Why we need more study of methods, not data, in computational historical linguistics

Philipp Rönchen[1] Tilo Wiklund[2]

[1]Department of Linguistics and Philology, Uppsala University, Sweden, philipp.ronchen@lingfil.uu.se
[2]Chief Data Scientist, UAB Sensmetry, previously Department of Mathematics, Uppsala University, Sweden

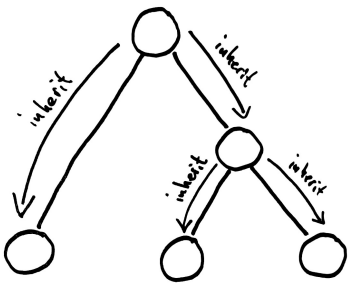
Maeiqcl - 25 February 2021

Starting point: Different conclusions

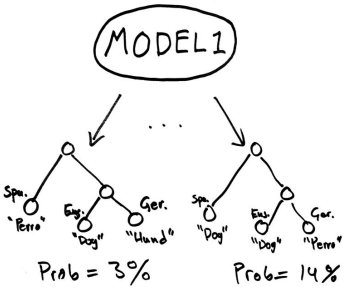
- Researchers have come to different conclusions using computational methods in historical linguistics
- Most famous example: The age of Indo-European, where Bouckaert et al. (2012) reached a different conclusion than Chang et al. (2015) even though they used rather similar methods
- Growing body of evidence that computational methods are less "stable" than previously thought (cf. Rama 2018, Ritchie and Ho 2019, Maurits et al. 2020)

Cognate data/family trees

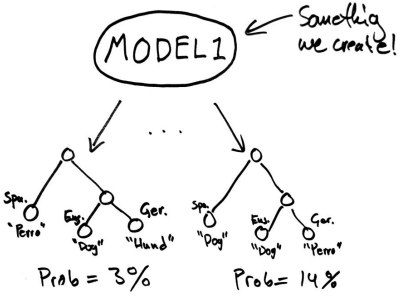
Language	Meaning 1	Meaning 2	...
English	three	dog	...
German	drei	Hund	...
Spanish	tres	perro	...



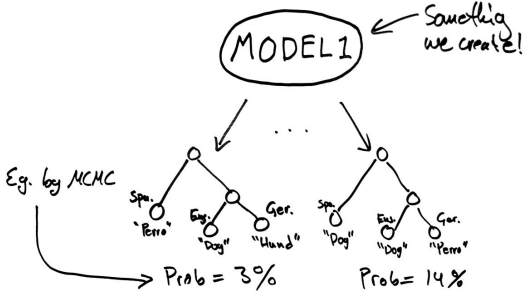
Likelihood inference



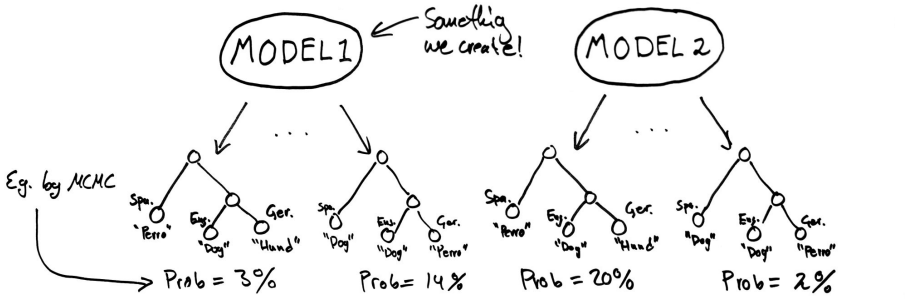
Likelihood inference



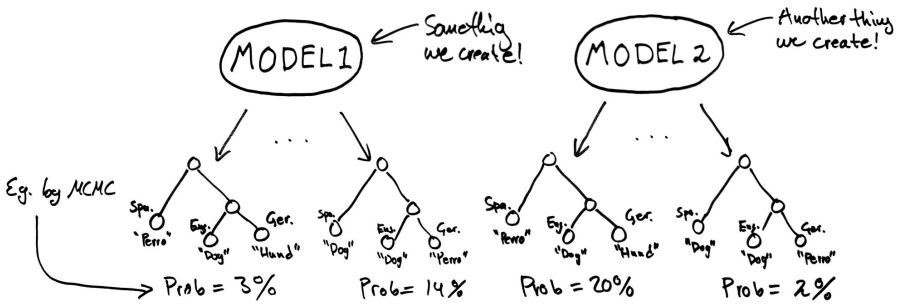
Likelihood inference



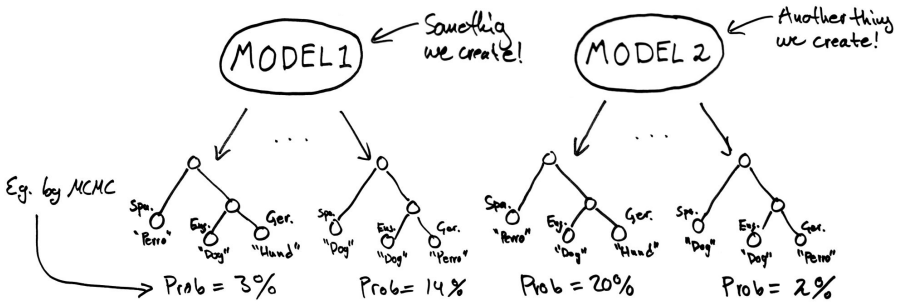
Likelihood inference



Likelihood inference

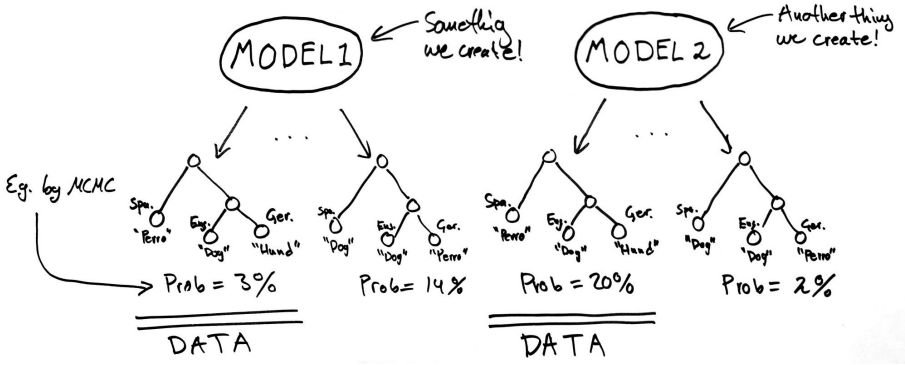


Likelihood inference



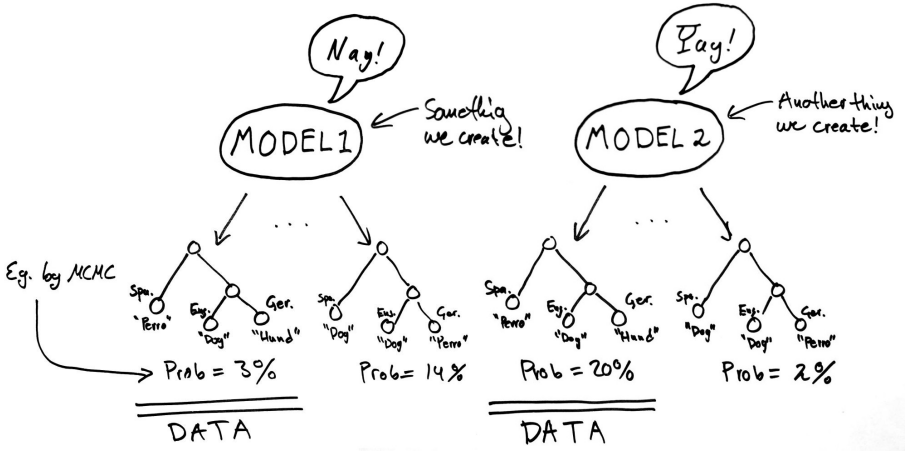
Different evolution dynamics, tree ages, topologies, ...

Likelihood inference



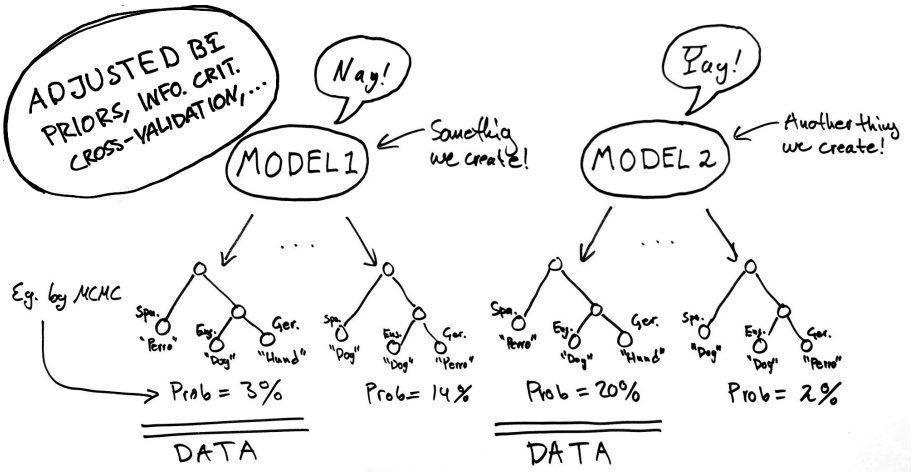
Different evolution dynamics, tree ages, topologies, . . .

Likelihood inference



Different evolution dynamics, tree ages, topologies, ...

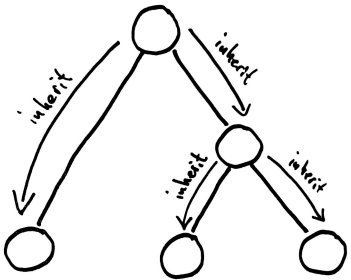
Likelihood inference



Different evolution dynamics, tree ages, topologies, ...

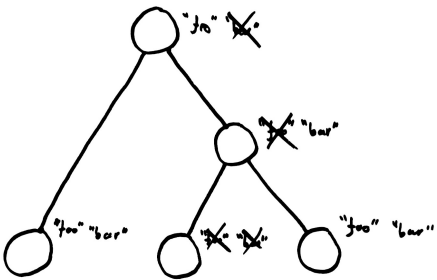
Cognate evolution models

Let's look at some examples of cognate evolution models



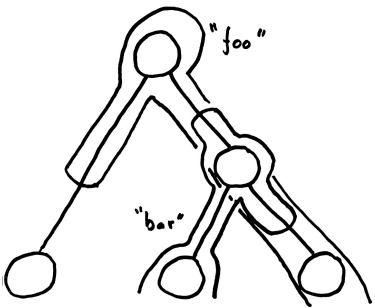
Cognate evolution models

Binary CTMC model



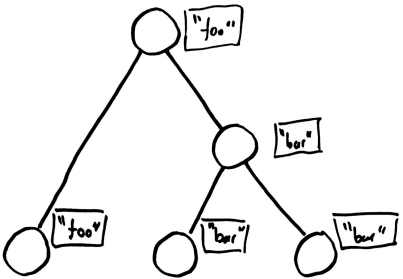
Cognate evolution models

Stochastic Dollo model



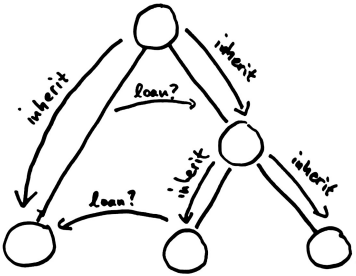
Cognate evolution models

Multistate model (with unlimited states)



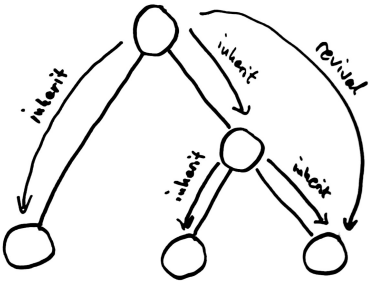
Cognate evolution models

A model with loans

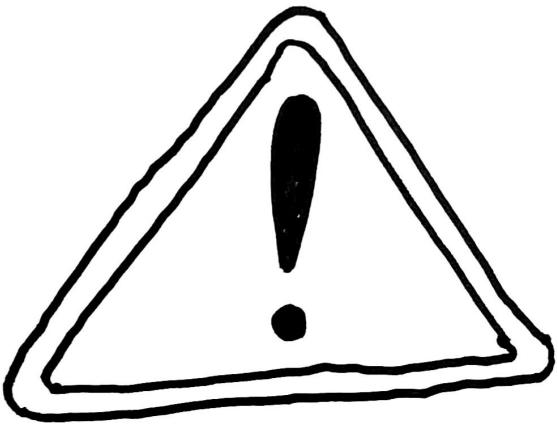


Cognate evolution models

A model with archaic words being "revived"

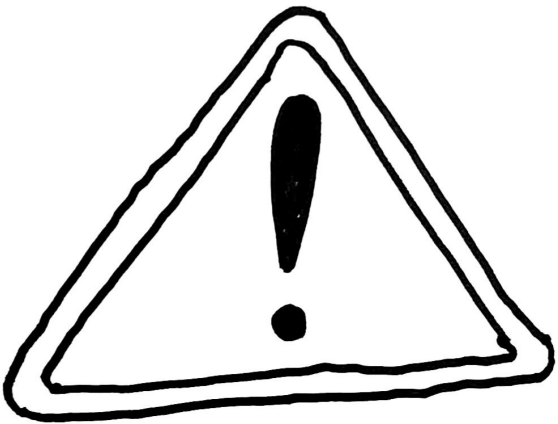


Issues with likelihood inference



- Schematic (but we think valid)
- Not the first to observe in general!

Issues with likelihood inference



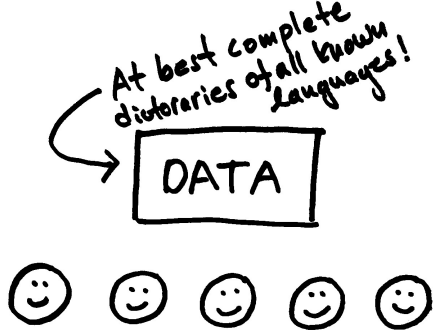
- Schematic (but we think valid)
- Not the first to observe in general!
- Took us time to clarify our thoughts!

How (linguistic) science works

One data, many people, many studies

How (linguistic) science works

One data, many people, many studies

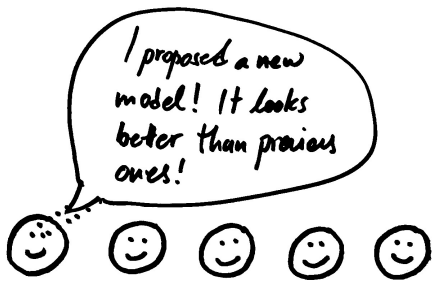


How (linguistic) science works

One data, many people, many studies

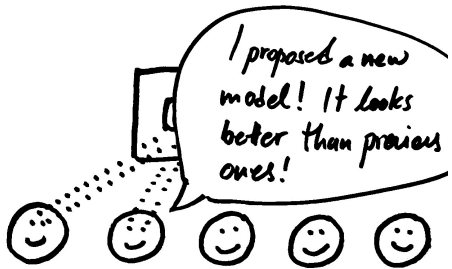
How (linguistic) science works

One data, many people, many studies



How (linguistic) science works

One data, many people, many studies

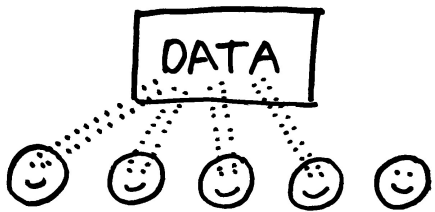


How (linguistic) science works

One data, many people, many studies

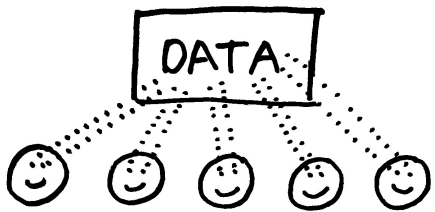
How (linguistic) science works

One data, many people, many studies



How (linguistic) science works

One data, many people, many studies



Problematic if “purely data driven“

If model selection is based only on likelihood:

- Too many models get tested
- Likelihoods (like probabilities) contain randomness
- If the likelihoods of enough models are compared, sooner or later one could find an “accidentally” well fitting model

Problematic if “purely data driven“

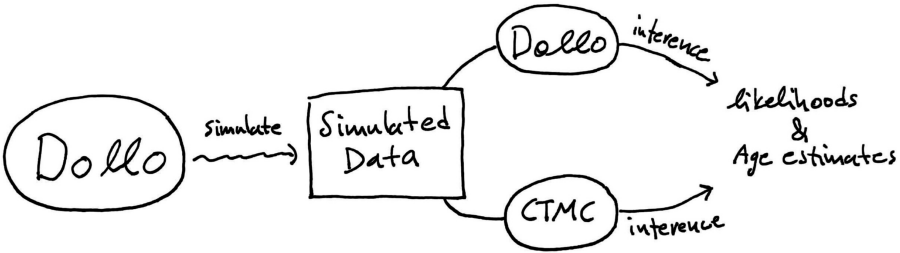
If model selection is based only on likelihood:

- Too many models get tested
- Likelihoods (like probabilities) contain randomness
- If the likelihoods of enough models are compared, sooner or later one could find an “accidentally” well fitting model
- Like meta-overfitting or multiple testing

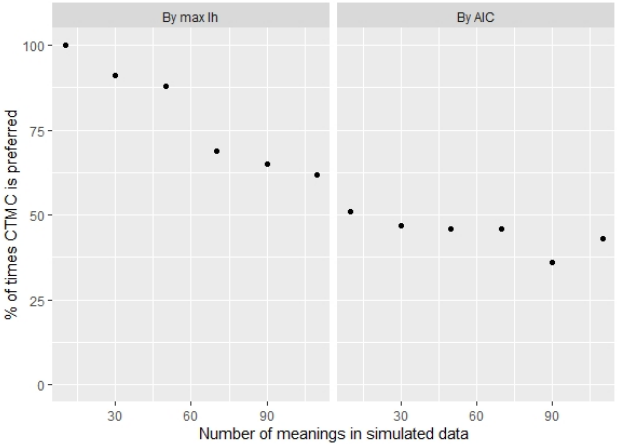
Small experiment

Imagine a perfect world where a certain model (Dollo) is actually true.

What conclusions would someone applying another popular model (CTMC) make?



Simulation results, SD/CTMC



Quite often, the wrong model looks better, even when adjusting for model complexity!

Caveats: tested only for small tree and some parameters, also hard to compare different models correctly

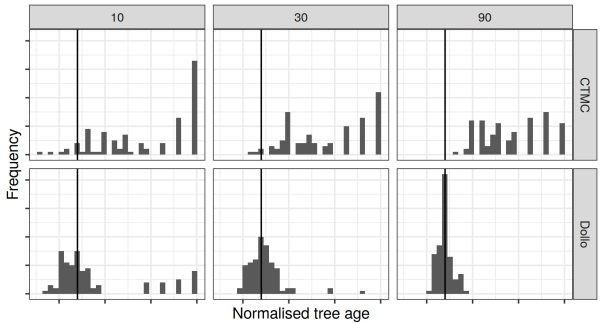
Looks often better, but wrong inference

Does it matter, just two descriptions of the same thing?

Looks often better, but wrong inference

Does it matter, just two descriptions of the same thing?

No, leads to wrong conclusions on tree height!



MAY FIT BETTER without giving good inferences!

Many studies = multiple testing

Even if on "real" data a "wrong" model looks good only very seldom (say 1% of times), we have a problem - there are too many models that can be tried

Many studies = multiple testing

Even if on "real" data a "wrong" model looks good only very seldom (say 1% of times), we have a problem - there are too many models that can be tried

- Different evolution models: Stochastic Dollo

$$\text{Probability of “bad inference”} = 1 - 0.99 = 0.01$$

Many studies = multiple testing

Even if on "real" data a "wrong" model looks good only very seldom (say 1% of times), we have a problem - there are too many models that can be tried

- Different evolution models: Stochastic Dollo , CTMC

$$\text{Probability of “bad inference”} = 1 - (0.99 \cdot 0.99) \approx 0.02$$

Many studies = multiple testing

Even if on "real" data a "wrong" model looks good only very seldom (say 1% of times), we have a problem - there are too many models that can be tried

- Different evolution models: Stochastic Dollo , CTMC , Covarion

$$\text{Probability of “bad inference”} = 1 - (0.99 \cdot 0.99 \cdot 0.99) \approx 0.04$$

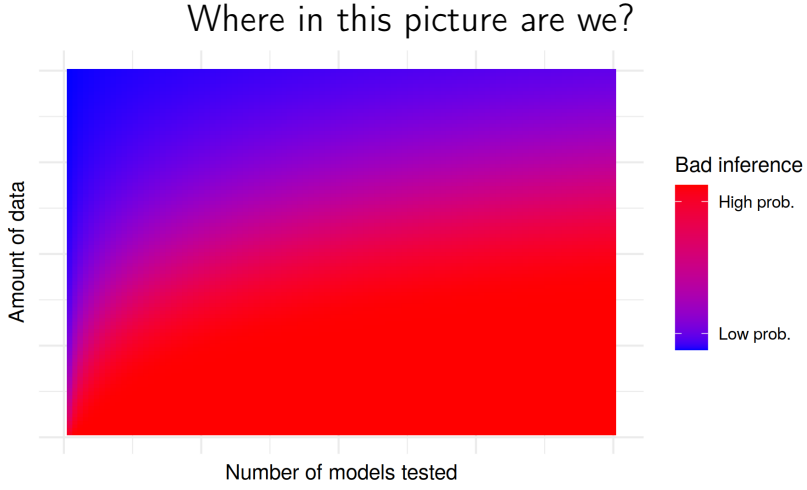
Many studies = multiple testing

Even if on "real" data a "wrong" model looks good only very seldom (say 1% of times), we have a problem - there are too many models that can be tried

- Different evolution models: Stochastic Dollo , CTMC , Covarion, etc. . . .
- Different tree assumptions: Topology constraints, age constraints, treatment of poorly attested languages . . .
- Different tree priors, parameter priors . . .

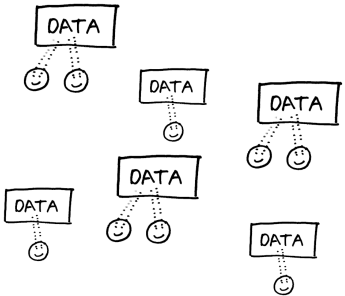
$$\text{Probability of “bad inference”} = 1 - (0.99 \cdot 0.99 \cdot 0.99 \cdot \cdots \cdot) \approx \text{high!}$$

More data = better, more models/people = worse



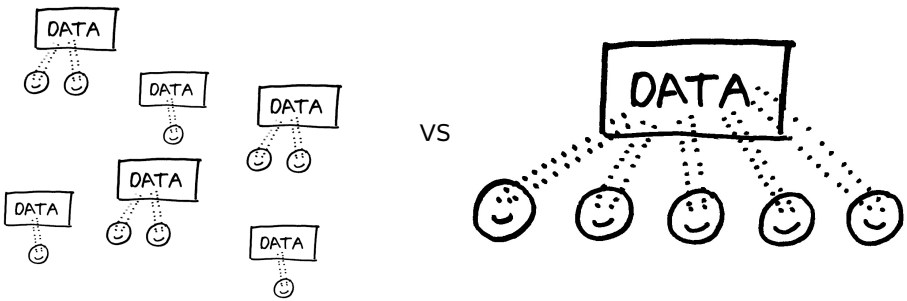
Difference to, e.g. chemistry: No replication

- In much of e.g. chemistry, psychology you can have replication studies



Difference to, e.g. chemistry: No replication

- In much of e.g. chemistry, psychology you can have replication studies
- In historical linguistics just one realisation of the data ("only one history of languages")



Data is valuable

The *less often* we use our data, the *stronger* our inferences get!

Are we lost? No!

Some suggestions

- Report more negative results
- Evaluate models not only by (adjusted) likelihood (i. e. not purely "by data")
- Do simulation studies
- Do analytic studies
- Build linguistically motivated models

Build linguistically motivated models

- We have reasons other than data to trust linguistically justified models.
- If viable for inference, more trustworthy inference, not just "what the data says".
- Even if inference not viable, can be used for simulation studies!

Ex linguistically motivated models for inference

- Ex 1: Kelly and Nicholls (2017) have found a way to systematically augment the Stochastic Dollo model to allow for loanwords (on small trees)
- Ex 2: We have developed an inference algorithm for a Multistate model of linguistic evolution (but we cannot yet deal with loanwords)

Build linguistically motivated models for simulation

Even if we inference not viable we can still use them as benchmarks for other models!

- Example from before: We believe the Stochastic Dollo model and the Multistate model to be realistic enough that for *any method (e.g. CTMC/Covarion) to be considered trustworthy* it should work for data generated by them!

Build linguistically motivated models for simulation

Even if we inference not viable we can still use them as benchmarks for other models!

- Example from before: We believe the Stochastic Dollo model and the Multistate model to be realistic enough that for *any method (e.g. CTMC/Covarion) to be considered trustworthy* it should work for data generated by them!
  (Does not imply Stochastic Dollo/Multistate are robust enough to be used for inference!)

Build linguistically motivated models for simulation

Even if we inference not viable we can still use them as benchmarks for other models!

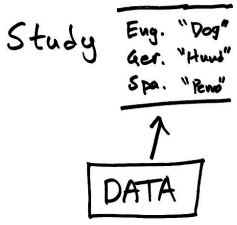
- Example from before: We believe the Stochastic Dollo model and the Multistate model to be realistic enough that for *any method (e.g. CTMC/Covarion) to be considered trustworthy* it should work for data generated by them!
  (Does not imply Stochastic Dollo/Multistate are robust enough to be used for inference!)
- See Bradley (2016) and Murawaki (2015) for more examples of evaluating models by simulation studies

Project proposal: Simulation replication

Take existing studies and try to replicate on simulation

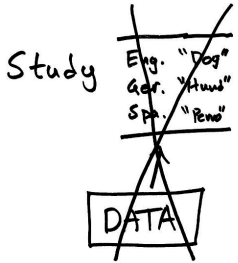
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



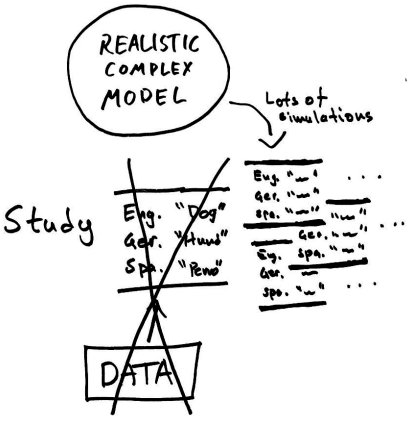
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



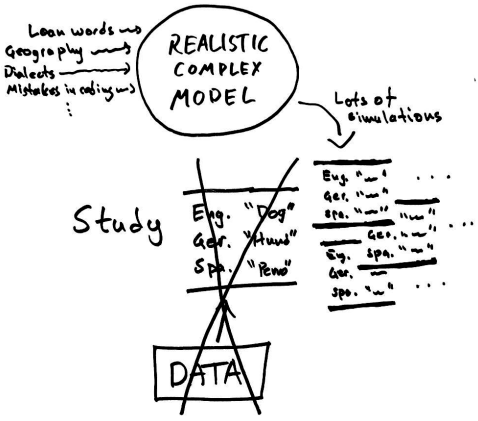
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



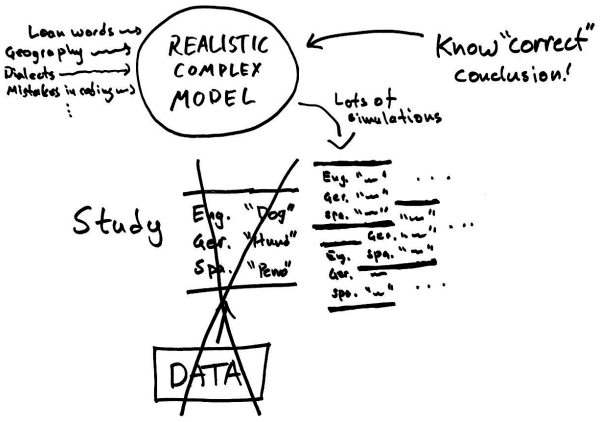
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



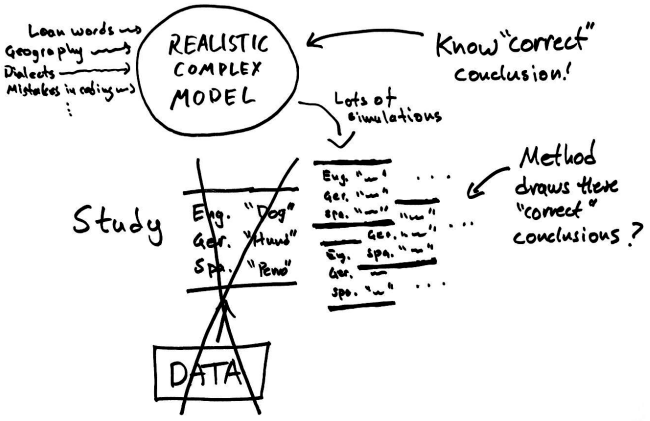
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



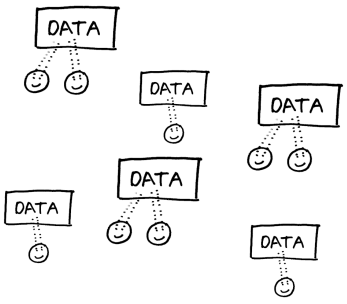
Project proposal: Simulation replication

Take existing studies and try to replicate on simulation



Why doing simulations is safe

- Simulations can always be rerun/replicated
- Potentially "infinite amounts" of simulated data
- Therefore simulations a powerful way to test the consequences of model assumptions, robustness, and reliability

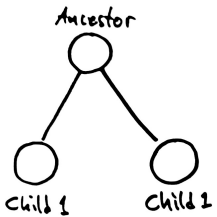


Analytic studies of models

- Even better if we can prove analytic or numeric results (no data, no simulation).
- If models always give same inference (no matter observation) $\Rightarrow$ no multiple model issue.

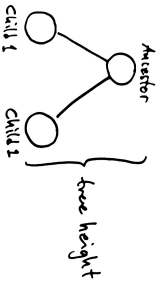
Analytic studies of models

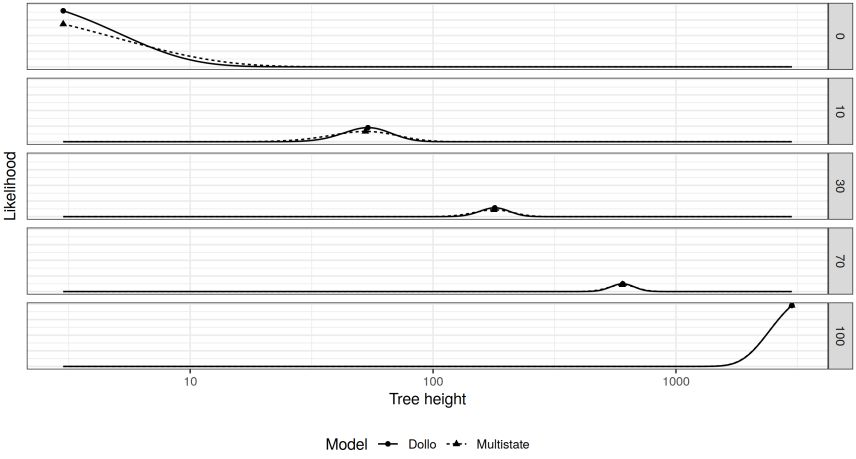
- Even better if we can prove analytic or numeric results (no data, no simulation).
- If models always give same inference (no matter observation) $\Rightarrow$ no multiple model issue.
- Ex: Small study to compare Multistate and Dollo analytically for tiny Cherry tree.



Ex: Interchangeability of models

Likelihood of different tree heights for different potential observations.





References I

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.

Bradley, S. (2016). Synthetic language generation and model validation in beast2. *arXiv preprint arXiv:1607.07931*.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, 91(1):194–244.

Kelly, L. J. and Nicholls, G. K. (2017). Lateral transfer in stochastic dollo models. *The Annals of Applied Statistics*, 11(2):1146–1168.

Maurits, L., de Heer, M., Honkola, T., Dunn, M., and Vesakoski, O. (2020). Best practices in justifying calibrations for dating language families. *Journal of Language Evolution*, 5(1):17–38.

References II

Murawaki, Y. (2015). Spatial structure of evolutionary models of dialects in contact. *Plos one*, 10(7):e0134335.

Rama, T. (2018). Three tree priors and five datasets: A study of Indo-European phylogenetics. *Language Dynamics and Change*, 8(2):182–218.

Ritchie, A. M. and Ho, S. Y. (2019). Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. *Journal of Language Evolution*, 4(2):108–123.