言語系統論への
計算的アプローチの可能性

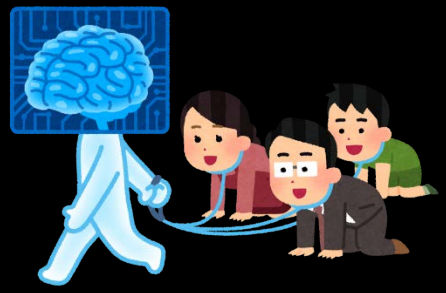
京都大学
村脇 有吾

日本歴史言語学会2017年大会 公開シンポジウム 言語系統論の過去(これまで)と未来(これから)@大阪学院大学 (2017年12月9日)

自己紹介: 村脇 有吾

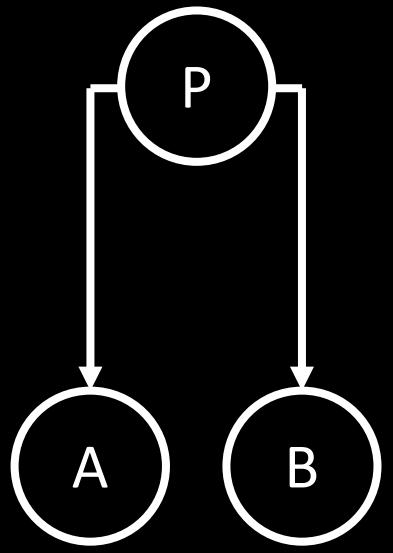
- 京都大学 大学院情報学研究科 知能情報学専攻 助教
  工学部電気電子工学科 兼担

- 専門: 計算言語学と自然言語処理
  - 人工知能バブル
  - 言語処理学会: 年次大会参加者 >900
  - 2000年代から統計・機械学習を本格導入
  - 表の仕事: ゼロ照応解析、談話構造解析、テキストからの知識獲得ほか
  - 今日前半で紹介する研究は生物学由来で、実は専門外



かつての言語年代学 (Glottochronology)

$$t = \frac{\log c}{2 \log r}$$



- t: 祖語Pの年代 (単位: 千年)
- c: 言語A, Bの基礎語彙共有率
- r: 基礎語彙の残存率 (200項目で0.81)

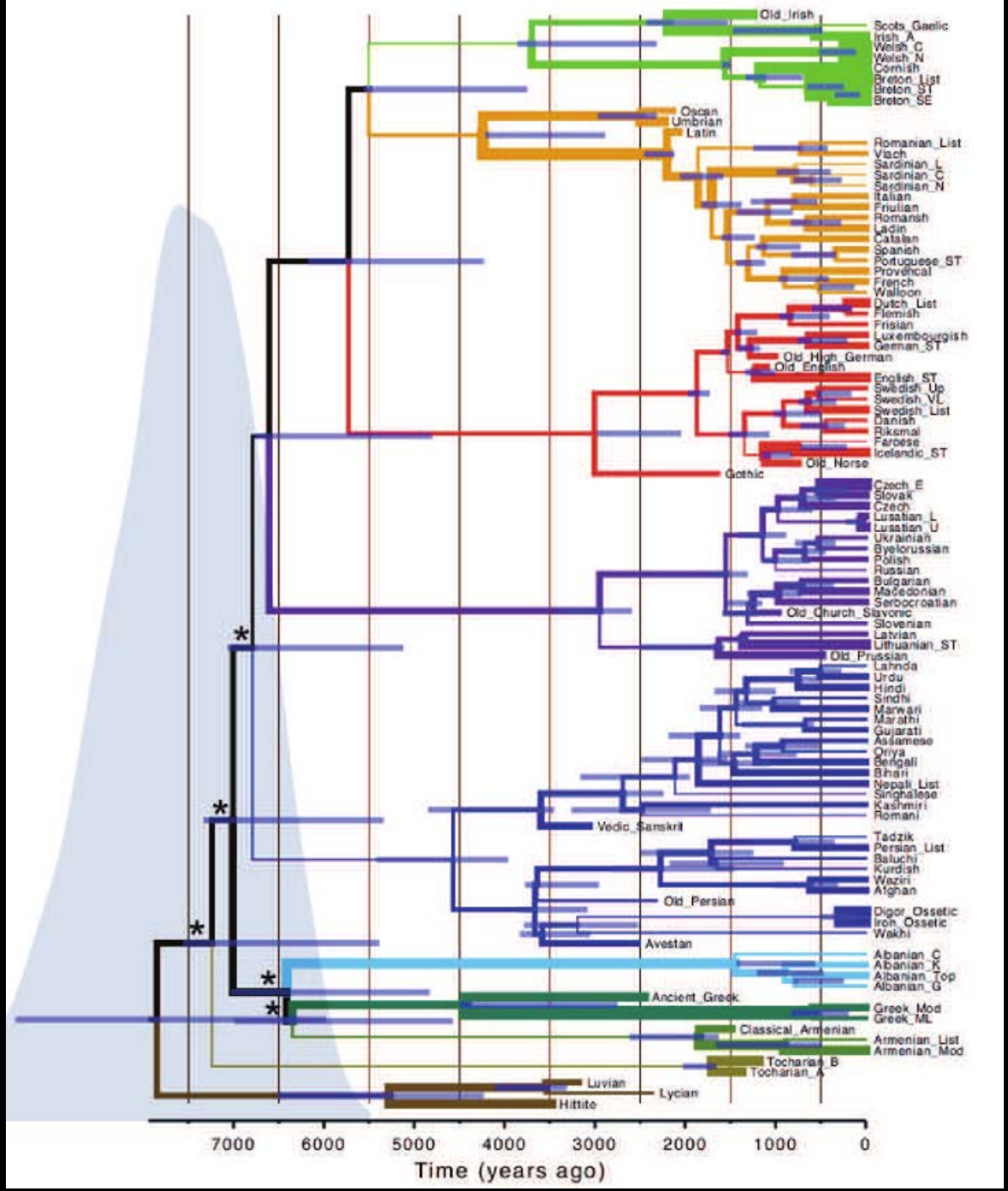
年代	1K	2K	3K	4K	5K	6K	7K	8K	9K
共有率 (r=.81)	.66	.43	.28	.19	.12	.08	.05	.03	.02

[Swadesh, 1952]

今日の目標:
→の作り方の雰囲気をつかむ

- ベイズ統計による確率モデル
- 最大節約、近隣結合は取り上げない

[Bouckaert+, Science 2012]

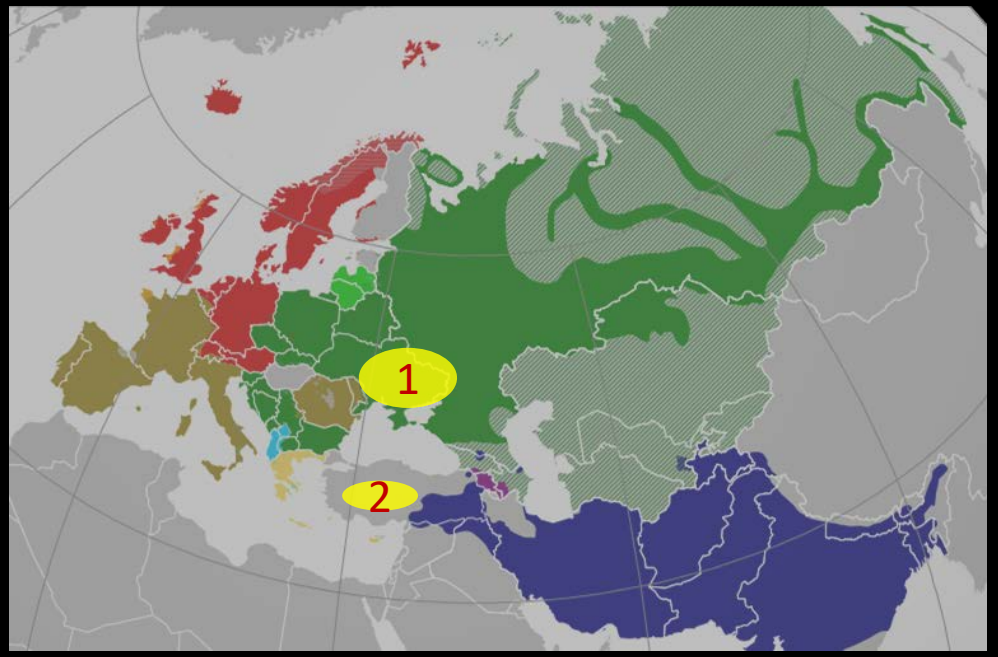


From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

印欧祖語の年代と故地

1. クルガン仮説
   - 5,000-6,000年前
   - 黒海周辺のステップ
   - 遊牧民の軍事的征服

2. アナトリア仮説
   - 8,000-9,500年前
   - アナトリア
   - 農耕とともに拡大
     - Renfrew (考古学者) の農耕・言語同時伝播モデル



Source:
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

印欧祖語の年代と故地

1. クルガン仮説
   - 5,000-6,000年前
   - 黒海周辺のステップ
   - 遊牧民の軍事的征服

2. アナトリア仮説
   - 8,000-9,500年前
   - アナトリア
   - 農耕とともに拡大
     - Renfrew (考古学者) の農耕・言語同時伝播モデル

Source:
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

印欧祖語の年代と故地

1. クルガン仮説
   - 5,000-6,000年前
   - 黒海周辺のステップ
   - 遊牧民の軍事的征服

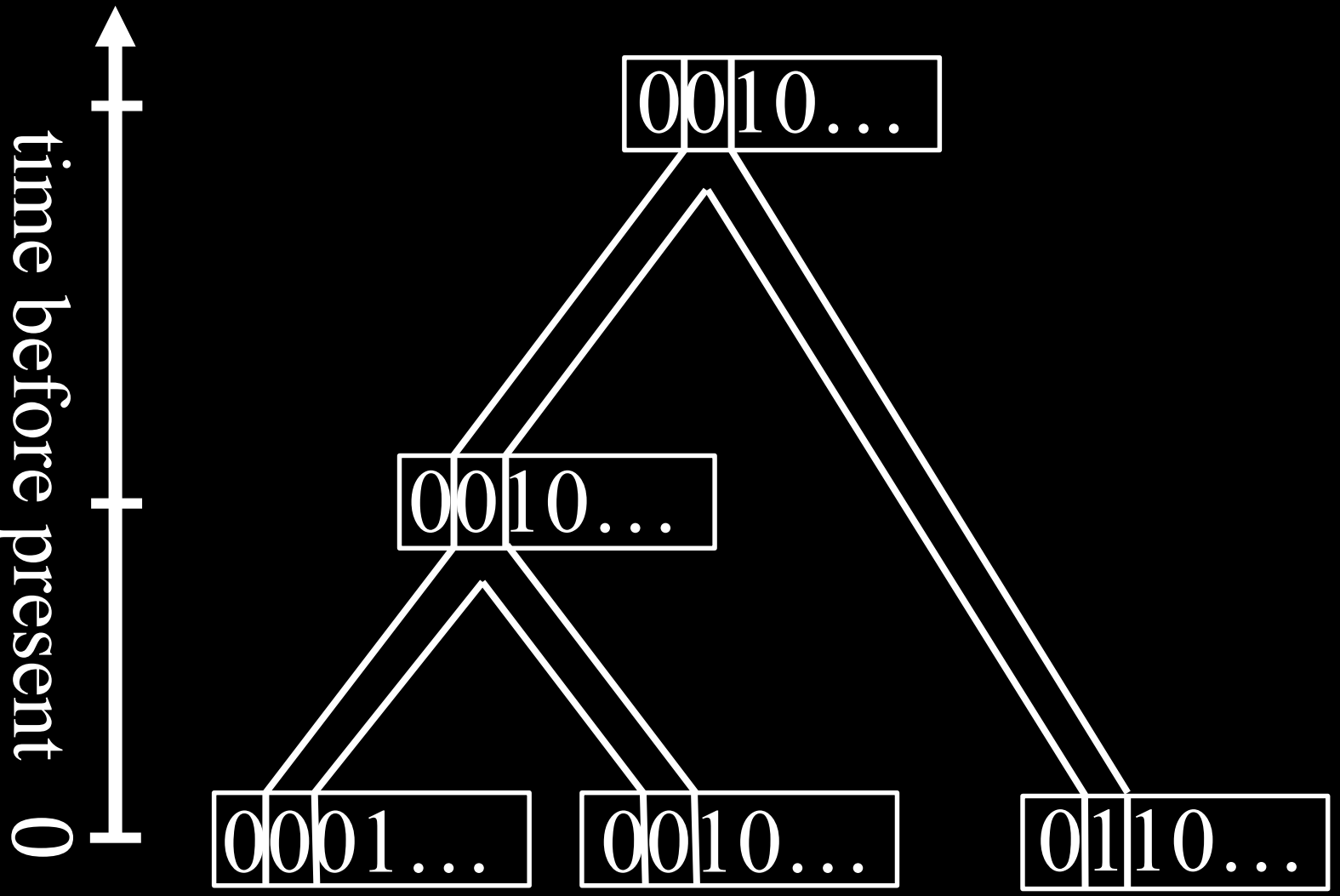
2. アナトリア仮説
   - 8,000-9,500年前
   - アナトリア
   - 農耕とともに拡大
     - Renfrew (考古学者) の農耕・言語同時伝播モデル

Source:
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

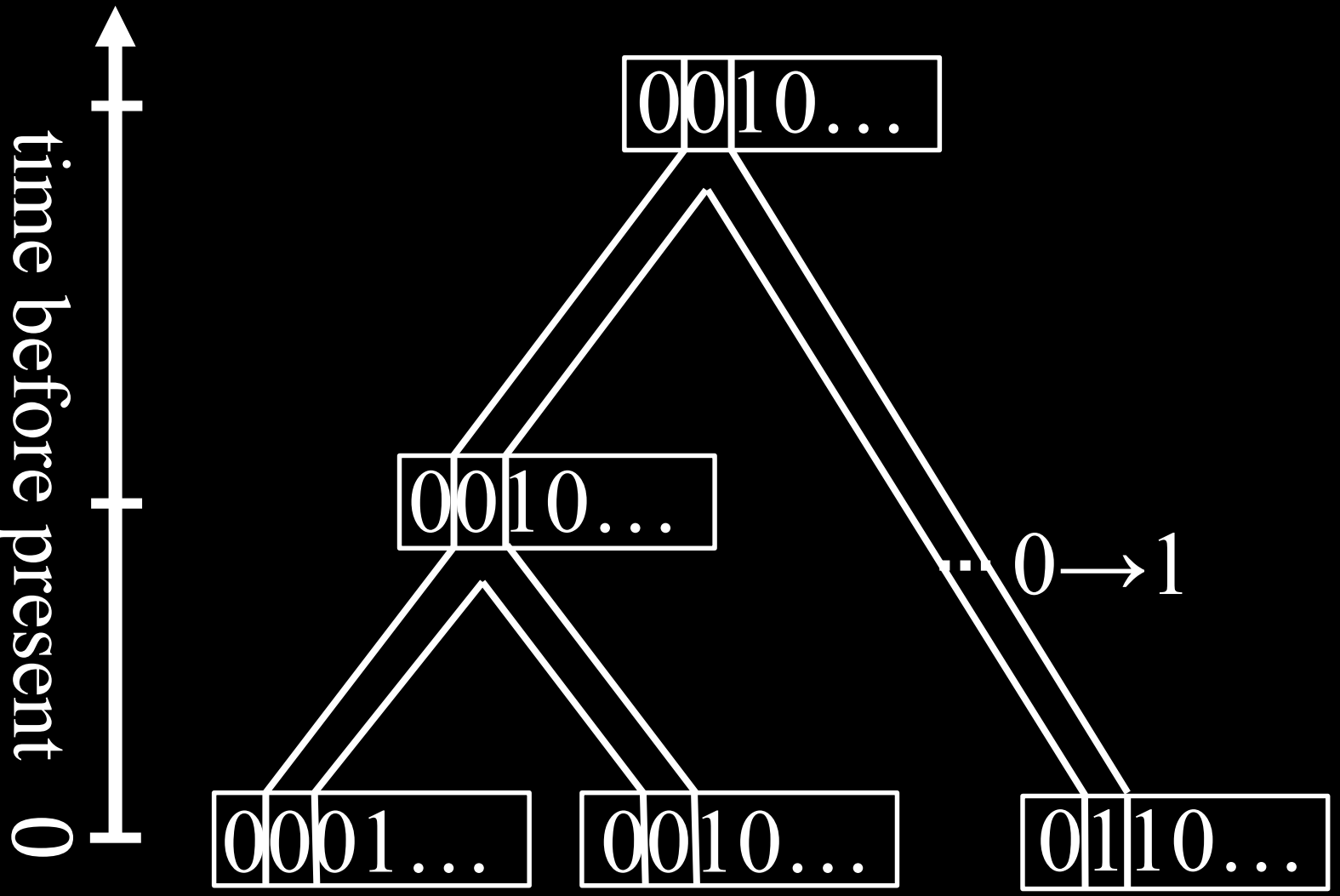
年代つき系統樹

0001…　0010…　0110…

年代つき系統樹



年代つき系統樹



言語の状態: 符号化された語彙

	WATER	BIG	
英語	water	big	
ドイツ語	Wasser	gross	
ロシア語	вода	большой	великий
フランス語	eau	grand	
イタリア語	acqua	grande	

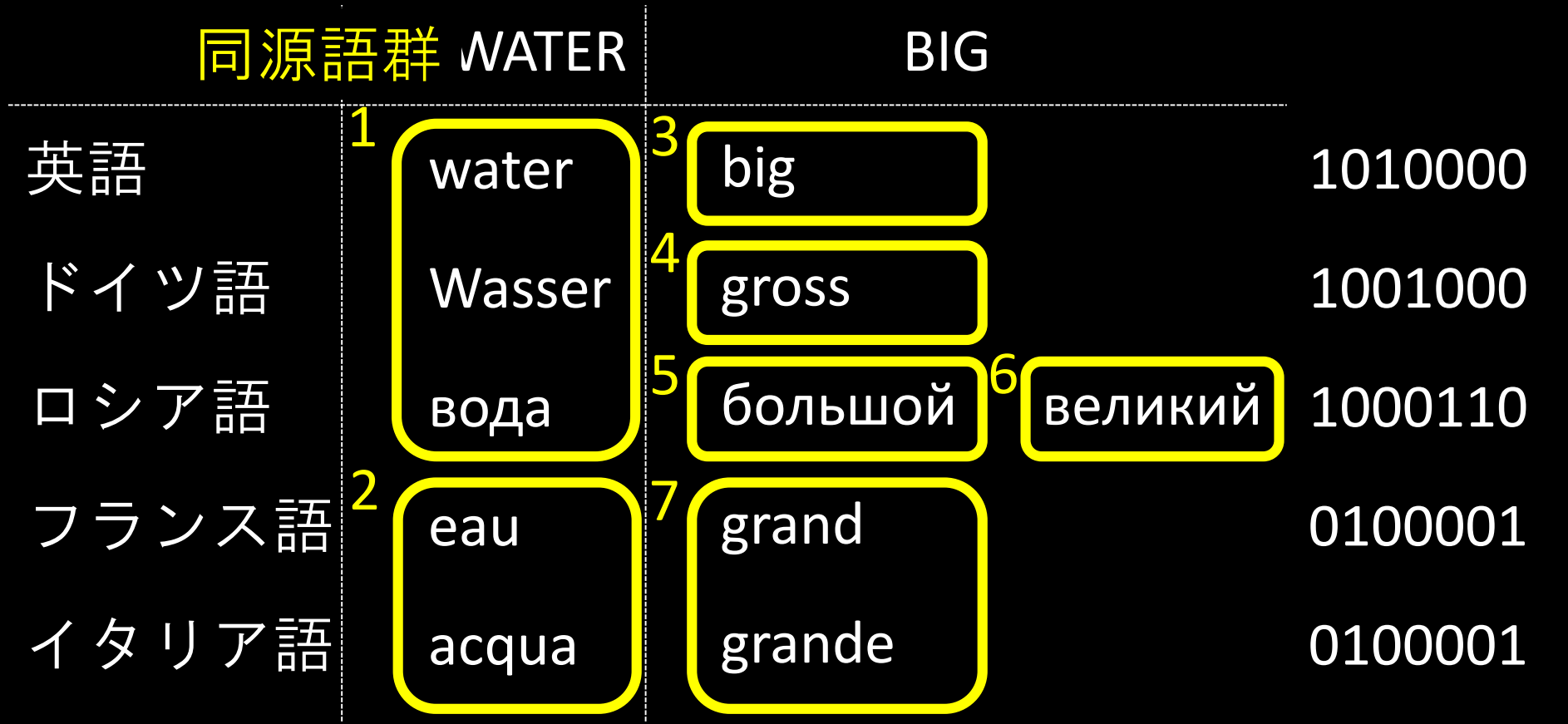
言語の状態: 符号化された語彙

同源語群	WATER	BIG	
英語	water	big	
ドイツ語	Wasser	gross	
ロシア語	вода	большой	великий
フランス語	eau	grand	
イタリア語	acqua	grande	

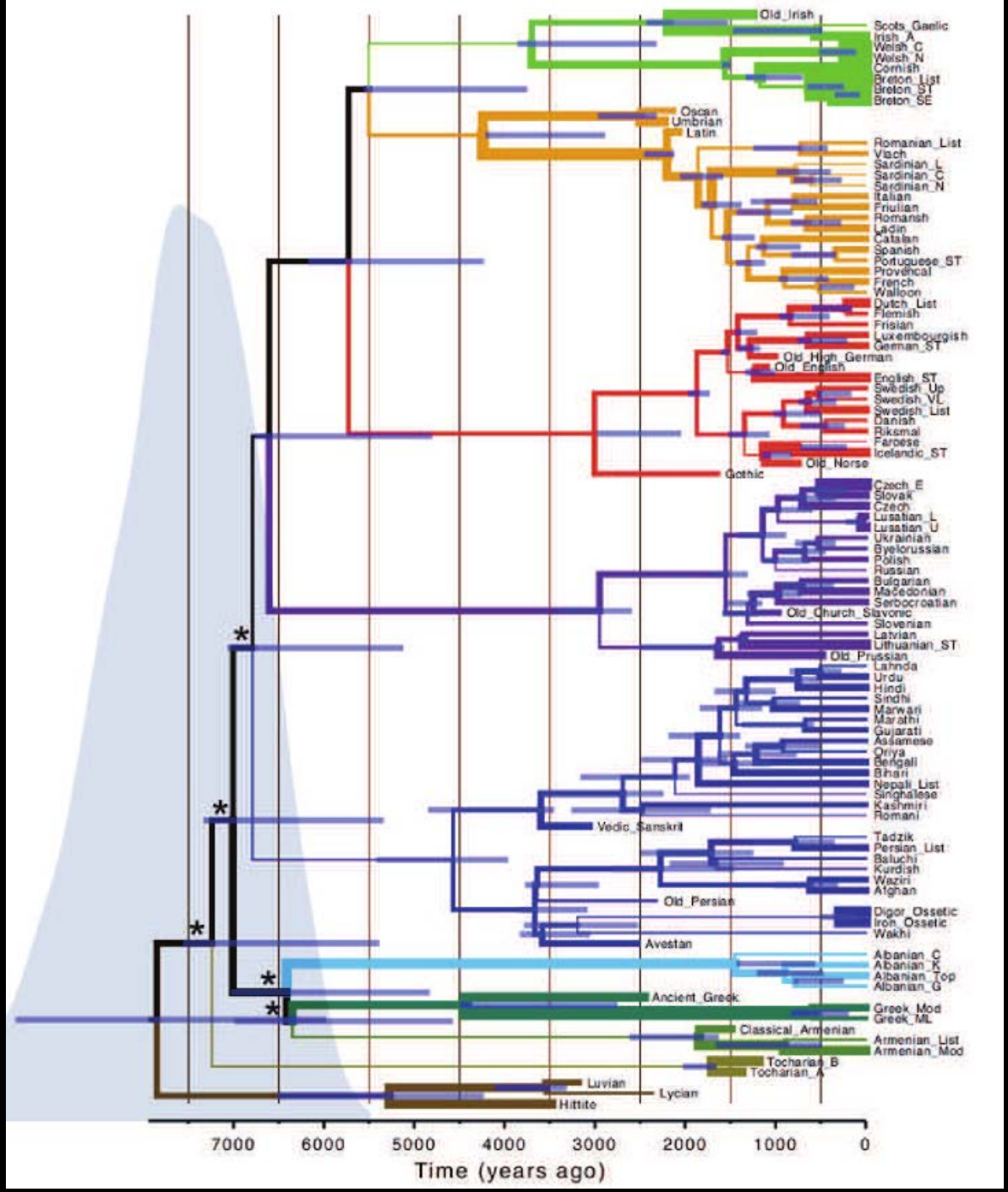
言語の状態: 符号化された語彙

	同源語群 WATER		BIG		
英語	1	water	3	big	{1, 3}
ドイツ語		Wasser	4	gross	{1, 4}
ロシア語		вода	5	большой	6 великий {1, 5, 6}
フランス語	2	eau	7	grand	{2, 7}
イタリア語		acqua		grande	{2, 7}

言語の状態: 符号化された語彙



問題設定



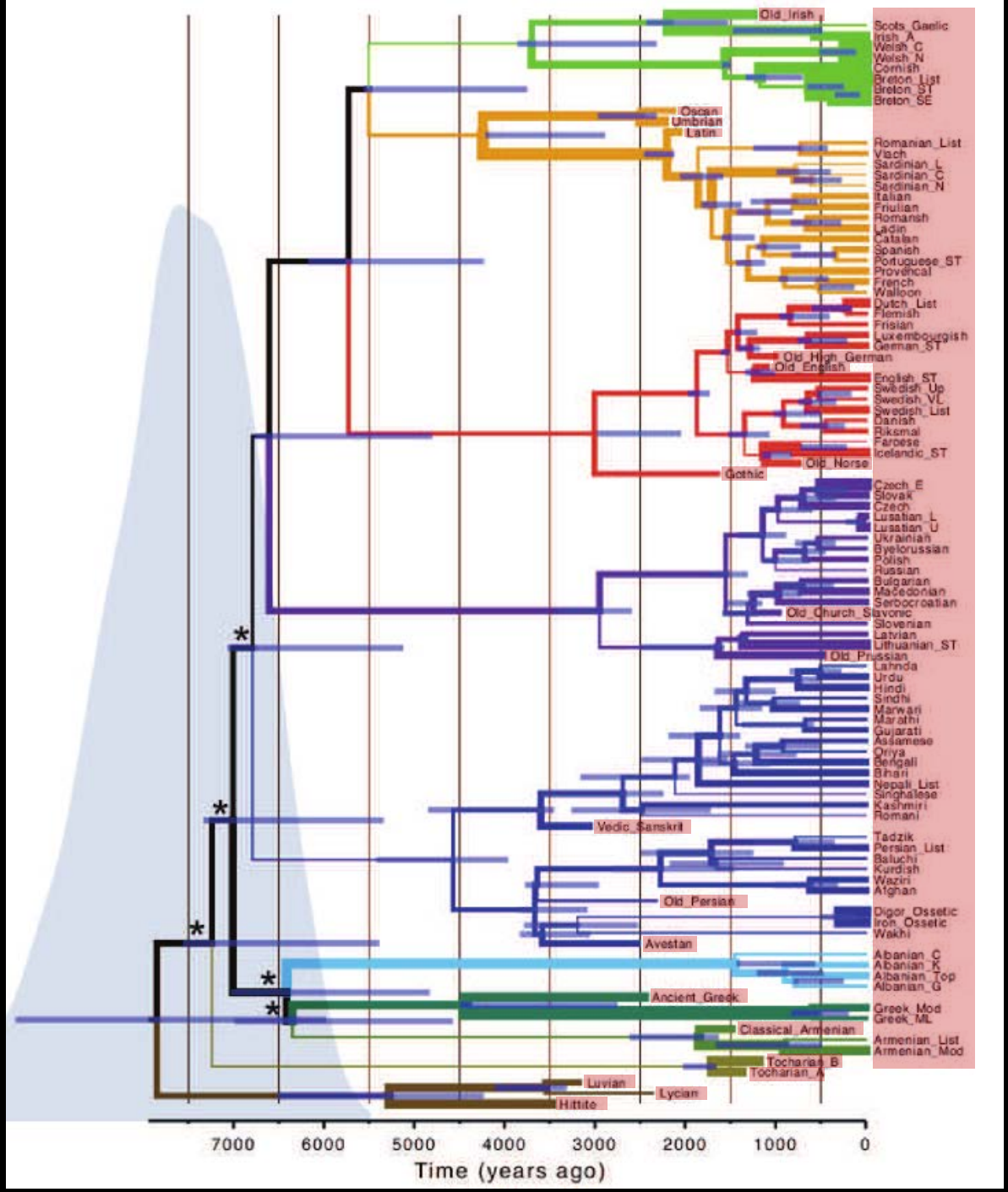
[Bouckaert+, Science 2012]

From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

問題設定

観測データ

- 葉ノードの状態



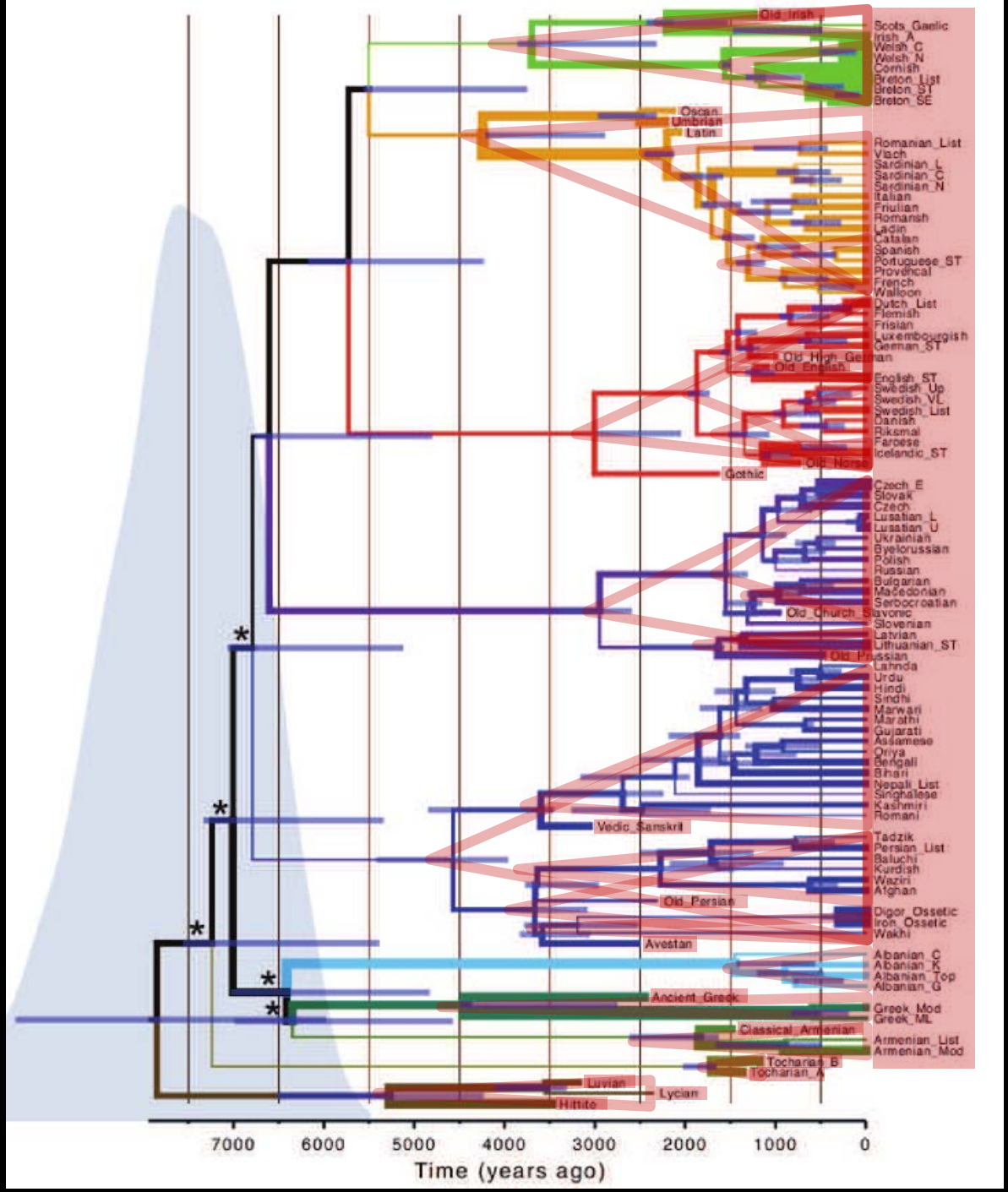
[Bouckaert+, Science 2012]

From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

問題設定

観測データ
- 葉ノードの状態

制約
- いくつかの単系統群



[Bouckaert+, Science 2012]

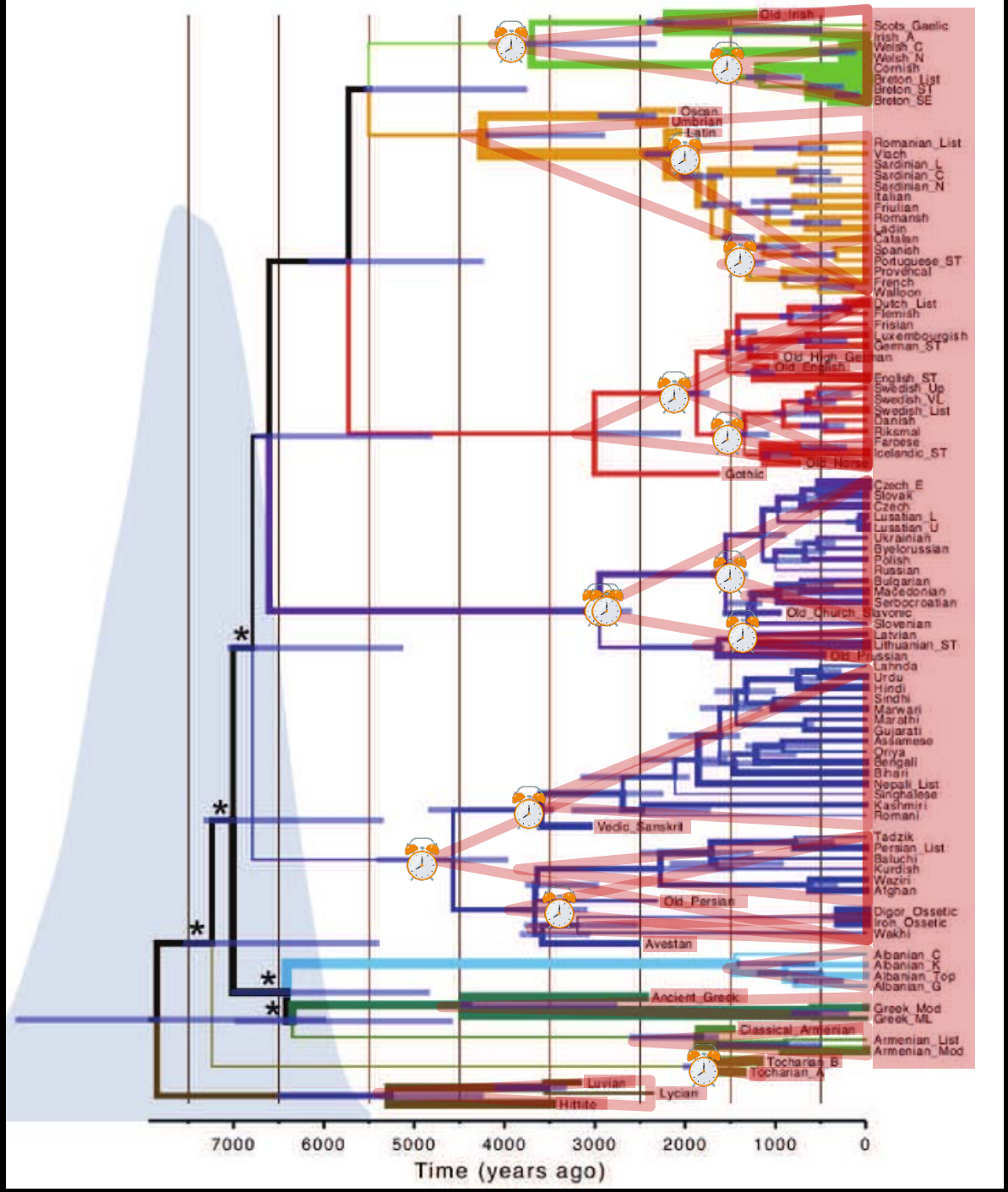
From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

問題設定

観測データ
- 葉ノードの状態

制約
- いくつかの単系統群
- いくつかの系統群のおおよその年代

[Bouckaert+, Science 2012]



From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

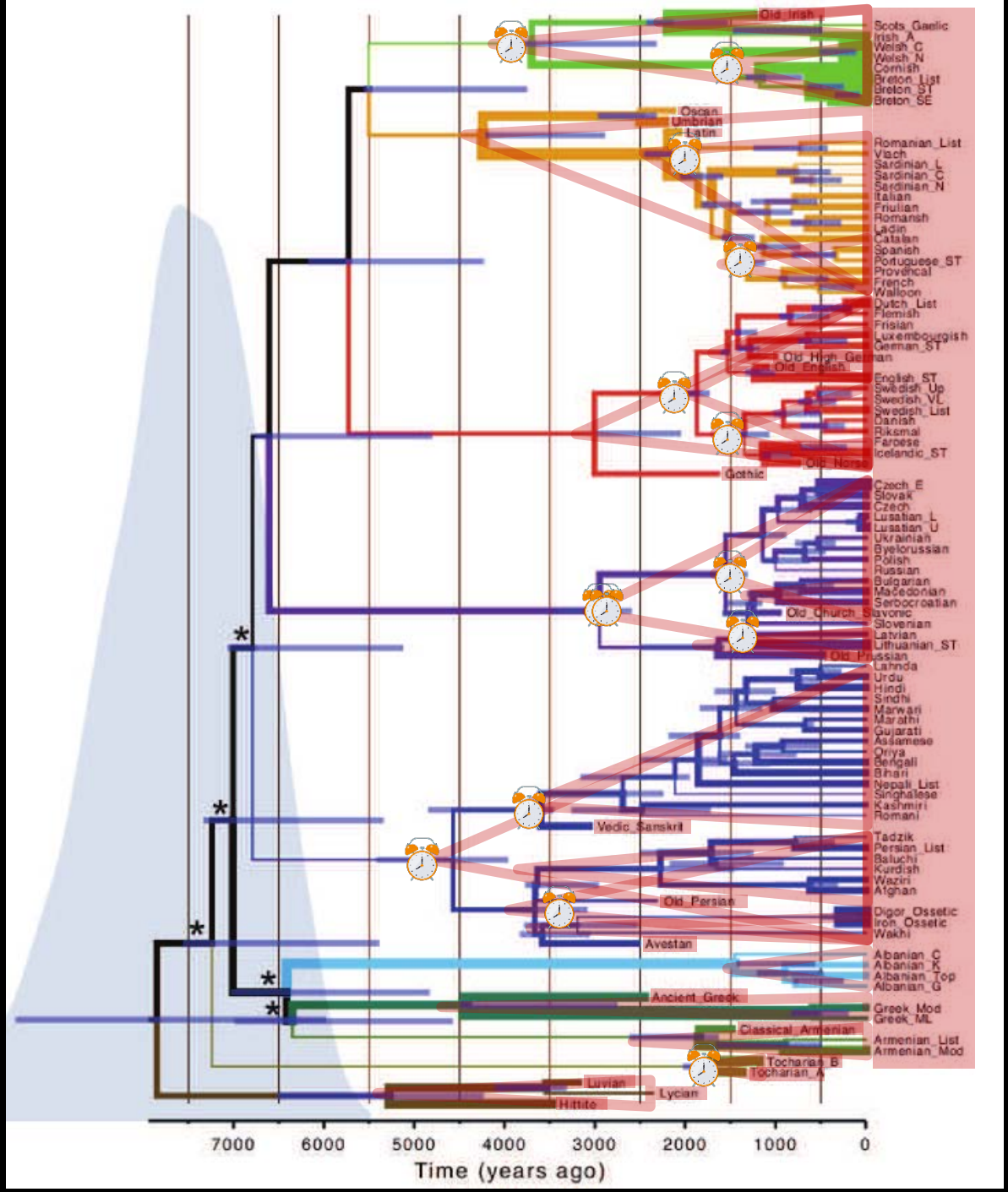
問題設定

観測データ
- 葉ノードの状態

制約
- いくつかの単系統群
- いくつかの系統群のおおよその年代

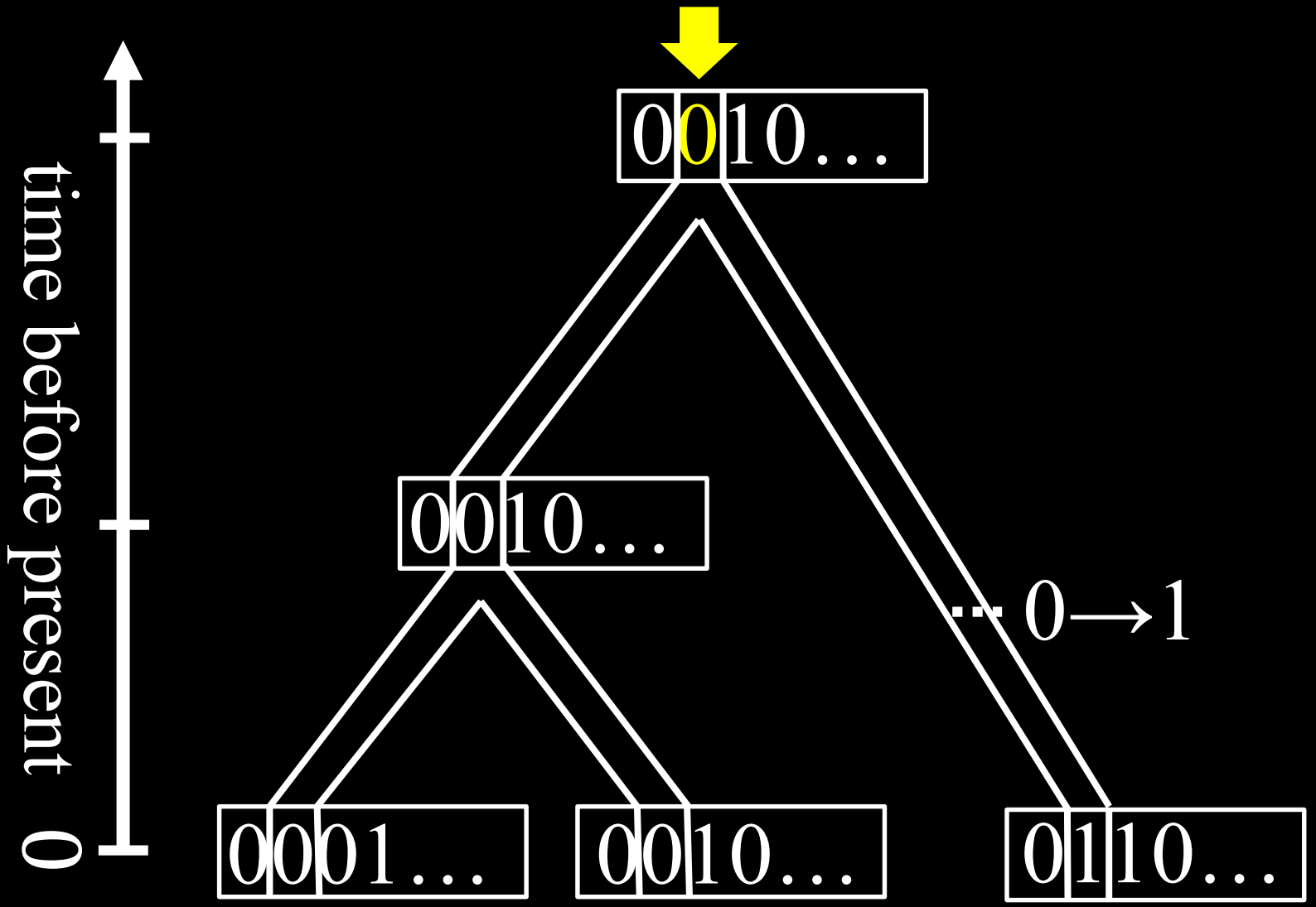
推論対象
- 木のトポロジー
  - 単系統群内部も
- 内部ノードの状態
- 内部ノードの年代

[Bouckaert+, Science 2012]

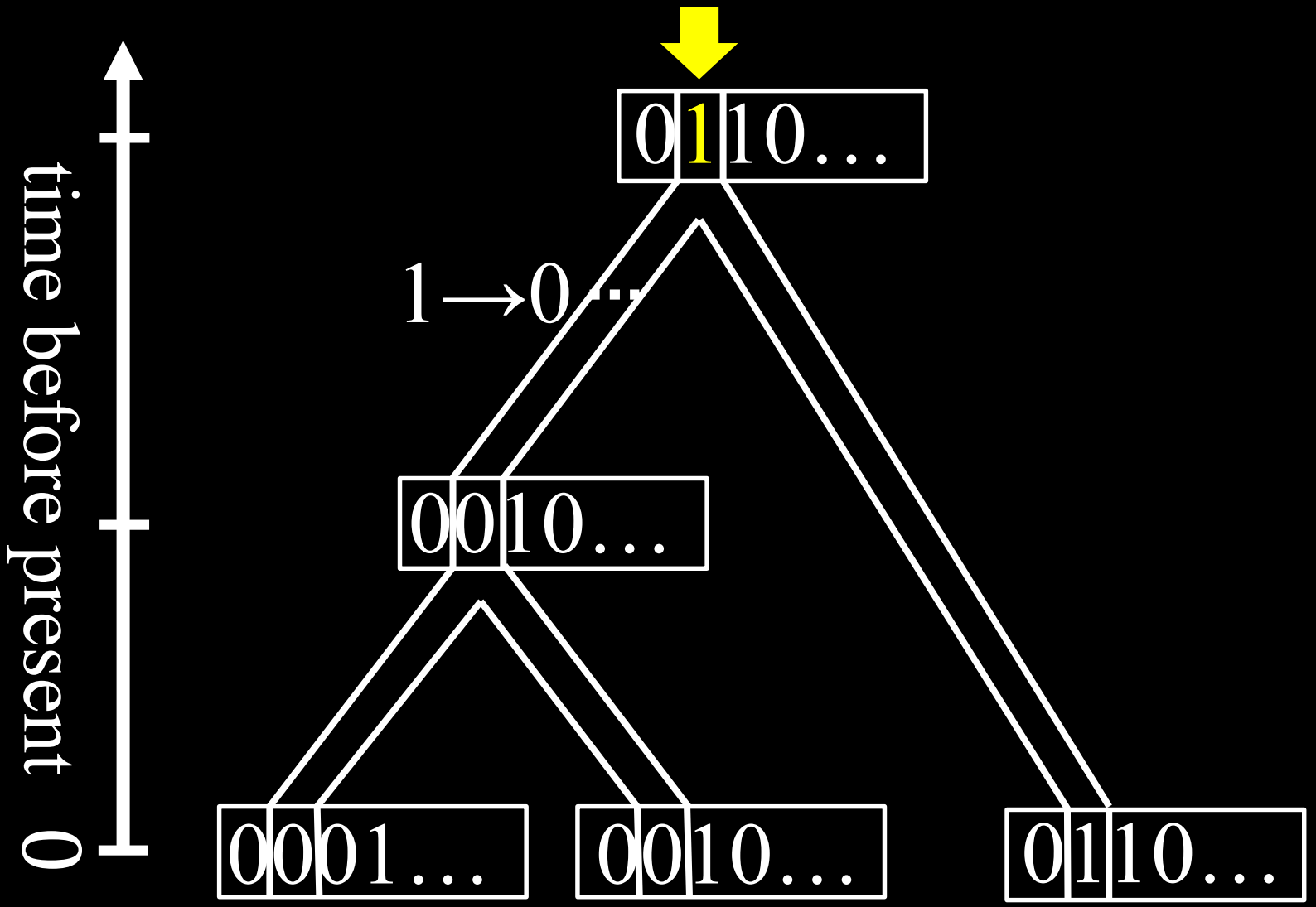


From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

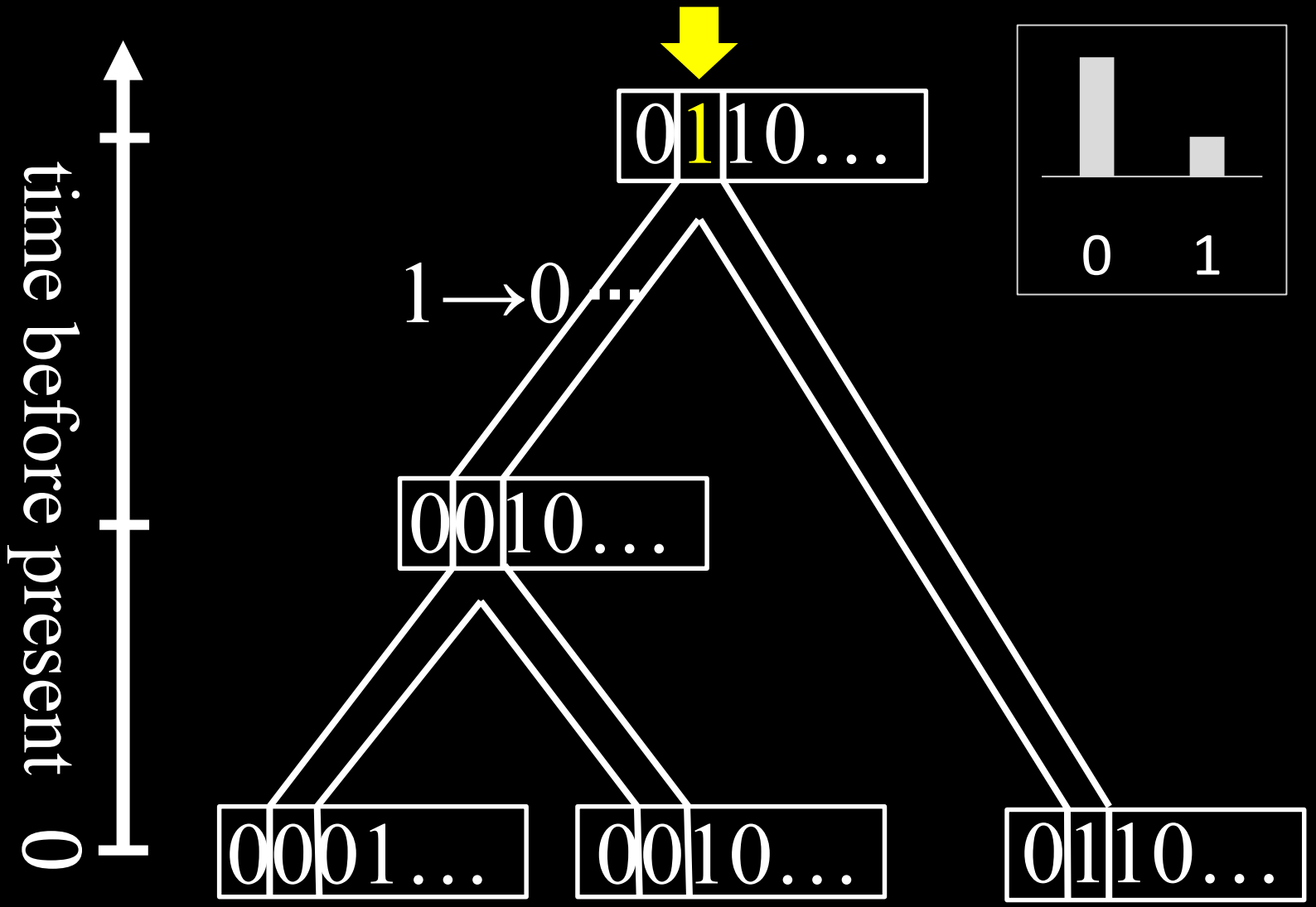
Q: 手持ちの手がかりだけでは
トポロジー/状態/年代を決めようがないのでは?



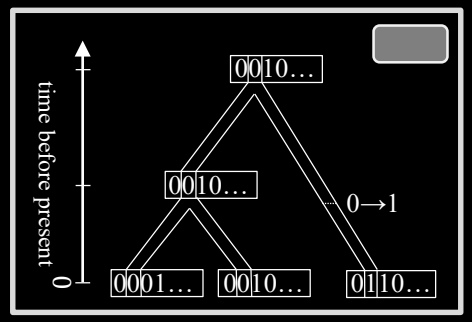
Q: 手持ちの手がかりだけでは
トポロジー/状態/年代を決めようがないのでは?



Q: 手持ちの手がかりだけでは
トポロジー/状態/年代を決めようがないのでは?

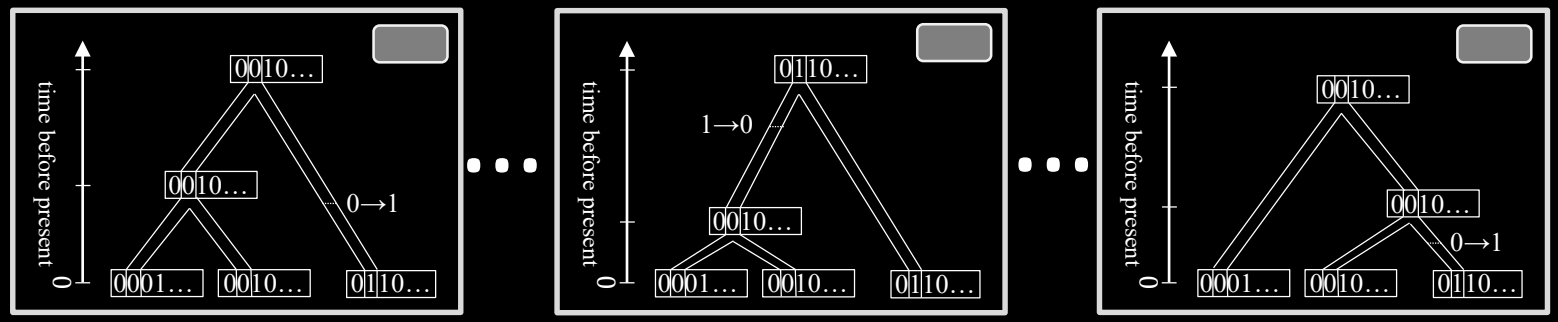


マルコフ連鎖モンテカルロ法 (MCMC)



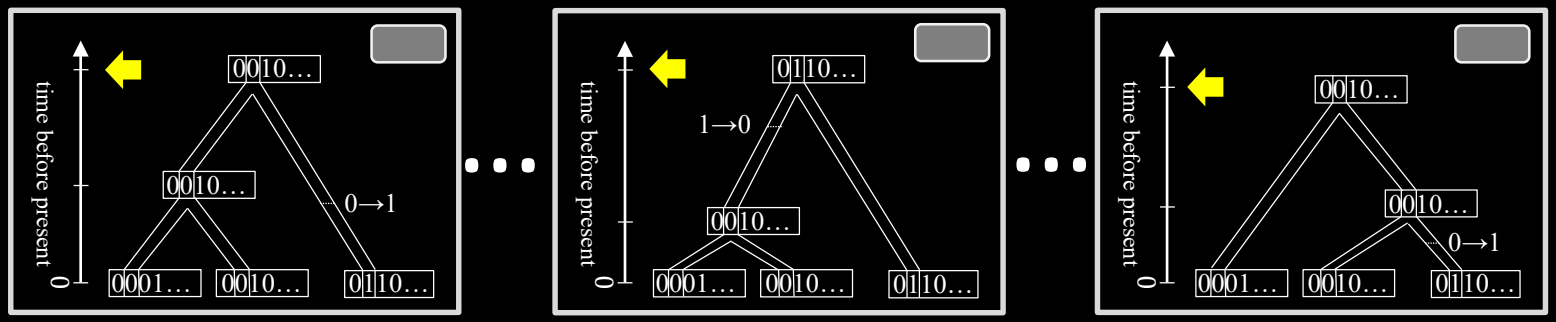
- 良さそうな木を計算機に大量に生成させる
  - 観測データは固定
  - 推論対象データを一定の条件にしたがってランダムに動かす

マルコフ連鎖モンテカルロ法 (MCMC)



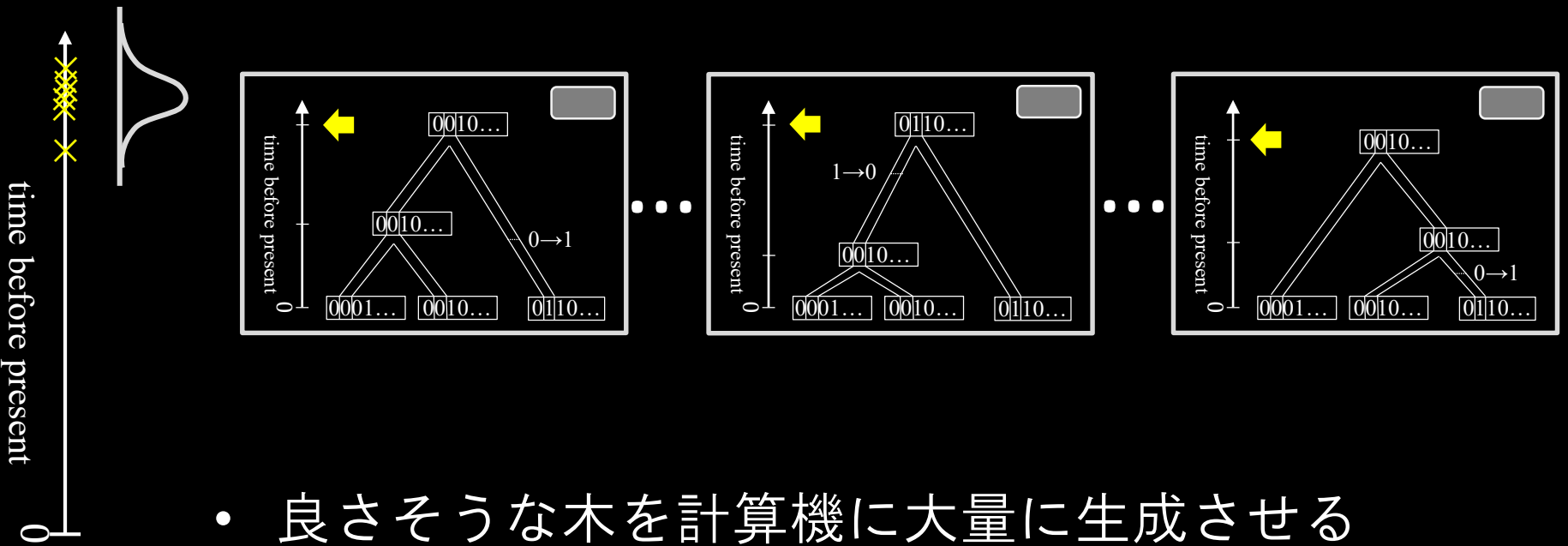
- 良さそうな木を計算機に大量に生成させる
  - 観測データは固定
  - 推論対象データを一定の条件にしたがってランダムに動かす

マルコフ連鎖モンテカルロ法 (MCMC)



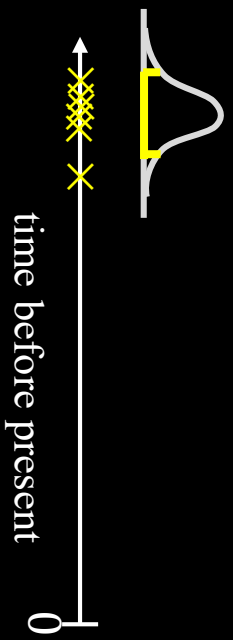
- 良さそうな木を計算機に大量に生成させる
  - 観測データは固定
  - 推論対象データを一定の条件にしたがってランダムに動かす
- 生成された大量の木を人間に見せるために要約

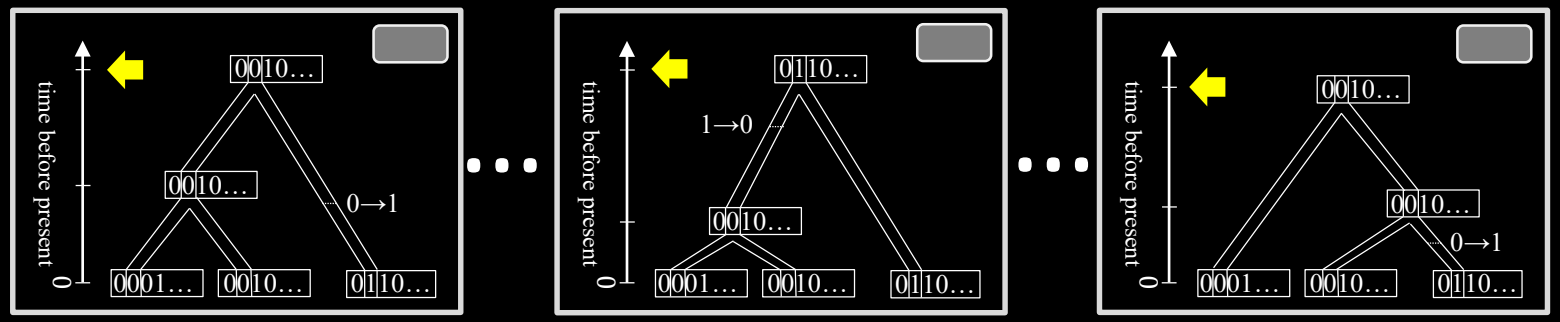
マルコフ連鎖モンテカルロ法 (MCMC)



- 良さそうな木を計算機に大量に生成させる
  - 観測データは固定
  - 推論対象データを一定の条件にしたがってランダムに動かす
- 生成された大量の木を人間に見せるために要約

マルコフ連鎖モンテカルロ法 (MCMC)



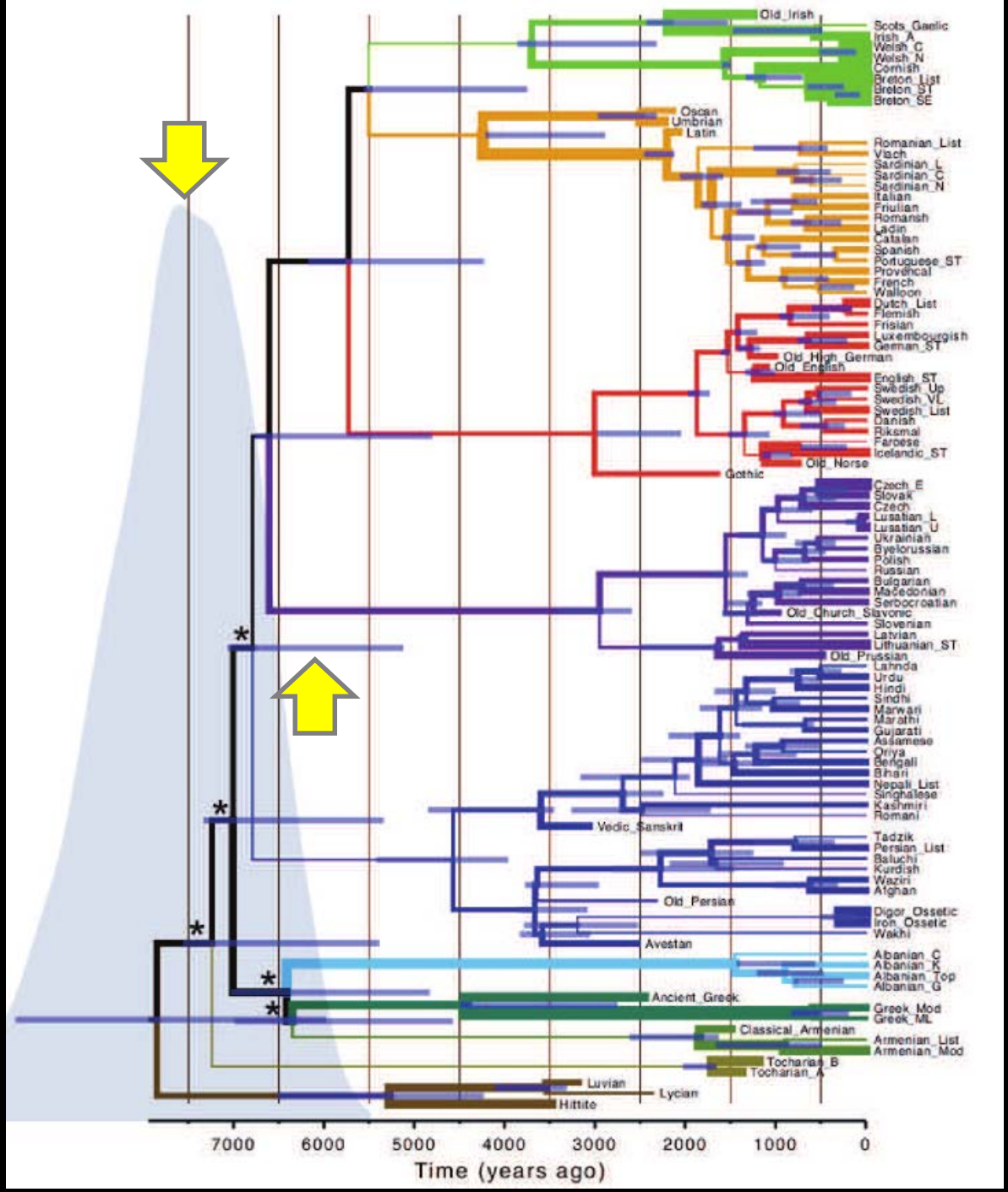


- 良さそうな木を計算機に大量に生成させる
  - 観測データは固定
  - 推論対象データを一定の条件にしたがってランダムに動かす
- 生成された大量の木を人間に見せるために要約

印欧語族年代つき系統樹 (再掲)

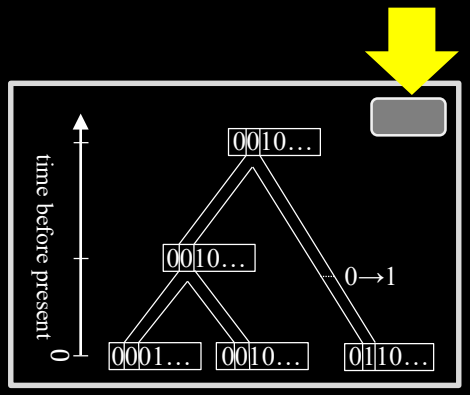
- 木のトポロジーも複数の木を要約したもの (最大系統群信頼度木)

[Bouckaert+, Science 2012]



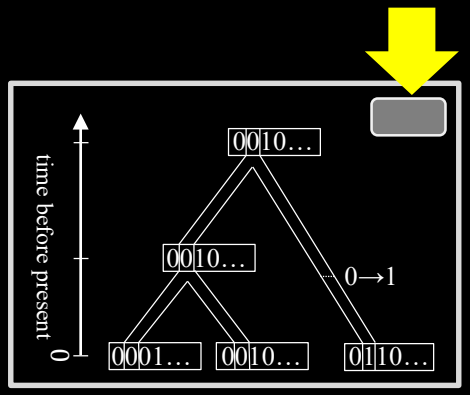
From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

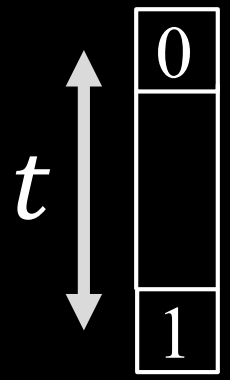
年代つき系統樹の良さ



- 確率的生成モデルは年代つき系統樹の良さを採点

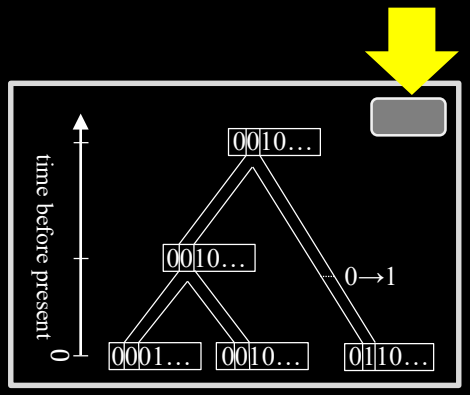
年代つき系統樹の良さ

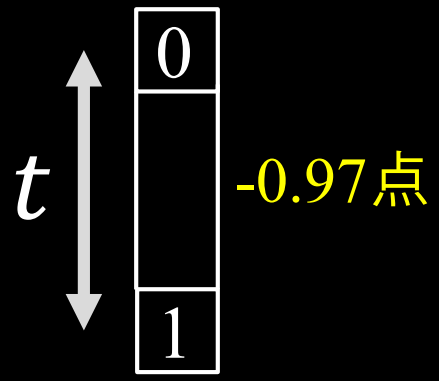




- 確率的生成モデルは年代つき系統樹の良さを採点

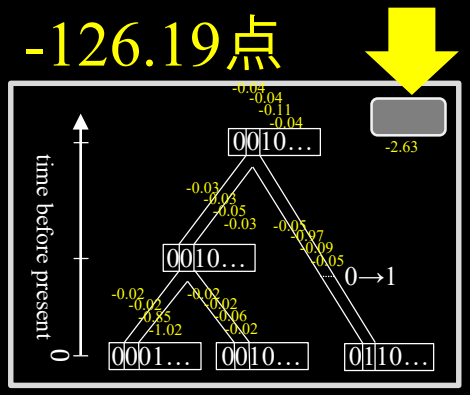
年代つき系統樹の良さ

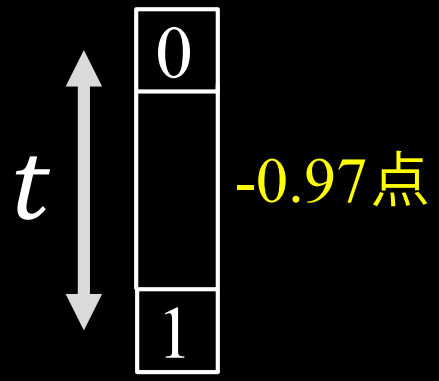




- 確率的生成モデルは年代つき系統樹の良さを採点

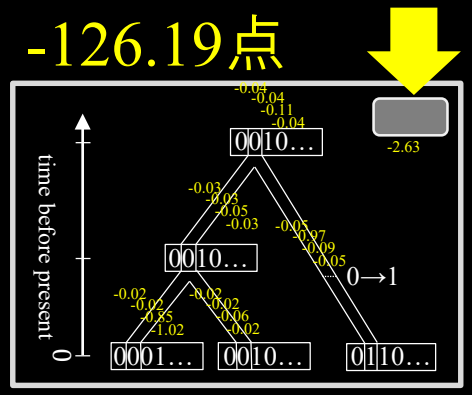
年代つき系統樹の良さ

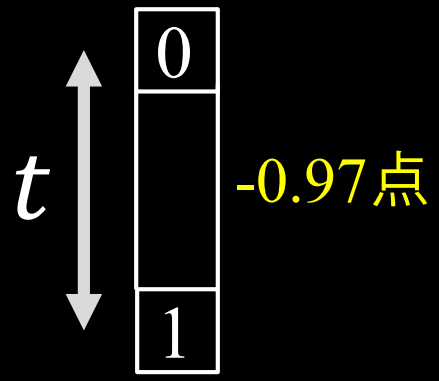




- 確率的生成モデルは年代つき系統樹の良さを採点

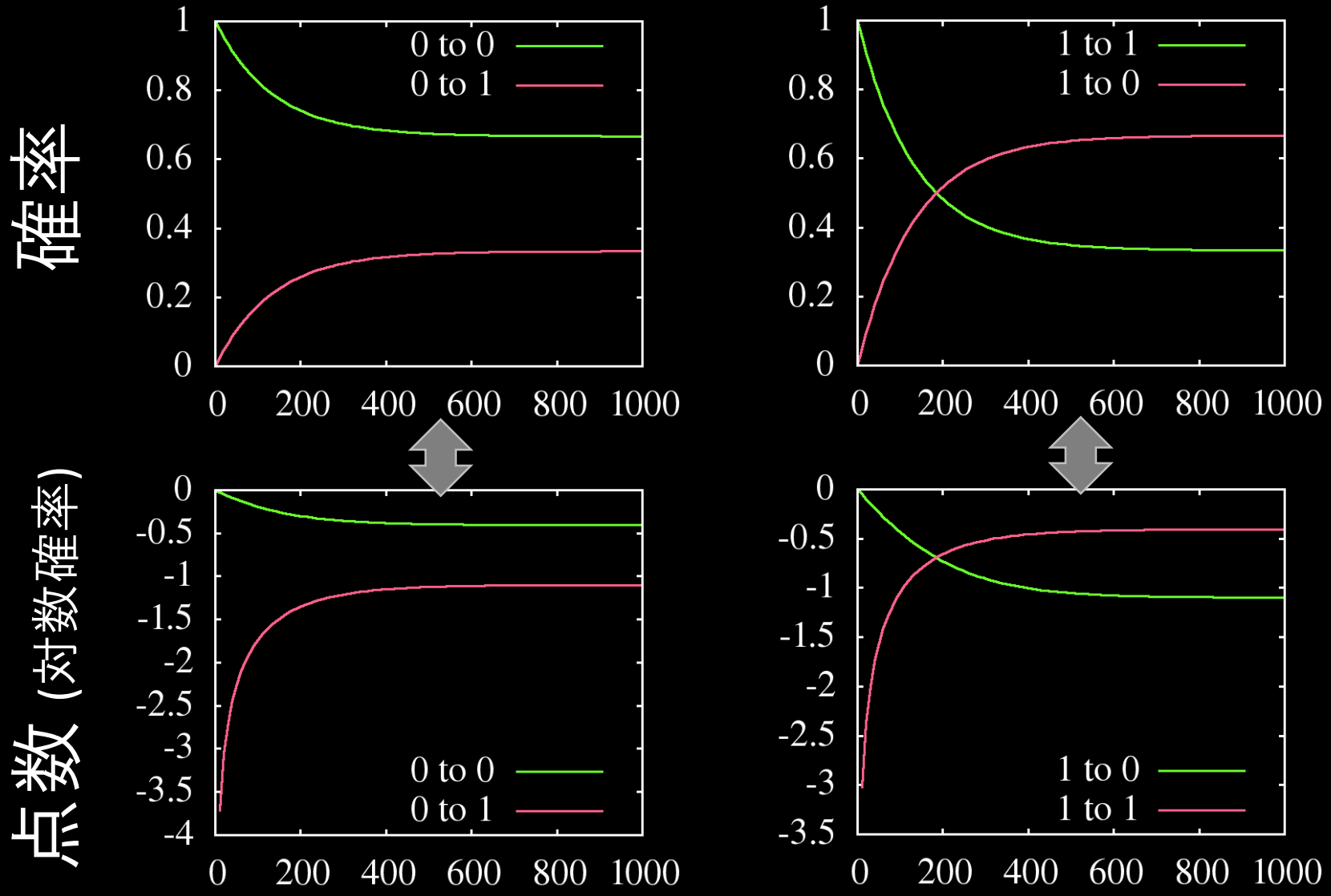
年代つき系統樹の良さ



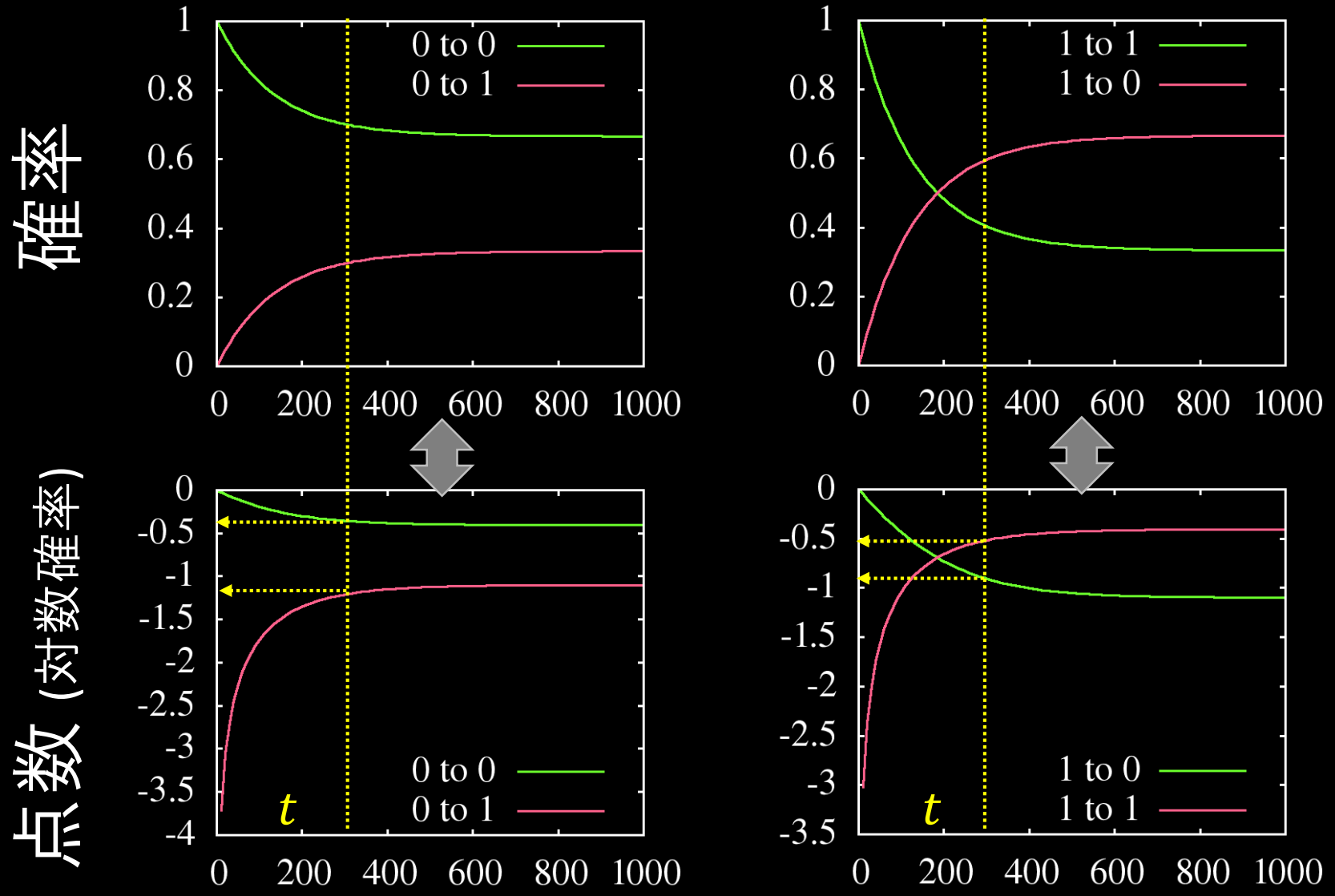


- 確率的生成モデルは年代つき系統樹の良さを採点
- モデルはデータの振る舞いに対する仮定
  - 年代つき系統樹モデルは数多くの仮定の産物

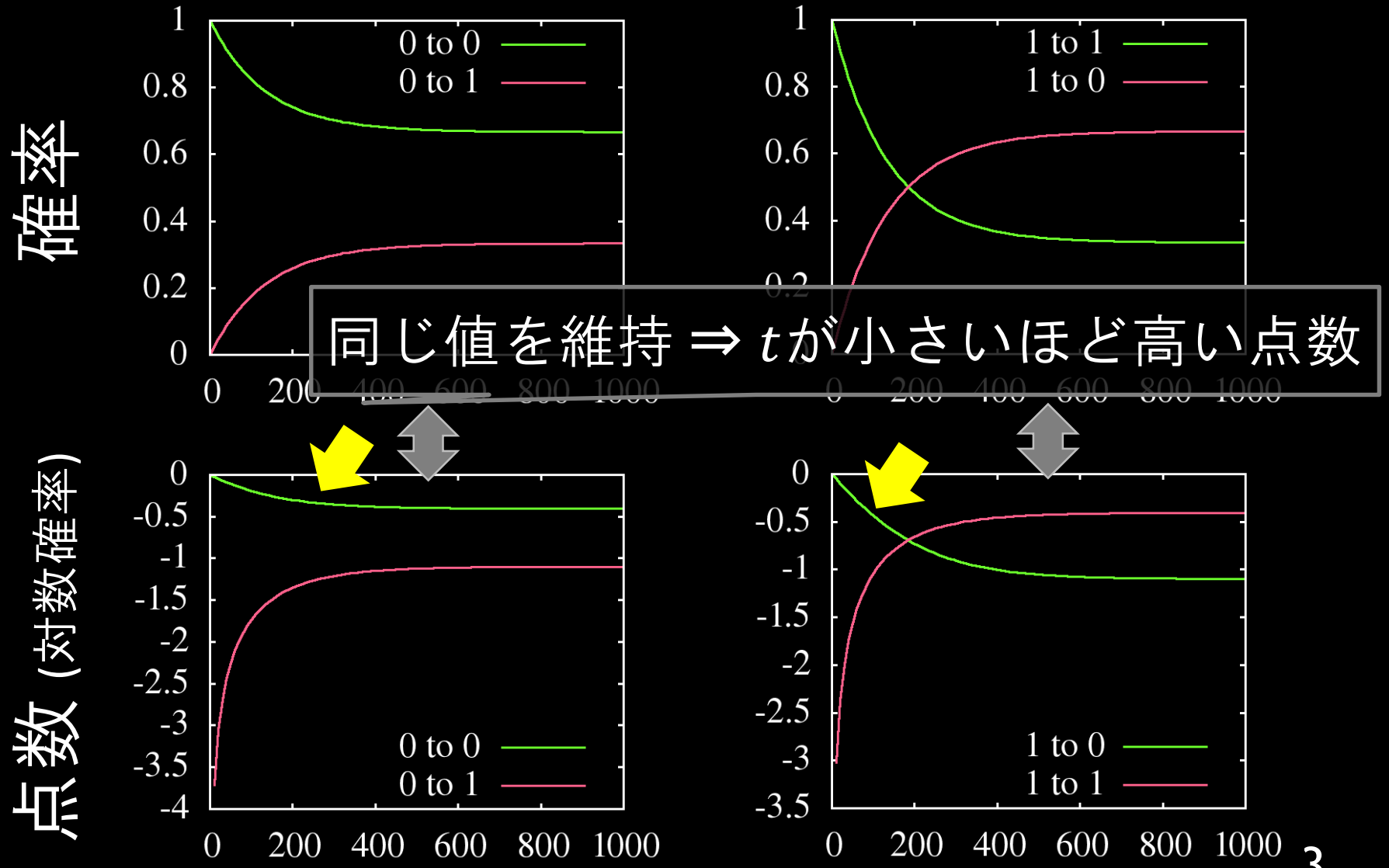
置換モデル (連続時間マルコフモデル)



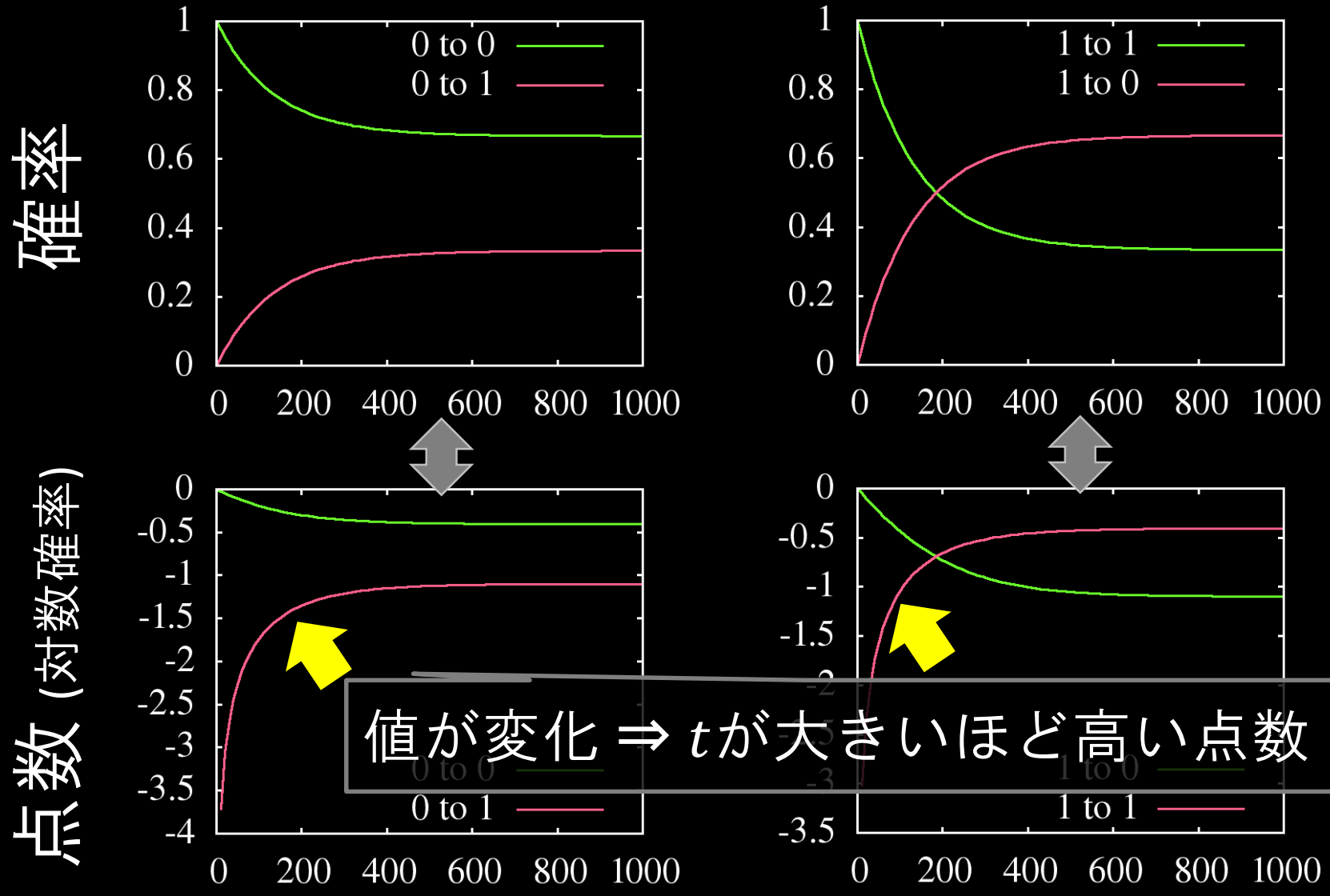
置換モデル (連続時間マルコフモデル)



置換モデル (連続時間マルコフモデル)

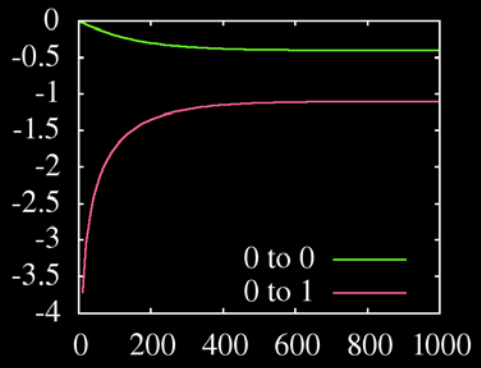


置換モデル (連続時間マルコフモデル)

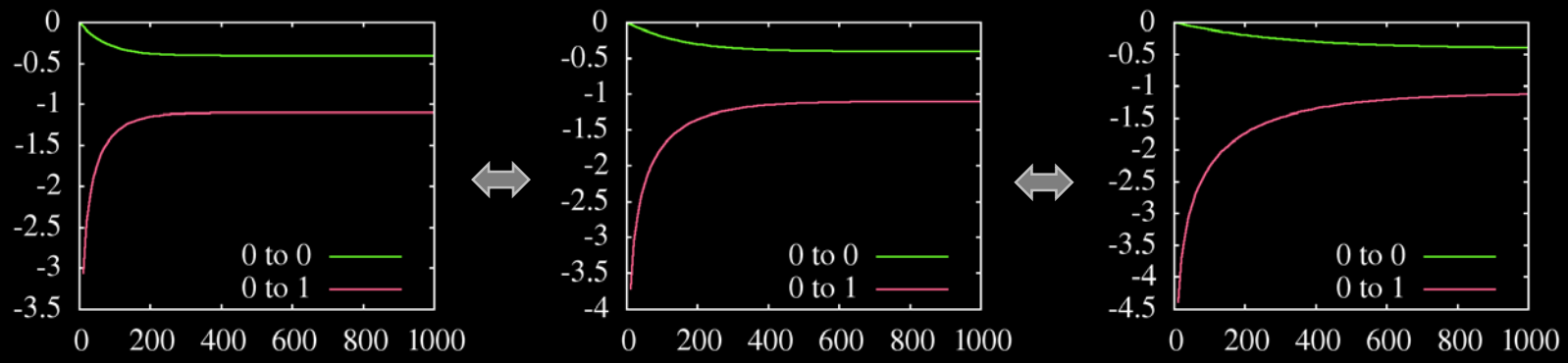


値が変化 ⇒ tが大きいほど高い点数

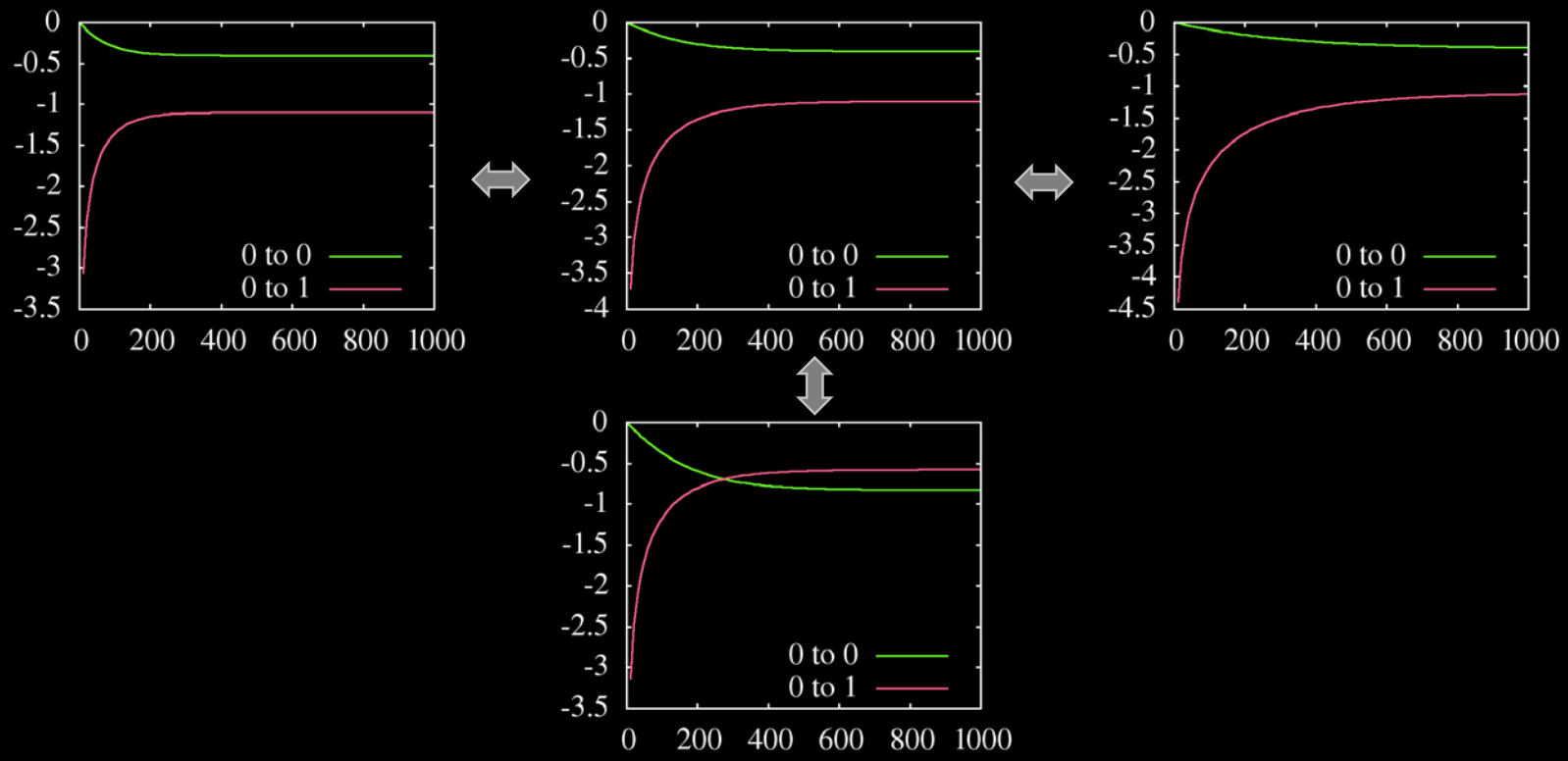
置換モデル自体も採点対象



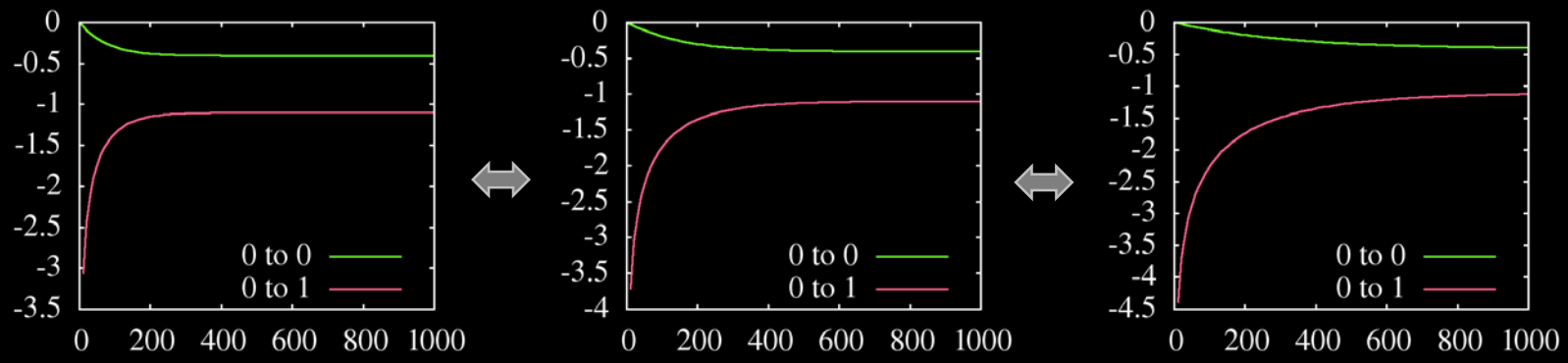
置換モデル自体も採点対象



置換モデル自体も採点対象

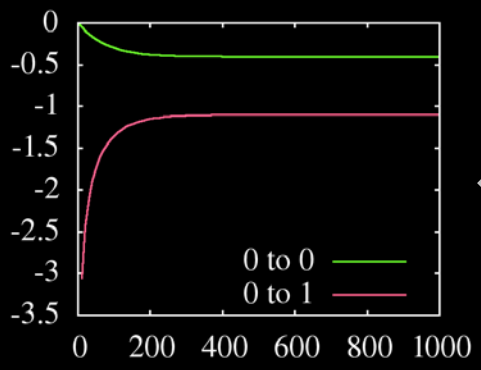


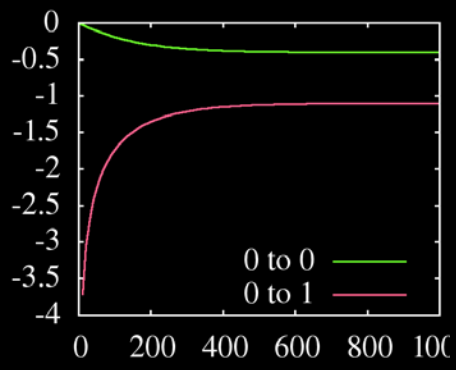
置換モデル自体も採点対象

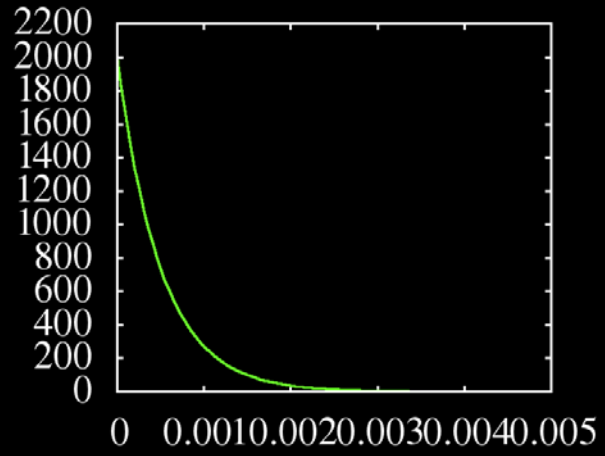


$$\begin{pmatrix} * & \alpha \\ \beta & * \end{pmatrix}$$

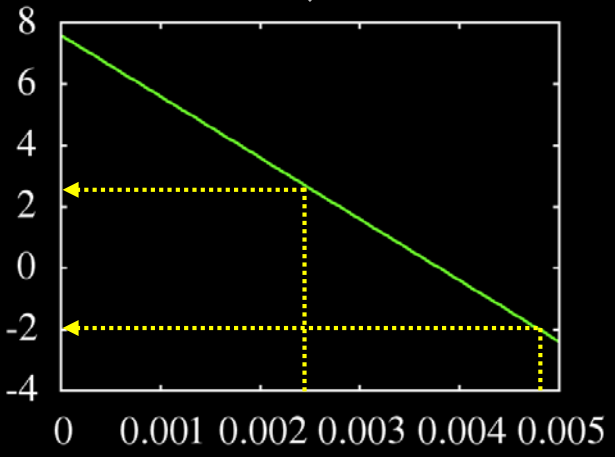
置換モデル自体も採点対象







$$\begin{pmatrix} * & \alpha \\ \beta & * \end{pmatrix}$$



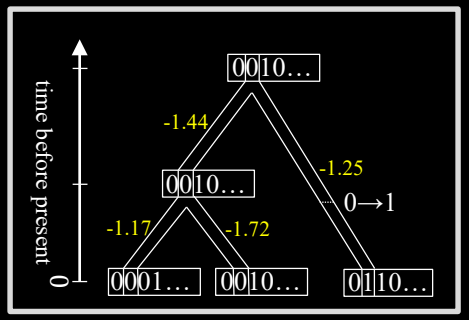
モデルやデータに対する様々な疑問

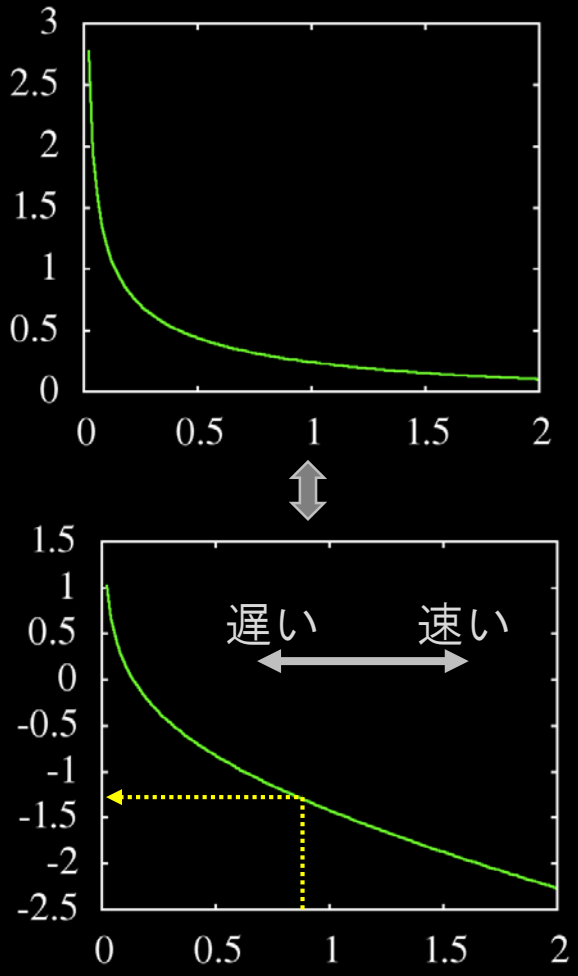
- 変化の速さは共通?
  - 言語によって異なるのでは?
  - 語によって異なるのでは?
- 系統樹モデルは接触の影響を無視するけど、本当は無視できないのでは?
- 借用語は事前に取り除いているというけど、本当に充分?
- 単純な置換モデルでは、誕生→死亡→誕生という変化があり得るけど、一度消えた復活するのは不自然では?
- 何か体系的なバイアスがデータに潜んでいない?

Q: 変化の速さはすべての言語に共通?

- アイスランド語は超保守的
  [Bergsland+, 1962]
- 系統樹の枝kごとに変異率r_kを導入 (緩和時計モデル)

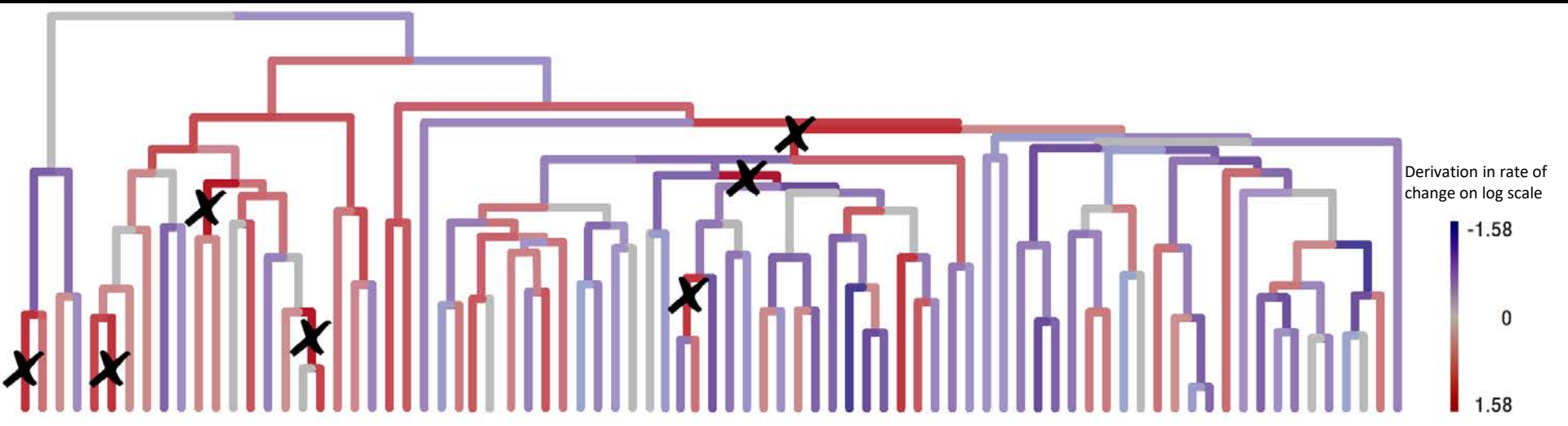
$$r_k \begin{pmatrix} * & \alpha \\ \beta & * \end{pmatrix}$$





Q: 変化の速さはすべての言語に共通?

- オーストロネシア語族における変化の緩急 [Greenhill+, PNAS 2017]



- 古い時代は変化の速さを決める手がかりにとぼしいので、緩和時計モデルを用いると、共通祖語の年代の分布はすそが広くなるはず

Q: 意味変化による体系的バイアス?

[Chang+, Language 2015]

- 成因的相同 (homoplasy): 同じ特徴が独立に発生
- 実は無視できないほど頻出
  - IELEXのロマンス諸語の基礎語彙の8.1%

*$d^h\acute{g}^h$om-, ADULT MALE

現代アイルランド語
duine

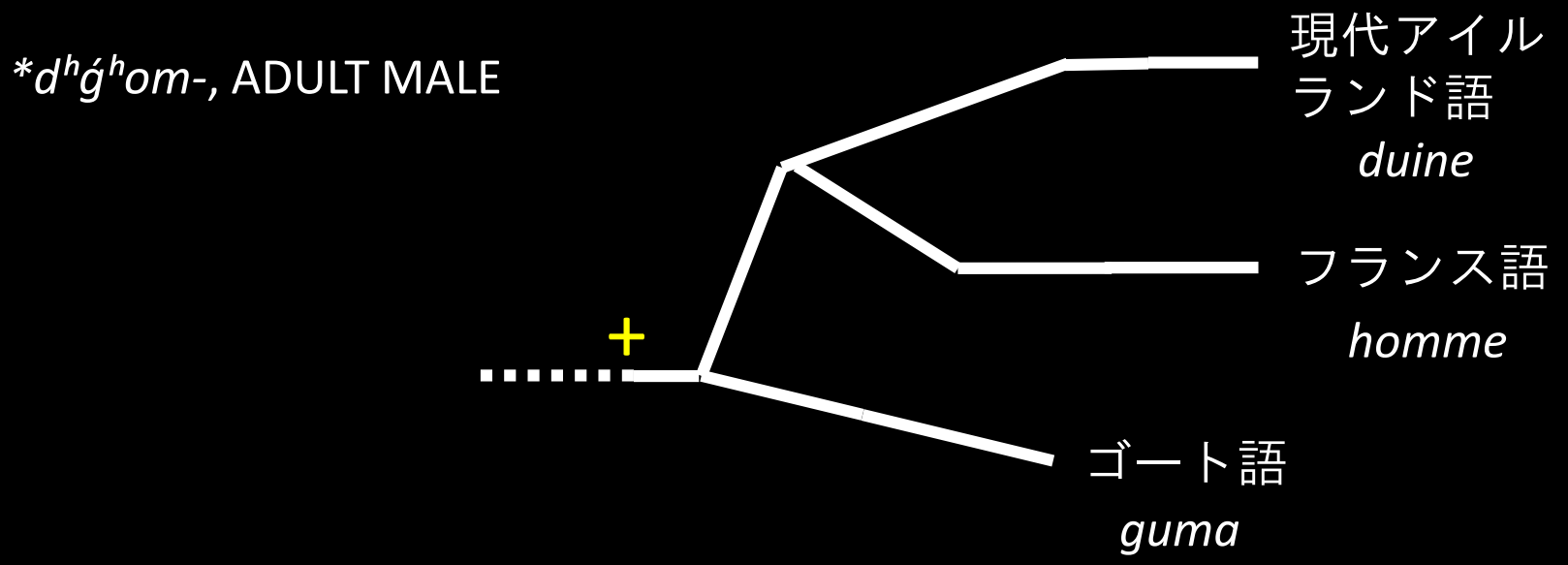
フランス語
homme

ゴート語
guma

Q: 意味変化による体系的バイアス?

[Chang+, Language 2015]

- 成因的相同 (homoplasy): 同じ特徴が独立に発生
- 実は無視できないほど頻出
  - IELEXのロマンス諸語の基礎語彙の8.1%



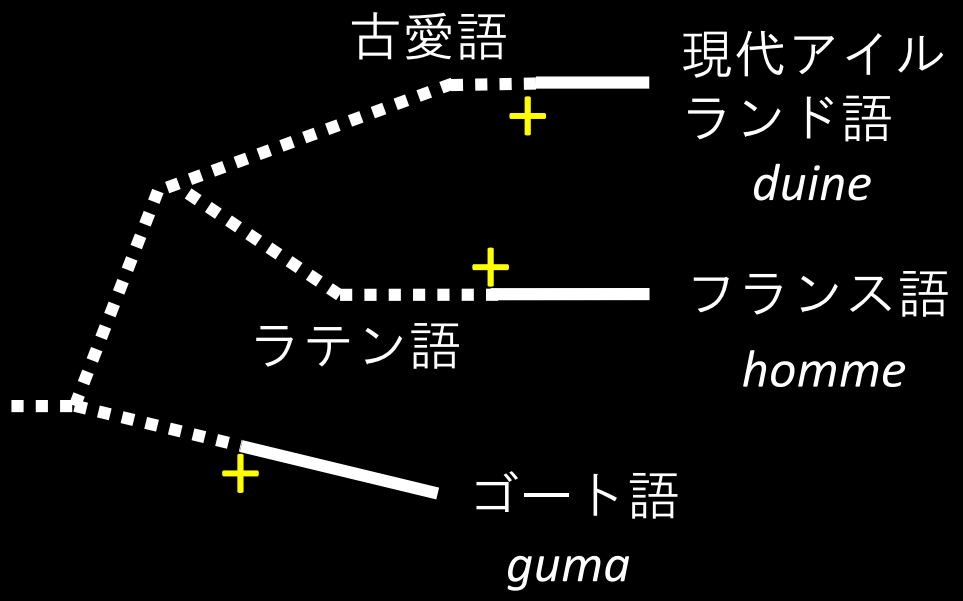
Q: 意味変化による体系的バイアス?

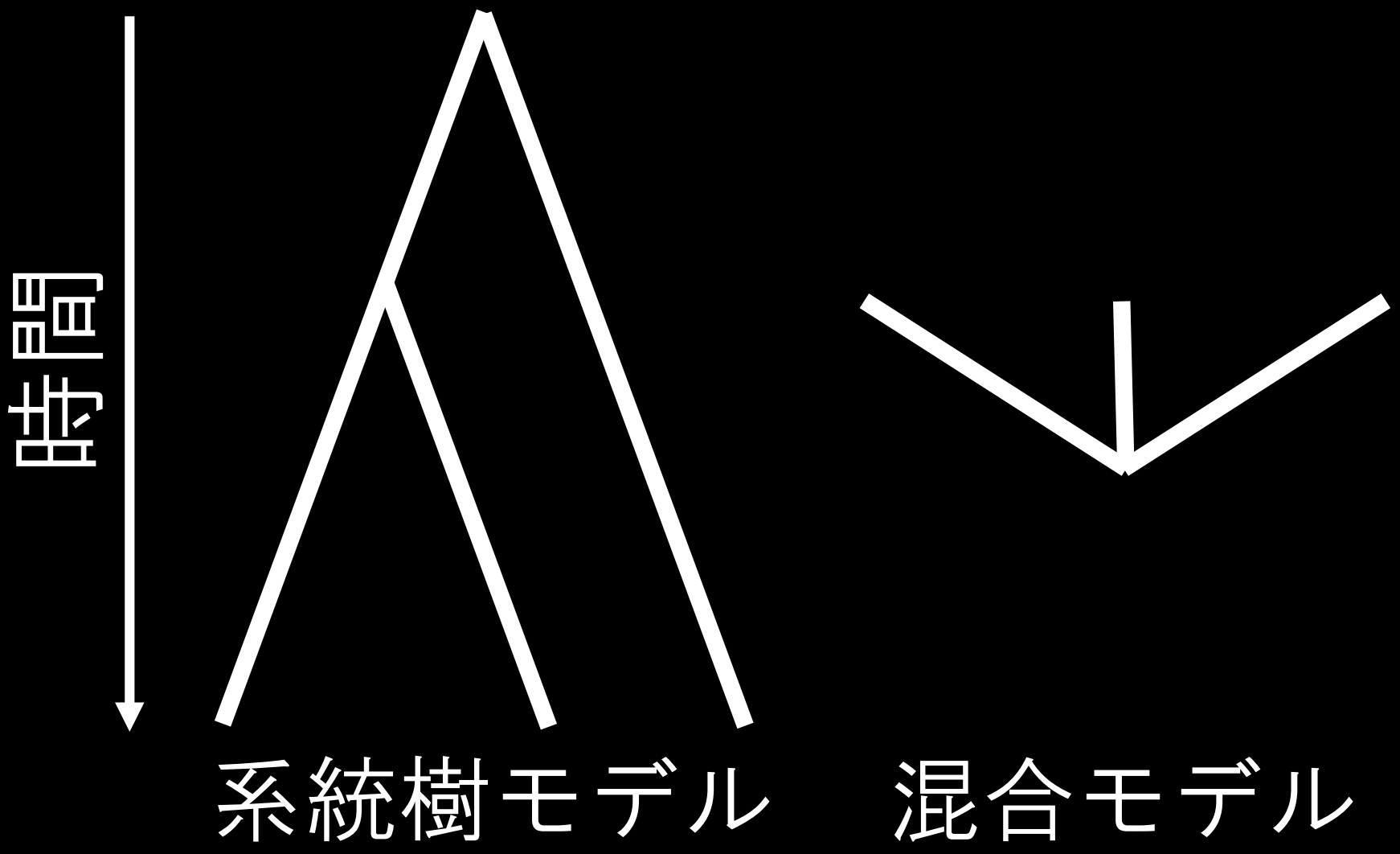
[Chang+, Language 2015]

- 提案手法: 古代語を制約として使う
- 結果: 印欧祖語の年代は約6,500年前に繰り上がり、クルガン仮説寄りに

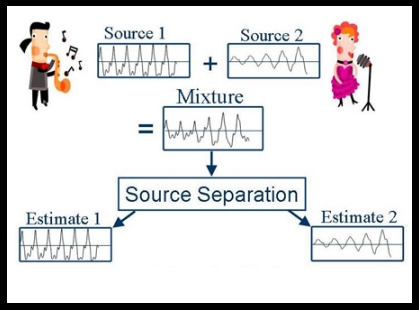
$d^h\acute{g}^h$om-, ADULT MALE

PERSON → ADULT MALE
の意味変化が独立に発生
- ラテン語: *homo*, PERSON
- 古愛語: *duine*, PERSON

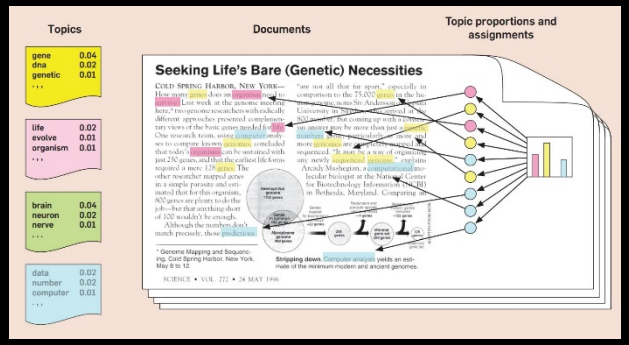



時間
系統樹モデル
混合モデル

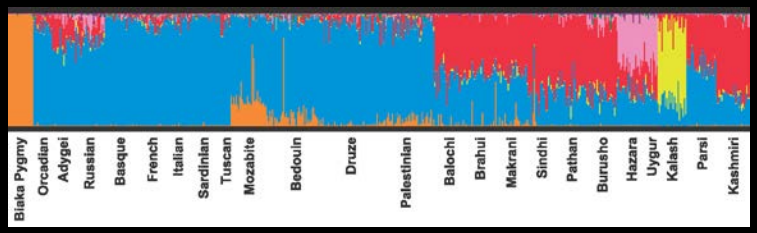
混合モデル



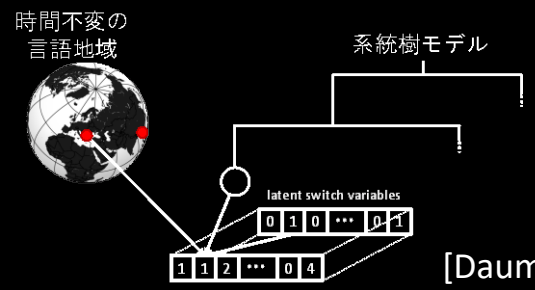
音源分離



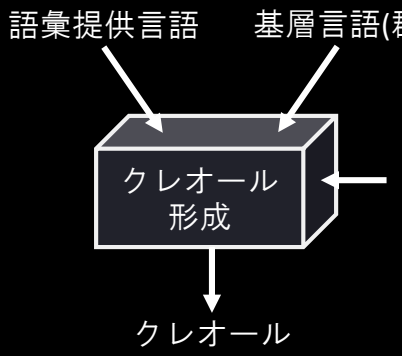
文書のトピックモデル [Blei+, 2012]



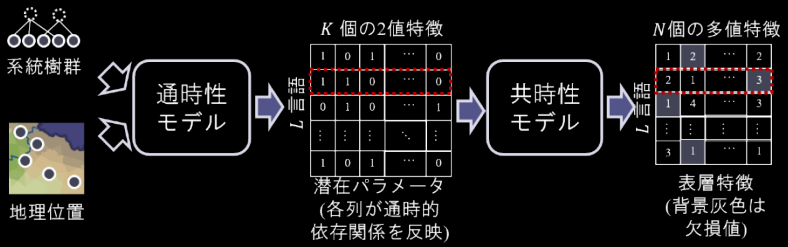
集団遺伝学における混合分析 [Pritchard+, 2000]



[Daumé III, NAACL2009]

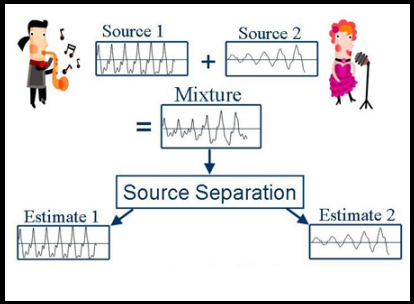


[Murawaki, NAACL2016]

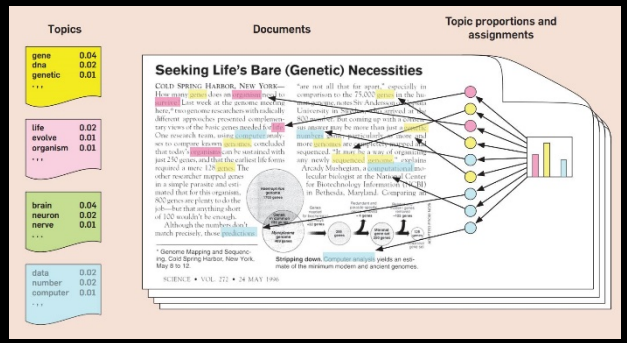


[Yamauchi+, COLING2016] [Murawaki, IJCNLP2017]

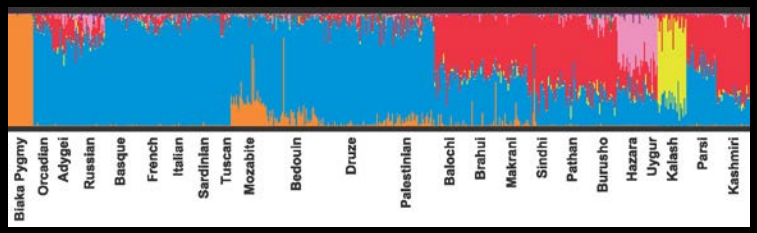
混合モデル



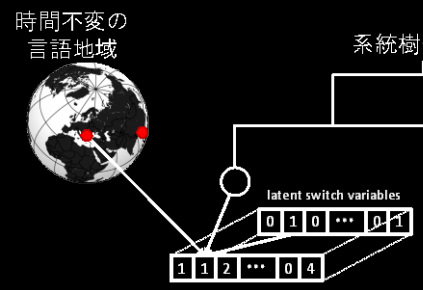
音源分離



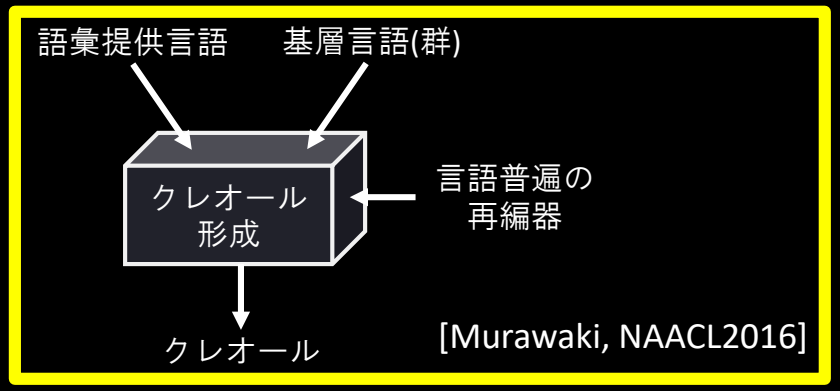
文書のトピックモデル [Blei+, 2012]



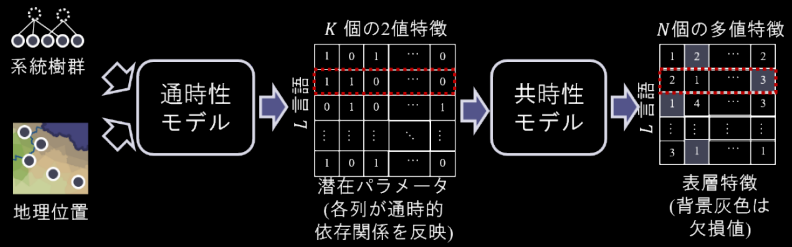
集団遺伝学における混合分析 [Pritchard+, 2000]



[Daumé III, NAACL2009]

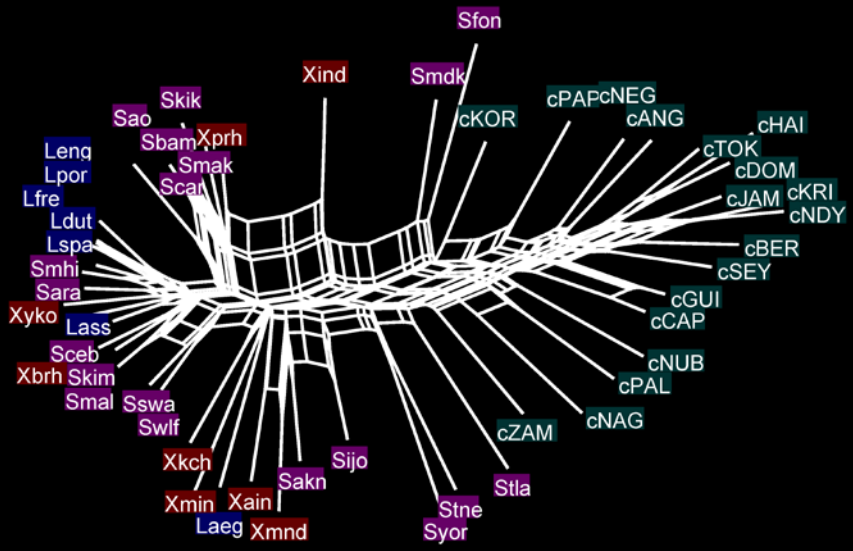


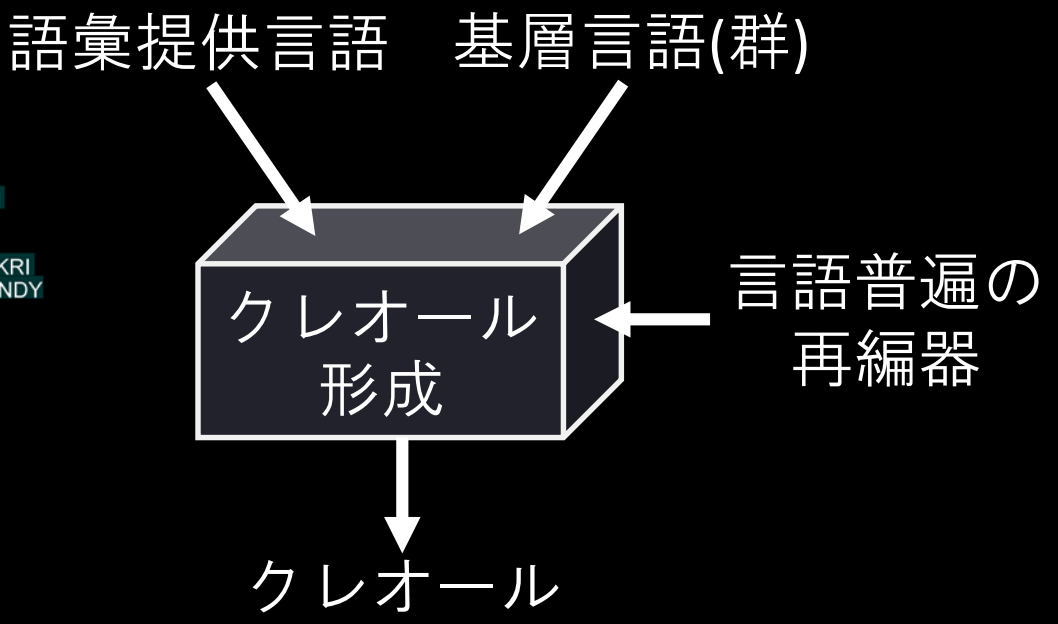
[Murawaki, NAACL2016]



[Yamauchi+, COLING2016] [Murawaki, IJCNLP2017]

クレオール形成のモデル化





Daval-Markussen+ (2012):
クレオールと非クレオールの(根なし)系統樹モデル

Murawaki (NAACL2016):
3種類の言語からクレオールを確率的に生成する混合モデル

まとめ

- 人間は苦手だが計算機なら解ける問題は計算機にやらせればよい
  - ベイズ統計は不確実性を自然に扱う枠組み
- モデルとはデータに対する仮定
  - 仮定次第で結果が大きく変わる (場合がある)
  - すべてが明示されていて検証可能
- 個人的な課題は系統樹モデルの克服
- 統一的な基準で作成された言語データを機械可読な形でどんどん公開してほしい