

See also 20 & 27 June 2018 at
<http://phyloseminar.org/recorded.html>

Bayesian Phylogenetics

Workshop on Molecular Evolution
Woods Hole, Massachusetts

27 May 2024

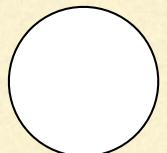
Paul O. Lewis
Department of Ecology & Evolutionary Biology



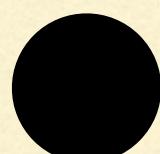
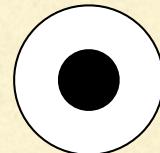
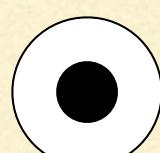
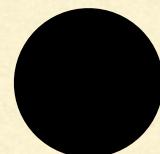
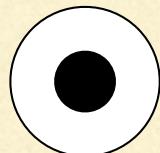
Bayesian inference

Joint probabilities

White,Solid



White,Dotted



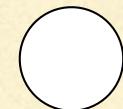
Black,Dotted

Black,Solid

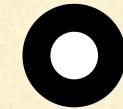
10 marbles in a bag
Sampling with replacement



$\Pr(B,S) = 0.4$



$\Pr(W,S) = 0.1$

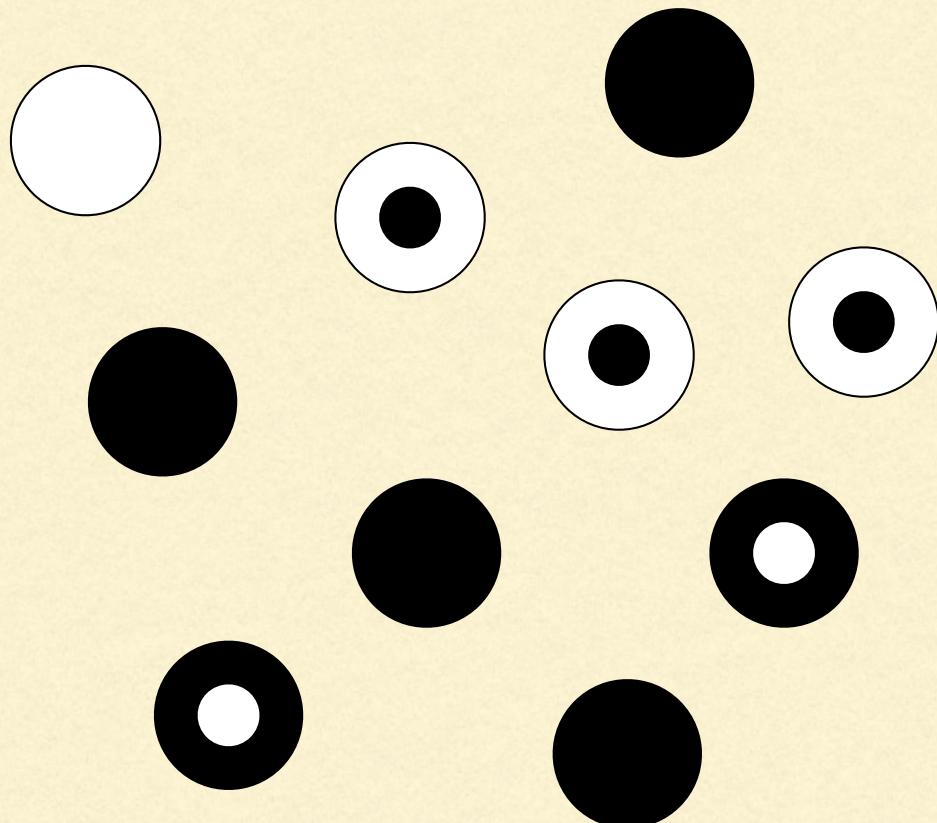


$\Pr(B,D) = 0.2$



$\Pr(W,D) = 0.3$

Conditional probabilities



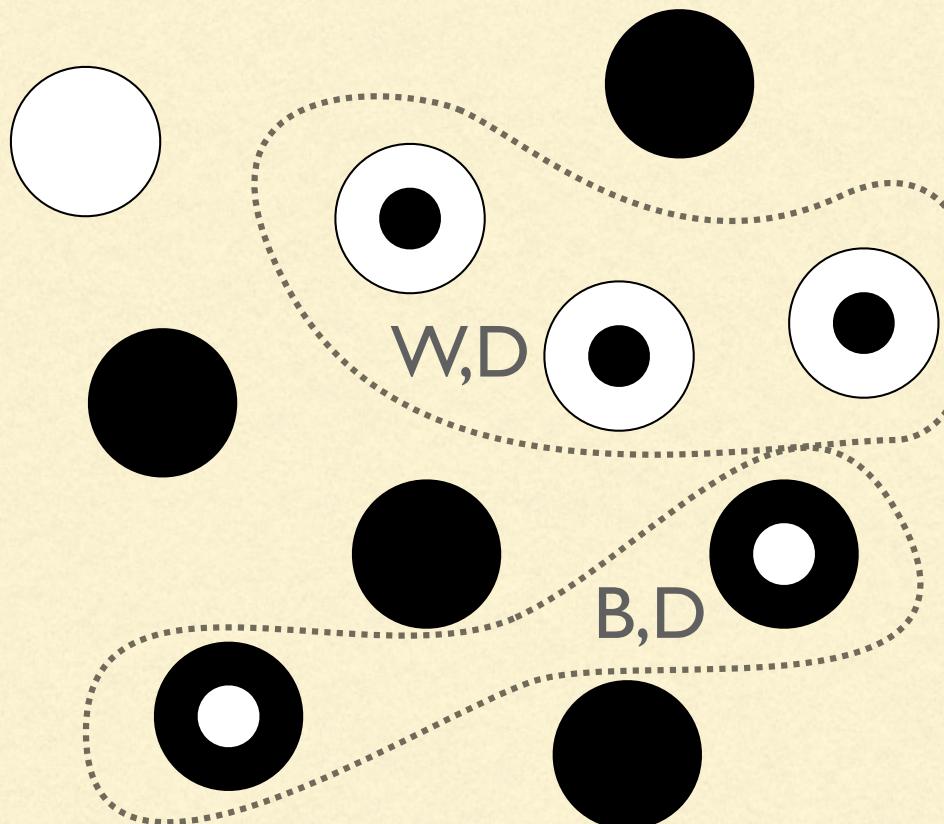
What's the probability that a marble is black given that it is dotted?

5 marbles satisfy the condition (D)

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

2 remaining marbles are black (B)

Marginal probabilities



Marginalizing over color yields the total probability that a marble is dotted (D)

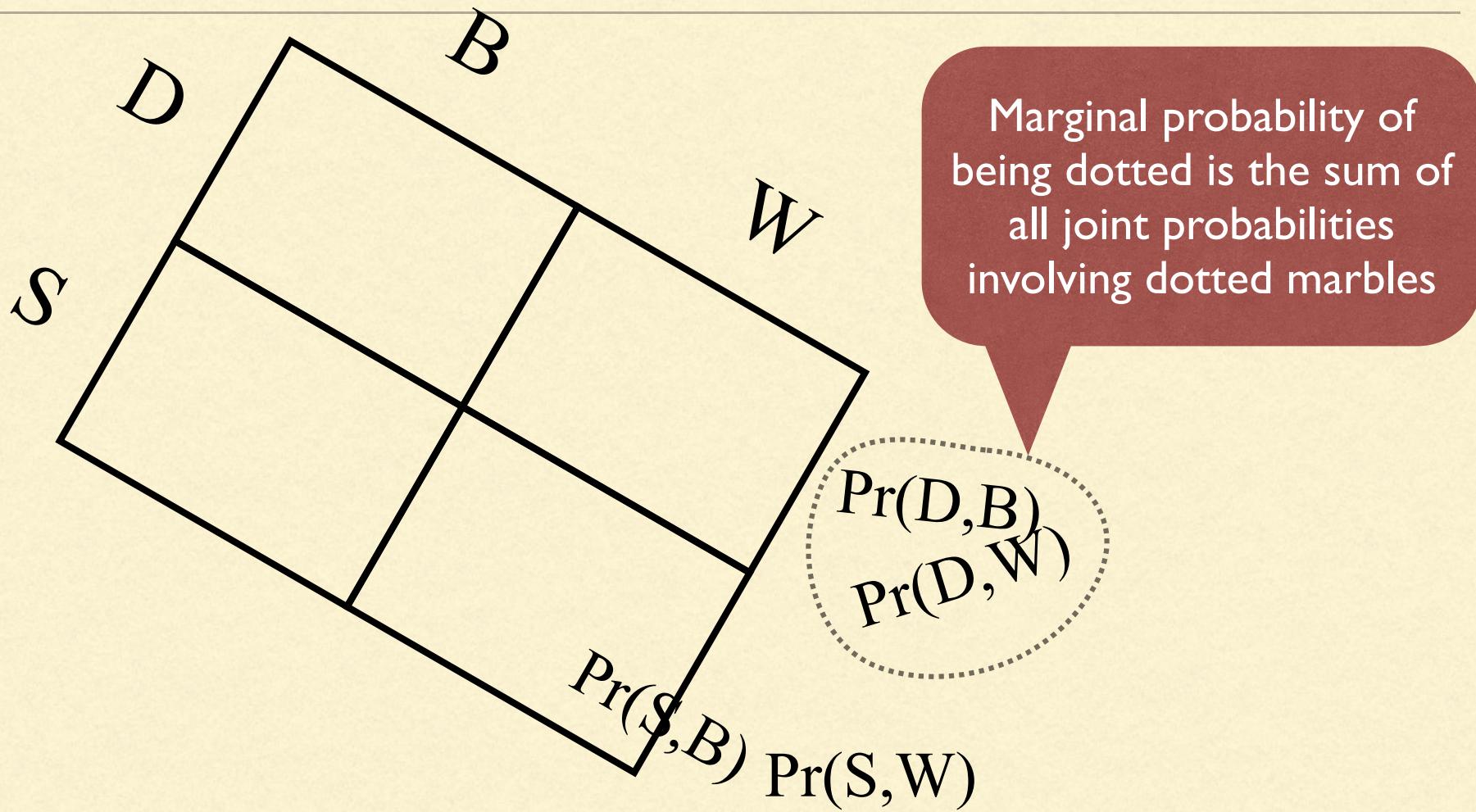
$$\begin{aligned}\Pr(\mathbf{D}) &= \Pr(\mathbf{B}, \mathbf{D}) + \Pr(\mathbf{W}, \mathbf{D}) \\ &= 0.2 + 0.3 \\ &= 0.5\end{aligned}$$

Marginalization involves summing all joint probabilities containing D

Marginalization

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

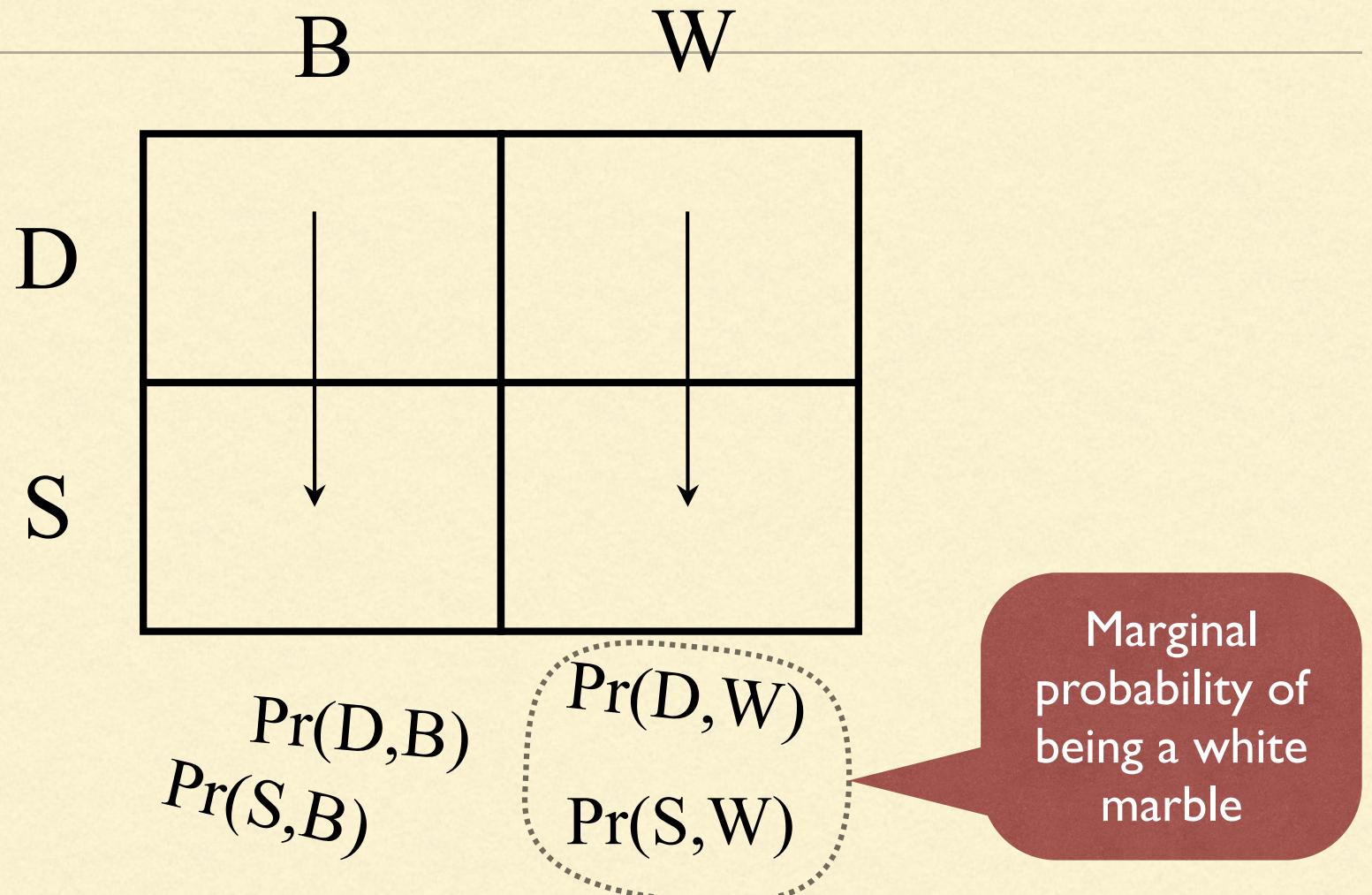
Marginalizing over colors



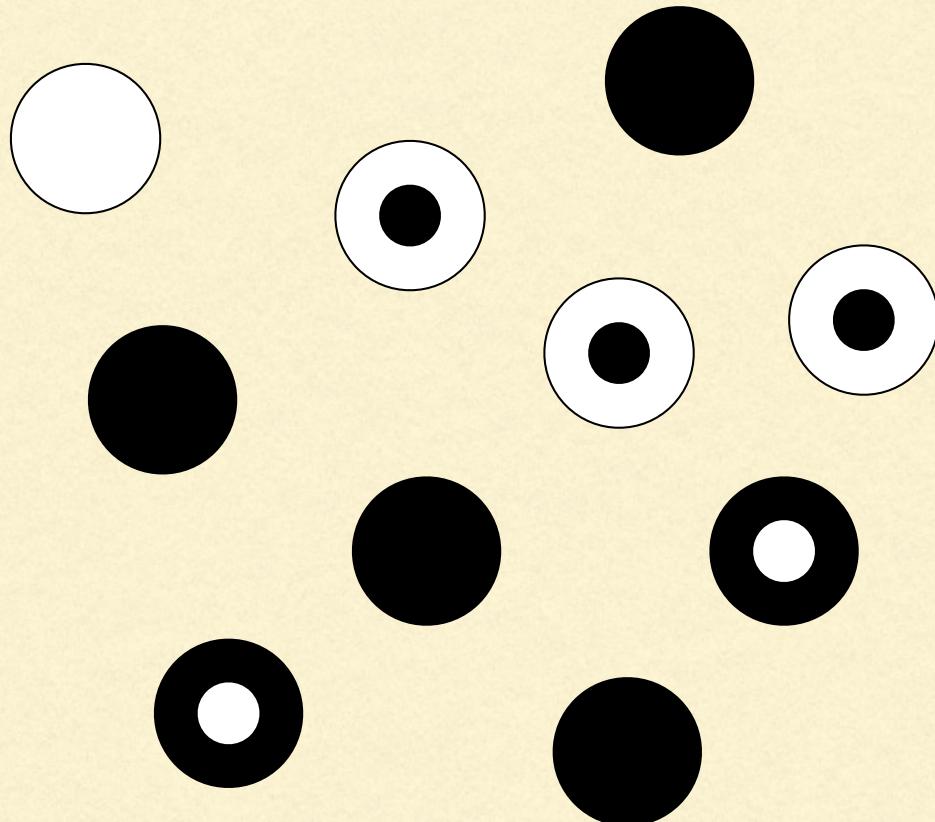
Joint probabilities

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

Marginalizing over "dottedness"



Bayes' rule



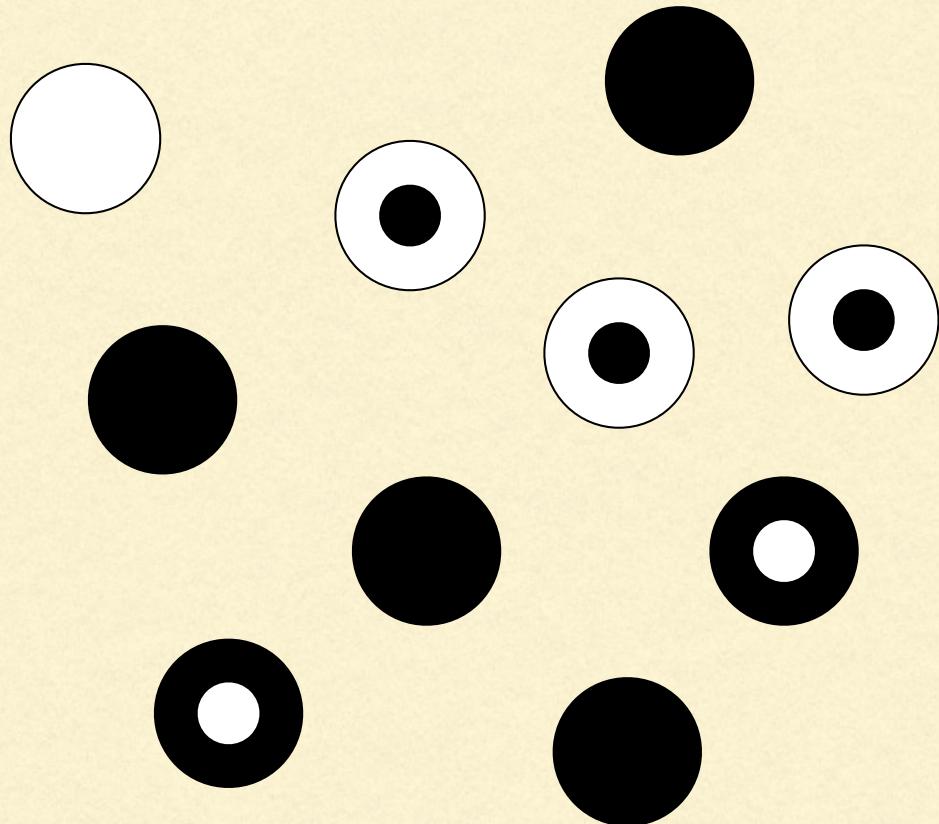
The joint probability $\Pr(B,D)$ can be written as the product of a *conditional probability* and the *probability of that condition*

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

Either B or D
can be the condition

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

Bayes' rule



Equate the two ways of writing $\Pr(B,D)$

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

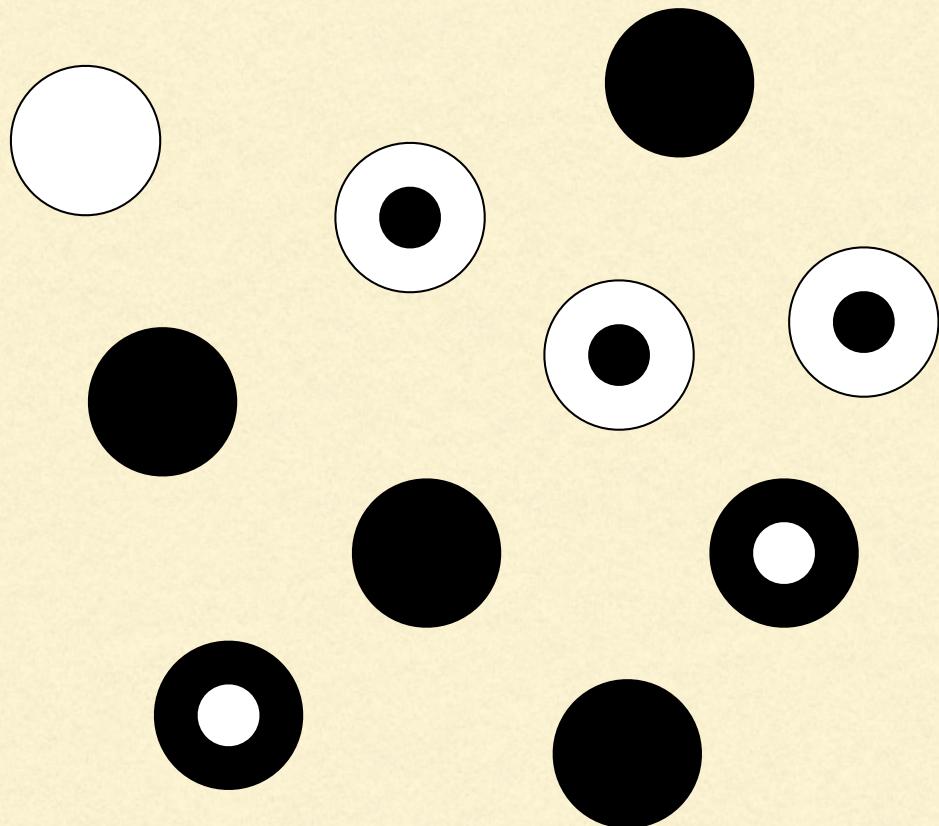
Divide both sides by $\Pr(D)$

$$\frac{\Pr(B|D) \cancel{\Pr(D)}}{\cancel{\Pr(D)}} = \frac{\Pr(D|B) \Pr(B)}{\Pr(D)}$$

Bayes' rule

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

Bayes' rule



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

$$\frac{2}{5} = \frac{2}{\text{Bayes' rule}}$$



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

Bayes' rule (variations)

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

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

Bayes' rule (variations)

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

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

$$= \frac{\Pr(D|B) \Pr(B)}{\sum_{\theta \in \{B, W\}} \Pr(D|\theta) \Pr(\theta)}$$

Bayes' rule in statistics

Likelihood of hypothesis θ

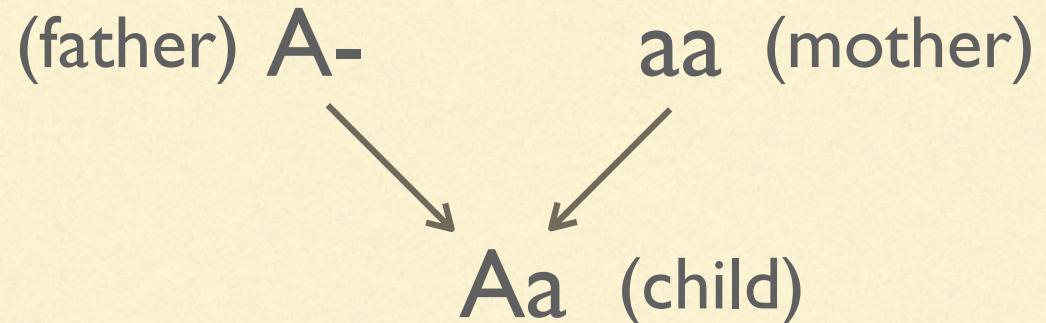
Prior probability of hypothesis θ

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

Posterior probability of hypothesis θ

Marginal probability of the data (marginalizing over hypotheses)

Paternity example



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

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

Bayes' rule: continuous case

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

↑

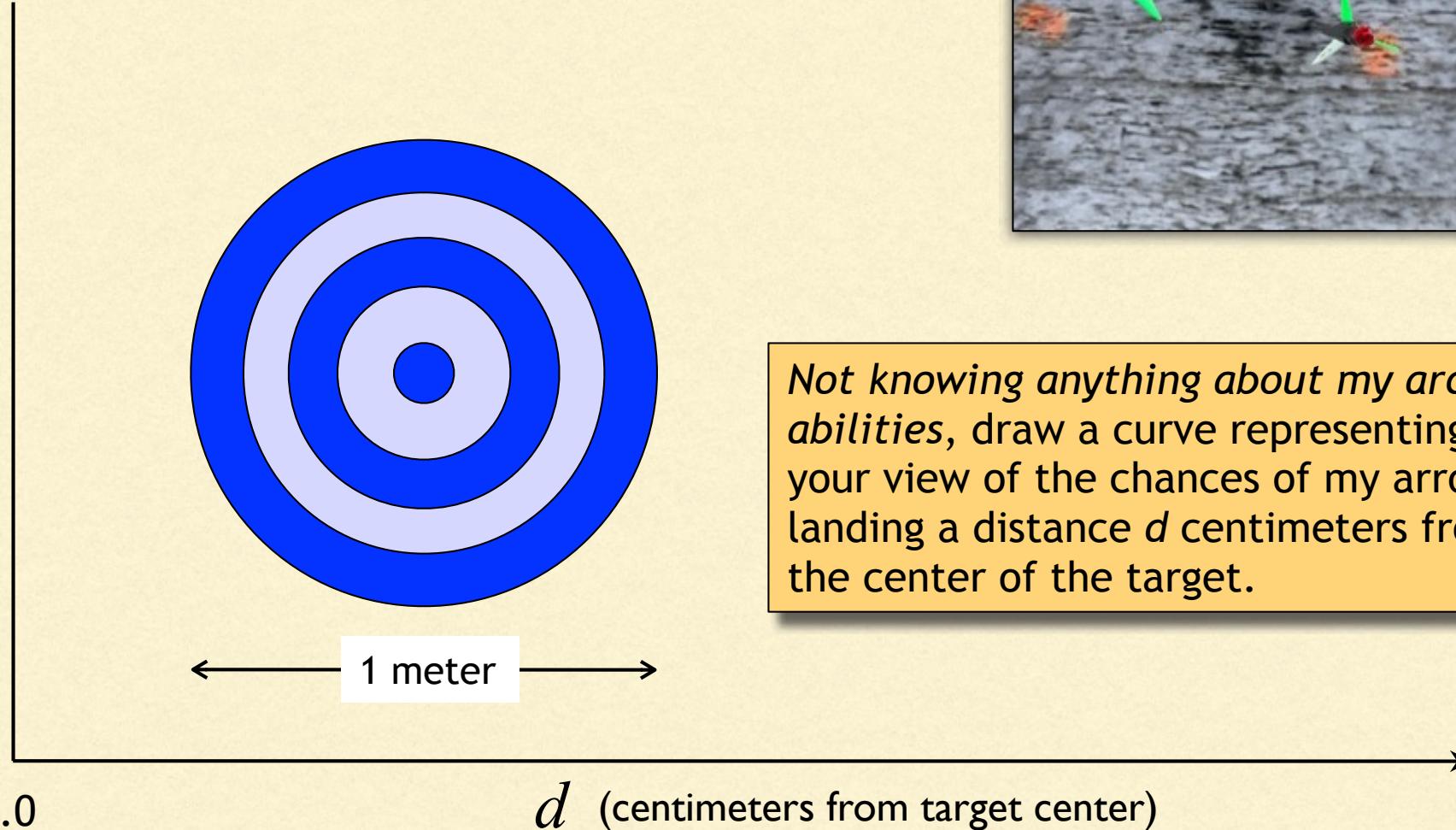
Likelihood Prior probability ***density***

Posterior probability
density

Marginal probability
of the data

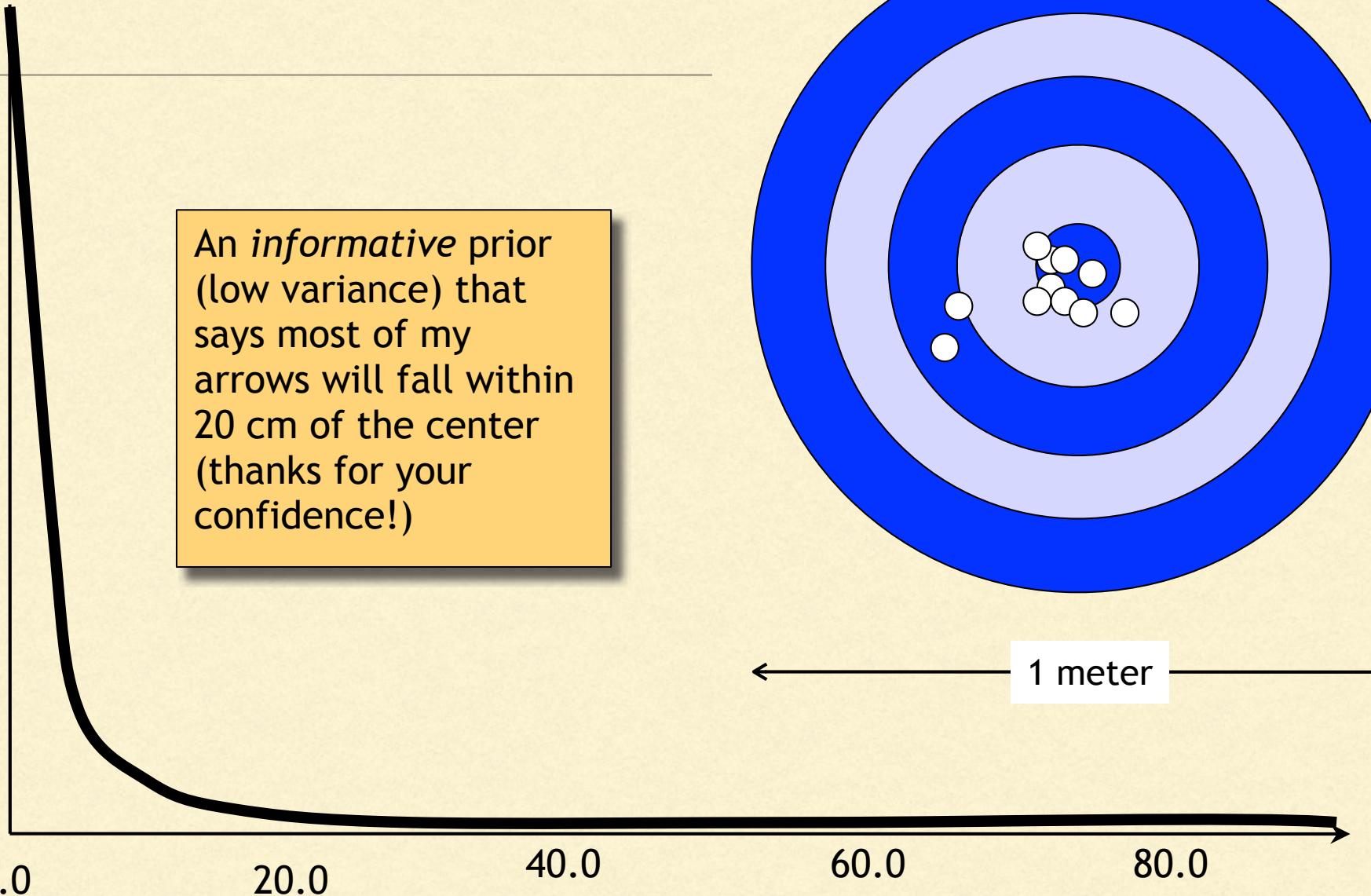
```
graph TD; A[p(D|θ)p(θ)] --> B[p(θ|D)]; A --> C[p(θ)]; D[Posterior probability  
density] --> E; F[Marginal probability  
of the data] --> G[integral p(D|θ)p(θ)dθ]
```

If you had to guess...

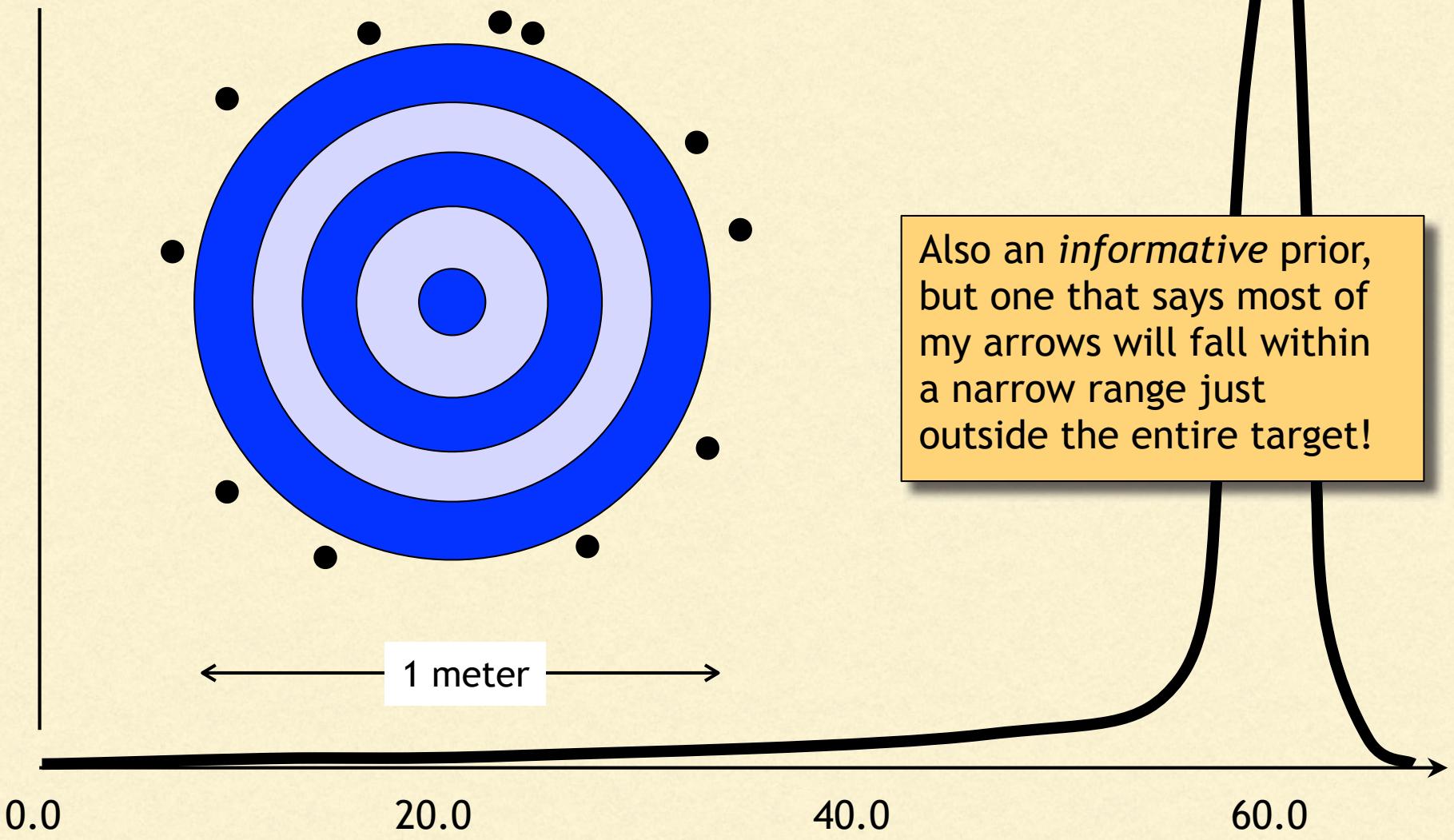


Not knowing anything about my archery abilities, draw a curve representing your view of the chances of my arrow landing a distance d centimeters from the center of the target.

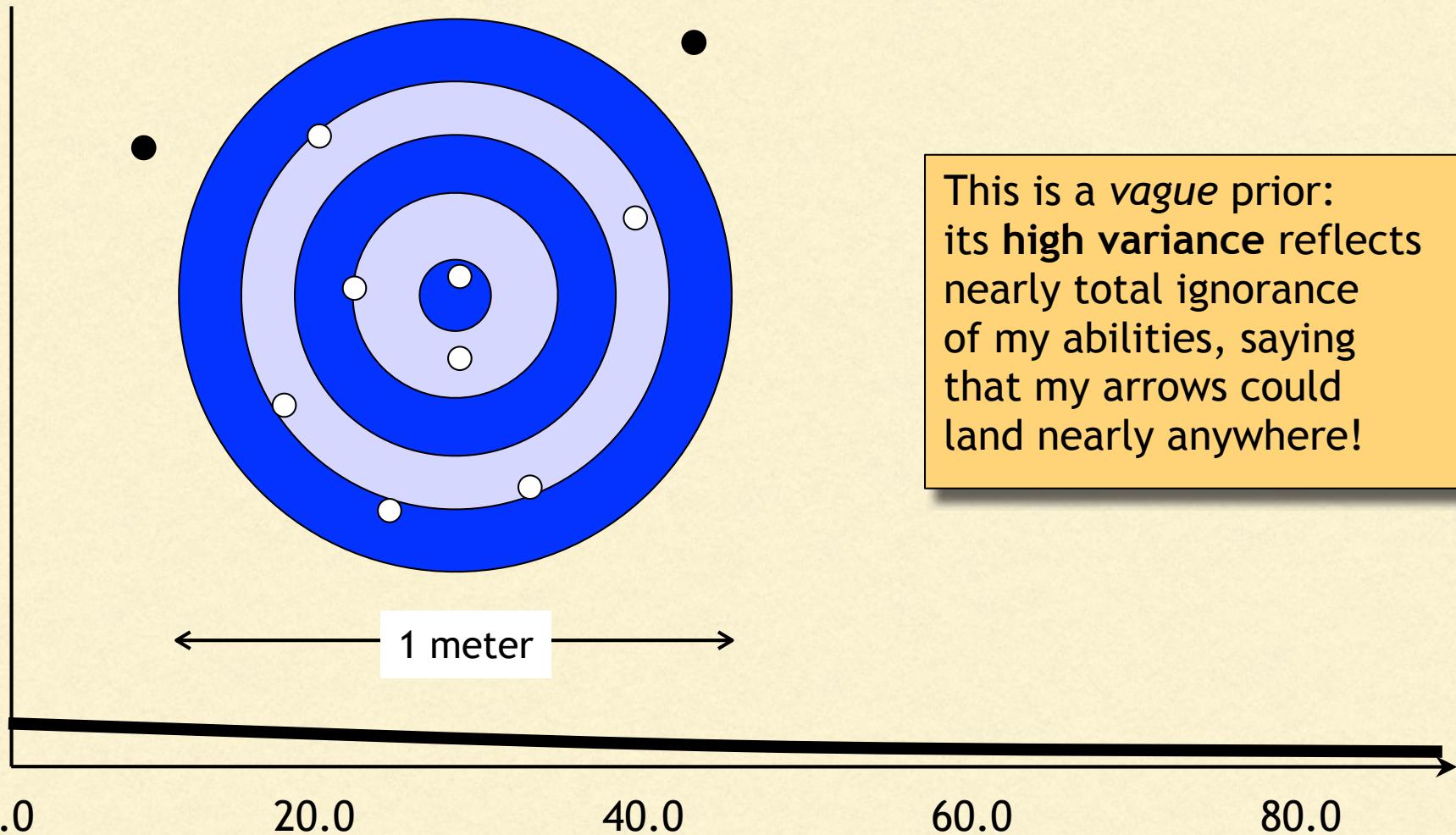
Case I: assume I have talent



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



Case 3: assume I have no talent

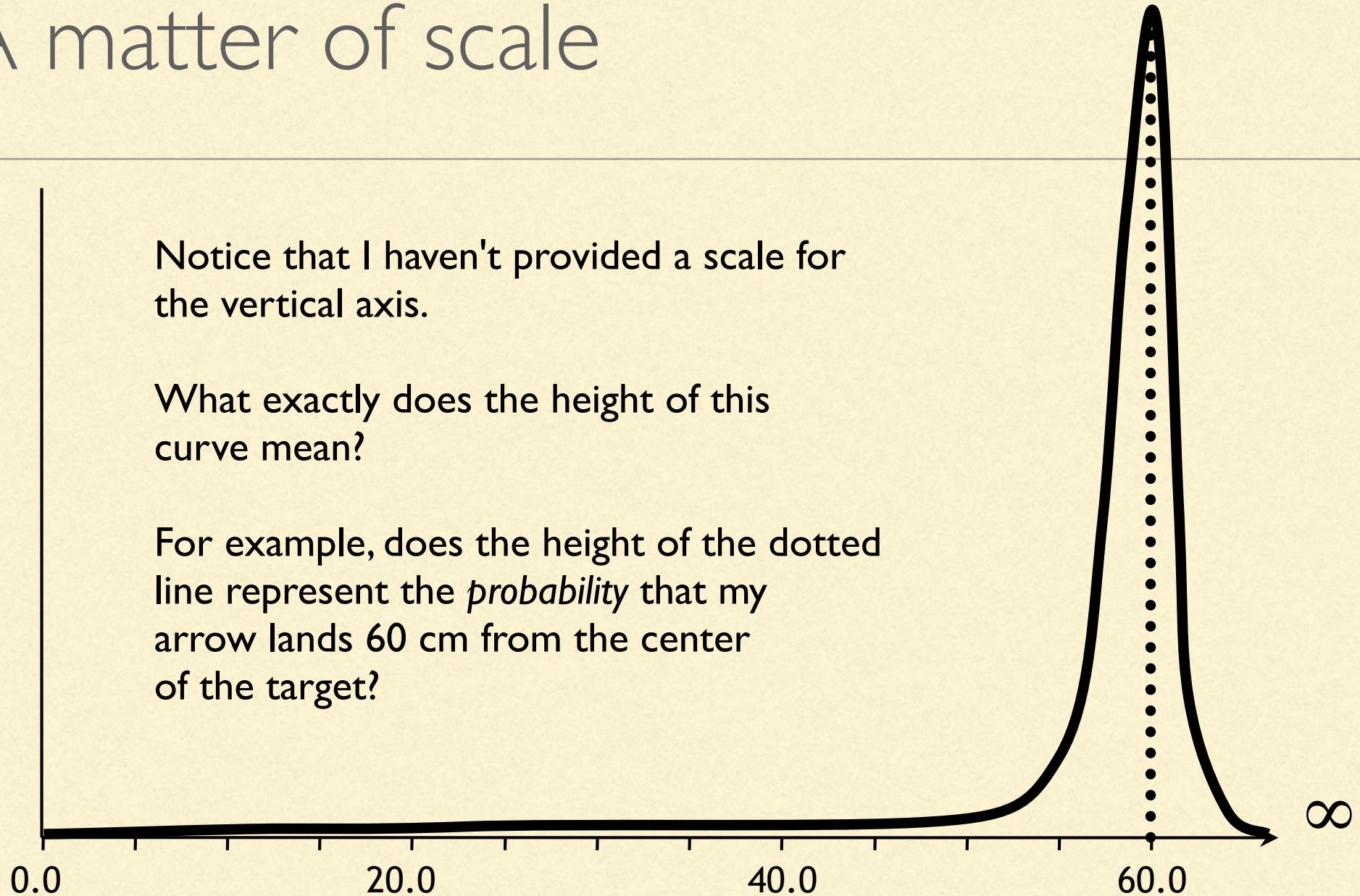


A matter of scale

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?

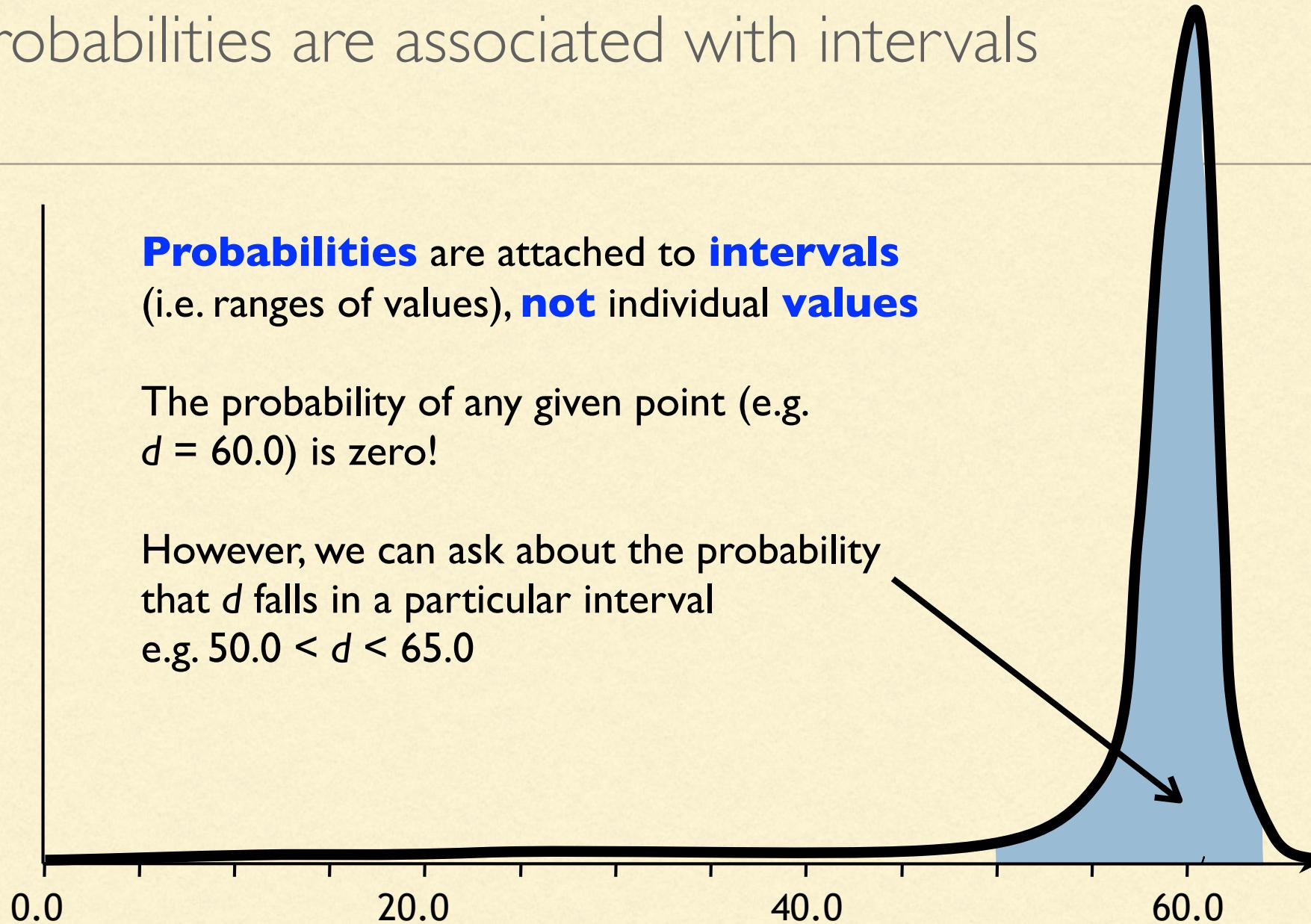


Probabilities are associated with intervals

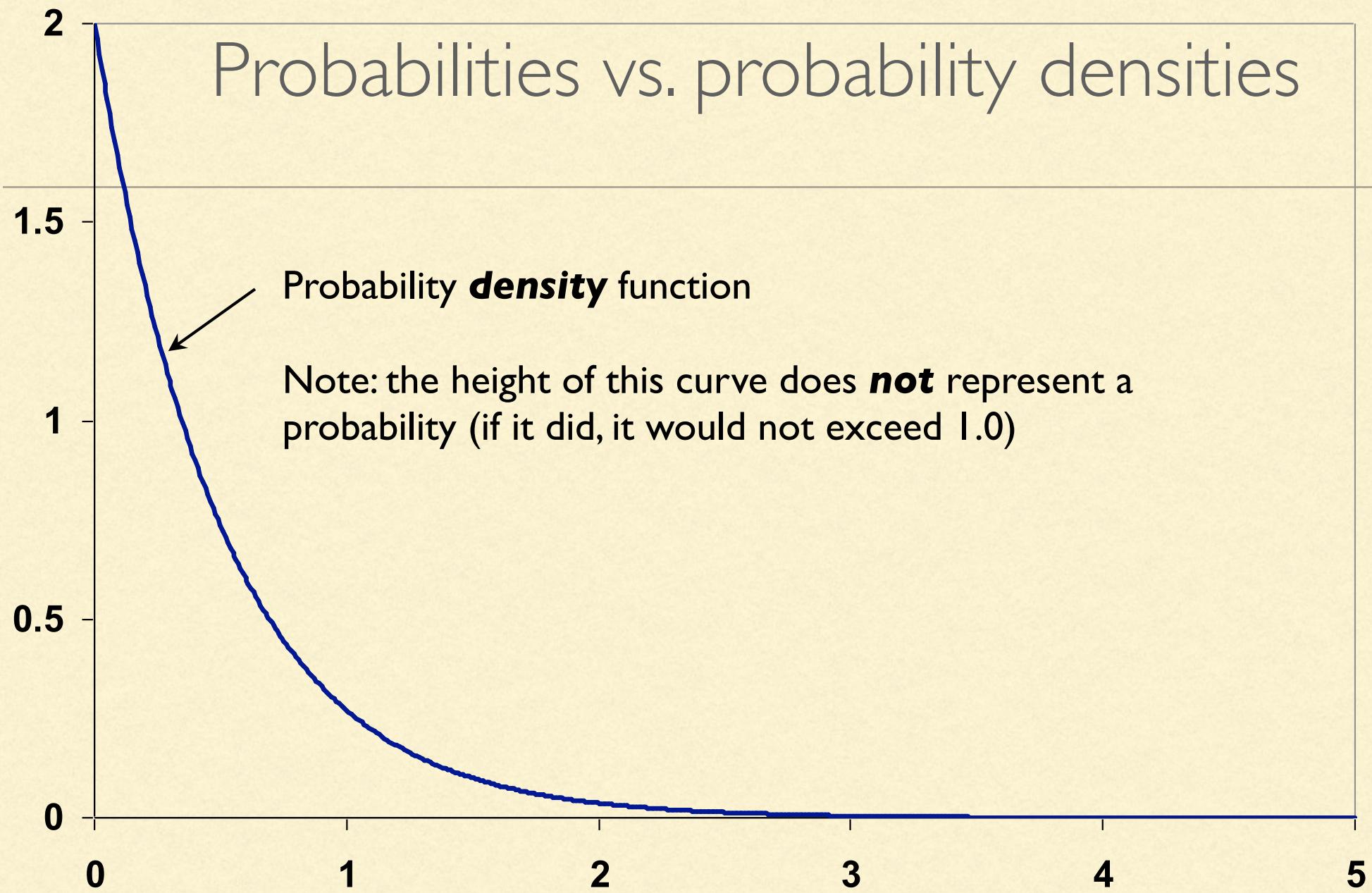
Probabilities are attached to **intervals**
(i.e. ranges of values), **not** individual **values**

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 interval
e.g. $50.0 < d < 65.0$



Probabilities vs. probability densities

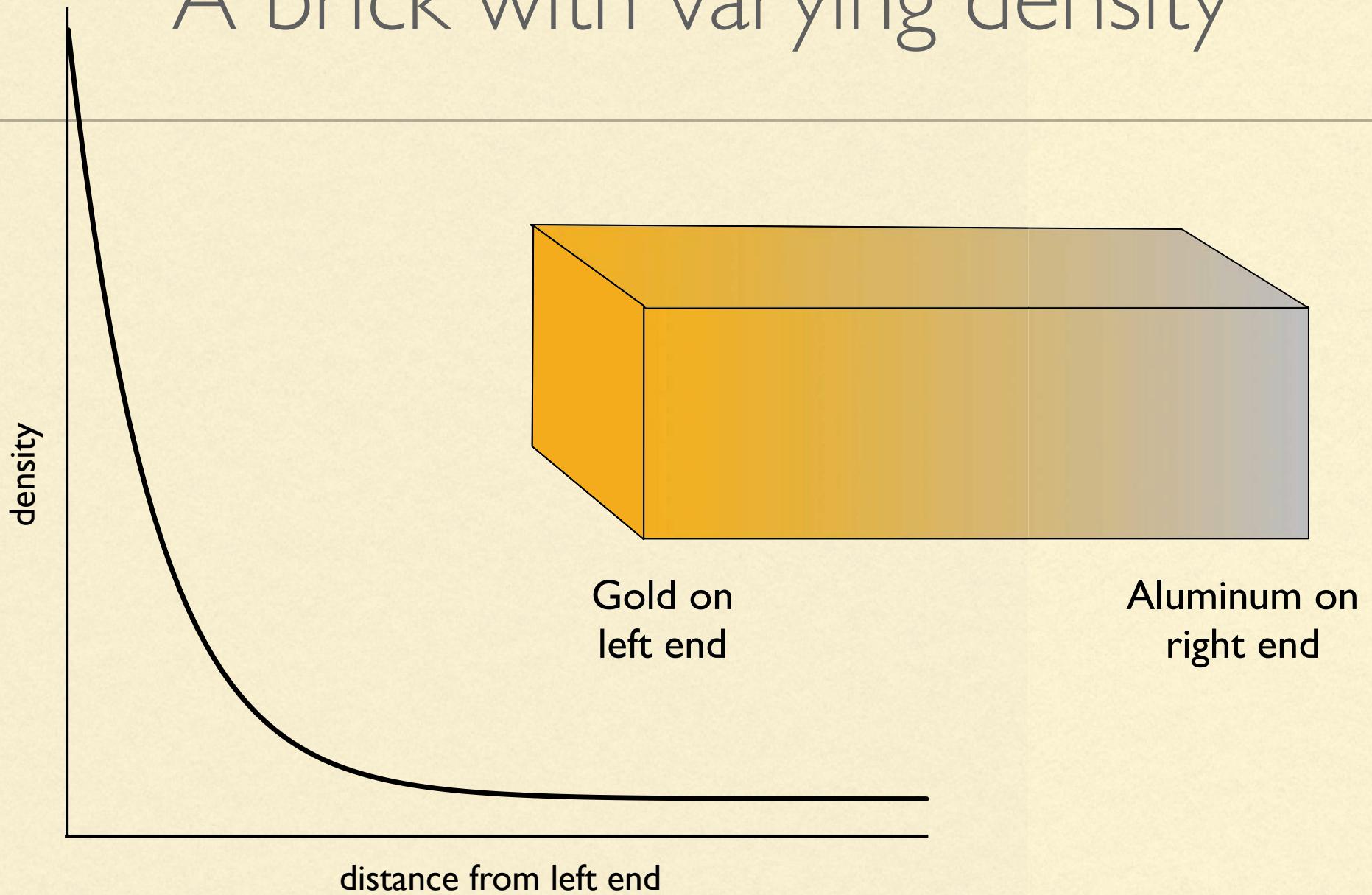


Densities of various substances

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

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

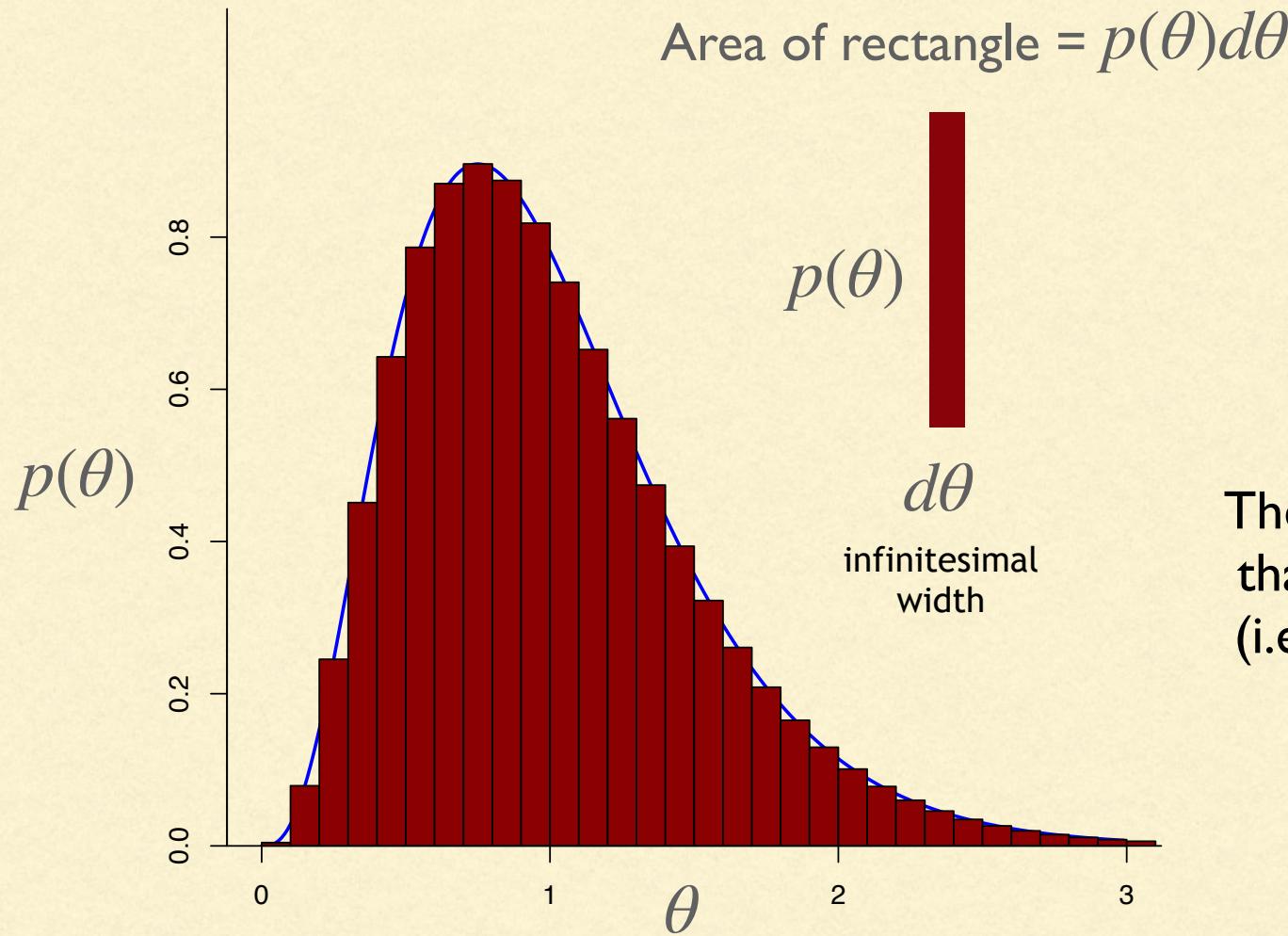
A brick with varying density



Density Rain Applet

<https://plewis.github.io/applets/density-rain/>

Integrating a density yields a probability

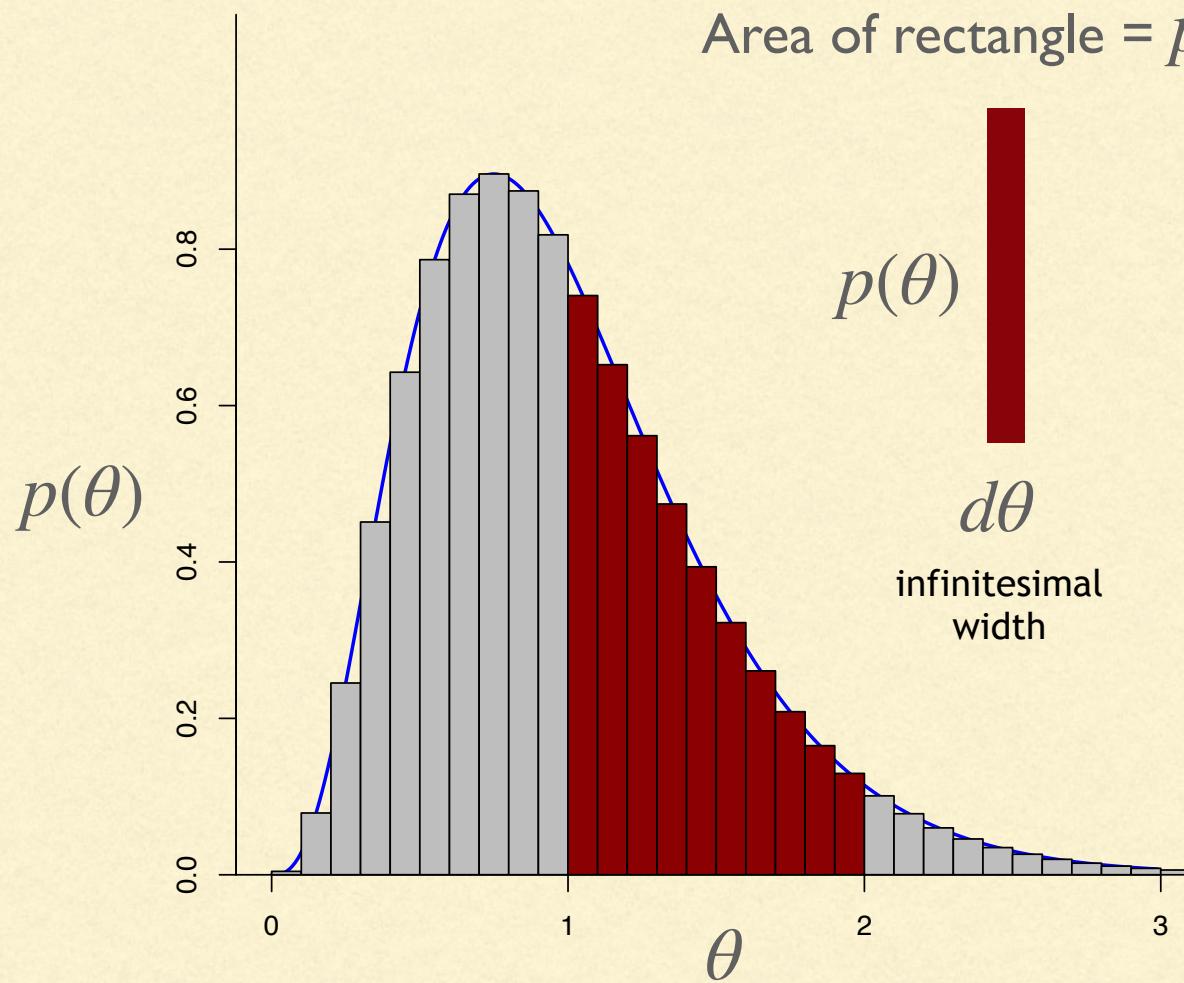


Long s from U.S. Bill of Rights

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

The density curve is scaled so that the value of this integral (i.e. the total area) equals 1.0

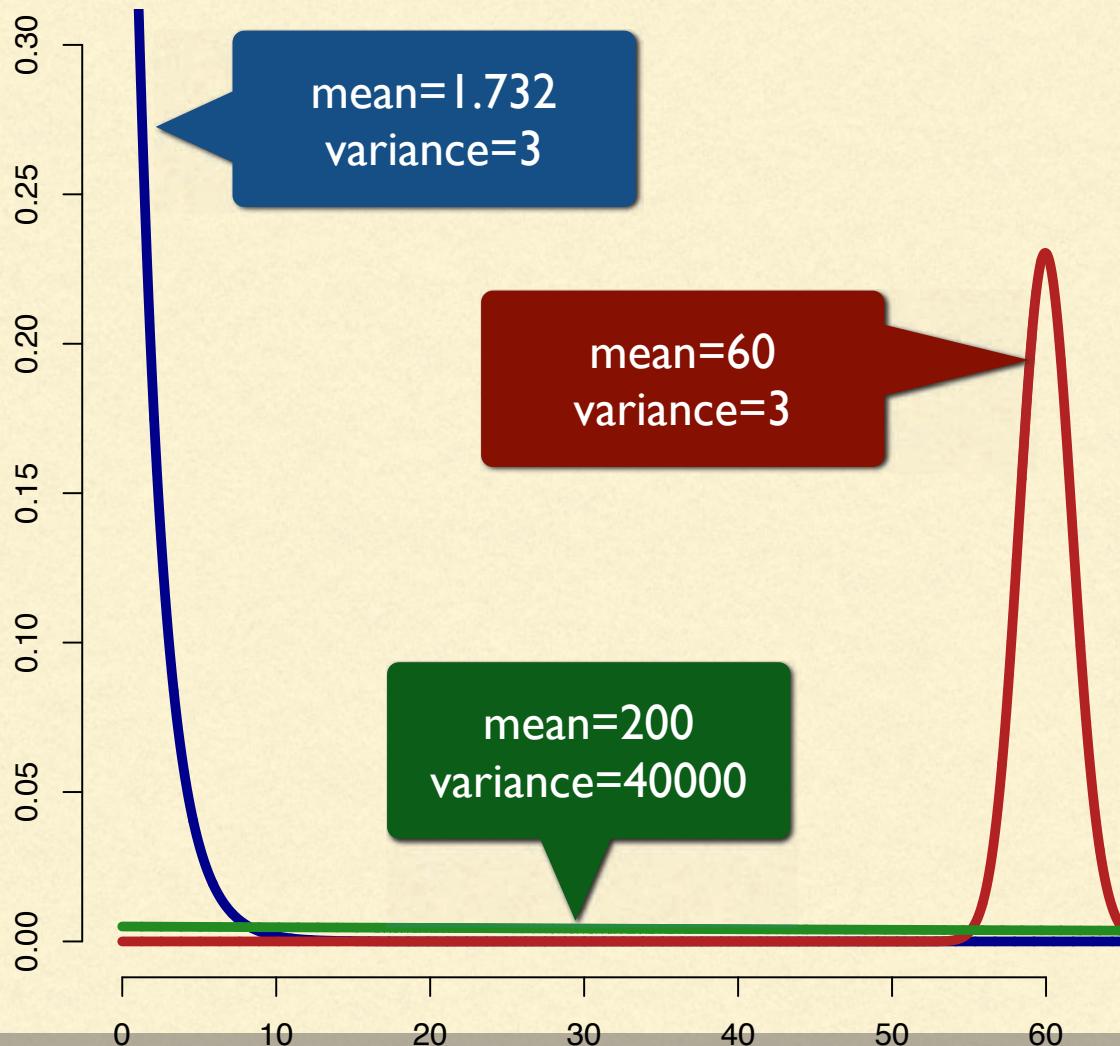
Integrating a density yields a probability



$$0.39109 = \int_1^2 p(\theta)d\theta$$

The **area** under the density curve from 1 to 2 is the **probability** that θ is between 1 and 2

Archery priors revisited



These density curves are all variations of a **gamma probability distribution**.

We could have used a gamma distribution to specify each of the prior probability distributions for the archery example.

Note that **higher variance** means **less informative**

Usually there are many parameters...

A 2-parameter example

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

↑

Posterior probability density

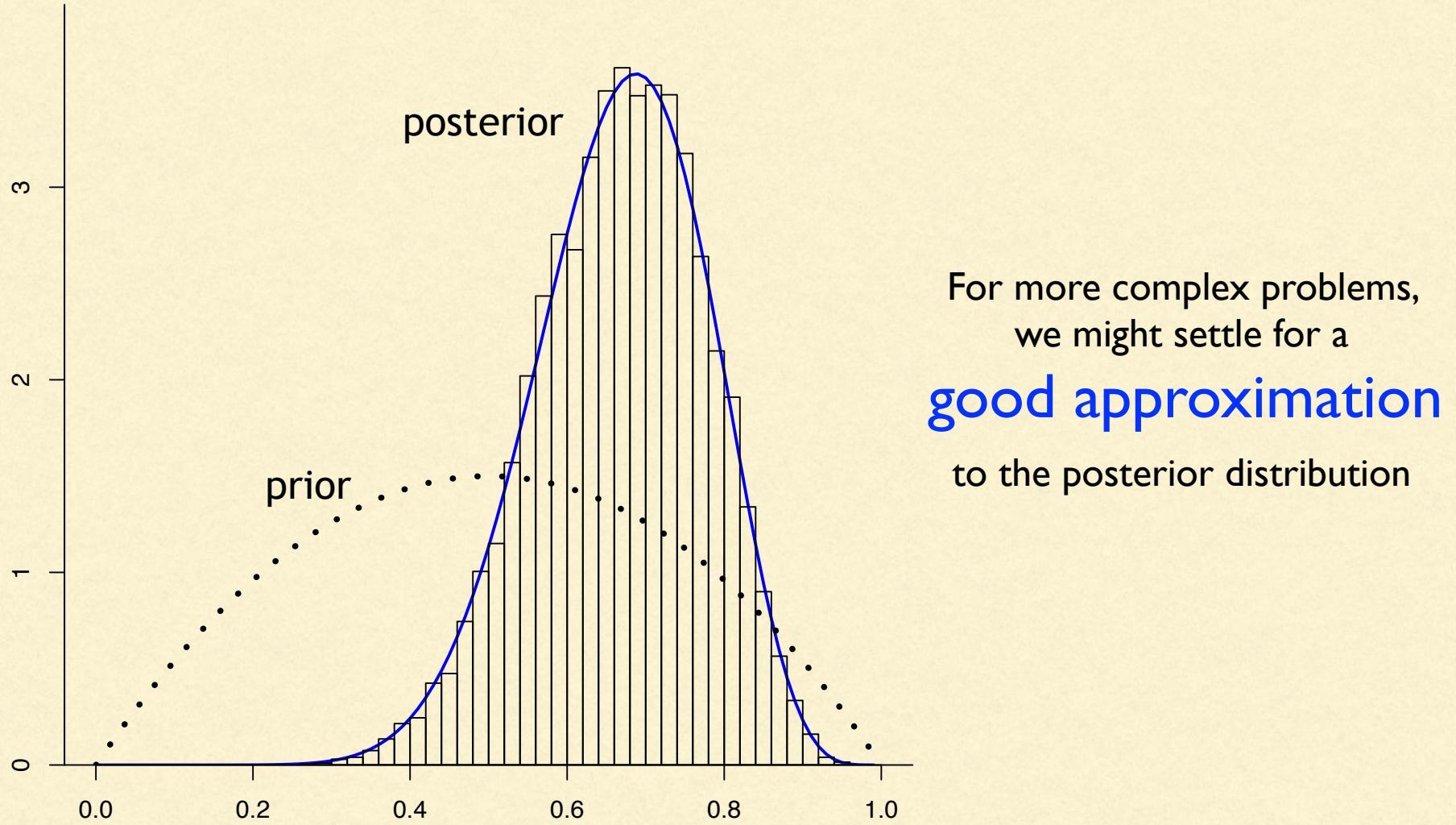
Likelihood Prior density

Marginal probability of data

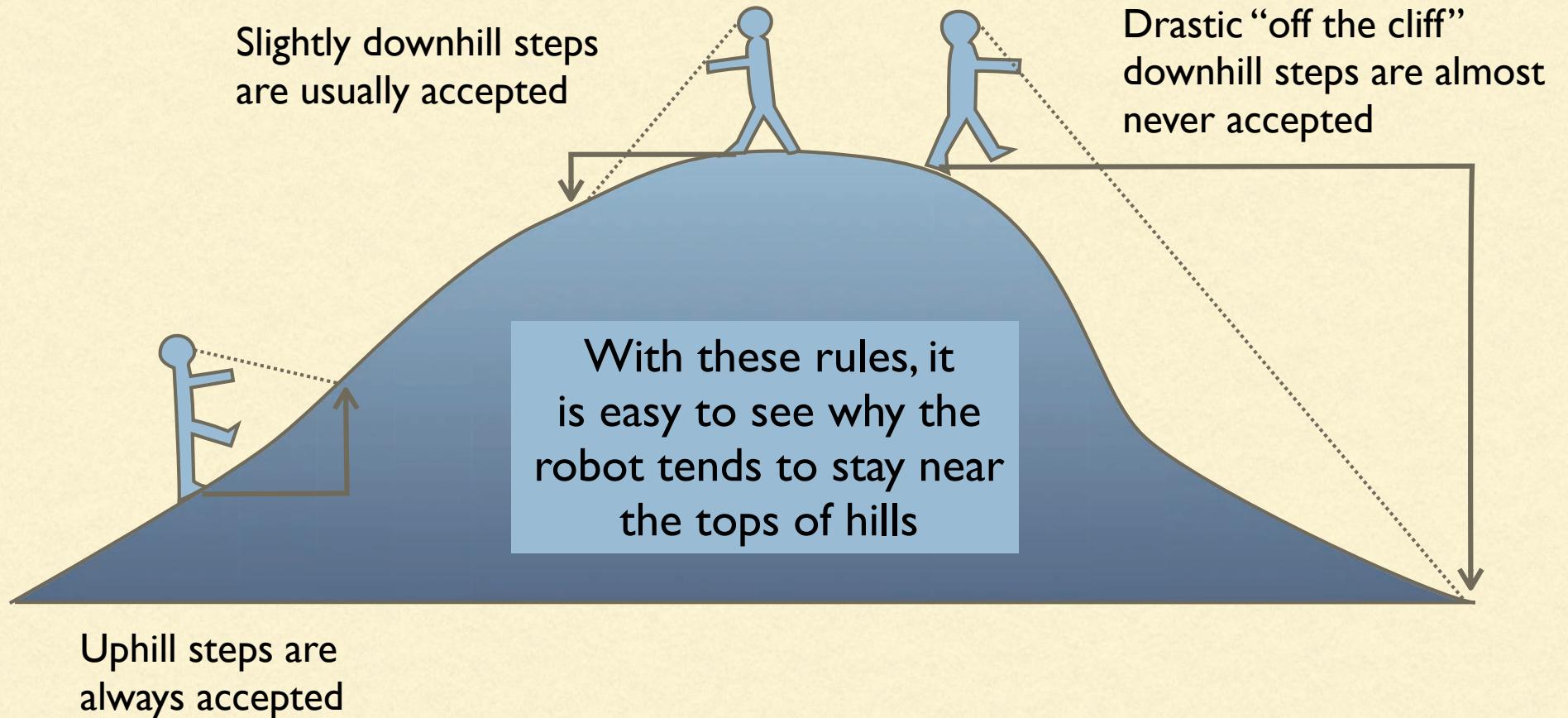
An analysis of **100 sequences** under the simplest model (JC69) requires 197 branch length parameters. The denominator would require a **197-fold integral** inside a sum over **all possible tree topologies!** It would thus be nice to avoid having to calculate the marginal probability of the data...

Markov chain Monte Carlo (MCMC)

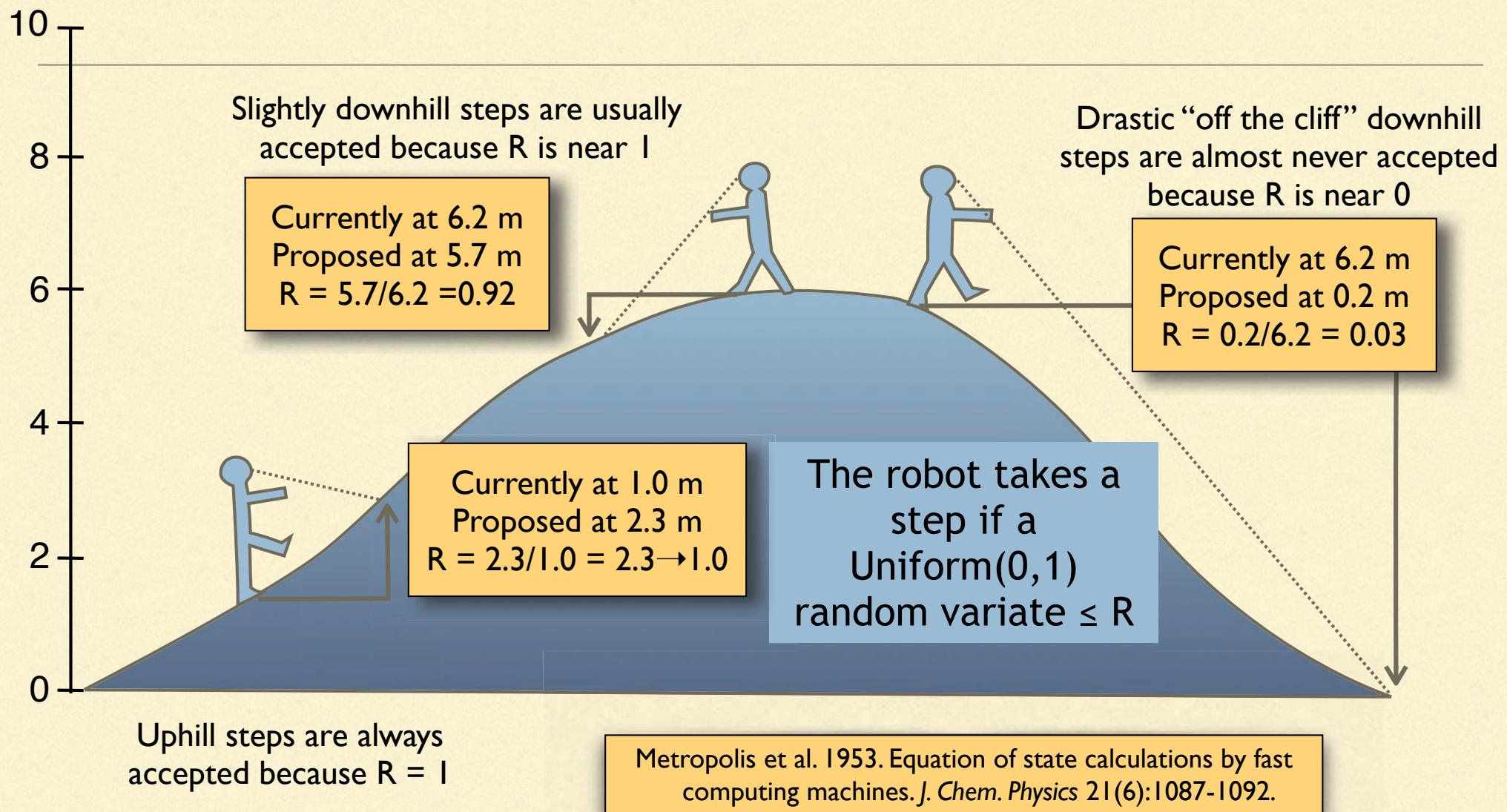
Markov chain Monte Carlo (MCMC)



MCMC robot's rules



Actual rules (Metropolis algorithm)



Cancellation of marginal likelihood

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

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

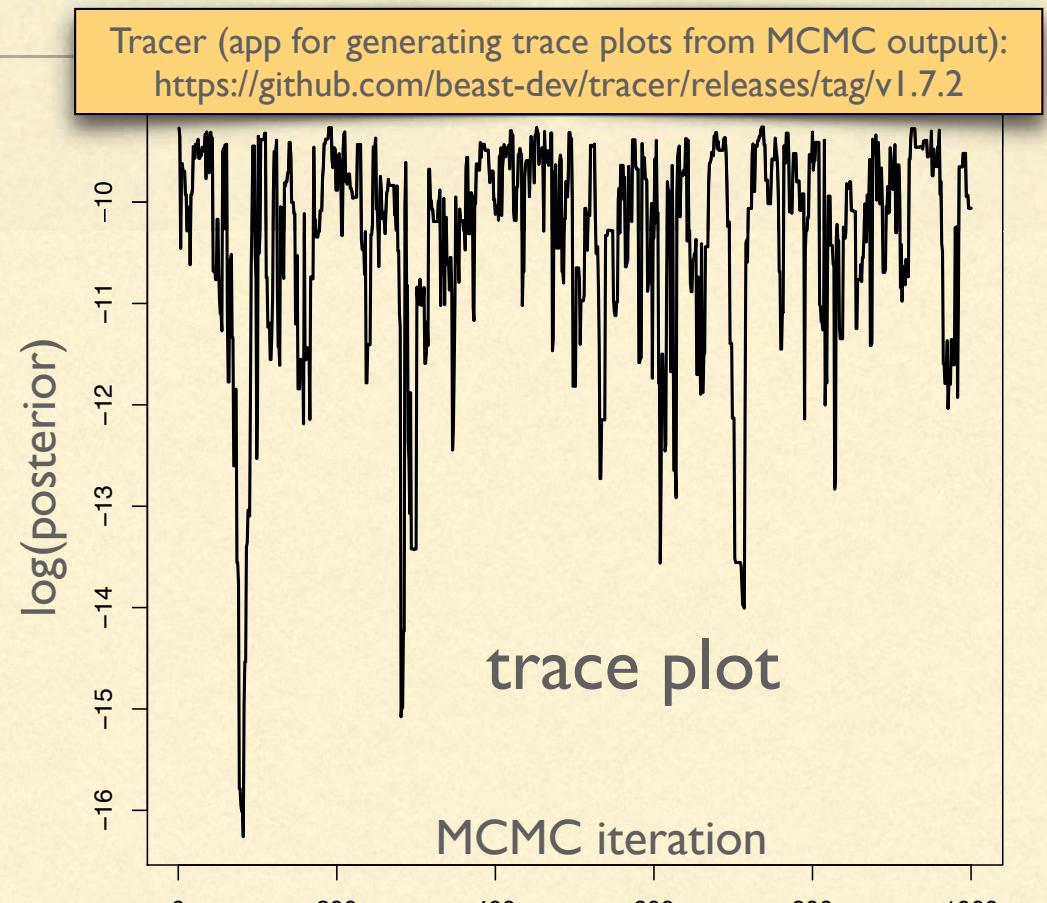
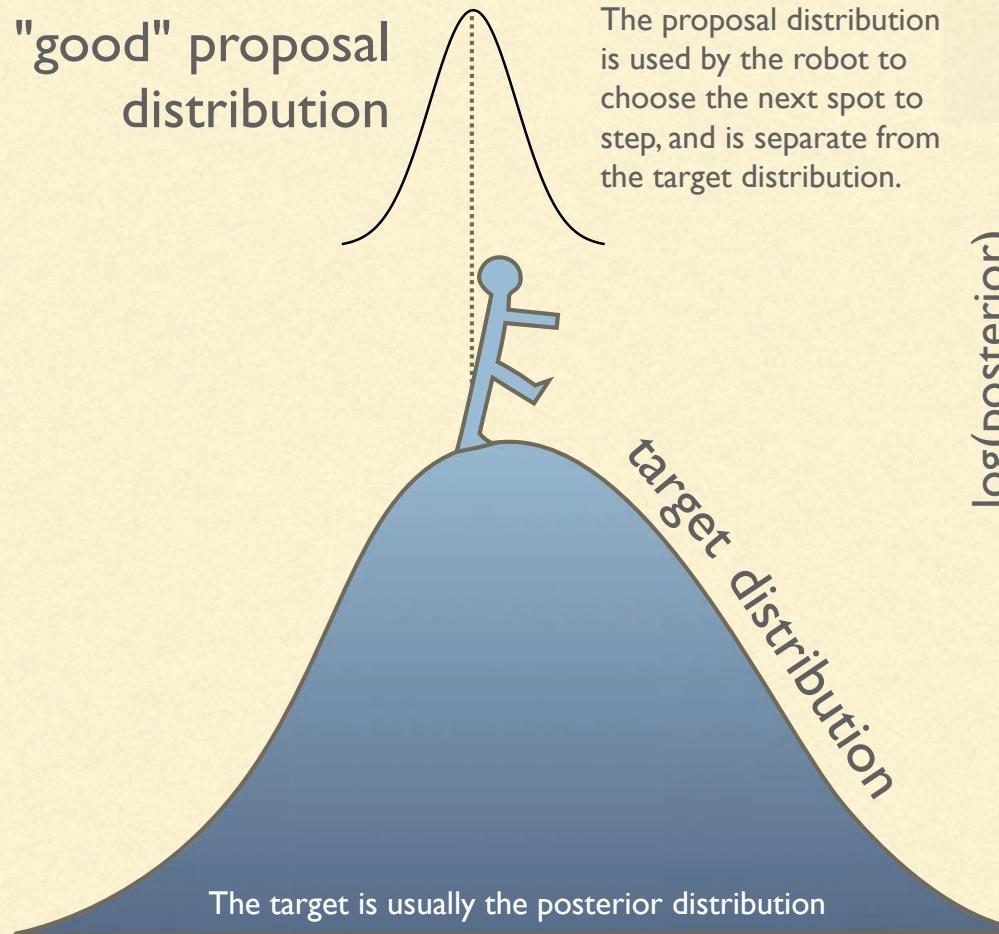
Posterior
ratio

Apply Bayes' rule to
both top and bottom

Likelihood
ratio

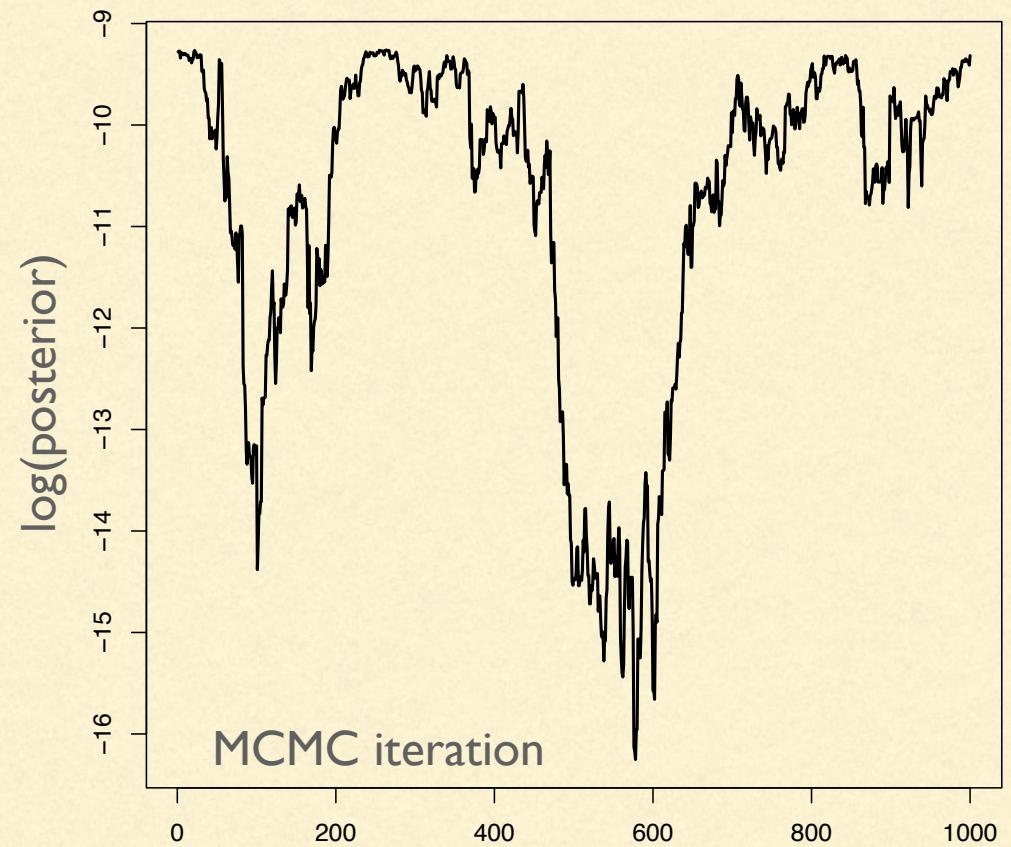
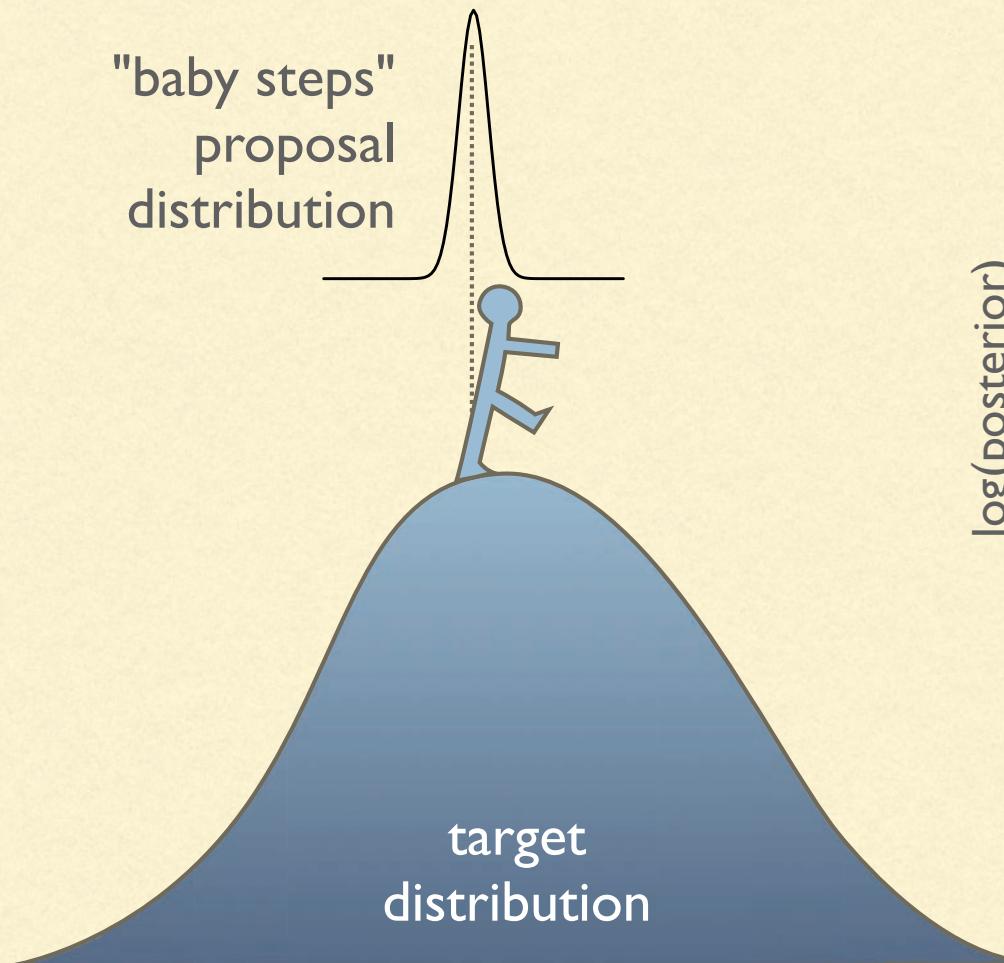
Prior
ratio

Target vs. Proposal Distributions



White noise appearance is a sign of **good mixing**

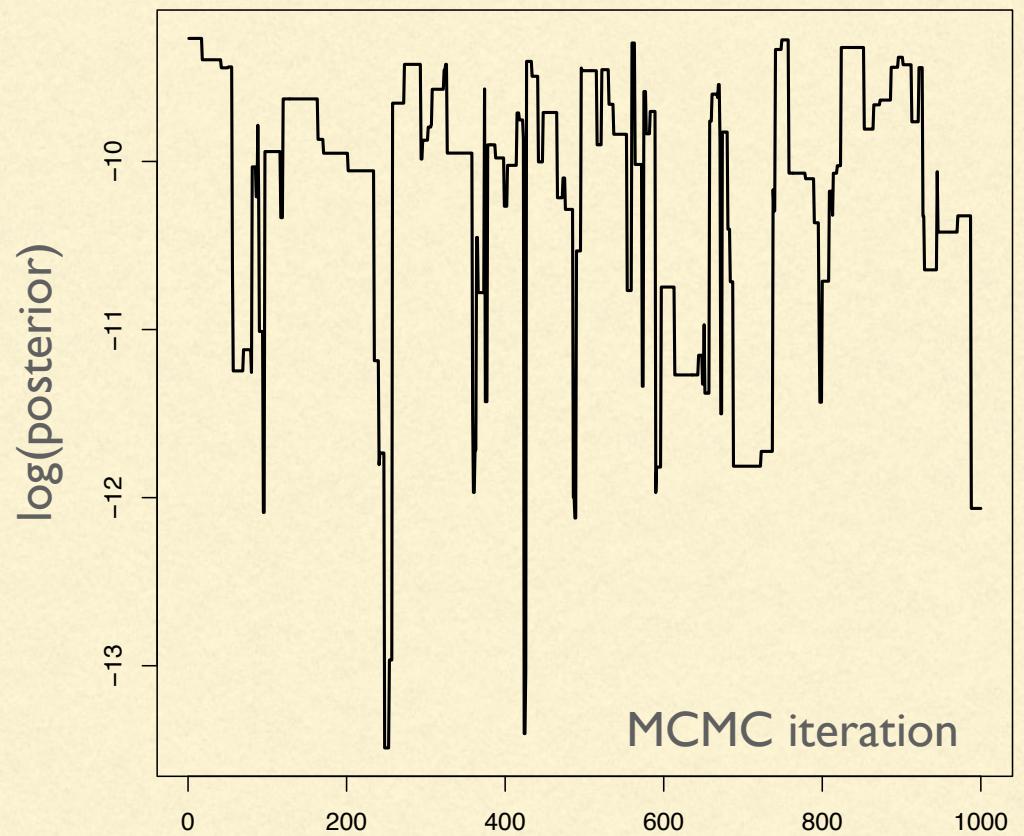
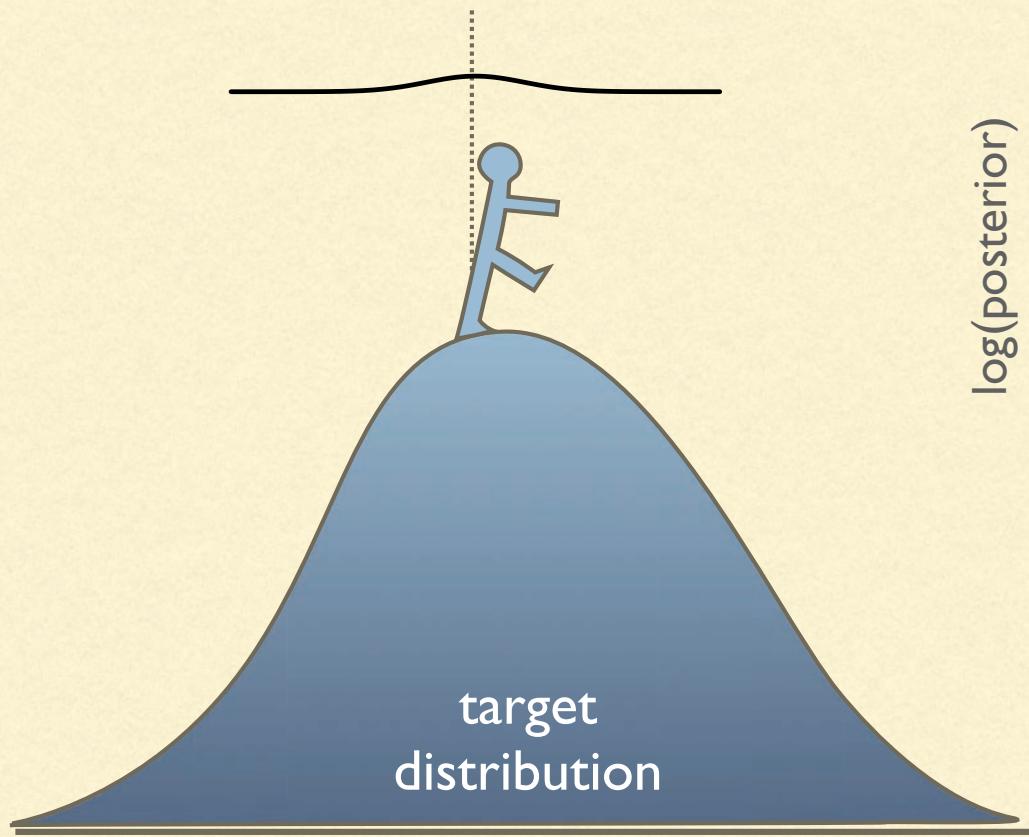
Target vs. Proposal Distributions



Big waves in trace plot indicate
robot is crawling around

Target vs. Proposal Distributions

"overly bold" proposal distribution



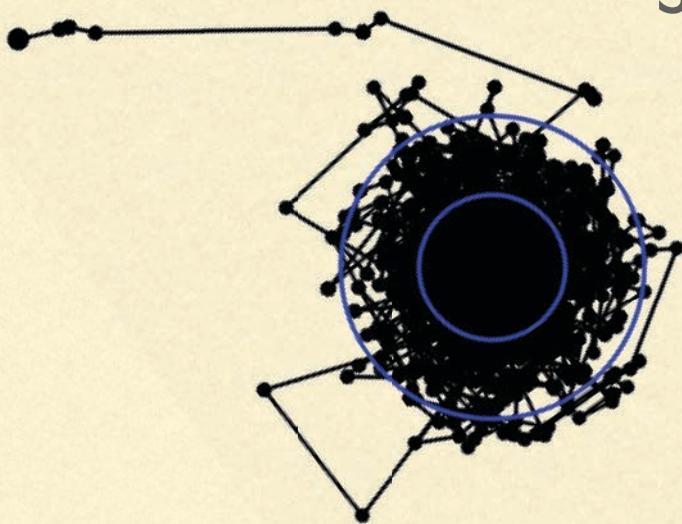
Plateaus in trace plot indicate robot is often stuck in one place

MCRobot (or "MCMC Robot")

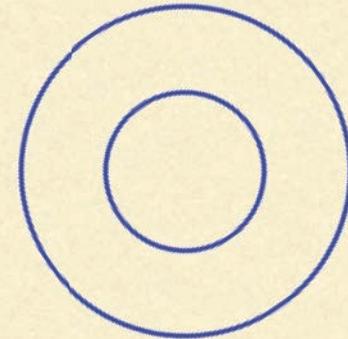
Javascript version used today will run in most web
browsers and is available here:

<https://plewis.github.io/applets/mcmc-robot/>

Metropolis-coupled Markov chain Monte Carlo (MCMCMC)



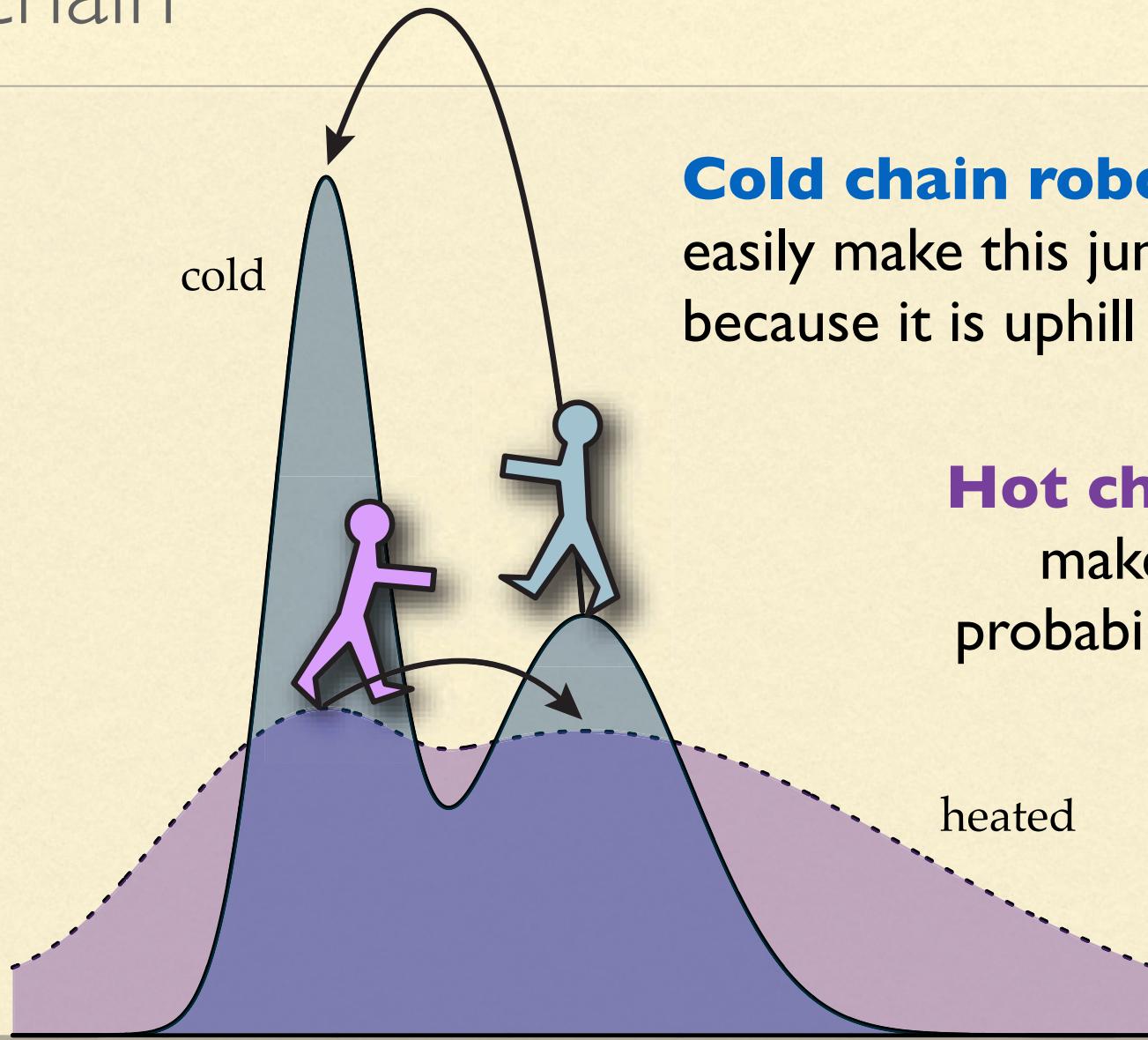
Sometimes the robot needs some help,



MCMCMC introduces helpers in the form of "heated chain" robots that can act as scouts.

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

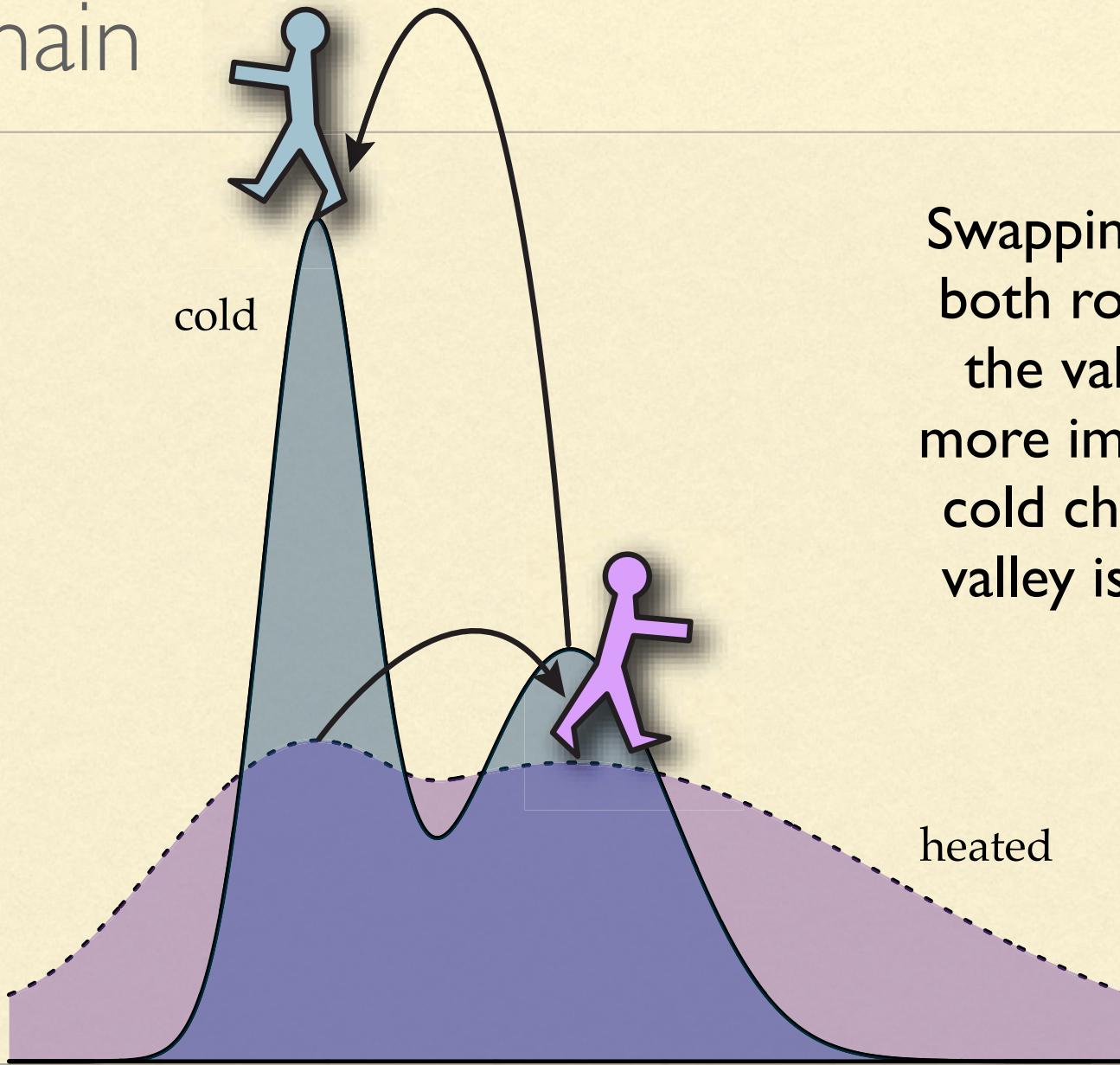
Heated chains act as scouts for the cold chain



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

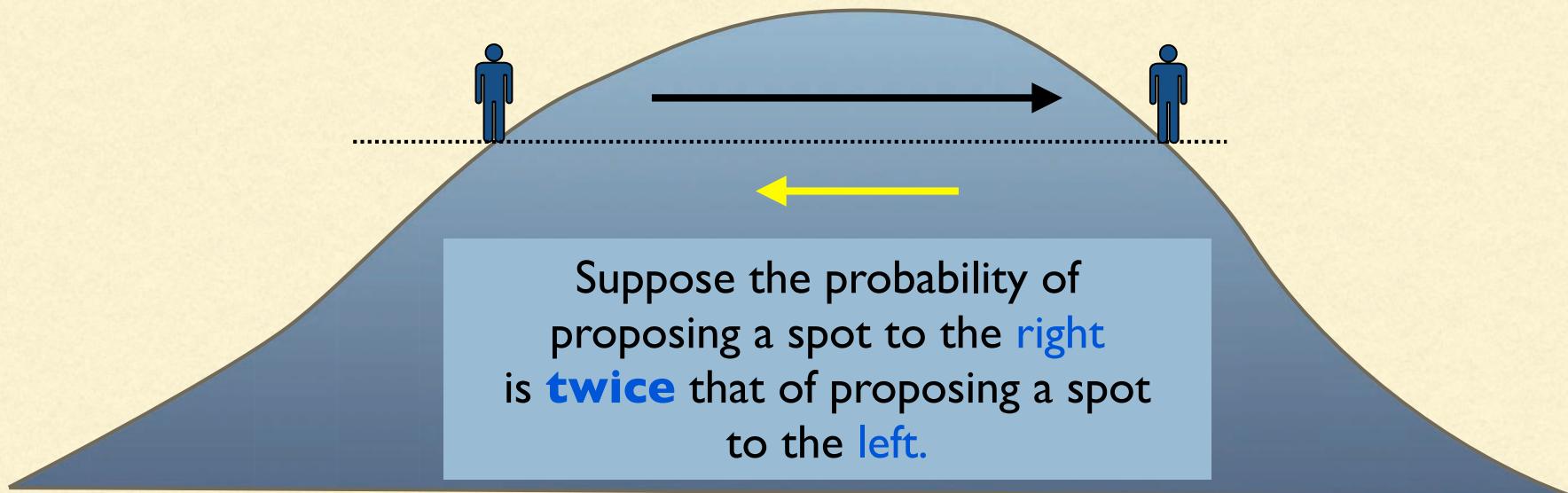
Hot chain robot can also make this jump with high probability because it is only slightly downhill

Heated chains act as scouts for the cold chain



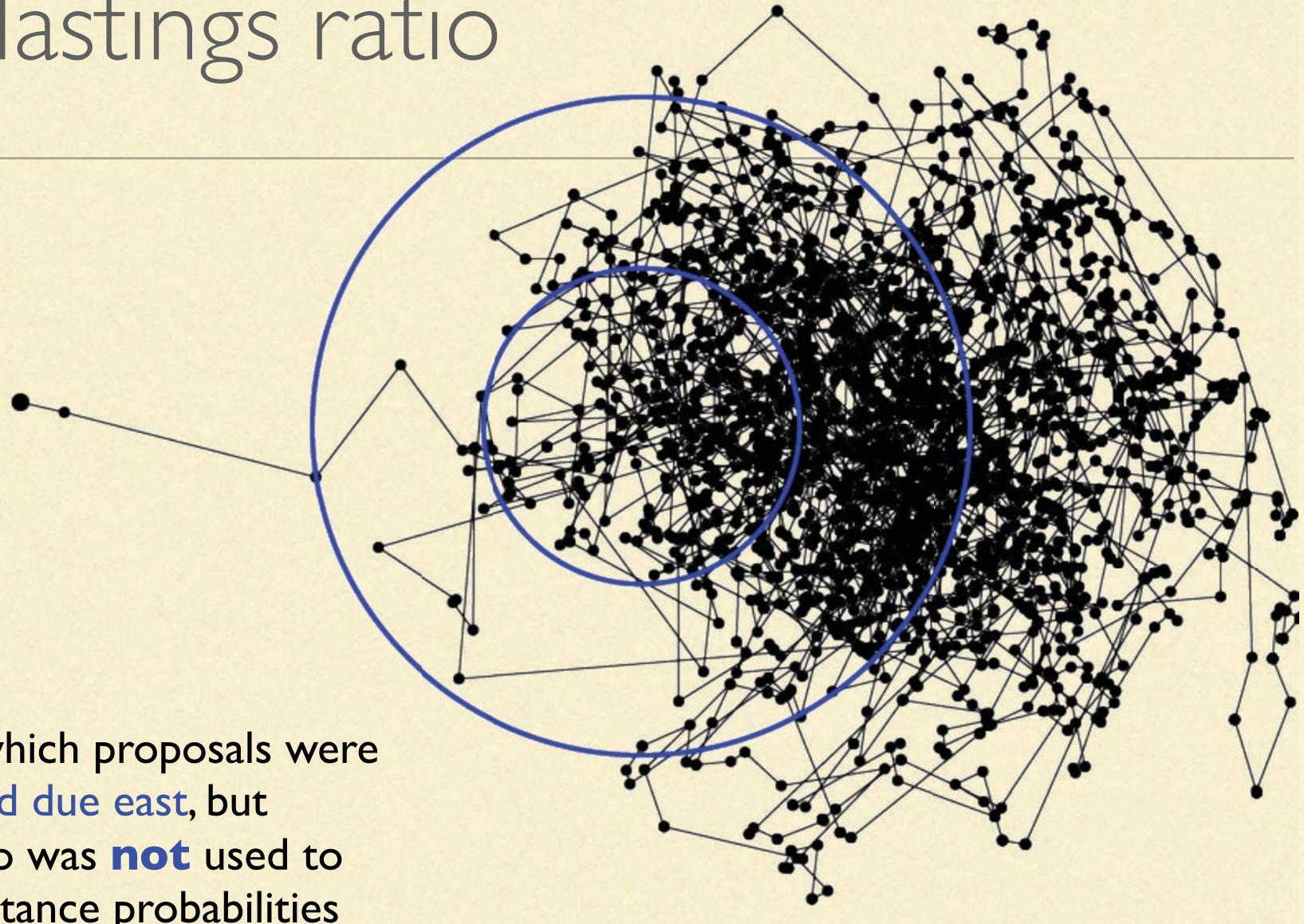
Swapping places means both robots can cross the valley, but this is more important for the cold chain because its valley is much deeper.

The Hastings ratio



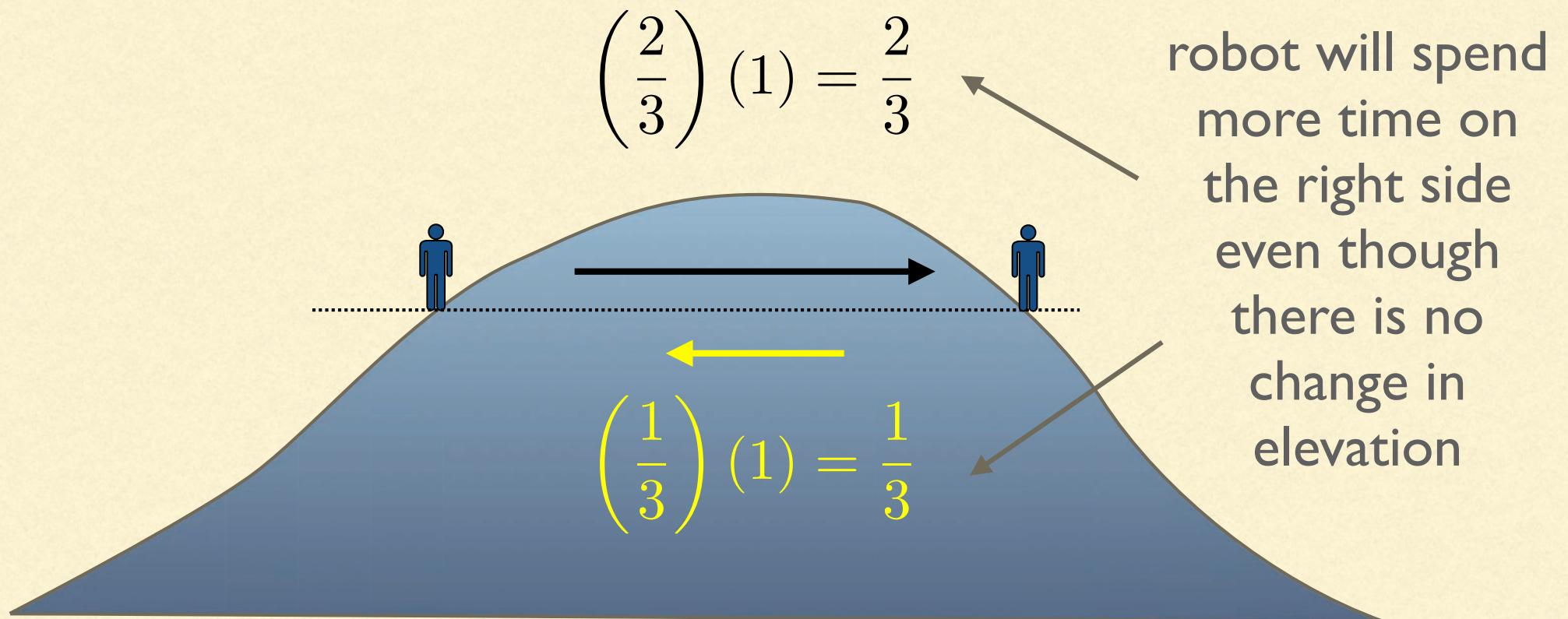
Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97-109.

The Hastings ratio



Example in which proposals were
biased toward due east, but
Hastings ratio was **not** used to
modify acceptance probabilities

The Hastings ratio



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

Hastings Ratio

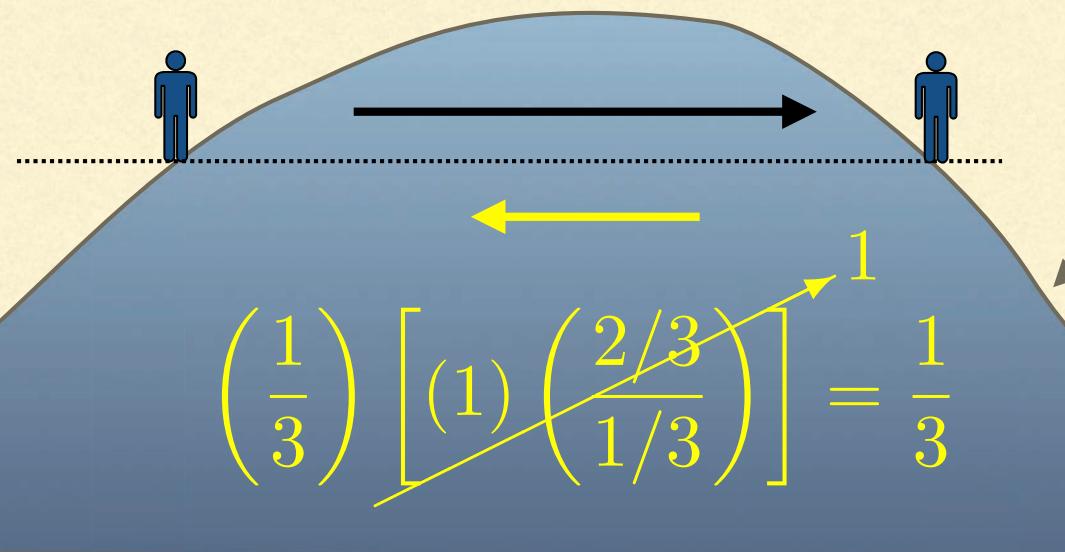
$$R = \min \left\{ 1, \left[\frac{p(D | \theta^*) p(\theta^*)}{p(D | \theta) p(\theta)} \right] \left[\frac{q(\theta | \theta^*)}{q(\theta^* | \theta)} \right] \right\}$$

posterior ratio Hastings ratio

Note that the Hastings ratio is 1.0 if $q(\theta^* | \theta) = q(\theta | \theta^*)$

The Hastings ratio

$$\left(\frac{2}{3}\right) \left[(1) \left(\frac{1/3}{2/3} \right) \right] = \frac{1}{3}$$



robot spends same amount of time on both sides, as it should

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