Statistical Shape Analysis using Topological Data Analysis

Part I: Introduction to Topological Data Analysis

Chul Moon

ECSSC 2021
25 July 2021

Idea of Topological Data Analysis (TDA)

Data have shape,

shape has meaning,

meaning brings value

Idea of Topological Data Analysis (TDA)

Data have shape,

$\rightarrow$ Modeling

shape has meaning,

$\rightarrow$ Inference/Learning

meaning brings value

$\rightarrow$ Collaboration

Contents of Today's Lecture

1. Part I: Introduction to TDA
2. Part II: Representations and Modeling
3. Part III: Applications
   - Fingerprint Data
   - Time Series Data
   - Material and Tumor Images

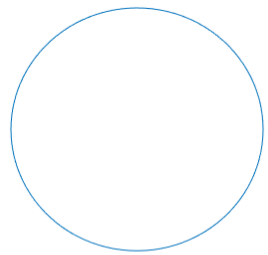
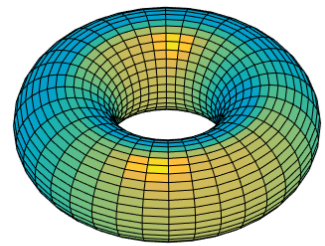
Outline

Topology and Topological Data Analysis (TDA)

- Topology and topological invariance
- TDA analyzes the "shape" of data
- Uses techniques from topology; homology and persistent homology
- Homology analyzes topological spaces

Betti Numbers

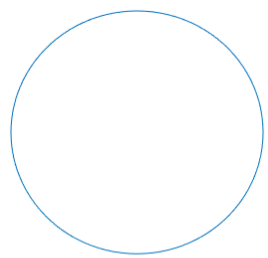
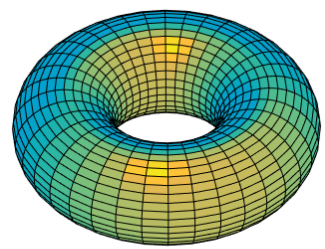
Betti number β_k, counts the number of k-dimensional holes

		
Dim. 0 (Conn. comp.)	1	1
Dim. 1 (Loops)	1	2
Dim. 2 (Voids)	0	1

Euler Characteristic

Euler characteristic χ is defined

$$\begin{aligned}\chi &= \sum_{i=0}^{d}(-1)^i\beta_i \\ &= \text{num. of vertices} - \text{num. of edges} + \text{num. of faces}\end{aligned}$$

		
Dim. 0 (Conn. comp.)	1	1
Dim. 1 (Loops)	1	2
Dim. 2 (Voids)	0	1
Euler characteristic	$1-1+0=0$	$1-2+1=0$

Toy Example

- How many of holes are there? What would be the Euler characteristics?

	A	B	U	⊠	◎
Dim. 0 (Conn. comp.)					
Dim. 1 (Loops)					
Dim. 2 (Voids)					
Euler characteristic					

Shape of Data?

- Shape $\longrightarrow$ Data $\longrightarrow$ Complex

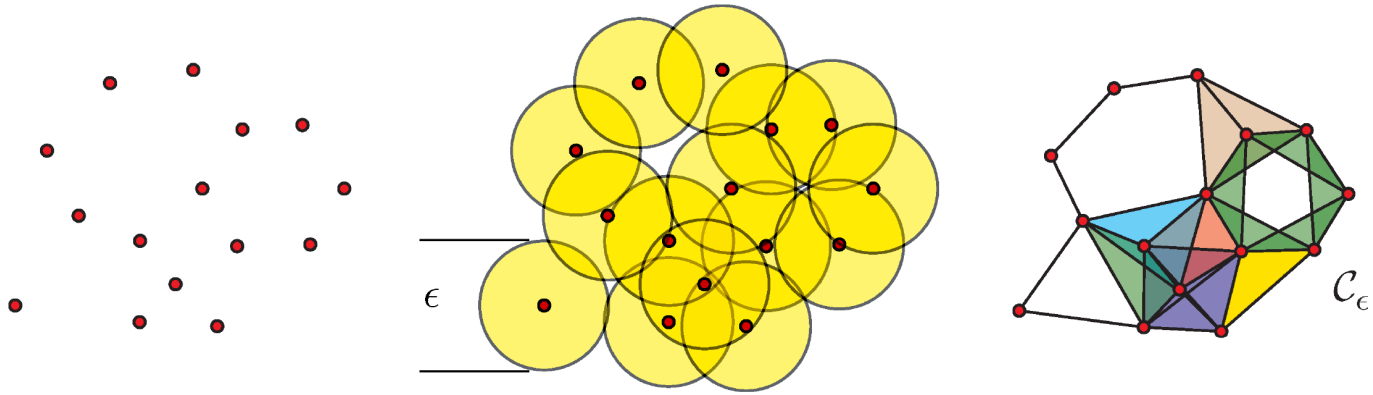


Figure: Example of Čech complex. Figure from Ghrist (2008)

- Čech complex is defined by the intersection between those balls
- If there are m balls of diameter ϵ that share the intersection region, then the corresponding m data points form an $m-1$-simplex (simplex will be covered within a few slides)
- Čech complex at the filtration value ϵ is a union of those simplexes and is represented as $\mathcal{C}_\epsilon$

Nerve Theorem

- Is it okay to analyze the complex instead of the topological space?
- A simplicial complex from an open covering of a topological space preserves topological properties

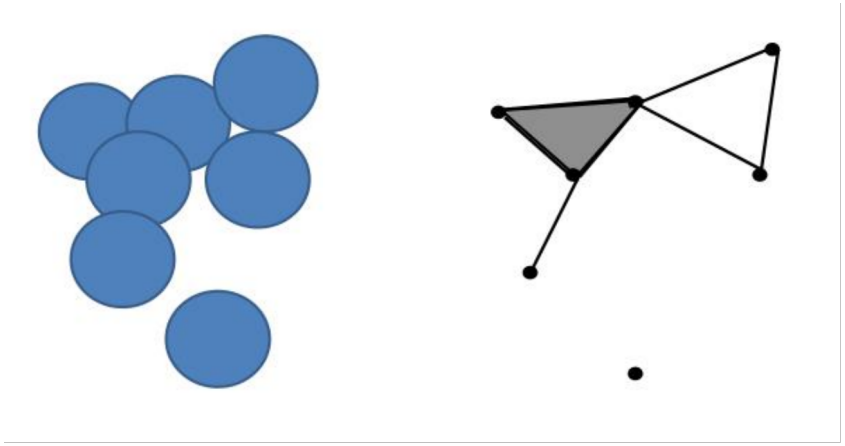


Figure: Figure from `https://www.slideserve.com/mandy/sampling-conditions-and-topological-guarantees-for-shape-recons`

- Čech complex satisfies the nerve theorem: if we sample enough data from a topological space $\mathcal{X}$ and construct the Čech complex, then the complex reflects the topology of $\mathcal{X}$

Vietoris-Rips Complex

- Drawback of the Čech complex is that it is computationally expensive
- Vietoris-Rips (Rips) complex is defined by pairwise distances
- For given ϵ, m data points whose pairwise distances are smaller than ϵ form a $(m-1)$-simplex
- Rips complex of X given ϵ is a union of the simplexes and expressed $\mathcal{R}_\epsilon$

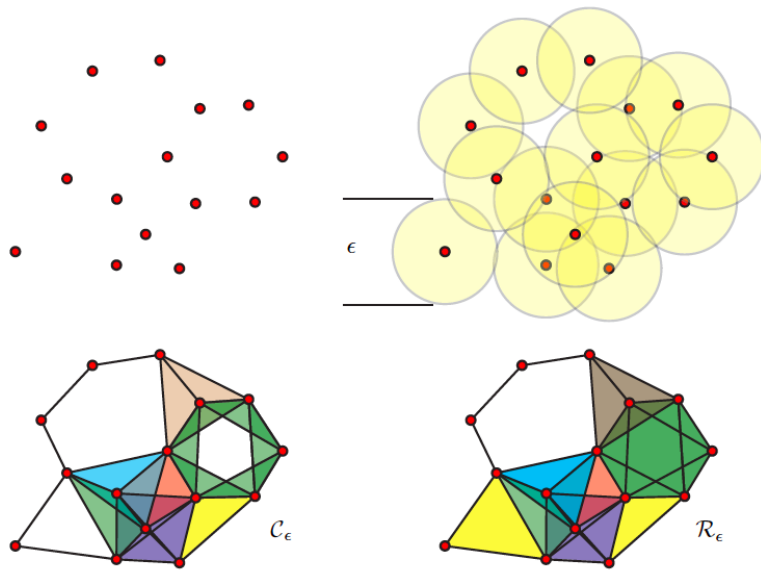


Figure: Figure from Ghrist (2008)

Čech vs. Rips Complexes

- Rips complex does not satisfy the nerve theorem
- Rips complex still carries topological properties of point cloud data. For all $\epsilon > 0$, the following relationships hold:

$$\mathcal{C}(X, \epsilon) \subset \mathcal{R}(X, \epsilon) \subset \mathcal{C}(X, 2\epsilon).$$

- Rips complex contains topological information of the Čech complexes somewhere between $\mathcal{C}(X, \epsilon)$ and $\mathcal{C}(X, 2\epsilon)$ and used as a good alternative to the Čech

Simplices and Simplicial Complexes

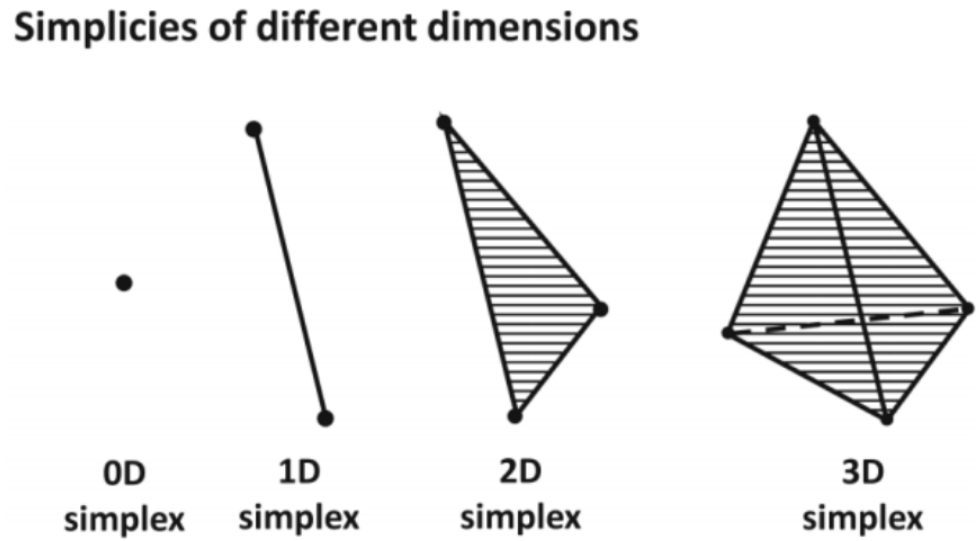


Figure: Figure from Wang et al. (2019)

- Simplices are building blocks of a simplicial complex: vertices (0-simplex), edges (1-simplex), triangles (2-simplex), tetrahedra (3-simplex), and higher-dimensional simplices

Simplices and Simplicial Complexes (2)

- In the construction of the simplicial complex, simplices are glued together in a way that two adjacent simplices share their faces

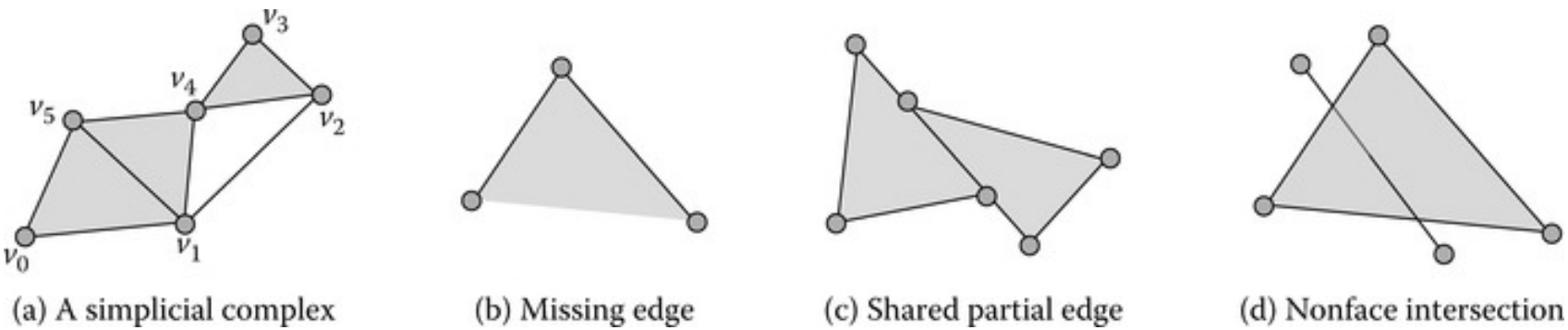


Figure: Figure from Atallah and Blanton (2009)

Homology

- Homology enables to analyze a topological space $\mathcal{X}$ by examining its k-dimensional holes

$$H_k = \ker(\partial_k)/\mathrm{im}(\partial_{k+1}) = \mathsf{Z}_k/\mathsf{B}_k$$

- The rank of $H_k(\mathcal{X})$ is the Betti number β_k
- Computation of homology is done by simple linear algebra

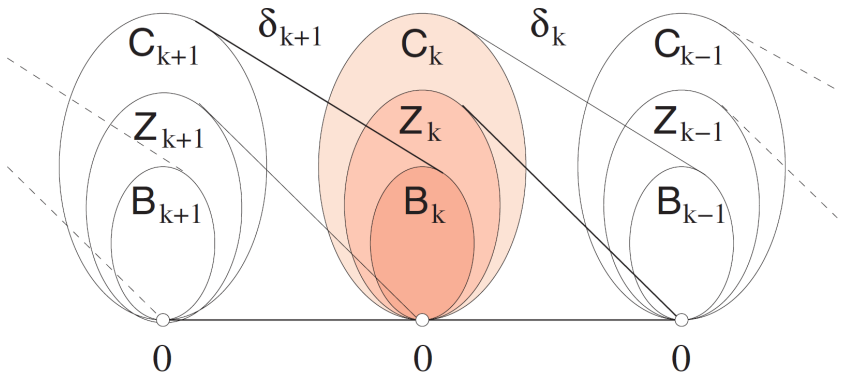


Figure: Chain, cycle, and boundary groups. Figure from Zomorodian and Carlsson (2005)

Is Homology Enough?

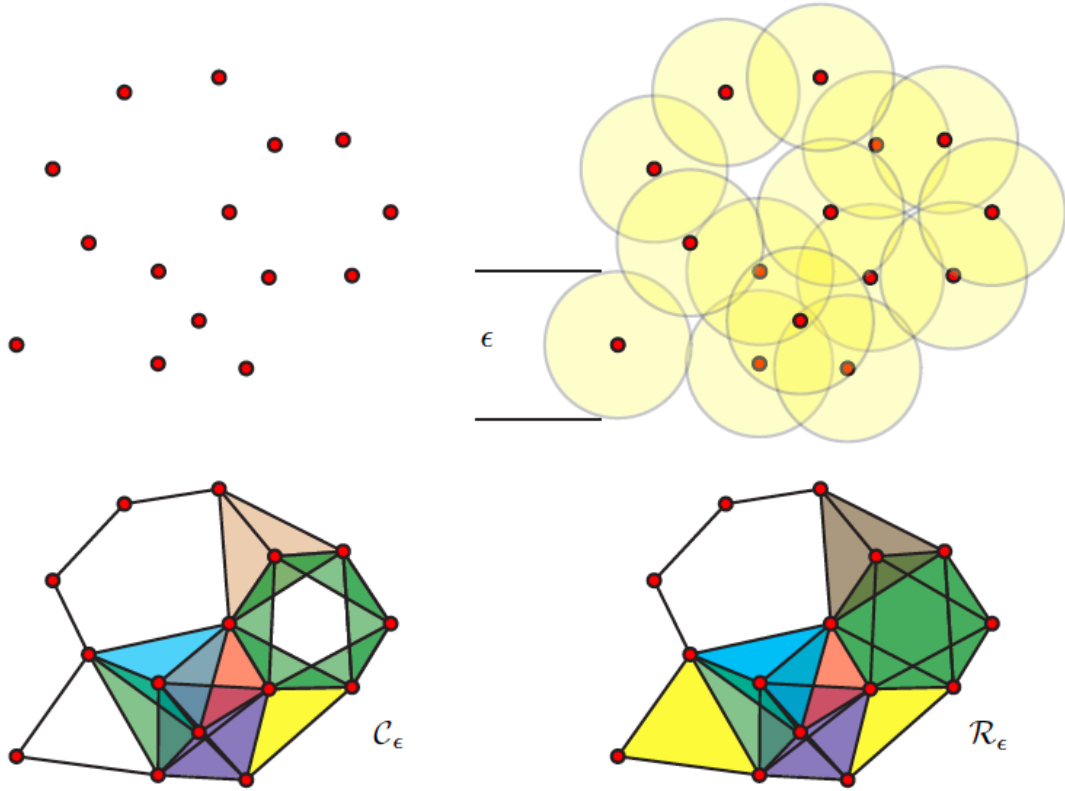


Figure: Figure from Ghrist (2008)

- Question: which ϵ do we need to use to reveal the underlying shape?

Persistent Homology

- Persistent homology keeps track of the homology of a sequence of topological spaces
- Let $\emptyset = \mathcal{X}_0 \subseteq \mathcal{X}_1 \subseteq \cdots \subseteq \mathcal{X}_n = \mathcal{X}$ be a non-decreasing sequence of topological spaces
- The inclusion of $\mathcal{X}_i$ in $\mathcal{X}_j$ for $0 \leq i \leq j \leq n$ induces a homomorphism between the homology groups, $\iota_k(i,j) : H_k(\mathcal{X}_i) \to H_k(\mathcal{X}_j)$

$$0 = H(\mathcal{X}_0) \to H(\mathcal{X}_1) \to \cdots \to H(\mathcal{X}_n) = H(\mathcal{X})$$

- Some topological features in a topological space $\mathcal{X}_s$ may born or die along the sequence. The image of the homomorphism include such information
- k-dimensional (i,j) persistent homology groups is the image of the homomorphism $\iota_k(i,j) : H_k(\mathcal{X}_i) \to H_k(\mathcal{X}_j)$
- The parameter that controls the sequential changes (ϵ) is called the filtration

Techniques from Topology: Persistent Homology (2)

- Tracks evolution of homology as ϵ (filtration) changes
- Persistence homology results are given as an interval of [Birth, Death]
- In general, true topological properties persist for a wide range of filtrations

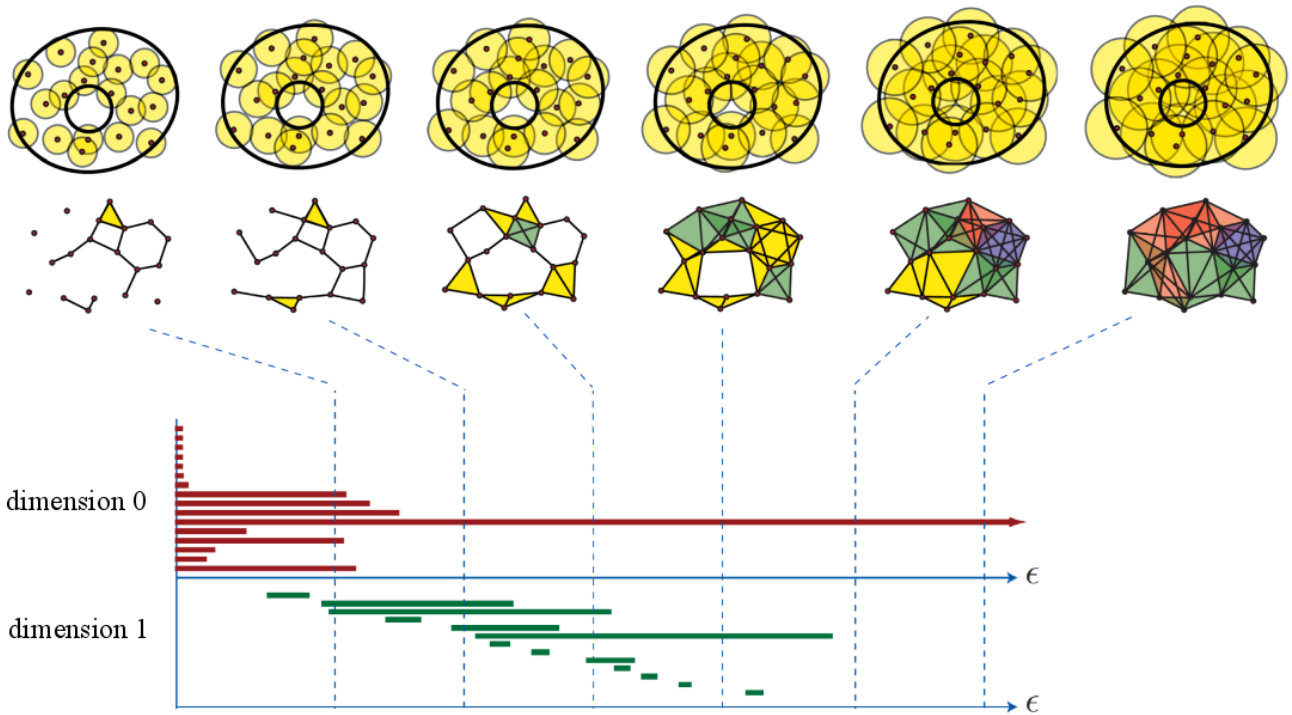
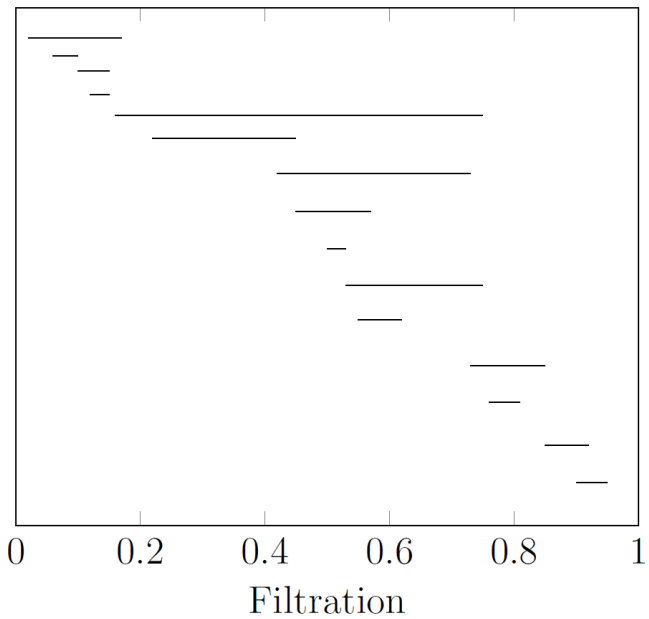


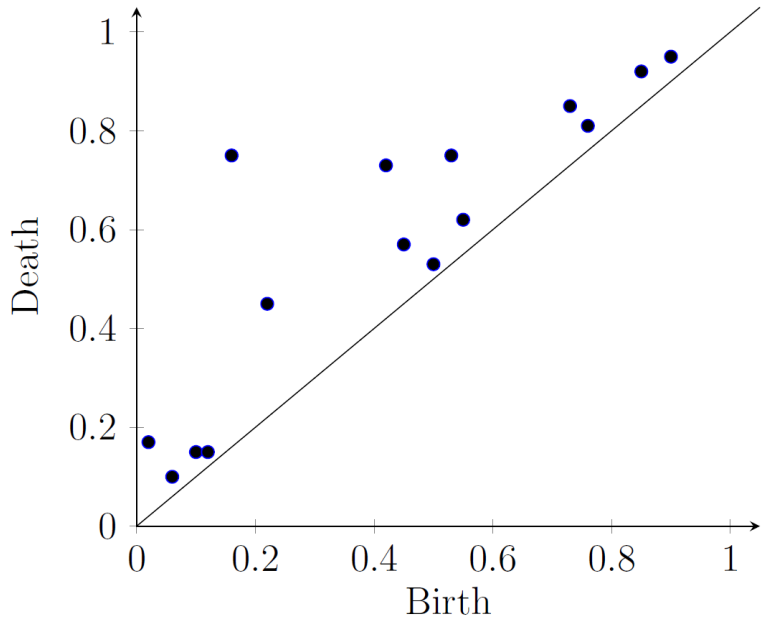
Figure: Figure from Ghrist (2008)

Summary of Persistent Homology Result

- Barcode plot: as a straight line of Birth———Death
- Persistence diagram: as a point of (Birth, Death)



(a) Barcode Plot



(b) Persistence Diagram

Popular Packages for Persistent Homology

- Python
  - Dionysus 2
  - GUDHI
  - scikit-tda
  - ripser
  - giotto-tda
- R
  - TDA

Pipeline

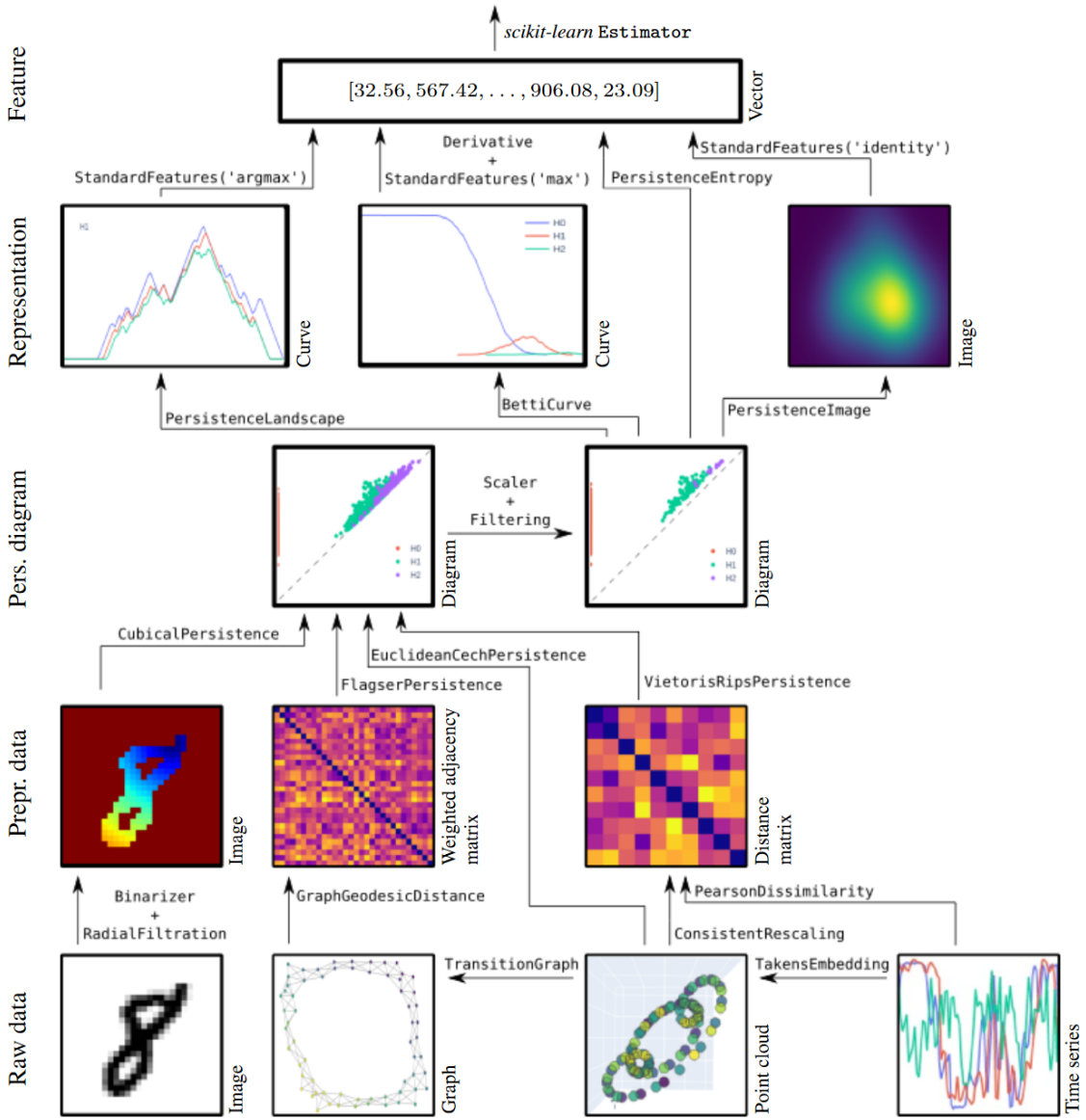


Figure 1: Non-exhaustive depiction of *giotto-tda* capabilities. Arrows represent operations available as `transformers` and paths potential `pipelines`.

Outline

Examples of Point Cloud Data

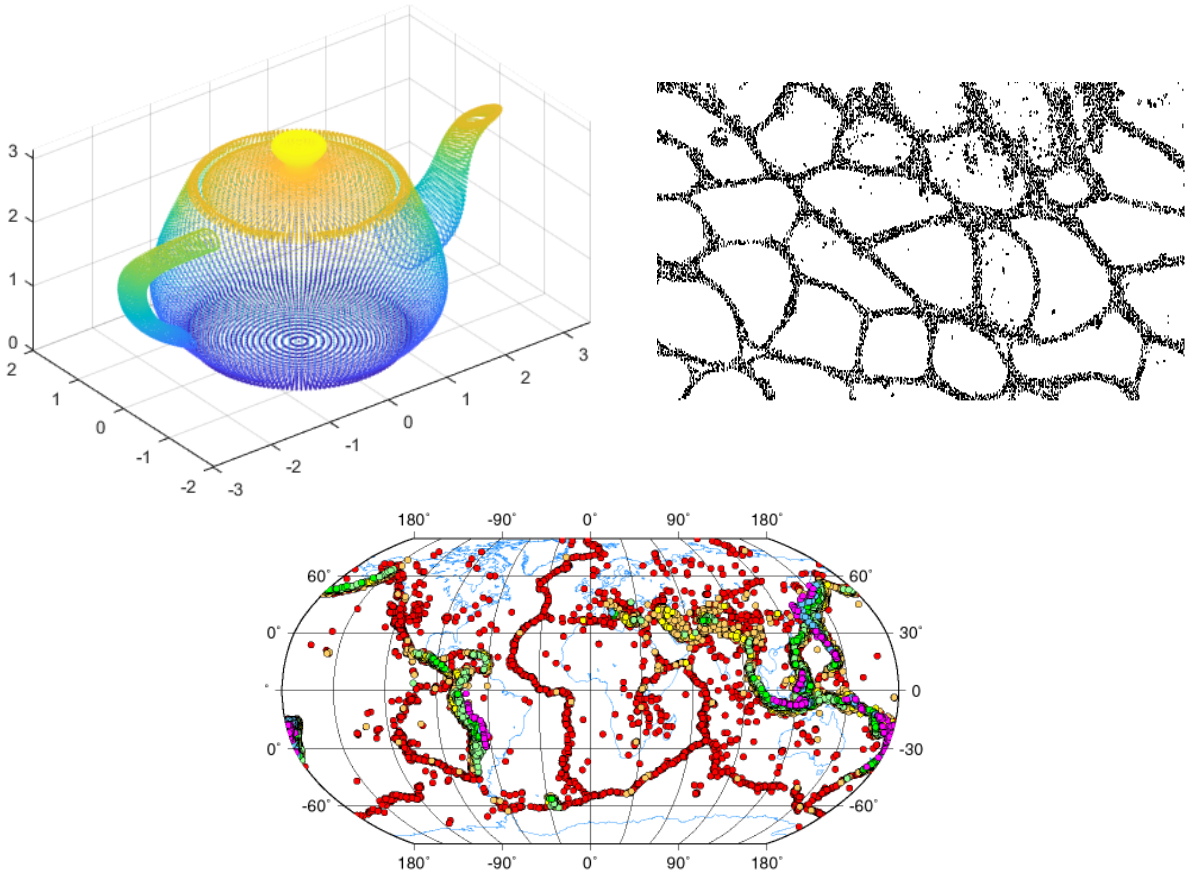


Figure: Figures from MATLAB (2016) (top-left), Mula et al. (2013) (top-right), and Pidwirny (2006) (bottom)

Limitations of Čech complex on Point Cloud Data

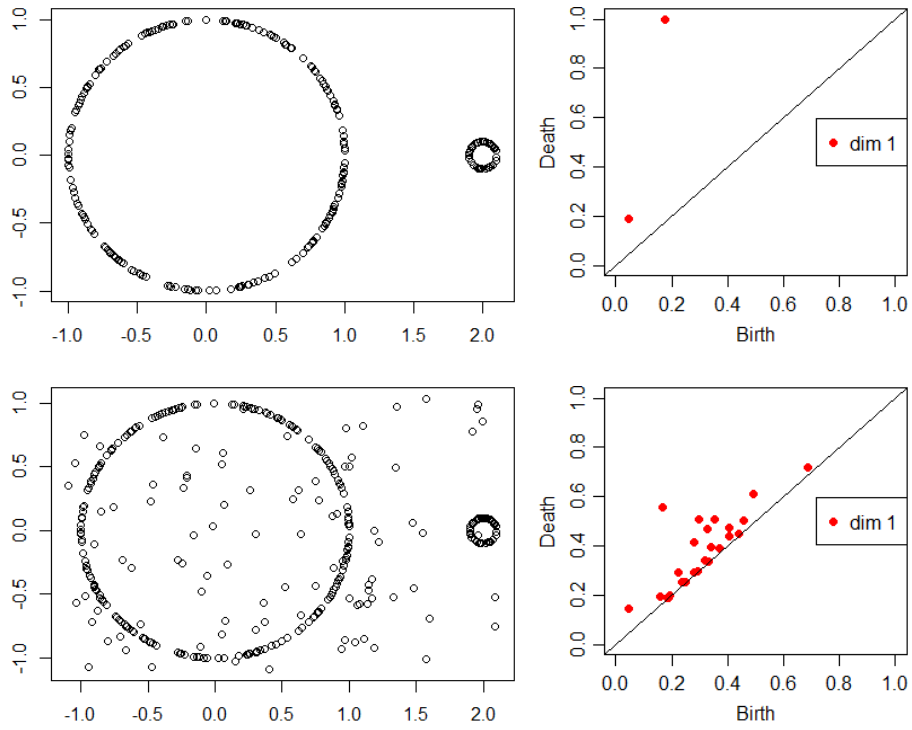
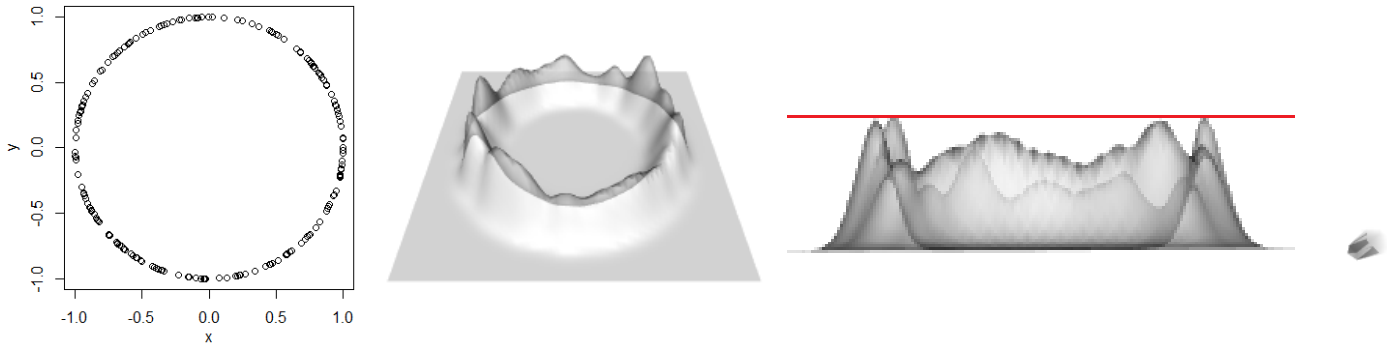


Figure: Data without noise (top row) and with noise (bottom row)

- Persistence is related to the size of the feature
- Small-sized features do not persist for a long range of filtrations
- Sensitive to noise

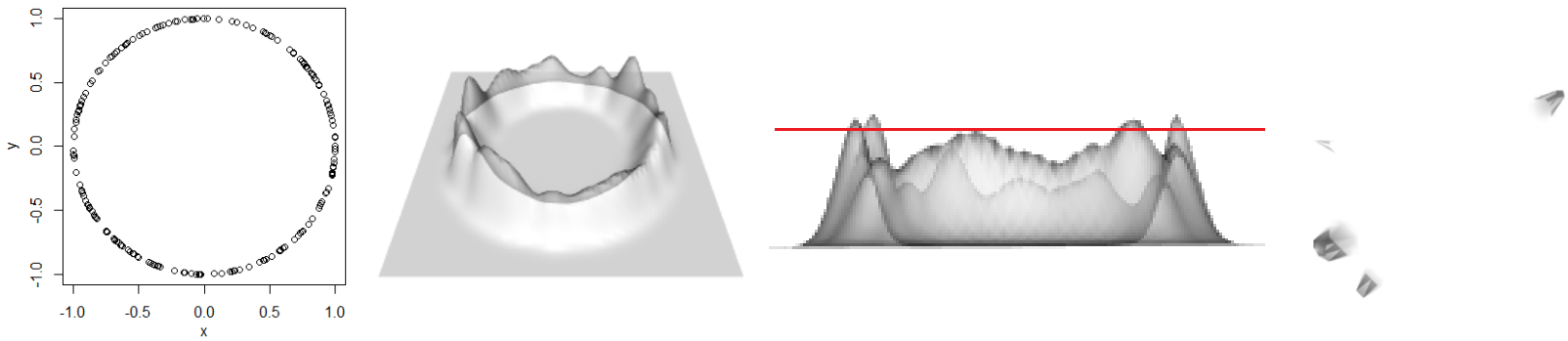
Robust Estimation using Smoothing Function



(a) Scatterplot (b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

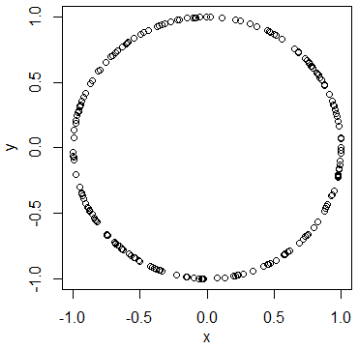
Robust Estimation using Smoothing Function



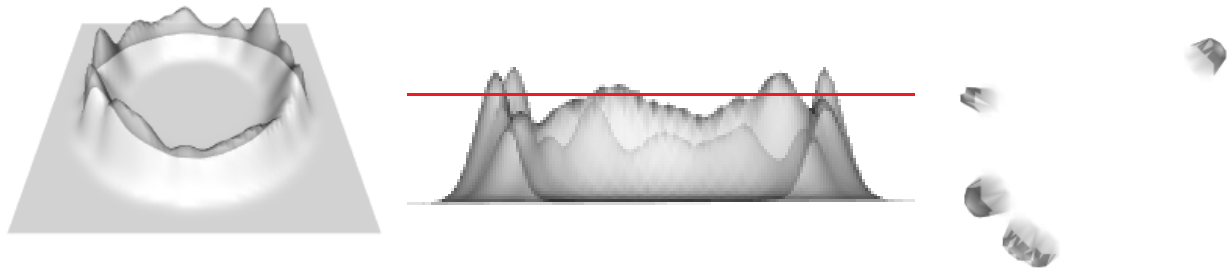
(a) Scatterplot (b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



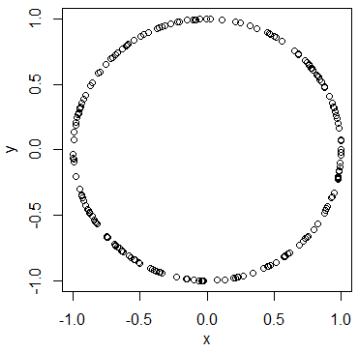
(a) Scatterplot



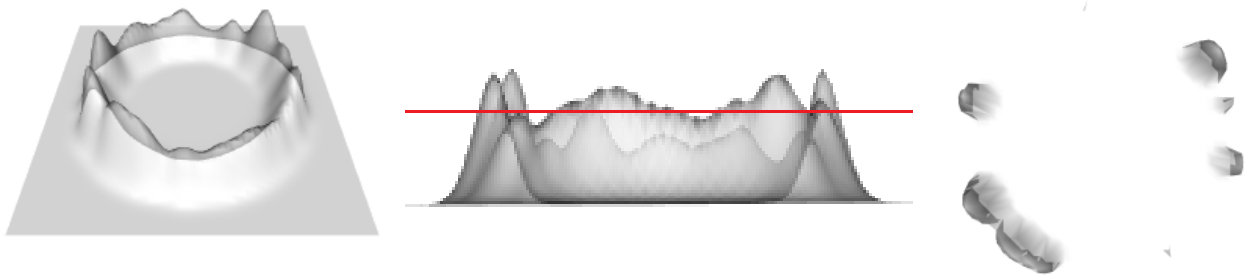
(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



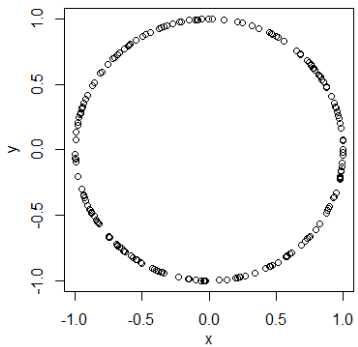
(a) Scatterplot



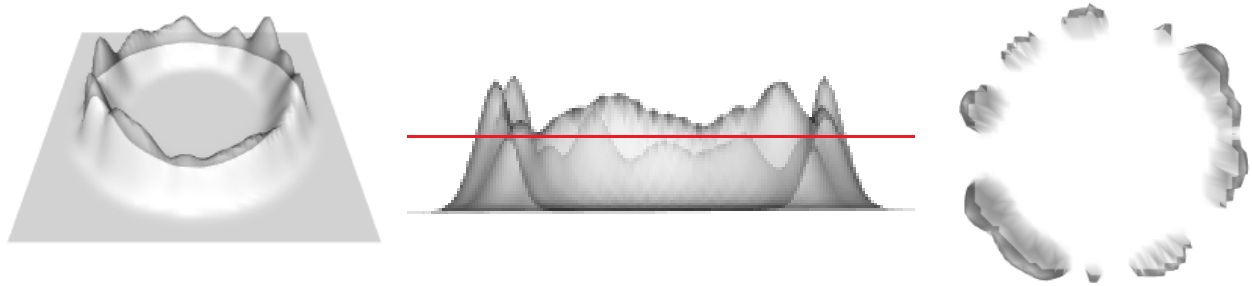
(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



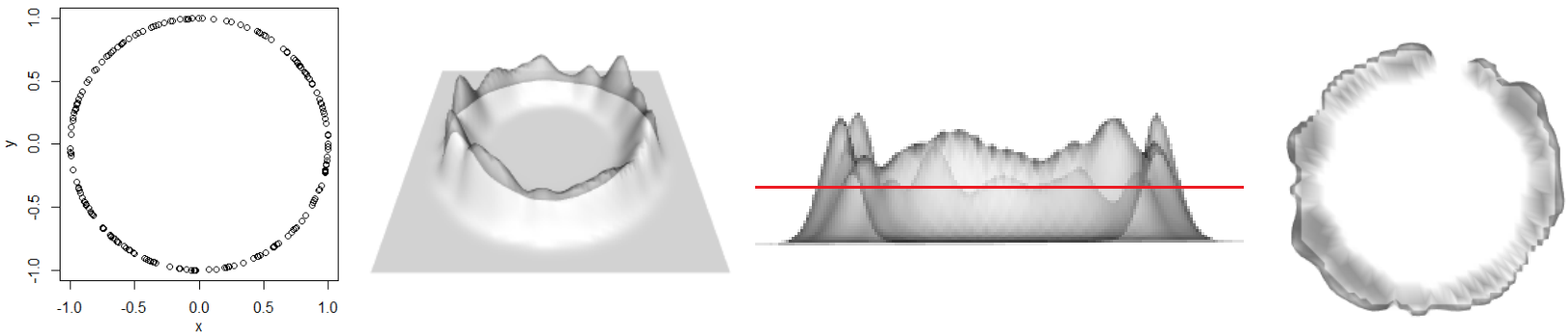
(a) Scatterplot



(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

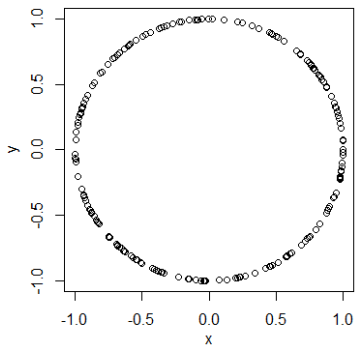
Robust Estimation using Smoothing Function



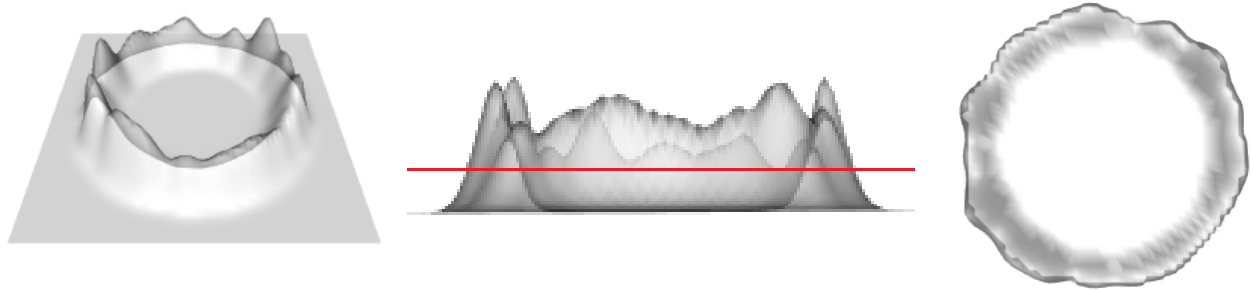
(a) Scatterplot (b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



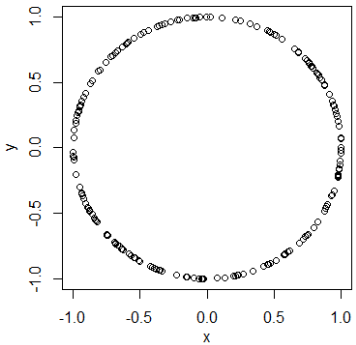
(a) Scatterplot



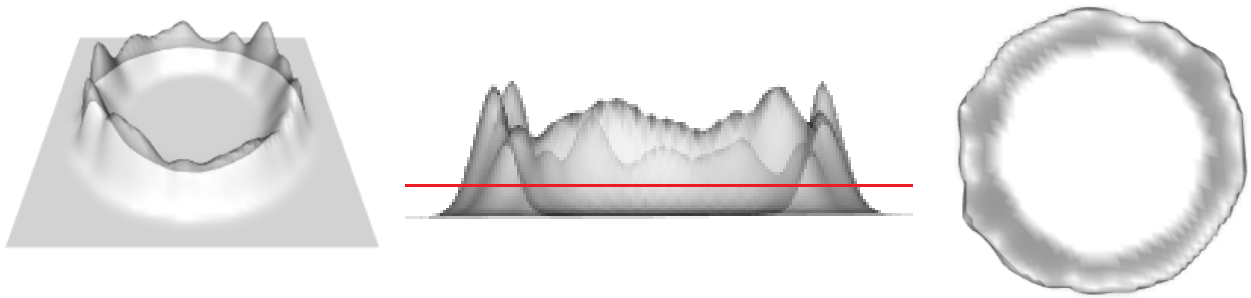
(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



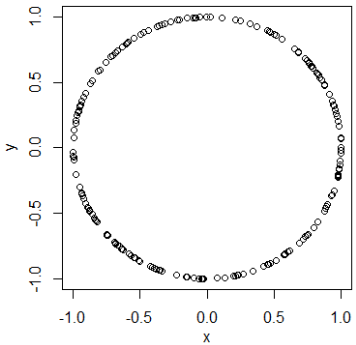
(a) Scatterplot



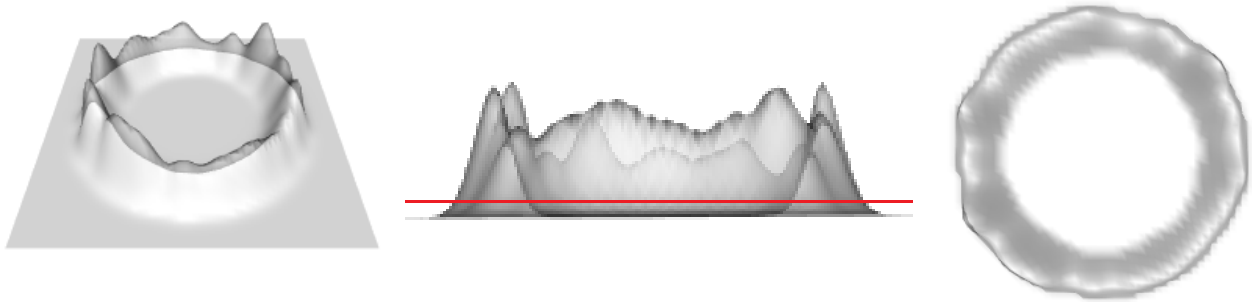
(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



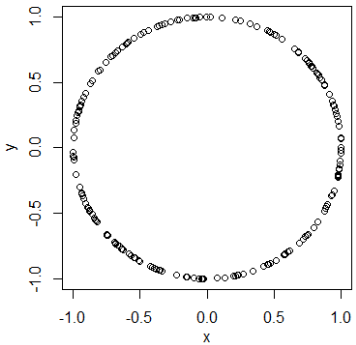
(a) Scatterplot



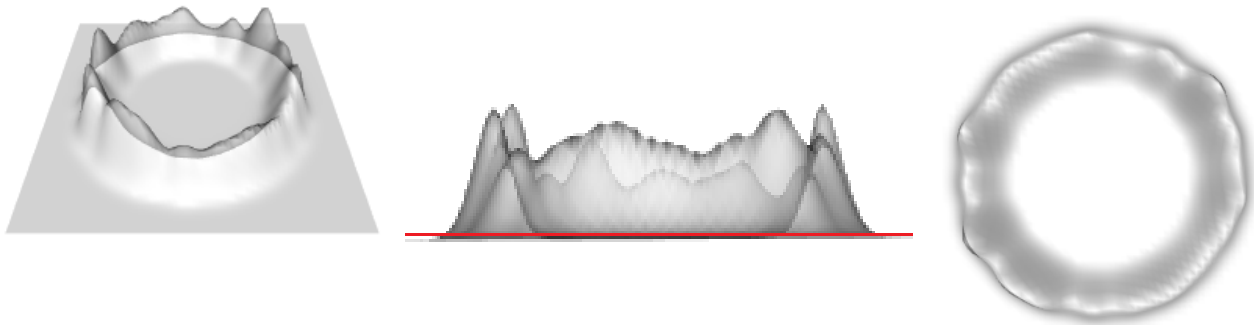
(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Robust Estimation using Smoothing Function



(a) Scatterplot



(b) Manifold

- Transform a discrete point cloud into a continuous manifold via a smoothing function
- Set a filtration as a level of manifold
- Optimal smoothing parameter selection
- Lose size information

Various Smoothing Functions

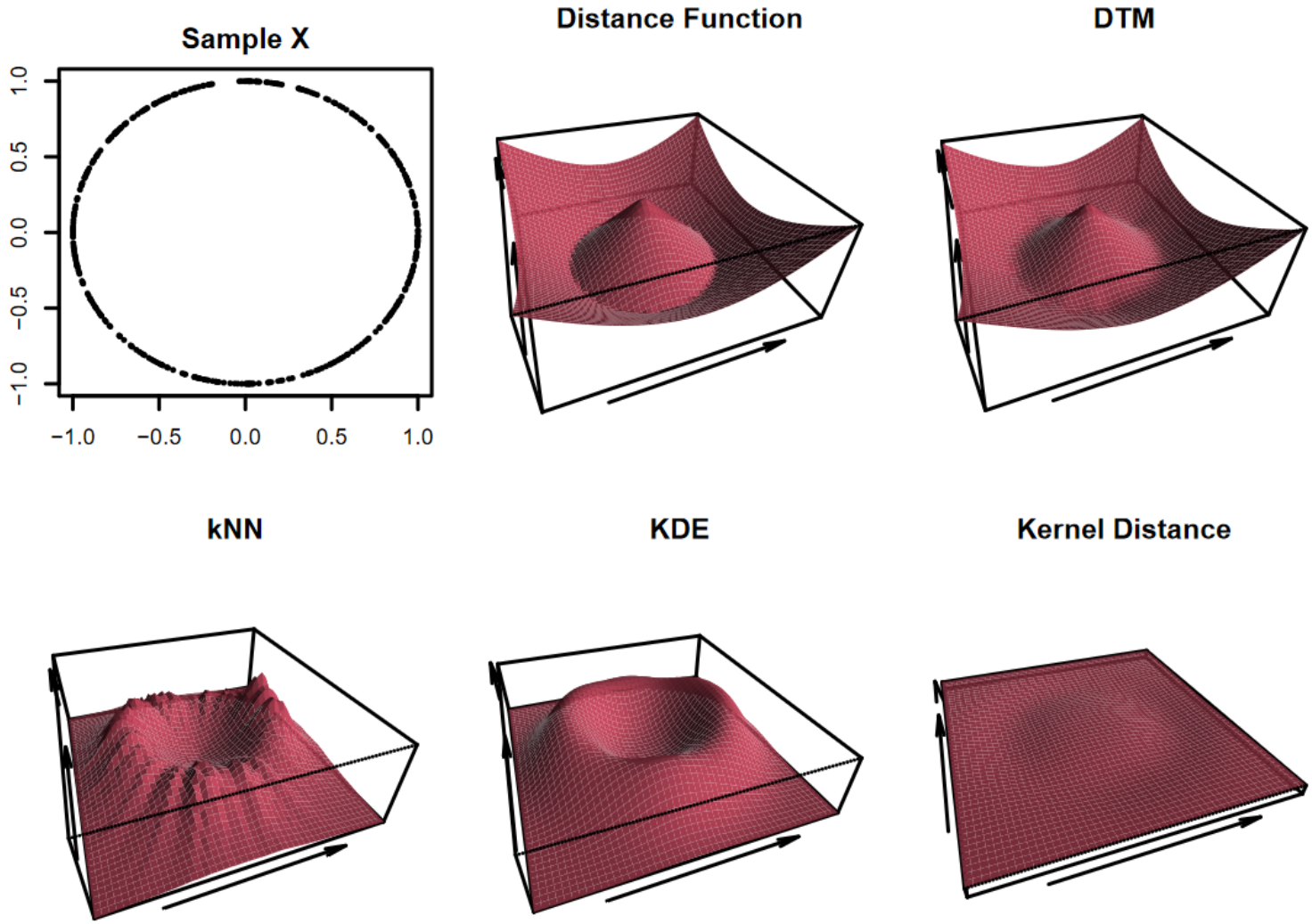


Figure: Figure from Fasy et al. (2014)

Outline

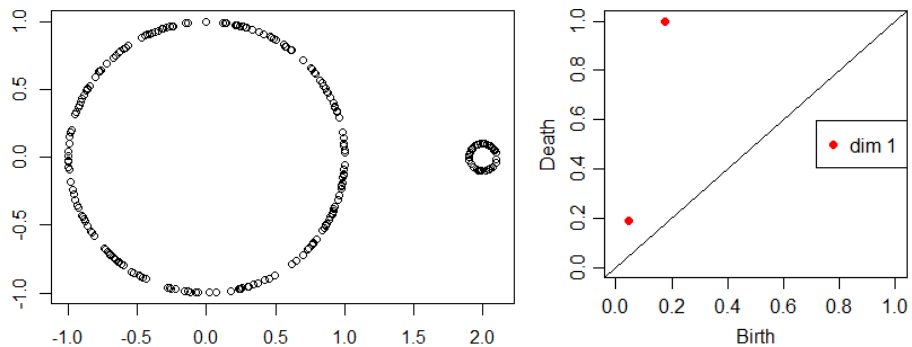
Reliability of Persistent Homology Results

Can persistence diagrams be used to measure geometric or topological differences between spaces?

- Let D_X and D_Y be the persistence diagrams of metric space X and Y
- For given metrics d_1 on persistence diagrams and d_2 on the metric spaces, the stability of persistence diagrams holds when

$$d_1(D_X, D_Y) \leq d_2(X, Y).$$

- If the persistence diagram is stable, then small changes in metric spaces (measured as $d_2(X, Y)$) induce small changes in persistence diagrams (measured as $d_1(D_X, D_Y)$)



Persistence Diagram Distances

- The persistence diagram distances $d_1(D_X, D_Y)$ measure the dissimilarity between diagrams
- The space where persistence diagrams live is unique
  - The persistence diagram distance needs to compare diagrams that include different numbers of points
  - Number of barcodes created varies from dataset to dataset
  - Only half of the 2-dimensional space will be used
- Also, the distances need to satisfy an important property called stability under perturbations.

Bottleneck Distance

- Let $x_i = (a_i, b_i)$, $y_j = (a_j, b_j)$, and $d(x_i, y_j) = ||x_i - y_j||_\infty = \max(|b_i - b_j|, |a_i - a_j|)$
- Bottleneck distance between persistence diagrams D_X and D_Y is defined as

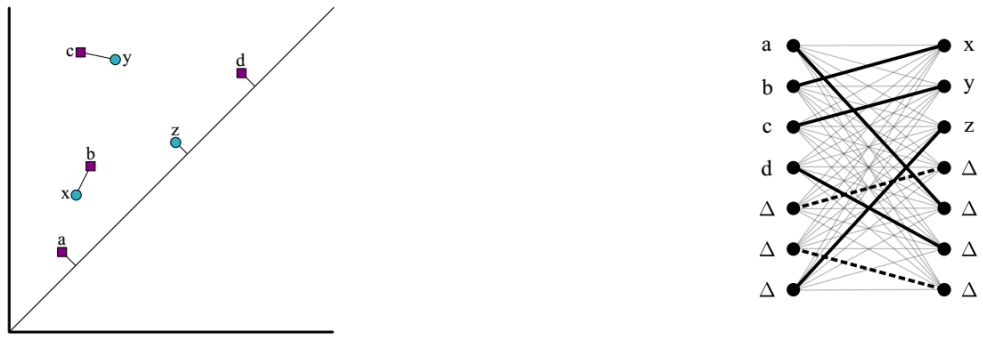
$$W_\infty(D_X, D_Y) = \inf_\gamma \sup_{x \in D_X} ||x - \gamma(x)||_\infty$$

  where γ is the bijections from D_X to D_Y
- Bottleneck distance between two persistence diagrams has an upper bound by L_∞-distance
- The stability of the bottleneck distance is shown in Cohen-Steiner et al. (2007) and Chazal et al. (2012)

Wasserstein Distance

p-Wasserstein distance between D_X and D_Y is given by

$$W_p(D_X, D_Y) = \left(\inf_{\gamma} \sum_{x \in D_X} ||x - \gamma(x)||_\infty^p \right)^{\frac{1}{p}}$$



(a) Two persistence diagrams (b) Weighted graph

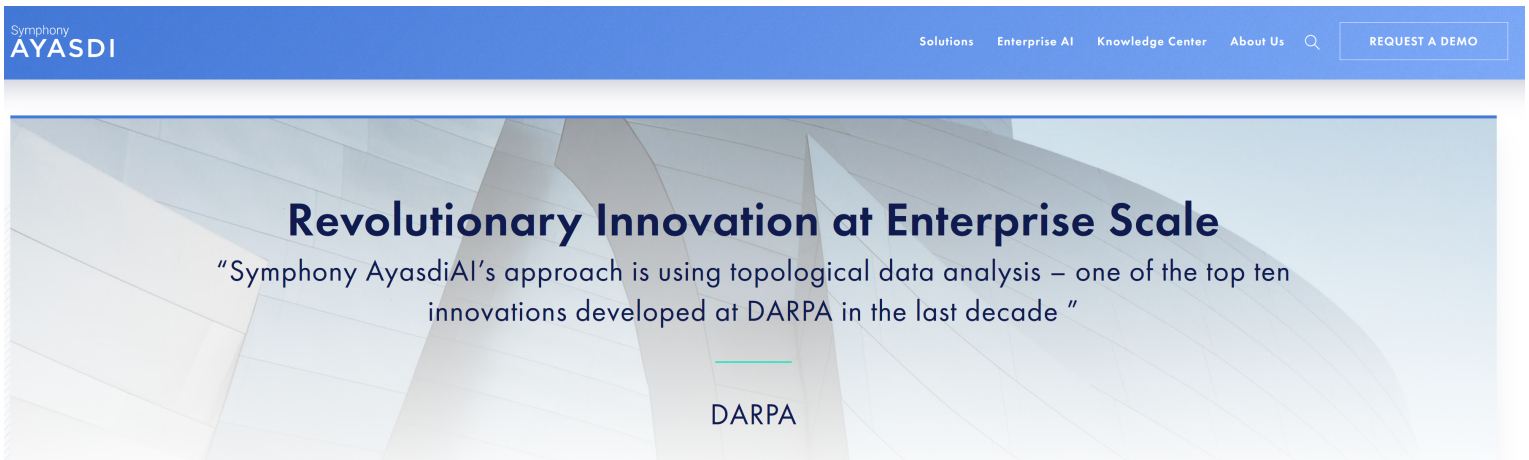
Figure: Two persistence diagrams of purple squares and green circles (left) and the weighted graph for Wasserstein distance computation (right). Figure from Munch et al. (2015).

- Wasserstein distance's stability conditions and results are shown in Cohen-Steiner et al. (2010)

Outline

Mapper Algorithm

- Proposed in Singh et al. (2007)
- The authors, Gunnar Carlsson, Gurjeet Singh, and Facundo Mémoli, who were professors of mathematics at Stanford founded the startup Symphony Ayasdi



Mapper Algorithm

- Data: point cloud data in a metric space
- Define an appropriate filter function
- Put data in overlapping bins
- Find cluster in each bin and construct a network

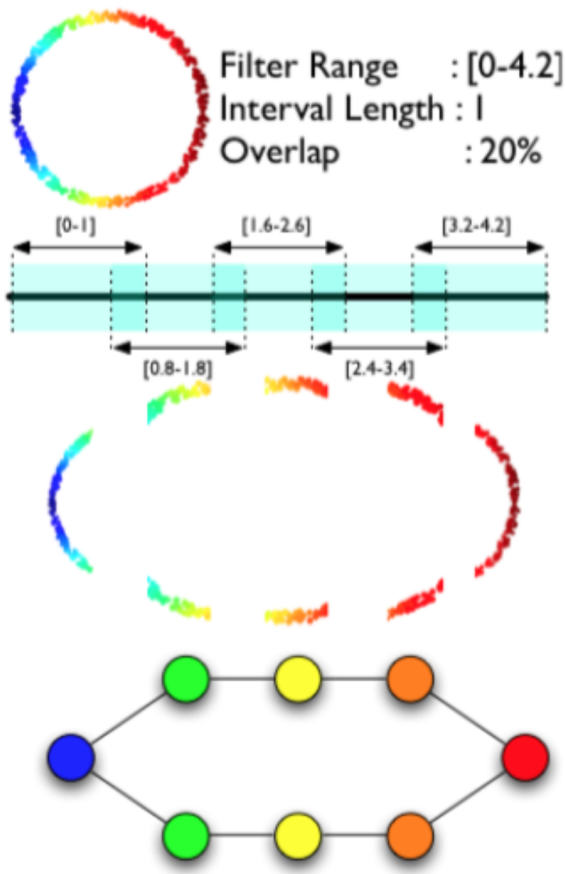


Figure: Figure from Singh et al. (2007)

Mapper Algorithm Toy Example

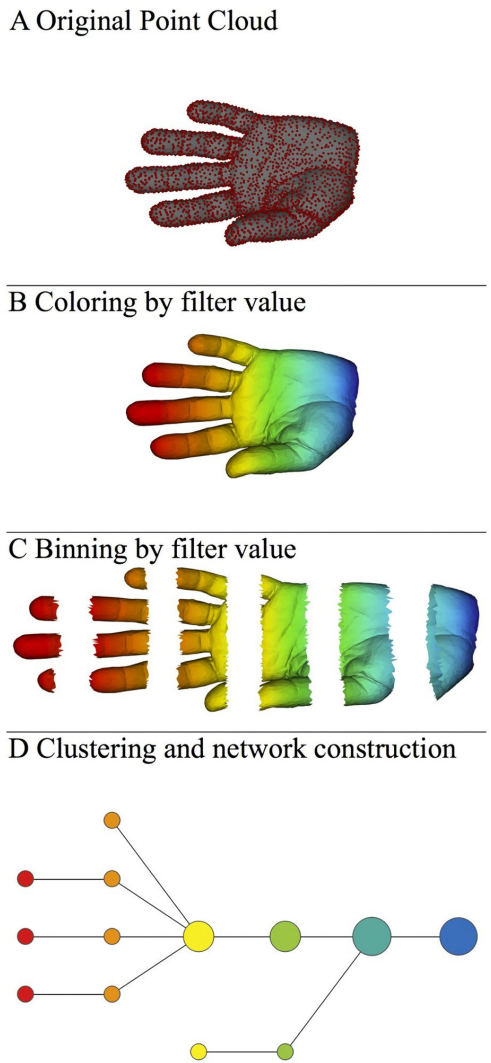


Figure: Figure from Lum et al. (2013)

Mapper Algorithm Toy Example

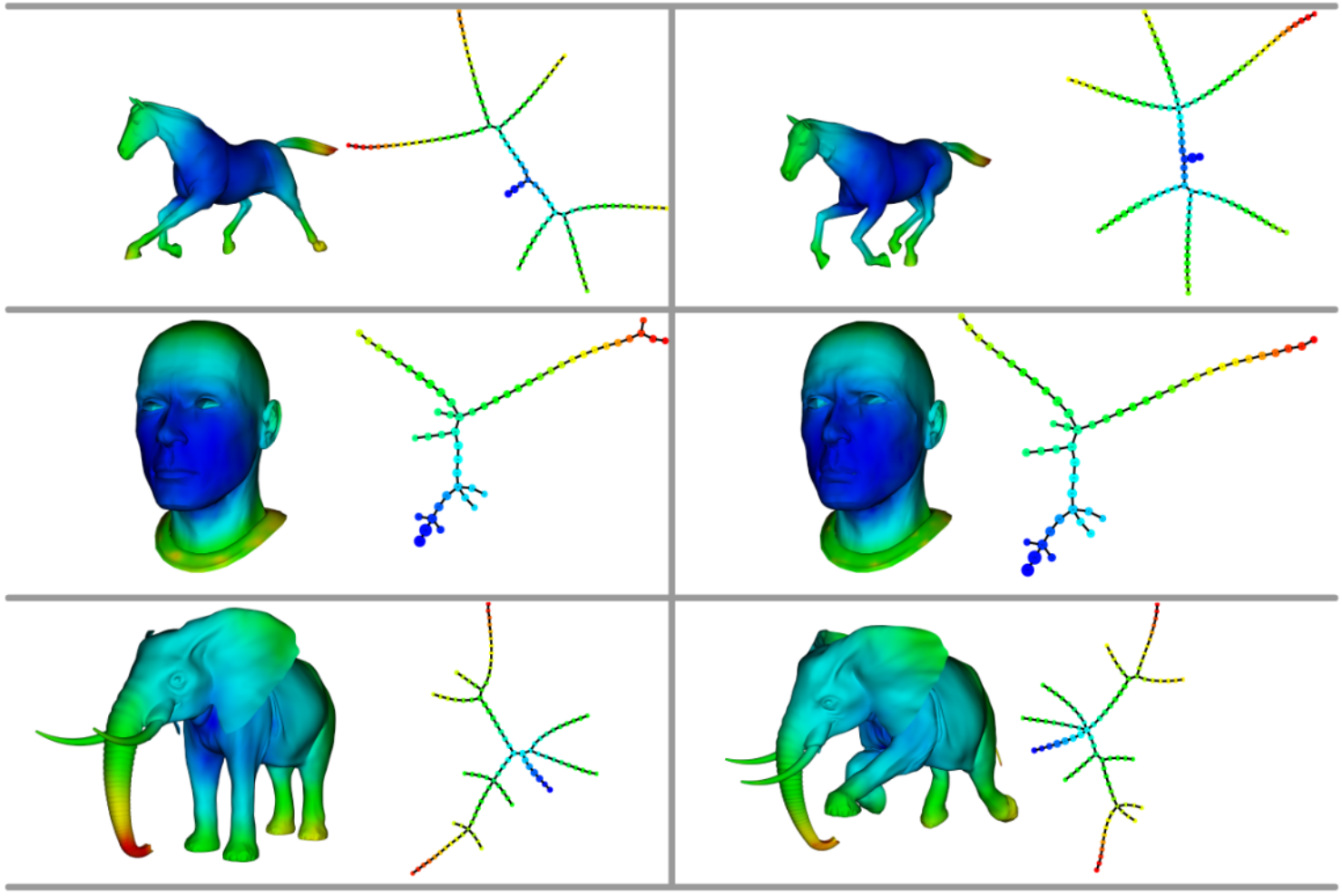


Figure 9: *Refer to Section 5.3 for details. Each row of this image shows two poses of the same shape along with the* `Mapper` *result which is computed as described in Section 5.3. For each* `Mapper` *computation, we used 15 intervals in the range of the filter with a 50% overlap.*

Figure: Figure from Singh et al. (2007)

Applications of Mapper

- Financial fraud detection
  - Ayasdi combined Mapper with machine learning algorithms
- Breast cancer subgroups stratification
- Types of NBA players

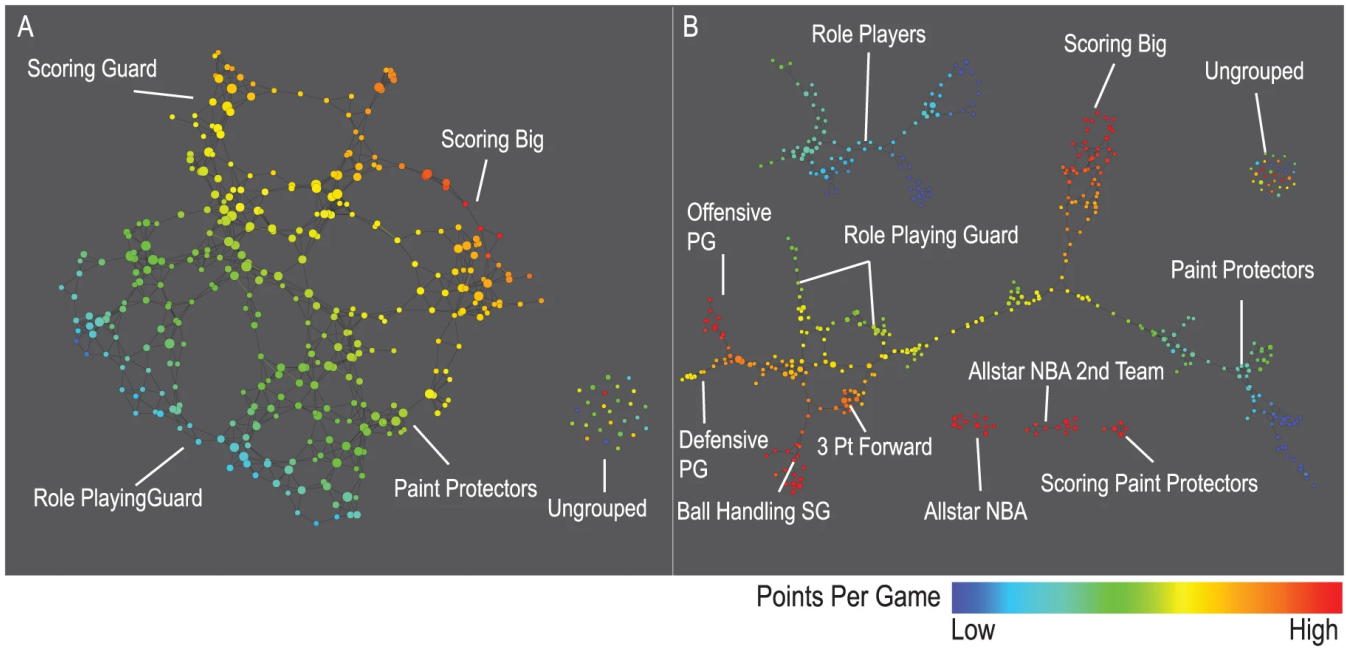


Figure: Figure from Lum et al. (2013)

Applications of Mapper

- Politics

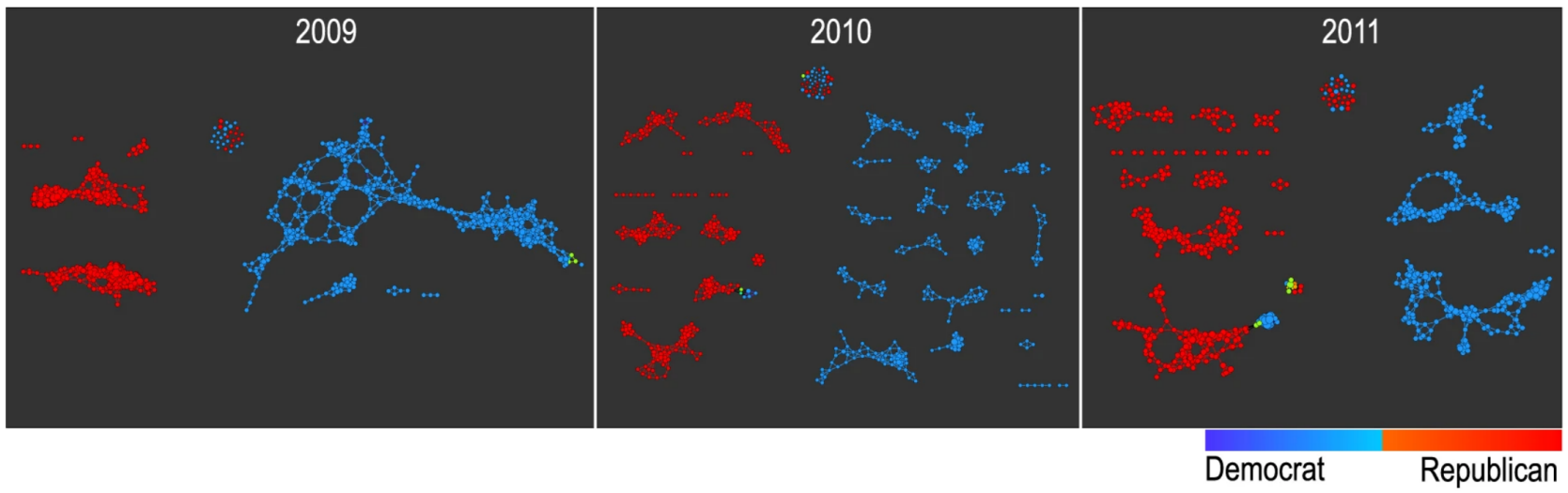


Figure: Figure from Lum et al. (2013)

Usage of Mapper

Packages

- Python
  - Mapper `http://danifold.net/mapper/index.html`
  - KeplerMapper
- R
  - TDAmapper (available via CRAN)
  - Mapper (`https://peekxc.github.io/Mapper/`)

Limitations

- Right filter function?
- How many intervals to use?
- How much buffer to allow?
- Clustering algorithm?

Outline

Motivating Example

- How to identify different shapes?

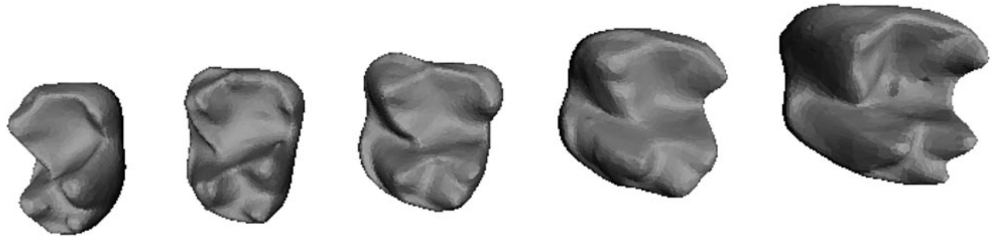


FIG. 1. Images of the meshes of five teeth. A common problem in morphology is to measure distances between these five teeth.

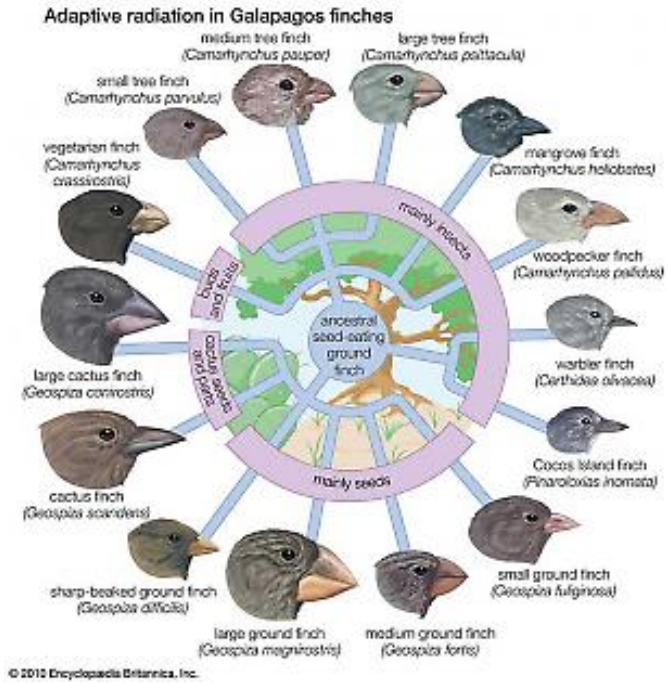


Figure: Figures from Turner et al. (2014) and `http://www.australasianscience.com.au/article/issue-novdec-2018/cultural-evolution-darwin%E2%80%99s-finches.html`

Classical Shape Analysis

- Shape analysis using landmark points

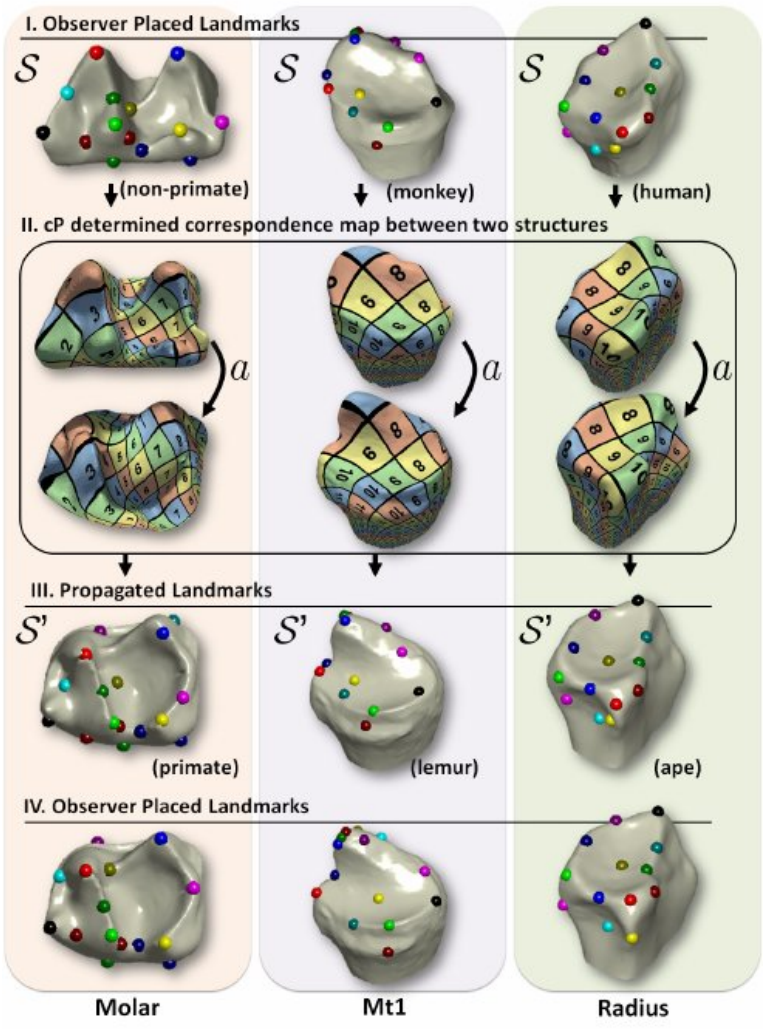


Figure: Figure from Boyer et al. (2011)

PHT and ECT

- Persistent homology transform (PHT) and Euler characteristic transform (ECT) are proposed in Turner et al. (2014)
- Swipe the object in multiple directions and compute persistent homology or Euler characteristics

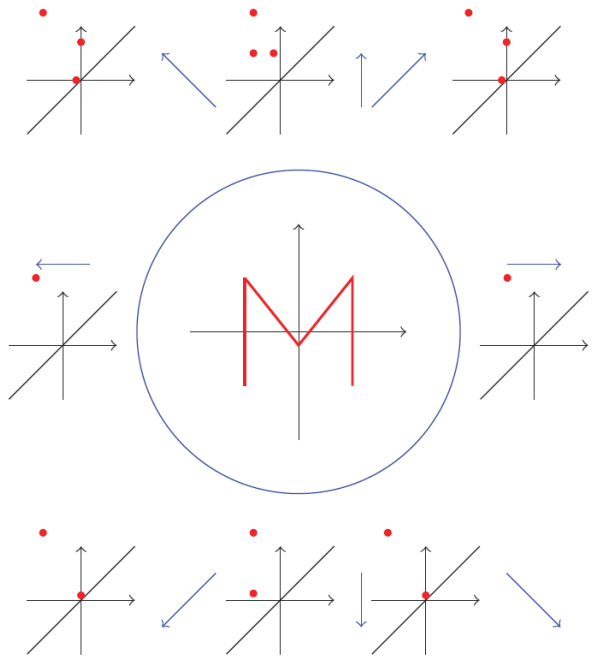


FIG. 5. The 0th dimensional persistence diagrams corresponding to filtrations of the letter M by height functions of eight different directions. The points with ∞ in the second coordinate are represented by a point with second coordinate just above the axes.

Figure: Figure from Turner et al. (2014)

PHT and ECT

- For PHT, persistence diagram is obtained from each direction
- For ECT, an Euler curve is obtaind from each direction
- Both PHT and ECT are injective for two and three dimensional shapes
  - For a given PHT and ECT statistics, there exists a unique corresponding shape
- PHT and ECT enable shape summary without landmarks

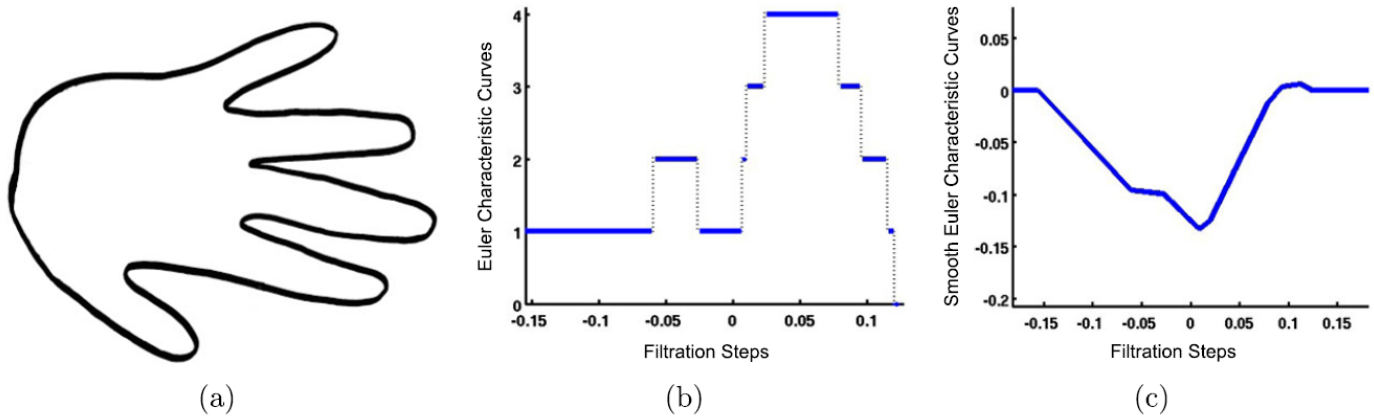
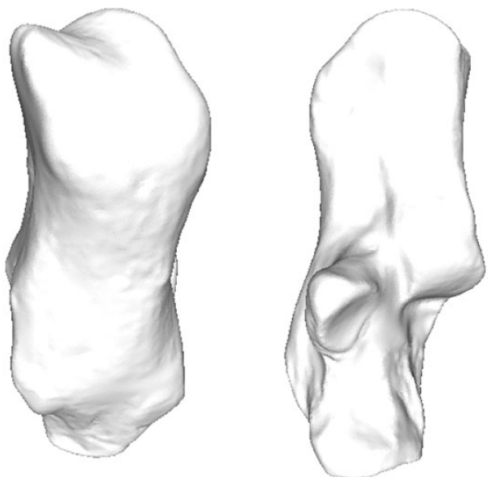


Figure: Figure from Turner et al. (2014)

Applications of PHT

- Would there be variations in heel bones according to genetic variation in species?
- 106 heel bones of extant and extinct primates



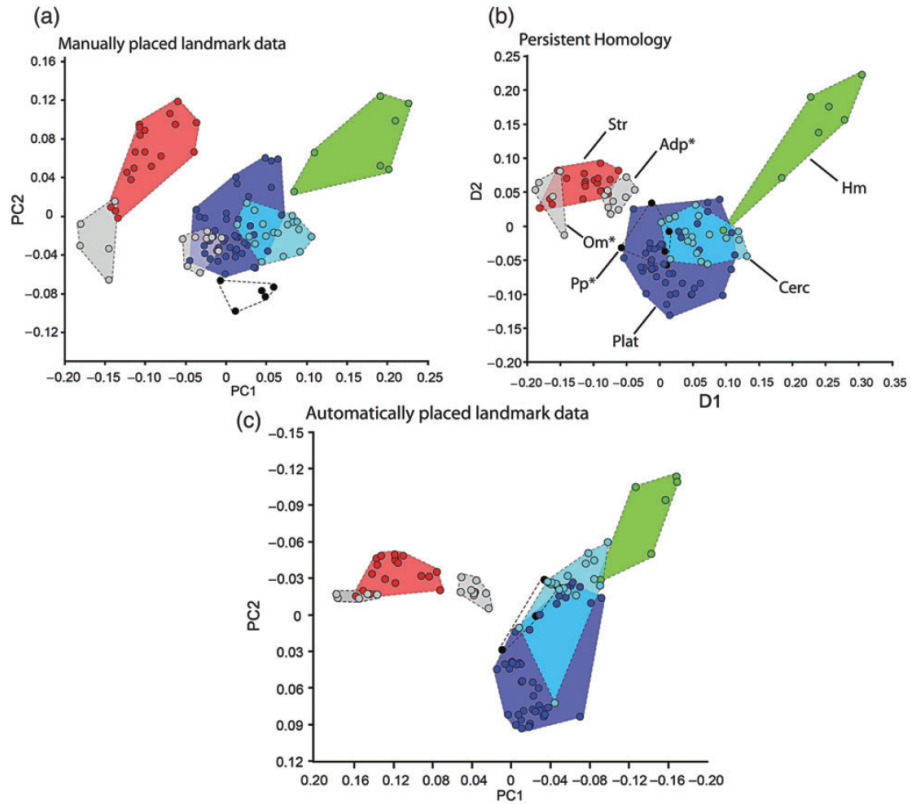


Figure: Figures from Turner et al. (2014)

Applications of (Smoothed) ECT

- Prediction of survival of brain tumor patients using tumor shapes in MRI images
- Smoothed ECT is a good predictor of survival outcomes

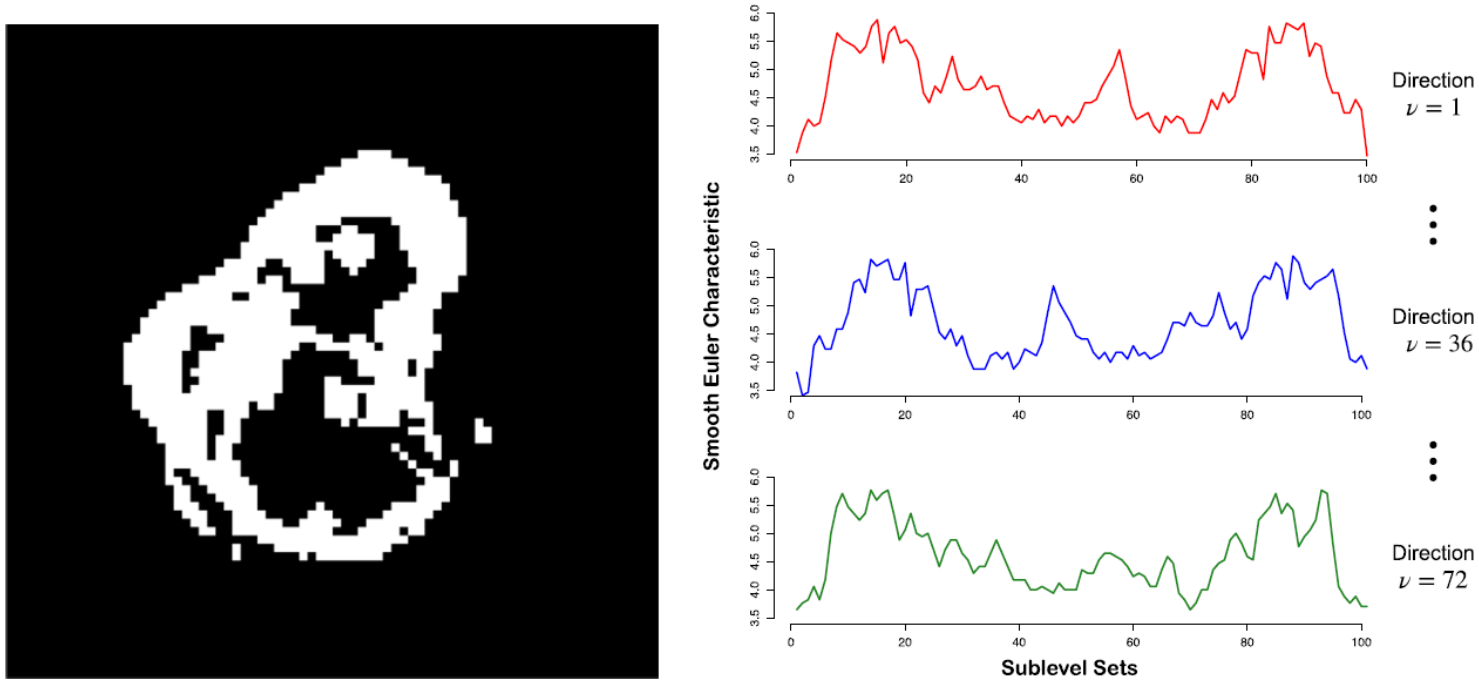


Figure: Figure from Crawford et al. (2020)

References I

Atallah, M. J. and Blanton, M. (2009), *Algorithms and theory of computation handbook, volume 2: special topics and techniques*, CRC press.

Boyer, D. M., Lipman, Y., Clair, E. S., Puente, J., Patel, B. A., Funkhouser, T., Jernvall, J., and Daubechies, I. (2011), "Algorithms to automatically quantify the geometric similarity of anatomical surfaces," *Proceedings of the National Academy of Sciences*, 108, 18221–18226.

Chazal, F., de Silva, V., Glisse, M., and Oudot, S. (2012), "The Structure and Stability of Persistence Modules," *ArXiv e-prints*.

Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007), "Stability of Persistence Diagrams," *Discrete & Computational Geometry*, 37, 103–120.

Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. (2010), "Lipschitz Functions Have L_p-Stable Persistence," *Foundations of Computational Mathematics*, 10, 127–139.

Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R. (2020), "Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis," *Journal of the American Statistical Association*, 115, 1139–1150.

References II

Fasy, B. T., Kim, J., Lecci, F., and Maria, C. (2014), "Introduction to the R package TDA," *arXiv preprint arXiv:1411.1830*.

Ghrist, R. (2008), "Barcodes: The Persistent Topology of Data," *Bulletin of the American Mathematical Society*, 45, 61–75.

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G. (2013), "Extracting insights from the shape of complex data using topology," *Scientific reports*, 3, 1–8.

MATLAB (2016), *version 9.10.0 (R2016b)*, Natick, MA: The MathWorks Inc.

Mula, J., Lee, J. D., Liu, F., Yang, L., and Peterson, C. A. (2013), "Automated Image Analysis Of Skeletal Muscle Fiber Cross-sectional Area," *Journal of Applied Physiology*, 114, 148–155.

Munch, E., Turner, K., Bendich, P., Mukherjee, S., Mattingly, J., and Harer, J. (2015), "Probabilistic Fréchet Means for Time Varying Persistence Diagrams," *Electronic Journal of Statistics*, 9, 1173–1204.

Pidwirny, M. (2006), *Fundamentals of Physical Geography*, `http://www.physicalgeography.net/`.

References III

Singh, G., Mémoli, F., Carlsson, G. E., et al. (2007), "Topological methods for the analysis of high dimensional data sets and 3d object recognition." *PBG@ Eurographics*, 2.

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., and Hess, K. (2021), "giotto-tda:: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration." *J. Mach. Learn. Res.*, 22, 39–1.

Turner, K., Mukherjee, S., and Boyer, D. M. (2014), "Persistent homology transform for modeling shapes and surfaces," *Information and Inference: A Journal of the IMA*, 3, 310–344.

Wang, F., Deng, C., Yuan, B., and Chen, C. (2019), "Hardware Acceleration of Persistent Homology Computation," in *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, Springer, pp. 81–88.

Zomorodian, A. and Carlsson, G. (2005), "Computing Persistent Homology," *Discrete & Computational Geometry*, 33, 249–274.