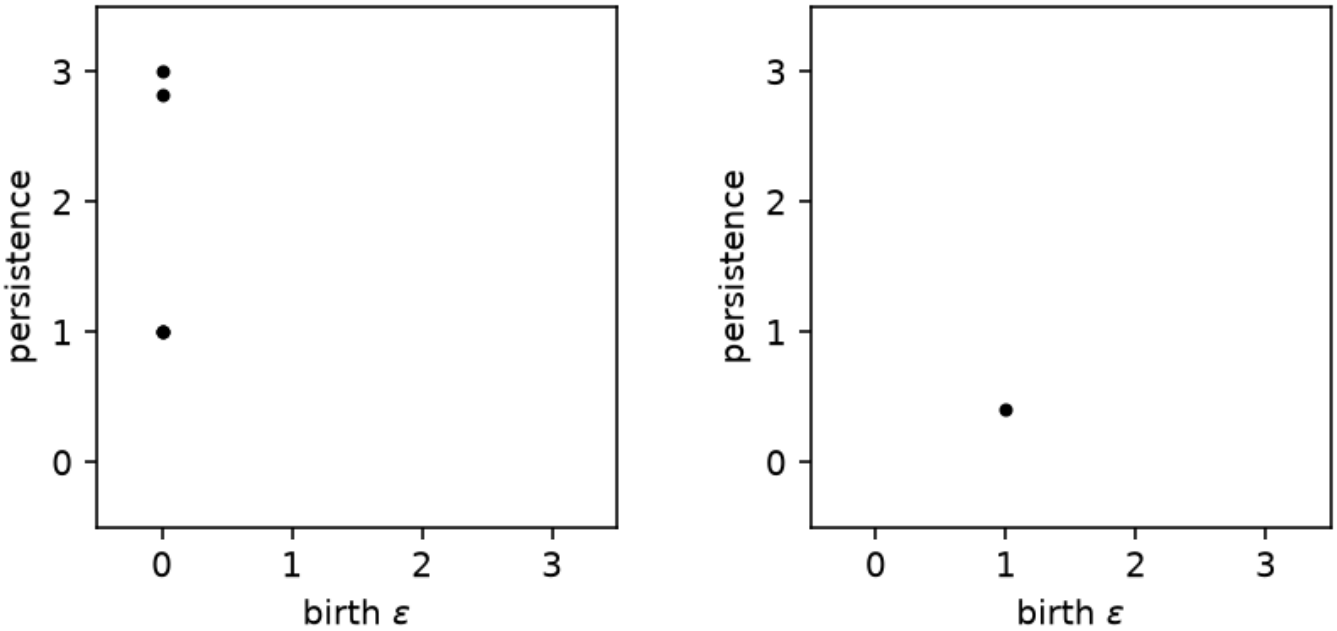
II: Birth-death diagram

Another way to visualize the topological features



II: Persistence diagram

Persistence diagram is similar: persistence = death - birth



II: Outline

I Motivation

II A non-technical primer

Munch (2017). A user's guide to topological data analysis.

Journal of Learning Analytics, 4(2), 47–61. doi: 10.18608/jla.2017.42.6

III Demonstration on zebrafish patterns

Liu & Volkening (2026). Multi-objective Bayesian inference in an agent-based model of zebrafish patterns via topological data analysis. arXiv:2605.18685

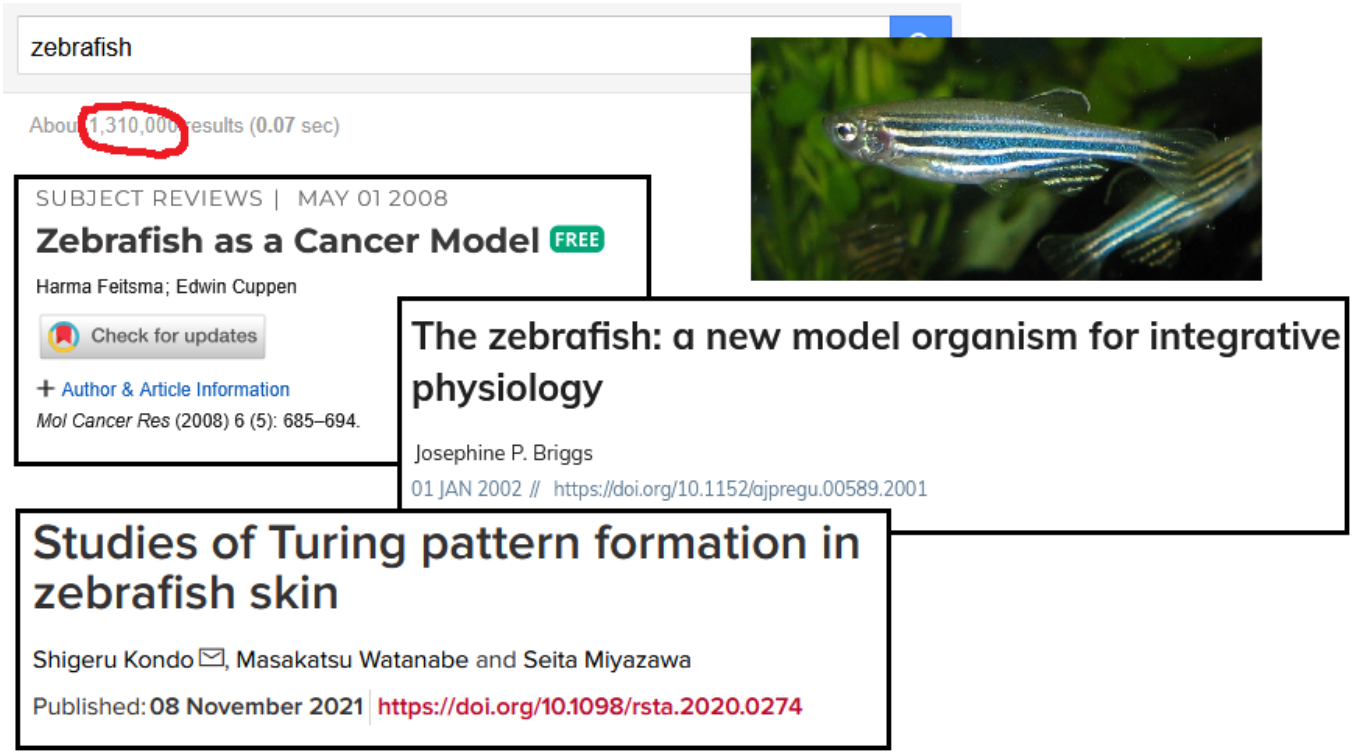
IV Mathematical foundations

Topaz et al (2015). Topological Data Analysis of Biological Aggregation Models.

PLOS ONE e0126383. doi: journal.pone.0126383

V Beyond Vietoris-Rips and persistent homology

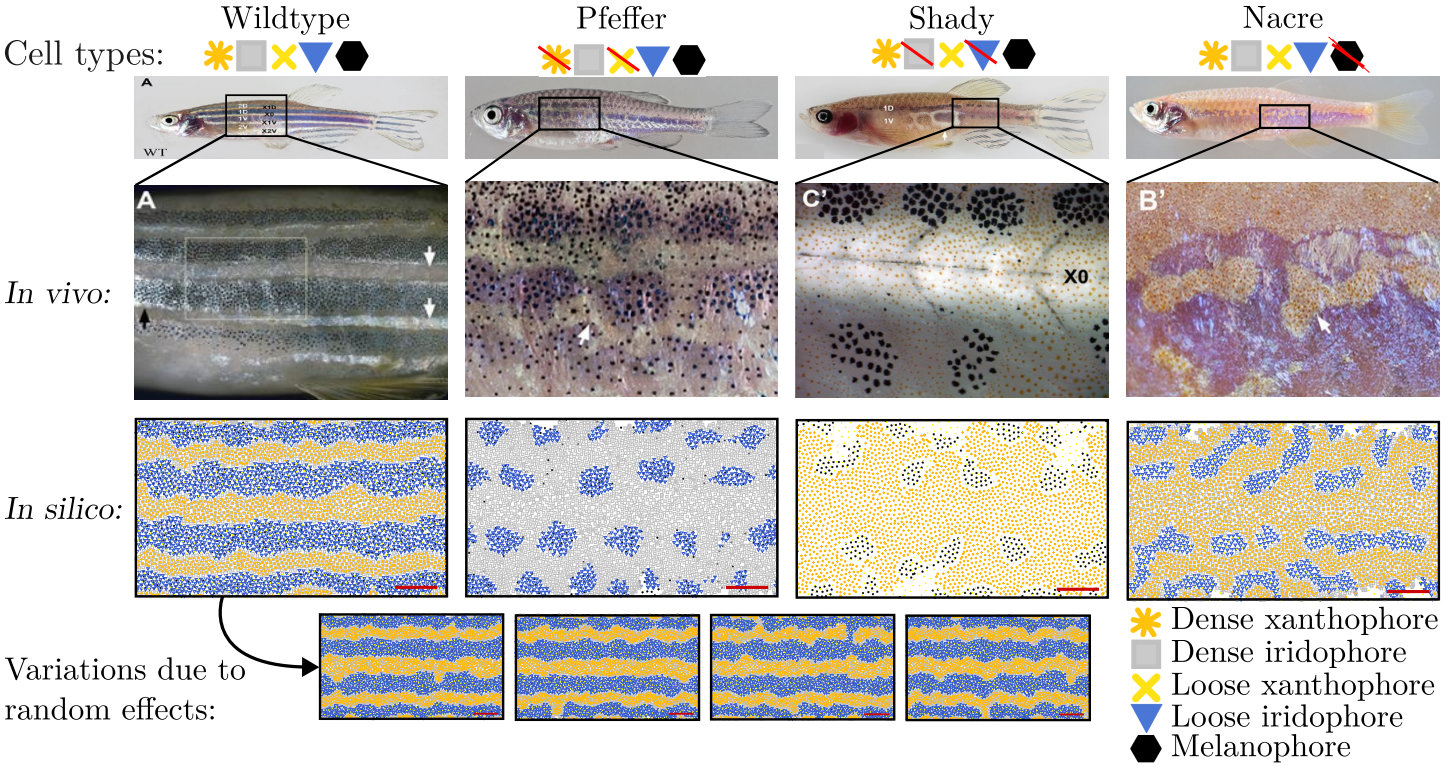
III: Example application: how do zebrafish grow their stripes?



Zebrafish photo: Azul, Wikipedia, 2005

III: Agent-based model of zebrafish stripes

Model by (Volkening & Sandstede, 2018) captures key pattern characteristics



Fish and *in vivo* pictures: Frohnhöfer et al, *Development*, 2013

III: The model is highly complex

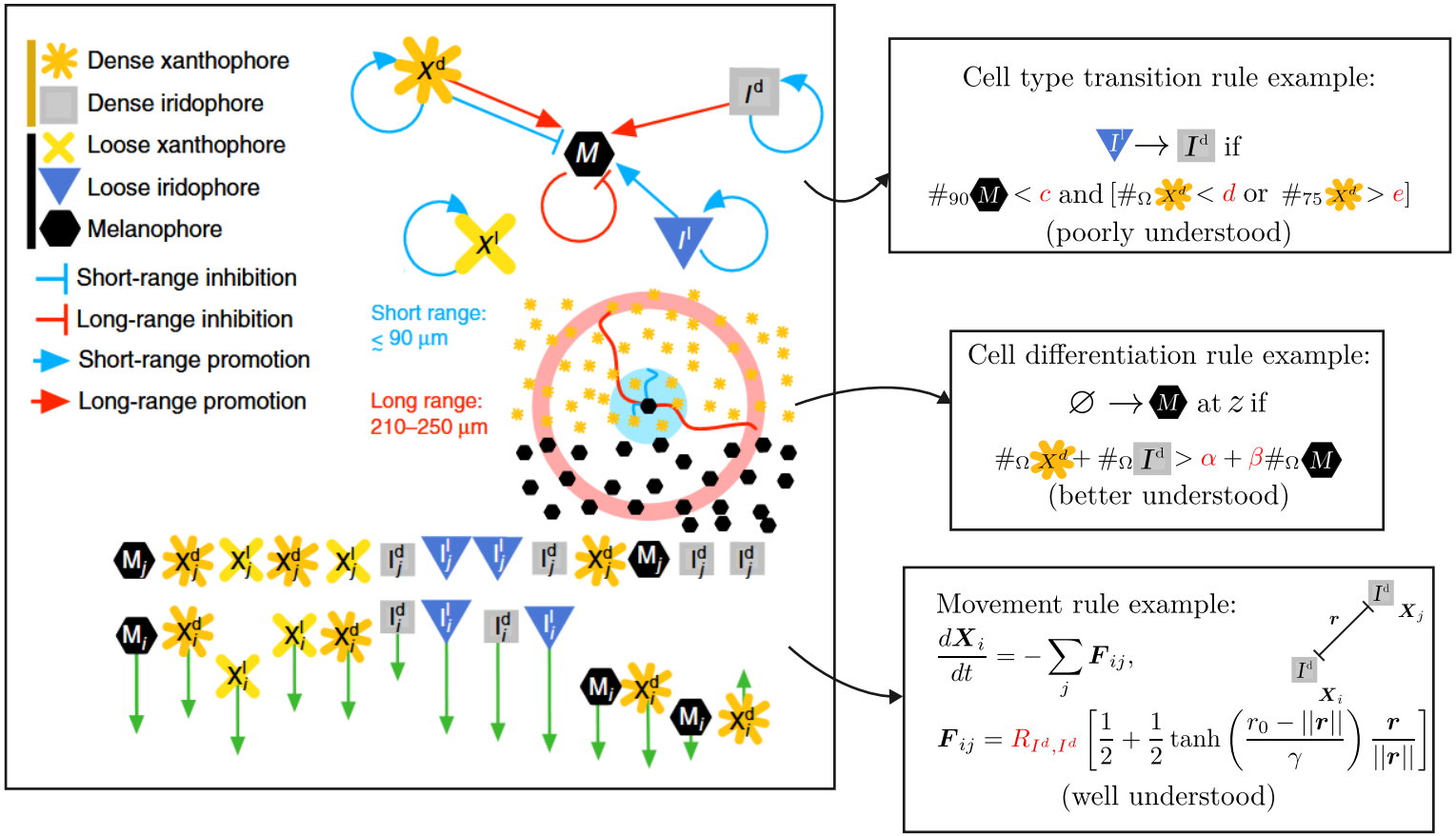
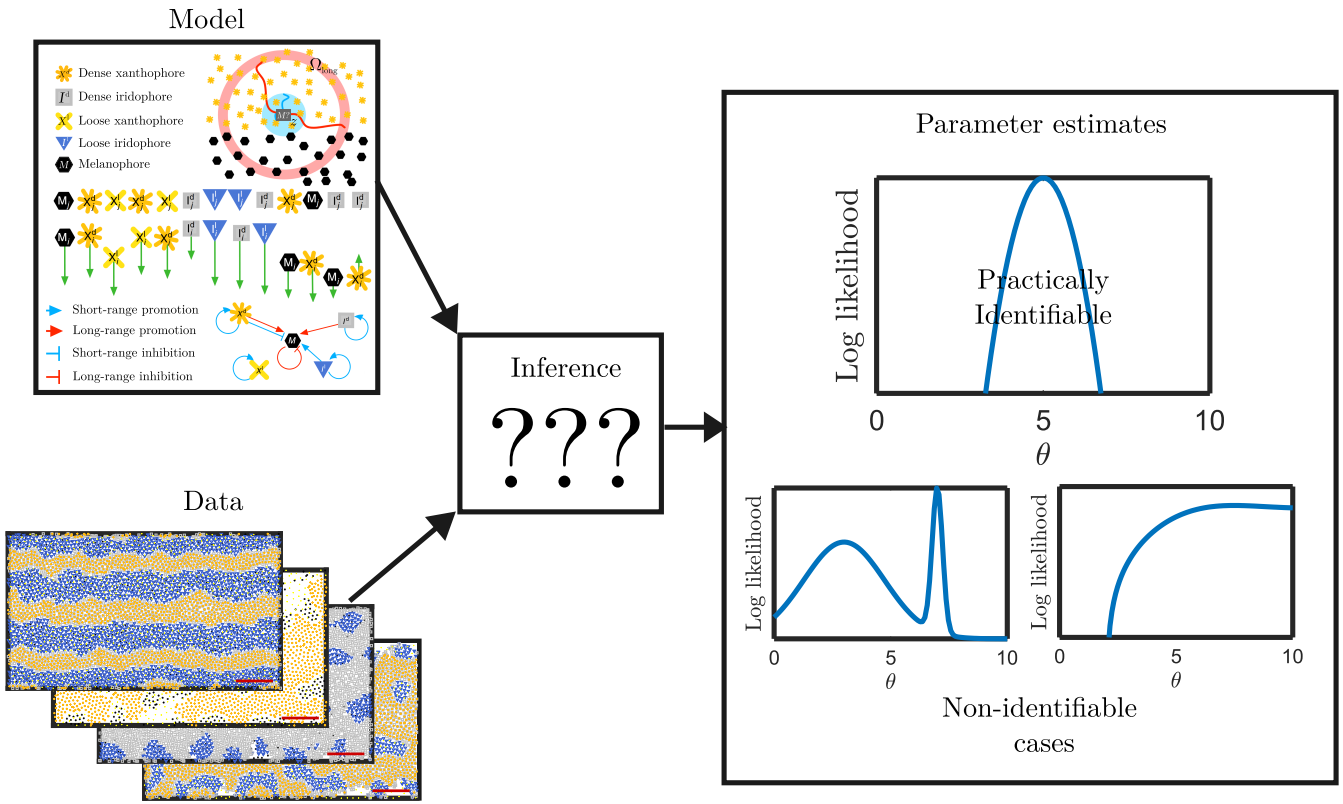


Figure adapted from Volkening & Sandstede, *Nat. Comm.*, 2018

III: Parameter inference

How to infer values of model parameters from experimental data, and quantify its uncertainty?



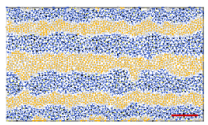
III: Inference requires comparing patterns

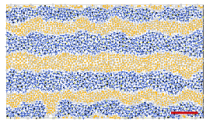
Likelihood-based approaches are intractable due to the complexity of the model

$$L(\theta) = P(x_{\text{obs}}|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{-(x_{\text{obs},i} \quad x(t_i))^2}{2\sigma^2}\right]$$

Likelihood-free based approaches requires us to be able to compare patterns

$$d(\quad,\quad) = ???$$

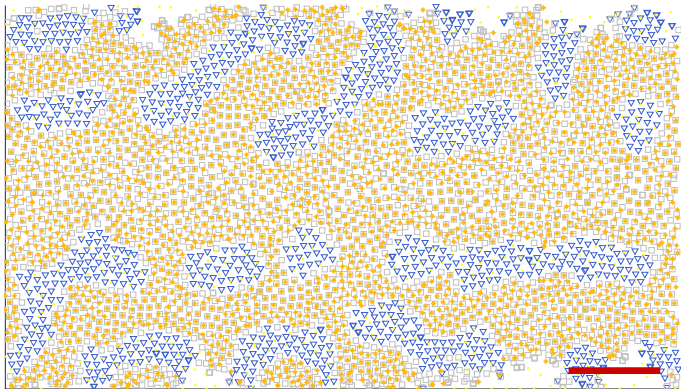


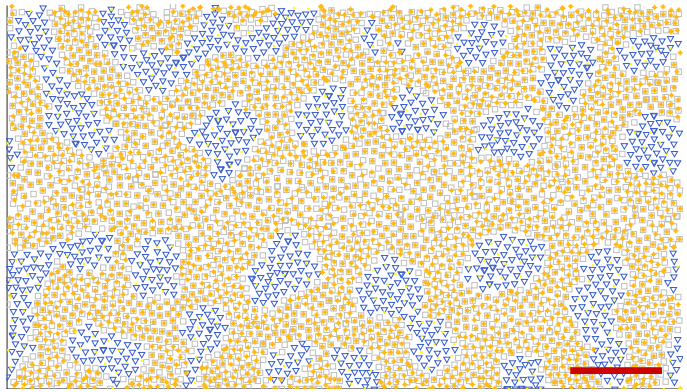


But how?

III: How to compare patterns?

Pixel-wise comparison does not respect the qualitative "essence" of patterns. Summary stats are often inadequet

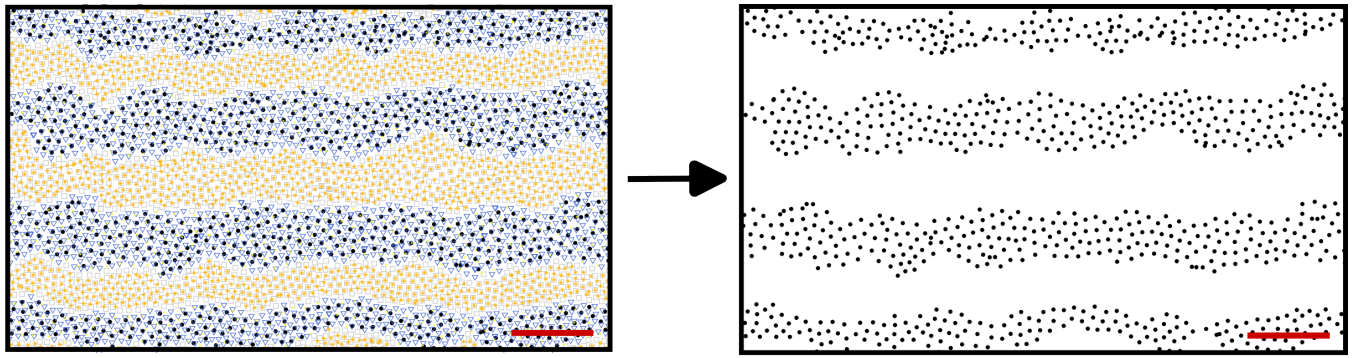




These two patterns are produced with the exact same model and parameter values, only random seed differ. Pixel-wise difference and summary stats fails to convey meaningful information about parameter values
Enters TDA!

III: Cell type selection

Step 1: extract the cell type of interest as point-cloud data from a pattern

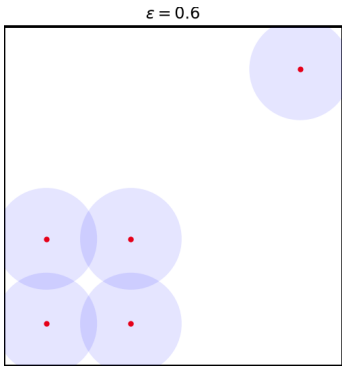


Full Pattern (wild-type simulation at t=45)

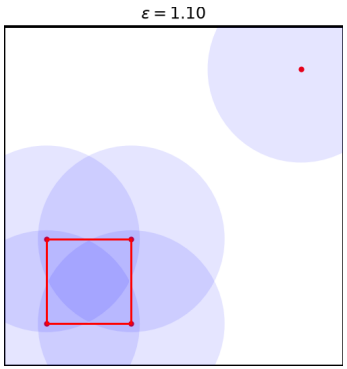
Melanophores (black pigment cells) only

III: Vietoris-Rips

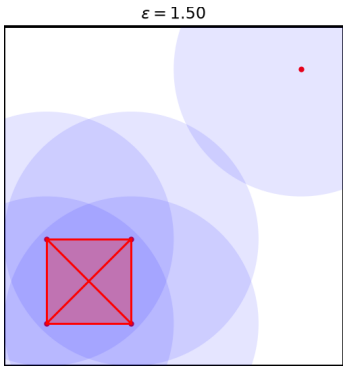
Find topological features using the Vietoris Rips filtration



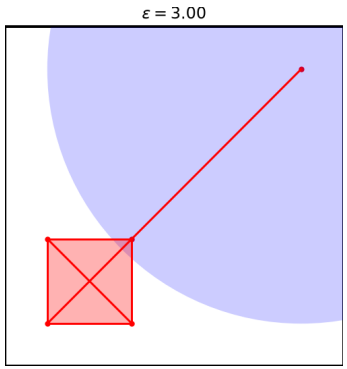
(a)
$\beta_0 = 5, \beta_1 = 0$



(b)
$\beta_0 = 2, \beta_1 = 1$



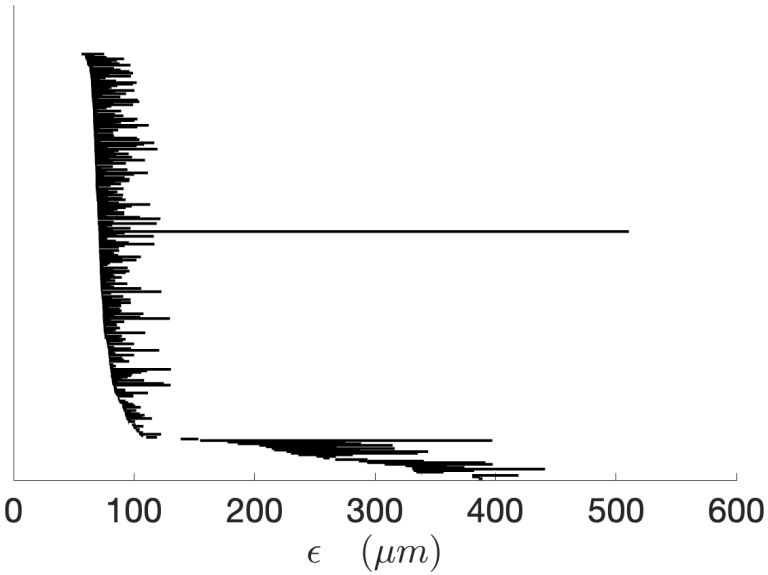
(c)
$\beta_0 = 2, \beta_1 = 0$

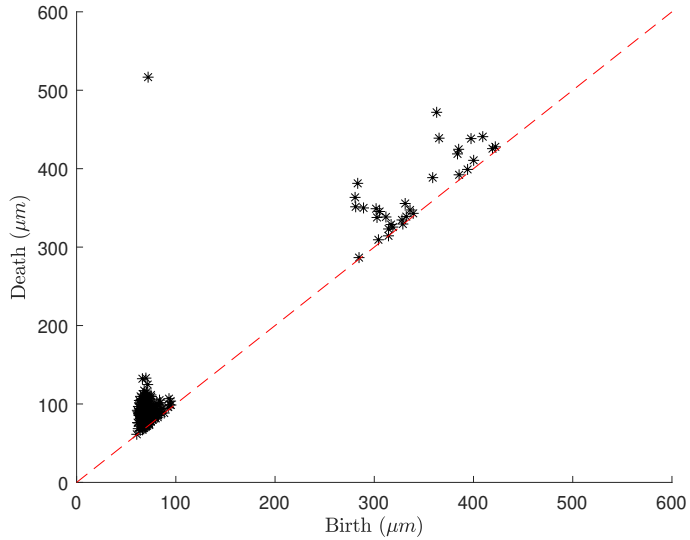


(d)
$\beta_0 = 1, \beta_1 = 0$

III: Visualize topological features

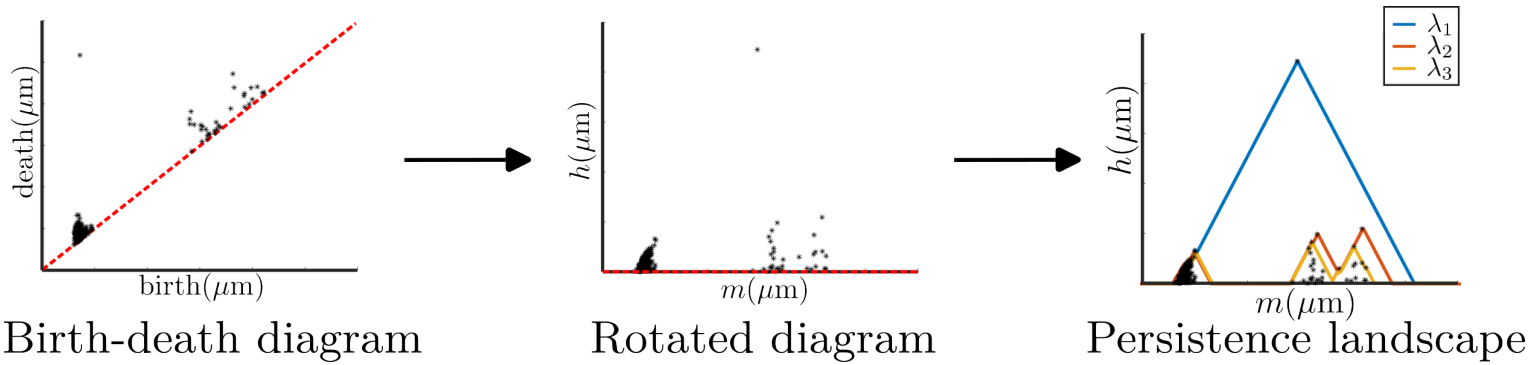
Barcodes and birth-death diagram for dim 0 features (connected components)





III: Persistence landscape

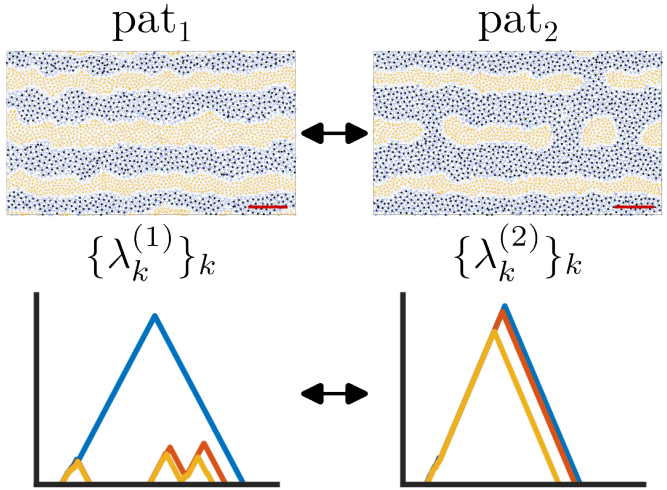
To compare two patterns, we can simply compare their corresponding barcodes. But the ordering of barcodes is not robust. Instead we compute persistence landscape



Birth-death diagram | Rotated diagram | Persistence landscape

III: Metric for comparing patterns

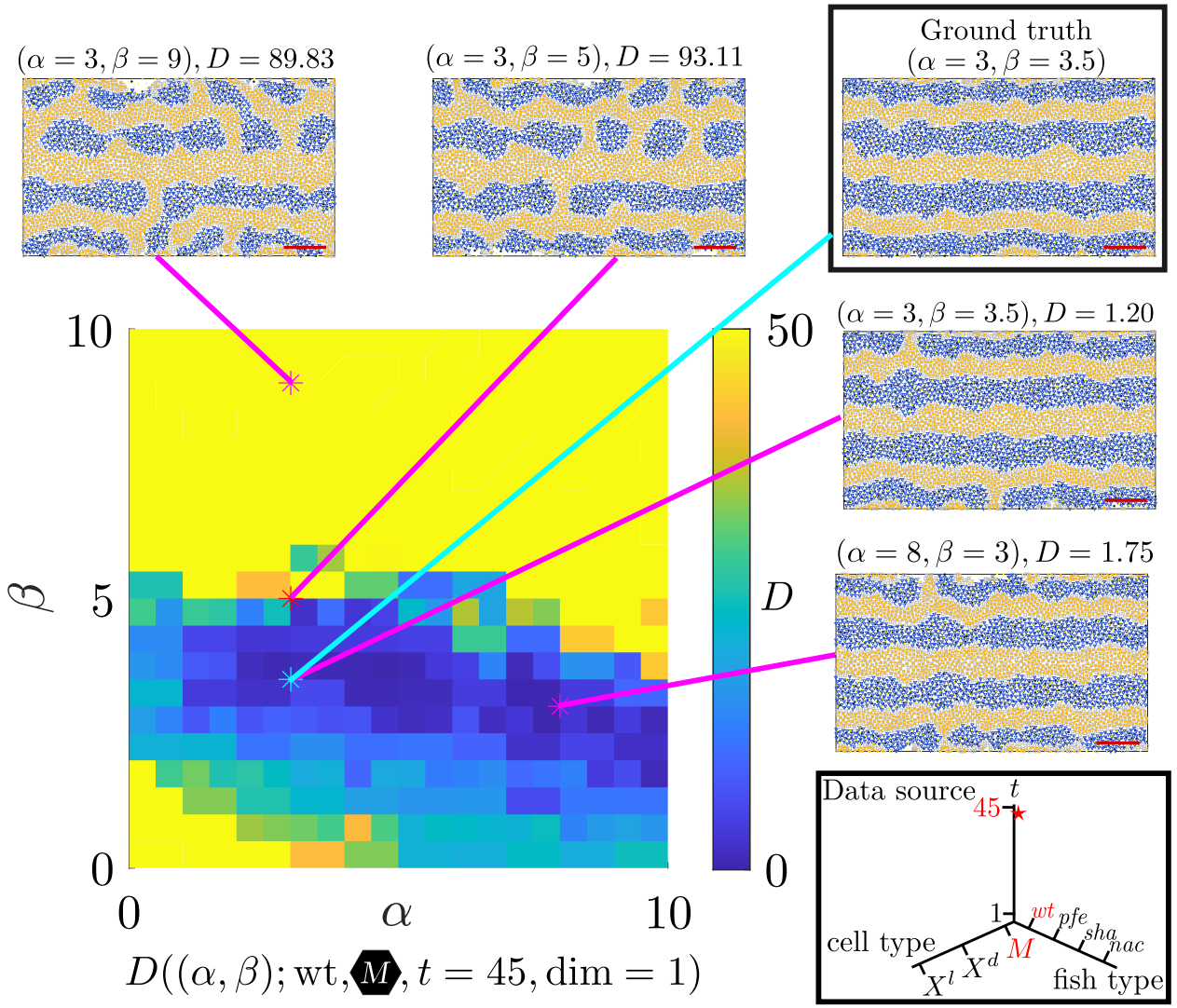
Persistence landscape offers a robust way to compare spatial data (Bubenik 2015, 2020), and consequently define score function for parameter values



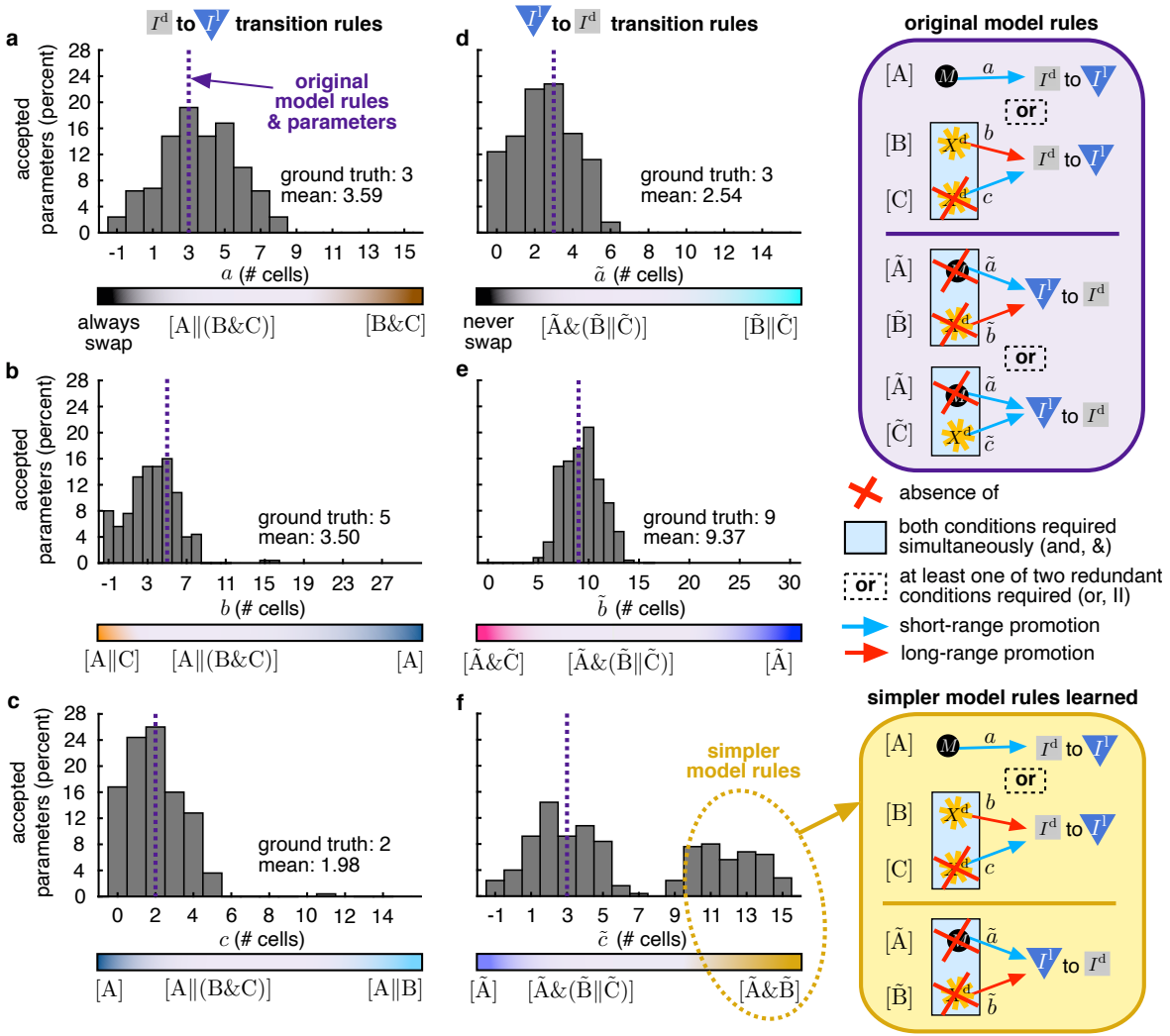
$$d_{\mathrm{land}}(\mathrm{pat}_1, \mathrm{pat}_2)^2 = \sum_{k=1}^{\infty} d_{L^2}\left(\lambda_k^{(1)}, \lambda_k^{(2)}\right)^2$$

$$D(\theta) = d_{\mathrm{land}}(\mathrm{pat}(\theta), \mathrm{pat}_{\mathrm{data}}), \quad \mathrm{pat}_{\mathrm{data}} \text{ is synthetic}$$

III: TDA captures key qualitative characteristics



III: TDA helped us to discover parameter values corresponding to alternative mechanisms



IV: Outline

I Motivation

II A non-technical primer

Munch (2017). A user's guide to topological data analysis.

Journal of Learning Analytics, 4(2), 47–61. doi: 10.18608/jla.2017.42.6

III Demonstration on zebrafish patterns

Liu & Volkening (2026). Multi-objective Bayesian inference in an agent-based model of zebrafish patterns via topological data analysis. arXiv:2605.18685

IV Mathematical foundations

Topaz et al (2015). Topological Data Analysis of Biological Aggregation Models.

PLOS ONE e0126383. doi: journal.pone.0126383

V Beyond Vietoris-Rips and persistent homology

IV: What is a topology, anyways?

Topology tries to capture the qualitative structures (shape, connectedness, holes, boundaries, continuity) without relying on a metric, by specifying τ, the collection of open subsets of X

Definition

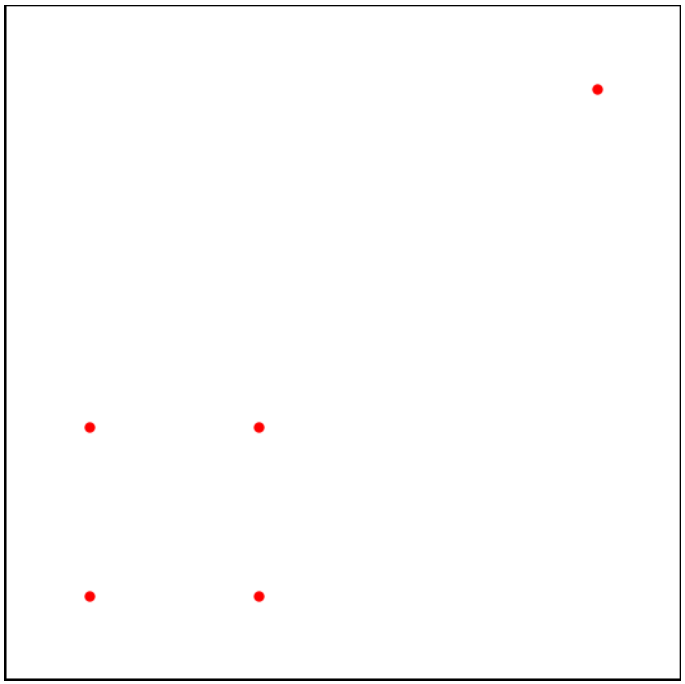
A topology (aka topological space) is (X, τ), where X is a set and $\tau \subseteq \mathcal{P}(X)$, such that

- $\emptyset \in \tau,\ X \in \tau$
- $A_i \in \tau \Rightarrow \bigcup_{i \in I} A_i \in \tau$ (closure under arbitrary union)
- $A_i \in \tau \Rightarrow \bigcap_i^n A_i \in \tau$ (closure under finite intersection)

Note: A metric space (X, d) always induces a topology (X, τ_d) where τ_d is generated by open balls under d

IV: What is the topology of a point cloud?

Under the Euclidean metric, the topology of a point cloud $X = \{x_i\}_{i=1}^{N}$ is the discrete topology:
every point is alone in its neighbourhood,
τ consists of every possible subset: $\tau = \mathcal{P}(X)$



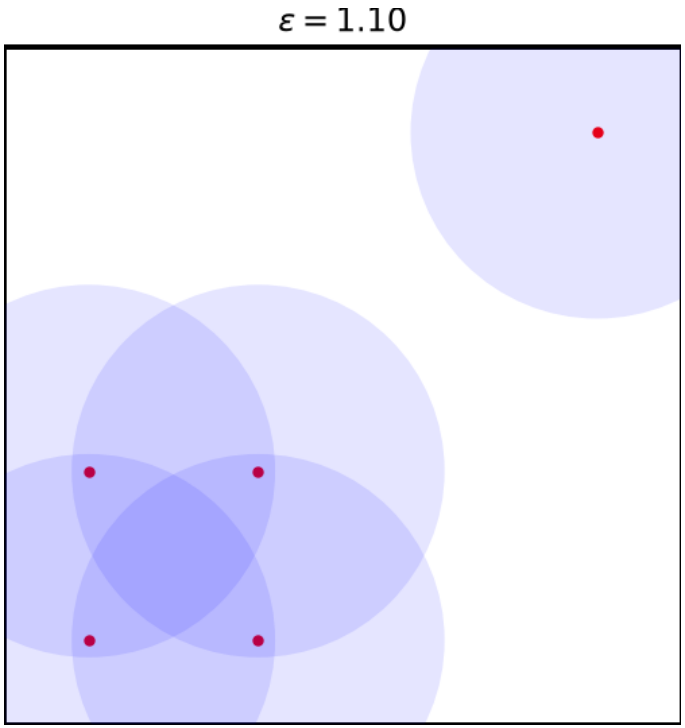
This tells us nothing interesting about the shape of the data

IV: Union of open balls

Idea: Instead of seeing X as a point cloud, consider

$$U_\varepsilon(X) = \bigcup_{i=1}^{N} B(x_i, \varepsilon)$$

Then $(U_\varepsilon(X), \tau_{L^2})$ induced by the Euclidean L^2 metric on $\mathbb{R}^n$ is a much more interesting topology



IV: From U_ε to Vietoris-Rips

$(U_\varepsilon(X), \tau_{L^2})$ is difficult to work with. Introduce Simplicial Complex:

Definition

A $k-$Simplex is a k-dimensional polytope that is the convex hull of its k + 1 vertices.
Example: $k = 0$: point; $k = 1$: line segment; $k = 2$: triangle; $k = 3$: tetrahedron...

Definition

A Face of simplex σ is a simplex formed by a subset of vertices of σ
Example: e.g. a triangle has itself ($k = 2$), 3 edges ($k = 1$), 3 vertices ($k = 0$) as faces

Definition

A Simplicial Complex $\mathcal{K}$ is a structured set of simplices such that

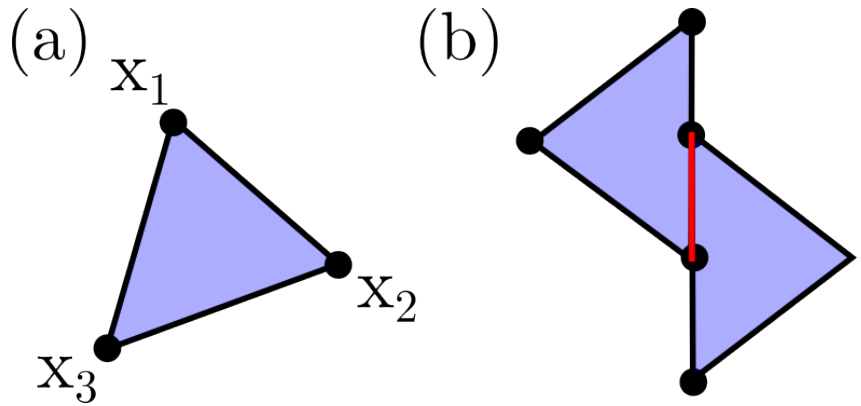
- Every face of simplex $\sigma \in \mathcal{K}$ is also in $\mathcal{K}$
- If $\sigma_1, \sigma_2 \in \mathcal{K}$ and $\sigma_1 \cap \sigma_2 \neq \emptyset$, then $\sigma_1 \cap \sigma_2$ is a face of both σ_1, σ_2

IV: Simplicial Complex

Definition

A Simplicial Complex $\mathcal{K}$ is a structured set of simplices such that

- Every face of simplex $\sigma \in \mathcal{K}$ is also in $\mathcal{K}$
- If $\sigma_1, \sigma_2 \in \mathcal{K}$ and $\sigma_1 \cap \sigma_2 \neq \emptyset$, then $\sigma_1 \cap \sigma_2$ is a face of both σ_1, σ_2



(a) First condition says if triangle $[123] \in \mathcal{K}$, then line segments $[12], [23], [13]$ and points x_1, x_2, x_3 must also be in $\mathcal{K}$
(b) is not a proper simplicial complex because the intersection of the two triangle (red line segment) is not a face of either triangle

IV: Čech and Vietoris-Rips

Recall: The Vietoris-Rips Complex $VR_\varepsilon(X)$ is constructed by

1. Add edge $[ij]$ if $d(x_i, x_j) < \varepsilon$
2. Add simplex $[i_1, \ldots, i_k]$ if its vertices are pairwise joined

The Čech Complex $\check{C}_\varepsilon(X)$ is very similar:

1. Add edge $[ij]$ if $d(x_i, x_j) < 2\varepsilon$
2. Add simplex $[i_1, \ldots, i_k]$ if $\bigcap_{l=1}^{k} B(x_{i_l}, \varepsilon) \neq \emptyset$

Both are instances of Simplicial Complices.

Theorem (Nerve Theorem)

$U_\varepsilon(X) \simeq \check{C}_\varepsilon(X)$, i.e. the Čech Complex is homotopy-equivalent to the union of balls, i.e. they essentially have the same topological features (under the topology induced by L^2 metric)

Theorem (VR approximates Č)

$\check{C}_\varepsilon(X) \subseteq VR_{2\varepsilon}(X) \subseteq \check{C}_{2\varepsilon}(X)$.

VR is much cheaper to compute, making it the practical choice

IV: From simplicial complex to homology

Given a simplicial complex $\mathcal{K}$ (for us $\mathcal{K} = VR_\varepsilon(X)$), let field $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$, define:

- Chain space $C_k = \mathrm{span}\{s : s \in \mathcal{K}, \dim(s) = k\}$
- Boundary maps $\partial_k : C_k \to C_{k-1}$

$$\partial_k[v_0, \ldots, v_k] = \sum_{i=0}^{k} [v_0, \ldots, \widehat{v_i}, \ldots, v_k]$$

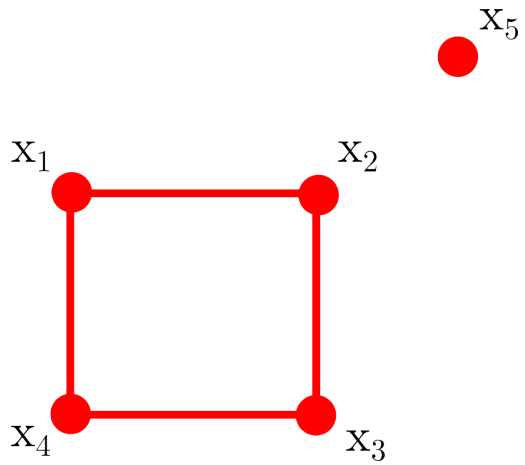
$$C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \to \cdots \to C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} \{0\}$$

- Cycles $Z_k = \ker(\partial_k)$
- Boundary spaces $B_k = \mathrm{im}(\partial_{k+1})$
- Homology group $H_k = Z_k/B_k$, basis of H_k corresponds to topological features
- Betty number $\beta_k = \dim(H_k) = \#$topo-features

Note that C_k, Z_k, B_k and H_k are both vector spaces over $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$, and group (with group operation $+$)

IV: Concrete example

Let's do a concrete example to compute homology group of



Simplicial Complex:

$$\mathcal{K} = \{x_1, x_2, x_3, x_4, x_5,$$
$$[12], [23], [34], [41],$$
$$(\text{no higher dimensional simplices}) \}$$

IV: Concrete example

Chain spaces:

$$C_0 = \text{span}\{x_1, \ldots, x_5\} \cong \mathbb{F}^5$$
$$C_1 = \text{span}\{[12], [23], [34], [41]\} \cong \mathbb{F}^4$$
$$C_k = \{0\} \text{ for } k \geq 2$$

To be more explicit:

$$C_0 = \left\{ \sum_{i=1}^{5} a_k x_k : a_k \in \mathbb{F} \right\}$$

Note that the span is a formal sum, where x_i is identified with $e_i \in \mathbb{F}^5$. We are not adding up the point's coordinates!

IV: Boundary map

$\partial_k : C_k \to C_{k-1}$ returns the boundary of a chain

- $\partial_0(x_i) = 0$ for all i (points have no boundaries)
- $\partial_1([ij]) = x_i + x_j \in C_0$
- $\partial_k(0) = 0, k \geq 2$ (note $C_k = \{0\}$ for $k \geq 2$ so ∂_k is trivial)
- Extend ∂_k linearly to C_k, i.e.:

$$\partial_1 \left(\sum_{ij} a_{ij}[ij] \right) = \sum a_{ij}(x_i + x_j)$$

∂_1 is the only non-trivial map, can be expressed in matrix form:

$$\partial_1 \sim \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cccc} [12] & [23] & [34] & [41] \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \sim \mathbb{F}^{5\times 4}$$

IV: Cycles and boundaries

Cycles $Z_k = \ker(\partial_k)$ corresponds to "chain with no boundary"

- $Z_0 = \ker(\partial_0) = \mathrm{span}\{x_1, \ldots, x_5\} = C_0 \cong \mathbb{F}^5$
- $Z_1 = \ker(\partial_1) = \mathrm{span}\{[12] + [23] + [34] + [41]\} \cong \mathbb{F}^1$
- $Z_k = \{0\}$ for $k \geq 2$

Boundary spaces $B_k = \mathrm{im}(\partial_{k+1})$ "chains that bounds a higher dimensional chain"

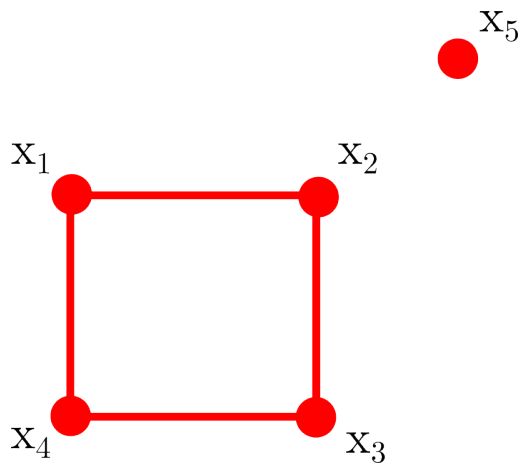
- $B_0 = \mathrm{im}(\partial_1) = \mathrm{span}\{x_1 + x_2, x_2 + x_3, x_3 + x_4, x_4 + x_1\} \cong \mathbb{F}^3$ (note we only have 3 linear independent basis element)
- $B_k = \mathrm{im}(\partial_k) = \{0\}$ for $k \geq 1$

Homology group $H_k = Z_k / B_k$ are the topological features

- $H_0 = Z_0 / B_0 \cong \mathbb{F}^5 / \mathbb{F}^3 \cong \mathbb{F}^2$, has basis $\{x_1, x_5\}$. Hence we have 2 connected components, one represented by x_1, the other by x_5 and Betty number $\beta_0 = 2$
- $H_1 = Z_1 / B_1 = Z_1 \cong \mathbb{F}^1$, corresponding to the single loop. Hence Betty number $\beta_1 = 1$
- $H_k = \{0\}$ for $k \geq 2$, we have no higher dimensional topo-features and $\beta_k = 0$

IV: Concrete example

We have jumping through a lot of hoops just to compute $\beta_0 = 2,\ \beta_1 = 1$ for



But these formalisms is easy for a computer: all spaces involved are linear spaces and ∂_k are linear maps, reducing the whole problem to linear algebra, which computers are very good at

V: Outline

I Motivation

II A non-technical primer

Munch (2017). A user's guide to topological data analysis.

Journal of Learning Analytics, 4(2), 47–61. doi: 10.18608/jla.2017.42.6

III Demonstration on zebrafish patterns

Liu & Volkening (2026). Multi-objective Bayesian inference in an agent-based model of zebrafish patterns via topological data analysis. arXiv:2605.18685

IV Mathematical foundations

Topaz et al (2015). Topological Data Analysis of Biological Aggregation Models.

PLOS ONE e0126383. doi: journal.pone.0126383

V Beyond Vietoris-Rips and persistent homology

V: Persistent homology with other filtration

VR is good for point-cloud datasets. Other ways to construct topological spaces are useful for other kinds of data:

- Sweeping plane: good for networks/branching structures (e.g. vasculature)



- Cubical complex: good for pixel images or data on regular lattice (e.g. medical images)
- Alpha complex: alternative for point cloud, can perform better if data lie in low dimension
- many others

Different constructions yield different filtered complexes, but computation for homology and downstream analysis is similar.

V: Beyond persistent homology

Persistent homology is by far the most common TDA tool. But there are others:

- Mapper: summarizes high-dimensional data as a graph to visualize branches, loops, and connected regions in the data, akin to graph-based clustering
- Reed graph: Summarize a space using the connected components of level sets of a function, capturing how they split and merge as the level cut-off changes
- Morse theory: Studies the topology of a space by looking at critical points (max, min, saddles) of a function defined on it
- Many others

TDA is a large and diverse set of approaches. Tools are constantly being brought out of the depth of pure math into the applied world. Good review paper:

Su et al (2025), Topological data analysis and topological deep learning beyond persistent homology: a review.

Artif Intell Rev 59, 58 (2026)