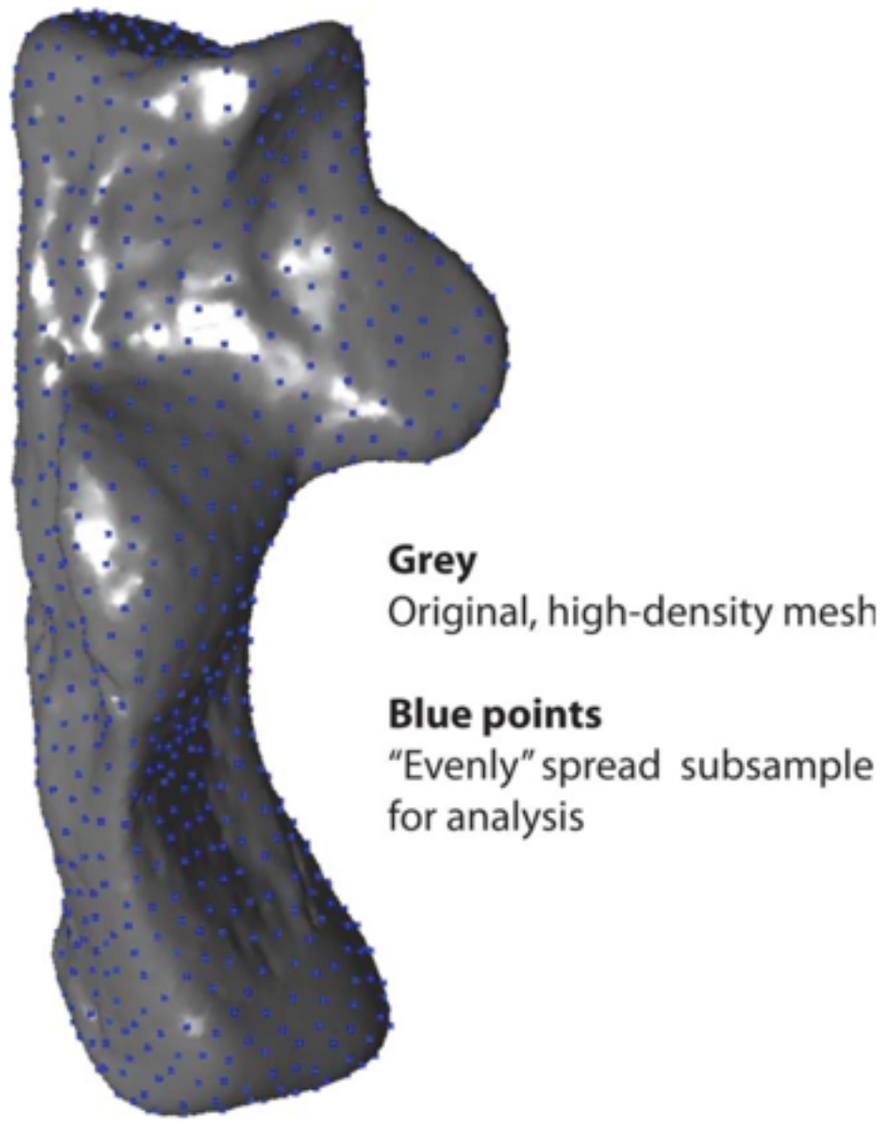
Applications of SSA: Phylogenetics

Shan Shan
Department of Mathematics
Duke University

Auto3dgm recap

- **Resampling**
- Pairwise alignment
- Global alignment

FPS: Farthest Point Sampling

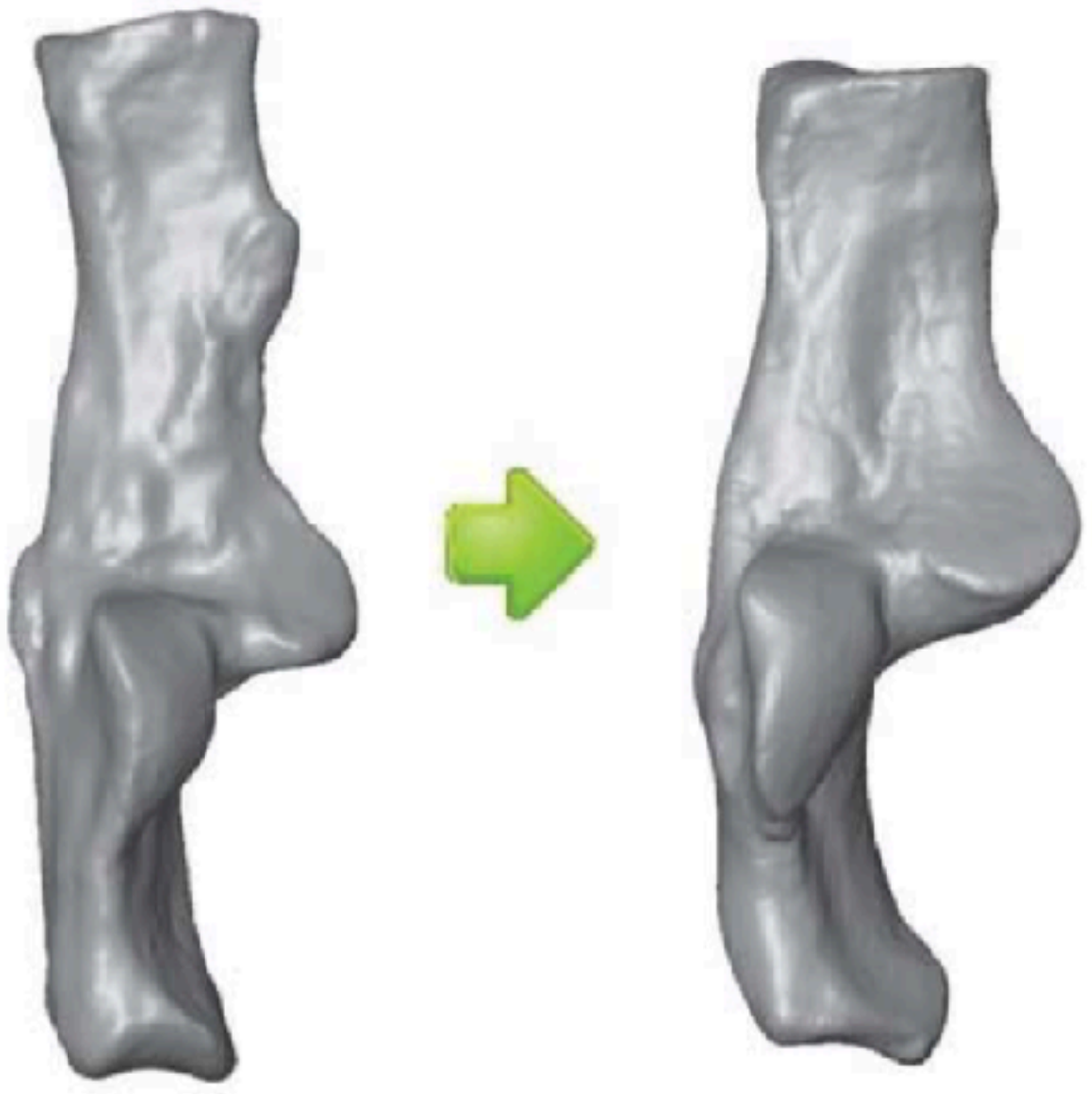


Boyer, Doug M., et al. "A new fully automated approach for aligning and comparing shapes." *The Anatomical Record* 298.1 (2015): 249-276.

Auto3dgm recap

- Resampling
- **Pairwise alignment**
- Global alignment

Iteratively search:
Best possible rotation via Kabsch
Best possible permutation via Hungarian

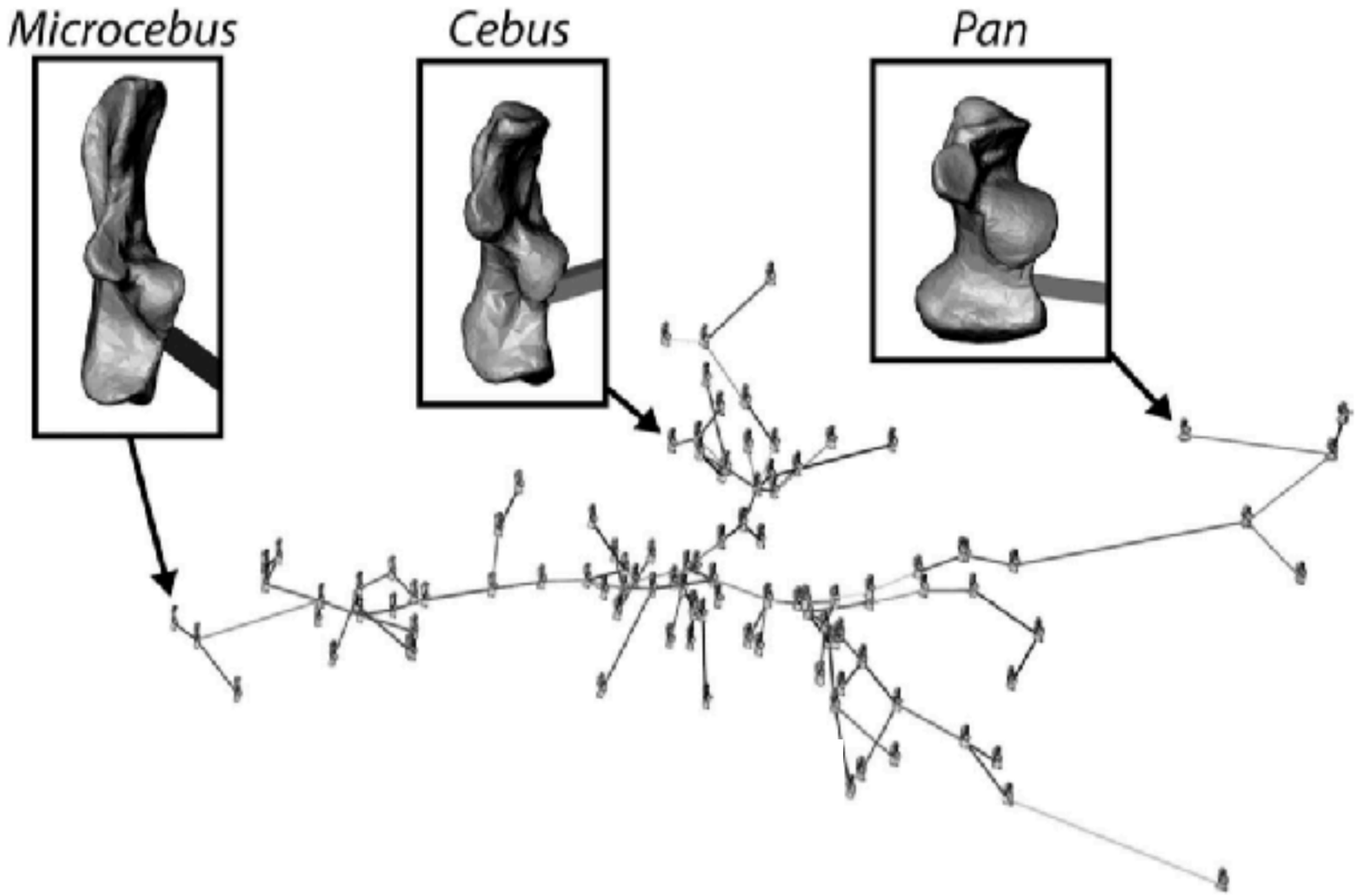


Boyer, Doug M., et al. "A new fully automated approach for aligning and comparing shapes." *The Anatomical Record* 298.1 (2015): 249-276.

Auto3dgm recap

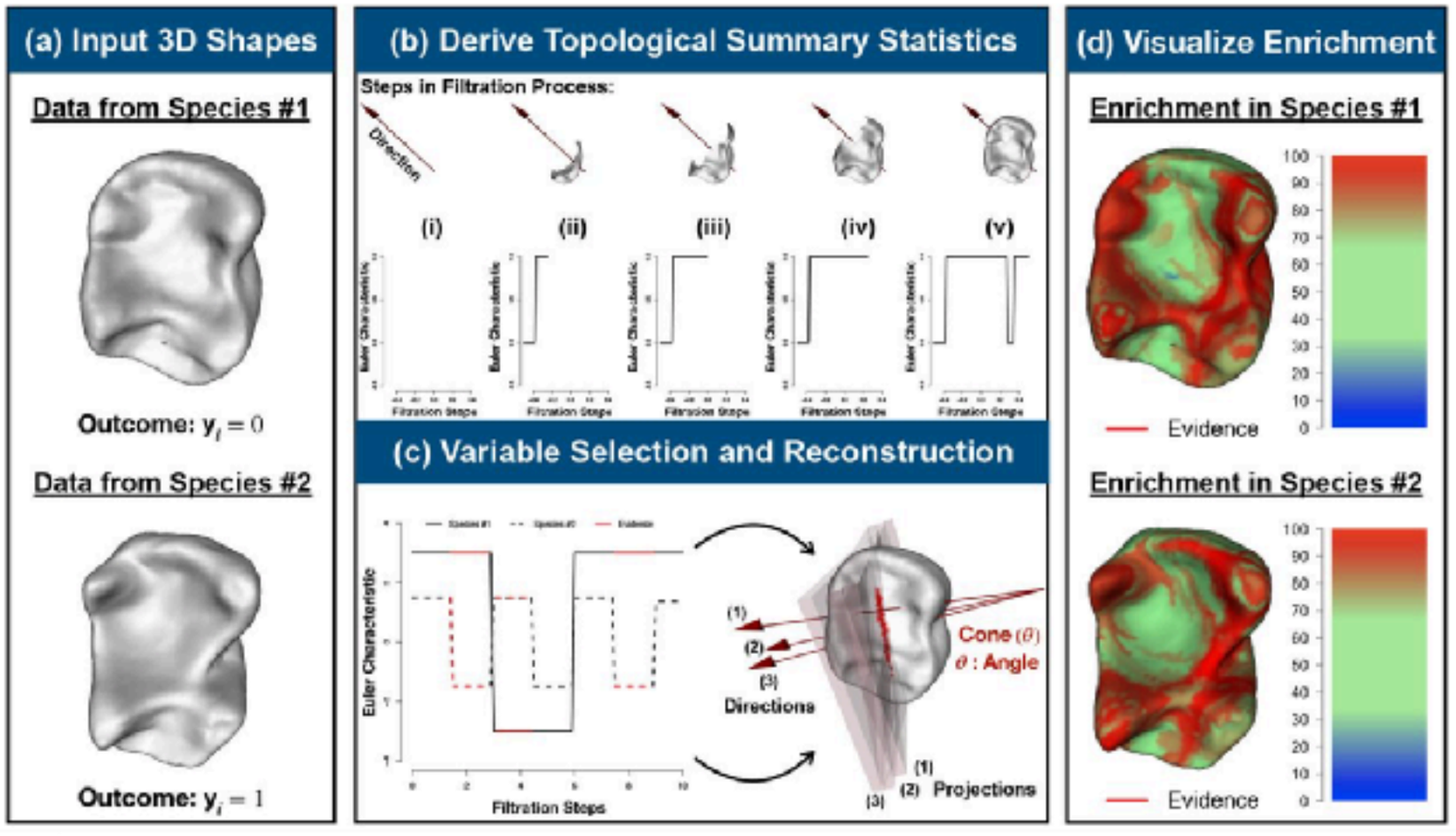
- Resampling
- Pairwise alignment
- **Global alignment**

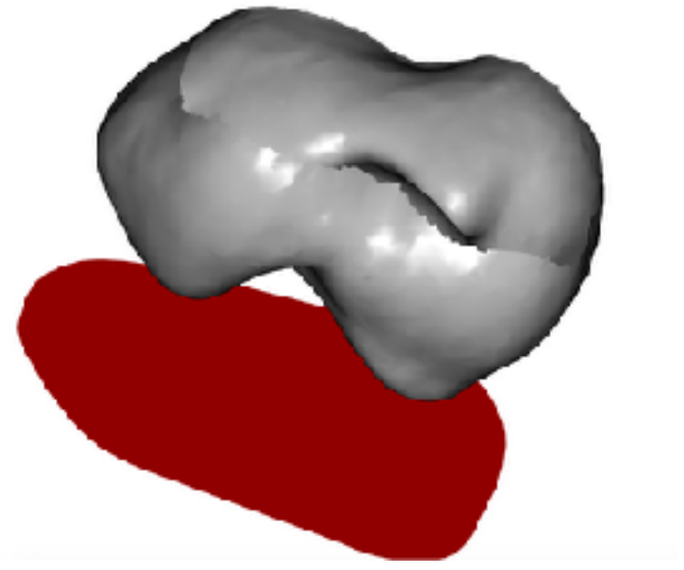
MST: Minimum Spanning Tree



Boyer, Doug M., et al. "A new fully automated approach for aligning and comparing shapes." *The Anatomical Record* 298.1 (2015): 249-276.

Application 1: Data Pre-processing





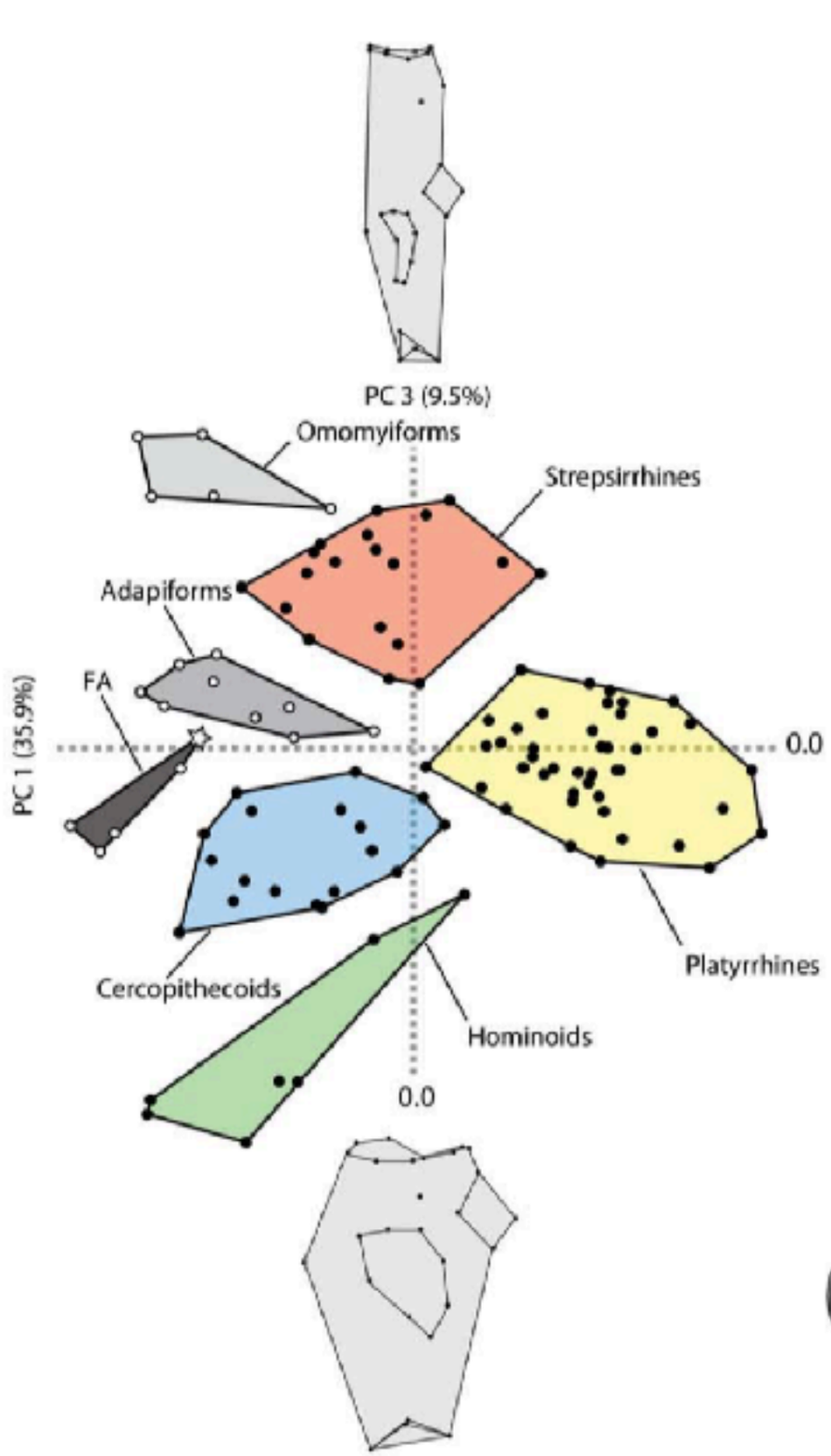
SINATRA: a statistical framework for feature selection

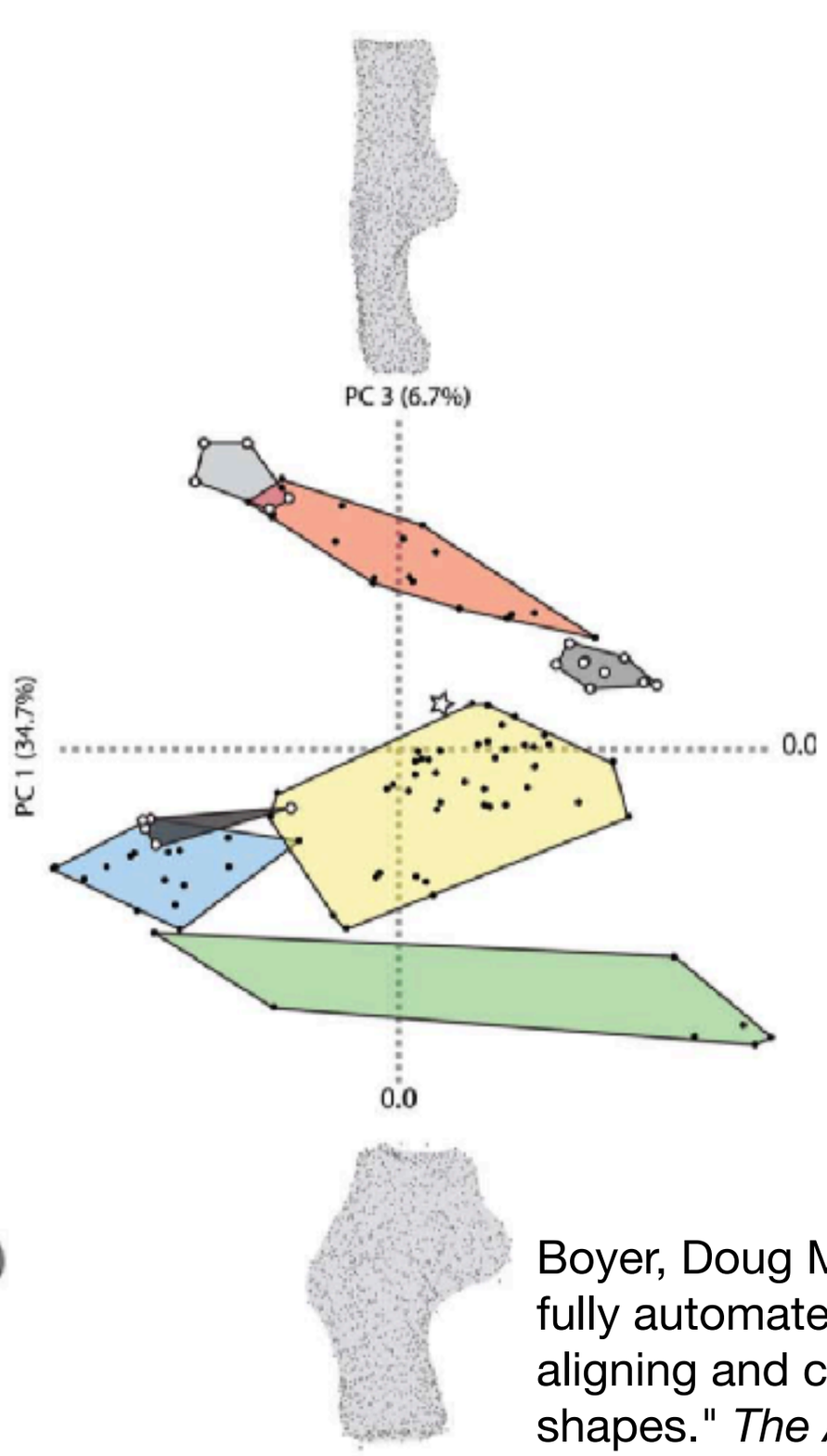
RFI: Relief Index

Wang, Bruce, et al. "SINATRA: A Sub-Image Analysis Pipeline for Selecting Features that Differentiate Classes of 3D Shapes." *bioRxiv* (2019): 701391.

Boyer, Doug M. "Relief index of second mandibular molars is a correlate of diet among prosimian primates and other euarchontan mammals." *Journal of Human Evolution* 55.6 (2008): 1118-1137.

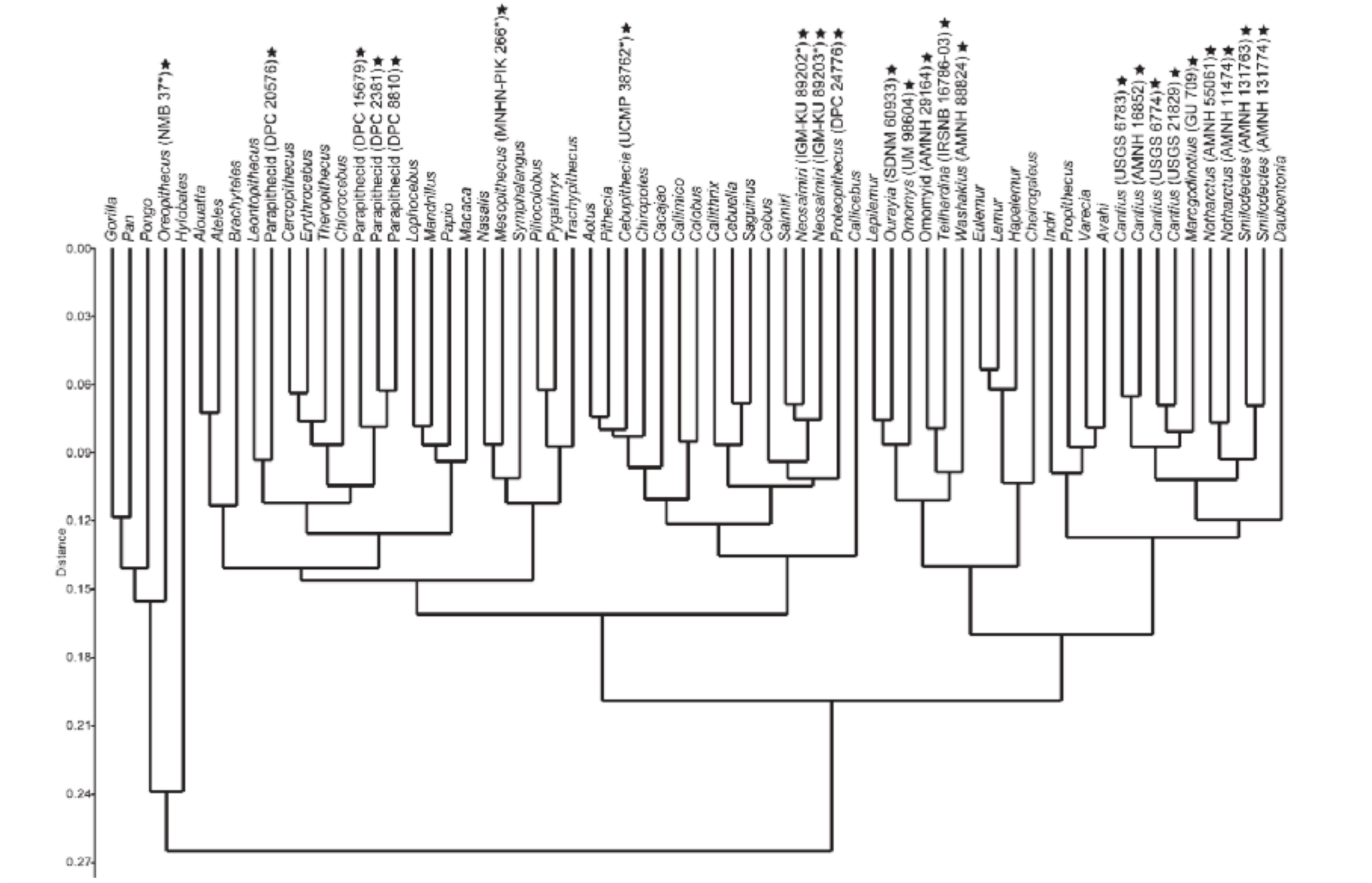
Application 2: Shape space





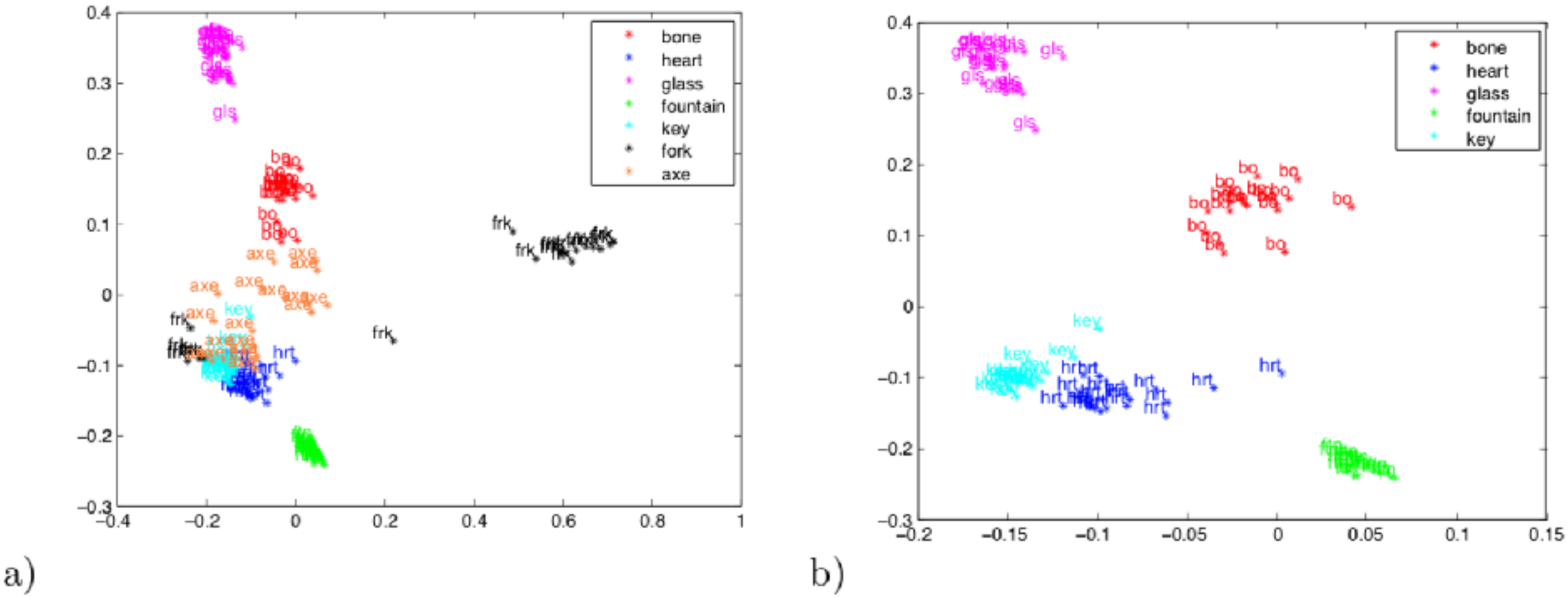
Boyer, Doug M., et al. "A new fully automated approach for aligning and comparing shapes." *The Anatomical Record* 298.1 (2015): 249-276

Application 3: Phenotypic affinities



Boyer, Doug M., et al. "A new fully automated approach for aligning and comparing shapes." *The Anatomical Record* 298.1 (2015): 249-276.

Application 4: Align shape spaces



Turner, Katharine, Sayan Mukherjee, and Doug M. Boyer. "Persistent homology transform for modeling shapes and surfaces." *Information and Inference: A Journal of the IMA* 3.4 (2014): 310-344.

Application 5: Phylogenetics

How do we study the evolutionary process without landmarks?

Application 5: Phylogenetics

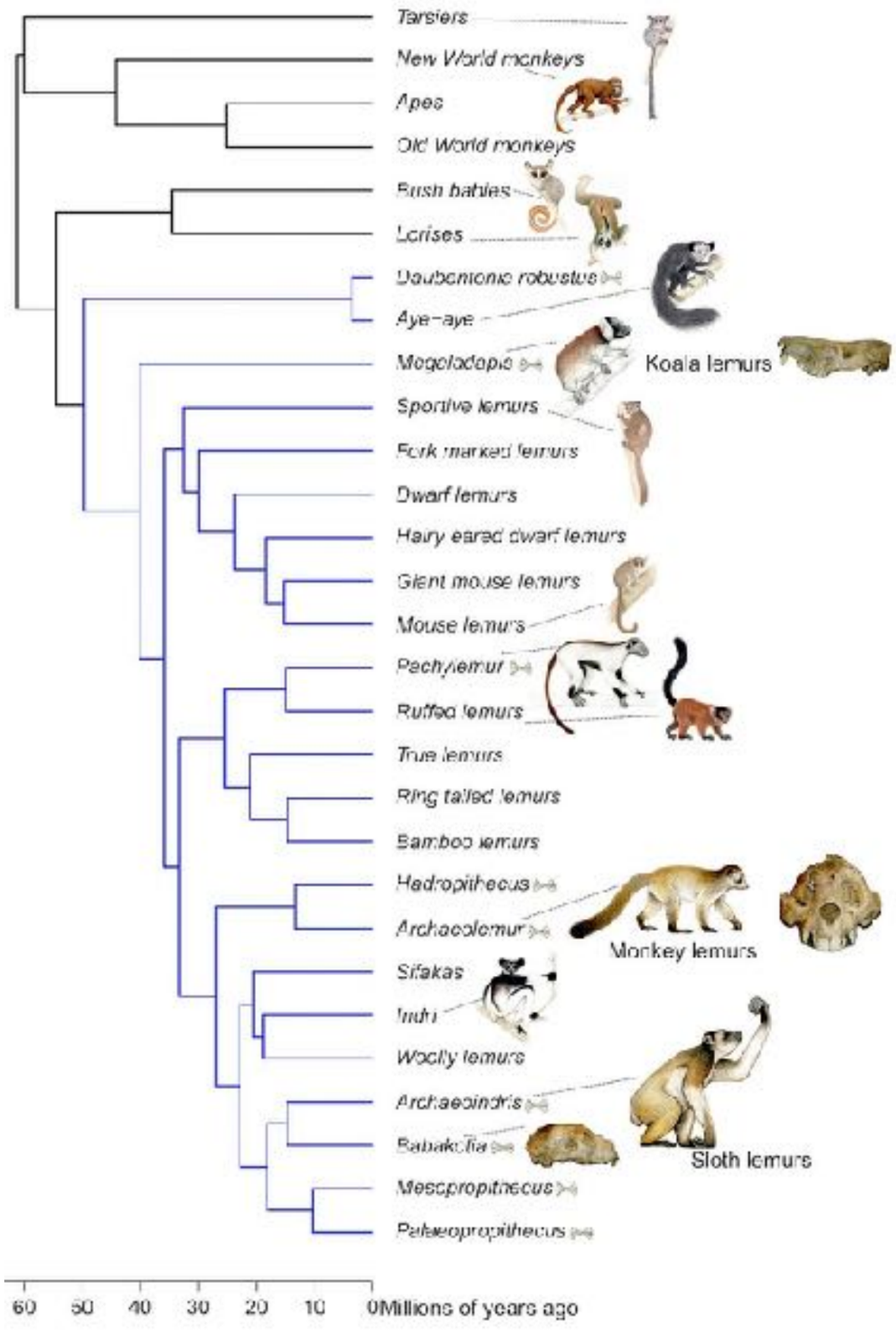
How do we study the evolutionary process without landmarks?

- Traditional phylogenetic comparative methods **with** landmarks
- Study evolution with aligned shapes and no landmarks

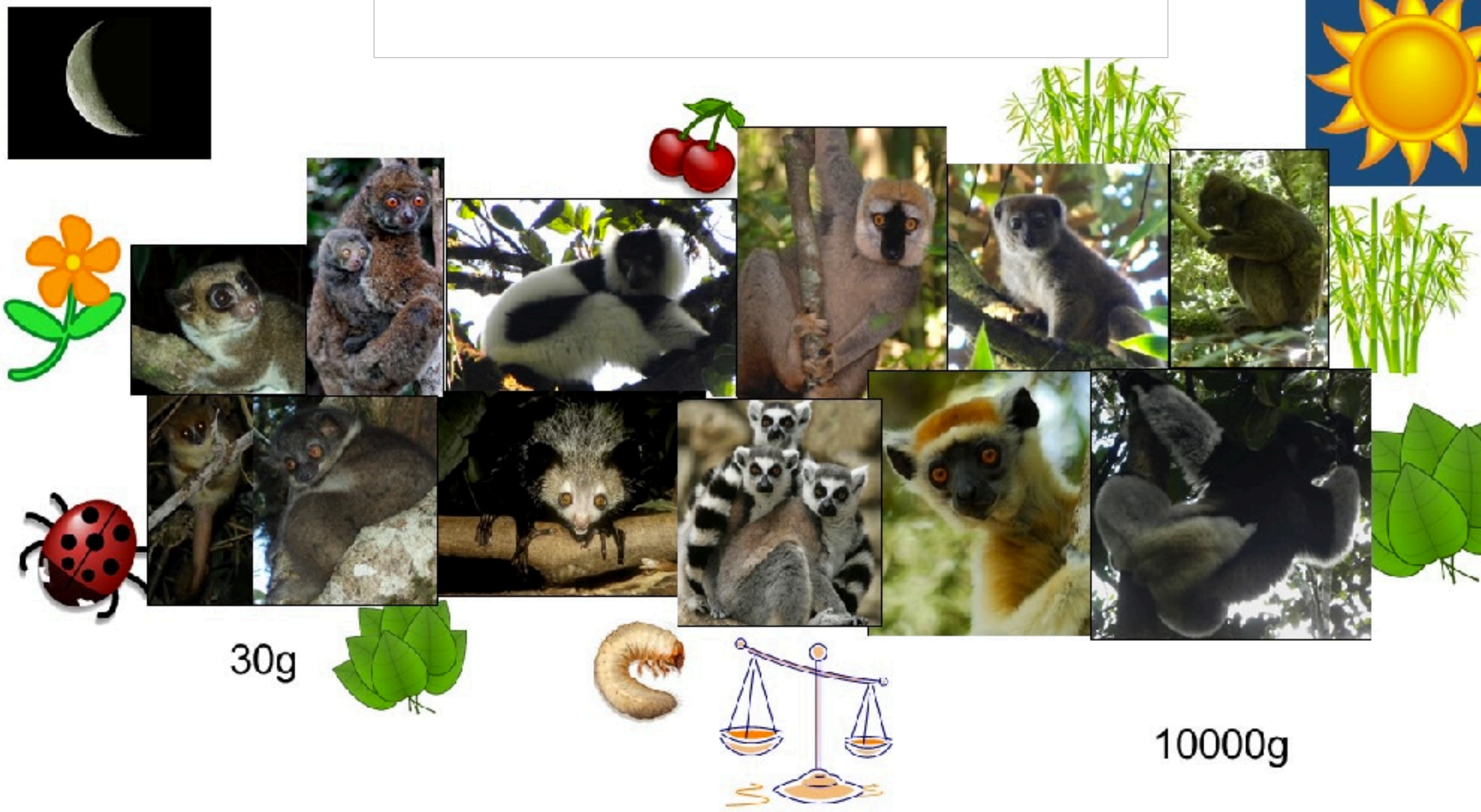
Part I:

How do animals evolve?

Diversity of life: Lemurs of Madagascar



Diversity of life: Lemurs of Madagascar

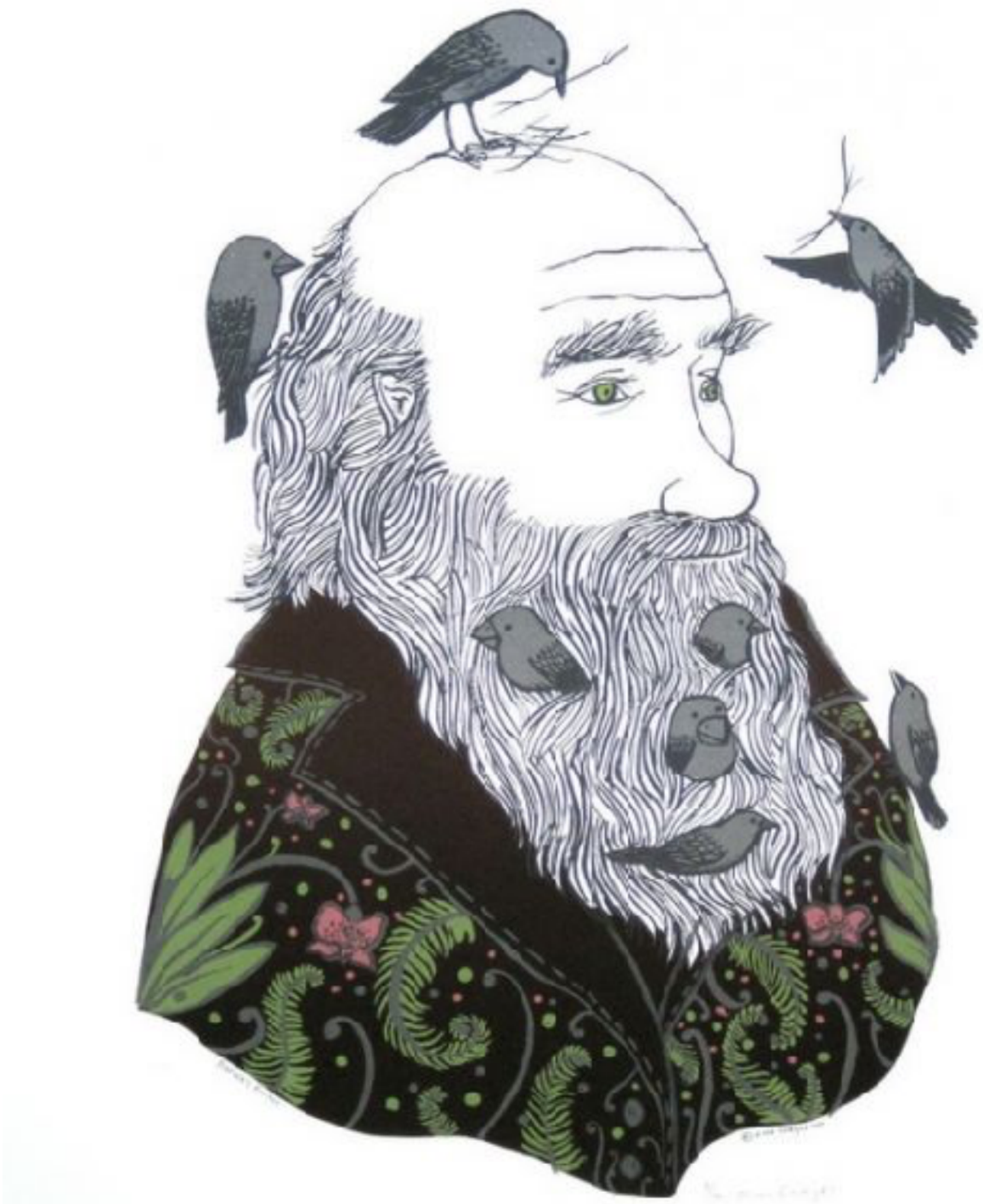


30g

10000g

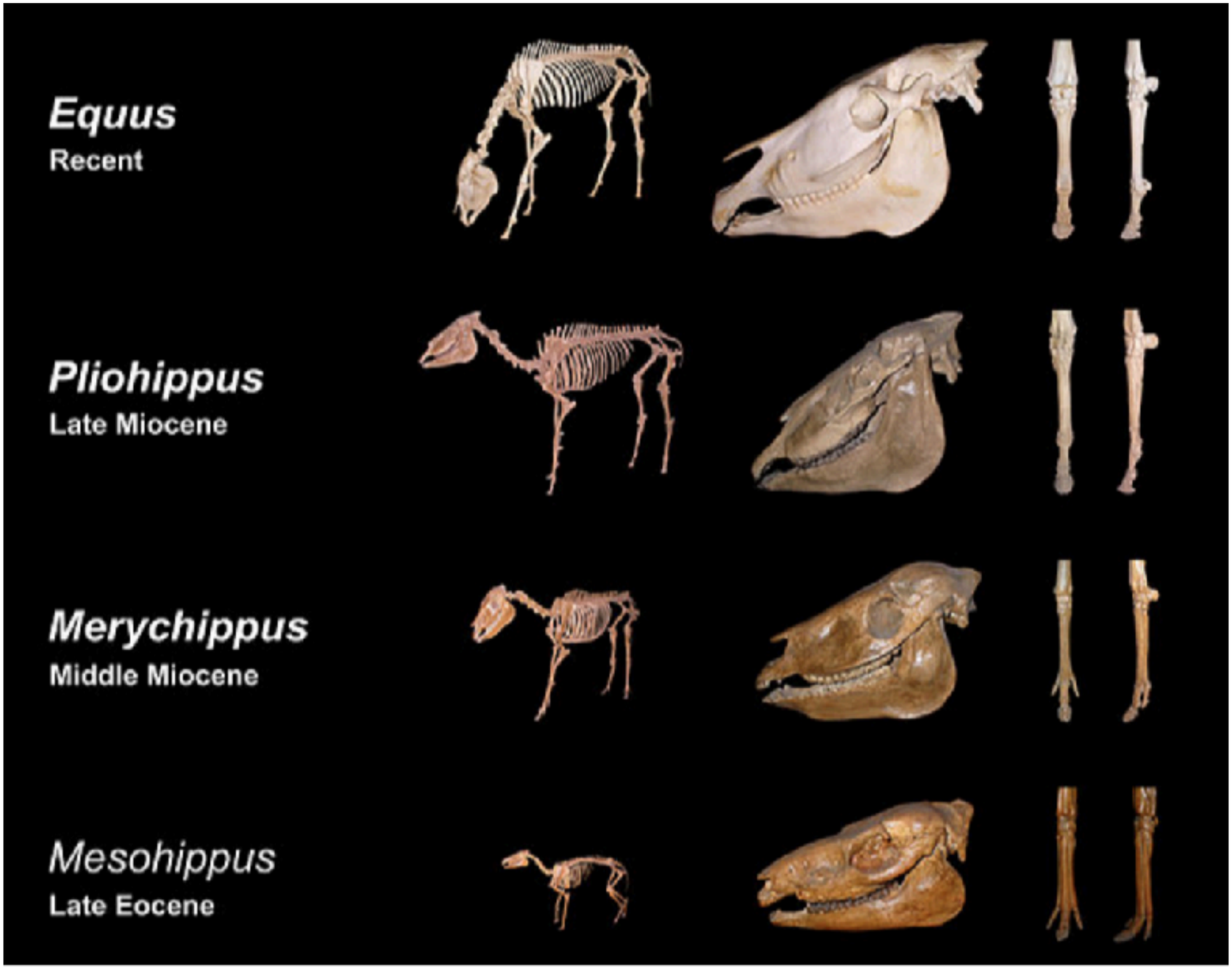
Diversity of life: Darwin's finches

Diversity of life

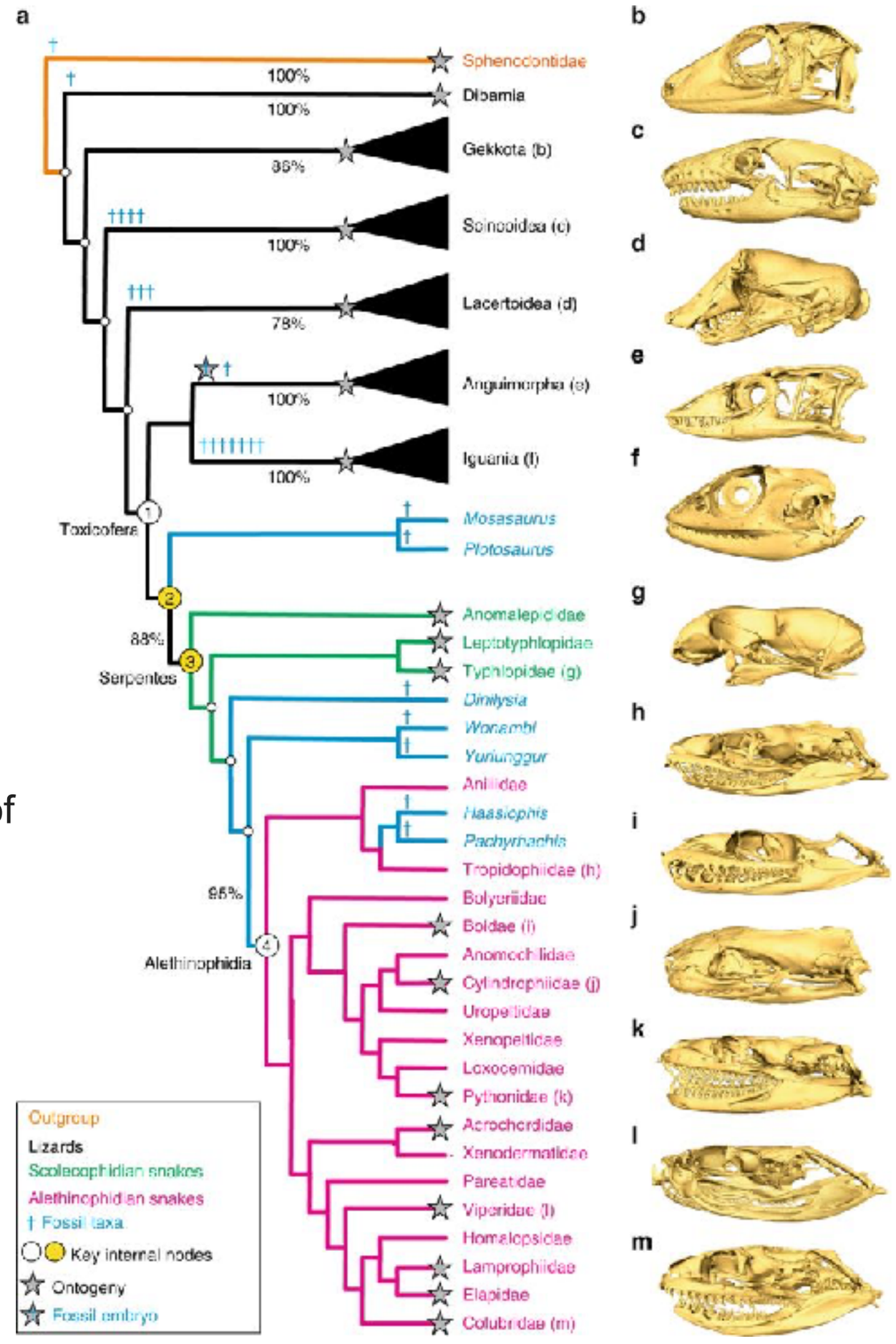


- How did the diversity of species' traits evolve?
- How did these traits first come to be?
- How fast did these traits change?
- How can they explain the diversity we see on earth today?

Shapes document change in morphology through time



Using shapes for evolution of snakes

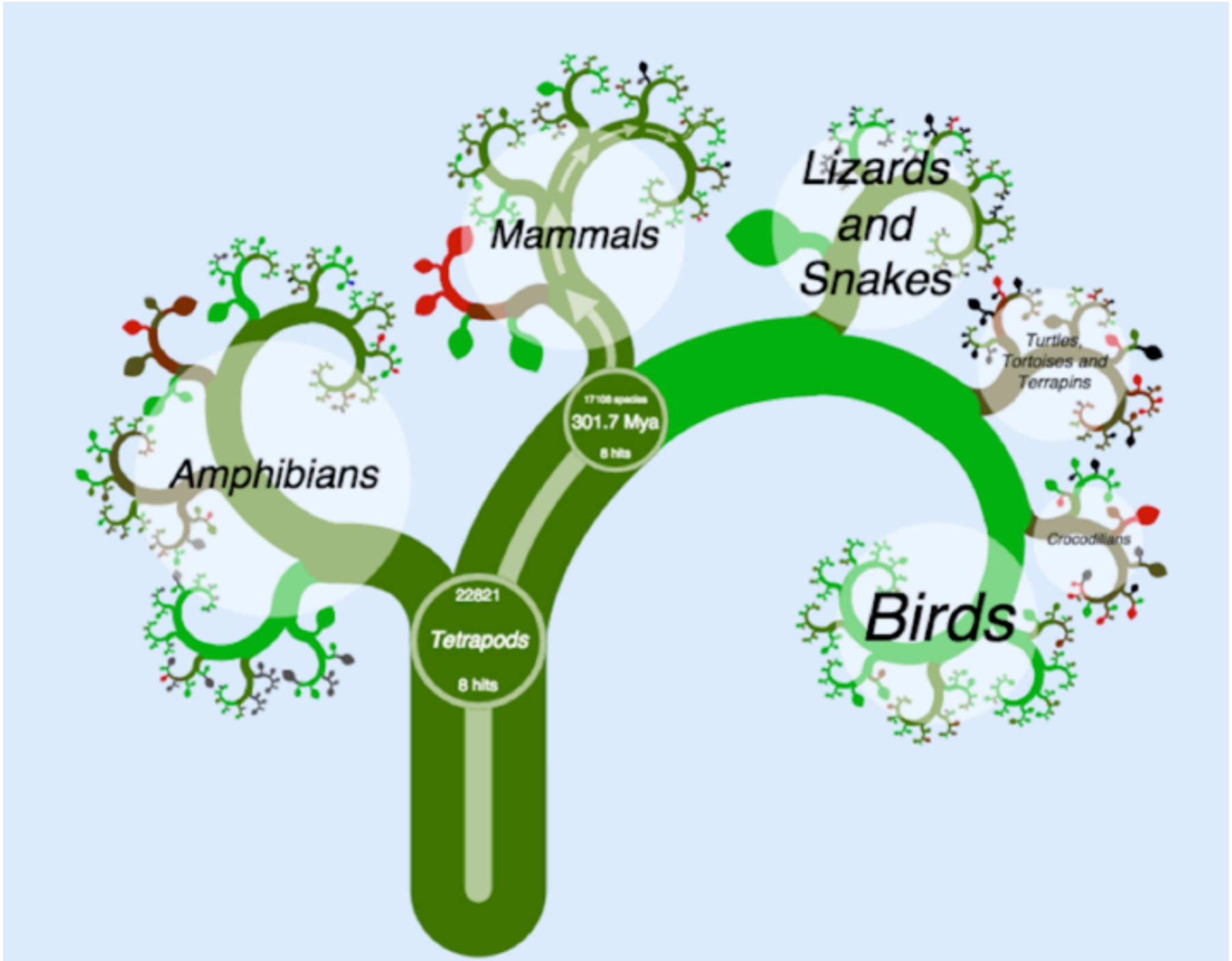


Da Silva, Filipe O., et al. "The ecological origins of snakes as revealed by skull evolution." *Nature communications* 9.1 (2018): 376.

Part I:

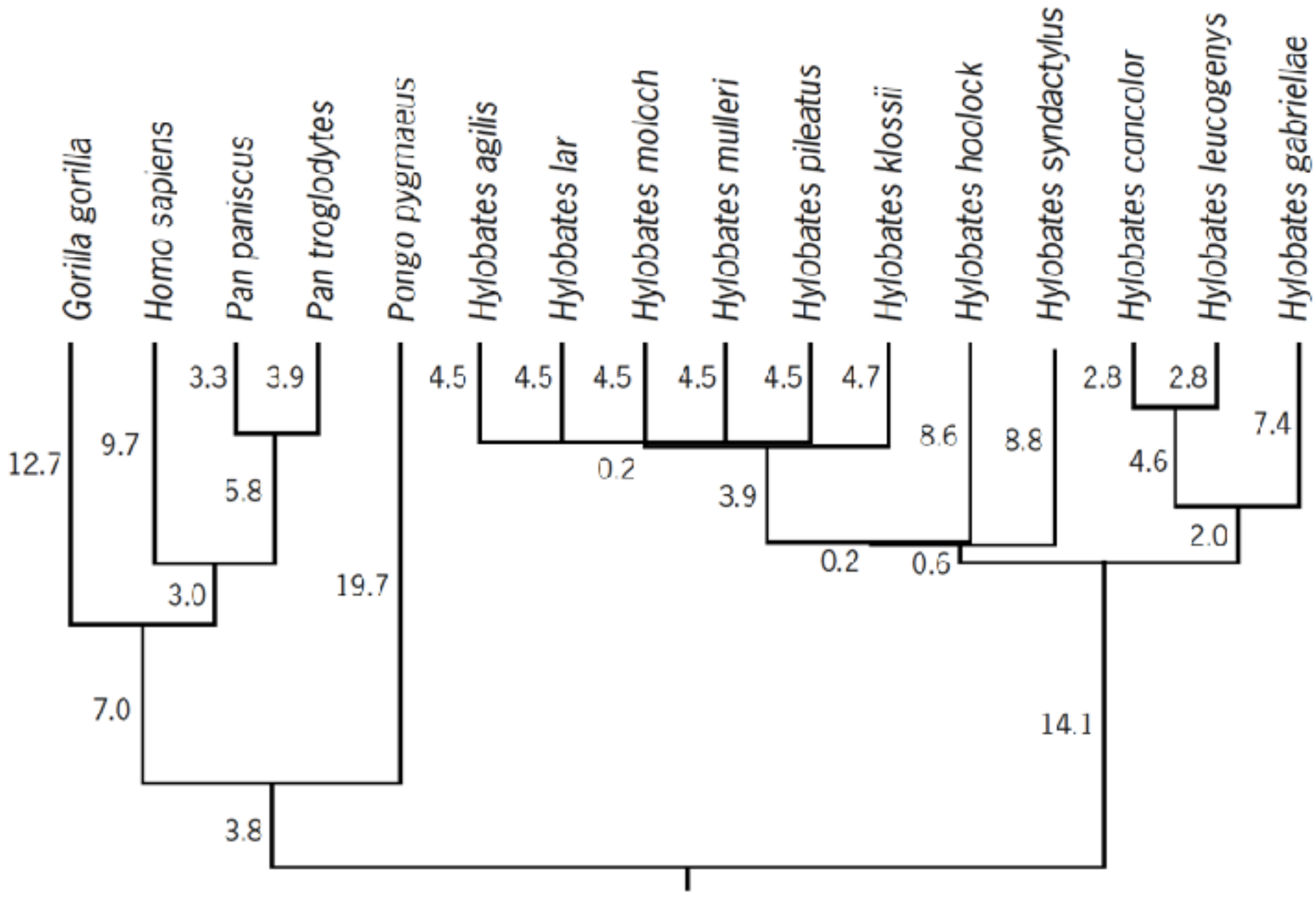
How do we model evolution?

Evolution: tree of life



Rosindell, J., and L. J. Harmon. 2012. OneZoom: A fractal explorer for the tree of life. PLoS Biol. 10:e1001406.

Phylogenetic comparative methods

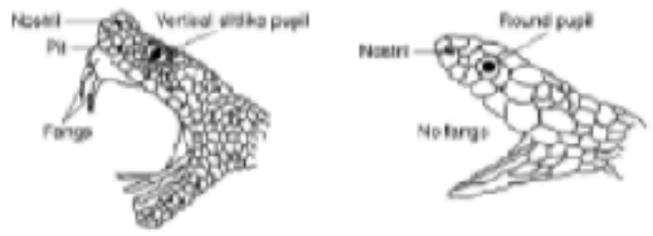


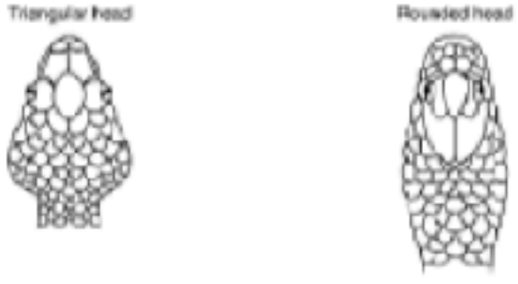
Lack of independence in traits due to shared evolutionary history.

Continuous trait is a real **number**

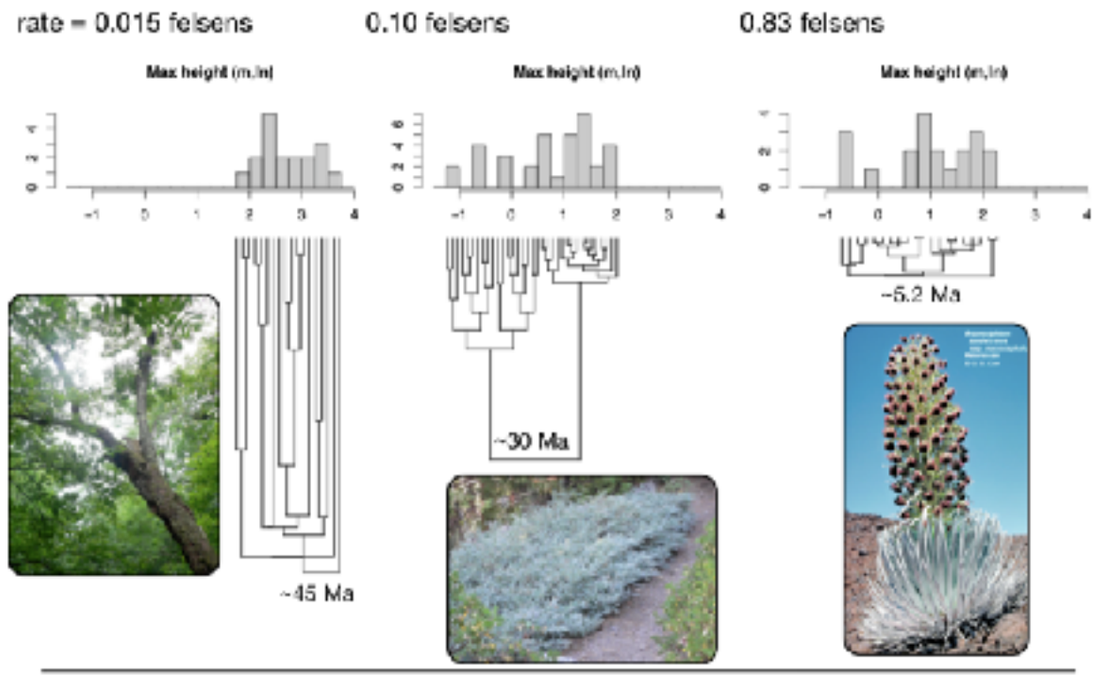
Pit Viper

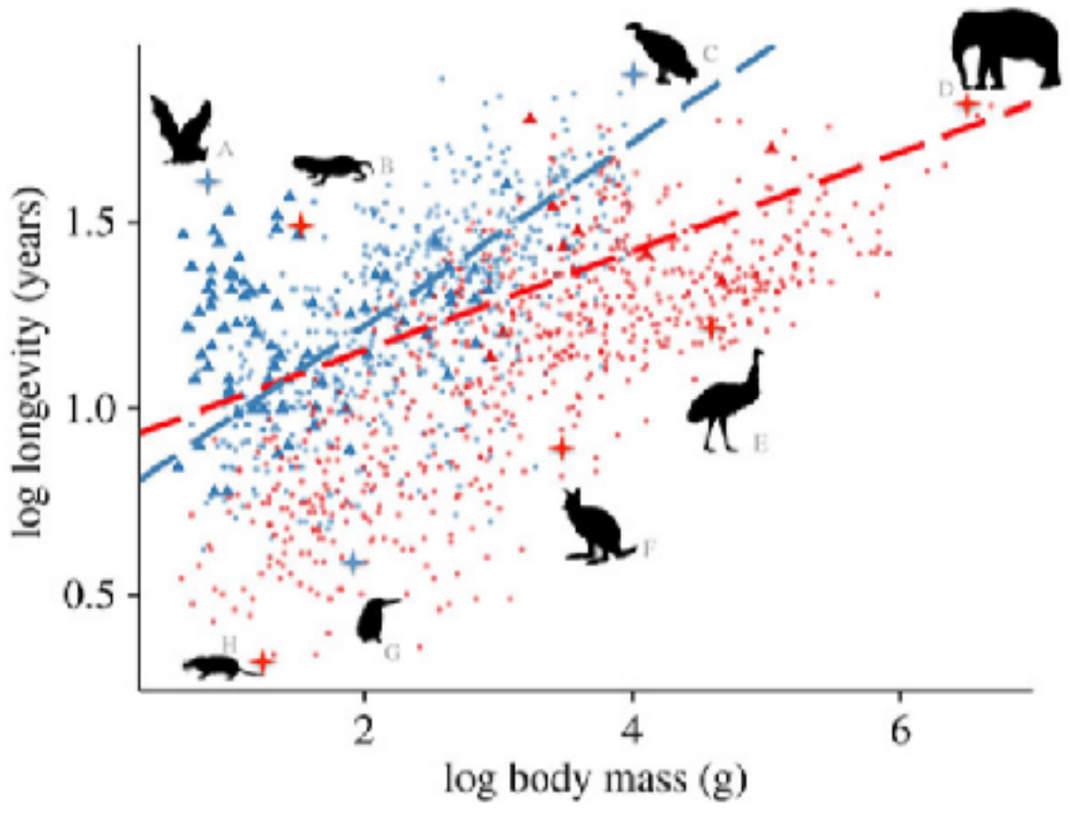
Nonvenemous Snake





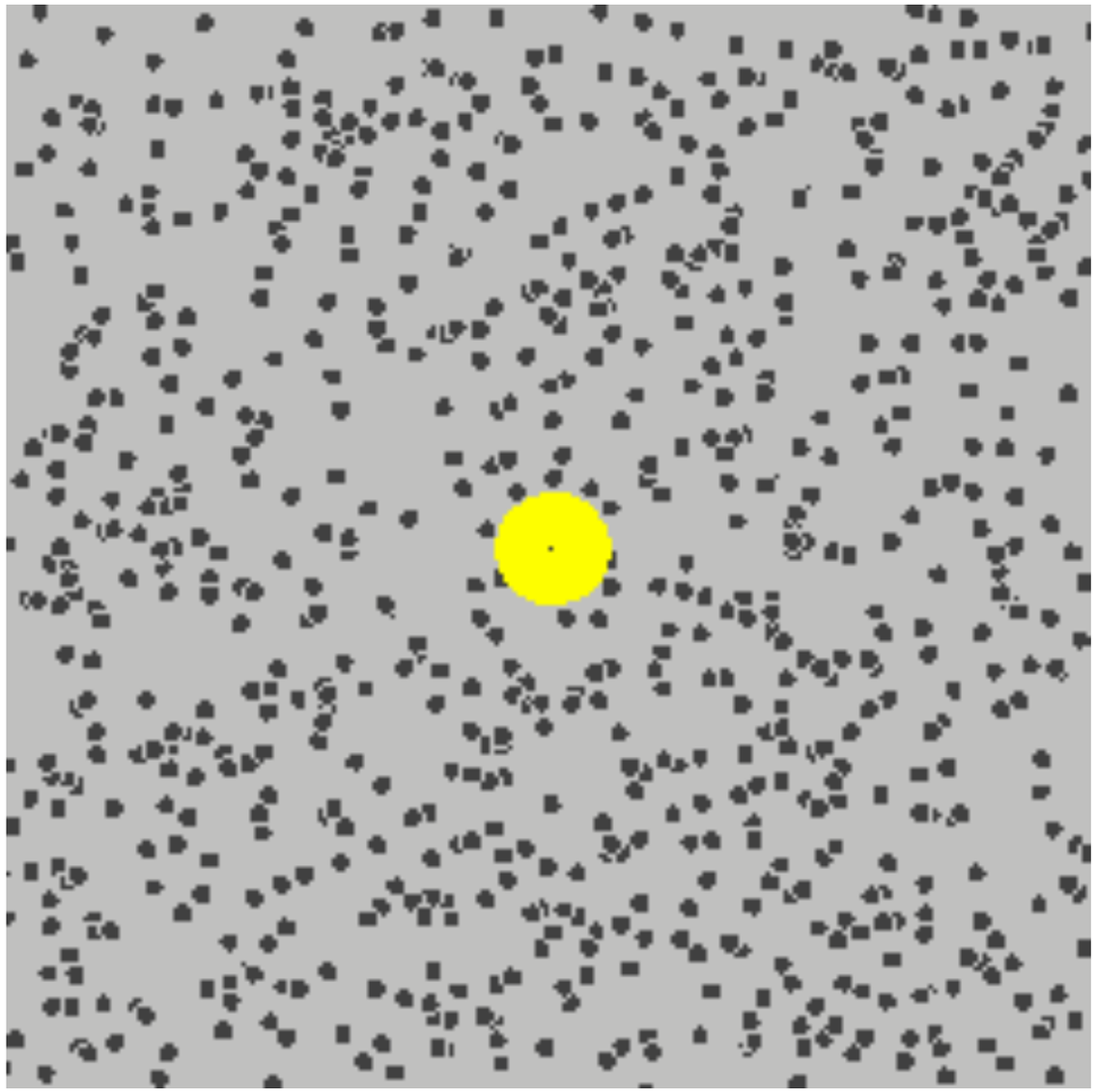
Pupil shape in snakes





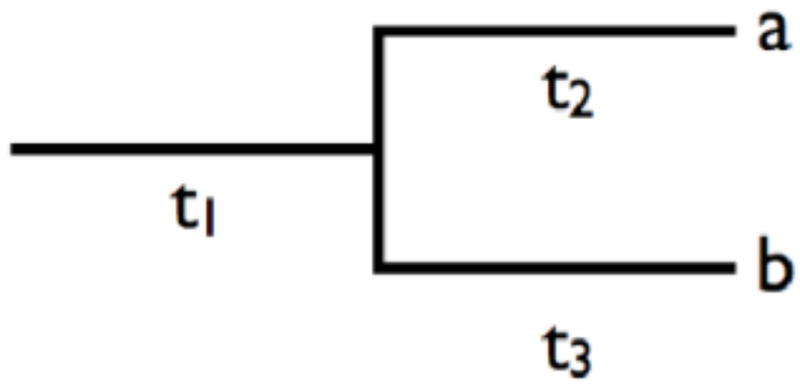
Brownian motion for continuous trait

- Example of Random walk
- Motion of object is due to sum of large number of very small random forces
- **Popular model in comparative biology**
- Nice statistical properties



Brownian motion on a phylogenetic tree

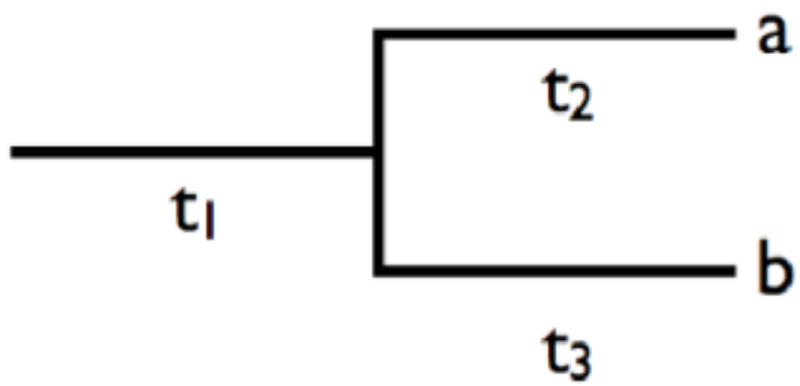
Simple tree with two species



Trait values as random variables

Brownian motion on a phylogenetic tree

Simple tree with two species

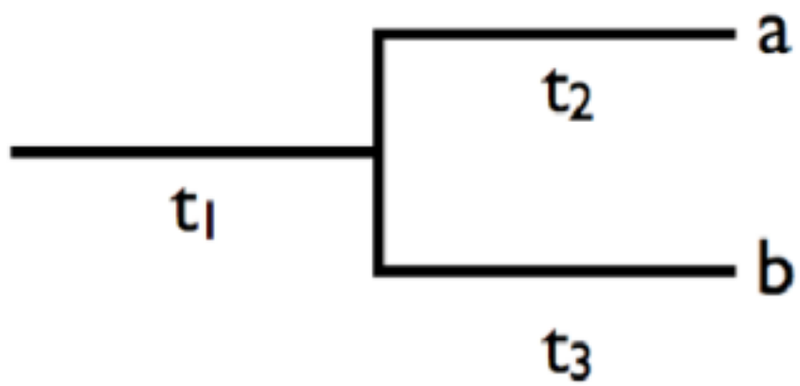


Trait values as random variables

$$\begin{bmatrix} Y_a \\ Y_b \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \sigma^2 \mathbf{V} \right)$$

Brownian motion on a phylogenetic tree

Simple tree with two species



Trait values as random variables

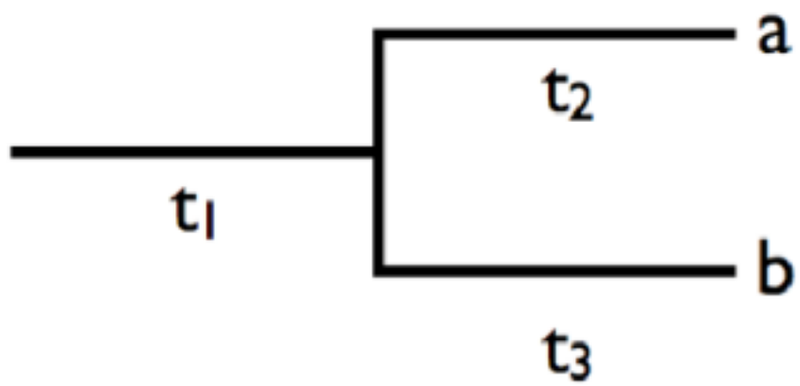
$$\begin{bmatrix} Y_a \\ Y_b \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \sigma^2 \mathbf{V} \right)$$

Variance-Covariance matrix

$$\mathbf{V} = \begin{bmatrix} t_1 + t_2 & t_1 \\ t_1 & t_1 + t_3 \end{bmatrix}$$

Brownian motion on a phylogenetic tree

Simple tree with two species



Trait values as random variables

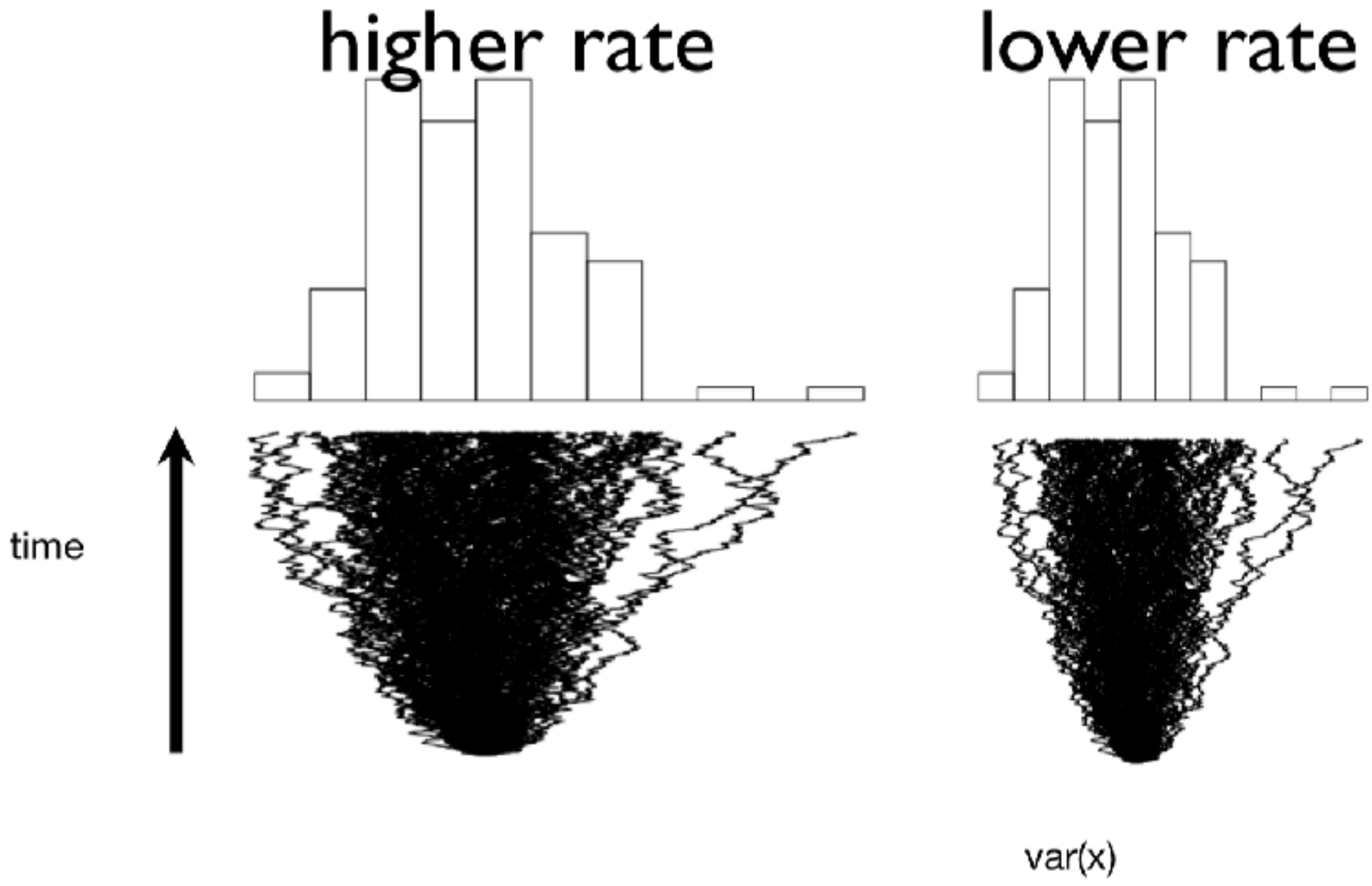
$$\begin{bmatrix} Y_a \\ Y_b \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \sigma^2 \mathbf{V} \right)$$

Evolution rate

Variance-Covariance matrix

$$\mathbf{V} = \begin{bmatrix} t_1 + t_2 & t_1 \\ t_1 & t_1 + t_3 \end{bmatrix}$$

Rate of Evolution σ^2



Finding σ^2 is a parameter tuning problem

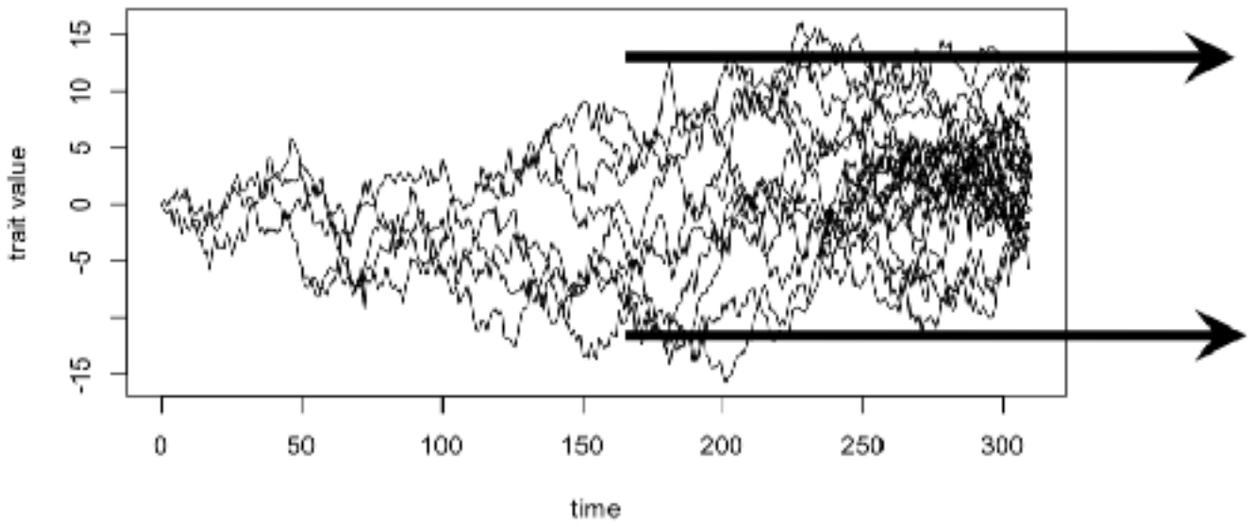
- Maximum Likelihood:
  $P(Y \mid \sigma^2)$ = probability of getting the data Y (observed trait values) given a value for σ^2
  software: **phytools**
- Bayesian methods:
  $P(\sigma^2 \mid Y)$ = posterior of σ^2 after observing the data Y
  software: **MRBAYES**

Liam J. Revell 2012. **phytools**: an R package for phylogenetic comparative biology (and other things)

John P. Huelsenbeck, Fredrik Ronquist, **MRBAYES**: Bayesian inference of phylogenetic trees , *Bioinformatics*, Volume 17, Issue 8, August 2001

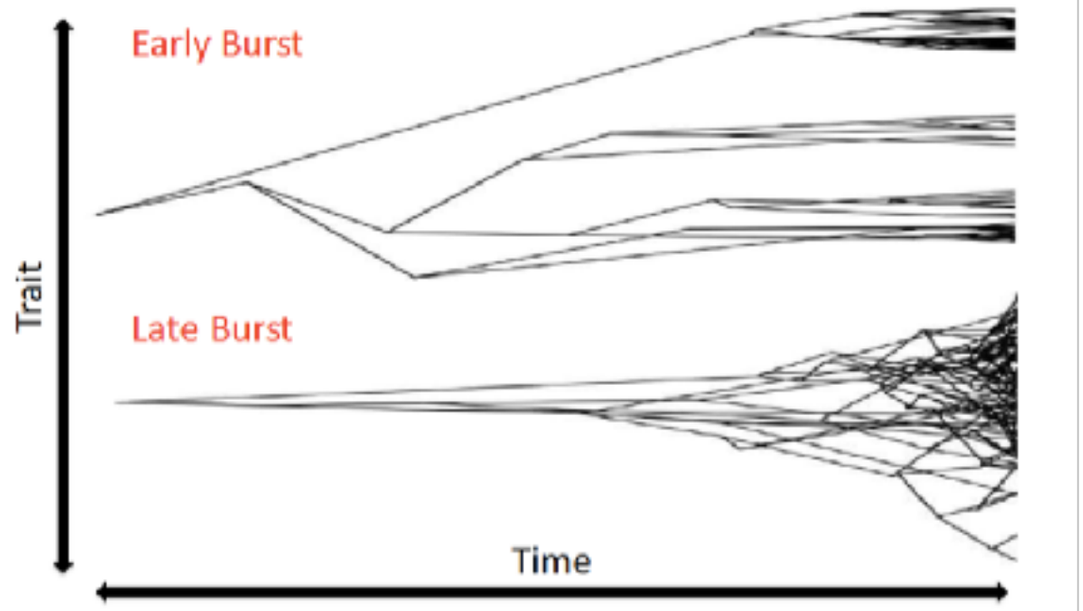
Other evolution models

- Ornstein-Uhlenbeck (**OU**)



Evolution has a tendency to move towards some medial value.

- Early Burst (**EB**)



Early burst models the rate of evolution is slowing through time

Why these three?

- Brownian Motion (**BM**) is assumed by almost all phylogenetic comparative methods
- Ornstein-Uhlenbeck (**OU**) may capture the importance of constraints on evolution
- Early Burst (**EB**) corresponded to one idea of adaptive radiation

BM, OU, EB differ only in covariance matrix

- Ornstein-Uhlenbeck (**OU**)

$$V_{ij} = \frac{\sigma^2}{\alpha} e^{-2\alpha(T - s_{ij})}(1 - e^{-2\alpha s_{ij}})$$

Three parameters:

- rate (σ)
- optimal value (θ)
- pull towards optimal (α)

T is depth of the tree.

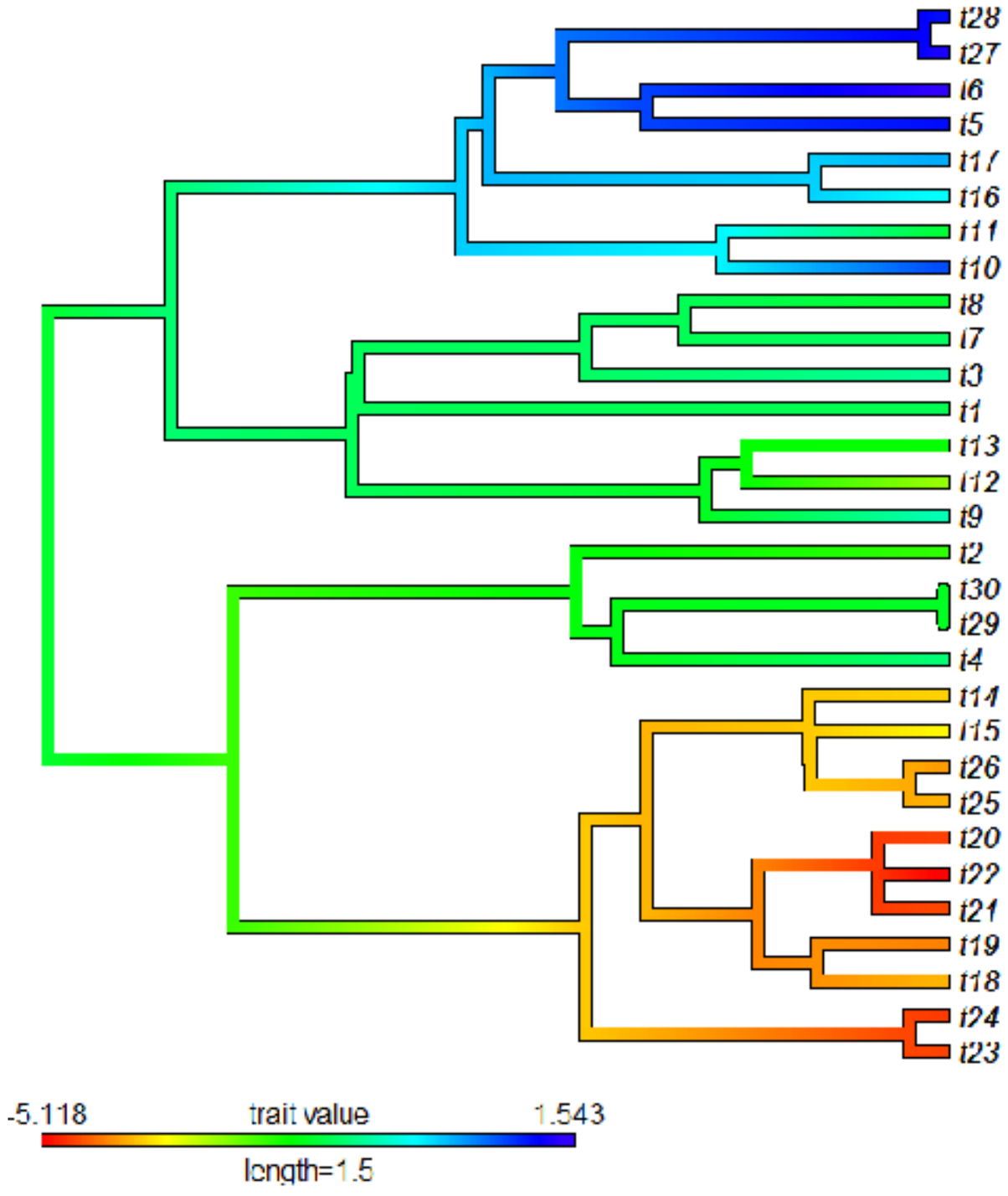
- Early Burst (**EB**)

$$V_{ij} = \int_0^{s_{ij}} \sigma_0^2 e^{rt}\, dt = \sigma_0^2 \frac{e^{r s_{ij}}}{r}$$

Two parameters:

- starting rate (σ_0)
- rate change (r)

Application: ancestral state reconstruction



Given

- a phylogenetic tree
- trait values
- an evolution model

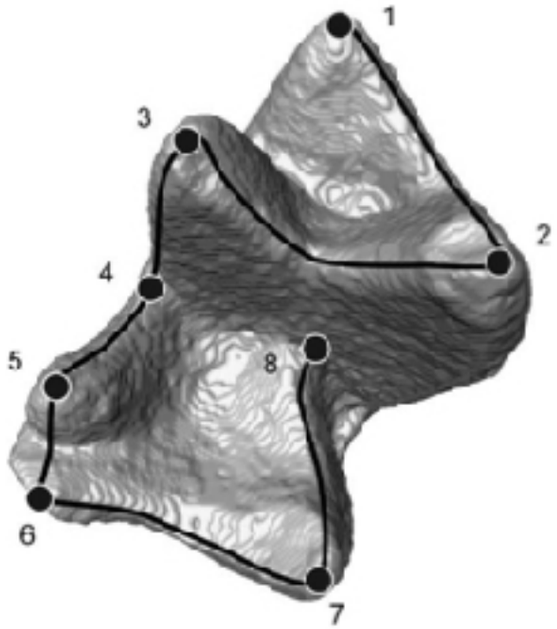
Goal

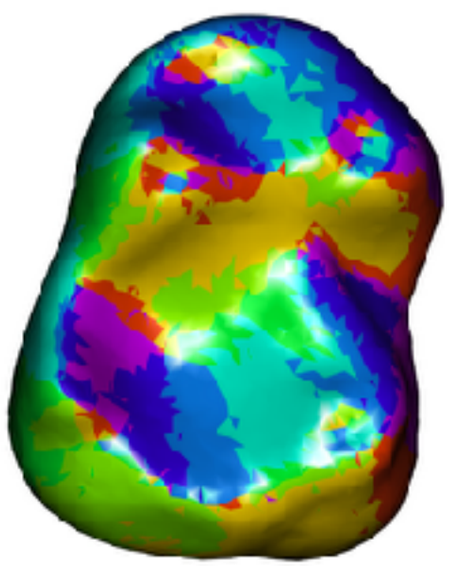
to obtain estimates for trait value at the ancestral nodes or along the branches of phylogeny

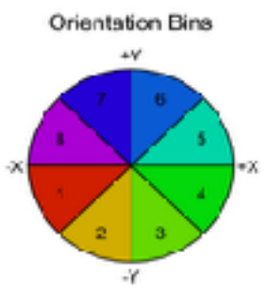
Part I:

How to study evolution on shapes?

Features on shapes



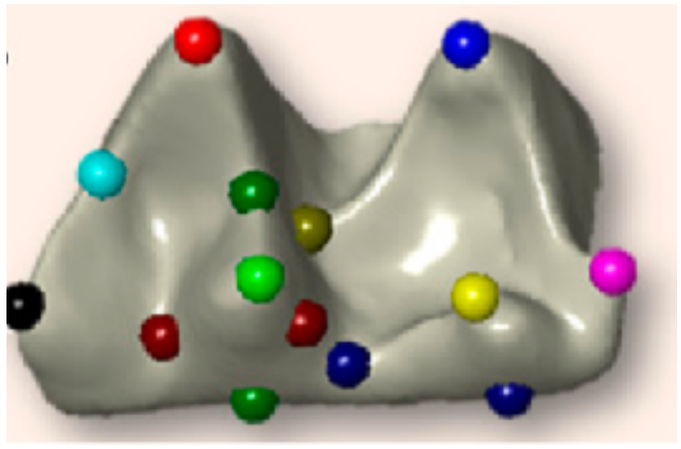


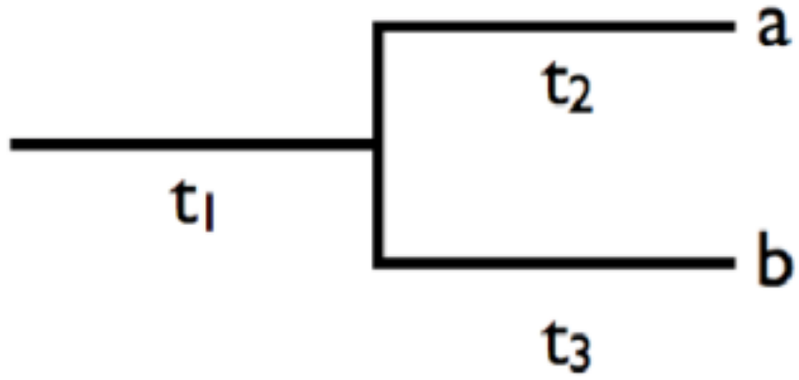


Shape landmarks are multivariate traits

a

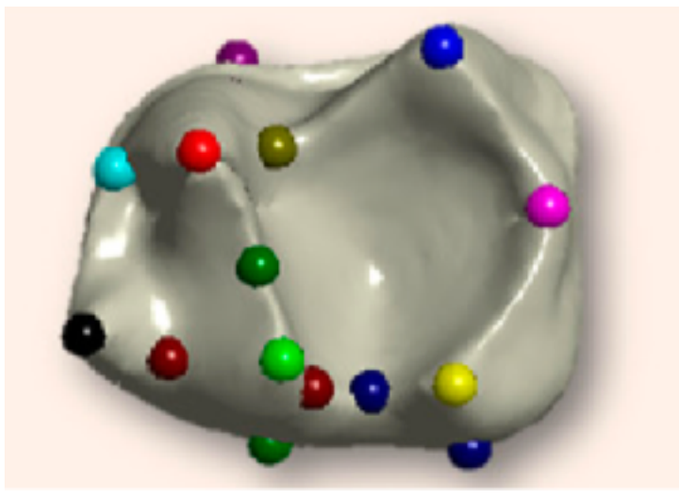
$(x_1^a, y_1^a, z_1^a) \quad (x_2^a, y_2^a, z_2^a)$





b

$(x_1^b, y_1^b, z_1^b) \quad (x_1^b, y_1^b, z_1^b)$



$$a: \begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \end{bmatrix} \qquad b: \begin{bmatrix} y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix}$$

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

Stack the landmark coordinates into a long vector

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

vector of length 30

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

30 by 30 matrix

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

$$C = V \otimes R$$

2 by 2 matrix phylogenetic information

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

$$C = V \otimes R$$

15 by 15 matrix trait/landmark correlation

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

$$C = V \otimes R$$

Kronecker product

$$\begin{bmatrix} t_1 + t_2 & t_1 \\ t_1 & t_1 + t_3 \end{bmatrix} \otimes \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix}$$

$$\begin{bmatrix} t_1 + t_2 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} & t_1 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} \\ t_1 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} & t_1 + t_3 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} \end{bmatrix}$$

Multivariate evolution models

$$\begin{bmatrix} y_1^a \\ y_2^a \\ \vdots \\ y_{15}^a \\ y_1^b \\ y_2^b \\ \vdots \\ y_{15}^b \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

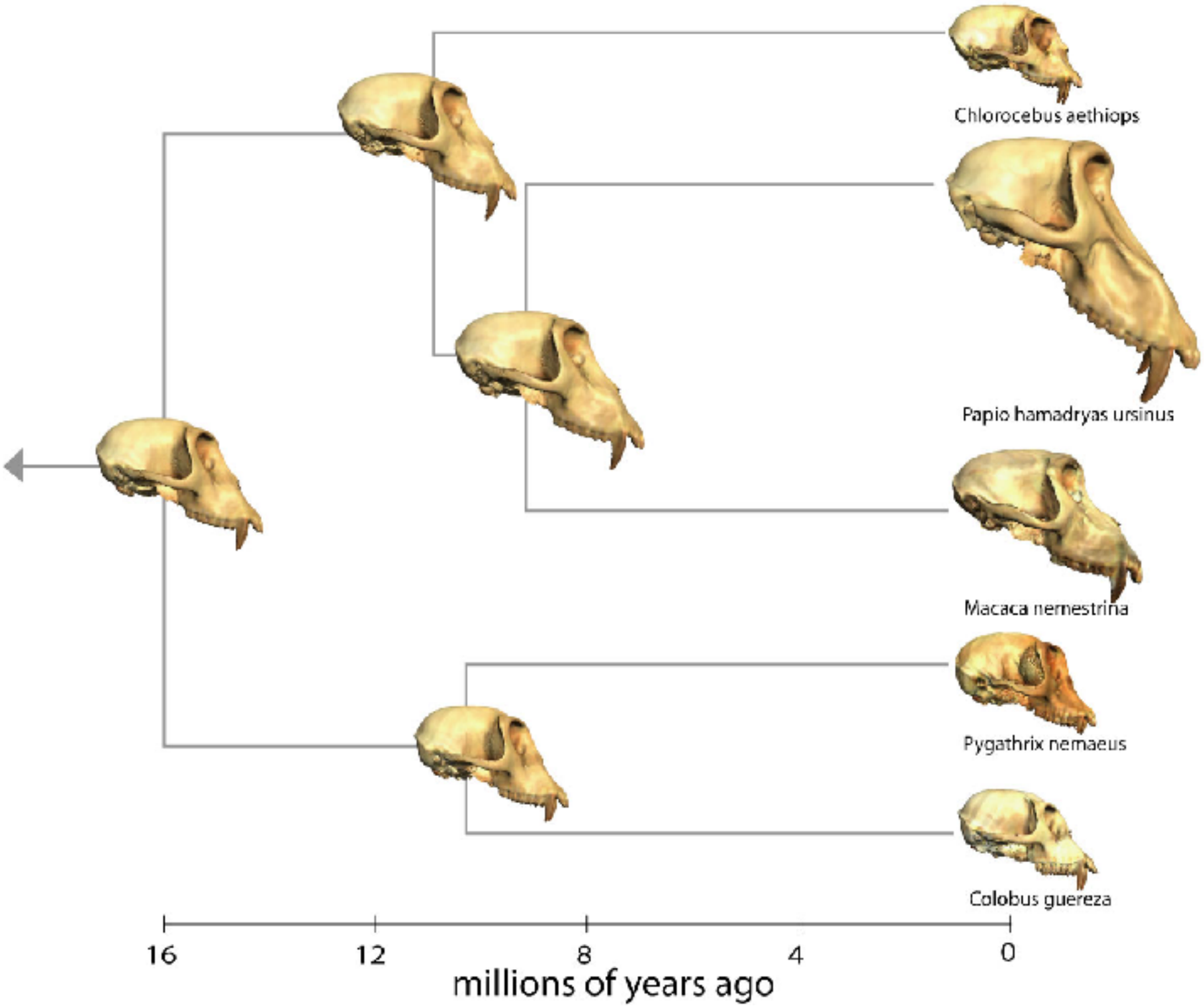
$$C = V \otimes R$$

Evolution ⊗ Trait

Dean C. Adams, Michael L. Collyer, **Multivariate Phylogenetic Comparative Methods**: Evaluations, Comparisons, and Recommendations, *Systematic Biology*

Adams, D. C. and Collyer, M. L. (2018), **Phylogenetic ANOVA**: Group-clade aggregation, biological challenges, and a refined permutation procedure. Evolution

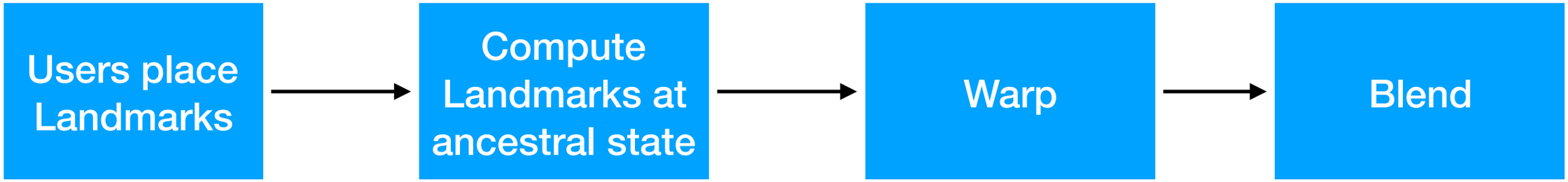
Application: reconstruct ancestral surface

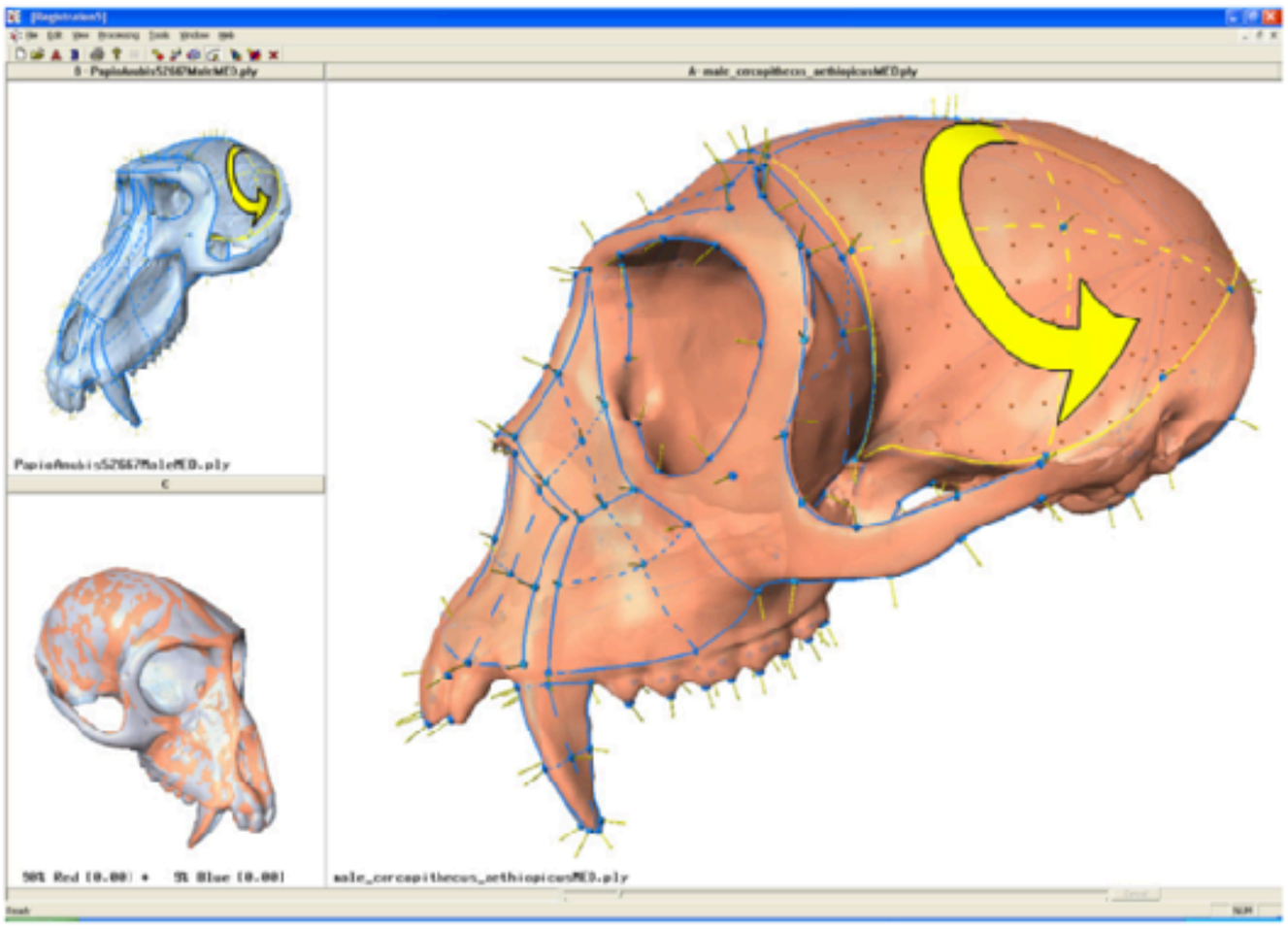


Wiley, David F., et al. "Evolutionary morphing." *VIS 05. IEEE Visualization, 2005.*. IEEE, 2005.

Application: reconstruct ancestral surface

Wiley, David F., et al. "Evolutionary morphing." *VIS 05. IEEE Visualization, 2005.*. IEEE, 2005.





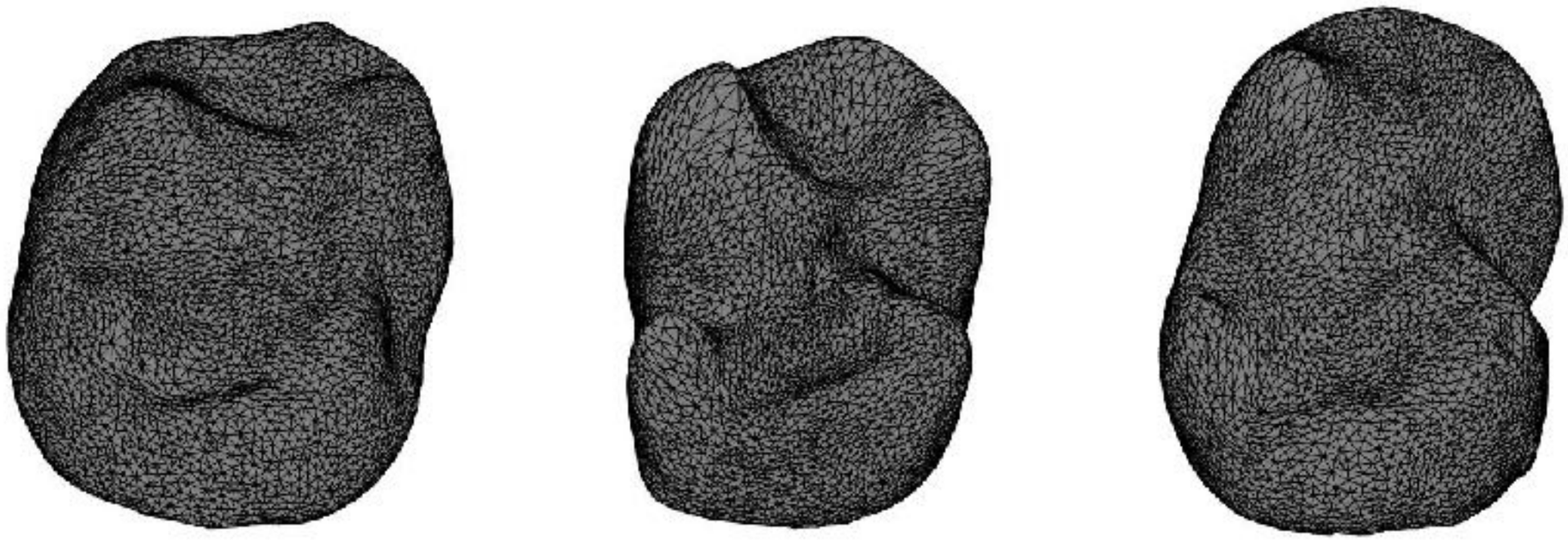
Manually putting landmarks are time consuming and requires domain knowledge.

We often require **many** landmarks.

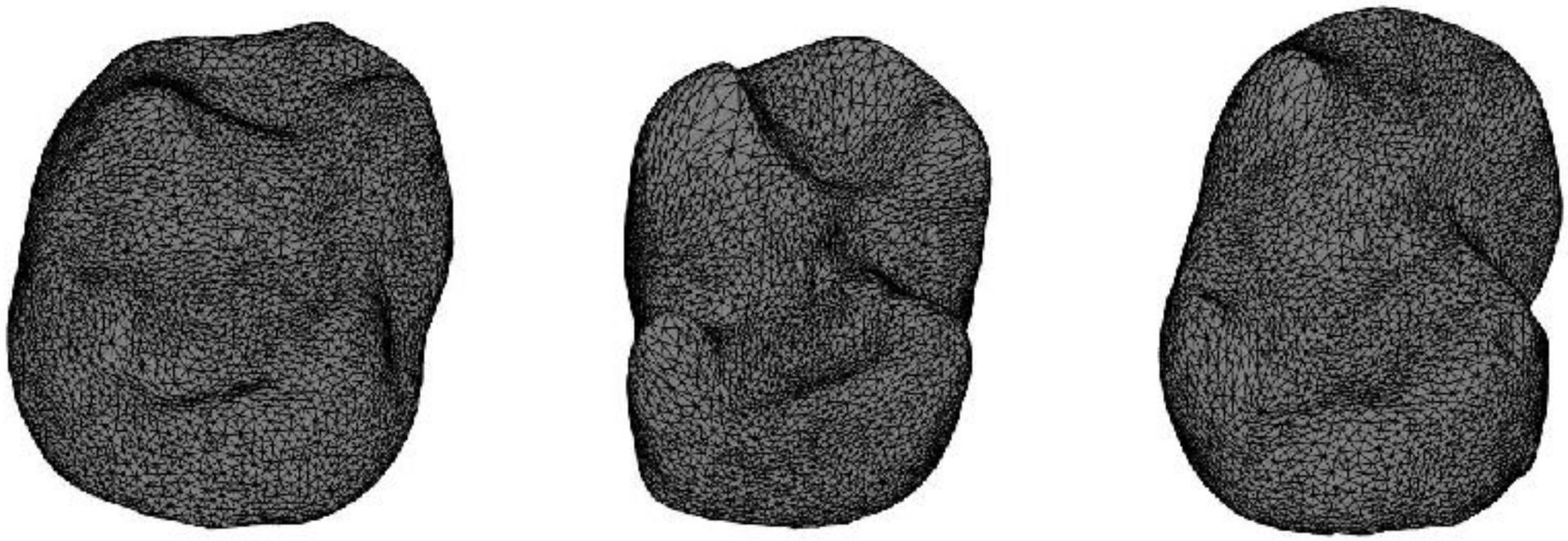
Part II:

How to study evolution without landmarks?

Aligned shapes from Auto3dgm



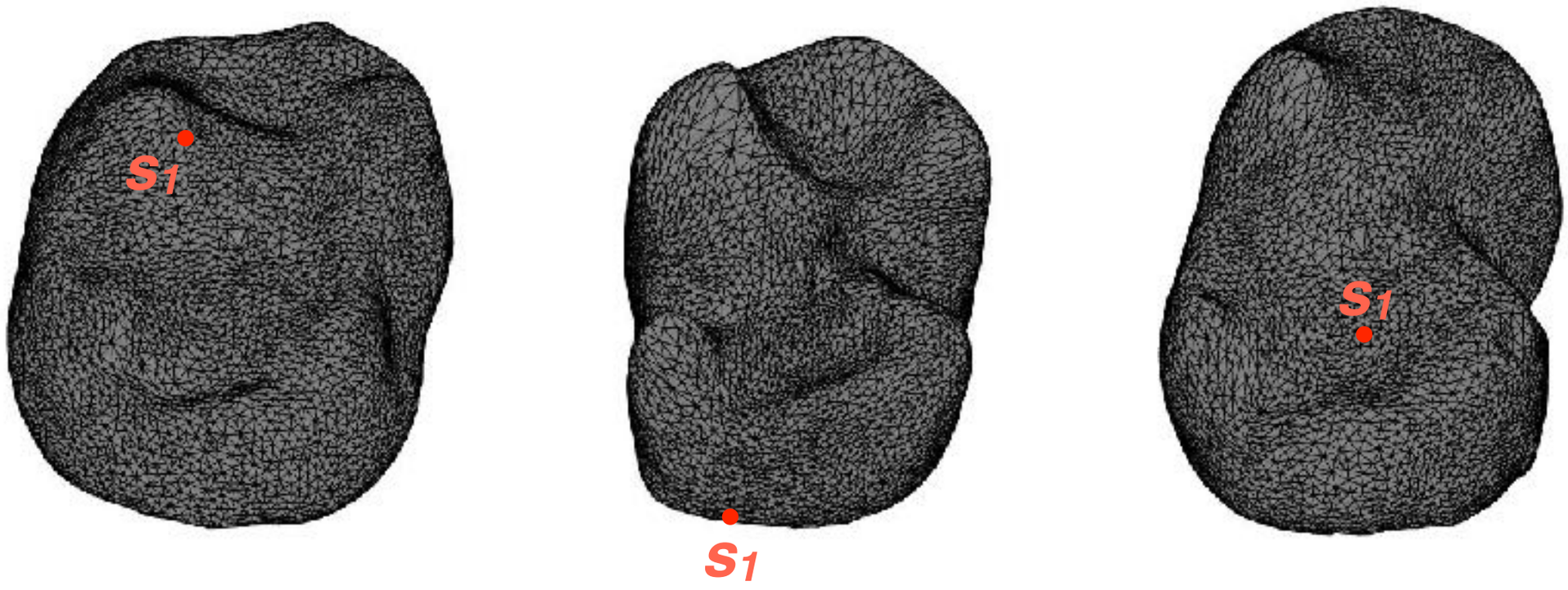
Challenges of studying evolution with the entire shape



5000 points ***4893 points*** ***4998 points***

Not all shapes are represented by the same number of points.

Challenges of studying evolution with the entire shape



The points are stored in random order.

(The points are not in correspondence.)

A naive model suffers from the curse of high dimensionality

$$\begin{bmatrix} y_1^a \\ \vdots \\ y_{5000}^a \\ y_1^b \\ \vdots \\ y_{4893}^b \\ y_1^c \\ \vdots \\ y_{4998}^c \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

14,891 by 14,891 matrix

A naive model suffers from the curse of high dimensionality

$$\begin{bmatrix} y_1^a \\ \vdots \\ y_{5000}^a \\ y_1^b \\ \vdots \\ y_{4893}^b \\ y_1^c \\ \vdots \\ y_{4998}^c \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

14,891 by 14,891 matrix

- Overfitting

  $14891 \times 14891 \approx 2 \times 10^8$

  free parameters to tune

A naive model suffers from the curse of high dimensionality

$$\begin{bmatrix} y_1^a \\ \vdots \\ y_{5000}^a \\ y_1^b \\ \vdots \\ y_{4893}^b \\ y_1^c \\ \vdots \\ y_{4998}^c \end{bmatrix} \sim \mathcal{N}(\mu, \mathbf{C})$$

14,891 by 14,891 matrix

- Overfitting
  $14891 \times 14891 \approx 2 \times 10^8$
  free parameters to tune.

- Where to put phylogenetic information?

Covariance matrix has a block structure

$$
\begin{bmatrix}
C(y_1^a, y_1^a) & \cdots & C(y_1^a, y_{5000}^a) & C(y_1^a, y_1^b) & \cdots & C(y_1^a, y_{4893}^b) & C(y_1^a, y_1^c) & \cdots & C(y_1^a, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{5000}^a, y_1^a) & \cdots & C(y_{5000}^a, y_{5000}^a) & C(y_{5000}^a, y_1^b) & \cdots & C(y_{5000}^a, y_{4893}^b) & C(y_{5000}^a, y_1^c) & \cdots & C(y_{5000}^a, y_{4998}^c) \\
C(y_1^b, y_1^a) & \cdots & C(y_1^b, y_{5000}^a) & C(y_1^b, y_1^b) & \cdots & C(y_1^b, y_{4893}^b) & C(y_1^b, y_1^c) & \cdots & C(y_1^b, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{4893}^b, y_1^a) & \cdots & C(y_{4893}^b, y_{4893}^b) & C(y_{4893}^b, y_1^b) & \cdots & C(y_{4893}^b, y_{4893}^b) & C(y_{4893}^b, y_1^c) & \cdots & C(y_{4893}^b, y_{4998}^c) \\
C(y_1^c, y_1^a) & \cdots & C(y_1^c, y_{5000}^a) & C(y_1^c, y_1^b) & \cdots & C(y_1^c, y_{4893}^b) & C(y_1^c, y_1^c) & \cdots & C(y_1^c, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{4998}^c, y_1^a) & \cdots & C(y_{4998}^c, y_{4998}^c) & C(y_{4998}^c, y_1^b) & \cdots & C(y_{4998}^c, y_{4893}^b) & C(y_{4998}^c, y_1^c) & \cdots & C(y_{4998}^c, y_{4998}^c)
\end{bmatrix}
$$

The block (i,j) indicates correlation between shapes i and j.
The entry $C(y_n^i, y_m^j)$ indicates correlation between the n-th point on shape i and and the m-th point on shape j.

Covariance matrix has a block structure

$$
\begin{bmatrix}
C(y_1^a, y_1^a) & \cdots & C(y_1^a, y_{5000}^a) & C(y_1^a, y_1^b) & \cdots & C(y_1^a, y_{4893}^b) & C(y_1^a, y_1^c) & \cdots & C(y_1^a, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{5000}^a, y_1^a) & \cdots & C(y_{5000}^a, y_{5000}^a) & C(y_{5000}^a, y_1^b) & \cdots & C(y_{5000}^a, y_{4893}^b) & C(y_{5000}^a, y_1^c) & \cdots & C(y_{5000}^a, y_{4998}^c) \\
C(y_1^b, y_1^a) & \cdots & C(y_1^b, y_{5000}^a) & C(y_1^b, y_1^b) & \cdots & C(y_1^b, y_{4893}^b) & C(y_1^b, y_1^c) & \cdots & C(y_1^b, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{4893}^b, y_1^a) & \cdots & C(y_{4893}^b, y_{4893}^b) & C(y_{4893}^b, y_1^b) & \cdots & C(y_{4893}^b, y_{4893}^b) & C(y_{4893}^b, y_1^c) & \cdots & C(y_{4893}^b, y_{4998}^c) \\
C(y_1^c, y_1^a) & \cdots & C(y_1^c, y_{5000}^a) & C(y_1^c, y_1^b) & \cdots & C(y_1^c, y_{4893}^b) & C(y_1^c, y_1^c) & \cdots & C(y_1^c, y_{4998}^c) \\
\vdots & & & & & & & & \\
C(y_{4998}^c, y_1^a) & \cdots & C(y_{4998}^c, y_{4998}^c) & C(y_{4998}^c, y_1^b) & \cdots & C(y_{4998}^c, y_{4893}^b) & C(y_{4998}^c, y_1^c) & \cdots & C(y_{4998}^c, y_{4998}^c)
\end{bmatrix}
$$

$$
\begin{bmatrix}
t_1 + t_2 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} & t_1 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} \\
t_1 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix} & t_1 + t_3 \begin{bmatrix} \delta_{1,1} & \delta_{1,2} & \cdots & \delta_{1,15} \\ \vdots & & & \vdots \\ \delta_{15,1} & \delta_{15,2} & \cdots & \delta_{15,15} \end{bmatrix}
\end{bmatrix}
$$

Geometrically inspired covariance matrix

Let $y_n^i = (z_i, u_n)$. We define the covariance matrix by

$$C(y_n^i, y_m^j) = C(z_i, u_n; z_j, u_m) = V(z_i, z_j) \cdot R(u_n, u_m)$$

Evolution

Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

Geometrically inspired covariance matrix

Let $y_n^i = (z_i, u_n)$. We define the covariance matrix by

$$C(y_n^i, y_m^j) = C(z_i, u_n; z_j, u_m) = V(z_i, z_j) \cdot R(u_n, u_m)$$

Geometry

We define the geometry part by

$$R(u_n, u_m) = \exp\left(-\frac{d^2\left(u_n, u_m\right)}{t}\right)$$

Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

Geometrically inspired covariance matrix

Let $y_n^i = (z_i, u_n)$. We define the covariance matrix by

$$C(y_n^i, y_m^j) = C(z_i, u_n; z_j, u_m) = V(z_i, z_j) \cdot R(u_n, u_m)$$

Geometry

We define the geometry part by

$$R(u_n, u_m) = \exp\left(-\frac{d^2(u_n, u_m)}{t}\right)$$

To understand evolutionary process is equivalent to understand the parameters.

Statistical model for evolution on shapes

Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

Geometrically inspired covariance matrix

Let $y_n^i = (z_i, u_n)$. We define the covariance matrix by

$$C(y_n^i, y_m^j) = C(z_i, u_n; z_j, u_m) = V(z_i, z_j) \cdot R(u_n, u_m)$$

Geometry

We define the geometry part by

$$R(u_n, u_m) = \exp\left(-\frac{d^2(u_n, u_m)}{t}\right)$$

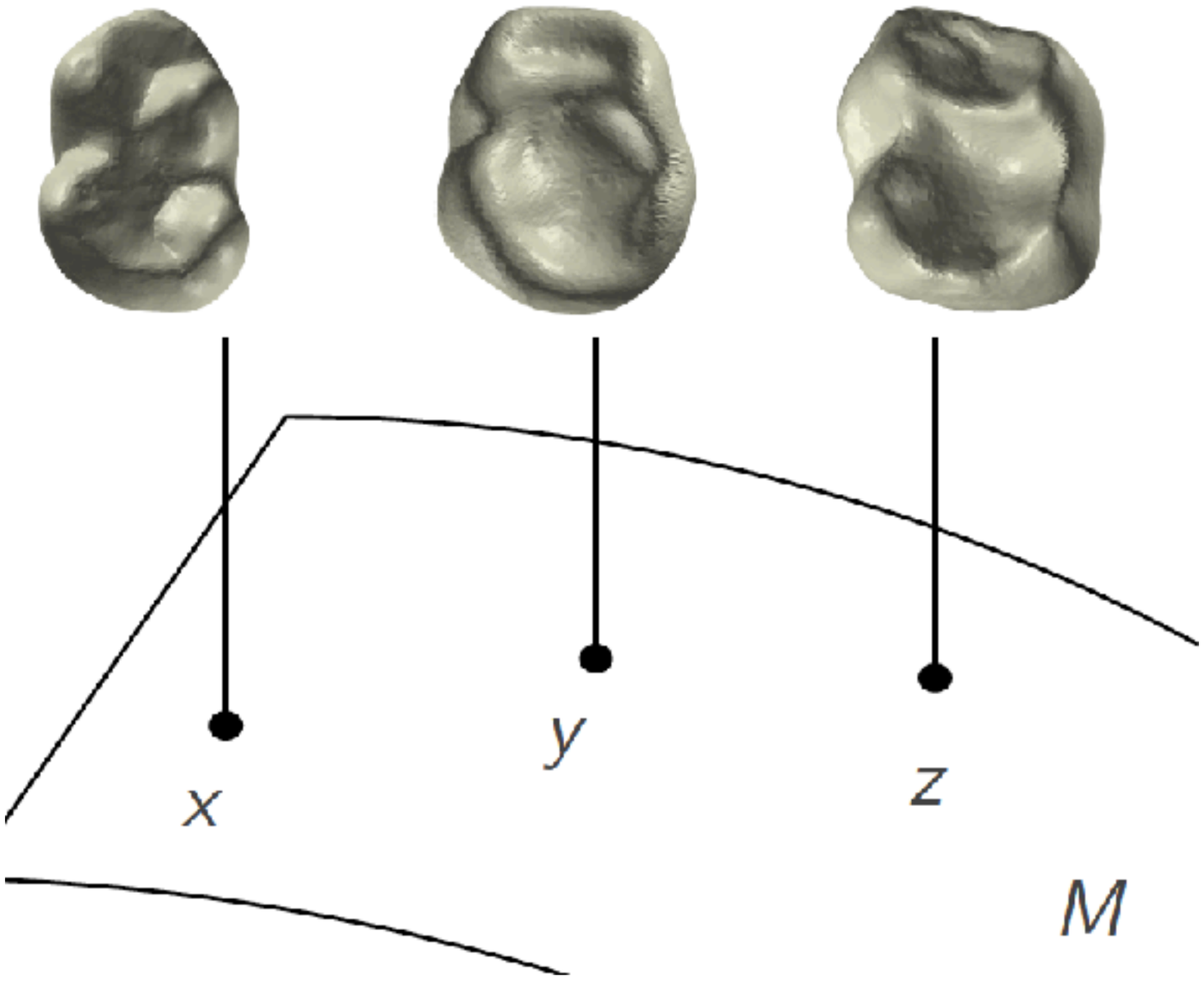
The total number of parameters is

(# of para in V) + 1

Significant dimension reduction

Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

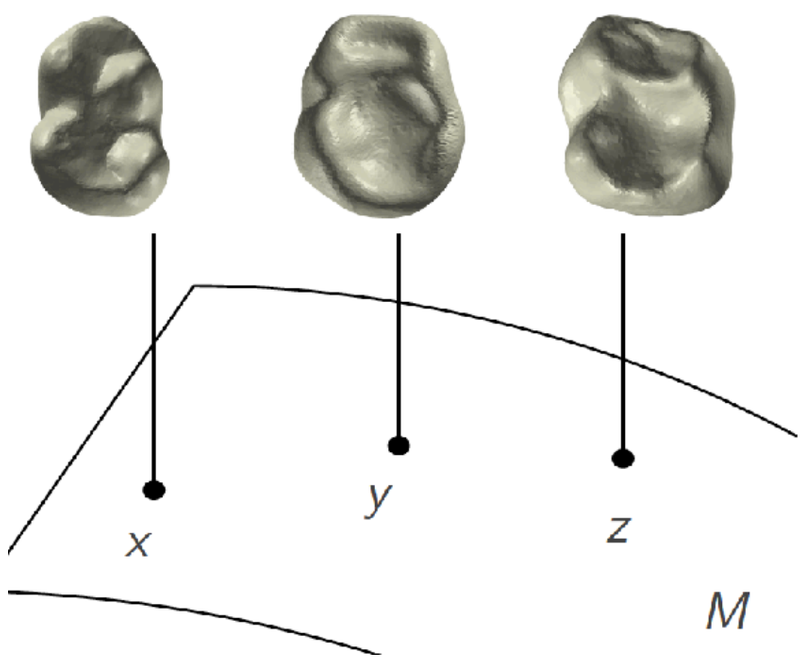
View the shape space as a fibre bundle



Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

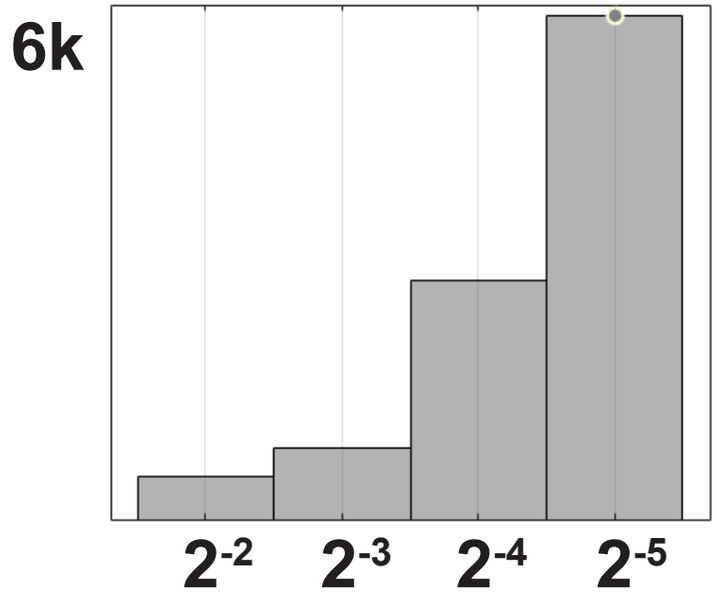
Covariance matrix is from diffusion operator on fibre bundles

- Diffusion operator gives rise to a geometric prior
- Computational advantage
- Natural hierarchical structure in fibre bundles



Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

Do lemurs of Madagascar follow an EB model?



r close to zero means less likely to be an EB model

Shan Shan, “Probabilistic Models on Fibre bundles.” Ph.D. Thesis, Duke University (2019)

Application 5: Phylogenetics

How do we study the evolutionary process without landmarks?

- Traditional phylogenetic comparative methods **with** landmarks
  - Single-variate and multivariate model
  - Features and landmarks for shape evolution
- Study evolution with aligned shapes and no landmarks
  - Fibre bundle approach

Future work

Future work