Statistical/computational phylogenetics

Alexandre Bouchard-Côté

Department of Statistics,

University of British Columbia

Organization

- Background in phylogenetics
- Non-standard application areas:
  - Clonal evolution inside an individual cancer tumour
  - Reconstructing ancient languages
- Current research on Divide-and-Conquer Sequential Monte Carlo methods

Background on phylogenetics

Notation/background on trees

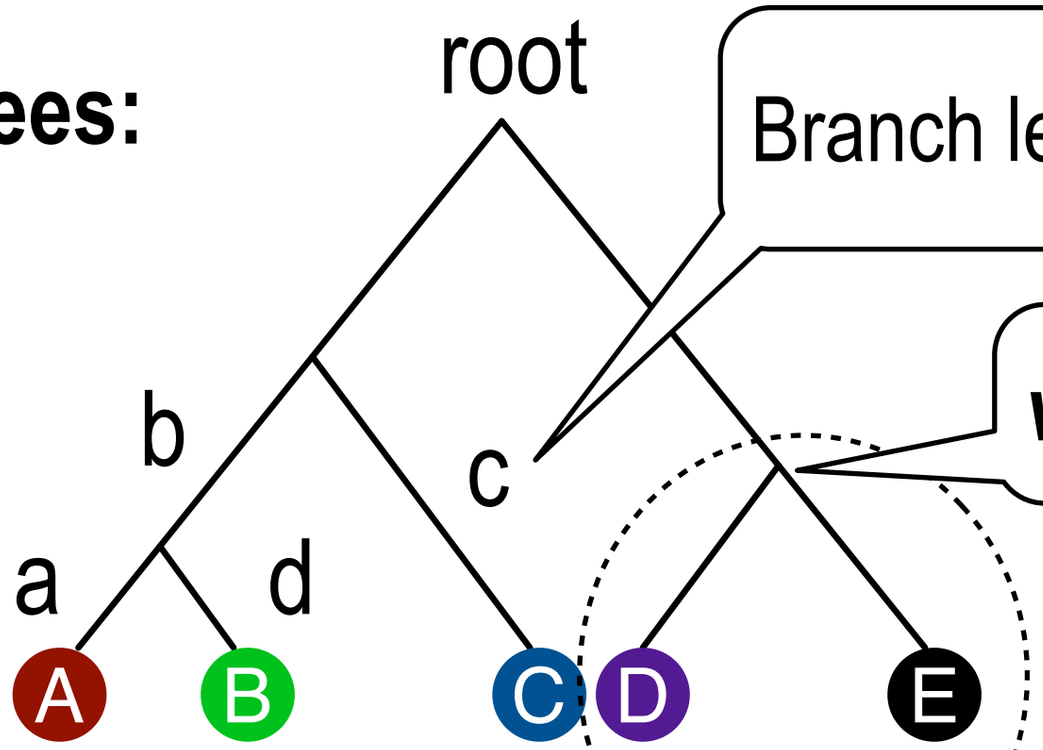
Clock trees:

a+b = c

d+b = c

a = d

...



Branch lengths

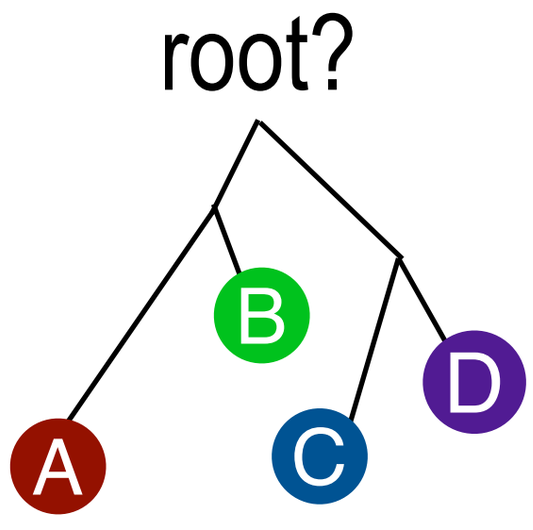
$v \in V$

Hidden sequences (typically)

Observations (typically)

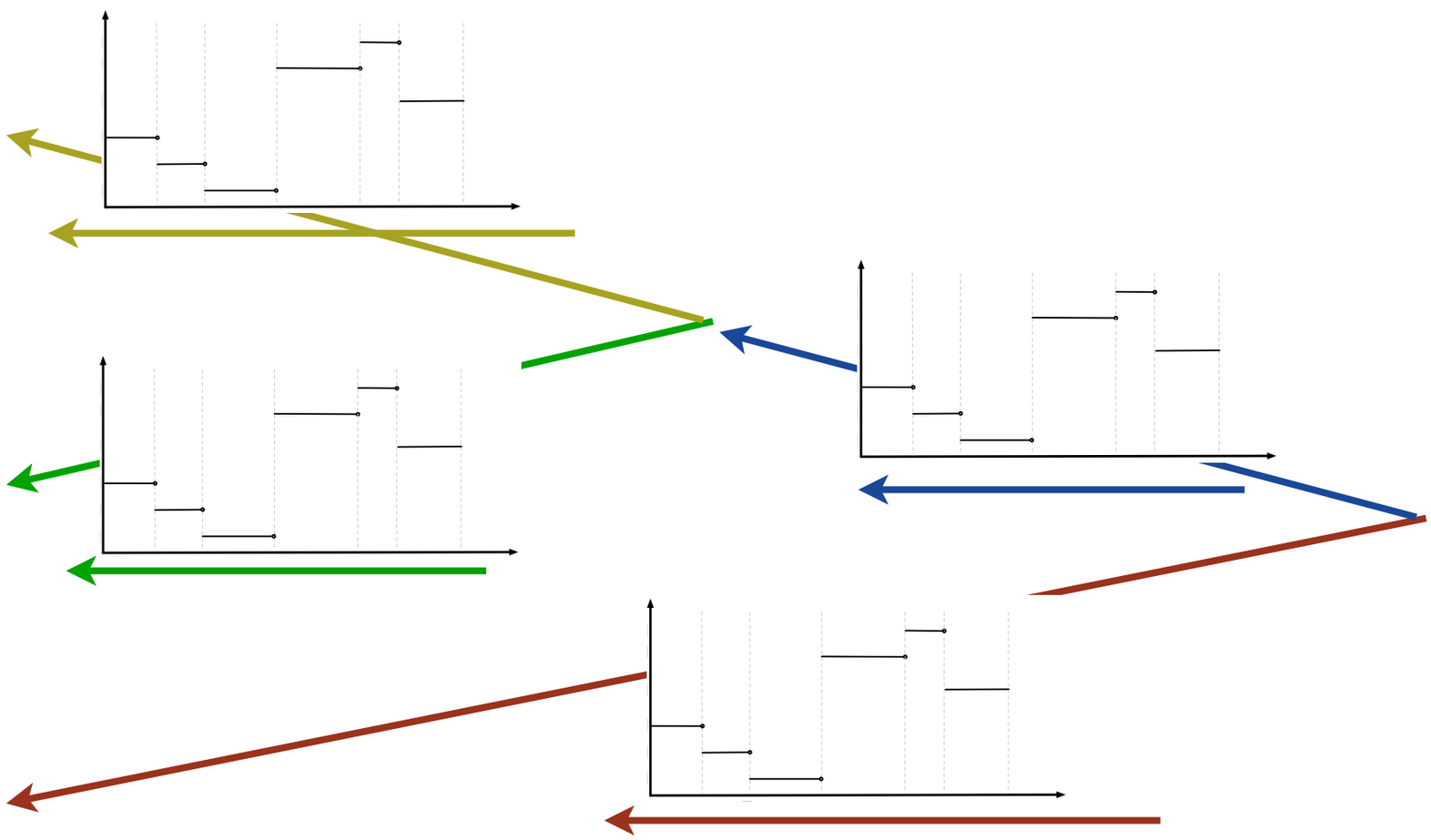
Clade: group containing a taxa and all its descendants. (I will restrict them to leaves, e.g. {D,E})

Non-clock tree: remove additivity restrictions on branch lengths



Stochastic process on a tree

- Evolution modeled by a stochastic process
  - Discrete data: Continuous time Markov chains (CTMC)
  - Continuous data: Diffusions



Phylogenetics: challenges

reality / data → estimated tree → biological/ clinical discovery

Where can this go wrong?

reality → probabilistic model → tree estimator → approximation algorithm

modeling errors

statistical issues

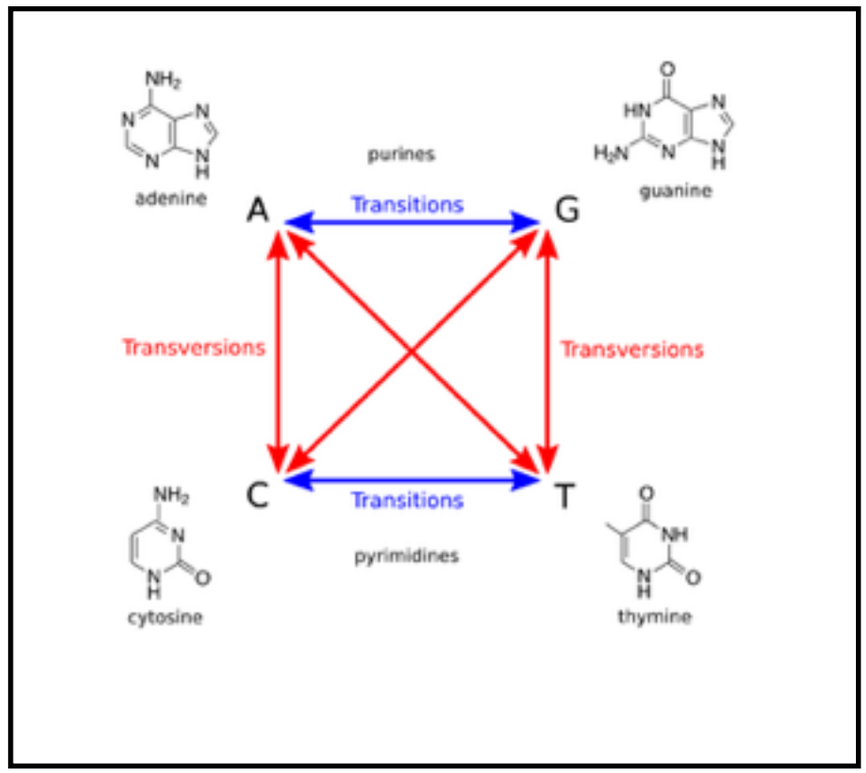
intractability, approximation/ optimization errors

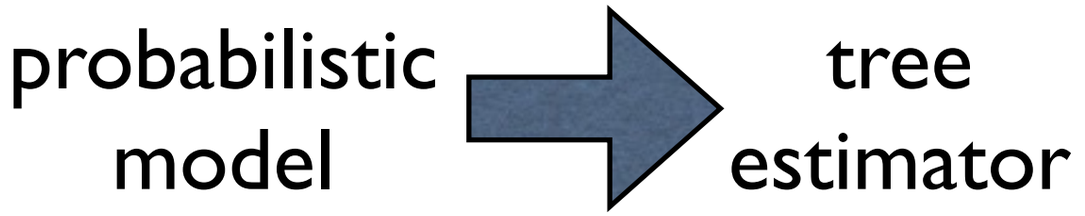
→ computer implementation → estimated tree

bugs!

reality → probabilistic model

- **Common assumption:** we have replicates (loci/sites) of the evolutionary stochastic process (often iid given tree)

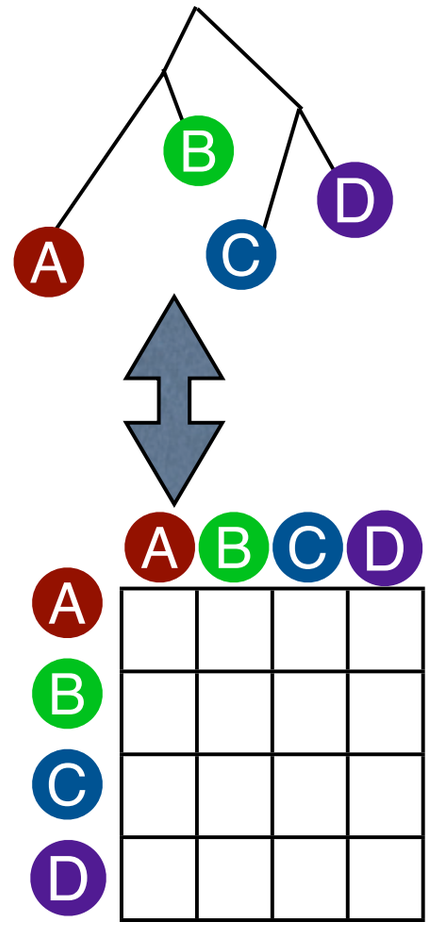




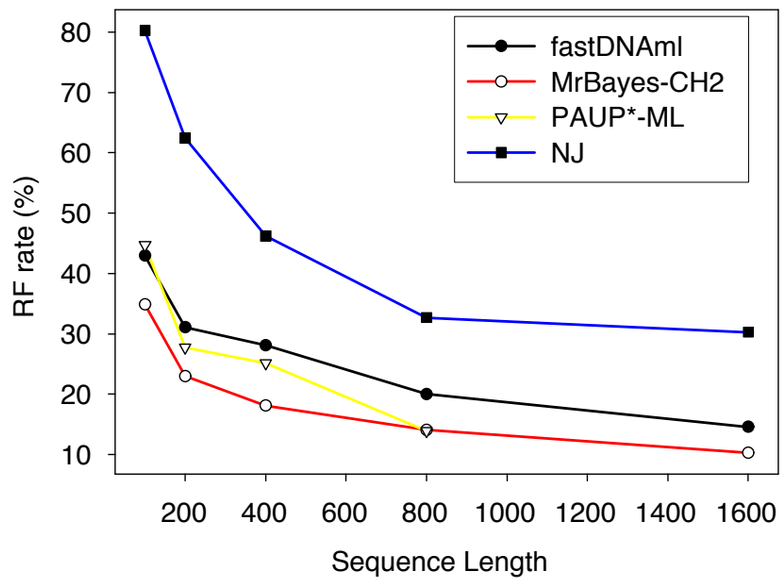
statistical issues

- Identifiability/consistency: can we find the tree given ‘infinite data’
  - What does infinite data mean?
    - # loci/sites
    - # leaves
    - sequencing depth
- Efficiency: how fast will the error decay as we add more data? (cf. computational efficiency)

probabilistic model → tree estimator

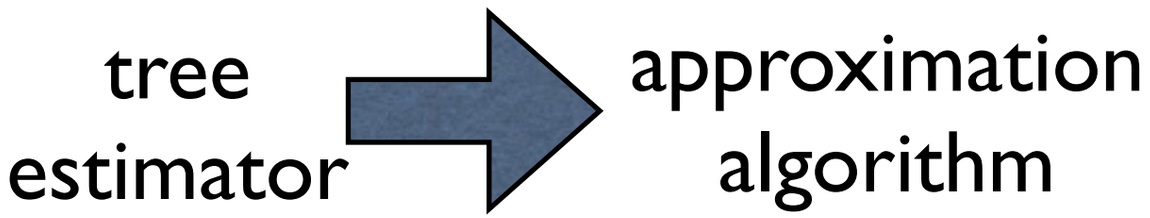


- Earlier methods: parsimony, distance-based
- State-of-the-art: likelihood-based (maximum likelihood and Bayesian method)
- key idea: score many hypothetical trees, marginalizing over unknown ancestral states



Taxa = 40, Height = 1.0

Williams and Moret, 03

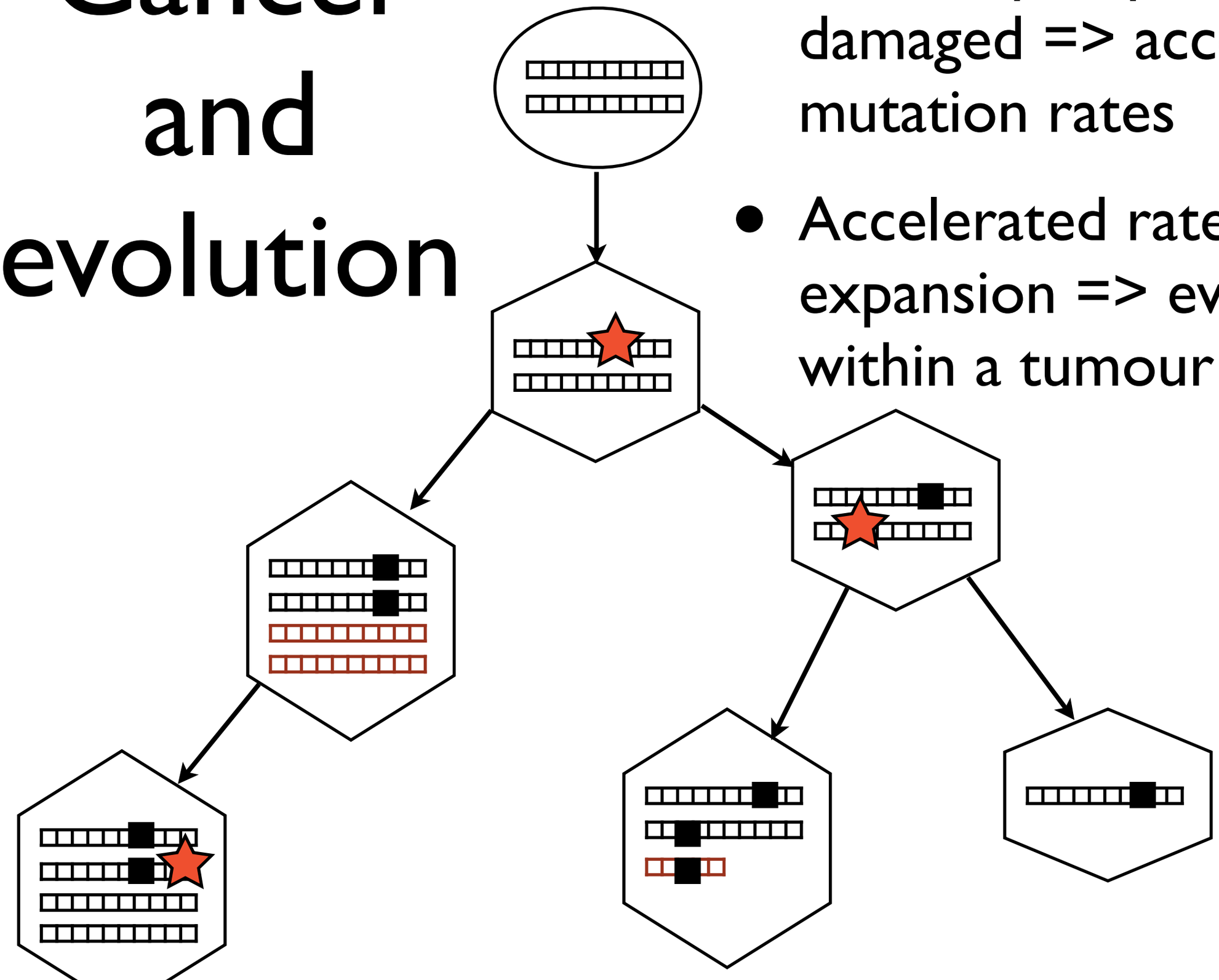


- Statistically efficient tree reconstruction methods are generally computationally inefficient
- Parallelization, distribution
- Approximation methods:
  - MCMC
  - Sequential Monte Carlo
  - Annealing

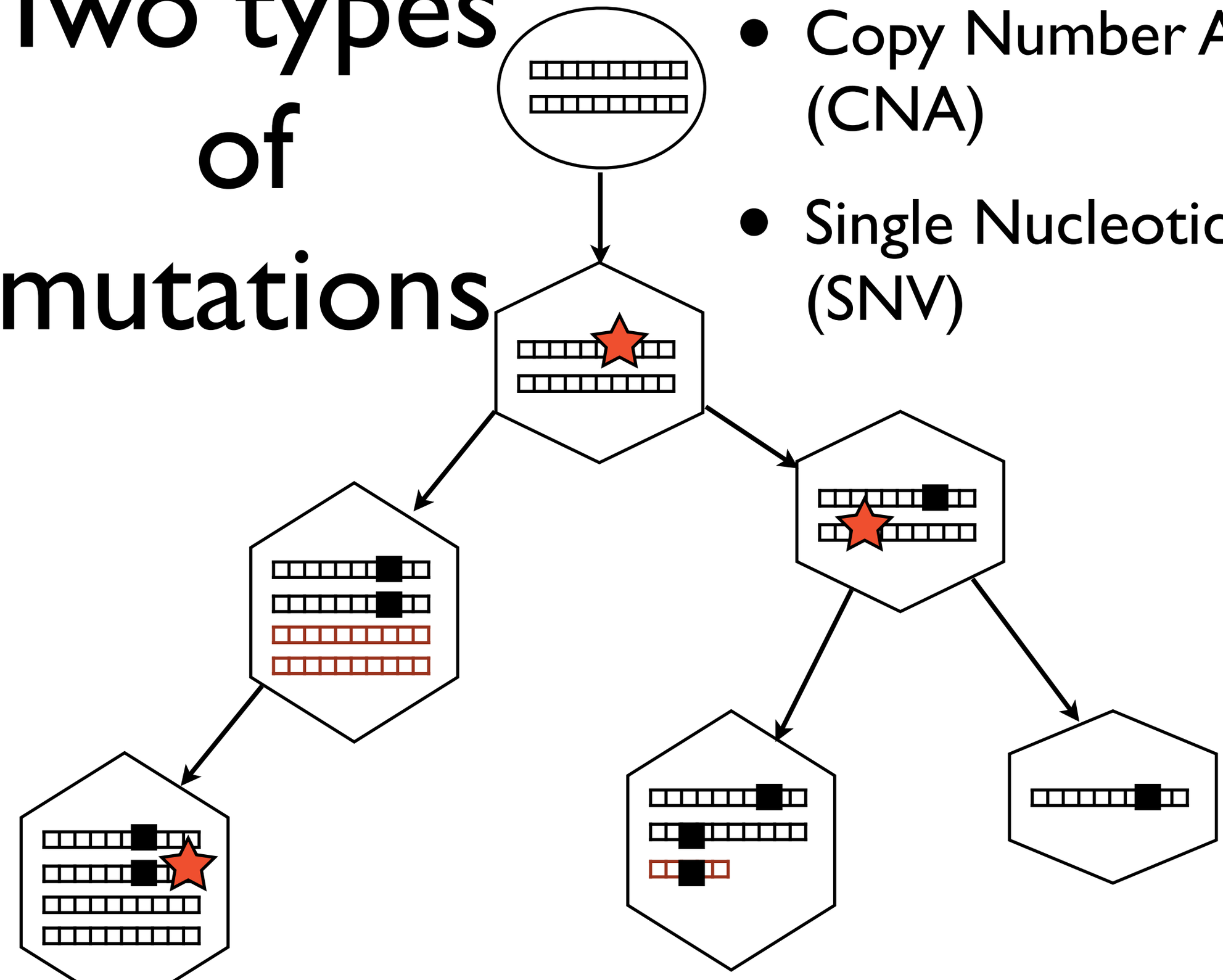
Application: Inferring the Phylogeny of Cancer Cells

Cancer and evolution

- DNA repair pathways damaged => accelerated mutation rates
- Accelerated rate + clonal expansion => evolution within a tumour



Two types of mutations

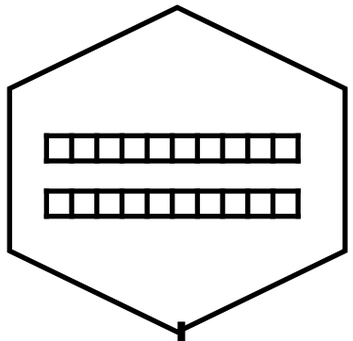


- Copy Number Alterations (CNA)
- Single Nucleotide Variant (SNV)

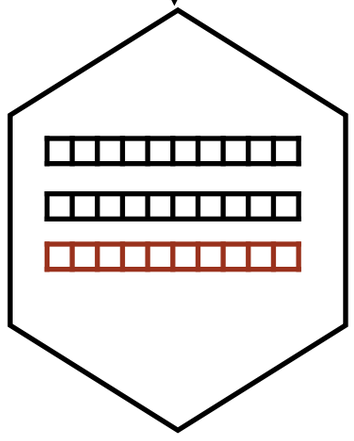
Copy number alterations

Whole chromosome

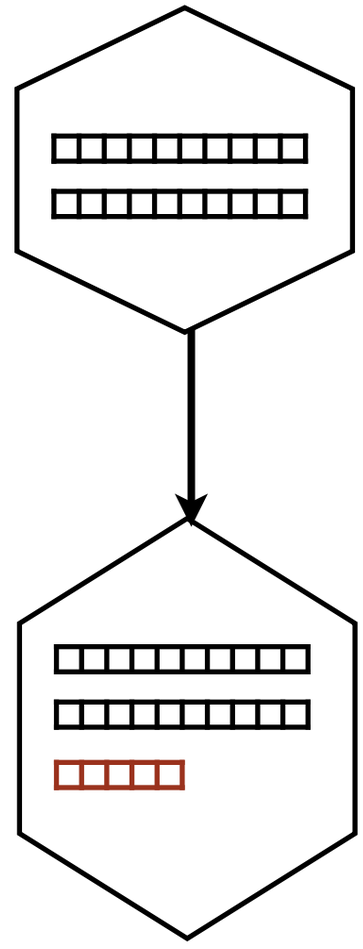
Parent cancer cell



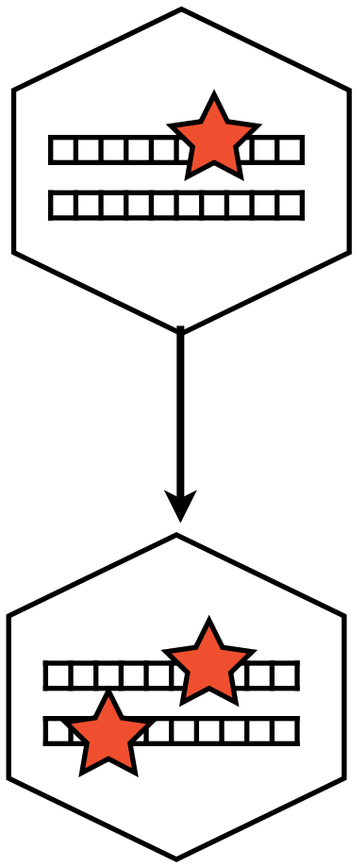
Child cancer cell



Local aberration

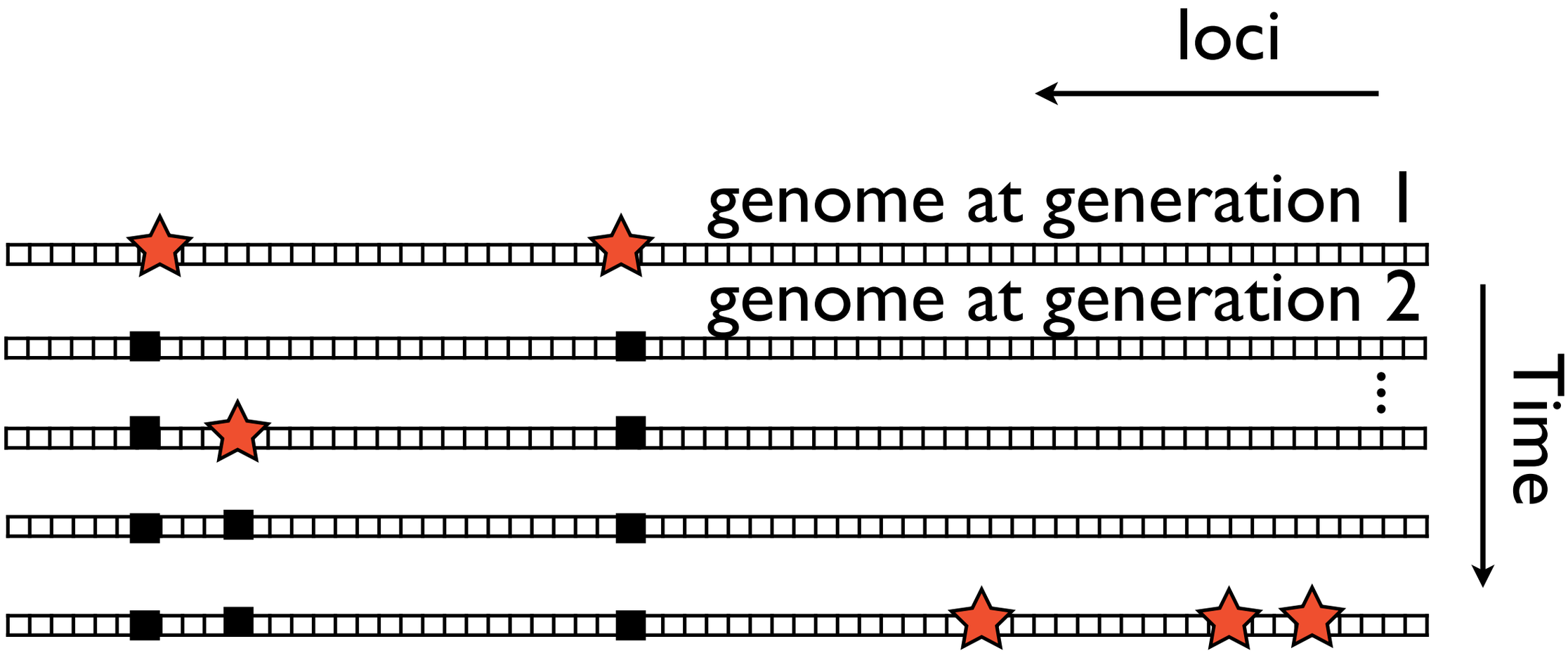


Single nucleotide variants (SNVs)



- Like SNPs, but arise in tumour instead of already in healthy cells (somatic instead of germline)

Accumulation of SNVs



Mathematical simplification: at most one single nucleotide hit per locus (‘**infinite site model**’, $p^2 << p$ when p is small)

=> binary state for SNVs

Motivations for studying this process

- Clinical:
  - Evolution => Adaptation => Relapse
- Scientific:
  - Evolution in the fitness landscape far from a local optimum

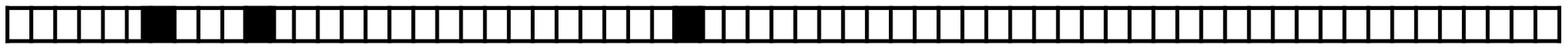
Bayesian non-parametric clonal inference

Trying to identify the 'nodes' of the tree

Simplification (pedagogical)

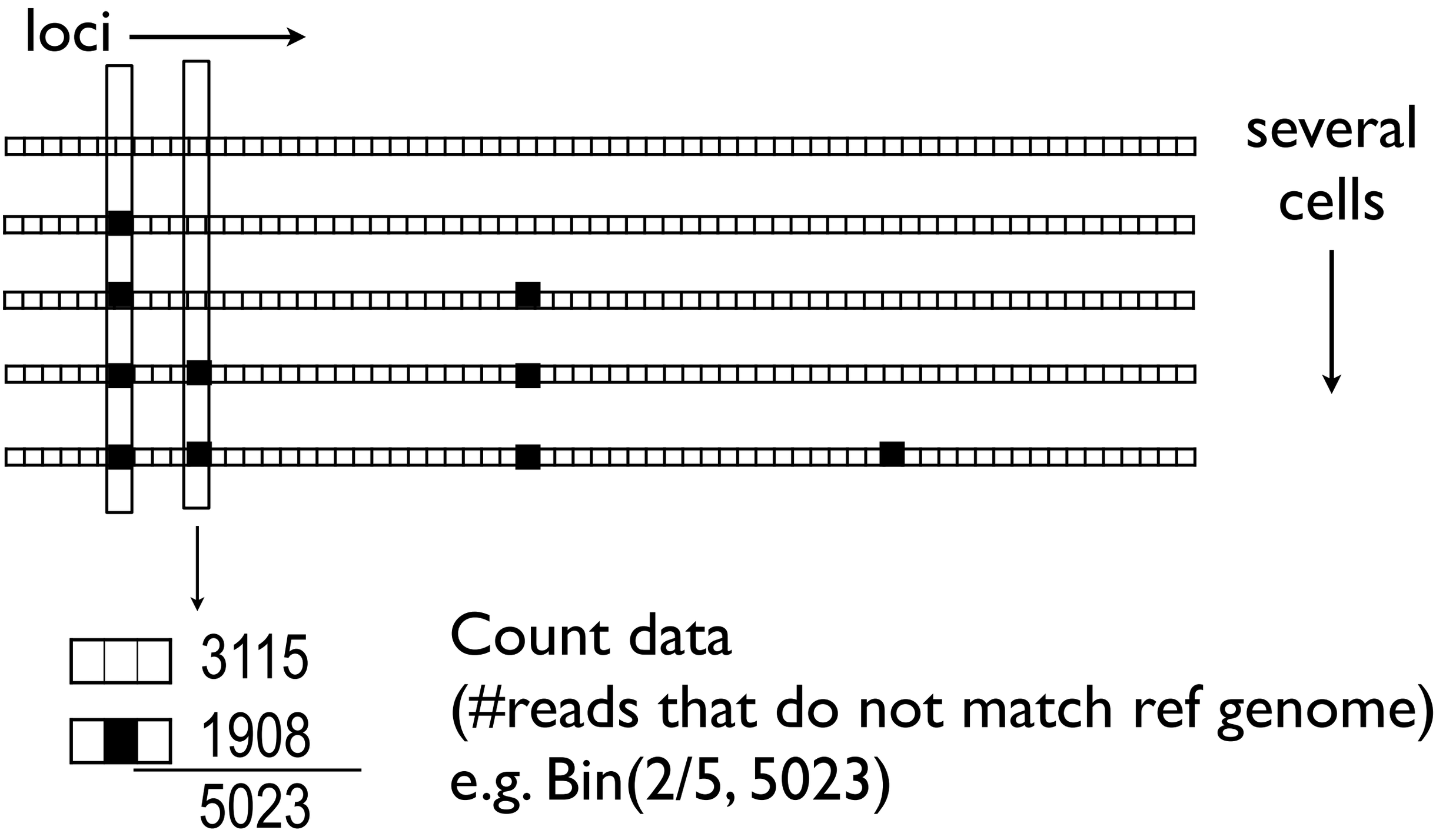
Simplification: for now, 1 chromosome, no CNA

one cell



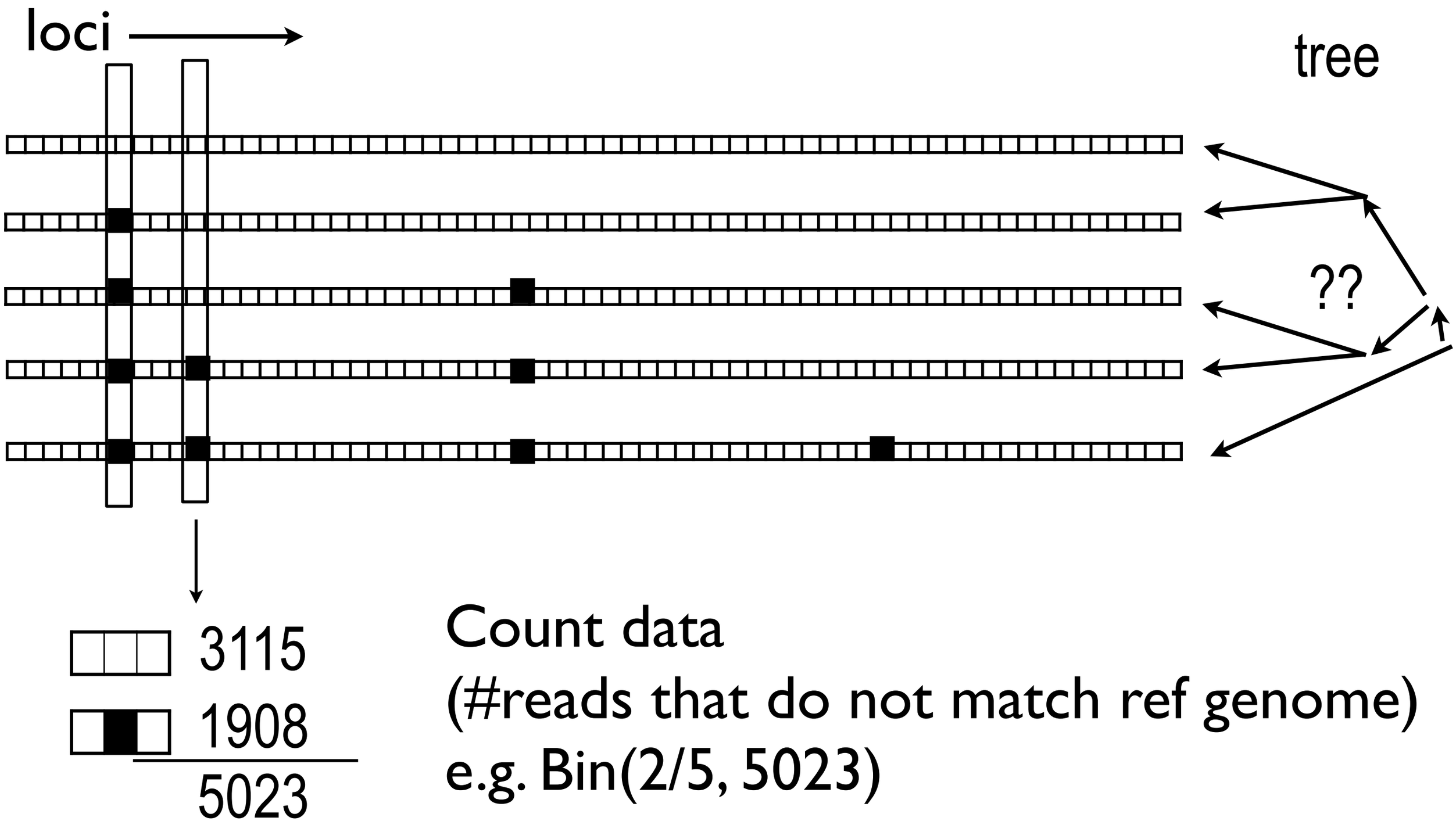
loci ⟶

Observation: ‘bulk data’



Count data
(#reads that do not match ref genome)
e.g. Bin(2/5, 5023)

Observation: ‘bulk data’

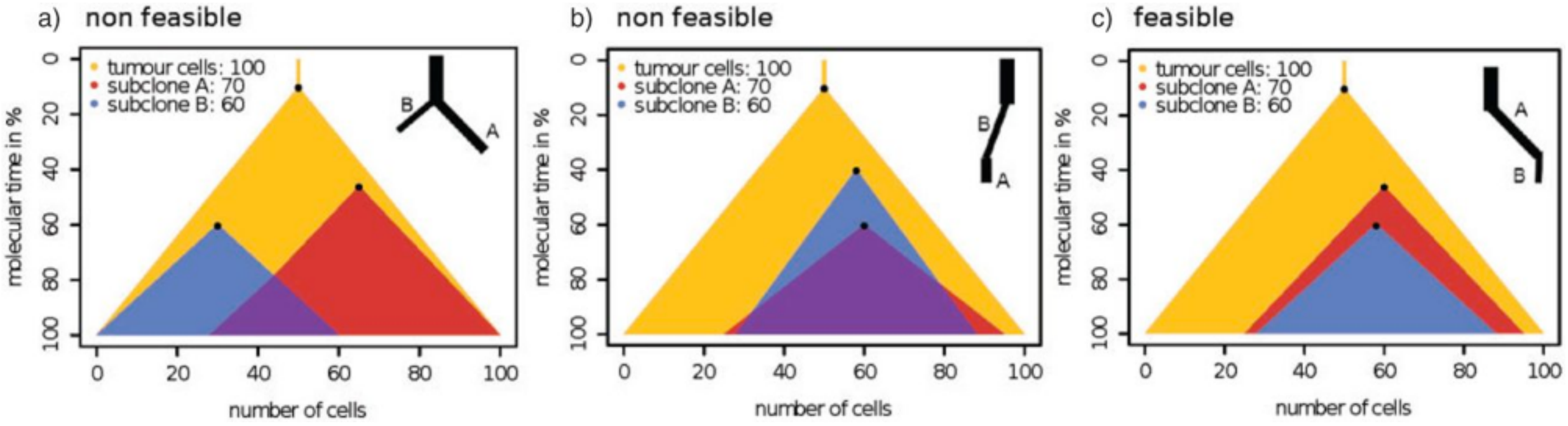


Count data
(#reads that do not match ref genome)
e.g. Bin(2/5, 5023)

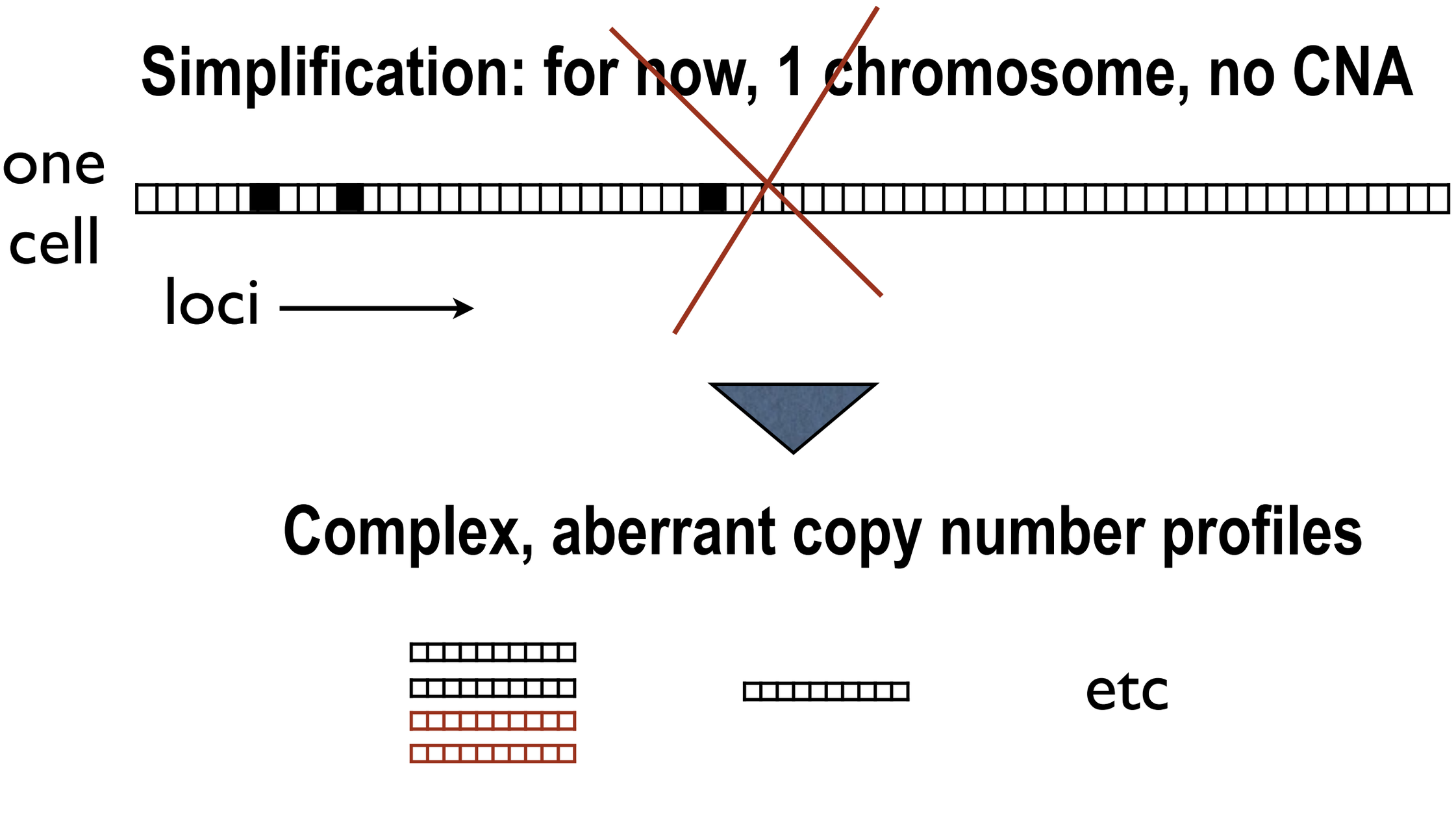
Mutation (partial) order from clonal prevalence

Example: one mutation in 70% of cells, another in 60%

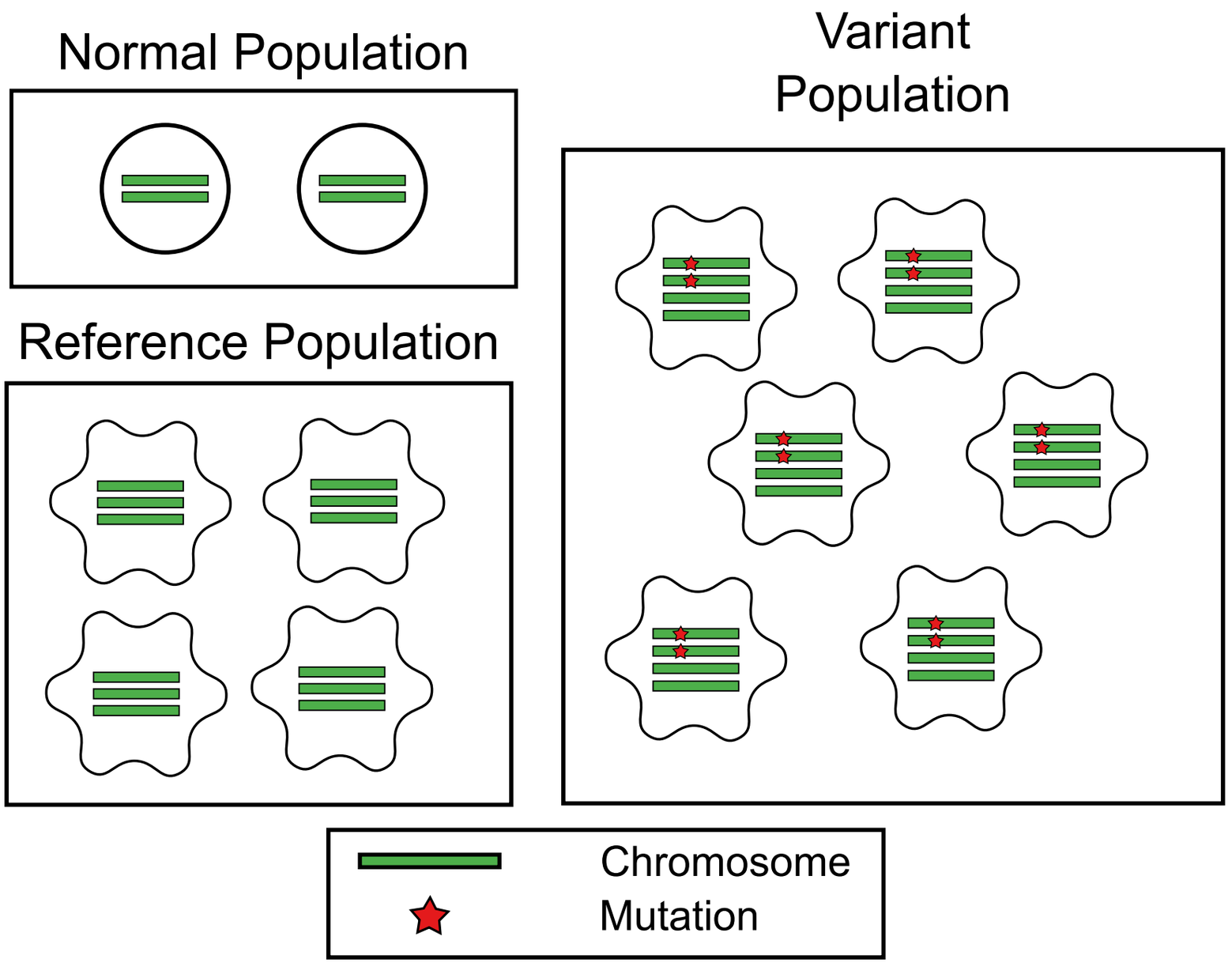
Consequence of infinite site assumption & pigeon hole principle:



Simplification (pedagogical)



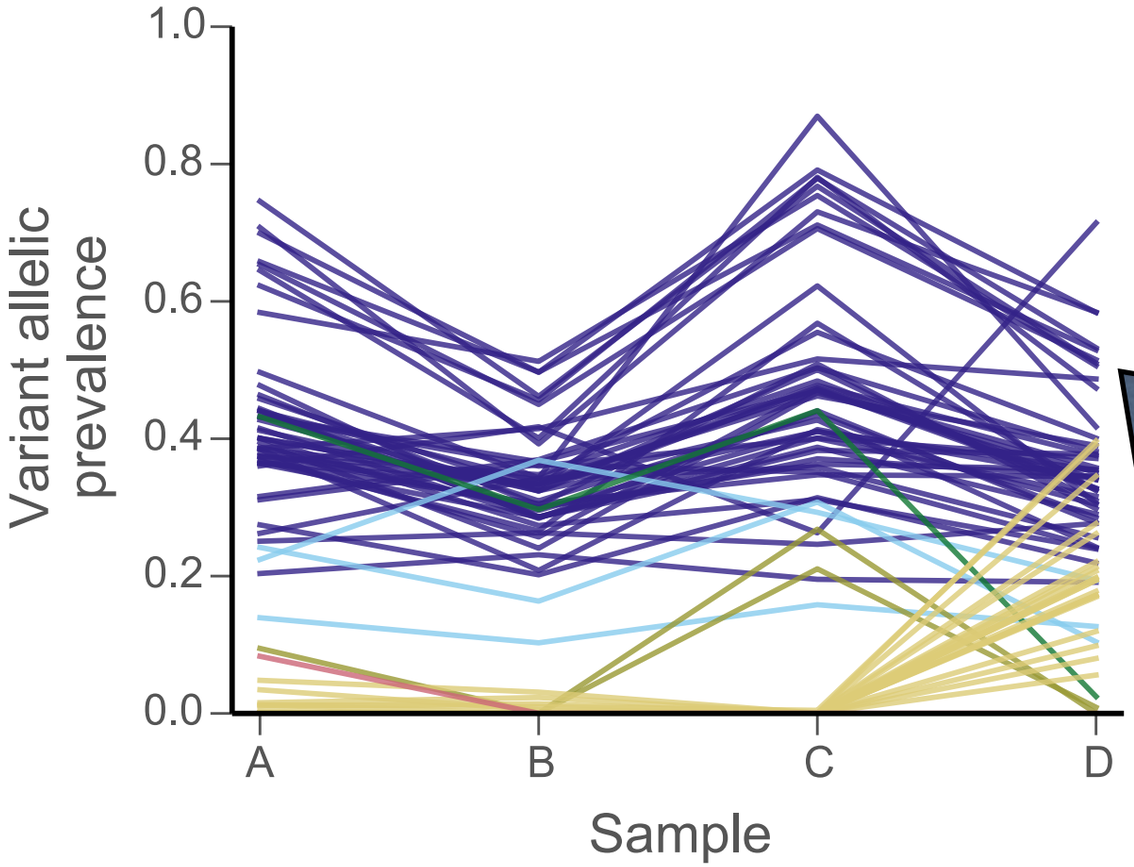
Deconvolution of bulk data



Variant allelic prevalence
= # variant strands / # strands
= 12/40 ('observed')

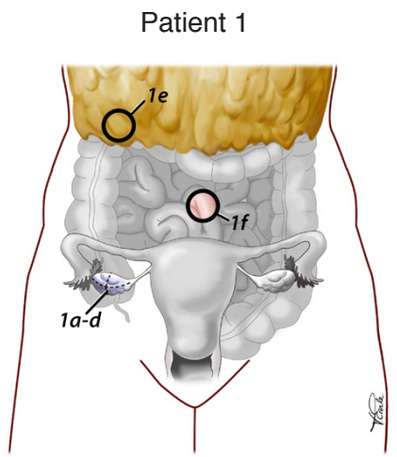
Variant cellular prevalence
= # variant cells / # cells
= 6/12 (inferred)

Example of data

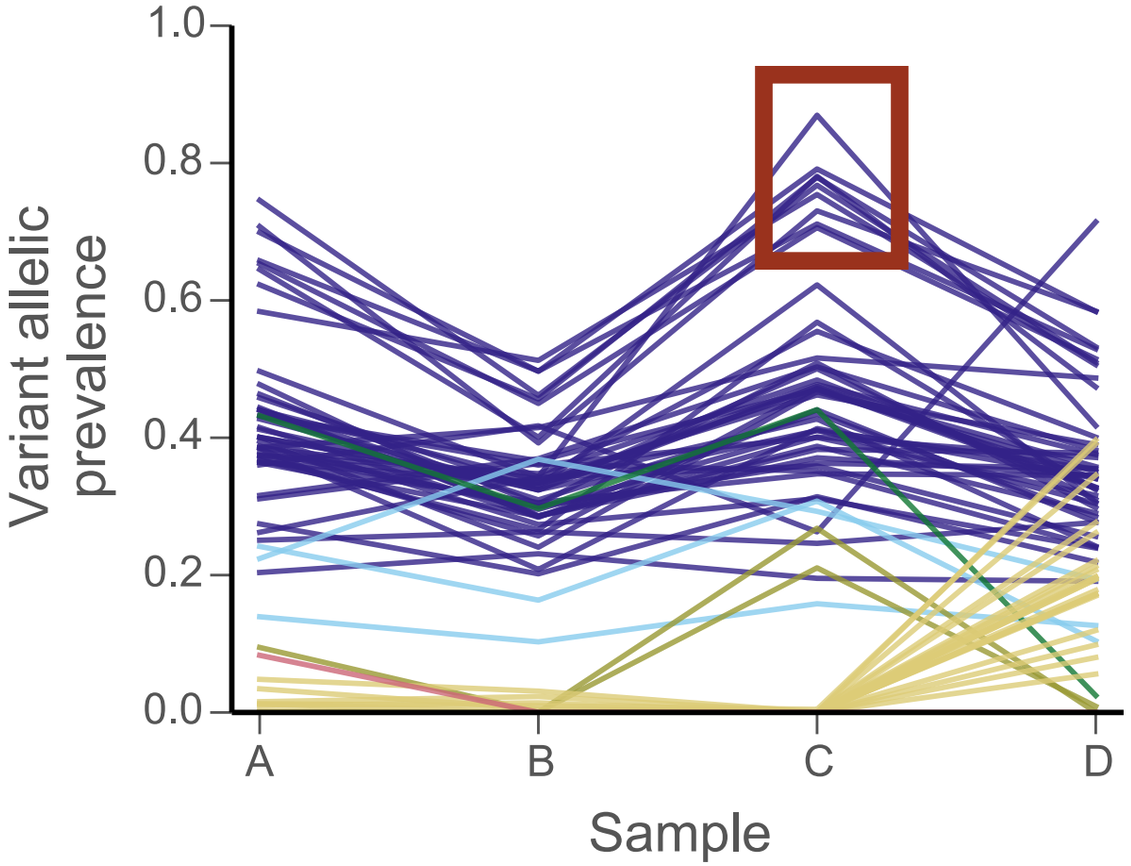


Each curve shows the allele prevalence of a mutation (depth not shown)

4 biopsies (can be temporal or spatial)

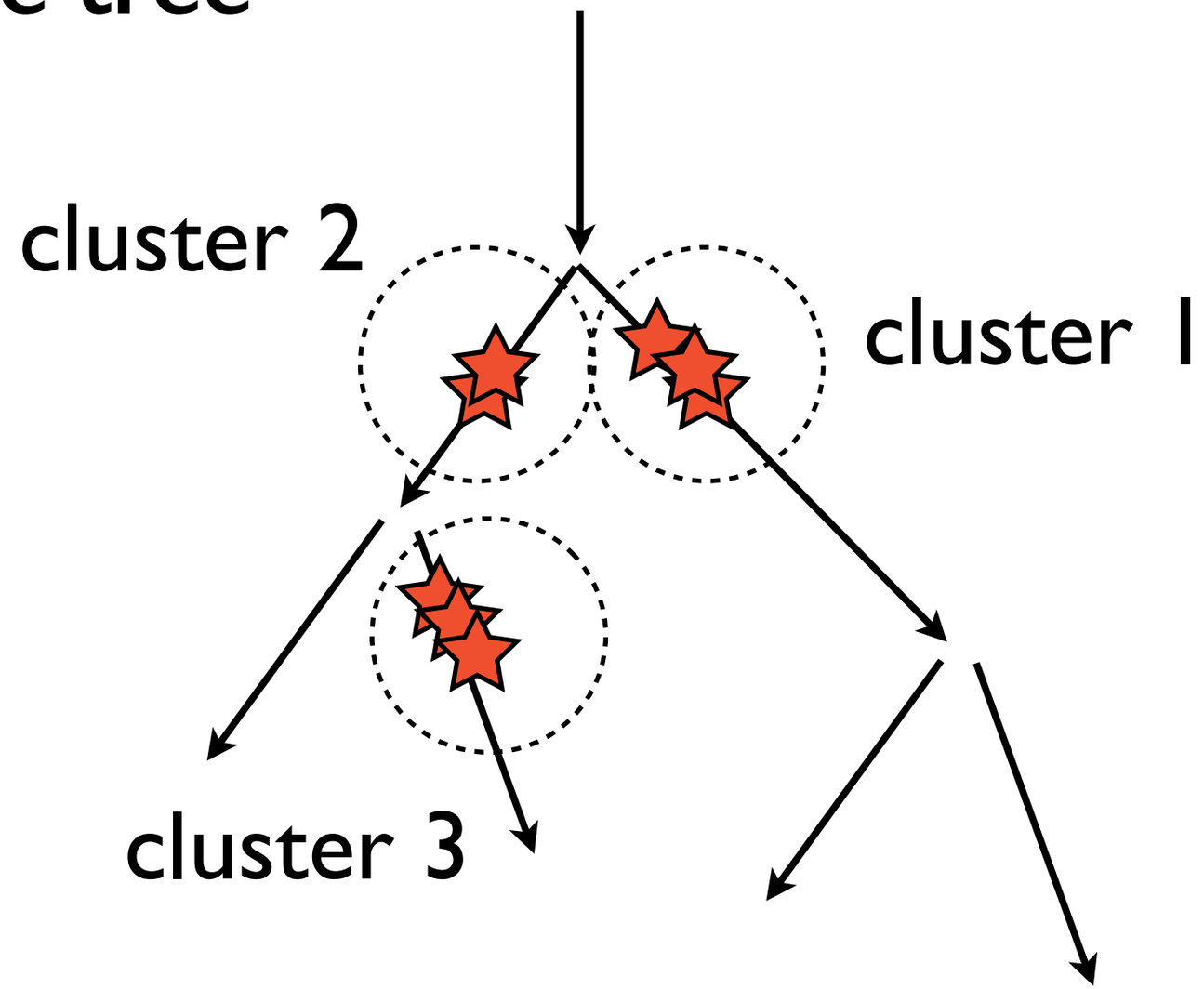


Clustering: motivation



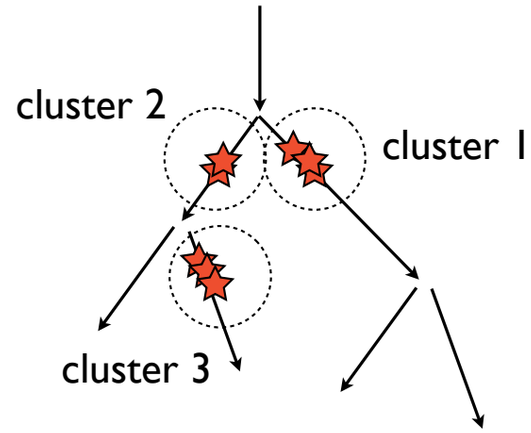
Clustering: motivation

Cluster ≈ edge of the tree

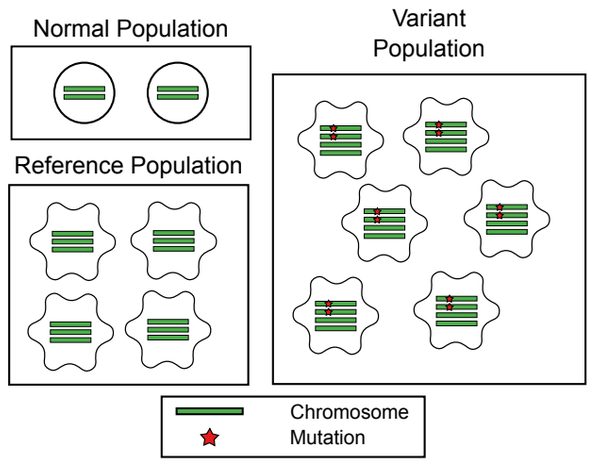


Inferential questions

- Cluster mutations



- Cellular prevalence of each sub-population.



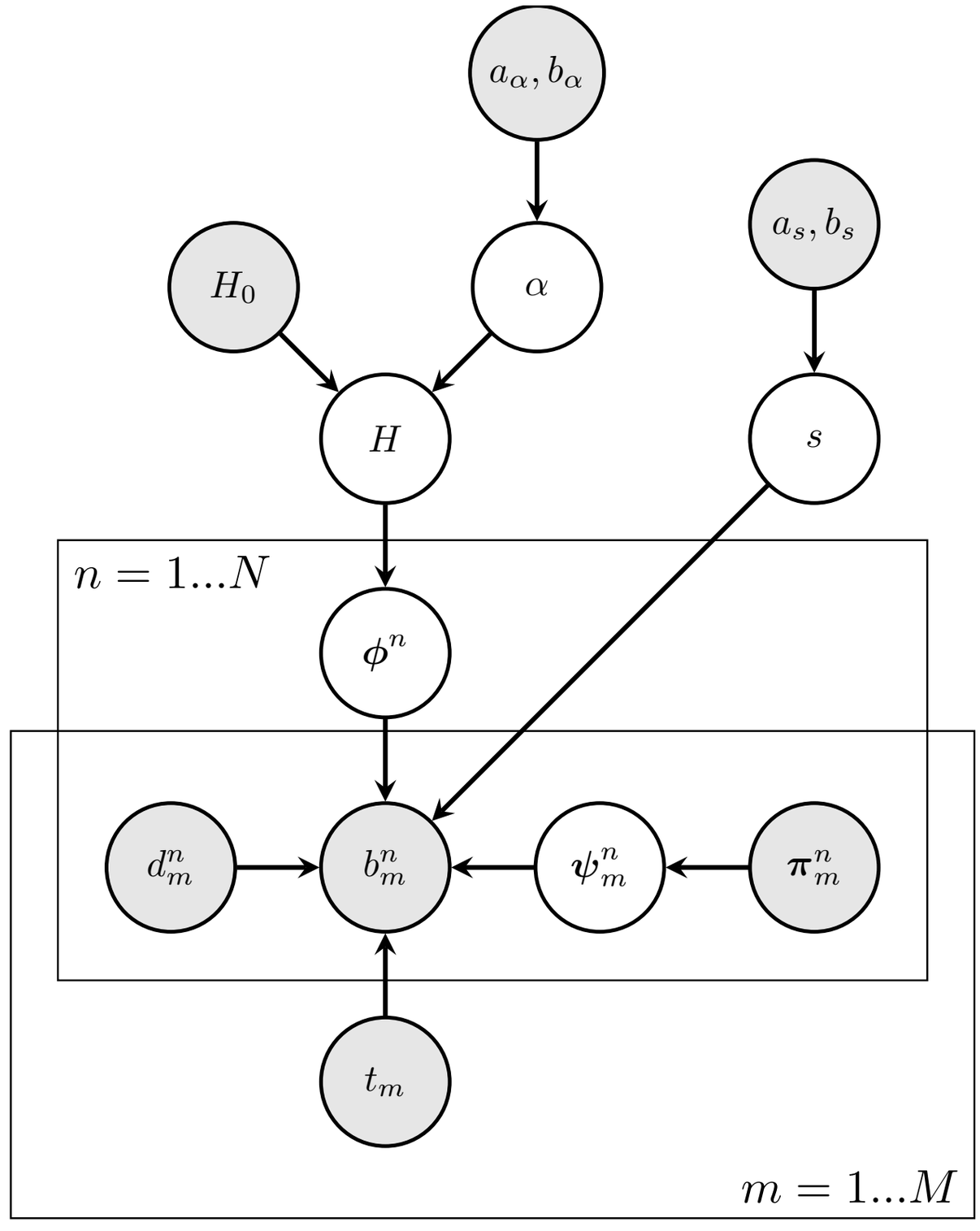
Challenges

- Noisy data
- High level of uncertainty, partial identifiability
- Many data types and sources of prior information

  => Good fit for Bayesian non-parametrics

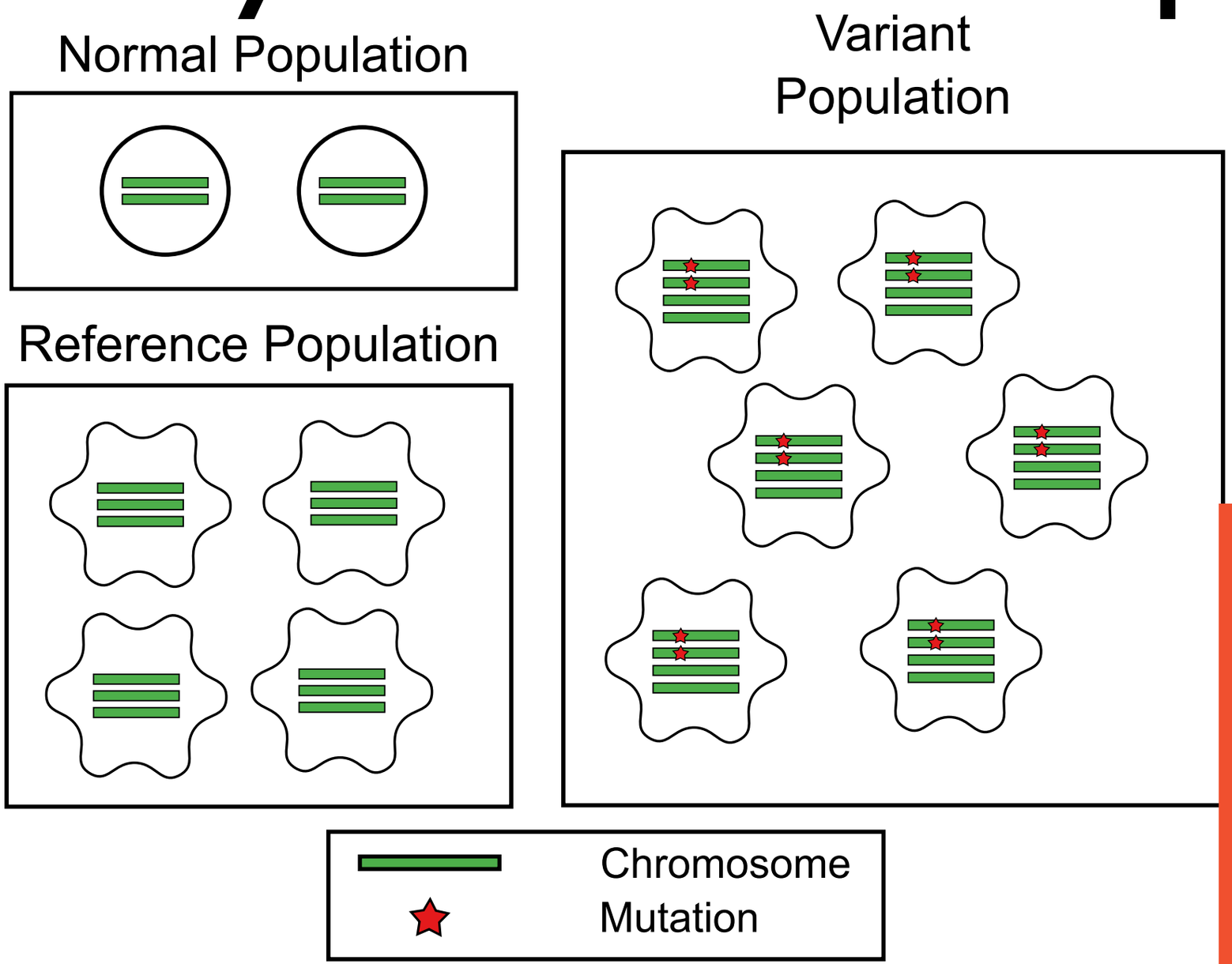
PyClone

- Input: bulk ultra-deep sequencing data
  [+ some prior info]
- Output: posterior distribution over clustering and cellular prevalences
- Based on the Dirichlet Process mixture model



Roth et al. (2014) Nature Methods

PyClone: main parameter



Variant allelic prevalence
= # variant strands / # strands
= 12/40 ('observed')

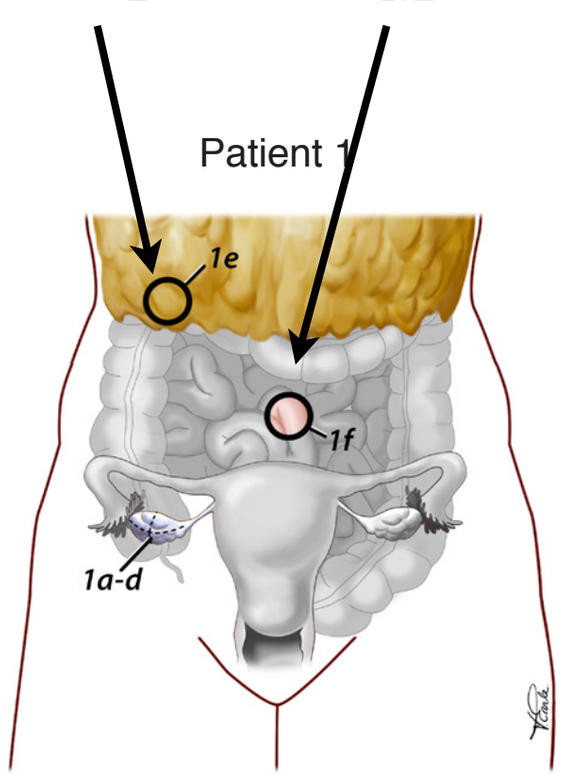
Variant cellular prevalence
= # variant cells / # cells
= 6/12 (inferred)

ϕ^n: cellular prevalence vector of mutation n

PyClone: main parameter

Note: the cellular prevalence varies across anatomical samples

$$\phi^n = (\phi_1^n, \ldots, \phi_M^n)$$

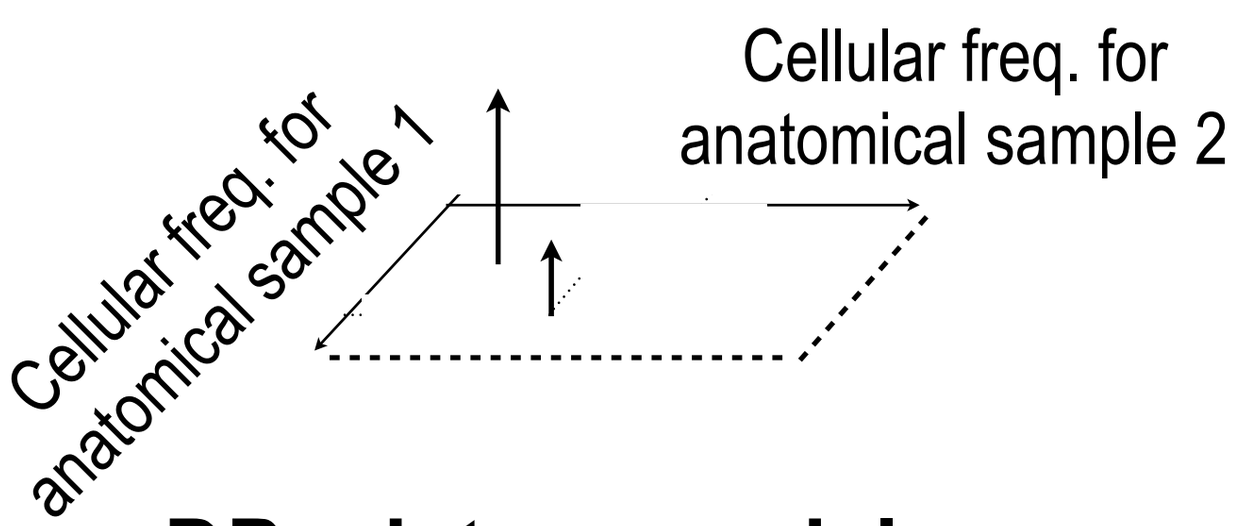


Variant cellular prevalence
= # variant cells / # cells
= 6/12 (inferred)

ϕ^n: cellular prevalence vector of mutation n

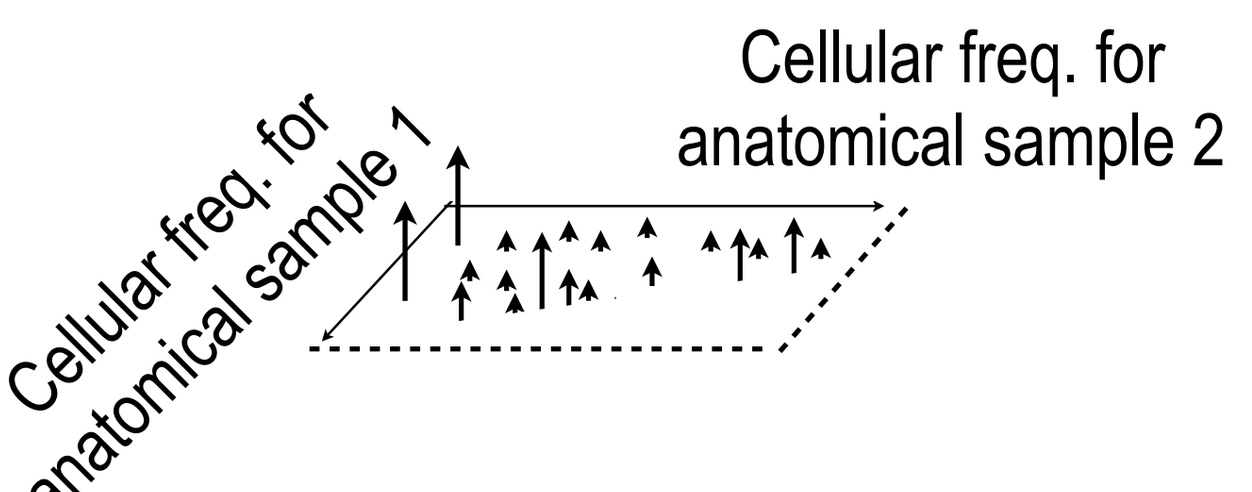
DP mixture model

Parametric mixture model:



Parameter for each datapoint (mutation) obtained by sampling the discrete measure on the right

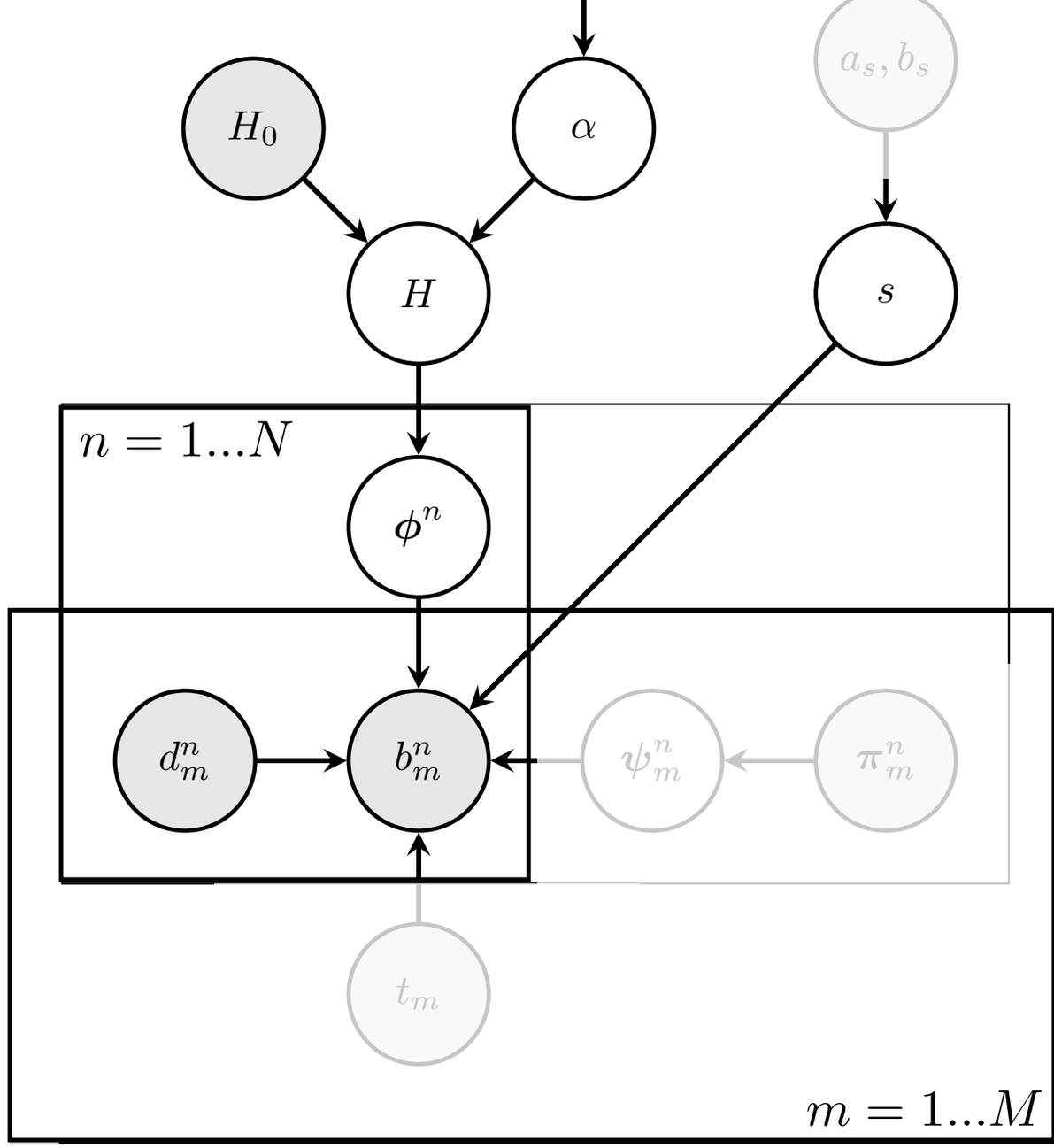
DP mixture model:



same

DP mixture

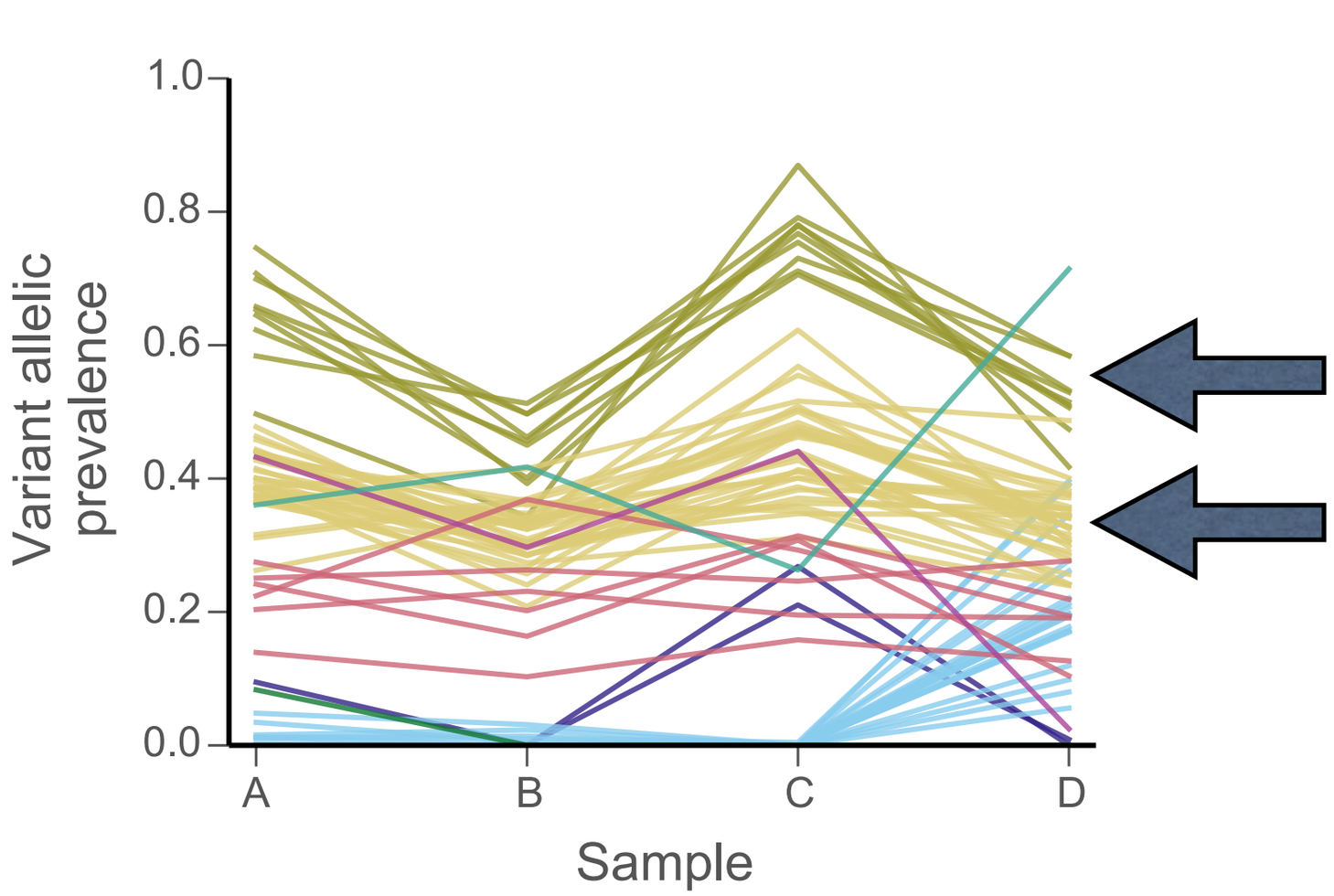
- n: index of the SNV loci
- m: index of sample (biopsy)
- ϕ^n: cellular prevalence vector of mutation n
- d: depth (total # of reads)
- b: number of mutant

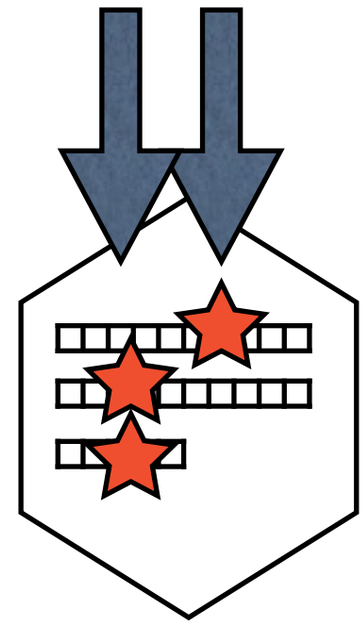


Posterior inference

- MCMC: standard auxiliary variable method for non-conjugate models (‘algorithm 8’)
  - Issue: scaling to large # of mutations
  - Latest work with Andy Roth and Arnaud Doucet: a particle MCMC split merge algorithm

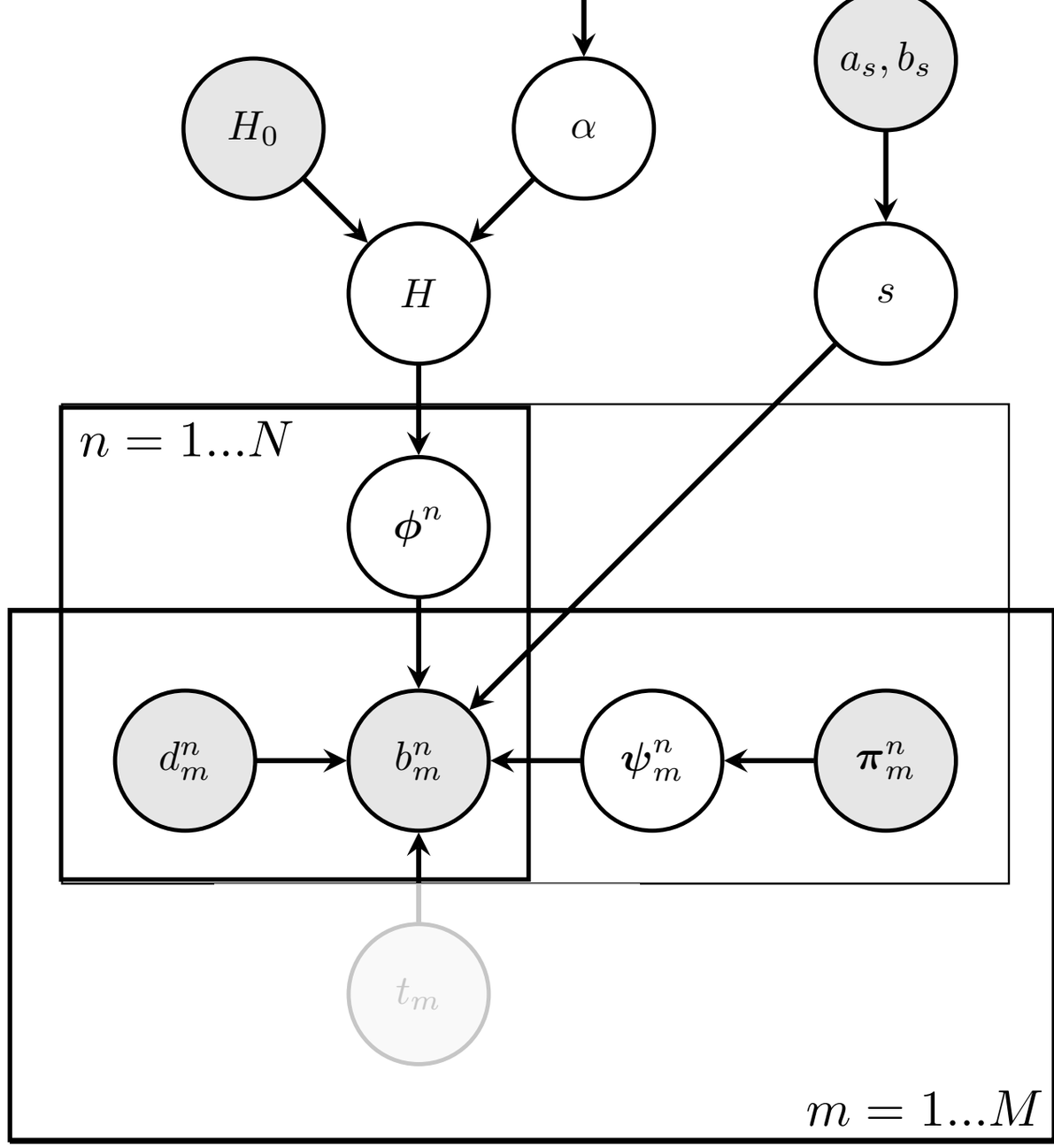
Example of clustering with the model so far

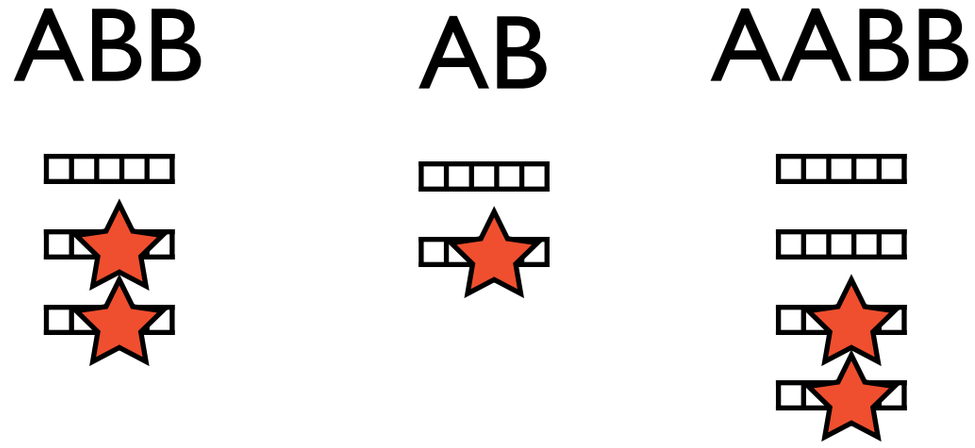




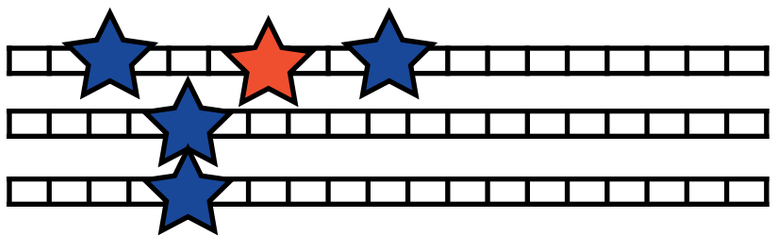
Genotypes

- n: index of the SNV loci
- m: index of sample (biopsy)
- ϕ^n: cellular prevalence vector of mutation n
- d: depth (total # of reads)
- b: number of mutant alleles
- s: overdispersion parameter
- ψ: genotype ($\in$ {AB, AAB, AABB, ...})





Informed priors over genotypes



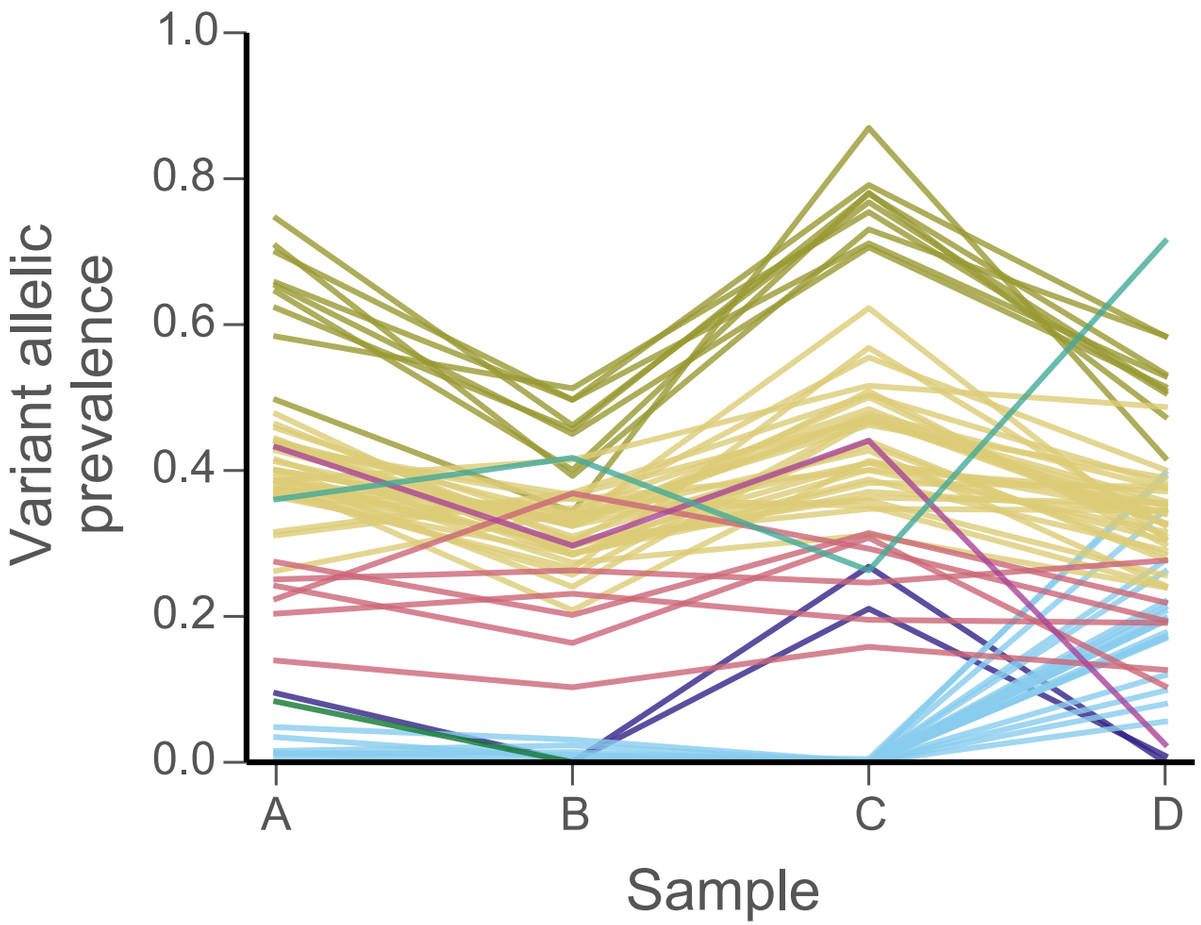
SNV (cancer)

SNP (from your parents)

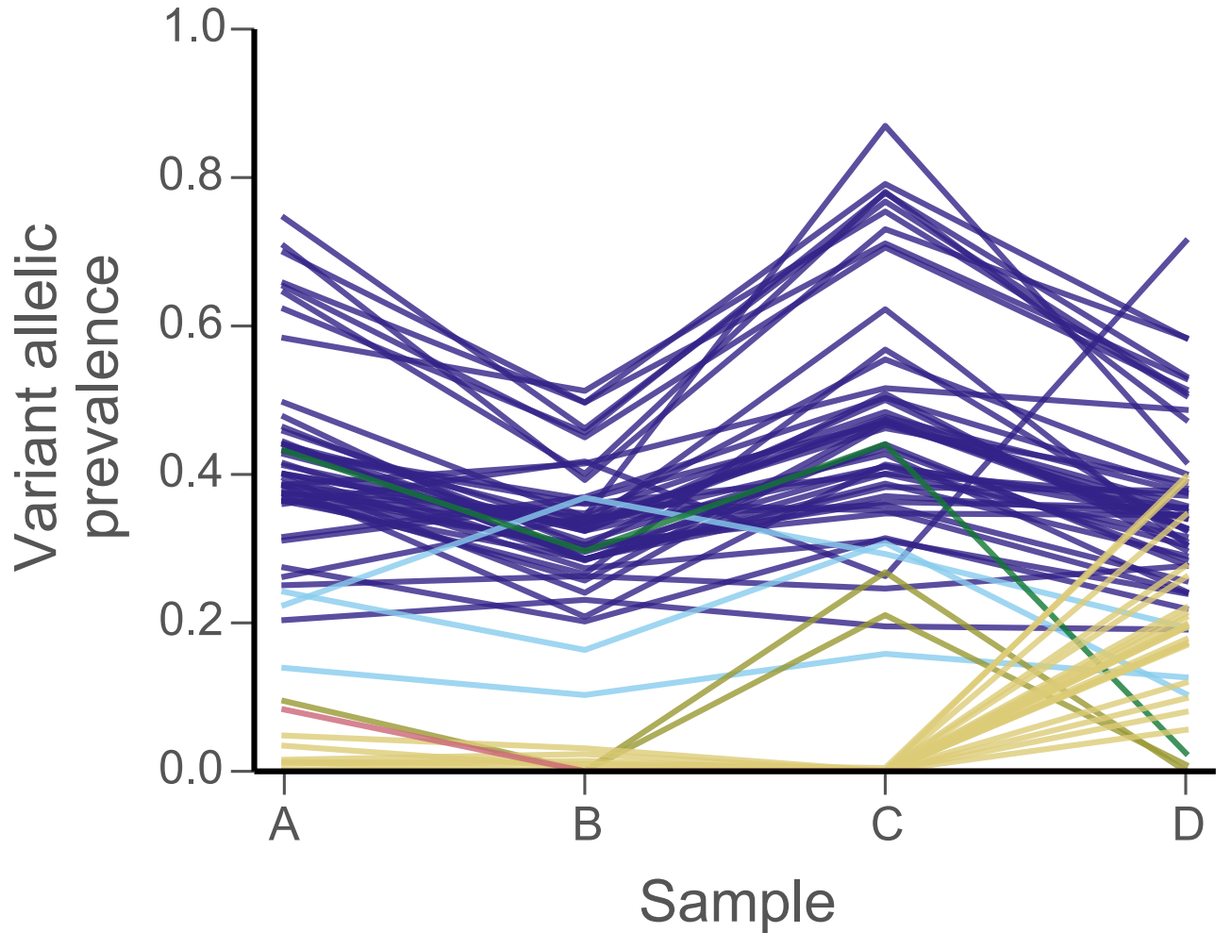
- Goal: reduce the space of possible genotypes
- Key idea: we can also use germline SNPs to get extra information on the copy number at a loci
- Hijack standard micro-array platforms to get log intensity in the neighborhood of the SNV of interest

Clustering of mutations

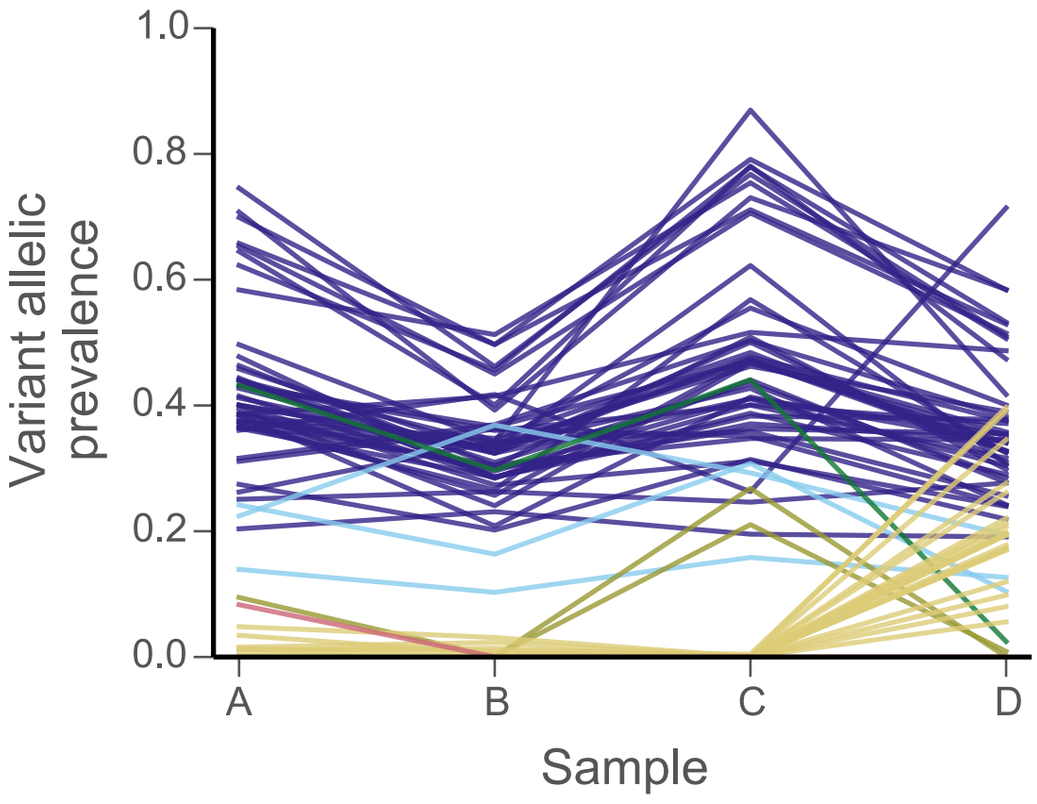
Without genotypes

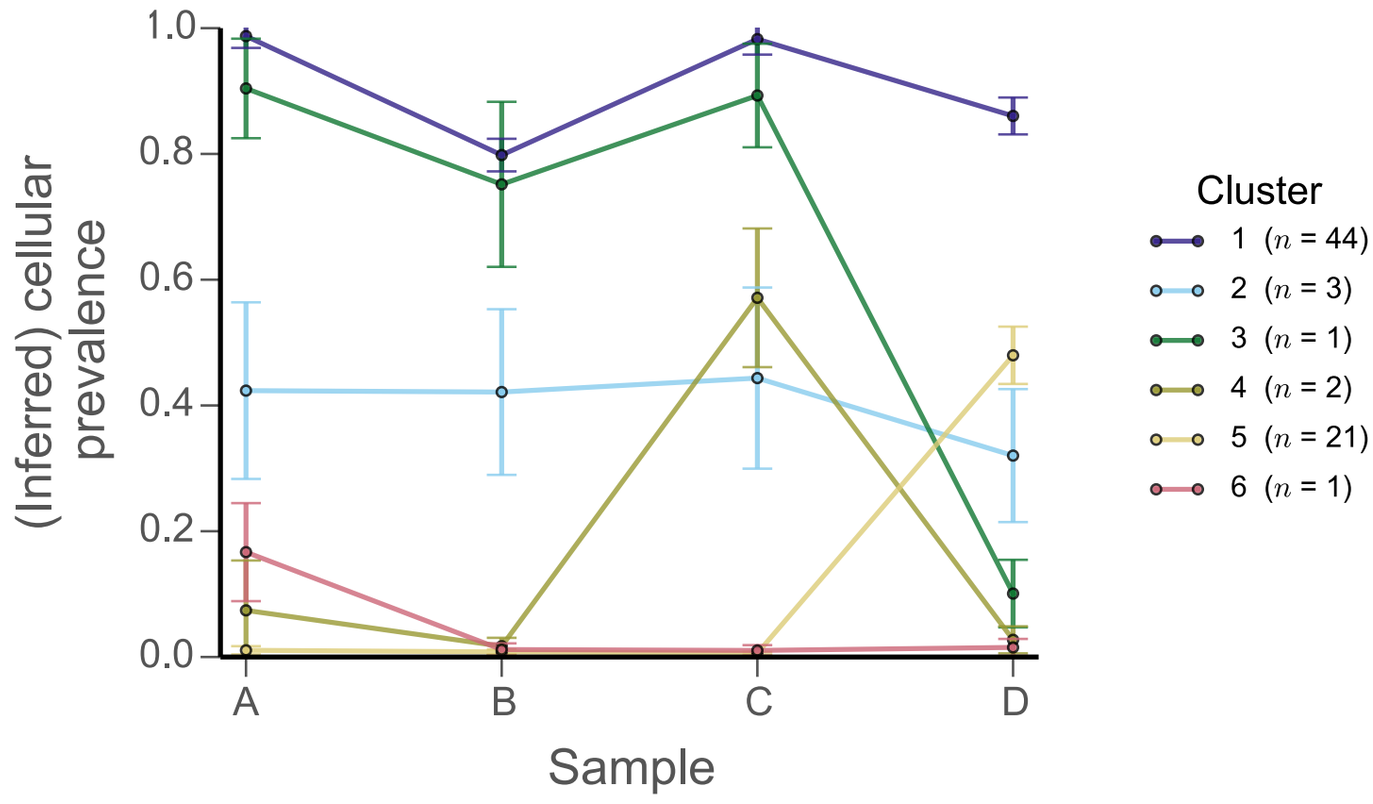


With genotypes



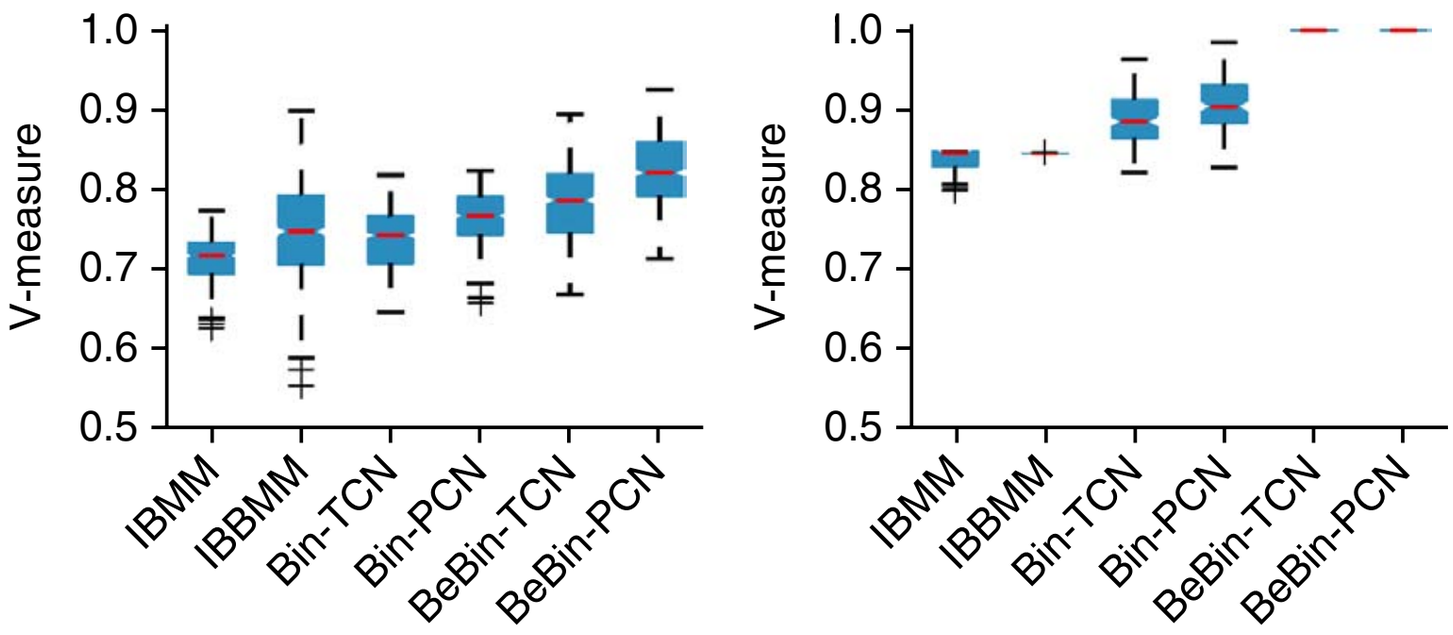
Full output (consensus)





How to validate?

- Cross validation
- Synthetic data (in silico and in vitro)



Single-cell sequencing

loci

BeBin-PCN
IBBMM

Absent
Present
Unknown

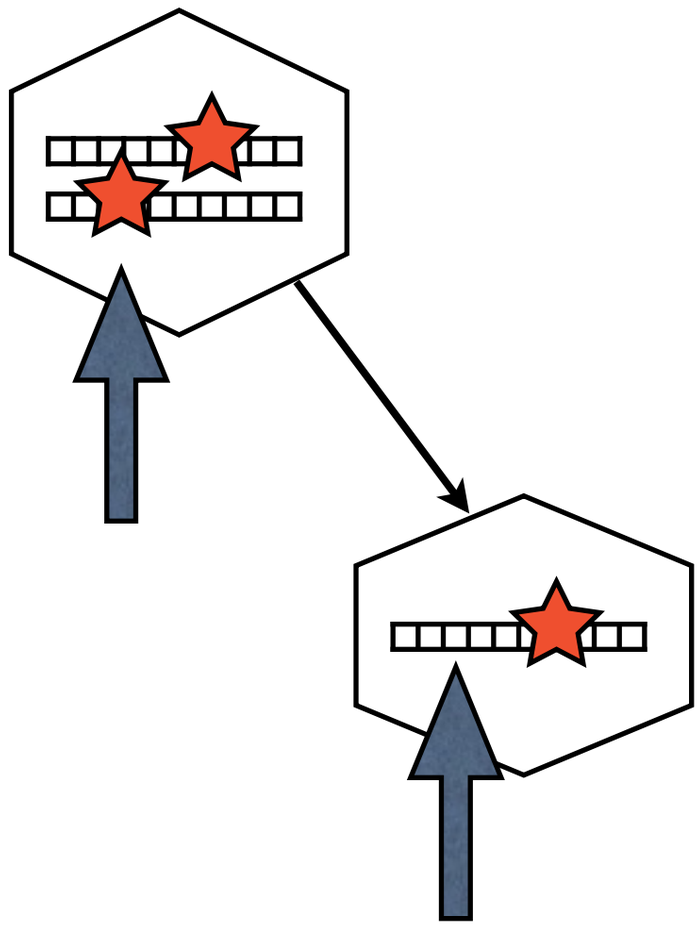
Cell type

Control
Cancer
Normal

gDNA 1
Nucleus 41
Nucleus 4
Nucleus 39
Nucleus 36
Nucleus 35
Nucleus 34
Nucleus 33
Nucleus 32
Nucleus 31
Nucleus 9
Nucleus 3
Nucleus 5
Nucleus 29
Nucleus 22
Nucleus 20
Nucleus 2
Nucleus 18
Nucleus 17
Nucleus 16
Nucleus 15
Nucleus 14
Nucleus 13
Nucleus 10
Nucleus 27
Nucleus 7
Nucleus 8
Nucleus 1
Nucleus 19
Nucleus 6
Nucleus 30

cell

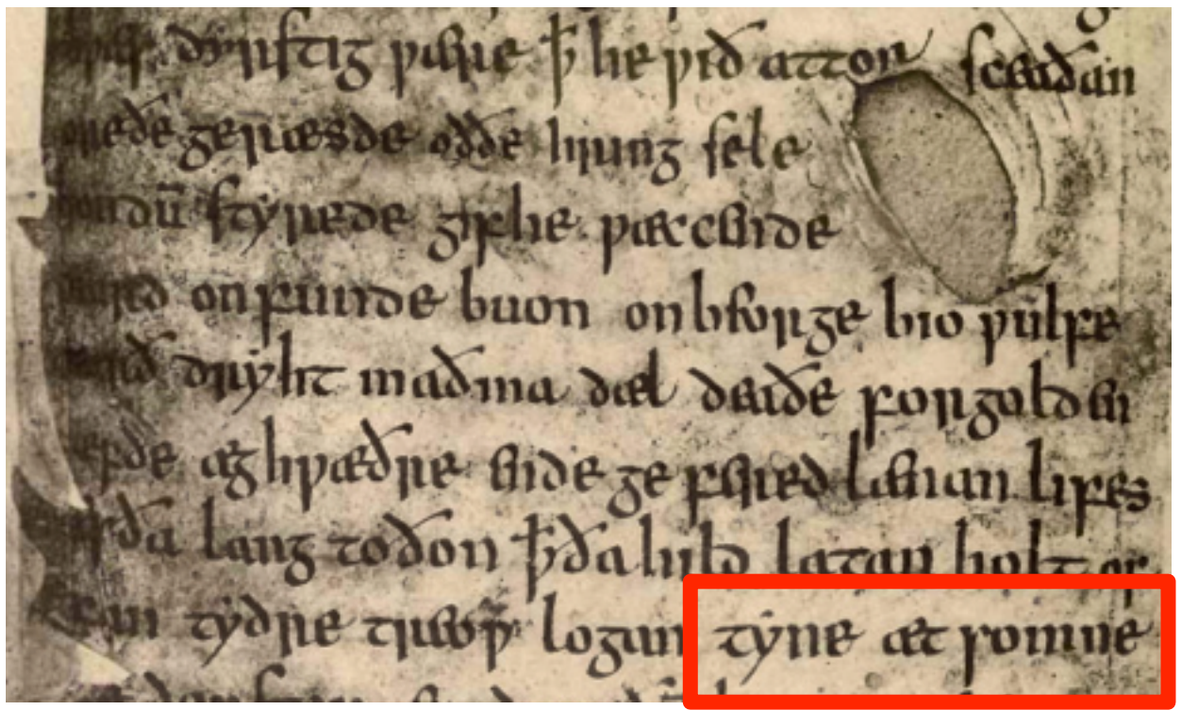
Issue: loss of SNVs



Application: reconstructing ancient languages

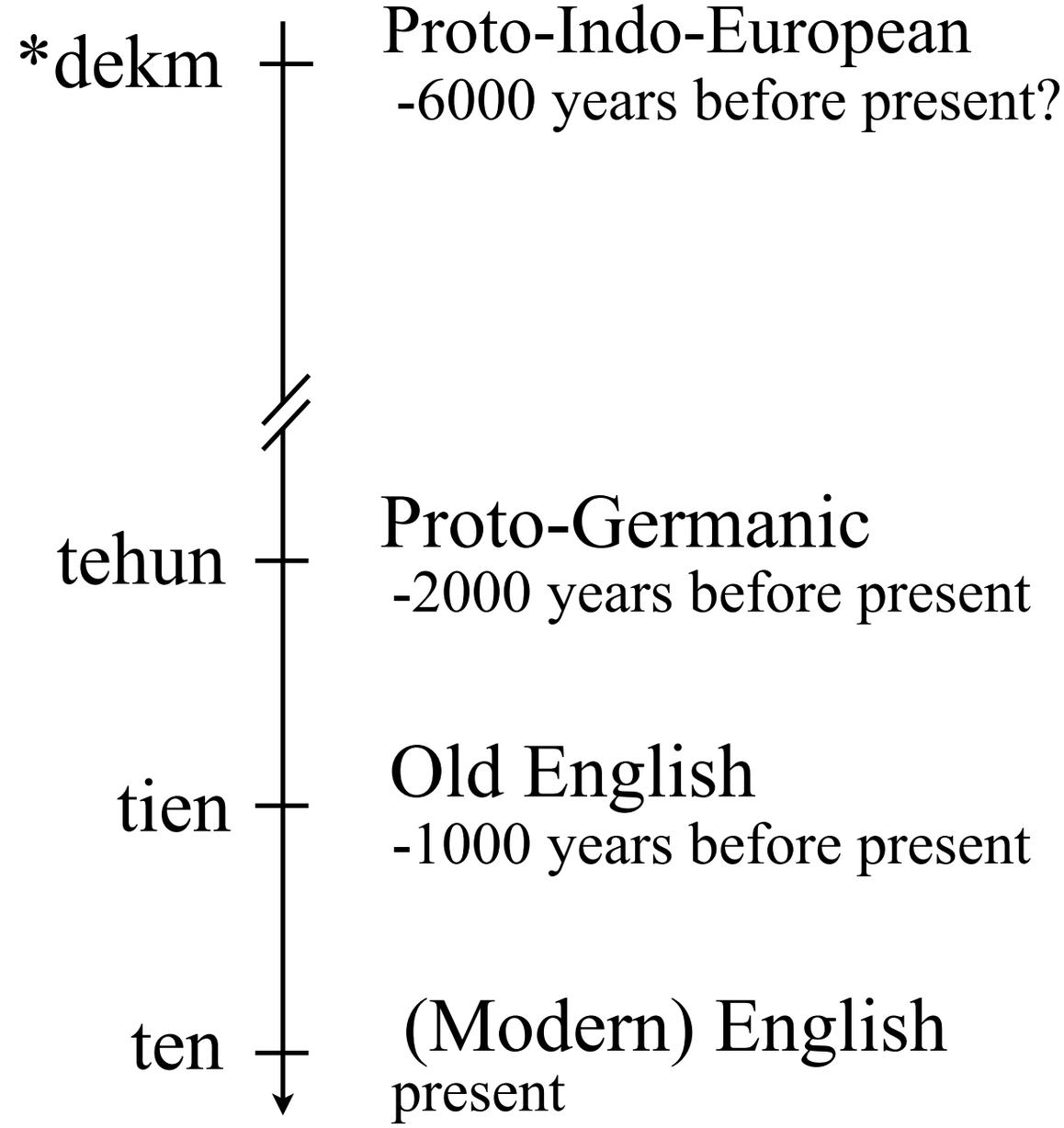
Examples of language change

Beowulf (~ 8th - 11th century)

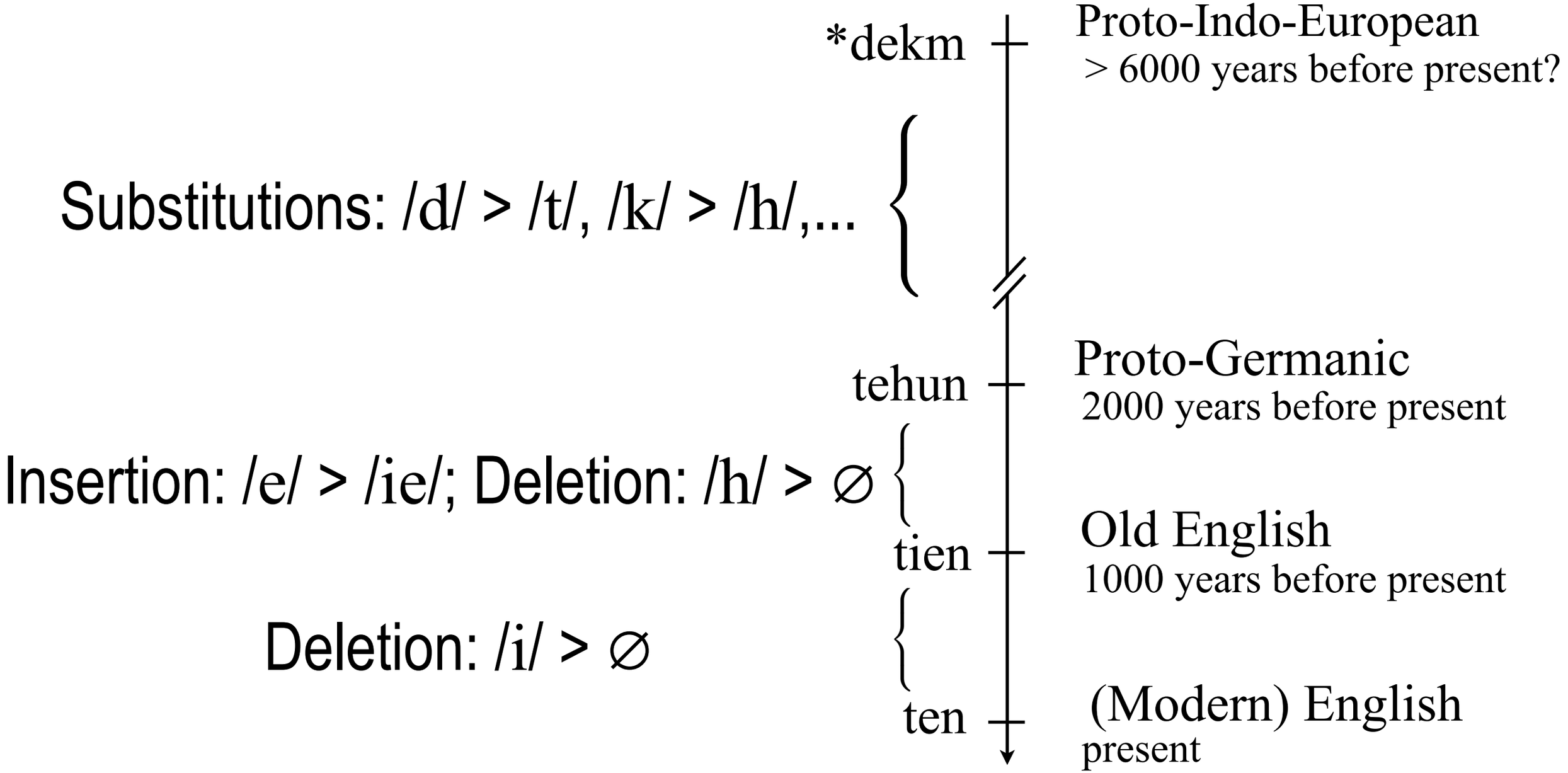


tȳne ætsomne

‘ten altogether’



Examples of language change



Example of data

Austronesian dataset:

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Proto-Oceanic	*ika	*mataku

…

⋮

Size: 706 languages, 150k word forms in IPA

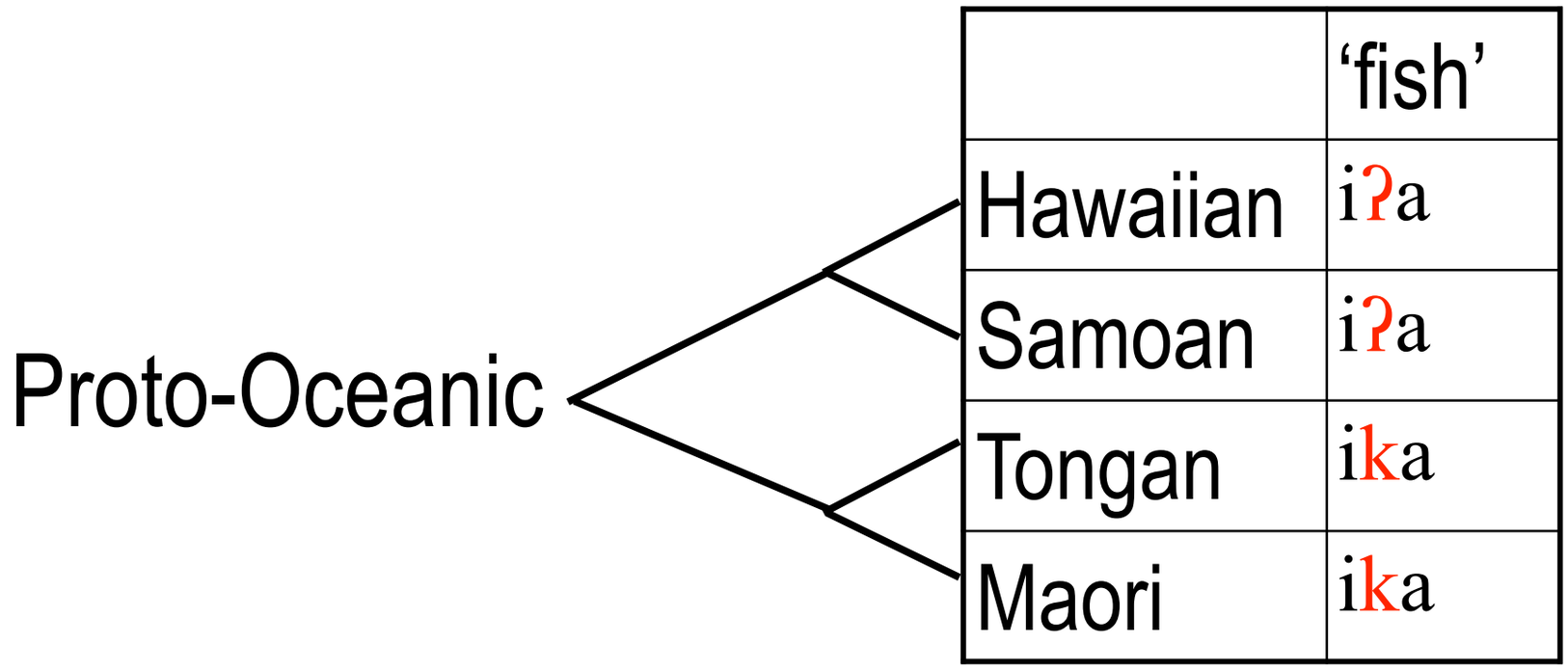
[Greenhill et al, ’08]

Oceanic language

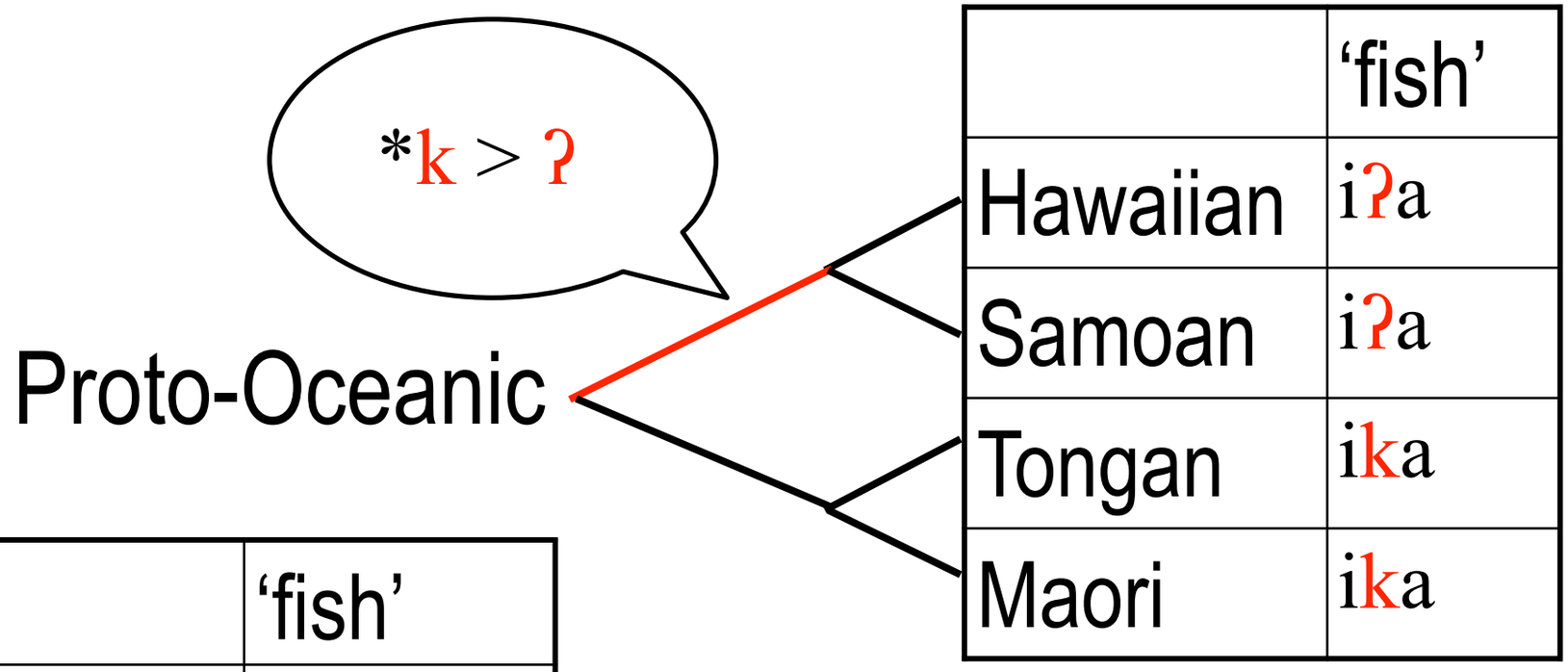
Task: comparative reconstruction

	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

Task: comparative reconstruction



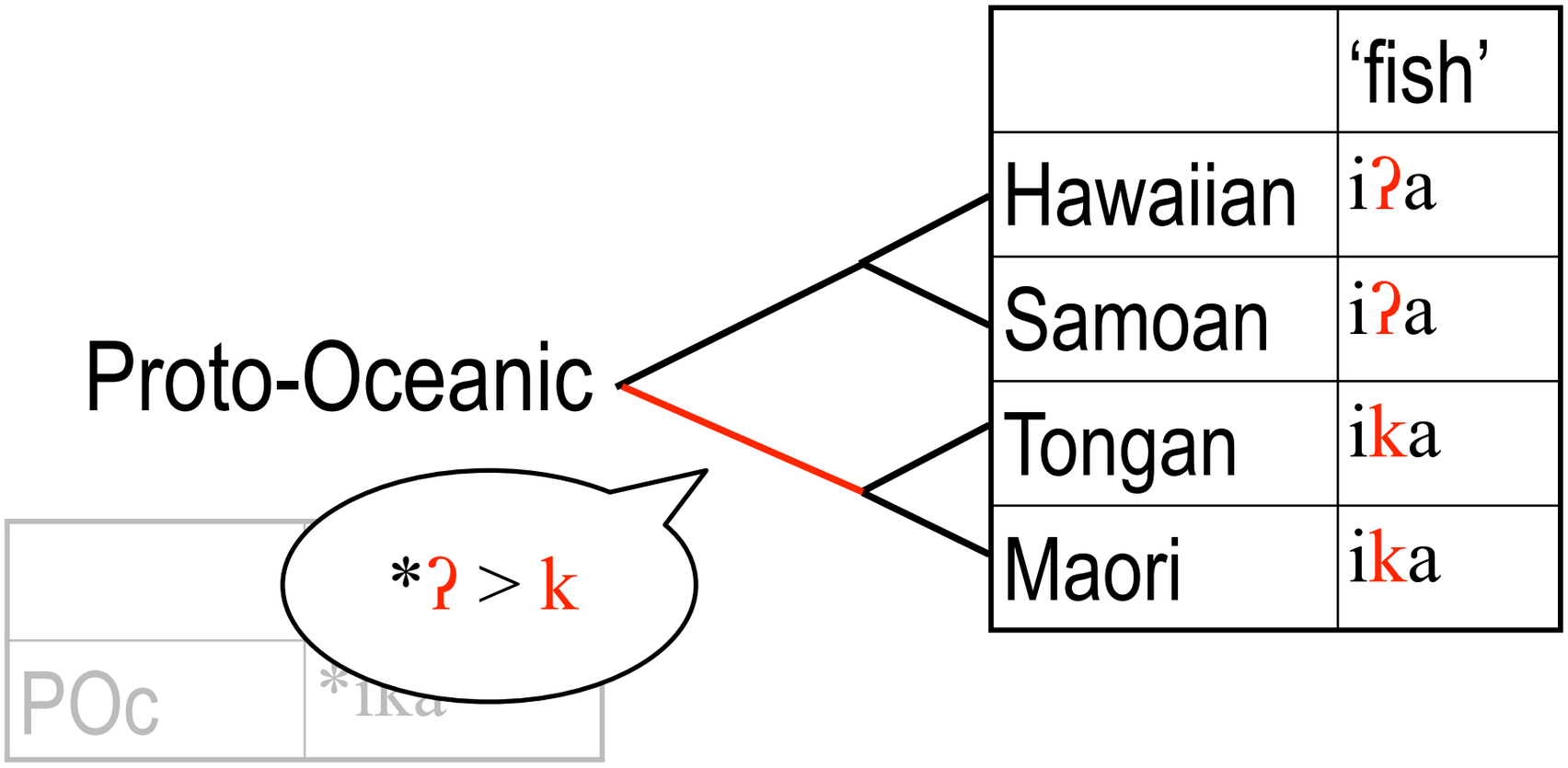
Task: comparative reconstruction



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

	‘fish’
POc	*ika

Task: comparative reconstruction



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

or

	‘fish’
POc	*iʔa

Task: comparative reconstruction

Proto-Oceanic

	‘fish’
POc	*ika

or

	‘fish’
POc	*iʔa

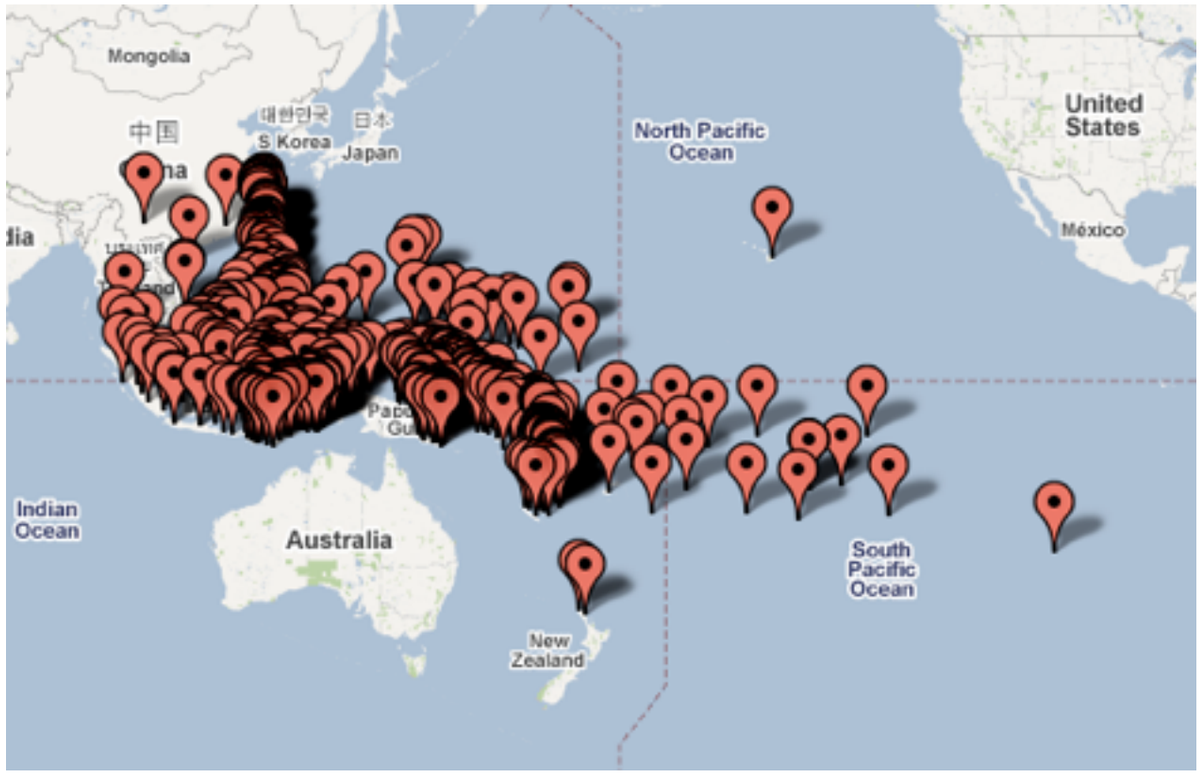
⋮

	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika

Previous work/main stream approach: manually reconstruct

Question: Can we leverage machine learning methods?

Can we harness more languages?



	‘fish’
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Maori	ika
Geser	ikan
Rapanui	ika
Nukuoro	iga
Niue	ika

Task 2: cognate inference

Fact: New words are invented & old words fall out of usage

Consequence: not all words forms for a given concept will have cousins in related languages.

	‘fish’	‘fear’	‘fear’
Hawaiian (G)	iʔa	makaʔu	
Samoan (S)	iʔa	mataʔu	
Tongan (T)	ika		manavahē
Maori (M)	ika	mataku	

Column: ‘cognate set’ (think as a cluster)

Task 2: cognate inference

Hawaiian	iʔa, makaʔu, ...
Samoan	iʔa, makaʔu, ...
Tongan	ika, manavahē, ...
Maori	ika, mataku, ...

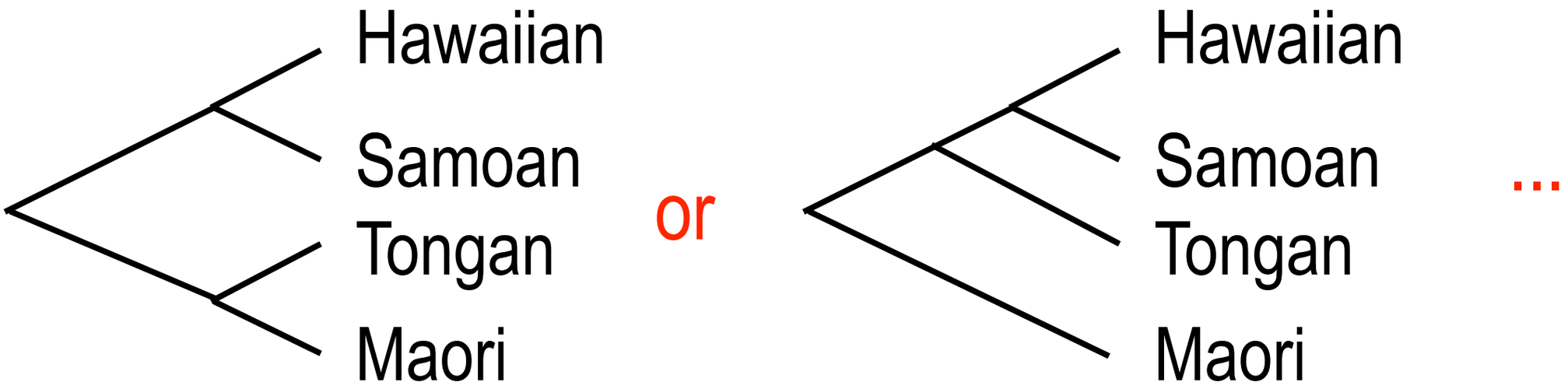
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	manavahē
Maori	ika	mataku

or

Hawaiian	iʔa	makaʔu	
Samoan	iʔa	mataʔu	
Tongan	ika		manavahē
Maori	ika	mataku	

...

Task 3: tree inference

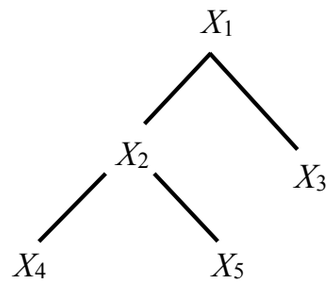


Model-based approach

Input: modern words

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	

Probabilistic models of change



$$X_2|X_1 \sim \mathrm{Mut}(X_1;\eta)$$
$$X_3|X_1 \sim \mathrm{Mut}(X_1;\eta)$$
$$\vdots$$
$$X_i|X_{\mathrm{par}(i)} \sim \mathrm{Mut}(X_{\mathrm{par}(i)};\eta)$$

Reconstruction

	‘fish’
POc	*ika

or

	‘fish’
POc	*iʔa

⋮

Cognates

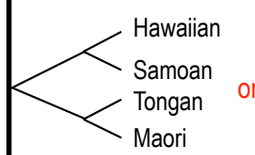
	‘fish’ (1)
Hawaiian	iʔa
Samoan	iʔa
Tongan	ika
Marshallese	yapil

or

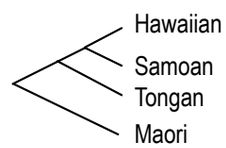
	‘fish’ (1)	‘fish’ (2)
Hawaiian	iʔa	
Samoan	iʔa	
Tongan	ika	
Marshallese		yapil

...

Trees



or

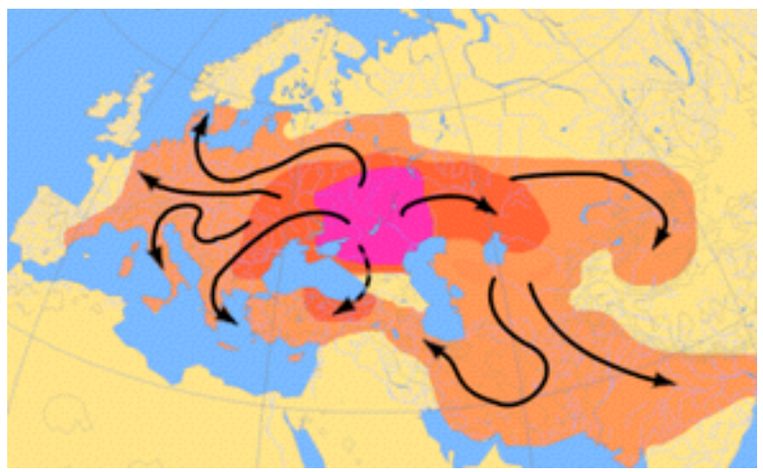


...

Motivation

Why reconstruct?

- Can answer a large number of questions about our past
  - Learn about ancient populations' migrations
  - Decipherment of ancient scripts



Motivation: typology

- What are the statistical regularities of sound change?
  - Role of functional load in sound change
- Useful for synchronic typology as well
  - Is an observed pattern the result of structural constraints, or the result of common descent?

Motivation: machine translation (MT)

- Long-term challenge: usable machine translation between *all* pairs of languages
  - Existing MT rely on large corpora in both languages in the pair,
  - in particular, *bitexts* (e.g. europarl)

- Concrete step:
  large scale cognate detection

Hawaiian	iʔa, makaʔu, ...
Samoan	iʔa, makaʔu, ...
Tongan	ika, manavahē, ...
Maori	ika, mataku, ...

Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	manavahē
Maori	ika	mataku

or

Hawaiian	iʔa	makaʔu	
Samoan	iʔa	mataʔu	
Tongan	ika		manavahē
Maori	ika	mataku	

Background

International Phonetic Alphabet

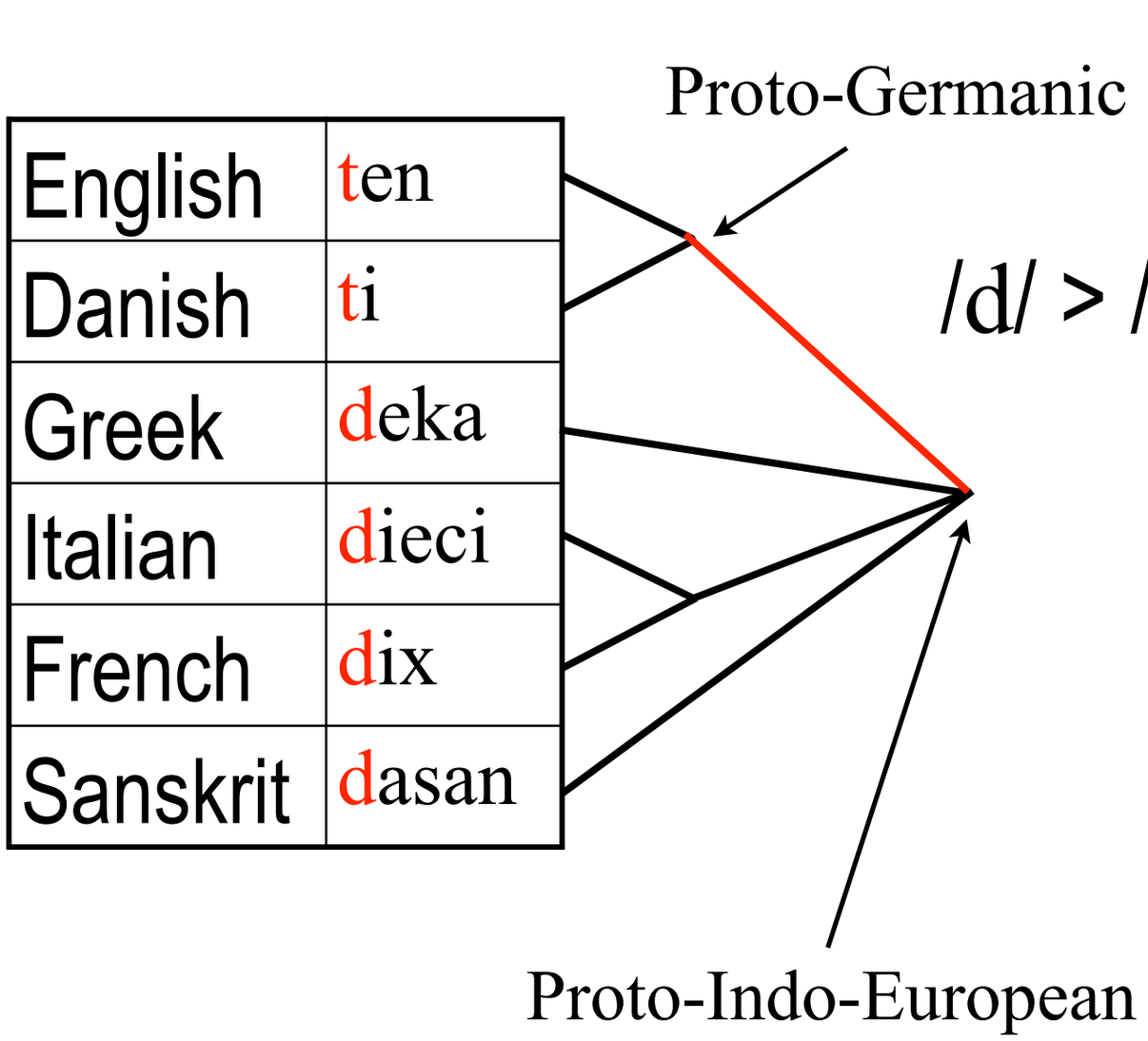
Instead of {A,C,G,T}, here the alphabet is a set of basic sound units (*phonemes*):

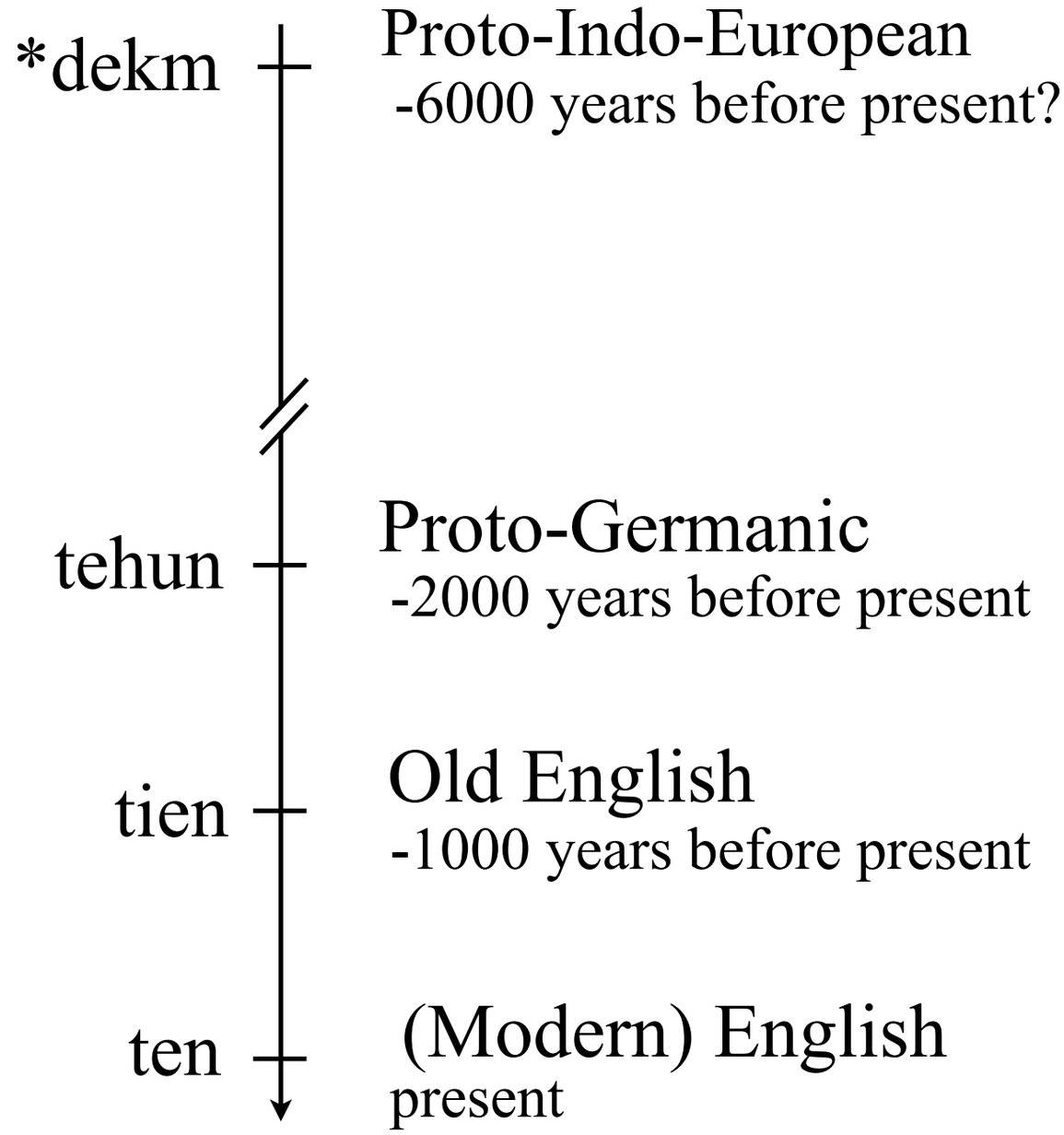
	m	ɱ	n	ŋ	ɴ	
p	b		t	d	k g	q ɢ	ʔ
ɸ	β	f	v	s	z	x ɣ	ʁ	h	ɦ

i	y			ɨ	ʉ			ɯ	u	
	e	ø			ɘ	ɵ		ɤ	o	
		ɛ	œ			ɜ	ɞ		ʌ	ɔ
			a	œ				ɑ	ɒ	

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
	iʔa	mataʔu
	ika	
to-Oceanic	ika	mataku

Sound changes





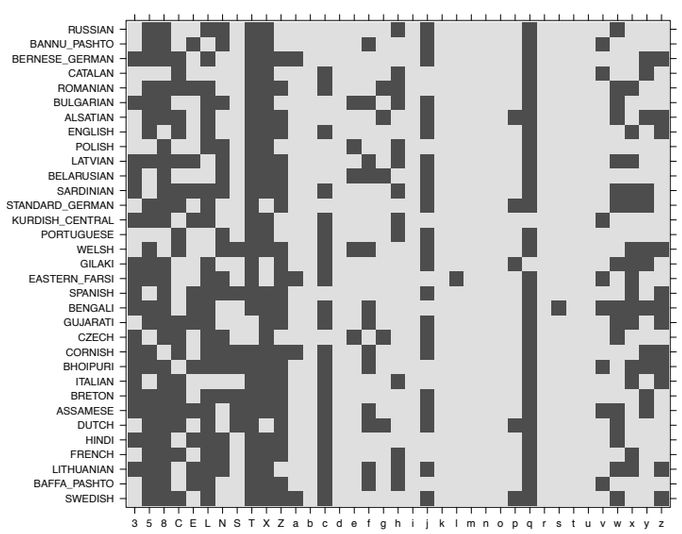
Regularity in sound change

English	ten	tooth
Danish	ti	tand
Greek	deka	donti
Italian	dieci	dente
French	dix	dent
Sanskrit	dasan	danta

/d/ > /t/ in word initial position

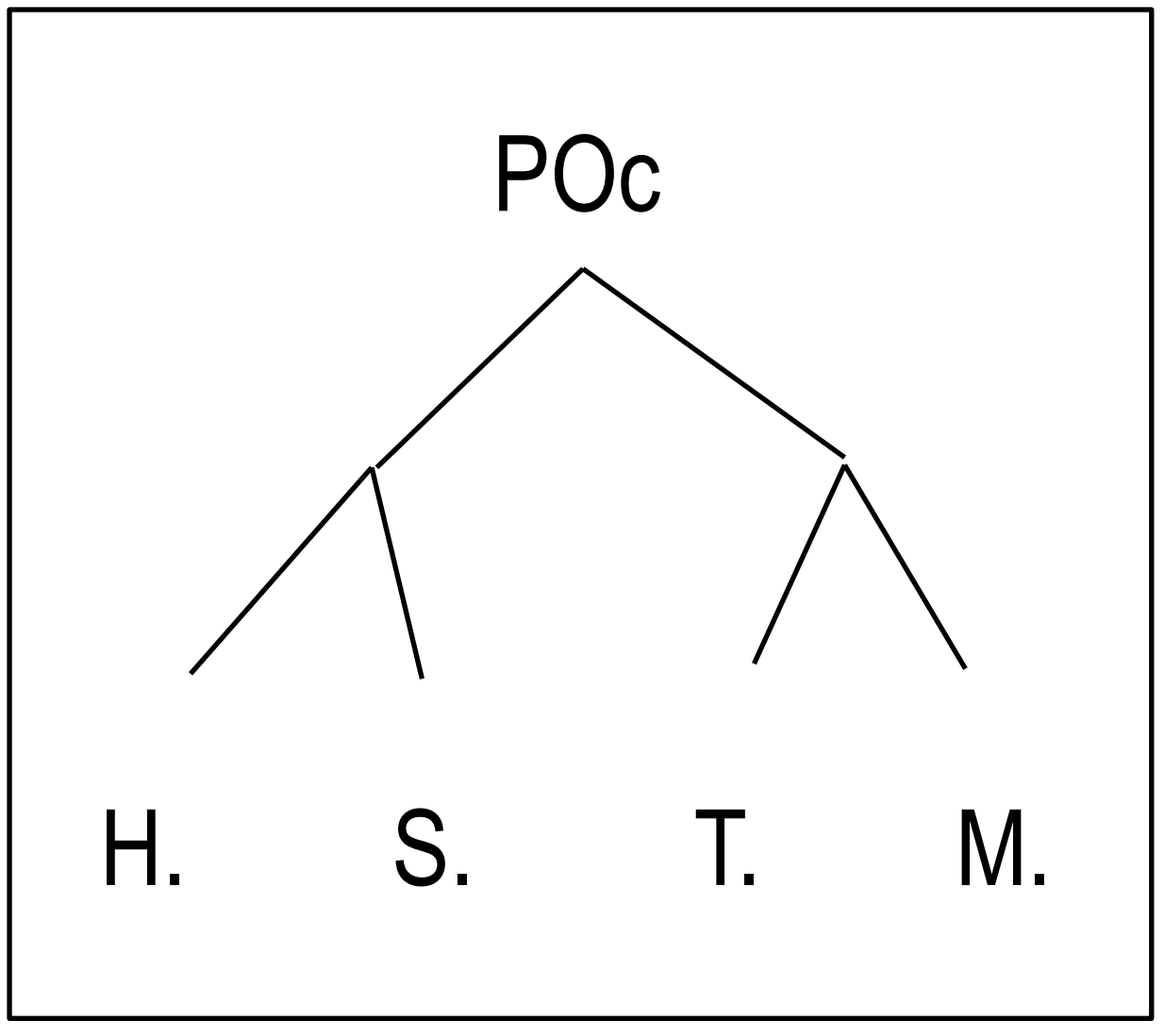
Proto-Indo-European

- Probabilistic models of inventory changes?
  - Open computational problem



Modelling phonological change

Simplifying assumptions

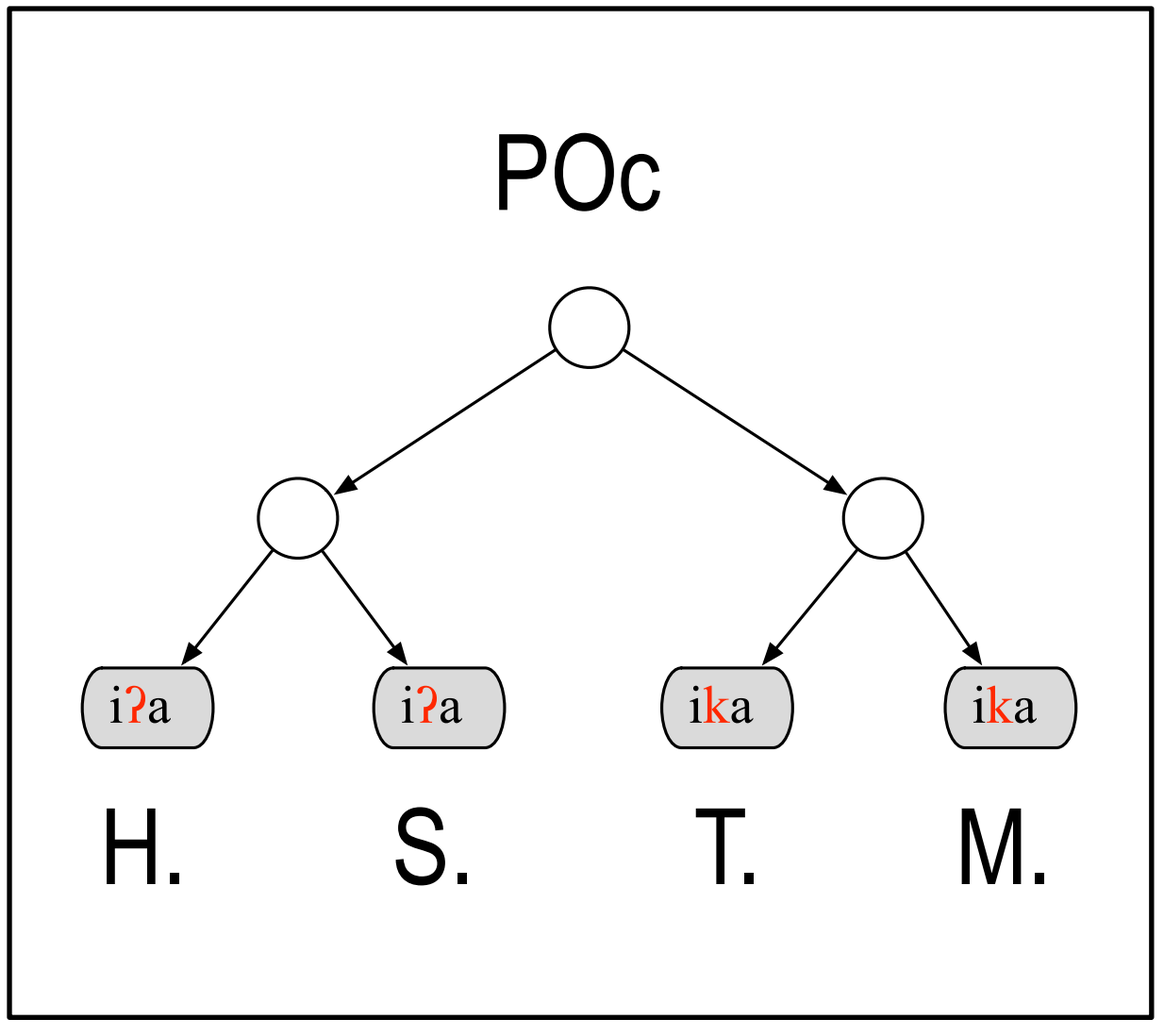


Fixed tree

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

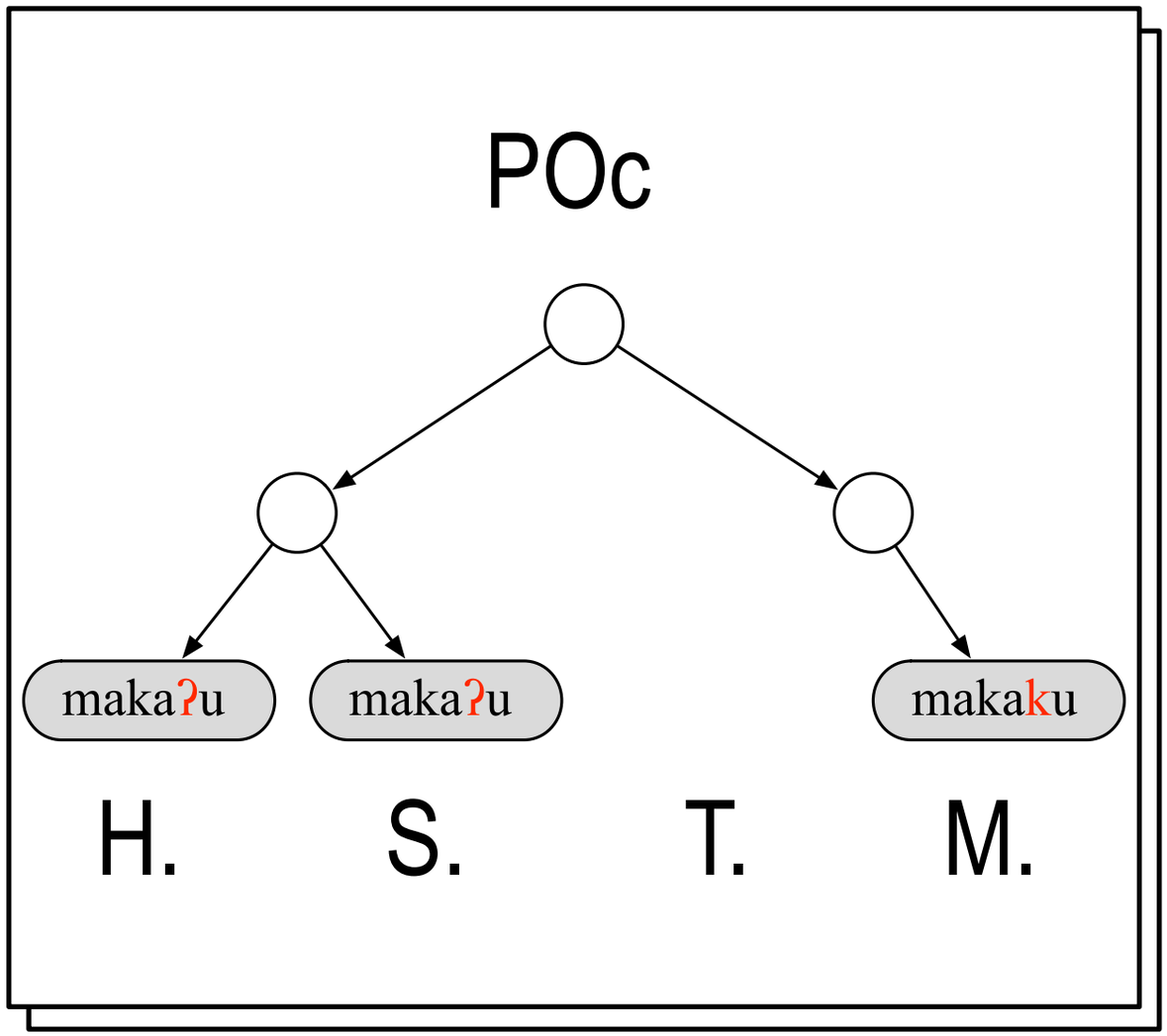
Fixed cognate sets

Graphical model



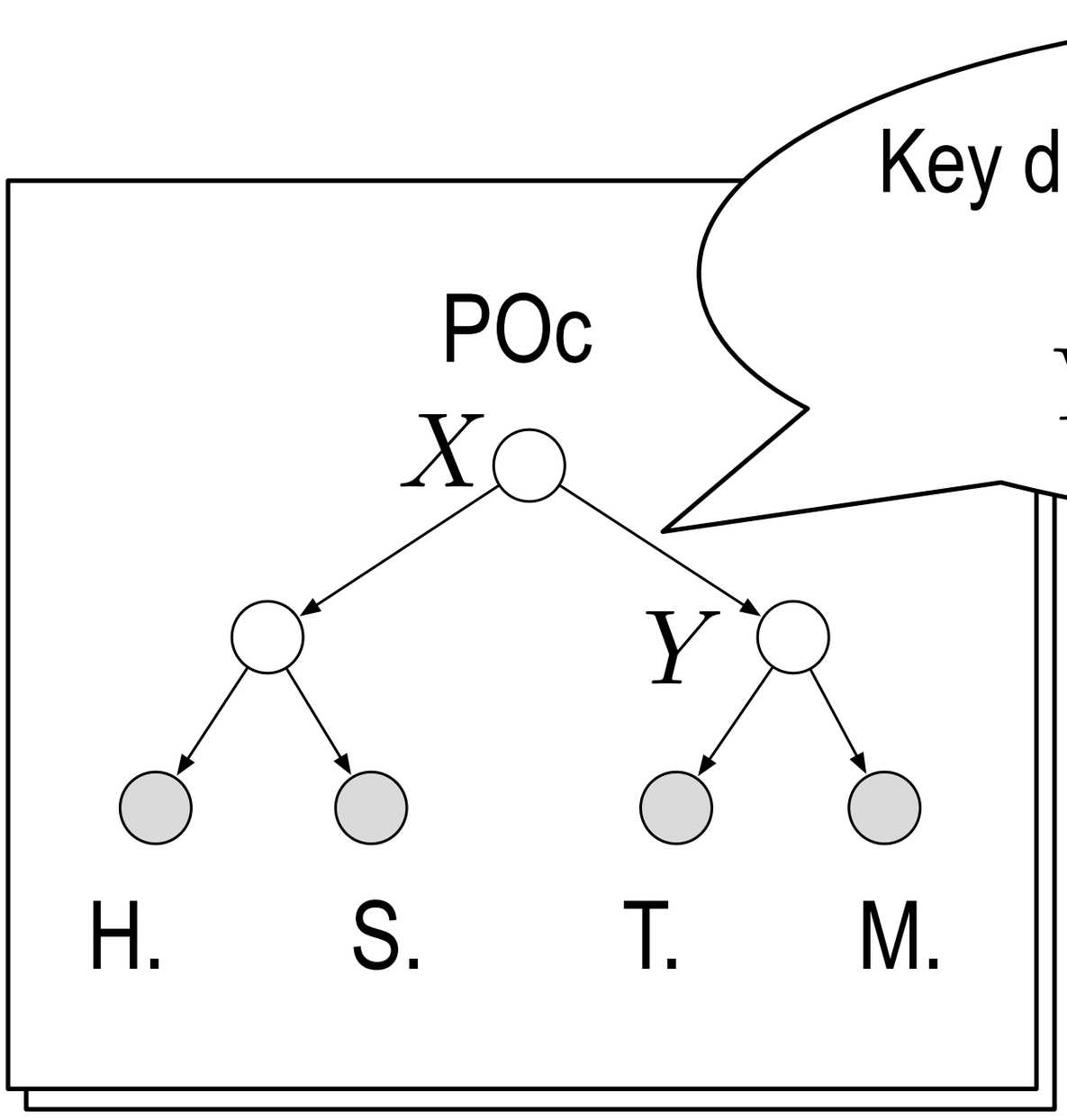
	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Graphical model



Hawaiian		makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Maori	ika	mataku

Modeling string mutation

- What kind of string mutations need to be captured?
  - Substitution

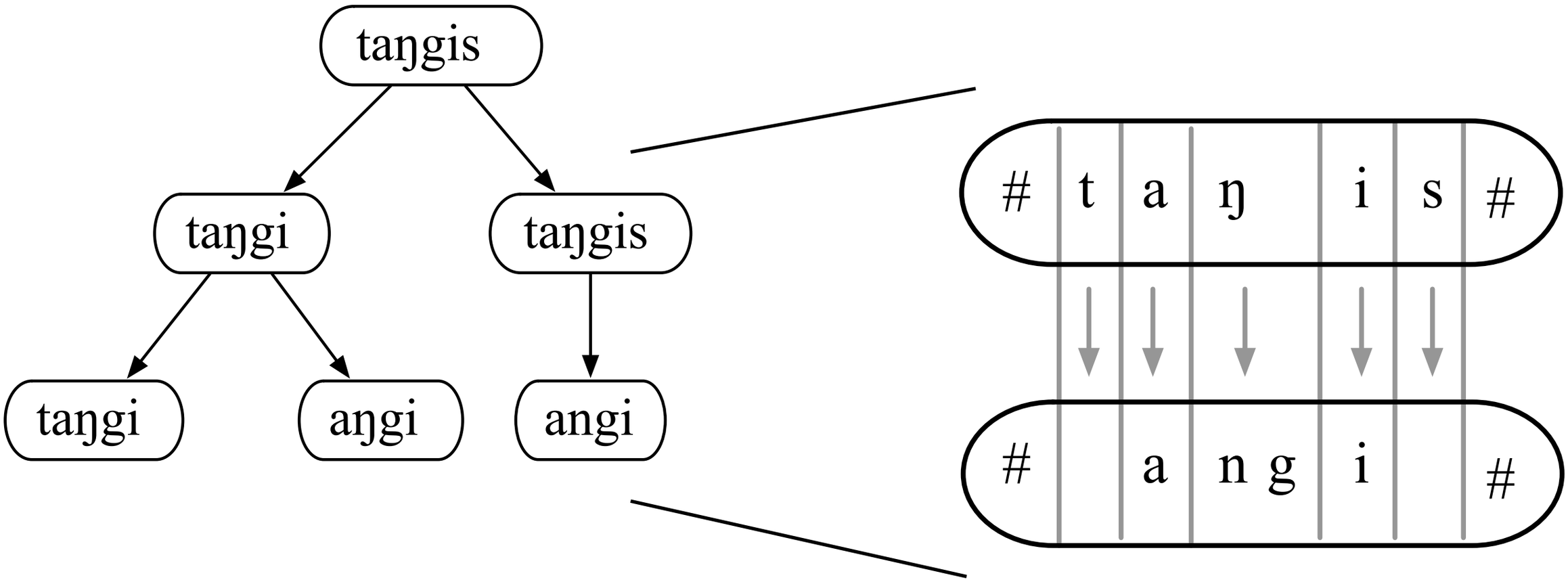
    *k > ʔ
  - Deletion (& insertions)

    *ʔ > Ø
  - Contextualized change

    *ʔ > Ø / V _ V

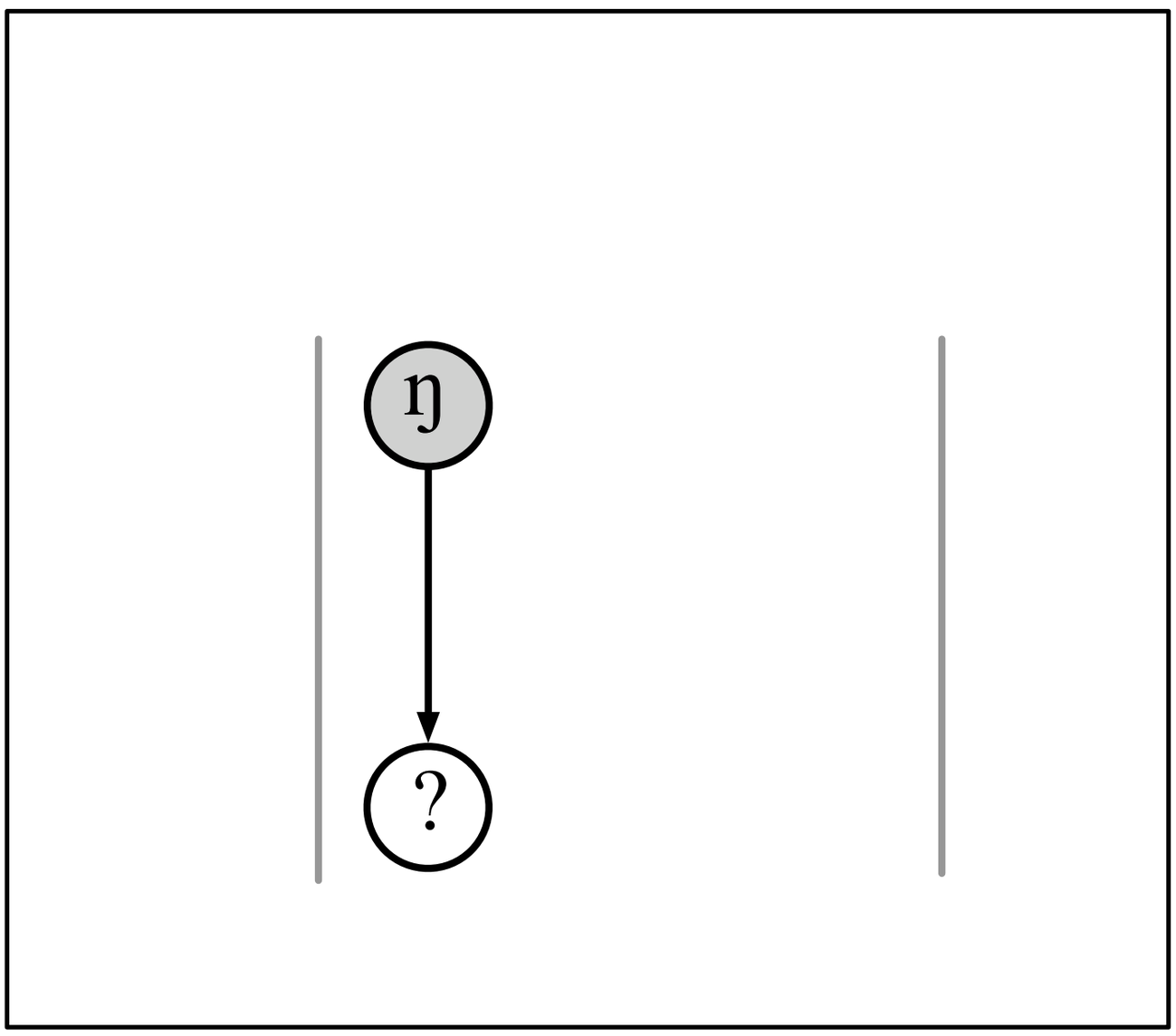
	‘fish’	‘voice’
Hawaiian	iʔa	leo
Samoan	iʔa	leo
Tongan	ika	leʔo
Maori	ika	reo

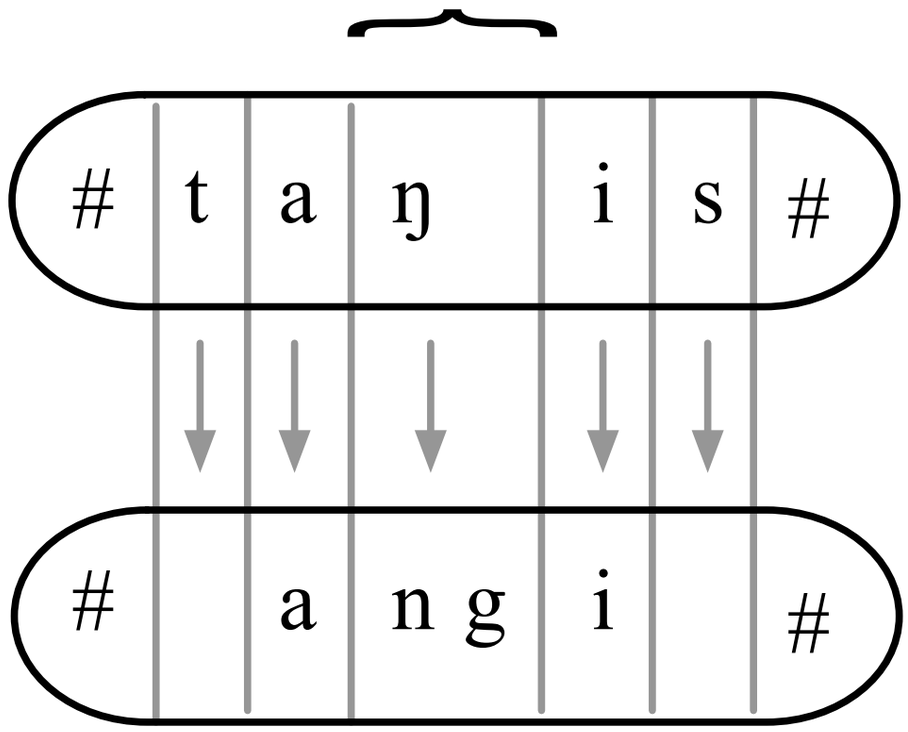
String transducer



‘to cry’

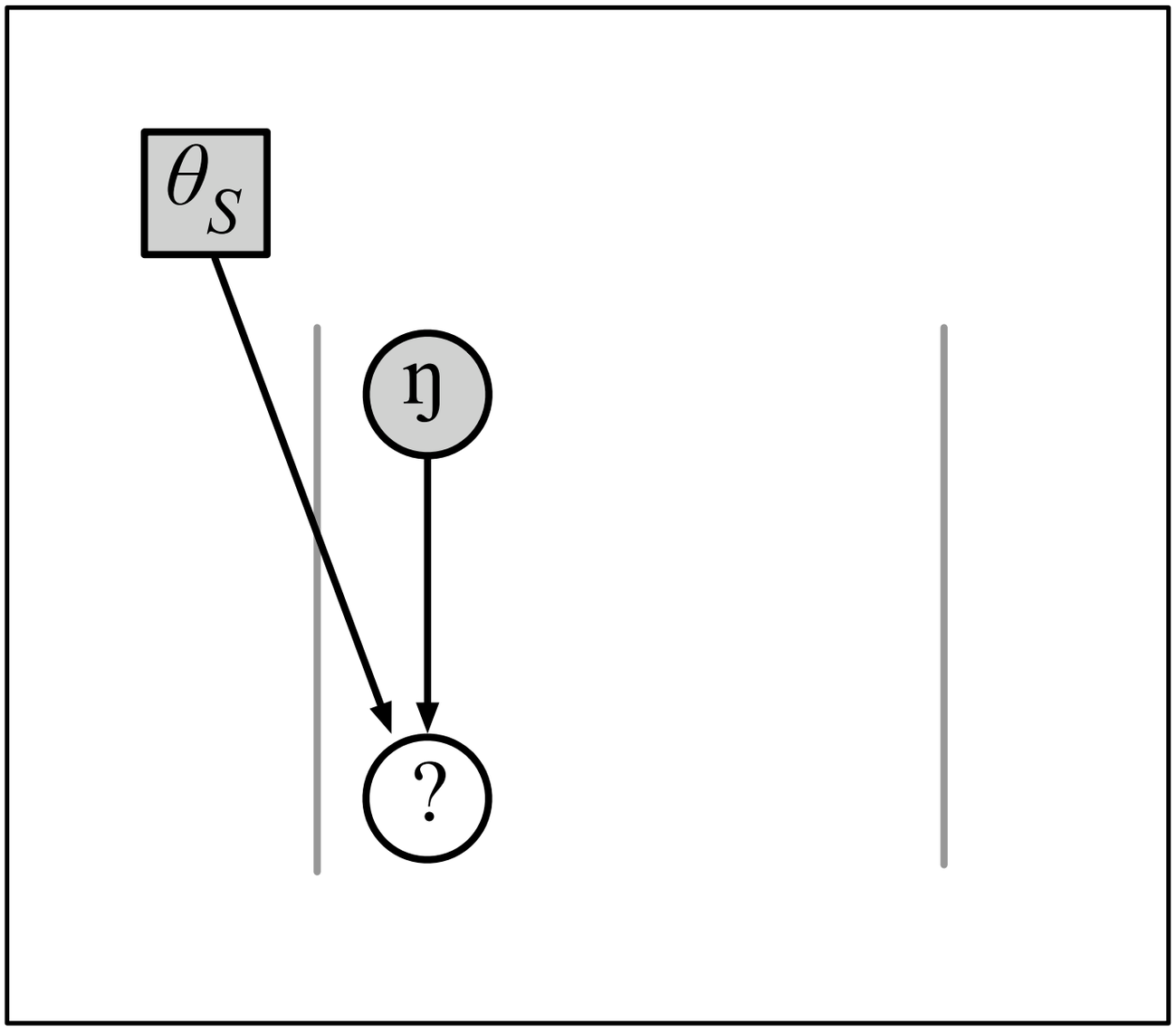
String transducer

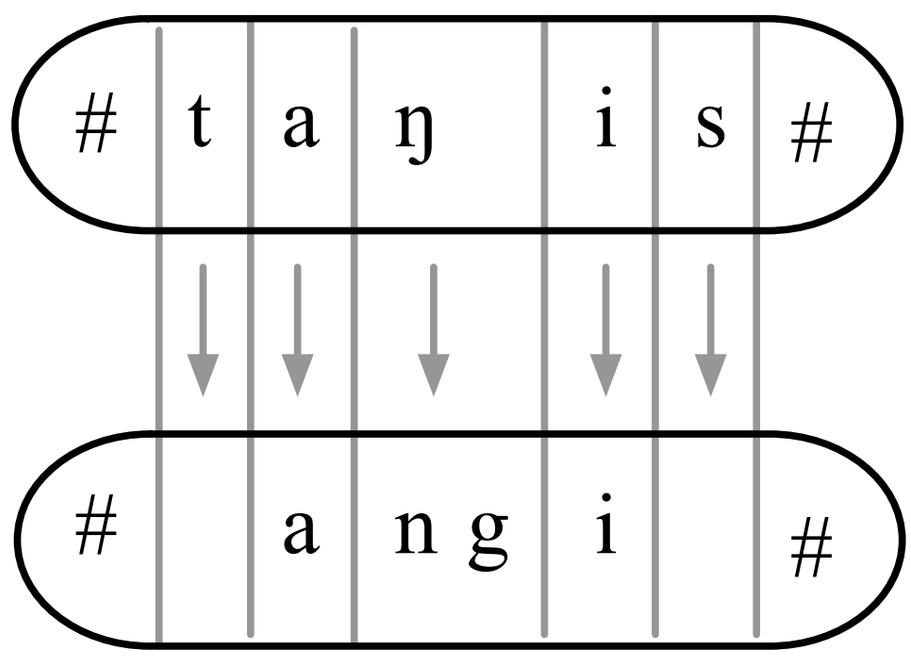




String transducer

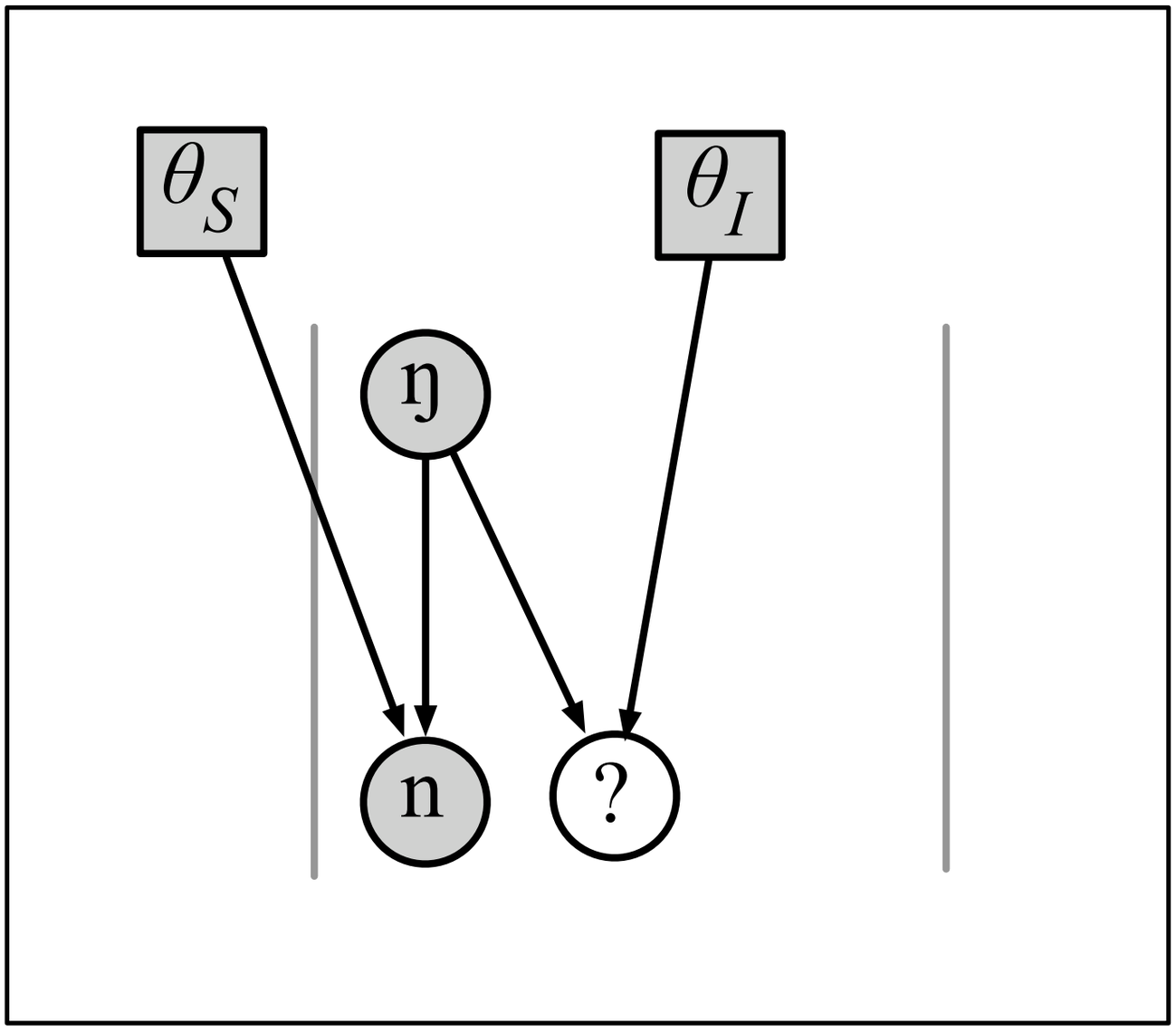
θ_S: Substitution/Deletion Parameters

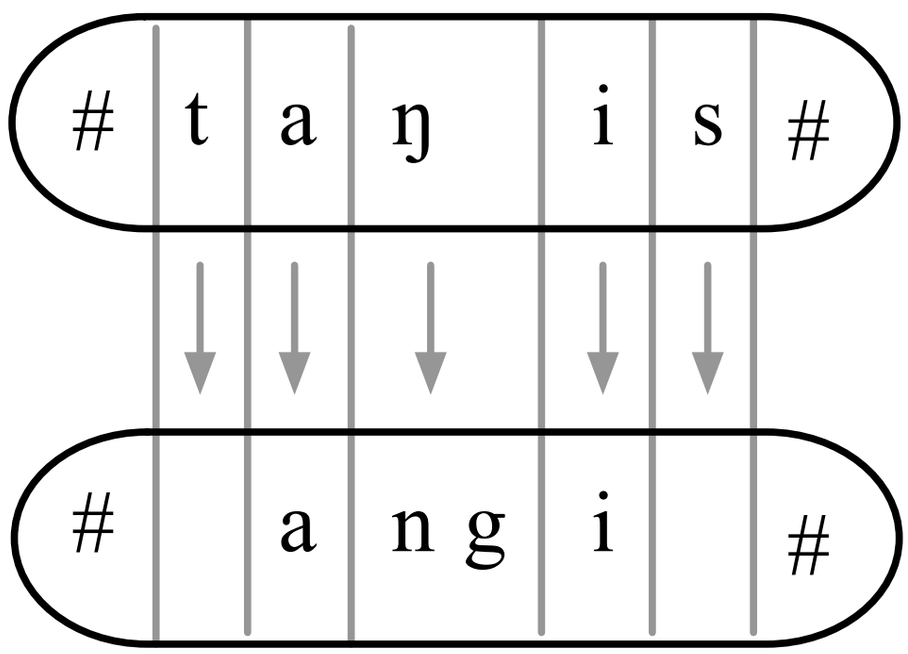




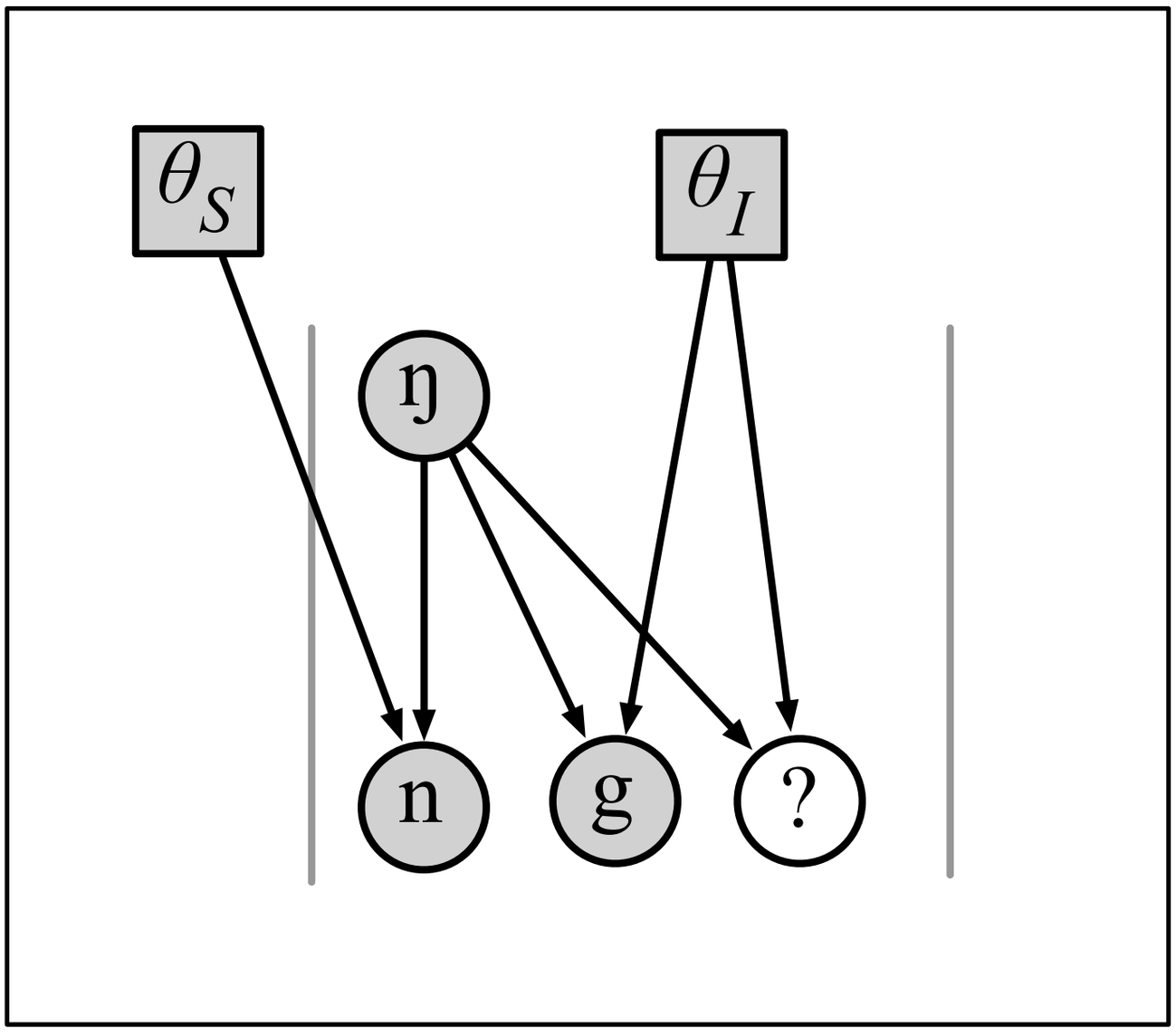
String transducer

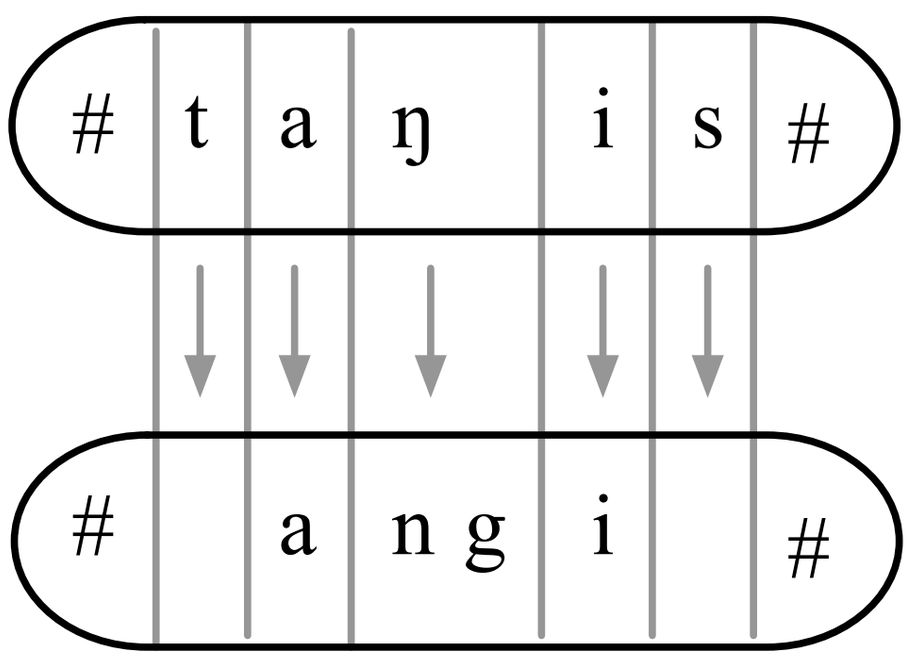
θ_I: Insertion Parameters





String transducer

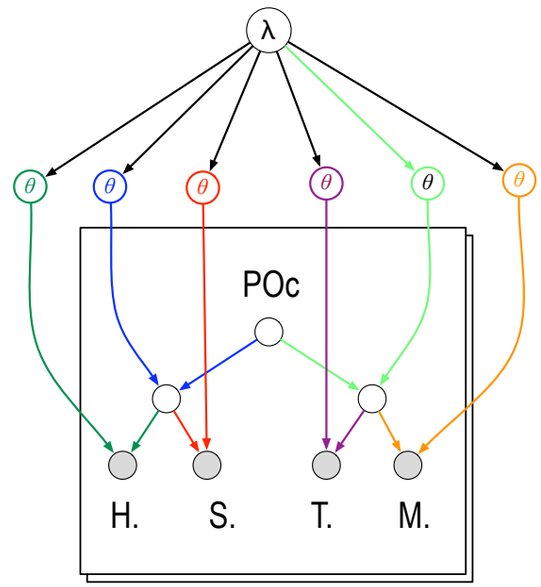




Bayesian inference

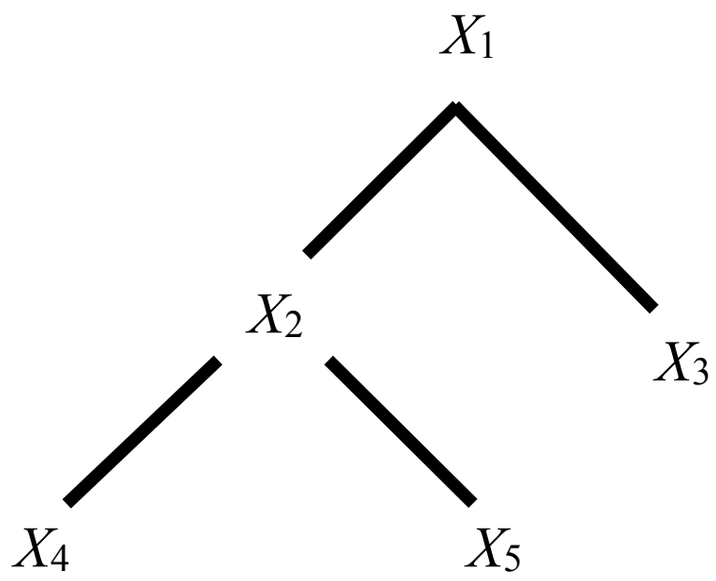
Computational bottleneck

Parameters fitting: can be approached using standard tools



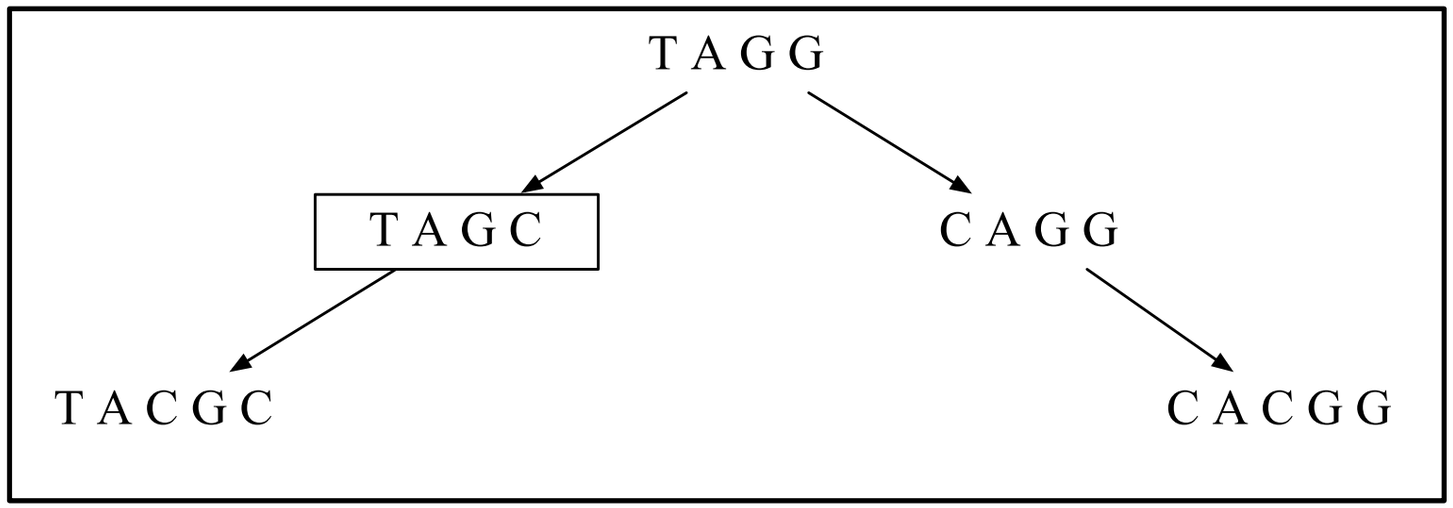
More difficult: string-valued expectations

$$X_1, X_2 | X_3, X_4, X_5$$

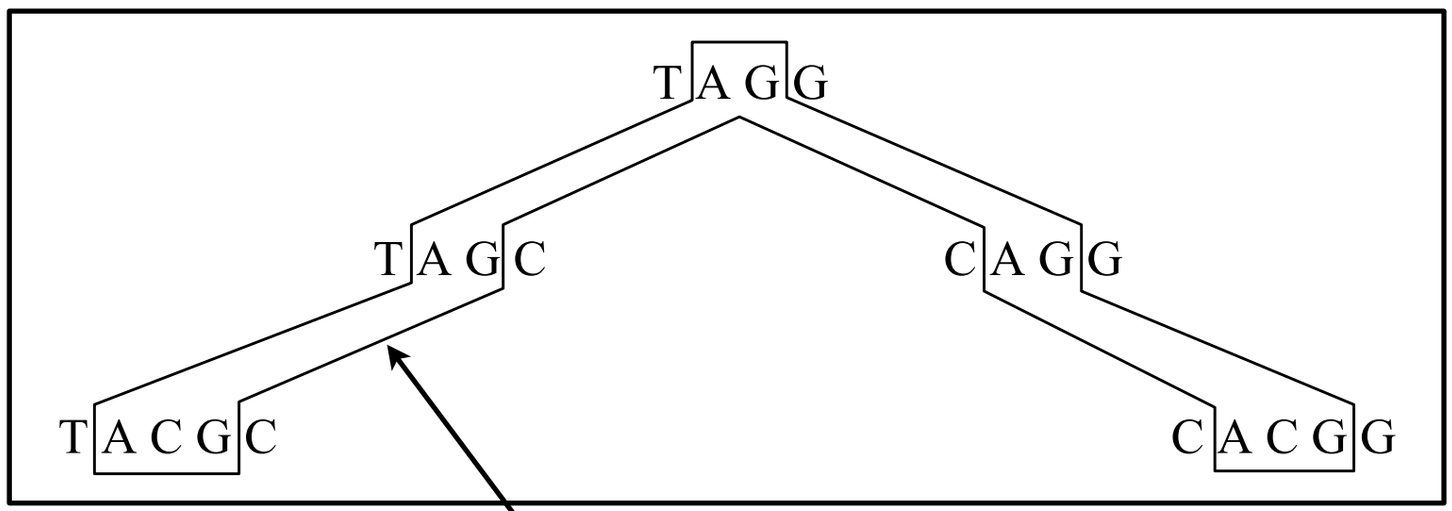


Various MCMC algorithms

Gibbs sampling



Ancestry resampling



Thin vertical section

Sequential Monte Carlo

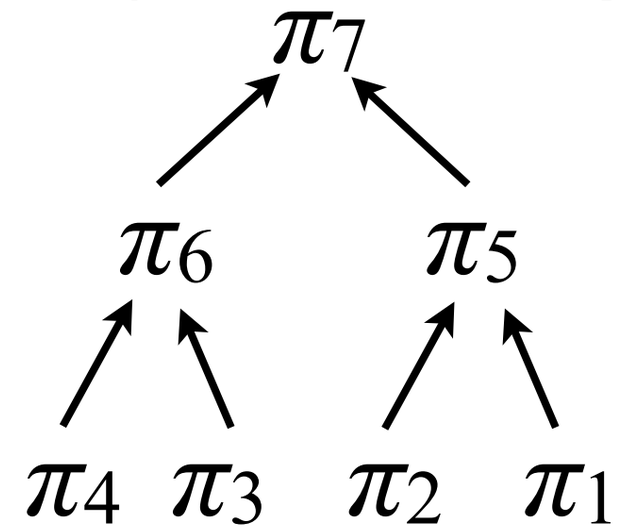
- SMC: An alternative to MCMC
  - Approximates posterior with particles and weights
- D&C SMC: generalization more suitable to phylogenies
- tinyurl.com/divide-and-conquer-smc
- Fredrik Lindsten, Adam M. Johansen, Christian A. Naesseth, Bonnie Kirkpatrick, Thomas B. Schön, John Aston, and Alexandre Bouchard-Côté. (2016) Divide-and-Conquer with Sequential Monte Carlo. JCSG, In Press.

Standard SMC samplers

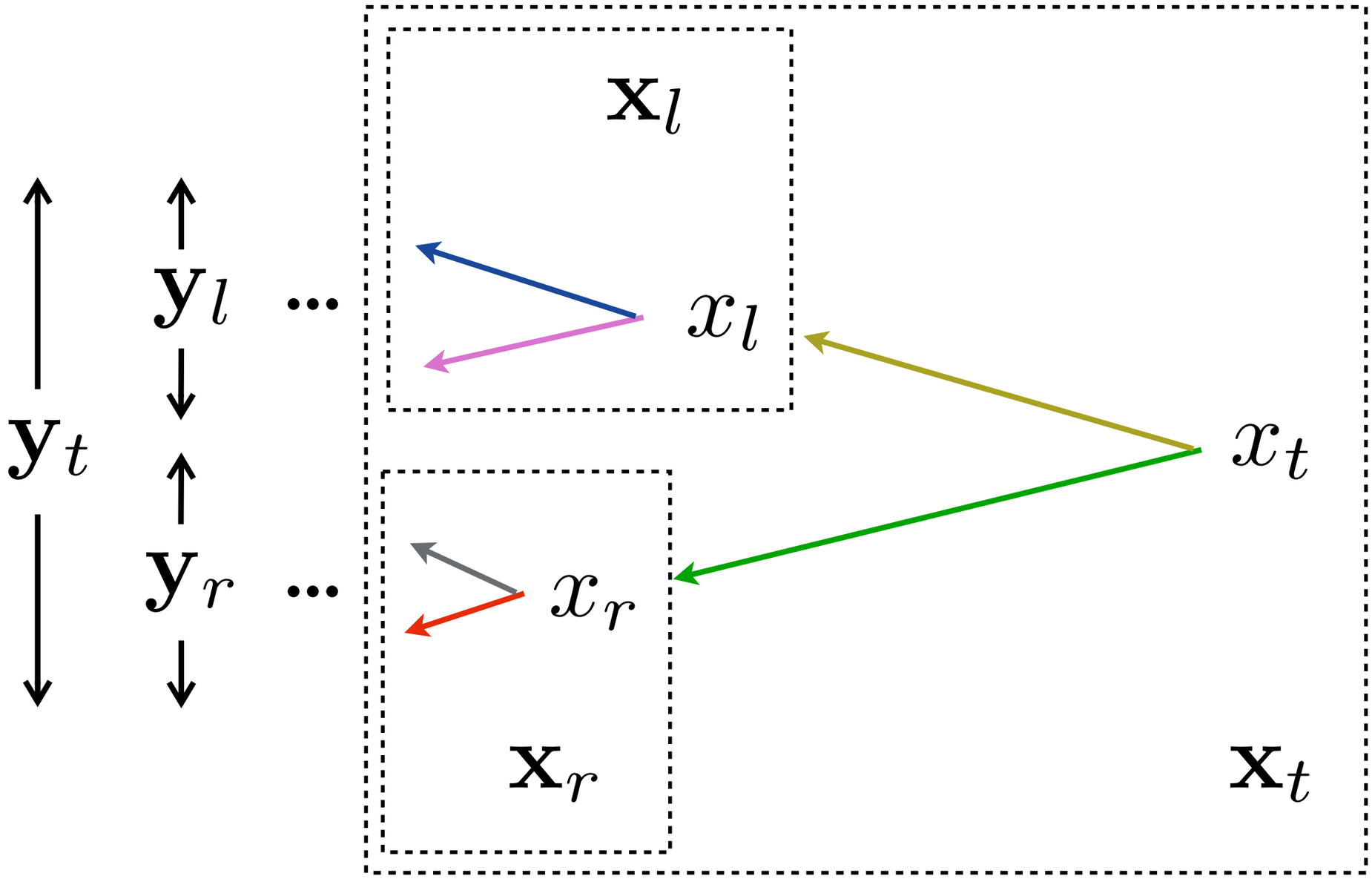
$\pi_1 \longrightarrow \pi_2 \longrightarrow \pi_3 \longrightarrow \ldots$

Divide and Conquer

(D&C) SMC samplers



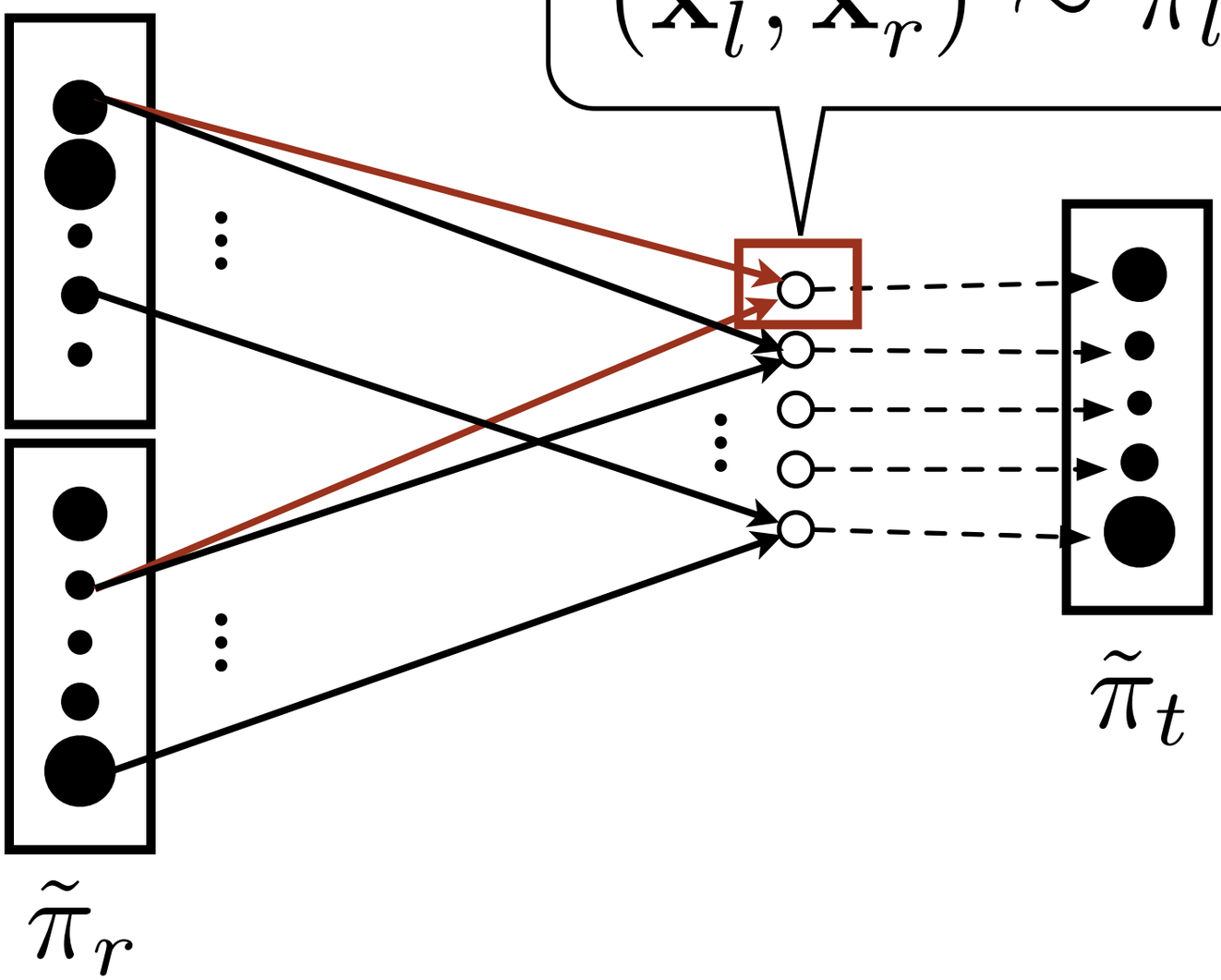
D&C SMC: notation



D&C (SIR):

1. Assume inductively we have $\tilde{\pi}_l$ & $\tilde{\pi}_r$
2. Sample from the product measure:
   - amounts to two independent multinomial sampling steps

$$(\tilde{\mathbf{x}}_l^i, \tilde{\mathbf{x}}_r^i) \sim \tilde{\pi}_l \times \tilde{\pi}_r$$

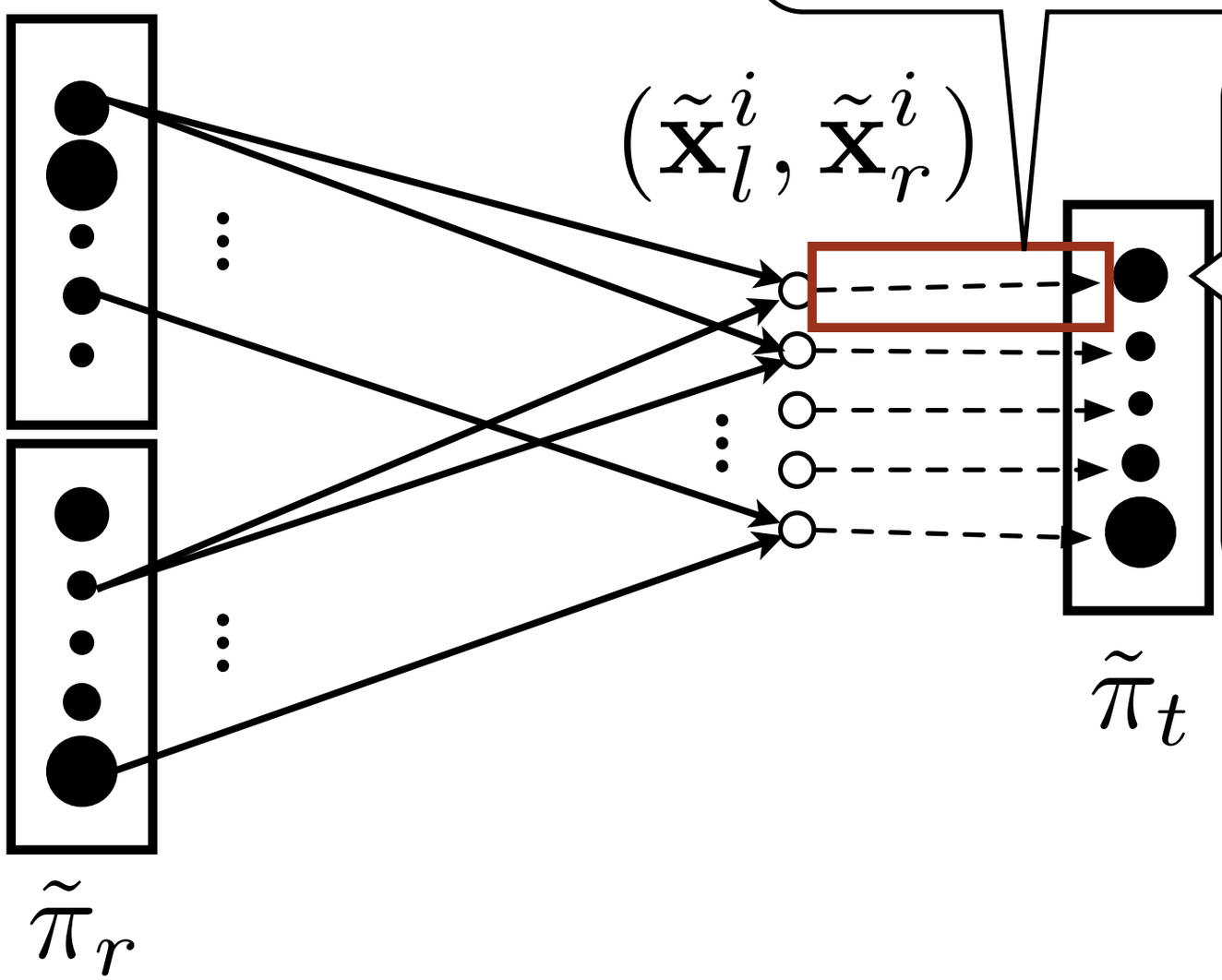


D&C (SIR):

1. Assume inductively...
2. Sample from the product measure:
3. Propose (extend):

$$x_t^i | (\tilde{\mathbf{x}}_l^i, \tilde{\mathbf{x}}_r^i) \sim q_t(\cdot | \tilde{\mathbf{x}}_l^i, \tilde{\mathbf{x}}_r^i)$$

?



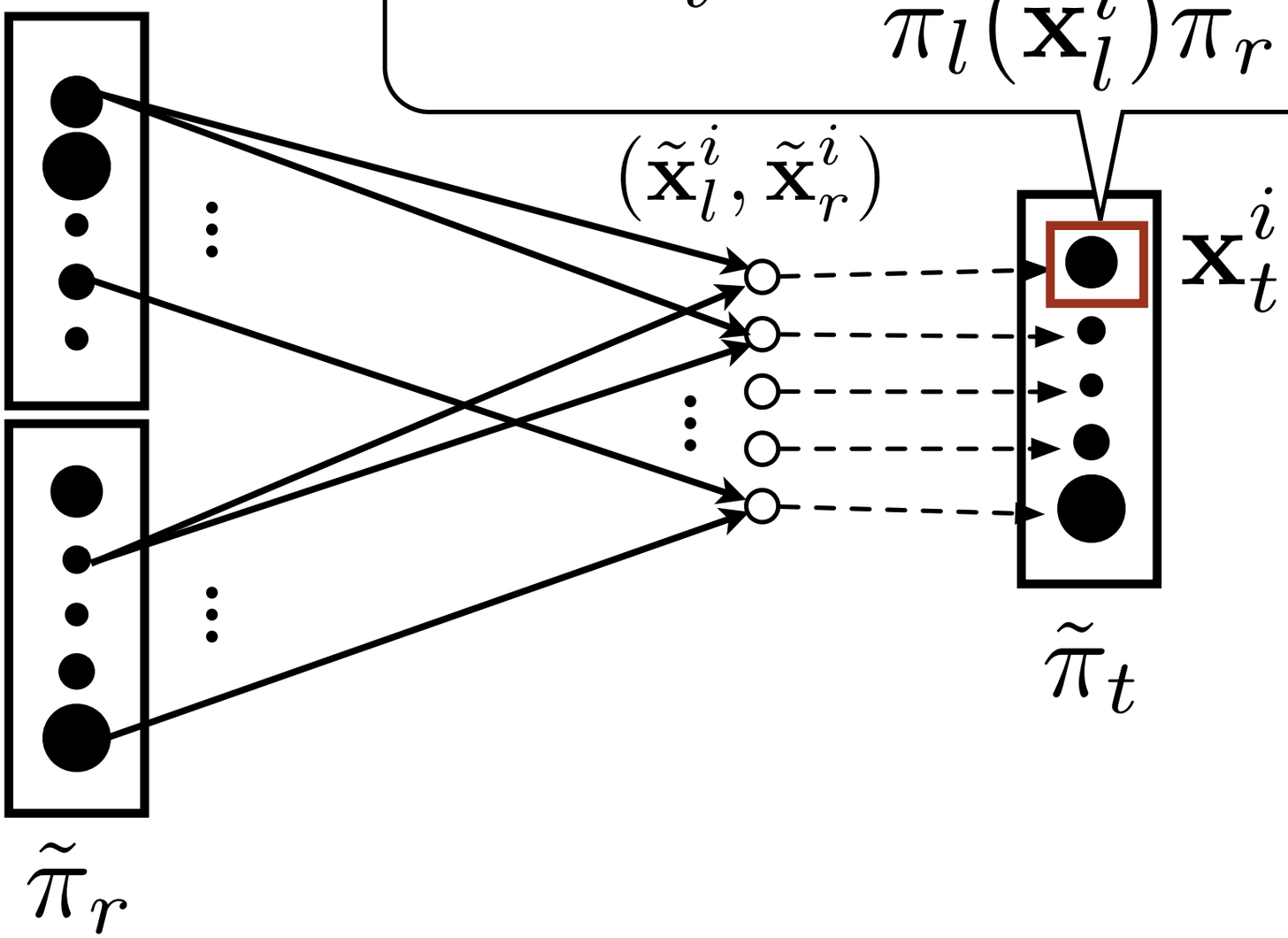
Concatenate:

$$\mathbf{x}_t^i = (x_t^i, -\tilde{\mathbf{x}}_l^i -, -\tilde{\mathbf{x}}_r^i -)$$

D&C (SIR):

1. Assume inductively...
2. Sample from the product measure
3. Propose (extend)
4. Reweigh:

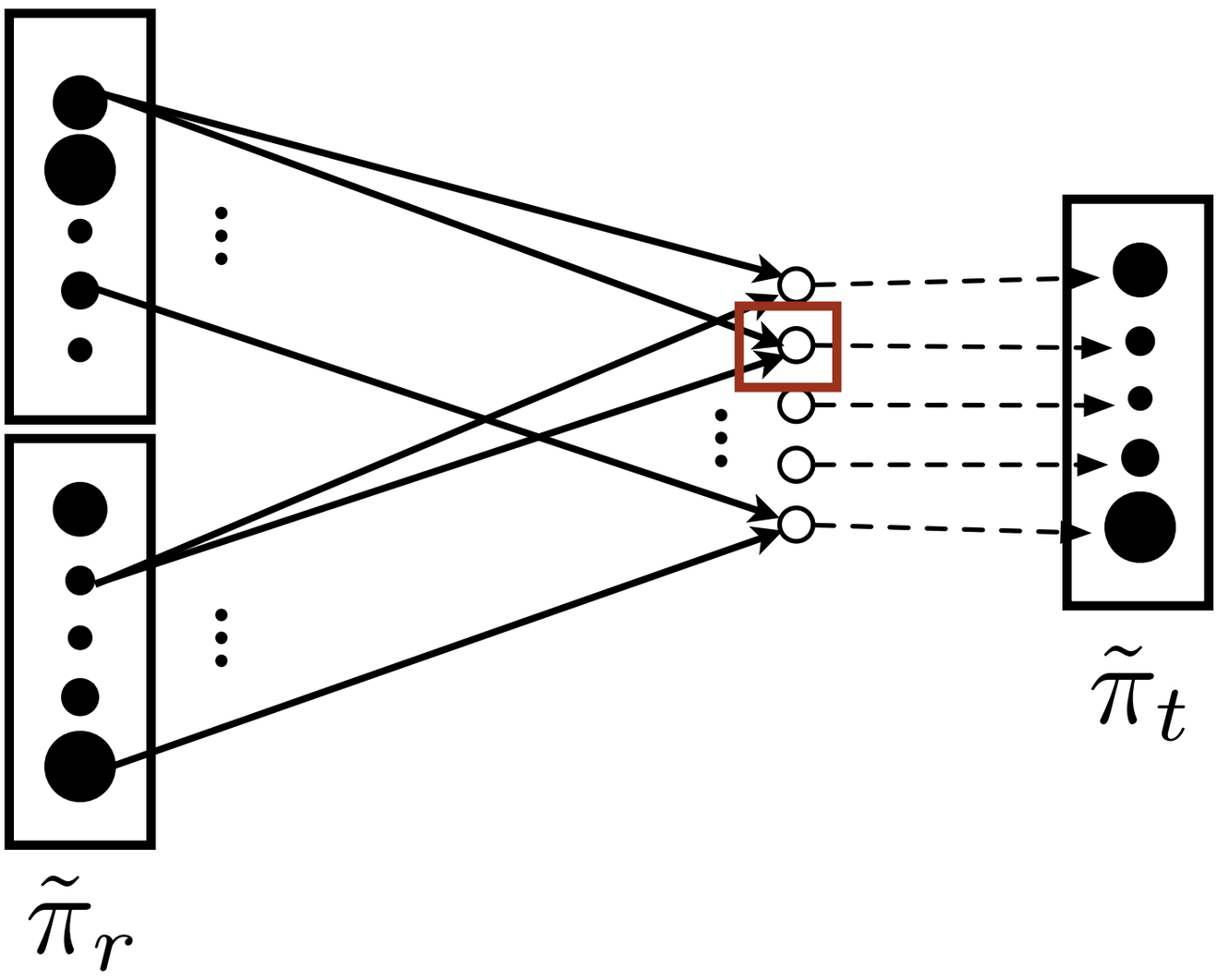
$$w_t^i = \frac{\pi_t(\mathbf{x}_t^i)}{\pi_l(\tilde{\mathbf{x}}_l^i)\pi_r(\tilde{\mathbf{x}}_r^i)} \frac{1}{q_t(x_t^i|\tilde{\mathbf{x}}_l^i, \tilde{\mathbf{x}}_r^i)}$$



D&C (SIR):

1. Assume inductively...
2. Sample from the product measure
3. Propose (extend)
4. Reweigh

Repeat for each particle



Potential applications

- Inference over non-local/catastrophic events (explicit sound change operators)
- Joint cognate inference
- Statistical multiple sequence alignment
- Relaxed clock models
- Many other models that would otherwise have to be handled with very expensive methods such as Approximate Bayesian Computation

Numerical examples

Validation methodology

Holding-out the manual reconstructions

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Proto-Oceanic	*ika	*mataku

	‘fish’	‘fear’
Hawaiian	iʔa	makaʔu
Samoan	iʔa	mataʔu
Tongan	ika	
Proto-Oceanic	???	???

Evaluation criterion: Edit distance

Smallest number of substitutions, insertions and deletions needed to go from one string to the other

e.g.: d(/ika/ , /ga/) = 2

/ika/ → /iga/ → /ga/

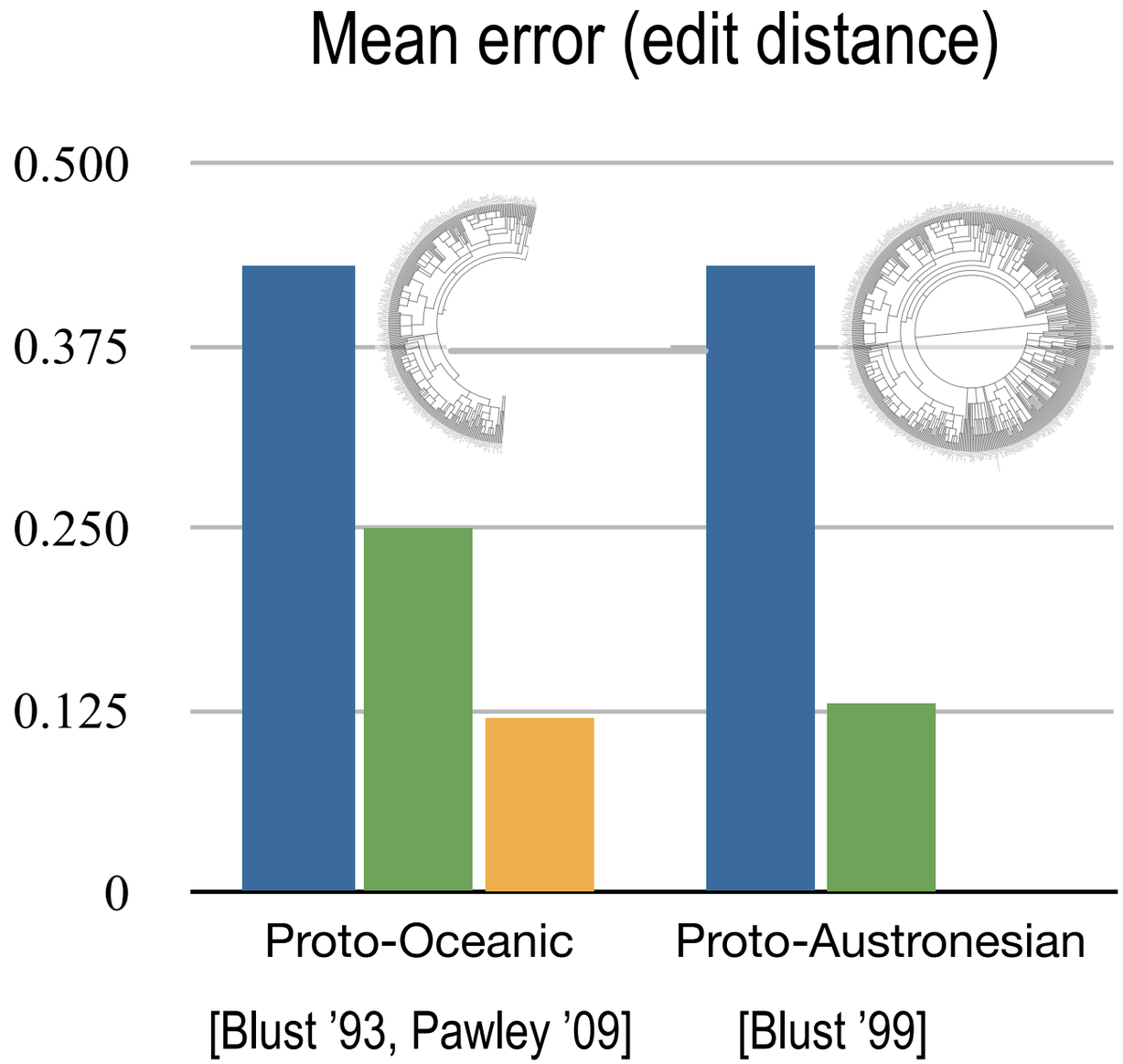
Datasets used for reconstruction

- Austronesian languages dataset [Greenhill et al. 08]
  - 706 languages, 150k words
  - 17 annotated reconstructions
  - Ideal for testing large scale reconstruction
  - Has continued to grow since the time experiments were performed

Comparisons in large phylogenies

Comparisons on Proto-Oceanic and Proto-Austronesian

- Baseline: sampling from the observed words
- This work
- Distance between two linguists' reconstructions



Examples of reconstruction

	Known Modern Languages				Reconstructed Ancestors		
Gloss	Fijian	Pazeh	Melanau	Inabaknon	Manual	Automated	Δ
star	kalokalo	mintol	biten	bitu'on	*bituqen	*bituqen	0
to hold	taura	maːraʔ	magem	kumkom	*gemgem	*gemgem	
house	vale	xumaʔ	lebuʔ	ruma	*ʀumaq	*ʀumaq	
bird	manumanu	aiam	manuk	manok	*qayam	*qayam	
to cut, hack	tata	taːtatak	tutek	hadhad	*taʀaq	*taʀaq	
at	e	N/A	gaʔ	N/A	*i	*i	
what?	cava	ʔaxai	uaʔ inew	ay	*nanu	*anu	1
this	oqo	ʔimini	itew	yayto	*ini	*ani	
wind	cagi	varə	paŋay	bariyo	*bali	*beliu	2

Colors here indicate cognate sets

Stratified sampling by edit distance