

An Introduction to Bayesian Phylogenetics

30 January 2015

Paul O. Lewis

Department of Ecology & Evolutionary Biology



Workshop on Molecular Evolution
Český Krumlov



An Introduction to Bayesian Phylogenetics

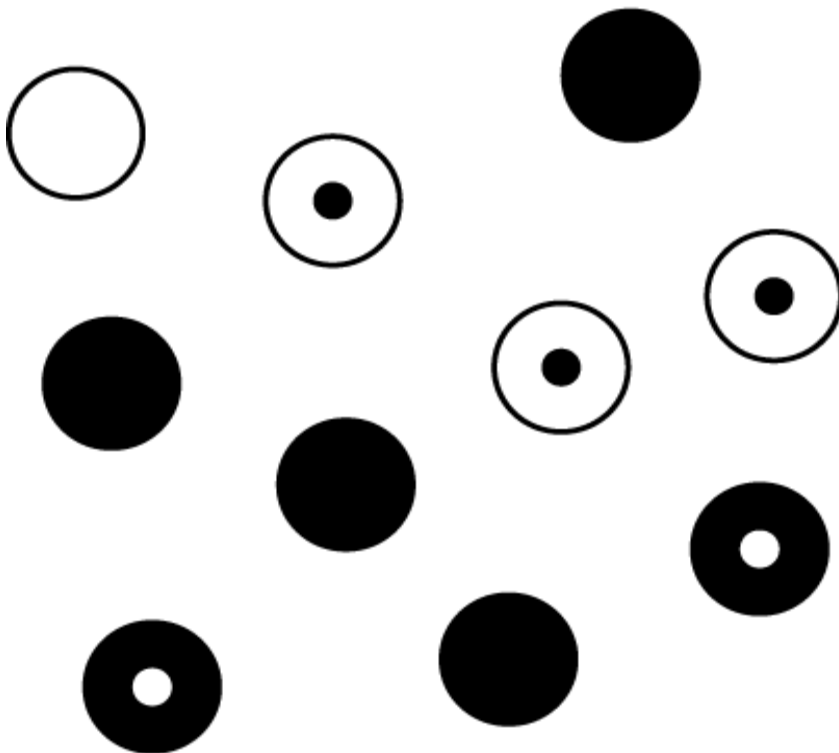
- Bayesian inference in general
- Markov chain Monte Carlo (MCMC)
- Bayesian phylogenetics
- Prior distributions
- Bayesian model selection

I. Bayesian inference in general

Joint probabilities

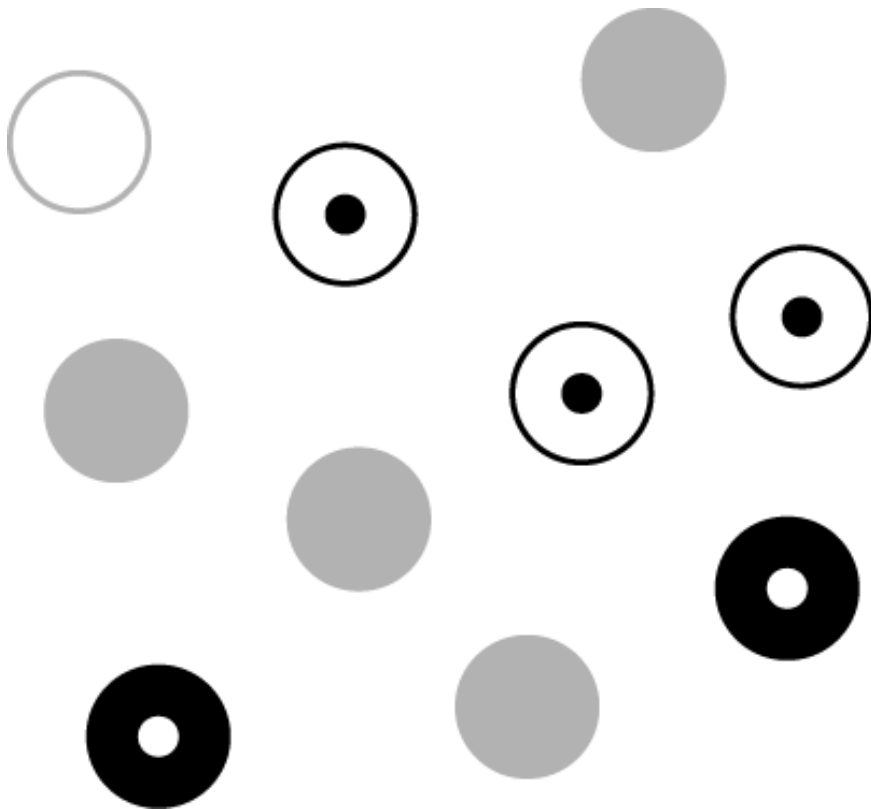
B = Black S = Solid
W = White D = Dotted

$$\begin{aligned}\Pr(B) &= 0.6 & \Pr(S) &= 0.5 \\ \Pr(W) &= 0.4 & \Pr(D) &= 0.5\end{aligned}$$



$$\begin{aligned}\Pr(\bullet) &= \Pr(B, D) = 0.2 \\ \Pr(\bullet) &= \Pr(B, S) = 0.4 \\ \Pr(\odot) &= \Pr(W, D) = 0.3 \\ \Pr(\bigcirc) &= \Pr(W, S) = 0.1\end{aligned}$$

Conditional probabilities



$$\Pr(B|D) = \frac{2}{5} = 0.4$$

Hide all solid marbles
(leaving 5 with dot)

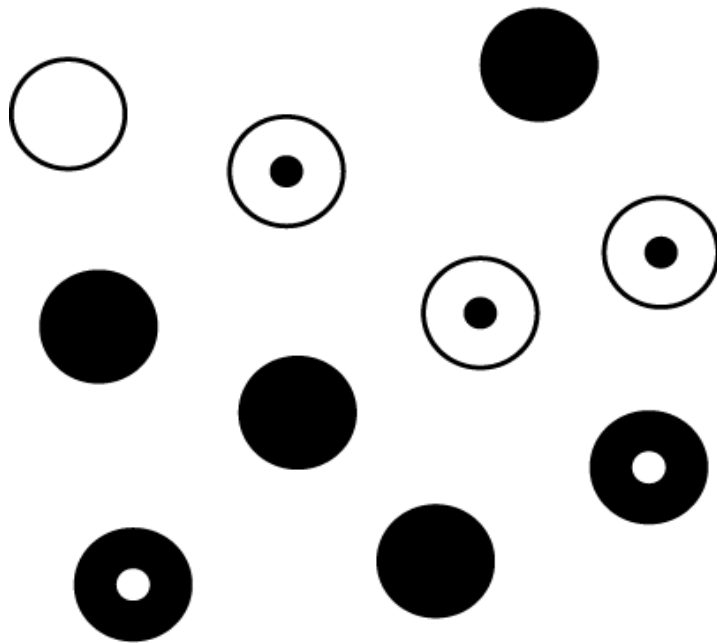
Of those left, 2 are black

Bayes' rule

$\Pr(B, D)$

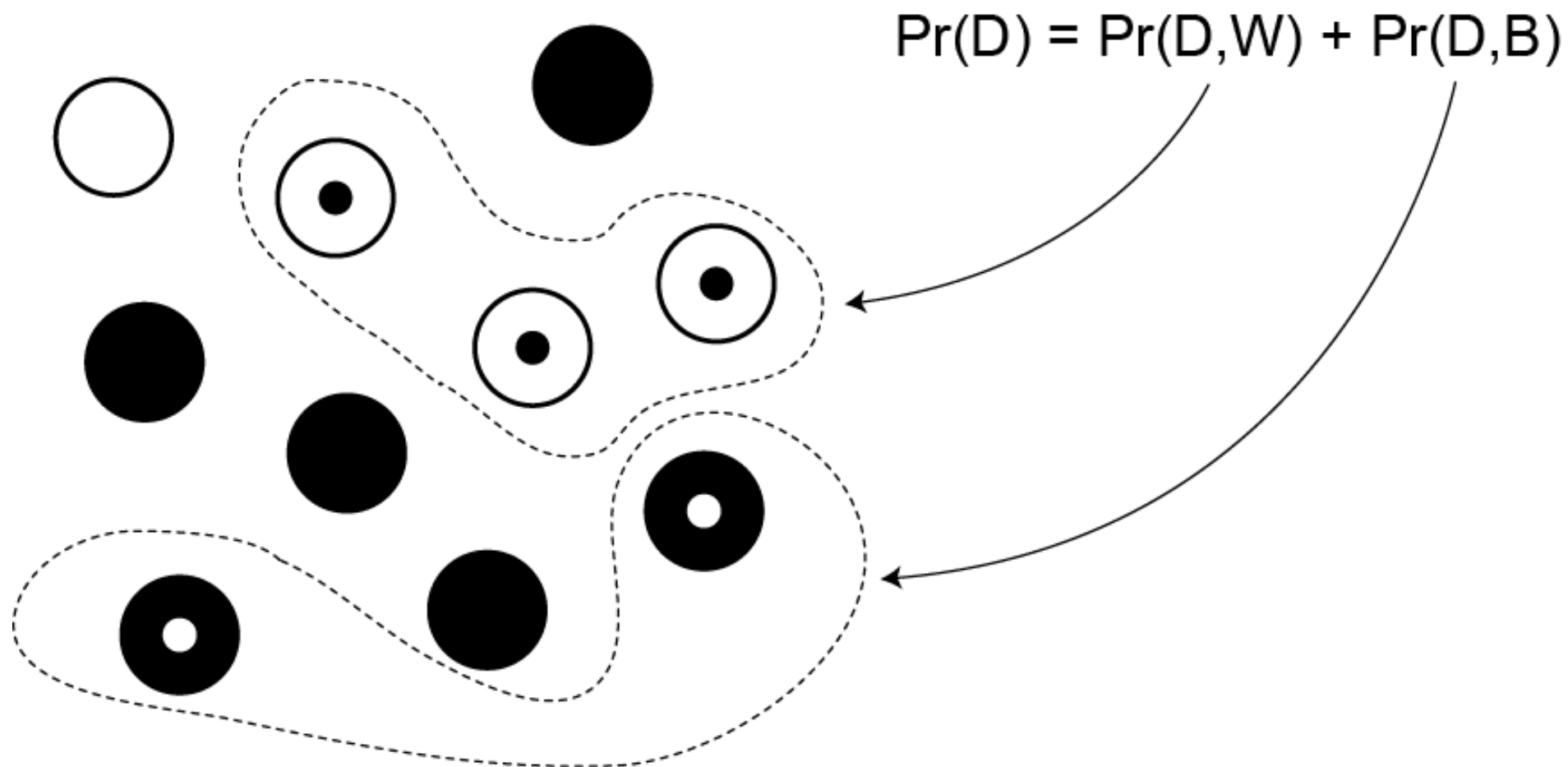
$$\Pr(D) \Pr(B|D) = \Pr(B) \Pr(D|B)$$

$$\frac{1}{2} \times \frac{2}{5} = \frac{3}{5} \times \frac{1}{3}$$



$$\begin{aligned} \Pr(B|D) &= \frac{\Pr(B) \Pr(D|B)}{\Pr(D)} \\ &= \frac{\frac{3}{5} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{5} \end{aligned}$$

Probability of "Dotted"



Bayes' rule (cont.)

$$\begin{aligned}\Pr(B|D) &= \frac{\Pr(B) \Pr(D|B)}{\Pr(D)} \\ &= \frac{\Pr(D, B)}{\Pr(D, B) + \Pr(D, W)}\end{aligned}$$

$\Pr(D)$ is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

Bayes' rule (cont.)

It is easy to see that $\Pr(D)$ serves as a *normalization constant*, ensuring that $\Pr(B|D) + \Pr(W|D) = 1.0$

$$\Pr(B|D) = \frac{\Pr(D, B)}{\Pr(D, B) + \Pr(D, W)} \longleftarrow \Pr(D)$$

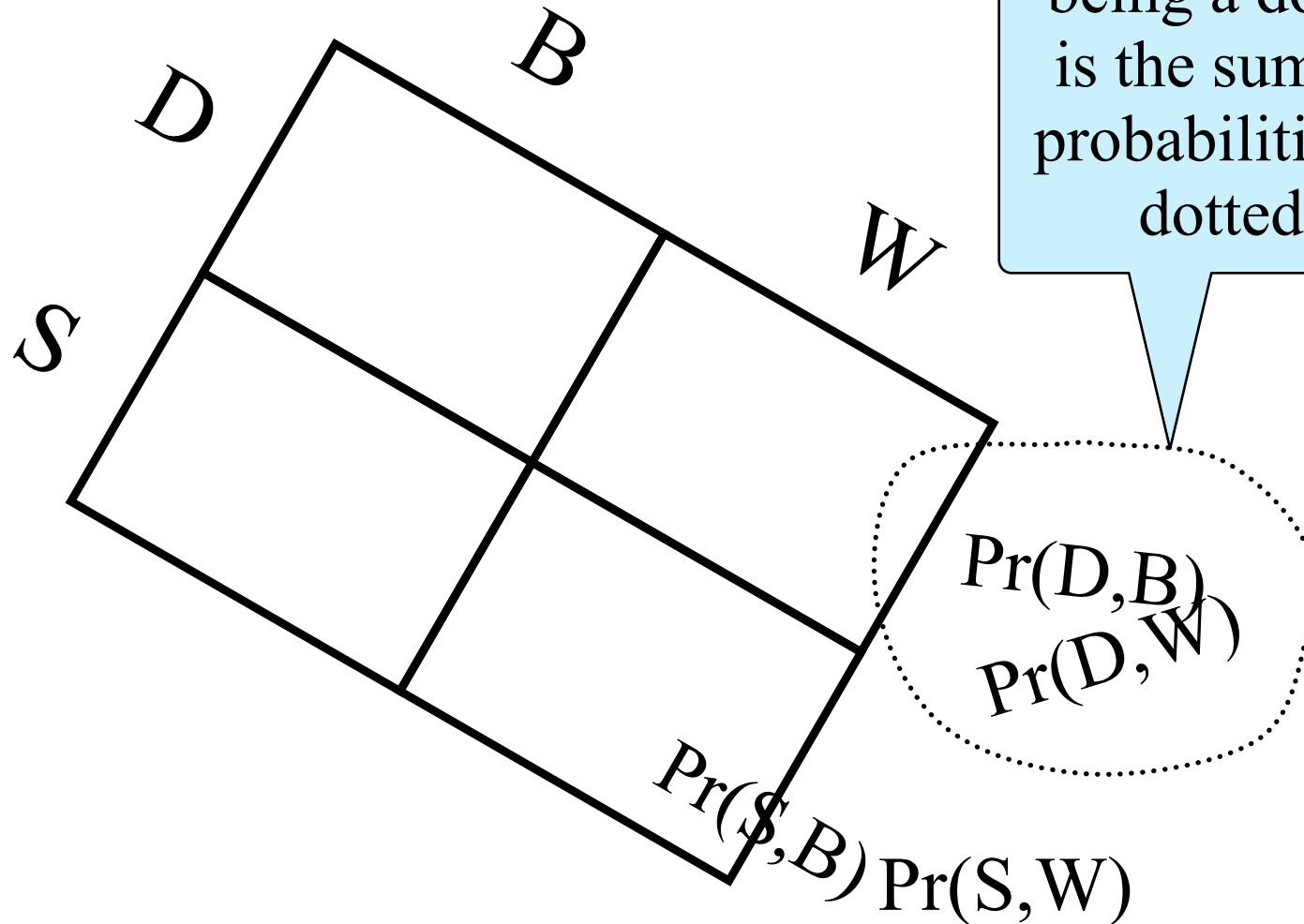
$$\Pr(W|D) = \frac{\Pr(D, W)}{\Pr(D, B) + \Pr(D, W)} \longleftarrow \Pr(D)$$

$$\Pr(B|D) + \Pr(W|D) = \frac{\cancel{\Pr(D, B)} + \cancel{\Pr(D, W)}}{\cancel{\Pr(D, B)} + \cancel{\Pr(D, W)}} = 1$$

Joint probabilities

	B	W
D	$\Pr(D,B)$	$\Pr(D,W)$
S	$\Pr(S,B)$	$\Pr(S,W)$

Marginalizing over colors



Marginal probability of being a dotted marble is the sum of all joint probabilities involving dotted marbles

Marginal probabilities

	B	W	
D			$\Pr(D)$ = marginal probability of being dotted
S			$\Pr(S)$ = marginal probability of being solid

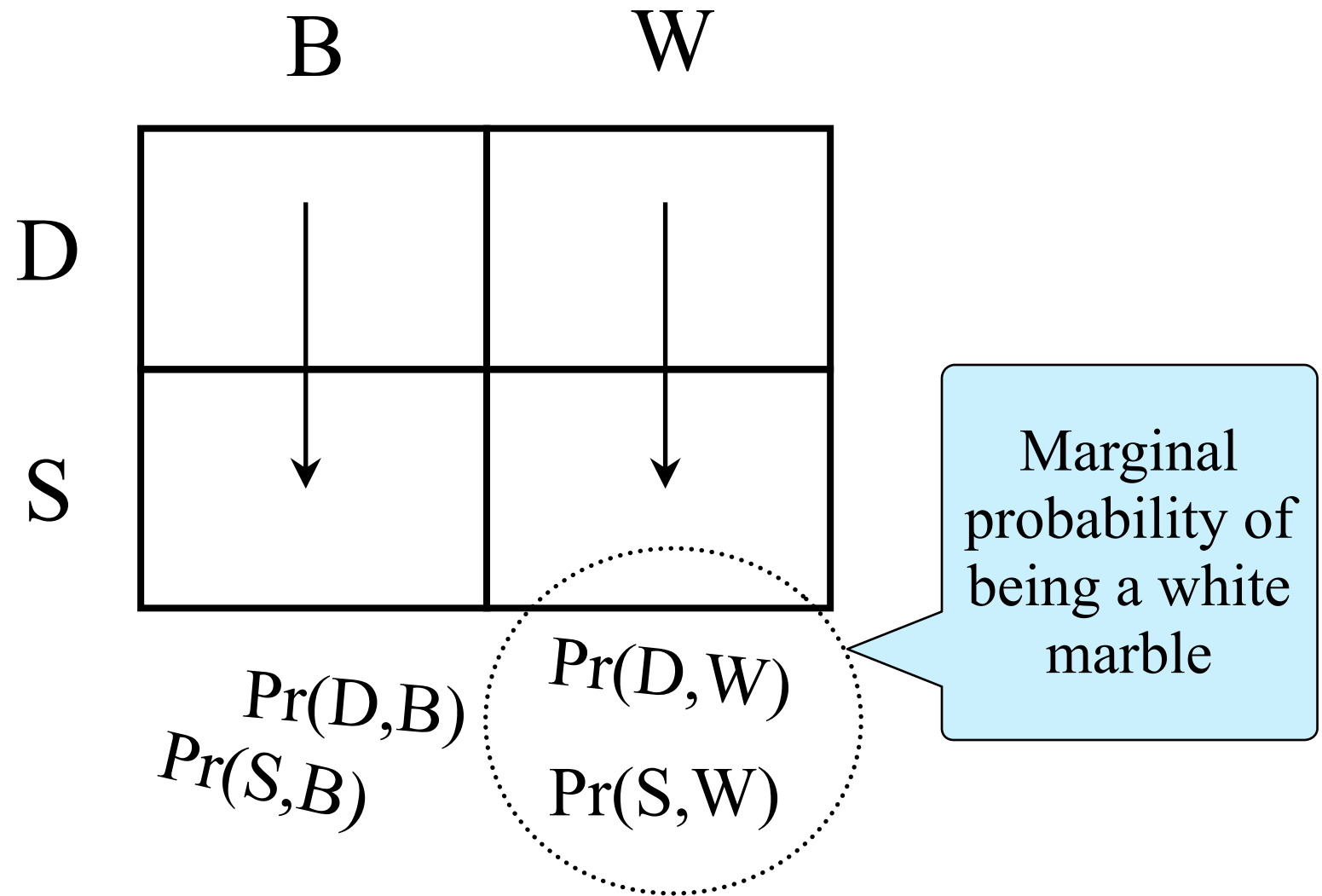
$\Pr(D,B) + \Pr(D,W)$

$\Pr(S,B) + \Pr(S,W)$

Joint probabilities

	B	W
D	$\Pr(D,B)$	$\Pr(D,W)$
S	$\Pr(S,B)$	$\Pr(S,W)$

Marginalizing over "dottedness"



Bayes' rule (cont.)

$$\begin{aligned}\Pr(B|D) &= \frac{\Pr(B) \Pr(D|B)}{\Pr(D, B) + \Pr(D, W)} \\ &= \frac{\Pr(B) \Pr(D|B)}{\Pr(B) \Pr(D|B) + \Pr(W) \Pr(D|W)} \\ &= \frac{\Pr(B) \Pr(D|B)}{\sum_{\theta \in \{B, W\}} \Pr(\theta) \Pr(D|\theta)}\end{aligned}$$

Bayes' rule in Statistics

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$

D refers to the "observables" (i.e. the **Data**)

θ refers to one or more "unobservables"
(i.e. **parameters** of a model, or the **model itself**):

- *tree model* (i.e. tree topology)
- *substitution model* (e.g. JC, F84, GTR, etc.)
- *parameter* of a substitution model (e.g. a branch length, a base frequency, transition/transversion rate ratio, etc.)
- *hypothesis* (i.e. a special case of a model)
- *a latent variable* (e.g. ancestral state)

Bayes' rule in statistics

Likelihood of hypothesis θ

Prior probability of hypothesis θ

$$\Pr(\theta|D) = \frac{\Pr(D|\theta) \Pr(\theta)}{\sum_{\theta} \Pr(D|\theta) \Pr(\theta)}$$

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

The diagram illustrates Bayes' rule with the following components and annotations:

- Posterior probability of hypothesis θ** : A purple box containing the expression $\Pr(\theta|D)$.
- Likelihood of hypothesis θ** : A blue box containing the expression $\Pr(D|\theta)$.
- Prior probability of hypothesis θ** : An orange box containing the expression $\Pr(\theta)$.
- Marginal probability of the data (marginalizing over hypotheses)**: A green box containing the expression $\sum_{\theta} \Pr(D|\theta) \Pr(\theta)$.

Arrows indicate the flow of information: from the likelihood and prior boxes to the numerator of the fraction, and from the denominator box to the equals sign. An arrow also points from the posterior box to the equals sign.

Simple (albeit silly) paternity example

θ_1 and θ_2 are assumed to be the only possible fathers, **child** has genotype **Aa**, **mother** has genotype **aa**, so child must have received allele **A** from the true father. Note: the **data** in this case is the child's genotype (**Aa**)

Possibilities	θ_1	θ_2	Row sum
Genotypes	AA	Aa	---
Prior	1/2	1/2	1
Likelihood	1	1/2	---
Prior X Likelihood	1/2	1/4	3/4
Posterior	2/3	1/3	1

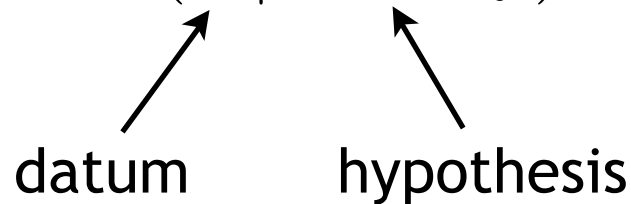
The prior can be your friend

Suppose the test for a **rare** disease is 99% accurate.

$$\Pr(+|\text{disease}) = 0.99$$

$$\Pr(+|\text{healthy}) = 0.01$$

datum hypothesis



Suppose further I **test positive** for the disease.
How worried should I be?

(Note that we do not
need to consider the case
of a negative test result.)

It is very tempting to (mis)interpret the likelihood as a posterior probability and conclude that there is a 99% chance that I have the disease.

Want to know $\Pr(\text{disease} | +)$, not $\Pr(+ | \text{disease})$

The prior can be your friend

The posterior probability is 0.99 only if the **prior probability** of having the disease is 0.5:

$$\begin{aligned}\Pr(\text{disease}|+) &= \frac{\Pr(+|\text{disease}) \left(\frac{1}{2}\right)}{\Pr(+|\text{disease}) \left(\frac{1}{2}\right) + \Pr(+|\text{healthy}) \left(\frac{1}{2}\right)} \\ &= \frac{(0.99) \left(\frac{1}{2}\right)}{(0.99) \left(\frac{1}{2}\right) + (0.01) \left(\frac{1}{2}\right)} = 0.99\end{aligned}$$

If, however, the prior odds against having the disease are 1 million to 1, then the posterior probability is much more reassuring:

$$\begin{aligned}\Pr(\text{disease}|+) &= \frac{(0.99) \left(\frac{1}{1000000}\right)}{(0.99) \left(\frac{1}{1000000}\right) + (0.01) \left(\frac{999999}{1000000}\right)} \\ &\approx 0.0001\end{aligned}$$

An important caveat

This (rare disease) example involves a **tiny amount of data** (one observation) and an extremely **informative prior**, and gives the impression that maximum likelihood (ML) inference is not very reliable.

However, in phylogenetics, we often have **lots of data** and use much **less informative priors**, so in phylogenetics ML inference is generally **very reliable**.

Discrete vs. Continuous

- So far, we've been dealing with **discrete hypotheses** (e.g. either this father or that father, have disease or don't have disease)
- In phylogenetics, substitution models represent an **infinite number of hypotheses** (each combination of parameter values is in some sense a separate hypothesis)
- How do we use Bayes' rule when our hypotheses form a continuum?

Bayes' rule: continuous case

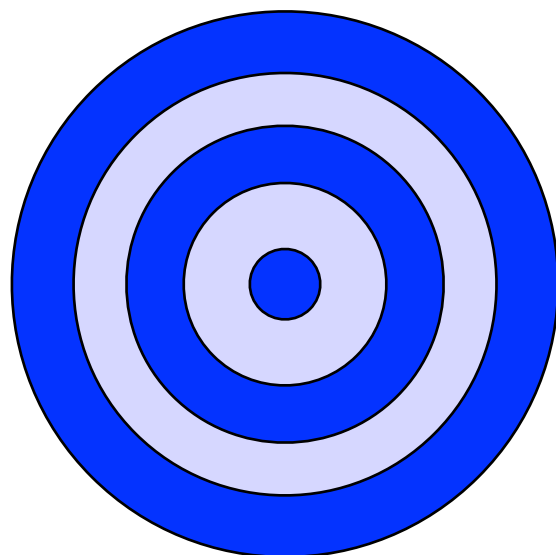
The diagram illustrates Bayes' rule for the continuous case. It features the equation $f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$. The components are color-coded and labeled with arrows: the numerator's first term $f(D|\theta)$ is in a blue box labeled 'Likelihood'; the second term $f(\theta)$ is in an orange box labeled 'Prior probability density'; the denominator $\int f(D|\theta)f(\theta)d\theta$ is in a green box labeled 'Marginal probability of the data'; and the entire left side $f(\theta|D)$ is in a purple box labeled 'Posterior probability density'.

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

Labels and arrows:

- Likelihood** points to $f(D|\theta)$
- Prior probability *density*** points to $f(\theta)$
- Posterior probability *density*** points to $f(\theta|D)$
- Marginal probability of the data** points to the denominator $\int f(D|\theta)f(\theta)d\theta$

If you had to guess...



← 1 meter →

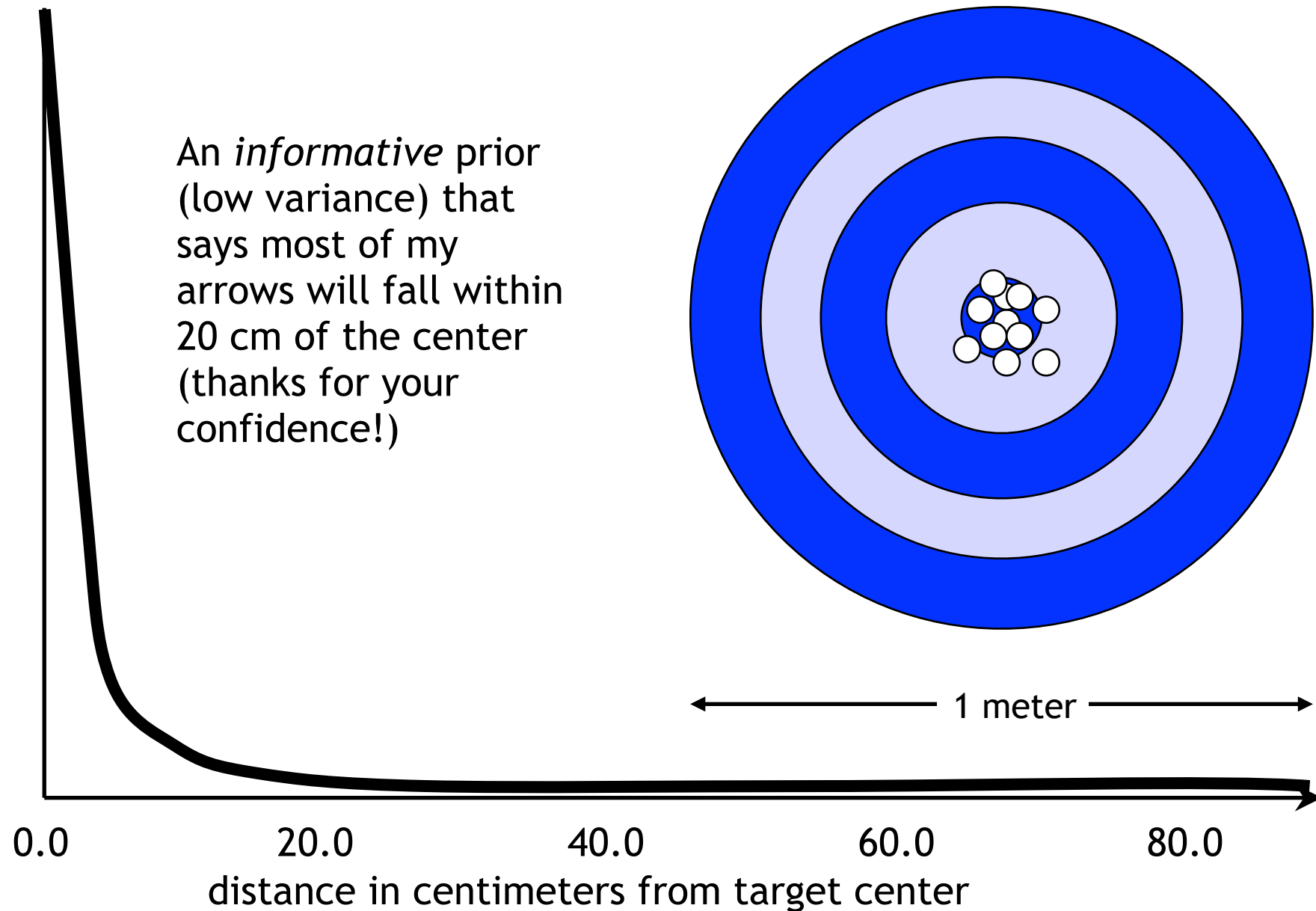
Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance d from the center of the target (if it helps, I'm standing 50 meters away from the target)

0.0

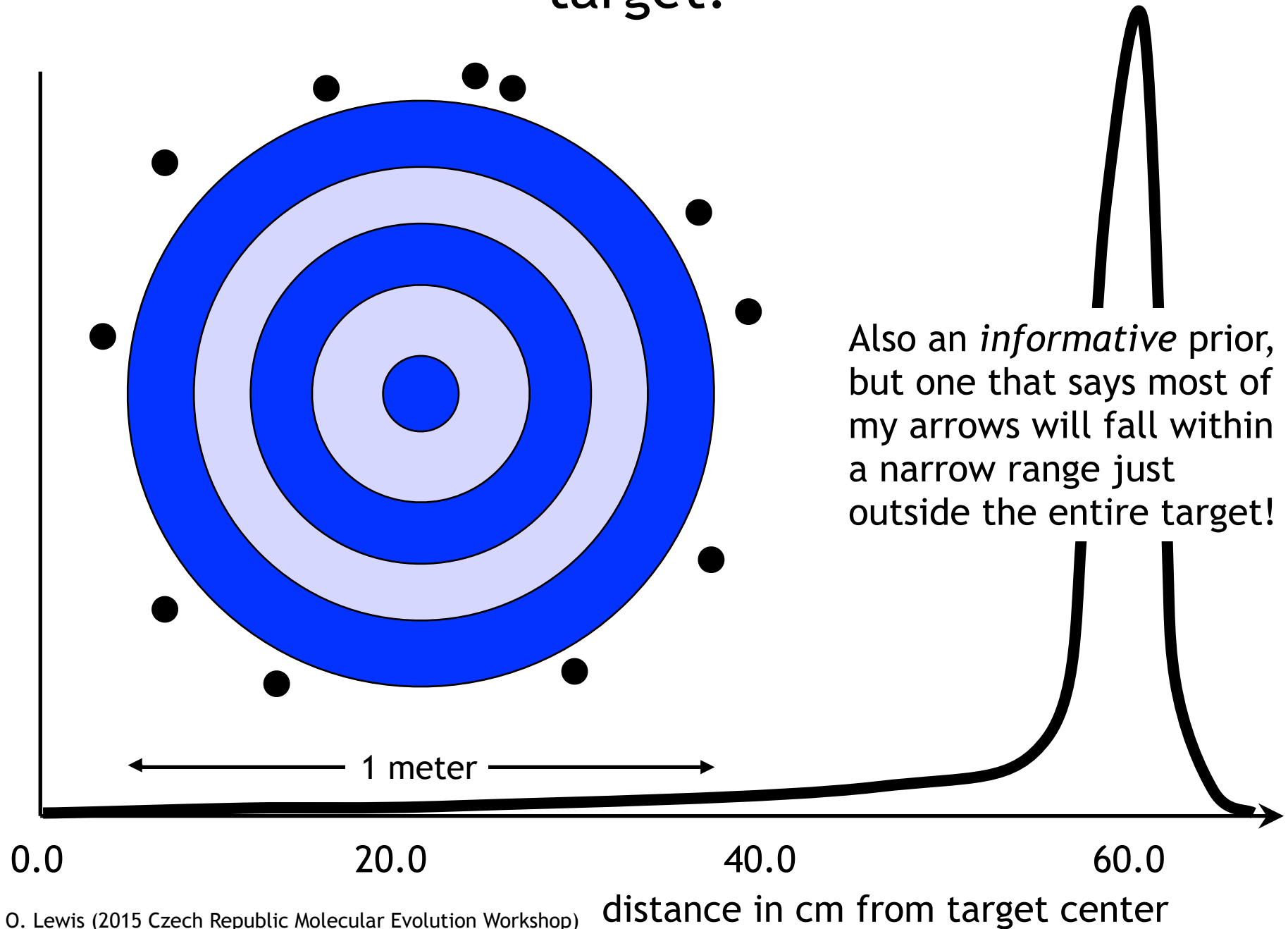
d

∞

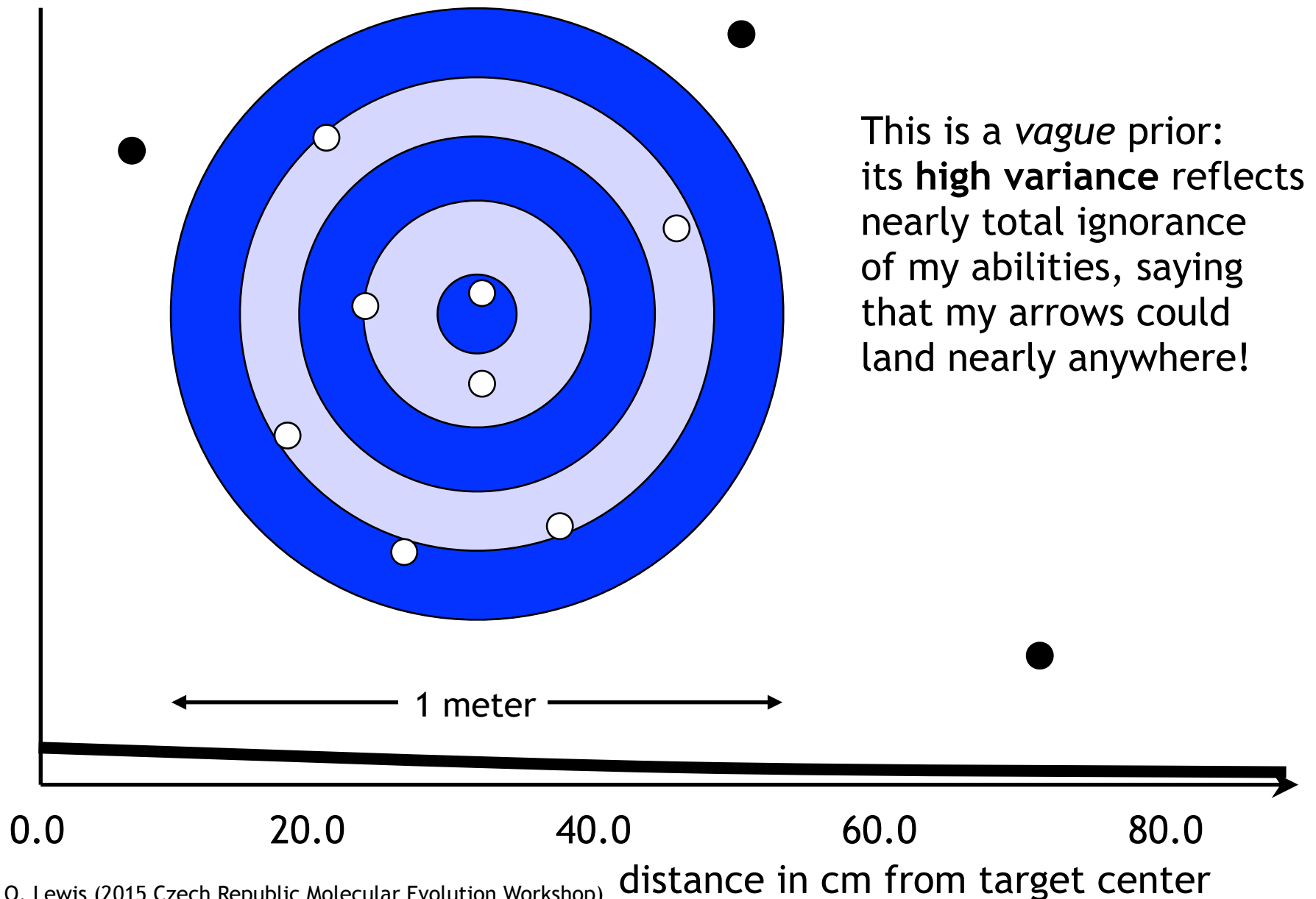
Case 1: assume I have talent



Case 2: assume I have a talent for missing the target!



Case 3: assume I have no talent



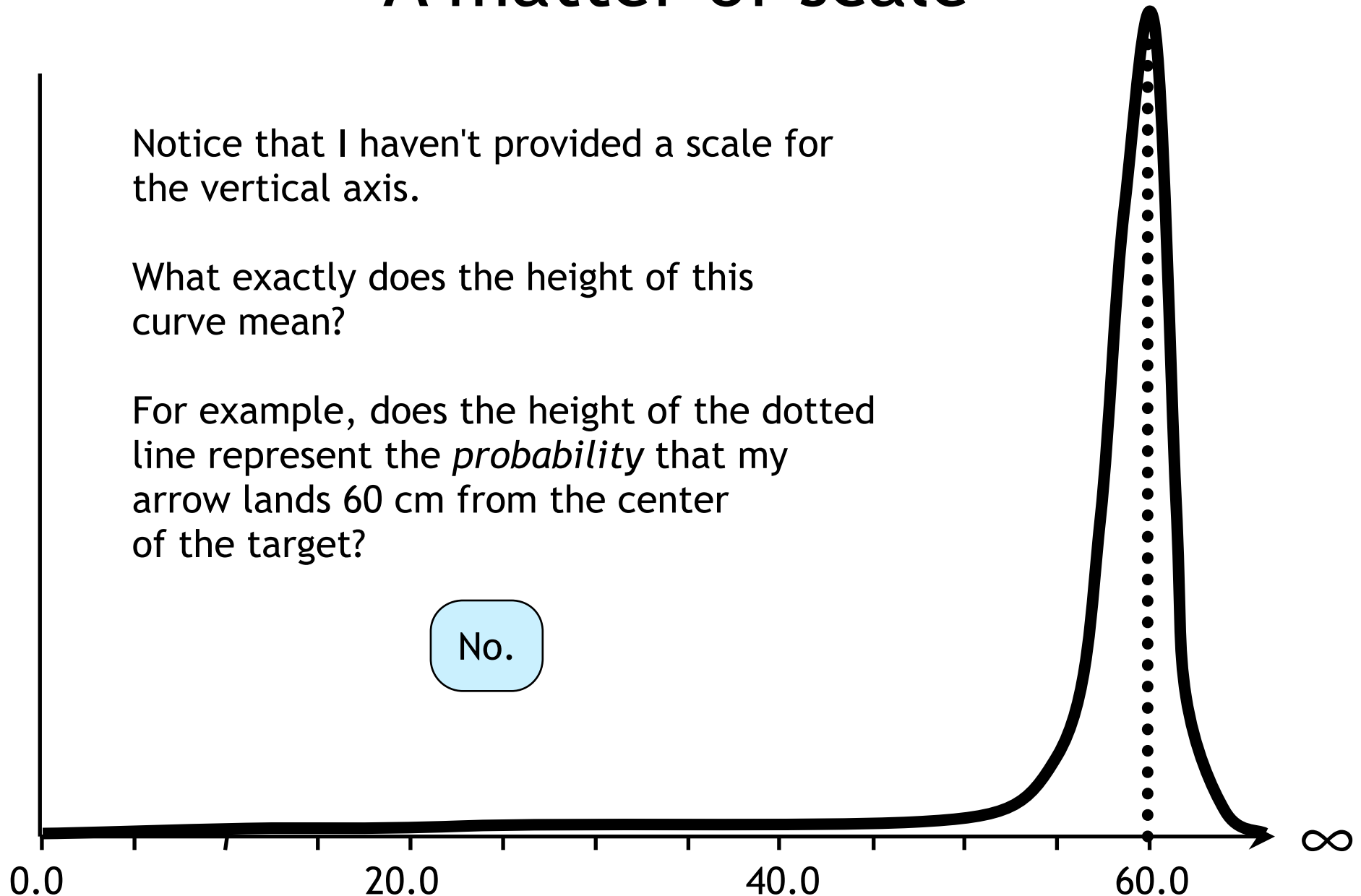
A matter of scale

Notice that I haven't provided a scale for the vertical axis.

What exactly does the height of this curve mean?

For example, does the height of the dotted line represent the *probability* that my arrow lands 60 cm from the center of the target?

No.

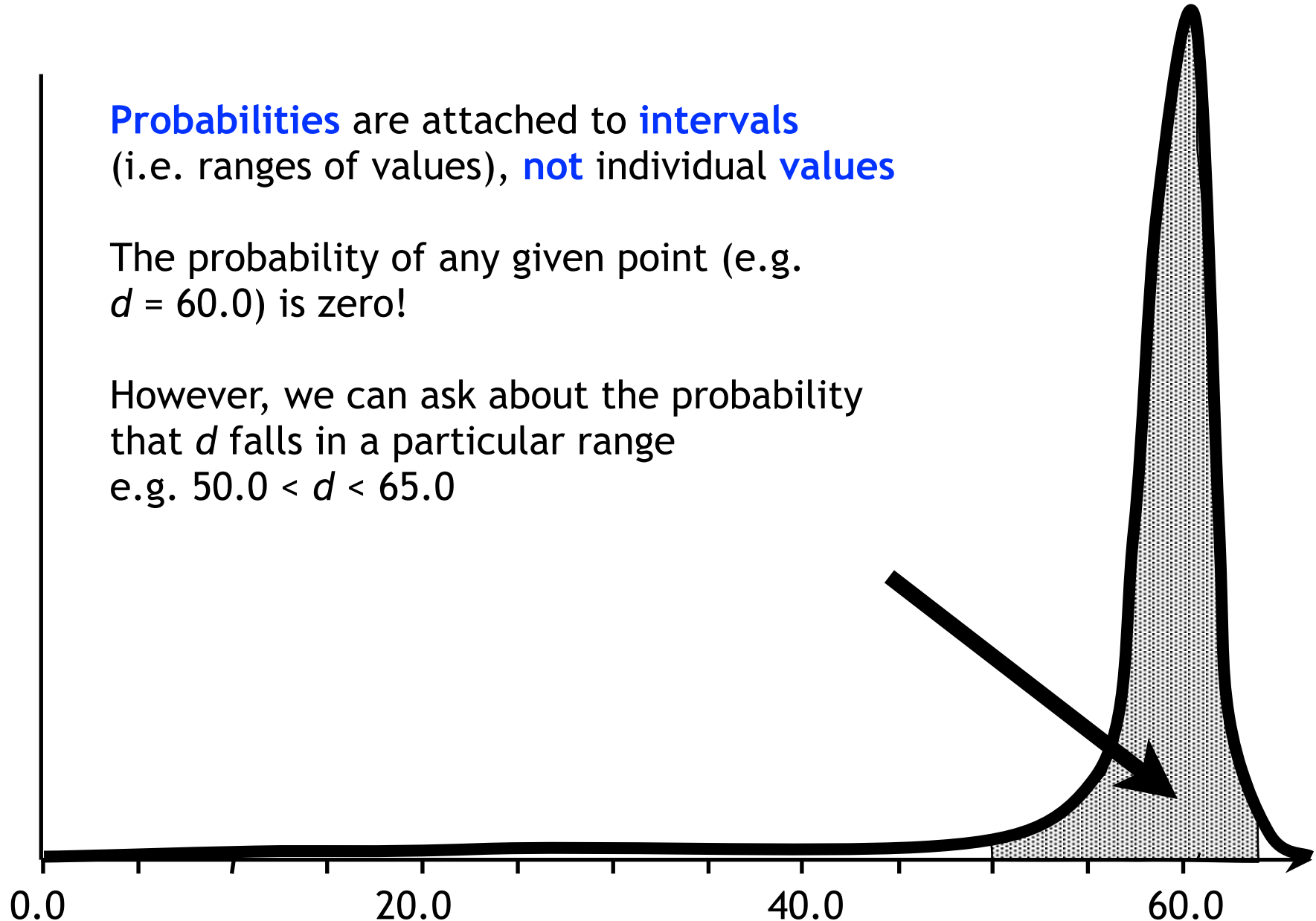


Probabilities are associated with intervals

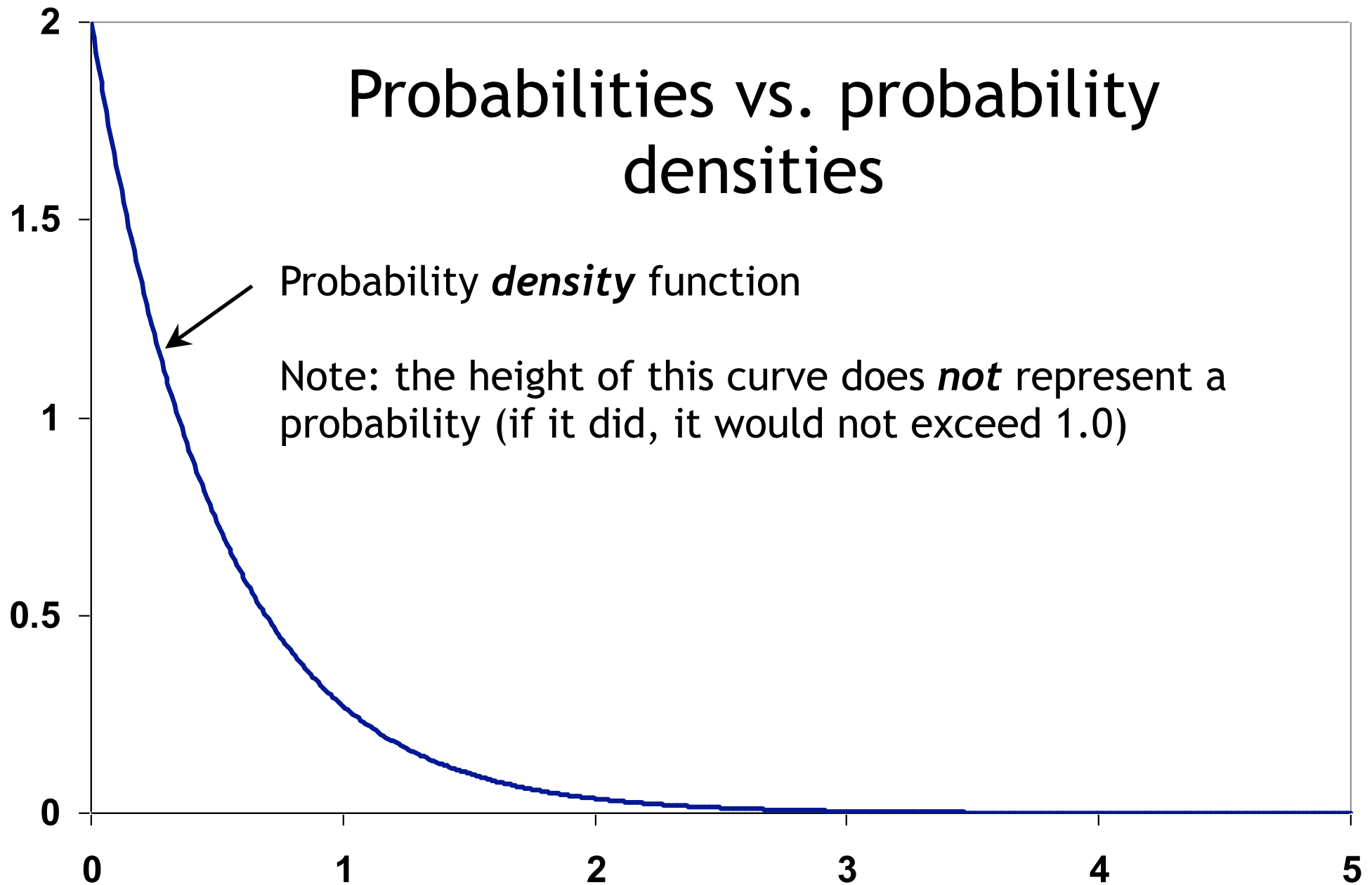
Probabilities are attached to **intervals**
(i.e. ranges of values), **not** individual **values**

The probability of any given point (e.g.
 $d = 60.0$) is zero!

However, we can ask about the probability
that d falls in a particular range
e.g. $50.0 < d < 65.0$



Probabilities vs. probability densities



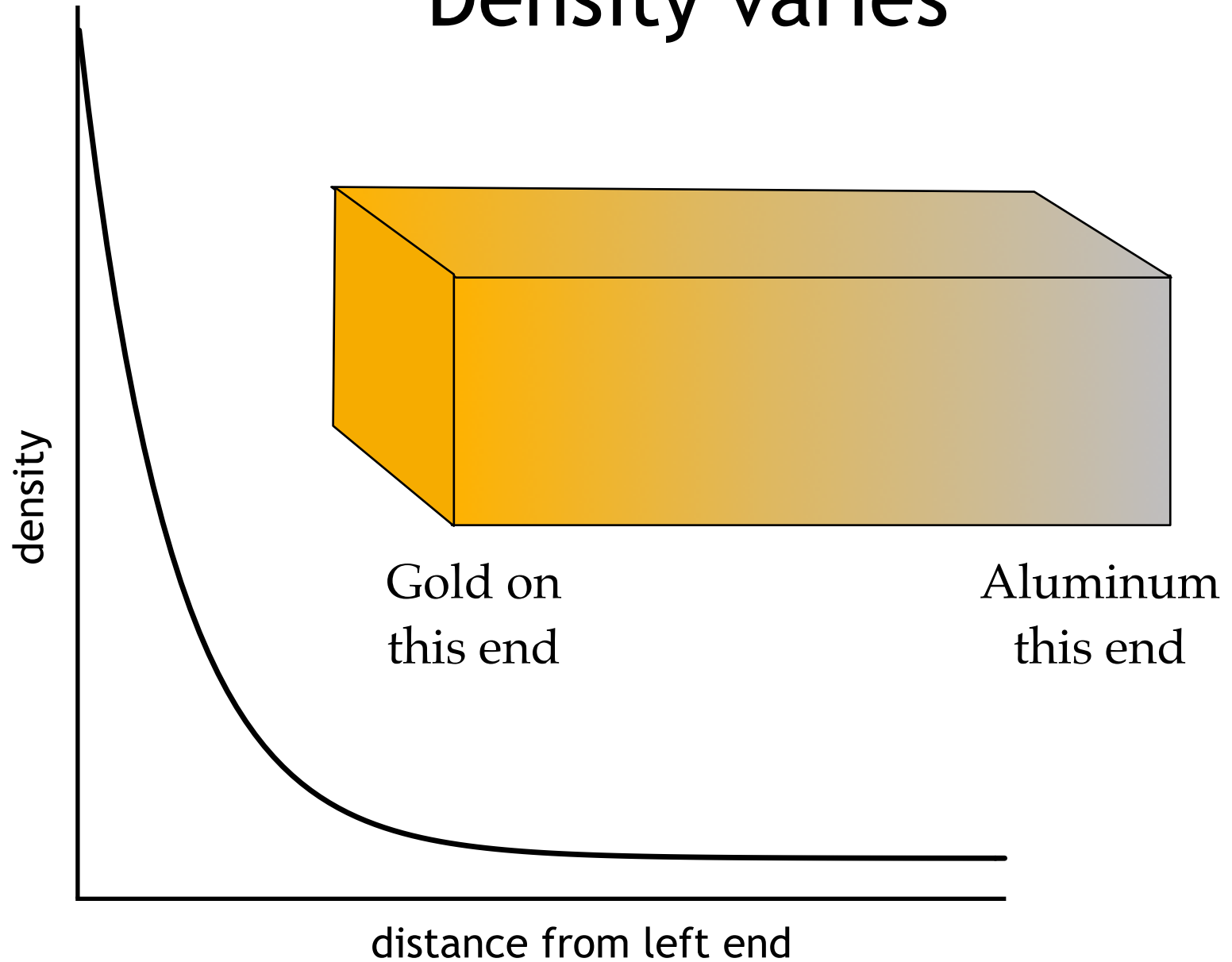
Densities of various substances

Substance	Density (g/cm ³)
Cork	0.24
Aluminum	2.7
Gold	19.3

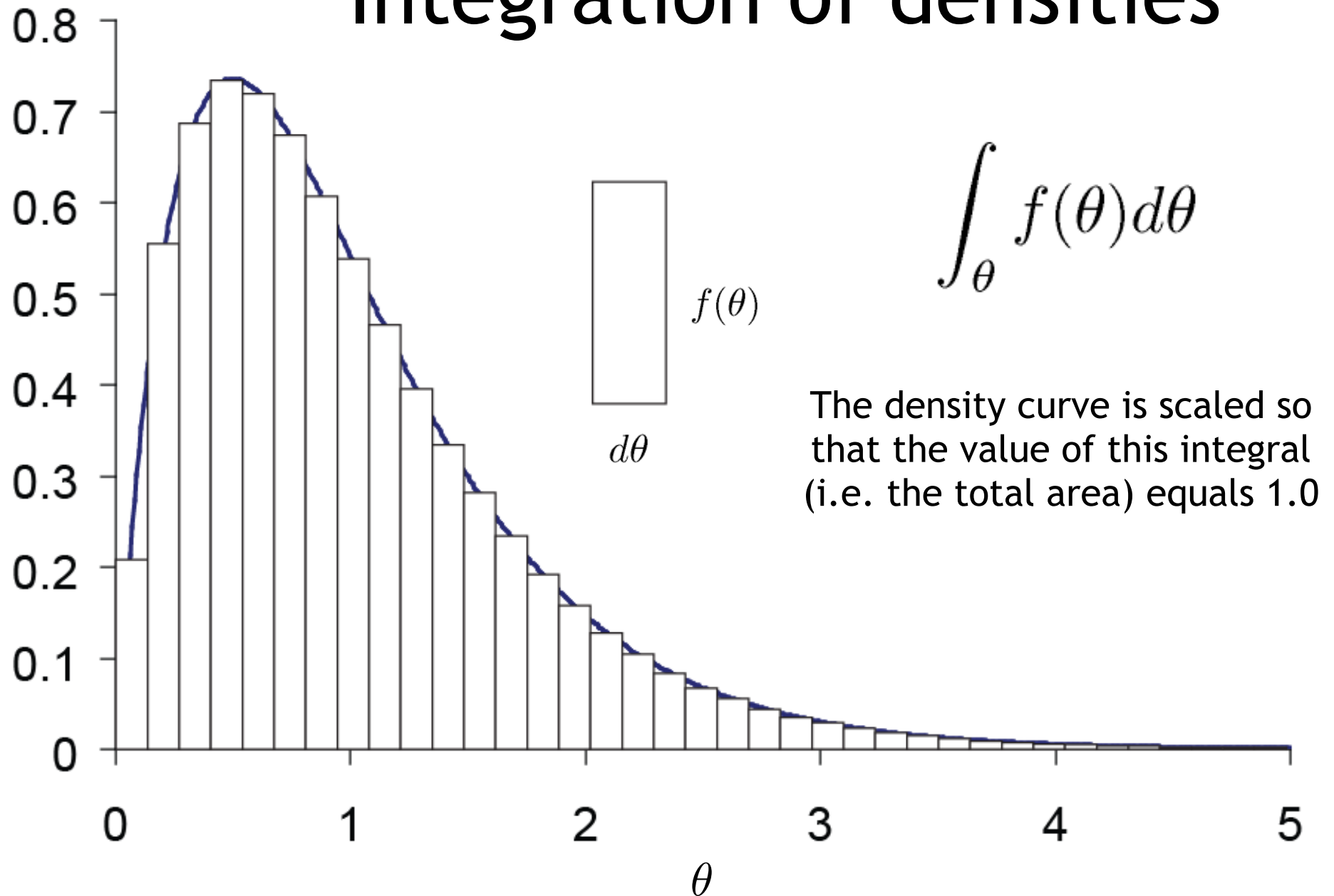
Density does not equal mass
mass = density × volume

Note: *volume* is appropriate for objects of dimension 3 or higher
For 2-dimensions, *area* takes the place of volume
For 1-dimension, *linear distance* replaces volume.

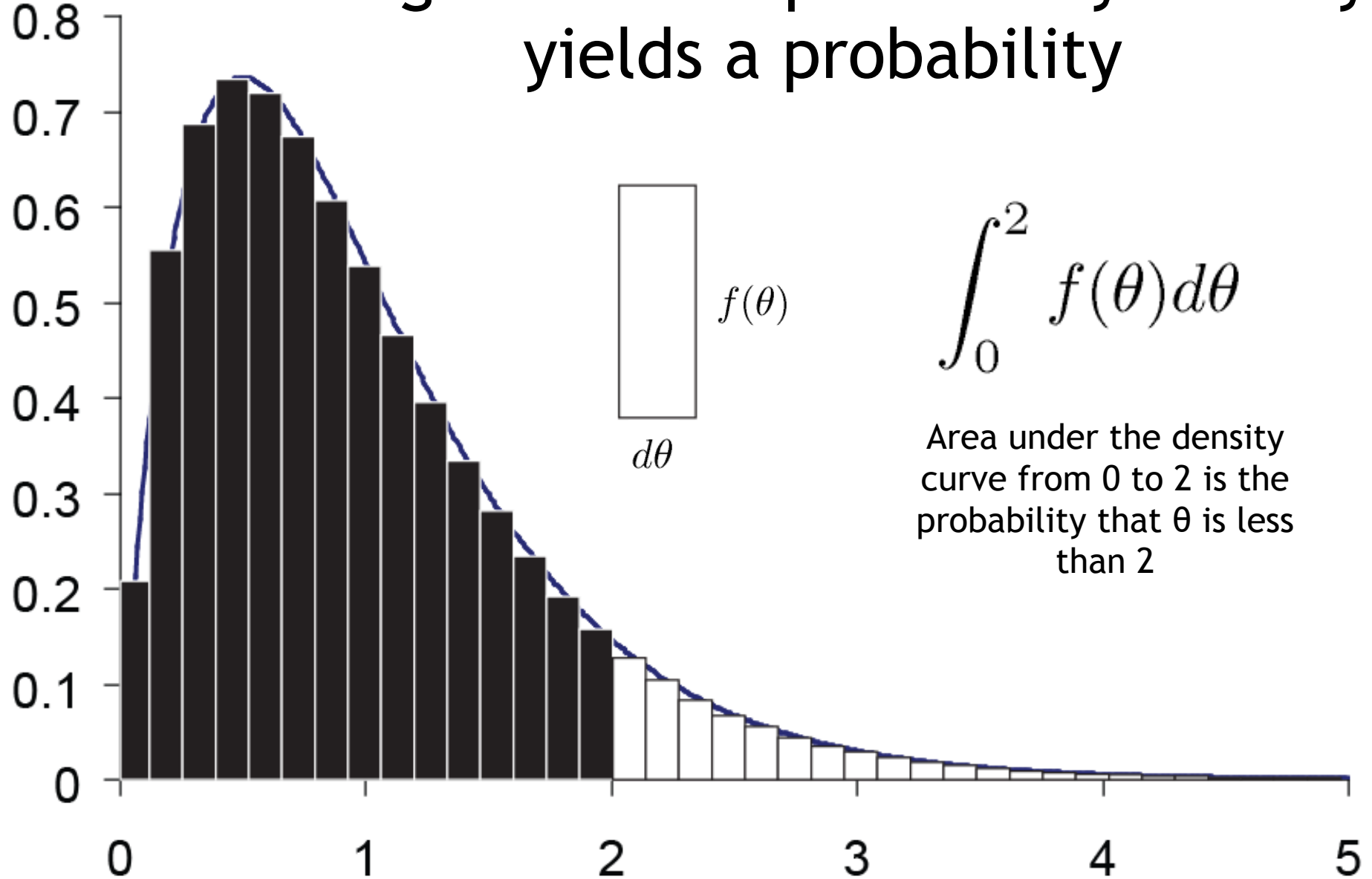
Density varies



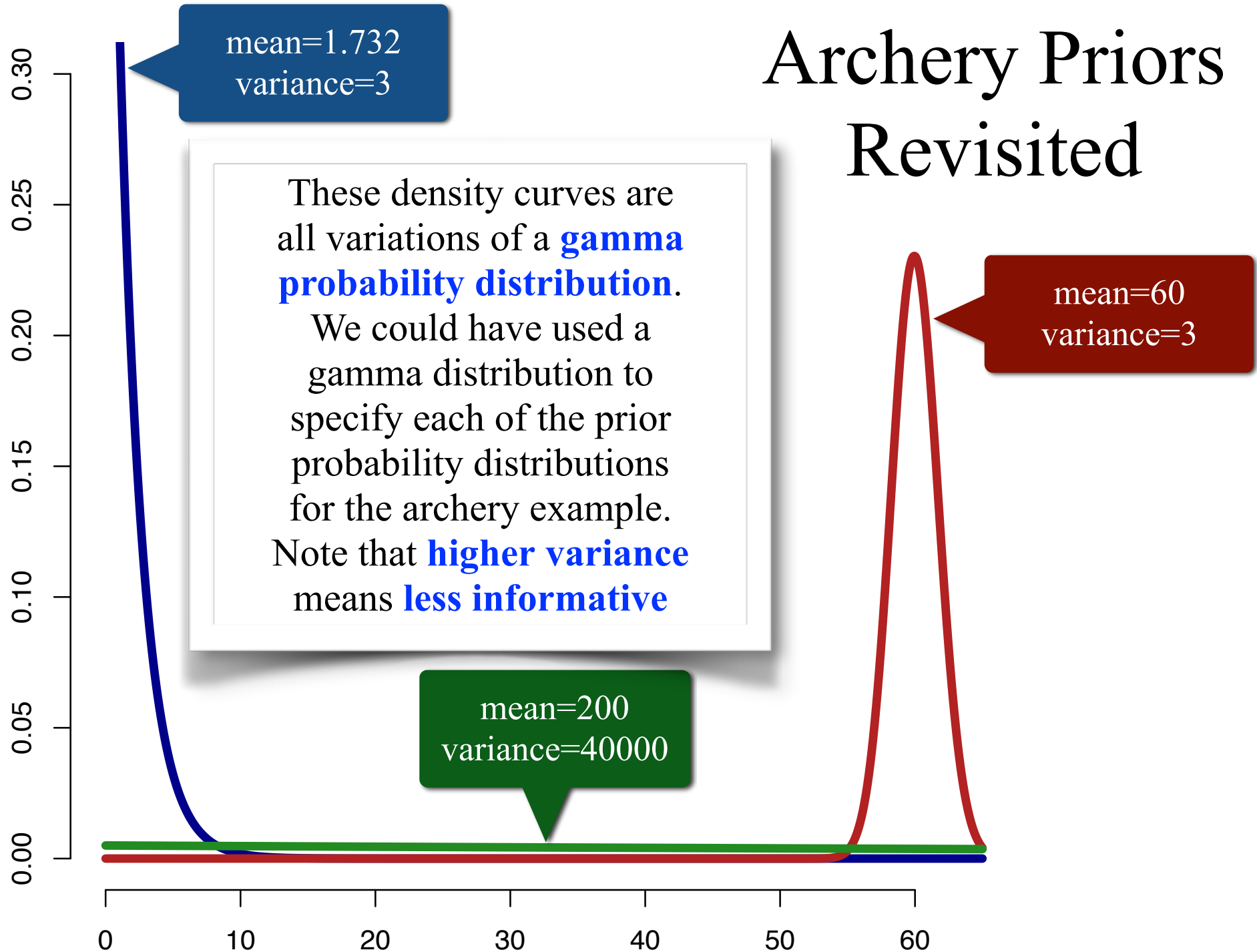
Integration of densities



Integration of a probability density yields a probability



Archery Priors Revisited



Coin-flipping

y = observed number of heads

n = number of flips (sample size)

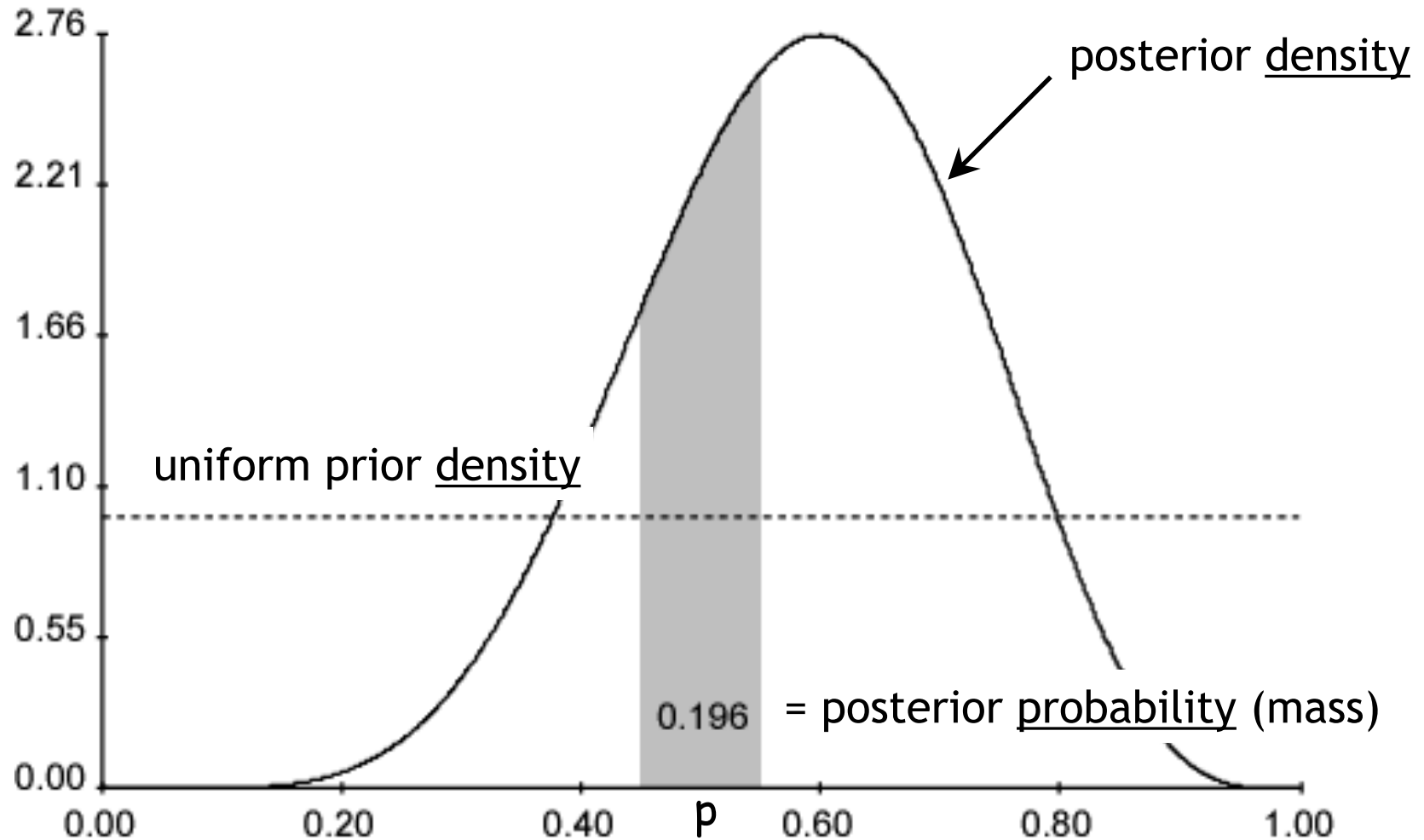
p = (unobserved) proportion of heads

$$\Pr(y|p) = \binom{n}{y} p^y (1 - p)^{n-y}$$

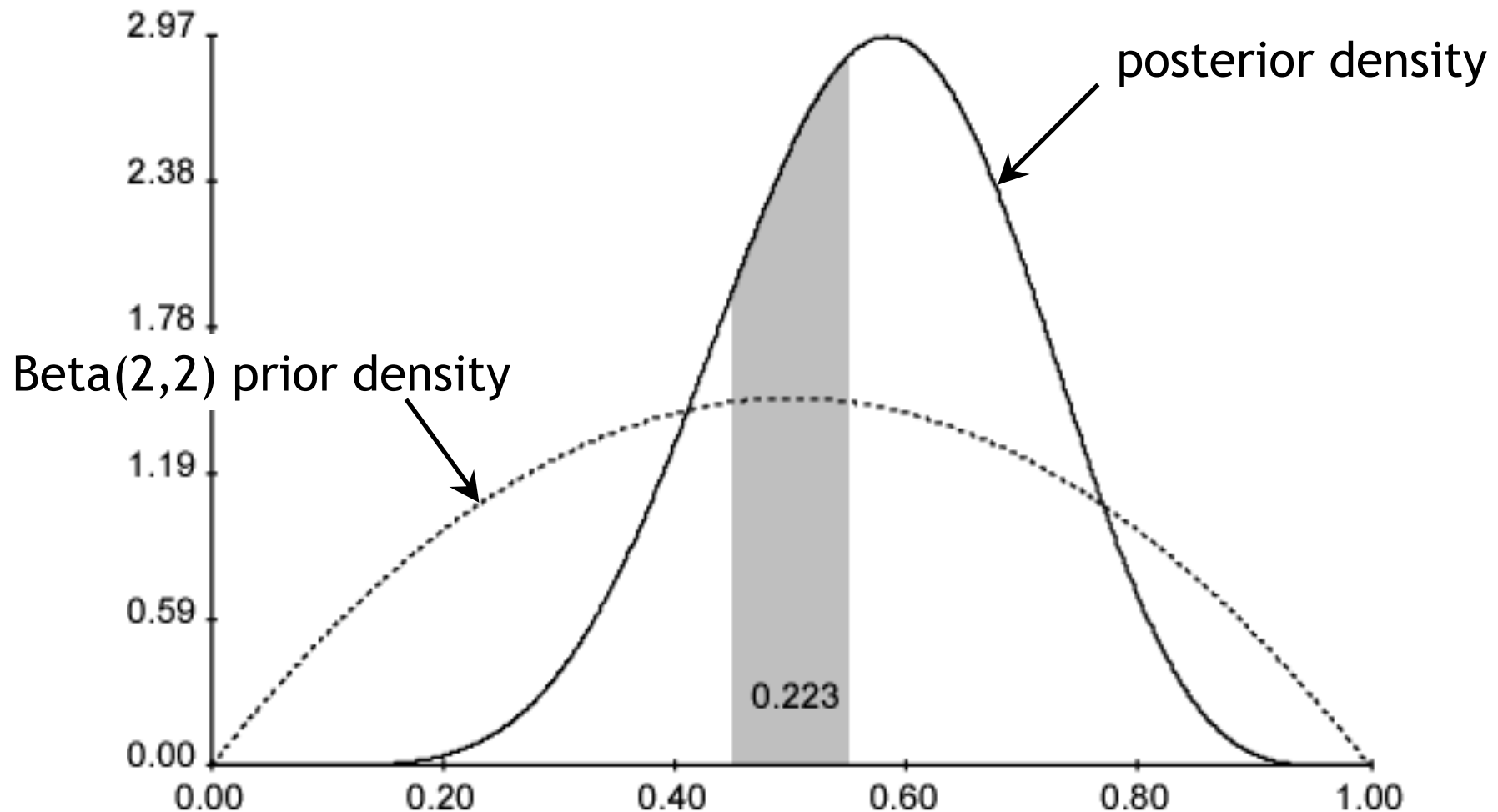
Note that the same formula serves as both the:

- probability of y (if p is fixed)
- likelihood of p (if y is fixed)

The posterior is (almost always) more informative than the prior



Beta(2,2) prior is vague but not flat



Posterior probability of p between 0.45 and 0.55 is **0.223**

Usually there are many parameters...

A 2-parameter example

Prior probability density

Likelihood

$$f(\theta, \phi | D) = \frac{f(D | \theta, \phi) f(\theta) f(\phi)}{\int_{\theta} \int_{\phi} f(D | \theta, \phi) f(\theta) f(\phi) d\theta d\phi}$$

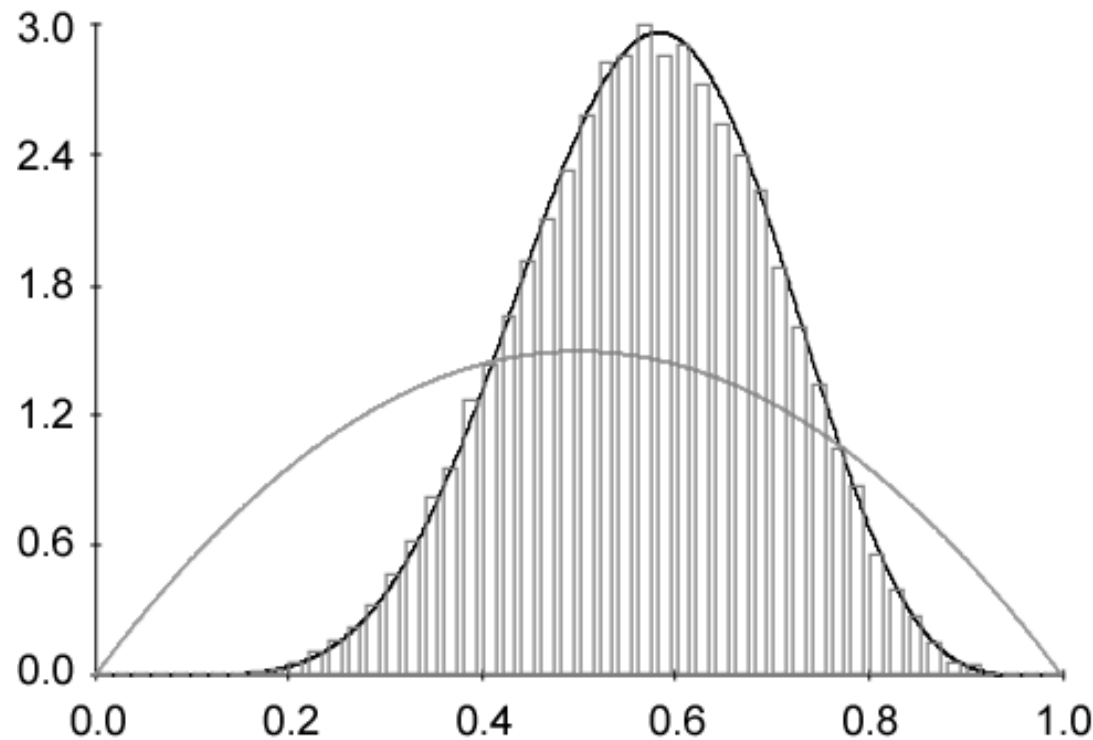
Posterior probability density

Marginal probability of data

An analysis of **100 sequences** under the simplest model (JC69) requires 197 branch length parameters. The denominator is a **197-fold integral** in this case! Now consider summing over **all possible tree topologies**! It would thus be nice to avoid having to calculate the marginal probability of the data...

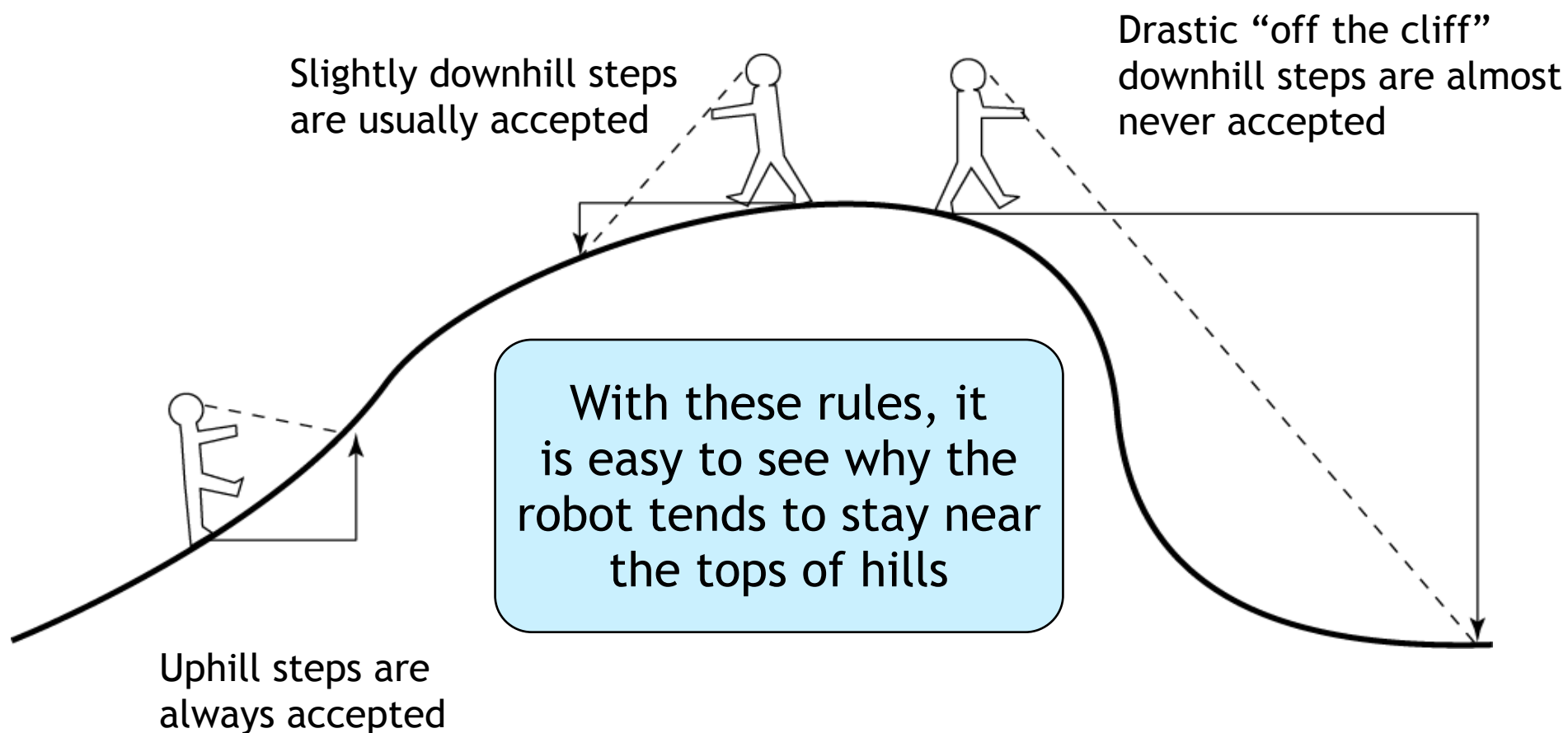
II. Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC)

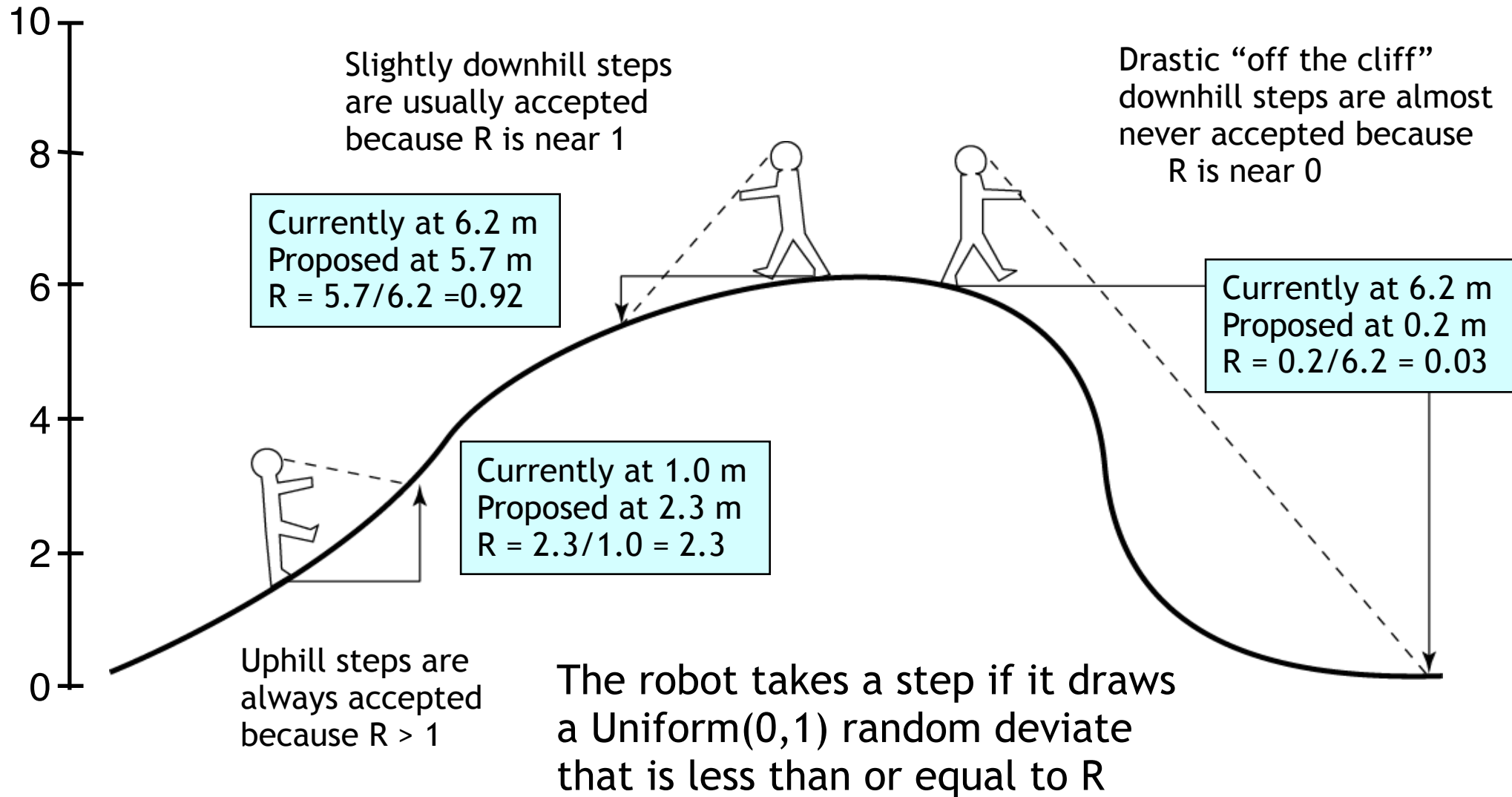


For more complex problems, we might settle for a
good approximation
to the posterior distribution

MCMC robot's rules



(Actual) MCMC robot rules



Cancellation of marginal likelihood

When calculating the ratio R of posterior densities, the marginal probability of the data cancels.

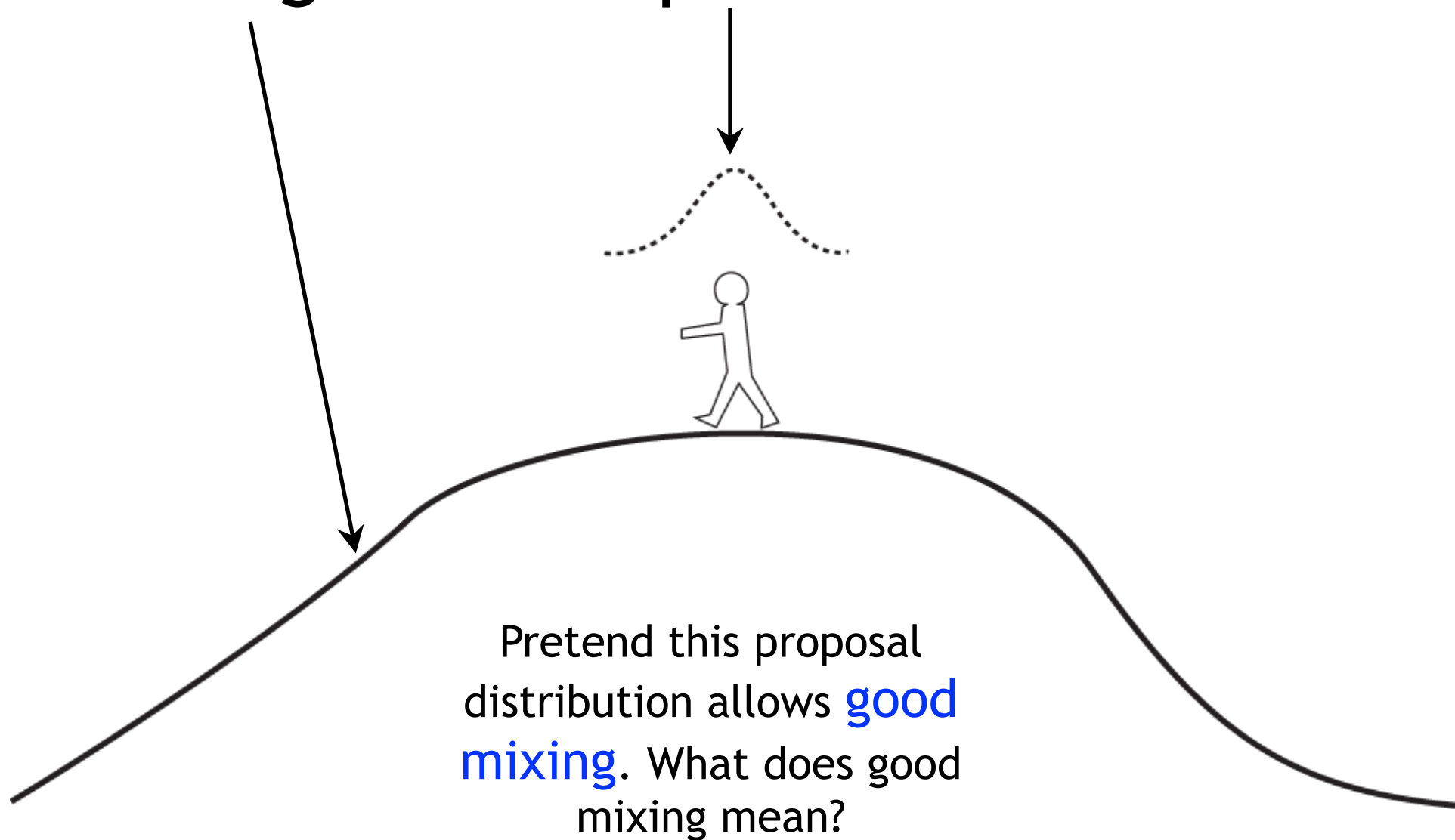
$$\frac{f(\theta^* | D)}{f(\theta | D)} = \frac{\frac{f(D|\theta^*)f(\theta^*)}{\cancel{f(D)}}}{\frac{f(D|\theta)f(\theta)}{\cancel{f(D)}}} = \frac{f(D|\theta^*)f(\theta^*)}{f(D|\theta)f(\theta)}$$

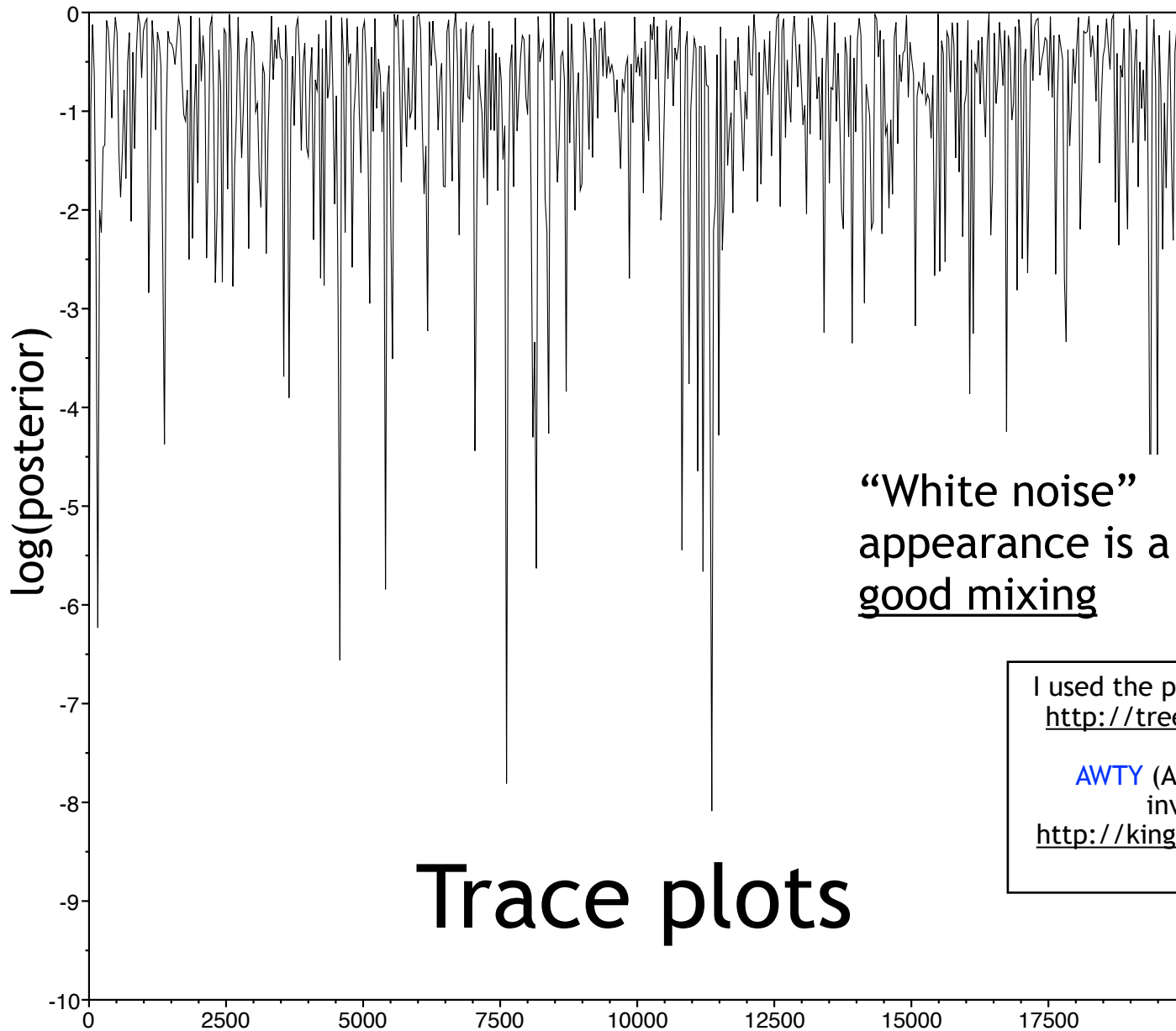
Posterior
odds

Likelihood
ratio

Prior odds

Target vs. Proposal Distributions





“White noise”
appearance is a sign of
good mixing

I used the program [Tracer](http://tree.bio.ed.ac.uk/software/tracer/) to create this plot:
<http://tree.bio.ed.ac.uk/software/tracer/>

[AWTY](http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php) (Are We There Yet?) is useful for
investigating convergence:
[http://king2.scs.fsu.edu/CEBProjects/awty/
awty_start.php](http://king2.scs.fsu.edu/CEBProjects/awty/awty_start.php)

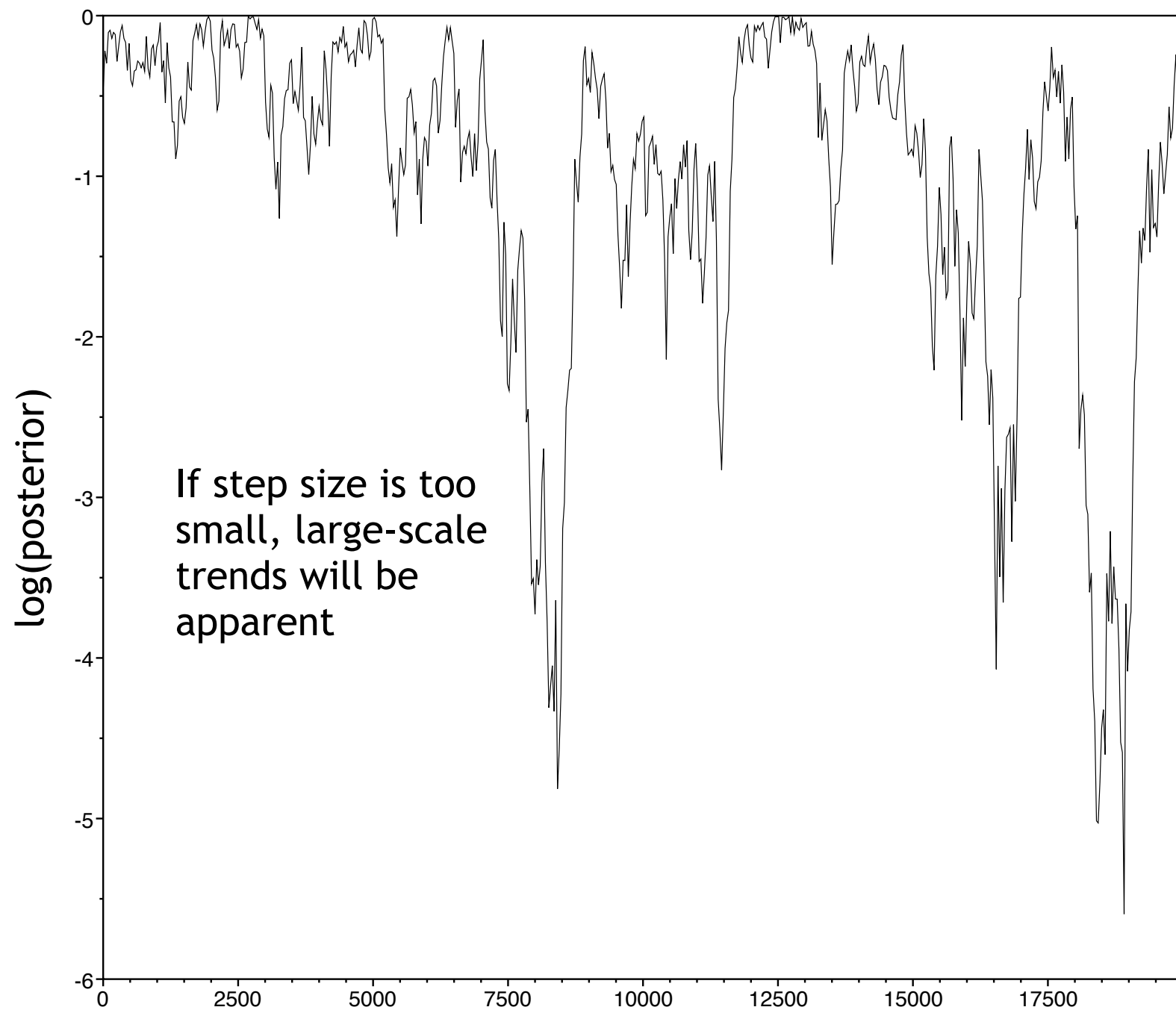
Target vs. Proposal Distributions

Proposal distributions
with **smaller variance**...



Disadvantage: robot takes smaller steps, more time required to explore the same area

Advantage: robot seldom refuses to take proposed steps

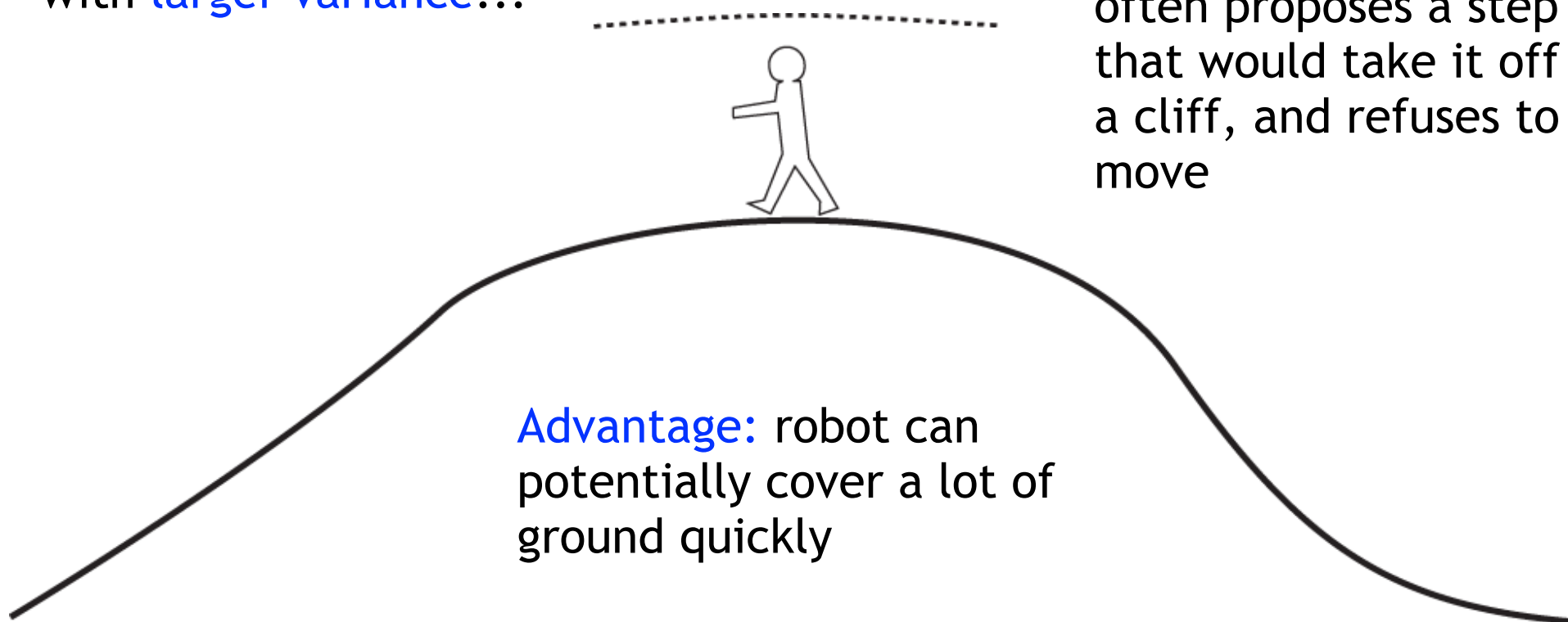


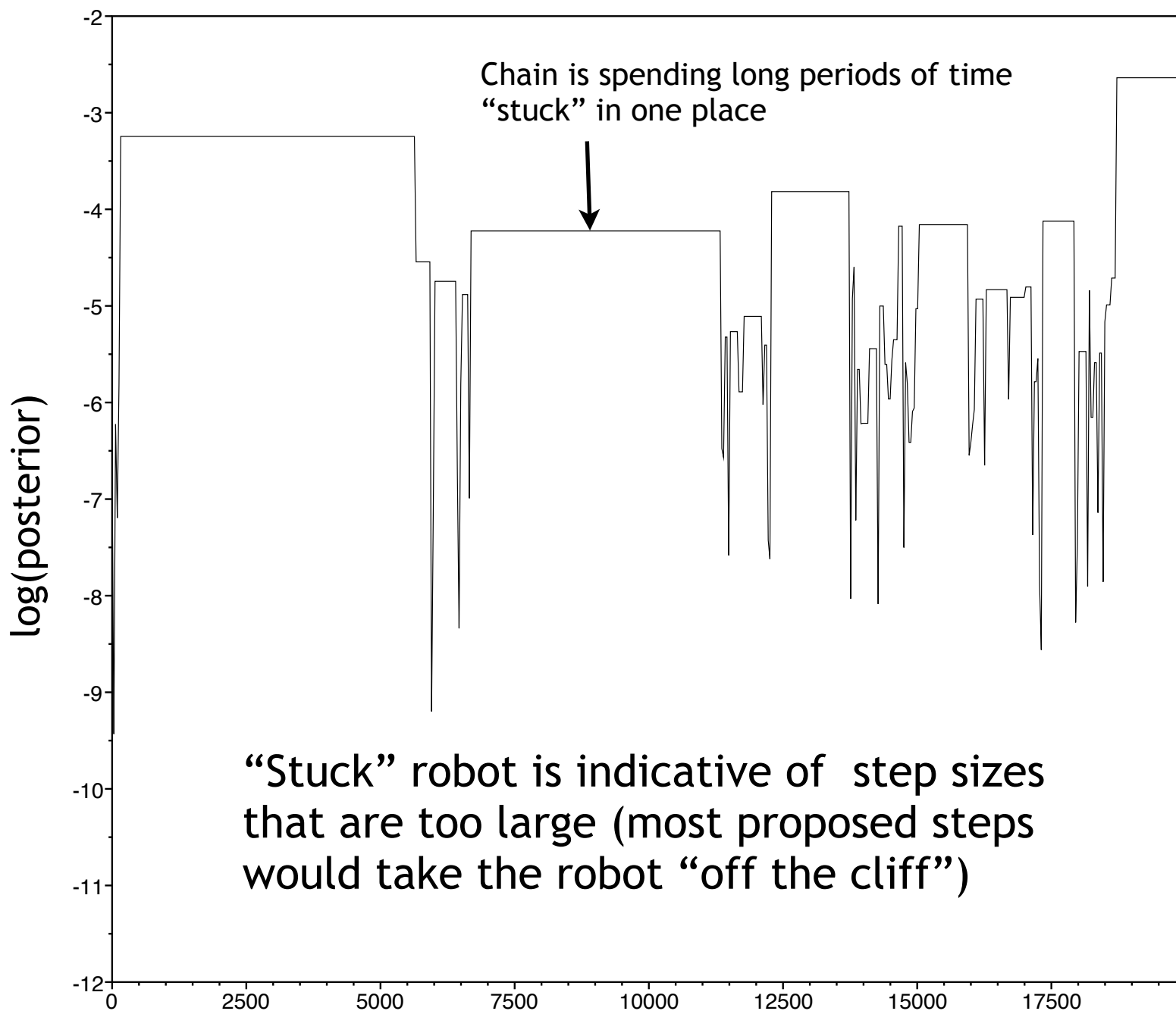
Target vs. Proposal Distributions

Proposal distributions with **larger variance**...

Disadvantage: robot often proposes a step that would take it off a cliff, and refuses to move

Advantage: robot can potentially cover a lot of ground quickly





MCRobot (or "MCMC Robot")

Free apps for **Windows** or **iPhone/iPad** available
from <http://mcmicrobot.org/>
(note: iOS 8 has caused some problems)

Android: hopefully by summer

Mac version: maybe some day
(but see John Huelsenbeck's iMCMC app for MacOS:
<http://cteg.berkeley.edu/software.html>)

Tradeoff

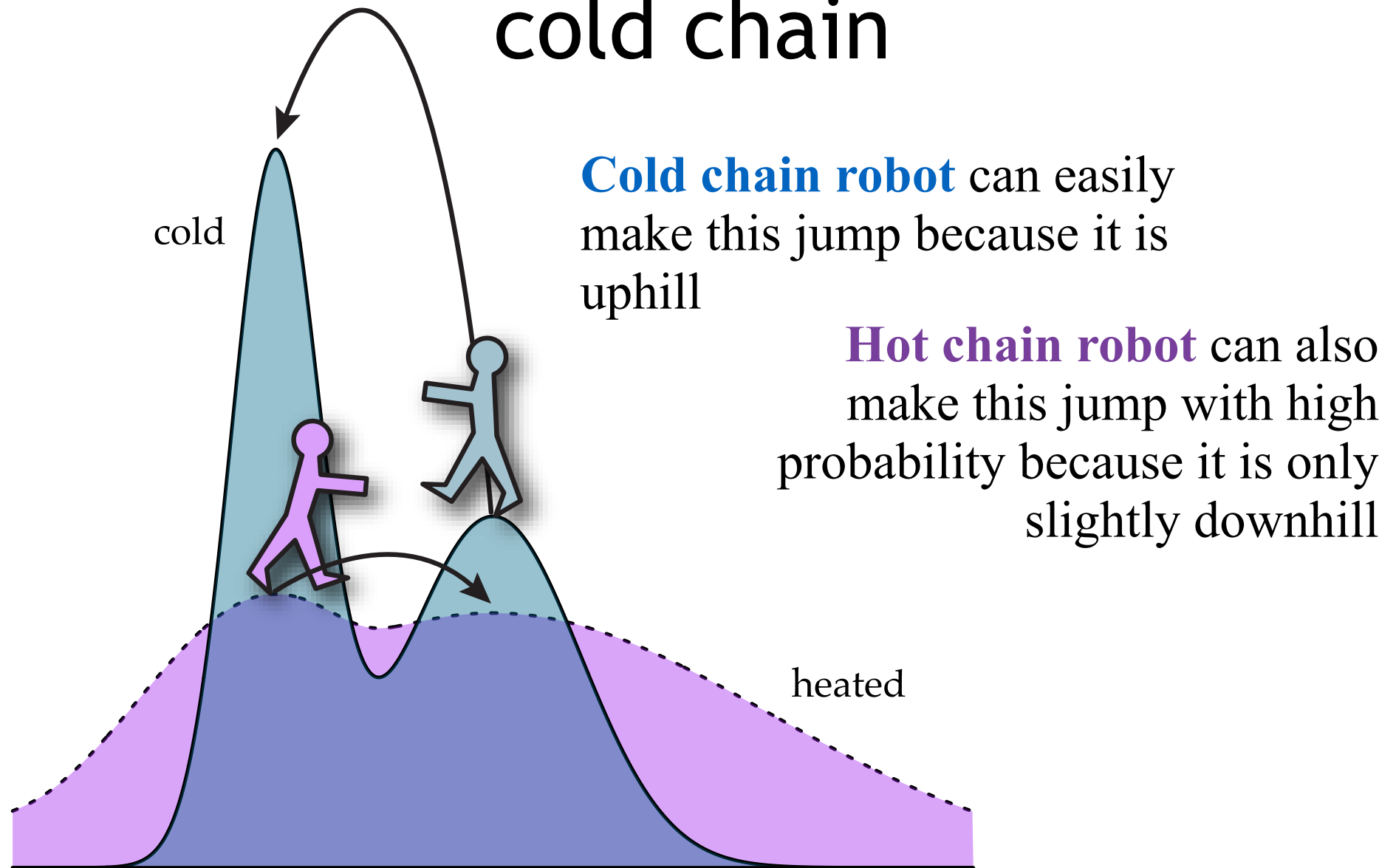
- Taking **big steps** helps in jumping from one “island” in the posterior density to another
- Taking **small steps** often results in better mixing
- How can we overcome this tradeoff? **MCMCMC**

Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

- MCMCMC involves running **several chains simultaneously**
- The **cold chain** is the one that counts, the rest are **heated chains**
- Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)

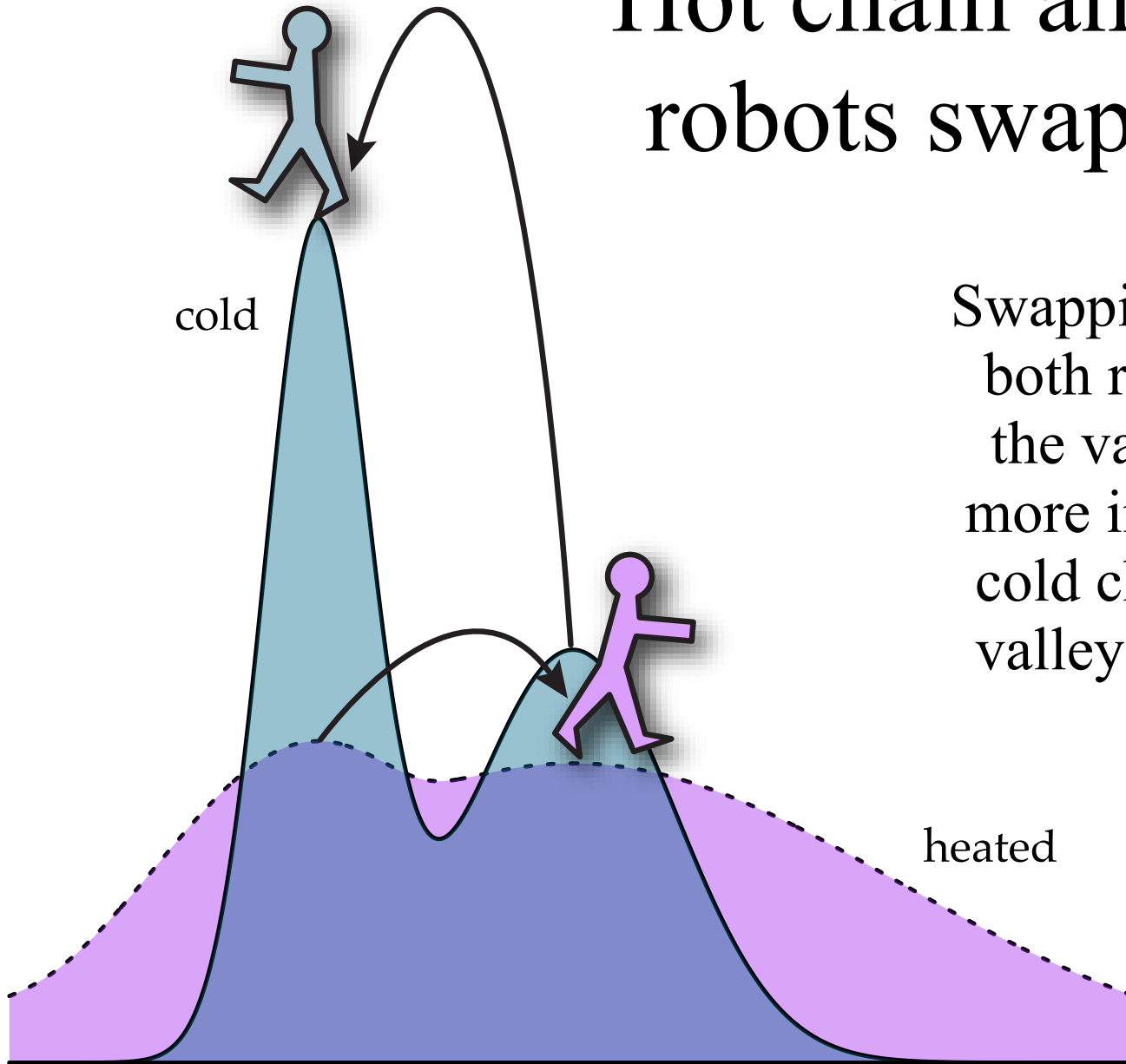
Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood for dependent data. Pages 156-163 *in* Computing Science and Statistics (E. Keramidas, ed.).

Heated chains act as scouts for the cold chain



Hot chain and cold chain robots swapping places

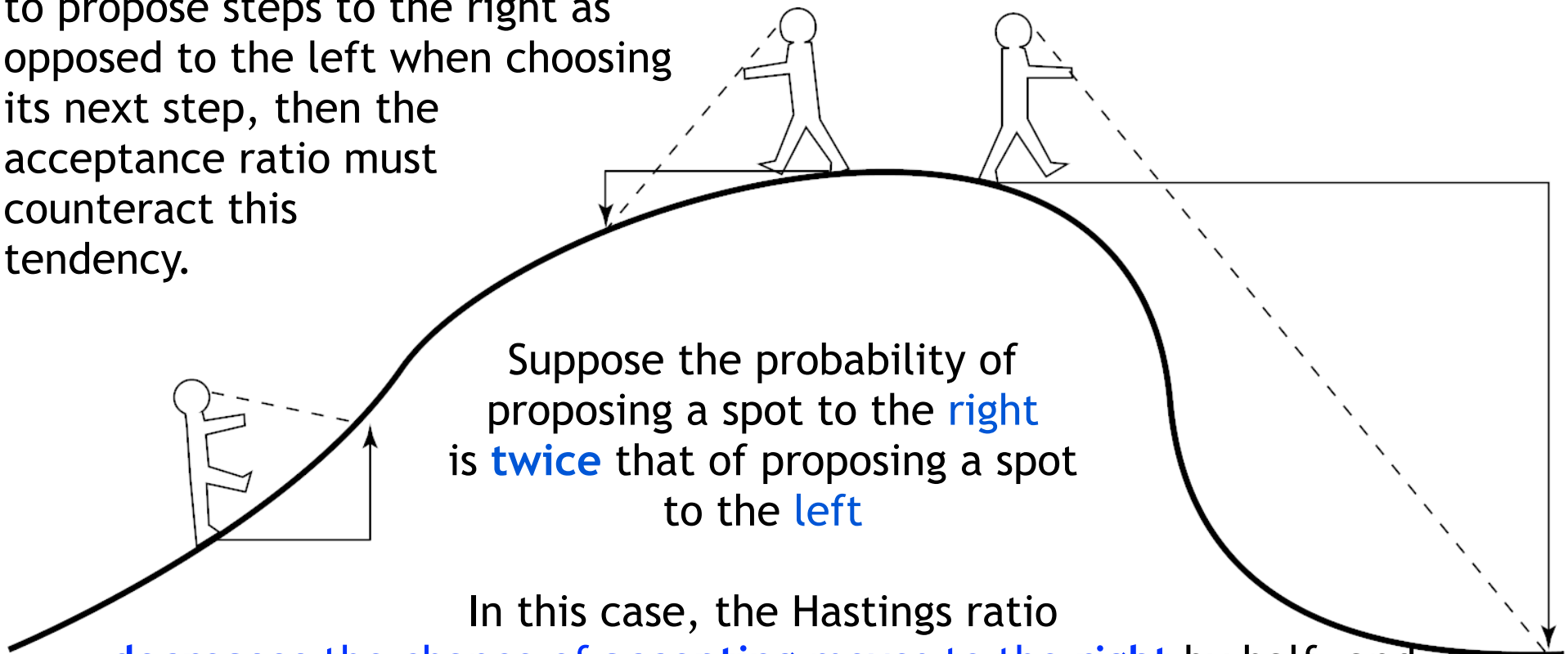
Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper



Back to MCRobot...

The Hastings ratio

If robot has a greater tendency to propose steps to the right as opposed to the left when choosing its next step, then the acceptance ratio must counteract this tendency.

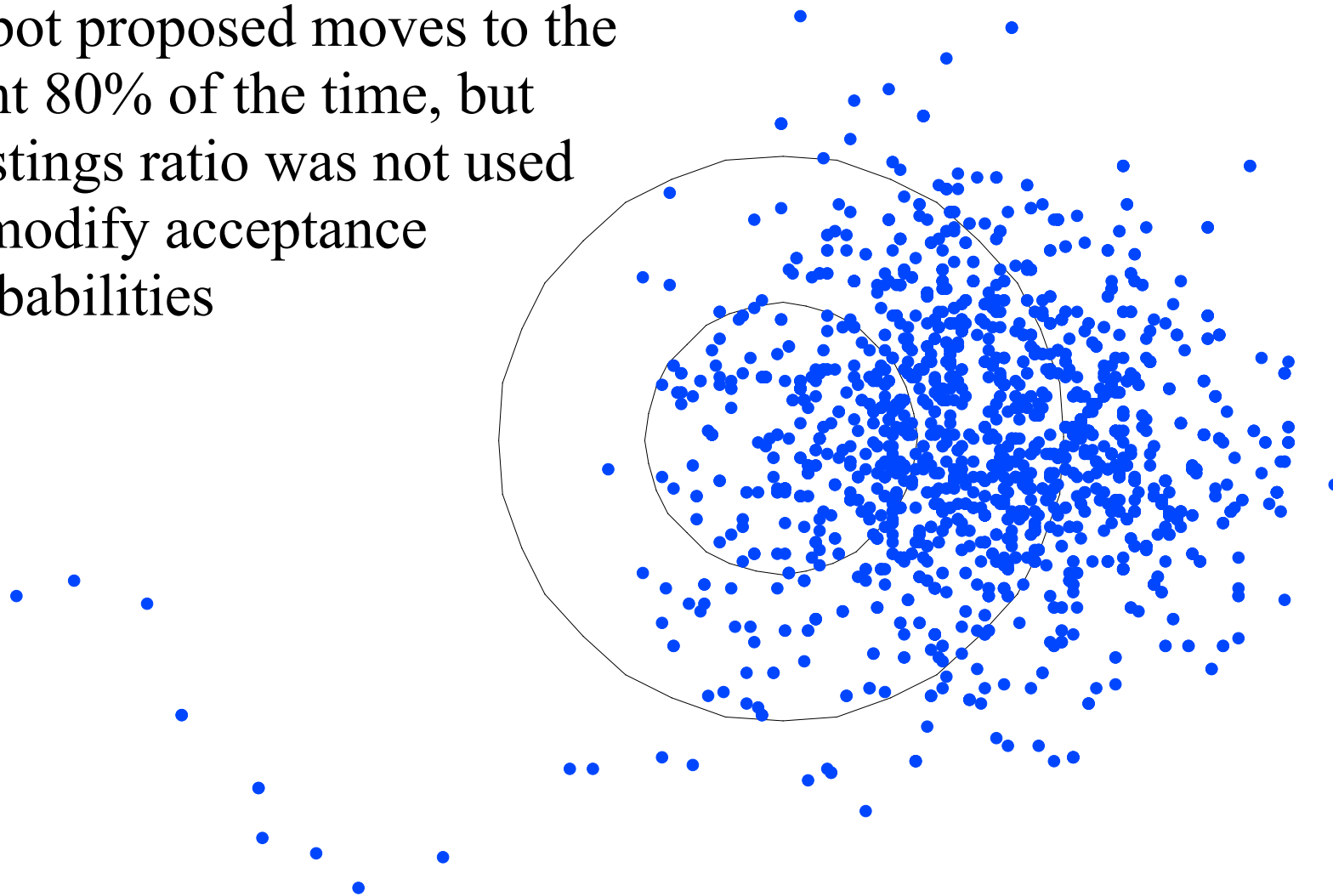


Suppose the probability of proposing a spot to the **right** is **twice** that of proposing a spot to the **left**

In this case, the Hastings ratio **decreases the chance of accepting moves to the right** by half, and **increases the chance of accepting moves to the left** (by a factor of 2), thus **exactly compensating** for the asymmetry in the proposal distribution.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Example where MCMC
Robot proposed moves to the
right 80% of the time, but
Hastings ratio was not used
to modify acceptance
probabilities



Hastings Ratio

$$R = \left[\frac{f(D|\theta^*) f(\theta^*)}{f(D|\theta) f(\theta)} \right] \left[\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)} \right]$$

Acceptance
ratio

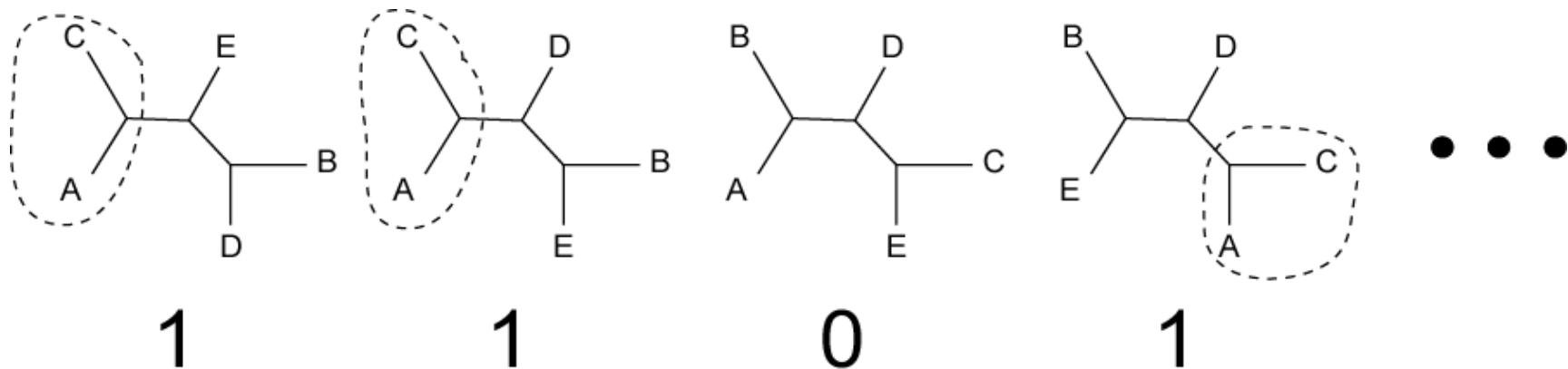
Posterior ratio

Hastings ratio

Note that if $q(\theta|\theta^*) = q(\theta^*|\theta)$, the Hastings ratio is 1

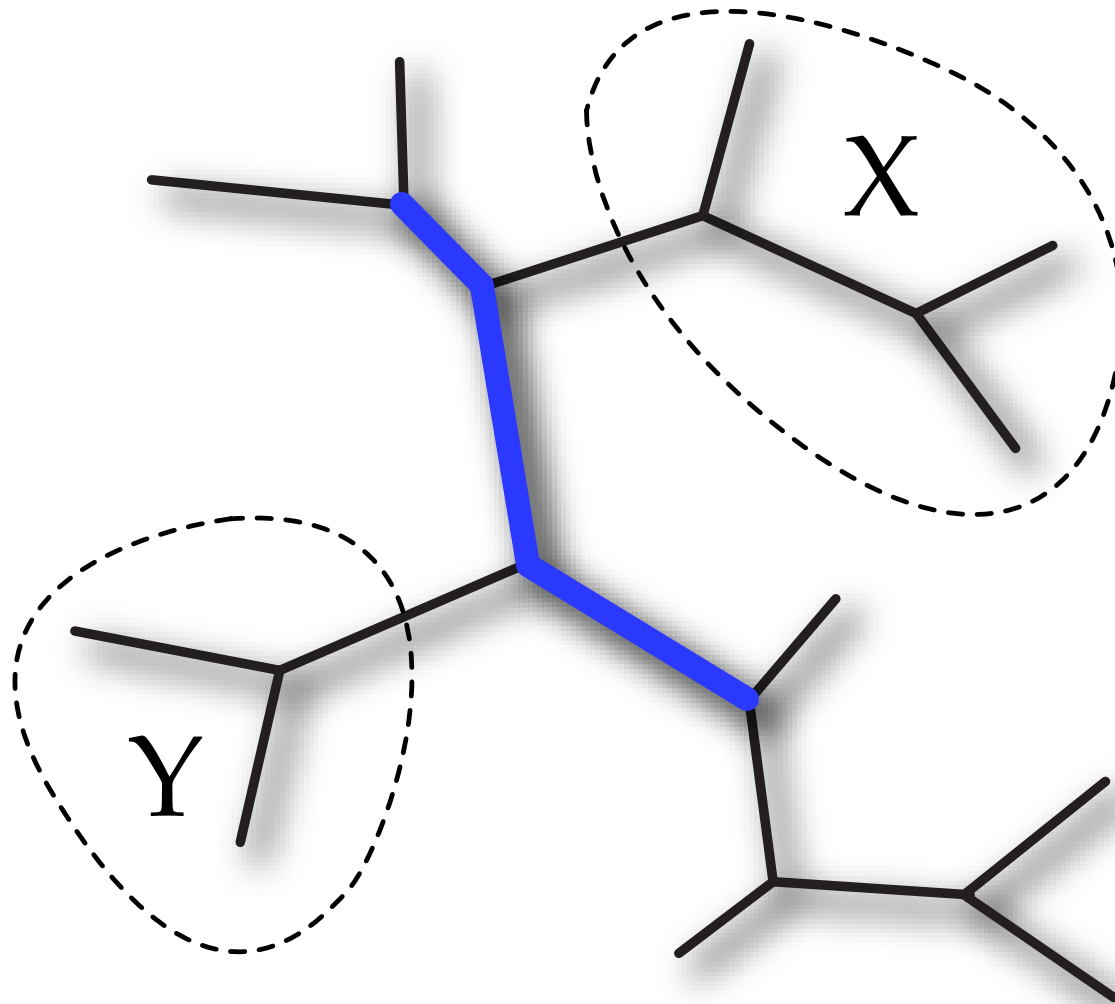
III. Bayesian phylogenetics

So, what's all this got to do with phylogenetics?



Imagine pulling out trees at random from a barrel. In the barrel, some trees are represented numerous times, while other possible trees are not present. Count 1 each time you see the split separating just A and C from the other taxa, and count 0 otherwise. Dividing by the total trees sampled approximates the **true proportion of that split in the barrel**.

Moving through treespace



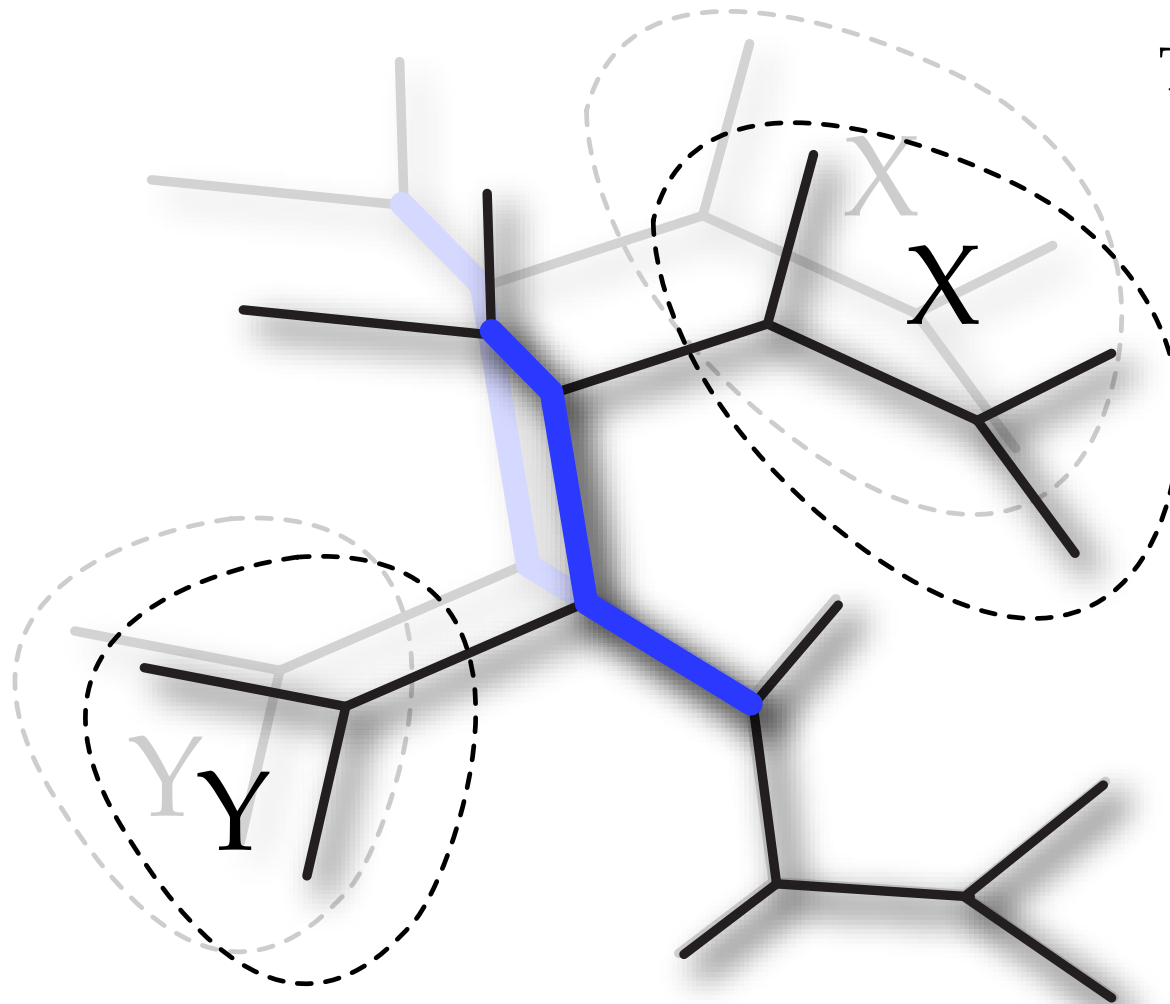
The Larget-Simon move

Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

*Larget, B., and D. L. Simon. 1999. Markov chain monte carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16: 750-759. See also: Holder et al. 2005. *Syst. Biol.* 54:

Moving through treespace



The Target-Simon move

Step 1:

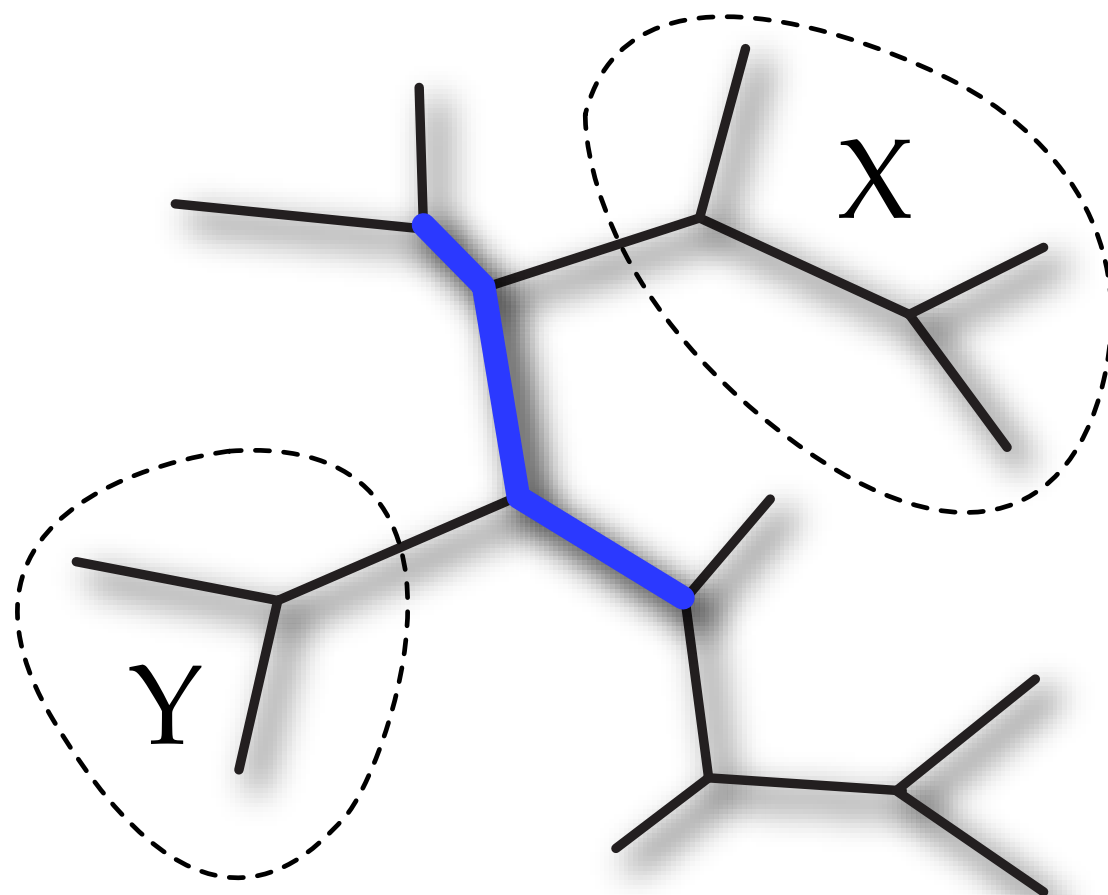
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:

Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

The Larget-Simon move



Step 1:

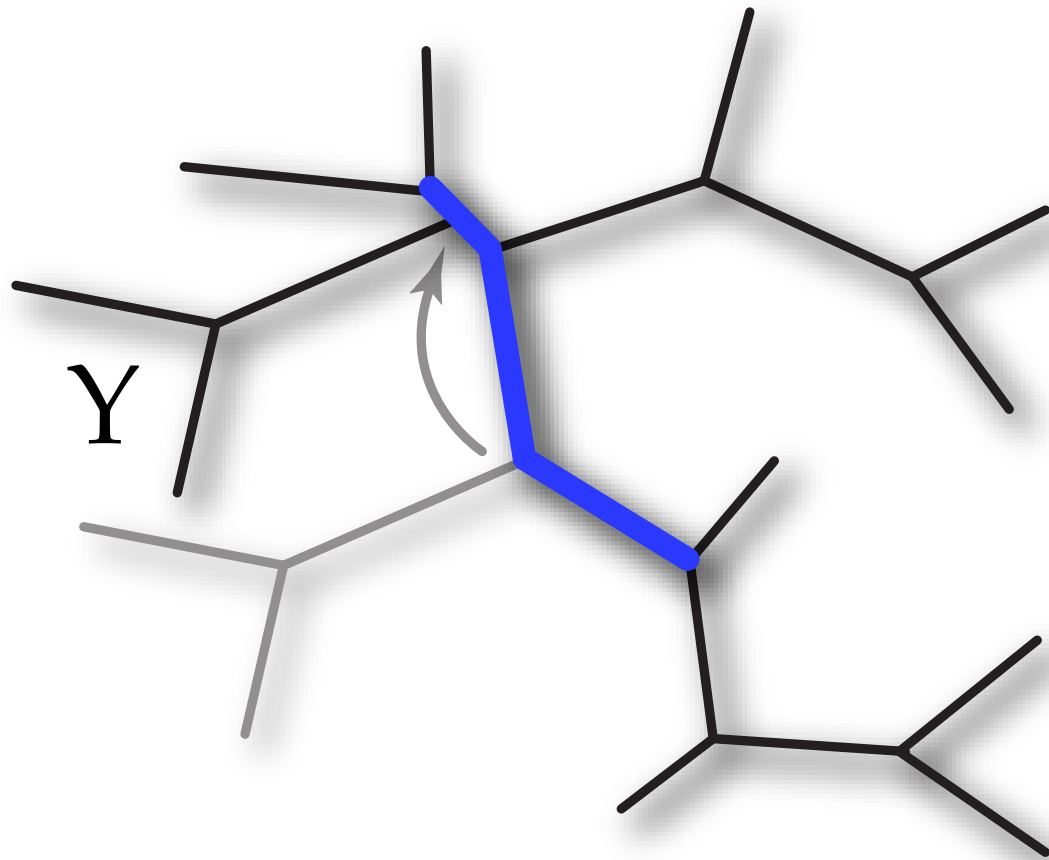
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:

Shrink or grow selected 3-edge segment by a random amount

Moving through treespace

The Larget-Simon move



Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

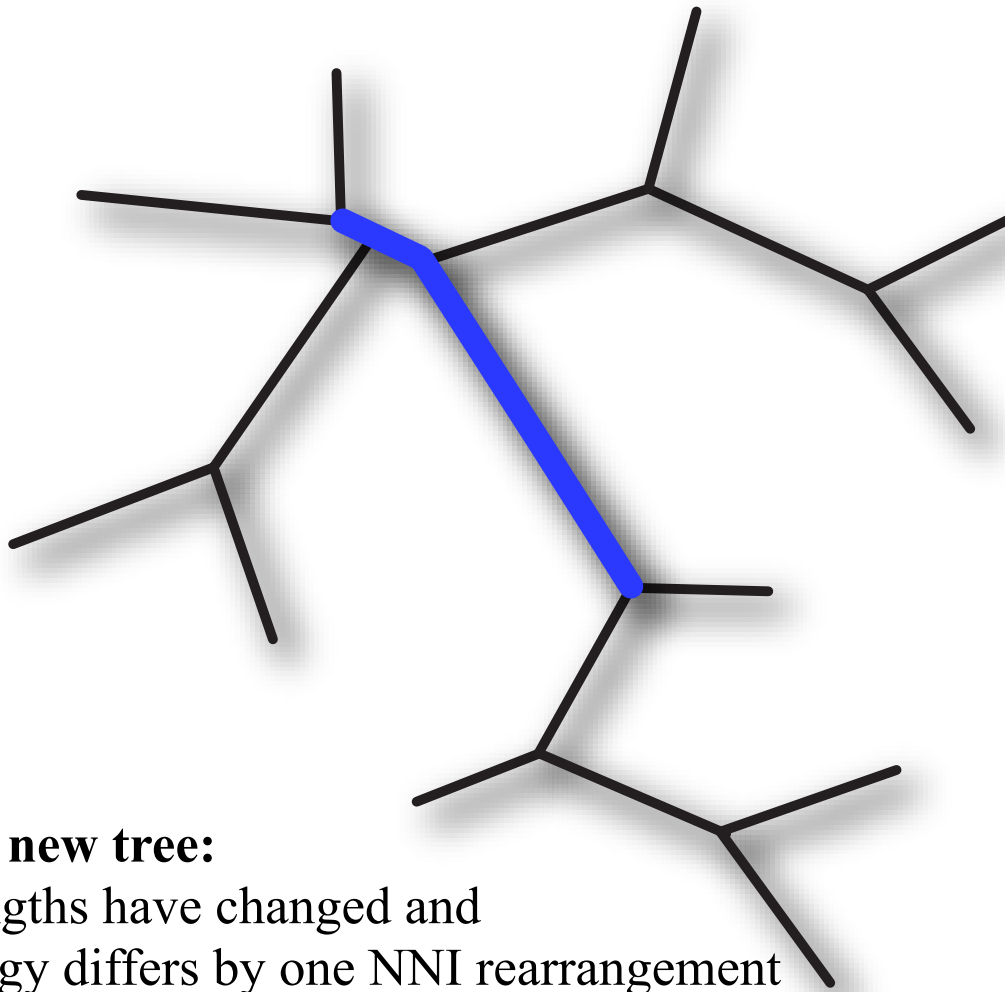
Step 2:

Shrink or grow selected 3-edge segment by a random amount

Step 3:

Choose X or Y randomly, then reposition randomly

Moving through treespace



Proposed new tree:
3 edge lengths have changed and
the topology differs by one NNI rearrangement

The Target-Simon move

Step 1:

Pick 3 contiguous edges randomly, defining two subtrees, X and Y

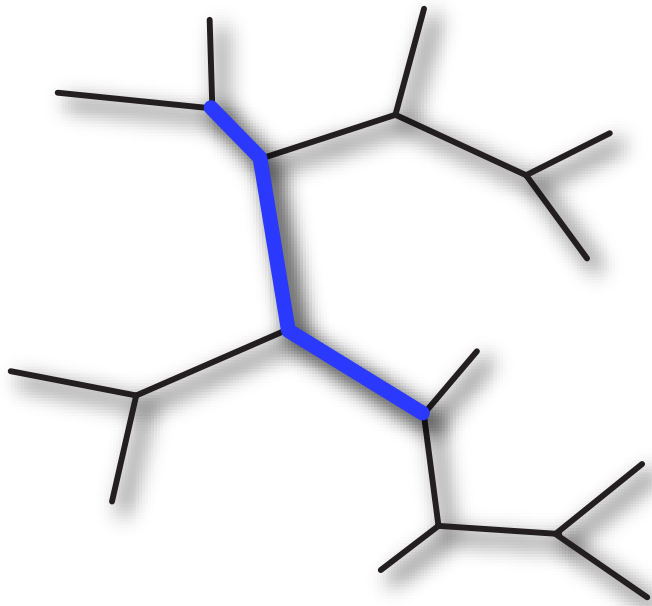
Step 2:

Shrink or grow selected 3-edge segment by a random amount

Step 3:

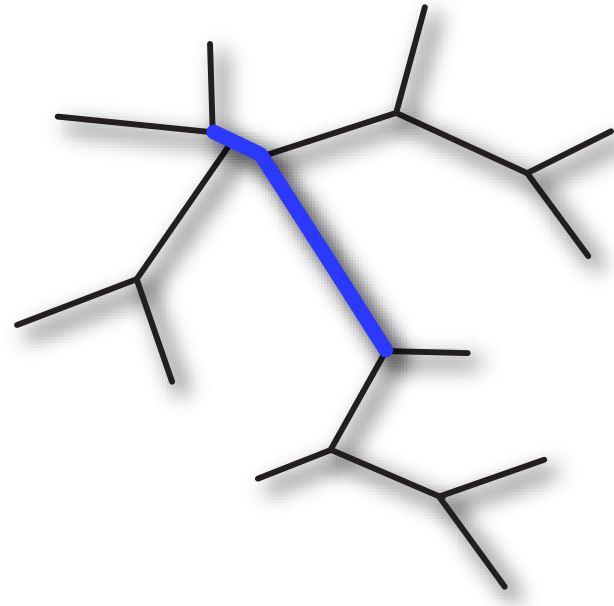
Choose X or Y randomly, then reposition randomly

Moving through treespace



Current tree

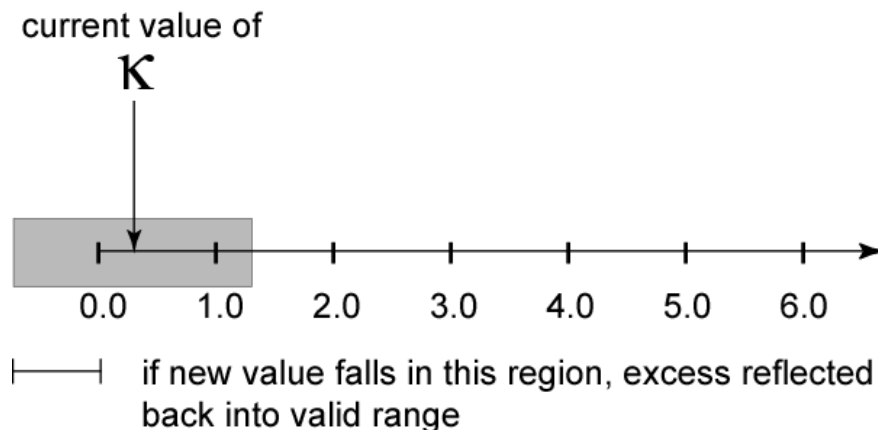
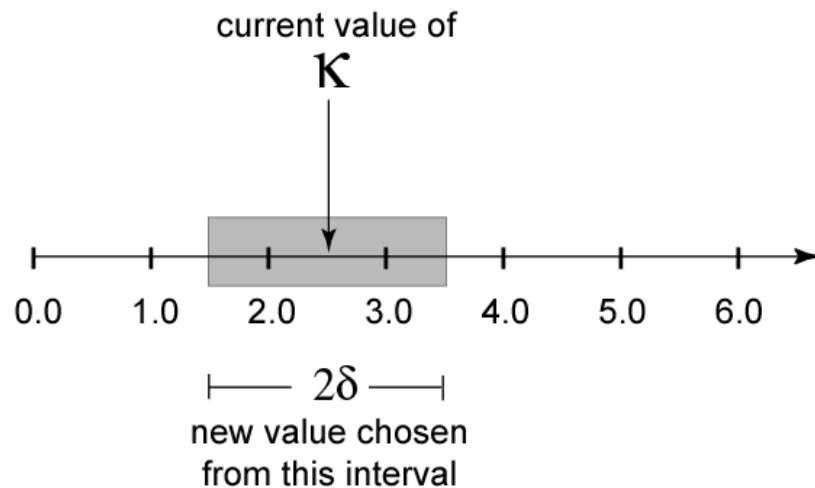
log-posterior = -34256



Proposed tree

log-posterior = -32519
(better, so accept)

Moving through parameter space



Using κ (ratio of the transition rate to the transversion rate) as an example of a model parameter.

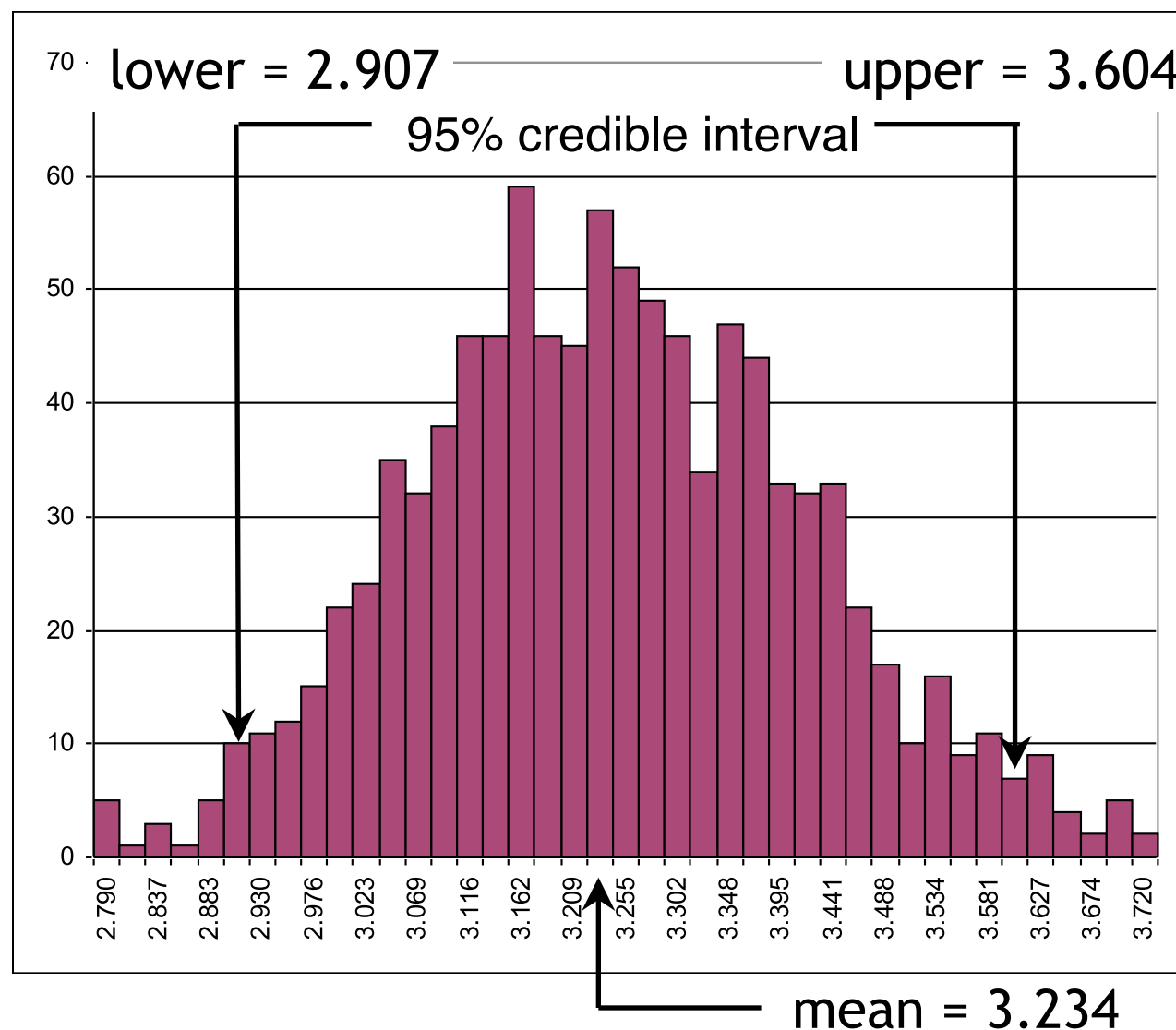
Proposal distribution is the uniform distribution on the interval $(\kappa - \delta, \kappa + \delta)$

The “step size” of the MCMC robot is defined by δ : a larger δ means that the robot will attempt to make larger jumps on average.

Putting it all together

- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
 - Propose a **new tree** (e.g. Largert-Simon move) and either accept or reject the move
 - Propose (and either accept or reject) a **new model parameter value**
- Every k generations, save tree topology, branch lengths and all model parameters (i.e. **sample the chain**)
- After n generations, **summarize sample** using histograms, means, credible intervals, etc.

Marginal Posterior Distribution of κ



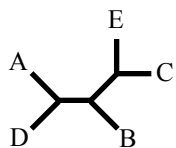
Histogram created from a sample of 1000 kappa values.

IV. Prior distributions

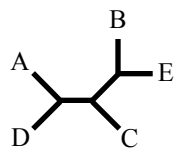
Common Priors

- **Discrete uniform** for topologies
 - exceptions becoming more common
- **Beta** for proportions
- **Gamma** or **Log-normal** for branch lengths and other parameters with support $[0, \infty)$
 - Exponential is common special case of the gamma distribution
- **Dirichlet** for state frequencies and GTR relative rates

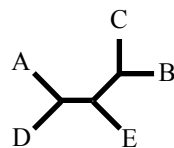
Discrete Uniform distribution for topologies



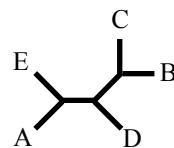
$$\frac{1}{15}$$



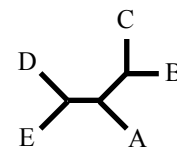
$$\frac{1}{15}$$



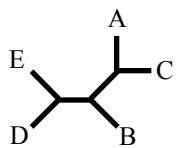
$$\frac{1}{15}$$



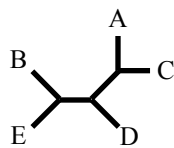
$$\frac{1}{15}$$



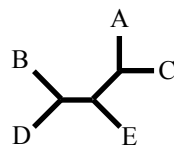
$$\frac{1}{15}$$



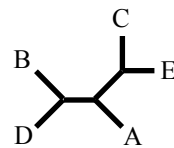
$$\frac{1}{15}$$



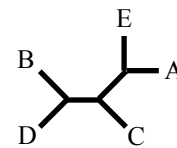
$$\frac{1}{15}$$



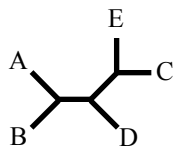
$$\frac{1}{15}$$



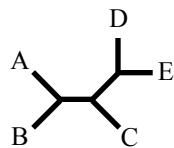
$$\frac{1}{15}$$



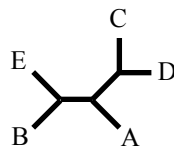
$$\frac{1}{15}$$



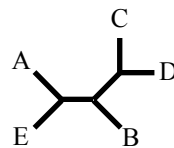
$$\frac{1}{15}$$



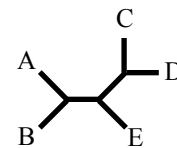
$$\frac{1}{15}$$



$$\frac{1}{15}$$

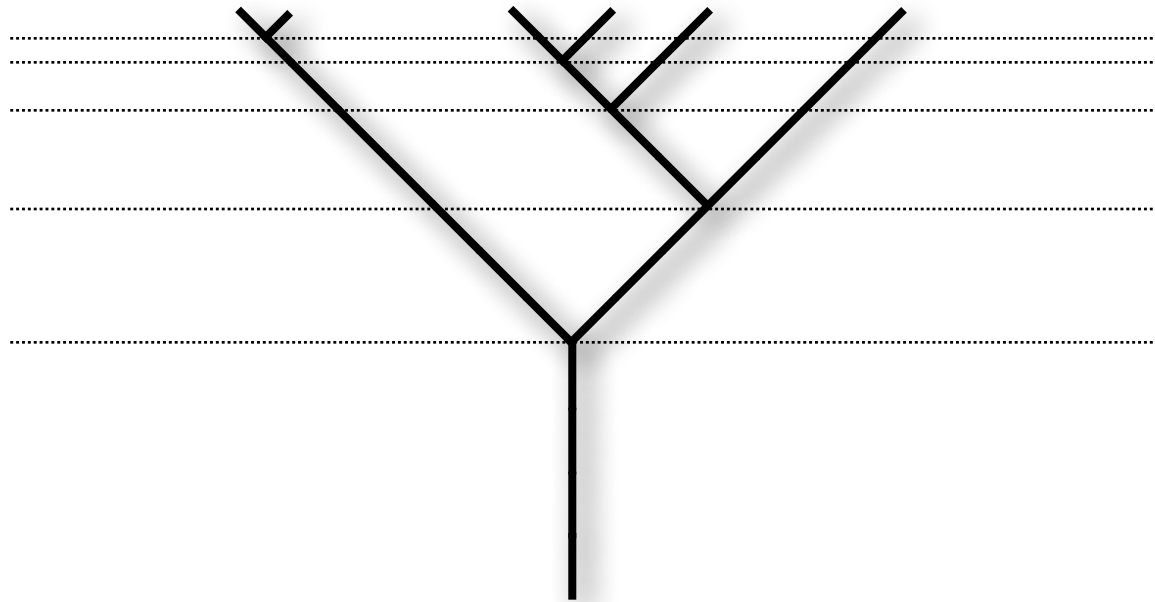


$$\frac{1}{15}$$



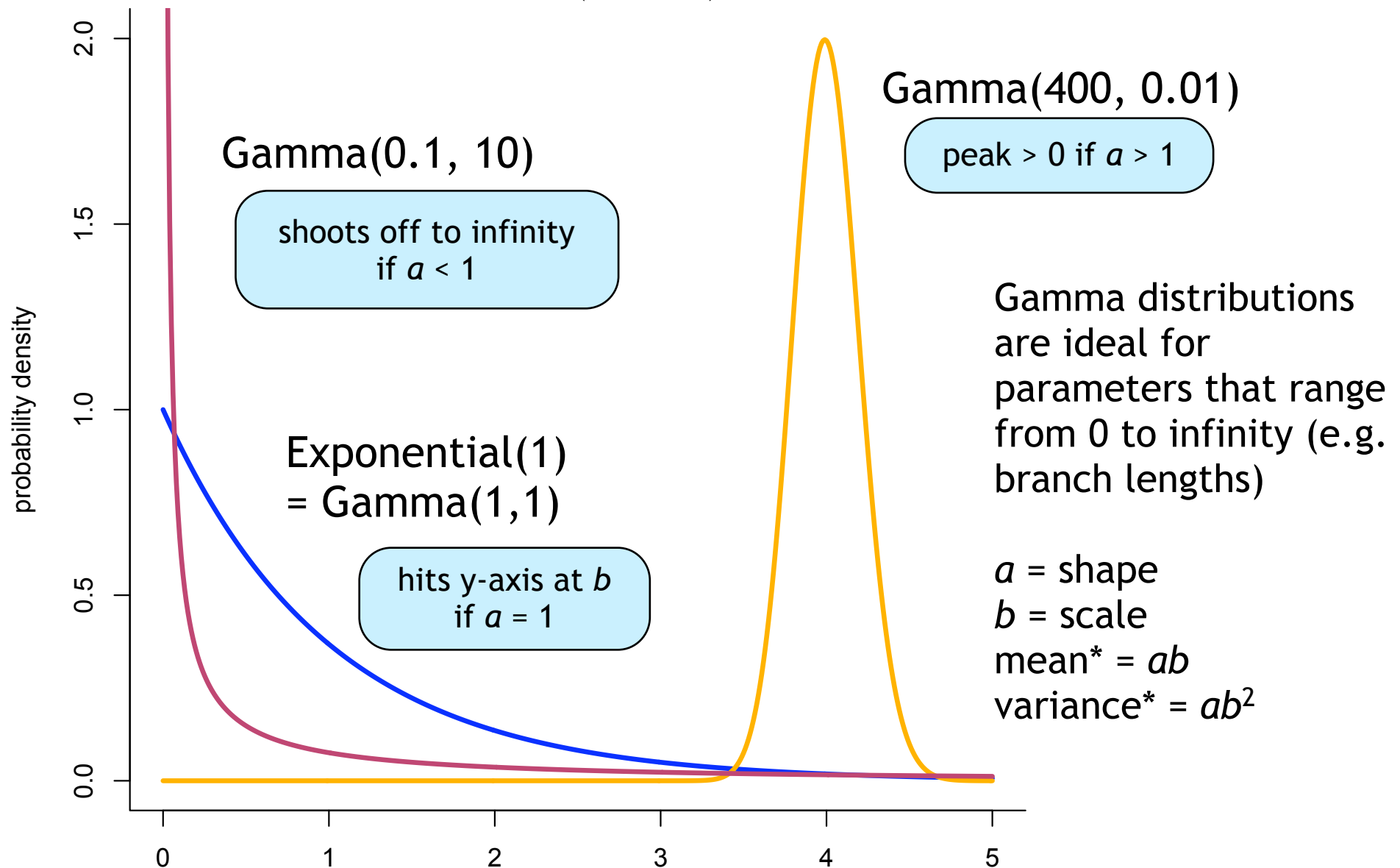
$$\frac{1}{15}$$

Yule model provides joint prior for both topology and divergence times



The rate of speciation under the Yule model (λ) is constant and applies equally and independently to each lineage. Thus, speciation events get closer together in time as the tree grows because more lineages are available to speciate.

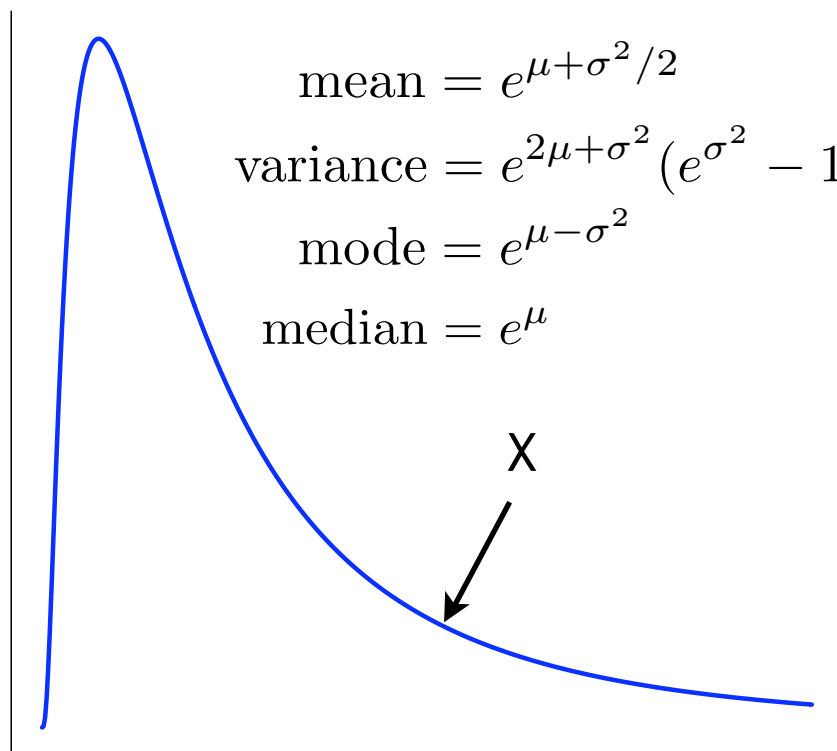
Gamma(a, b) distributions



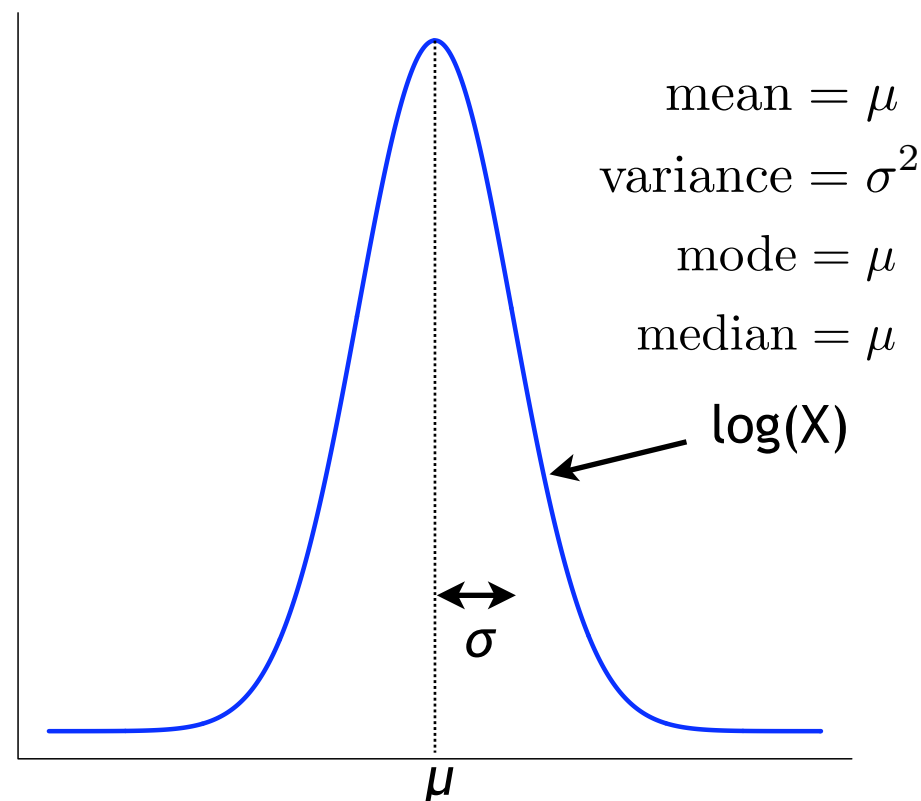
*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value b used in this slide! In this case, the mean and variance would be a/b and a/b^2 , respectively.

Log-normal distribution

If X is log-normal with parameters μ and σ ...



...then $\log(X)$ is normal with mean μ and standard deviation σ .



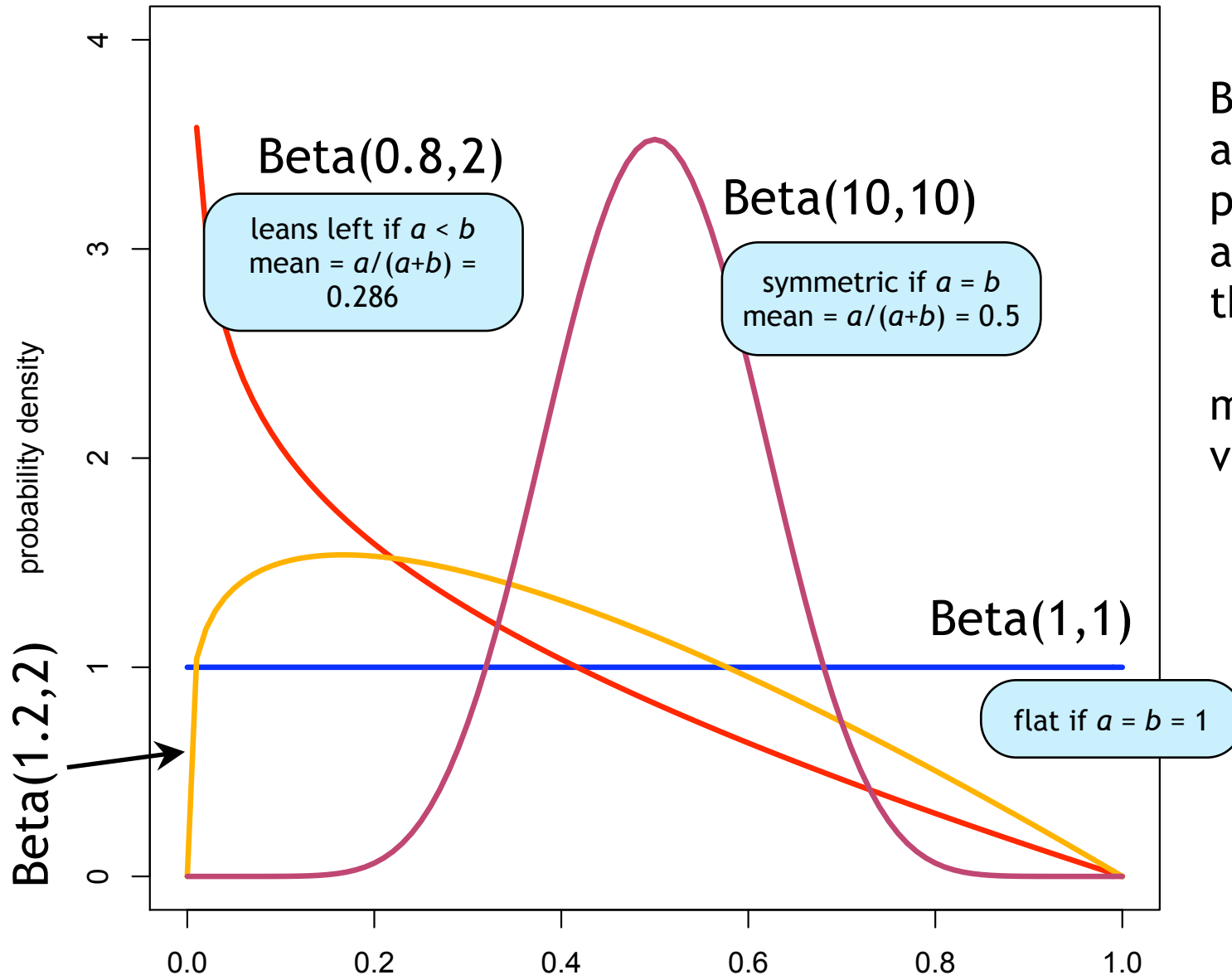
Important: μ and σ do *not* represent the mean and standard deviation of X : they are the mean and standard deviation of $\log(X)$!

To choose μ and σ to yield a particular mean (m) and variance (v) for X , use these formulas:

$$\mu = \log(m^2) - \log(m) - \frac{\log(v + m^2) - \log(m^2)}{2}$$

$$\sigma^2 = \log(v + m^2) - \log(m^2)$$

Beta(a,b) gallery



Beta distributions are appropriate for proportions, which are constrained to the interval $[0,1]$.

$$\text{mean} = a/(a+b)$$
$$\text{variance} = \frac{ab}{[(a+b)^2(a+b+1)]}$$

Dirichlet(a, b, c, d) distribution

Used for nucleotide relative frequencies:

$$a \rightarrow \pi_A, b \rightarrow \pi_C, c \rightarrow \pi_G, d \rightarrow \pi_T$$

Flat prior:

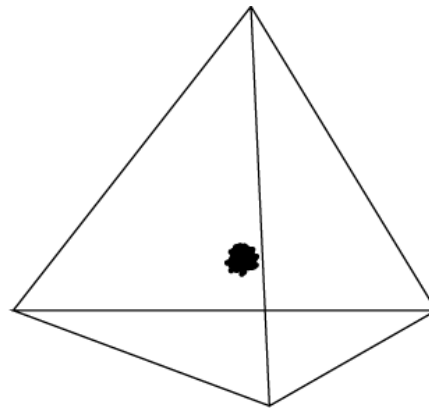
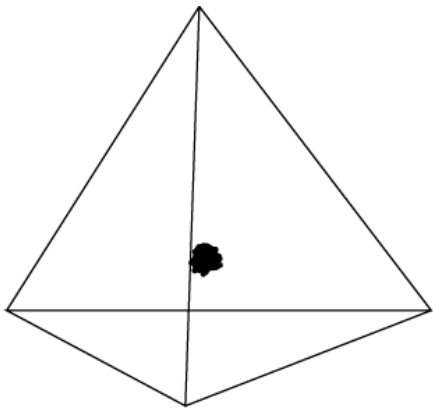
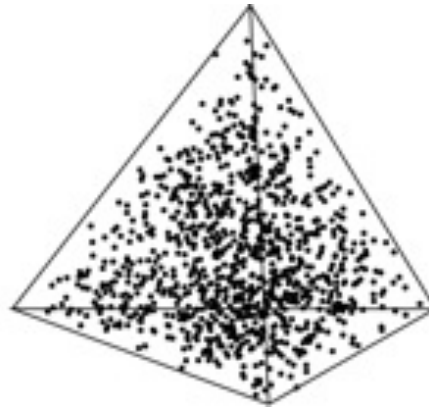
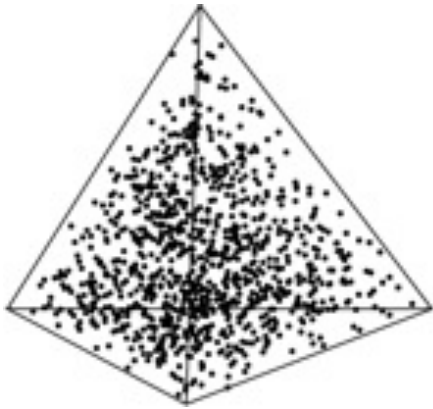
$$a = b = c = d = 1$$

(no scenario discouraged)

Informative prior:

$$a = b = c = d = 300$$


(equal frequencies strongly encouraged)

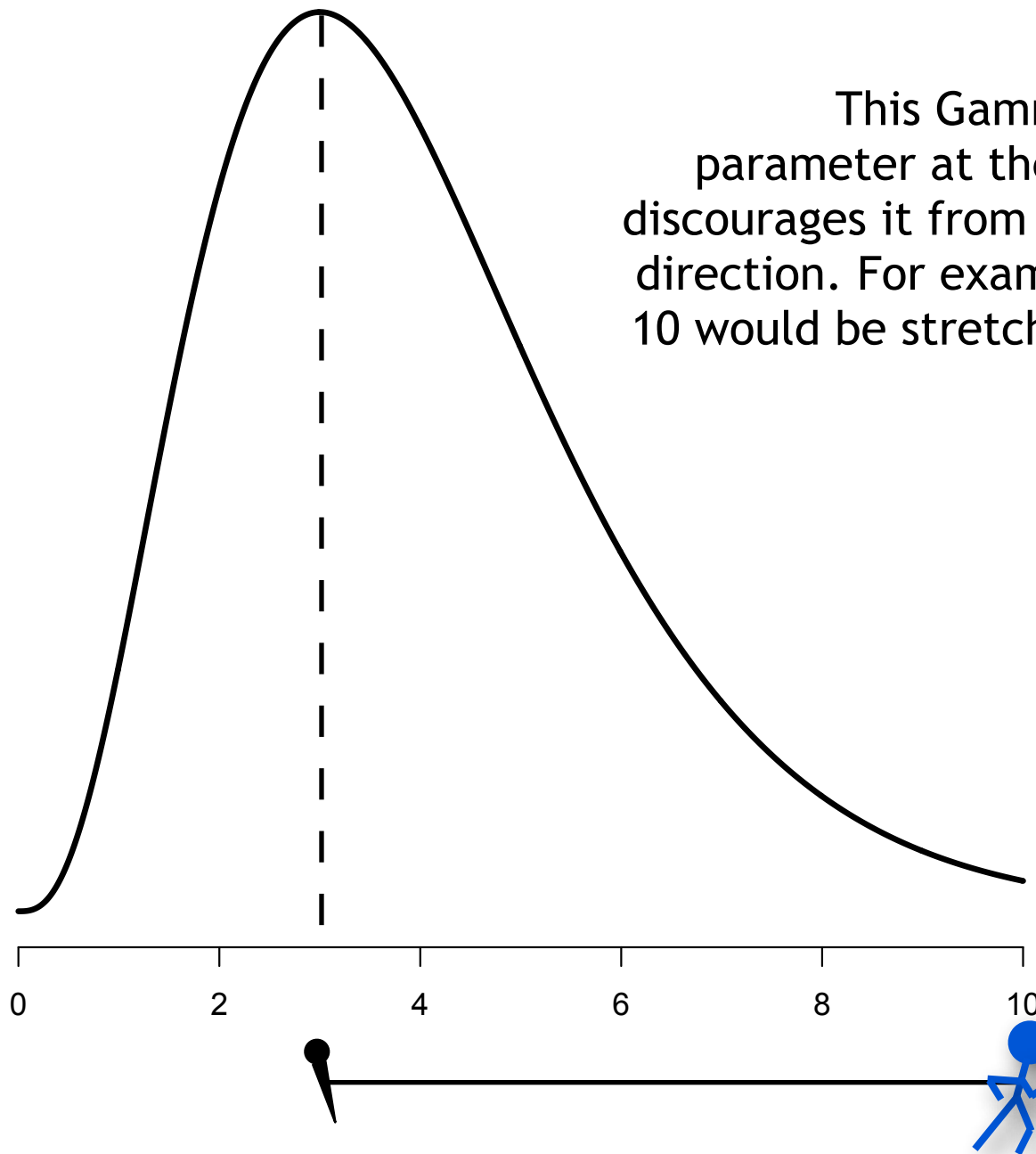


(stereo pairs)

Dirichlet(a, b, c, d, e, f) used for
GTR exchangeability parameters.

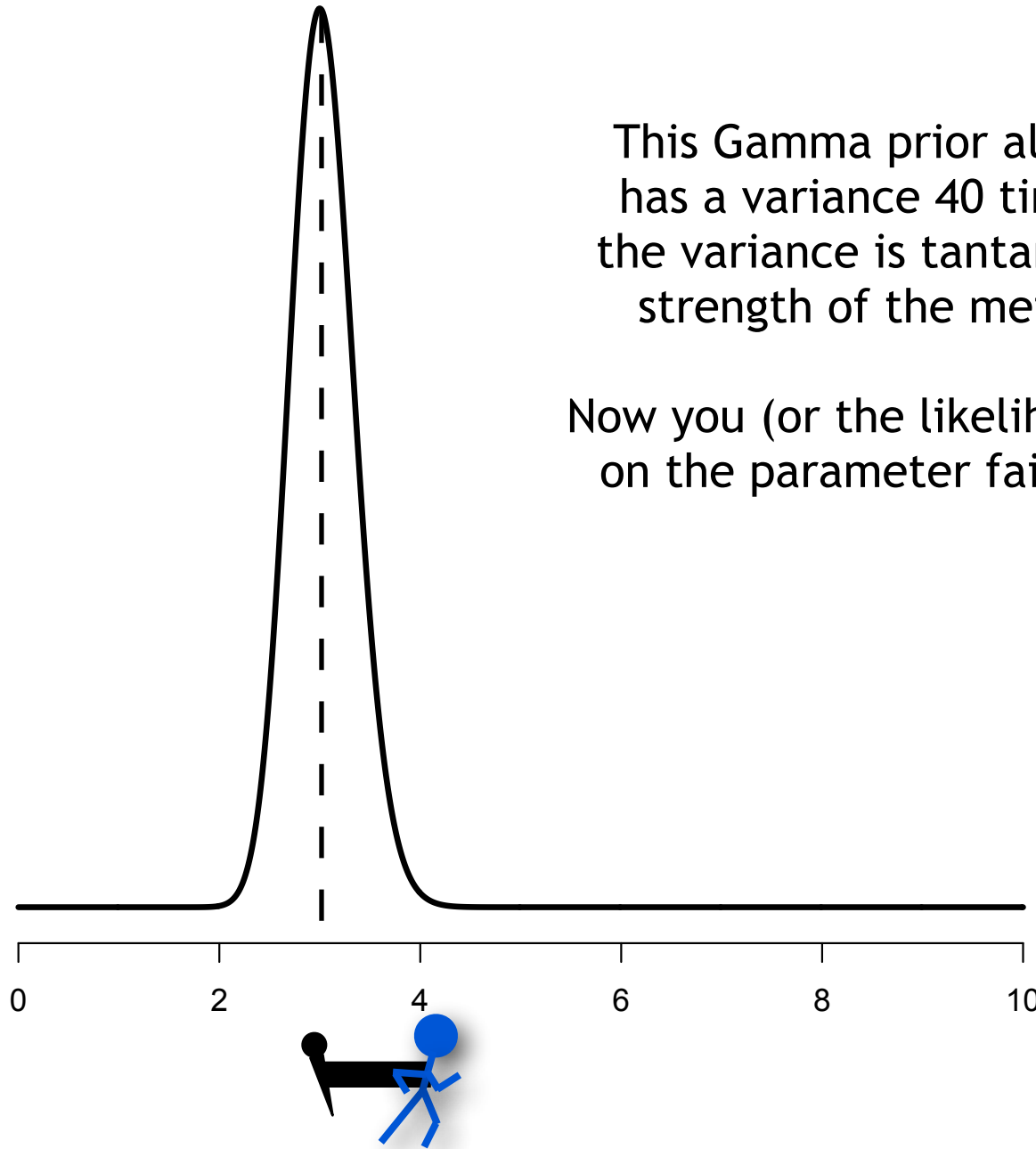
Prior Miscellany

- priors as rubber bands 
- running on empty
- hierarchical models
- empirical bayes



This Gamma(4,1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

The mode of a Gamma(a,b) distribution is $(a-1)b$ (assuming $a > 1$)

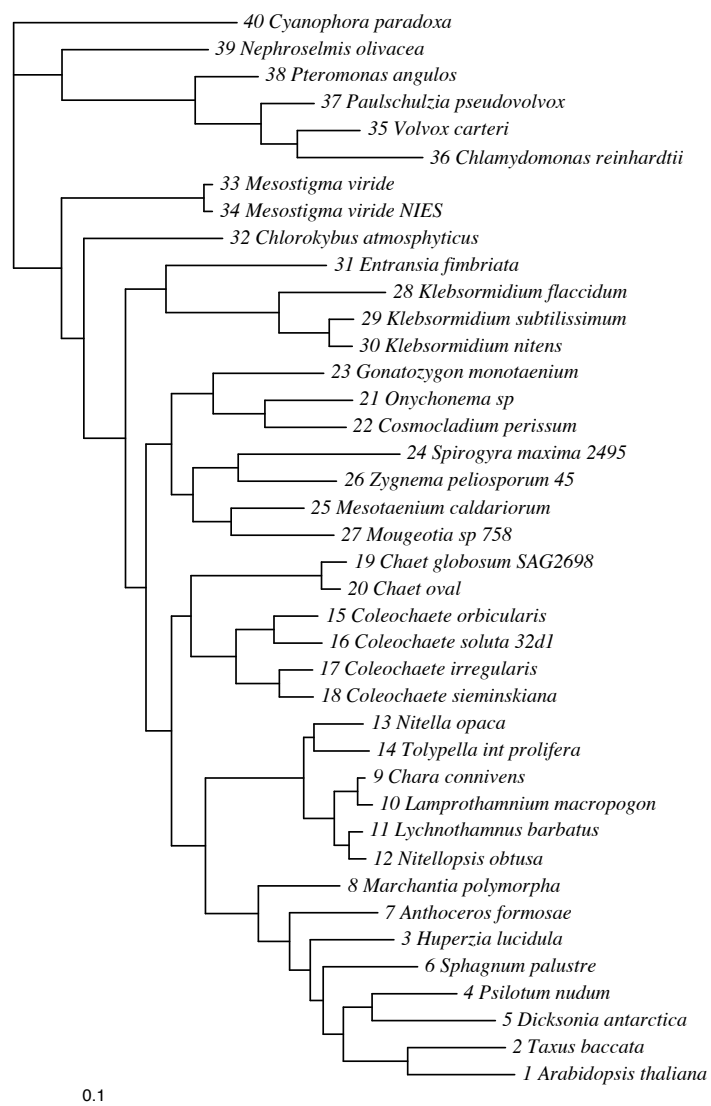


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

Now you (or the likelihood) would have to tug on the parameter fairly hard for it to have a value as large as 4.

This gamma distribution has shape 91.989 and scale 0.032971

Example: Internal Branch Length Priors

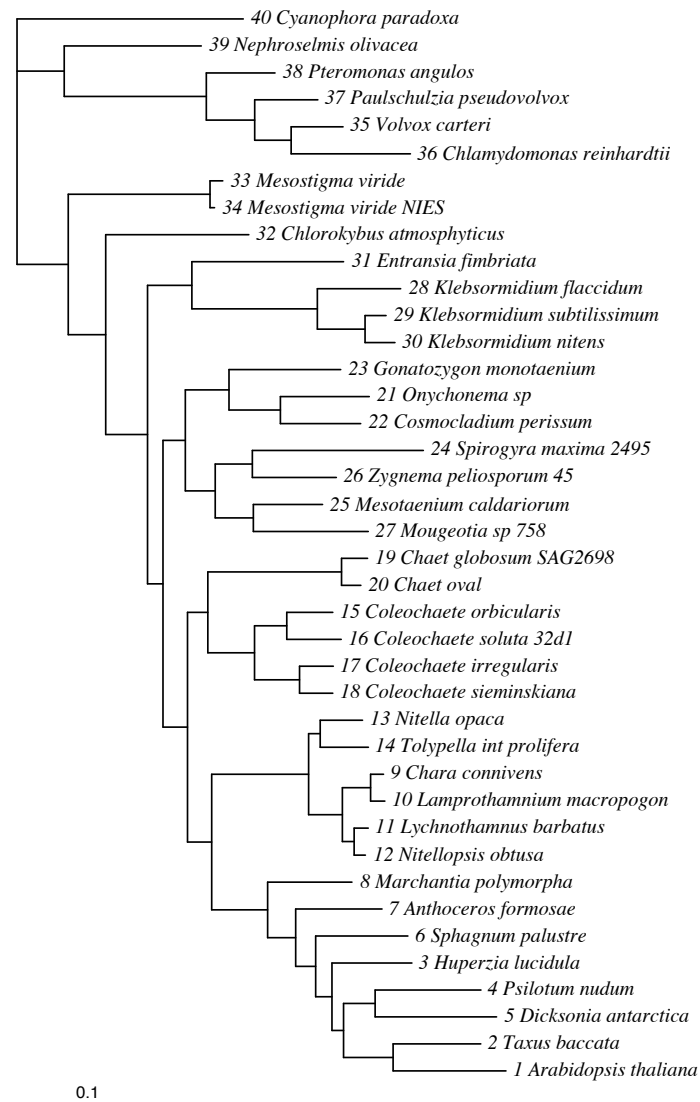


Separate priors applied to
internal and external branches

External branch length prior is
exponential with mean 0.1

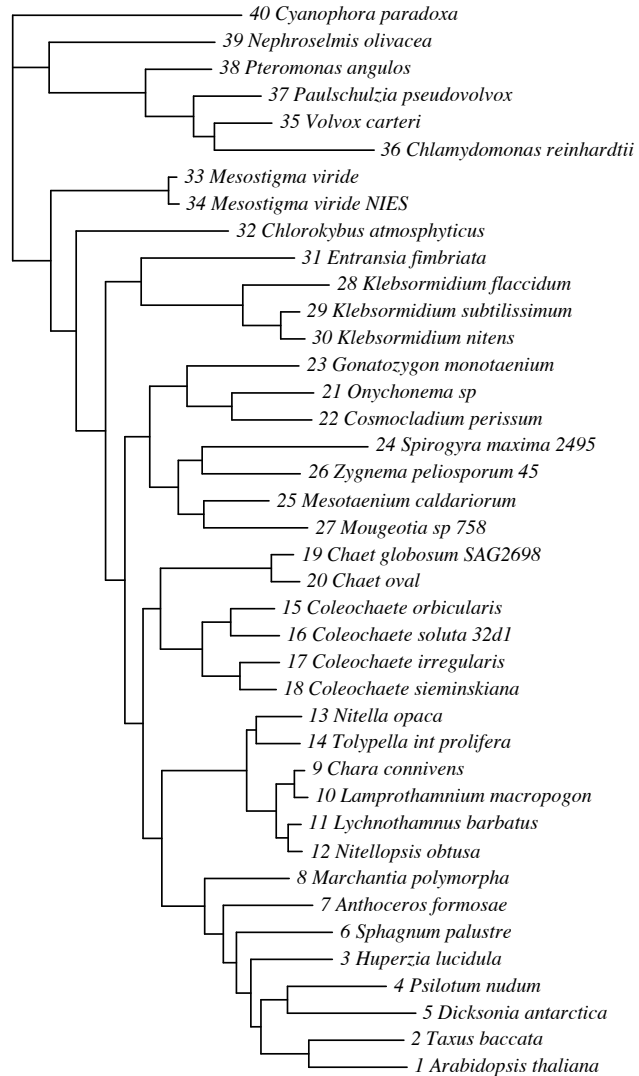
Internal branch length prior is
exponential with mean 0.1

This is a reasonably vague
internal branch length prior

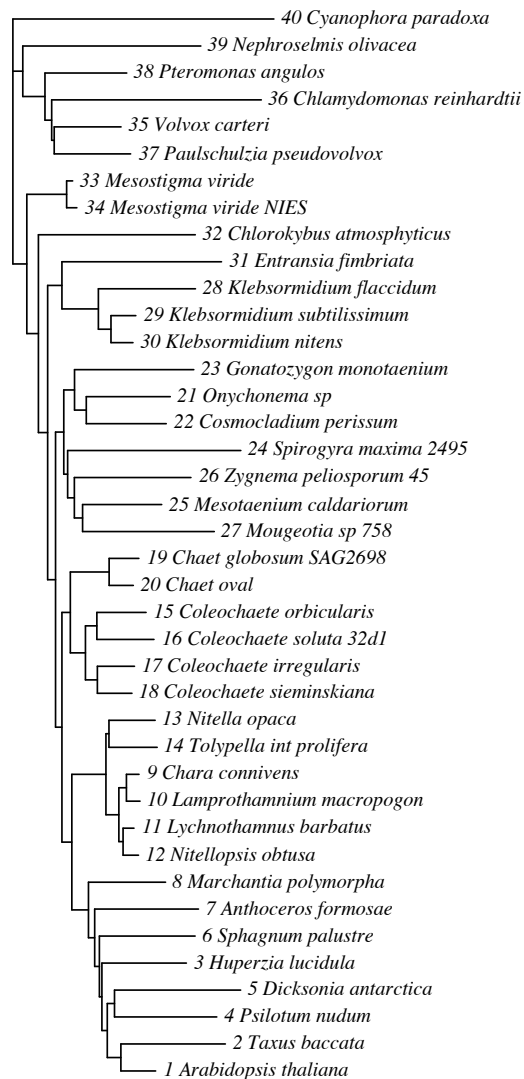


Internal branch length prior mean 0.01

(external branch length prior mean always 0.1)

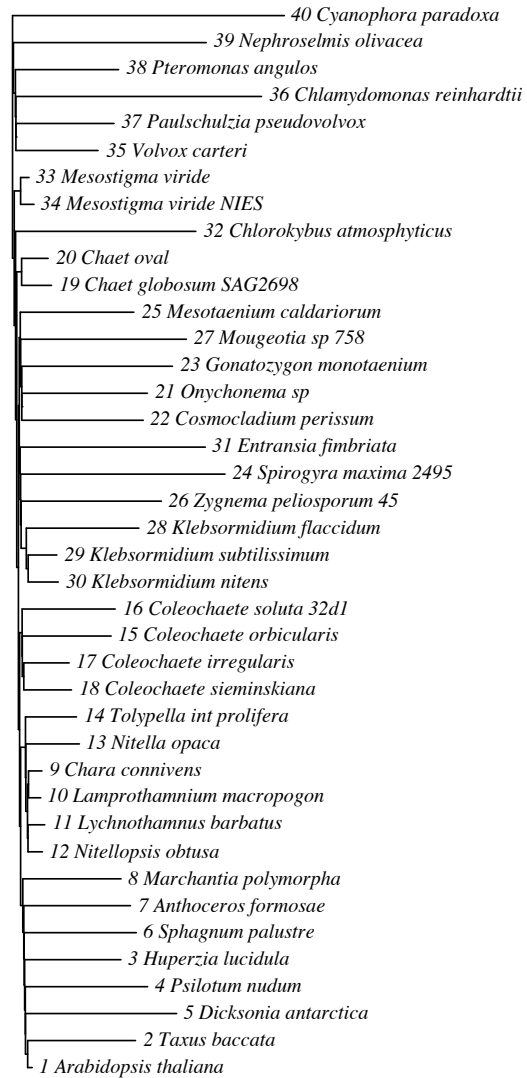


Internal branch length prior mean
0.001

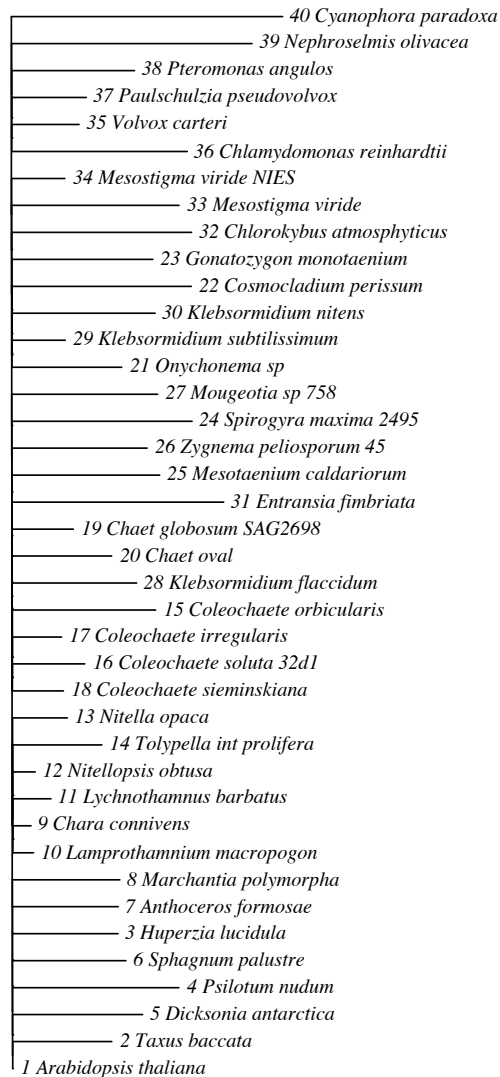


Internal branch length prior mean
0.0001

0.1



Internal branch length prior mean
0.00001



Internal branch length prior mean
0.000001

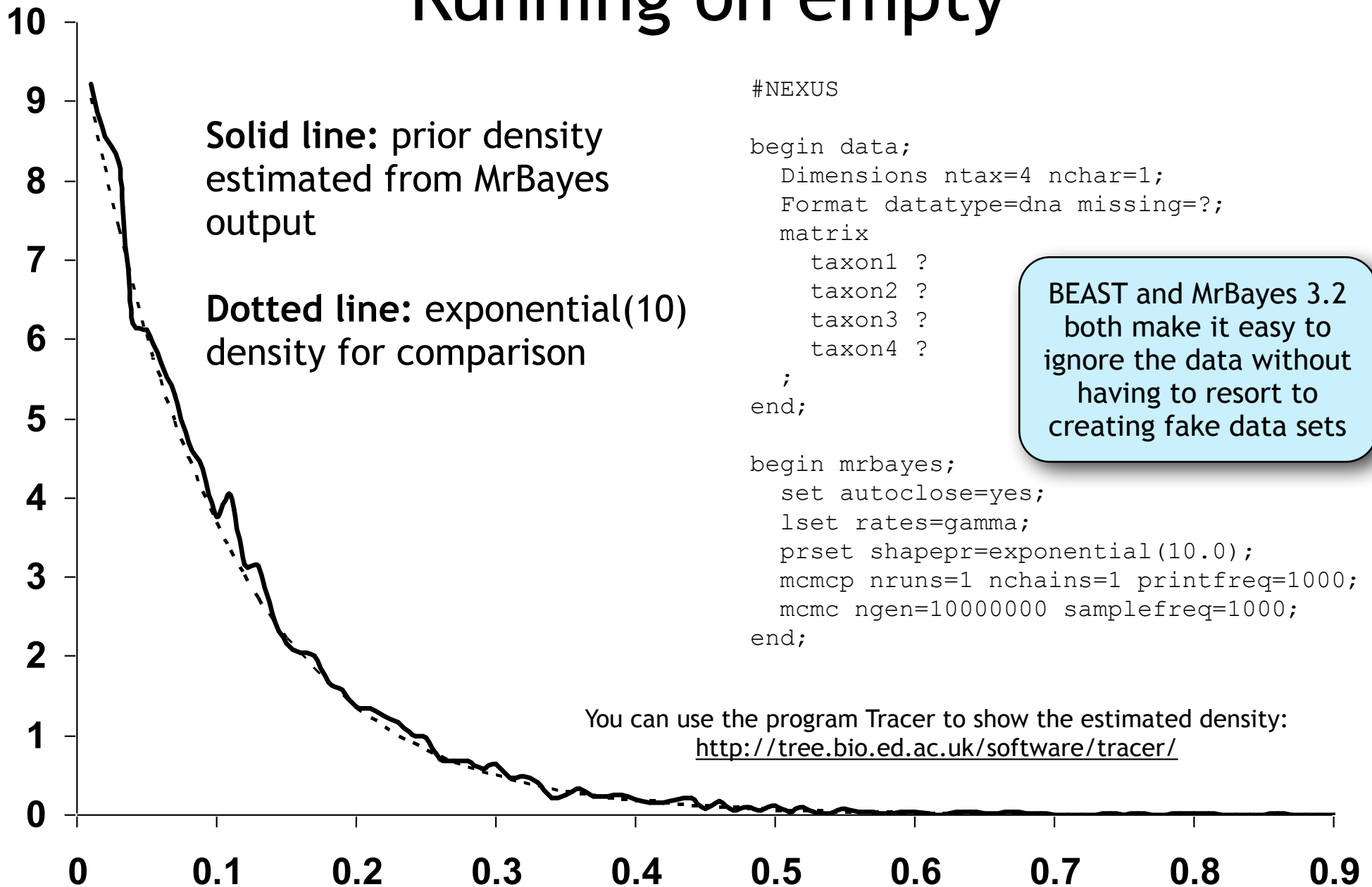
The internal branch length prior is
calling the shots now, and the
likelihood must obey.

Prior Miscellany

- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



Running on empty



Prior Miscellany

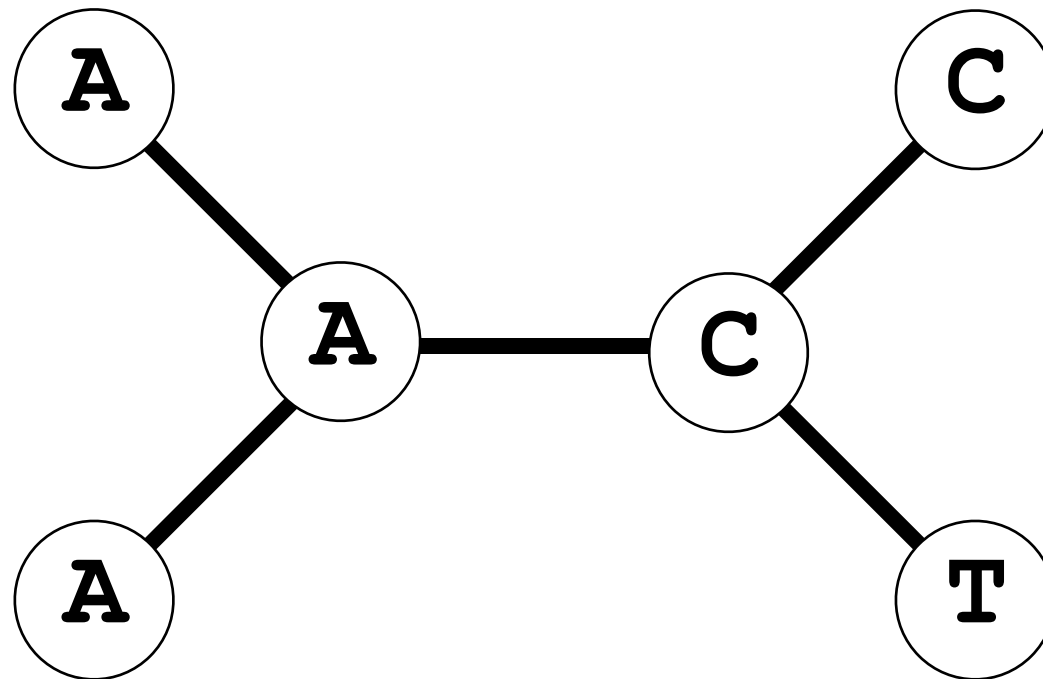
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



In a **non-hierarchical** model, all parameters are present in the likelihood function

Prior: Exponential, mean=0.1

$$L_k = \frac{1}{4} \left[\frac{1}{4} + \frac{3}{4} e^{-4v_1/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_2/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_3/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_4/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_5/3} \right]$$




Hierarchical models add *hyperparameters* not present in the likelihood function

μ is a *hyperparameter* governing the mean of the edge length prior

hyperprior



Prior: Exponential, mean μ


$$L_k = \frac{1}{4} \left[\frac{1}{4} + \frac{3}{4} e^{-4v_1/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_2/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_3/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_4/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_5/3} \right]$$

During an MCMC analysis, μ will hover around a reasonable value, sparing you from having to decide what value is appropriate. You still have to specify a hyperprior, however.

Prior Miscellany

- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes



Empirical Bayes

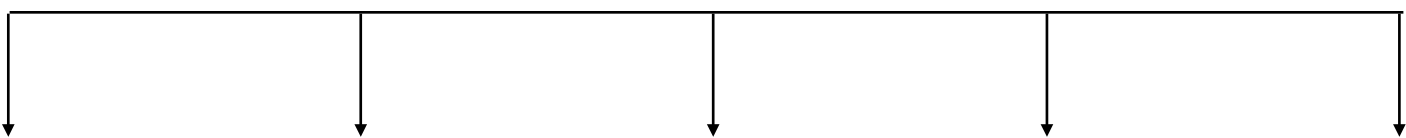
Empirical Bayes uses the data to determine some aspects of the prior, such as the prior mean.

Pure Bayesian approaches choose priors without reference to the data.

An empirical Bayesian would use the maximum likelihood estimate (MLE) of the length of an average branch here



Prior: Exponential, mean=MLE


$$L_k = \frac{1}{4} \left[\frac{1}{4} + \frac{3}{4} e^{-4v_1/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_2/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_3/3} \right] \left[\frac{1}{4} - \frac{1}{4} e^{-4v_4/3} \right] \left[\frac{1}{4} + \frac{3}{4} e^{-4v_5/3} \right]$$

V. Bayesian model selection

AIC is not Bayesian. Why?

$$AIC = 2k - 2 \log(\max L)$$

number of free (estimated) parameters maximized log likelihood

AIC is not Bayesian because the **prior is not considered**
(and the prior is an important component of a Bayesian model)

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}$$

The **marginal likelihood** (denominator in Bayes' Rule) is
commonly used for Bayesian model selection

Represents the (weighted) **average fit of the model** to the
observed data (weights provided by the prior)

An evolutionary distance example



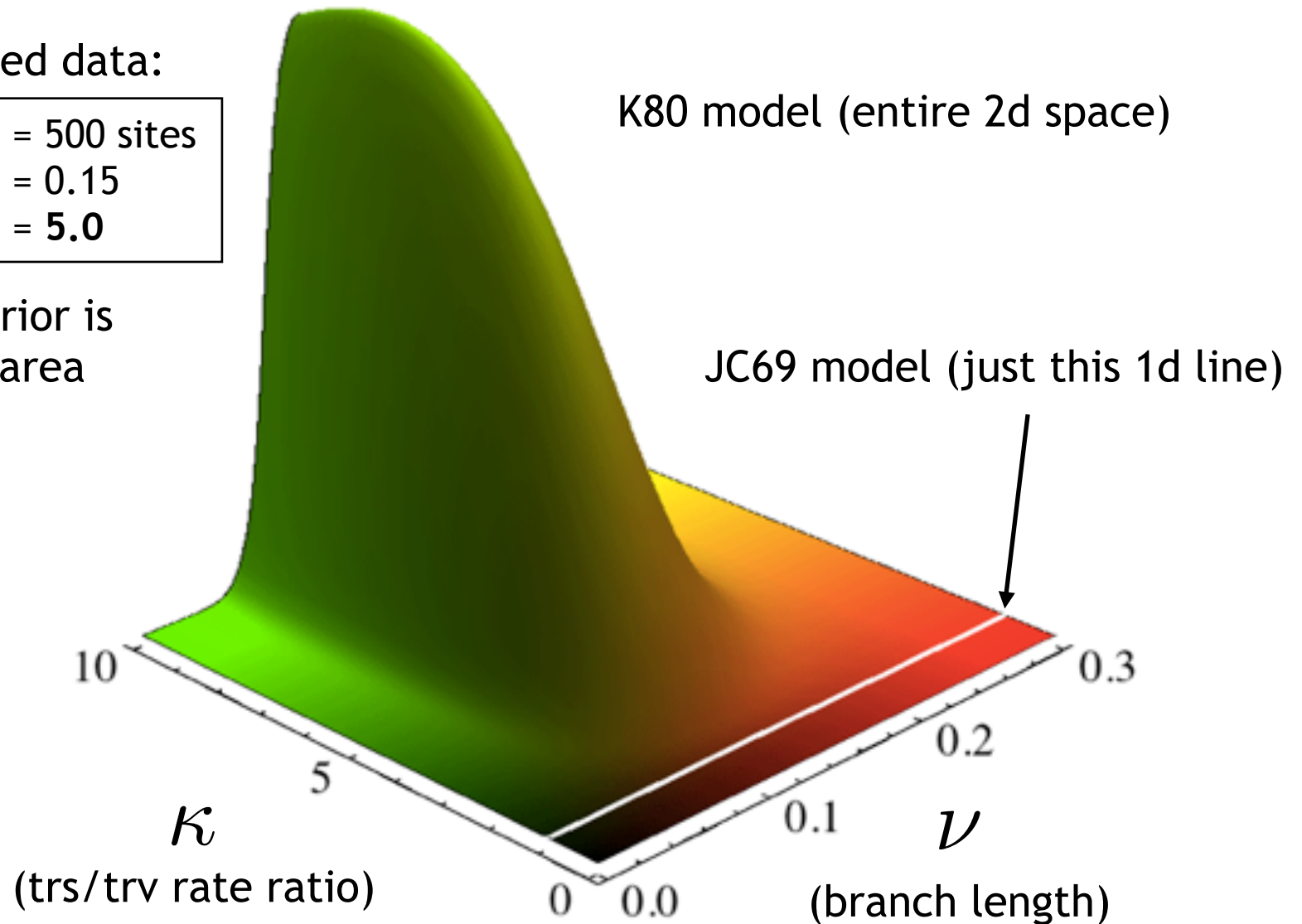
- Let's compare models JC69 vs. K80
- Parameters:
 - v is edge length (expected no. substitutions/site)
 - free in both JC69 and K80 models
 - k is transition/transversion rate ratio
 - free in K80, set to 1.0 in JC69

Likelihood Surface when K80 true

Based on simulated data:

sequence length	= 500 sites
true branch length	= 0.15
true kappa	= 5.0

Assume joint prior is
flat over the area
shown.



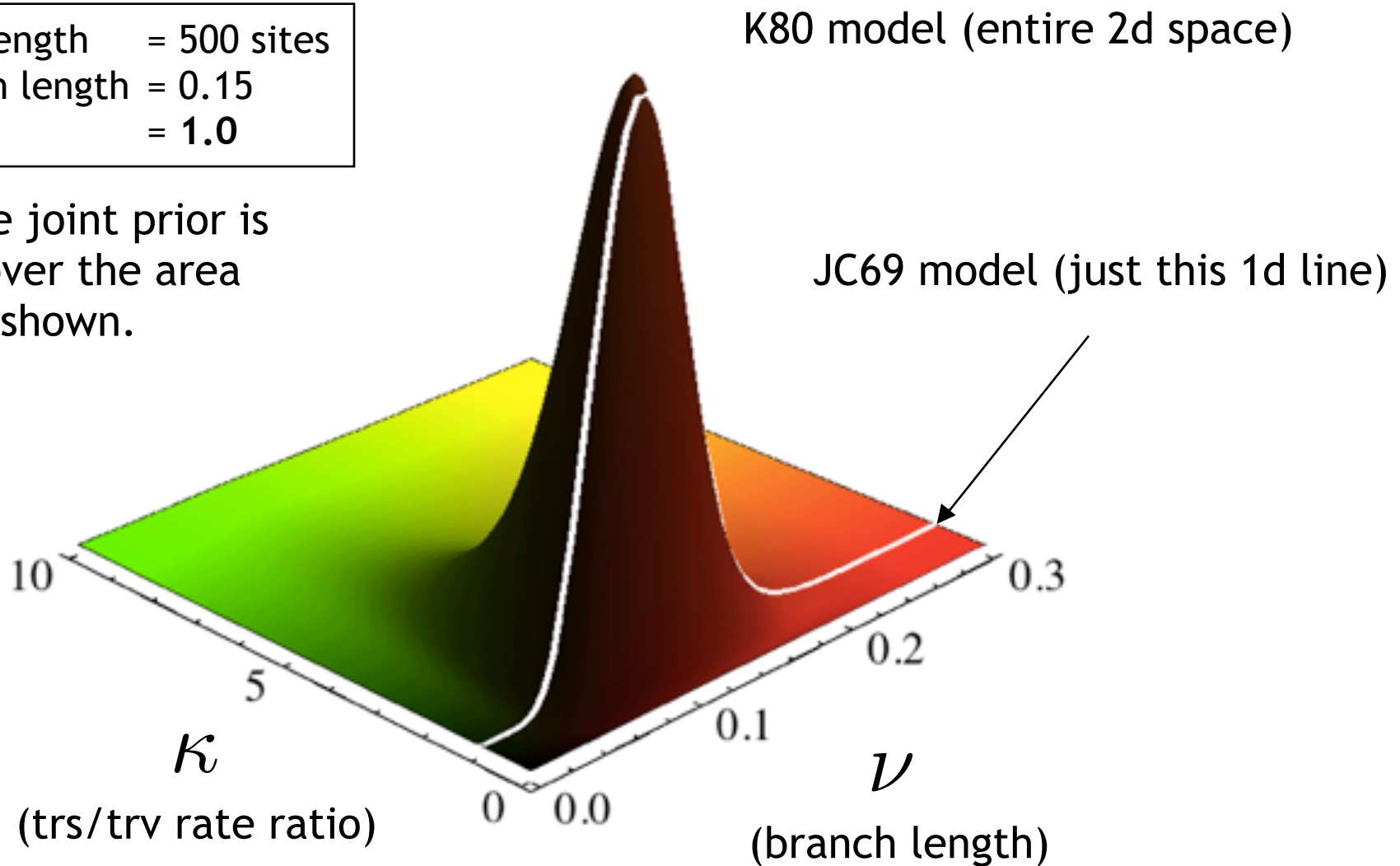
K80 wins

Likelihood Surface when JC true

Based on simulated data:

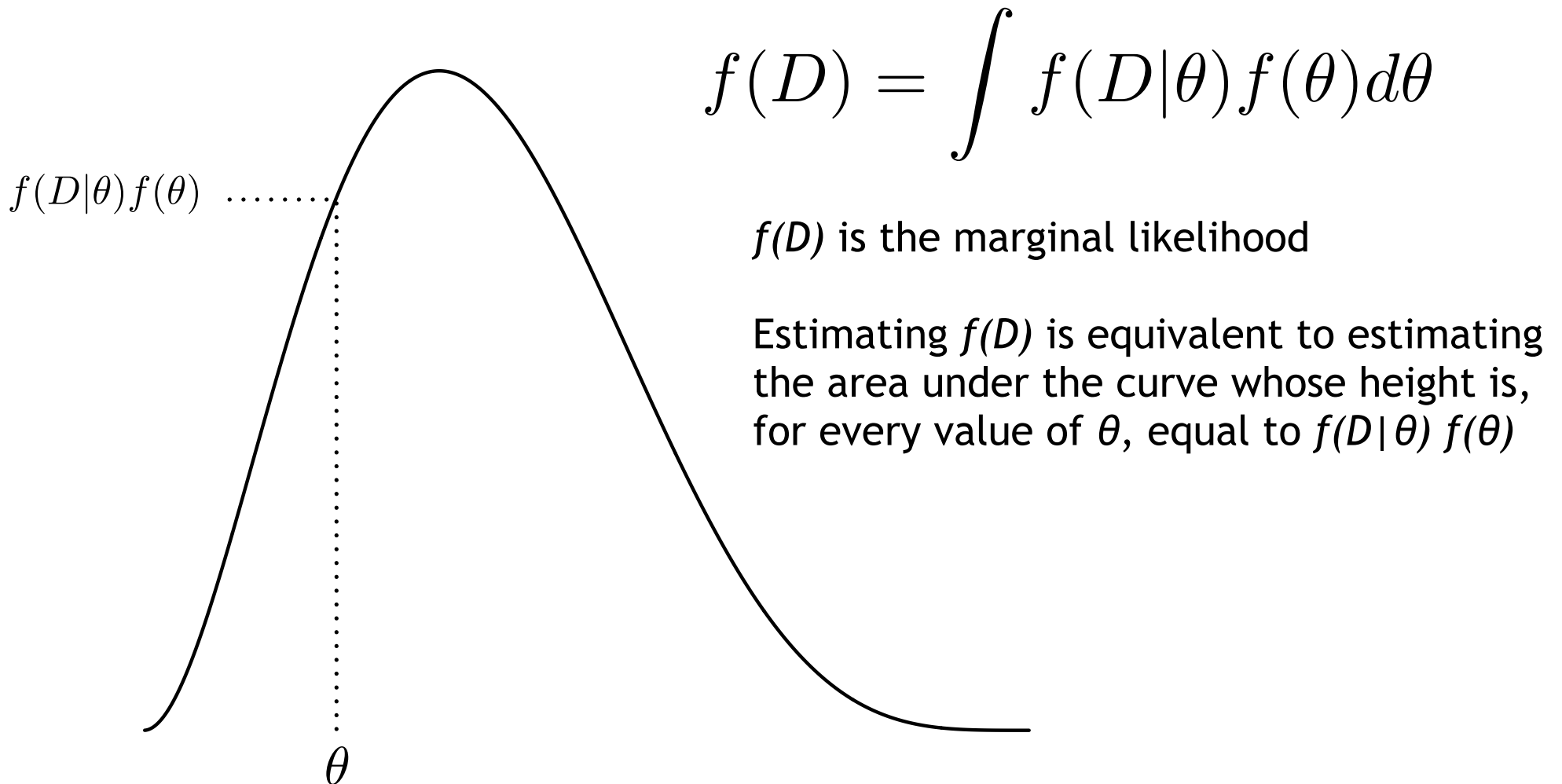
sequence length	= 500 sites
true branch length	= 0.15
true kappa	= 1.0

Assume joint prior is flat over the area shown.



JC69 wins

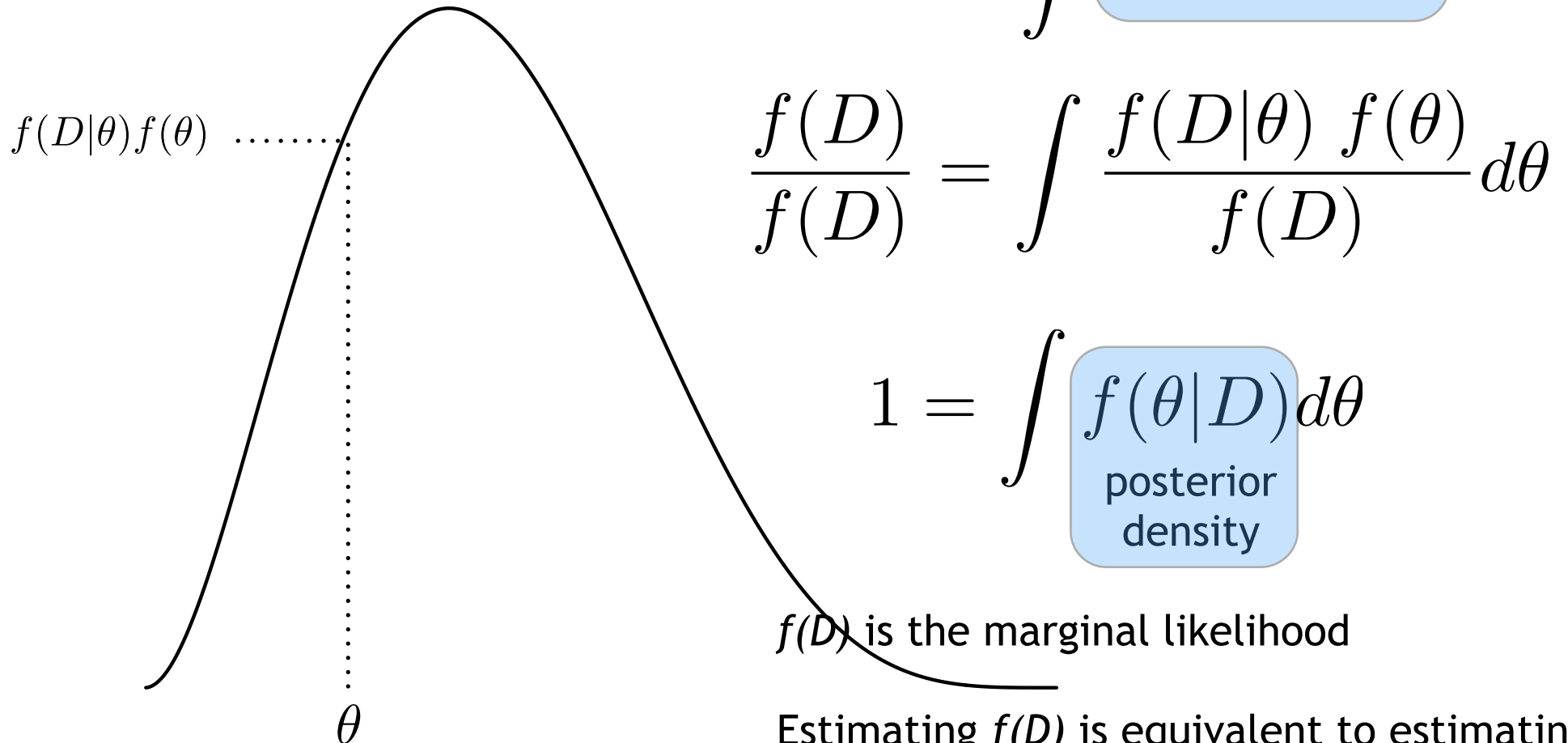
Estimating the marginal likelihood



Estimating the marginal likelihood

Remember that $f(D)$ is the normalizing constant that turns the posterior *kernel* into a posterior *density*.

$$f(D) = \int \overbrace{f(D|\theta)f(\theta)}^{\text{posterior kernel}} d\theta$$



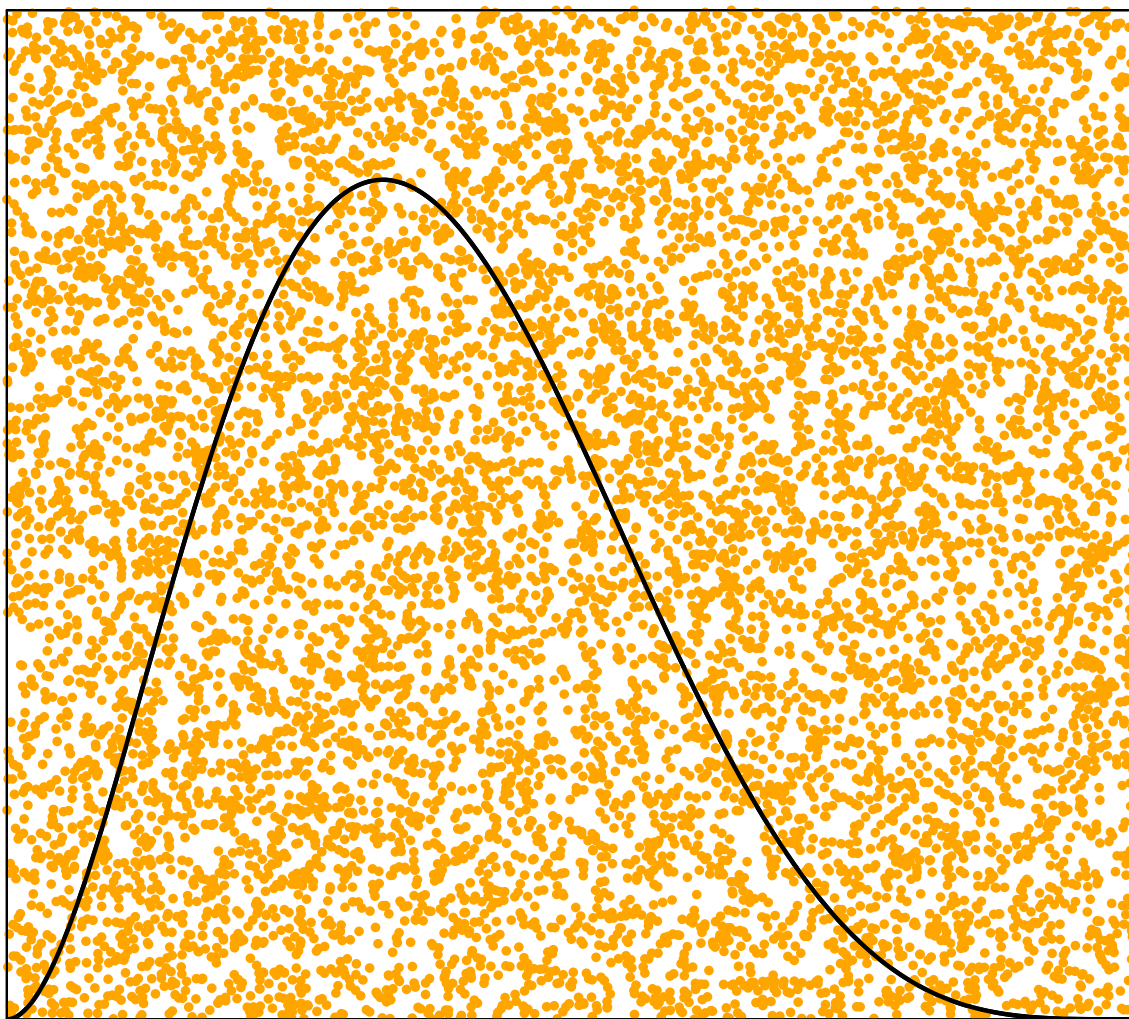
$$\frac{f(D)}{f(D)} = \int \frac{f(D|\theta) f(\theta)}{f(D)} d\theta$$

$$1 = \int \underbrace{f(\theta|D)}_{\text{posterior density}} d\theta$$

$f(D)$ is the marginal likelihood

Estimating $f(D)$ is equivalent to estimating the area under the curve whose height is, for every value of θ , equal to $f(D|\theta) f(\theta)$

Estimating the marginal likelihood



Sample evenly from a box with known area A that completely encloses the curve.

Area under the curve is just A times the fraction of sampled points that lie under the curve.

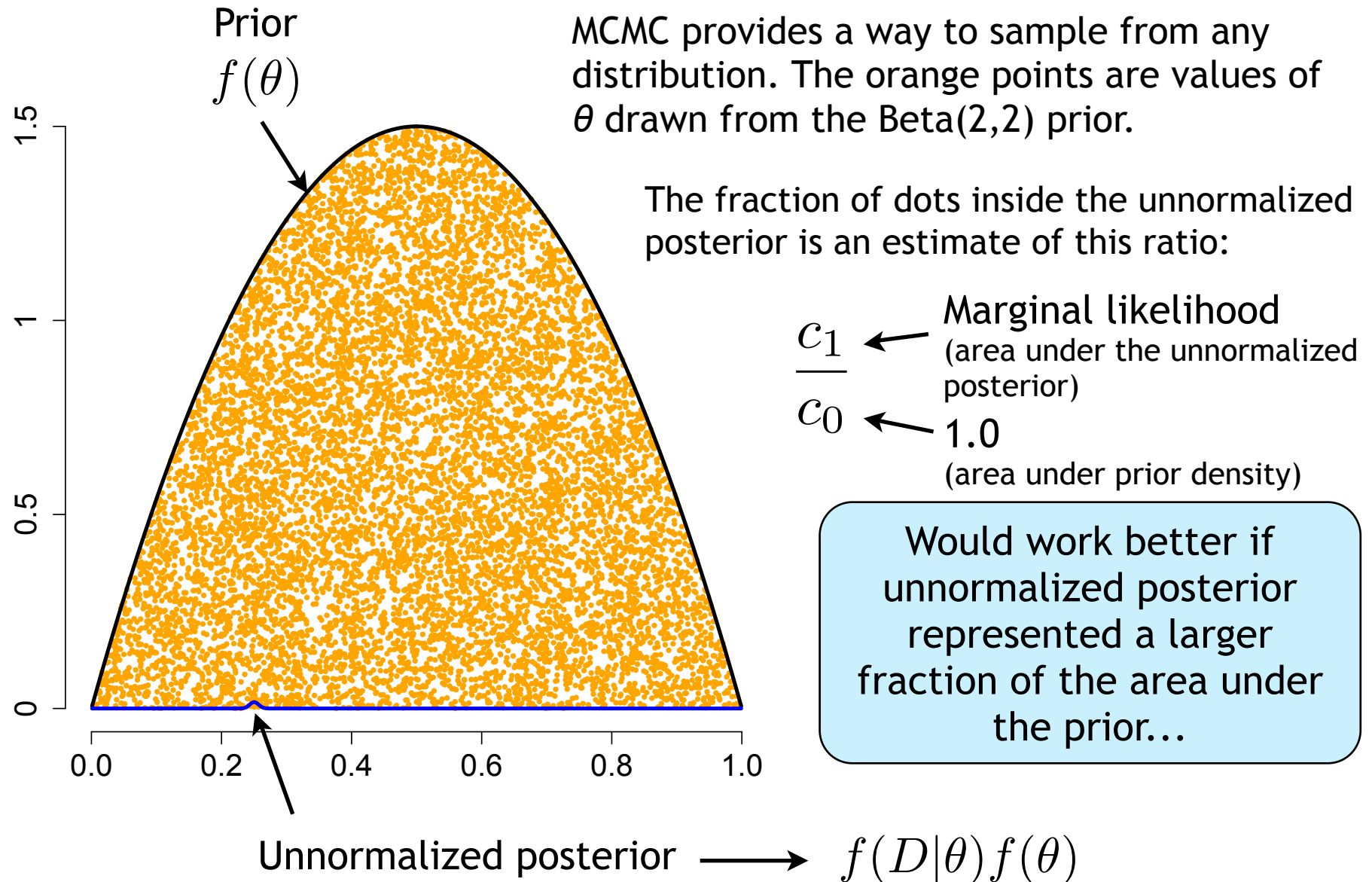
While not a box, the prior $f(\theta)$ does have area 1.0 and completely encloses the curve:

$$1.0 = \int f(\theta) d\theta$$

$$f(D) = \int f(D|\theta) f(\theta) d\theta$$

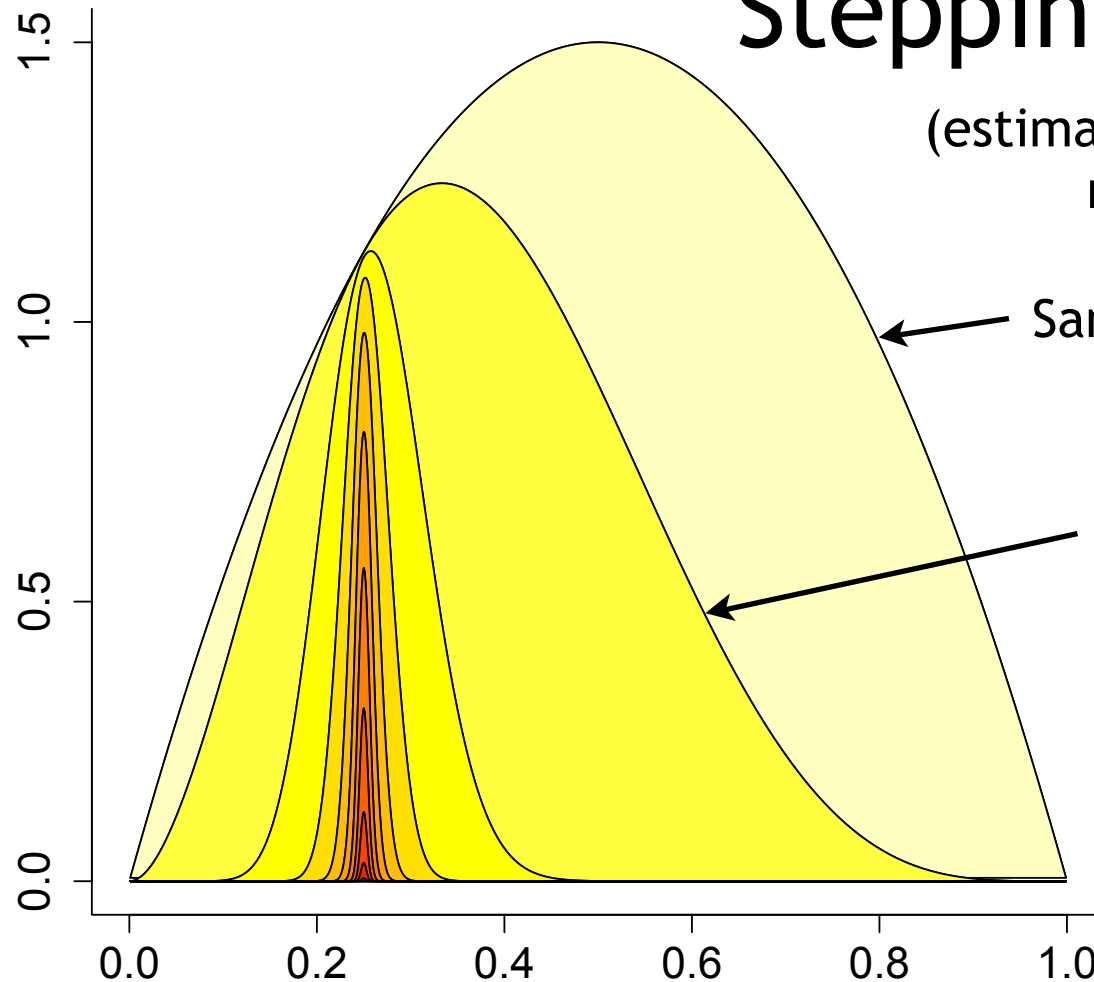
↑

Estimating the marginal likelihood



Stepping-stone method

(estimates a series of ratios that each represent smaller jump)



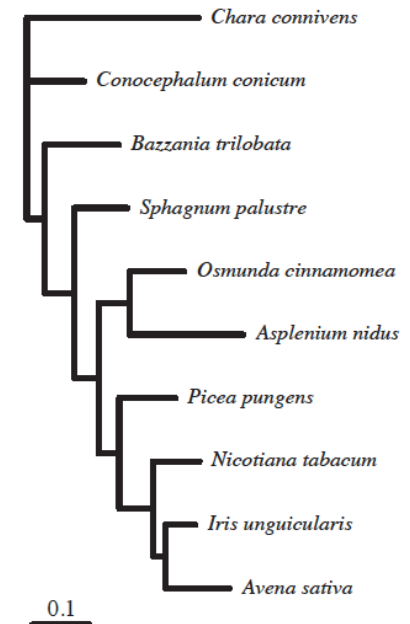
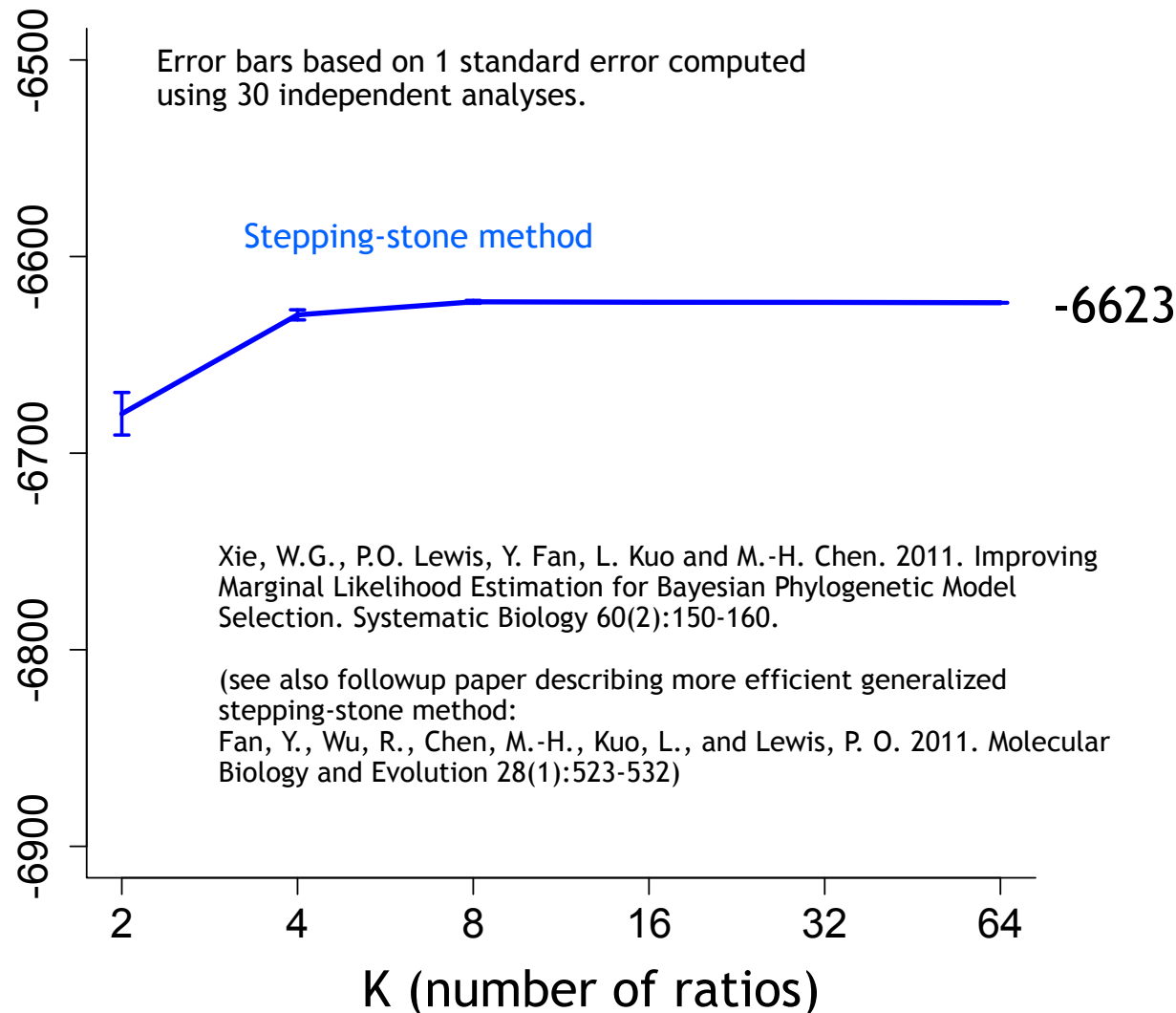
Sample from this distribution

See what fraction of samples are under this density curve

This fraction is an estimate of this ratio

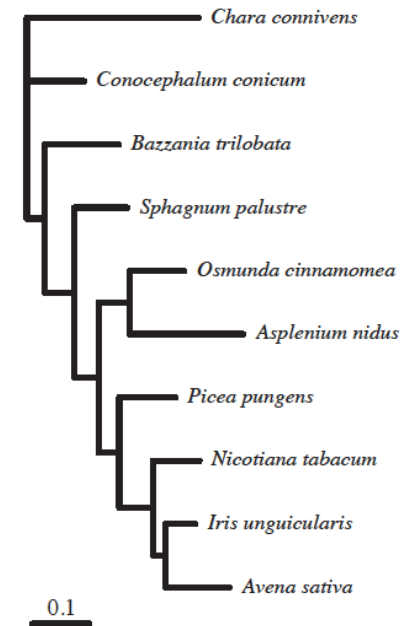
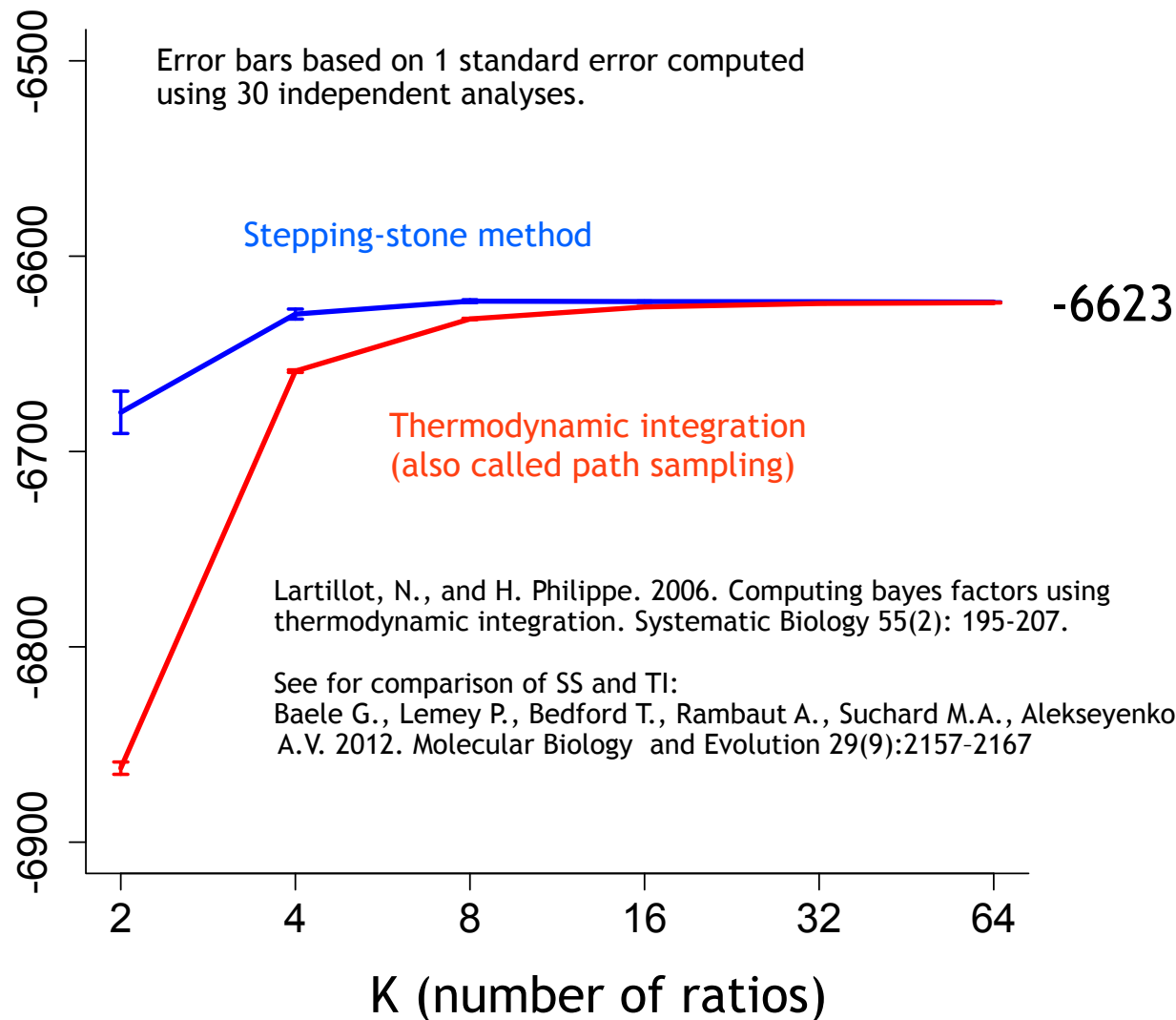
$$\frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{c_{0.9}} \right) \left(\frac{c_{0.9}}{c_{0.8}} \right) \left(\frac{c_{0.8}}{c_{0.7}} \right) \left(\frac{c_{0.7}}{c_{0.6}} \right) \left(\frac{c_{0.6}}{c_{0.5}} \right) \left(\frac{c_{0.5}}{c_{0.4}} \right) \left(\frac{c_{0.4}}{c_{0.3}} \right) \left(\frac{c_{0.3}}{c_{0.2}} \right) \left(\frac{c_{0.2}}{c_{0.1}} \right) \left(\frac{c_{0.1}}{c_{0.0}} \right)$$

How many “stepping stones” (i.e. ratios) are needed?



- *rbcL* data
- 10 green plants
- GTR+G model
- 1000 samples/steppingstone

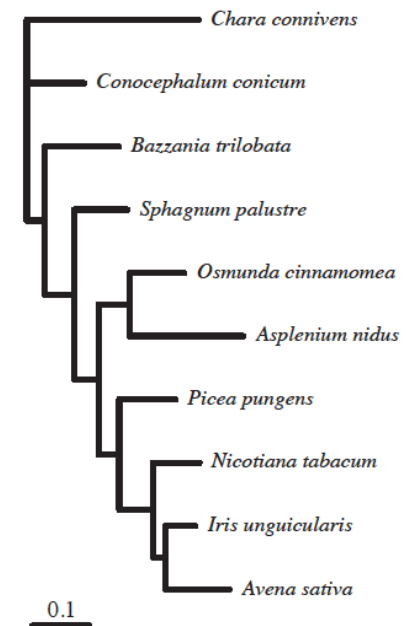
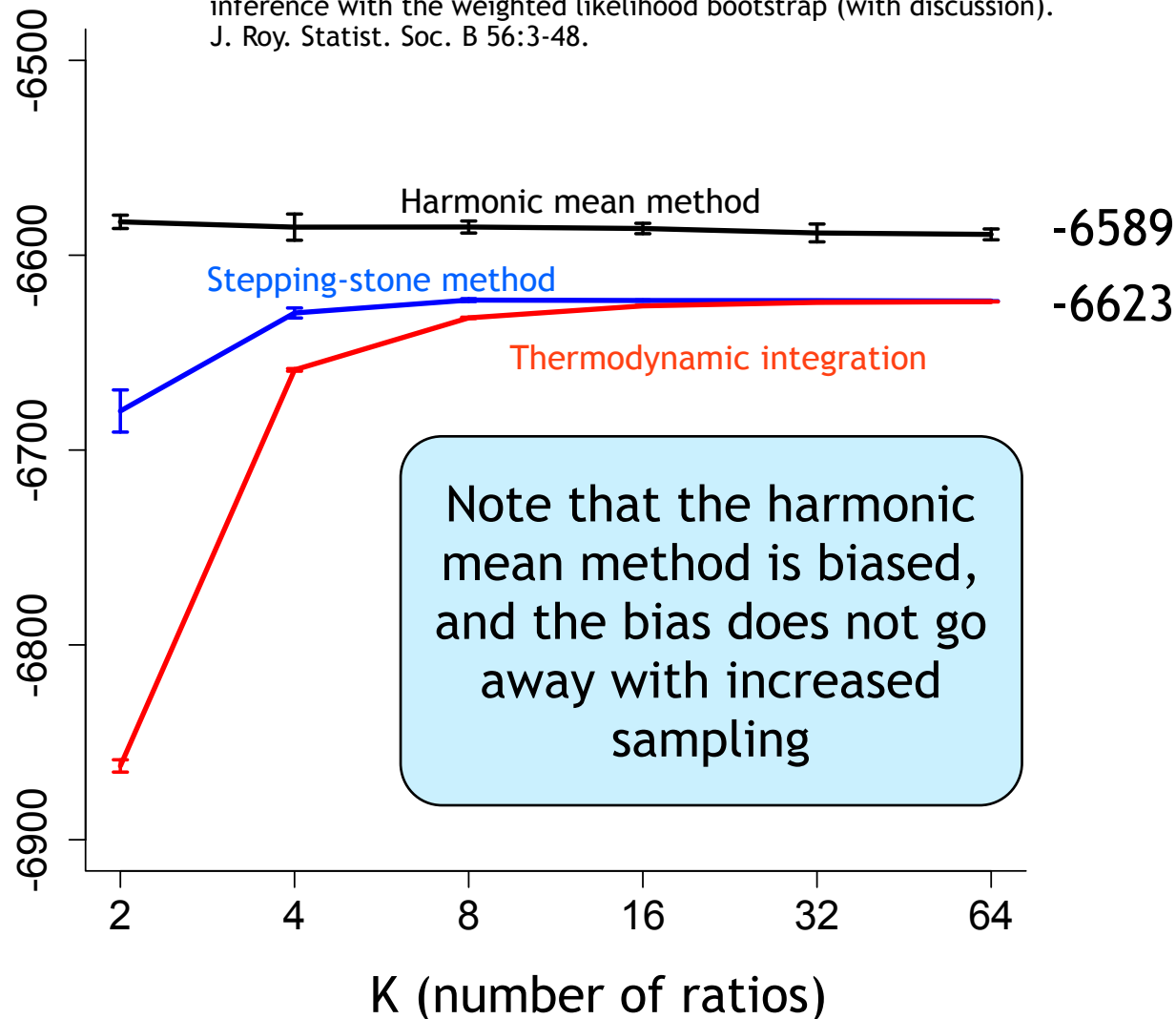
Is steppingstone sampling accurate?



- *rbcL* data
- 10 green plants
- GTR+G model
- 1000 samples/steppingstone

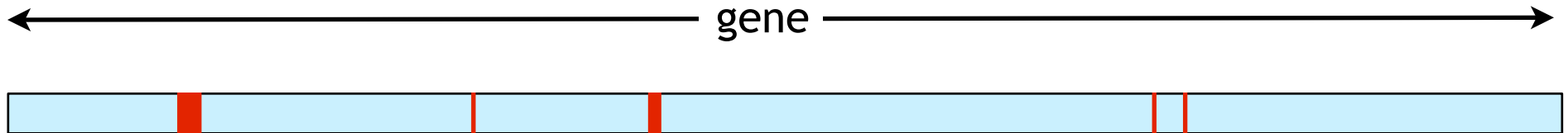
How about the harmonic mean method?

Newton, M. A., and A. E. Raftery. 1994. Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J. Roy. Statist. Soc. B 56:3-48.



- *rbcL* data
- 10 green plants
- GTR+G model
- 1000 samples/steppingstone

The problem that DP models help solve

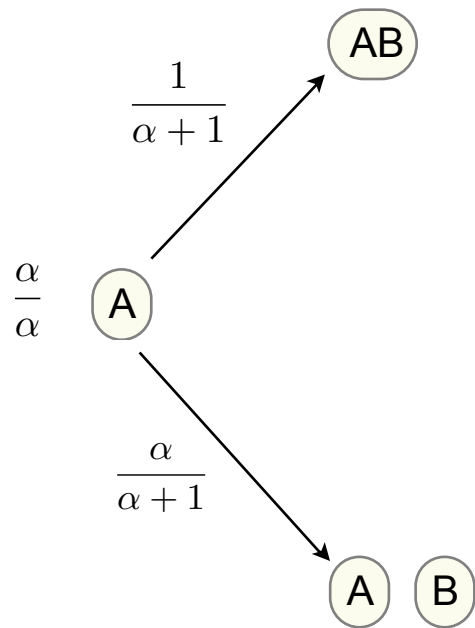


Red depicts sites with, for example:

- an unusually high or low rate
- unusual equilibrium base (or amino acid) frequencies
- an unusually high or low nonsynon./synon. rate ratio
- some other unusual feature

Desired: a prior model that:

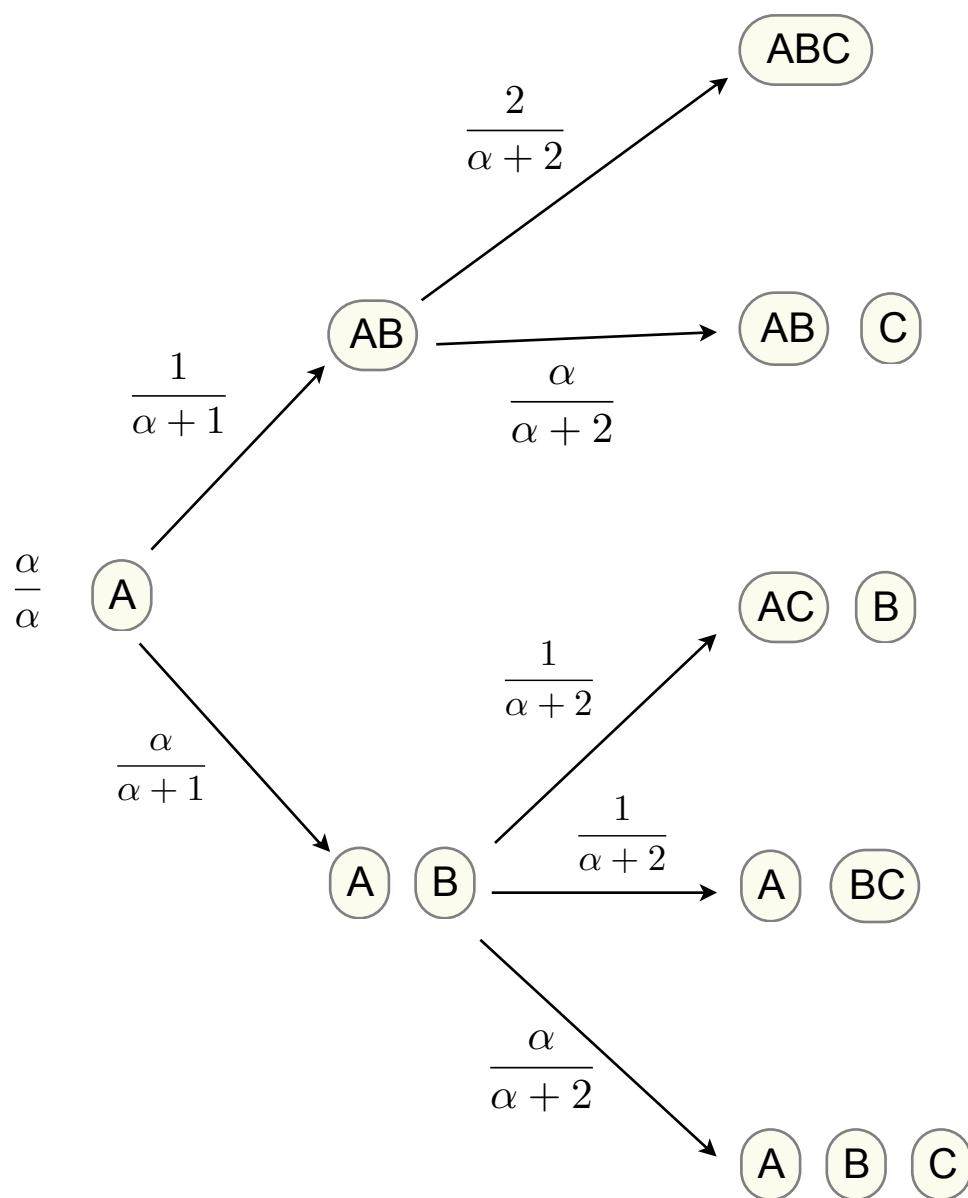
- classifies sites into meaningful categories
- discourages large numbers of categories
- assigns reasonable parameter values to each of the categories
- does all this automatically



Imagine you have a collection of objects (e.g. sites, codons) labeled A, B, C, ...

B can either be added to A's group or form its own group

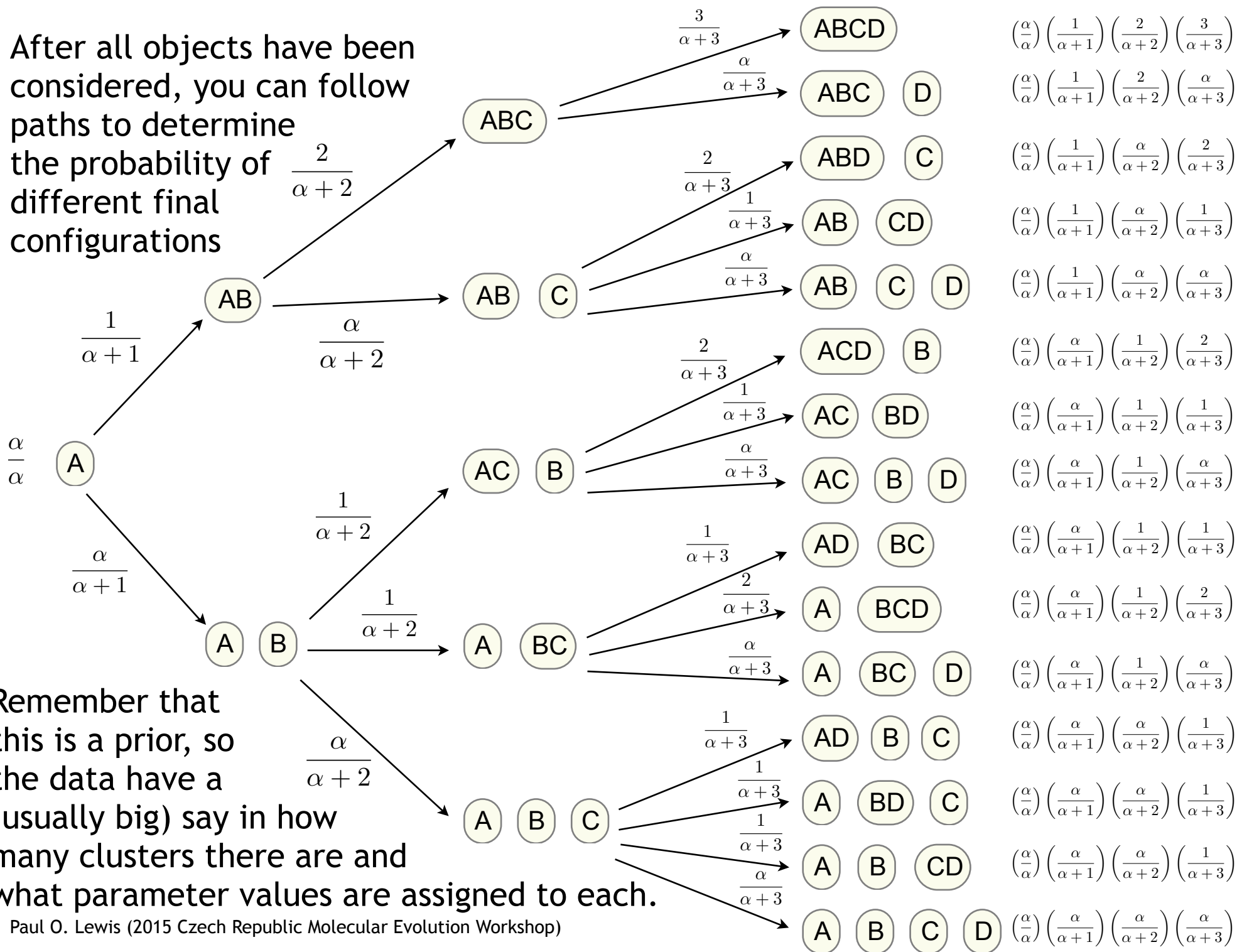
The parameter α determines the propensity for forming a new group



The third object C can either be added to an existing group...

...or form its own group

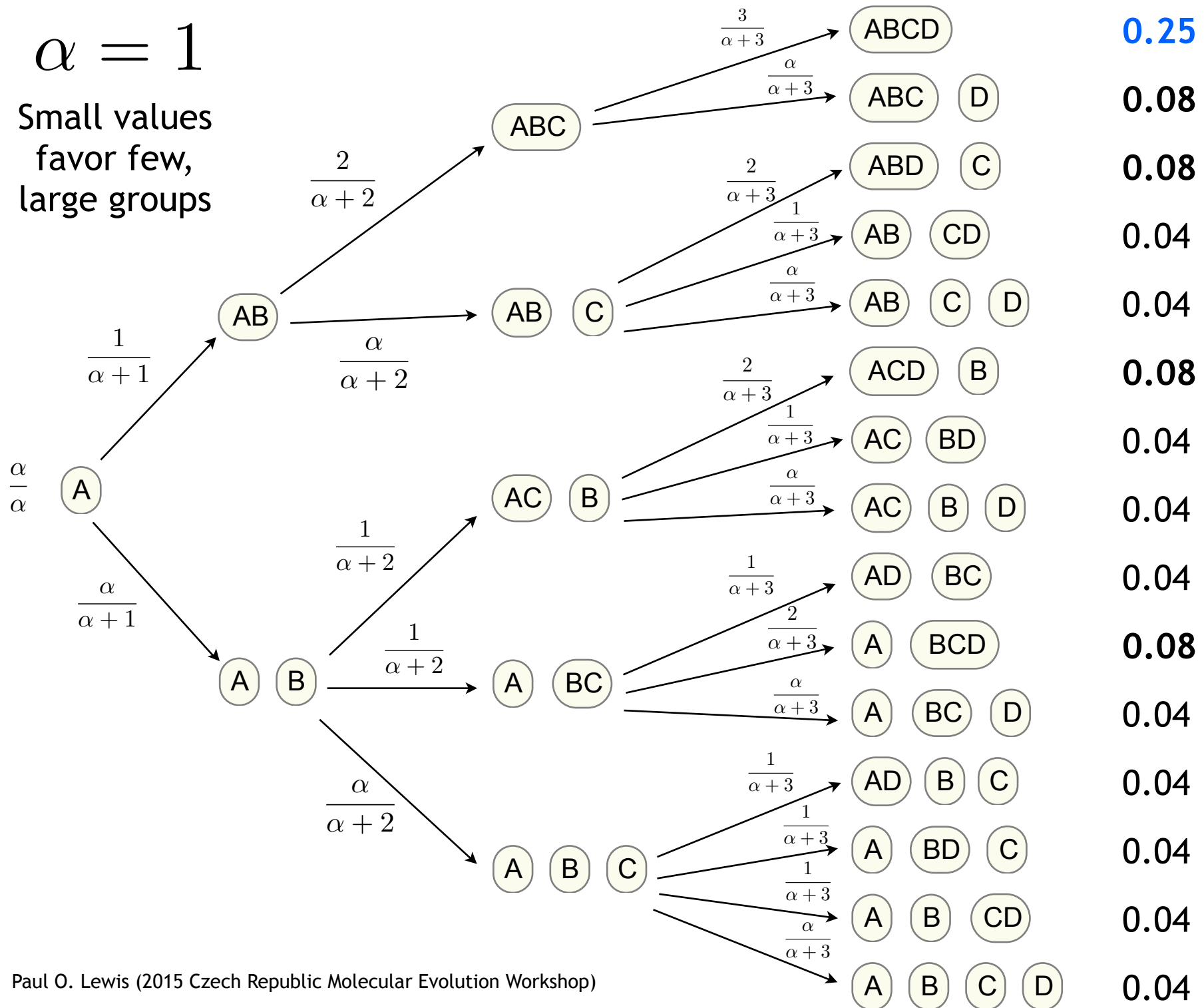
After all objects have been considered, you can follow paths to determine the probability of different final configurations



Remember that this is a prior, so the data have a (usually big) say in how many clusters there are and what parameter values are assigned to each.

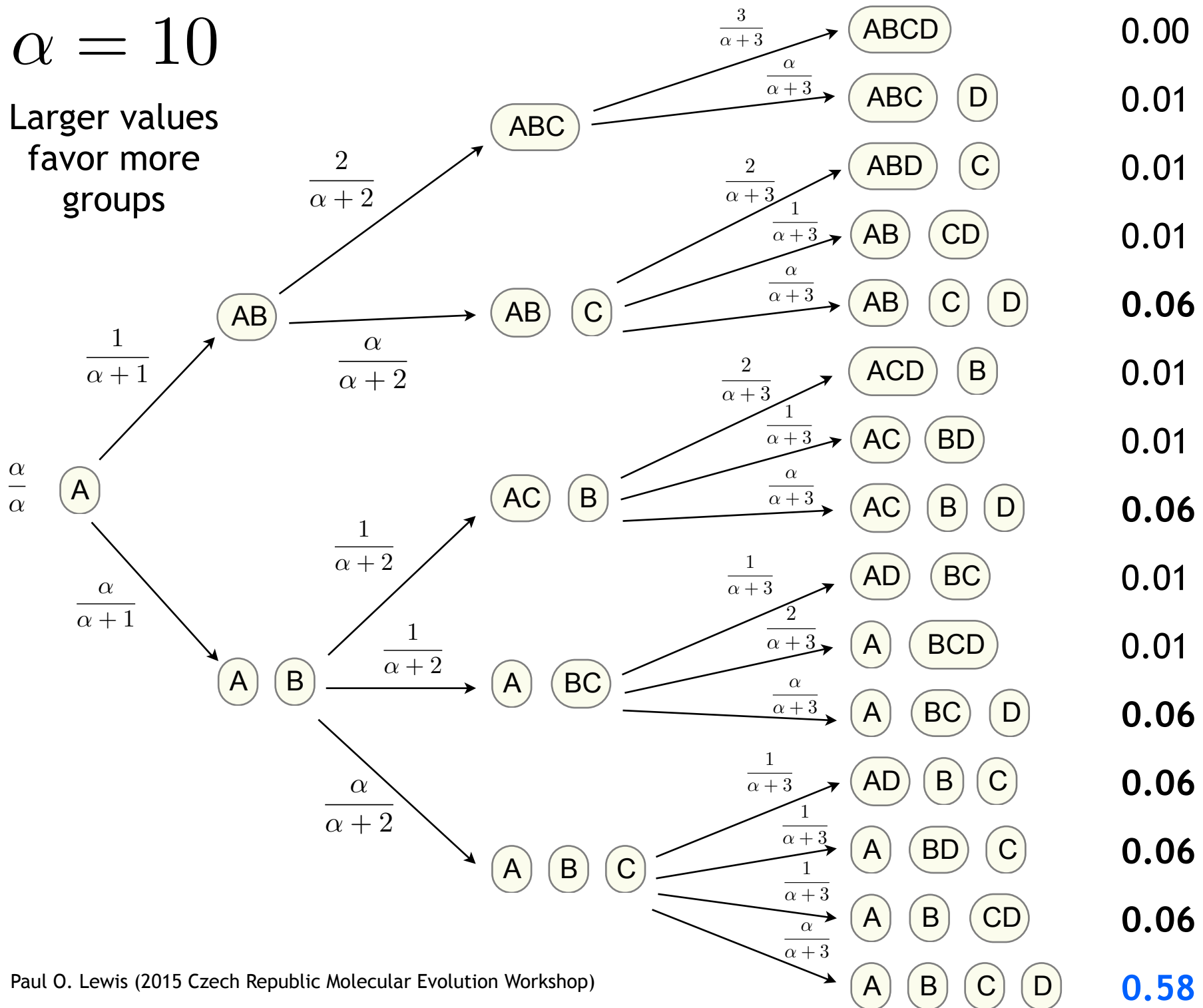
$$\alpha = 1$$

Small values
favor few,
large groups



$$\alpha = 10$$

Larger values
favor more
groups



Dirichlet Process Priors

- To encourage **few, large** groups, use a **small** alpha value
- To encourage **lots of small** groups, use a **large** alpha value
- In practice, **hierarchical models** are often used (i.e. alpha is a hyperparameter that is estimated, so you need not worry about choosing the appropriate value for alpha)
- Bottom line: DP models are very nice for automatically grouping sites into clusters that have some property in common

The End