Bayesian phylogenetics

Workshop on Molecular Evolution
Woods Hole, MA
Claudia Solis-Lemus

Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\overset{\textbf{Likelihood}}{\prod P(X = x_i|\lambda)}\,\overset{\textbf{Prior}}{p(\lambda)}}{\underset{\textbf{Marginal}}{p(\mathbf{X})}} \longleftarrow \text{Many times, intractable}$$

Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\overset{\text{Likelihood}}{\prod P(X=x_i|\lambda)}\,\overset{\text{Prior}}{p(\lambda)}}{\underset{\text{Marginal}}{p(\mathbf{X})}} \longleftarrow \text{Many times, intractable}$$

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i+\alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i+\alpha, n+\beta\right)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

Conjugate prior

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i + \alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n + \beta\right)$$

Posterior distribution

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(\alpha, \beta)$$

Conjugate prior

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|\mathbf{X}) = \frac{(n+\beta)^{(\sum x_i+\alpha)}}{\Gamma(\sum x_i + \alpha)}\lambda^{\sum x_i+\alpha-1}e^{-(n+\beta)\lambda}$$

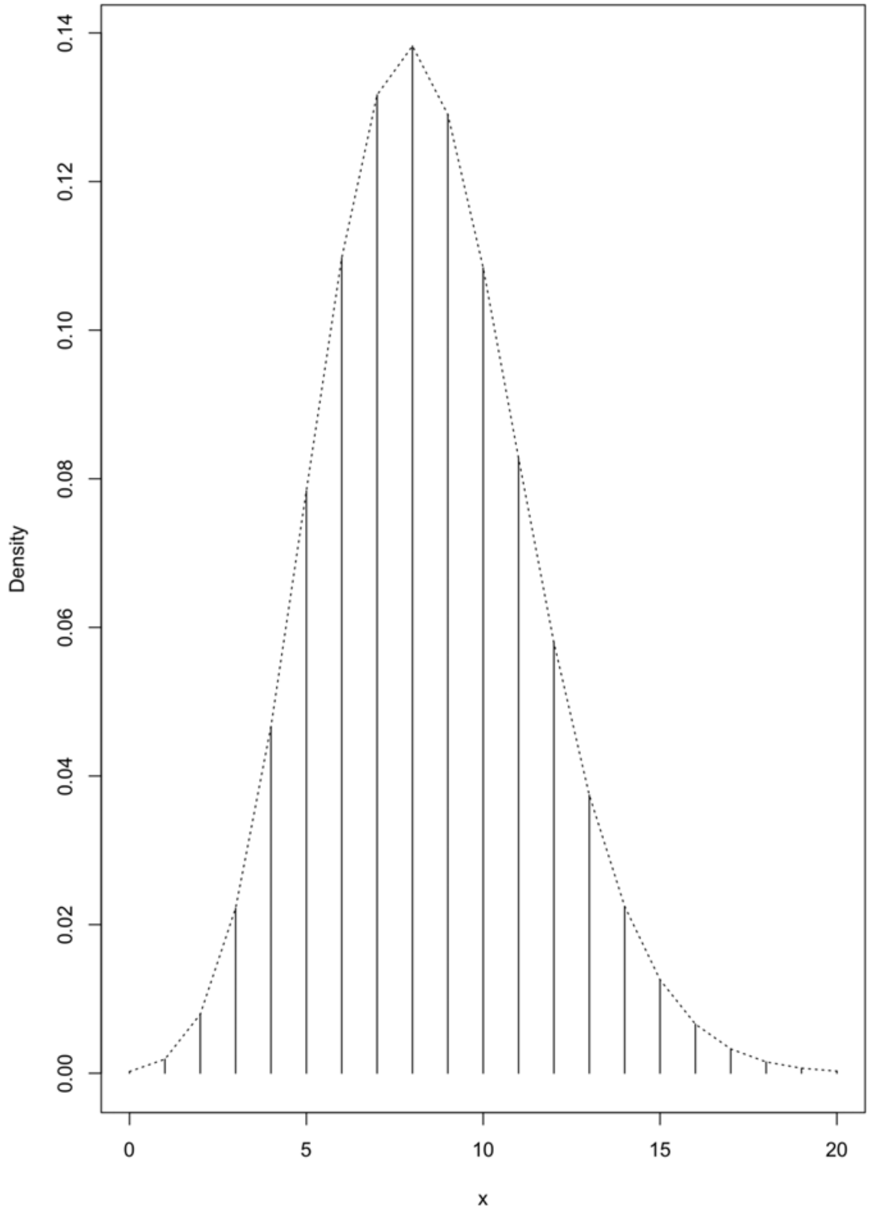
$$\lambda|\mathbf{X} \sim Gamma\left(\sum x_i + \alpha, n+\beta\right)$$

Posterior mean: $\dfrac{\sum x_i + \alpha}{n+\beta} = \dfrac{n}{n+\beta}\left(\dfrac{\sum x_i}{n}\right) + \dfrac{\beta}{n+\beta}\left(\dfrac{\alpha}{\beta}\right)$

Bayesian: Likelihood 2.0

Information in the data

Likelihood

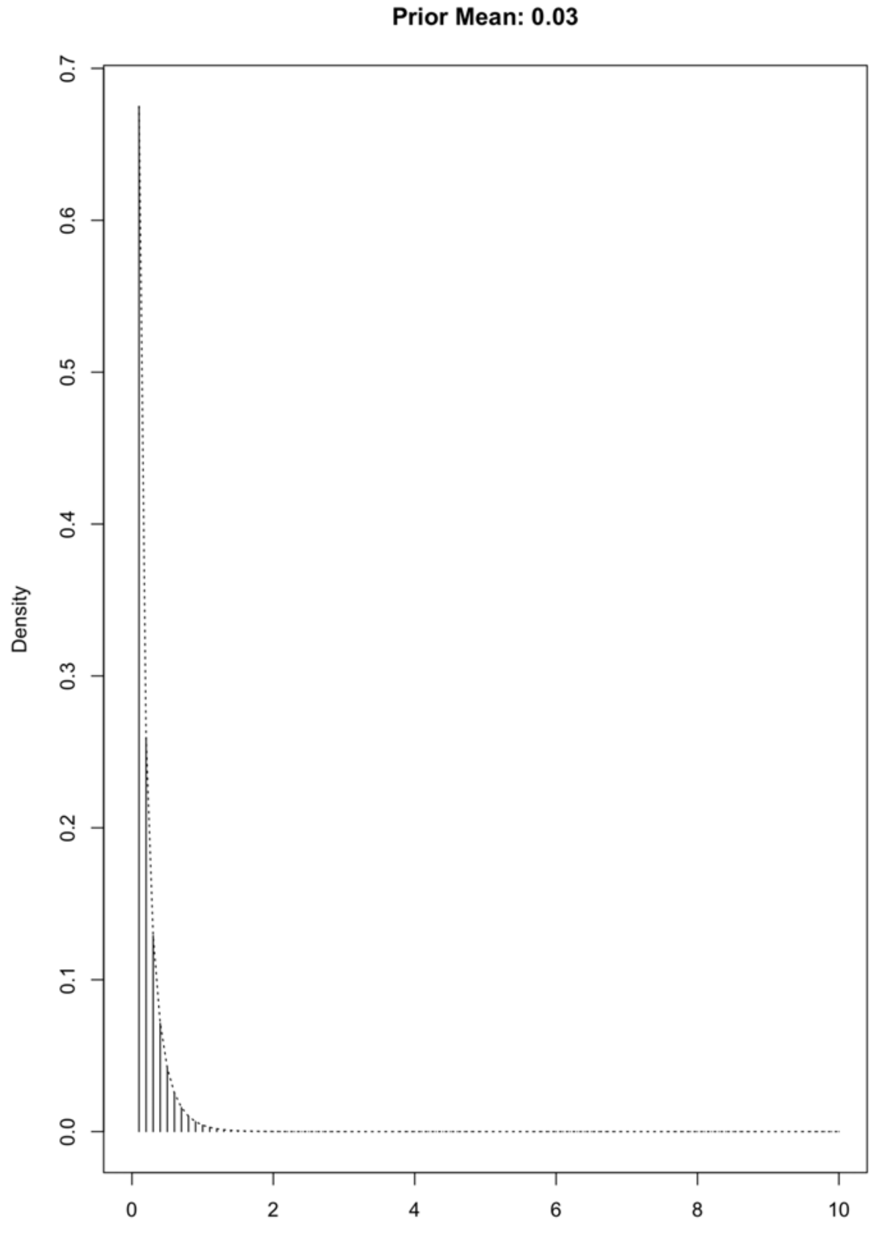


$$\hat{\lambda} = 8.4$$

Bayesian: Likelihood 2.0

Your knowledge

Prior

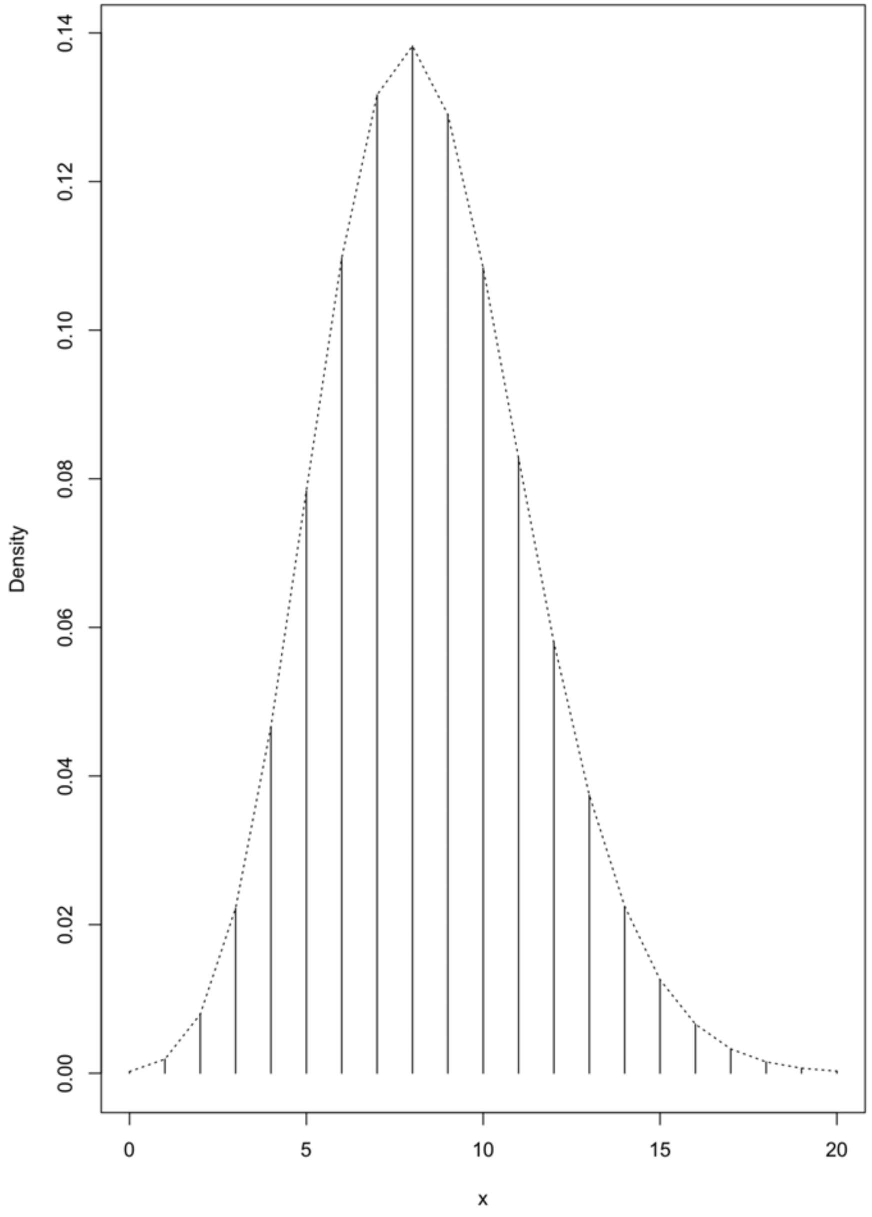


Lambda

$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$

Information in the data

Likelihood

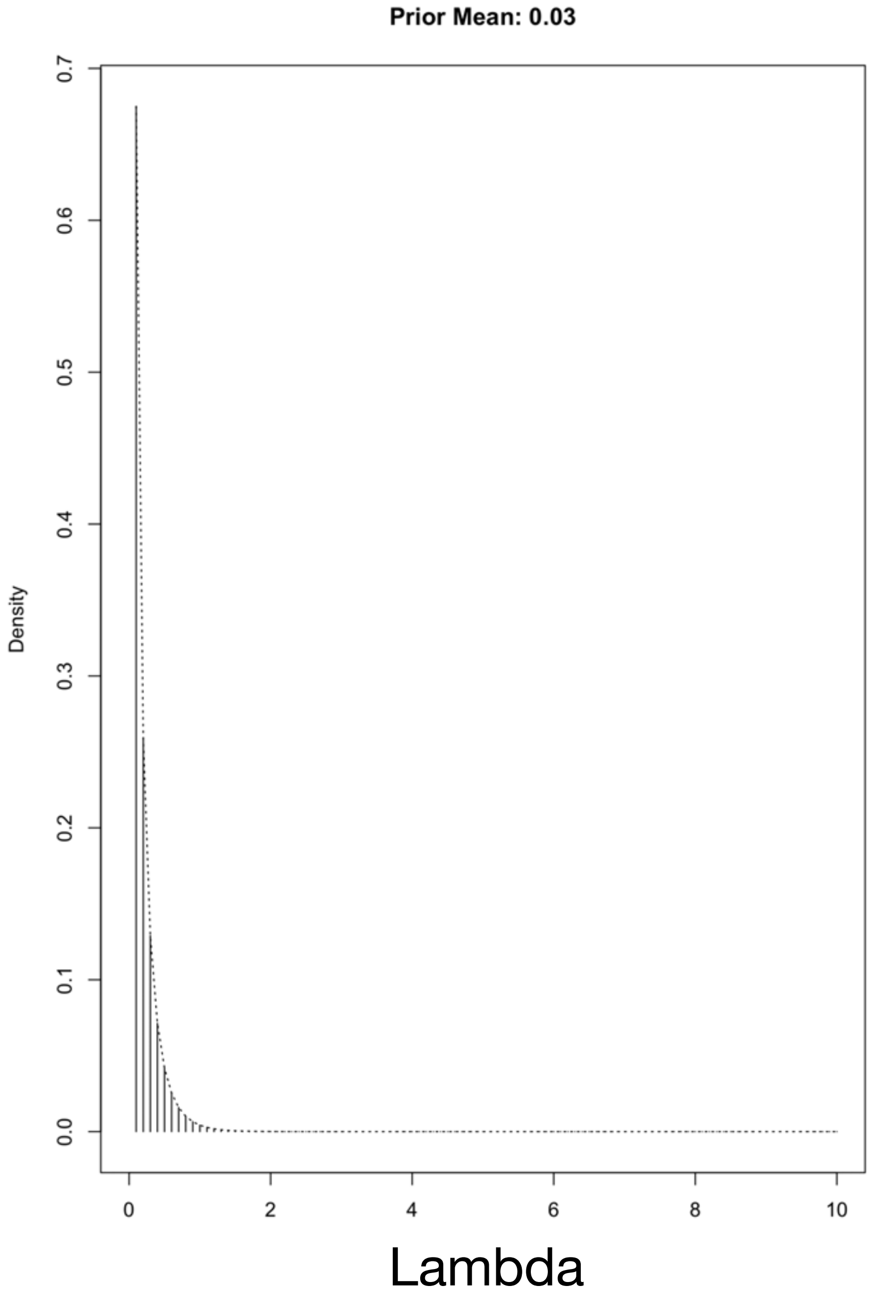


$$\hat{\lambda} = 8.4$$

Bayesian: Likelihood 2.0

Your knowledge

Prior

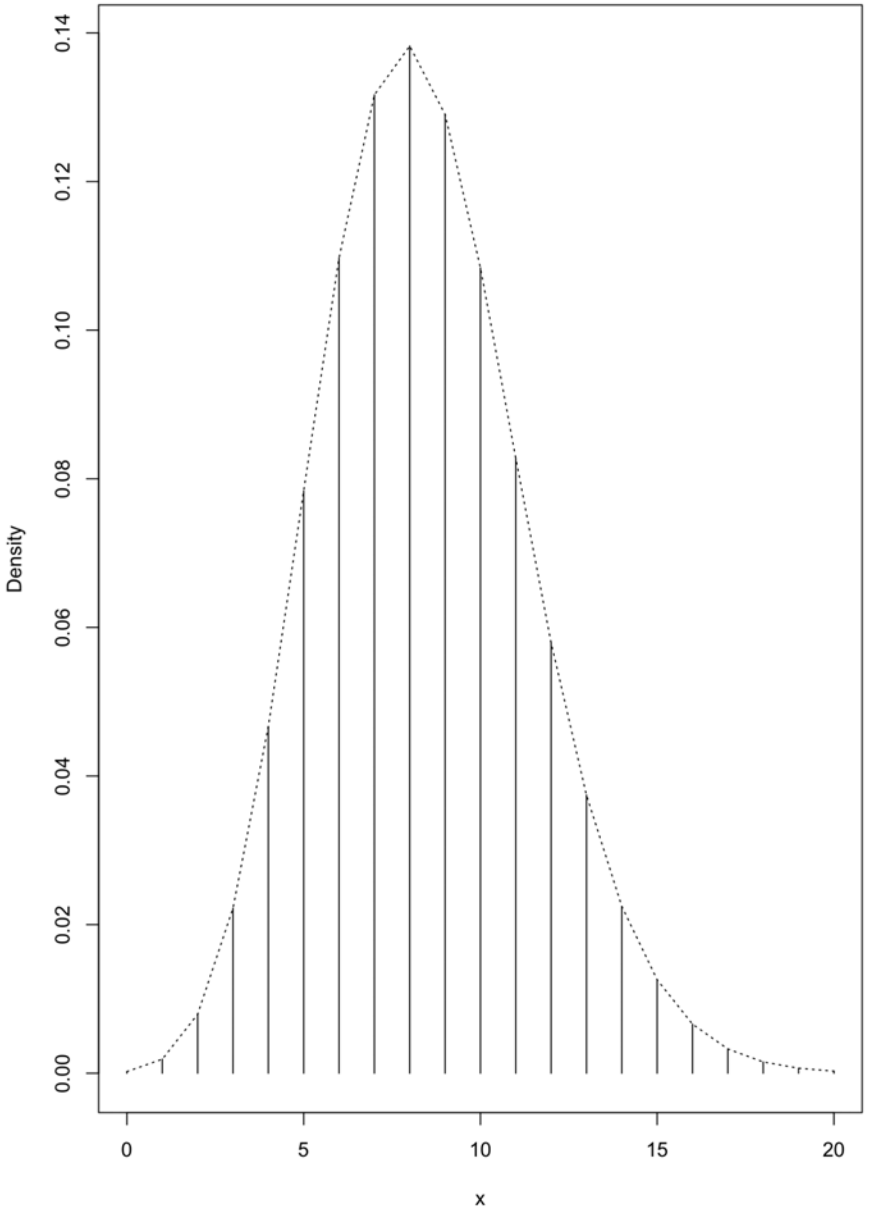


$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

+

Information in the data

Likelihood

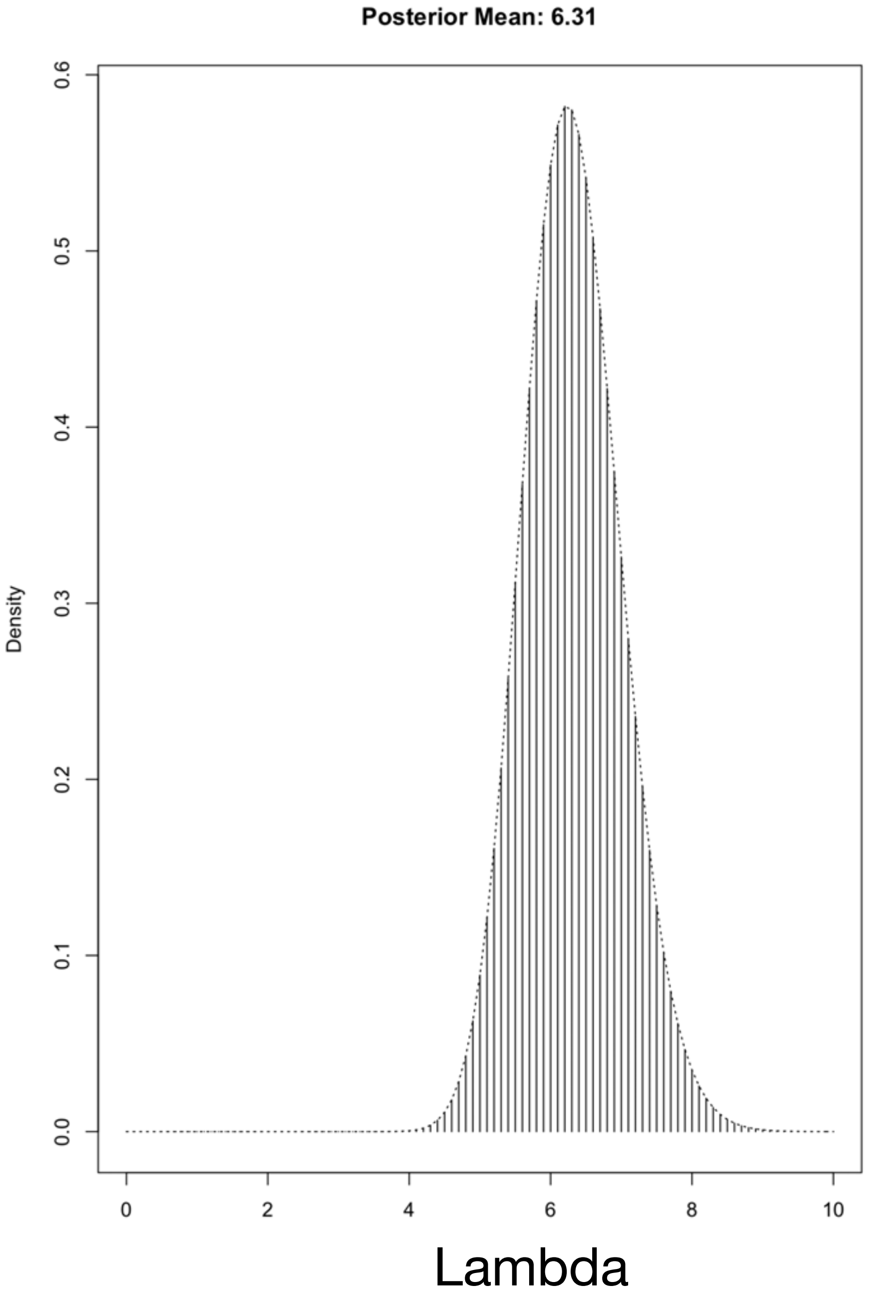


$$\hat{\lambda} = 8.4$$

=

Inference on parameter

Posterior



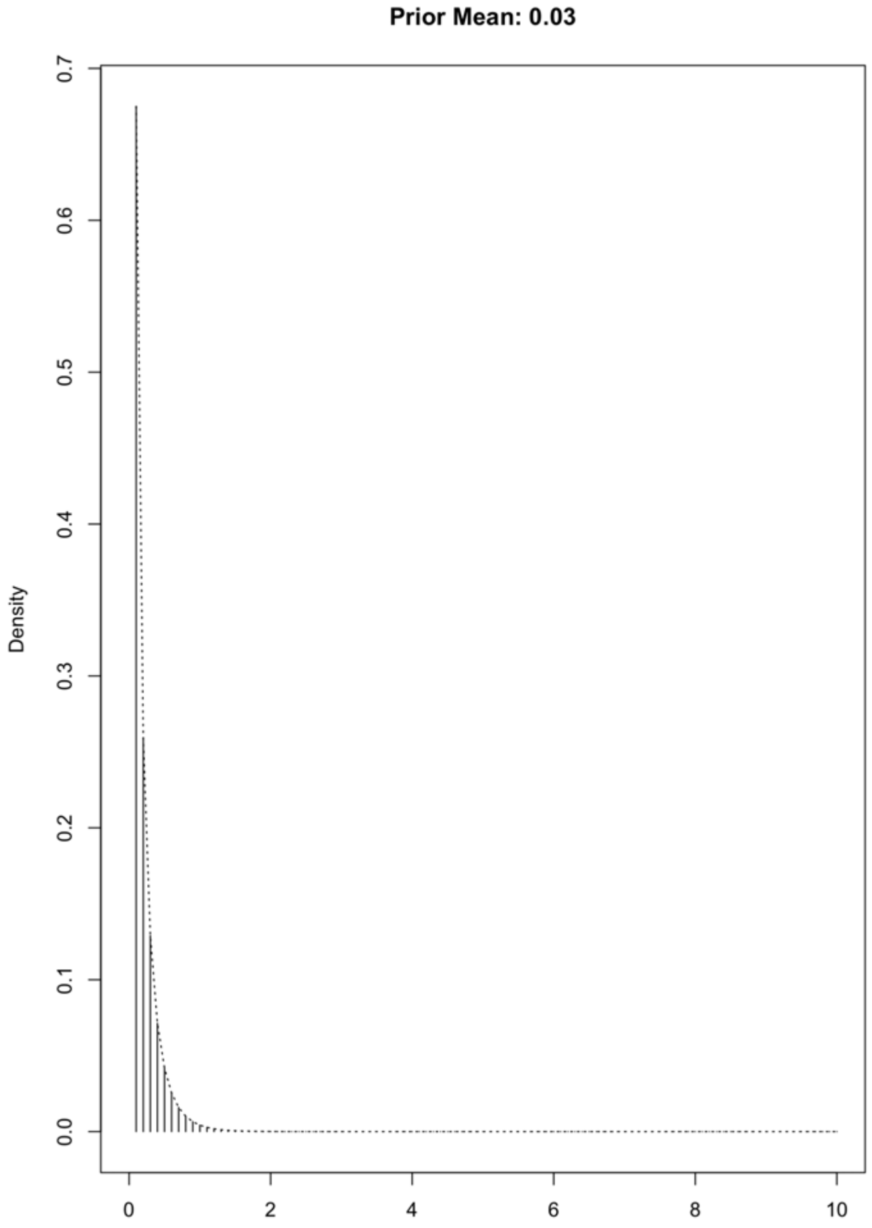
$$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior



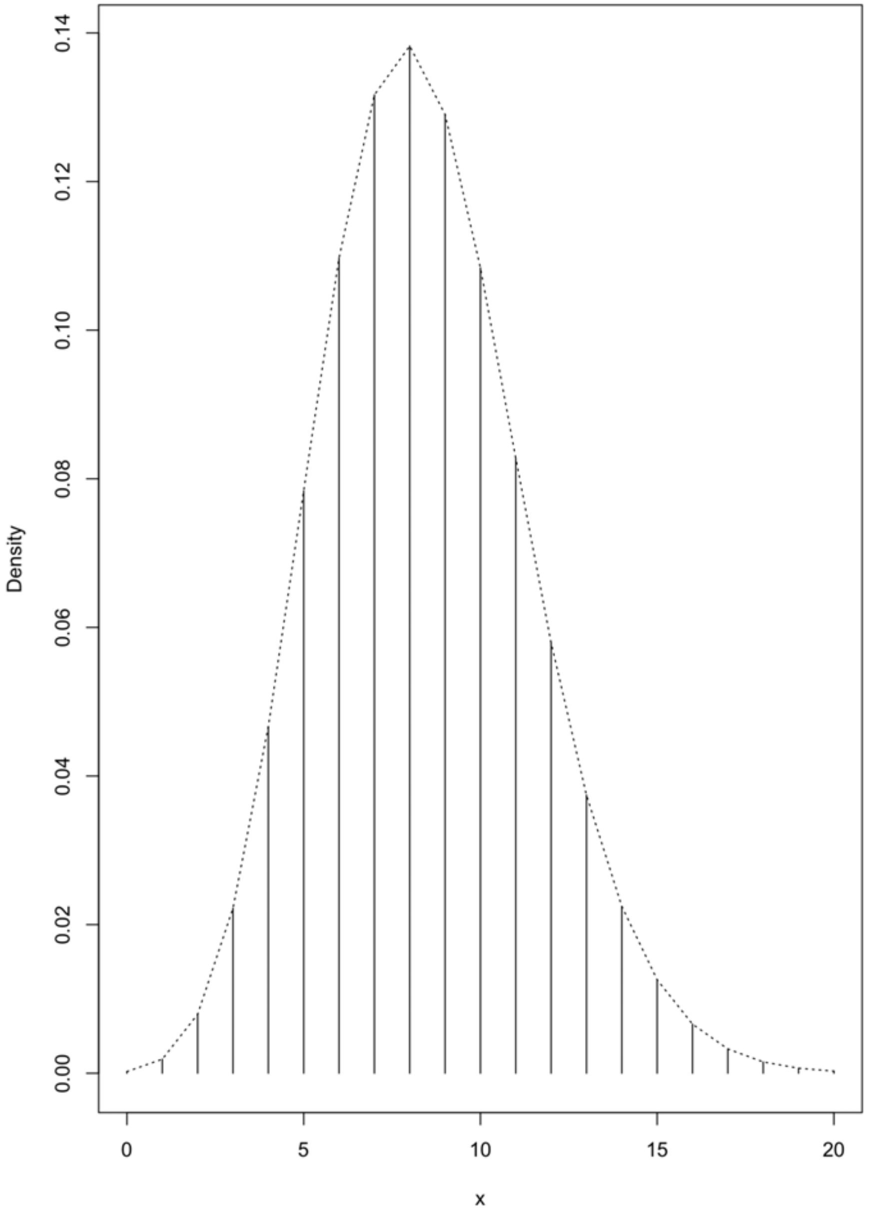
Lambda

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

+

Information in the data

Likelihood

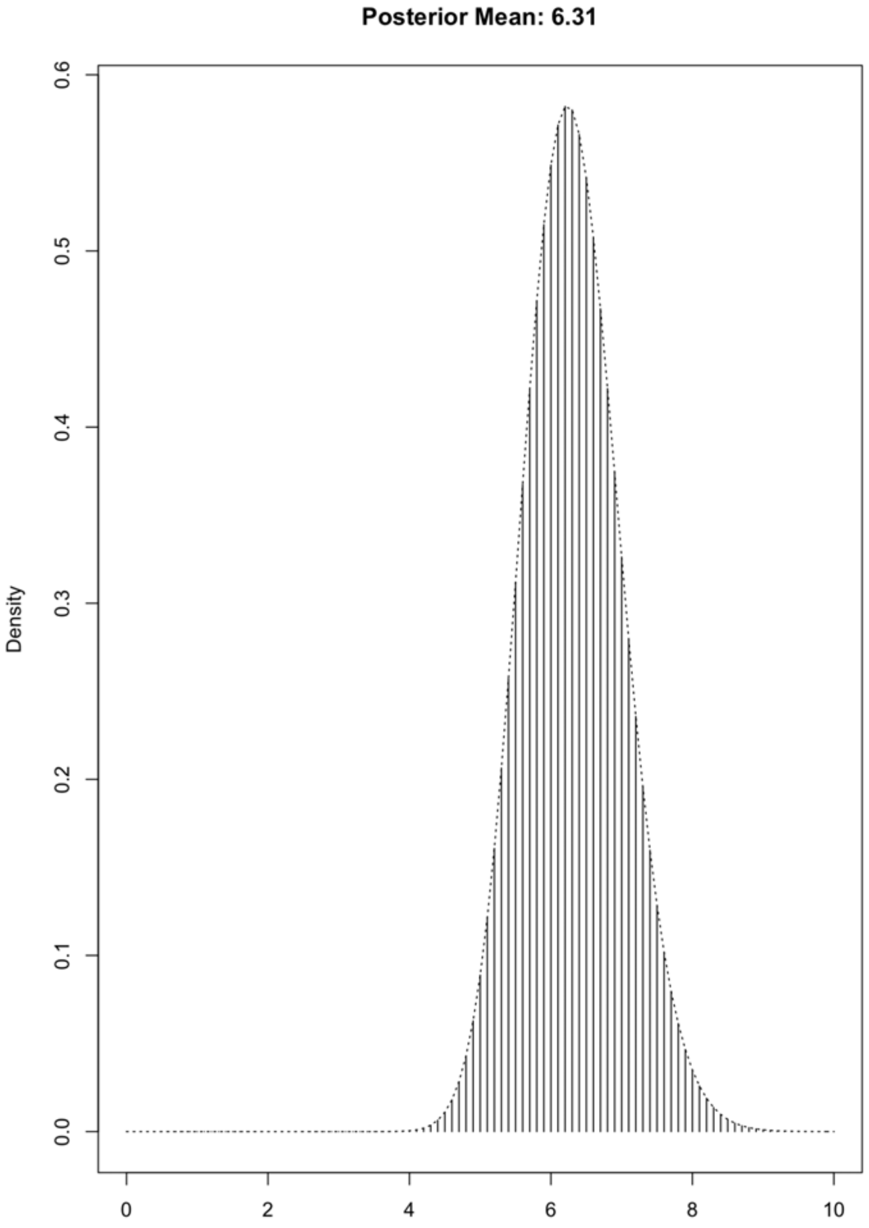


$$\hat{\lambda} = 8.4$$

=

Inference on parameter

Posterior



Lambda

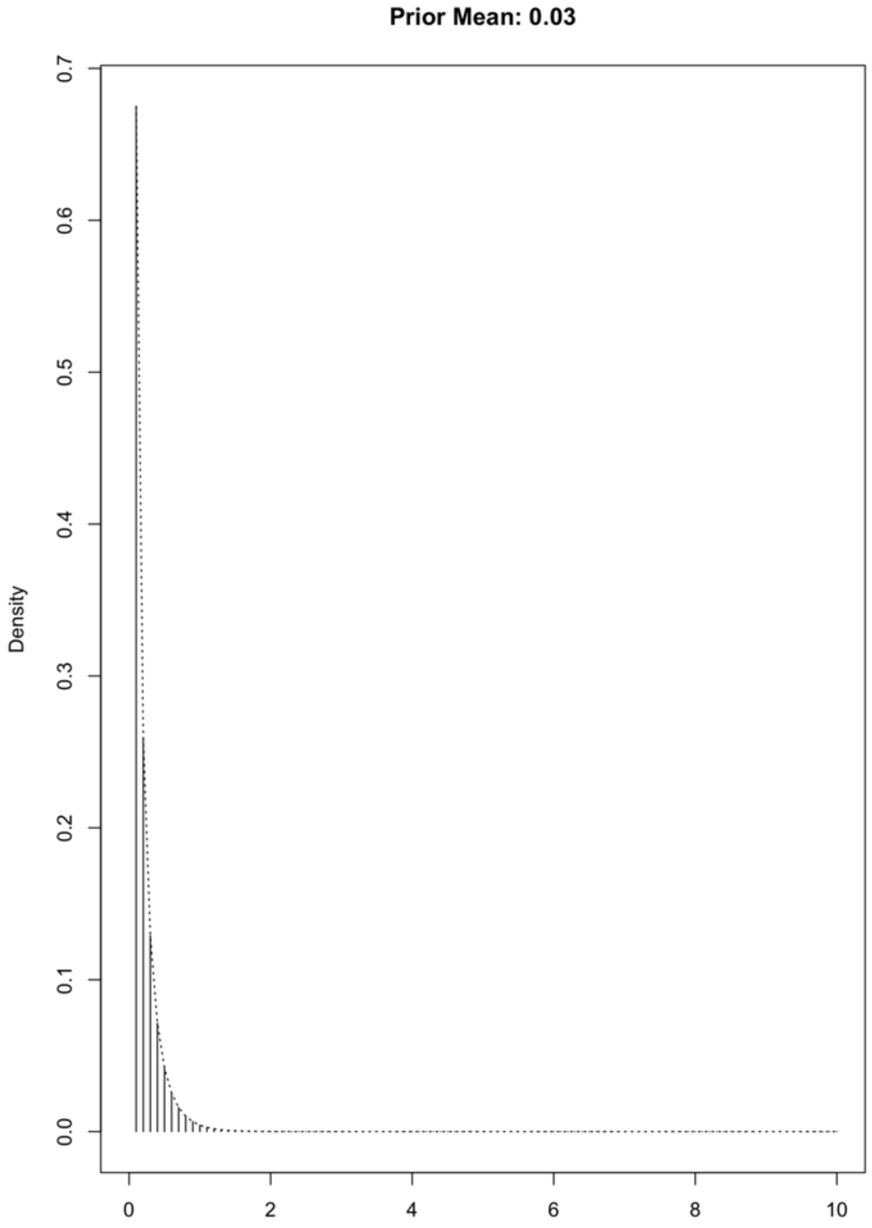
$$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

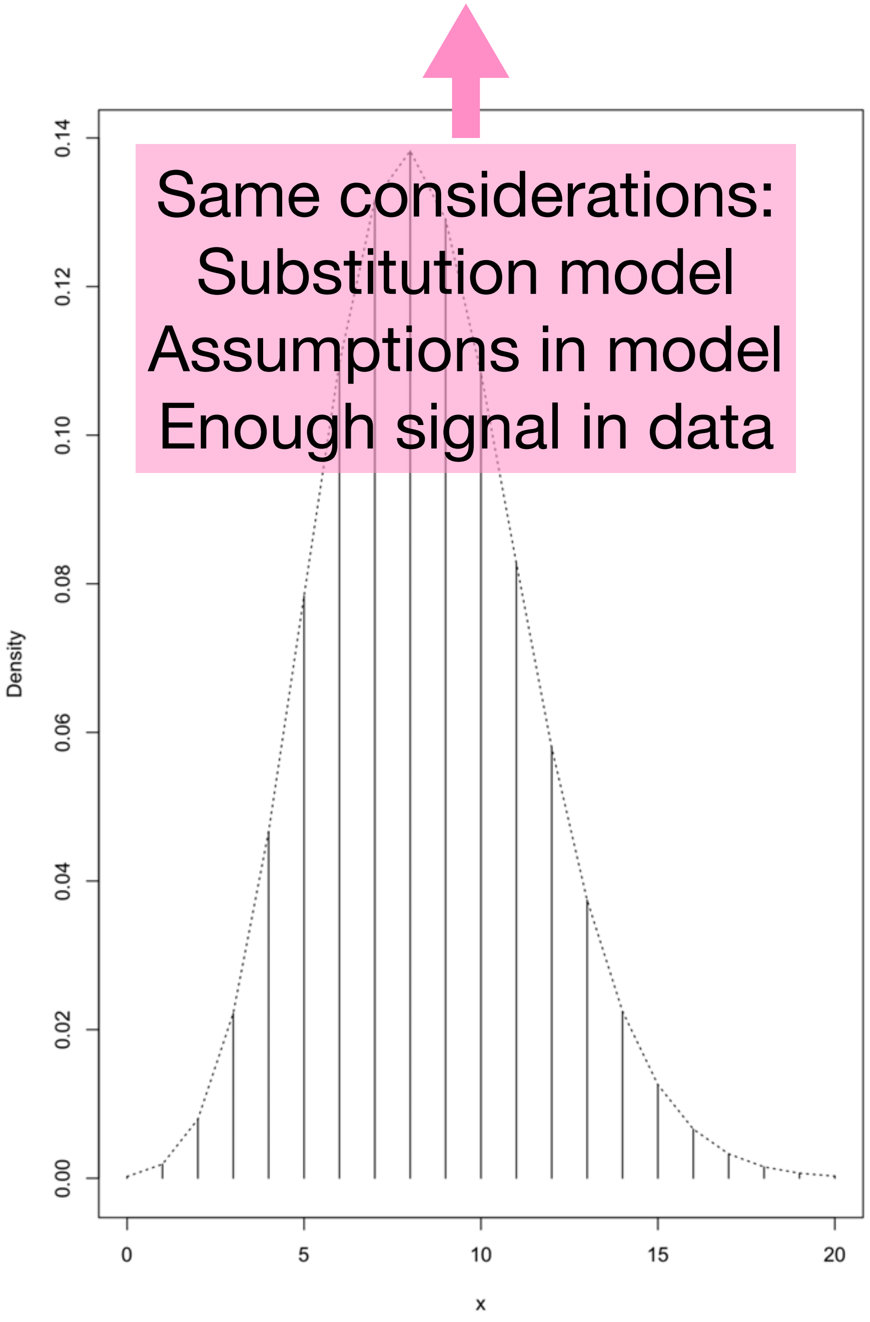
Your knowledge | **Information in the data** | **Inference on parameter**

Prior + **Likelihood** = **Posterior**



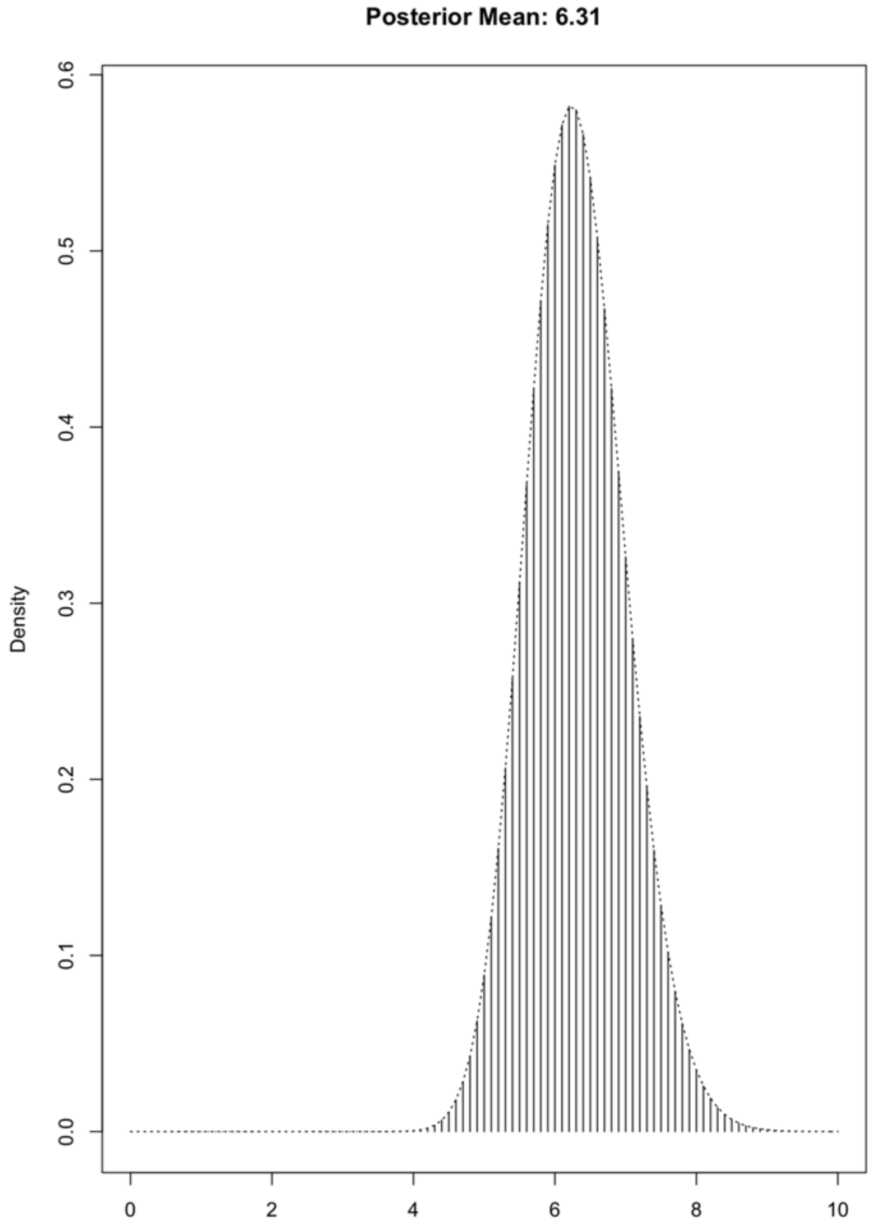
Lambda

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$



Same considerations:
Substitution model
Assumptions in model
Enough signal in data

$$\hat{\lambda} = 8.4$$



Lambda

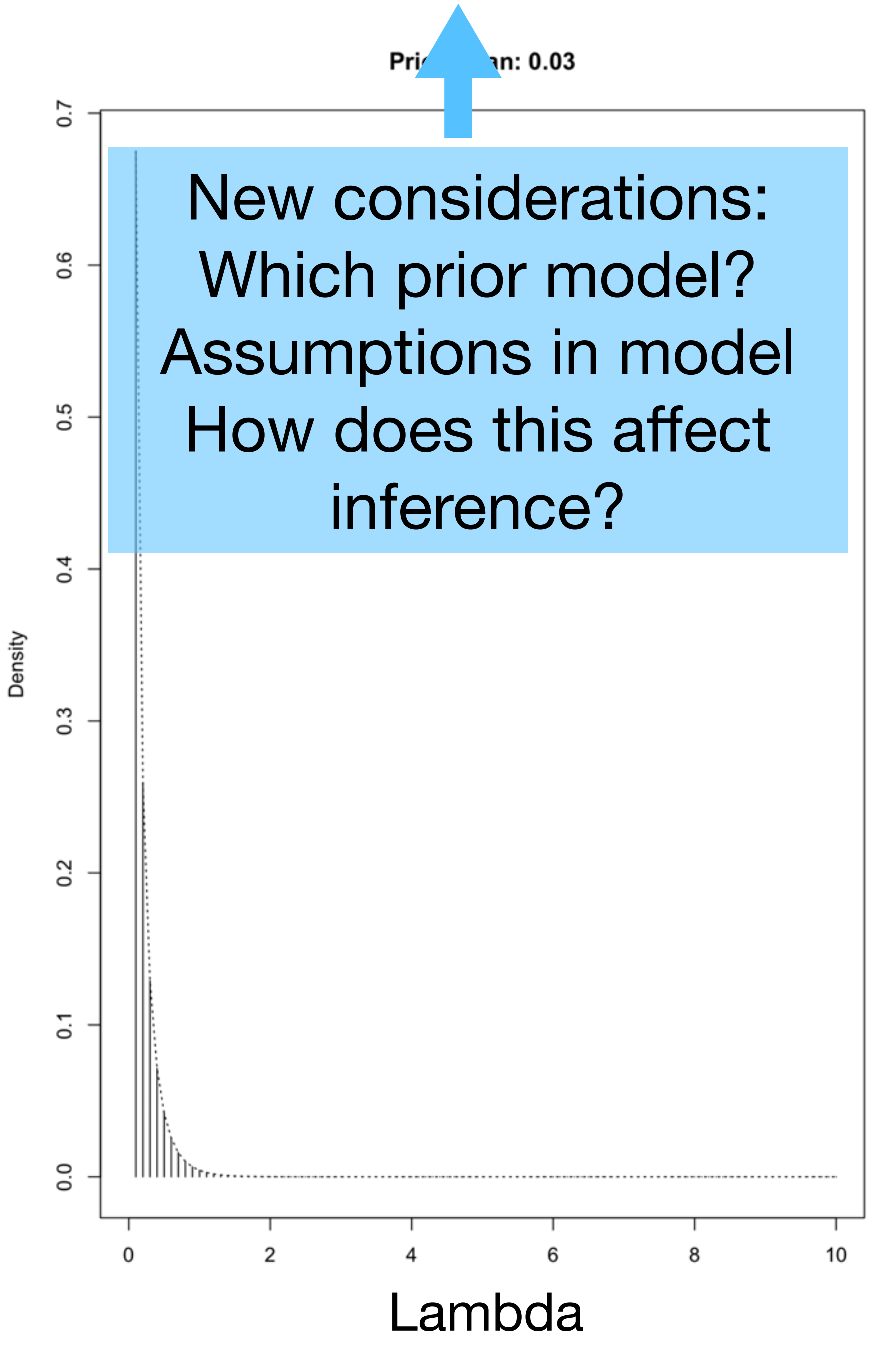
$$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

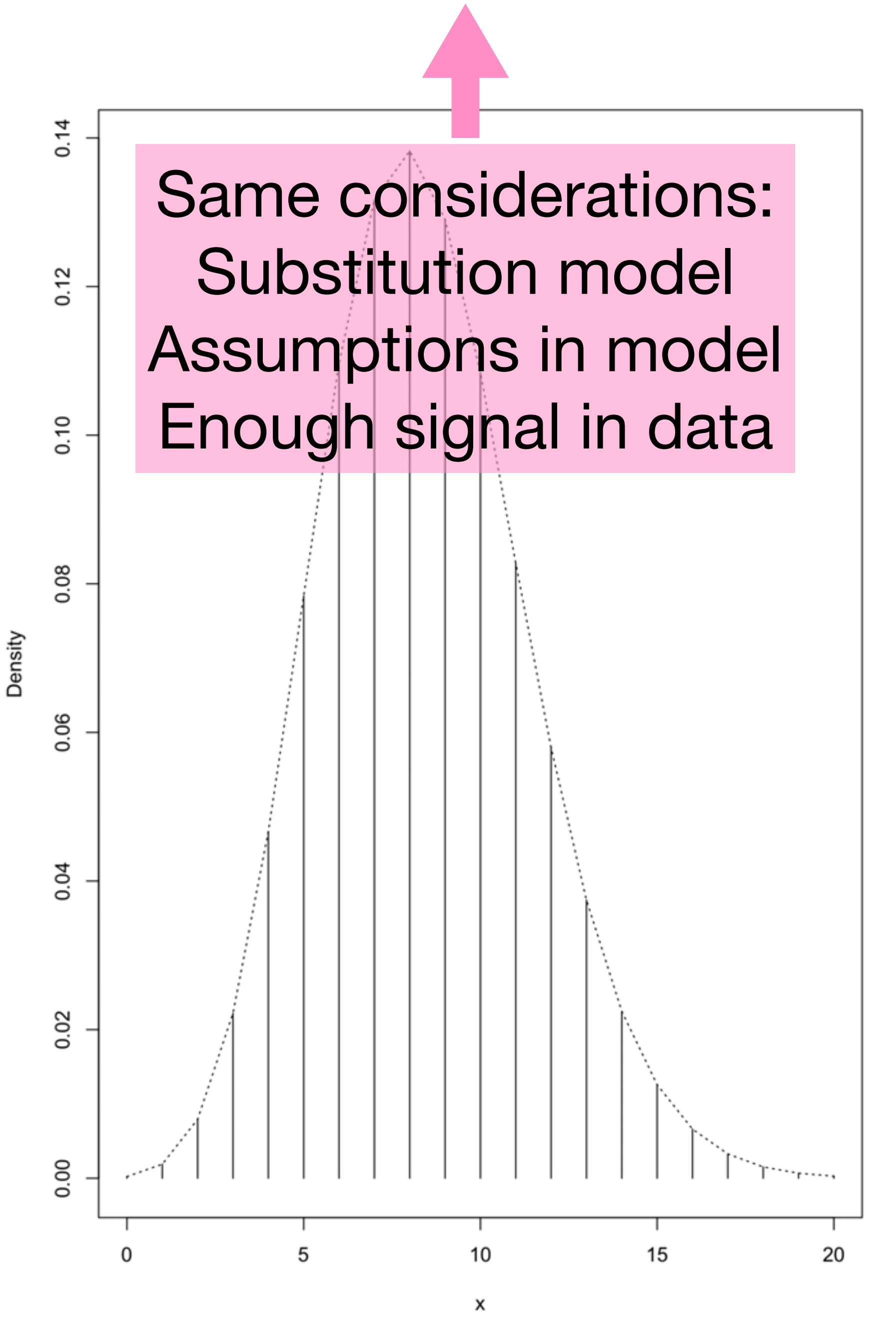
Your knowledge | **Information in the data** | **Inference on parameter**

Prior + **Likelihood** = **Posterior**



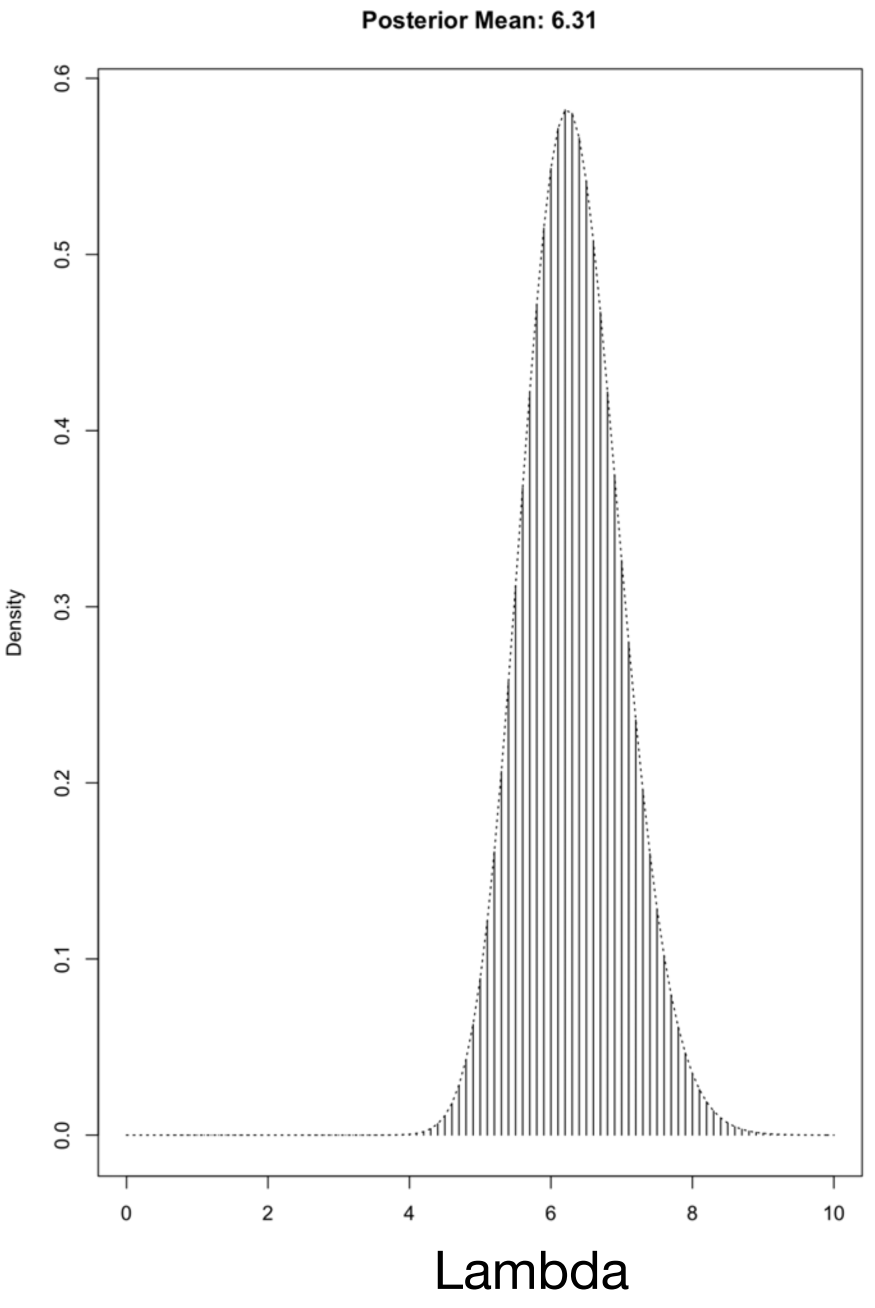
New considerations:
Which prior model?
Assumptions in model
How does this affect inference?

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

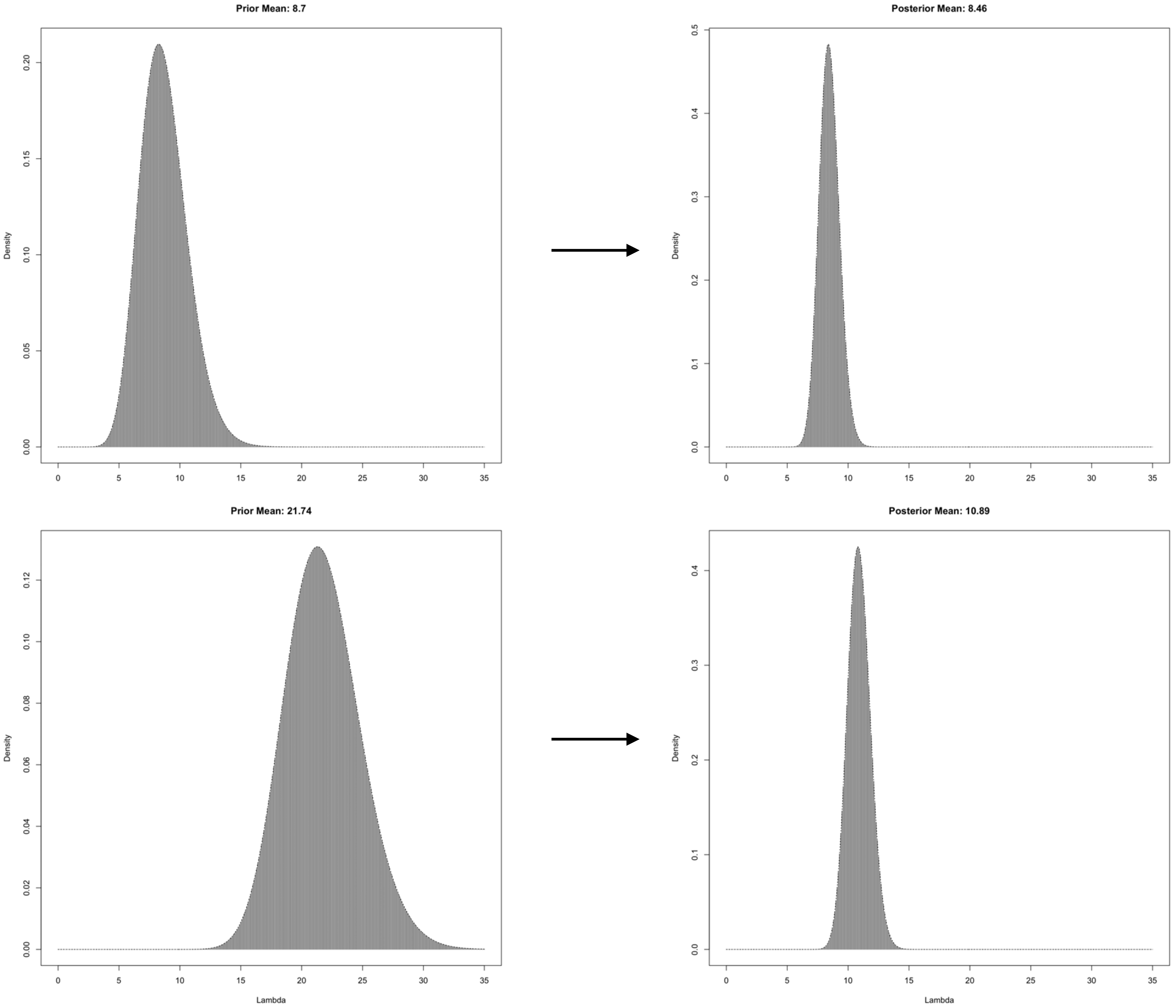


Same considerations:
Substitution model
Assumptions in model
Enough signal in data

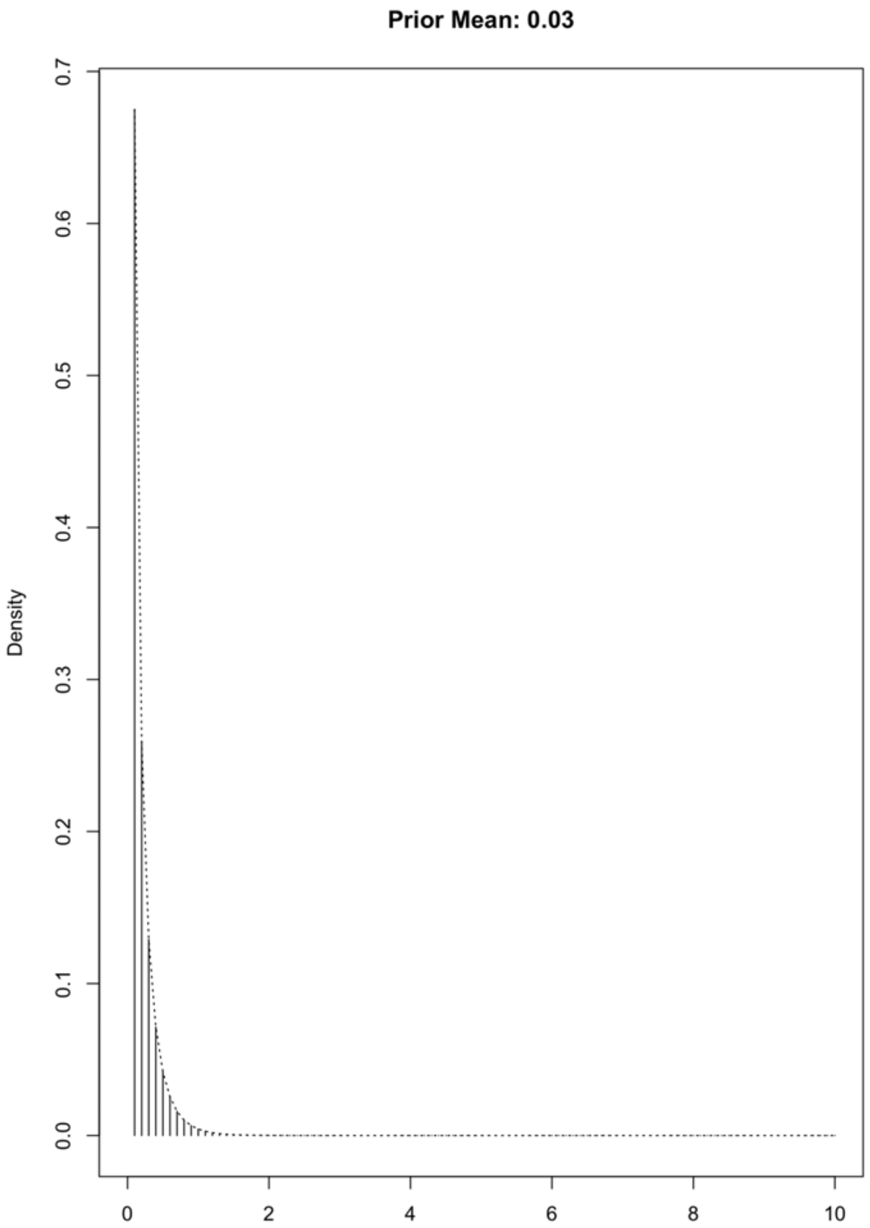
$$\hat{\lambda} = 8.4$$



$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$


Prior Mean: 8.7
Posterior Mean: 8.46
Prior Mean: 21.74
Posterior Mean: 10.89
Density
Lambda
0
5
10
15
20
25
30
35
0.00
0.05
0.10
0.15
0.20
0.0
0.1
0.2
0.3
0.4
0.5
0.02
0.04
0.06
0.08
0.12

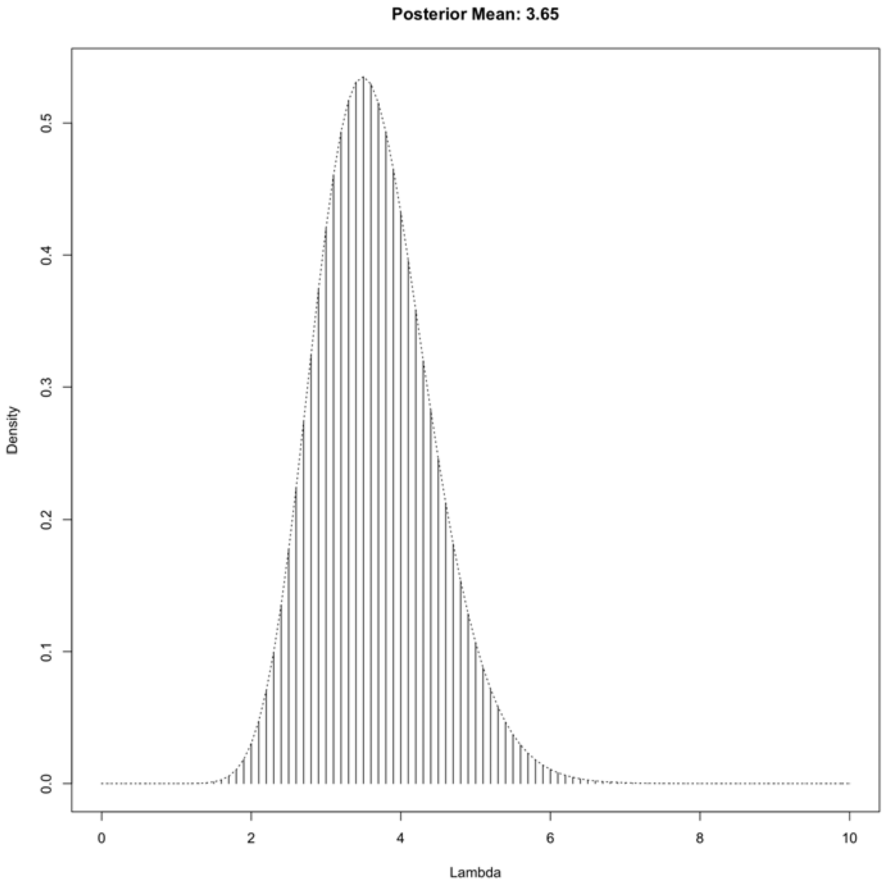
Prior



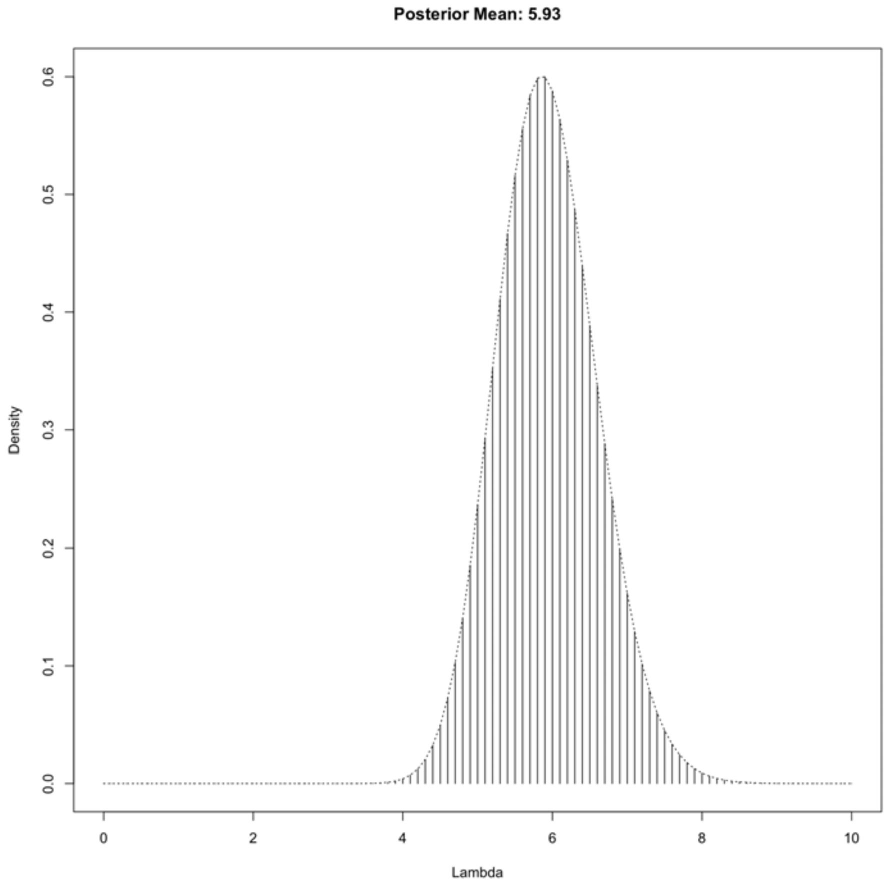
Lambda

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

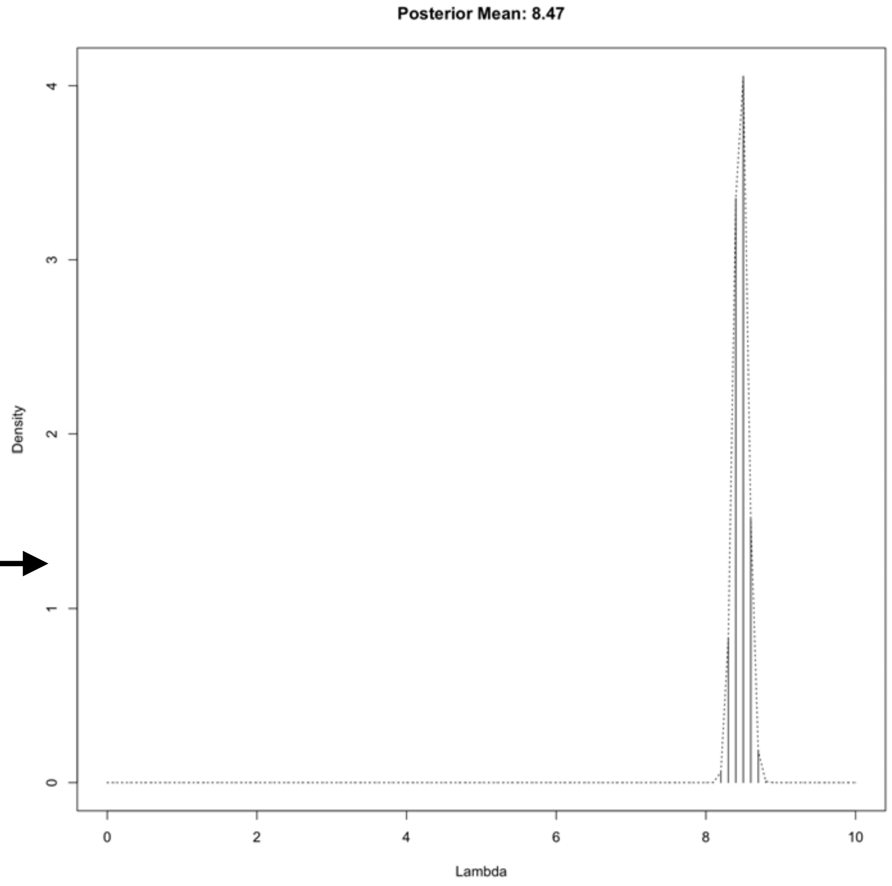
n=3



n=10



n=1000

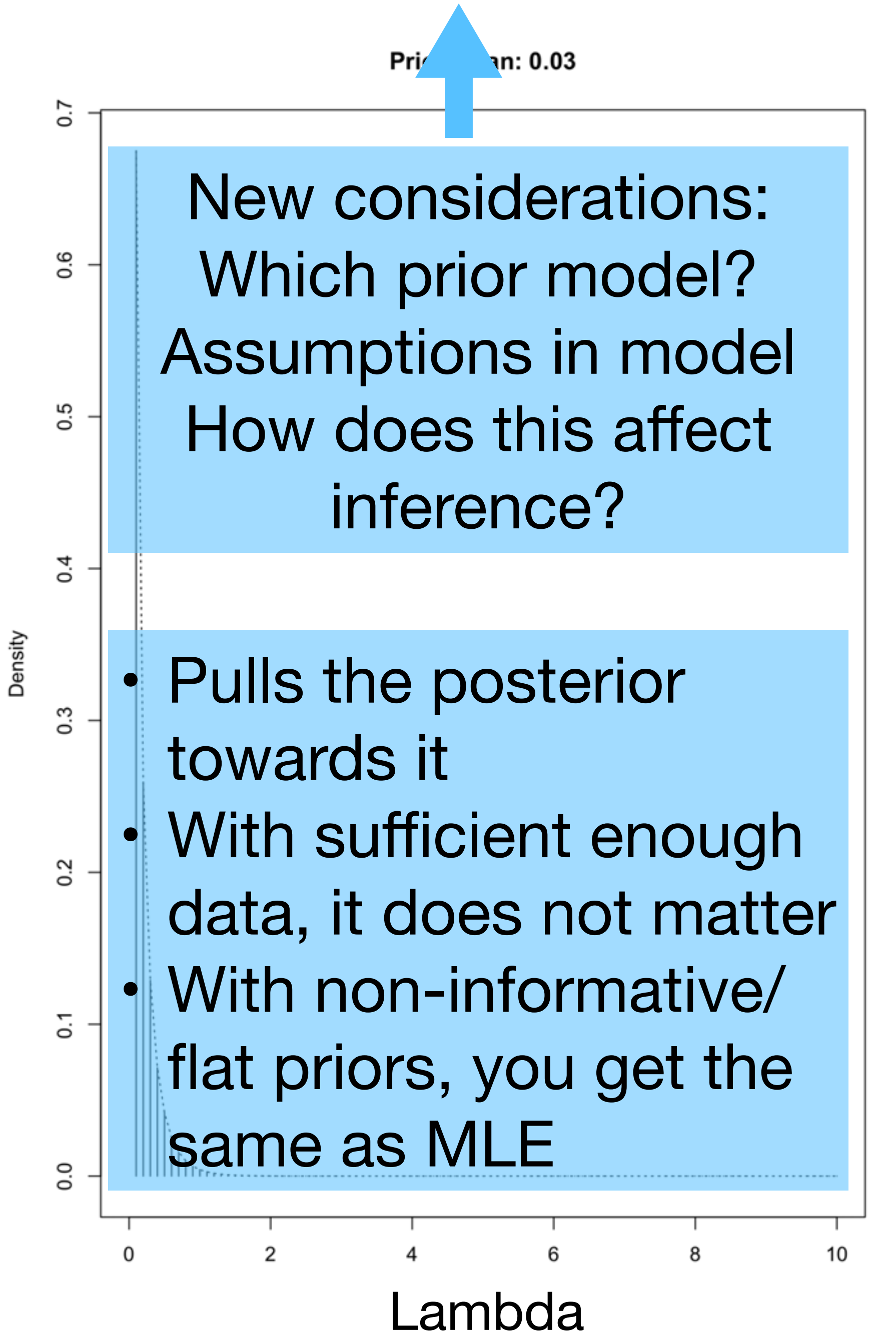


Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge | **Information in the data** | **Inference on parameter**

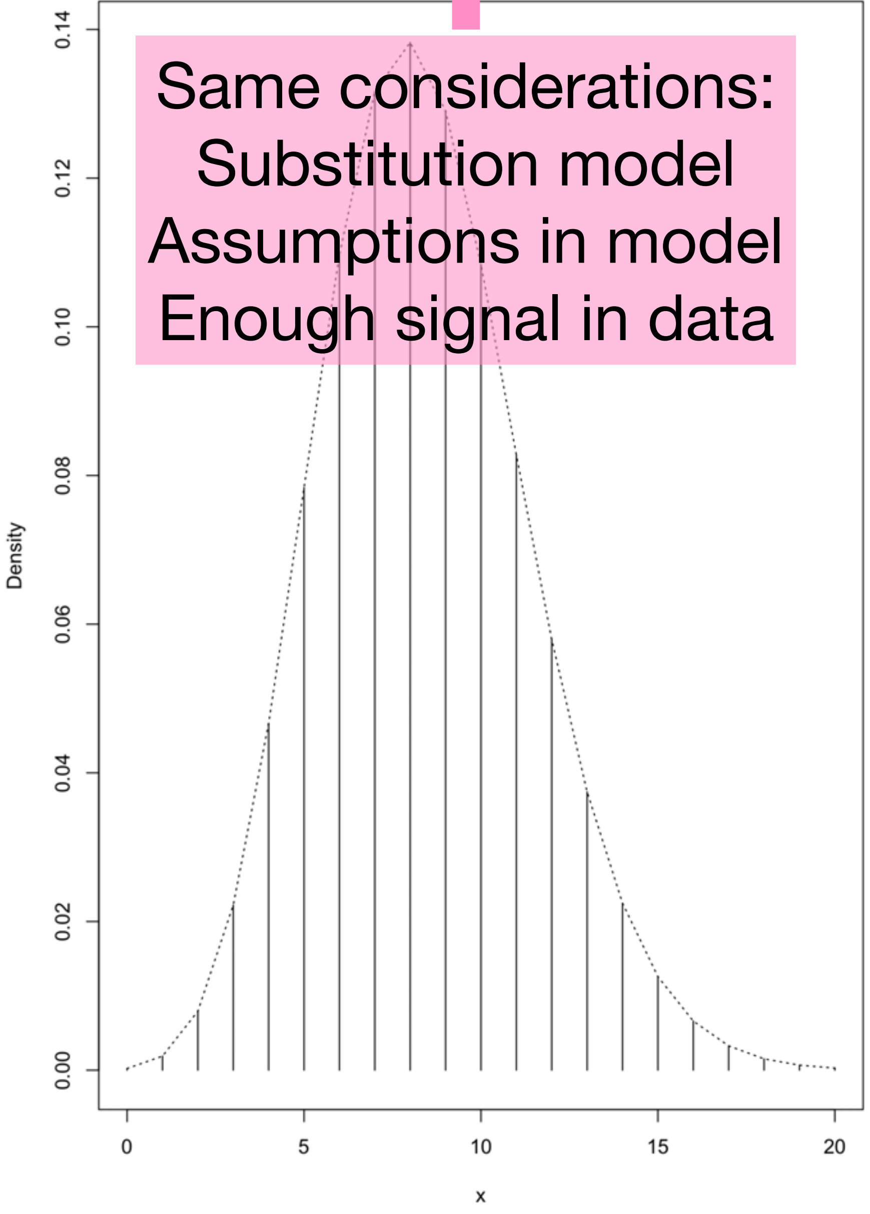
Prior + **Likelihood** = **Posterior**



New considerations:
Which prior model?
Assumptions in model
How does this affect inference?

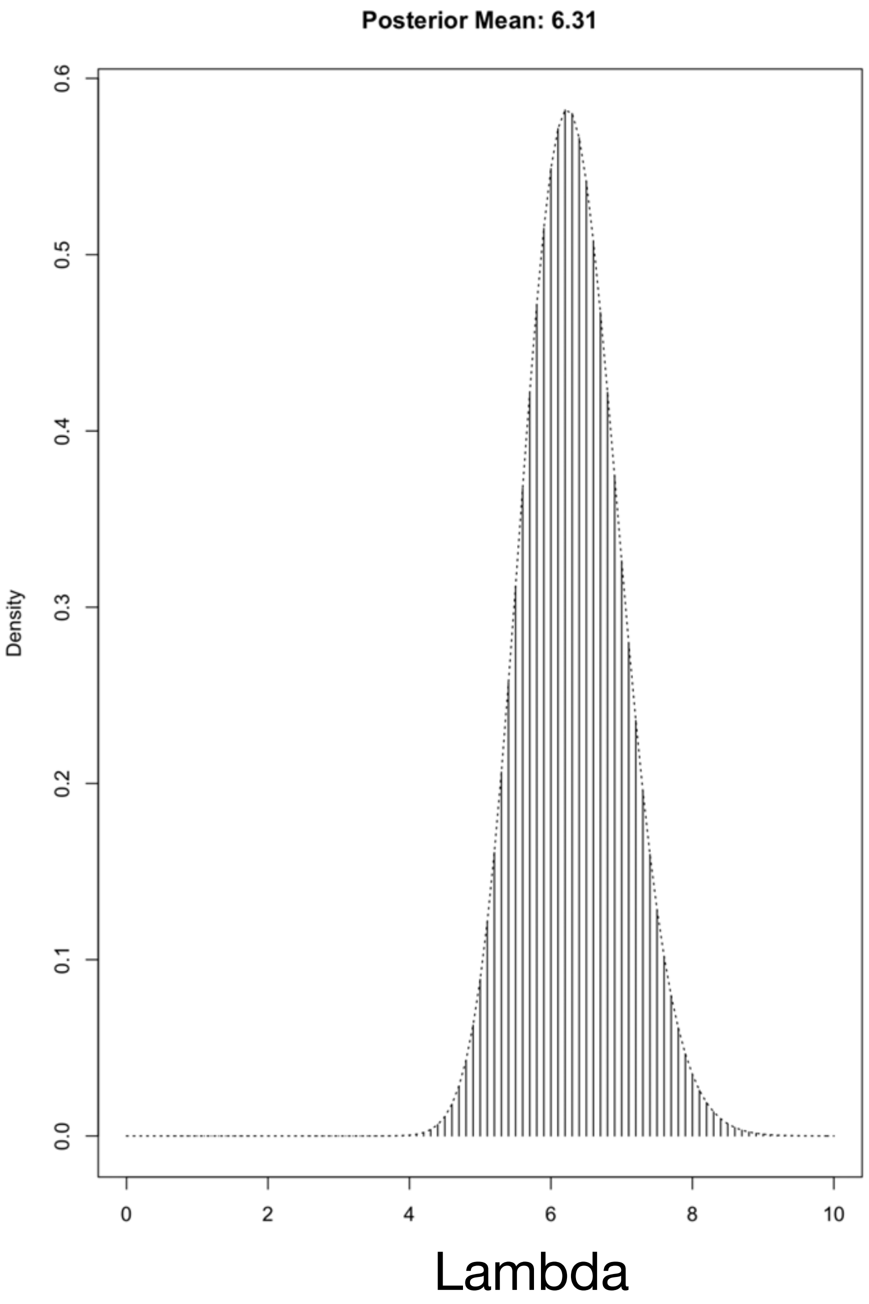
- Pulls the posterior towards it
- With sufficient enough data, it does not matter
- With non-informative/ flat priors, you get the same as MLE

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$



Same considerations:
Substitution model
Assumptions in model
Enough signal in data

$$\hat{\lambda} = 8.4$$



$$\lambda|\mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Intractable posterior distribution

$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Intractable posterior distribution

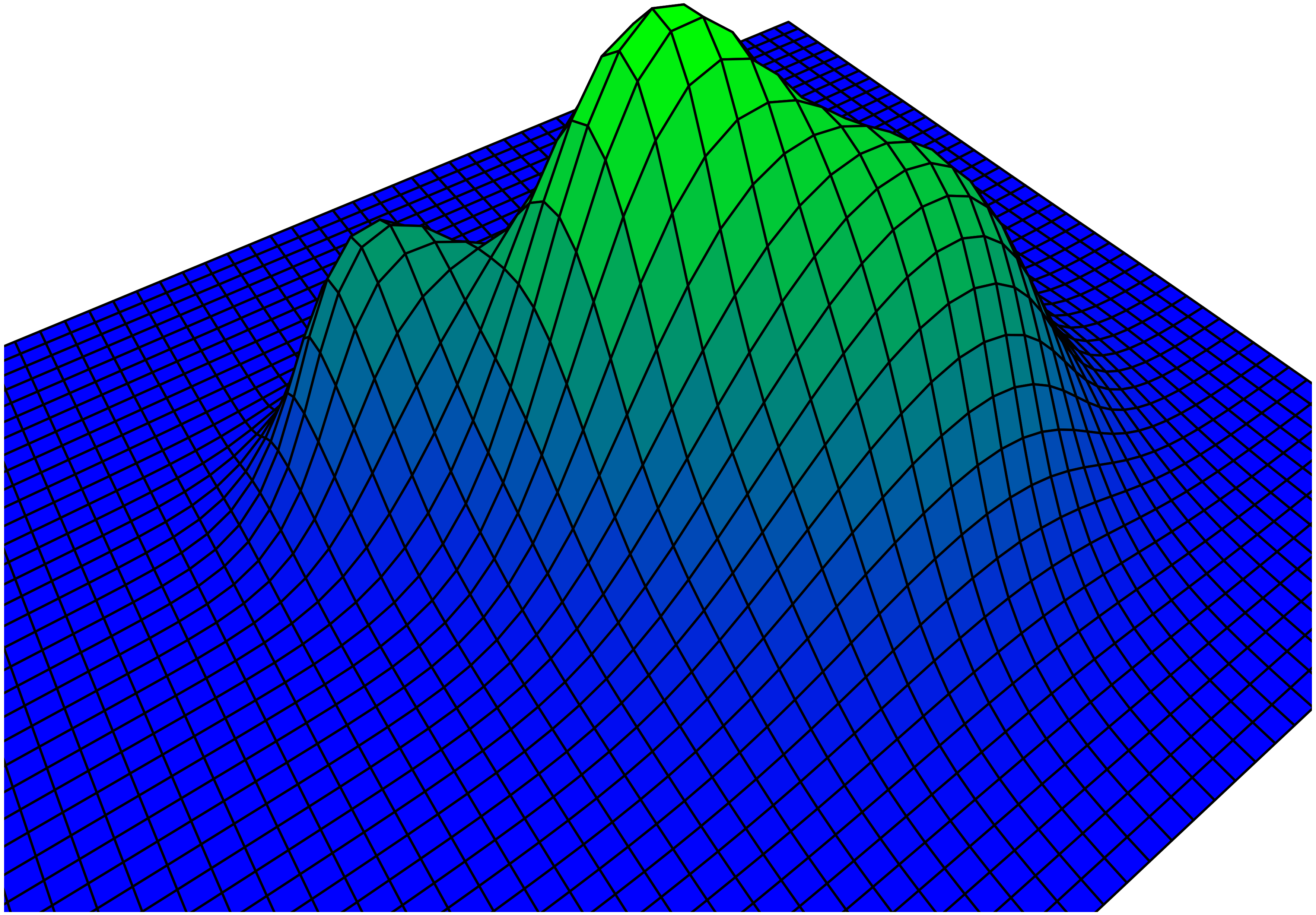
$$X_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Lognormal(\mu, \sigma)$$

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(- \frac{(\log \lambda - \mu)^2}{2\sigma^2} \right)$$

Not a known distribution anymore
We need to approximate it

MCMC: A way to approximate intractable posterior distributions



MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution

Usually symmetric, but could be asymmetric (Hasting ratio)

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \frac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution

Usually symmetric, but could be asymmetric (Hasting ratio)

Acceptance probability

Posterior ratio

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Proposal distribution

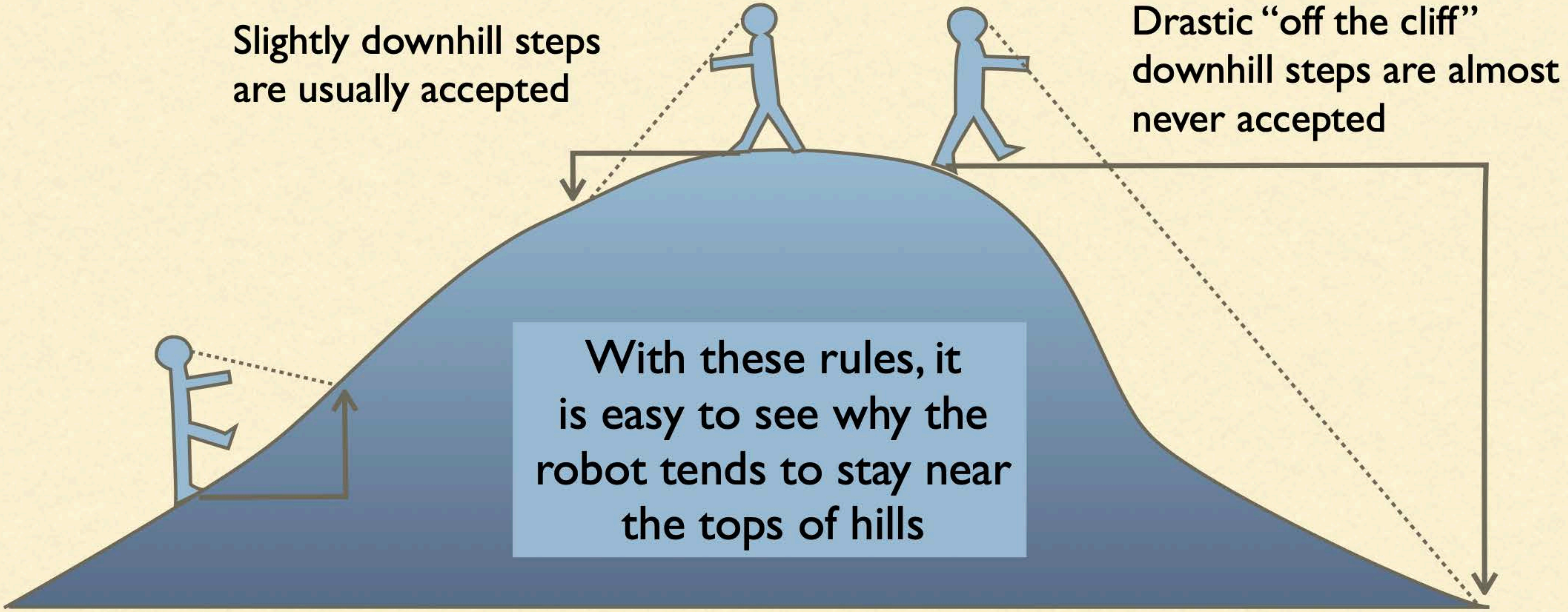
Usually symmetric, but could be asymmetric (Hasting ratio)

Acceptance probability

Posterior ratio

The chain tends to stay on regions of high posterior

MCMC robot's rules



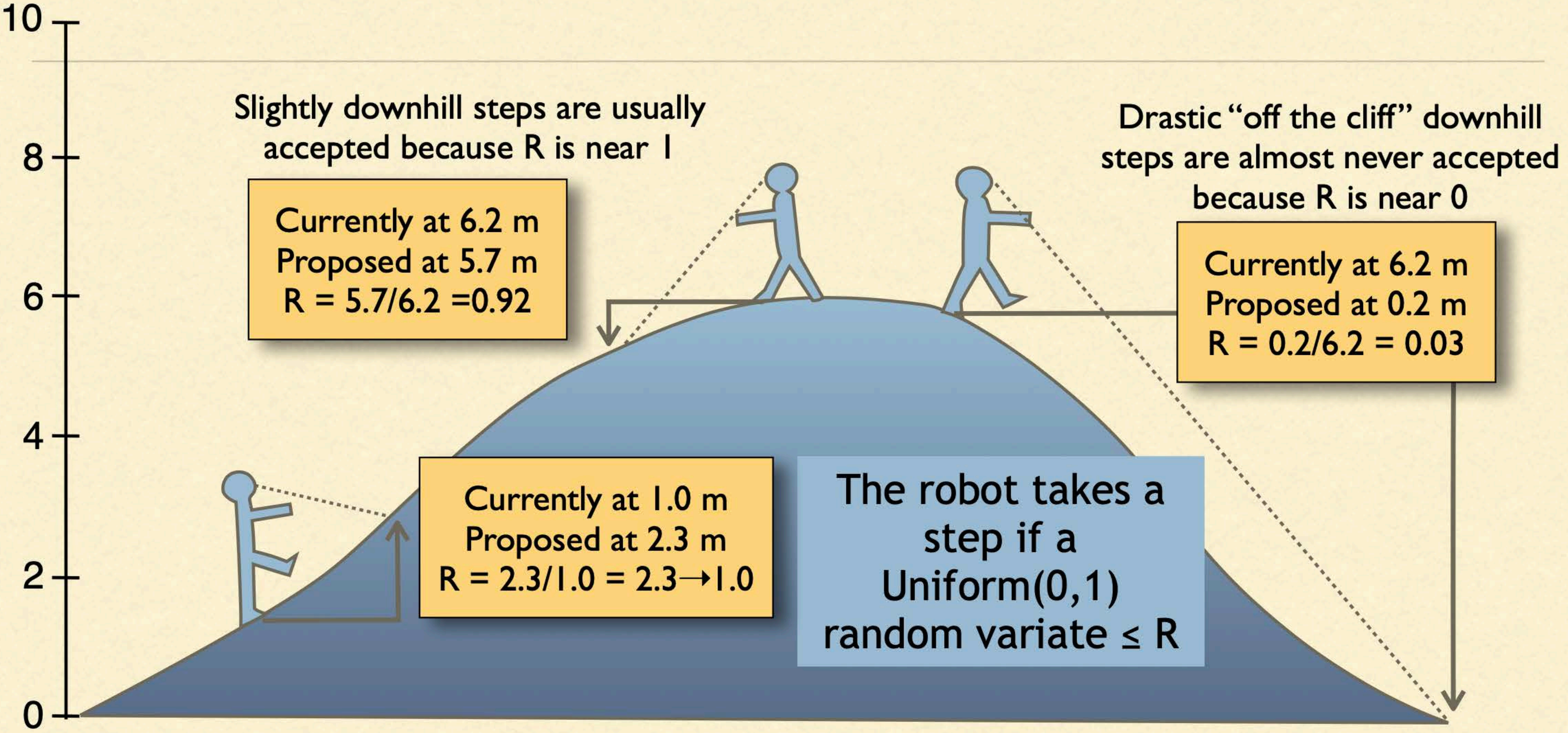
Slightly downhill steps
are usually accepted

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

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

Uphill steps are
always accepted

Actual rules (Metropolis algorithm)



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→1.0

The robot takes a step if a Uniform(0,1) random variate ≤ R

Uphill steps are always accepted because R = 1

Metropolis et al. 1953. Equation of state calculations by fast computing machines. *J. Chem. Physics* 21(6):1087-1092.

Hasting ratio

$$\alpha = \left[\frac{P(\lambda^*|\mathbf{X})}{P(\lambda|\mathbf{X})}\right] \left[\frac{q(\lambda|\lambda^*)}{q(\lambda^*|\lambda)}\right] \longleftarrow \text{Proposal distribution}$$

Posterior ratio **Hasting ratio**

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

Let’s implement the MCMC algorithm for the example you went over with Rosana

Try it out yourself first and then we can check out the solution: https://go.wisc.edu/k5n7su

MCMC considerations

- Choice of priors
- Mixing
- Convergence
- Burnin

MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence
- Burnin

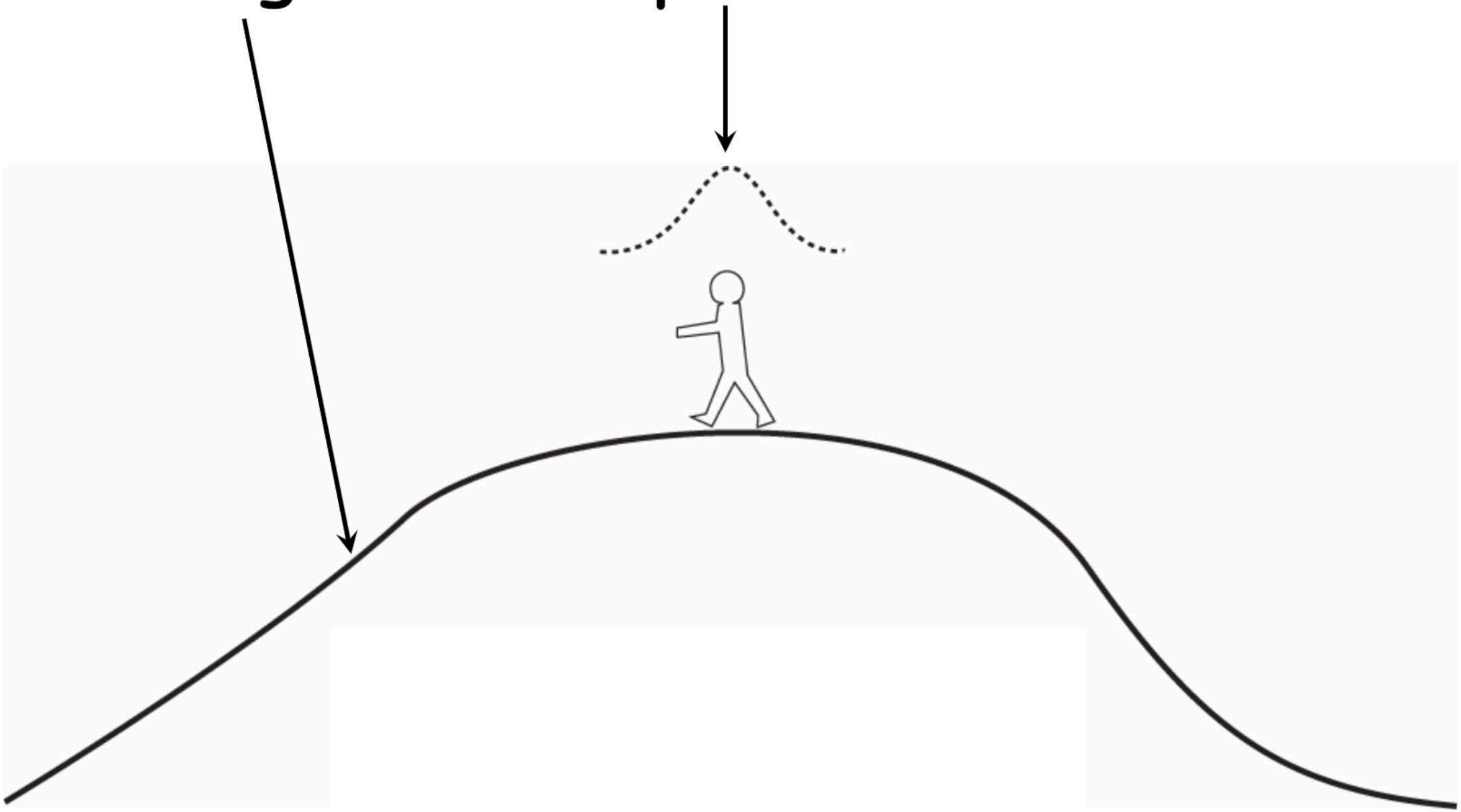
MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin

MCMC considerations

- Choice of priors
- Mixing ← How well you navigate the parameter space?
- Convergence ← How well you reach all regions of high posterior values?
- Burnin ← How long it takes to reach regions of high posterior values?

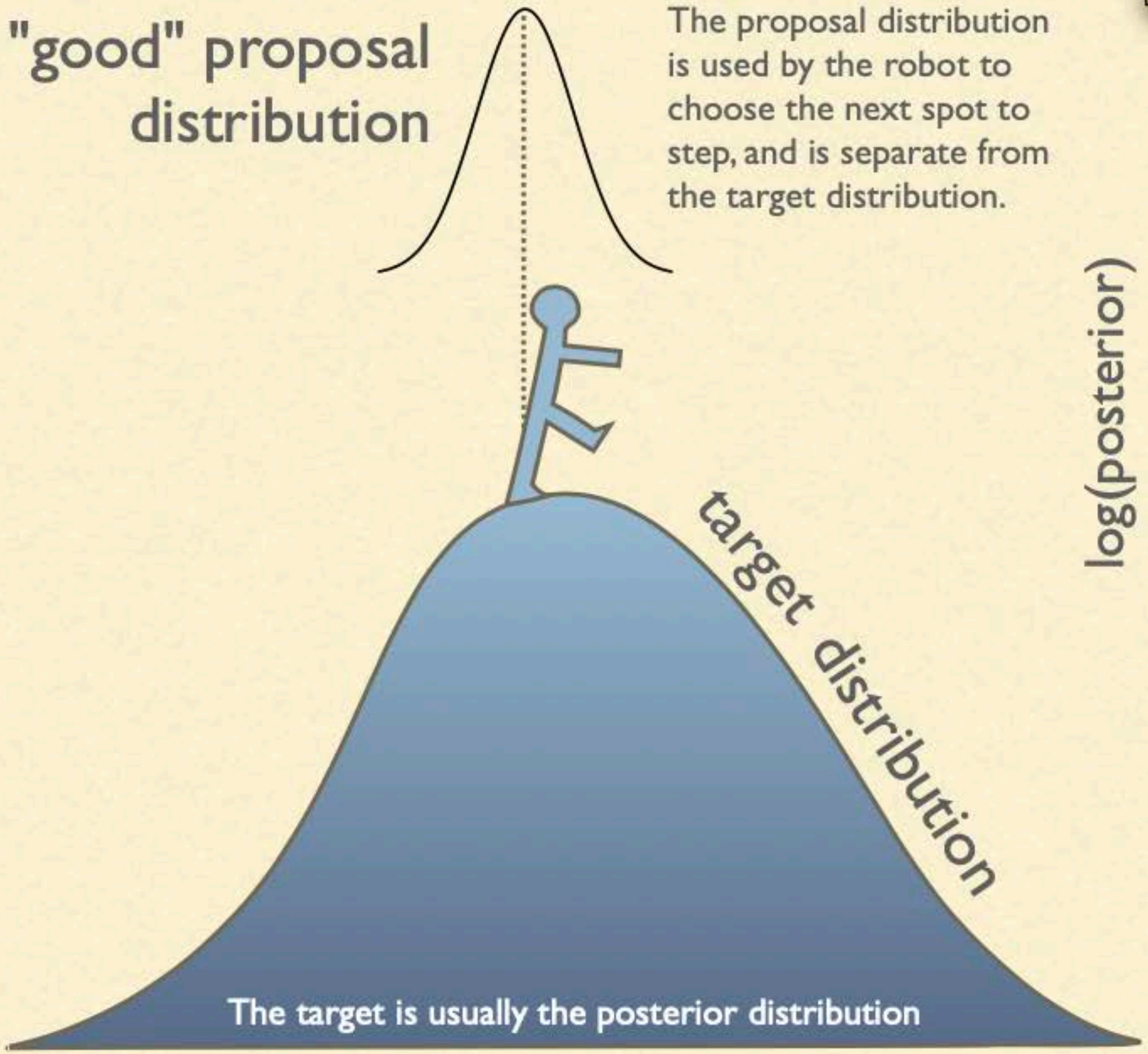
Target vs. Proposal Distributions



Target vs. Proposal Distributions

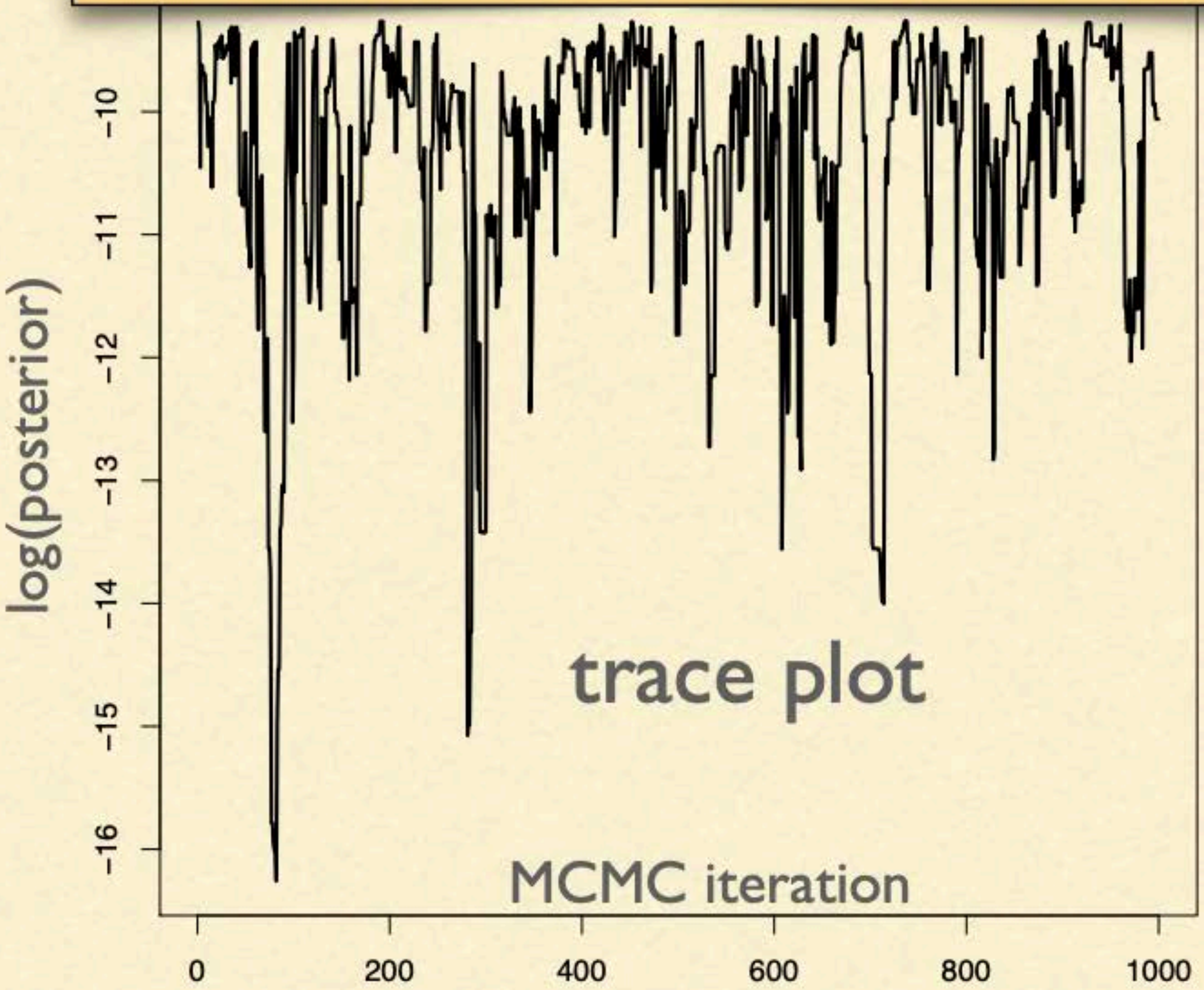
"good" proposal distribution

The proposal distribution is used by the robot to choose the next spot to step, and is separate from the target distribution.



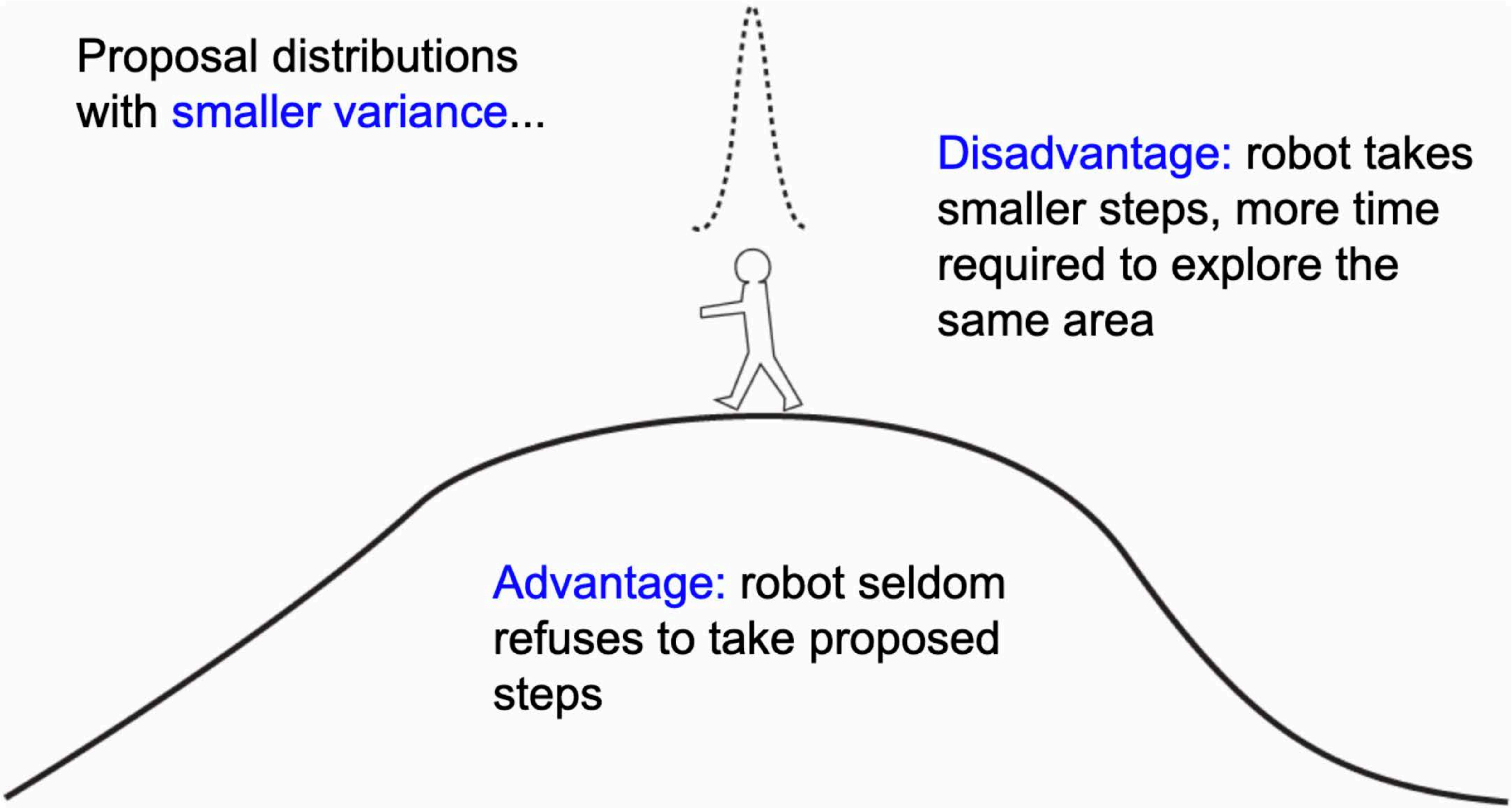
The target is usually the posterior distribution

Tracer (app for generating trace plots from MCMC output):
https://github.com/beast-dev/tracer/releases/tag/v1.7.2



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

Target vs. Proposal Distributions

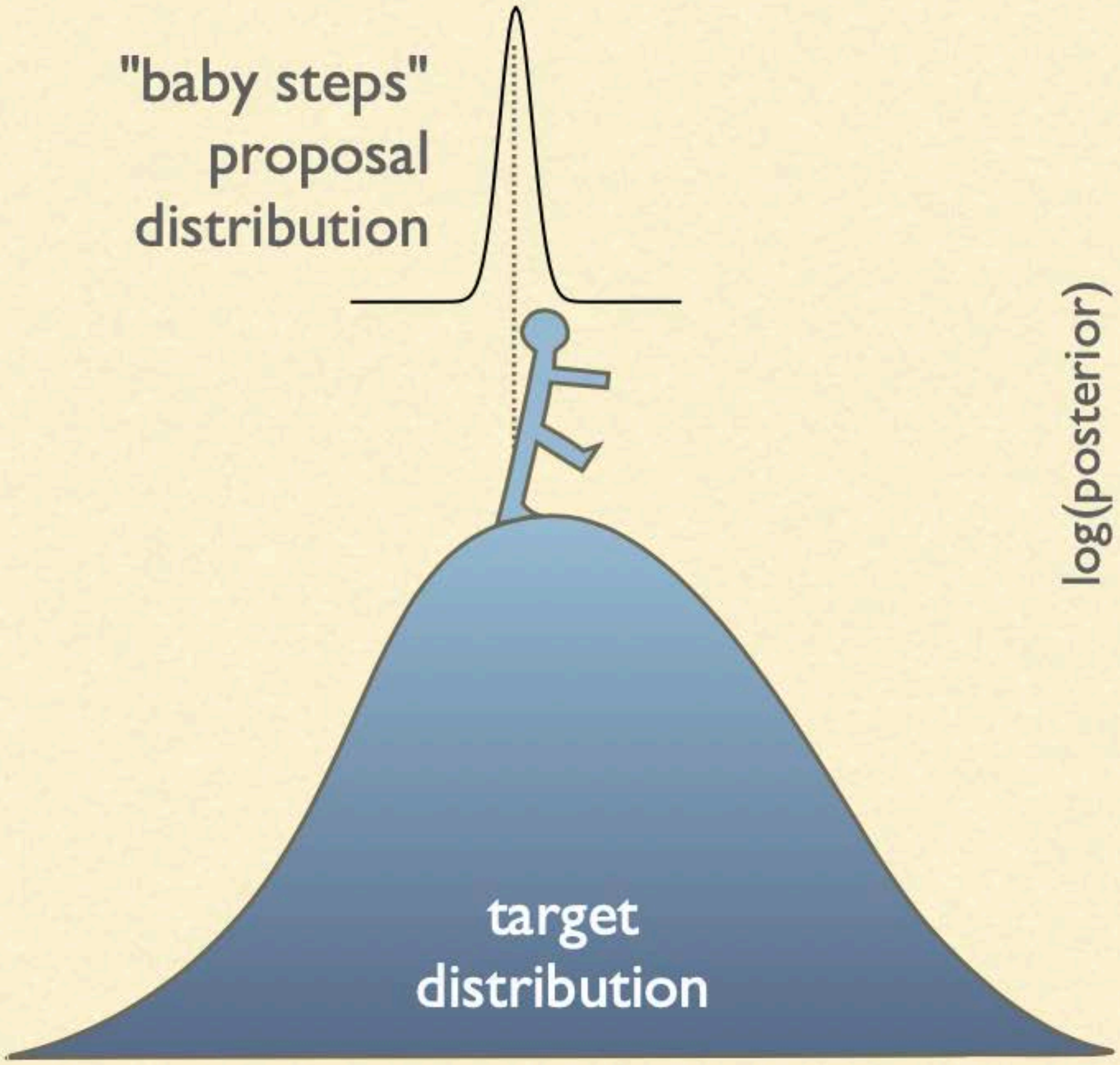


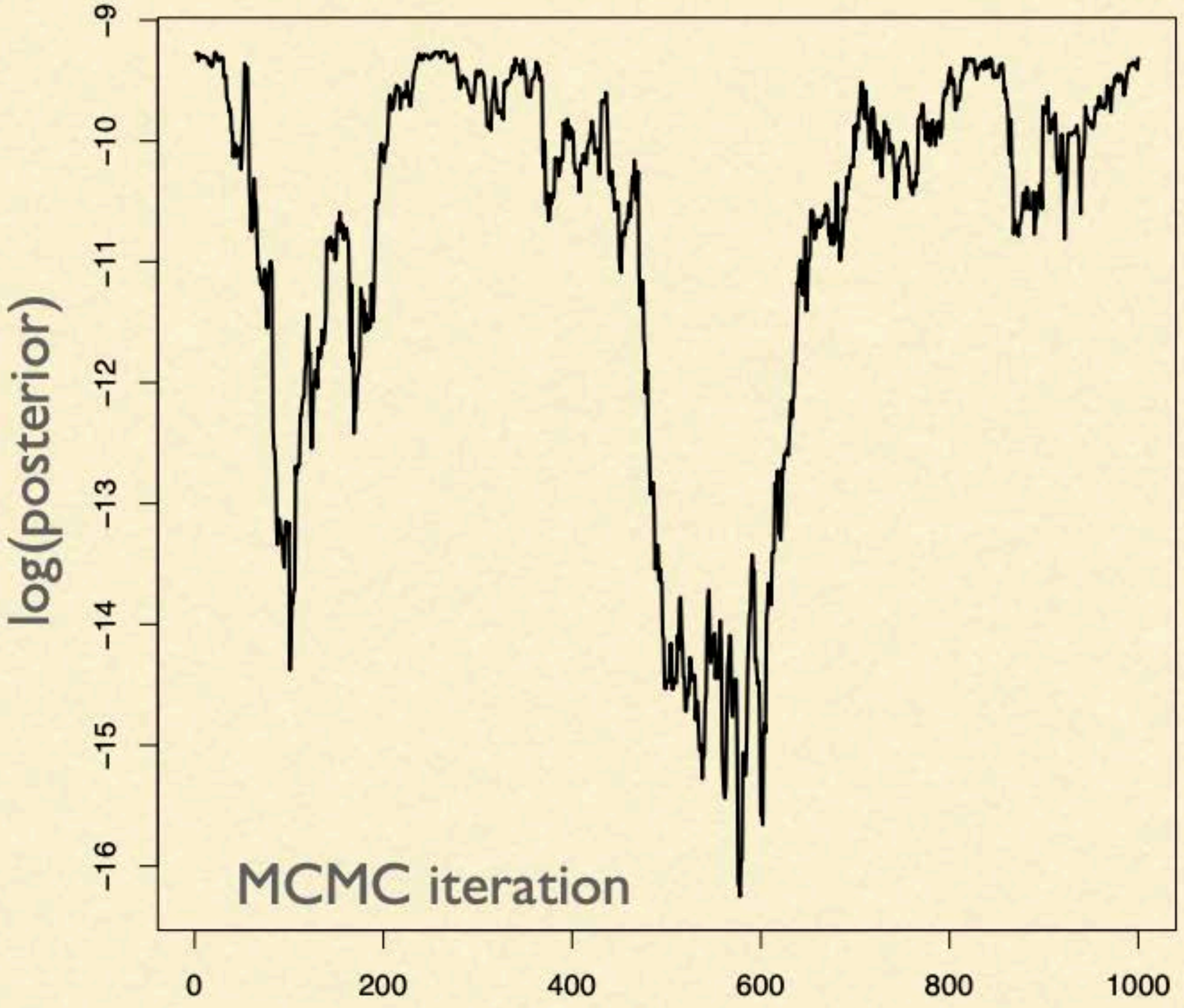
Proposal distributions with smaller variance...

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

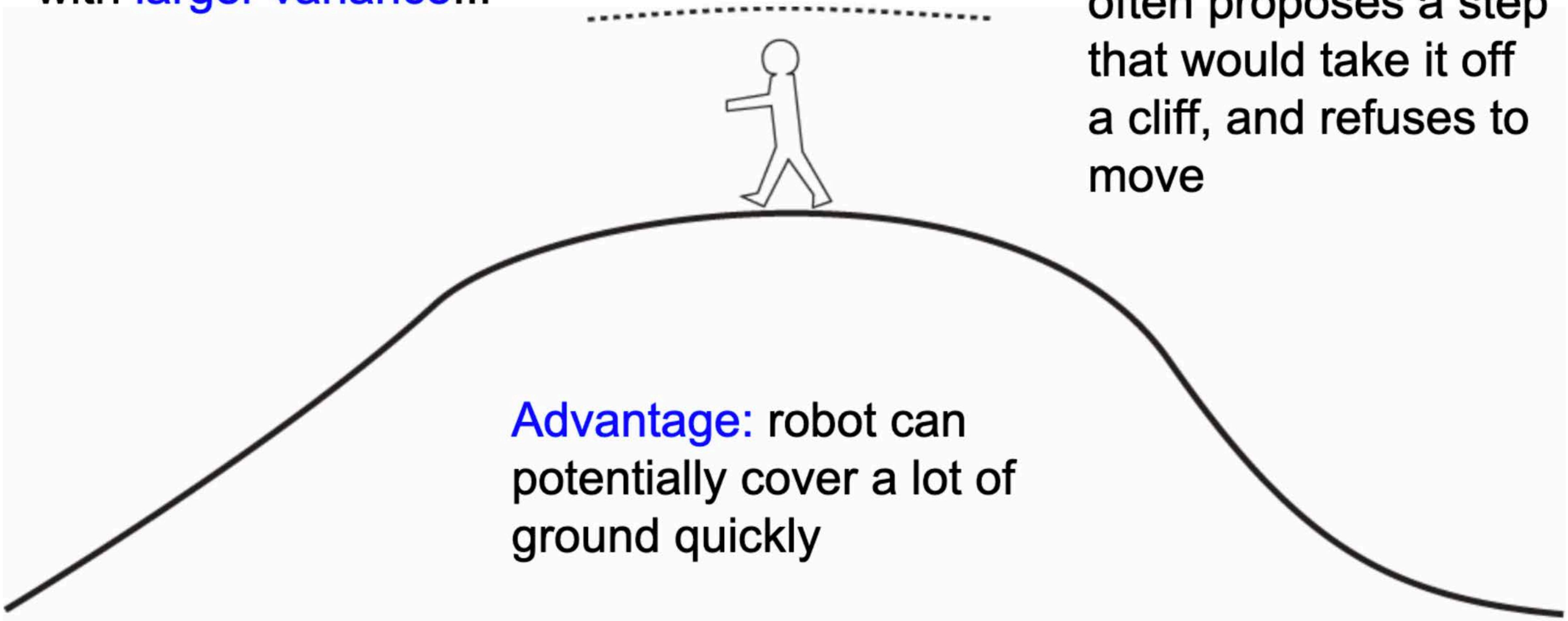




Big waves in trace plot indicate robot is crawling around

Target vs. Proposal Distributions

Proposal distributions with larger variance...

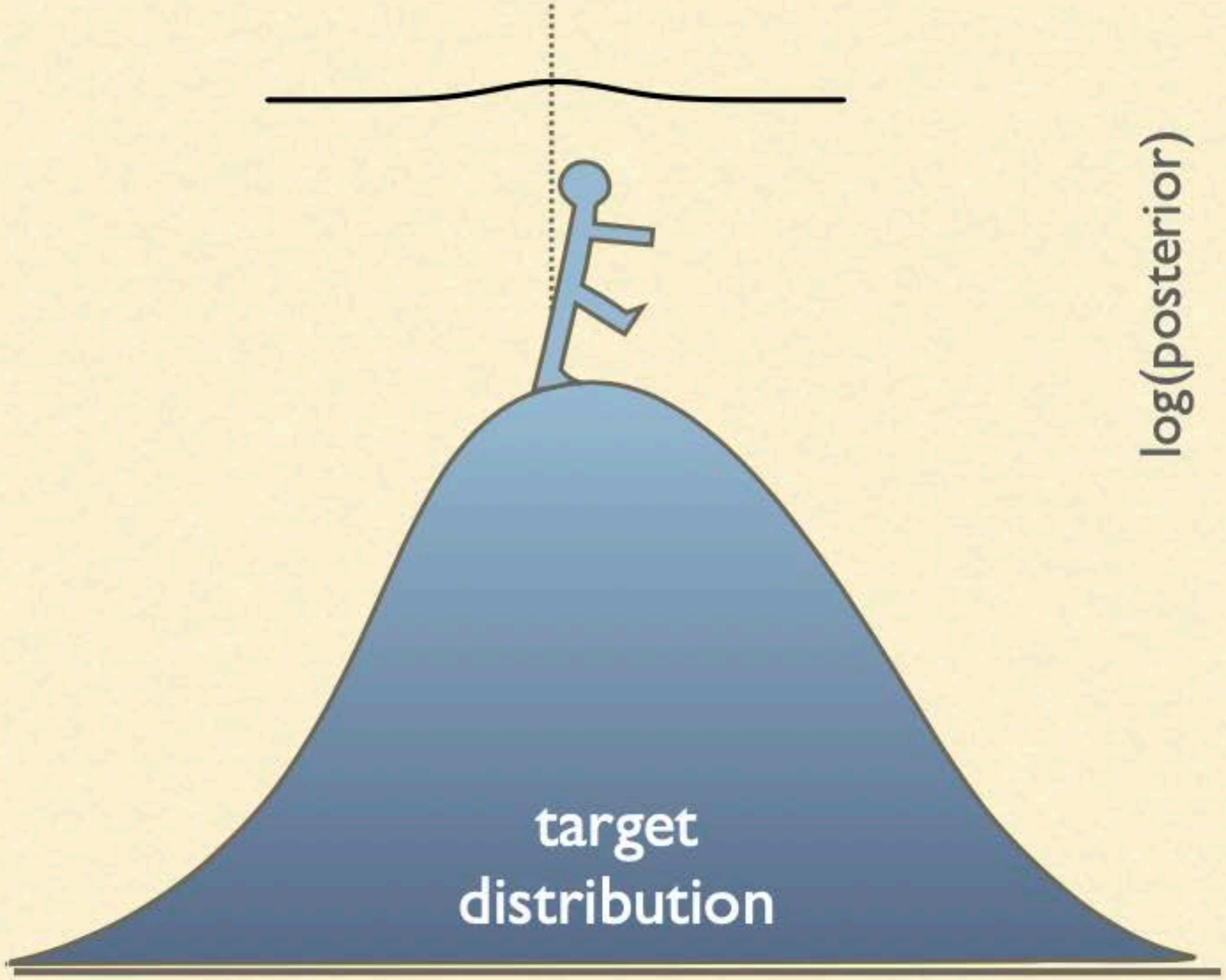


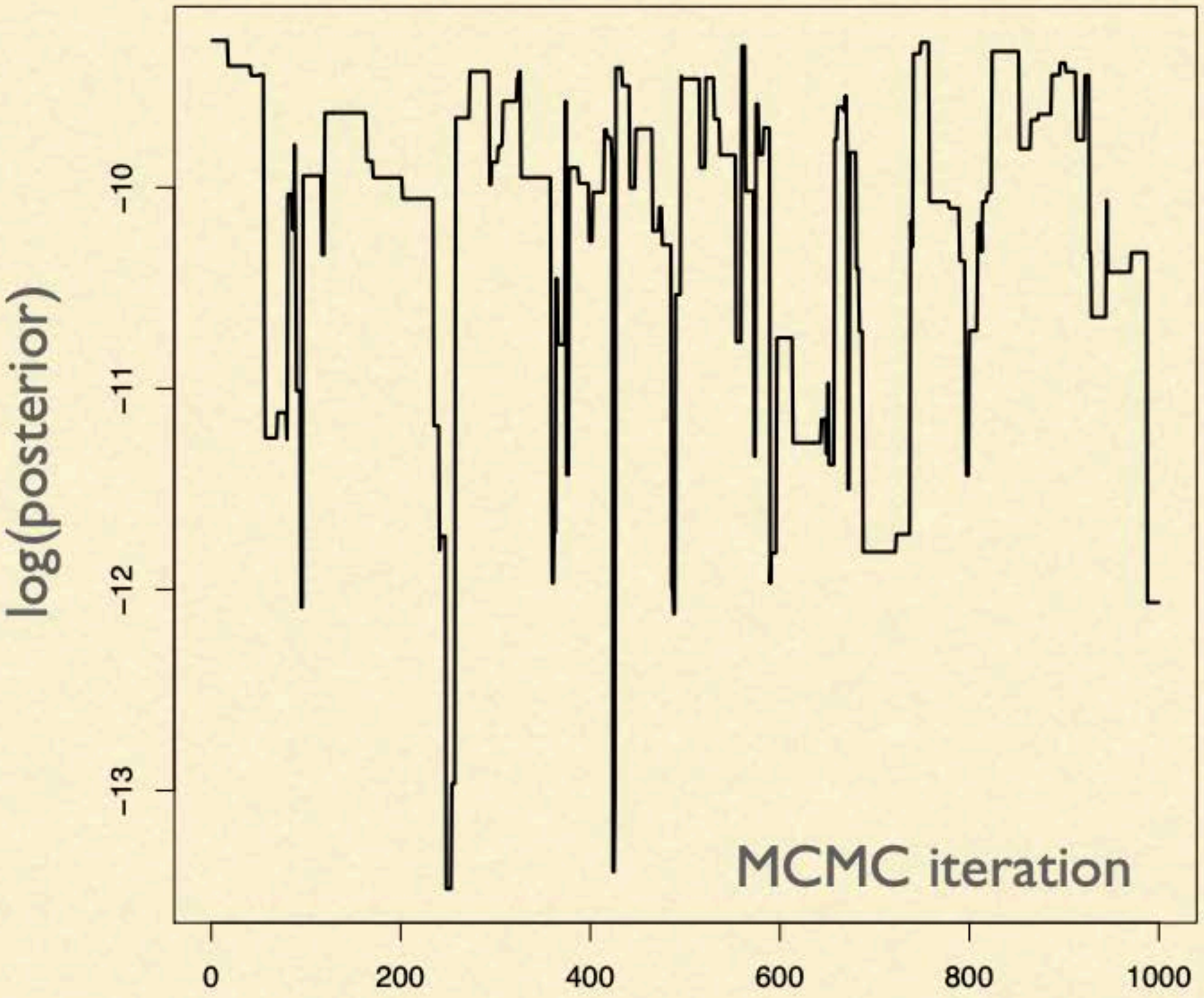
Disadvantage: robot often proposes a step that would take it off a cliff, and refuses to move

Advantage: robot can potentially cover a lot of ground quickly

Target vs. Proposal Distributions

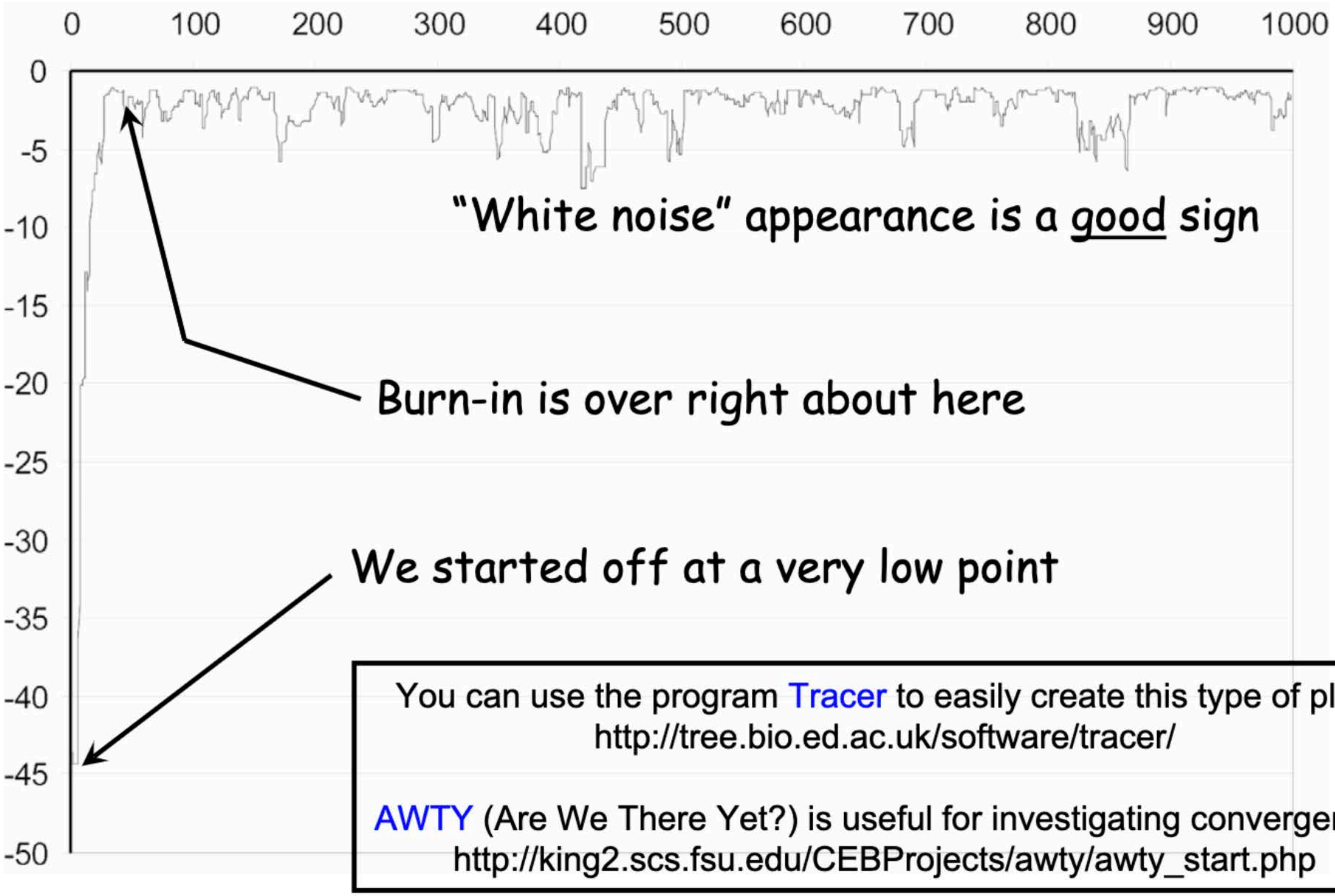
"overly bold" proposal distribution





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

Trace plots



"White noise" appearance is a <u>good</u> sign

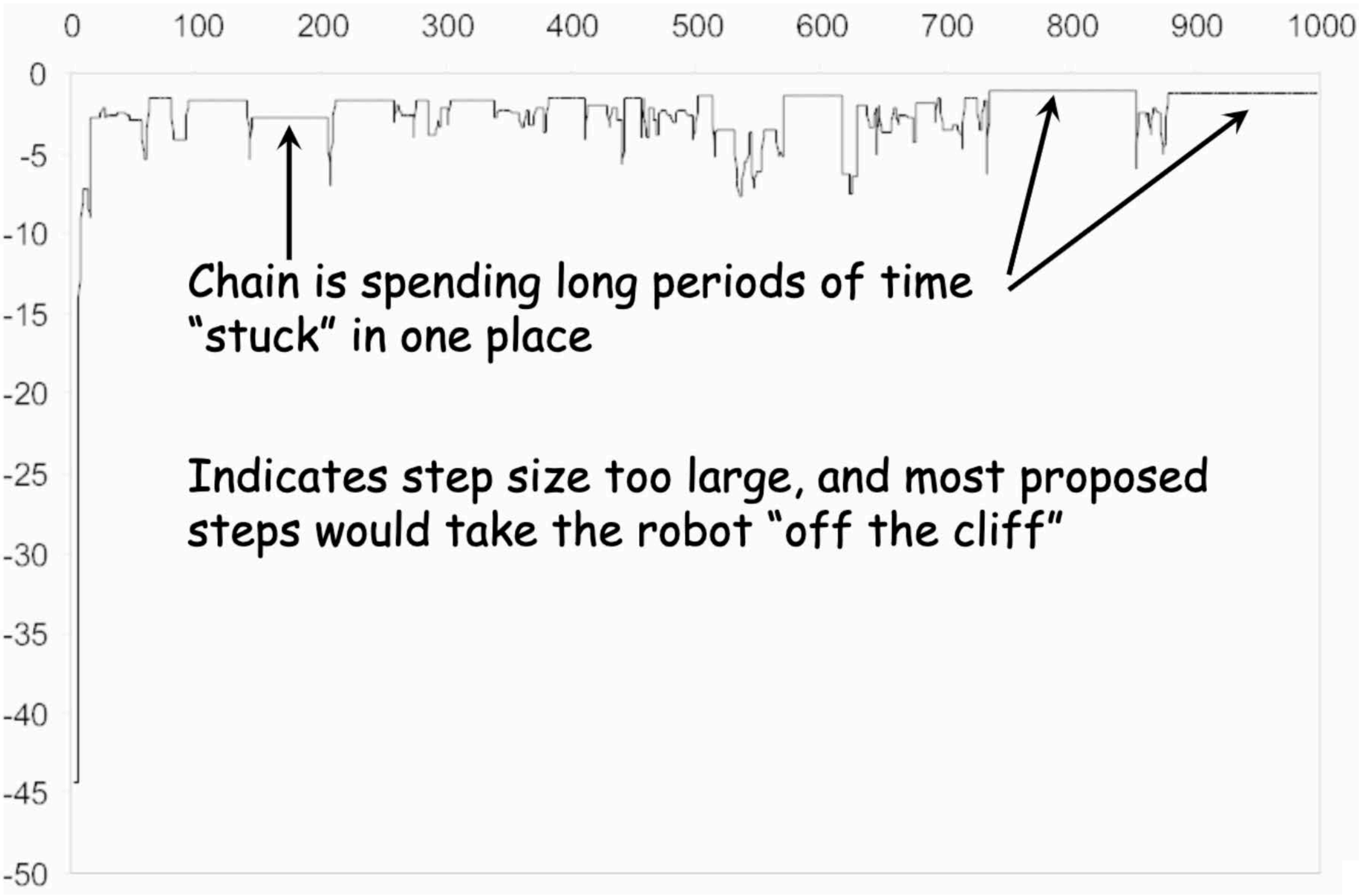
Burn-in is over right about here

We started off at a very low point

You can use the program Tracer to easily create this type of 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

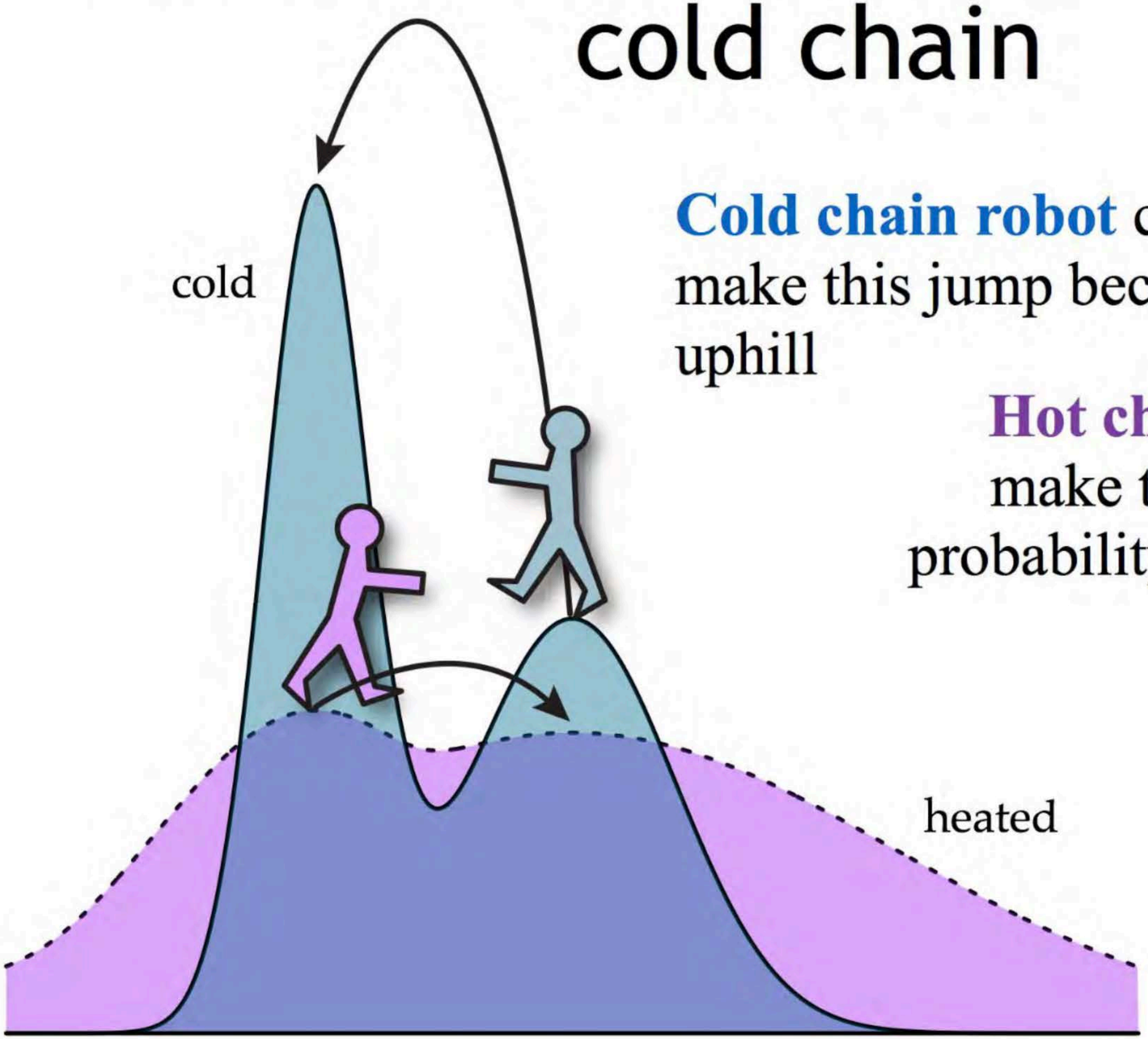
Poor mixing



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

Indicates step size too large, and most proposed steps would take the robot "off the cliff"

Heated chains act as scouts for the cold chain

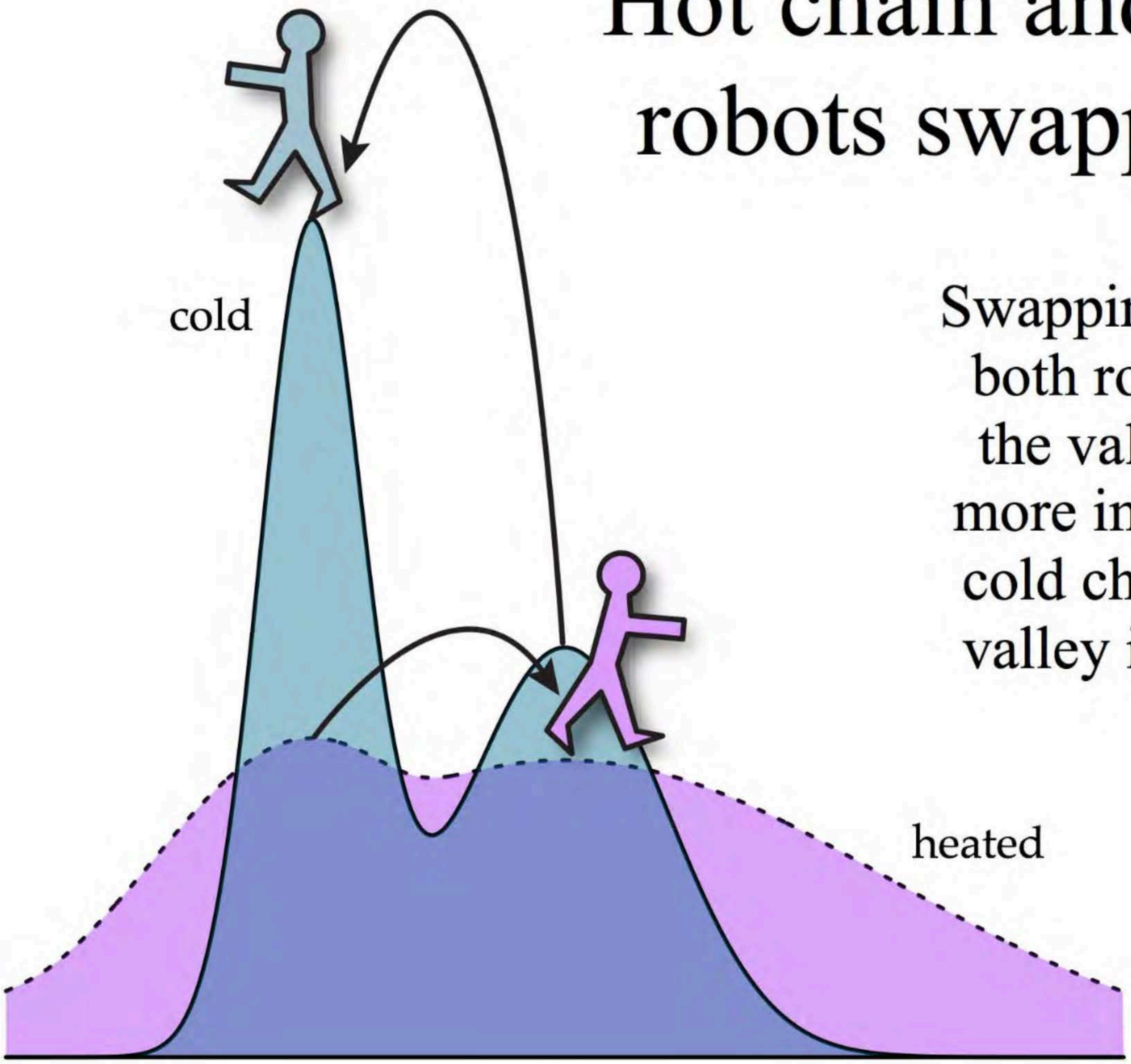


Cold chain robot can easily make this jump because it is uphill

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

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

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

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

“Metropolis algorithm will produce a precise and accurate approximation of the posterior distribution if run long enough”. - Paul Lewis
“People always forget how long of a time infinity really is” - paraphrased from Dave Swofford

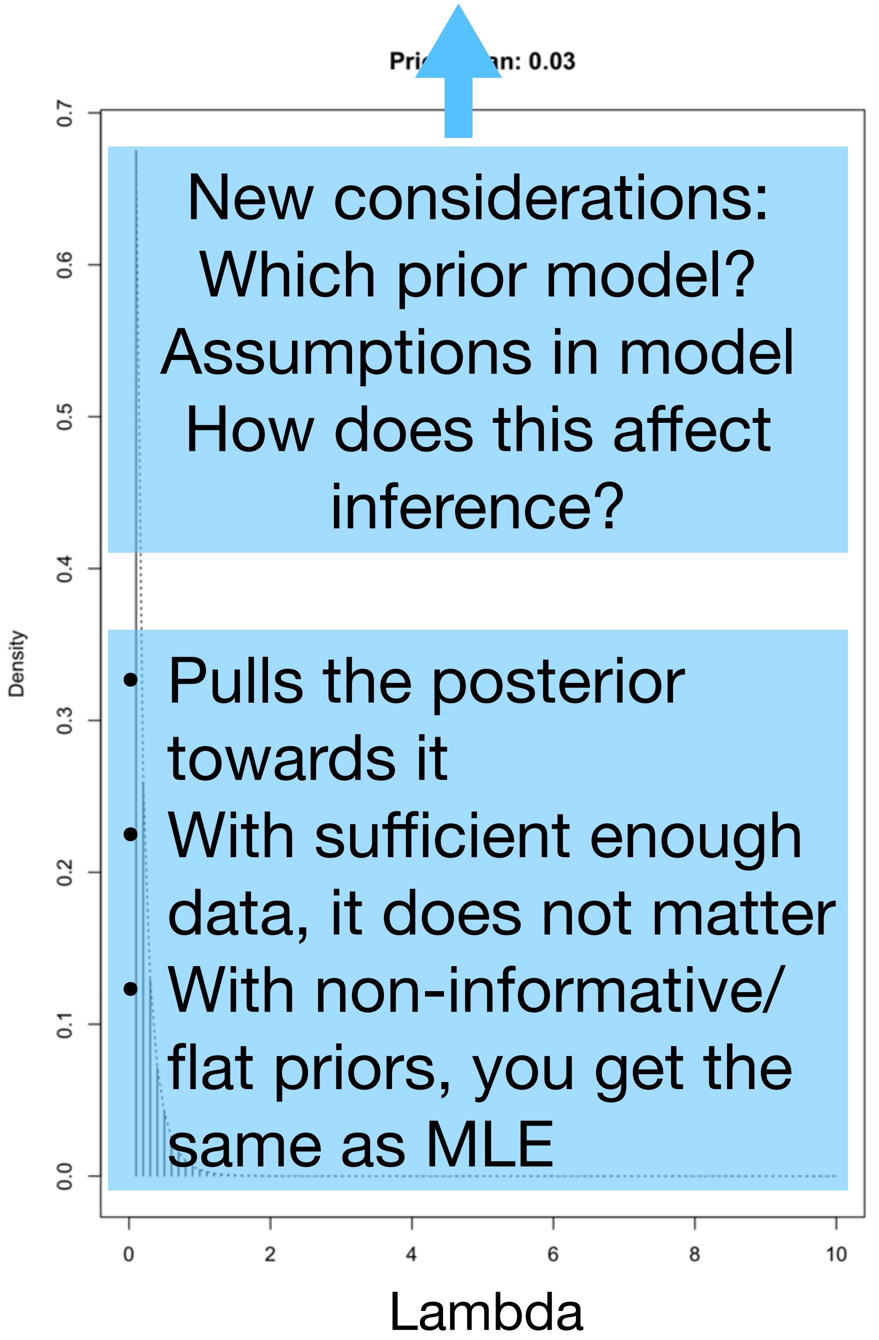
Back to phylogenetics

Bayesian: Likelihood 2.0

- Incorporate prior knowledge
- You get a distribution, not just a point estimate

Your knowledge

Prior



New considerations:
Which prior model?
Assumptions in model
How does this affect inference?

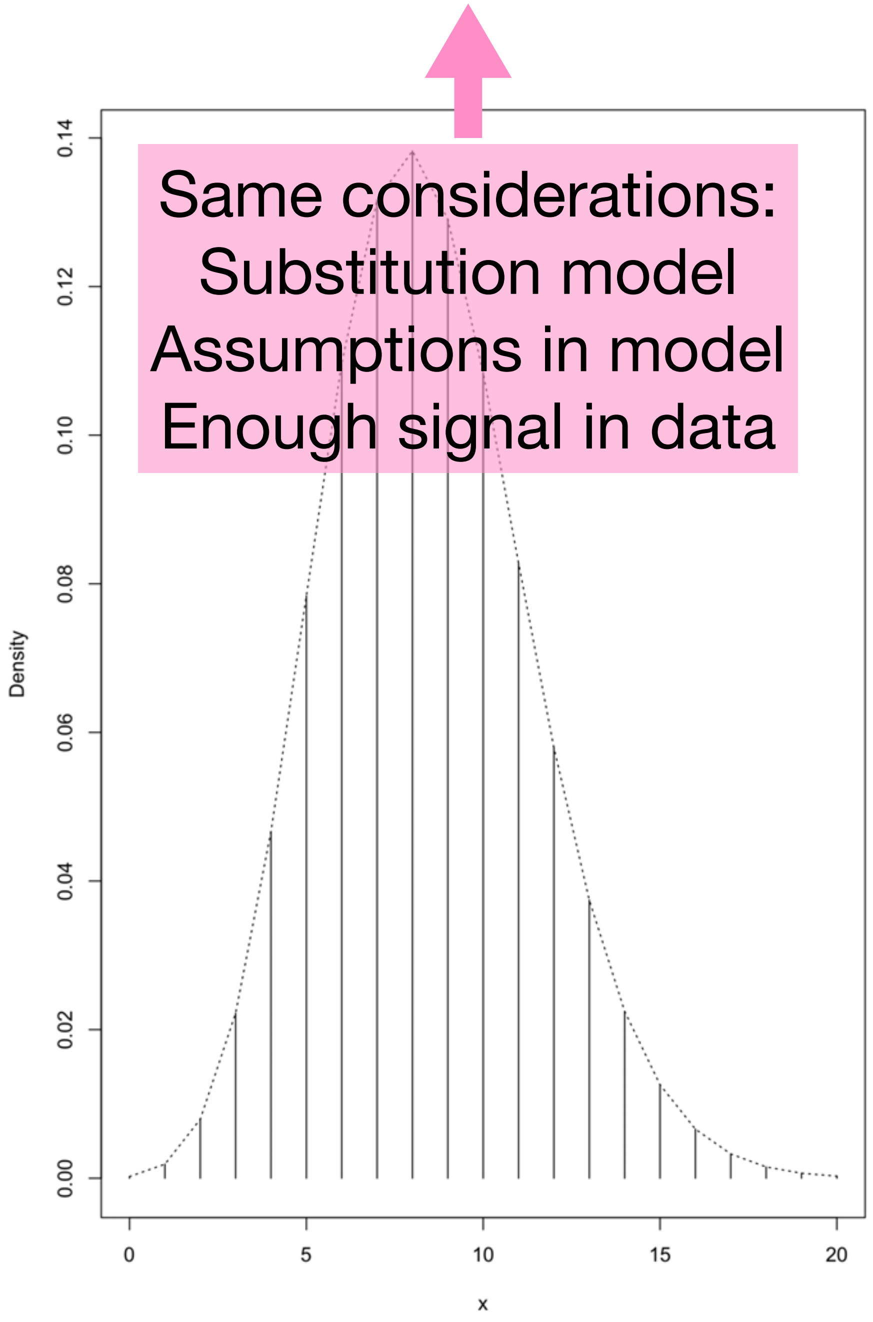
- Pulls the posterior towards it
- With sufficient enough data, it does not matter
- With non-informative/ flat priors, you get the same as MLE

$$\lambda \sim Gamma(\alpha = 0.1, \beta = 3.3)$$

+

Information in the data

Likelihood



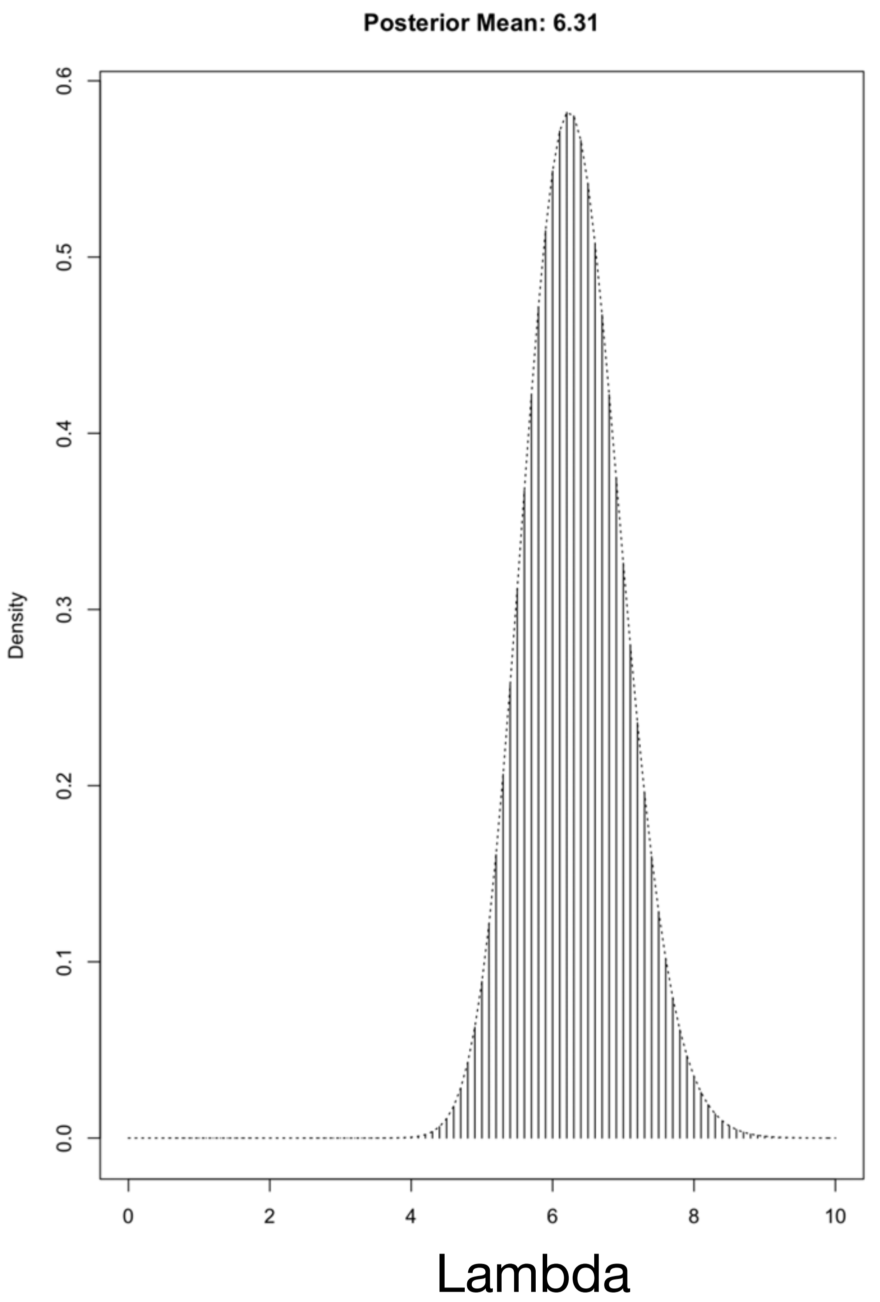
Same considerations:
Substitution model
Assumptions in model
Enough signal in data

$$\hat{\lambda} = 8.4$$

=

Inference on parameter

Posterior



$$\lambda | \mathbf{X} \sim Gamma(\alpha = 0.1 + \sum x_i, \beta = 3.3 + n)$$

Posterior distribution

$$P(\lambda|\mathbf{X}) = \frac{L_{\mathbf{X}}(\lambda)p(\lambda)}{p(\mathbf{X})} = \frac{\overbrace{\prod P(X = x_i|\lambda)}^{\text{Likelihood}}\overbrace{p(\lambda)}^{\text{Prior}}}{\underbrace{p(\mathbf{X})}_{\text{Marginal}}} \longleftarrow \text{Many times, intractable}$$

$$\Rightarrow P(\lambda|\mathbf{X}) \propto L_{\mathbf{X}}(\lambda)p(\lambda)$$

MCMC: A way to approximate intractable posterior distributions

$$P(\lambda|\mathbf{X}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{1}{\lambda\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log\lambda - \mu)^2}{2\sigma^2}\right)$$

Initialization: Start at a random λ_0

Loop: For *i*, propose a new lambda

$$\lambda^* \sim Uniform(\lambda_{i-1} - w/2, \lambda_{i-1} + w/2)$$

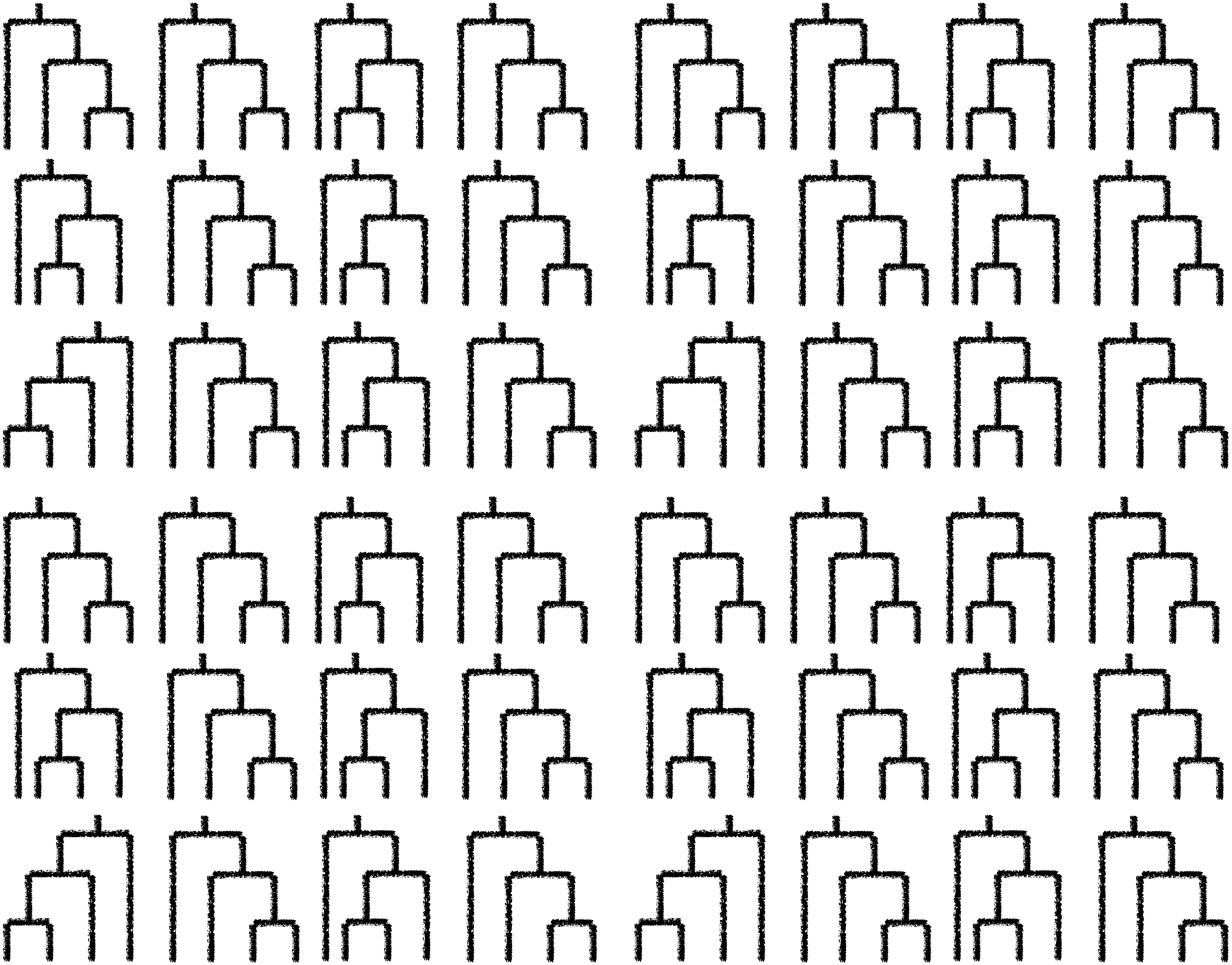
- If $P(\lambda^*|\mathbf{X}) > P(\lambda_{i-1}|\mathbf{X}) \Rightarrow \lambda_i = \lambda^*$
- Else, accept $\lambda_i = \lambda^*$ with probability $\alpha = \dfrac{P(\lambda^*|\mathbf{X})}{P(\lambda_{i-1}|\mathbf{X})}$
- Otherwise, $\lambda_i = \lambda_{i-1}$

- **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.

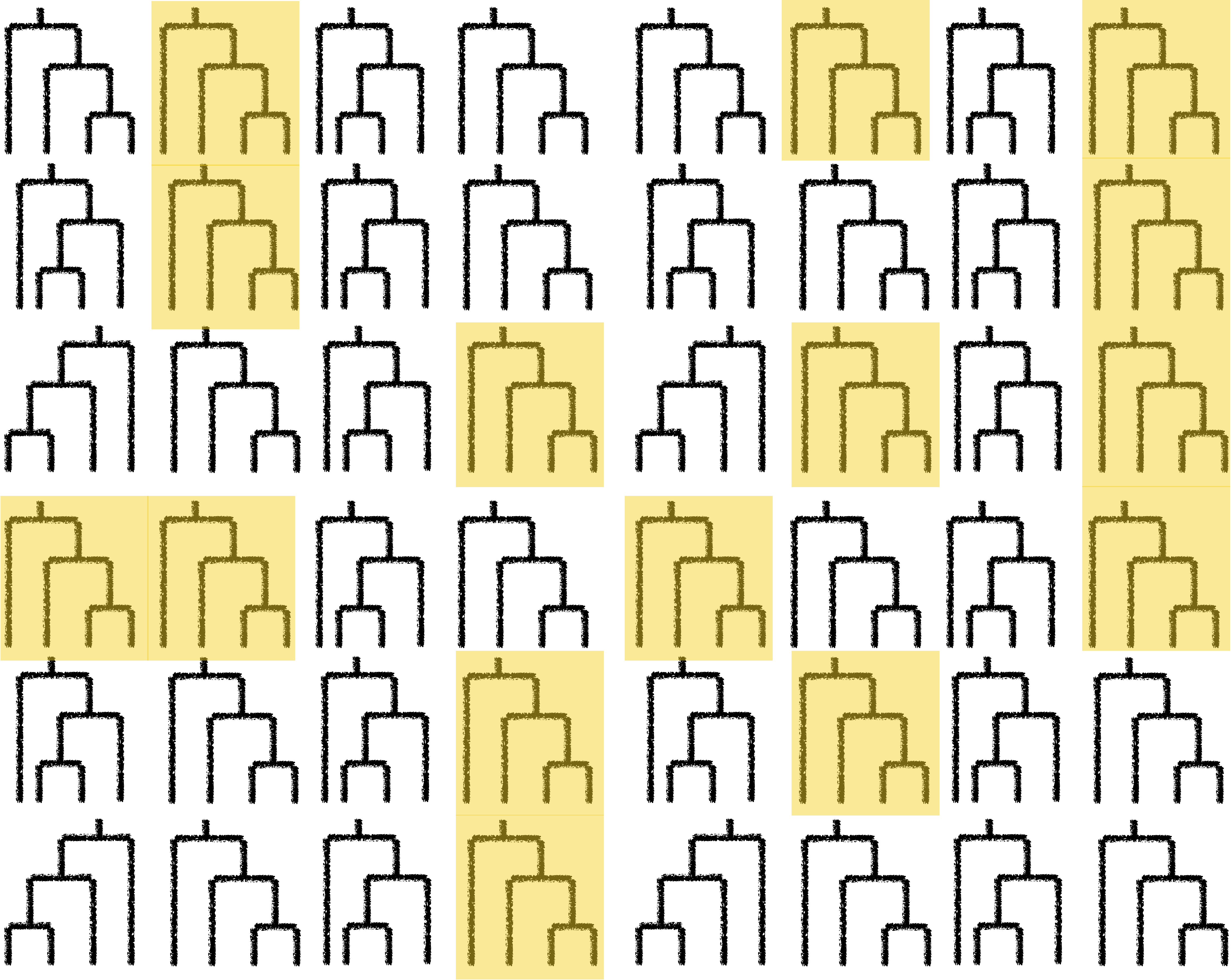
Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

After MCMC…

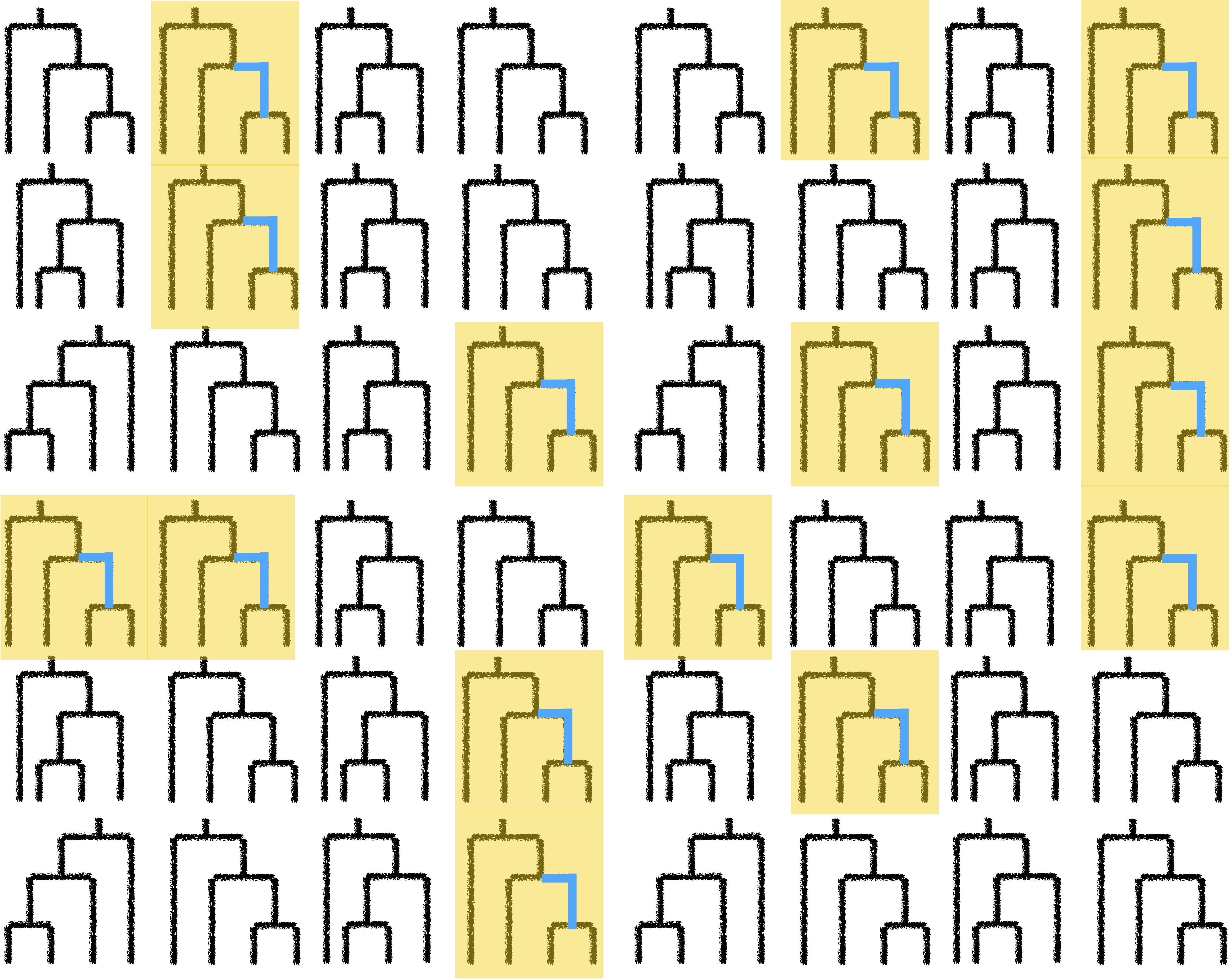
After MCMC…

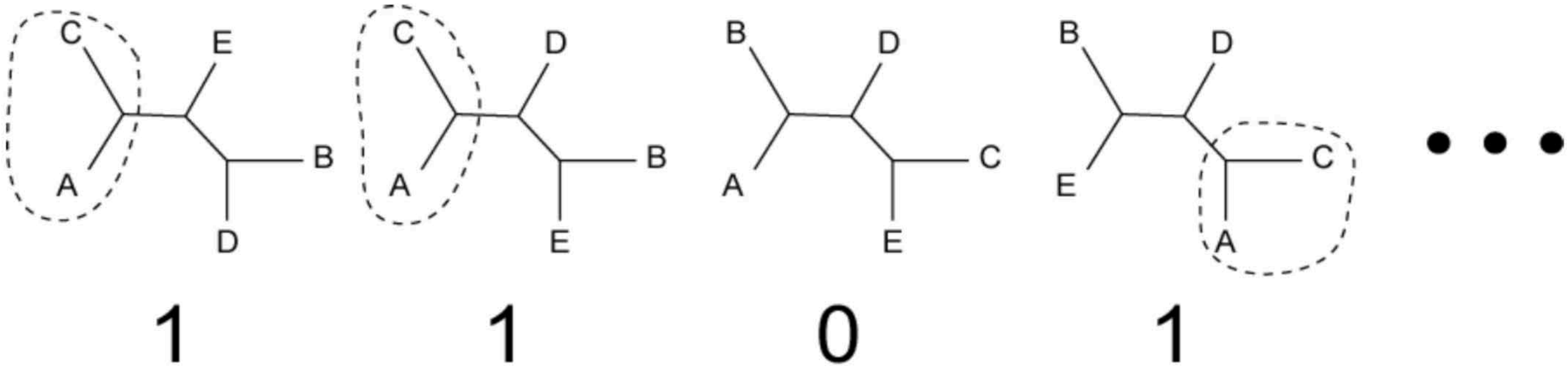


After MCMC...



After MCMC...

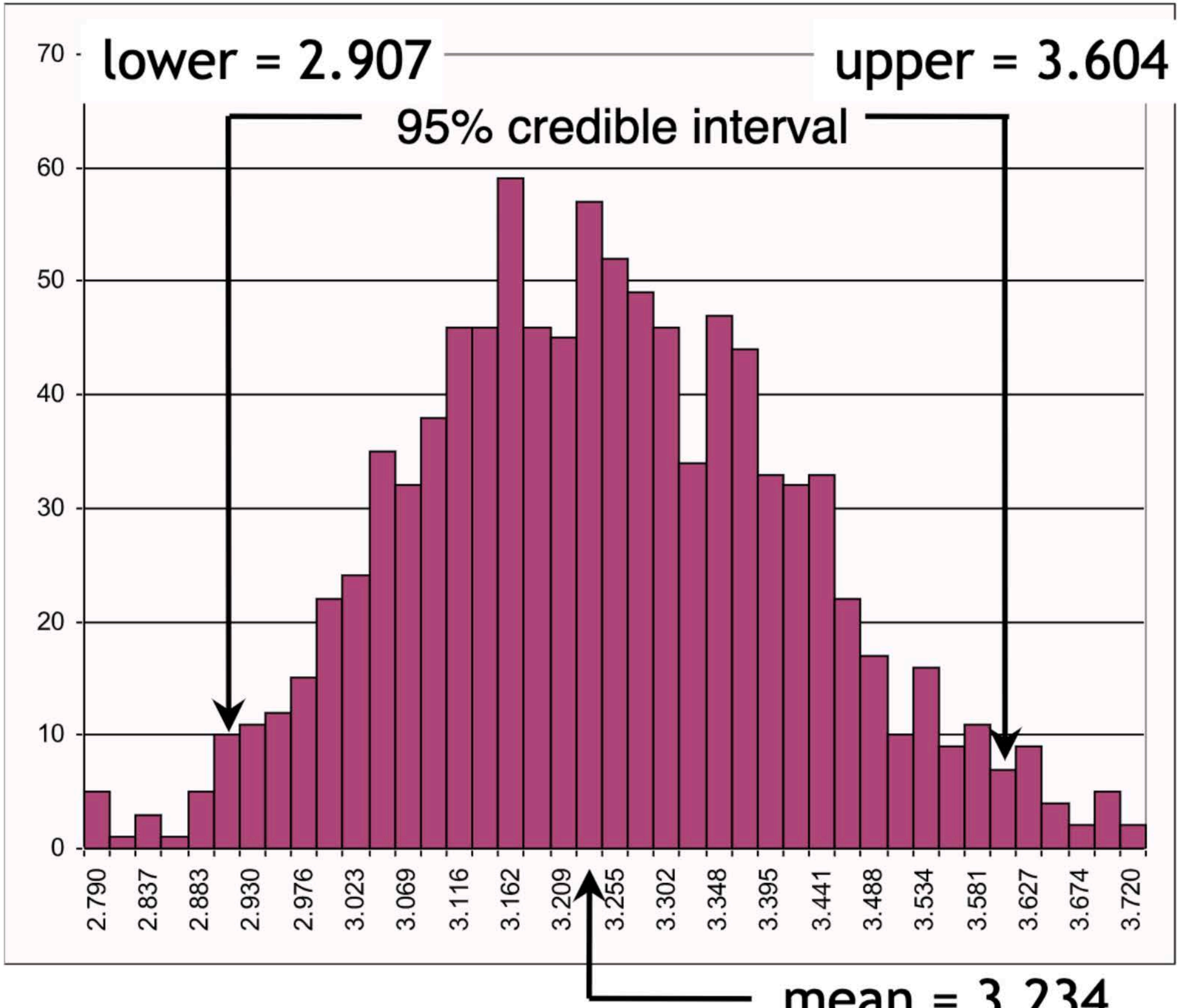




The posterior probability of the split AC|BDE may be approximated by the fraction of trees sampled from the posterior that contain that split.

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Marginal Posterior Distribution of κ



lower = 2.907

upper = 3.604

95% credible interval

mean = 3.234

Histogram created from a sample of 1000 kappa values.

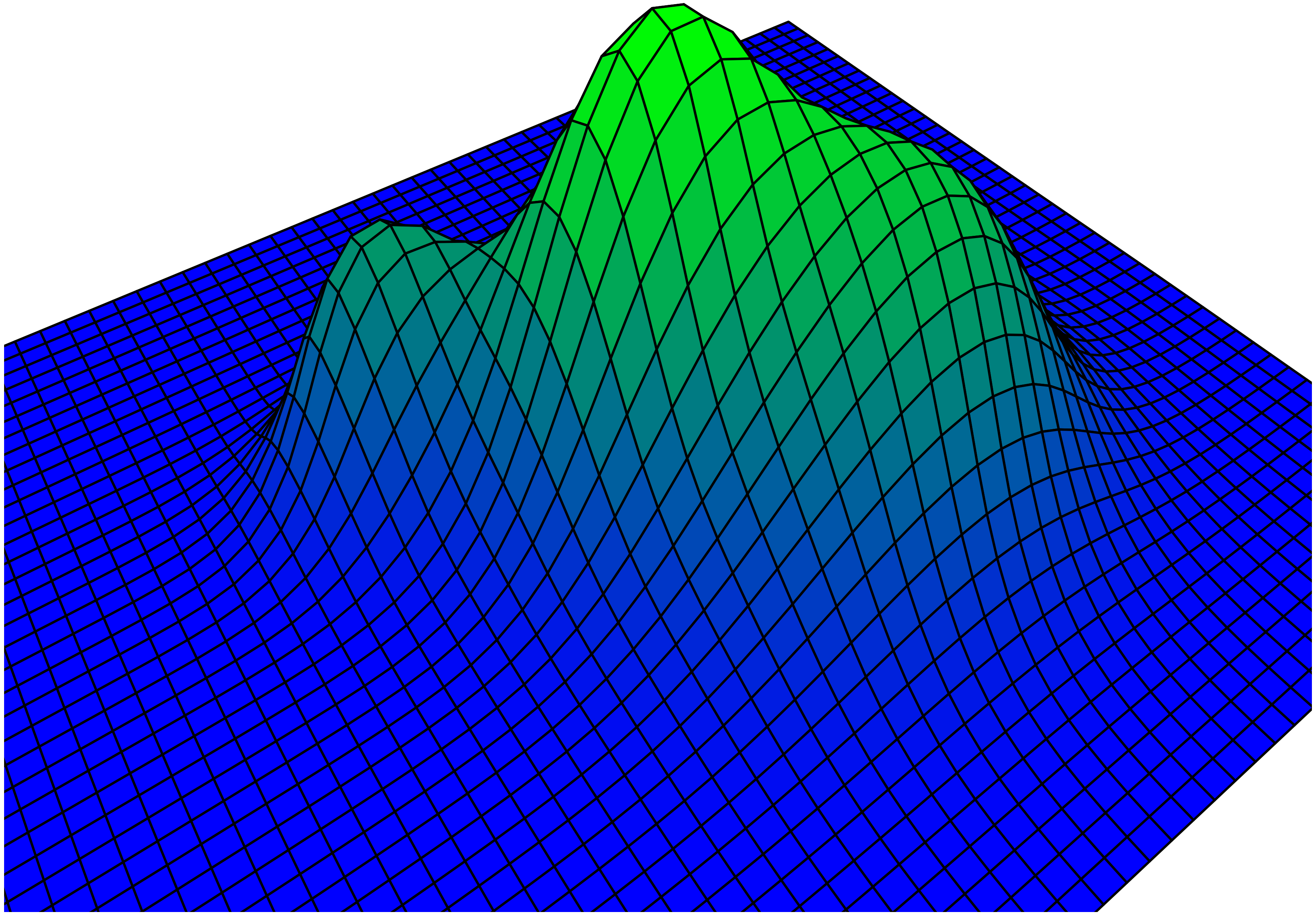
Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Data from Lewis, L., and Flechtner, V. 2002. Taxon 51: 443-451.

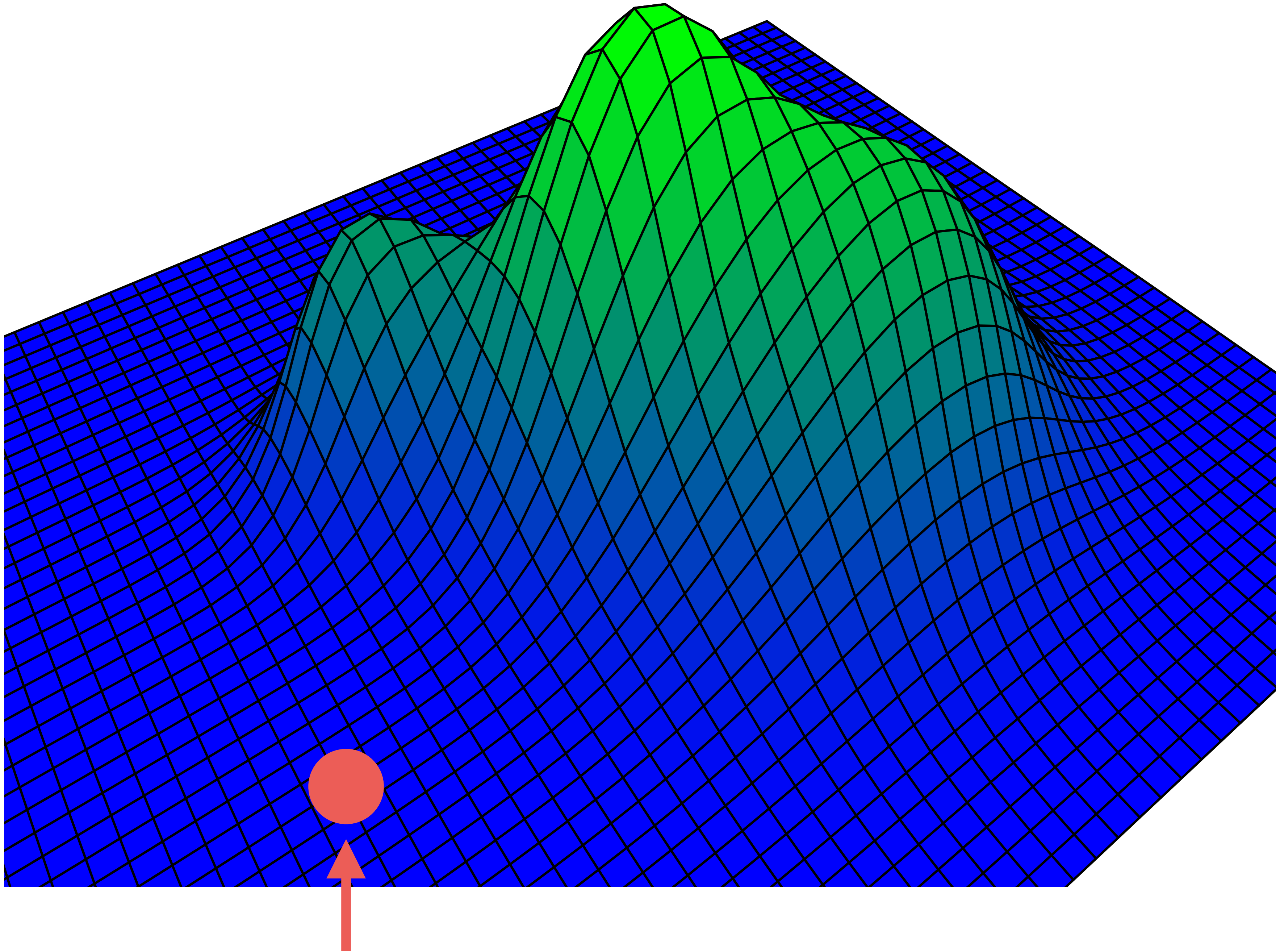
Why is MCMC so slow?

Why is MCMC so slow? Traverse tree space

Why is MCMC so slow? Traverse tree space

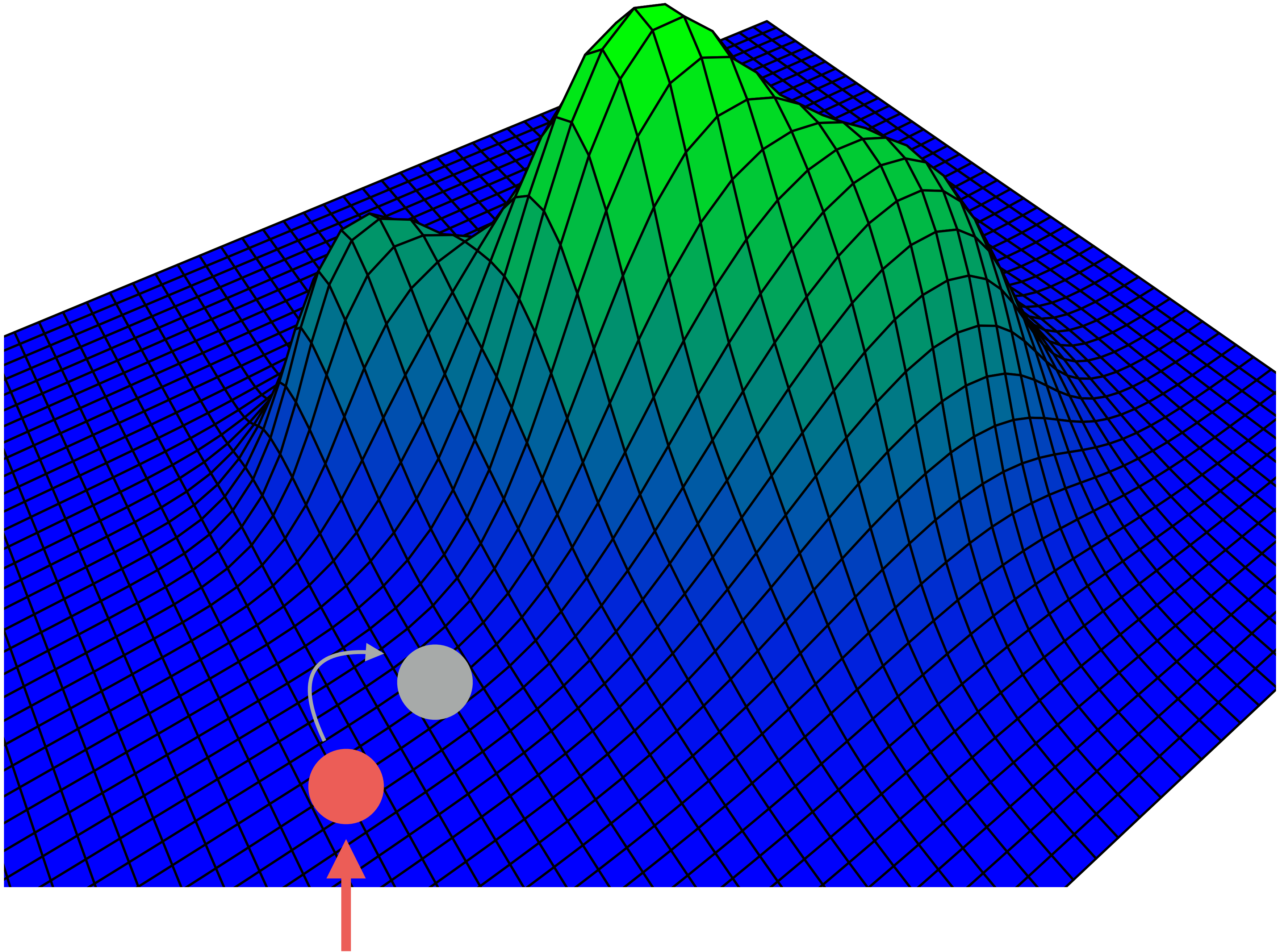


Why is MCMC so slow? Traverse tree space



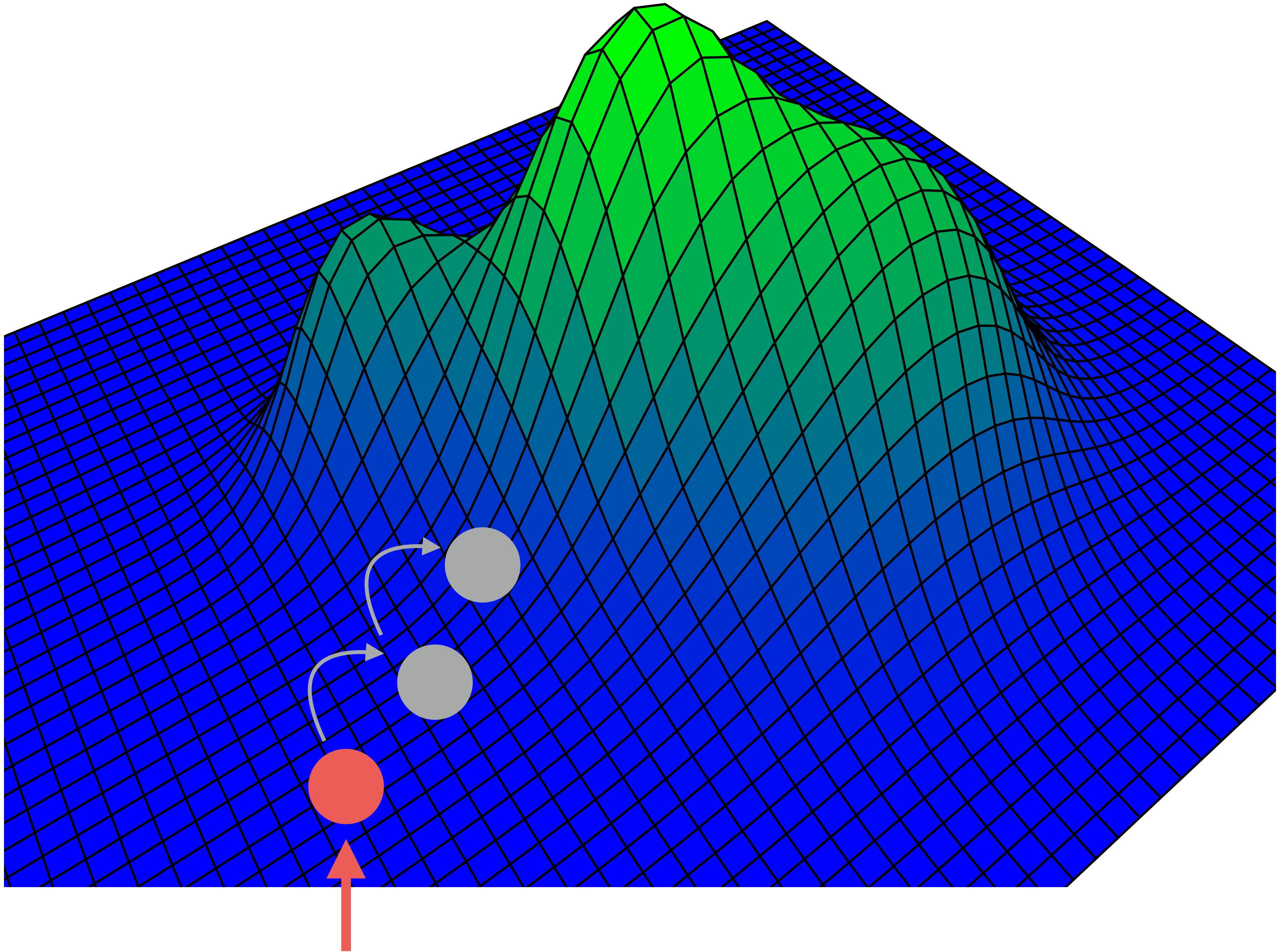
Starting tree

Why is MCMC so slow? Traverse tree space



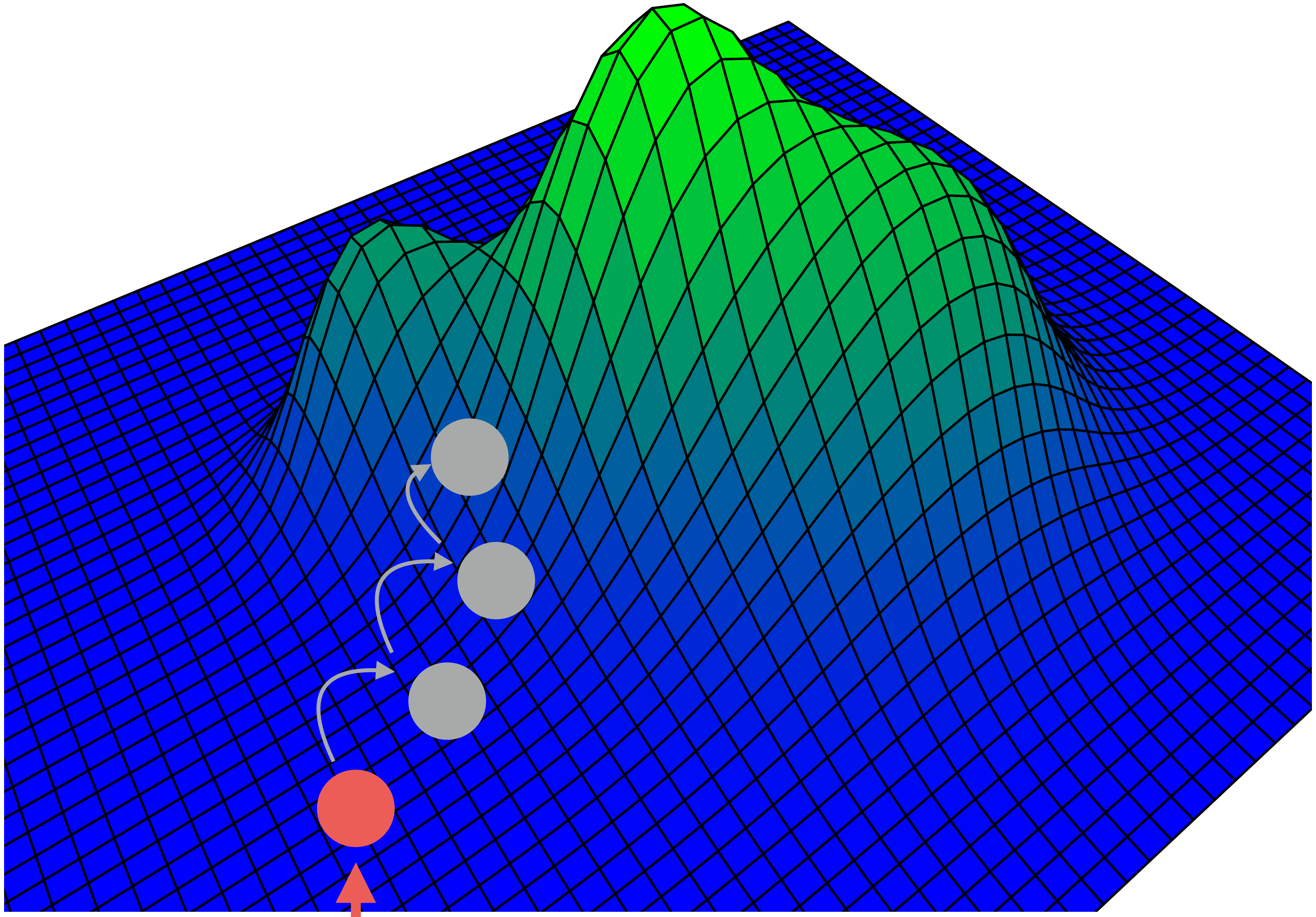
Starting tree

Why is MCMC so slow? Traverse tree space



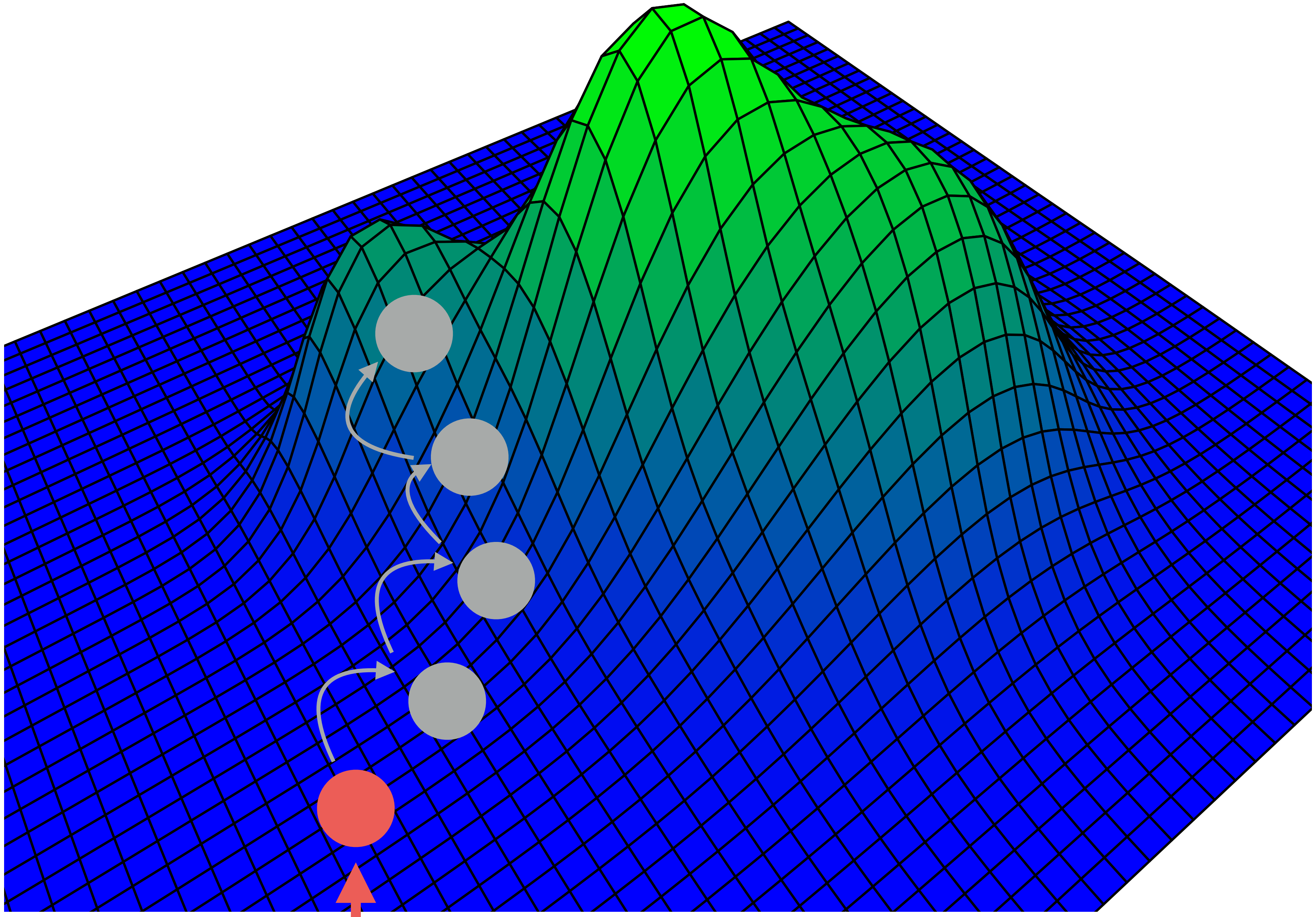
Starting tree

Why is MCMC so slow? Traverse tree space



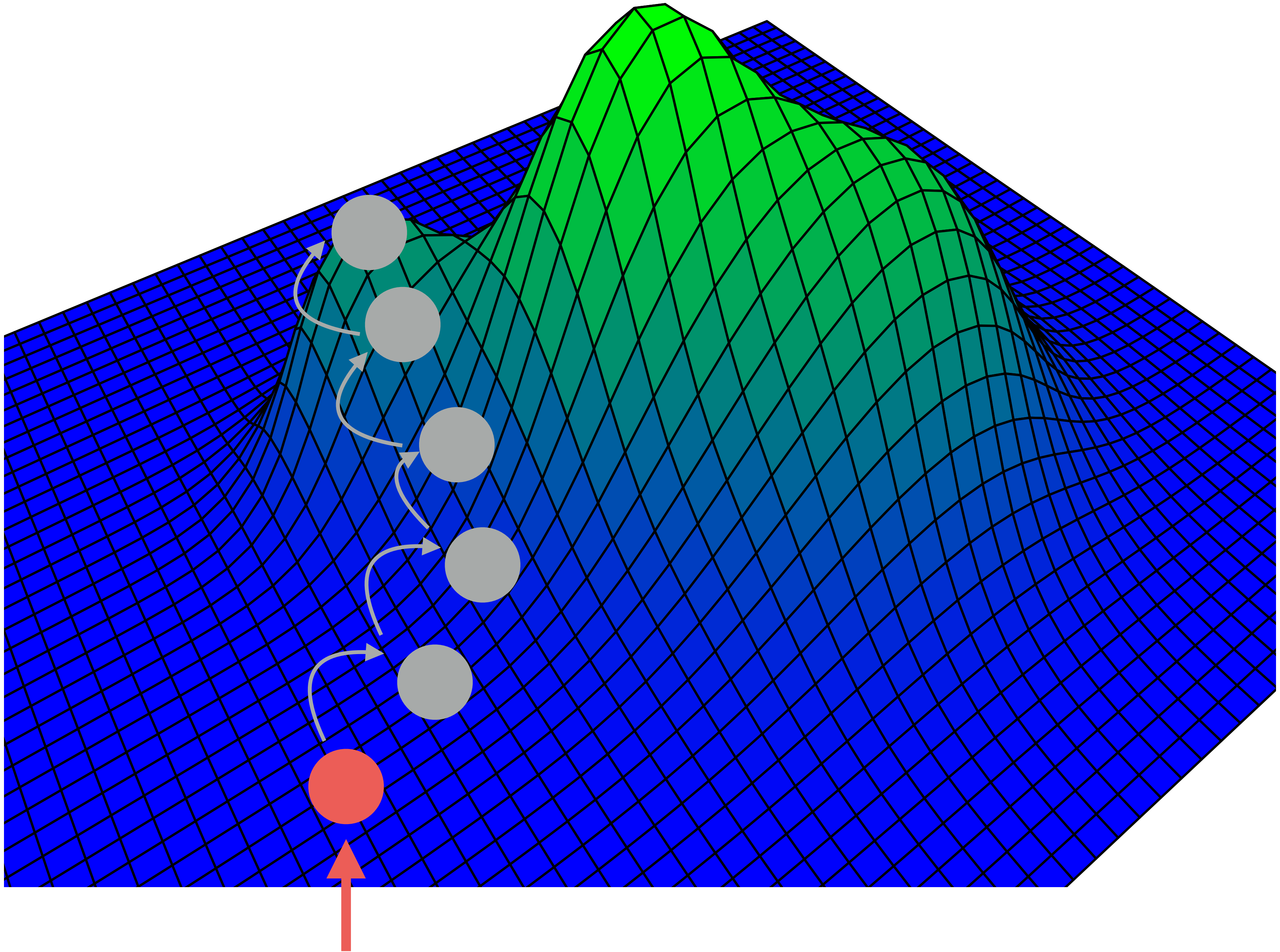
Starting tree

Why is MCMC so slow? Traverse tree space



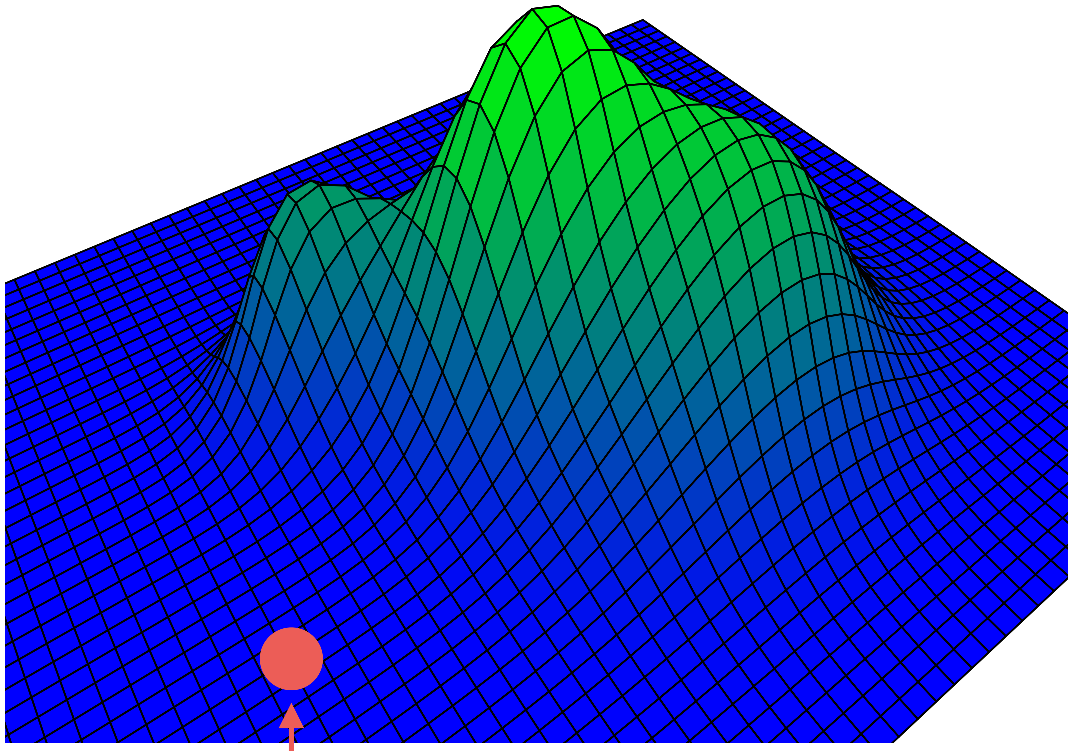
Starting tree

Why is MCMC so slow? Traverse tree space

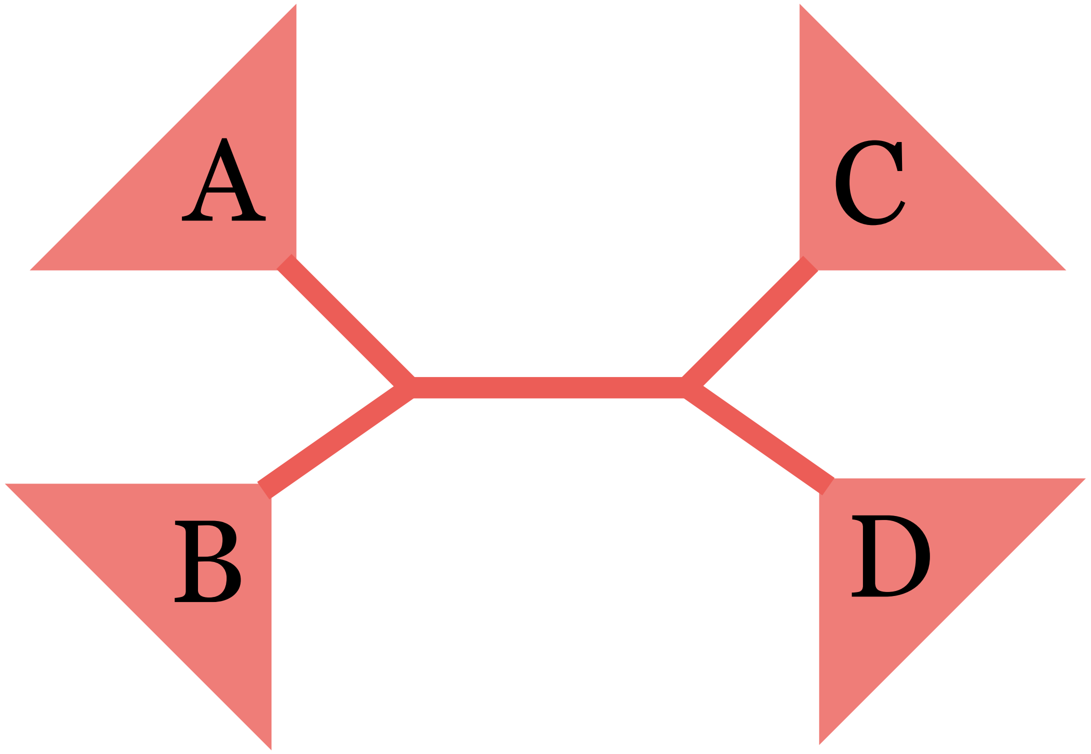


Starting tree

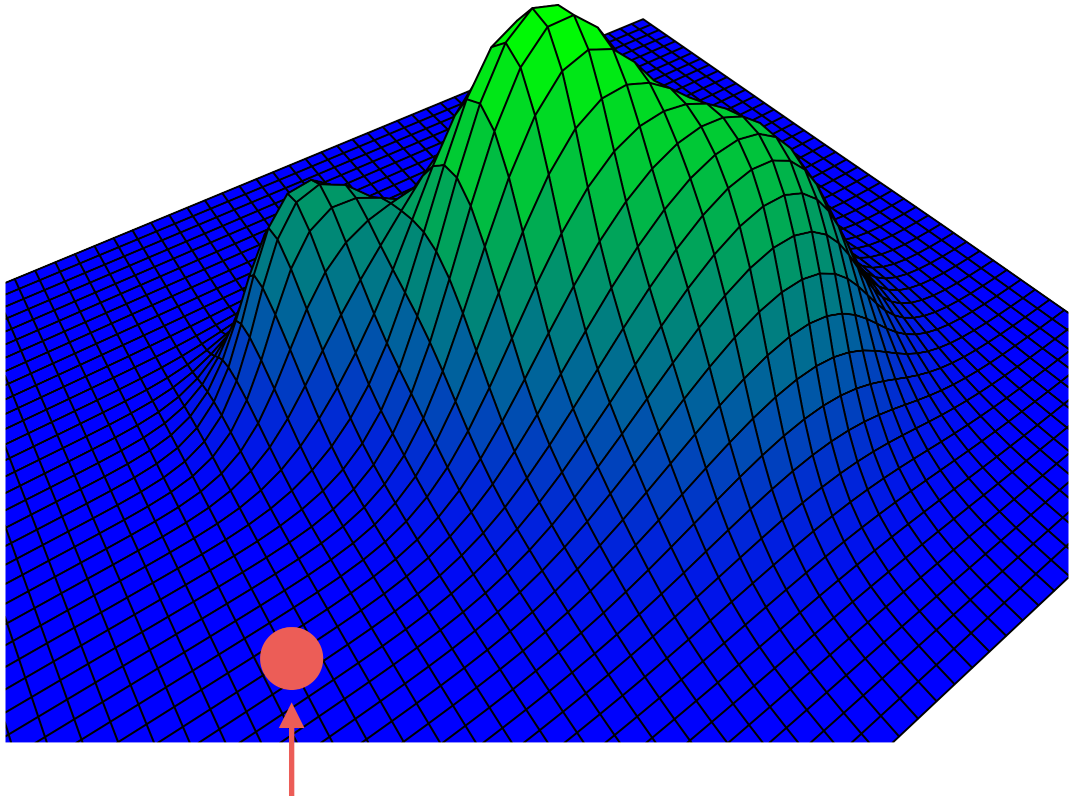
Why is MCMC so slow? Traverse tree space



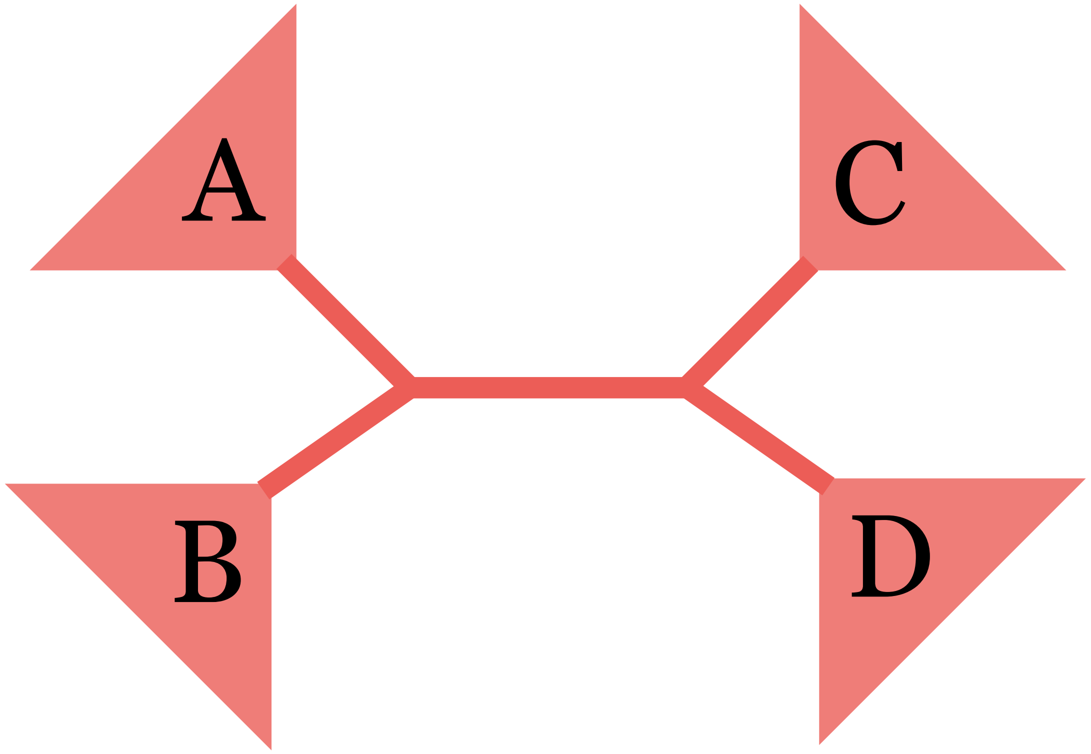
Starting tree



Why is MCMC so slow? Traverse tree space



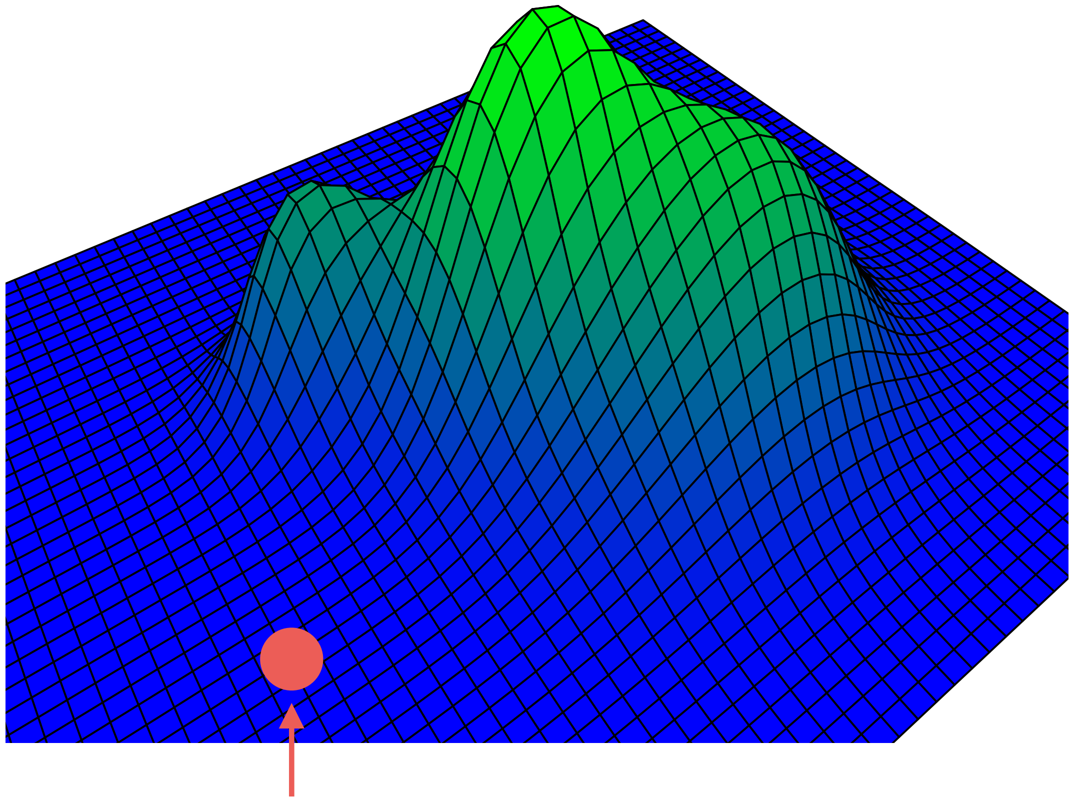
Starting tree

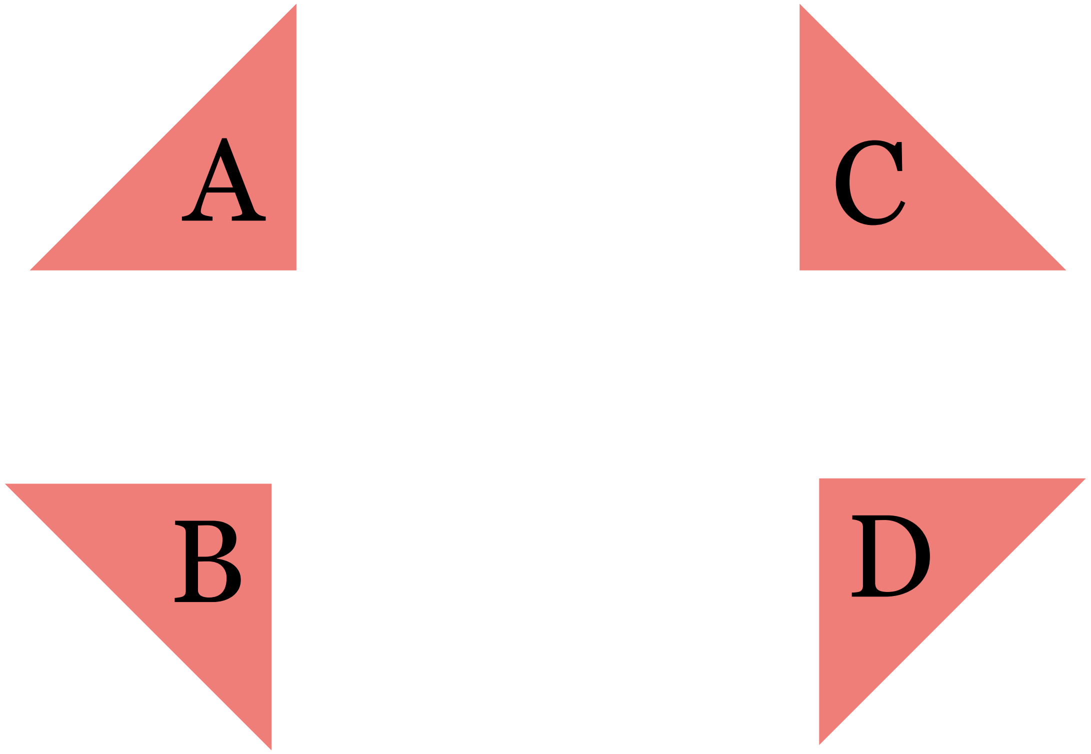


Nearest Neighbor
Interchange

Why is MCMC so slow?

Traverse tree space

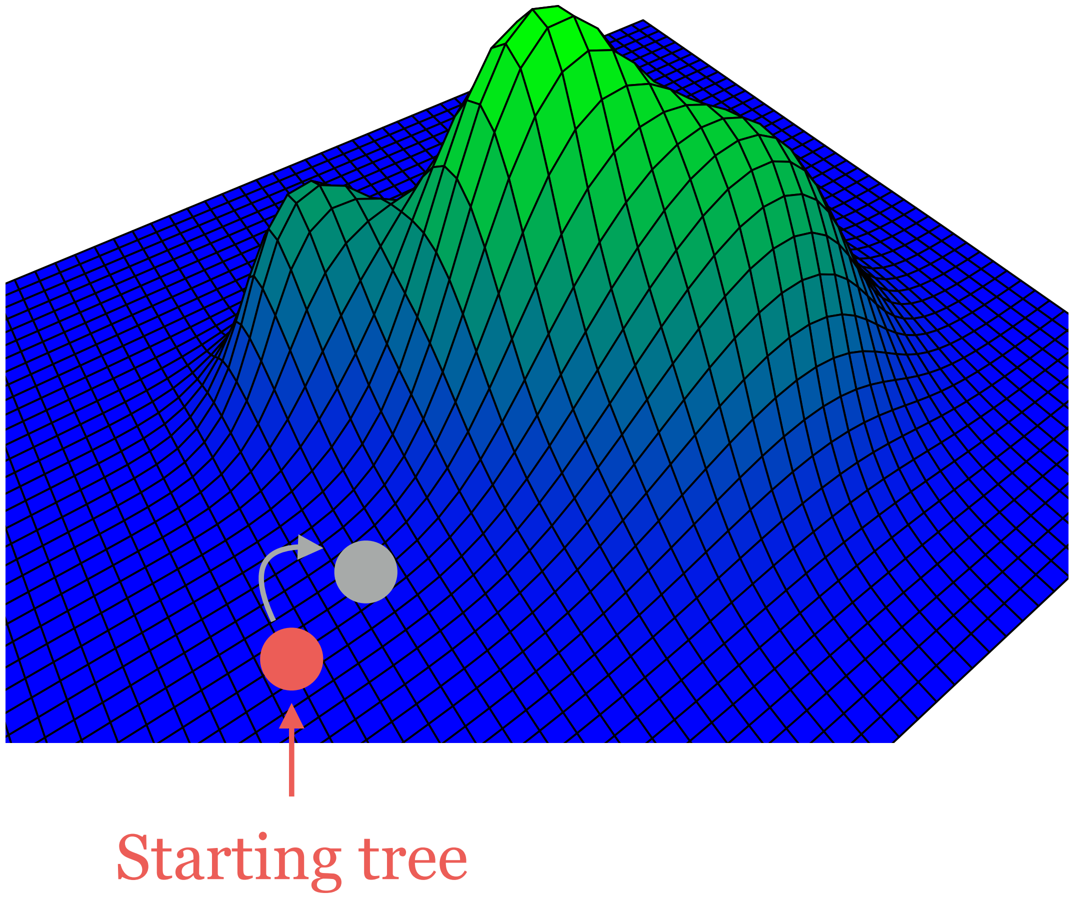

Starting tree

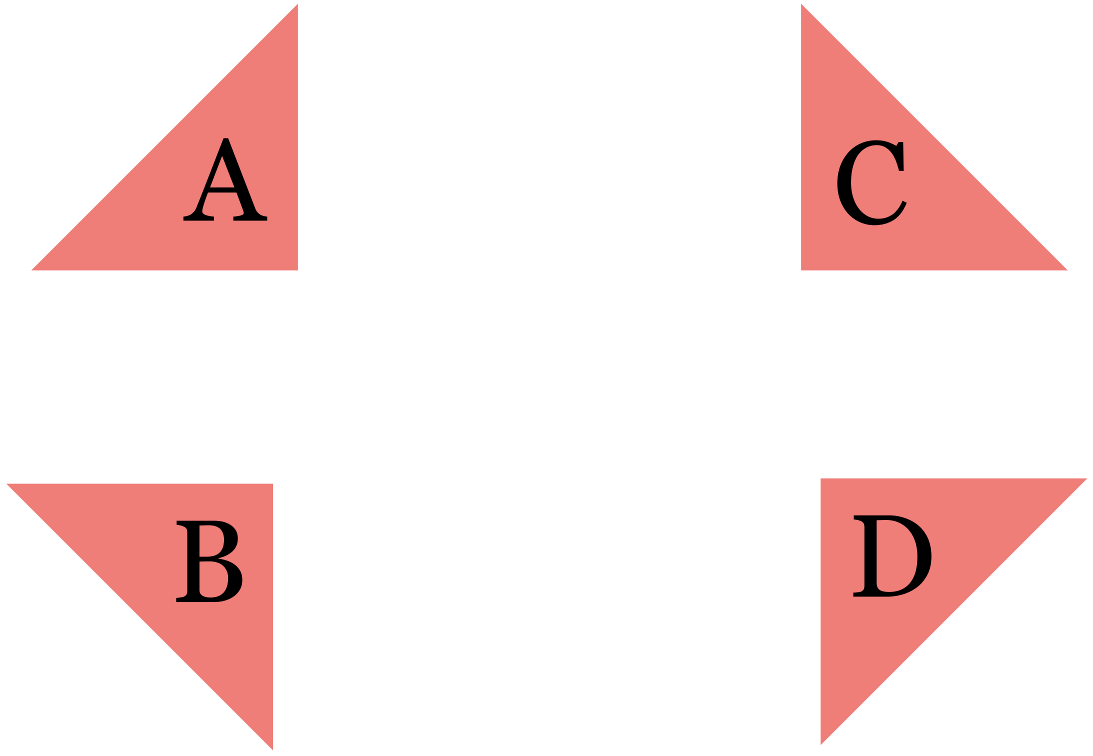


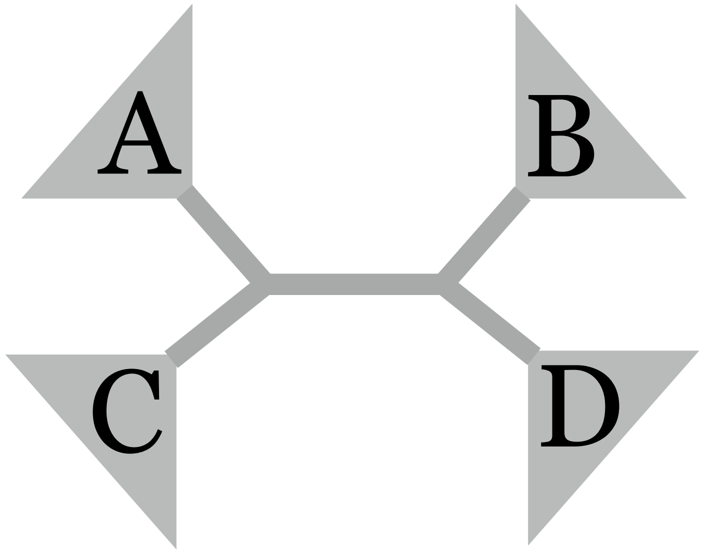
Nearest Neighbor Interchange

Why is MCMC so slow?

Traverse tree space

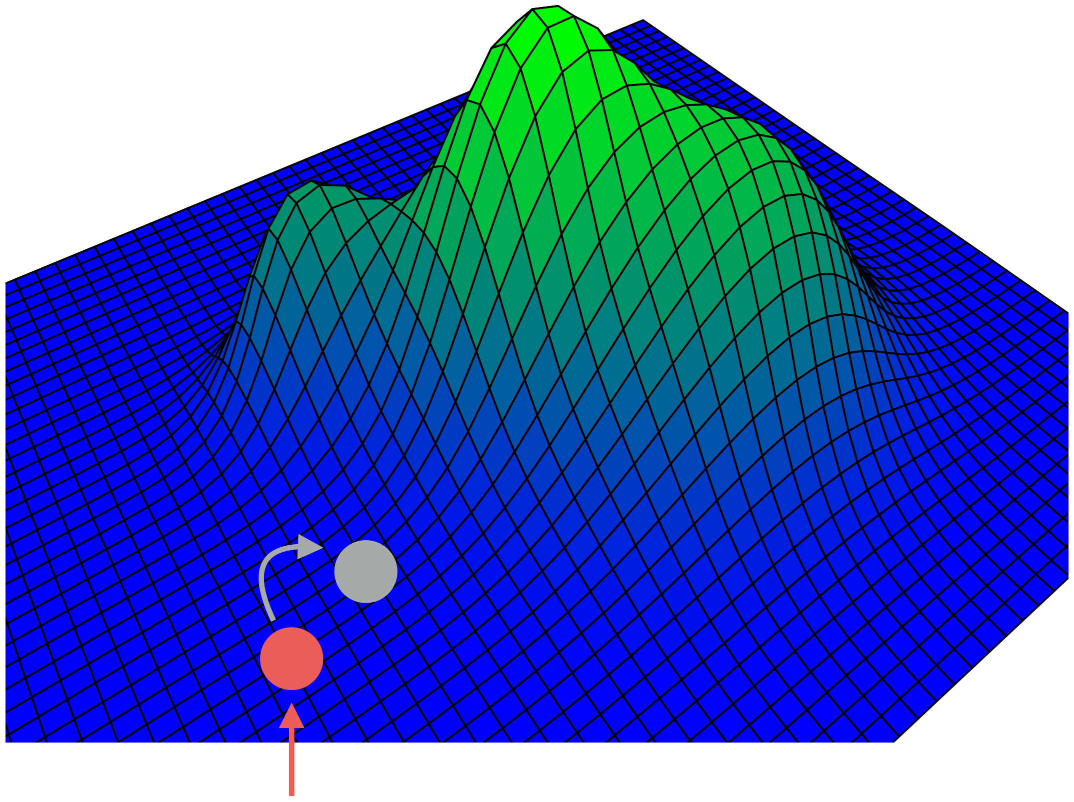






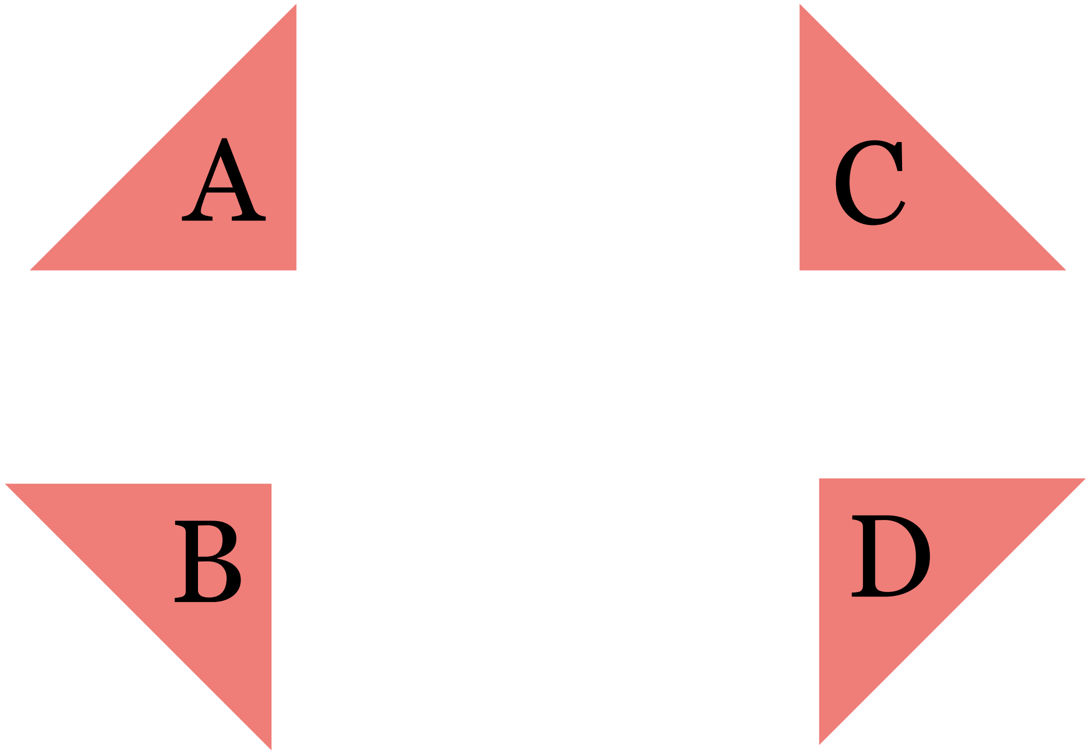
Nearest Neighbor Interchange

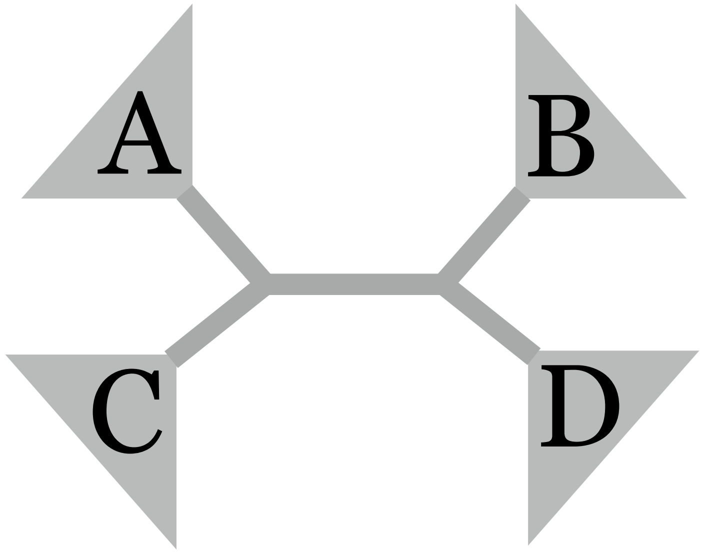
Why is MCMC so slow? Traverse tree space

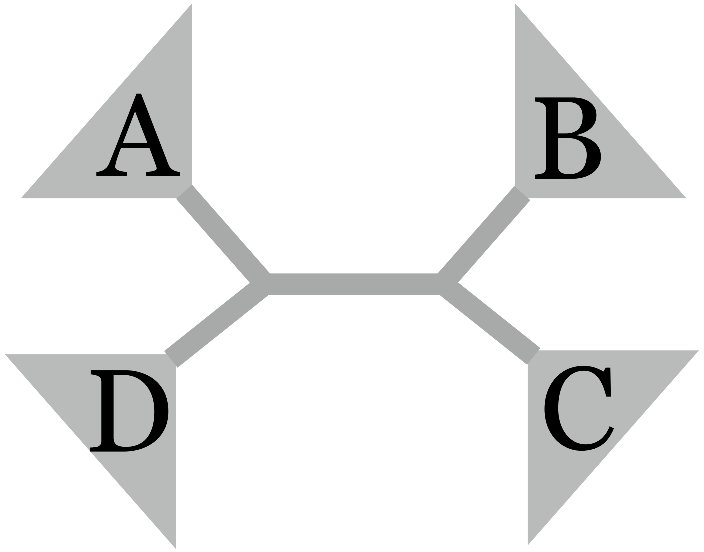


Starting tree

Nearest Neighbor Interchange







NNI

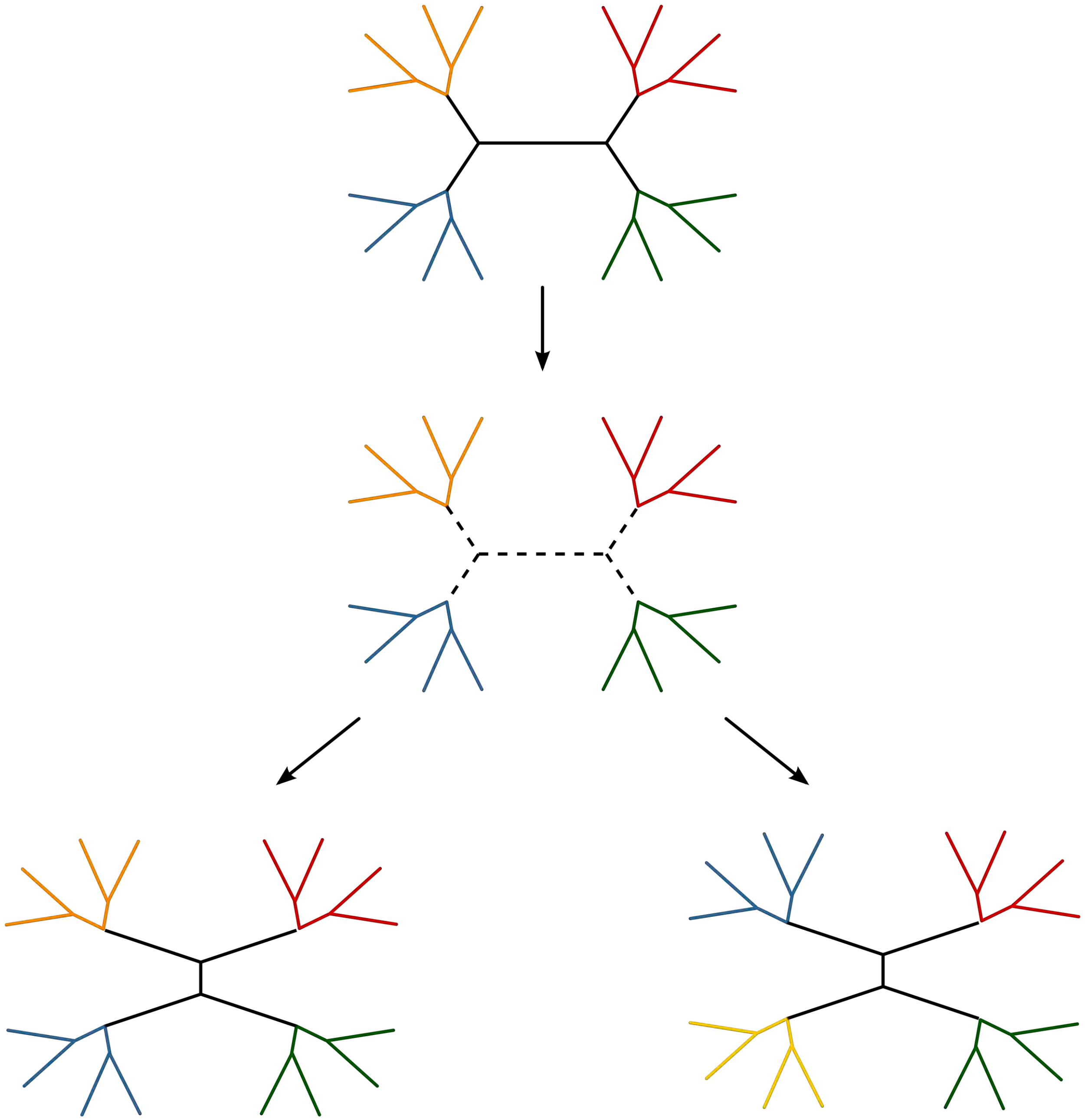


Image: Wikipedia

SPR

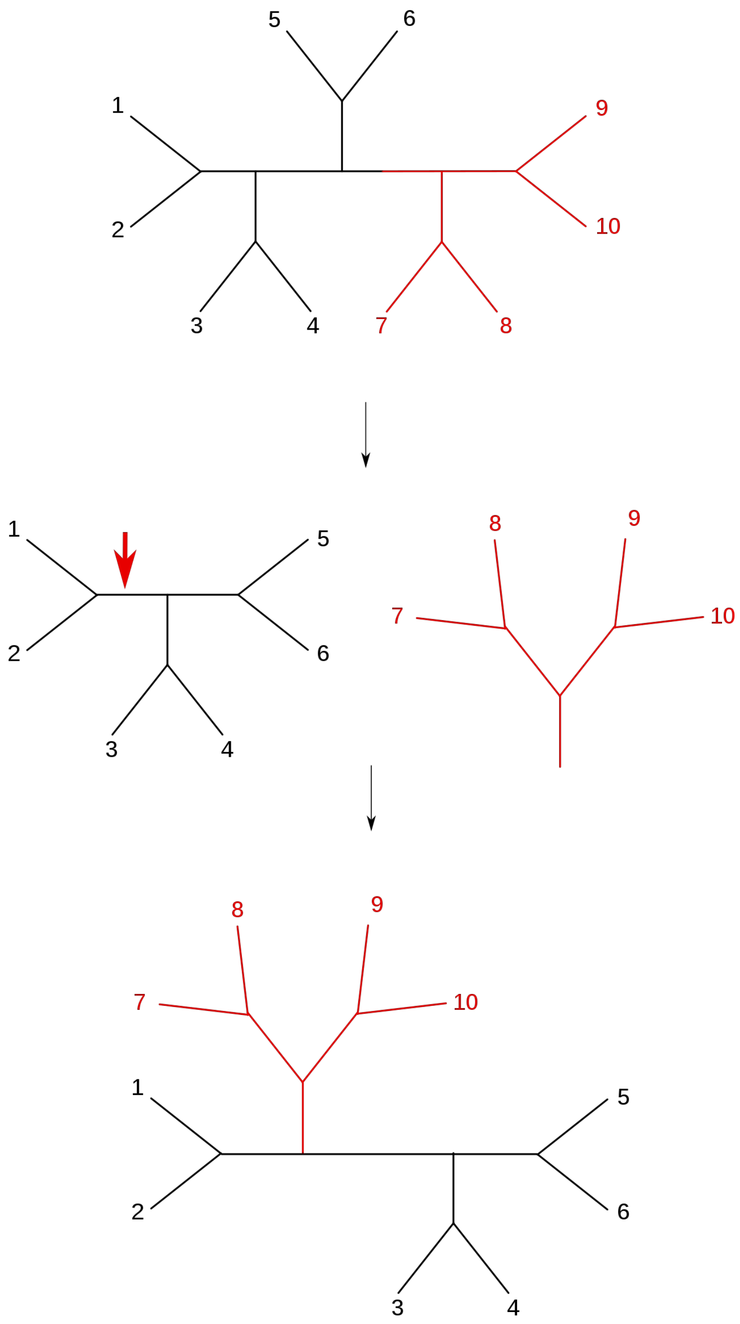


Image: Wikipedia

TBR

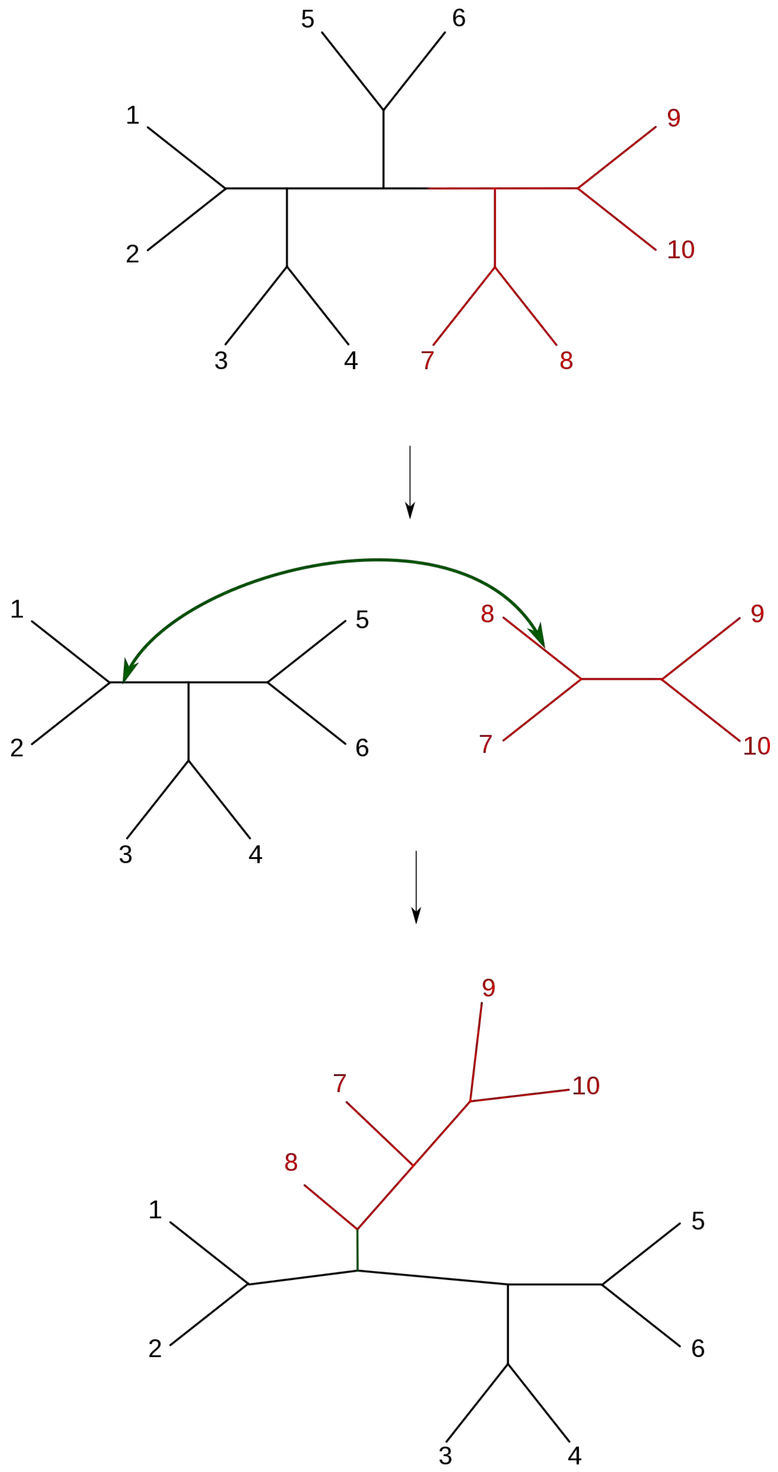
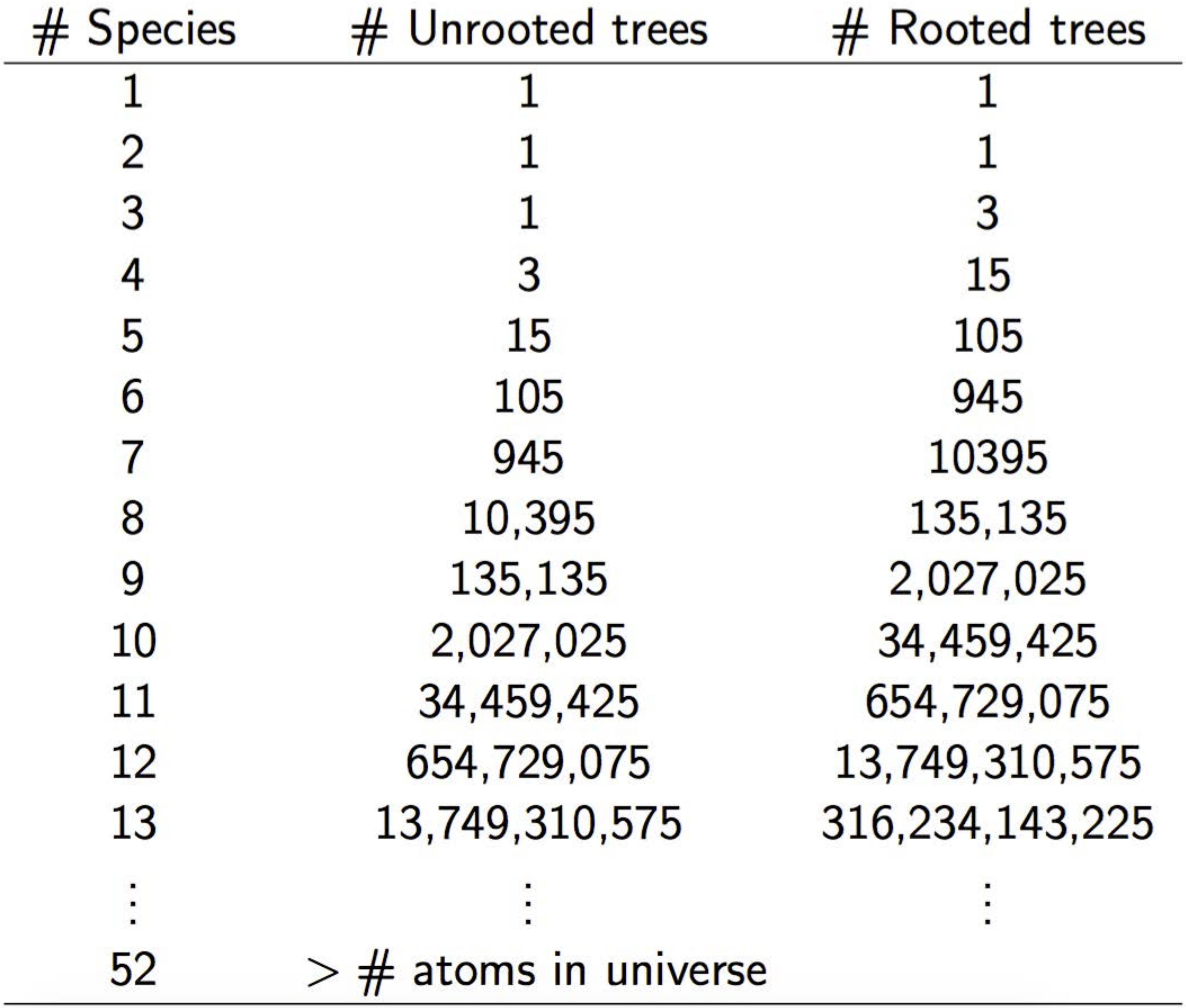


Image: Wikipedia

Why is MCMC so slow?

Why is MCMC so slow? Tree space is huge

Why is MCMC so slow? Tree space is huge

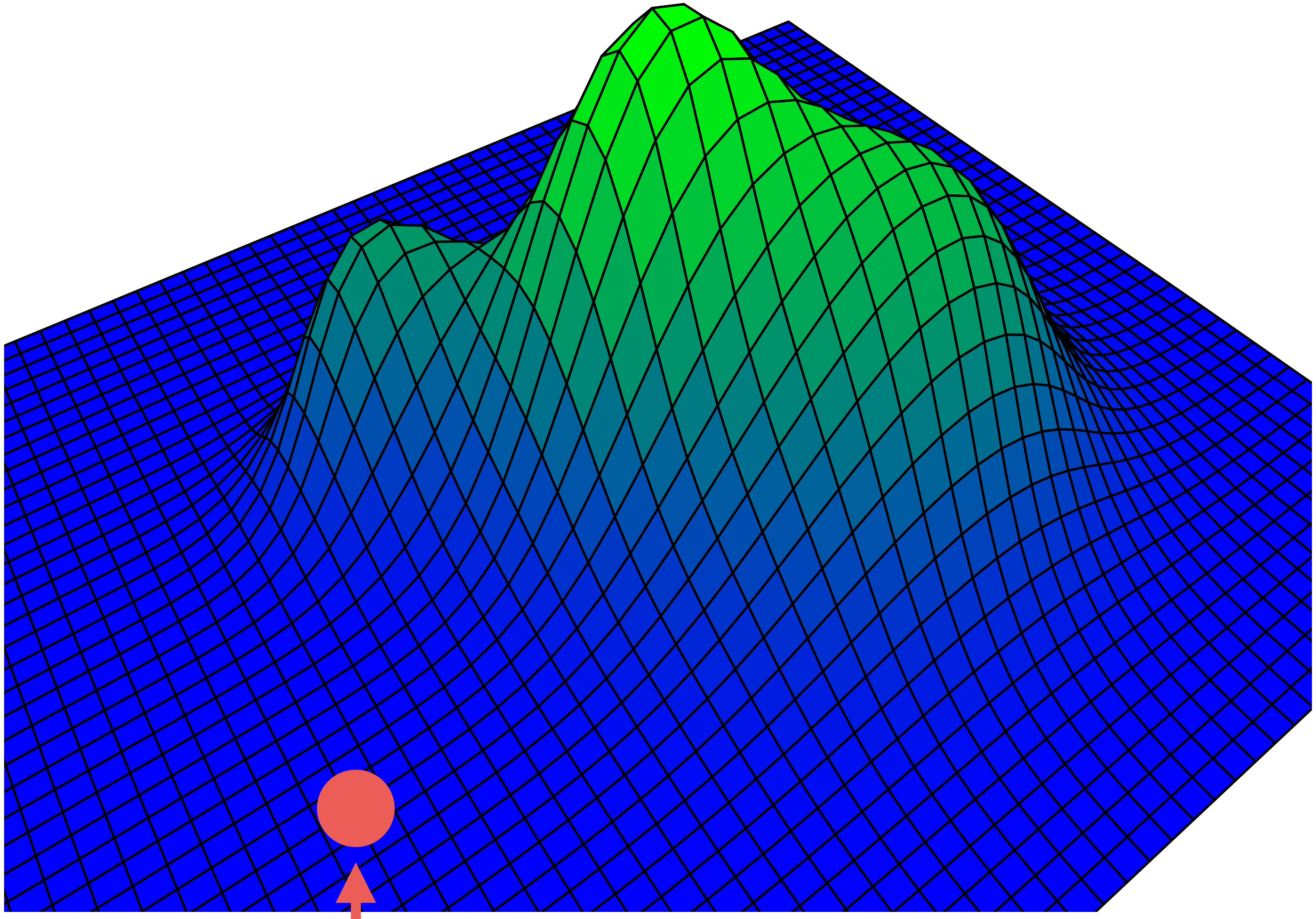


# Species	# Unrooted trees	# Rooted trees
1	1	1
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10,395	135,135
9	135,135	2,027,025
10	2,027,025	34,459,425
11	34,459,425	654,729,075
12	654,729,075	13,749,310,575
13	13,749,310,575	316,234,143,225
⋮	⋮	⋮
52	> # atoms in universe	

Why is MCMC so slow?

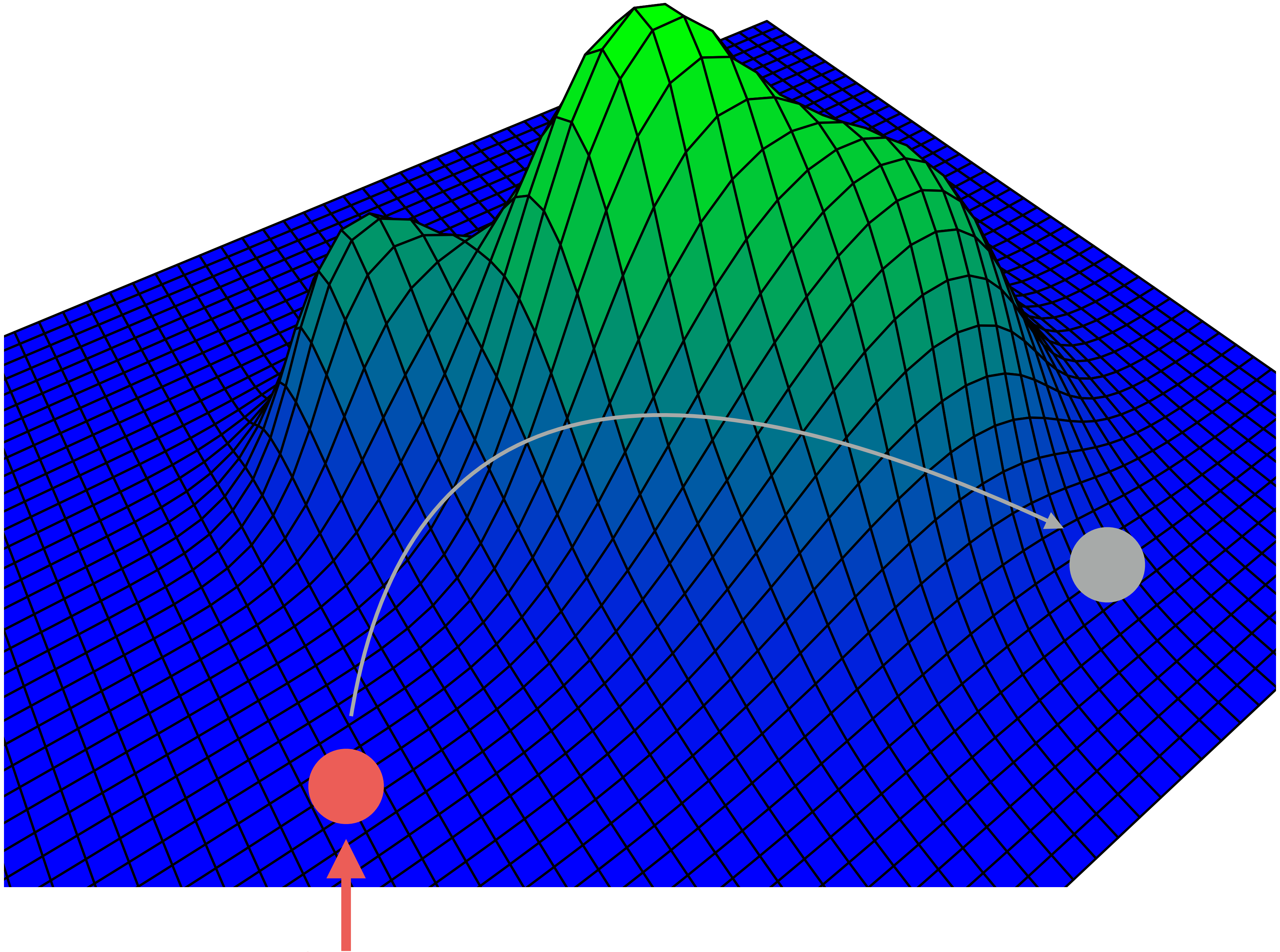
Why is MCMC so slow? Low acceptance of moves

Why is MCMC so slow? Low acceptance of moves



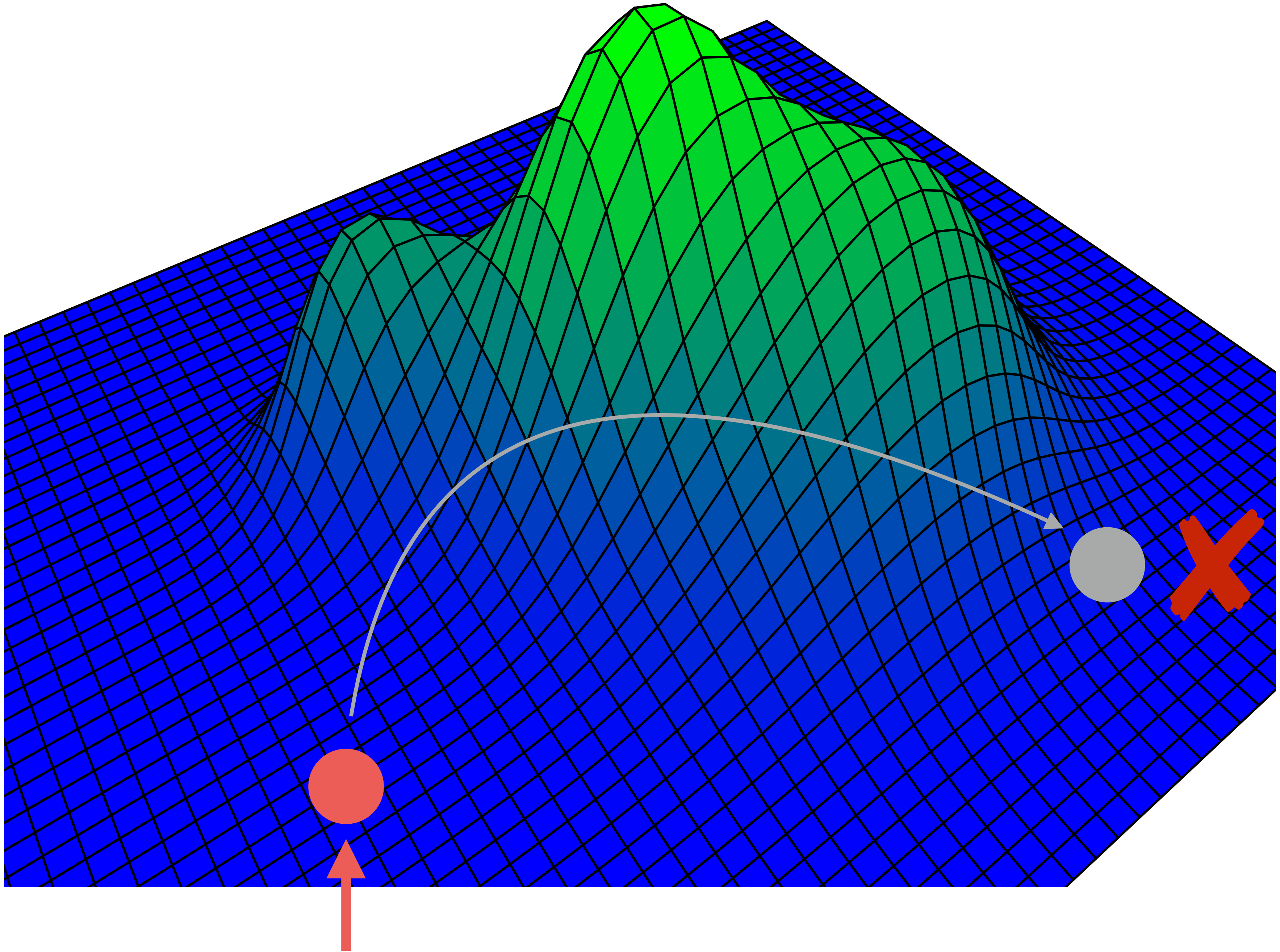
Starting tree

Why is MCMC so slow? Low acceptance of moves



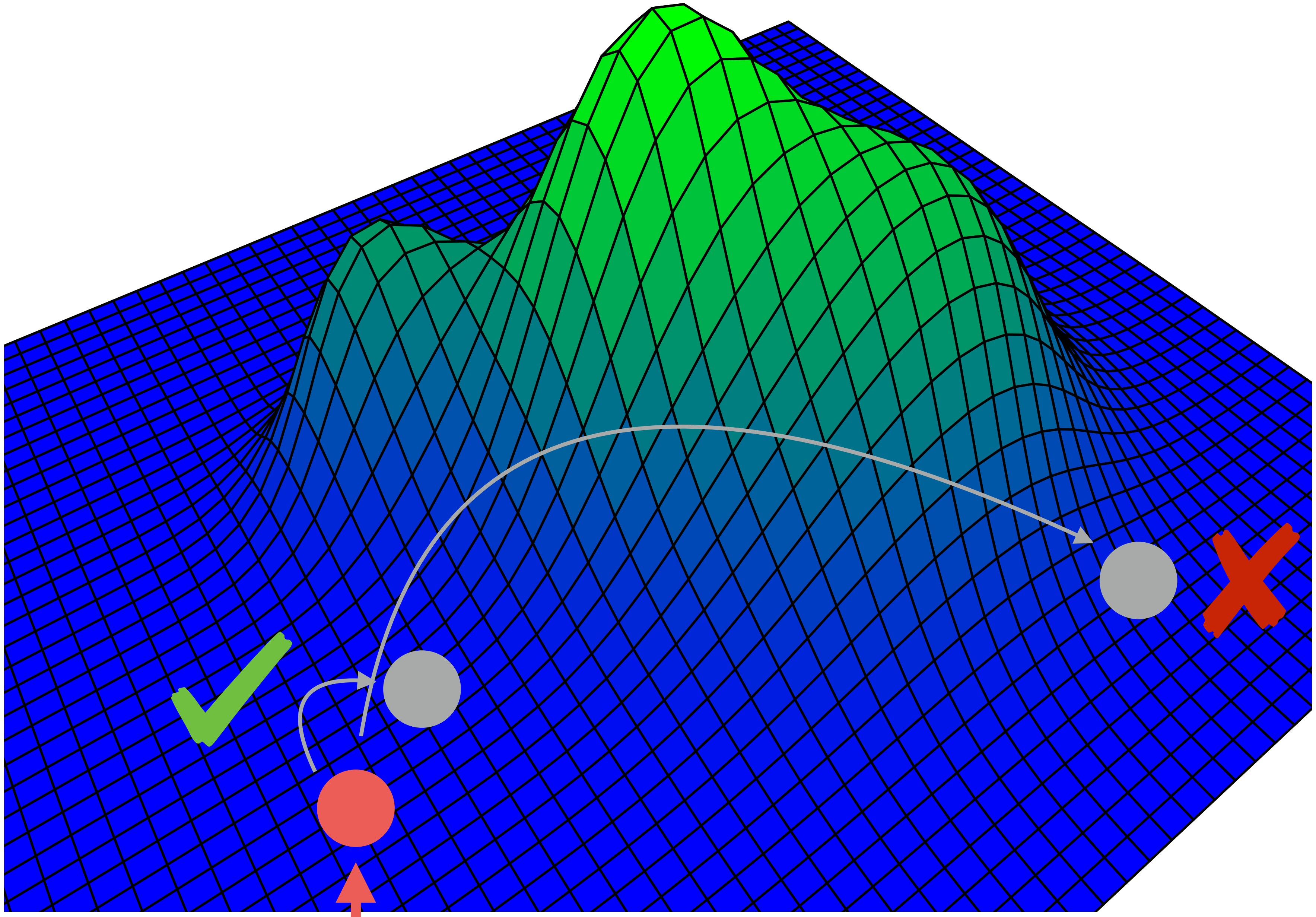
Starting tree

Why is MCMC so slow? Low acceptance of moves



Starting tree

Why is MCMC so slow? Low acceptance of moves



Starting tree

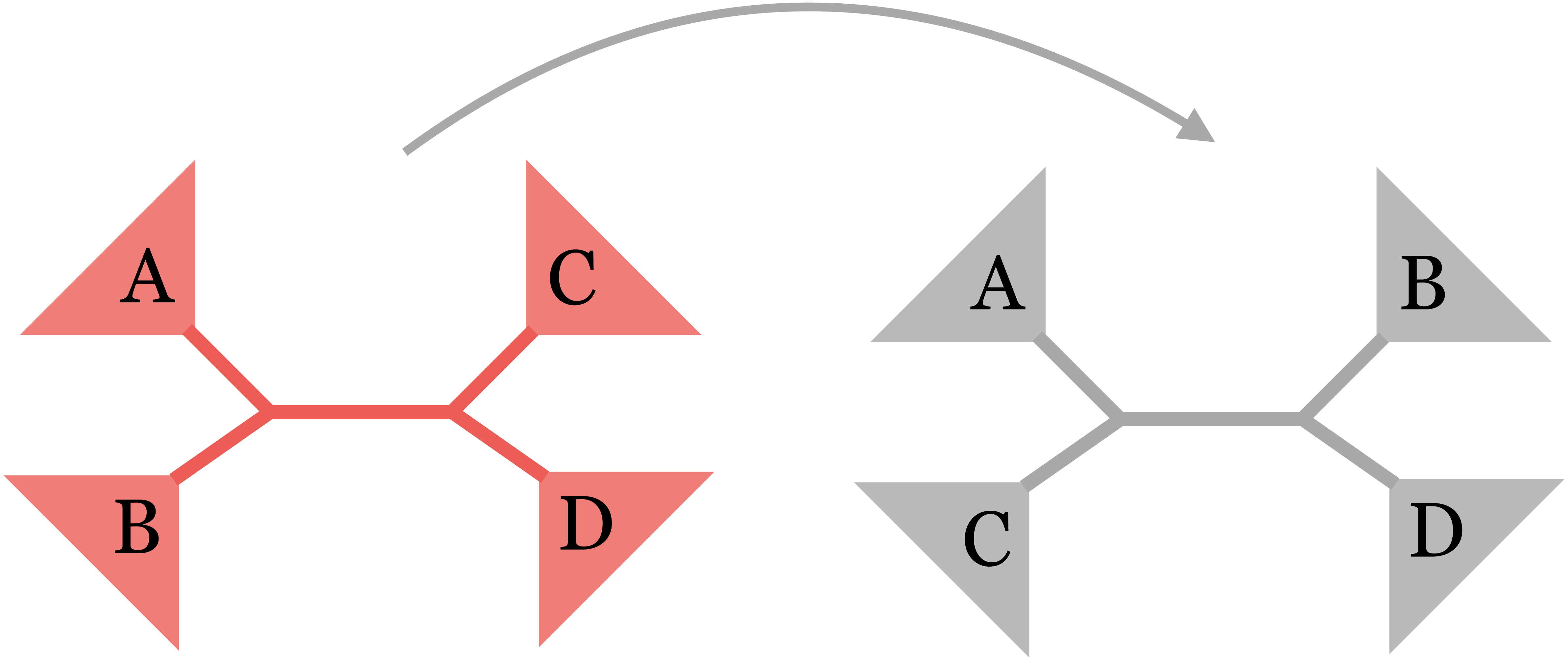
Why is MCMC so slow?

Why is MCMC so slow?

Small neighborhood implies very dependent sample

Why is MCMC so slow?

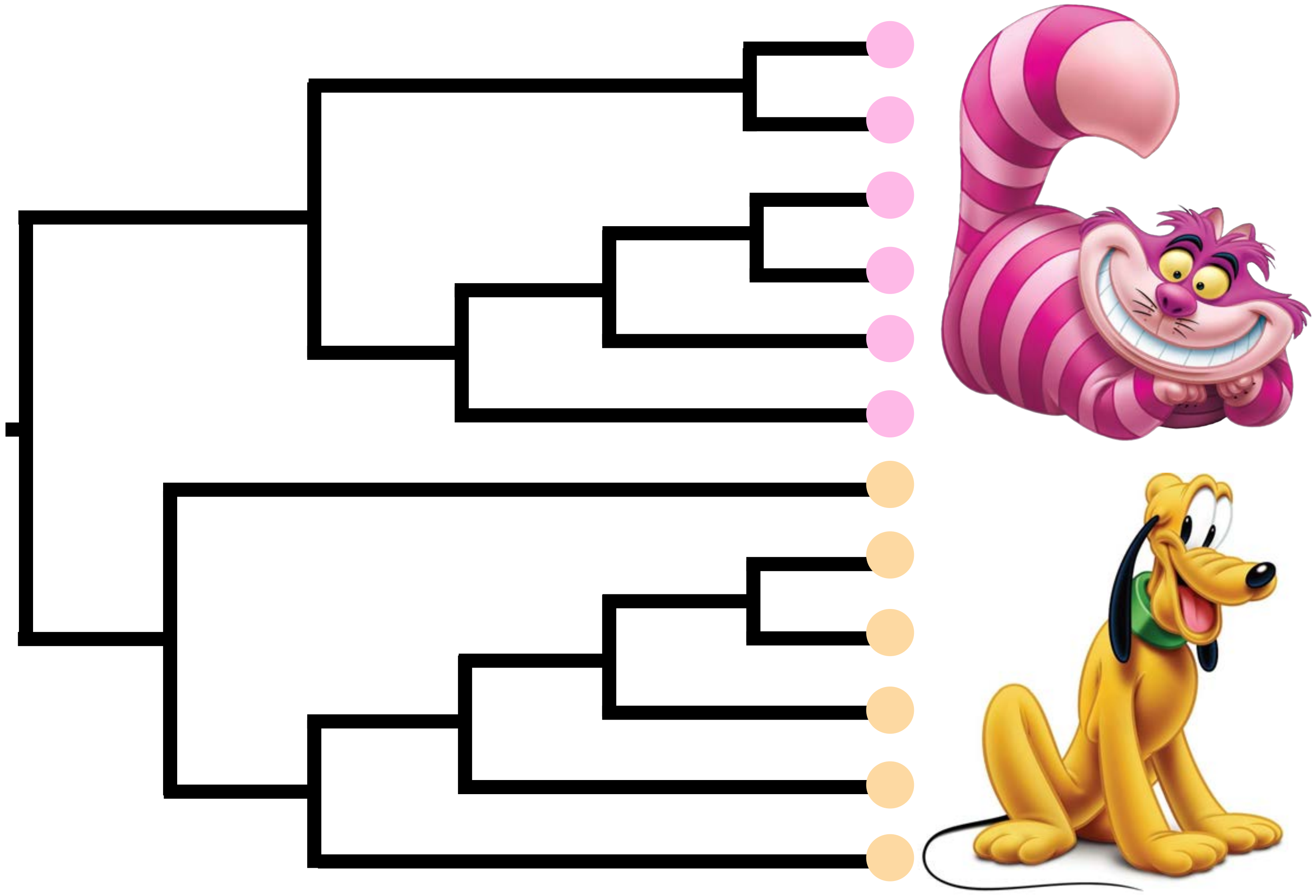
Small neighborhood implies very dependent sample



Why is MCMC so slow?

1) Huge tree space size
2) Low acceptance of moves unless small neighborhood
3) Small neighborhood implies very dependent sample, which means small effective sample size

We need a gigantic chain because the space is huge and we are making tiny moves



12 taxa *Carnivora*

MCMC efficiency ~0.025%

(250 from 1 million post-burnin generations)

↑

ESS

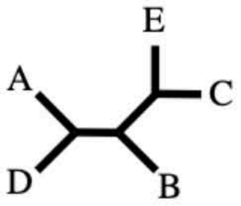
(Larget, 2013)

Priors

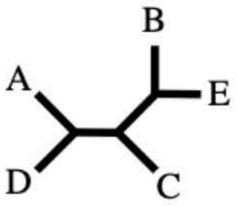
Common Priors

- **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

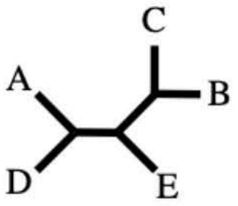
Discrete Uniform distribution for topologies



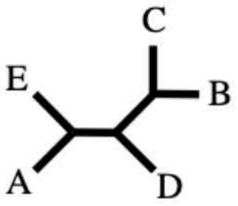
$$\frac{1}{15}$$



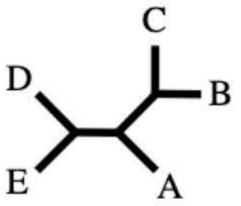
$$\frac{1}{15}$$



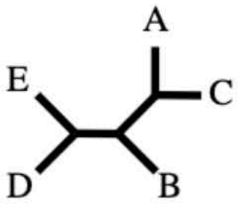
$$\frac{1}{15}$$



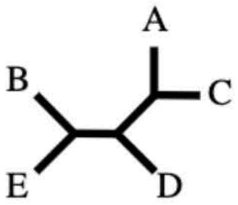
$$\frac{1}{15}$$



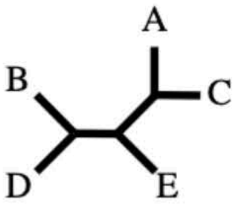
$$\frac{1}{15}$$



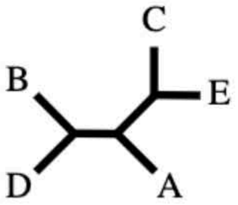
$$\frac{1}{15}$$



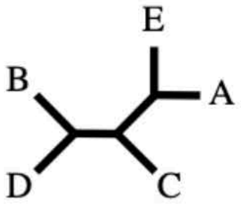
$$\frac{1}{15}$$



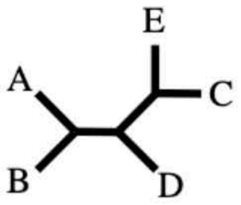
$$\frac{1}{15}$$



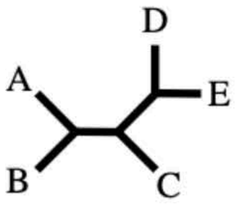
$$\frac{1}{15}$$



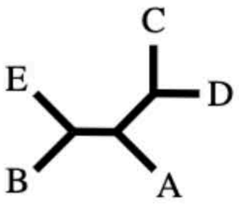
$$\frac{1}{15}$$



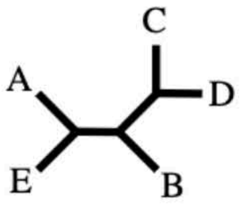
$$\frac{1}{15}$$



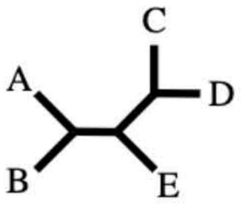
$$\frac{1}{15}$$



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



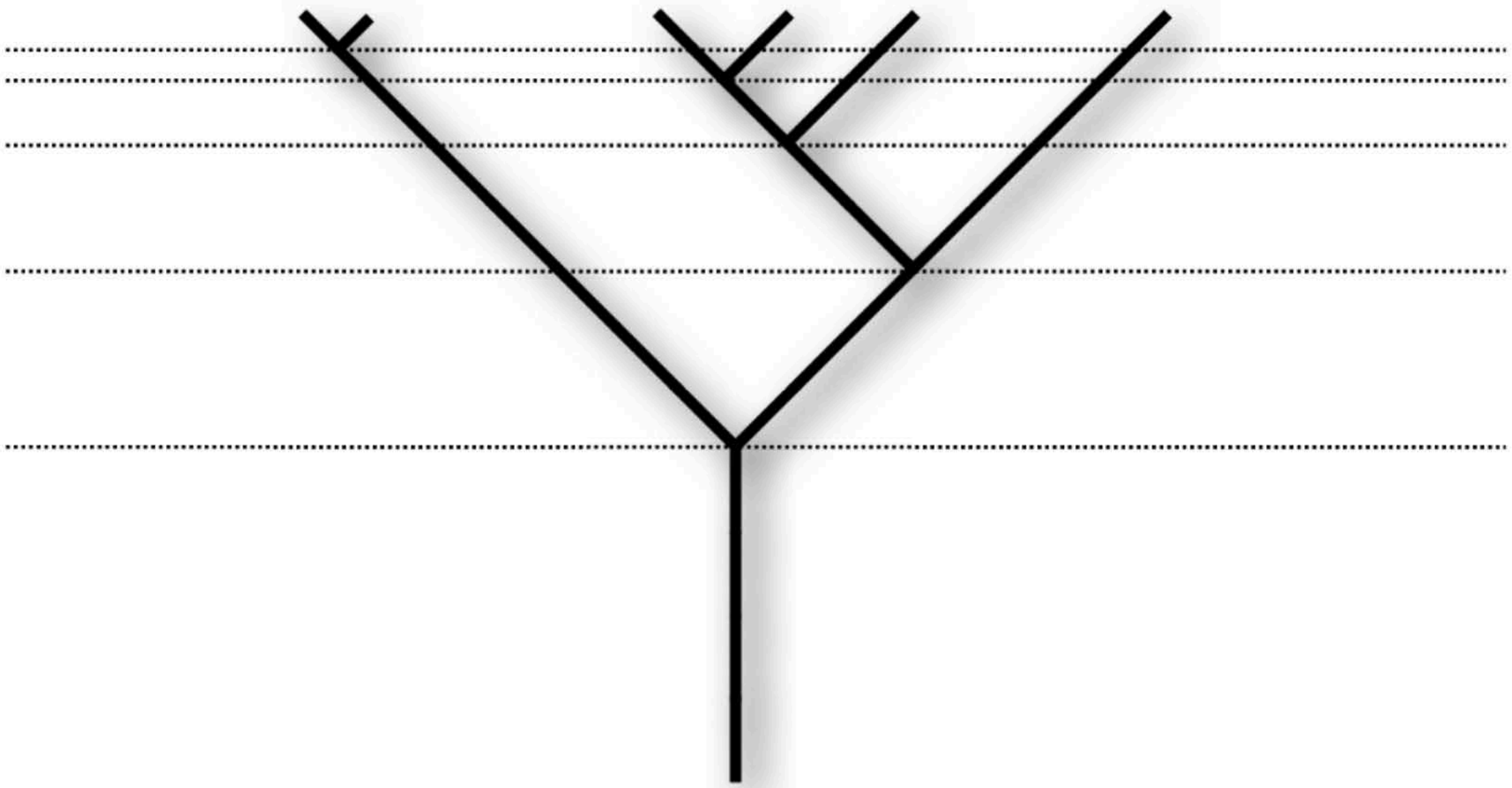
$$\frac{1}{15}$$



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

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

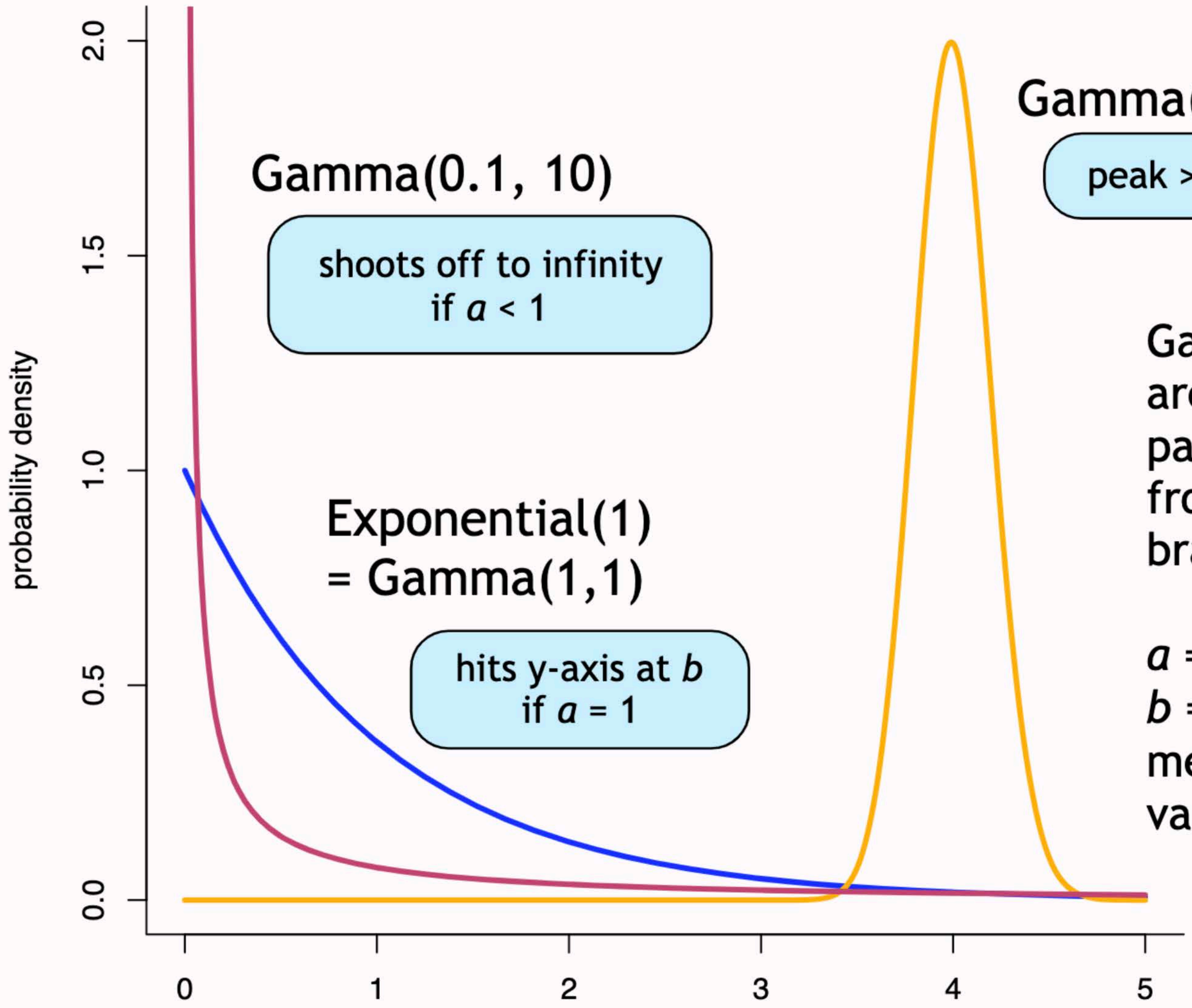
Yule model provides joint prior for both **topology** and **divergence times**



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.

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Gamma(a,b) distributions



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(400, 0.01)

peak > 0 if $a > 1$

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

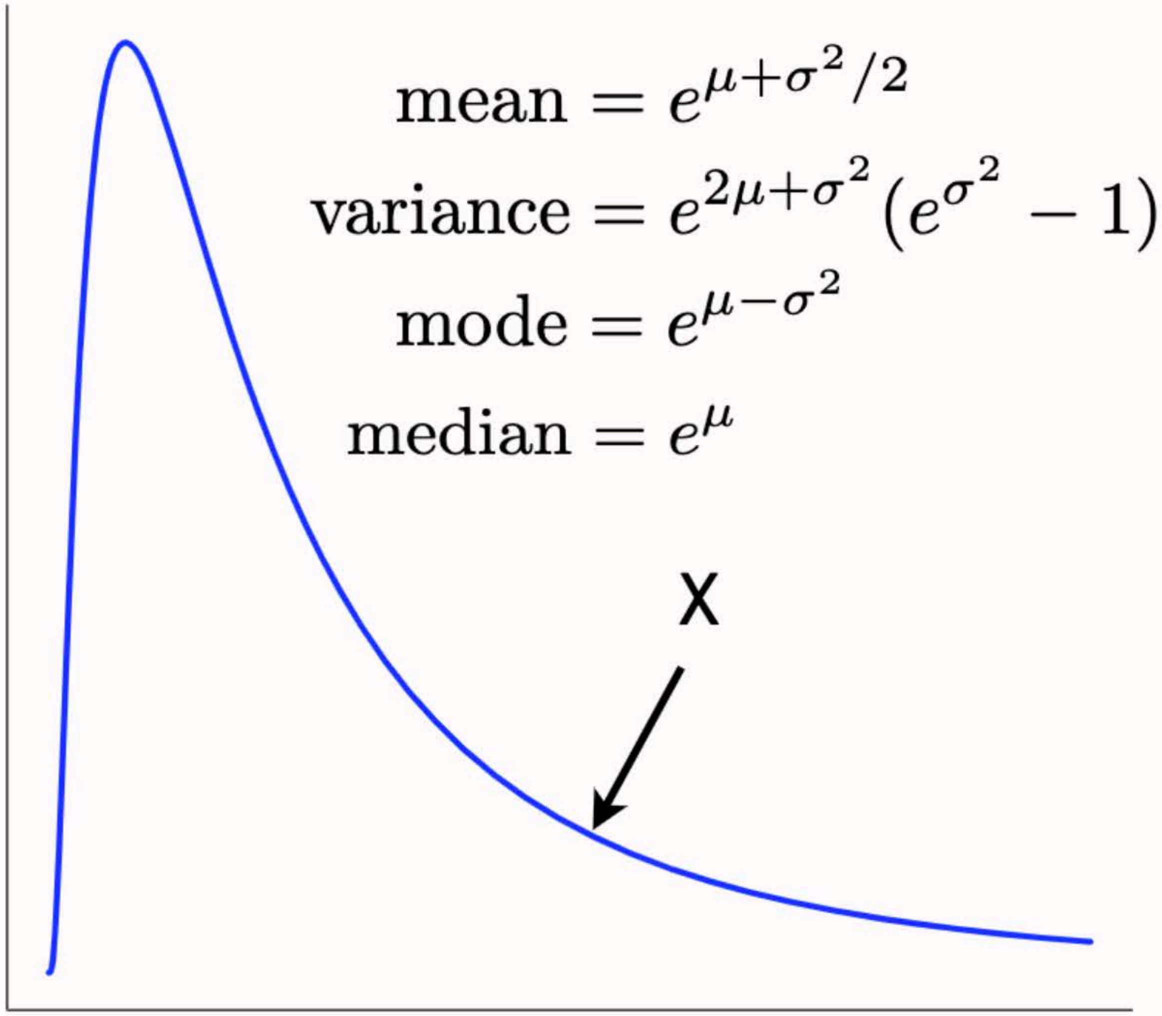
a = shape
b = scale
mean* = ab
variance* = ab^2

*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.

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

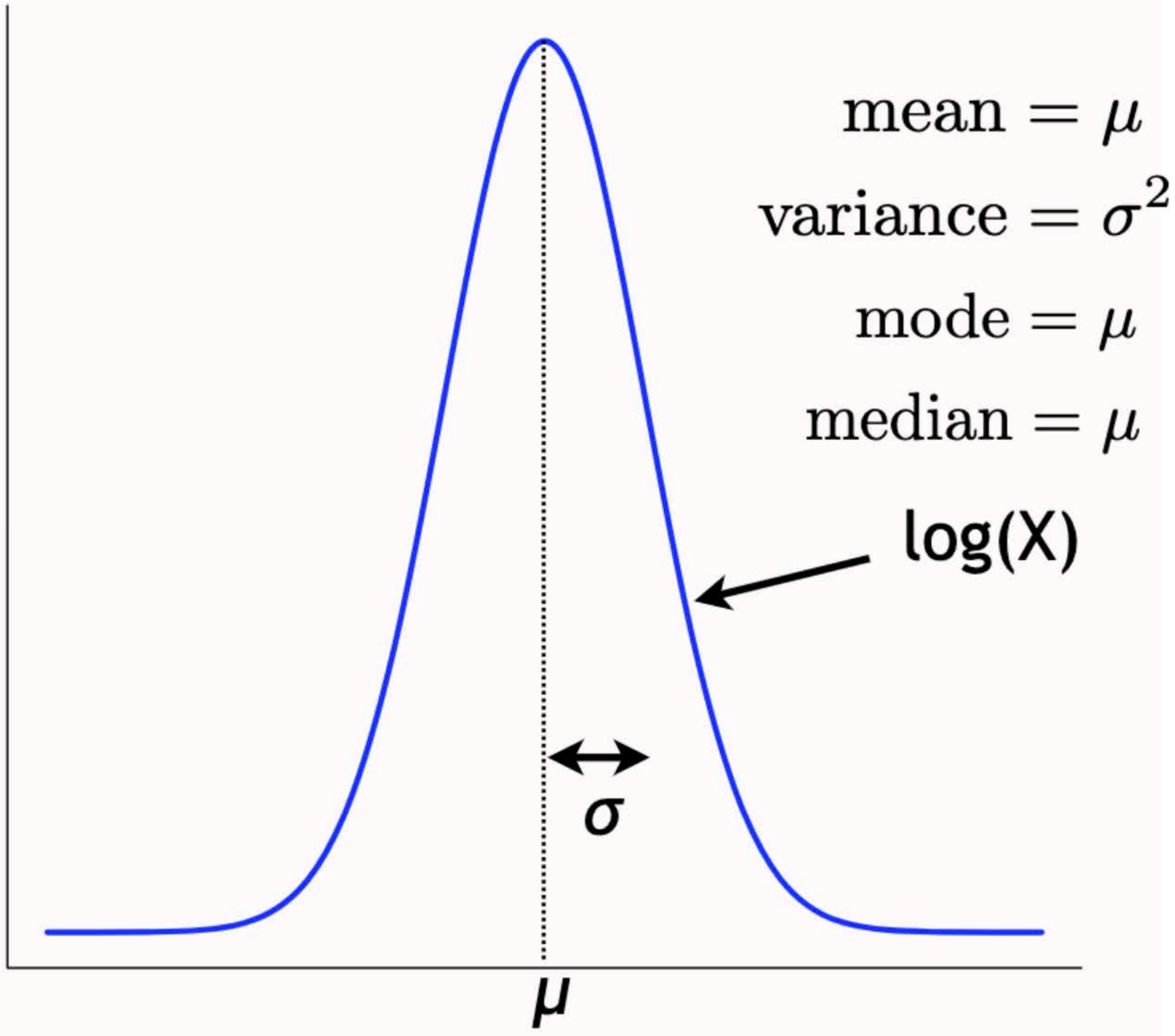
Log-normal distribution

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.



$$\text{mean} = \mu$$
$$\text{variance} = \sigma^2$$
$$\text{mode} = \mu$$
$$\text{median} = \mu$$

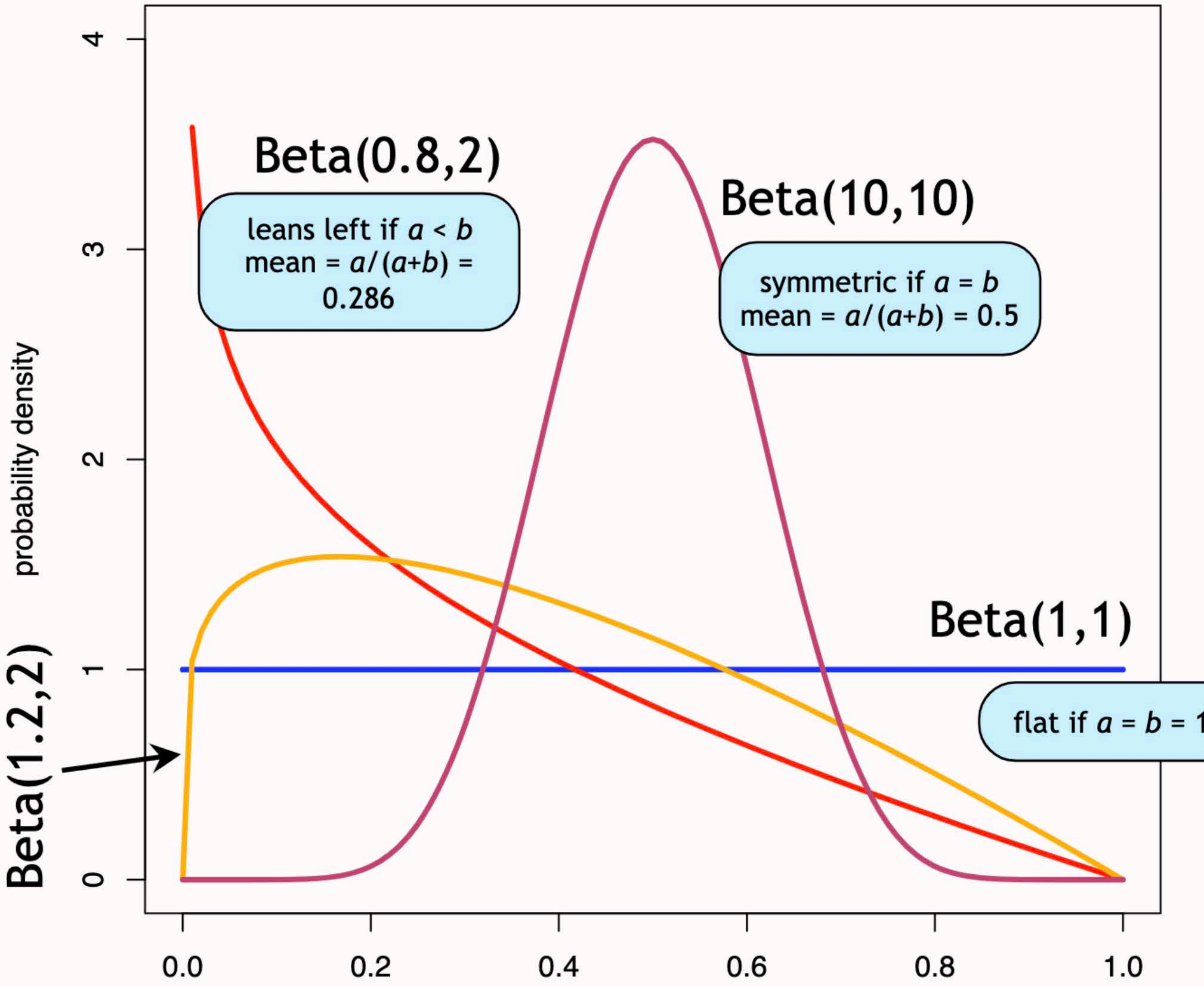
Important: *μ and σ do **not** represent the mean and standard deviation of X: they are the mean and standard deviation of log(X)!*

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

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

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

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

Non-informative prior

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

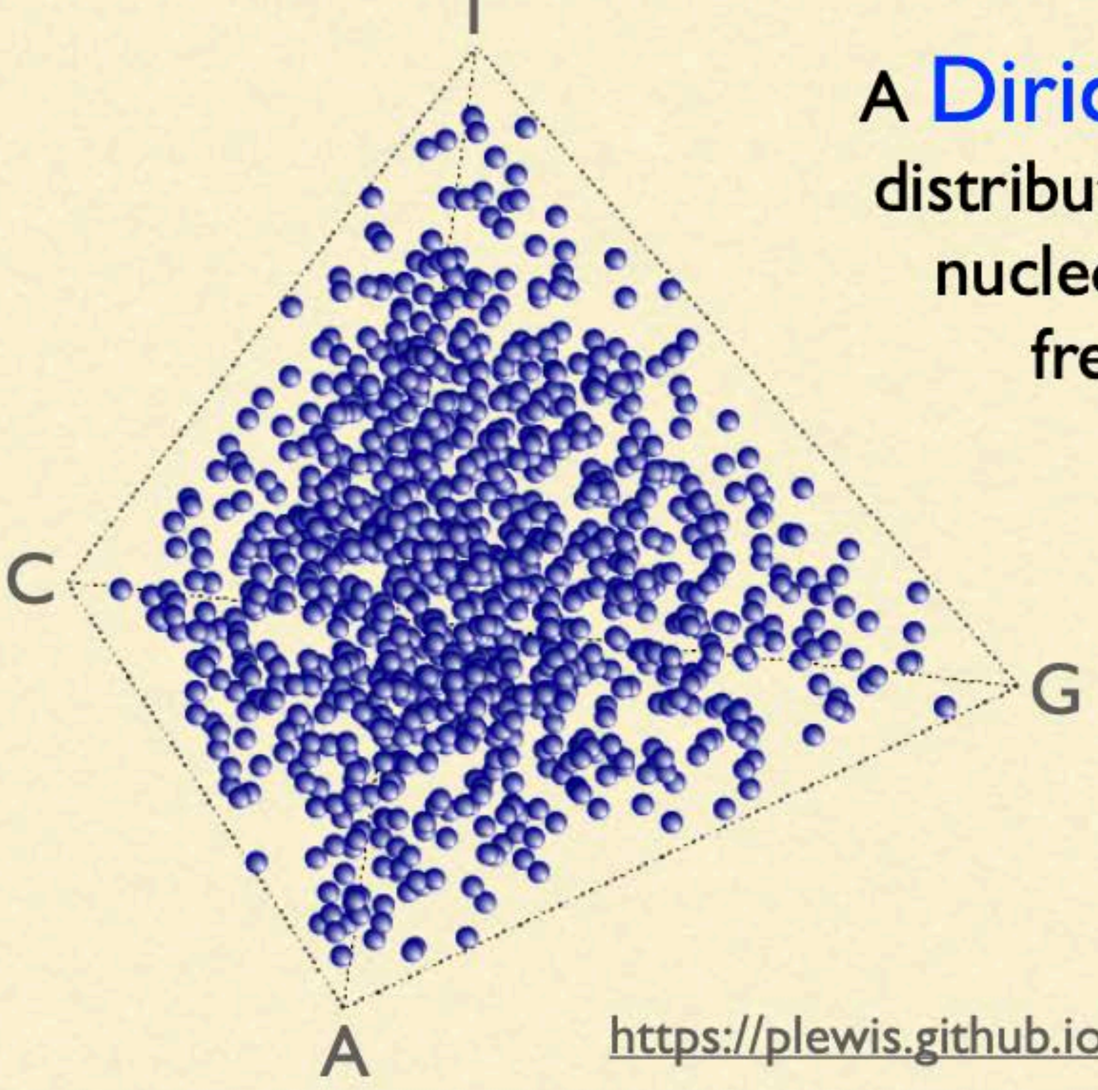
mean = a/(a+b)
variance =
ab/[(a+b)^2(a+b+1)]

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Dirichlet(*a*,*b*,*c*,*d*) distribution

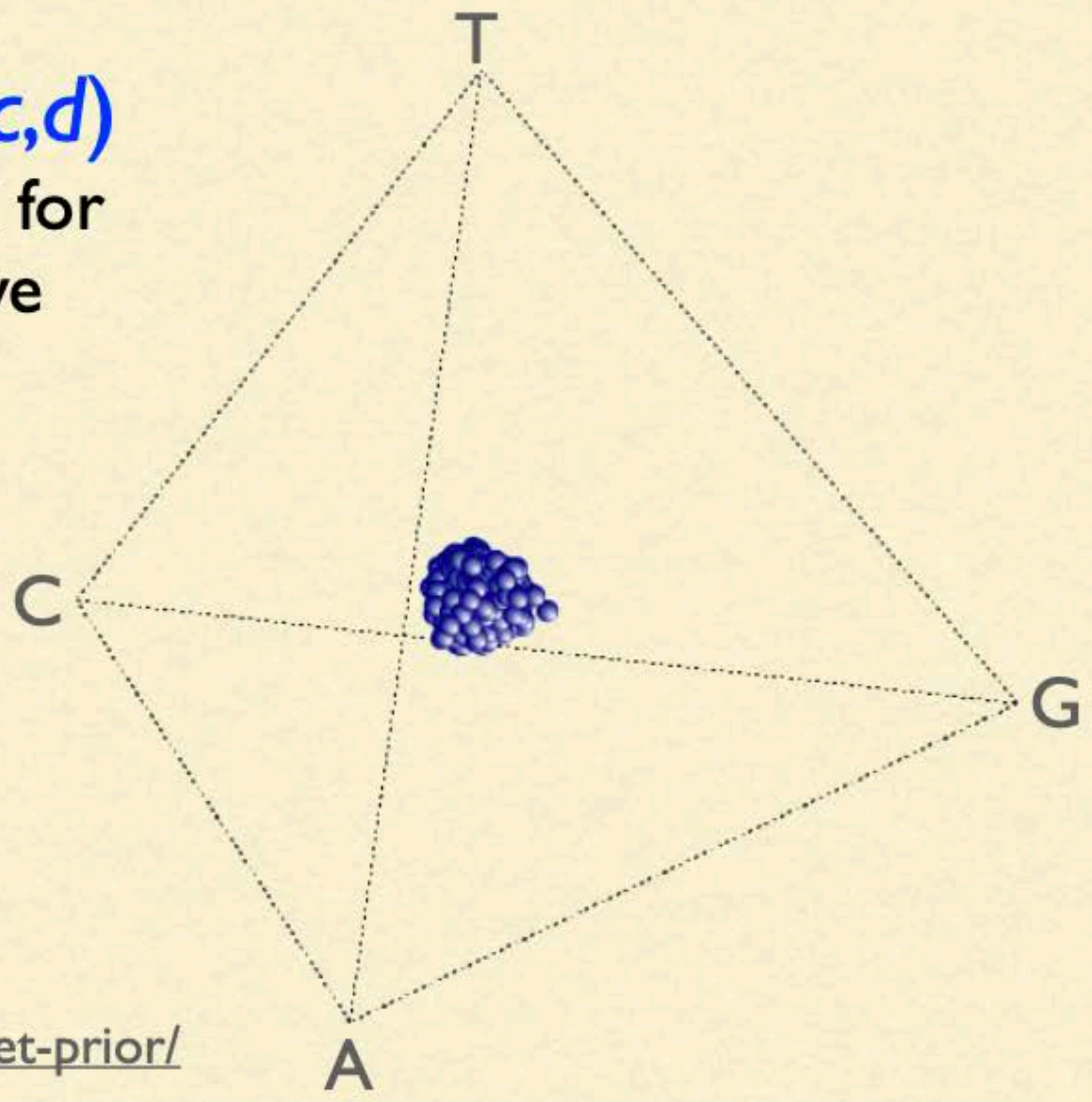
Flat: *a* = *b* = *c* = *d* = 1

(every combination equally probable)



Informative: *a* = *b* = *c* = *d* = 100

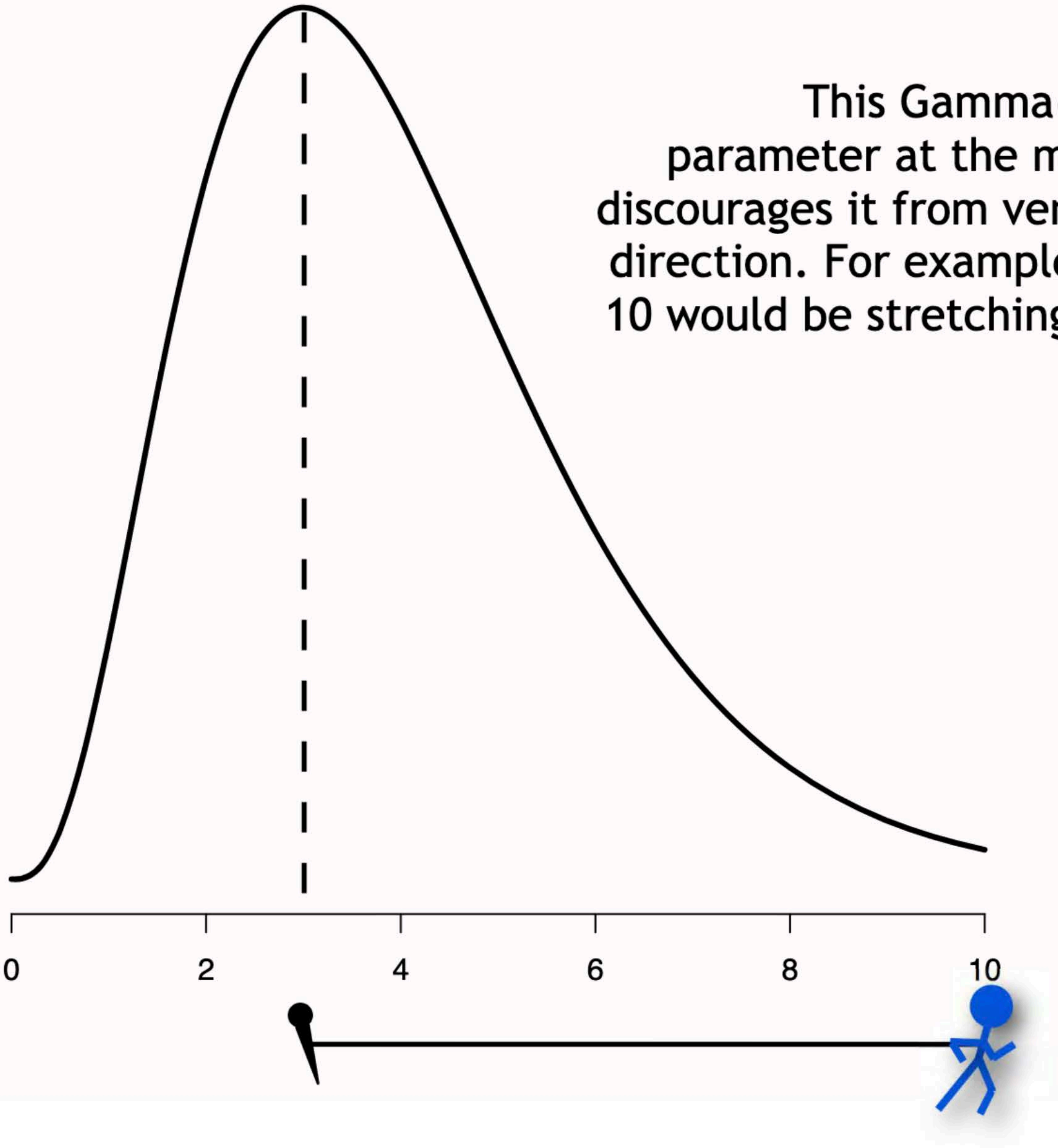
(frequencies tend to be nearly equal)



A **Dirichlet(*a*,*b*,*c*,*d*)** distribution is ideal for nucleotide relative frequencies.

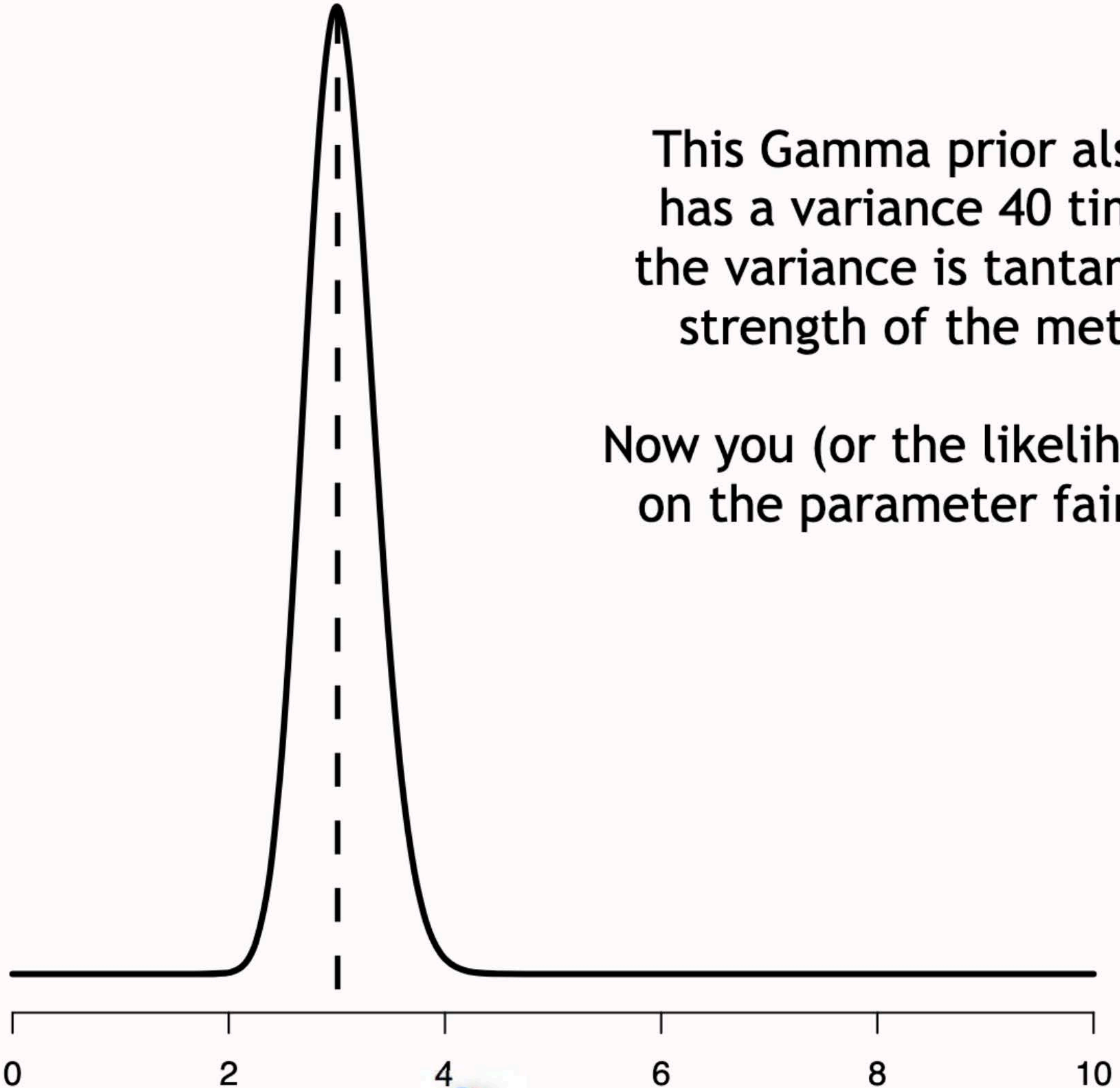
https://plewis.github.io/applets/dirichlet-prior/

Priors as rubber bands



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)



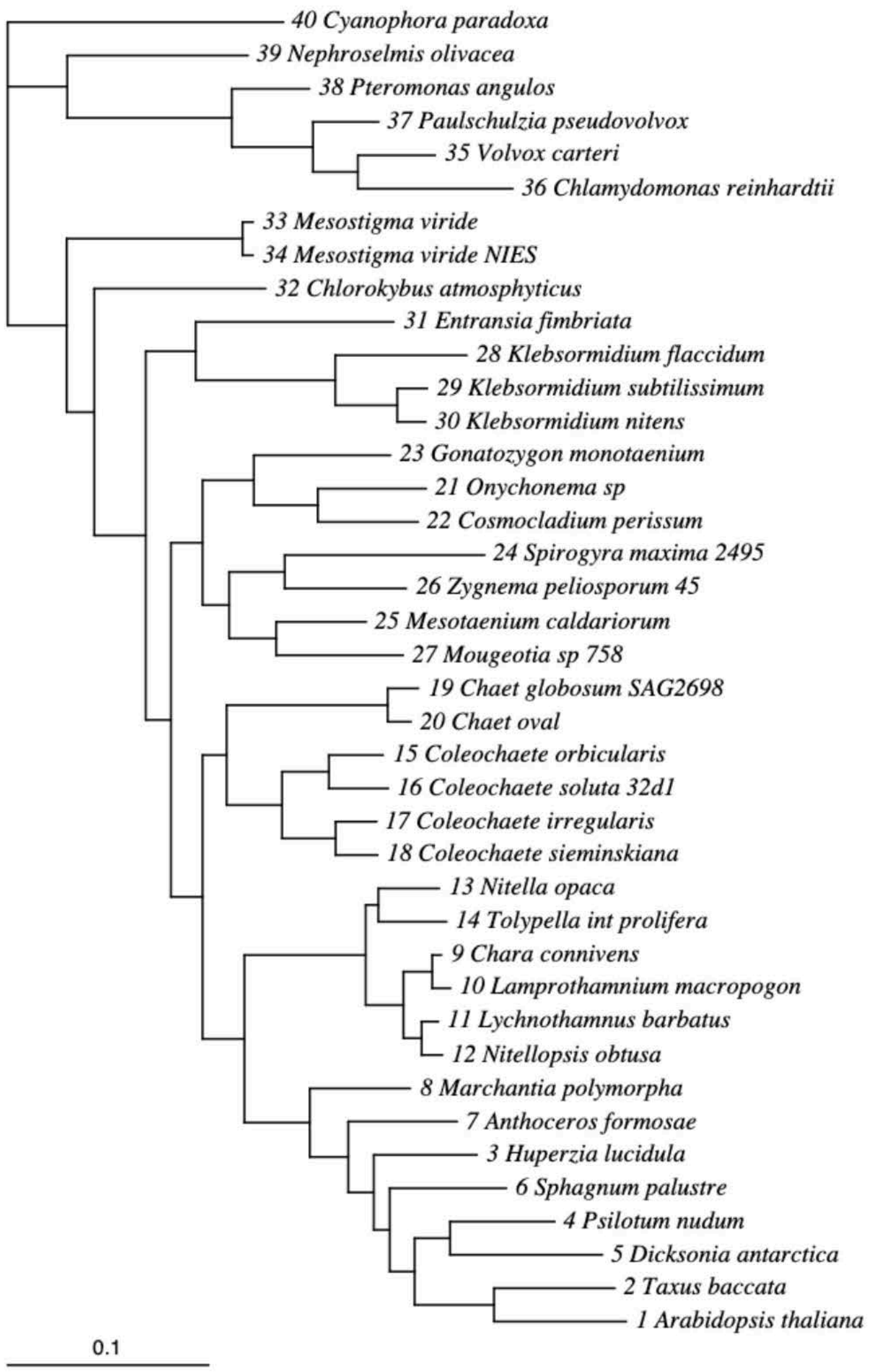
This Gamma prior also has a mode at 3, but has a variance 40 times smaller. Decreasing the variance is tantamount to increasing the strength of the metaphorical rubber band.

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

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Example: Internal Branch Length Priors



Separate priors applied to internal and external branches

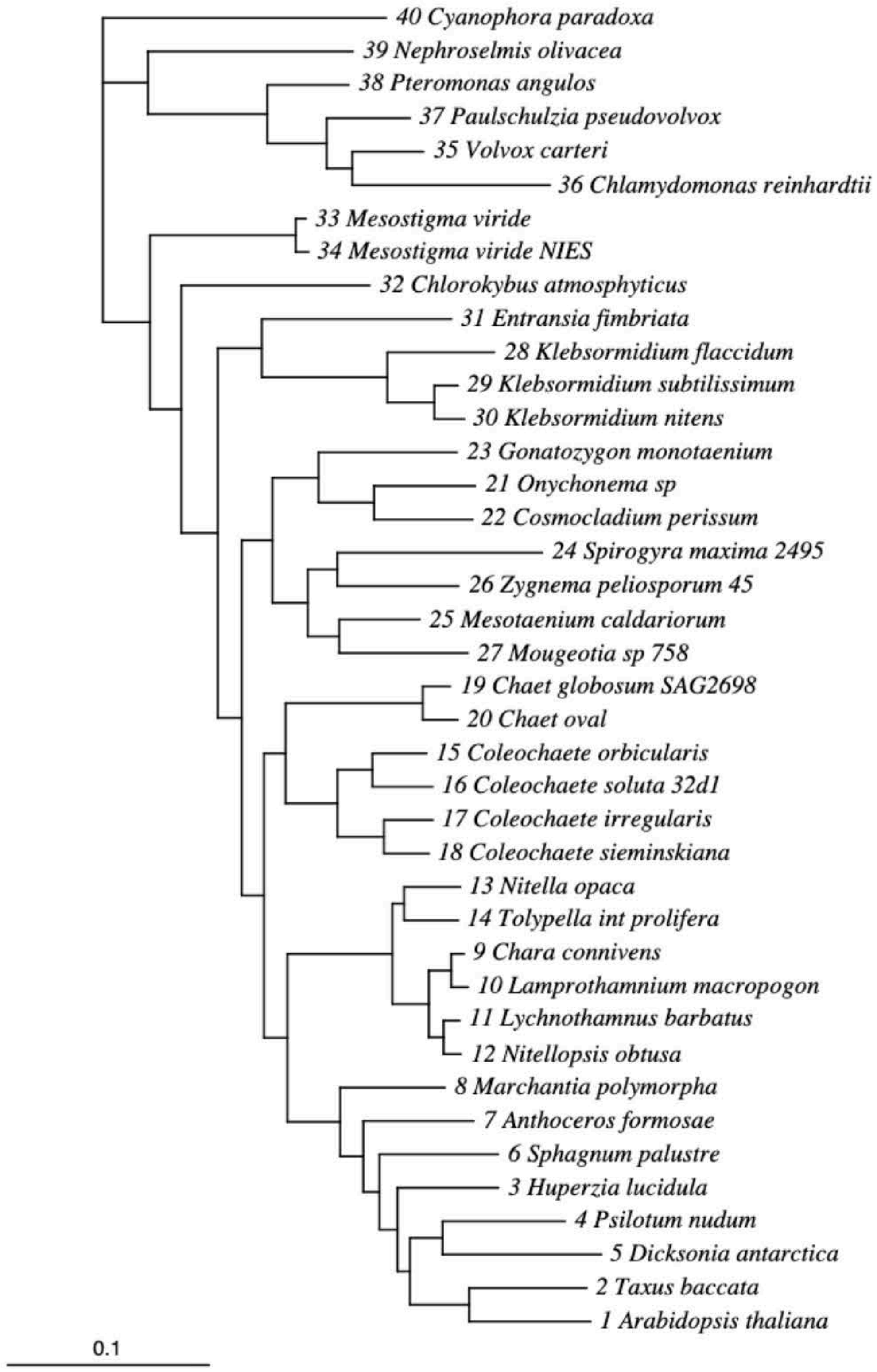
External branch length prior is exponential with mean 0.1

Internal branch length prior is exponential with mean 0.1

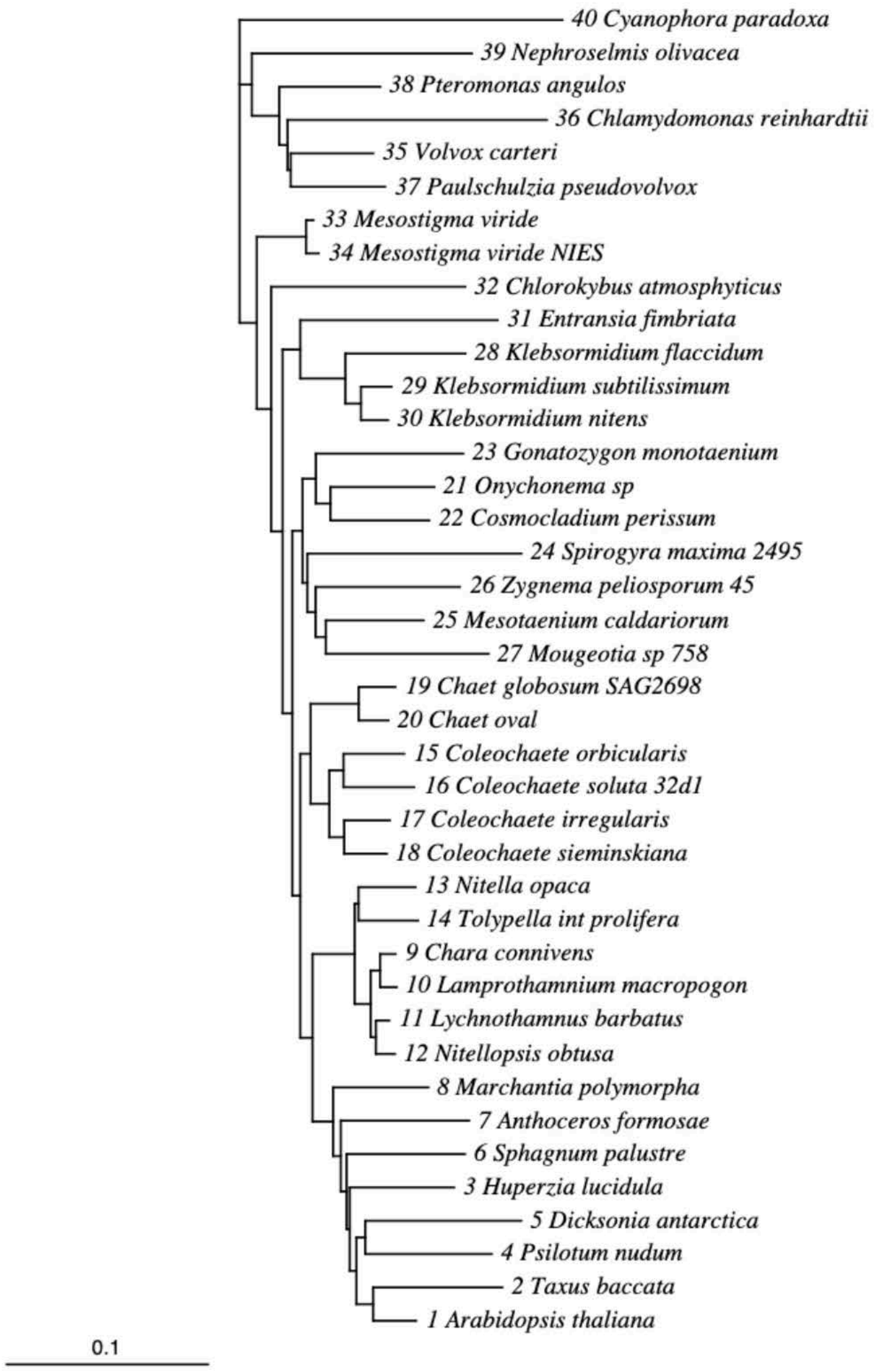
This is a reasonably vague internal branch length prior

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Internal branch length prior mean 0.001

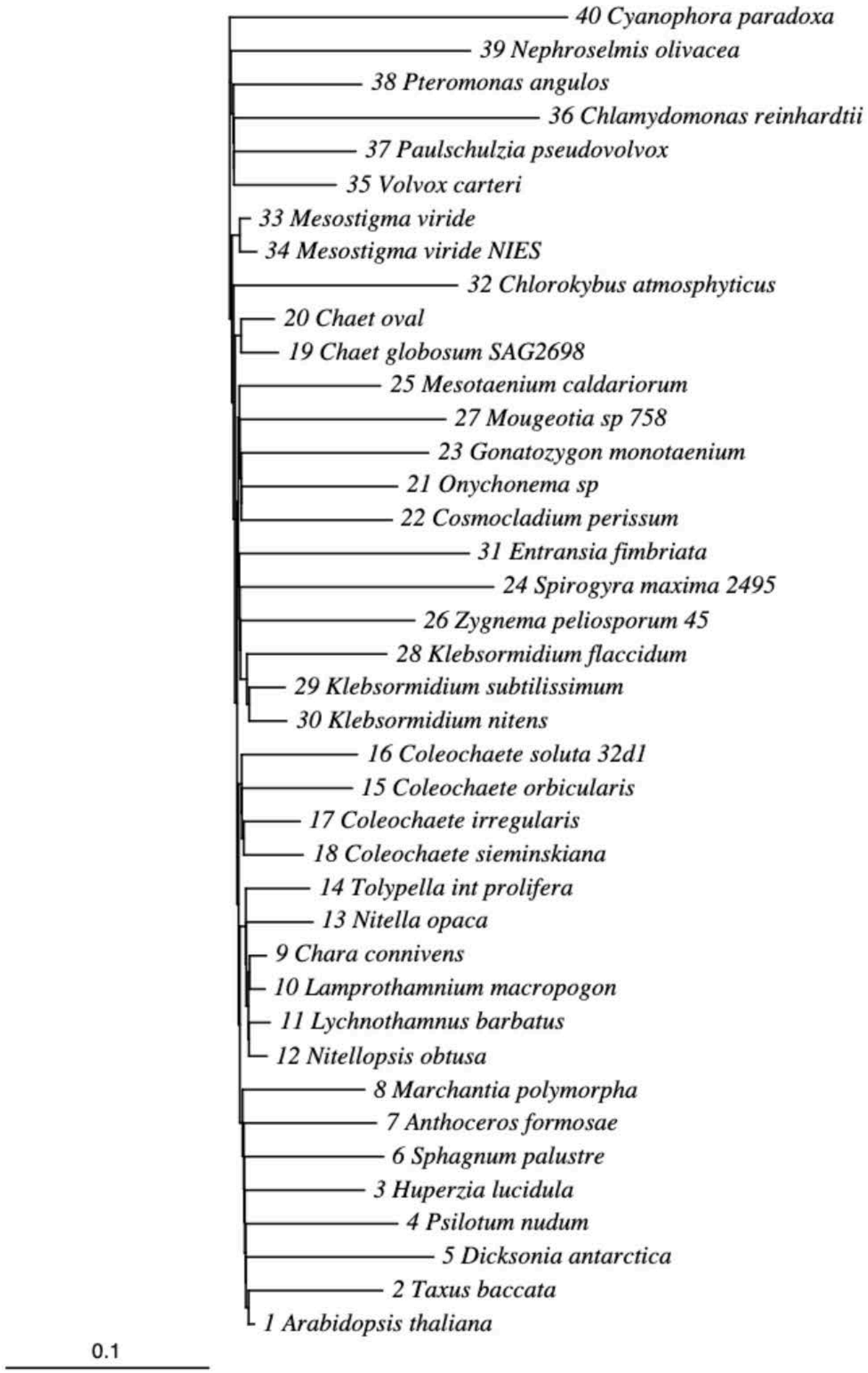


Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)



Internal branch length prior mean
0.0001

Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)



Internal branch length prior mean
0.00001

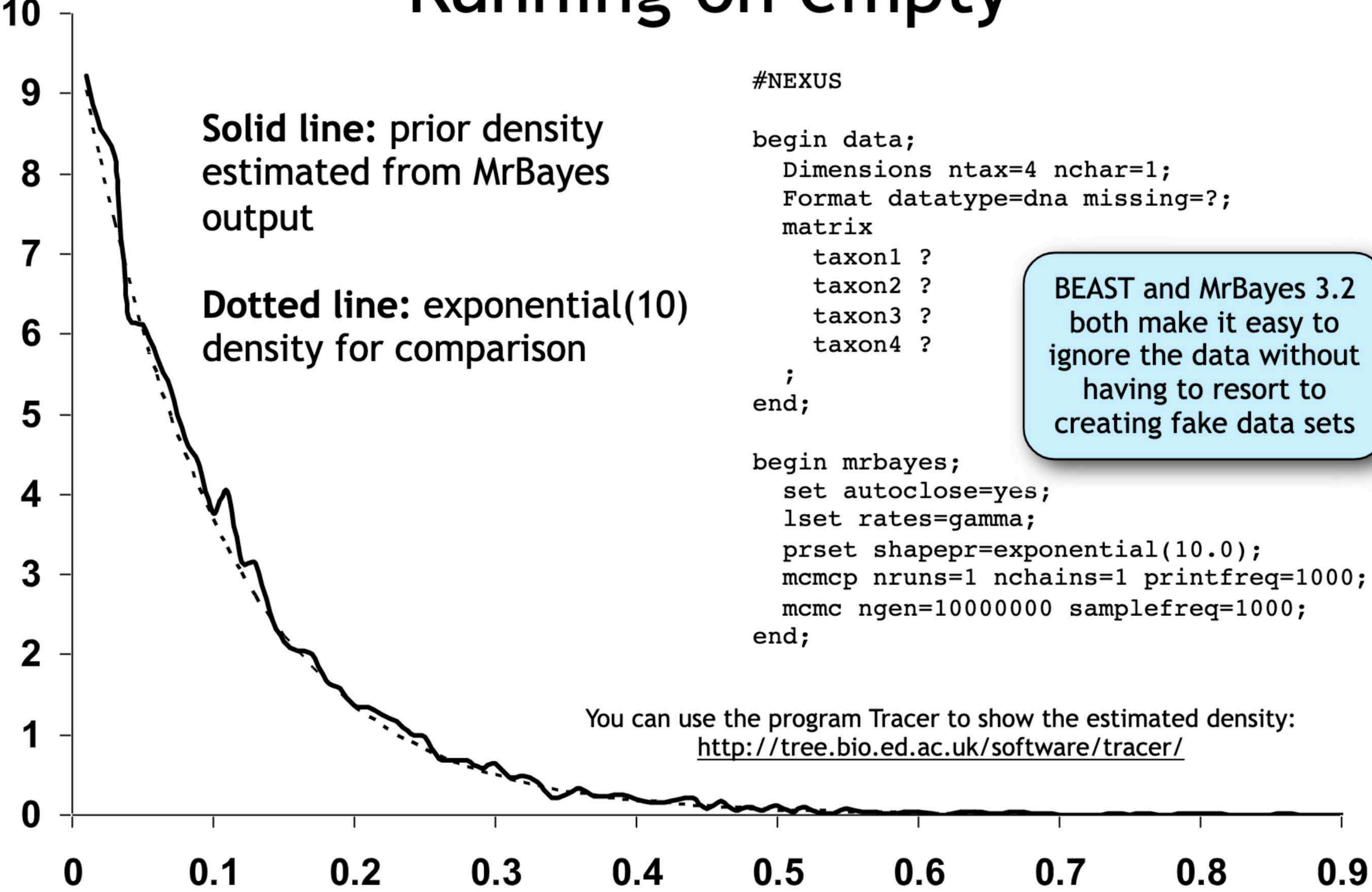
Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

Running on empty

- Run MCMC without data
- Some programs generate dummy “empty” alignments that can be used
- Why run MCMC on the prior?
    - Check correctness of the software (do the results match the theoretical prior?)
    - Compare prior run with posterior run:
        - If prior and posterior are too similar, then data might contain little information
        - If prior and posterior overlap, but the posterior is more concentrated, then data are informative and prior is reasonable
        - If prior and posterior do not overlap well, then the prior might be misspecified

- CAUTION: do not set a prior to match the posterior! The prior is meant to reflect our knowledge BEFORE the analysis of the data

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/

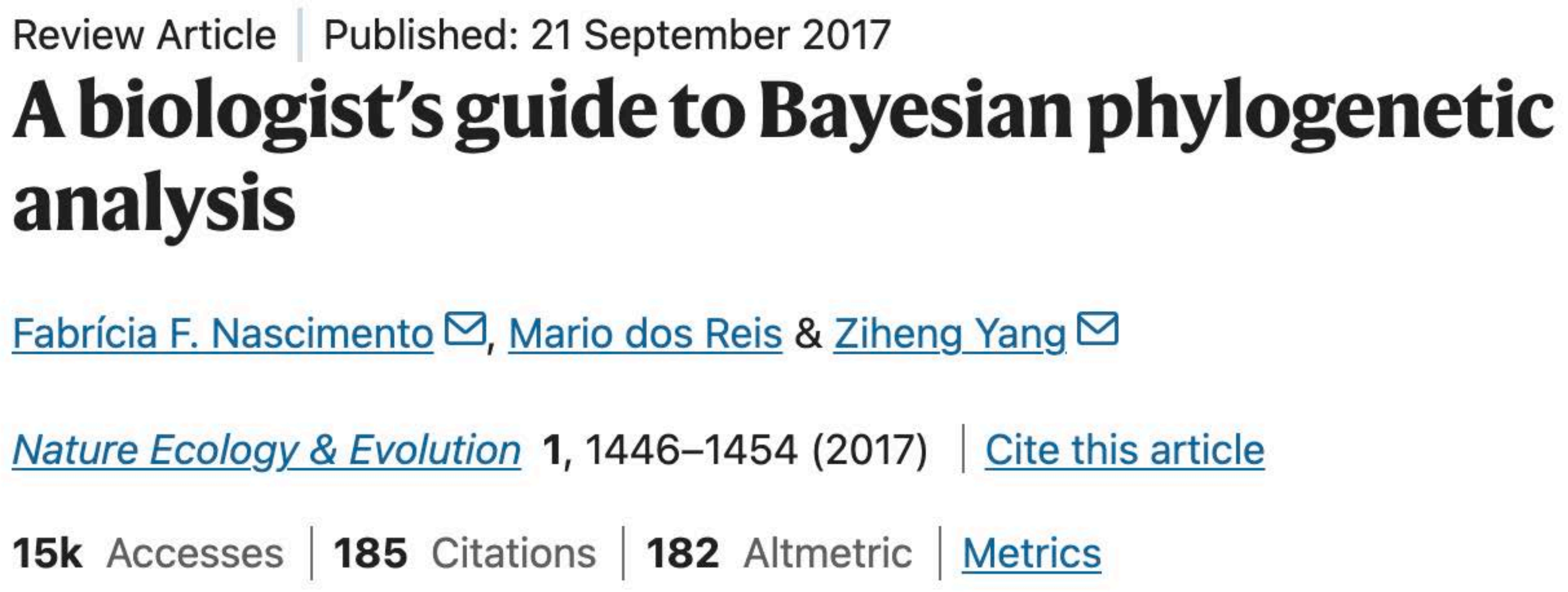
Paul O. Lewis (2017 Woods Hole Molecular Evolution Workshop)

JOURNAL ARTICLE

Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. FREE

Z Yang, B Rannala

Molecular Biology and Evolution, Volume 14, Issue 7, Jul 1997, Pages 717–724,
https://doi.org/10.1093/oxfordjournals.molbev.a025811
Published: 01 July 1997



Review Article | Published: 21 September 2017

A biologist's guide to Bayesian phylogenetic analysis

Fabrícia F. Nascimento, Mario dos Reis & Ziheng Yang

Nature Ecology & Evolution **1**, 1446–1454 (2017) | Cite this article

15k Accesses | **185** Citations | **182** Altmetric | Metrics