CONFIDENCE SETS FOR PHYLOGENETIC TREES

Amy Willis

~~Department of Statistical Science, Cornell University~~

Department of Biostatistics, University of Washington

RESEARCH INTERESTS

- Collaborative work
  - Bataille, Willis, Yang & Liu (2017). **Continental Crust Composition: A Major Control of Past Global Weathering Rates.** *Science Advances.*
  - Chan, Willis, Kornhauser, Vadhat, et al. (2016). **Influencing the tumor microenvironment.** *Clinical Cancer Research.*
  - Vanden Brink, Willis, Jarrett & Lujan (2015). **Sonographic markers of ovarian morphology.** *Fertility & Sterility.*
  - Christ, Willis, Brooks, et al. (2014). **Follicle number represents the best ultrasonographic marker of PCOS.** *Fertility & Sterility.*

RESEARCH INTERESTS

- Microbial ecology
  - Willis, Bunge & Whitman (2017+). **Improved detection of changes in species richness in high-diversity microbial communities.** *JRSS-C.*
  - Willis (2016). **Extrapolating abundance curves has no predictive power for estimating microbial biodiversity.** *PNAS.*
  - Willis & Bunge (2015). **Estimating diversity via frequency ratios.** *Biometrics.*
  - Bunge, Willis & Walsh (2014). **Estimating the Number of Species in Microbial Diversity Studies.** *Ann Rev Stat.*

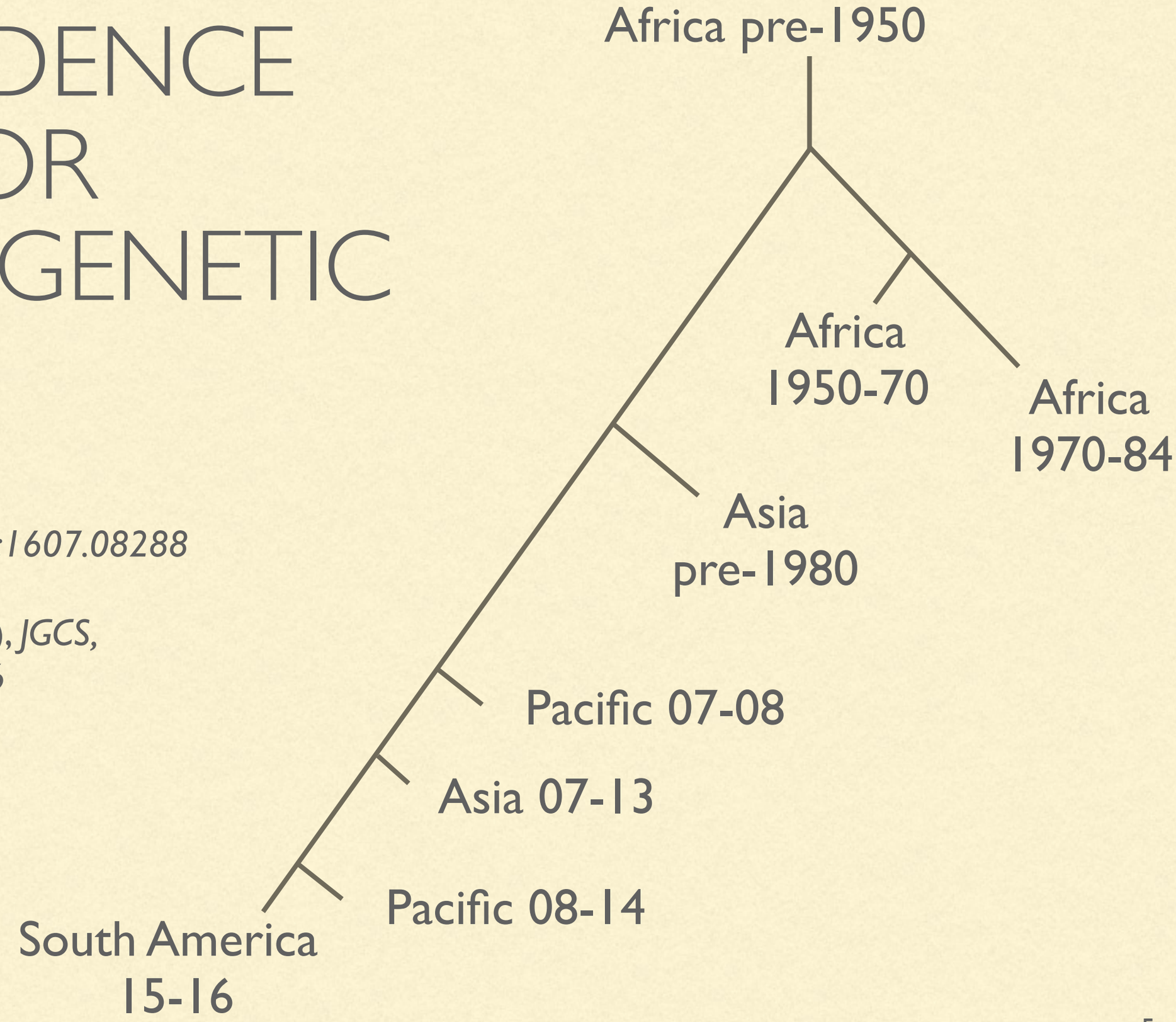
RESEARCH INTERESTS

- Statistical phylogenetics
  - RoyChoudhury, Willis & Bunge (2014). **Consistency of a phylogenetic tree maximum likelihood estimator.** *J Stat Plan Inference.*
  - Willis (2016+). **Confidence sets for phylogenetic trees.** *Under Review.*
  - Willis & Bell (2017+). **Uncertainty in phylogenetic tree estimates.** *J Comput Graph Stat.*

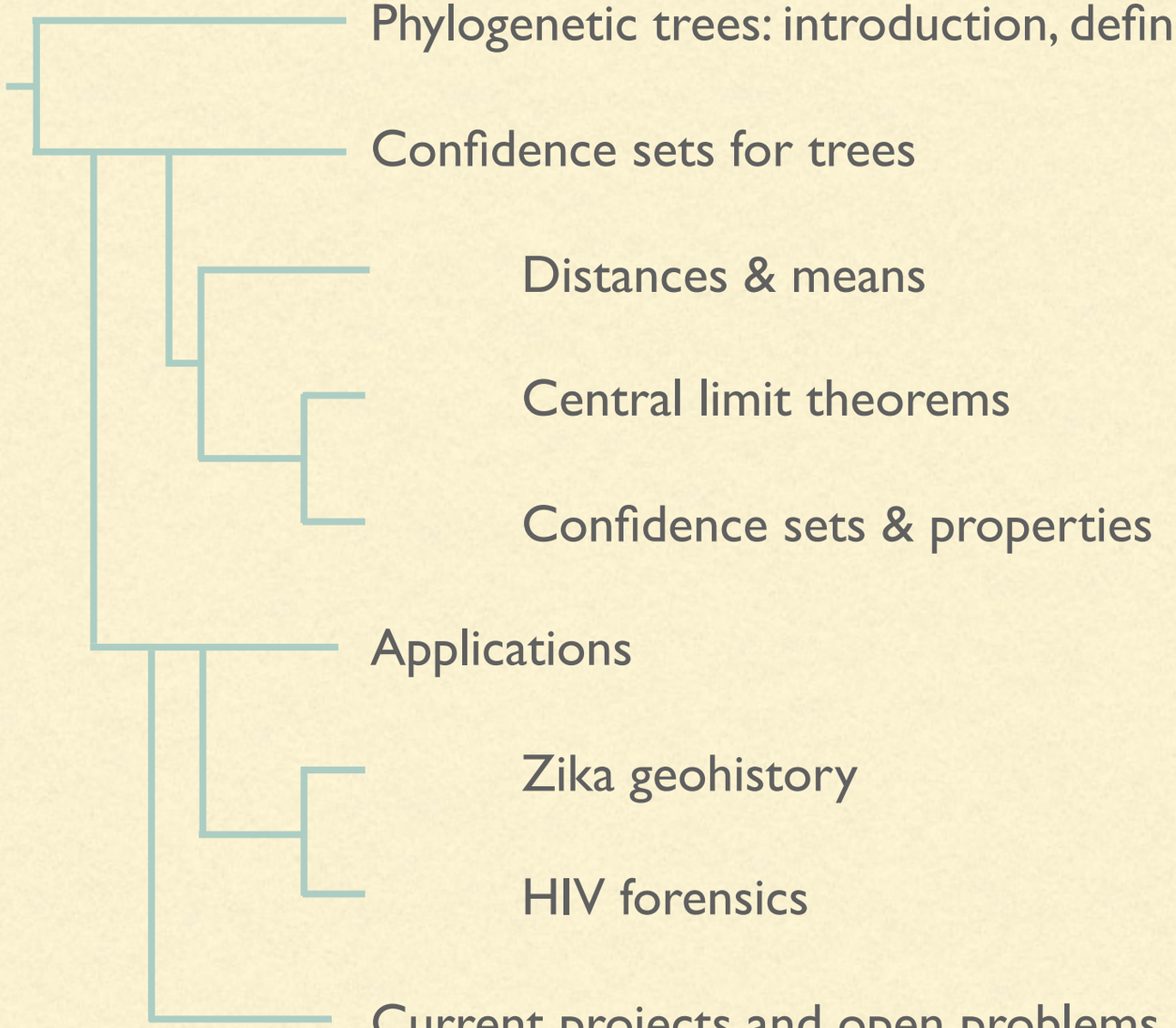
CONFIDENCE SETS FOR PHYLOGENETIC TREES

Willis (2016+), *arXiv:1607.08288*

Willis & Bell (2017+), *JGCS, arXiv:1611.03456*

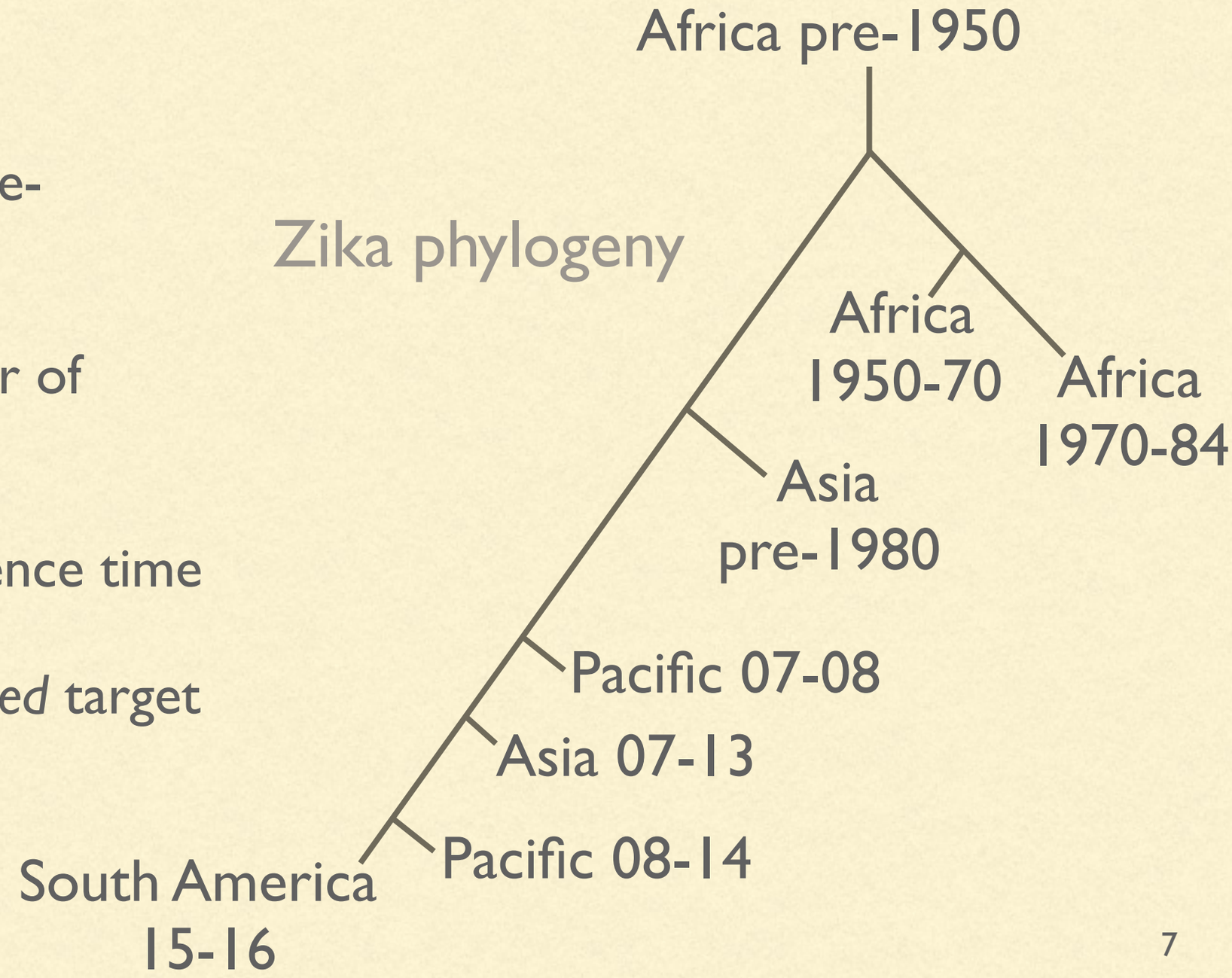


OUTLINE


Phylogenetic trees: introduction, definition, motivation
Confidence sets for trees
Distances & means
Central limit theorems
Confidence sets & properties
Applications
Zika geohistory
HIV forensics
Current projects and open problems

EVOLUTIONARY HISTORIES

- phylogenetic tree: edge-weighted tree graph
- branch structure: order of divergence
- branch lengths: divergence time
- Multivariate, *graph-valued* target of interest



MOTIVATION

- Save the rainforest
  - Conservation demands species delimitation
- Free the innocent
  - “Bulgarian nurses affair”
- Improve healthcare
  - Identify & prevent disease transmission by healthcare workers

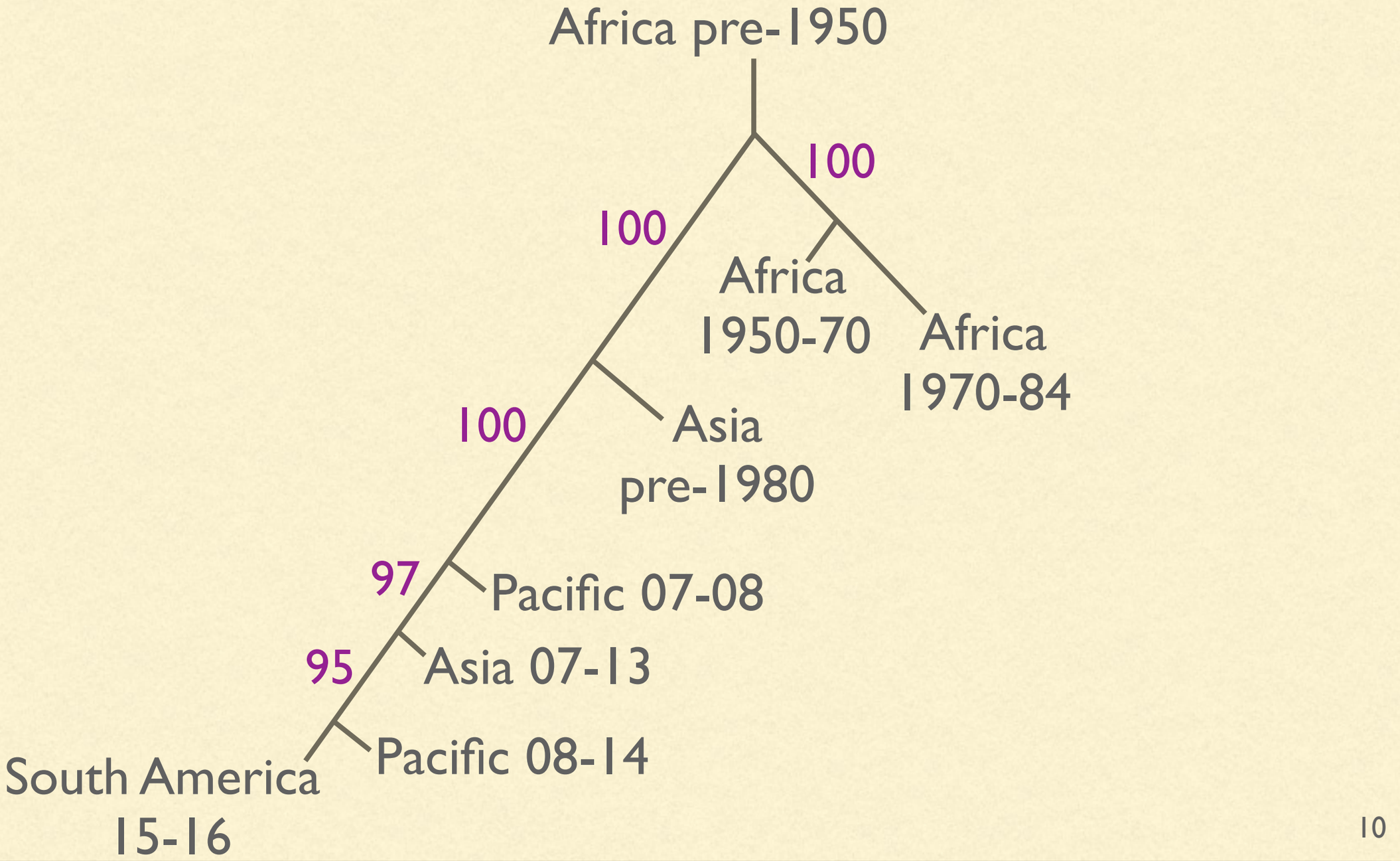
Su et al 2015, de Oliveira et al 2006, Ou et al 1992, Scaduto et al 2010

MOTIVATION

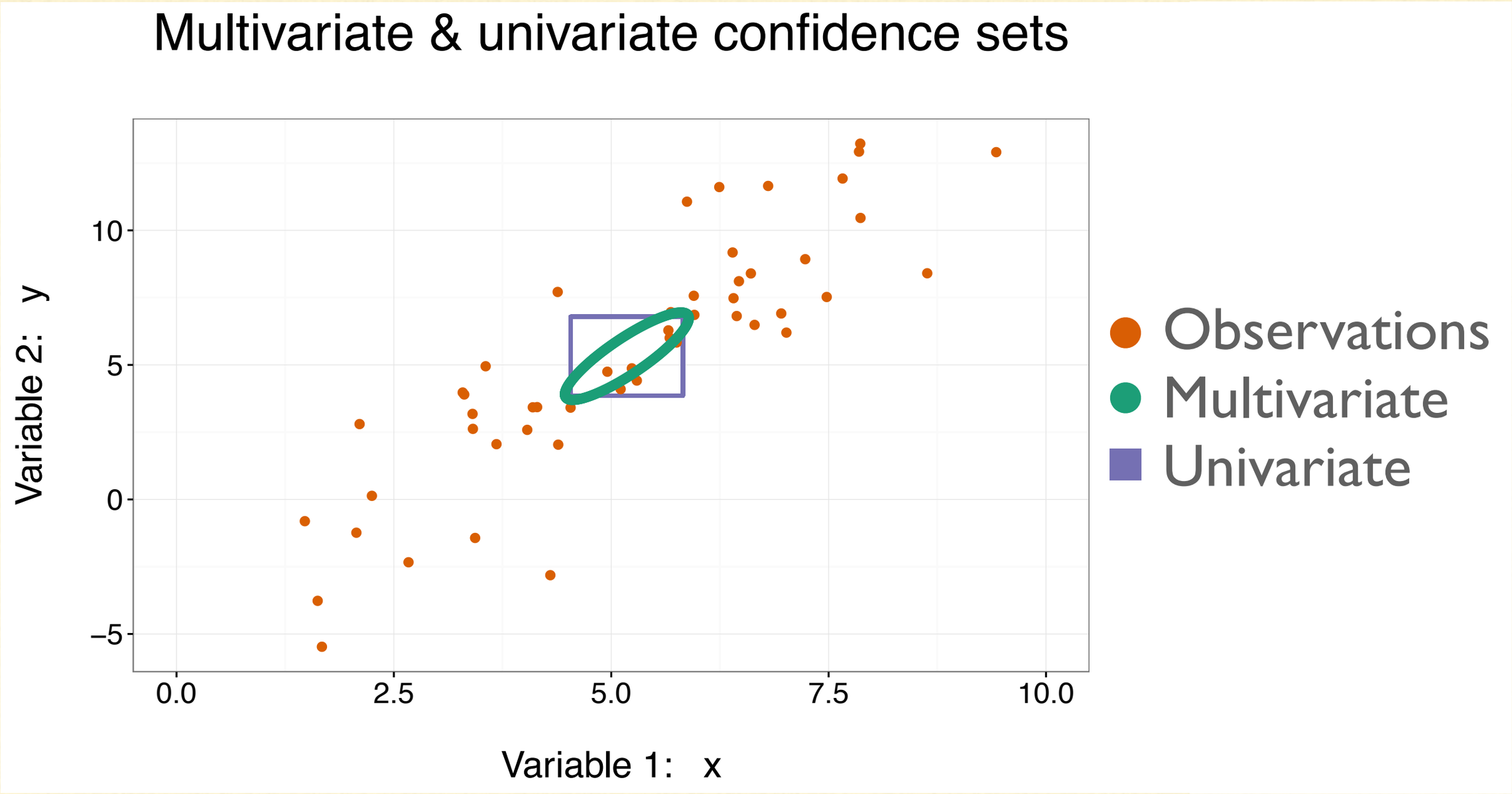
- Save the rainforest
  - Conservation demands species delimitation
- Free the innocent
  - “Bulgarian nurses affair”
- Improve healthcare
  - Identify & prevent disease transmission by healthcare workers

Q: How confident are we in these estimates?

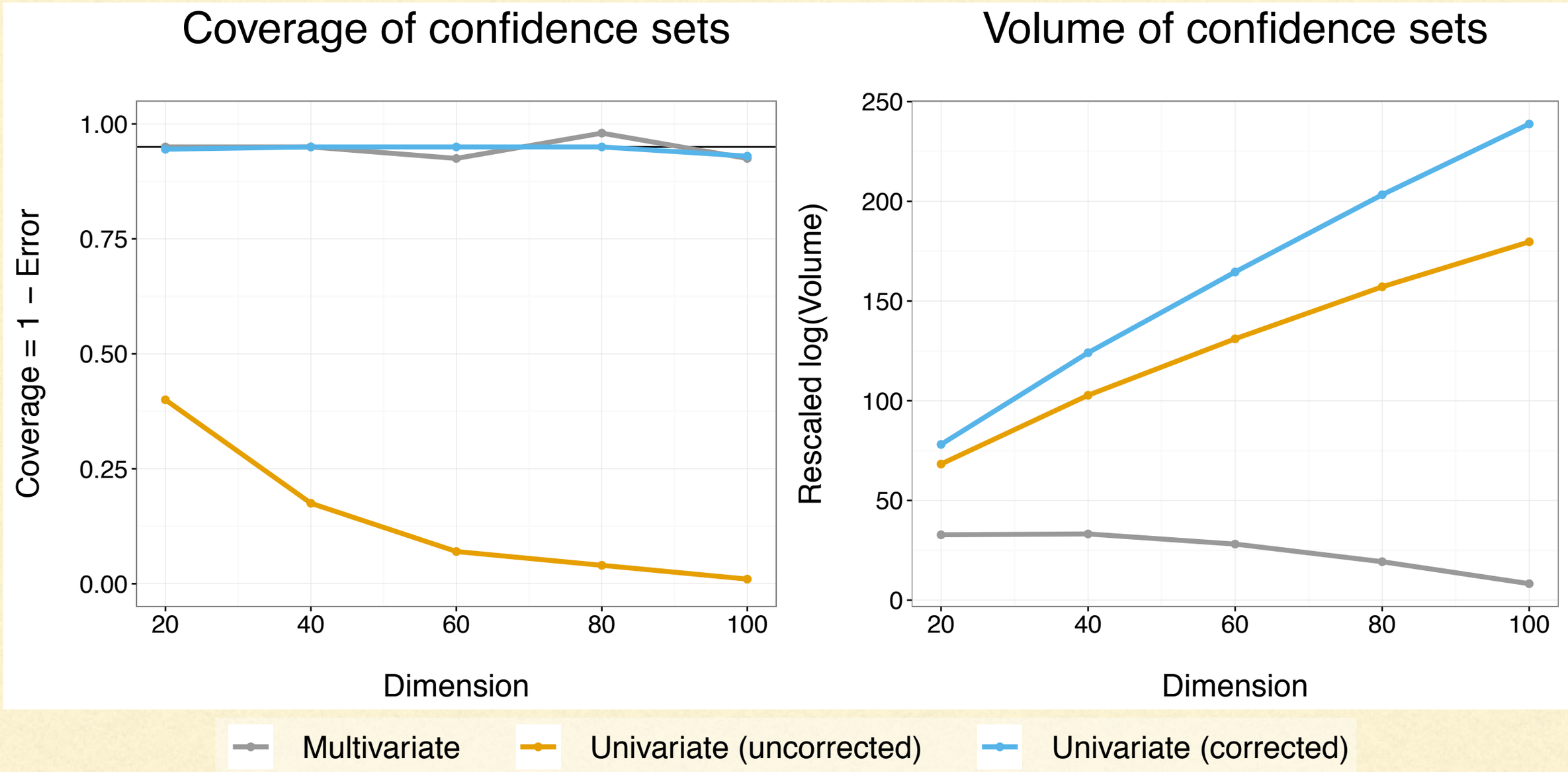
TREE UNCERTAINTY



MULTIVARIATE UNCERTAINTY



MULTIVARIATE UNCERTAINTY



TREE-VALUED DATA

- Collections of trees arise as
  - Gene trees
  - Posterior samples
  - Within-species replicates

GOAL

- Given trees $T_1, T_2, \ldots, T_n$ representing n different histories of the same m species, we want to

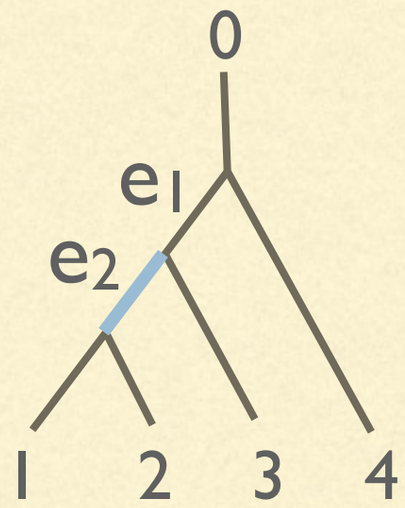
construct a confidence set for the true mean phylogenetic tree

OUTLINE

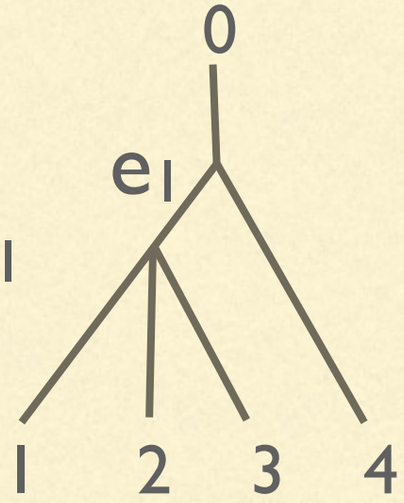
- "Constructing a confidence set for a mean tree"
  - Mean?
    - Distance?
      - Metric space of trees?

TREE SPACE

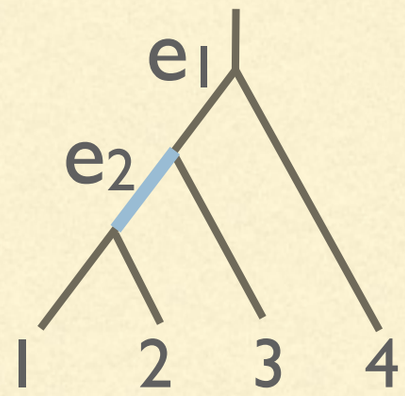
$T_1 =$



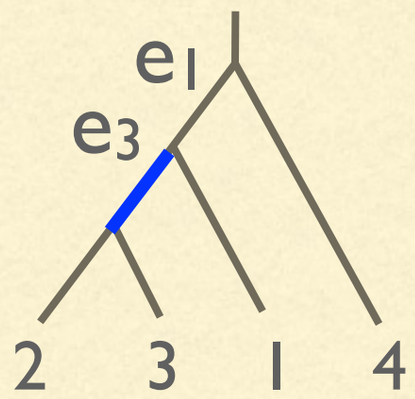
$\lim_{e2 \to 0} T_1$



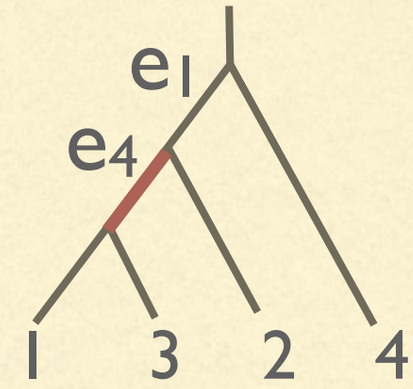
$T_1 =$



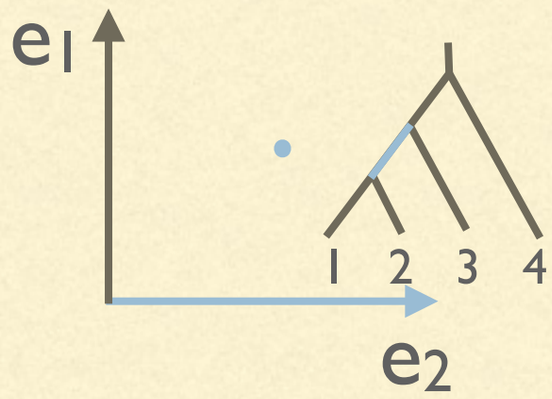
$T_2 =$

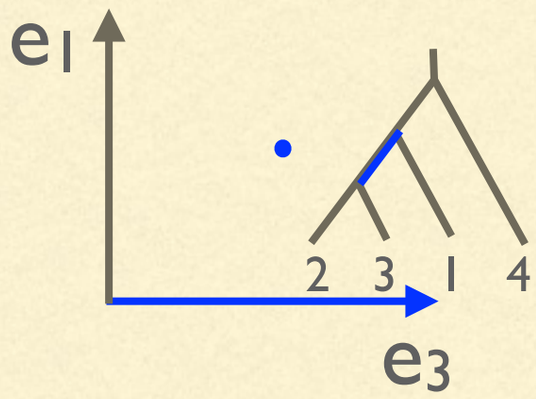


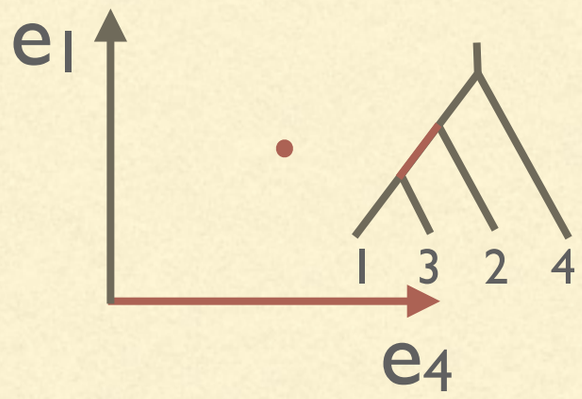
$T_3 =$



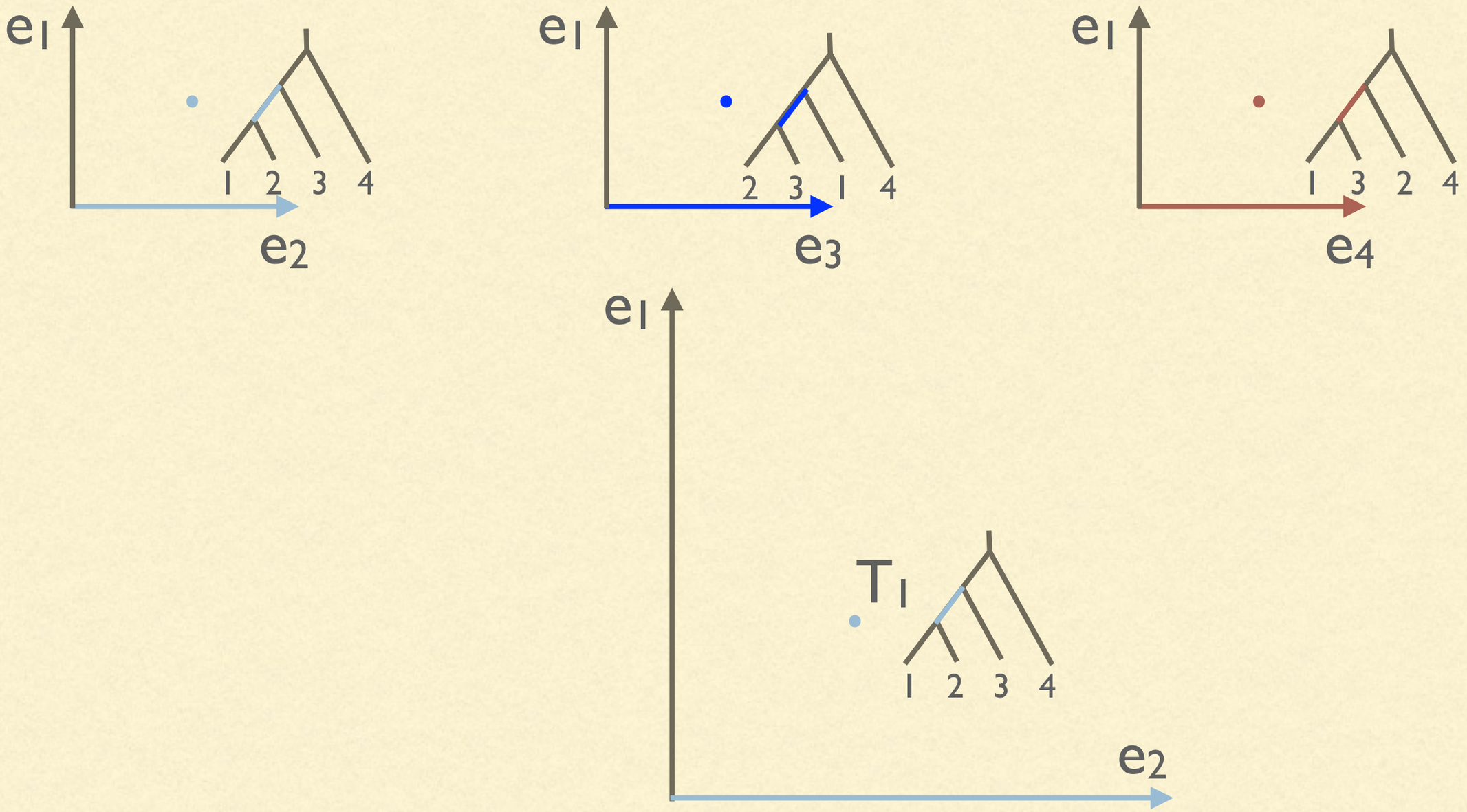
TREE SPACE



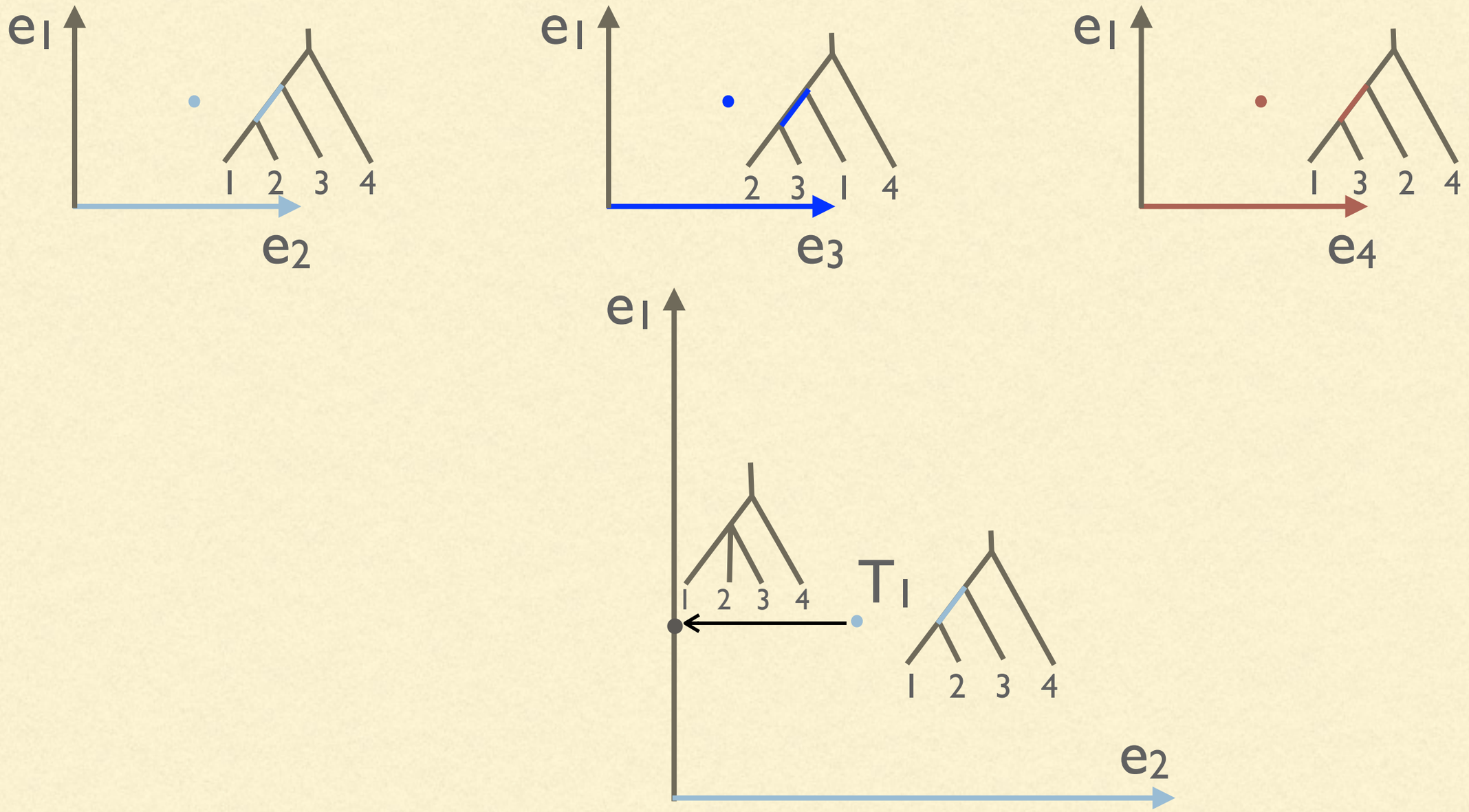




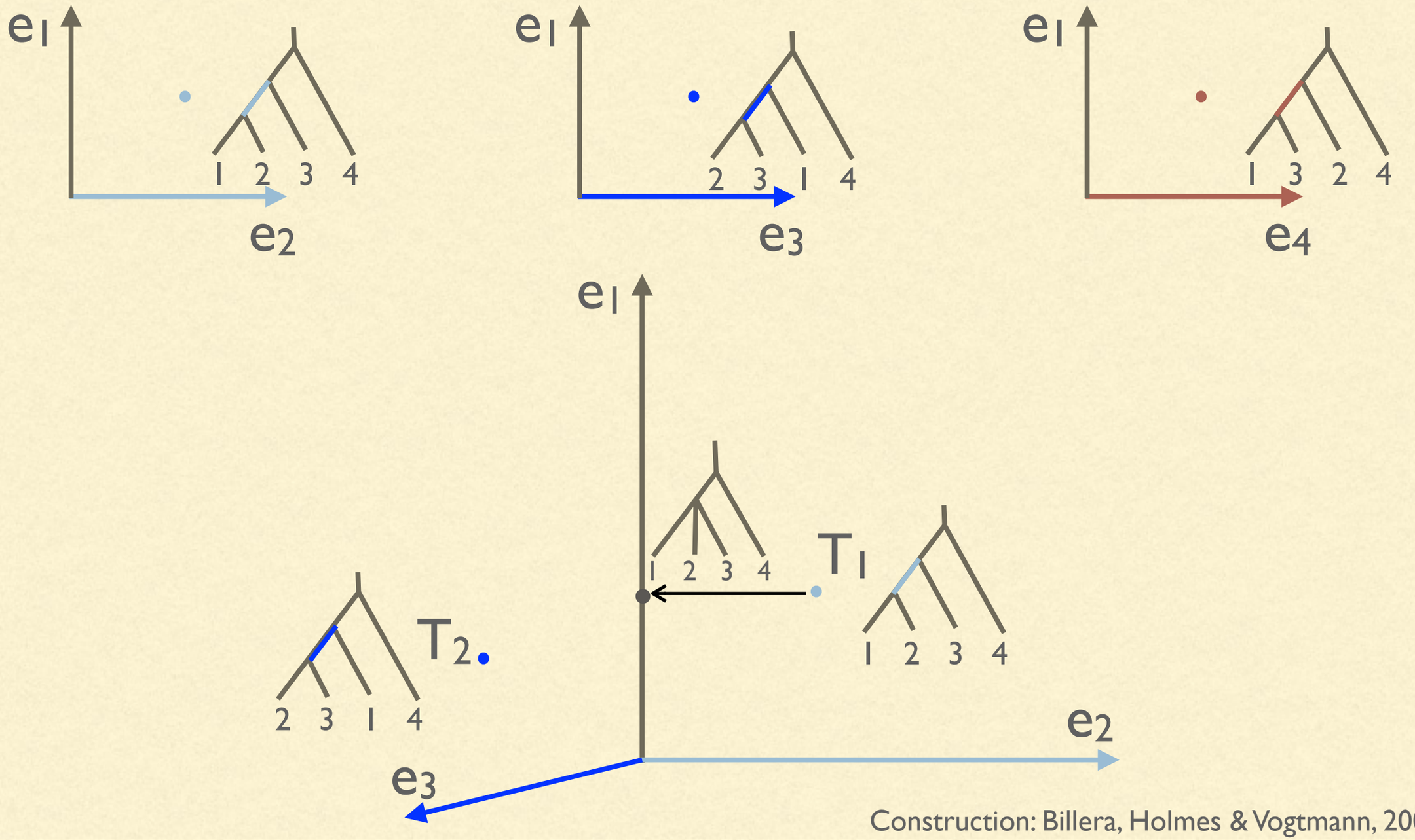
TREE SPACE



TREE SPACE



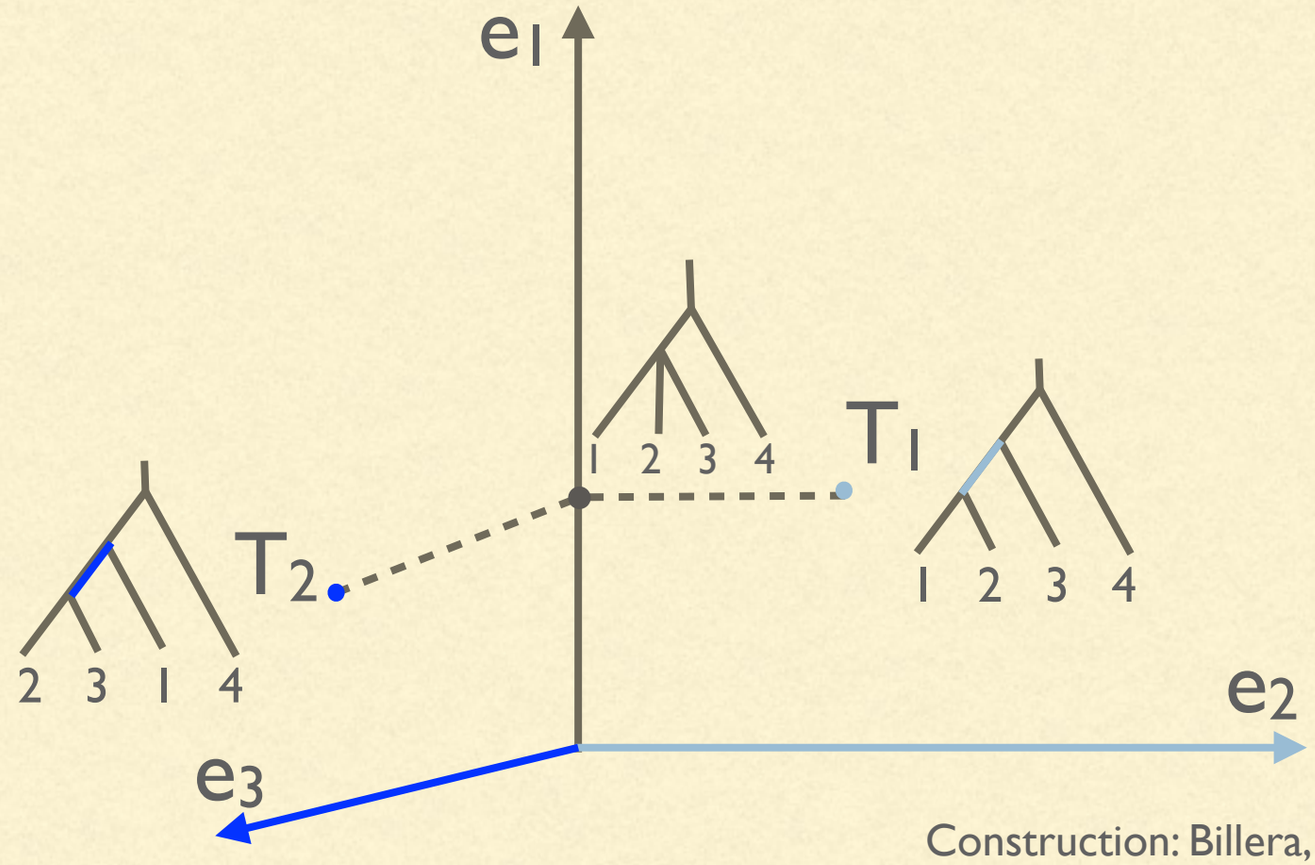
TREE SPACE



Construction: Billera, Holmes & Vogtmann, 2001

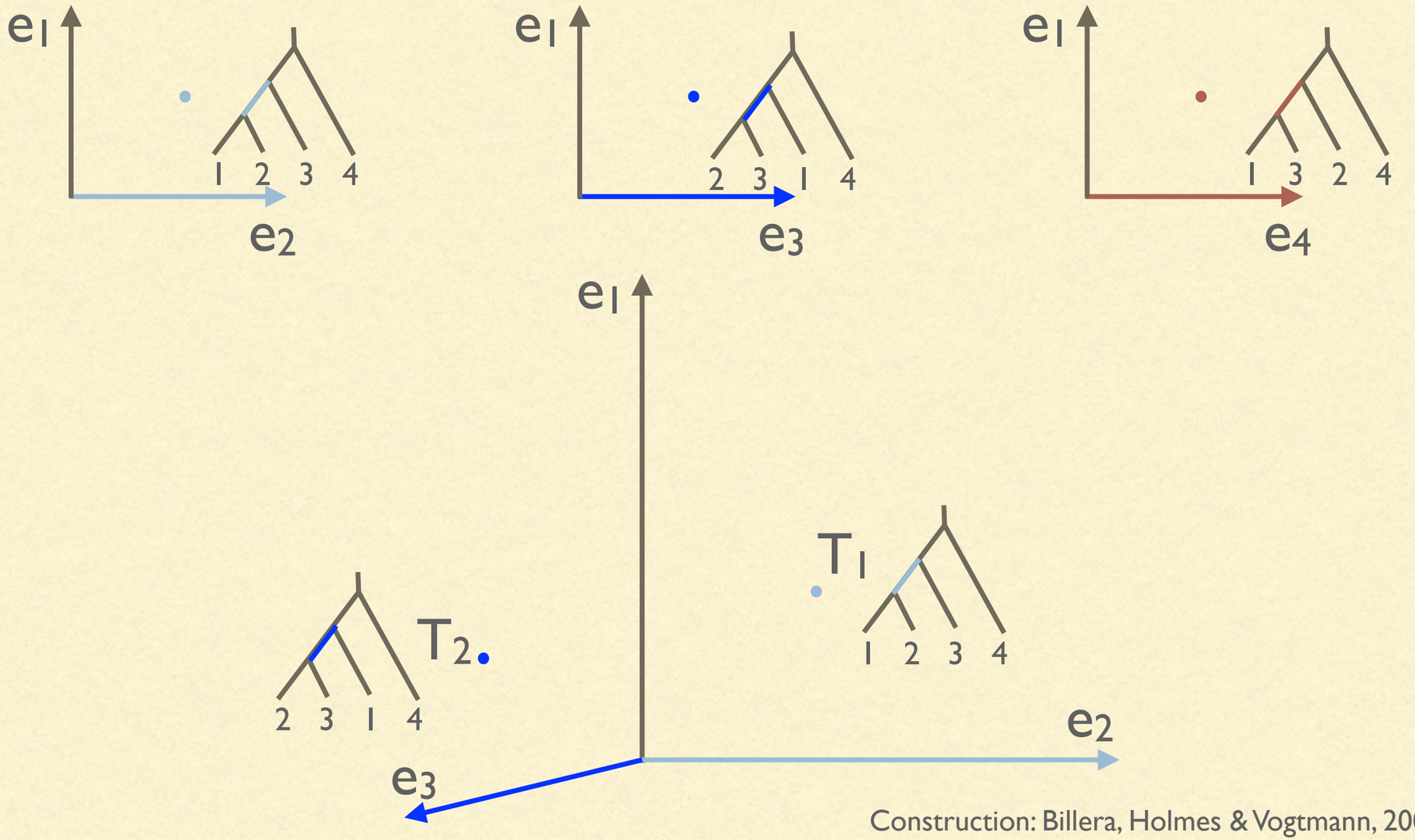
METRIC DISTANCE

$\gamma(T_1, T_2)$ = length of shortest path between T_1 and T_2



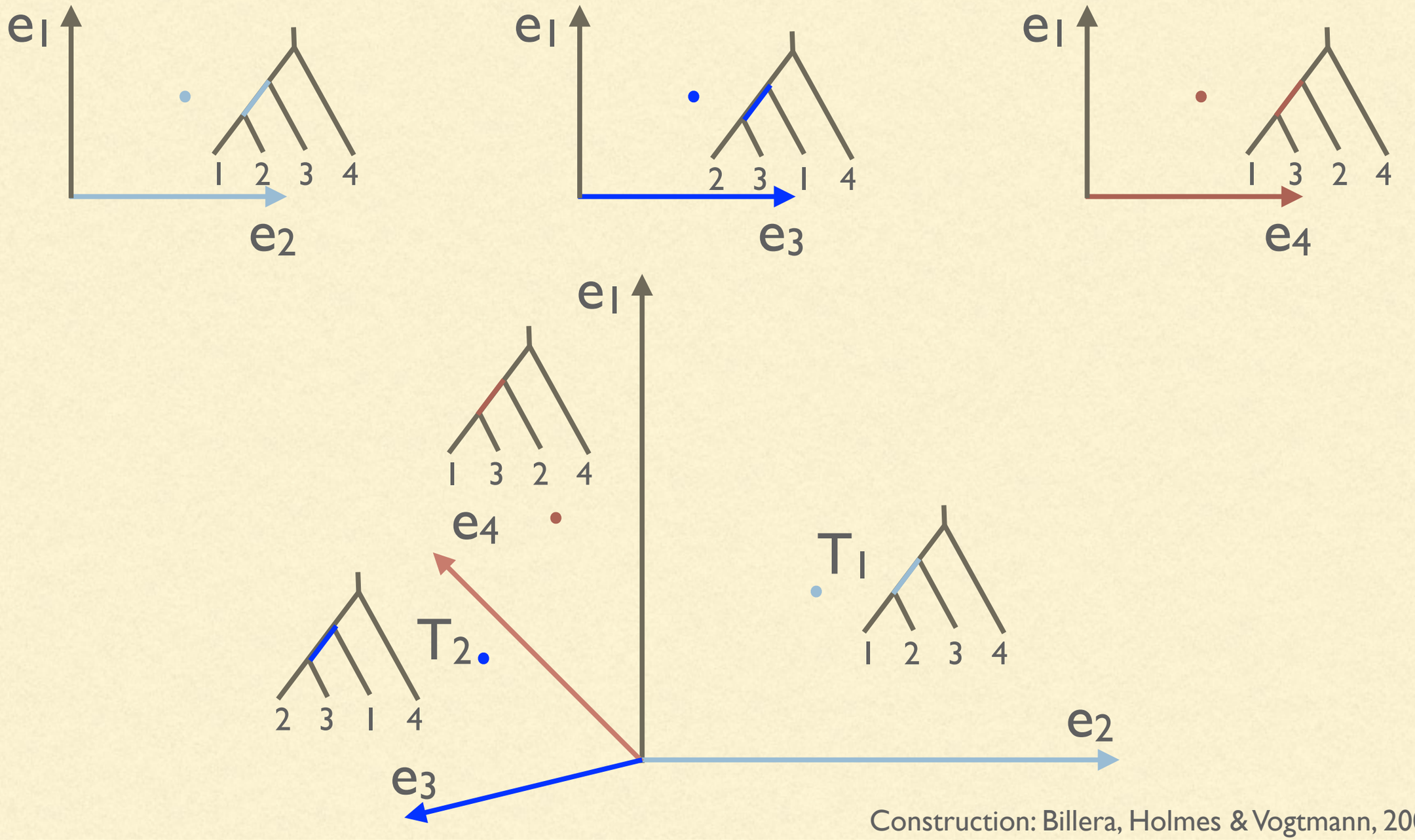
Construction: Billera, Holmes & Vogtmann, 2001

TREE SPACE



Construction: Billera, Holmes & Vogtmann, 2001

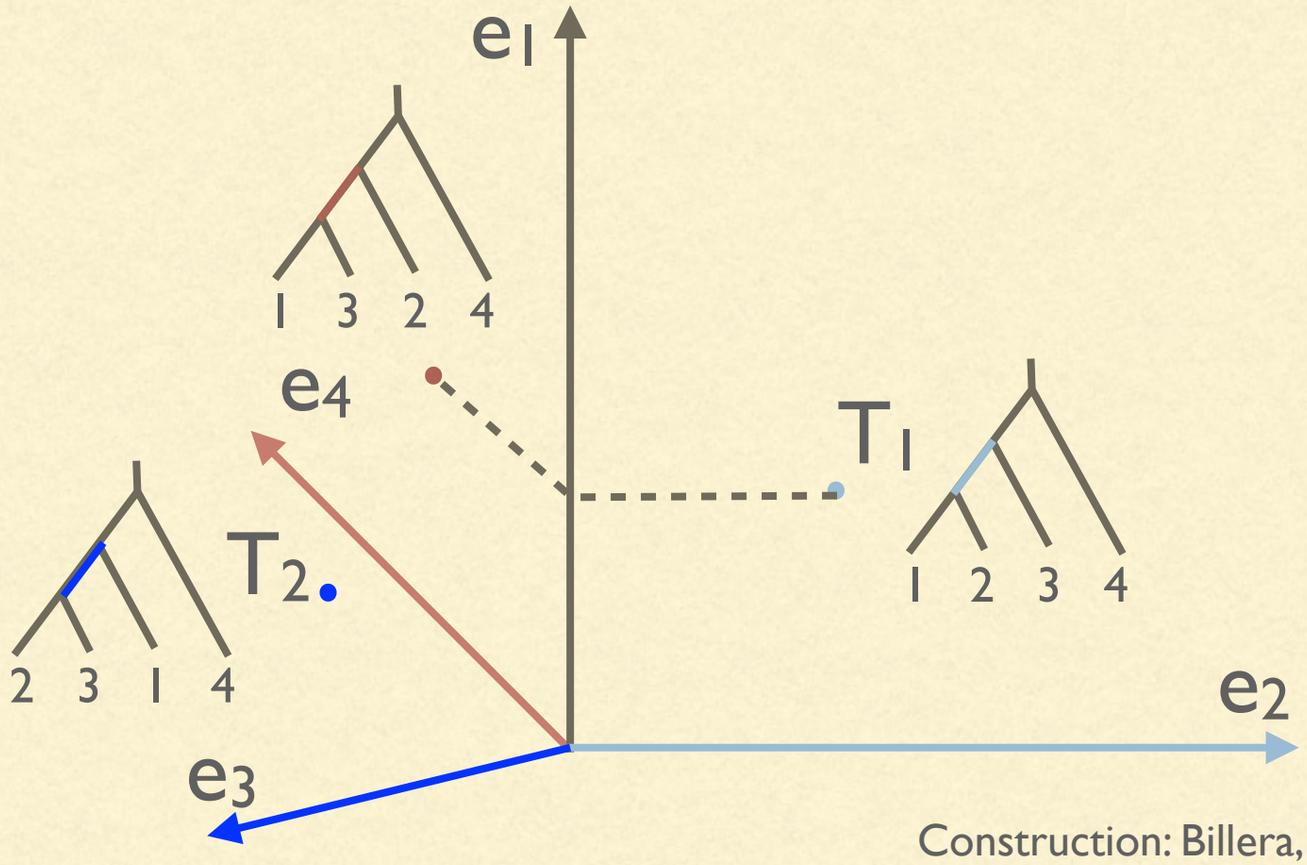
TREE SPACE



Construction: Billera, Holmes & Vogtmann, 2001

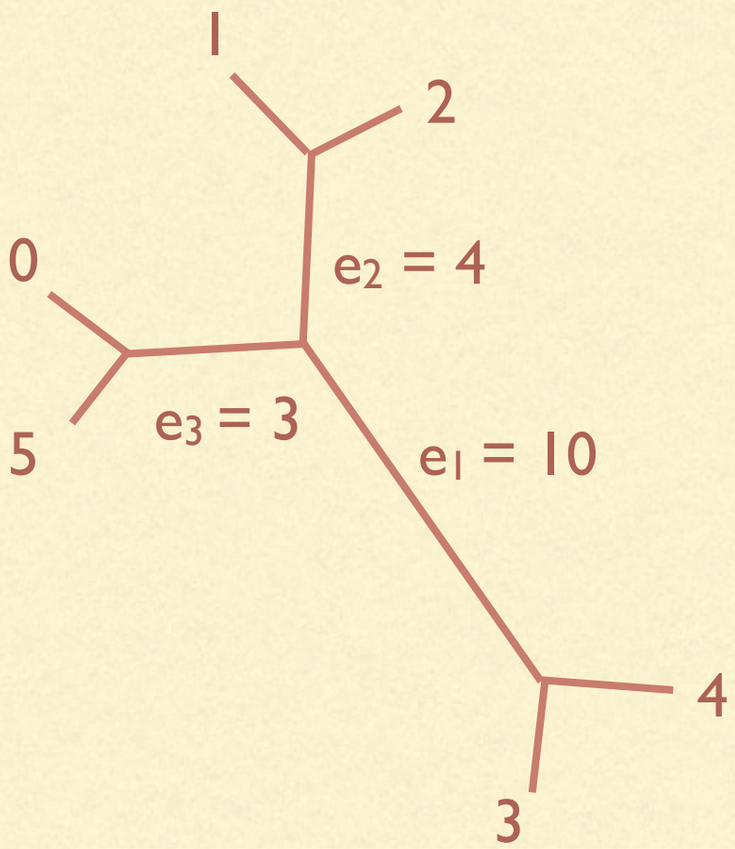
METRIC DISTANCE

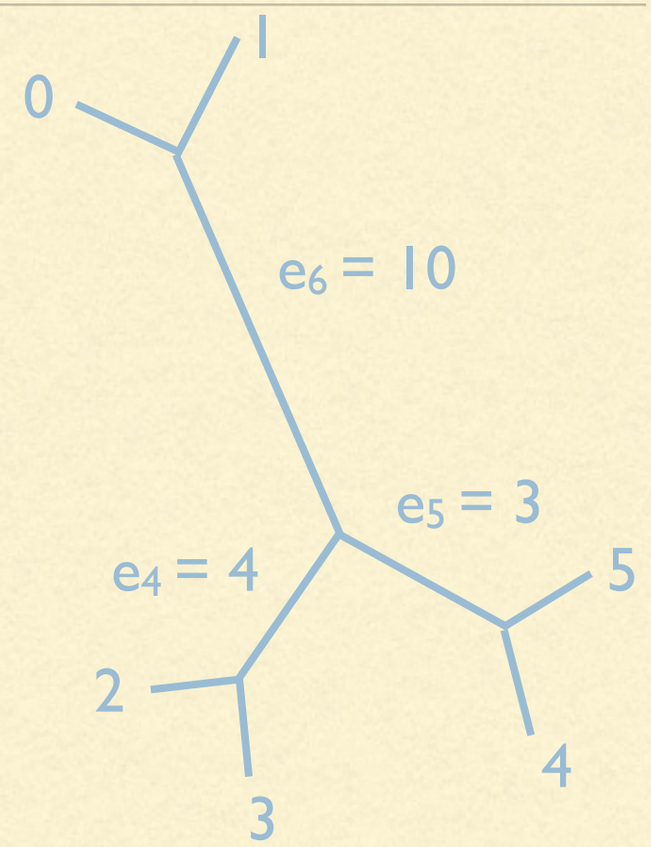
$\gamma(T_1, T_3)$ = length of shortest path between T_1 and T_3



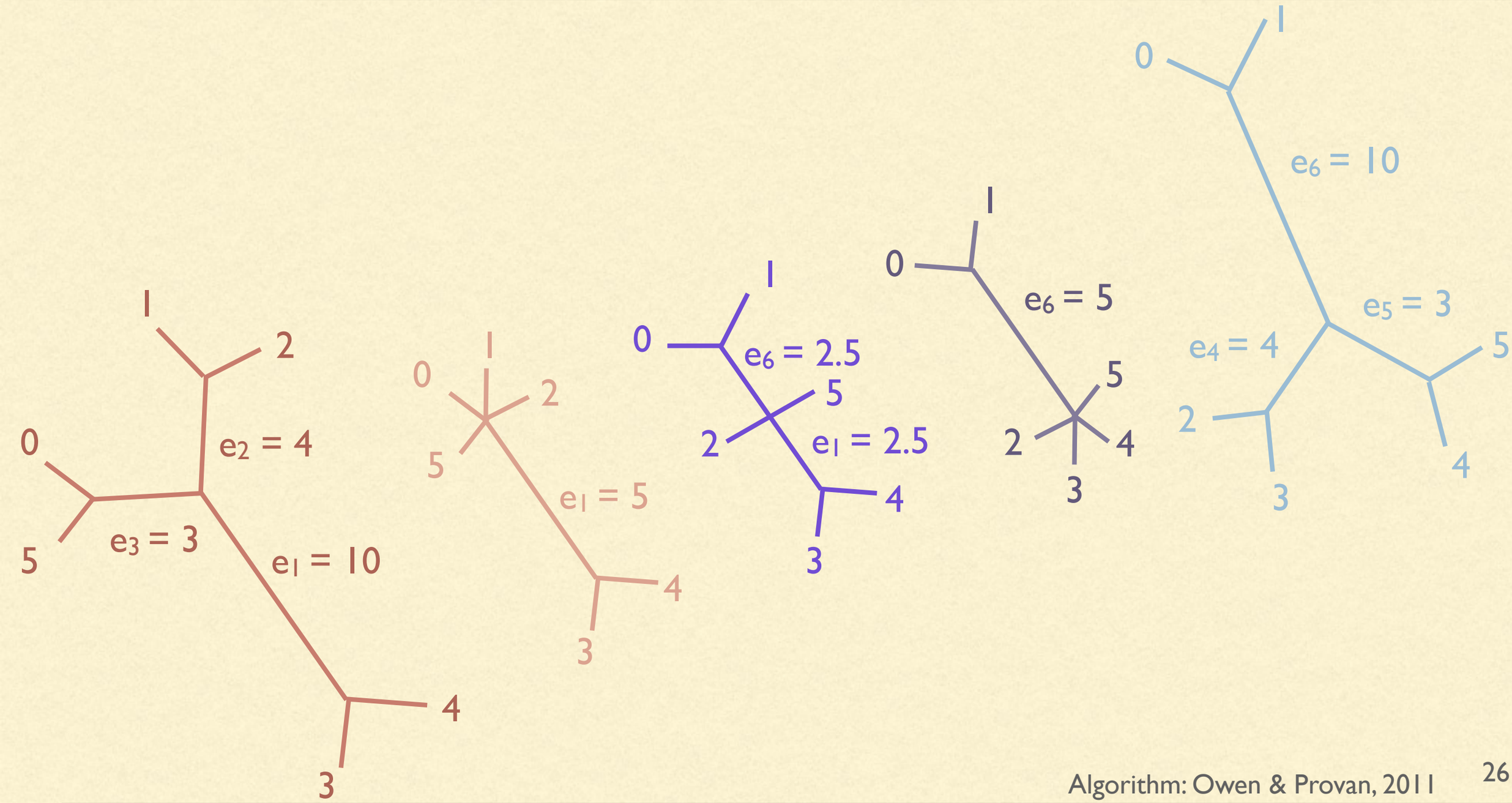
Construction: Billera, Holmes & Vogtmann, 2001

PATHS BETWEEN TREES



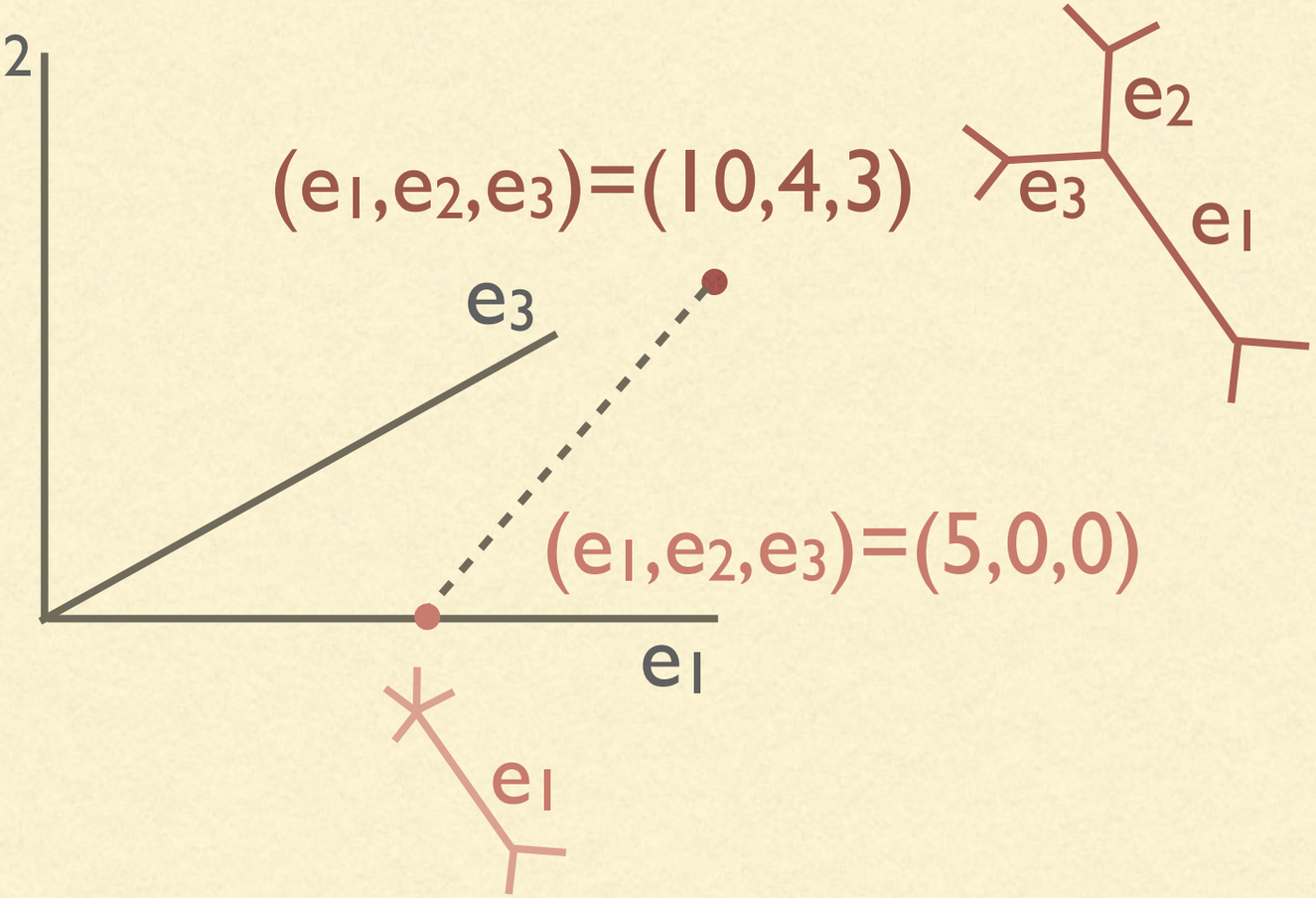


PATHS BETWEEN TREES

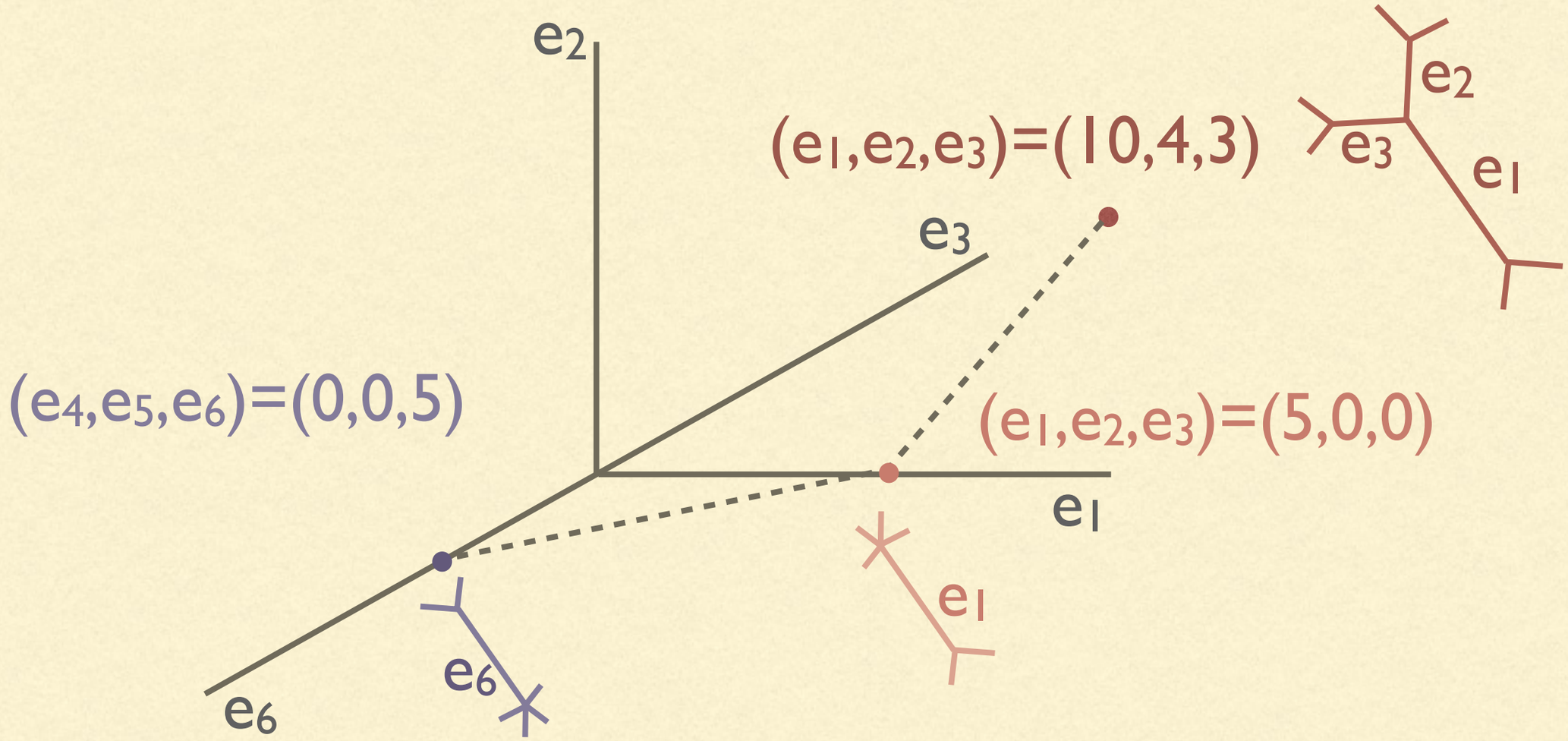


Algorithm: Owen & Provan, 2011

PATHS BETWEEN TREES

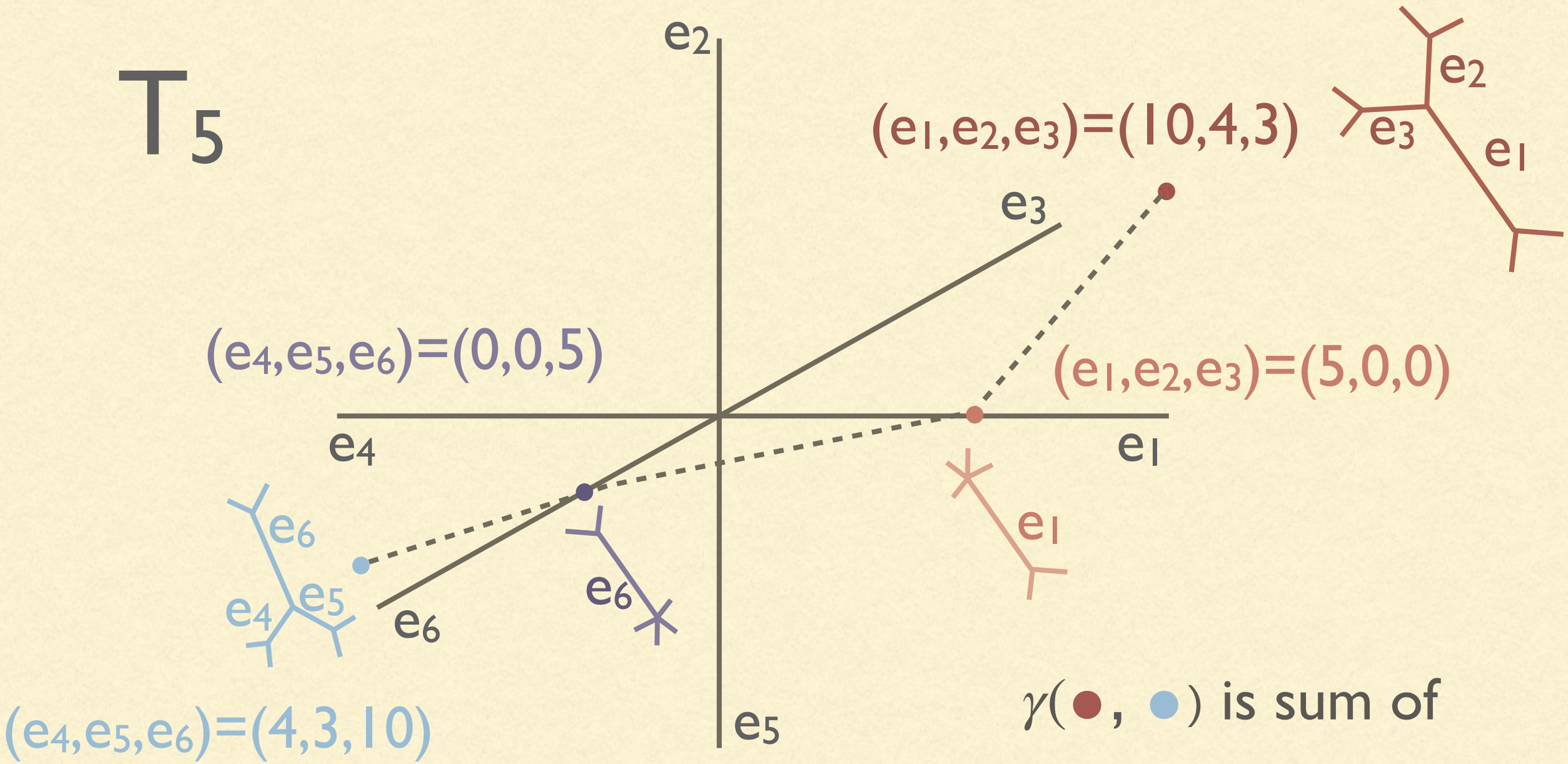


PATHS BETWEEN TREES



PATHS BETWEEN TREES

T_5



$\gamma(\bullet, \bullet)$ is sum of segment lengths

TREE MEANS

- Definition: The Fréchet mean of a distribution *F* on $\mathcal{T}_m$

$$\mu = \underset{u \in \mathcal{T}_m}{arg\,min} \int \gamma(q,u)^2 F(dq)$$

- Sample analogue:

$$\hat{T}_n(T_1, \ldots, T_n) = \underset{u \in \mathcal{T}_m}{arg\,min} \sum_{i=1}^{n} \gamma(T_i, u)^2$$

TREE MEAN VARIABILITY

- Tree space is a metric space, but not a vector space nor inner product space
- Multivariate variability not possible without orthogonality
- Need to *unfold* tree space to embed it in Euclidean space

LOG MAP

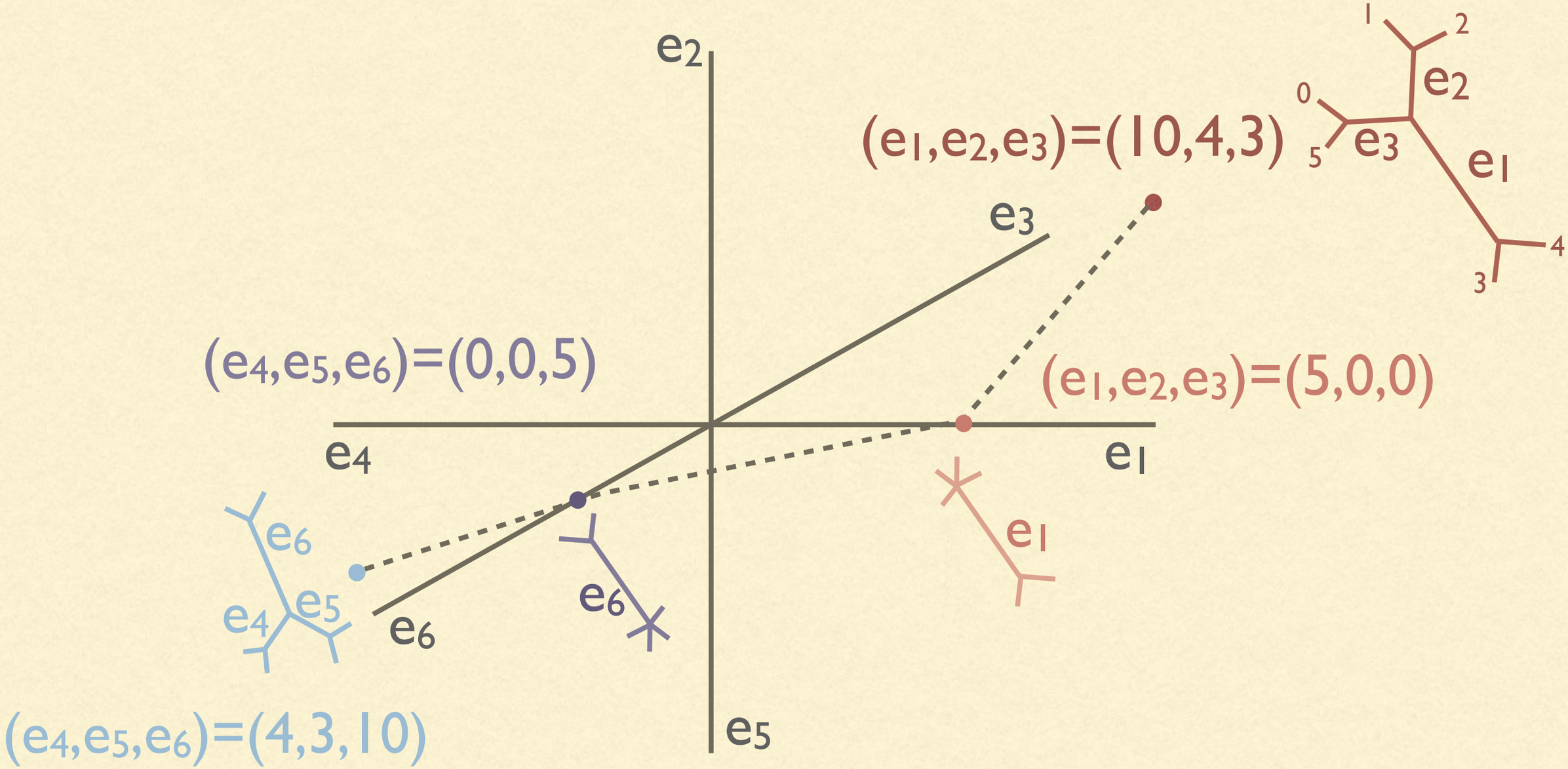
- The log map *unfolds* the combinatorial structure of tree space

$$\Phi_T(\cdot): T_M \rightarrow R^{M-2}$$

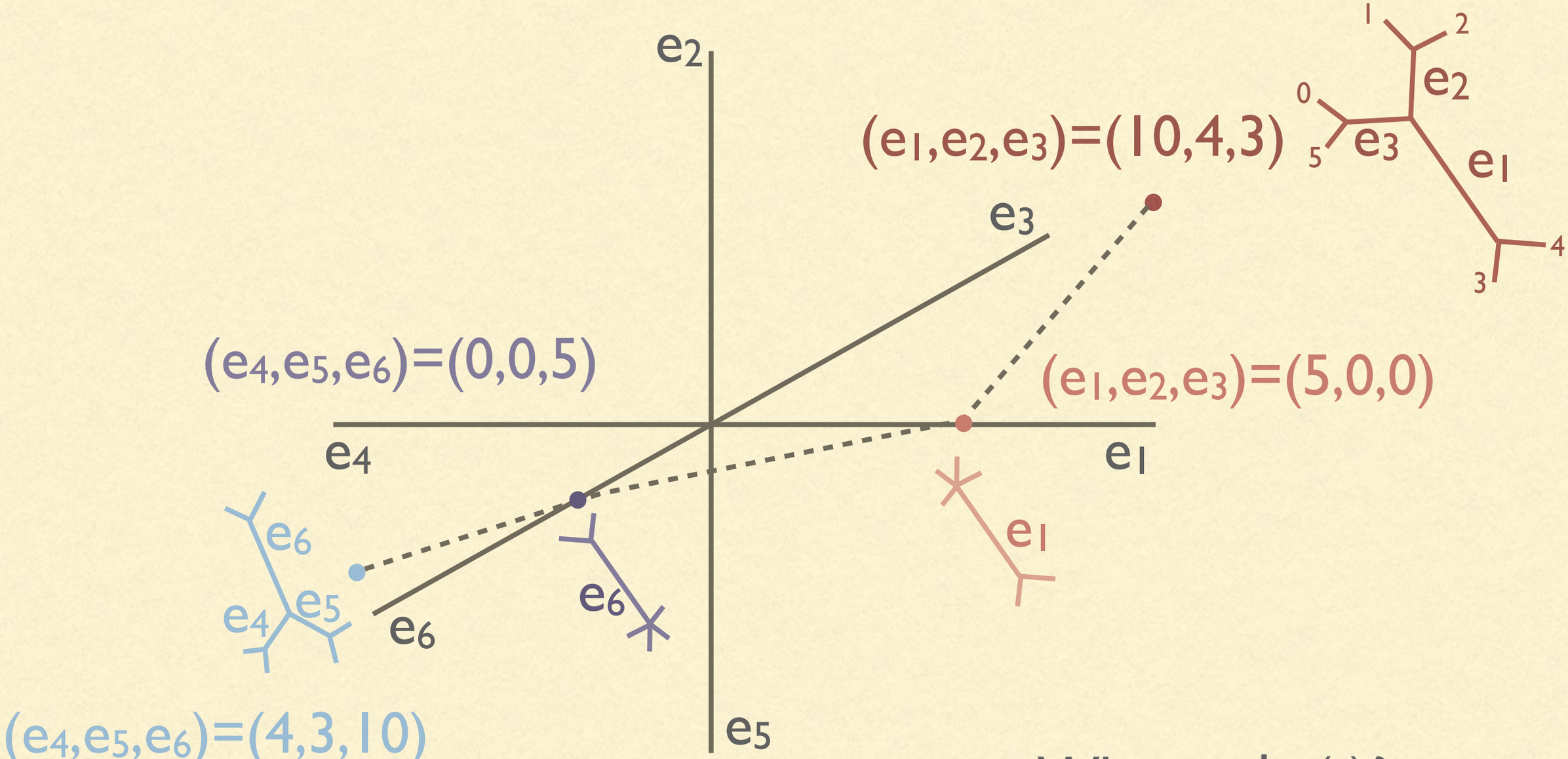
- It is *centred* at a base tree T

- Construction due to Barden, Le & Owen (2016)

LOG MAP


e2
e3
e4
e1
e5
e6
(e1,e2,e3)=(10,4,3)
(e4,e5,e6)=(0,0,5)
(e1,e2,e3)=(5,0,0)
(e4,e5,e6)=(4,3,10)
1
2
0
5
3
4

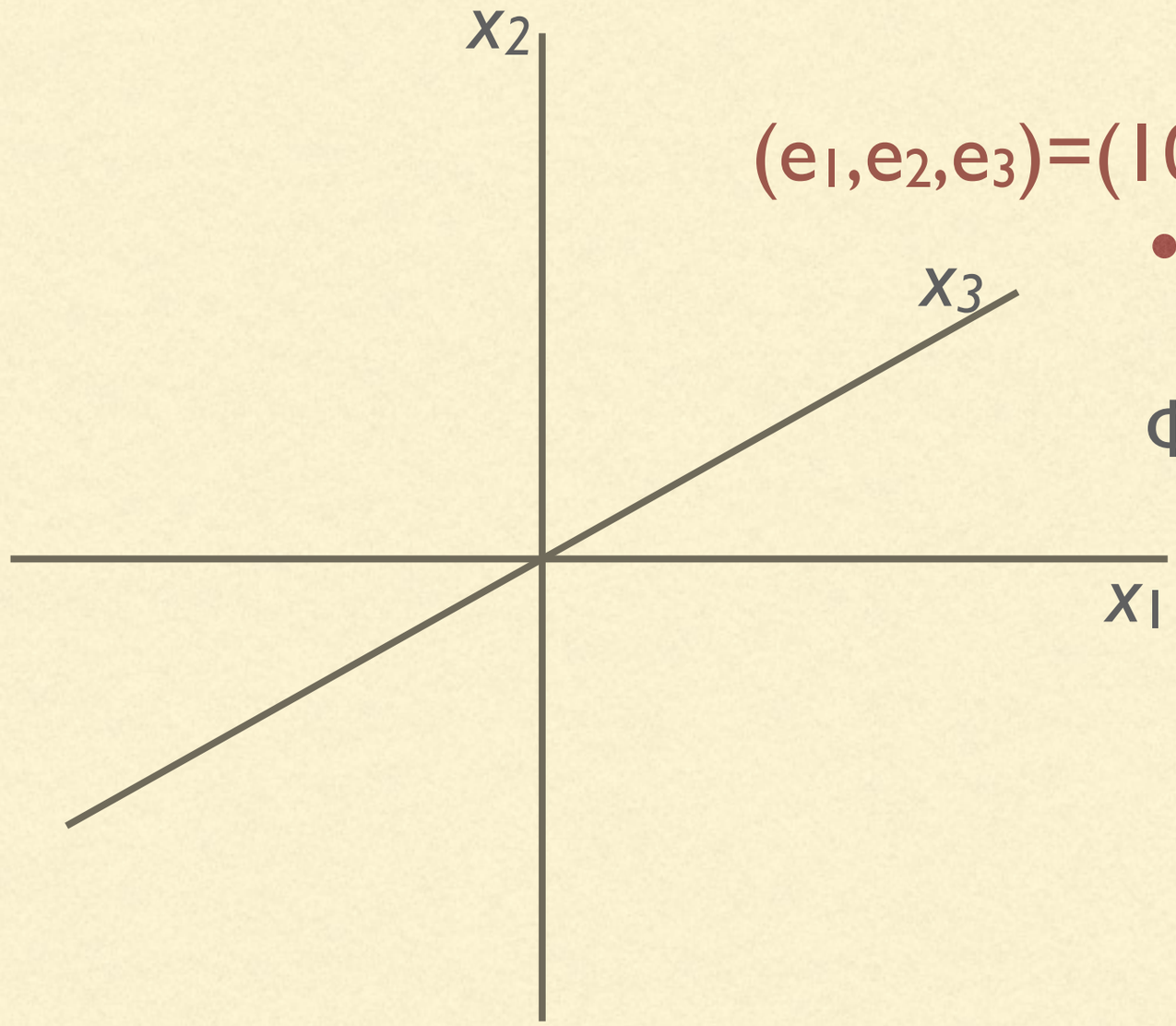
LOG MAP



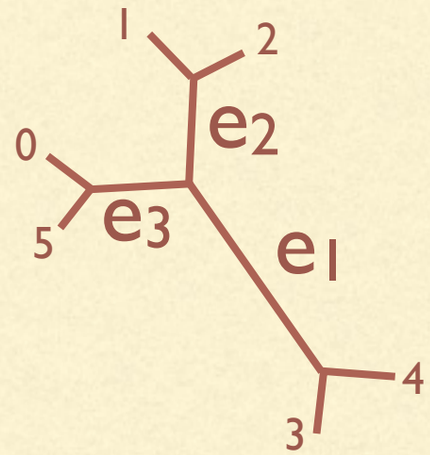
What is $\Phi_{\bullet}(\cdot)$?

LOG MAP

R^3



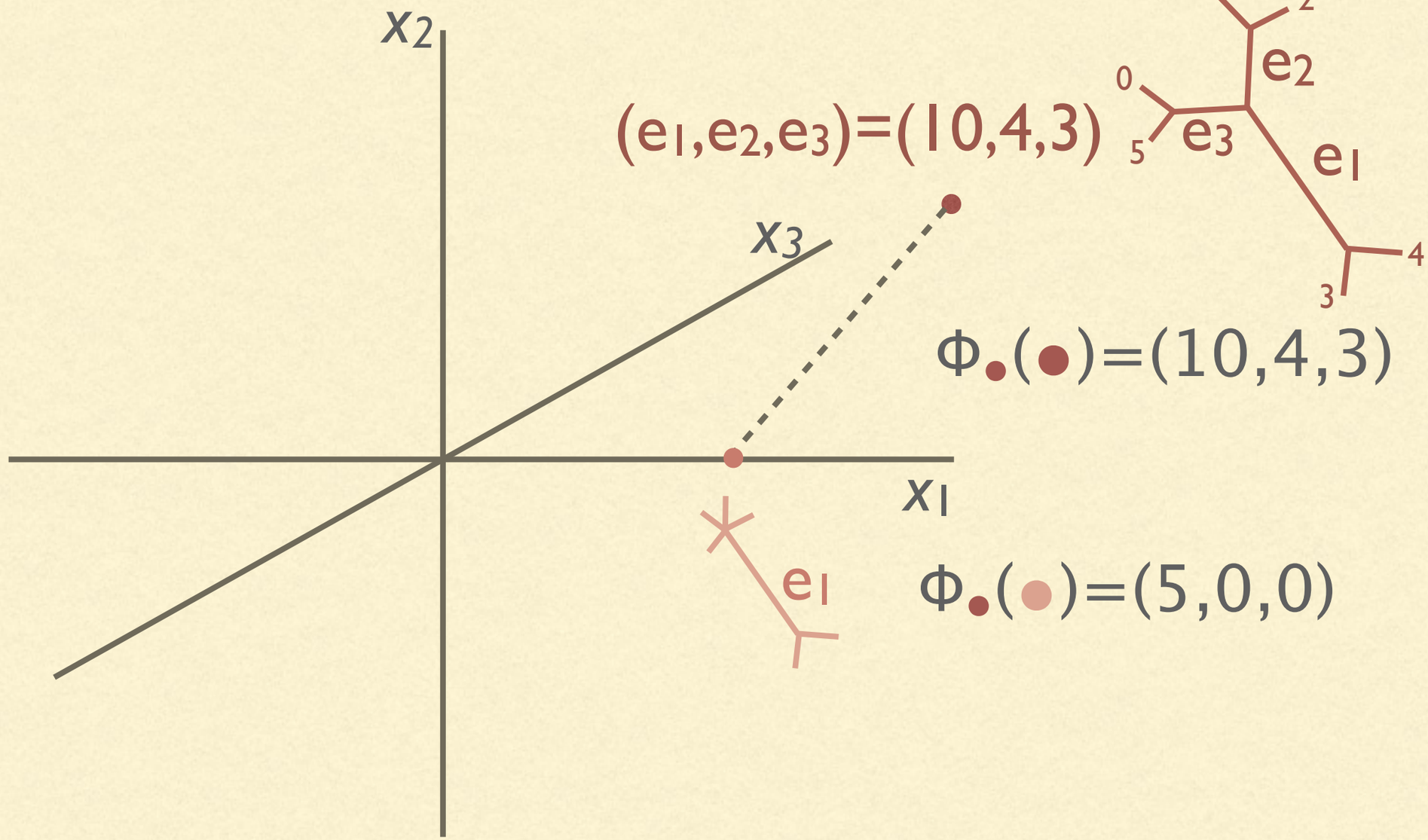
$(e_1,e_2,e_3)=(10,4,3)$



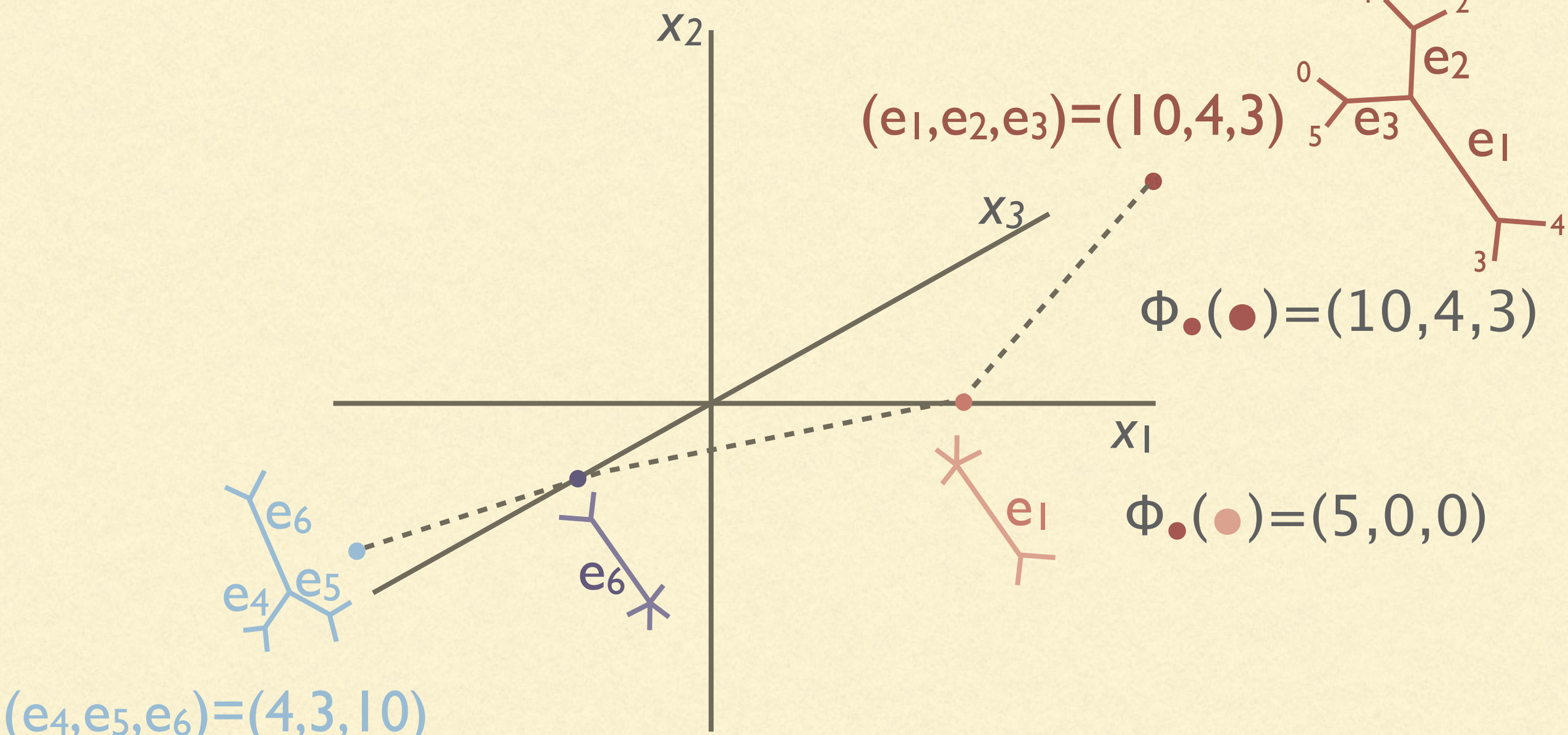
$\Phi_{\bullet}(\bullet)=(10,4,3)$

LOG MAP

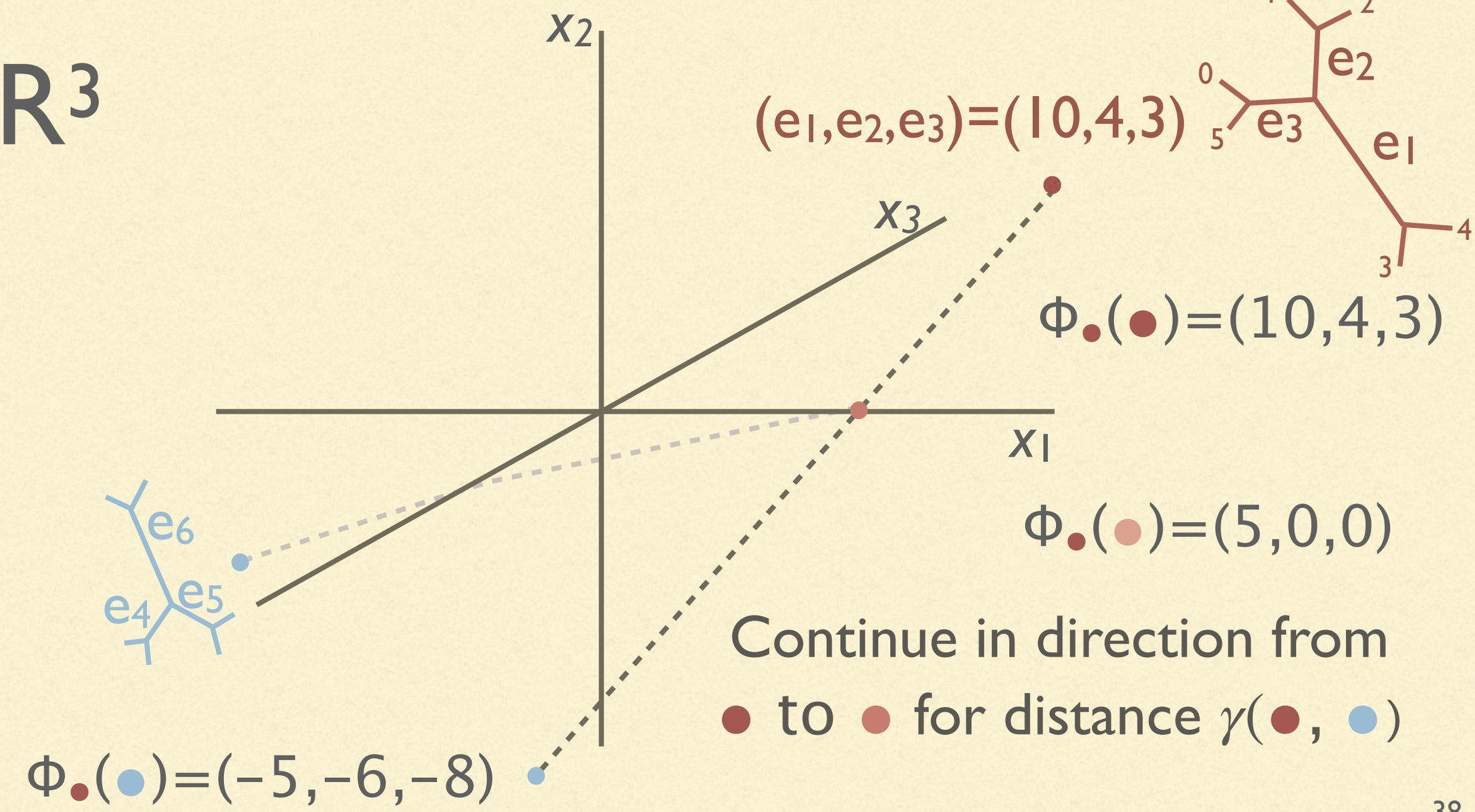
R^3



LOG MAP


x2
x3
x1
(e1,e2,e3)=(10,4,3)
e1
e2
e3
0
1
2
3
4
5
Φ•(●)=(10,4,3)
Φ•(●)=(5,0,0)
e1
e6
e4
e5
e6
(e4,e5,e6)=(4,3,10)
Φ•(●) = ?

LOG MAP UNFOLDS TREE SPACE



TREE CENTRAL LIMIT THEOREM

- Theorem: (Barden, Le & Owen, 2016)

  For $\{T_i\}_{i=1\ldots n}$ drawn iid from F with Fréchet mean T^*, as n becomes large,

$$\sqrt{n}(\Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{T^*}(T^*)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, A^T V A)$$

  where V is the covariance of $\Phi_{T^*}(T_1)$ and A is a rotation depending on F

TREE CENTRAL LIMIT THEOREM

- For trees $\{T_i\}_{i=1\ldots n}$ *with Fréchet mean* $\hat{T}_n$, define

$$\overline{\Phi_{\hat{T}_n}(T)} = \frac{1}{n}\sum_{i=1}^{n}\Phi_{\hat{T}_n}(T_i)$$

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\Phi_{\hat{T}_n}(T_i) - \overline{\Phi_{\hat{T}_n}(T)})(\Phi_{\hat{T}_n}(T_i) - \overline{\Phi_{\hat{T}_n}(T)})^T$$

as the sample mean and covariance of the log-mapped observations, $\Phi_{\hat{T}_n}(T_i)$

CONFIDENCE SETS FOR TREES

- Theorem: (Willis, 2016)

  A 100(1-α)% confidence hull for the Fréchet mean of *F* is

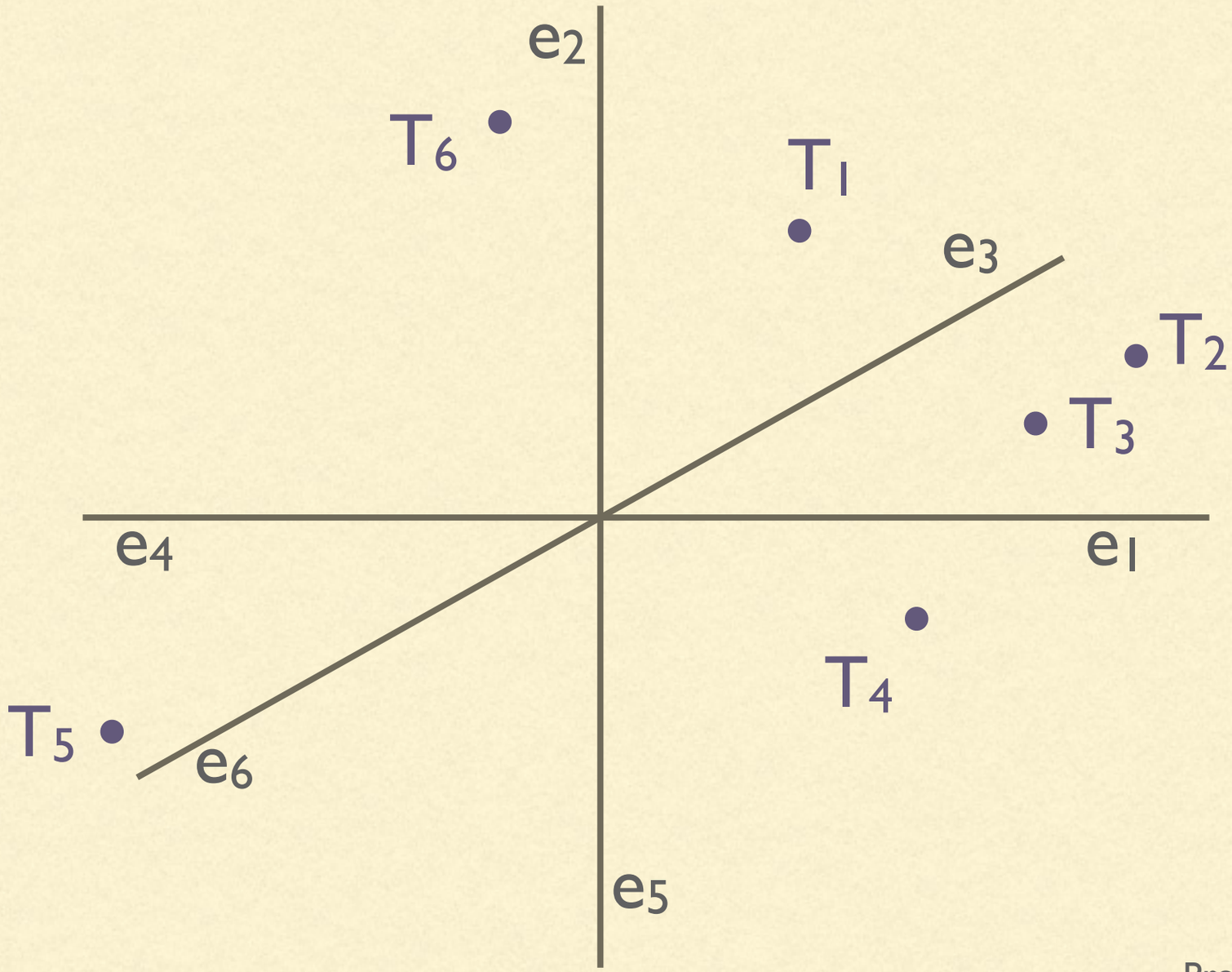
$$\left\{ T_0 \in \mathcal{T}_{m+2} : \right.$$

$$\left(\overline{\Phi_{\hat{T}_n}(T)} - \Phi_{\hat{T}_n}(T_0) \right)^T S^{-1} \left(\overline{\Phi_{\hat{T}_n}(T)} - \Phi_{\hat{T}_n}(T_0) \right)$$

$$\left. < \frac{m(n-1)}{n(n-m)} F_{m,n-m}(1-\alpha) \right\}$$

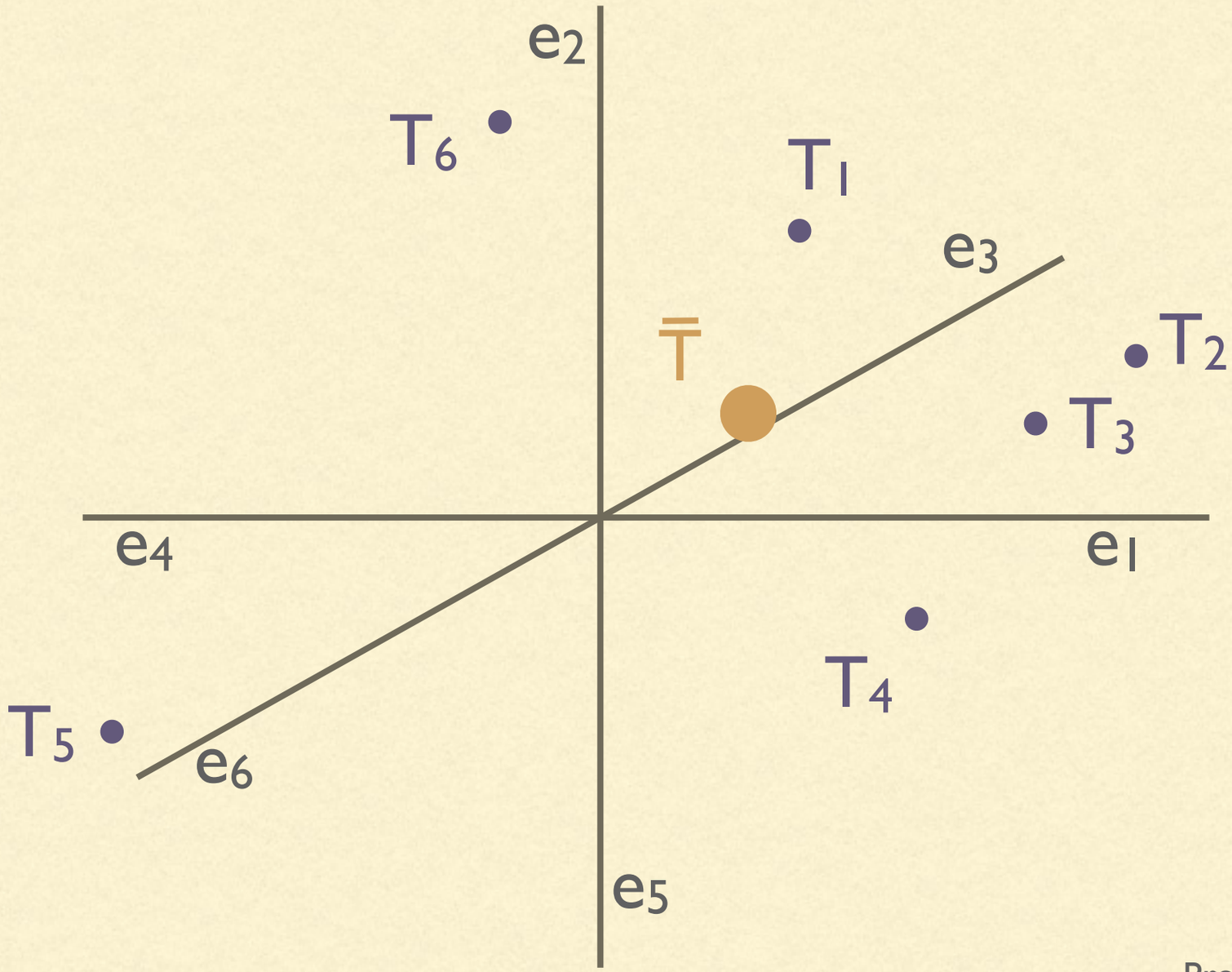
where S and $\overline{\Phi_{\hat{T}_n}(T)}$ are the sample mean and covariance of $\{\Phi_{\hat{T}_n}(T_i)\}_{i=1,\ldots,n}$

CONFIDENCE SETS FOR TREES



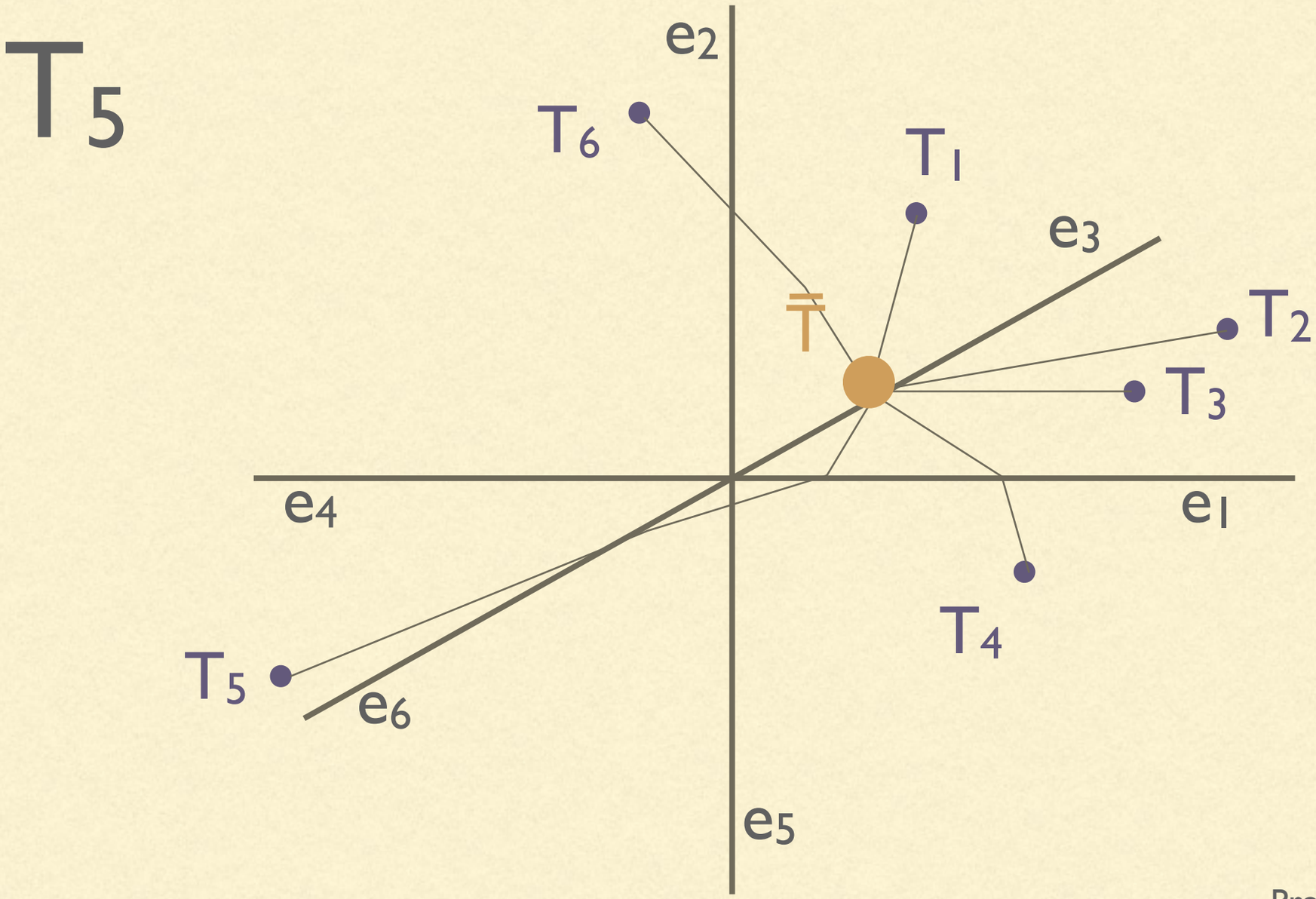
Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



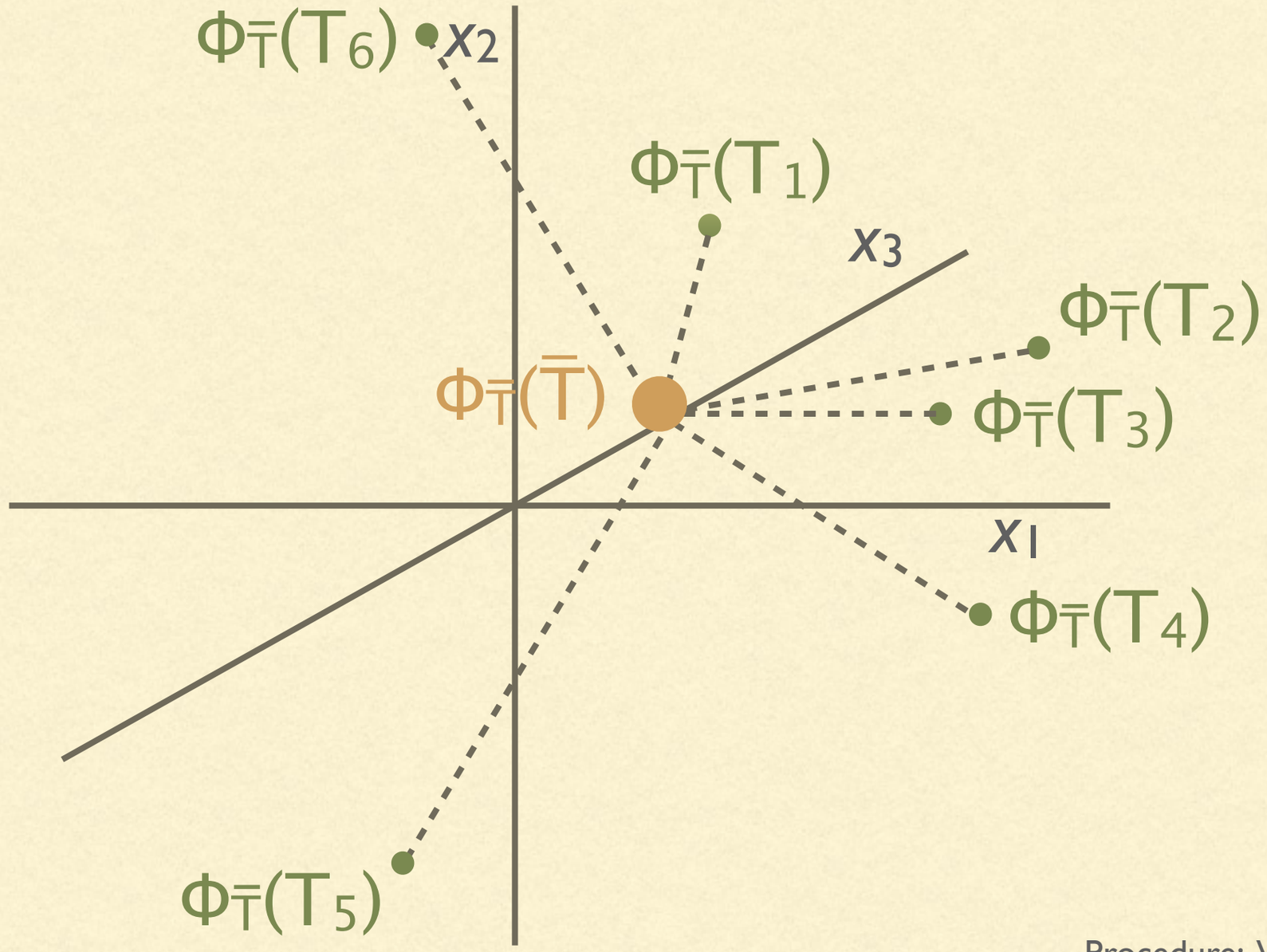
Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



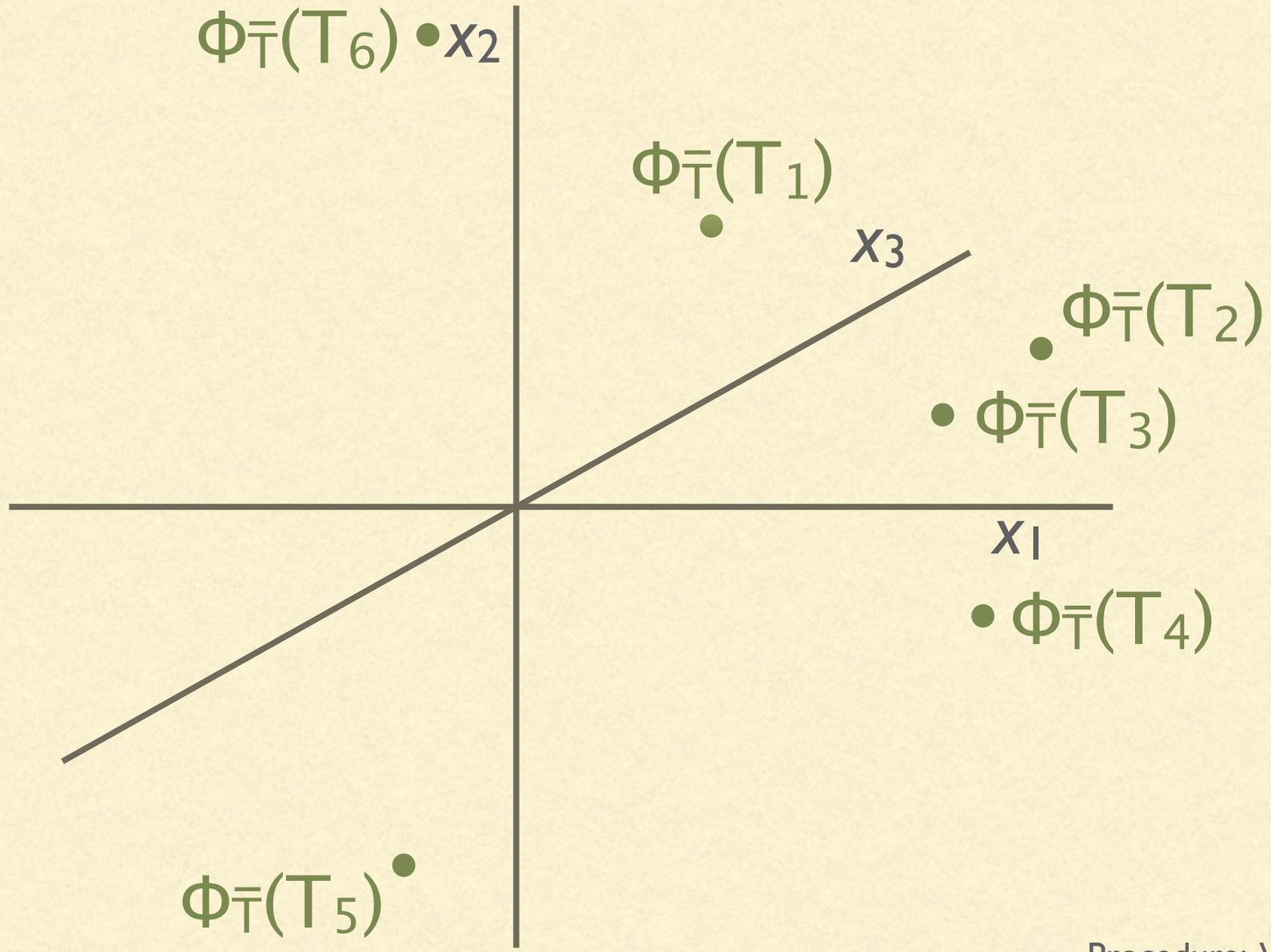
Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



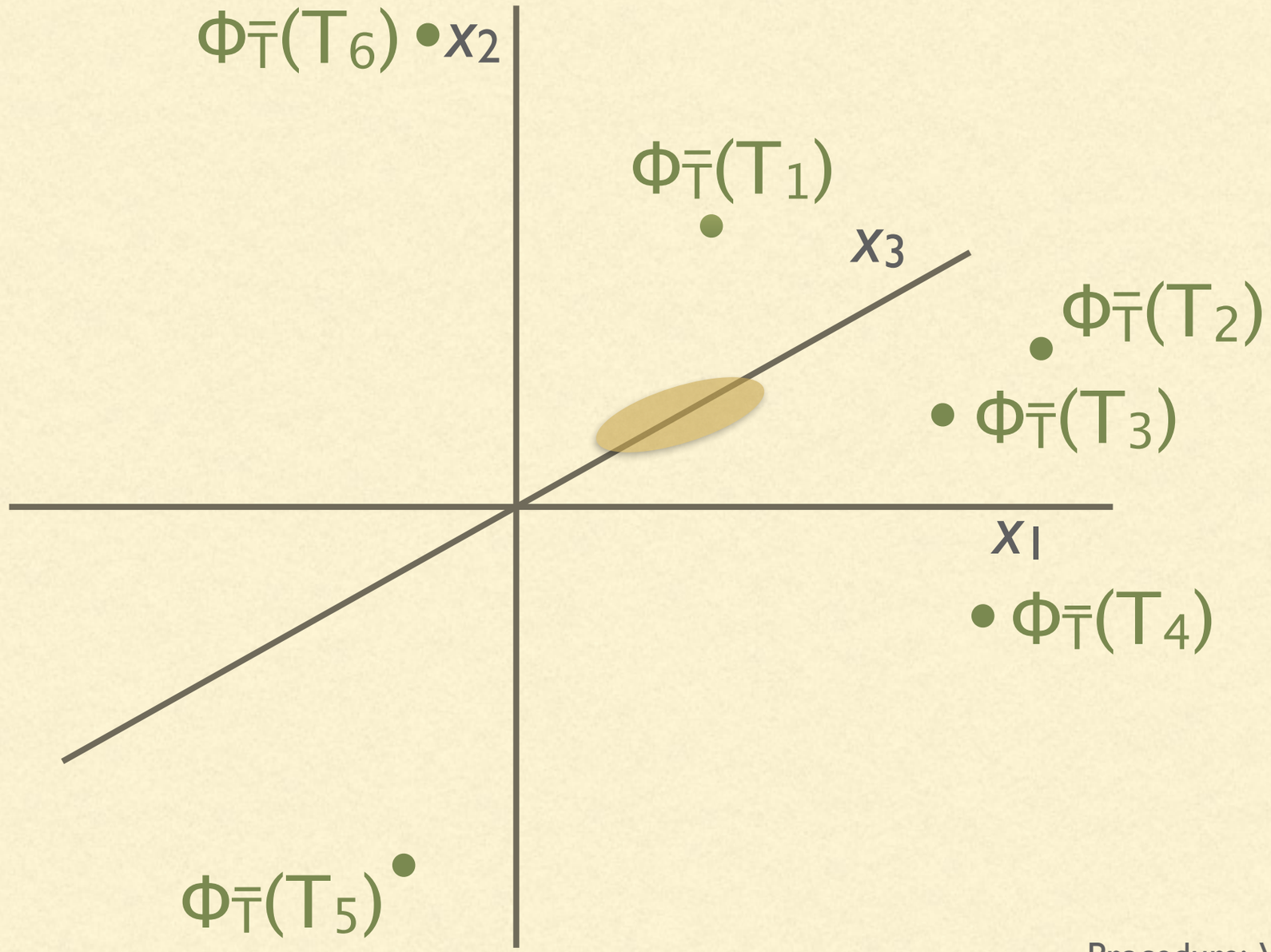
Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



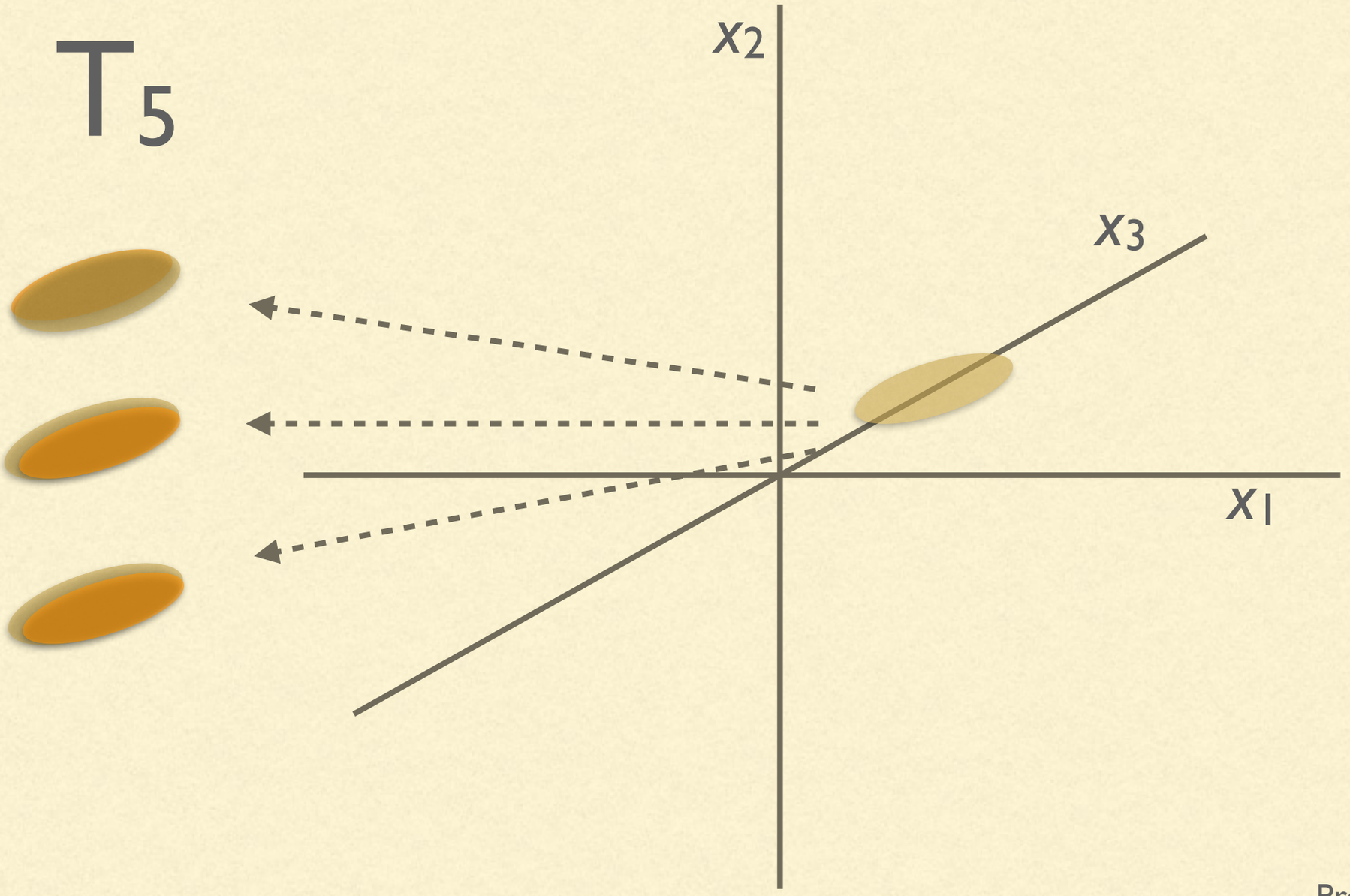
Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



Procedure: Willis, 2016

CONFIDENCE SETS FOR TREES



Procedure: Willis, 2016

SET PROPERTIES: VOLUME

- First confidence set for a tree!
- Multivariate variability is formally incorporated
- Sets are small (informative)

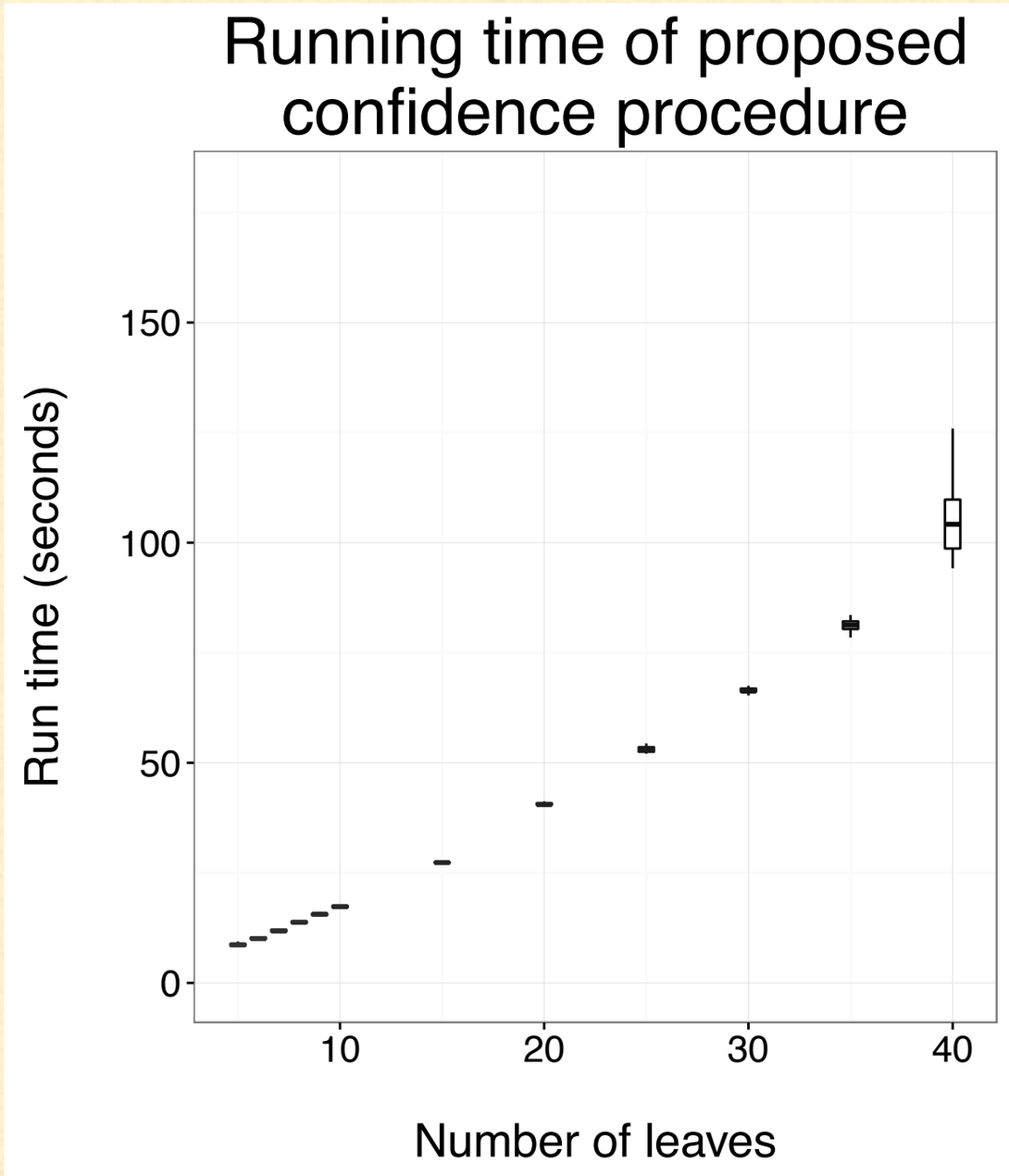
SET PROPERTIES: SPEED

- Construction: two steps

1. Calculate $\overline{T}$

   - Expensive!
   - Compute to within tolerance

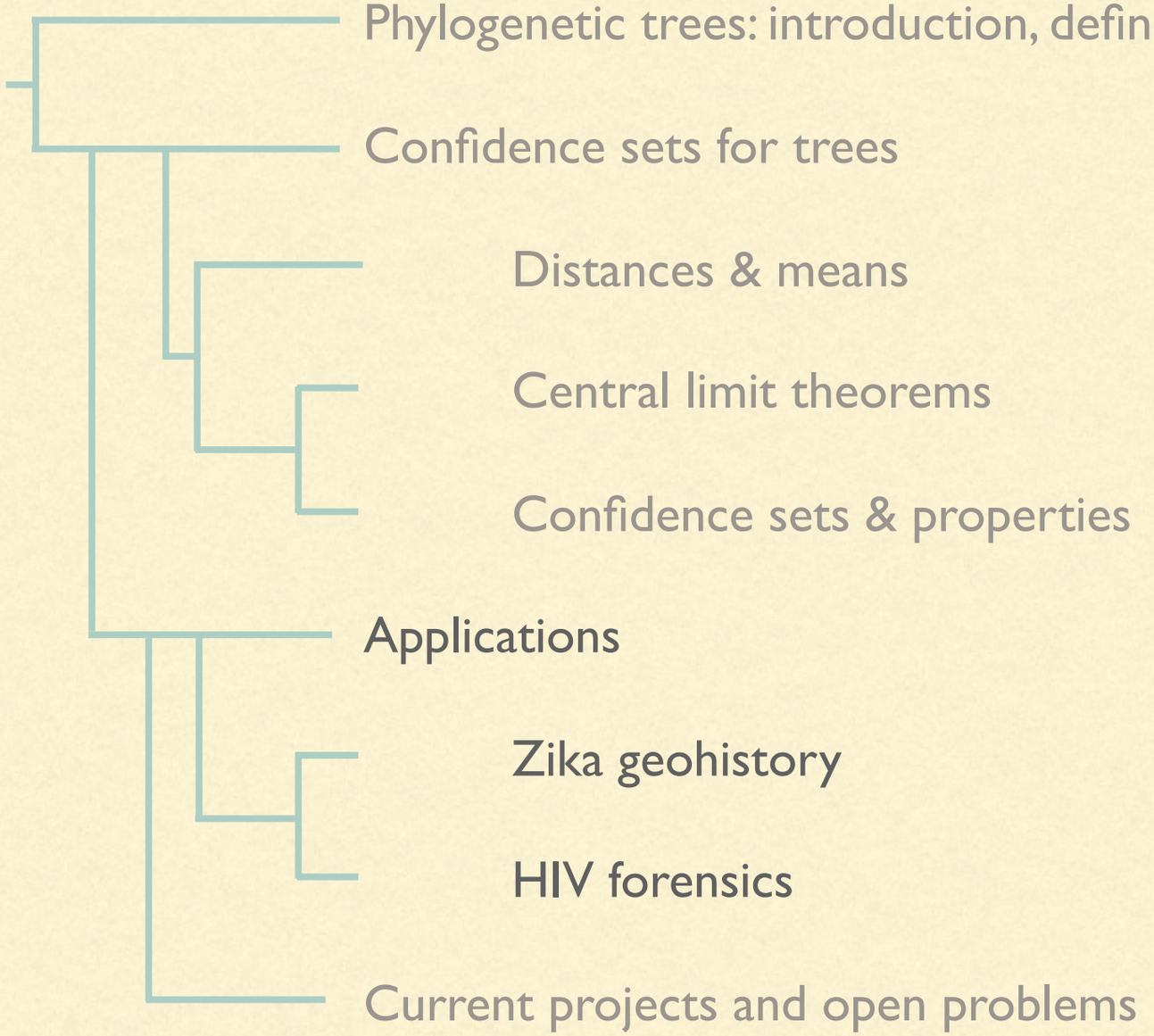
2. Calculate $\Phi_{\overline{T}}(T_i)$: O(nm⁴)



SET PROPERTIES: COVERAGE

- Simulations on Zika tree: m = 6, α = 0.05
  - 86% for n = 50, 87% for n = 100
- Mitigates concerns about structural assumptions
- Simulations incorporate uncertainty in tree building

OUTLINE

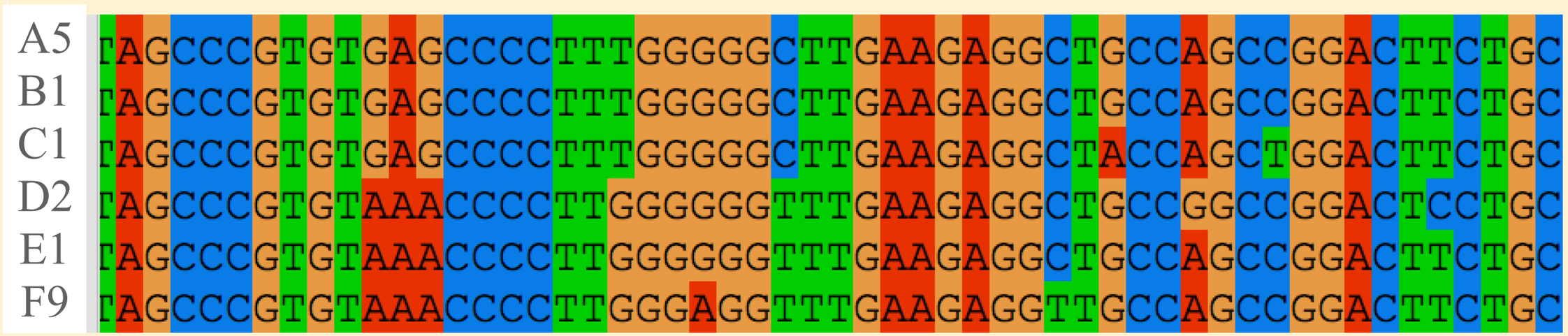

Phylogenetic trees: introduction, definition, motivation
Confidence sets for trees
Distances & means
Central limit theorems
Confidence sets & properties
Applications
Zika geohistory
HIV forensics
Current projects and open problems

ZIKA TREE

- Instances of Zika can be grouped by spatiotemporal properties

Africa pre-1984	A1… A13
Asia pre-1970	B1
Pacific pre-08	C1
Asia 2008-13	D1, D2, D3
Pacific 2008-14	E1
South America 2015-16	F1…F50

- Each instance is the sequenced genome of the Zika virus



ZIKA TREE

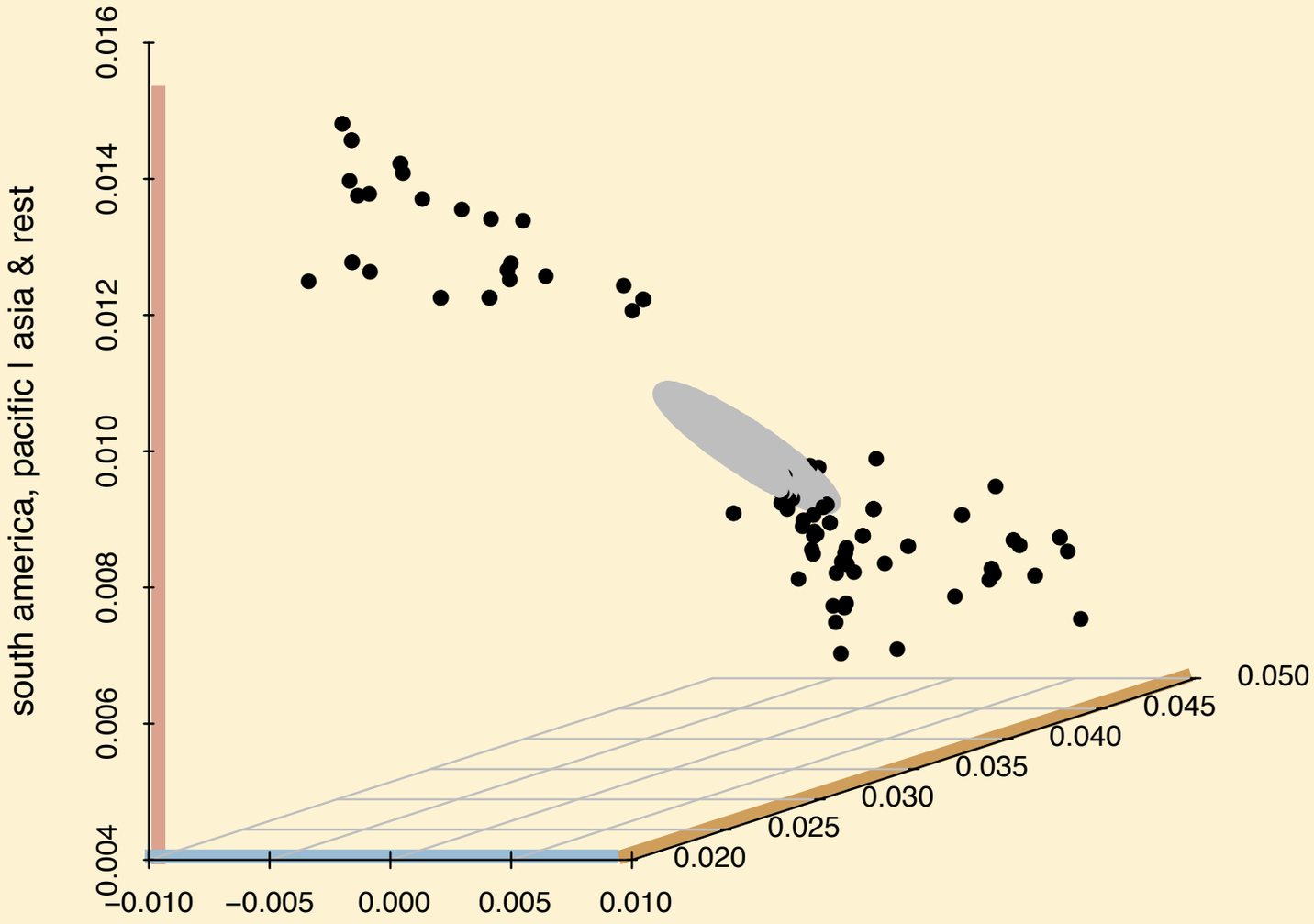
- By choosing one representative from each group, we can estimate the tree, T_1

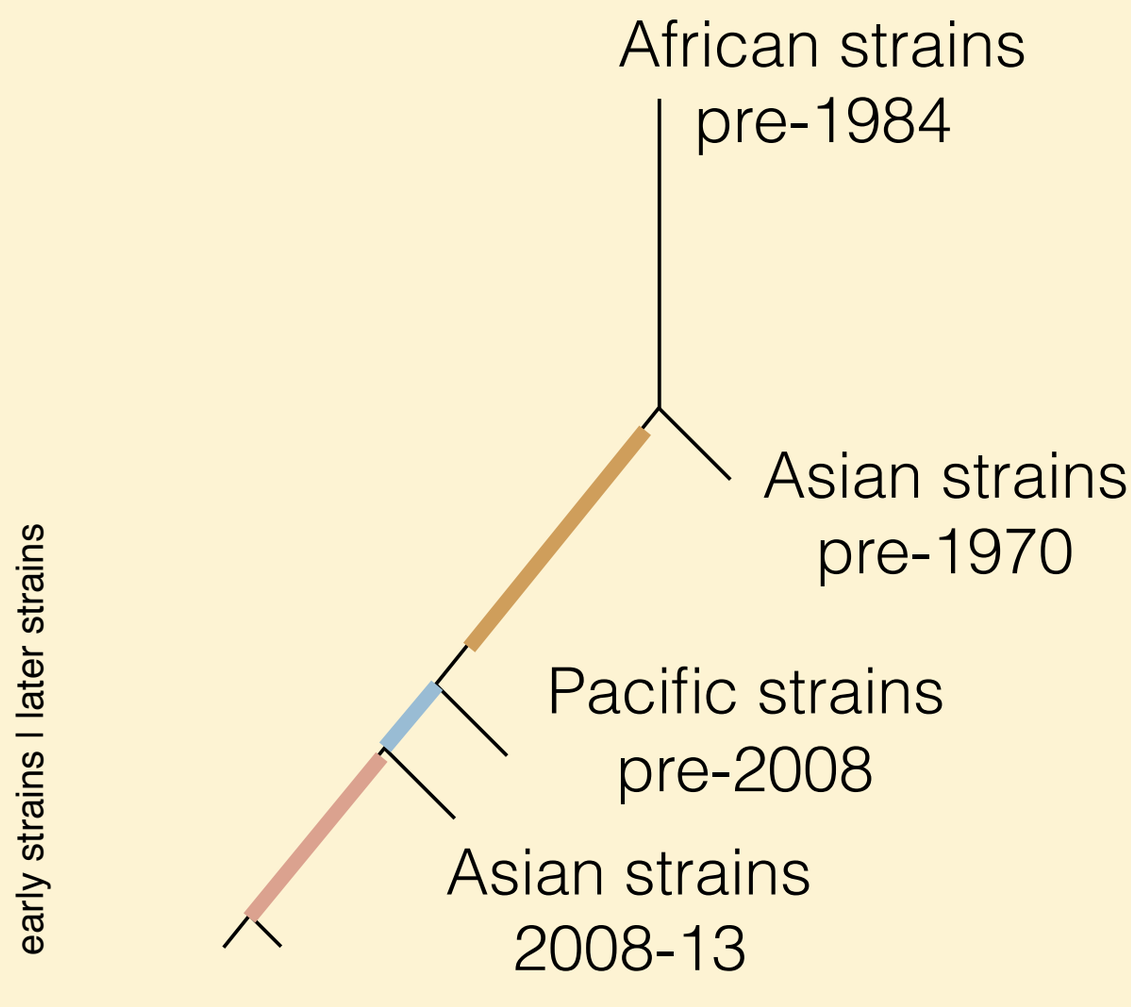
- By choosing a different representative from each group, we can estimate the tree again, T_2

- Repeating this for different representatives allows us to generate a collection of trees $T_1 \ldots T_m$

  - Trees reflect *within-group variability*

ZIKA TREE



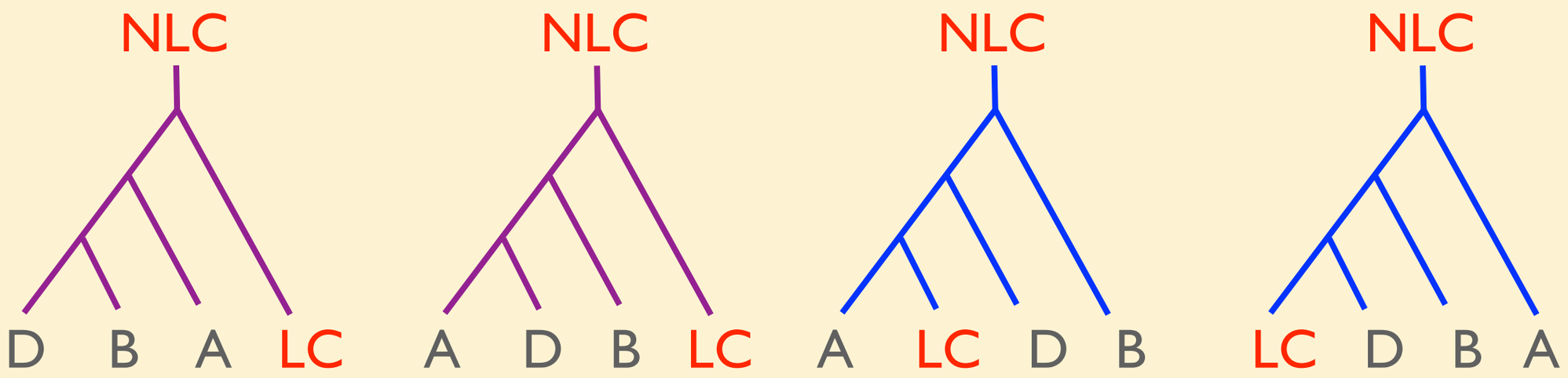


South America 2015-16

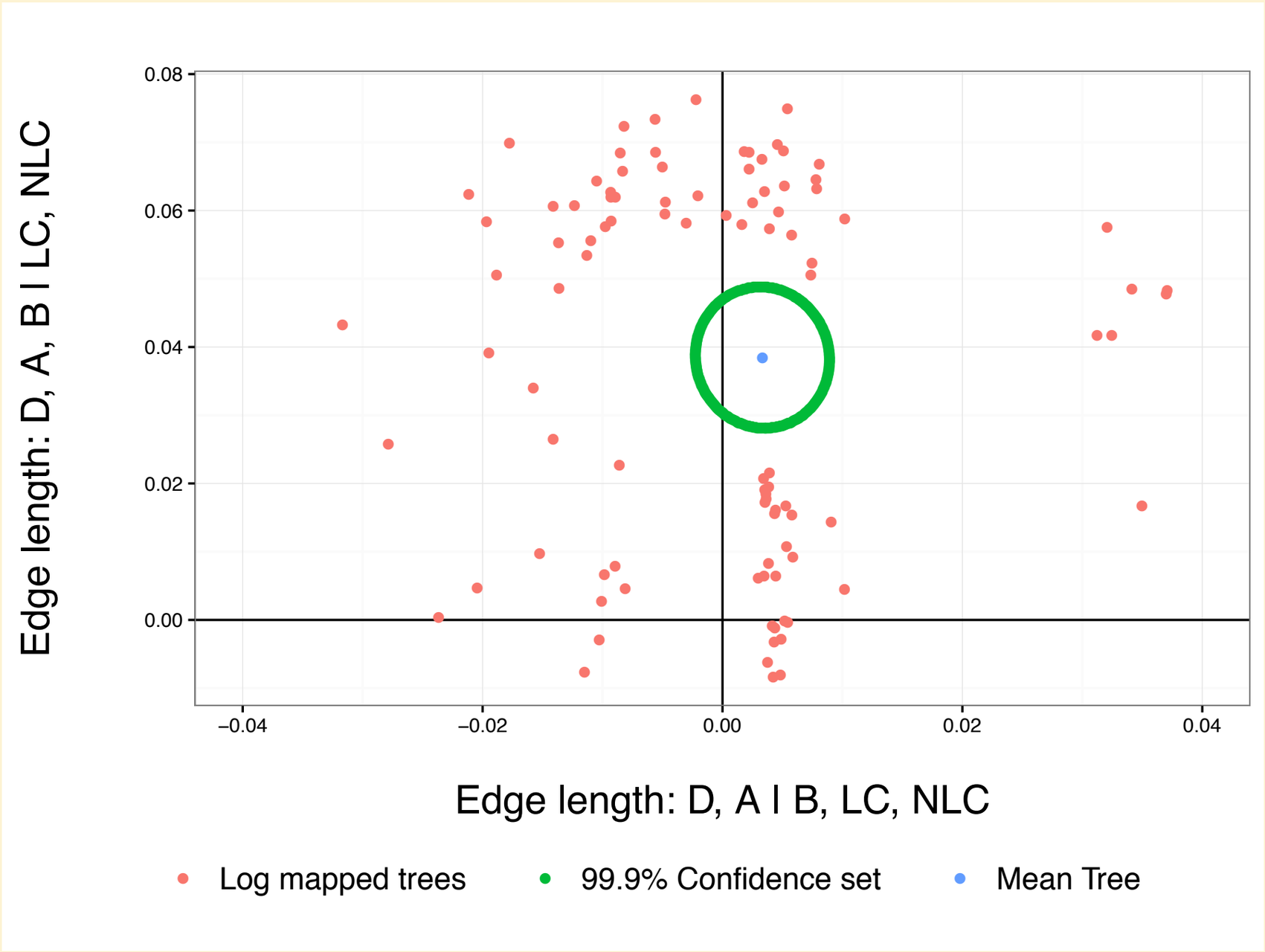
Pacific strains 2008-14

HIV FORENSICS

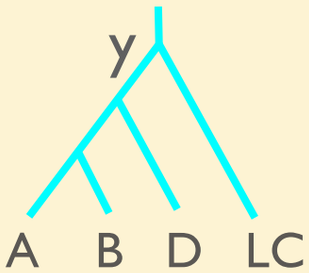
- 19 yo with no risk factors found to have HIV
- Transmission event hypothesized to originate from dentist
- Multiple sequences obtained from patients (A & B), dentist (D), local controls (LC) and non-local controls (NLC)

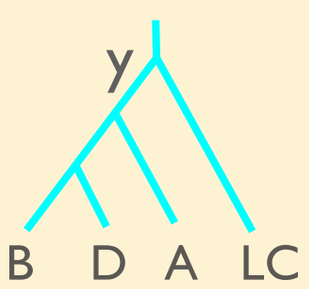


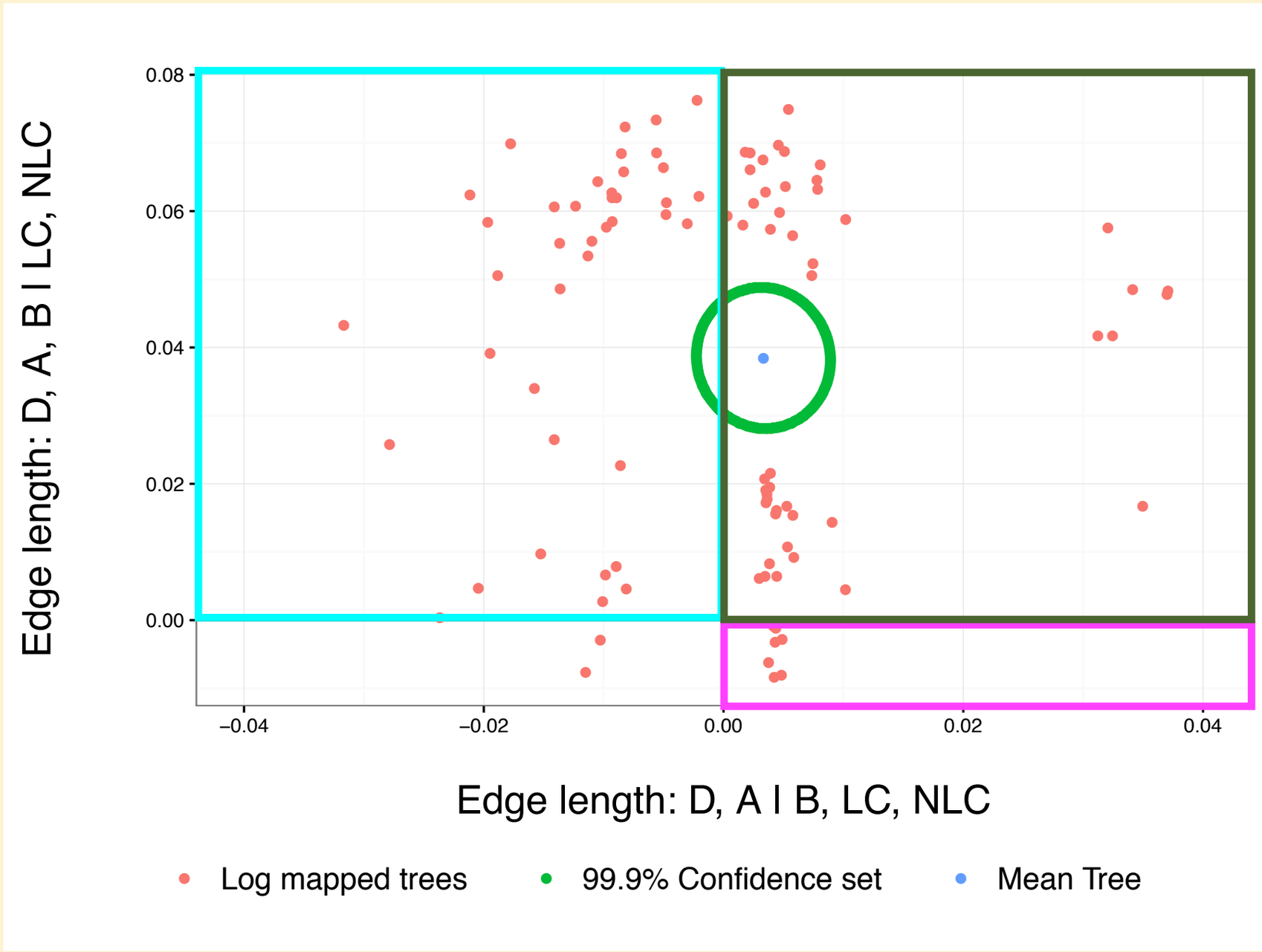
HIV FORENSICS

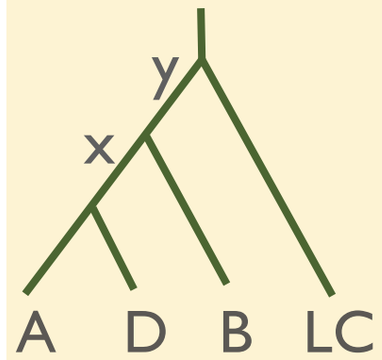


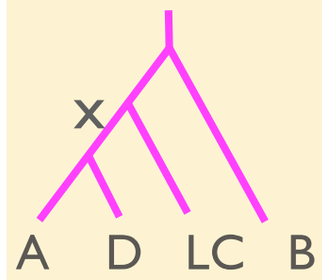
HIV FORENSICS

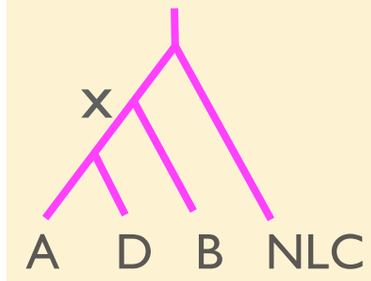












OUTSTANDING QUESTIONS

- Trees are estimated with noise: heteroskedastic case
- CLT extension to non-independence
- Coverage robustness with heavier tails
- Other types of sets eg. minimum diameter, log map-free…

PHYLOGENETICS QUESTIONS

- "All of Statistics" on $\mathcal{T}_m$
- Incorporating both bias and variance (risk-optimal shrinkage trees?)
- Experimental design in $\mathcal{T}_m$
- Tree variability to give errors on tree summary statistics (eg. BWPD)

CONFIDENCE SETS FOR PHYLOGENETIC TREES

Amy Willis
~~Department of Statistical Science, Cornell University~~
Department of Biostatistics, University of Washington

SUPPLEMENTARY FIGURES

Amy Willis

Department of Biostatistics, University of Washington

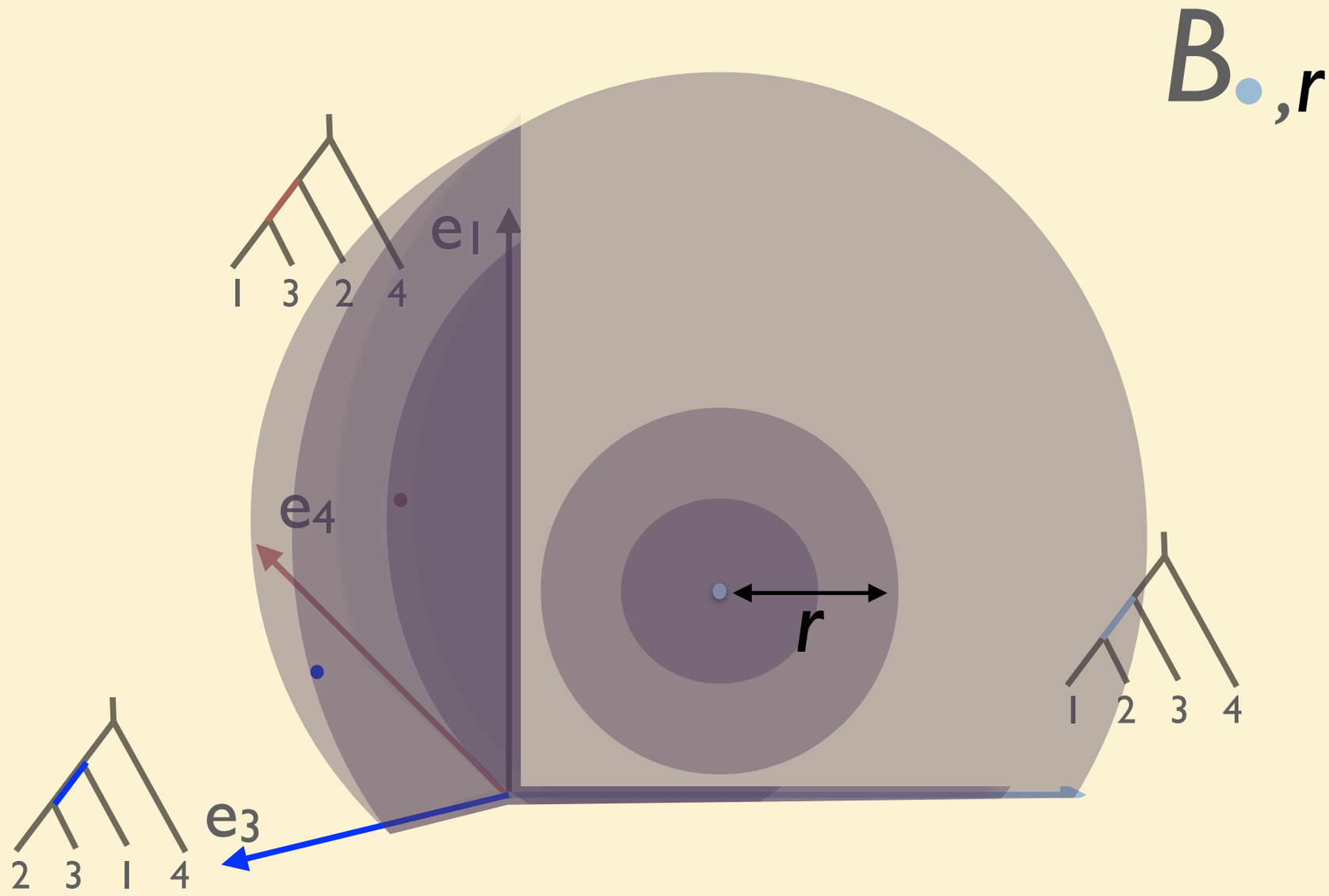
NOISY TREES

- But trees are estimated!

- Brownian motion (BM) on tree space (Nye, 2015) provides a framework for considering trees as noisy

  - Tree BM behaves like Euclidean BM within orthants, but with uniform-at-random orthant crossings

NOISE MODEL



Brownian motion on tree space

Construction: Nye, 2016

NOISY CENTRAL LIMIT THEOREM

- Theorem: (Willis, 2017)

Consider $\{S_i\}_{i=1\ldots n}$ drawn iid from F with Fréchet mean T^*.

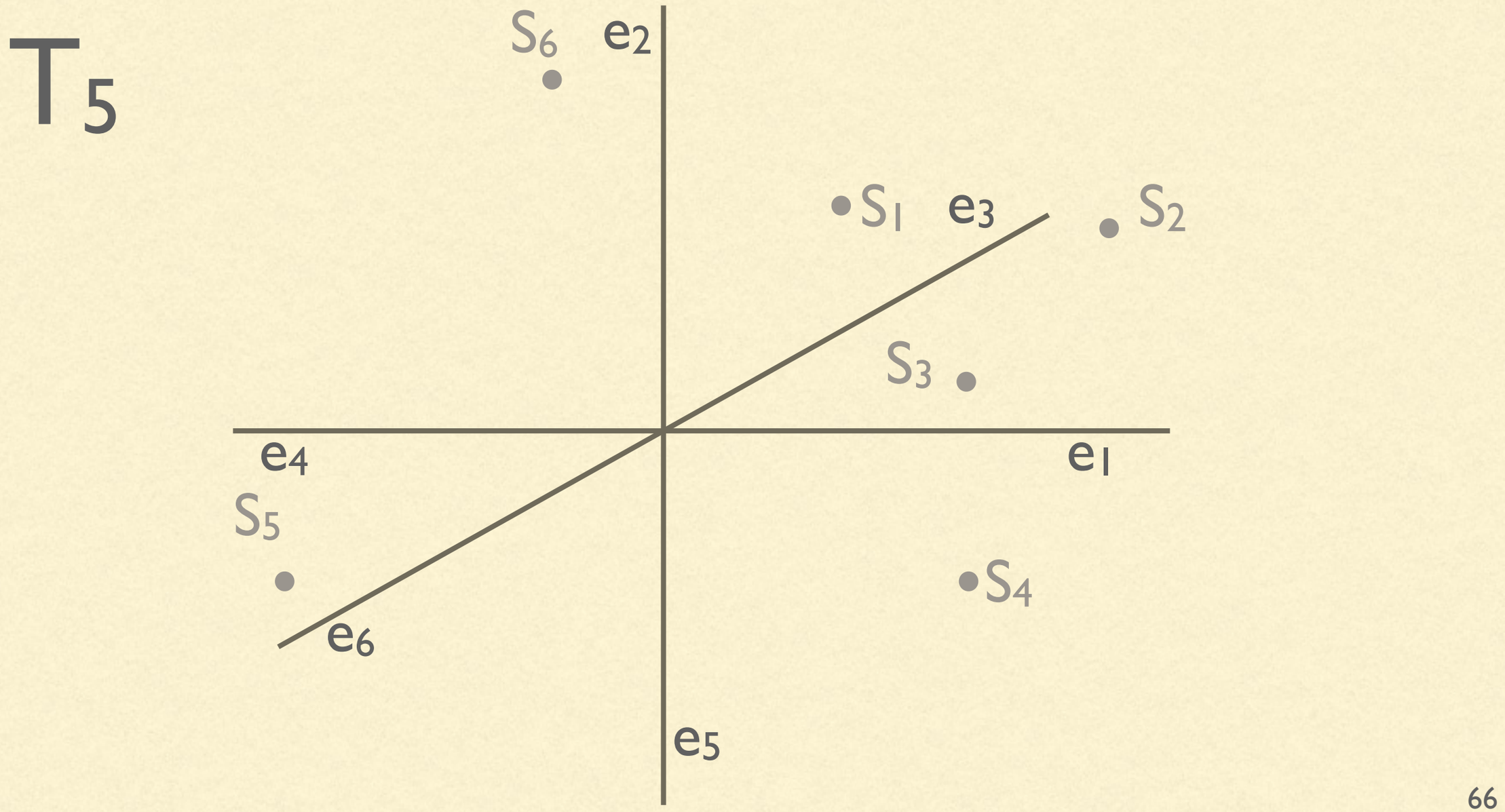
Consider perturbing each tree by a Brownian motion: $T_i = B_{S_i,r}$.

Then $\hat{T}_n$ *is consistent for* T^* *and as* n becomes large,

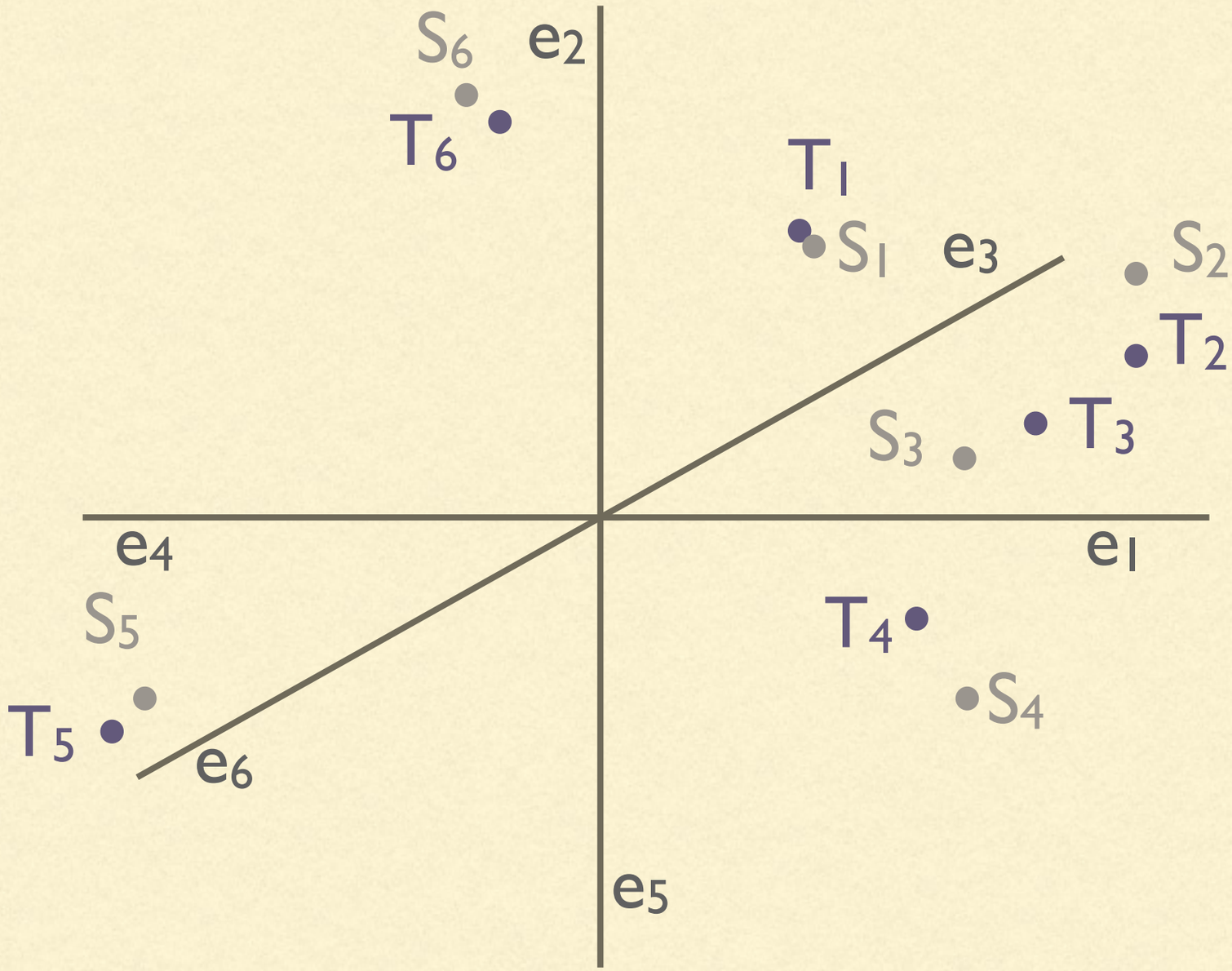
$$\sqrt{n}\left(\Phi_{\hat{T}_n}(\hat{T}_n) - \Phi_{T^*}(T^*)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$$

where Σ depends on F and r

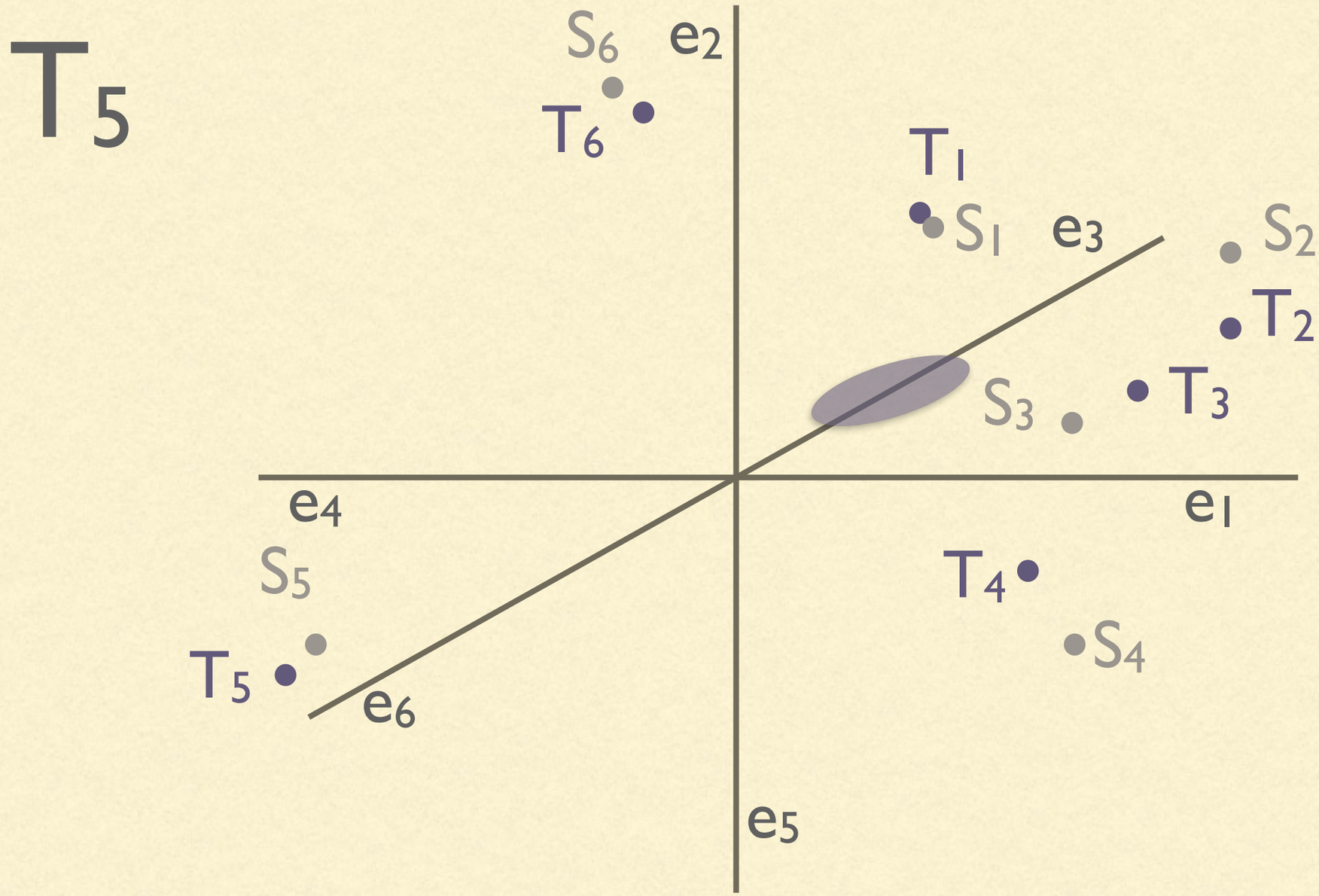
CONFIDENCE SETS FOR TREES



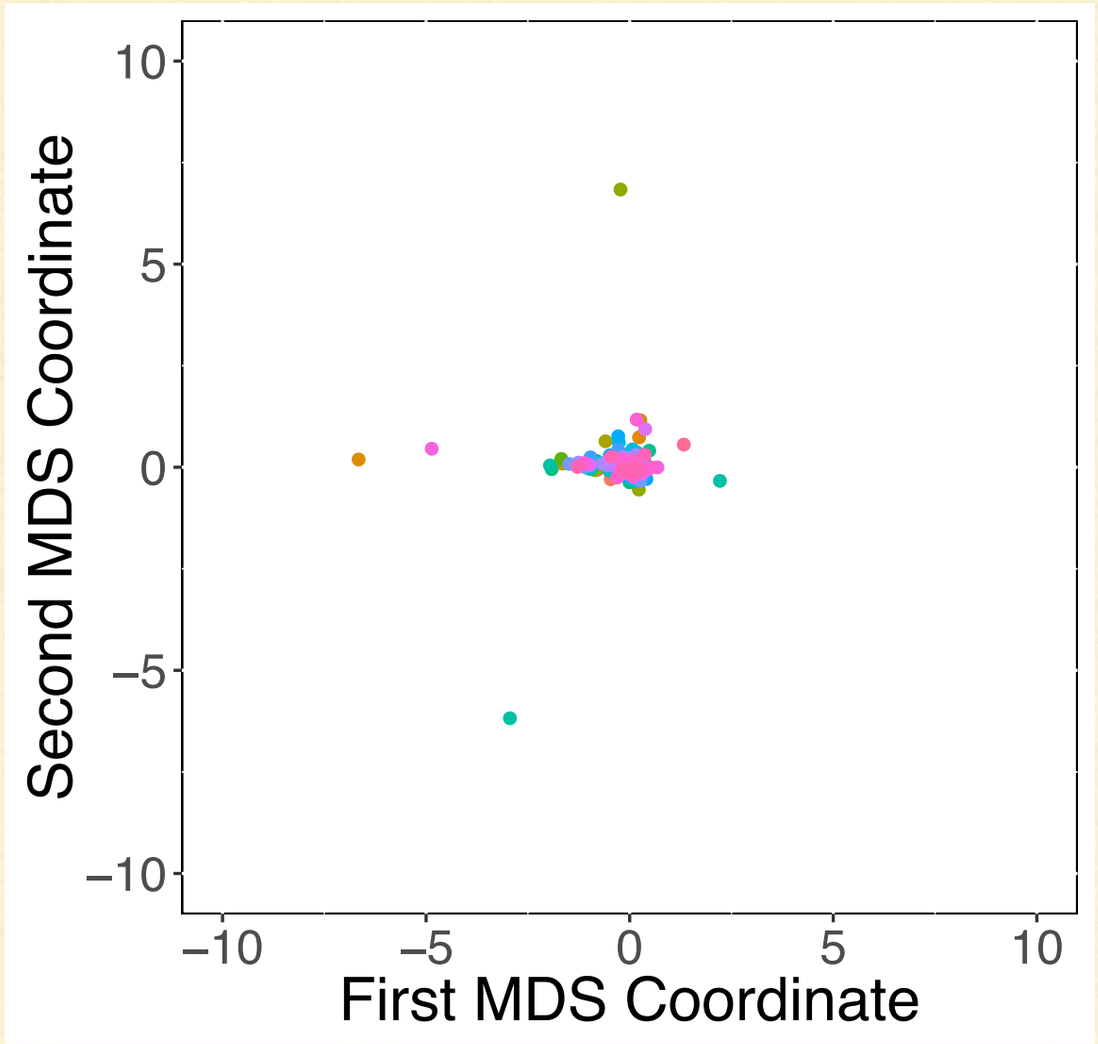
CONFIDENCE SETS FOR TREES



CONFIDENCE SETS FOR TREES

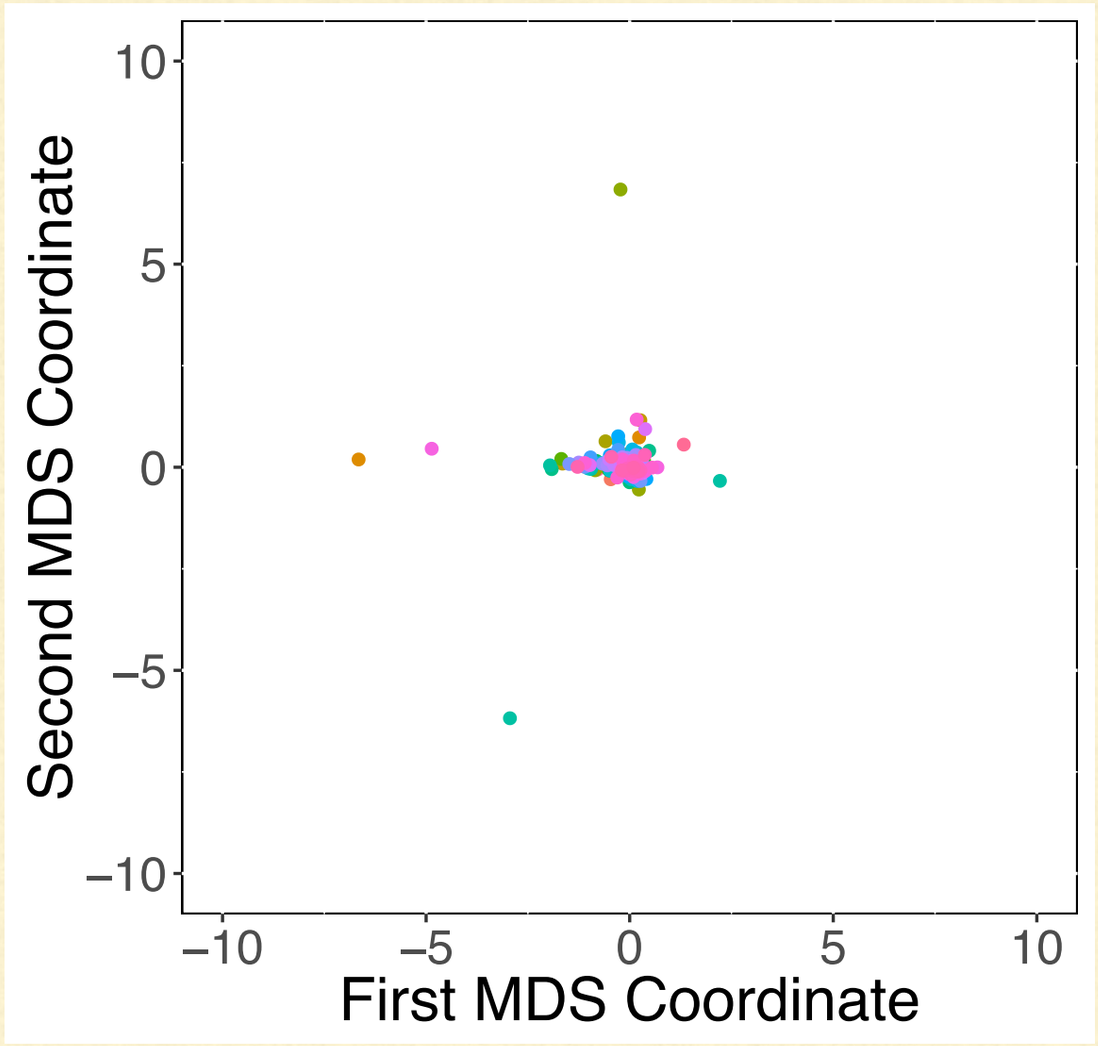


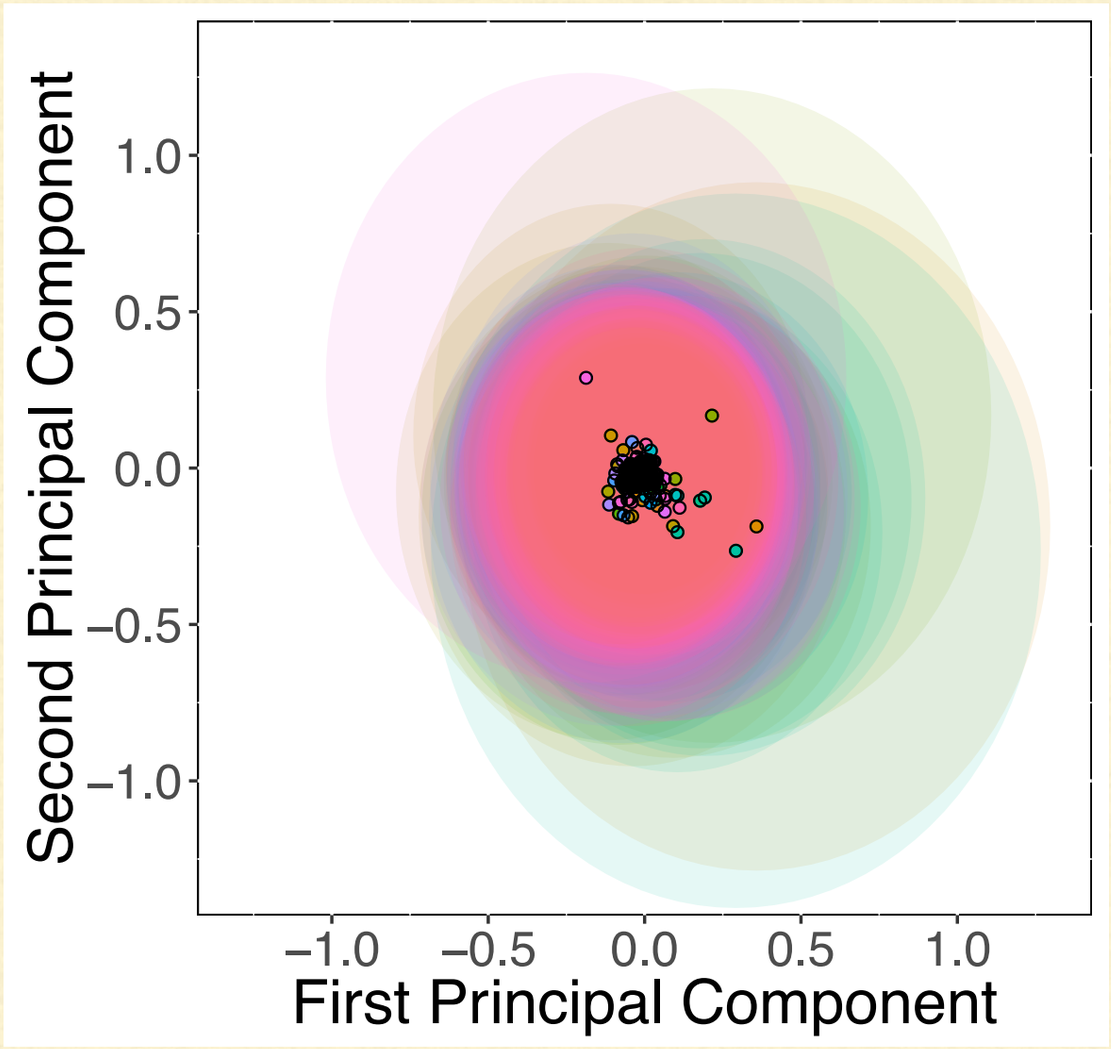
VISUALIZING UNCERTAINTY



Willis & Bell, 2017

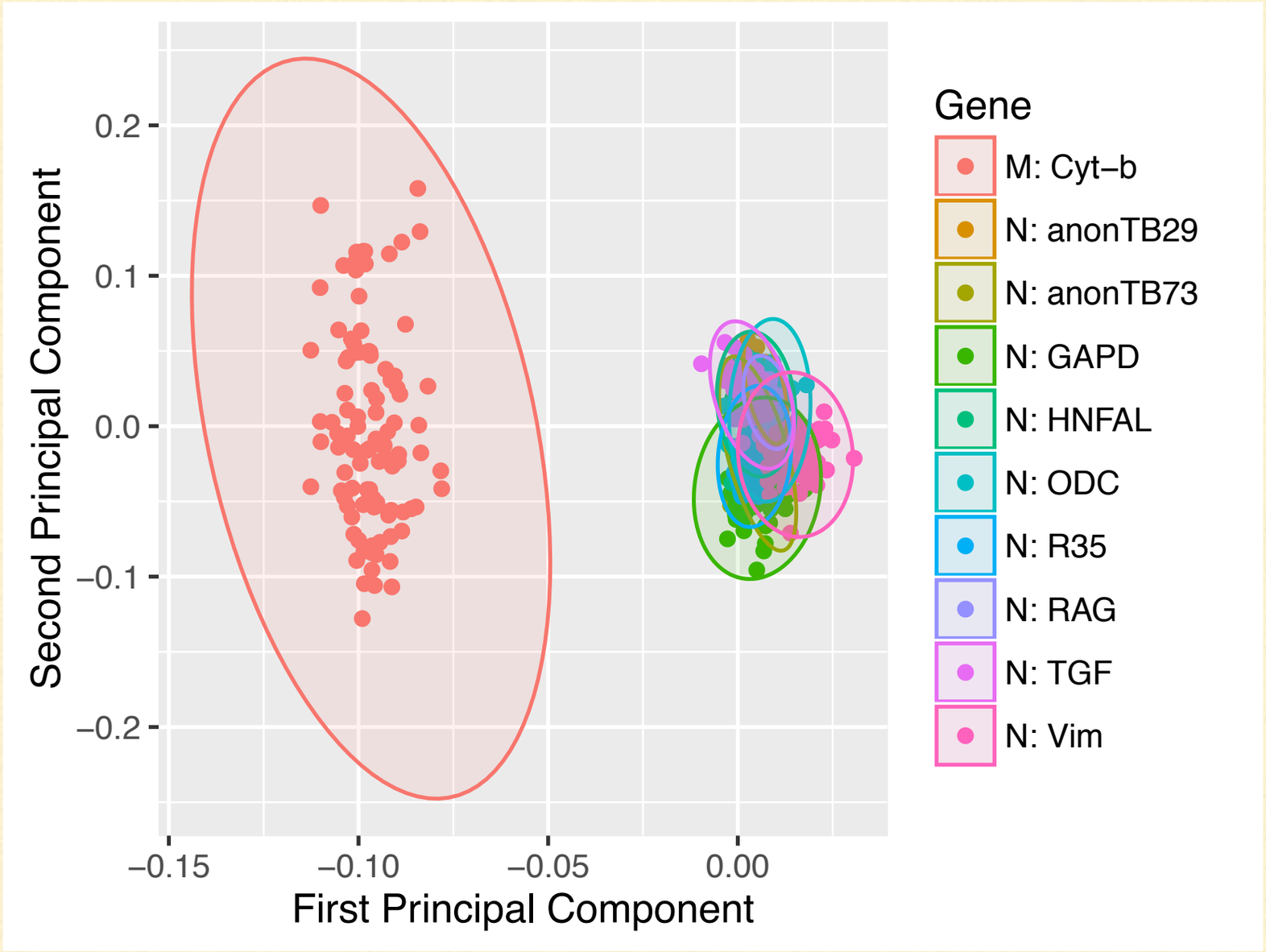
VISUALIZING UNCERTAINTY





Willis & Bell, 2017

VISUALIZING UNCERTAINTY



Willis & Bell, 2016