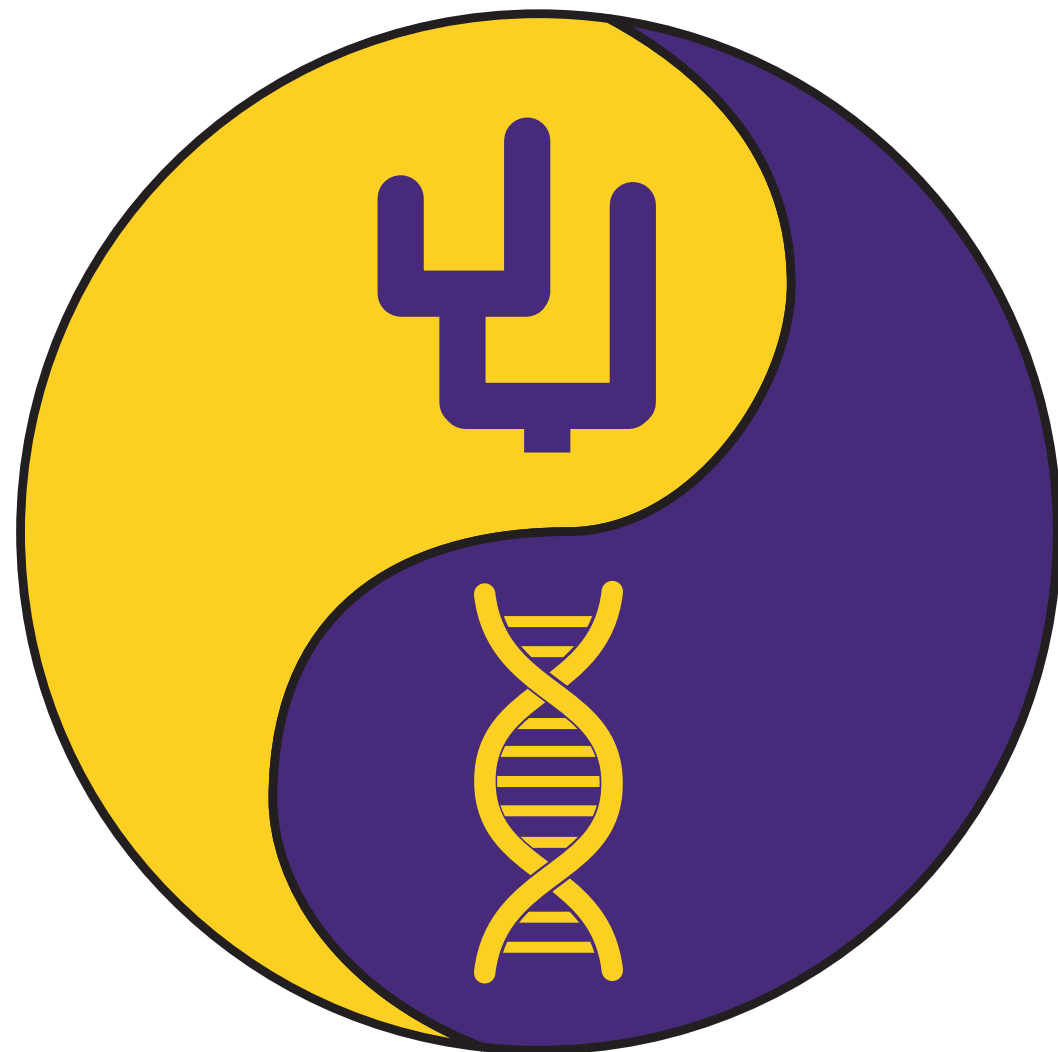# Model Complexity and Big Data
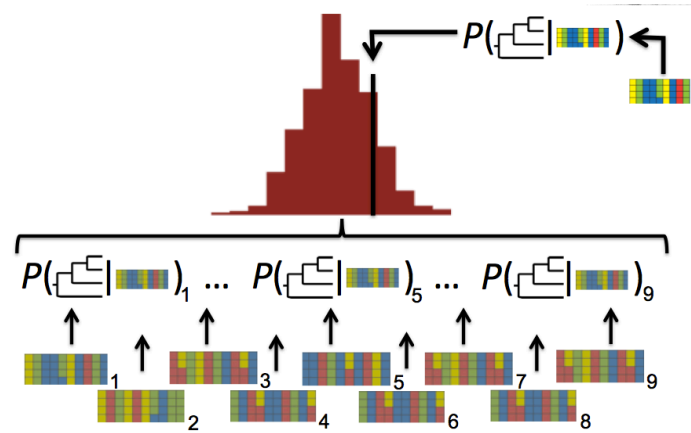
Jeremy M. Brown @jembrown
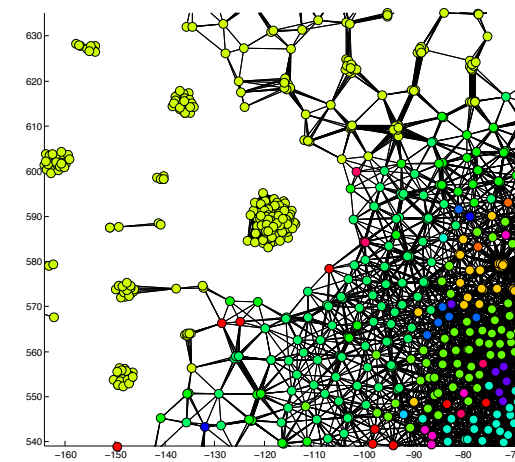
Robert C. Thomson @rcthomson
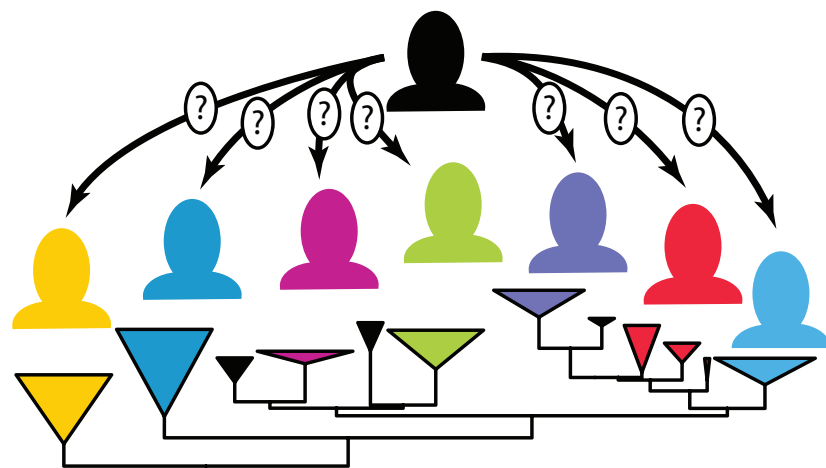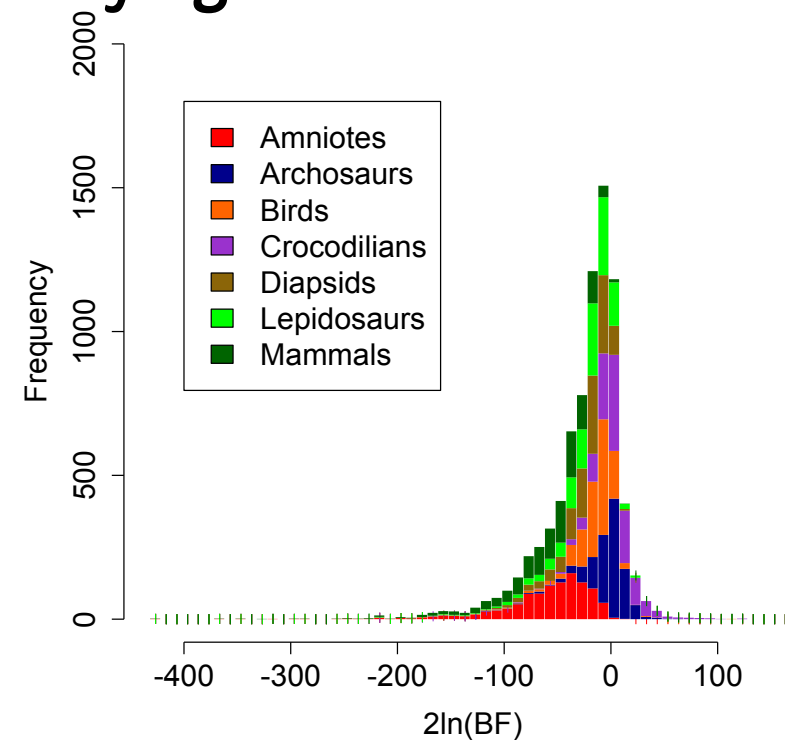
Some Interests of Mine

Evaluating Model Fit

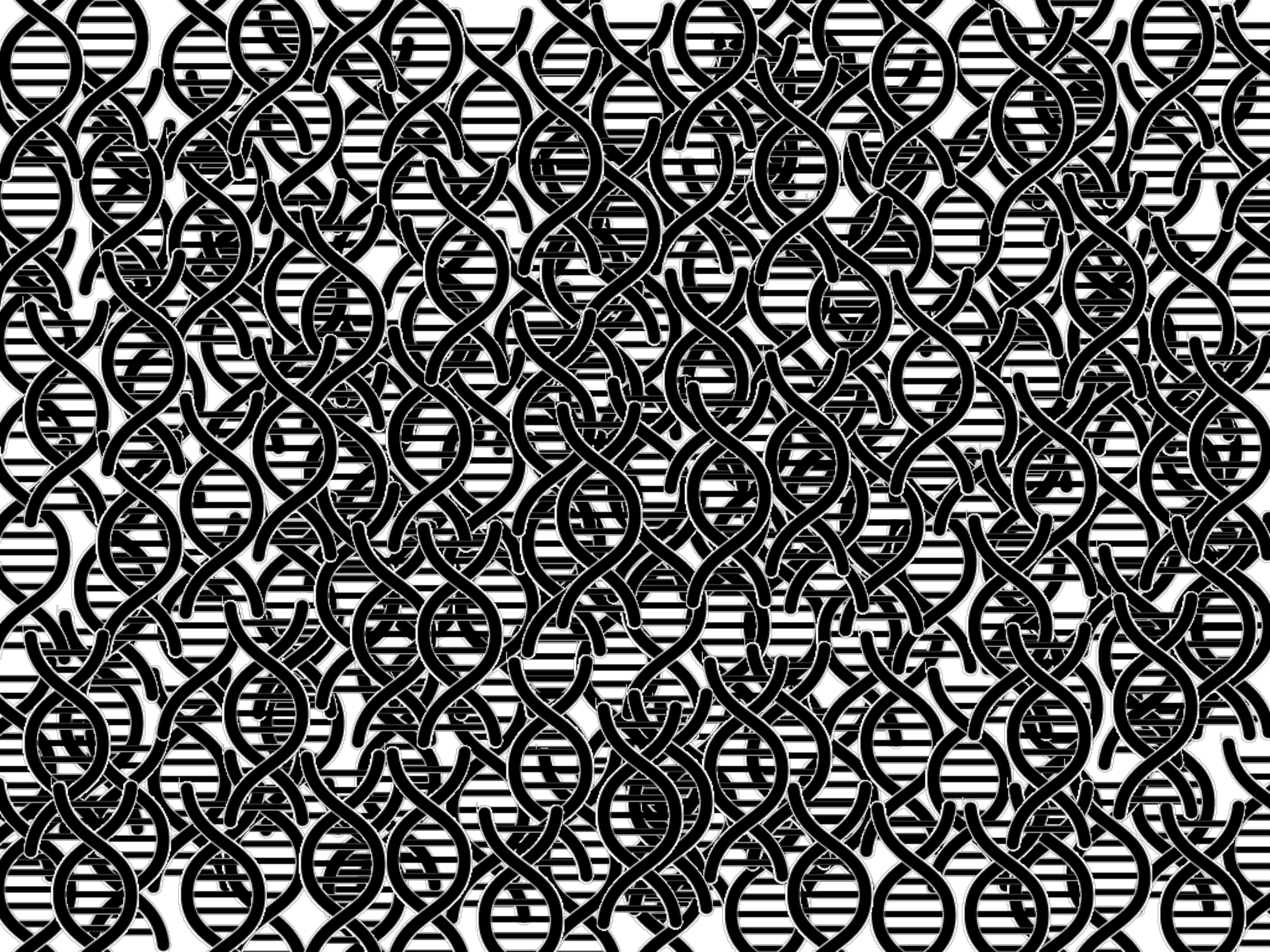## Visualizing and Exploring Tree Sets

Viral molecular evolution, epidemiology, and forensics

Phylogenetic Information
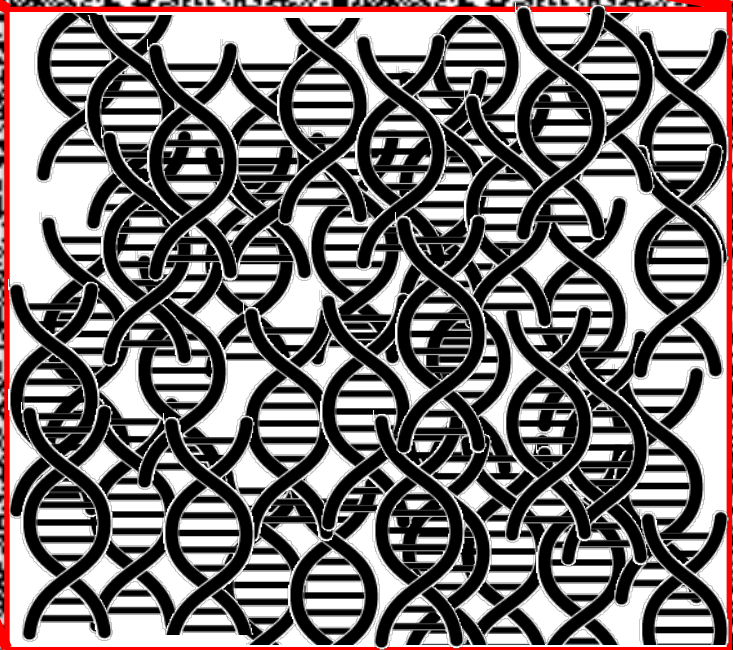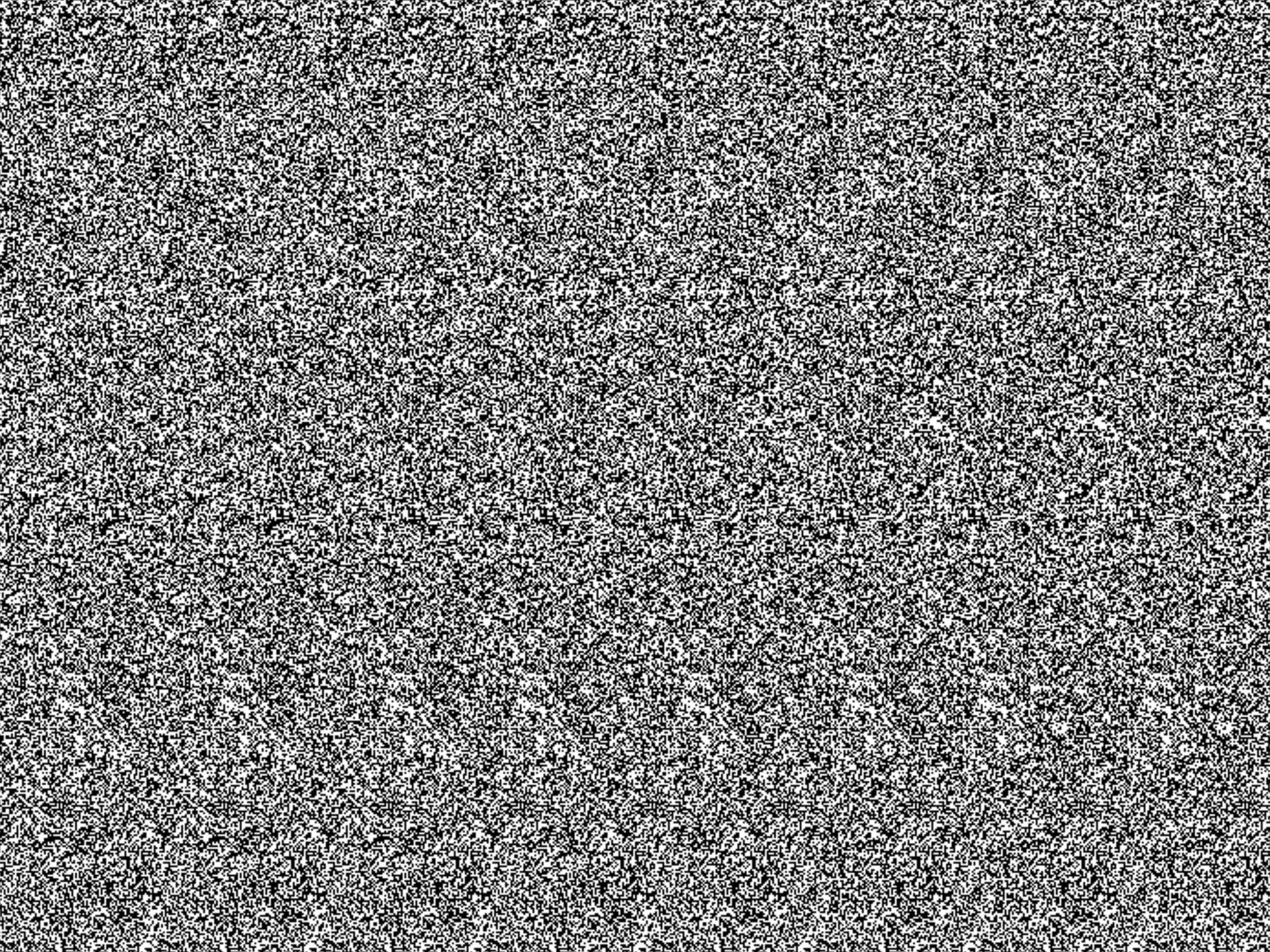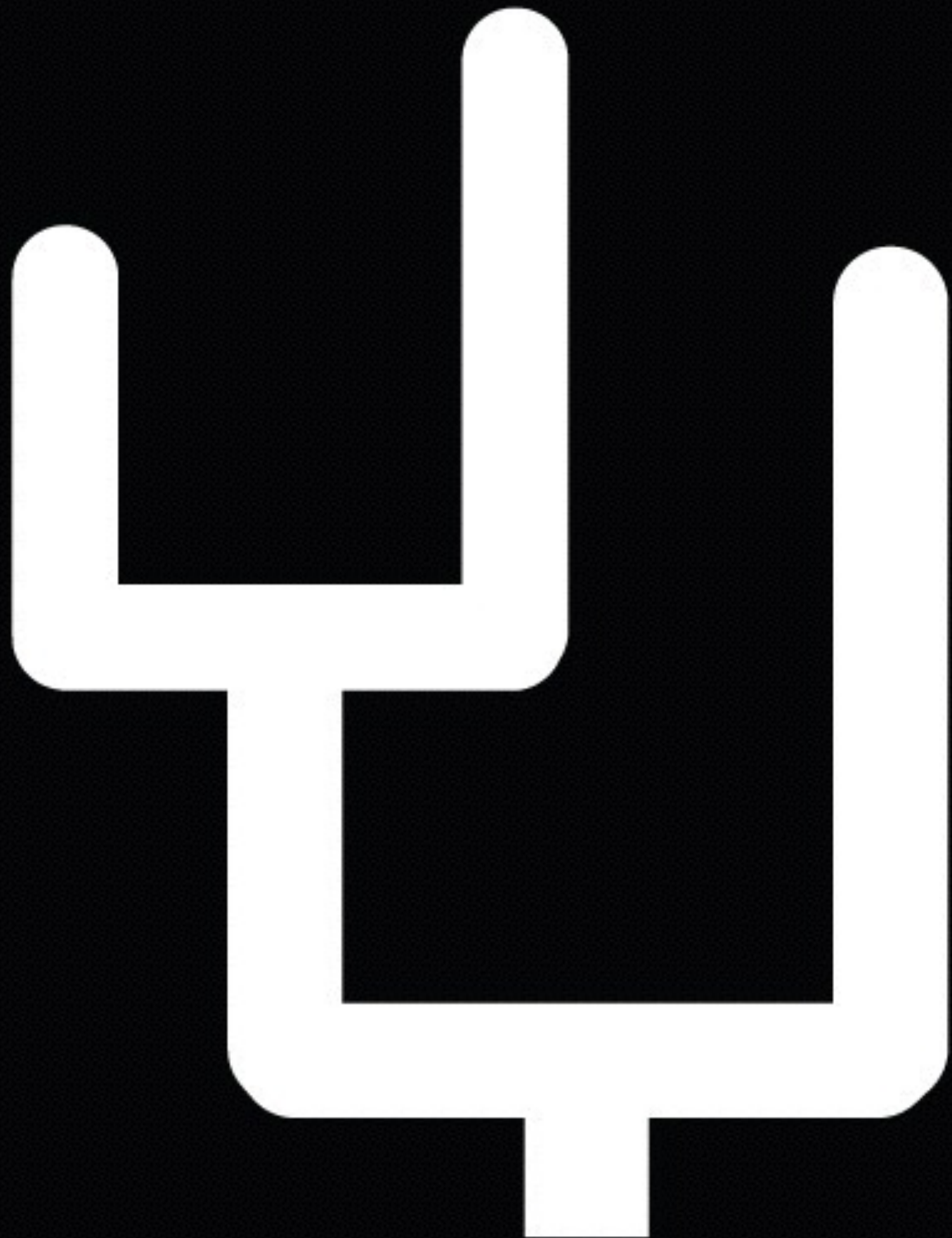
So...genomes, eh?

- GTR+I+Γ seems pretty complicated!

- 10 parameters to describe change in 4 nucleotides

- Surely that's enough.
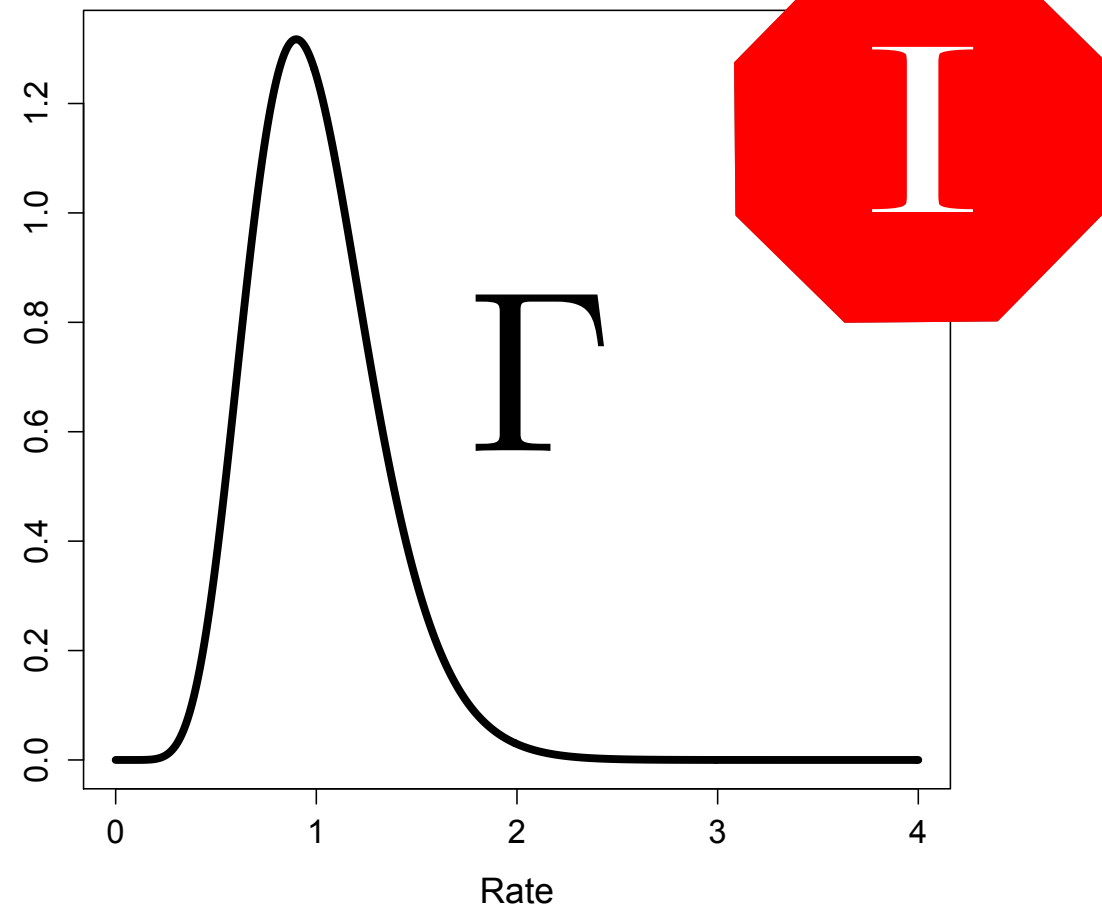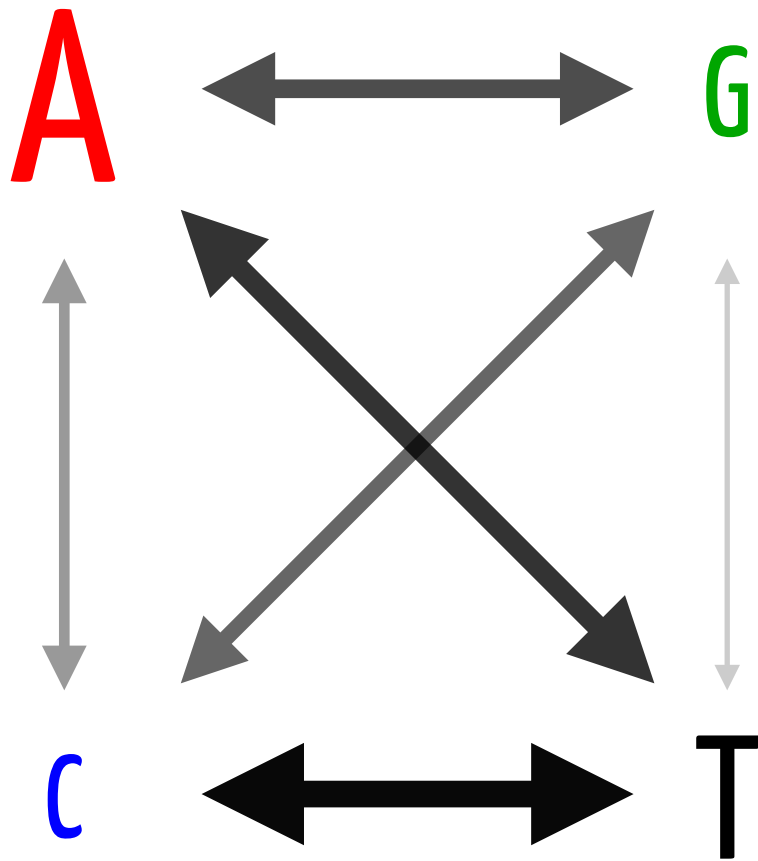
Challenge 1: Genes Vary in Rate
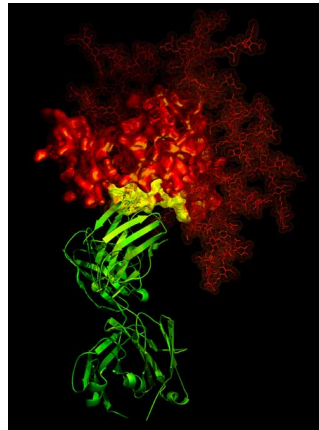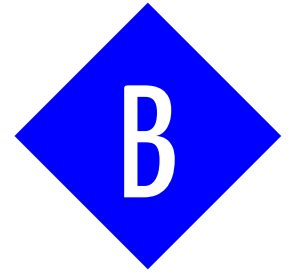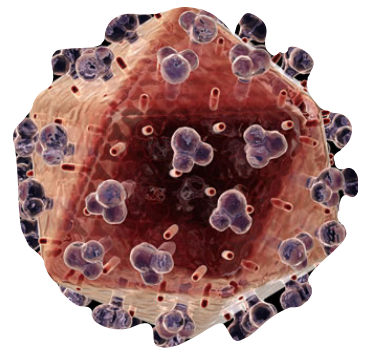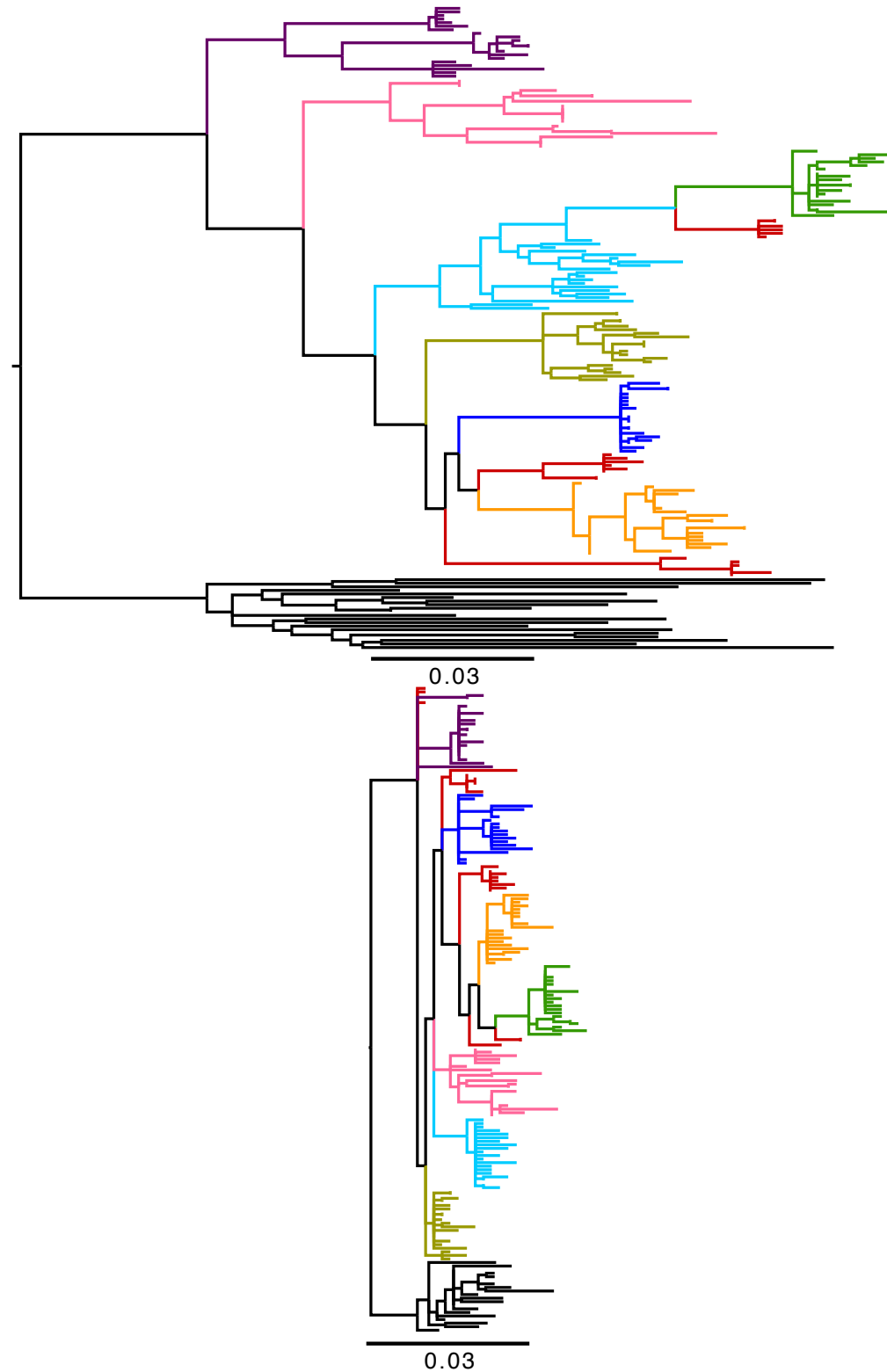
Challenge 2: Genes Vary in Model/Parameters
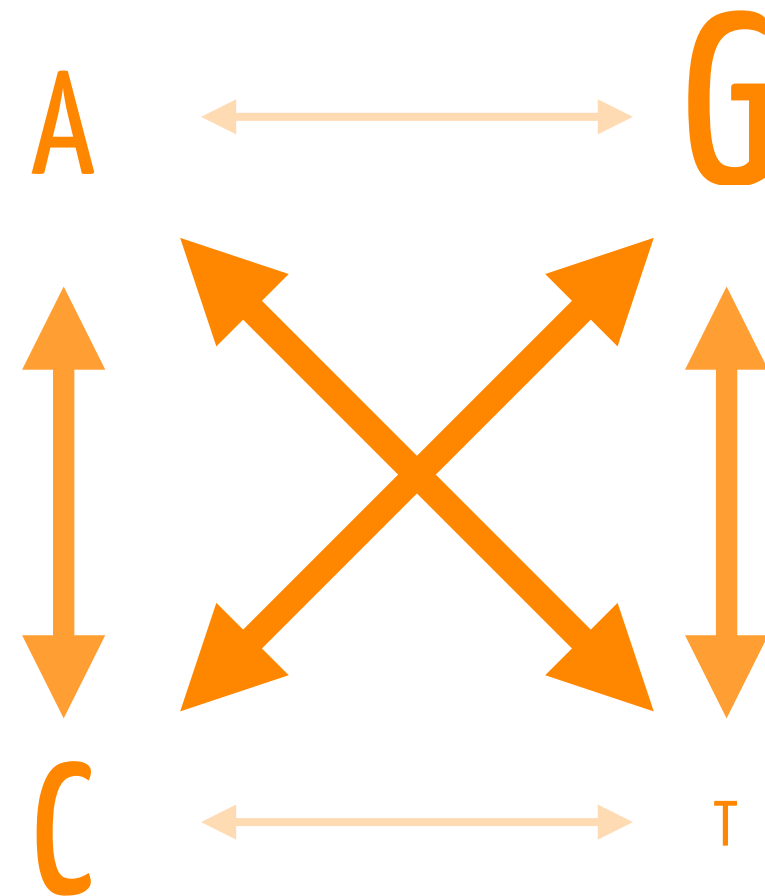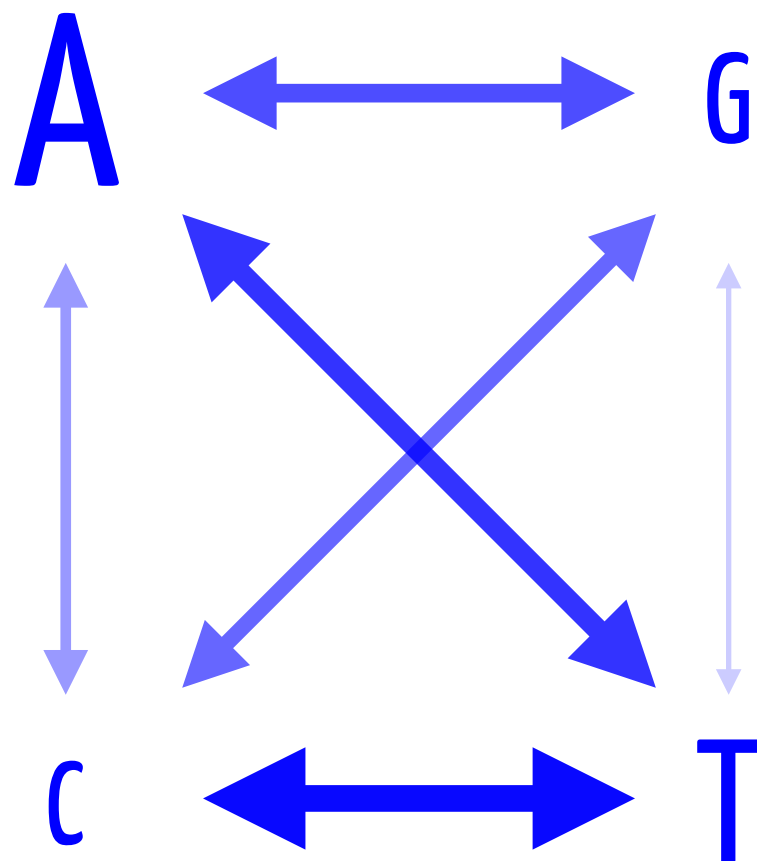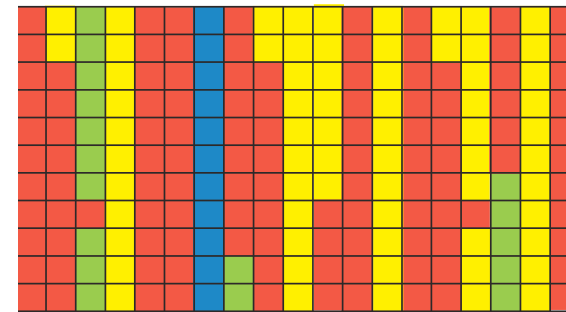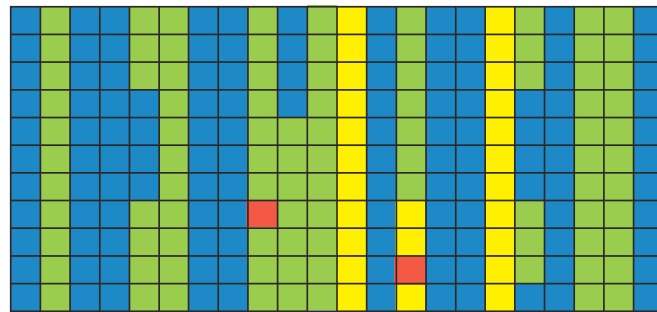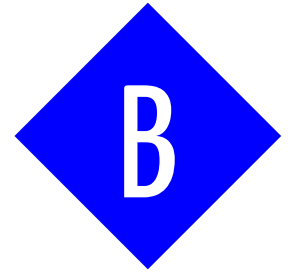
Challenge 3: Genes Vary in Topology ◆ B
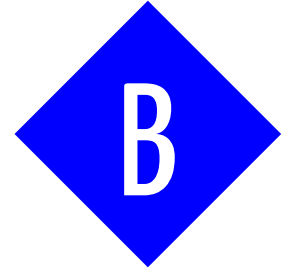
- Incomplete Lineage Sorting

- Hybridization

- Horizontal Gene Transfer

Challenge 4: Variation in Gene-Model Fit M

- Genes and models should fit together like a hand in a glove. A glove abstracts a hand, but in a useful way.

- When fit is poor, the glove may not function properly.

Challenge 4: Variation in Gene-Model Fit ◆ M

- Nearly all of our models (or at least the ones we usually consider) still assume a lot of things:

  - Independence of sites

  - Constant site rates across the tree

  - Constant base frequencies across the tree

  - Consistent evolutionary dynamics across the tree

Challenge 5: Non-homology of sites and genes M

- An alignment is a statement of homology.

- We are saying that we are **certain** that nucleotides in a column have a common ancestor that diverged due to a speciation event (usually).

- This is commonly violated in at least two circumstances:

  - Alignments can be uncertain

  - Paralogy (can exert undue influence)

Types of Variation Across Genes

B
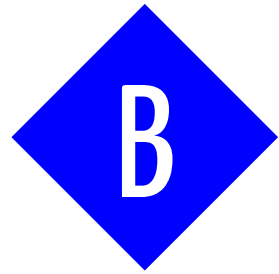
Rate

Topology

Model Parameters
(evolutionary dynamics)

M

Absolute Model Fit

Incorrect Homology

So...how do we deal with this variation?

- We develop elegant models that relax these assumptions!

- Now we do 3 things with our models:

  - Select the best available model (model selection)

So...how do we deal with this variation?

- We develop elegant models that relax these assumptions!

- Now we do 3 things with our models:

  - Select the best available model (model selection)

  - Critically evaluate the fit of this model (model adequacy)

So...how do we deal with this variation?

- We develop elegant models that relax these assumptions!

- Now we do 3 things with our models:

  - Select the best available model (model selection)

  - Critically evaluate the fit of this model (model adequacy)

  - Accept, refine, or reject (the art)

Model Selection

The Fundamental Tradeoff
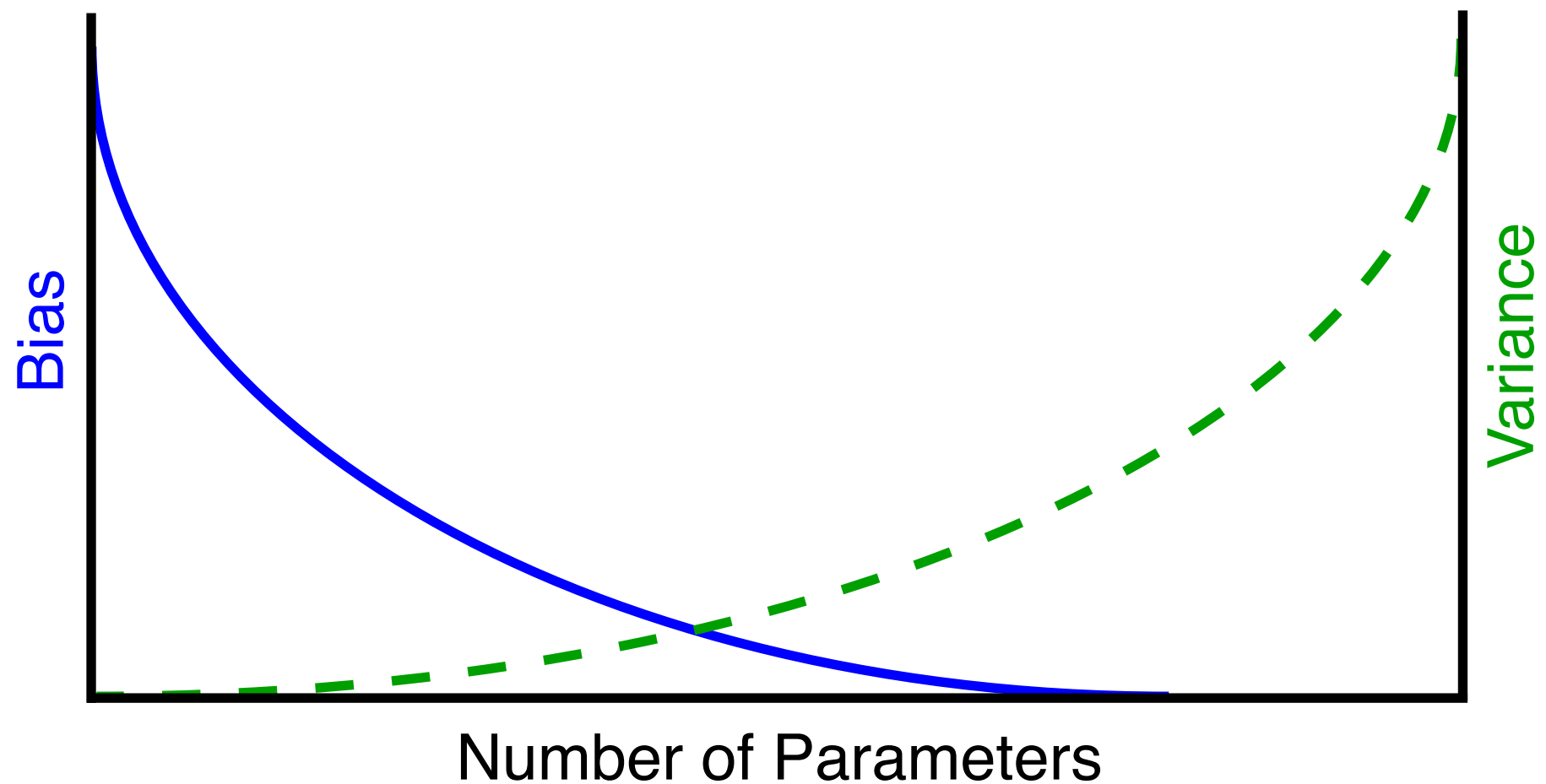
Model Selection

Model too simple!
We're misinterpreting the data.

Model Selection
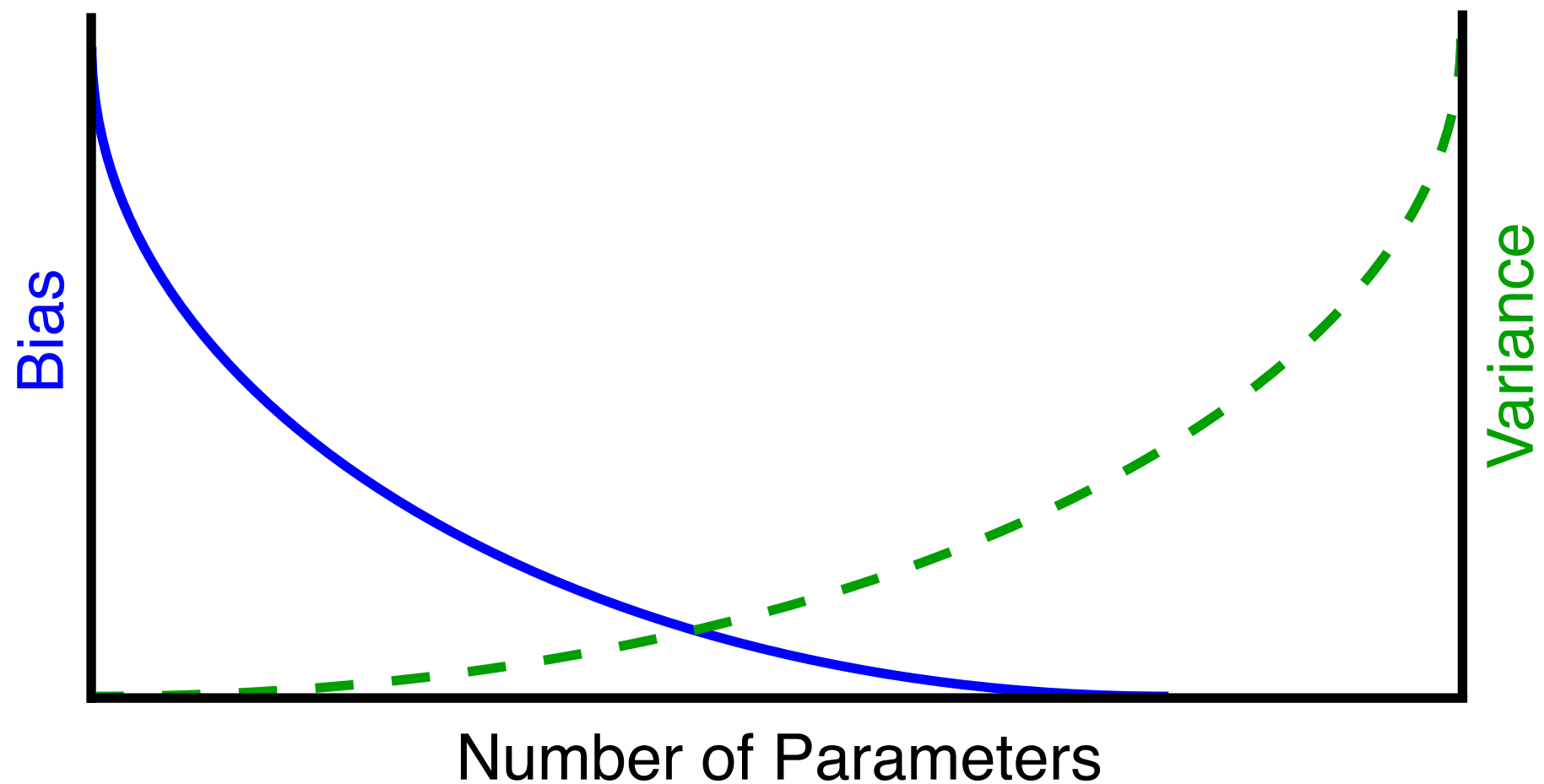
Model Selection
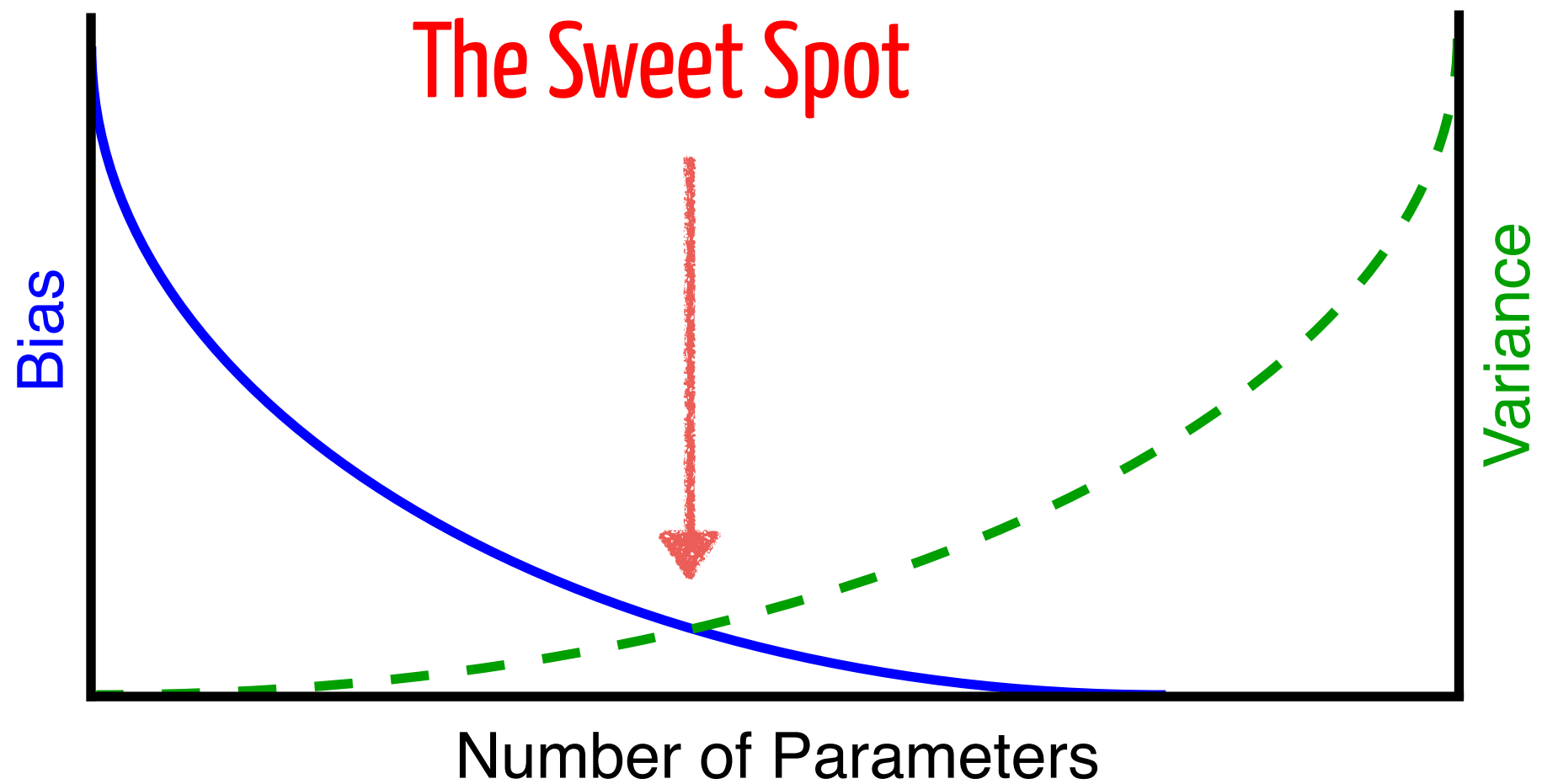
Model Selection

Bias and **Variance** can be traded off in different ways.

This leads to **multiple criteria** for model selection.
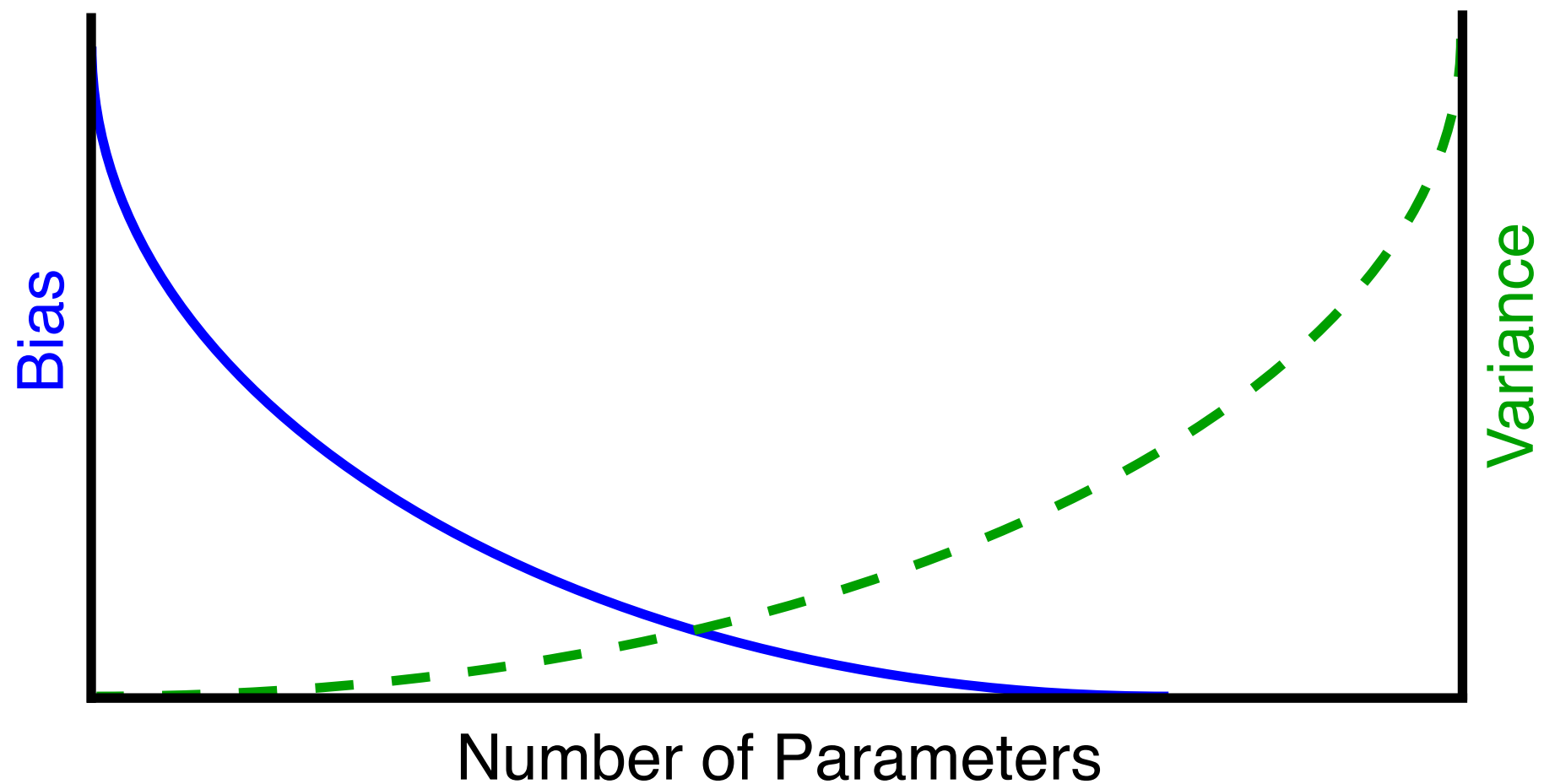
The Likelihood
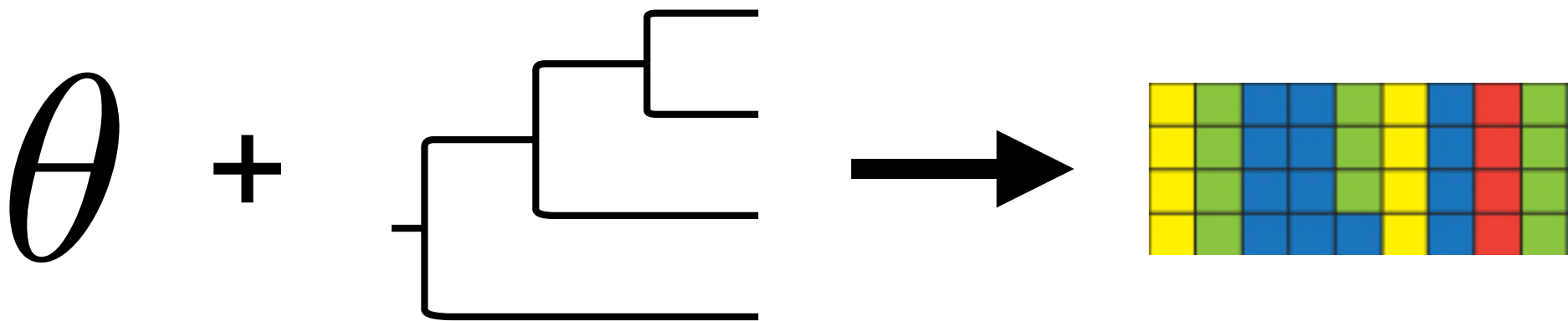
The Likelihood
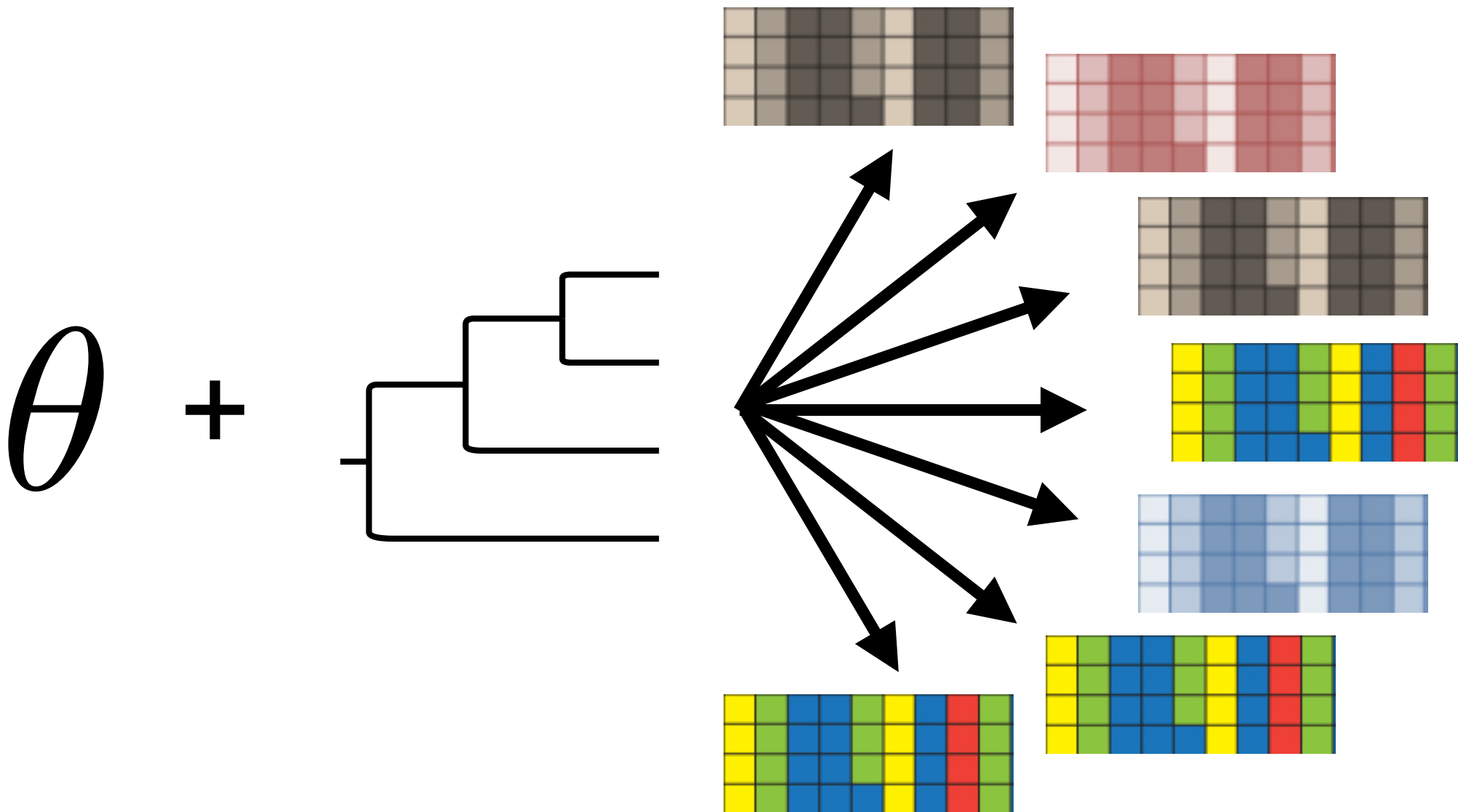
The Likelihood Function

$$P(\text{ } | \theta , \text{ })$$

Read as "**the probability of the sequence data given a tree and model**".

The quantity by which the data provide information.

Compares how well different trees and models predict the observed data or as a "**measure of relative surprise**".

Maximum Likelihood
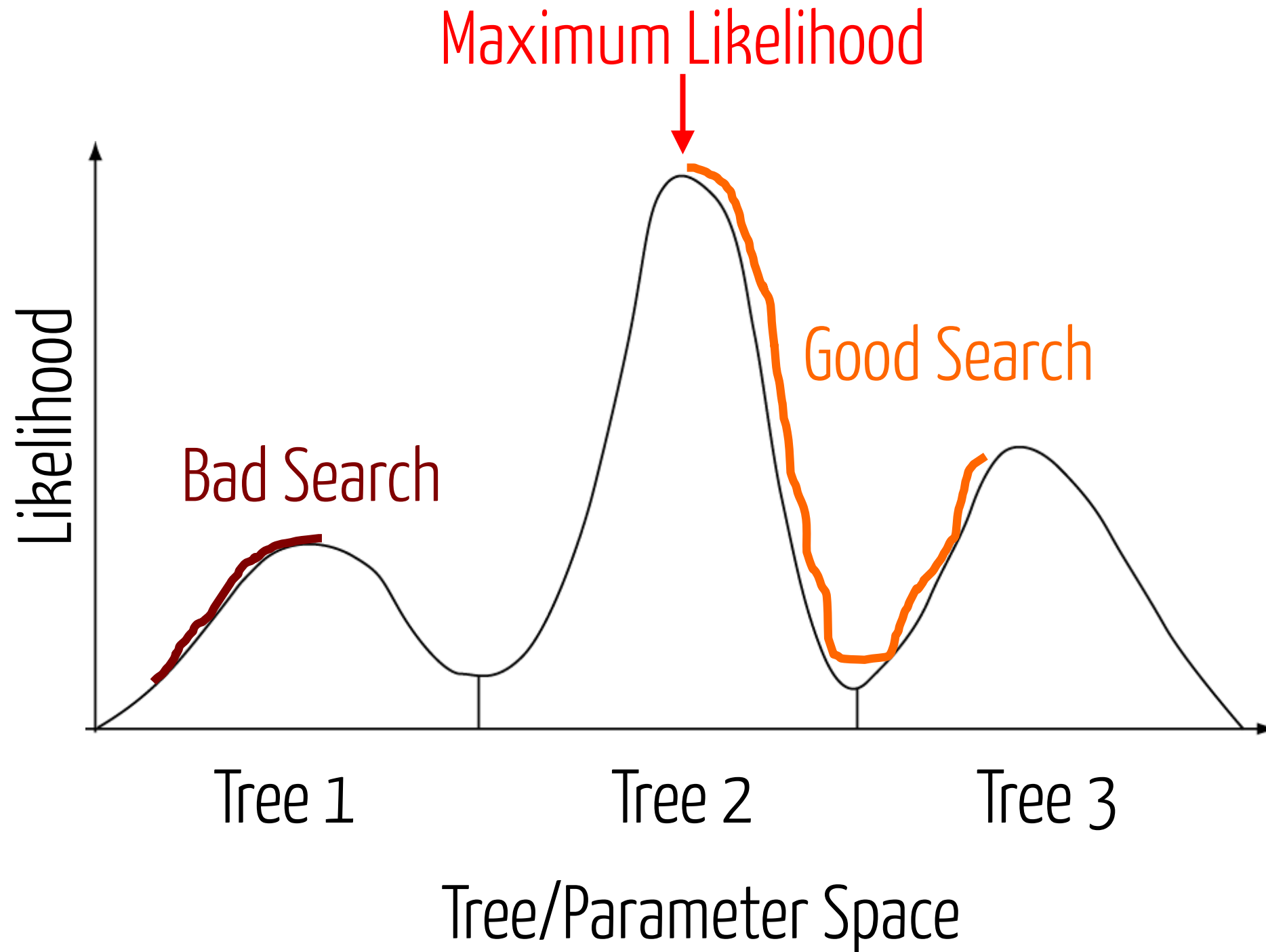
More Parameters = Better Likelihood
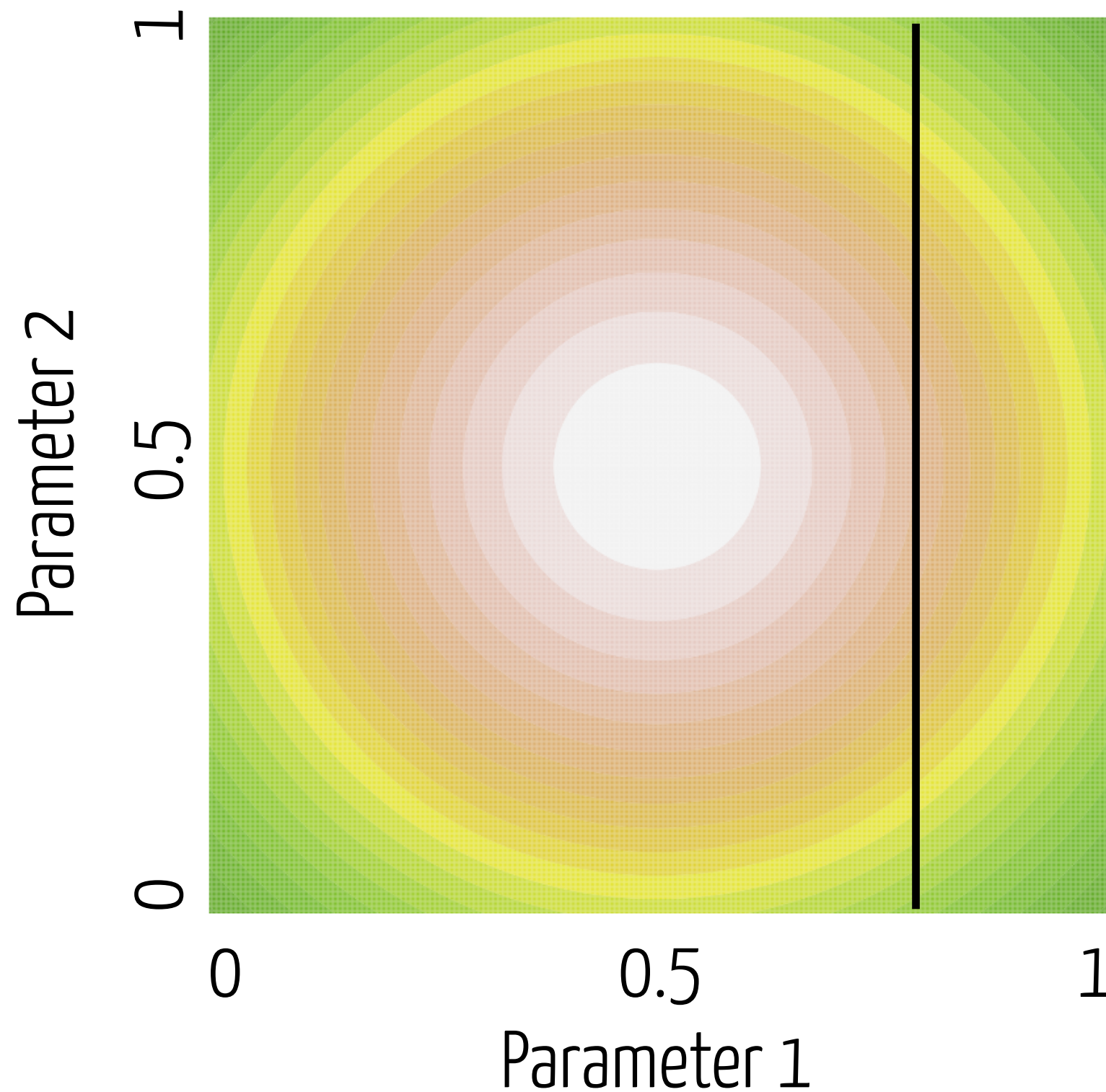
More Parameters = Better Likelihood
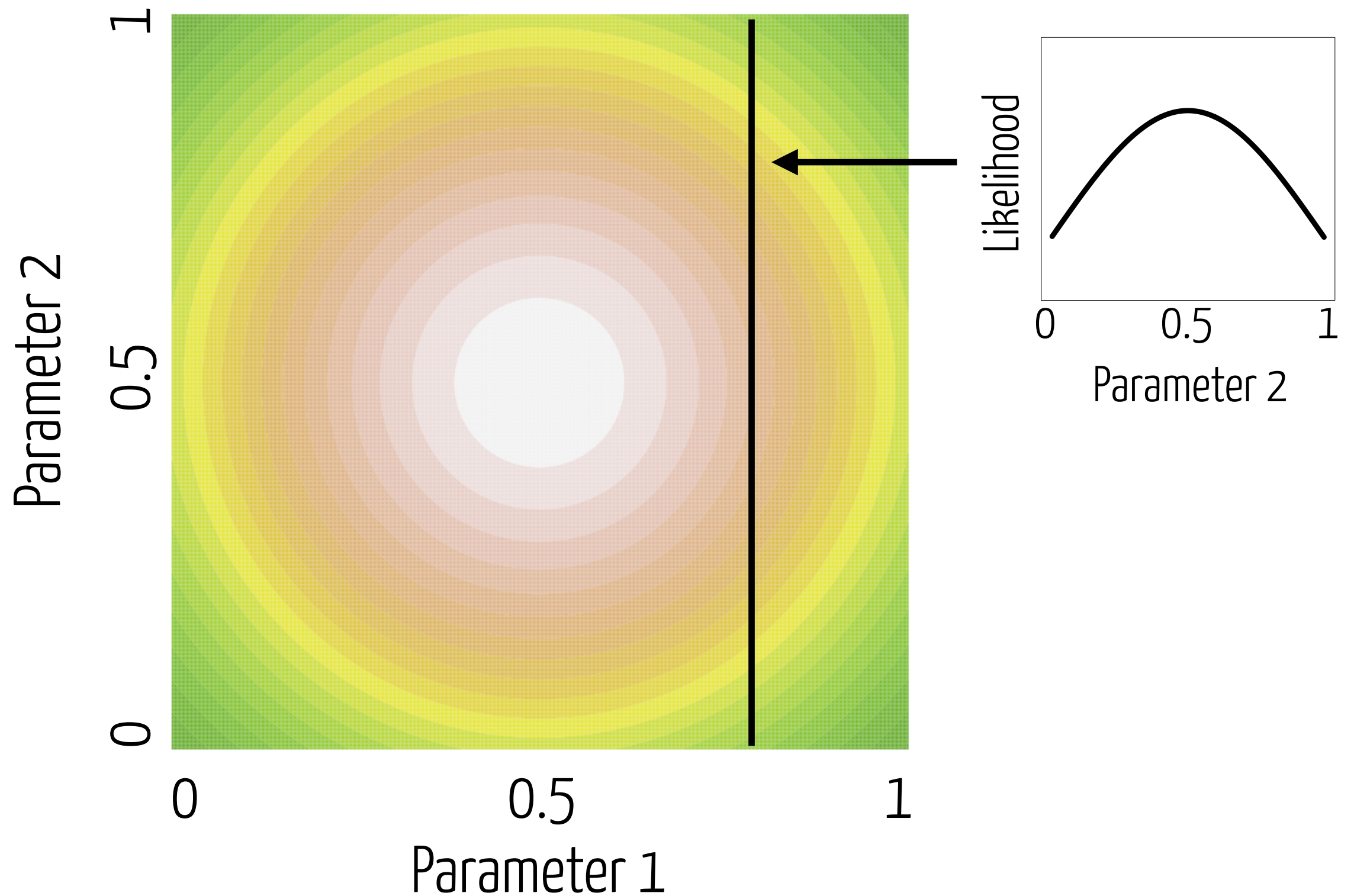
More Parameters = Better Likelihood
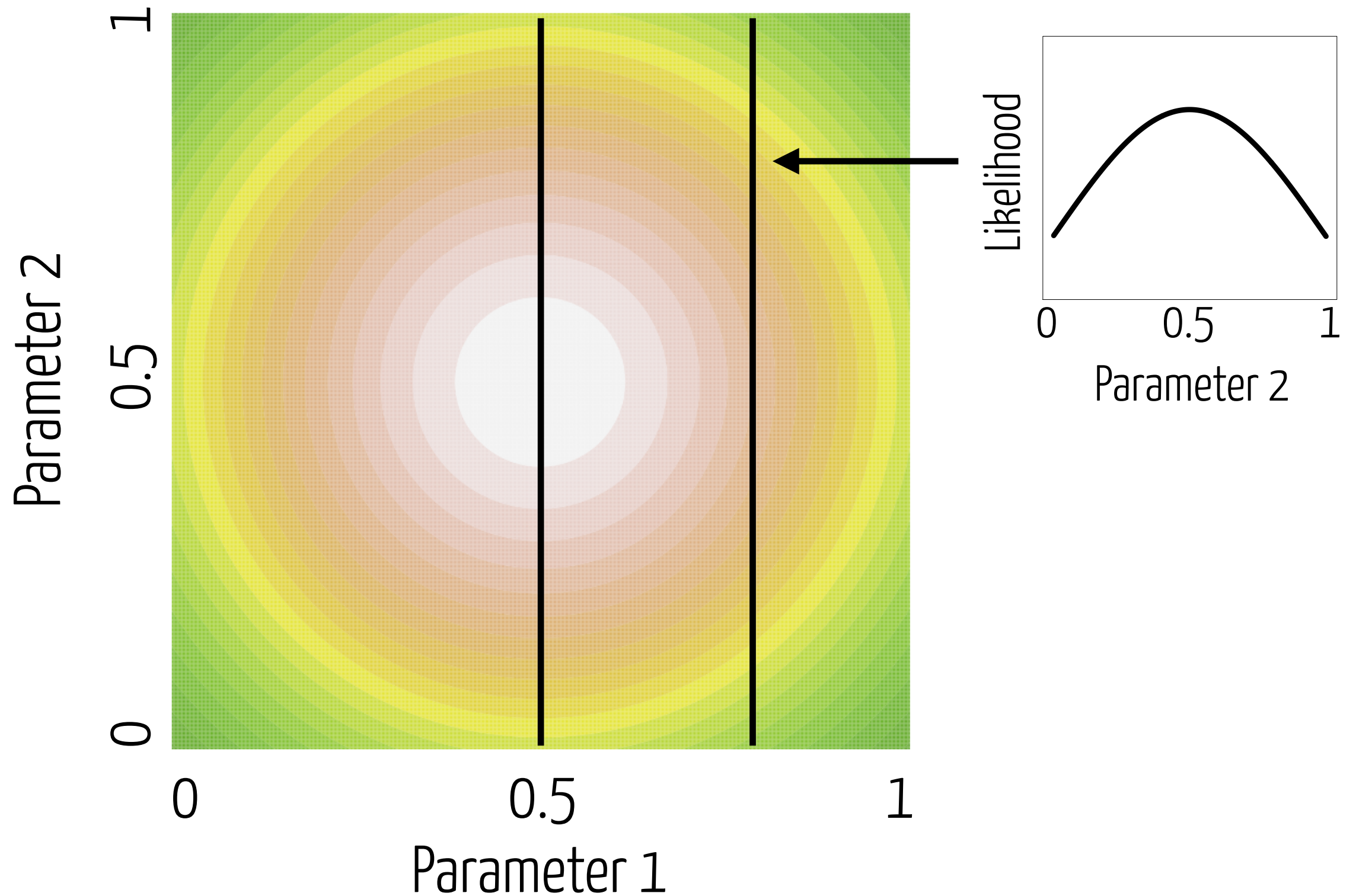
More Parameters = Better Likelihood
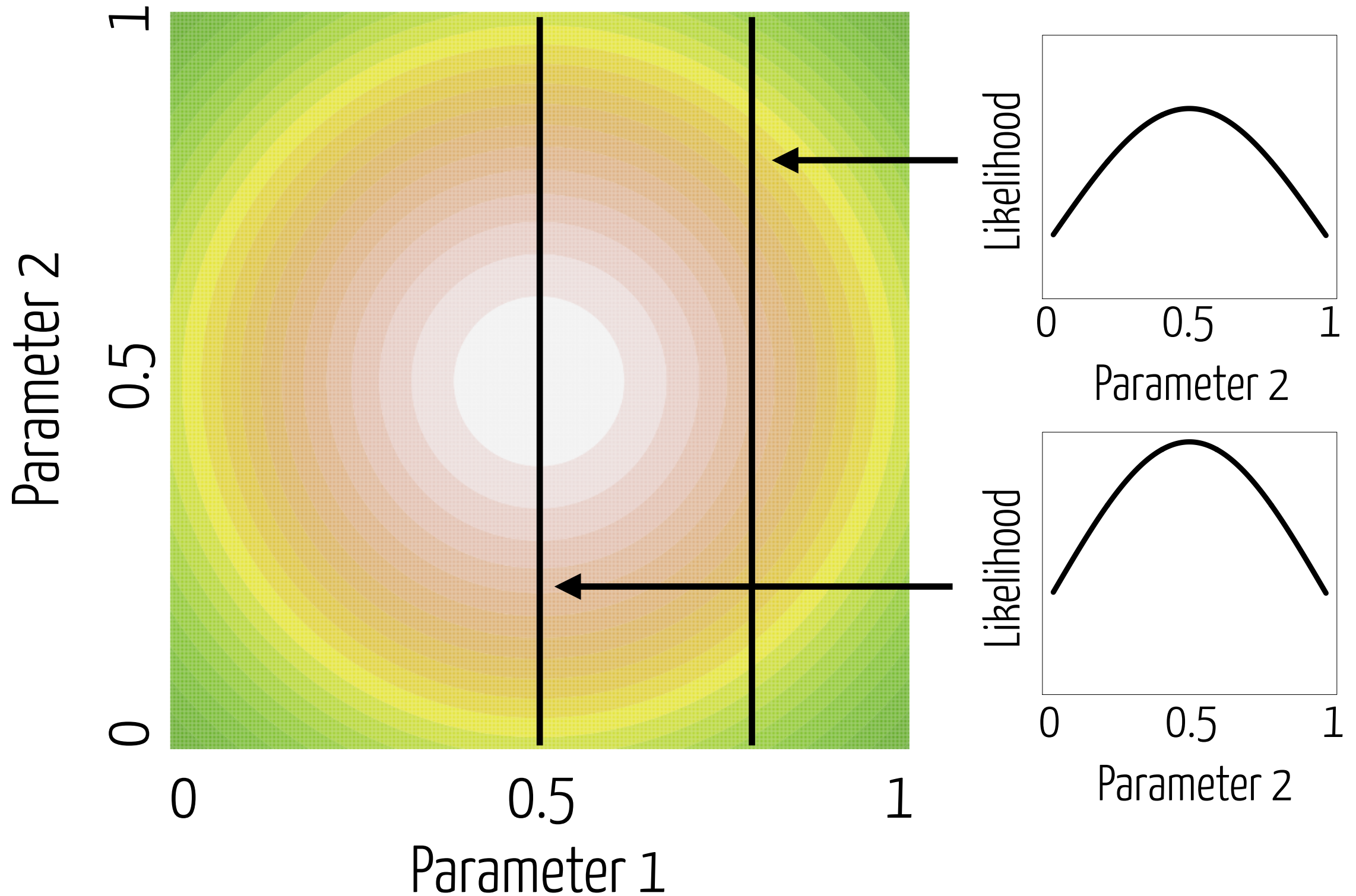
ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.

ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.

- ◦ Akaike's Information Criterion (**AIC**)

- ◦ Bayesian Information Criterion (**BIC**)
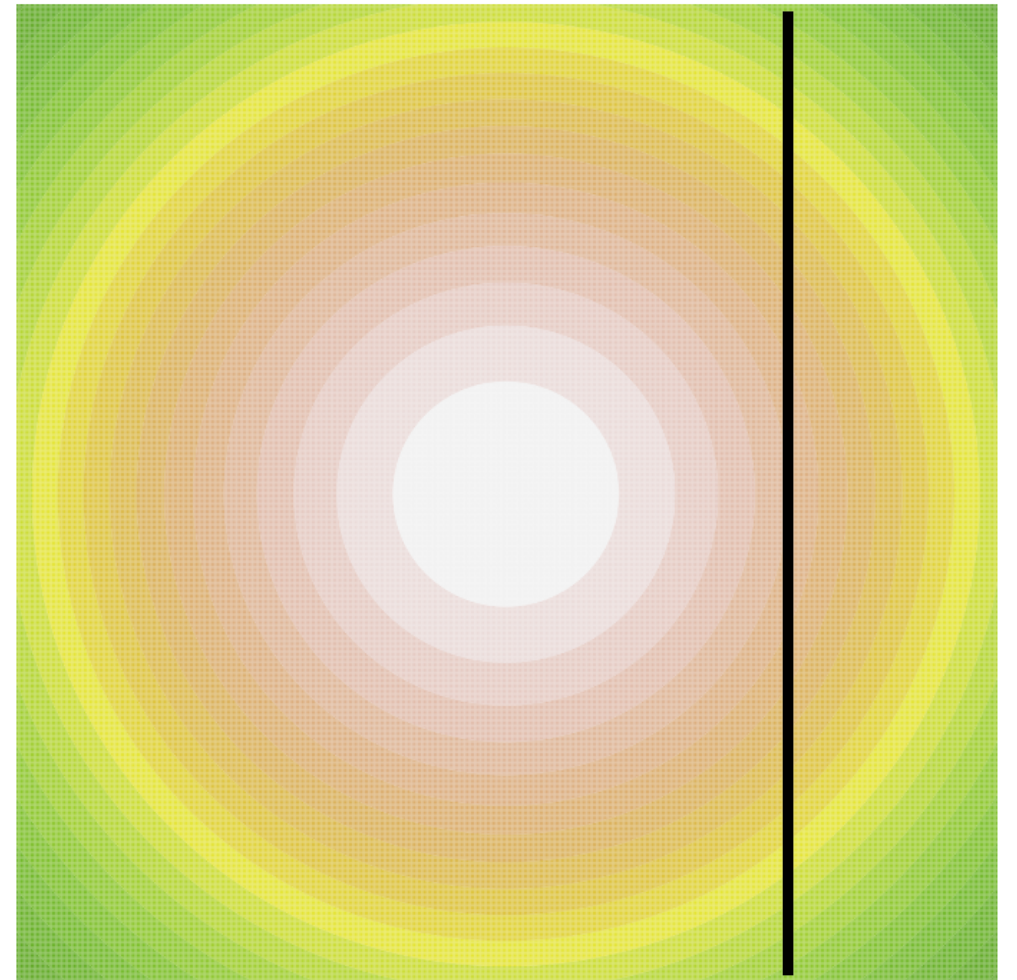
- ◦ Likelihood Ratio Test (**LRT**)

ML-based Model Selection

If the **more complex** model always gives a likelihood that is **at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.

- Akaike's Information Criterion (**AIC**)

- Bayesian Information Criterion (**BIC**)
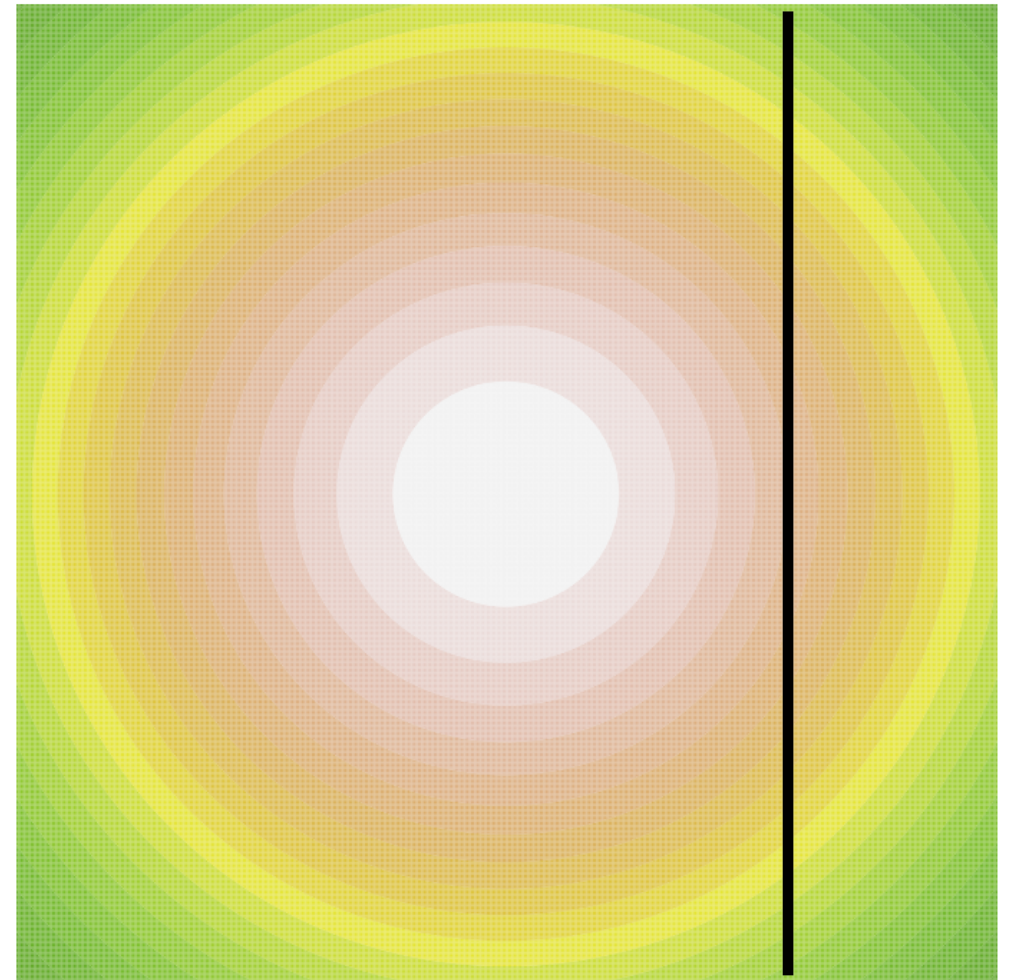
- Likelihood Ratio Test (**LRT**)
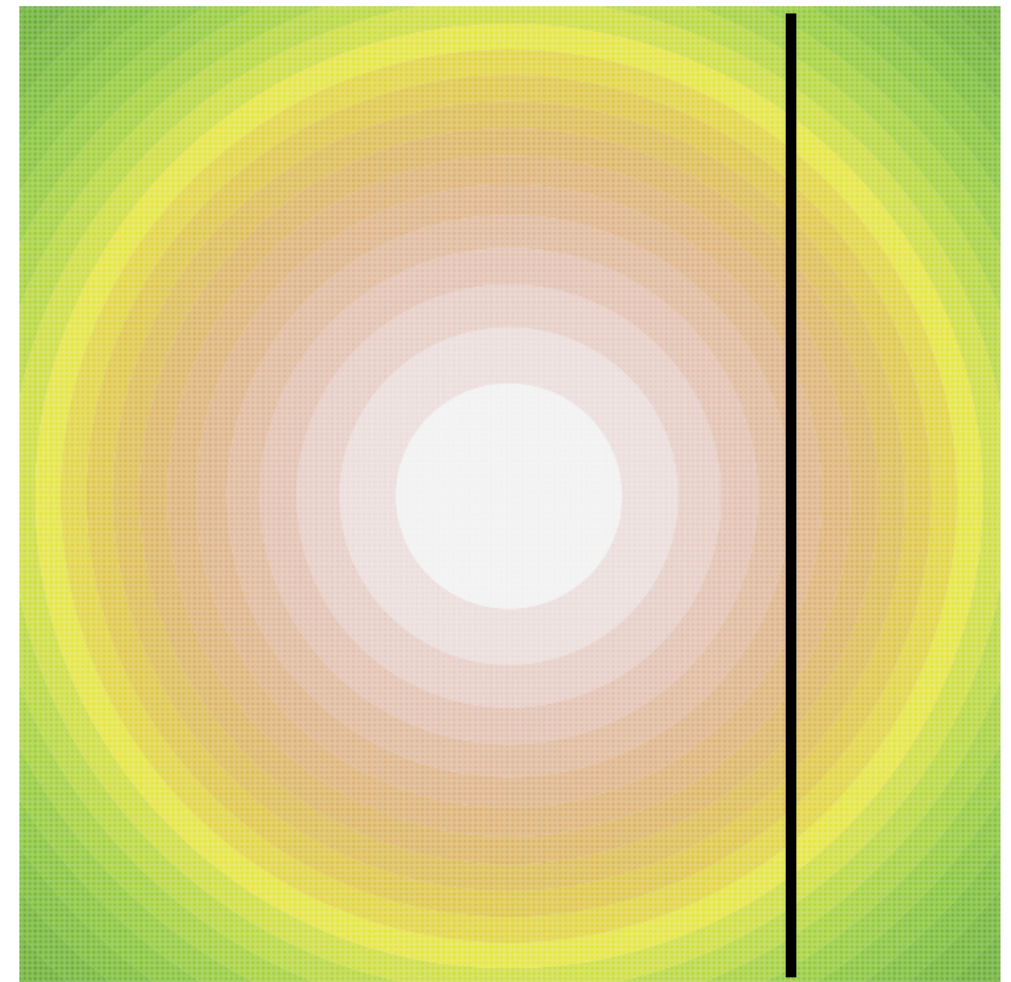
Different penalties for extra parameters.
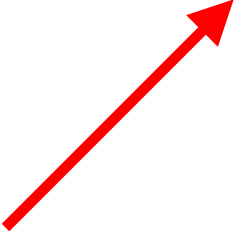
ML-based Model Selection

Akaike's Information Criterion (**AIC**)

Minimum AIC preferred.

$$\text{AIC} = \textcolor{red}{2k} - \textcolor{blue}{2ln(\hat{L})}$$

Penalty for more parameters (k).

Likelihood term becomes more negative when $\hat{L}$ worse.

ML-based Model Selection

$$\text{AIC} = \color{red}{2k} \color{black}{-} \color{blue}{2ln(\hat{L})}$$

$$\text{BIC} = \color{red}{ln(n)k} \color{black}{-} \color{blue}{2ln(\hat{L})}$$

LRT

Hypothesis test

Pairwise

Difference in free parameters
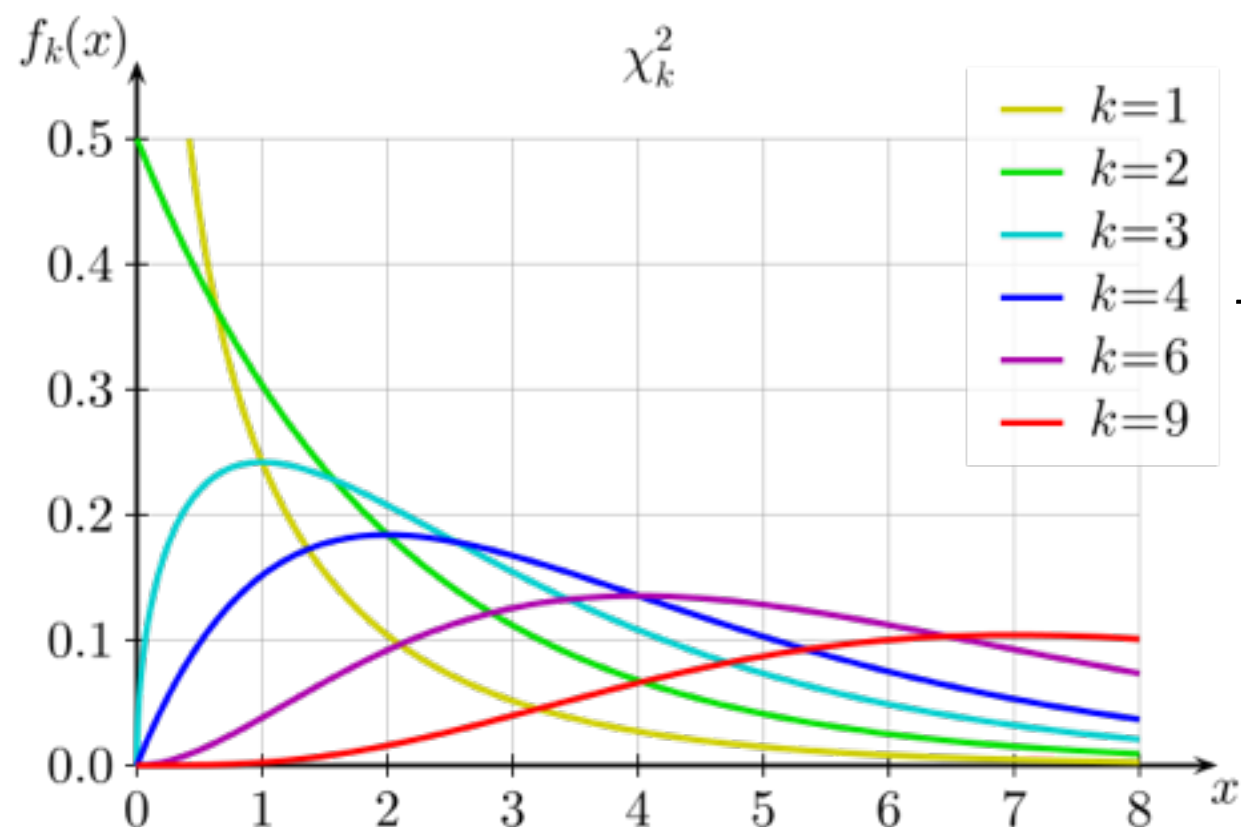
Bayesian Model Selection

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

Bayesian Model Selection

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

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}$$

Bayesian Model Selection

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

$$P(\theta|D,M) = \frac{P(D|\theta,M)P(\theta|M)}{\int P(D|\theta,M)P(\theta|M)d\theta}$$

Marginal Likelihood

Probability of the data given the model, considering uncertainty in model parameters.

Bayesian Model Selection

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

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{\int P(D|\theta, M)P(\theta|M)d\theta}$$

Marginal Likelihood

Essentially, the **weighted average likelihood**, weighted by the prior probability of different parameter values.

Marginal Likelihood Example

Evolutionary Distance

Sp. A ———————————— Sp. B

Compare **JC** and **K80** models

v: edge length
estimated in both models

k: transition-transversion ratio
estimated in K80 and fixed at 1 for JC

Marginal Likelihood Example

Simulation Conditions
Sequence length: 500 bp
True v: 0.15
True k: 5.0

Prior is flat over whole area.

K80 (whole 2D plane)

JC (just 1D line)

k

v

10

5

0

0.0

0.1

0.2

0.3

Marginal Likelihood Example

Simulation Conditions
Sequence length: 500 bp
True v: 0.15
True k: **5.0**

Prior is flat over whole area.

K80 (whole 2D plane)

JC (just 1D line)

k

v

K80 wins!

Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

Simulation Conditions
Sequence length: 500 bp
True v: 0.15
True k: **1.0**

Prior is flat over whole area.

K80 (whole 2D plane)

JC (just 1D line)

Marginal Likelihood Example

Simulation Conditions
Sequence length: 500 bp
True v: 0.15
True k: **1.0**

Prior is flat over whole area.

K80 (whole 2D plane)

JC (just 1D line)

k

v

10

5

0

0.0

0.1

0.2

0.3
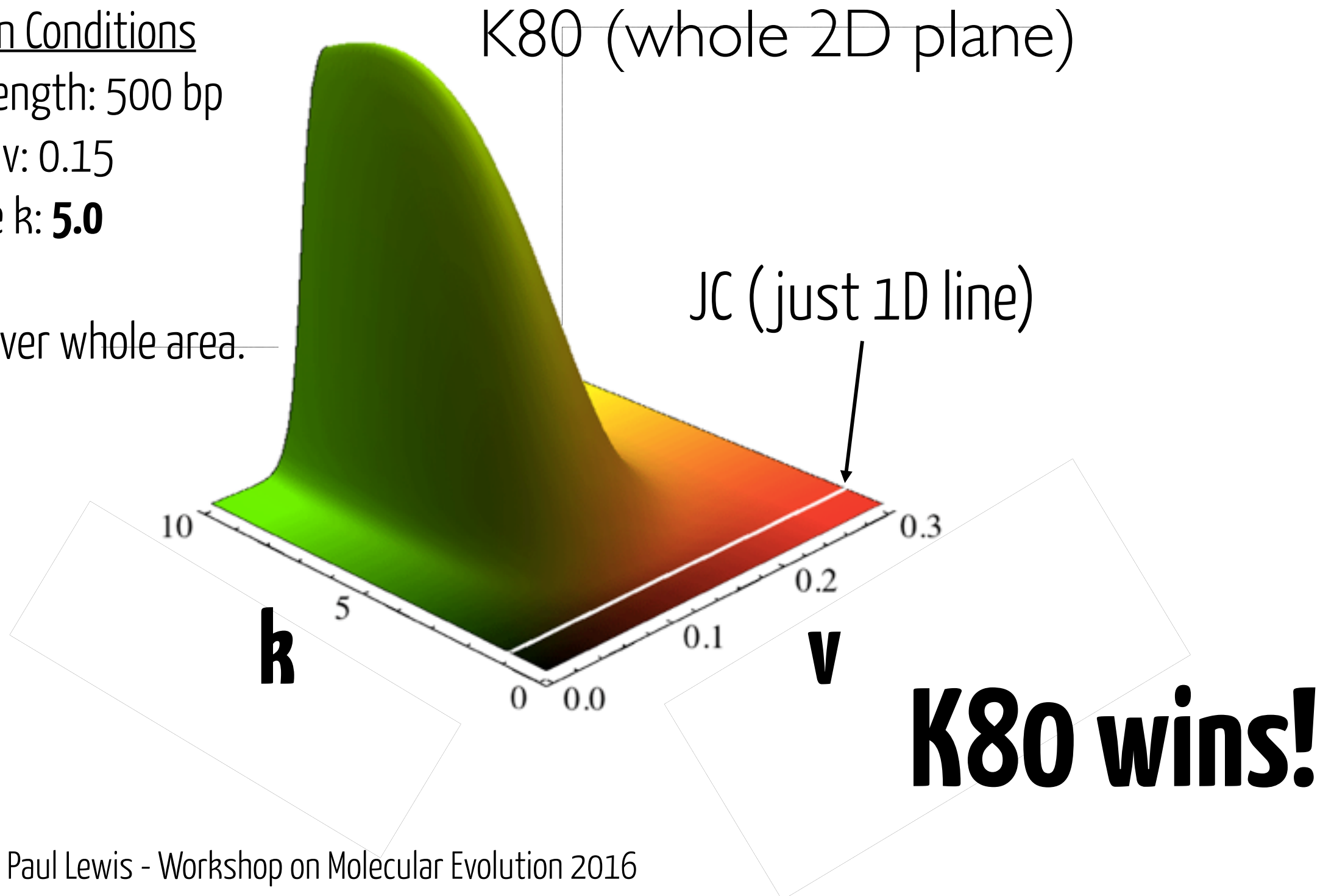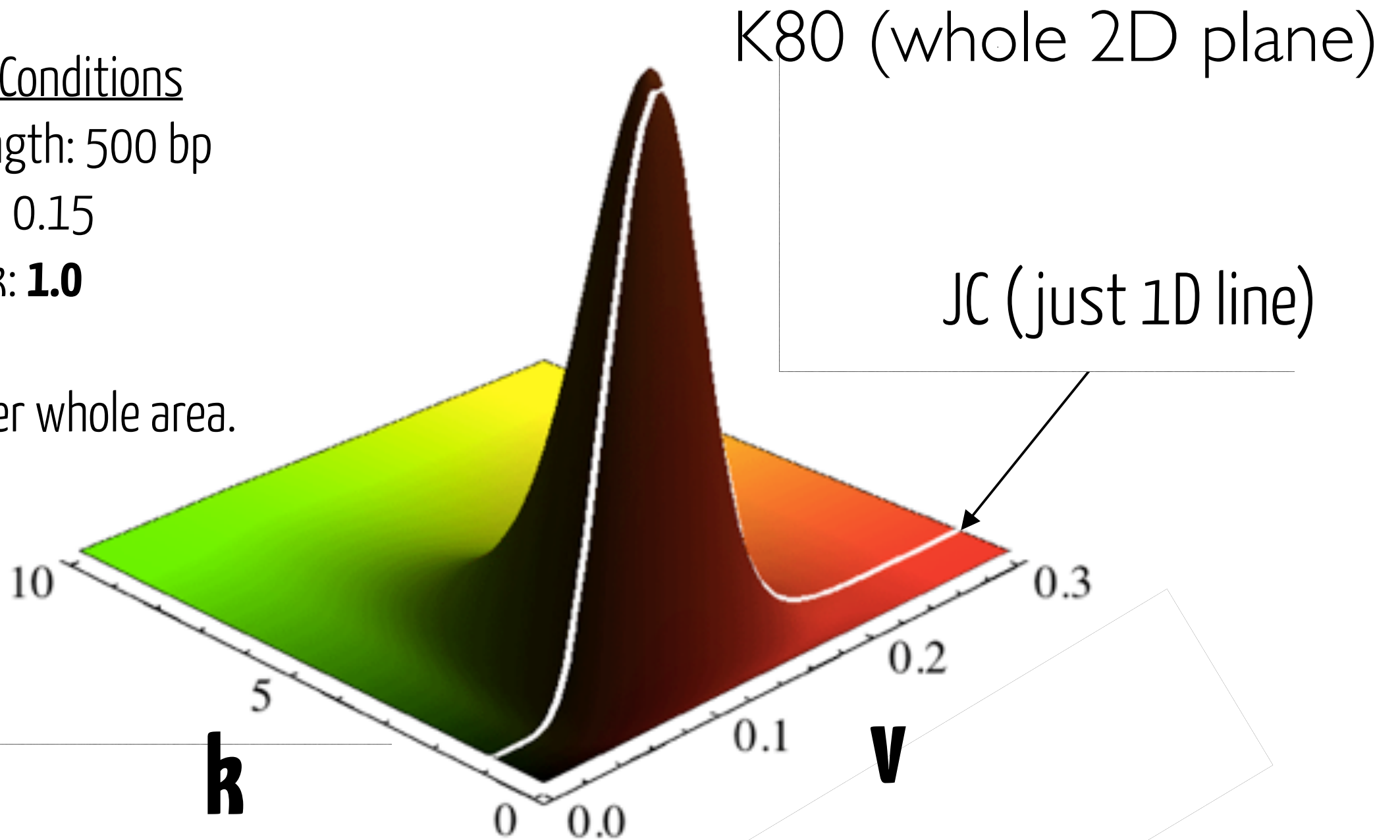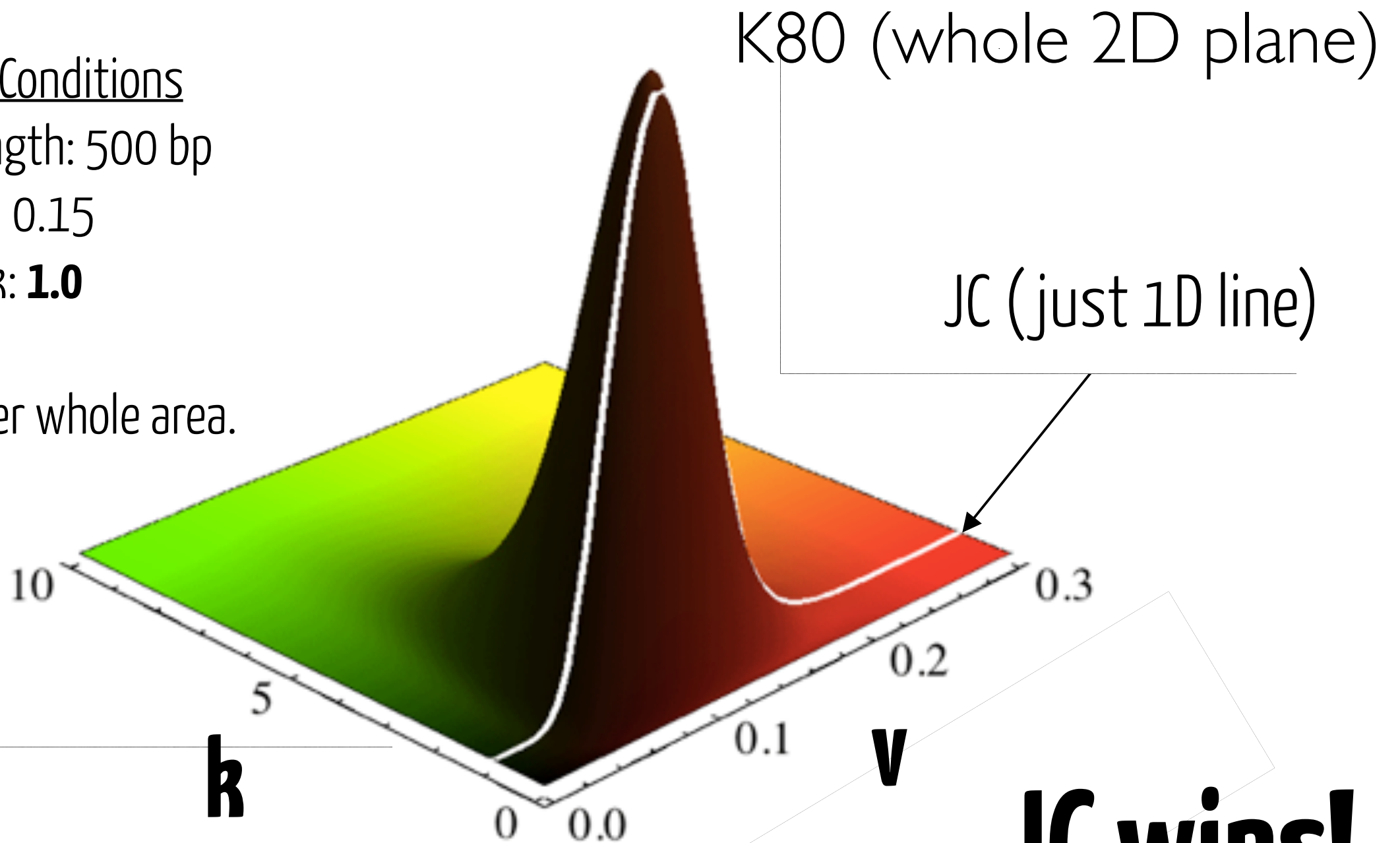
JC wins!

Marginal Likelihood Example

Important contrast with ML-based model selection: by averaging, rather than maximizing, marginal likelihoods automatically account for extra parameters.

More complicated models can have lower marginal likelihoods.

Calculating Marginal Likelihoods
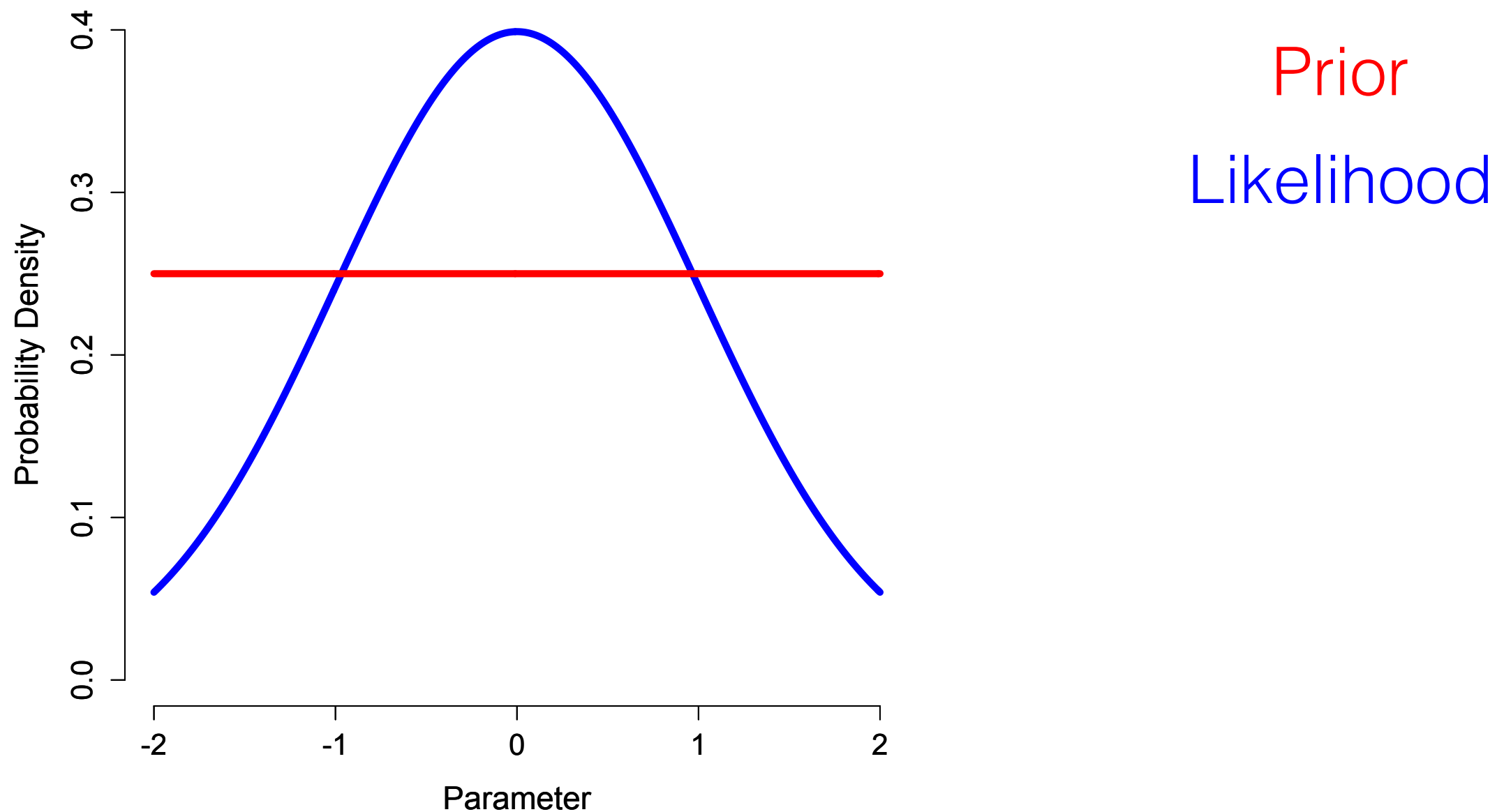
Easy Approach 1 - Sample from the prior

Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior
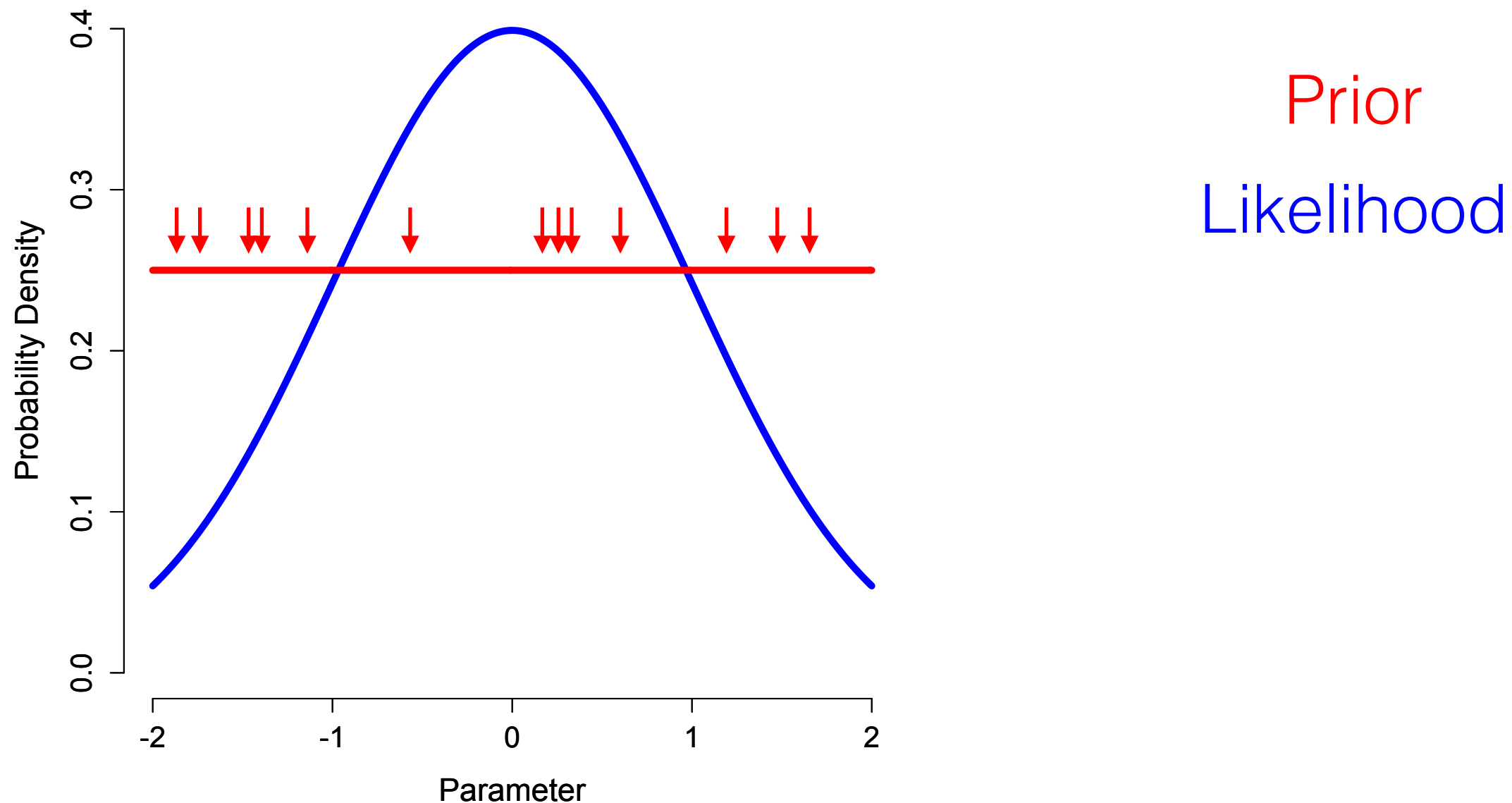
Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior
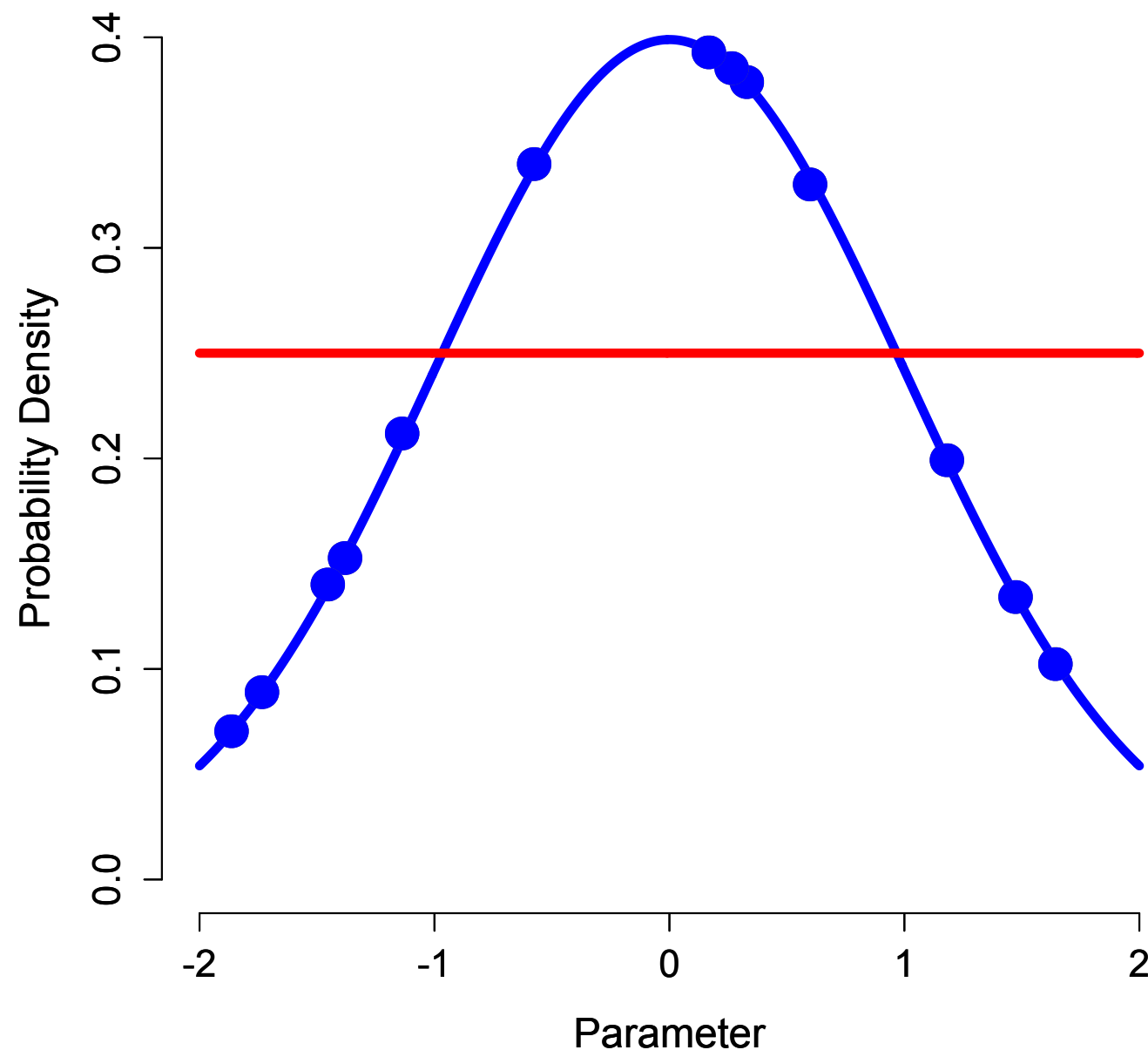
Prior

Likelihood

Take average of blue dots

Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior

Calculating Marginal Likelihoods

Easy Approach 1 - Sample from the prior
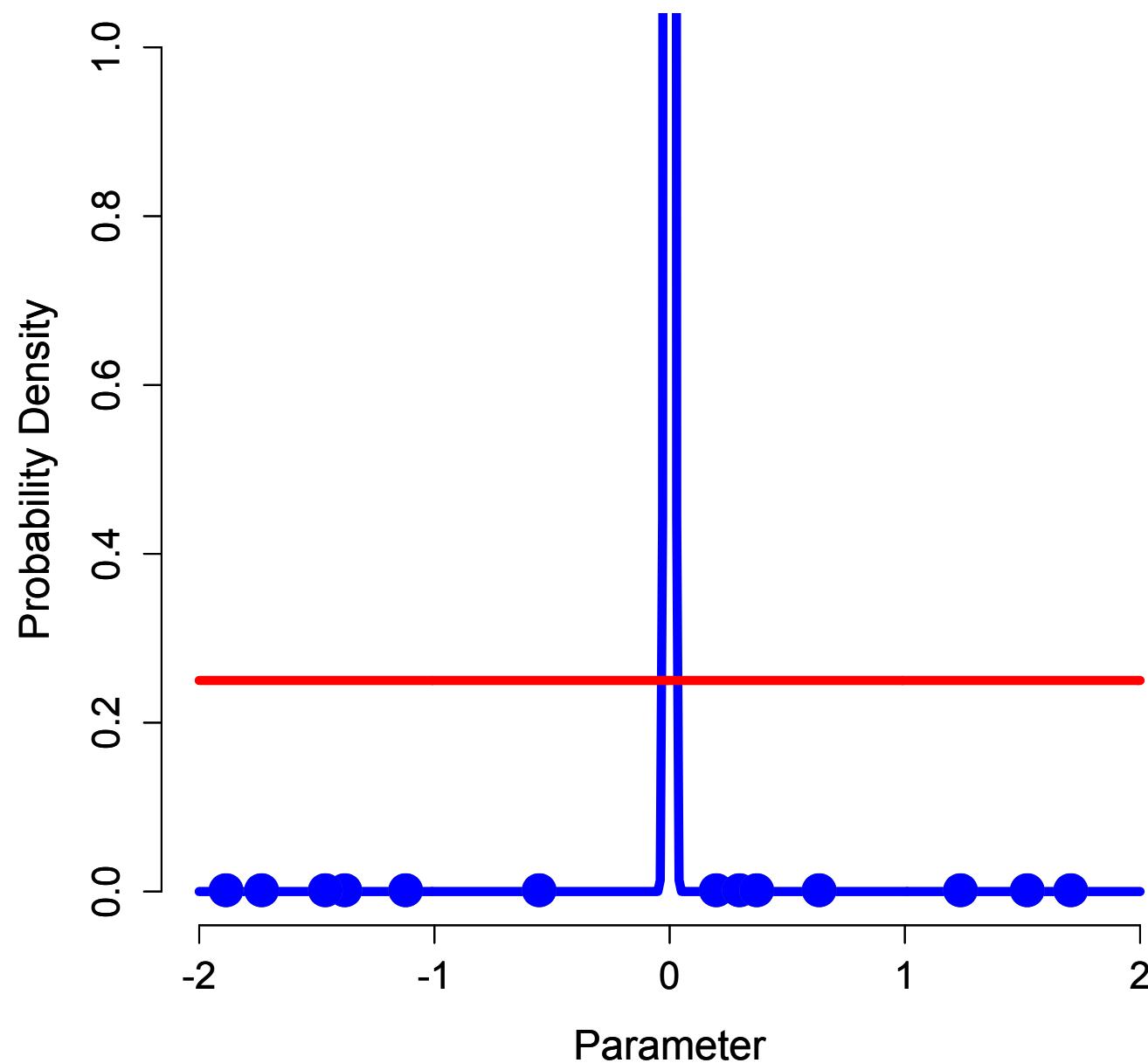
Prior

Likelihood

Take average of blue dots**??!!**

We'd like to make sure we're sampling high likelihood parts of space with reasonable frequency.

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior
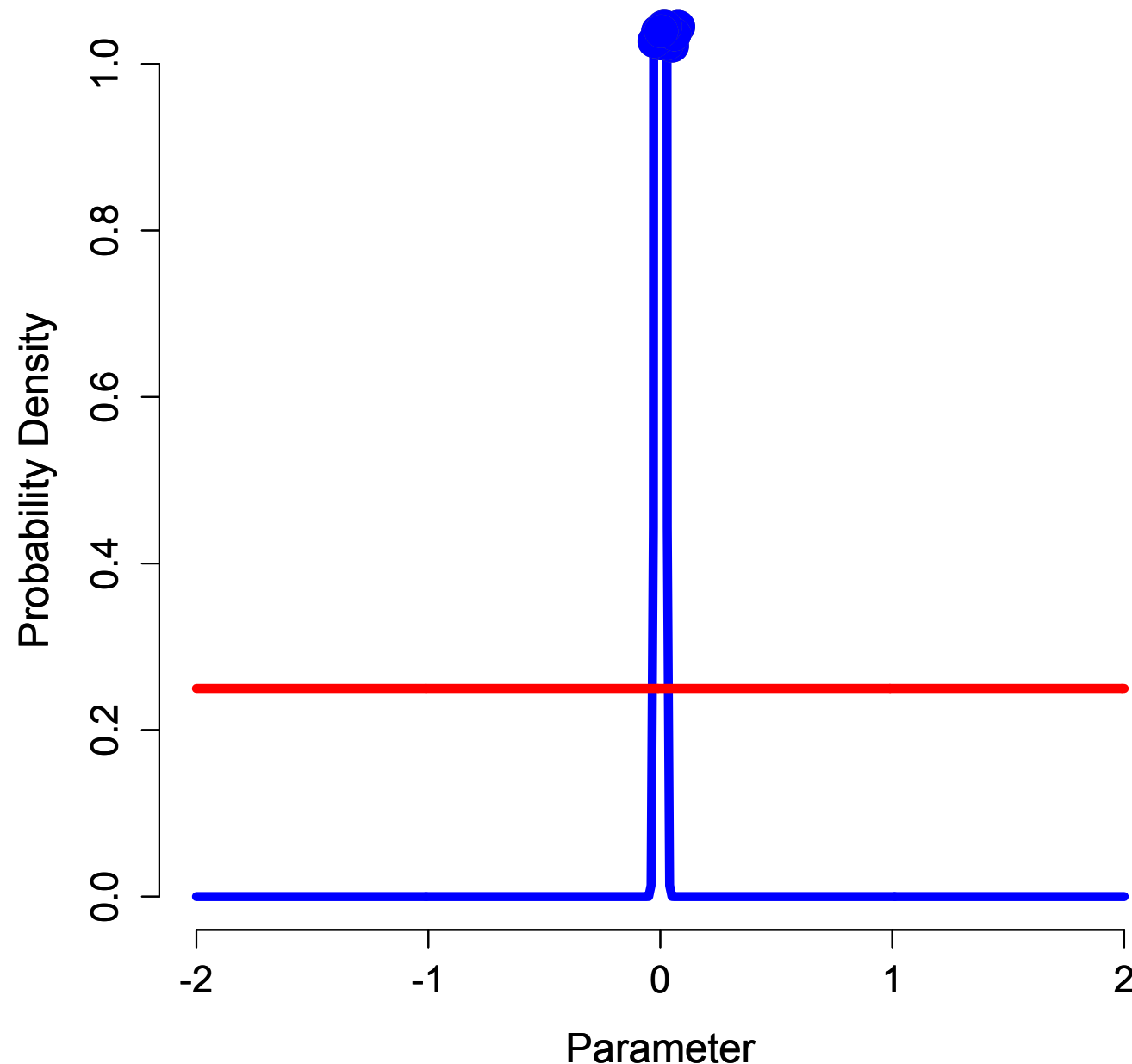
Prior

Posterior (~Likelihood)

Since we're supposed to be integrating across the prior, we need to correct for the fact that our samples are from the posterior.

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior

The Harmonic Mean Method

$$\frac{1}{ML} = \frac{1}{L_1} + \frac{1}{L_2} + \ldots + \frac{1}{L_3}$$

What's an important property of harmonic means?

Anyone remember discussing bottlenecks in pop gen?

Calculating Marginal Likelihoods

Less-Naive Approach 2- Sample from the posterior
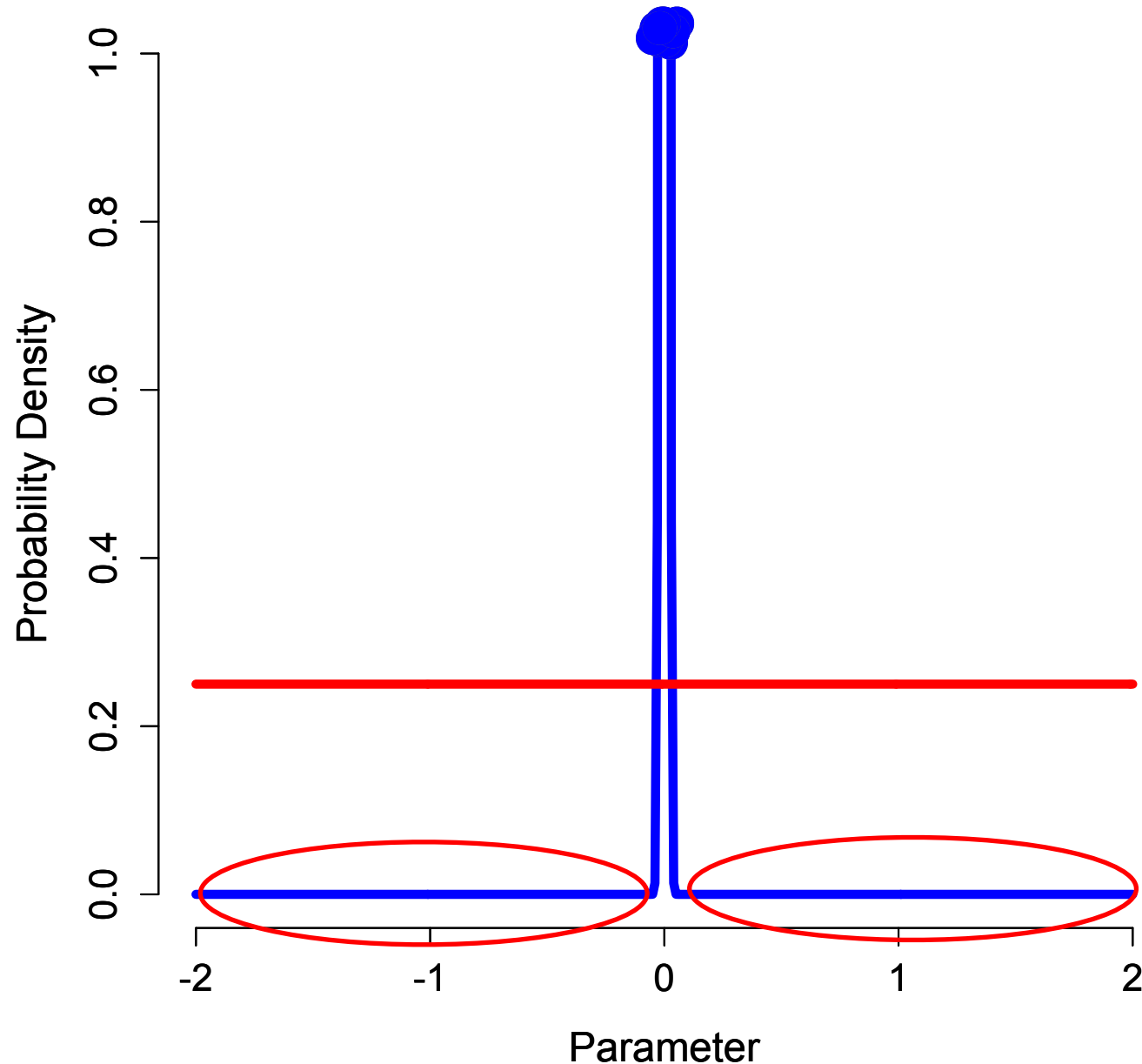
The reverse problem to our first naive approach!

Rarely sampled low likelihoods have a big influence on estimates.

Very unstable.

Calculating Marginal Likelihoods

Approach 3* - Sample from a series of distributions

Steppingstone or path sampling

Steppingstone Sampling

$$P(D) = \int P(D|\theta)P(\theta)$$

$P(D|\theta)P(\theta)$ ········

Equivalent to estimating the area under this curve.

θ
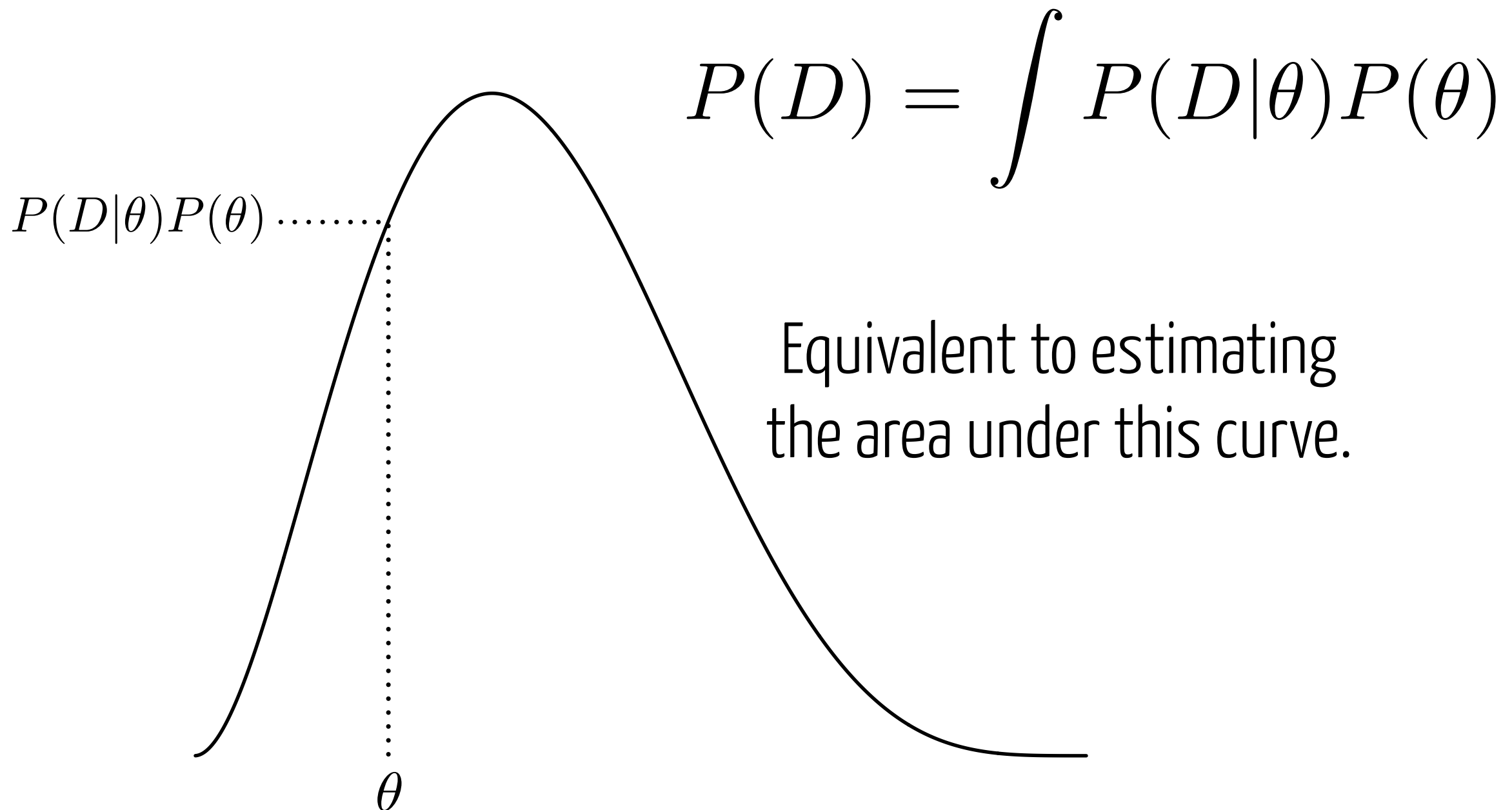
Steppingstone Sampling
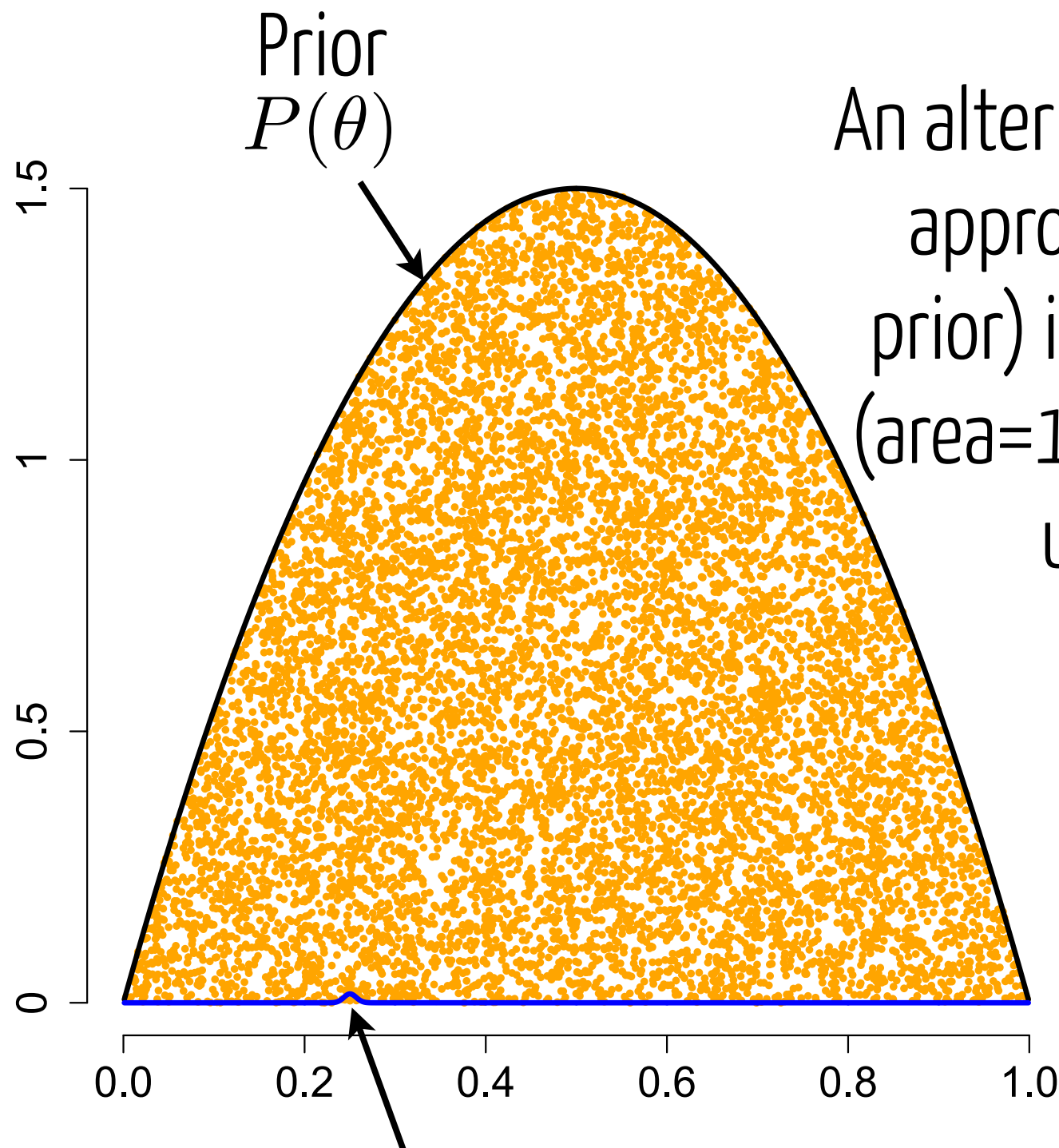
Prior
$P(\theta)$

An alternative way to think about our first approach (sampling directly from the prior) is to sample points from the prior (area=1.0), then ask what proportion fall under the curve of interest.

Unfortunately, not many!
As before, this is unstable.

Unnormalized Posterior ⟶ $P(D|\theta)P(\theta)$

Steppingstone Sampling

$P(D)$
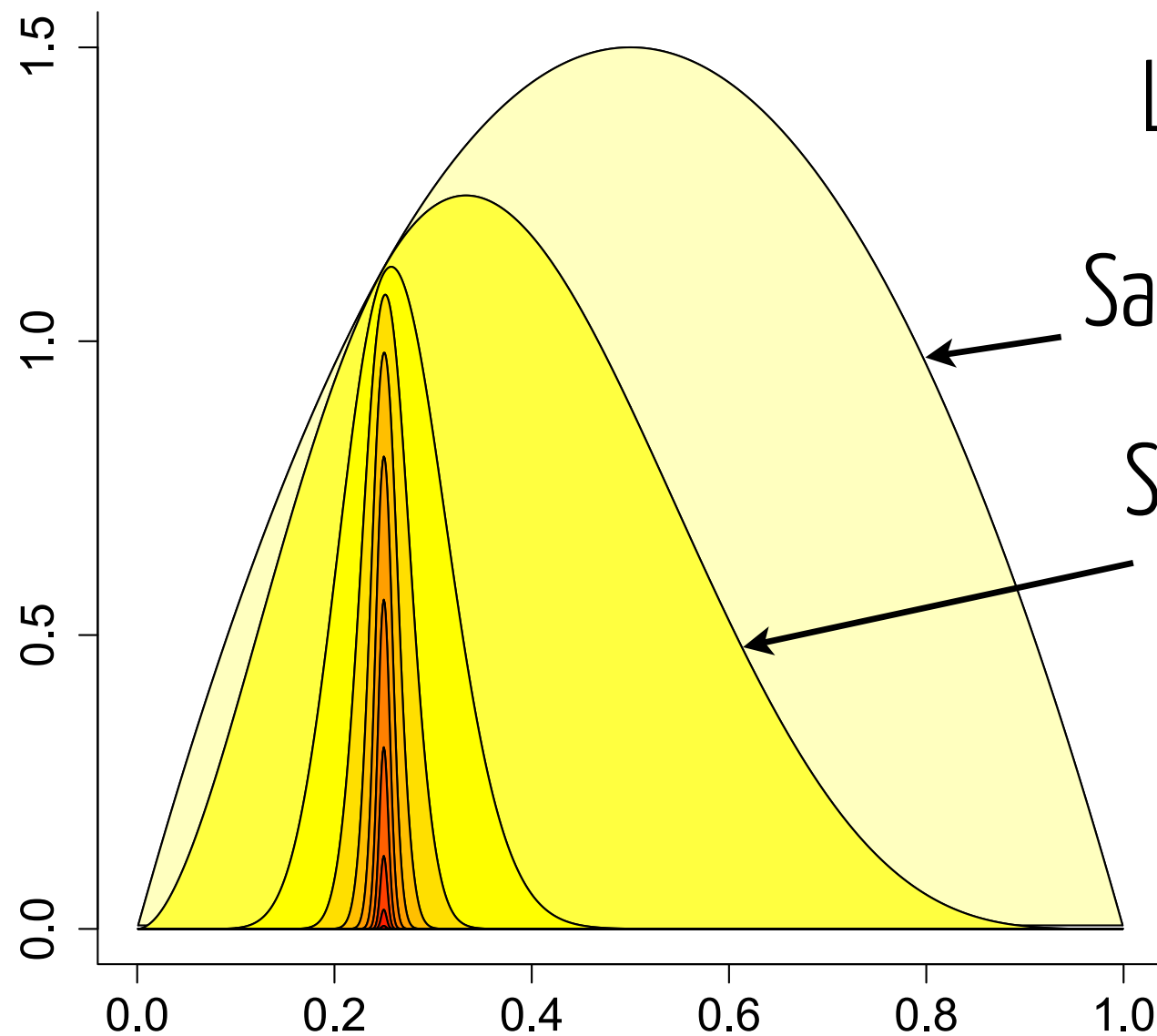
Let's try it in **steps**!

Sample from this distribution

See what fraction of samples are under this curve

That fraction is an estimate of this ratio

$$\frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{c_{0.9}}\right) \left(\frac{c_{0.9}}{c_{0.8}}\right) \left(\frac{c_{0.8}}{c_{0.7}}\right) \left(\frac{c_{0.7}}{c_{0.6}}\right) \left(\frac{c_{0.6}}{c_{0.5}}\right) \left(\frac{c_{0.5}}{c_{0.4}}\right) \left(\frac{c_{0.4}}{c_{0.3}}\right) \left(\frac{c_{0.3}}{c_{0.2}}\right) \left(\frac{c_{0.2}}{c_{0.1}}\right) \left(\frac{c_{0.1}}{c_{0.0}}\right)$$

Power Posteriors

$P(D)$

$$\frac{c_{1.0}}{c_{0.0}} = \left(\frac{c_{1.0}}{c_{0.9}}\right)\left(\frac{c_{0.9}}{c_{0.8}}\right)\left(\frac{c_{0.8}}{c_{0.7}}\right)\left(\frac{c_{0.7}}{c_{0.6}}\right)\left(\frac{c_{0.6}}{c_{0.5}}\right)\left(\frac{c_{0.5}}{c_{0.4}}\right)\left(\frac{c_{0.4}}{c_{0.3}}\right)\left(\frac{c_{0.3}}{c_{0.2}}\right)\left(\frac{c_{0.2}}{c_{0.1}}\right)\left(\frac{c_{0.1}}{c_{0.0}}\right)$$

1

Posterior

$\beta = 1$

Prior

$\beta = 0$

$$P(D|\theta)_\beta \propto P(D|\theta)^\beta P(\theta)$$

Power Posteriors

$$\frac{c_{1.0}}{c_{0.0}} = \text{Stable estimate of marginal likelihood!}$$

But it requires a **specific type of analysis**,
independent of standard MCMC.

Bayes Factors

Posterior

Likelihood

Prior

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

Normalizing Constant
(Marginal Likelihood)

Bayes Factors

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

Bayes Factors

$$\frac{\color{red}P(H_1|D)}{\color{red}P(H_2|D)} = \frac{\dfrac{\color{blue}P(H_1)\color{green}P(D|H_1)}{\color{orange}\cancel{P(D)}}}{\dfrac{\color{blue}P(H_2)\color{green}P(D|H_2)}{\color{orange}\cancel{P(D)}}}$$

Bayes Factors

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

Bayes Factors

Posterior Odds   Prior Odds   Bayes Factor

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

Bayes Factors

$$\underbrace{\frac{P(H_1)}{P(H_2)}}_{\text{Prior Odds}} \underbrace{\frac{P(D|H_1)}{P(D|H_2)}}_{\text{Bayes Factor}} = \underbrace{\frac{P(H_1|D)}{P(H_2|D)}}_{\text{Posterior Odds}}$$

Bayes Factors

2ln(BF)	BF	Strength of evidence
0-2	1-3	Barely worth mentioning.
2-6	3-20	Positive
6-10	20-150	Strong
>10	>150	Very Strong

Robert E. Kass & Adrian E. Raftery (1995). "Bayes Factors". Journal of the American Statistical Association. 90:791

Bayes Factors

For now, we're going to use these to **compare different models of sequence evolution** as our hypotheses.
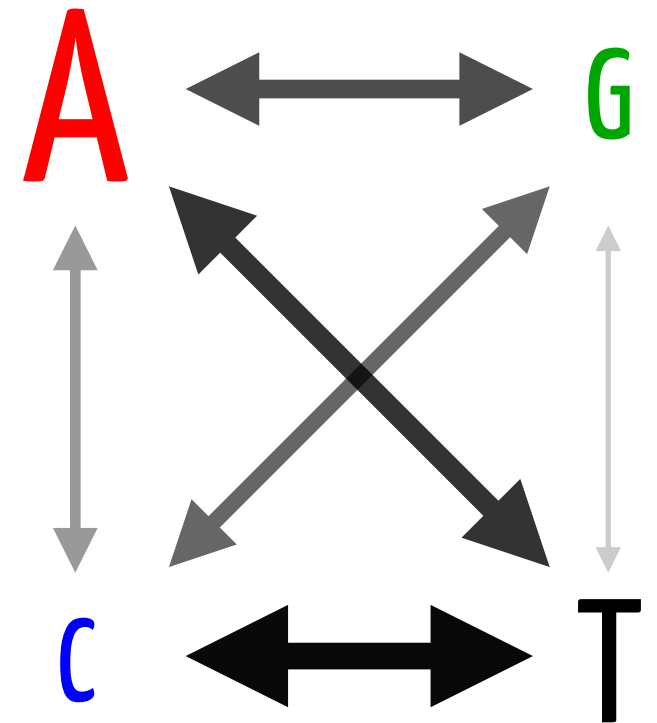
However, BFs can also be used for other hypotheses, like **topological relationships** (this afternoon).
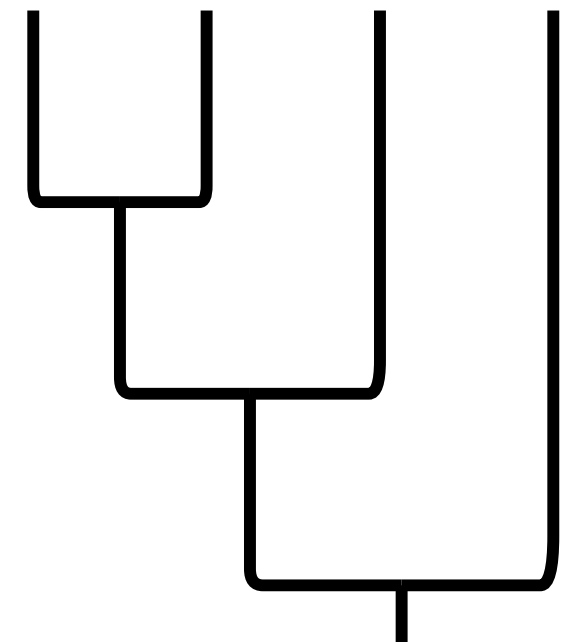
Bayes Factors

Tutorial (in just a second)

Or...don't choose a model!

Reversible Jump MCMC

Instead of picking a model, include MCMC moves that jump between. Integrate out uncertainty about which model is best. This is a Bayesian form of **model averaging**.

We already do this for trees. Can also do this for models.

Or...don't choose a model!

Reversible Jump MCMC

Instead of picking a model, include MCMC moves that jump between. Integrate out uncertainty about which model is best. This is a Bayesian form of **model averaging**.

We already do this for trees. Can also do this for models.

*Disclaimer: Setting up proper reversible jump moves can often be **very challenging**.