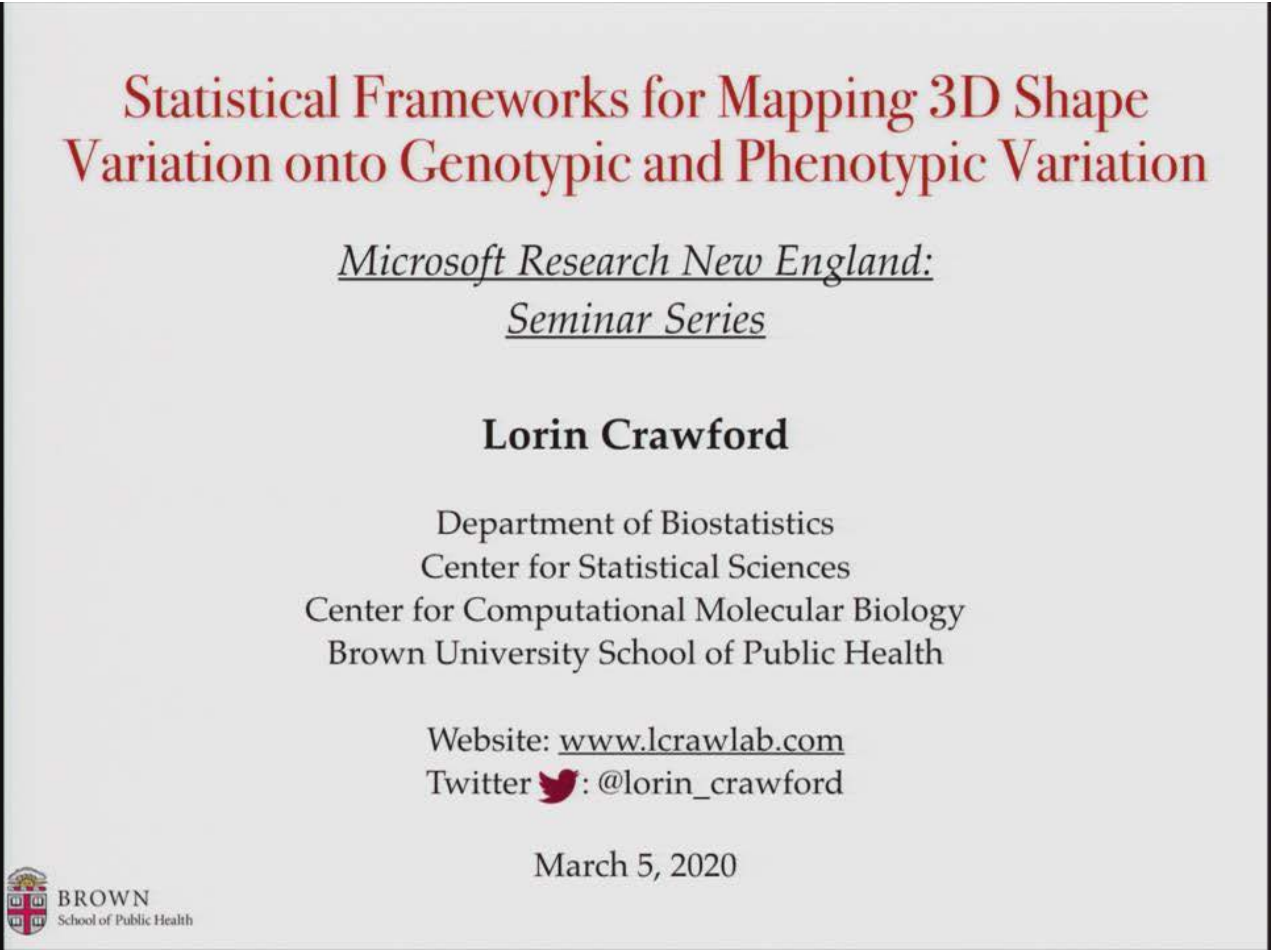


Statistical Frameworks for Mapping 3D Shape Variation onto Genotypic and Phenotypic Variation

Microsoft Research New England:
Seminar Series

Lorin Crawford

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

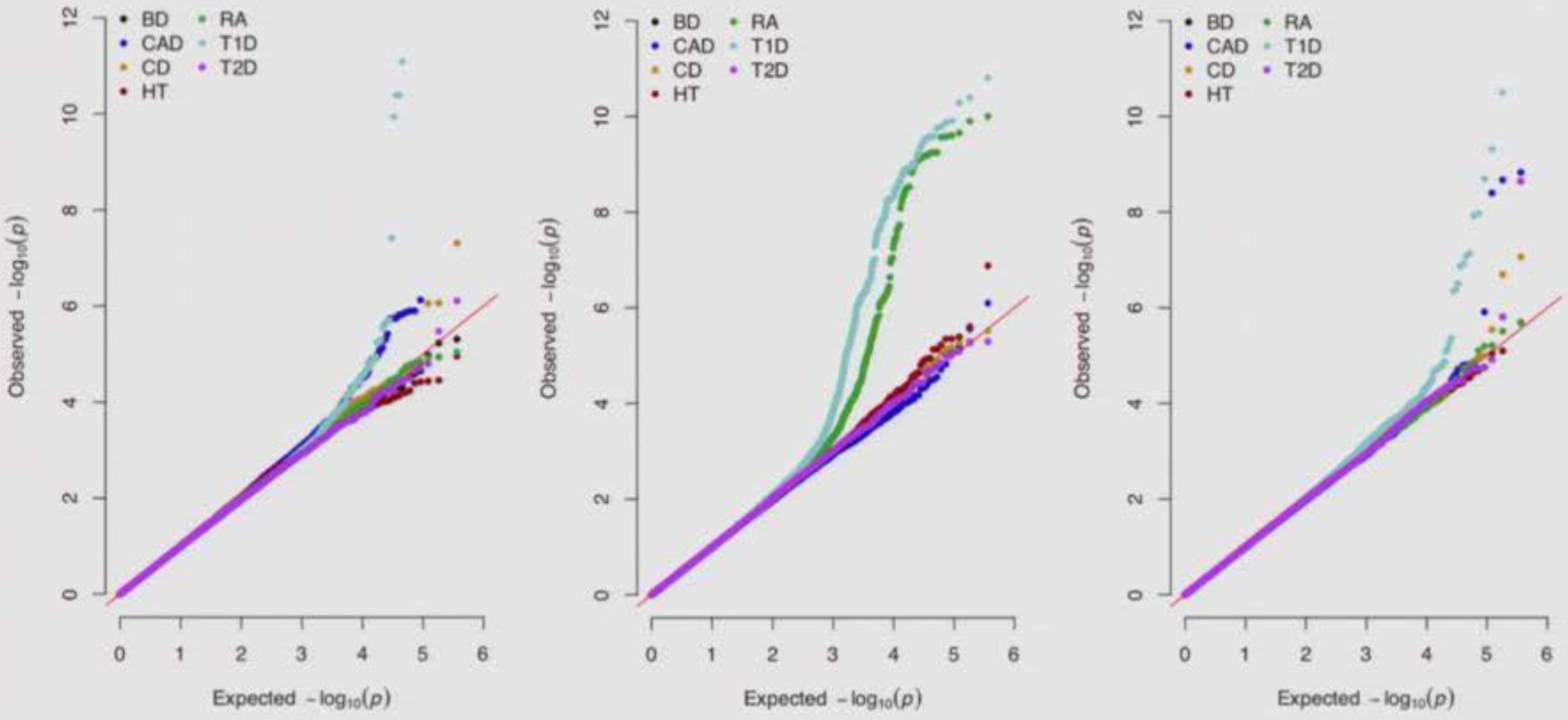
Website: www.lcrawlab.com
Twitter 🐦: @lorin_crawford

March 5, 2020

BROWN
School of Public Health

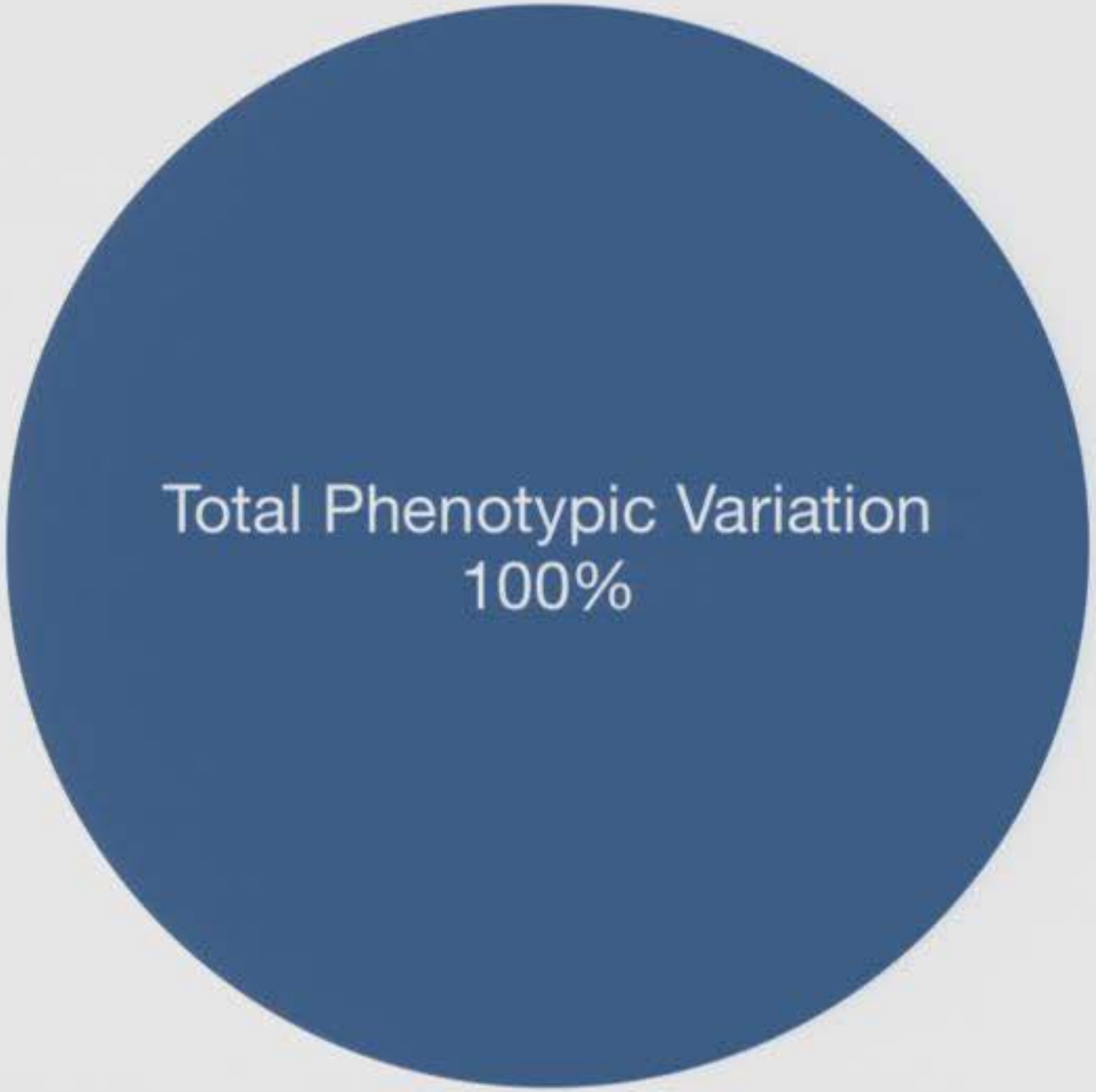
Crawford Lab Motto

Take modern computational approaches and develop theory that enable their interpretations to be related back to classical genomic principles.



(a) Results with $\mathbf{K}_{GW}$ (b) Results with $\mathbf{K}_{cis}$ (c) Results with $\mathbf{K}_{trans}$

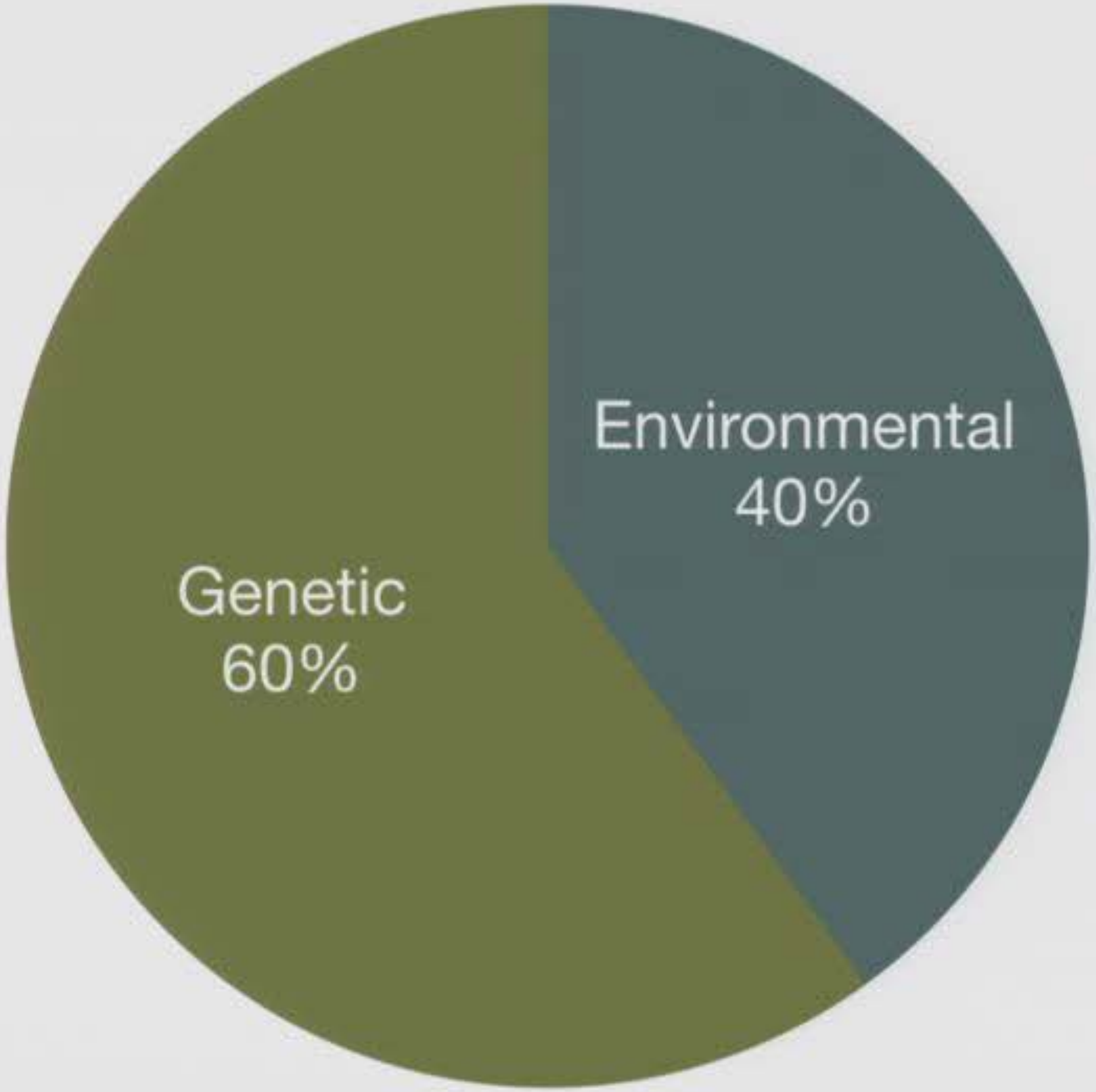
Lab Theme: Dissecting Phenotypic Variation



Total Phenotypic Variation
100%

Lab Theme: Dissecting Phenotypic Variation

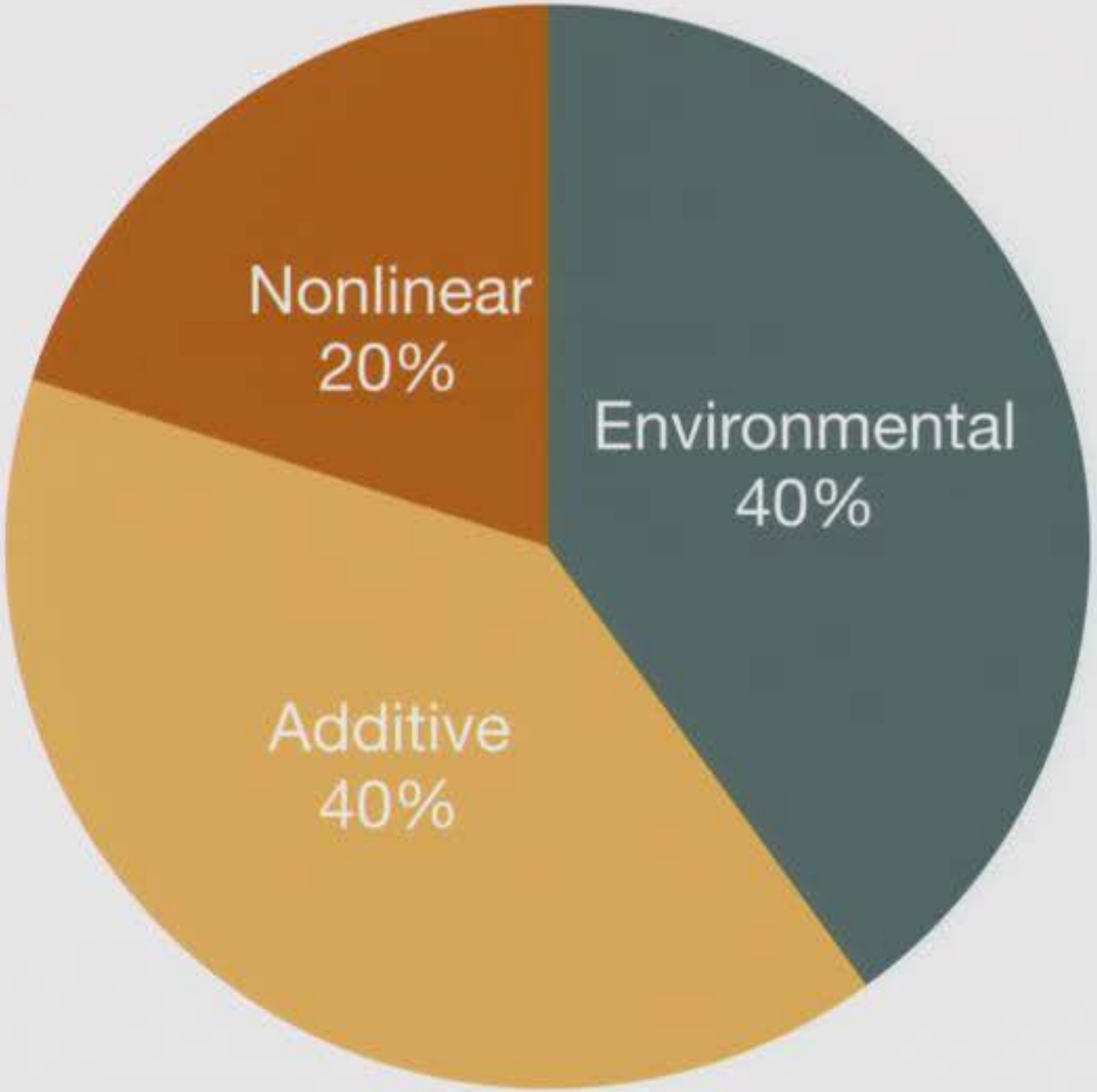
- The phenotypic variance is made up of genetic and environmental effects.



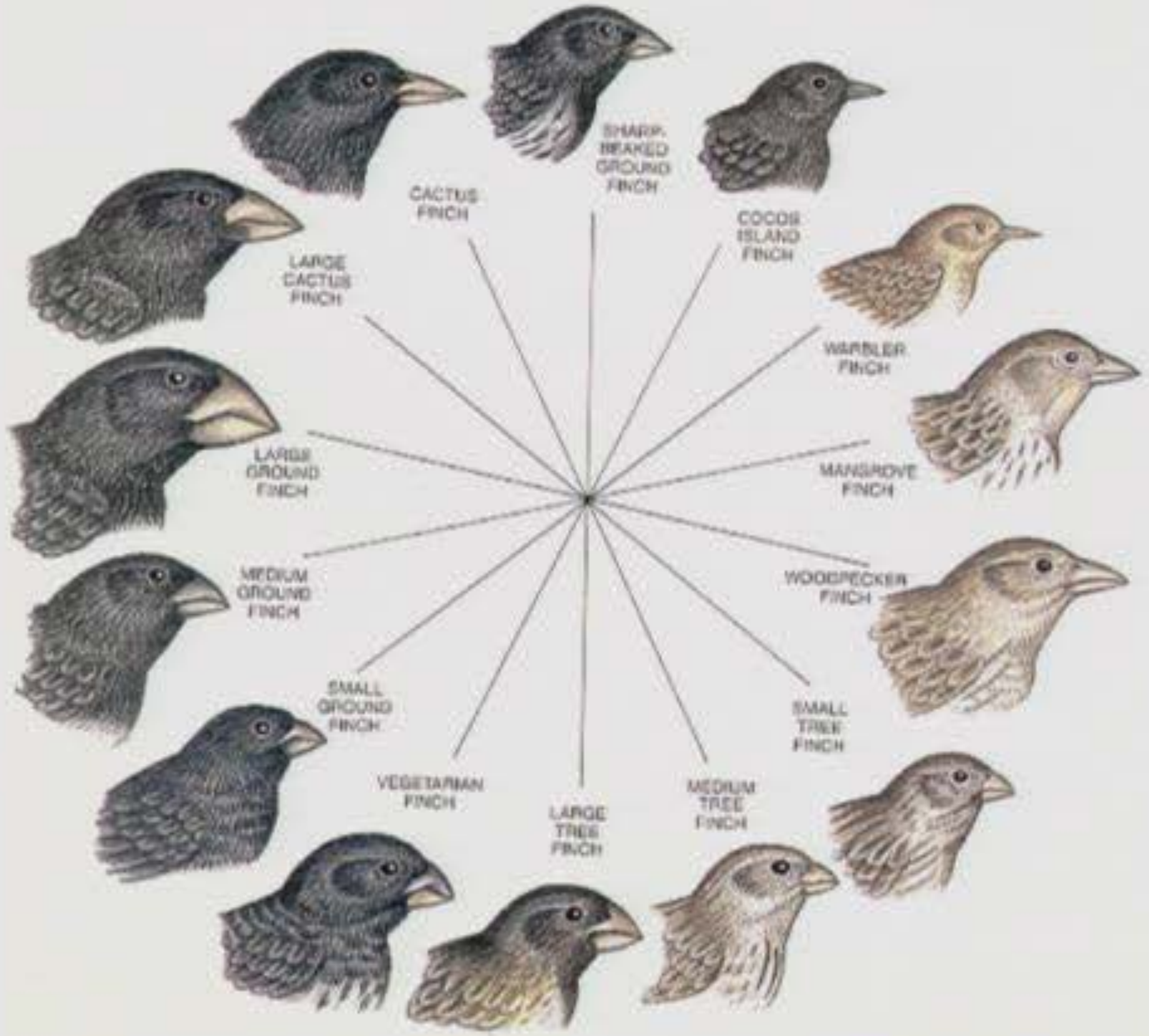
Lab Theme: Dissecting Phenotypic Variation

- The phenotypic variance is made up of genetic and environmental effects.

- Genotypic variation can be dissected into additive effects and nonlinear interactions.

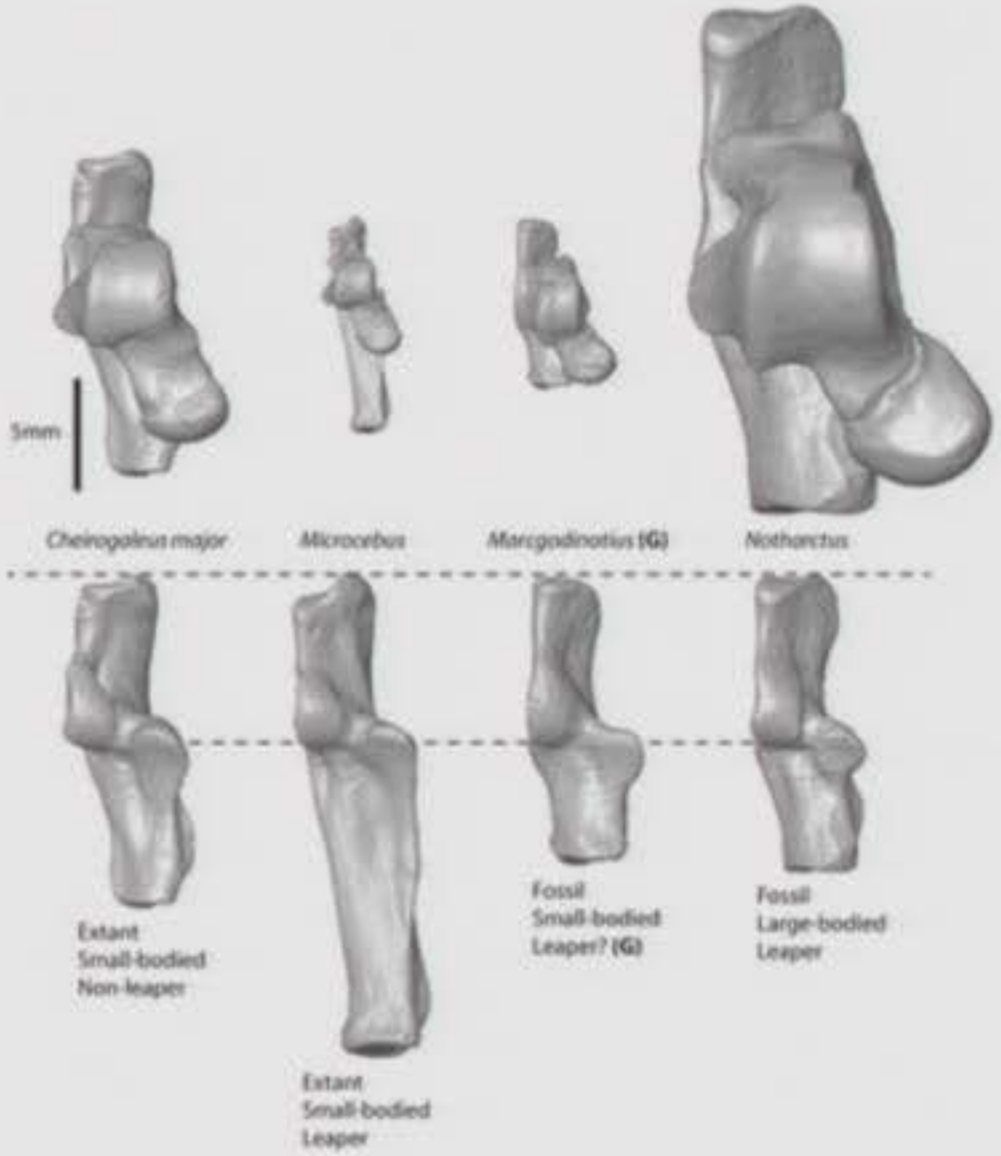


Modeling Variation across Shapes



Phylogeny of Darwin's Finch Beaks

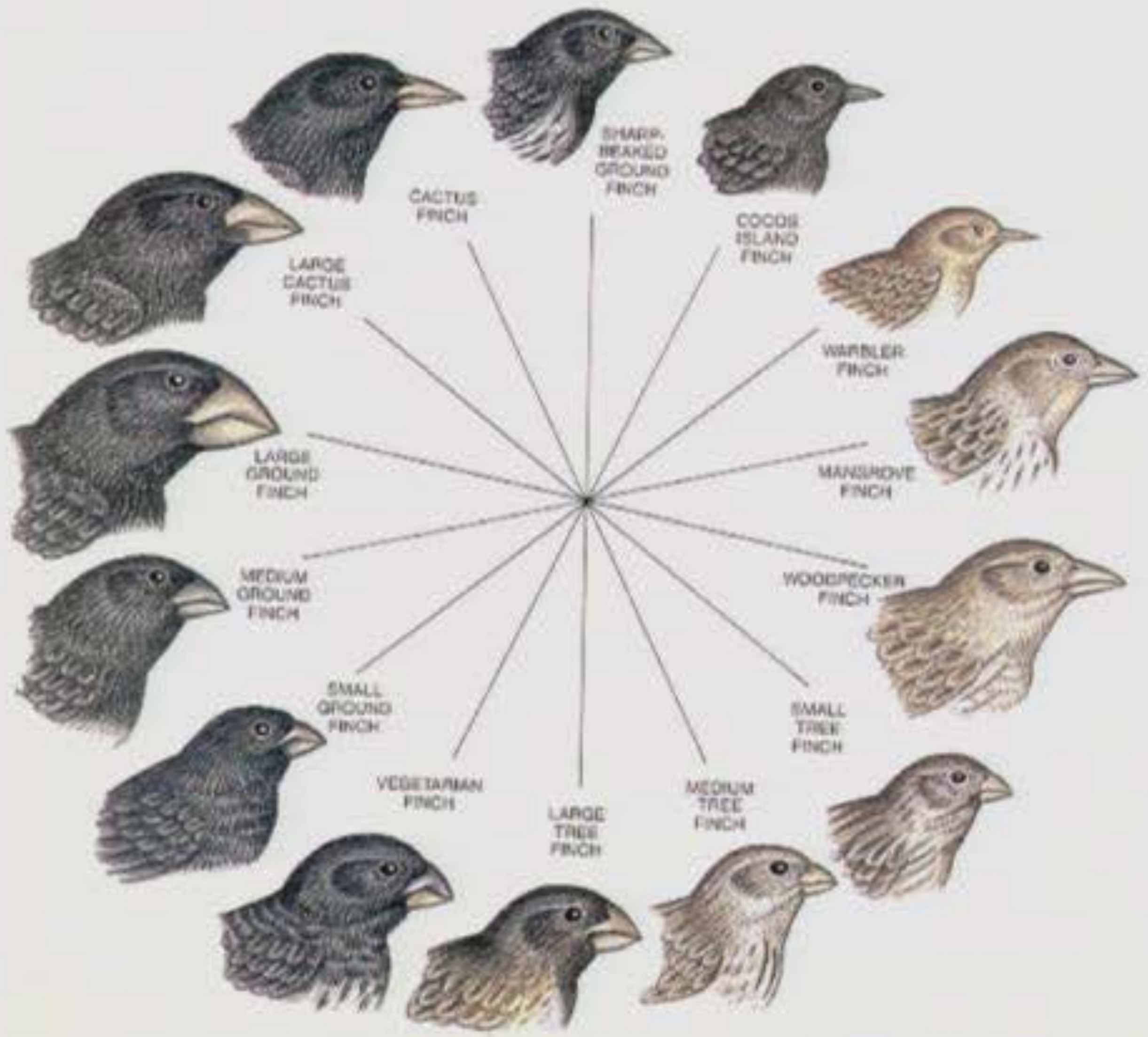
[Gould (1977), *Ontogeny and Phylogeny*]


5mm
Cheirogaleus major
Microcebus
Marcgodinotius (G)
Notharctus
Extant Small-bodied Non-leaper
Extant Small-bodied Leaper
Fossil Small-bodied Leaper? (G)
Fossil Large-bodied Leaper

Fossil Classification

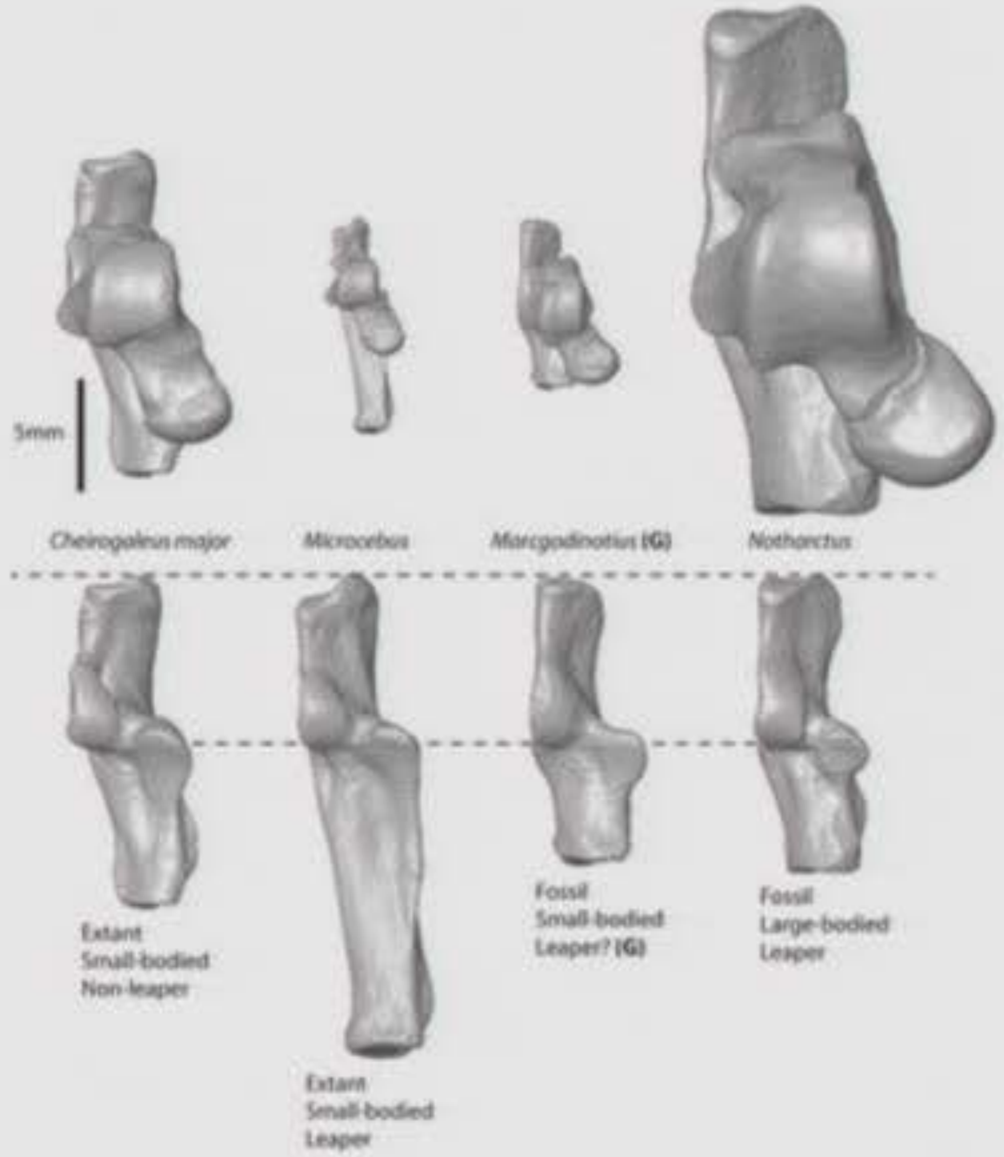
[Boyer et al. (2011), *PNAS*]

Modeling Variation across Shapes



Phylogeny of Darwin's Finch Beaks

[Gould (1977), *Ontogeny and Phylogeny*]


5mm
Cheirogaleus major
Microcebus
Marcgodinotius (G)
Notharctus
Extant Small-bodied Non-leaper
Extant Small-bodied Leaper
Fossil Small-bodied Leaper? (G)
Fossil Large-bodied Leaper

Fossil Classification

[Boyer et al. (2011), *PNAS*]

Presentation Outline

- **Part I: Previous Work with Shapes in Statistics**
  - History of Comparing Shapes
  - Topological Summary Statistics
  - Prediction-Driven Application in Radiomics

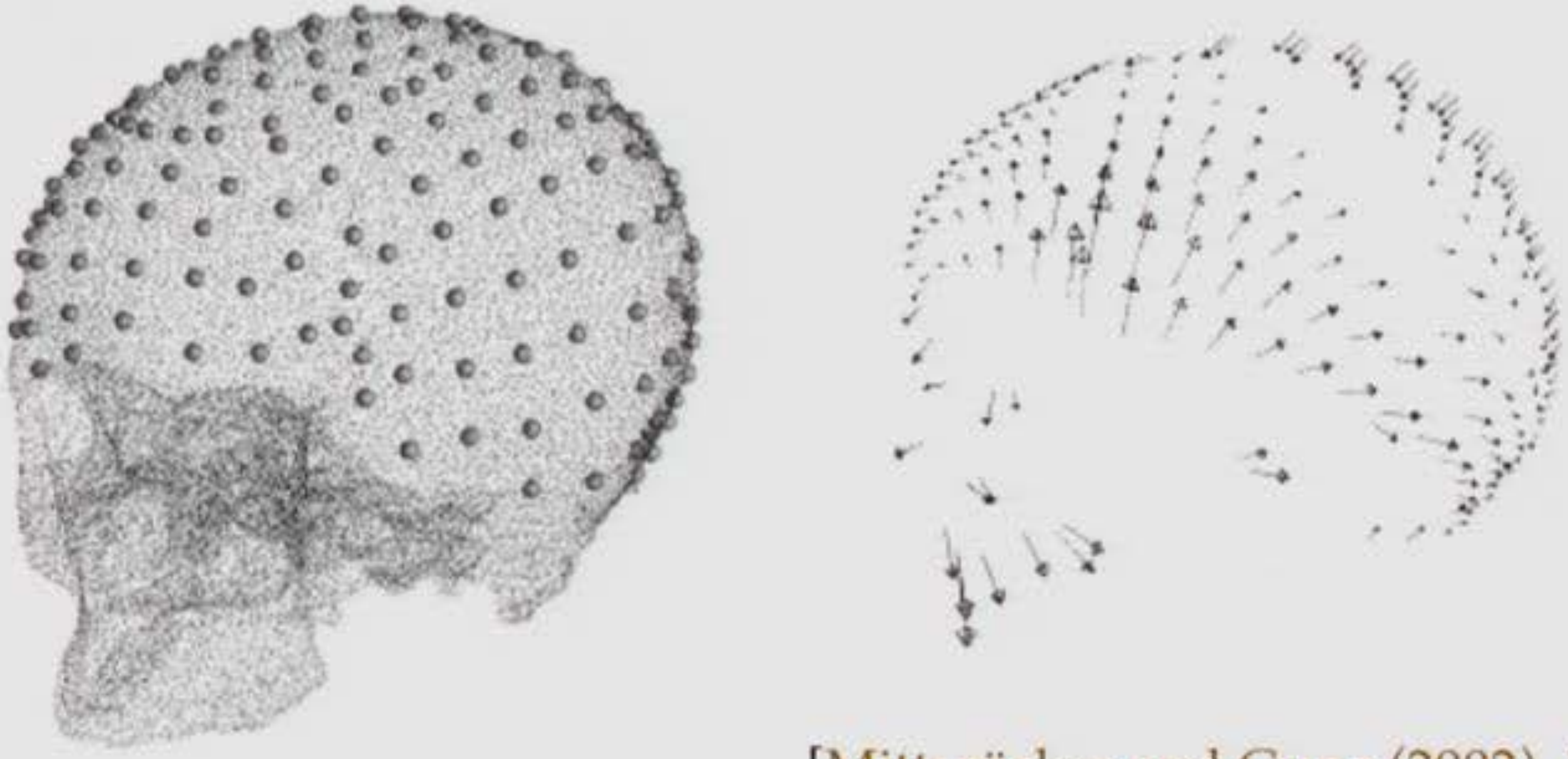
Presentation Outline

- **Part I: Previous Work with Shapes in Statistics**
  - History of Comparing Shapes
  - Topological Summary Statistics
  - Prediction-Driven Application in Radiomics
- **Part II: SINATRA Pipeline for Variable Selection with 3D Shapes**
  - Algorithmic Overview
  - Entropy and RelATive cEntrality (RATE) Measures
  - Reconstruction and Visualization of Enrichment
  - Simulations and Real Data Classification of Shapes

History of Shape Statistics

- Classical shape statistics represented 3D shapes as user-defined landmark points placed on the shape.

- Methods that incorporated information of 3D structure simply did not exist.

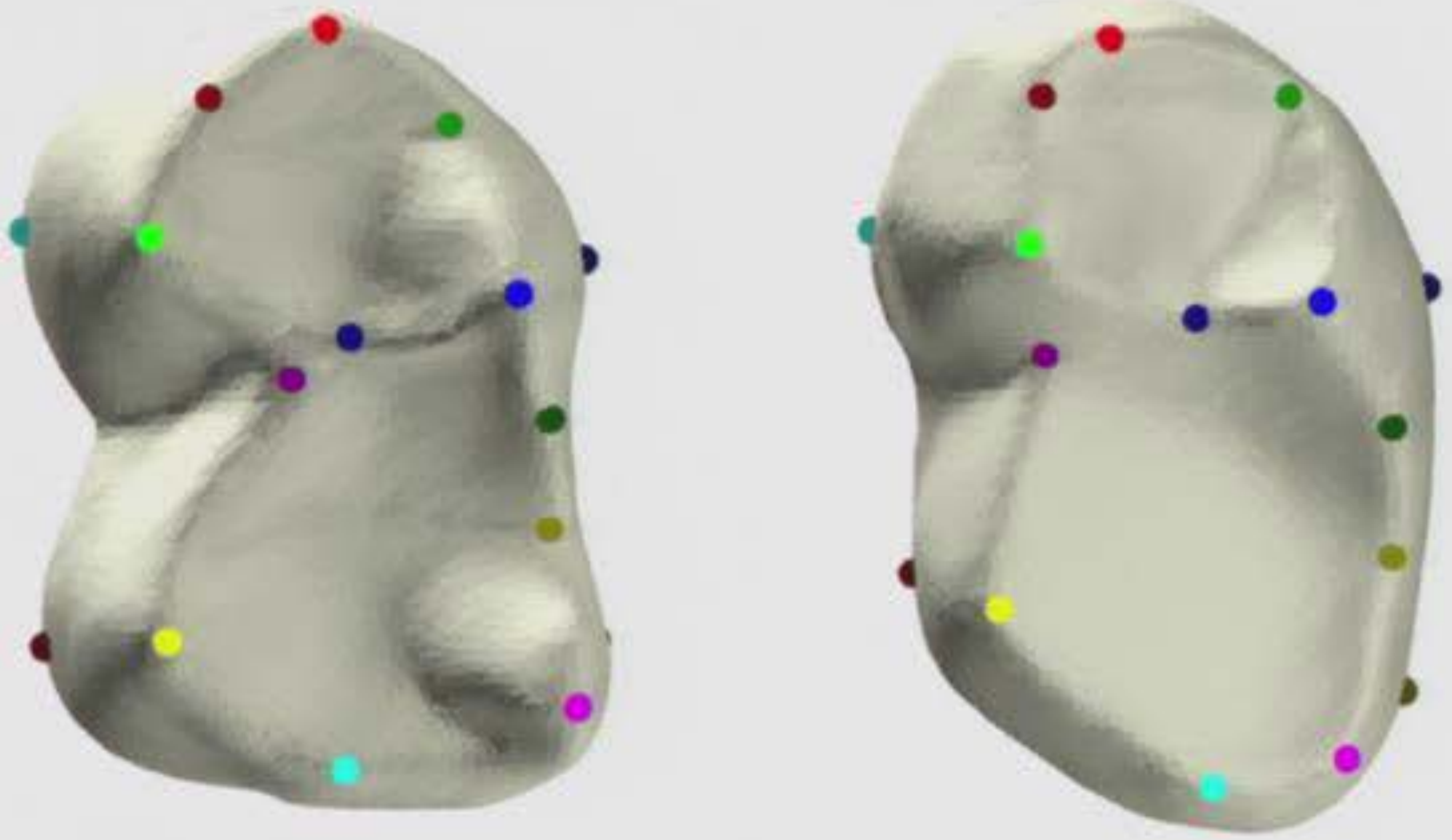


[Mitteröcker and Gunz (2002), *J Phys Anthropol*]

Classic Shape Comparisons

- Recent methods generate (semi-)automatically defined landmark points and bypass the variability caused by user-specifications.

- **Application**: Biological Morphometrics

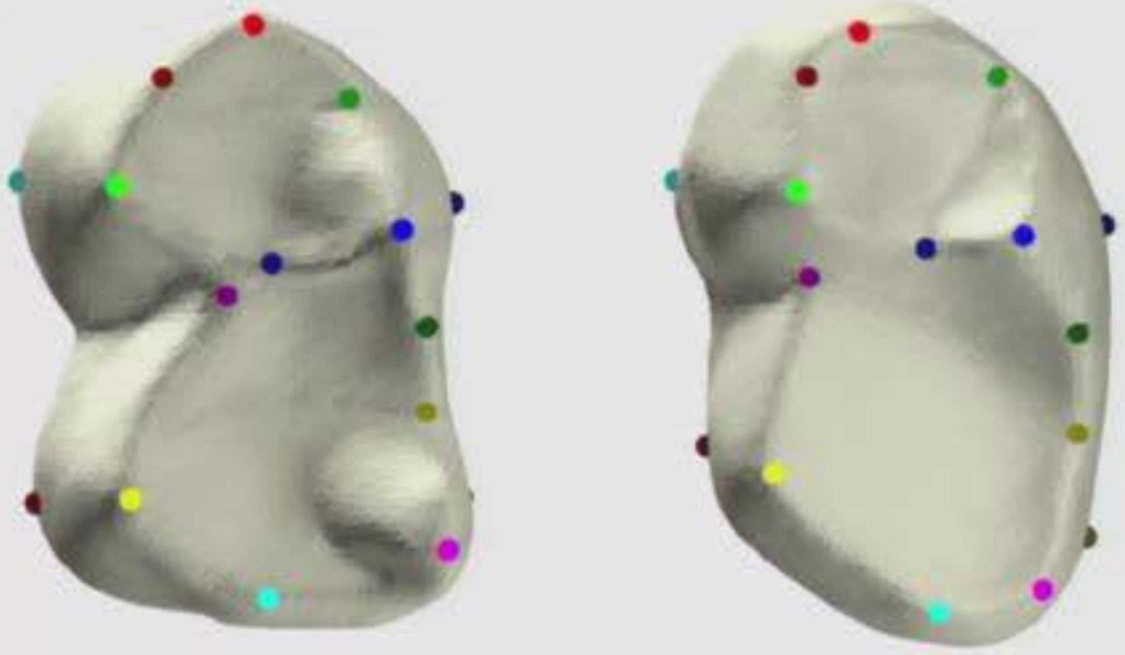


[Boyer et al. (2011), *PNAS* ; Gao et al. (2016), *Anat Rec (Hoboken)*]

Classic Shape Comparisons

- Collect landmarks and compare shapes via some distance metric.

- **Example:** Procrustes Distance

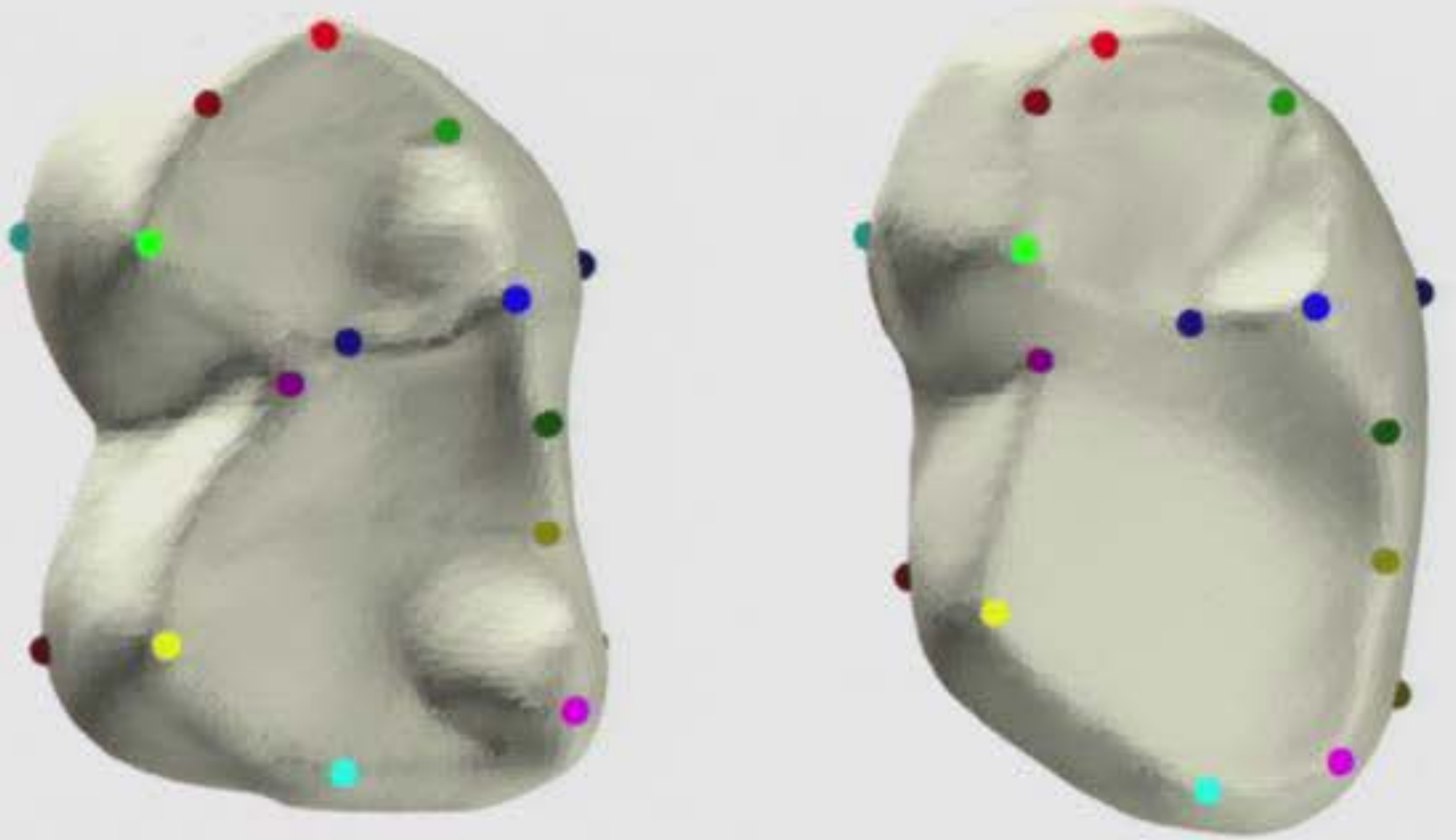


$$\mathcal{D} = \{\mathbf{x}, \mathbf{x}'\}$$

$$\mathrm{d}(\mathbf{x}, \mathbf{x}') = \inf_{r \in R} \left(\sum_{i=1}^{n} \left\| r \frac{x_i}{S_x} - \frac{x'_i}{S_{x'}} \right\|^2 \right)^{1/2}$$

Classic Shape Comparisons

- Recent methods generate (semi-)automatically defined landmark points and bypass the variability caused by user-specifications.
- **Application**: Biological Morphometrics



[Boyer et al. (2011), *PNAS* ; Gao et al. (2016), *Anat Rec (Hoboken)*]

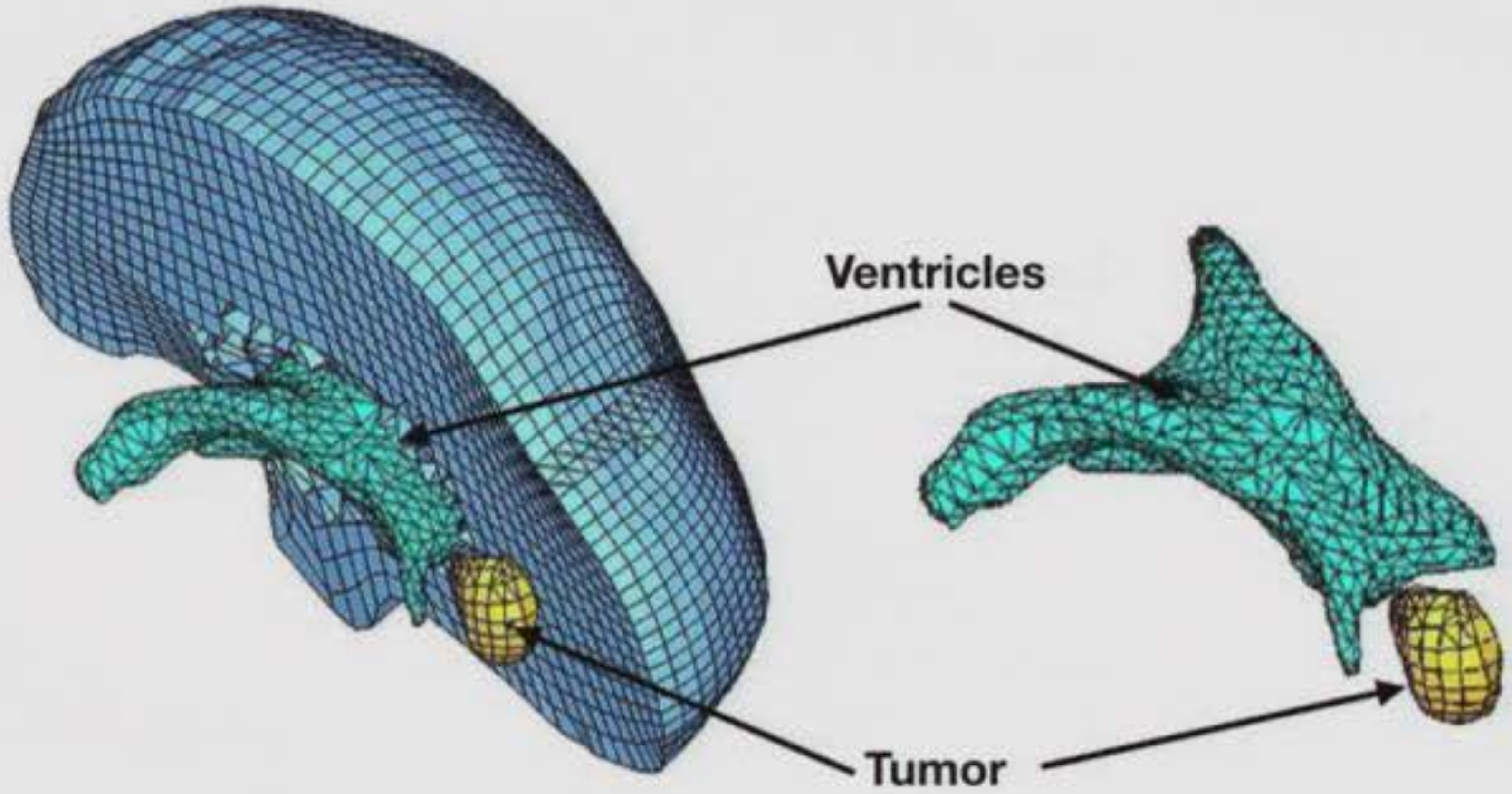
Issues with Landmark-Based Methods

- Current methods for geometric-morphometrics are currently limited to simple pairwise comparisons and often rely on expert-derived landmarks (e.g. Gao et al. (2016), *Anat Rec (Hoboken)*).

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

Shape Representations

- Improved imaging technologies allow 3D shapes to be represented as meshes --- a collection of vertices (V), faces (F), and edges (E).

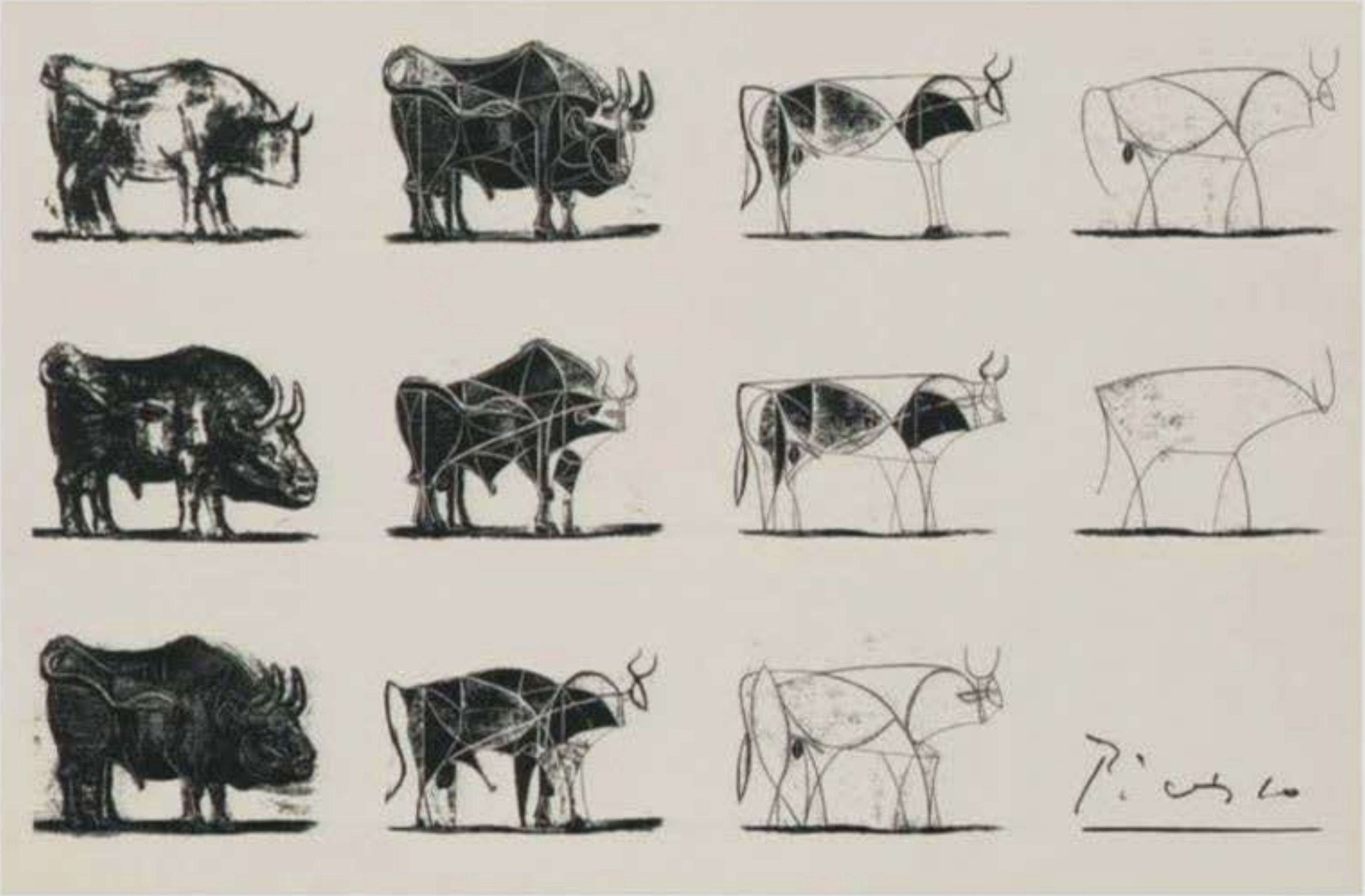


[Boyer et al. (2011), *PNAS*; Crawford et al. (2020), *JASA*]

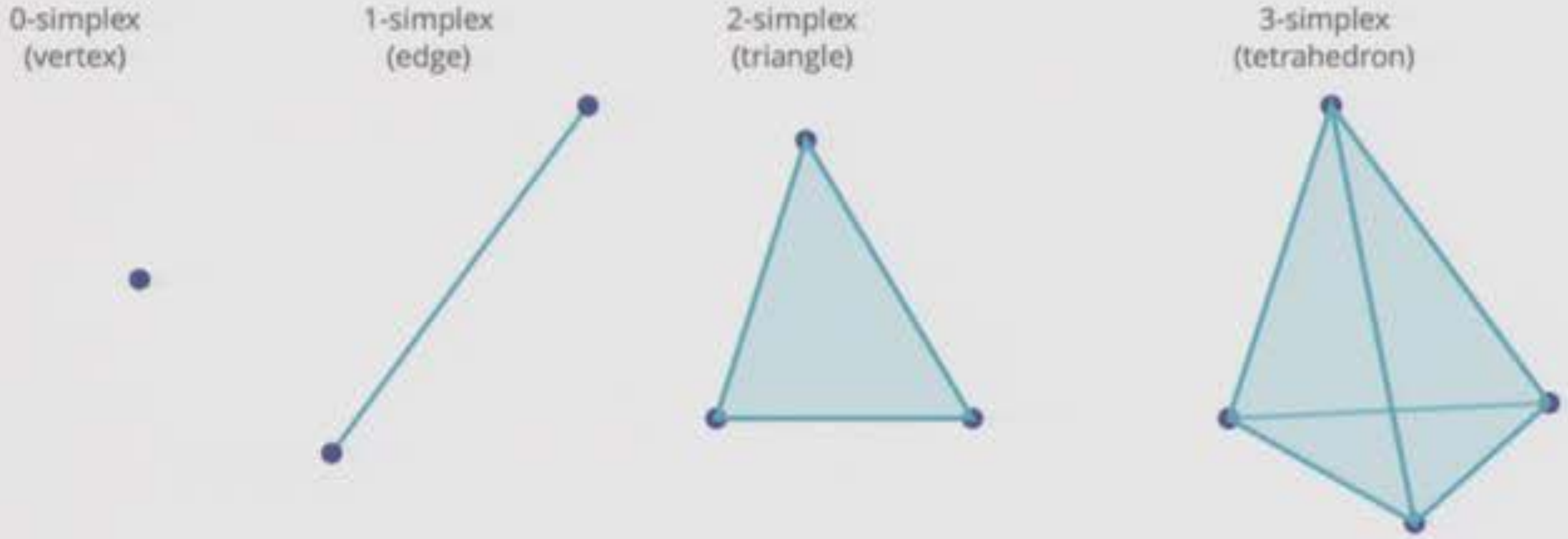
Main Objective(s)

- **Alternative transformation that can be used in wide range of regression and machine learning methods:**
  - Generalized linear models (GLMs)
  - Neural Networks
- **Desired Transformation Properties:**
  - Injective mapping or (even better) explicitly invertible
  - Compute distances and define probabilities in the transformed space
- **Topological Summaries:**
  - Persistence Landscapes (PL)
  - Persistent Homology Transform (PHT)
  - Euler Characteristic Transform (ECT)

Motivating Topology with Picasso

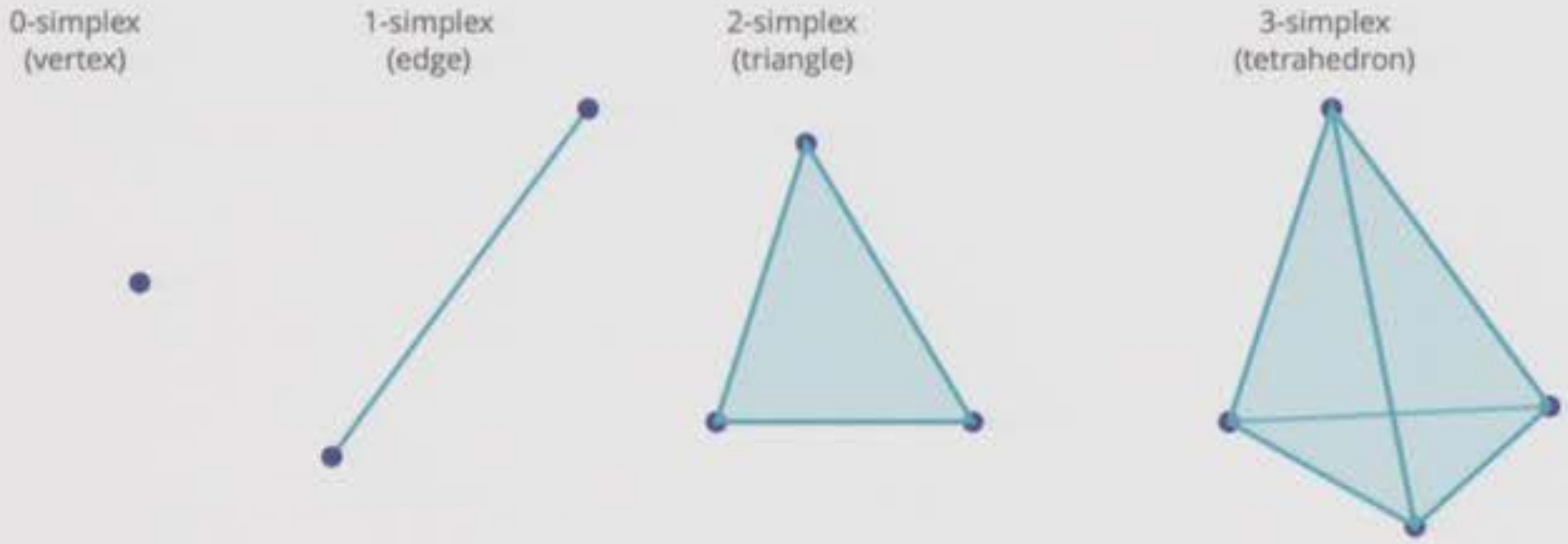


Persistent Homology



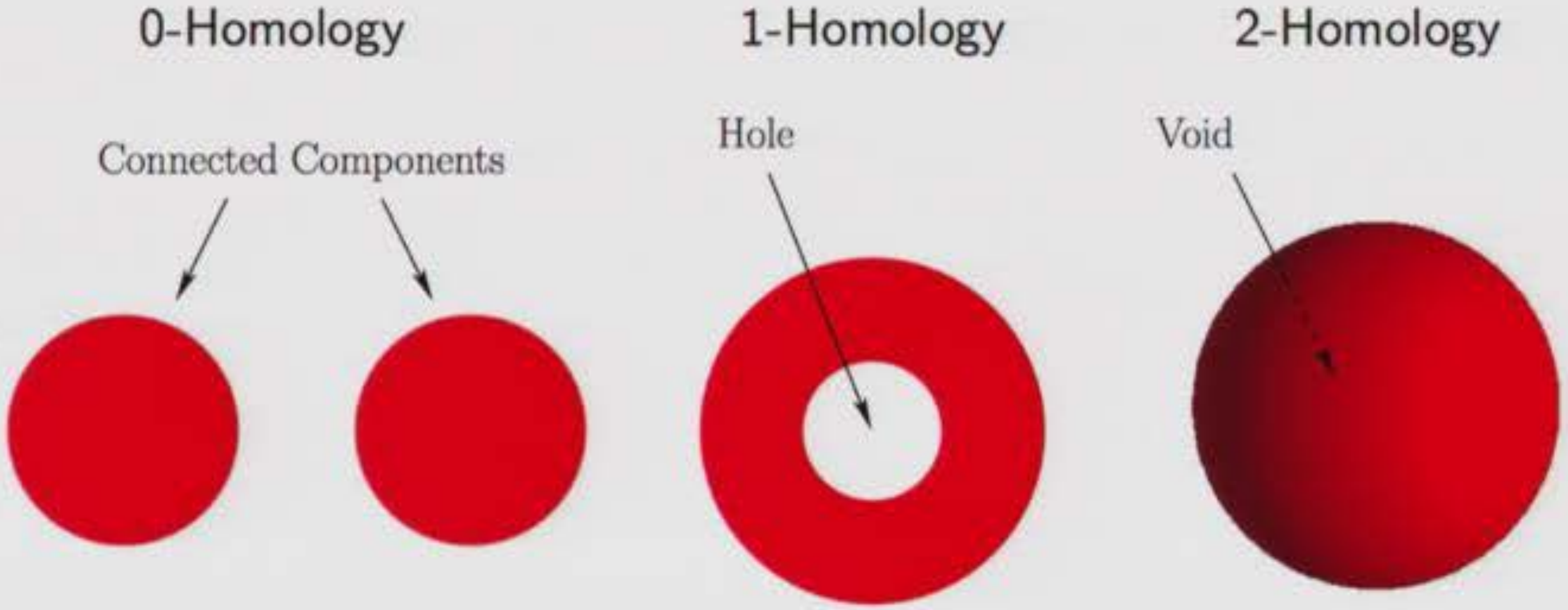
(A) Simplicial Complexes

Persistent Homology



(A) Simplicial Complexes

(B) Homology Groups and Betti Numbers



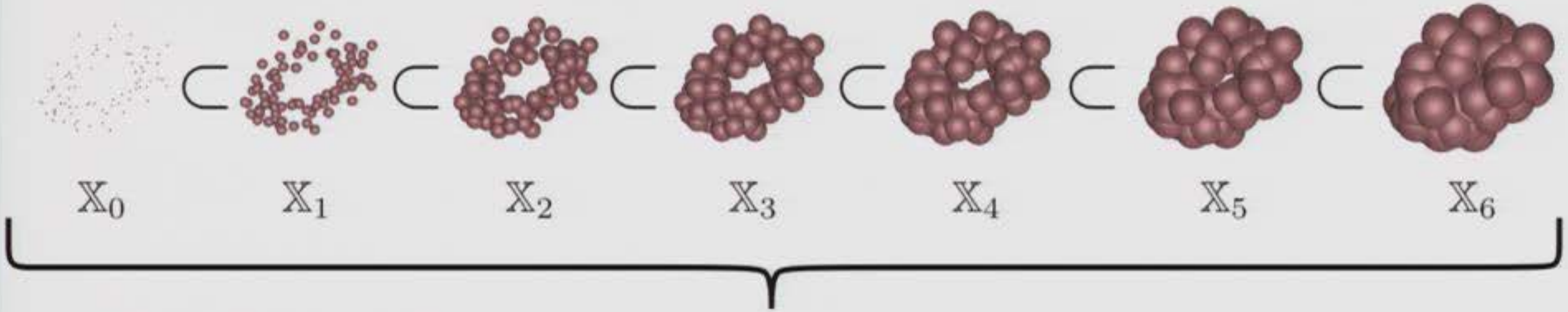
$\beta_0 = 2,\ \beta_1 = 0,\ \beta_2 = 0$

$\beta_0 = 1,\ \beta_1 = 1,\ \beta_2 = 0$

$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1$

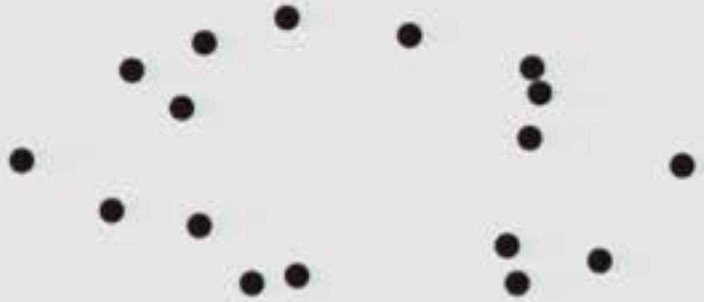
Persistent Homology

Construct some filtration operator…

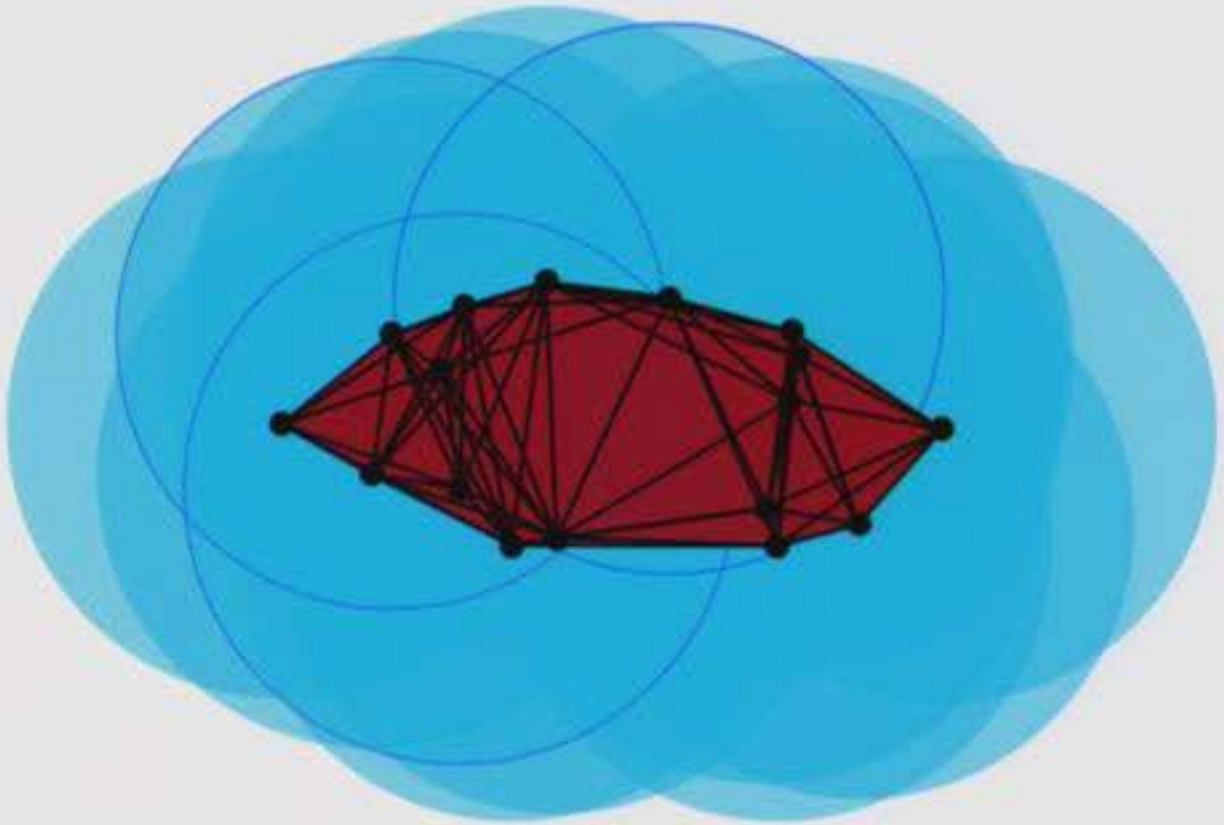


Persistent homology tracks the evolution of homology via collections of simplicial complexes

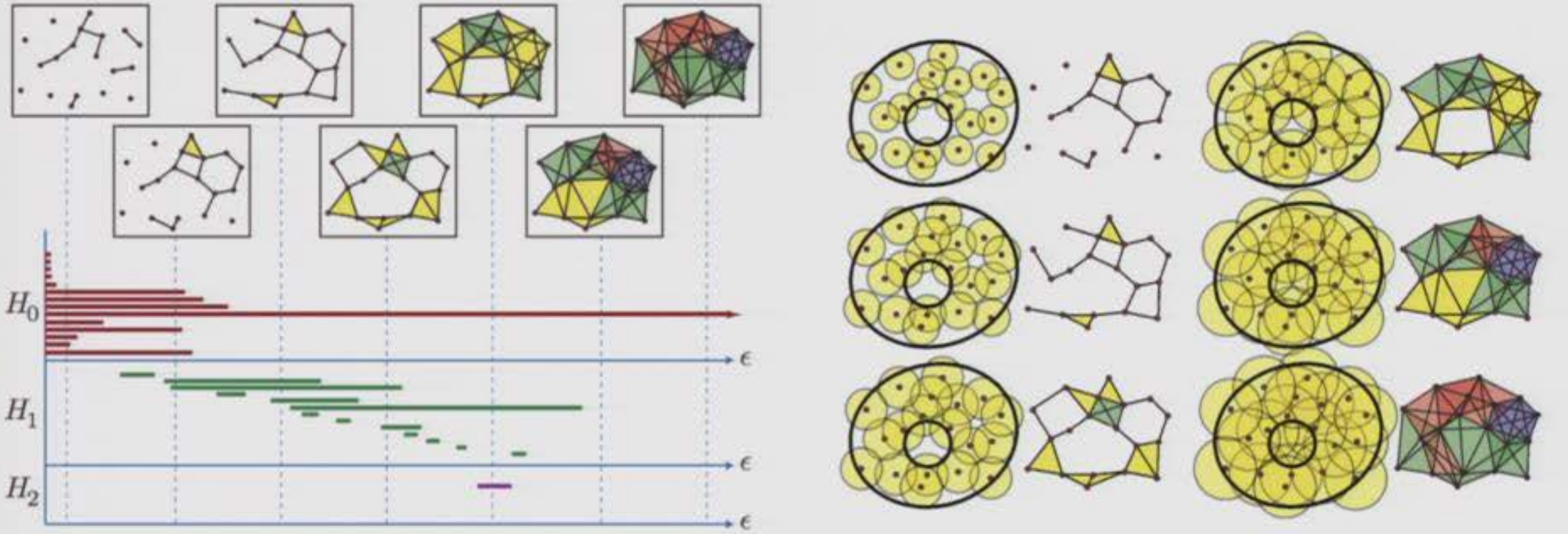
Persistent Homology: A Visual Demonstration



Persistent Homology: A Visual Demonstration



Persistent Homology



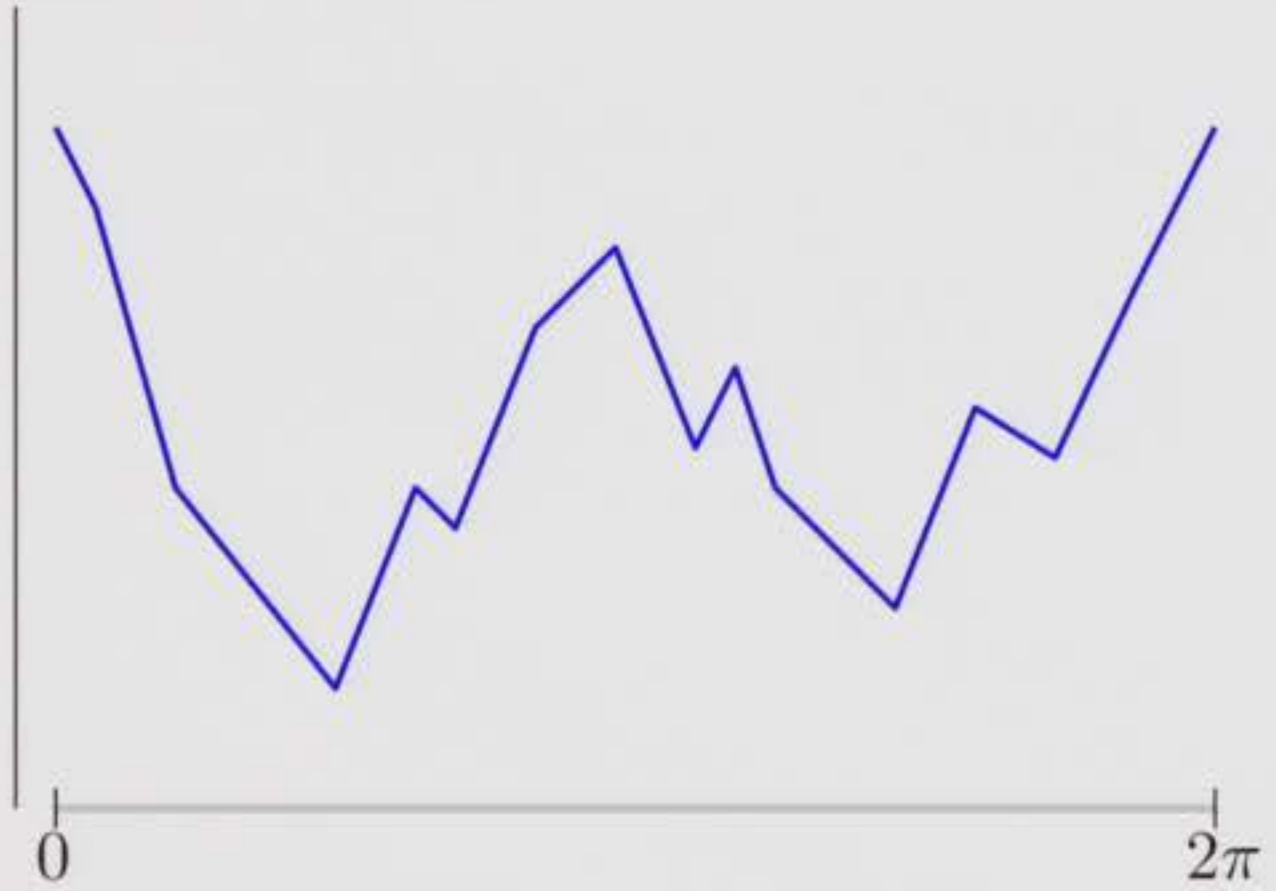
(a)

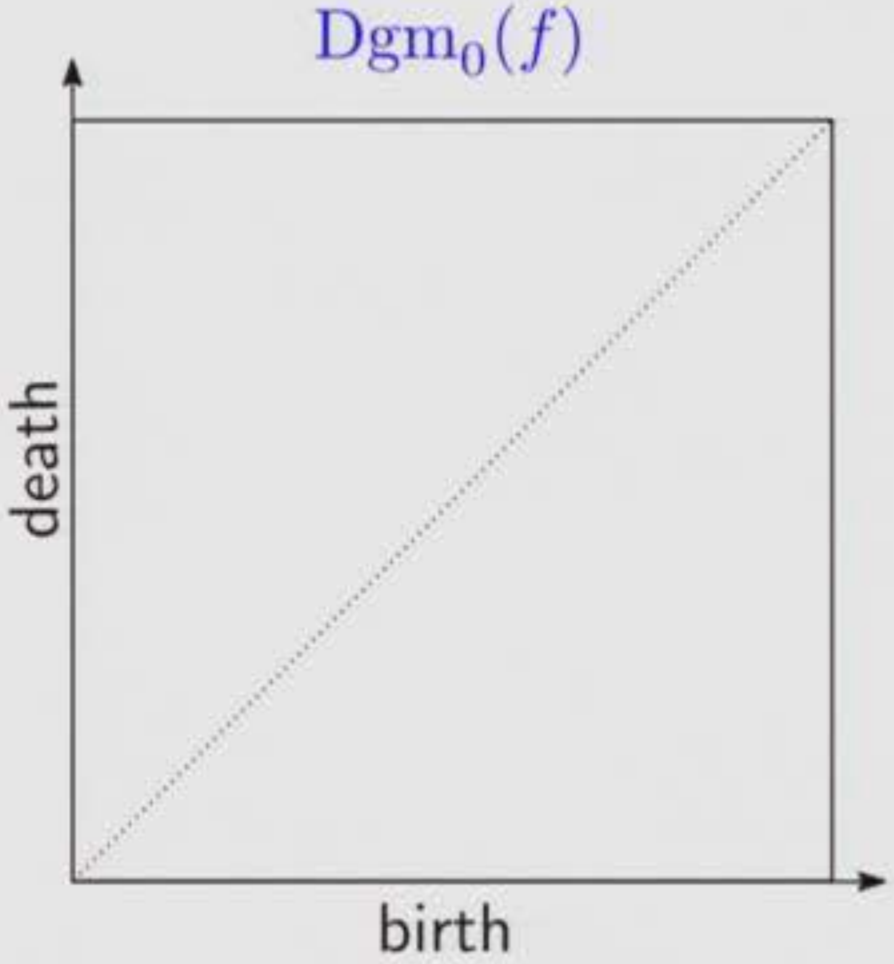
(b)

[Carlsson (2014), *Acta Numer*]

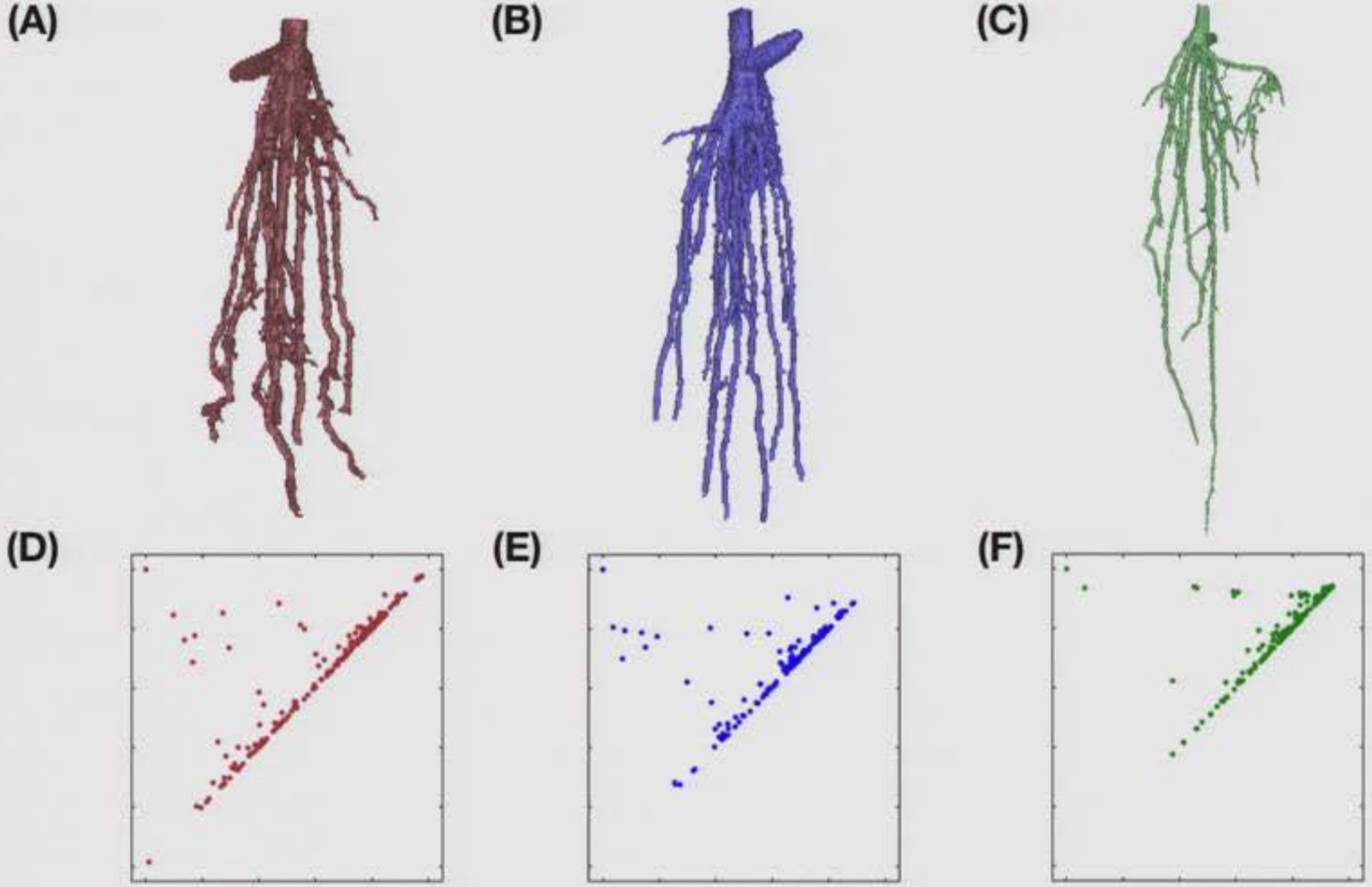
Persistent Homology: A Visual Demonstration

Evolution of homology as a birth-death pair.





Practical Example: 2D Maize Roots



Persistent Homology Transform for 3D Shapes

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.

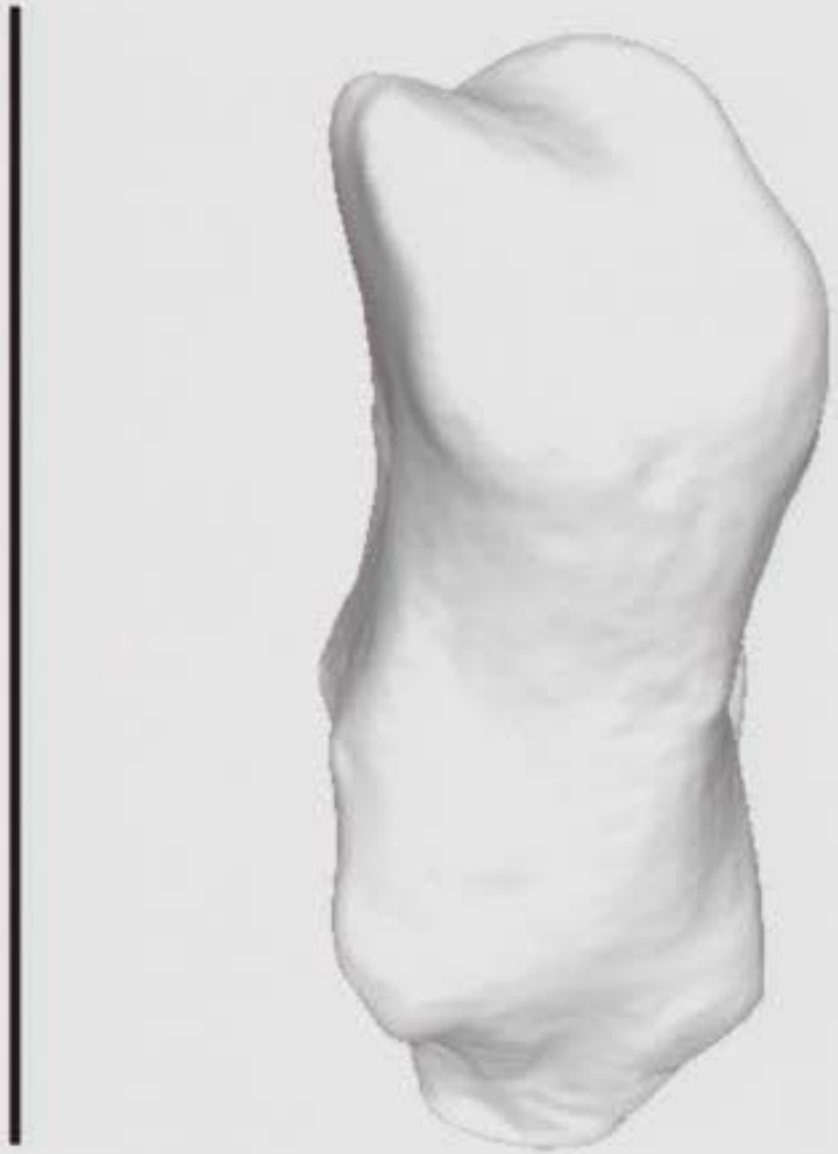
[Turner et al. (2014), *Inf Inference*]

Persistent Homology Transform for 3D Shapes

For direction ν_1:

Persistent Homology Transform for 3D Shapes

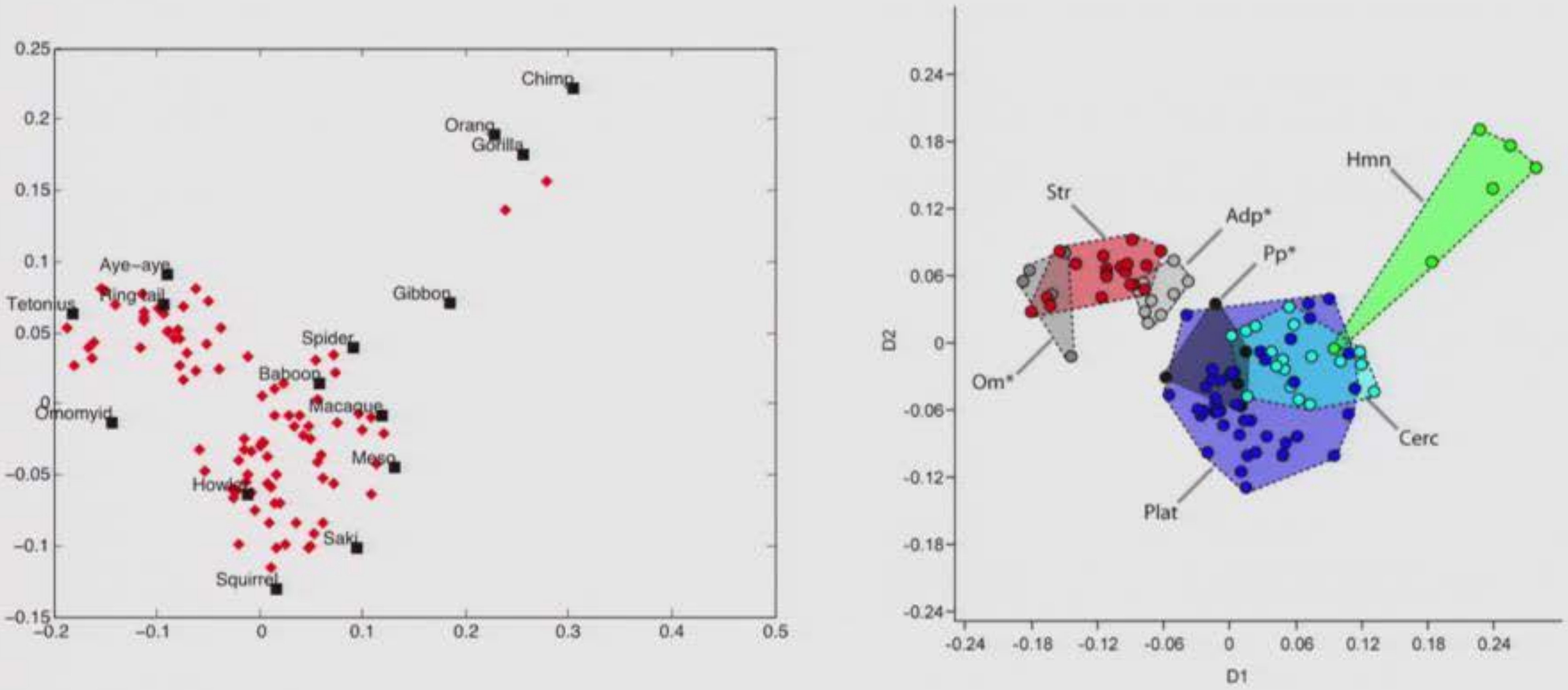
For direction ν_1:



Persistent Homology Transform for 3D Shapes

For direction ν_2:

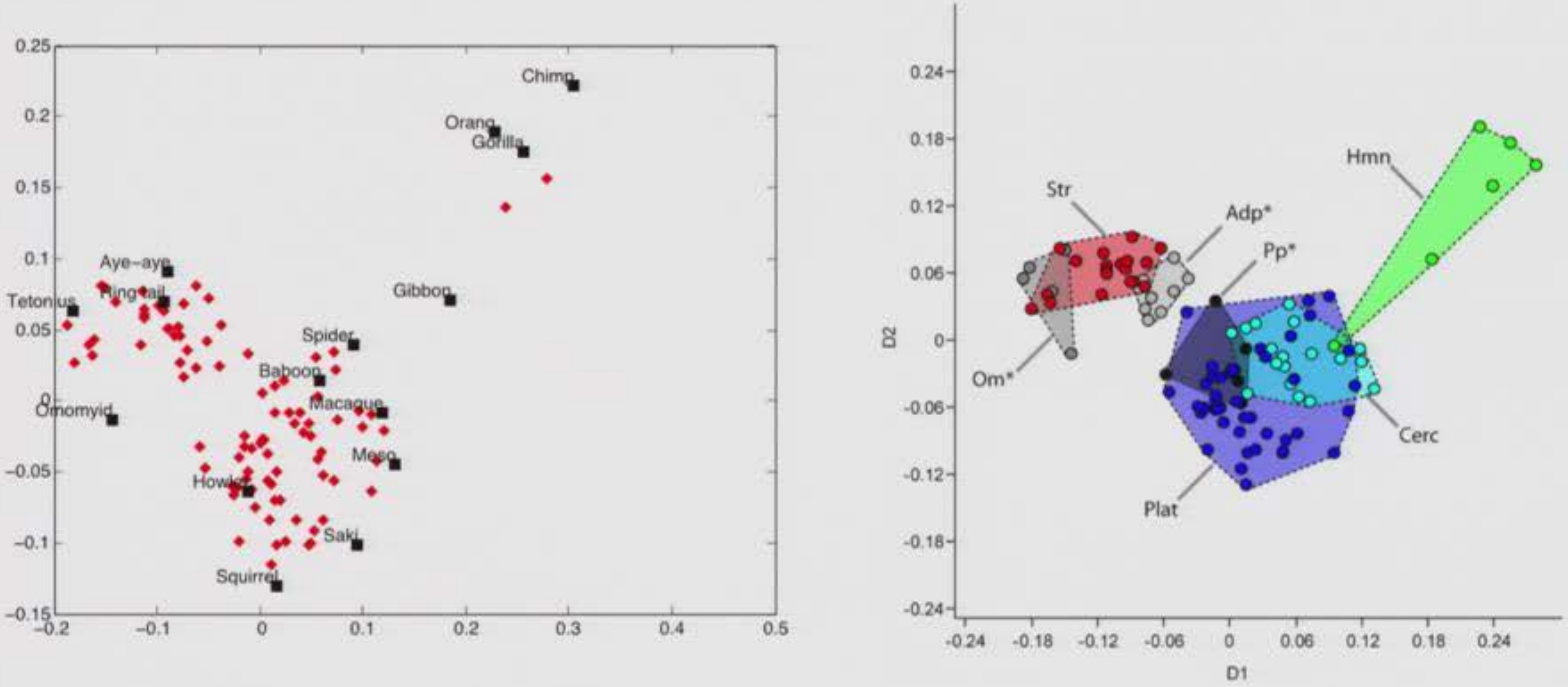
Shape Analysis Using the PHT



Ex: Phylogenetic groups of primate calcanei with 67 genera.

[Turner et al. (2014), *Inf Inference*]

Shape Analysis Using the PHT



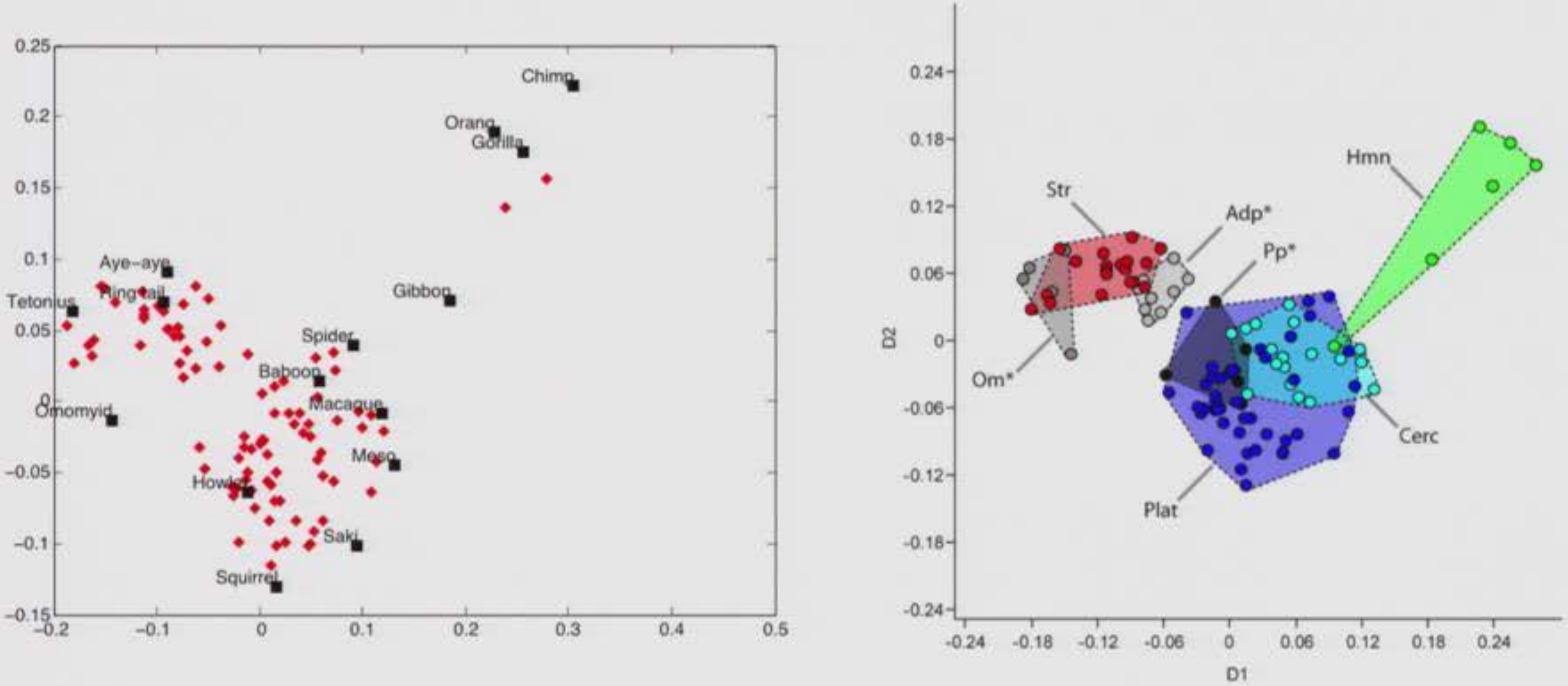
Ex: Phylogenetic groups of primate calcanei with 67 genera.

[Turner et al. (2014), *Inf Inference*]

Disadvantages/Pitfalls of the PHT

- Common regression models use covariates that have an inner product structure defined in Hilbert space.
- The PHT does not admit a simple inner product structure as it is a collection of persistence diagrams.
- **Example:** What is the interpretation of an effect size for an ordered (birth and death time) pair?

Shape Analysis Using the PHT



Ex: Phylogenetic groups of primate calcanei with 67 genera.

[Turner et al. (2014), *Inf Inference*]

Disadvantages/Pitfalls of the PHT

- Common regression models use covariates that have an inner product structure defined in Hilbert space.

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

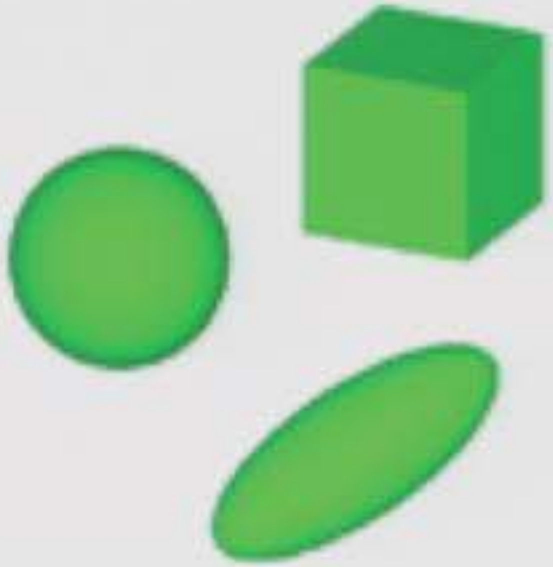
- **Example:** What is the interpretation of an effect size for an ordered (birth and death time) pair?

The Euler Characteristic

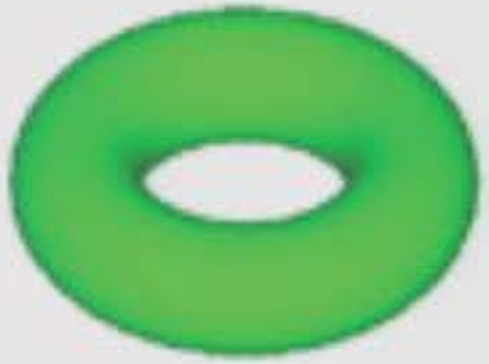
The Euler characteristic (EC) χ for a finite simplicial complex K^d for $d = 3$ is defined by:

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

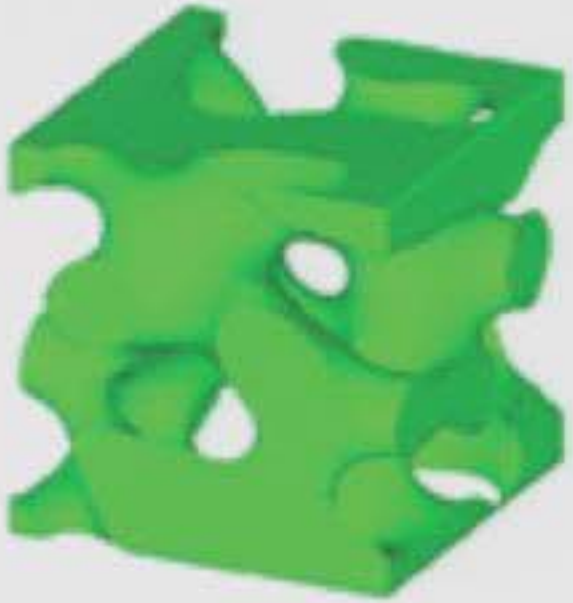
where V, E, and F are the numbers of vertices, edges, and faces, respectively.



χ=2

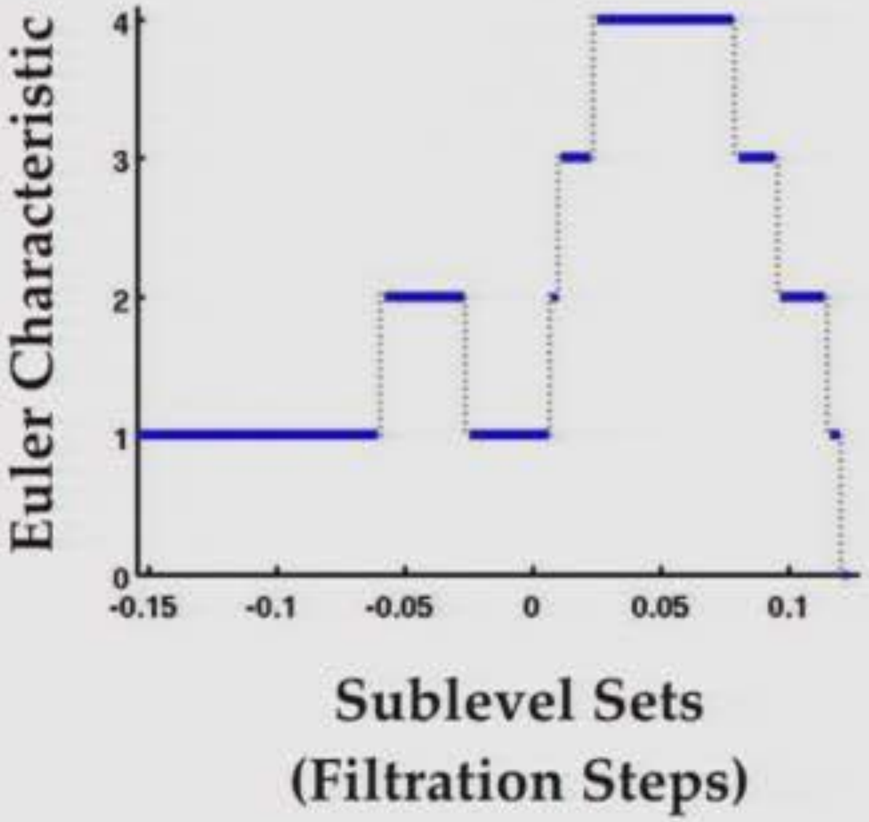


χ=0



χ=−34

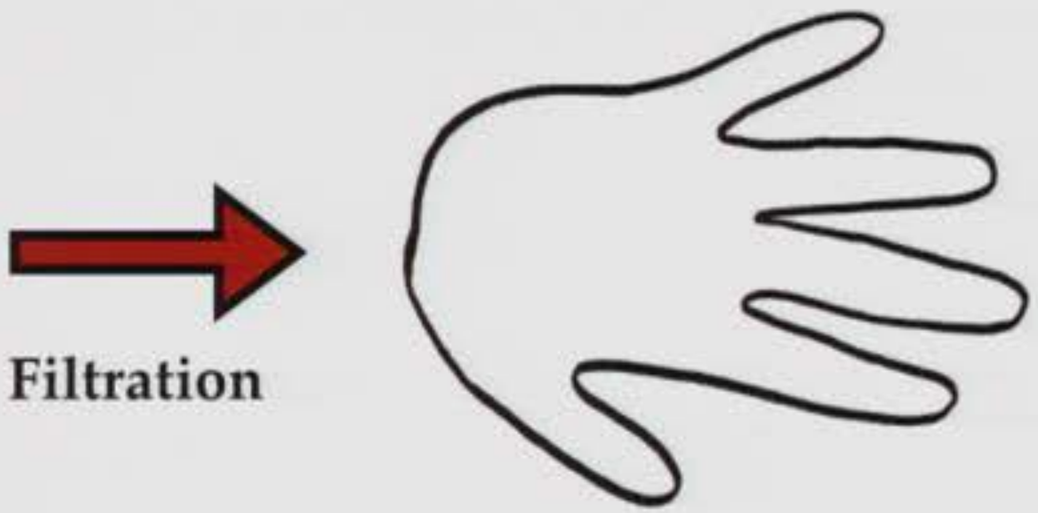
The Euler Characteristic Curve

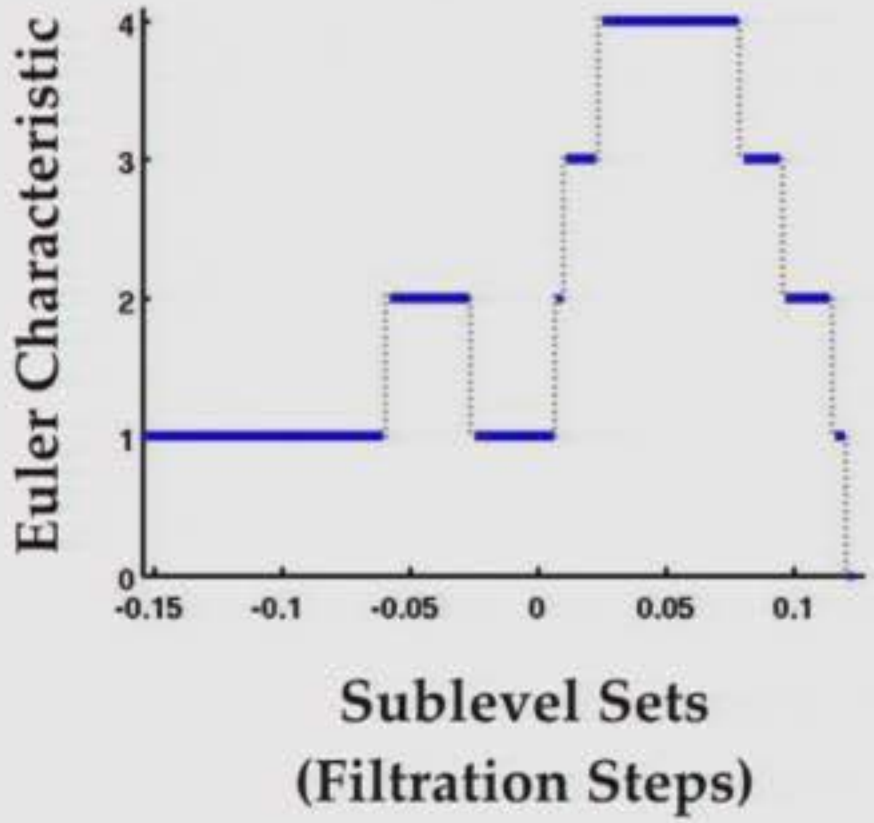


[Turner et al. (2014), *Inf Inference;* Crawford et al. (2020), *JASA*]

The Euler Characteristic Curve

- Concatenate curves over all directions to obtain a vector representation of the shape.





- **End result**: A matrix where each row is the concatenated EC curve of one shape in our dataset.

[Turner et al. (2014), *Inf Inference*; Crawford et al. (2020), *JASA*]

Properties of the Euler Characteristic Transform

- The Euler characteristic transform results in a collection of curves — this represents the topological summary statistic of a 3D shape.

- An EC curve has a simple inner product structure.

- Allows for quantitative comparisons using the full scope of parametric and nonparametric regression methodology.

Application to Radiogenomics

Predicting Clinical Outcomes in Radiomics

- Magnetic resonance images (MRIs) of primary glioblastoma multiforme (GBM) tumors were collected from ~40 patients

- Data archived by the The Cancer Imaging Archive (TCIA)

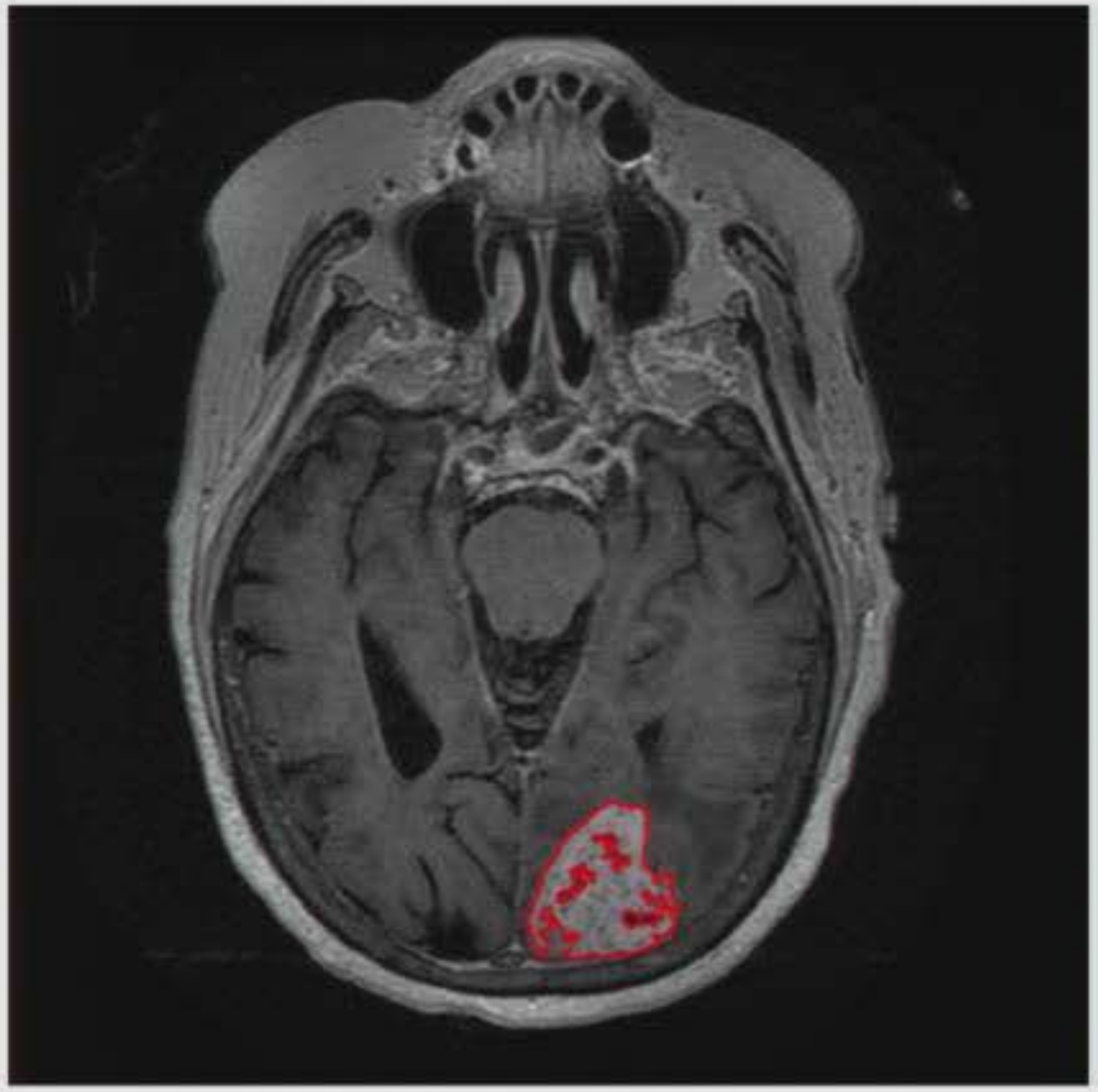
Predicting Clinical Outcomes in Radiomics

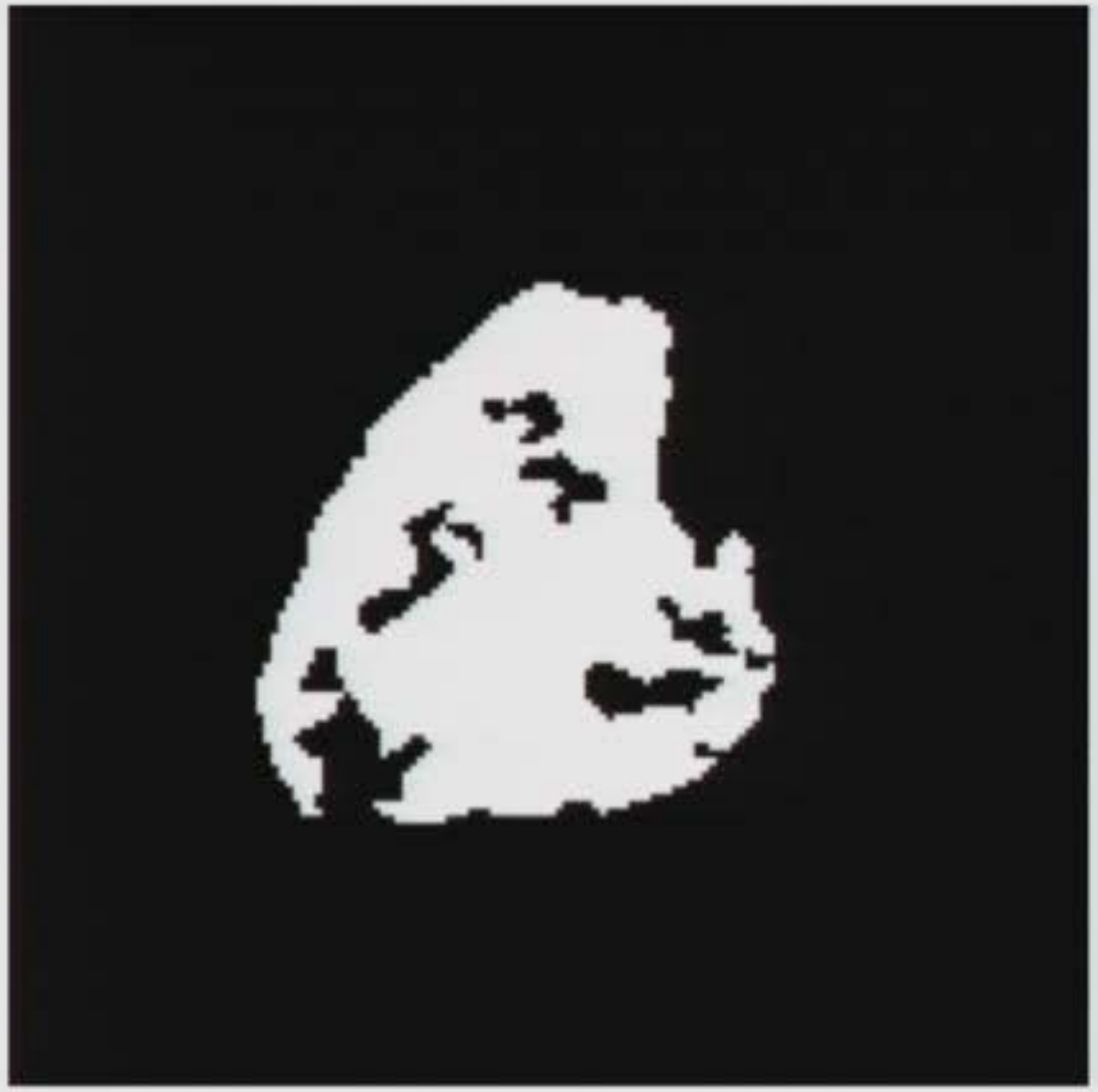
- Magnetic resonance images (MRIs) of primary glioblastoma multiforme (GBM) tumors were collected from ~40 patients

- Data archived by the The Cancer Imaging Archive (TCIA)

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

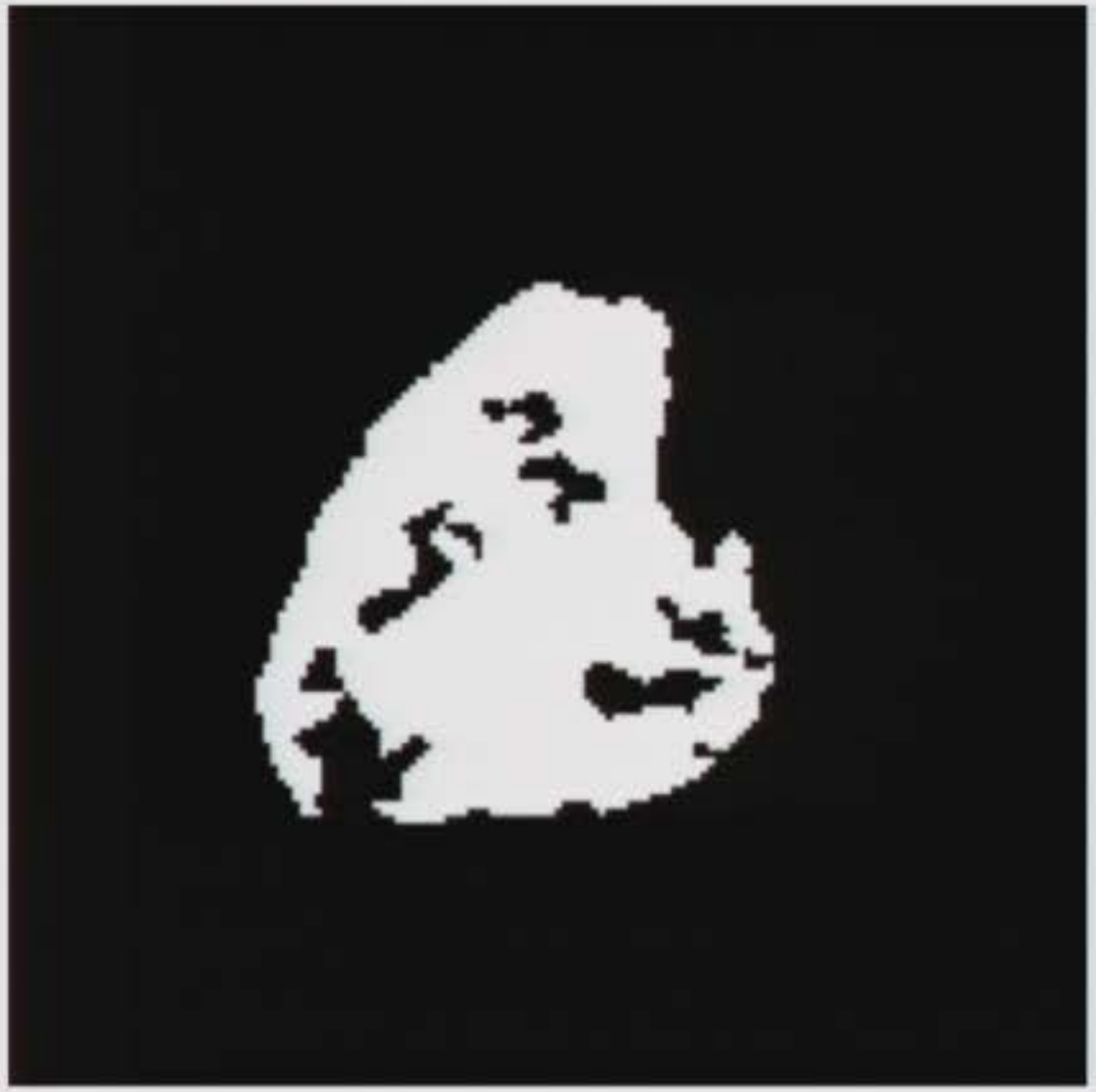
Application to Glioblastoma Multiforme

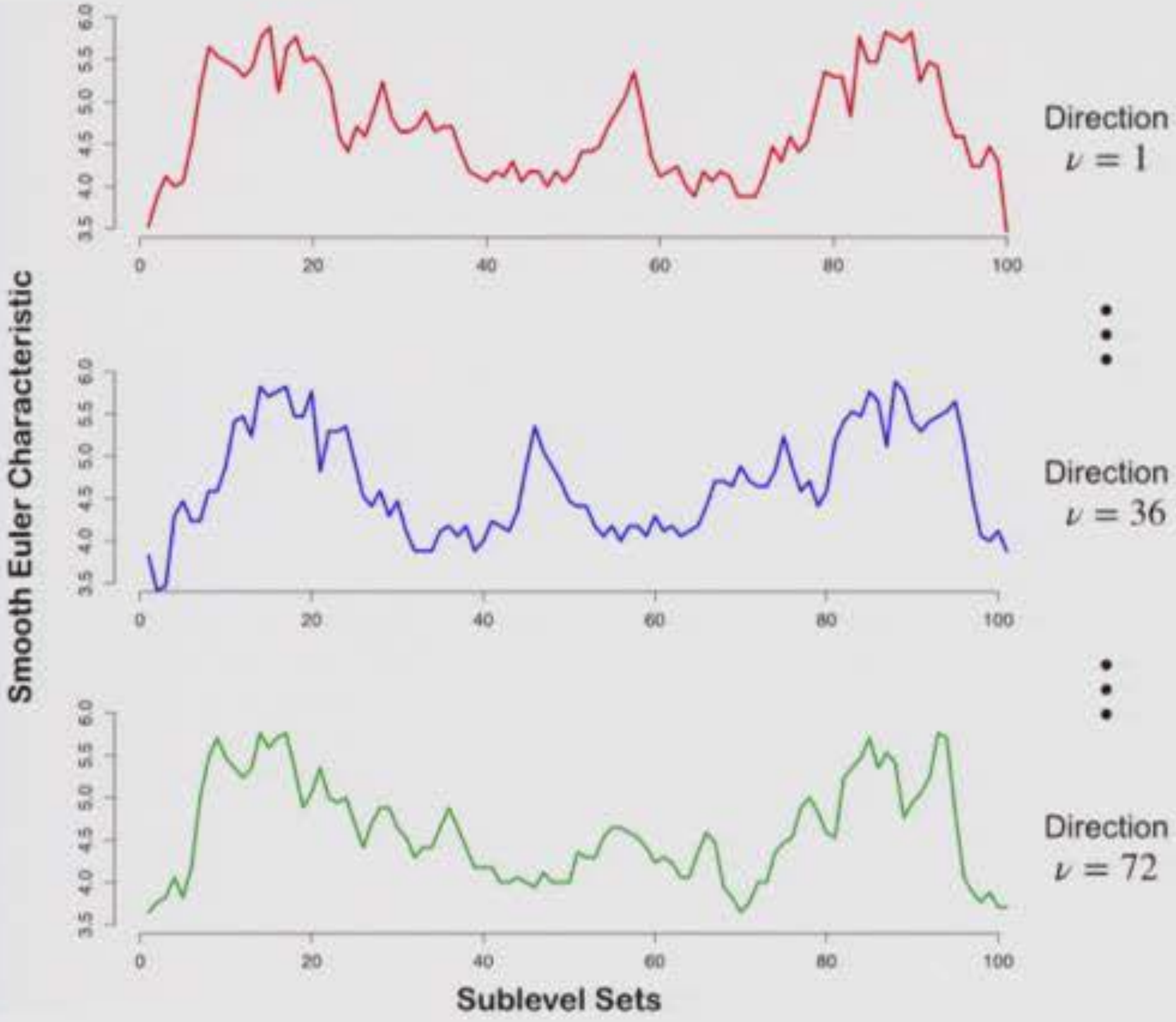




[Crawford et al. (2020), *JASA*]

Application to Glioblastoma Multiforme





[Crawford et al. (2020), *JASA*]

Nonlinear Regression Methods

Nonlinear models perform better for phenotypic prediction

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

Nonlinear Regression Methods

Nonlinear models perform better for phenotypic prediction

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

Gaussian processes specify prior distribution over the function space directly

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')),$$

where:

$$\mathbf{K} = \begin{pmatrix} k(\mathrm{x}_1, \mathrm{x}_1) & k(\mathrm{x}_1, \mathrm{x}_2) & \cdots & k(\mathrm{x}_1, \mathrm{x}_n) \\ k(\mathrm{x}_2, \mathrm{x}_1) & k(\mathrm{x}_2, \mathrm{x}_2) & \cdots & k(\mathrm{x}_2, \mathrm{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathrm{x}_n, \mathrm{x}_1) & k(\mathrm{x}_n, \mathrm{x}_2) & \cdots & k(\mathrm{x}_n, \mathrm{x}_n) \end{pmatrix}$$

Predicting Clinical Outcomes in Radiogenomics

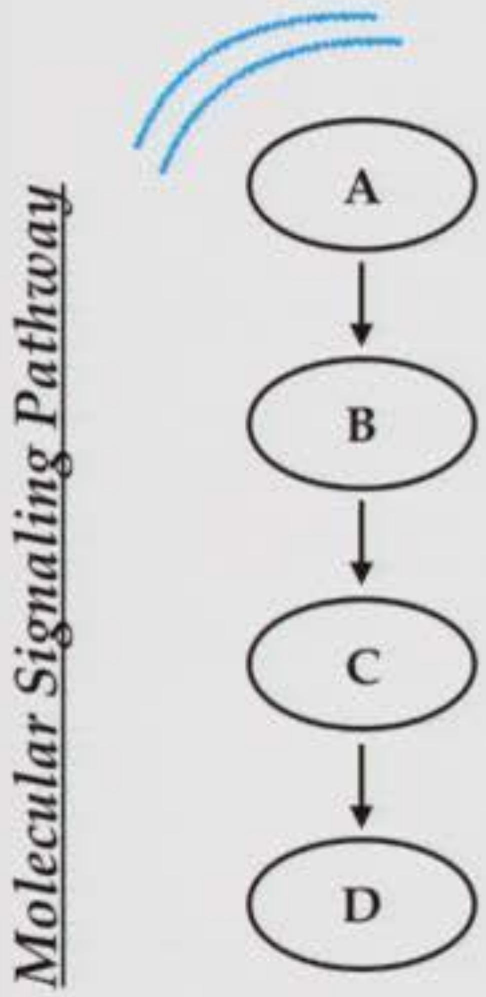
- Compare ECs with three key types of tumor characteristics:
  - mRNA Gene Expression Measurements
  - Tumor Morphometry
  - Tumor Volume and Geometrics
- Predict two clinical outcomes:
  - **Disease Free Survival (DFS)**
  - **Overall Survival (OS)**
- Perform 80-20 (in/out of sample) splits; 100 times
- **Predictive Measure:** Root Mean Square Error of Prediction (RMSEP)

Prediction Results

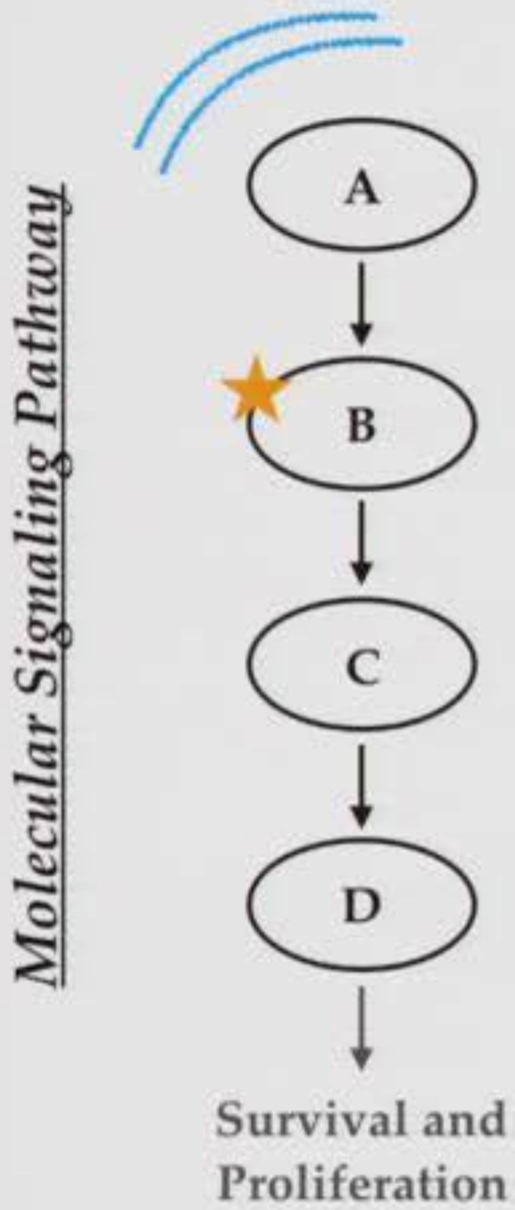
	Disease Free Survival		*Overall Survival*	
Data Type	RMSEP	Pr[Optimal]	RMSEP	Pr[Optimal]
Gene Expression	0.944 (0.035)	0.20	0.981 (0.030)	0.27
Morphometrics	0.942 (0.035)	0.07	0.965 (0.029)	0.15
Volume	0.939 (0.035)	0.06	0.964 (0.029)	0.16
SECT	**0.803 (0.035)**	**0.69**	**0.958 (0.028)**	**0.42**

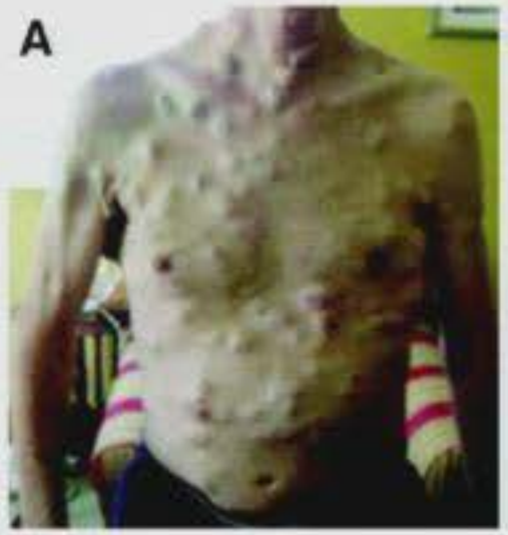
Average RMSPE across both clinical outcomes. The number in parenthesis is the standard error due to random sampling

Oncogene Activity and Therapeutic Resistance

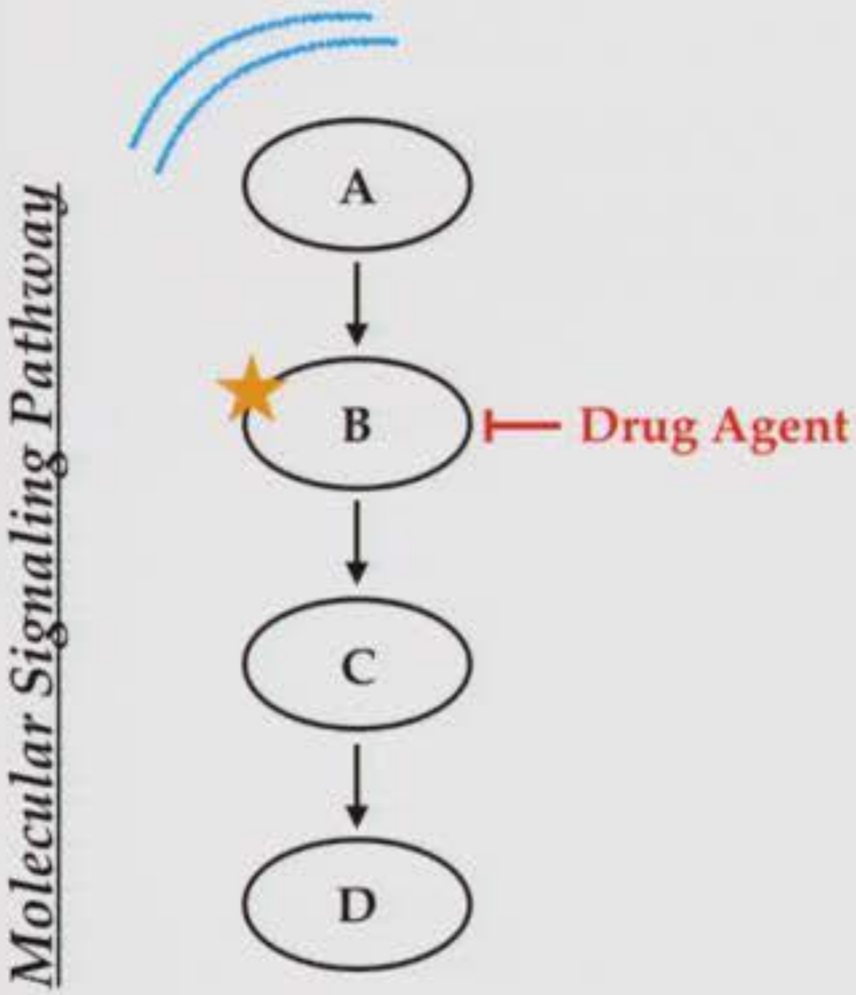


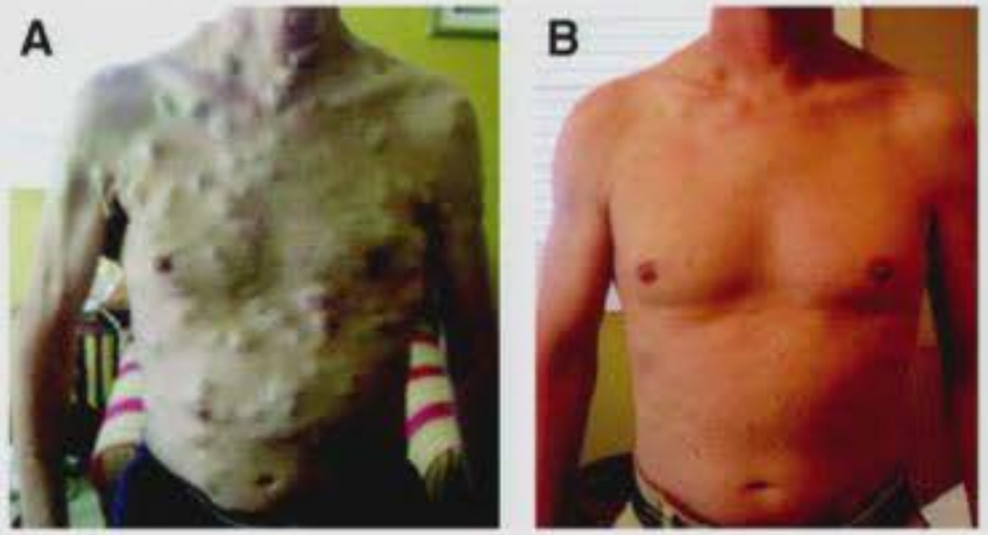
Oncogene Activity and Therapeutic Resistance



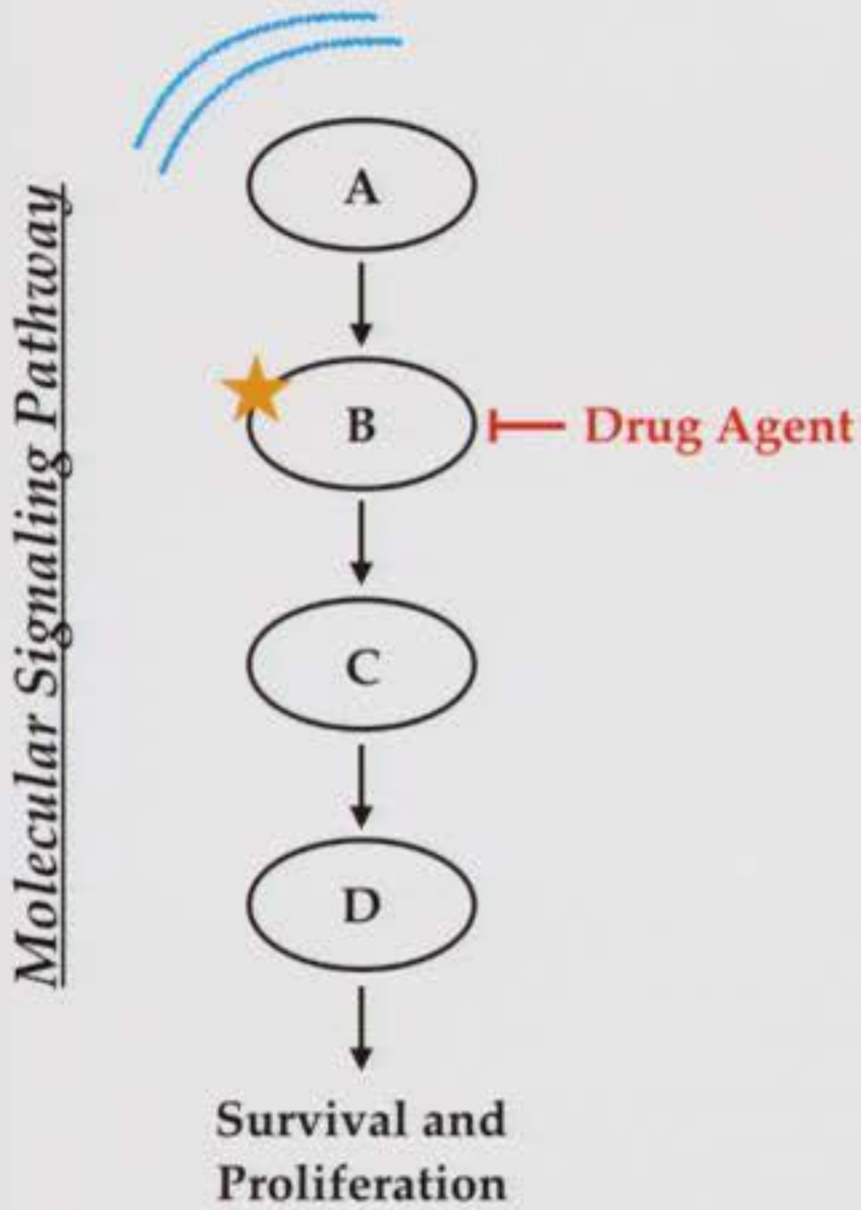


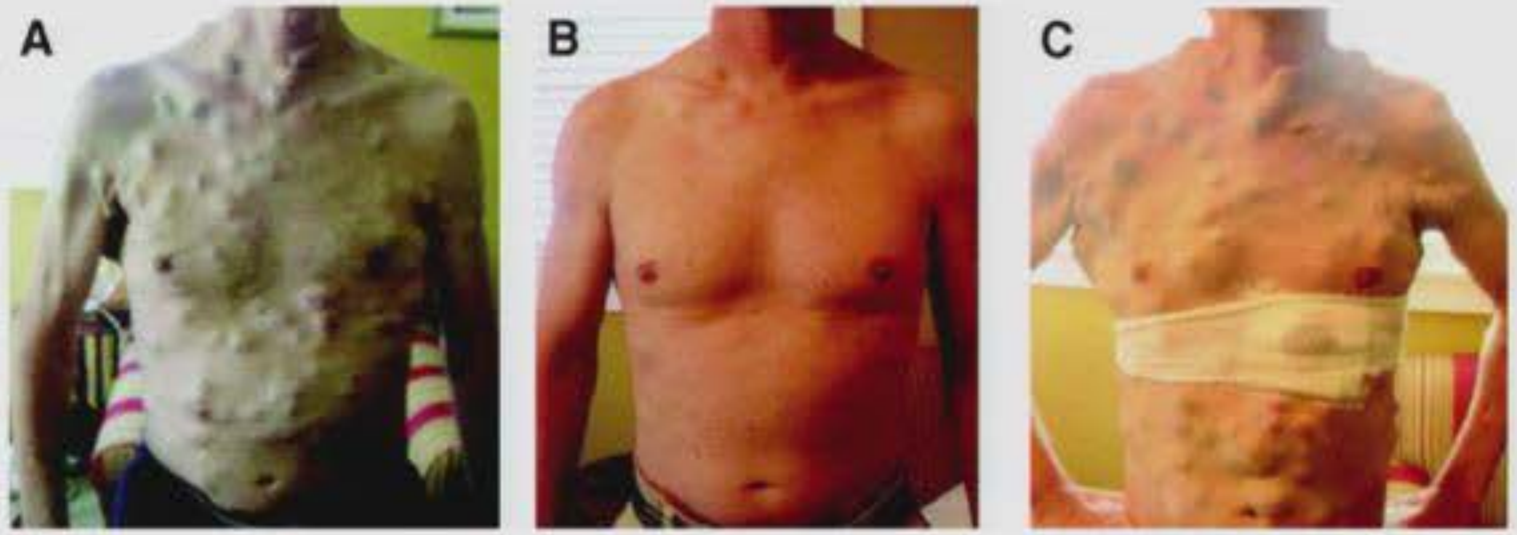
Oncogene Activity and Therapeutic Resistance



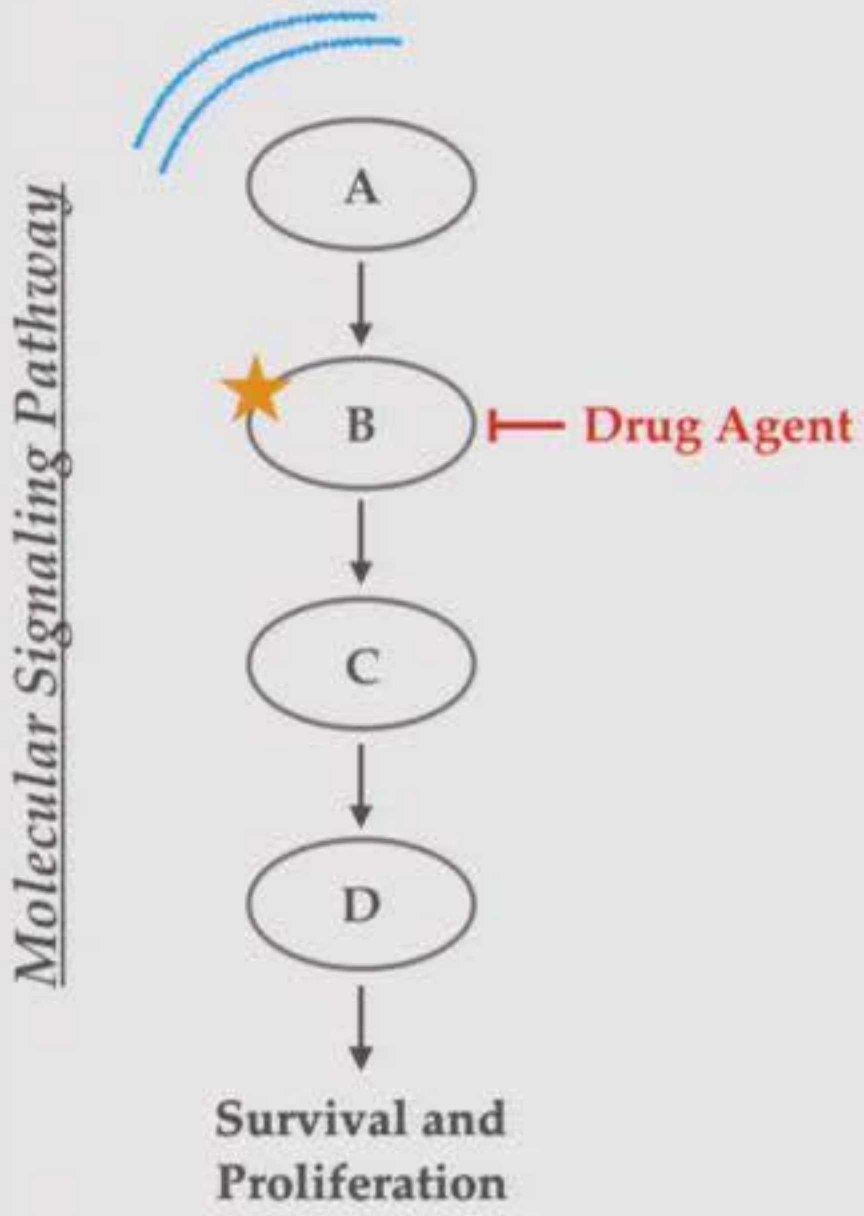


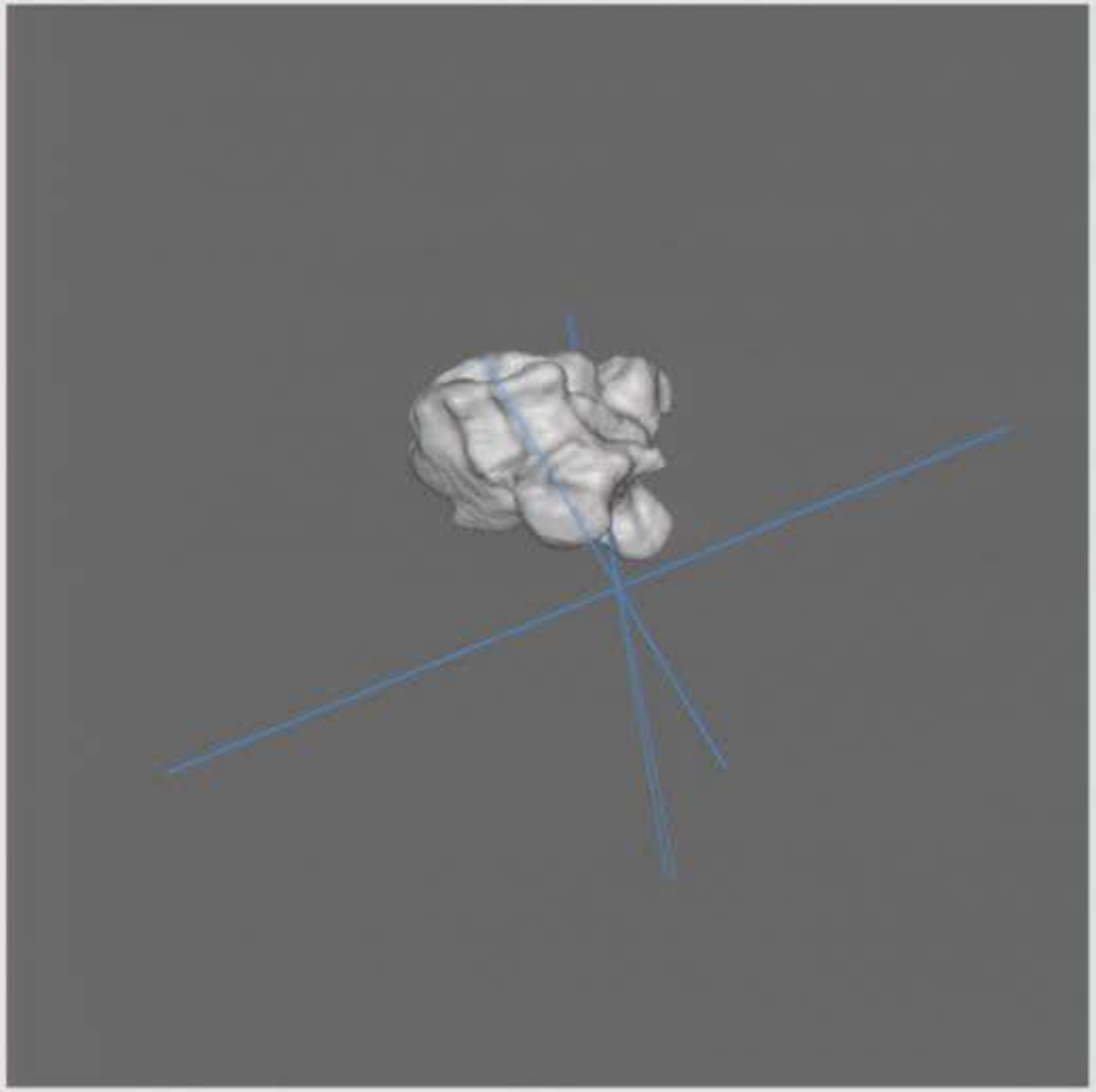
Oncogene Activity and Therapeutic Resistance





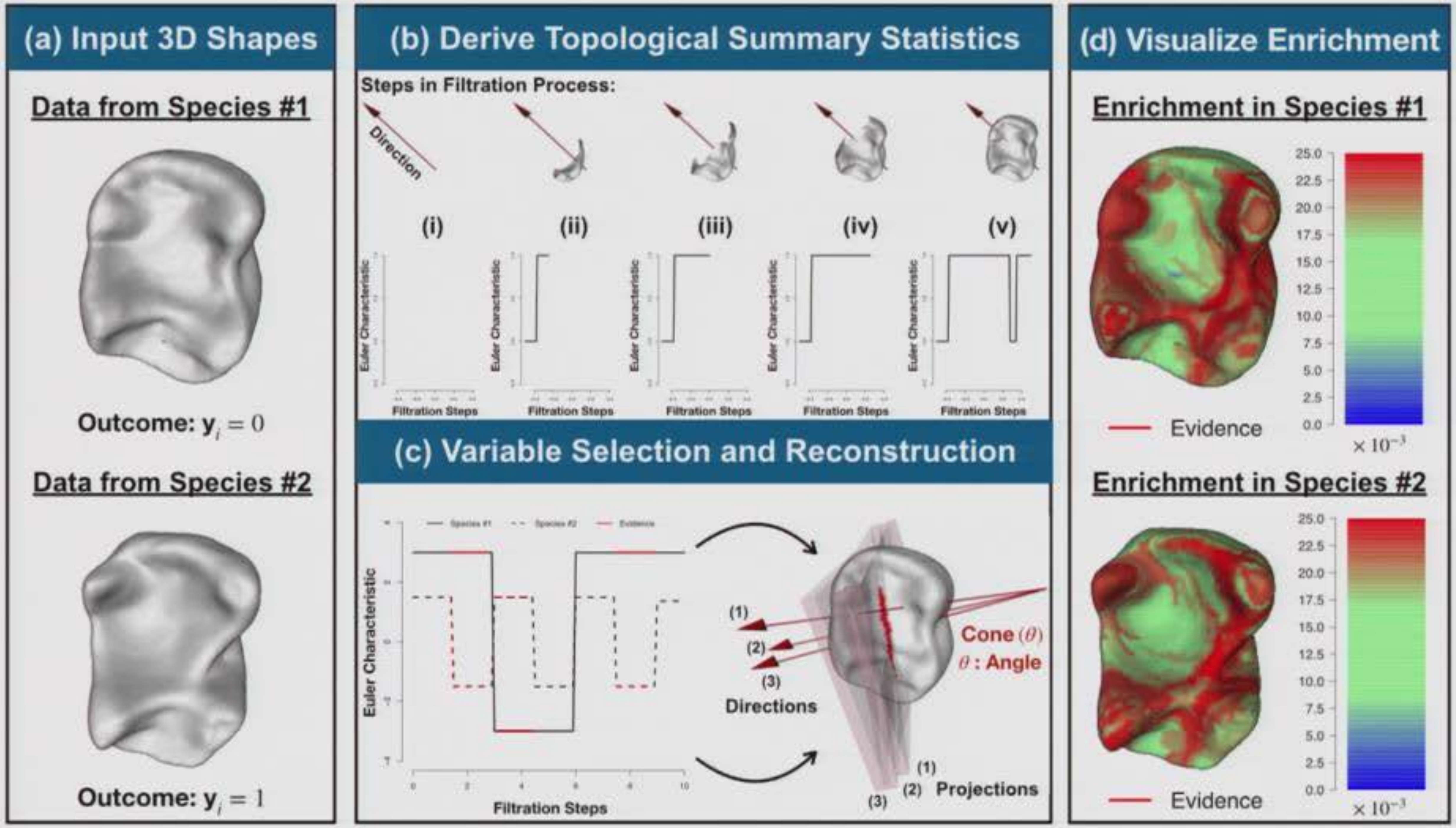
Shape Variation to Explain Biological Phenomena





Sub-Image Analysis using Topological Summary Statistics (SINATRA)

The SINATRA Pipeline



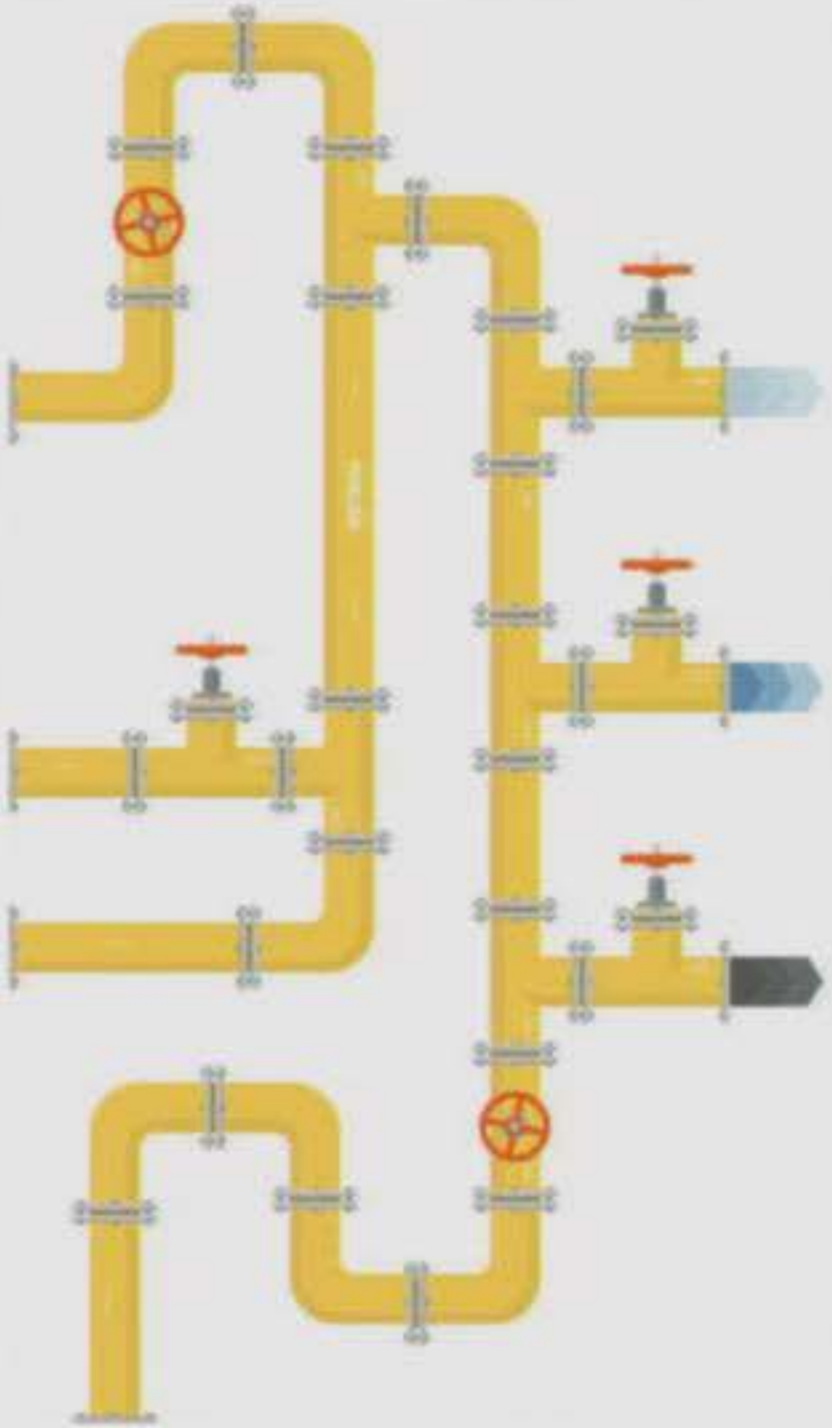
(a) Input 3D Shapes

(b) Derive Topological Summary Statistics

(c) Variable Selection and Reconstruction

(d) Visualize Enrichment

General Steps in the SINATRA Pipeline



- □ Represent shapes via statistics summarizing their topology / geometry;
- □ Use a statistical model and classify shapes based on these summary statistics;
- □ Derive an "evidence of association" metric for each topological / geometric feature;
- □ Project these association measures back onto the original shape.

Revisiting the Gaussian Process

Nonlinear models perform better for phenotypic prediction

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

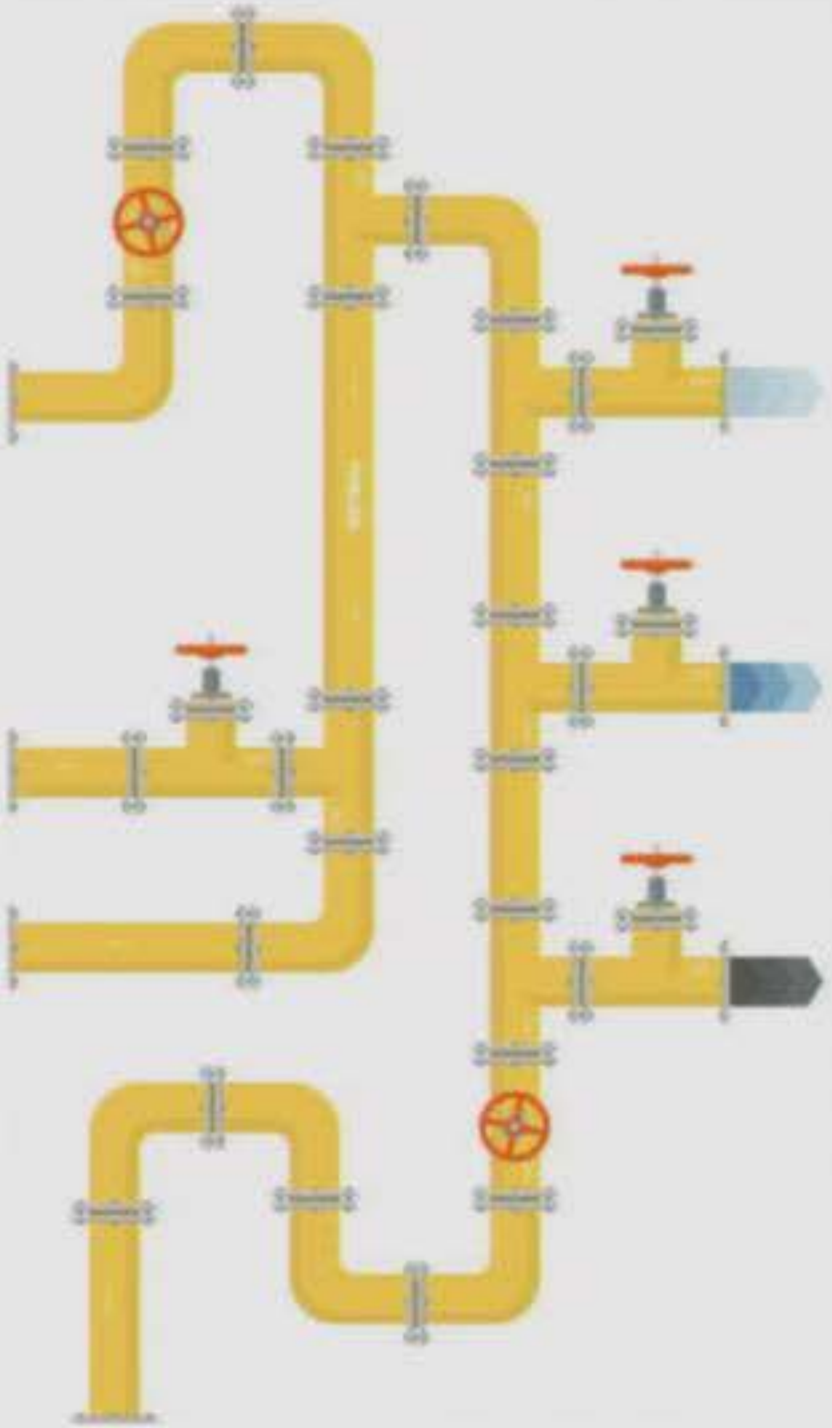
Gaussian processes specify prior distribution over the function space directly

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')),$$

where:

$$\mathbf{K} = \begin{pmatrix} k(\mathrm{x}_1, \mathrm{x}_1) & k(\mathrm{x}_1, \mathrm{x}_2) & \cdots & k(\mathrm{x}_1, \mathrm{x}_n) \\ k(\mathrm{x}_2, \mathrm{x}_1) & k(\mathrm{x}_2, \mathrm{x}_2) & \cdots & k(\mathrm{x}_2, \mathrm{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathrm{x}_n, \mathrm{x}_1) & k(\mathrm{x}_n, \mathrm{x}_2) & \cdots & k(\mathrm{x}_n, \mathrm{x}_n) \end{pmatrix}$$

General Steps in the SINATRA Pipeline



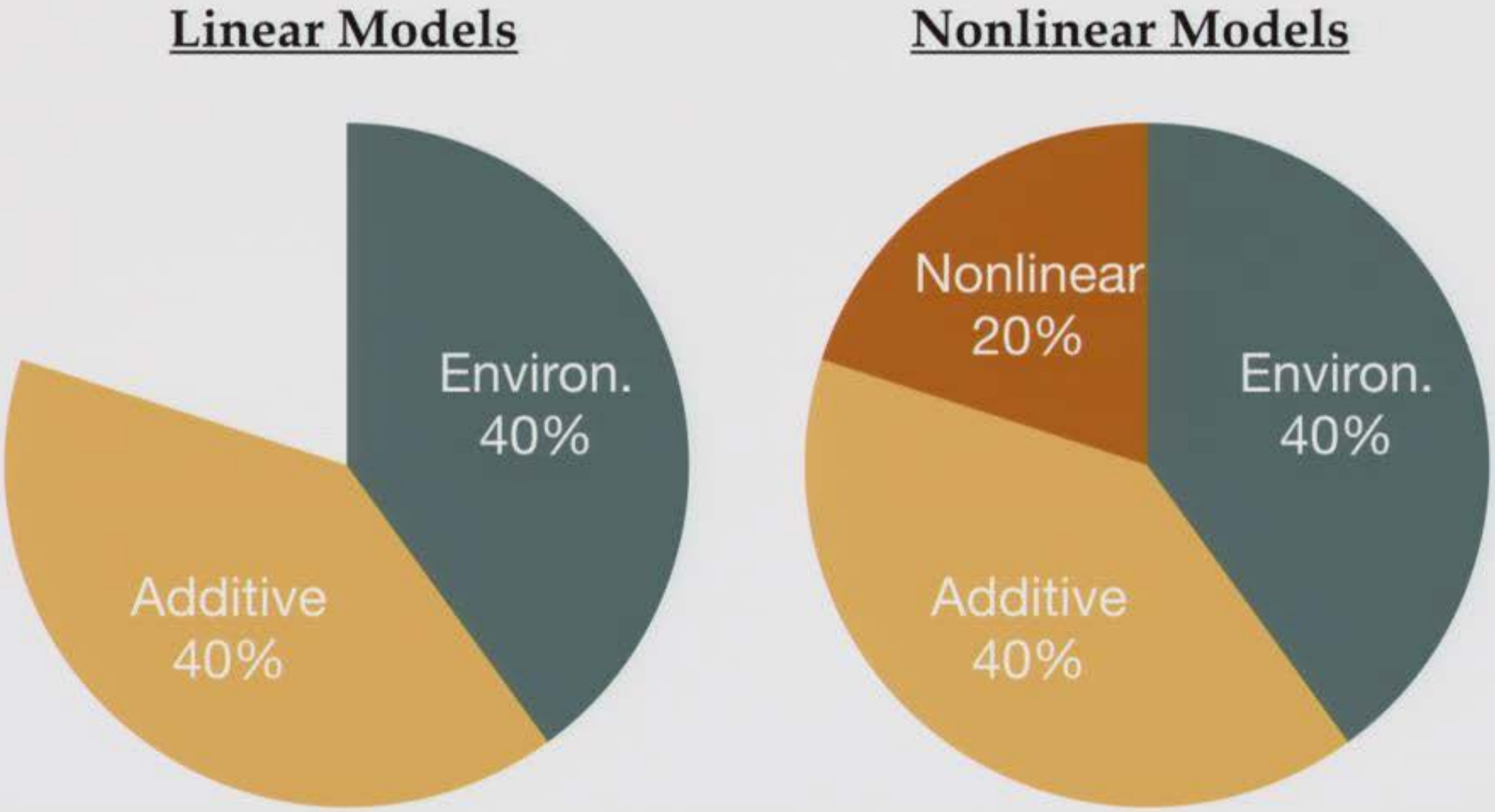
☑ Represent shapes via statistics summarizing their topology / geometry;

☑ Use a statistical model and classify shapes based on these summary statistics;

□ Derive an "evidence of association" metric for each topological / geometric feature;

□ Project these association measures back onto the original shape.

Linear vs. Nonlinear Models



The “Kernel Trick” Issue

original p-dimensional space

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

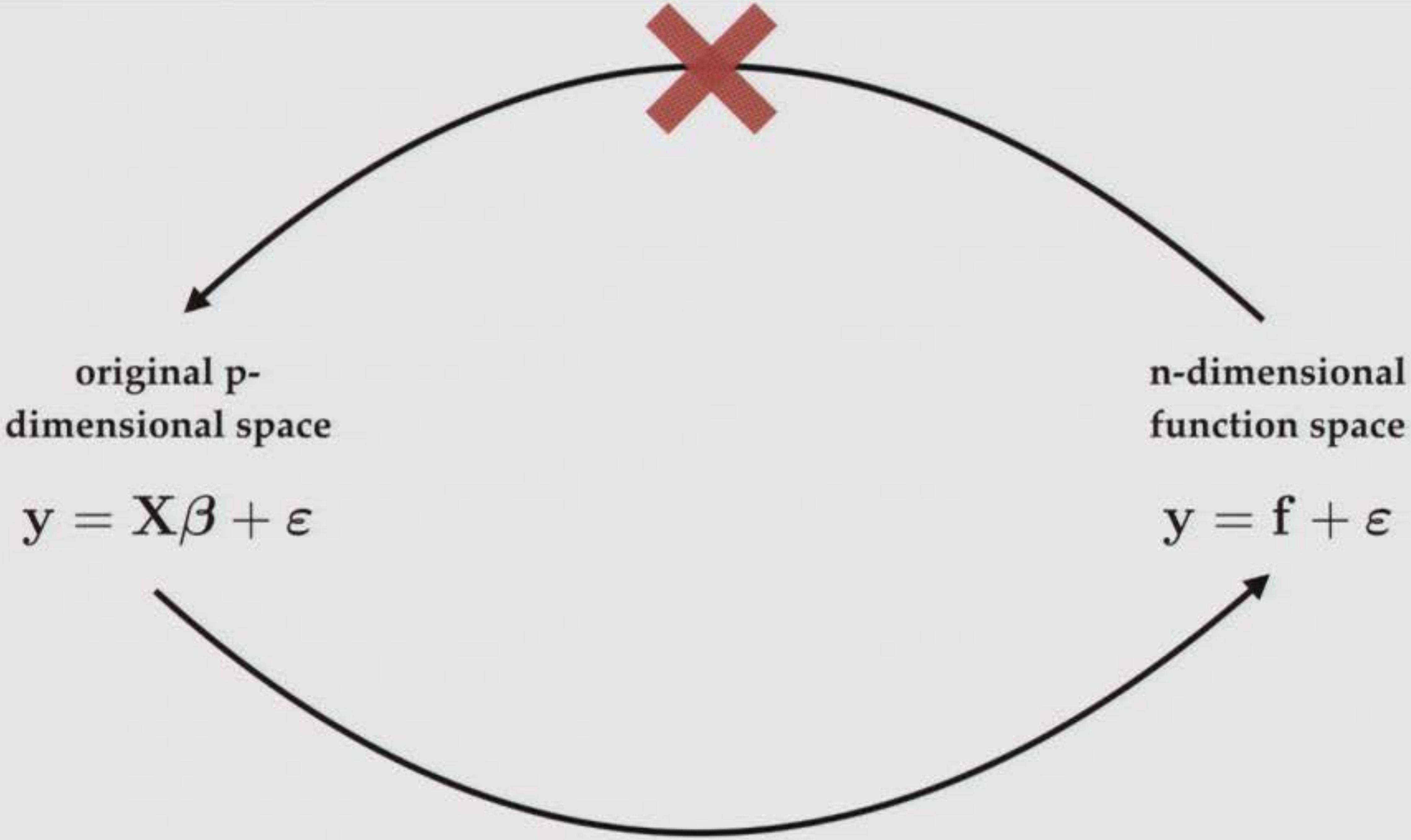
The “Kernel Trick” Issue

original p-dimensional space

n-dimensional function space

$$\begin{tikzcd} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \arrow[r, bend right] & \mathbf{y} = \mathbf{f} + \varepsilon \end{tikzcd}$$

The “Kernel Trick” Issue



original p-dimensional space

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

n-dimensional function space

$$\mathbf{y} = \mathbf{f} + \varepsilon$$

The Effect Size Analog

Linear Models

Nonlinear Models

The Effect Size Analog

Linear Models

- A regression model is takes the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- An **effect size** is the linear projection onto the phenotype:

$$\widehat{\boldsymbol{\beta}} = \mathrm{Proj}(\mathbf{X}, \mathbf{y})$$

- One standard projection operation is uses generalized inverses:

$$\mathrm{Proj}(\mathbf{X}, \mathbf{y}) = \mathbf{X}^{\dagger}\mathbf{y}$$

Nonlinear Models

The Effect Size Analog

Linear Models

- A regression model is takes the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- An **effect size** is the linear projection onto the phenotype:

$$\widehat{\boldsymbol{\beta}} = \mathrm{Proj}(\mathbf{X}, \mathbf{y})$$

- One standard projection operation is uses generalized inverses:

$$\mathrm{Proj}(\mathbf{X}, \mathbf{y}) = \mathbf{X}^{\dagger}\mathbf{y}$$

Nonlinear Models

- A regression model is takes the form:

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}$$

- An **effect size analog** is the projection onto the smooth nonlinear function:

$$\widetilde{\boldsymbol{\beta}} = \mathrm{Proj}(\mathbf{X}, \mathbf{f})$$

- We ***can*** use the same standard projection operations:

$$\mathrm{Proj}(\mathbf{X}, \mathbf{f}) = \mathbf{X}^{\dagger}\mathbf{f}$$

Posterior Inference and Sampling

Assume we have completely specified hierarchical model

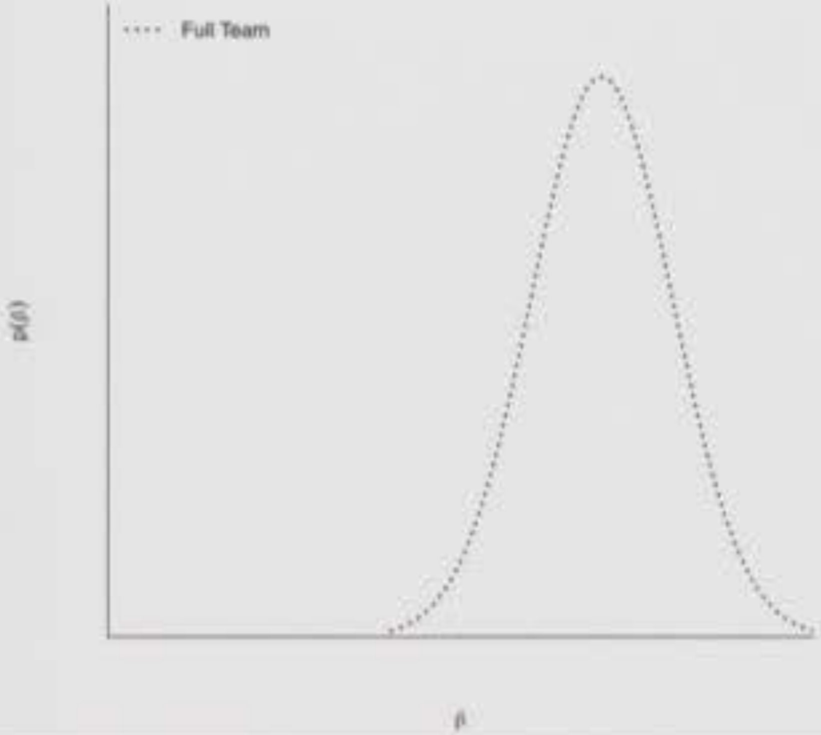
$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}, \quad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}), \quad \tau^2 \sim \text{Scale-Inv-}\chi^2(a, b).$$

MCMC for this regression model includes:

Illustration: Ranking Influential Players

Illustration: Ranking Influential Players

Illustration: Ranking Influential Players



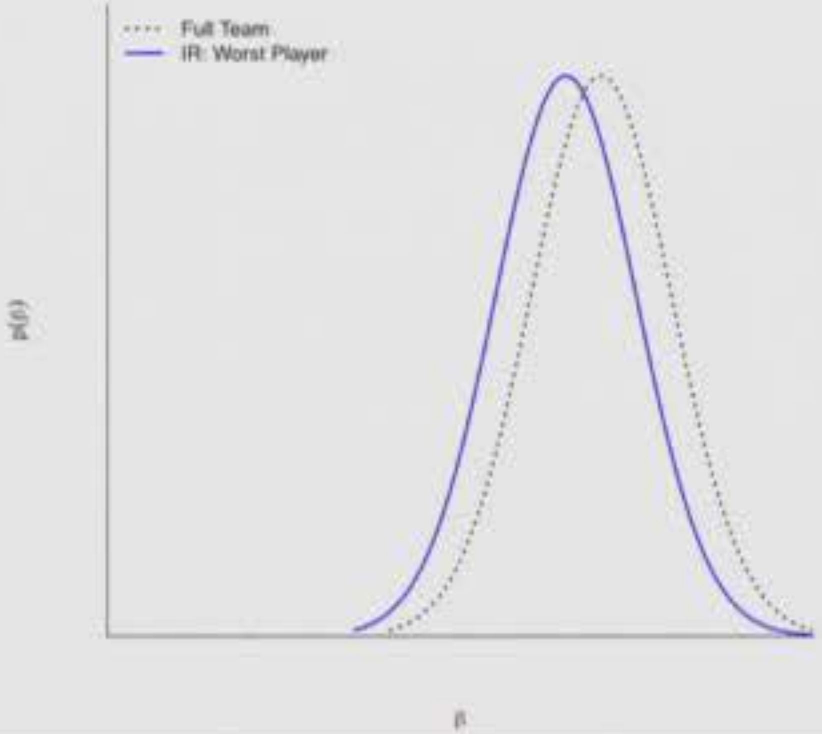
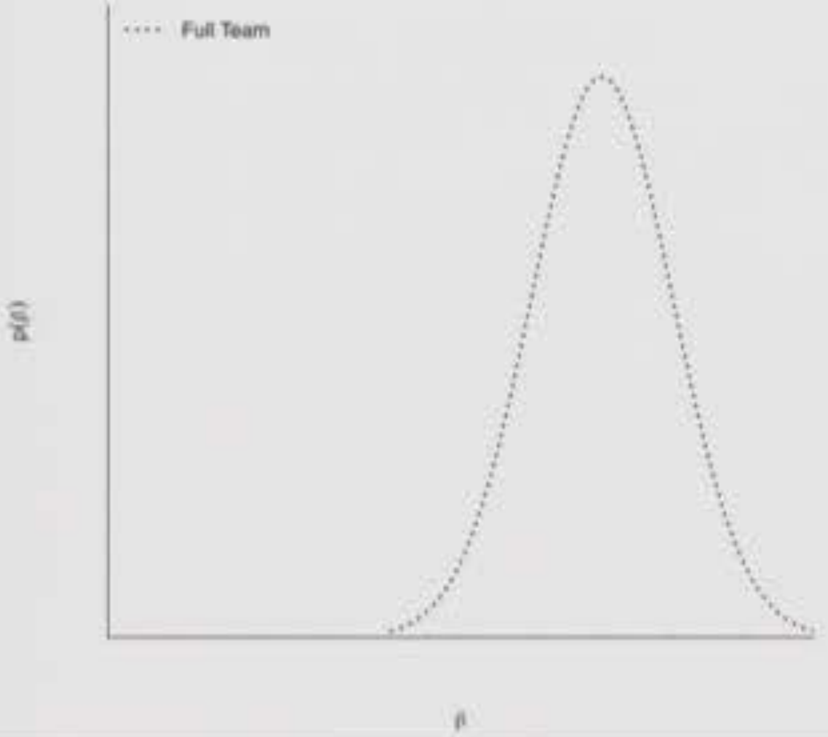
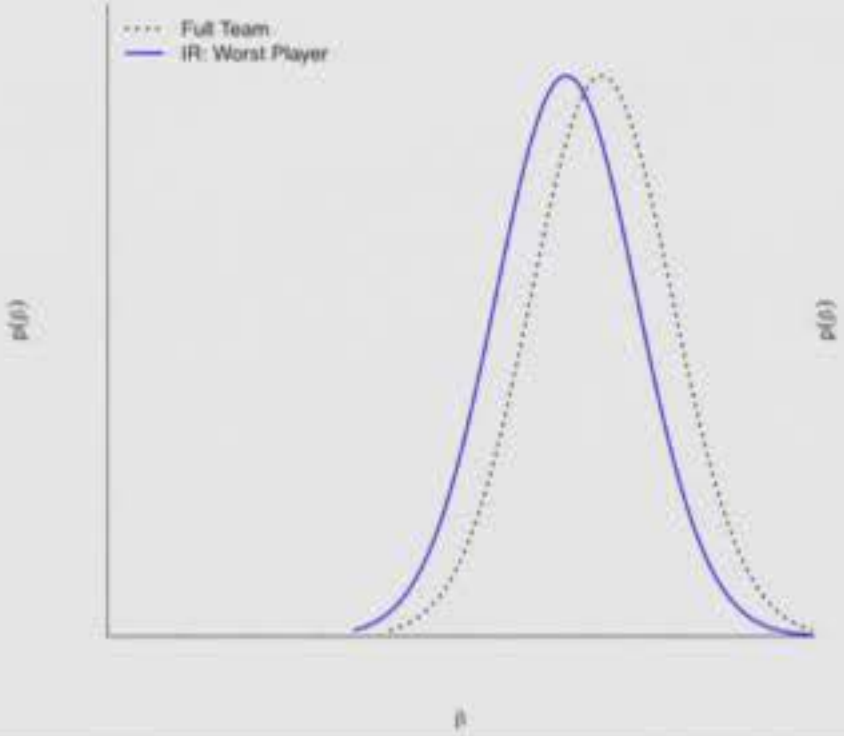
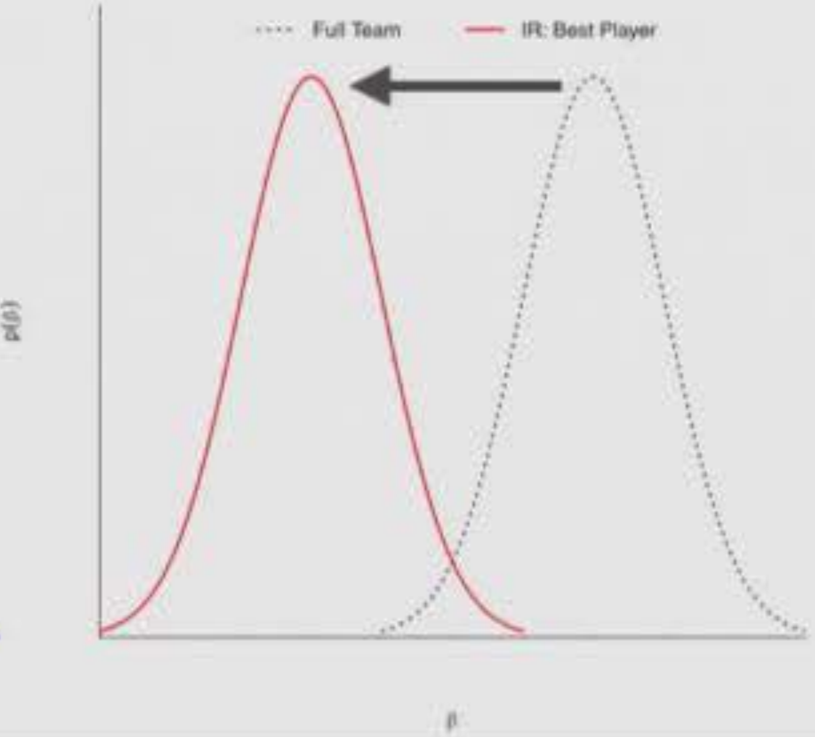


Illustration: Ranking Influential Players







Kullback-Leibler Divergence (KLD)

Summarize the influence of the variant $\mathbf{x}_j$ on the rest of the variants in $\mathbf{X}_{-j}$ via the KLD measuring the difference between $p(\boldsymbol{\beta}_{-j} \mid \beta_j)$ and $p(\boldsymbol{\beta}_{-j})$. Namely,

Kullback-Leibler Divergence (KLD)

Summarize the influence of the variant $\mathbf{x}_j$ on the rest of the variants in $\mathbf{X}_{-j}$ via the KLD measuring the difference between $p(\boldsymbol{\beta}_{-j} \mid \beta_j)$ and $p(\boldsymbol{\beta}_{-j})$. Namely,

$$\mathrm{KLD}(\beta_j) = \int_{\boldsymbol{\beta}_{-j}} \log\left(\frac{p(\boldsymbol{\beta}_{-j})}{p(\boldsymbol{\beta}_{-j} \mid \beta_j)}\right) p(\boldsymbol{\beta}_{-j})\, \mathrm{d}\boldsymbol{\beta}_{-j}.$$

where $\mathrm{KLD}(\beta_j) \in [0, \infty)$.

Here, $\mathrm{KLD}(\beta_j) = 0$ is interpreted as variant j not being a key explanatory variable relative to others.

Or alternatively, $\mathrm{KLD}(\beta_j) = 0$ if and only if $p(\boldsymbol{\beta}_{-j} \mid \beta_j) = p(\boldsymbol{\beta}_{-j})$.

RelATive cEntrality (RATE) Measures

One natural way for determining significance is to explore a variable's "RelATive cEntrality" or RATE

$$\text{RATE}(\beta_j) = \text{KLD}(\beta_j) / \sum \text{KLD}(\beta_\ell), \quad \sum \text{RATE}(\beta_j) = 1.$$

[Crawford et al. (2019), *AoAS*]

RelATive cEntrality (RATE) Measures

One natural way for determining significance is to explore a variable's "RelATive cEntrality" or RATE

$$\text{RATE}(\beta_j) = \text{KLD}(\beta_j)/\sum \text{KLD}(\beta_\ell), \quad \sum \text{RATE}(\beta_j) = 1.$$

A set of significant markers then have RATEs satisfying

$$\{j : \text{RATE}(\beta_j) > 1/p\}\,.$$

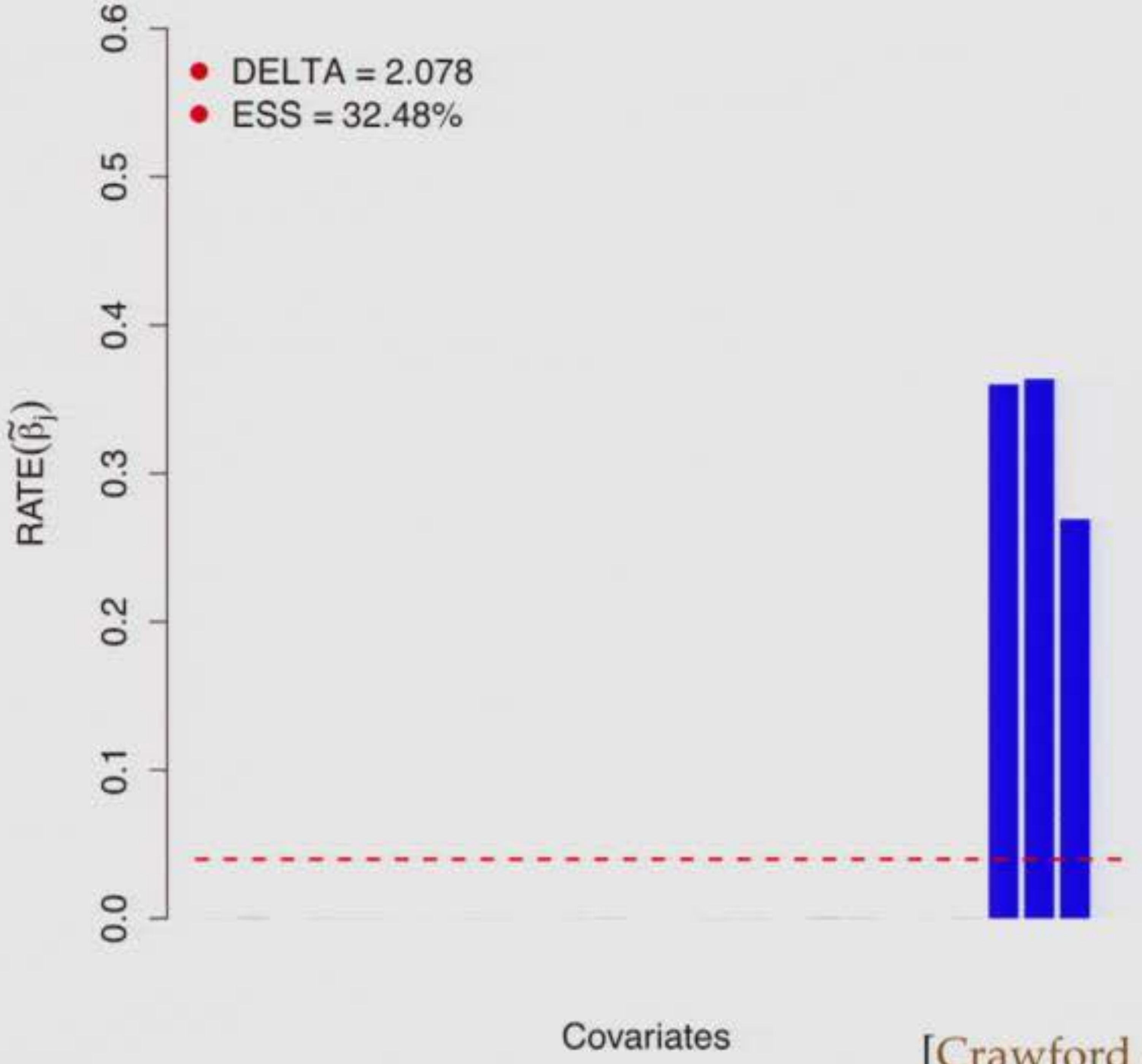
where $1/p$ represents the level at which there is relative equal importance across all variants.

[Crawford et al. (2019), *AoAS*]

RATE Example: Proof-of-Concept

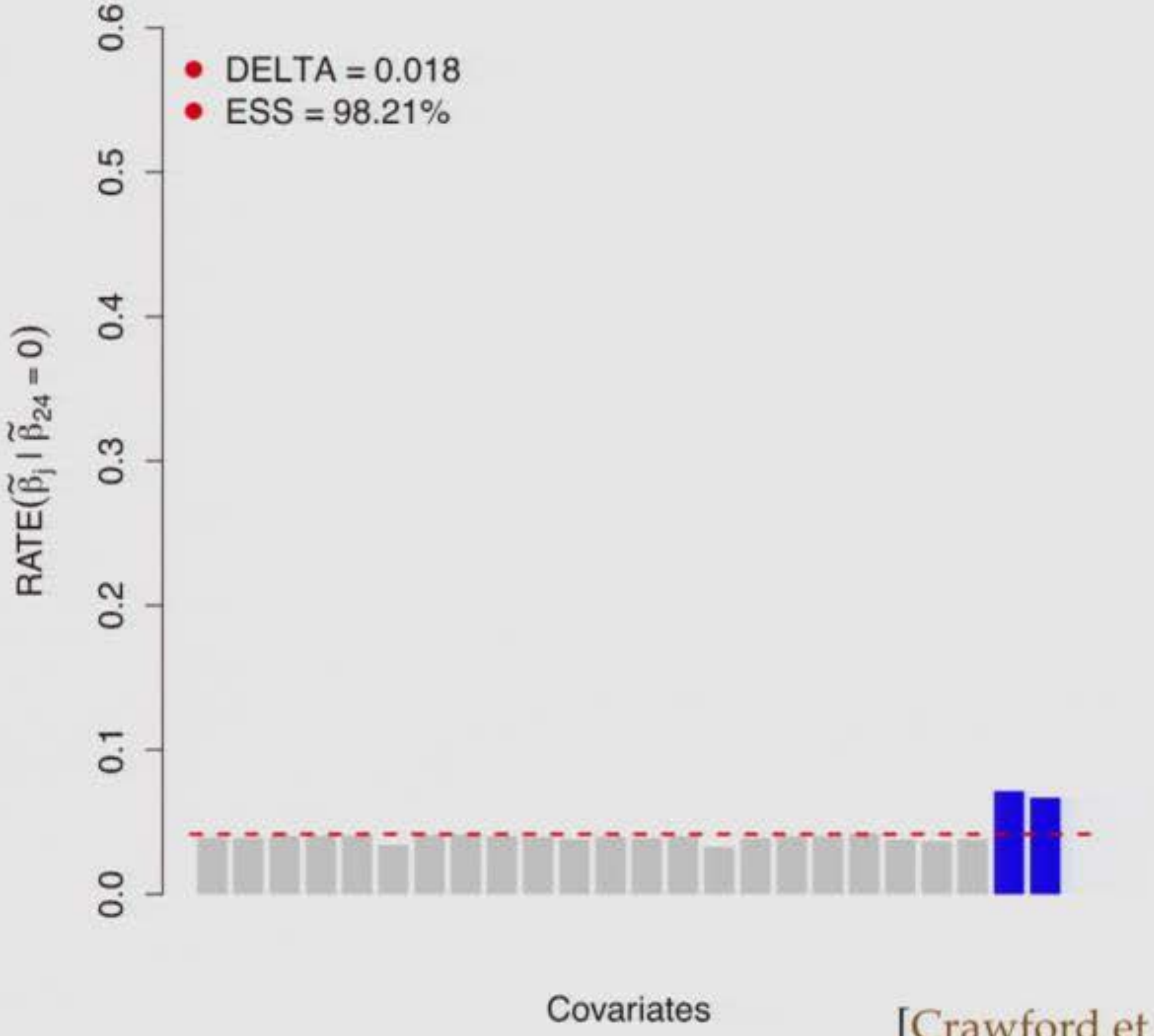
- Simulate datasets with $n = 2000$ samples and $p = 25$ features.
- Choose the last three features $j^* = \{23, 24, 25\}$ to be causal.
- Consider the following scenario to simulate phenotypes:
  - **All j^* variants have additive effects;**
  - **There is an interaction between these variables;**
  - **Interaction effects makes up 50% of the phenotypic variance.**
- Perform association mapping using RATE.

RATE Example: Proof-of-Concept



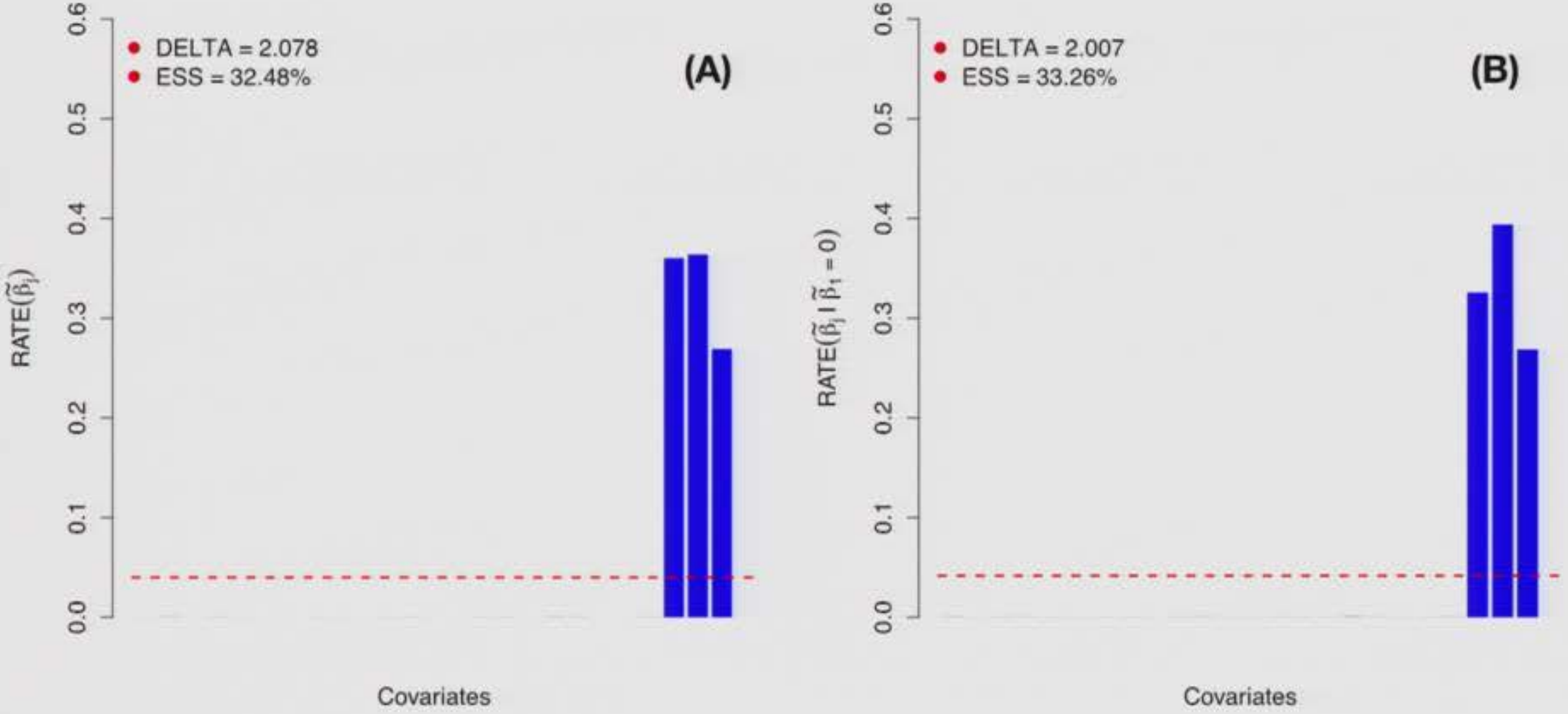
[Crawford et al. (2019), *AoAS*]

RATE Example: Proof-of-Concept



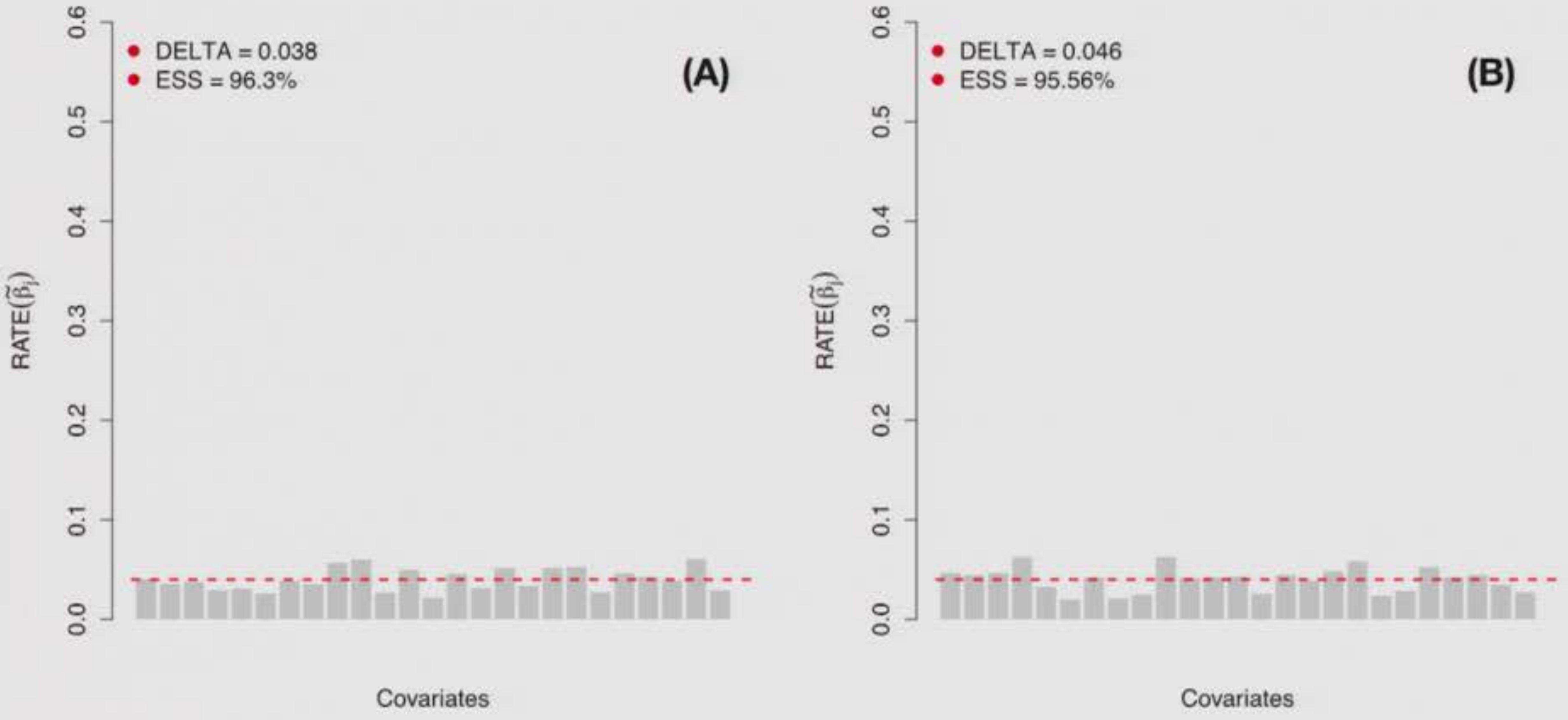
[Crawford et al. (2019), *AoAS*]

RATE Example: Proof-of-Concept



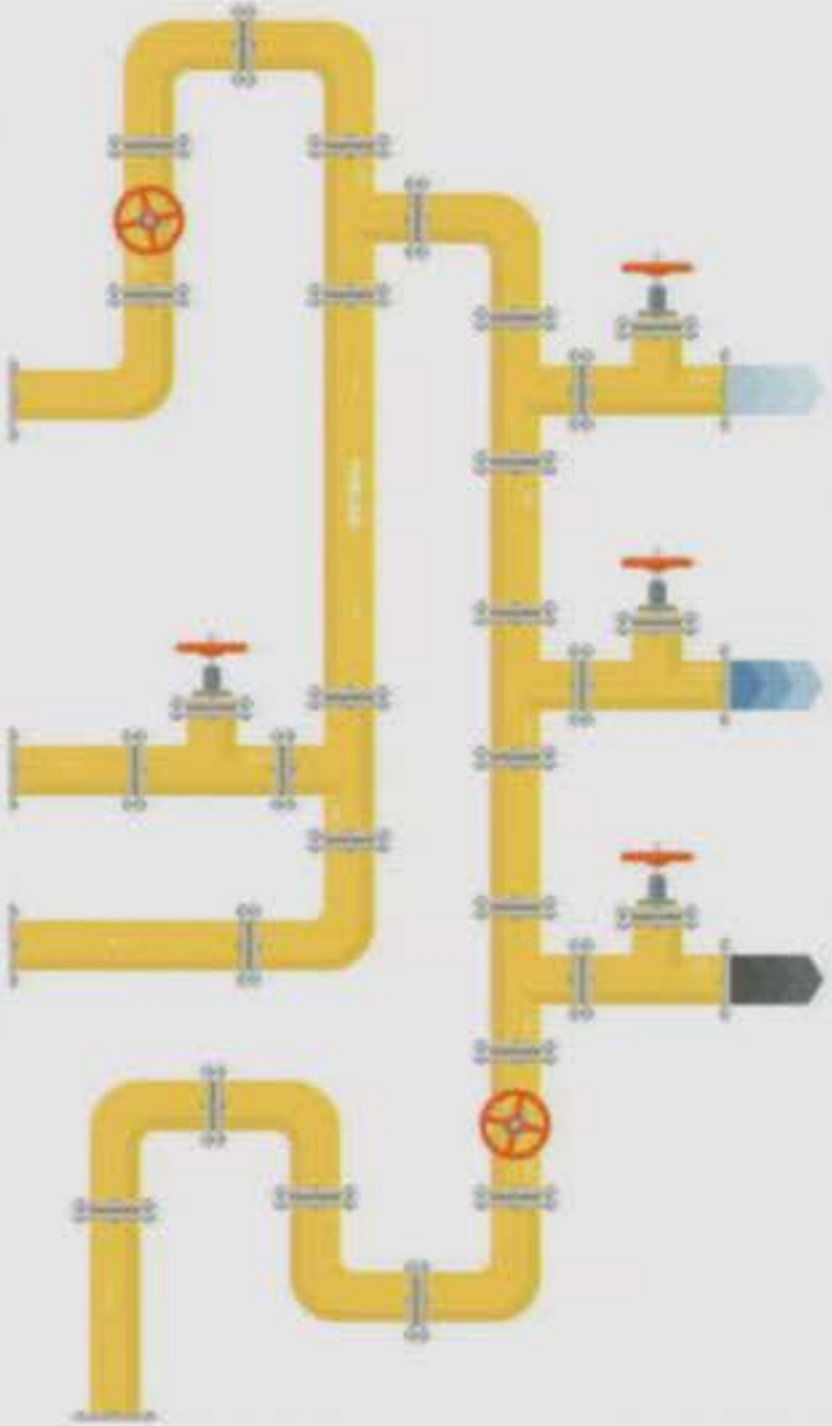
[Crawford et al. (2019), *AoAS*]

RATE Example: Null Hypothesis



[Crawford et al. (2019), *AoAS*]

General Steps in the SINATRA Pipeline



☑ Represent shapes via statistics summarizing their topology / geometry;

☑ Use a statistical model and classify shapes based on these summary statistics;

□ Derive an "evidence of association" metric for each topological / geometric feature;

□ Project these association measures back onto the original shape.

Shape Reconstruction Algorithm

- **Goal: Map the selected features back onto the shape.**

- Directions near each other will share similar information [Curry, Turner, and Mukherjee (2018)].

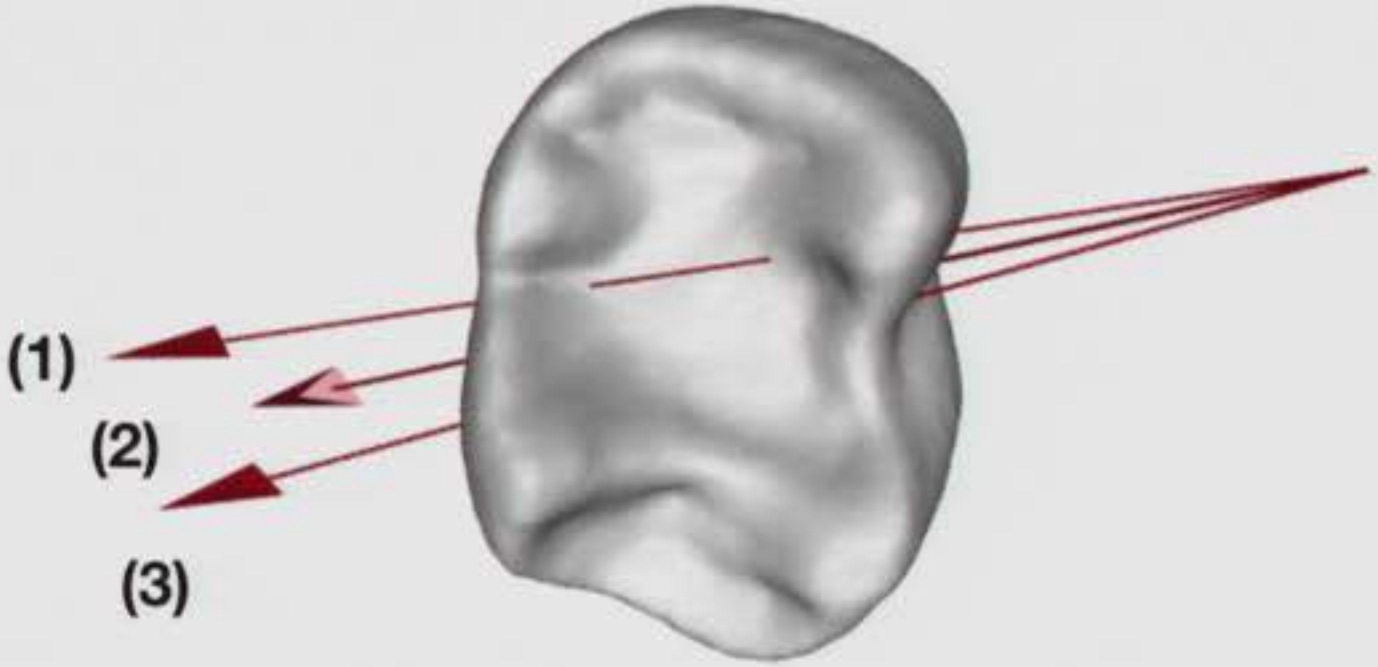
- **Reconstruction Algorithm** uses the following steps:

  (1) Pick a cone with a set of directions;

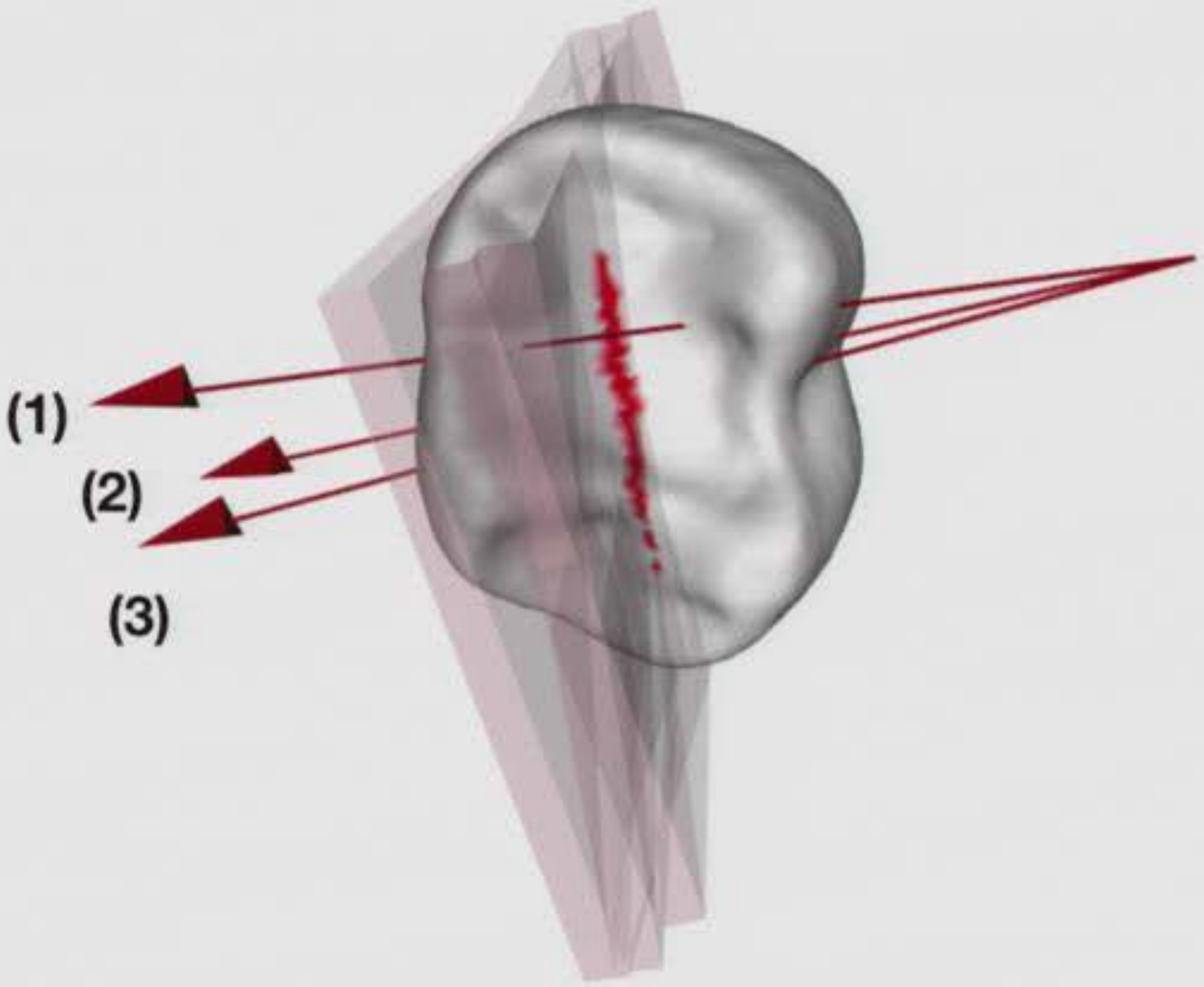
  (2) For each direction, find all vertices that correspond to the topological features selected by the GP;

  (3) Repeat this procedure for all cones;

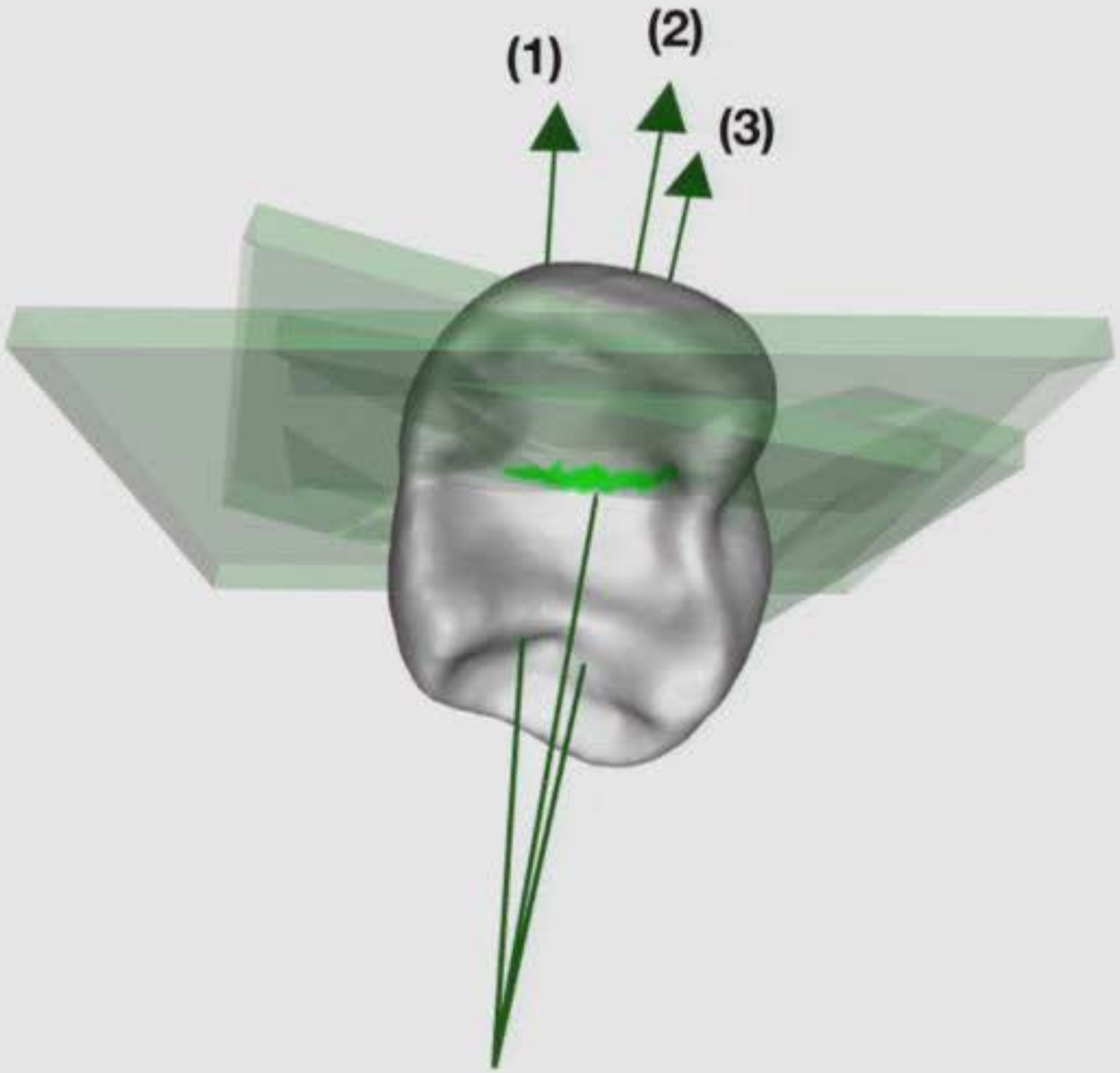
Shape Reconstruction Algorithm



Shape Reconstruction Algorithm



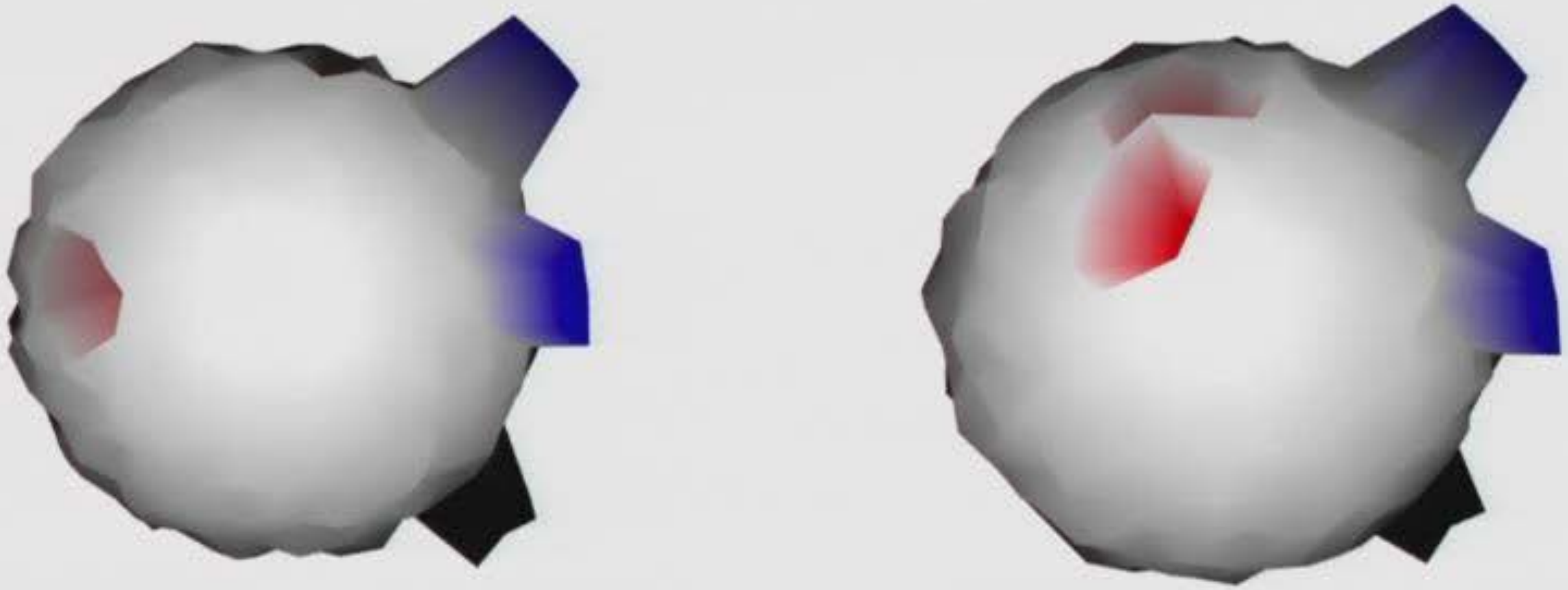
Shape Reconstruction Algorithm



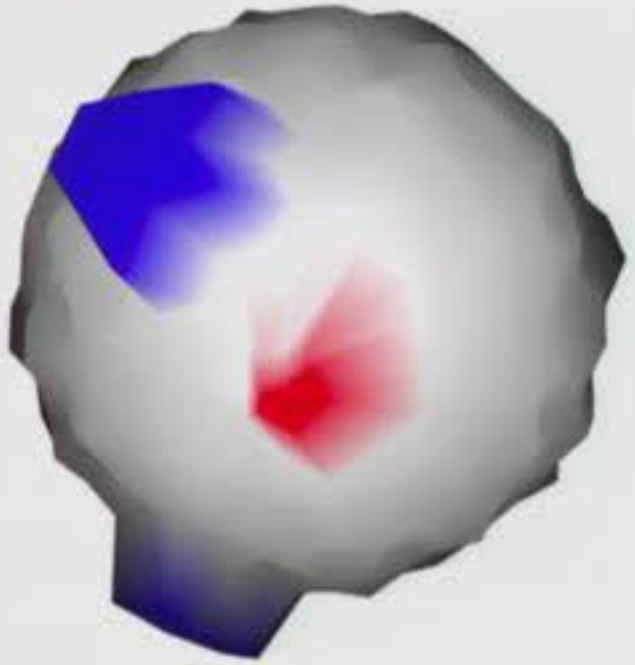
Proof-of-Concept Simulation Study

Proof-of-Concept Simulation Study

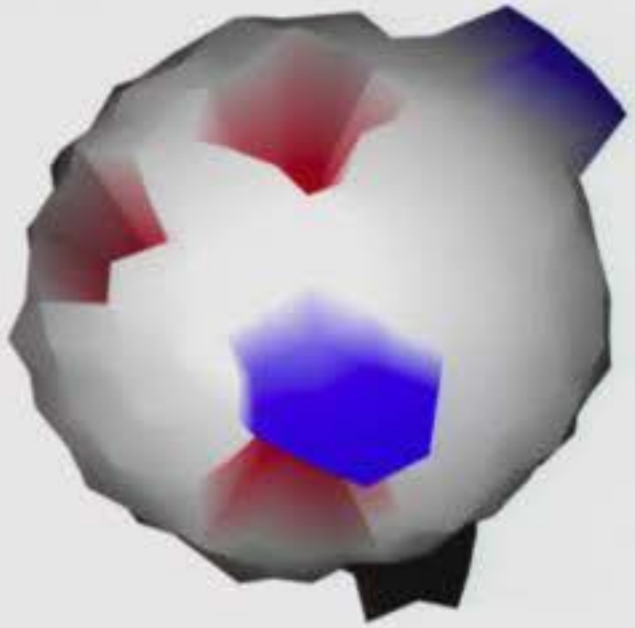
- Simulate datasets with $n = 100$ spheres split into two classes.
- Select a set of **shared regions** marked by cusps.
- Class-specific **causal regions** marked by dents.
- Assess the power of SINATRA via ROC curves (TPR vs. FPR).



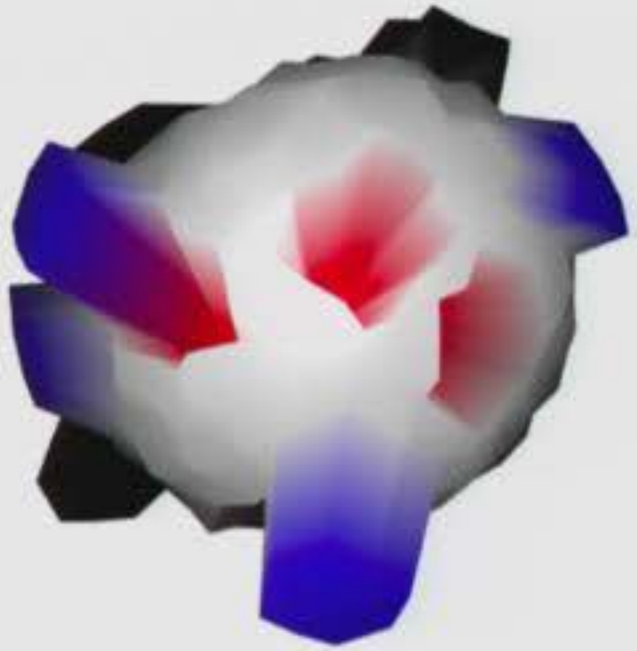
Simulation Study Results



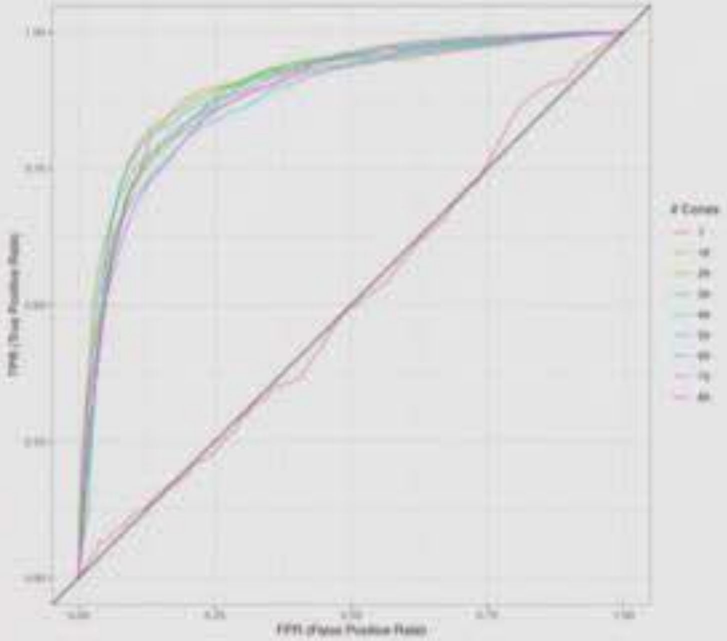
(a) Scenario I



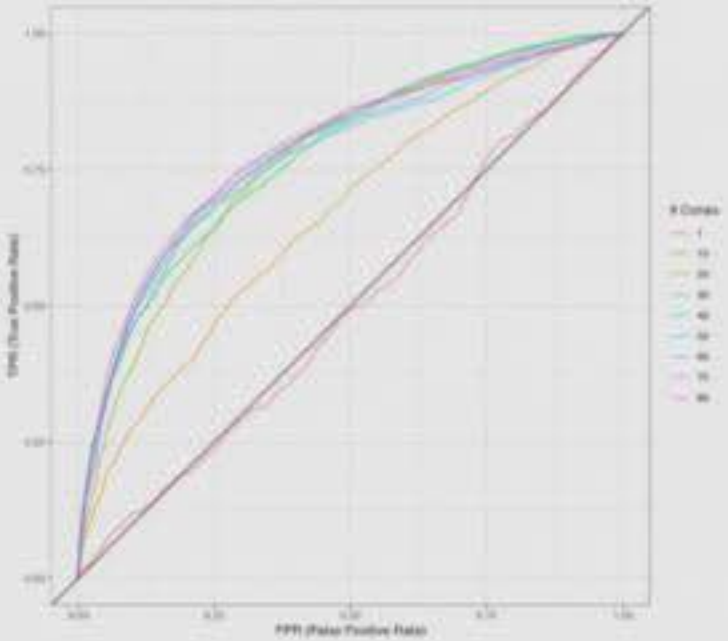
(b) Scenario II



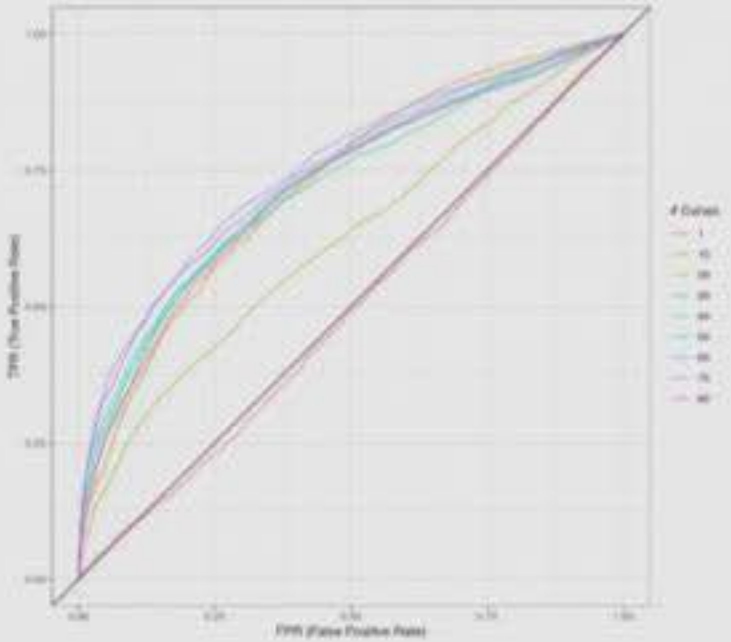
(c) Scenario III



(d) Scenario I

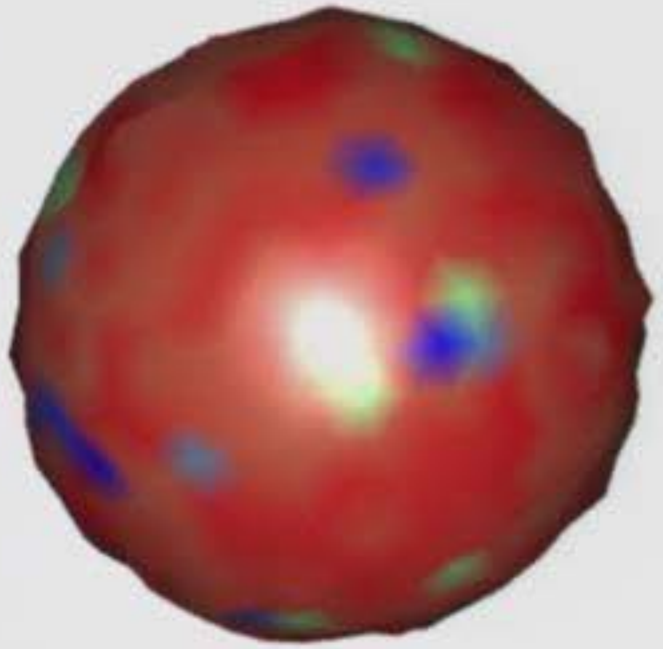


(e) Scenario II

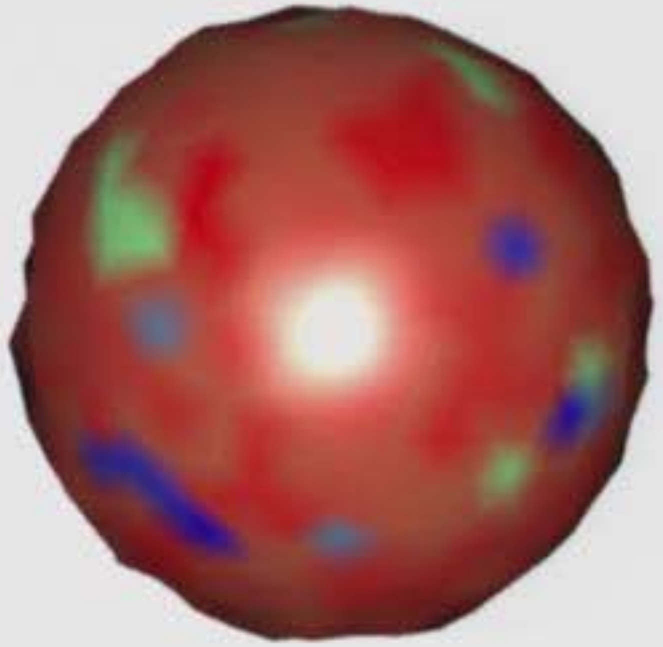


(f) Scenario III

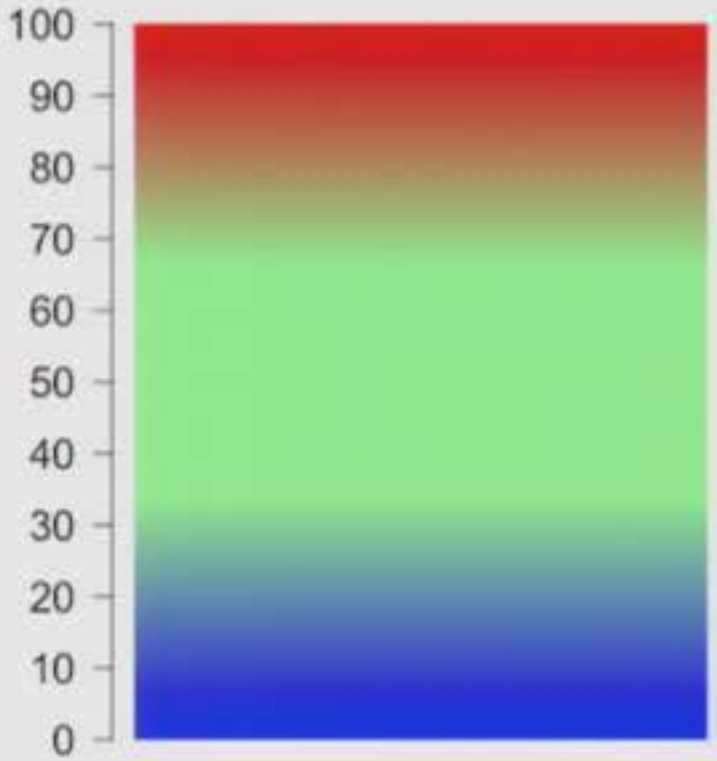
Null Hypothesis: Scenario #1



Class #1

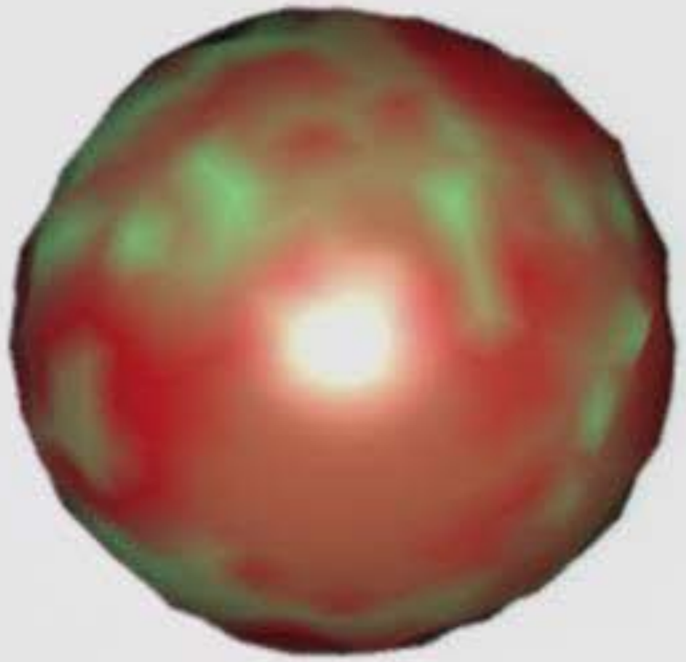


Class #2

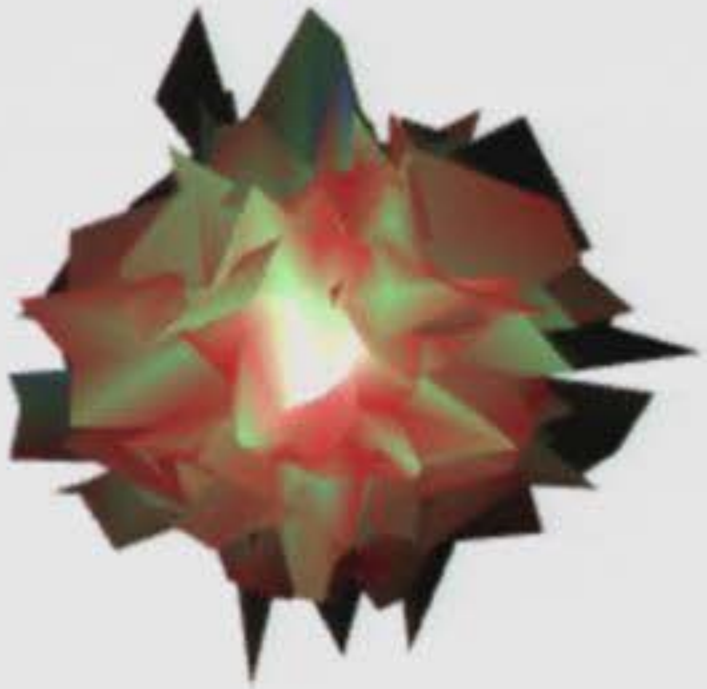


Evidence Scale (γ)

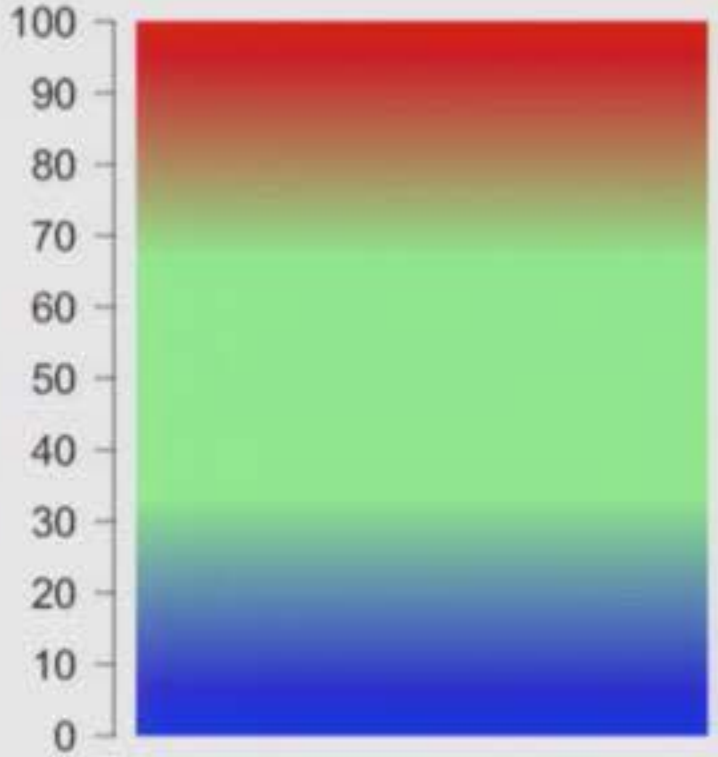
Null Hypothesis: Scenario #2



Class #1



Class #2



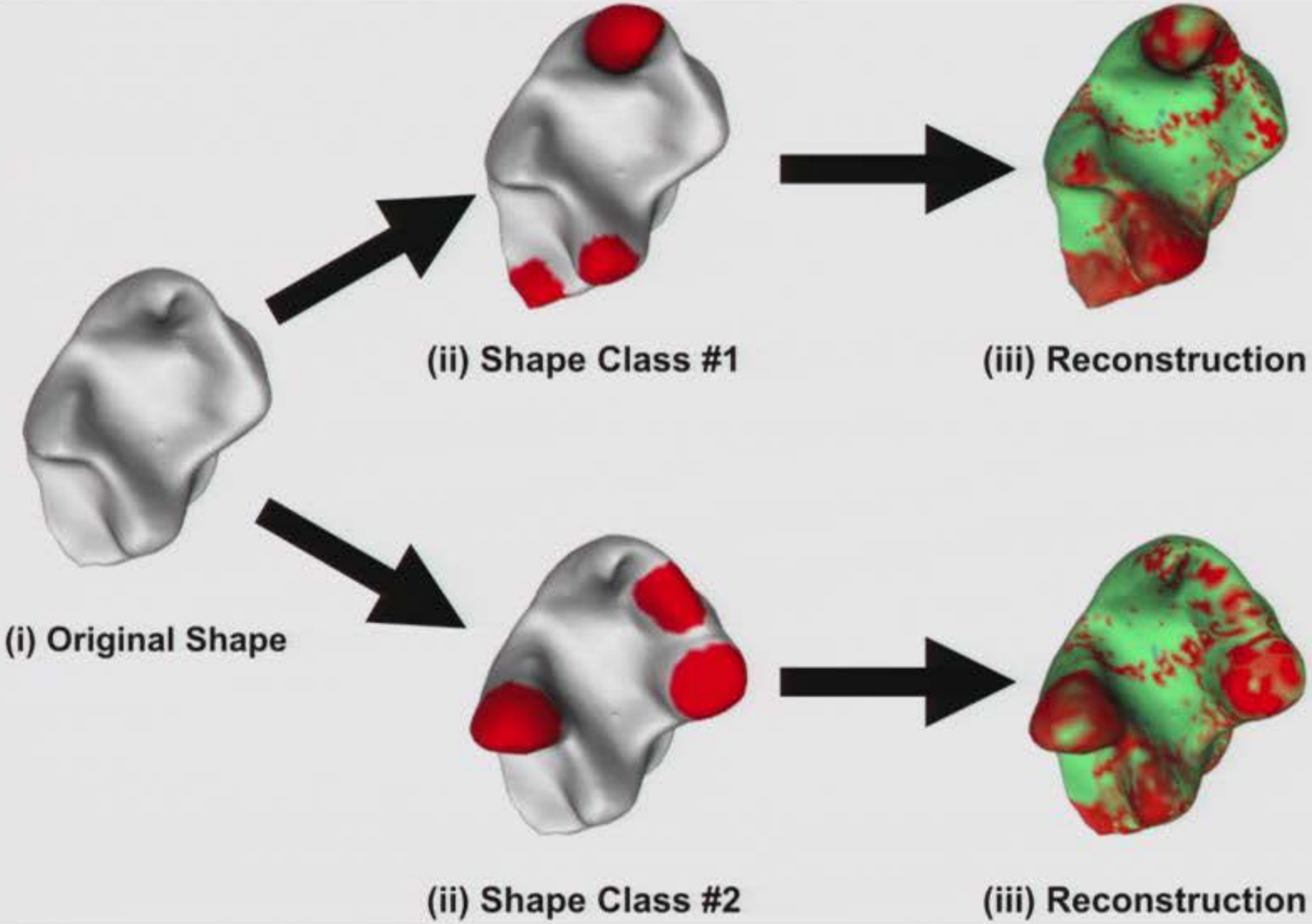
Evidence Scale (γ)

Simulation via Caricaturization of Real Data

- Computed tomography (CT) scans of real Lemuridae teeth (primates commonly known as lemurs).
- Classes are defined by creating causal and shared regions via caricaturization.
- This done by smoothly modifying regions of interest on the triangular mesh of the teeth (centered around expert-derived biological landmarks).

[Sela et al. (2015), *Comput Vis Image Underst*]

Caricature Simulation Flowchart



(i) Original Shape

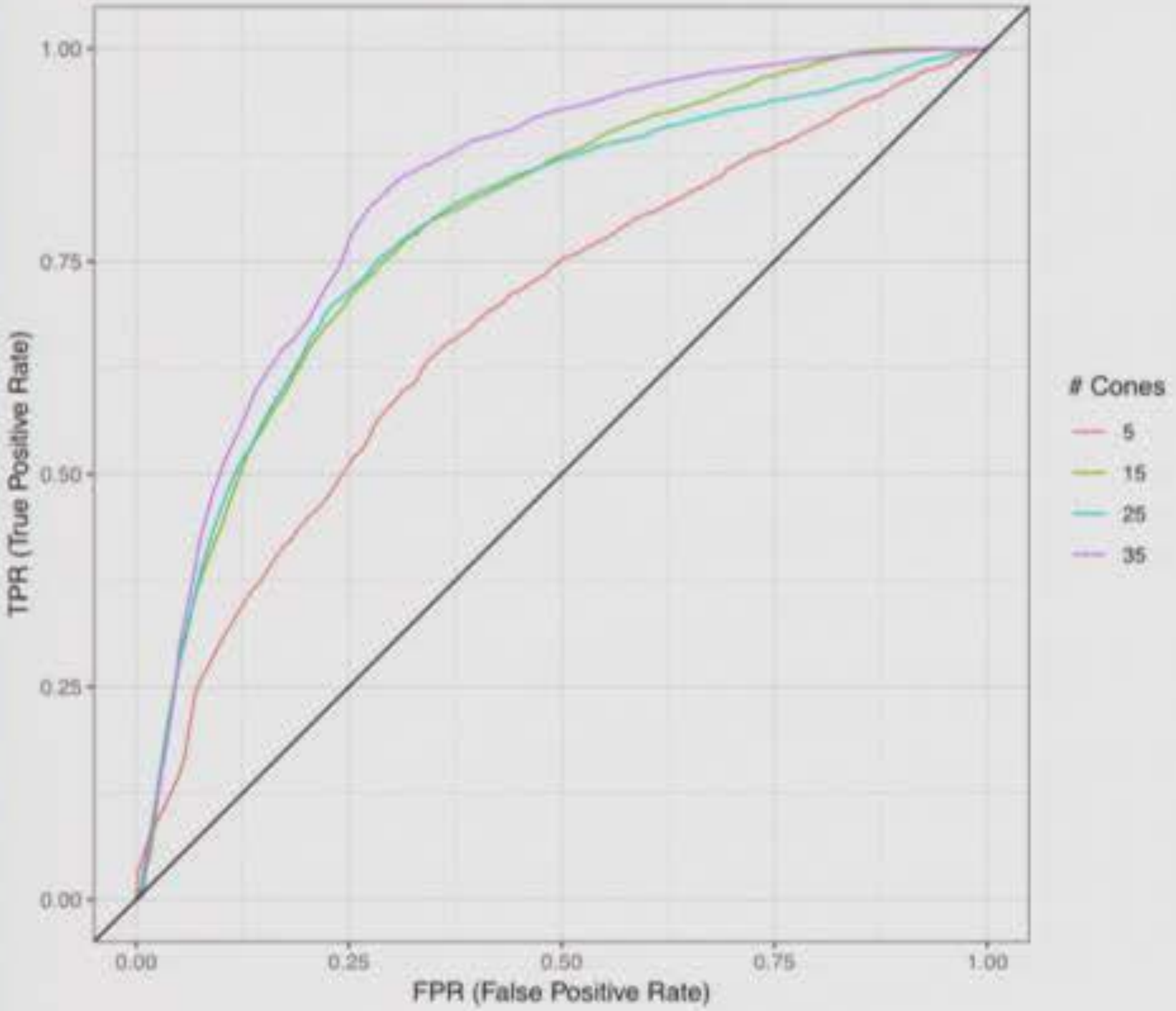
(ii) Shape Class #1

(iii) Reconstruction

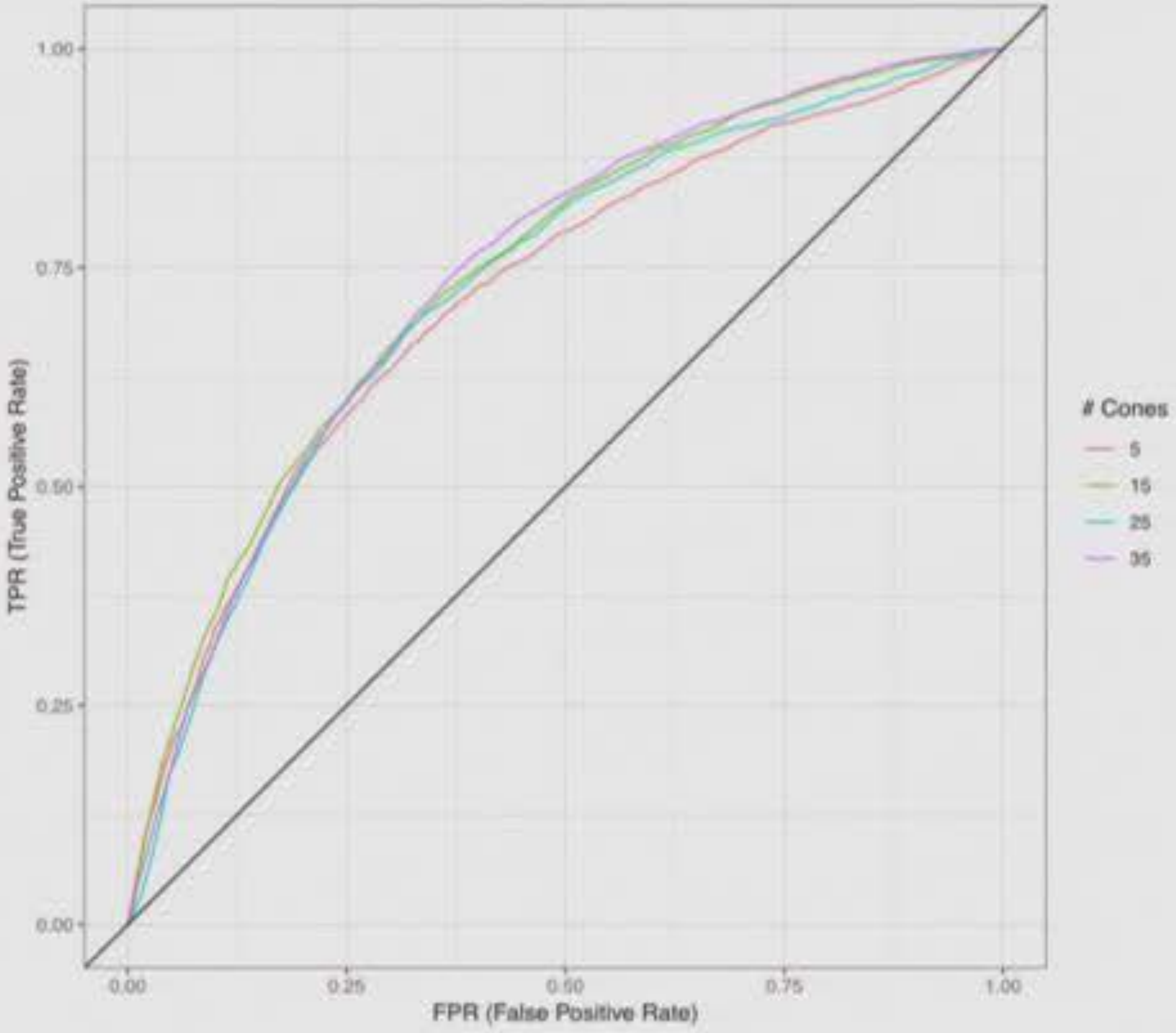
(ii) Shape Class #2

(iii) Reconstruction

Caricature Simulation Results



Easy Scenario (3 Peaks)



Difficult Scenario (5 Peaks)

Application: Recovering Known Morphological Variation

Morphological Variation Across Genera of Primates

- Data set with CT scans of $n = 59$ second mandibular molars from four genera of primates: *Tarsius*, *Saimiri*, *Microcebus*, and *Mirza*.

- **Ground Truth:** *Tarsius* have retained the paraconid (the cusp of a primitive lower molar), while the other primates have not.

- **Goal:** Assess if SINATRA recovers the information that the paraconids are specific to the *Tarsius* genus.

Morphological Variation Across Genera of Primates

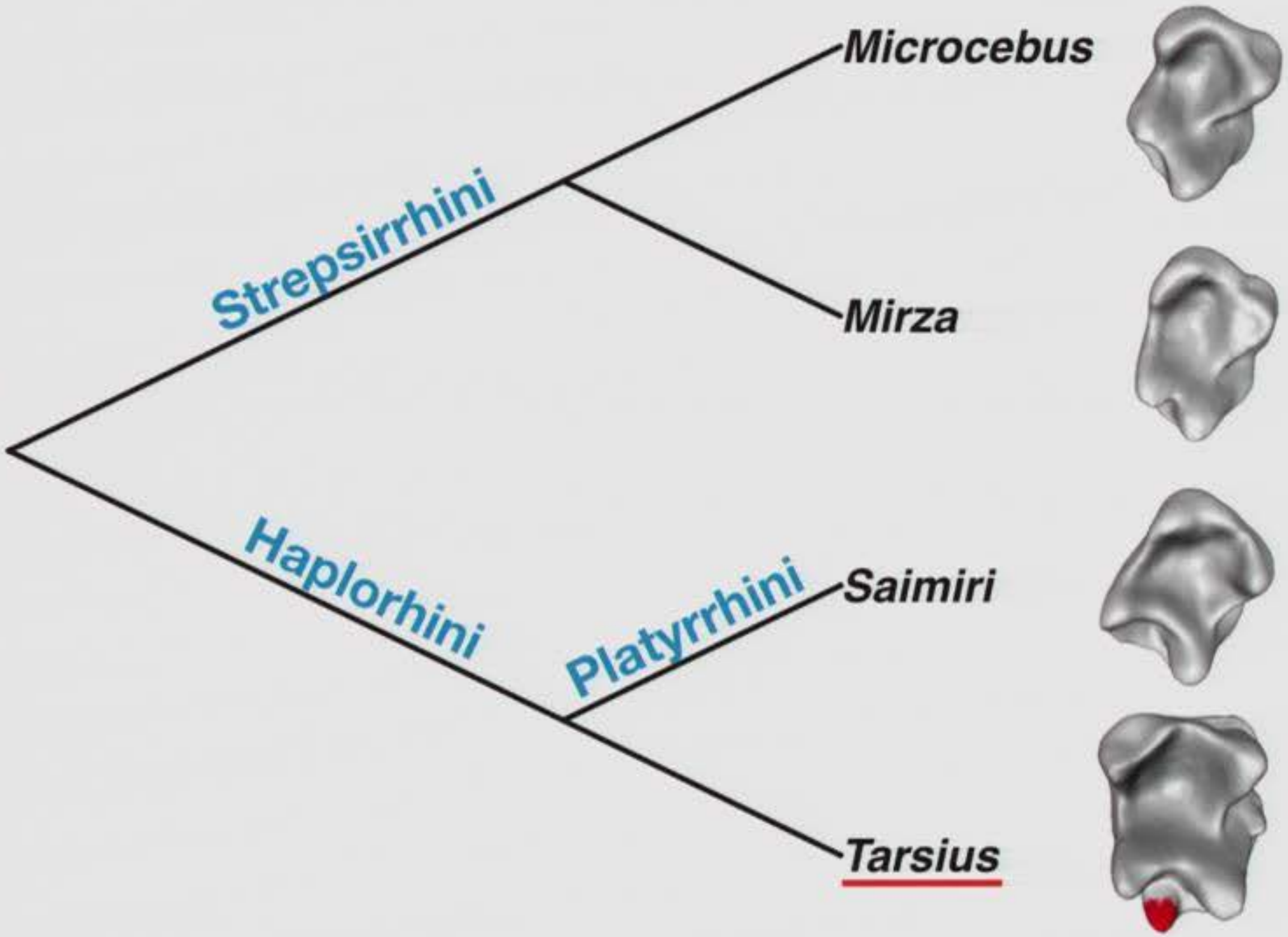
- Data set with CT scans of $n = 59$ second mandibular molars from four genera of primates: *Tarsius, Saimiri, Microcebus,* and *Mirza*.

- **Ground Truth:** *Tarsius* have retained the paraconid (the cusp of a primitive lower molar), while the other primates have not.

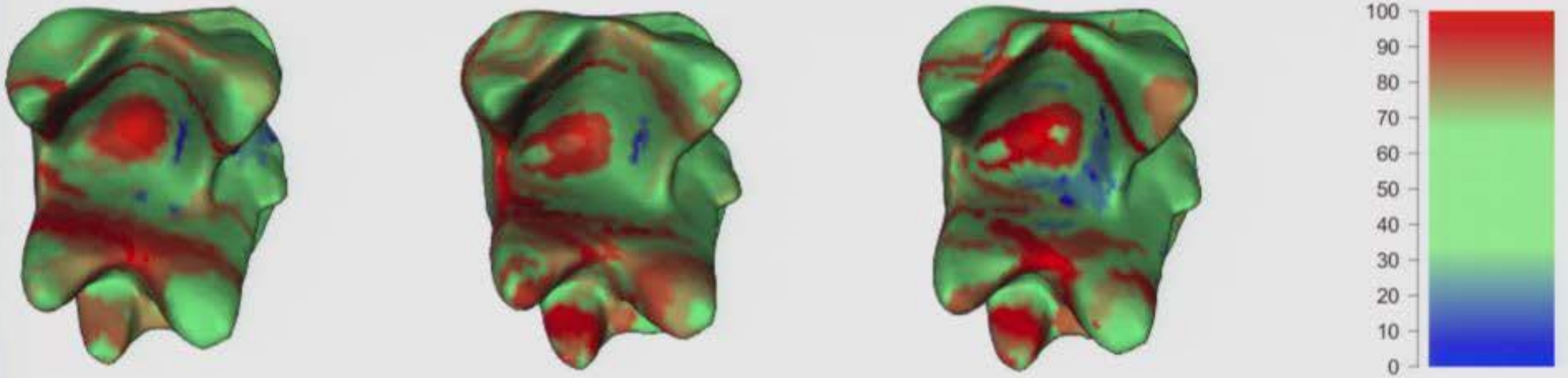
- **Goal:** Assess if SINATRA recovers the information that the paraconids are specific to the *Tarsius* genus.

- **Observation:** Determine whether variation across the molar is associated to the divergence time of the genera.

Phylogenetic Relationship Between Primates



Recovering the Region of Interest (ROI)



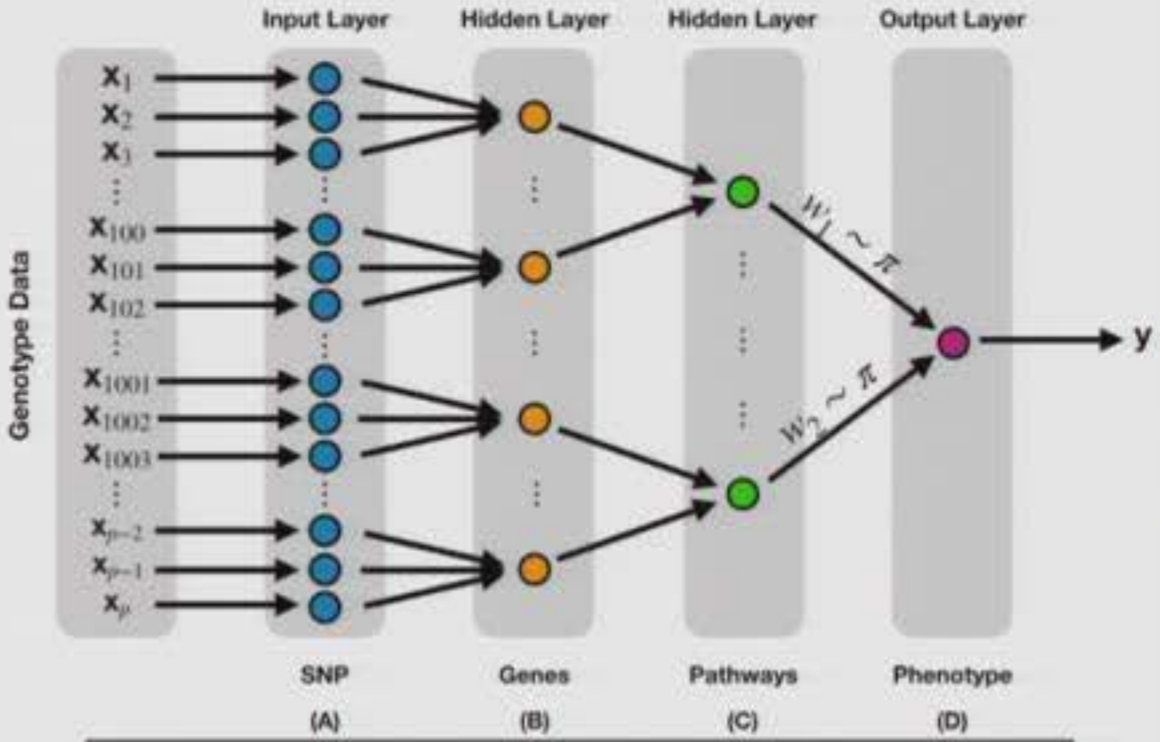
(i) Tarsius vs. Saimiri

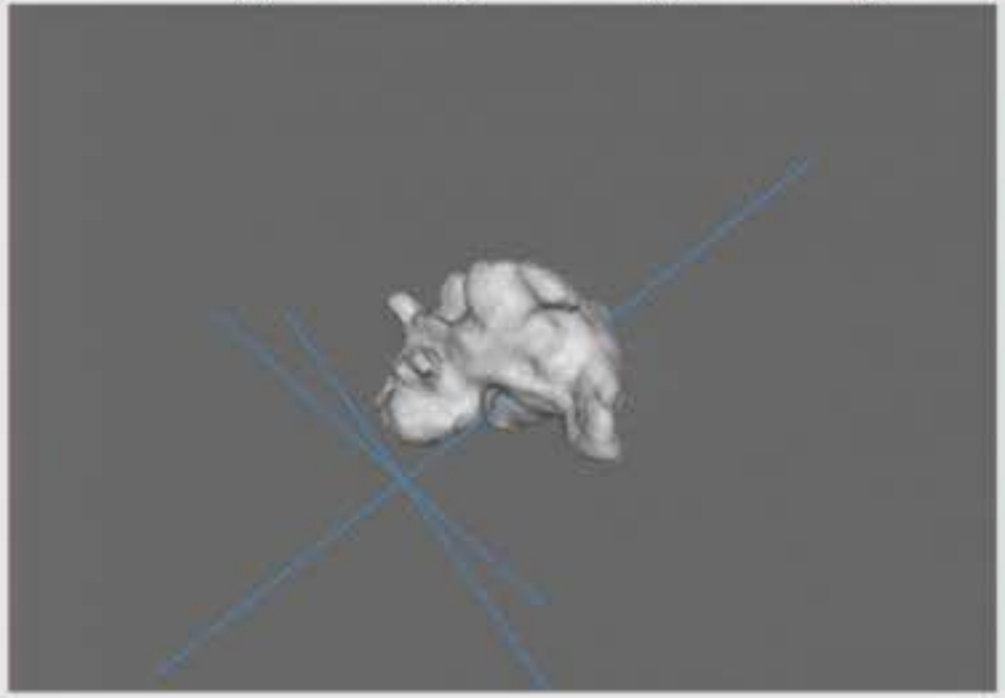
(ii) Tarsius vs. Mirza

(iii) Tarsius vs. Microcebus

Evidence Scale

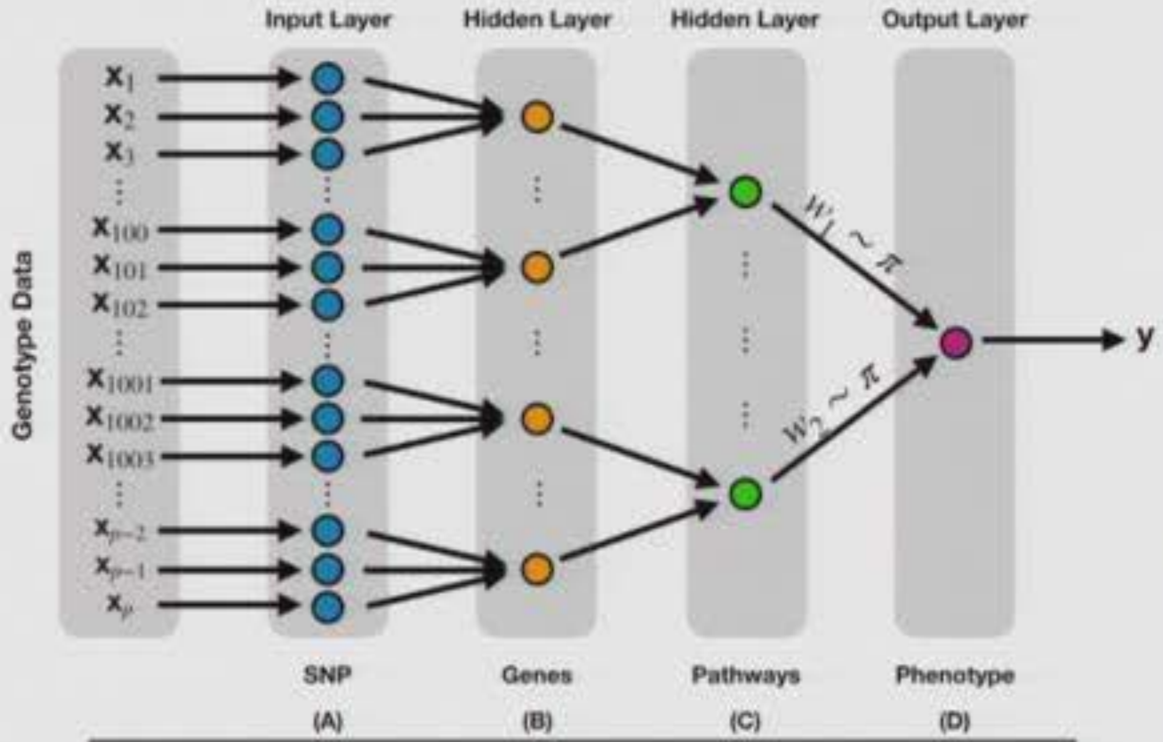
Ongoing Work in the Lab

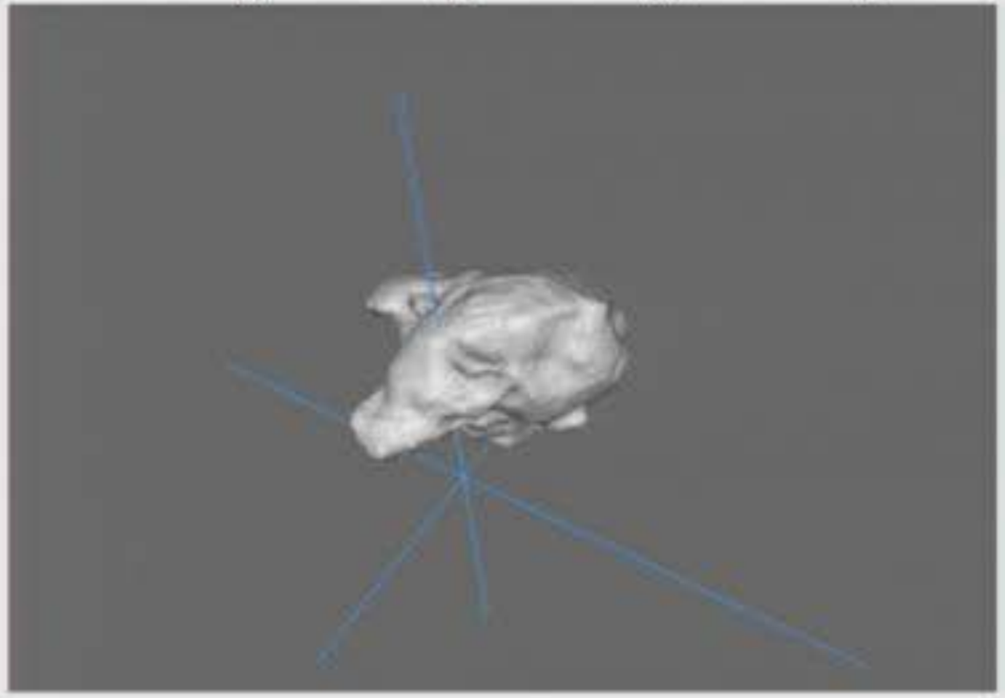




Ongoing Work in the Lab

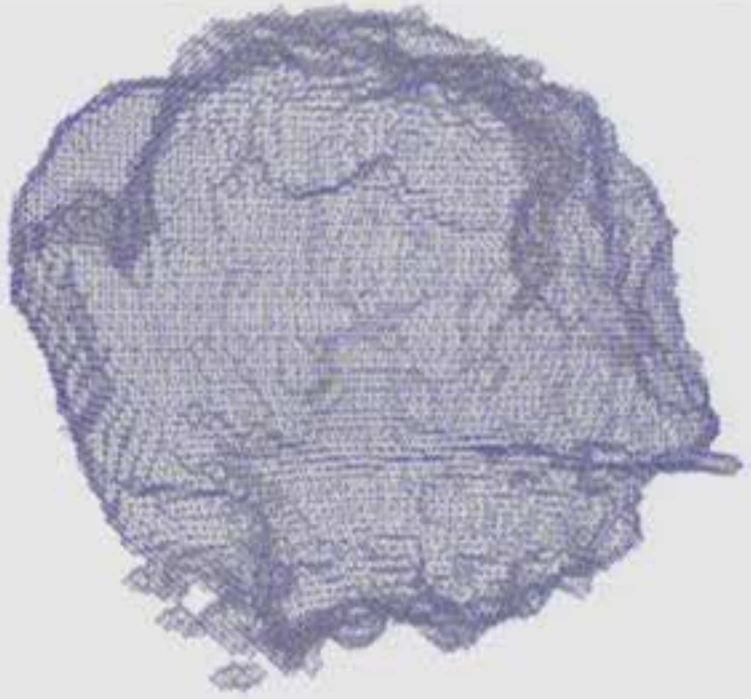
- **Explore pairing SINATRA with probabilistic deep learning methods:**
  - Biologically annotated neural networks (BANNs) provide a framework amenable for genomic studies with small sample sizes.
  - Extend the BANN framework to model multiple *-omic* and shape information simultaneously.
- **Association Analyses Using Shape Summary Statistics Derived from MRIs:**
  - Probe whether shape variation is correlated with genotypic/phenotypic variation.

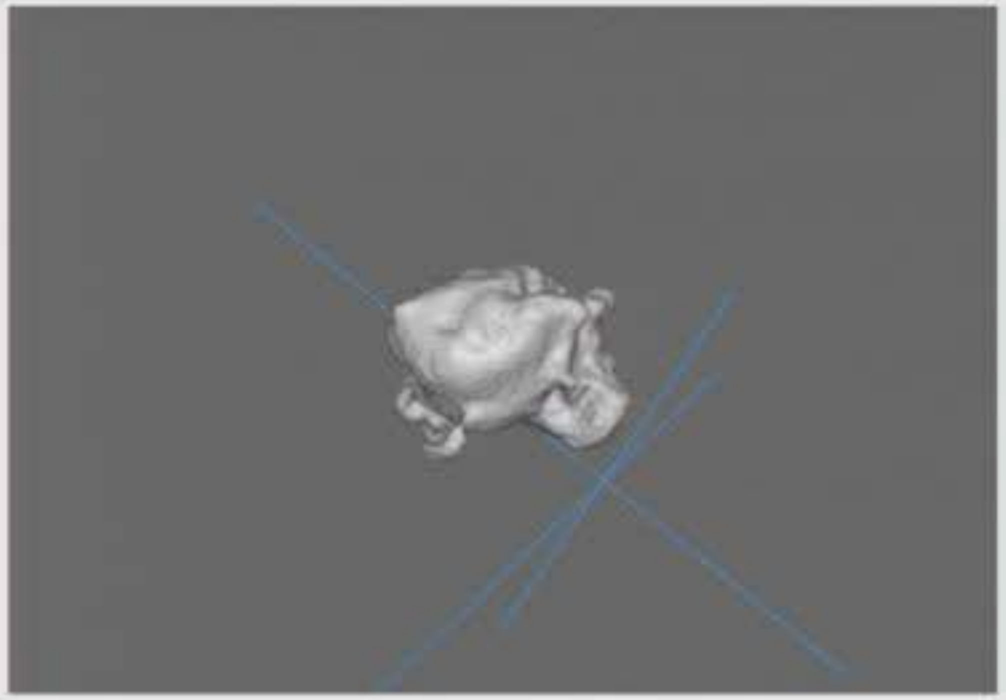




Ongoing Work in the Lab

- **Explore pairing SINATRA with probabilistic deep learning methods:**
  - Biologically annotated neural networks (BANNs) provide a framework amenable for genomic studies with small sample sizes.
  - Extend the BANN framework to model multiple *-omic* and shape information simultaneously.
- **Association Analyses Using Shape Summary Statistics Derived from MRIs:**
  - Probe whether shape variation is correlated with genotypic/phenotypic variation.
  - Identify physical characteristics of brain tumors that are linked to oncogenic signatures or underlying signaling cascades that have become activated.





Acknowledgements

- **SINATRA Collaborators:**
  - Bruce Wang (Princeton University)***
  - Timothy Sudijono (Brown University)***
  - Henry Kirveslahti (Duke University)***
  - Tingran Gao, Ph.D. (University of Chicago)
  - Doug M. Boyer, Ph.D. (Duke University)
  - Sayan Mukherjee, Ph.D. (Duke University)
- **Other Contributors:**
  - Katharine Turner, Ph.D. (ANU)
  - Anthea Monod, Ph.D. (Tel Aviv University)
  - Seth Flaxman, Ph.D. (ICL)
  - Dan Runcie, Ph.D. (UC Davis)
  - Mike West, Ph.D. (Duke University)
  - Christine Wall, Ph.D. (Duke University)
- **Data Availability:**
  - https://gaotingran.com/codes/codes.html
  - http://www.wisdom.weizmann.ac.il/~ylipman

Alfred P. Sloan
FOUNDATION

NIH
National Institute of
General Medical Sciences

BROWN
School of Public Health

Research partly supported by IDeA grants P20GM109035 (Center for Computational Biology of Human Disease) and P20GM103645 (Center for Central Nervous System Function) from NIH NIGMS. Content is sole responsibility of author(s).

Crawford Lab.

Relevant References

SINATRA Pipeline:

- B. Wang*, T. Sudijono*, H. Kirveslahti*, T. Gao, D.M. Boyer, S. Mukherjee, and **L. Crawford**. A statistical pipeline for identifying physical features that differentiate classes of 3D shapes. *bioRxiv*. 701391.

Topological Summary Statistics:

- 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.

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

RelATive cEntrality (RATE) Measures:

- **L. Crawford**, S.R. Flaxman, D.E. Runcie, and M. West (2019). Predictor variable prioritization in nonlinear models: a genetic association case study. *Annals of Applied Statistics*. **13**(2): 958-989.

SINATRA Software:

- https://github.com/lcrawlab/SINATRA