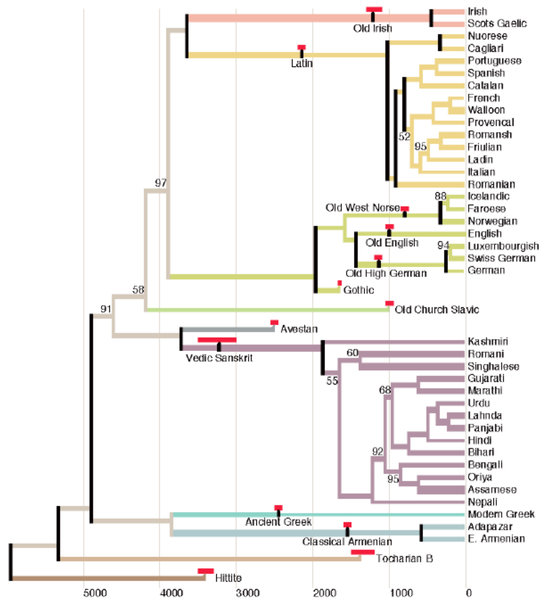
Principal Component Analysis as a Sanity Check for Bayesian Phylolinguistic Reconstruction

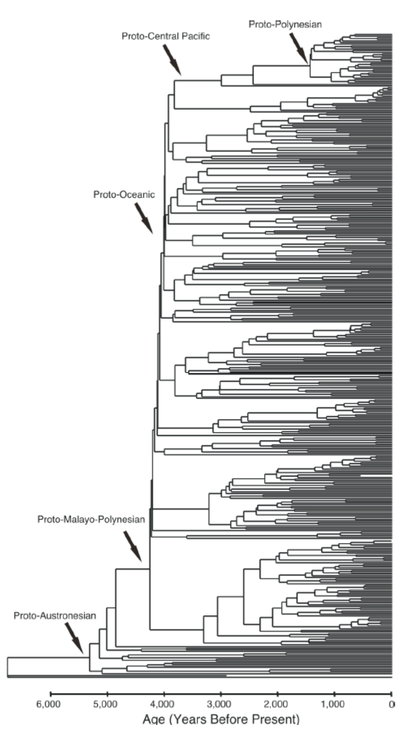
Yugo Murawaki Kyoto University

1. Bayesian Phylolinguistic Reconstruction

- Not a popular topic in NLP but primarily studied by evolutionary biologists
- Compute-heavy statistical method to applied to binary-coded lexical data
- Popular applications:



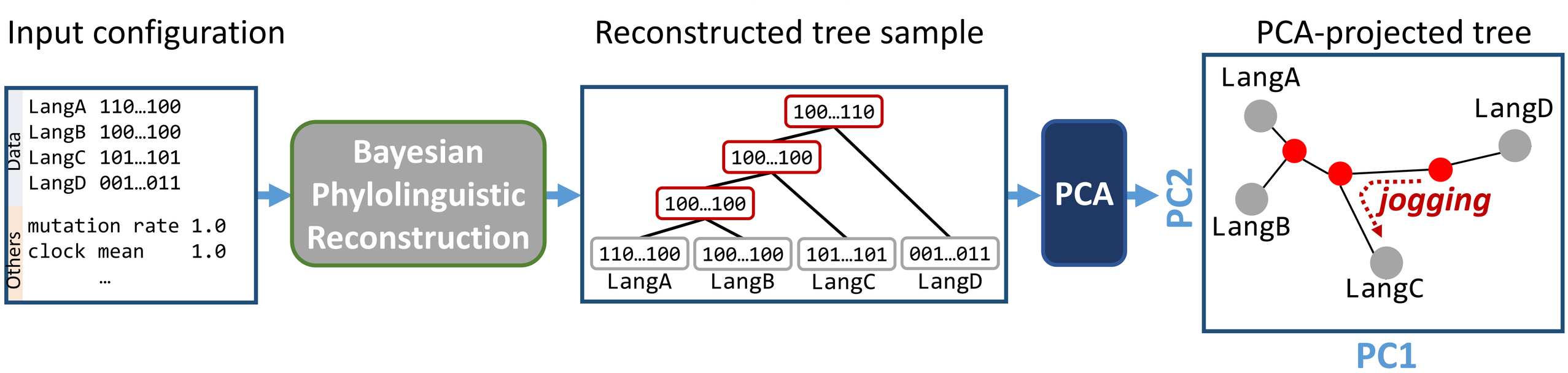
Indo-European root age
[Chang+, 2015]



Austronesian expansion
[Greenhill&Gray, 2009]

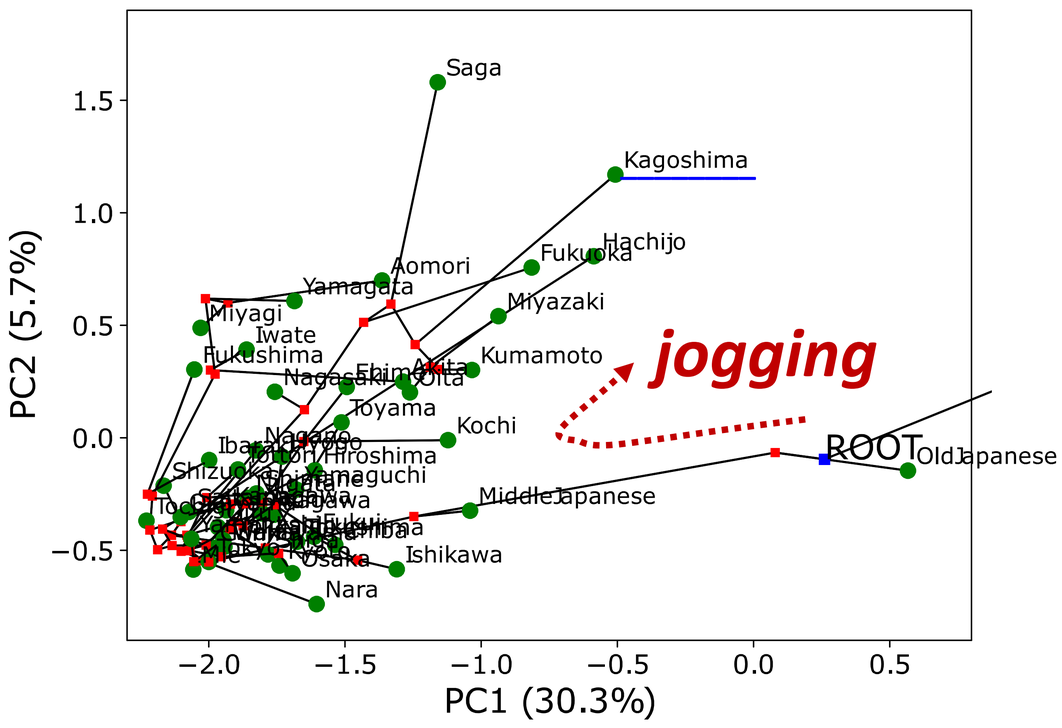
- The tree assumption is often violated to varying degrees
  - Assumption: Independent evolution after a branching event
  - Reality: Horizontal transmission obscures vertical signals
- Previous efforts to explore the boundaries of the tree model's applicability have two limitations:
  1. Rely on distance-based (not Bayesian) analytical tools
  2. Relative lack of uncertainty during inference is misconstrued as support for the tree model

2. Principal Component Analysis (PCA) as a Sanity Check



- Slightly edit the input config file
  - Often published as a supplementary material
- Run MCMC inference as usual
- Choose a single tree sample
- First apply PCA to the states of the leaf nodes
- Then project internal nodes
- Lexical data must roughly follow a unidirectional pattern along PC1
- Violations as jogging

3. Analysis of Real Data (Japonic as an Example)



- Lee and Hasegawa (2011) analyzed closely-related mainland Japanese dialects under intense contact
- PCA projection effectively highlights anomalies
  - Kagoshima at the southwestern tip of the mainland
    - Closest to Old Japanese along PC1
    - Second to last in terms of overall similarities
- Possible explanation:
  - Kagoshima was less affected by dialect leveling
    - Relatively rapid overall change
    - Retained features indicate archaism