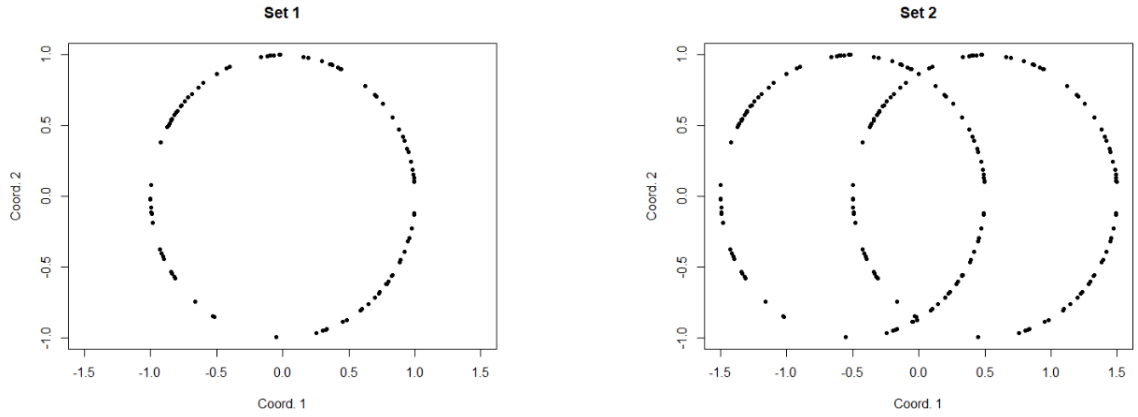
A Sketch of Topological Data Analysis for Statistics

Benjamin Beaudett STA-6557

April 12, 2022

Motivation: Detecting the Shape of a Manifold



Set	Sample Mean	Sample Cov.
1	$(0.0341 \quad 0.1038)^T$	$\begin{pmatrix} 0.5597 & -0.0768 \\ -0.0768 & 0.4384 \end{pmatrix}$
2	$(0.0341 \quad 0.1038)^T$	$\begin{pmatrix} 0.8081 & -0.0764 \\ -0.0764 & 0.4362 \end{pmatrix}$

Points sampled from a circle and the union of two circles. Simple summary statistics do not distinguish between them.

Images and table from Dixon et al. [4]

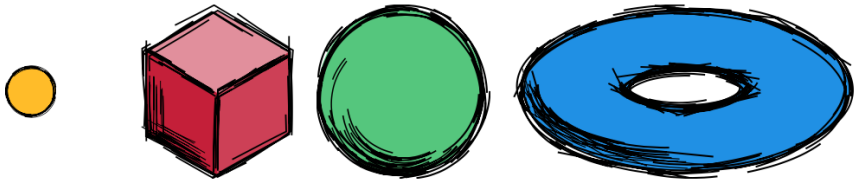
Approach: Classify Shapes Using Topological Features

Betti numbers

The d^{th} Betti number counts the number of d-dimensional holes. It can be used to distinguish between spaces.

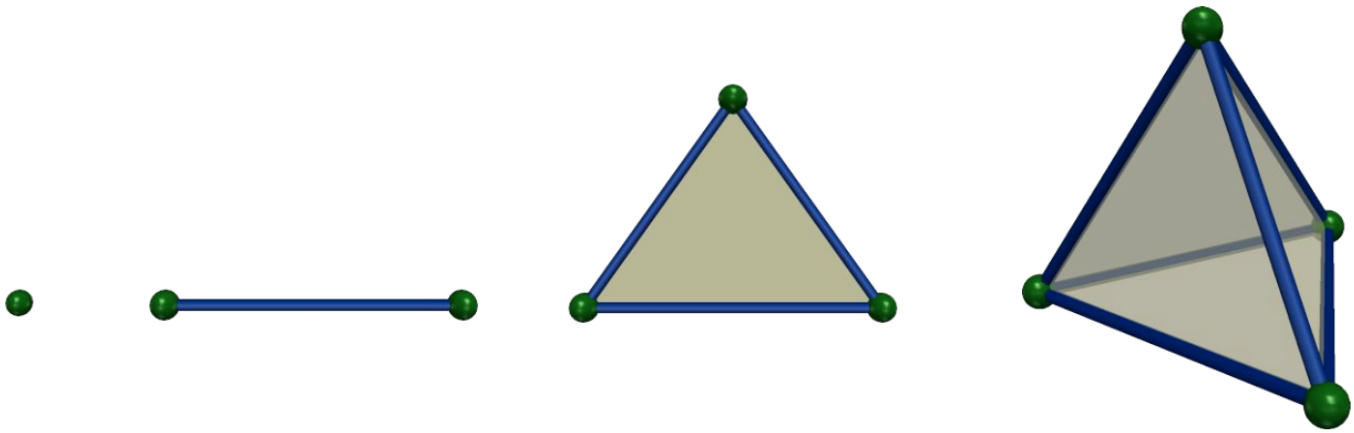
β_0 Connected components
β_1 Tunnels
β_2 Voids

Space	β_0	β_1	β_2
Point	1	0	0
Cube	1	0	1
Sphere	1	0	1
Torus	1	2	1



Slide from Rieck [6]

Simplices and Simplicial Complexes



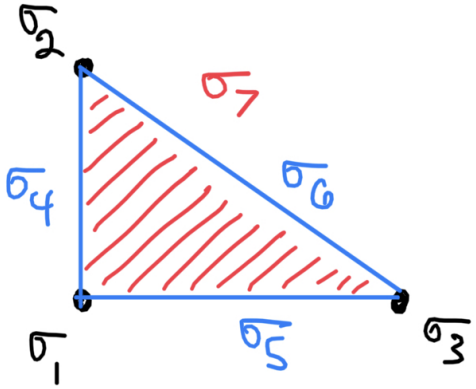
Simplices of dimension 0, 1, 2, and 3.

p-Simplex σ: The convex hull of $p+1$ points $\sigma = [v_0\, v_1 \cdots v_p]$ with $v_i \in \mathbb{R}^d$, $p \leq d$ which are affinely independent ($\{v_i - v_0\}_{i=1}^{p}$ forms a linearly independent set).

Face of a p-simplex σ: The simplex formed by the convex hull of a nonempty subset of the vertices of σ.

Proper faces of a p-simplex σ: Each of the $(p-1)$-simplices formed from removing exactly one of the vertices of σ and preserving the order of those remaining.

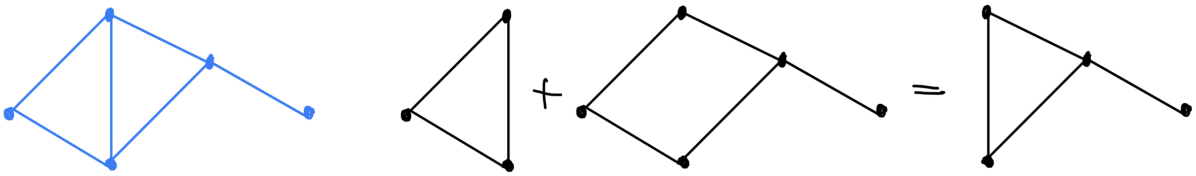
Simplicial complex K: A collection of simplices such that every face of a simplex from K is also in K, and the non-empty intersection of any two simplices in K is a face of both of those simplices.



Computer-Generated Images from Tierny [7]

Chains, Cycles, Boundaries, and Homology

The p-simplices of a simplicial complex K form the basis of an Abelian group C_p of p-chains under the operation + using coefficients in $\mathbb{Z}/2\mathbb{Z}$ so that for any $\sigma \in K$, $\sigma + \sigma = 0$.



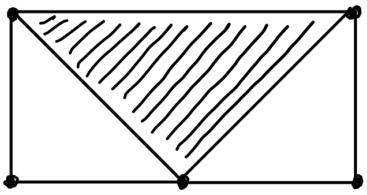
The boundary operator $\partial_p : C_p \to C_{p-1}$ maps a p-chain $c = \sum \sigma_i$ to the sum of all of the proper faces of all of the σ_i included in it.

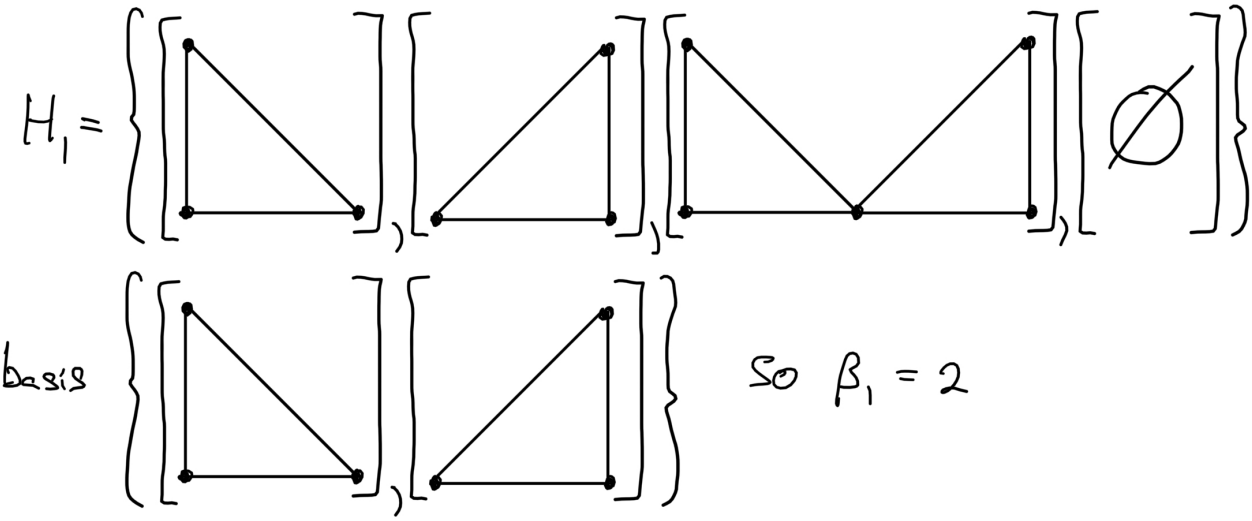
The cycle group Z_p is the kernel of ∂_p. The boundary group B_p is the image of ∂_{p+1}.

$B_p \subseteq Z_p \subseteq C_p$.

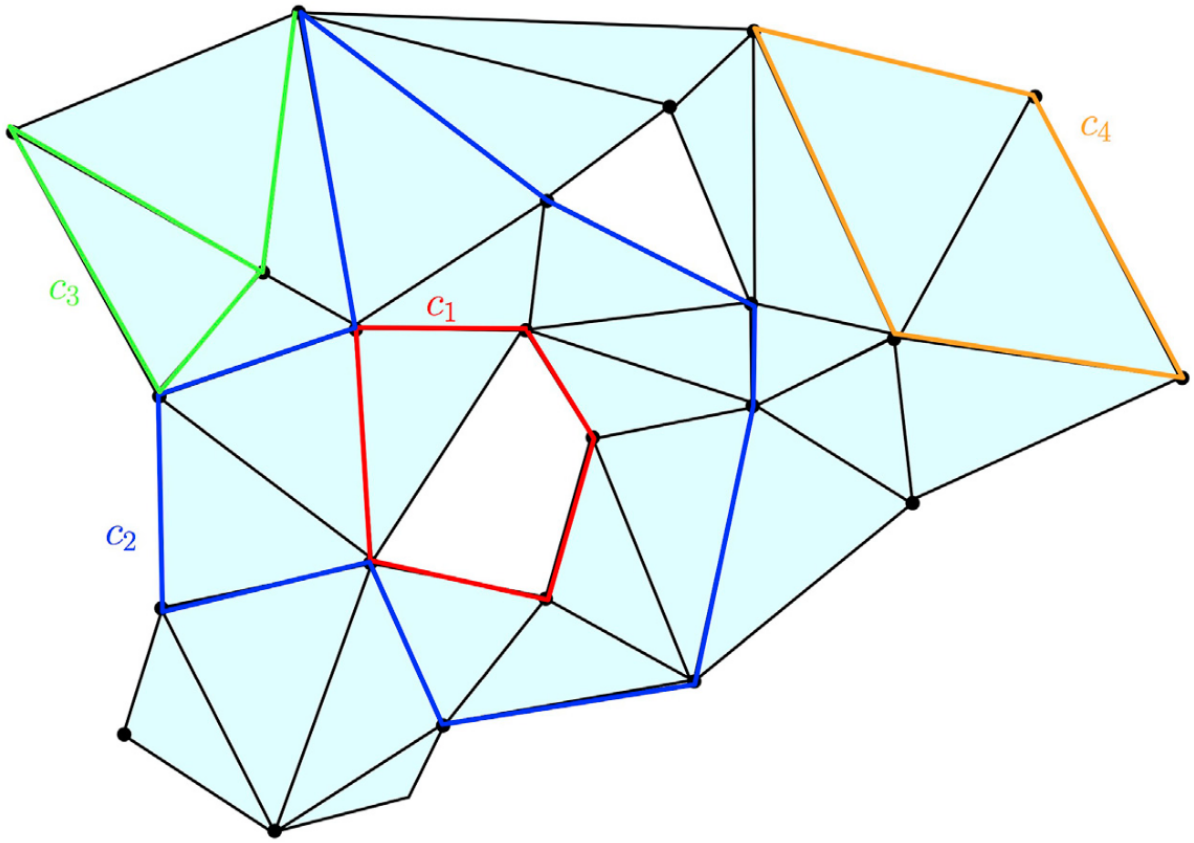
The p-th simplicial homology group is defined $H_p = Z_p/B_p$. Two cycles are homologous if $\zeta_1 + \zeta_2 \in B_p$.

The p-th betti number is defined $\beta_p = \dim(H_p)$.

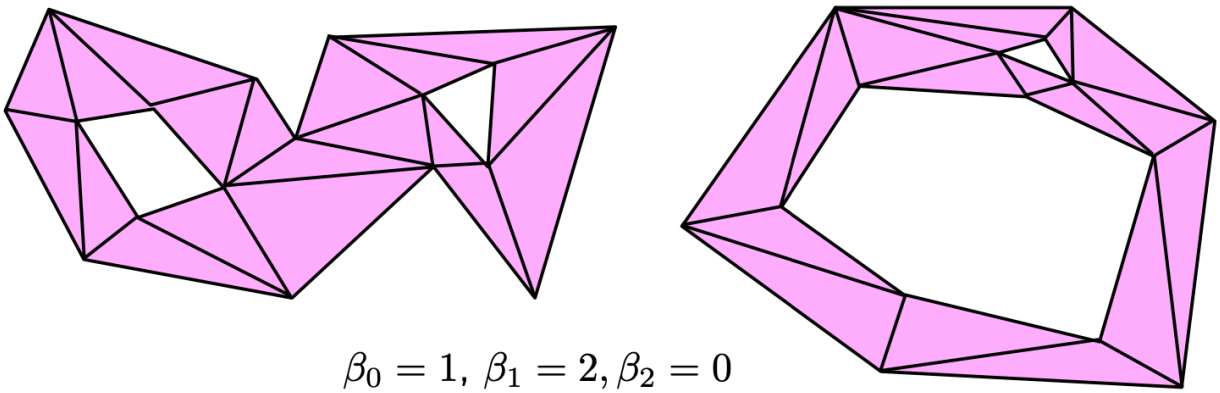




Chains, Cycles, Boundaries, and Homology



All c_i are one-chains. c_3 is not a cycle but the others are. c_4 is a boundary, but c_1 and c_2 are not. c_1 and c_2 are homologous to each other.



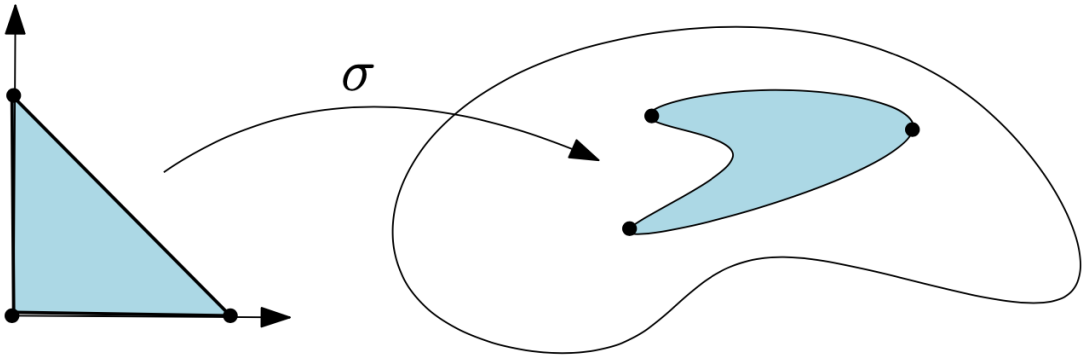
The Betti numbers of these simplicial complexes are equal and their homology groups are isomorphic.

Images from Chazal and Michel [3] and Chazal [2]

Topological Invariance and Singular Homology

The standard p-simplex is $\Delta_p = \{(t_0, t_1, \ldots, t_p) \in \mathbb{R}^{p+1} : t_i \geq 0,\ \sum_{i=0}^{p} t_i = 1\}$.

A singular p-simplex in a topological space M is a continuous map $\sigma : \Delta_p \to M$.



Singular homologies can be constructed in the same way as simplicial homologies. If M is a triangulable space, the pth simplicial and singular homology groups are isomorphic for all p.

So work can be done in terms of simplicial complexes even if the underlying manifold of interest is not one.

The triangulation of a manifold, called a piecewise linear manifold.

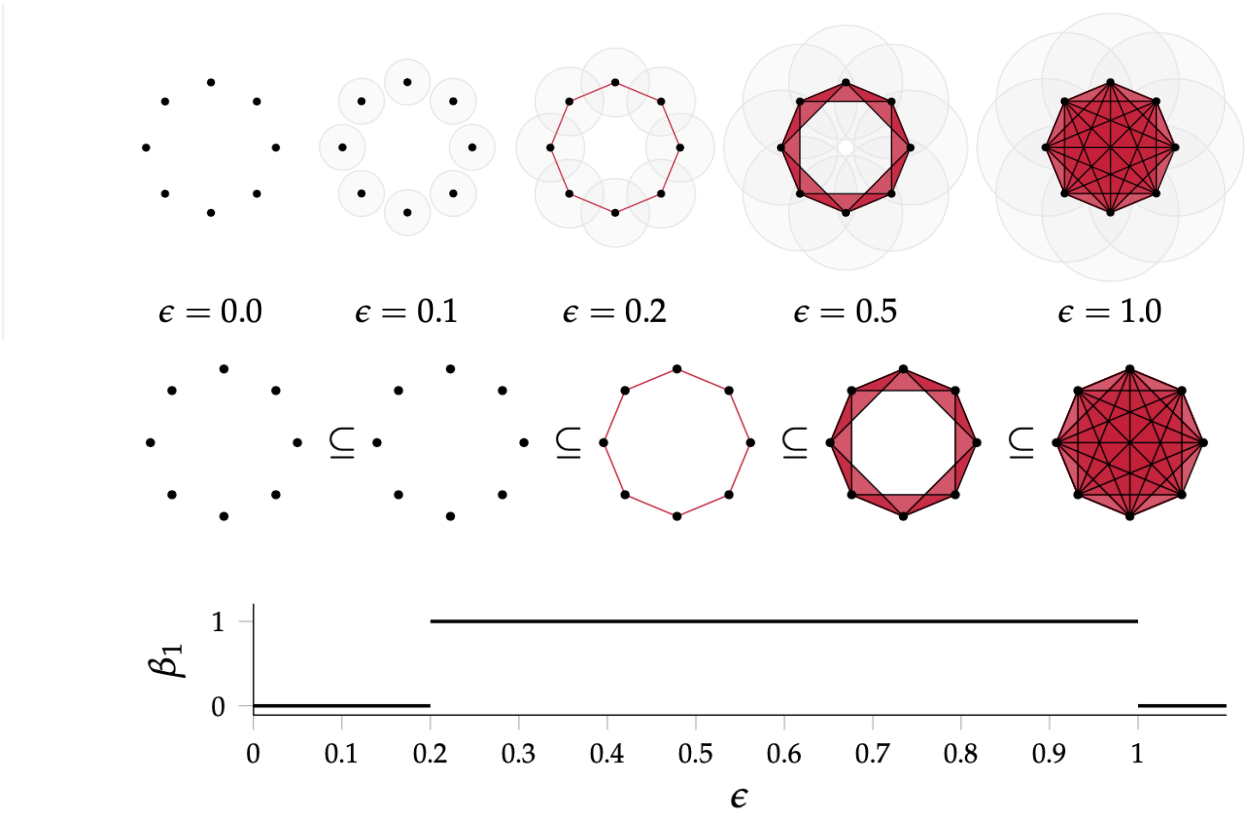
Images from Chazal [2] and Tierny [7], theorem from Dixon et al. [4]

The Vietoris-Rips Complex and Persistent Homology

A sample of data points can be used to define a large number of different simplicial complexes. Rather than picking one, work with a nested sequence of complexes (a filtration). The Vietoris-Rips complex $\mathcal{R}_\epsilon$ for a set of points in $\mathbb{R}^d$ consists of all p-simplices formed from connecting every set of $p+1$ points with pairwise distances all $\leq \epsilon$, $p = \overline{0, d}$.

$$\mathcal{R}_{\epsilon_0} \subseteq \mathcal{R}_{\epsilon_1} \subseteq \cdots \subseteq \mathcal{R}_{\epsilon_m} \qquad 0 = \epsilon_0 < \epsilon_1 < \cdots < \epsilon_m.$$

This produces sequences of homology groups $H_p^i = H_p(\mathcal{R}_{\epsilon_i})$ and Betti numbers $\beta_p^i = \dim(H_p^i)$.



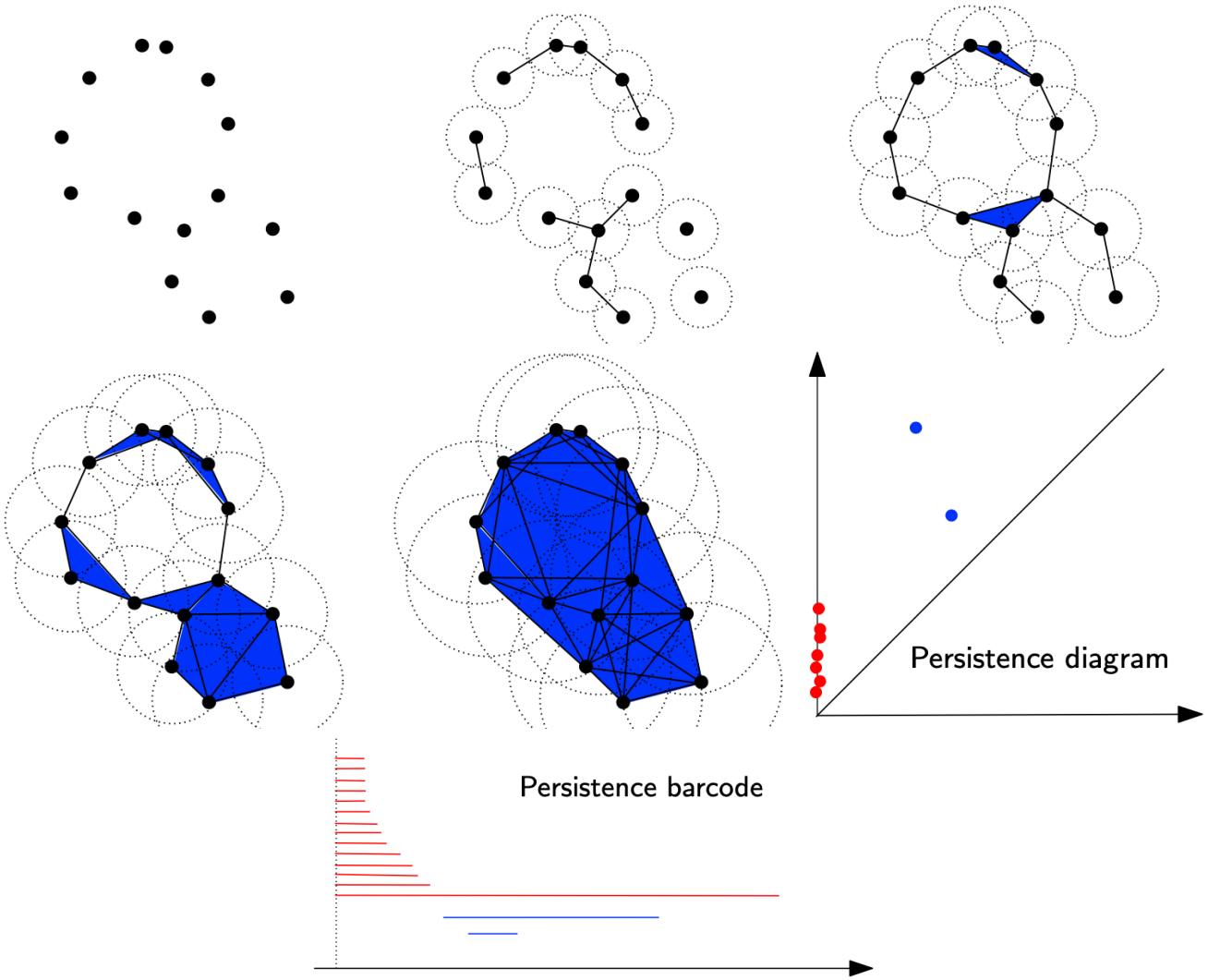
Successive nested Vietoris-Rips complexes form a filtration of the largest complex. Betti numbers change as ϵ increases.

The inclusion maps $\mathcal{R}_{\epsilon_i} \hookrightarrow \mathcal{R}_{\epsilon_j}$ $i < j$ induce linear maps $H_p^i \to H_p^j$. Their images are the persistent homology groups $H_p^{i,j} = Z_p^i / (B_p^j \cap Z_p^i)$ with similarly defined filtration cycle and boundary groups. These represent the homology classes of $\mathcal{R}_{\epsilon_i}$ that are still present in $\mathcal{R}_{\epsilon_j}$. The persistent Betti numbers are $\beta_p^{i,j} = \operatorname{rank}(H_p^i \to H_p^j) = \dim(H_p^{i,j})$.

Images from Rieck [6]

Persistence Diagrams

A homology feature is born at ϵ_i if it is present in $\mathcal{R}_{\epsilon_i}$ but not in any $\mathcal{R}_{\epsilon<\epsilon_i}$. It dies at ϵ_j if it is present in local $\mathcal{R}_{\epsilon<\epsilon_j}$ but not in $\mathcal{R}_{\epsilon_j}$. A feature that is born at $\epsilon = b$ and dies at $\epsilon = d$ has persistence $d - b$. Features are plotted with their birth parameter on the x-axis and death parameter on the y-axis to form a persistence diagram $\mathcal{D}$.



Images from Chazal [2]

Persistence Landscapes

Points in the persistence diagram can further be represented by birth-death peak functions

$$f_{(b,d)}(t) = \begin{cases} t-b & \text{if } b \le t < \frac{b+d}{2} \\ d-t & \text{if } \frac{b+d}{2} \le t < d \\ 0 & \text{otherwise} \end{cases}$$

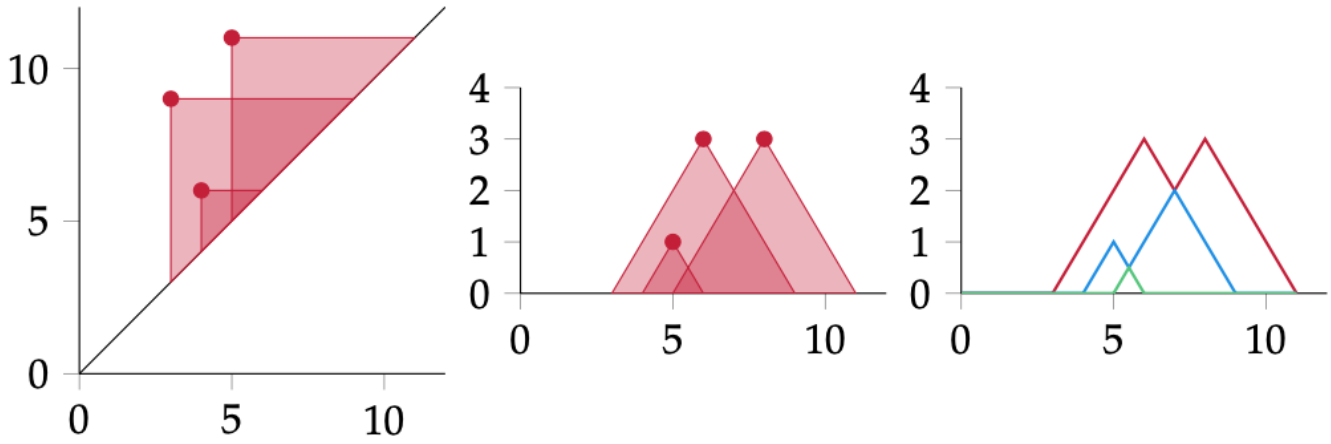
and these are used to define persistence landscape functions

$$\lambda_k(t) = \underset{(b_i,d_i)\in\mathcal{D}}{\text{kmax}} f_{(b_i,d_i)}(t).$$

These functions are decreasing in k :

$$\lambda_k(t) \ge \lambda_l(t) \quad k < l.$$

The persistence landscape Λ consists of the set of functions $\{\lambda_k\}$.



A persistence diagram, birth-death peaks, and persistence landscape.

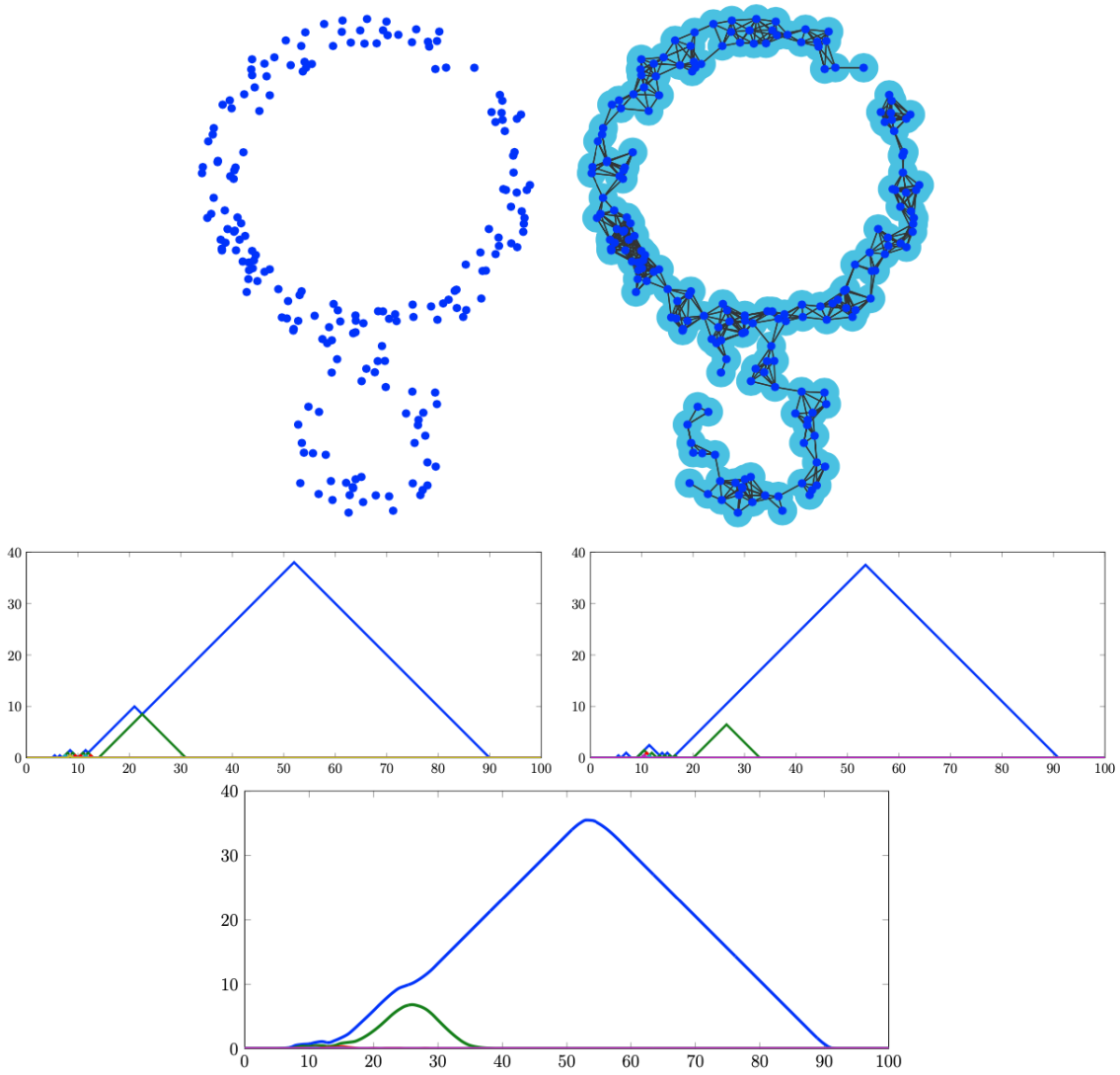
(The landscapes are plotted separately for different degrees p).

Images from Rieck [6]

Averaging Persistence Landscapes from a Sample

Take a sample of n observations $S_i \quad i = \overline{1,n}$ each consisting of a subsample of points from the manifold. Then to each S_i belongs a persistence diagram $\mathcal{D}_i$ and a persistence landscape Λ_i. An average landscape $\bar{\Lambda}$ can be formed using average landscape functions

$$\bar{\lambda_k}(t) = n^{-1} \sum_{i=1}^{n} \lambda_{k,i}(t).$$



100 samples were taken from a pair of linked annuli, each consisting of 200 points. Points from one sample and the 1-skeleton from a corresponding Čech complex are shown, followed by two degree-one persistence landscapes and the mean degree one persistence landscape for all samples. (A Čech complex is similarly to a Vietoris-Rips complex, considering distances between points mutually rather than pairwise).

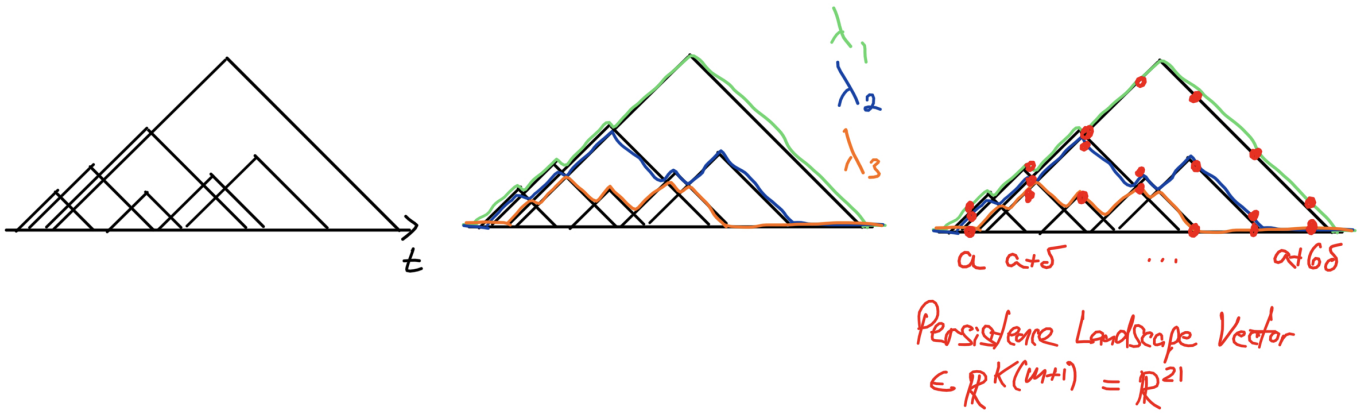
Images from Bubenik [1]

Discretization and Inference

The persistence landscapes need to be converted into scalar values to perform statistical inference. One approach involves discretizing the landscapes into vectors.

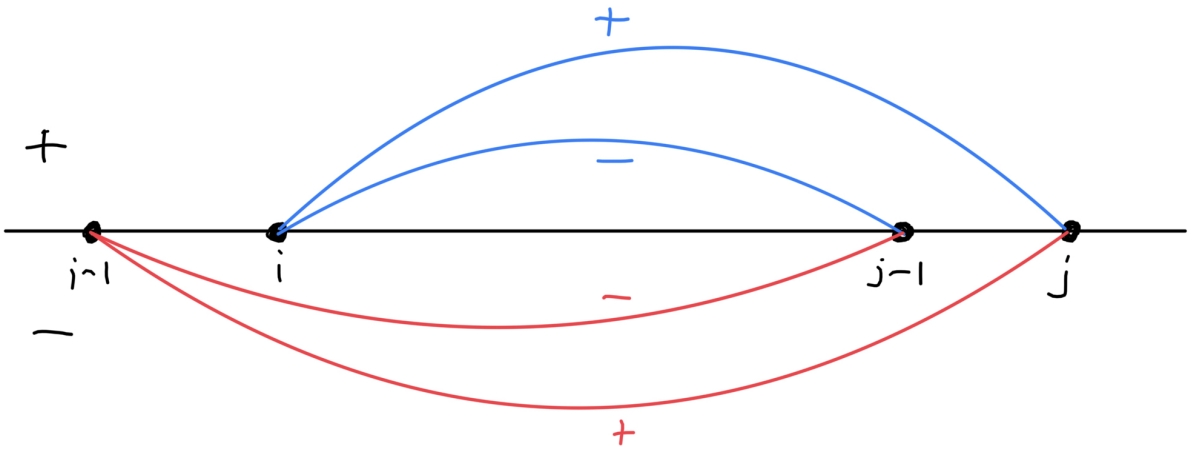
For degree zero simplices (points), all birth values are zero and a death vector is formed by simply recording the order statistics of the death values.

For higher degrees, suitable K, m, and δ are chosen so the values of the first K landscape functions are recorded over m intervals of size δ to create a landscape vector.



Other approaches involve defining a norm of the persistence landscape. Various distributional results are known, see for example Dixon et al. [4].

Multiplicity of Persistence Diagram Birth-Death Points

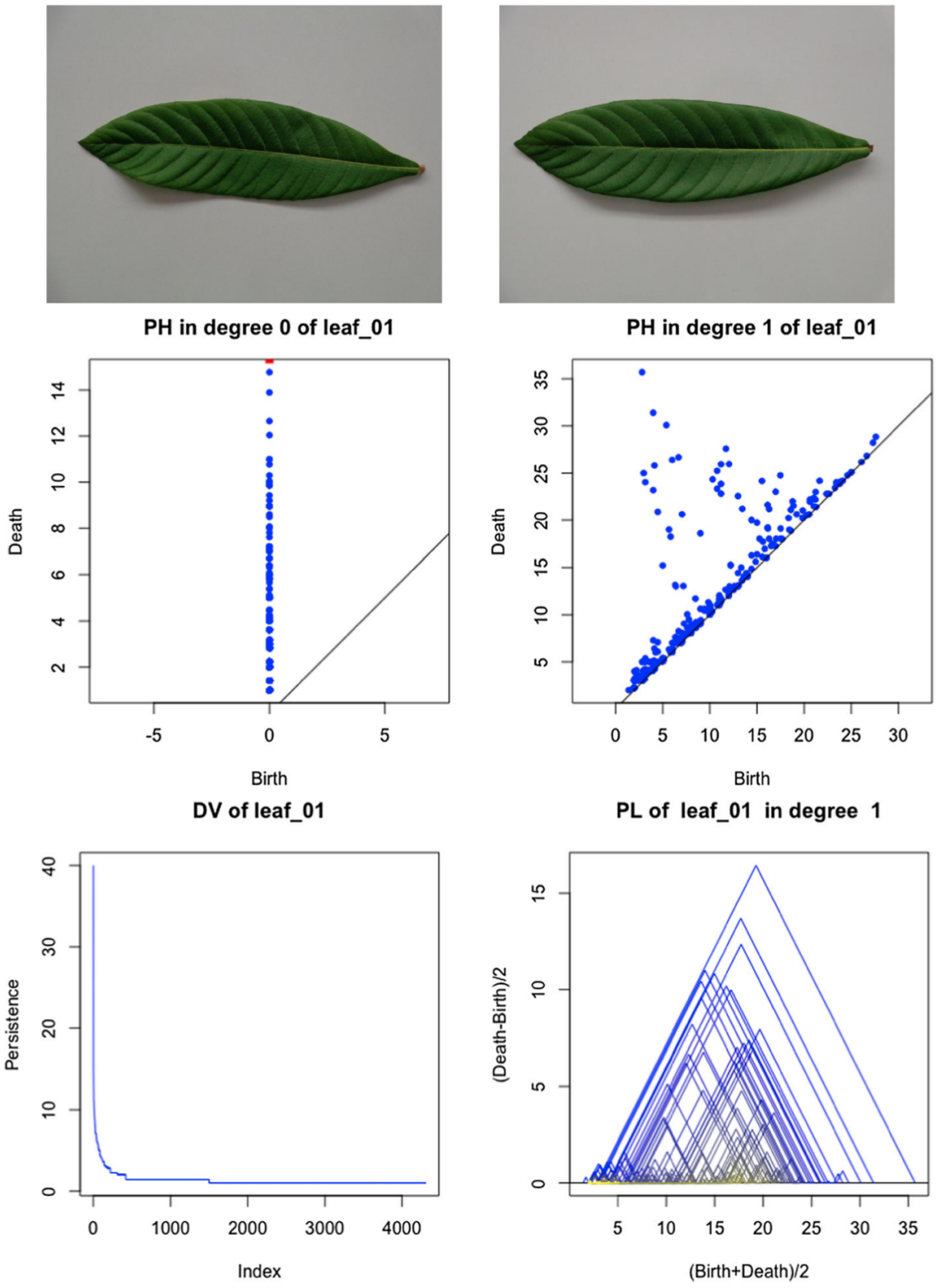


$$m_p^{i,j} = \underbrace{\left(\beta_p^{i,j-1} - \beta_p^{i,j}\right)}_{\text{Existed at } i,\ \text{died at } j} - \underbrace{\left(\beta_p^{i-1,j-1} - \beta_p^{i-1,j}\right)}_{\text{Existed at } i-1,\ \text{died at } j}$$

$$= \#\{\text{Died at } j,\ \text{existed at } i,\ \text{did not exist at } i-1\}$$

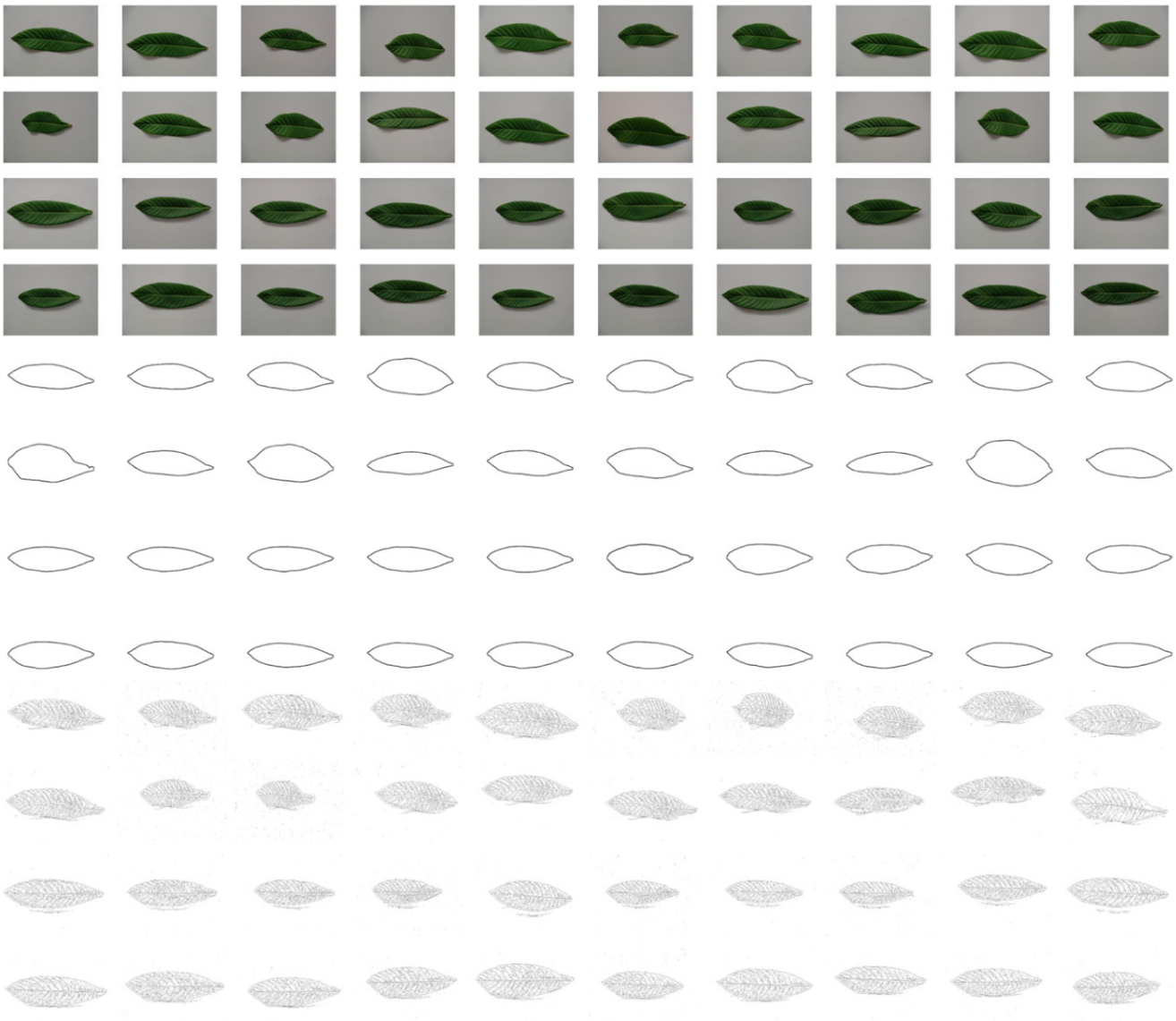
$$= \#\{\text{Born at } i,\ \text{died at } j\}$$

An Example from Patrangenaru et al.



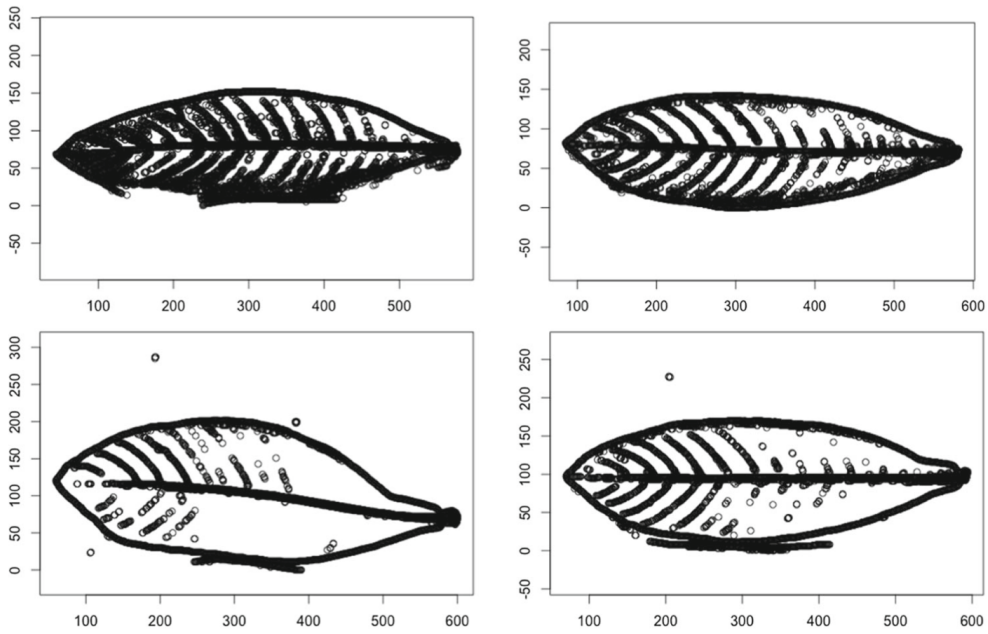
Sample original images of leaf A (left) and leaf B (right). Persistence diagrams, death vector, and persistence landscape for leaf A.

Images from Patrangenaru et al. [5]

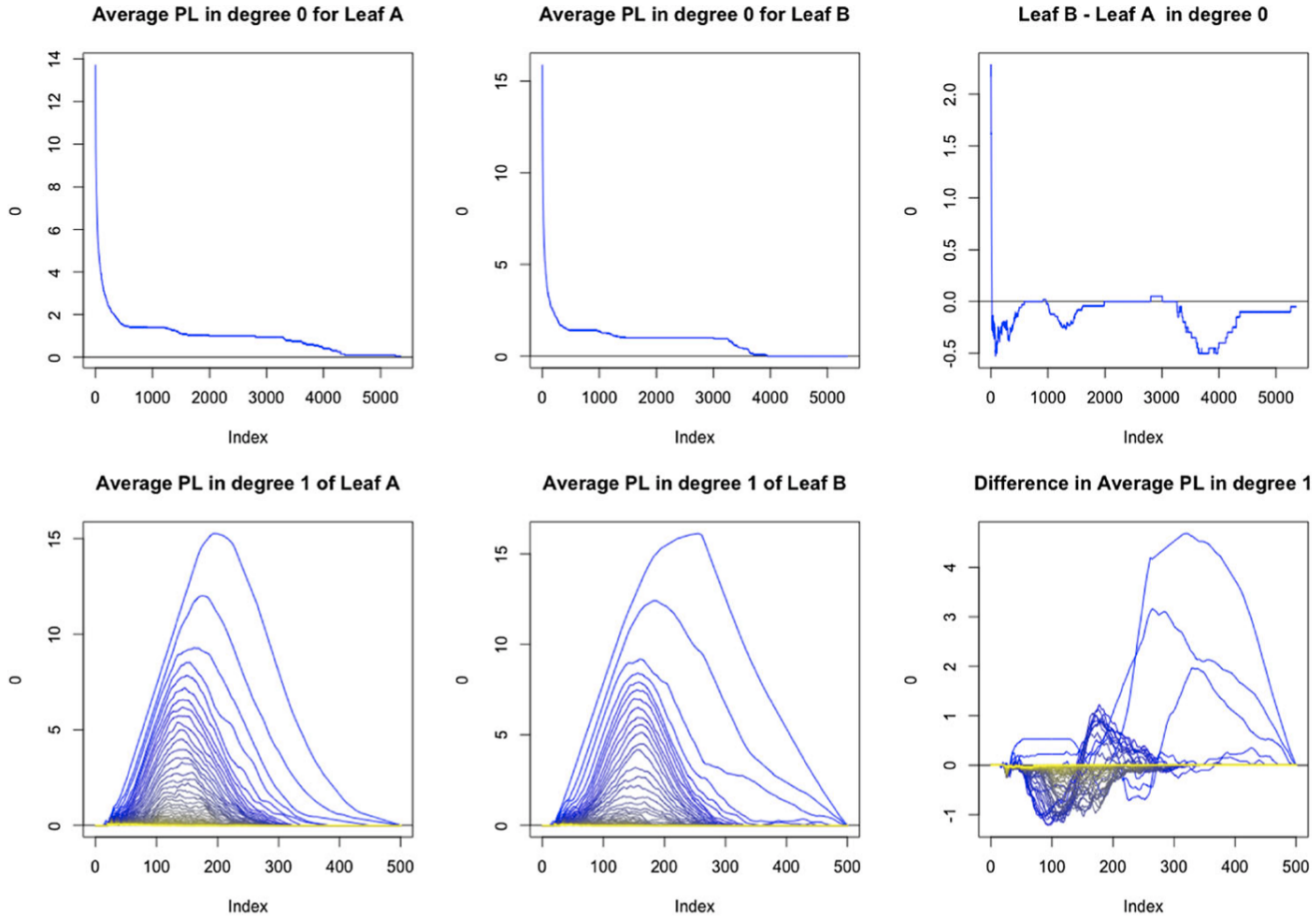


Original images, contours extracted from the original images, and leaf edges from the original images. In each figure the top 20 thumbnails are from leaf A and the bottom 20 are from leaf B.

Images from Patrangenaru et al. [5]



Sample point clouds. Left images from leaf A and right images from leaf B.



Averages of 20 death vectors and persistence landscapes for each leaf, and their differences.

Images from Patrangenaru et al. [5]

References

[1] Bubenik, Peter (2015). *Statistical Topological Data Analysis using Persistence Landscapes*. Journal of Machine Learning Research 16 (2015) 77-102.

[2] Chazal, Frederic (date unspecified). *Topological Data Analysis* Class lecture slides. Accessed at https://julien-tierny.github.io/topologicalDataAnalysisClass.html.

[3] Chazal, Frederic; Michel, Bertrand (2021). *An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists*. Front. Artif. Intell. 4:667963. doi: 10.3389/frai.2021.667963.

[4] Dixon, Adam; Patrangenaru, Victor; Shen, Chen (2020). *An Introduction to Topological Object Data Analysis*. Balkan Society of Geometers BGS Proceedings 28 (2021) 30-44.

[5] Patrangenaru, Vic; Bubenik, Peter; Paige, Robert; Osborne, Daniel (2018) *Challenges in Topological Object Data Analysis* Sankhya A. 81, 244-271.

[6] Rieck, Bastian (2020) *Topological Data Analysis for Machine Learning* Online lectures delivered 9/14/2020. Accessed at https://bastian.rieck.me/talks/ecml_pkdd_2020/.

[7] Tierny, Julien (date unspecified). *Topological Data Analysis* Class lecture slides. Accessed at https://julien-tierny.github.io/topologicalDataAnalysisClass.html.