

# Bayesian Phylogenetics

Paul O. Lewis

Department of Ecology & Evolutionary Biology  
University of Connecticut

28 January 2011

Workshop on Molecular Evolution  
Český Krumlov, Česká republika

# An Introduction to Bayesian Phylogenetics

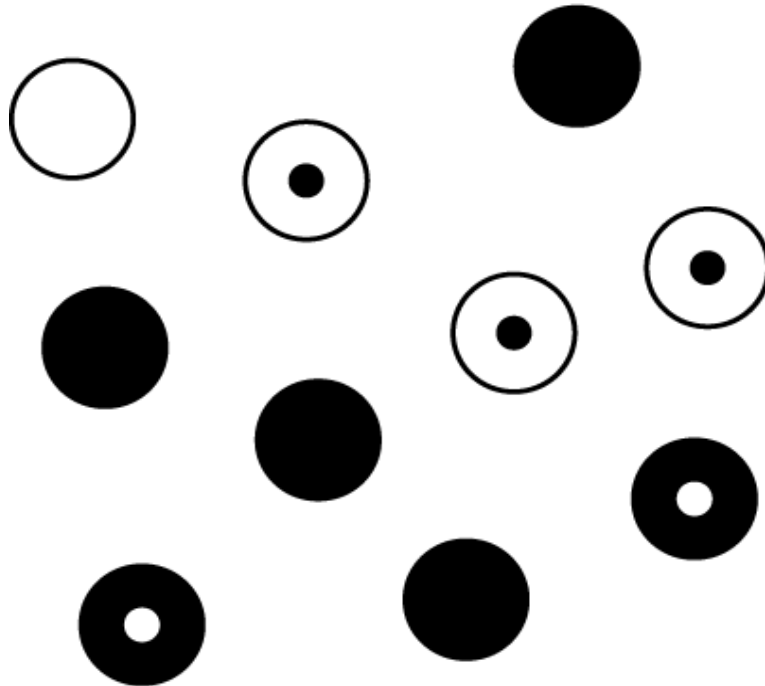
- Bayesian inference in general
- Markov chain Monte Carlo (MCMC)
- Bayesian phylogenetics
- Prior distributions
- Bayesian model selection

# I. Bayesian inference in general

# Joint probabilities

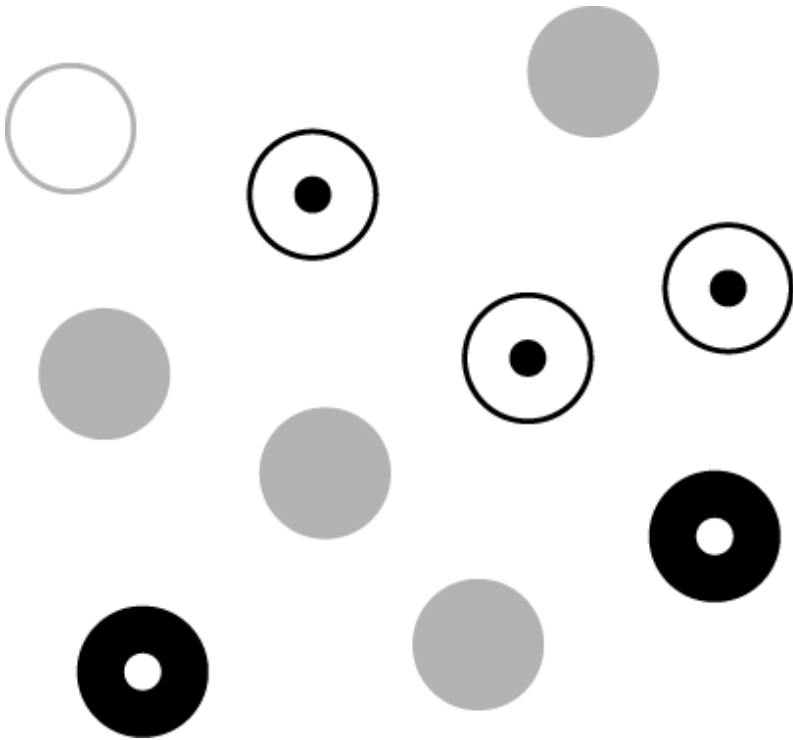
**B = Black**      **S = Solid**  
**W = White**      **D = Dotted**

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



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

# Conditional probabilities



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

Hide all solid marbles  
(leaving 5 with dot)

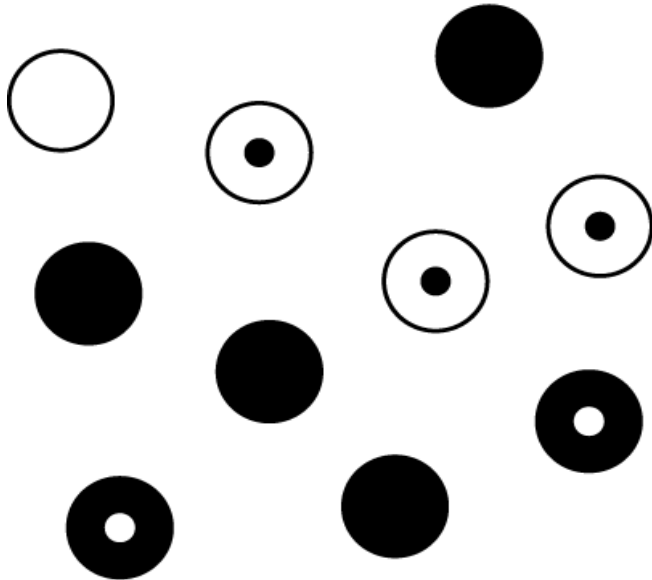
Of those left, 2 are black

# Bayes' rule

$\Pr(B, D)$

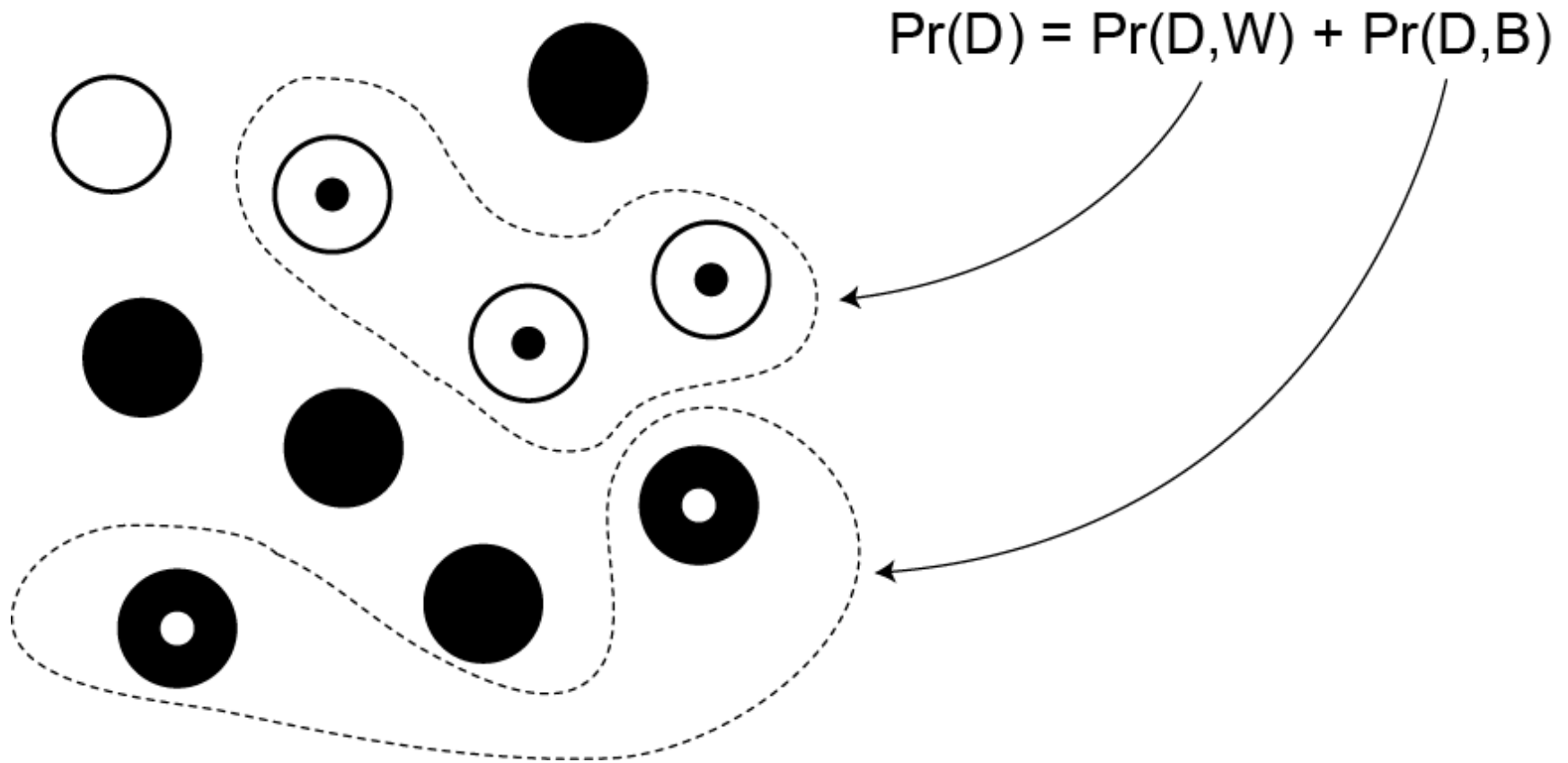
$$\Pr(D) \Pr(B|D) = \Pr(B) \Pr(D|B)$$

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



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

# Probability of "Dotted"



## Bayes' rule (cont.)

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

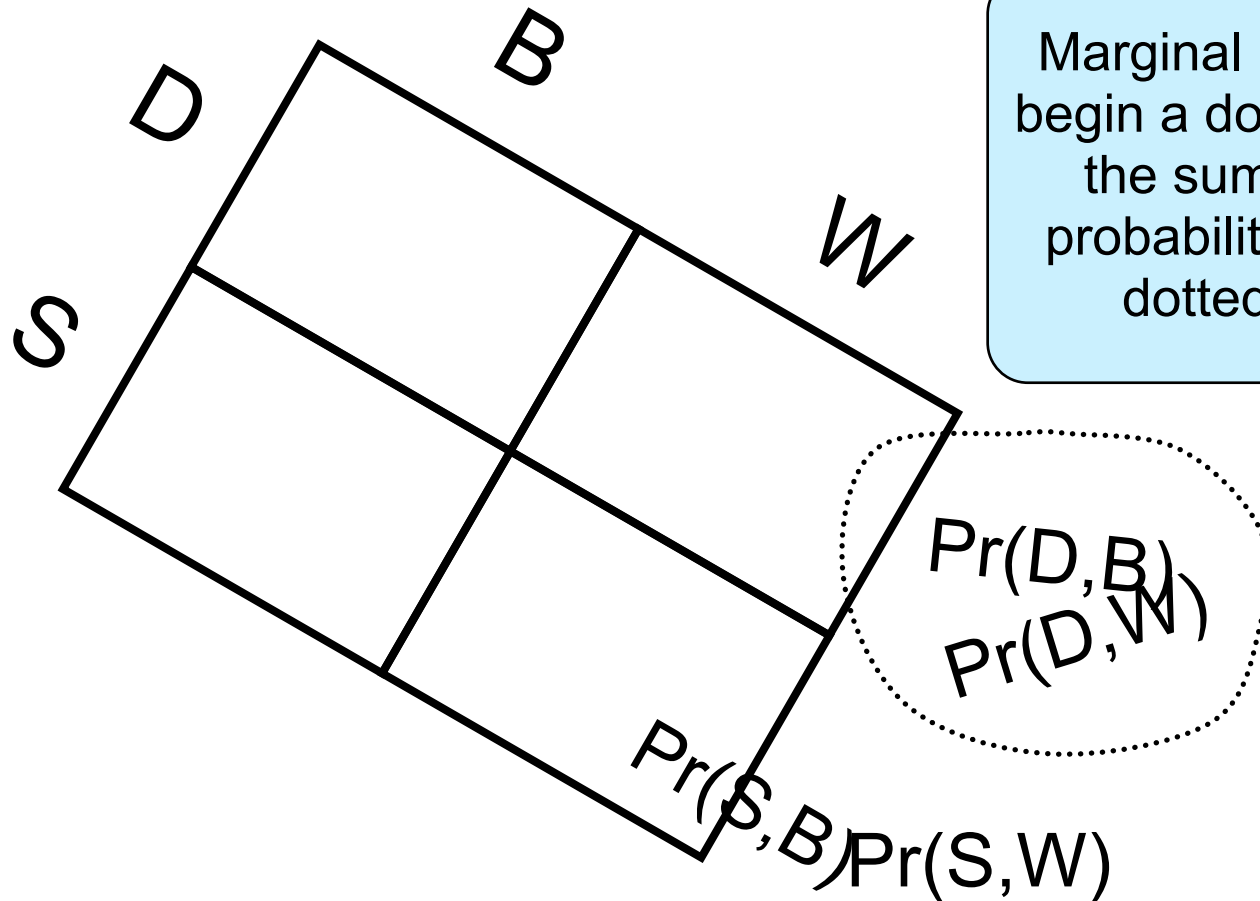
$\Pr(D)$  is the **marginal probability** of being dotted  
To compute it, we **marginalize over colors**



# Joint probabilities

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

# Marginalizing over colors

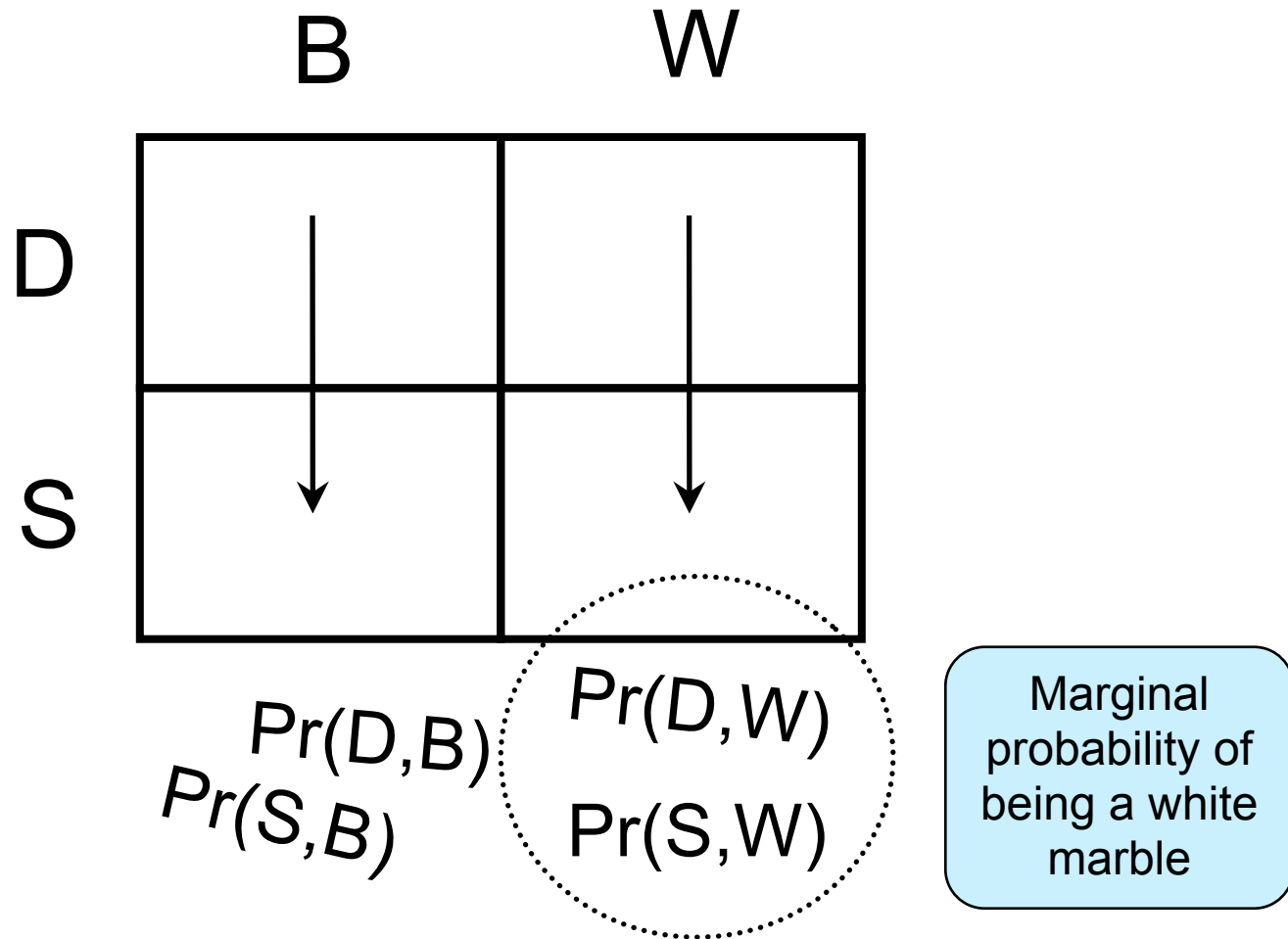


Marginal probability of begin a dotted marble is the sum of all joint probabilities involving dotted marbles

# Marginal probabilities

	B	W	
D			Marginal probability of being dotted $\Pr(D,B) + \Pr(D,W)$
S			$\Pr(S,B) + \Pr(S,W)$ Marginal probability of being solid

# 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**)

$\theta$  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  $\theta$       **Prior probability** of hypothesis  $\theta$

**Posterior probability** of hypothesis  $\theta$       **Marginal probability of the data** (marginalizing over hypotheses)

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

The diagram shows the equation for Bayes' rule. The numerator consists of two terms:  $\Pr(D|\theta)$  and  $\Pr(\theta)$ . The denominator is a sum over all hypotheses  $\theta$  of the product  $\Pr(D|\theta) \Pr(\theta)$ . Arrows point from the labels to these components: 'Likelihood of hypothesis  $\theta$ ' points to  $\Pr(D|\theta)$ ; 'Prior probability of hypothesis  $\theta$ ' points to  $\Pr(\theta)$ ; 'Posterior probability of hypothesis  $\theta$ ' points to  $\Pr(\theta|D)$ ; and 'Marginal probability of the data (marginalizing over hypotheses)' points to the denominator.

## Simple (albeit silly) paternity example

$\theta_1$  and  $\theta_2$  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	$\theta_1$	$\theta_2$	Row sum
Genotypes	<b>AA</b>	<b>Aa</b>	---
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. (Note that we do not need to consider the case of a negative test result.)  
How worried should I be?

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

Likelihood

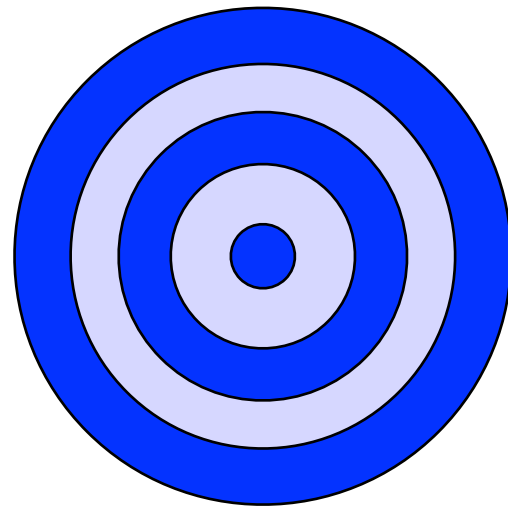
Prior probability density

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

Posterior probability density

Marginal probability of the data

# If you had to guess...



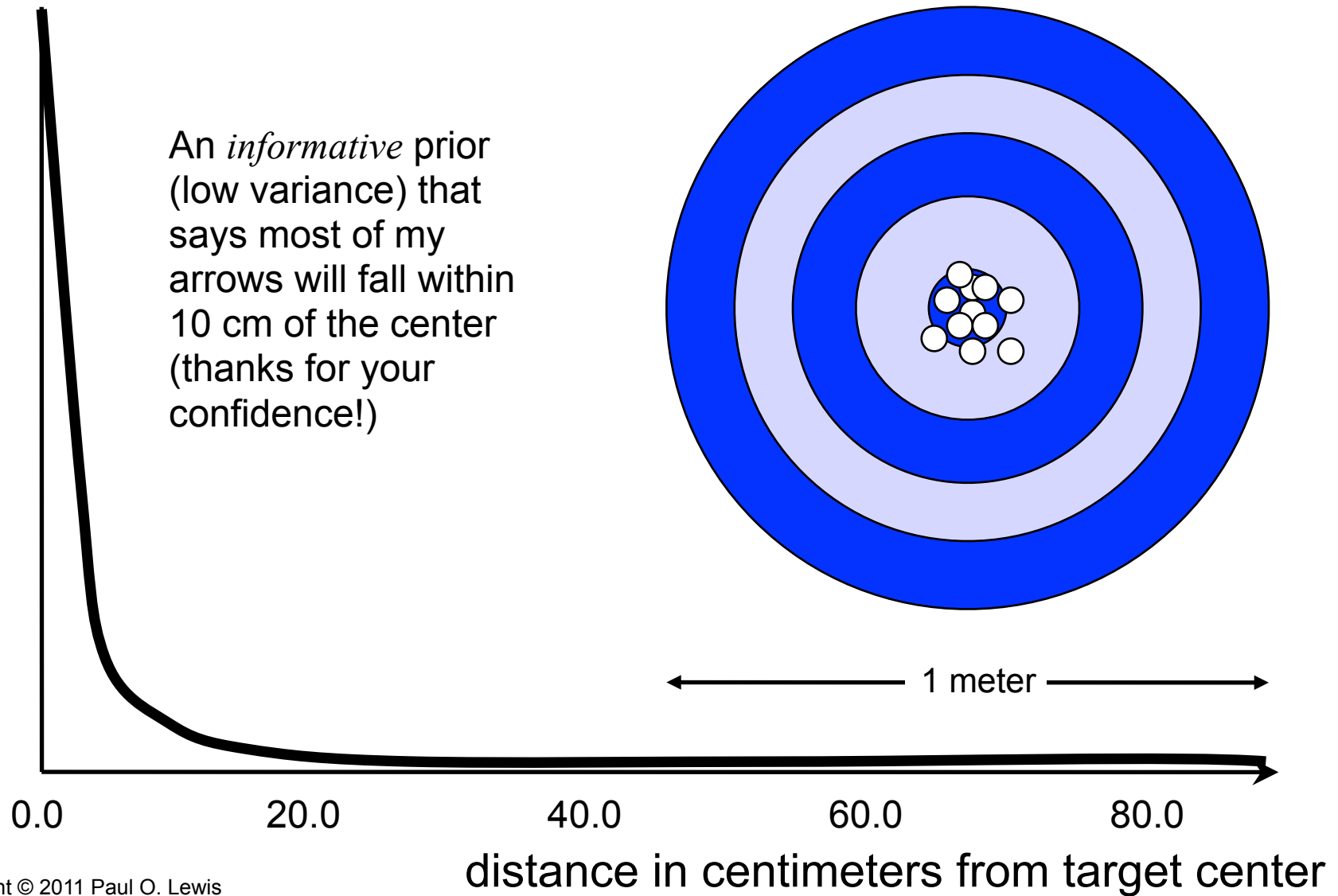
*Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance  $d$  from the center of the target (if it helps, I'm standing 50 meters away from the target)*

0.0

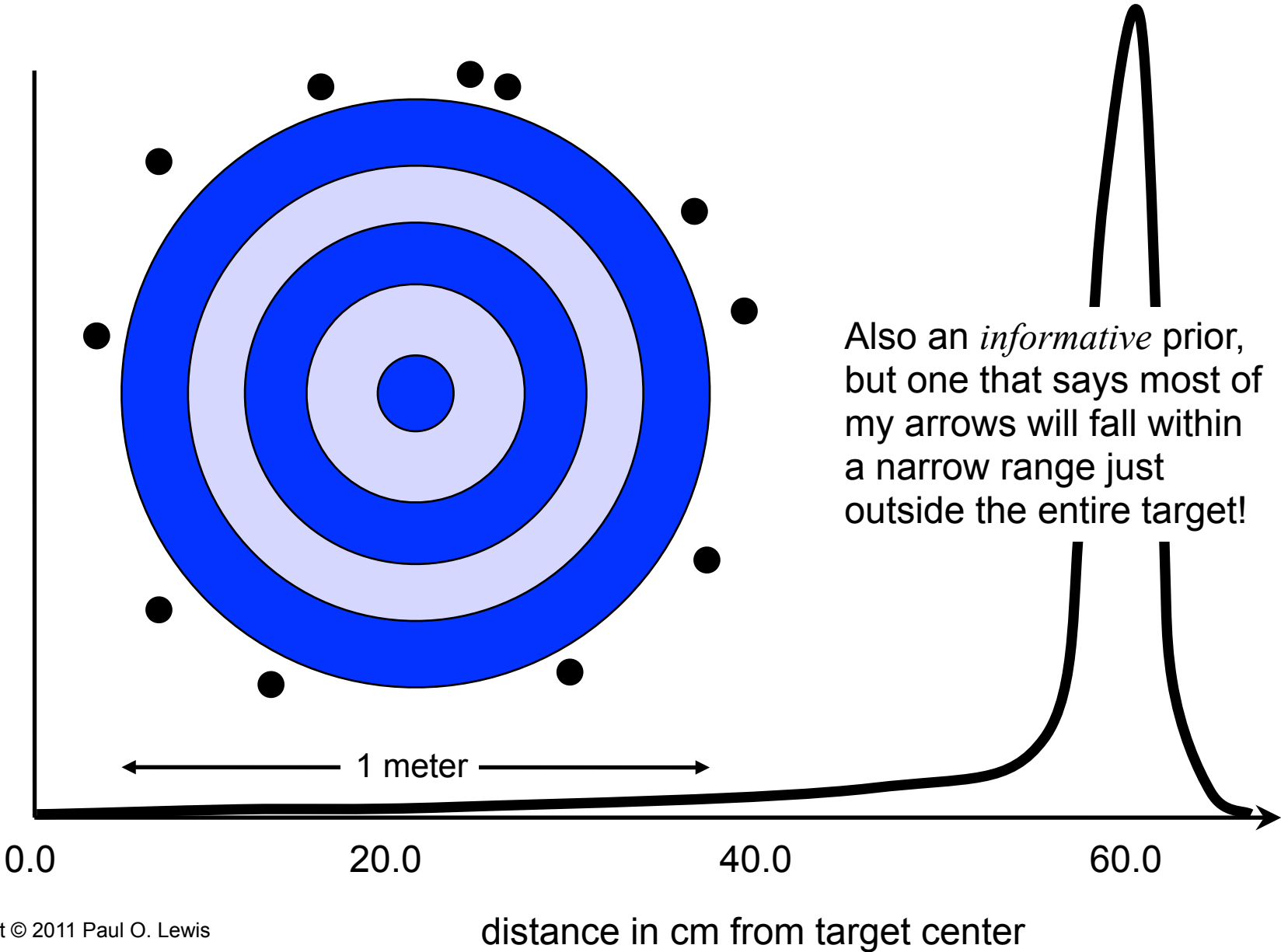
$d$

$\infty$

# Case 1: assume I have talent

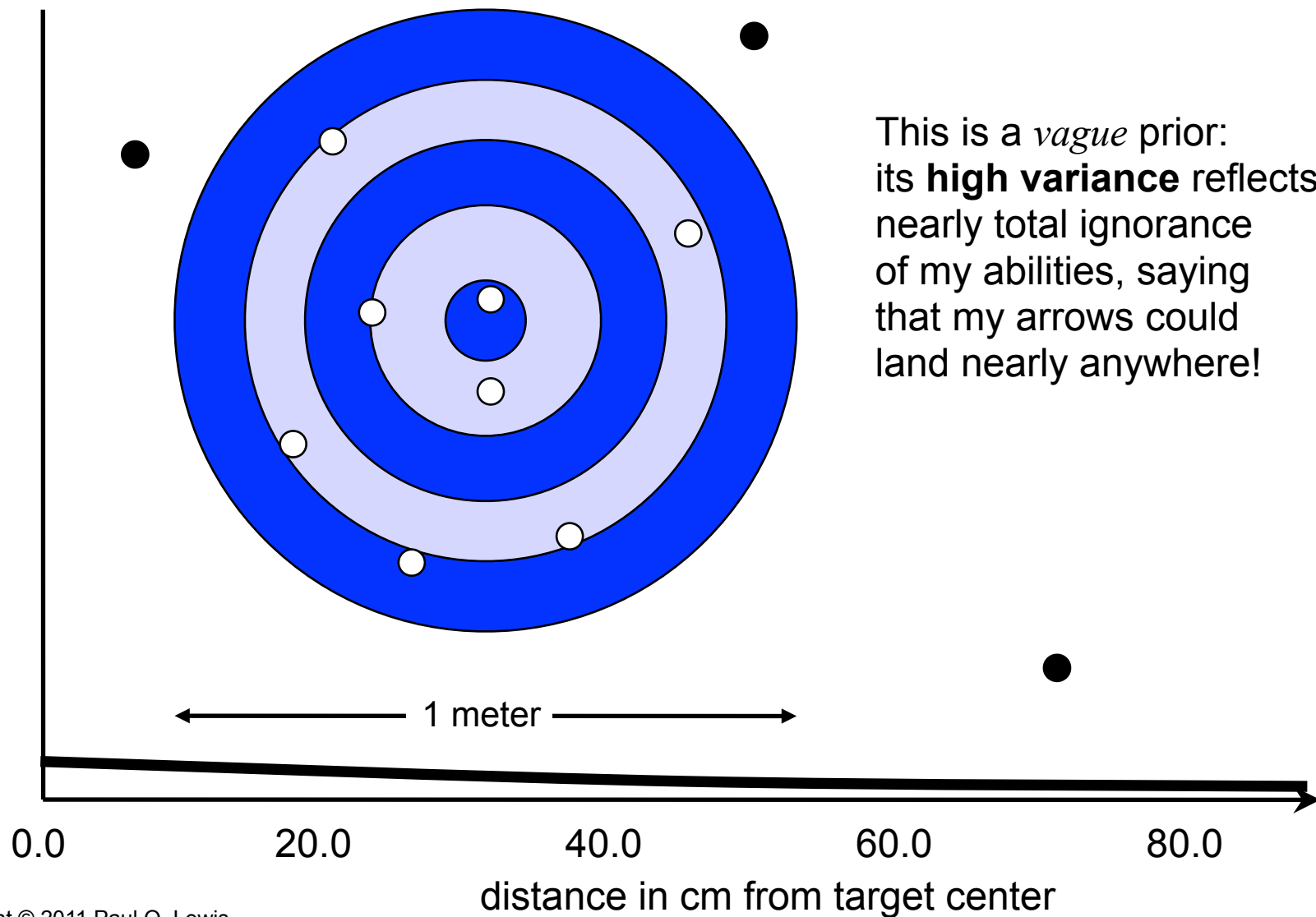


# Case 2: assume I have a talent for missing the target!





### Case 3: assume I have no talent

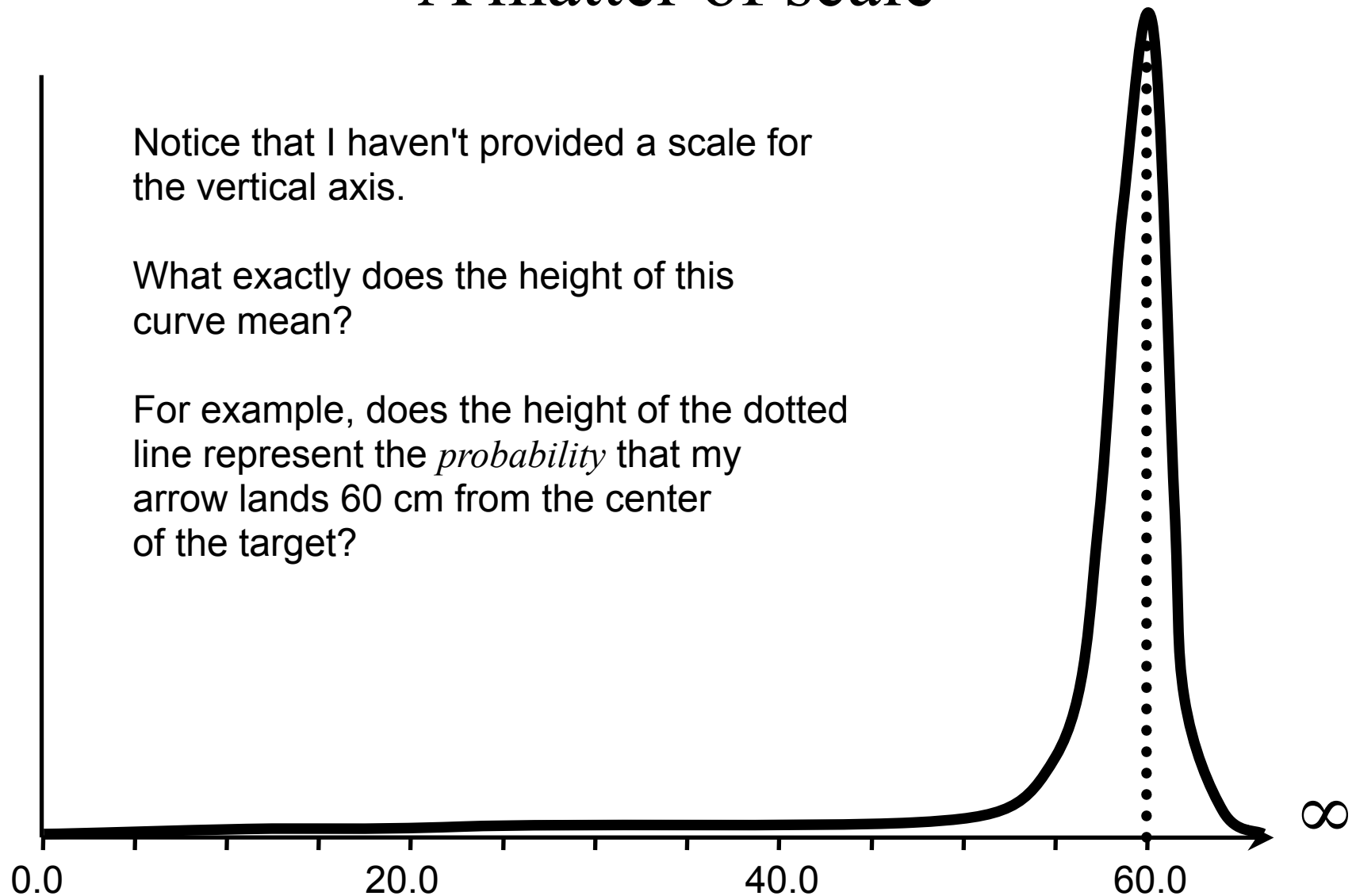


# A matter of scale

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?

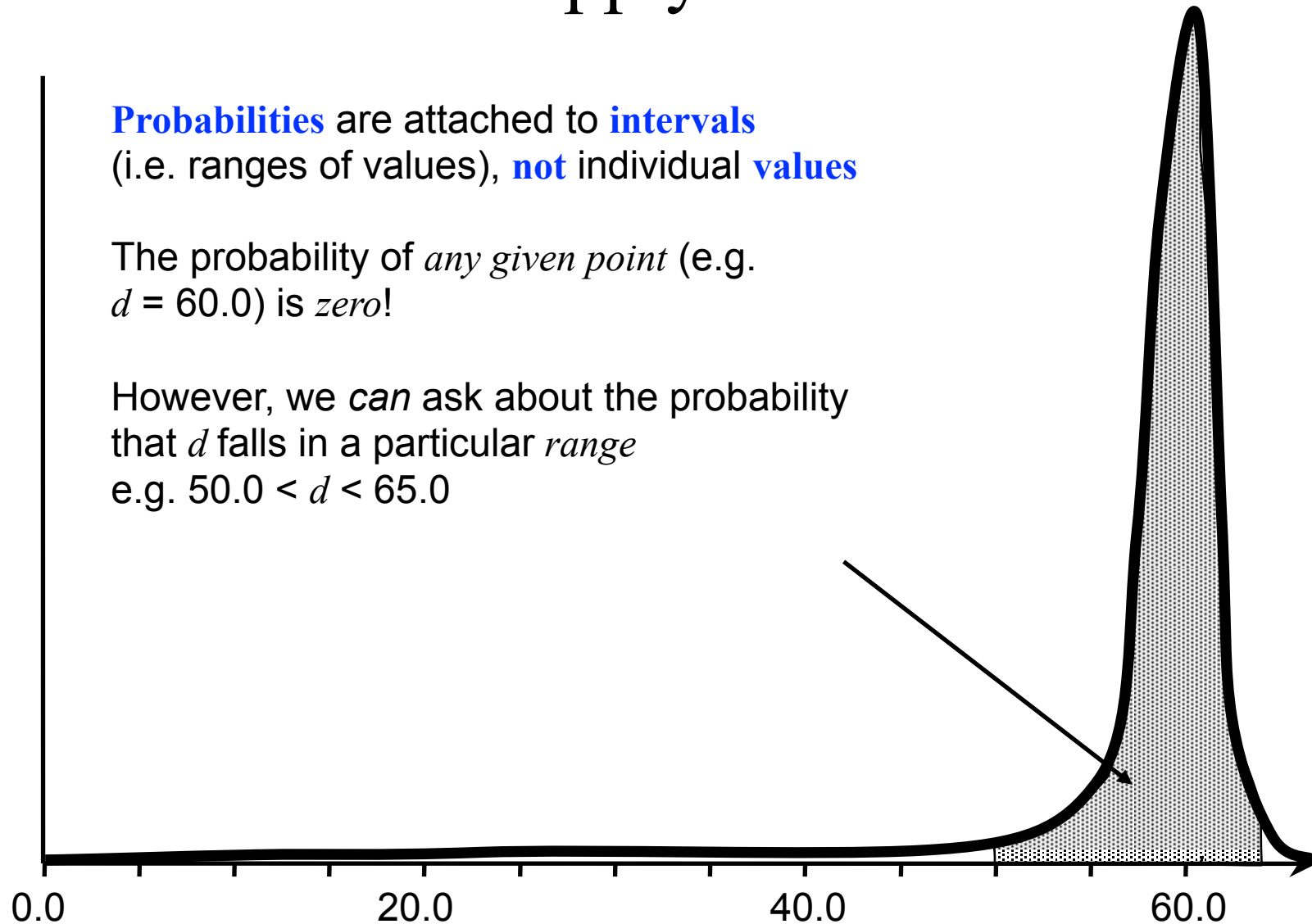


# Probabilities apply to intervals

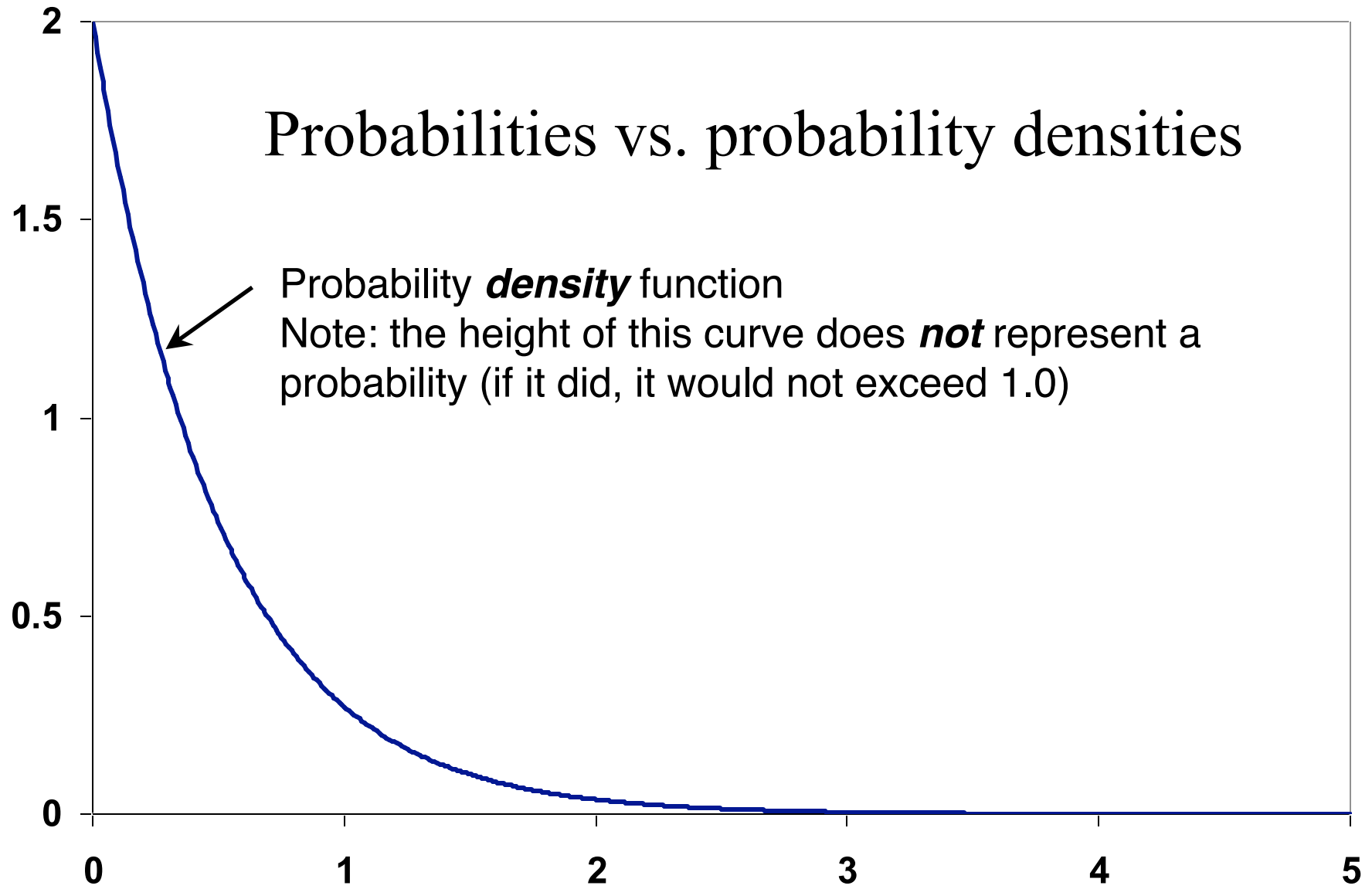
**Probabilities** are attached to **intervals**  
(i.e. ranges of values), **not** individual **values**

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



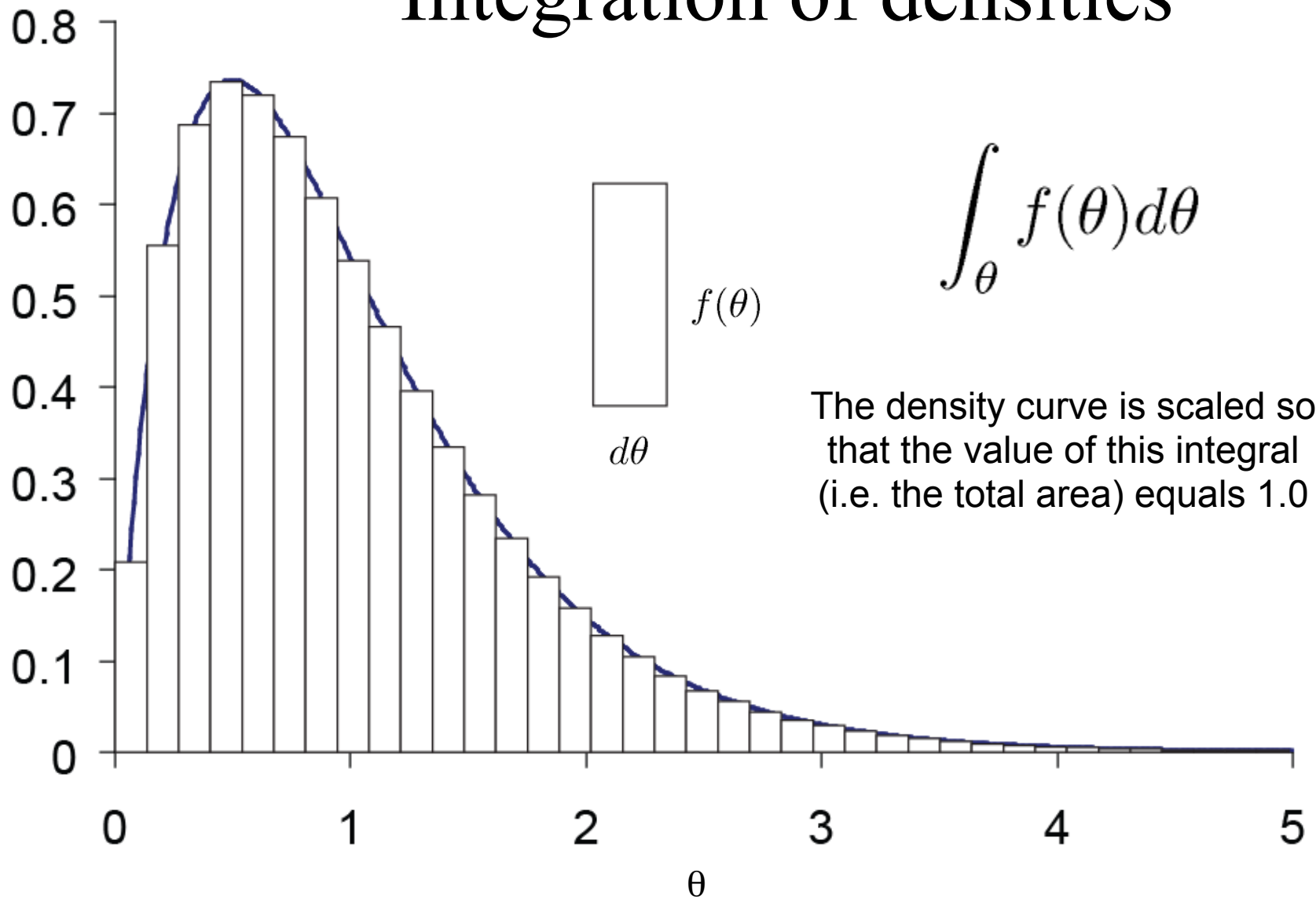
# Densities of various substances

Substance	Density (g/cm <sup>3</sup> )
Cork	0.24
Aluminum	2.70
Gold	19.30

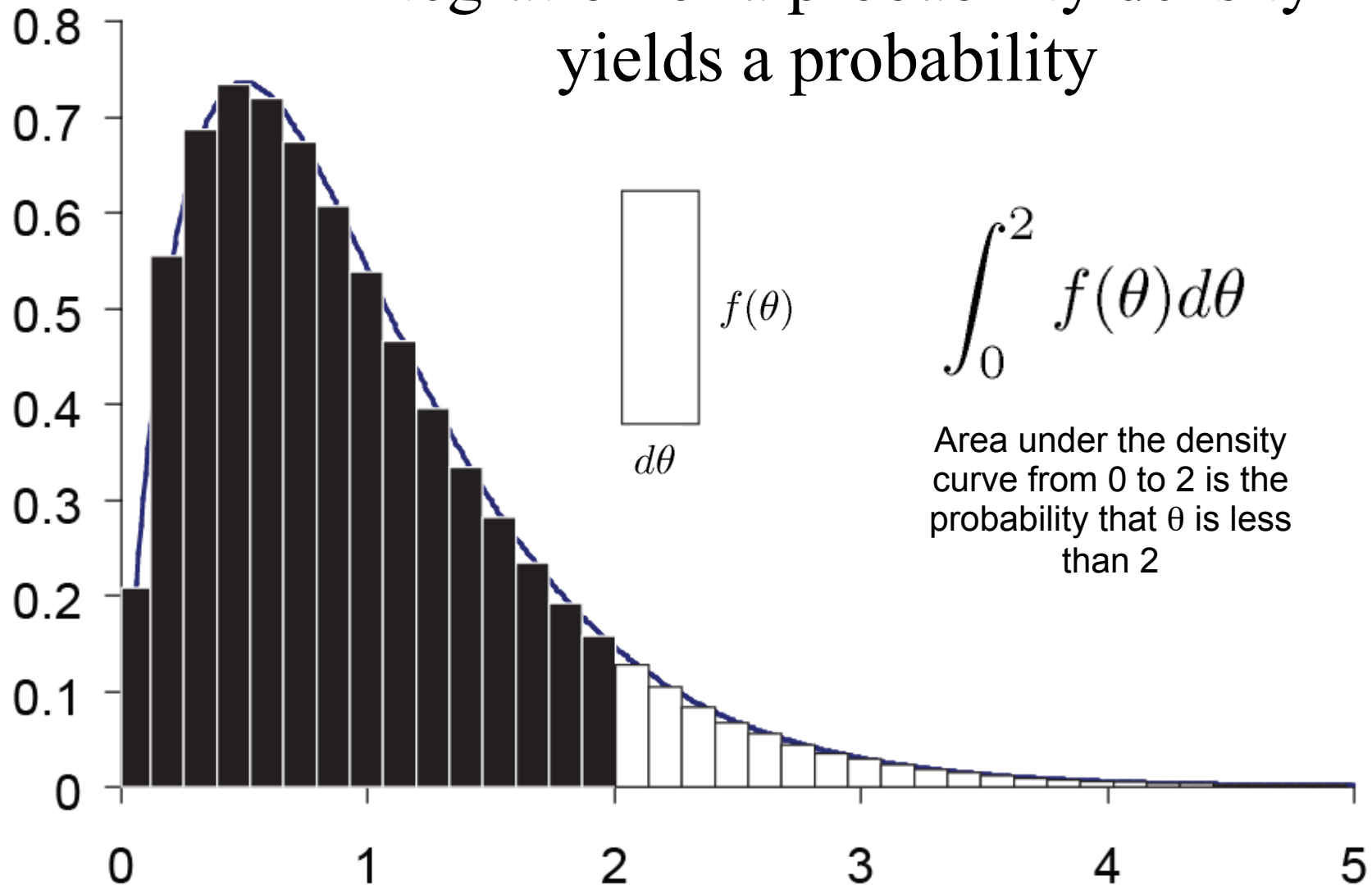
*Density does not equal mass*  
mass = density × volume

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

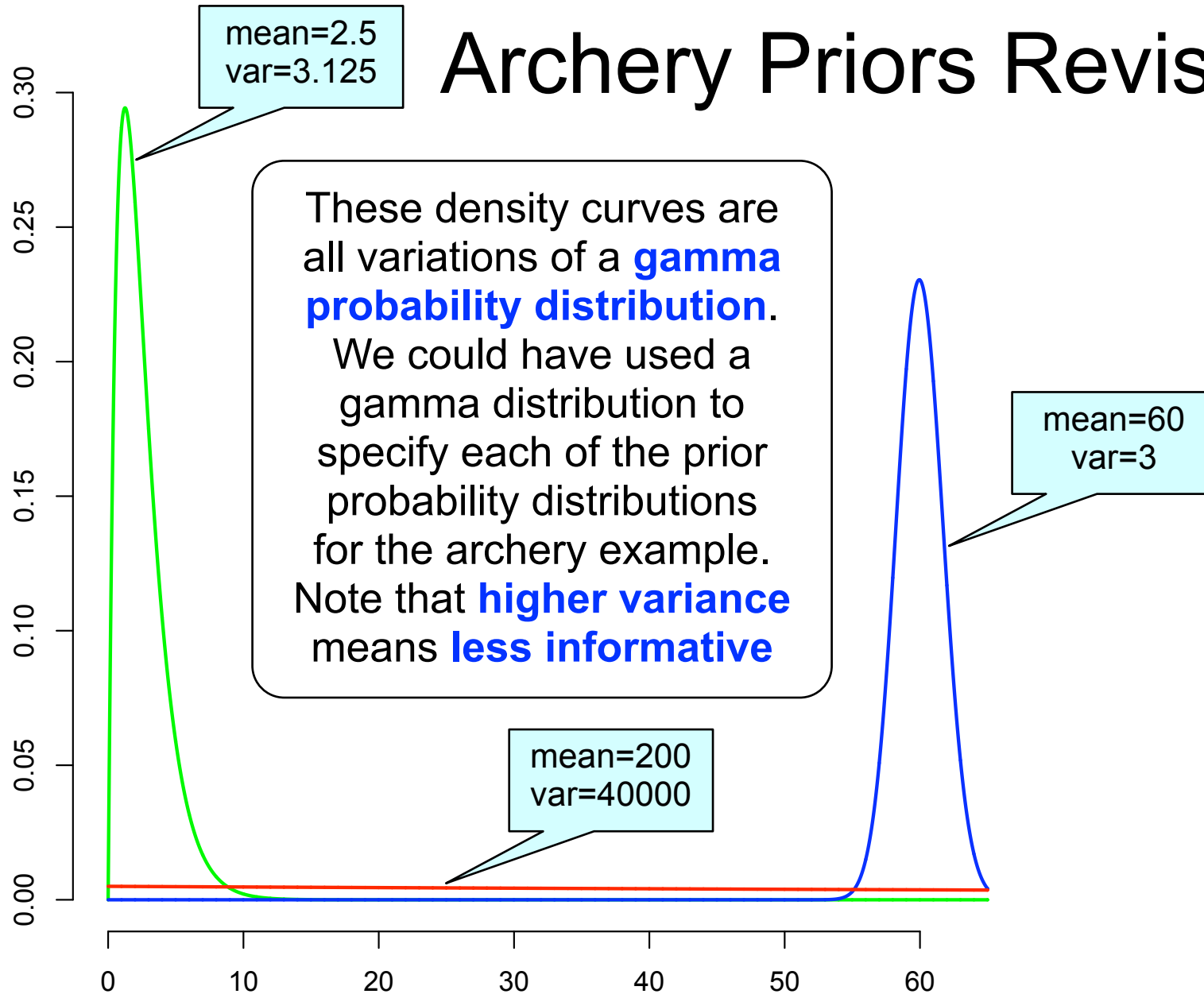
# Integration of densities



# 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

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

Note that the same formula serves as both the:

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

# Likelihood vs. Probability

Outcome	Fair coin model	Two-heads model
H	0.5	1.0
T	0.5	0.0
	1.0	1.0

$L(M | D)$

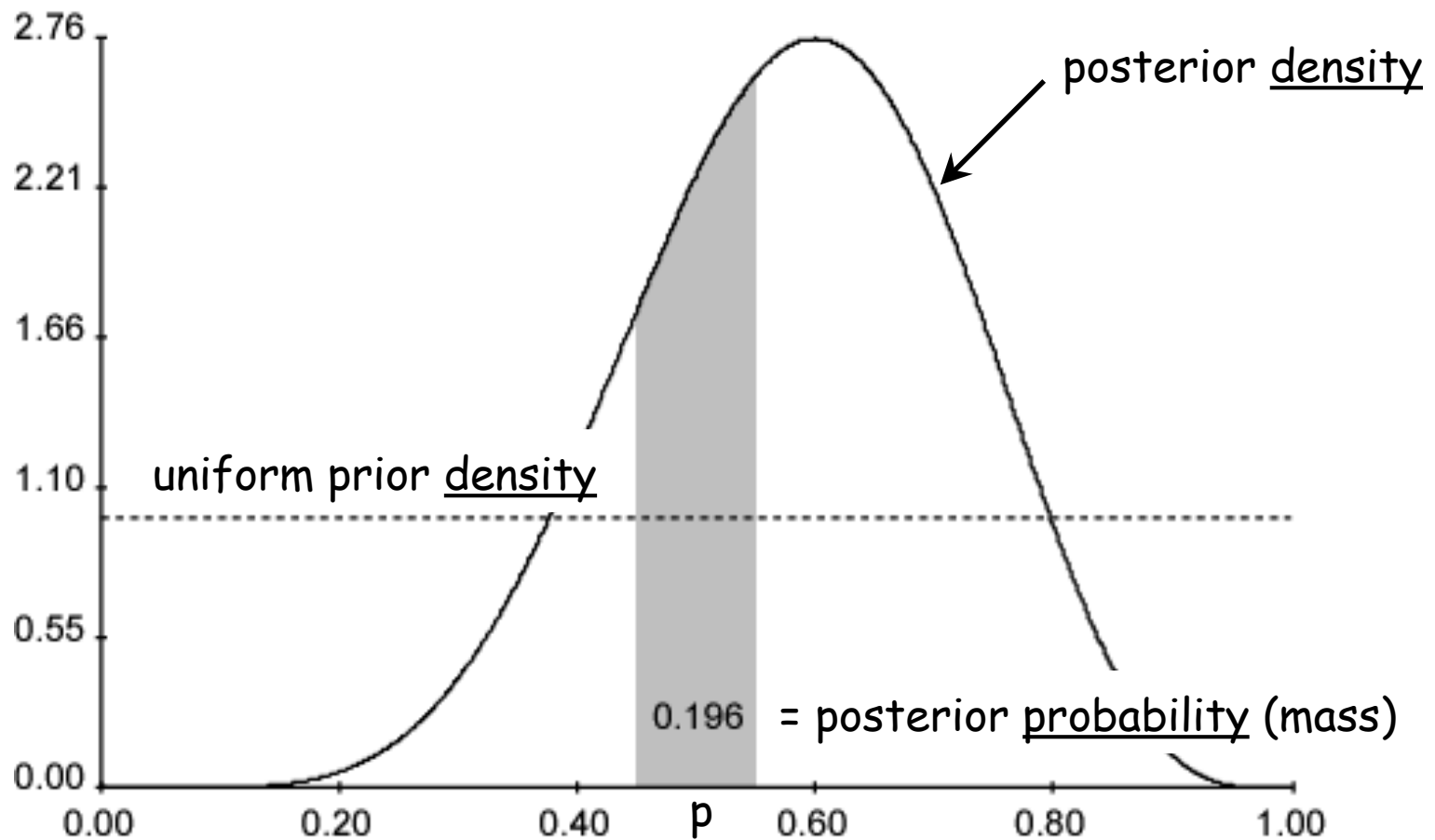
$\Pr(D | M)$

**Probabilities**  
are functions of  
the data (the  
model is fixed)

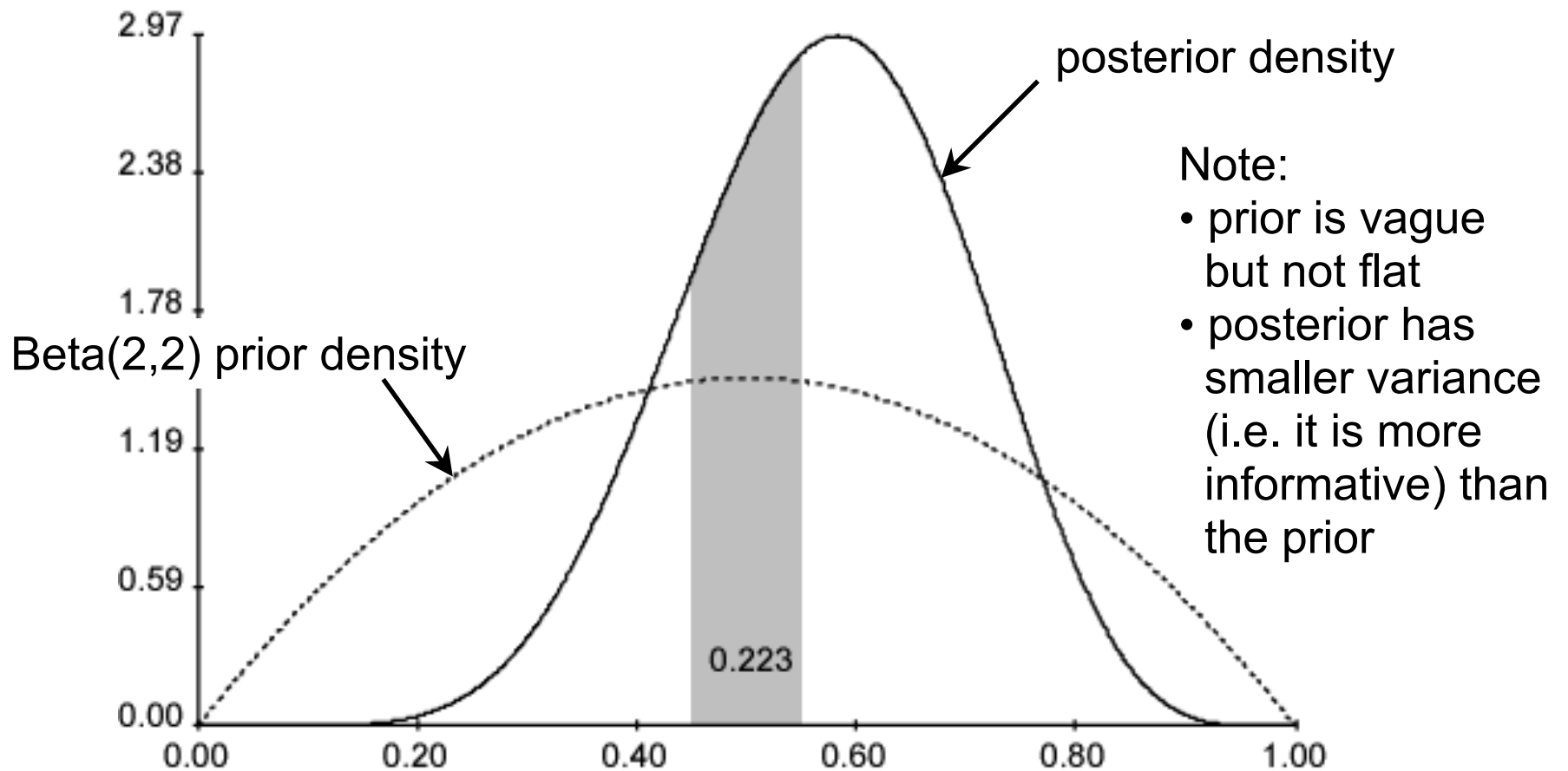
**Likelihoods**  
are functions of  
models (the  
data is fixed)

- Symbol  $\Pr(D|M)$  often used to represent the likelihood (I will be following convention in this regard)

The posterior is generally more informative than the prior (data contains information)



# Beta prior gives more flexibility



Note:

- prior is vague but not flat
- posterior has smaller variance (i.e. it is more informative) than the prior

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

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

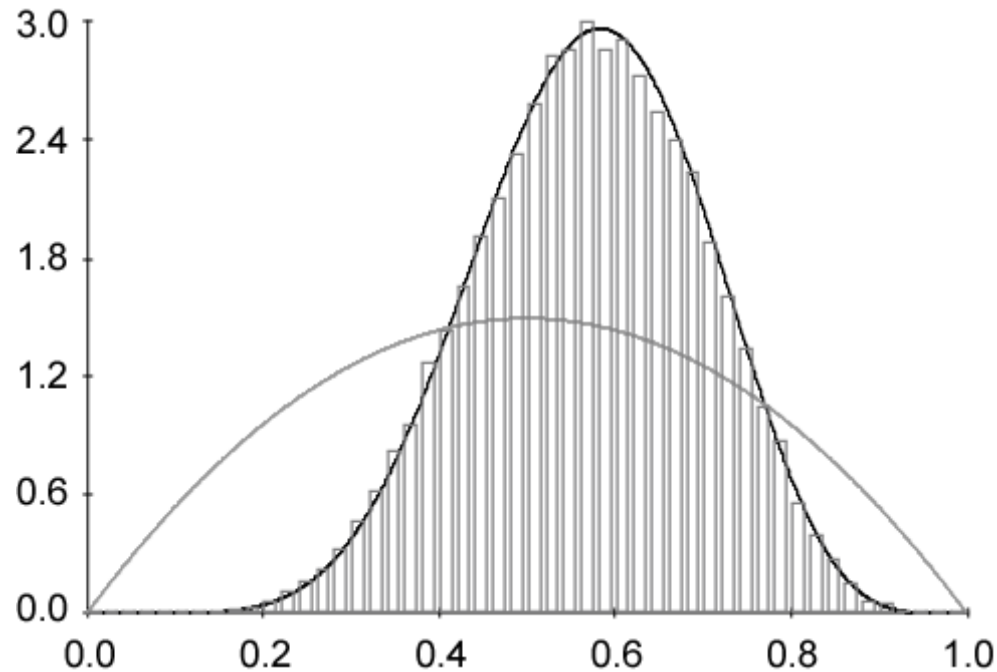
Marginal probability of data

↑  
Posterior  
probability  
density

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)

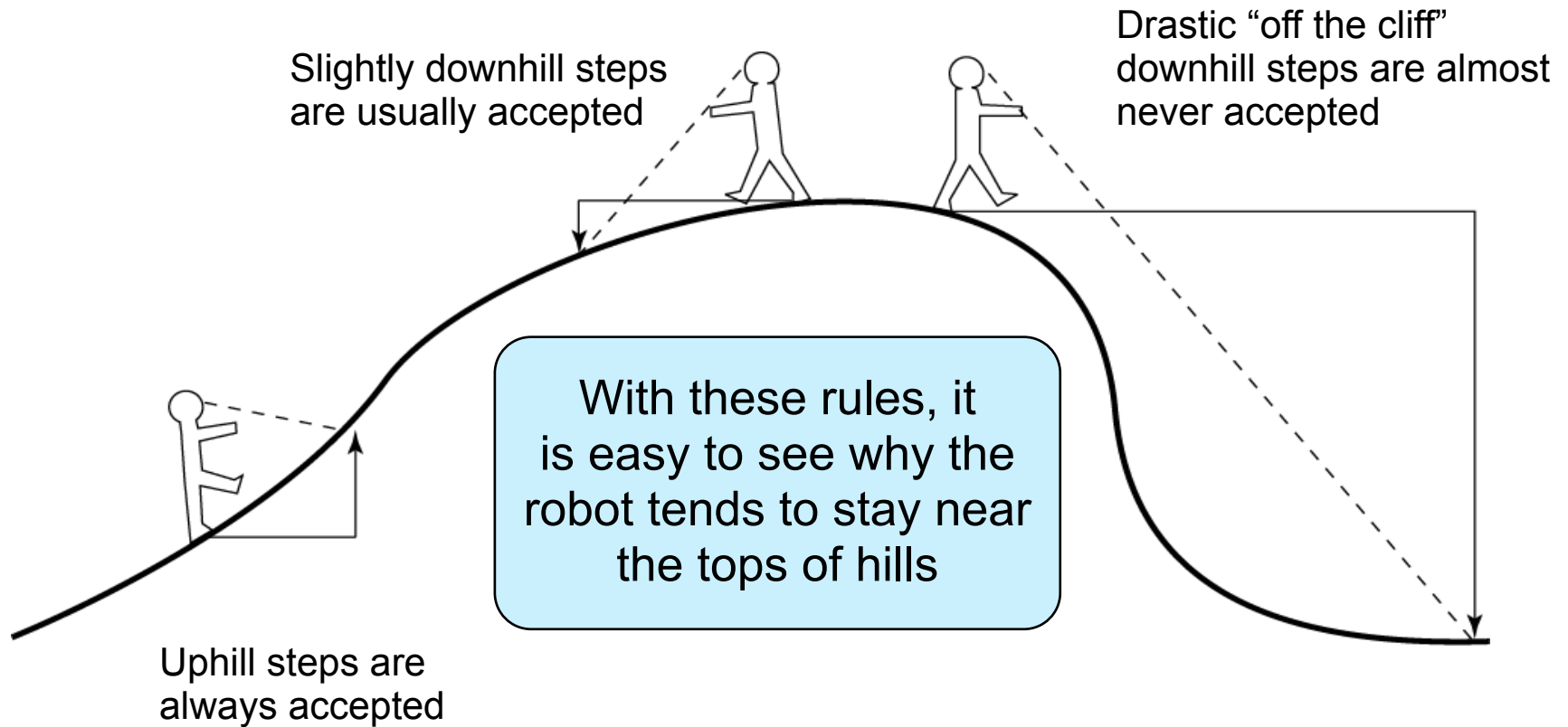


For more complex problems, we might settle for a

**good approximation**

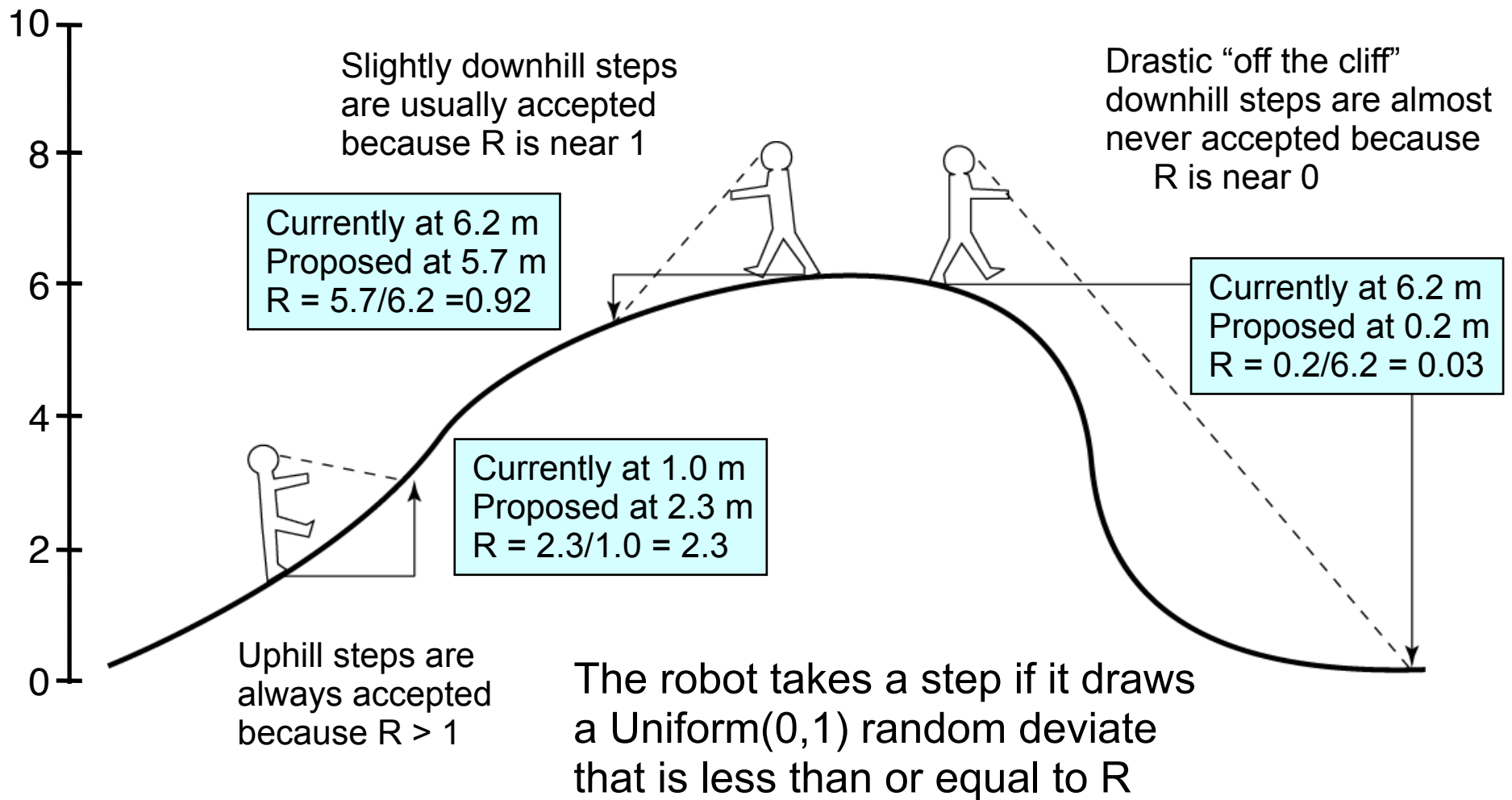
to the posterior distribution

# 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

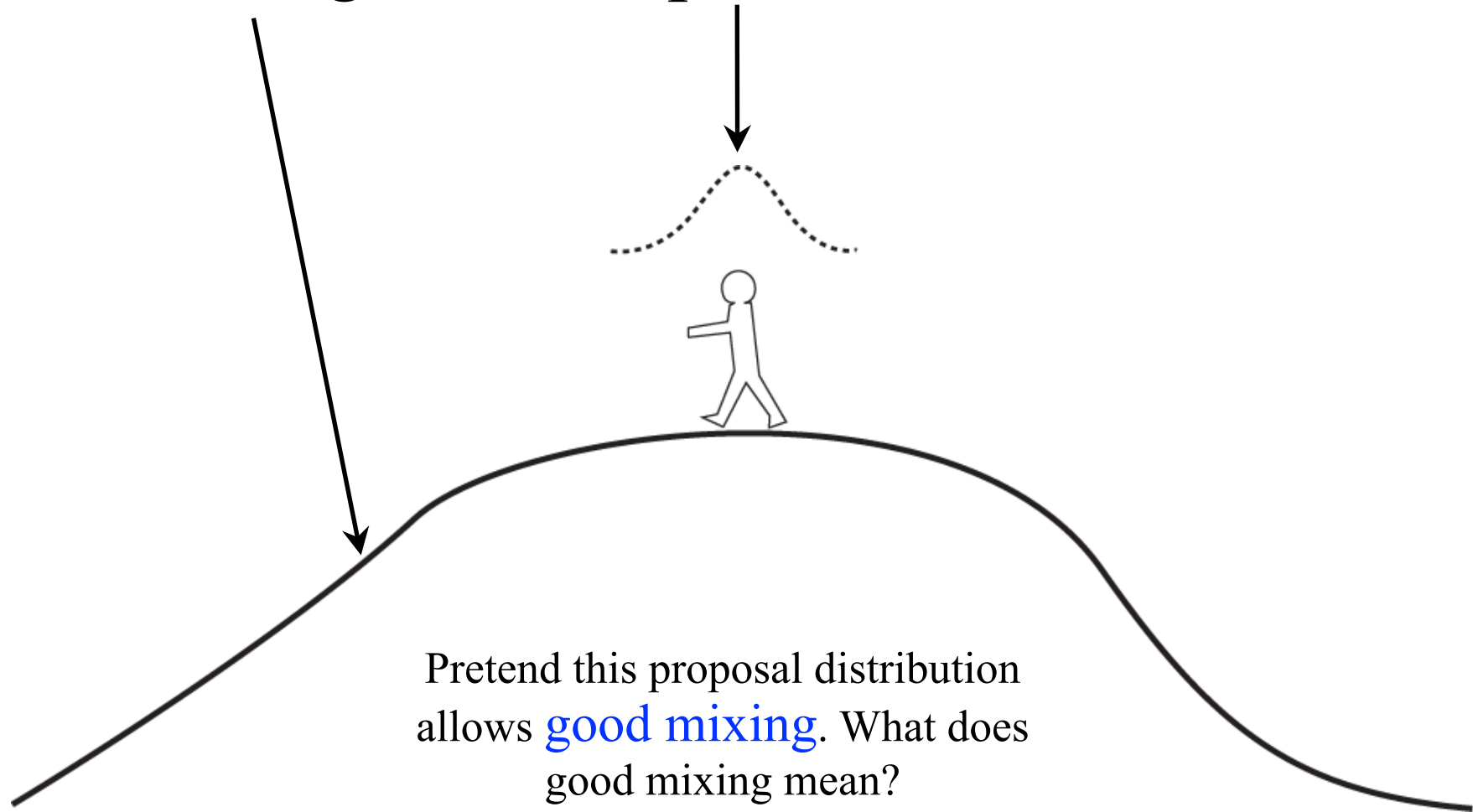
Likelihood  
ratio

Prior odds

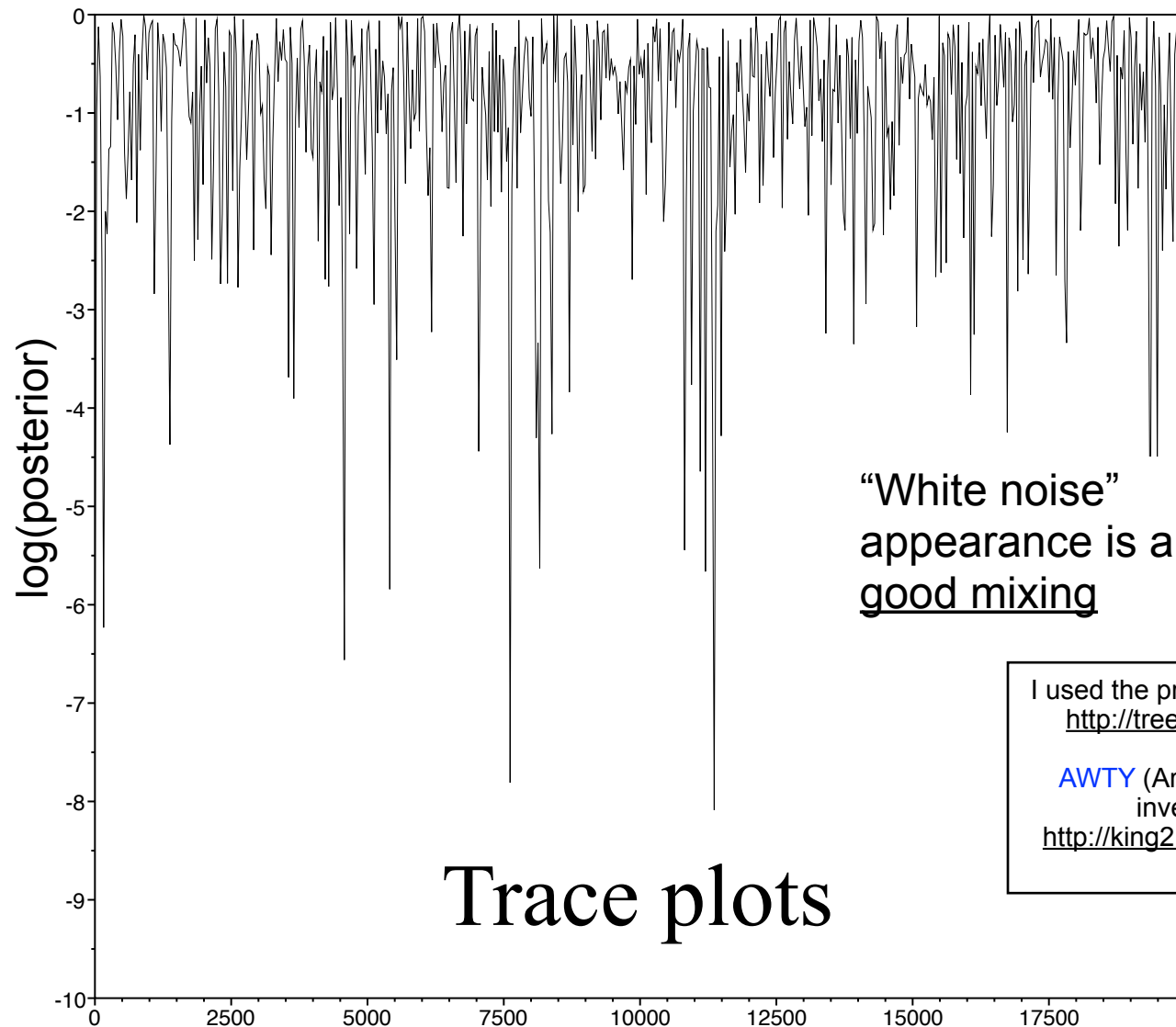
# Target vs. proposal distributions

- The **target distribution** is the posterior distribution of interest
- The **proposal distribution** is used to decide which point to try next
  - you have much flexibility here, and the choice affects only the **efficiency** of the MCMC algorithm
  - MCMC using a **symmetric** proposal distribution is the Metropolis algorithm (Metropolis et al. 1953)
  - Use of an **asymmetric** proposal distribution requires a modification proposed by Hastings (1970), and is known as the Metropolis-Hastings algorithm

# Target vs. Proposal Distributions



Pretend this proposal distribution allows **good mixing**. What does good mixing mean?



“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)

## Trace plots

# 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

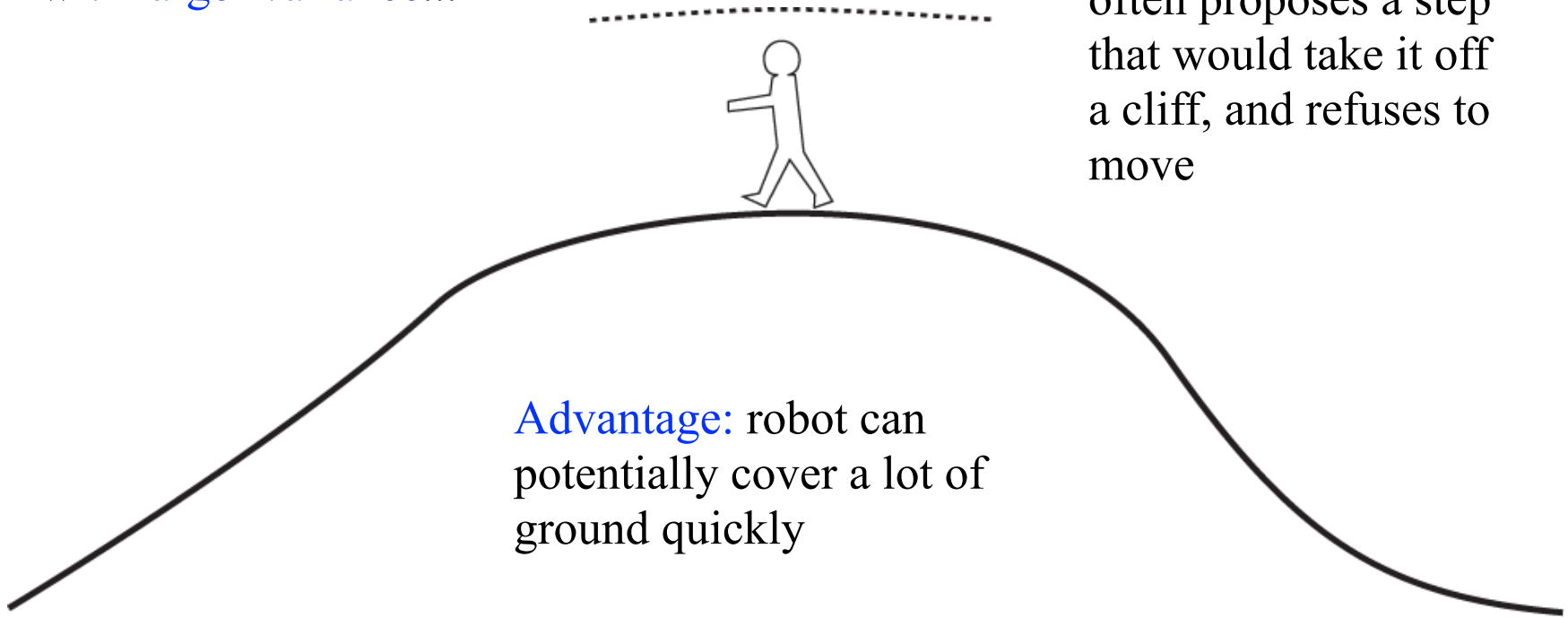


# Target vs. Proposal Distributions

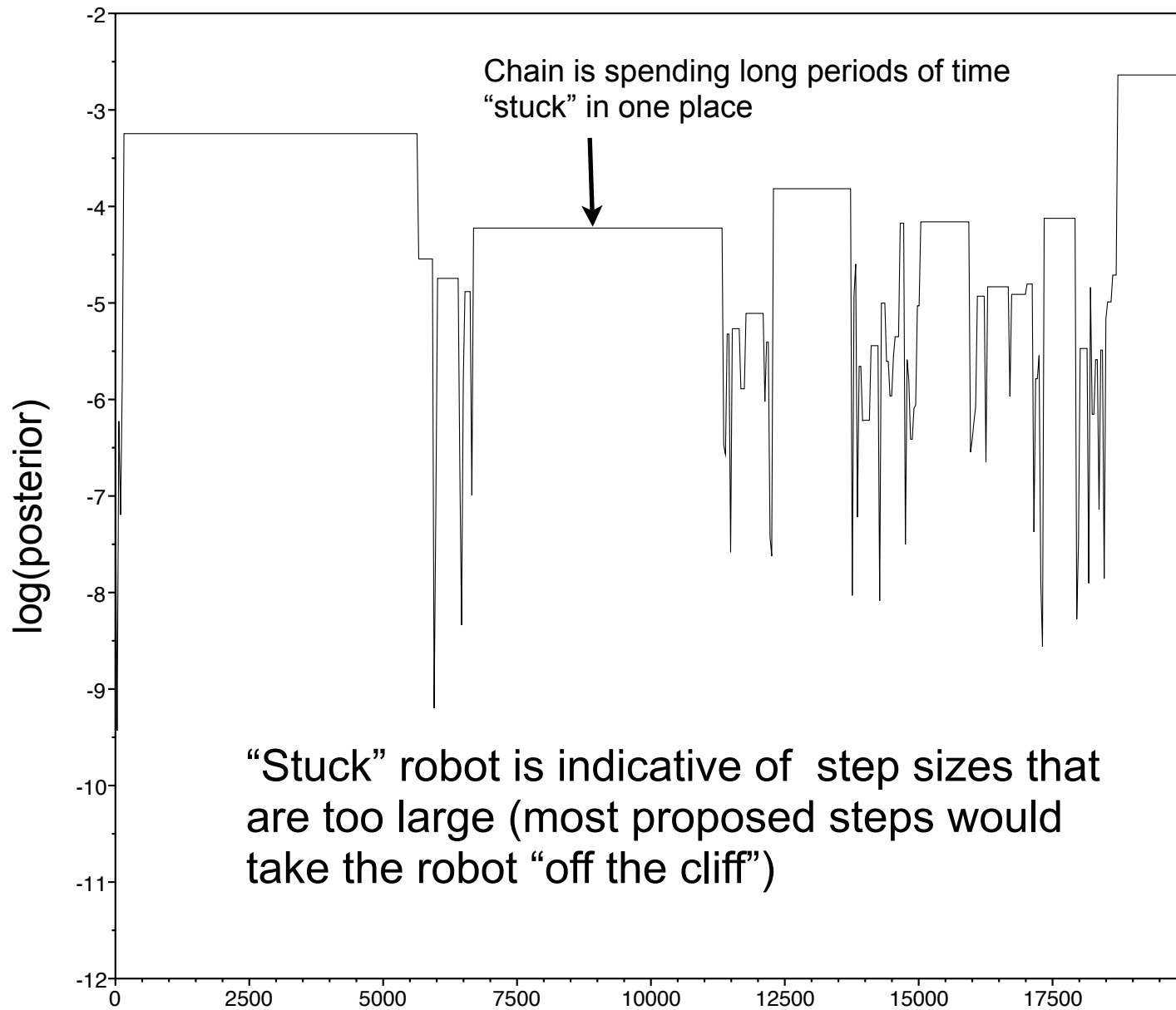
Proposal distributions with larger variance...

**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

Windows program download from:  
<http://www.eeb.uconn.edu/people/plewis/software.php>

# Tradeoff

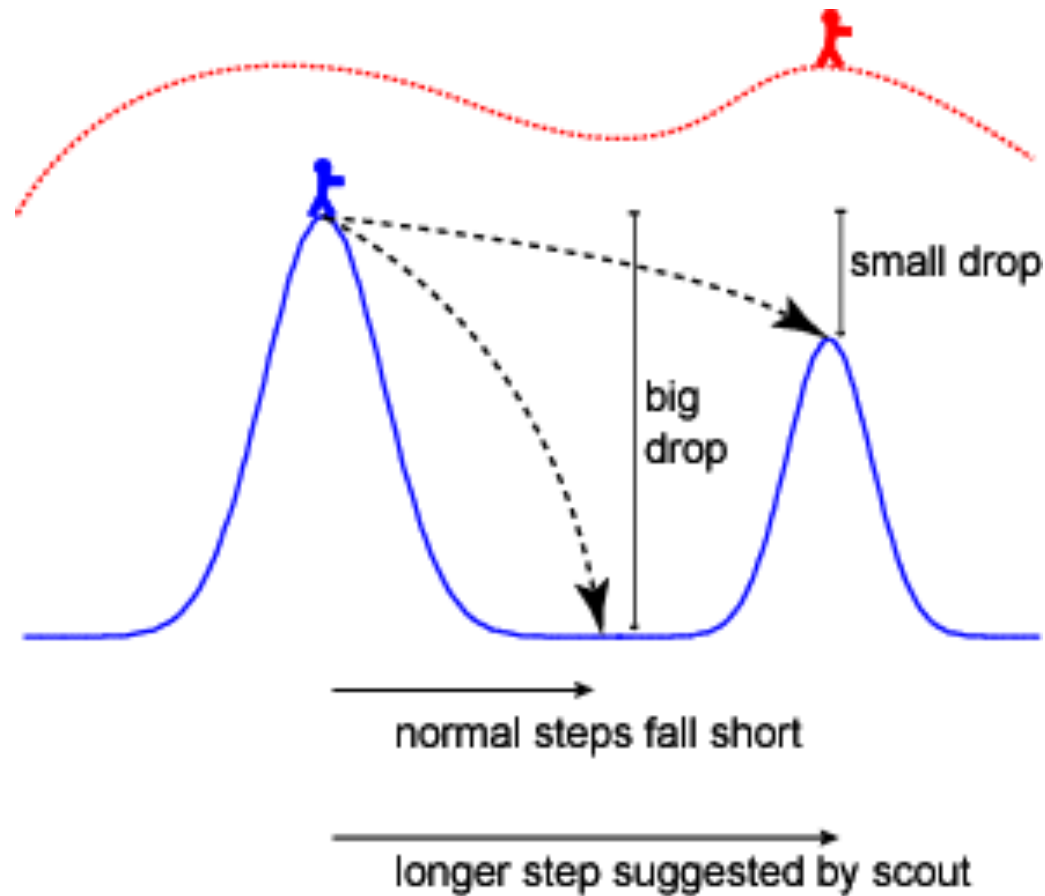
- Taking **big steps** helps in jumping from one “island” in the posterior density to another
- Taking **small steps** often results in better mixing
- How can we overcome this tradeoff? **MCMCMC**

# Metropolis-coupled Markov chain Monte Carlo (MCMCMC)

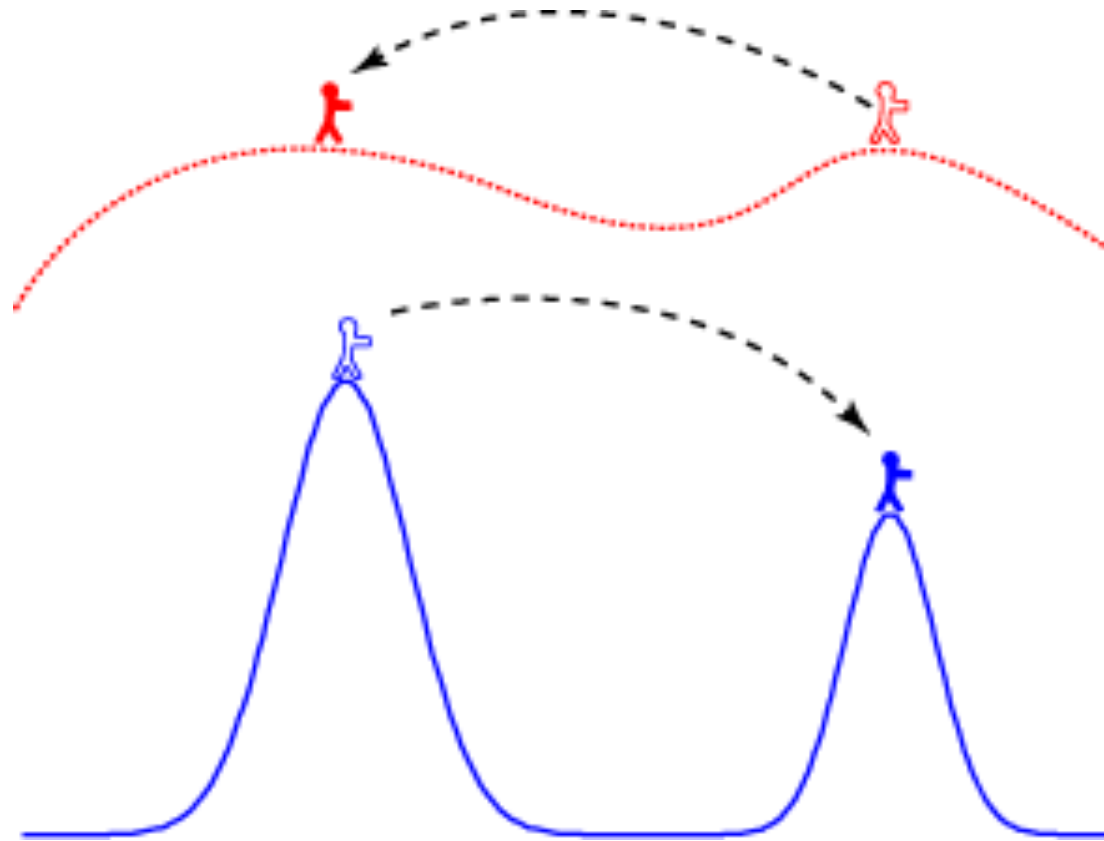
- MCMCMC involves running **several chains simultaneously**
- The **cold chain** is the one that counts, the rest are **heated chains**
- Chain is heated by raising densities to a power less than 1.0 (values closer to 0.0 are warmer)

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



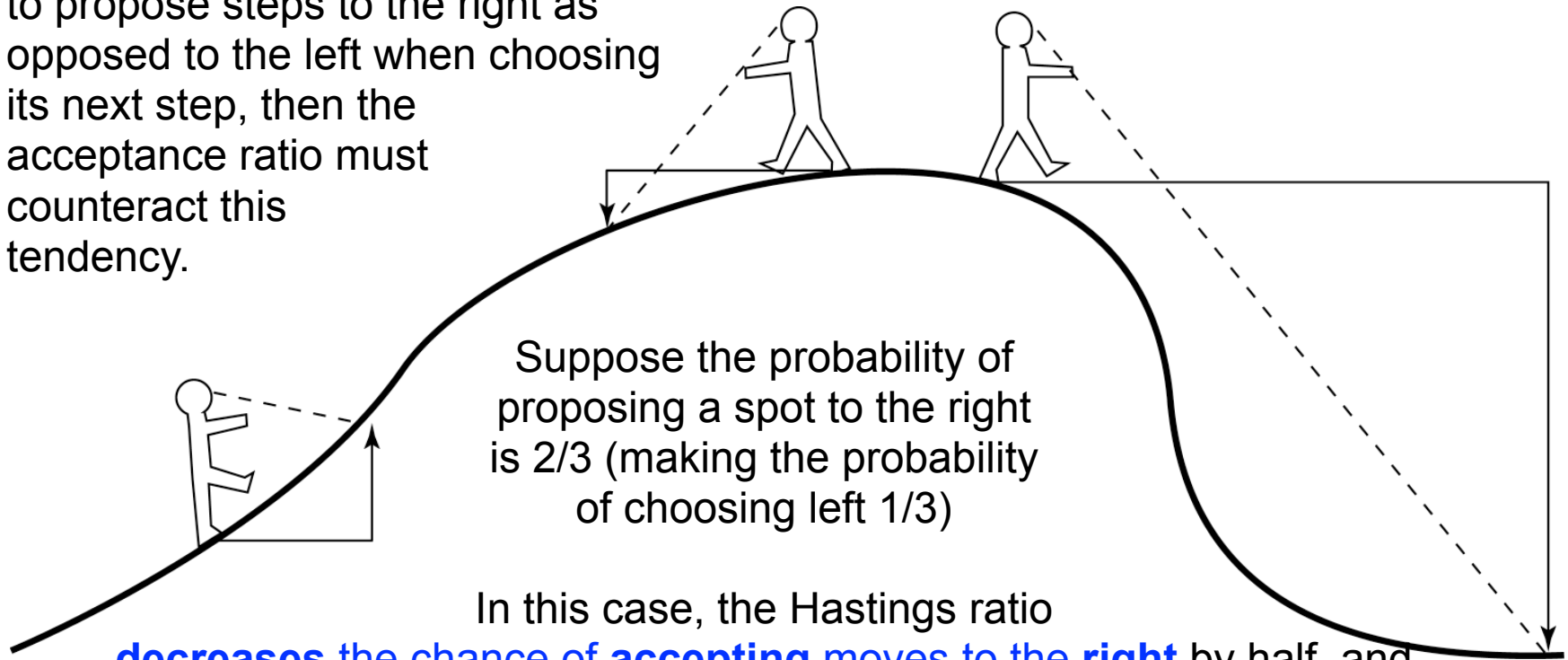
# Cold and hot chains swapped



# 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  $2/3$  (making the probability of choosing left  $1/3$ )

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.



# 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

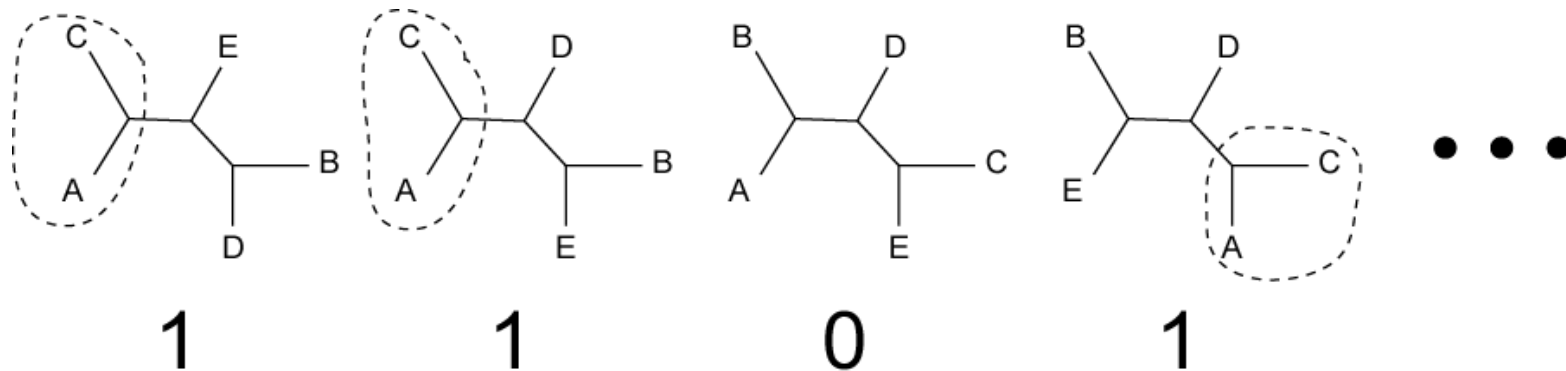
Posterior ratio

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

Step 1: select 3 contiguous branch segments (bolded)

Step 2: shrink or expand selected segment by a random amount

$$m^* = m e^{\lambda(u - 1/2)}$$

Step 3: select one of 2 groups attached to selected segment at random and prune (group X selected here)

Step 4: reattach pruned group to selected segment at a random point (this will change topology of tree if reattachment occurs in this region)

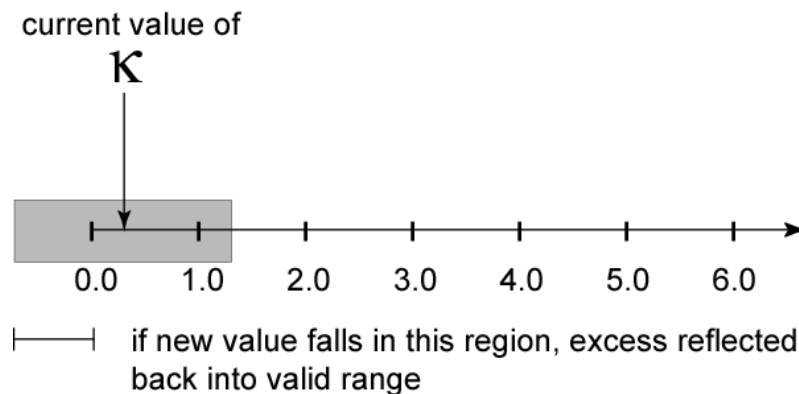
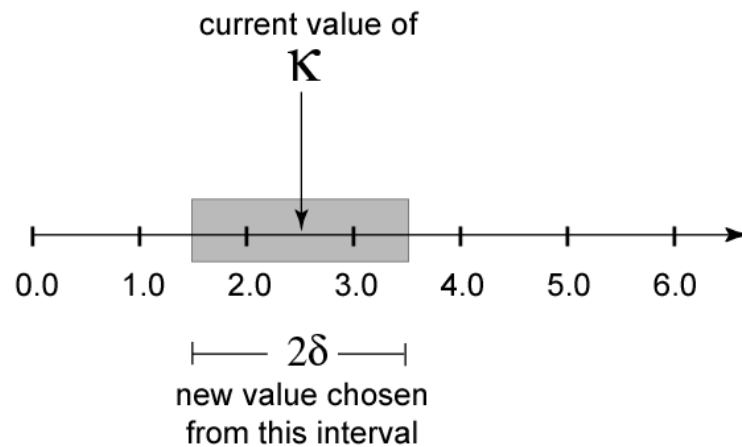
## The Larget-Simon\* move

\*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: 961-965.

This shows the tree after the proposed move has been accepted. The selected segment has been shortened, and group X ended up on a different segment, thus changing the topology

# Moving through parameter space



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

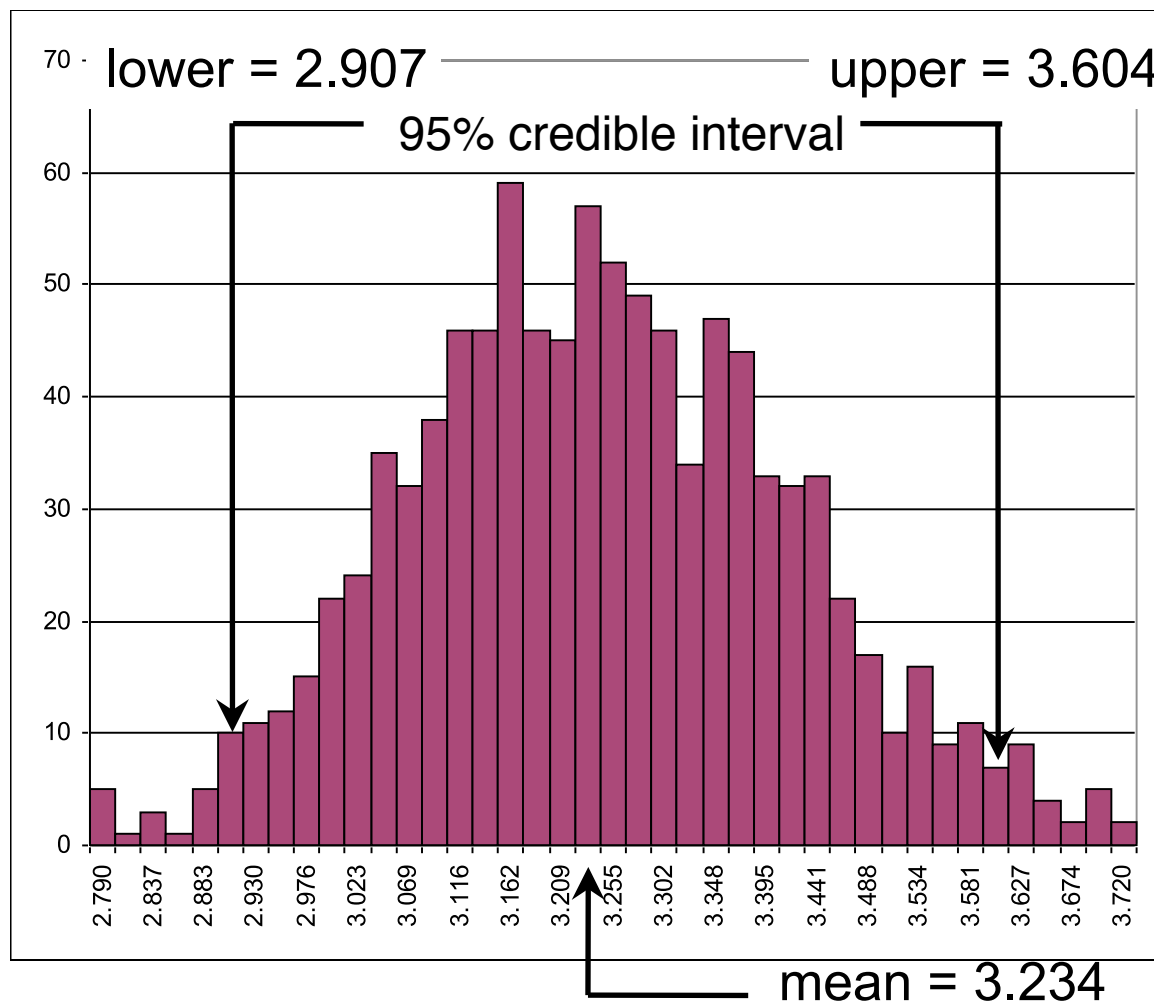
Proposal distribution is the uniform distribution on the interval  $(\kappa-d, \kappa+d)$

The “step size” of the MCMC robot is defined by  $d$ : a larger  $d$  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. Larget-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 $\kappa$



Histogram created from a sample of 1000 kappa values.

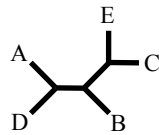
# IV. Prior distributions



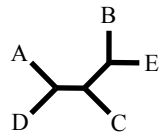
# Common Priors

- **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

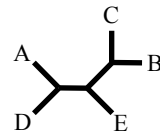
# Discrete Uniform distribution for topologies



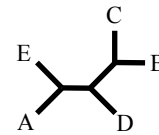
$$\frac{1}{15}$$



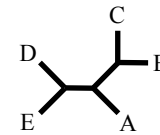
$$\frac{1}{15}$$



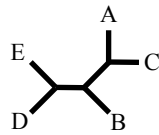
$$\frac{1}{15}$$



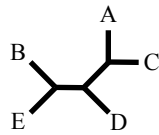
$$\frac{1}{15}$$



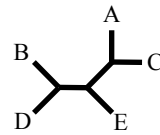
$$\frac{1}{15}$$



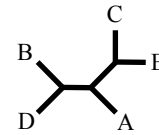
$$\frac{1}{15}$$



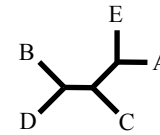
$$\frac{1}{15}$$



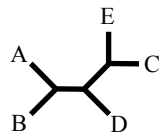
$$\frac{1}{15}$$



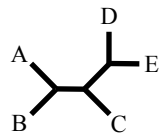
$$\frac{1}{15}$$



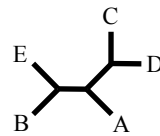
$$\frac{1}{15}$$



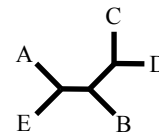
$$\frac{1}{15}$$



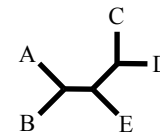
$$\frac{1}{15}$$



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

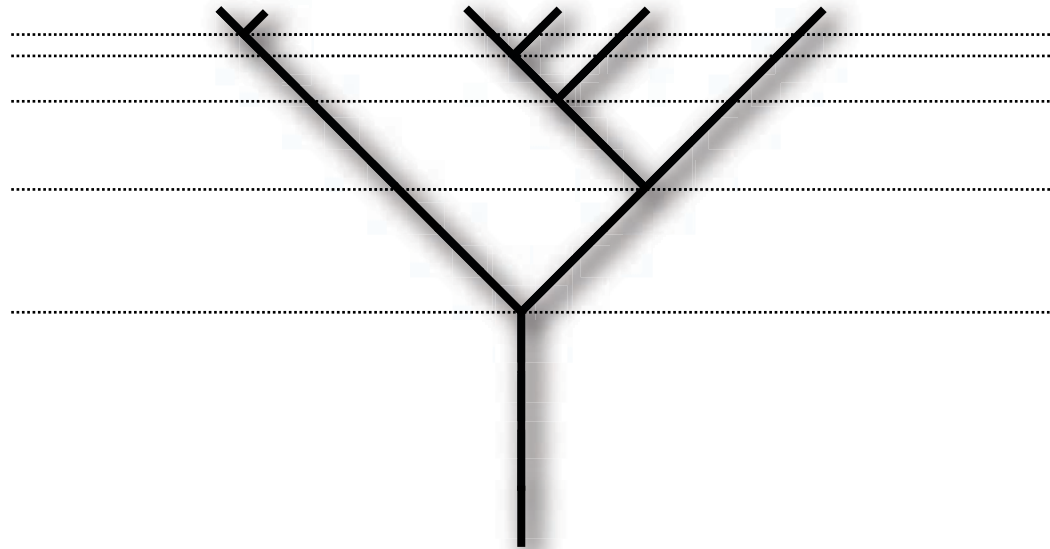


$$\frac{1}{15}$$



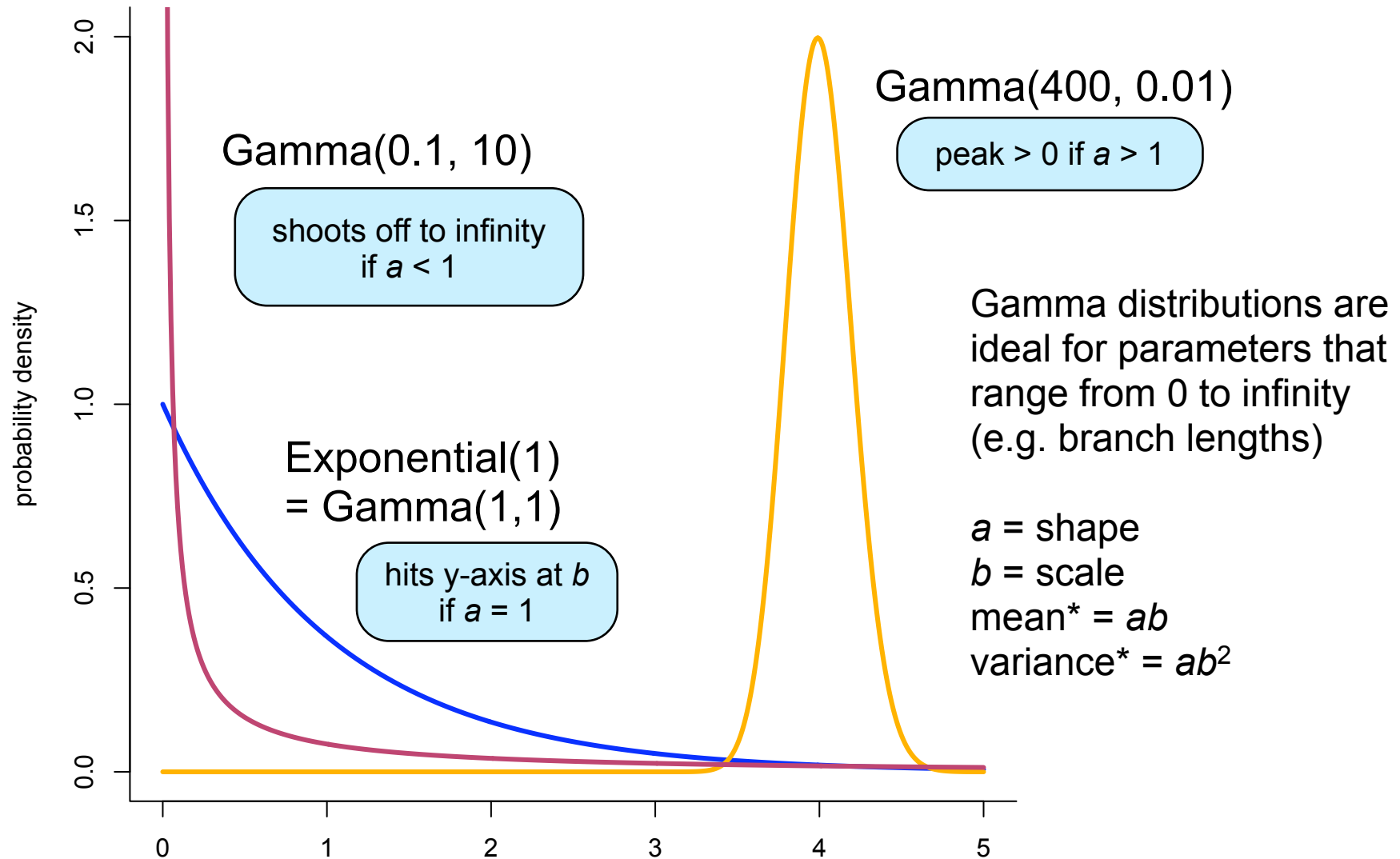
$$\frac{1}{15}$$

# Yule model provides joint prior for both **topology** and **branch lengths**



The rate of speciation under the Yule model ( $\lambda$ ) 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.

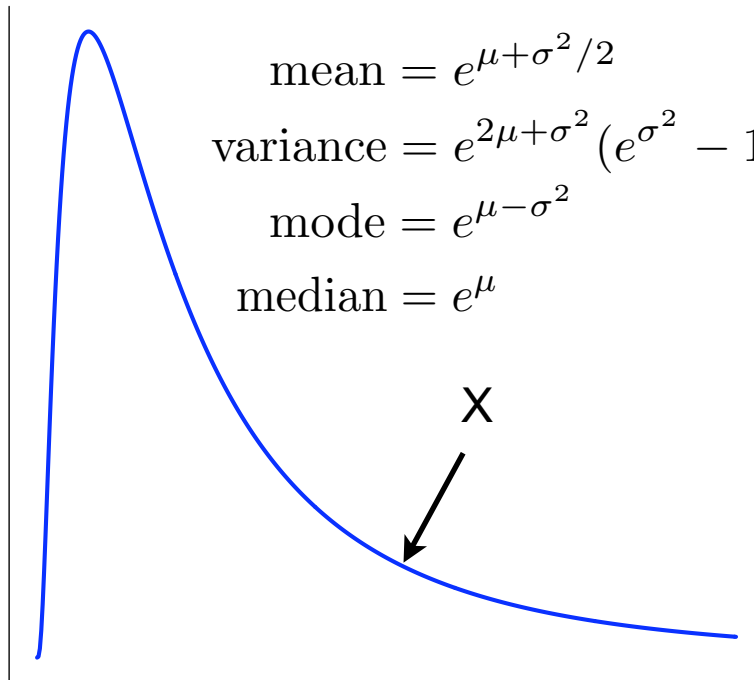
# Gamma( $a, b$ ) distributions



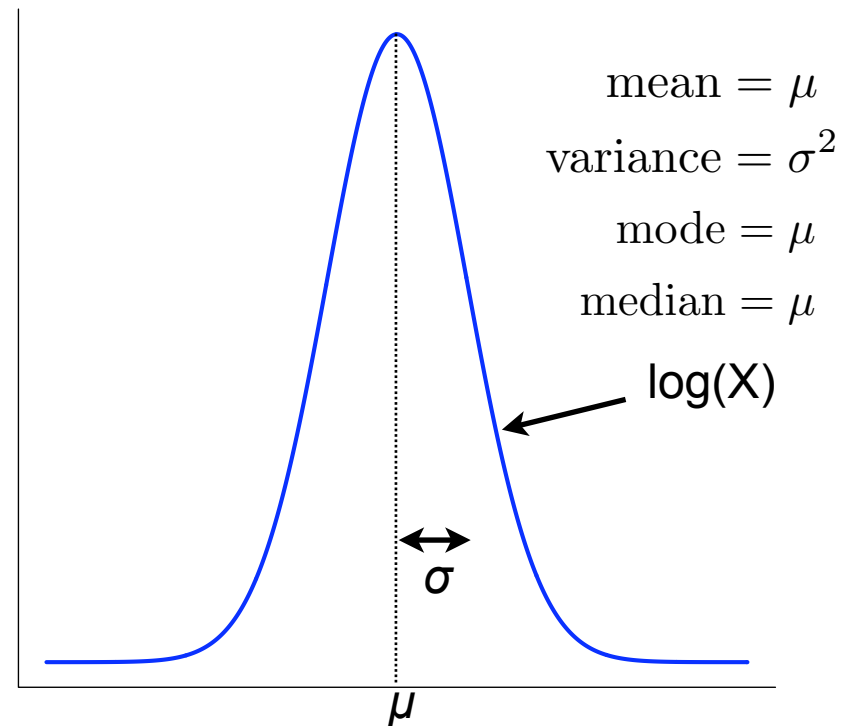
\*Note: be aware that in many papers the Gamma distribution is defined such that the second (scale) parameter is the *inverse* of the value  $b$  used in this slide! In this case, the mean and variance would be  $a/b$  and  $a/b^2$ , respectively.

# Log-normal distribution

If  $X$  is **log-normal** with parameters  $\mu$  and  $\sigma$ ...



...then **log(X)** is **normal** with mean  $\mu$  and standard deviation  $\sigma$ .



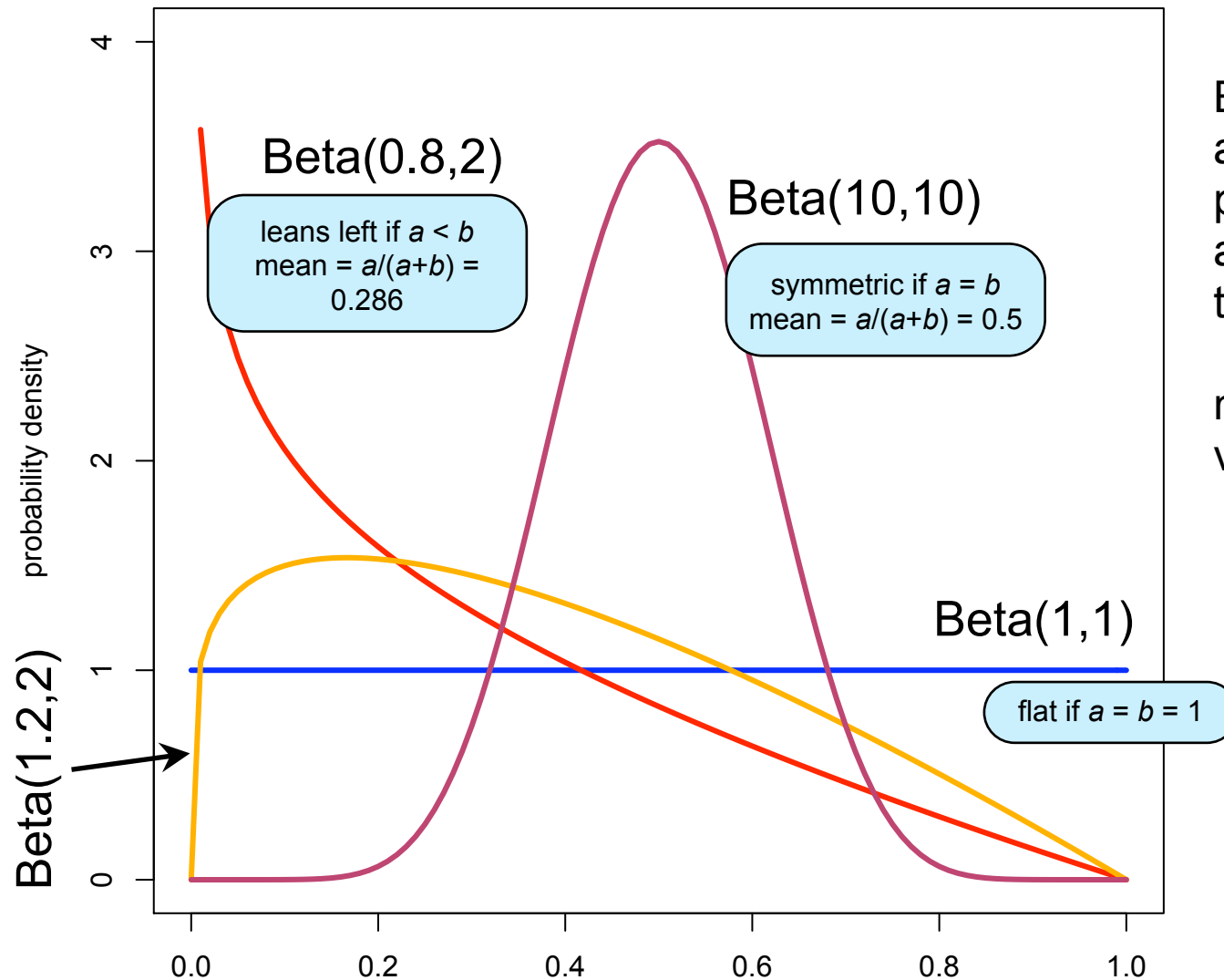
**Important:**  $\mu$  and  $\sigma$  do **not** represent the mean and variance of  $X$ : they are the mean and variance of  $\log(X)$ !

To choose  $\mu$  and  $\sigma$  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



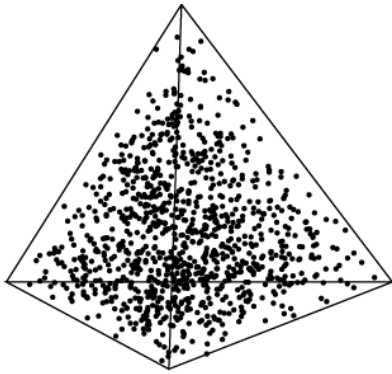
Beta distributions are appropriate for proportions, which are constrained to the interval  $[0, 1]$ .

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

# Dirichlet( $a, b, c, d$ ) distribution

Used for nucleotide relative frequencies:

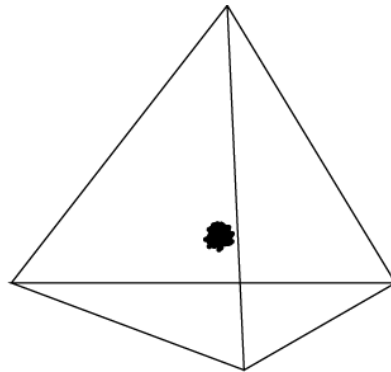
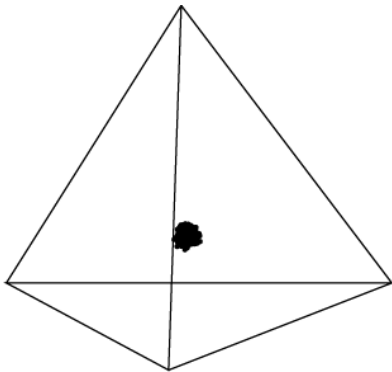
$$a \rightarrow \pi_A, b \rightarrow \pi_C, c \rightarrow \pi_G, d \rightarrow \pi_T$$



Flat prior:

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

(no scenario discouraged)



Informative prior:


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

(equal frequencies strongly encouraged)

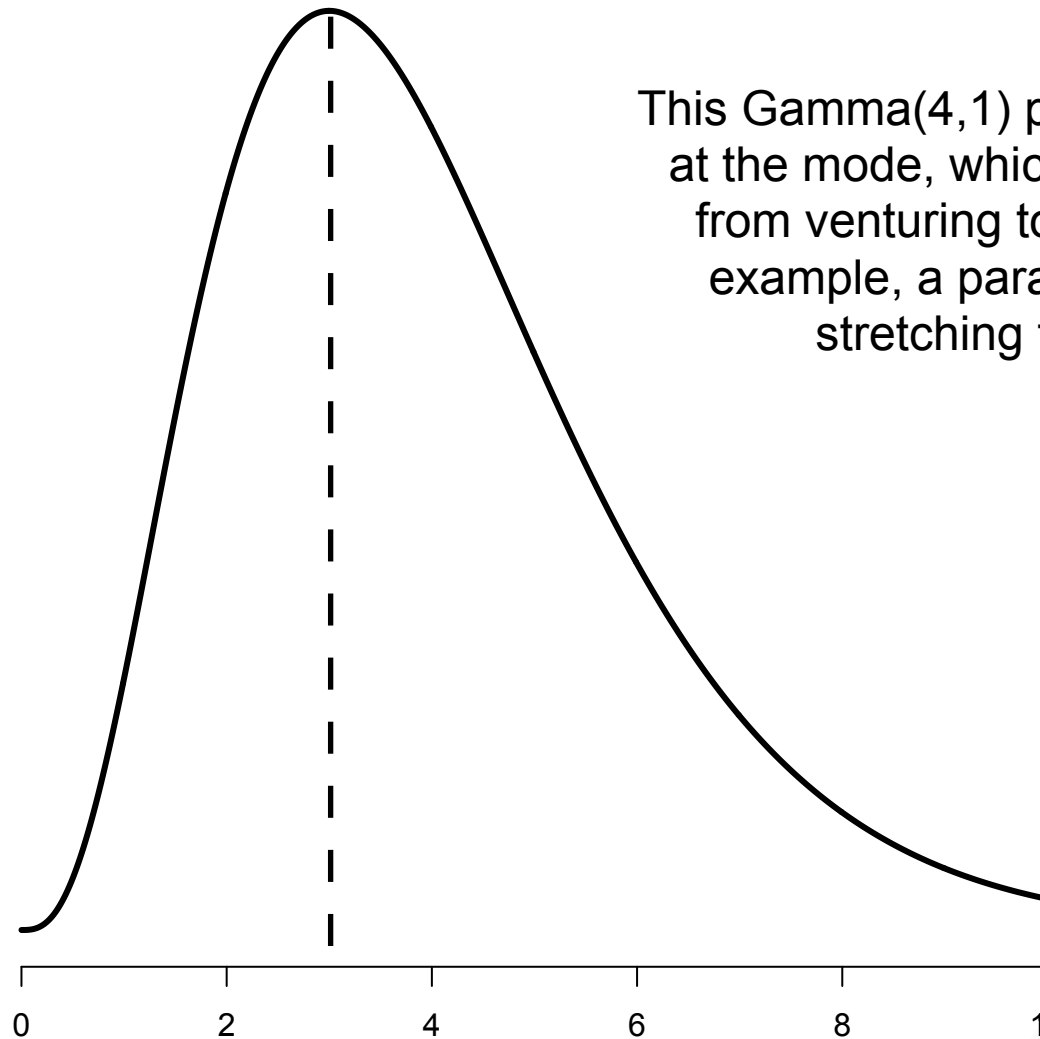
(stereo pairs)

Dirichlet( $a, b, c, d, e, f$ ) used for  
GTR relative rates

# Prior Miscellany

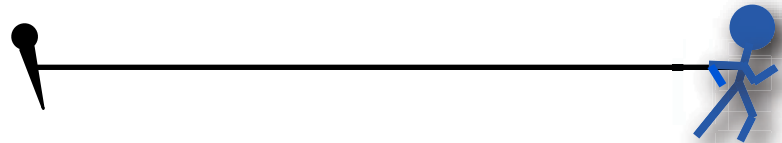
- priors as rubber bands 
- running on empty
- hierarchical models
- empirical bayes
- dirichlet process priors

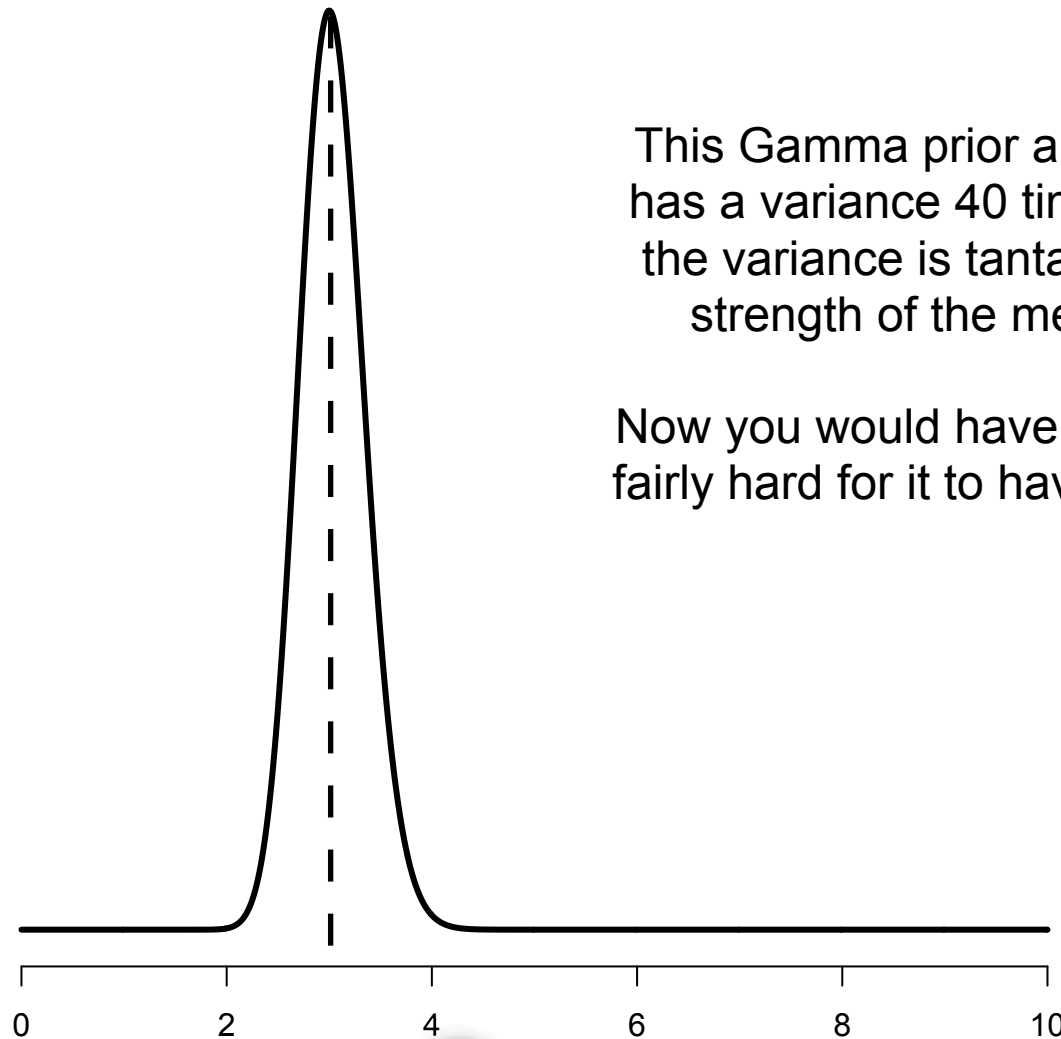




This Gamma(4,1) prior ties down its parameter at the mode, which is at 3, and discourages it from venturing too far in either direction. For example, a parameter value of 10 would be stretching the rubber band fairly tightly

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



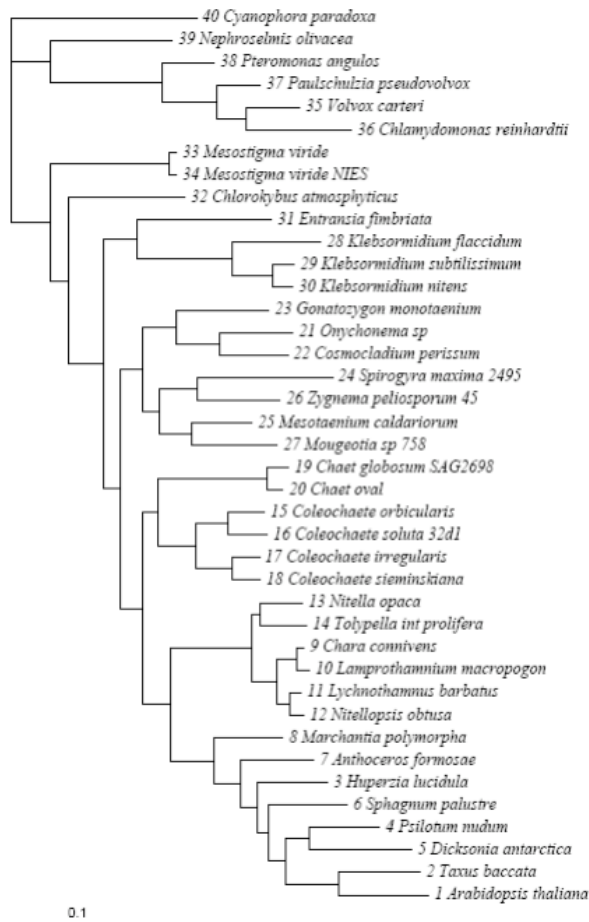


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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

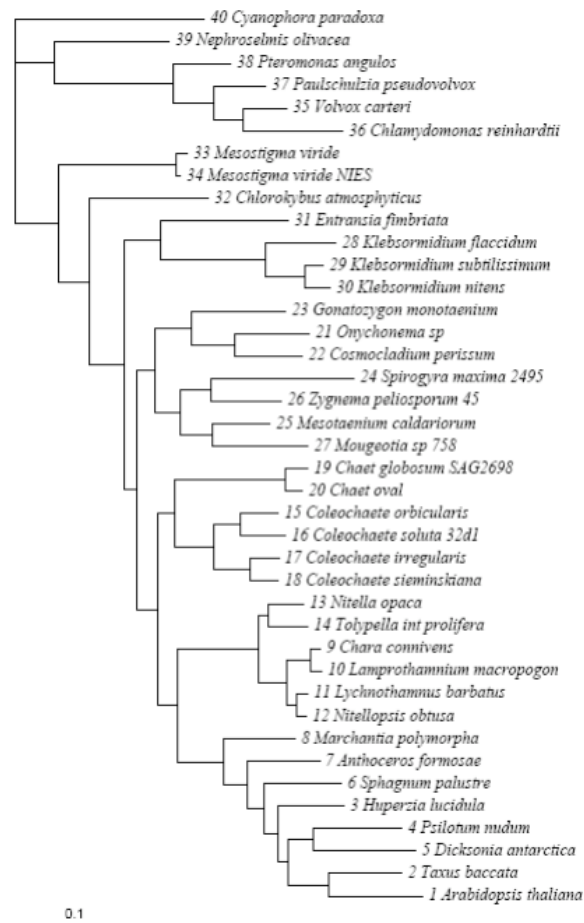
This gamma distribution has shape 91.989 and scale 0.032971

# Example: Internal Branch Length Priors

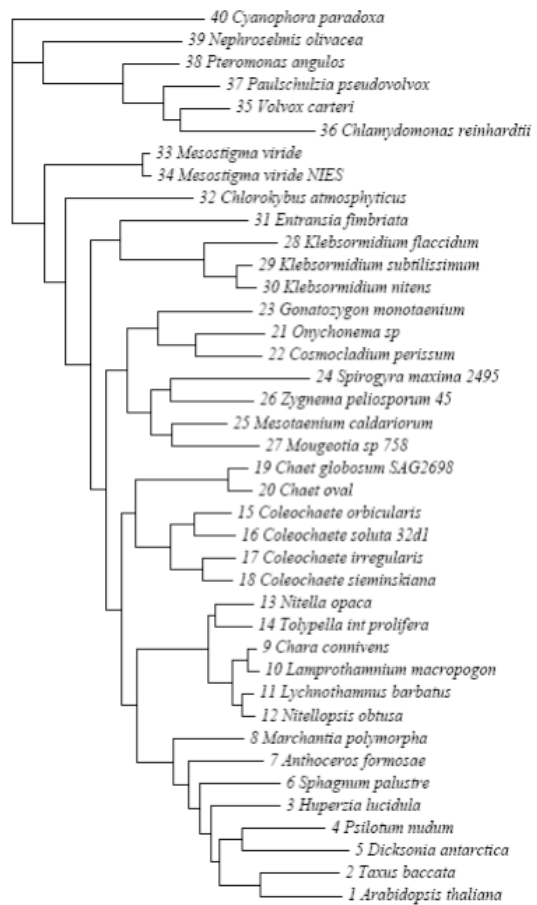


Internal branch length prior is exponential with mean 0.1

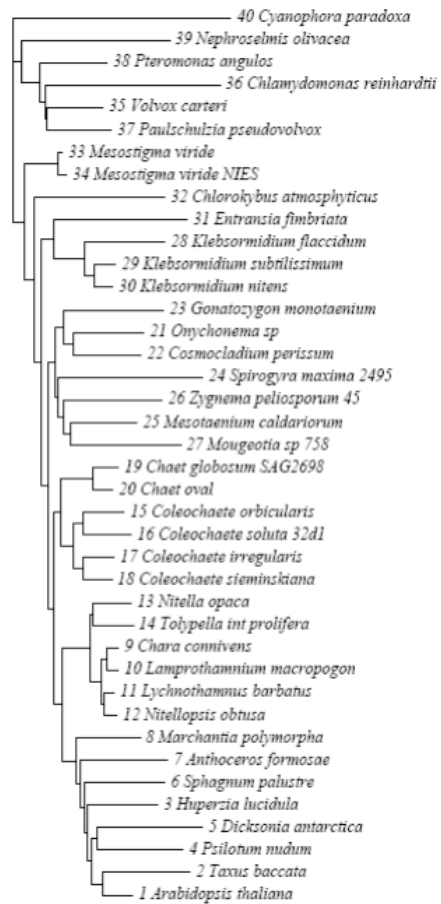
This is a reasonably vague internal branch length prior



Internal branch length prior mean 0.01

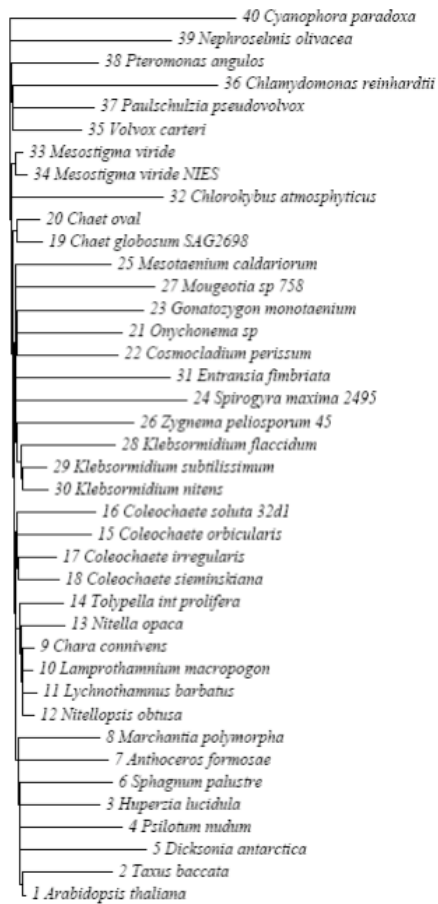


Internal branch length prior mean 0.001



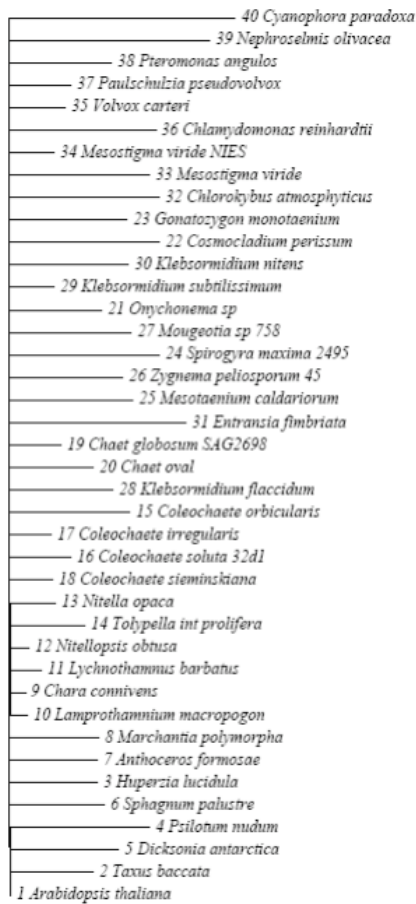
Internal branch length prior mean 0.0001

0.1



Internal branch length prior mean 0.00001

0.1



0.1

Internal branch length prior mean 0.000001

The internal branch length prior is calling the shots now.

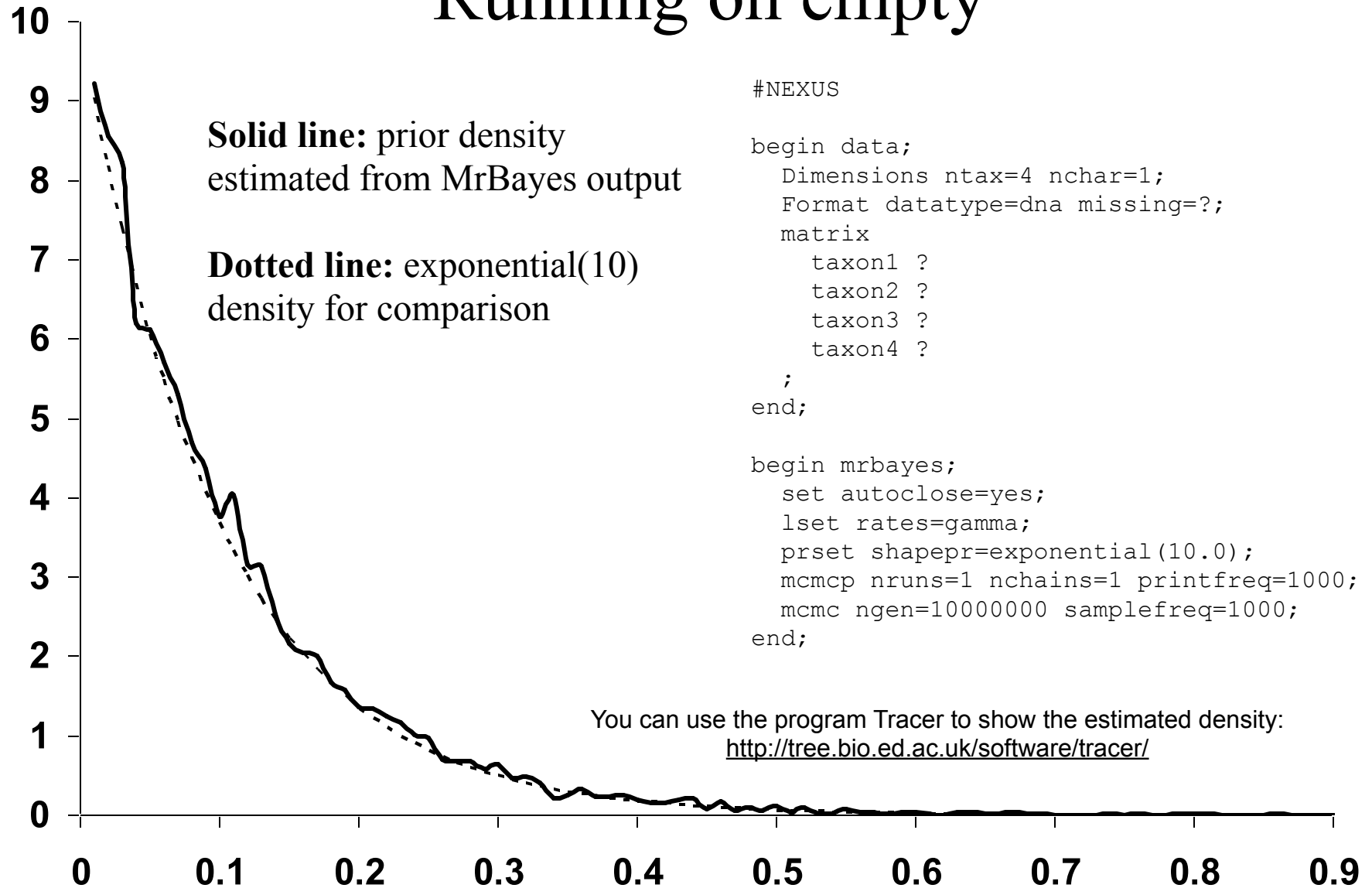


# Prior Miscellany

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



# Running on empty



# Prior Miscellany

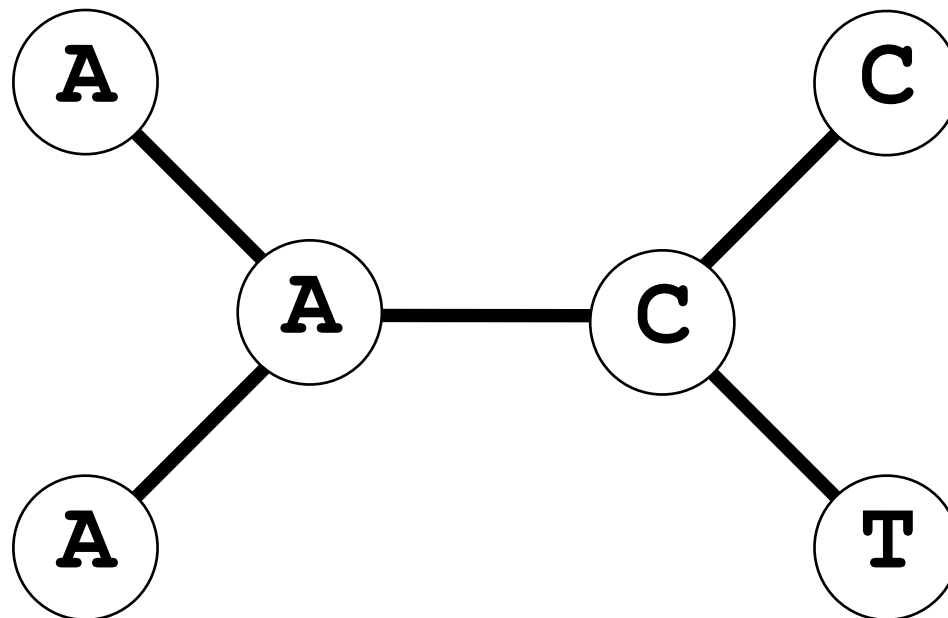
- priors as rubber bands
- running on empty
- hierarchical models
- empirical bayes
- dirichlet process priors



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

$\mu$  is a *hyperparameter* governing the mean of the edge length prior

*hyperprior*

Prior: Exponential, mean  $\mu$

The diagram illustrates a hierarchical model structure. At the top, the word "hyperprior" is written. A downward arrow points from "hyperprior" to the text "Prior: Exponential, mean  $\mu$ ". From this text, a horizontal line extends to the left and right, with five downward arrows pointing to five separate terms in a likelihood function  $L_k$ . Each term is a bracketed expression of the form  $\left[ \frac{1}{4} + \frac{3}{4} e^{-4v_i/3} \right]$  for  $i=1, 2, 3, 4, 5$ .

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

During an MCMC analysis,  $\mu$  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
- dirichlet process priors



# Empirical Bayes

Empirical Bayes uses the data to determine some aspects of the prior, such as the prior mean. This uses the data twice, which is not acceptable to Bayesian purists


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



Prior: Exponential, mean=MLE

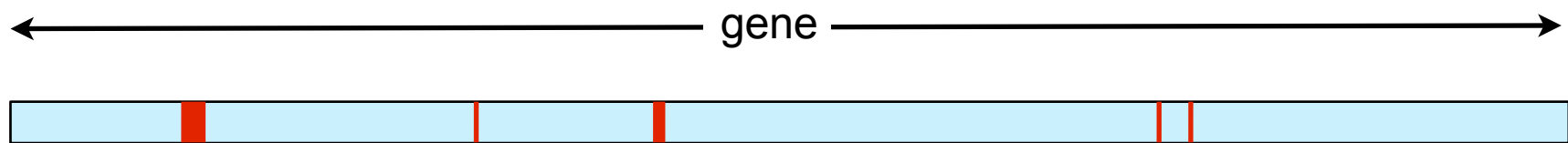
$$L_k = \frac{1}{4} \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_1/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_2/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4v_3/3} \right] \left[ \frac{1}{4} - \frac{1}{4} e^{-4v_4/3} \right] \left[ \frac{1}{4} + \frac{3}{4} e^{-4v_5/3} \right]$$

# Prior Miscellany

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



# The problem that DP models help solve

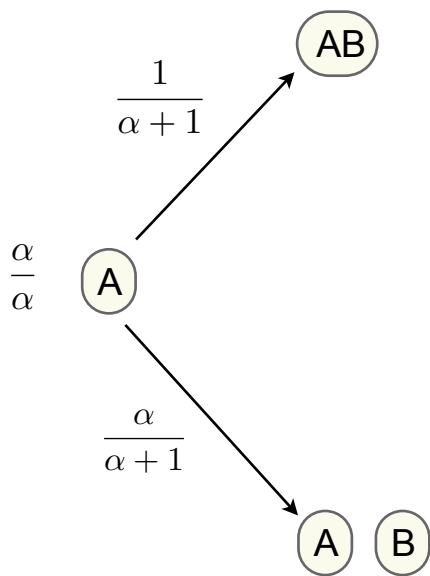


**Red** depicts sites with, for example:

- 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 model feature

Desired: a model that:

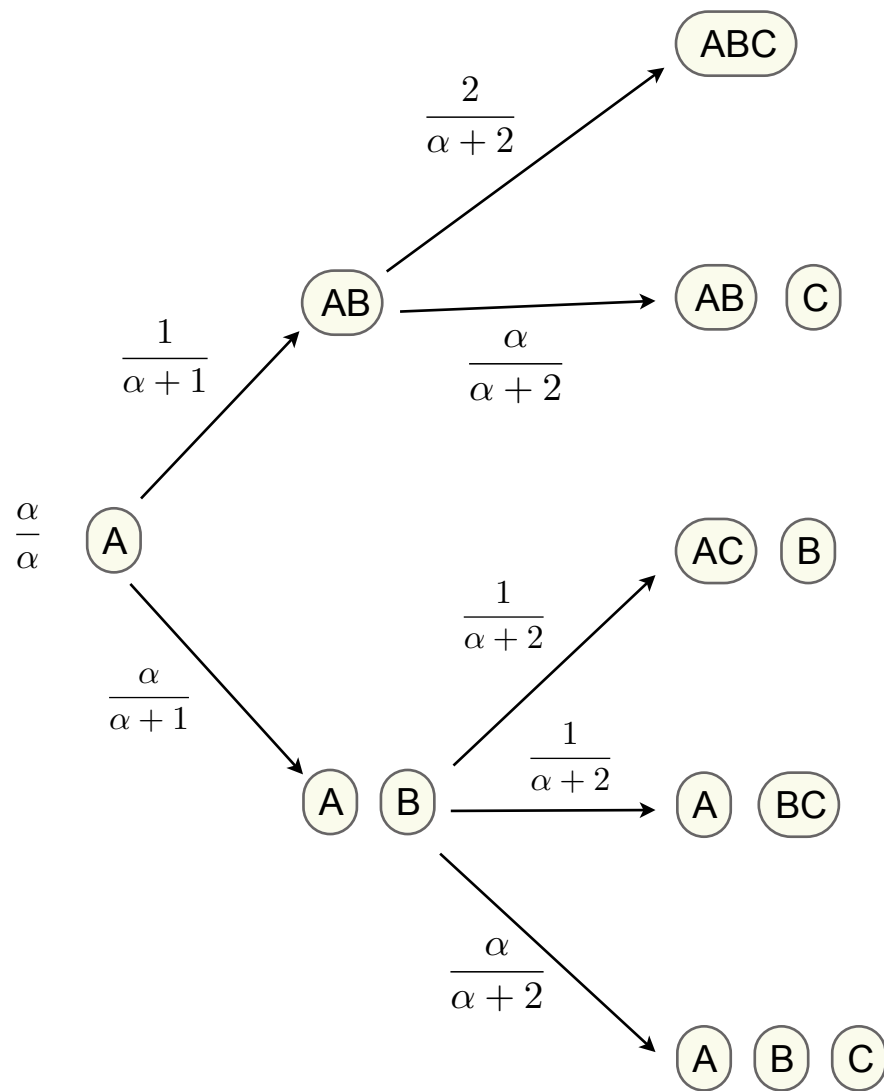
- classifies sites into meaningful categories
- discourages large numbers of categories (with the strength of discouragement determined by some value  $\alpha$ )
- assigns reasonable parameter values to each of the categories
- does all this automatically



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

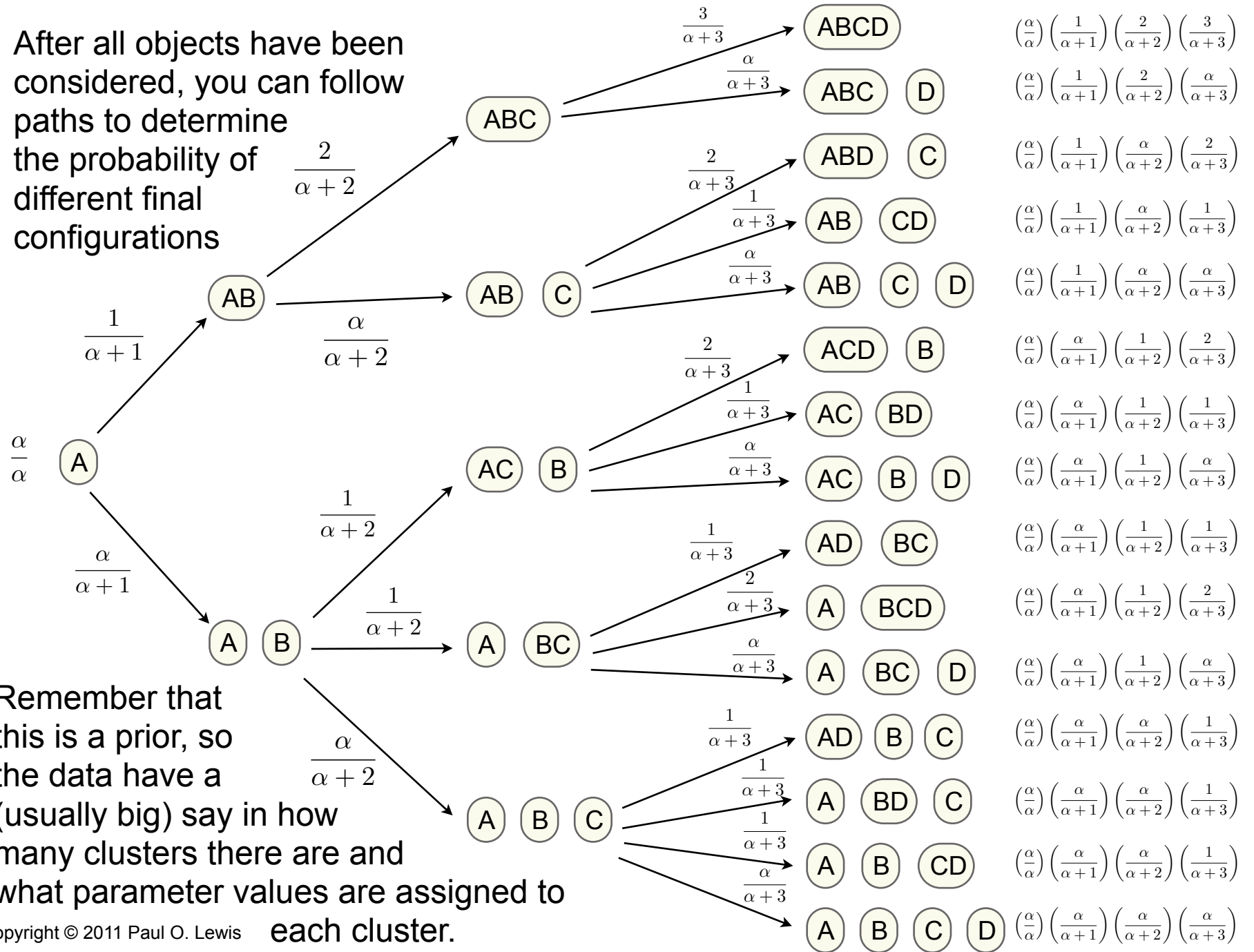
The parameter  $\alpha$  determines the propensity for forming a new group



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

...or form its own group

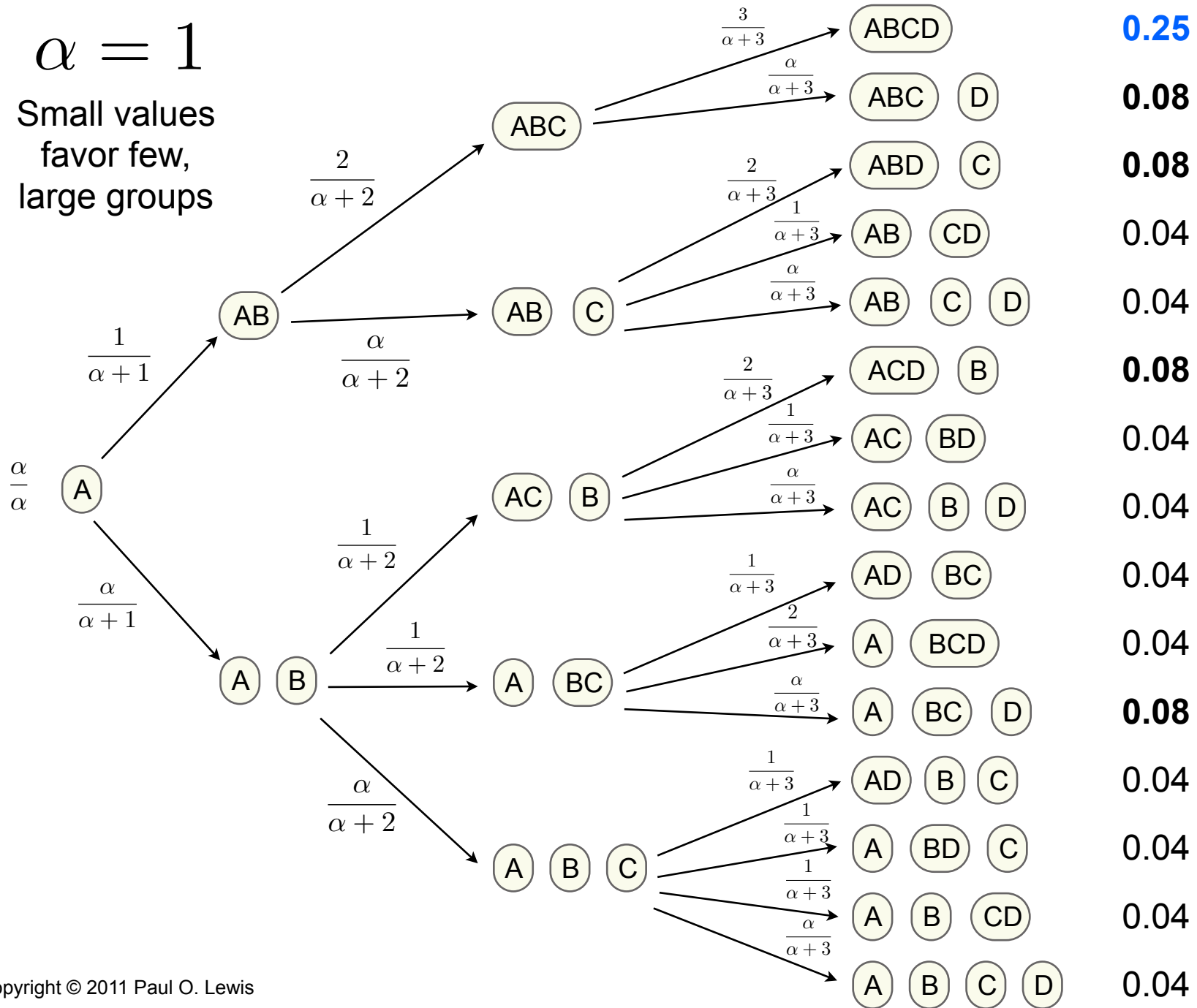
After all objects have been considered, you can follow paths to determine the probability of different final configurations



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 cluster.

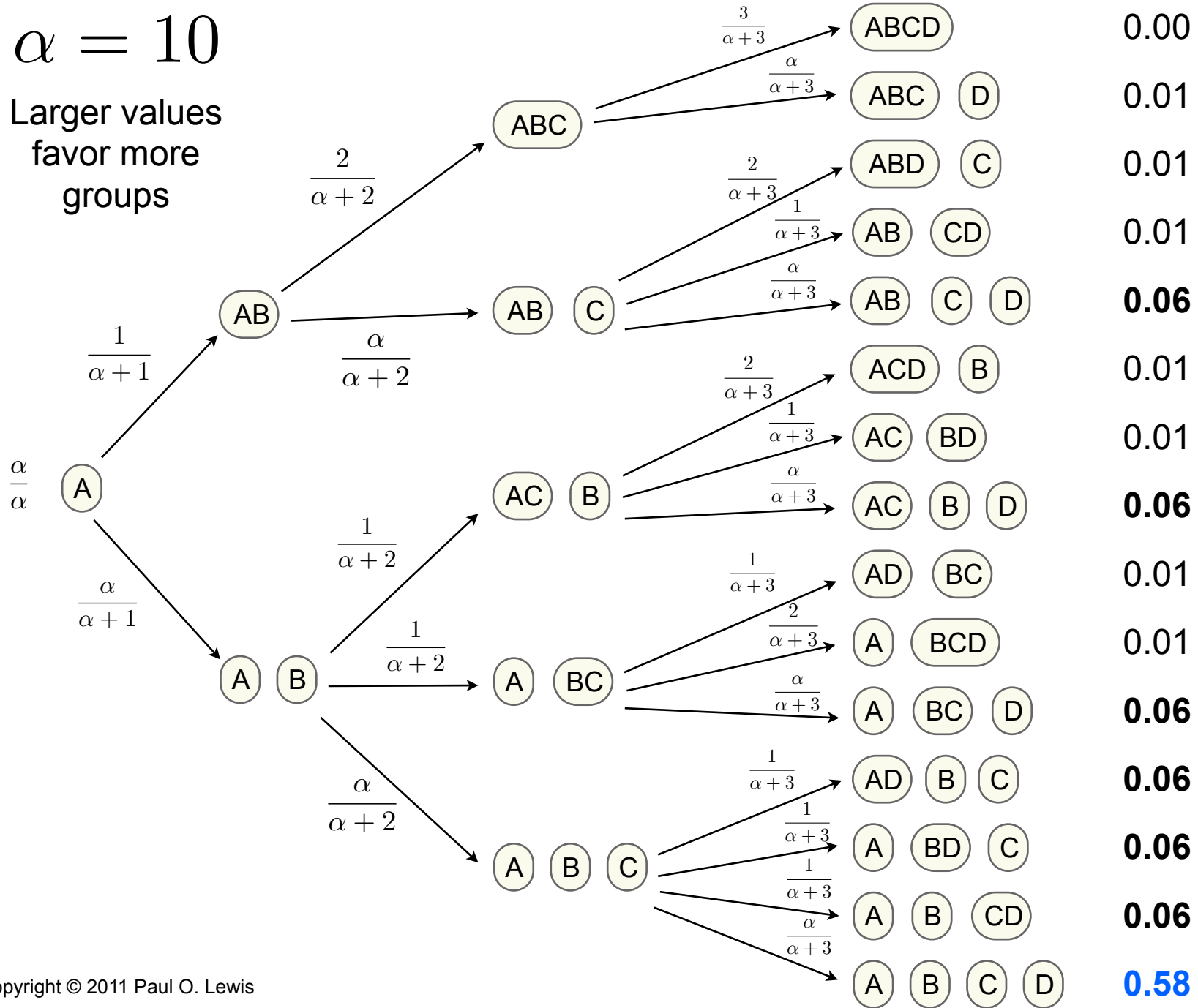
$$\alpha = 1$$

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 used (i.e. alpha is a hyperparameter that can be 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

# Where to find DP models

## **A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process**

*Nicolas Lartillot and Hervé Philippe*

Molecular Biology and Evolution (2004) vol. 21 (6) pp. 1095-1109

PhyloBayes

<http://www.atgc-montpellier.fr/phylobayes>

## **A Dirichlet process model for detecting positive selection in protein-coding DNA sequences**

John P. Huelsenbeck<sup>\*†</sup>, Sonia Jain<sup>‡</sup>, Simon W. D. Frost<sup>§</sup>, and Sergei L. Kosakovsky Pond<sup>§</sup>

Proceedings of the National Academy of Sciences (2006) vol. 103 (16) pp. 6263-6268

BUCKy

<http://www.stat.wisc.edu/~ane/bucky/>

## **Bayesian Estimation of Concordance among Gene Trees**

*Cécile Ané,<sup>\*†</sup> Bret Larget,<sup>\*†</sup> David A. Baum,<sup>†</sup> Stacey D. Smith,<sup>‡</sup> and Antonis Rokas<sup>§</sup>*

Molecular Biology and Evolution (2007) vol. 24 (2) pp. 412-426

## **A Nonparametric Method for Accommodating and Testing Across-Site Rate Variation**

JOHN P. HUELSENBECK,<sup>1</sup> AND MARC A. SUCHARD<sup>2,3,4</sup>

Systematic Biology (2007) vol. 56 (6) pp. 975-987



# V. Bayesian model selection

# Marginal likelihoods of models

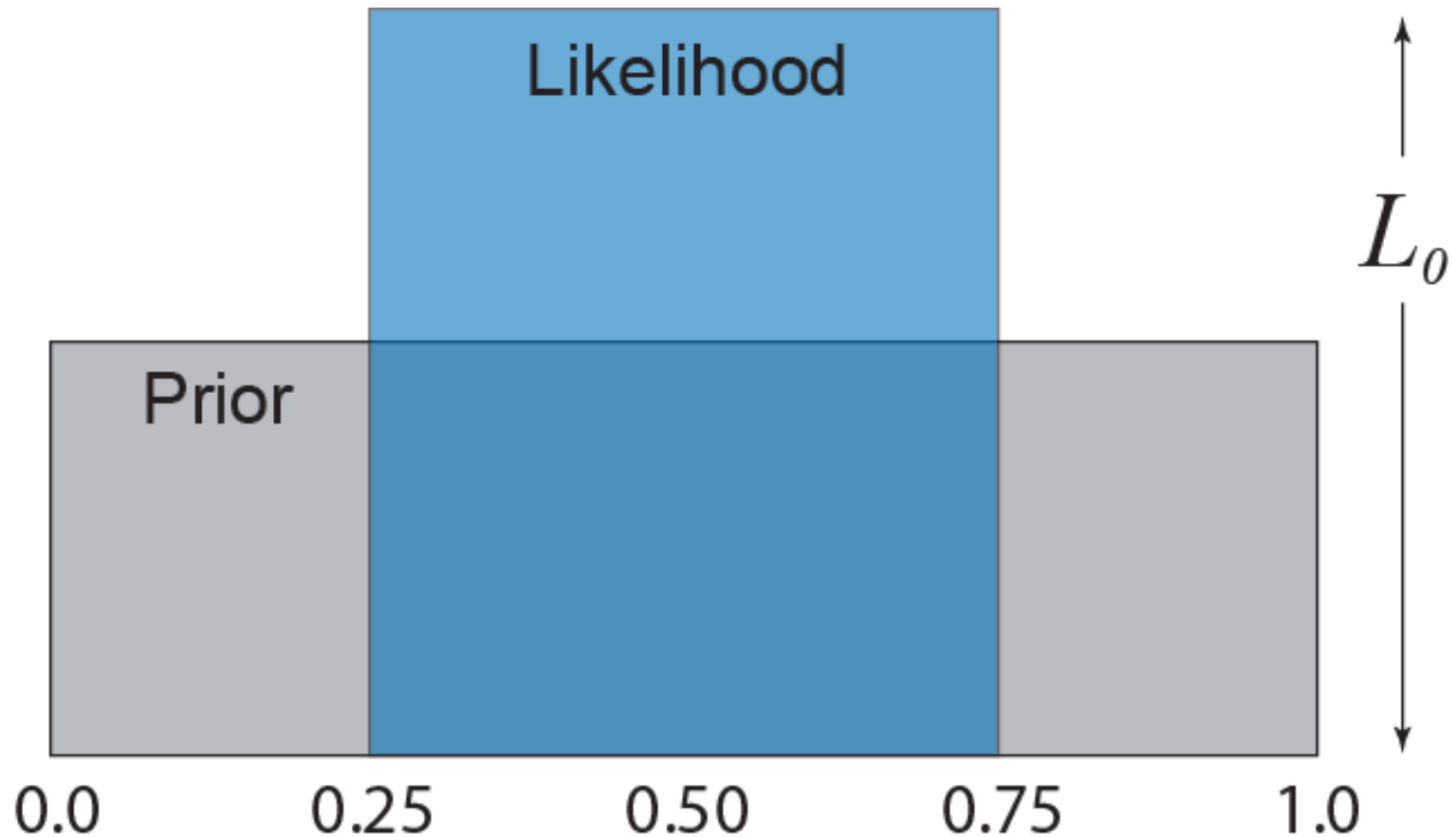
$$\Pr(D) = \int_{\theta} f(D|\theta) f(\theta) d\theta$$

Marginal probability of the data (denominator in Bayes' rule).  
This is a weighted average of the likelihood, where the weights are provided by the prior distribution.

$$\Pr(D|M) = \int_{\theta} f(D|\theta, M) f(\theta|M) d\theta$$

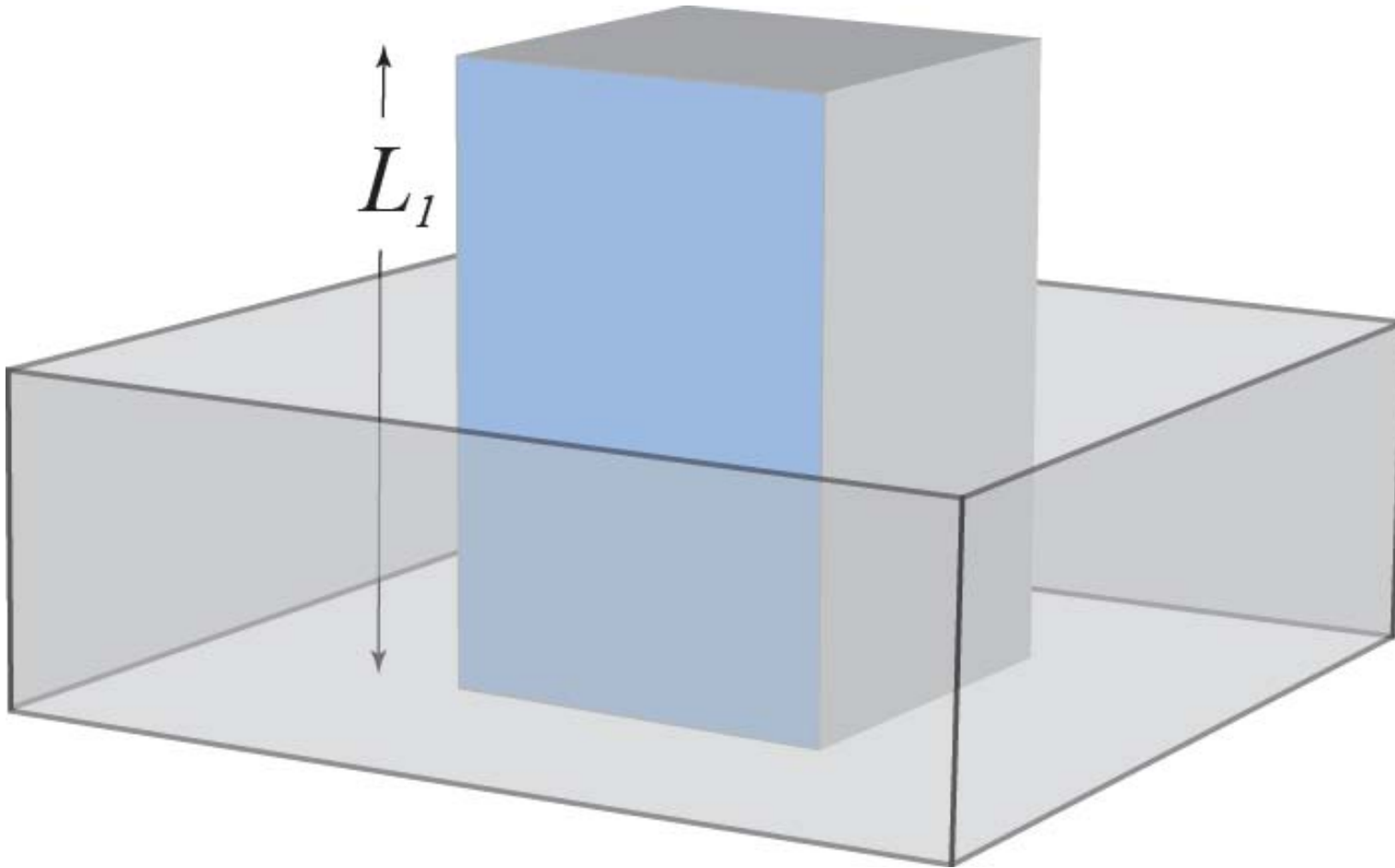
Often left out is the fact that we are also conditioning on  $M$ , the model used.  $\Pr(D|M_1)$  is comparable to  $\Pr(D|M_2)$  and thus the marginal probability of the data can be used to compare the average fit of different models as long as the data  $D$  is the same.  $\Pr(D|M)$  is also known as the **marginal likelihood** of the model  $M$ .

# Marginal likelihood (1-param. model)



$$\text{Average likelihood} = \left(\frac{1}{2}\right) L_0 + \left(\frac{1}{2}\right) (0)$$

# Marginal likelihood (2-param. model)



$$\text{Average likelihood} = \left(\frac{1}{2}\right)^2 L_1 + \left[1 - \left(\frac{1}{2}\right)^2\right] (0)$$

# The Bayes Factor is a ratio of marginal (model) likelihoods

$$\begin{array}{l} \text{1-parameter model } M_0: (1/2) L_0 \\ \text{2-parameter model } M_1: (1/4) L_1 \end{array} \quad \text{BF}_{01} = \frac{(1/2)L_0}{(1/4)L_1} = \frac{2L_0}{L_1}$$

$\text{BF}_{01}$  is the Bayes Factor in favor of model  $M_0$  against model  $M_1$ :

if  $\text{BF}_{01} > 1$ , model  $M_0$  wins

if  $\text{BF}_{01} < 1$ , model  $M_1$  wins

In this case,  $L_1$  would need to be *twice* as great as  $L_0$  in order for model  $M_1$  to win.

Notes about BF:

- automatically penalizes model for extra dimensions (parameters)
- severity of penalty depends on priors (under control of investigator, unlike AIC, BIC, LRT, etc., which assess a constant penalty for each additional parameter)

Recent work on Bayes factors with respect to phylogenetics:  
Huelsenbeck, Larget & Alfaro. 2004. MBE 21(6):1123-1133.  
Lartillot & Phillippe. 2005. Syst. Biol. 55(2):195-207.  
Fan, Wu, Chen, Guo & Lewis. 2011. MBE 28(1):523-532

# Something closer to reality

- Example:
  - Compare JC69 vs. K80 models
  - Parameters:
    - $\nu$  is edge length (expected no. substitutions/site)
      - free in both JC69 and K80 models
    - $\kappa$  is transition/transversion rate ratio
      - free in K80, set to 1.0 in JC69

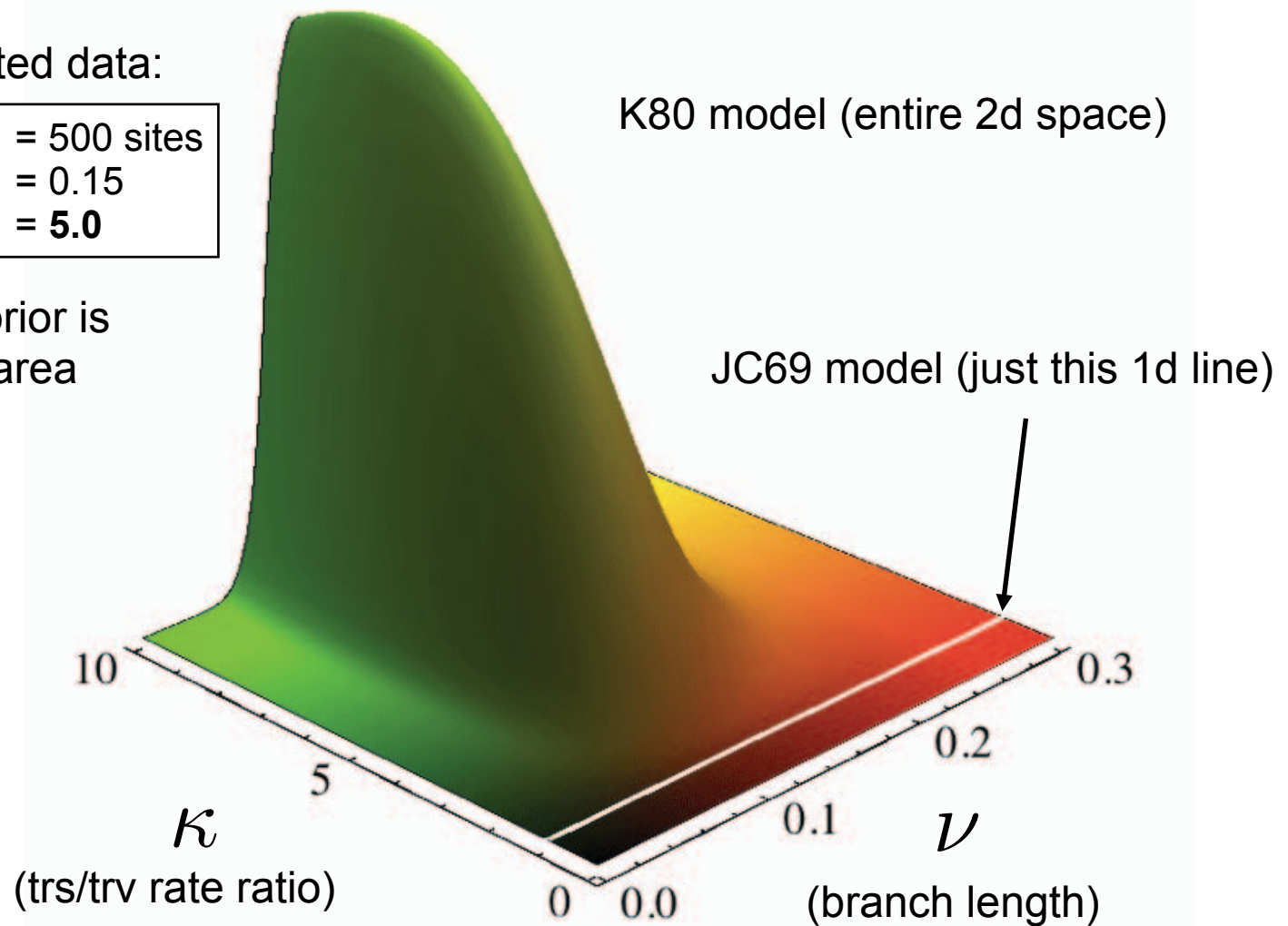


# Likelihood Surface when K80 true

Based on simulated data:

sequence length	= 500 sites
true branch length	= 0.15
true kappa	= <b>5.0</b>

Assume joint prior is flat over the area shown.



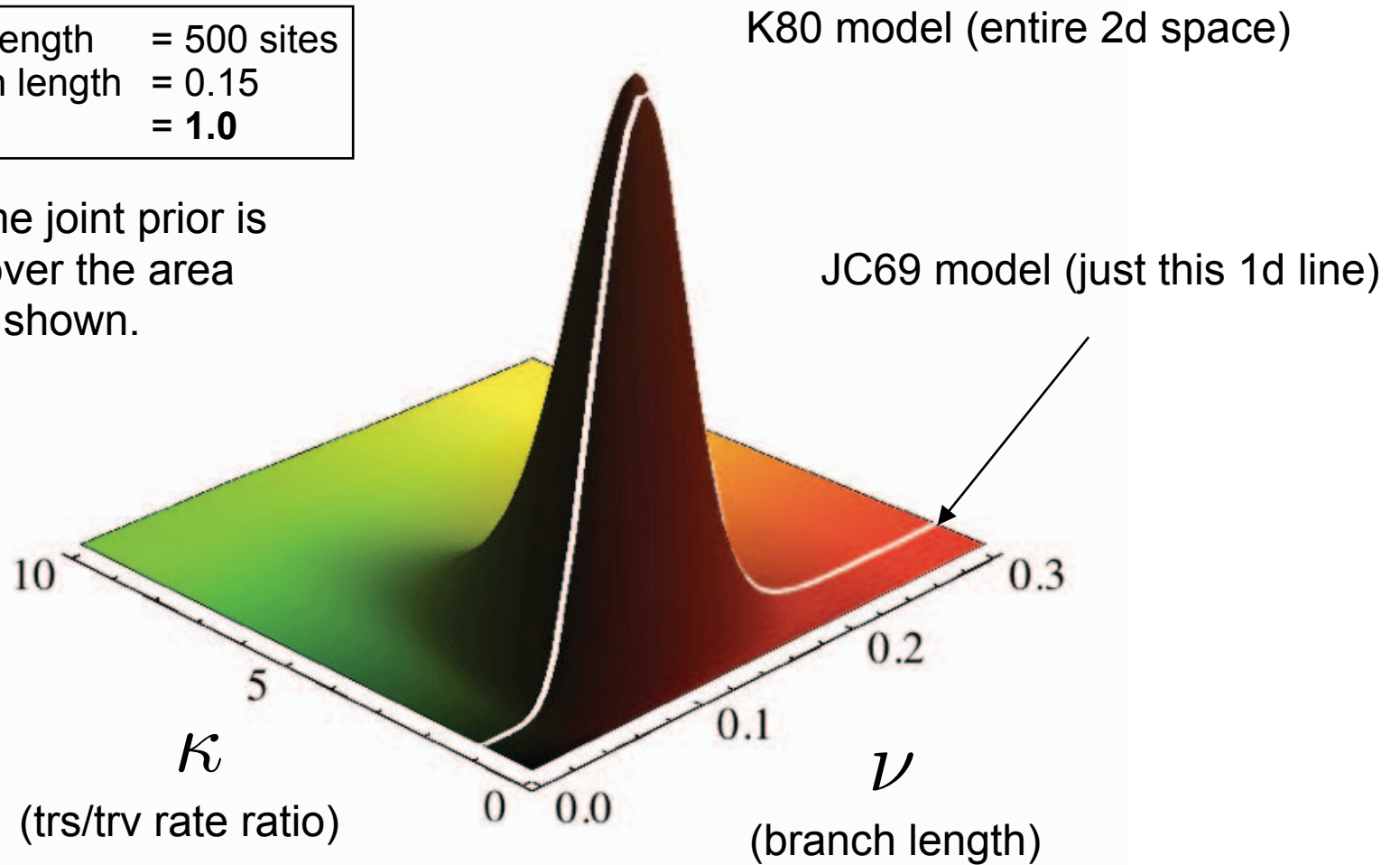
**K80 wins**

# 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