Lingüística histórica moderna
Métodos filogenéticos

Ezequiel Koile

Max-Planck-Institut für Menschheitsgeschichte (Jena, Alemania)

Primer Simposio "Análisis cuantitativo de datos lingüísticos"

Biblioteca Nacional,

Buenos Aires

Febrero de 2019

Motivation

- Studying HL with a cross-linguistic perspective is useful for:
  - Knowing the history of the languages
  - Literary studies / Phylologies
- But it also can provide insight into:
  - Linguistic typology
  - Human cognition
  - Psychology of language
  - Cultural evolution

What historical linguistics is about

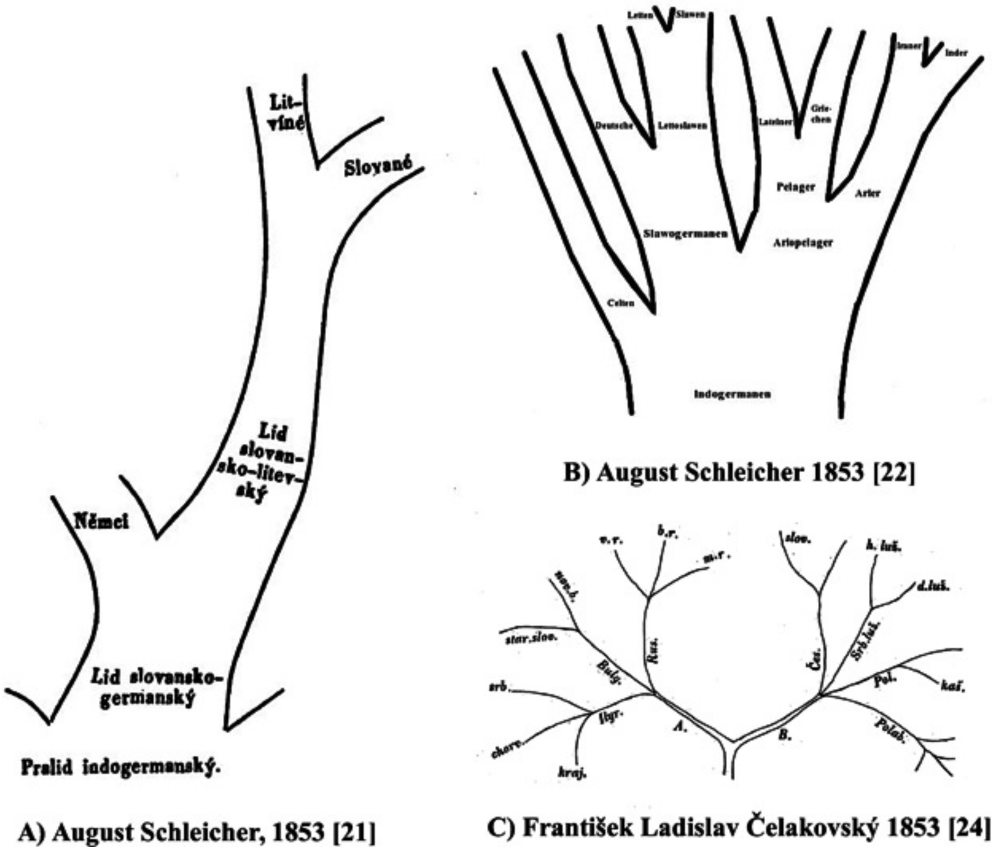
- In the past, HL's main concern was on *how* languages change
- Since the 1960s, *how* and *why* languages change
- Not studying individual etymologies of words, but the kinds of changes they have undergone and the techniques or methods we have at our disposal to recover this history

Introduction

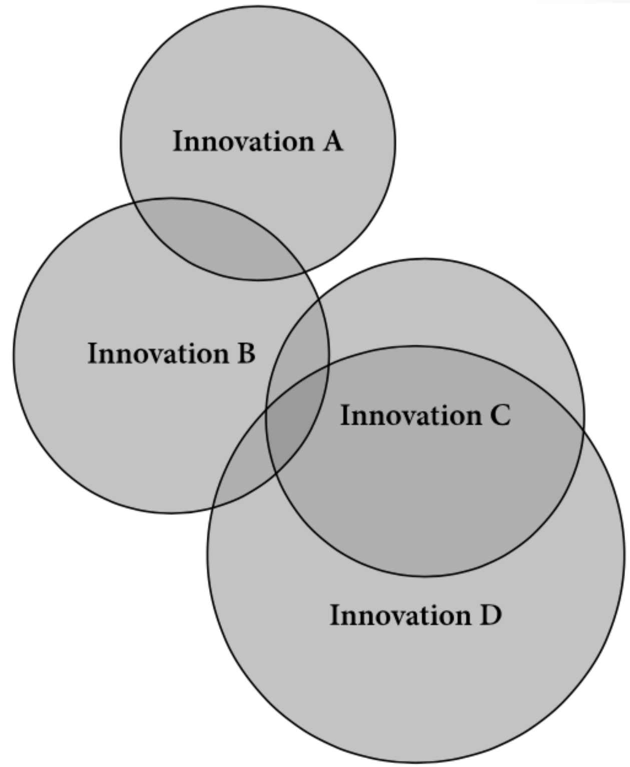
- Current evolutionary theory offers a rigorous, quantifiable approach to phylogenetic inference.
- Linguistic phylogenetics incorporates the whole approach of the phylogenetic comparative method
- The quantified, algorithmic approach to phylogenetics started in the early 1960s. Linguistics has been part of this movement twice: firstly with the development of lexicostatistics and glothochronology in the late 1960s, and again with the development of model based, hypothesis-testing (and usually Bayesian) approaches starting around 2000.

[Dunn 2013]

Historical linguistics



Schleicher’s trees
(Img from List et al 2014)



Schmidt’s waves

Linguistic reconstruction

TABLE 2.1: Sanskrit–Latin cognates showing Sanskrit merger of *e*, *o*, *a* > *a*

Sanskrit	*Latin*	*Proto-Indo-European*	
ad-	ed-	*ed-	'to eat'
danta	dent-	*dent-	'tooth'
avi-	ovi-	*owi-	'sheep'
dva-	duo	*dwo-	'two'
ajra-	ager	*agro-	'field' (compare *acre*)
apa	ab	*apo	'away, from'

	PIE	*Greek*	*Latin*	*Gothic*	*OHG*	*English*
*o	*oktō(u)-	oktṓ	octo	ahtau [axtau]	ahto	'eight'
*ə	*pəter-	patḗr	pater	fadar	fater	'father'
*a	*agro-	agrós	ager	akrs	ackar	'field' (acre)

Campbell 2013

Relations between languages

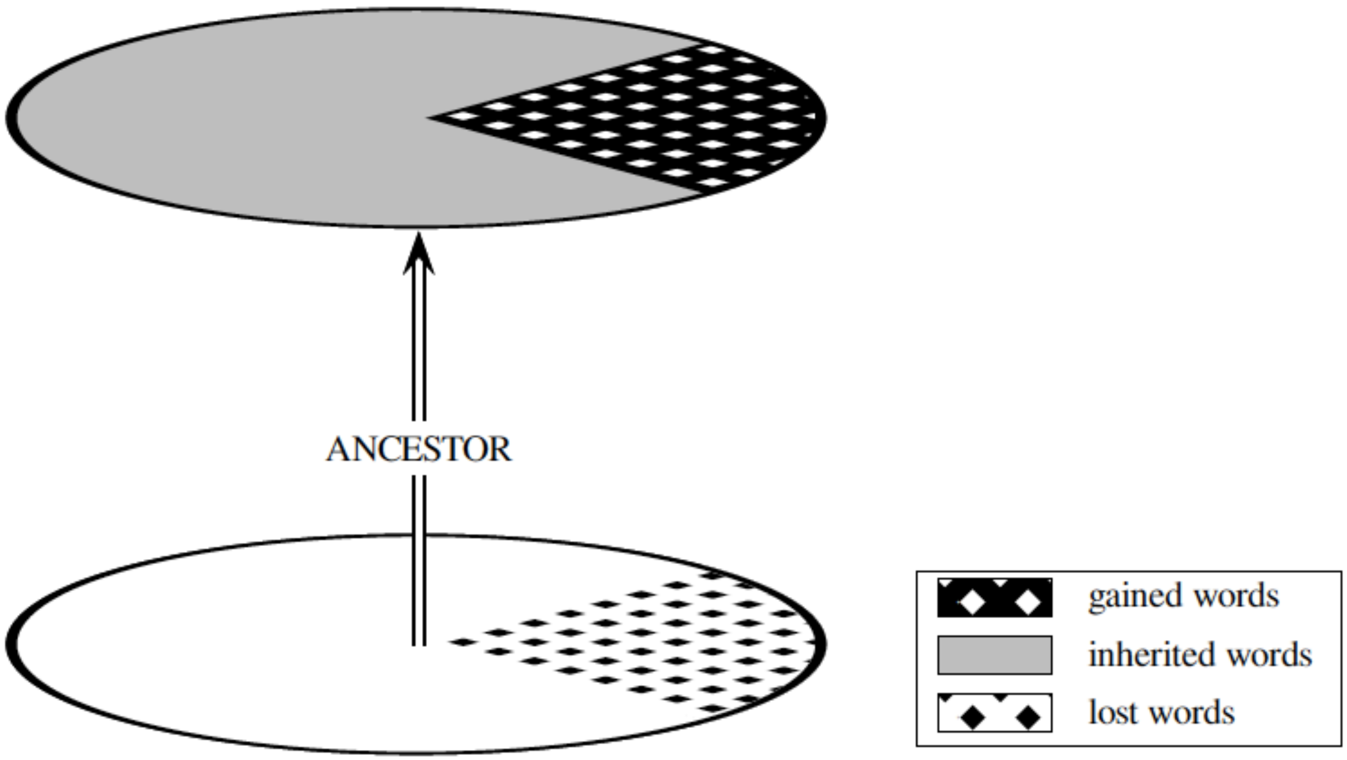


Figure 2.10: *Ancestor-descendant relation between languages*

[List 2014]

Relations between languages

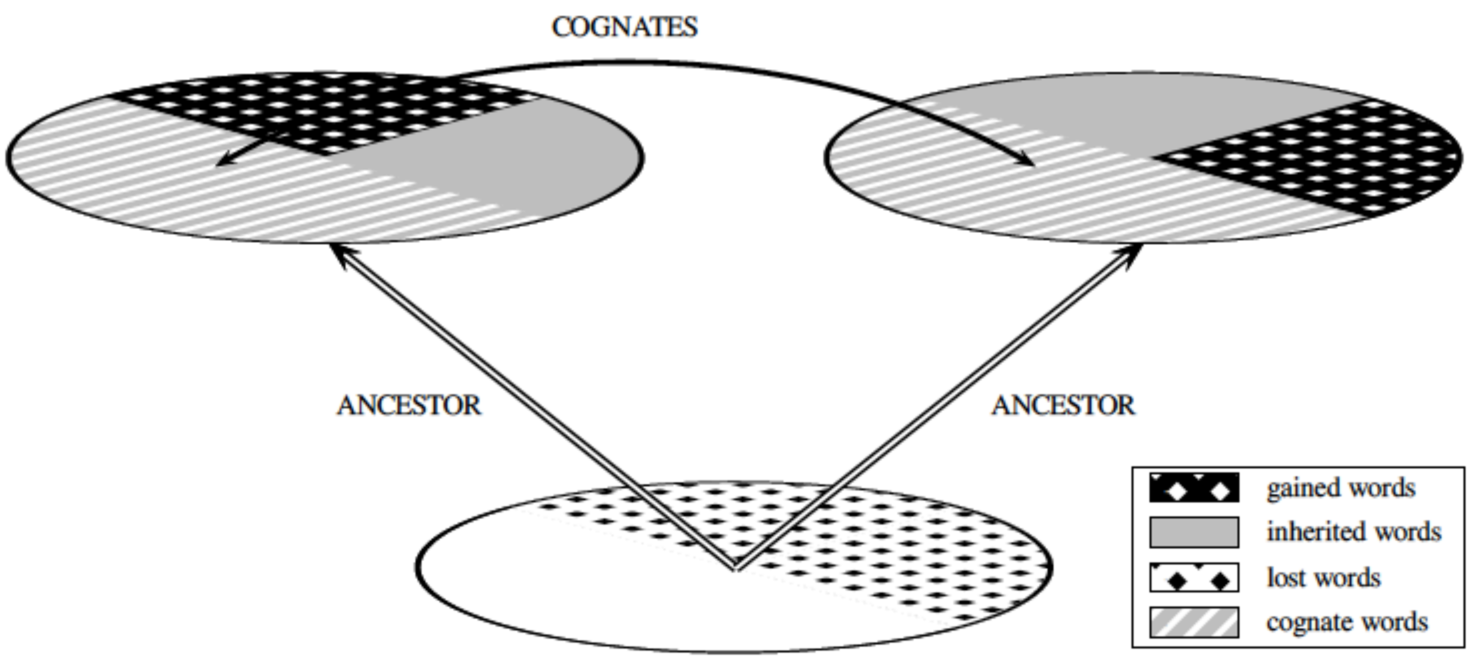


Figure 2.11: *Genetic relation between languages*

[List 2014]

Relations between languages

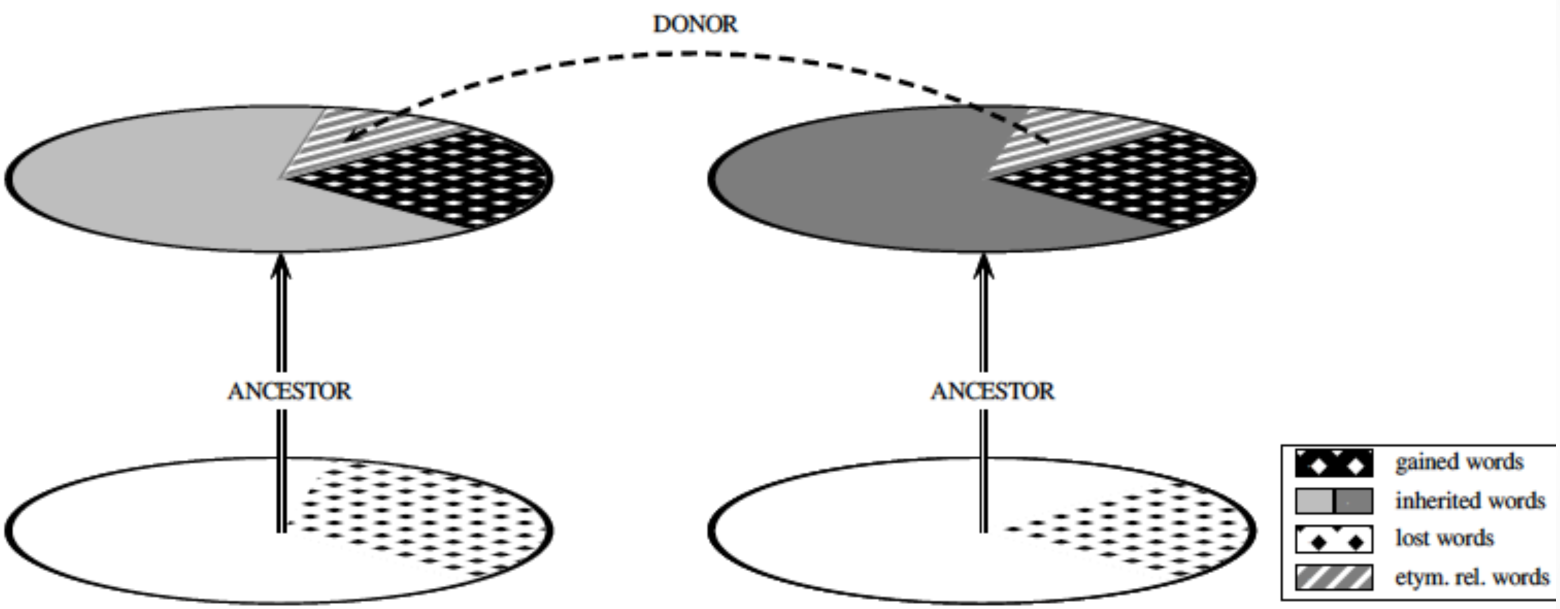


Figure 2.12: *Contact relation between languages*

[List 2014]

Sound change

- What is the rule for these changes?

Meaning	Latin	Italian
"feather"	plu:ma	pjuma
"flat"	pla:nus	pjano
"square"	plate:a	pjats:a

- Sound change is a *recurrent process*
- Sound change is a *contextually restricted process*

- Therefore: *regular sound change*

Cognate testing

[List 2014]

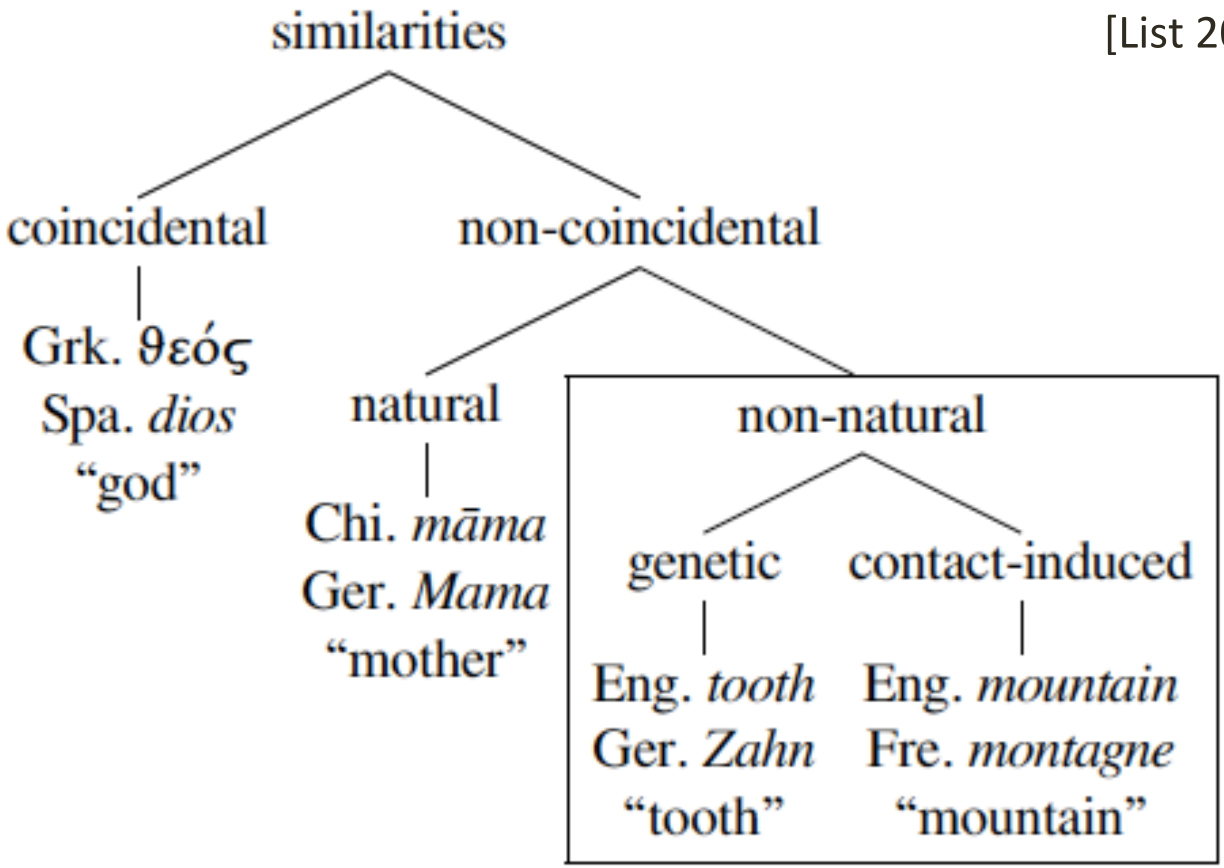


Figure 2.13: *Common causes for resemblances in the form material of languages: Both kinds of non-natural resemblances are "historical" and constitute one of the key objectives of historical linguistics.*

Steps of cognate detection

- Wordlist
- Pairwise comparison
- Pairwise distances between words
- Cognate clustering
- Cognate sets

Cognate detection

Goal:

ID	Taxa	Word	Gloss	GlossID	IPA	...
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	...
22	Dutch	vrouw	woman	20	vrɑu	...
23	English	woman	woman	20	wʊmən	...
24	Danish	kvinde	woman	20	kvenə	...
25	Swedish	kvinna	woman	20	kvi:na	...
26	Norwegian	kvine	woman	20	kʋinə	...
...	...	...	...	...	...	...

(a) *Input*

ID	Taxa	Word	Gloss	GlossID	IPA	CogID
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	1
22	Dutch	vrouw	woman	20	vrɑu	1
23	English	woman	woman	20	wʊmən	2
24	Danish	kvinde	woman	20	kvenə	3
25	Swedish	kvinna	woman	20	kvi:na	3
26	Norwegian	kvine	woman	20	kʋinə	3
...	...	...	...	...	...	...

(b) *Output*

Table 4.20: *Input* (a) *and output format* (b) *of LexStat. Four columns are required in the input: ID, Taxa, GlossID, and IPA. An additional column is added in the output. Each word is assigned a specific cognate ID (CogID). All words that have the same CogID have been identified as cognates by the algorithm.*

[List 2014]

An example

Cognate List	
German	*dünn*
English	*thin*
German	*Ding*
English	*thing*
German	*dumm*
English	*dumb*

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*dumm*	d	ʊ	m
English	*dumb*	d	ʌ	m

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*dumm*	d	ʊ	m
English	*dumb*	d	ʌ	m

Correspondence List

GER	ENG	Frequ.
d	θ	2 x
d	d	1 x
n	n	1 x
m	m	1 x
ŋ	ŋ	1 x

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*dumm*	d	ʊ	m
English	*dumb*	d	ʌ	m

Correspondence List

GER	ENG	Frequ.
d	θ	2 x
d	d	1 x
n	n	1 x
m	m	1 x
ŋ	ŋ	1 x

Eng\ Ger	d	n	m	ŋ
θ	2	0	0	0
d	1	0	0	0
n	0	1	0	0
m	0	0	1	0
ŋ	0	0	0	1

[List 2017]

An example

Cognate List		**Alignment**			**Correspondence List**		
German	*dünn*	d	ʏ	n	GER	ENG	Frequ.
English	*thin*	θ	ɪ	n	d	θ	2 x
German	*Ding*	d	ɪ	ŋ	d	d	1 x
English	*thing*	θ	ɪ	ŋ	n	n	1 x
German	*dumm*	d	ʊ	m	m	m	1 x
English	*dumb*	d	ʌ	m	ŋ	ŋ	1 x
German	*Dorn*	d	ɔɐ	n			
English	*thorn*	θ	ɔː	n			

[List 2017]

An example

Cognate List		**Alignment**			**Correspondence List**		
German	*dünn*	d	ʏ	n	GER	ENG	Frequ.
English	*thin*	θ	ɪ	n	d	θ	3 x
German	*Ding*	d	ɪ	ŋ	d	d	1 x ?
English	*thing*	θ	ɪ	ŋ	n	n	2 x
German	*dumm*	d	ʊ	m	m	m	1 x
English	*dumb*	d	ʌ	m	ŋ	ŋ	1 x
German	*Dorn*	d	ɔɐ	n			
English	*thorn*	θ	ɔː	n			

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*dumm*	d	ʊ	m
English	*dumb*	d	ʌ	m
German	*Dorn*	d	ɔɐ	n
English	*thorn*	θ	ɔː	n

Correspondence List

GER	ENG	Frequ.
d	θ	3 x
d	d	1 x ?
n	n	2 x
m	m	1 x
ŋ	ŋ	1 x

Eng\ Ger	d	n	m	ŋ
θ	3	0	0	0
d	1	0	0	0
n	0	2	0	0
m	0	0	1	0
ŋ	0	0	0	1

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
~~German~~	~~*dumm*~~	~~d~~	~~ʊ~~	~~m~~
~~English~~	~~*dumb*~~	~~d~~	~~ʌ~~	~~m~~
German	*Dorn*	d	ɔɐ	n
English	*thorn*	θ	ɔː	n

Correspondence List

GER	ENG	Frequ.
d	θ	3 x
d	d	1 x ?
n	n	2 x
m	m	1 x
ŋ	ŋ	1 x

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*Dorn*	d	ɔɐ	n
English	*thorn*	θ	ɔː	n

Correspondence List

GER	ENG	Frequ.
d	θ	3 x
n	n	2 x
ŋ	ŋ	1 x

[List 2017]

An example

Cognate List		Alignment		
German	*dünn*	d	ʏ	n
English	*thin*	θ	ɪ	n
German	*Ding*	d	ɪ	ŋ
English	*thing*	θ	ɪ	ŋ
German	*Dorn*	d	ɔɐ	n
English	*thorn*	θ	ɔː	n

Correspondence List

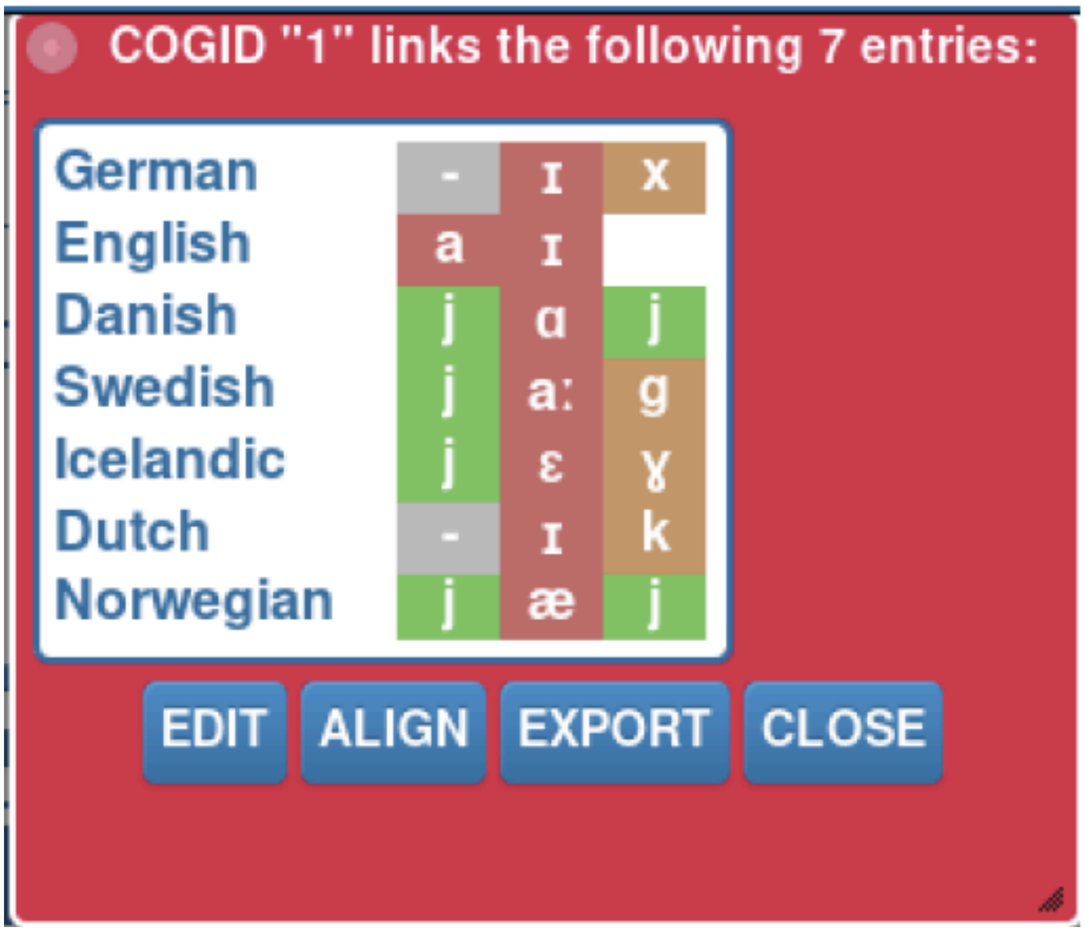
GER	ENG	Frequ.
d	θ	3 x
n	n	2 x
ŋ	ŋ	1 x

Eng\ Ger	d	n	ŋ
θ	3	0	0
n	0	2	0
ŋ	0	0	1

[List 2017]

Computer-Assisted Language Comparison

- EDICTOR and LingPy



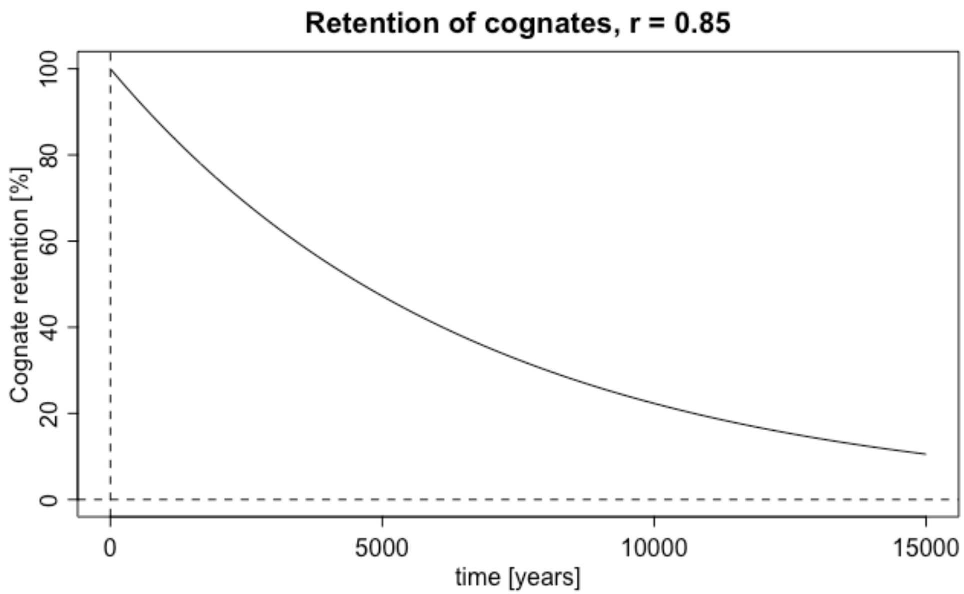
Lexicostatistics (a.k.a. Glottochronology)

- Wordlists of “basic” vocabulary
- Count shared cognates between language pairs (retention rate)
- Cluster languages with highest similarity

[Swadesh 1950, 1952, 1955]

History of glottochronology

- Loss of cognates happens at a constant rate (as inspired in radioactive decay: exponential)
- The rate of retention is about 80 to 85% (*i.e.* loss 15 to 20%) every 1000 years



[Swadesh 1950]

History of glottochronology

- Scholars were very excited in the first place with glottochrnonology (1960s)

In the last decade, glottochronology has excited international interest and acquired a literature of its own. To the antrhropologist it promises a measure of time depth for language families iwthout documented history, and yet another linguistic example of regularity in cultural phenomena

[Hymes 1960]

… 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]

History of glottochronology

- But this didn't last long:

- Relation between Old Norse and Icelandic :
  - According to glottochronology: 200 years
  - Historical records: 1000 years

On the Validity of Glottochronology

by Knut Bergsland and Hans Vogt [1962]

Our findings clearly disprove the basic assumption of glottochronology 'that fundamental vocabulary changes at a constant rate'

History of glottochronology

- And there is more…

A tradition of hostility towards probabilistic modelling in historical linguistics

[Sankoff 1973]

In summary, glottochronology is not accurate; all its basic assumptions have been severely criticized. It should not be accepted, it should be rejected

[Campbell 2004]

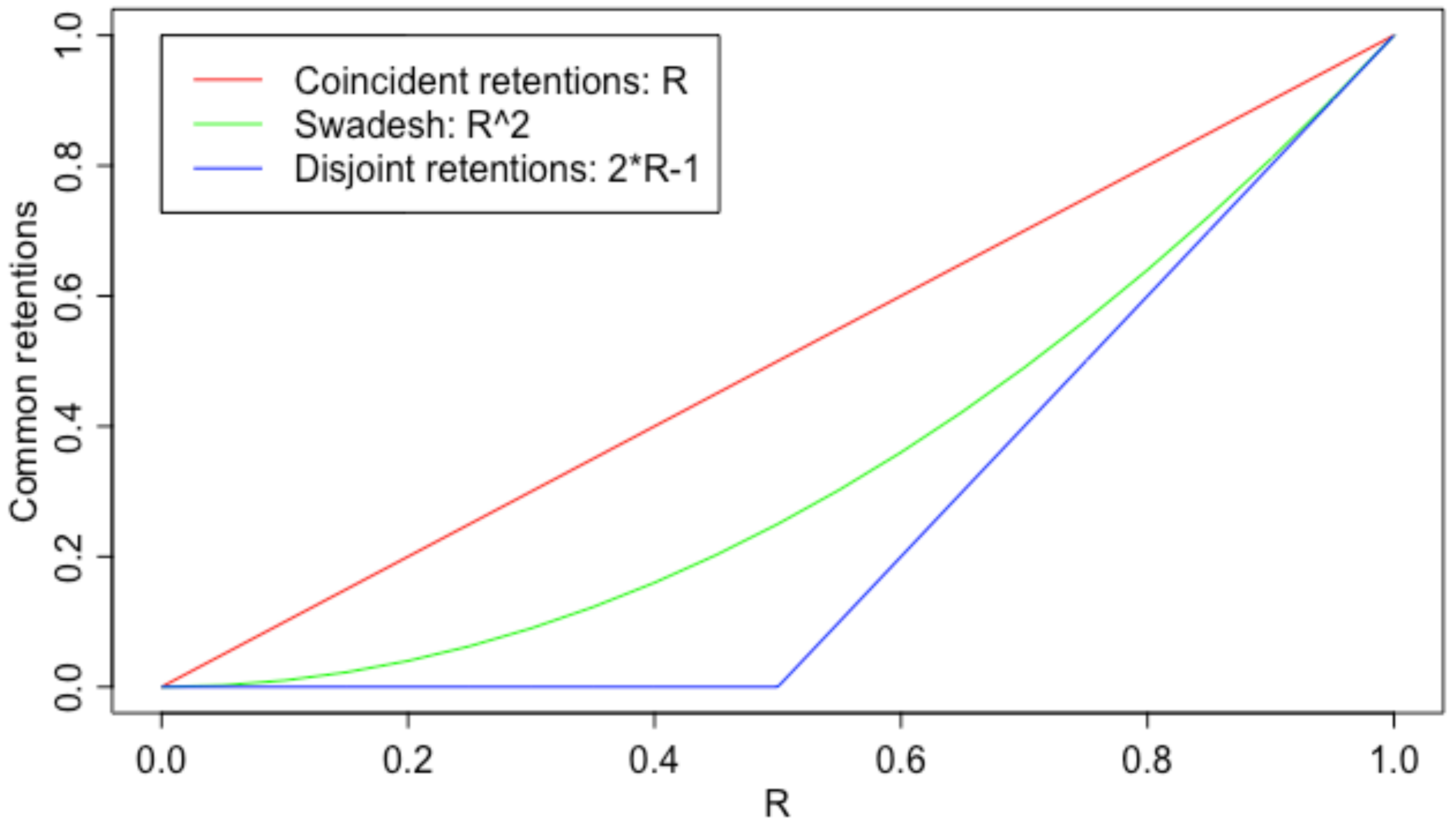
Linguists don't do dates

[McMahon & McMahon 2003]

The Swadesh List

- Morris ("Mauricio") Swadesh
- He started with 225 meanings of "basic vocabulary"
- Reduced to 165 for Salish languages [1950]
- Updated to 215 [1952]
- Last version of 200 [1955] "Swadesh 200"
- Last version of 100 [1971, posthumous] "Swadesh 100"
- Today, there are many "swadesh lists":
- http://concepticon.clld.org

The Swadesh List



Sound classes

No.	Cl.	Description	Examples
1	"P"	labial obstruents	p, b, f
2	"T"	dental obstruents	d, t, θ, ð
3	"S"	sibilants	s, z, ʃ, ʒ
4	"K"	velar obstruents, dental and alveolar affricates	k, g, ʦ, ʧ
5	"M"	labial nasal	m
6	"N"	remaining nasals	n, ɲ, ŋ
7	"R"	liquids	r, l
8	"W"	voiced labial fricative and initial rounded vowels	v, u
9	"J"	palatal approximant	j
10	"Ø"	laryngeals and initial velar nasal	h, ɦ, ŋ

Table 4.2: *Dolgopolsky's original sound class model*

[Dolgopolsky 1964]

[List 2012]

Sound classes

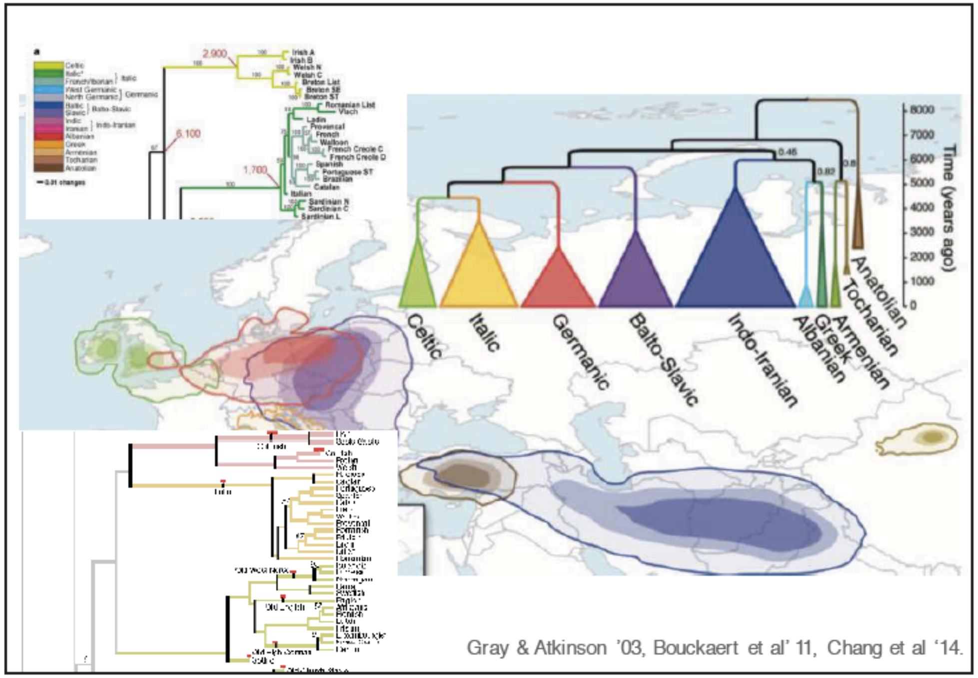
No.	Cl.	Description	Examples
1	"A"	unrounded back vowels	a, ɑ
2	"B"	labial fricatives	f, β
3	"C"	dental / alveolar affricates	ʦ, ʥ, ʧ, ʤ
4	"D"	dental fricatives	θ, ð
5	"E"	unrounded mid vowels	e, ɛ
6	"G"	velar and uvual fricatives	ɣ , x
7	"H"	laryngeals	h, ʔ
8	"I"	unrounded close vowels	i, ɪ
9	"J"	palatal approxoimant	j
10	"K"	velar and uvular plosives	k, g
11	"L"	lateral approximants	l
12	"M"	labial nasal	m
13	"N"	nasals	n, ŋ
14	"O"	rounded back vowels	Œ, ɒ
15	"P"	labial plosives	p, b
16	"R"	trills, taps, flaps	r
17	"S"	sibilant fricatives	s, z, ʃ, ʒ
18	"T"	dental / alveolar plosives	t, d
19	"U"	rounded mid vowels	ɔ , o
20	"W"	labial approx. / fricative	v, w
21	"Y"	rounded front vowels	u, ʊ, y
22	"0"	low even tones	11, 22
23	"1"	rising tones	13, 35
24	"2"	falling tones	51, 53
25	"3"	mid even tones	33
26	"4"	high even tones	44, 55
27	"5"	short tones	1, 2
28	"6"	complex tones	214

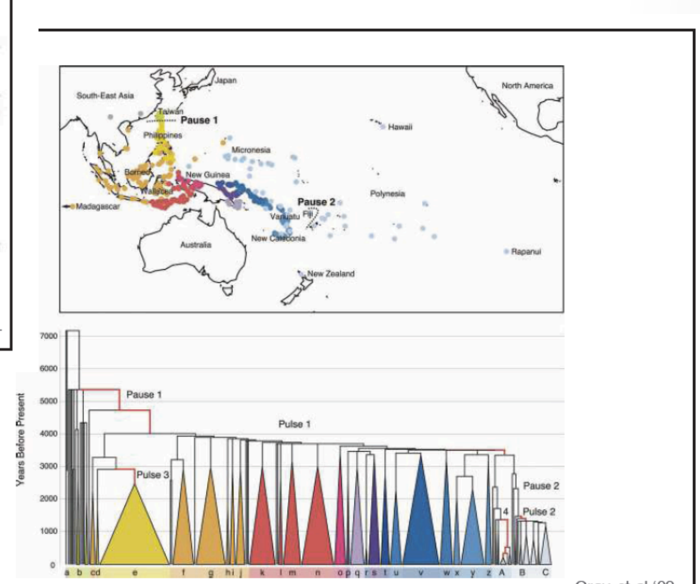
Table 4.3: *The SCA sound class model*

[List 2012]

Phylogenetics

- BEAST + Figtree





Inferring linguistic phylogenies

	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

```
#NEXUS

BEGIN CHARACTERS;
    concept_1=1-1; [taboo]
    concept_2=2-3; [blood]
    concept_3=4-6; [to suck]
END CHARACTERS;

BEGIN DATA;

DIMENSIONS NTAX=5 NCHAR=6;

FORMAT DTATYPE=STANDARD SYMBOLS="10" GAP=- MISSING=? INTERLEAVE= YES

MATRIX
Fijian      110100
Tahitian    101010
Maori       101010
Hawaiian    101001
Marquesan   101001

END;
```

	meaning 1	meaning 2		meaning 3		
Meaning	taboo	blood		to suck		
Cognate set	*tabu*	*dra*	*toto*	*sucu-ma*	*ngote*	*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

[Greenhill 2017]

Inferring linguistic phylogenies

- Ideally, **proven cognates** should be used
- In cases in which a proper cognate judgment can't be carried out, **cognate candidates** might be used as well, although this adds a further unquantified level of uncertainty.

- **Distance-based models of change**
  - Aggregate amount of difference between two languages.
  - Some kind of *distance metric* defined.
- **Character-based models of change**
  - Infers the plausible pathways by which each language evolved from their common ancestor
  - It is the shortest path between the languages
  - It is always equal or greater than a distance model for the same pair of languages
  - Are more realistic than the former

[Dunn 2013]

Distance-based models of change

1. Levenshtein distance (edit distance)

- Alignment analysis has two steps:
  - 1. Identify corresponding segments
  - 2. Introduce gaps for non-corresponding segments.
- Brute-force algorithm
  - Build all possible alignments between the two sequences
  - Define a *scoring scheme* to determine the similarity between the different correspondences (exact match, partial match, gap, mismatch)
  - Sum all individual segment scores to obtain the alignment score
  - Compare the alignment scores of each possible alignment
  - One such score is called Levenshtein distance or edit distance

  [V. I. Levenshtein 1965]

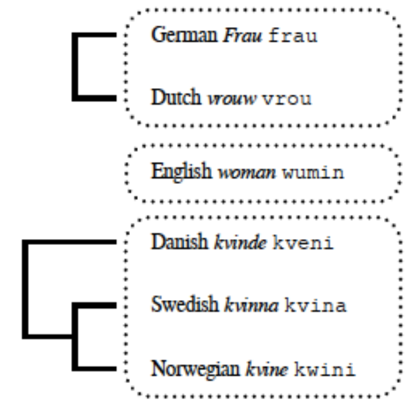
[List 2012]

Distance-based models of change

1. Levenshtein distance (edit distance)

ID	Taxa	Word	Gloss	GlossID	IPA
...	...	...	...	...	...
21	German	Frau	woman	20	frau
22	Dutch	vrouw	woman	20	vrɑu
23	English	woman	woman	20	wʊmən
24	Danish	kvinde	woman	20	kvenə
25	Swedish	kvinna	woman	20	kvi:na
26	Norwegian	kvine	woman	20	kvinə
...	...	...	...	...	...

	Swedish *kvinna*	English *woman*	Danish *kvinde*	Norwegian *kvine*	Dutch *vrouw*	German *Frau*
Swedish kvina	0.00	0.69	0.07	0.12	0.71	0.78
English wumin	0.69	0.00	0.66	0.57	0.68	0.87
Danish kveni	0.07	0.66	0.00	0.08	0.67	0.71
Norwegian kwini	0.12	0.57	0.08	0.00	0.75	0.74
Dutch frou	0.71	0.68	0.67	0.75	0.00	0.17
German frau	0.78	0.87	0.71	0.74	0.17	0.00



ID	Taxa	Word	Gloss	GlossID	IPA	CogID
...	...	...	...	...	...	...
21	German	Frau	woman	20	frau	1
22	Dutch	vrouw	woman	20	vrɑu	1
23	English	woman	woman	20	wʊmən	2
24	Danish	kvinde	woman	20	kvenə	3
25	Swedish	kvinna	woman	20	kvi:na	3
26	Norwegian	kvine	woman	20	kvinə	3
...	...	...	...	...	...	...

[List 2014]

Distance-based models of change

1. Levenshtein distance (edit distance)

Problems with Levenshtein distance:

- it is only a coherent measure of language change where the forms being compared are cognates
- Useful for dialectometry:
  - Most forms for a meaning are cognate
  - The variation rates are similar

Levenshtein Distances Fail to Identify Language Relationships Accurately

Simon J. Greenhill*
The University of Auckland

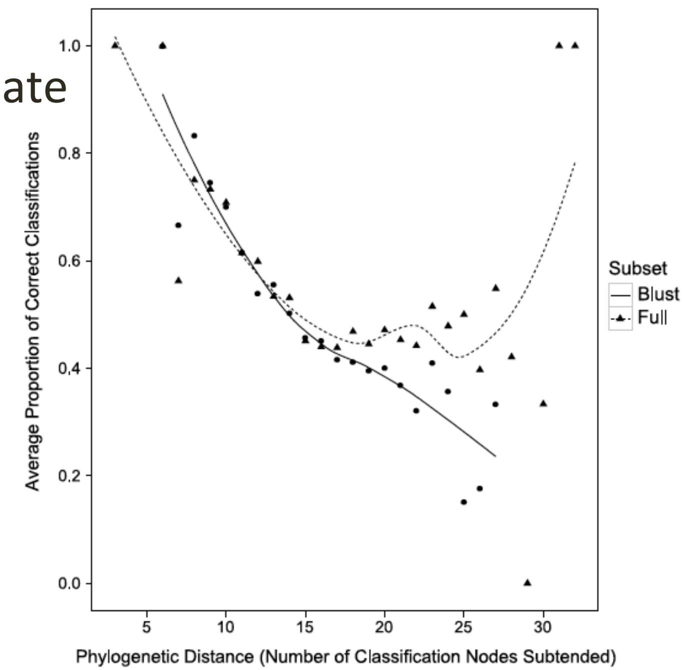


Figure 1
Scatter plot showing the accuracy of the Levenshtein classification approach as a function of phylogenetic distance. Phylogenetic distance is measured by the average number of Ethnologue classification nodes subtended by each language triplet. The points are drawn from the two language subsets spanning the largest range of subgroups (the full data set and the Blust subsample) with LOESS curves of best fit (Full data set: triangles, dotted line; Blust data set: circles, line).

[Dunn 2013 ; Greenhill 2011]

Distance-based models of change

2. Lexicostatistics (distance = cognate proportion)

- *d =* proportion of cognates which are NOT cognates
- We assume a constant rate of change (*à la* Swadesh 1950,1952,1955: retention rate around *r* = 0.81 per 1,000 years): *Glottochronology*
- This approach failed, as we saw before
- Distance-based clustering is extremely sensitive to differences in rate of change in different branches of the tree

[Dunn 2013 ; Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

- The parsimony method seeks a tree that explains a data set (*e.g.* a set of cognate judgments) by minimizing the number of evolutionary changes required to produce the observed states.

[Dunn 2013]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

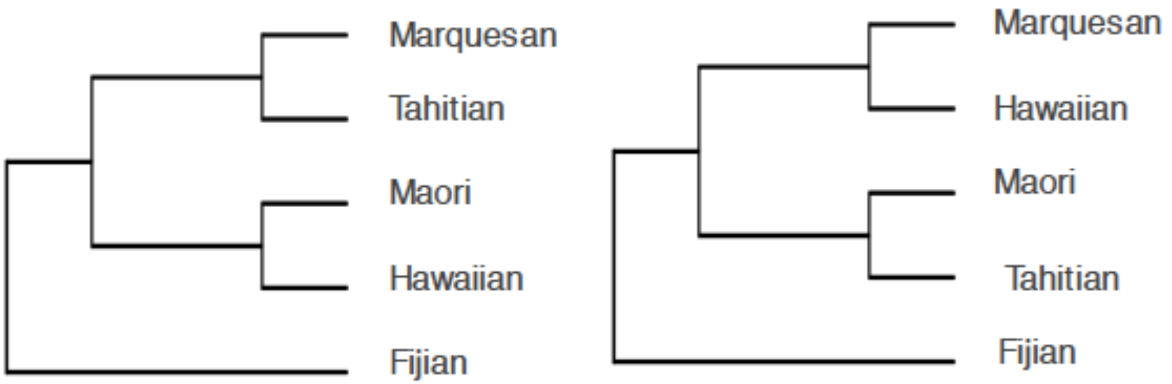
	Taboo	Blood	To Suck
Fijian	tabu	drā	sucu-ma
Tahitian	tapu	toto	ngote
Maori	tapu	toto	ngote
Hawaiian	kapu	koko	omo
Marquesan	tapu	toto	omo

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

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

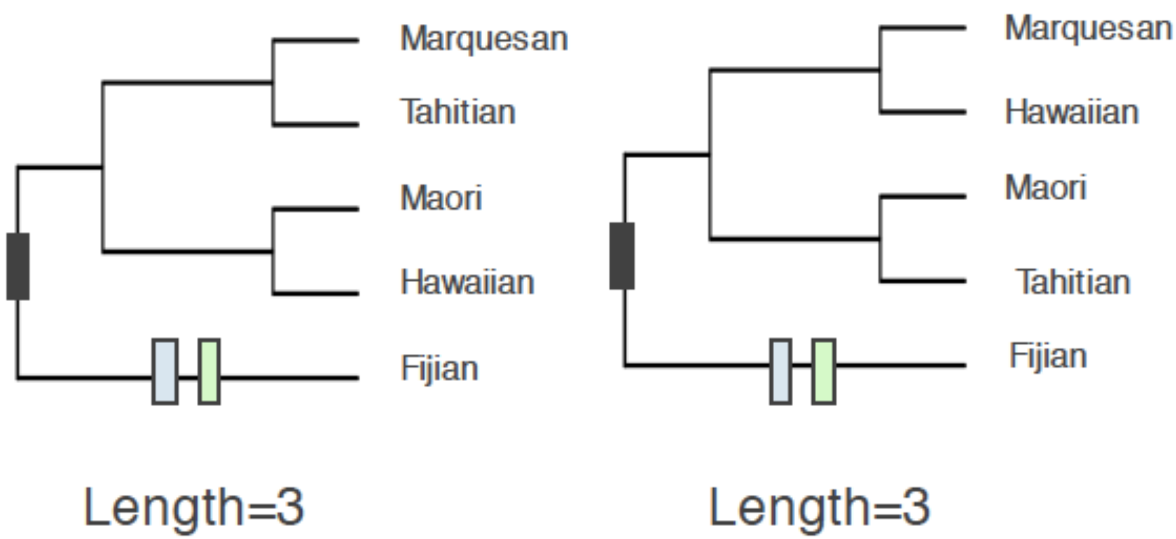


Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

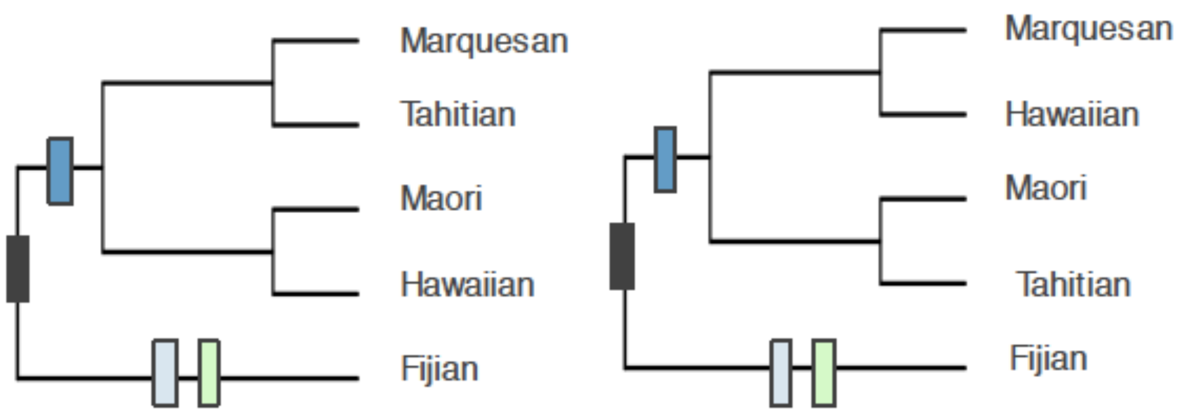


Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)



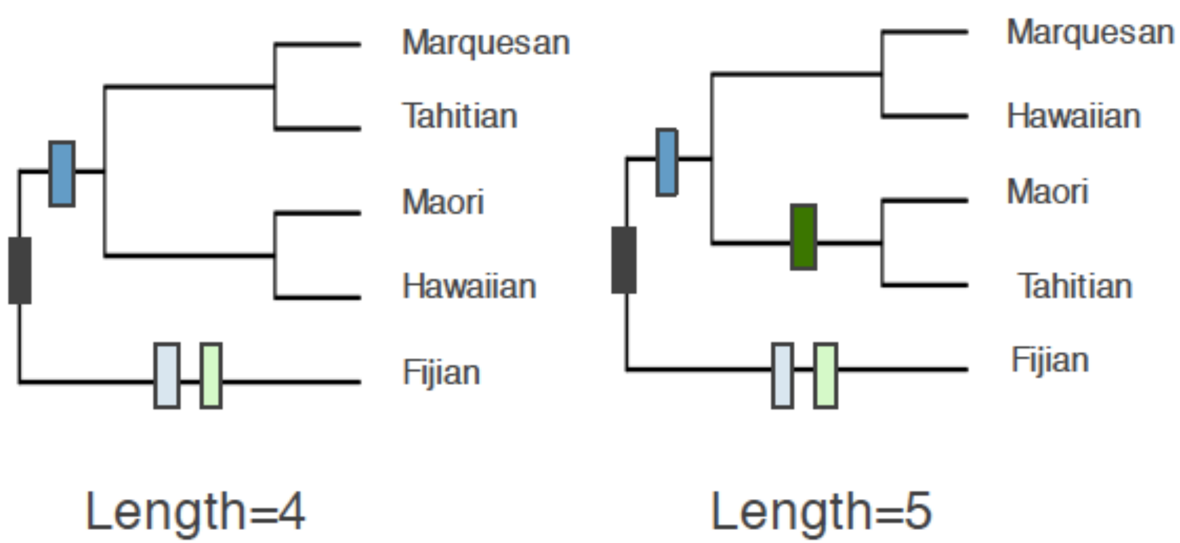
Length=4 Length=4

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

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

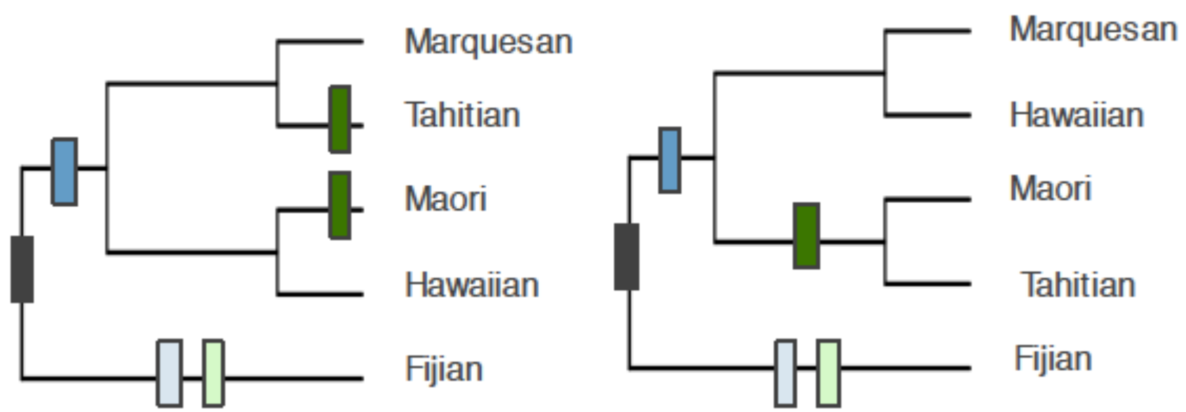


Fijian	1	1	0	1	0	0
Tahitian	1	0	1	0	1	0
Maori	1	0	1	0	1	0
Hawaiian	1	0	1	0	0	1
Marquesan	1	0	1	0	0	1

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)



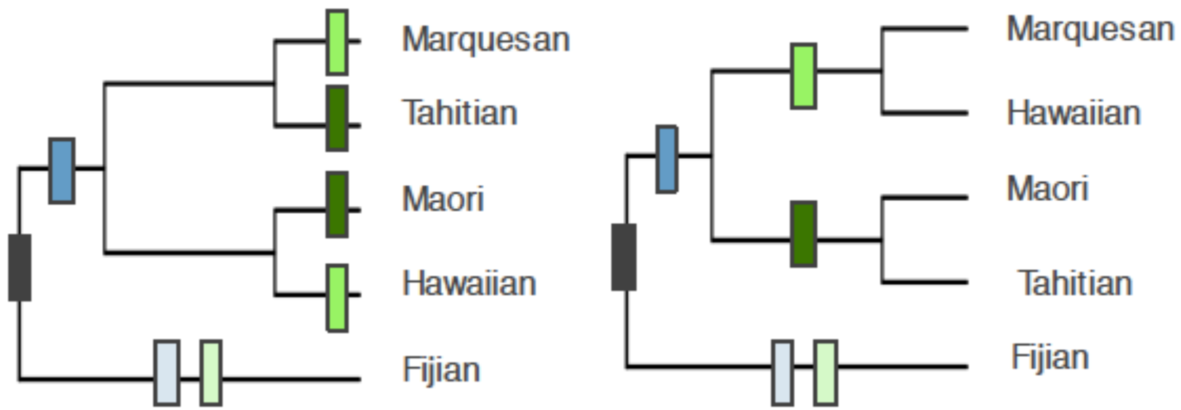
Length=6 Length=5

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

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)



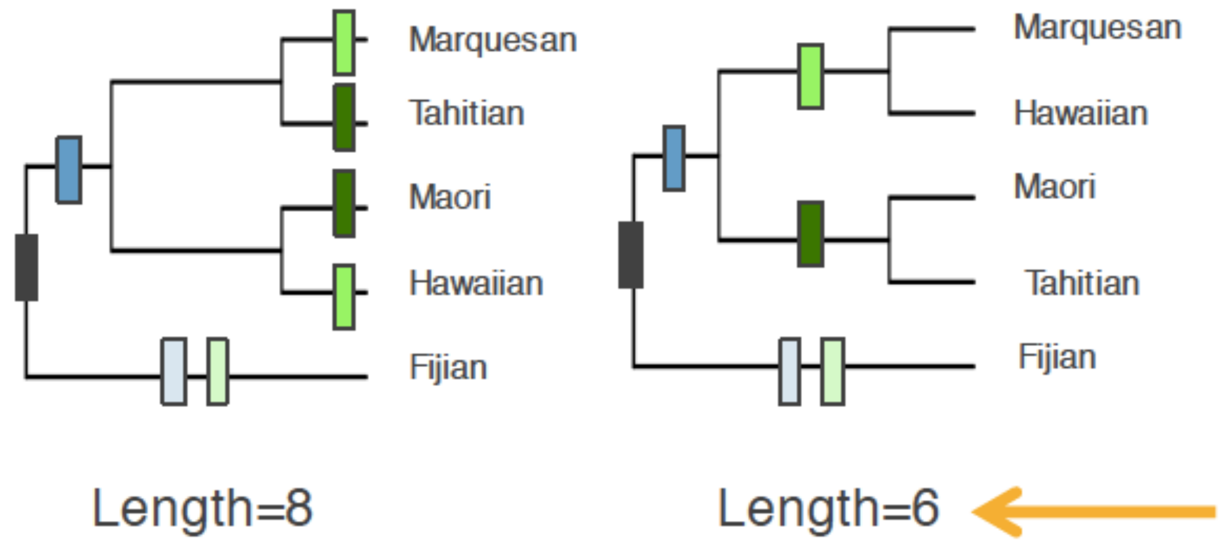
Length=8 Length=6

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

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)



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

Figures stolen from[Greenhill 2017]

Character-based models of change

1. Maximum parsimony (Ockham's razor)

- The parsimony method seeks a tree that explains a data set (e.g. a set of cognate judgments) by minimizing the number of evolutionary changes required to produce the observed states.
- **Problem:** *Long branch attraction:*
  - Long branches (much change) will tend to be clustered together even if they are only distantly related in the true evolutionary history.
  - When two branches have both undergone a lot of change, the most parsimonious ("cheap") account is always to bundle the two branches together as a single set of innovations at the end.

[Dunn 2013]

Character-based models of change

2. Maximum Likelihood

- Explain a set of observed data by quantifying how likely it was to have been produced by a particular process.
- The likelihood is the probability of seeing the observed data under a particular hypothetical mechanism, $L = P(D|H)$.
- Within phylogenetics, the hypothesised mechanism is an evolutionary process, "the model", which consists of a mathematical description of evolutionary change.
- A model includes tree topology, branch lengths, the probability that a new cognate set appears in the tree, the probability that a reflex of a cognate set is lost, etc.
- Given a topology, maximizing the likelihood of the other parameters of a tree is generally tractable to exact mathematical methods.
- However, finding the best tree topology out of the vast space of possible trees is extremely challenging: It is not possible (by now) to solve this using random sampling of tree likelihoods.

[Dunn 2013]

Character-based models of change

2. Maximum Likelihood

- However, finding the best tree topology out of the vast space of possible trees is extremely challenging: It is not possible (by now) to solve this using random sampling of tree likelihoods.

Table 2.2 The number of unlabelled rooted tree shapes, the number of labelled rooted trees, the number of labelled ranked trees (on contemporaneous tips) and the number of fully ranked trees (on distinctly timed tips) as a function of the number of taxa, n

n	#shapes	#trees, $\lvert\mathcal{T}_n\rvert$	#ranked trees, $\lvert\mathcal{R}_n\rvert$	#fully ranked trees, $\lvert\mathcal{F}_n\rvert$
2	1	1	1	1
3	1	3	3	4
4	2	15	18	34
5	3	105	180	496
6	6	945	2700	11 056
7	11	10 395	56 700	349 504
8	23	135 135	1 587 600	14 873 104
9	46	2 027 025	57 153 600	819 786 496
10	98	34 459 425	2 571 912 000	56 814 228 736

[Drummond and Bouckaert 2015]

Character-based models of change

2. Maximum Likelihood

- However, finding the best tree to … of possible trees is extremely challen… now) to solve this using random s…

Table 2.2 The number of unlabelled rooted tree shapes, the nu… trees, the number of labelled ranked trees (on contemporaneou… fully ranked trees (on distinctly timed tips) as a function of the r…

n	#shapes	#trees, $\lvert\mathcal{T}_n\rvert$	#ranked trees, $\lvert\mathcal{R}_n\rvert$	#f…
2	1	1	1	
3	1	3	3	
4	2	15	18	
5	3	105	180	
6	6	945	2700	
7	11	10 395	56 700	
8	23	135 135	1 587 600	
9	46	2 027 025	57 153 600	
10	98	34 459 425	2 571 912 000	

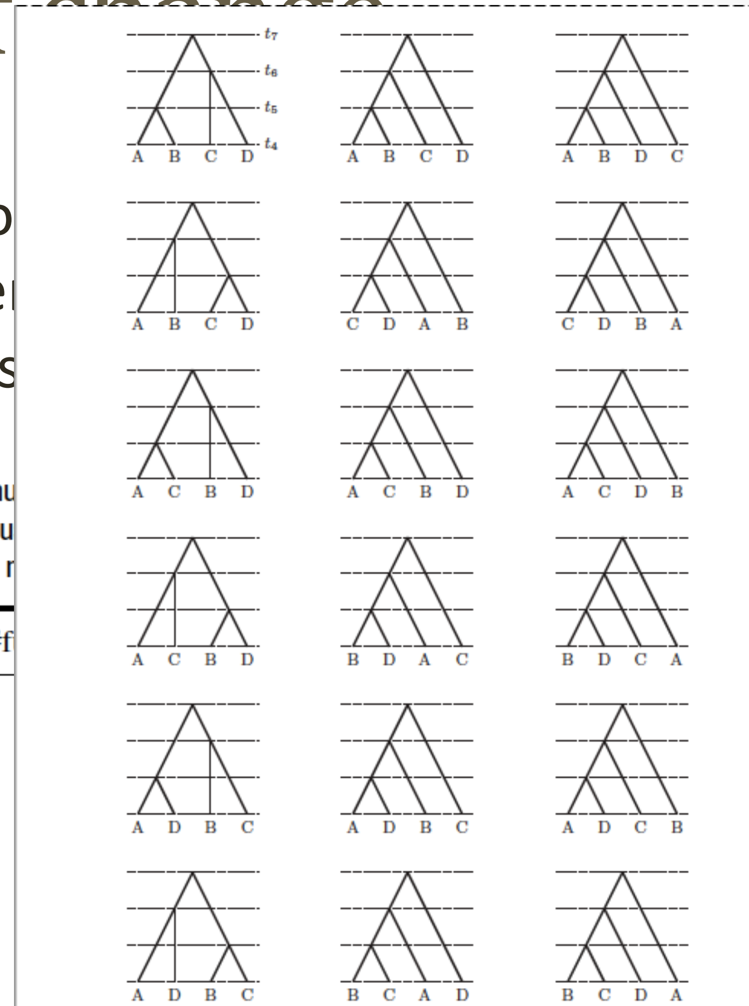
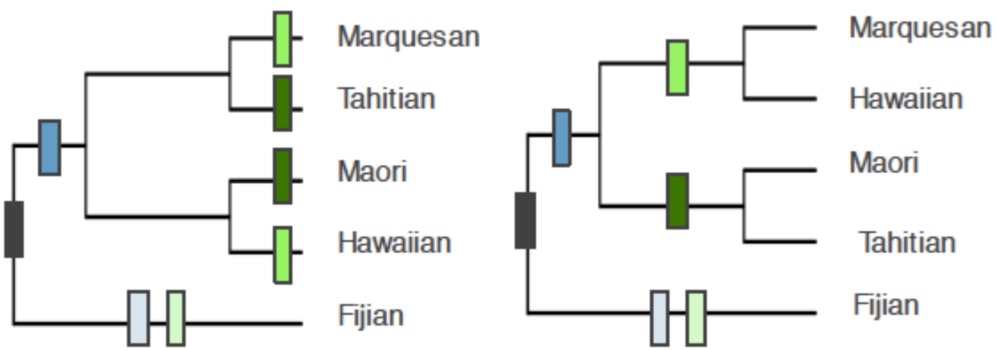


Figure 2.3 All ranked trees of size 4.

[Drummond and Bouckaert 2015]

Character-based models of change

2. Maximum Likelihood



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

[Greenhill 2017]

Character-based models of change

3. Evolutionary models (Bayesian phylogenetic analysis)

A. *Clock model:* rate of change (not fixed as in glottochronology!)
   - Strict clock: constant rate of change
   - Relaxed clock: allows rates to vary across the tree, chosen from a probability distribution.
   - Different kinds of probability distribution can model processes where rate change occurs continuously along a branch, or where rates change at nodes independently of branch length
   - The clock (either strict or relaxed) at any node, is the same for all cognate sets.

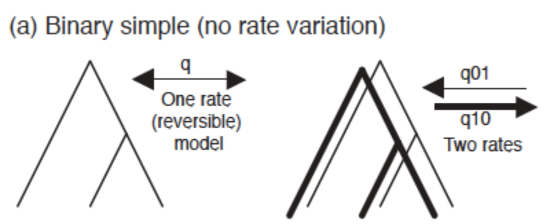
[Dunn 2013]

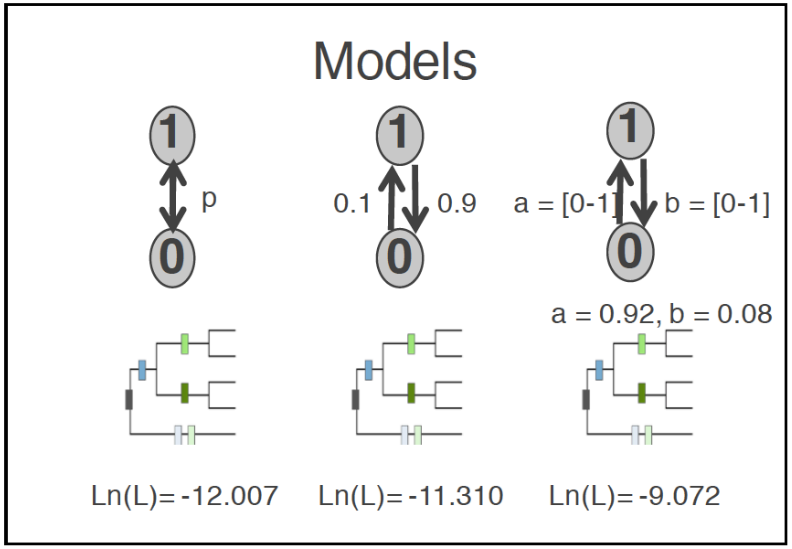
Character-based models of change

3. Evolutionary models (Bayesian phylogenetic analysis)

B. Substitution model: Specifies how rates differ among characters (*i.e.* among cognate sets).

- *Binary model*
  - *One-rate*
  - *Two-rates*





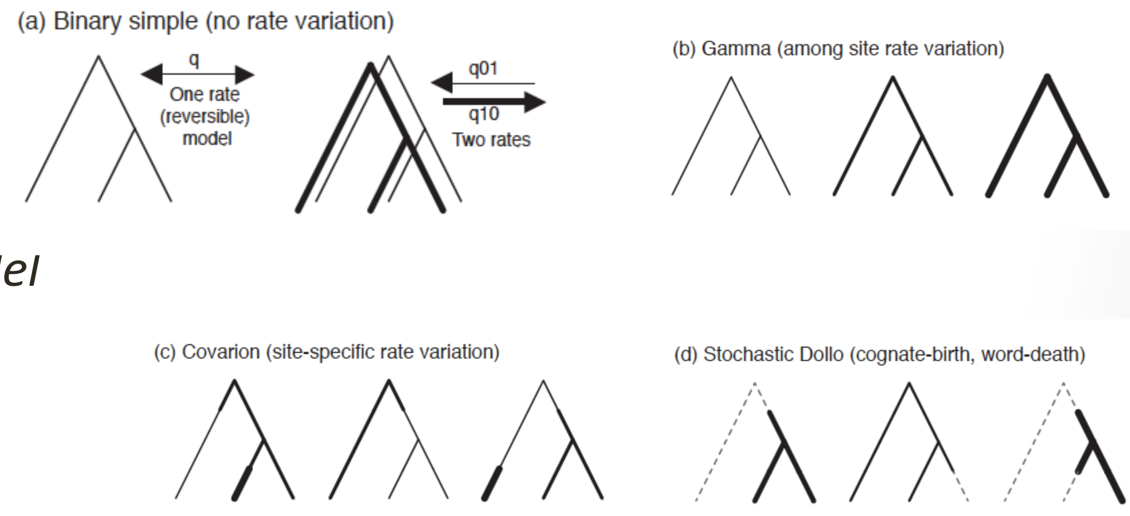
[Dunn 2013; Greenhill 2017]

Character-based models of change

3. Evolutionary models (Bayesian phylogenetic analysis)

B. *Substitution model:* Specifies how rates differ among characters (*i.e.* among cognate sets).

- *Binary model*
    - *One-rate*
    - *Two-rates*
- *Gamma model*
- *Covarion model*
- *Stochastic Dollo model*



[Dunn 2013]

Model selection

1. Likelihood and Bayes Factor

- The *likelihood score* of an analysis is the probability for the observed data to evolve given a particular model.
- Even assuming that for each model the optimal parameter values have been inferred, some models still fit better than others.
- The difference in acceptability of two models can be expressed by the *Bayes Factor:*

$$BF_{12} = \frac{L(H_1)}{L(H_2)}$$

- It is usually expressed as twice its natural logarithm:

$$2*log(BF_{12}) = 2*(log(L(H_1)) - log(L(H_2)))$$

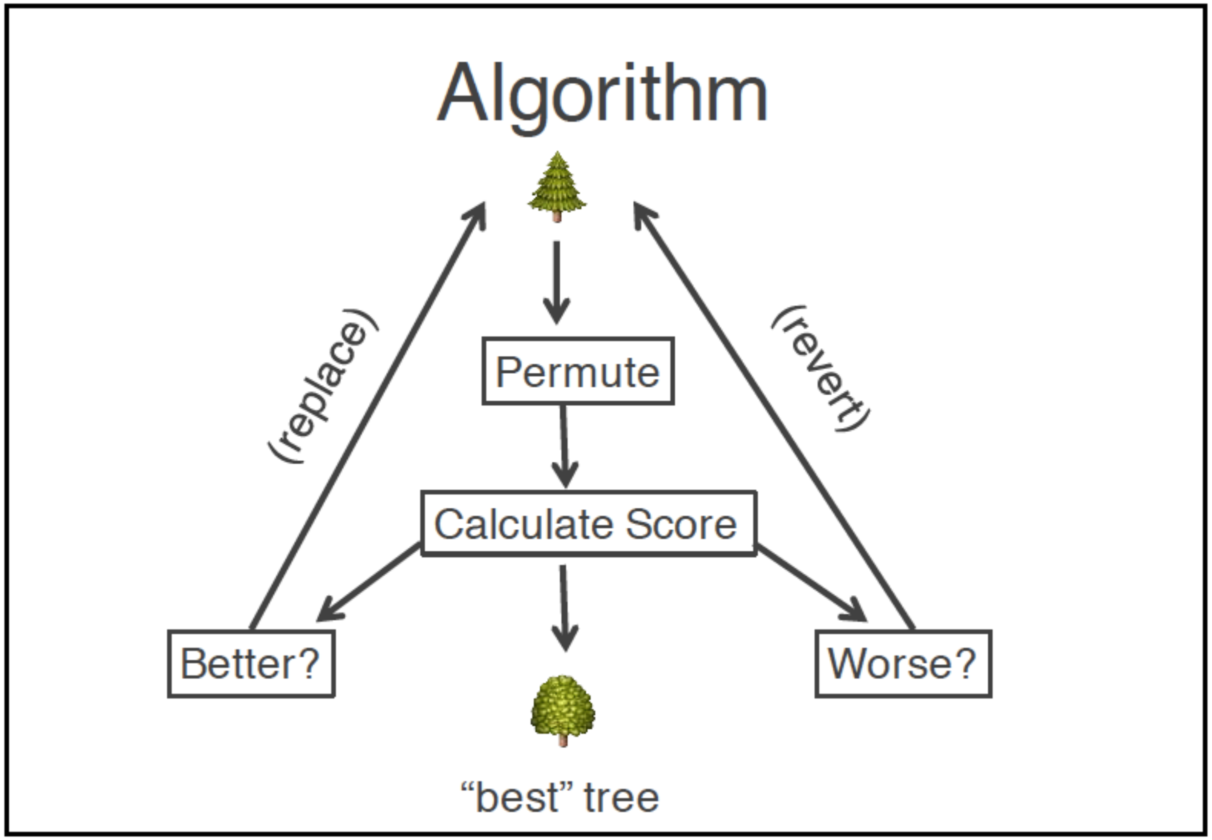
BF_{12}	$2logBF_{12}$	Evidence for H_1 over H_2
0 to 2	1 to 2	Negligible
3 to 20	2 to 6	Positive
20 to 50	6 to 10	Strong
>150	>10	Very strong

Table 4: Guidelines for the interpretation of Bayes Factors and Log Bayes Factors (after Kass and Raftery 1995: 777)

[Dunn 2013]

Model selection

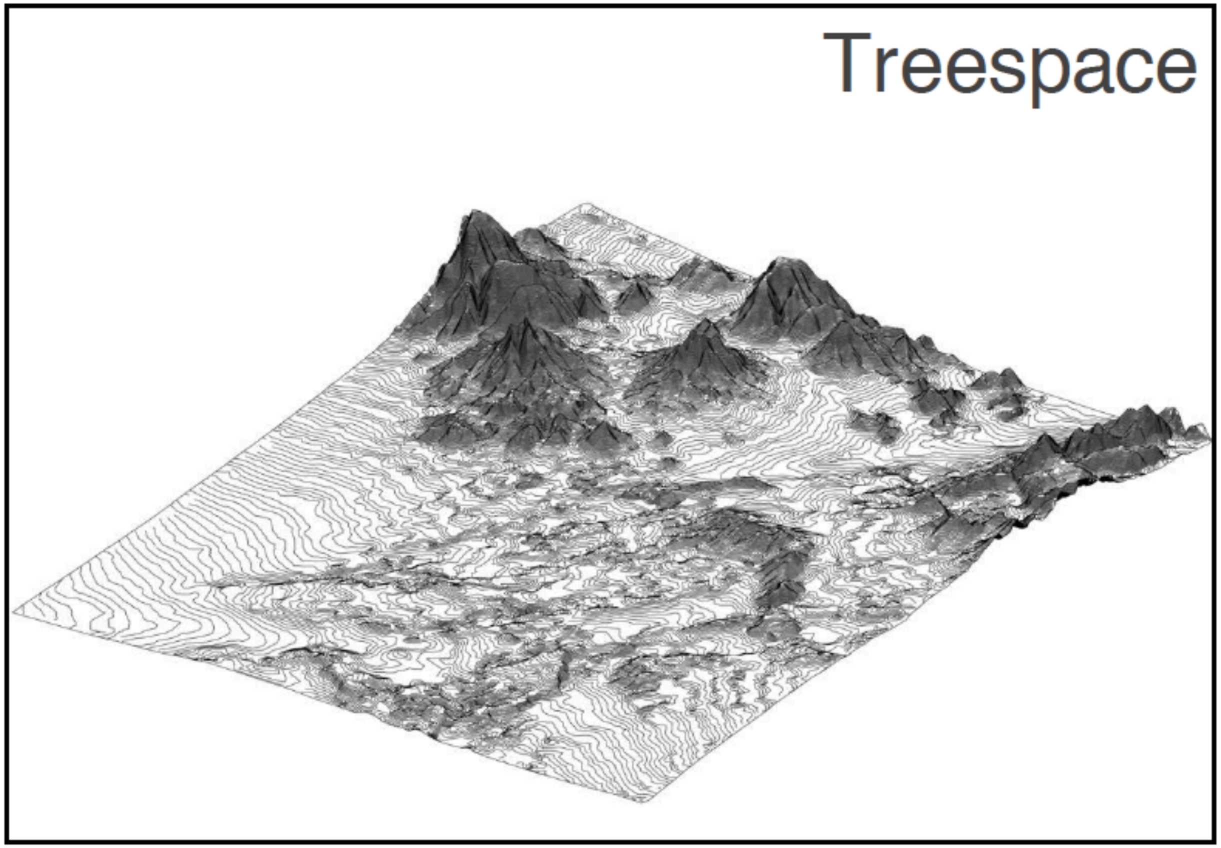
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

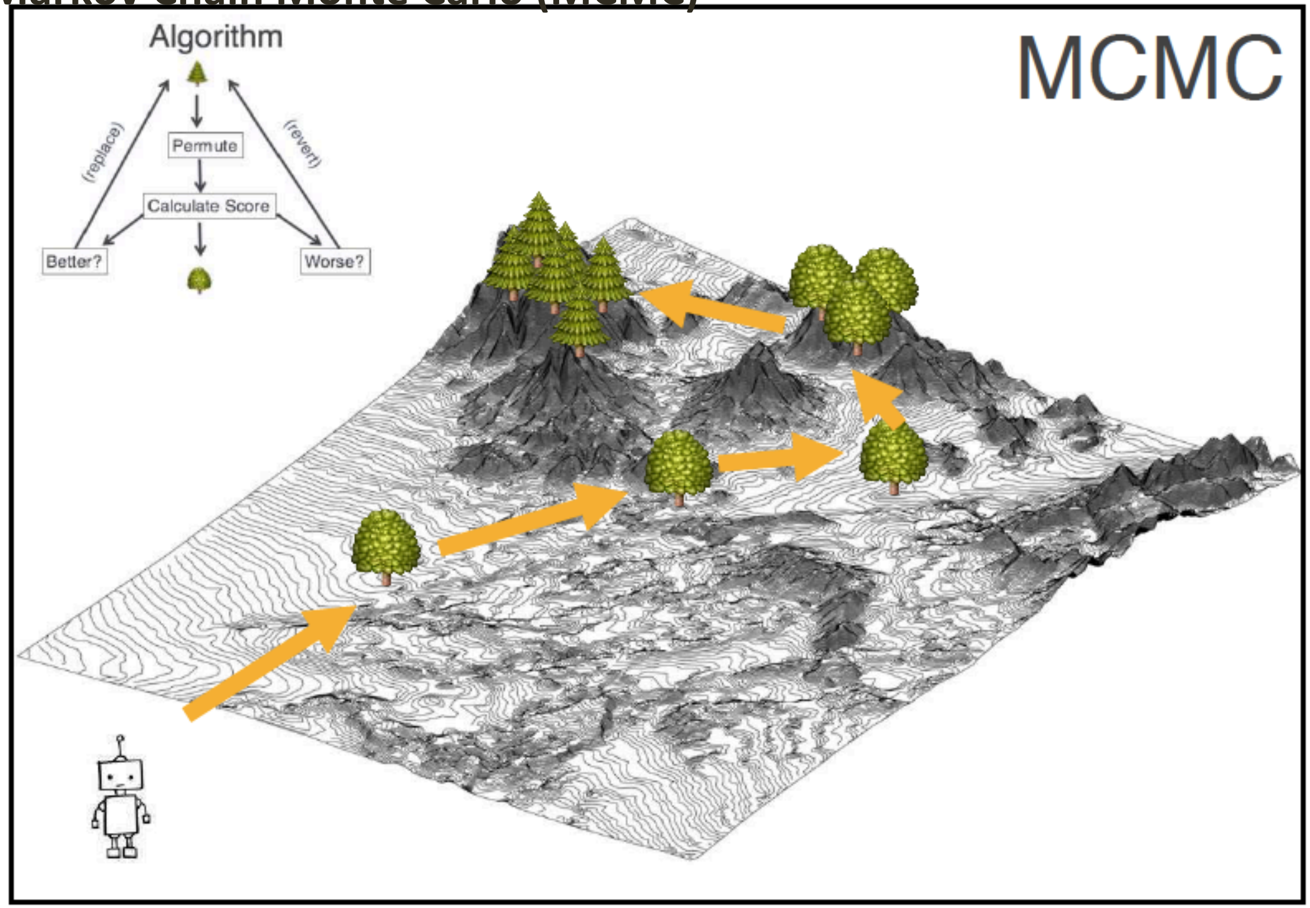
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

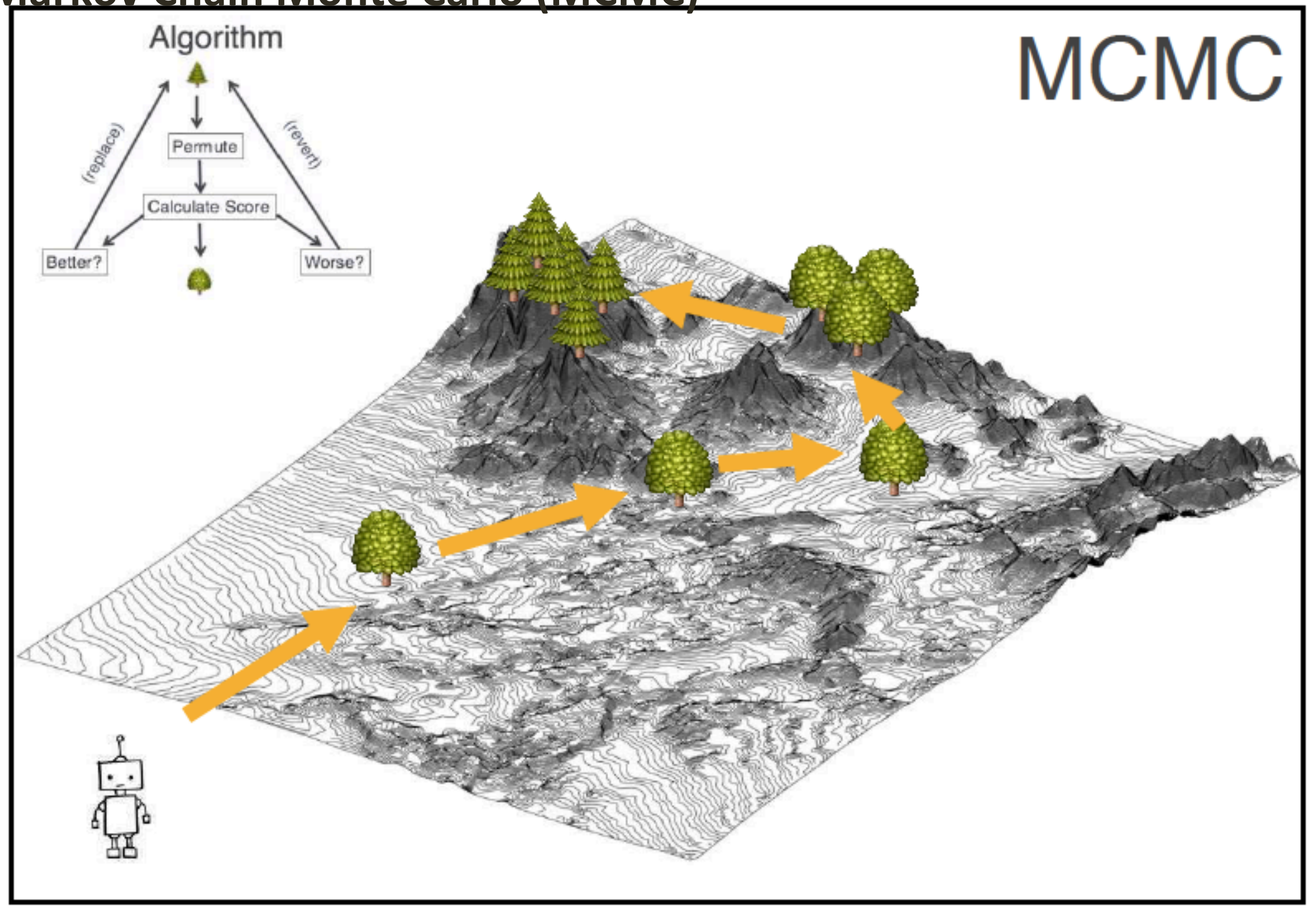
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

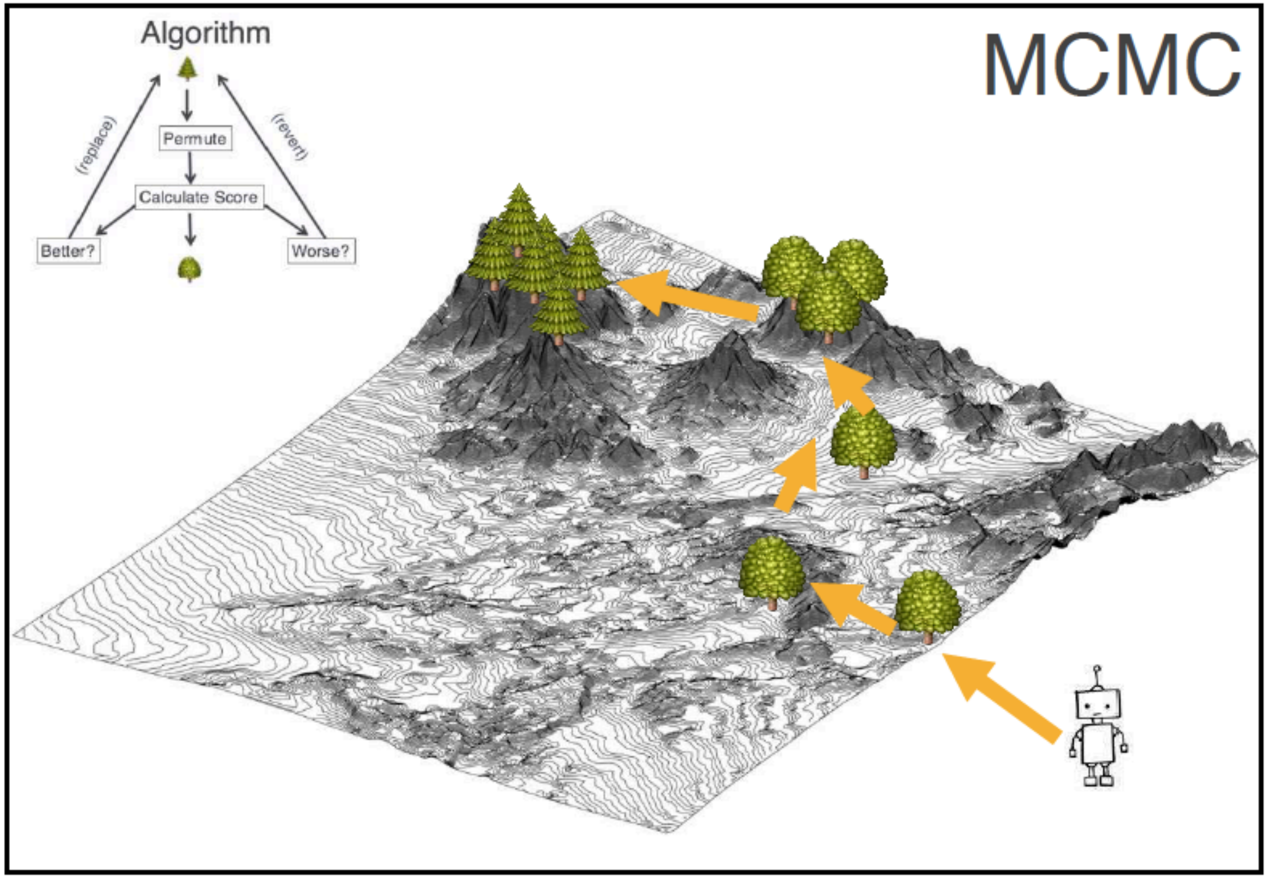
2. **Markov Chain Monte Carlo (MCMC)**



Figures stolen from [Greenhill 2017]

Model selection

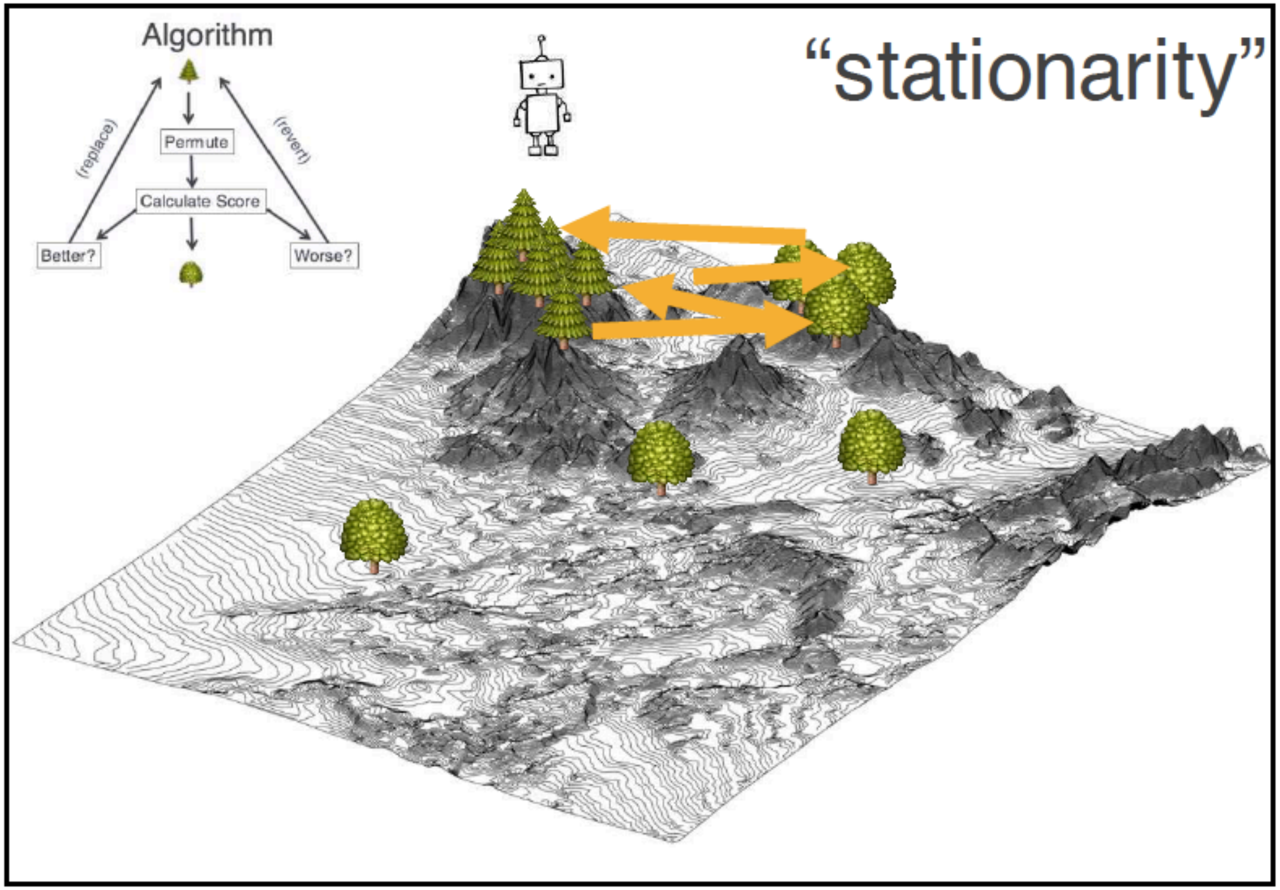
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

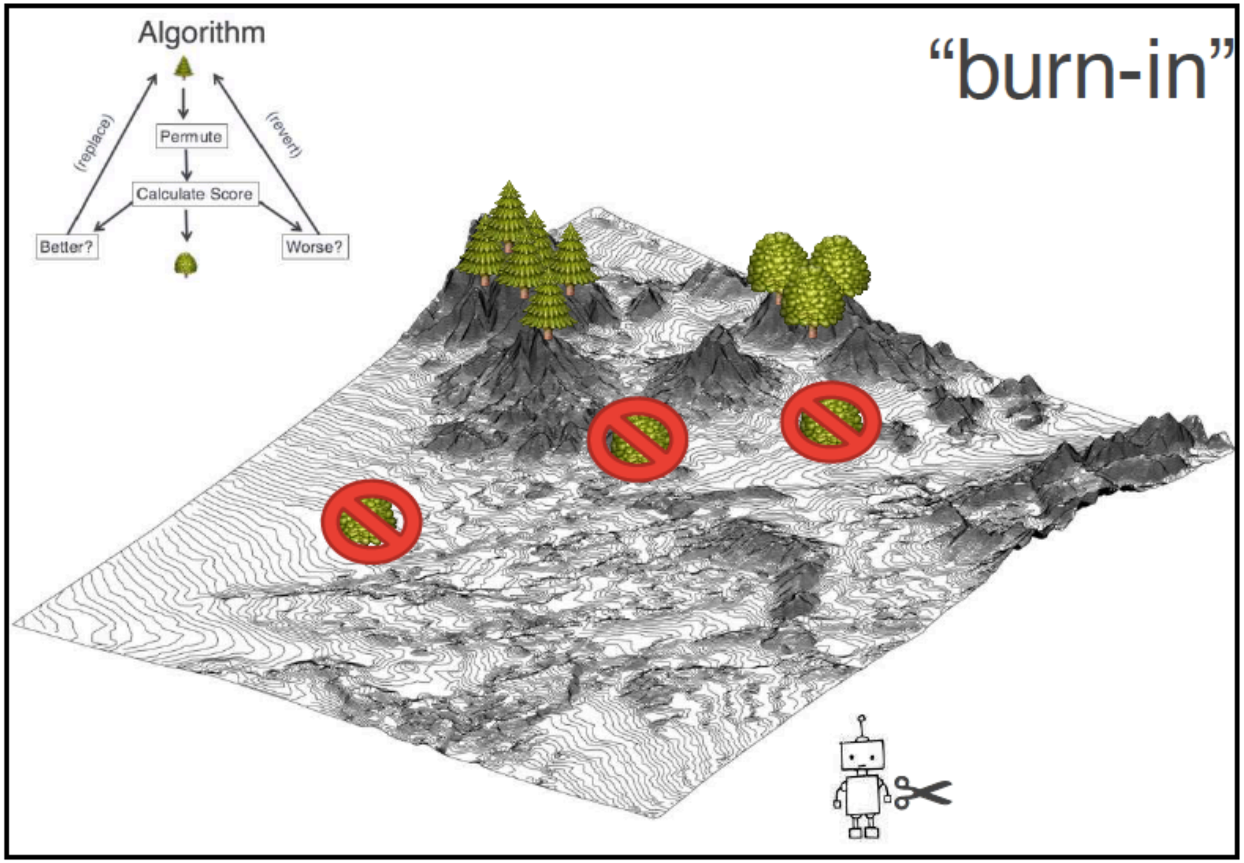
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

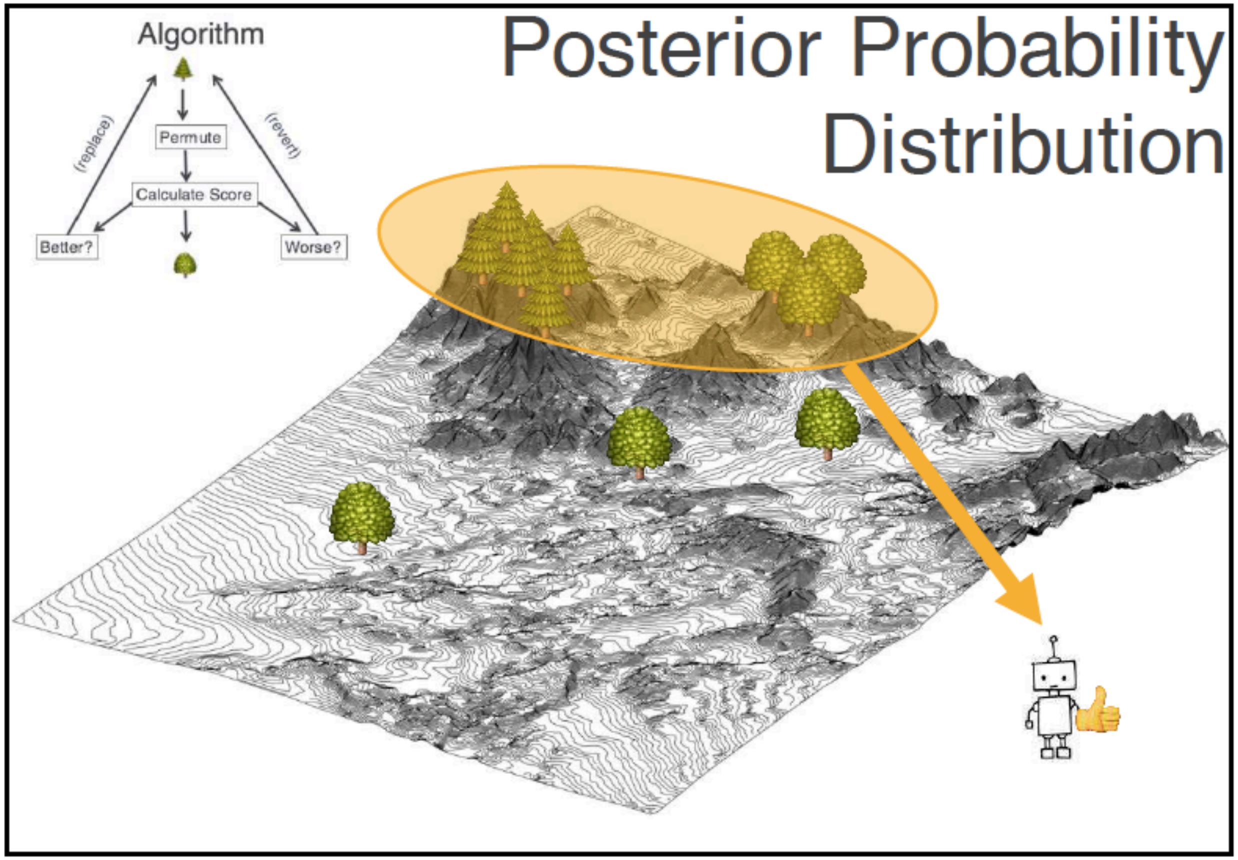
2. Markov Chain Monte Carlo (MCMC)



Figures stolen from [Greenhill 2017]

Model selection

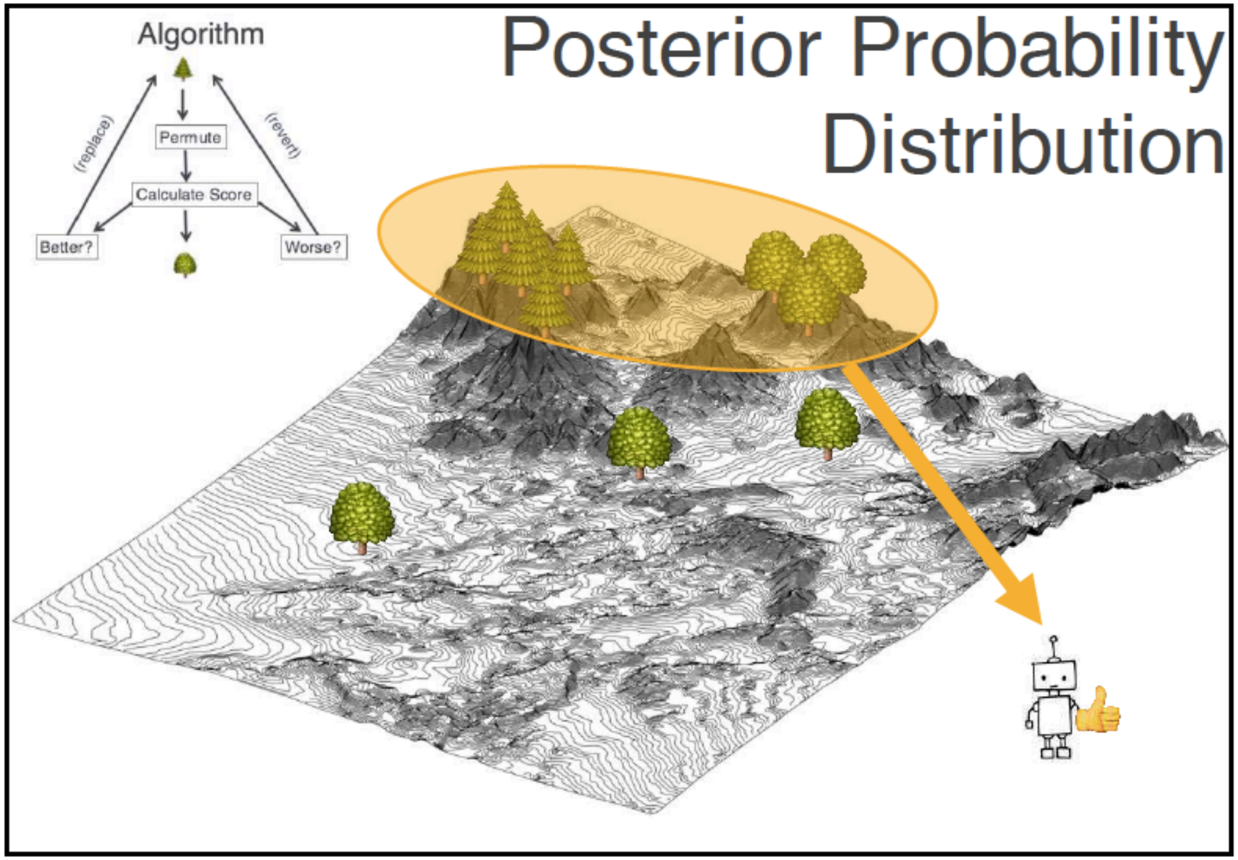
2. **Markov Chain Monte Carlo (MCMC)**



Figures stolen from [Greenhill 2017]

Model selection

2. **Markov Chain Monte Carlo (MCMC)**



Figures stolen from [Greenhill 2017]

Model selection

3. Summary tree

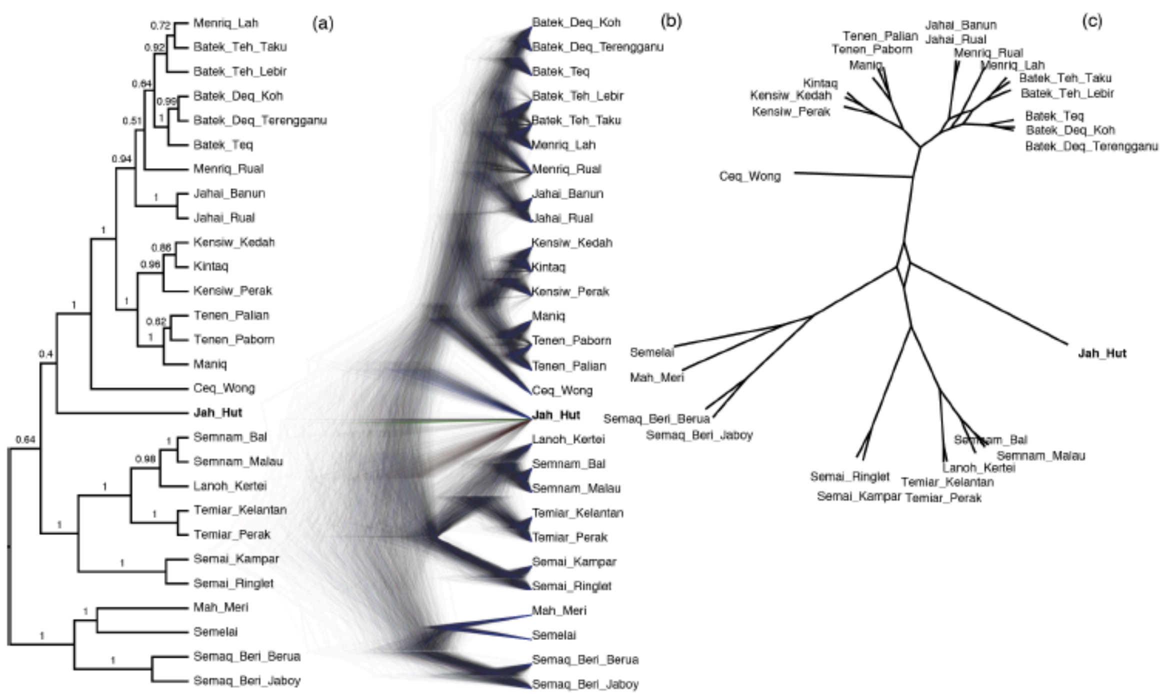


Figure 5: Summarizing the posterior tree sample; Aslian phylogenies (Dunn, Burenhult, et al. 2011) visualized with (a) Maximum Clade Credibility (MCC) Tree, (b) DensiTree, and (c) Consensus Network. Note how the uncertainty about the classification of the language "Jah Hut" is reflected by (a) low posterior probability values, (b) multiple points of origin, and (c) a box showing conflicting splits.

[Dunn 2013]

Model selection

4. Priors

- Bayes' Theorem:

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

[Greenhill 2017]

Model selection

4. Priors

- Bayes' Theorem:

$$\overbrace{\Pr(H|D)}^{\substack{\text{Posterior}\\\text{(Pr hypothesis given data)}}} = \frac{\overbrace{\Pr(H)}^{\text{prior}}\,\overbrace{\Pr(D|H)}^{\text{Lh}}}{\underbrace{\Pr(D)}_{\text{Prior Pr that data is true}}}$$

[Greenhill 2017]

Model selection

4. Priors

- Different kinds of priors can inform the tree:
  - Distributional priors on model parameters (clock and substitution model)
  - Elements of the tree structure (we restrict the search to these trees only):
    - Integrate subgrouping knowledge from classical linguistic comparative method (*e.g.* phonological and morphological innovations)
    - Integrate calibration points (*i.e.* date the documented nodes)
    - In more advanced analyses, geographical priors can be added (phylogeographic models)
- This allows us to generate trees that:
  - Go beyond the subgroupings provided by sound changes only
  - Comparative method trees with meaningful branch lengths and chronological calibration
  - Comparative method trees with quantified estimates of uncertainty and rate change

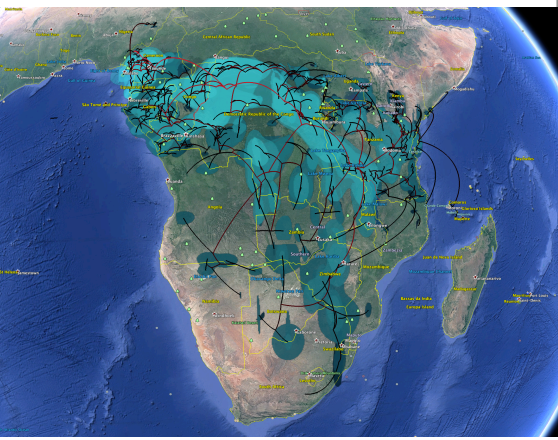
[Dunn 2013]

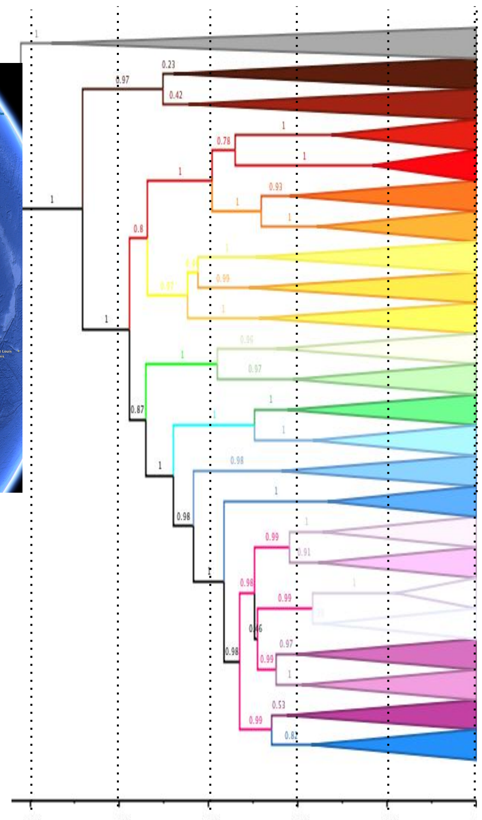
More advanced stuff

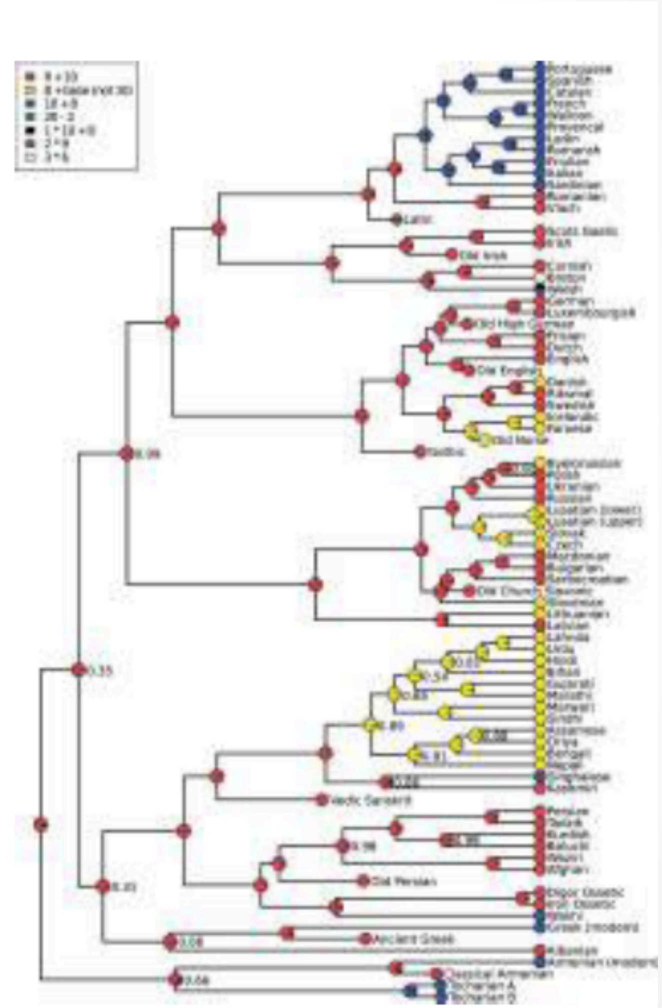
- Add phonetics
- Add morphology
- Add geography (phylogeography): priors can be plugged into the model
- (Try to) study the effects of borrowings and reticulation (family-internal borrowing)

Some applications

- Character evolution
- Phylogeography

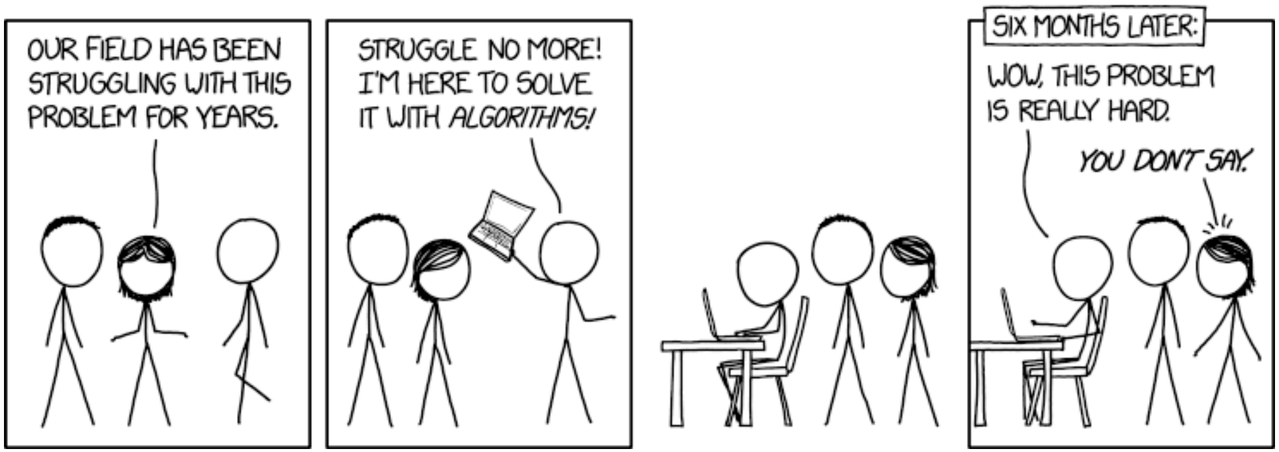






Calude and Verkerk '16

A final thought



“We TOLD you it was hard.” “Yes, but now that I’VE tried, I KNOW it’s hard-”

Bibliography

- Dunn, Michael 2015**. Language phylogenies**. In: Bowern & Evans (ed.s) Routledge Handbook of Historical Linguistics. London: Routledge
- Greenhill, Simon**, Phylogenetics & Tree-Thinking.** Spring School on Quantitative Methods, Jena 2017. Available at https://github.com/shh-dlce/qmss-2017/blob/master/TalkSlides/QMSS_2017_Greenhill-Phylogenetics.pdf
- Alexei J. Drummond, Remco Bouckaert - **Bayesian Evolutionary Analysis with BEAST** - Cambridge (2015)
- List, J.-M. (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
- List, J.-M. (2016): **Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction**. *Journal of Language Evolution* 1.2. 119-136.
- List, J.-M. (2017): **Introduction to computer-assisted language comparison** [Einführung in den computergestützten Sprachvergleich]. Institute of Linguistics: Nankai University (Tianjin, China).