Maximum Likelihood and Bayesian Inference

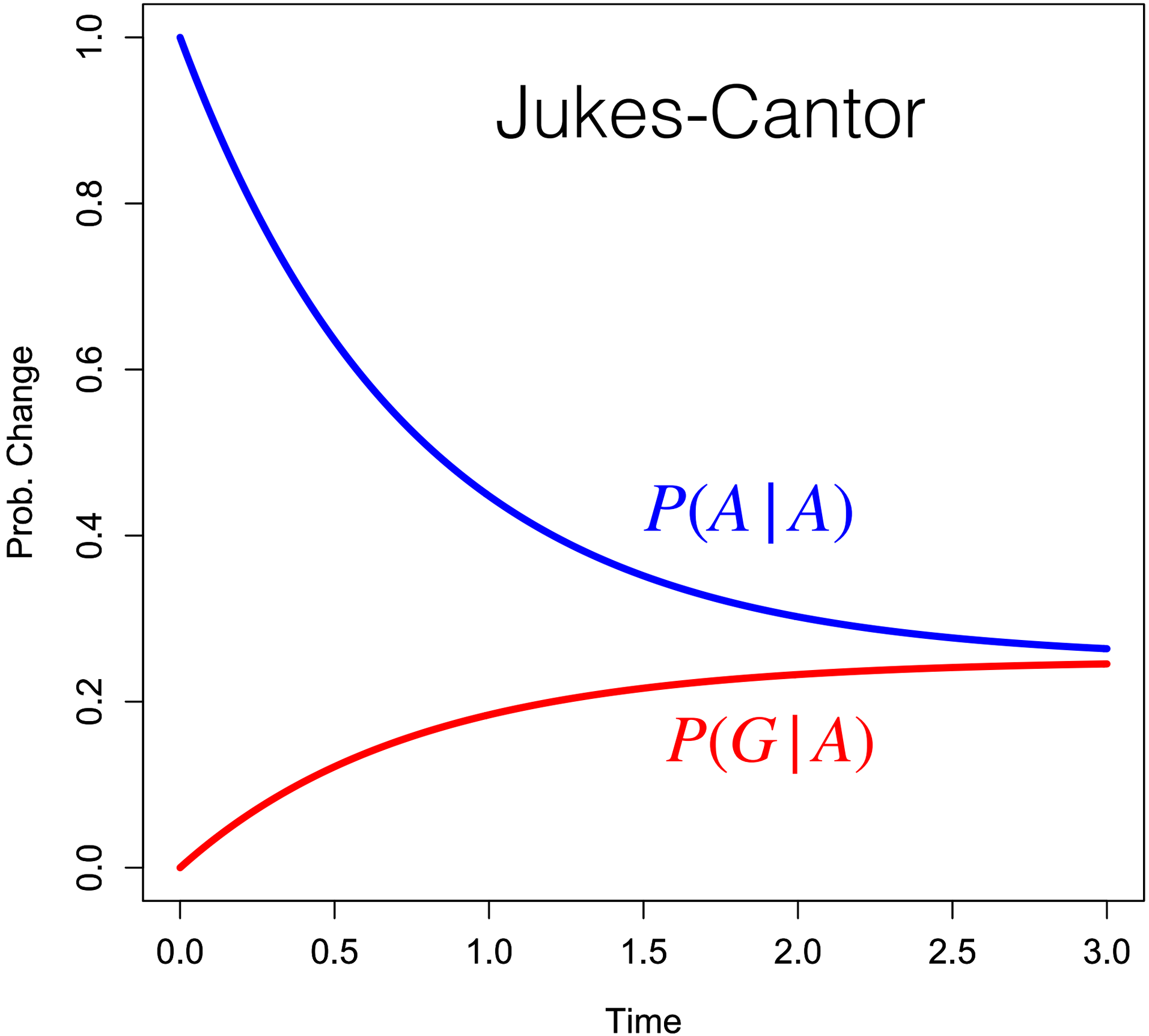
Probability of Starting and Ending

$$P(t) = e^{Qt}$$

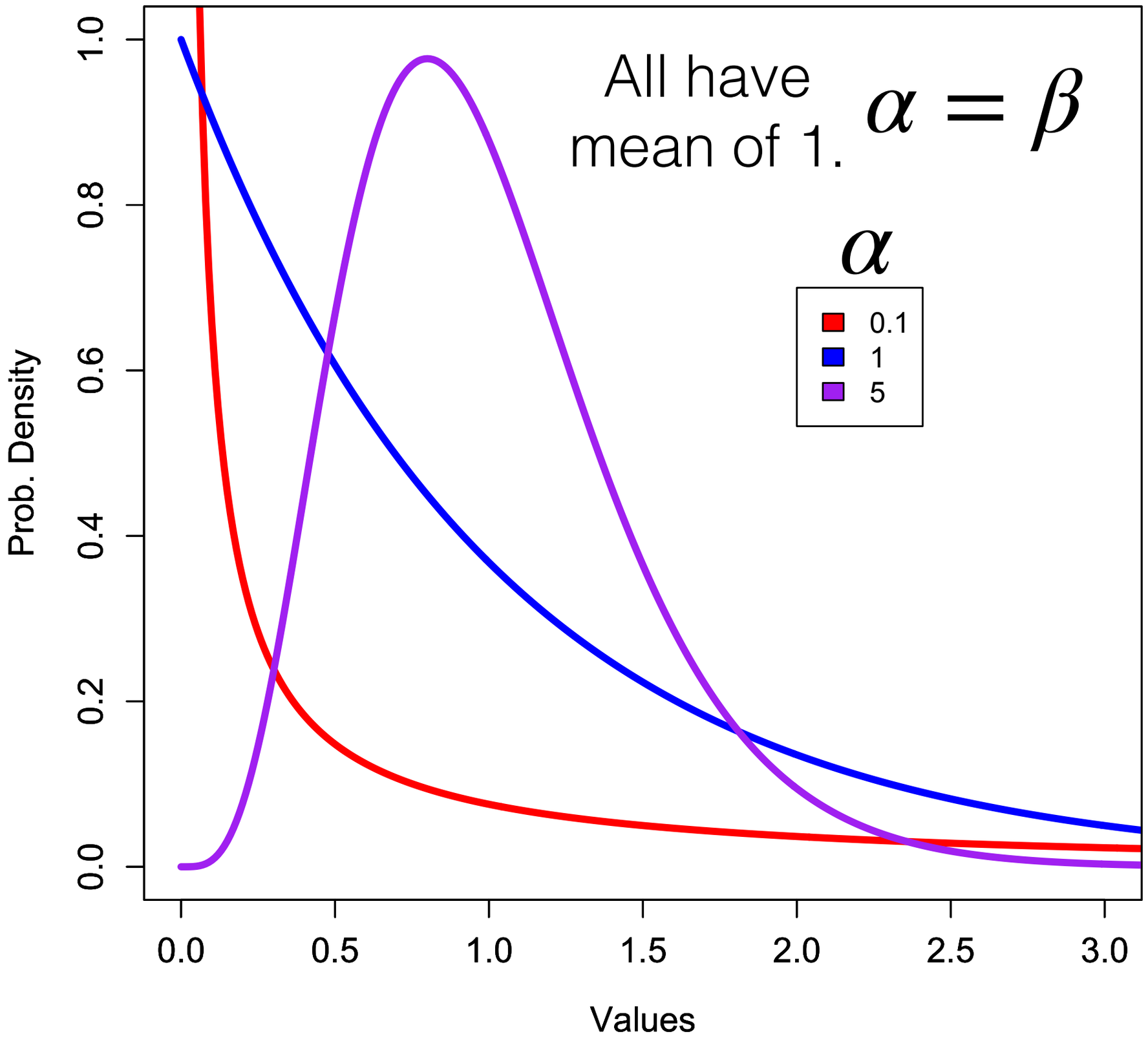
$$P(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad P(0.01) = \begin{bmatrix} 0.9901 & 0.0033 & 0.0033 & 0.0033 \\ 0.0033 & 0.9901 & 0.0033 & 0.0033 \\ 0.0033 & 0.0033 & 0.9901 & 0.0033 \\ 0.0033 & 0.0033 & 0.0033 & 0.9901 \end{bmatrix}$$

$$P(0.1) = \begin{bmatrix} 0.9064 & 0.0312 & 0.0312 & 0.0312 \\ 0.0312 & 0.9064 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.9064 & 0.0312 \\ 0.0312 & 0.0312 & 0.0312 & 0.9064 \end{bmatrix} \quad P(1) = \begin{bmatrix} 0.4477 & 0.1841 & 0.1841 & 0.1841 \\ 0.1841 & 0.4477 & 0.1841 & 0.1841 \\ 0.1841 & 0.1841 & 0.4477 & 0.1841 \\ 0.1841 & 0.1841 & 0.1841 & 0.4477 \end{bmatrix}$$

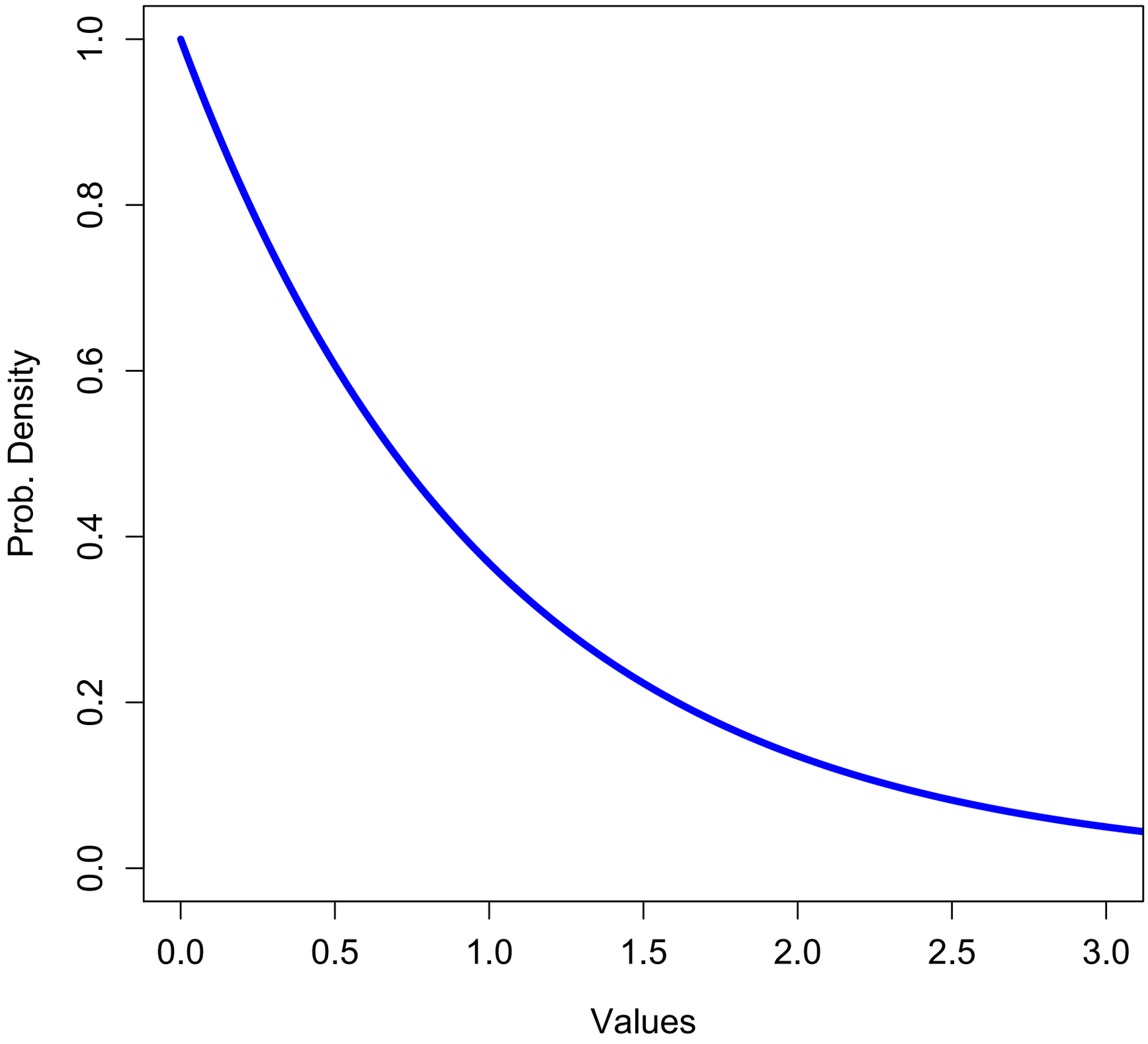
$$P(3) = \begin{bmatrix} 0.2637 & 0.2454 & 0.2454 & 0.2454 \\ 0.2454 & 0.2637 & 0.2454 & 0.2454 \\ 0.2454 & 0.2454 & 0.2637 & 0.2454 \\ 0.2454 & 0.2454 & 0.2454 & 0.2637 \end{bmatrix}$$


Jukes-Cantor
$P(A \mid A)$
$P(G \mid A)$
Prob. Change
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Time

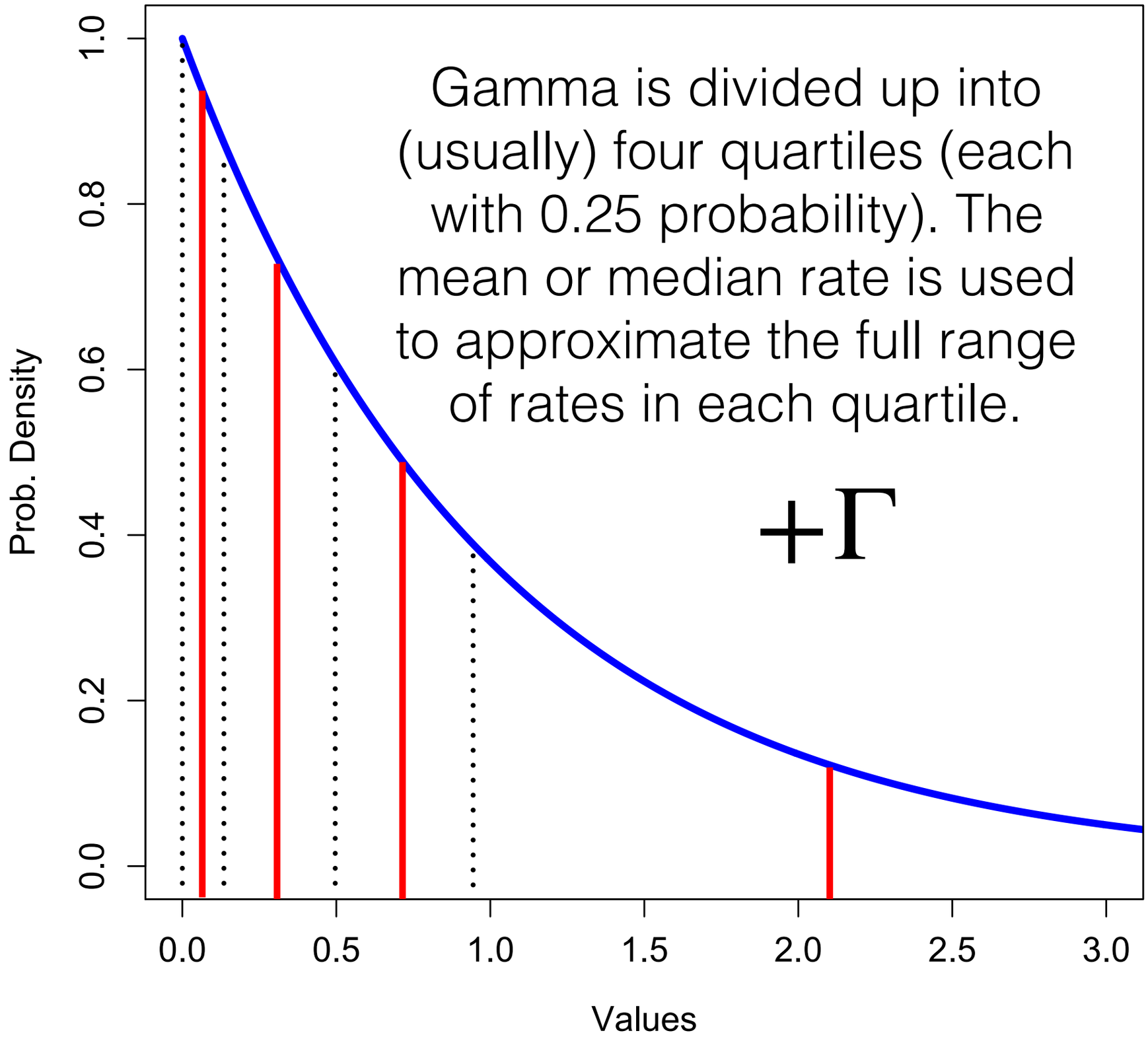
Rate Variation Across Sites (Gamma)



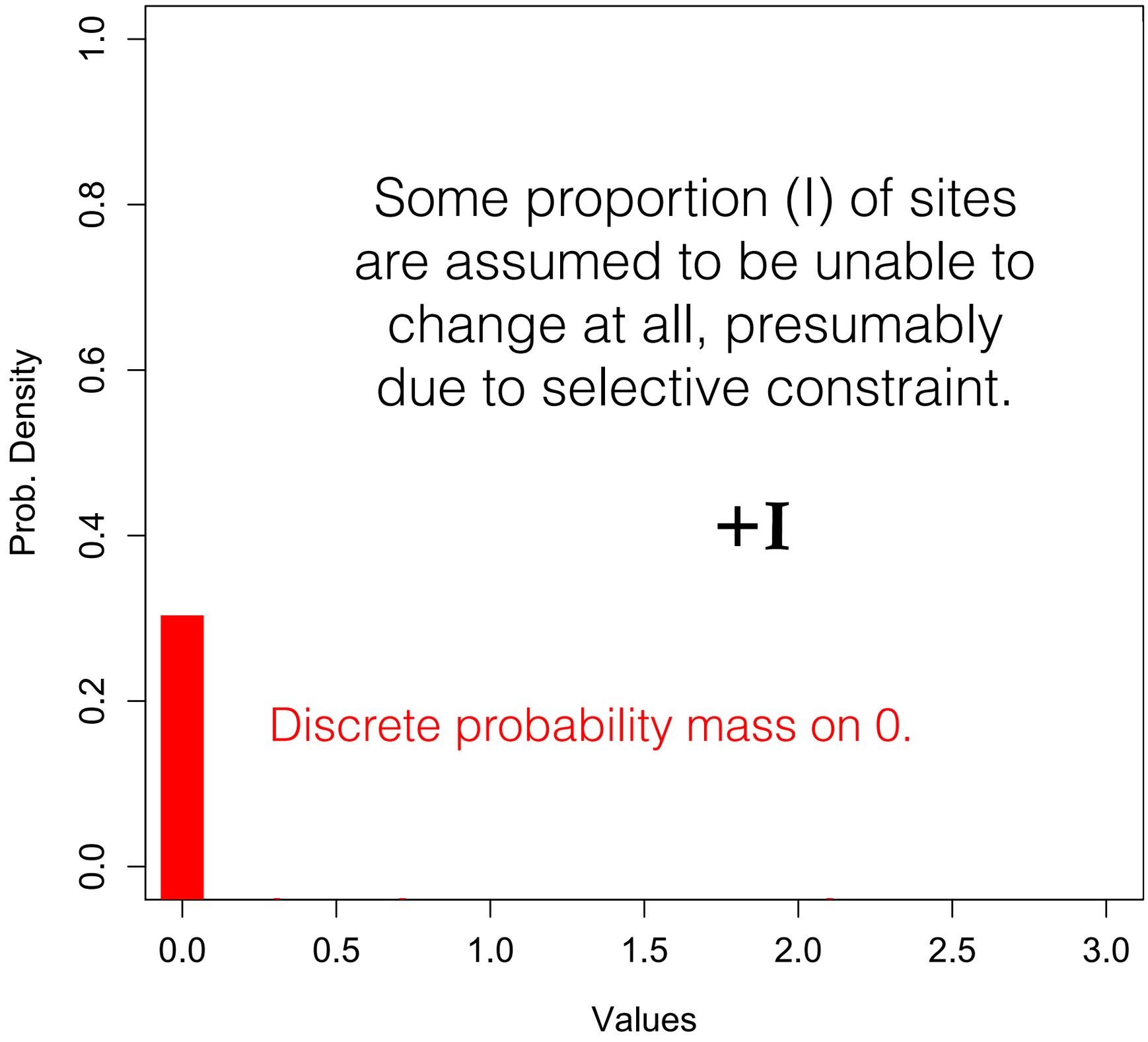
Rate Variation Across Sites (Gamma)



Rate Variation Across Sites (Discrete Gamma)

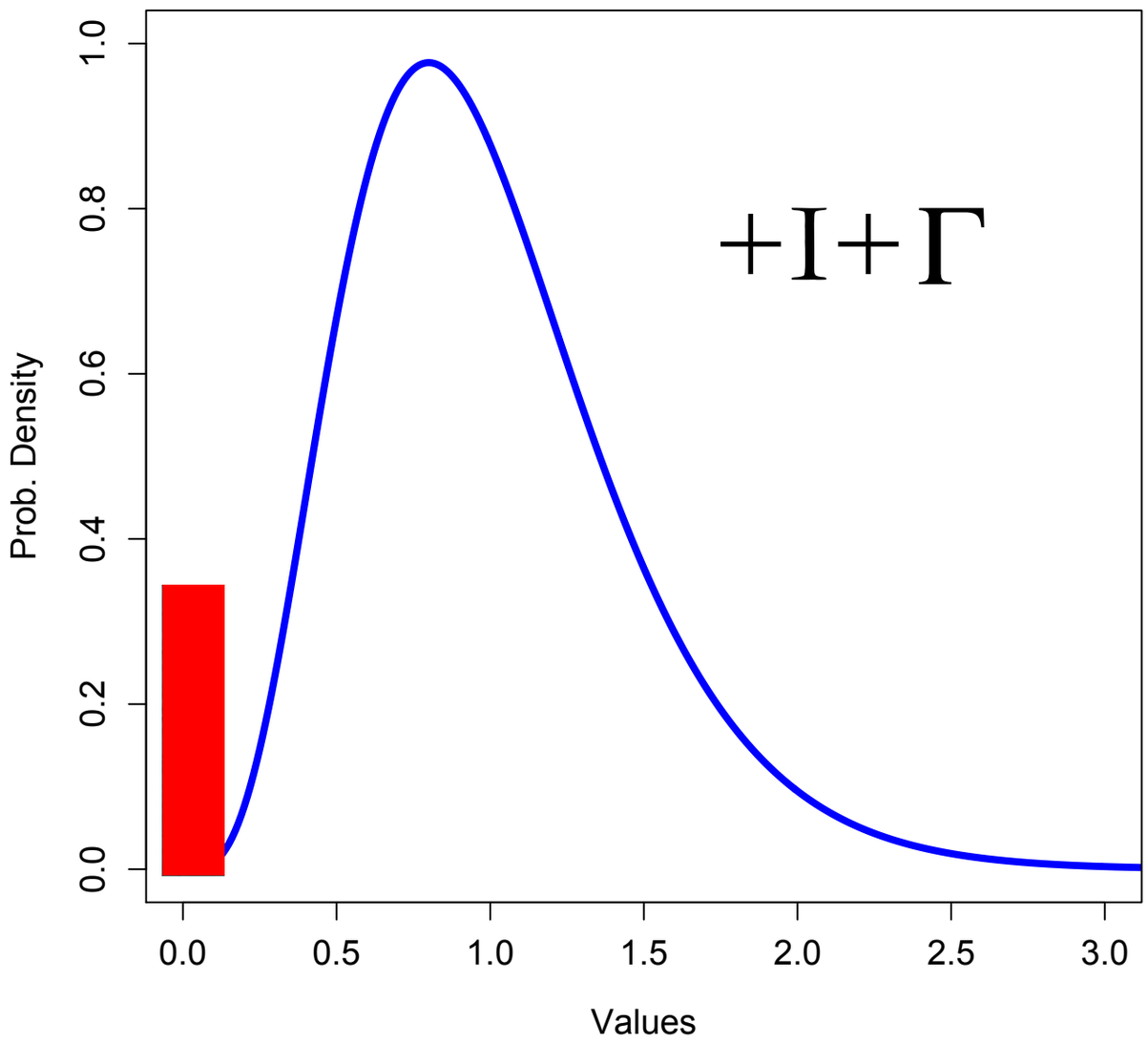


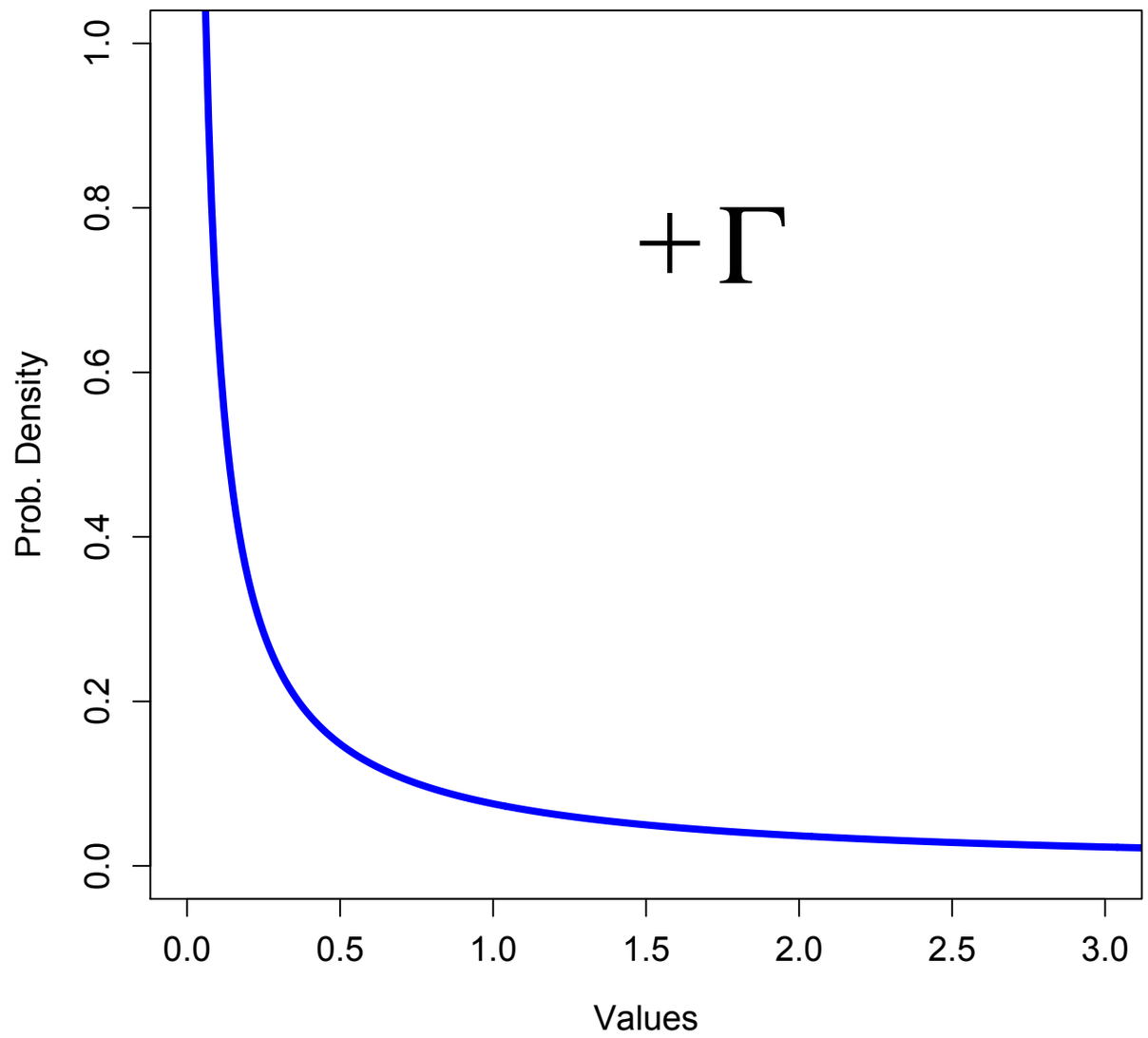
Rate Variation Across Sites (Proportion Invariable Sites)



Do these make **distinct** predictions about how rates of evolution vary across sites?

Challenge of **identifiability**.





Using models for: **Prediction** v Inference

If you know you have a fair coin (p=0.5), you can predict whether or not you should expect to observe different numbers of heads and tails.

$$P(k = 0 \,|\, n = 10, p = 0.5) = 0.001$$
$$P(k = 1 \,|\, n = 10, p = 0.5) = 0.010$$
$$P(k = 2 \,|\, n = 10, p = 0.5) = 0.044$$
$$P(k = 3 \,|\, n = 10, p = 0.5) = 0.117$$
$$P(k = 4 \,|\, n = 10, p = 0.5) = 0.205$$

...

Using models for: Prediction v **Inference**

But, what if instead of knowing (p=0.5), you know that k=3 and now you want to learn something about p?

Using models for: Prediction v **Inference**

But, what if instead of knowing (p=0.5), you know that k=3 and now you want to learn something about p?

Learning about some aspect of the process that is generating the data based on observed outcomes of that process is known as **inference**!

Even when we don't formalize it, we do this every day.

Using models for: Prediction v **Inference**

One way to think about this is to flip the conditionality in our previous probability statement:

$$P(p = 0.5 \,|\, n = 10, k = 3)$$

Using models for: Prediction v **Inference**

One way to think about this is to flip the conditionality in our previous probability statement:

$$P(p = 0.5 \,|\, n = 10, k = 3)$$

This is Bayesian inference!

However, because Bayesian inference associates probabilities with “belief” and requires statements of “prior belief”, it has made many people uncomfortable (R.A. Fisher included).

Maximum **Likelihood** Inference

To avoid the uncomfortable parts of Bayesian inference, Fisher invented *likelihood*.

Note that while *likelihood* and *probability* are used interchangeably in every day language, they mean distinct things in statistical inference.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\mathscr{L}(p = 0.1; n = 10, k = 3) = 0.057$$
$$\mathscr{L}(p = 0.2; n = 10, k = 3) = 0.201$$
$$\mathscr{L}(p = 0.3; n = 10, k = 3) = 0.267$$
$$\mathscr{L}(p = 0.4; n = 10, k = 3) = 0.215$$
$$\mathscr{L}(p = 0.5; n = 10, k = 3) = 0.117$$
$$\ldots$$

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\mathscr{L}(p = 0.1; n = 10, k = 3) = 0.057$$
$$\mathscr{L}(p = 0.2; n = 10, k = 3) = 0.201$$
$$\mathscr{L}(p = 0.3; n = 10, k = 3) = 0.267$$
$$\mathscr{L}(p = 0.4; n = 10, k = 3) = 0.215$$
$$\mathscr{L}(p = 0.5; n = 10, k = 3) = 0.117$$
$$\ldots$$

Now, *k* stays the same and *p* changes.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Looking at these likelihoods, what is different than when we looked at the list of conditional probabilities a few slides ago?

$$\begin{aligned}
\mathscr{L}(p = 0.1; n = 10, k = 3) &= 0.057 \\
\mathscr{L}(p = 0.2; n = 10, k = 3) &= 0.201 \\
\mathscr{L}(p = 0.3; n = 10, k = 3) &= 0.267 \\
\mathscr{L}(p = 0.4; n = 10, k = 3) &= 0.215 \\
\mathscr{L}(p = 0.5; n = 10, k = 3) &= 0.117
\end{aligned}$$

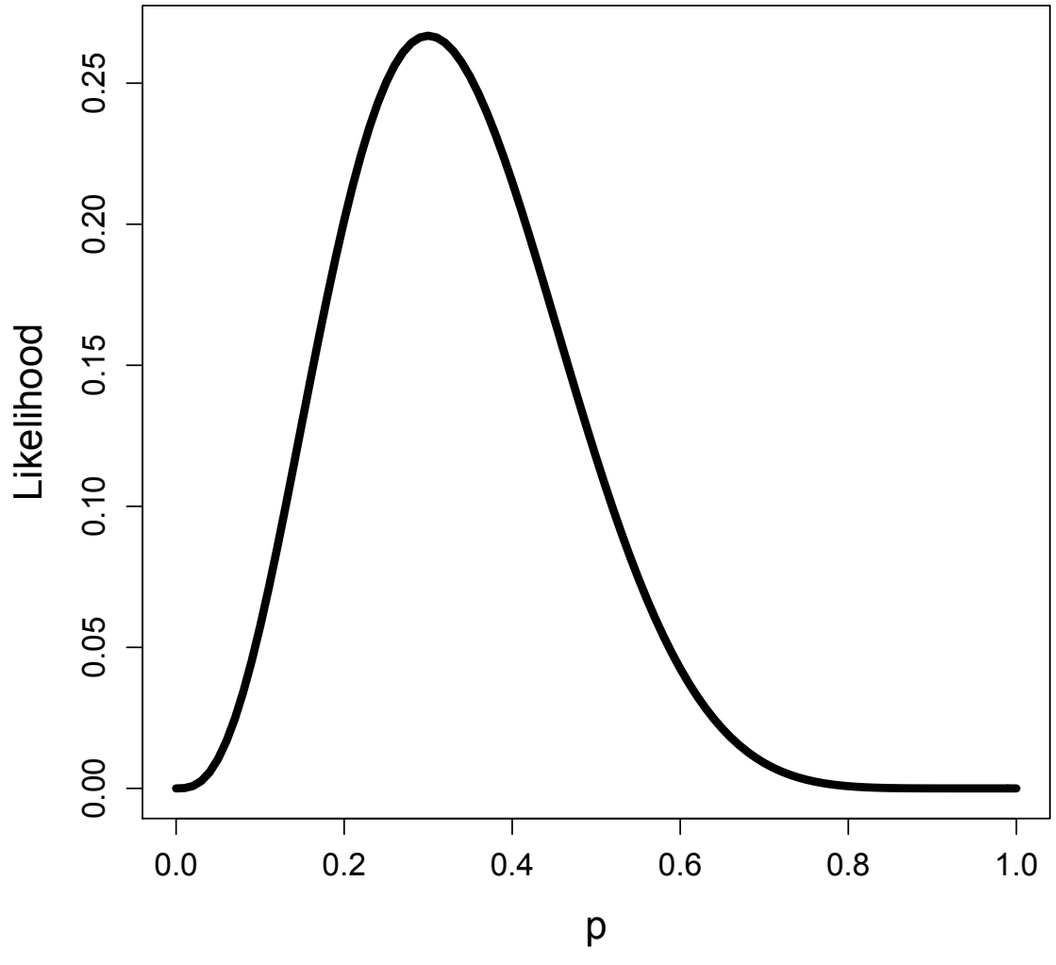
...

Also, likelihoods don't sum or integrate to 1.

Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

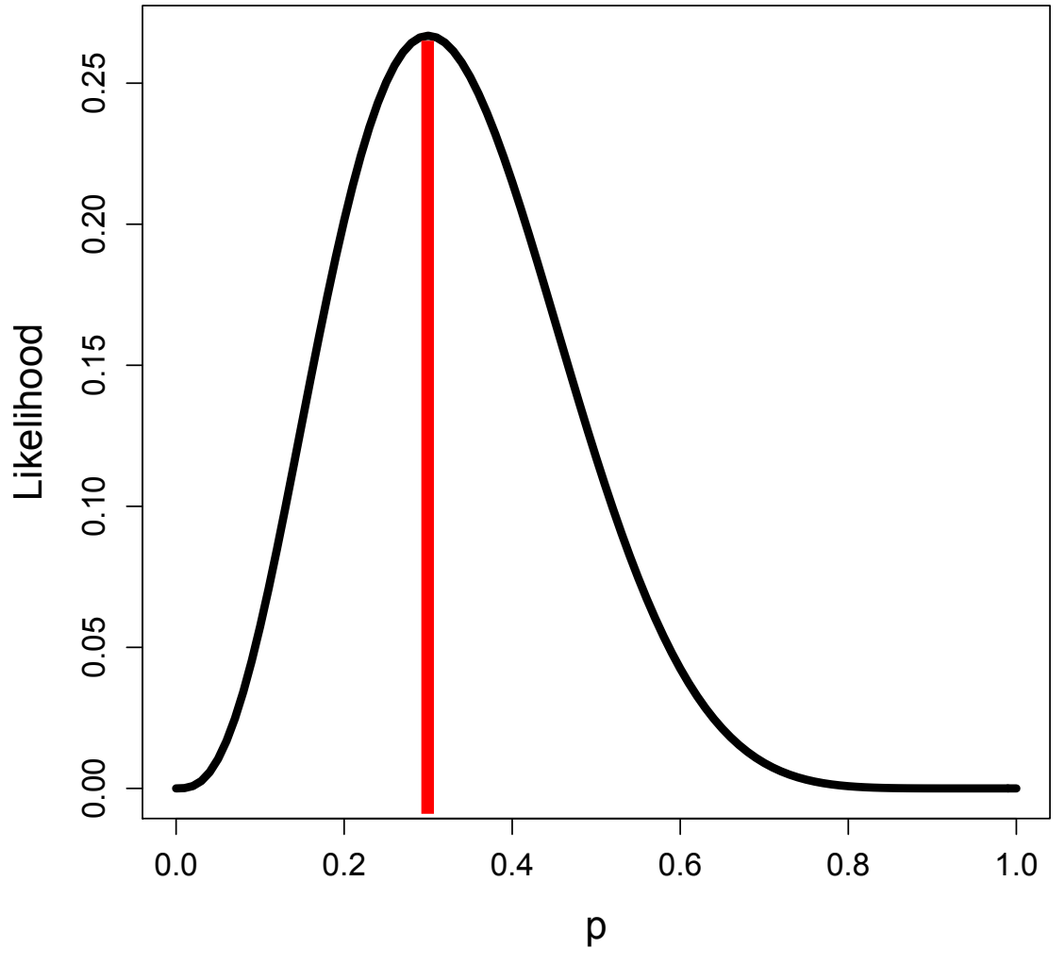
Fisher (and others) argued that likelihood is the most objective and principled way to evaluate evidence about parameters and models.



Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

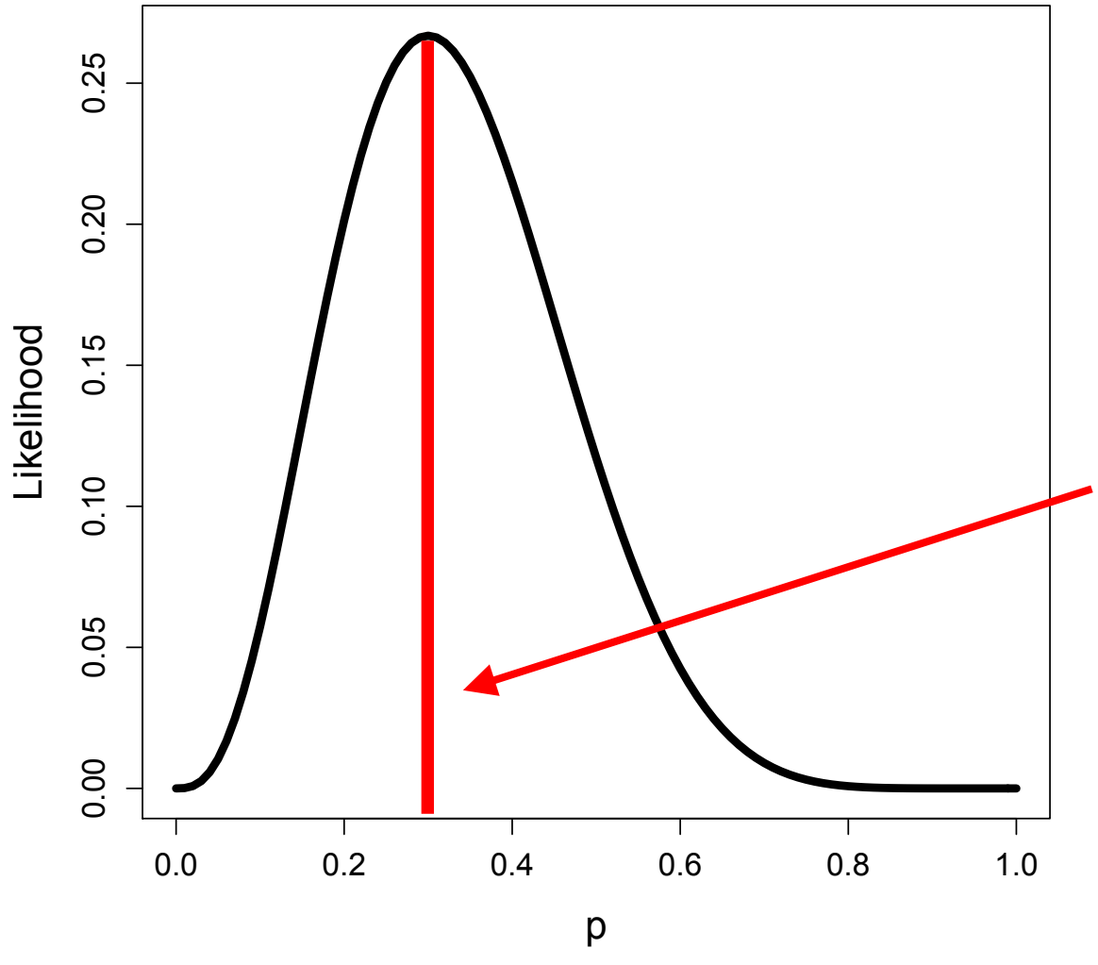
The value of a parameter that maximizes the likelihood is known as the **maximum-likelihood estimate (MLE)** of the "true" parameter value.



Maximum **Likelihood** Inference

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

The value of a parameter that maximizes the likelihood is known as the **maximum-likelihood estimate (MLE)** of the "true" parameter value.

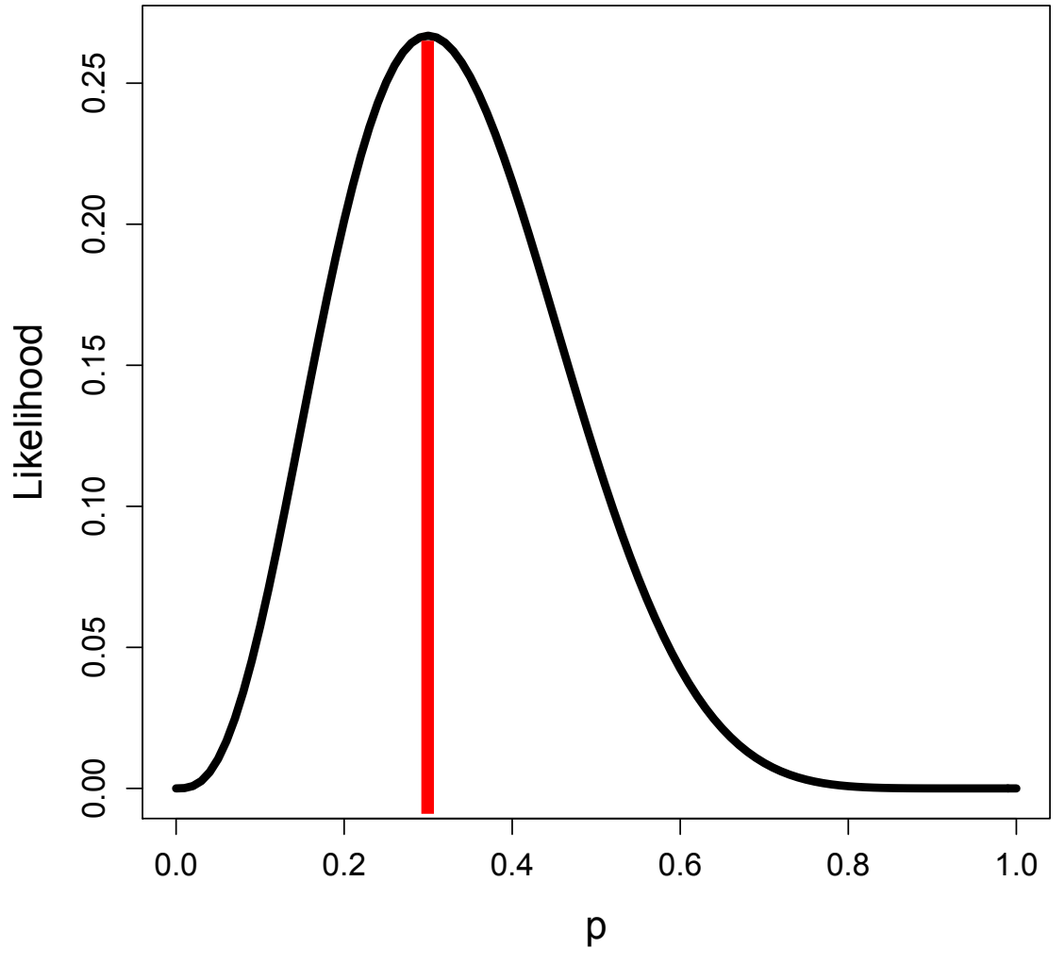


Can you guess the MLE for 3 heads in 10 coin flips?

Hill Climbing

$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

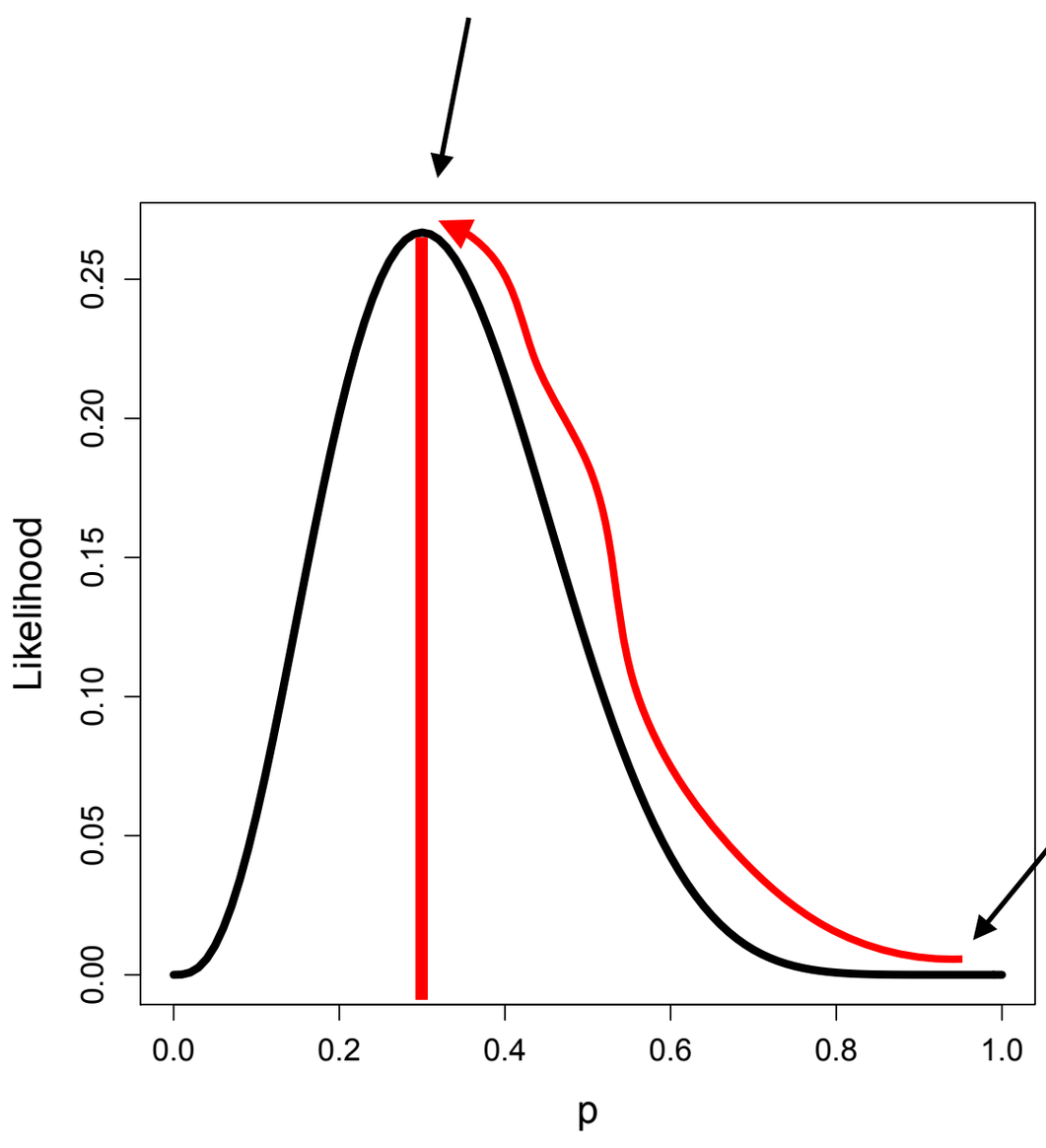
If we have a likelihood function, we will generally use some kind of hill climbing algorithm (there are lots!) to find the MLE.



Hill Climbing

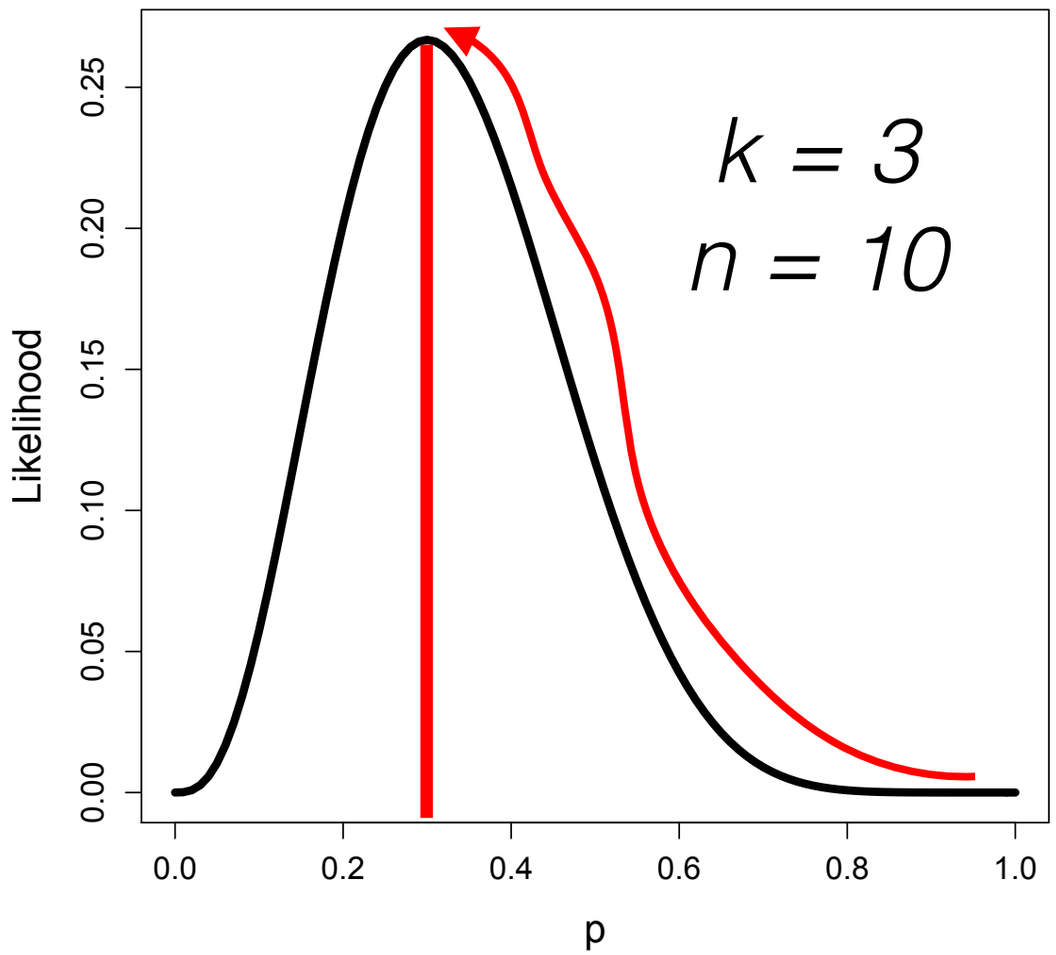
$$\mathscr{L}(p; n, k) = P(k \mid n, p)$$

Estimate of the MLE



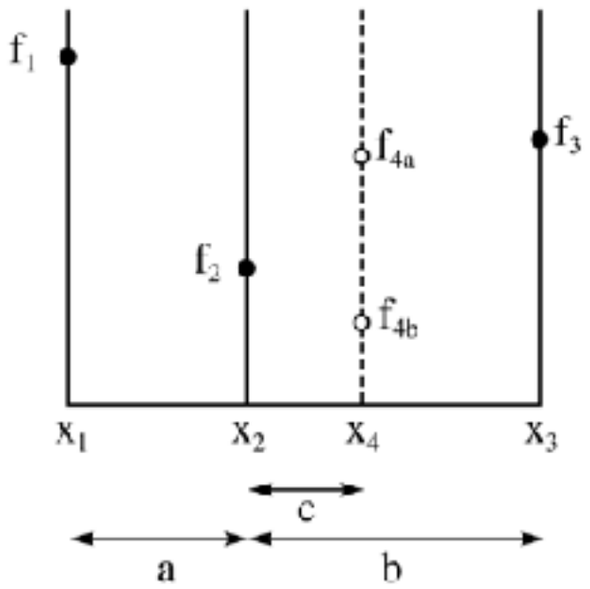
Initial guess or starting *p*.

ML Optimization



For simple cases, like the binomial shown here, analytical solutions exist to find the MLE. However, we are going to largely ignore that theory for now, because such solutions generally don't exist for phylogenetic problems. In these more complicated cases, we use stochastic search algorithms to climb hills in the likelihood landscape.

ML Optimization

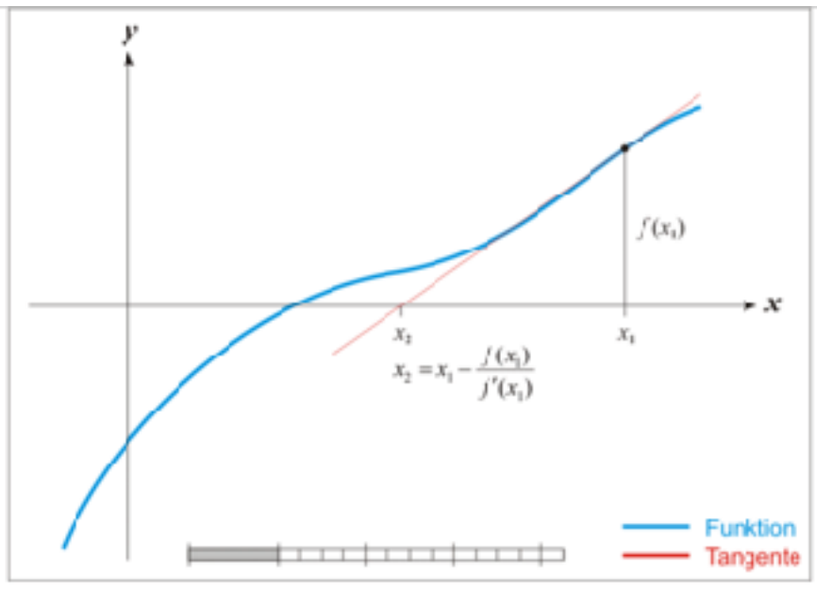


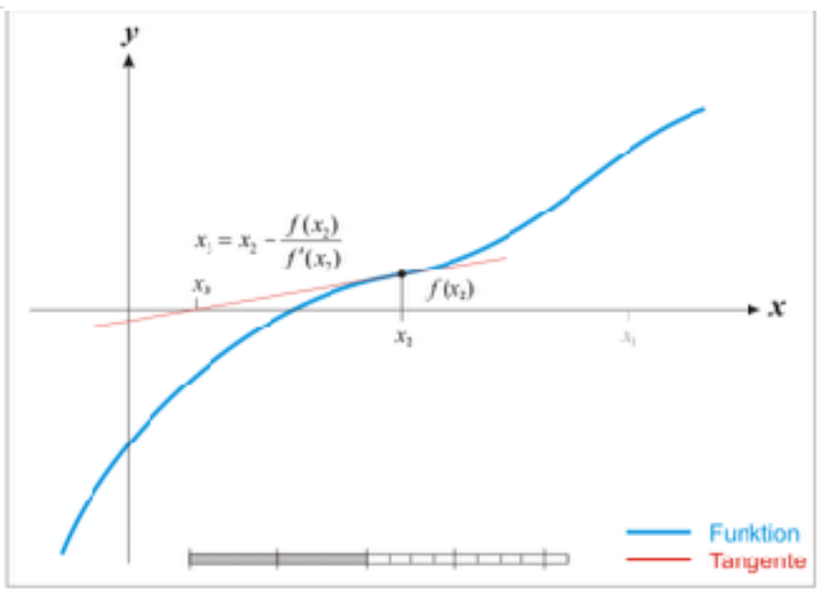
- Golden Section Search

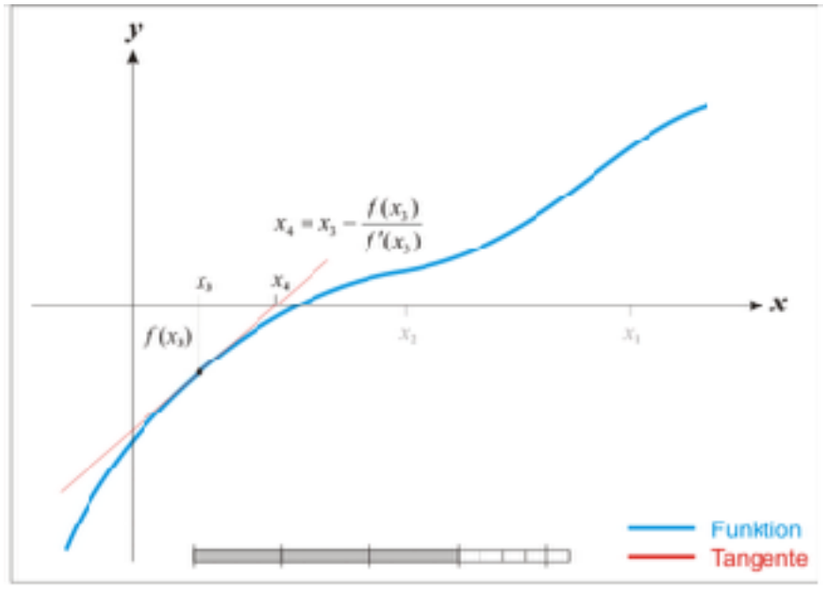
  If minimum is in $[f_1, f_3]$, we can reduce this interval by looking at f_2 and f_4. If $f_4 > f_2$, then minimum is in $[f_1, f_4]$. Else, minimum is in $[f_2, f_3]$.

- Newton-Raphson Optimization

  Approximate likelihood function with a polynomial. Then use its derivatives to estimate where first derivative of polynomial is 0. Keep iterating until you converge.

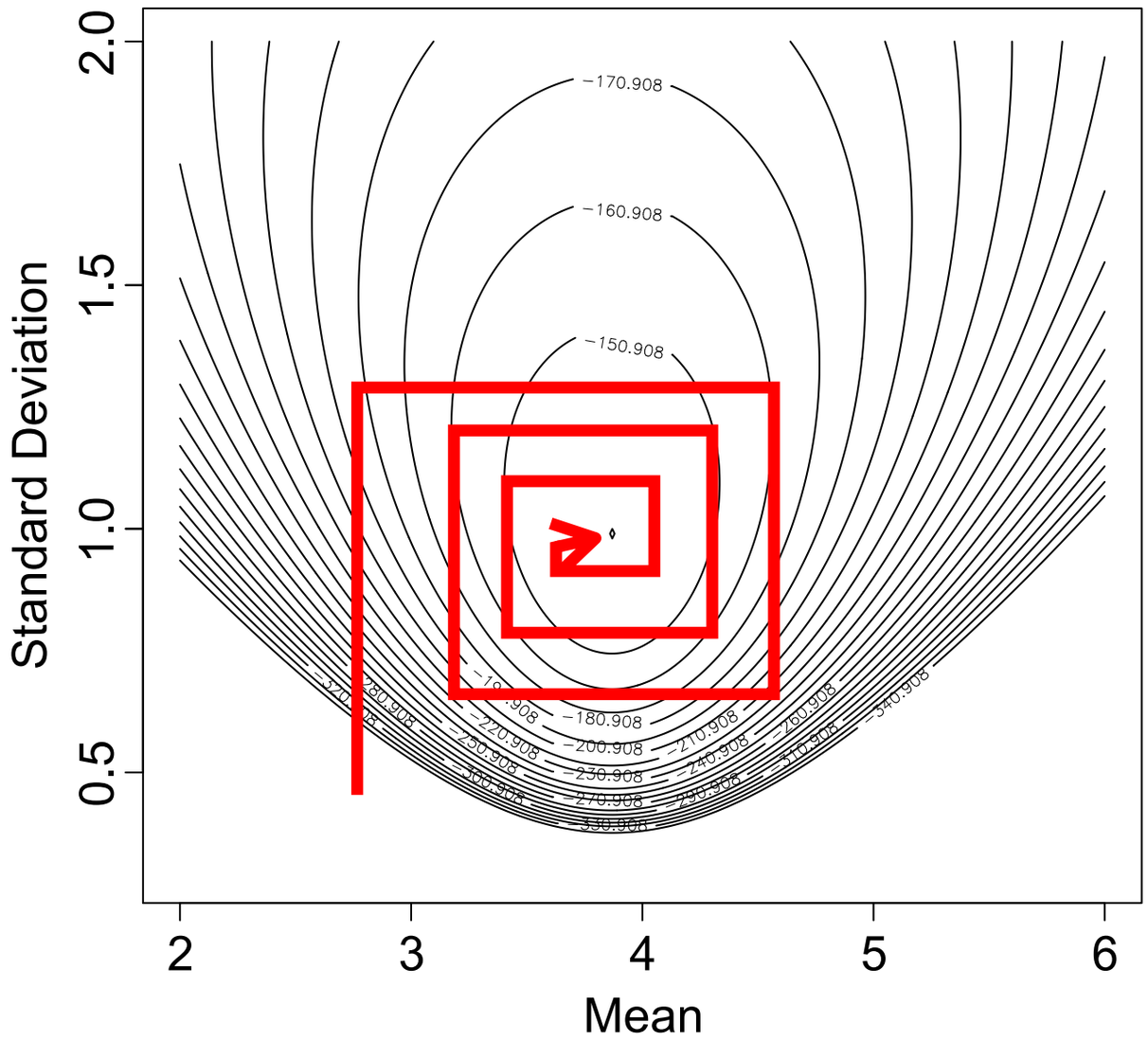


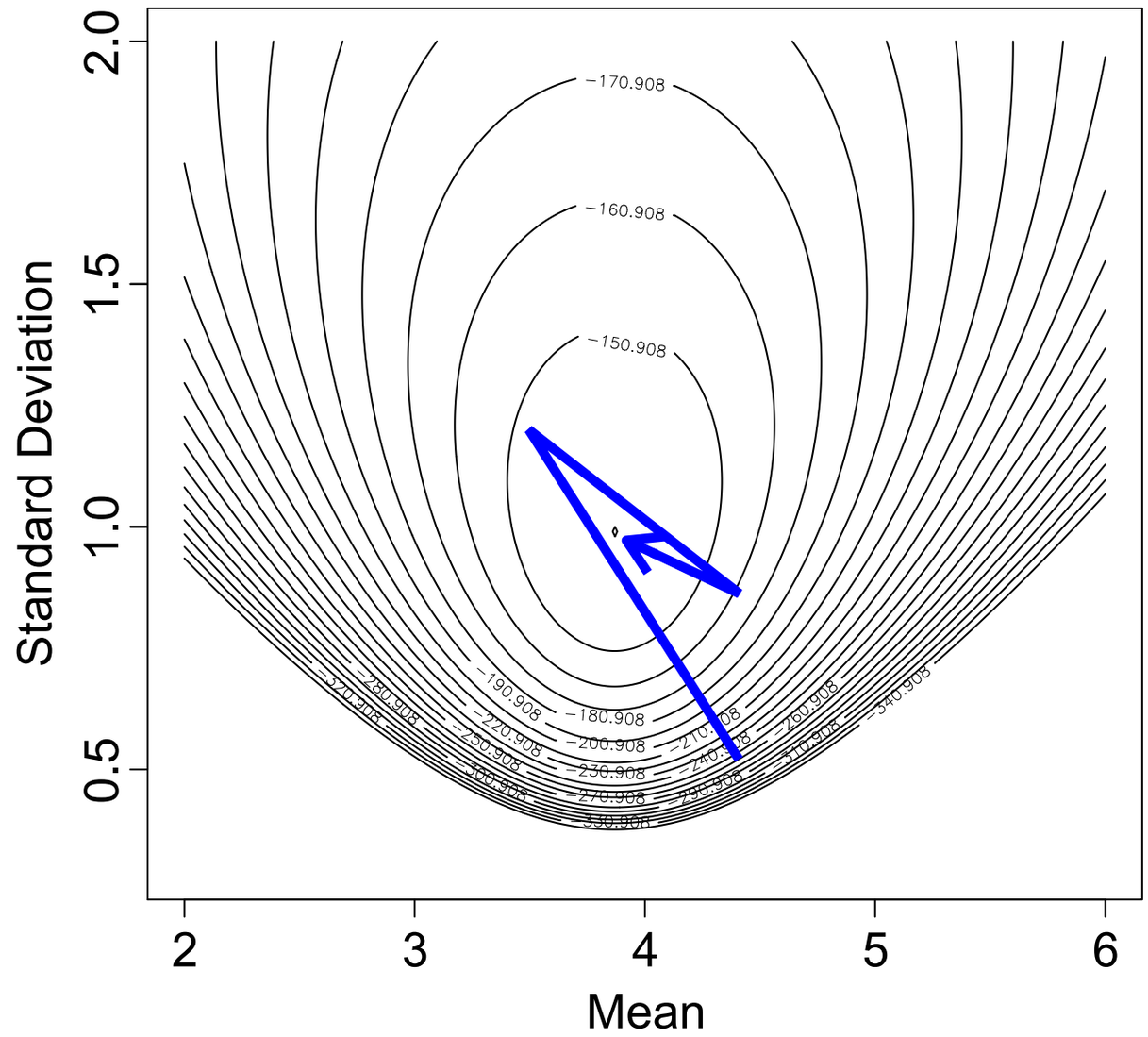




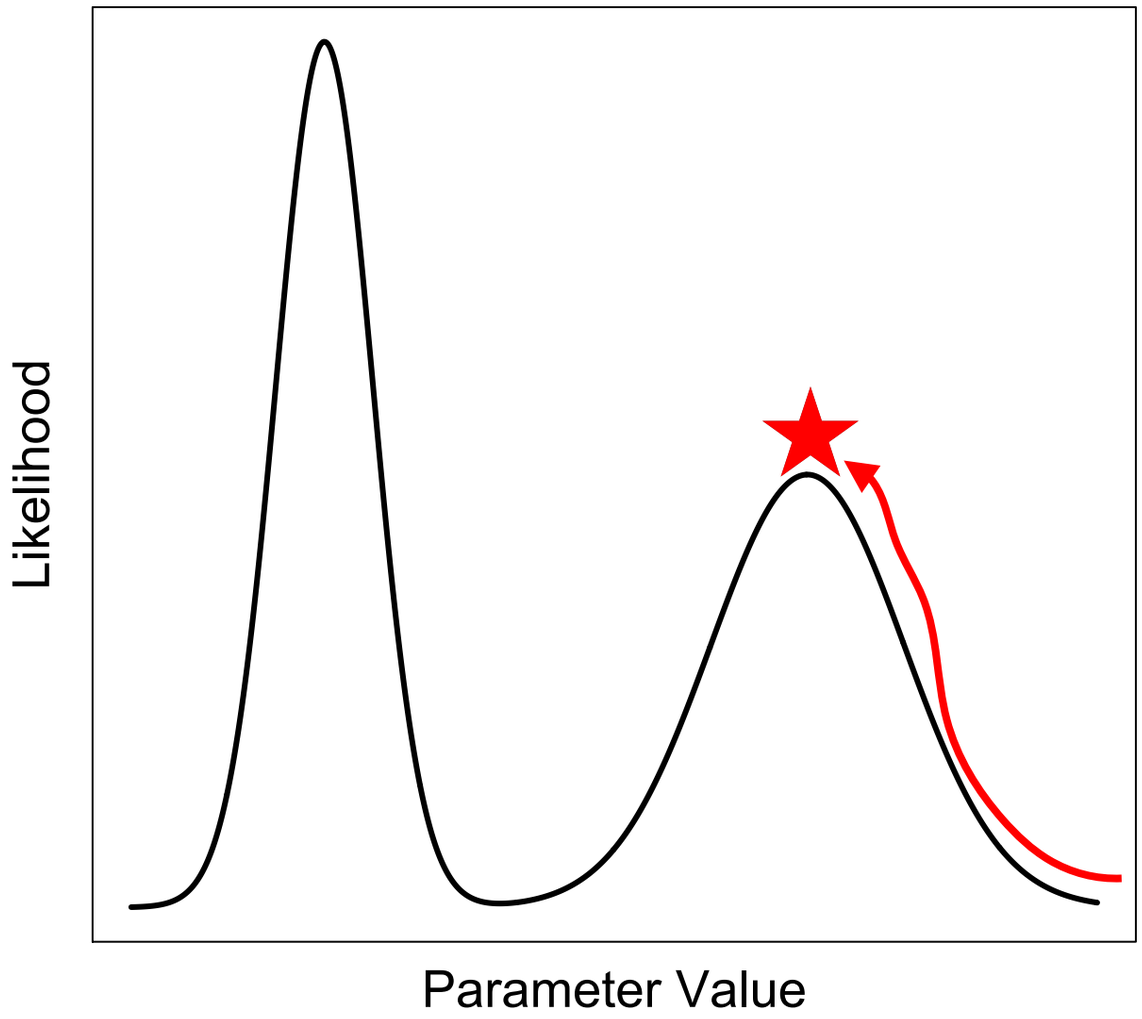
ML Optimization (Multiple Parameters)

Similar to methods for one parameter, but need to consider potential correlations among parameters in order to be efficient.



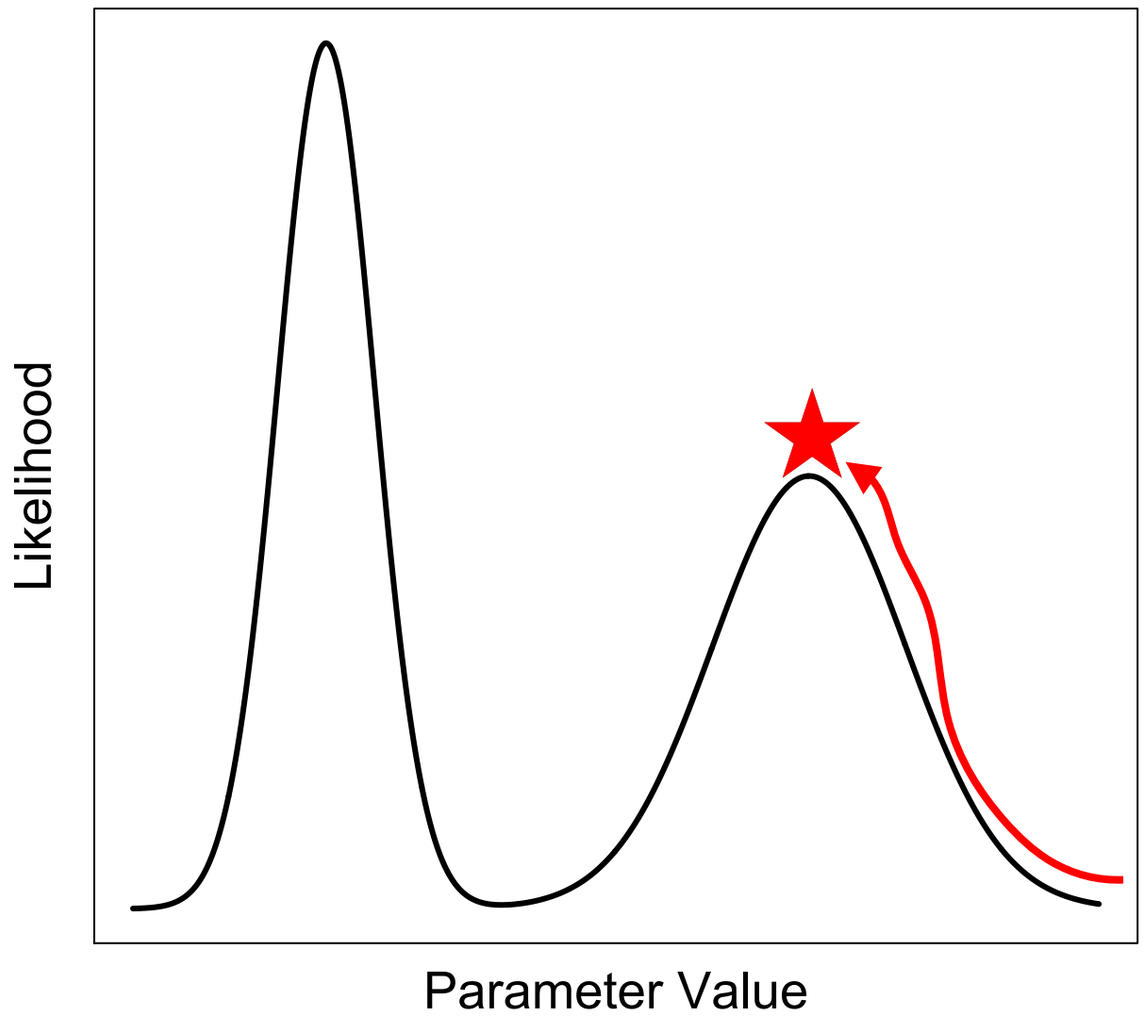


ML Optimization - WARNING!



What happened to our hill-climbing algorithm here?
Why is this a problem?

ML Optimization - WARNING!



To avoid this in practice:

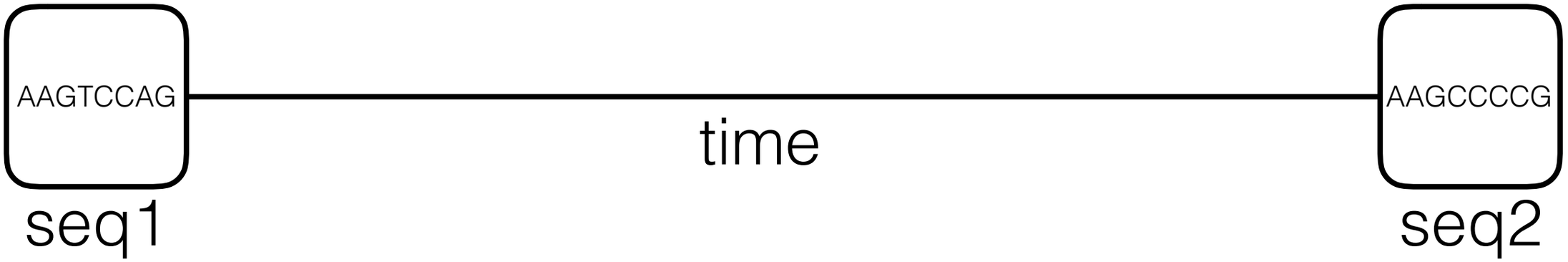
(1) Always use multiple searches with different starting points

(2) Use gut checks to make sure the answer make sense

(3) Use programs that employ robust algorithms

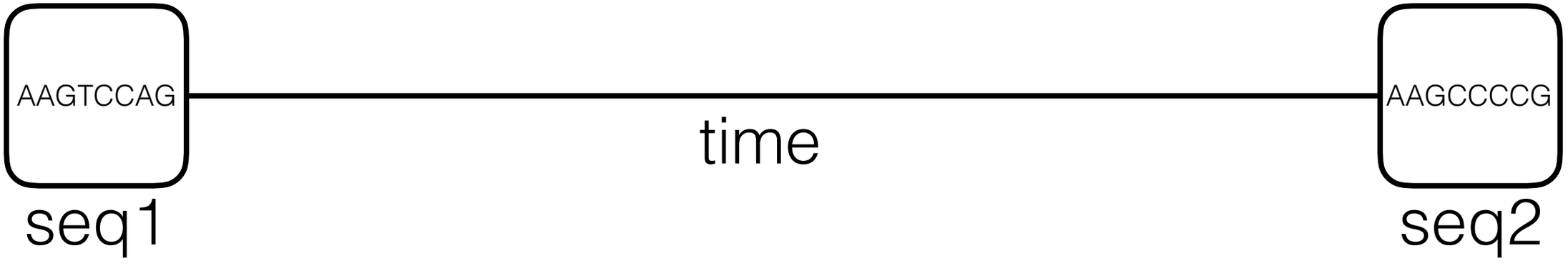
Maximum-likehood genetic distance estimation

$$\mathcal{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



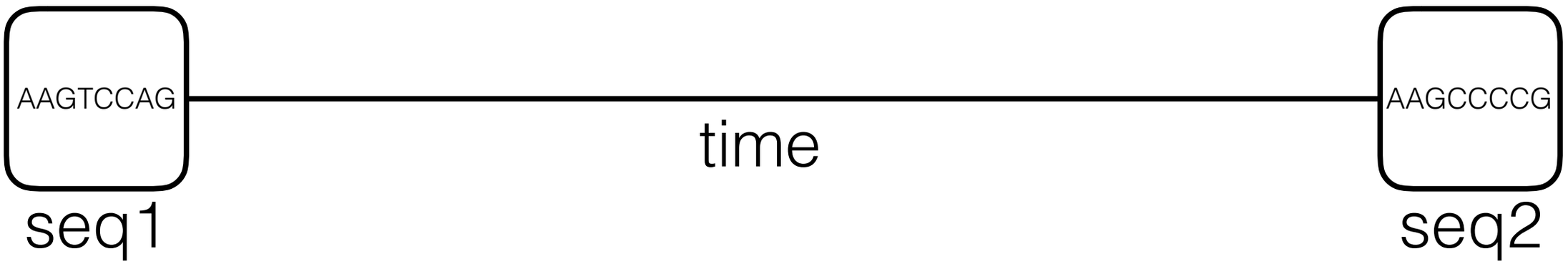
$$P(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

What is $\mathscr{L}(0; AAGTCCAG, AAGCCCCG)$?

Remember, sites are independent of one another!

Maximum-likehood genetic distance estimation

$$\mathscr{L}(t; seq1, seq2) = P(seq1, seq2 \mid t)$$



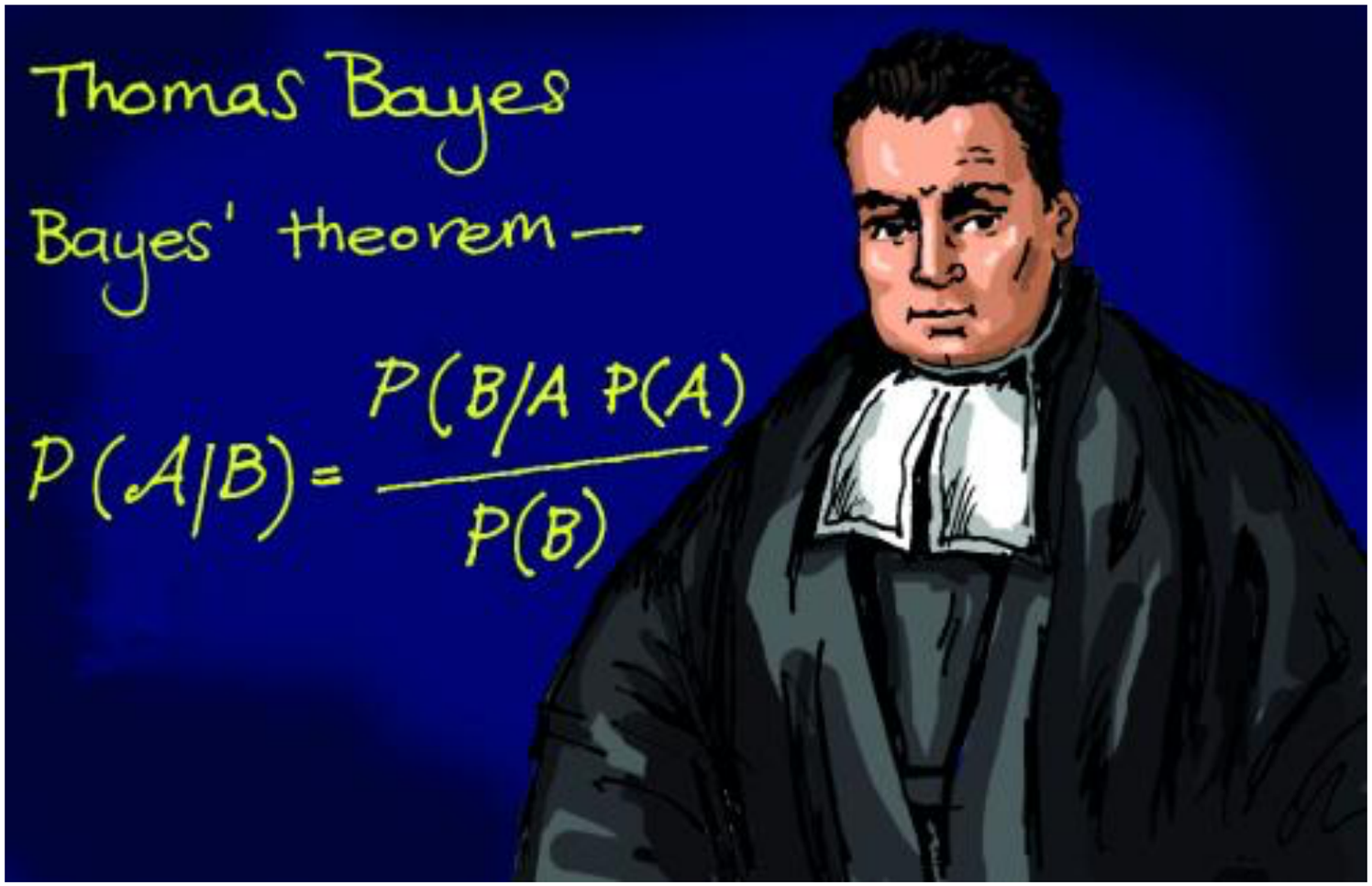
$$P(0.1) = \begin{bmatrix} 0.9064 & 0.0312 & 0.0312 & 0.0312 \\ 0.0312 & 0.9064 & 0.0312 & 0.0312 \\ 0.0312 & 0.0312 & 0.9064 & 0.0312 \\ 0.0312 & 0.0312 & 0.0312 & 0.9064 \end{bmatrix}$$

What is $\mathscr{L}(0.1; AAGTCCAG, AAGCCCCG)$?

RevBayes Exercise 8

Estimating a Jukes-Cantor Genetic Distance

An Overview of Bayesian Inference



http://www.pharmaceutical-journal.com/opinion/blogs/bayess-theorem-falls-foul-of-the-law/11089087.blog

Revisiting Some Facts About Probability

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

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

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

Bayes Theorem

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

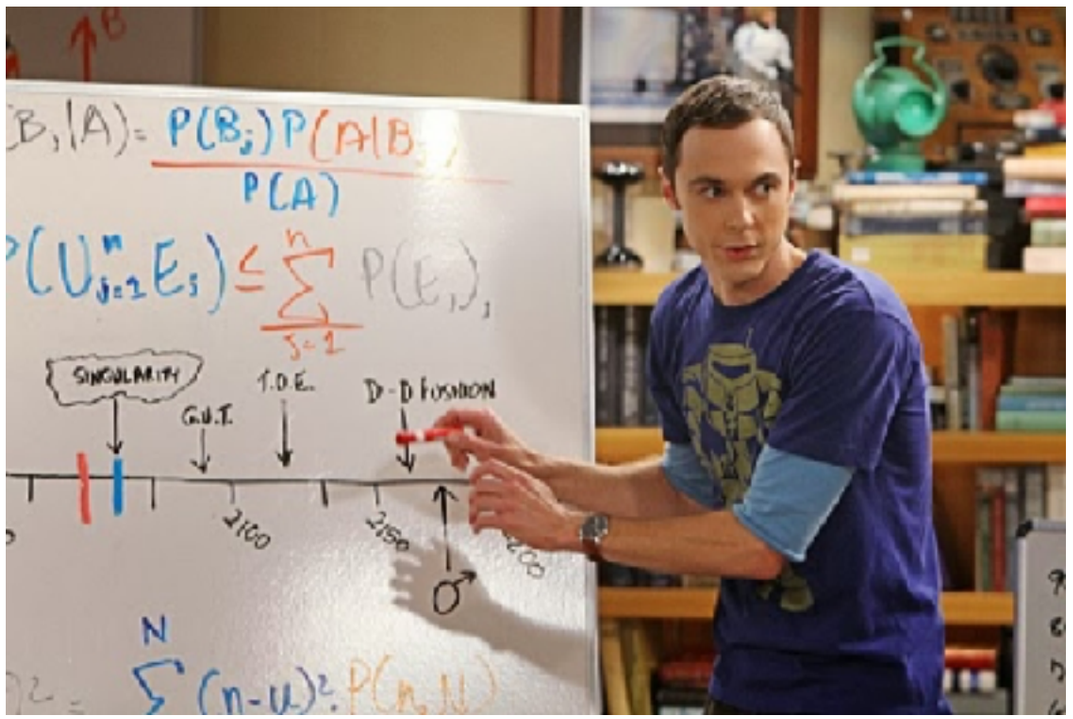
T. Bayes.

https://en.wikipedia.org/wiki/Thomas_Bayes

Bayes Theorem

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

T. Bayes.

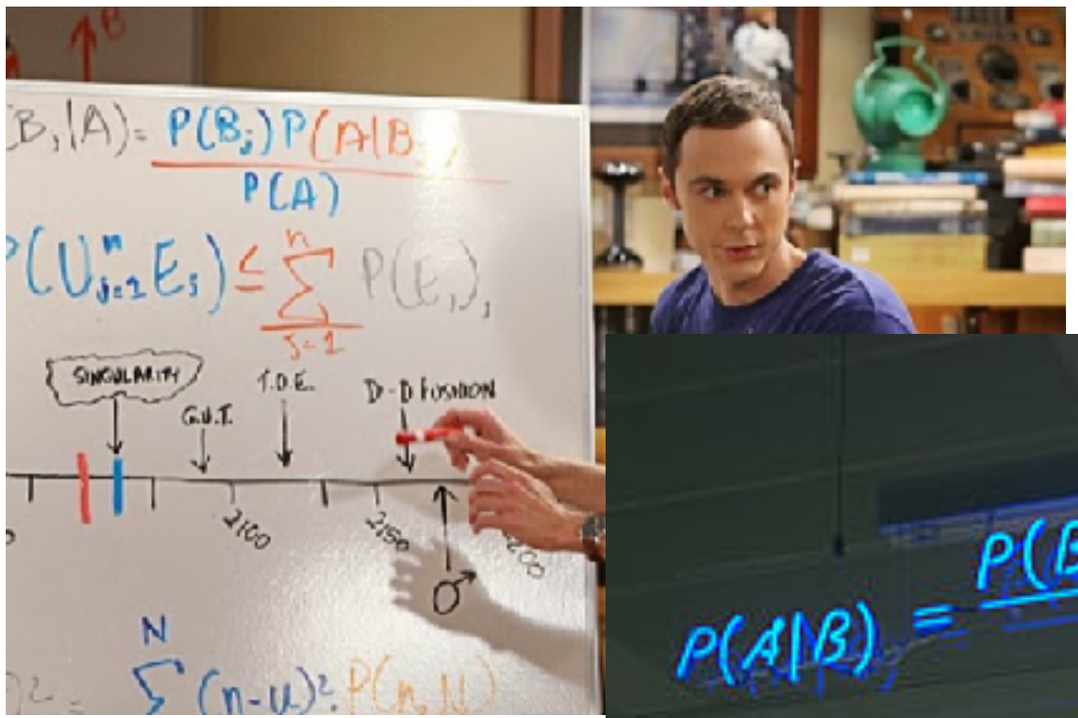


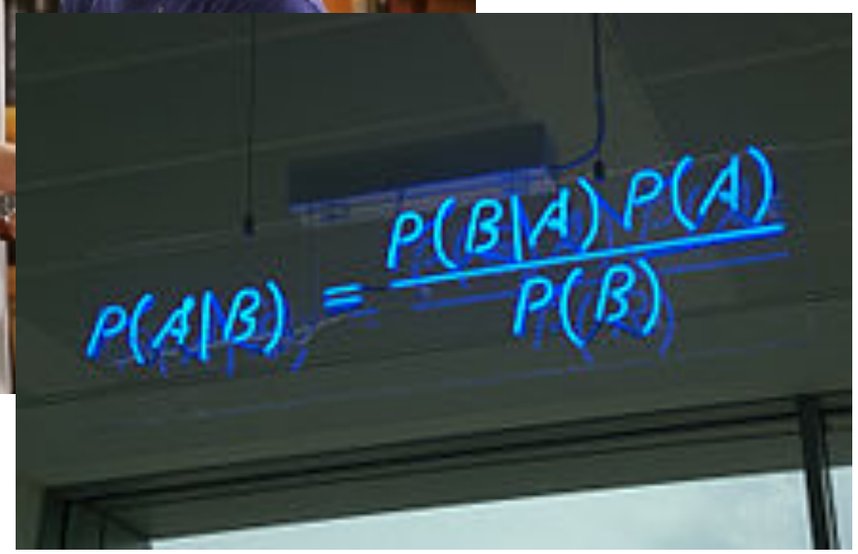
https://en.wikipedia.org/wiki/Thomas_Bayes

Bayes Theorem

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

T. Bayes.



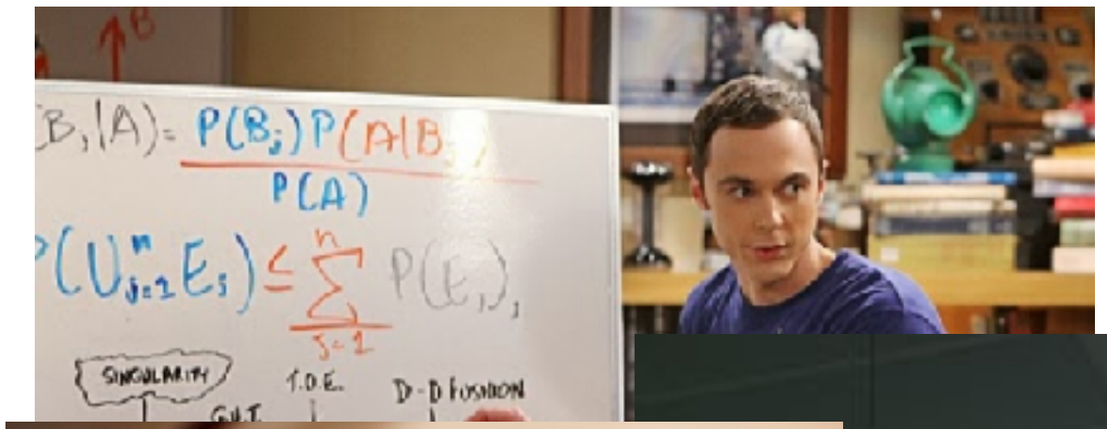


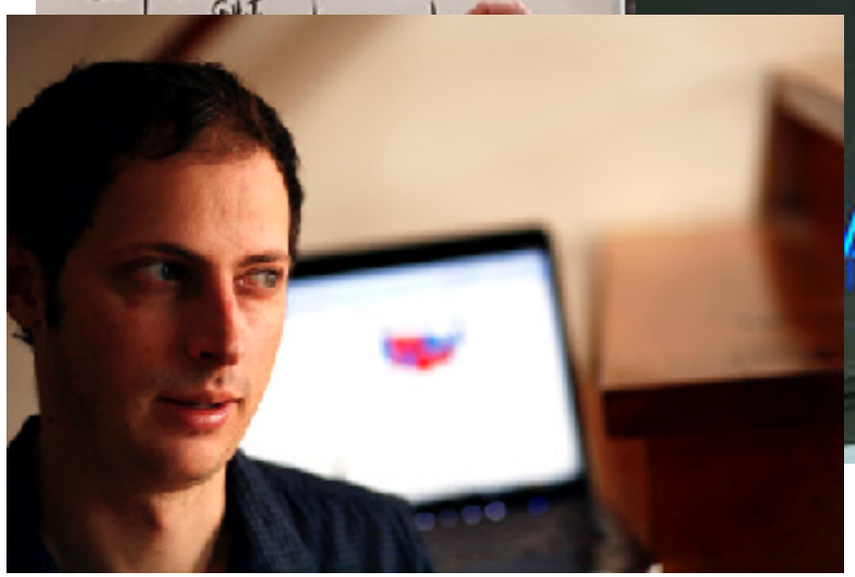
https://en.wikipedia.org/wiki/Thomas_Bayes

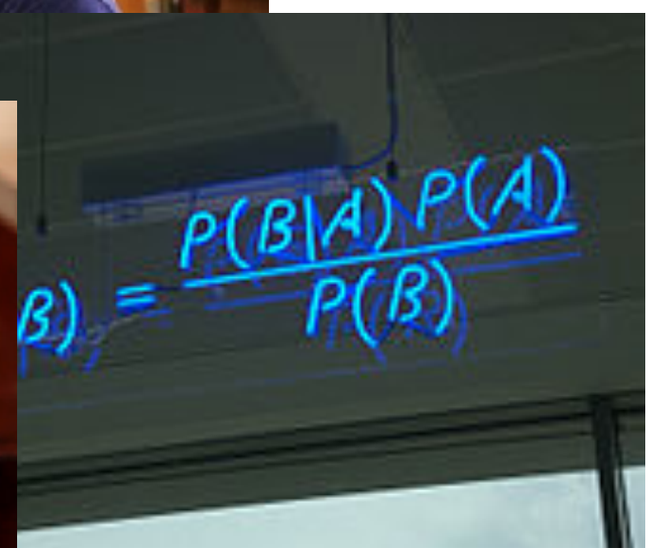
Bayes Theorem

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

T. Bayes.





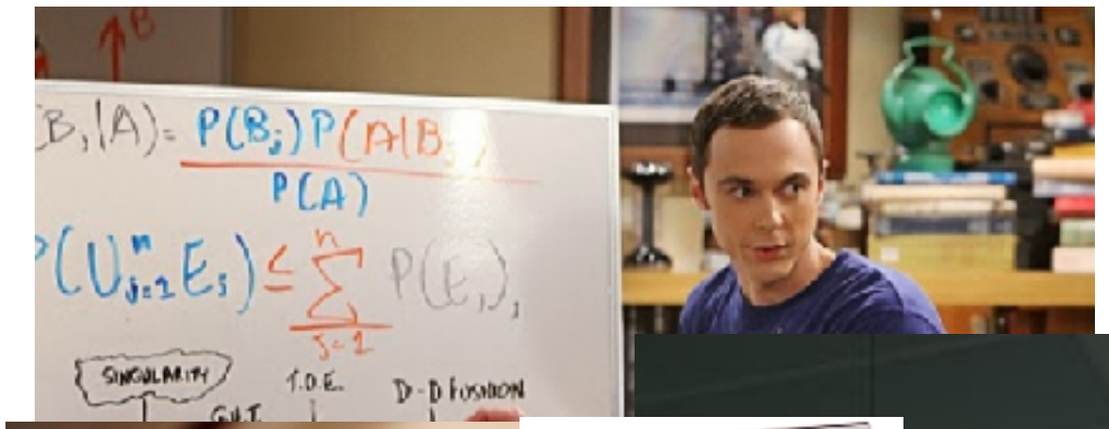


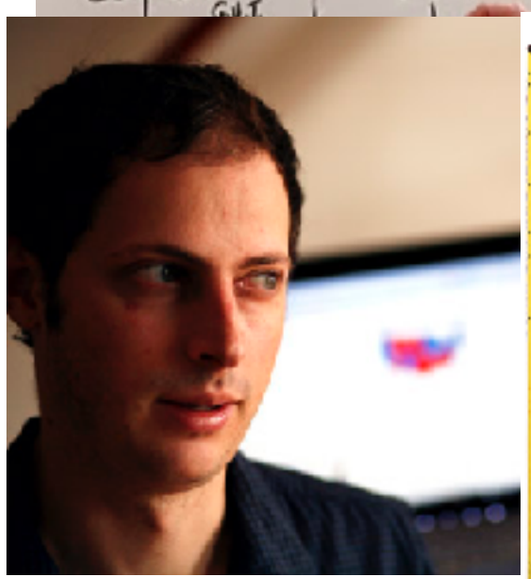
https://en.wikipedia.org/wiki/Thomas_Bayes

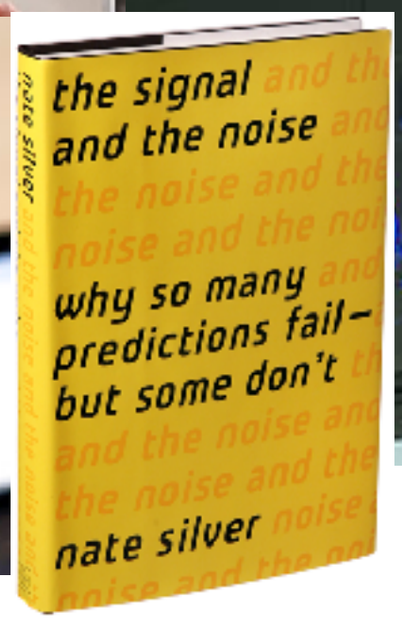
Bayes Theorem

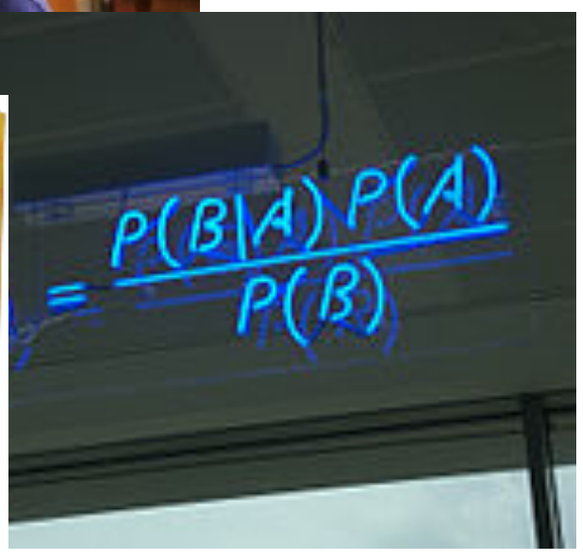
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

T. Bayes.





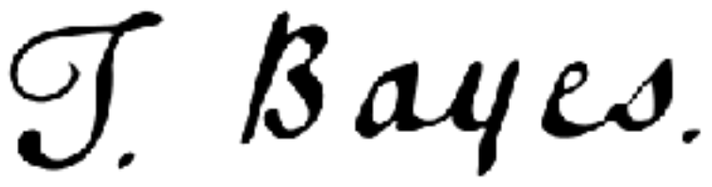




https://en.wikipedia.org/wiki/Thomas_Bayes

Bayes Theorem

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



Rev. Thomas Bayes
1701-1761

"An Essay towards solving a Problem in the Doctrine of Chances" published in 1763 (Richard Price)

Binomial with a uniform prior on *p*

https://en.wikipedia.org/wiki/Thomas_Bayes

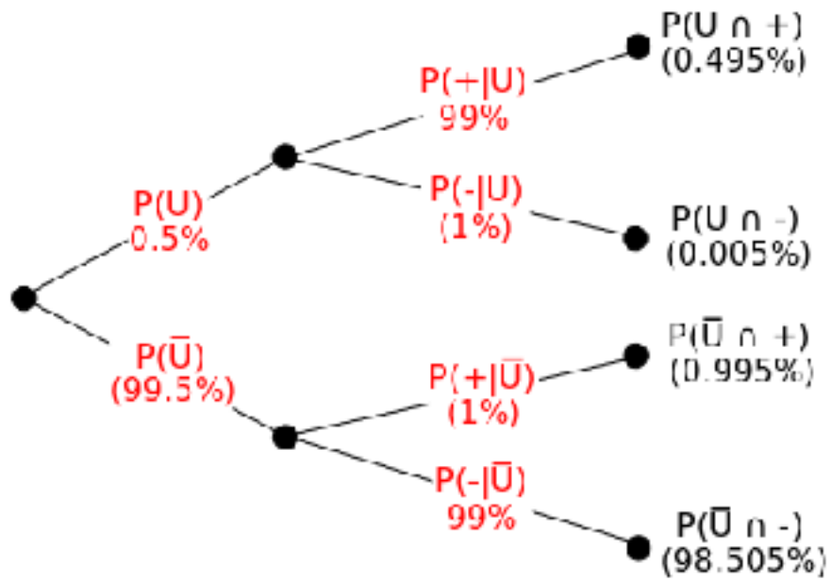
Bayes Theorem

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

Also used with frequentist probabilities!

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

Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability that he is a user?



$$P(\text{User} \mid +) = \frac{P(+ \mid \text{User})P(\text{User})}{P(+ \mid \text{User})P(\text{User}) + P(+ \mid \text{Non-user})P(\text{Non-user})}$$

$$= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}$$

$$\approx 33.2\%$$

https://en.wikipedia.org/wiki/Bayes'_theorem

Bayes Theorem

Likelihood

Posterior

Prior

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

Normalizing Constant
(Marginal Likelihood)

Odds Ratios

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{\frac{P(H_1)P(D|H_1)}{P(D)}}{\frac{P(H_2)P(D|H_2)}{P(D)}}$$

Odds Ratios

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{\frac{P(H_1)P(D|H_1)}{\cancel{P(D)}}}{\frac{P(H_2)P(D|H_2)}{\cancel{P(D)}}}$$

Odds Ratios

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)P(D|H_1)}{P(H_2)P(D|H_2)}$$

Odds Ratios

Posterior Odds | Prior Odds | Bayes Factor

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(H_1)}{P(H_2)} \frac{P(D|H_1)}{P(D|H_2)}$$

Odds Ratios

Prior Odds Bayes Factor Posterior Odds

$$\frac{P(H_1)}{P(H_2)} \frac{P(D|H_1)}{P(D|H_2)} = \frac{P(H_1|D)}{P(H_2|D)}$$

Marginalizing

Joint Probabilities

	σ_1	σ_2	σ_3	
μ_1	0.10	0.07	0.12	0.29
μ_2	0.05	0.22	0.06	0.33
μ_3	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

Marginal Probabilities

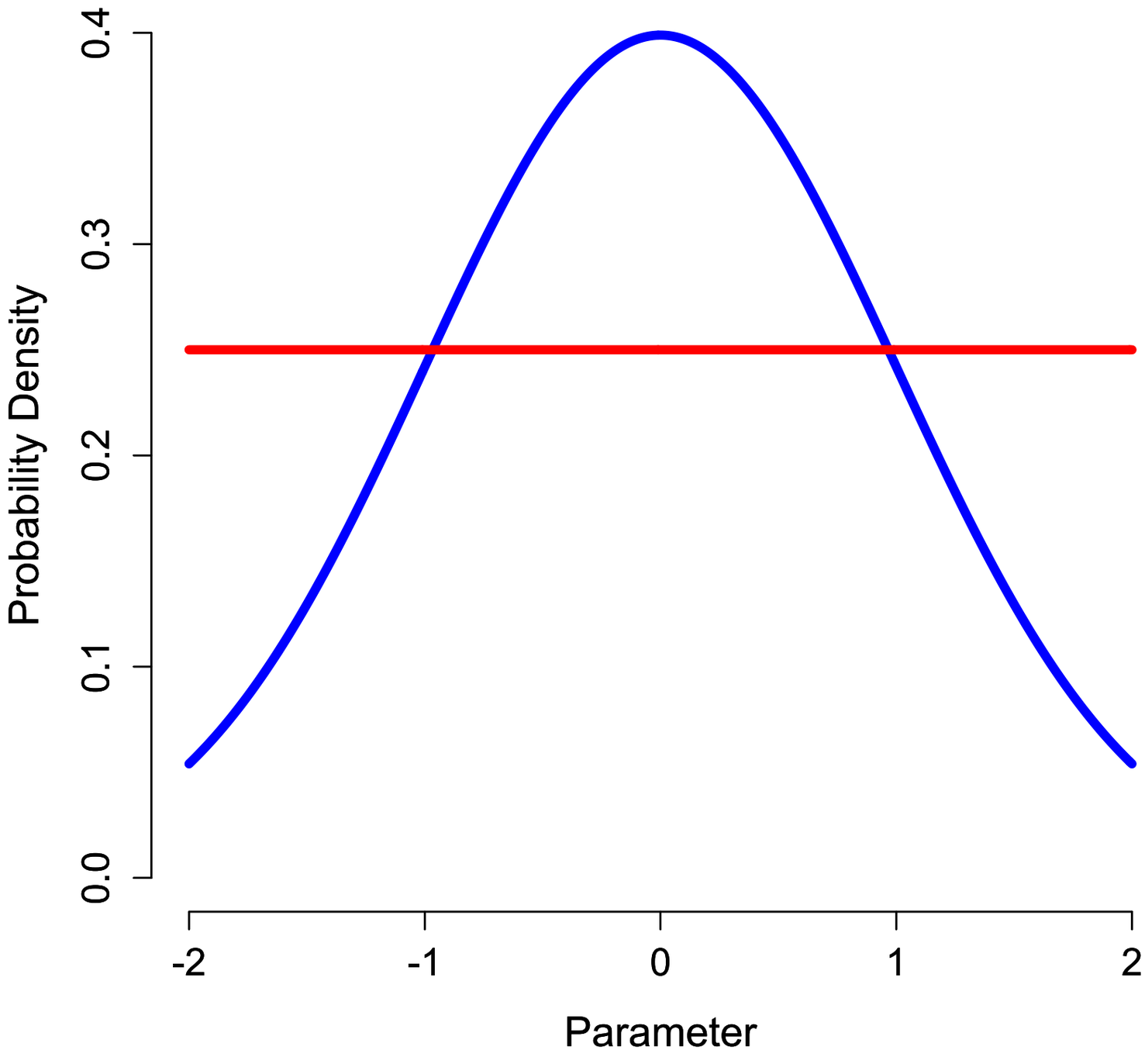
Monte Carlo Methods

An entire class of methods to draw samples from or estimate moments (mean, variance, etc.) from distributions when closed-form solutions are not available. Rely on drawing (pseudo-)random numbers.

(Monaco, not Las Vegas)

Monte Carlo Methods

Importance Sampling

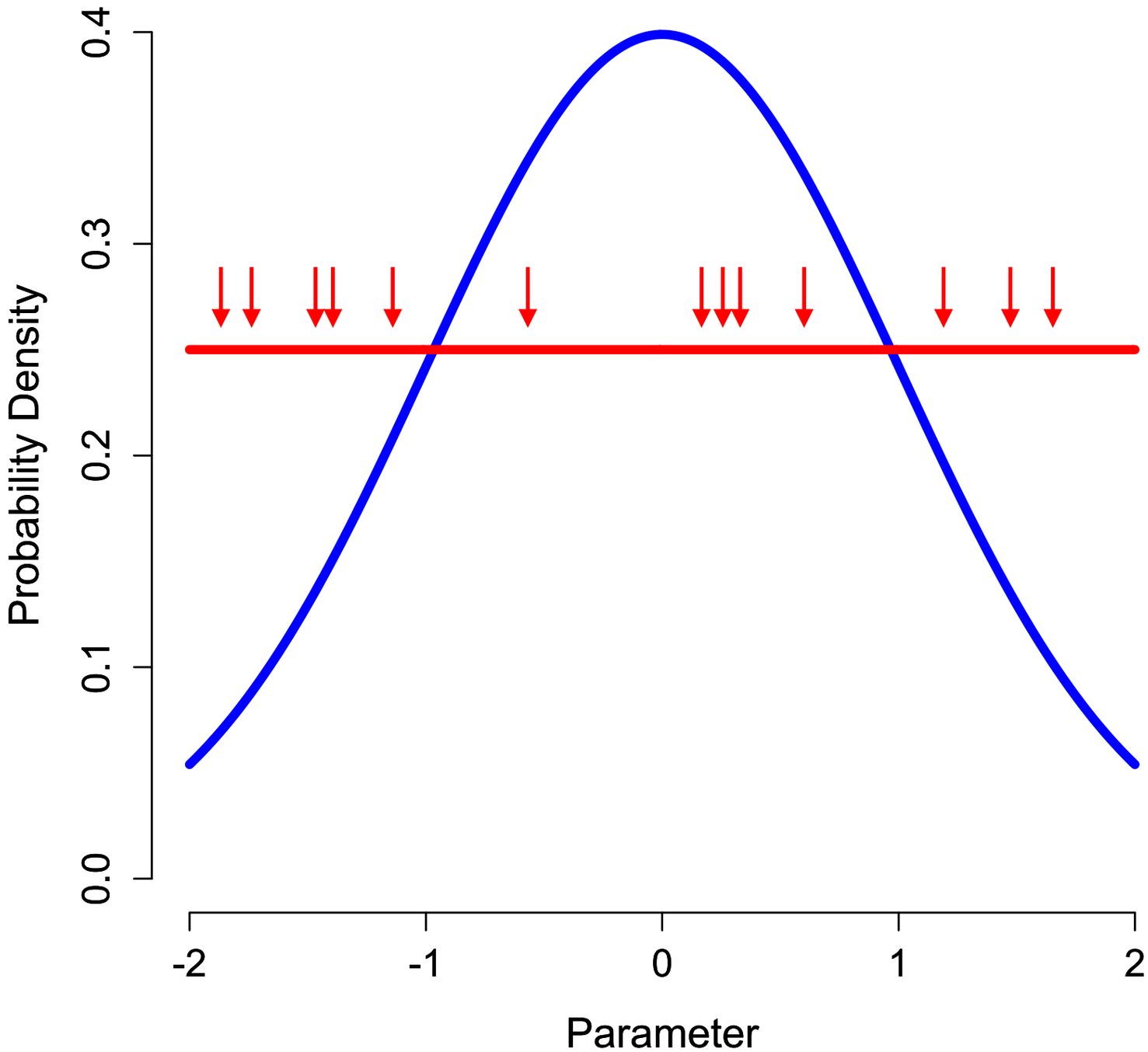


Sampling Distribution

Distribution of Interest

Monte Carlo Methods

Importance Sampling

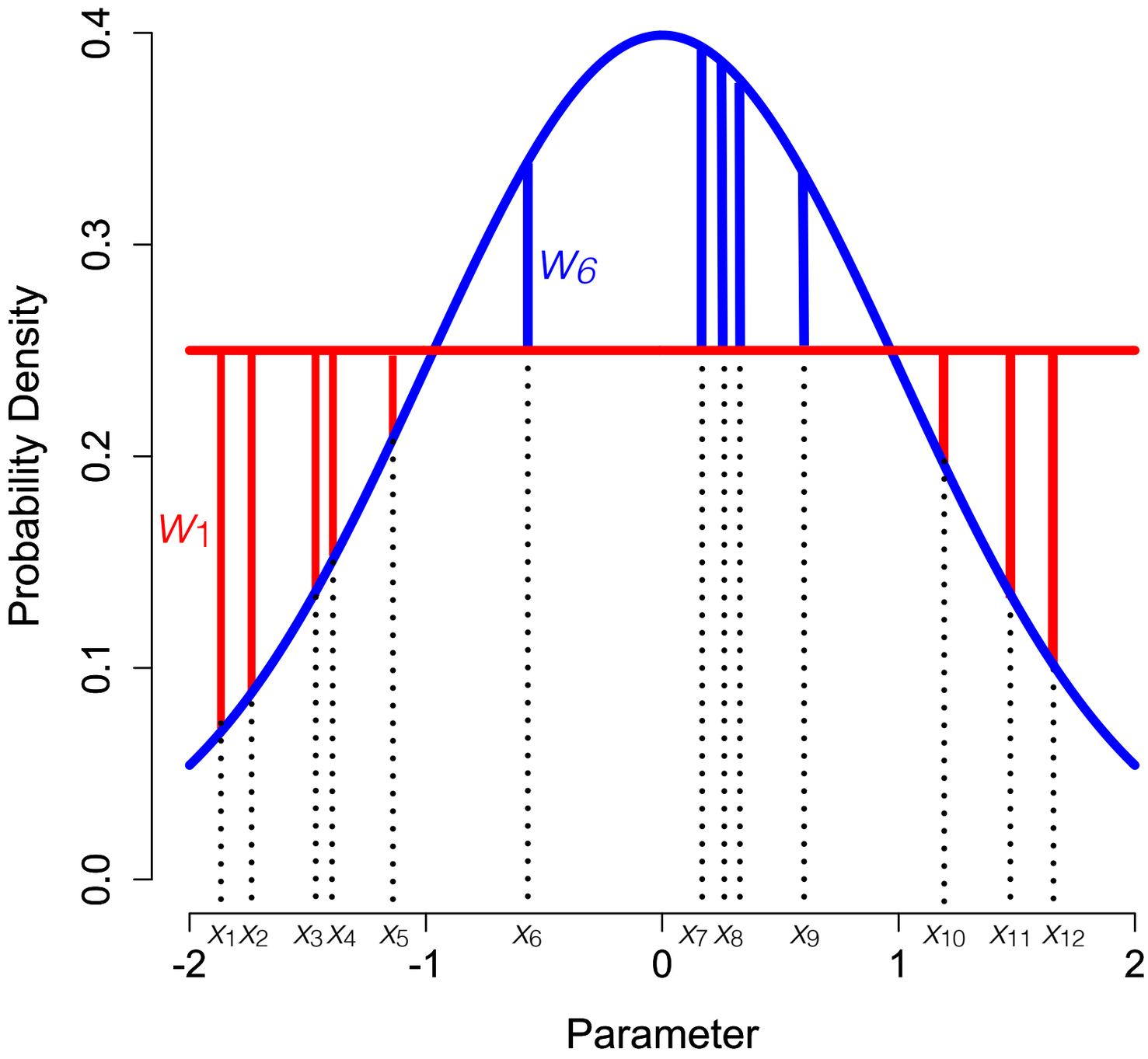


Sampling Distribution

Distribution of Interest

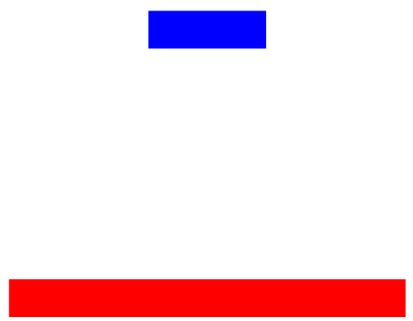
Monte Carlo Methods

Importance Sampling



Sampling Distribution

Distribution of Interest



A little bit

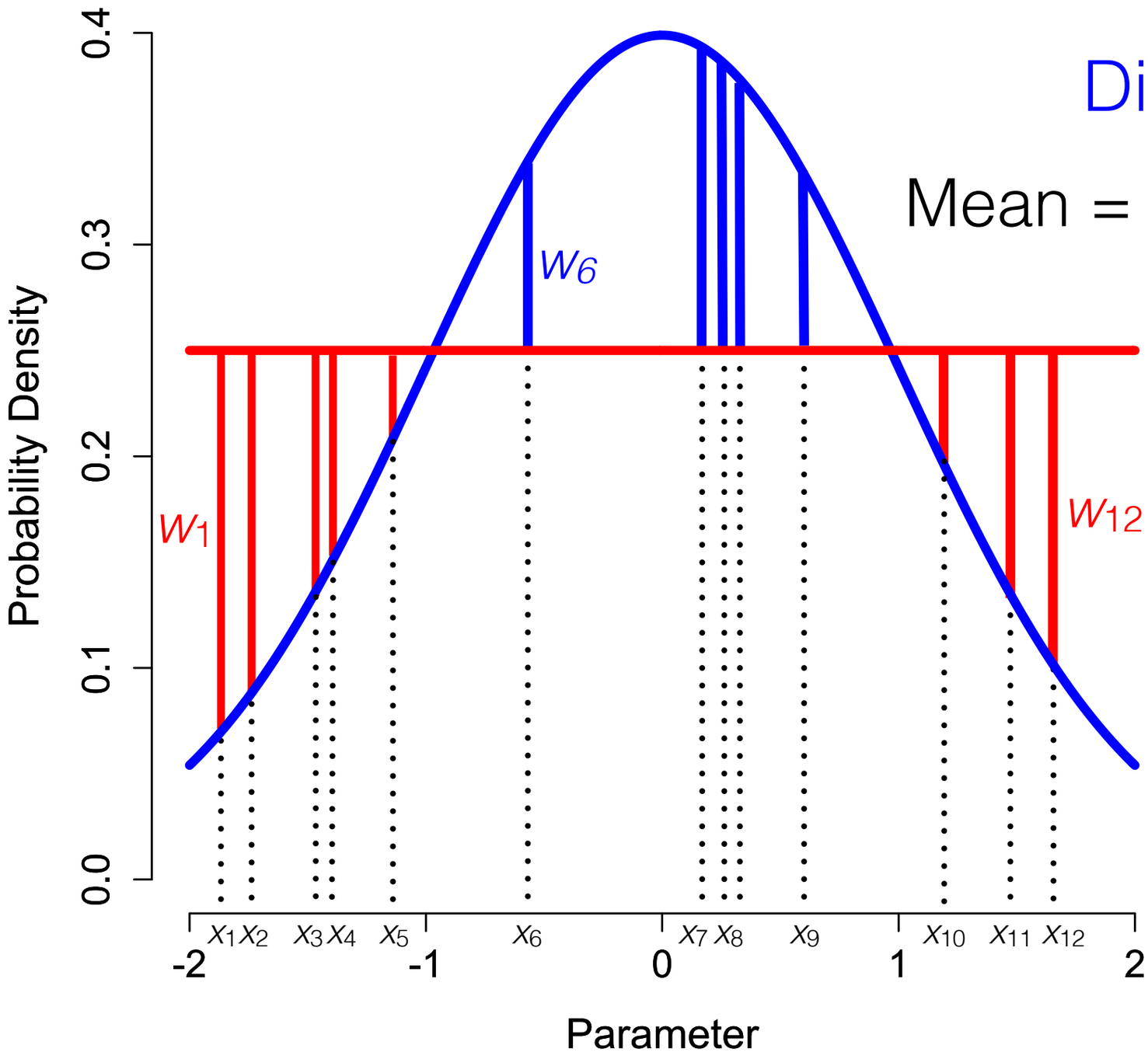
> 1

A lot

< 1

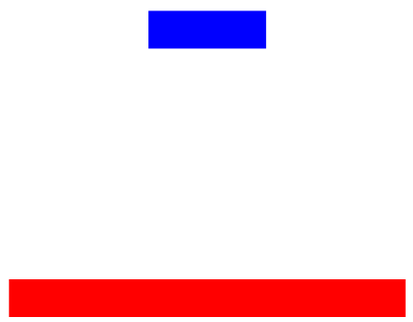
Monte Carlo Methods

Importance Sampling



Distribution of Interest

Mean = $x_1 w_1 + \ldots + x_6 w_6 + \ldots + x_{12} w_{12}$



A little bit

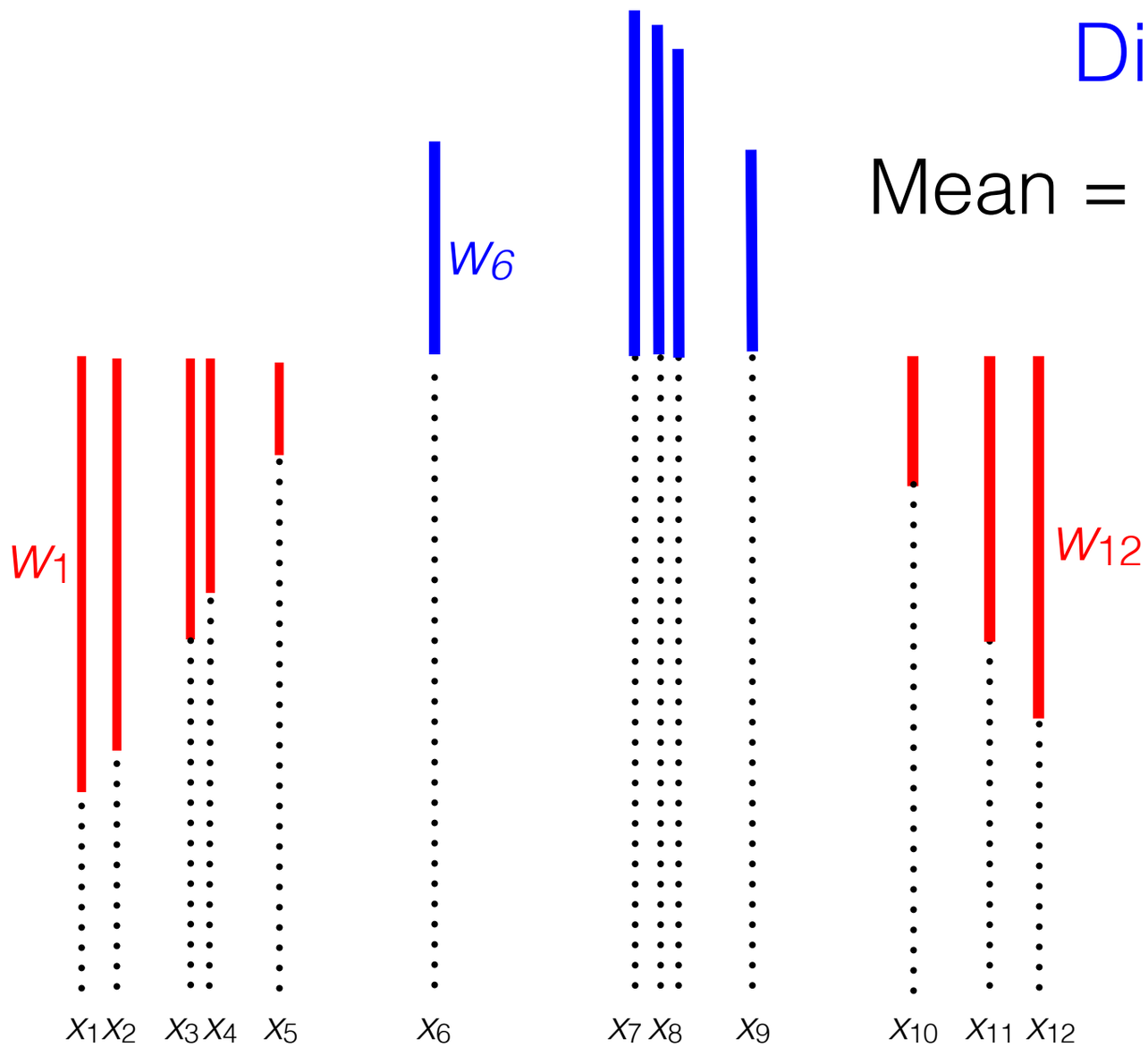
> 1

A lot

< 1

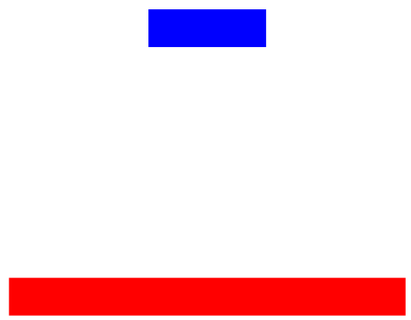
Monte Carlo Methods

Importance Sampling



Distribution of Interest

Mean = $x_1 w_1 + \ldots + x_6 w_6 + \ldots + x_{12} w_{12}$



A little bit

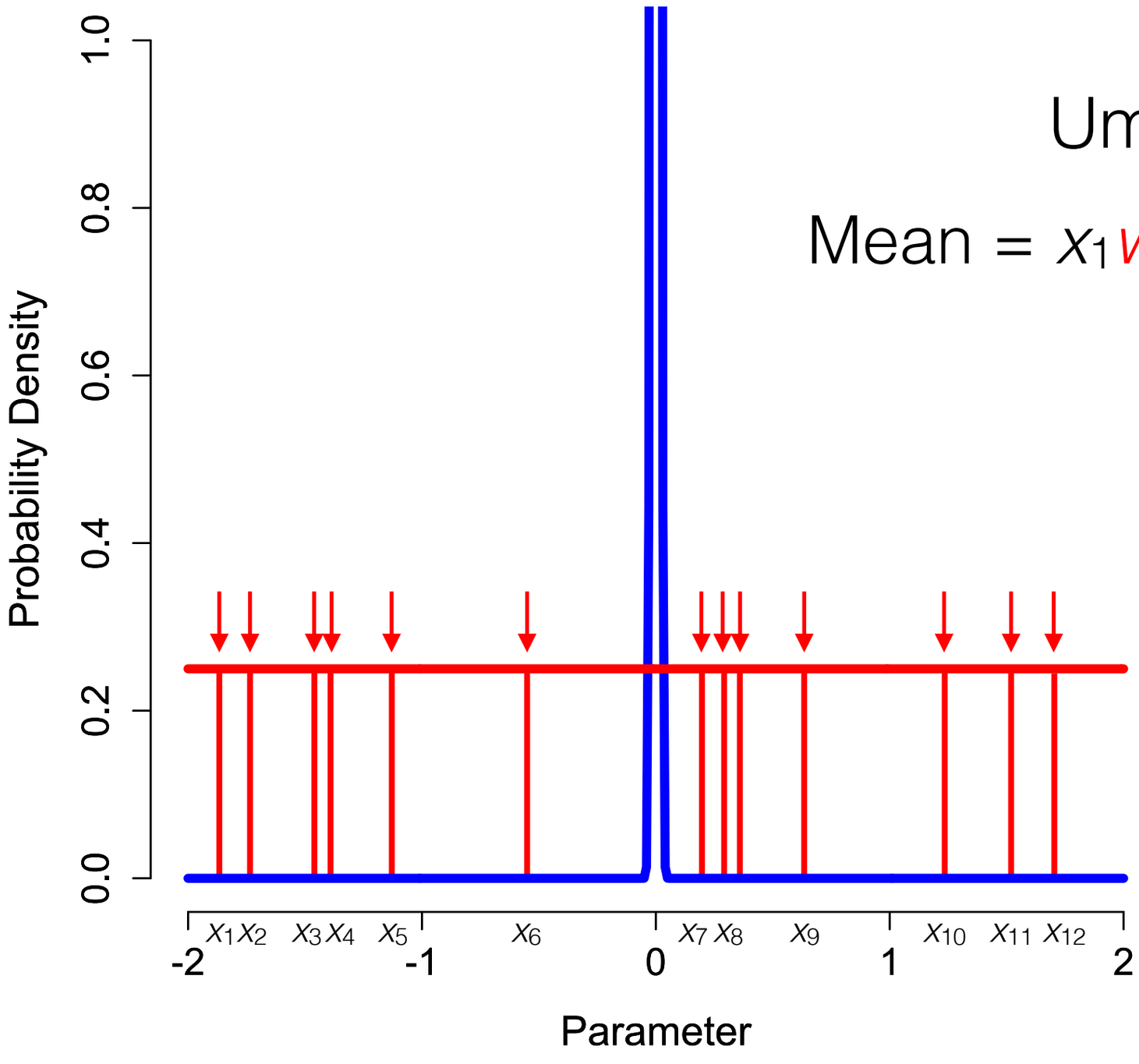
> 1

A lot

< 1

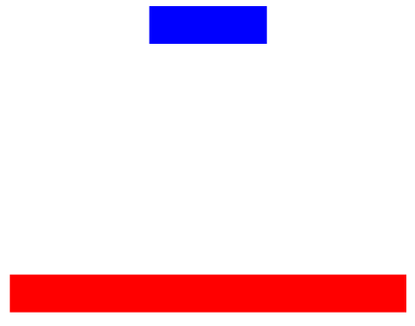
Monte Carlo Methods

Importance Sampling



Ummmm……

Mean = $x_1 w_1 + \ldots + x_6 w_6 + \ldots + x_{12} w_{12}$



A little bit

> 1

A lot

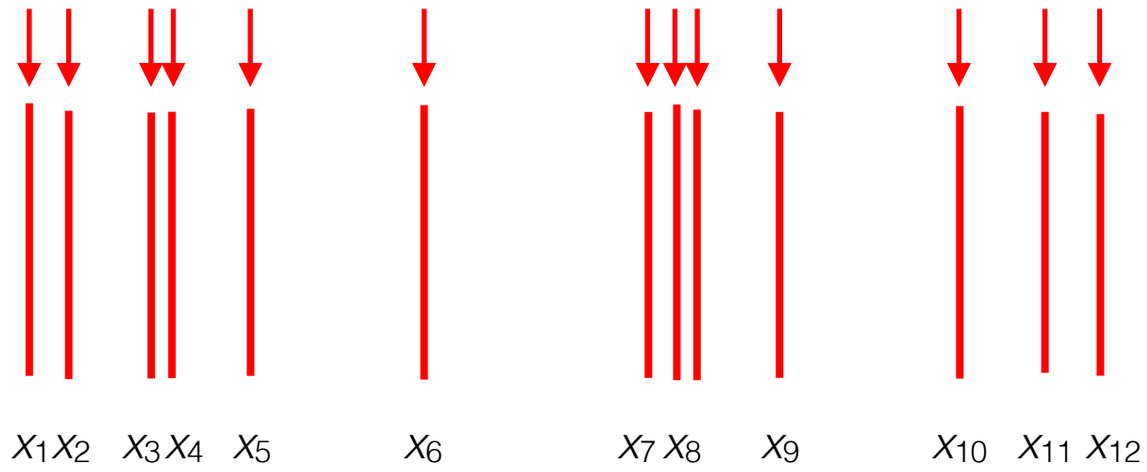
< 1

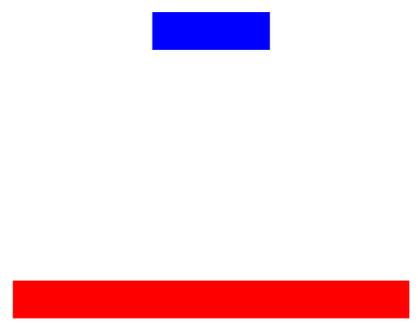
Monte Carlo Methods

Importance Sampling

Does this look like our
Distribution of Interest?

Mean = $x_1 w_1 + \ldots + x_6 w_6 + \ldots + x_{12} w_{12}$



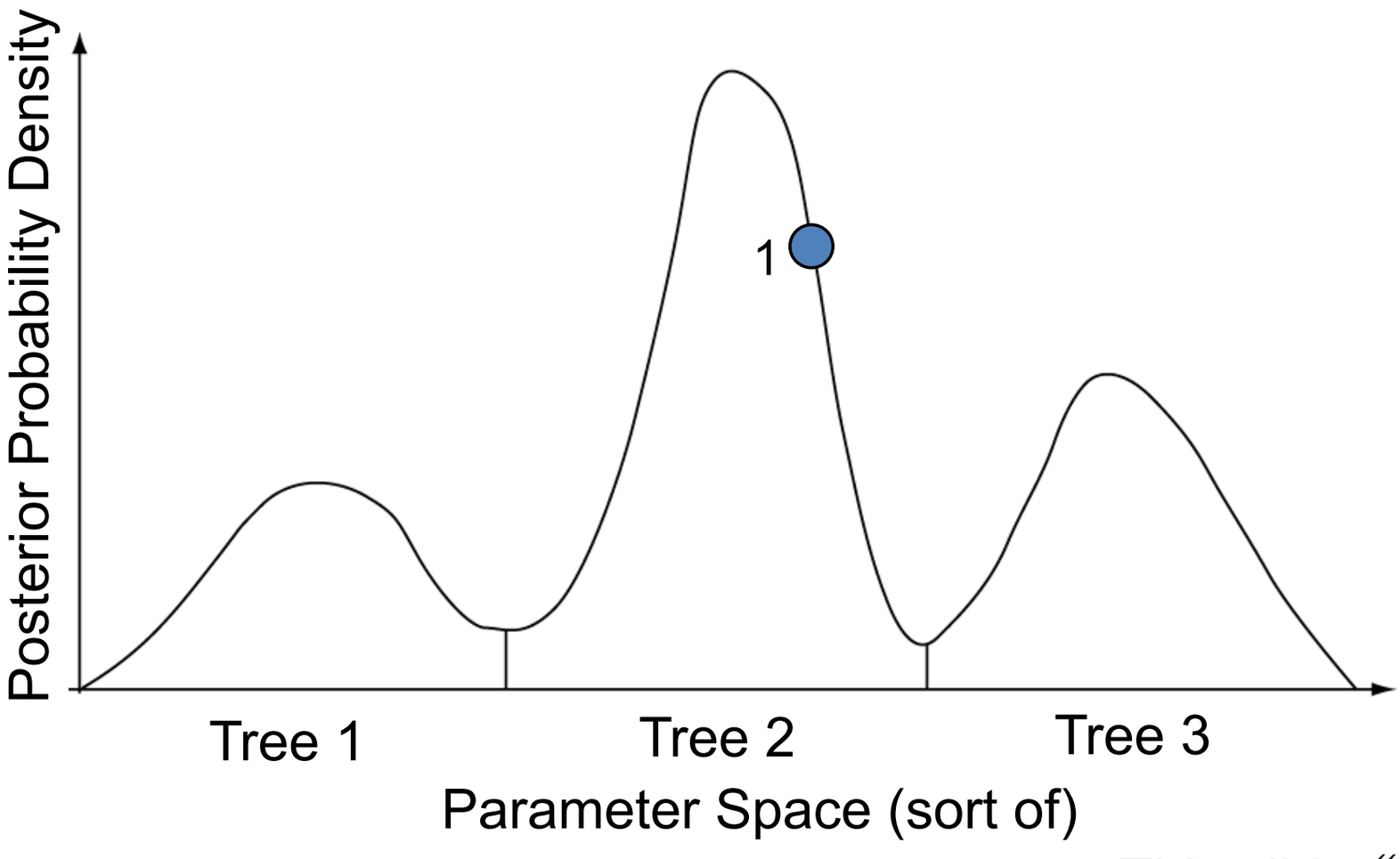


A little bit
> 1

A lot
< 1

Markov chain Monte Carlo

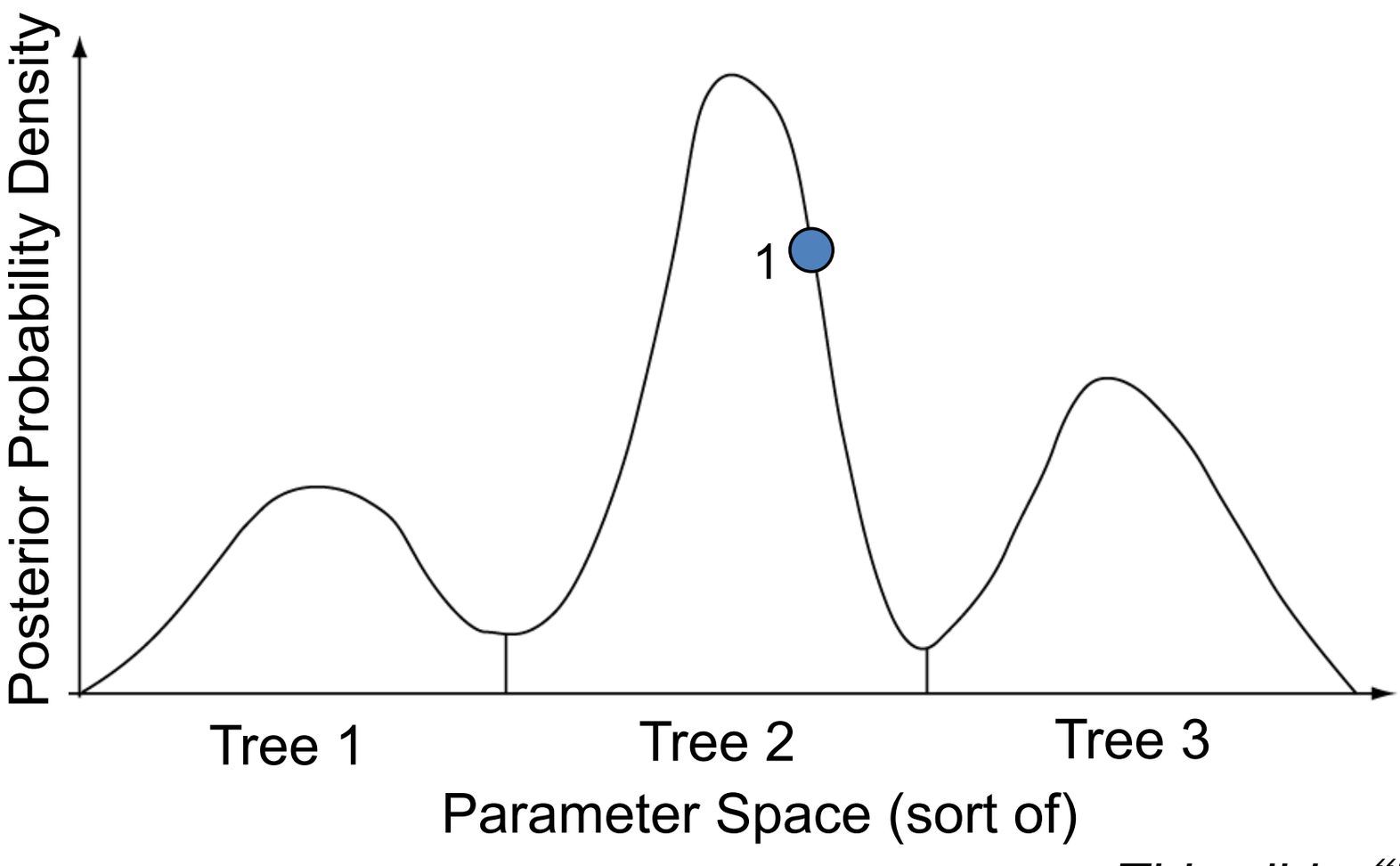
1. Start at an arbitrary point



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

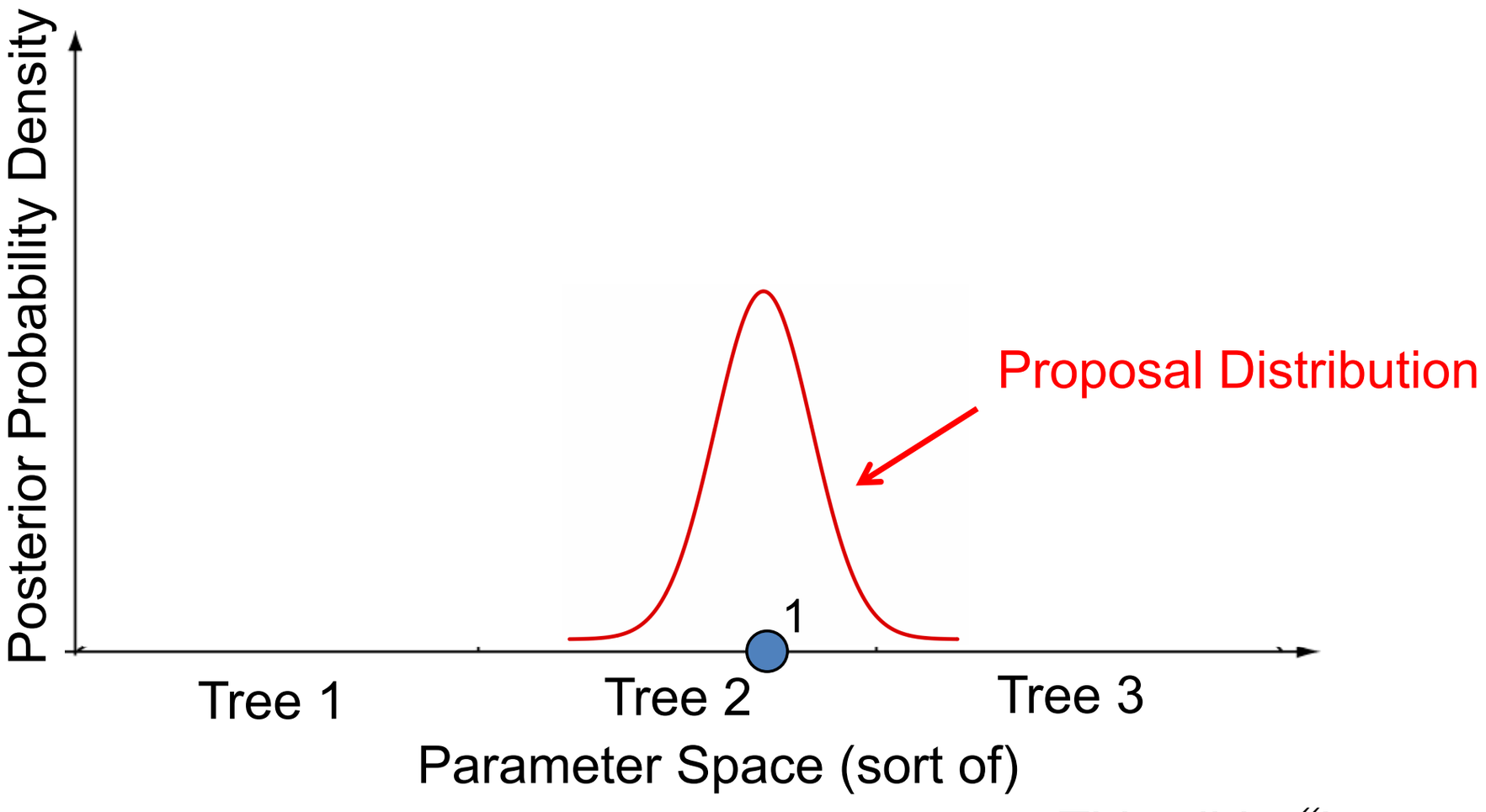
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

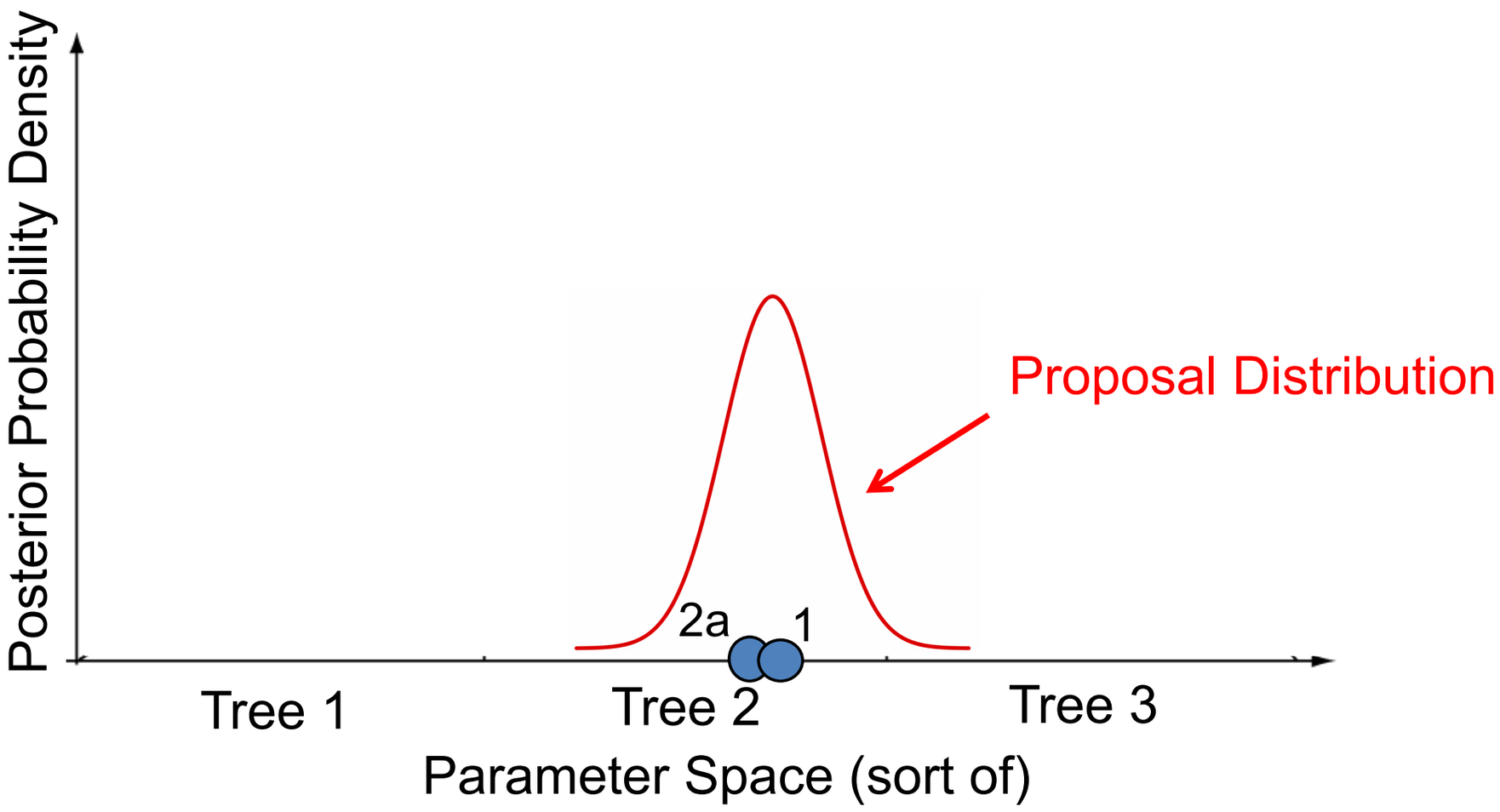
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

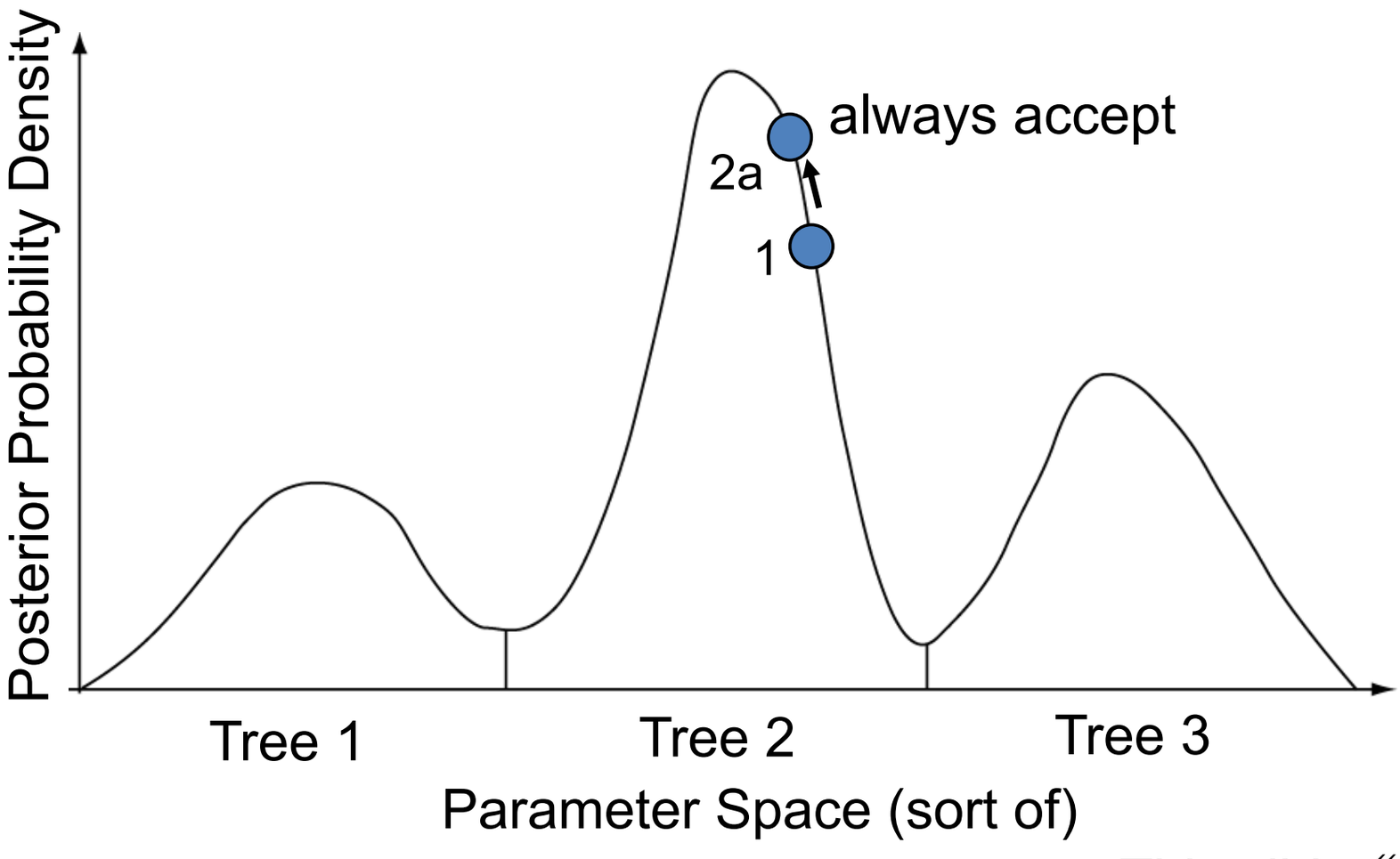
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

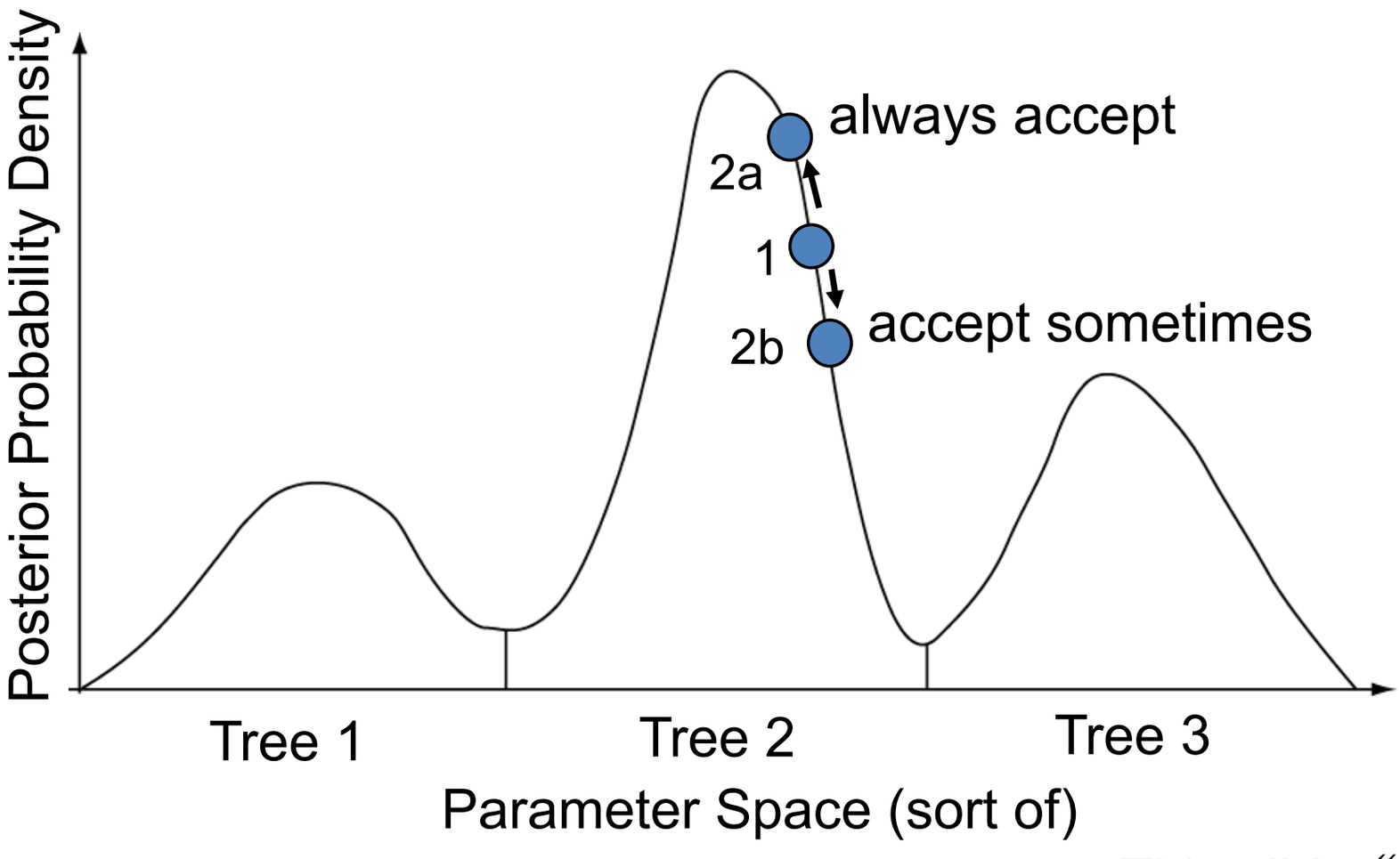
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

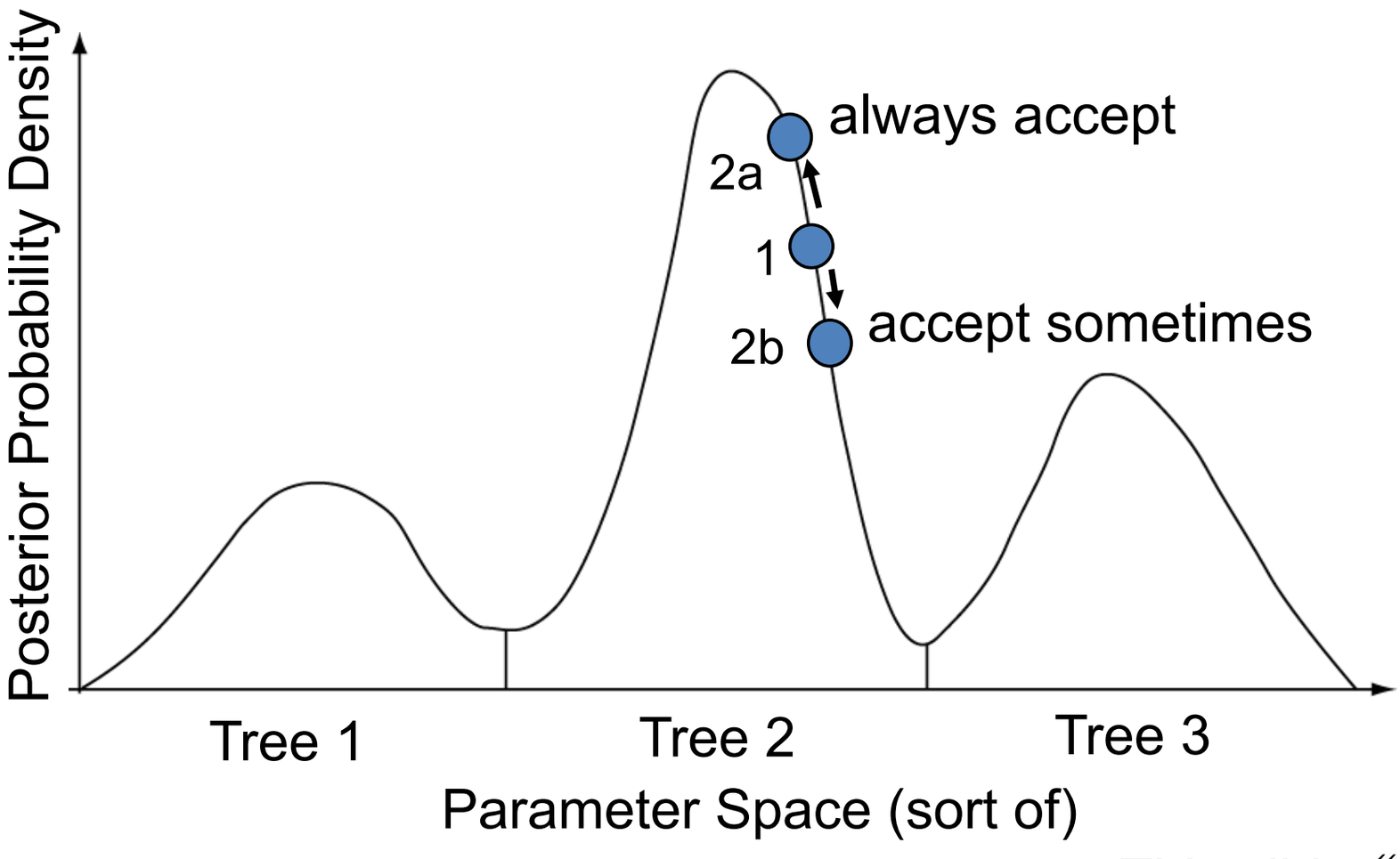
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

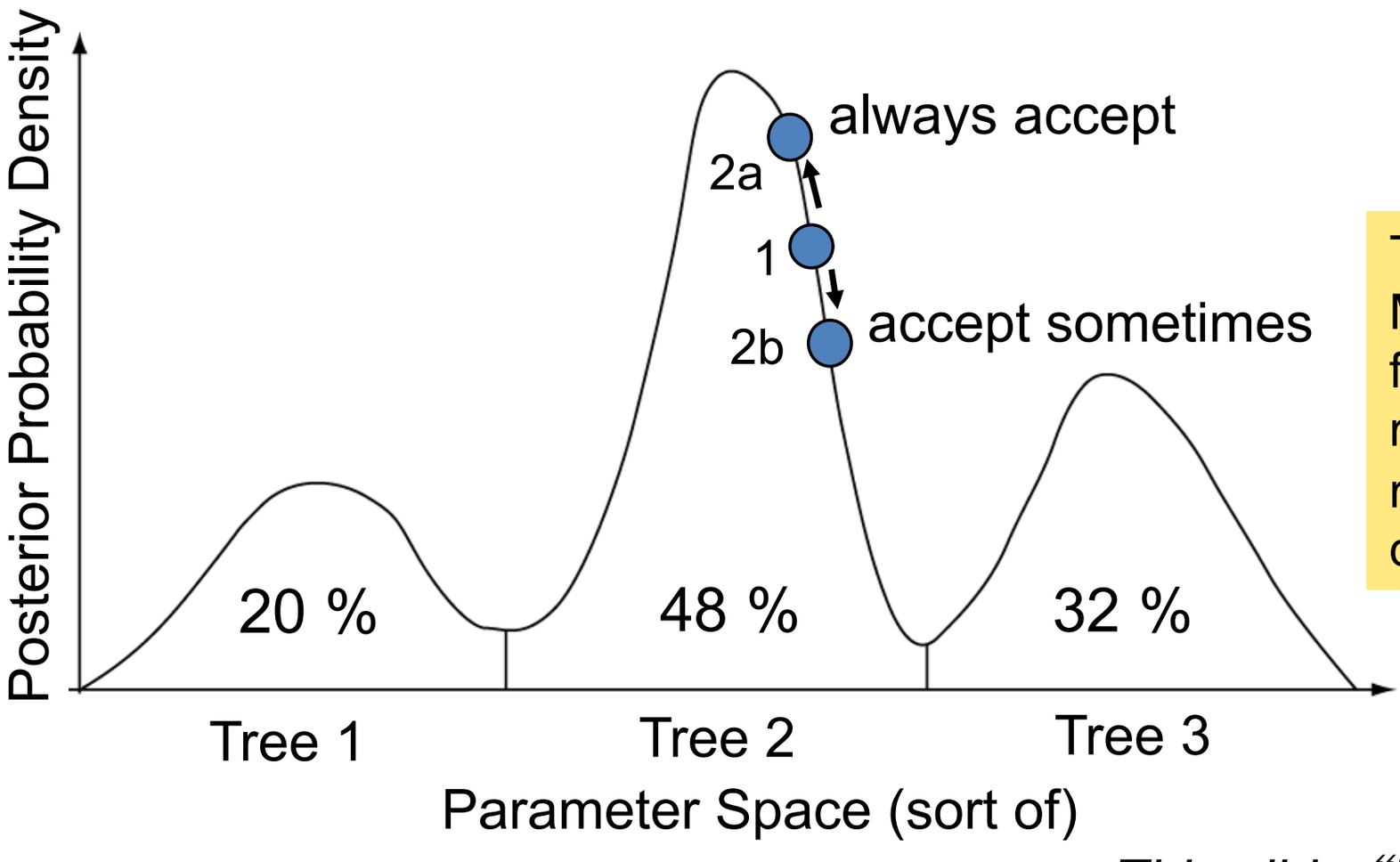
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH (x 10,000's – x 10,000,000's)



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH (x 10,000's – x 10,000,000's)



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

This slide "borrowed" from F. Ronquist

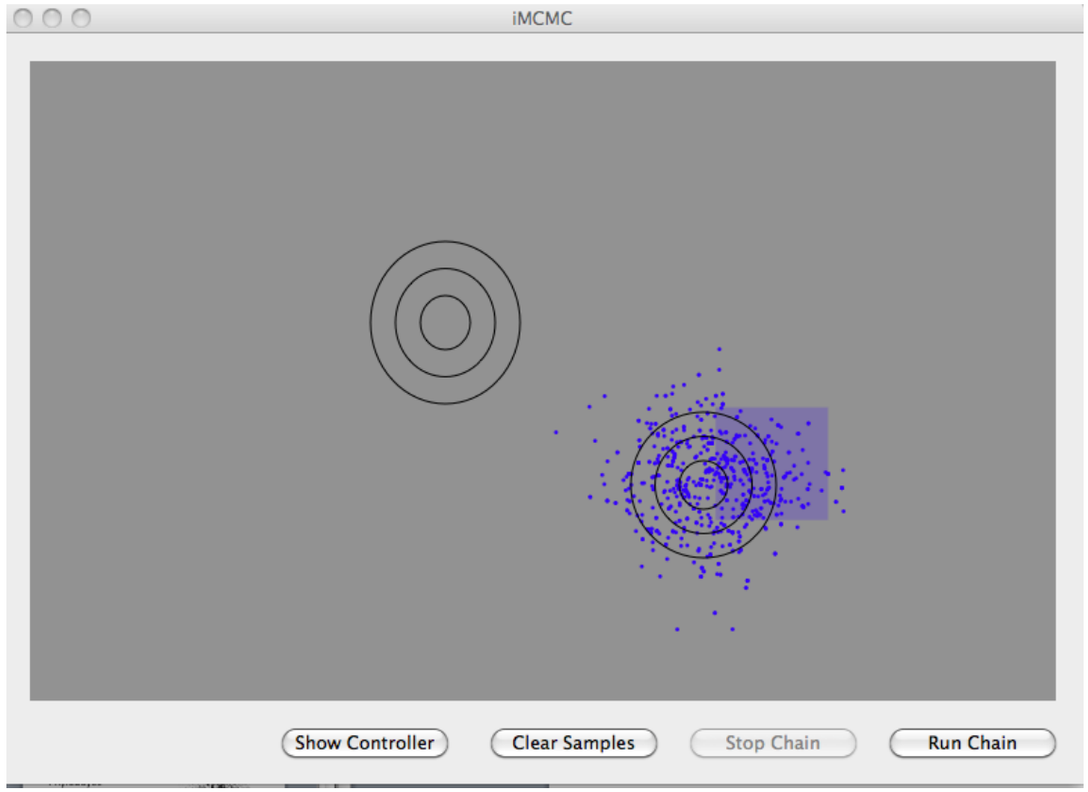
Toy MCMC Software

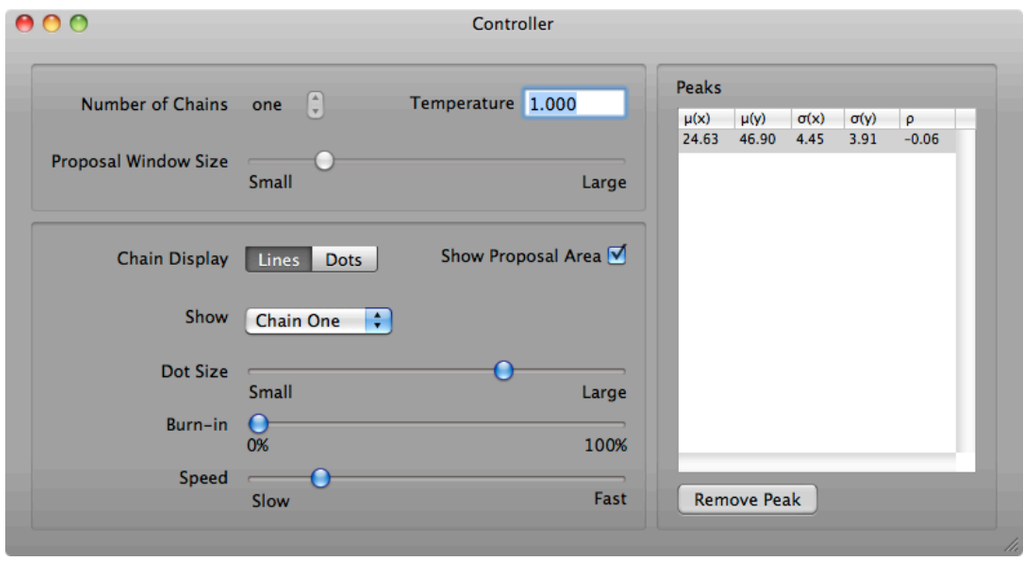
MCMC Robot (Paul Lewis; PC & iOS)

http://www.mcmcrobot.org

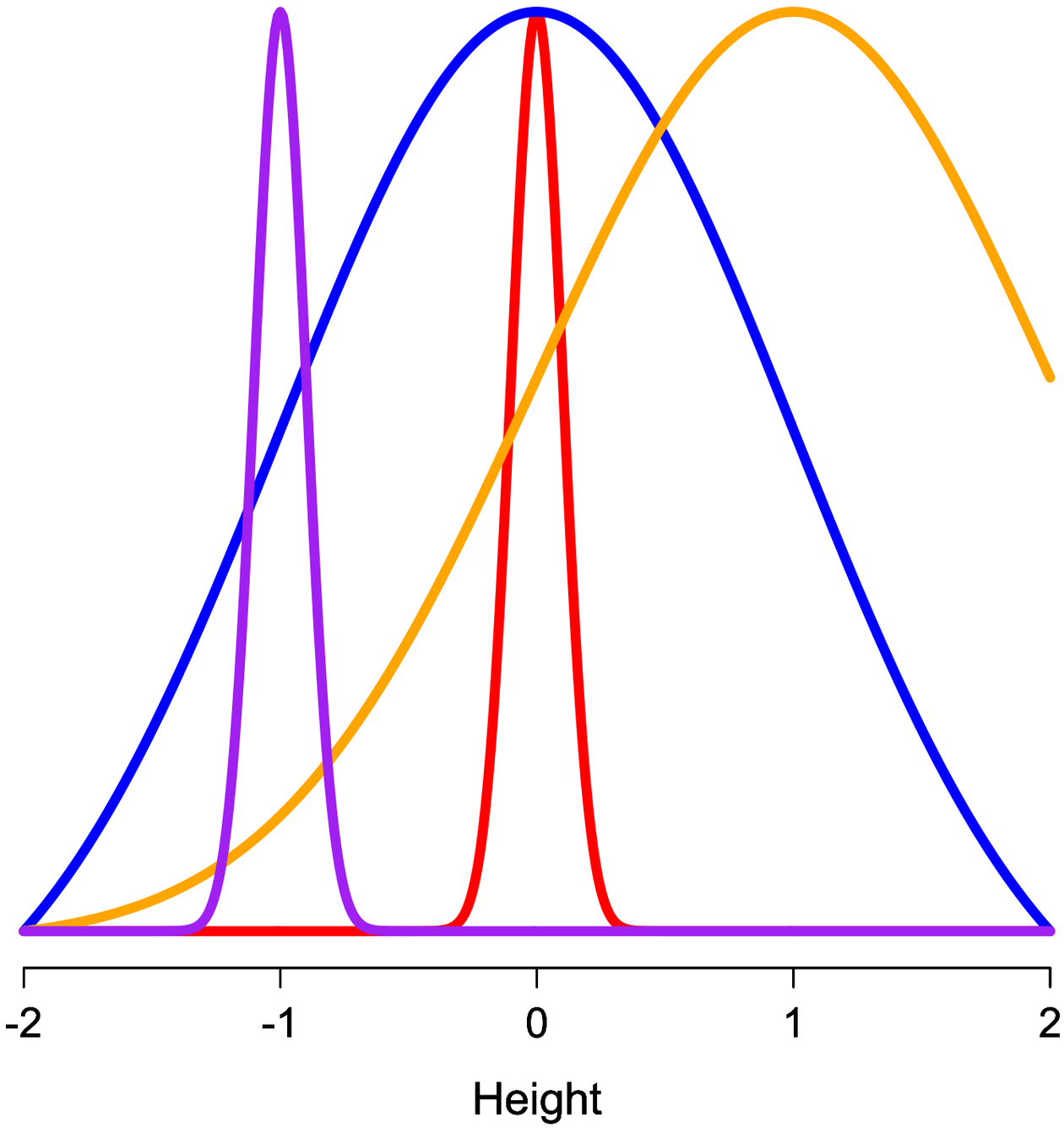
iMCMC (John Huelsenbeck; Mac)

http://cteg.berkeley.edu/software/huelsenbeck/McmcApp.zip





Estimating Parameters of a Normal Distribution



Mean (μ)

Standard Deviation (σ)

RevBayes Exercise 9

MCMC Parameter Estimation of a Normal Distribution

Convergence of Scalars - Tracer

Tracer

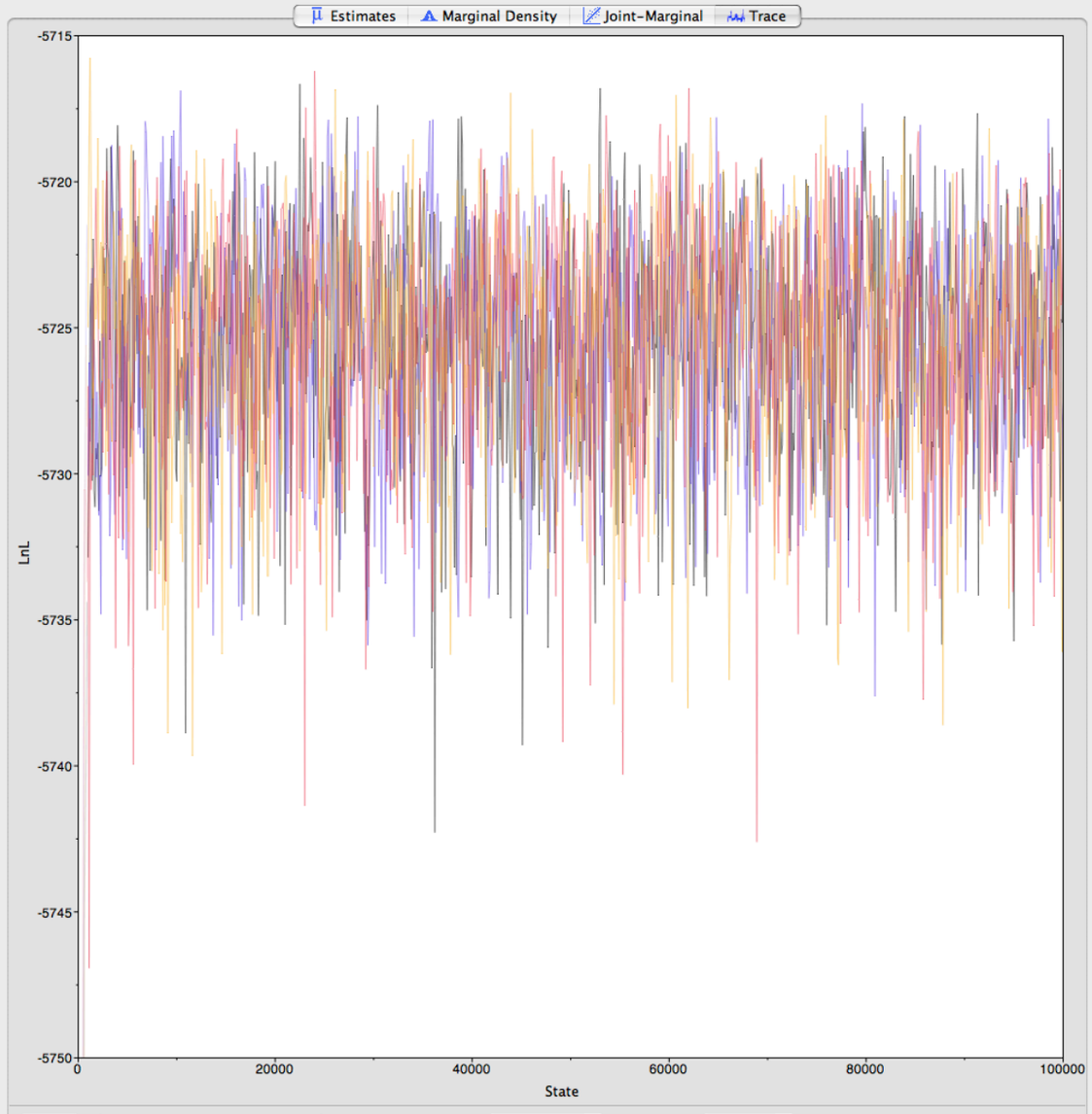
Trace Files:

Tree File	States	Burn-In
primates_HKYG.run1.p	100000	1000
primates_HKYG.run2.p	100000	1000
primates_HKYG.run3.p	100000	1000
primates_HKYG.run4.p	100000	1000
Combined	396400	-

Traces:

Statistic	Mean	ESS
LnL	-5725.48	487.117
TL	3.03	344.642
kappa	11.542	342.441
pi(A)	0.363	424.714
pi(C)	0.319	387.966
pi(G)	8.23E-2	422.1
pi(T)	0.236	407.606
alpha	0.385	473.782

Estimates | Marginal Density | Joint-Marginal | Trace



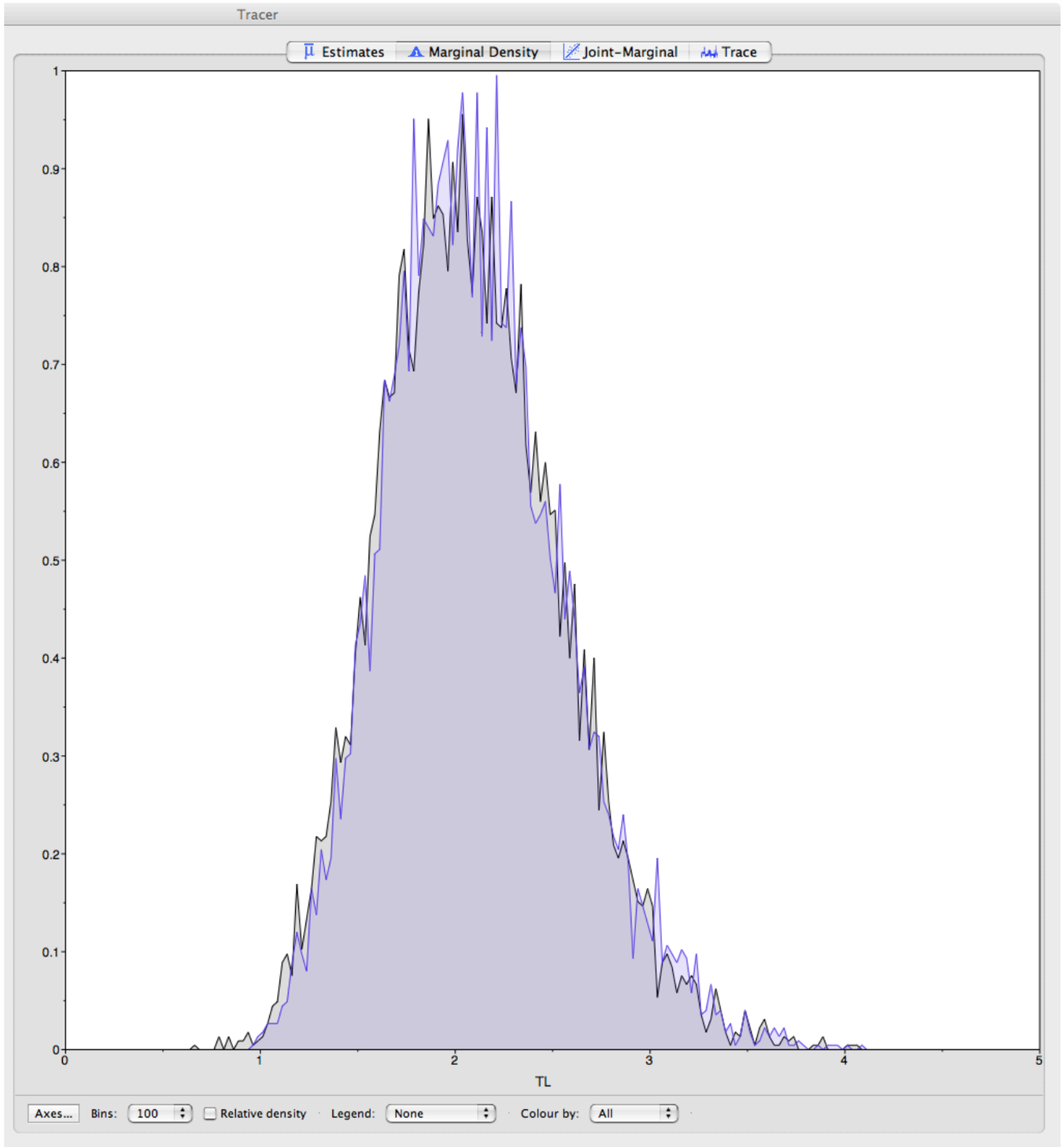
Axes... | Show Burn-in | Sample only | Draw line plot | Legend: None | Colour by: All

Running on Empty

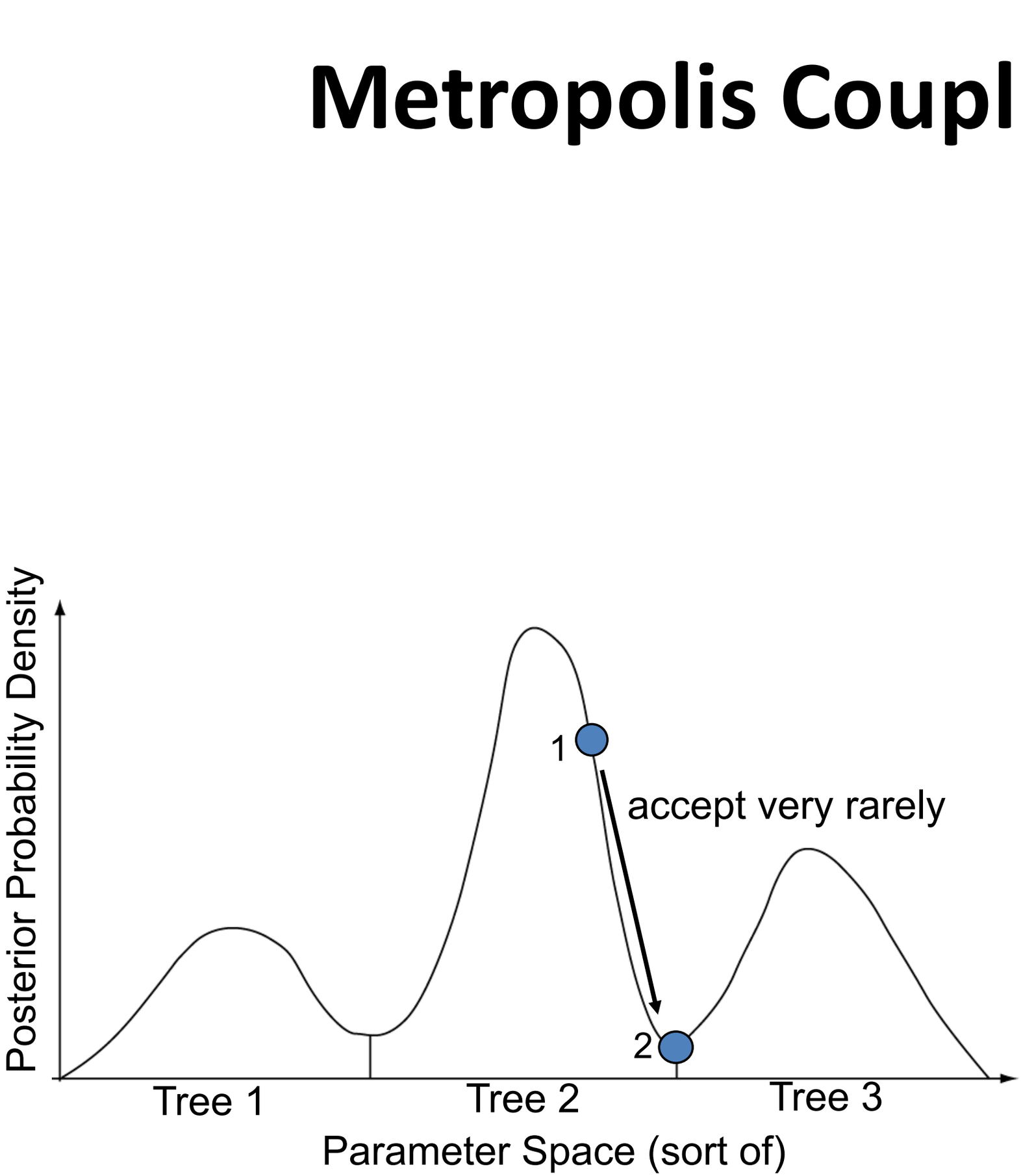
```
#NEXUS
begin data;
dimensions ntax=12 nchar=5;
format datatype=dna interleave=no gap=- missing=?;
matrix
Tarsius_syrichta    ?????
Lemur_catta         ?????
Homo_sapiens        ?????
Pan                 ?????
Gorilla             ?????
Pongo               ?????
Hylobates           ?????
Macaca_fuscata      ?????
M_mulatta           ?????
M_fascicularis      ?????
M_sylvanus          ?????
Saimiri_sciureus    ?????
;
end;
```

Or now in MrBayes:

mcmc data=no



Metropolis Coupling

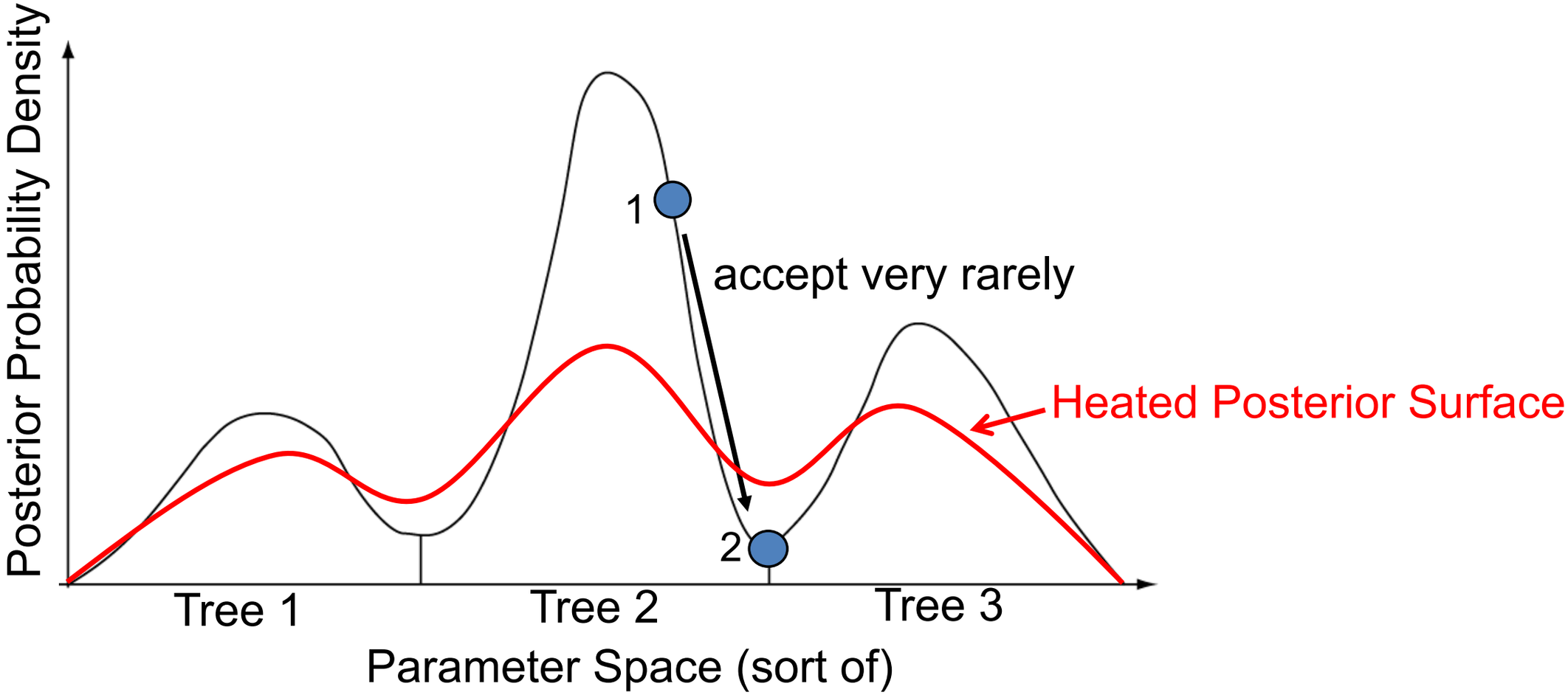


This slide "borrowed" from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases

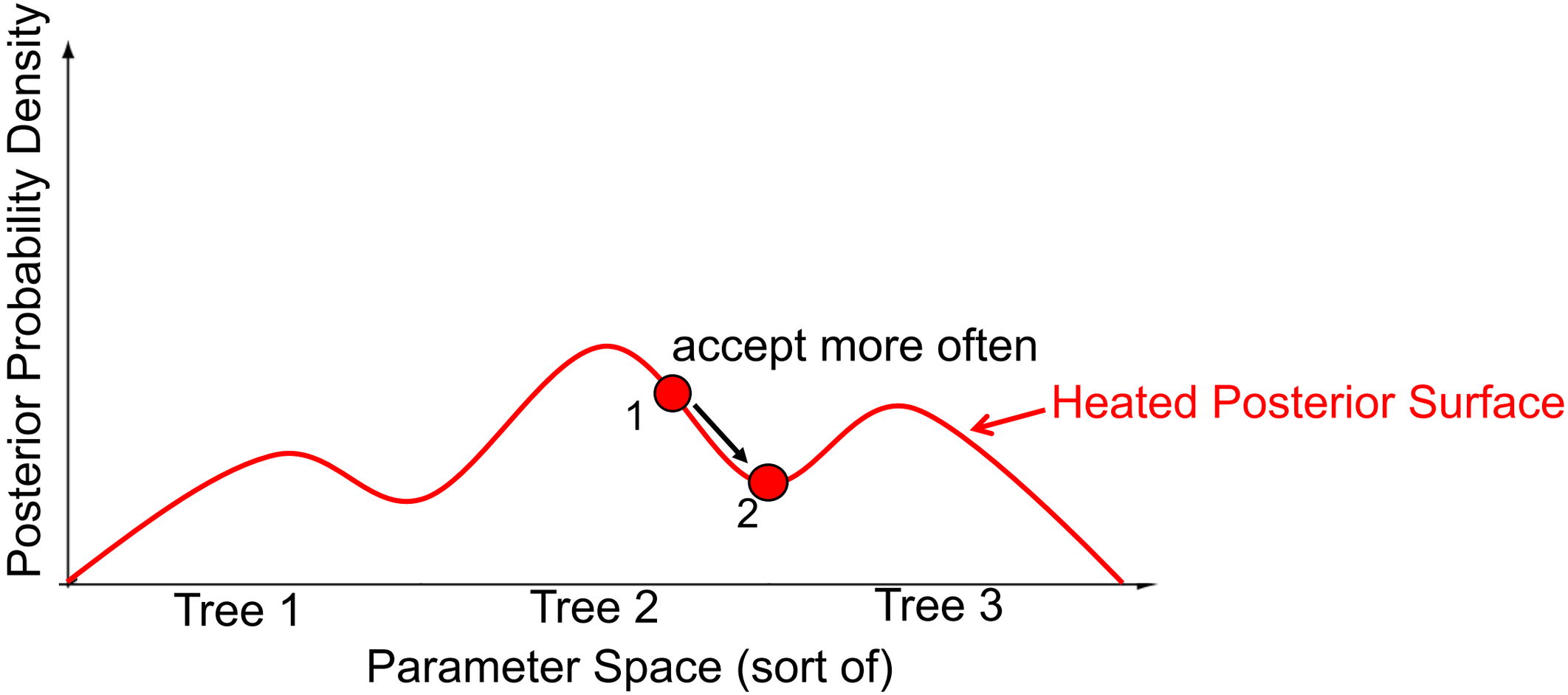


This slide “borrowed” from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases



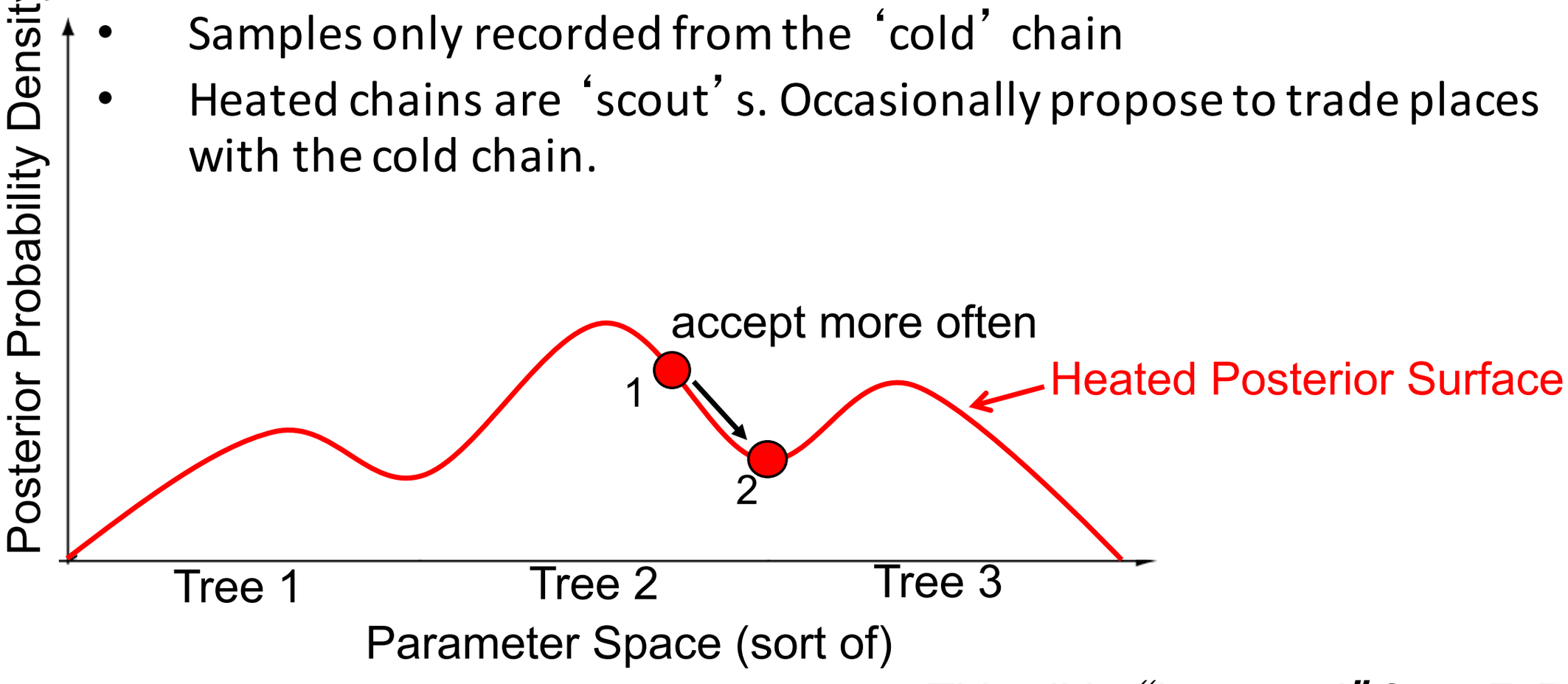
This slide “borrowed” from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases

- Samples only recorded from the ‘cold’ chain
- Heated chains are ‘scout’s. Occasionally propose to trade places with the cold chain.



This slide “borrowed” from F. Ronquist