Using simulated data to evaluate models of Indo-European vocabulary evolution

Philipp Rönchen[1], Oscar Billing[1] and Tilo Wiklund[2]

[1]Department of Linguistics and Philology, Uppsala University
[2]Chief Data Scientist, UAB Sensmetry, previously Department of Mathematics, Uppsala University

7 September 2023

Background: Using vocabulary to date proto-languages

- Since the emergence of glottochronology (Swadesh 1952), researchers have tried to use the vocabulary of modern languages to make inferences about the age of proto-languages

Background: Using vocabulary to date proto-languages

- Since the emergence of glottochronology (Swadesh 1952), researchers have tried to use the vocabulary of modern languages to make inferences about the age of proto-languages
- Basic idea: The more vocabulary a group of related languages have in common, the shorter the time since they split up

Background: Using vocabulary to date proto-languages

- Since the emergence of glottochronology (Swadesh 1952), researchers have tried to use the vocabulary of modern languages to make inferences about the age of proto-languages
- Basic idea: The more vocabulary a group of related languages have in common, the shorter the time since they split up
- Recent approaches tend to use extremely complicated methods of inference

Different inferences

- Different methods have tended to give different results

Different inferences

- Different methods have tended to give different results
- Recent examples for Indo-European:
  - Bouckaert et al. (2012, amended in Bouckaert et al. 2013) support "Anatolian hypothesis" with age larger than 8000 BP
  - Chang et al. (2015) support "Steppe hypothesis" with age around 6000 BP
  - Heggarty et al. (2023) infer yet another scenario, with age around 8000 BP

How to evaluate inferences – inspection of model?

How to evaluate inferences – inspection of model?

- Complicated inference methods make a lot of assumptions

How to evaluate inferences – inspection of model?

- Complicated inference methods make a lot of assumptions
- The effect of those is very hard to gauge by inspection only

Testing on known historic data?

- We can evaluate methods by testing whether they correctly infer known historic cases (for instance, age of Latin etc)

Testing on known historic data?

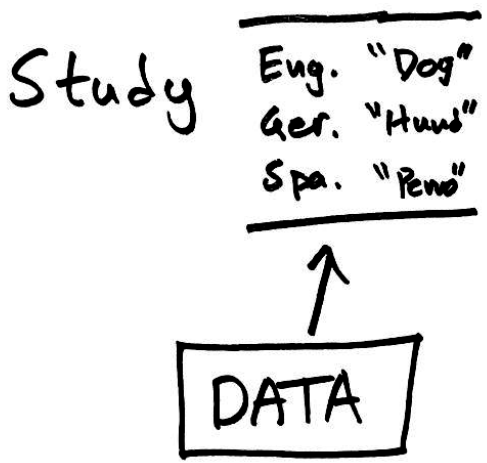
- We can evaluate methods by testing whether they correctly infer known historic cases (for instance, age of Latin etc)
- However, amount of historic data that can be used for testing is very limited

Testing on known historic data?

- We can evaluate methods by testing whether they correctly infer known historic cases (for instance, age of Latin etc)
- However, amount of historic data that can be used for testing is very limited
- Furthermore, methods usually use most historic data already for choosing the model parameters

Way forward: Realistic simulated data

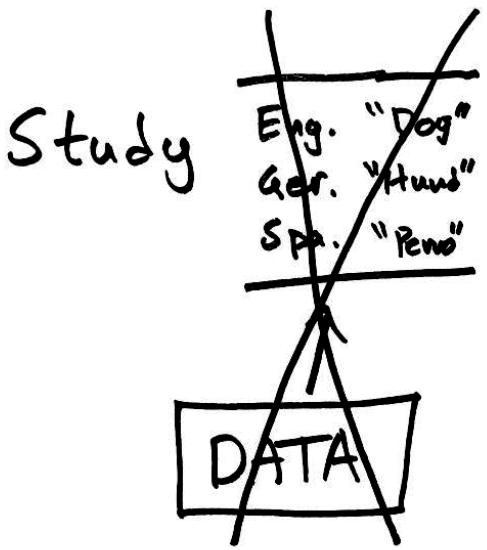
Way forward: Realistic simulated data



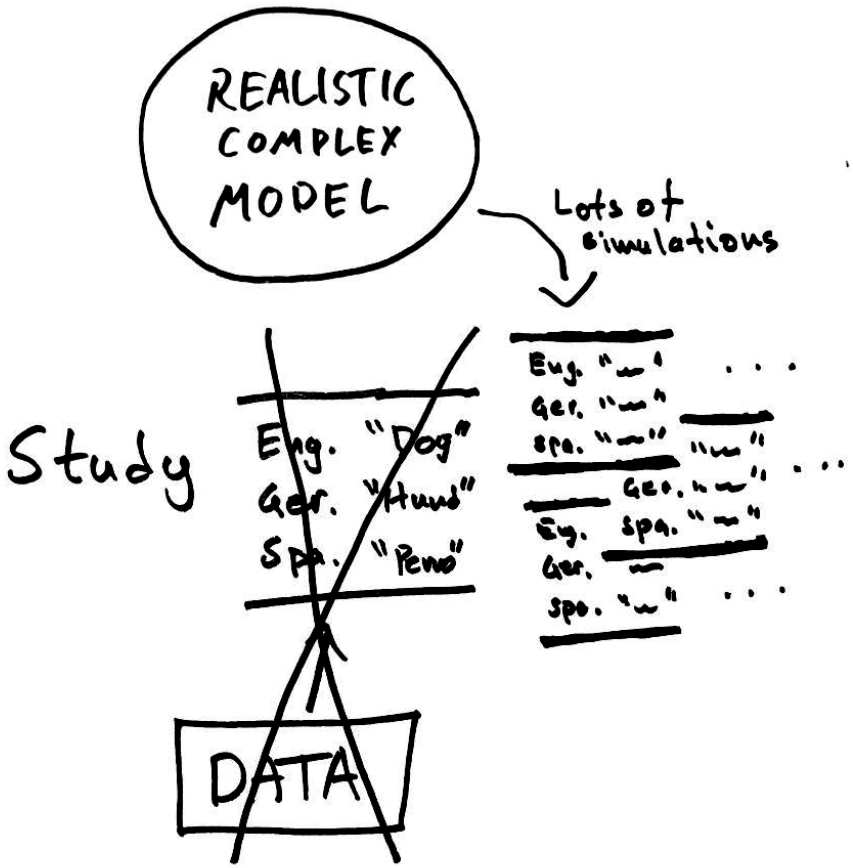
Way forward: Realistic simulated data

I.e. Chang et al (2015)
Bouckaert et al (2012)

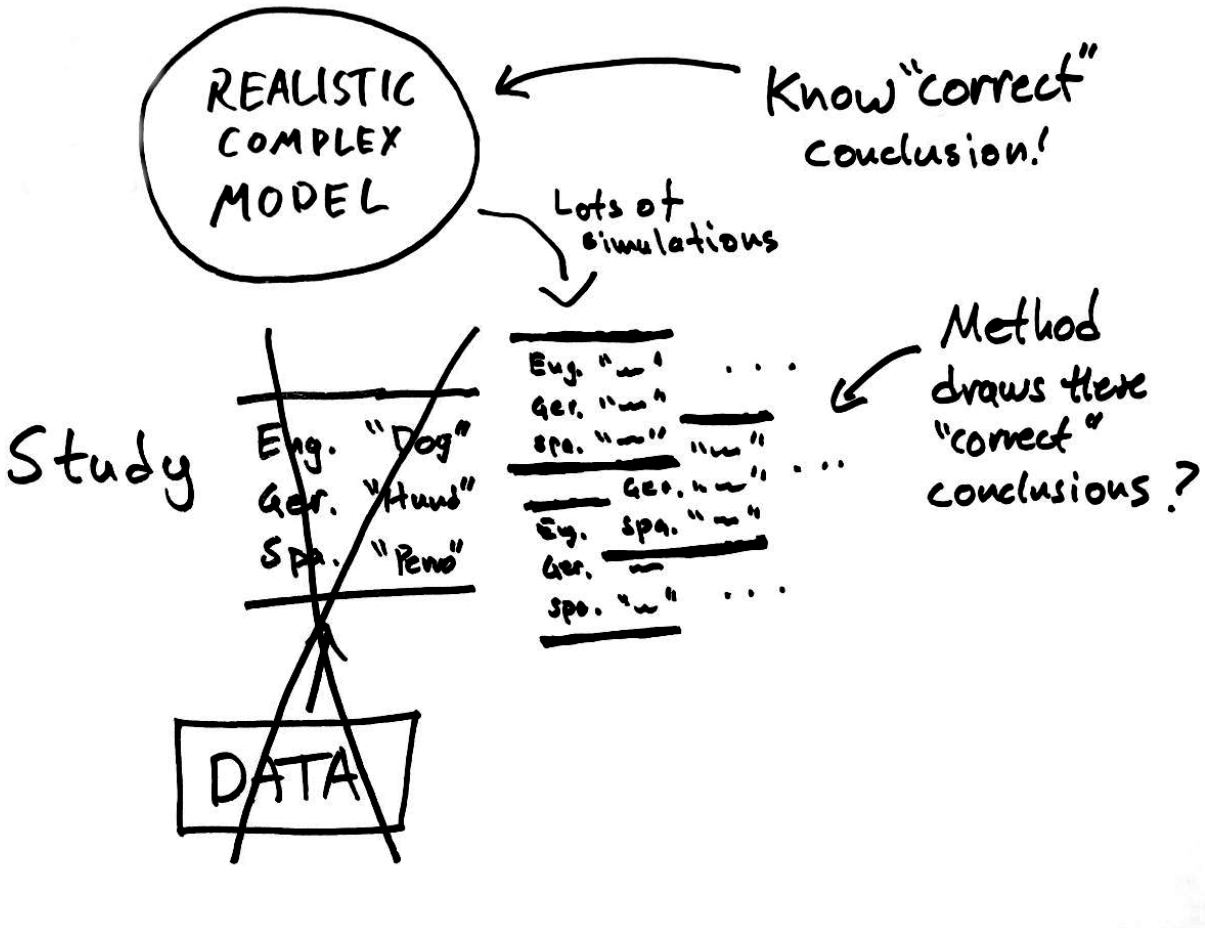
Way forward: Realistic simulated data



Way forward: Realistic simulated data



Way forward: Realistic simulated data



Principle: Uncertainty requires robustness

Principle: Uncertainty requires robustness

- We do not know the exact mechanisms by which Indo-European vocabulary evolved

Principle: Uncertainty requires robustness

- We do not know the exact mechanisms by which Indo-European vocabulary evolved
- That means: If we are to trust an inference method, it has to work on all data that *could have* evolved in a similar way as the Indo-European data evolved

Our simulation model

Ingredients:

Our simulation model

Ingredients:

- Fixed Indo-European tree that is reasonable (compatible with established research)

Our simulation model

Ingredients:

- Fixed Indo-European tree that is reasonable (compatible with established research)
- A sensible model of cognate evolution ("Multistate") with change rates estimated from historic data

Our simulation model

Ingredients:

- Fixed Indo-European tree that is reasonable (compatible with established research)
- A sensible model of cognate evolution ("Multistate") with change rates estimated from historic data
- Complemented by loan word events, with frequency of loan events estimated from World Loanword Database (Haspelmath and Tadmor, 2009)

Our simulation model

Ingredients:

- Fixed Indo-European tree that is reasonable (compatible with established research)
- A sensible model of cognate evolution ("Multistate") with change rates estimated from historic data
- Complemented by loan word events, with frequency of loan events estimated from World Loanword Database (Haspelmath and Tadmor, 2009)

We are using our simulation model to evaluate the methods of Chang et al. (2015) and Bouckaert et al. (2012) (not Heggarty et al. 2023 yet since it is too recent)

Tree construction – method

- Tree topology follows Olander (2018), in turn based on Ringe et al. (2002)

Tree construction – method

- Tree topology follows Olander (2018), in turn based on Ringe et al. (2002)
- Node dates follow "linguistic consensus" of specialists (usually a compromise!)
- Can be based on, e.g., contact event, linguistic palaeontology, "linguist's intuition", historical events (*termini post/ante quem*)

Tree construction – method

- Tree topology follows Olander (2018), in turn based on Ringe et al. (2002)
- Node dates follow "linguistic consensus" of specialists (usually a compromise!)
- Can be based on, e.g., contact event, linguistic palaeontology, "linguist's intuition", historical events (*termini post/ante quem*)
- Each node given as extensive justification in literature as possible
- We have not found dates for a small amount of nodes, then:
  - Check the data and make independent estimation
  - In difficult cases, apply arbitrary estimate

Tree construction – method

- Tree topology follows Olander (2018), in turn based on Ringe et al. (2002)
- Node dates follow "linguistic consensus" of specialists (usually a compromise!)
- Can be based on, e.g., contact event, linguistic palaeontology, "linguist's intuition", historical events (*termini post/ante quem*)
- Each node given as extensive justification in literature as possible
- We have not found dates for a small amount of nodes, then:
  - Check the data and make independent estimation
  - In difficult cases, apply arbitrary estimate
- NB these nodes do not invalidate analysis — a good model should be able to cope with any linguistic history!

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

- Toch. A and B dated 1350 BP as per Chang.

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

- Toch. A and B dated 1350 BP as per Chang.
- Toch. A and B "wholly separate" by first cents. CE (Pinault, 2002, 245).
- "unlikely that two languages as different as Tocharian A and B were separated by a couple of hundred years only" (Carling, 2005, 66).

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

- Toch. A and B dated 1350 BP as per Chang.
- Toch. A and B "wholly separate" by first cents. CE (Pinault, 2002, 245).
- "unlikely that two languages as different as Tocharian A and B were separated by a couple of hundred years only" (Carling, 2005, 66).
- Tocharian divergence between 500–1000y (Lane, 1966, 232).
- Tocharian divergence ca. 1000y (Ringe, 1995, 439).

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

- Toch. A and B dated 1350 BP as per Chang.
- Toch. A and B "wholly separate" by first cents. CE (Pinault, 2002, 245).
- "unlikely that two languages as different as Tocharian A and B were separated by a couple of hundred years only" (Carling, 2005, 66).
- Tocharian divergence between 500–1000y (Lane, 1966, 232).
- Tocharian divergence ca. 1000y (Ringe, 1995, 439).
- Mid/late 2nd mil. BCE (ca. 2000y of divergence), assuming later dialectal convergence (Adams, 2006, 388-389).

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

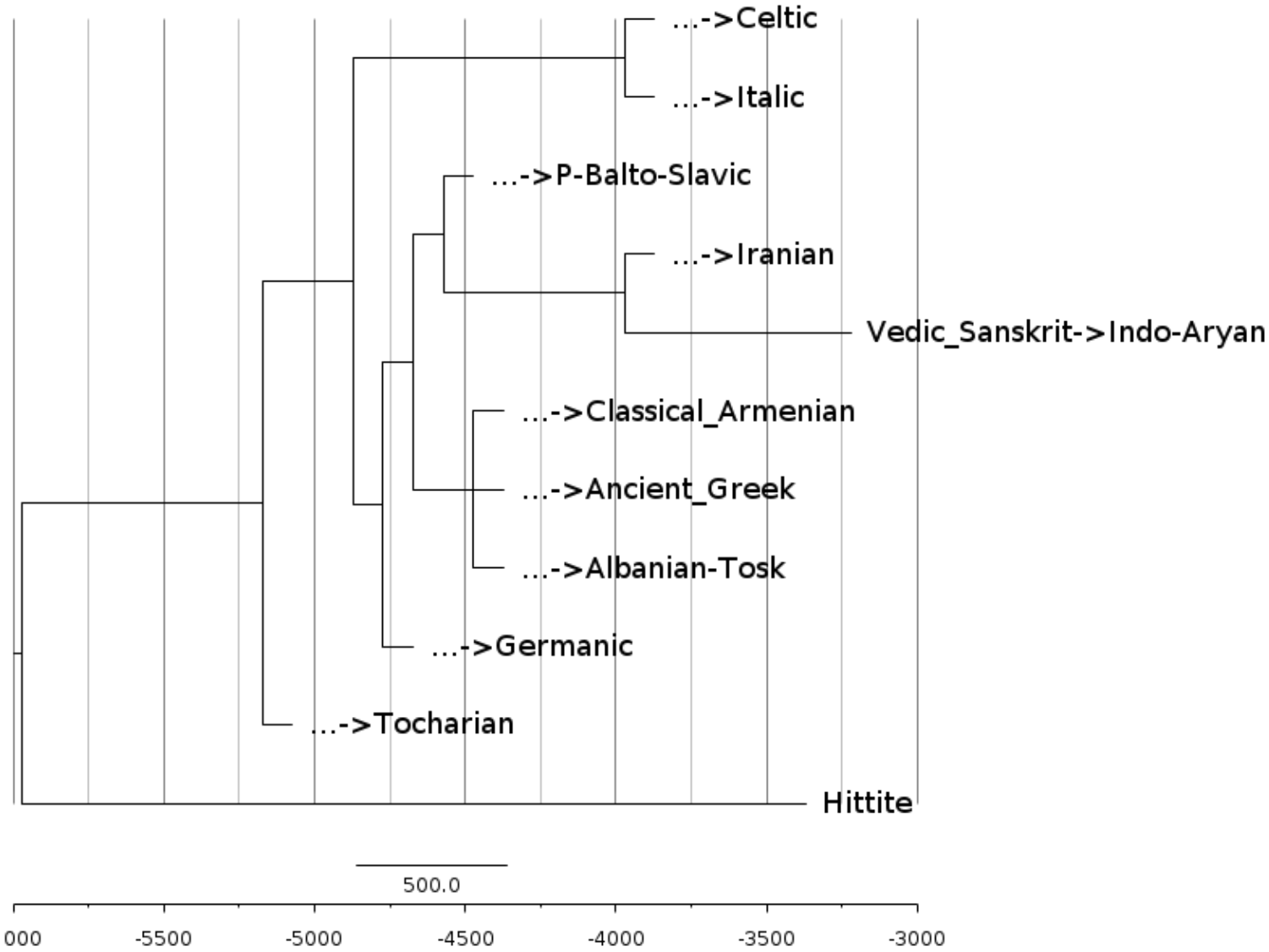
- Toch. A and B dated 1350 BP as per Chang.
- Toch. A and B "wholly separate" by first cents. CE (Pinault, 2002, 245).
- "unlikely that two languages as different as Tocharian A and B were separated by a couple of hundred years only" (Carling, 2005, 66).
- Tocharian divergence between 500–1000y (Lane, 1966, 232).
- Tocharian divergence ca. 1000y (Ringe, 1995, 439).
- Mid/late 2^{nd} mil. BCE (ca. 2000y of divergence), assuming later dialectal convergence (Adams, 2006, 388-389).
- P-Toch. beginning of 1^{st} mil. BCE, Iranian contact evidence adduced (Peyrot, 2022, 87).

Tree construction – example

Example: "P-Tocharian" (ancestor of Tocharian A and B)

- Toch. A and B dated 1350 BP as per Chang.
- Toch. A and B "wholly separate" by first cents. CE (Pinault, 2002, 245).
- "unlikely that two languages as different as Tocharian A and B were separated by a couple of hundred years only" (Carling, 2005, 66).
- Tocharian divergence between 500–1000y (Lane, 1966, 232).
- Tocharian divergence ca. 1000y (Ringe, 1995, 439).
- Mid/late 2nd mil. BCE (ca. 2000y of divergence), assuming later dialectal convergence (Adams, 2006, 388-389).
- P-Toch. beginning of 1st mil. BCE, Iranian contact evidence adduced (Peyrot, 2022, 87).
- Chosen date: **2750 BP** (1400y of divergence)

Our base tree – uppermost branches



Multistate model of cognate evolution

- Model suggested by Warnow et al. (2004)

Multistate model of cognate evolution

- Model suggested by Warnow et al. (2004)
- A language is modelled to have one primary word per meaning, with a certain probability (different for different meanings), this word is replaced over time

Multistate model of cognate evolution

- Model suggested by Warnow et al. (2004)
- A language is modelled to have one primary word per meaning, with a certain probability (different for different meanings), this word is replaced over time
- Whenever a replacement occurs, this results in the creation of a new cognate classes

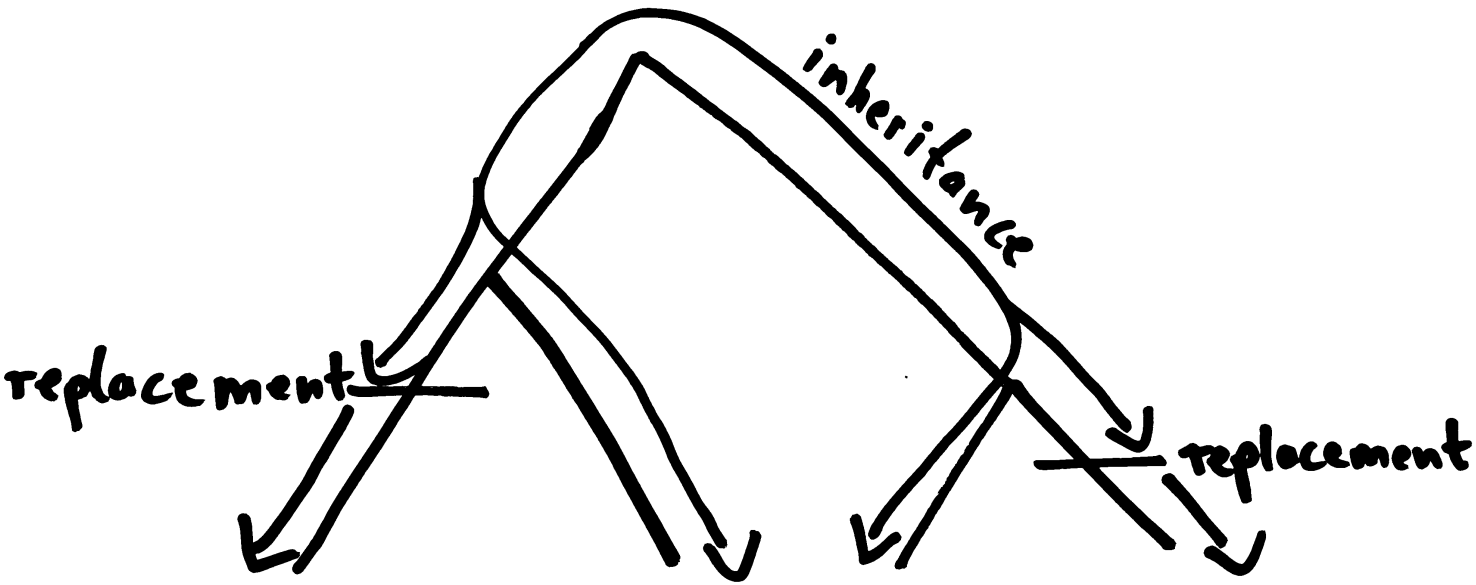
Multistate model of cognate evolution

- Model suggested by Warnow et al. (2004)
- A language is modelled to have one primary word per meaning, with a certain probability (different for different meanings), this word is replaced over time
- Whenever a replacement occurs, this results in the creation of a new cognate classes
- Multiple replacements along the same lineage are possible

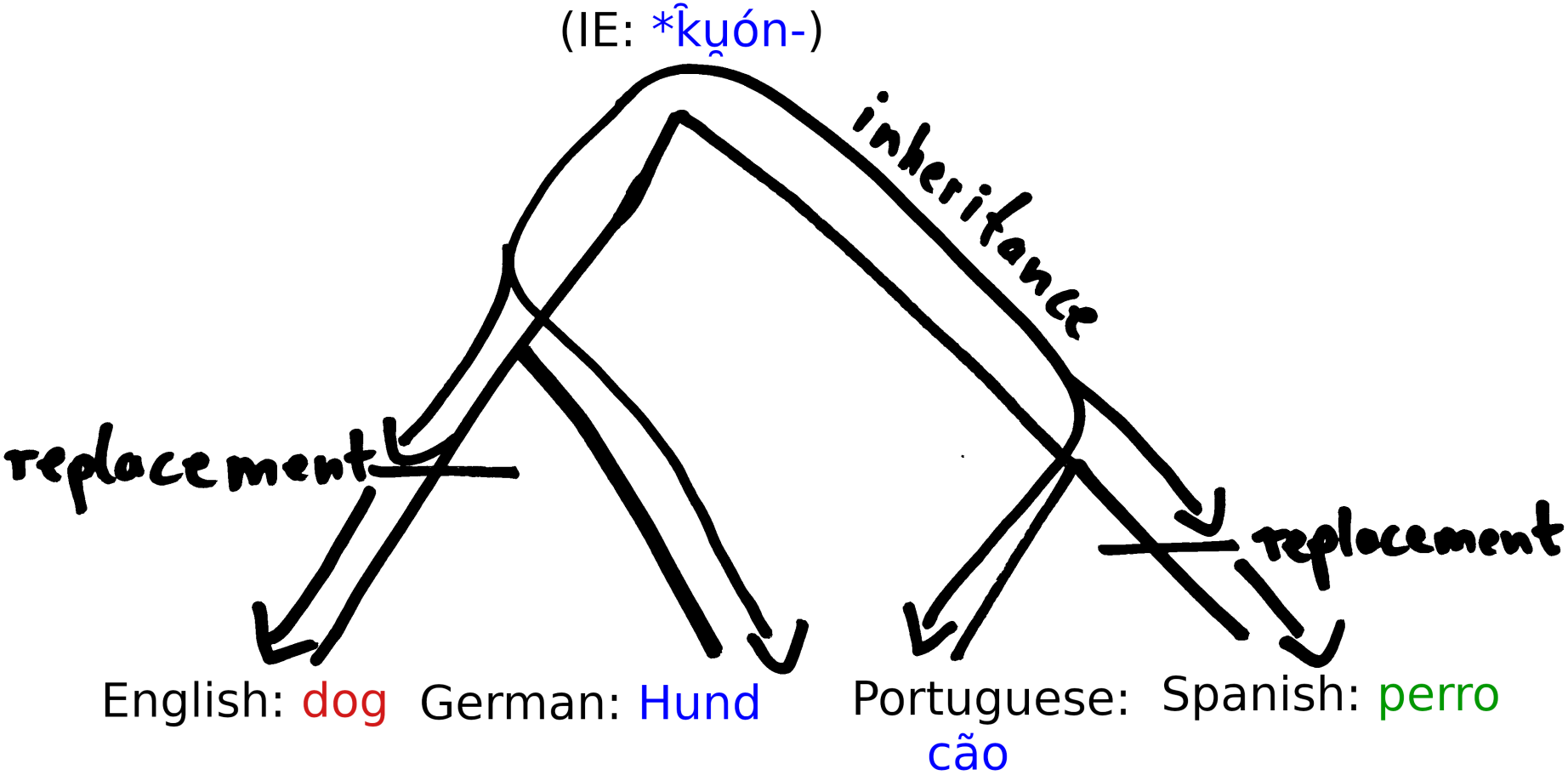
Multistate model of cognate evolution

- Model suggested by Warnow et al. (2004)
- A language is modelled to have one primary word per meaning, with a certain probability (different for different meanings), this word is replaced over time
- Whenever a replacement occurs, this results in the creation of a new cognate classes
- Multiple replacements along the same lineage are possible
- The change rates per meanings are estimated from historic data provided by Lees (1953) and Swadesh (1955)

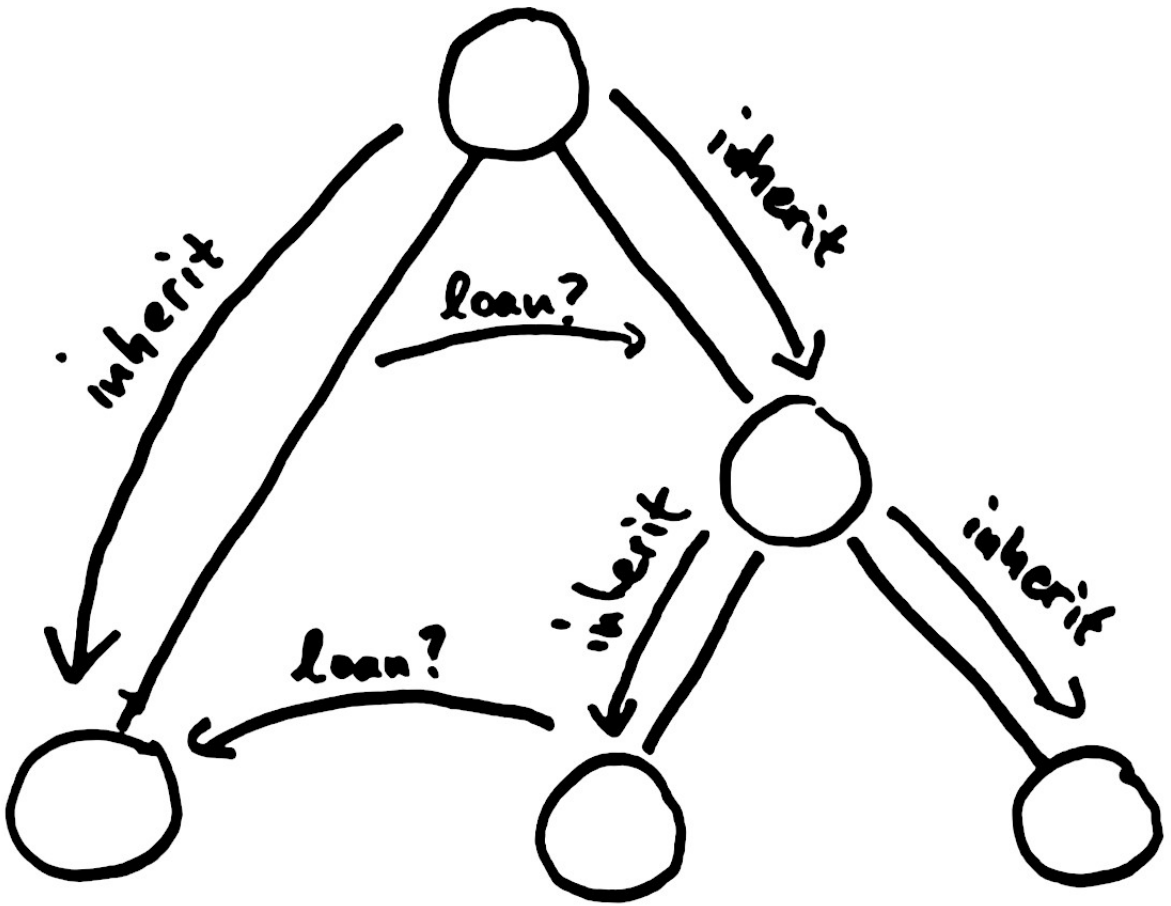
Multistate model



Multistate model - example



Including loan word process



Evaluating Chang et al.

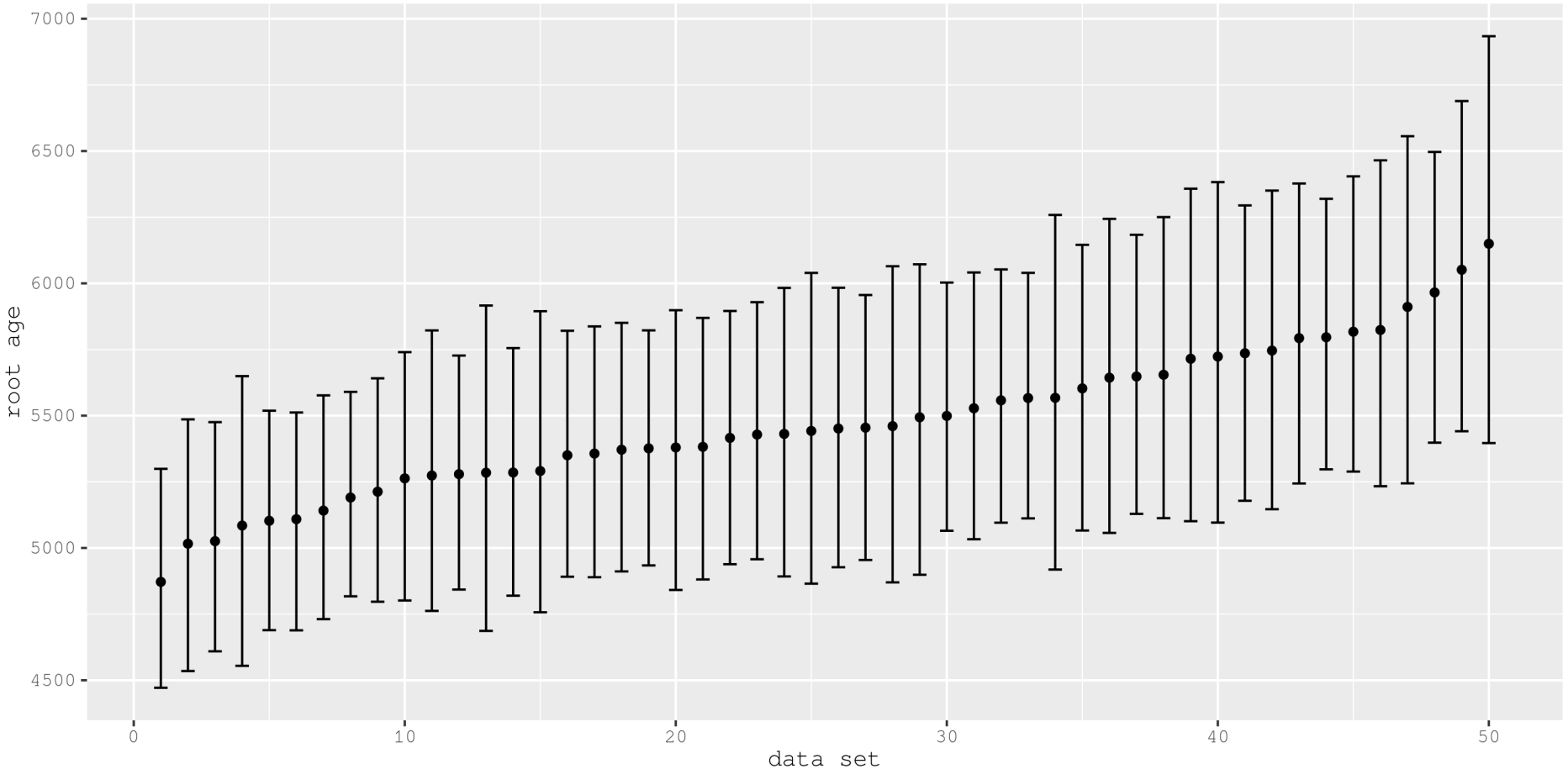
Evaluating Chang et al.

- We created 50 data sets with our simulation model and run Chang et al.'s main analysis on the data sets

Evaluating Chang et al.

- We created 50 data sets with our simulation model and run Chang et al.'s main analysis on the data sets
- We just replace the data in the XML-files used by Chang et al. and run them with the Beast-software (Drummond et al. 2012)

Main results for Chang et al.



Inferences of the age of Indo-European (HPD-intervalls and means) from Chang et al.'s main analysis A1 on 50 data sets simulated on our tree

Additional study: Old tree

- To be able to more directly see whether Chang et al. (2015)'s methods correctly arbitrates between different hypotheses we also created 50 data sets on an *older* tree

Additional study: Old tree

- To be able to more directly see whether Chang et al. (2015)'s methods correctly arbitrates between different hypotheses we also created 50 data sets on an *older* tree
- Created by combining upper branchings of Bouckaert et al. (2012)'s MCC tree with the lower branchings of our tree

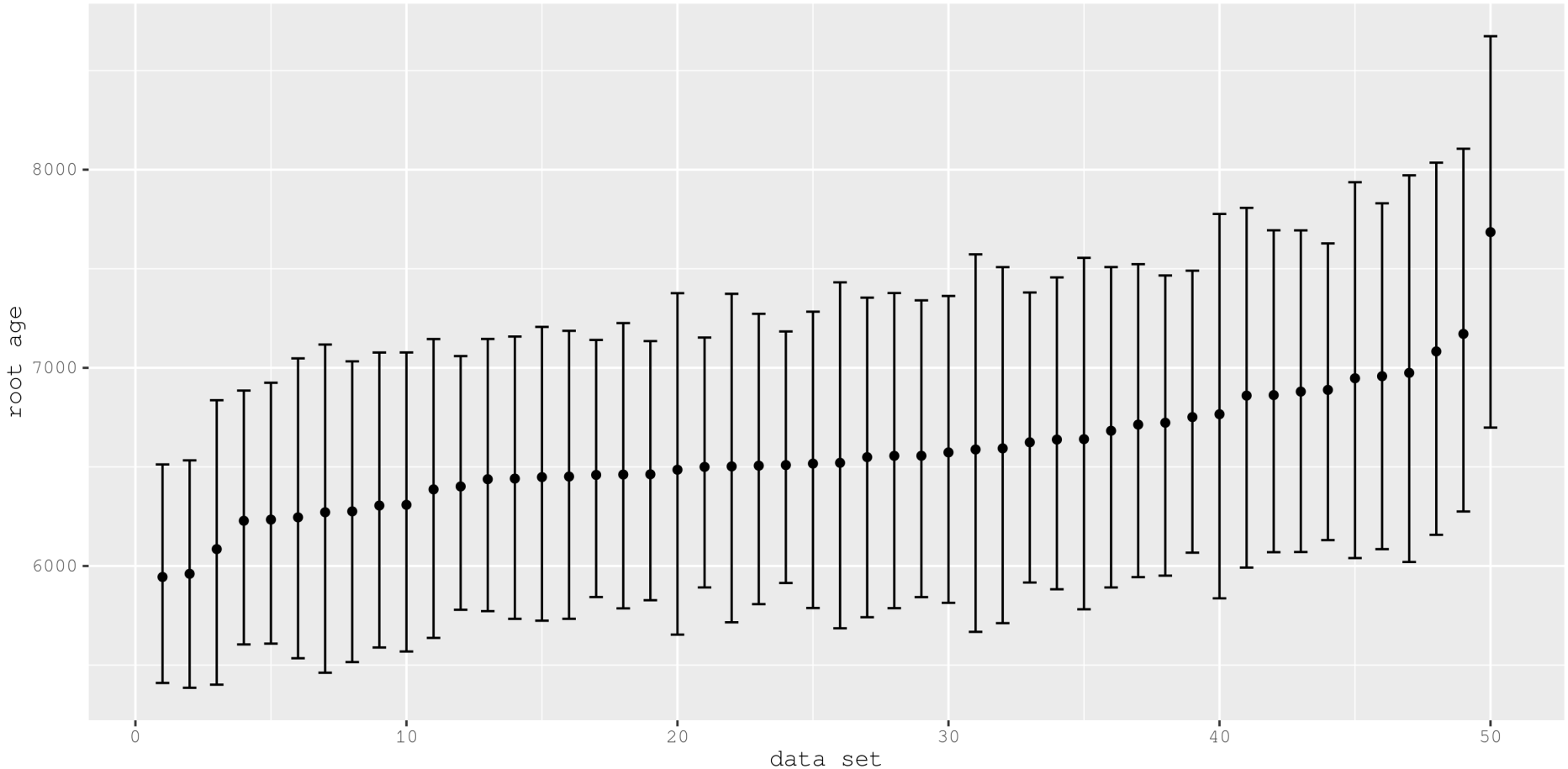
Additional study: Old tree

- To be able to more directly see whether Chang et al. (2015)'s methods correctly arbitrates between different hypotheses we also created 50 data sets on an *older* tree
- Created by combining upper branchings of Bouckaert et al. (2012)'s MCC tree with the lower branchings of our tree
- Split-up of PIE is set to 7850 years

Additional study: Old tree

- To be able to more directly see whether Chang et al. (2015)'s methods correctly arbitrates between different hypotheses we also created 50 data sets on an *older* tree
- Created by combining upper branchings of Bouckaert et al. (2012)'s MCC tree with the lower branchings of our tree
- Split-up of PIE is set to 7850 years
- *Not* as reasonable as our main tree, but still useful as a test

Results for Chang et al. on old tree



Inferences of the age of Indo-European (HPD-intervalls and means) from Chang et al.'s main analysis A1 on 50 data sets simulated on the older tree

Bayes factor test

- Using the methodology of Chang et al., we carry out a *Bayes factor test* for each of the simulated data sets produced on the older tree

Bayes factor test

- Using the methodology of Chang et al., we carry out a *Bayes factor test* for each of the simulated data sets produced on the older tree
- This is a measure for how much Chang et al.'s methodologies would choose the Steppe hypothesis over the Anatolian hypothesis

Bayes factor test

- Using the methodology of Chang et al., we carry out a *Bayes factor test* for each of the simulated data sets produced on the older tree
- This is a measure for how much Chang et al.'s methodologies would choose the Steppe hypothesis over the Anatolian hypothesis
- Since the root age of the old tree (7850 years) is closer to the Anatolian hypothesis, we would want Chang et al. to give negative Bayes factors, that is to not prefer the Steppe hypothesis

Bayes factor test – results

Method	very strong	strong	substantial	weak	negative
A1	39	5	3	2	1
A2	34	5	6	2	3
A3	42	1	6	0	1

Table: Support that the Bayes factor indicates for the Steppe over the Anatolian hypothesis

For most data sets, Chang et al. (2015)'s methods strongly favour the Steppe hypothesis, but there is some fluctuation

Evaluating Bouckaert et al.

- To evaluate Bouckaert et al. (2012), we create 50 more data sets on our main tree

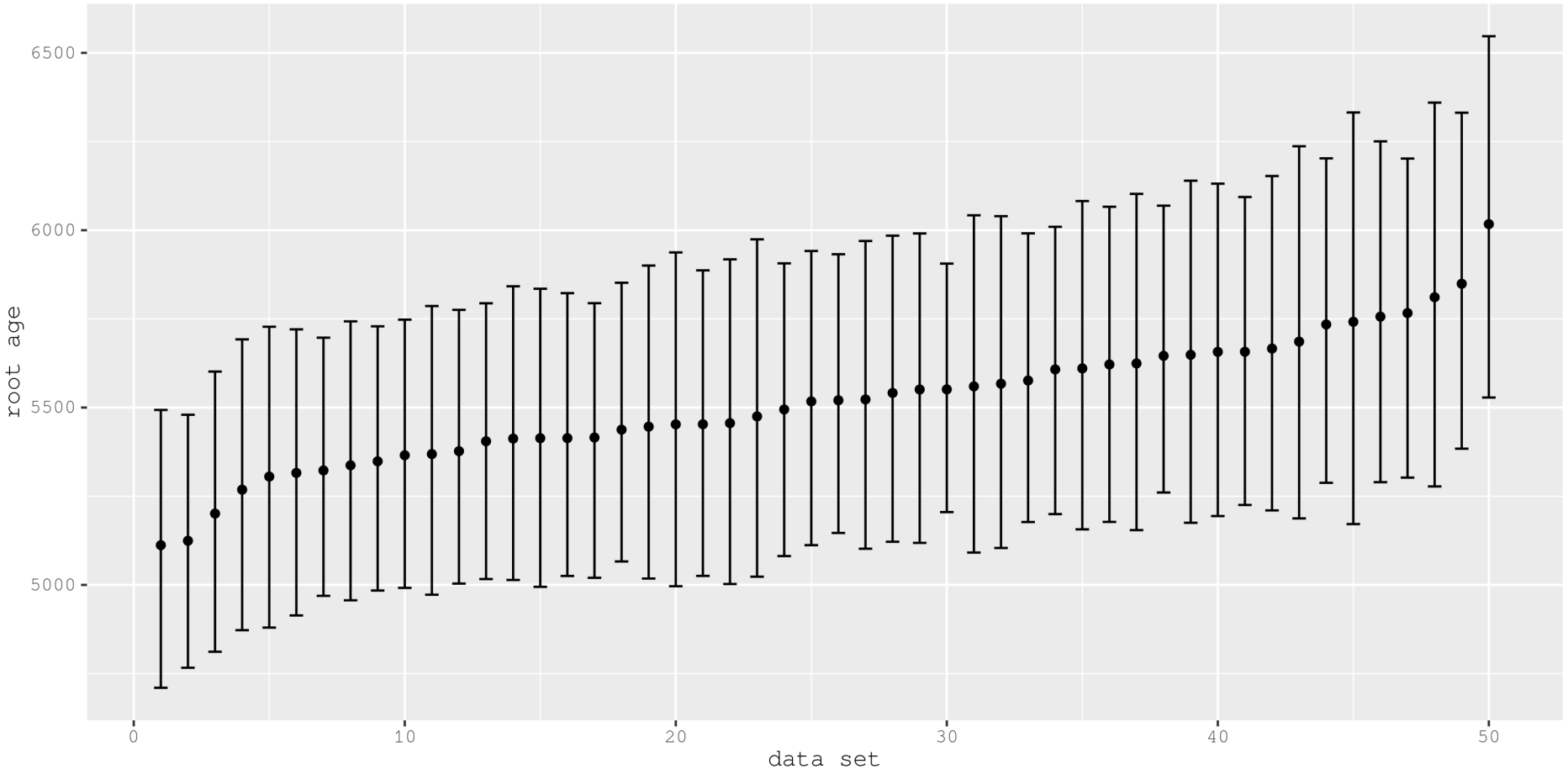
Evaluating Bouckaert et al.

- To evaluate Bouckaert et al. (2012), we create 50 more data sets on our main tree
- We slightly extended our simulation model to allow for some languages to have missing data and for some loan-words to be removed from the data

Evaluating Bouckaert et al.

- To evaluate Bouckaert et al. (2012), we create 50 more data sets on our main tree
- We slightly extended our simulation model to allow for some languages to have missing data and for some loan-words to be removed from the data
- We run Bouckaert et al. (2012)'s method on the data sets

Results Bouckaert et al.



Inferences of the age of Indo-European (HPD-intervalls and means) from Bouckaert et al.'s main analysis on 50 data sets simulated on our main tree

Evaluating Chang et al. and Bouckaert et al. – Summary

- Both Chang et al. and Bouckaert et al. tend to underestimate the true tree age by a large margin

Evaluating Chang et al. and Bouckaert et al. – Summary

- Both Chang et al. and Bouckaert et al. tend to underestimate the true tree age by a large margin
- There is considerable fluctuation of the inference depending on the data set (depending on the "randomness" in the data)

Evaluating Chang et al. and Bouckaert et al. – Summary

- Both Chang et al. and Bouckaert et al. tend to underestimate the true tree age by a large margin
- There is considerable fluctuation of the inference depending on the data set (depending on the "randomness" in the data)
- If our simulation model was the "true model", Chang et al.'s method would not be able to correctly distinguish between the Steppe and the Anatolian hypothesis

Our interpretation of results

One should be sceptical of the inferences of Chang et al. (2015) and Bouckaert et al. (2012) until that at least

Our interpretation of results

One should be sceptical of the inferences of Chang et al. (2015) and Bouckaert et al. (2012) until that at least

1. The realisticness of our simulation model has been disproven

Our interpretation of results

One should be sceptical of the inferences of Chang et al. (2015) and Bouckaert et al. (2012) until that at least

1. The realisticness of our simulation model has been disproven
2. It has been shown that that the methods give correct inferences on data sets produced by a variety of different realistic simulation models

Interpretation of underestimation

- Both Chang et al.'s and Bouckaert et al.'s underestimate the true tree age

Interpretation of underestimation

- Both Chang et al.'s and Bouckaert et al.'s underestimate the true tree age
- However, we believe this particular finding should not be overinterpreted – it does *not* imply that the true historic age of Indo-European must be high

Interpretation of underestimation

- Both Chang et al.'s and Bouckaert et al.'s underestimate the true tree age
- However, we believe this particular finding should not be overinterpreted – it does *not* imply that the true historic age of Indo-European must be high
- Using other realistic simulated data, maybe the methods would overestimate the tree age

Ideas for future work

Ideas for future work

- Try to isolate which parts of the methods are responsible for the wrong inferences

Ideas for future work

- Try to isolate which parts of the methods are responsible for the wrong inferences
  - We suspect that Covarion/CTMC models are generally unsuitable for age inferences on Multistate-generated data

Ideas for future work

- Try to isolate which parts of the methods are responsible for the wrong inferences
  - We suspect that Covarion/CTMC models are generally unsuitable for age inferences on Multistate-generated data
- Investigate under which circumstances Chang et al.'s and Bouckaert et al.'s methods *over*estimate the true tree age

Ideas for future work

- Try to isolate which parts of the methods are responsible for the wrong inferences
  - We suspect that Covarion/CTMC models are generally unsuitable for age inferences on Multistate-generated data
- Investigate under which circumstances Chang et al.'s and Bouckaert et al.'s methods *over*estimate the true tree age
- Build more refined realistic simulation models, including, for instance, semantic shifts and dialect continua

Ideas for future work

- Try to isolate which parts of the methods are responsible for the wrong inferences
  - We suspect that Covarion/CTMC models are generally unsuitable for age inferences on Multistate-generated data
- Investigate under which circumstances Chang et al.'s and Bouckaert et al.'s methods *over*estimate the true tree age
- Build more refined realistic simulation models, including, for instance, semantic shifts and dialect continua
- Try to evaluate more studies, such as Heggarty et al. (2023)

The end

Thank you for listening!

References I

Adams, D. Q. (2006). Some implications of the carbon-14 dating of Tocharian manuscripts. *The Journal of Indo-European Studies*, 34(3):381–389.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2013). Mapping the origins and expansion of the Indo-European language family – corrections and clarifications. *Science*, 342(6165):1446–1446.

References II

Carling, G. (2005). Proto-Tocharian, Common Tocharian, and Tocharian – on the value of linguistic connections in a reconstructed language. In Jones-Bley, K., Huld, M. E., Della Volpe, A., and Robbins Dexter, M., editors, *Proceedings of the Sixteenth Annual UCLA Indo-European Conference, Los Angeles, November 5–6, 2004*, pages 47–71. Institute for the Study of Man, Washington, DC.

Chang, W., Hall, D., Cathcart, C., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, pages 194–244.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*, 29(8):1969–1973.

Haspelmath, M. and Tadmor, U., editors (2009). *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

References III

Heggarty, P., Anderson, C., Scarborough, M., King, B., Bouckaert, R., Jocz, L., Kümmel, M. J., Jügel, T., Irslinger, B., Pooth, R., et al. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656):eabg0818.

Lane, G. S. (1966). On the Interrelationship of the Tocharian Dialects. In Birnbaum, H. and Puhvel, J., editors, *Ancient Indo-European Dialects*, pages 213–234. University of California Press, Berkeley.

Lees, R. B. (1953). The basis of glottochronology. *Language*, pages 113–127.

Olander, T. (2018). Connecting the Dots: The Indo-European Family Tree as a Heuristic Device. In Goldstein, D. M. and Jamison, S. W., editors, *Proceedings of the 29th Annual UCLA Indo-European Conference*, pages 181–202. Hempen, Bremen.

References IV

Peyrot, M. (2022). Tocharian. In Olander, T., editor, *The Indo-European Language Family: A Phylogenetic Perspective*, pages 83–101. Cambridge University Press, Cambridge.

Pinault, G.-J. (2002). Tocharian and Indo-Iranian: Relations between two linguistic areas. In Sims-Williams, N., editor, *Indo-Iranian Languages and Peoples*, pages 243–284. Oxford University Press, Oxford.

Ringe, D. (1995). Tocharians in xinjiang: The linguistic evidence. *Journal of Indo-European Studies*, 23(3-4):439–444.

Ringe, D., Warnow, T., and Taylor, A. (2002). Indo-european and computational cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Swadesh, M. (1952). Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

References V

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

Warnow, T., Evans, S. N., Ringe, D., and Nakhleh, L. (2004). Stochastic models of language evolution and an application to the Indo-European family of languages. Download at `http://www.stat.berkeley.edu/users/evans/659.pdf`.