Bayesian phylogenetics
Phylogenetic Biology

Biology 1425

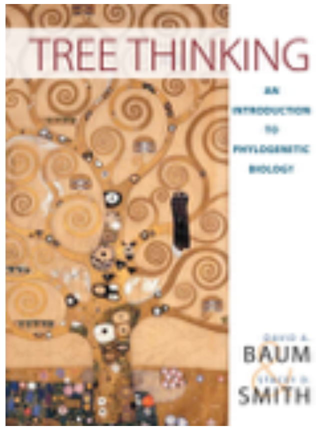
Professor: Casey Dunn, dunnlab.org

Brown University

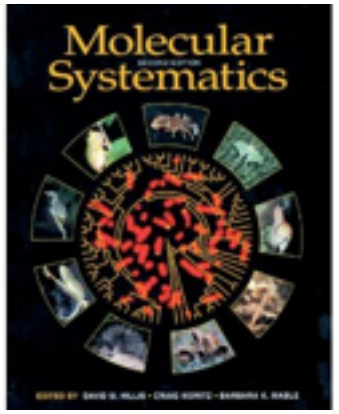
Front matter...

Sources

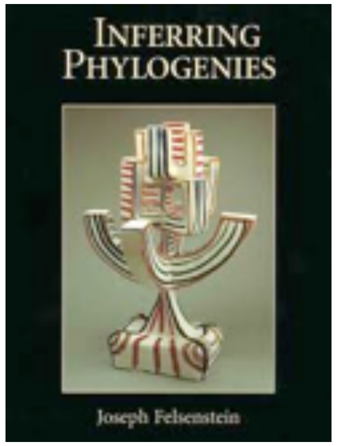
Some non-original content is drawn from:



Baum, D and S. Smith (2012) Tree Thinking: and Introduction to Phylogenetic Biology. Roberts and Company Publishers. ISBN 9781936221165



Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1996). Phylogenetic inference. In: Molecular Systematics, Second Edition. eds: D. M. Hillis, C Moritz, & B. K. Mable. Sinauer Associates. ISBN 9780878932825



Felsenstein, J. (2003) Inferring Phylogenies. Sinauer Associates. ISBN 978-0878931774

Other non-original content is referenced by url.

Sources

Some slides (identified by their footer) are from Paul Lewis's excellent Bayes lecture at the MBL Workshop on Molecular Evolution:

https://molevol.mbl.edu/index.php/Paul_Lewis

Likelihood

Likelihood is the probability of the data (D) given a hypothesis (H):

$$P(D|H)$$

In our case, the data is our aligned matrix (homologous characters and their observed states) and the hypothesis is a particular tree and model of character evolution.

Likelihood

To calculate likelihood, we need:

- Data (eg, character matrix)
- A model of evolution
- Hypothesis (eg, tree and model parameters)
- A mechanism to calculate the likelihood given the above

Maximum Likelihood

$$P(D|H)$$

The hypothesis with the highest likelihood given the data and model.

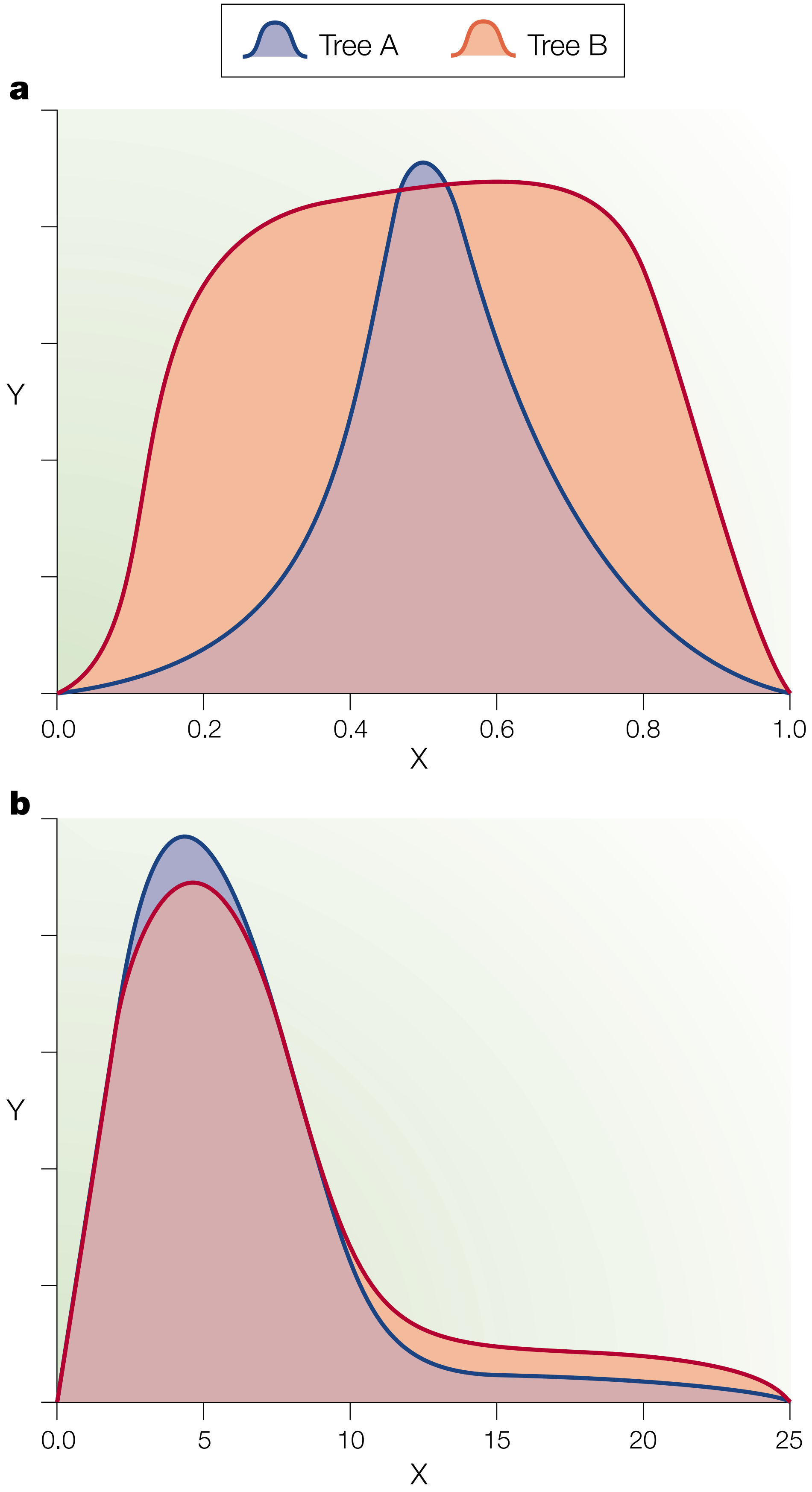


Figure 1 | **Contrast between marginal and joint estimation.** Panels **a** and **b** depict the likelihood profile for two trees versus a hypothetical parameter *x*. The *x* axis represents some nuisance parameter (for example, the ratio of the rate of transitions to the rate of transversions). The *y* axis represents the likelihood in the case of ML, or the posterior-probability density in a Bayesian approach. The area under the likelihood curve for tree A is shown in light blue, the area for tree B is shown in orange. Mauve regions are under the curve for both trees. In both cases, jointly estimating *x* and the tree favours tree A (that is, the highest peak is blue in both cases), but marginalizing over *x* favours tree B (that is, the orange area is greater than the blue area).

(Holder and Lewis 2003)

Bayesian Statistics

An observation about conditional probabilities:

$$P(B|A)P(A) = P(A|B)P(B)$$

The probability of B given A times the probability of A equals the probability of A given B times the probability of B

Bayesian Statistics

$$P(B|A)P(A) = P(A|B)P(B)$$

Rearrange:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The probability of B given A equals the probability of A given B times the probability of B divided by the probability of A

Bayesian Statistics

$$P(B|A)P(A) = P(A|B)P(B)$$

Rearrange:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

The probability of B given A equals the probability of A given B times the probability of B divided by the probability of A

Bayesian Statistics

A demonstration of this:

http://setosa.io/ev/conditional-probability/

Bayesian Statistics

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

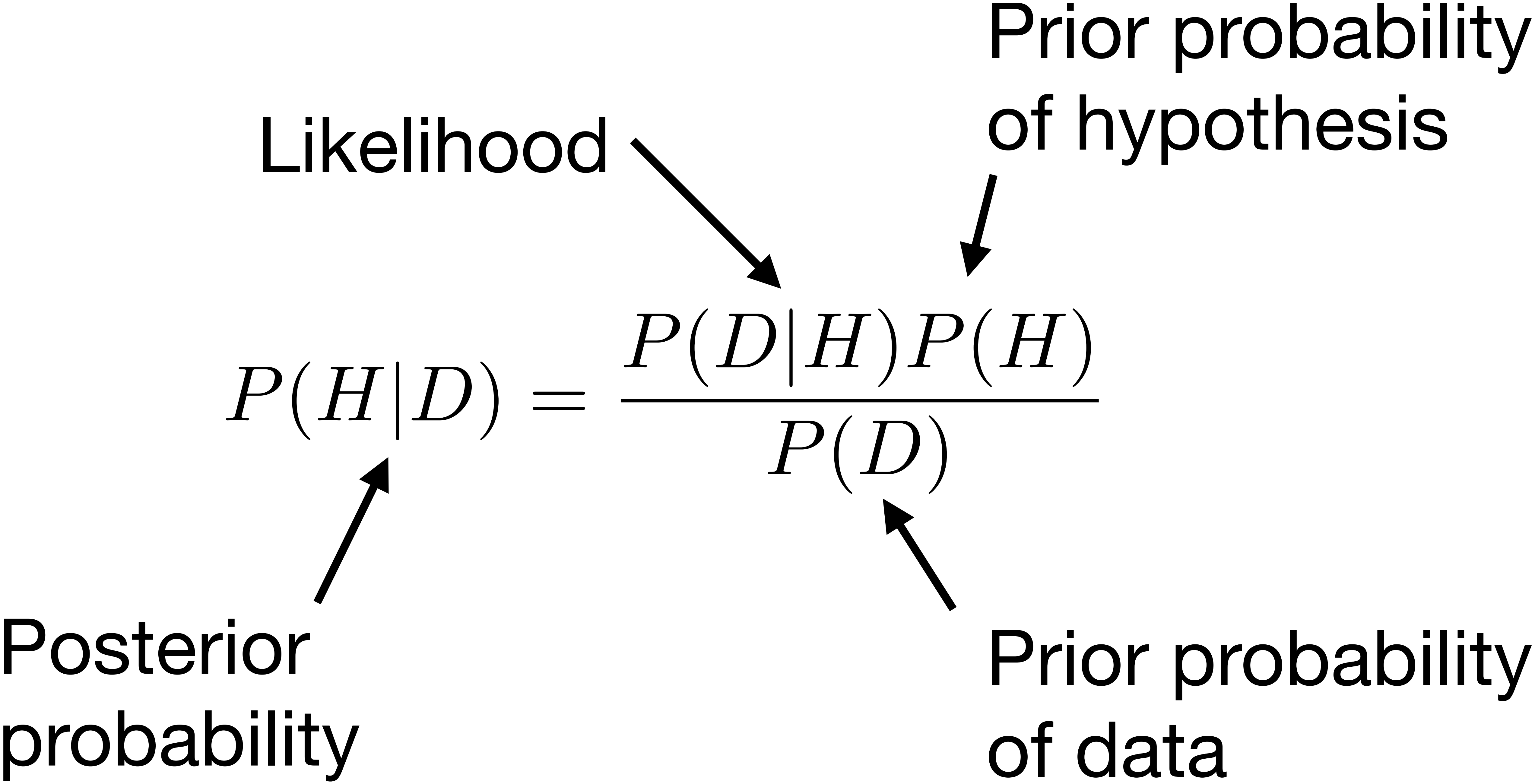
How does this related to phylogenetics? Instead of A and B, let’s talk about hypotheses and data.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Bayesian Statistics

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Bayesian Statistics



$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Likelihood

Prior probability of hypothesis

Posterior probability

Prior probability of data

Bayesian Statistics

The priors explain your expectations before you analyze the new data.

The prior can be your friend

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

$$\begin{aligned}\Pr(+|\text{disease}) &= 0.99\\ \Pr(+|\text{healthy}) &= 0.01\end{aligned}$$

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(disease|+), <u>not</u> Pr(+|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:

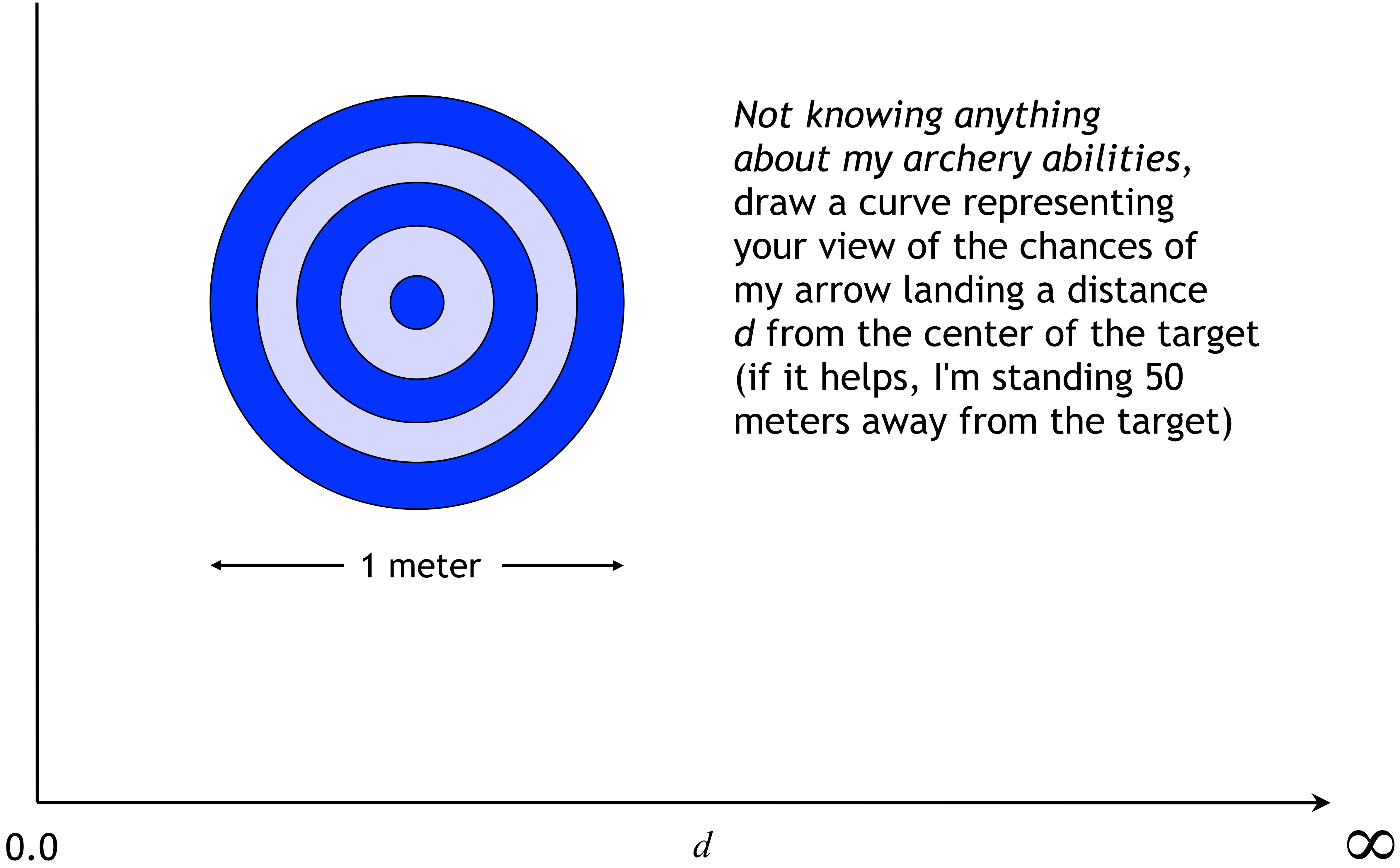
$$
\begin{aligned}
\Pr(\text{disease}|+) &= \frac{(0.99)\left(\frac{1}{1000000}\right)}{(0.99)\left(\frac{1}{1000000}\right) + (0.01)\left(\frac{999999}{1000000}\right)} \\
&\approx 0.0001
\end{aligned}
$$

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.

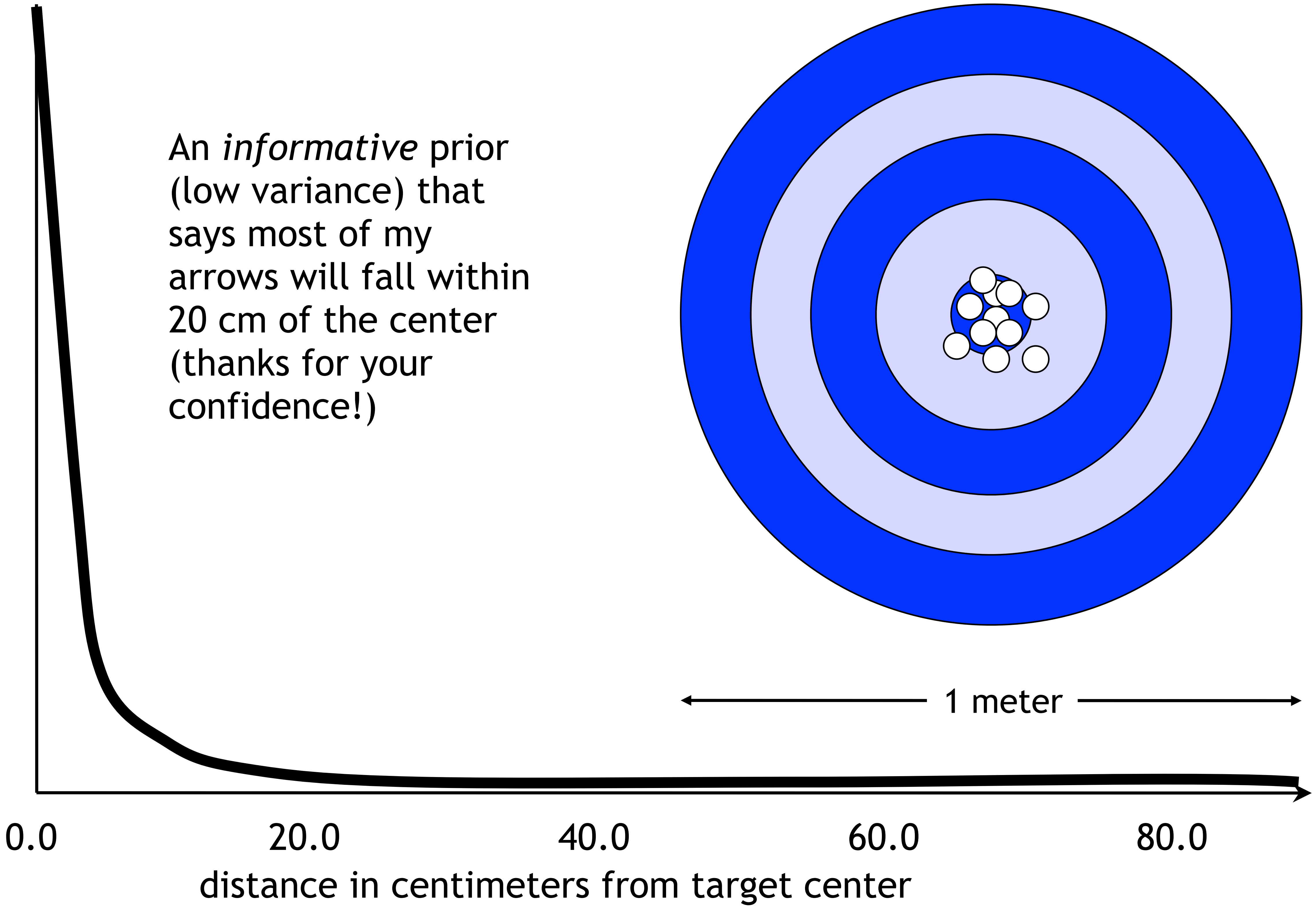
However, in phylogenetics, we often have **lots of data** and use much **less informative priors**, so in phylogenetics ML inference is generally **very reliable**.

If you had to guess...



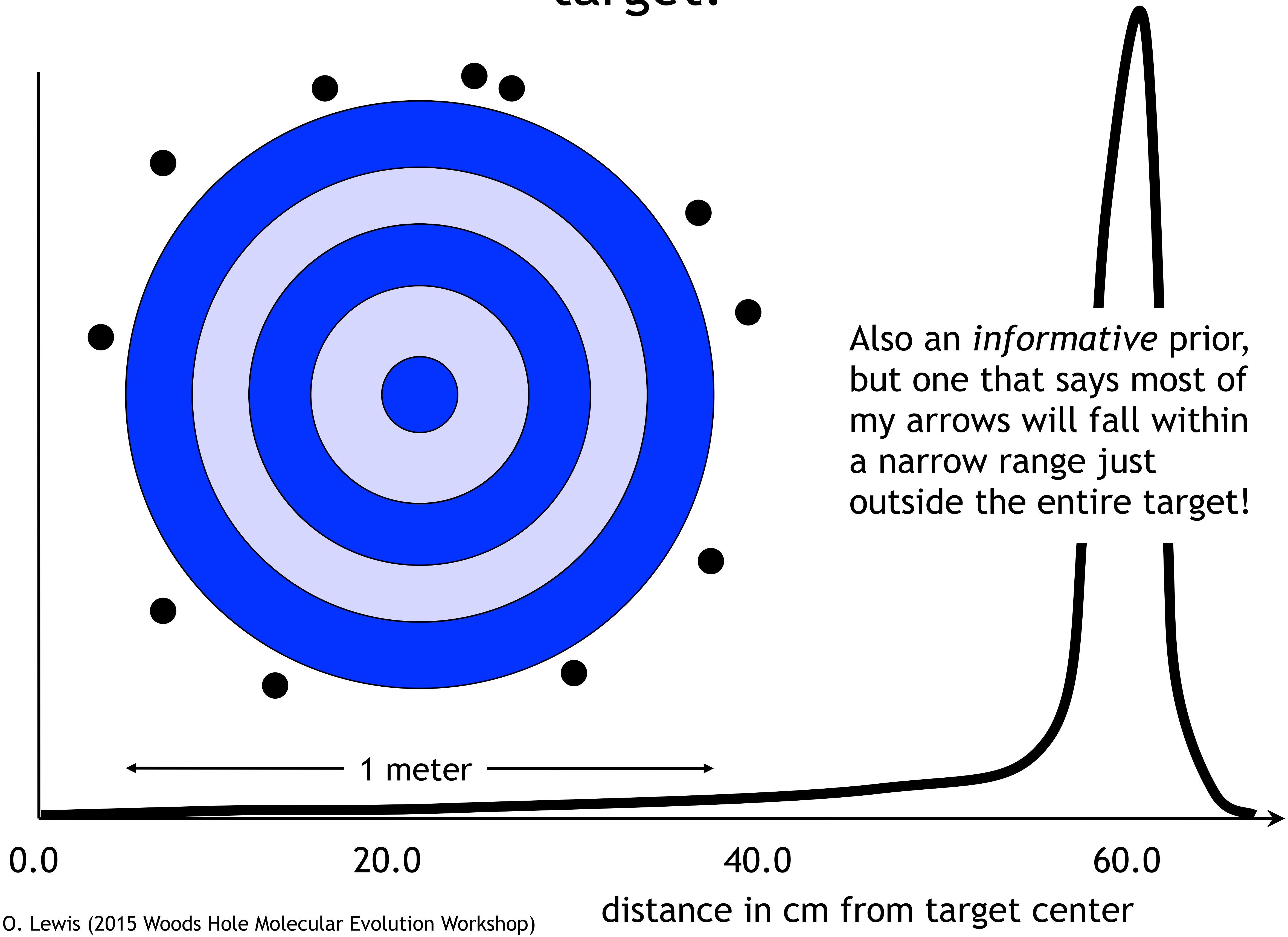
Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance *d* from the center of the target (if it helps, I'm standing 50 meters away from the target)

Case 1: assume I have talent



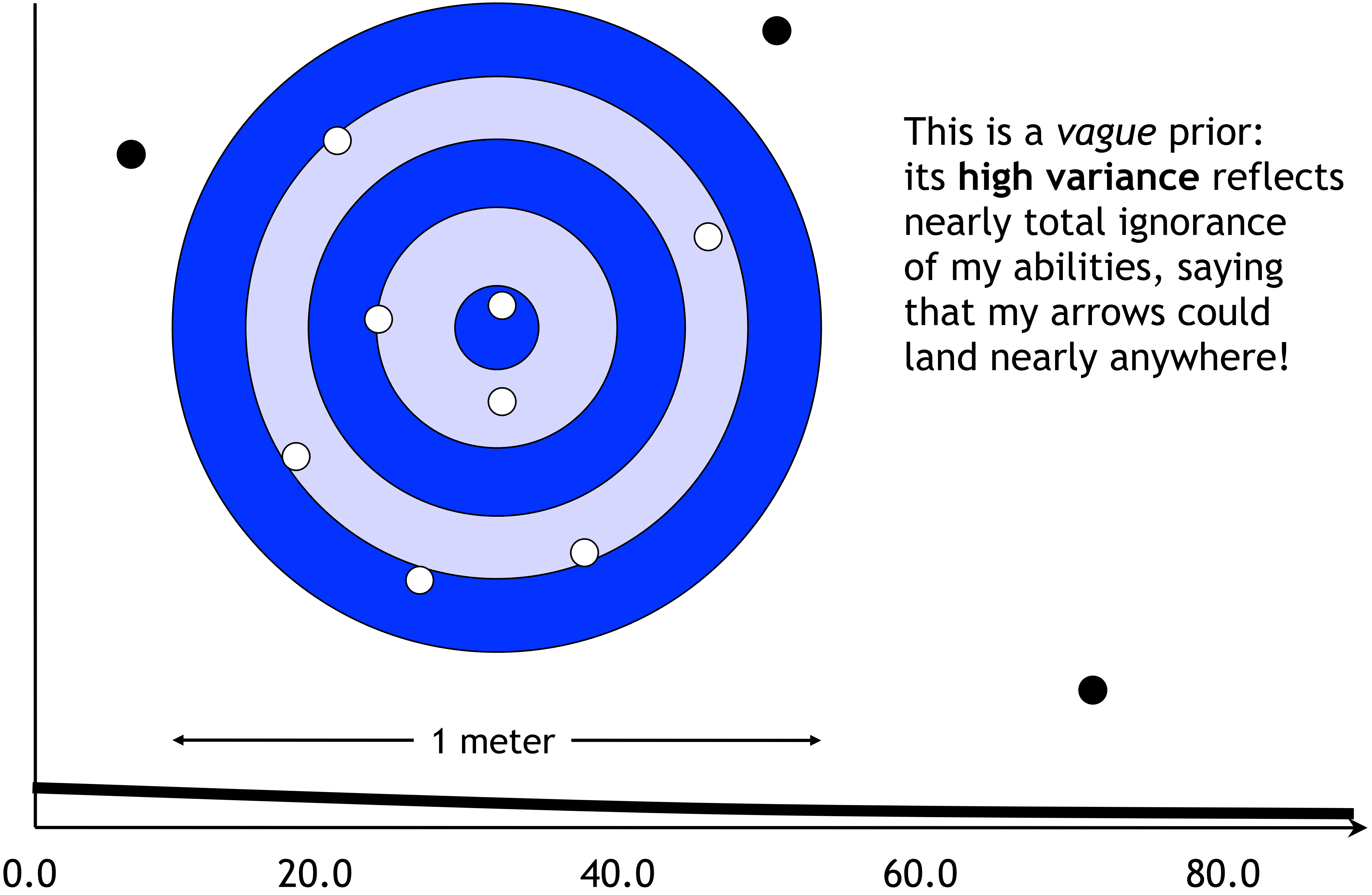
An *informative* prior (low variance) that says most of my arrows will fall within 20 cm of the center (thanks for your confidence!)

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



Also an *informative* prior, but one that says most of my arrows will fall within a narrow range just outside the entire target!

Case 3: assume I have no talent



This is a *vague* prior:
its **high variance** reflects
nearly total ignorance
of my abilities, saying
that my arrows could
land nearly anywhere!

Paul O. Lewis (2015 Woods Hole Molecular Evolution Workshop)

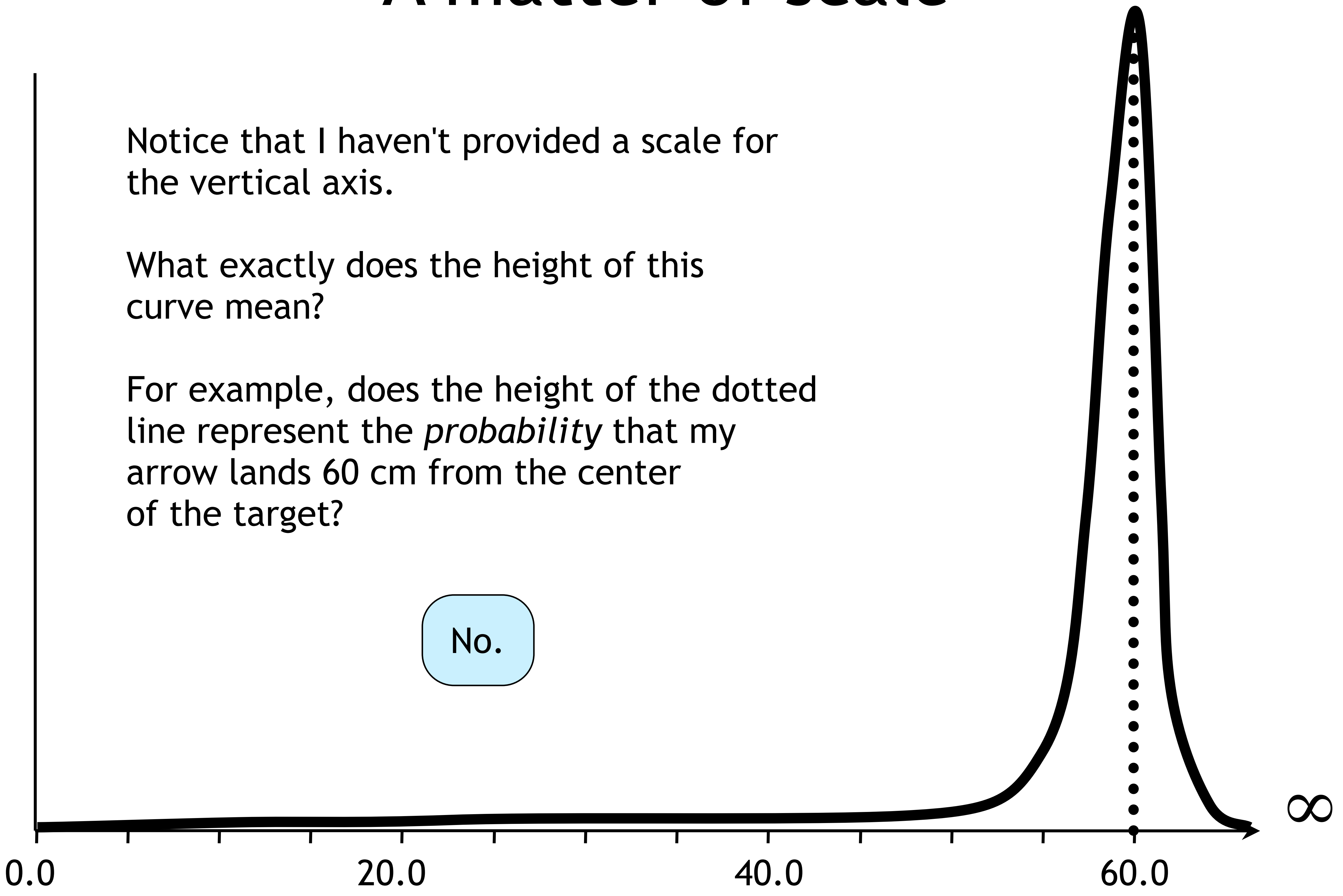
A matter of scale

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.

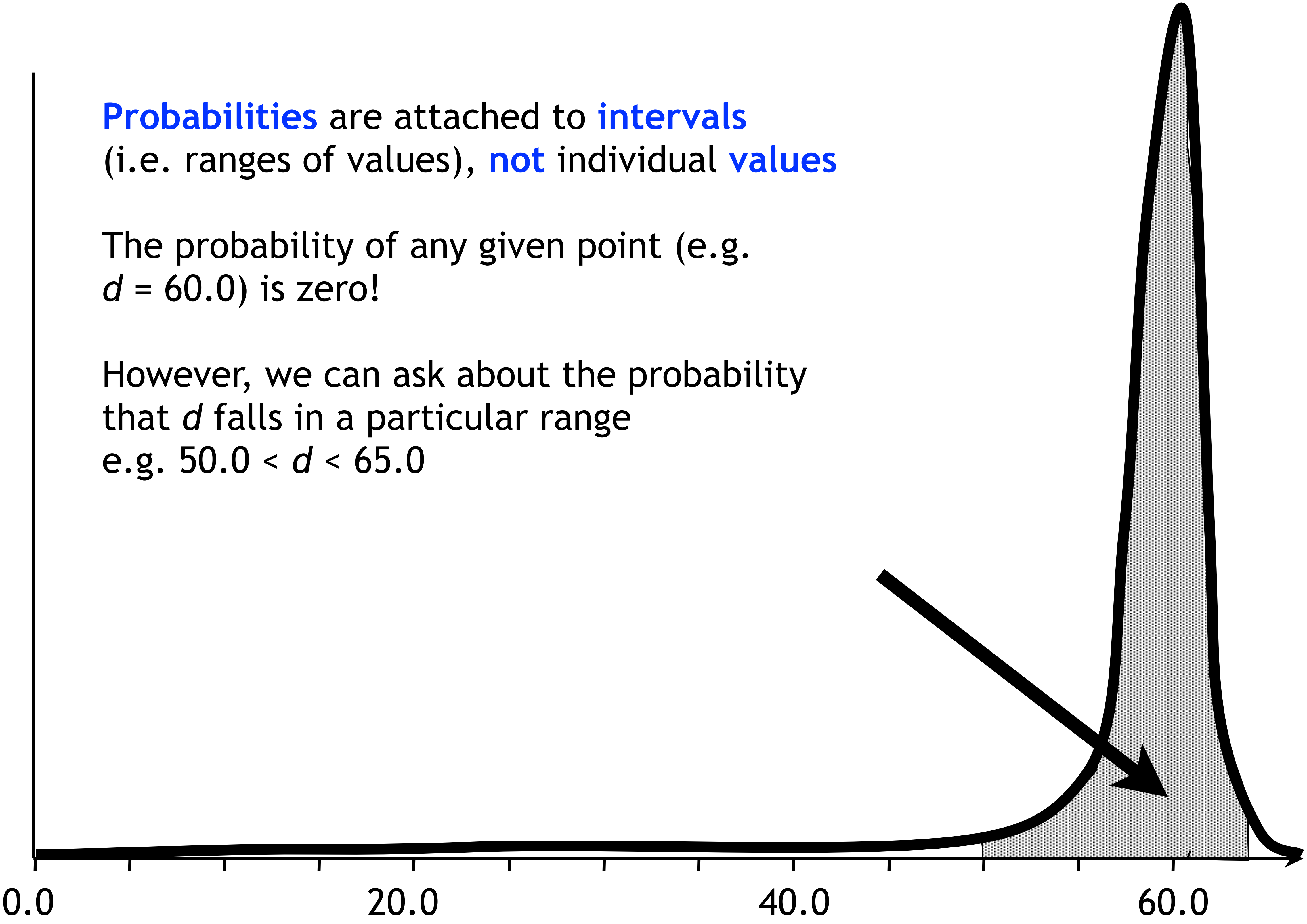


Probabilities are associated with intervals

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



Bayesian Statistics

How do we calculate the posterior probability?

$$\underbrace{P(H|D)}_{\text{Posterior probability}} = \frac{\overbrace{P(D|H)}^{\text{Likelihood}}\,\overbrace{P(H)}^{\text{Prior probability of hypothesis}}}{\underbrace{P(D)}_{\text{Prior probability of data}}}$$

Bayesian Statistics

How do we calculate the posterior probability?

Likelihood - we know how to do this already

Reflects any information we already had about the hypothesis

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Posterior probability

Normalizes the posterior distribution so that the area is 1

Bayesian Statistics

How do we calculate the prior probability of the data?

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_H P(D|H)P(H)}$$

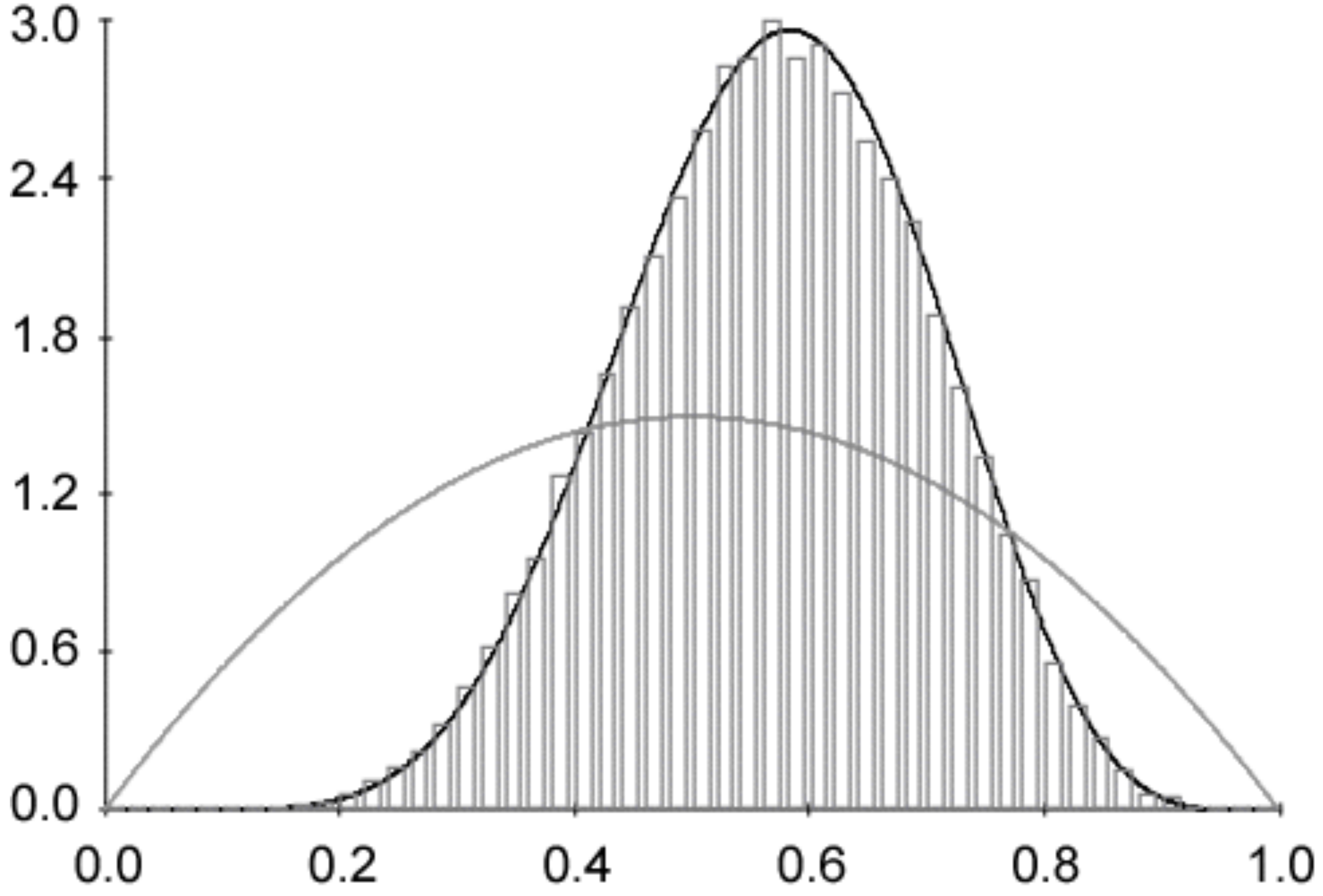
Bayesian Statistics

How do we calculate the prior probability of the data?

We don’t… we approximate the posterior in a way that never requires us to calculate the prior probability of the data.

II. Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC)

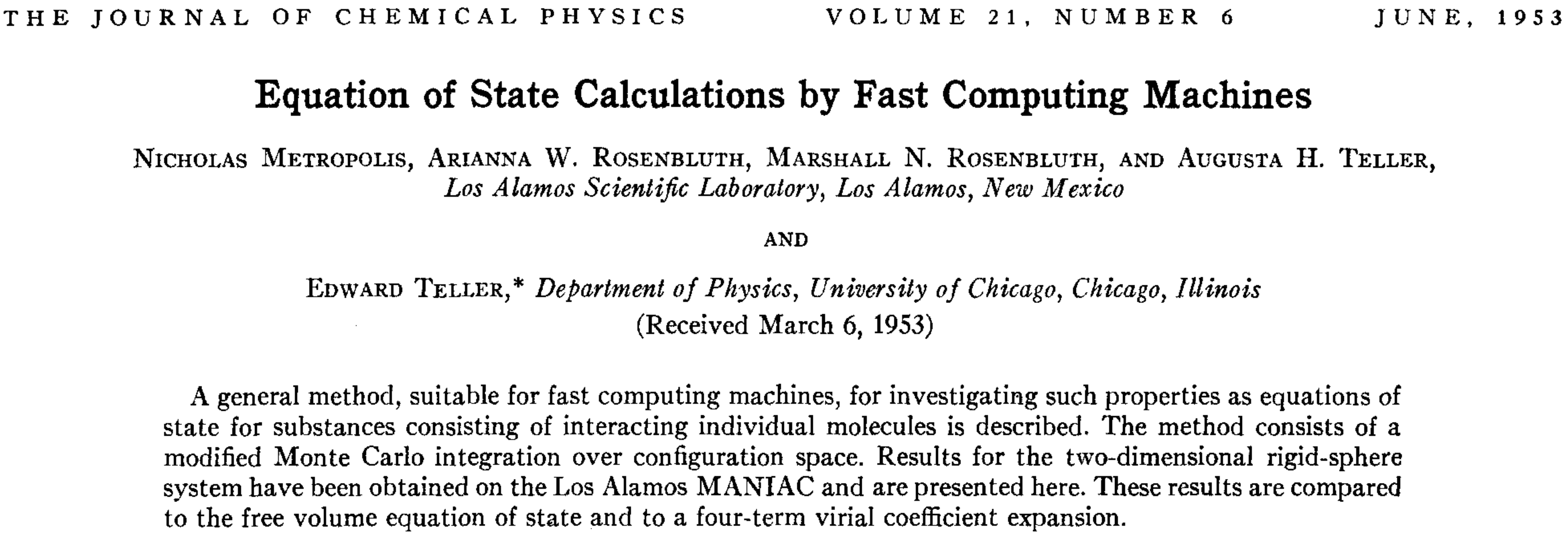


For more complex problems, we might settle for a

good approximation

to the posterior distribution

MCMC from the dawn of statistical computing



THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

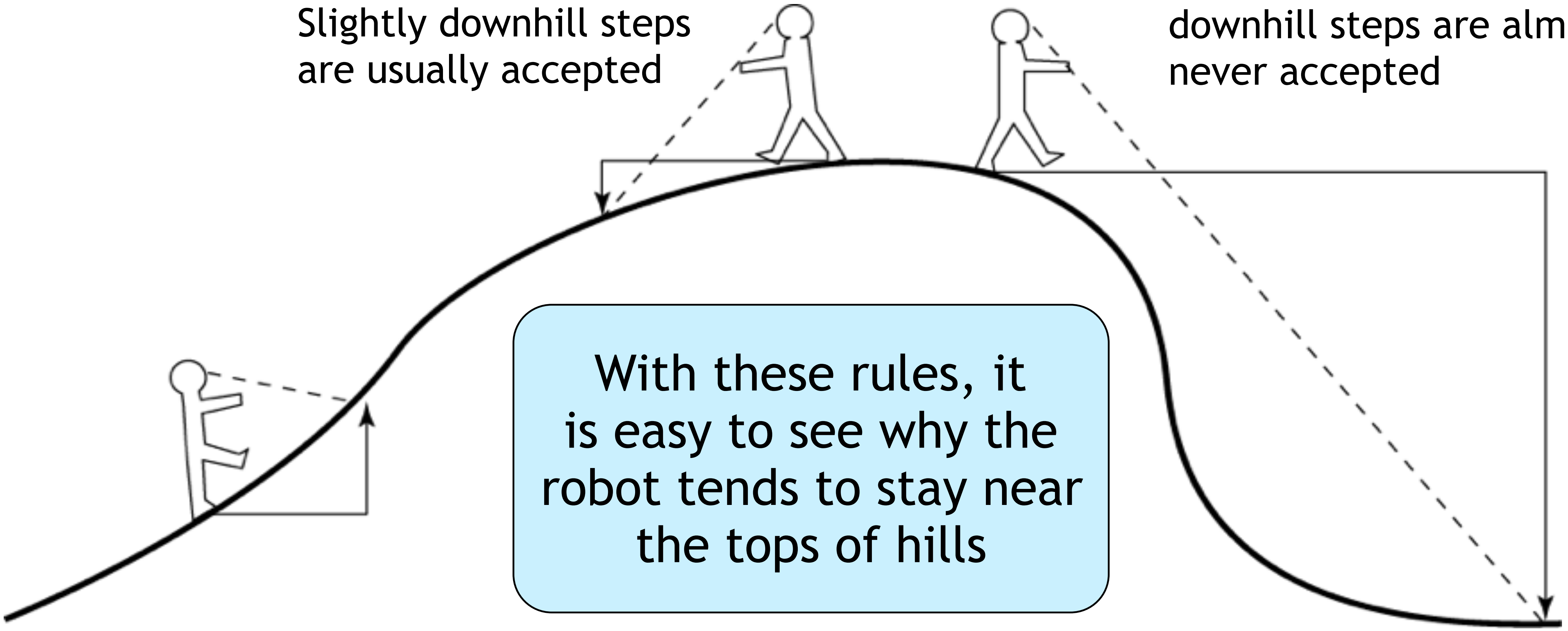
AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

MCMC robot's rules



Slightly downhill steps are usually accepted

Drastic "off the cliff" downhill steps are almost never accepted

Uphill steps are always accepted

With these rules, it is easy to see why the robot tends to stay near the tops of hills

What explicit criteria can be used to decide which steps to accept?

$$R = \frac{P(H^*|D)}{P(H|D)}$$

Where H is the current hypothesis and H* is the new hypothesis being considered.

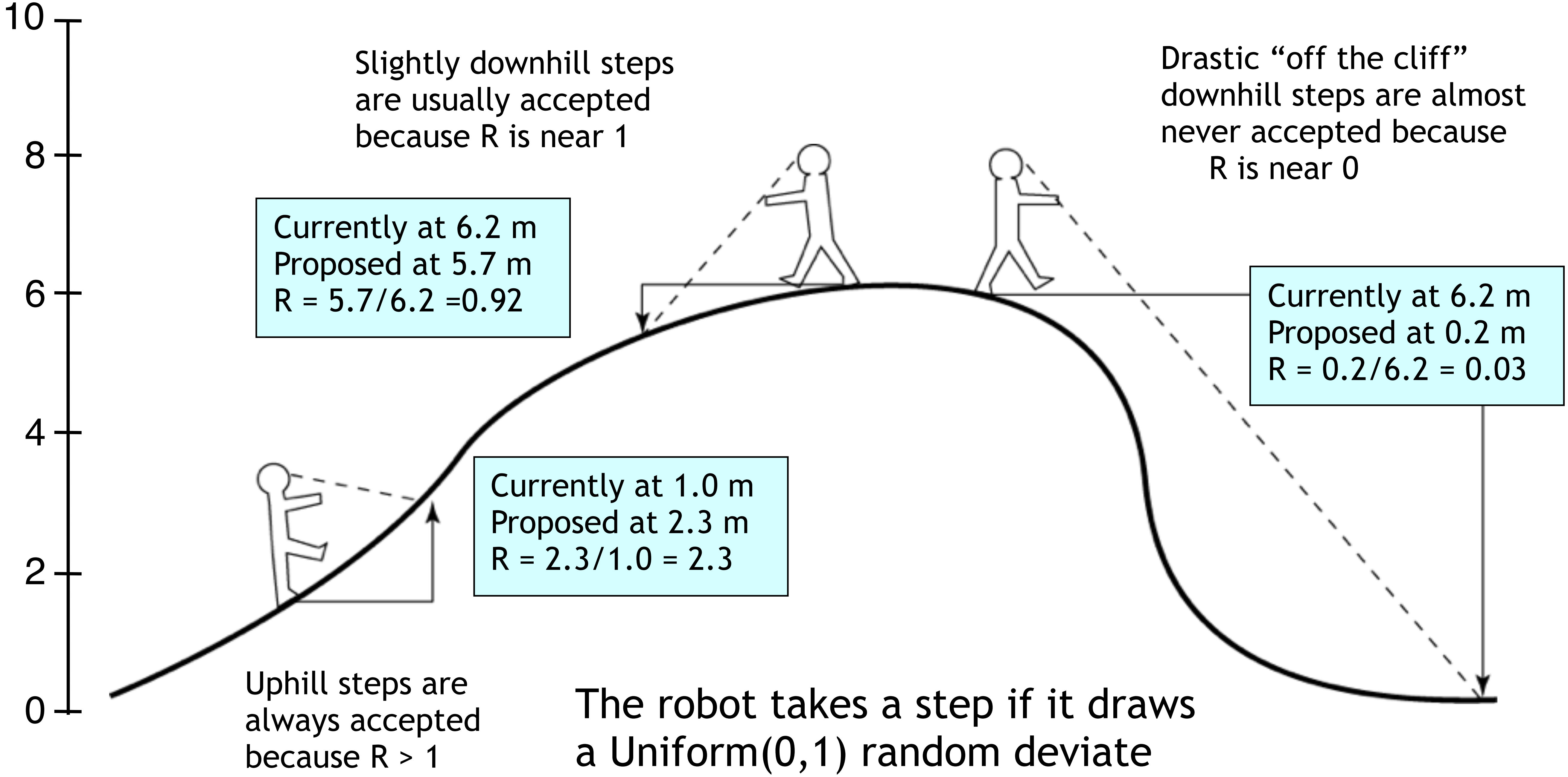
What explicit criteria can be used to decide which steps to accept?

$$R = \frac{P(H^*|D)}{P(H|D)}$$

If R > 1, accept the new hypothesis.

If R < 1, accept the new hypothesis with probability R.

(Actual) MCMC robot rules



Slightly downhill steps are usually accepted because R is near 1

Currently at 6.2 m
Proposed at 5.7 m
R = 5.7/6.2 =0.92

Drastic “off the cliff” downhill steps are almost never accepted because R is near 0

Currently at 6.2 m
Proposed at 0.2 m
R = 0.2/6.2 = 0.03

Currently at 1.0 m
Proposed at 2.3 m
R = 2.3/1.0 = 2.3

Uphill steps are always accepted because R > 1

The robot takes a step if it draws a Uniform(0,1) random deviate that is less than or equal to R

Why bother?

$$R = \frac{P(H^*|D)}{P(H|D)} = \frac{P(D|H^*)P(H^*)}{P(D)} \frac{P(D)}{P(D|H)P(H)}$$

$$R = \frac{P(H^*|D)}{P(H|D)} = \frac{P(D|H^*)P(H^*)}{P(D|H)P(H)}$$

P(D) cancels out!!!!!!!!!!

MCRobot (or "MCMC Robot")

Free app for **Windows** or **iPhone/iPad** available
from http://mcmcrobot.org/
(note: version currently on Apple Store does
not work under iOS 8 - will replace it soon)

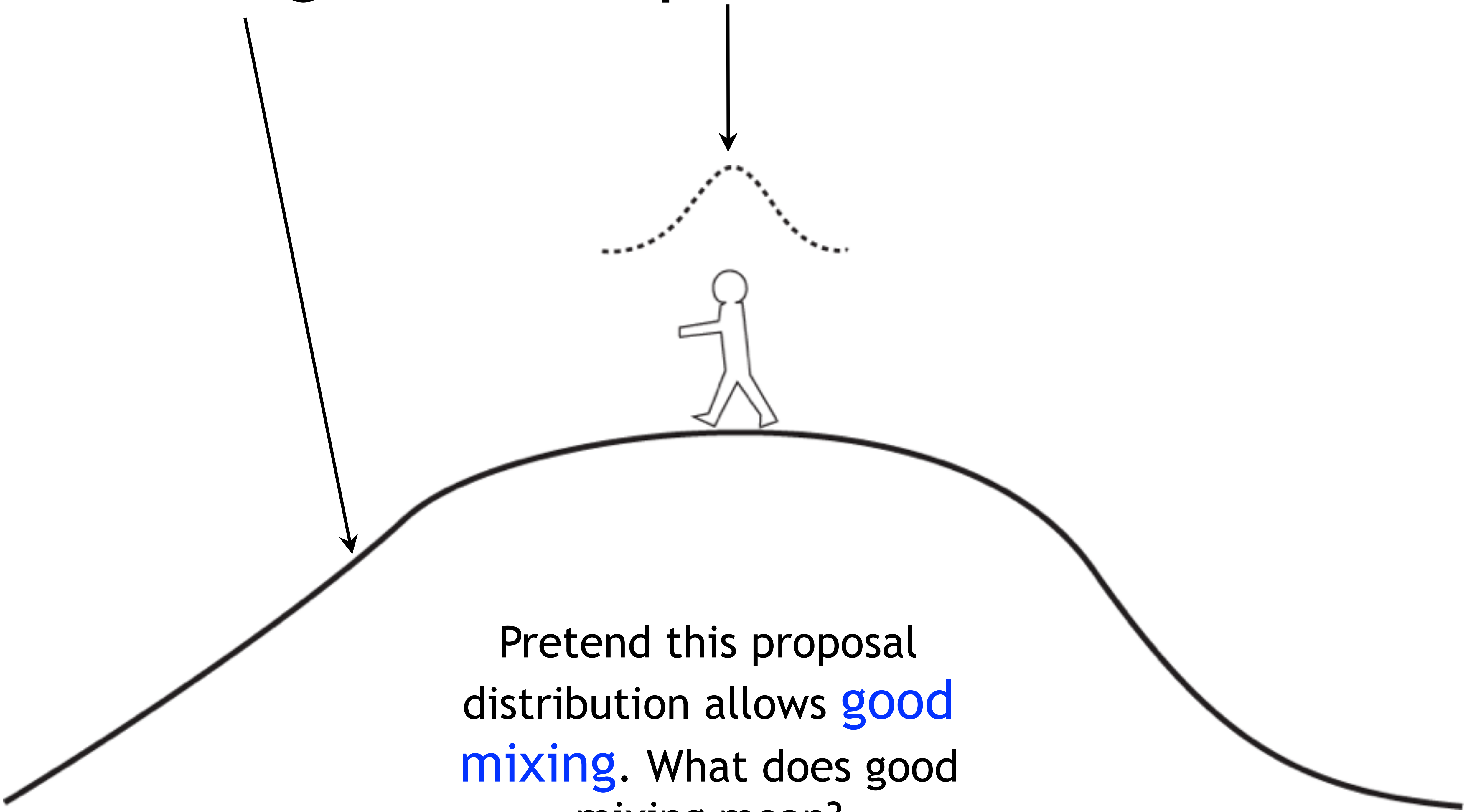
Android: next year?

Mac version: not soon
(but see John Huelsenbeck's iMCMC app for MacOS:
http://cteg.berkeley.edu/software.html)

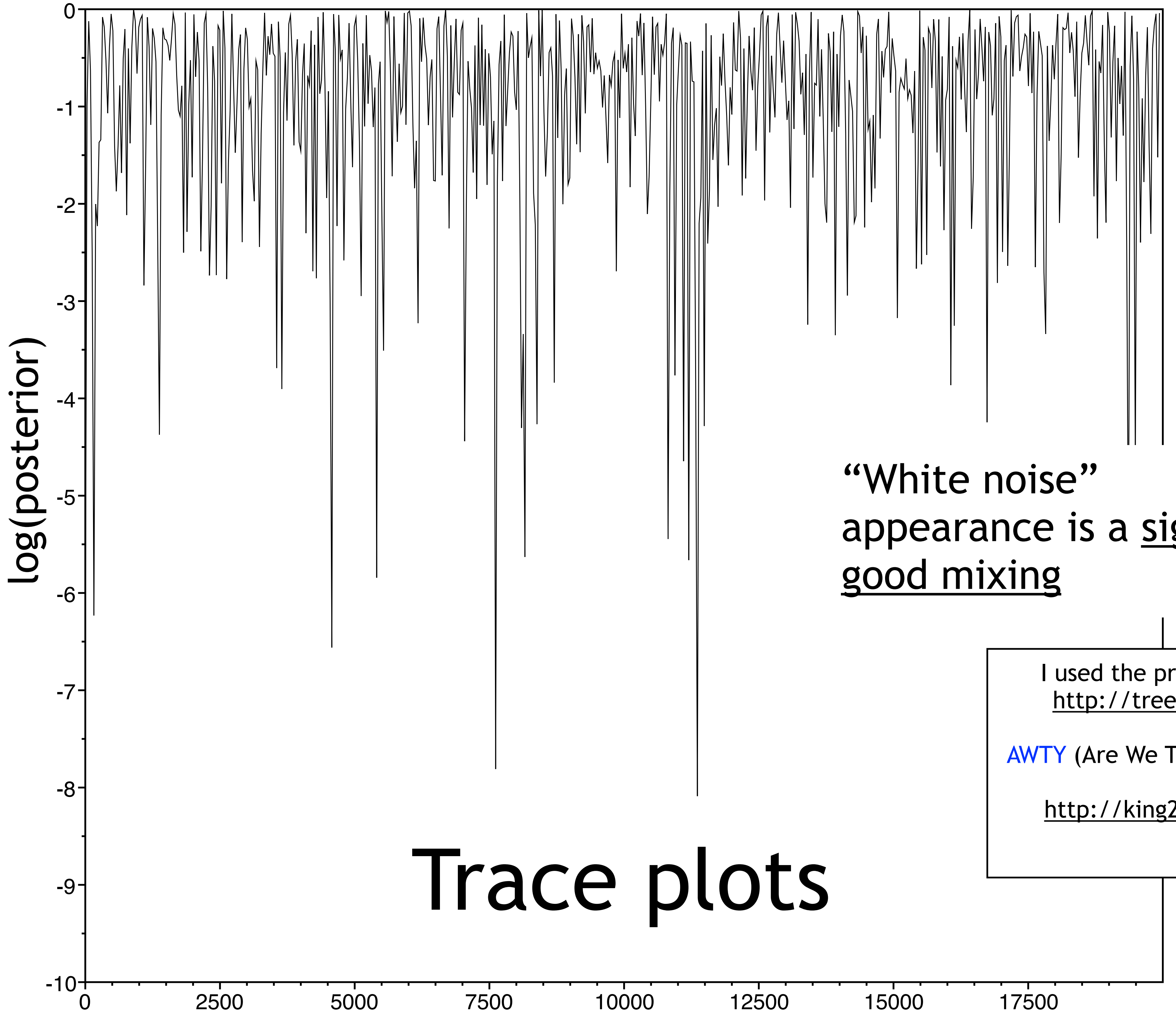
Target distribution - the distribution you are trying to estimate

Proposal distribution - the distribution of steps that are proposed by the robot

Target vs. Proposal Distributions



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



Trace plots

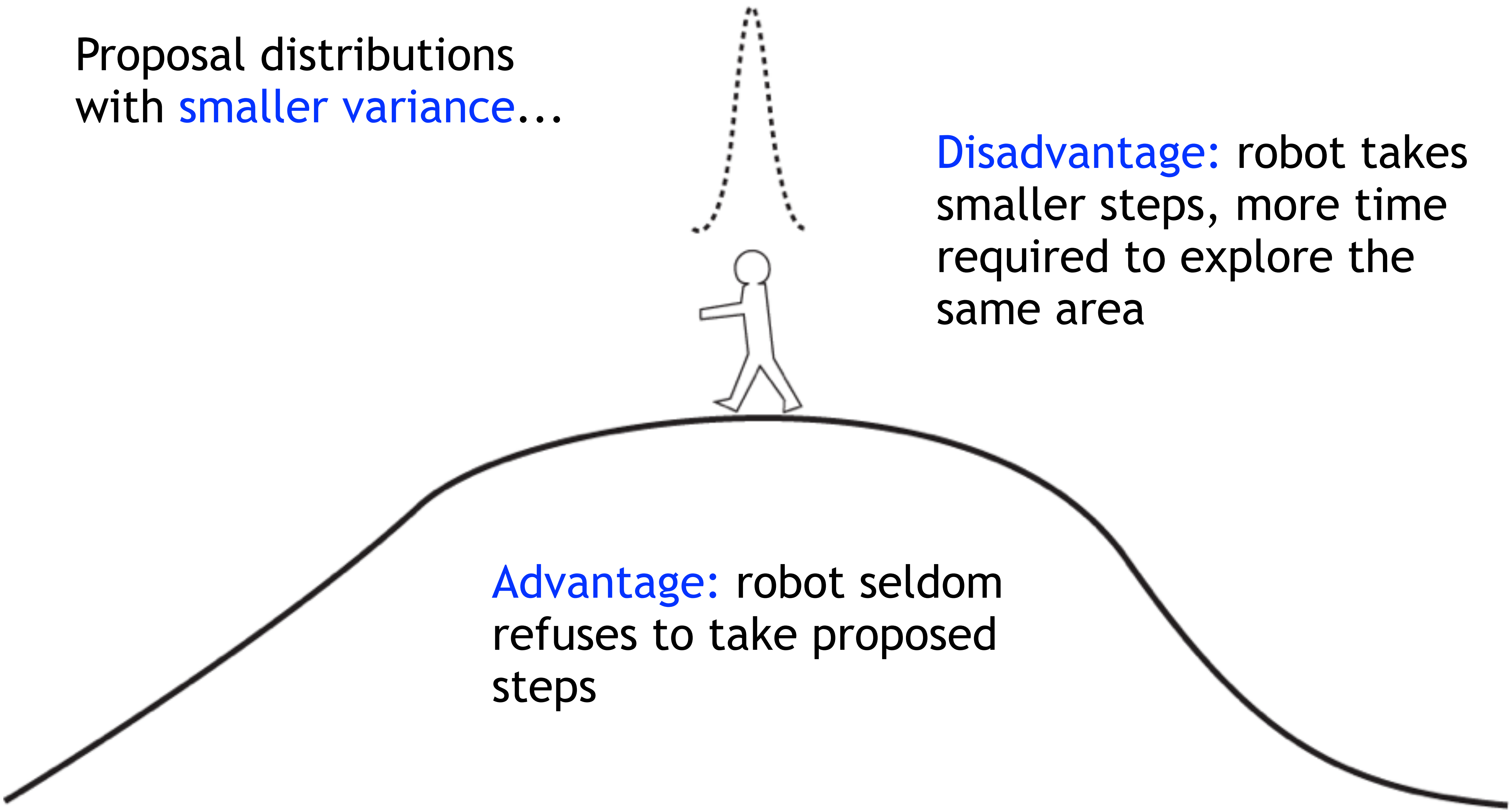
"White noise" appearance is a sign of good mixing

I used the program Tracer to create this plot:
http://tree.bio.ed.ac.uk/software/tracer/

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

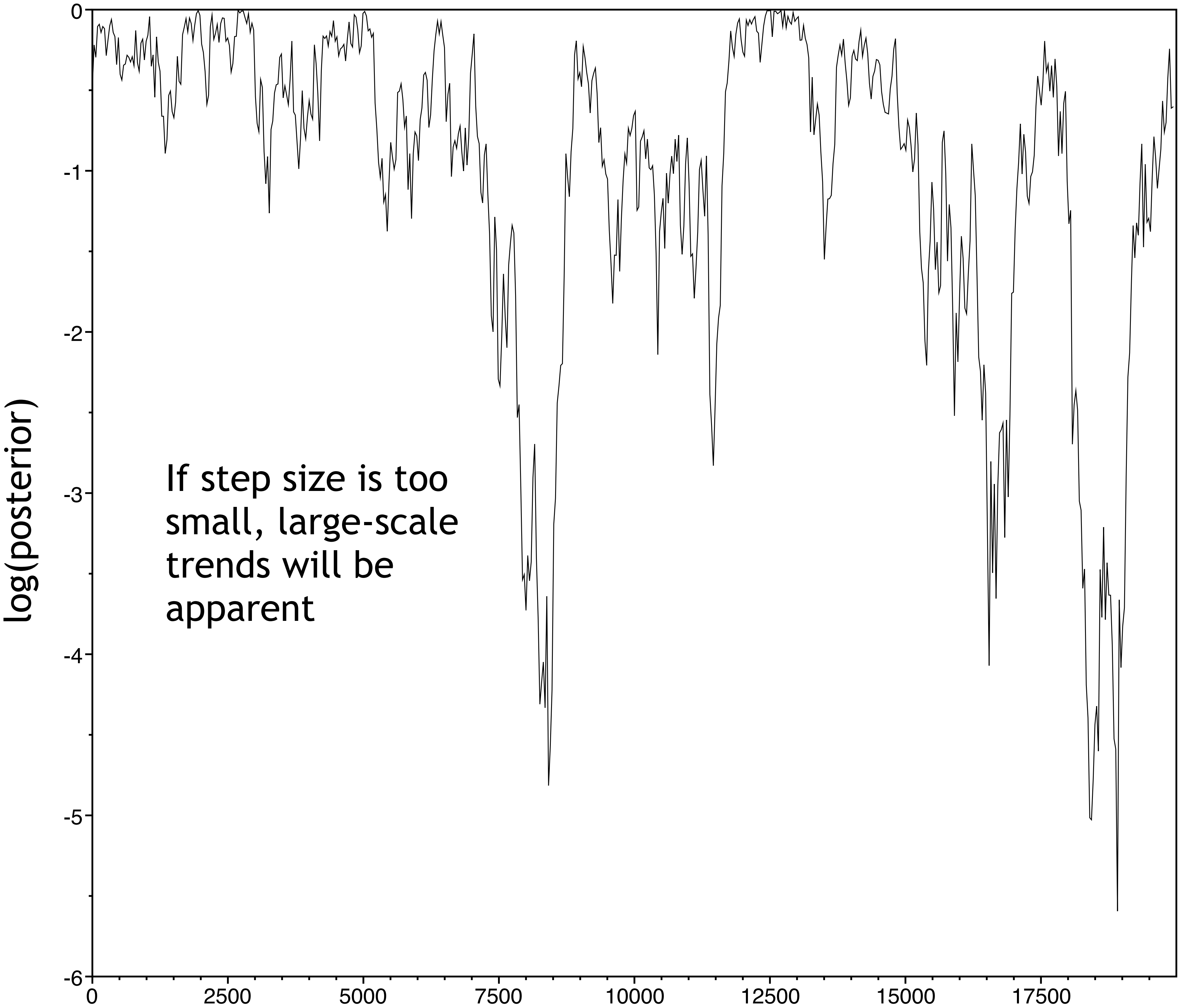
Target vs. Proposal Distributions

Proposal distributions with smaller variance...



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

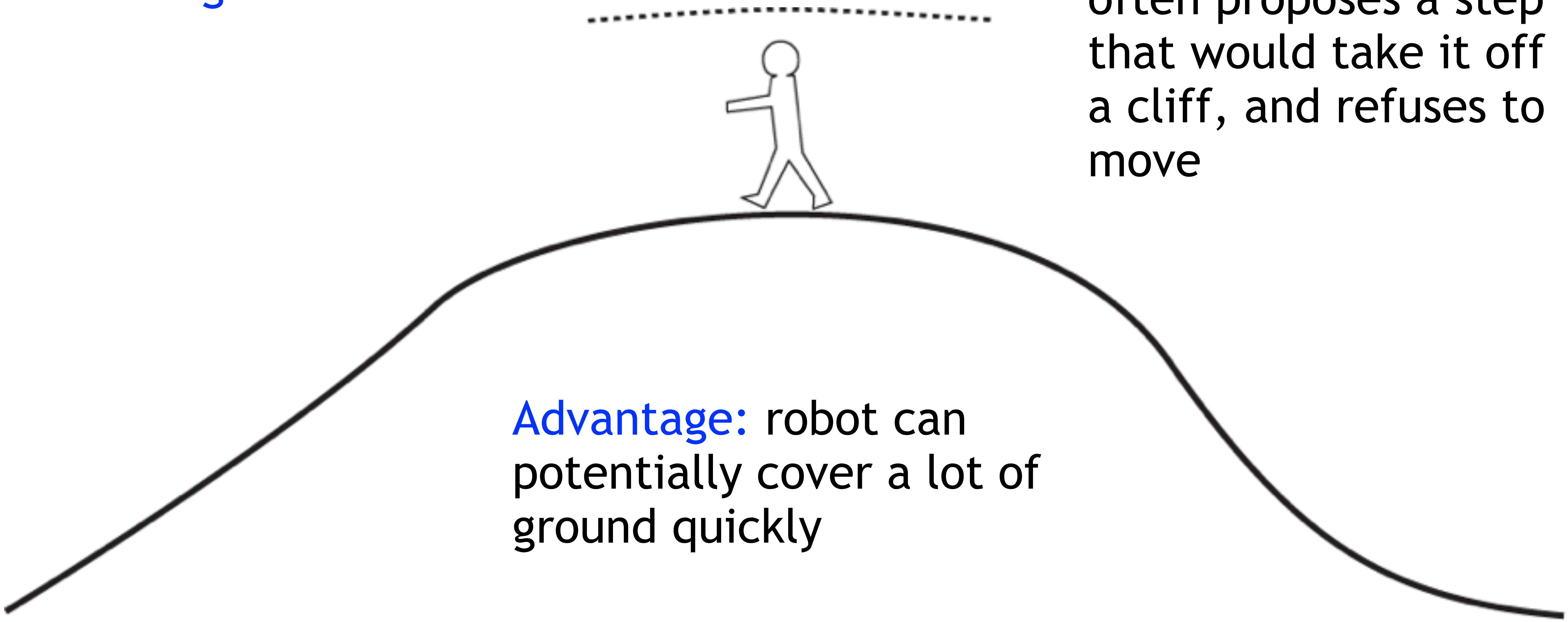
Advantage: robot seldom refuses to take proposed steps



If step size is too small, large-scale trends will be apparent

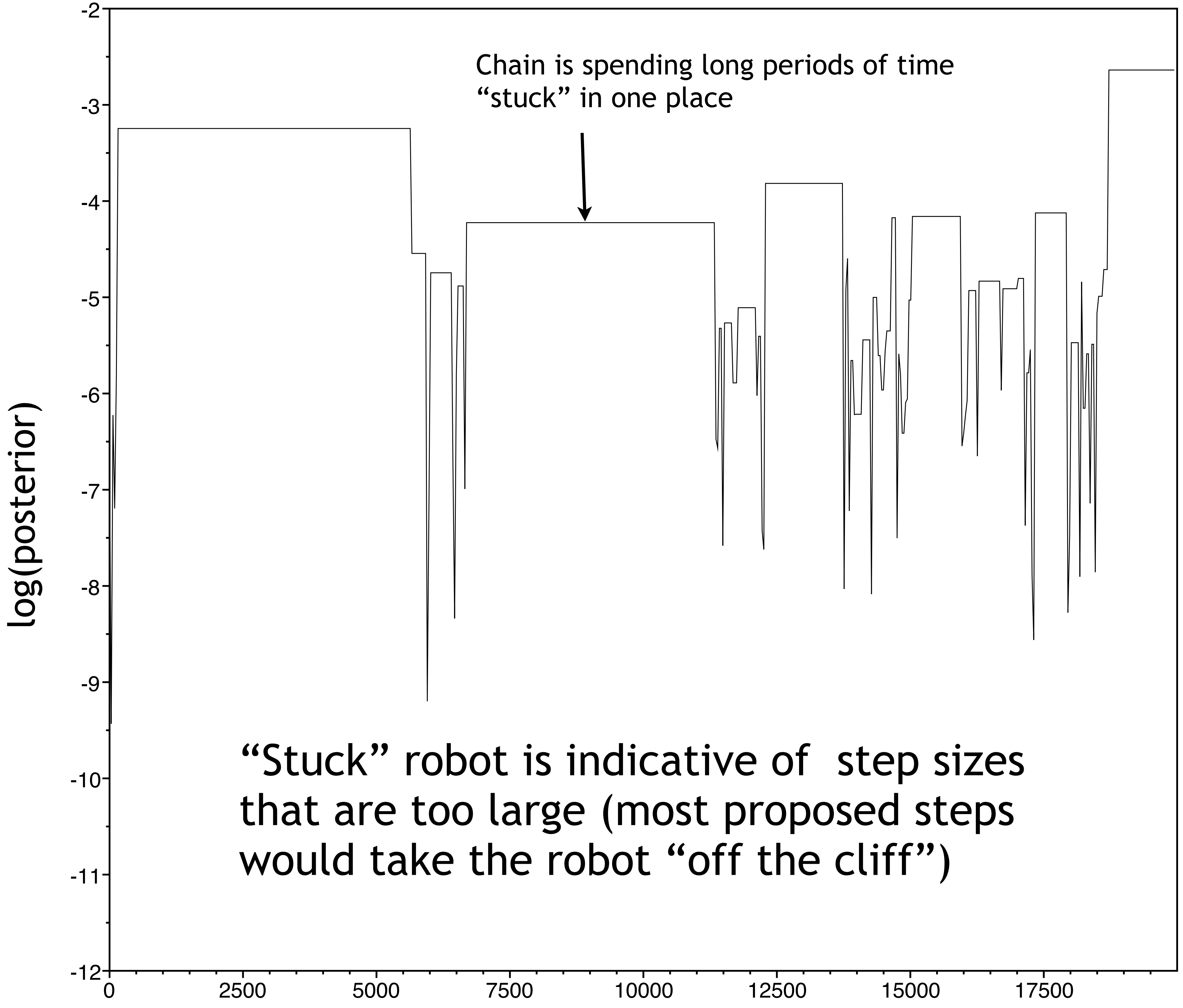
Target vs. Proposal Distributions

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



Chain is spending long periods of time "stuck" in one place

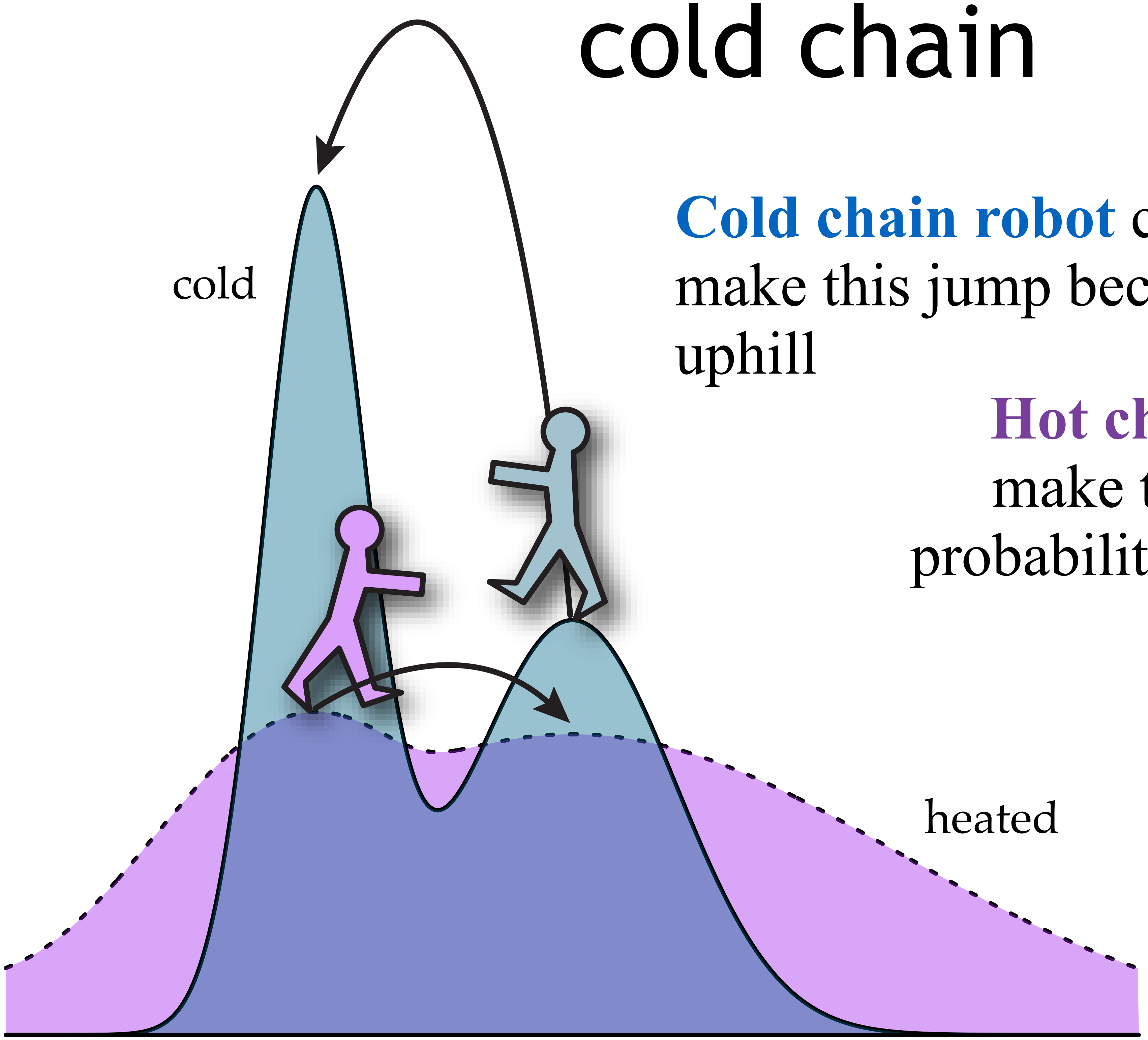
"Stuck" robot is indicative of step sizes that are too large (most proposed steps would take the robot "off the cliff")

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.).

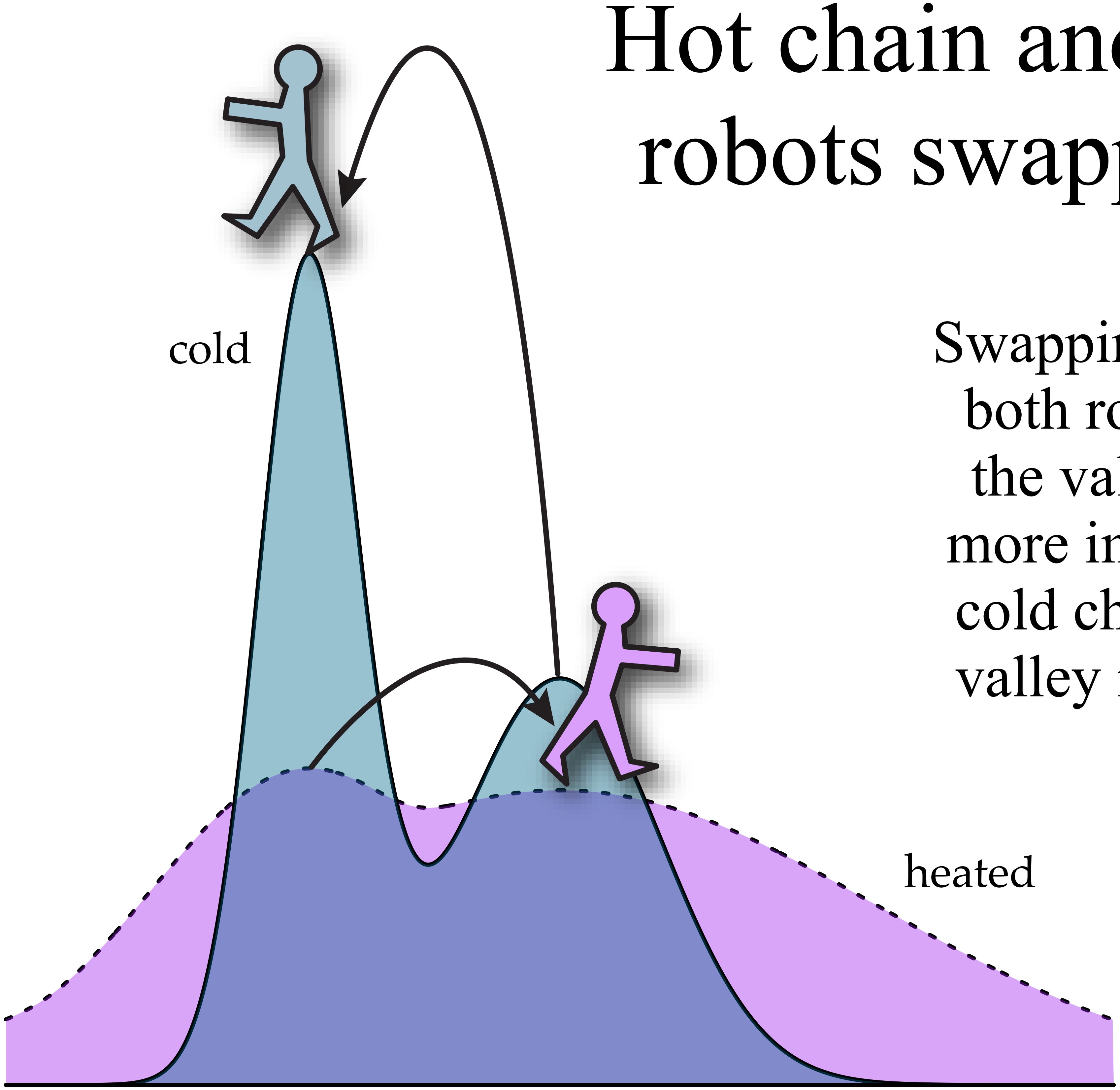
Heated chains act as scouts for the cold chain



Cold chain robot can easily make this jump because it is uphill

Hot chain robot can also make this jump with high probability because it is only slightly downhill

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

Bayesian Statistics

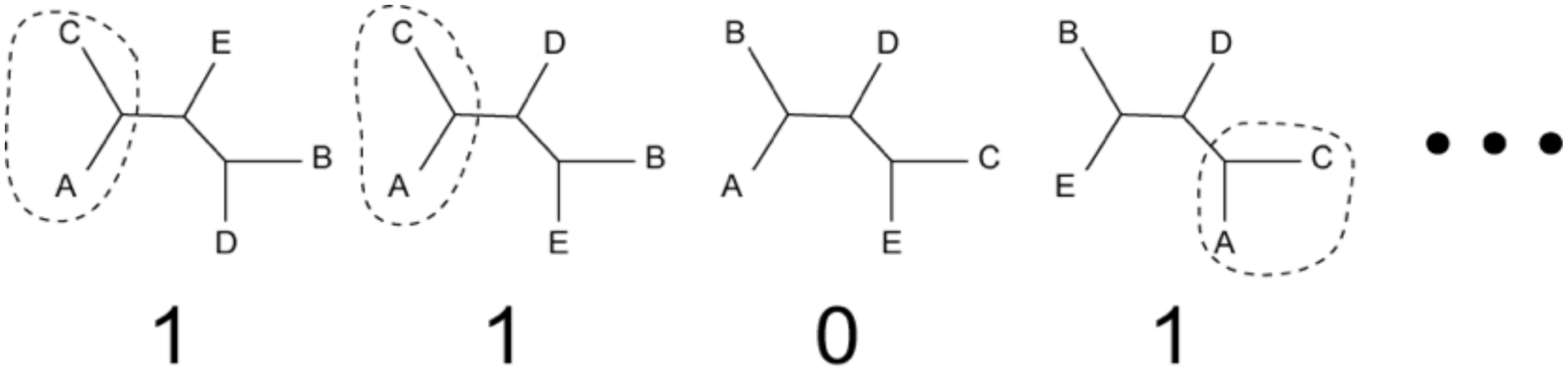
Burnin - Throw away the first steps that were made before the robot converged on the target distribution

Bayesian Statistics

Burnin - Throw away the first steps that were made before the robot converged on the target distribution

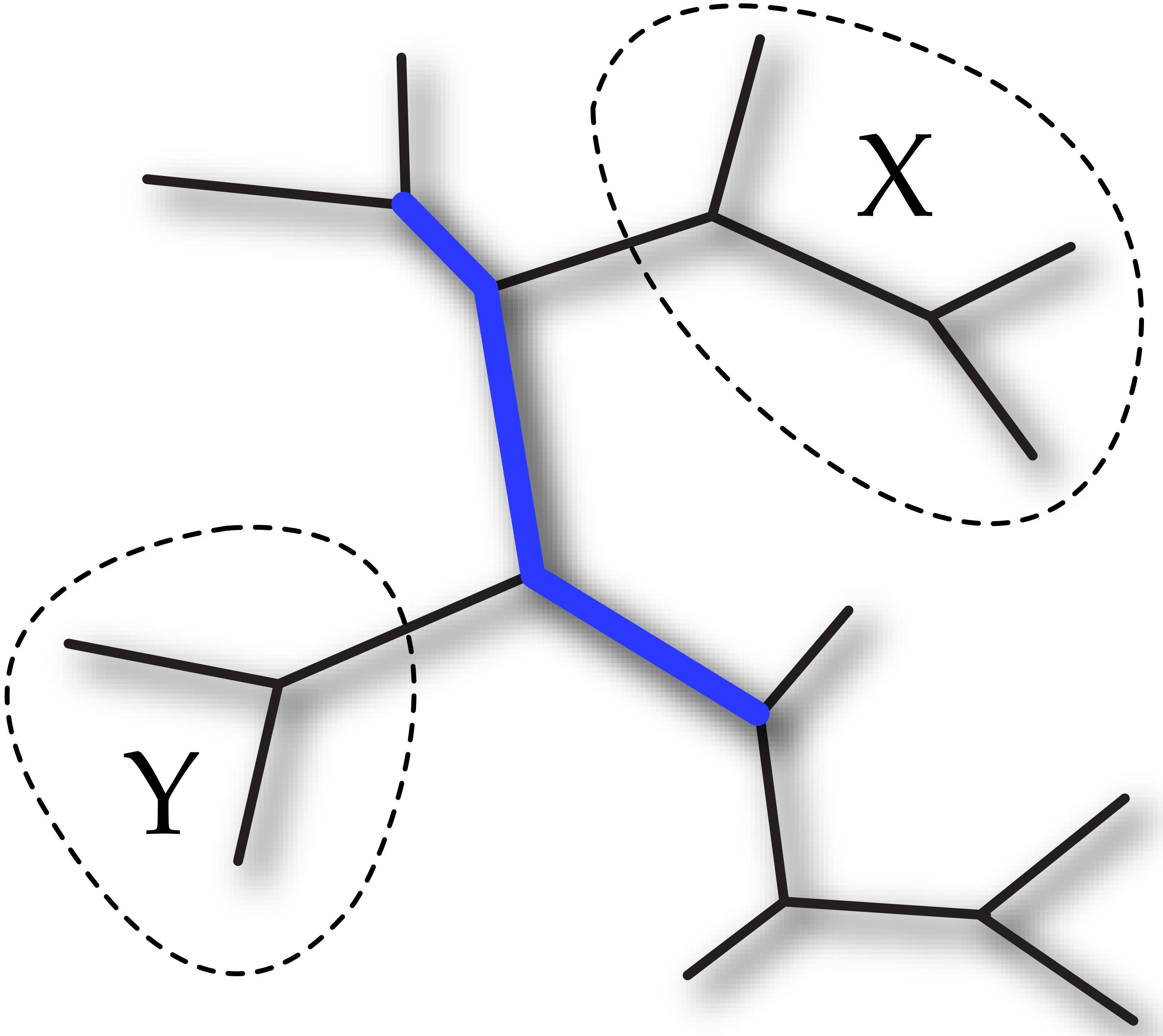
III. Bayesian phylogenetics

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

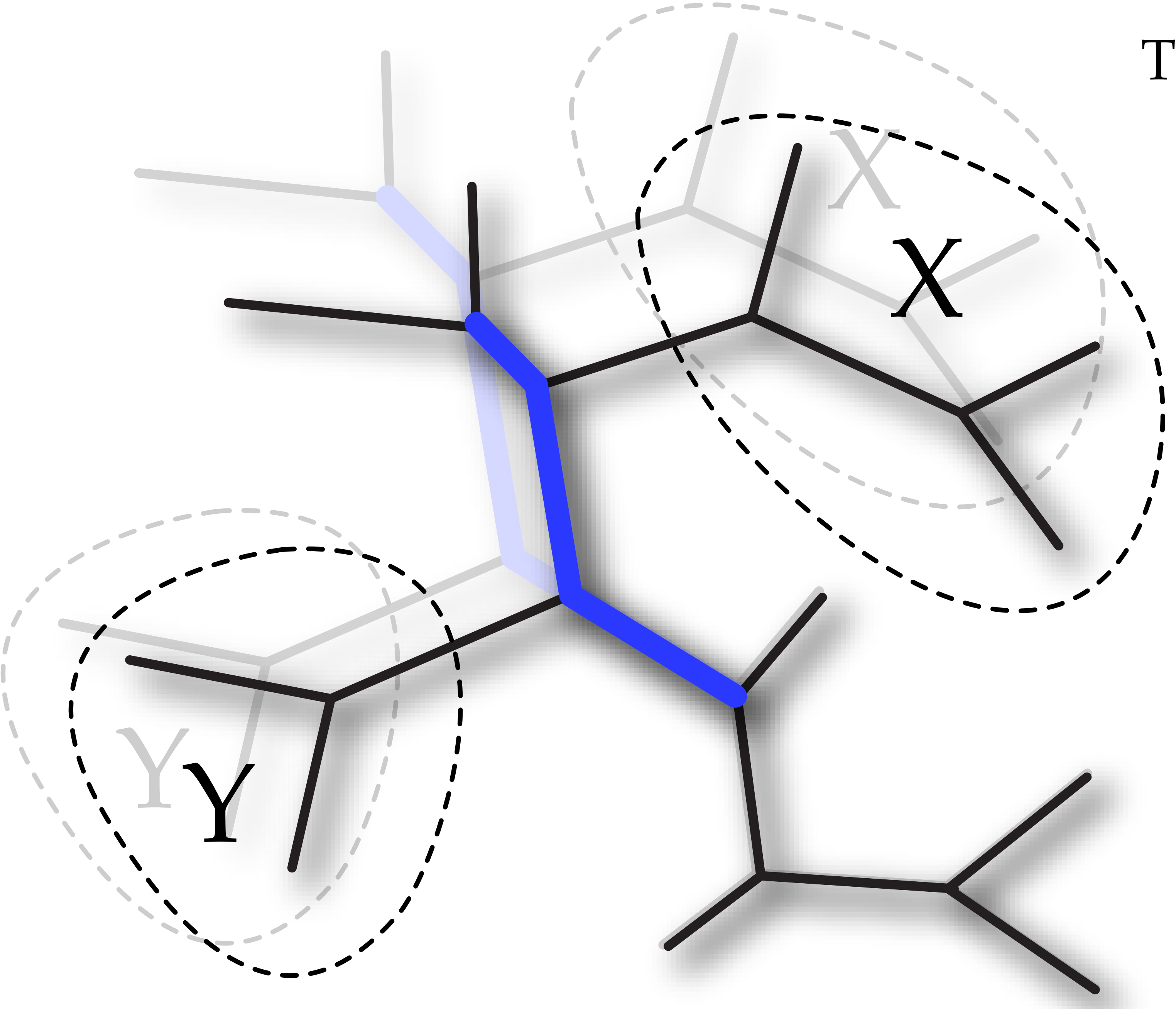


The Larget-Simon move

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: 961-965.

Moving through treespace

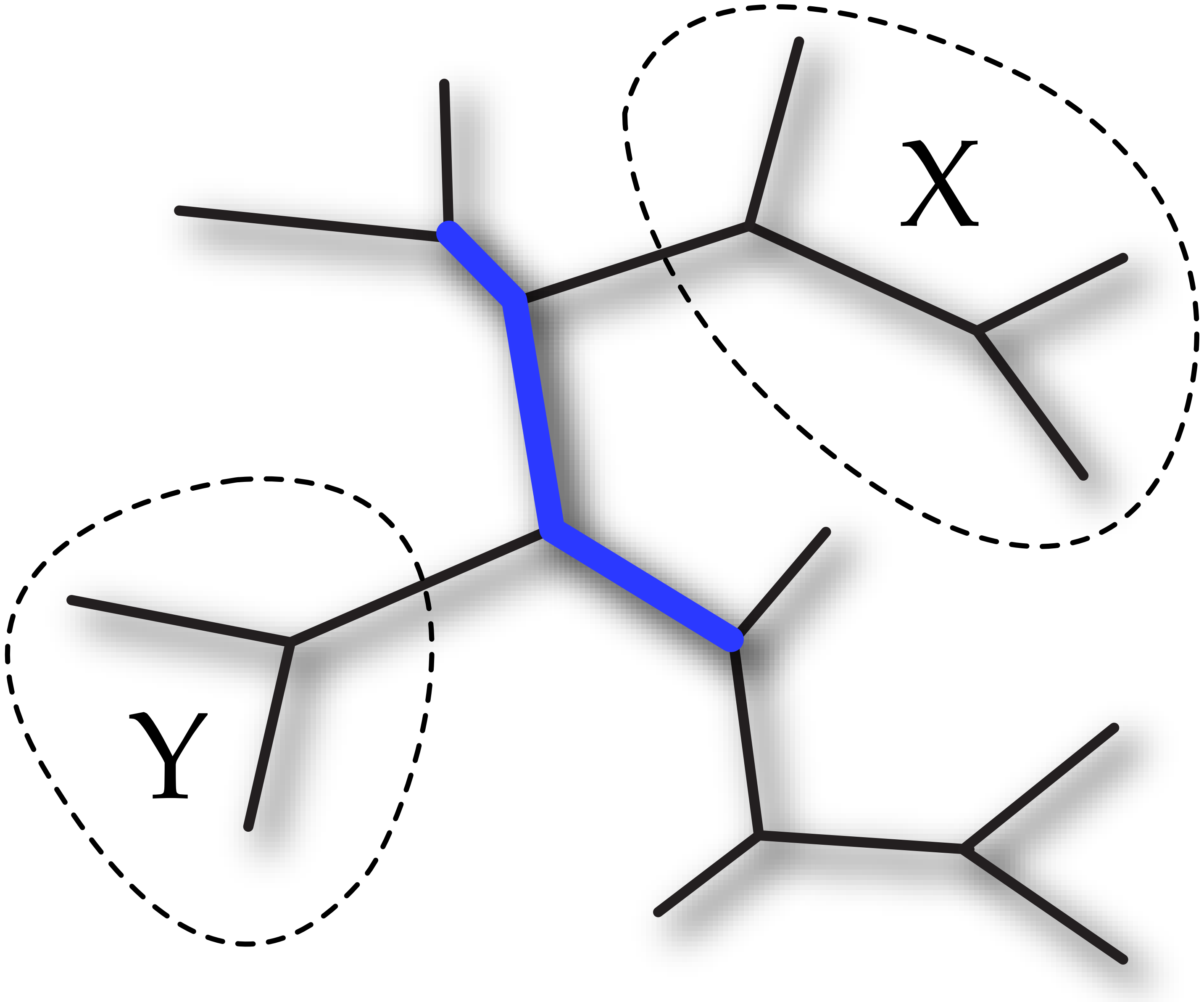


The Larget-Simon move

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

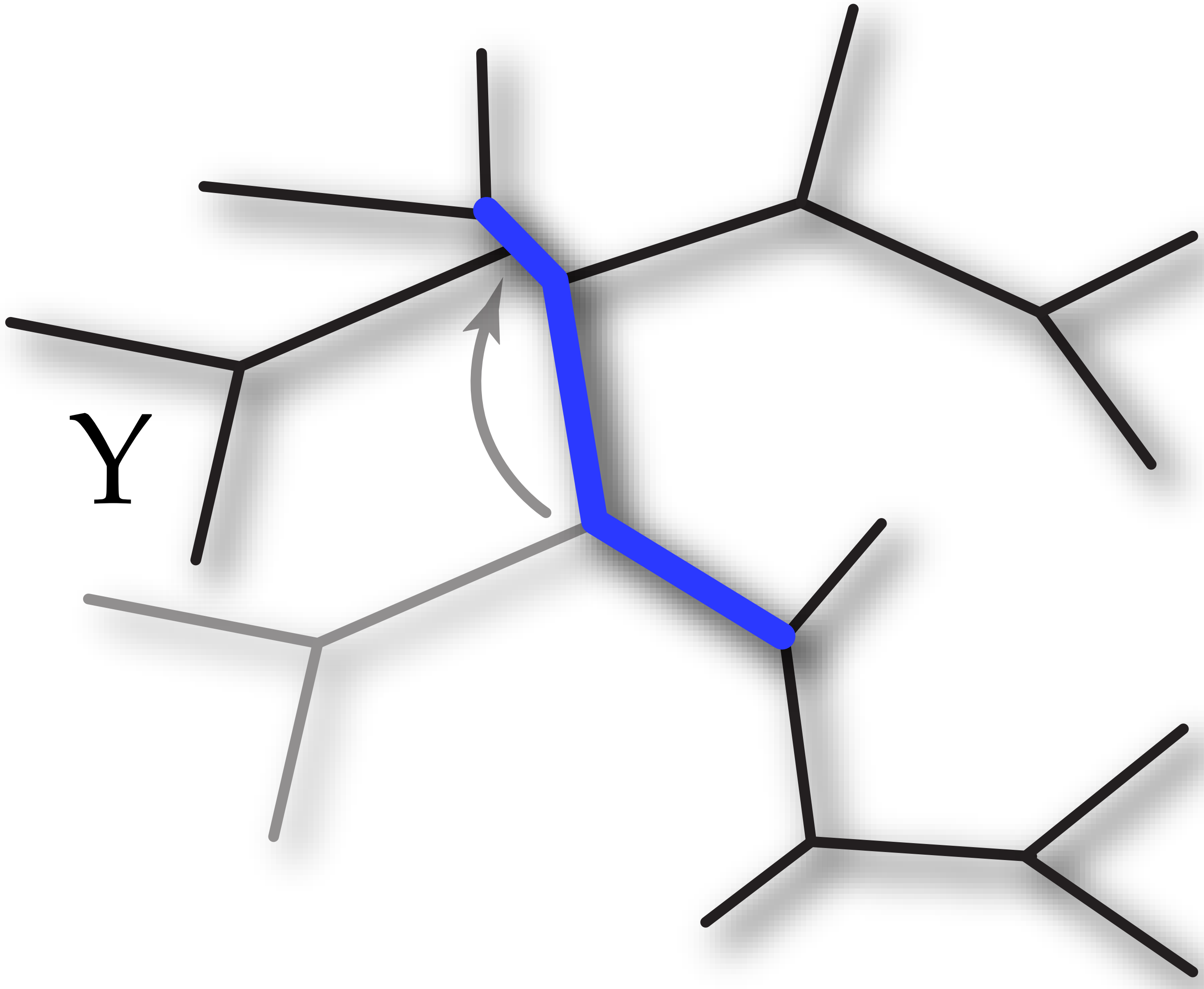


The Larget-Simon move

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

Moving through treespace



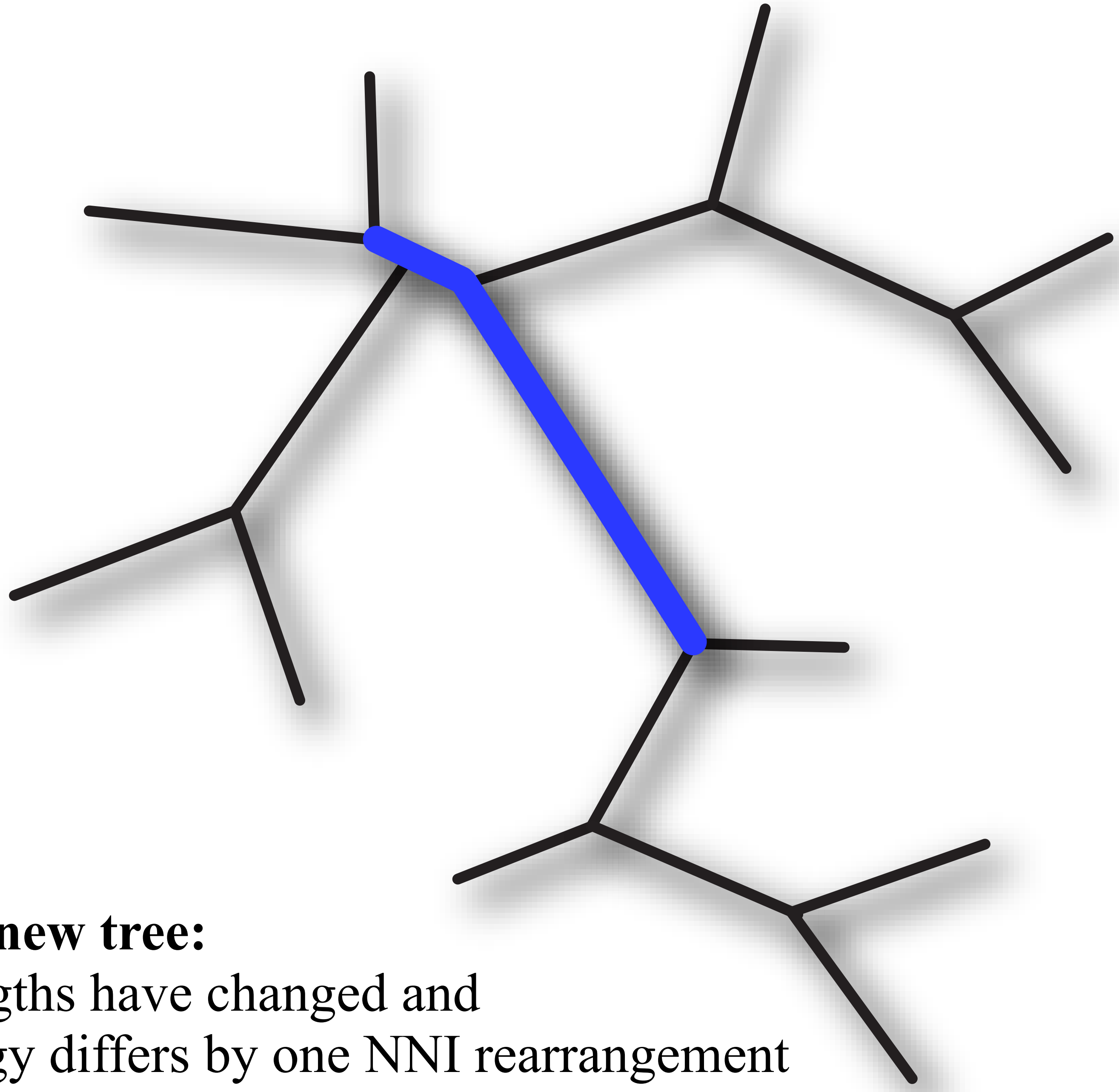
The Large-Simon move

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



The Larget-Simon move

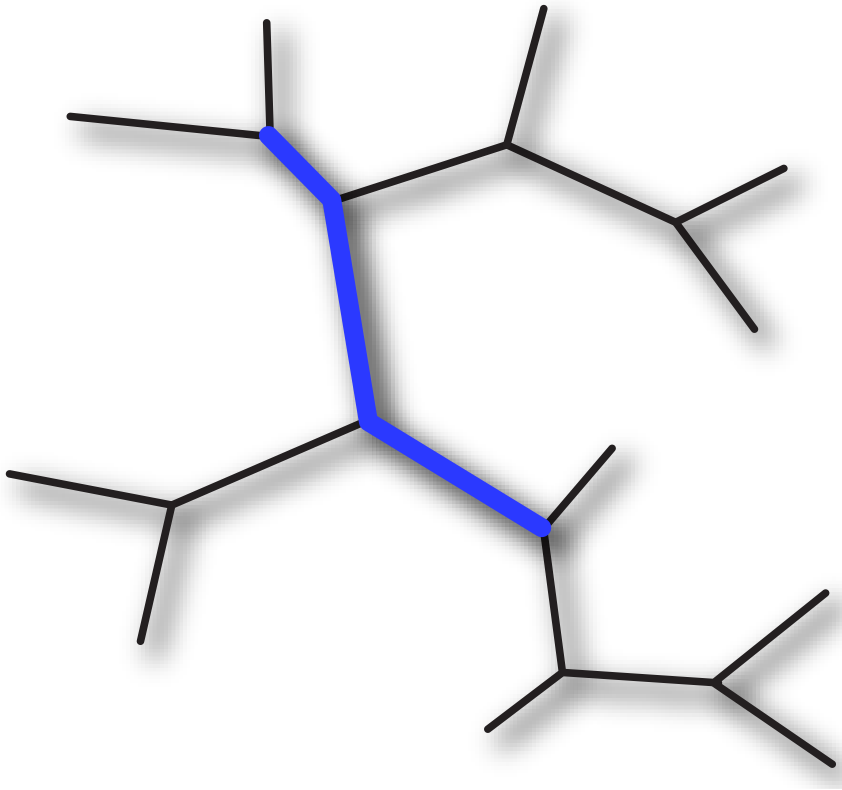
Step 1:
Pick 3 contiguous edges randomly, defining two subtrees, X and Y

Step 2:
Shrink or grow selected 3-edge segment by a random amount

Step 3:
Choose X or Y randomly, then reposition randomly

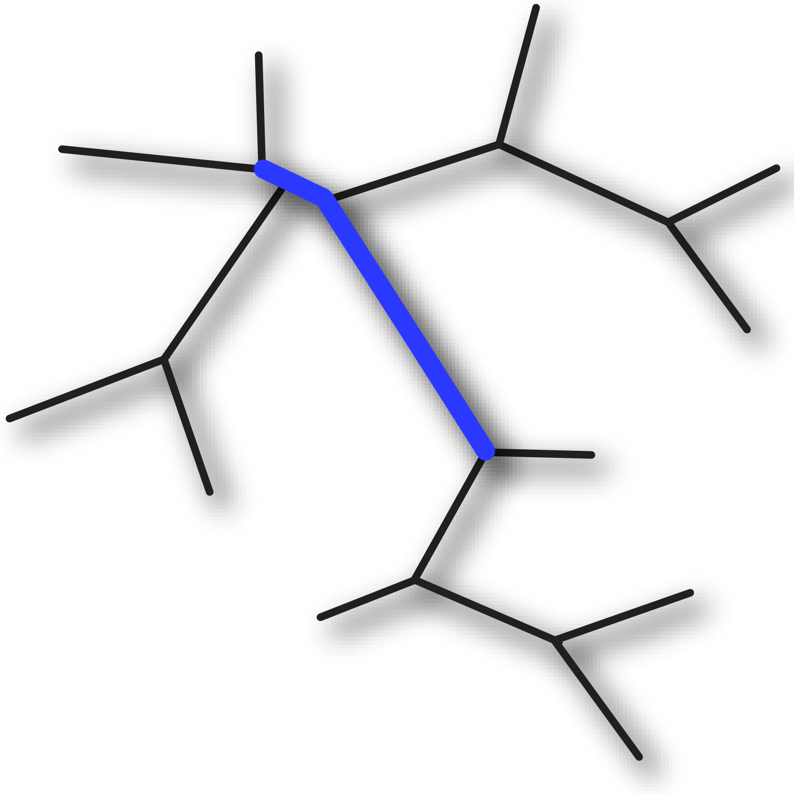
Proposed new tree:
3 edge lengths have changed and
the topology differs by one NNI rearrangement

Moving through treespace



Current tree

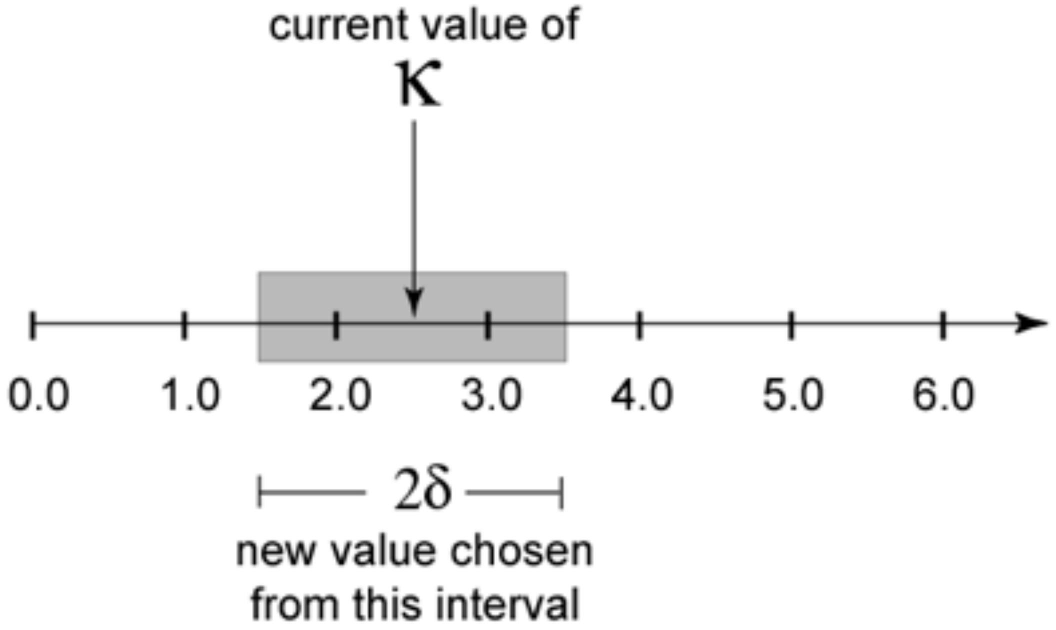
log-posterior = -34256

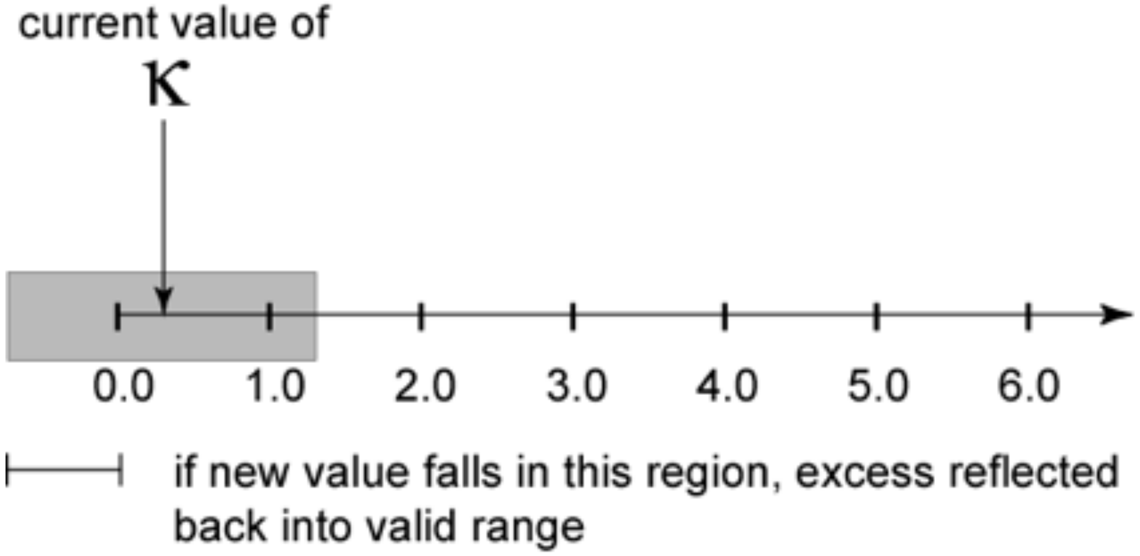


Proposed tree

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

Moving through parameter space




current value of
κ
0.0
1.0
2.0
3.0
4.0
5.0
6.0
if new value falls in this region, excess reflected back into valid range

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 (κ-δ, κ+δ)

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

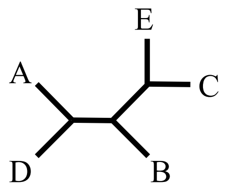
- **Start with** random tree and arbitrary initial values for branch lengths and model parameters
- **Each generation** consists of one of these (chosen at random):
  - Propose a **new tree** (e.g. Larget-Simon move) and either accept or reject the move
  - Propose (and either accept or reject) a **new 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.

IV. Prior distributions

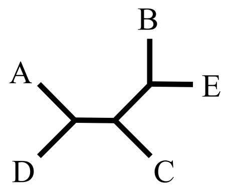
Common Priors

- **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

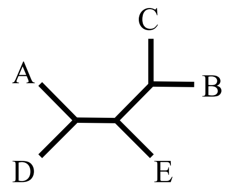
Discrete Uniform distribution for **topologies**



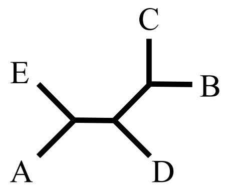
$$\frac{1}{15}$$



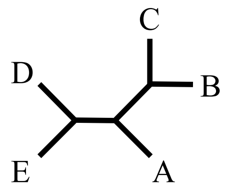
$$\frac{1}{15}$$



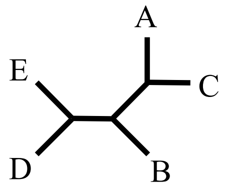
$$\frac{1}{15}$$



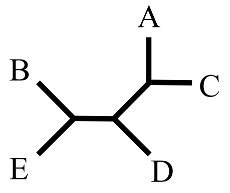
$$\frac{1}{15}$$



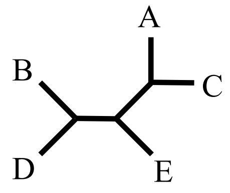
$$\frac{1}{15}$$



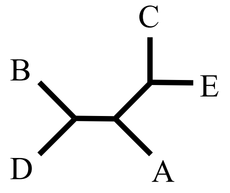
$$\frac{1}{15}$$



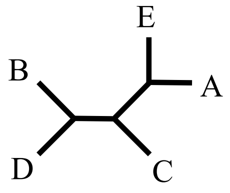
$$\frac{1}{15}$$



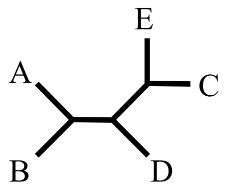
$$\frac{1}{15}$$



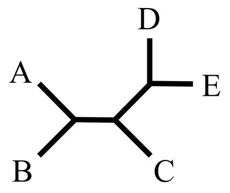
$$\frac{1}{15}$$



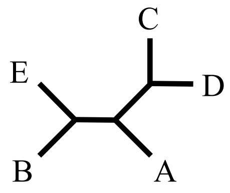
$$\frac{1}{15}$$



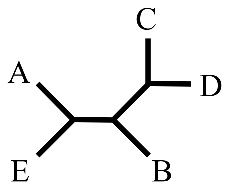
$$\frac{1}{15}$$



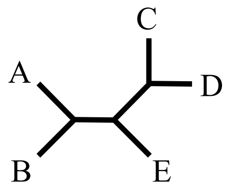
$$\frac{1}{15}$$



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

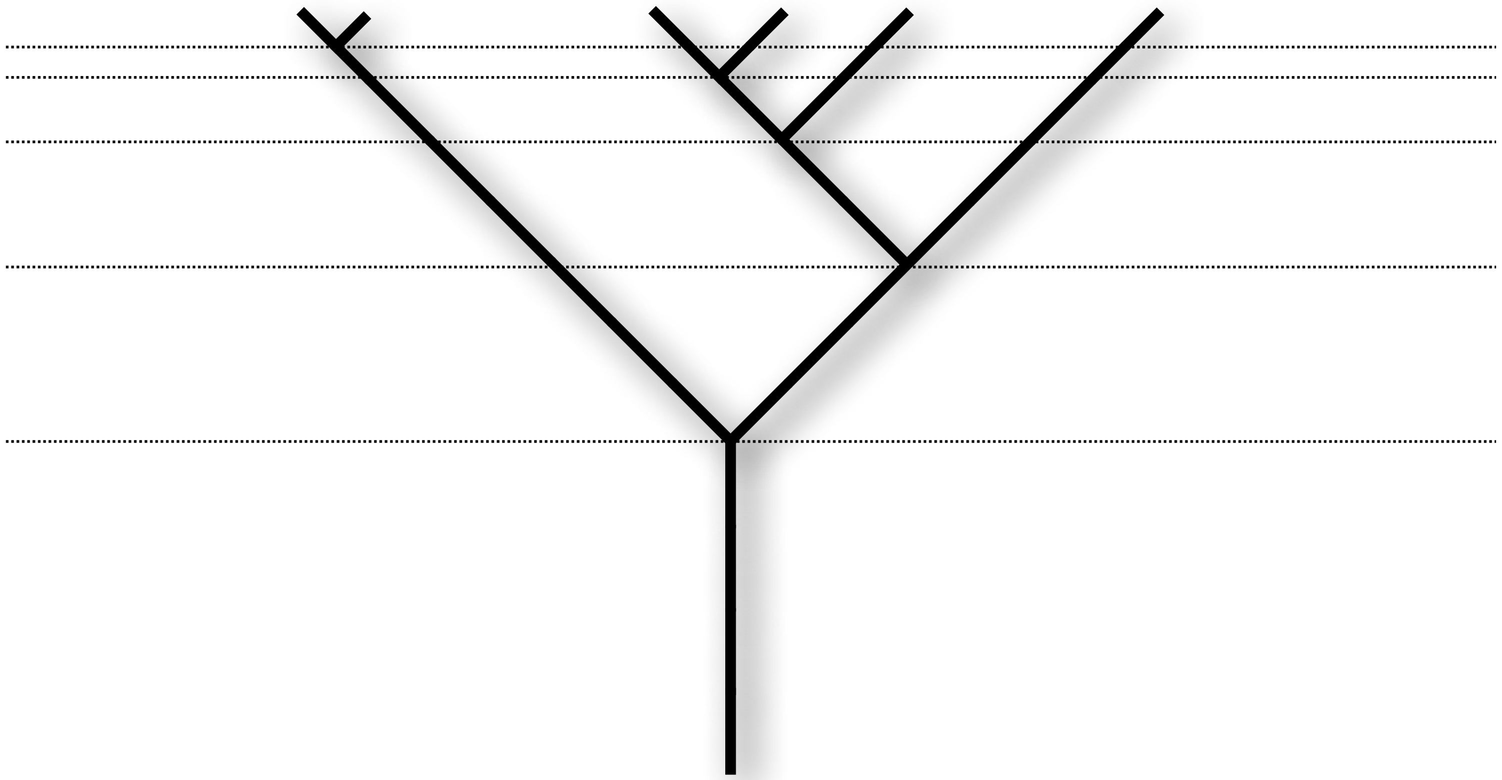


$$\frac{1}{15}$$



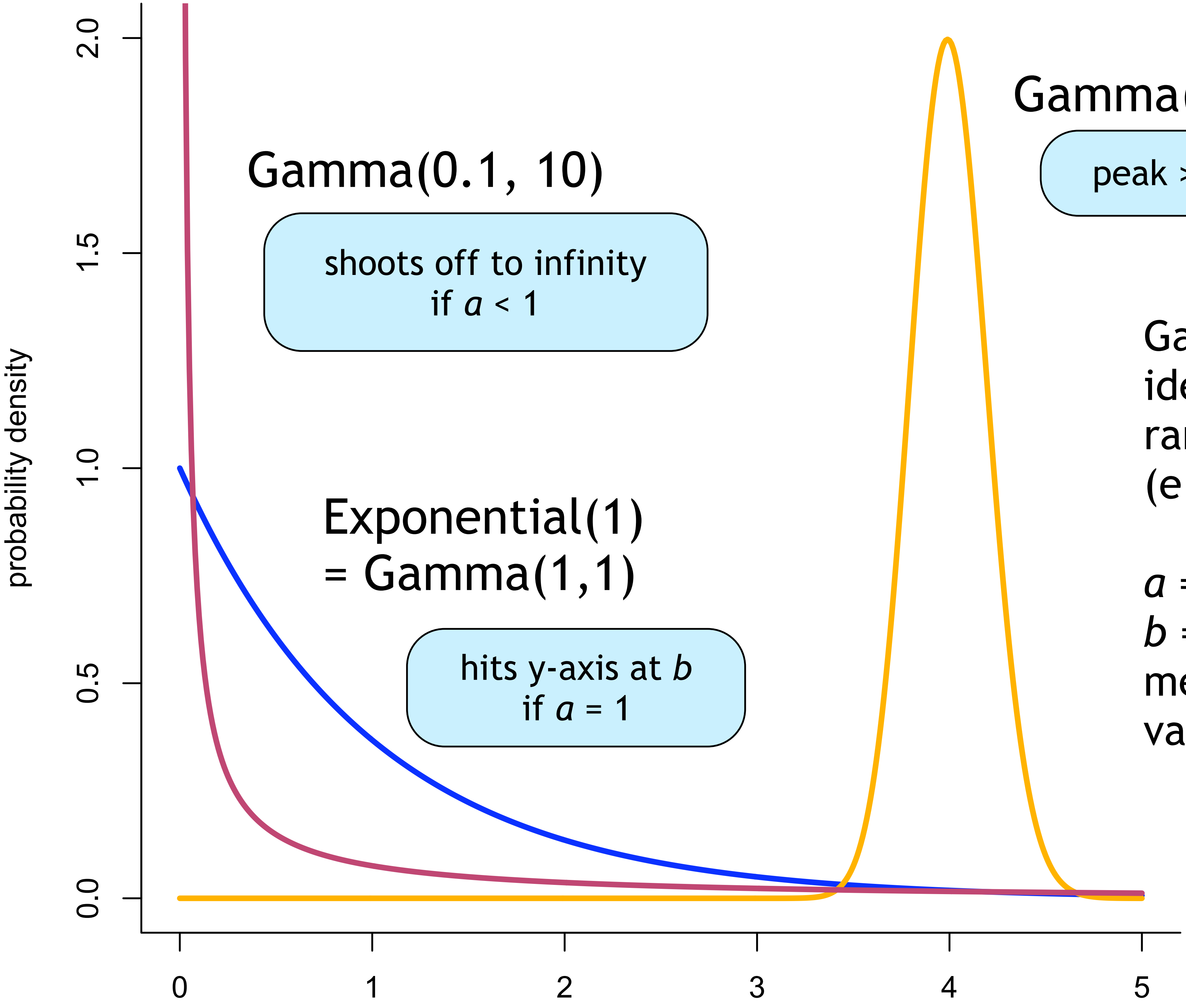
$$\frac{1}{15}$$

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



Gamma(400, 0.01)

peak > 0 if *a* > 1

Gamma(0.1, 10)

shoots off to infinity if *a* < 1

Exponential(1) = Gamma(1,1)

hits y-axis at *b* if *a* = 1

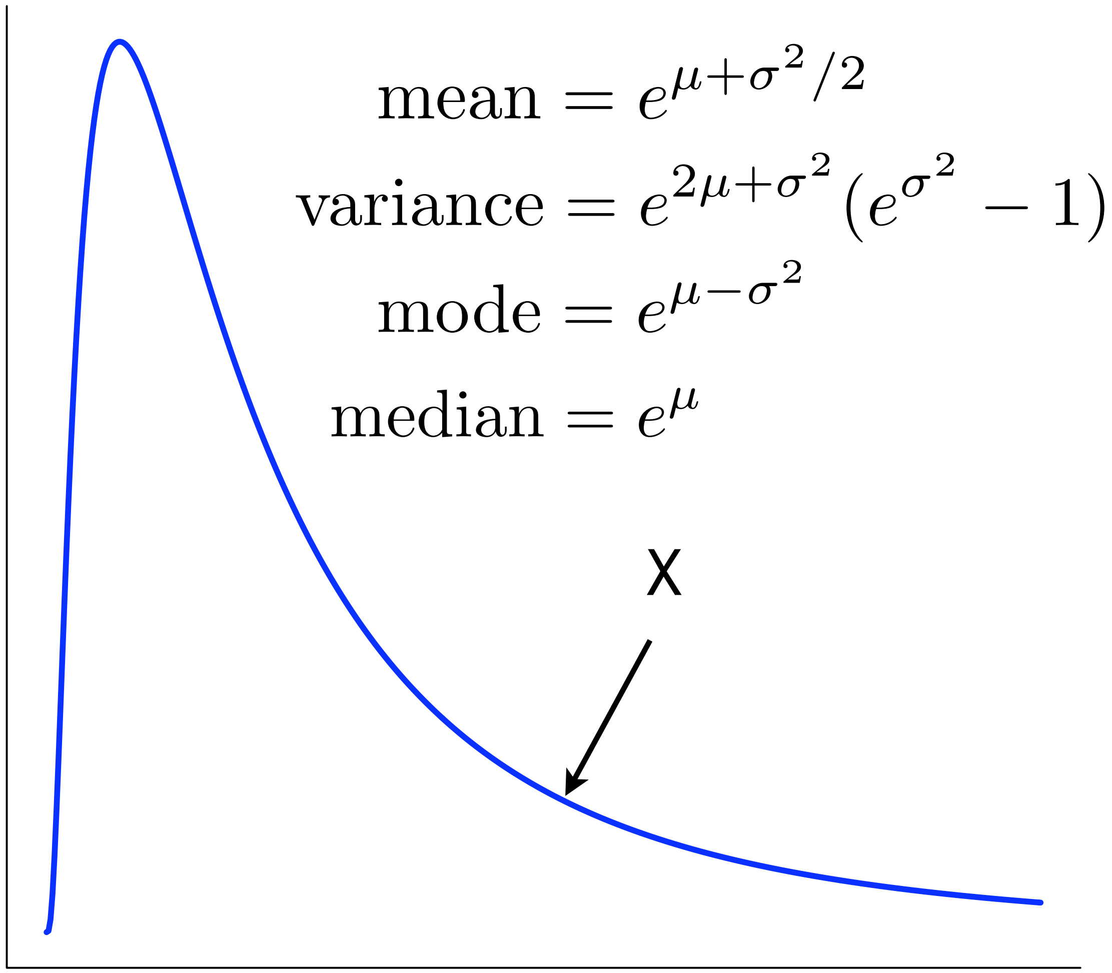
Gamma distributions are ideal for parameters that range from 0 to infinity (e.g. branch lengths)

a = shape
b = scale
mean* = *ab*
variance* = ab^2

*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*…



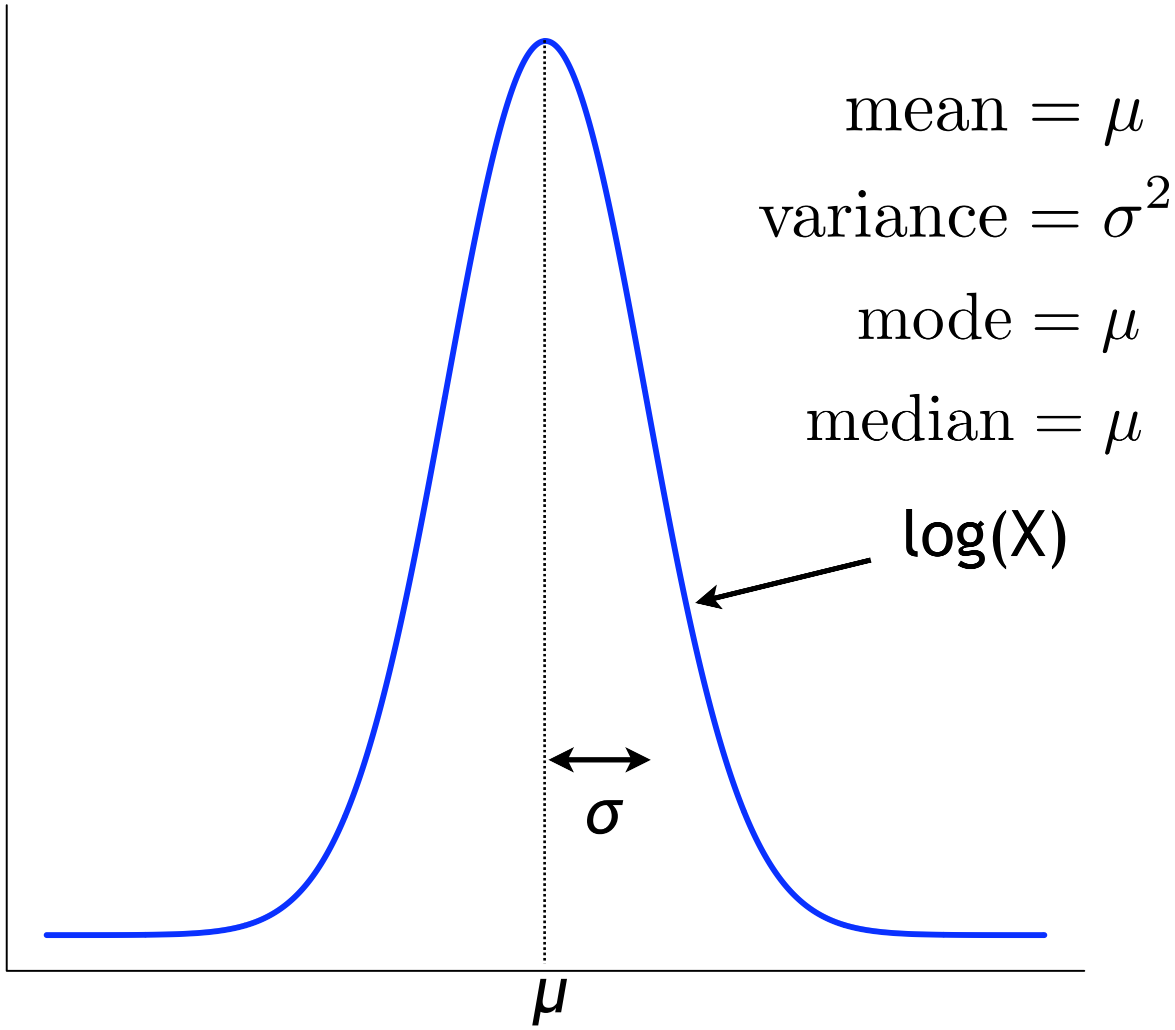
$$\text{mean} = e^{\mu+\sigma^2/2}$$

$$\text{variance} = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$$

$$\text{mode} = e^{\mu-\sigma^2}$$

$$\text{median} = e^{\mu}$$

…then **log(X)** is **normal** with *mean μ* and *standard deviation σ.*



$$\text{mean} = \mu$$

$$\text{variance} = \sigma^2$$

$$\text{mode} = \mu$$

$$\text{median} = \mu$$

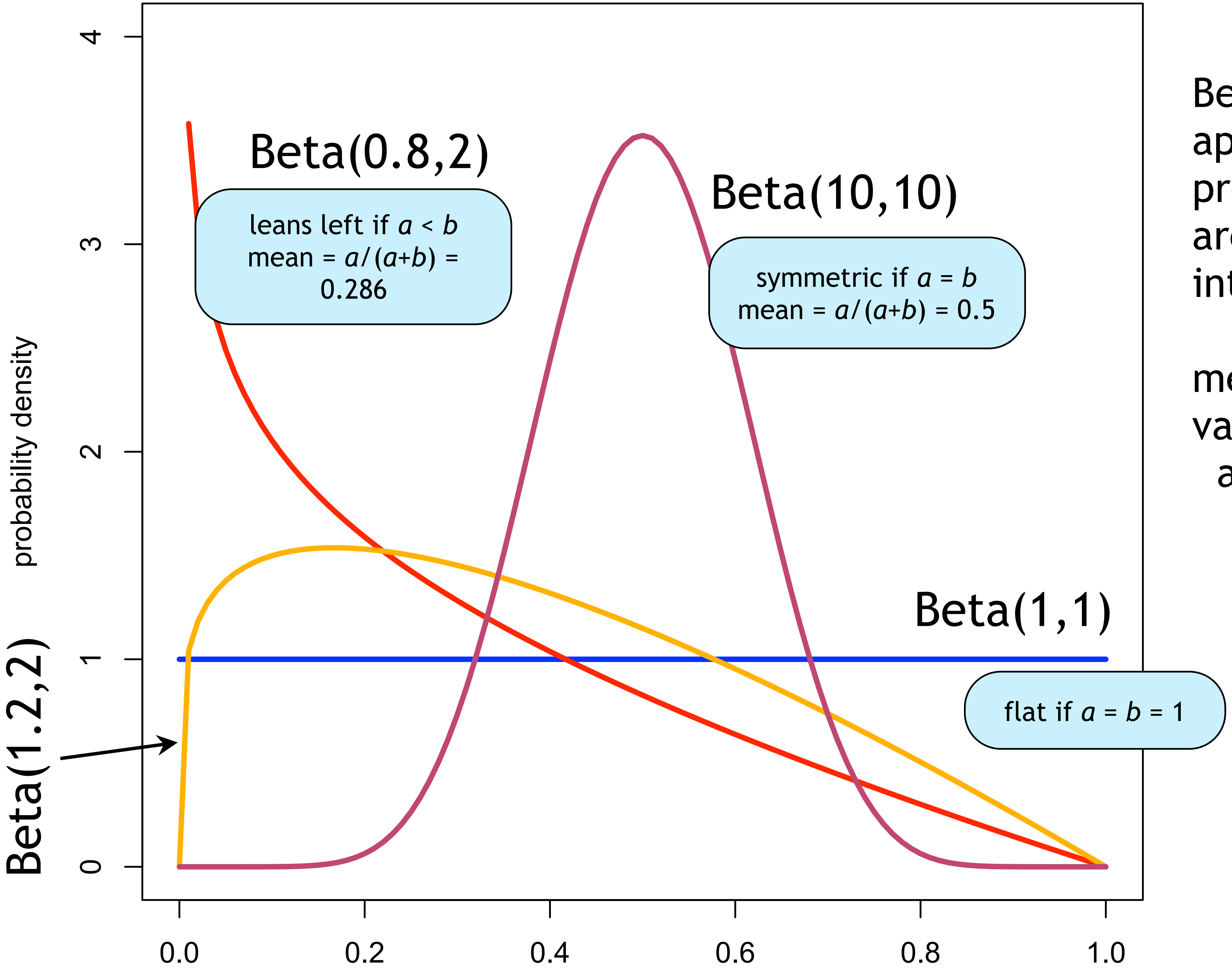
Important: *μ and σ do* ***not*** *represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

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



leans left if $a < b$
mean = $a/(a+b)$ = 0.286

symmetric if $a = b$
mean = $a/(a+b) = 0.5$

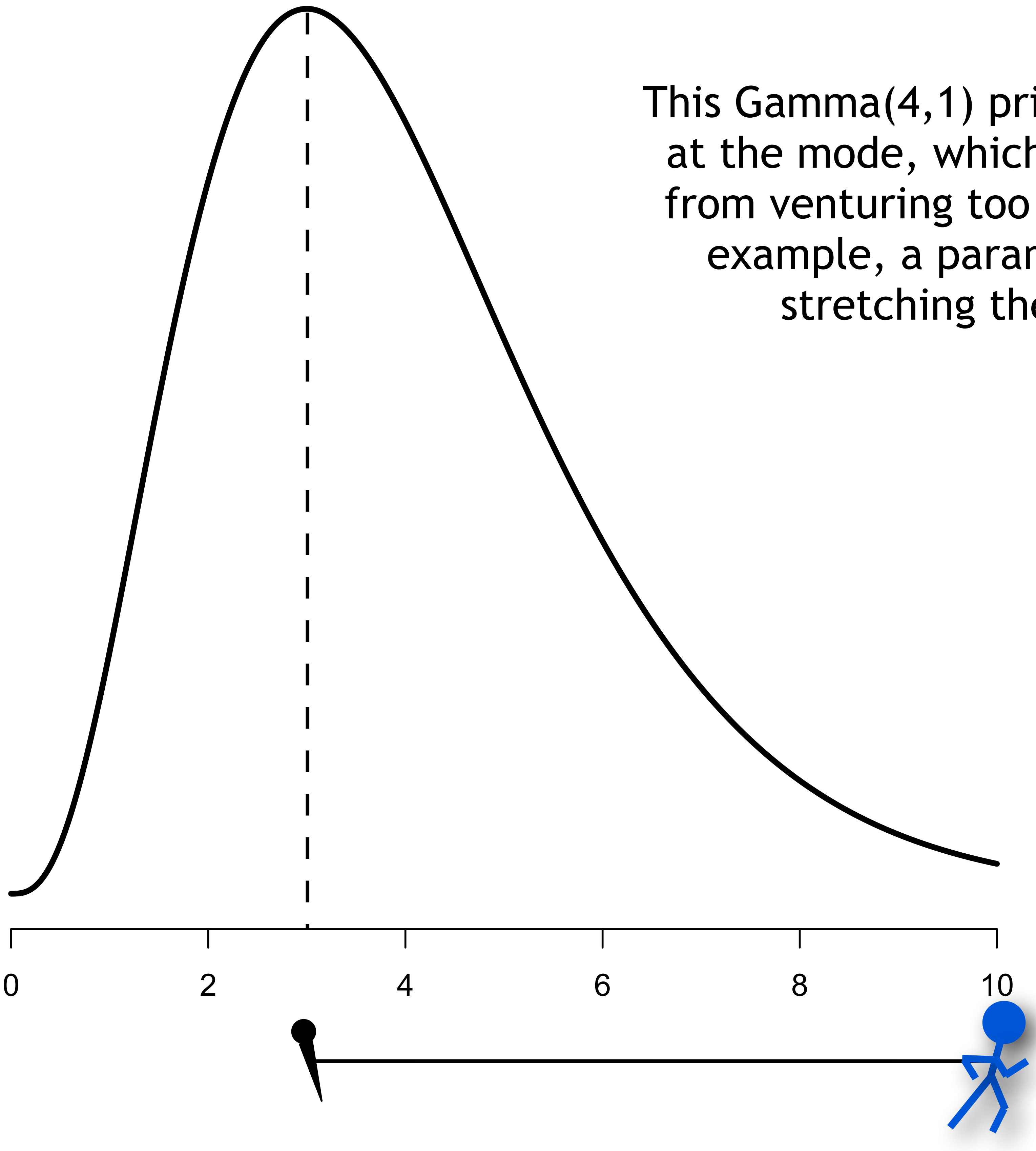
flat if $a = b = 1$

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

mean = a/(a+b)
variance =
ab/[(a+b)^2(a+b+1)]

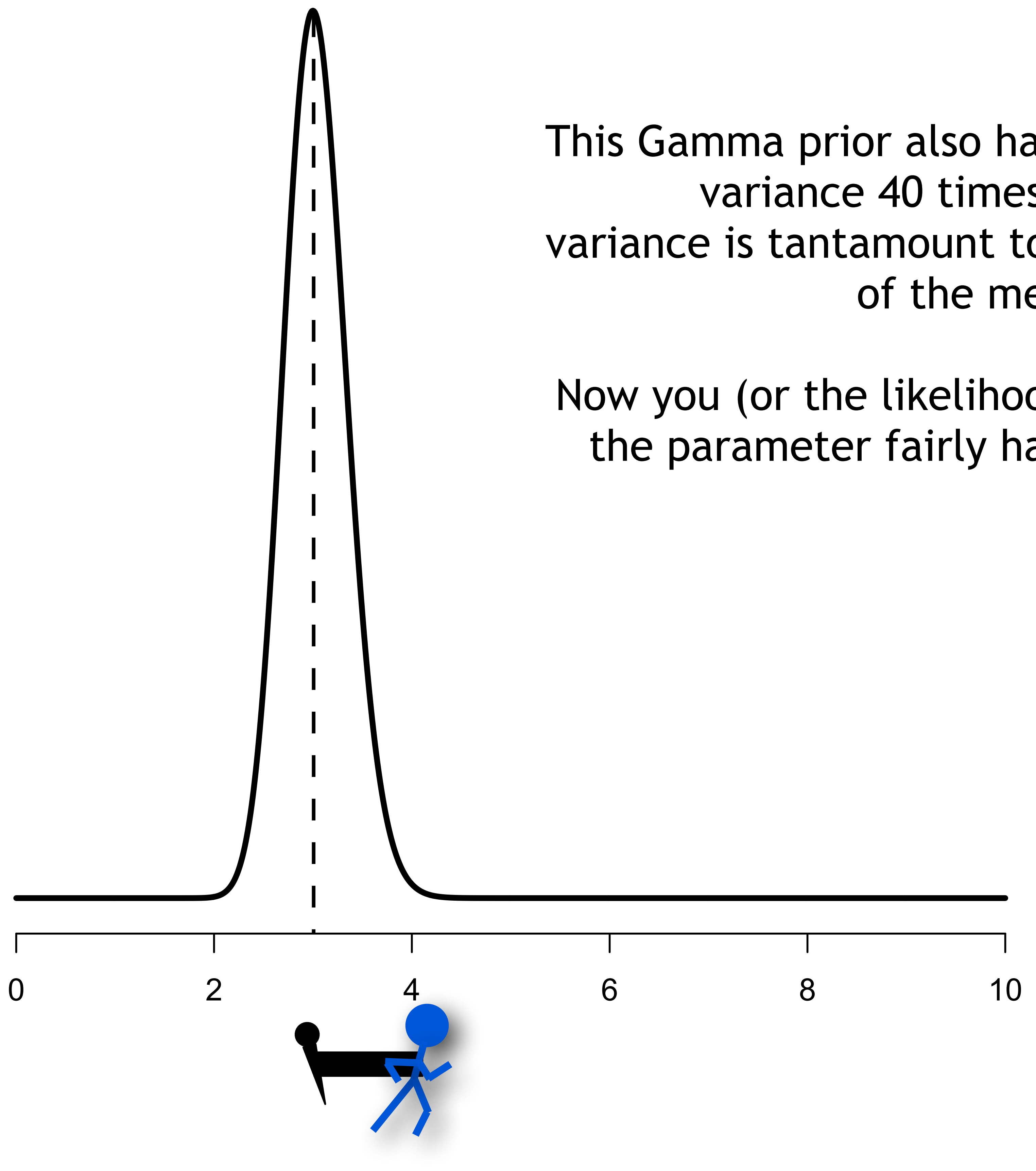
Prior Miscellany

- 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)

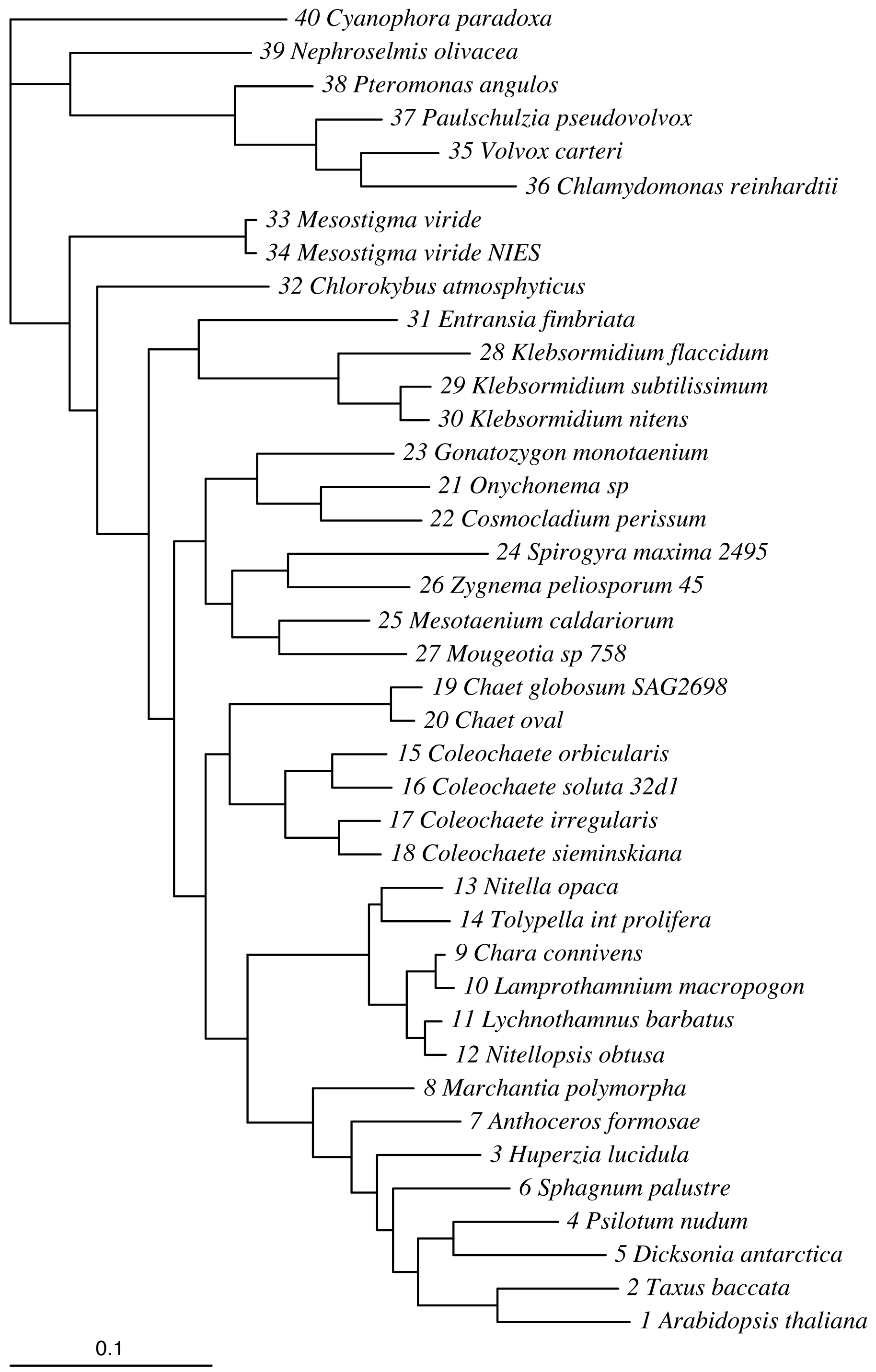


This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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

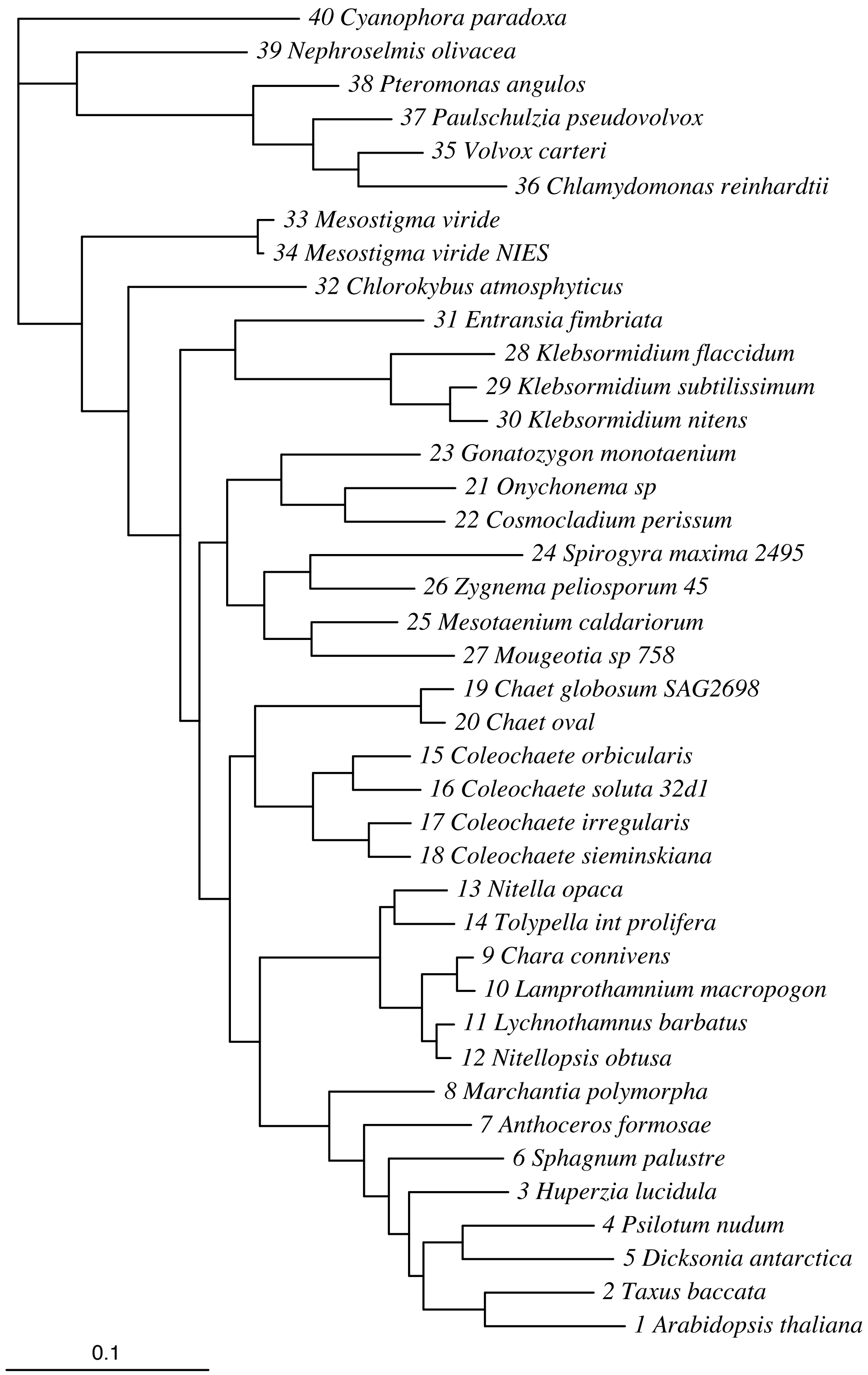


Separate priors applied to internal and external branches

External branch length prior is exponential with mean 0.1

Internal branch length prior is exponential with mean 0.1

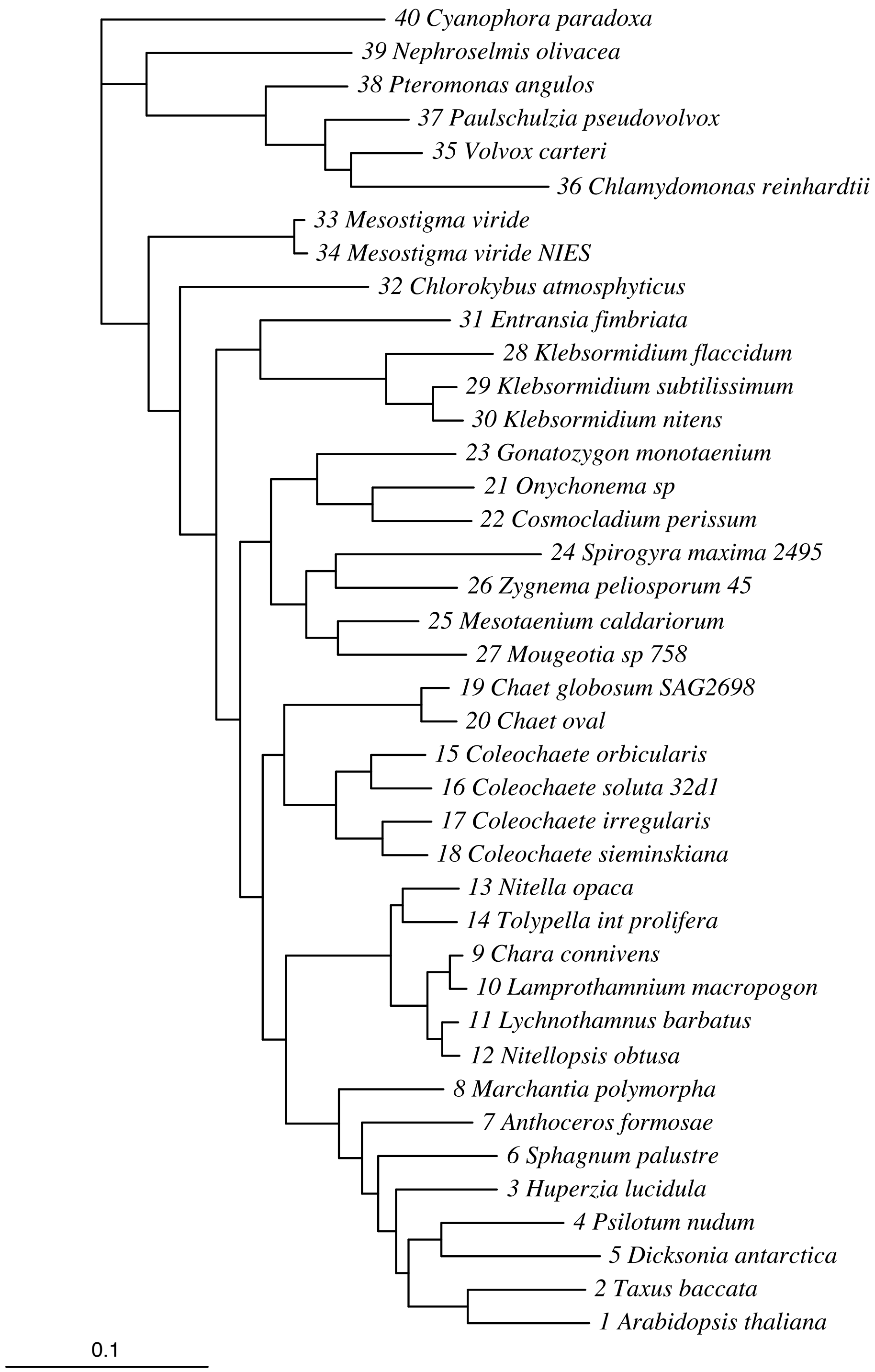
This is a reasonably vague internal branch length prior



Internal branch length prior mean 0.01

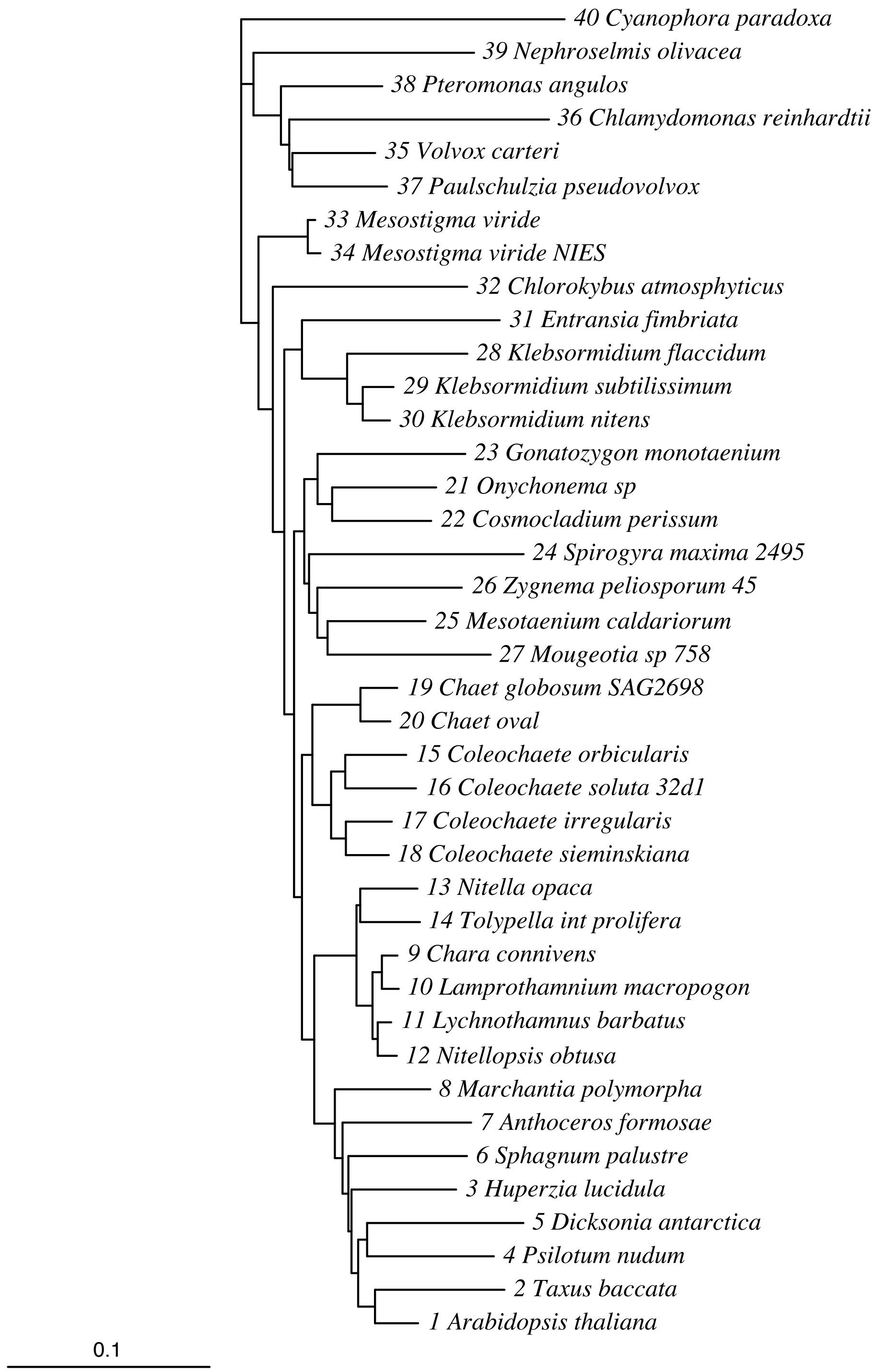
(external branch length prior mean always 0.1)

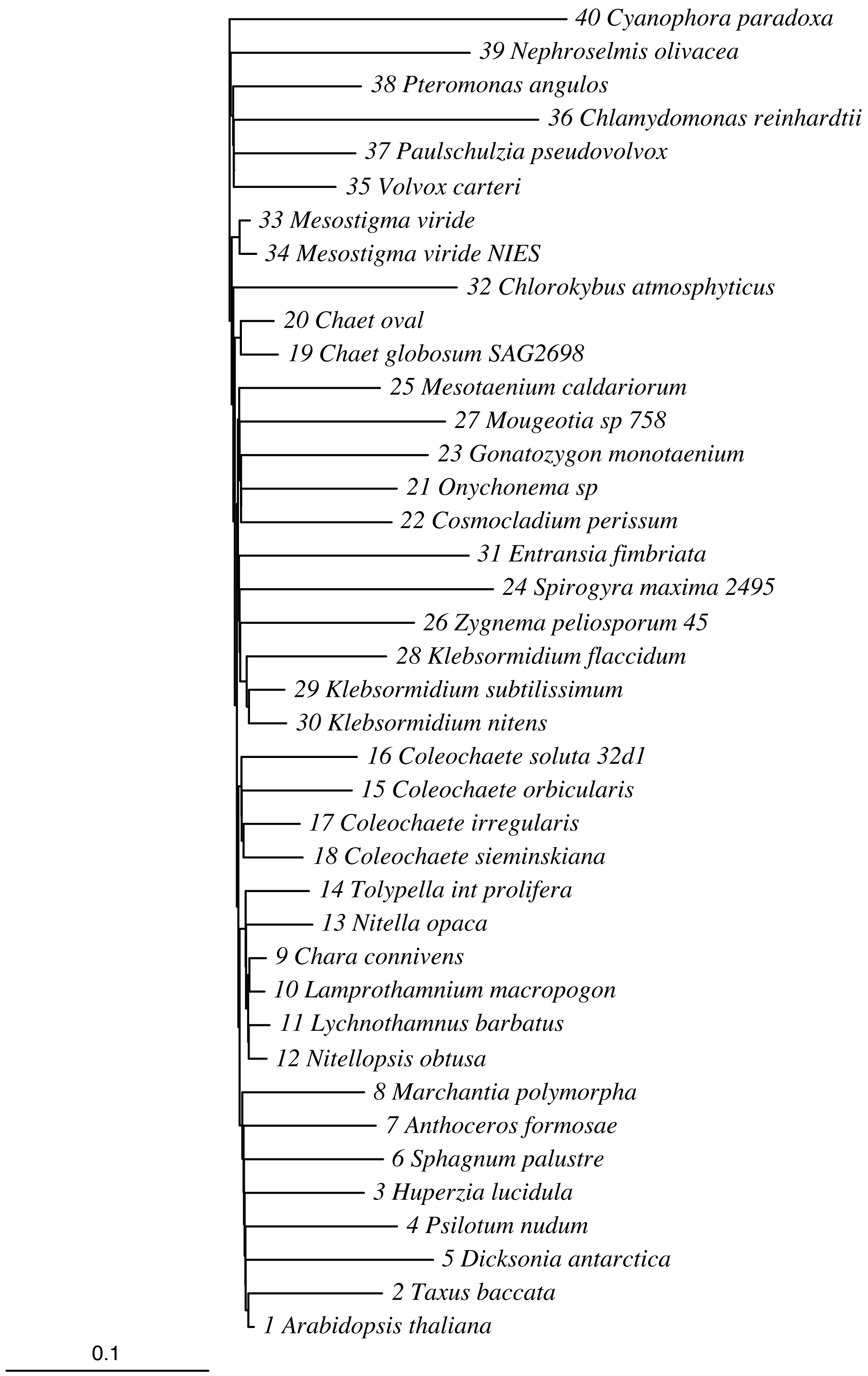
Internal branch length prior mean 0.001



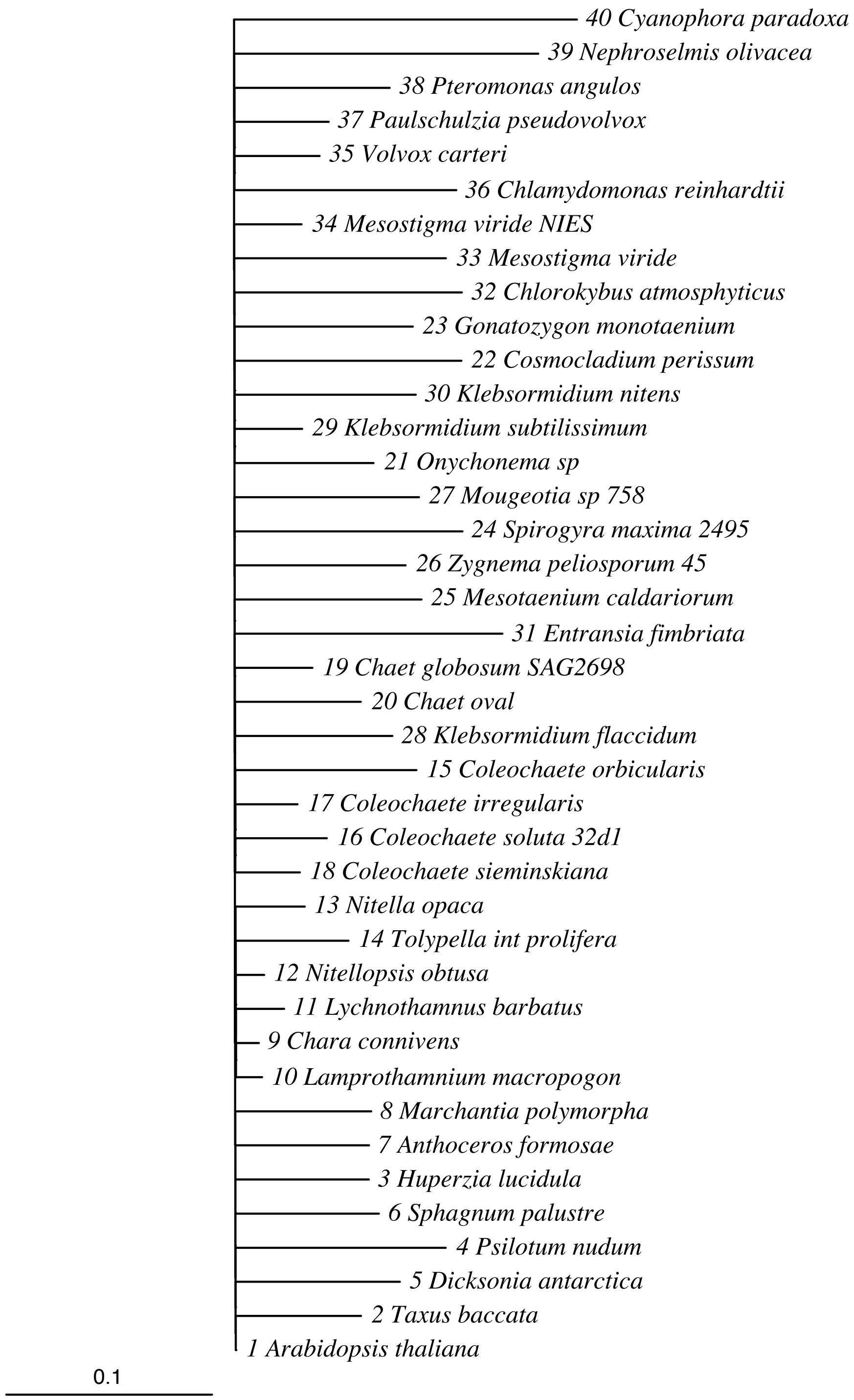
Paul O. Lewis (2015 Woods Hole Molecular Evolution Workshop)

Internal branch length prior mean 0.0001





Internal branch length prior mean 0.00001



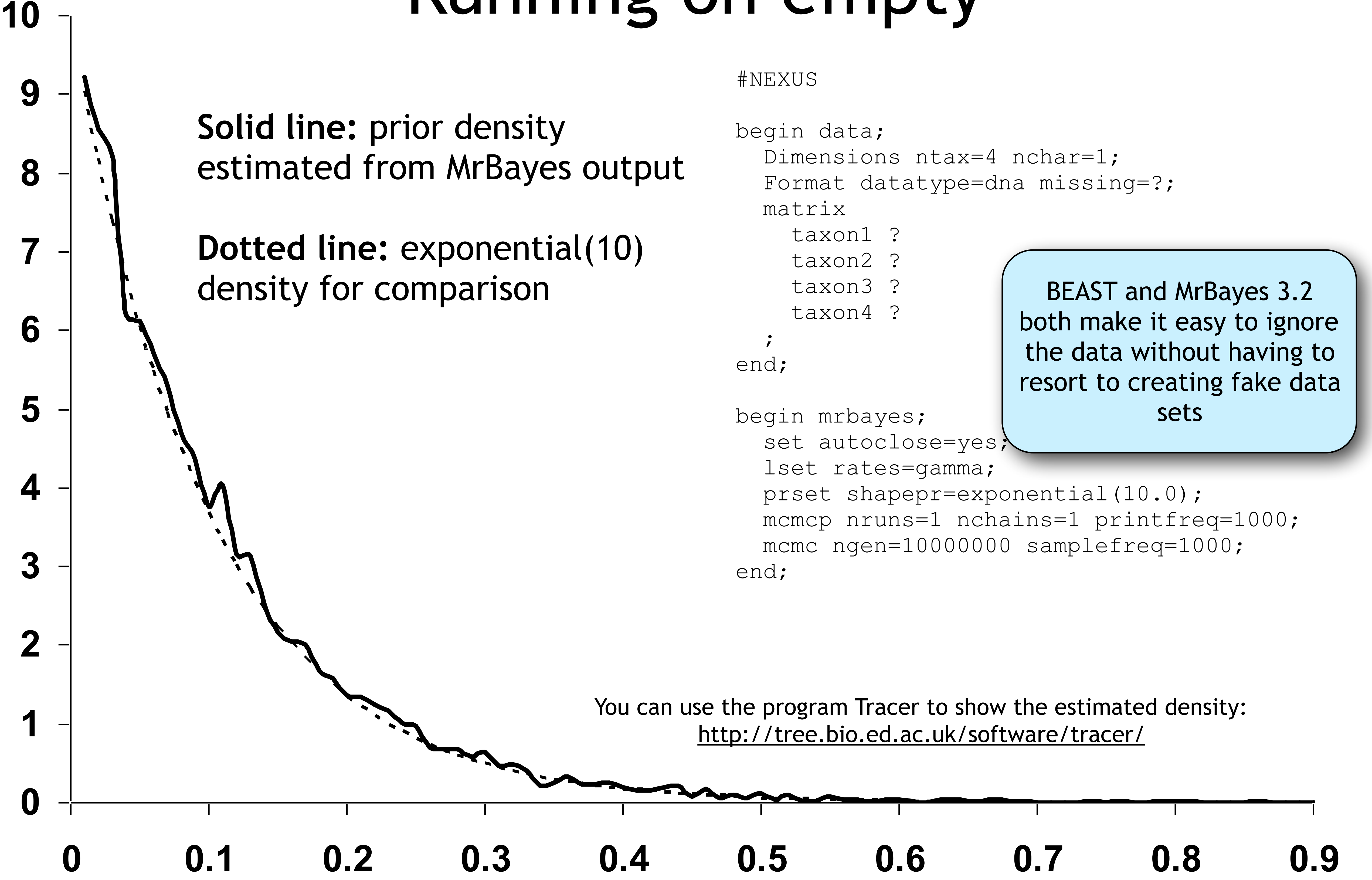
Internal branch length prior mean 0.000001

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



Solid line: prior density estimated from MrBayes output

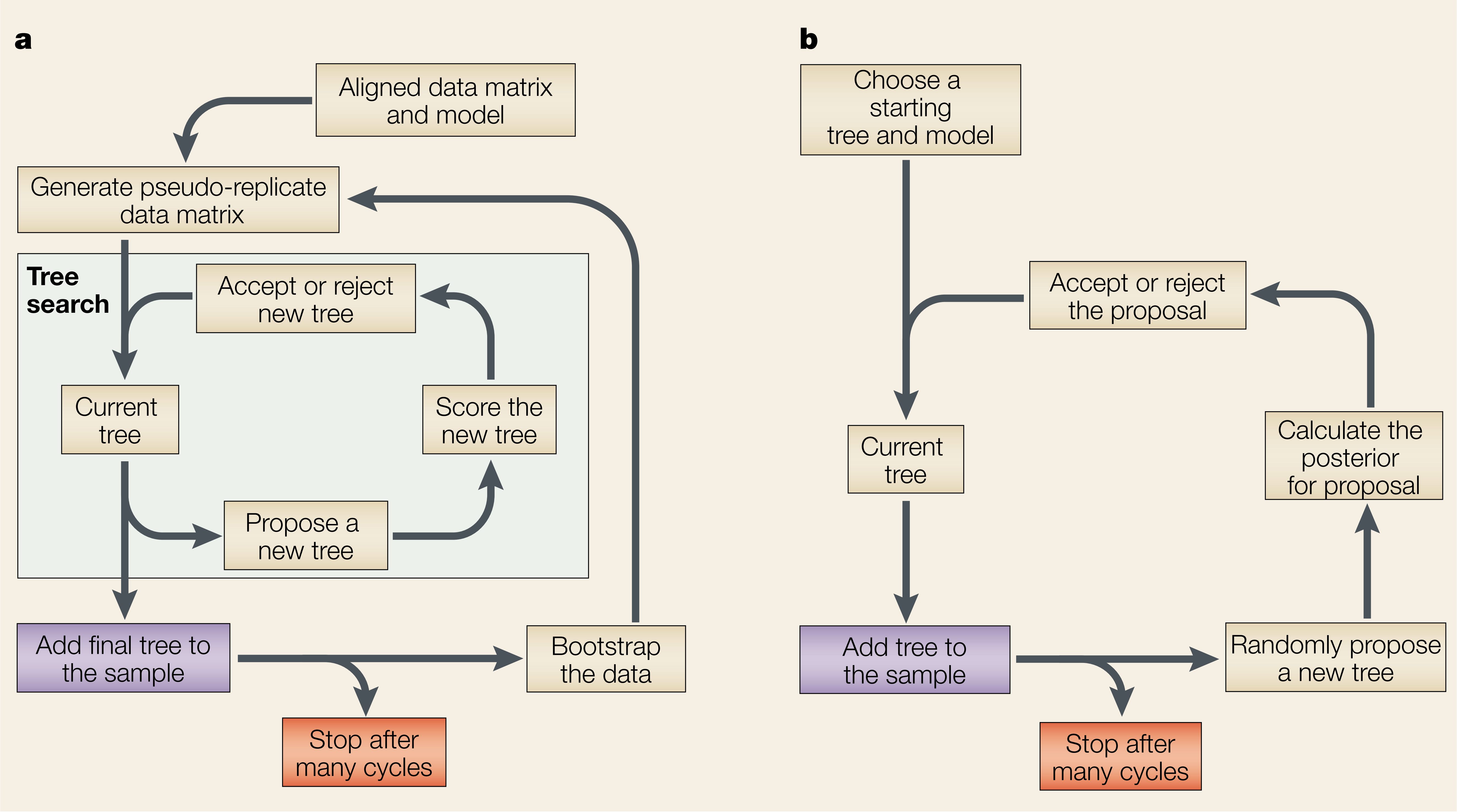
Dotted line: exponential(10) density for comparison

```
#NEXUS

begin data;
  Dimensions ntax=4 nchar=1;
  Format datatype=dna missing=?;
  matrix
    taxon1 ?
    taxon2 ?
    taxon3 ?
    taxon4 ?
  ;
end;

begin mrbayes;
  set autoclose=yes;
  lset rates=gamma;
  prset shapepr=exponential(10.0);
  mcmcp nruns=1 nchains=1 printfreq=1000;
  mcmc ngen=10000000 samplefreq=1000;
end;
```

BEAST and MrBayes 3.2 both make it easy to ignore the data without having to resort to creating fake data sets

You can use the program Tracer to show the estimated density:
http://tree.bio.ed.ac.uk/software/tracer/

Maximum Likelihood Bayesian


a
Aligned data matrix and model
Generate pseudo-replicate data matrix
Tree search
Accept or reject new tree
Current tree
Score the new tree
Propose a new tree
Add final tree to the sample
Bootstrap the data
Stop after many cycles
b
Choose a starting tree and model
Accept or reject the proposal
Current tree
Calculate the posterior for proposal
Add tree to the sample
Randomly propose a new tree
Stop after many cycles

(Holder and Lewis 2003)