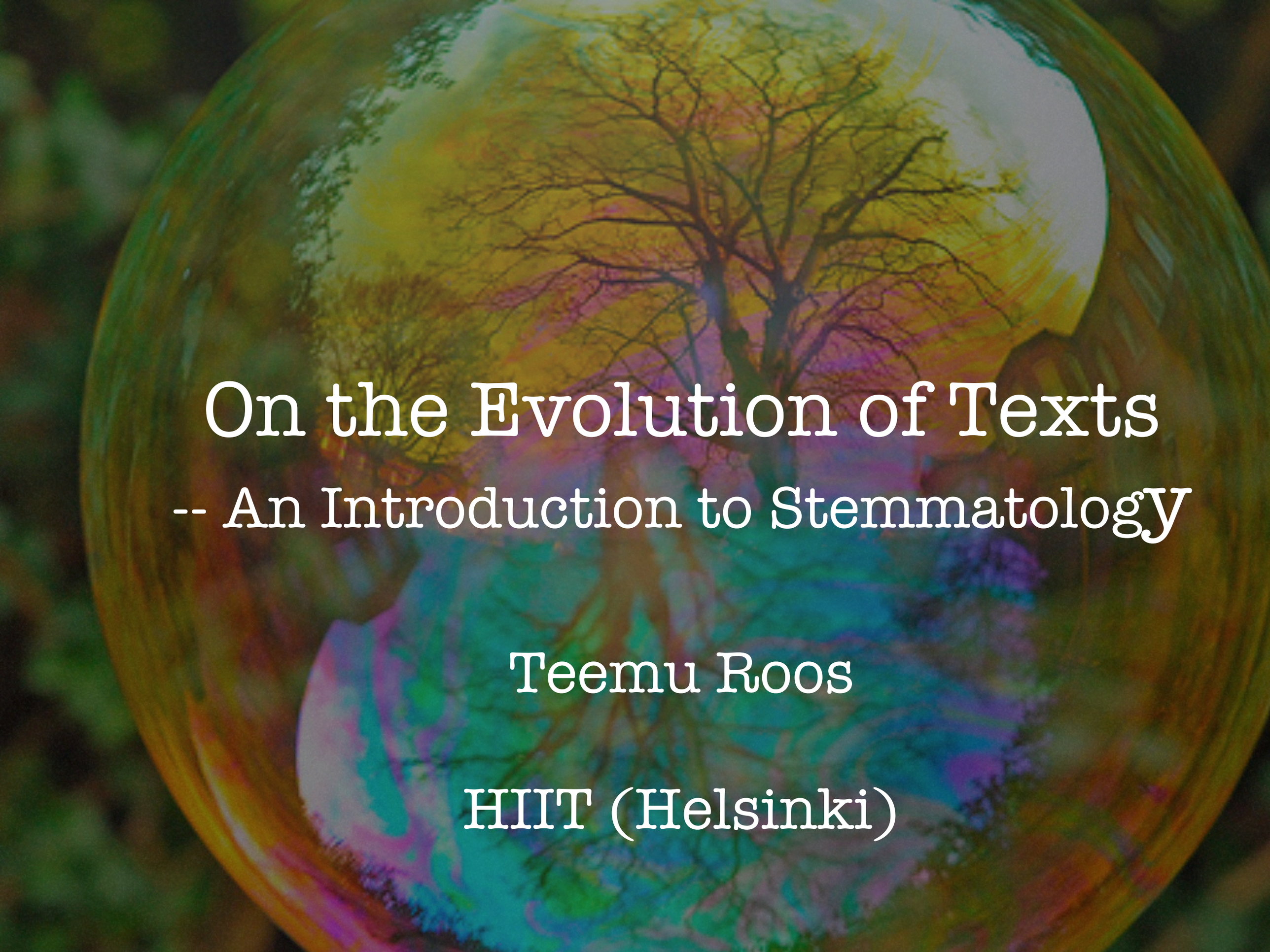


On the Evolution of Texts

-- An Introduction to Stemmatology

Teemu Roos

HIIT (Helsinki)

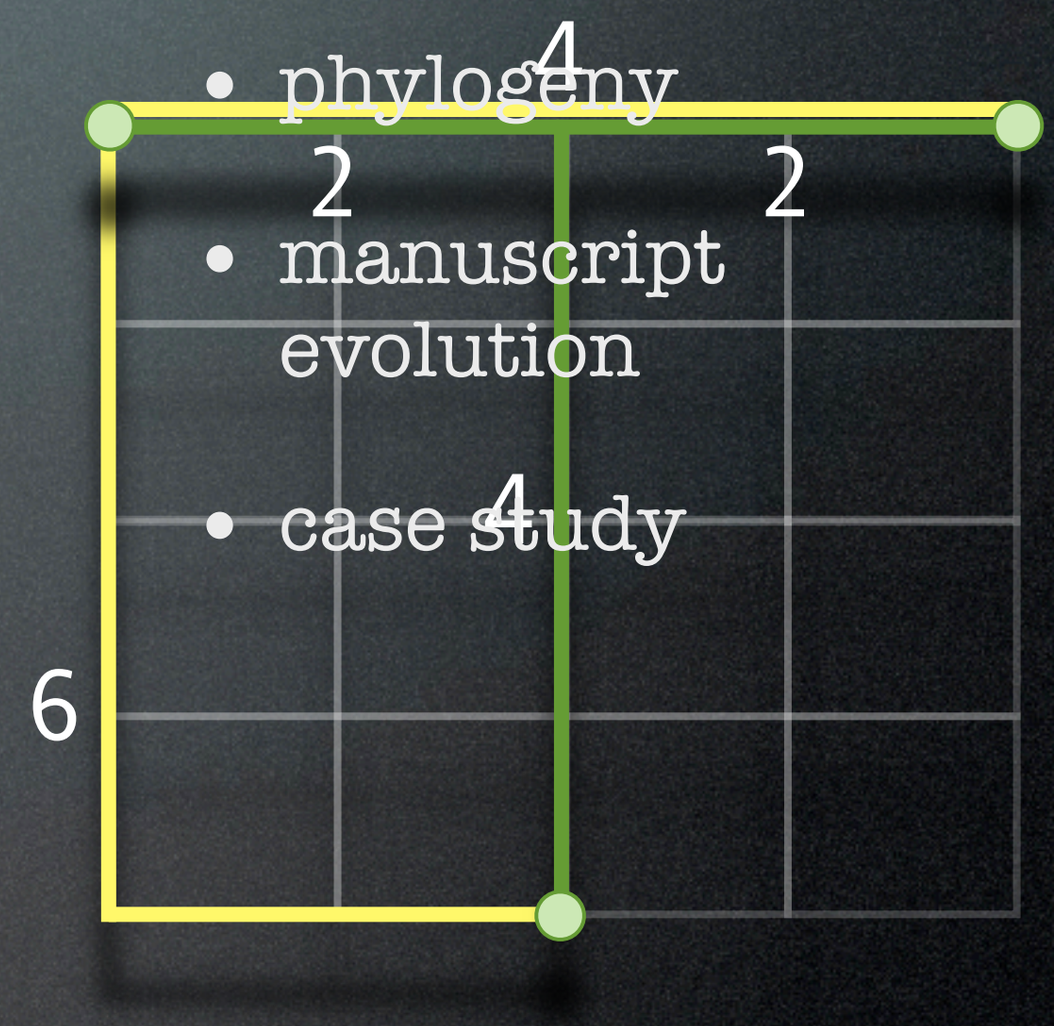
Plan

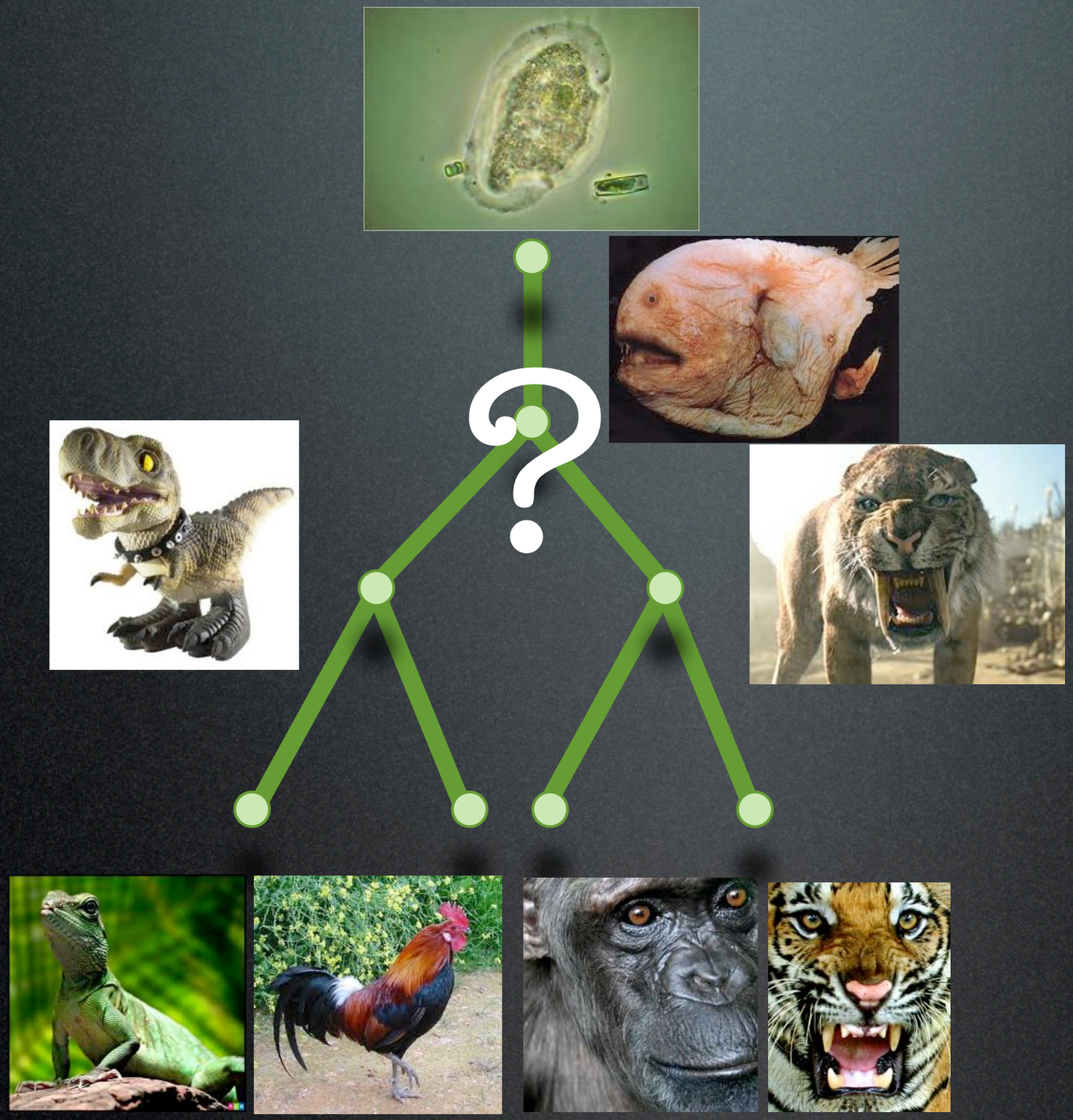
1. Finding trees:

- spanning trees
- Steiner trees
- structural EM

2. Evolution of texts:

- phylogeny
- manuscript evolution
- case study





Generative model

Typical assumptions:

1. tree structure $T = (V, E)$, $|V| = n+1$
2. $p(\text{node} = x) = p_x$
3. $p(\text{child} = x \mid \text{parent} = y) = p_{y \to x}$
4. $p_y \, p_{y \to x} = p_x \, p_{x \to y} = p_{x,y}$

$$p(x_0,\ldots,x_n \mid T) = p_{x_0} \prod_{i=1}^{n} p_{pa_i \to x_i}$$

p(x0) prod p(xi,pai) / p(pai)
= prod p(xi) prod p(xi,xj) / p(xi)p(xj)

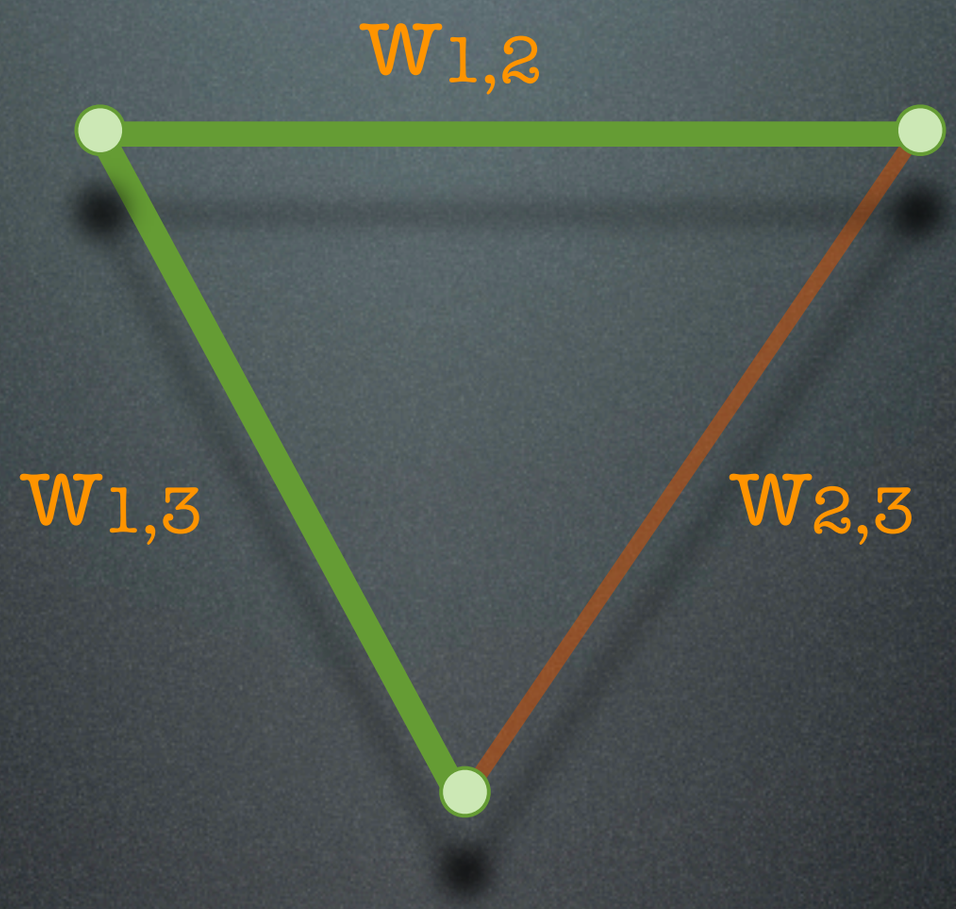
$$= \Big[\prod_{i \in V} p_{x_i} \Big] \Big[\prod_{(i,j) \in E} p_{x_i,x_j} / (p_{x_i} \, p_{x_j}) \Big]$$

$$\log \left[\prod_{i \in V} p_{x_i} \right] \left[\prod_{(i,j) \in E} p_{x_i,x_j} / (p_{x_i}\, p_{x_j}) \right]$$

$$= \sum_{(i,j) \in E} \underbrace{\log p_{x_i,x_j} / (p_{x_i}\, p_{x_j})}_{w_{i,j}} + \text{const}$$

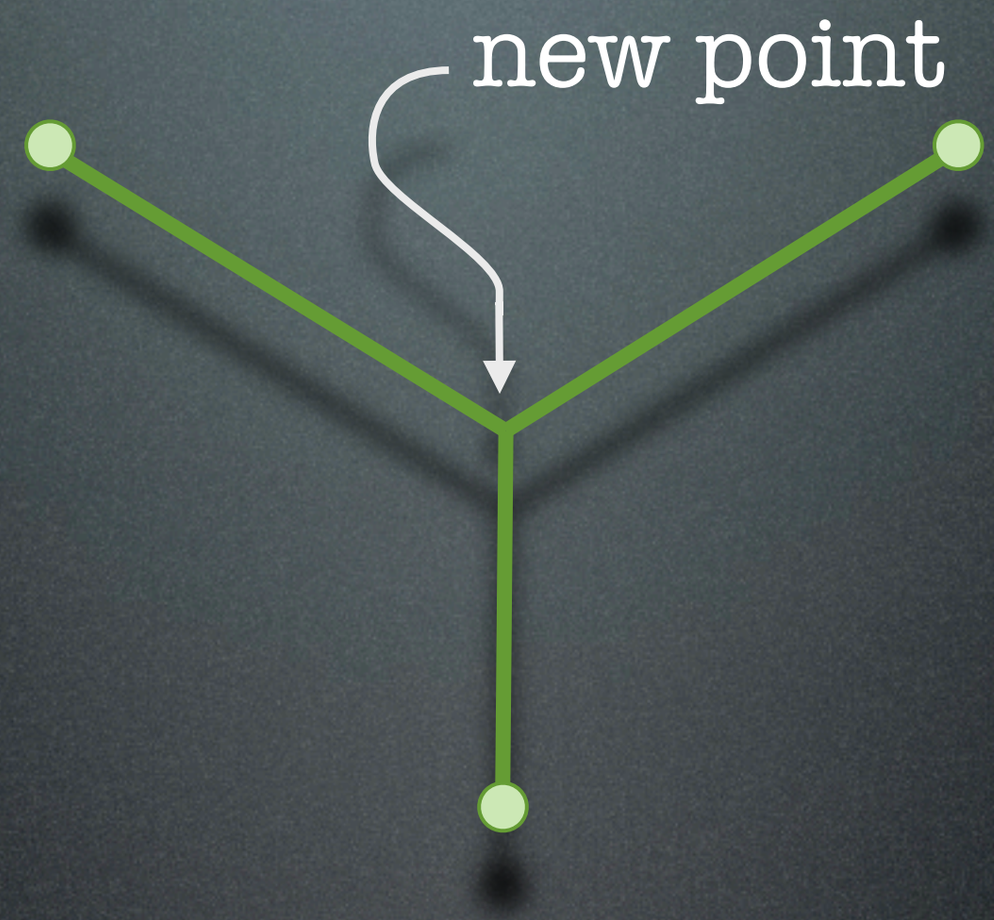
$$\log P(x_0,\ldots,x_n \mid T) = \sum_{(i,j) \in E} w_{i,j} + \text{const}$$

example: $w_{i,j} = \begin{cases} \alpha & \text{if } x_i = x_j \quad (\alpha > 0) \\ \beta & \text{if } x_i \neq x_j \quad (\beta < 0) \end{cases}$

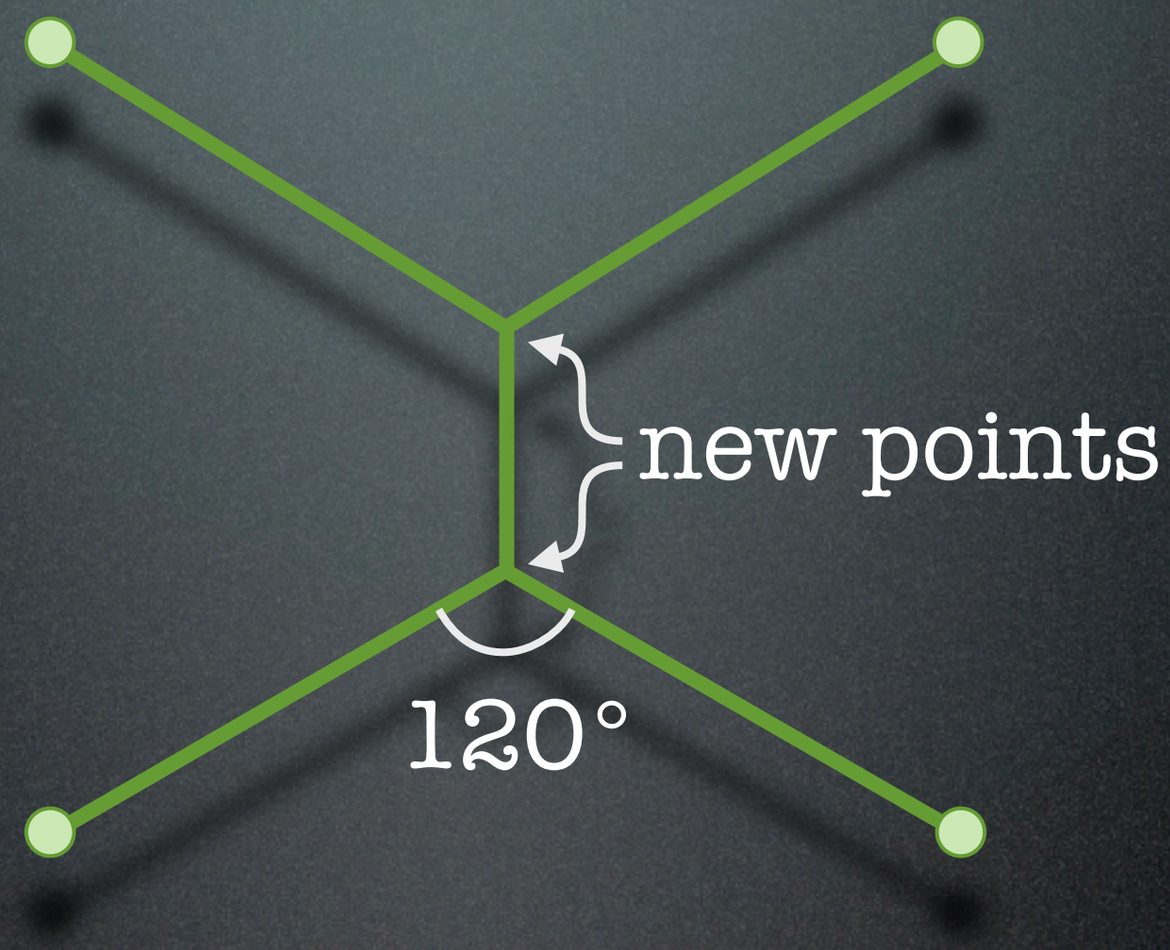


$$MST = \max_{T} P(x_0,\ldots,x_n \mid T) = \min_{T} |(i,j) \in E : x_i \neq x_j|$$

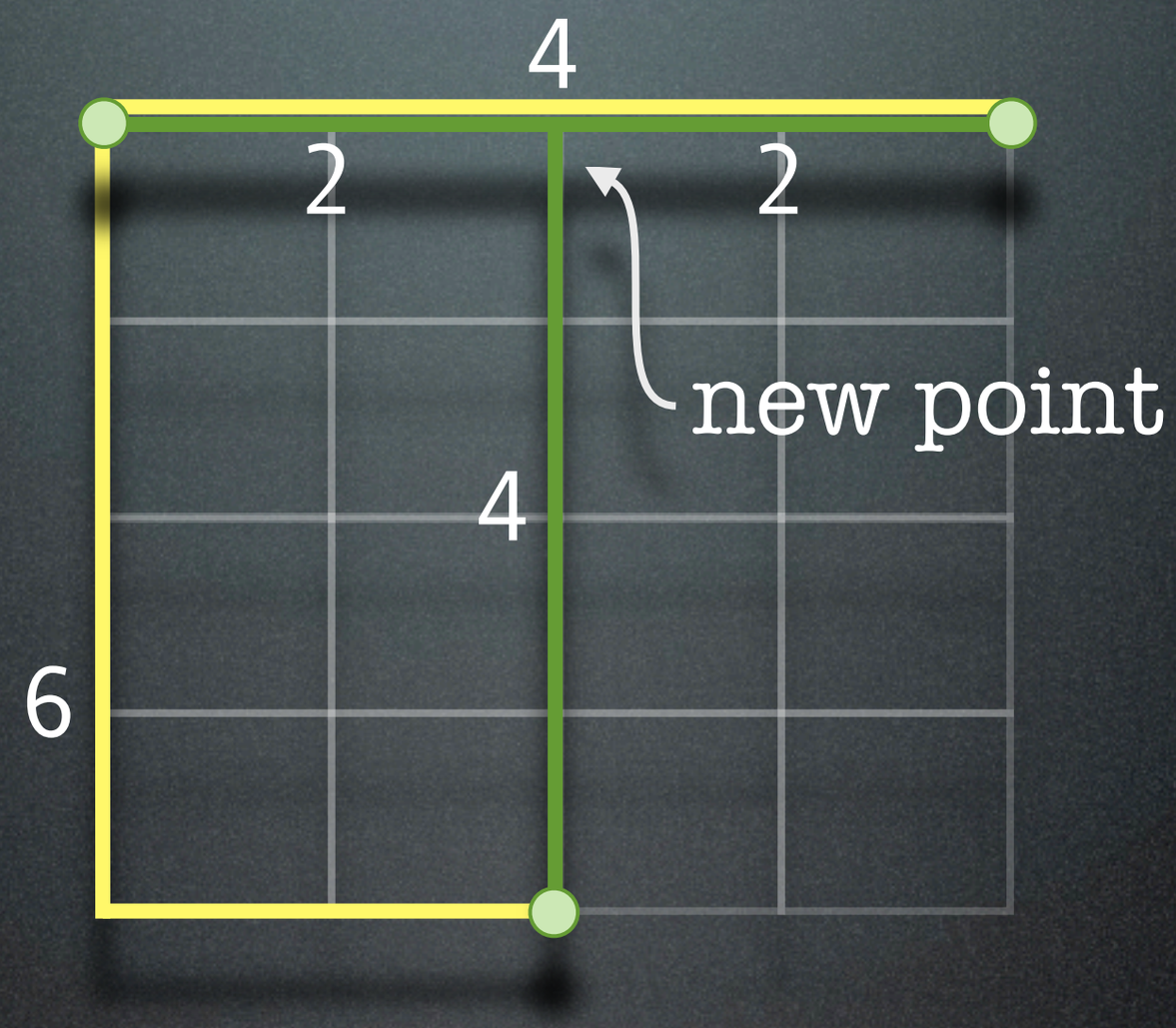
spanning trees



Steiner tree problem



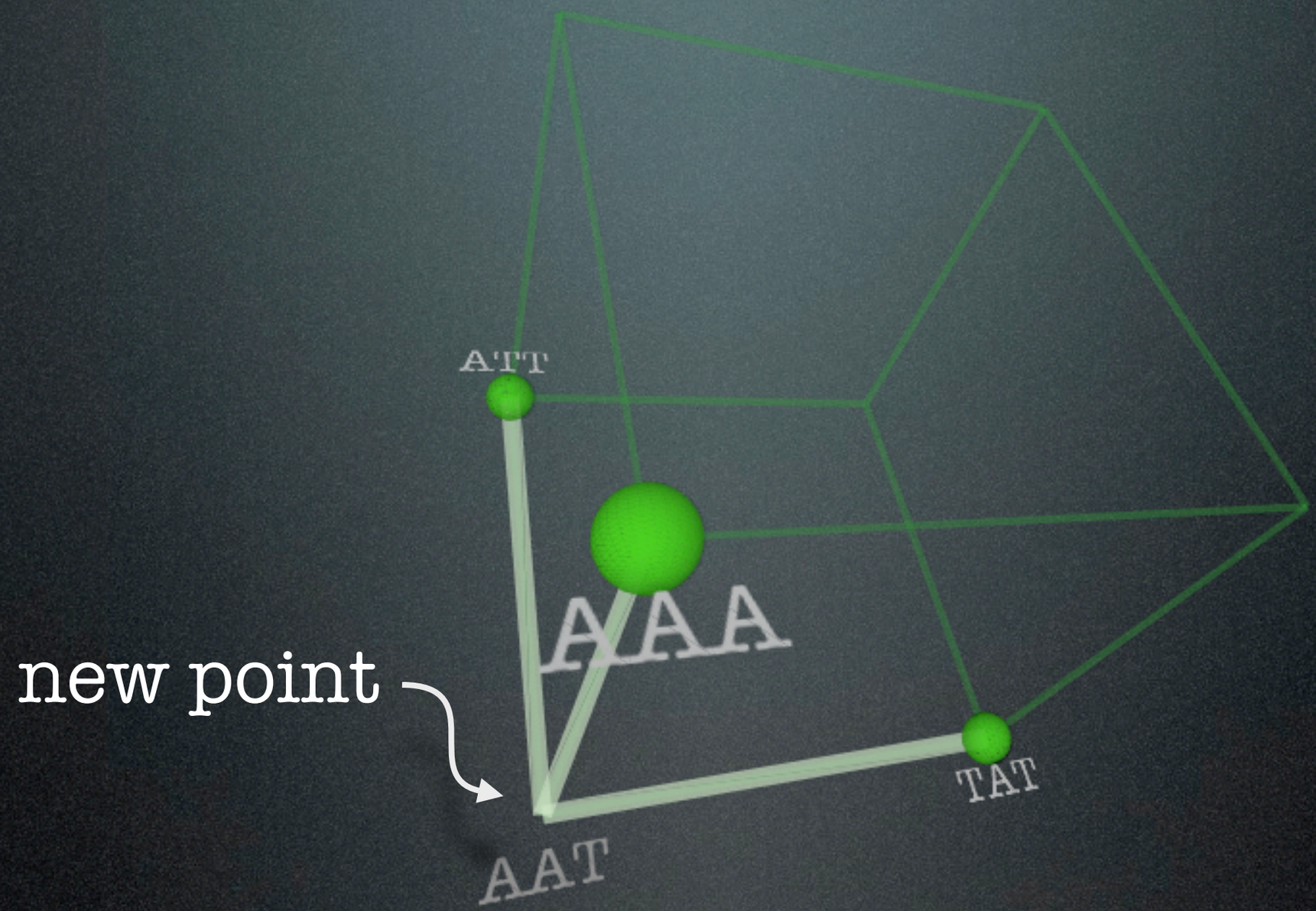
Steiner tree problem



Steiner
8

MST
10

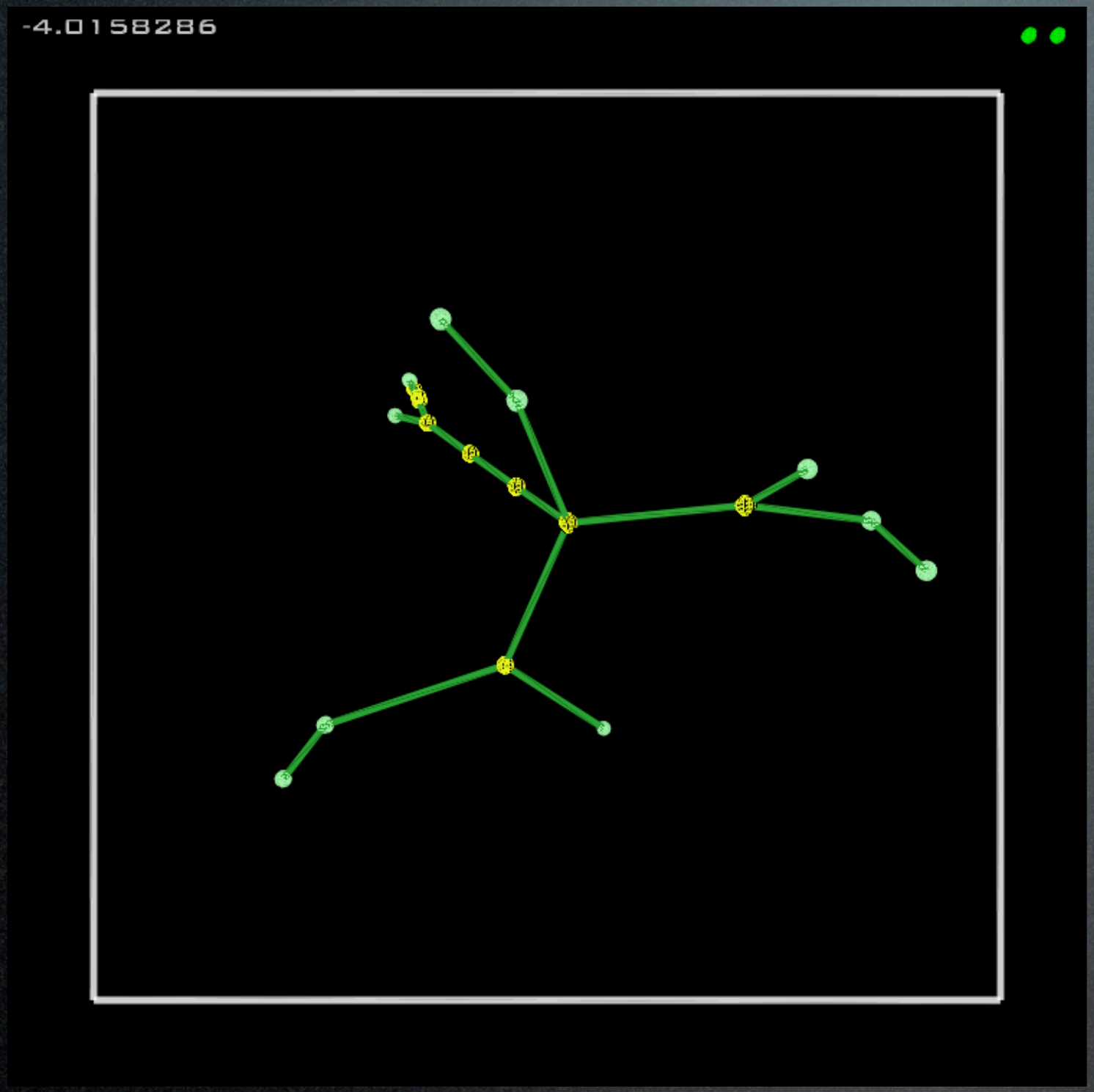
Steiner tree problem



Steiner
3

MST
4

Steiner tree problem



Step 0: Random initial tree

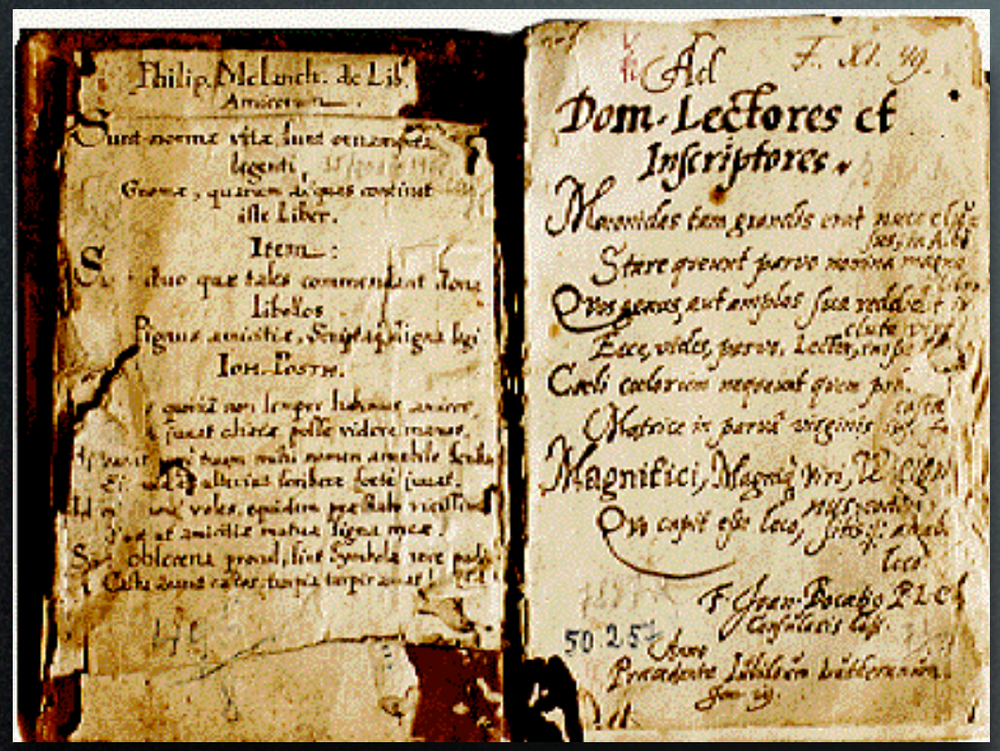
Step 1: Inference

Step 2: MST

Goto step 1

structural EM

Plan



2. Evolution of texts:

- tree of life
- manuscripts
- data sets

Statistical Inference of Phylogenies

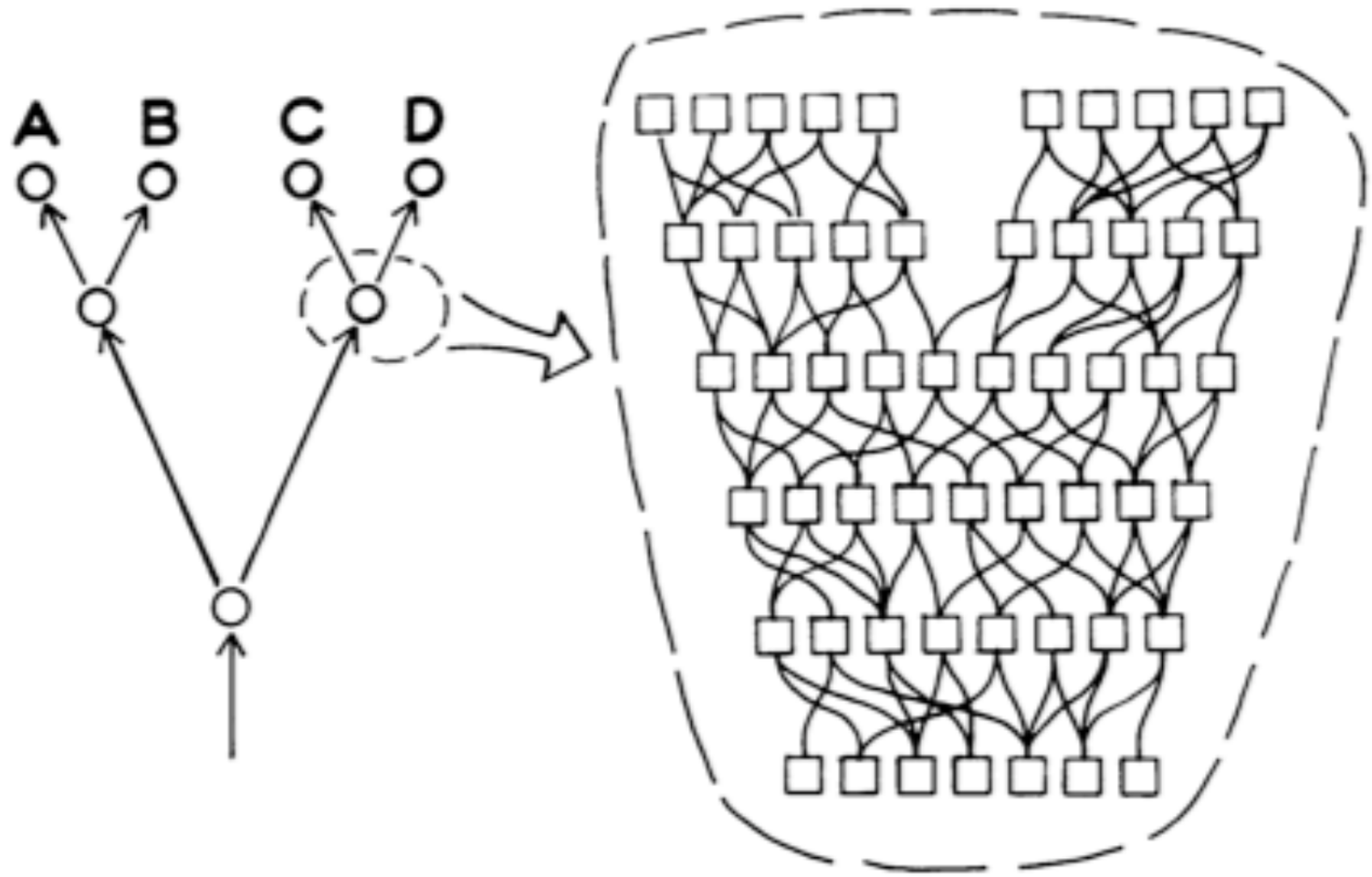
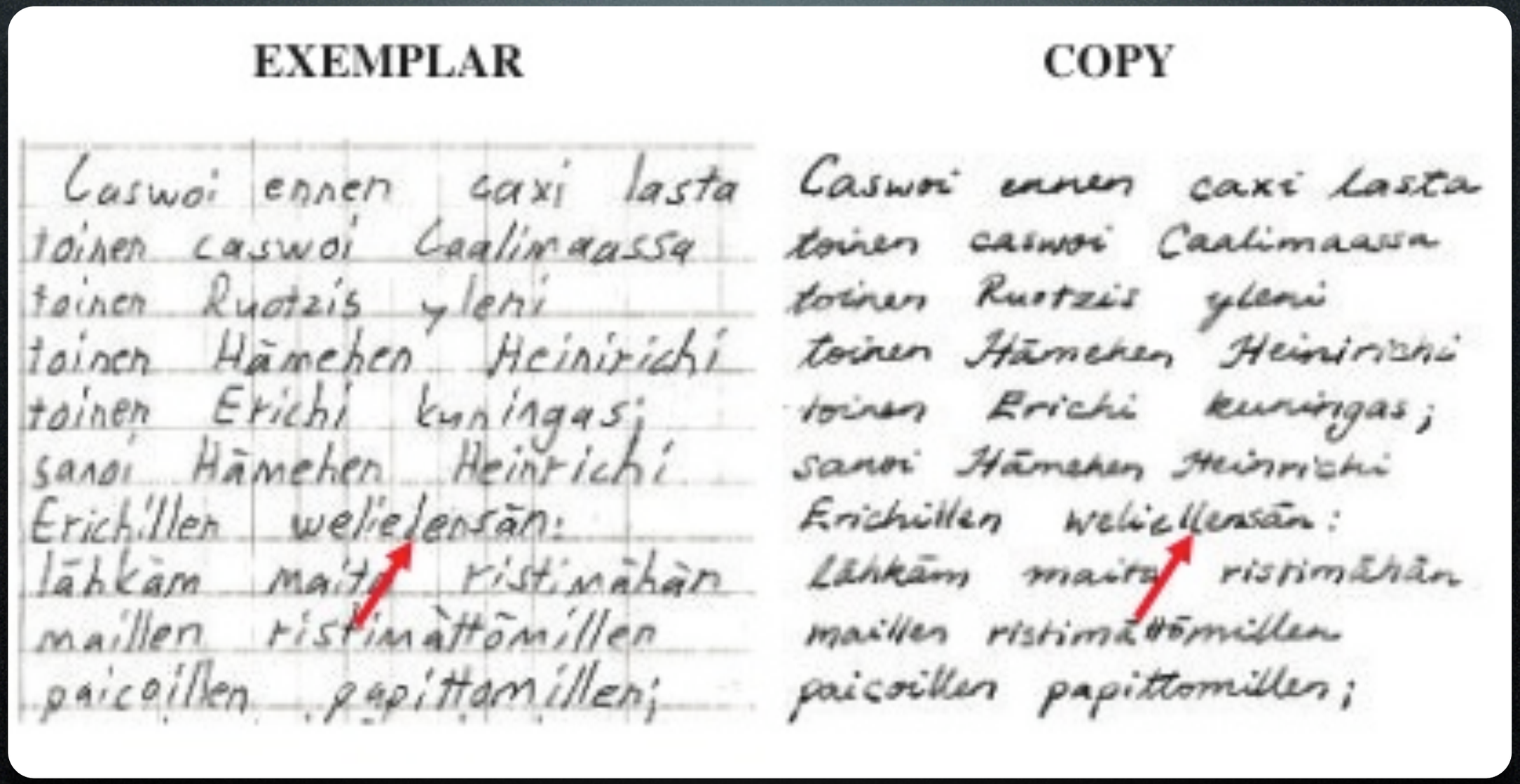


Fig. 1. A phylogeny, showing its relationship to the underlying genealogy.

evolution of texts?



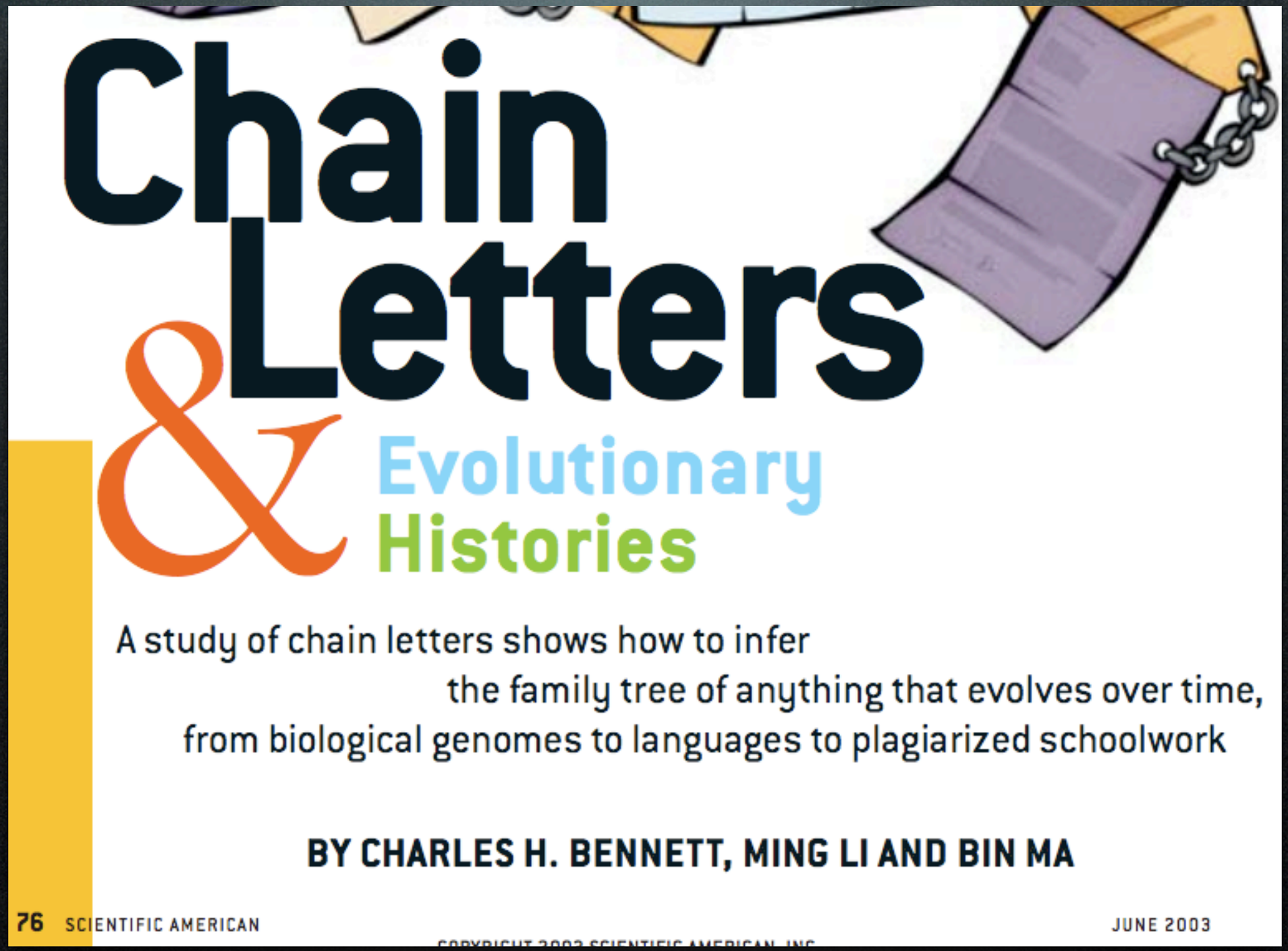
typical change

Similarities

- copying - reproduction
- “errors” - mutations

Differences

- non-randomness of changes
- old manuscripts remain, species evolve
- multifurcation - bifurcation

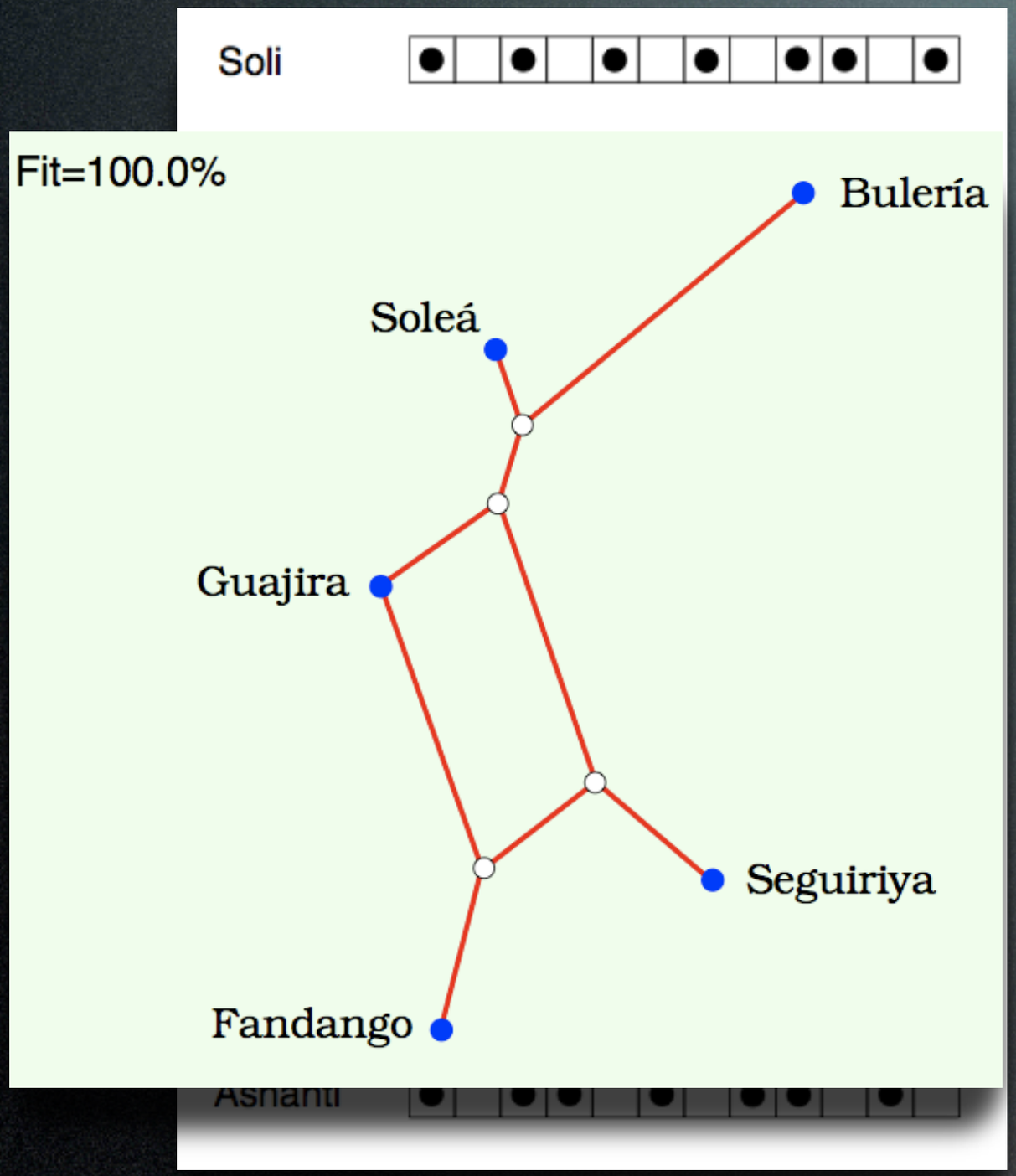


Chain Letters & Evolutionary Histories

A study of chain letters shows how to infer the family tree of anything that evolves over time, from biological genomes to languages to plagiarized schoolwork

BY CHARLES H. BENNETT, MING LI AND BIN MA

chain letters

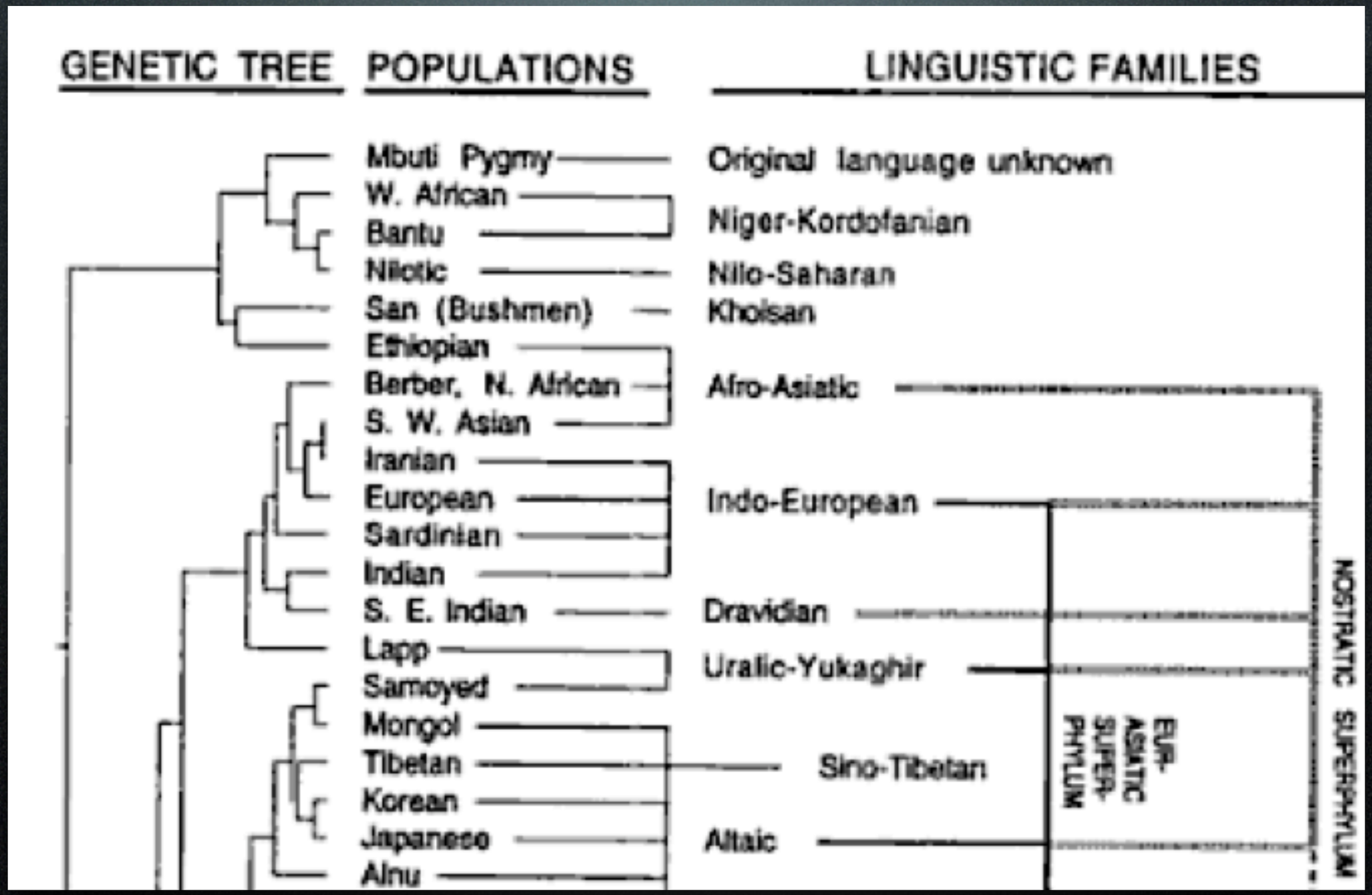


Toussaint, “Classification and phylogenetic analysis of **African ternary rhythm** timelines”, 2003.

Diaz-Banez et al., “El Compas **Flamenco**: a phylogenetic analysis”, 2003.

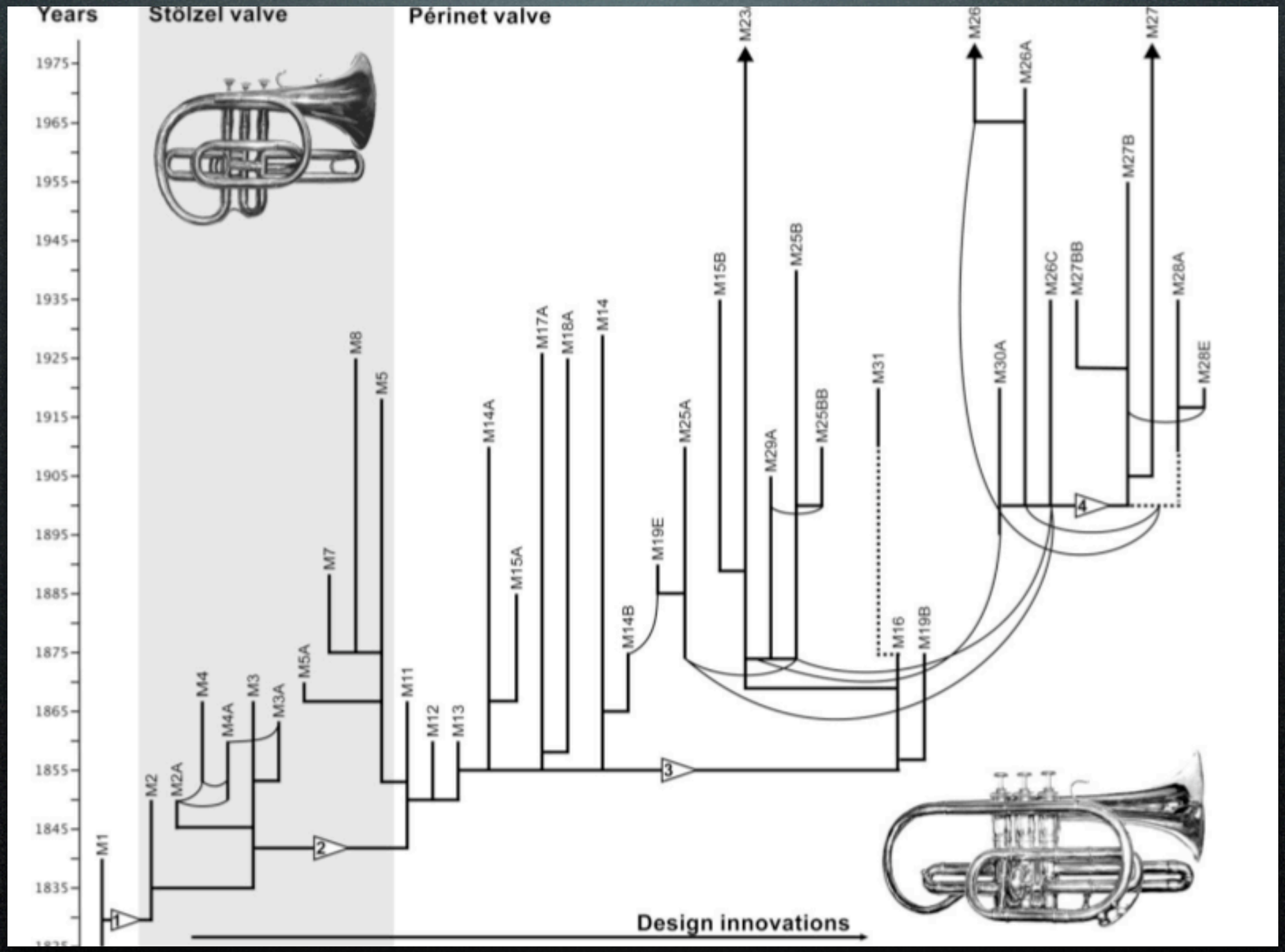
etc.

music



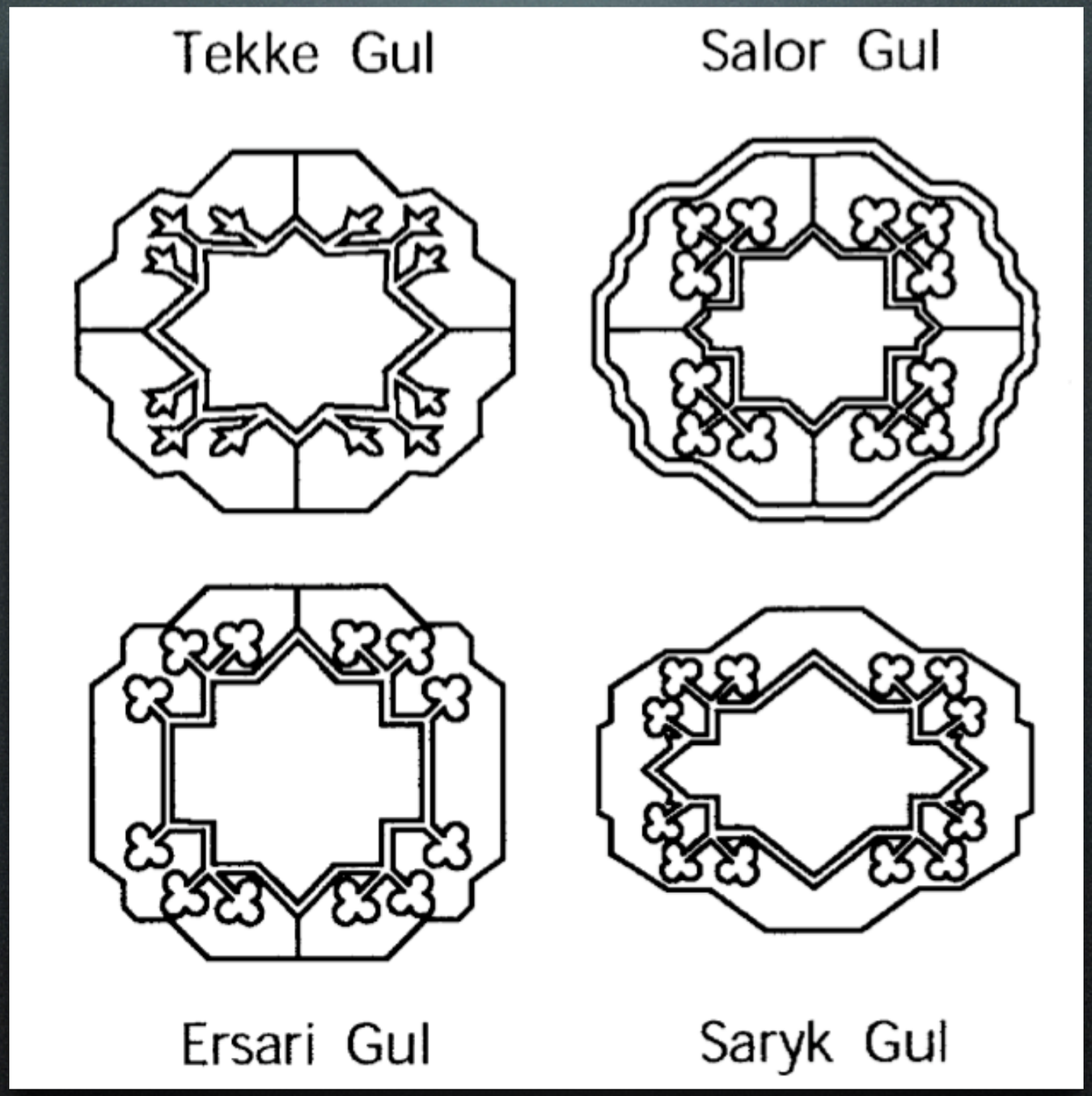
Cavalli-Sforza et al. , The History and Geography of Human Genes, 1993

languages & peoples



Temkin and Eldredge, “Phylogenetics and Material Culttural Evolution”, 2007

design

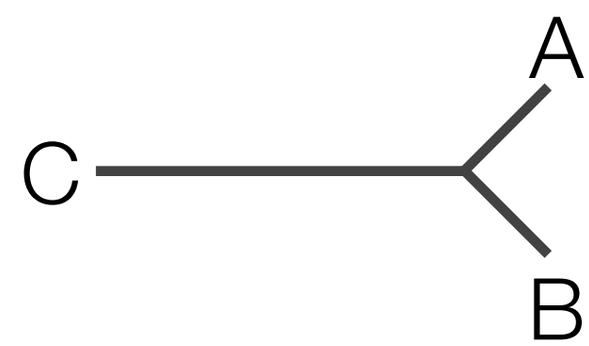


Tehrani and Collard, “Investigating cultural evolution through biological phylogenetic analyses of Turkmen textiles”, 2002.

skills

Basic principles

When do we introduce a hyparchetype?



“Steiner tree”

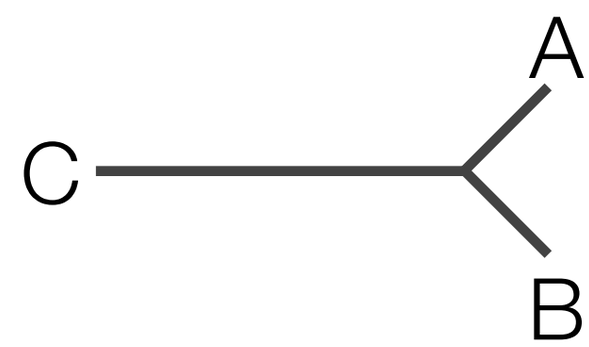
Basic principles

When do we introduce a hyparchetype?

A = kissa

B = kissa

C = koira



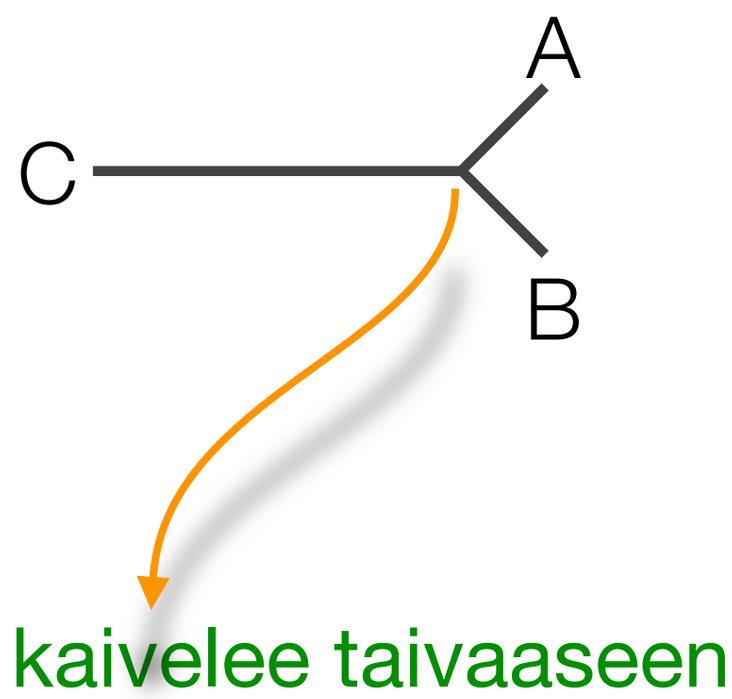
Basic principles

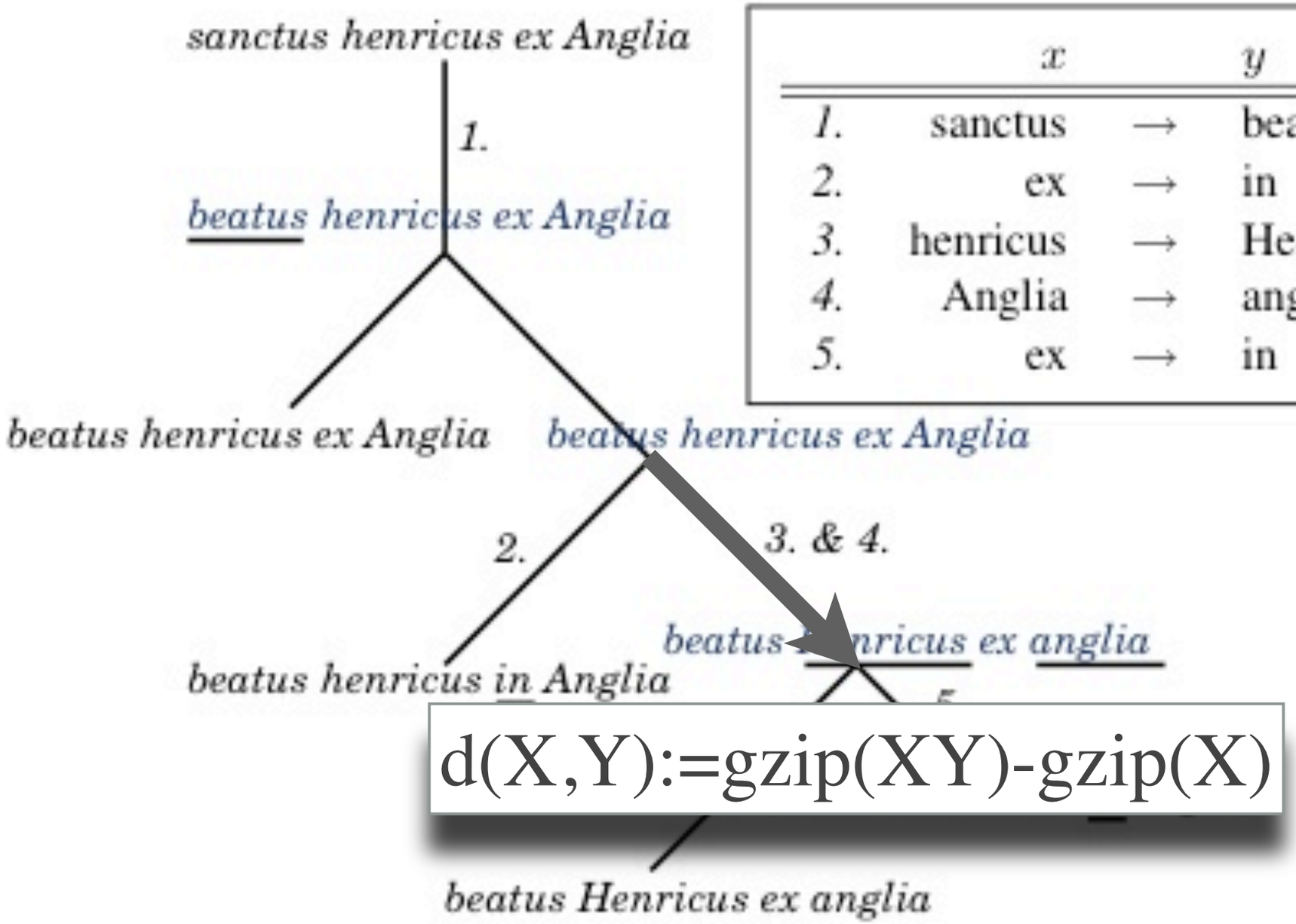
When do we introduce a hyparchetype?

A = kissa kävelee taivaaseen

B = kissa kaivelee teeveeseen

C = koira kaivelee taivaaseen





	x		y	C
1.	sanctus	→	beatus	5
2.	ex	→	in	3
3.	henricus	→	Henricus	3
4.	Anglia	→	anglia	3
5.	ex	→	in	3

maximum parsimony / MDL

text phylogeny method

$\mathcal{X}_1 =$

$\mathcal{X}_2 =$

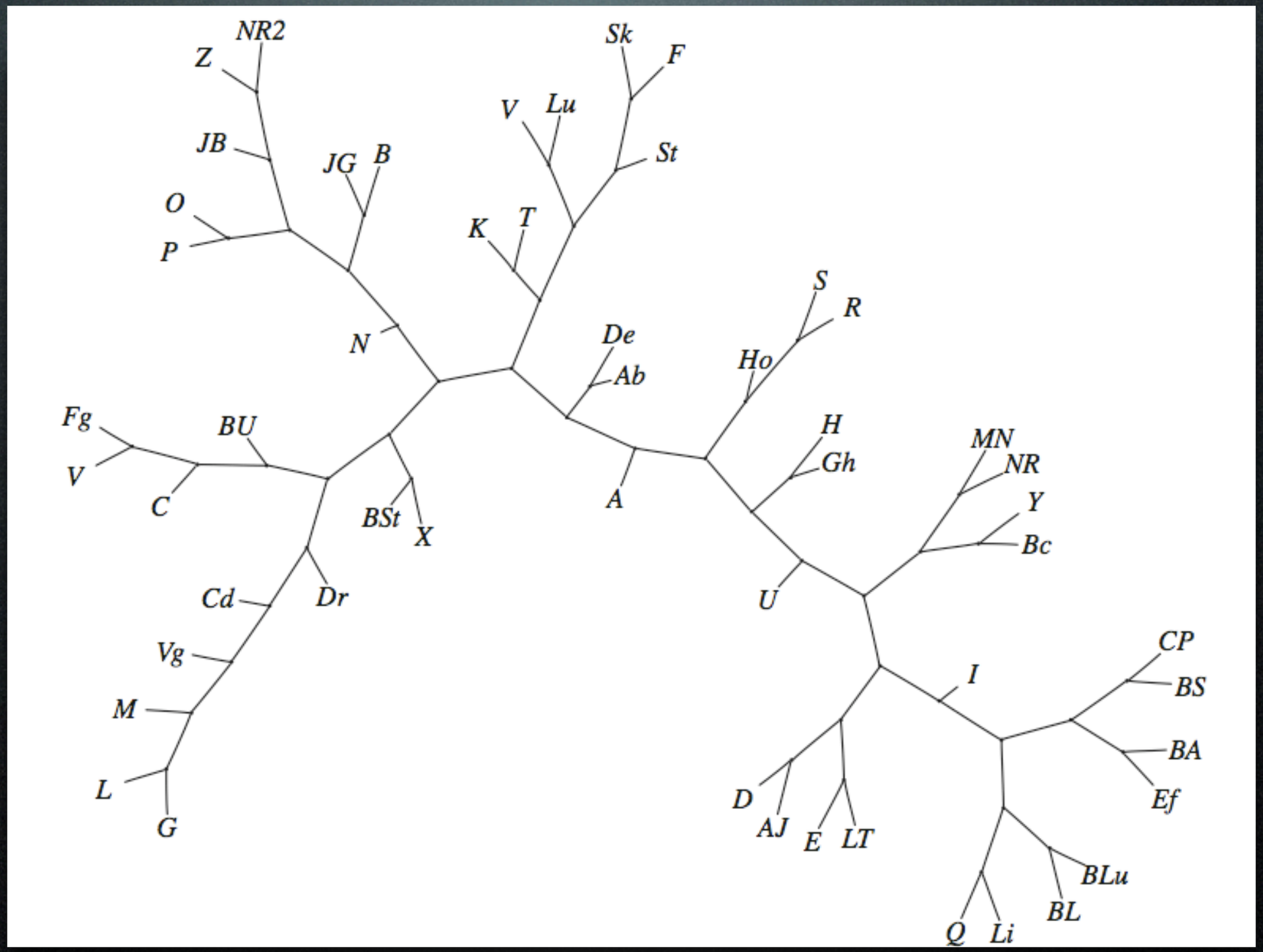
A	B	C	D	E	F
rege	rege	rege	rege	rege	rege
sancto	sancto	sancto	sancto	sancto	sancto
erico	erico	erico	erico	erico	erico
in	in	in	in	in	in
suecia	suecia	swecia	swetia	suecia	suetia
uenerabilis		venerabilis	venerabilis		
pontifex		pontifex	pontifex		
beatus		beatus	beatus	sanctus	sanctus
henricus		henricus	heinricus	henricus	henricus
de		de	de	in	ex
anglia		anglia	anglia	anglia	anglia
oriundus		oriundus	oriundus	oriundus	oriundus
conspicuus		conspicuus	conspicuus	conspicuus	conspicuus
uite		vite	vite	vite	vite
sanctitate		sanctitate	sanctitate	sanctitate	sanctitate
et		&	&	et	&
morum		morum	morum	morum	morum
honestate		honestate	honestate	honestate	honestate
		preclarus	preclarus	preclarus	preclarus
vpsalensem		upsalensem	vpsalensem	vpalensem	vpsalensem
regebat		regebat	regebat	regebat	regebat
ecclesiam		ecclesiam	ecclesiam	ecclesiam	ecclesiam

maximum parsimony / MDL

text phylogeny method

Real data
12th-15th century
53+ manuscripts
objective: locate origin &
reconstruct original

Case 1: Saint Henry



Case 1: Saint Henry

Domain experts happy
Origin: most likely Finland

Reasonable reconstruction

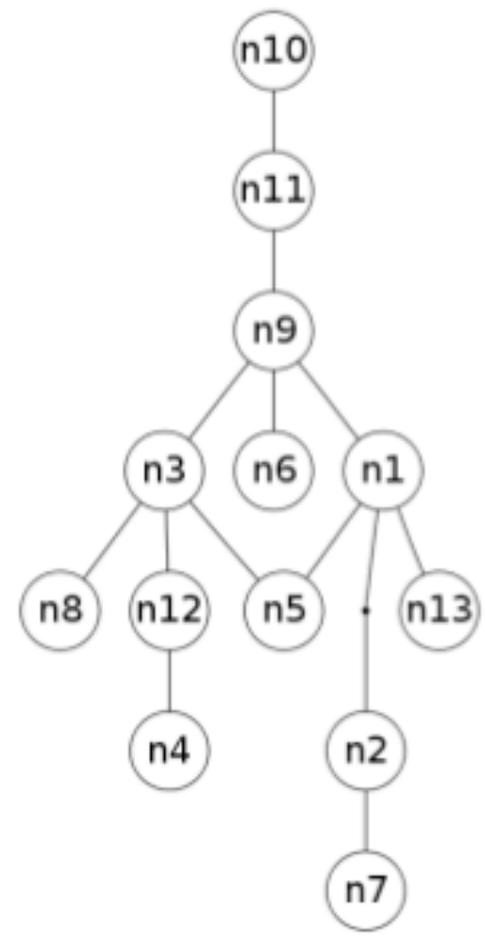
Caveat: root?
(score symmetrical; cf. causal discovery)

Case 1: Saint Henry

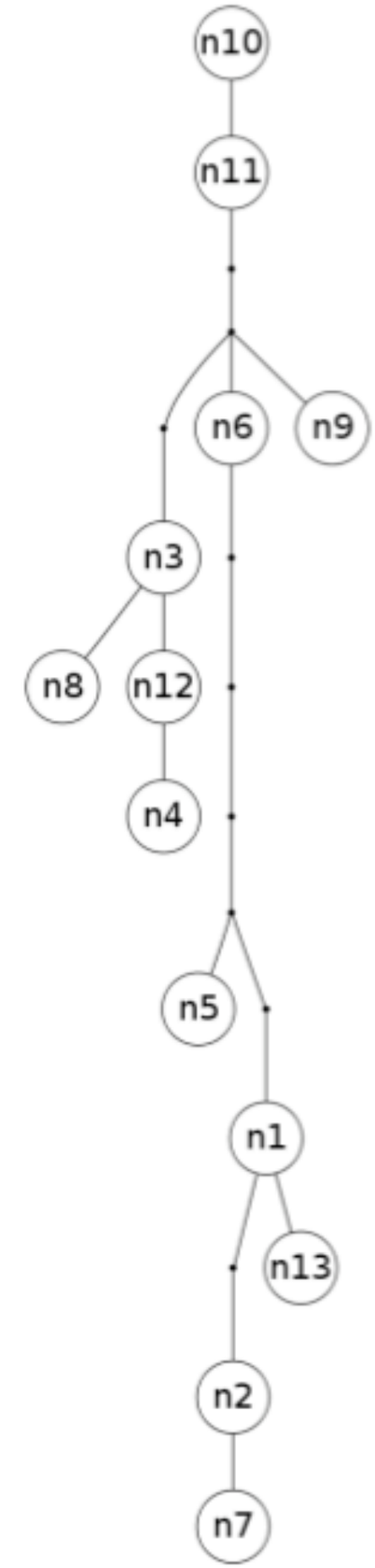
Case 1: Notre Besoin

Scores:

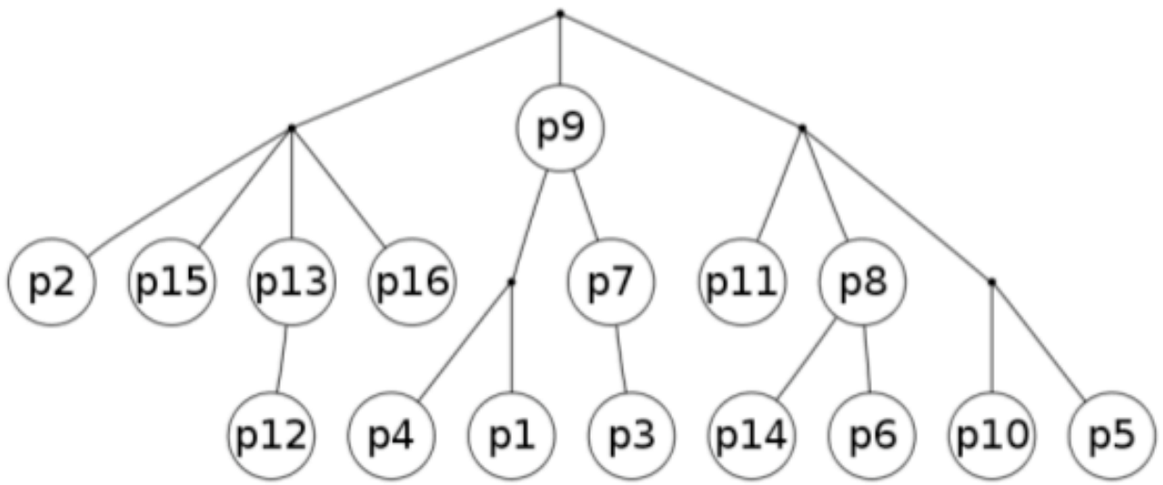
PAUP	75%
RHM	77%
SEM	73%

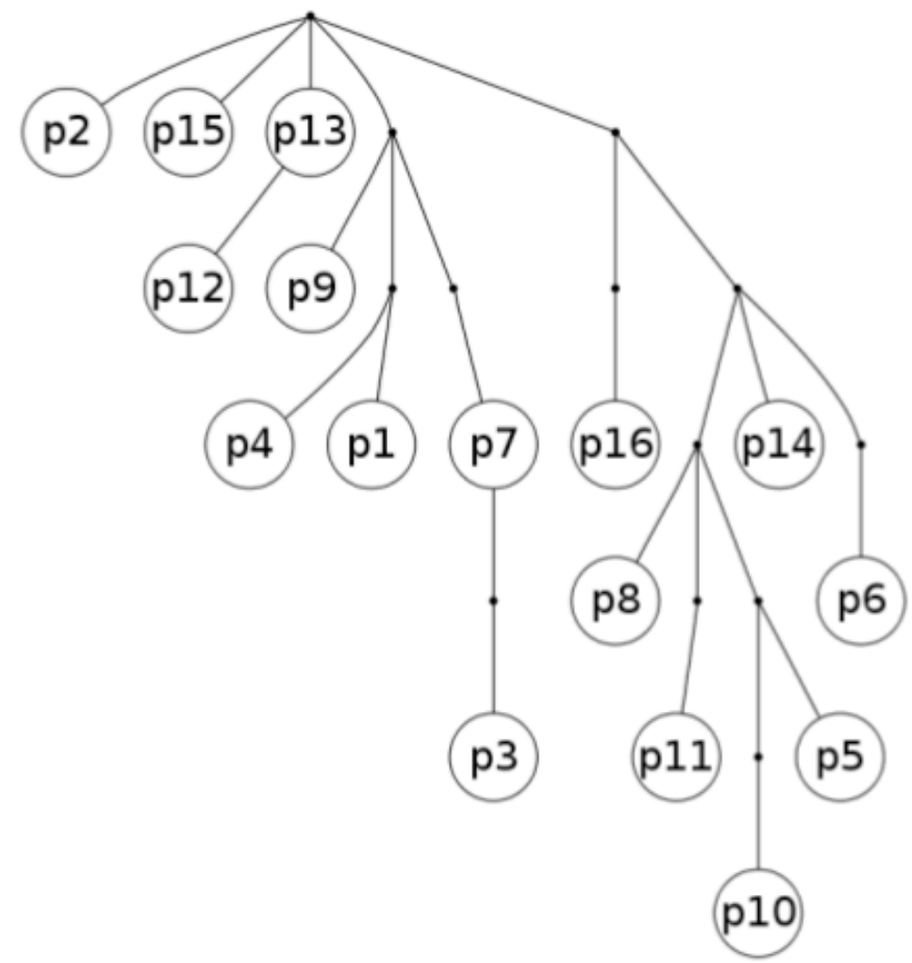


Correct stemma



SEM



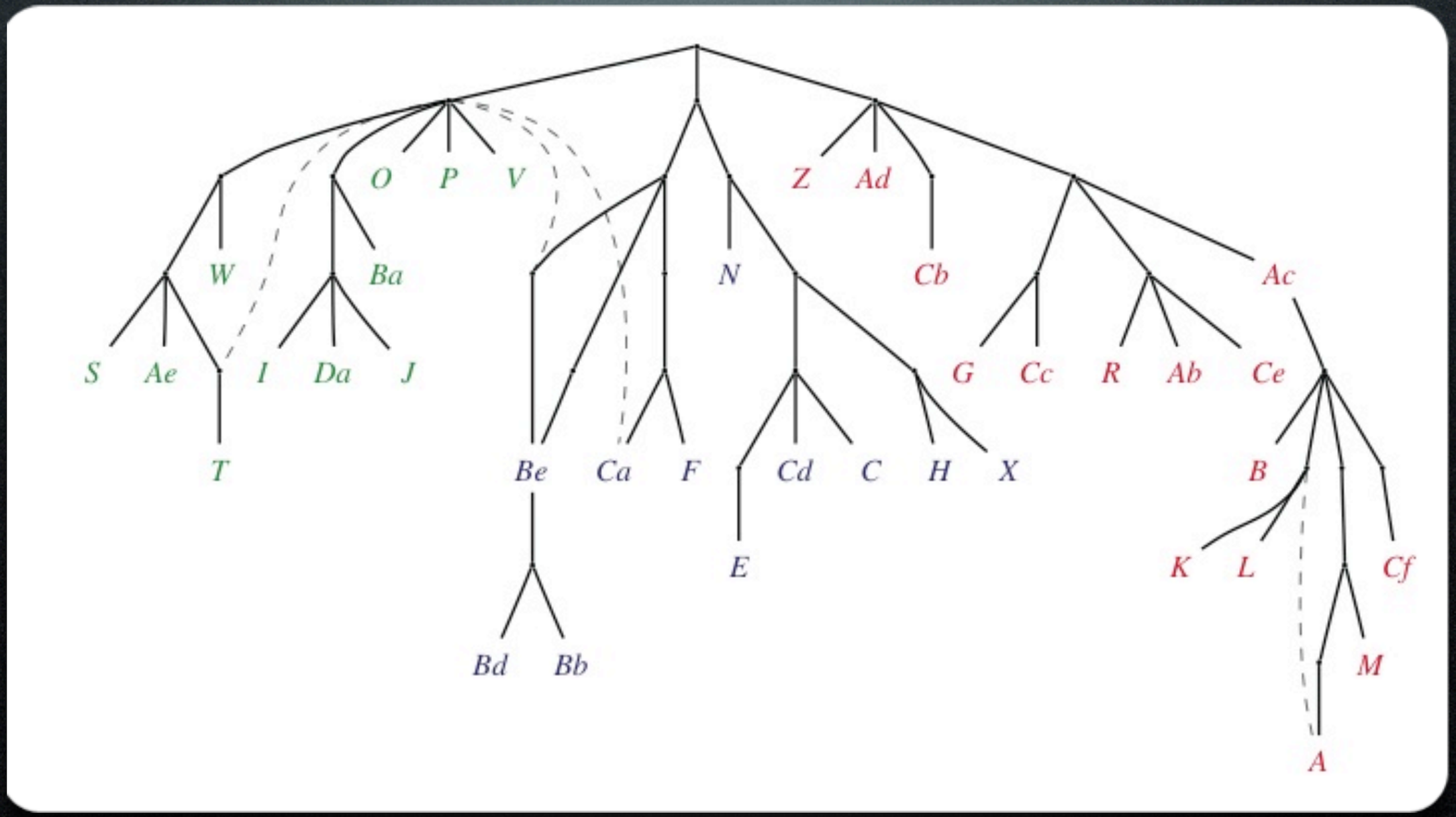


Correct stemma

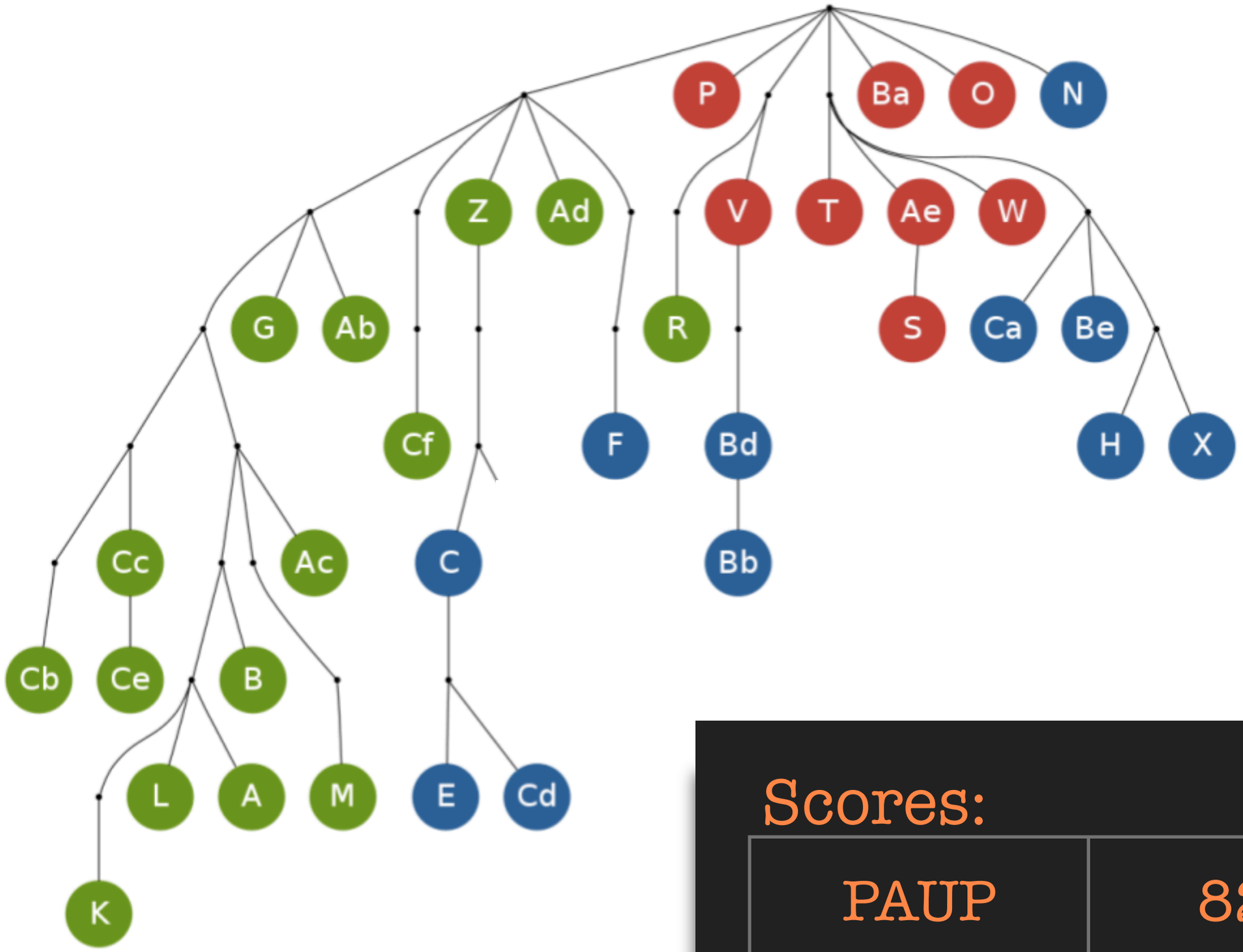
Case 2: Parzival

Scores:

	SEM
PAUP	78%
RHM	80%
SEM	79%



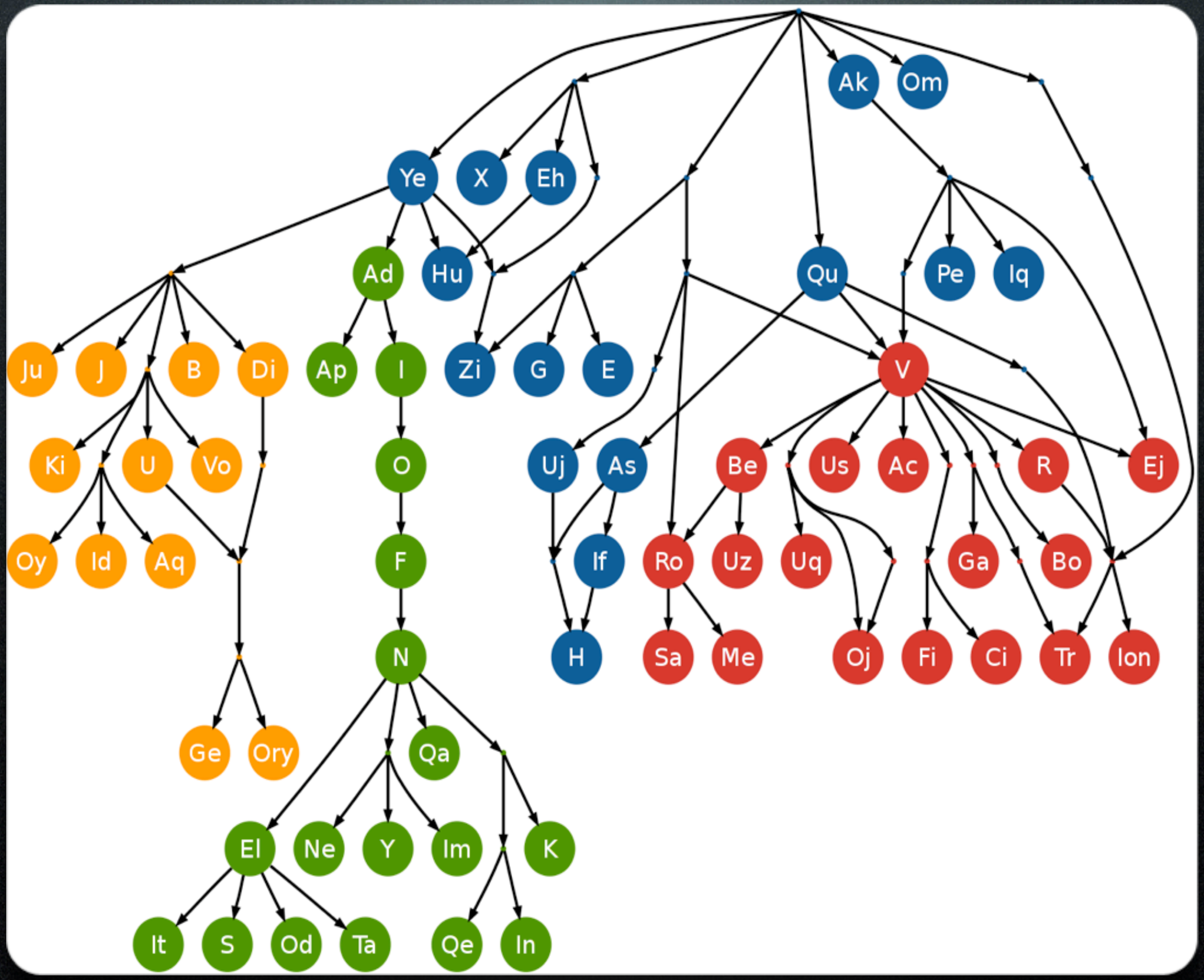
Case 3: Heinrichi



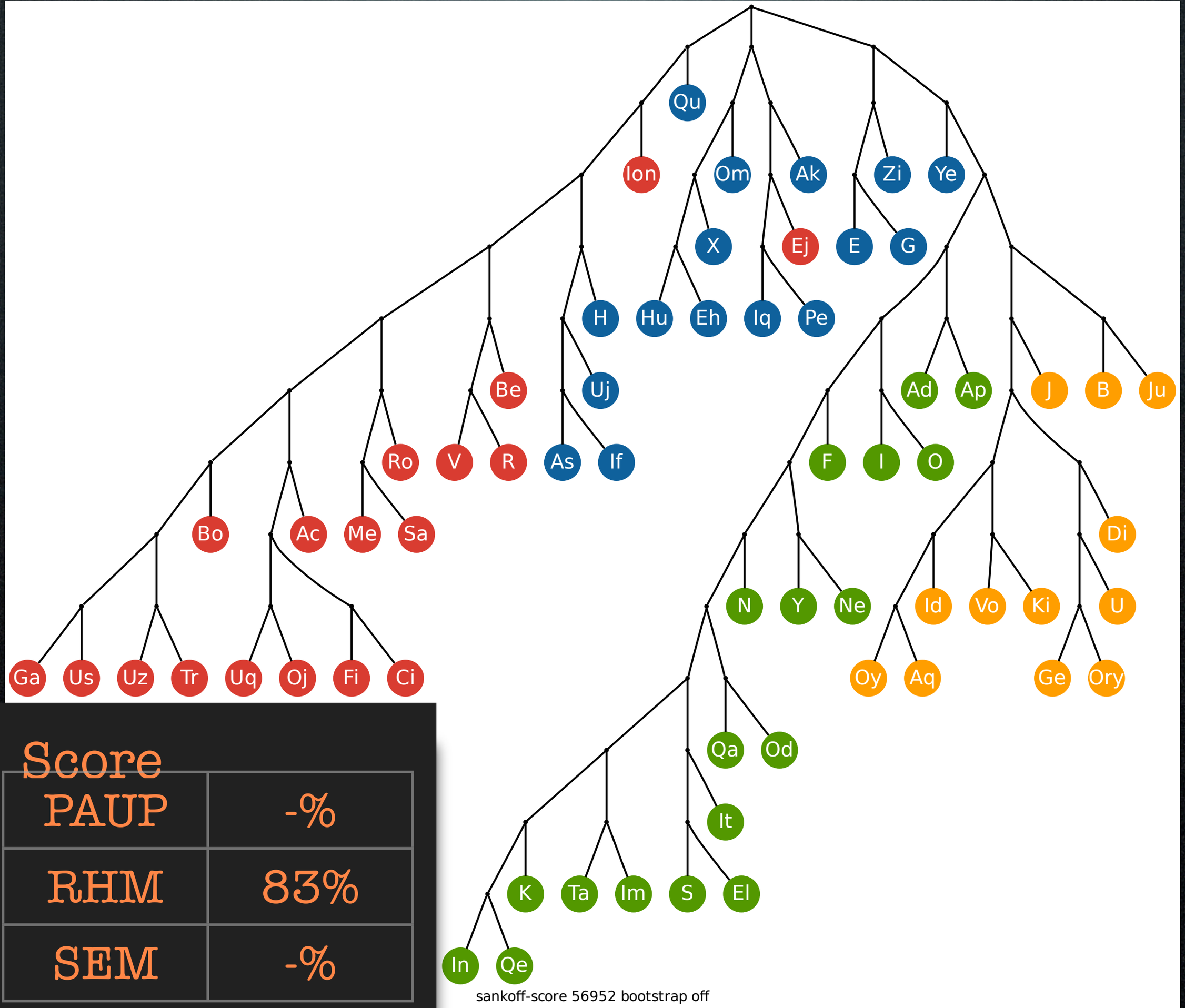
Case 3: Heinrichi

Scores:

PAUP	82%
RHM	82%
SEM	76%

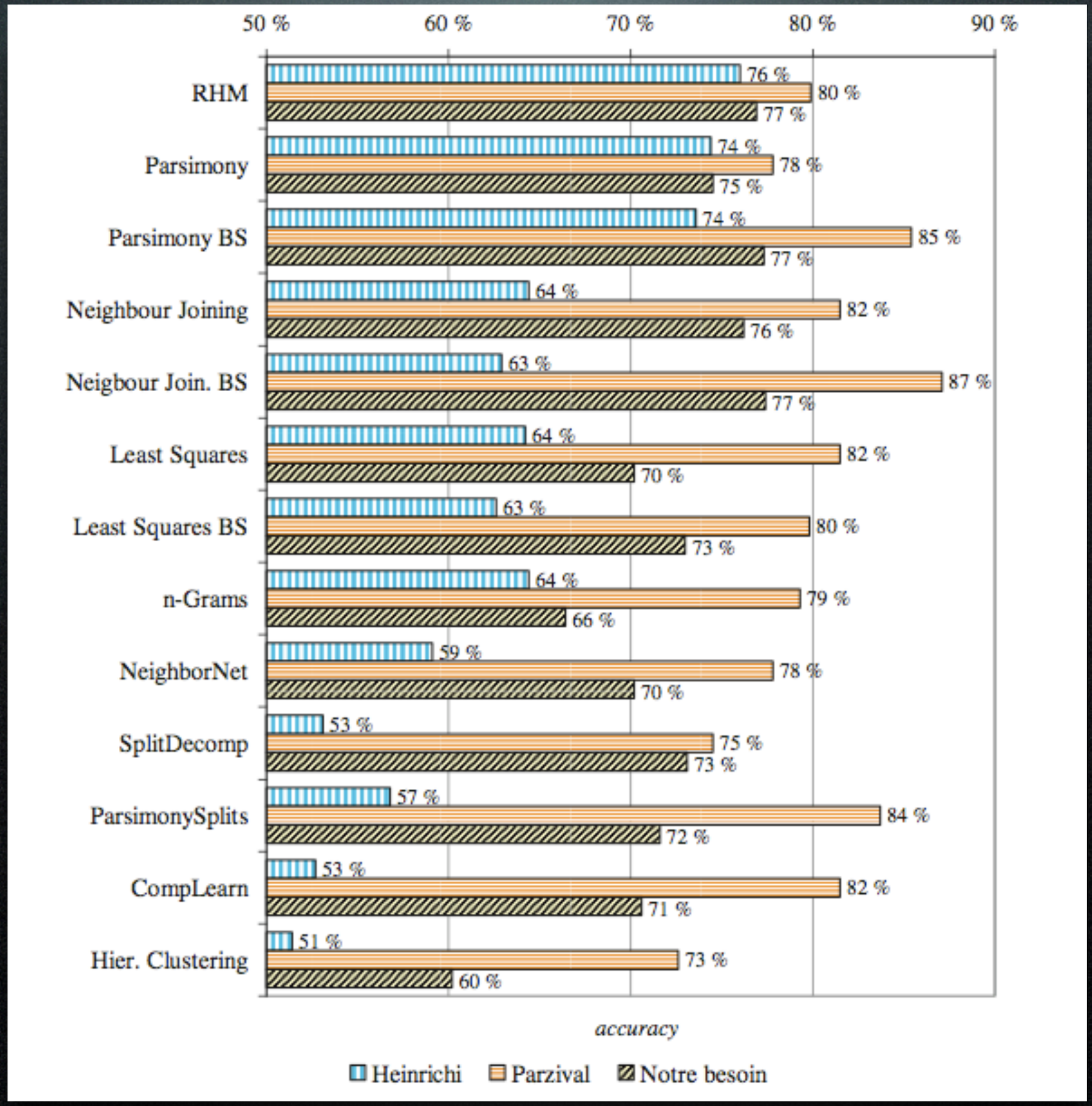


Case 2: Julius Caesar



Score

PAUP	-%
RHM	83%
SEM	-%


50 %
60 %
70 %
80 %
90 %
RHM
76 %
80 %
77 %
Parsimony
74 %
78 %
75 %
Parsimony BS
74 %
85 %
77 %
Neighbour Joining
64 %
82 %
76 %
Neigbour Join. BS
63 %
87 %
77 %
Least Squares
64 %
82 %
70 %
Least Squares BS
63 %
80 %
73 %
n-Grams
64 %
79 %
66 %
NeighborNet
59 %
78 %
70 %
SplitDecomp
53 %
75 %
73 %
ParsimonySplits
57 %
84 %
72 %
CompLearn
53 %
82 %
71 %
Hier. Clustering
51 %
73 %
60 %
accuracy
Heinrichi
Parzival
Notre besoin

open problems

- improving models (of copying)
  -> better scores
- model checking: is it really a tree?

- orientation of edges

$$p_{x,y} \neq p_{y,x}$$

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A	0	1	2	3	4	4	4	3	4	5	6	7	5	5
B	1	0	1	2	3	3	3	2	3	4	5	6	4	4
C	2	1	0	1	2	2	2	1	2	3	4	5	3	3
D	3	2	1	0	1	1	1	2	3	4	5	6	4	4
E	4	3	2	1	0	2	2	3	4	5	6	7	5	5
F	4	3	2	1	2	0	2	3	4	5	6	7	5	5
G	4	3	2	1	2	2	0	3	4	5	6	7	5	5
H	3	2	1	2	3	3	3	0	1	2	3	4	2	2
I	4	3	2	3	4	4	4	1	0	1	2	3	1	1
J	5	4	3	4	5	5	5	2	1	0	1	2	2	2
K	6	5	4	5	6	6	6	3	2	1	0	1	3	3
L	7	6	5	6	7	7	7	4	3	2	1	0	4	4
M	5	4	3	4	5	5	5	2	1	2	3	4	0	2
N	5	4	3	4	5	5	5	2	1	2	3	4	2	0

Acknowledgments

- Tuomas Heikkilä, Dept. of History, UH
- Petri Myllymäki, HIIT
- Yuan Zou, HIIT