

# Growing Trees from Big Data

## Bayesian Phylogeny for Historical Linguistics

Gereon Kaiping

2017-07-18

## 1 Solutions to All Your Problems!

- Crunching Numbers
- Bayesian ...
- ... Phylogenetics

## 2 [Examples]

- Austronesian: Branches and times
- Bantu: Phylogeography
- Indo-European: Ancient written sources

## 3 And why actually not.

## 4 Conclusions

- Further Reading

# Problem

Using the comparative method is hard and limited, because

- it is a lot of painstaking work,
- we don't know how to weigh the evidence,
- is a mix of hypothesis generation and validation,
- loan words and chance resemblances make our lives more difficult,
- cognates may have changed meanings (but how far?).

And then it doesn't even give us dates, just "not before" or "not after" if we are lucky.

# Problem

Using the comparative method is hard and limited, because

- it is a lot of painstaking work,
- we don't know how to weigh the evidence,
- is a mix of hypothesis generation and validation,
- loan words and chance resemblances make our lives more difficult,
- cognates may have changed meanings (but how far?).

And then it doesn't even give us dates, just "not before" or "not after" if we are lucky.

# Solution

## Tree reconstruction methods from **Bioinformatics**

cctccacgcc aacggagcct cattcttctt  
cctccacgcc aacaaagcct cattctt---  
cctccacgcc aacaaaggcct cattcttctt →  
cctccacgcc aacggagcct caggtgtctt  
cc---actcc aacggagcct caggtgtctt

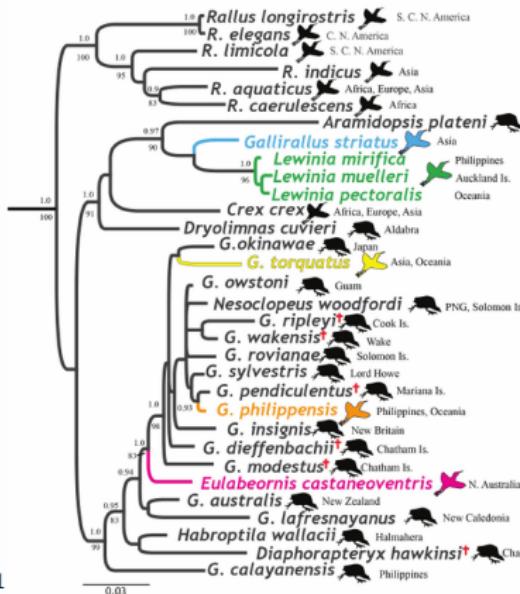
---

<sup>1</sup>Garcia-Ramirez et al. 2015

# Solution

## Tree reconstruction methods from Bioinformatics

```
cctccacgccc aacggagcct cattcttctt →
cctccacgccc aacaaaggcct cattctt---
cctccacgccc aacaaaggcct cattcttctt
cctccacgccc aacggagcct caggtgtctt
cc---actcc aacggagcct caggtgtctt
```

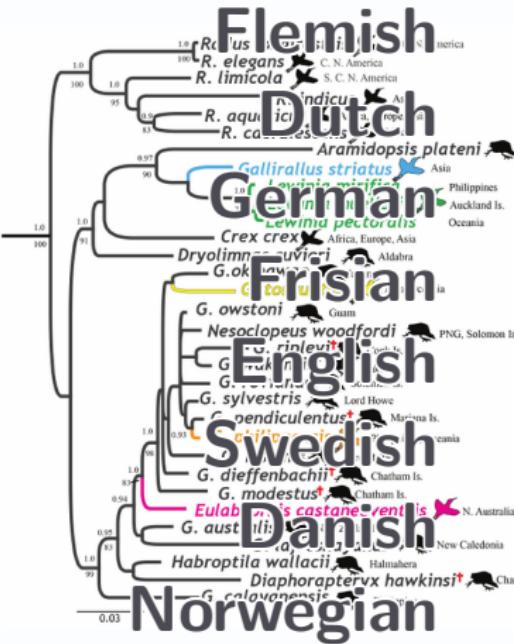


<sup>1</sup>Garcia-Ramirez et al. 2015

# Solution

## Tree reconstruction methods from Bioinformatics

h o n - e r - t \_ v c u - t  
 h o n d e - - t \_ v c:--t  
 h o n d e - - t \_ β æ - - t →  
 h e n d - u e d \_ w 3:--d  
 h o n d - r a - - o:--d



1

<sup>1</sup>Garcia-Ramirez et al. 2015

# Evolution as a random process

**Idea** Evolution = a random process on a tree<sup>2</sup>.

Generate a tree using dice rolls.



## Example

- Start with a single language, and proceed for 3 generations.
- In each generation and language, roll a dice: On 6, split the current language in 2.

Possible trees:



"Likelihood":  $P(\text{Data} \mid \text{Model})$

<sup>2</sup>or a network or a population, as long as we can formalize it.

# Evolution as a random process

**Idea** Evolution = a random process on a tree<sup>2</sup>.

Generate a tree using dice rolls.



## Example

- Start with a single language, and proceed for 3 generations.
- In each generation and language, roll a dice: On  $\geq 4$ , split the current language in 2.

Possible trees:



“Likelihood”:  $P(\text{Data} \mid \text{Model})$

<sup>2</sup>or a network or a population, as long as we can formalize it.

# Evolution as a random process

**Idea** Evolution = a random process on a tree<sup>2</sup>.

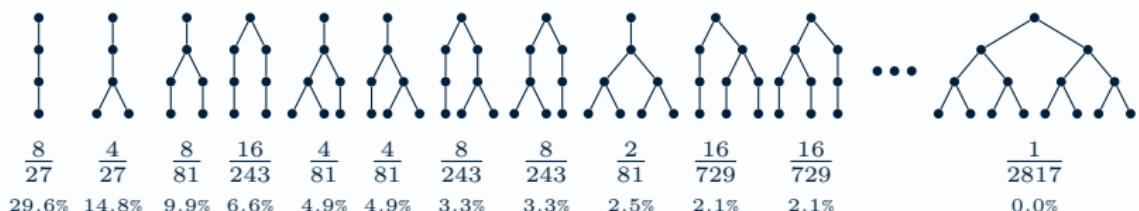
Generate a tree using dice rolls.



## Example

- Start with a single language, and proceed for 3 generations.
- In each generation and language, roll a dice: On  $\geq 4$ , split the current language in 2.

Possible trees:



“Likelihood”:  $P(\text{Data} \mid \text{Model})$

<sup>2</sup>or a network or a population, as long as we can formalize it.

# Evolution as a random process

**Idea** Evolution = a random process on a tree<sup>2</sup>.

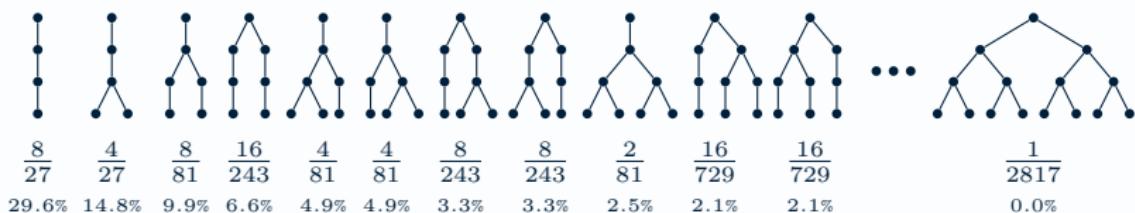
Generate a tree using dice rolls.



## Example

- Start with a single language, and proceed for 3 generations.
- In each generation and language, roll a dice: On 6, split the current language in 2.

Possible trees:

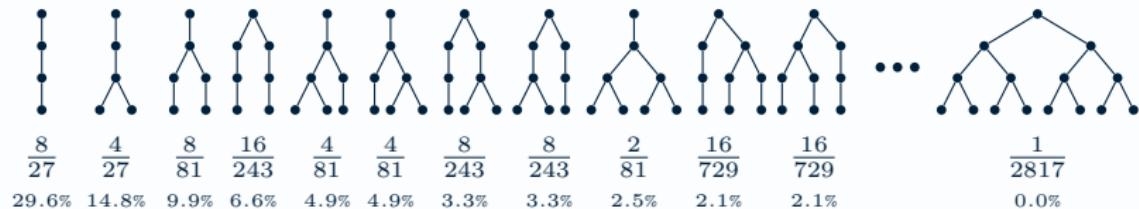


“Likelihood”:  $P(\text{Data} \mid \text{Model})$

<sup>2</sup>or a network or a population, as long as we can formalize it.

# Going back from data

Is given data compatible with this model?



“I generated a tree with three recent languages.”

“Also, my first roll was one of ☀☀☀”

How compatible is this data with this or that model? Which models should I believe in?

Probabilities = confidence of belief. Not: repeatable random experiment.

“Posterior probability”:  $P(\text{Model} \mid \text{Data})$

## Going back from data

Is given data compatible with this model?



“I generated a tree with three recent languages.”

“Also, my first roll was one of ☀☀☀”

How compatible is this data with this or that model? Which models should I believe in?

Probabilities = confidence of belief. Not: repeatable random experiment.

“Posterior probability”:  $P(\text{Model} \mid \text{Data})$

## Going back from data

Is given data compatible with this model?



"I generated a tree with three recent languages."

"Also, my first roll was one of ☀☀☀"

How compatible is this data with this or that model? Which models should I believe in?

Probabilities = confidence of belief. Not: repeatable random experiment.

"Posterior probability":  $P(\text{Model} \mid \text{Data})$

# Bayes' Theorem

$$P(\text{Model} \mid \text{Data}) \propto P(\text{Data} \mid \text{Model}) \times P(\text{Model}) \quad (3)$$

“What did the language history look like?”

=

“What trees are compatible with the data and my idea of language change?”

=

“Weighted by how ‘strange’ they are, how well does each tree explain my data?”

Bayesian inference may look complicated, but it is

- model-based
- can incorporate prior knowledge
- outputs result uncertainty
- gives implicit weights from first principles



# Bayes' Theorem

$$P(\text{Model} \mid \text{Data}) \propto P(\text{Data} \mid \text{Model}) \times P(\text{Model}) \quad (3)$$

“What did the language history look like?”

=

“What trees are compatible with the data and my idea of language change?”

=

“Weighted by how ‘strange’ they are, how well does each tree explain my data?”

Bayesian inference may look complicated, but it is

- model-based
- can incorporate prior knowledge
- outputs result uncertainty
- gives implicit weights from first principles

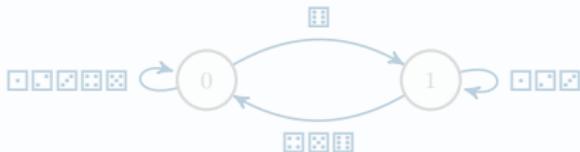


# Computational Phylogenetics

"Roll dice to generate trees, but only keep the good ones"

Need:

- simple stochastic model(s) of language evolution, with parameters.  
Example: *generalized binary model*



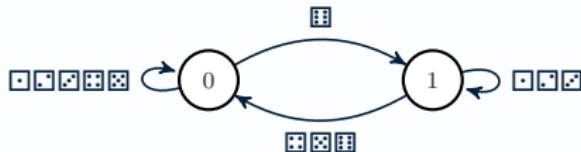
- intuition ("prior") of what parameters look like
- large dataset of model-compatible data

# Computational Phylogenetics

"Roll dice to generate trees, but only keep the good ones"

Need:

- simple stochastic model(s) of language evolution, with parameters.  
Example: *generalized binary model*



- intuition ("prior") of what parameters look like
- large dataset of model-compatible data

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	<i>hand</i>	<i>two</i>
Dutch	hant	tue:
English	hænd	tu:
French	mẽ	dø
Indonesian	taŋan	dua

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	<i>hand</i>	<i>two</i>
Dutch	hant	tue:
English	hænd	tu:
French	mě	dø
Indonesian	taŋan	dua

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	<i>hand</i>	<i>two</i>
Dutch		
English		
French		
Indonesian		

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	hand	two
Dutch	1	4
English	1	4
French	2	4
Indonesian	3	5

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	hand	two
Dutch	1 0 0	1 0
English	1 0 0	1 0
French	0 1 0	1 0
Indonesian	0 0 1	0 1

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# Data for Computational Phylogenetics

- Swadesh lists: models based on semantic change (like Glottochronology – but much more flexible)

Language	hand	two
Dutch	1 0 0	1 0
English	1 0 0	1 0
French	0 1 0	1 0
Indonesian	0 0 1	0 1

- Geography: various models
- Phonetic alignments: still in infancy
- Typological data: Some approaches, problems with universals/pathways
- Morphosyntax: ??????

# So much the theory.

- Austronesian: Branches and times
- Bantu: Phylogeography
- Indo-European: Ancient written sources

# Example 1: Austronesian

abv.org › Austronesian › Trees › (Beta)

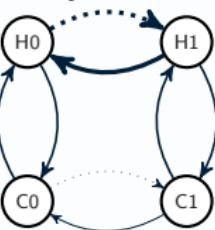
## Austronesian Basic Vocabulary Database

Word: hand

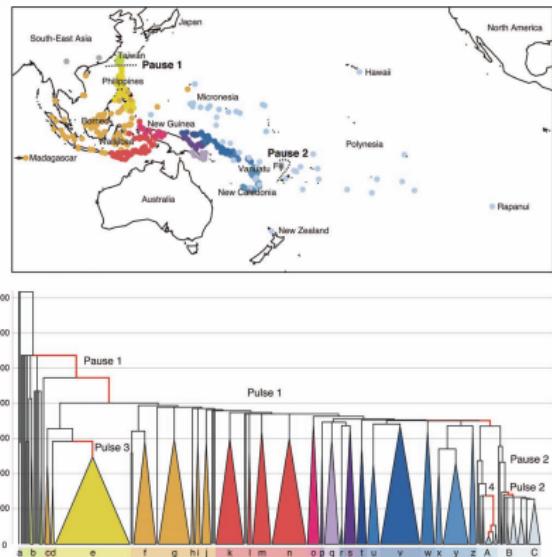
Entries for "hand":

ID	Language	Item	Annotate	Cognacy	Classification	Loan
<b>Hand</b>						
317959	Noroi (Bangg)	ka'w	ka'w			
317960	Noroi (Bangg)	ka'w	ka'w			
33064	Mohor (Tasitol)	kamay				
359818	Proto-Oceanic	*kamay				
327485	Osagai-Mindanao (1773)	isengat				
215785	Proto-Min-Khaser	gahay				
205937	Proto-Min-Khaser	*gahay				
186016	Chewung	cas				
177236	Han	cas				
177236	Hut	cas				
289824	Nukuhiva, Central	ka'				
246693	Hang (Ampong)	terus				
246693	Hang (Ampong)	terus				
246427	Batak	terus				
247574	Bugis (Nalati)	teus				
247574	Bugis (Paramping)	teus				
123576	Malagasy	tsia				
308716	Sarawak	ta				
205	Proto-Austronesian	*tasi				
184446	Proto-Austronesian (Lexic)	*kandu		1	Austronesian	
184446	Proto-Austronesian (Lexic)	*kandu		1, 79	Austronesian	
216	Malay (Central)	pasar		2	K-A-C-P	
106921	Malay - Cleft USA	gala!		2	K-A-C-P!	
205353	Malay - Cleft USA	gla?		2	K-A-C-P!	
71424	Malay - Soqali Fili	gasa?		2	K-A-N-SqS	
71425	Malay - Soqali Fili	rapa?		2	K-A-N-SqS	
71426	Malay - Soqali Fili	kasa		2	K-A-N-SqS	
71427	Malay - Soqali Fili	ketem		2	K-A-N-SqS	
71428	Malay - Soqali Fili	avet		2	K-A-N-SqS	
205720	Malay (Central)	baet?		2	K-A-N-SqS	
235	Sundan (Sunda)	baen?		51	K-A-Sundan	
207720	Sundan-LB1 (Pawon)	baeq?		51	K-A-Sundan	
207520	Sundan-LB1 (Totan)	baeq		51	K-A-Sundan	
207674	Sundan-LB1 (Husua)	baeq		51	K-A-Sundan	
207833	Sundan-LB1 (Husua)	baeq		51	K-A-Sundan	
71419	Burun Htb, Southern	inot		3	A-Burun	
71420	Burun Htb, Southern	lapid				
38602	Burun (Tabukau) LBS	ka'w?		1	A-Burun	
203720	Burun (Tabukau) LBS	ka'w?		1	A-Burun	
203946	Burun (Tabukau)	ka'wa		1	A-Burun	
204355	Burun (Tabukau)	pime		1	A-Burun	
60822	Aring (Central)	ka'may		13	A-B-RC	
203945	Aring (Central)	ka'may		13	A-B-RC	
203941	Aring (Central)	ka'may		13	A-B-RC	
203941	Aring (Central)	ka'may?		1	A-B-RC	
203941	Aring (Central)	ka'may?		1	A-B-RC	
210373	Borneo T'boli (D)	chme		1	A-C-D-Dene	
210595	Borneo T'boli (D)	bahe		2	A-C-D-Dene	
210595	Borneo T'boli (D)	ene		1	A-C-D-Dene	
63899	Kawioi Htb	lins?		1	A-C-H-Kawioi	
63899	Kawioi Htb	gralp		3	A-C-H-Kawioi	

- Austronesian Basic Vocabulary Database: several 1000 cognate classes for 210 meanings in 400 langs
- Plus two “outgroup” langs, minus borrowings
- Binary covariation model
- Calibrations and variable replacement rates



## Example 1: Austronesian



"The invention of the outrigger canoe and its sail may have enabled the Austronesians to move across this channel before spreading rapidly over the 7000 km from the Philippines to Polynesia (4). This is supported by linguistic reconstructions showing that the terminology associated with the outrigger canoe complex can only be traced back to Proto-Malayo-Polynesian and not Proto-Austronesian (41)."

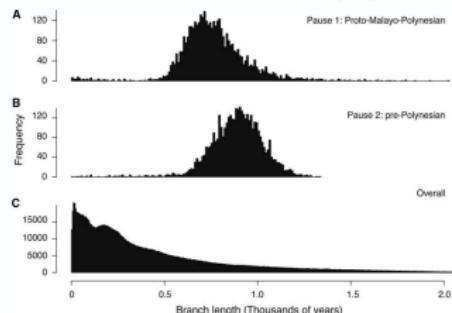


Fig. 3. Histograms of the branch length distributions. (A) The distribution of the Proto-Malayo-Polynesian pause, (B) the distribution of the pre-Polynesian pause, and (C) the overall branch-length distribution.

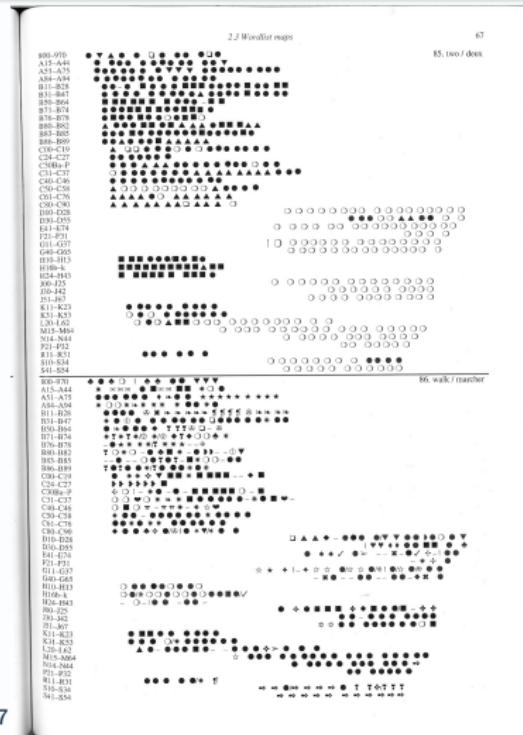
<sup>6</sup>Gray, Drummond & Greenhill 2009

## Example 1: Austronesian – Critique

- Pauses and pulses appear with high posterior probability
- Prior? Do the results follow from data or original guess?
- Some subgroupings not linguistically supported – Data contains sociogeography
- How realistic is binary covarion?

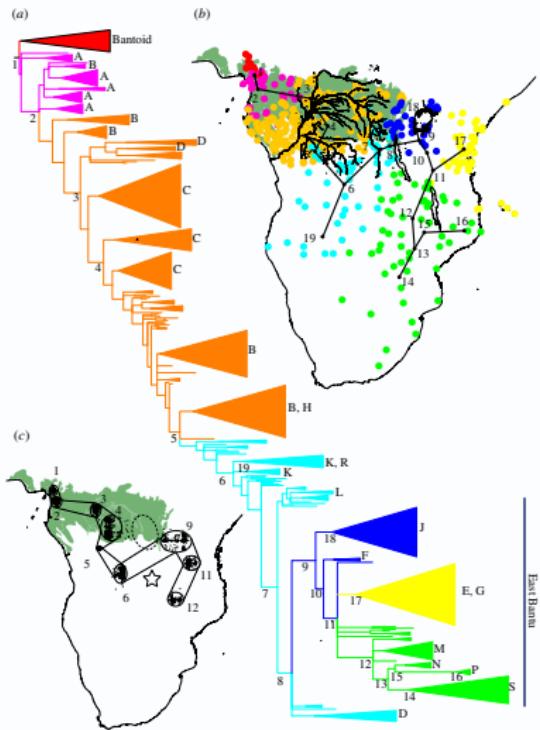
## Example 2: Bantu

- 2908 cognate classes for 90 meanings in 542 varieties of Bantu/Bantoid, with geographical point-coordinates
- Binary covariation with 6 (empirical) rate categories
- Brownian motion ancestral state reconstruction of latitudes and longitudes on 500 best trees
- Branch-dependent speed of movement and lexical change
- Other statistical analyses



<sup>7</sup>Bastin, Coupez & Mann 1999

## Example 2: Bantu



8

<sup>8</sup>Currie et al. 2013

"These debates have implications regarding the origin and spread of important cultural innovations, such as metallurgy and cattle-keeping."

"[...] explicit mapping of ancestral locations to make inferences about the specific route taken during the dispersal of Bantu languages. The results clearly support the 'pathway through the rainforest' scenario for the expansion of Bantu through much of sub-Saharan Africa. There is no support in these analyses for an early, deep split between East and West Bantu languages and a movement by one branch north of the rainforest."

## Example 2: Bantu – Critique

- Several robustness checks of parameters
- Prior? Geography without lexical data?
- The tree is unrooted.
- How good is Brownian motion as model for language spread?  
Language shift and post-split contact might affect geographic inference.  
(Though the fundamental results look robust.)

## Example 3: Indo-European – Data and Prologue

a 066 HAND

b

066 73 Ossetic  
 066 59 Gujarati

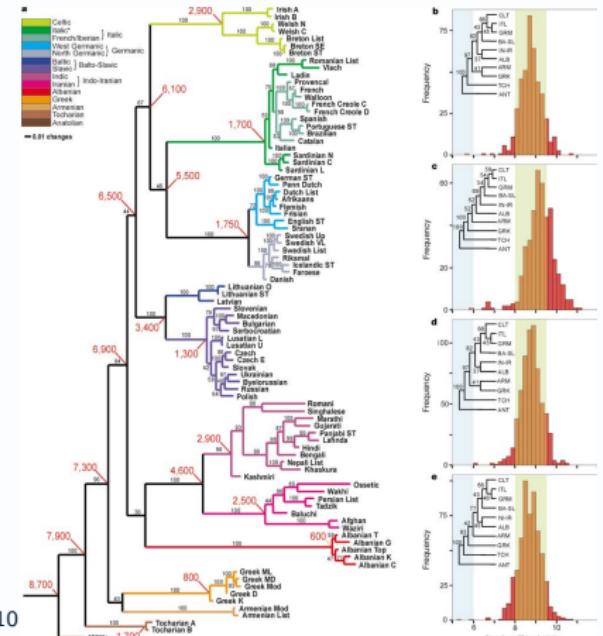
001

K"YX  
 NATH

c

066 17 Sardinian N	MANU
066 18 Sardinian L	MANU
066 09 Vlach	MYNE
066 22 Brazilian	MAO
066 21 Portuguese ST	MAO
066 15 French Creole C	LAME
066 13 French	MAIN
066 16 French Creole D	LAME
066 14 Walloon	MIN
066 12 Provencal	MAIN
066 20 Spanish	MANO
066 23 Catalan	MA
066 10 Italian	MANO
066 19 Sardinian C	MANU
066 11 Ladin	MAUN
066 08 Rumanian List	MINA

9



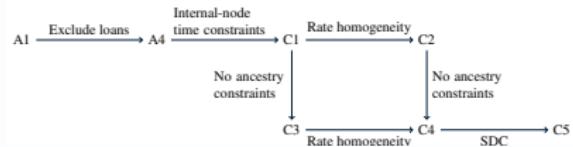
<sup>9</sup>Dyen 1997

<sup>10</sup>Gray & Atkinson 2003

## Example 3: Indo-European

IELex hand login

ID	Language	Source Form	Phonological Form	Notes	Cognate Class
11.4	Proto-Indo-European	*mon-u-			E
11.4	Proto-Indo-European	*gʰ̥es-r(o)-			C
11.4	Proto-Indo-European	*gʰ̥es-t(o)-			
80	Hittite	keššar			C
133	Luvian	išaris			C
134	Lycian	izre			C
81	Tocharian A	tsar			C
82	Tocharian B	ṣar			C
88	Albanian	dorë		A singularised neut. plural PAIb ...	C
143	Standard Albanian	dorë			C
2	Albanian Sicily	dorë		A singularised neut. plural PAIb ...	C
4	Albanian Corinth	dorë		A singularised neut. plural PAIb ...	C
3	Albanian Gheg	dorë		A singularised neut. plural PAIb ...	C
6	Albanian Tsk	dorë		A singularised neut. plural PAIb ...	C
173	Mycenaean Greek	ke-º	kʰer-	Attested as an element in ...	C
110	Ancient Greek	χείρ	kʰé:r	G.sg. χειρός	C
152	Tsakonian	χερά			C
32	Greek	χερί	'çeri		C
31	Greek Lesbos	CHERI			C
129	Classical Armenian	ծեռն	jeñn		C
8	Armenian Eastern	ծերգ	džerkʰ		C
7	Armenian Western	ծերգ	ts'erkʰ		C
11	128 Avestan	zastō			C

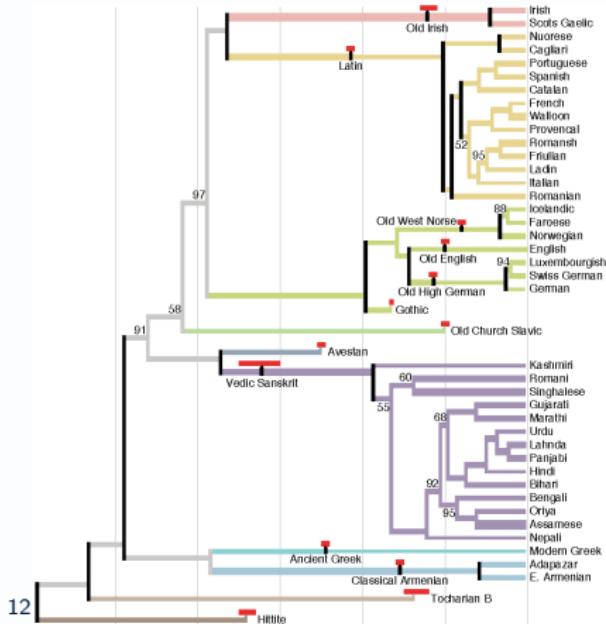


Starting from a replication of previous work (Bouckaert et al. 2012), improve

- data
- methodology
- tree prior
- post-processing

comparing each step.

## Example 3: Indo-European



"Here we present a phylogenetic analysis in which ancestry constraints permit more accurate inference of rates of change, based on observed changes between ancient or medieval languages and their modern descendants, and we show that the result strongly supports the steppe hypothesis."

"Because previous statistical phylogenetic research supported the Anatolian hypothesis, linguists who find that hypothesis implausible for other reasons may dismiss statistical analyses that purport to determine ancestral chronology. [...] statistical phylogenetic analysis can yield reliable information about pre-historic chronology, at least where all of the available data is taken into consideration."

## Example 3: Indo-European – Critique

- Very explicit about methodology (small steps, driver files available)
- Careful description of data coding
- Would someone have been this careful if the original results *had* matched the linguists' expectations?
- Ancestral constraints are very strong, and somewhat artificial in the model.

Further reading: Verkerk (2017)

# It will not solve all problems

## Some papers

- disregard prior knowledge
- use models that don't fit their data
- don't show their priors

# Not even in a better world.

## State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

## Not even in a better world.

### State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

## Not even in a better world.

### State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

## Not even in a better world.

### State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

## Not even in a better world.

### State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

### I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

## Not even in a better world.

### State of the art models

- can only build trees, no language contact
- only support cognate data (baby steps towards phonetic data and typology), no distinction between innovation and retention
- are not realistic / not calibrated / have biases
- can't actually decide high-level relationships

I think we will always

- need domain experts
- have problems modelling morphosyntax
- have qualitative data that are hard to integrate

# Conclusions

- It is useful to talk about probabilities of events in the past
- Computer models can help make sense of large data sets
- The computer only tests consistency or helps build intuition, it does not replace expertise
- Very few language-appropriate models so far
- Building a *good* inference is hard!
- Mathematical models can handle and combine new types of data for new types of results

If you disagree with results, *what parameters or choices do you disagree with?*

## Sources and Further Reading I

-  Bastin, Yvonne, André Coupez & Michael Mann. 1999. *Continuity and divergence in the bantu languages: perspectives from a lexicostatistic study.* Musée royal de l'Afrique centrale.
-  Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 337(6097). 957–960. <https://doi.org/10.1126/science.1219669>. <http://www.sciencemag.org/content/337/6097/957> (3 December, 2014).
-  Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. <https://doi.org/10.1353/lan.2015.0005>. <http://www.linguisticsociety.org/files/news/ChangEtAlPreprint.pdf> (27 February, 2015).

## Sources and Further Reading II

-  Currie, Thomas E., Andrew Meade, Myrtille Guillon & Ruth Mace. 2013. Cultural phylogeography of the bantu languages of sub-saharan africa. *Proceedings of the Royal Society of London B: Biological Sciences* 280(1762). <https://doi.org/10.1098/rspb.2013.0695>. <http://rspb.royalsocietypublishing.org/content/280/1762/20130695>.
-  Dunn, Michael. 2015. *IELex – Indo-European Lexical Cognacy Database*. <http://ielex.mpi.nl/> (12 March, 2015).
-  Dyen, Isidore. 1997. *COMPARATIVE INDOEUROPEAN DATABASE COLLECTED BY ISIDORE DYEN*. <http://www.wordgumbo.com/ie/cmp/iedata.txt>.
-  Garcia-Ramirez, Juan C, Graeme Elliott, Kath Walker, Isabel Castro & Steven A Trewick. 2015. Trans-equatorial range of a land bird lineage (aves: rallidae) from tropical forests to subantarctic grasslands. *Journal of Avian Biology*.

## Sources and Further Reading III

-  Gray, R. D., A. J. Drummond & S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913). 479–483. <https://doi.org/10.1126/science.1166858>. <http://science.scienmag.org/content/323/5913/479>.
-  Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965). 435–439. <https://doi.org/10.1038/nature02029>. <http://www.nature.com/nature/journal/v426/n6965/abs/nature02029.html> (27 November, 2014).
-  Greenhill, S.J., R Blust & R.D. Gray. 2008. The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*. 271–283.
-  McMahon, April & Robert McMahon. 2005. *Language classification by numbers*. Oxford University Press. 285 pp.

## Sources and Further Reading IV

-  Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Sérgio Meira, Vivian Wauters & Zachary O'Hagan. 2015. A Bayesian Phylogenetic Classification of Tupí-Guaraní. *LIAMES: Línguas Indígenas Americanas* 15(2). 193–221.  
<https://doi.org/10.20396/liames.v15i2.8642301>.  
<https://periodicos.sbu.unicamp.br/ojs/index.php/liames/article/view/8642301> (29 August, 2017).
-  Verkerk, Annemarie. 2017. Phylogenies: Future, not fallacy. *Language Dynamics and Change* 7(1). 127–140.  
<https://doi.org/10.1163/22105832-00601013>.  
<http://booksandjournals.brillonline.com/content/journals/10.1163/22105832-00601013> (9 October, 2017).