Functional Data Analysis Using a Topological Summary Statistic: The Smooth Euler Characteristic Transform

Anthea Monod

Department of Systems Biology
Columbia University

am4691@cumc.columbia.edu

Joint work with Lorin Crawford (Brown University), Andrew X. Chen, Sayan Mukherjee (Duke University), Raúl Rabadán

14 February 2018

Key Concepts and Terms

Topological Data Analysis (TDA):

- Adapts concepts from algebraic topology to point clouds to rigorously quantify the "shape" and "size" of data

Functional Data Analysis (FDA):

- Analyzes data that provides information about curves and surfaces, or more generally, variables over a given continuum

Topology & Data

Data: Often very high dimensional, may come with a notion of distance, proximity or similarity

- Coordinate systems may not be practical
- Metric may not be well-defined
- Require qualitative information $\Longrightarrow$ *Summaries* are more valuable

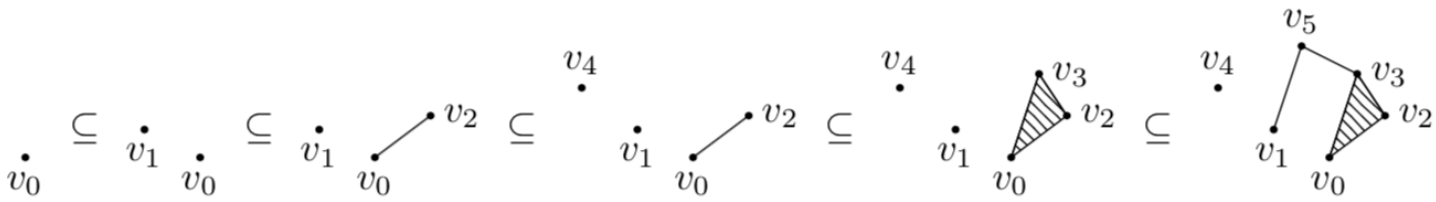
Topology: Provides qualitative analysis and geometric pattern recognition

- Coordinate-free
- Insensitive to metric
- Studies connectivity information

$\Longrightarrow$ Study point cloud data (i.e. finite set of points with a distance measure) using geometric and topological methods as a means to reduce dimension and summarize pertinent information

Persistent Homology

Homology groups were developed in classical topology to "measure" the shape of spaces by abstractly counting the occurrences of patterns (connected components, loops, voids, etc.)

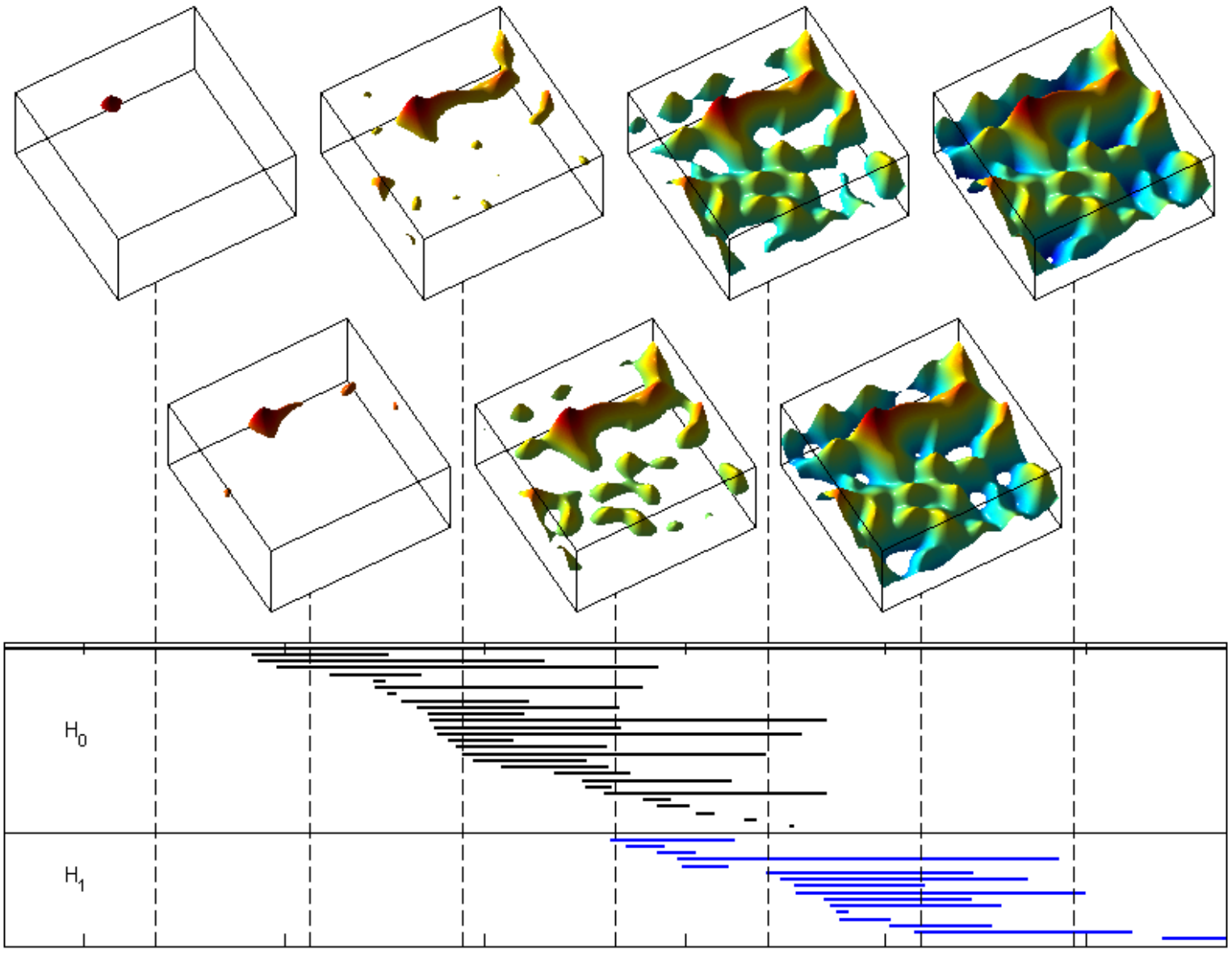


Persistent homology keeps tracks of the evolution of homology groups with respect to a *filtration*

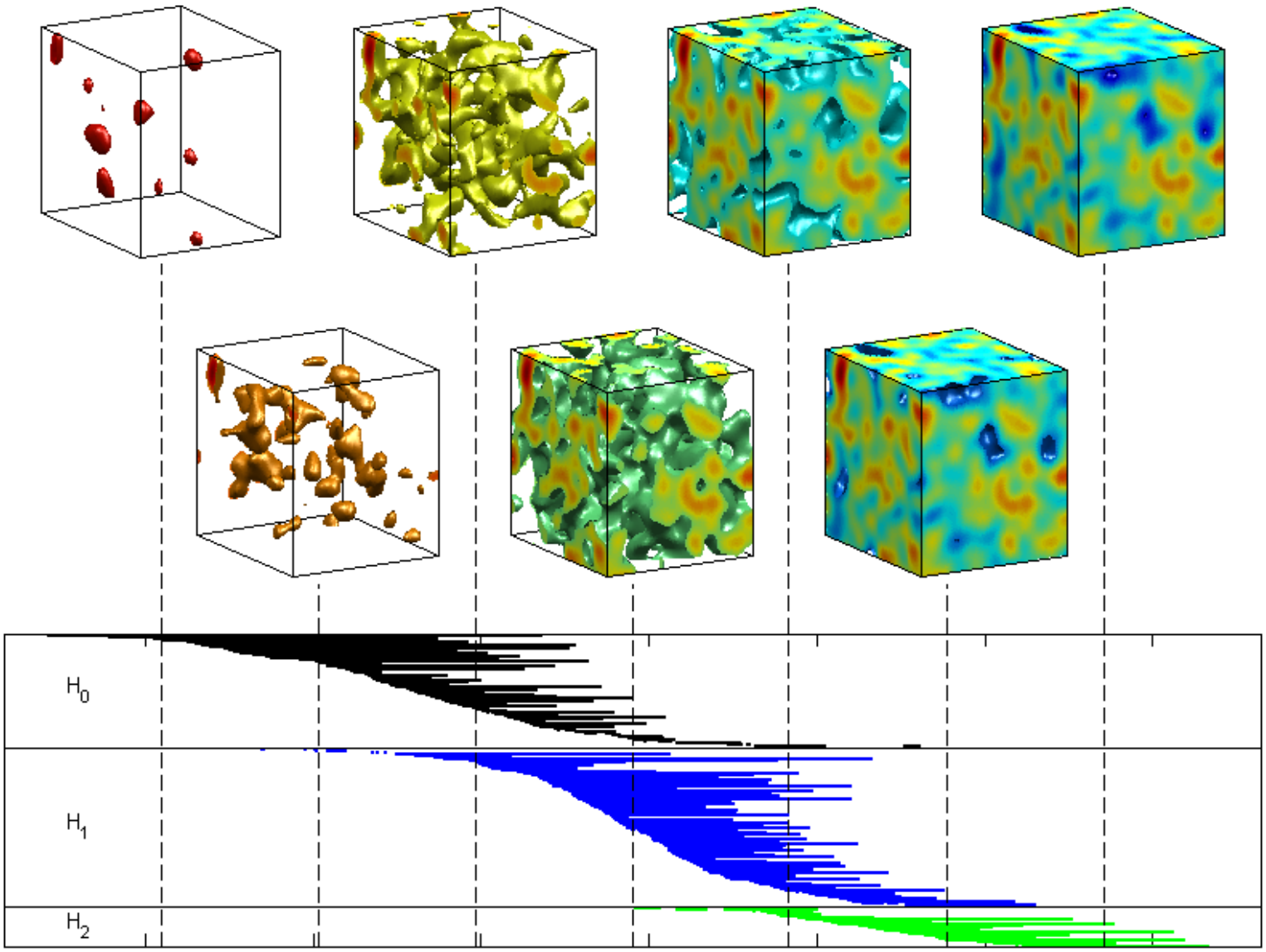
- It is useful in data analysis because it adapts homology to finite metric spaces = point clouds
- It measures the "shape" and "size" of data

Persistent homology is encoded in a *barcode* or *persistence diagram*

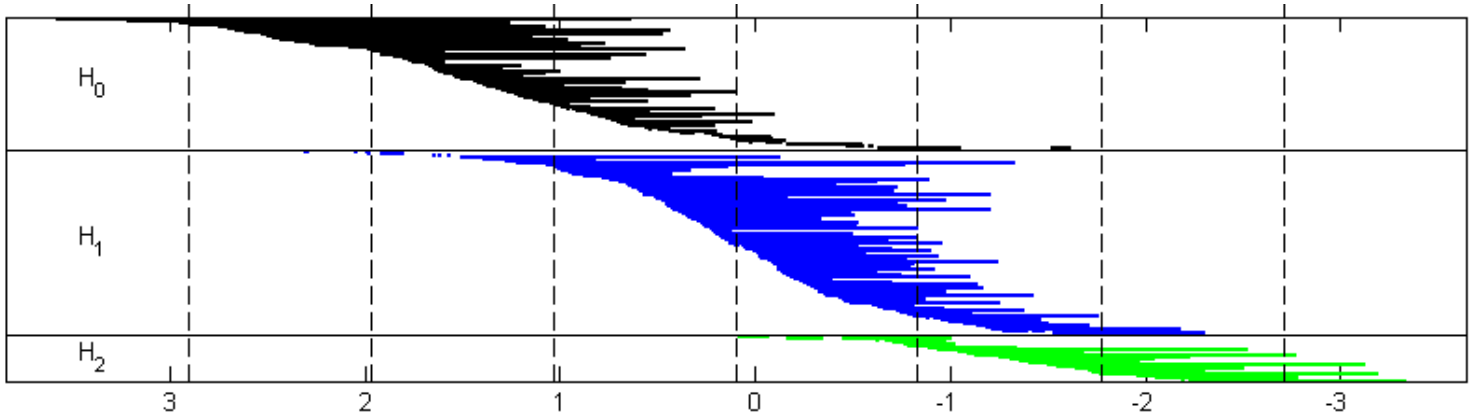
Example: Persistence in 2 Dimensions

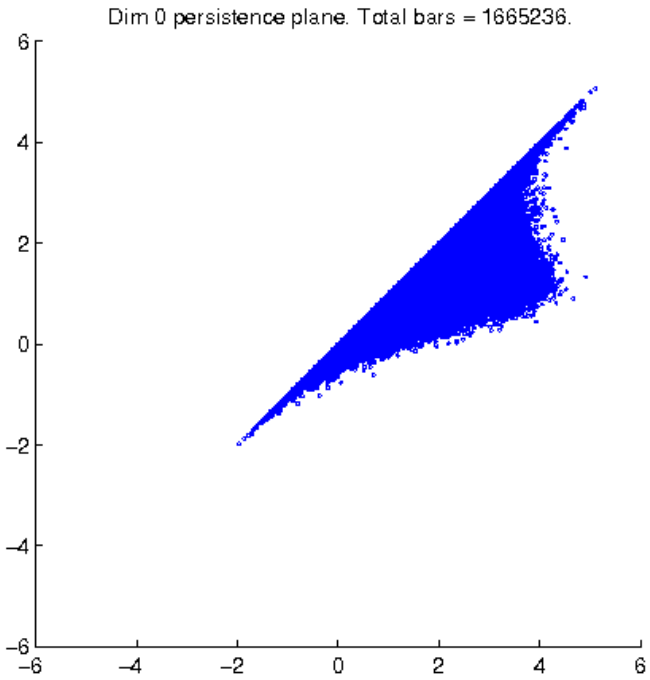


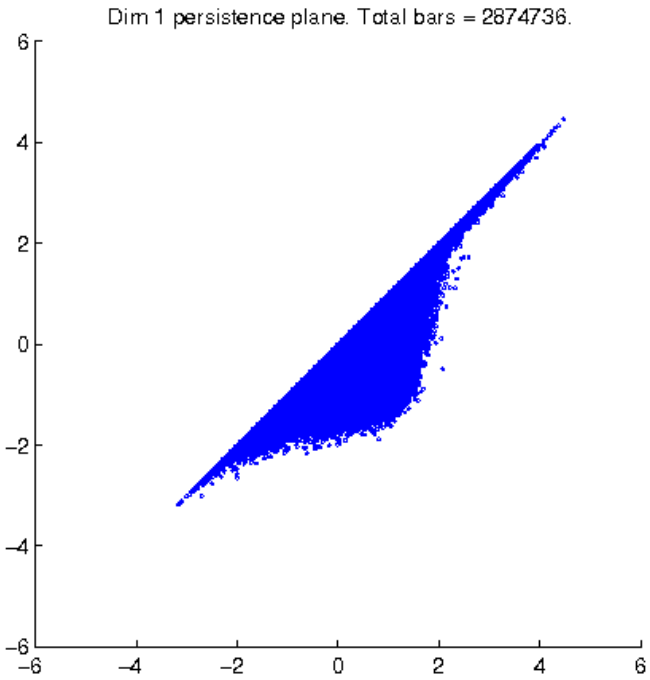
Example: Persistence in 3 Dimensions

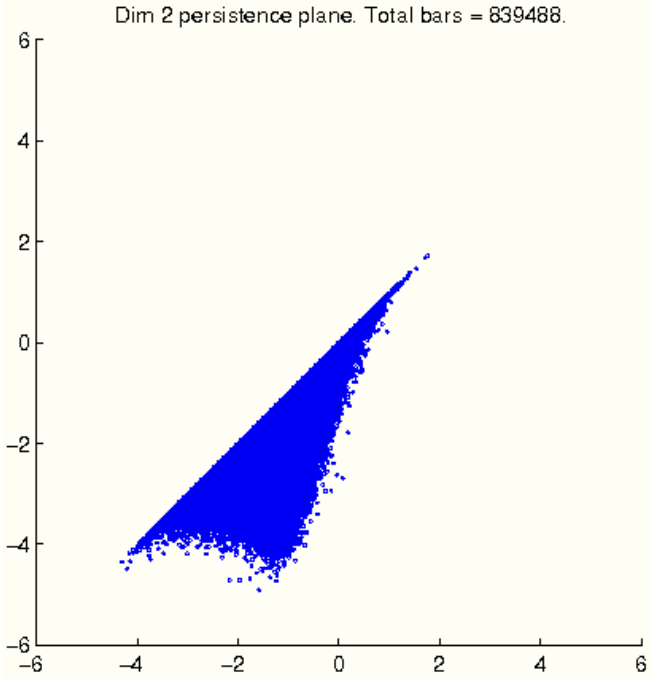


Barcodes & Persistence Diagrams

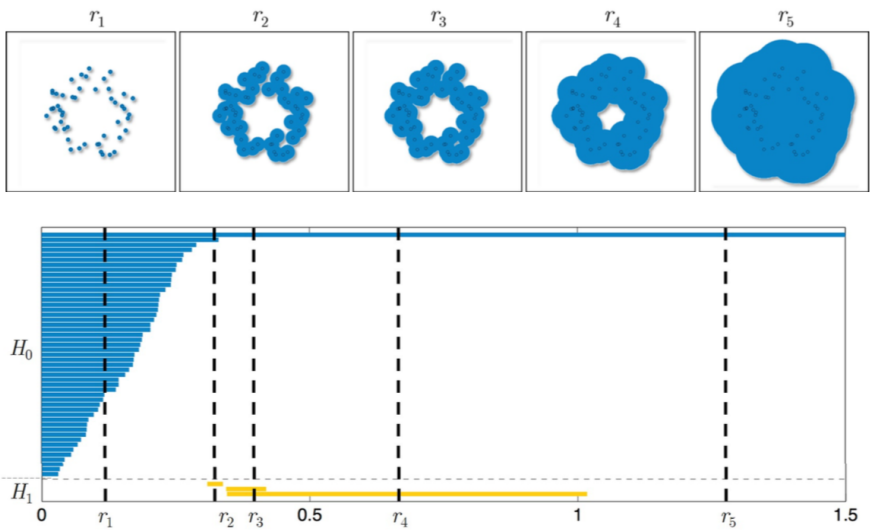


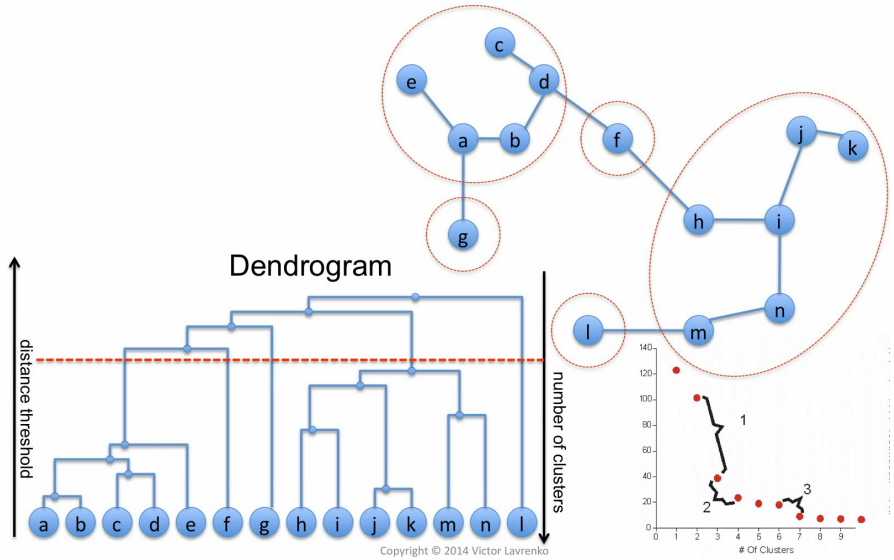






Persistent Homology vs. Clustering





H_0 Persistence $\leftrightarrow$ Single-Linkage Clustering

Filtration Parameter $\leftrightarrow$ Distance Threshold

Barcode $\leftrightarrow$ Dendrogram

# of Clusters $\leftrightarrow$ # of H_0 Bars

Computing Homology & Persistent Homology

- To study complex objects (e.g. shapes, surfaces, spaces), it is often convenient to *discretize*
- Represent the object as a union of simple "building blocks" cleanly glued together; e.g. a *simplicial complex* is a collection of *simplices* (satisfying certain properties)
- There are efficient algorithms to compute homology when the object is a simplicial complex
- The Nerve Theorem guarantees that the homologies of the original object and its discretized form will be the same

The Space of Barcodes & Persistence Diagrams

Important Properties:

- *Metric space*
  For regularity considerations, add the diagonal $\Delta = \{(x, y) \in \mathbf{R}^2 : x = y\}$, where each point on the diagonal has infinite multiplicity
  - Wasserstein p-distance, $p \geq 1$:

$$\mathrm{d}_p^W(D_1, D_2) = \left(\inf_{\gamma} \sum_{x \in D_1} \|x - \gamma(x)\|_\infty^p \right)^{\frac{1}{p}}$$

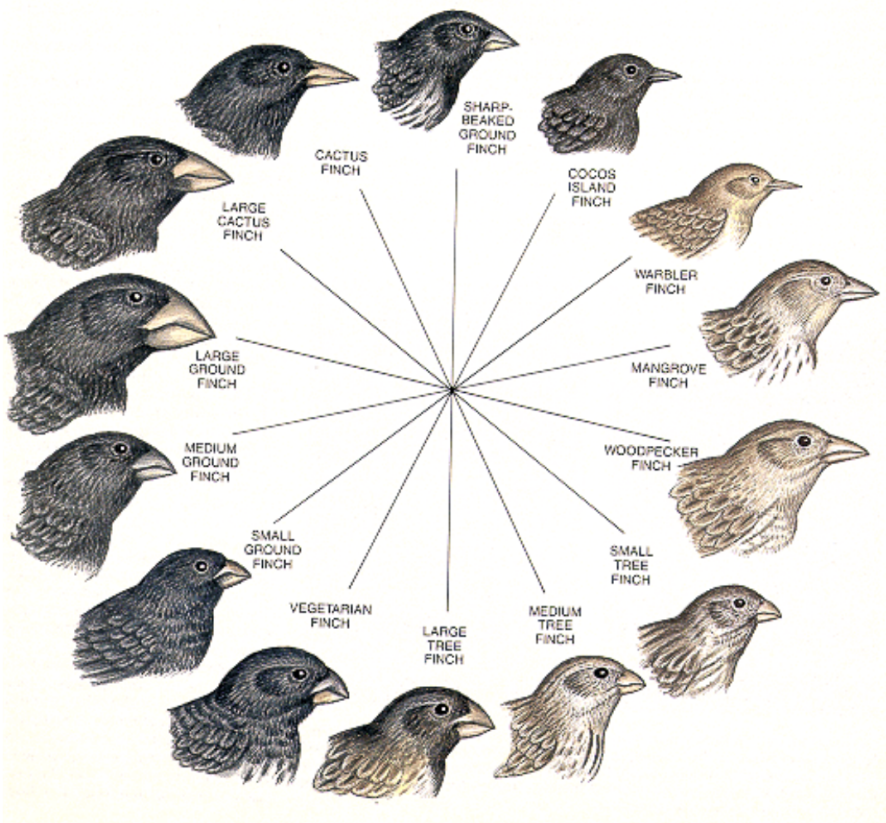
  - Bottleneck distance

$$\mathrm{d}^B(D_1, D_2) = \inf_{\gamma} \sup_{x} \|x - \gamma(x)\|_\infty$$

- *Completeness*, *Separability*
- *Existence of compact subspaces*
- *Alexandrov space with curvature bounded from below*

Variation in Shapes and Surfaces in Biology

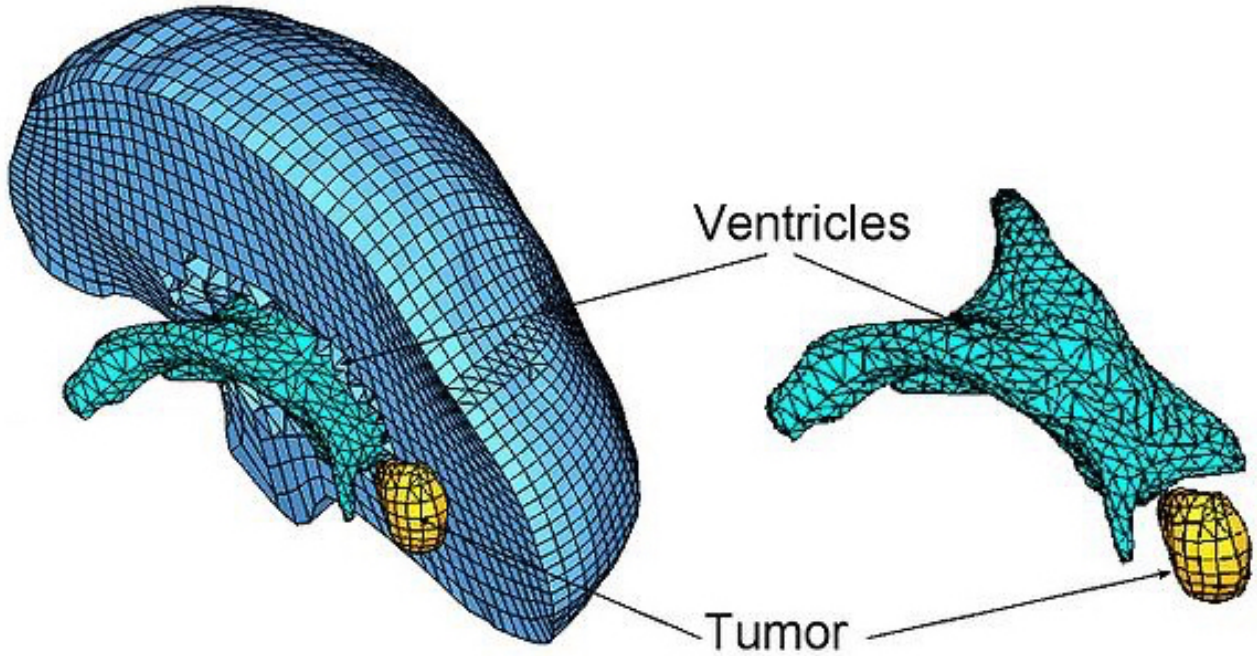
Phylogeny of Darwin's Finch Beaks (Gould, 1977)



Fossil Classification (Boyer et al., 2011)

Representing Shapes

Improved imaging technologies represent 3D shapes to be represented as meshes — collections of vertices, faces, and edges



These make methods from TDA directly applicable (Turner et al., 2014)!

Motivation

- Methods for geometric morphometrics are known to suffer from structural errors when comparing shapes that are highly dissimilar
- Comparing shapes requires a metric, which is not always straightforward to specify
- Turner et al. (2014) developed a statistical summary for shapes known as the *persistent homology transform* (PHT):
  $\Longrightarrow$ The PHT summarizes the shape information robustly for highly dissimilar and non-isomorphic shapes

But more is needed to fully integrate TDA measures into FDA methods...

Main Shape Statistic Objectives

- Represent shapes or images quantitatively to use in a wide range of FDA methods (e.g. generalized functional linear models, GFLMs)
- Desired Properties:
  - Injective mapping = Summary statistics
  - Distances are computable
  - Ability to define probabilistic models in the transformed space
- Topological Summaries:
  - Persistent Homology Transform (PHT)
  - Smooth Euler Characteristic Transform (SECT)

Persistent Homology Transform

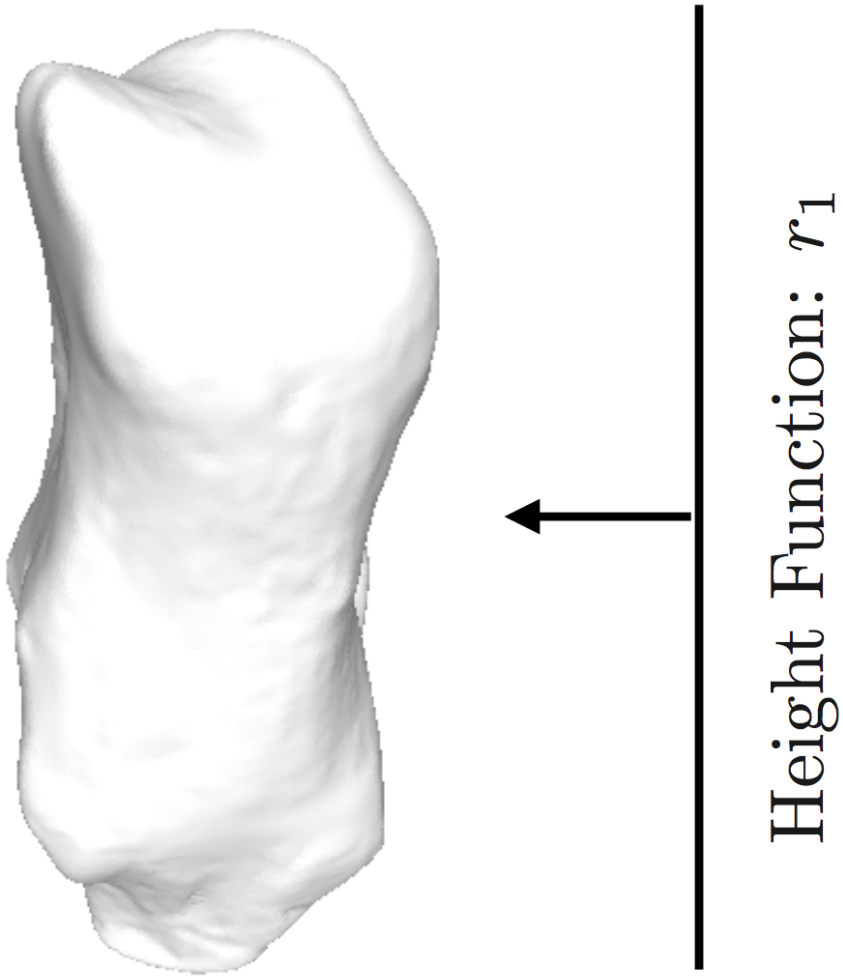
- Let M be a shape in $\mathbf{R}^d$ that can be written as a finite simplicial complex K
- Let $\nu \in S^{d-1}$ be any unit vector over the unit sphere
- Define a filtration $K(\nu)$ of K parameterized by a height function r:

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

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

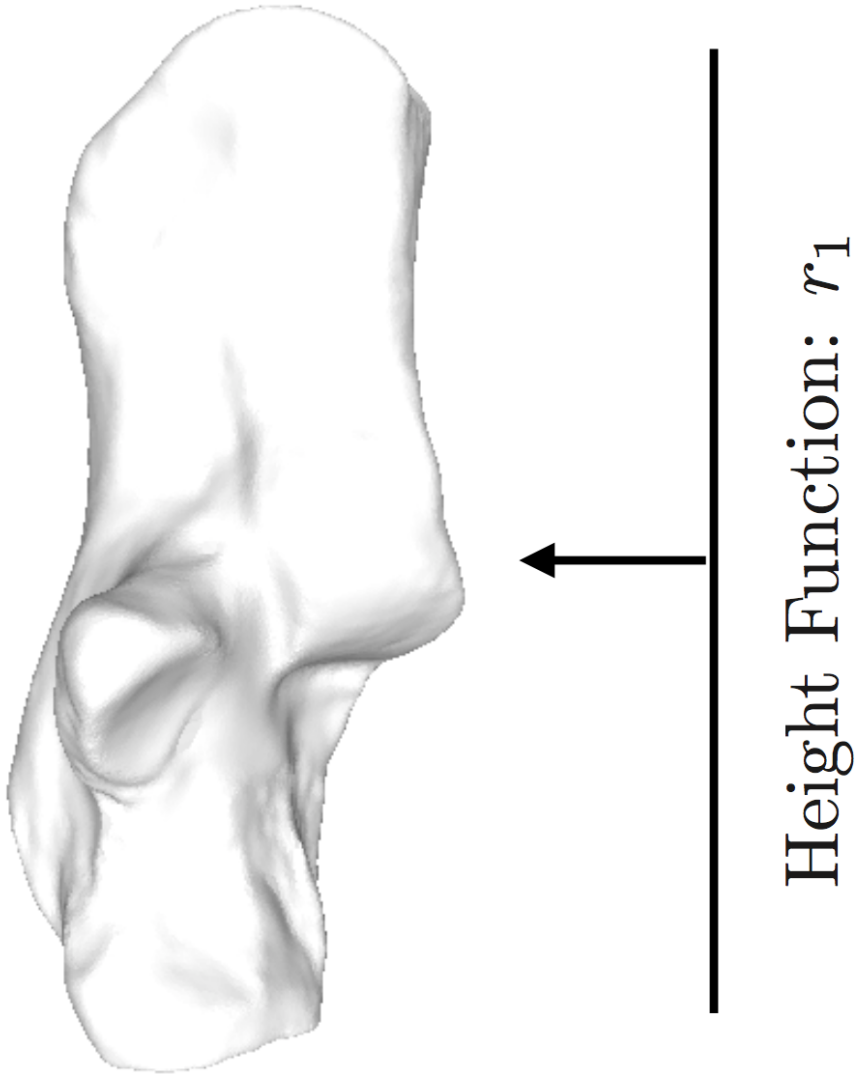
Persistent Homology Transform

For direction ν_1:



Persistent Homology Transform

For direction ν_2:



Persistent Homology Transform

Definition (Turner et al., 2014)

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

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

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

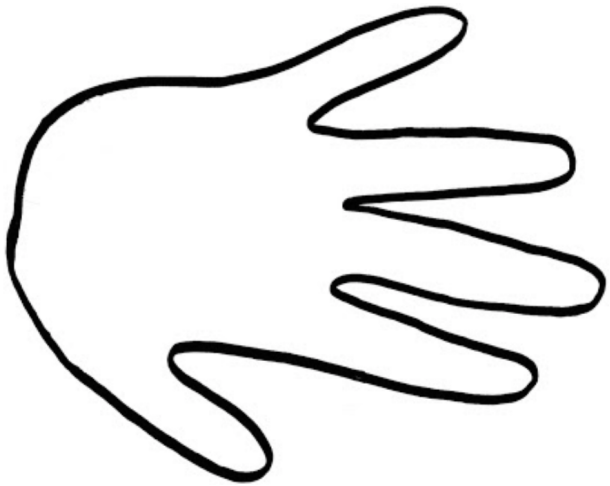
Shortcomings of the PHT

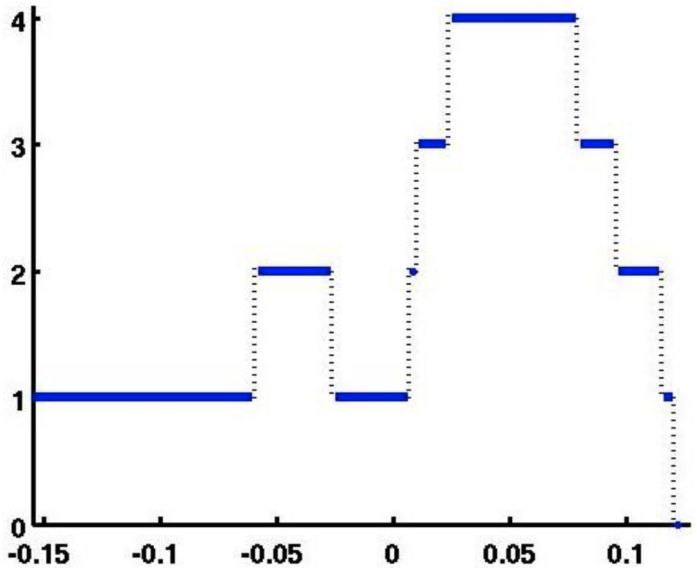
- Most widely used functional regression models use covariates that have an inner product structure defined in Hilbert space
- The geometry of the space of PDs is computationally prohibitive
- The PHT does not admit a simple inner product structure (it is a collection of PDs)
- Therefore, it is challenging to use in standard functional data analytic methods

The Euler Characteristic Curve

Euler Characteristic (EC) : $\chi(K^3) = \#(\text{Vertices}) - \#(\text{Edges}) + \#(\text{Faces})$
EC Curve, $\chi(K_x)$: Calculate EC with sublevel set filtration

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

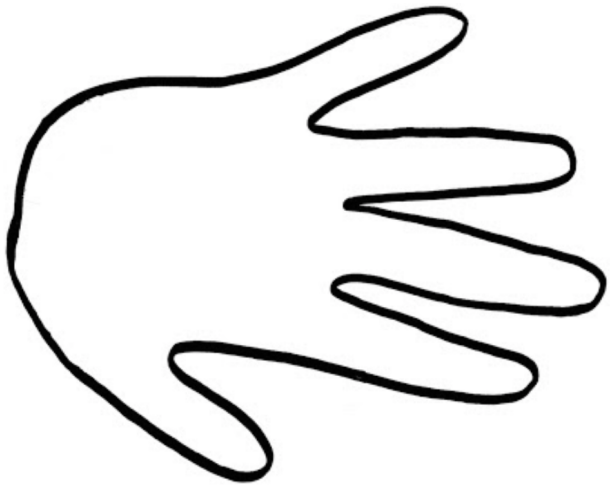


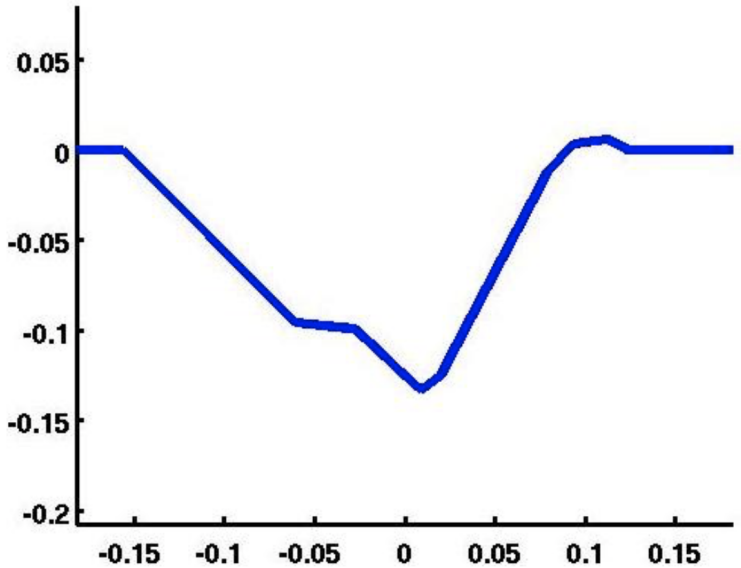


The Smooth Euler Characteristic Curve

The *smooth Euler Characteristic (SEC) curve* is computed by

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





Functional Data Analysis

- SECT summaries are a collection of curves
  $\Longrightarrow$ Infinite-dimensional topological summary statistic
- By construction, the SECT is a continuous, linear function that is an element of Hilbert space $\mathbf{L}^2$ with an 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)

Glioblastoma Multiforme (GBM)

Arising from a class of tumors that affect the central nervous system:

- Most common and most aggressive in humans
- $\sim$15'000 cases in the US in 2016
- Post-diagnosis survival of 12-15 months
- Probability of 5-year survival is $< 10\%$
- Treatment: Stupp Protocol (2005)
  Surgery + Radiotherapy + Chemotherapy $\Longrightarrow$ 26.5% 2-year survival
- No effective treatment for recurrent tumors

Available Data:

- Molecular: e.g. mutation status, DNA methylation, gene expression
  $\Longrightarrow$ Obtained by surgery
- Imaging: e.g. MRI images $\Longrightarrow$ Obtained from radiology

Imaging Data & Radiogenomics

Imaging Data:

- Obtained via radiology
- Relatively accessible, cf. molecular assays

Radiogenomics aims to understand the relationship between clinical imaging data and functional genomic variation:

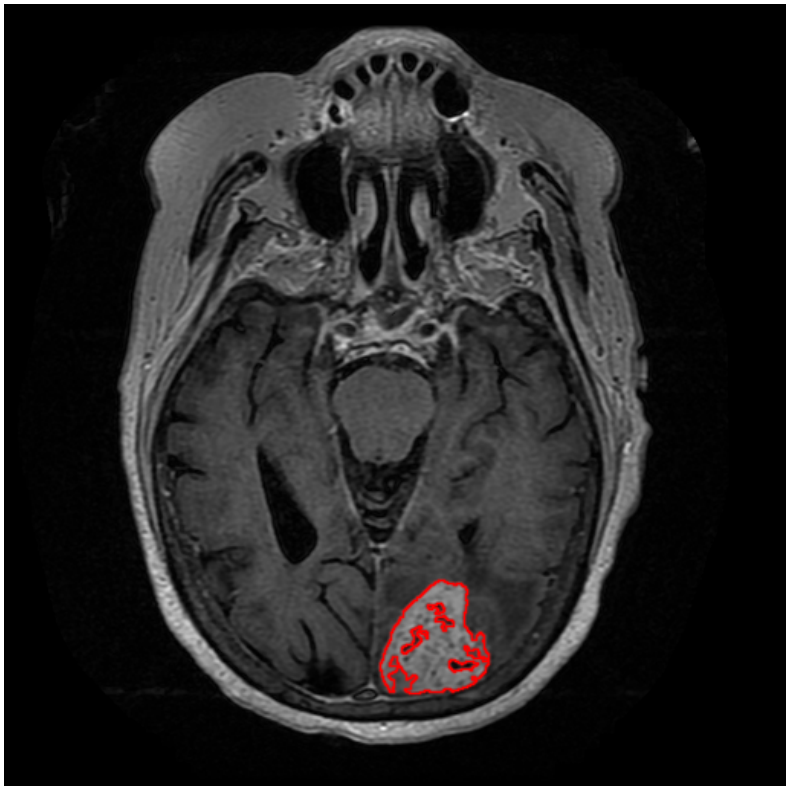
$$\text{Images} + \text{Molecular Data} \Longrightarrow \text{Prognosis, Clinical Decisions, etc.}$$

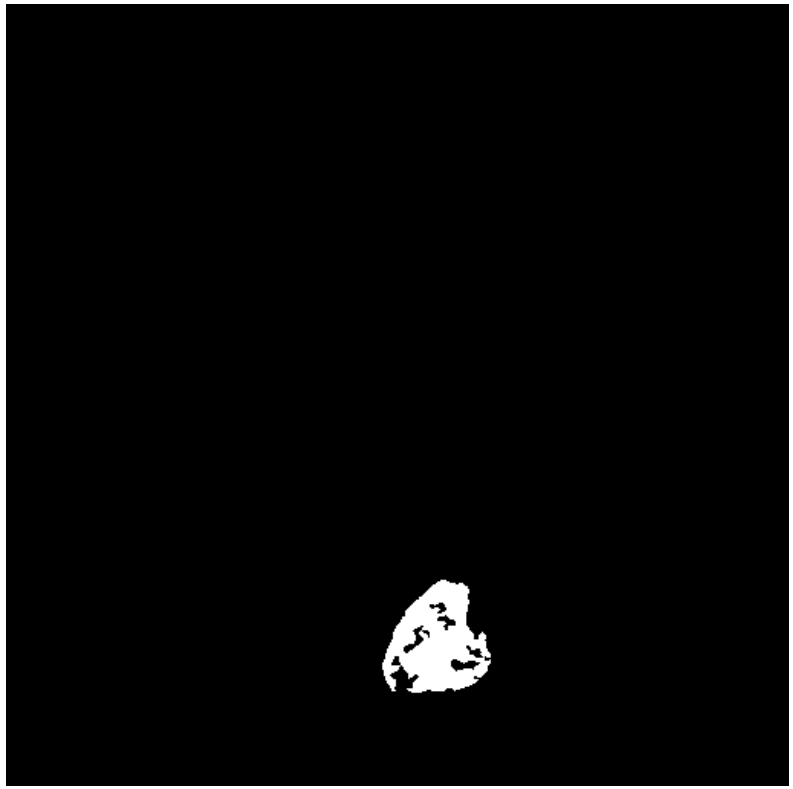
Computational radiogenomics requires the quantification of images

Existing Measures:

- Geometry, e.g. Volume, longest diameter
- Morphometry, e.g. Shape, texture
- Spatial features, e.g. Location of recurrent lesions

MRIs of GBM





Topological Features in Radiogenomics

Topological Characteristics of GBM:

- Multifocality, i.e. existence of multiple lesions $\Longrightarrow H_0$
- Necrosis, i.e. existence of necrotic regions within a tumor $\Longrightarrow H_2$

$\Longrightarrow$ Integrate *topological* features into radiogenomic analysis

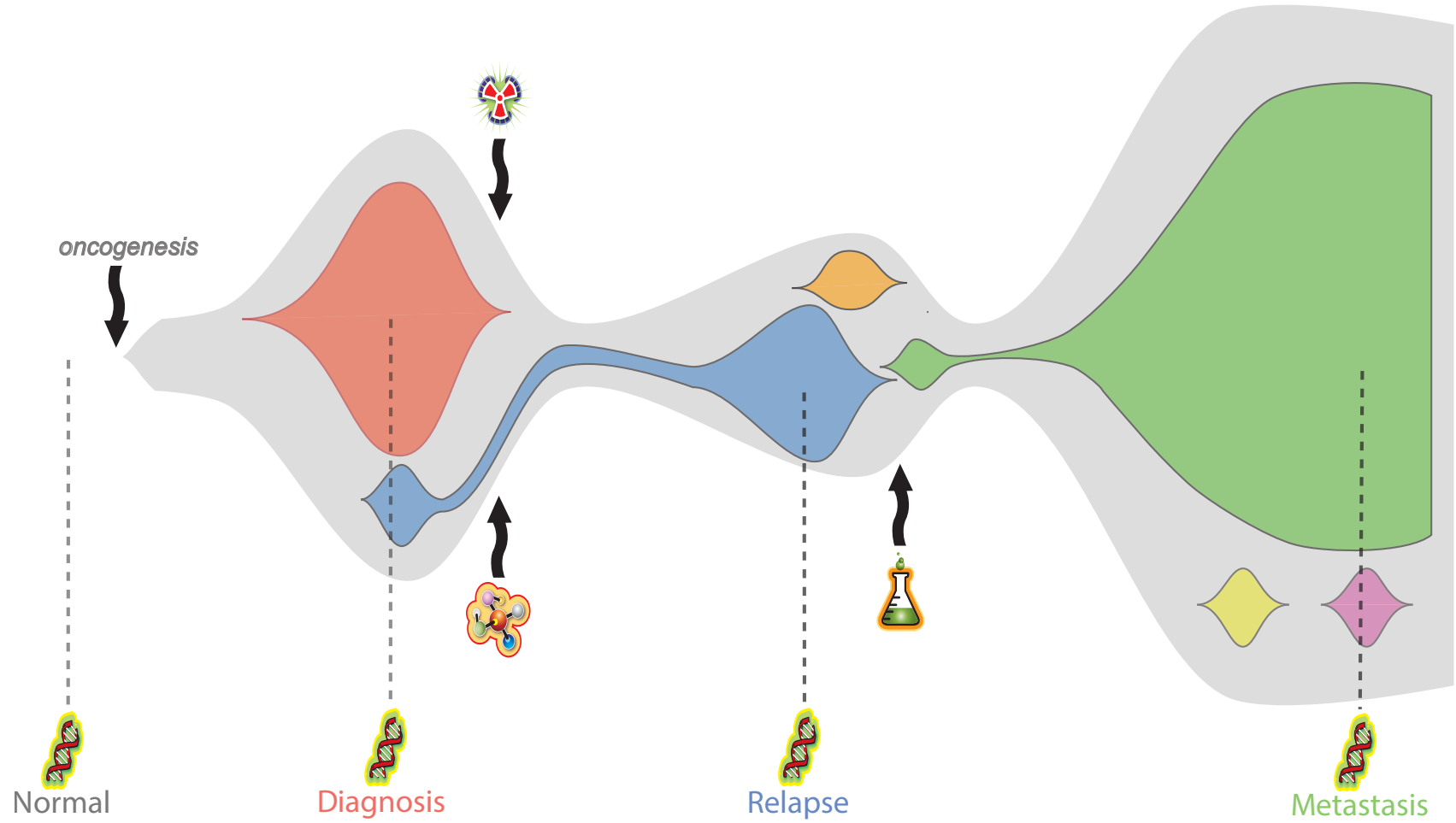
- MRIs of primary GBM tumors were collected from ~ 40 patients in The Cancer Imaging Archive (TCIA)
- These patients also had matched genomic and clinical data in The Cancer Genome Atlas (TCGA)

Goal: Use the SECT to predict clinical outcomes:

- Overall Survival (OS)
- Disease Free Survival (DFS)

Prediction of Clinical Outcomes

- Overall Survival (OS): Treatment $\rightarrow$ Death
- Disease Free Survival (DFS): Successful Treatment $\rightarrow$ Relapse

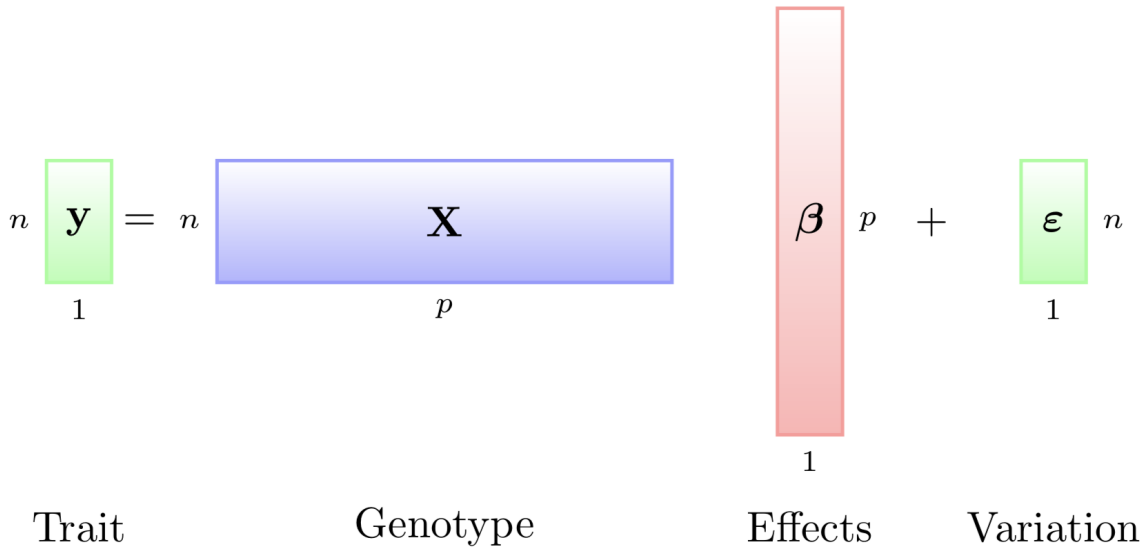


Description of Data & Linear Models

Trait or Phenotype: $\boldsymbol{y}$ is $n \times 1$, with each $y_i \in \mathbf{R}$ (e.g. survival time, crop yield, height, etc.): $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$

Genotype: $\boldsymbol{X}$ is $n \times p$ with each $x_{ij} \in \mathbf{R}$ (e.g. gene expression):

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$



Nonlinear Functional Models for Phenotypic Prediction

Conventional Wisdom in Genetics: Nonlinear functions perform better in genomic selection (i.e. phenotypic prediction)

Conventional Wisdom in Statistics: In high-dimensional regression, smooth nonlinear functions are more predictive than linear functions

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad f \in \mathbf{H}$$

State-of-the-art nonlinear and functional regression frameworks (e.g. kernel models or Gaussian processes) make use of the *reproducing kernel Hilbert space (RKHS)* structure, **H**

Reproducing Kernel Hilbert Spaces

Given a *kernel* (i.e. similarity measure) $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$,

$$\lambda_j \psi_j(\boldsymbol{u}) = \int_{\mathcal{X}} k(\boldsymbol{u}, \boldsymbol{v}) \psi_j(\boldsymbol{v}) d\boldsymbol{v},$$

where $\{(\lambda_i, \psi_j)\}_j$ are eigenvalues and eigenfunctions with

$$k(\boldsymbol{u}, \boldsymbol{v}) = \sum_{j=1}^{\infty} \lambda_k \langle \psi_j(\boldsymbol{u}), \psi_j(\boldsymbol{v}) \rangle,$$

a *reproducing kernel Hilbert space* is defined as

$$\mathbf{H} = \overline{\left\{ f : f(\boldsymbol{x}) = \sum_{j=1}^{\infty} c_j \psi_j(\boldsymbol{x}) \ \forall \ \boldsymbol{x} \in \mathcal{X} \text{ and } \|f\|_{\mathbf{H}}^2 = \sum_{j=1}^{\infty} \frac{c_j^2}{\lambda_j^2} < \infty \right\}}$$

The Representer Theorem & "Kernel Trick"

Key Idea: Use a penalized loss function in

$$\min_{f \in \mathbf{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} L(f, \text{data}) + \lambda \|f\|_{\mathbf{H}}^2 \right\}$$

Theorem (Kimeldorf & Wahba (1971); Schölkopf, Herbrich & Smola (2001))

The solution to the above optimization problem is

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

An infinite-dimensional optimization problem becomes an n-dimensional optimization problem:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Longrightarrow \boldsymbol{y} = \boldsymbol{K}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

A Generalized Functional Kernel Regression Model

Assume the target function $f \in \mathbf{H}$, and using the Representer Theorem and "kernel trick," the estimator function is then

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

We can posit a generalized functional kernel regression model

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{K}) \Longleftrightarrow \boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{K}^{-1})$$

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

We can take:

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

Predicting Clinical Outcomes in Radiogenomics

Compare the SECT with 3 key types of GBM tumor characteristics:

- mRNA Gene Expression Measurements
- Tumor Morphometry
- Tumor Volume and Geometrics

We attempt to predict two clinical outcomes:

- Overall Survival (OS)
- Disease Free Survival (DFS)

Perform 75-25 (in/out of sample) splits, 250 times, in the Bayesian setting

Performance Measure: Root Mean Squared Error of Prediction (RMSEP)

Prediction Results

Gaussian Kernel: $k(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\intercal \boldsymbol{v}/p + h$

	Disease Free Survival	
Data Type	RMSEP	Pr(Optimal)
Gene Expression	0.944 (0.035)	0.20
Morphometrics	0.942 (0.035)	0.07
Volumetrics	0.939 (0.035)	0.06
SECT	**0.803 (0.035)**	**0.69**

	Overall Survival	
Data Type	RMSEP	Pr(Optimal)
Gene Expression	0.981 (0.030)	0.27
Morphometrics	0.965 (0.029)	0.15
Volumetrics	0.964 (0.029)	0.16
SECT	**0.958 (0.028)**	**0.42**

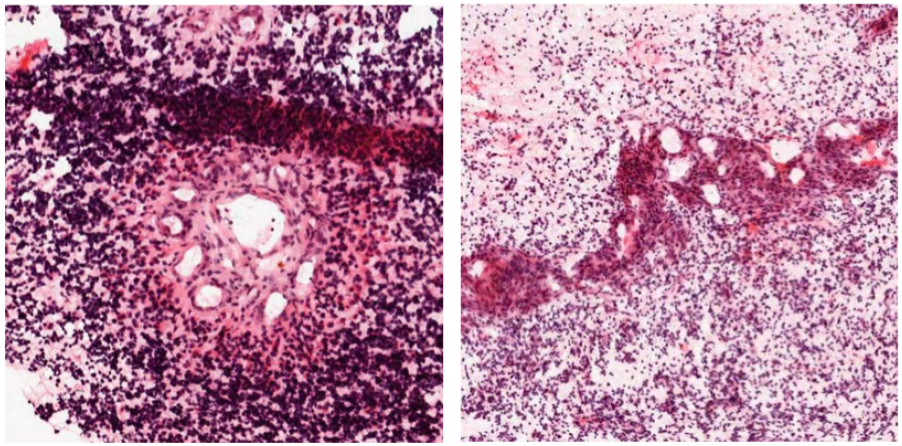
Future Directions and Ongoing Work

From Radiogenomics to Histogenomics:

Joint work with U. Bauer, C. Chen

- Currently, quantitative histopathological analysis for GBM is extremely limited
- Existing methodologies are difficult to apply due to genetic heterogeneity, rapid evolution, and spatial diffusivity
- Histology images of cellular nuclei show a clear spatial architecture

Use TDA to summarize and quantify histology images



From Summary Statistics to Sufficient Statistics

Idea: Sufficient statistics allow for a lower dimensional or less complex representation of data without the loss of information

- Sufficiency for a family/class of distributions via a *statistic*
  e.g. Exponential family, distributions on spaces, order statistics
- Sufficiency for a parameter that defines a distribution
  e.g. $\bar{x}$ for μ in $\mathcal{N}(\mu, \sigma^2)$

M., Kališnik Verovšek, Patiño-Galindo, Crawford: Sufficient statistics for persistent homology based on tropical geometry

$\Longrightarrow$ Allows for parametric modeling of intra- & inter-subtype reassortment in RNA viruses (HIV, avian influenza)

Relevant References

The Persistent Homology Transform (PHT):

- K. Turner, S. Mukherjee, 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, 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, L. Crawford (2017). Tropical sufficient statistics for persistent homology. *arXiv.* 1709.02647.

Publicly Available Source Code & Data Availability

- The Smooth Euler Characteristic Transform (SECT): `https://github.com/RabadanLab/SECT`
- Bayesian Approximate Kernel Regression (BAKR): `https://github.com/lorinanthony/BAKR`

- The Cancer Imaging Archive (TCIA)
- The Cancer Genome Atlas (TCGA)

Thanks

Collaborators:

- Lorin Crawford, Ph.D. (Brown University)
- Andrew X. Chen (Columbia University)
- Sayan Mukherjee, Ph.D. (Duke University)
- Raúl Rabadán, Ph.D. (Columbia University)

Funding: National Cancer Institute Physical Sciences in Oncology Network (NCI PS–ON) 5U54CA193313-02