Functional Data Analysis using Topological Summary Statistics

NIPS 2017: Synergies in Geometric Data Analysis Workshop

Lorin Crawford

Department of Biostatistics
Center for Statistical Sciences
Center for Computational Molecular Biology
Brown University

In Collaboration with:
Anthea Monod, Andrew Chen, Raúl Rabadán (Columbia University),
and Sayan Mukherjee (Duke University)

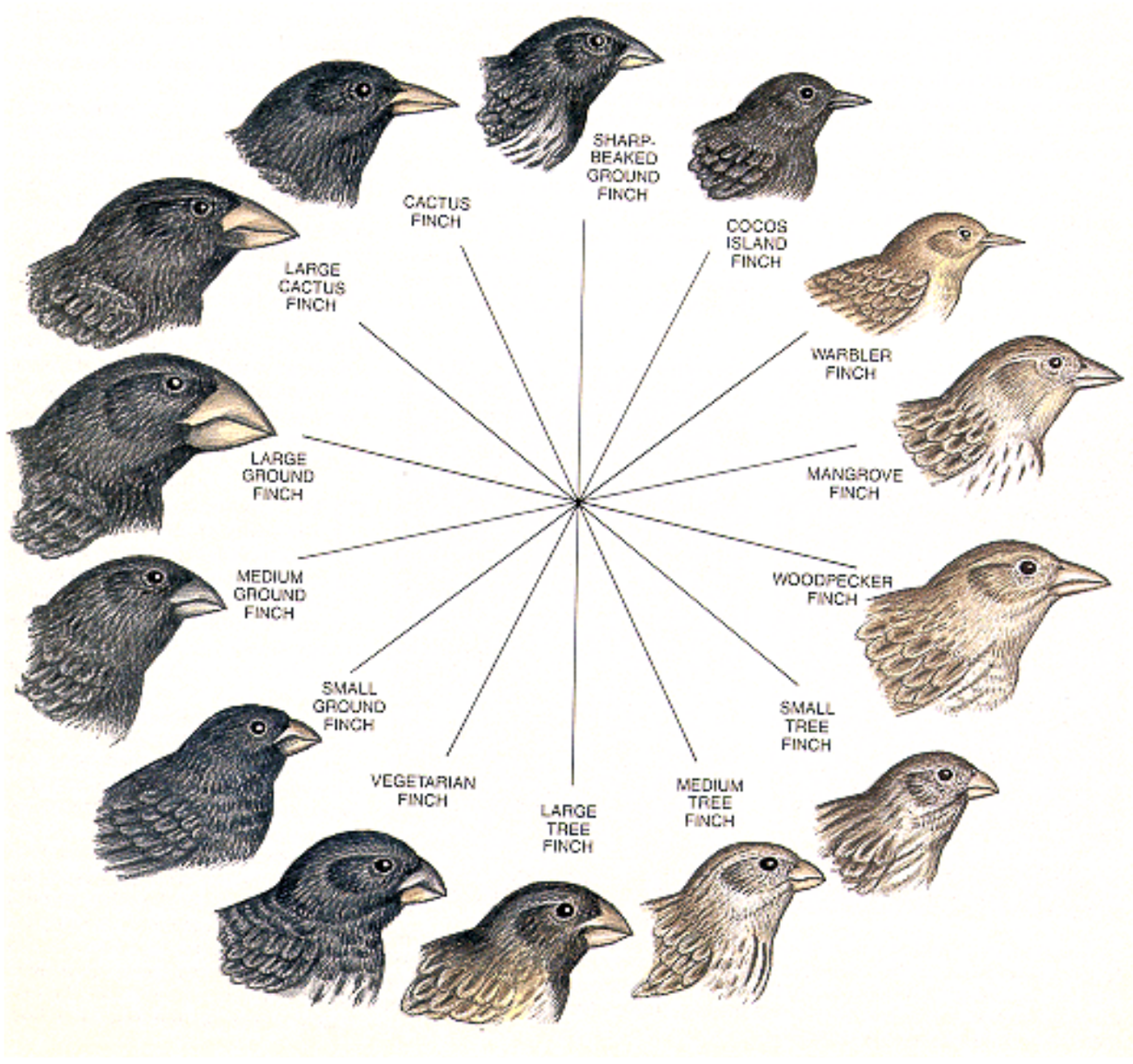
December 8, 2017

BROWN
School of Public Health

Key Concepts and Terms

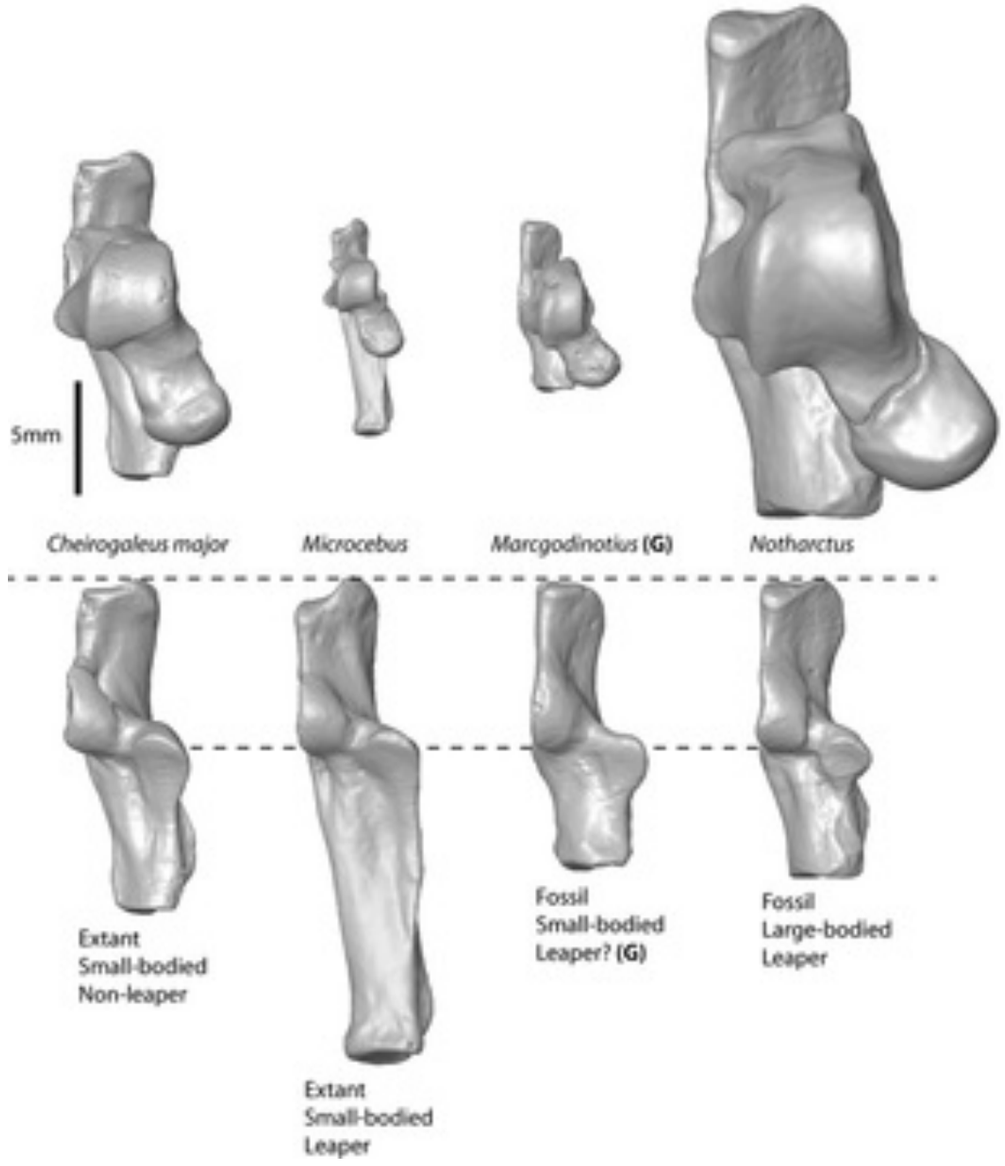
- **Topological Data Analysis (TDA):**
  - Combines algebraic topology and other tools from pure mathematics to give mathematically rigorous and quantitative study of “shape”
- **Functional Data Analysis (FDA):**
  - An area of statistics where it is of key interest to analyze data providing information about curves, surfaces, images, and any other variables that vary over a given continuum

Modeling Variation across Shapes



Phylogeny of Darwin's Finch Beaks

[Gould (1977)]



Fossil Classification

[Boyer et al. (2011)]

History of Shape Statistics

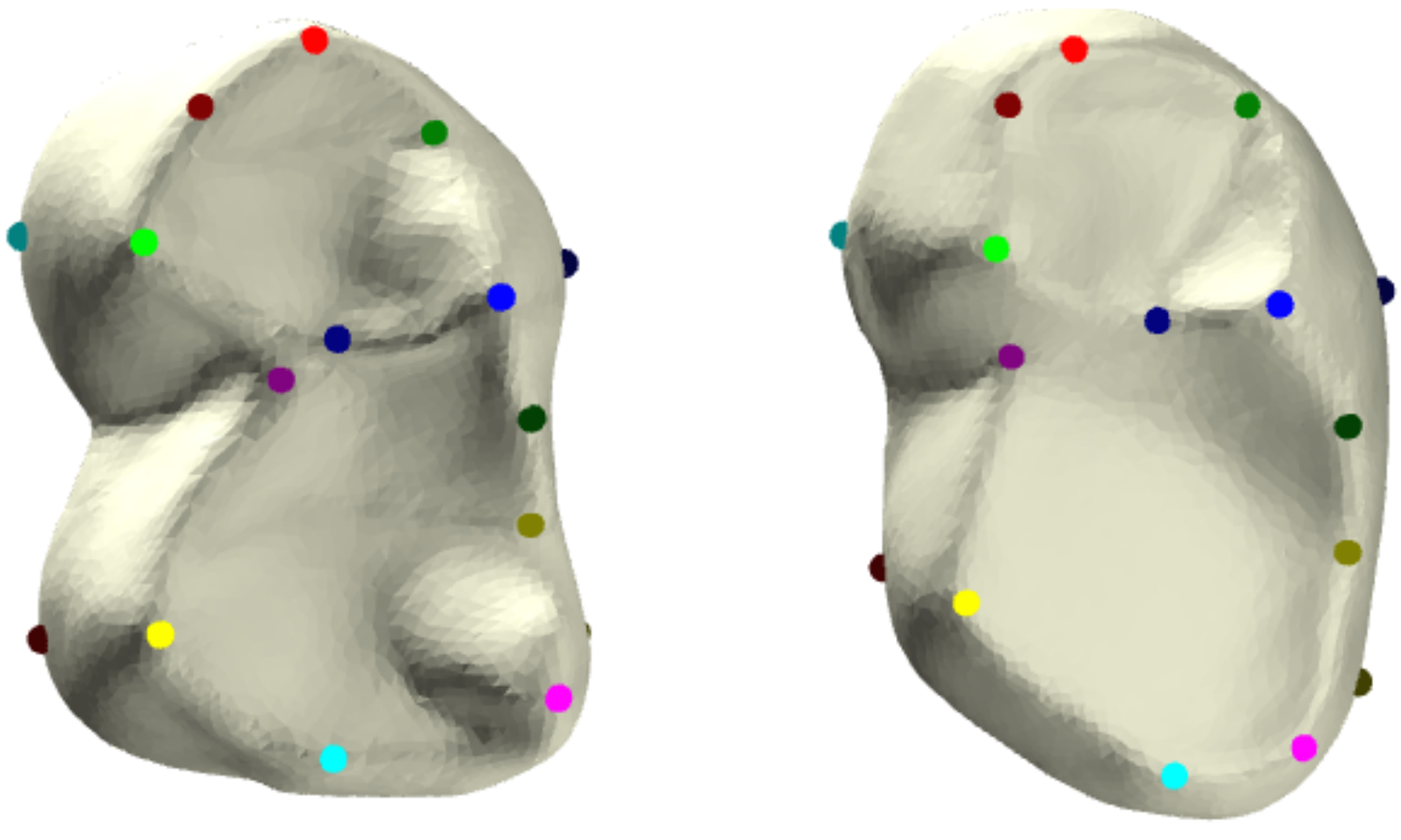
- Classical shape statistics represented three-dimensional shapes as user defined landmark points placed on the shape.

- This representation was partly due to the limited imaging and processing technology of the time.

- Computational methodology that effectively incorporate information embedded in three-dimensional shapes simply did not exist.

Shape Representations

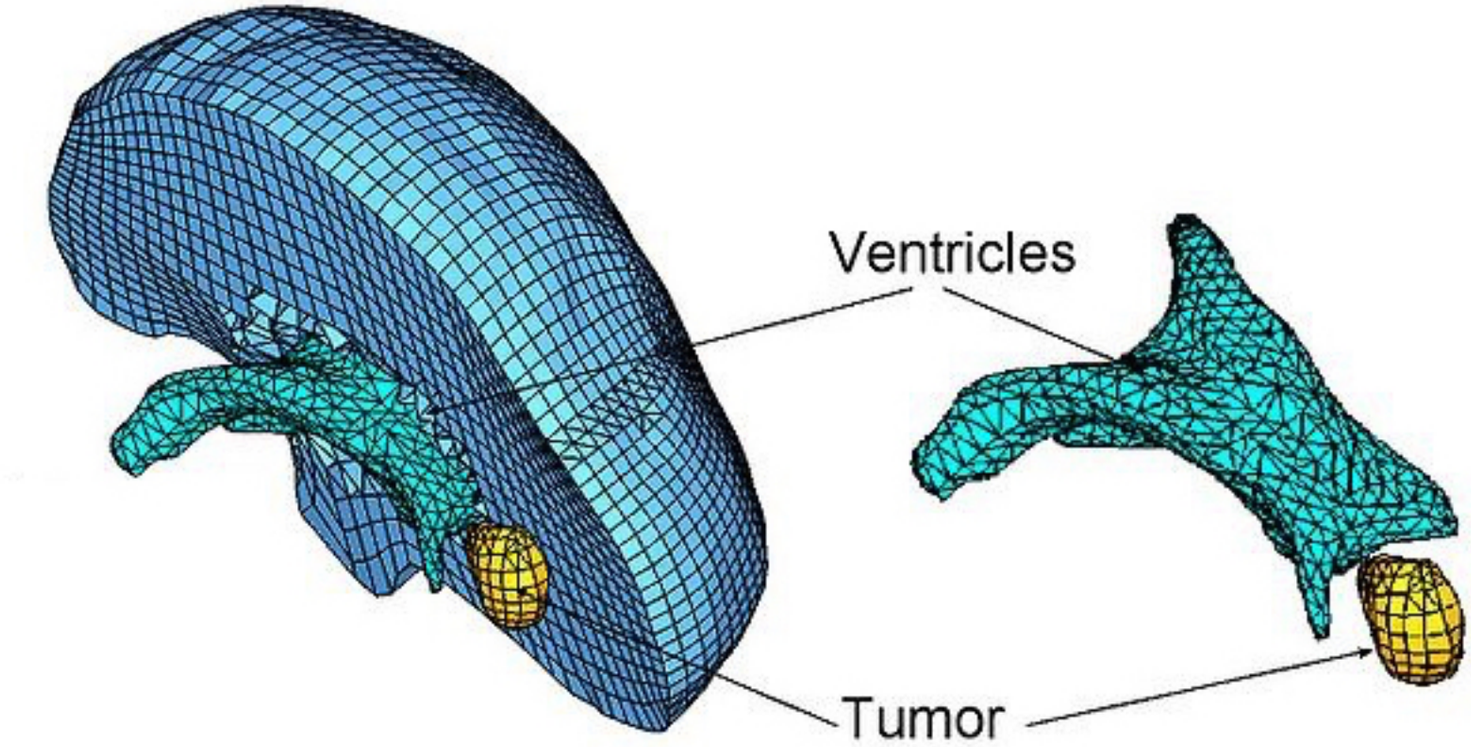
- Methods have been developed to generate automated geometric morphometrics for shapes, bypassing the need for user-specified landmarks



[Boyer et al. (2011)]

Shape Representations

- Currently, much improved imaging technologies allow three-dimensional shapes to be represented as meshes --- a collection of vertices, faces, and edges



[Boyer et al. (2011)]

Motivation

- Methods for geometric morphometrics are known to suffer from structural errors when comparing shapes that are highly dissimilar.

- These analyses require the specification of a metric, which is not always a straightforward task.

- Turner et al. (2014) developed a statistical summary of shape data known as the persistent homology transform (PHT).

  - The PHT bypasses the need to specify landmarks, and is robust to highly dissimilar and non-isomorphic shapes.

Motivation

But more needs to be done to fully integrate TDA measures with FDA methods…

Main Objective(s)

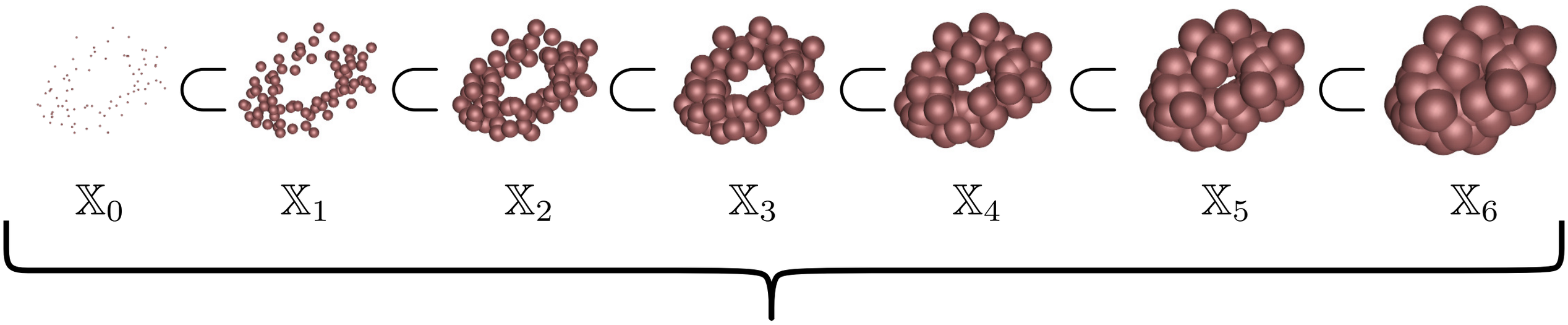
- **Transform shapes or images into a representation that can be used in wide range of functional data analytic methods (e.g. generalized functional linear models, GFLMs)**

- **Desired Transformation Properties:**
  - Injective mapping, so that the resulting measures are summary statistics
  - We want to be able to compute distances or define probabilistic models in the transformed space

- **Topological Summaries:**
  - Persistent Homology Transform (PHT)
  - Smooth Euler Characteristic Transform (SECT)

Persistent Homology

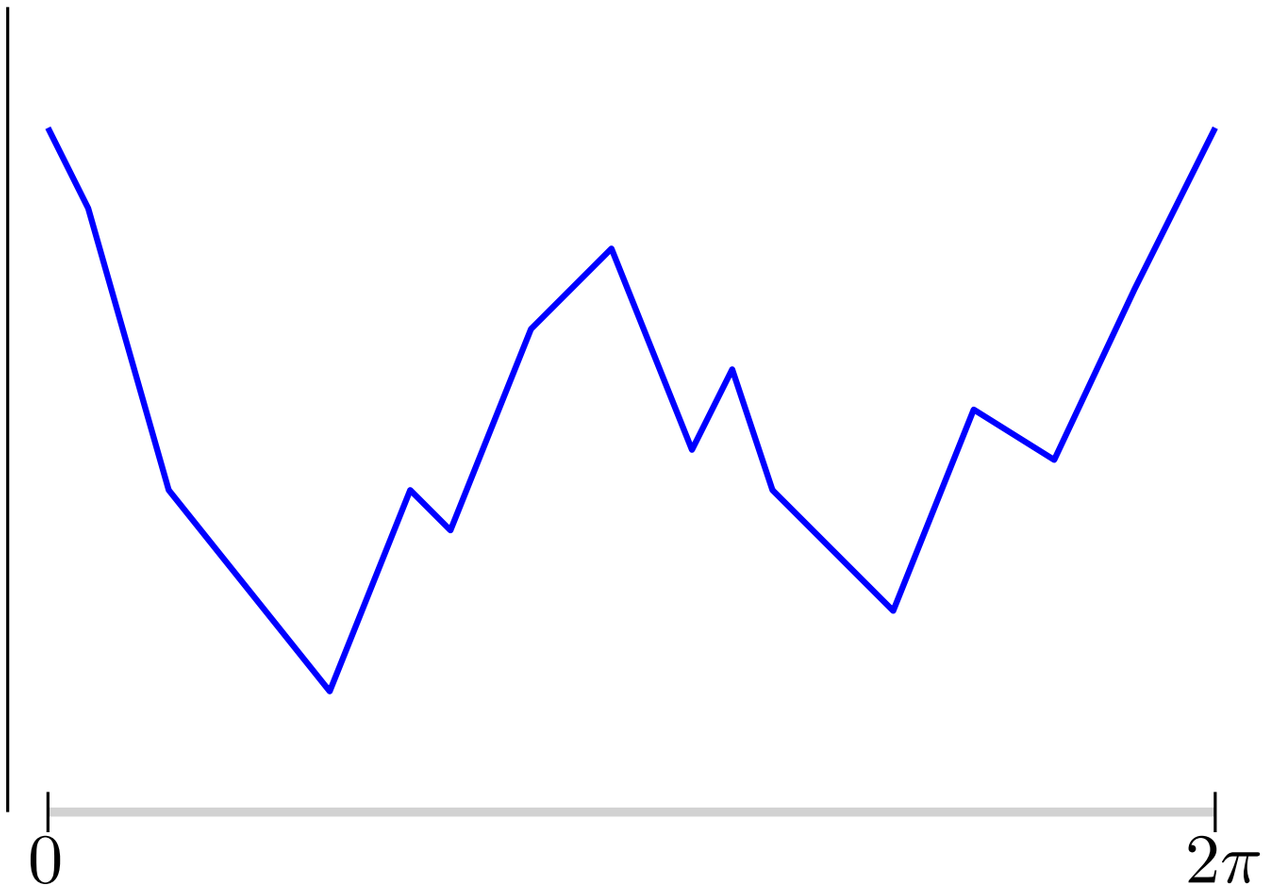
Construct a filtration $\mathcal{K}$

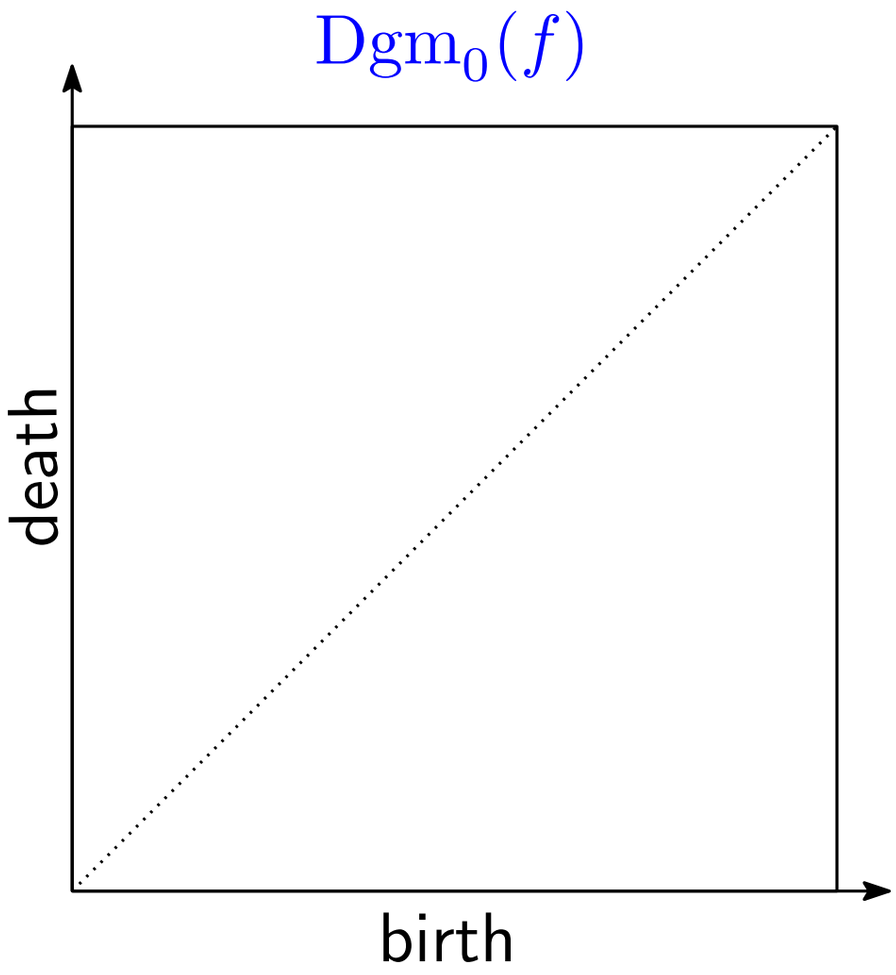


The *persistent homology* of $\mathcal{K}$, denoted by $\mathrm{PH}_*(\mathcal{K})$, keeps track of the progression of homology groups generated by the filtration

Persistent Homology

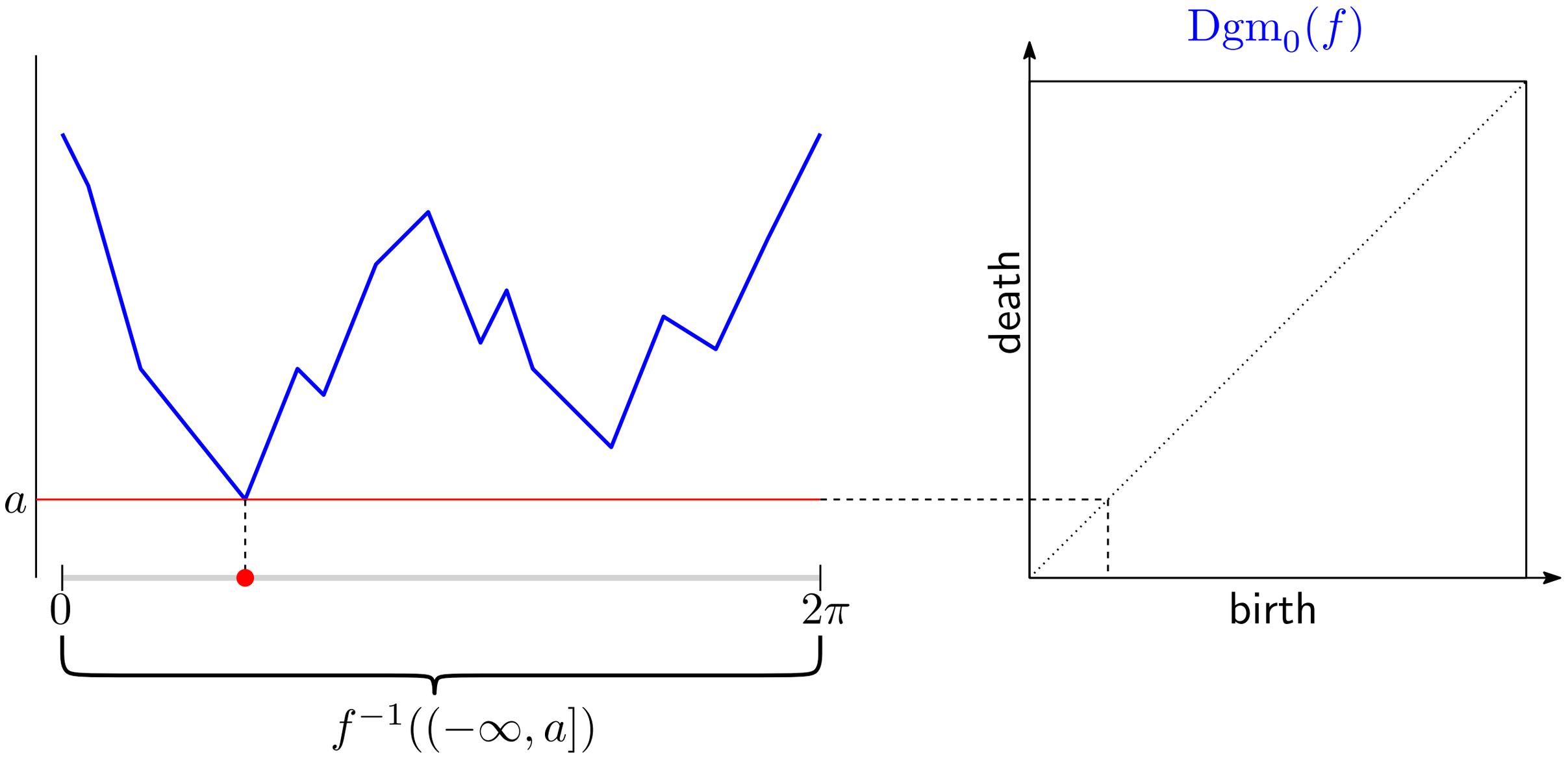
Evolution of homology as a birth-death pair.





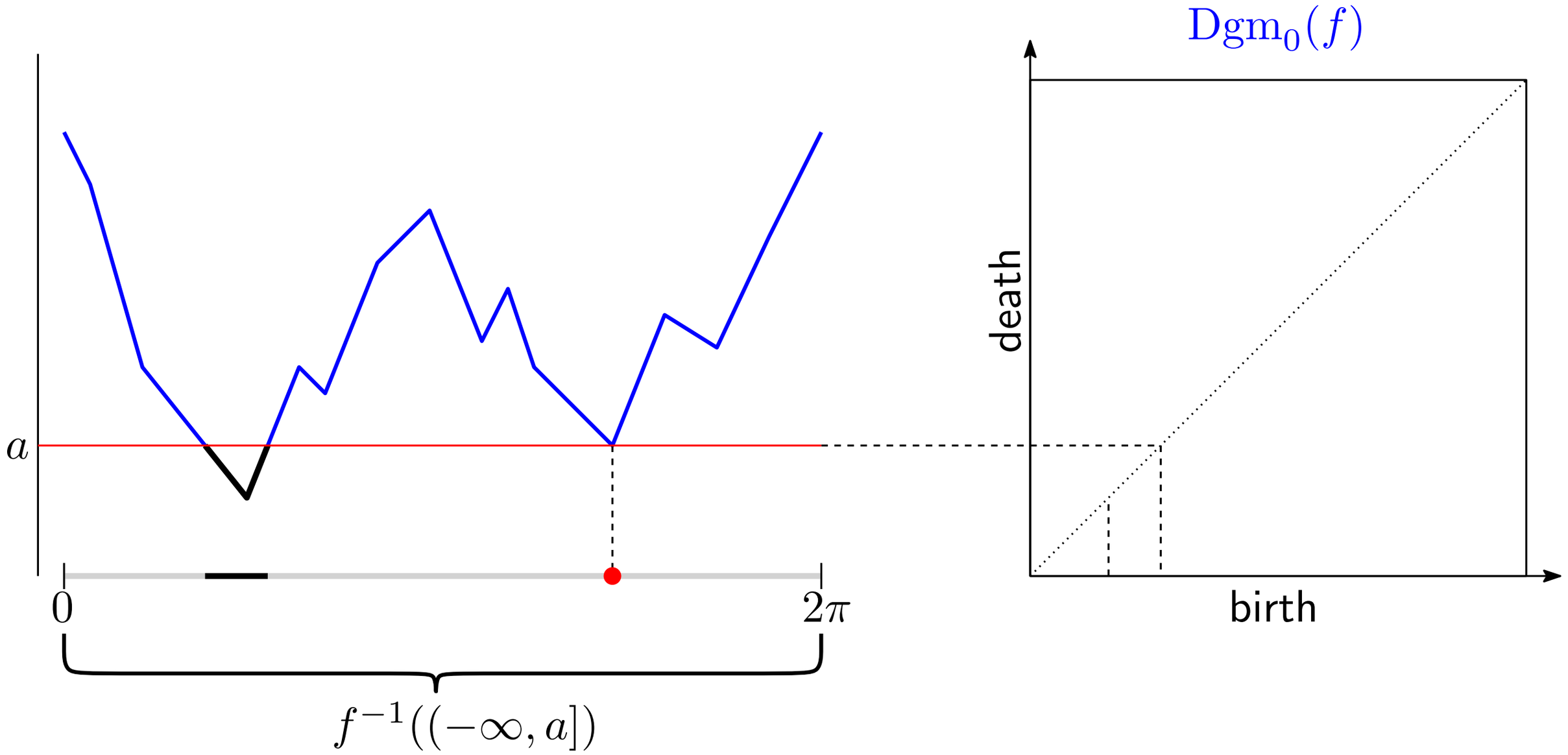
Persistent Homology

Evolution of homology as a birth-death pair.



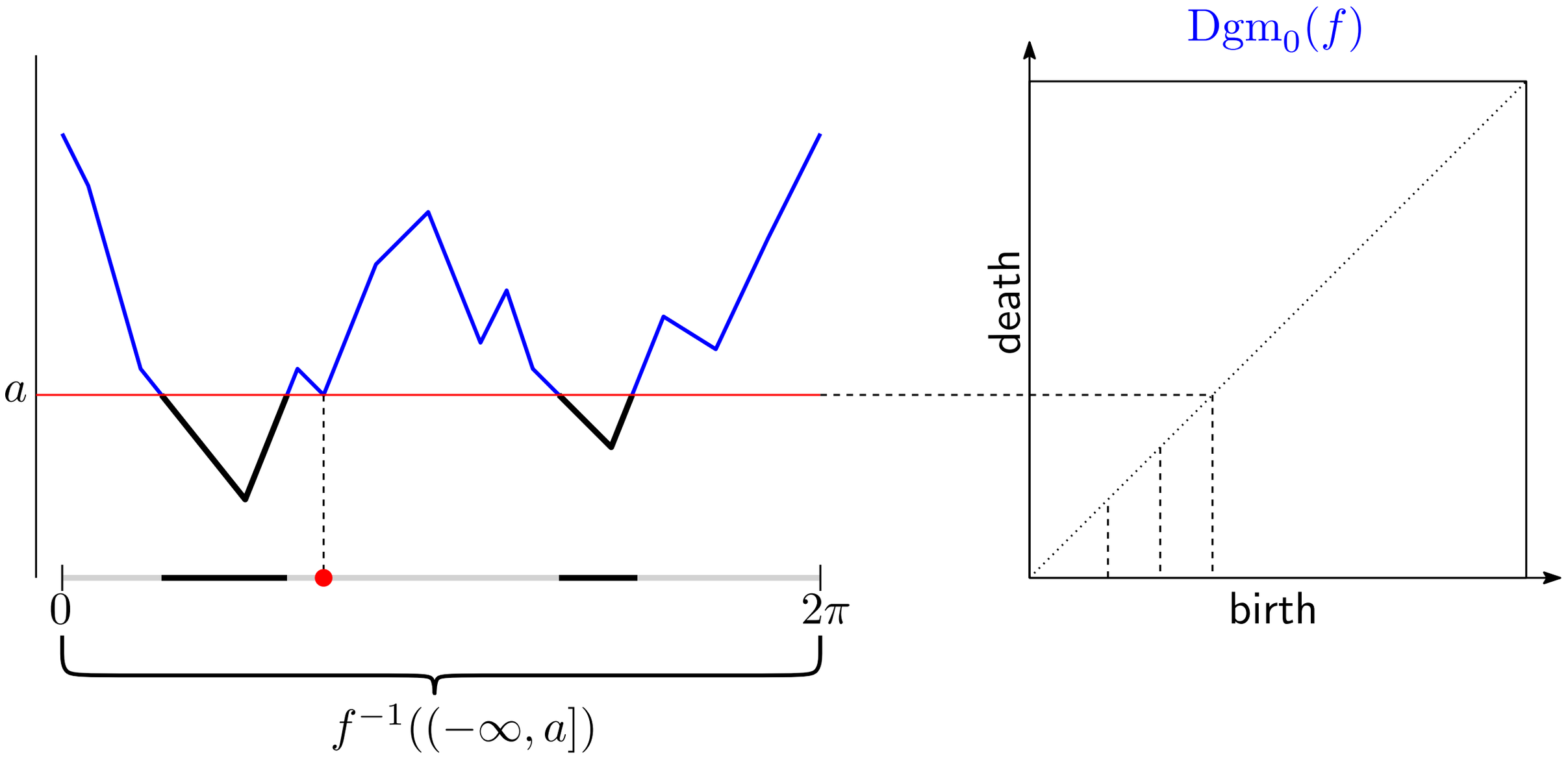
Persistent Homology

Evolution of homology as a birth-death pair.



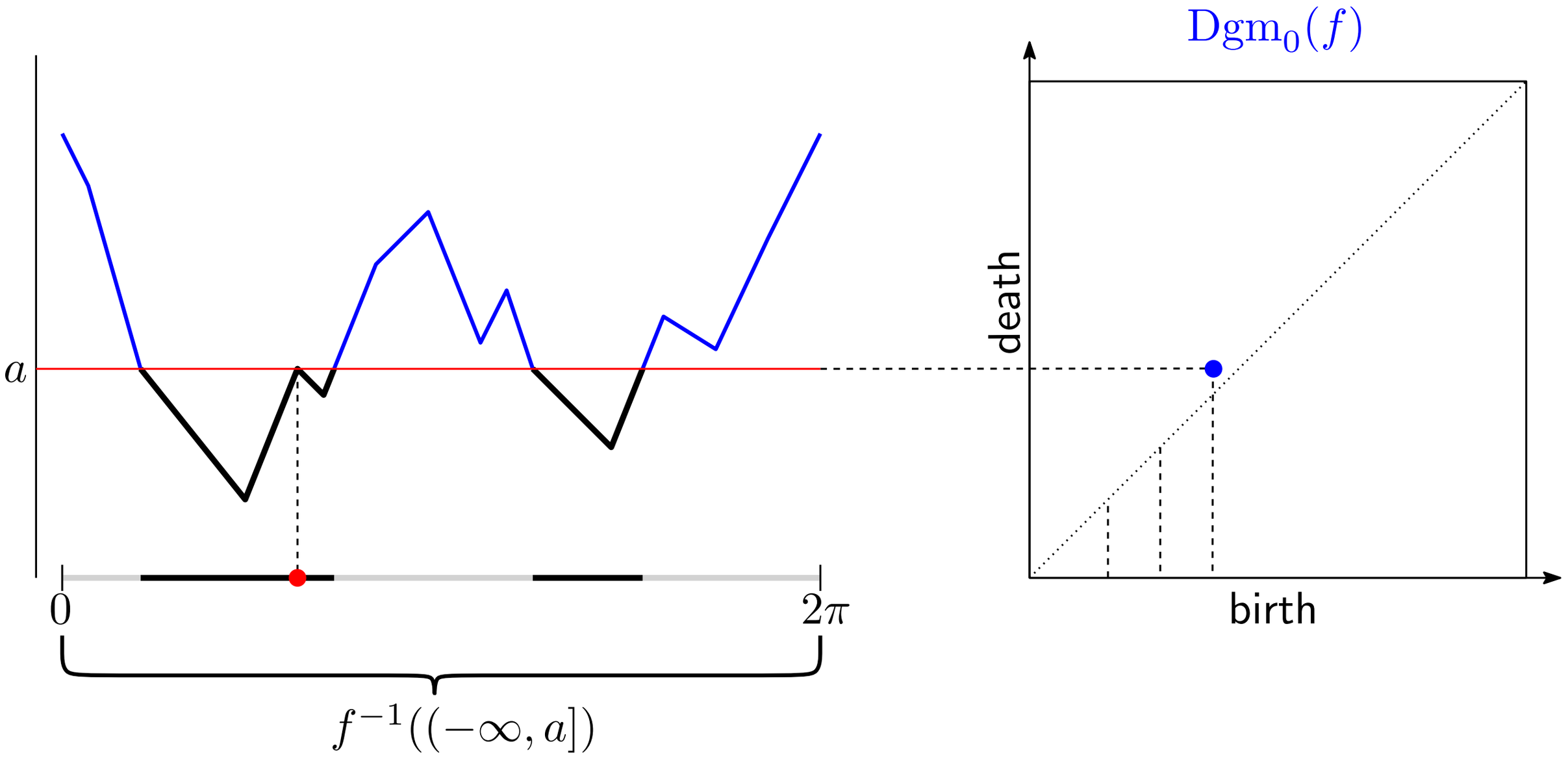
Persistent Homology

Evolution of homology as a birth-death pair.



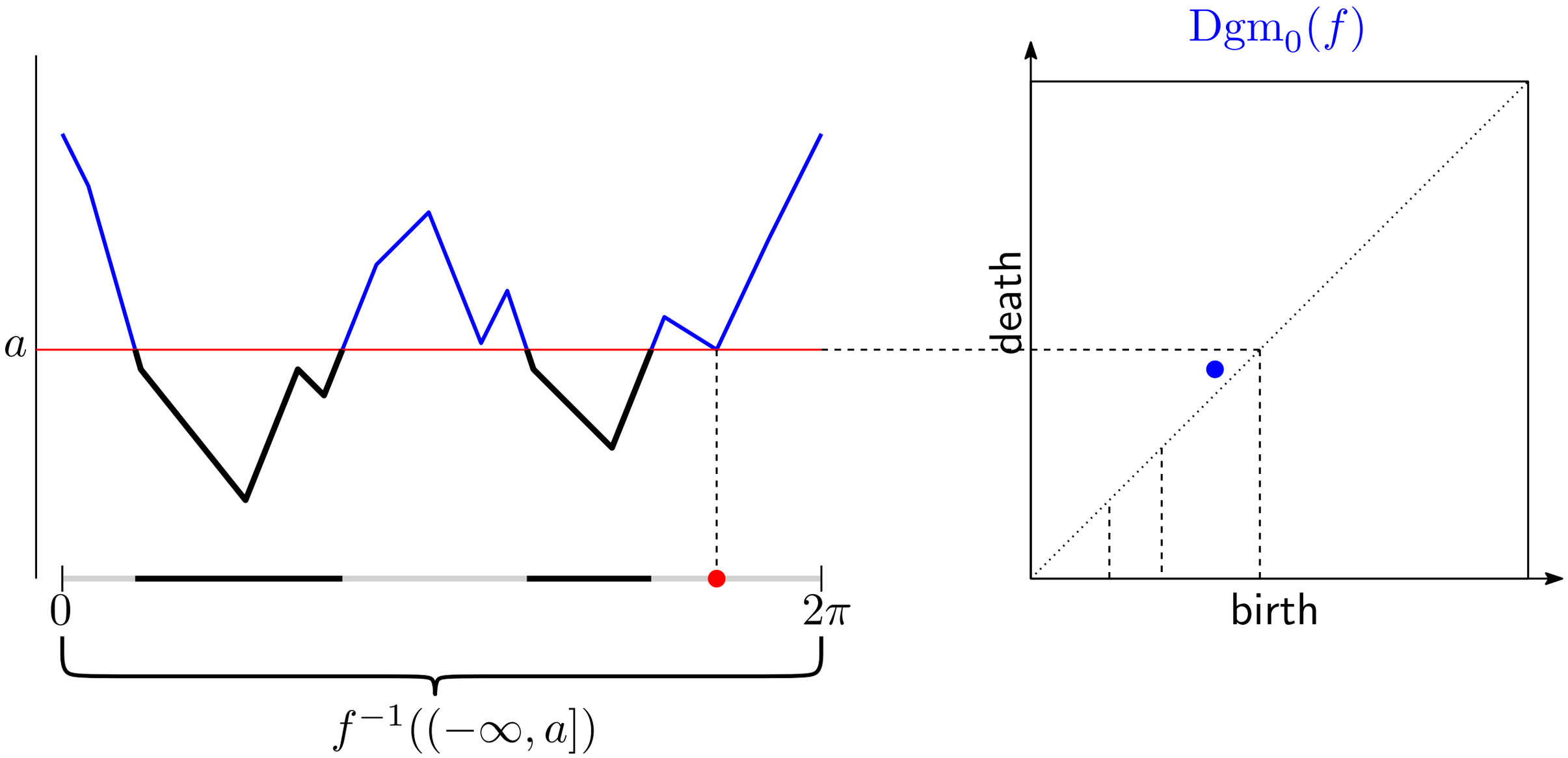
Persistent Homology

Evolution of homology as a birth-death pair.



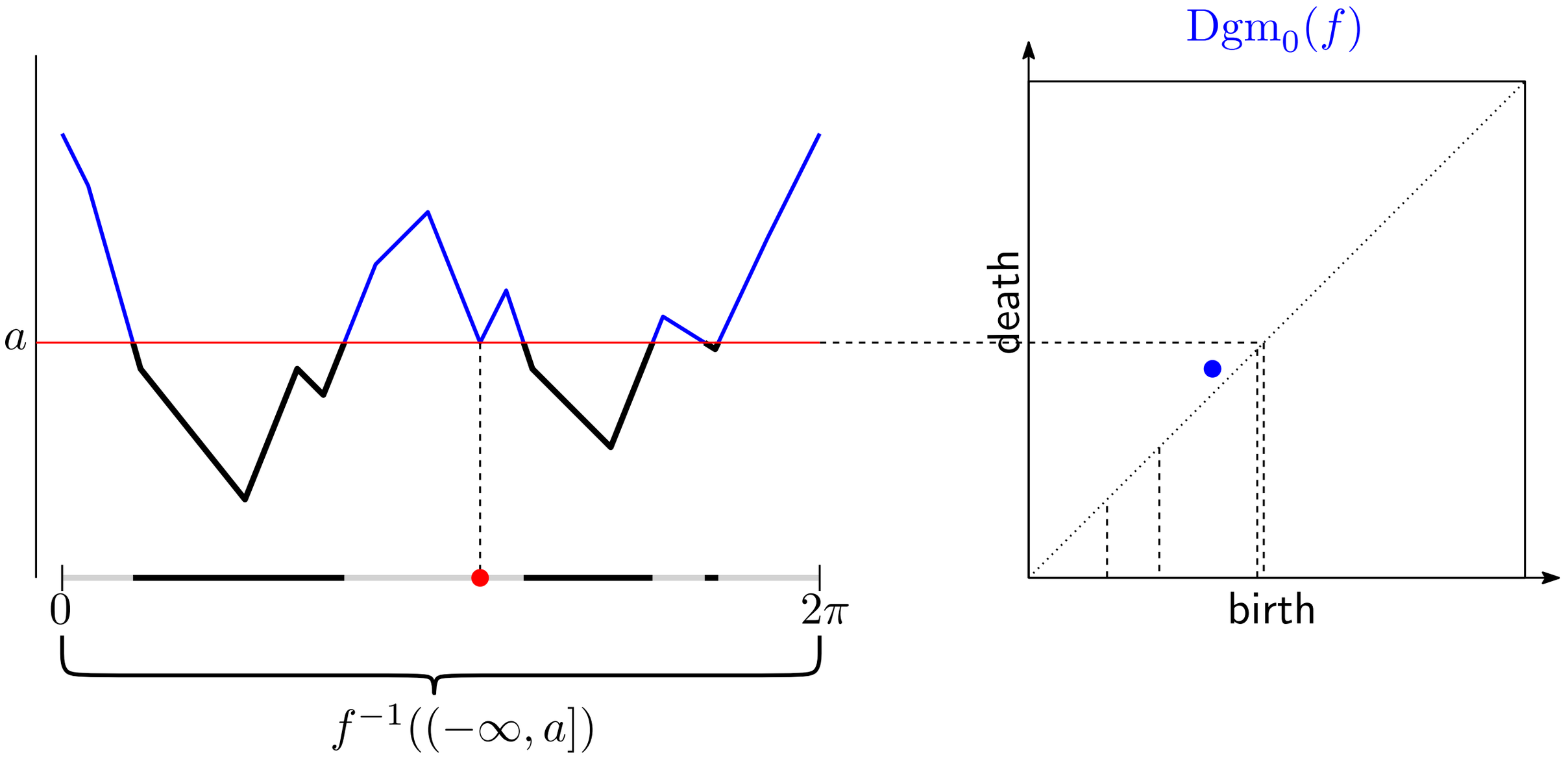
Persistent Homology

Evolution of homology as a birth-death pair.



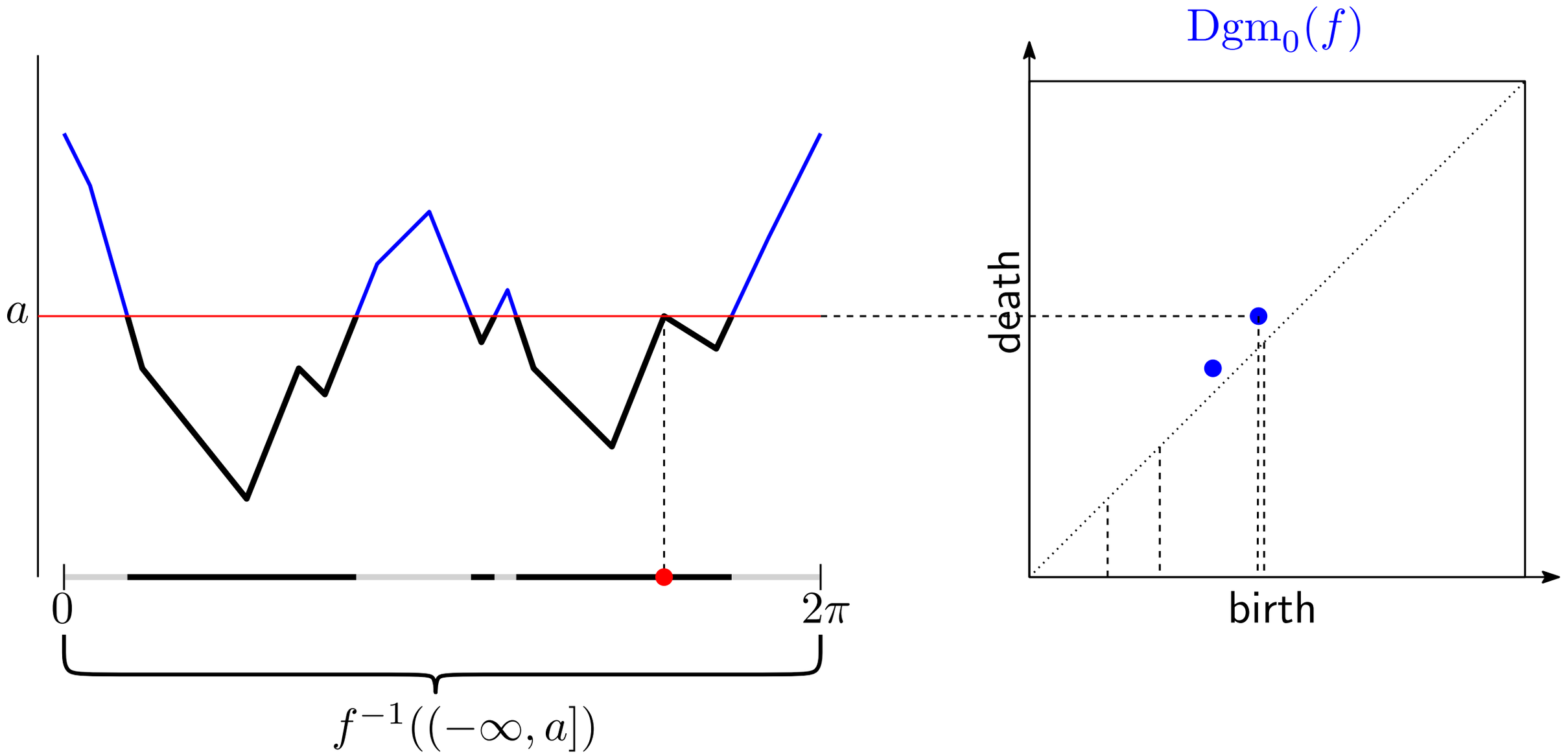
Persistent Homology

Evolution of homology as a birth-death pair.



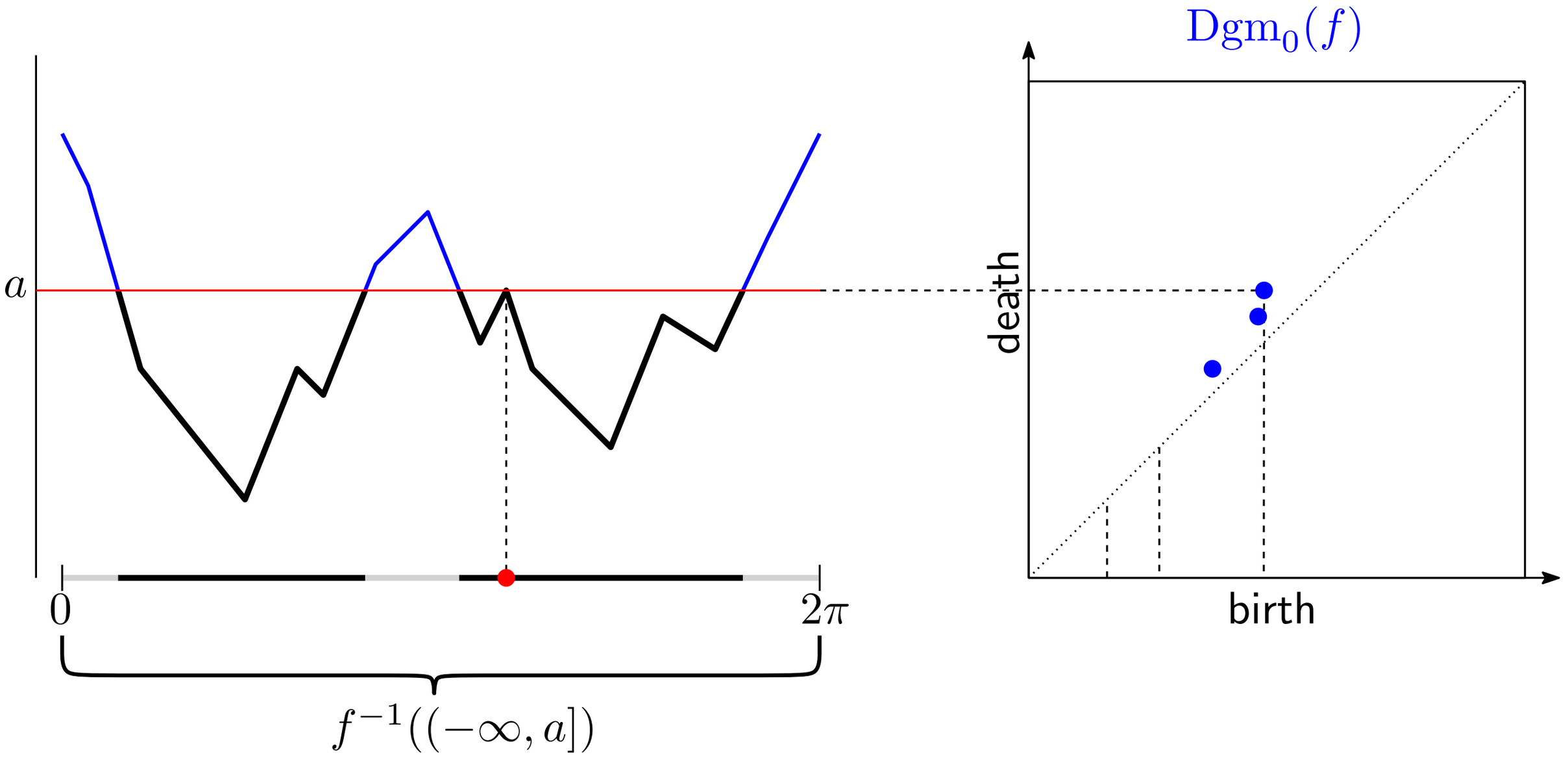
Persistent Homology

Evolution of homology as a birth-death pair.



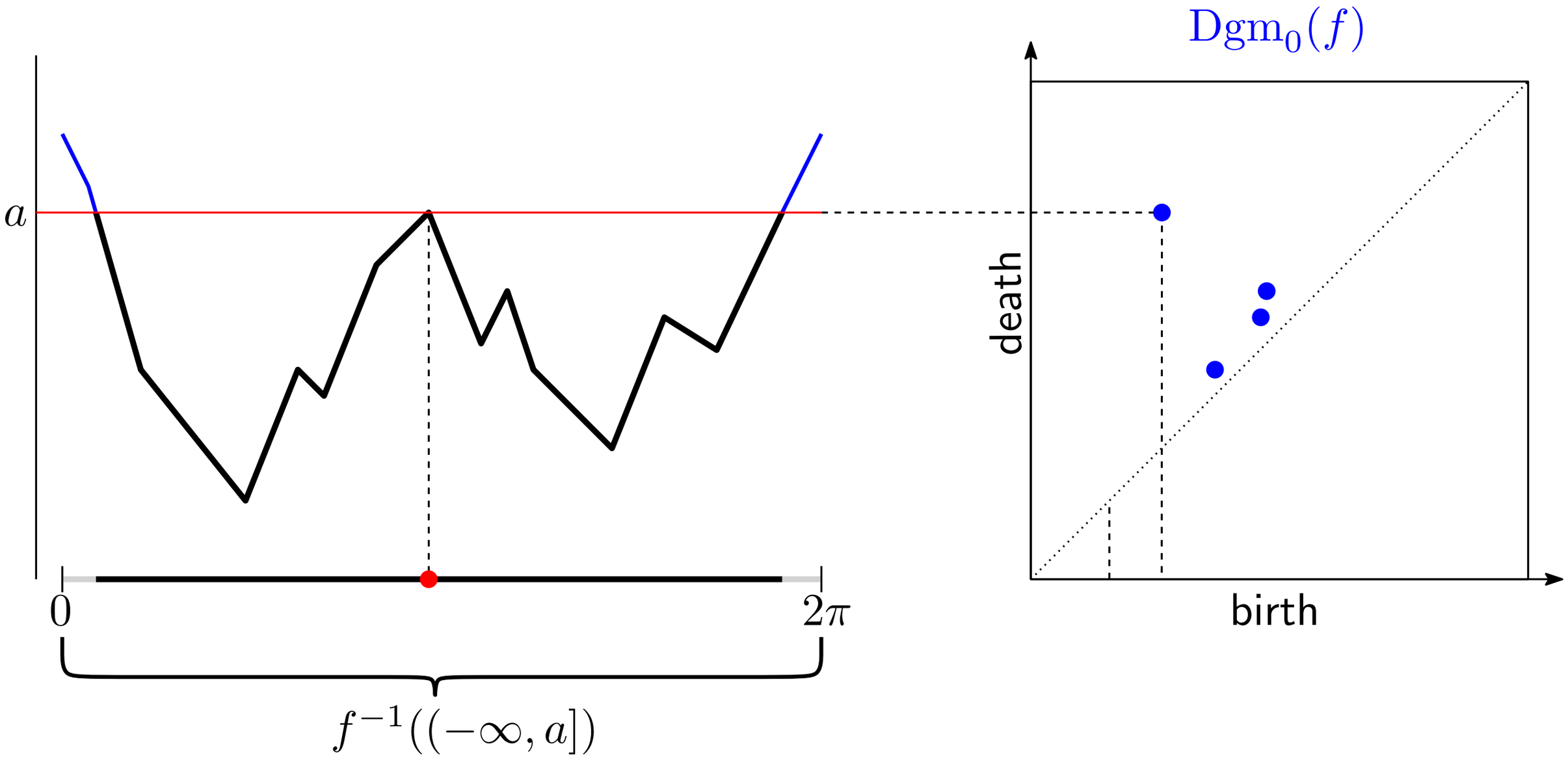
Persistent Homology

Evolution of homology as a birth-death pair.



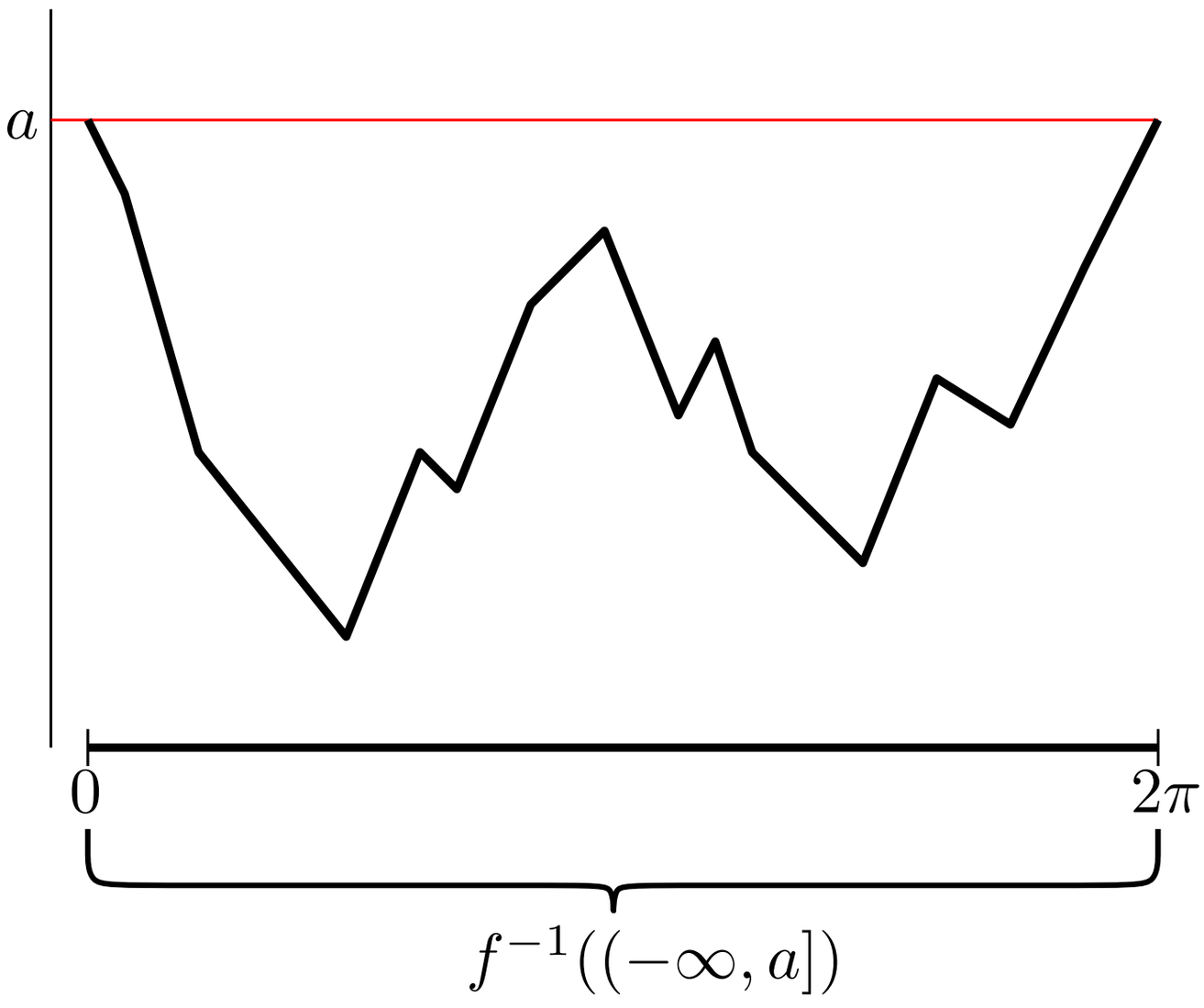
Persistent Homology

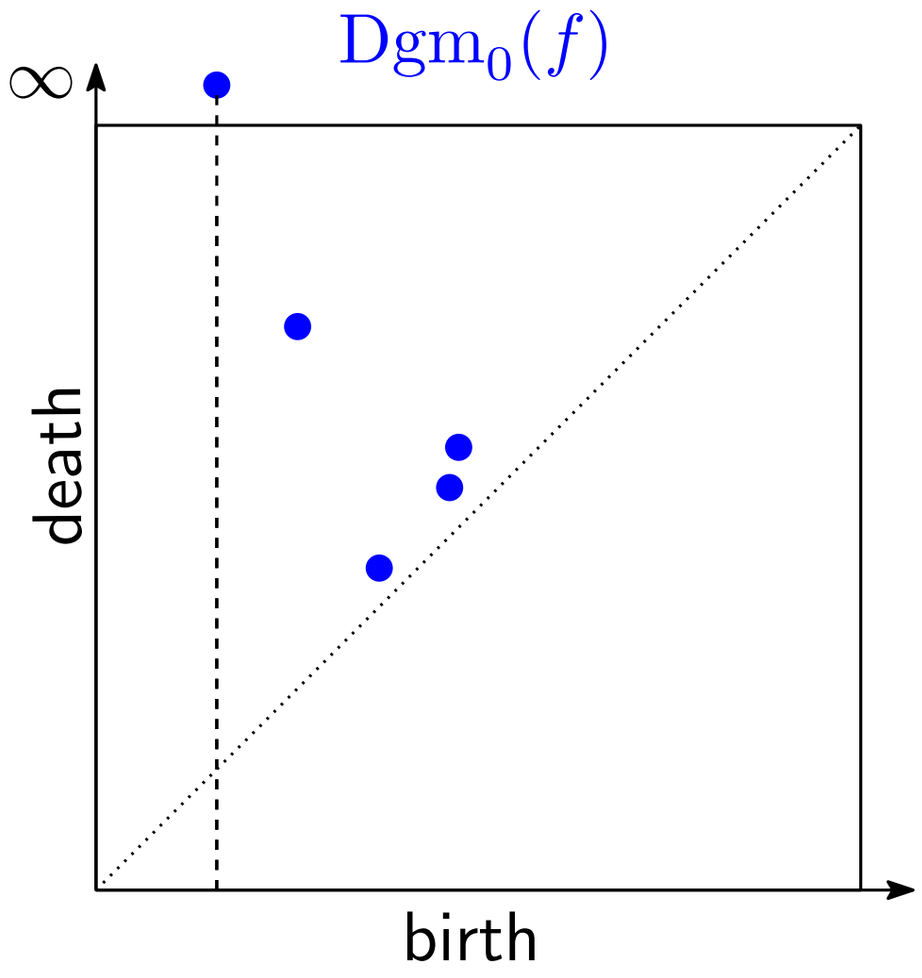
Evolution of homology as a birth-death pair.



Persistent Homology

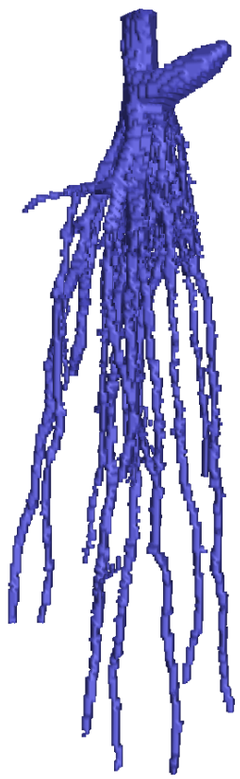
Evolution of homology as a birth-death pair.

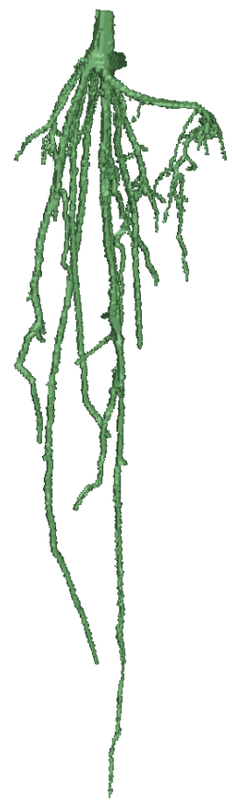


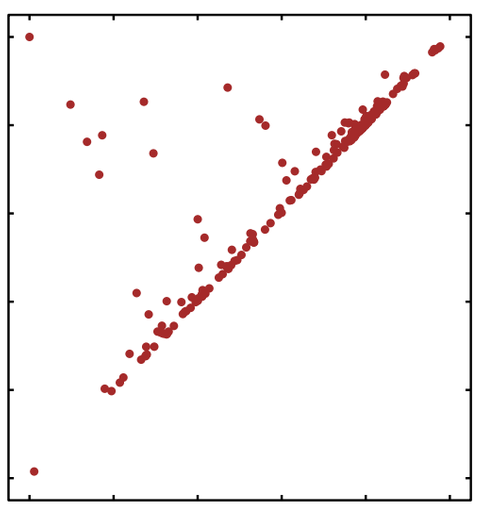


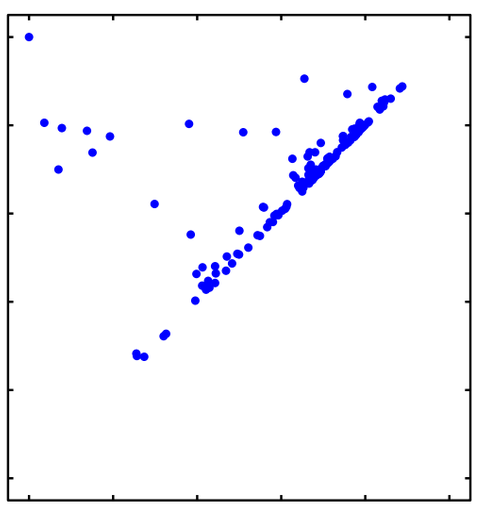
Persistent Homology

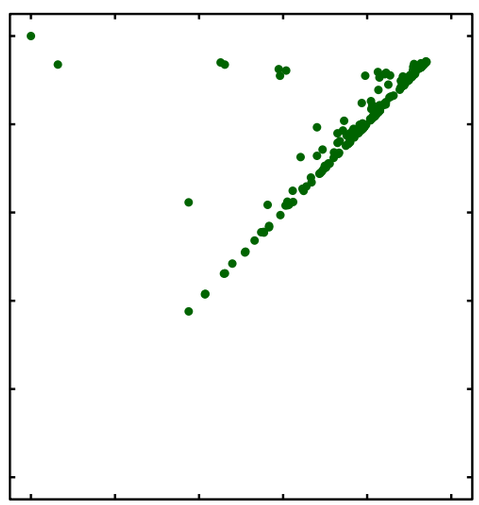
In practice…











Persistent Homology Transform

Let M be a shape of $\mathbf{R}^d$ that can be written as a finite simplicial complex K.

And let $\nu \in S^{d-1}$ be any unit vector over the unit sphere.

Persistent Homology Transform

Let M be a shape of $\mathbf{R}^d$ that can be written as a finite simplicial complex K.

And let $\nu \in S^{d-1}$ be any unit vector over the unit sphere.

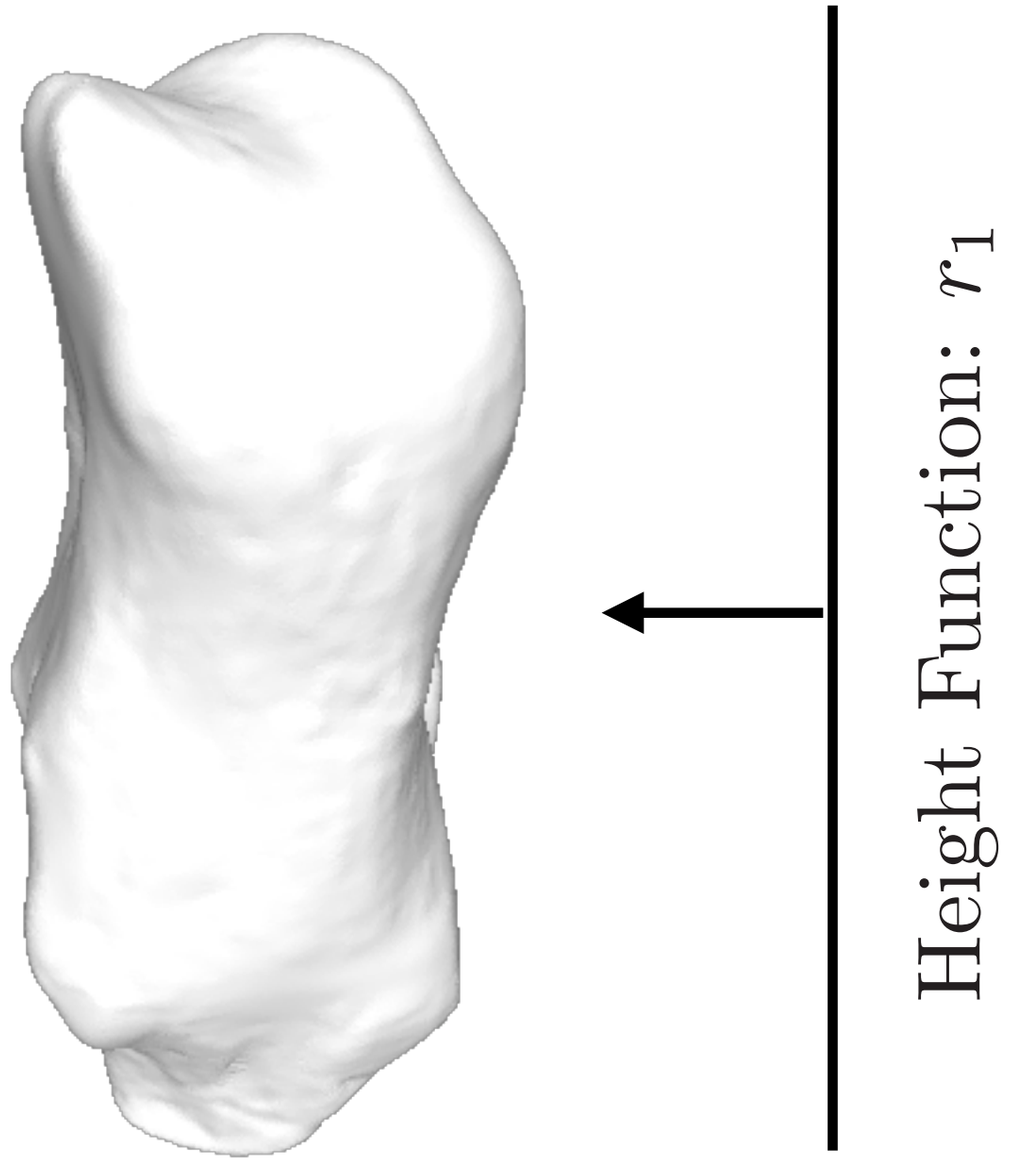
We define a filtration $K(\nu)$ of K parameterized by a height function r as

$$K(\nu)_r = \{x \in K : x \cdot \nu \leq r\}$$

The k-th dimensional persistence diagram $X_k(K, \nu)$ summarizes how the topology of the filtration $K(\nu)$ changes over the height parameter r

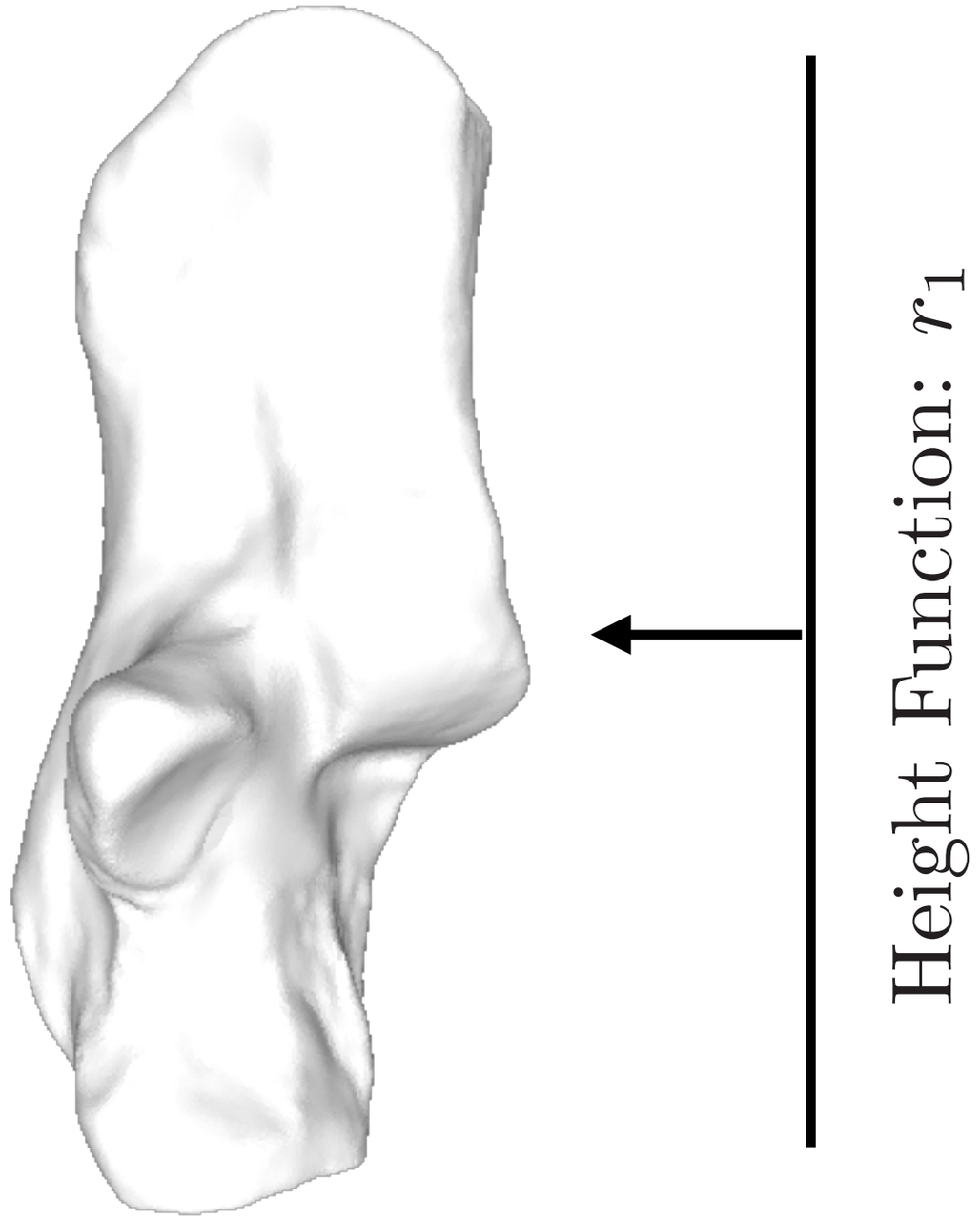
Persistent Homology Transform

For direction ν_1:



Persistent Homology Transform

For direction ν_2:



Persistent Homology Transform

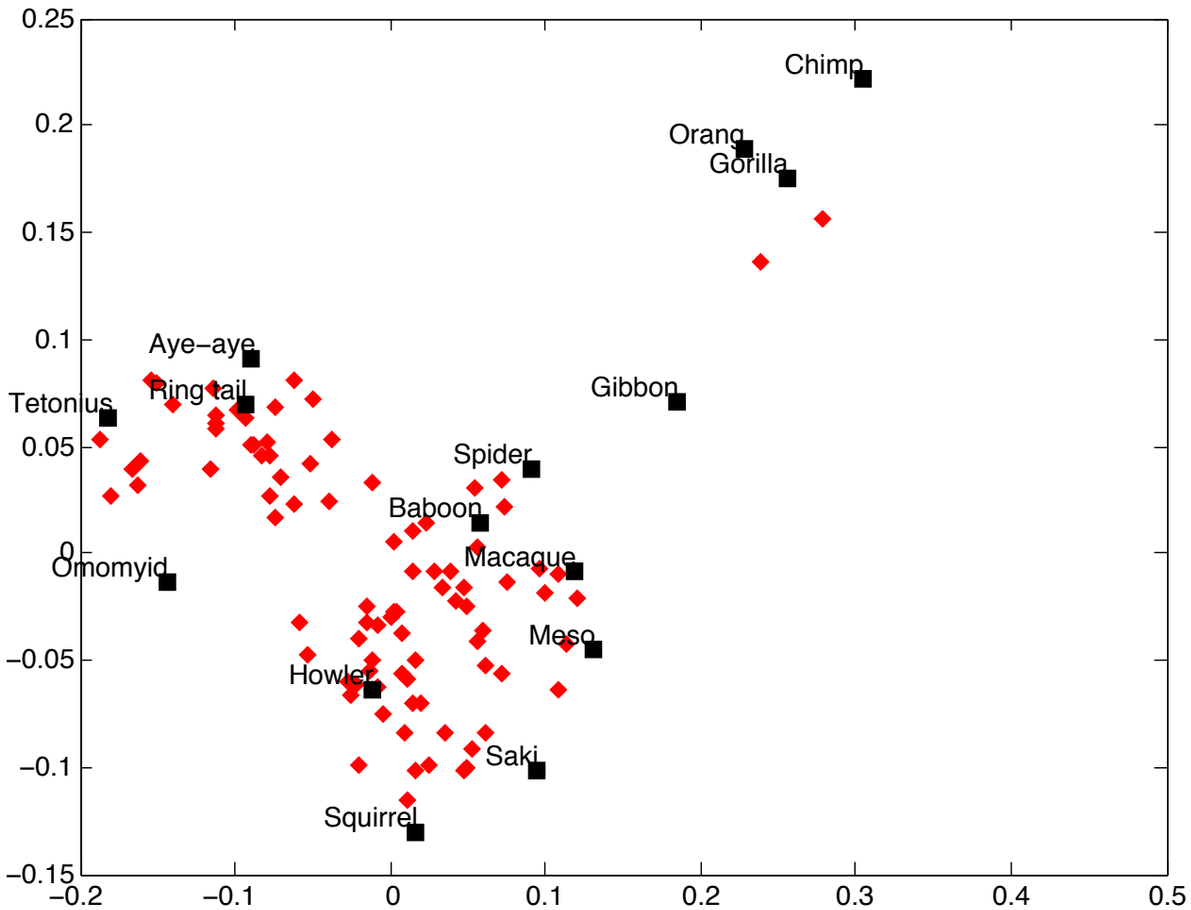
Definition: The *persistent homology transform (PHT)* of $K \subset \mathbf{R}^d$ is the function

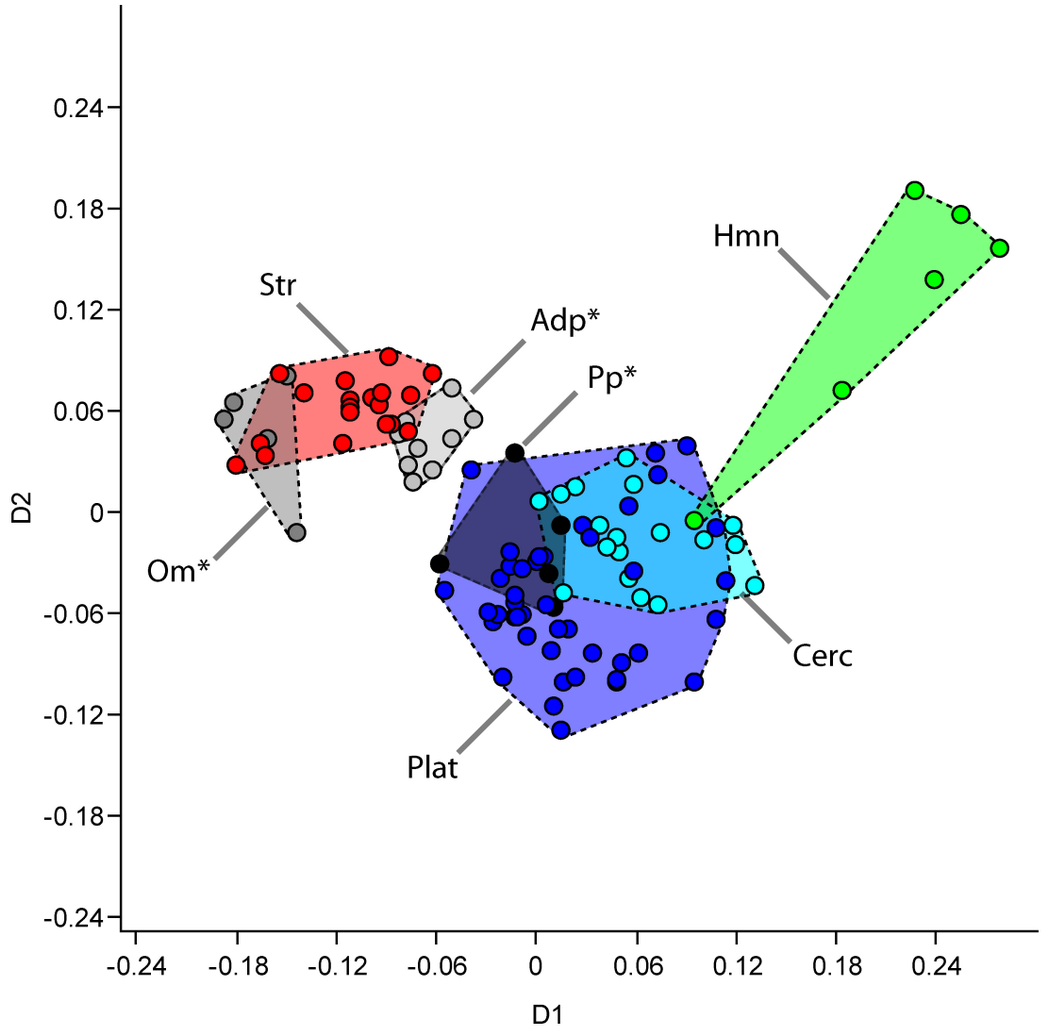
$$\begin{aligned} \mathrm{PHT}(K) : S^{d-1} &\to \mathcal{D}^d \\ \nu &\mapsto \big(X_0(K,\nu), X_1(K,\nu), \ldots, X_{d-1}(K,\nu)\big). \end{aligned}$$

- The PHT measures the change in homology by the height filtration over all directions on the unit sphere.
- It allows for the comparisons and similarity studies between shapes.
- The PHT preserves information, and a notion of statistical sufficiency was suggested for the PHT.

[Turner et al. (2014)]

Example Using the PHT





Ex: Phylogenetic groups of primate calcanei with 67 genera.

Pitfalls of the PHT

- Most widely used functional regression models use covariate that have an inner product structure defined in the Hilbert space.

- The geometry of the space of persistence diagrams is known to be a Alexandrov space with curvature bounded from below.

- The PHT does not admit a simple inner product structure as it is a collection of persistence diagrams.

- Therefore, it is challenging to use in all standard functional data analytic methods.

The Euler Characteristic

The Euler characteristic (EC) χ for a finite simplicial complex K^d for $d = 3$ is defined by:

$$\chi(K^3) = V - E + F,$$

where V, E, and F are the numbers of vertices, edges, and faces, respectively.

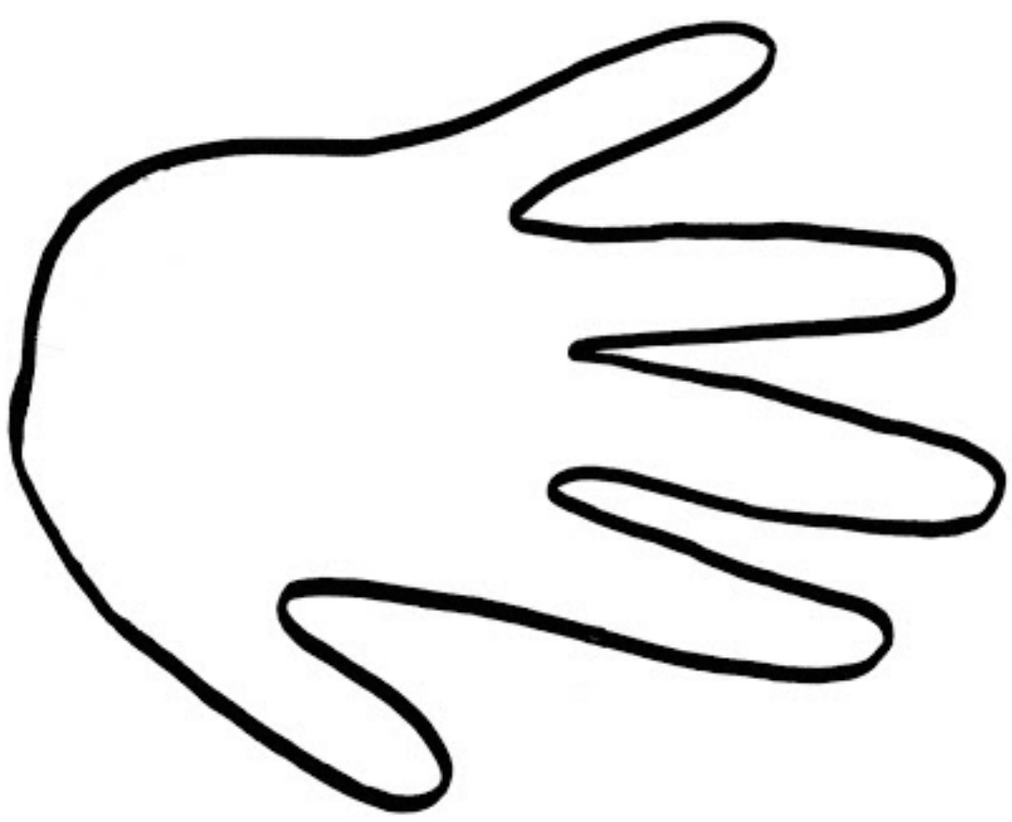
Euler Characteristic Curve

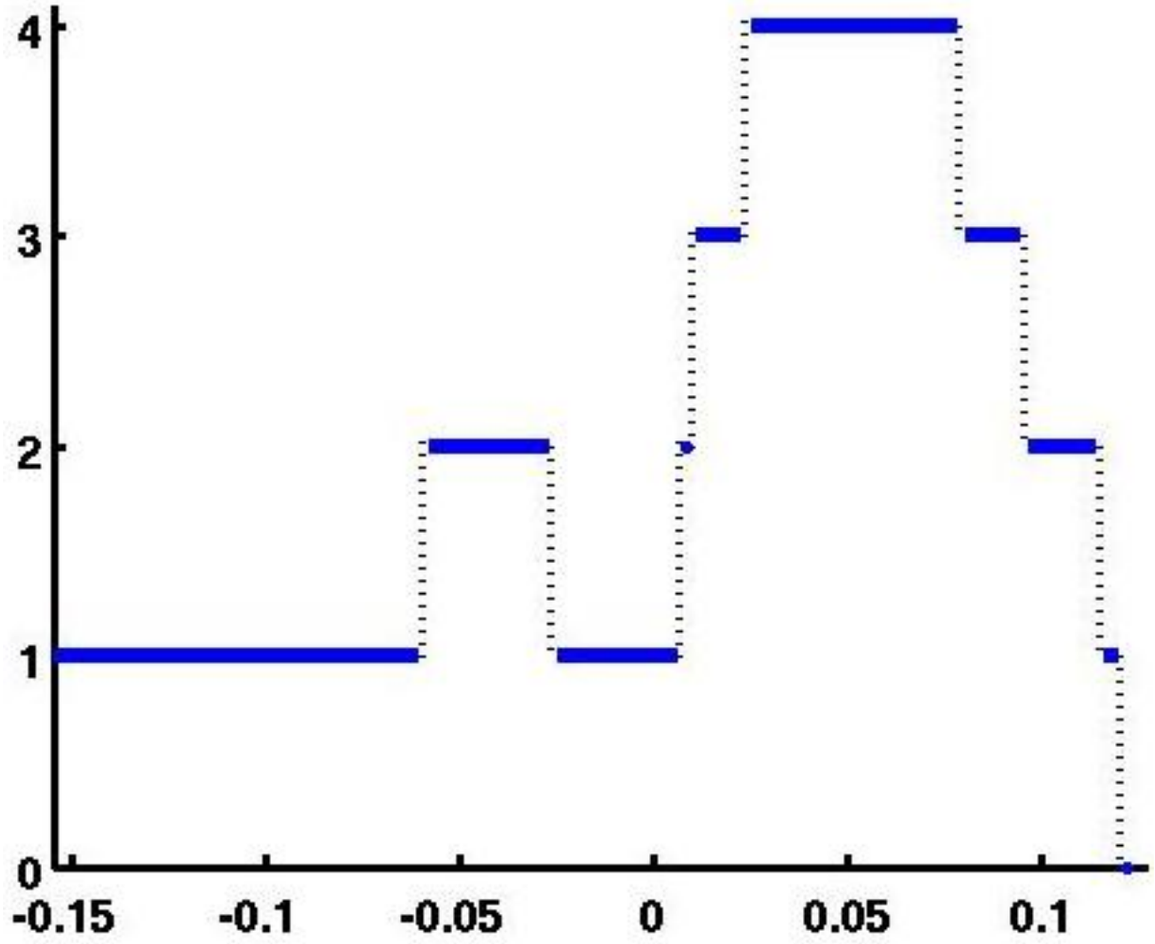
Definition: The *EC curve* is defined by:

$$\chi_\nu^K : [a_\nu, b_\nu] \to \mathbf{Z} \subset \mathbf{R}$$
$$x \mapsto \chi\left(K_\nu^x\right).$$

[Turner et al. (2014)]

Euler Characteristic Curve





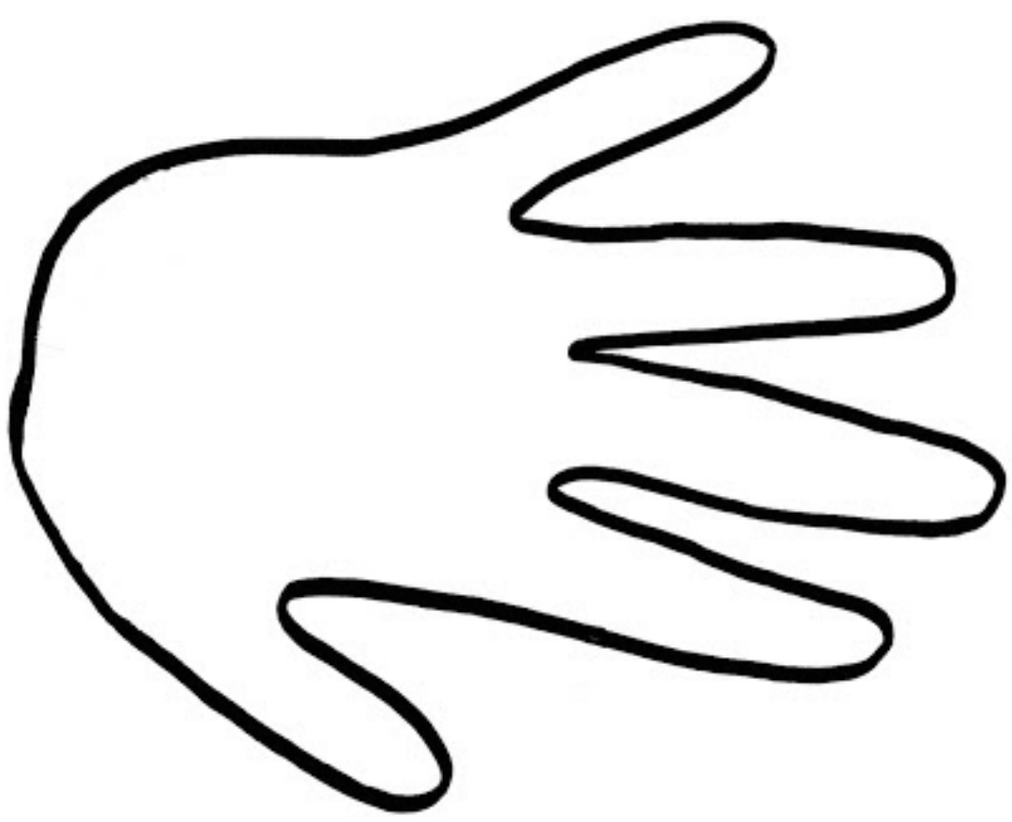
[Turner et al. (2014)]

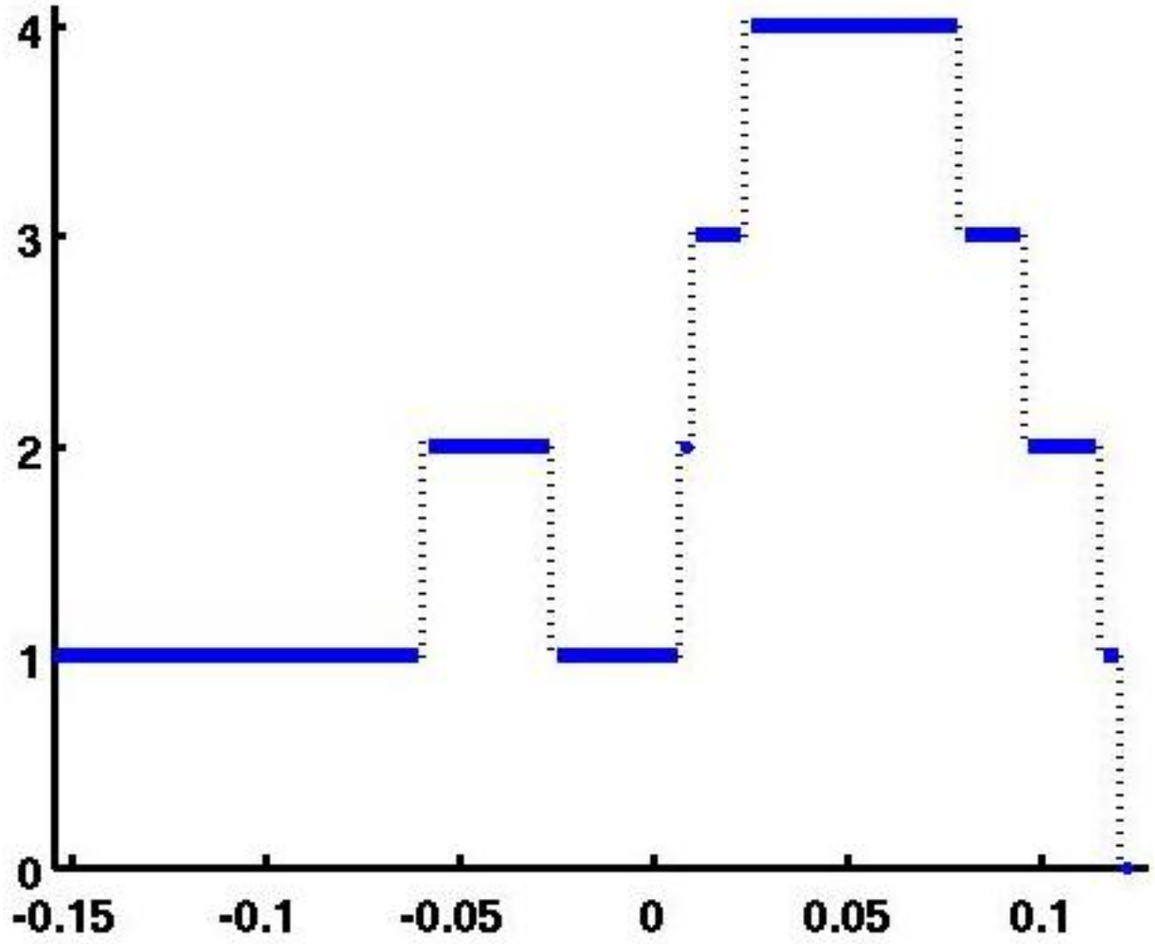
Smooth Euler Characteristic Curve

The *smooth Euler characteristic (SEC) curve* is computed by:

1. Taking the mean value of the EC curve $\bar{\chi}_\nu^K$ over $[a_\nu, b_\nu]$
2. Subtracting it from the value of the EC curve $\chi_\nu^K(x)$ at every $x \in [a_\nu, b_\nu]$

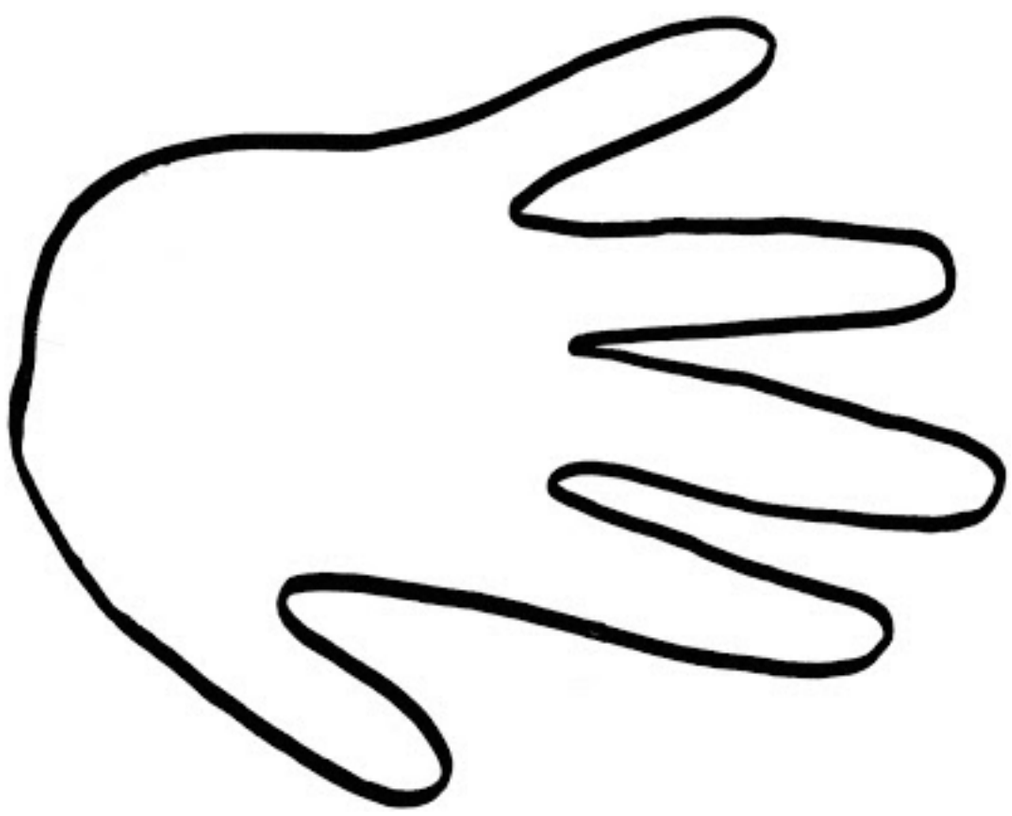
Euler Characteristic Curve

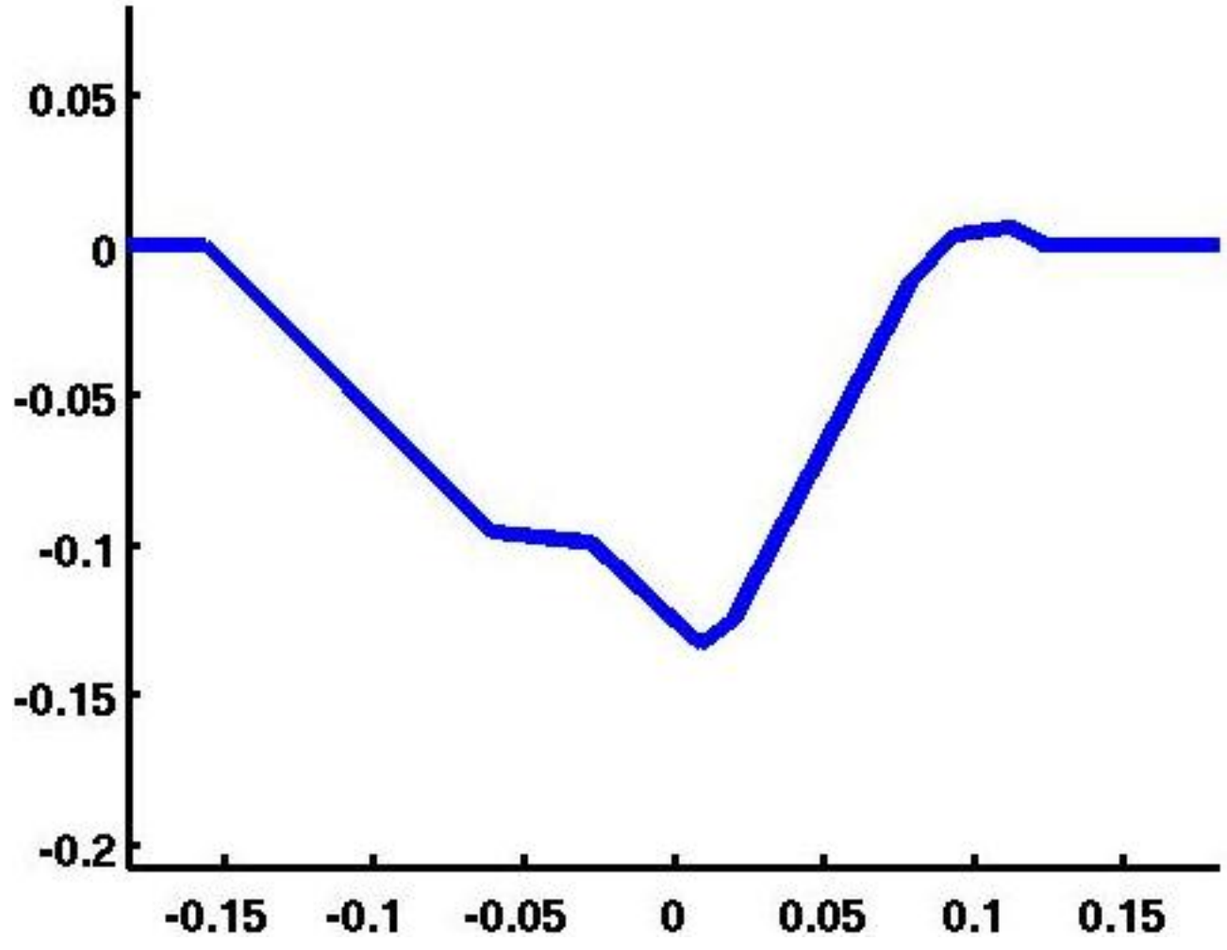




[Turner et al. (2014)]

Smooth Euler Characteristic Curve





Conventional Wisdom in Statistics

- SECT summaries are a collection of curves — this is a decidedly infinite-dimensional topological summary statistic.

- By construction, the SECT is a continuous, linear function that is an element of the Hilbert space $\mathbf{L}^2$ with a simple inner product structure.

- This means that their structure allows for quantitative comparisons using the full scope of functional and nonparametric regression methodology.

- This is the basis of functional data analysis (FDA).

Predicting Clinical Outcomes in Radiogenomics

- **Radiomics:** A newer subfield of genetics and genomics which focuses on the study of correlations between *imaging or network features* and *genetic variation*.

- Gliomas are a collection of tumors arising from glia or their precursors within the central nervous system.

- Of all gliomas, glioblastoma multiforme (GBM) is the most aggressive and most common in humans.

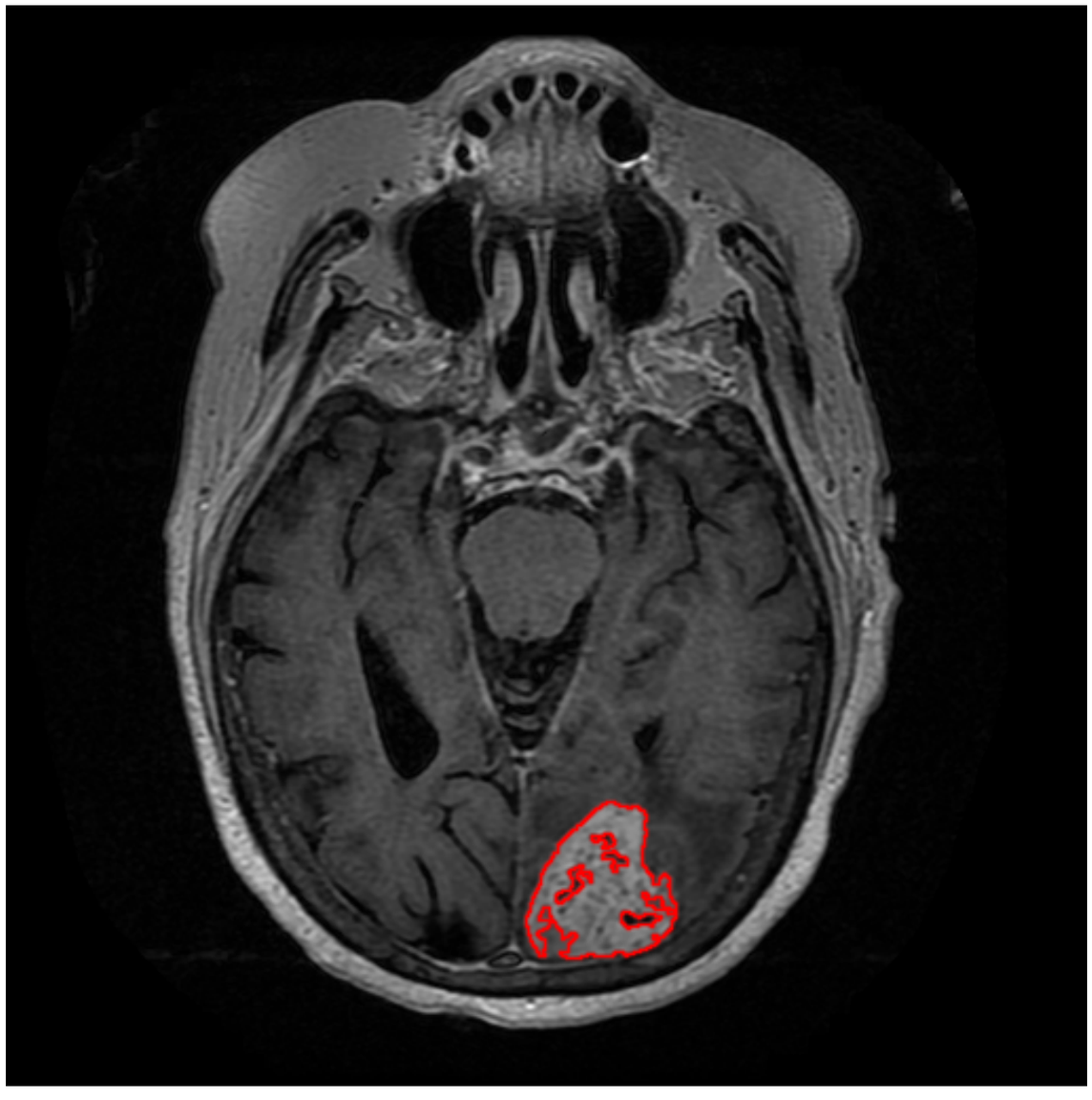
Predicting Clinical Outcomes in Radiogenomics

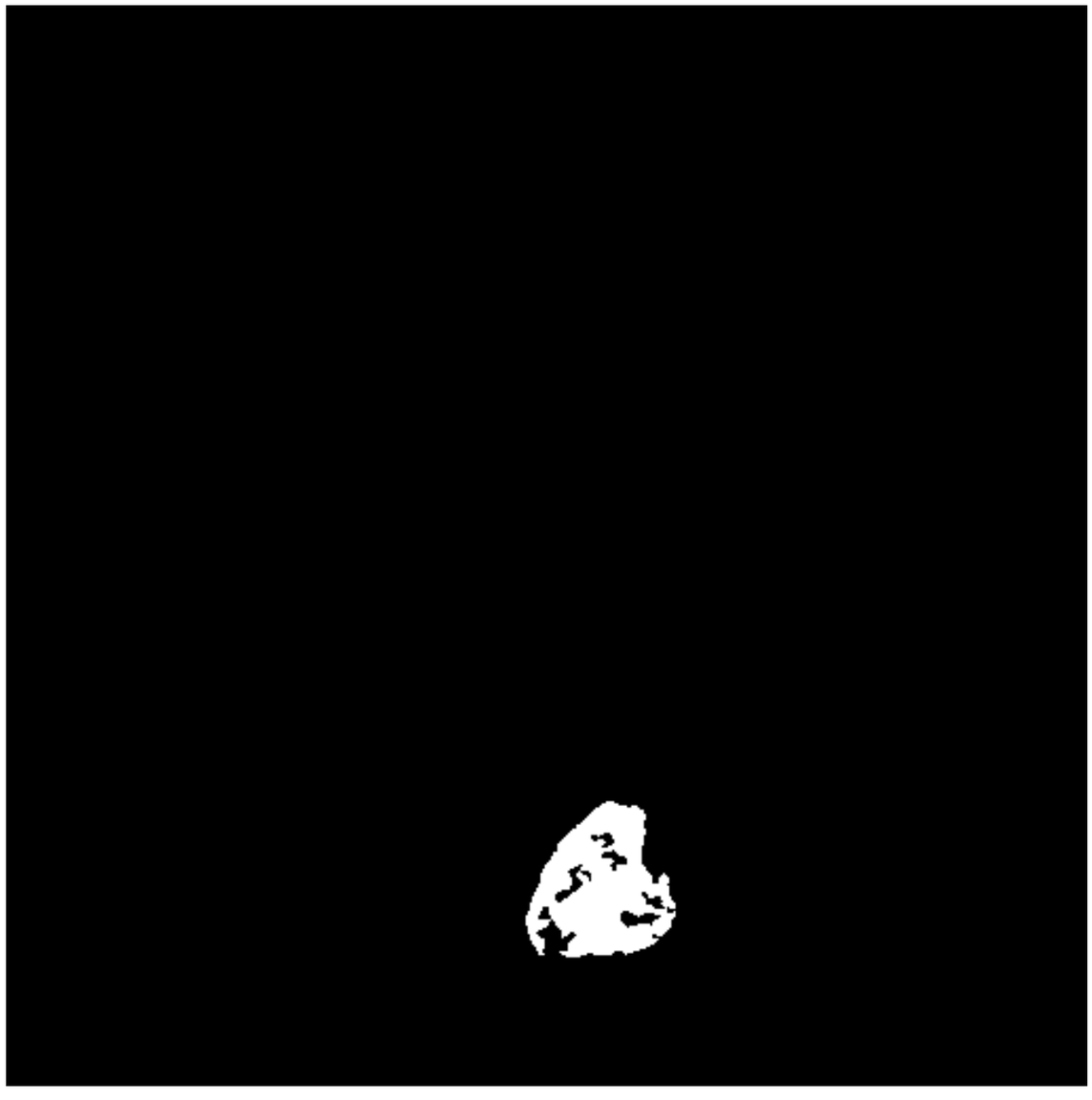
- Magnetic resonance images (MRIs) of primary GBM tumors were collected from ~40 patients archived by the The Cancer Imaging Archive (TCIA)

- These patients also had matched genomic and clinical data collected by The Cancer Genome Atlas (TCGA)

- **Goal:** We want to use the SECT to predict clinical outcomes:
  - Overall Survival (OS)
  - Disease Free Survival (DFS)

Application to Glioblastoma Multiforme





Regression with Functional Covariates

Assume that we have a finite response $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$.

Denote the SECT features as square integrable functions $F_\nu(t)$ on the real interval domain $\mathcal{T}$ where $t \in \mathcal{T}$.

Regression with Functional Covariates

Assume that we have a finite response $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$.

Denote the SECT features as square integrable functions $F_\nu(t)$ on the real interval domain $\mathcal{T}$ where $t \in \mathcal{T}$.

Given a real-valued measure $\mathrm{d}w$, a functional regression model takes on the form

$$\boldsymbol{y} \sim p(\boldsymbol{y} \,|\, \boldsymbol{\mu}), \quad g^{-1}(\boldsymbol{\mu}) = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \int_{\mathcal{T}} \sum_{\nu=1}^{m} f\left(F_\nu(t)\right) \mathrm{d}w(t) + \boldsymbol{\varepsilon}.$$

Here f is a smooth operator from $\mathbf{L}^2$ to $\mathbf{R}$ to be estimated over m directions.

Functional Linear Models

Classical parametric inferences assume that f is linear in the covariates:

$$\boldsymbol{\eta} = \sum_{\nu=1}^{m} \langle F_\nu(t), \boldsymbol{\beta}_\nu(t) \rangle,$$

Functional Linear Models

Classical parametric inferences assume that f is linear in the covariates:

$$\boldsymbol{\eta} = \sum_{\nu=1}^{m} \langle F_\nu(t), \boldsymbol{\beta}_\nu(t) \rangle,$$

where unlike traditional linear regression,

- $\boldsymbol{\beta}_\nu(t)$ is an unknown smooth parameter function that is also square integrable on the domain $\mathcal{T}$;
- $\langle \cdot, \cdot \rangle$ denotes an inner product in the Hilbert space $\mathbf{L}^2$.

Limitations for Functional Linear Models

- In many applications, it is considered too restrictive to only assume linear effects on the functional covariates.

- For example, it is reasonable to assume that interactions between modes of brain activity extend well beyond additivity.

- Nonlinear kernel regression models serve as a natural alternative choice, as they often display greater predictive accuracy than linear models.

Functional Kernel Models

Assume the target function f to be an element of the reproducing kernel Hilbert space (RKHS) $\mathbf{H}$ equipped with an inner product, with

Functional Kernel Models

Assume the target function f to be an element of the reproducing kernel Hilbert space (RKHS) $\mathbf{H}$ equipped with an inner product, with

$$\mathbf{H} = \left\{ f \mid f\left(F_\nu(t)\right) = \sum_{j=1}^{\infty} c_j \psi_j\left(F_\nu(t)\right) \text{ and } \|f\|_{\mathbf{H}}^2 = \sum_{j=1}^{\infty} c_j^2/\lambda_j < \infty \right\},$$

and estimator function

$$\widehat{f}\left(F_\nu(t)\right) = \sum_{i=1}^{n} \alpha_i \, k\left(F_\nu(t), F_{\nu,i}(t)\right).$$

Functional Kernel Models

We can posit a generalized functional kernel regression model

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{K})$$

where $\boldsymbol{K}$ is a symmetric and positive-definite covariance (kernel) matrix with elements $\boldsymbol{K}_{ij} = k\left(F_{\nu,i}(t), F_{\nu,j}(t)\right)$.

Functional Kernel Models

We can posit a generalized functional kernel regression model

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{K})$$

where $\boldsymbol{K}$ is a symmetric and positive-definite covariance (kernel) matrix with elements $\boldsymbol{K}_{ij} = k\left(F_{\nu,i}(t), F_{\nu,j}(t)\right)$.

Here we may consider for example:

1. $k(\boldsymbol{s}, \boldsymbol{v}) = \boldsymbol{s}^\intercal \boldsymbol{v}/p + h$;
2. $k(\boldsymbol{s}, \boldsymbol{v}) = \exp\{-h\|\boldsymbol{s} - \boldsymbol{v}\|^2\}$;
3. $k(\boldsymbol{s}, \boldsymbol{v}) = \log(\|\boldsymbol{s} - \boldsymbol{v}\|^h + 1)$.

Bayesian Functional Kernel Regression

When modeling continuous outcomes

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}),$$

Bayesian Functional Kernel Regression

When modeling continuous outcomes

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}),$$

where each parameter is assumed to come from the following prior distributions

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{K}), \quad \sigma^{-2}, \tau^{-2} \sim \mathcal{G}(\kappa_1, \kappa_2).$$

We will exclusively consider the posterior distribution that arises in the limits $\kappa_1 \to 0$ and $\kappa_2 \to 0$.

Posterior Inference and Sampling

Markov chain Monte Carlo (MCMC) via a Gibbs sampler for the regression model:

(1) $\boldsymbol{\eta} \,|\, \boldsymbol{y}, \omega, \sigma^2, \tau^2 \sim \mathcal{N}(\boldsymbol{m}^*, \boldsymbol{V}^*)$ where $\boldsymbol{m}^* = \tau^{-2}\boldsymbol{V}^*\boldsymbol{y}$ and $\boldsymbol{V}^* = \tau^2\sigma^2(\tau^2\boldsymbol{K} + \sigma^2\boldsymbol{I}_n)^{-1}$;

(2) $\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{\eta}, \omega, \tau^2 \sim \mathcal{G}(a^*, b^*)$ where $a^* = n/2$ and $b^* = \boldsymbol{\eta}^\intercal \boldsymbol{K}^{-1}\boldsymbol{\eta}/2$;

(3) $\tau^2 \,|\, \boldsymbol{y}, \boldsymbol{\eta}, \omega, \sigma^2 \sim \mathcal{G}(a^*, b^*)$ where $a^* = n/2$ and $b^* = \boldsymbol{y}^\intercal\boldsymbol{y}/2$.

Posterior Predictive Distribution

To predict outcomes for individuals in a test set T, based on what we observe in the sample set S, let

$$\{\boldsymbol{y}_T^{(b)} = \boldsymbol{\eta}_T^{(b)}\}_{b=1}^{B}$$

[Speed and Balding (2014)]

Posterior Predictive Distribution

To predict outcomes for individuals in a test set T, based on what we observe in the sample set S, let

$$\{\boldsymbol{y}_T^{(b)} = \boldsymbol{\eta}_T^{(b)}\}_{b=1}^{B}$$

where, for B MCMC samples, we define

$$\boldsymbol{\eta}_T^{(b)} = \boldsymbol{K}_{TS}\boldsymbol{K}_{SS}^{-1}\boldsymbol{\eta}_S^{(b)}, \quad b = 1, \ldots, B$$

with $\boldsymbol{K}_{TS}$ and $\boldsymbol{K}_{SS}$ being submatrices that are found by first computing $\boldsymbol{K}^* = [\boldsymbol{K}_{SS}; \boldsymbol{K}_{ST}; \boldsymbol{K}_{TS}; \boldsymbol{K}_{TT}]$.

[Speed and Balding (2014)]

Predicting Clinical Outcomes in Radiogenomics

- Compare the SECT with three key types of glioblastoma tumor characteristics:
  - mRNA Gene Expression Measurements
  - Tumor Morphometry
  - Tumor Volume and Geometrics

- We attempt to predict two clinical outcomes:
  - **Disease Free Survival (DFS)**
  - **Overall Survival (OS)**

- Perform 80-20 (in/out of sample) splits; 100 times

- **Predictive Measure:** Root Mean Square Error of Prediction (RMSEP)

Prediction Results

	Disease Free Survival		*Overall Survival*	
Data Type	**RMSEP**	**Pr[Optimal]**	**RMSEP**	**Pr[Optimal]**
Gene Expression	0.944 (0.035)	0.20	0.981 (0.030)	0.27
Morphometrics	0.942 (0.035)	0.07	0.965 (0.029)	0.15
Volume	0.939 (0.035)	0.06	0.964 (0.029)	0.16
SECT	**0.803 (0.035)**	**0.69**	**0.958 (0.028)**	**0.42**

Average RMSPE across both clinical outcomes. The number in parenthesis is the standard error due to random sampling

Future Directions and Ongoing Work

- **Proving Sufficiency for Summary Statistics of 3D Shapes:**
  - An important open problem is proving that the transformations defined by the SECT and PHT are capturing all *sufficient information* needed to fully characterize a given shape.

- **Improving Phenotypic prediction with Manifold Approximation and Multiple Kernel Learning:**
  - Begin to learn about the manifold underlying the 3D shapes in order to extract information about their intrinsic geometries

- **Gene Set Enrichment Analysis Using Sufficient Shape Statistics:**
  - It is of natural interest to probe whether variation in shape is correlated with molecular signaling pathway dysregulation.
  - Build a framework for analyzing the heterogeneity of fitness trajectories in cells exposed to therapy (i.e. stress).

Relevant References

The Persistent Homology Transform (PHT):

- Turner, K., S. Mukherjee, and D. M. Boyer (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*. **3**(4): 310–344.

The Smooth Euler Characteristic Transform (SECT):

- L. Crawford, A. Monod, A.X. Chen, S. Mukherjee, and R. Rabadán (2017). Functional data analysis using a topological summary statistic: the smooth Euler characteristic transform. *arXiv*. 1611.06818.

Tropical Sufficient Statistics for Persistent Homology (Tropix):

- A. Monod, S. Kališnik Verovšek, J.Á. Patiño-Galindo, and L. Crawford (2017). Tropical sufficient statistics for persistent homology. *arXiv*. 1709.02647.

Available Source Code

Crawford Lab Website:

- http://www.lcrawlab.com

The Smooth Euler Characteristic Transform (SECT):

- https://github.com/RabadanLab/SECT

Bayesian Approximate Kernel Regression (BAKR):

- https://github.com/lorinanthony/BAKR

Acknowledgements

- **Collaborators:**
  - **Andrew Chen (Columbia University)**
  - **Anthea Monod, Ph.D. (Columbia University)**
  - **Sayan Mukherjee, Ph.D. (Duke University)**
  - **Raúl Rabadán, Ph.D. (Columbia University)**
- **Contributors:**
  - **Nicolas Garcia Trillos, Ph.D. (Brown University)**
  - **ECOG-ACRIN Cancer Research Group**
- **Data Availability:**
  - **The Cancer Imaging Archive (TCIA)**
  - **The Cancer Genome Atlas (TCGA)**

BROWN
School of Public Health

Duke Statistical Science

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Center for Topology of Cancer Evolution and Heterogeneity

A member of the National Cancer Institute's Physical Sciences in Oncology Network

THANK YOU!