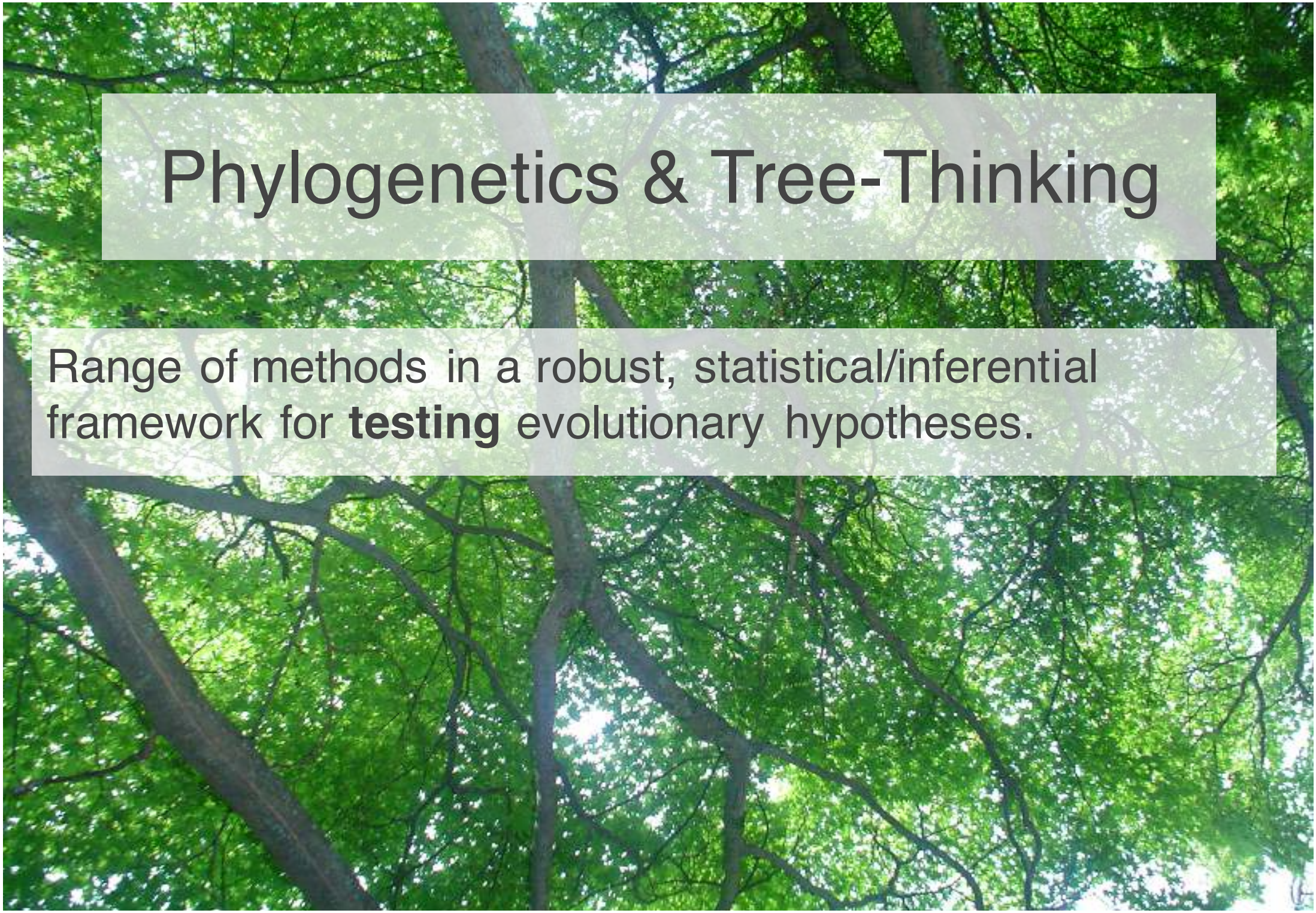
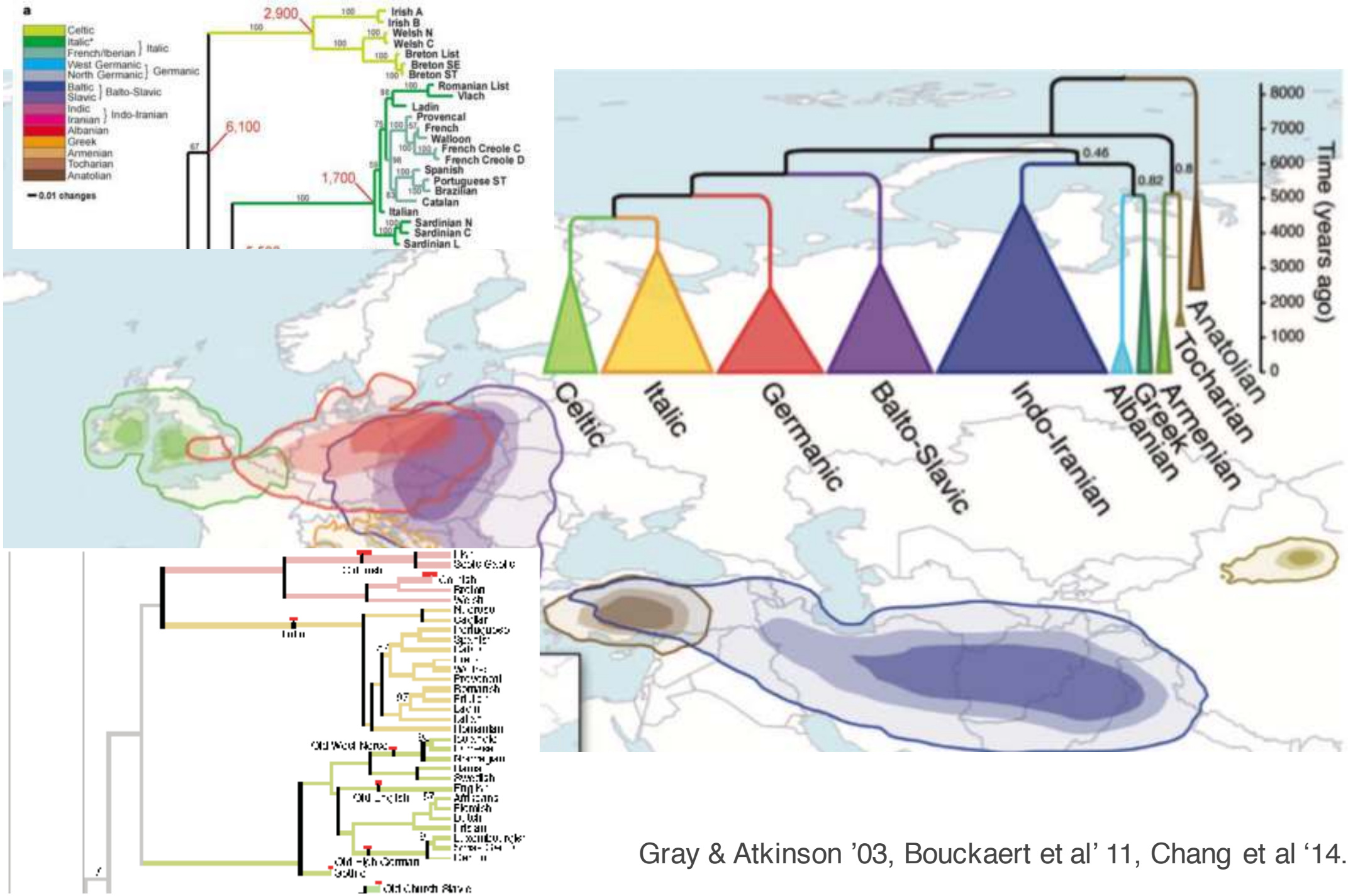
Phylogenetics & Tree-Thinking

Simon J. Greenhill

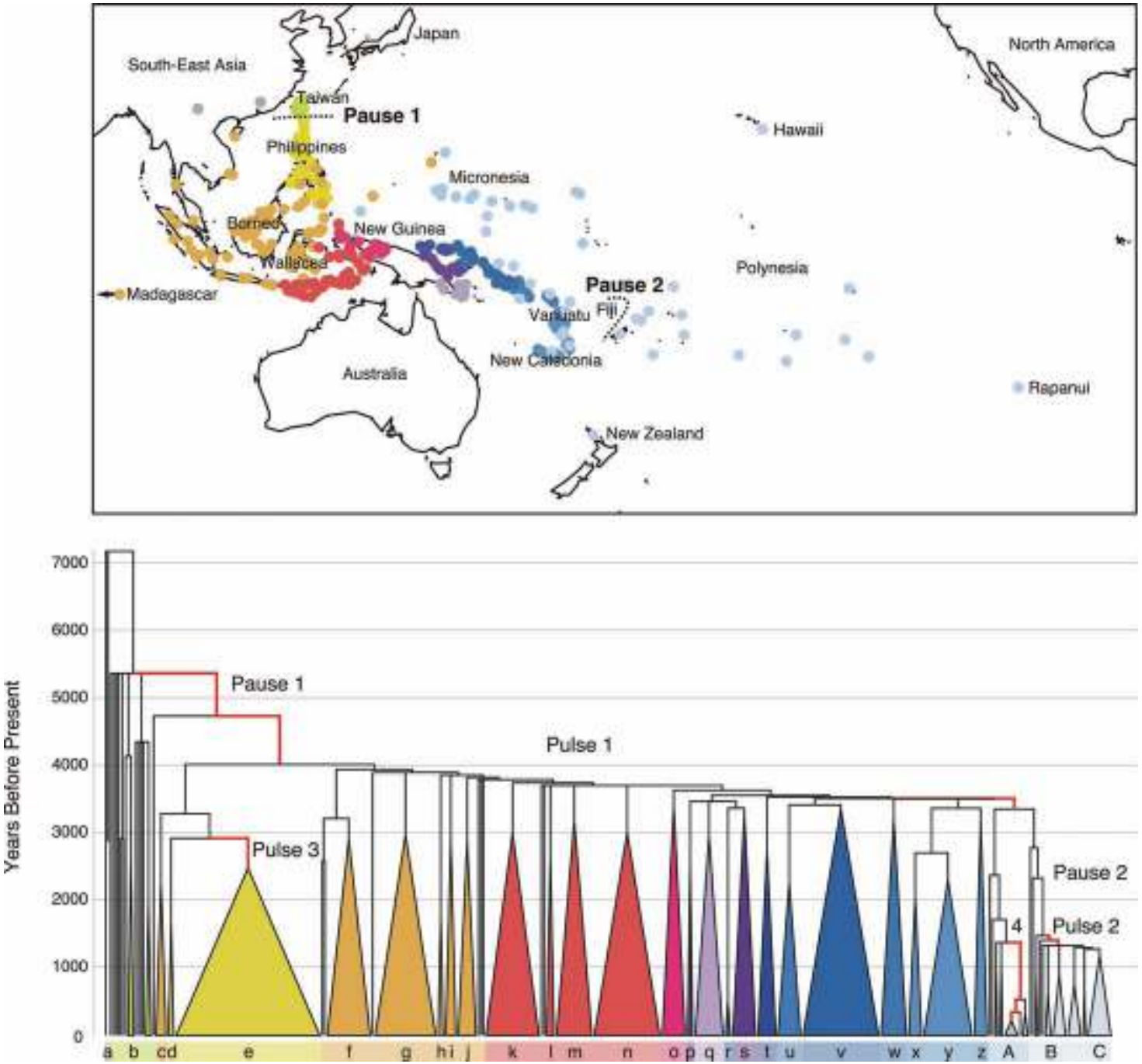
ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Max Planck Institute for the
Science of Human History

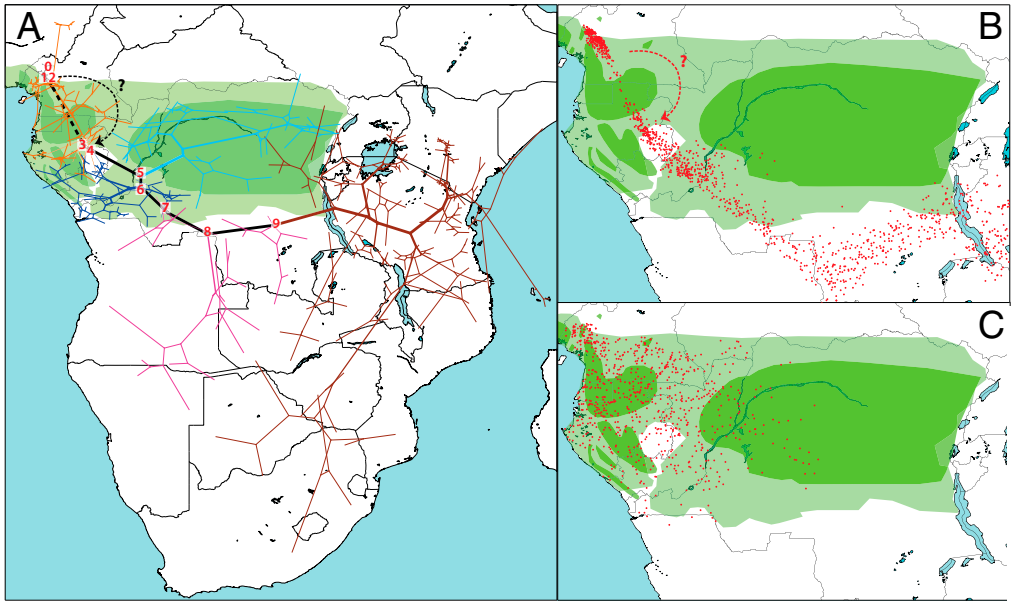

Phylogenetics & Tree-Thinking
Range of methods in a robust, statistical/inferential framework for **testing** evolutionary hypotheses.

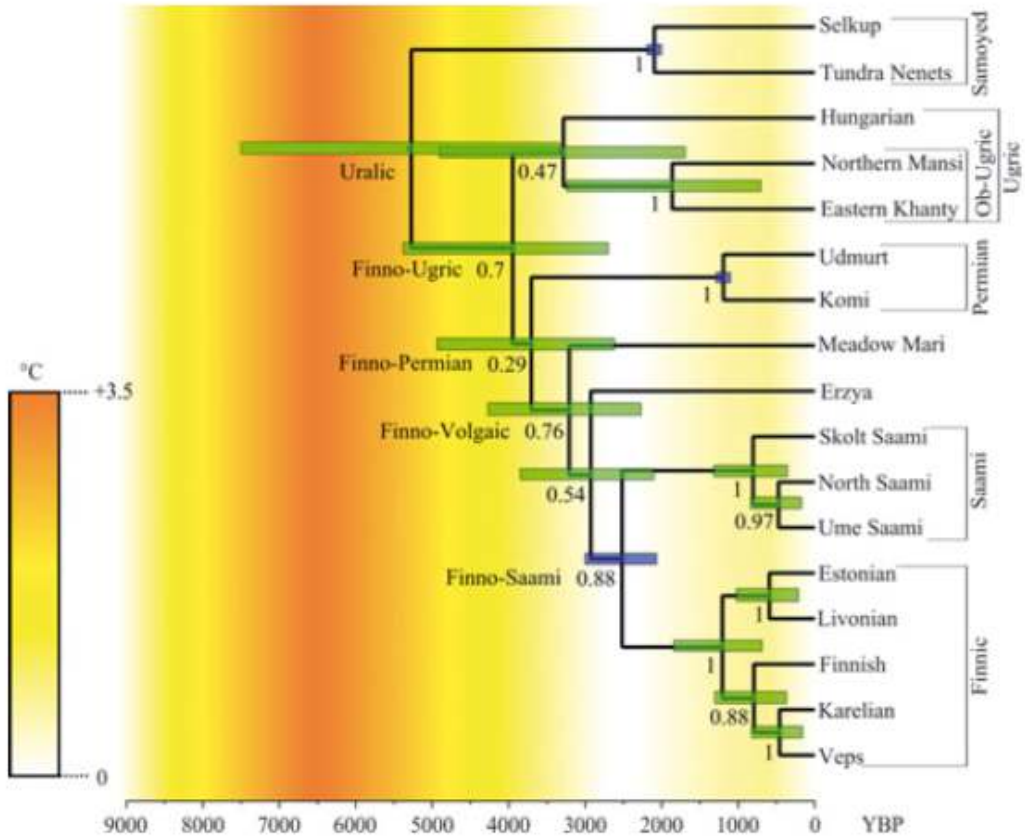


Gray & Atkinson ’03, Bouckaert et al’ 11, Chang et al ‘14.


Japan
South-East Asia
North America
Taiwan
Pause 1
Hawaii
Philippines
Micronesia
Borneo
New Guinea
Wallacea
Polynesia
Pause 2
Madagascar
Vanuatu
Fiji
New Caledonia
Australia
Rapanui
New Zealand
Years Before Present
7000
6000
5000
4000
3000
2000
1000
0
Pause 1
Pulse 1
Pulse 3
Pause 2
4
Pulse 2
a b cd e f g hi j k l m n o p q r s t u v w x y z A B C

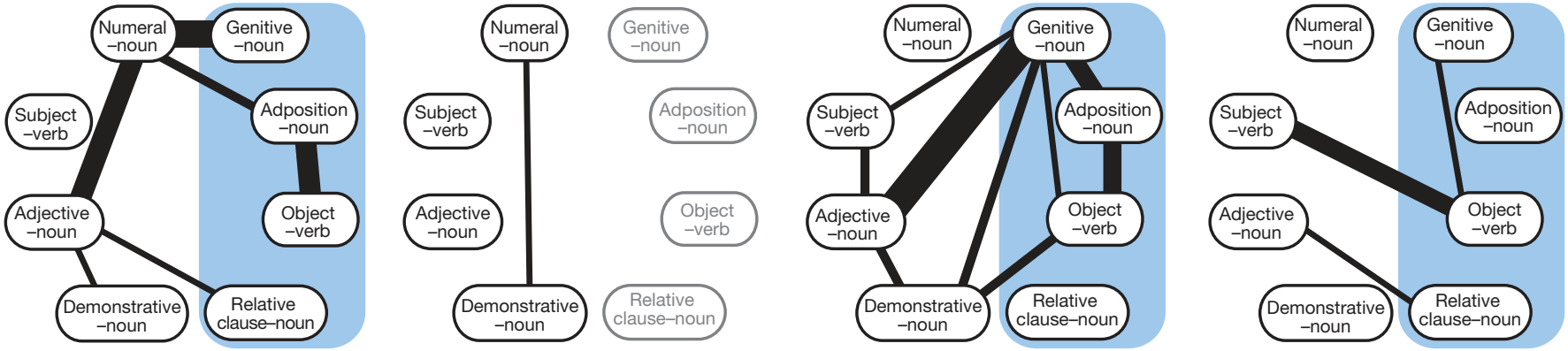
Gray et al ‘09



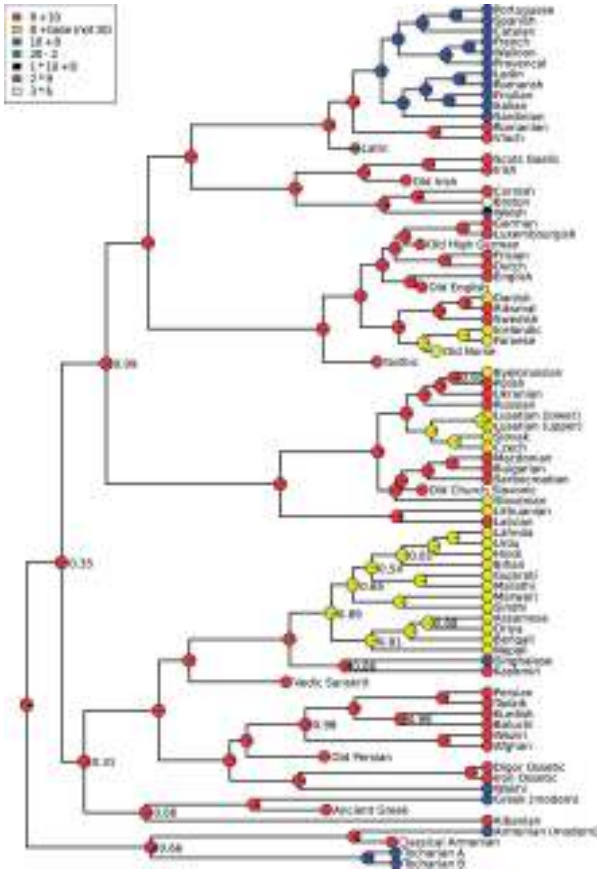


Grollemund et al. ‘15

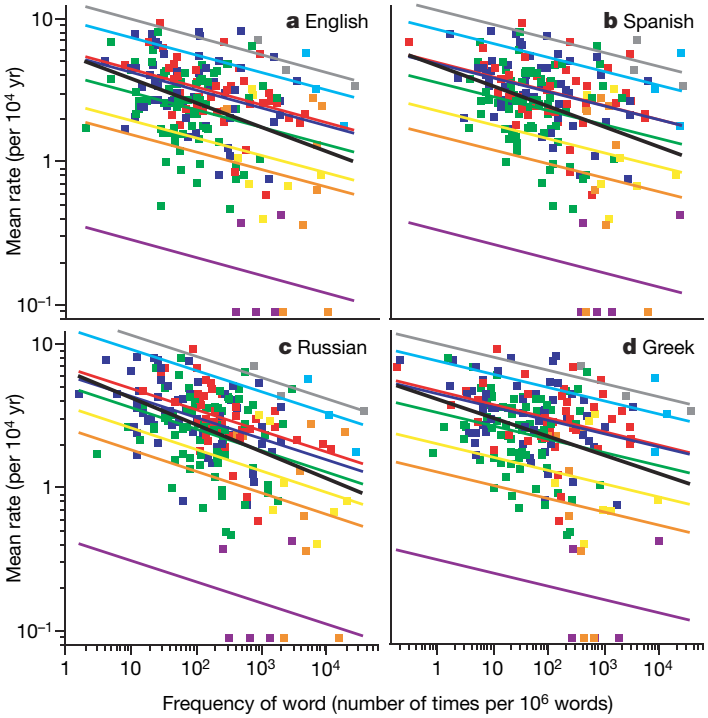
Honkola et al ‘13



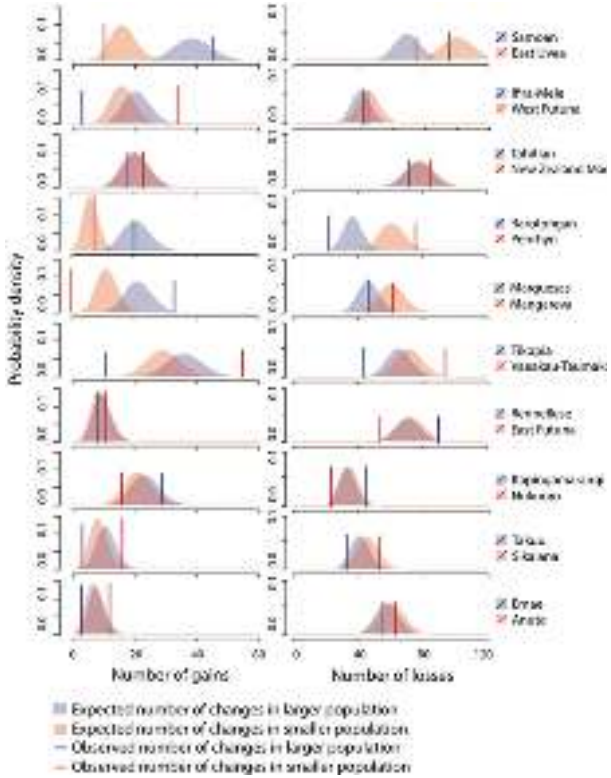
Dunn et al. ‘15



Calude and Verkerk ‘16

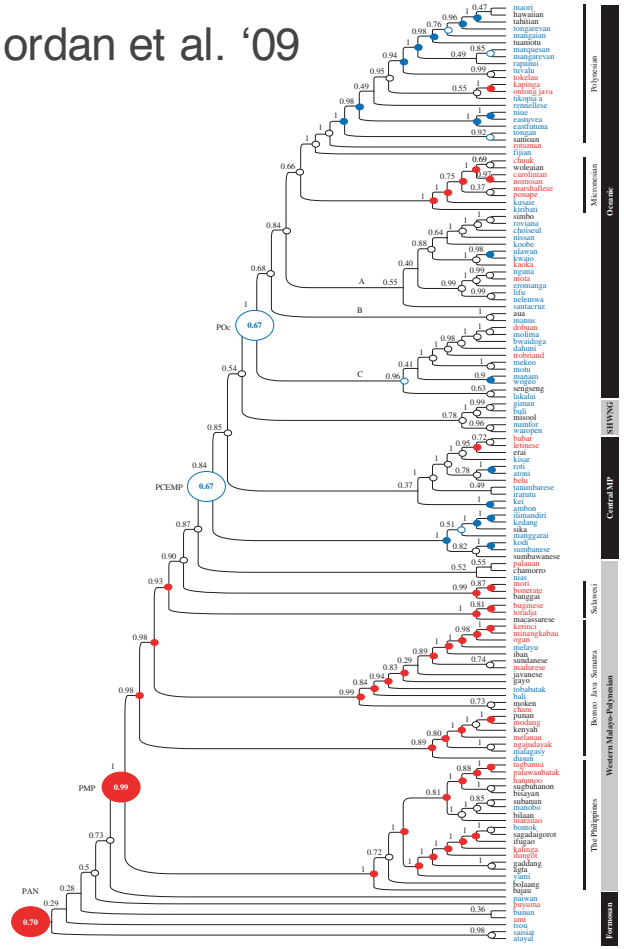


Pagel et al ’07

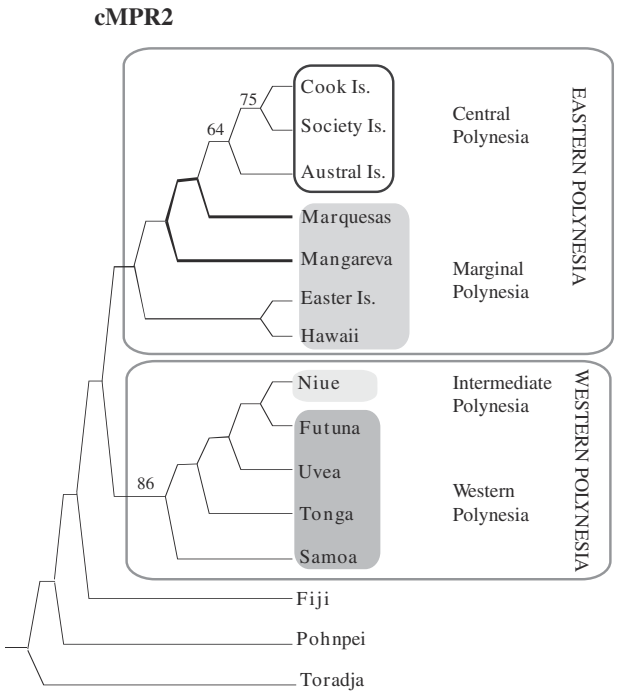


Bromham et al. ’15

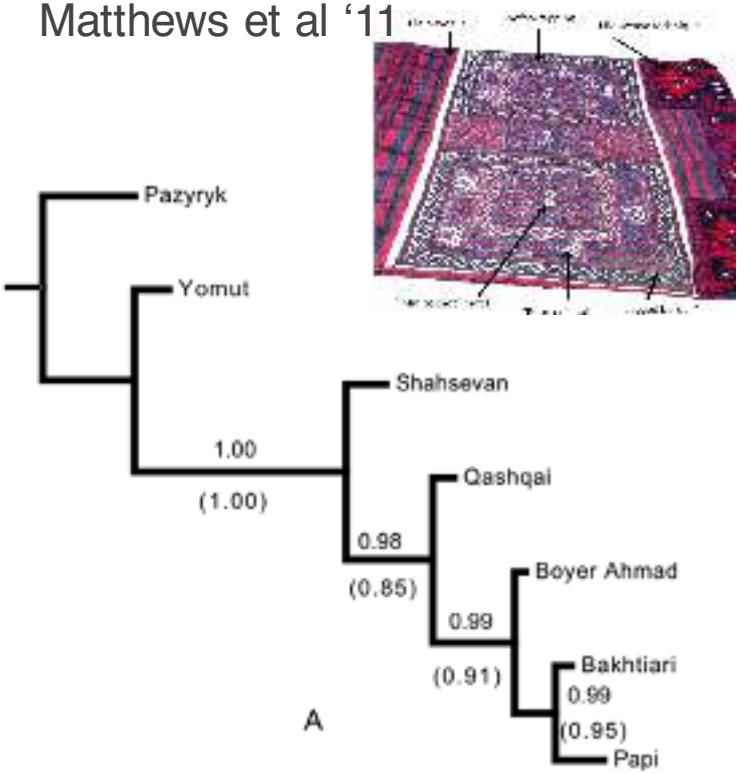
Jordan et al. ‘09

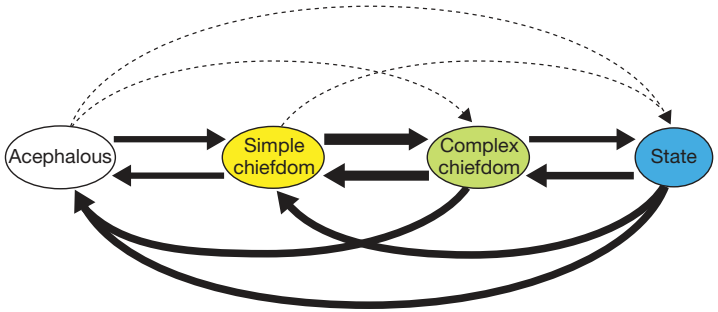


Larsen ‘11

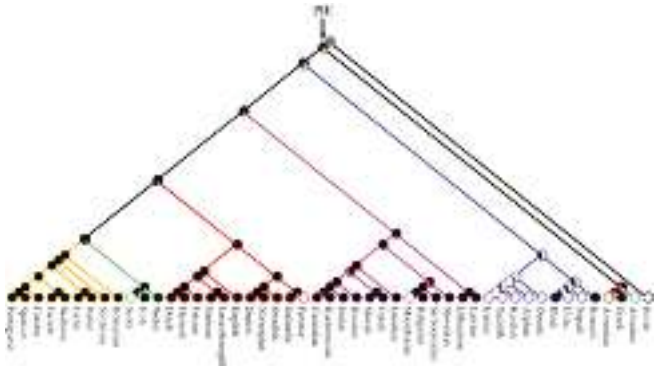


Matthews et al ‘11

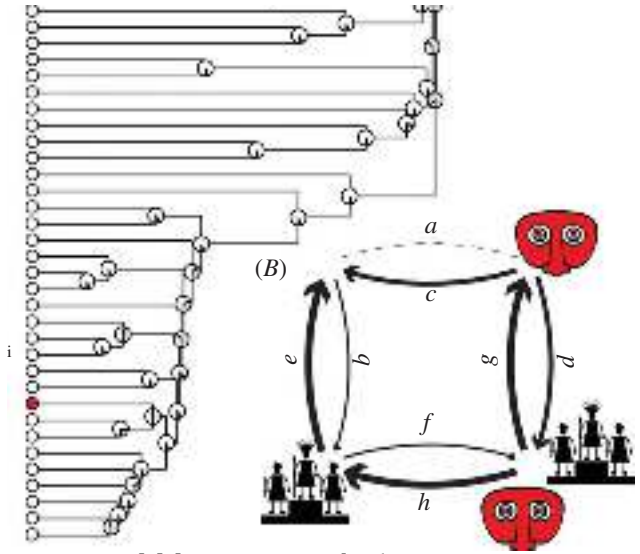




Currie et al. ‘10



Da Silva & Tehrani ‘16



Watts et al. ‘15

Controversial

"most vibrant stream of contemporary linguistics"

"Computational methodologies of this kind can only be helpful for historical linguistics"

"languages and biomolecular sequences evolve in very different ways"

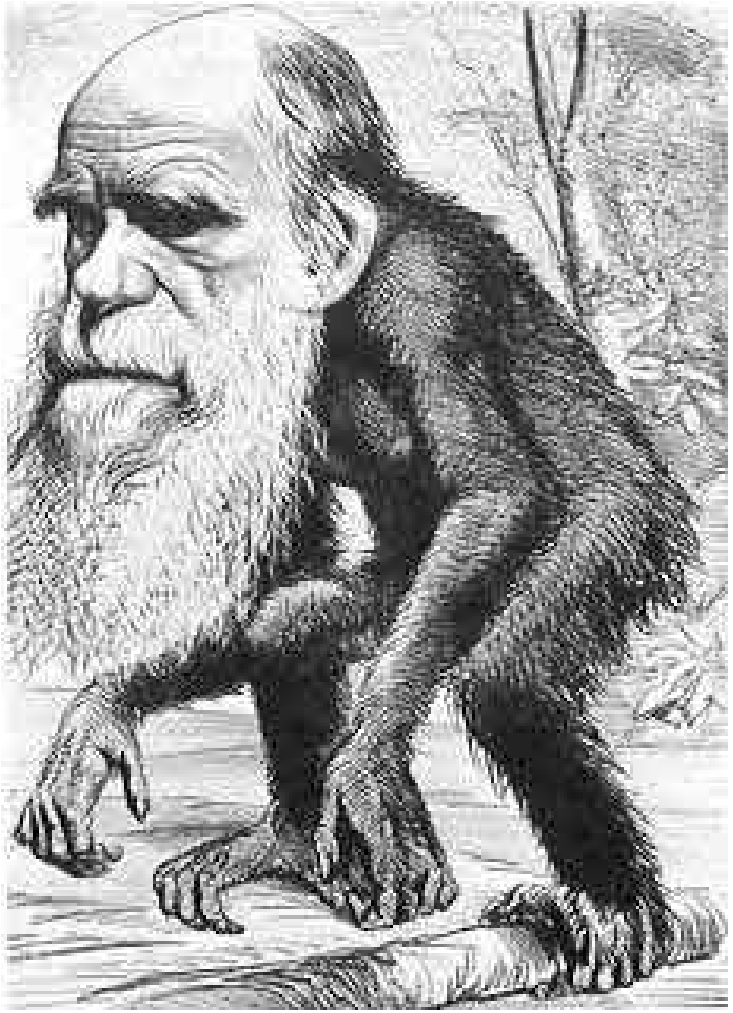
"more questions than answers"

"utter bollocks"

"biggest intellectual fraud since Chomsky"

"this isn't history, it's history put in nested boxes!"

What is **evolution**?



Variation

Heritability

Differential survival

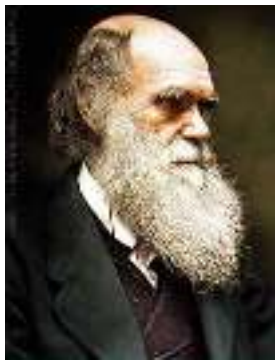
When and where did 🐨👧💬 **originate**?

What **differences** are there between 🐨👧💬's?

How are 🐨👧💬 **related** to other 🐨👧💬's?

What **processes** shaped 🐨👧💬?

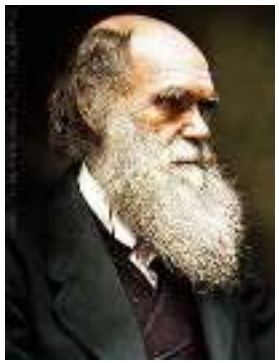
Can we infer what 🐨👧💬 were in the **past**?



Darwin (1871)

"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”



Darwin (1871)

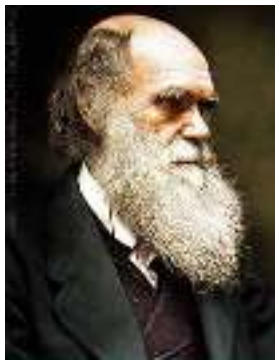
"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”

Schleicher (1863)

Darwinism Tested by the Science of Language

“same process has long been generally assumed for linguistic organisms”



Darwin (1871)

"Languages, like organic beings, can be classed in groups under groups…

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are **curiously parallel**”

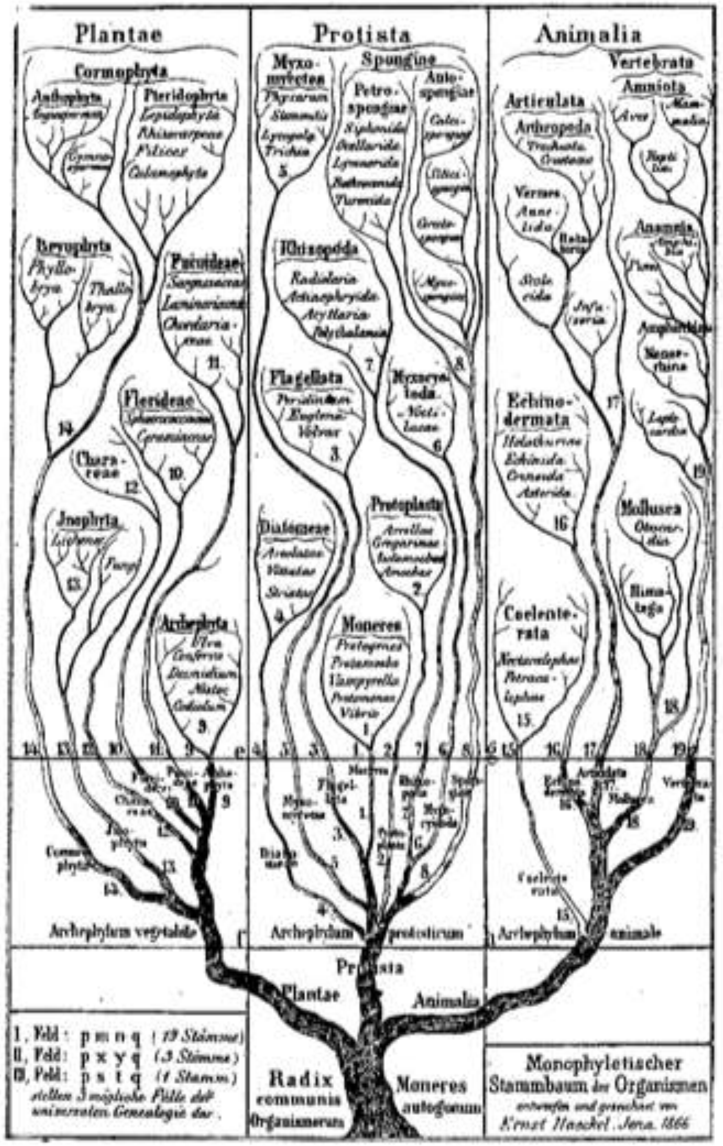
Schleicher (1863)

Darwinism Tested by the Science of Language

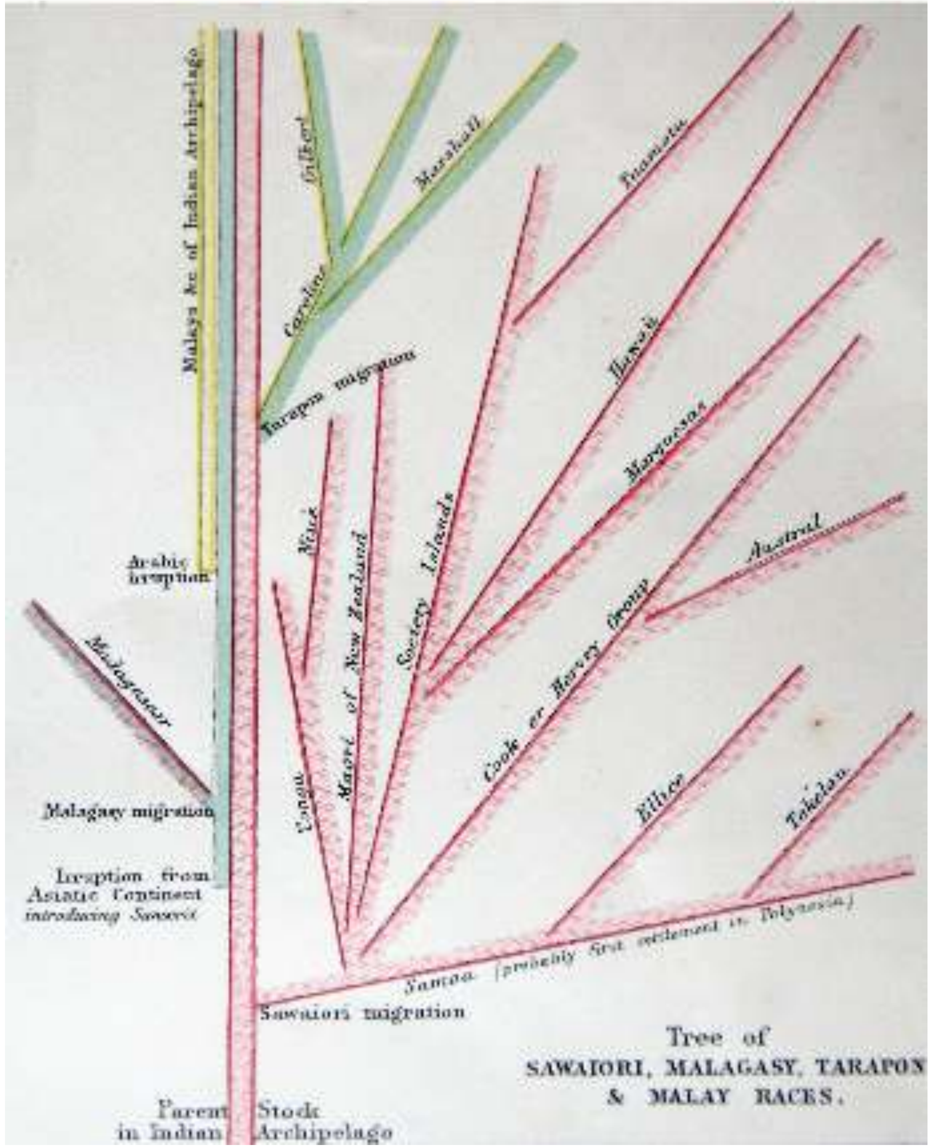
“same process has long been generally assumed for linguistic organisms”

Brugmann (1884)

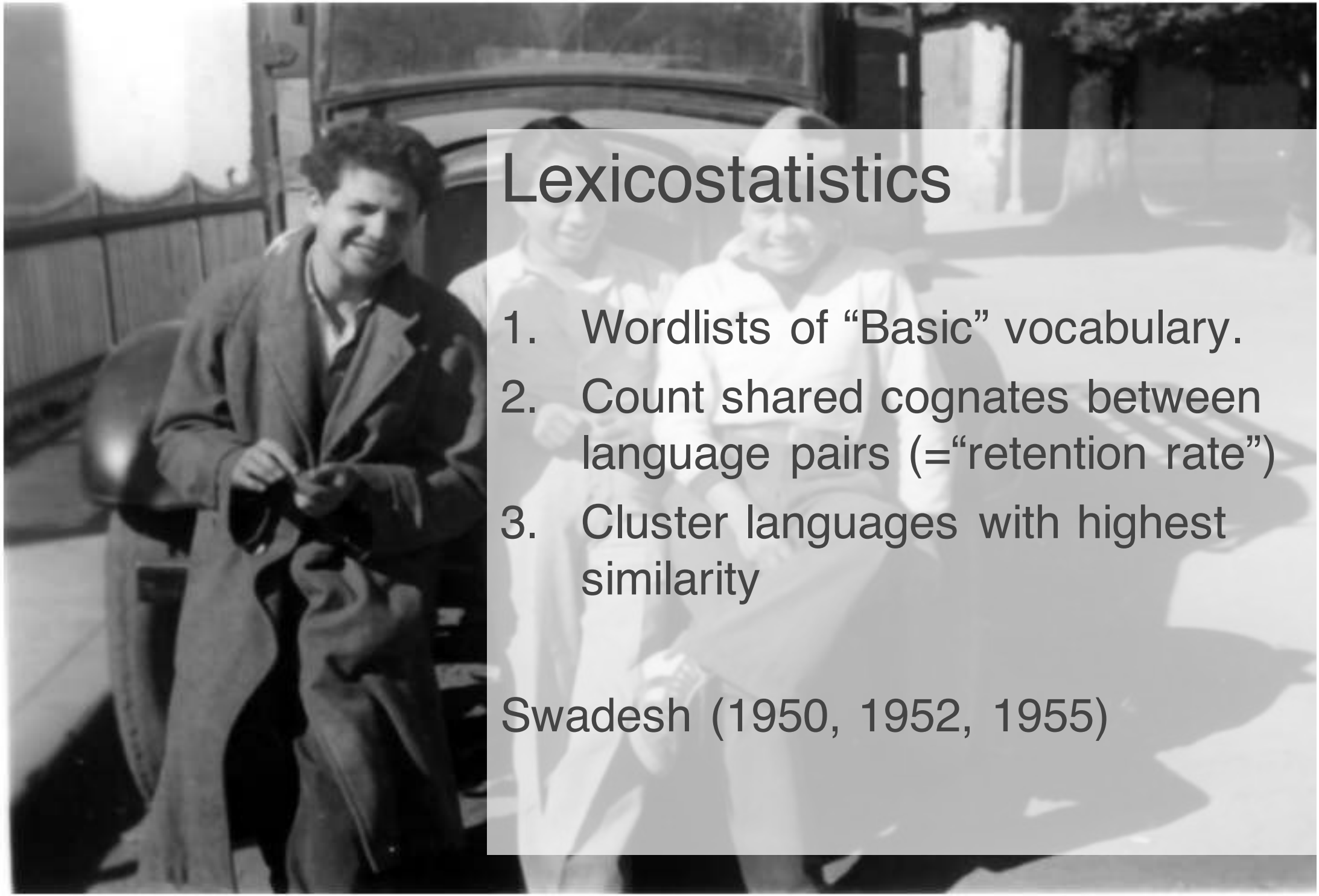
Importance of using “shared innovations” to identify clades and not “shared retentions”



Plantae
Protista
Animalia
Monophyletischer Stammbaum der Organismen
entworfen und gezeichnet von Ernst Haeckel. Jena. 1866
Radix communis Organismorum
Moneres autogonum



Tree of SAWAIORI, MALAGASY, TARAPON & MALAY RACES.
Parent Stock in Indian Archipelago
Sawaiori migration
Samoa (probably first settlement in Polynesia)
Malagasy migration
Madagascar
Irruption from Asiatic Continent introducing Sanscrit
Arabic Irruption
Tarapon migration
Malays &c. of Indian Archipelago
Caroline
Gilbert
Marshall
Paumotu
Hawaii
Marquesas
Austral
Cook or Hervey Group
Society Islands
Maori of New Zealand
Niue
Tonga
Ellice
Tokelau



Lexicostatistics

1. Wordlists of “Basic” vocabulary.
2. Count shared cognates between language pairs (=“retention rate”)
3. Cluster languages with highest similarity

Swadesh (1950, 1952, 1955)

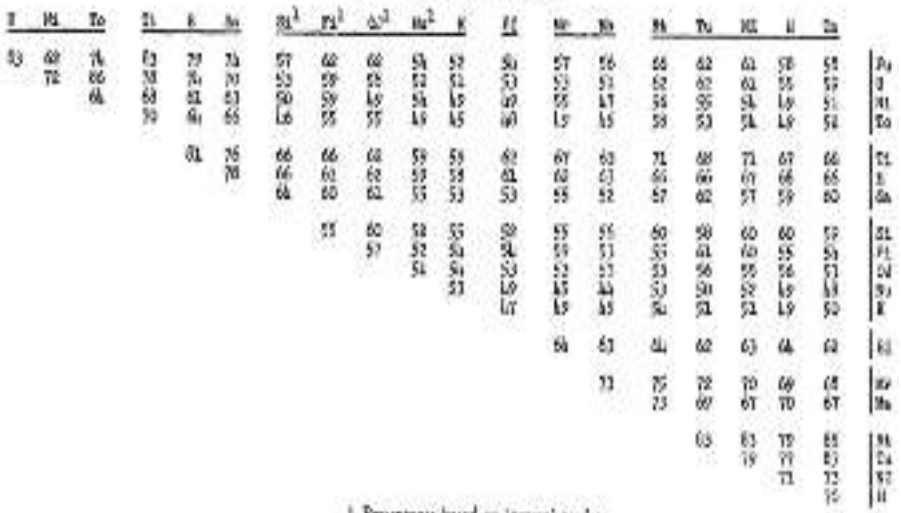
	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

Identified by Systematic Sound Correspondences
- e.g. Maori “t” = “k” in Hawaiian.

Elbert 1953

TABLE 2

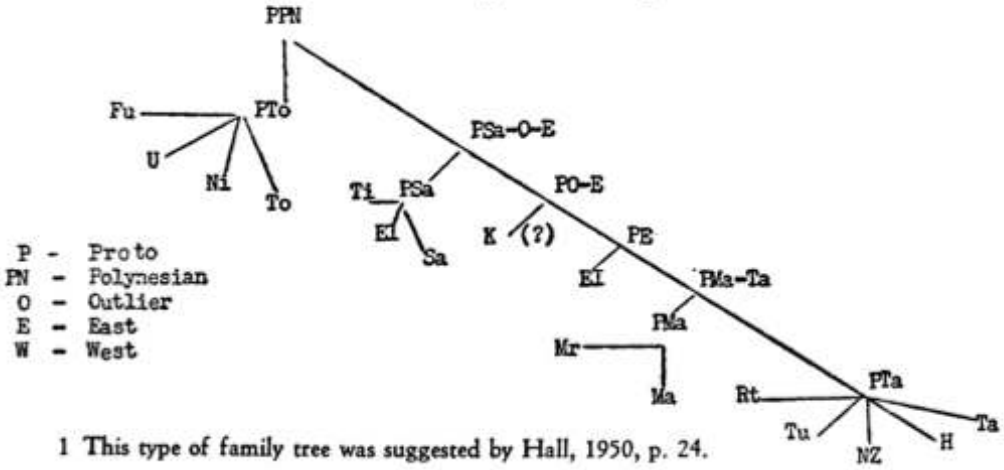
Polynesian cognate percentages



1 Percentages based on incomplete data

TABLE 4

A tentative family tree for Polynesia[1]



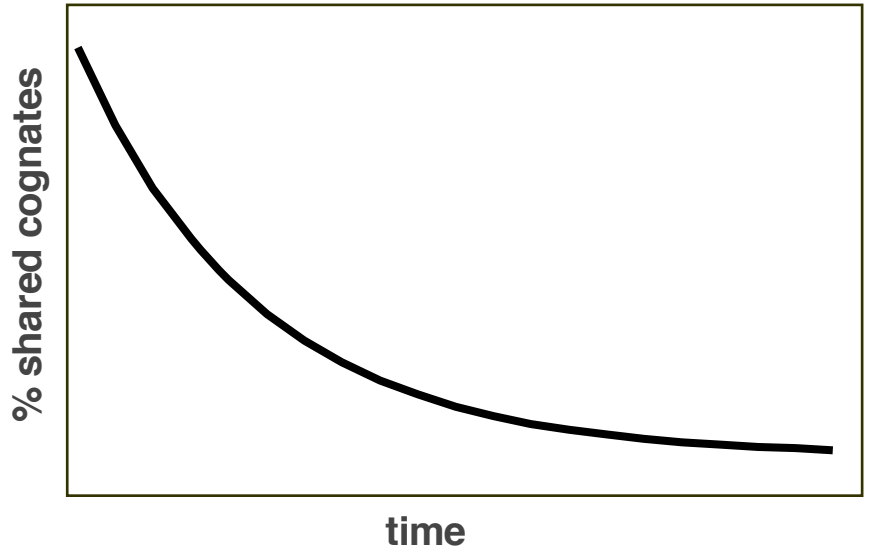
1 This type of family tree was suggested by Hall, 1950, p. 24.

Glottochronology

Loss of cognates happens at a constant rate
(=radioactive decay)

19% loss per 1000 years (Lees 1953)

$$time = \frac{\log(\%sharedcognates)}{2\log(retentionrate)}$$



The Rise of Lexicostatistics…

IN THE LAST DECADE glottochronology has excited international interest and acquired a literature of its own. To the anthropologist it promises a measure of time depth for language families without documented history, and yet another linguistic example of regularity in cultural phenomena.

Hymes (1960): "Lexicostatistics so far"

"... a significant work—one which may conceivably be as revolutionary for Oceanic linguistics and culture history as was the work of Greenberg (1949–54) for the interpretation of African languages and cultures"

Murdock (1964) p.117

…and the fall of Lexicostatistics

Major Criticism: Universality of Rates

Old Norse & Icelandic?

- Glottochronology: 200 years.
- Reality: 1000 years

Bergsland & Vogt 1962: “Our findings clearly disprove the basic assumption of glottochronology 'that fundamental vocabulary changes at a constant rate’ ”

Jungner, Hugo; Elisabeth Svärdström (1940-1971). Sveriges runinskrifter: V. Västergötlands runinskrifter. Stockholm: Kungl. Vitterhets Historie och Antikvitets Akademien. ISSN 0562-8016. p. 260

Fallout.

"a tradition of hostility towards probabilistic modelling in historical linguistics" (Sankoff '73)

"In summary, glottochronology is not accurate; all its basic assumptions have been severely criticized. It should not be accepted, it should be rejected" (Campbell '04)

"Linguists don't do dates" (McMahon & McMahon '03)

U.P.G.M.A

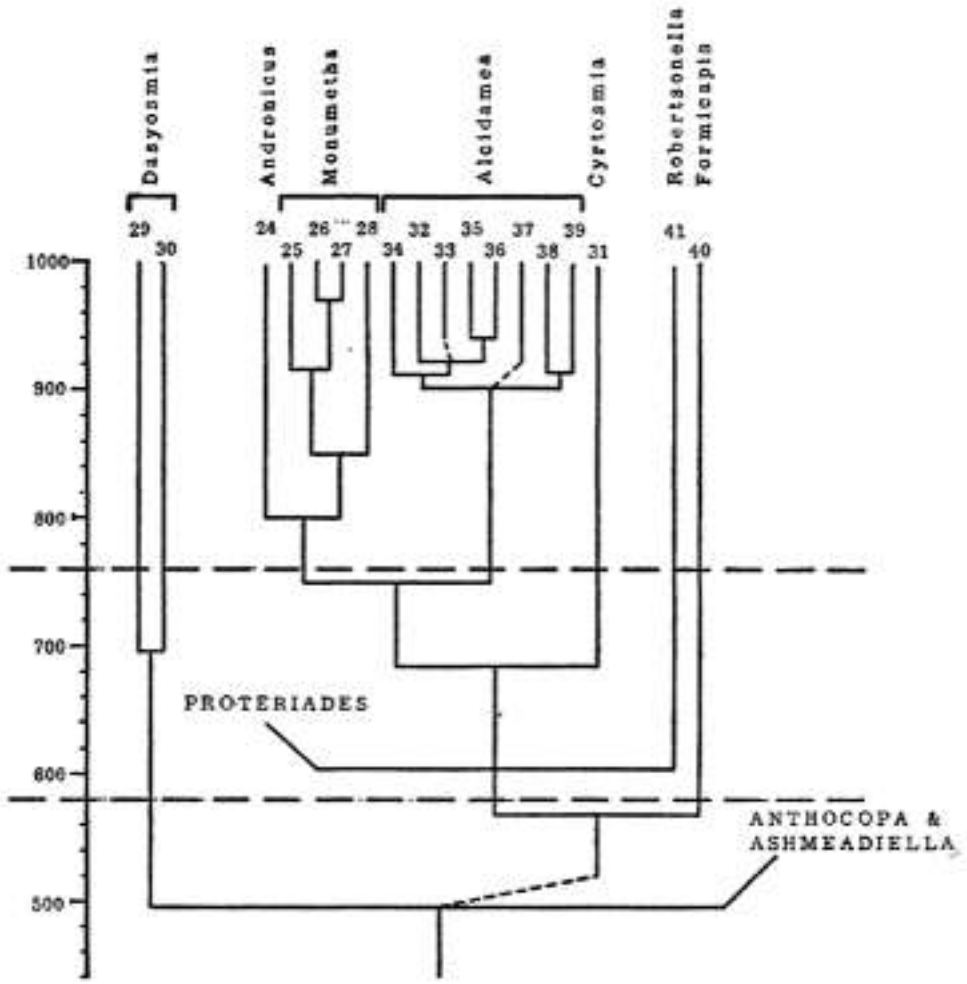


FIG. 6. Diagram of relationships for the genus *Hoplitis* obtained by the weighted variable group method.

A QUANTITATIVE APPROACH TO A PROBLEM IN CLASSIFICATION [1]

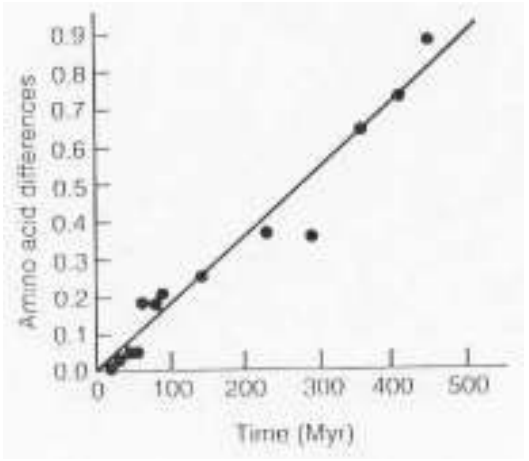
CHARLES D. MICHENER AND ROBERT R. SOKAL [2]

Department of Entomology, University of Kansas, Lawrence

Molecular Clock

Zuckerkandl & Pauling 1962

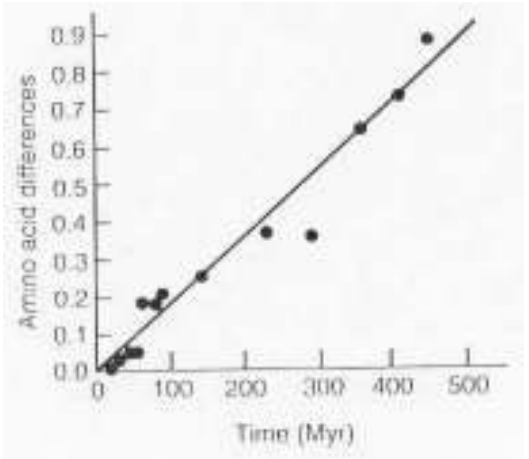
Number of AA differences were proportional to species divergence times.



Molecular Clock

Zuckerkandl & Pauling 1962

Number of AA differences were proportional to species divergence times.



Kimura 1968

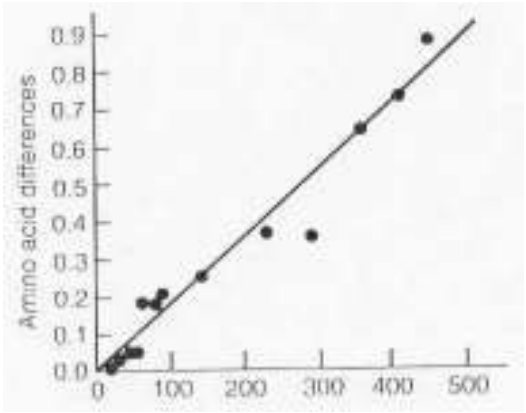
The average time taken for one base pair replacement within a genome is therefore

$$28 \times 10^6 \text{ yr} \div \left(\frac{4 \times 10^9}{300}\right) \div 1{\cdot}2 \doteqdot 1{\cdot}8 \text{ yr}$$

Molecular Clock

Zuckerkandl & Pauling 1962

Number of AA differences were proportional to species divergence times.



Kimura 1968

The average time taken for one base pair replacement within a genome is therefore

$$28 \times 10^6 \text{ yr} \div \left(\frac{4 \times 10^9}{300}\right) \div 1{\cdot}2 \doteqdot 1{\cdot}8 \text{ yr}$$

Sarich & Wilson 1967

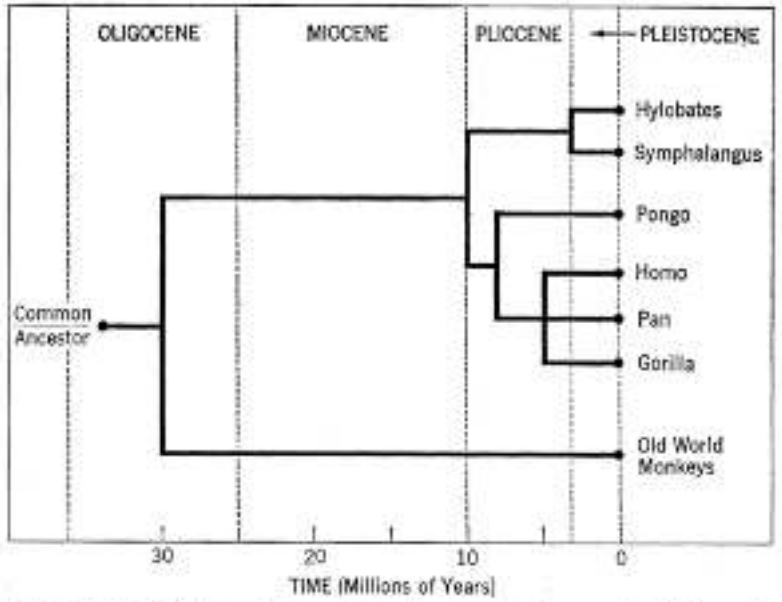


Fig. 1. Times of divergence between the various hominoids, as estimated from immunological data. The time of divergence of hominoids and Old World monkeys is assumed to be 30 million years.

Problems?

Kirsch 1969

SEROLOGICAL DATA AND PHYLOGENETIC INFERENCE:
THE PROBLEM OF RATES OF CHANGE

JOHN A. W. KIRSCH

Felsenstein 1978

CASES IN WHICH PARSIMONY OR COMPATIBILITY
METHODS WILL BE POSITIVELY MISLEADING[1]

JOSEPH FELSENSTEIN

Britten1986

**Rates of DNA Sequence Evolution Differ
Between Taxonomic Groups**

ROY J. BRITTEN

The Cladistics Wars

Solutions.

Cavalli-Sforza & Edwards 1967

Phylogenetic Analysis
Models and Estimation Procedures

L. L. CAVALLI-SFORZA AND A. W. F. EDWARDS*

Yang 1993

Maximum-Likelihood Estimation of Phylogeny from DNA Sequences When Substitution Rates Differ over Sites[1]

Ziheng Yang

Sanderson 1997

A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy

Michael J. Sanderson

Drummond et al. 2006

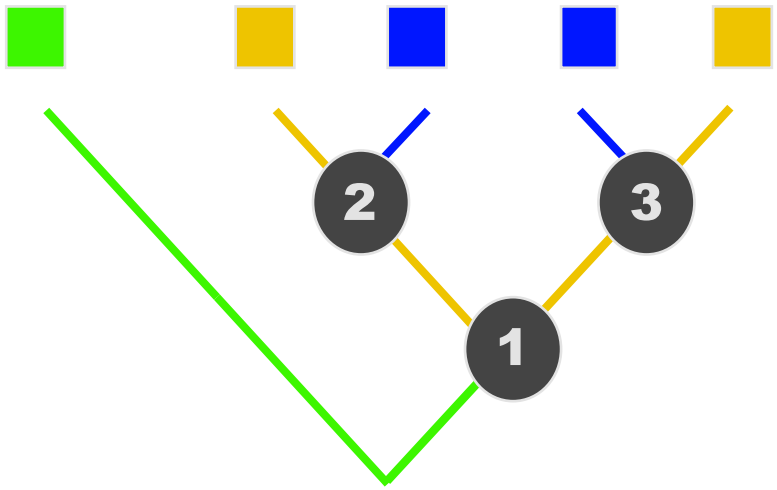
Relaxed Phylogenetics and Dating with Confidence

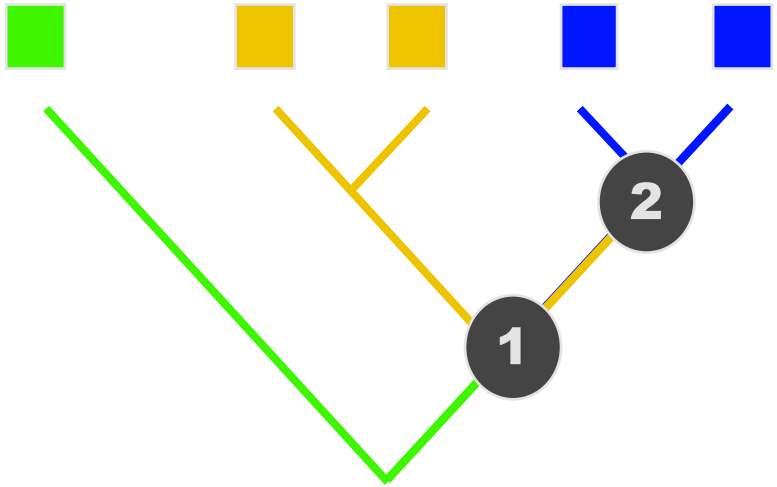
Alexei J. Drummond, Simon Y. W. Ho, Matthew J. Phillips, Andrew Rambaut

How do we **build** trees?

1. Distance Methods (Chiara)
2. Maximum Parsimony.
3. Maximum Likelihood.
4. **Bayesian Phylogenetic Analyses.**

Maximum Parsimony





Unlikely that complex traits should arise more than once

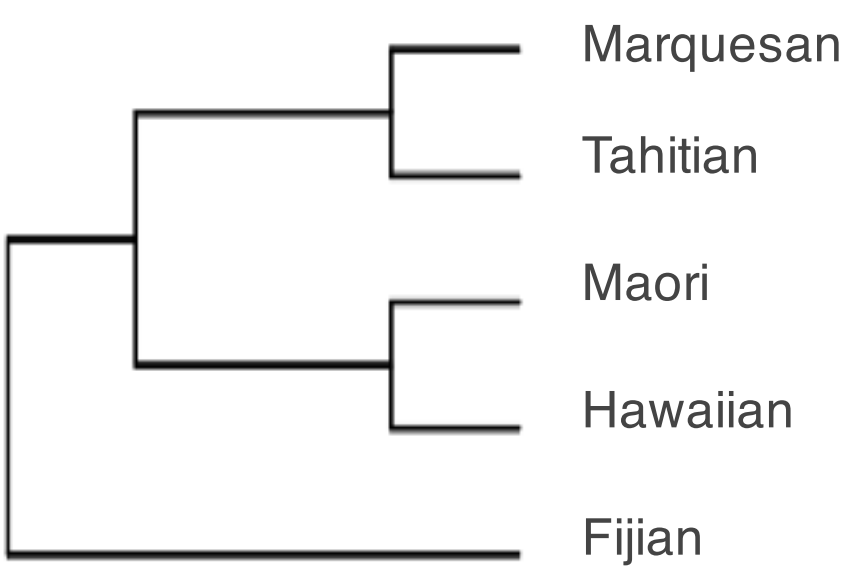
=> Best tree is maximally parsimonious

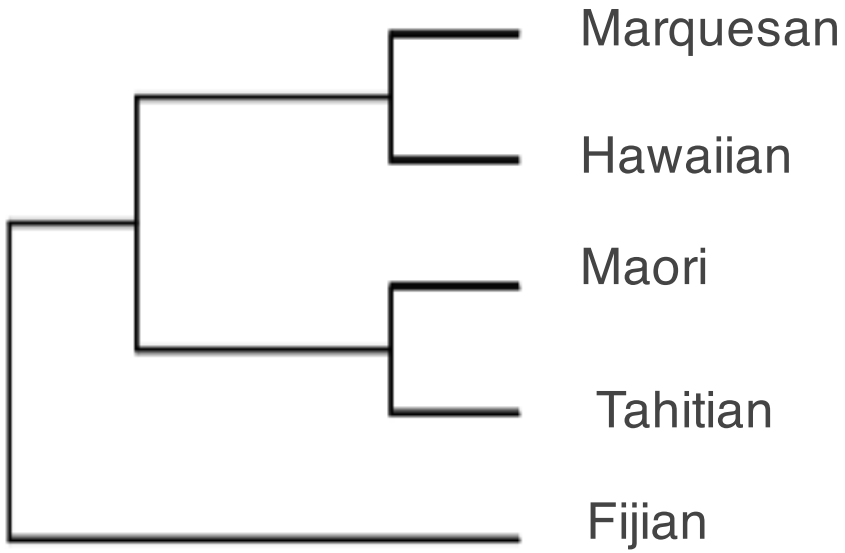
=> Smallest amount of evolution (Fewest number of character state changes)

	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

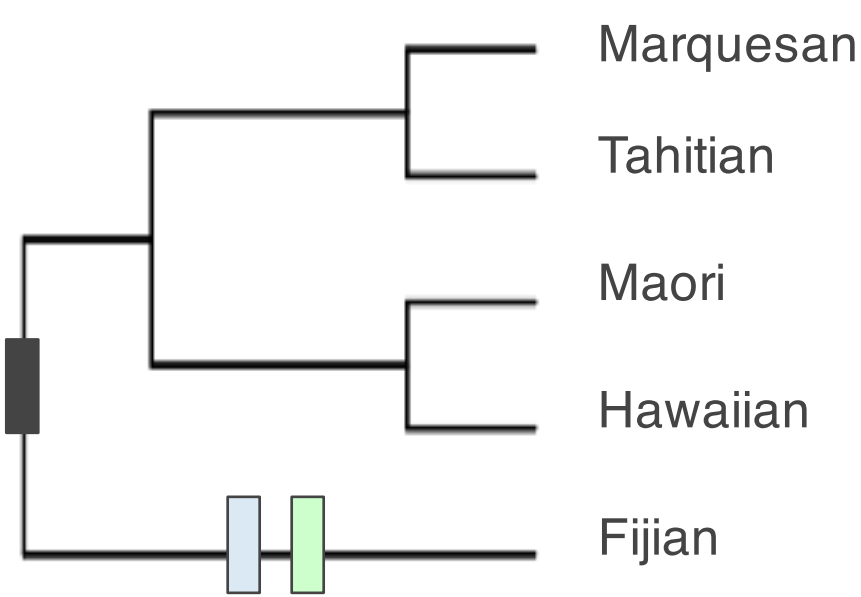
	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

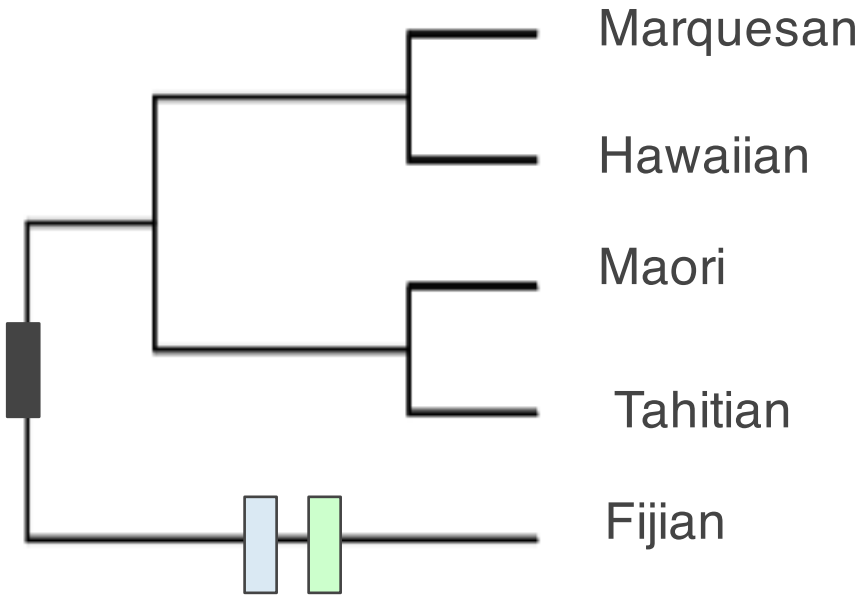




Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

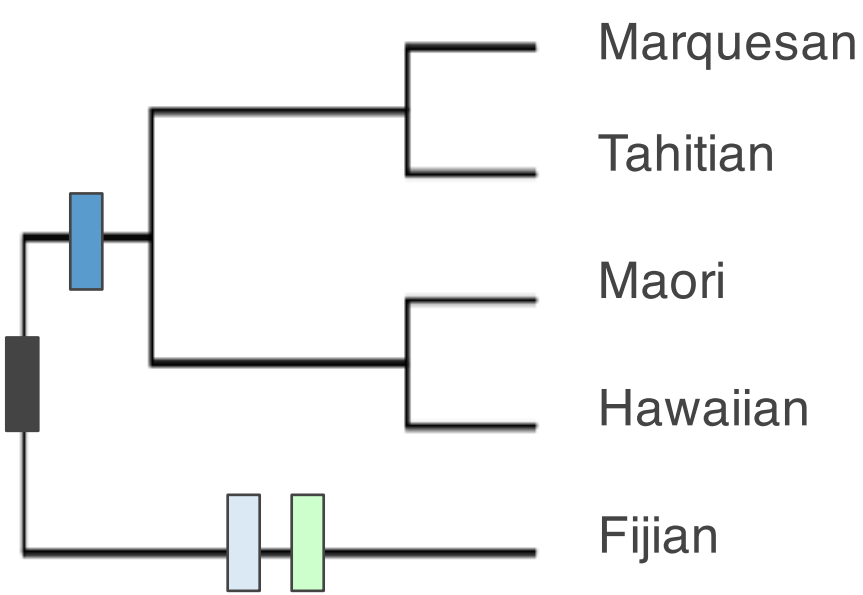


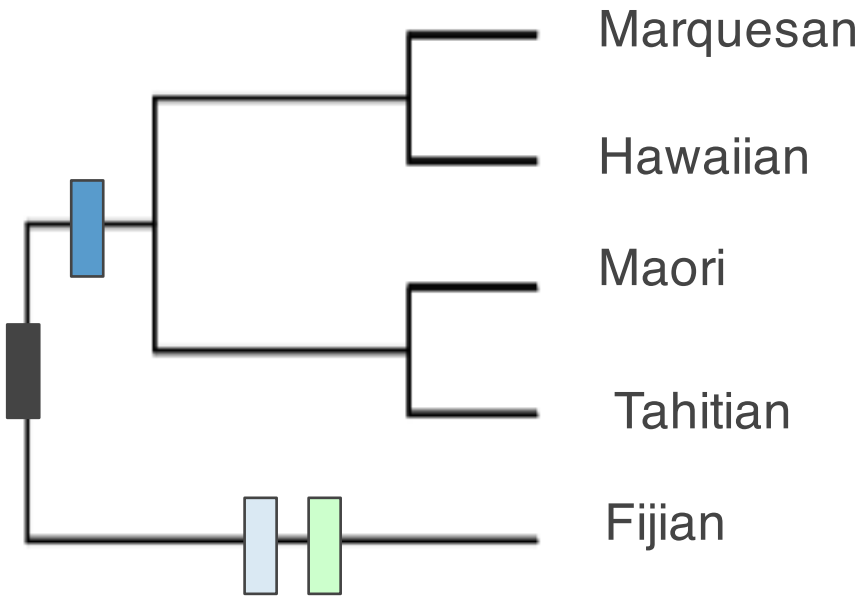
Length=3



Length=3

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

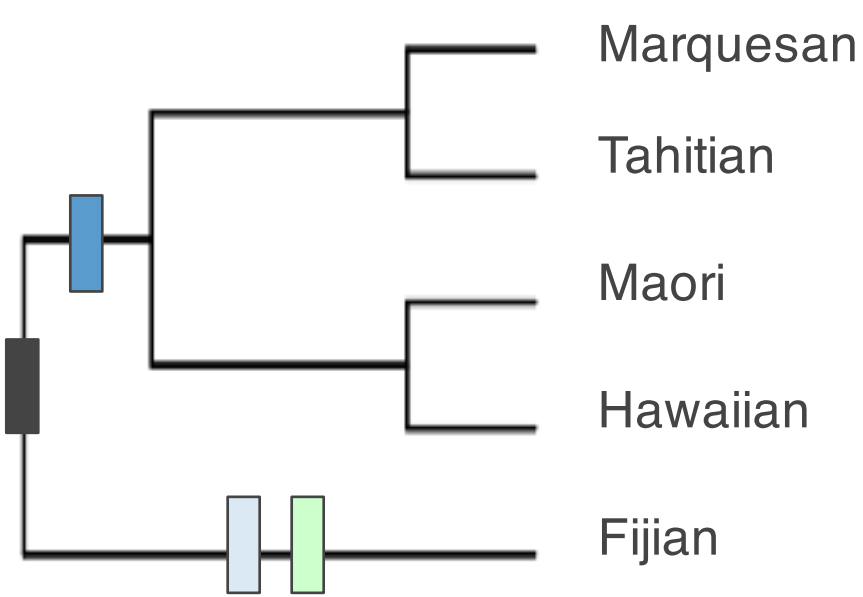


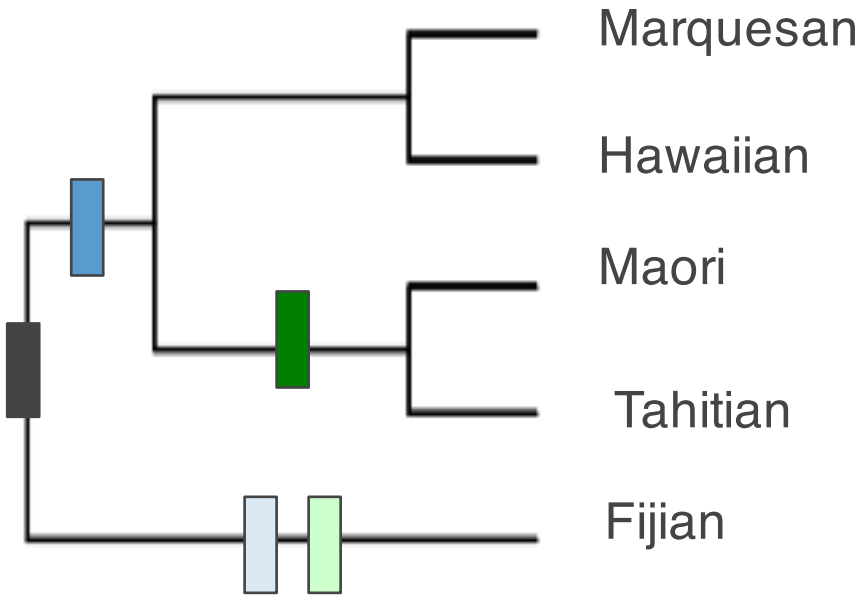


Length=4

Length=4

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

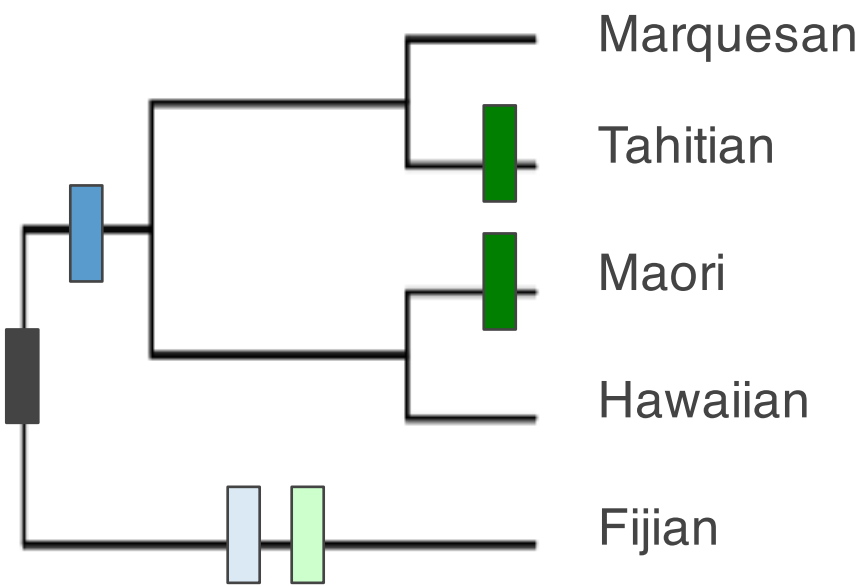




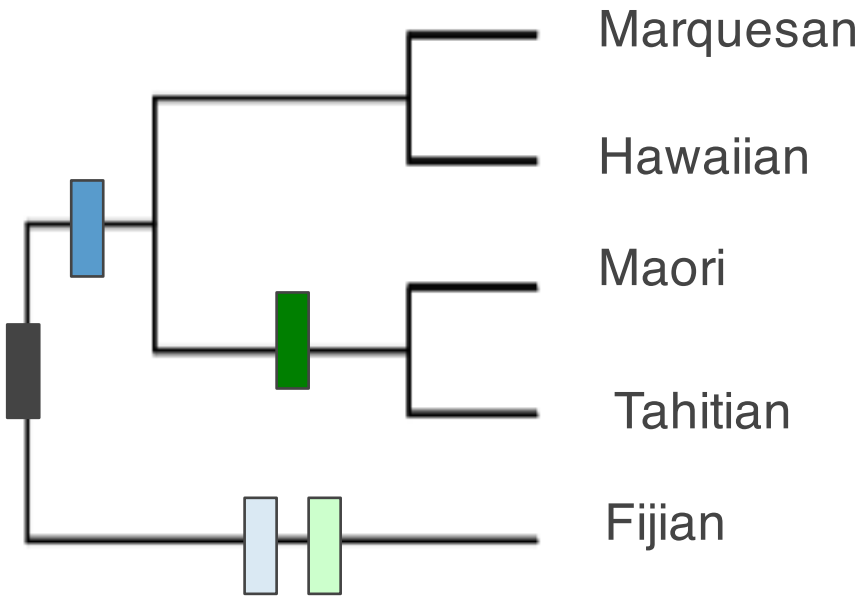
Length=4

Length=5

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

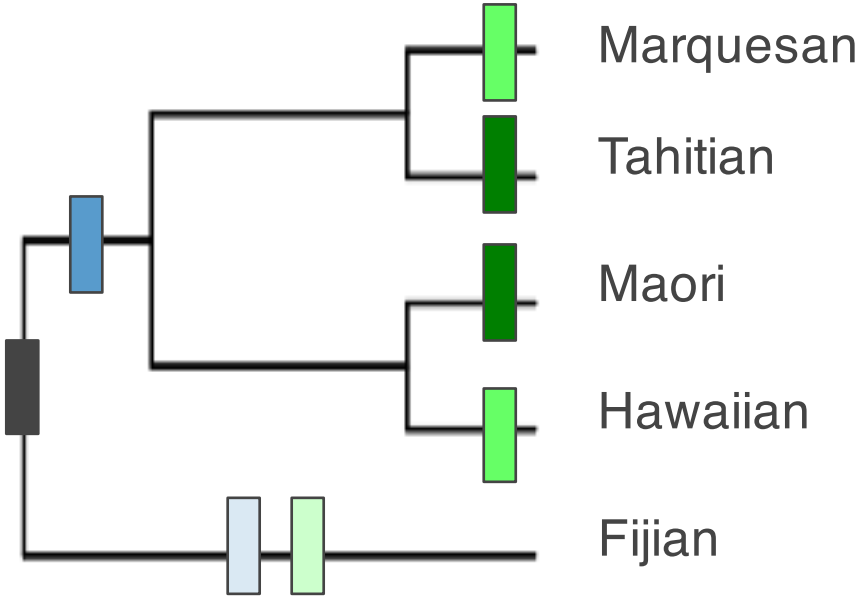


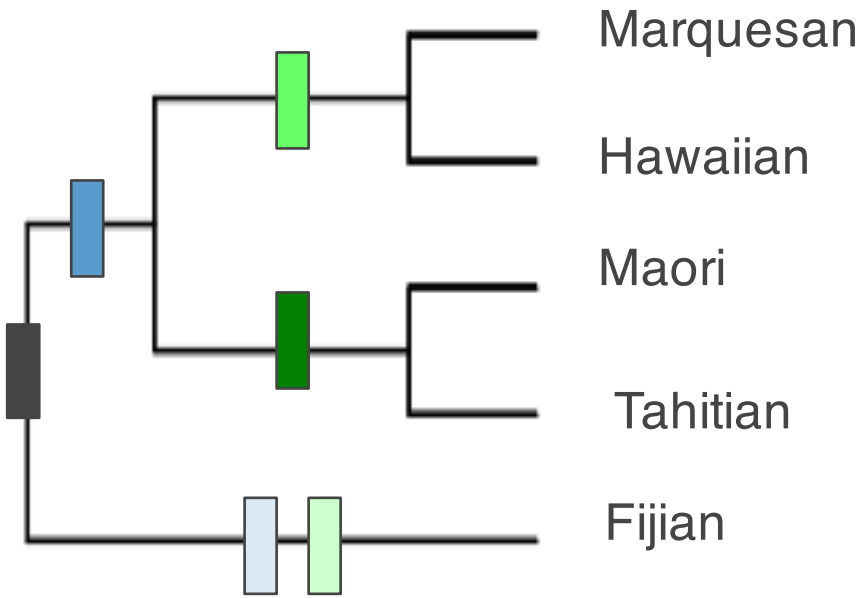
Length=6



Length=5

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**

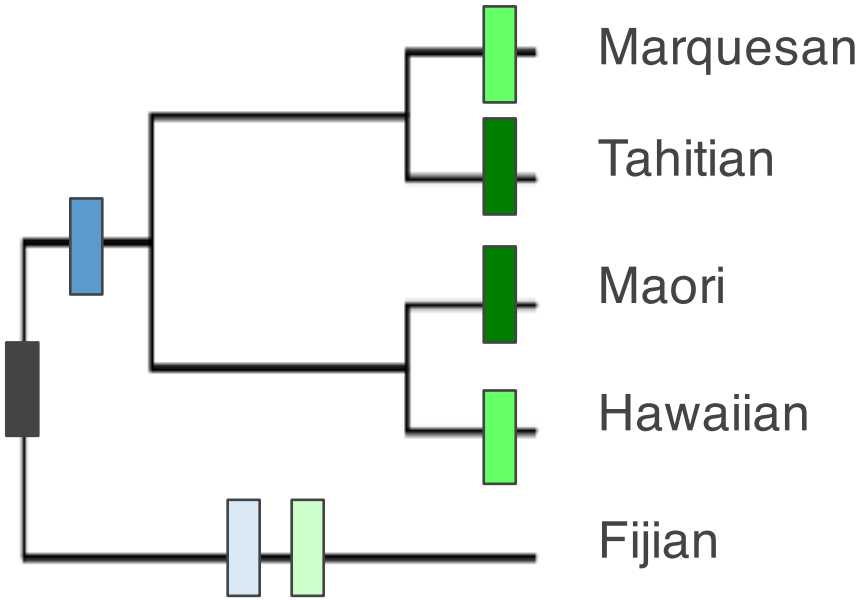


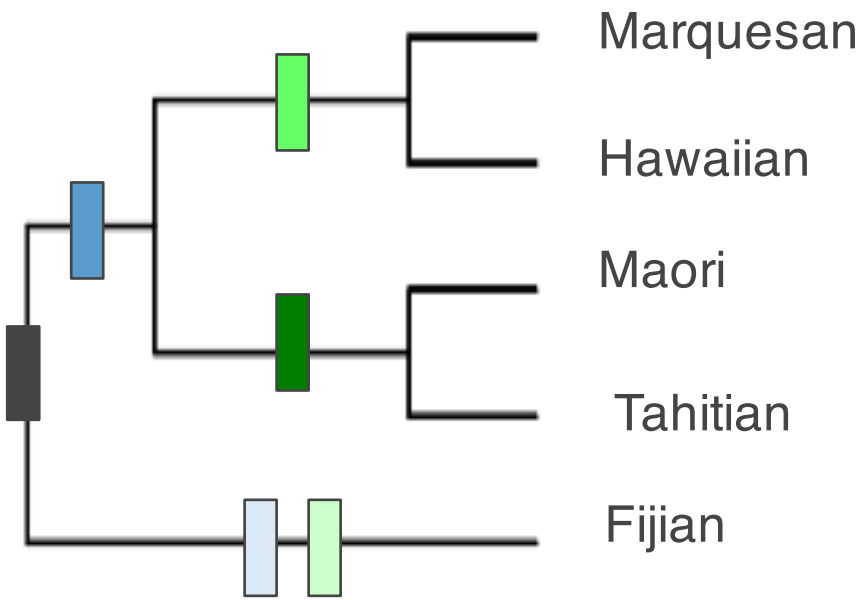


Length=8

Length=6

Fijian	**1**	**1**	**0**	**1**	**0**	**0**
Tahitian	**1**	**0**	**1**	**0**	**1**	**0**
Maori	**1**	**0**	**1**	**0**	**1**	**0**
Hawaiian	**1**	**0**	**1**	**0**	**0**	**1**
Marquesan	**1**	**0**	**1**	**0**	**0**	**1**



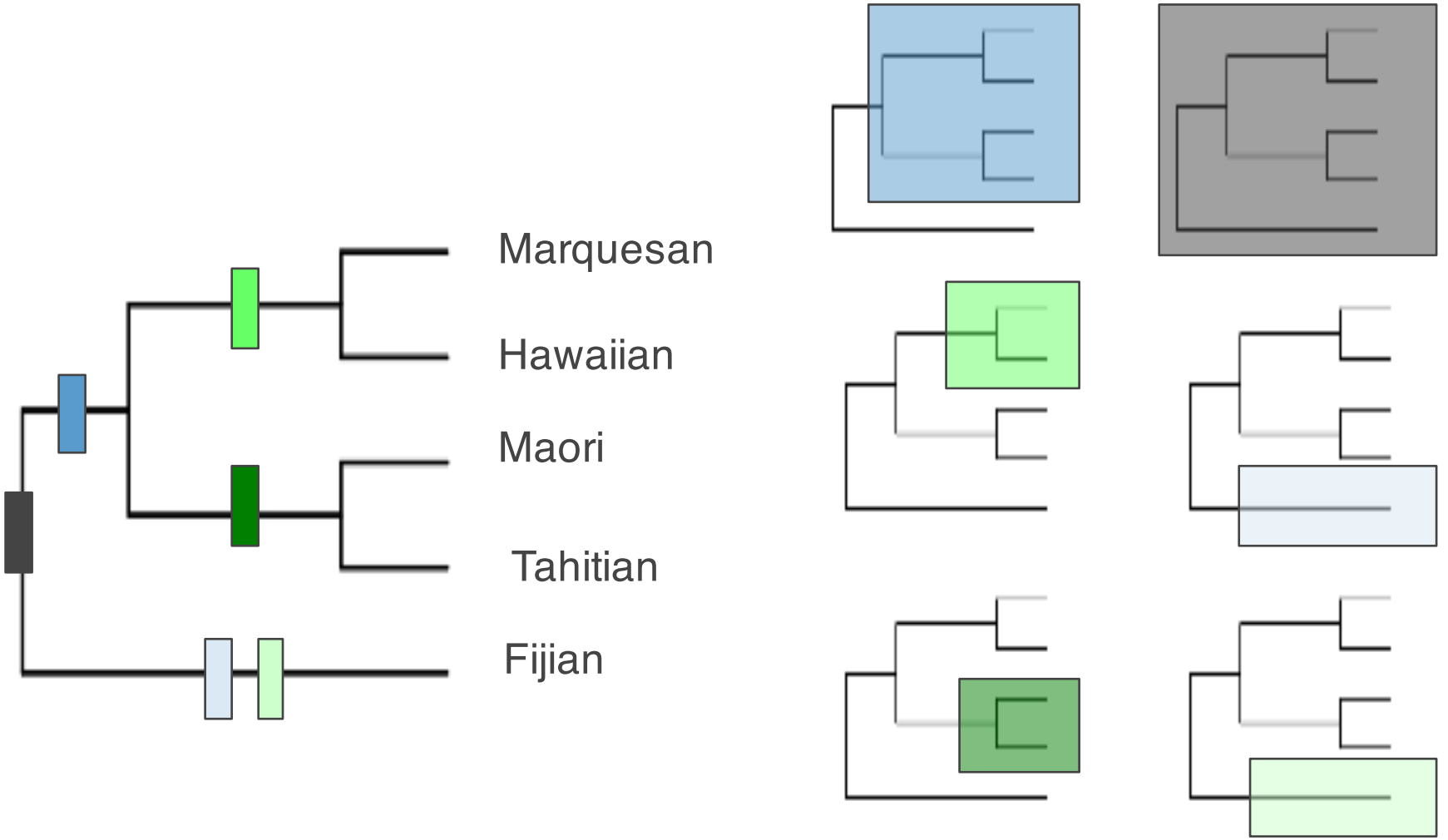


Length=8

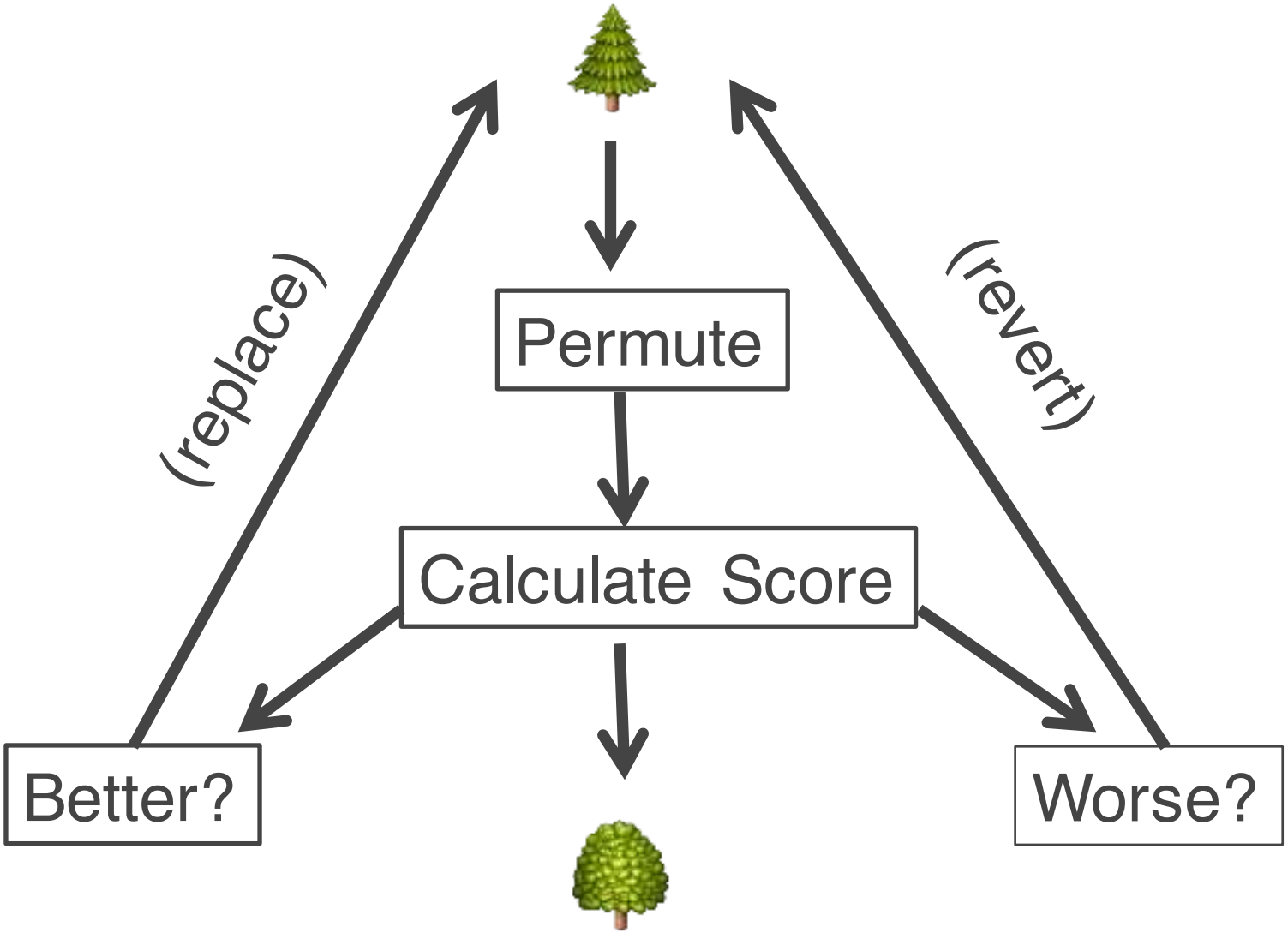
Length=6

Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

Innovations vs. Retentions

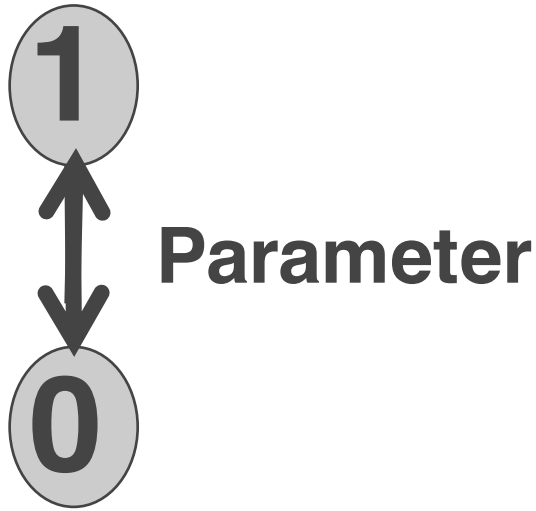


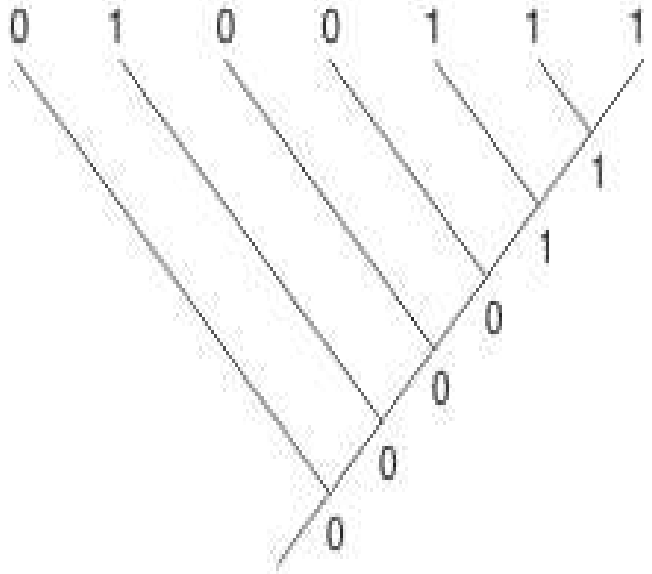
Algorithm


Permute
Calculate Score
Better?
Worse?
(replace)
(revert)
"best" tree

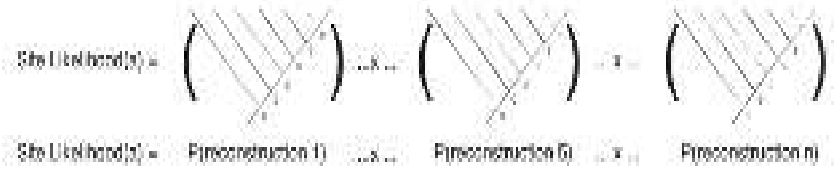
Maximum Likelihood

- Builds on Max. Parsimony
- Stochastic **model** of change
  (=assumptions about how evolution works)
- => Find tree that maximises the likelihood
- **Likelihood** = fit of data to tree under a model.
  - Very small number = log(Lh)
  - Closer to zero = better fit.





$$L(a) = P(0\rightarrow 0|b1) \times P(0\rightarrow 0|b2) \times P(1\rightarrow 1|b3) \times P(1\rightarrow 0|b4) \times P(0\rightarrow 0|b5) \times P(0\rightarrow 0|b6) \times P(0\rightarrow 0|b7) \times P(0\rightarrow 1|b8) \times P(1\rightarrow 1|b9) \times P(1\rightarrow 1|b10) \times P(1\rightarrow 1|b11) \times P(1\rightarrow 1|b12)$$



Lh =
P of being in state 0, and staying state 0 on branch 1.

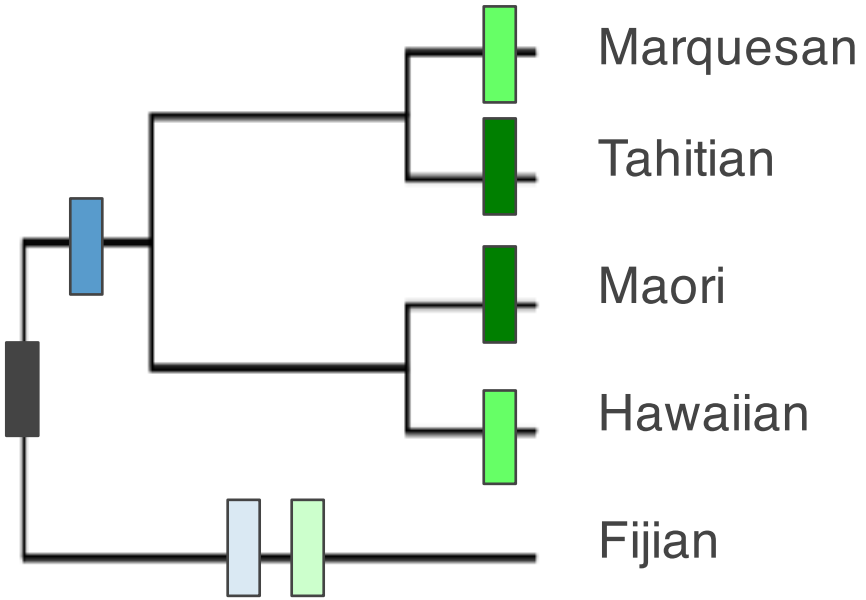
x

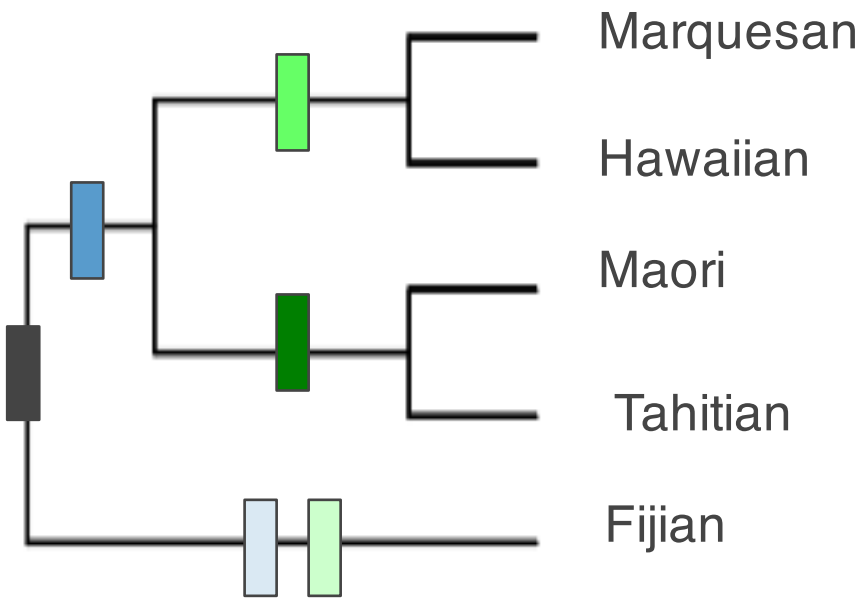
P of being in state 0, and staying state 0 on branch 2.

x

P of being in state 0, and staying state 0 on branch 2.

…. etc



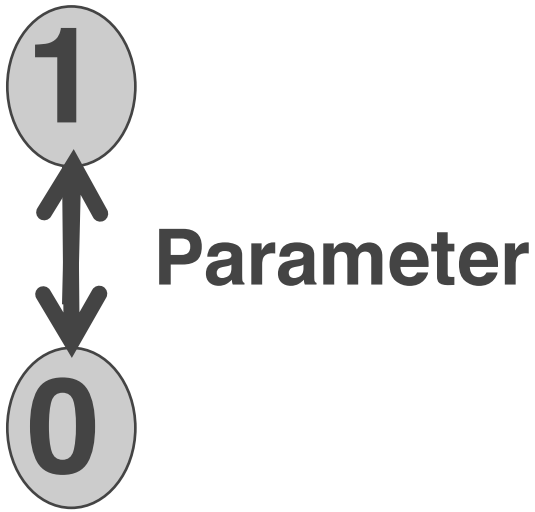


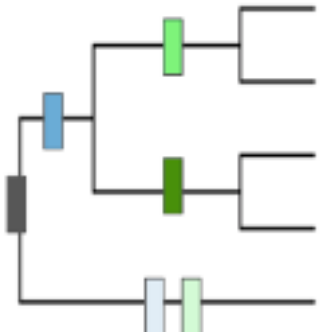
Ln(L)= -14.804

Ln(L)= -12.007

Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

Model Comparison with Lh

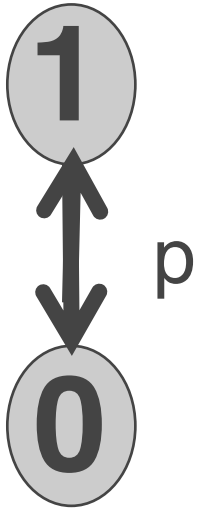


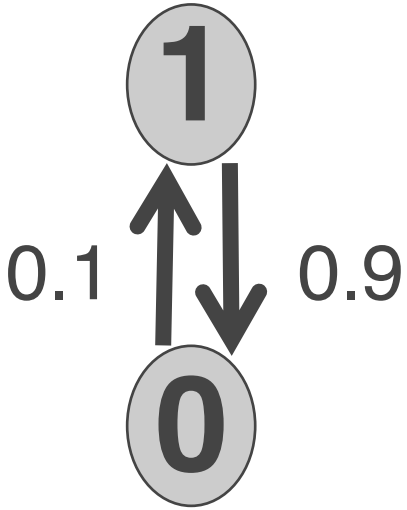


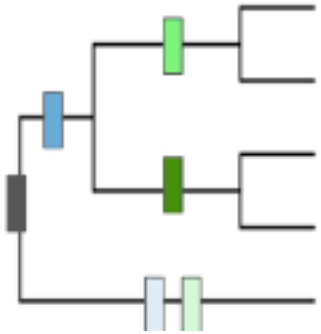
Can we modify our **assumptions** and get a **better explanation of the data?**

Ln(L)= -12.007

Models

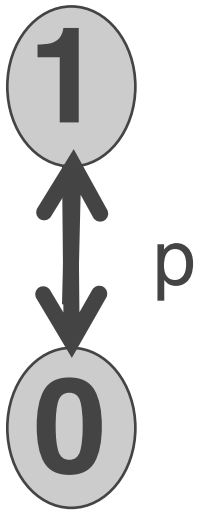


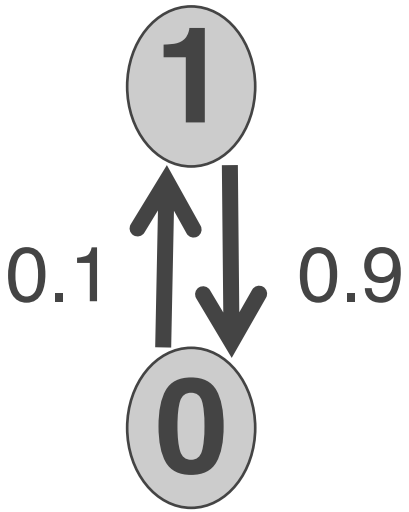


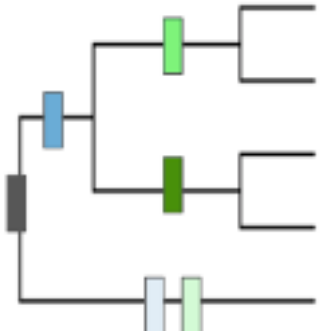


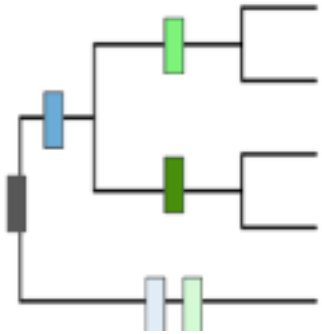
Ln(L)= -12.007

Models





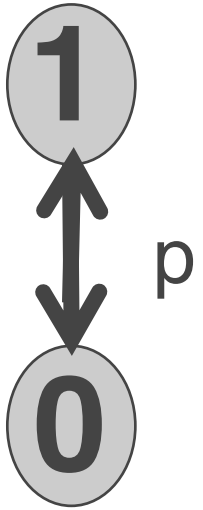


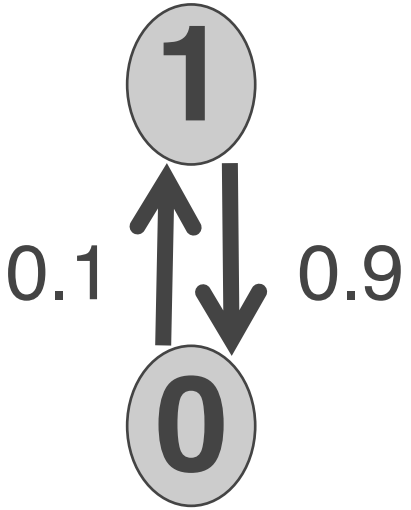


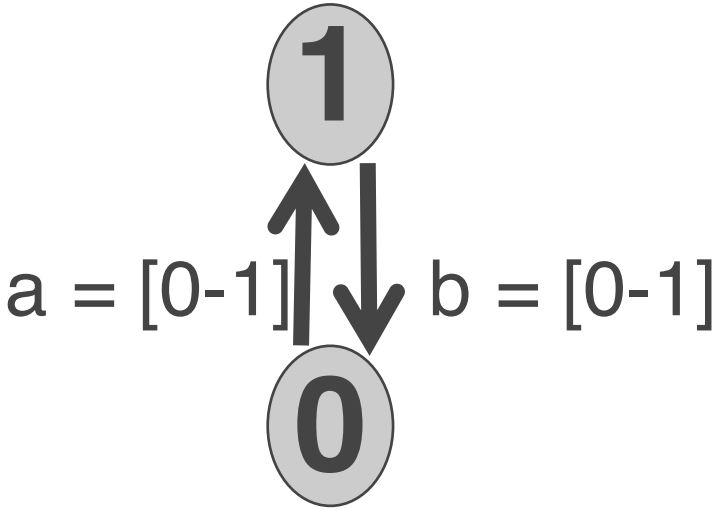
Ln(L)= -12.007

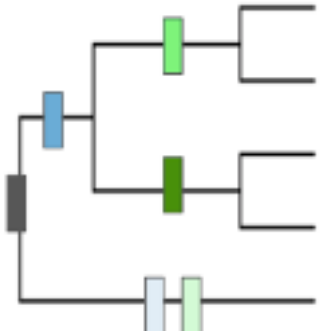
Ln(L)= -11.310

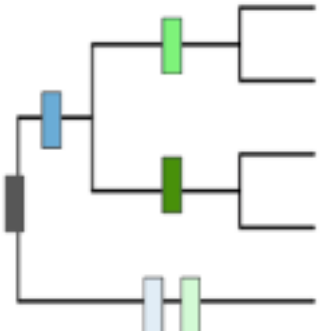
Models







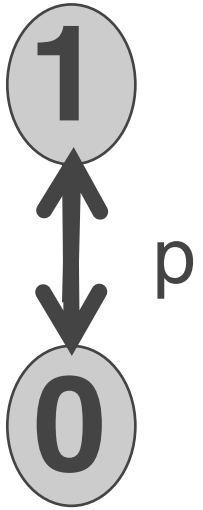


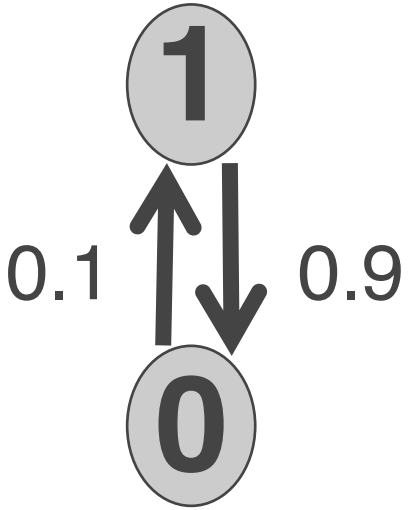


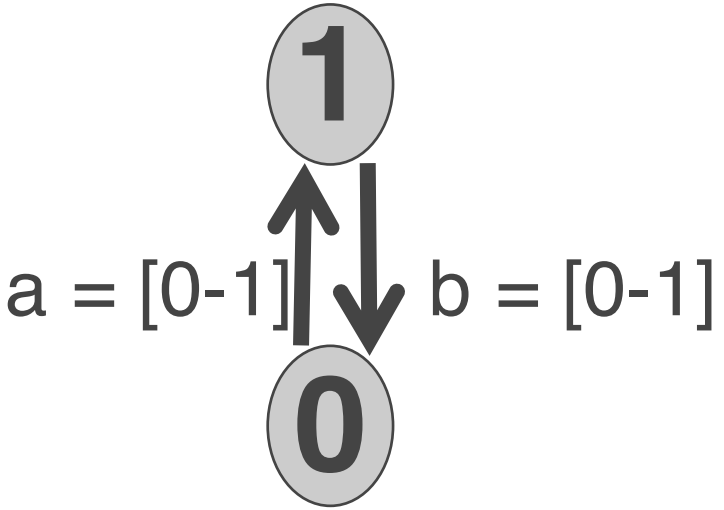
Ln(L)= -12.007

Ln(L)= -11.310

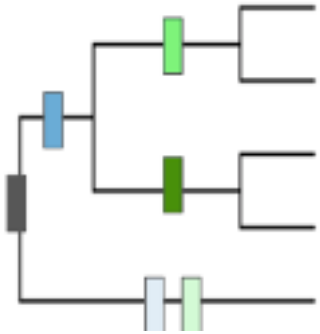
Models

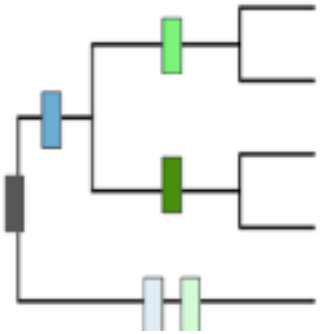






a = 0.92, b = 0.08

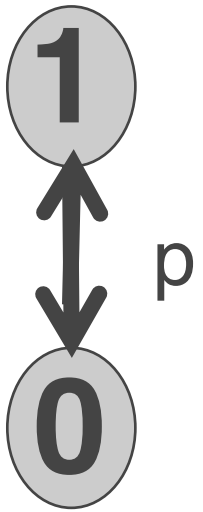


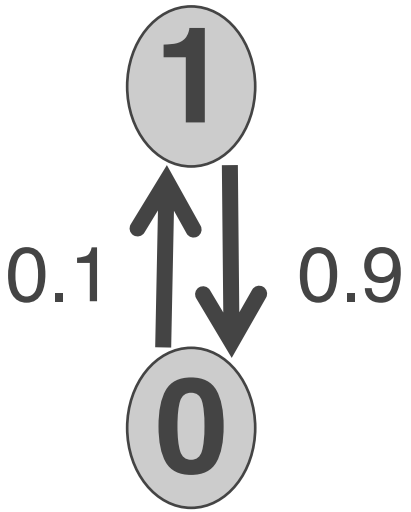


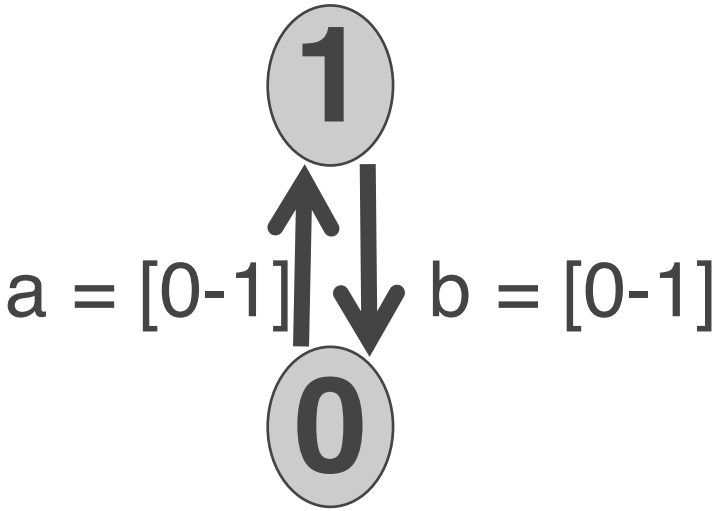
Ln(L)= -12.007

Ln(L)= -11.310

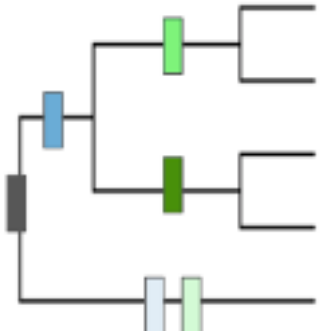
Models

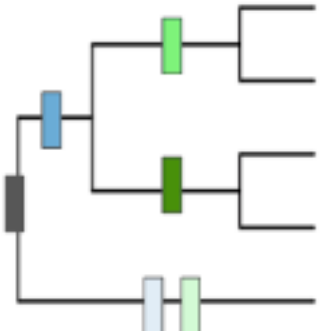


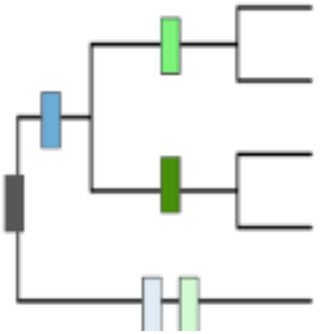




a = 0.92, b = 0.08





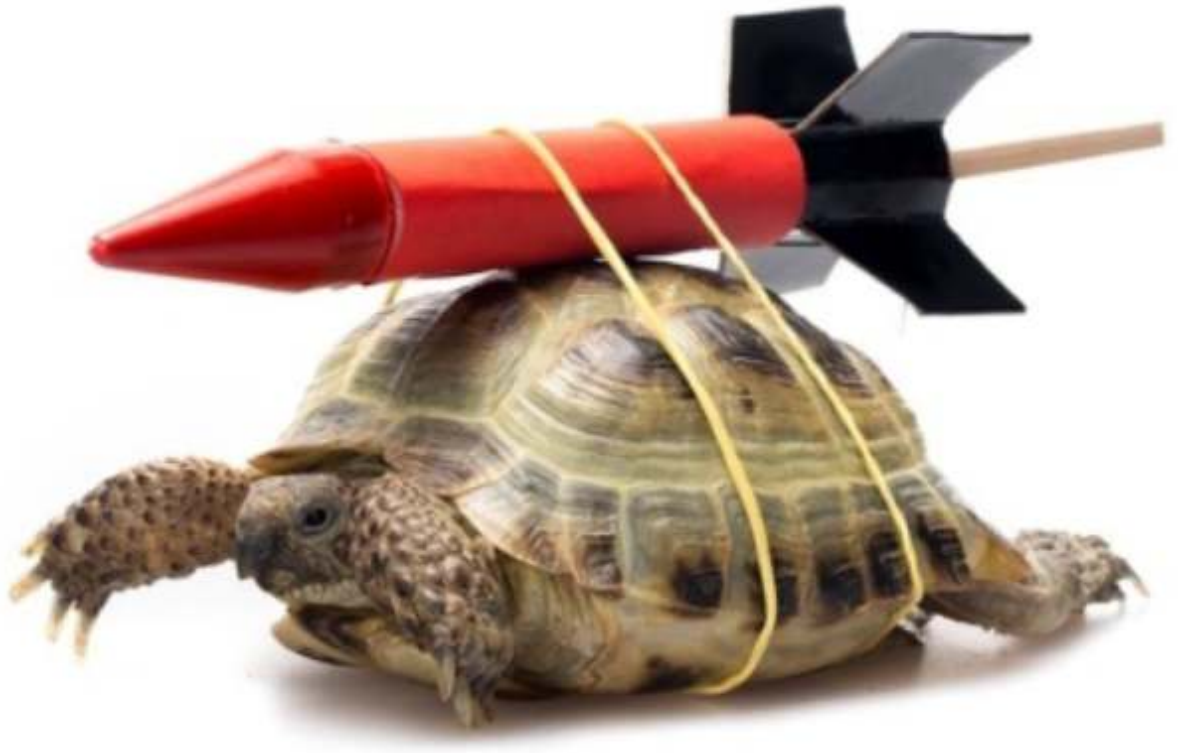


Ln(L)= -12.007

Ln(L)= -11.310

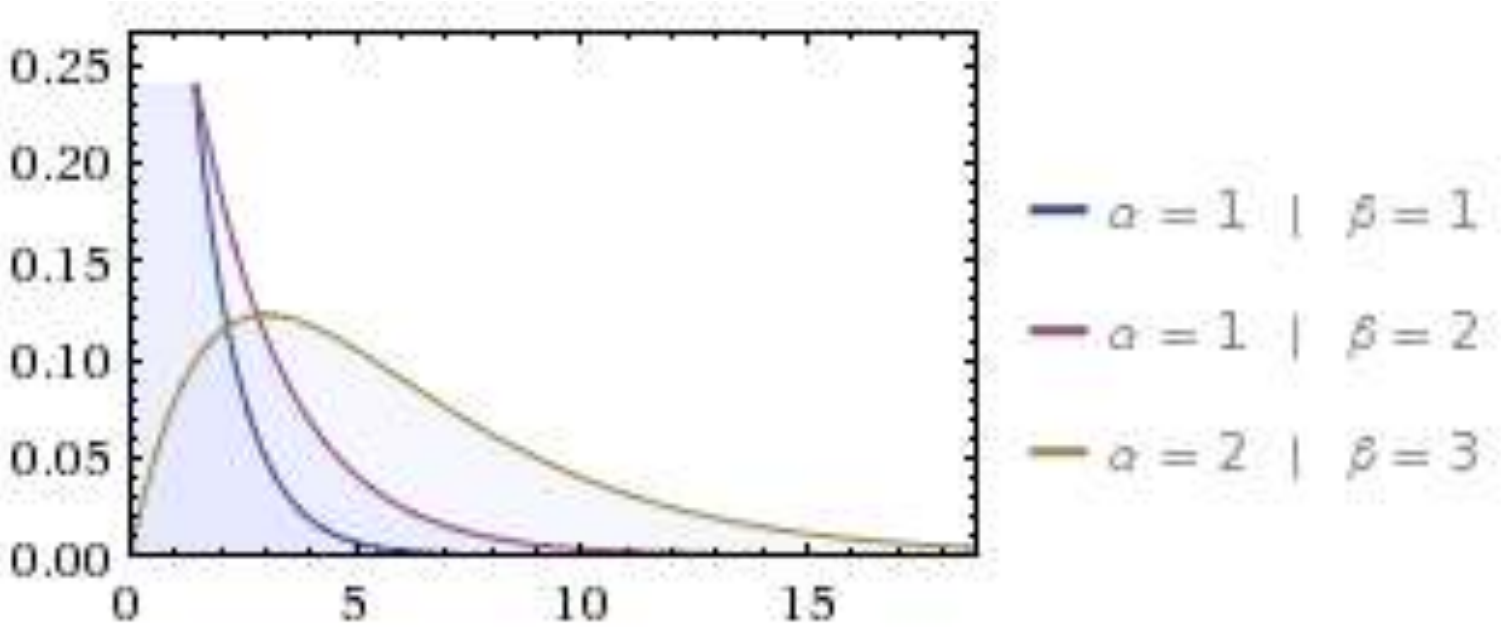
Ln(L)= -9.072

But Rates Vary!



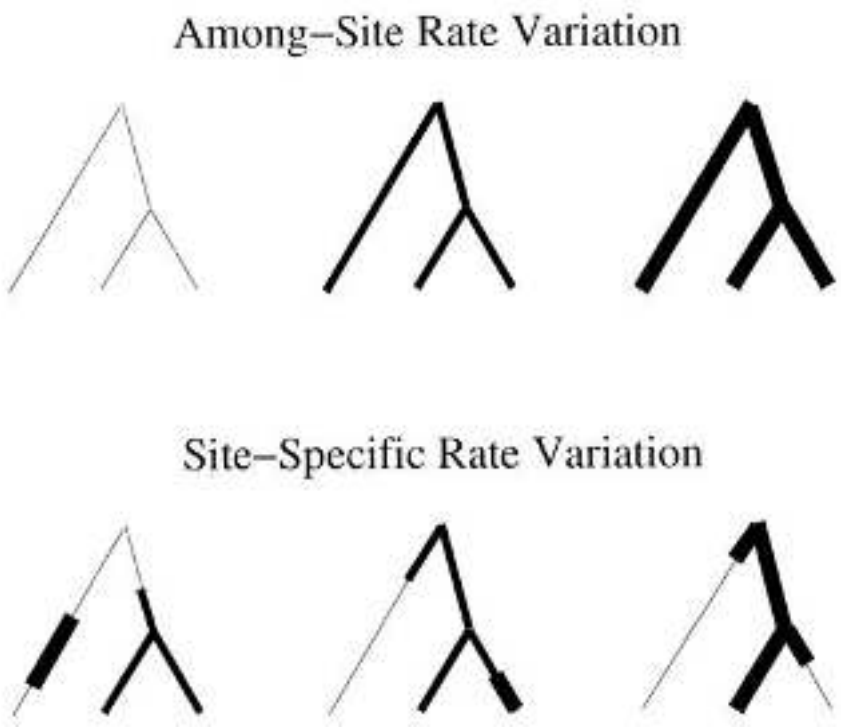
Rate Variation - Gamma

- "Site-Specific Rates": Rates vary across **sites**
- Gamma Distribution: One parameter, α, controls the shape
- Discretized into *n* categories and Characters assigned a category
- Estimate the best value for α using Lh.



Rate Variation - Covarion

- Rates change across **phylogeny**
- Each site is either “on” or “off”
  - In “on” state it evolves according to one set of rates
  - In “off” state it evolves according to another (often 0)
- Analysis estimate when sites are “on” and “off” and switching rates.
- = Each site can evolve at different rates at different times on the tree.



Gautier 2001

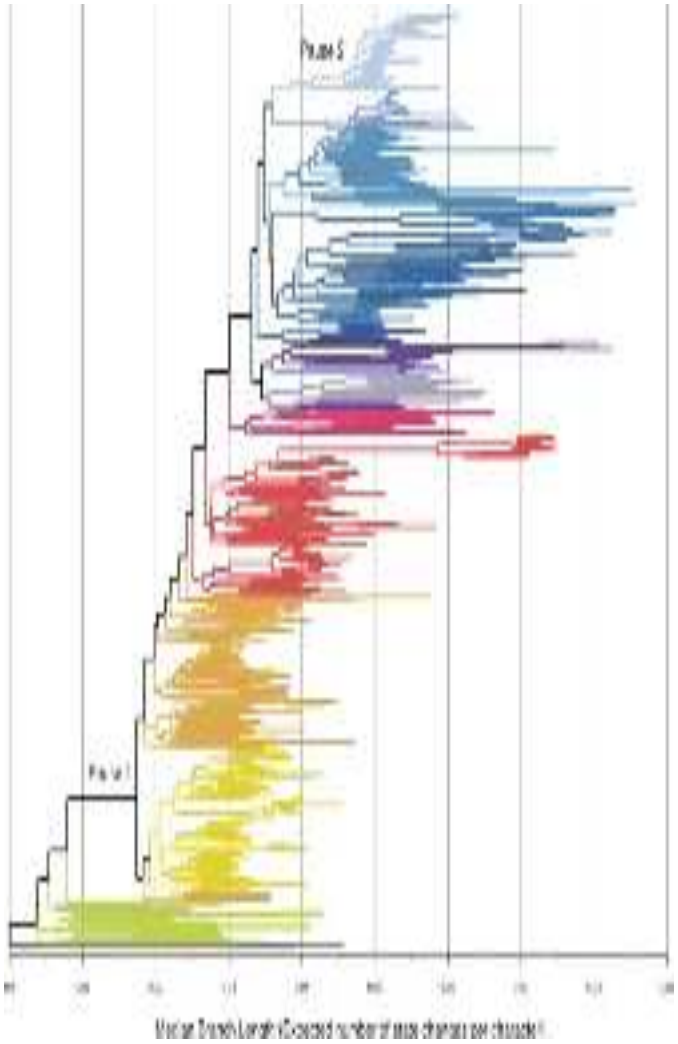
“Linguists don’t do dates.”

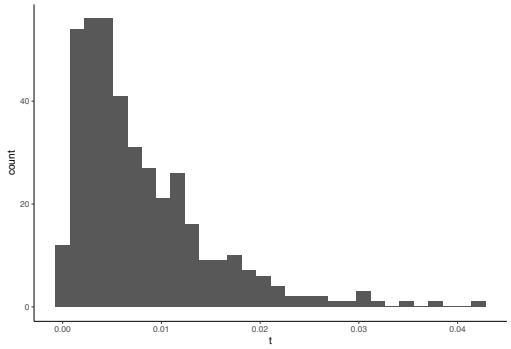
Phylogenetic Dating

ML estimates amount of change along lineage

= (number of changes per character)

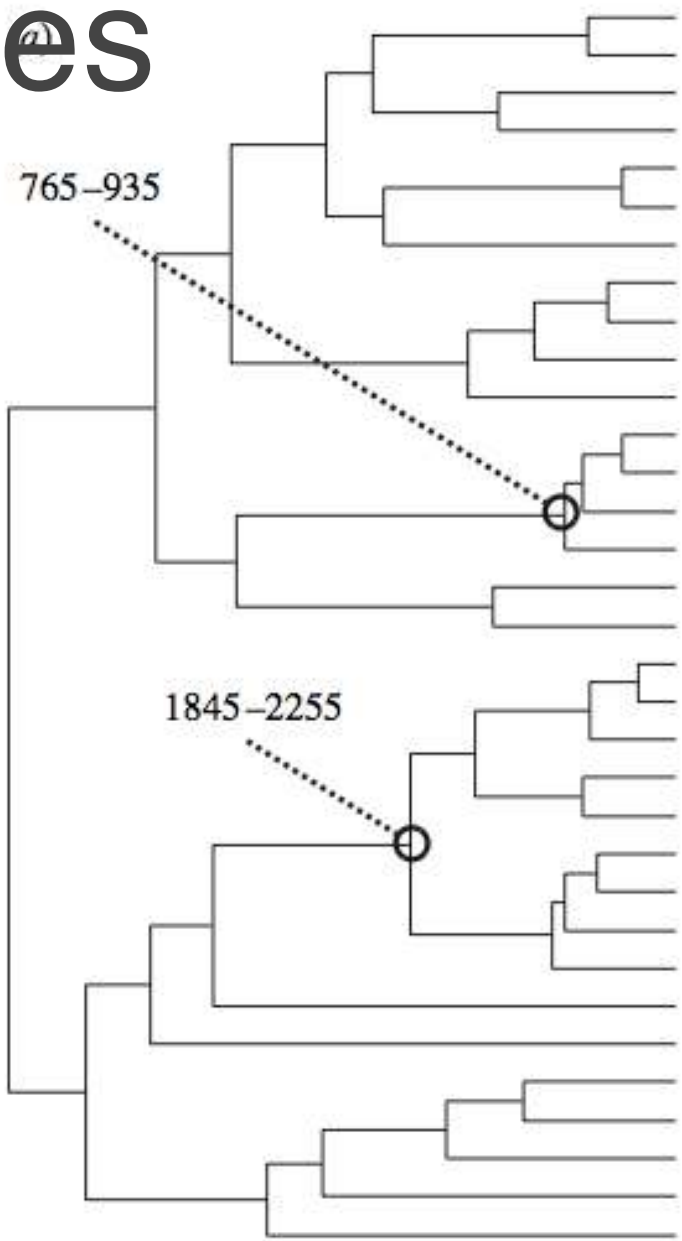
Not a global retention rate but a **per-lineage** estimate of the amount of change.





Convert Rates to Dates

- (pre)historical information to **calibrate** nodes
  - e.g. Archaeology suggests initial settlement was..
  - e.g. Historical evidence says that X and Y were separate by…
- Smooth rates over these calibrations



Strict Clock

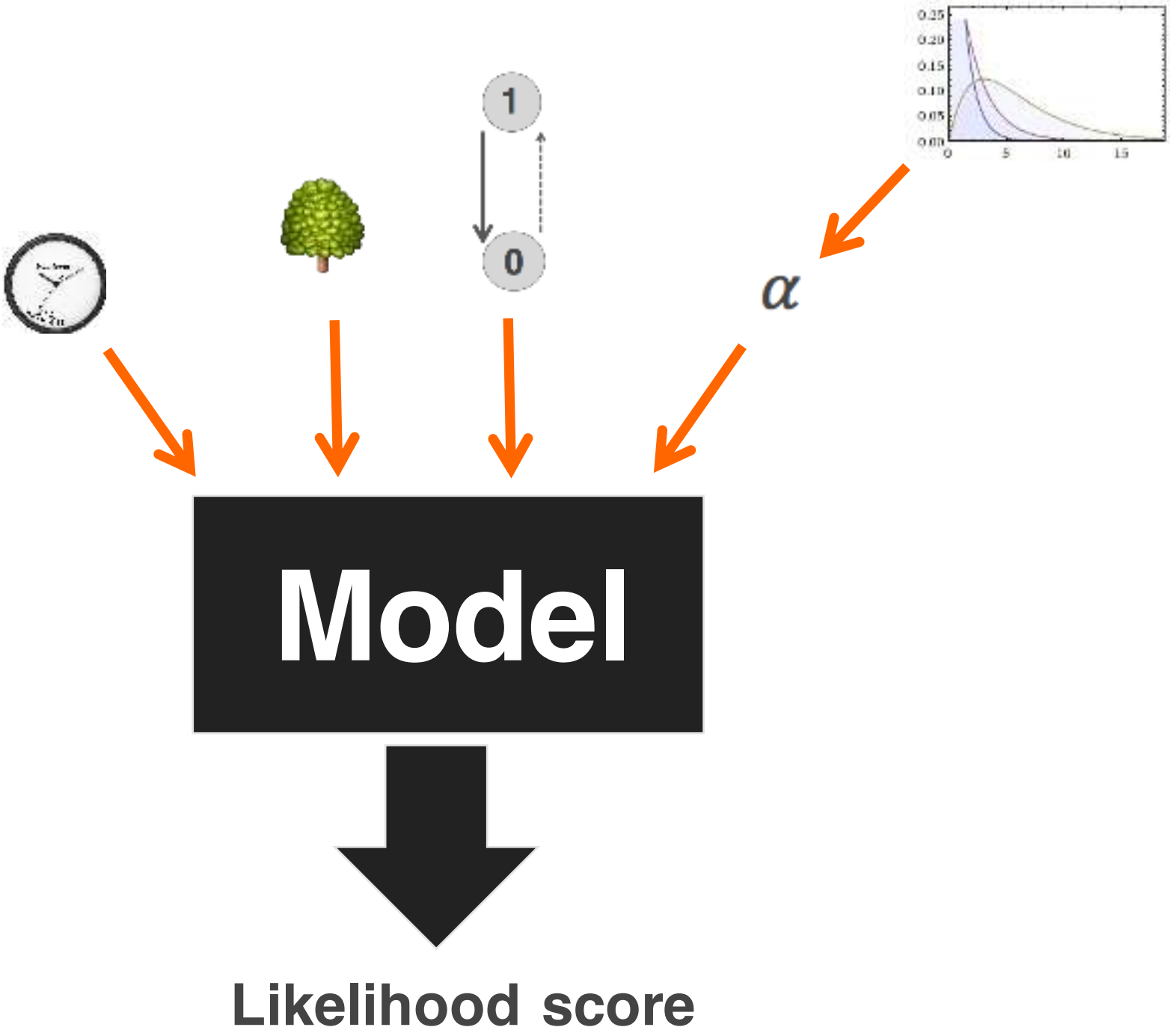
- One rate for all languages.
- No variation
- ~= glottochronology

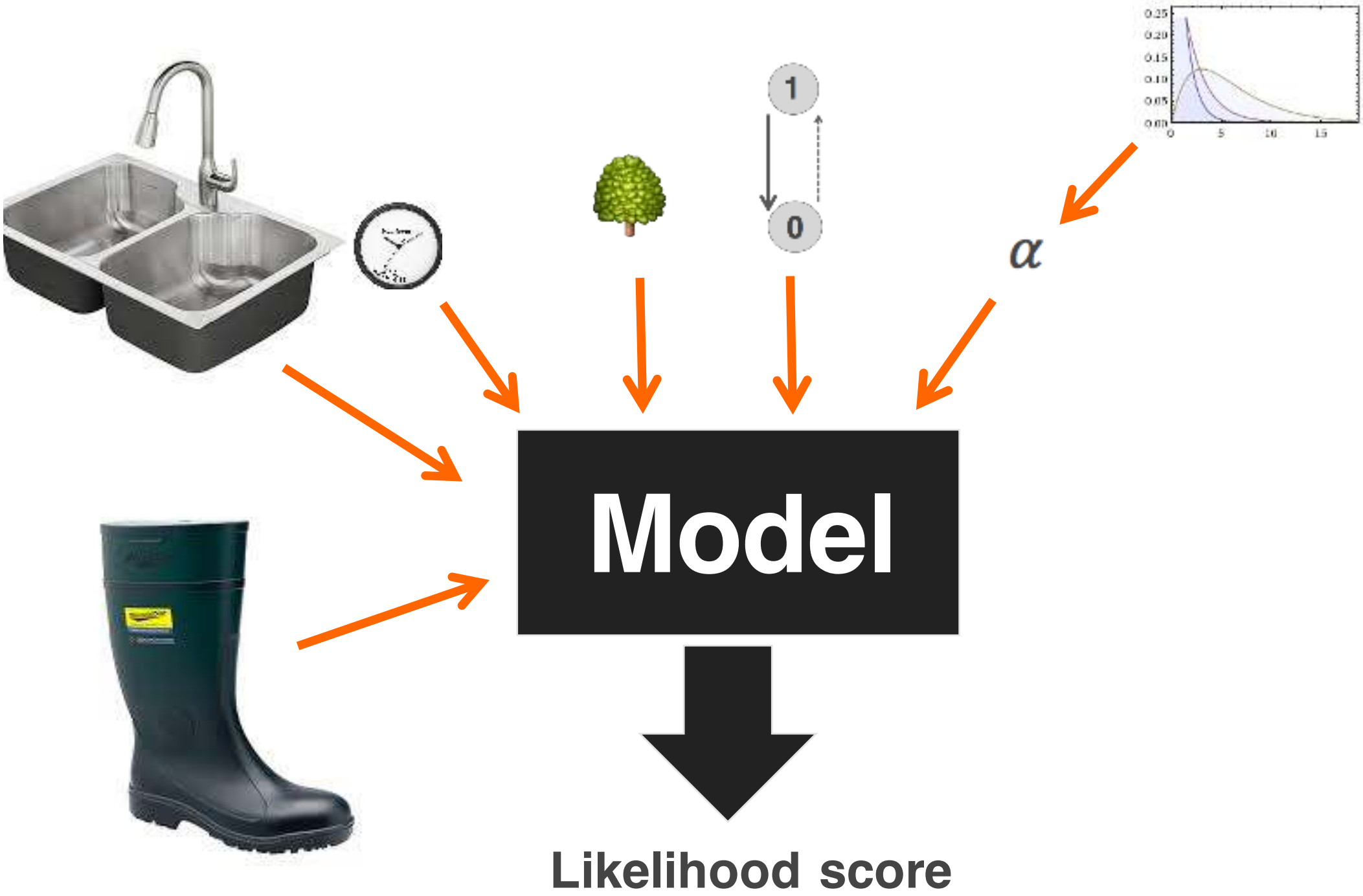
Strict Clock

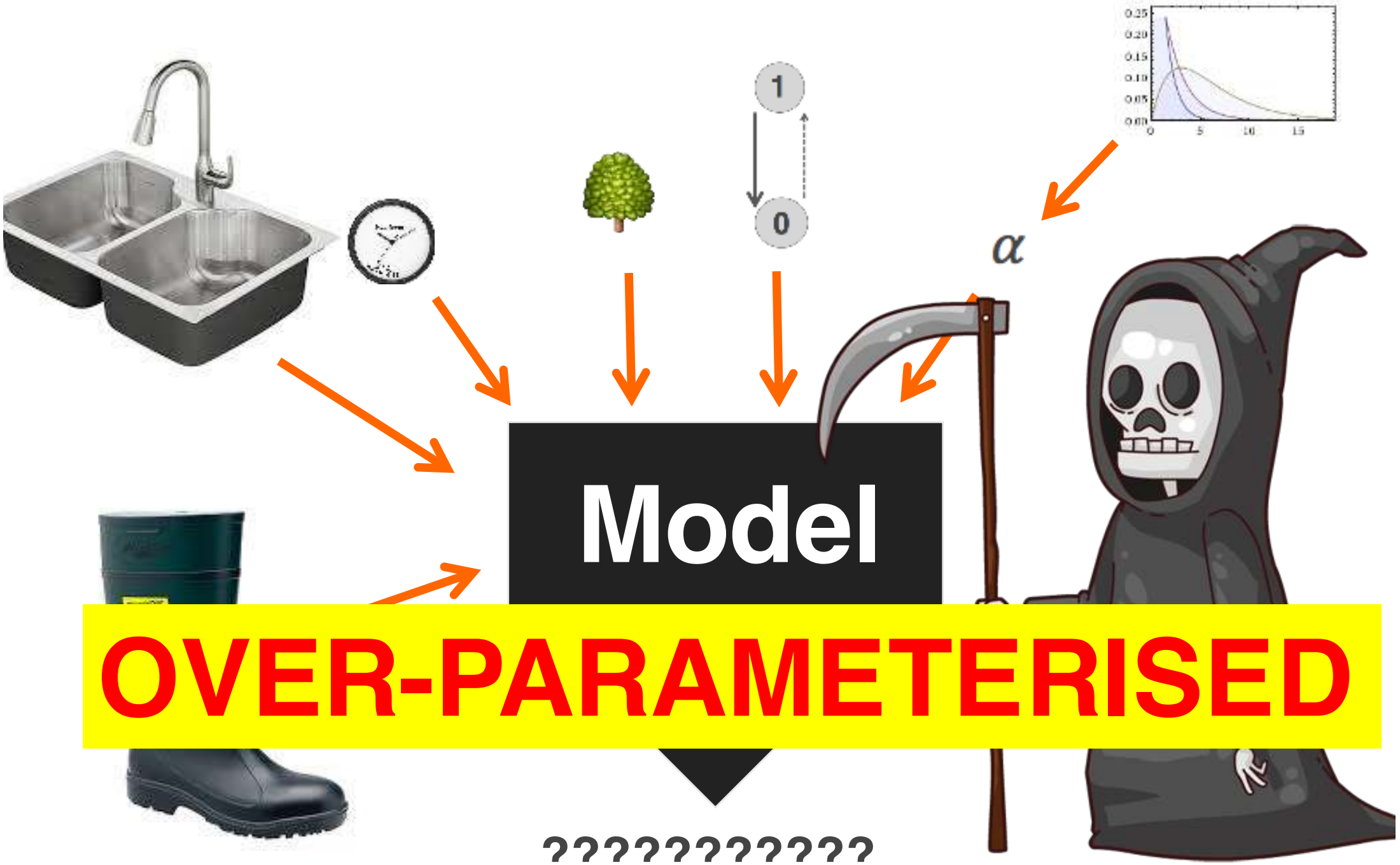
- One rate for all languages.
- No variation
- ~= glottochronology

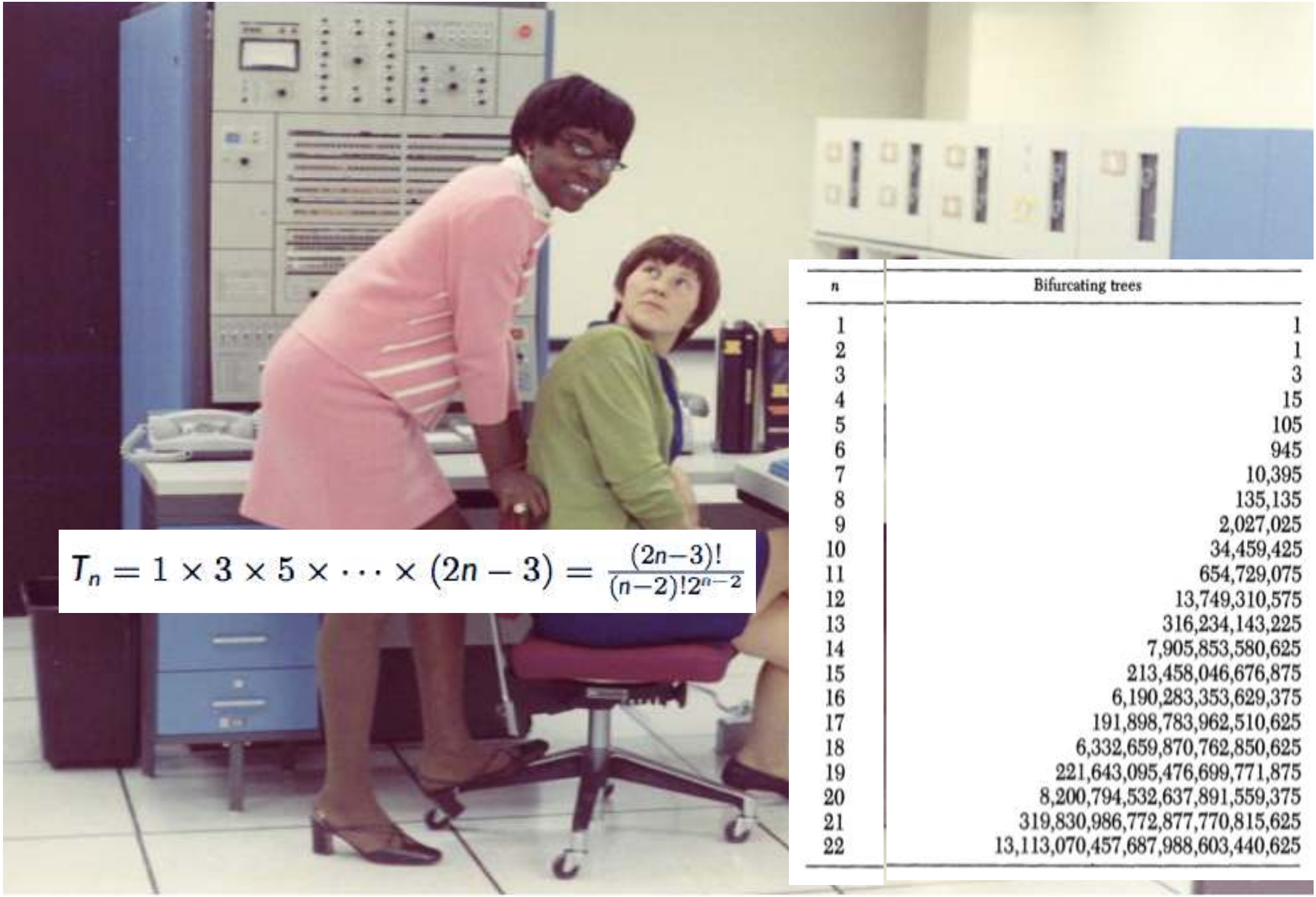
Relaxed Clock

- Estimate distribution of *branch* rates from data
- Rates drawn from a parametric distribution estimated from the data
  - (~=gamma)
  - **LogNormal**, Exponential
- Allows rate to vary across branches (each branch could have own rate)


1
0
α
Model
Likelihood score


1
0
α
Model
Likelihood score


1
0
α
Model
OVER-PARAMETERISED
???????????



$$T_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{(n-2)!2^{n-2}}$$

n	Bifurcating trees
1	1
2	1
3	3
4	15
5	105
6	945
7	10,395
8	135,135
9	2,027,025
10	34,459,425
11	654,729,075
12	13,749,310,575
13	316,234,143,225
14	7,905,853,580,625
15	213,458,046,676,875
16	6,190,283,353,629,375
17	191,898,783,962,510,625
18	6,332,659,870,762,850,625
19	221,643,095,476,699,771,875
20	8,200,794,532,637,891,559,375
21	319,830,986,772,877,770,815,625
22	13,113,070,457,687,988,603,440,625

Need to handle complexity

- ML and MP give **a point** estimate.

- But a single tree is not enough.
  - Reality is complicated.
  - Need to estimate uncertainty around that estimate.
  - “confidence intervals” = how confident can I be about my estimate.

Bayesian methods

- Extend ML methods to explicitly account for uncertainty

$$\Pr(H|D) = \frac{\Pr(H)\Pr(D|H)}{\Pr(D)}$$

General intro: McElreath “Statistical Rethinking” (http://xcelab.net/rm/statistical-rethinking/)
Phylogenetics: Brown “The State of Bayesian Phylogenetics: Bayes for the Uninitiated” (https://tinyurl.com/ly2cvfd)

Bayesian methods

- Extend ML methods to explicitly account for uncertainty

Posterior
(Pr hypothesis given data)

prior

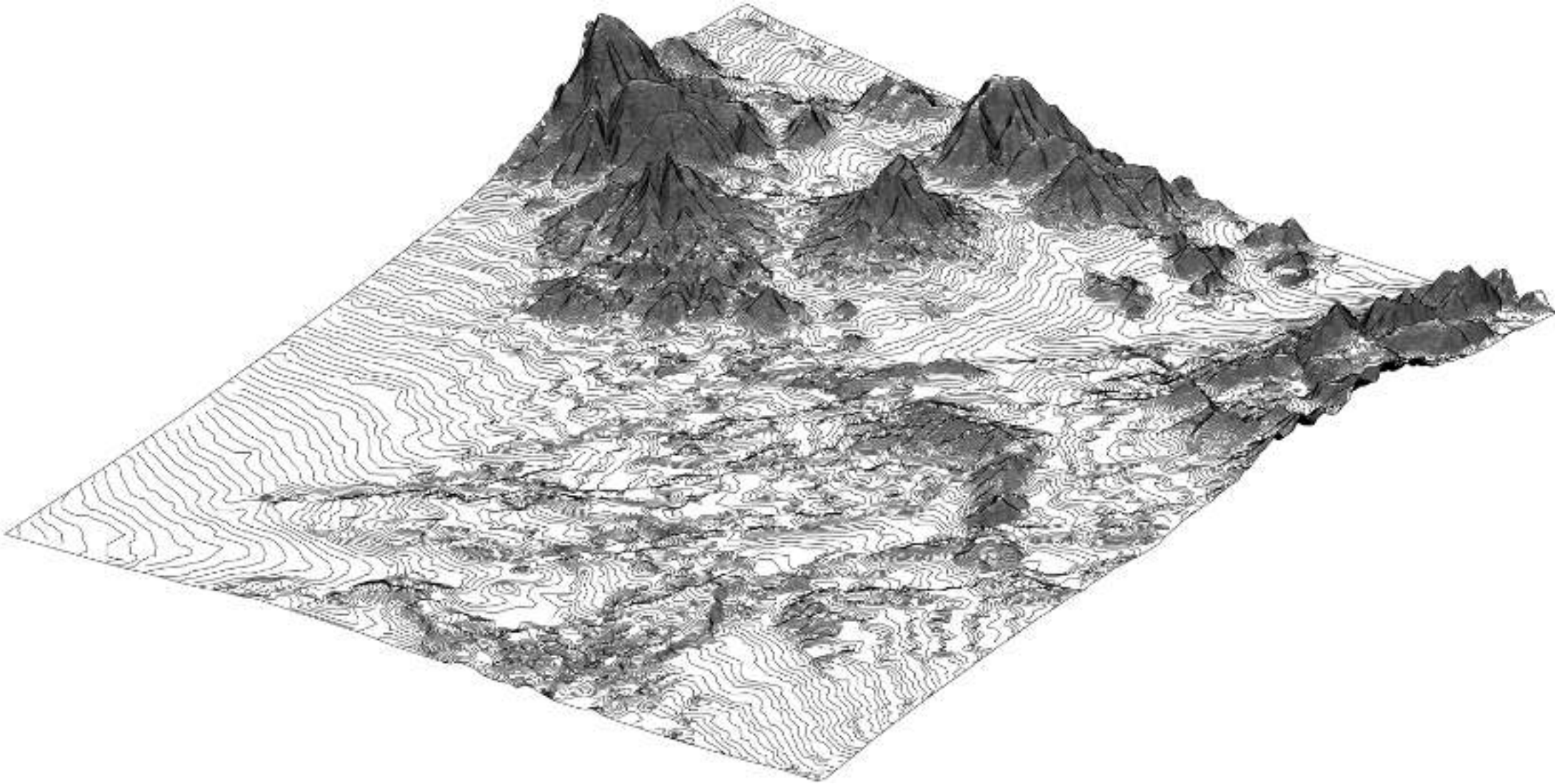
Lh

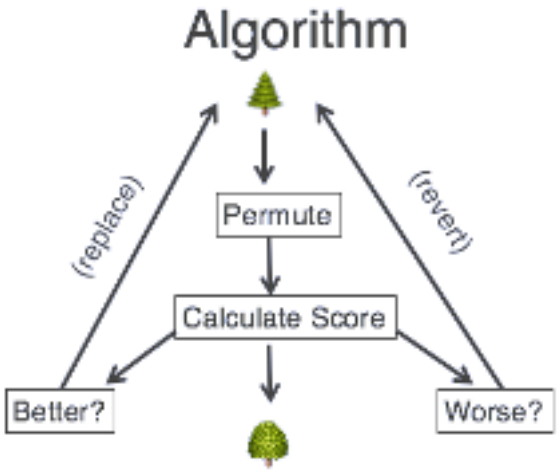
$$\Pr(H|D) = \frac{\Pr(H)\,\Pr(D|H)}{\Pr(D)}$$

Prior Pr that data is true

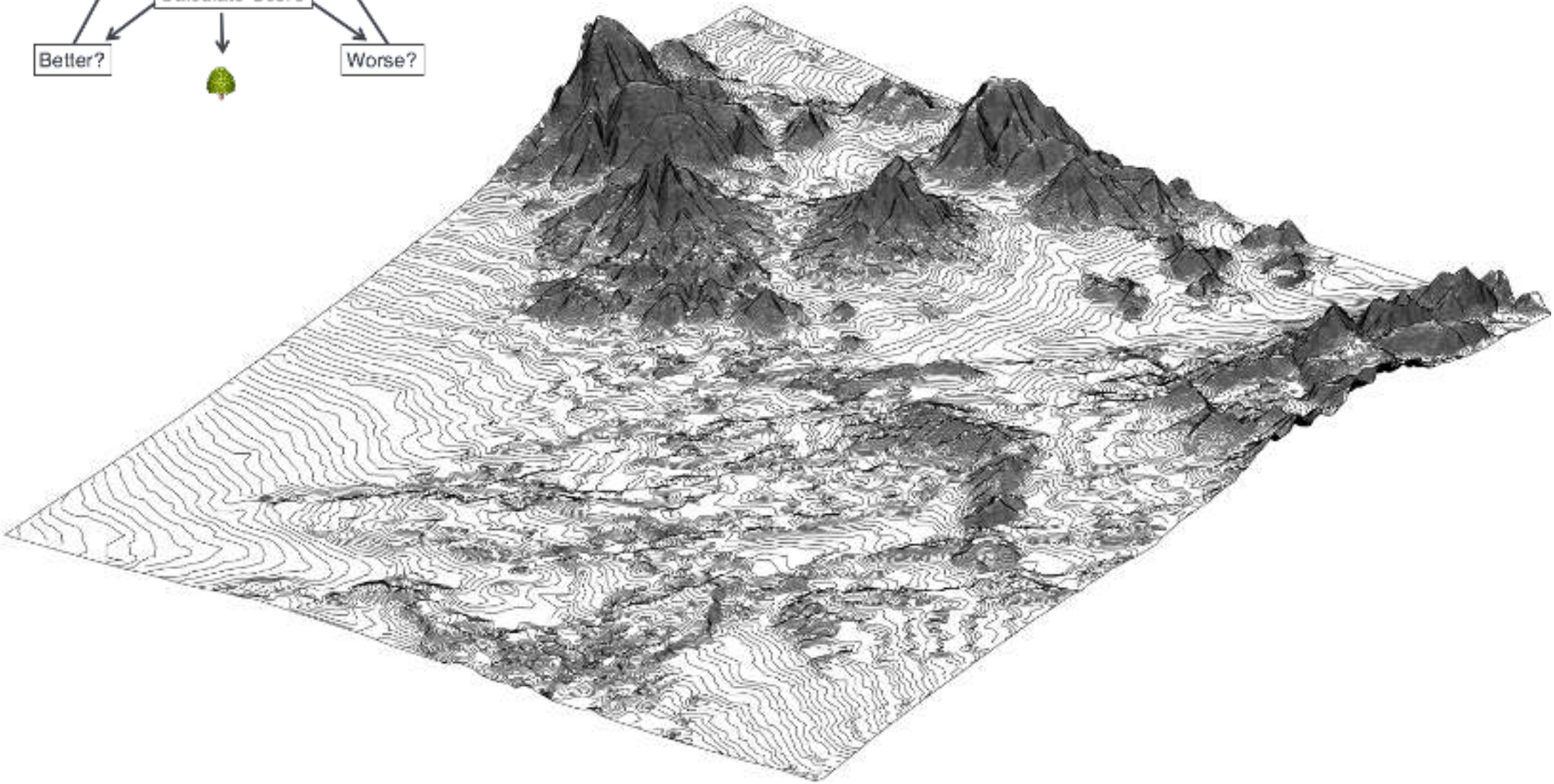
General intro: McElreath “Statistical Rethinking” (http://xcelab.net/rm/statistical-rethinking/)
Phylogenetics: Brown “The State of Bayesian Phylogenetics: Bayes for the Uninitiated” (https://tinyurl.com/ly2cvfd)

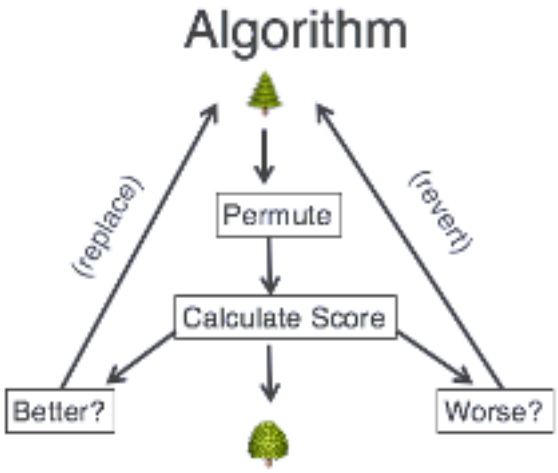
Treespace



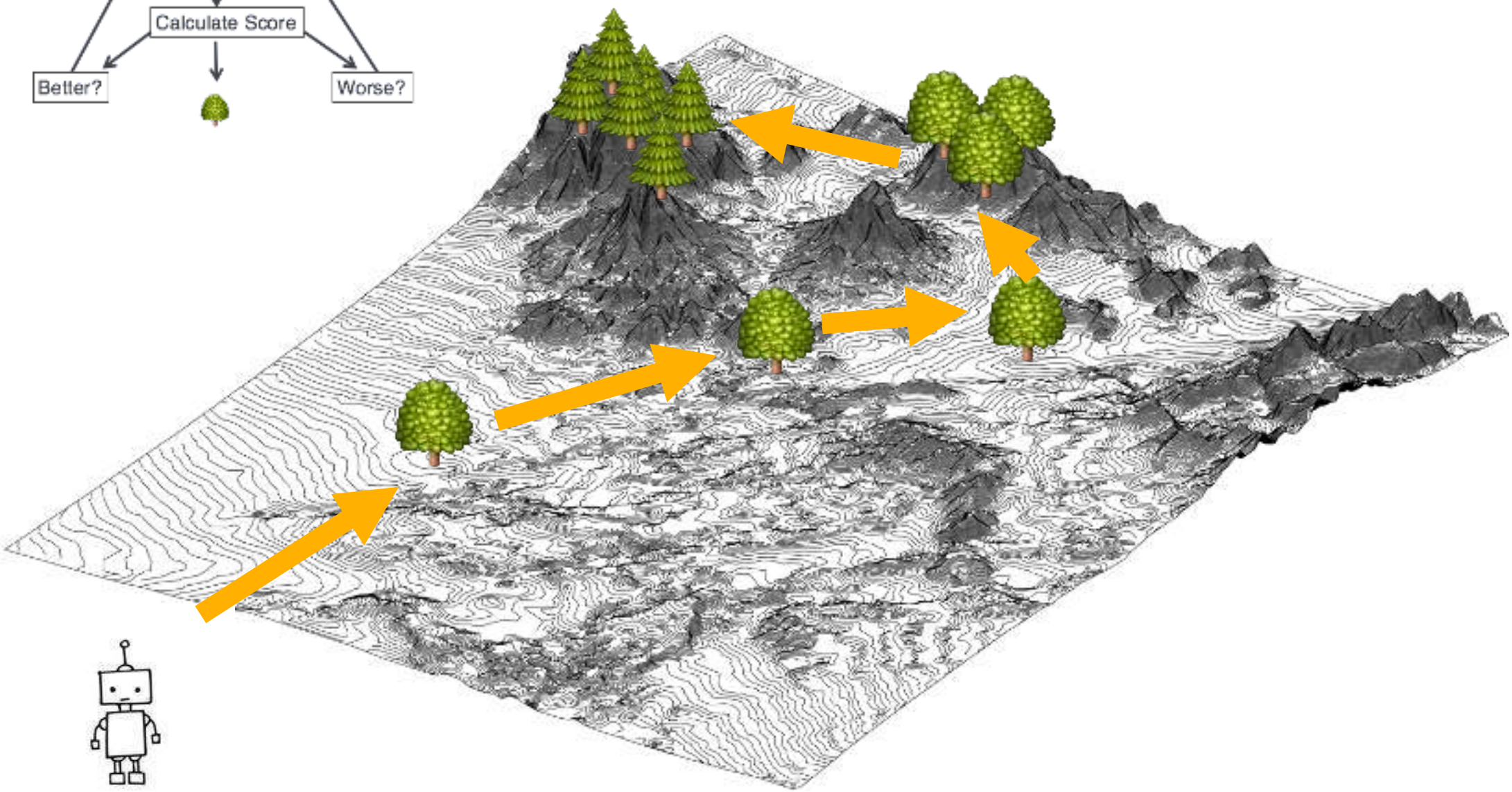

Algorithm
Permute
Calculate Score
Better?
Worse?
(replace)
(revert)

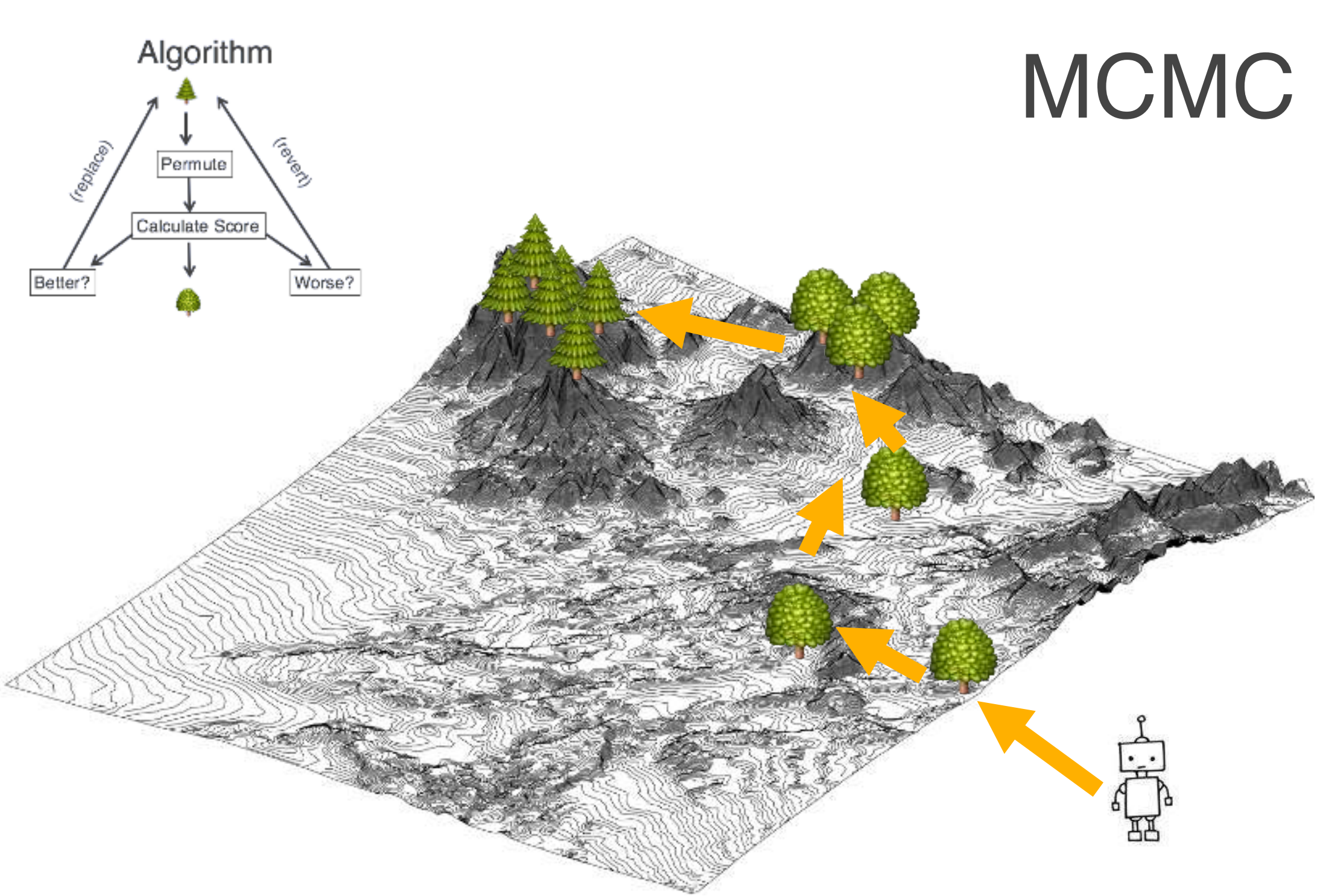
MCMC

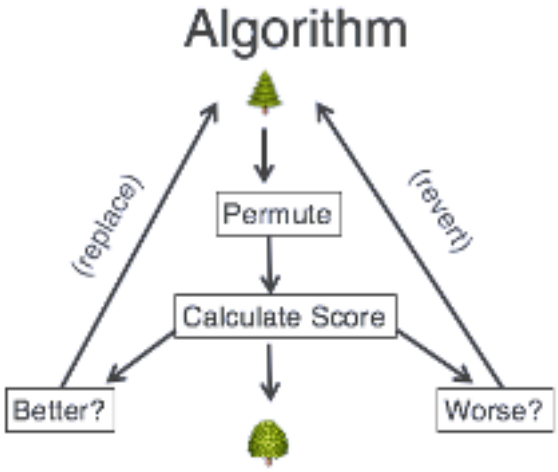


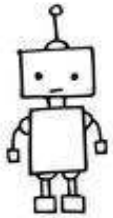

Algorithm
Permute
Calculate Score
Better?
Worse?
(replace)
(revert)

MCMC

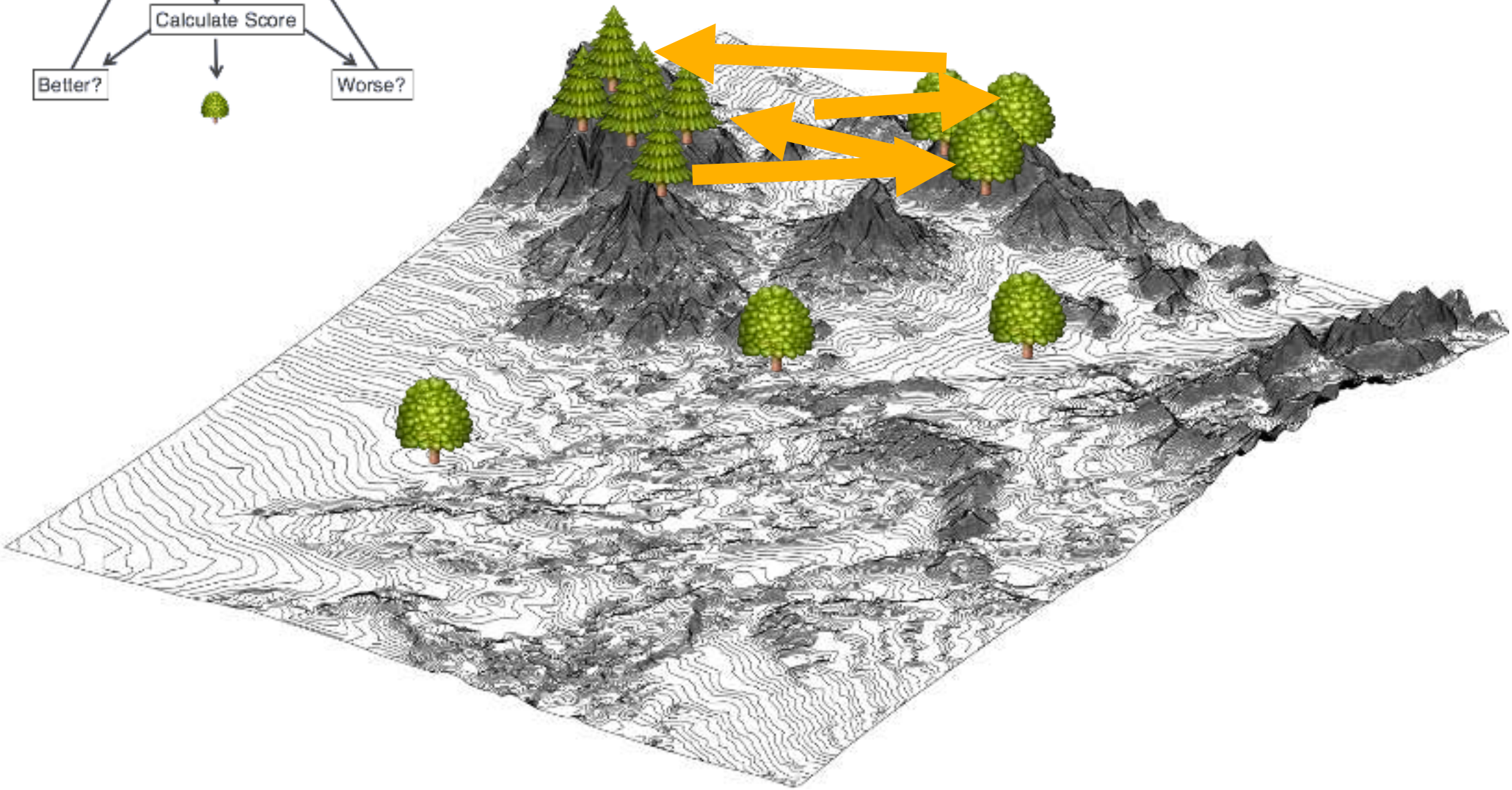


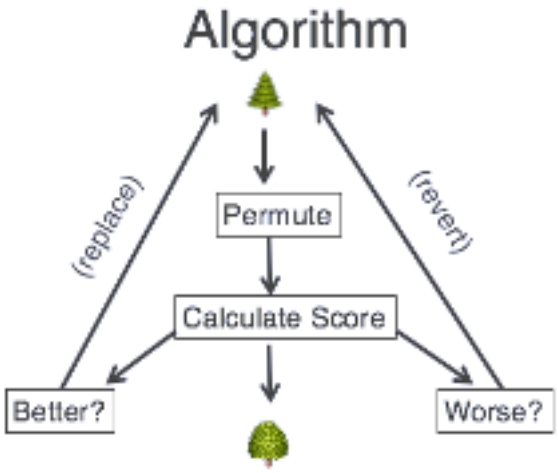

Algorithm
Permute
Calculate Score
Better?
Worse?
(replace)
(revert)
MCMC





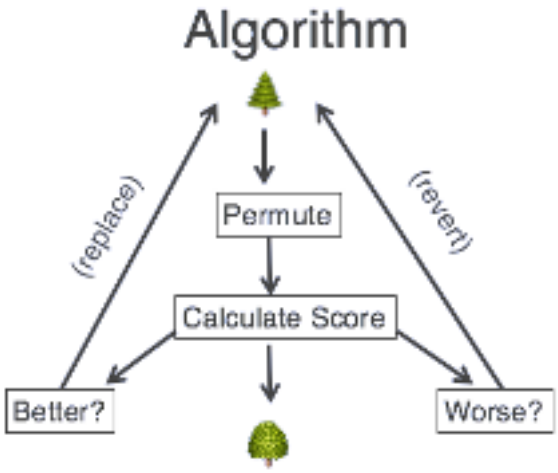
“stationarity”

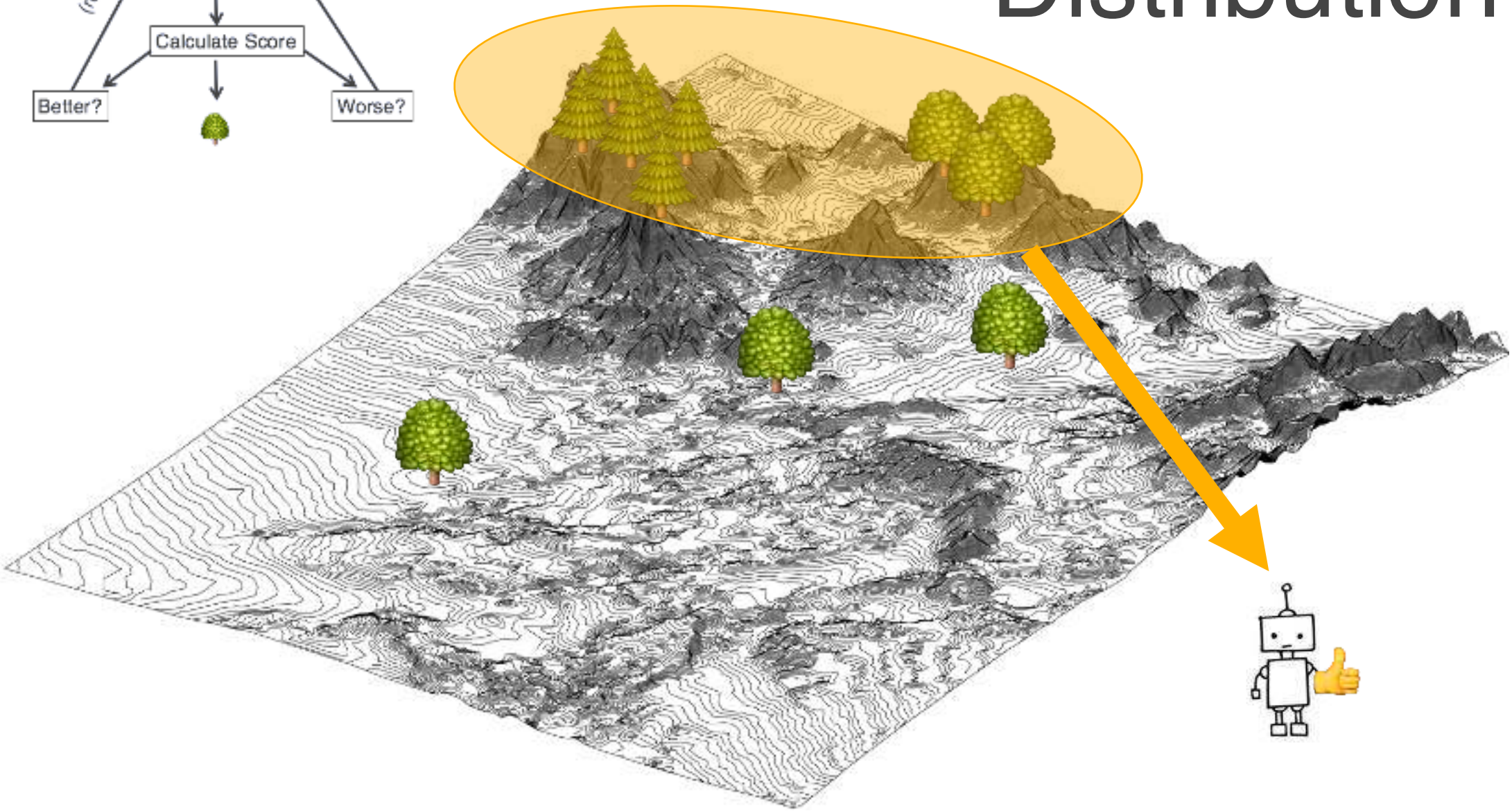



Algorithm
Permute
Calculate Score
Better?
Worse?
(replace)
(revert)

“burn-in”

Posterior Probability Distribution





Posterior Probability Distribution:

Uto-Aztecan Languages

Covarion & Relaxed Clock + etc

Sampled 10,000 trees.

Posterior Probability Distribution:

Uto-Aztecan Languages

Covarion & Relaxed Clock + etc
Sampled 10,000 trees.
Draw each one (Densitree, Bouckaert ‘10)
Some well supported regions
Some less…
Some conflict

Posterior Probability Distribution:

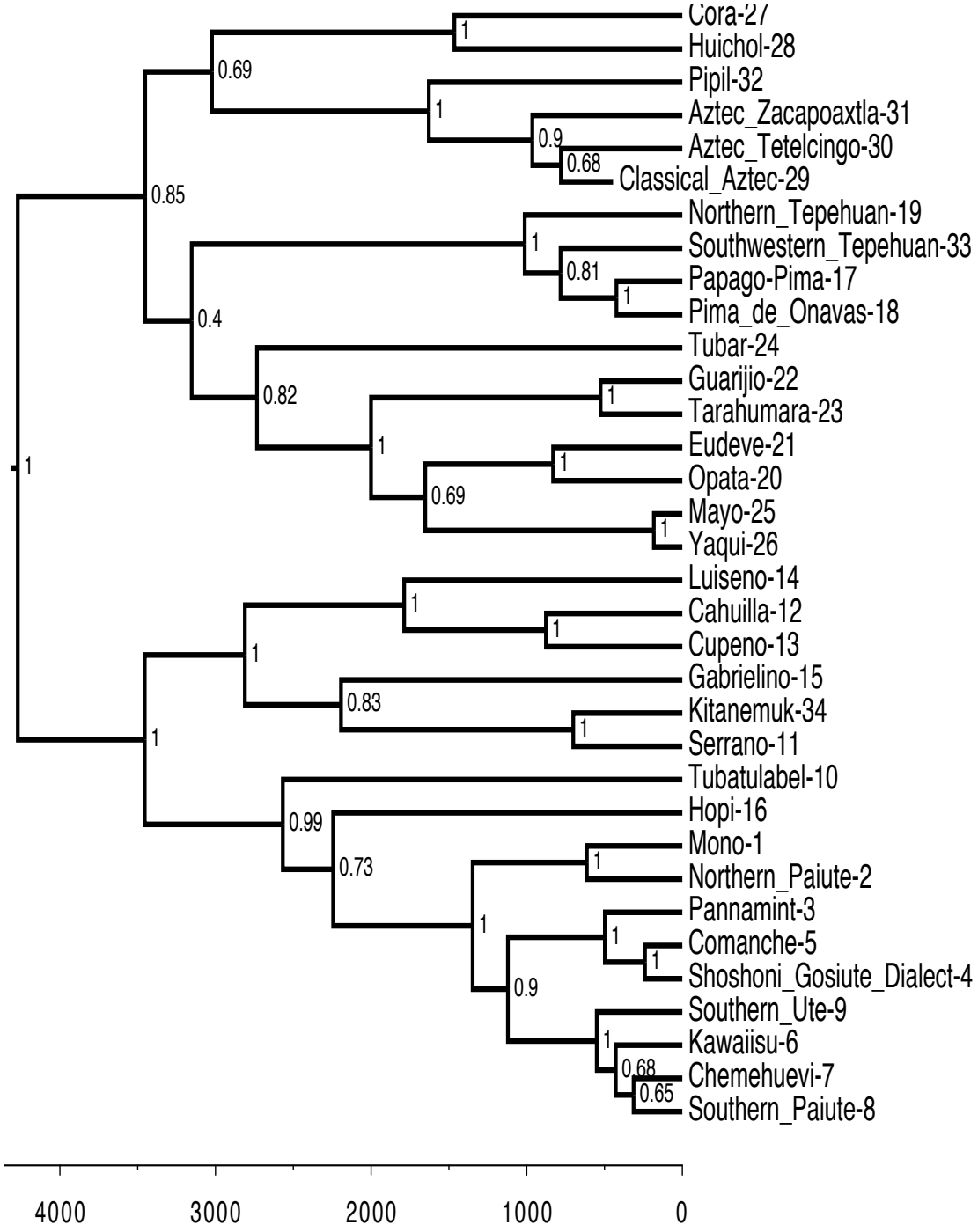
Uto-Aztecan Languages

Maximum Clade Credibility Tree

Reduce posterior to single summary tree

Values are “Posterior probabilities”

- 1.00 = present in ALL trees in P.
- 0.50 = only half of the trees have this.



Posterior Probability Distribution:

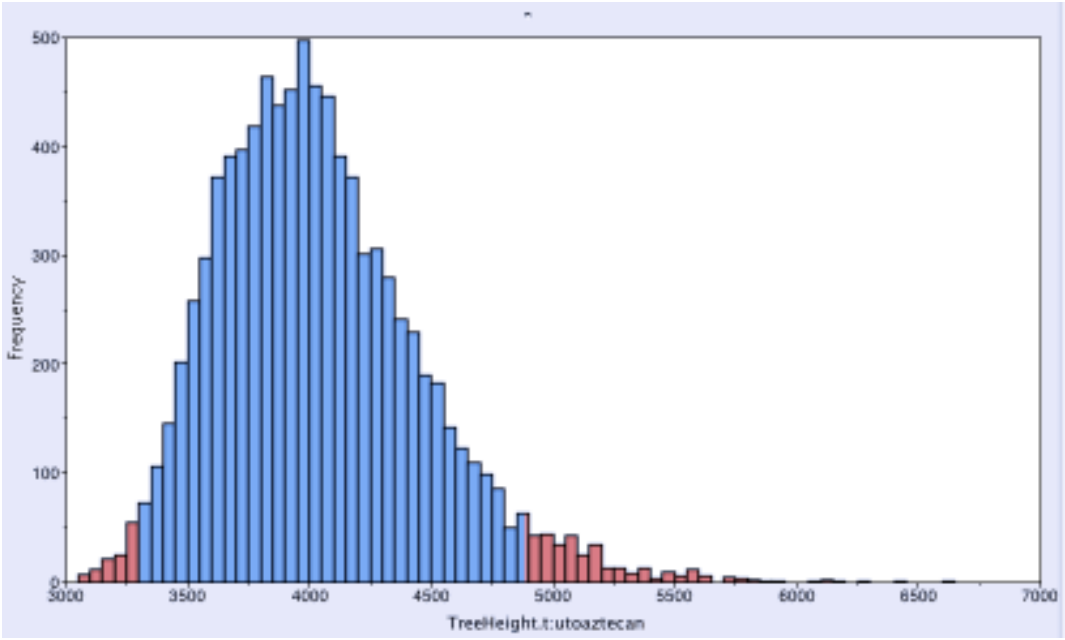
Uto-Aztecan Languages

…and any statistic we calculate has a PP distribution.

- mean
- 95% highest posterior density interval

  (95% probability that μ falls in range)

Mean = 4334 years



95% HPD: 3302 – 4884 years

95% HPD: 3302 – 4884 years

Whirlwind tour

- Disciplinary parallels & similarity of questions
- Lexicostatistics and Glottochronology
- Parallel evolution of methods in Biology
- Maximum Parsimony
- Maximum Likelihood
- Bayesian Phylogenetics

- Theory

- Practical: My BEAST tutorial tomorrow.
  - https://taming-the-beast.github.io/