BAYESIAN MCMC

aims

To come away with an intuitive* understanding of:

(1) maximum likelihood
(2) Bayes Theorem
(3) what they mean for phylogenetic inference

(4) a basis for understanding how** they work

*as little mathematics as possible
**requires equations

We find the set of trees that maximise the Lh

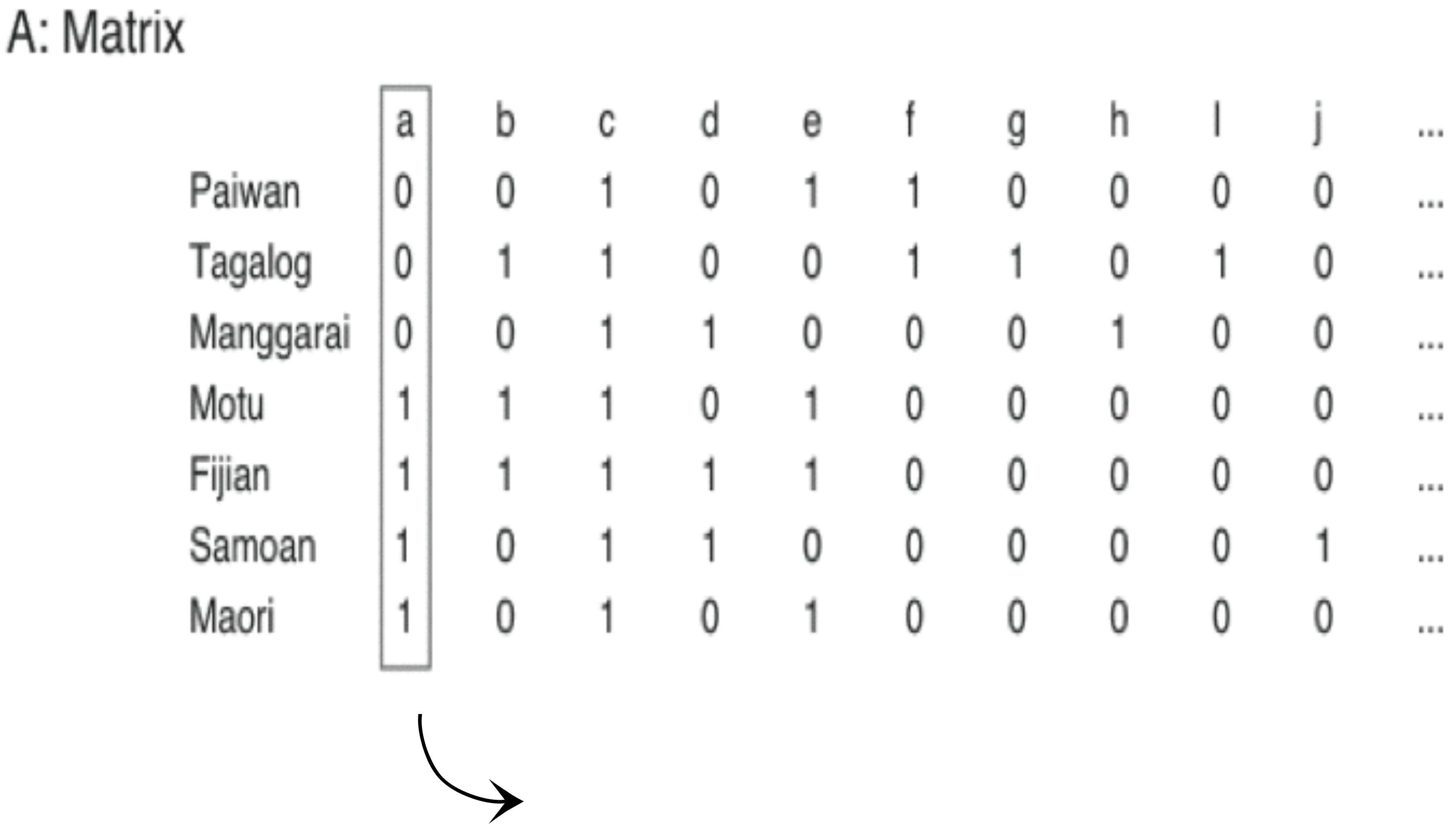
$$p(X|\theta) = Lh$$

X = data
θ = model parameters

{tree topology, branch lengths, node values}
{mathematical description of change}

Take a deep breath

Calculating the likelihood
X = observed data

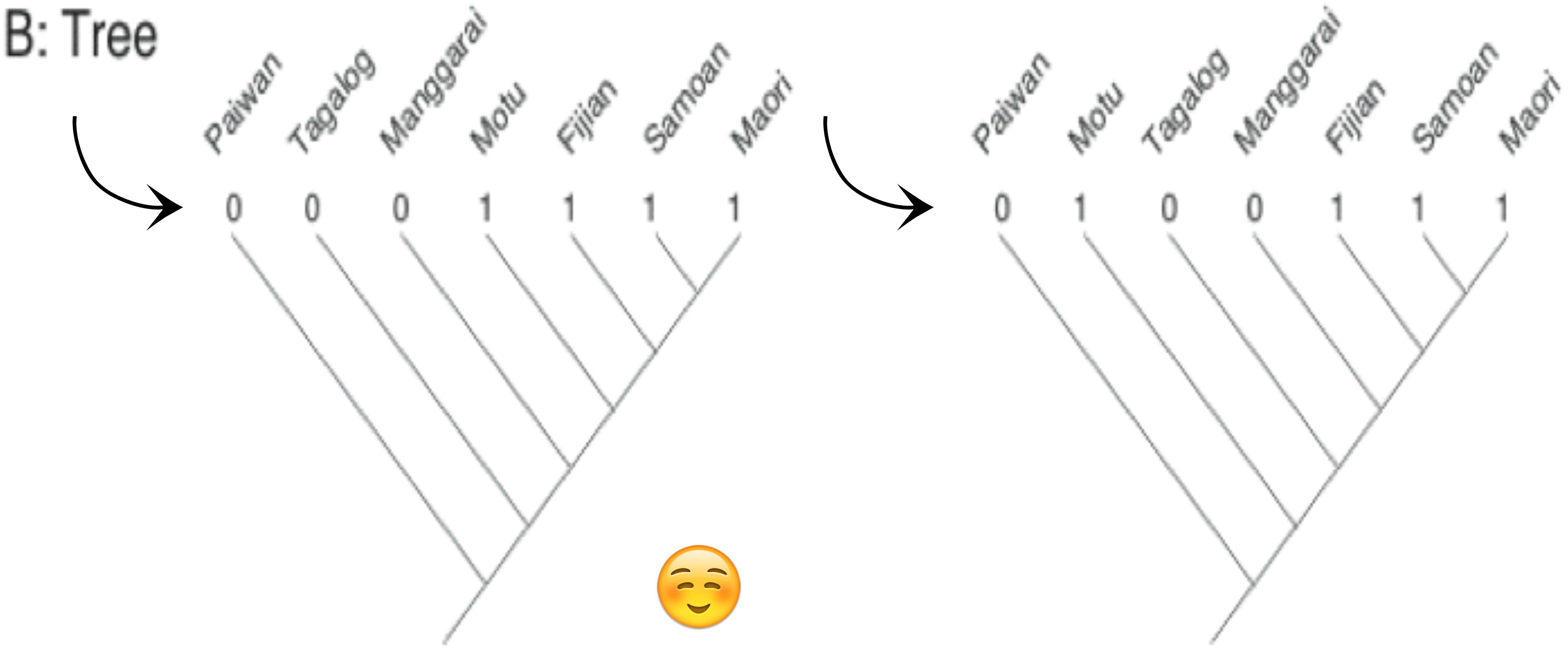


A: Matrix

	a	b	c	d	e	f	g	h	l	j	...
Paiwan	0	0	1	0	1	1	0	0	0	0	...
Tagalog	0	1	1	0	0	1	1	0	1	0	...
Manggarai	0	0	1	1	0	0	0	1	0	0	...
Motu	1	1	1	0	1	0	0	0	0	0	...
Fijian	1	1	1	1	1	0	0	0	0	0	...
Samoan	1	0	1	1	0	0	0	0	0	1	...
Maori	1	0	1	0	1	0	0	0	0	0	...

from Greenhill & Gray 2009

Calculating the site likelihood
θ_1 = trees (some hypotheses about history)



☺️

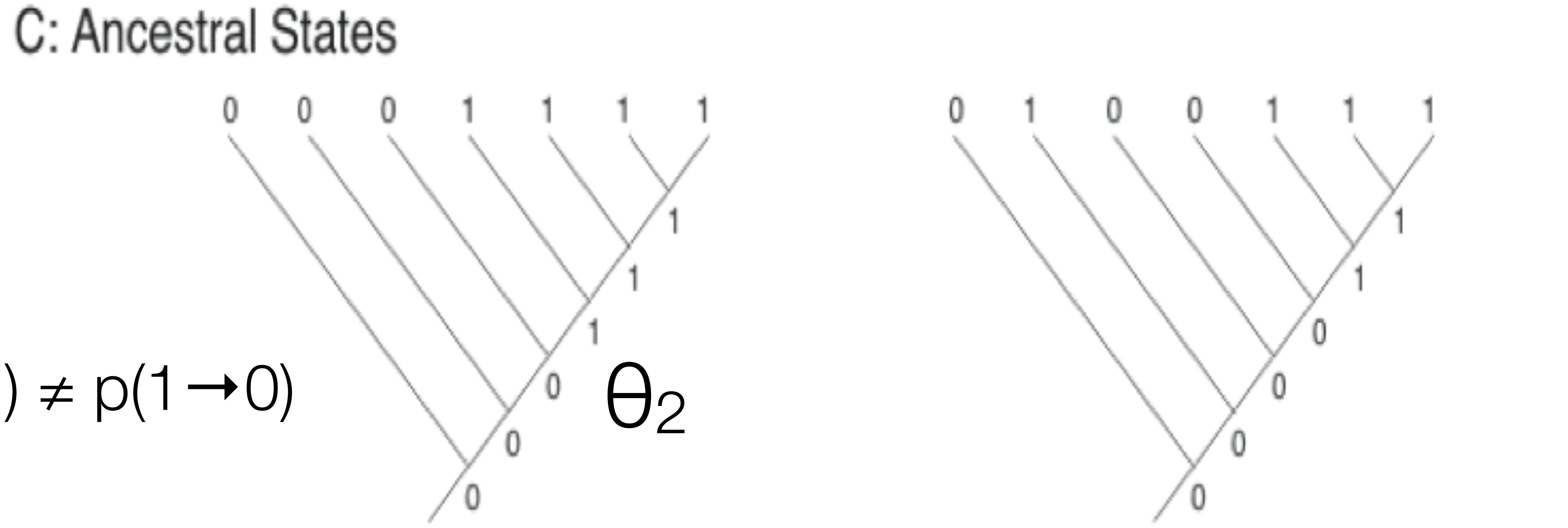
from Greenhill & Gray 2009

Calculating the site likelihood
θ_2 = node states and θ_3 = model of change

We combine a model of character change with some node values

θ_3

$p(0 \rightarrow 1) \neq p(1 \rightarrow 0)$



θ_2

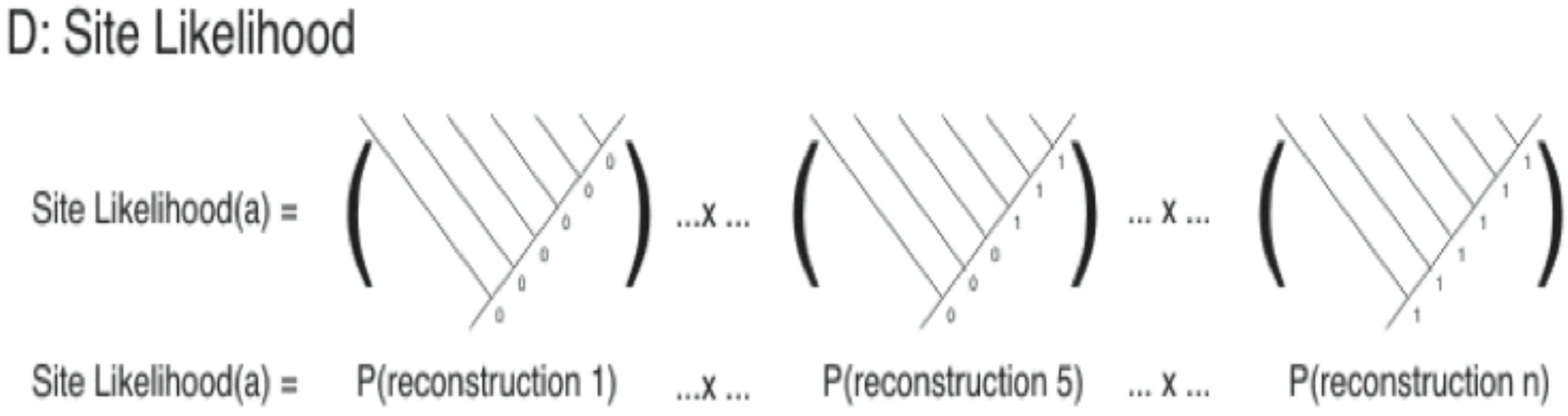
$L(a) = P(0 \rightarrow 0|b1) \times P(0 \rightarrow 0|b2) \times P(0 \rightarrow 0|b3) \times P(0 \rightarrow 0|b4) \times P(0 \rightarrow 0|b5) \times P(0 \rightarrow 1|b6) \times P(1 \rightarrow 1|b7) \times P(1 \rightarrow 1|b8) \times P(1 \rightarrow 1|b9) \times P(1 \rightarrow 1|b10) \times P(1 \rightarrow 1|b11) \times P(1 \rightarrow 1|b12)$

$L(a) = P(0 \rightarrow 0|b1) \times P(0 \rightarrow 0|b2) \times P(1 \rightarrow 1|b3) \times P(1 \rightarrow 0|b4) \times P(0 \rightarrow 0|b5) \times P(0 \rightarrow 0|b6) \times P(0 \rightarrow 0|b7) \times P(0 \rightarrow 1|b8) \times P(1 \rightarrow 1|b9) \times P(1 \rightarrow 1|b10) \times P(1 \rightarrow 1|b11) \times P(1 \rightarrow 1|b12)$

IN WORDS: the likelihood of character (a) is equal to the probability of (a) staying in state 0 along branch 1, multiplied by the probability of (a) staying in state 0 along branch 2 (...), multiplied by the probability of (a) changing to state 1 on branch 6, multiplied by the probability of it staying in state 0 on branch 7 ...

from Greenhill & Gray 2009

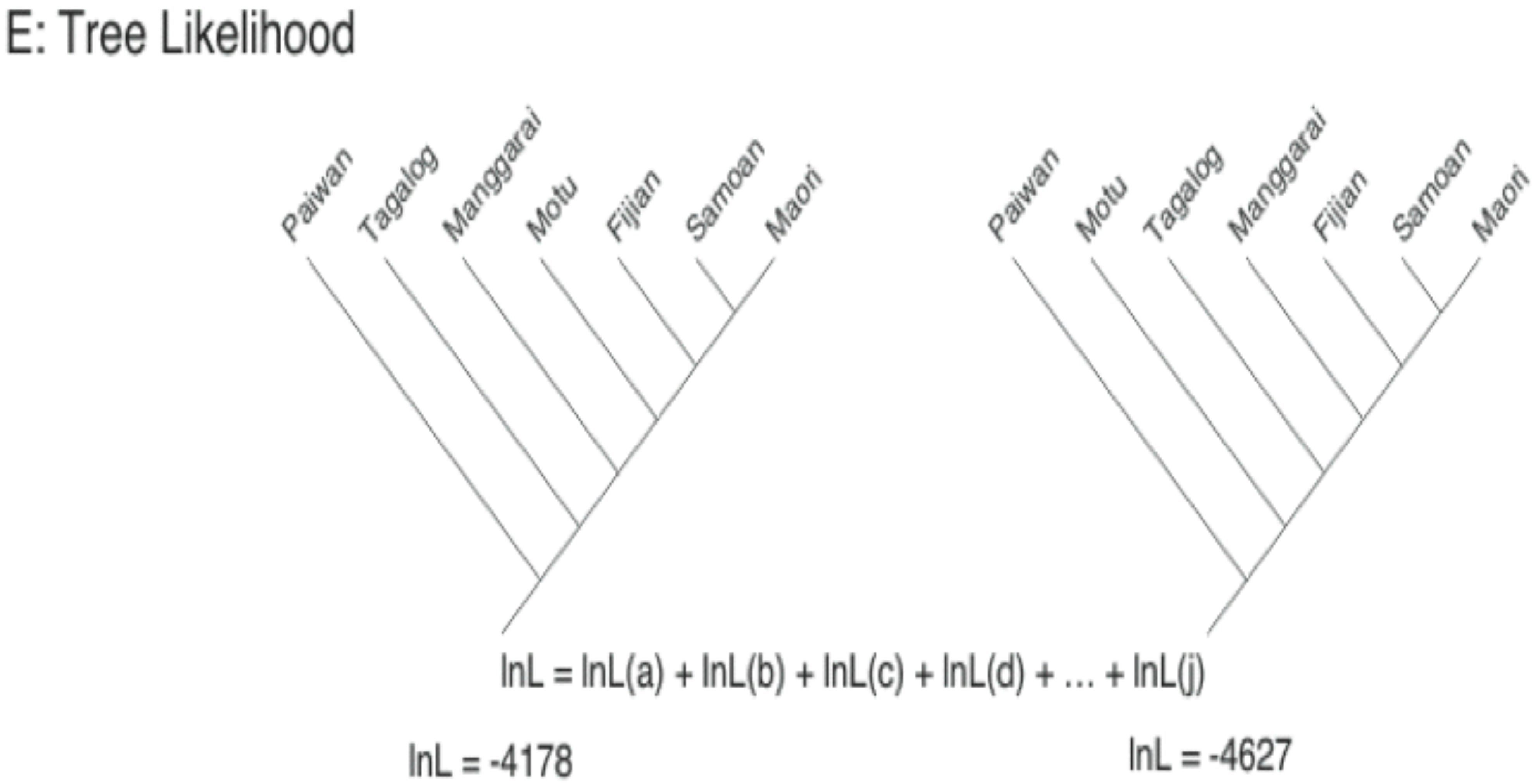
Calculating the site likelihood
Lh = node values over all possible values ($\Pi\theta_2$)



... multiplied over all the possible node values for the character.

from Greenhill & Gray 2009

Calculating the tree likelihood
Lh = Σ site likelihoods


E: Tree Likelihood

$$\ln L = \ln L(a) + \ln L(b) + \ln L(c) + \ln L(d) + \ldots + \ln L(j)$$

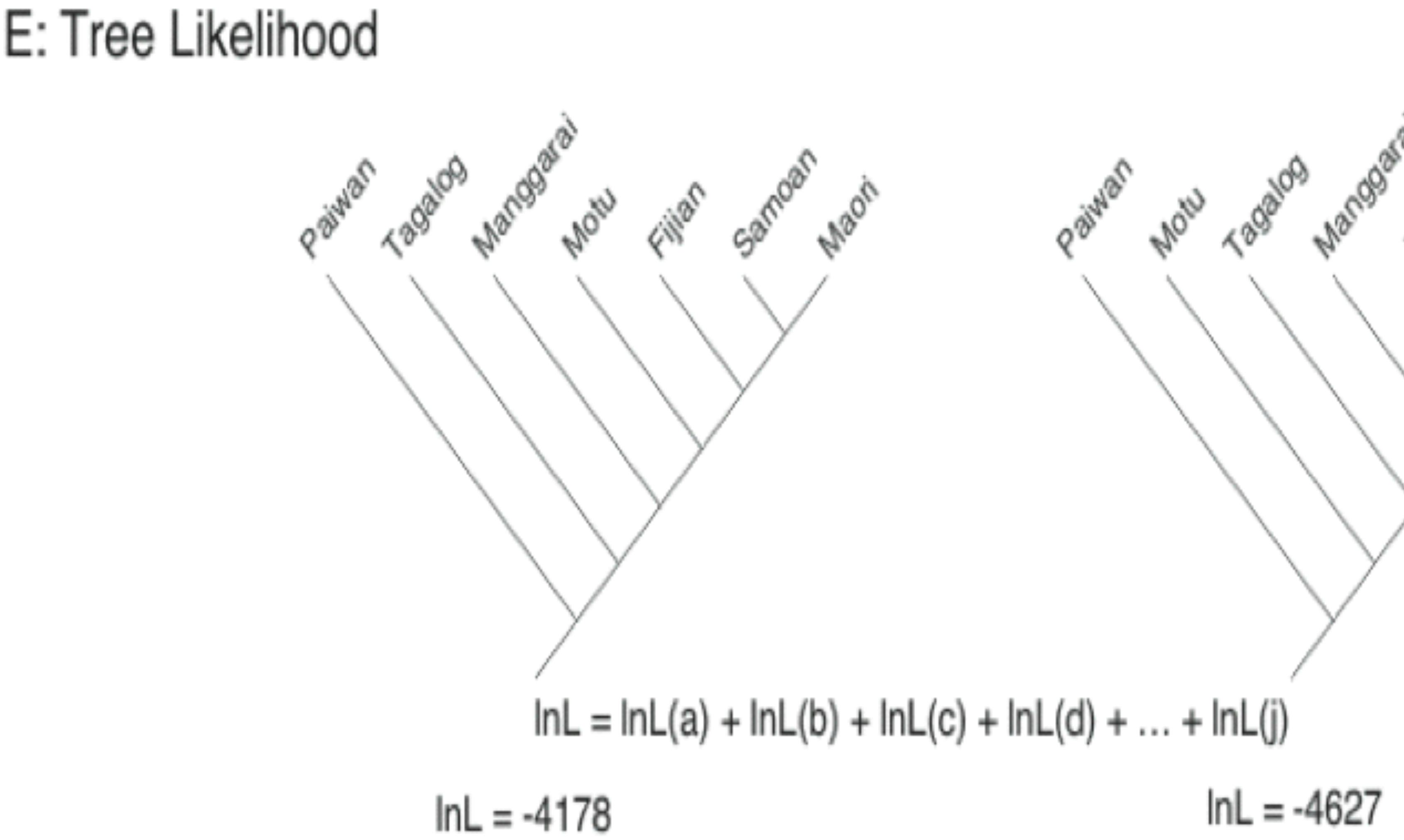
$\ln L = -4178$

$\ln L = -4627$

IN WORDS: The likelihood of the tree is the sum of the likelihood of each site on that tree.

from Greenhill & Gray 2009

Calculating the maximum likelihood
Find the tree with the best Lh



The higher the likelihood (closer to zero) the more we prefer a tree. We retain that tree and tweak it in an iterative fashion.

We infer character state changes, so innovations and retentions are incorporated in the model.

from Greenhill & Gray 2009

Finding the maximum likelihood is computationally expensive

the good and the not-so-good of maximum likelihood

- desirable statistical properties
- explicit expression of model
- model-testing framework (LRT)
- lots of data > converge on MLE

- non-intuitive
- computationally expensive
- how do we integrate over trees?
- how to account for uncertainty

Huelsenbeck & Ronquist 2009 in *The Phylogenetic Handbook*

the Bayesian approach

- explicit expression of model
- model-testing framework
- retains advantages of ML
- computationally efficient
- can integrate over trees
- can account for uncertainty
- best for linguistic/cultural data

- intuitive statistical reasoning?

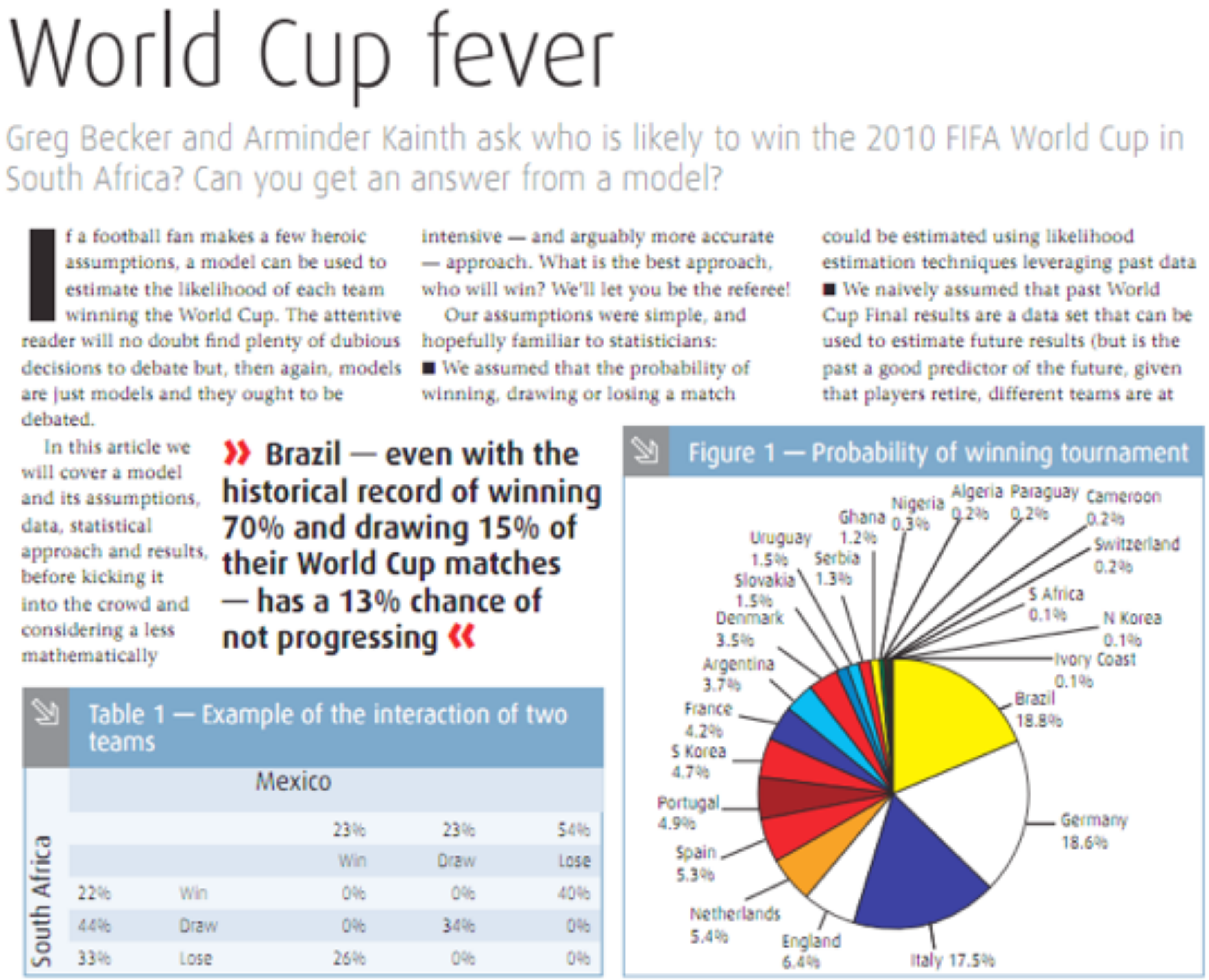
Huelsenbeck & Ronquist 2009 / Ronquist 2004

the Bayesian outlook

WORLD CUP WINNERS
1930
HOSTED IN URUGUAY
FINAL MATCH
URUGUAY - 4 vs. ARGENTINA - 2
1934
ITALY - 2 vs. CZECHOSLOVAKIA - 1
HOSTED IN ITALY
1938
HOSTED IN FRANCE
ITALY - 4 vs. FRANCE - 2
1950
URUGUAY* vs. BRAZIL
HOSTED IN BRAZIL
1954
HOSTED IN SWITZERLAND
WEST GERMANY - 3 vs. SWITZERLAND - 2
1958
BRAZIL - 5 vs. SWEDEN - 2
HOSTED IN SWEDEN
1962
HOSTED IN CHILE
BRAZIL - 3 vs. CHILE - 1
1966
ENGLAND - 4 vs. GERMANY - 2
HOSTED IN ENGLAND
1970
HOSTED IN MEXICO
BRAZIL - 4 vs. ITALY - 1
1974
WEST GERMANY - 2 vs. NETHERLANDS - 1
HOSTED IN WEST GERMANY
1978
HOSTED IN ARGENTINA
ARGENTINA - 3 vs. NETHERLANDS - 1
1982
ITALY - 3 vs. WEST GERMANY - 1
HOSTED IN SPAIN
1986
HOSTED IN MEXICO
ARGENTINA - 3 vs. WEST GERMANY - 2
1990
GERMANY - 1 vs. ARGENTINA - 0
HOSTED IN ITALY
1994
HOSTED IN UNITED STATES
BRAZIL* - 0 vs. ITALY - 0
1998
FRANCE - 3 vs. BRAZIL - 0
HOSTED IN FRANCE
2002
HOSTED IN KOREA REPUBLIC
BRAZIL - 2 vs. GERMANY - 0
2006
ITALY - 1 vs. FRANCE - 1
HOSTED IN GERMANY
2010
HOSTED IN SOUTH AFRICA
?
TBD vs. TBD
WORLD CUP

*For 1994, the winner was determined by a shoot out, as the game ended in a tie at the end of extra time.

* There was no final game for this World Cup, rather the winner was determined by a final round-robin group contested by four teams (Uruguay, Brazil, Sweden, and Spain); however, Uruguay's 2-1 victory over Brazil is considered the decisive match in the final round.



World Cup fever

Greg Becker and Arminder Kainth ask who is likely to win the 2010 FIFA World Cup in South Africa? Can you get an answer from a model?

If a football fan makes a few heroic assumptions, a model can be used to estimate the likelihood of each team winning the World Cup. The attentive reader will no doubt find plenty of dubious decisions to debate but, then again, models are just models and they ought to be debated.

In this article we will cover a model and its assumptions, data, statistical approach and results, before kicking it into the crowd and considering a less mathematically intensive — and arguably more accurate — approach. What is the best approach, who will win? We'll let you be the referee!

Our assumptions were simple, and hopefully familiar to statisticians:

- We assumed that the probability of winning, drawing or losing a match could be estimated using likelihood estimation techniques leveraging past data
- We naively assumed that past World Cup Final results are a data set that can be used to estimate future results (but is the past a good predictor of the future, given that players retire, different teams are at

» Brazil — even with the historical record of winning 70% and drawing 15% of their World Cup matches — has a 13% chance of not progressing «

Table 1 — Example of the interaction of two teams

		Mexico		
		23%	23%	54%
		Win	Draw	Lose
South Africa 22%	Win	0%	0%	40%
44%	Draw	0%	34%	0%
33%	Lose	26%	0%	0%

Figure 1 — Probability of winning tournament

reasoning about probability and updating our reasoning as new information becomes available

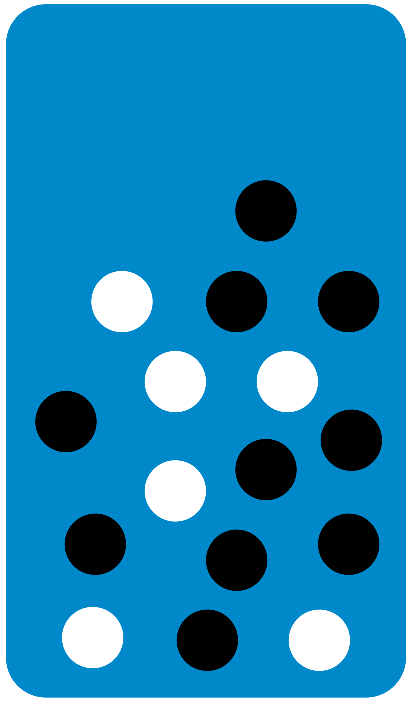
talking in maths*

P	probability
t	time
Σ	sum
θ	"model"
q	"rate"
X	data
$\int$	integral
$\mid$	conditional

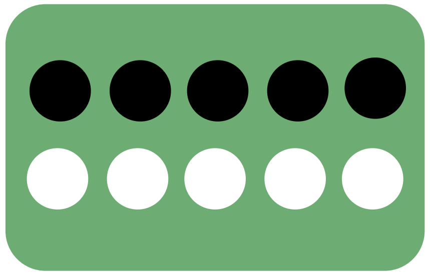
*apologies to Thom Yorke

probability

“forward” probability



M = 10 black, 5 white



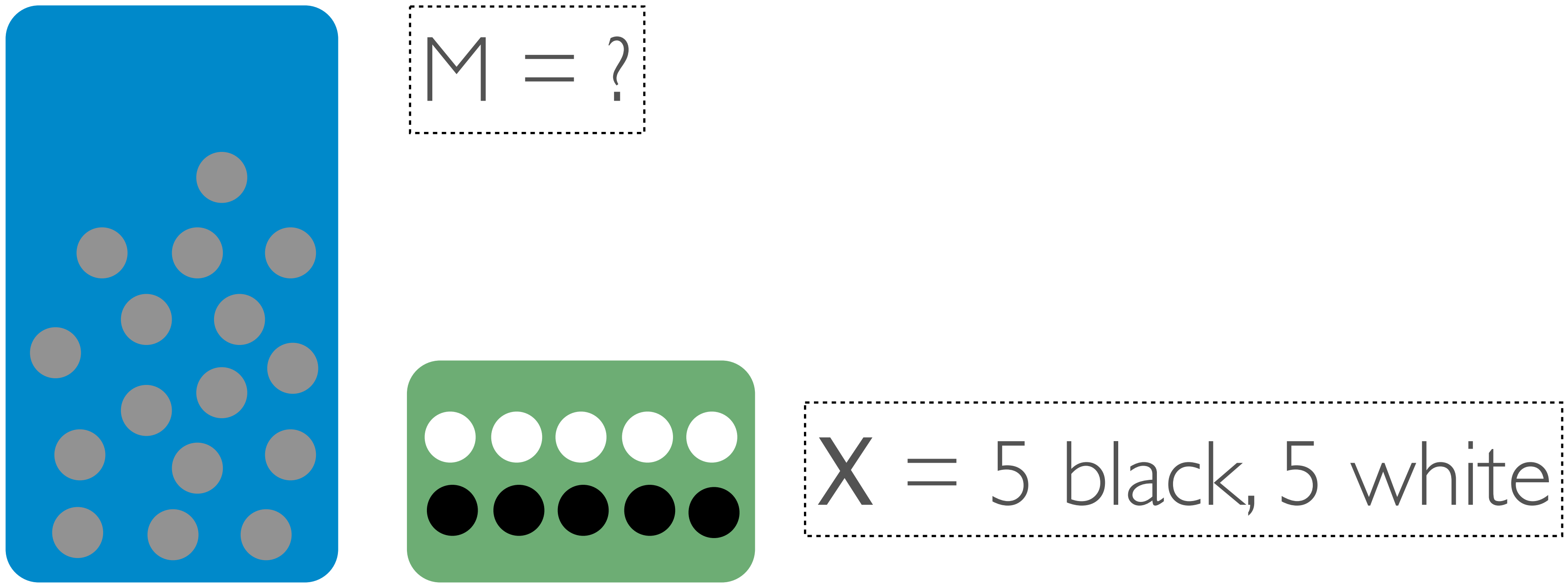
X = 5 black then 5 white

$$p(\mathbf{X}|M) = 10/15 * 9/15 * \ldots * 1/6 = 0.00033$$

i.e. what is the probability of these balls (data) given this model (urn)?

probability

“converse” probability



M = ?

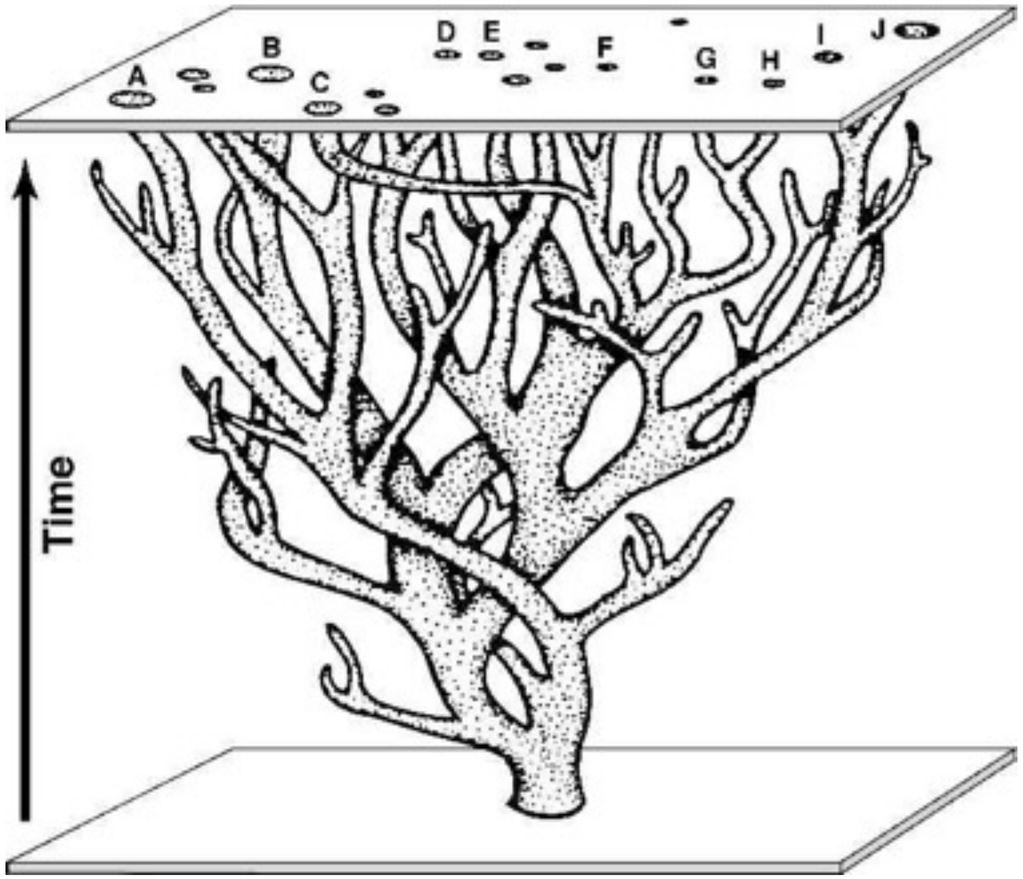
X = 5 black, 5 white

How do we find p(M|**X**) ???
Impossible unless we know something about M!

i.e. given these balls (data) what can we infer about what’s in the urn (model)?

probability

“converse” probability



The historical sciences regularly deal with this situation.

i.e. given these balls (data) what can we infer about what’s in the urn (model)?

1797

Bayes Theorem I

X are the Data
Θ the model Parameters

$$\underbrace{f(\theta \mid X)}_{\text{posterior}} = \frac{\overset{\text{prior}}{f(\theta)}\,\overset{\text{likelihood}}{f(X \mid \theta)}}{\underbrace{\int f(\theta)\,f(X \mid \theta)\,d\theta}_{\text{normalizing constant}}}$$

T. Bayes.

Take another deep breath

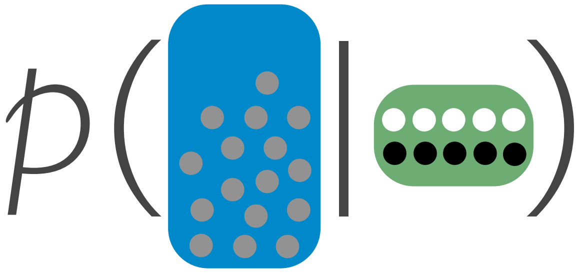
Bayes Theorem 2

conditional probability $$p(X|M) = \frac{p(XM)}{p(M)}$$

joint probability $$p(M) \times p(X|M) = p(XM)$$

similarly $$p(X) \times p(M|X) = p(XM)$$

remember the balls (X) and urn (M) situation:



Bayes Theorem 3

joint probability $p(M) \times p(X|M) = p(XM)$

similarly $p(X) \times p(M|X) = p(XM)$

$$p(M|X) = \frac{p(XM)}{p(X)}$$

$$p(M|X) = \frac{p(M) \times p(X|M)}{p(X)}$$

Bayes Theorem $$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{p(X)}$$

Bayes Theorem 4

Bayes Theorem

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{p(X)}$$

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{\sum \{(p(X|\theta_1)p(\theta_1)) + (p(X|\theta_2)p(\theta_2)) + \ldots\}}$$

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{\int p(X) \text{ over all } \theta}$$

$$p(\text{you}) = \text{argh!} = 1.0$$

Bayes Theorem 5

Bayes Theorem

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{p(X)}$$

the probability of the data given the hypothesis

the probability of the hypothesis

$$\text{posterior probability} = \frac{\text{likelihood of data} \times \text{prior}}{\text{marginal likelihood}}$$

the probability of the hypothesis, given the data

the unconditional (over all hypotheses) probability of the data

the troublesome marginal likelihood

calculating the Lh(site) is **analytically intractable**

for 100 taxa, there are 4.02 x 10^{59}
possible ancestral state configurations for one site

4.02 x 100,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000!

we need to **sample** this parameter space

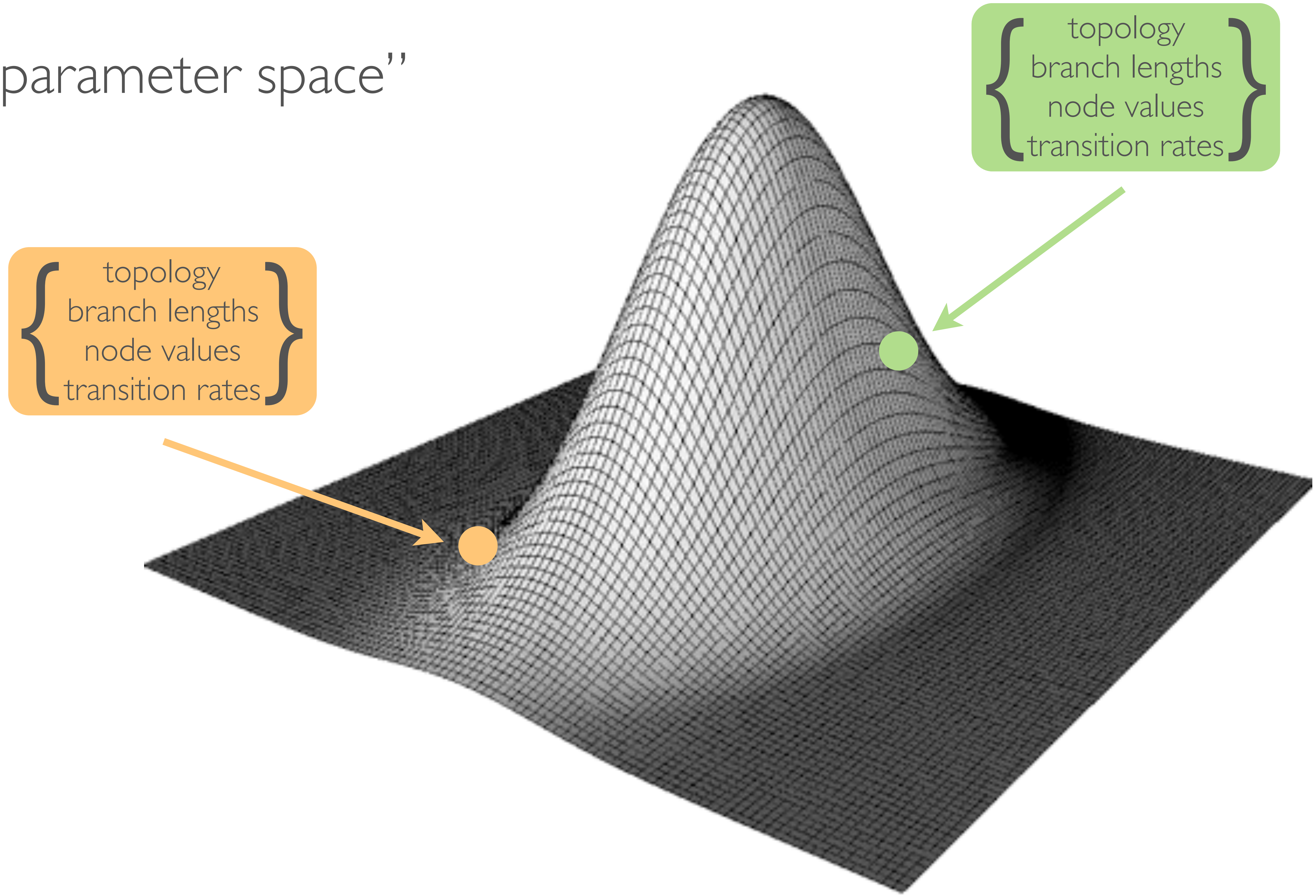
random sampling is wasteful (huge space, lots of low probability)

MCMC to the rescue!

Geyer 1992

the posterior probability distribution as a landscape

“parameter space”

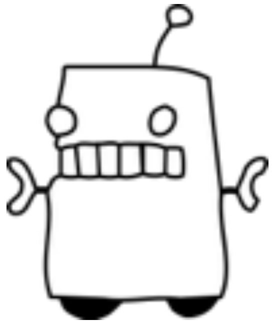


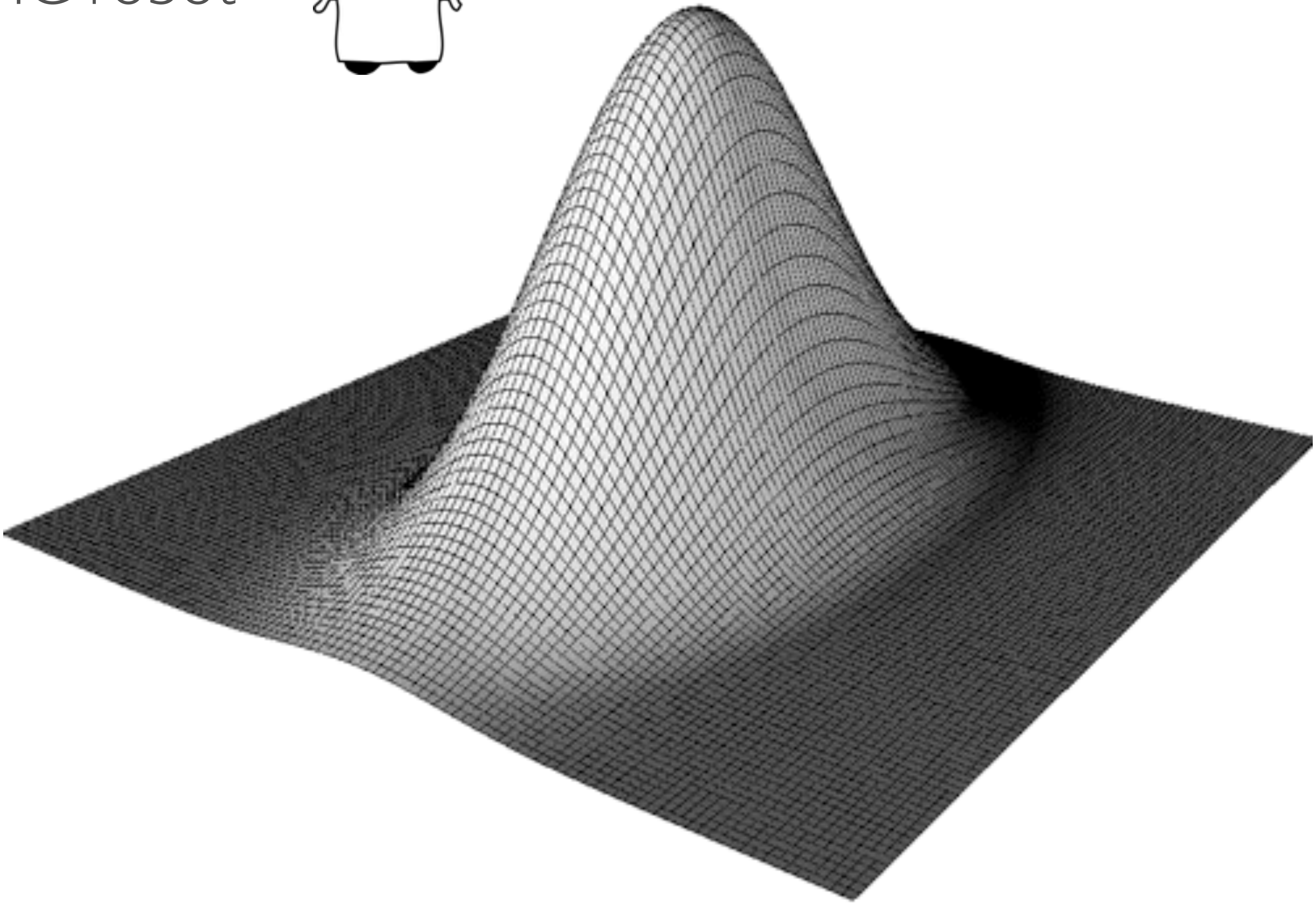
{ topology
branch lengths
node values
transition rates }

{ topology
branch lengths
node values
transition rates }

the posterior probability distribution as a landscape

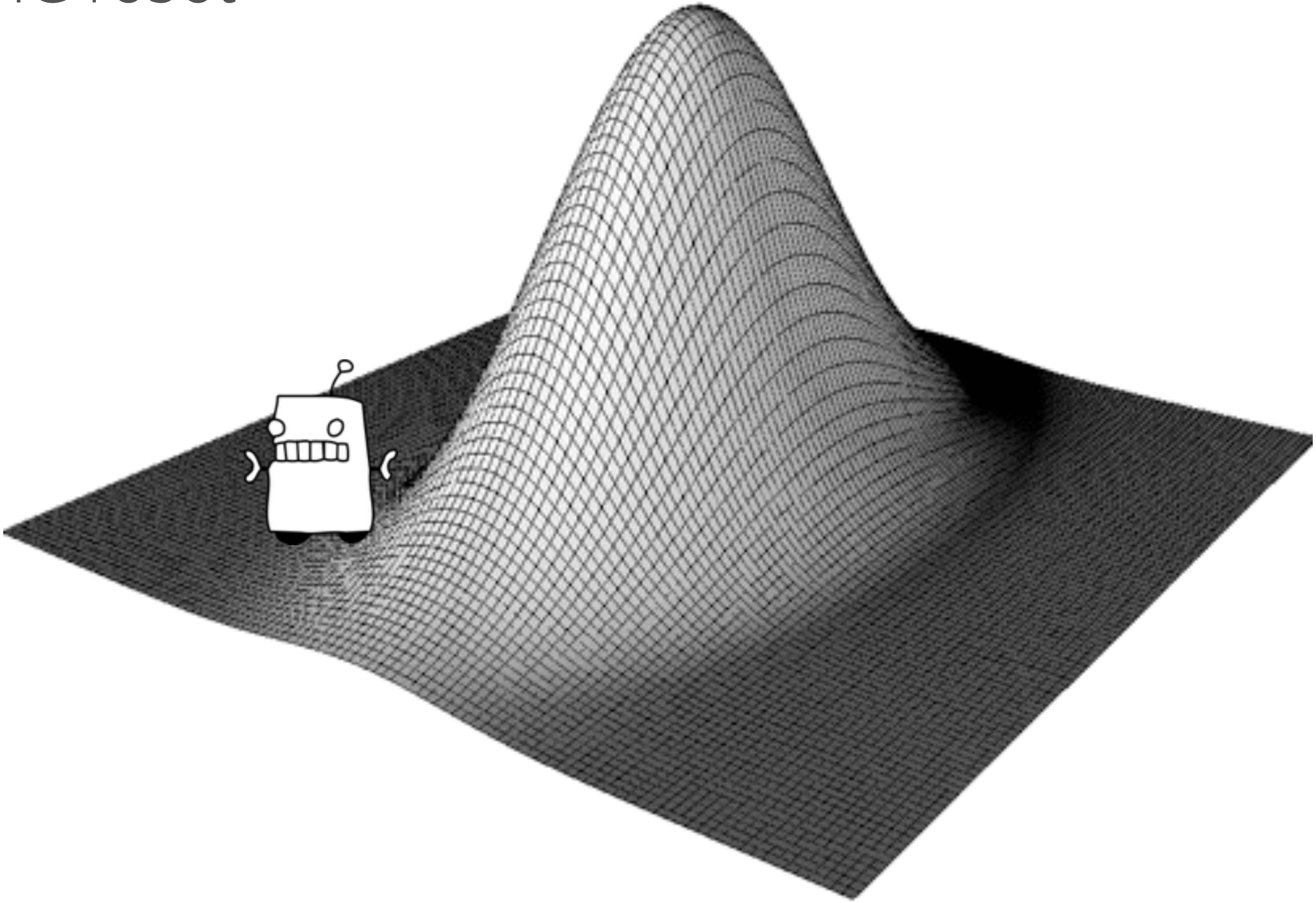
MCMC robot





the posterior probability distribution as a landscape

MCMC robot



Markov chain Monte Carlo sampler

parameter space

transition rates

Markov process

posterior probability distribution

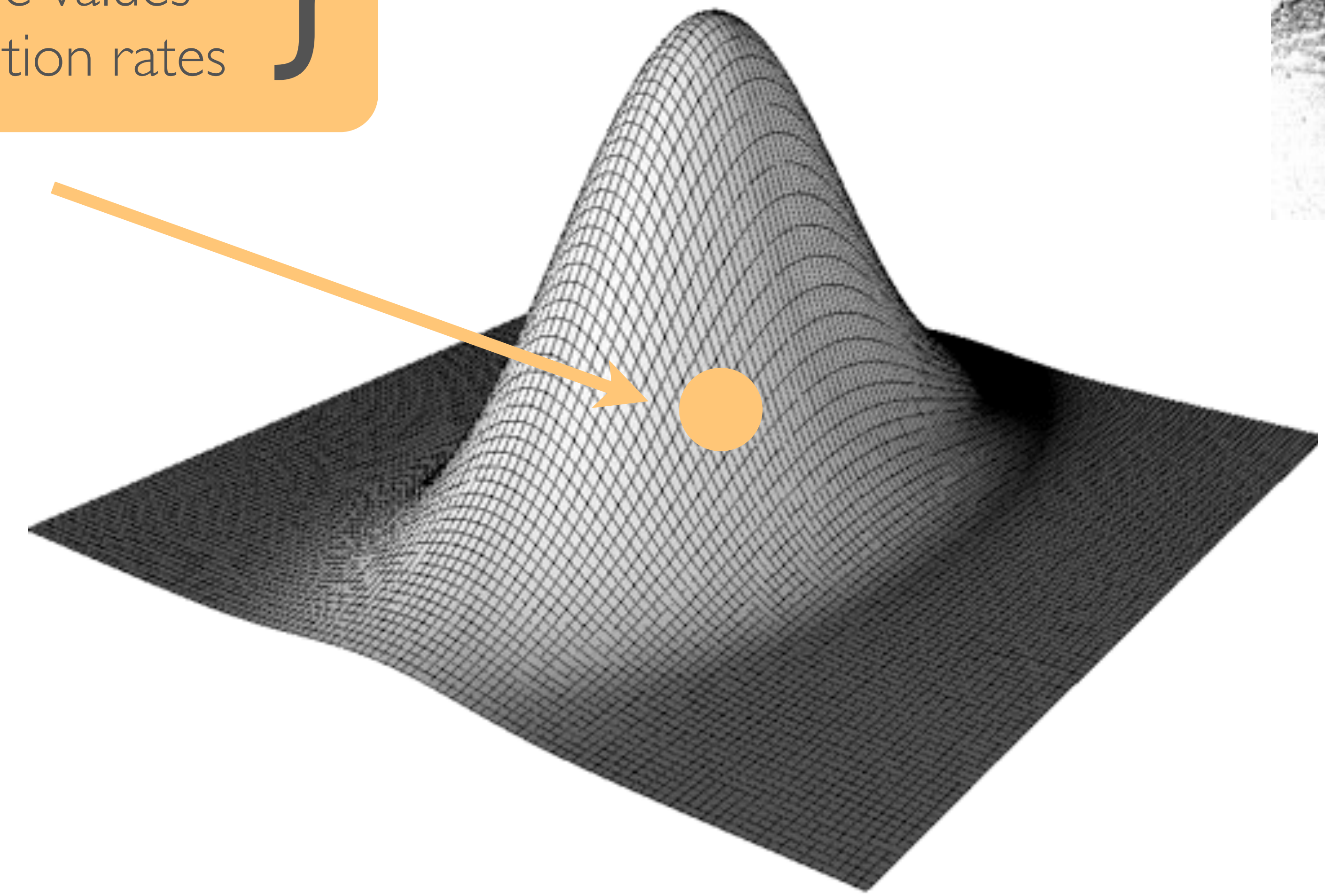
Monte Carlo:
repeated random sampling

parameters

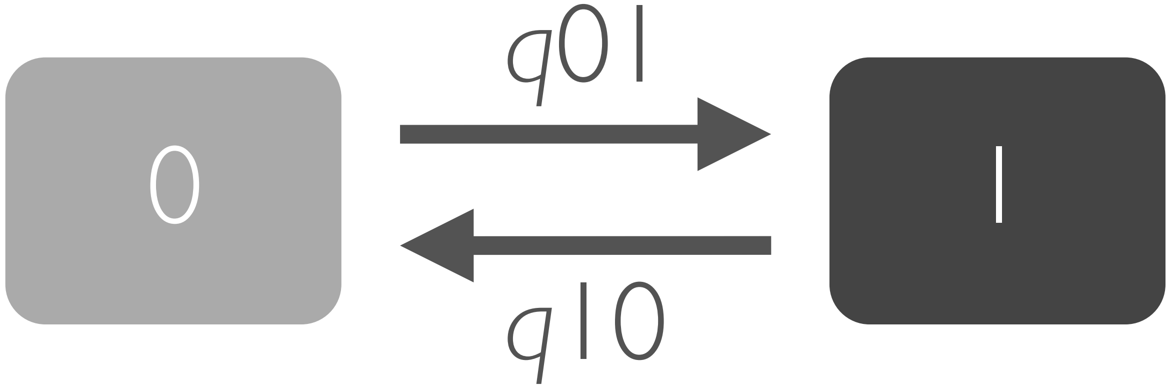
X are the Data
Θ the model Parameters

T. Bayes.

{ topology
branch lengths
node values
transition rates }

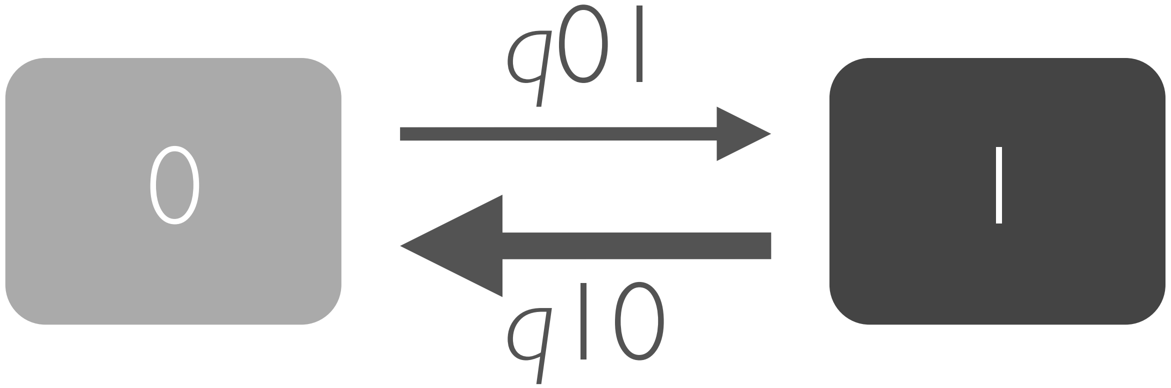


modelling change in discrete characters I



$q01 = q10$

a one-parameter model



$q01 \neq q10$

a two-parameter model

$$q01 = \alpha$$
$$q10 = \beta$$

we want to estimate these transition rate parameters!

Pagel 1994

Markov process

"A mathematical model of infrequent changes of discrete states over time, in which future events occur by chance and depend only on the current state, and not on the history of how that state was reached."

state at t_{t+1} only depends on state t

Markov chain
process has a finite state-space
(i.e. you can count it)

Whelan et al 2001

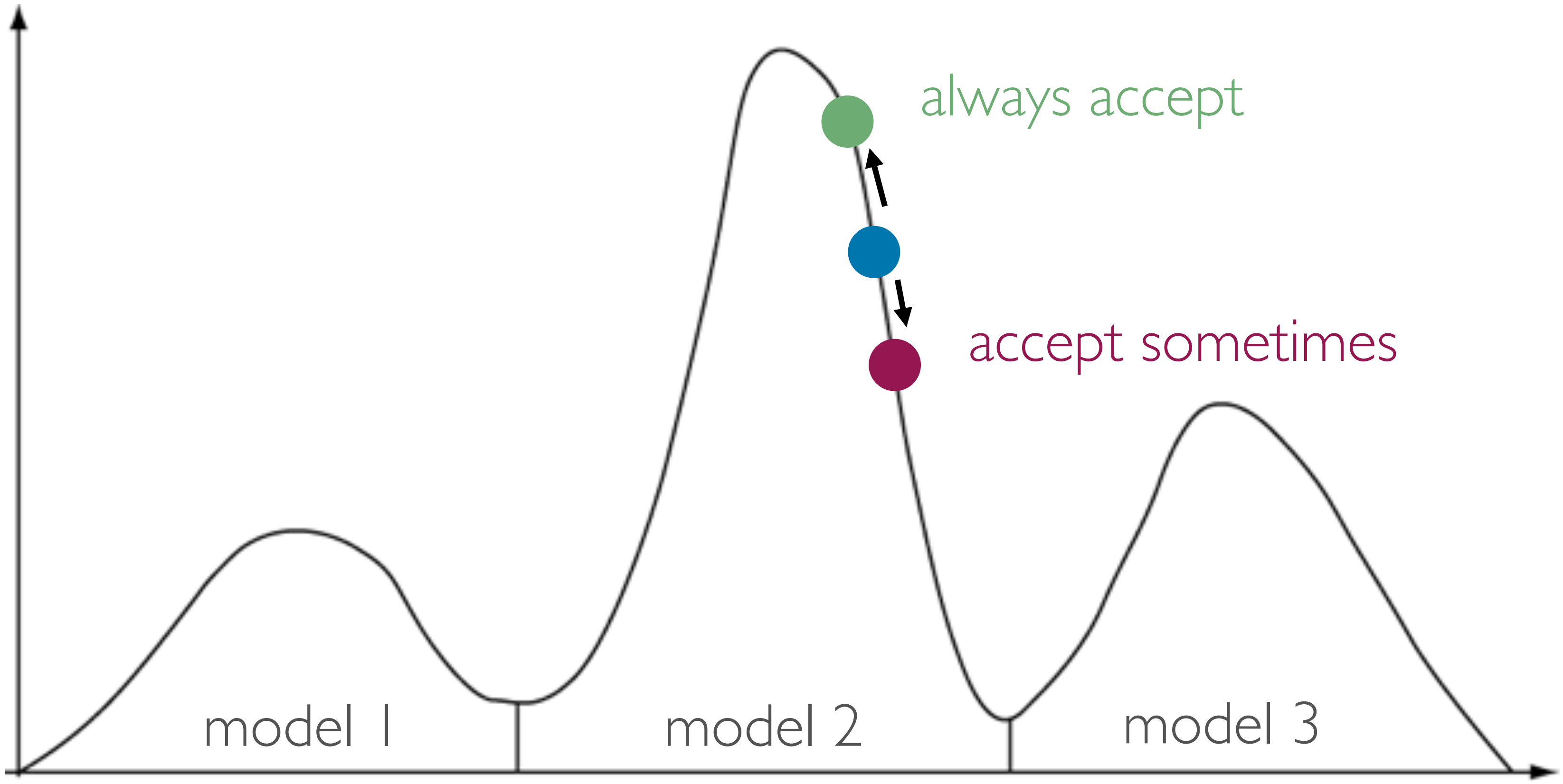
Markov chain Monte Carlo sampler I

[1] Start at an arbitrary point in parameter space

[2] Make a small random move in one parameter

[3] Calculate proposal ratio (R) of new state to old state:

- [a] if R > 1 then the new state is accepted
- [b] if R < 1 then the new state is accepted with probability R
- [c] If new state not accepted, stay in the old state



slide adapted from Frederick Ronquist

Markov chain Monte Carlo sampler 2

[3] Calculate proposal ratio (R) of new state to old state

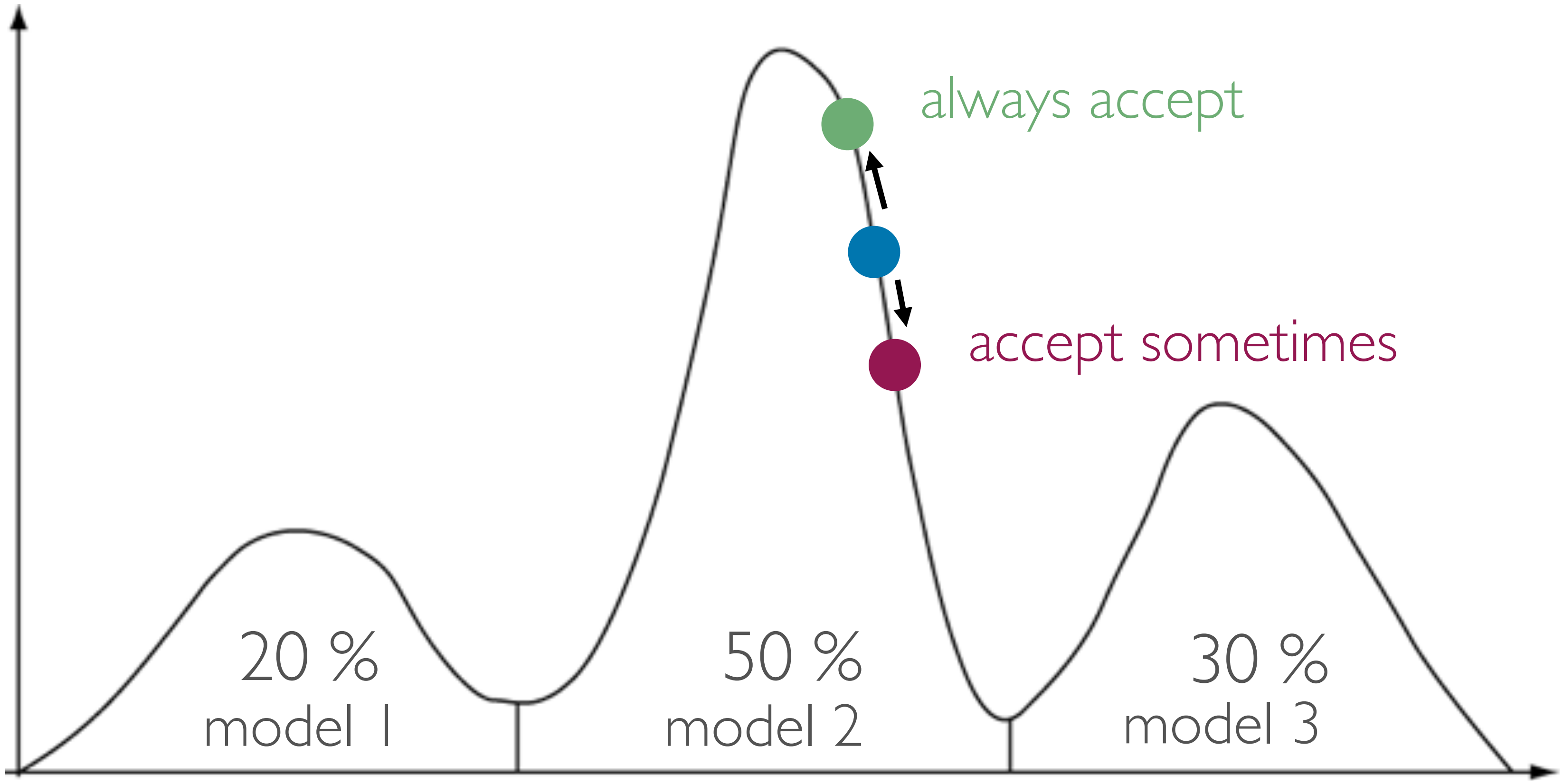
R = min[1, ~~Lh ratio~~ x ~~prior ratio~~ x proposal ratio]

[a] if R > 1 then the new state is accepted

[b] Generate random variable U[0,1]*

If R > U then the new state is accepted

[c] If new state not accepted, stay in the old state



* allows us to fully characterise the marginal distribution

slide adapted from Frederick Ronquist

Markov chain Monte Carlo sampler 3

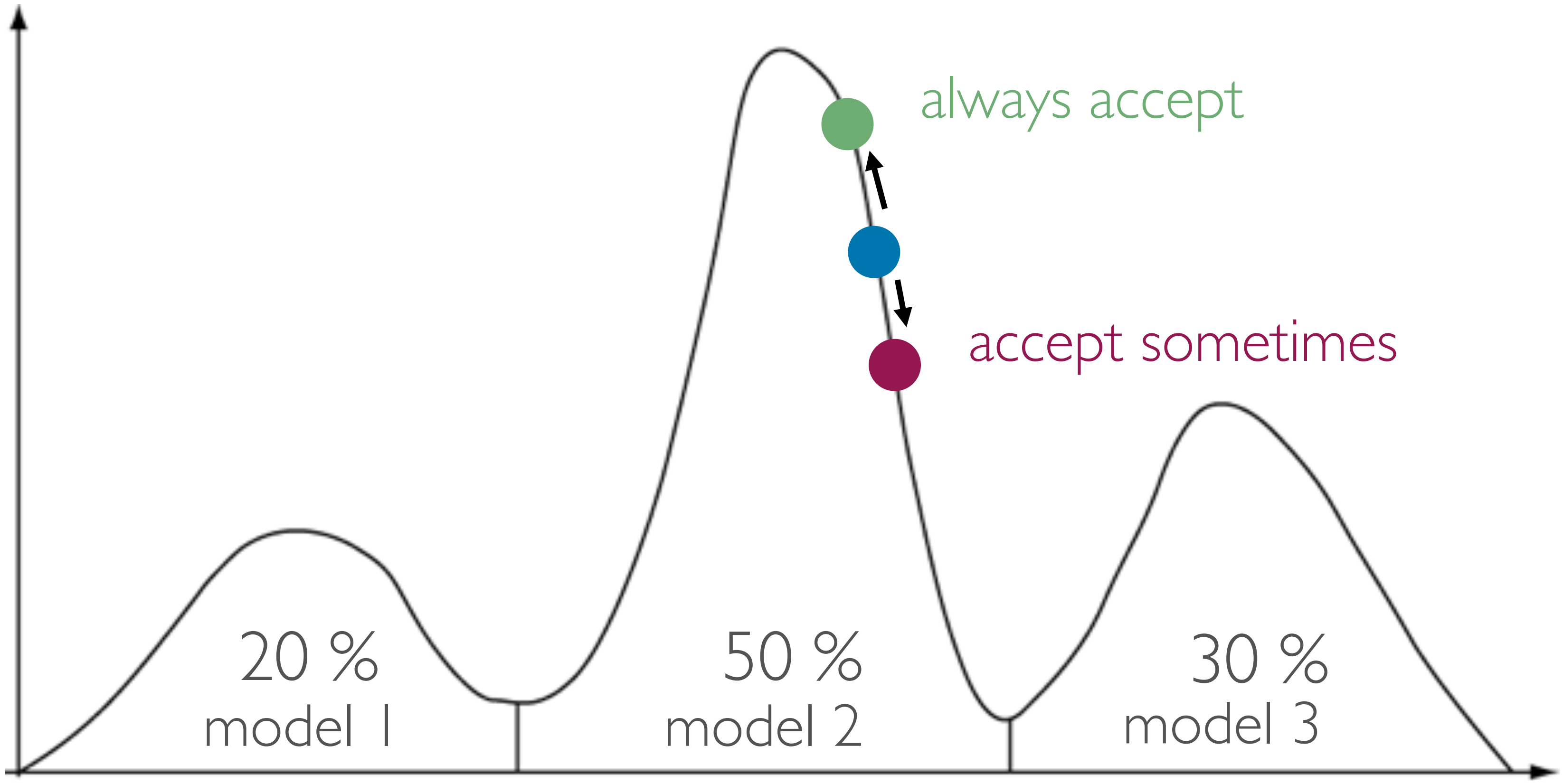
[1] Start at an arbitrary point in parameter space

[2] Make a small random move in one parameter

[3] Calculate proposal ratio (R) of new state to old state:

Accept or not based on R, U

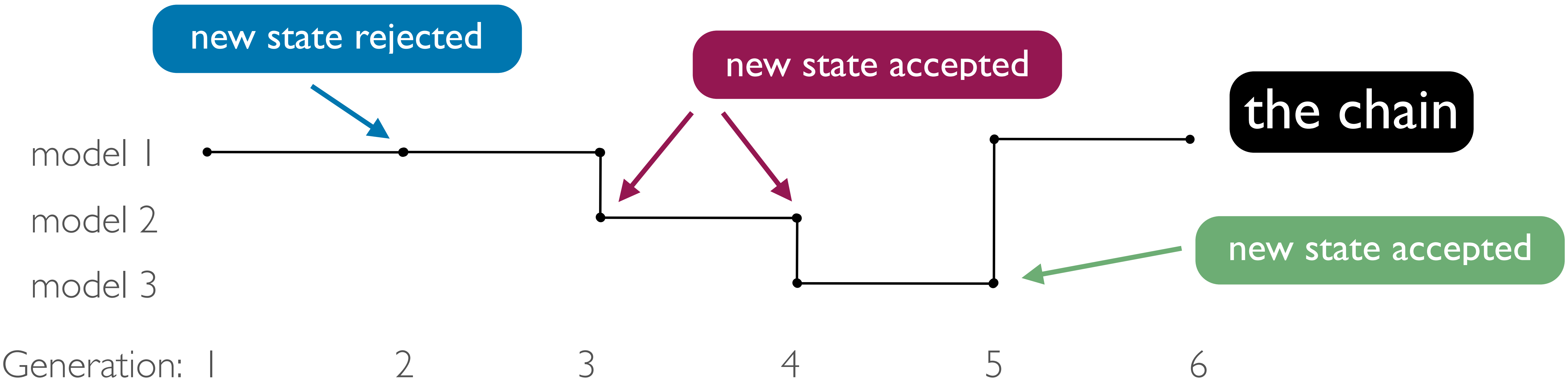
[4] Repeat many, many times: a Markov chain



The proportion of time the MCMC procedure samples from a particular parameter region is an estimate of that region's posterior probability density

slide adapted from Frederick Ronquist

Markov chain Monte Carlo sampler 4

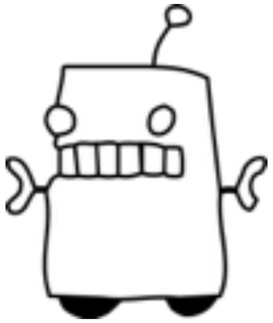


Bayesian Posterior Probability for model 1 ($BPP_{model\ 1}$)= 4/6

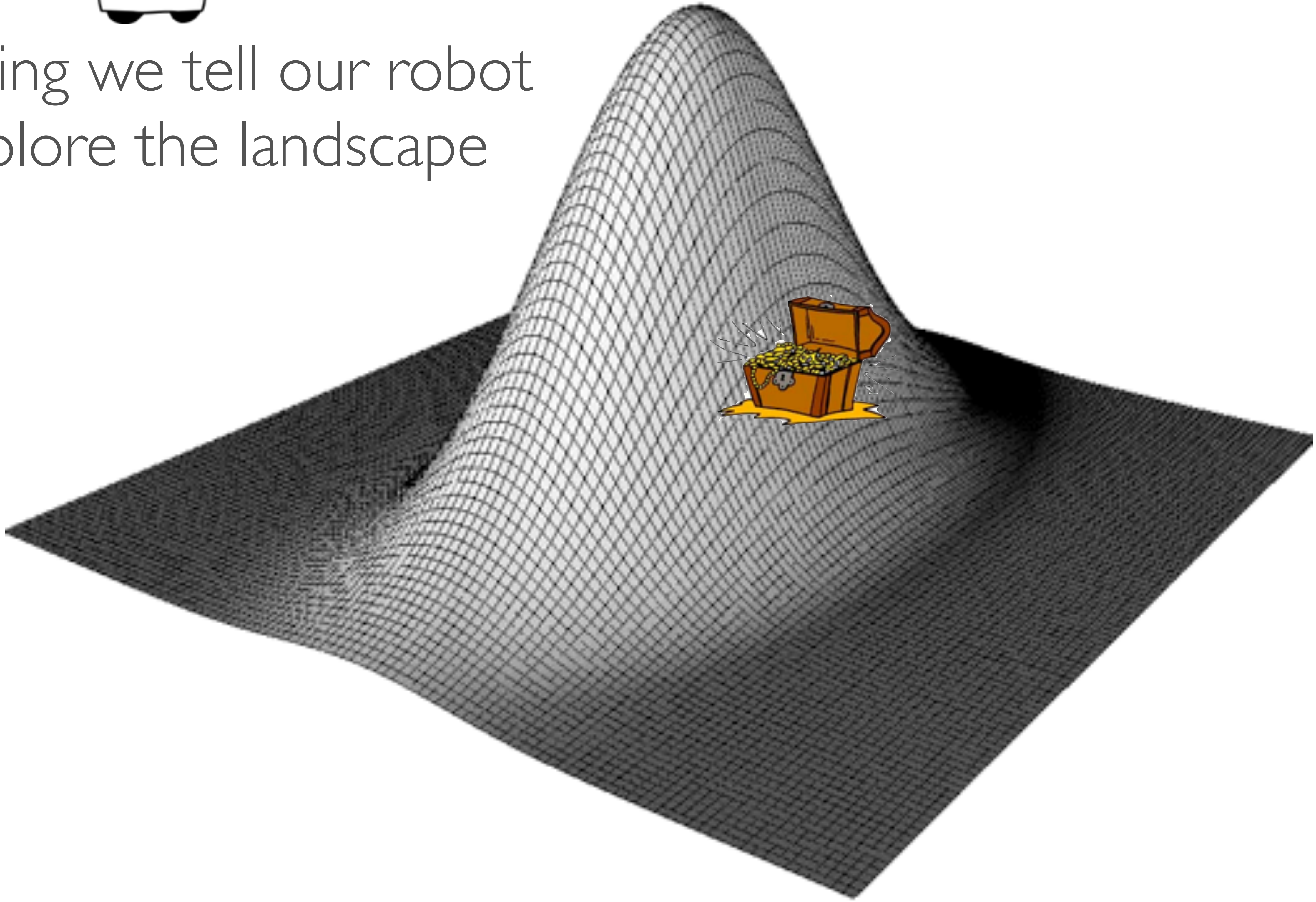
Sampling the MCMC provides a valid approximation for the posterior distribution of trees (over 100,000s – 1,000,000s of generations) without having to know the denominator

slide adapted from Frederick Ronquist

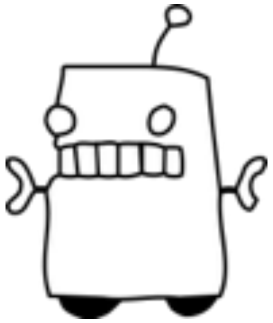
but what is a prior?



something we tell our robot to explore the landscape



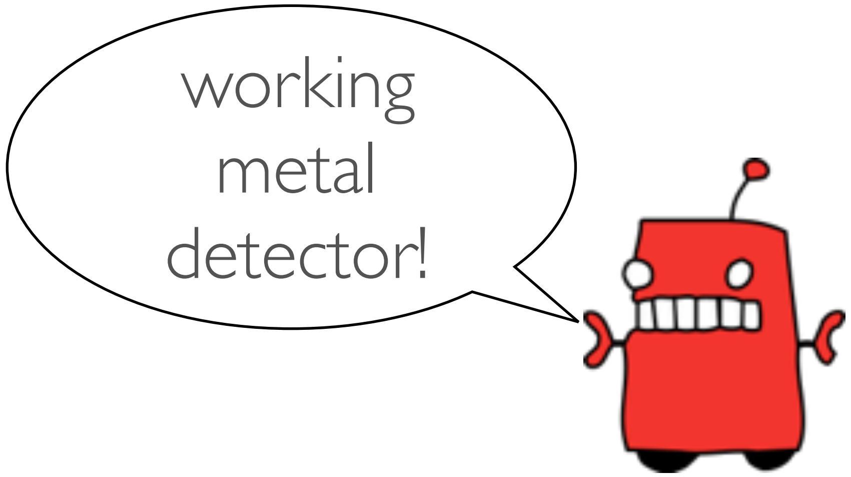
priors I

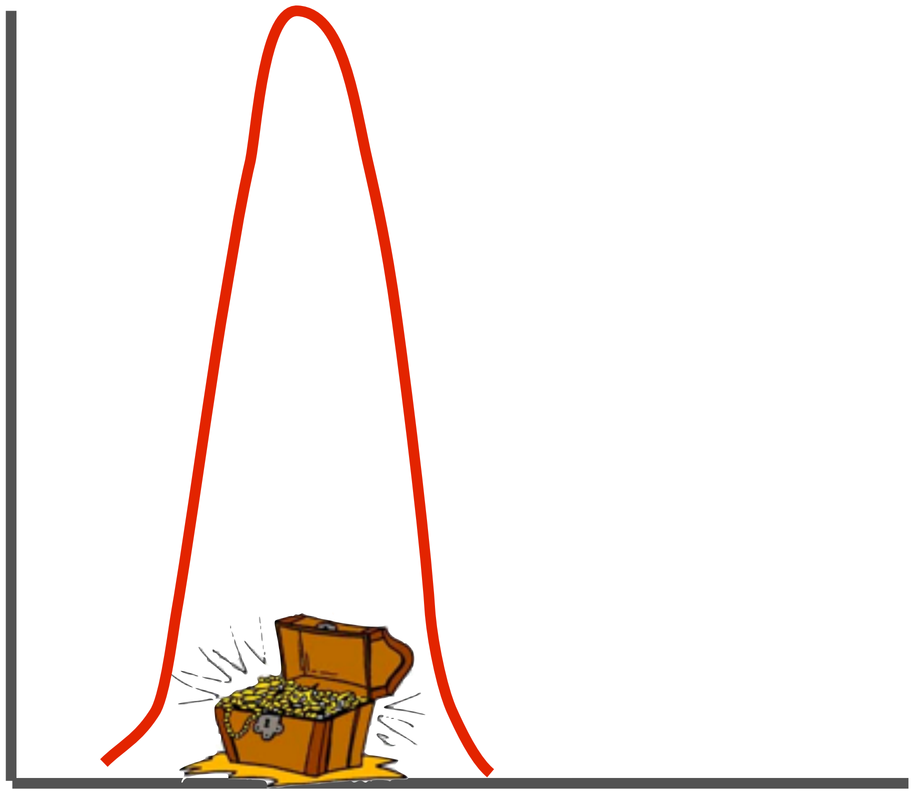


information formalised as a distribution

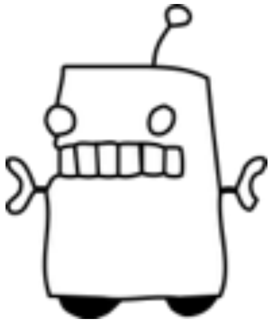
informative prior

we have some good information





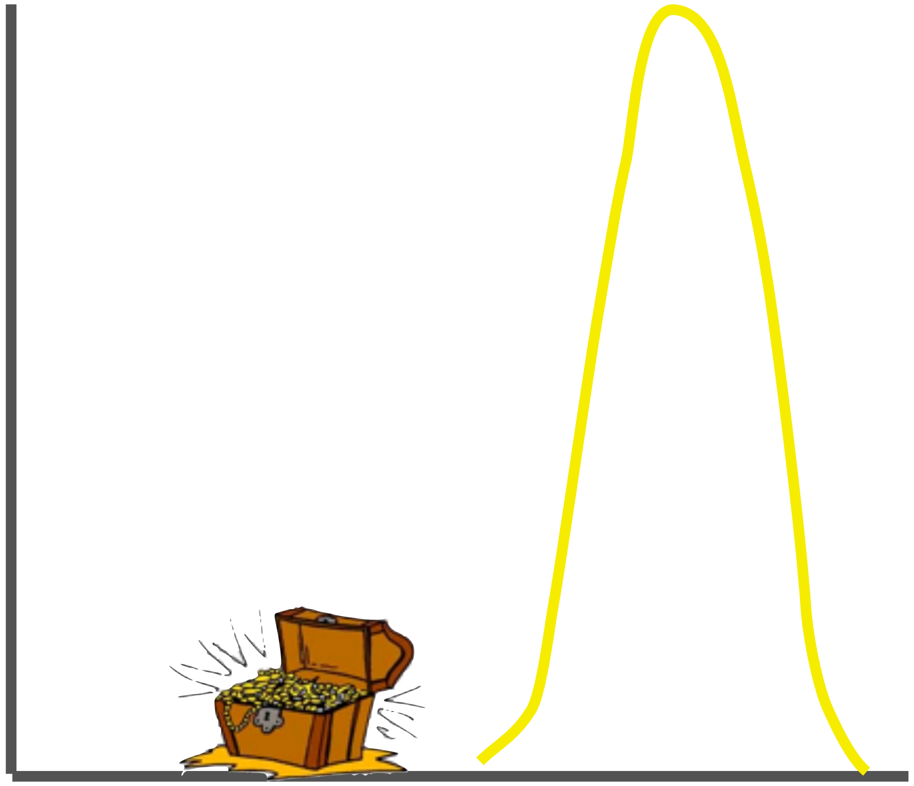
priors 2



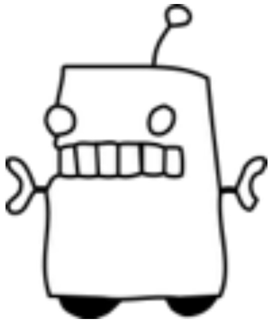
information formalised as a distribution

informative prior

we have some bad information



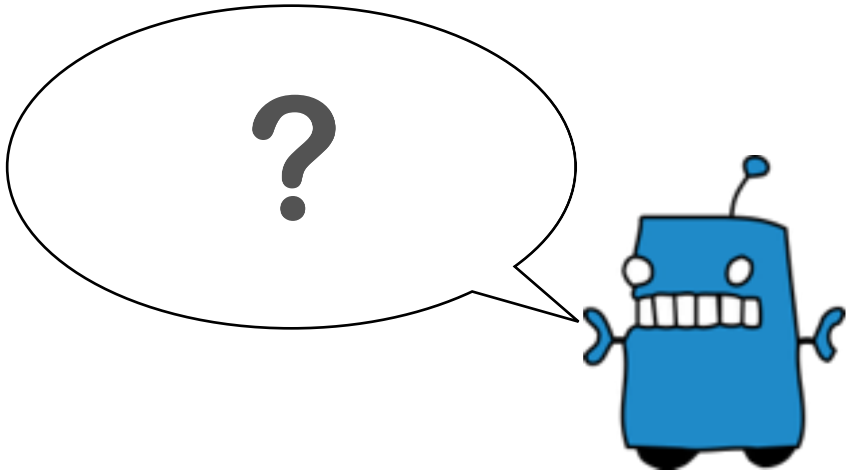
priors 3

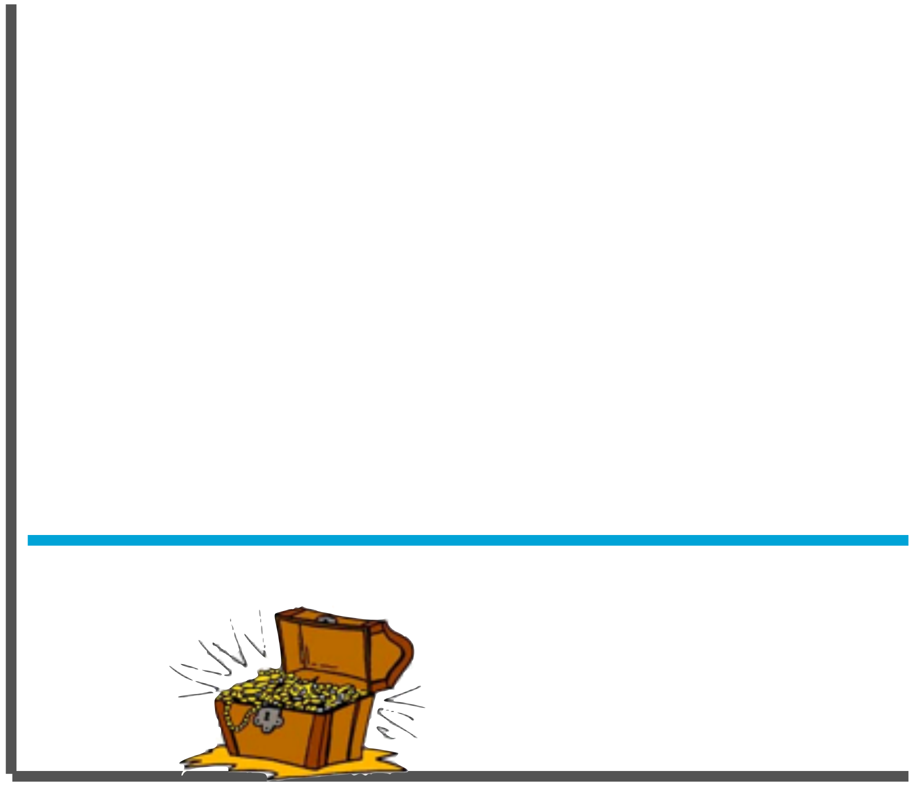


information formalised as a distribution

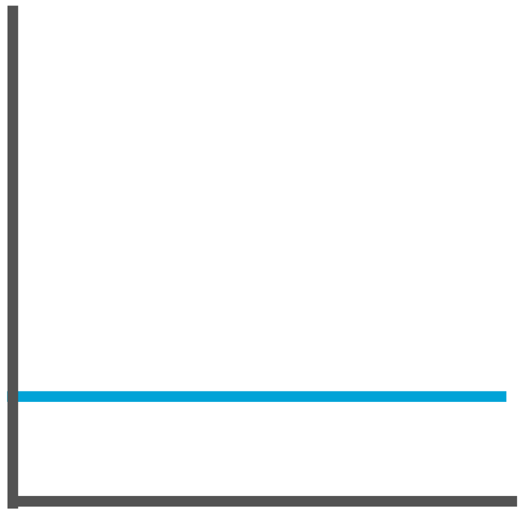
uniform prior

sometimes we don’t know anything

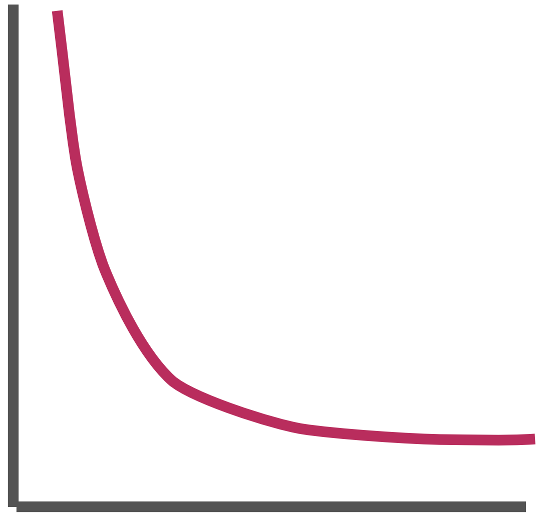




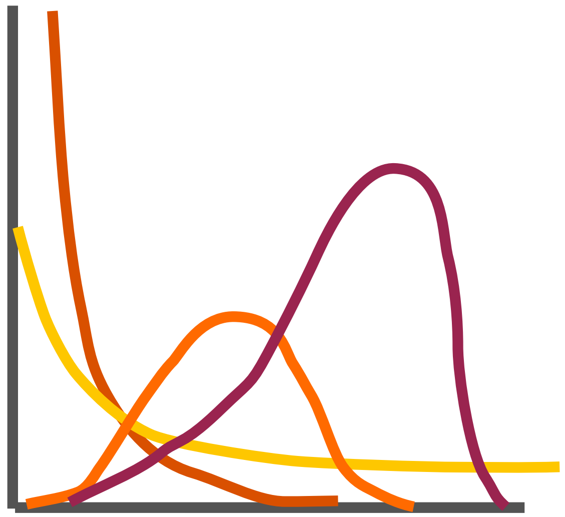
priors 4



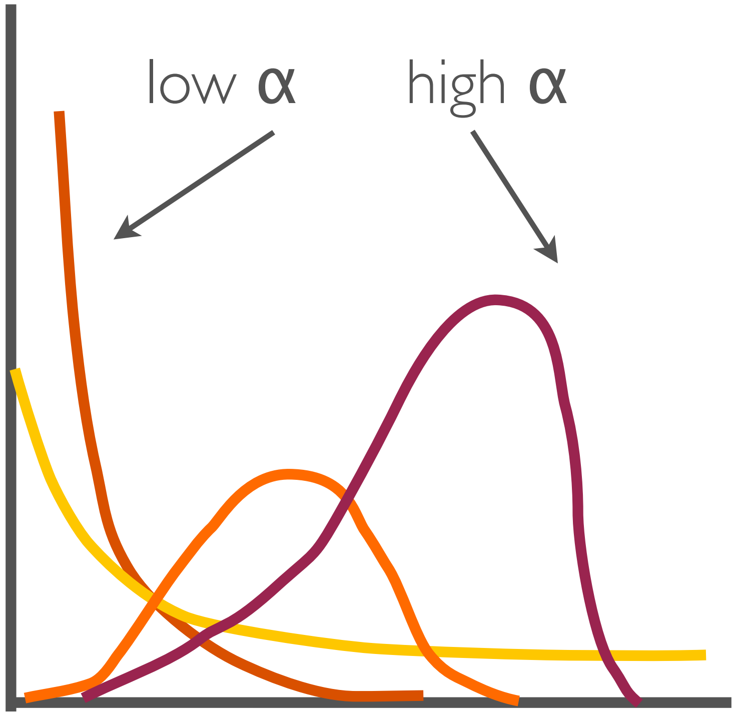
uniform



exponential



gamma



modelling change in discrete characters 2

Parameters can vary:

(1) across sites (characters)

(2) across the tree (heterotachy)

$$q01 = \alpha$$
$$q10 = \beta$$

rate variation across sites

residence	descent	
$q01 = \alpha = 1$ $q10 = \beta = 2$	$q01 = \alpha = 1$ $q10 = \beta = 2$	characters have same rates
$q01 = \alpha = 1$ $q10 = \beta = 2$	$q01 = \alpha = 4$ $q10 = \beta = 1$	characters have different rates

The Gamma Solution:
a relative rate multiplier for branches

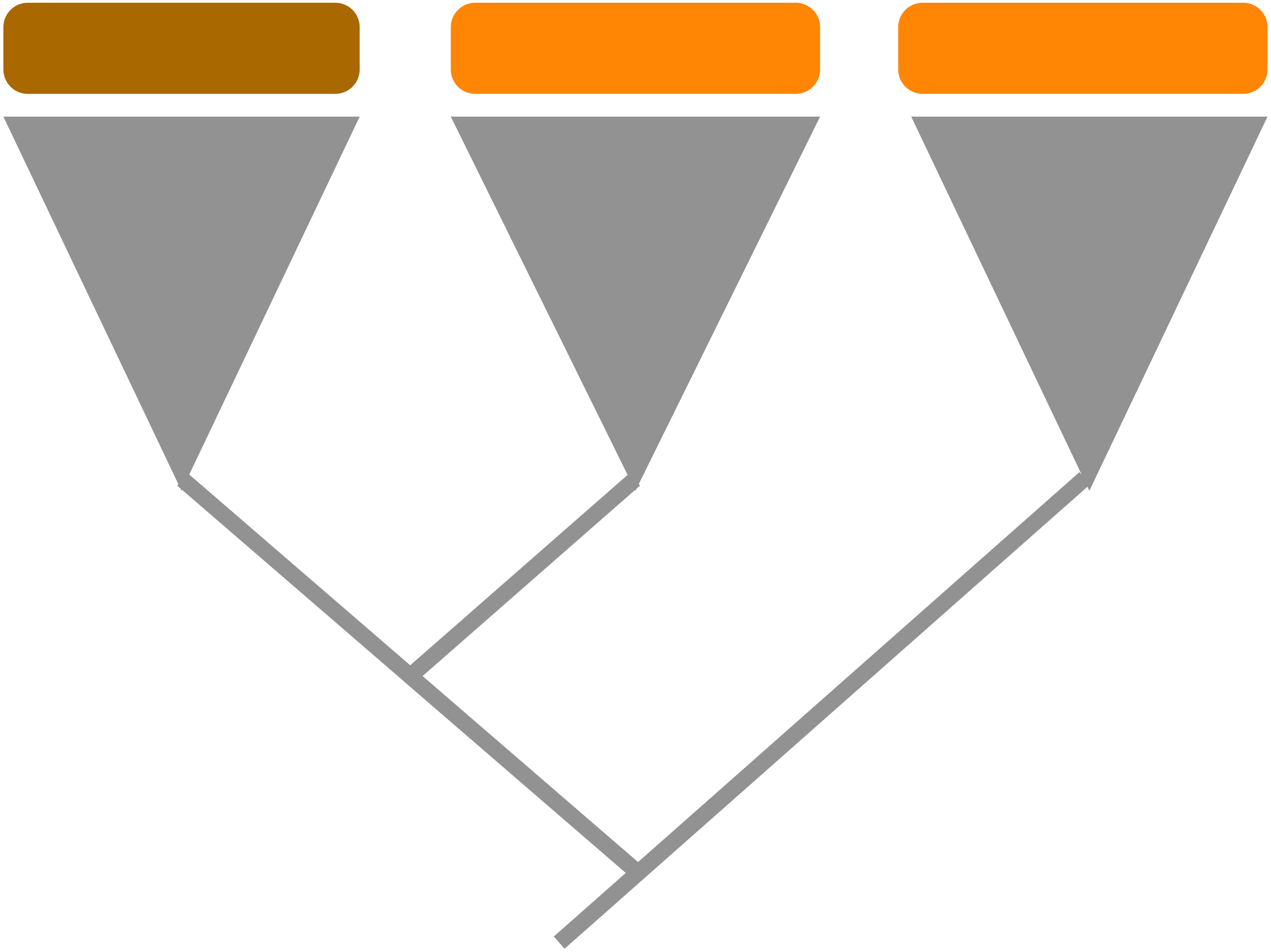
- shrinks branches for faster rates of evolution, stretches for lower
- draw multipliers from a gamma distribution, magically approximated by actually just four rates

Yang 1994

rate variation across the tree

residence	descent	
$q01 = \alpha = 1$ $q10 = \beta = 2$	$q01 = \alpha = 1$ $q10 = \beta = 2$	characters have same rates ...
$q01 = \alpha = 4$ $q10 = \beta = 1$	$q01 = \alpha = 4$ $q10 = \beta = 1$	... but rates differ across the tree

rate variation across the tree



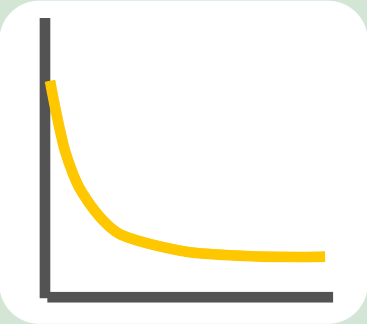
rate variation across the tree

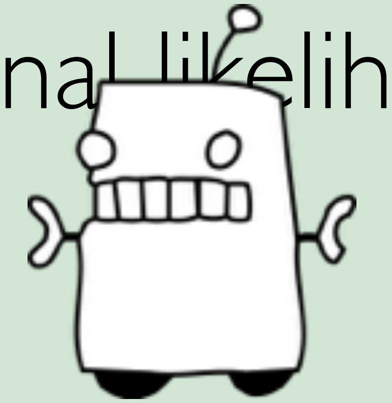
residence	descent	
$q01 = \alpha = 1$ $q10 = \beta = 2$	$q01 = \alpha = 1$ $q10 = \beta = 2$	characters have same rates ...
$q01 = \alpha = 4$ $q10 = \beta = 1$	$q01 = \alpha = 4$ $q10 = \beta = 1$	... but rates differ across the tree

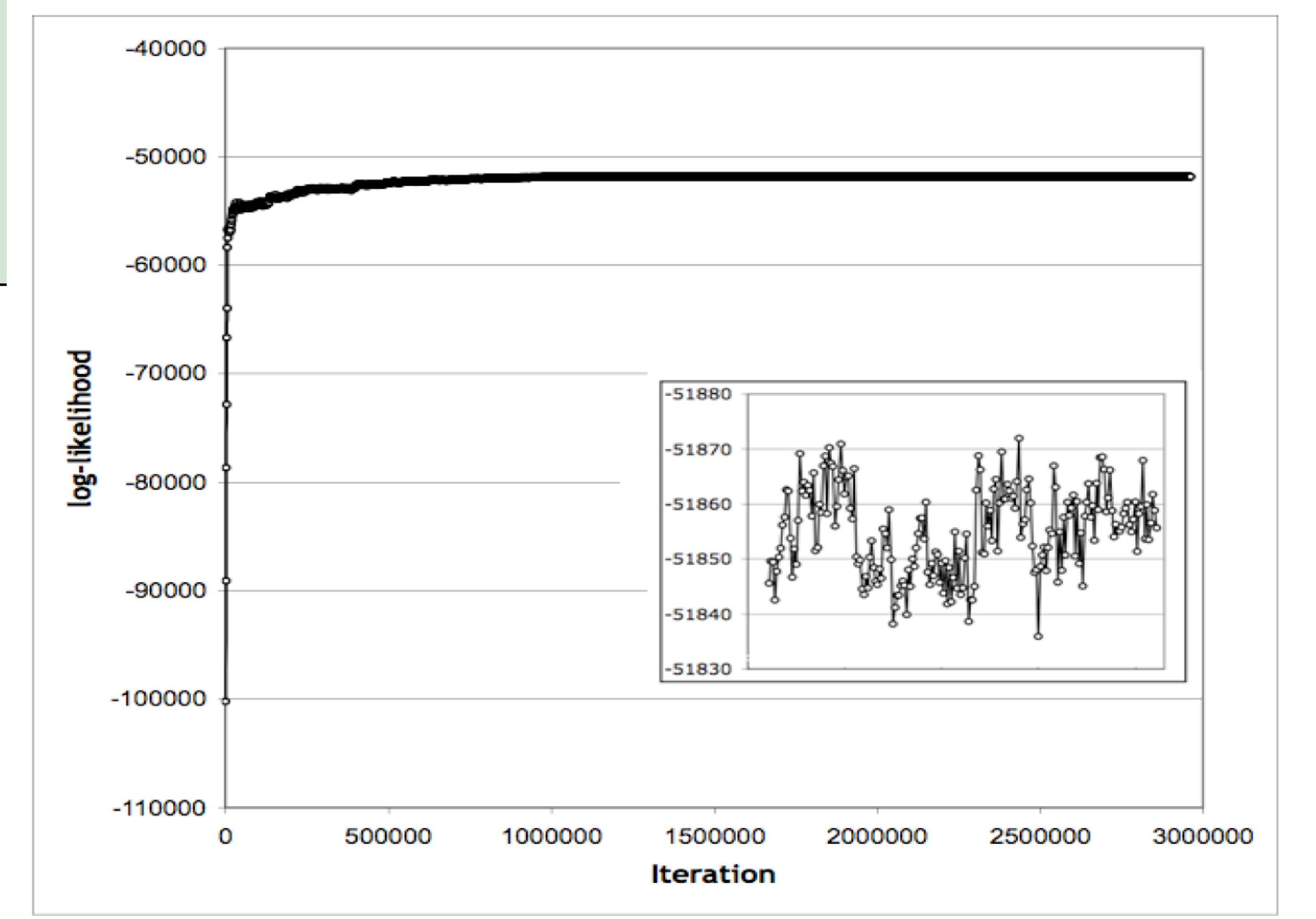
The Covarion Solution:

sites are switched on/off at different parts of the tree

posterior probability distribution I

posterior probability distribution = likelihood of data ×  / marginal likelihood





posterior probability distribution 2

diagnostics

chains
burn-in
convergence
acceptance rates
sampling / ESS
prior influence
MC^3

summary

Lh tracer plots
all parameters
histograms
credibility interval
marginal likelihood

interpret

Bayes factors
posterior:prior
model plots

Tracer is very useful
http://tree.bio.ed.ac.uk/software/tracer/

Tracer: Rambaut & Drummond

diagnostics I: chains

[1] As many as you can

[2] As long as you can run it

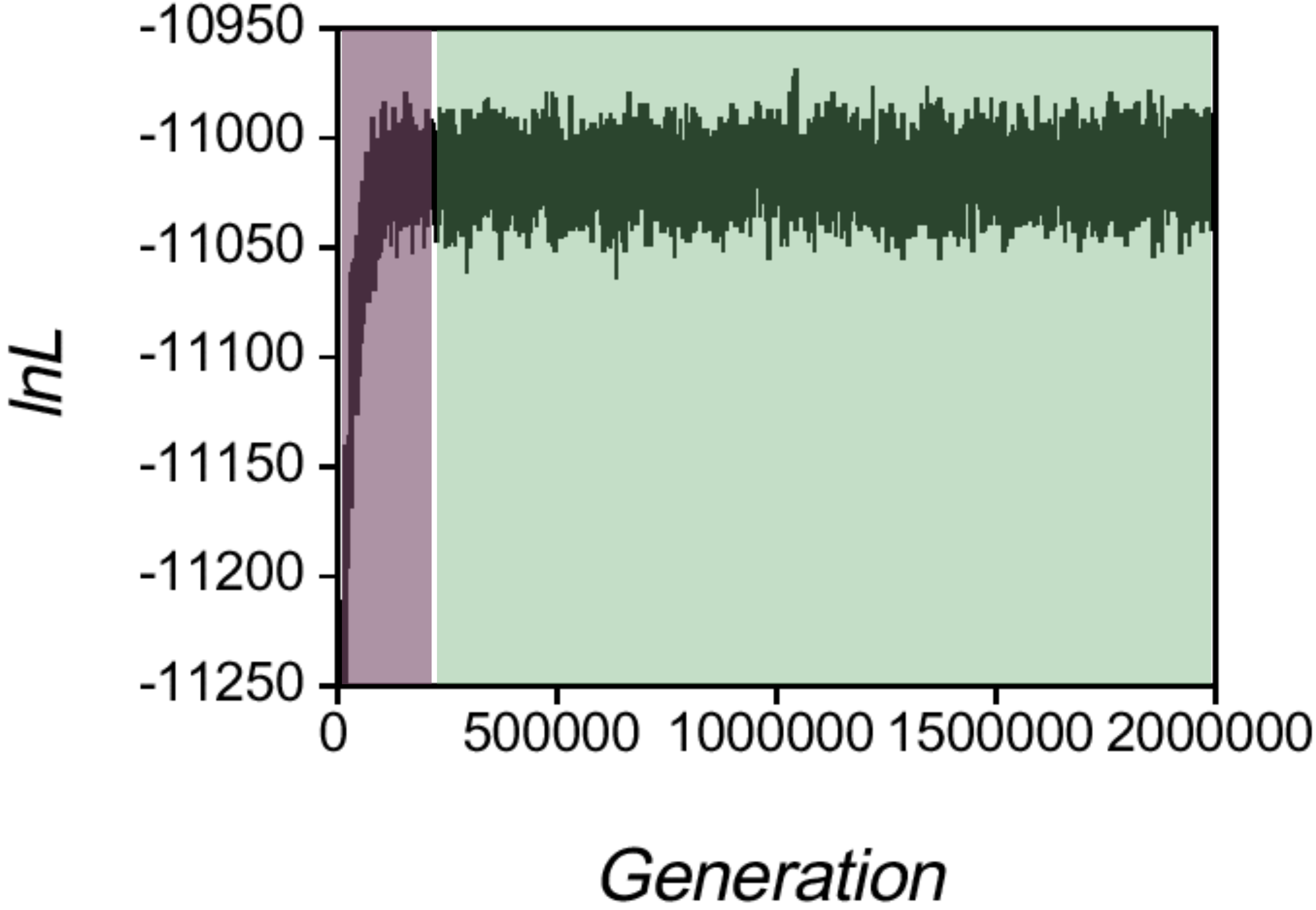
[3] Sample to reduce autocorrelation

[4] Use MC^3 for hot/cold chains in one run if possible

diagnostics 2: trace plots

burn-in

stationary phase
random fluctuations with no trend



diagnostics 3: acceptance rates & effective sample size

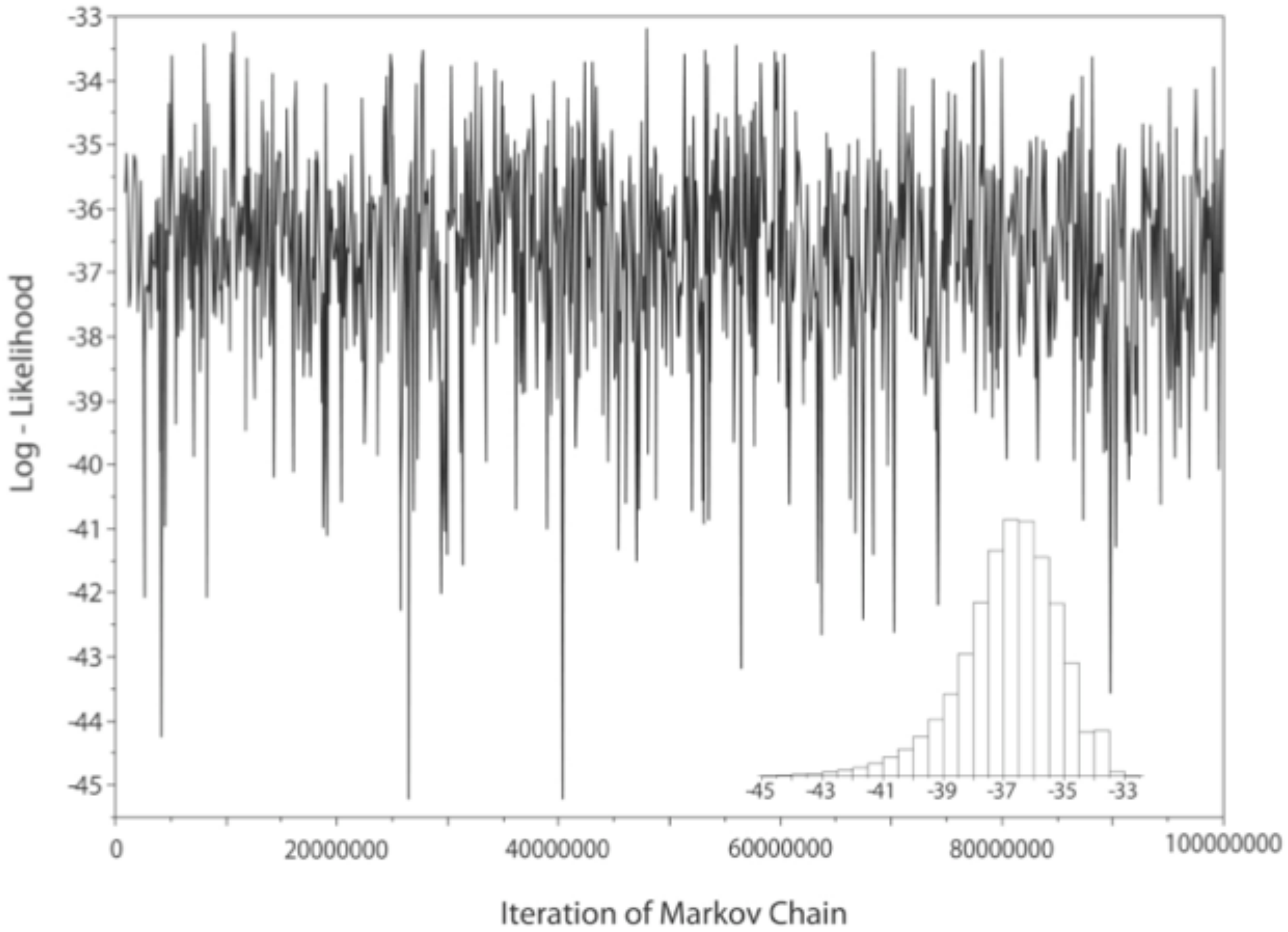
ACCEPTANCE

[1] Acceptance of the proposed move has a rate
[2] Aim for 20-70% acceptance
[3] Use the tuning parameter to hit this target (e.g. *ratedev* in BayesTraits)

ESS

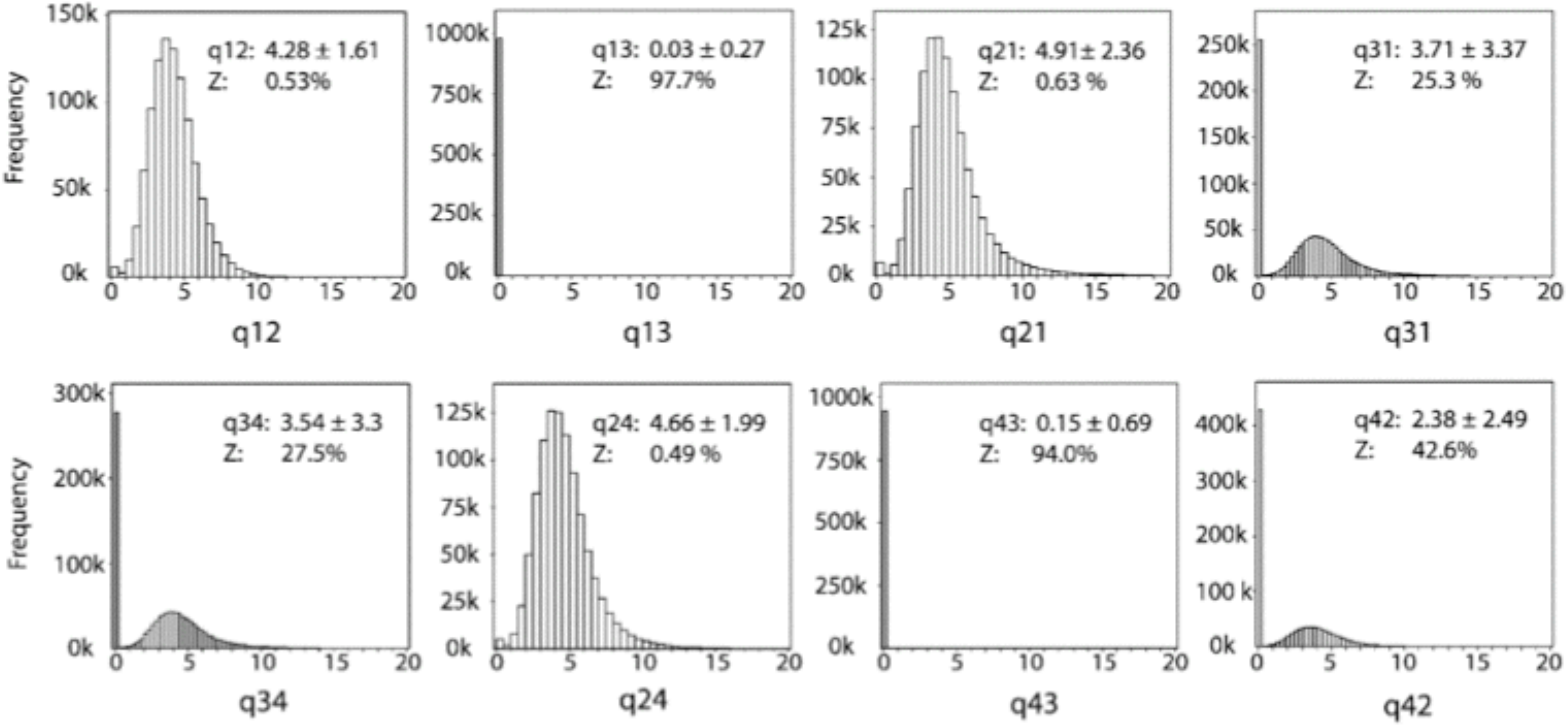
[1] ESS: effective sample size
[2] Each chain step is correlated with the previous
[3] Sample from the chain at large-ish intervals to reduce autocorrelation
[4] Plot θ_t by θ_{t+1} to assess

diagnostics 4: examples



[1] As many as you can
[2] As long as you can run it
[3] Sample to reduce autocorrelation
[4] Use MC^3 for hot/cold chains in one run if possible

summary 1: PPD histograms and 95%CI



95% credibility interval

contains the true values with 95% probability

Pagel & Meade 2006

summary 2: the marginal likelihood

Approximate the marginal likelihood of the PPD with the **harmonic mean** of the likelihoods

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} 1/x_i}, \qquad x_i > 0 \text{ for all } i.$$

harmonic mean = reciprocal of the arithmetic mean of all reciprocals

[1] gives more weight to small values

[2] minimises the effect of large values

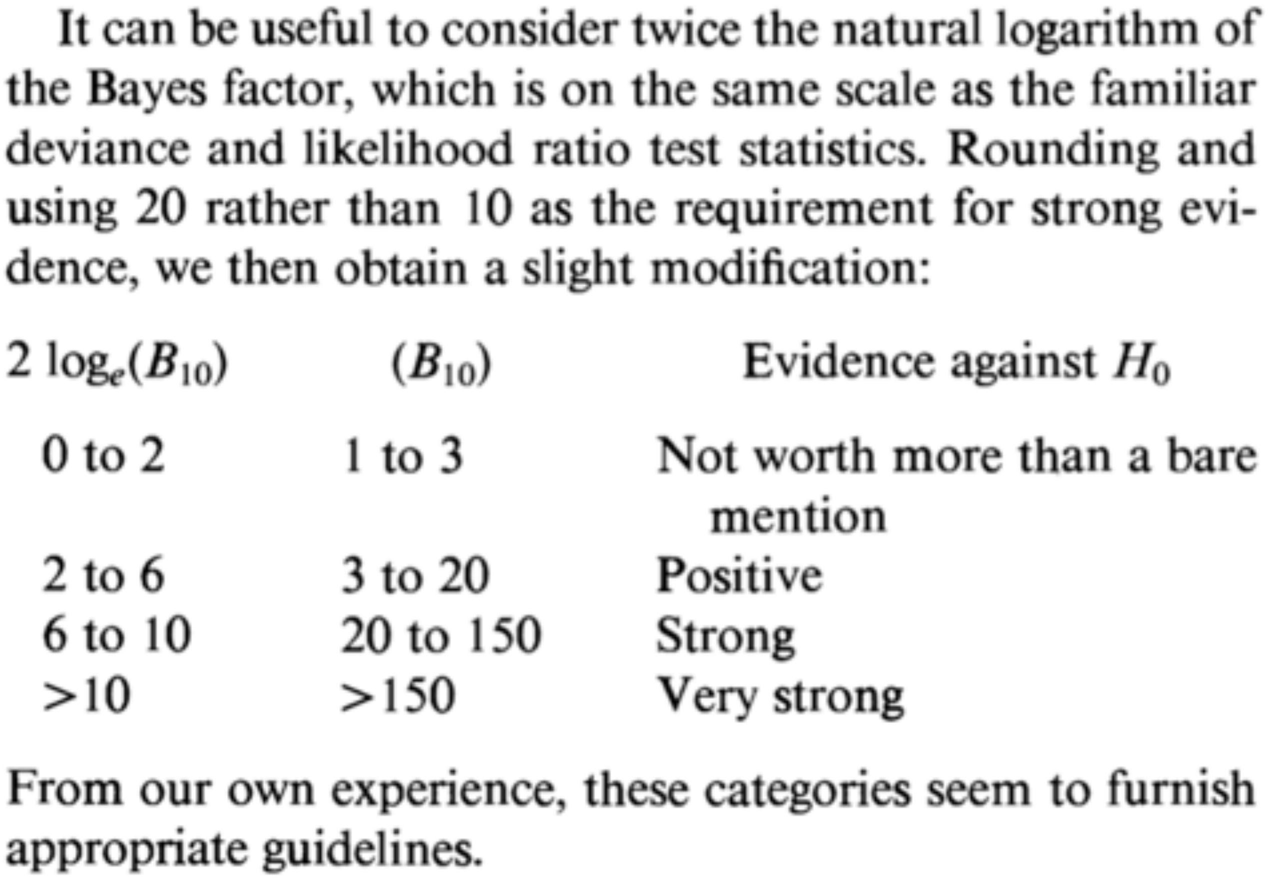
[3] not without contention!

Kass & Raftery 1995 / Newton & Raftery 1994

interpretation 1: model comparison and the Bayes factor

$$\mathrm{BF}_{ij} = \frac{P(D|M_i)}{P(D|M_j)}$$

$$2\log_e(\mathrm{BF})$$



It can be useful to consider twice the natural logarithm of the Bayes factor, which is on the same scale as the familiar deviance and likelihood ratio test statistics. Rounding and using 20 rather than 10 as the requirement for strong evidence, we then obtain a slight modification:

$2\log_e(B_{10})$	(B_{10})	Evidence against H_0
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

From our own experience, these categories seem to furnish appropriate guidelines.

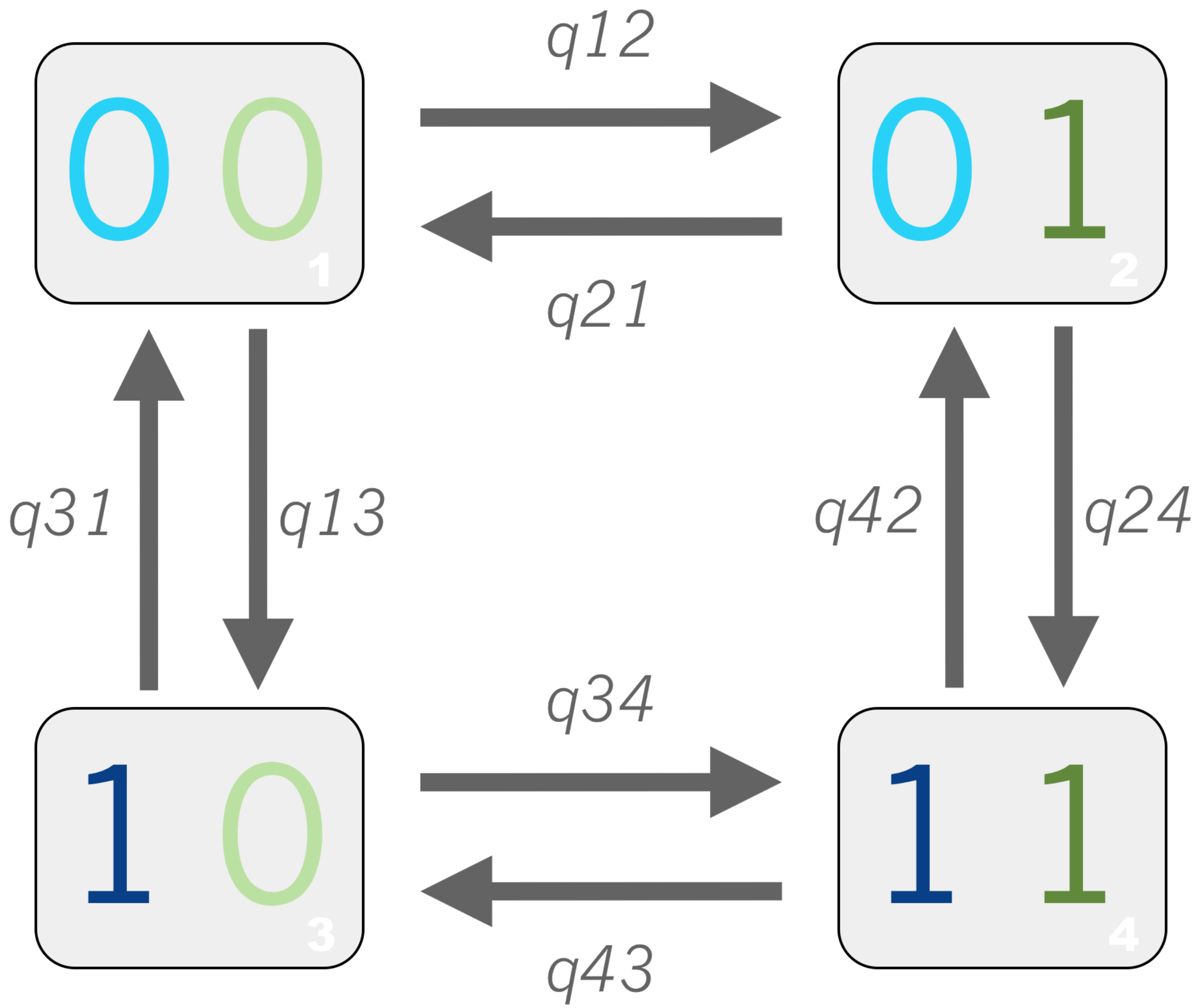
Bayes factor

An evaluation of the support for one model over another.

No penalties are needed for extra parameters, because a more parameter-rich model has a larger parameter space and therefore a lower prior probability density.

Kass & Raftery 1995 / Newton & Raftery 1994 / Raftery 1996

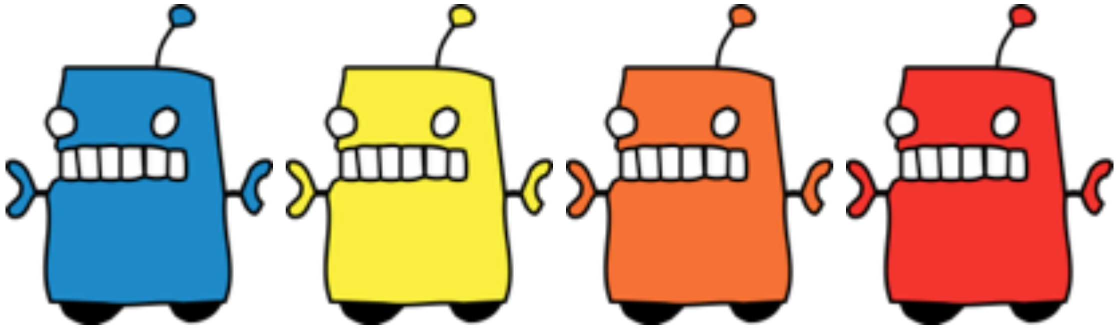
interpretation 2: model plots for BayesTraits



Metropolis-coupled MCMC (MC3)

Imagine the PPD surface is made of wax

Multiple MCMC robots descend



A "hot" chain robot acts as a scout for the "cold" captain robot

Heat distorts the joint PPD and flattens it
> this makes it easier to explore regions of low probability

The cold chain (which could be stuck in a local optimum) can escape when a proposed swap with a hot chain is successful.