Comparing Models and Evaluating Their Fit

Jeremy M. Brown

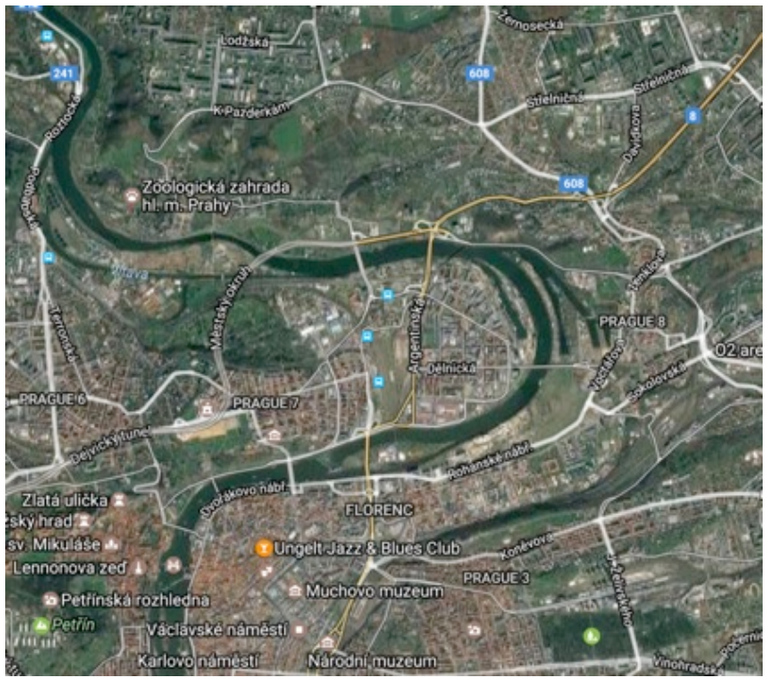
Louisiana State University

Model Selection

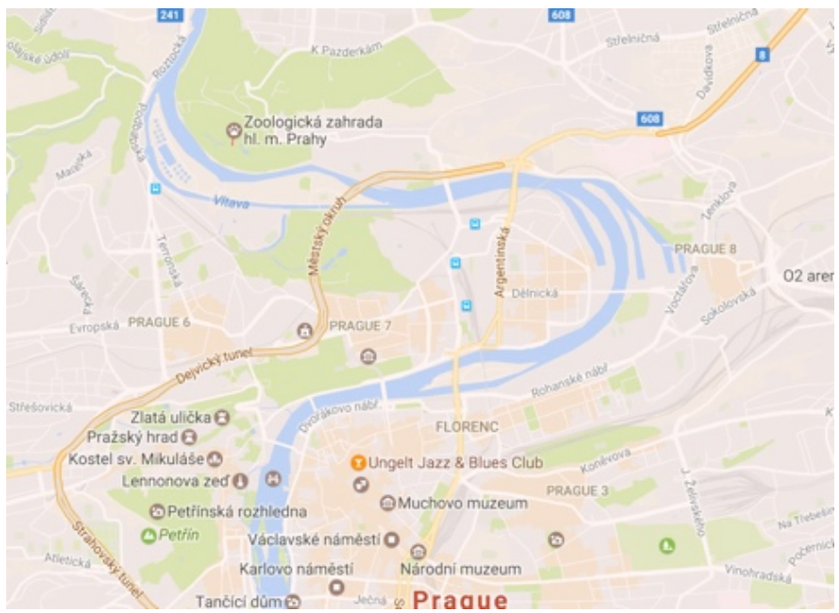
To do statistical inference we must have a model

- What model should that be?
- Our goal should be to have a model that is complex enough to capture the "important" variation in the data, but not be more complex than it needs to be.

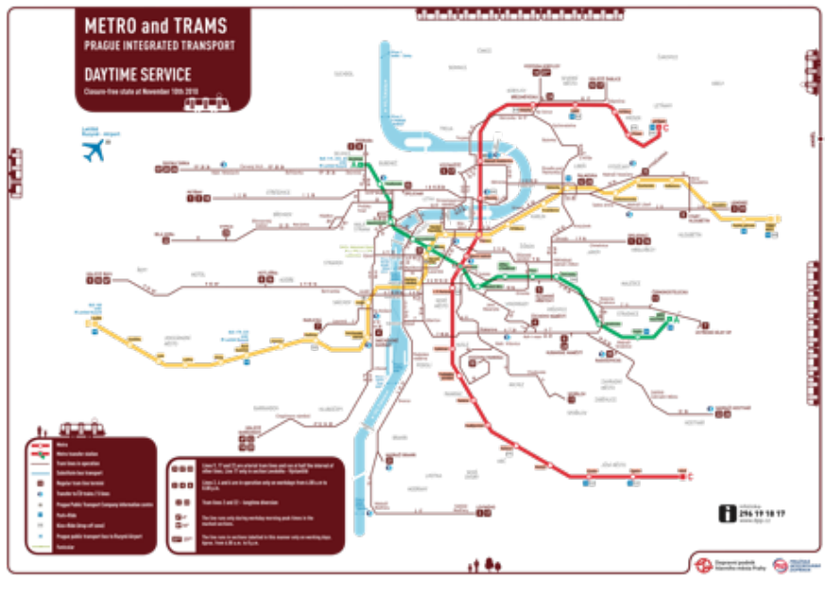
Which is more useful?



“Reality”



Detailed map



Detailed public transportation

Simplified metro

Taken from Dave Swofford

Concept credit: Paul Lewis

Models don’t need to reflect reality

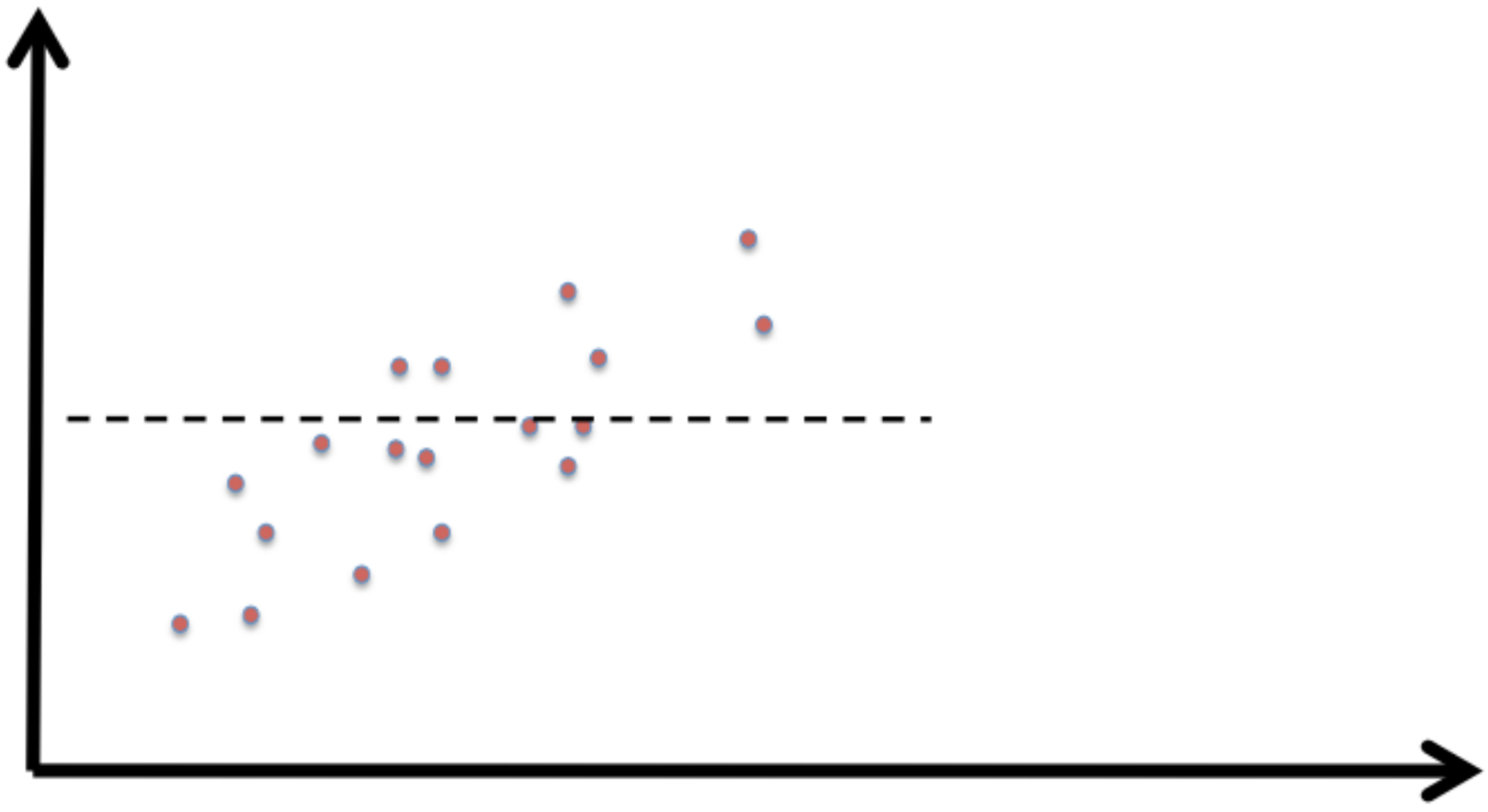
"The most that can be expected from any model is that it can supply a useful approximation to reality: **All models are wrong; some models are useful"**. (George E. P. Box, 1987)

Model selection is a process of seeking the least inadequate model from a predefined set, all of which may be grossly inadequate as a representation of reality. (J. J. Welch, 2006)

Taken from Dave Swofford

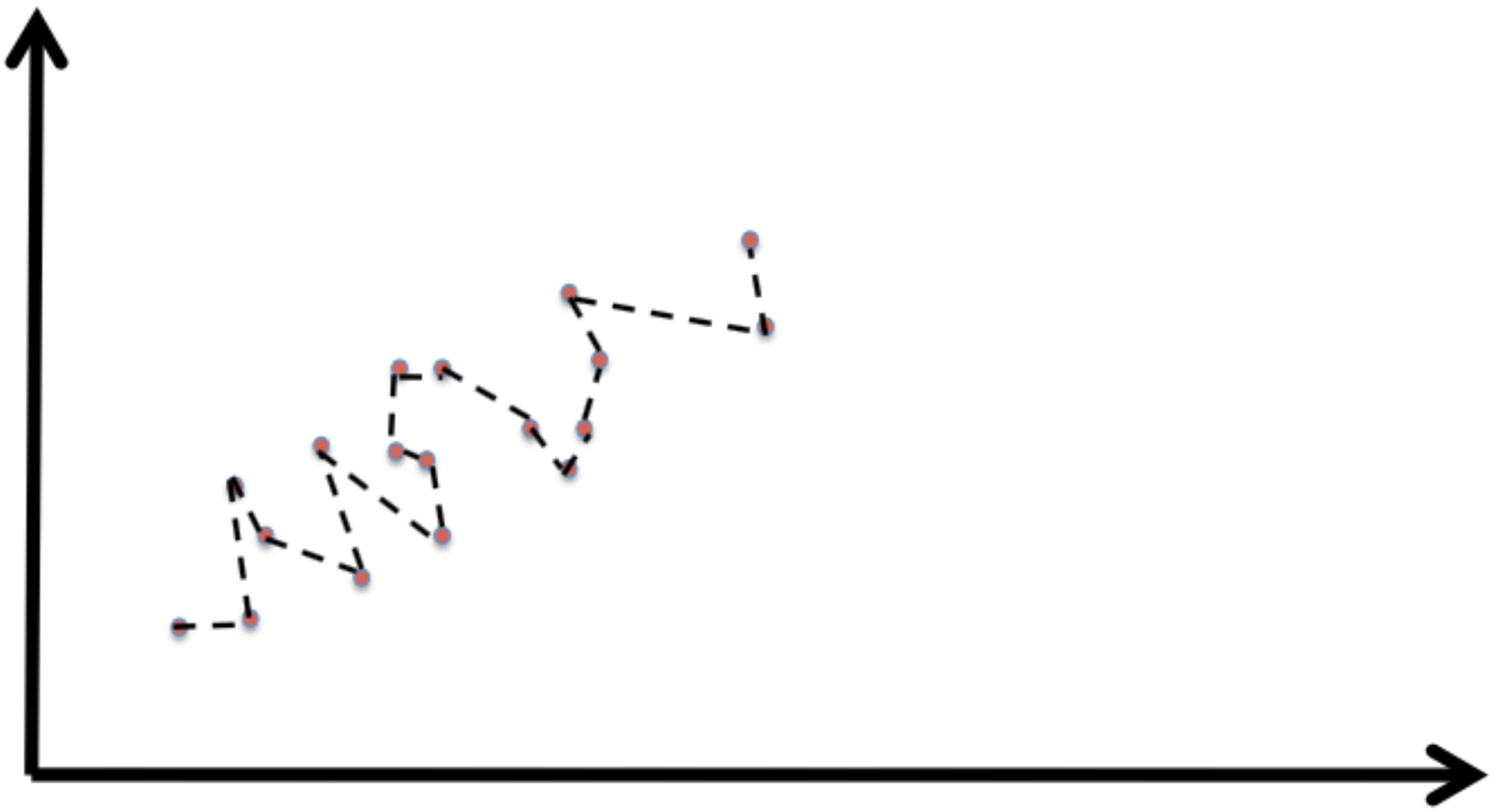
Model Selection

Underfitting model: does not capture important variation in the data



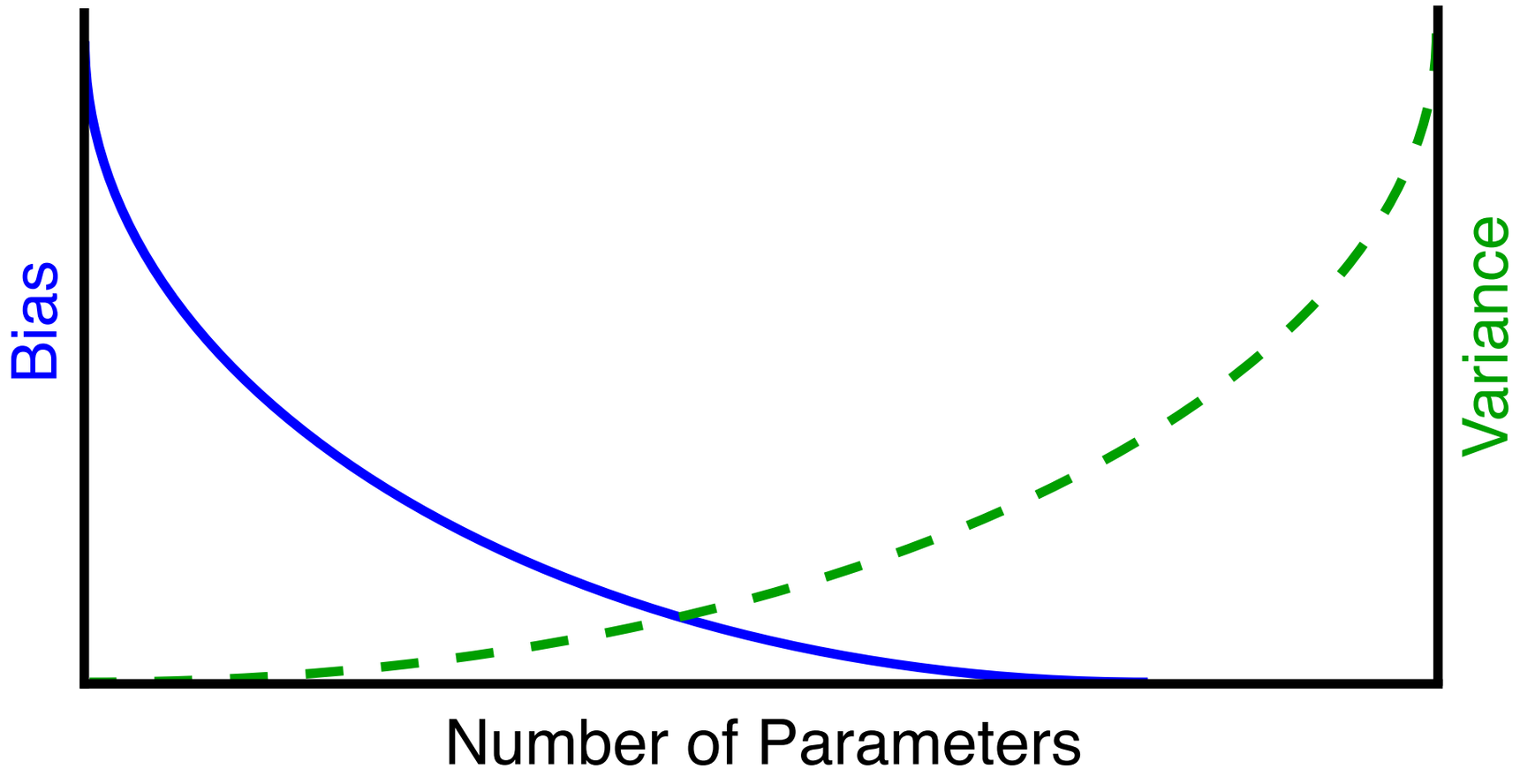
Model Selection

Overfitting model: model captures all variation in the data, but is not a realistic description of the underlying process



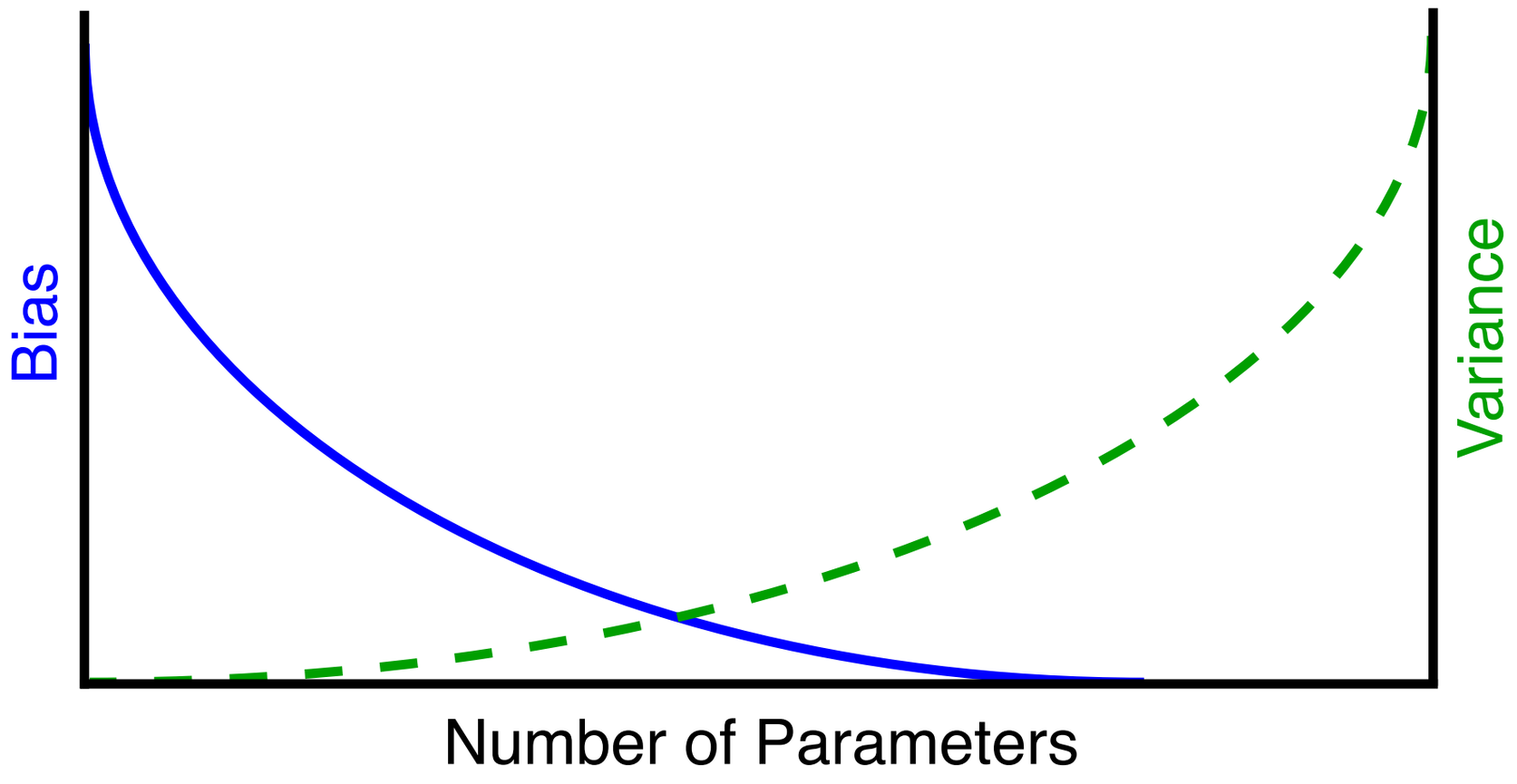
Model Selection

The Fundamental Tradeoff

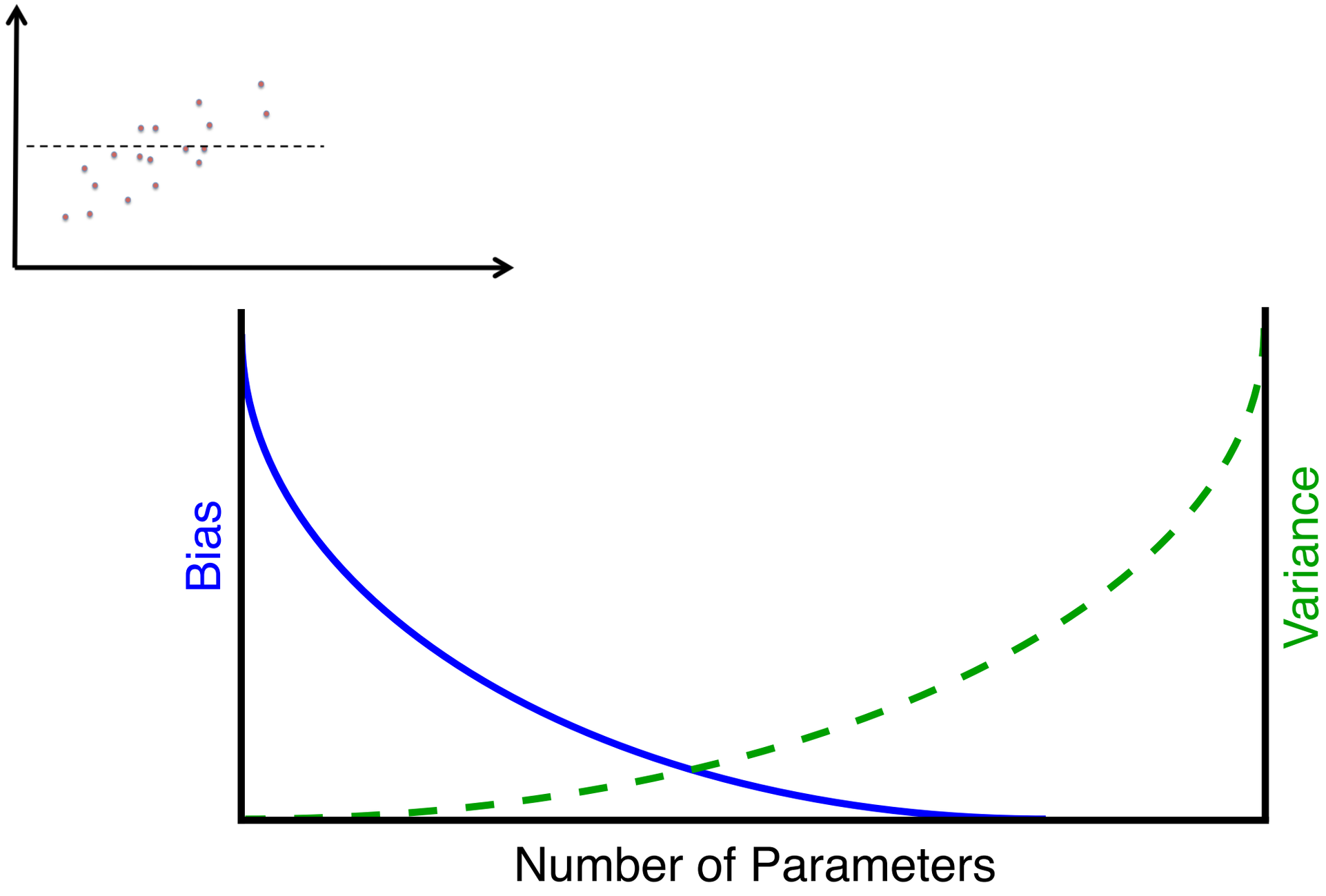


Bias-Variance Tradeoff

Model too simple!
We're misinterpreting the data.

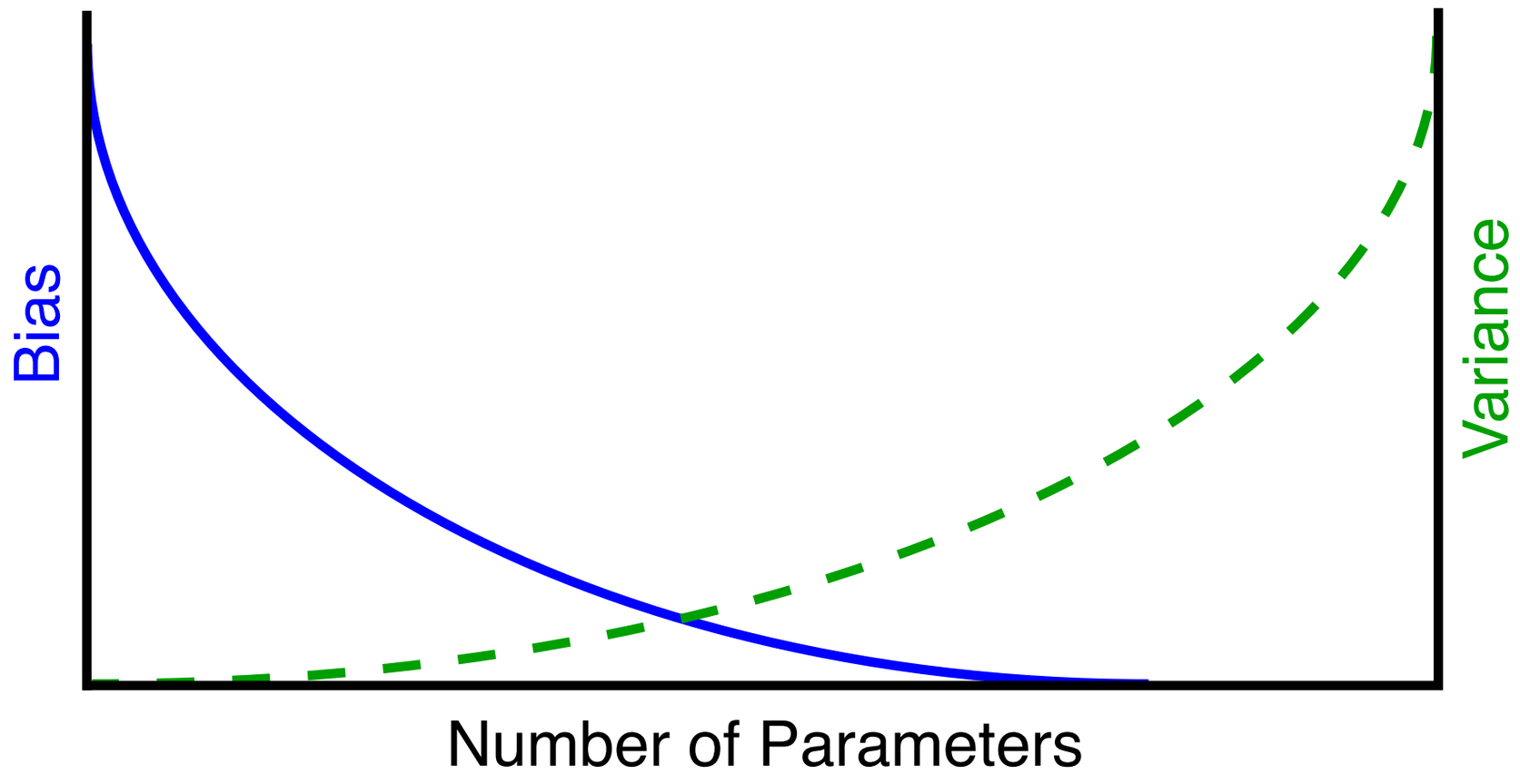


Bias-Variance Tradeoff

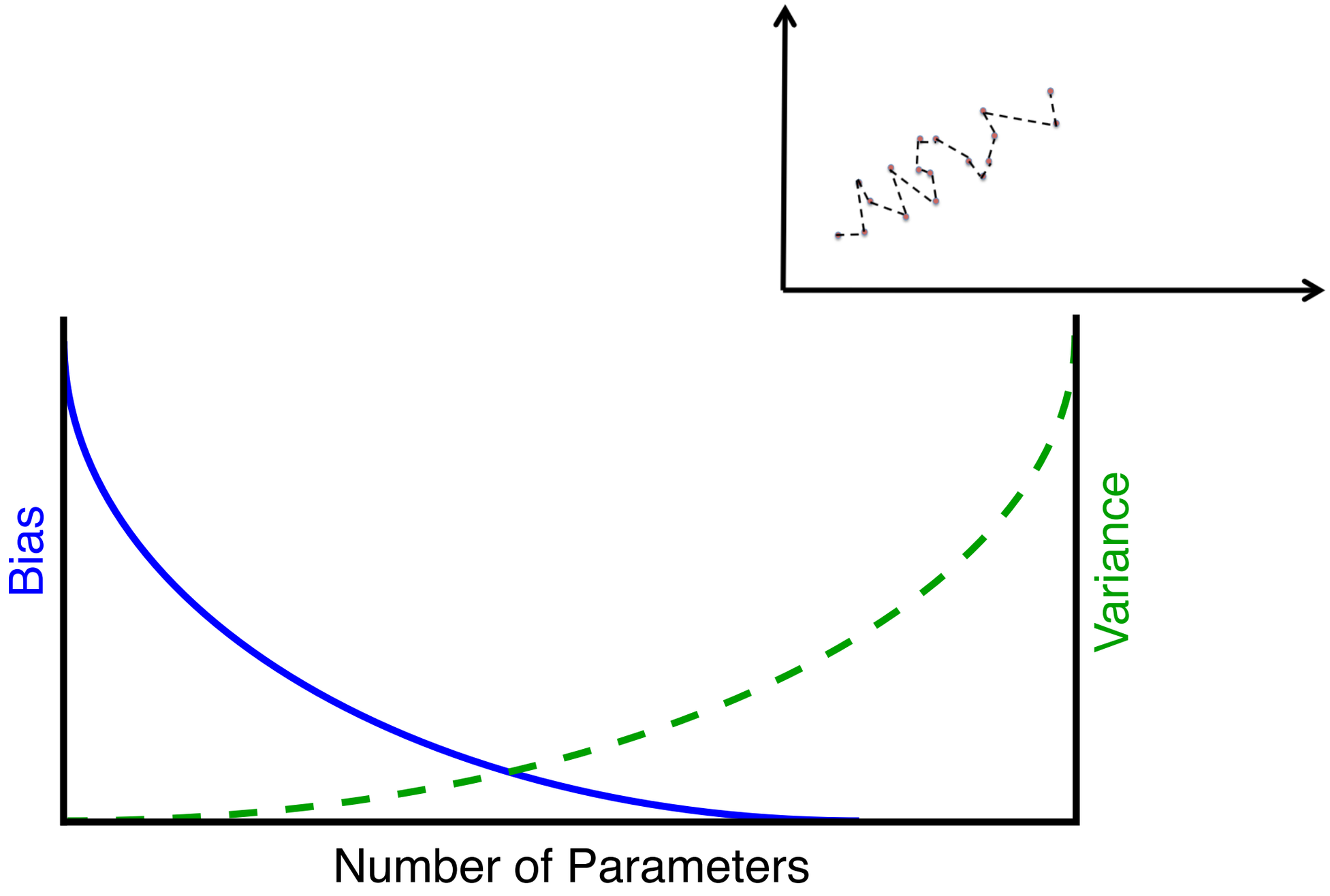


Bias-Variance Tradeoff

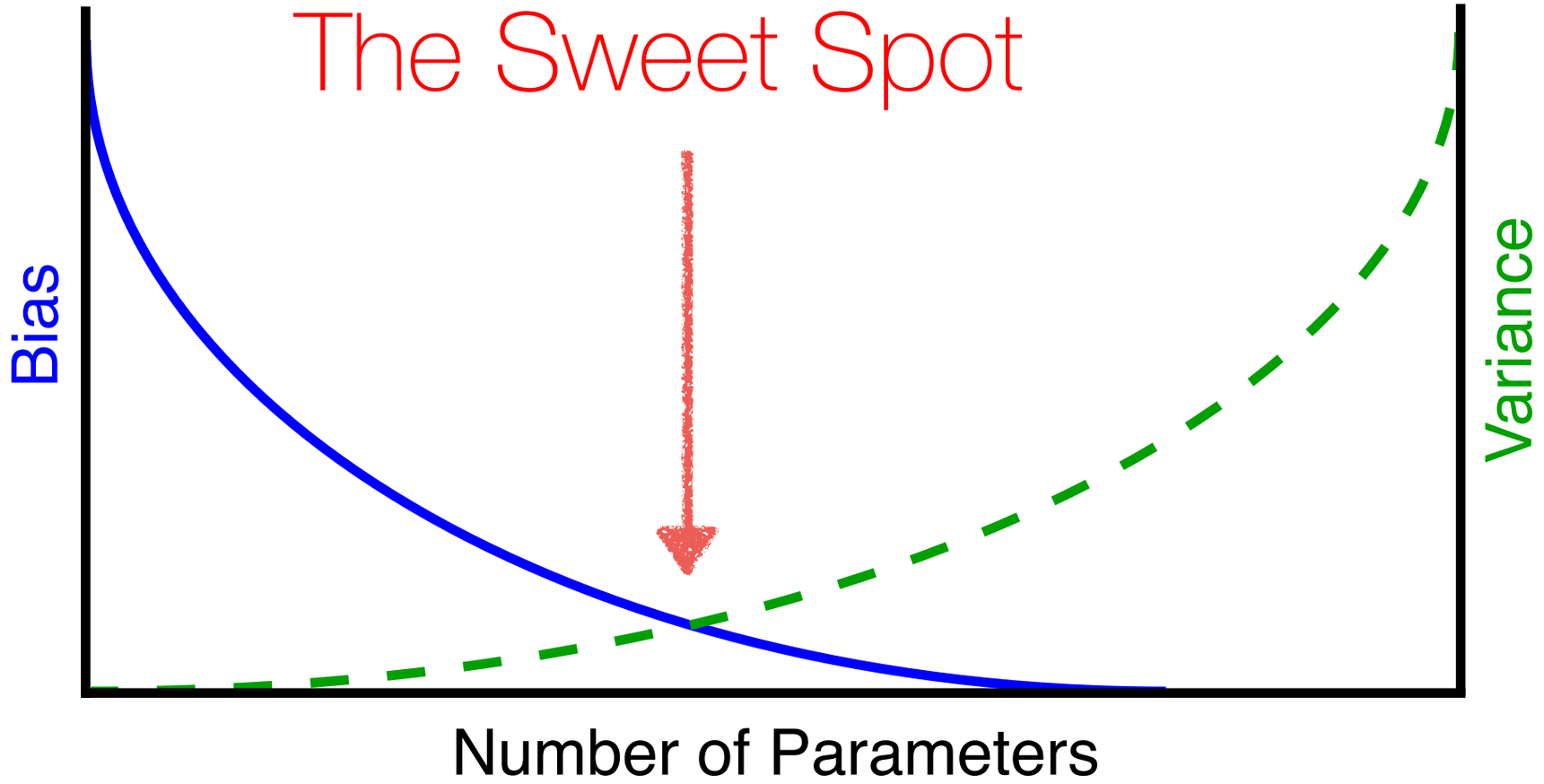
Model too complicated!
We don't have enough information.



Bias-Variance Tradeoff



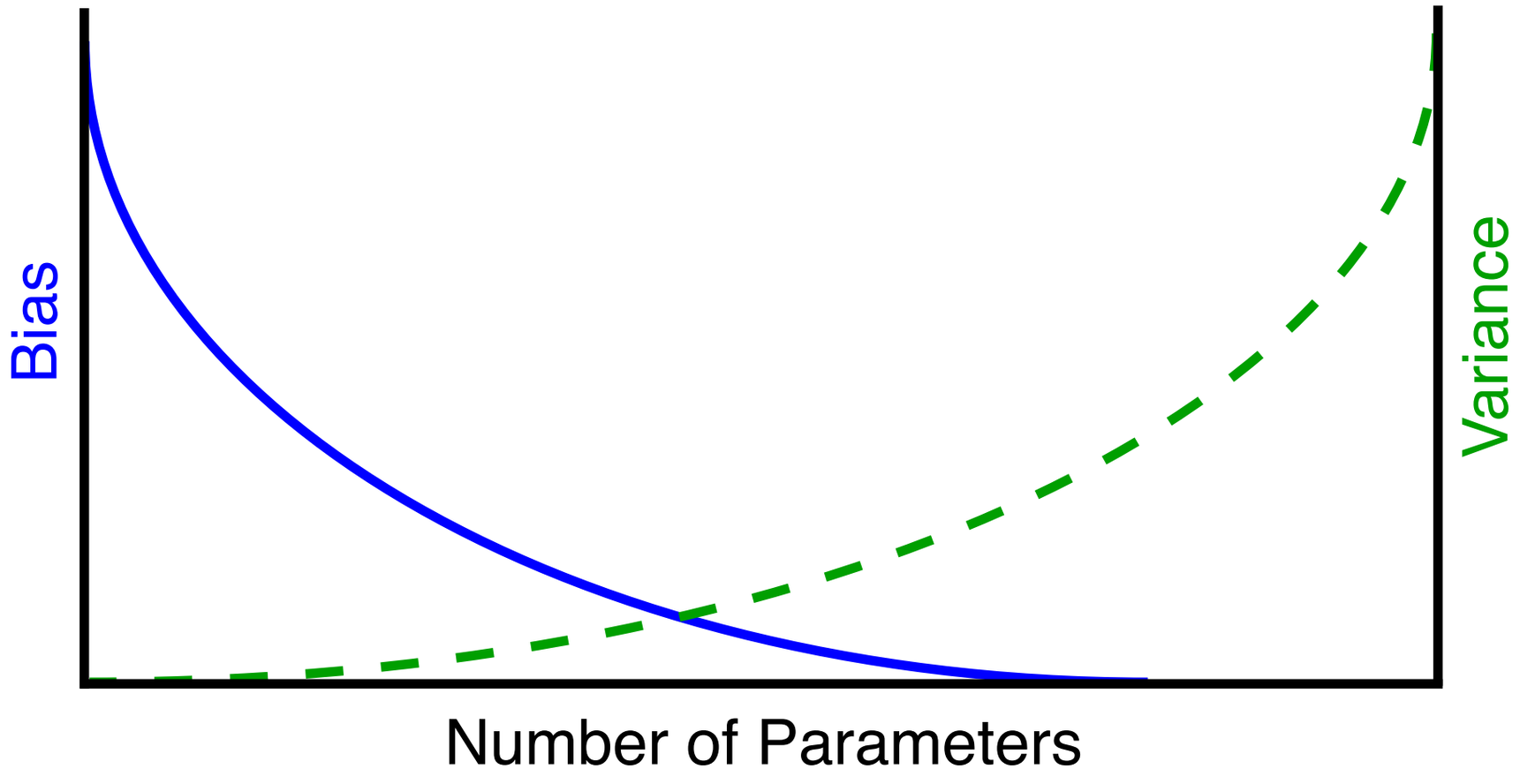
Model Selection



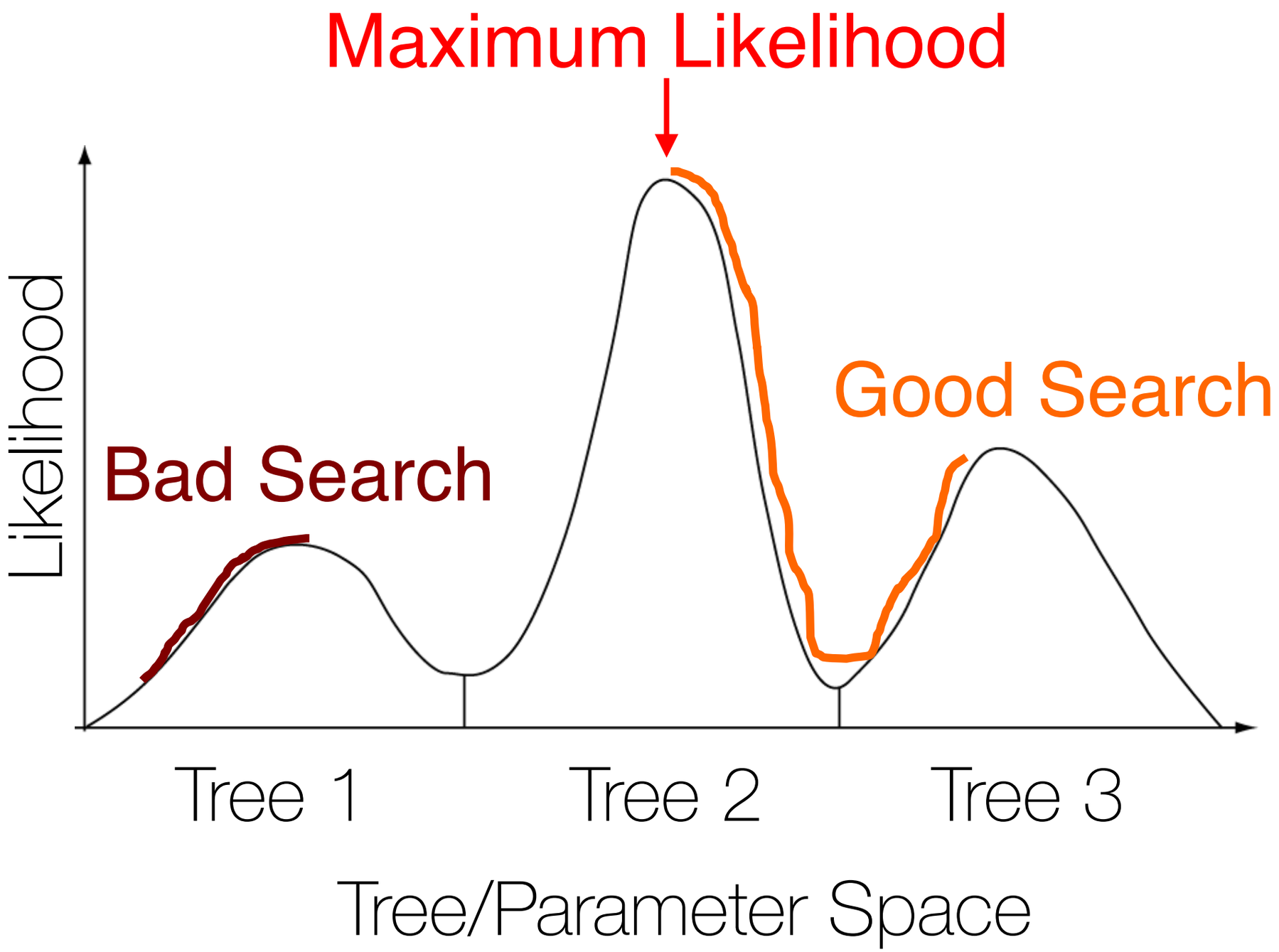
Model Selection

Bias and Variance can be traded off in different ways.

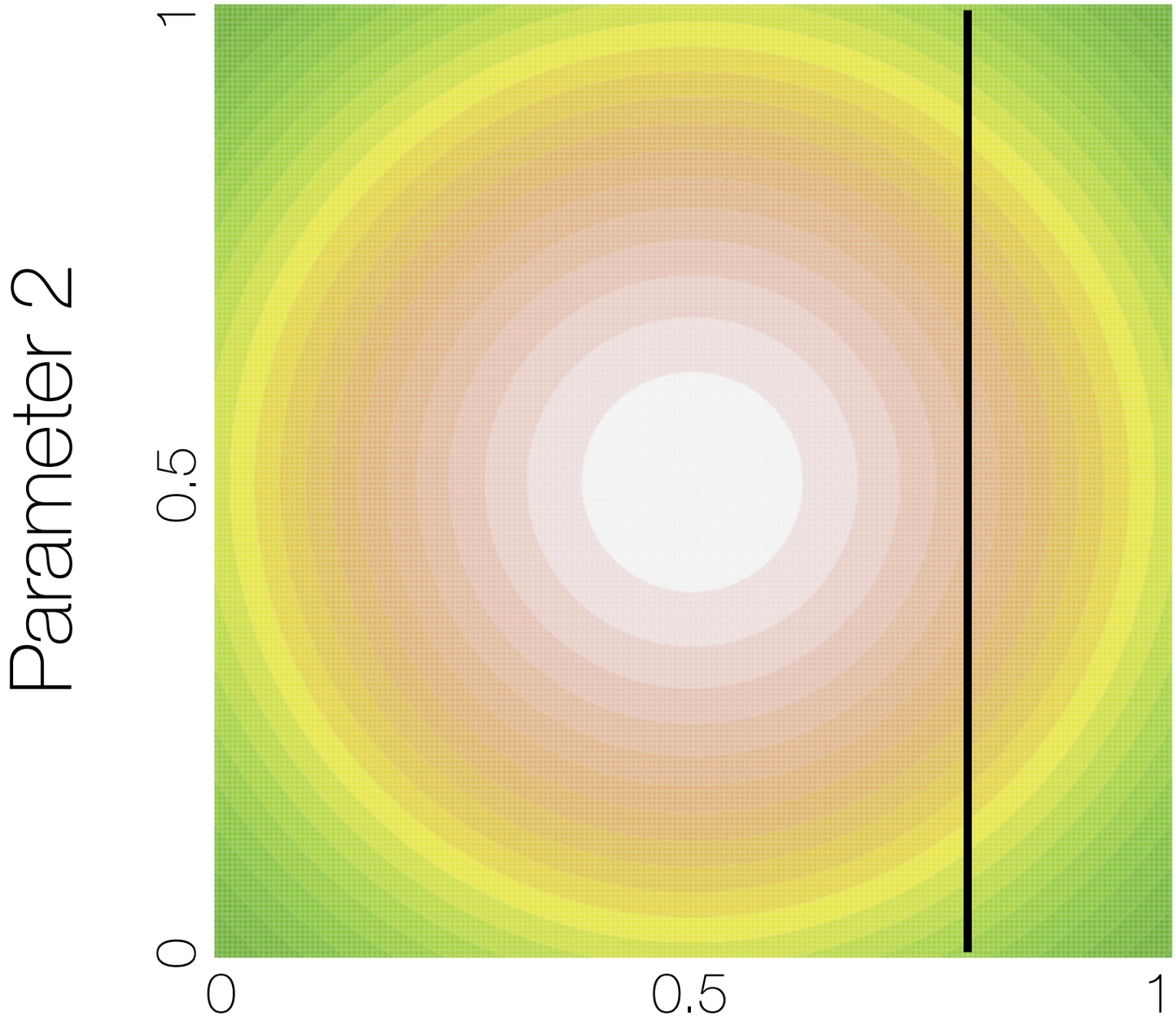
This leads to multiple criteria for model selection.



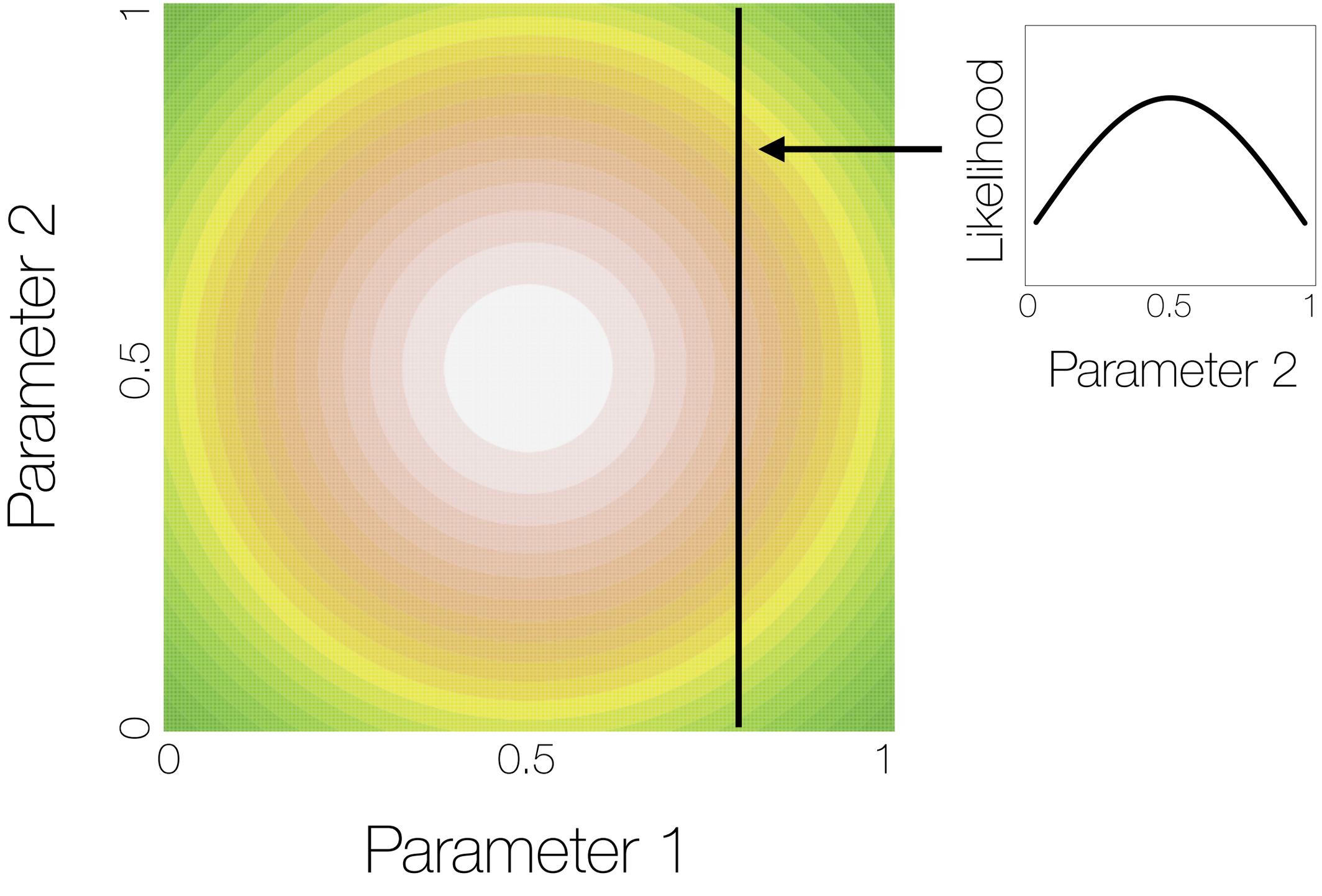
Maximum Likelihood



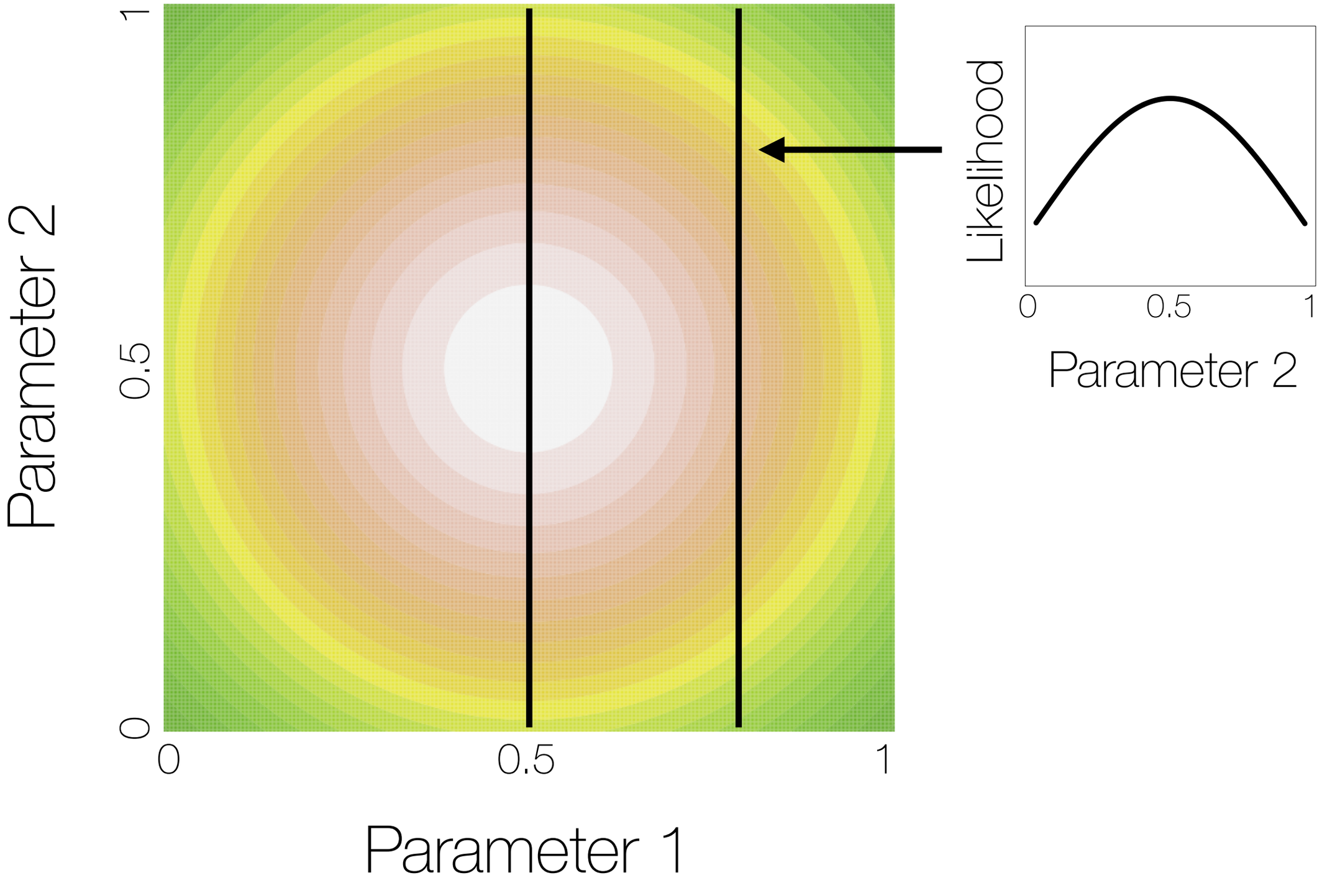
More Parameters = Better Likelihood



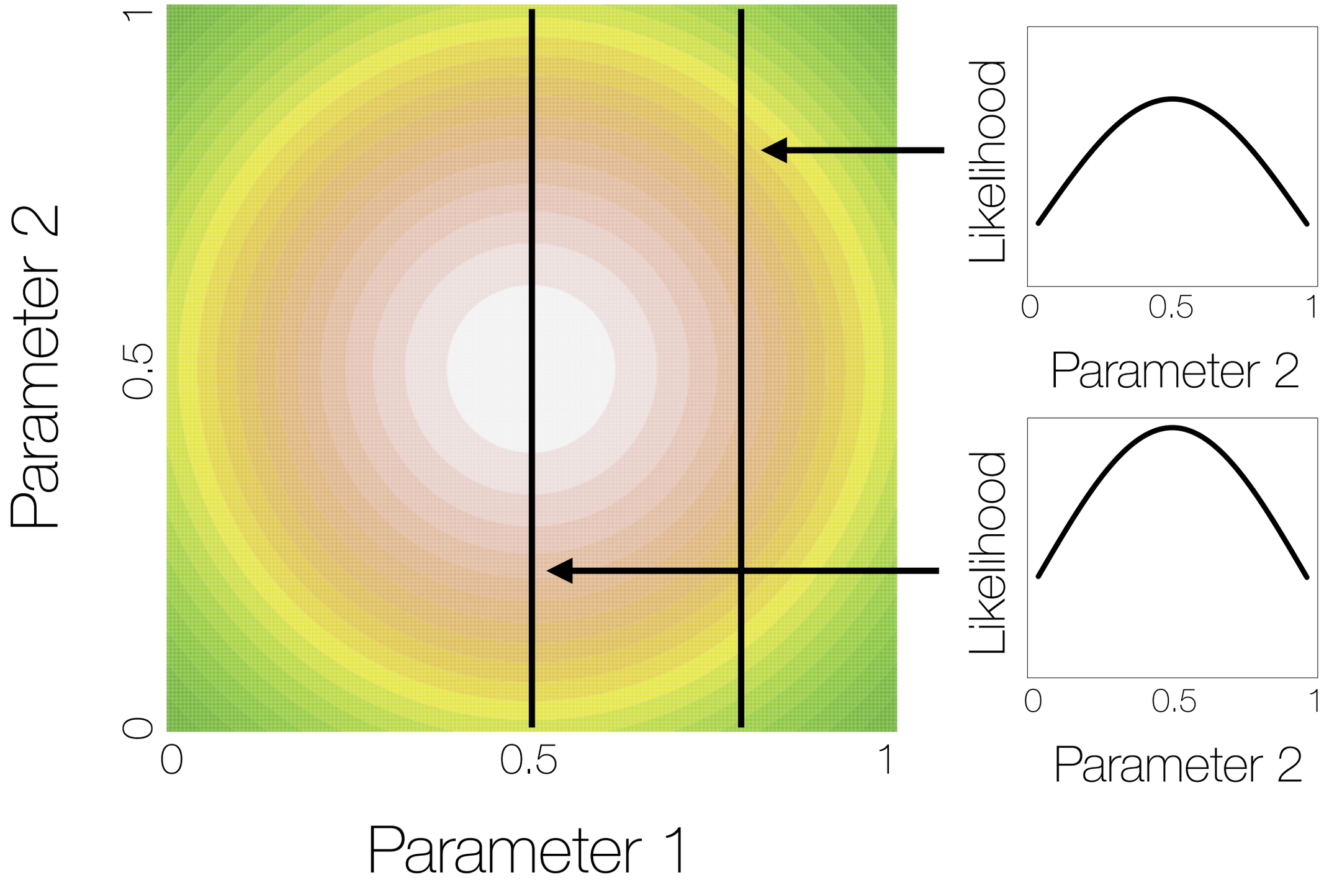
More Parameters = Better Likelihood



More Parameters = Better Likelihood

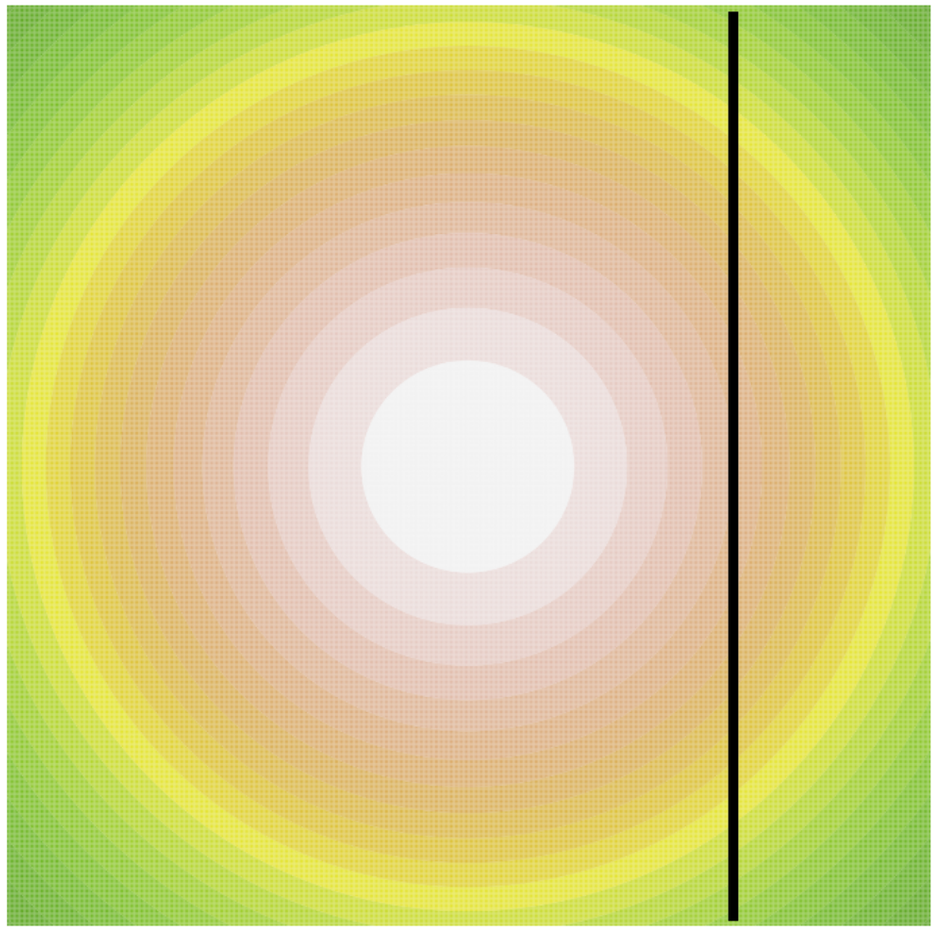


More Parameters = Better Likelihood



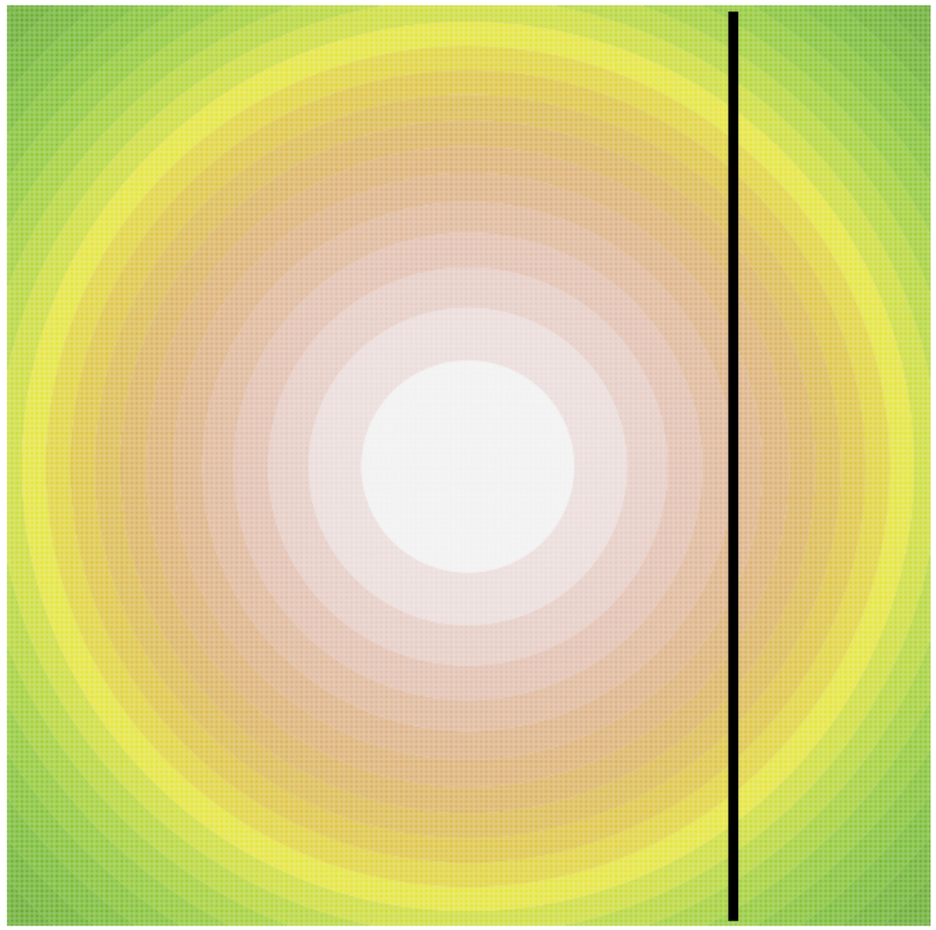
ML-based Model Selection

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

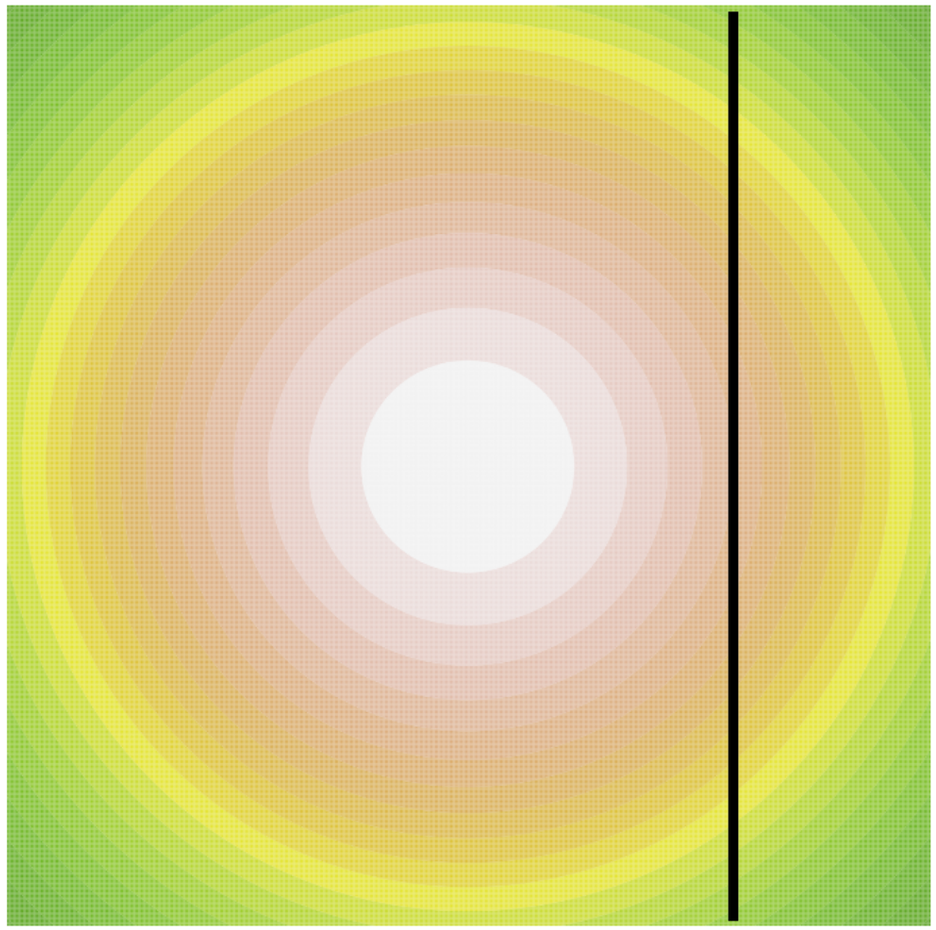
If the **more complex** model always gives **a likelihood that is at least as good** as a simpler model, even if the simpler one is true, we need ways to assess **whether it's enough better to warrant our attention**.



- 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.

$$\mathrm{AIC} = 2k - 2ln(\hat{L})$$

Penalty for more parameters (k).

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

ML-based Model Selection

$$\mathrm{AIC} = 2k - 2ln(\hat{L})$$

$$\mathrm{BIC} = ln(n)k - 2ln(\hat{L})$$

ML-based Model Selection

Likelihood ratio tests

- Calculate "delta" statistic

$$\delta = 2(\ln L_1 - \ln L_0)$$

If model 0 (simpler) is nested within model 1 (more complex), δ is asymptotically distributed as a χ^2 random variable with degrees-of-freedom equal to the difference in number of free parameters.*

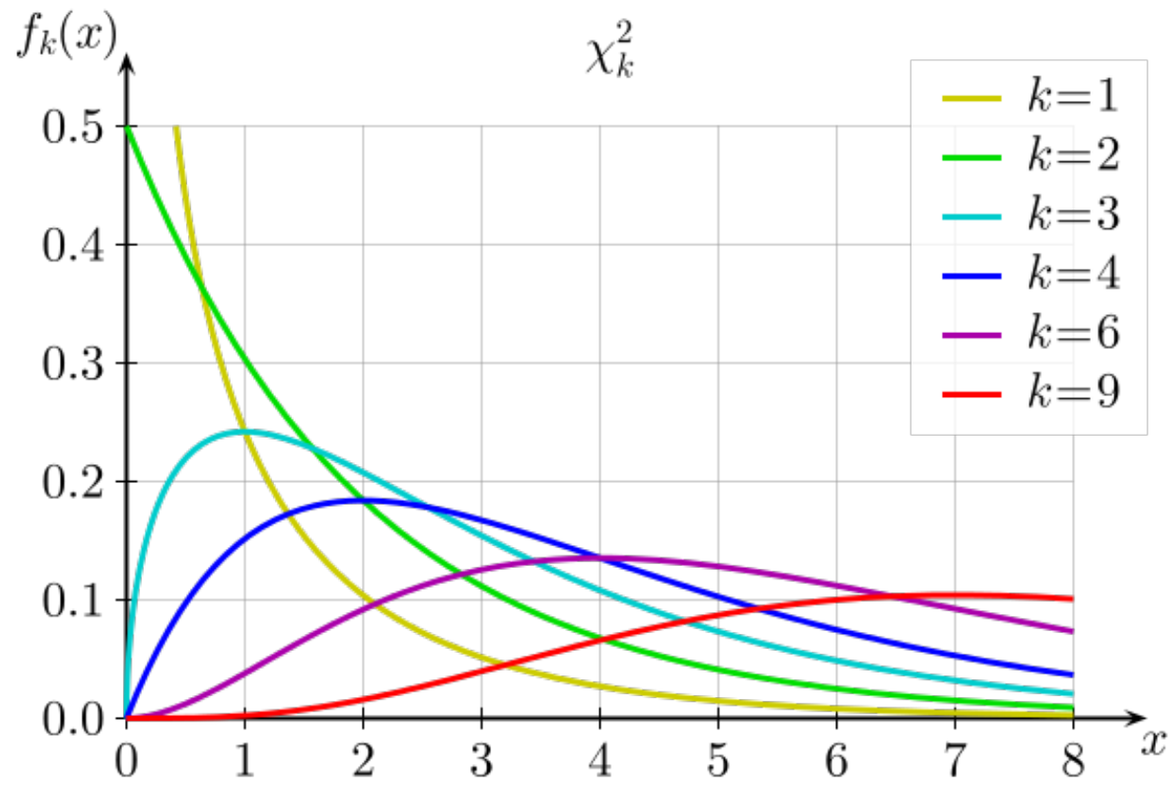
Taken from Dave Swofford

ML-based Model Selection

LRT

Hypothesis test

Pairwise

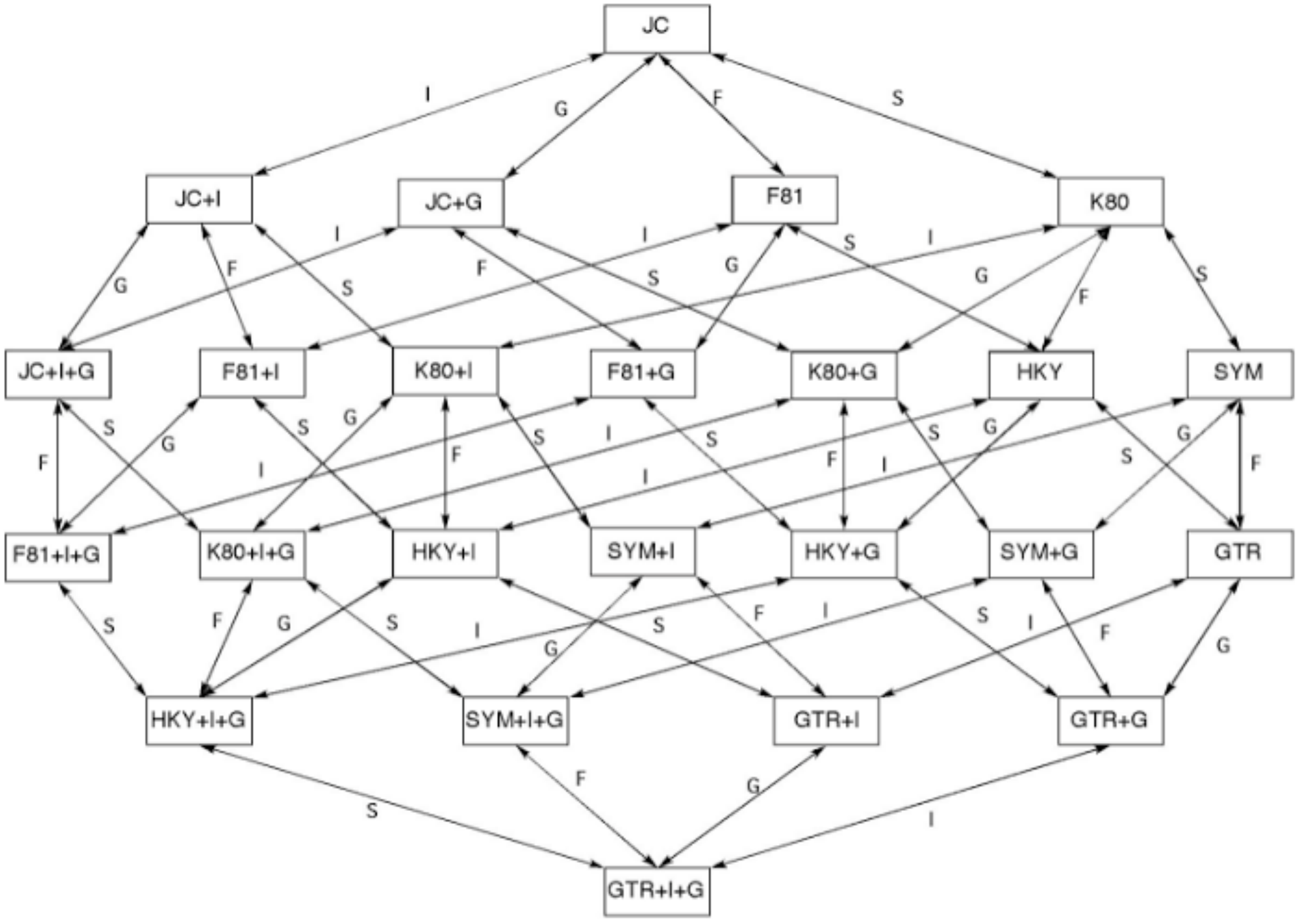


Difference in free parameters

If the simpler model is true, twice the difference in log-likelihoods between the true and more complex model will follow a **Chi-squared** distribution with d.f. = the difference in complexity between the models.

Only for **nested** models (simple = restriction of complex)

Hierarchy of Phylogenetic Models



jModelTest Manual

For phylogenetic model comparisons, need to use the same tree topology!

No.	Model	−LnL	df	AIC	AICc	BIC
1	GTR+F	121805.443	37	243684.886	243685.021	243978.802
2	GTR+F+I	116709.062	38	233494.125	233494.268	233795.984
3	GTR+F+G4	116297.301	38	232670.602	232670.745	232972.462
4	GTR+F+I+G4	116239.427	39	232556.854	232557.004	232866.657
8	SYM+I+G4	116442.995	36	232957.991	232958.119	233243.963
12	TVM+F+I+G4	116387.972	38	232851.945	232852.087	233153.804
16	TVMe+I+G4	116491.447	35	233052.895	233053.016	233330.923
20	TIM3+F+I+G4	116266.123	37	232606.247	232606.382	232900.162
24	TIM3e+I+G4	116539.014	34	233146.029	233146.143	233416.113
28	TIM2+F+I+G4	116294.540	37	232663.081	232663.216	232956.997
32	TIM2e+I+G4	116457.329	34	232982.659	232982.773	233252.744
36	TIM+F+I+G4	116319.406	37	232712.813	232712.948	233006.729
40	TIMe+I+G4	116552.585	34	233173.170	233173.285	233443.255
44	TPM3u+F+I+G4	116411.915	36	232895.829	232895.957	233181.801
48	TPM3+I+G4	116588.326	33	233242.652	233242.760	233504.793
52	TPM2u+F+I+G4	116436.717	36	232945.434	232945.562	233231.406
56	TPM2+I+G4	116505.262	33	233076.524	233076.632	233338.665
60	K3Pu+F+I+G4	116458.351	36	232988.703	232988.831	233274.675
64	K3P+I+G4	116601.248	33	233268.496	233268.604	233530.637
68	TN+F+I+G4	116322.091	36	232716.182	232716.310	233002.154
72	TNe+I+G4	116553.614	33	233173.228	233173.336	233435.369
76	HKY+F+I+G4	116460.979	35	232991.958	232992.079	233269.986
80	K2P+I+G4	116602.466	32	233268.933	233269.034	233523.130
84	F81+F+I+G4	120265.516	34	240599.033	240599.147	240869.117
88	JC+I+G4	120344.412	31	240750.823	240750.919	240997.077

Akaike Information Criterion: GTR+F+I+G4
Corrected Akaike Information Criterion: GTR+F+I+G4
Bayesian Information Criterion: GTR+F+I+G4
Best-fit model: GTR+F+I+G4 chosen according to BIC

Bayesian Model Selection

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

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

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

Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

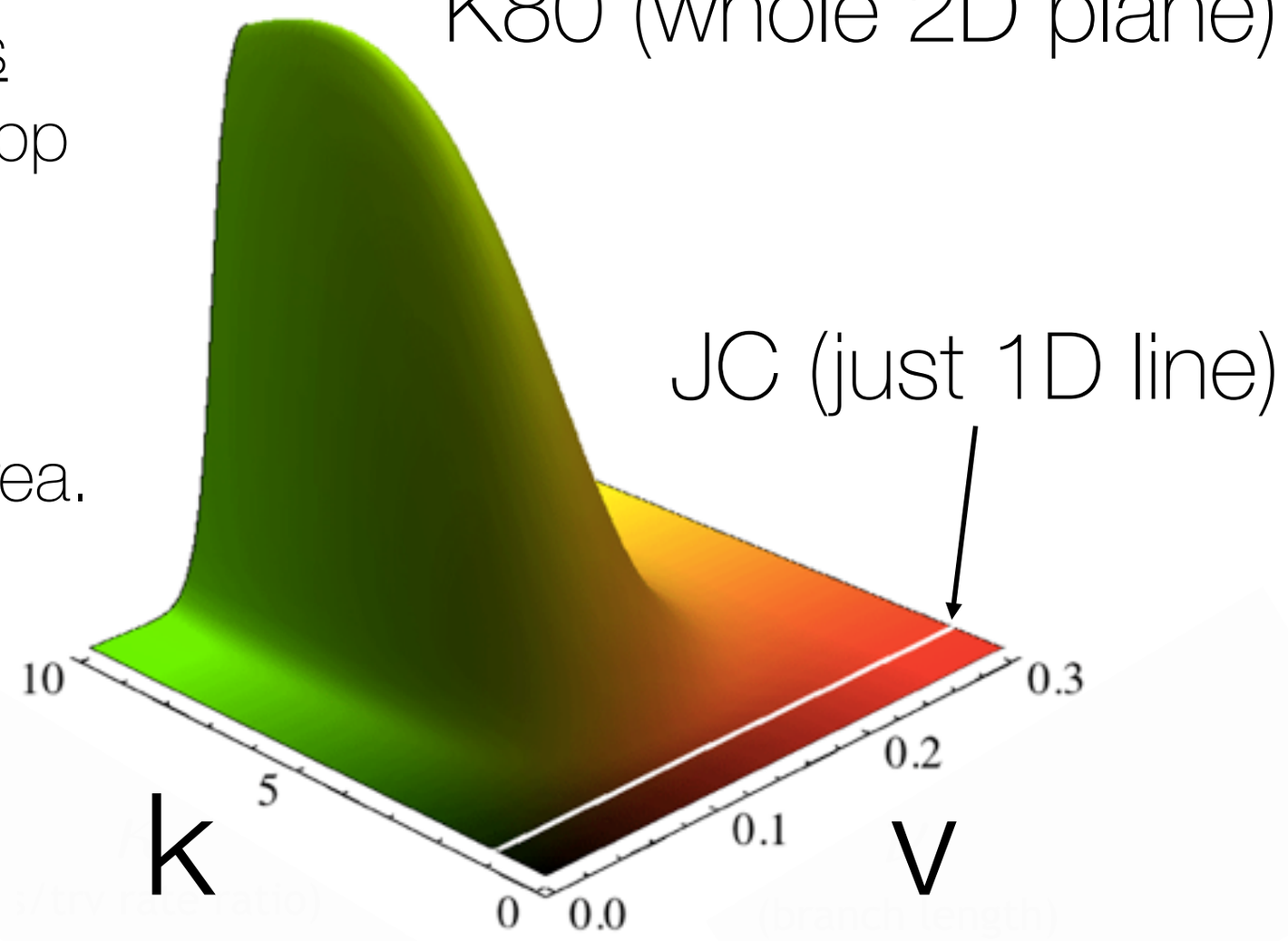
<u>Simulation Conditions</u>

Sequence length: 500 bp

True v: 0.15

True k: **5.0**

Prior is flat over whole area.

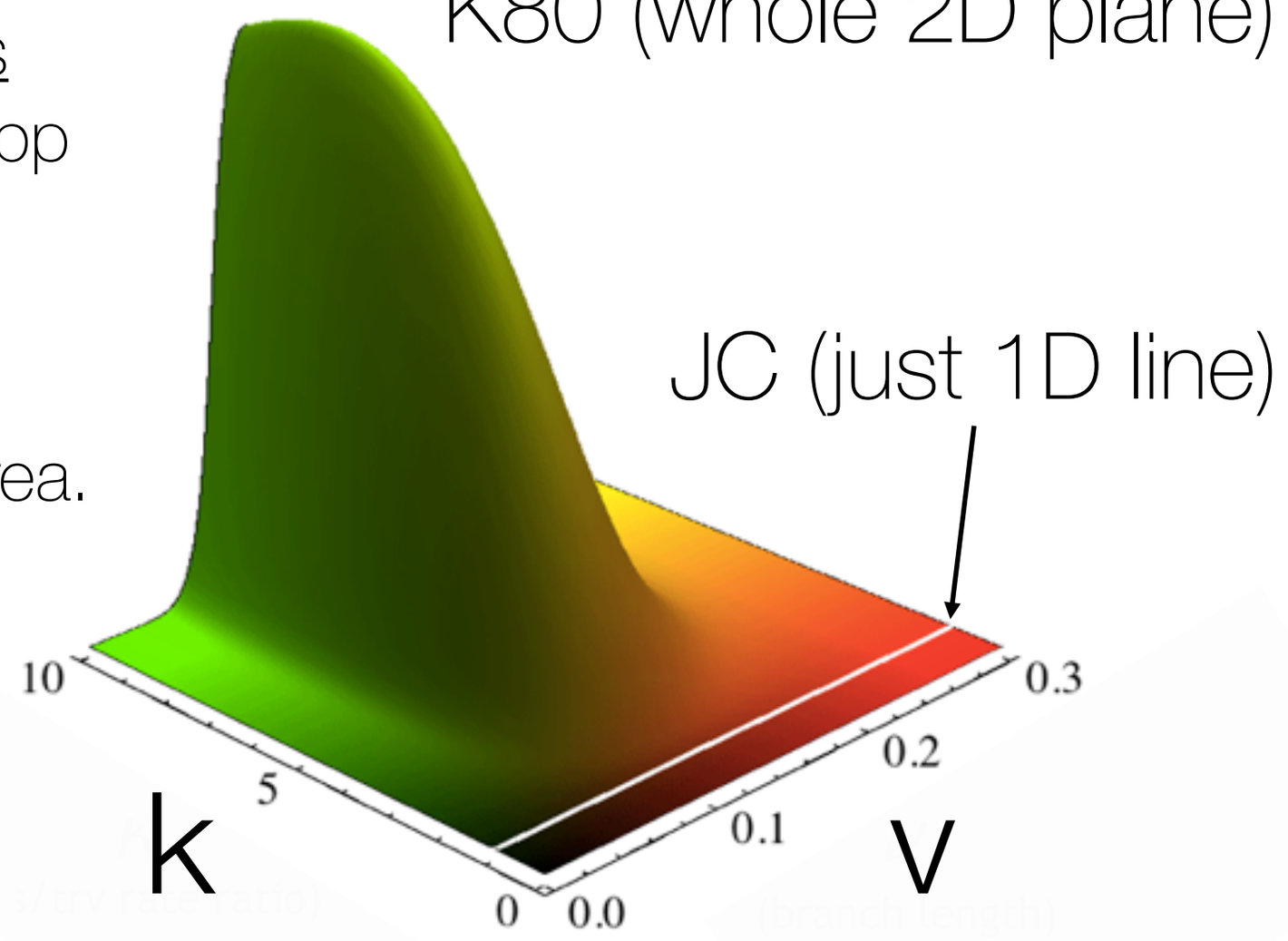


Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

<u>Simulation Conditions</u>
Sequence length: 500 bp
True v: 0.15
True k: **5.0**

Prior is flat over whole area.



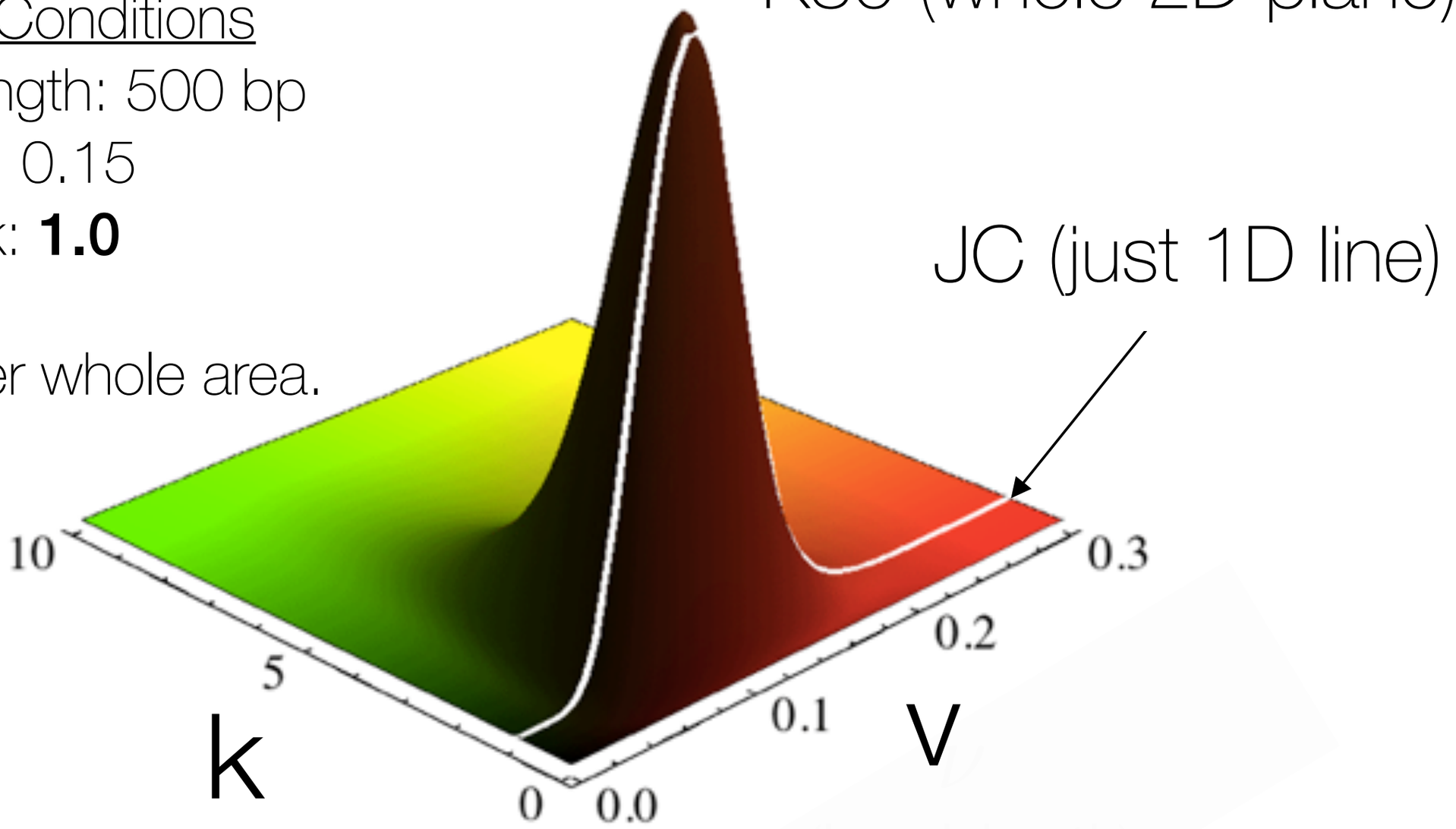
K80 wins!

Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

<u>Simulation Conditions</u>
Sequence length: 500 bp
True v: 0.15
True k: **1.0**

Prior is flat over whole area.



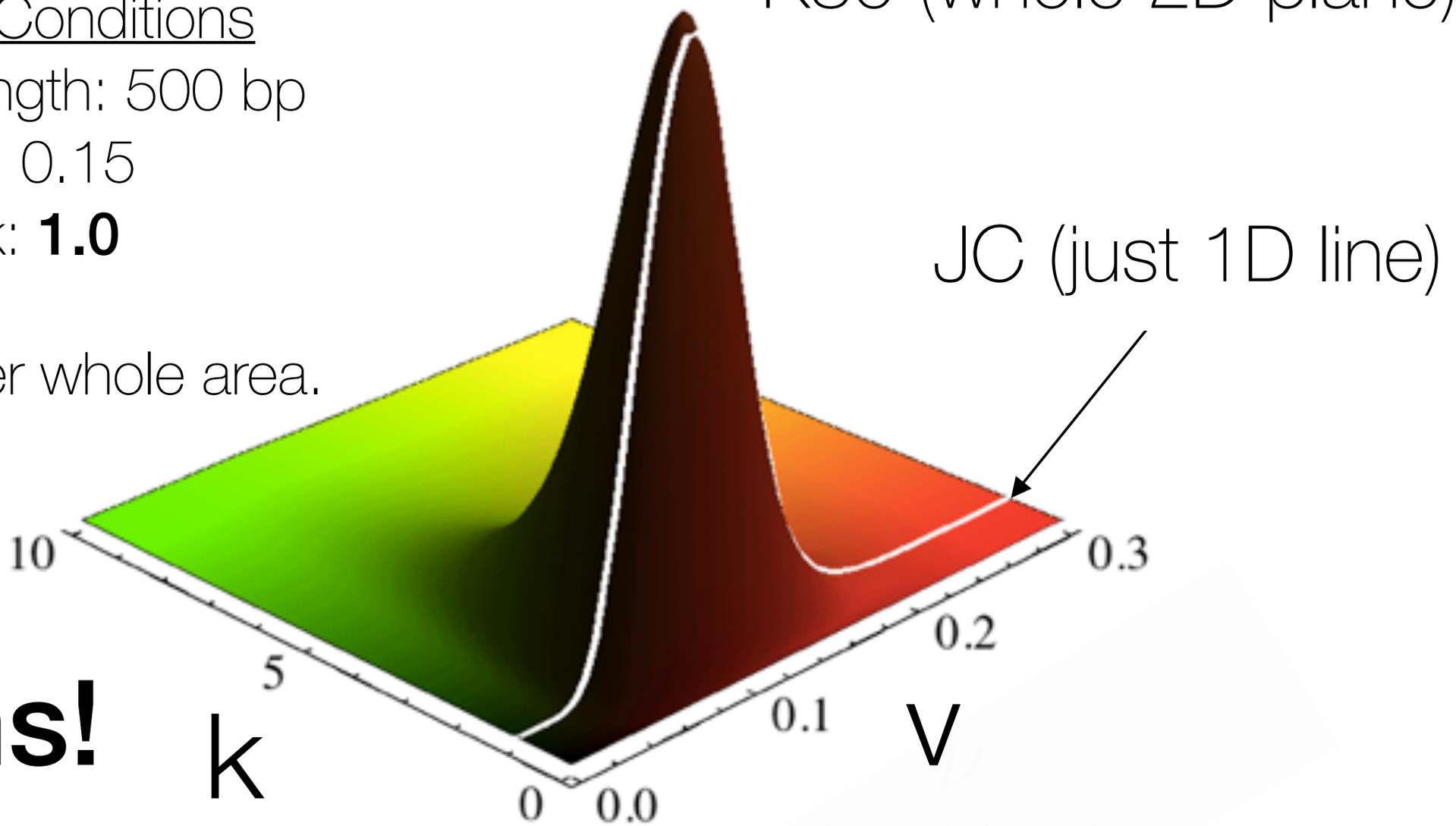
Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

<u>Simulation Conditions</u>
Sequence length: 500 bp
True v: 0.15
True k: **1.0**

Prior is flat over whole area.

JC wins!



Example taken from Paul Lewis - Workshop on Molecular Evolution 2016

Marginal Likelihood Example

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

More complicated models can have lower marginal likelihoods.

The Bayes Factor

$$K = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|\theta, M_1)P(\theta|M_1)d\theta}{\int P(D|\theta, M_2)P(\theta|M_2)d\theta}$$

Ratio of the probability of the data under two models

Note that this is related to the LRT

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

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

Interpreting Bayes Factors

Strength of evidence	BF(M_0,M_1)**	log(BF(M_0,M_1))	$log_{10}(BF(M_0,M_1))$
Negative (supports M_1)	< 1	< 0	< 0
Barely worth mentioning	1 to 3.2	0 to 1.16	0 to 0.5
Substantial	3.2 to 10	1.16 to 2.3	0.5 to 1
Strong	10 to 100	2.3 to 4.6	1 to 2
Decisive	> 100	> 4.6	> 2

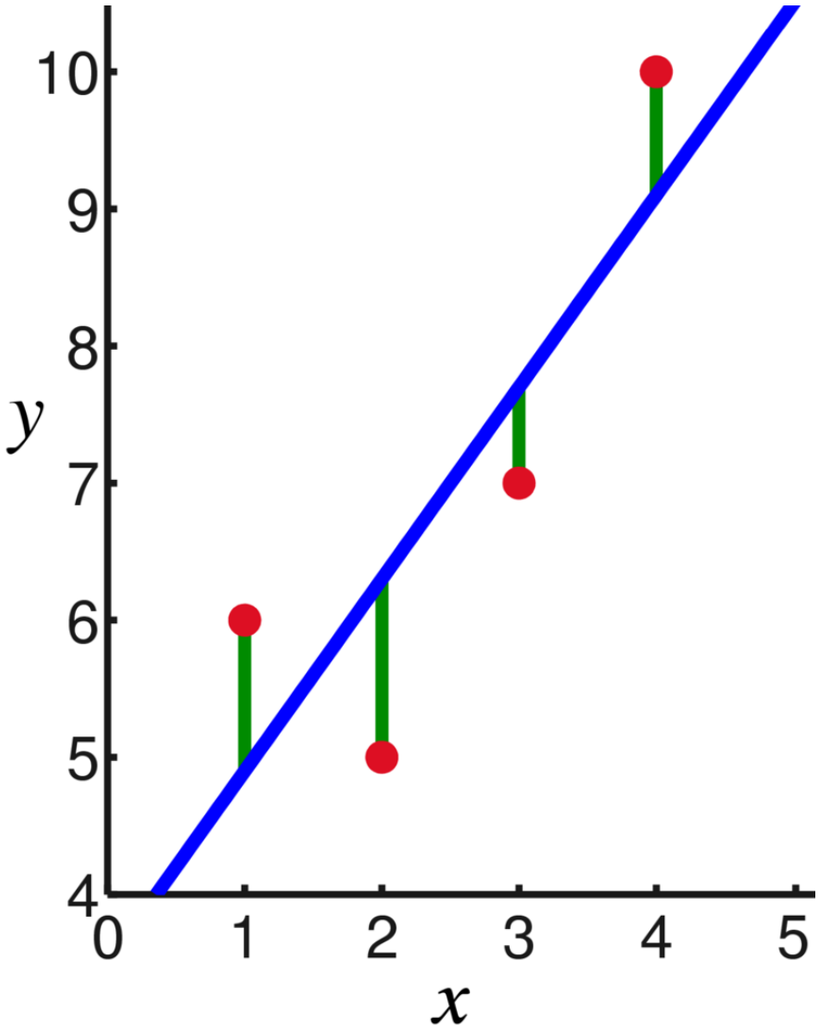
https://revbayes.github.io/tutorials/model_selection_bayes_factors/bf_intro.html

Now we have tools to select the “best” or most useful model from our set…

Now we have tools to select the “best” or most useful model from our set…

…but does that model capture the important features of our data well?

Assessing Model Fit is Standard Practice in Many Areas



Standard Linear Regression Assumptions

Weak exogeneity
(predictor variables have no error)

Linearity

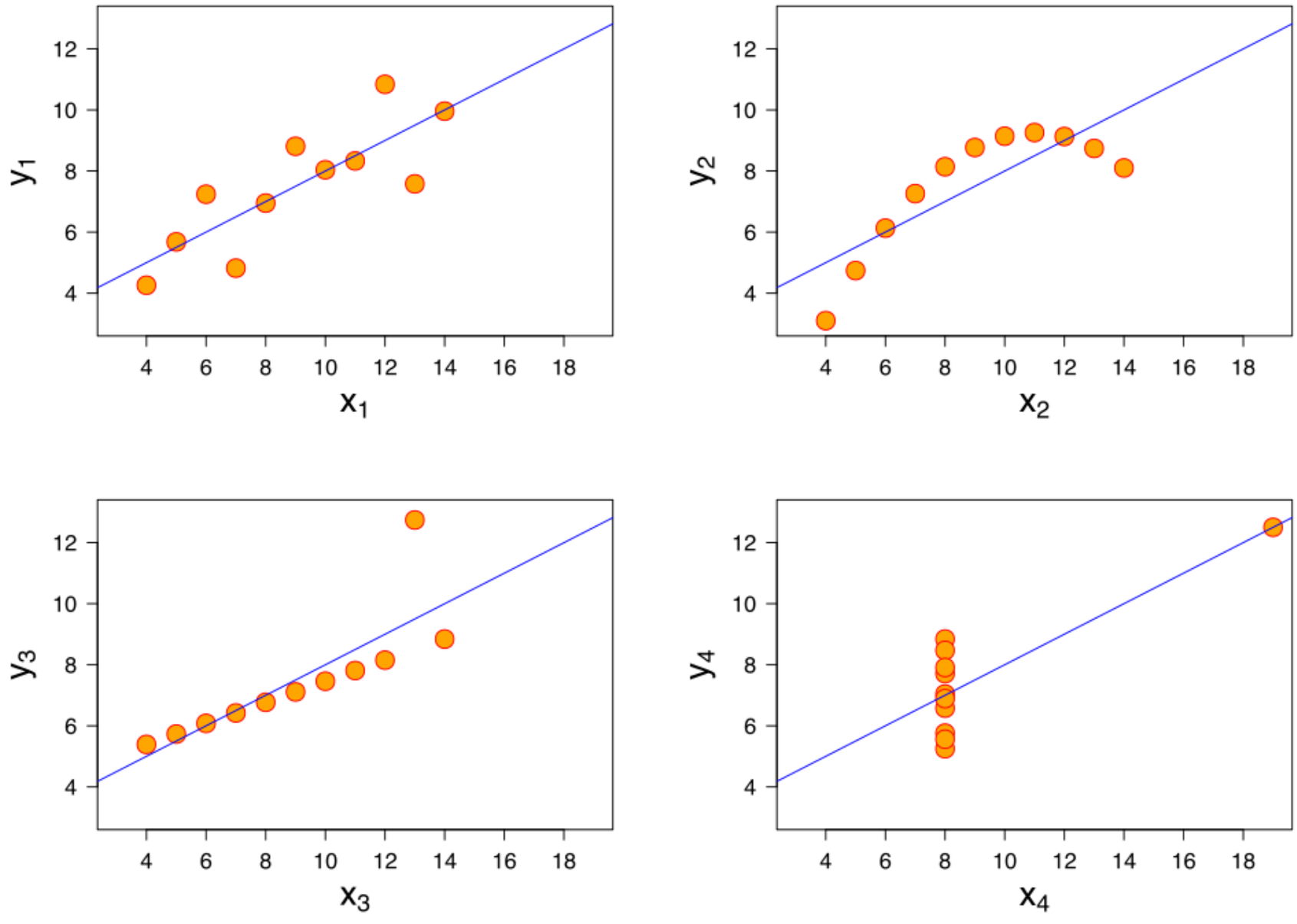
Constant variance

Independent errors
(phylogenetic comparative methods)

Figure and assumptions from: https://en.wikipedia.org/wiki/Linear_regression

Being Able to Fit a Model Does **Not** Mean That it Fits Well

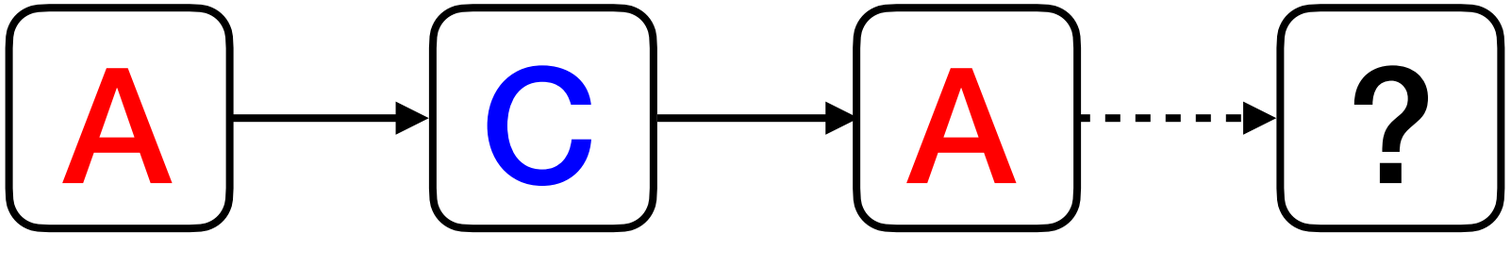
Anscombe's Quartet



From https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Common Phylogenetic Assumptions

(Homogenous GTR-class)



Evolution is **Markovian**

Sites Evolve **Independently**

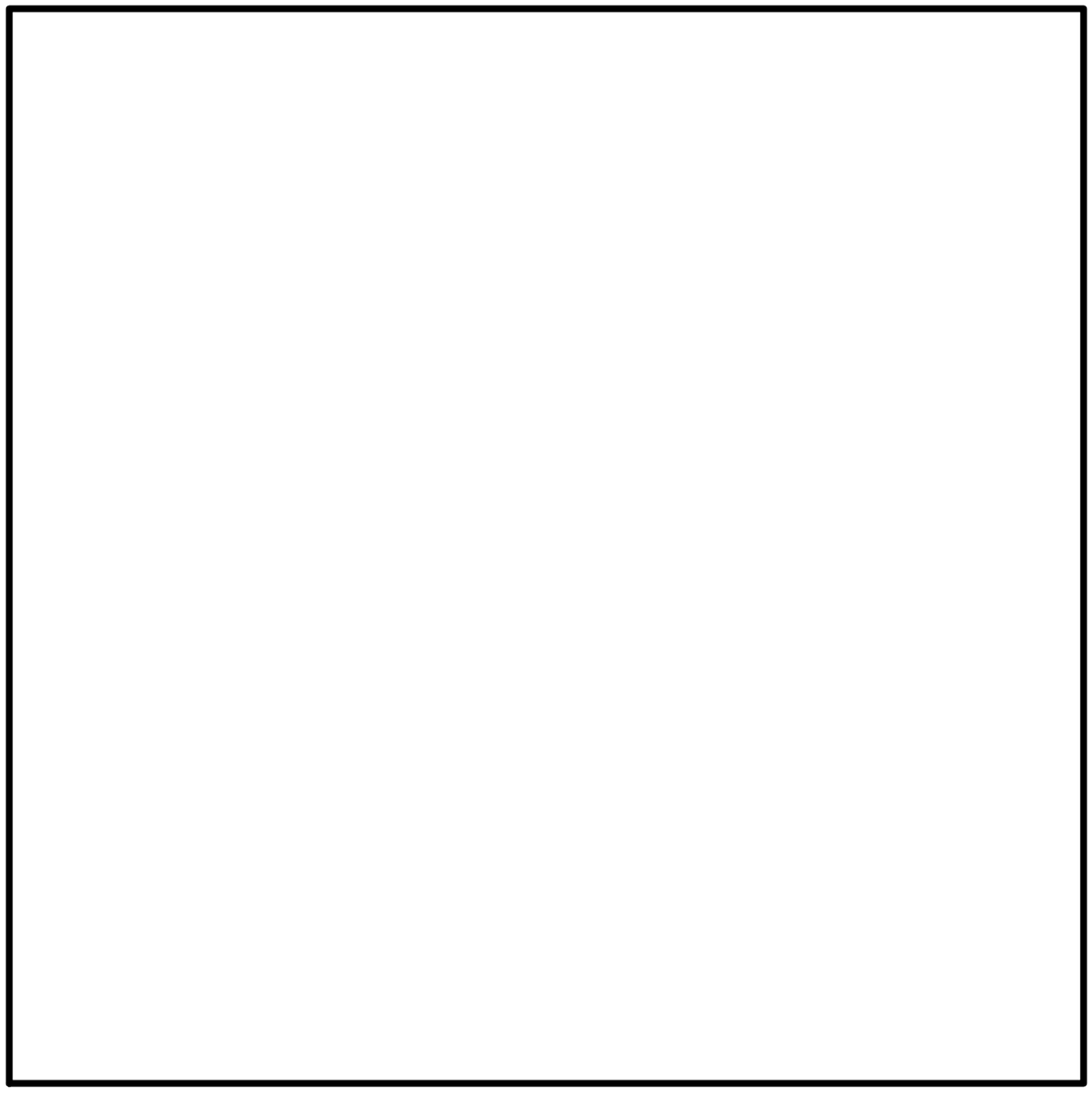
Base Frequencies Constant Across Taxa

(Groups of) Sites Evolve Under **Same Dynamics**

Sites Do Not Change How They **Evolve Across a Tree**

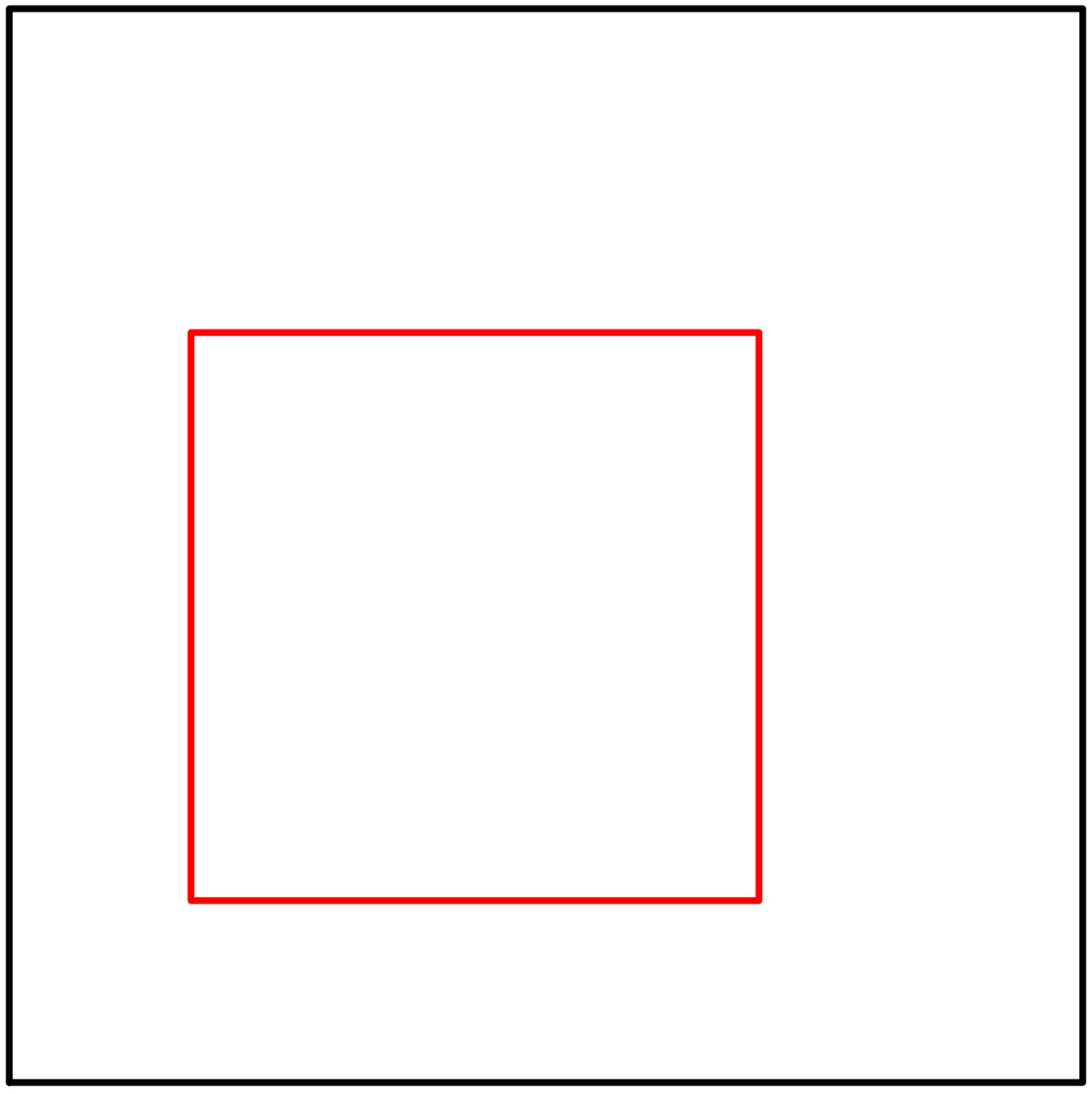
Model Space (conceptual)

All Possible Ways Genes Could Evolve



Model Space (conceptual)

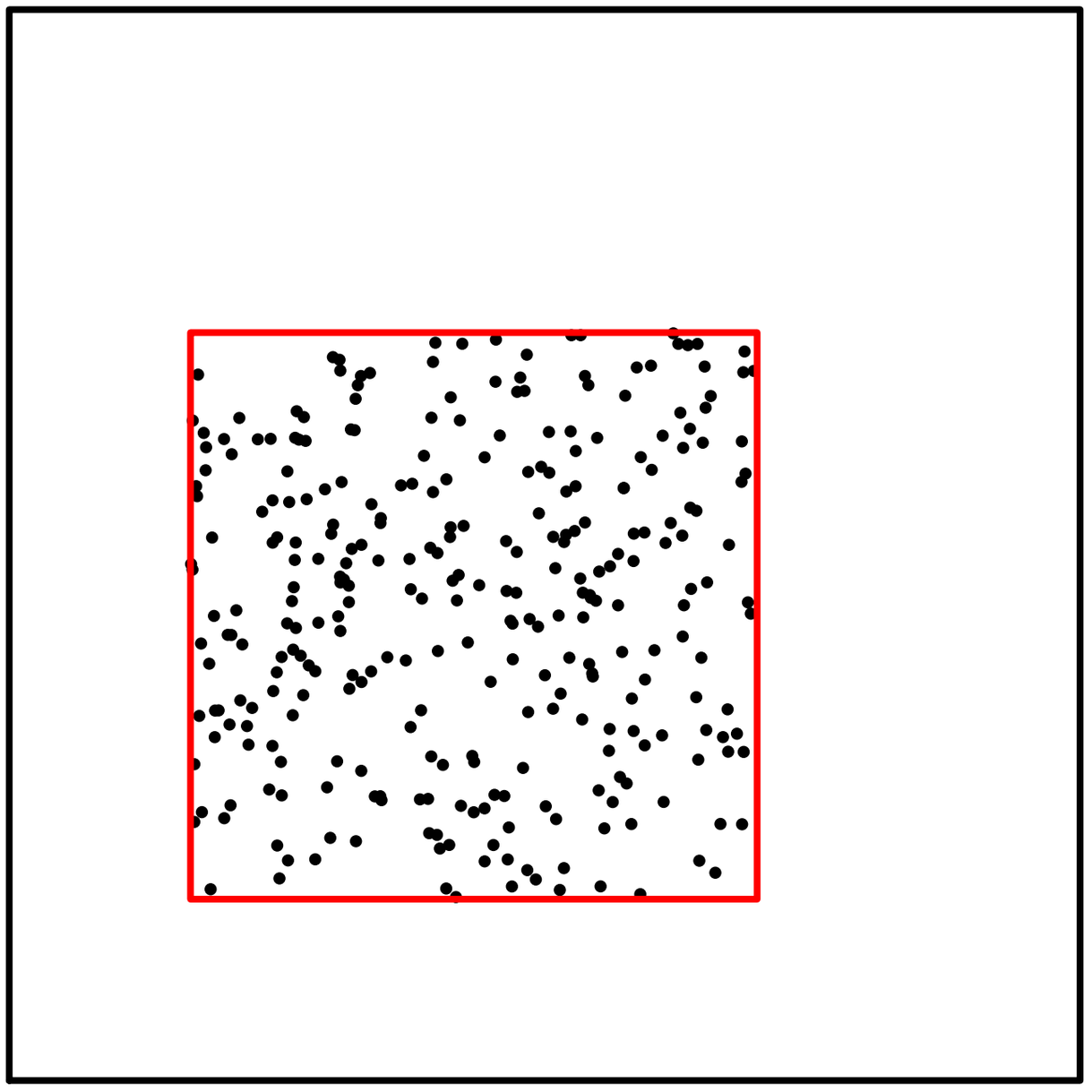
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Our Hope

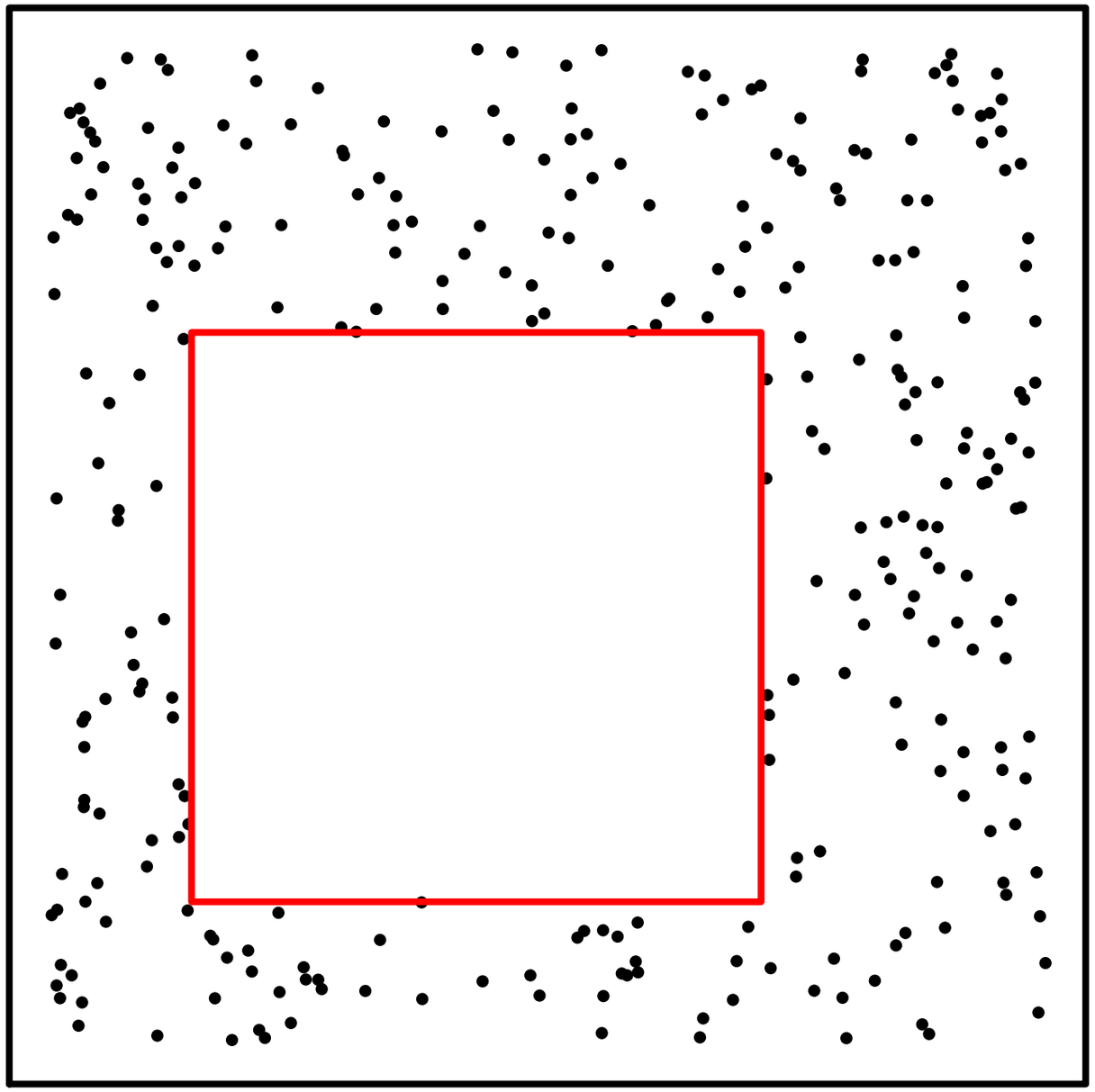
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Our Fears

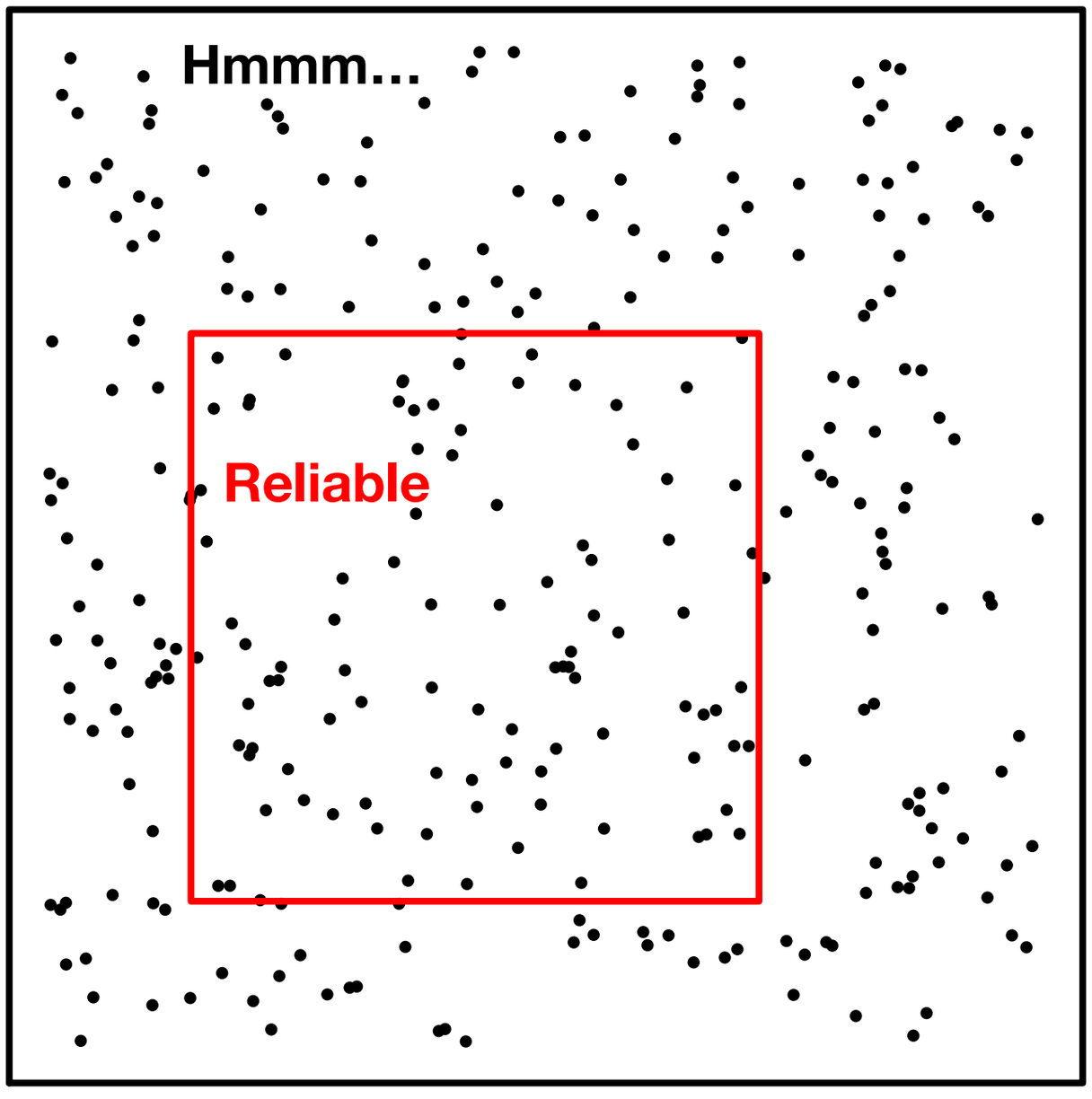
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

The Probable Truth

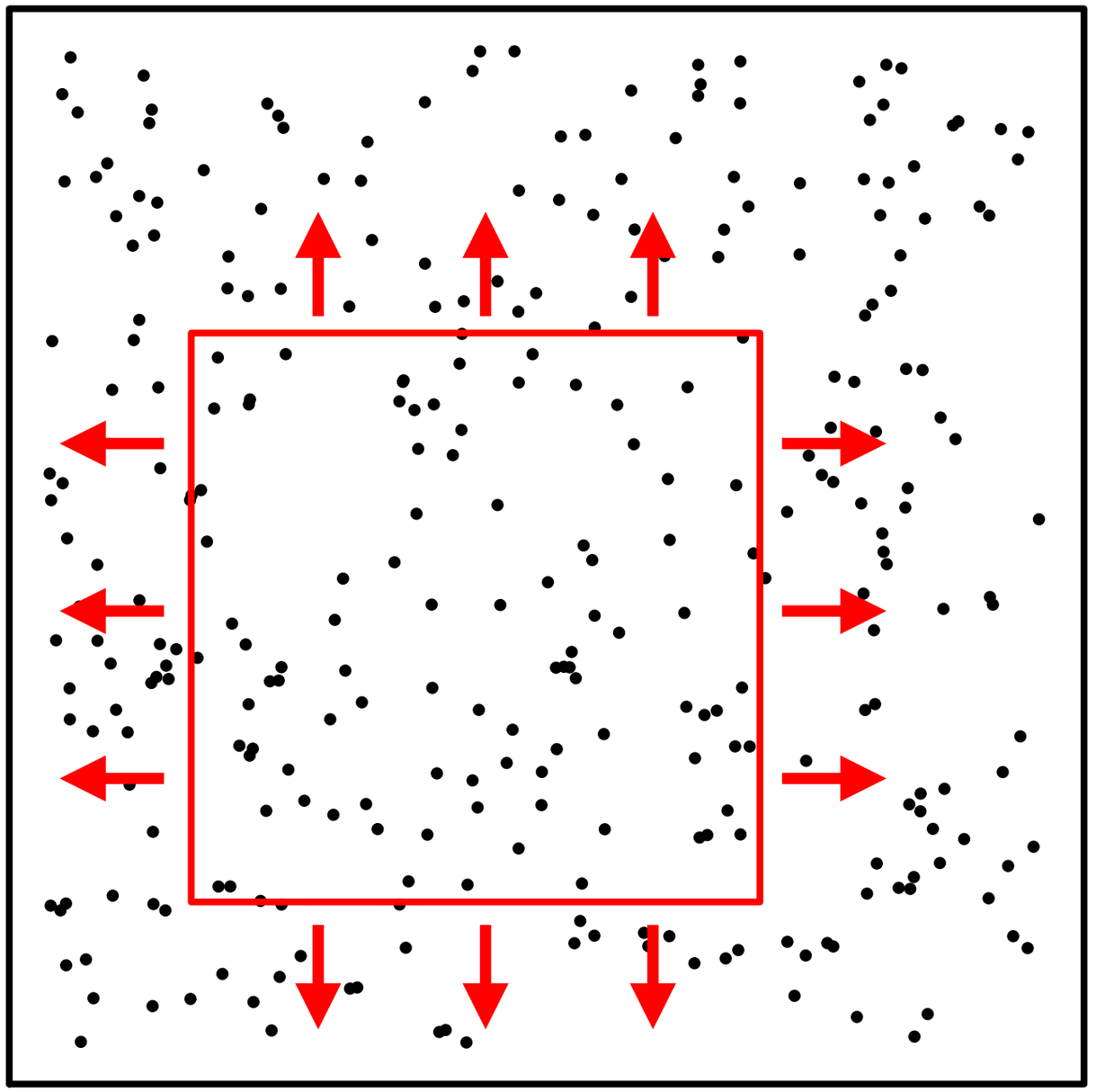
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Model Development

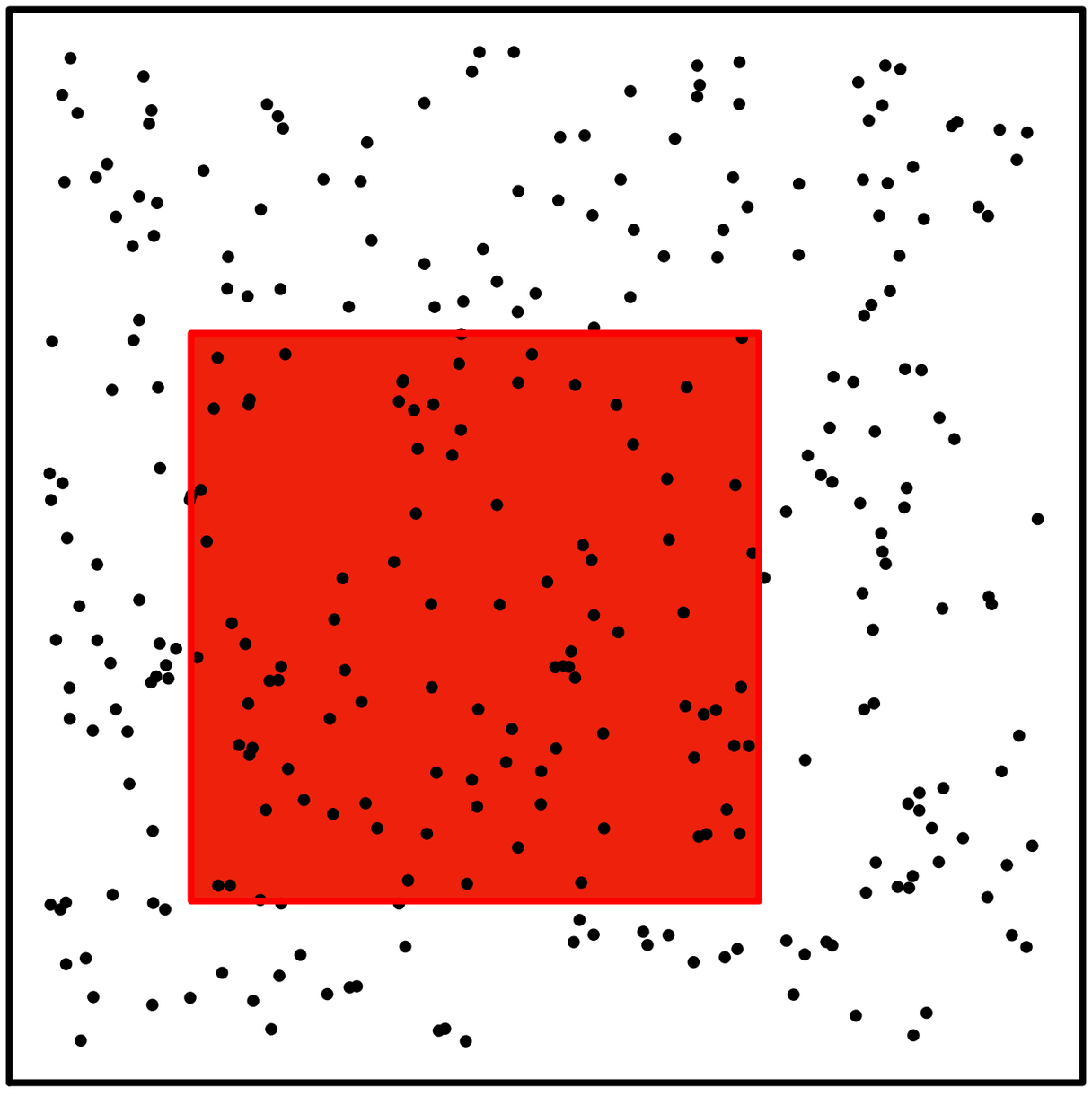
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Knowing When We're Doing Well

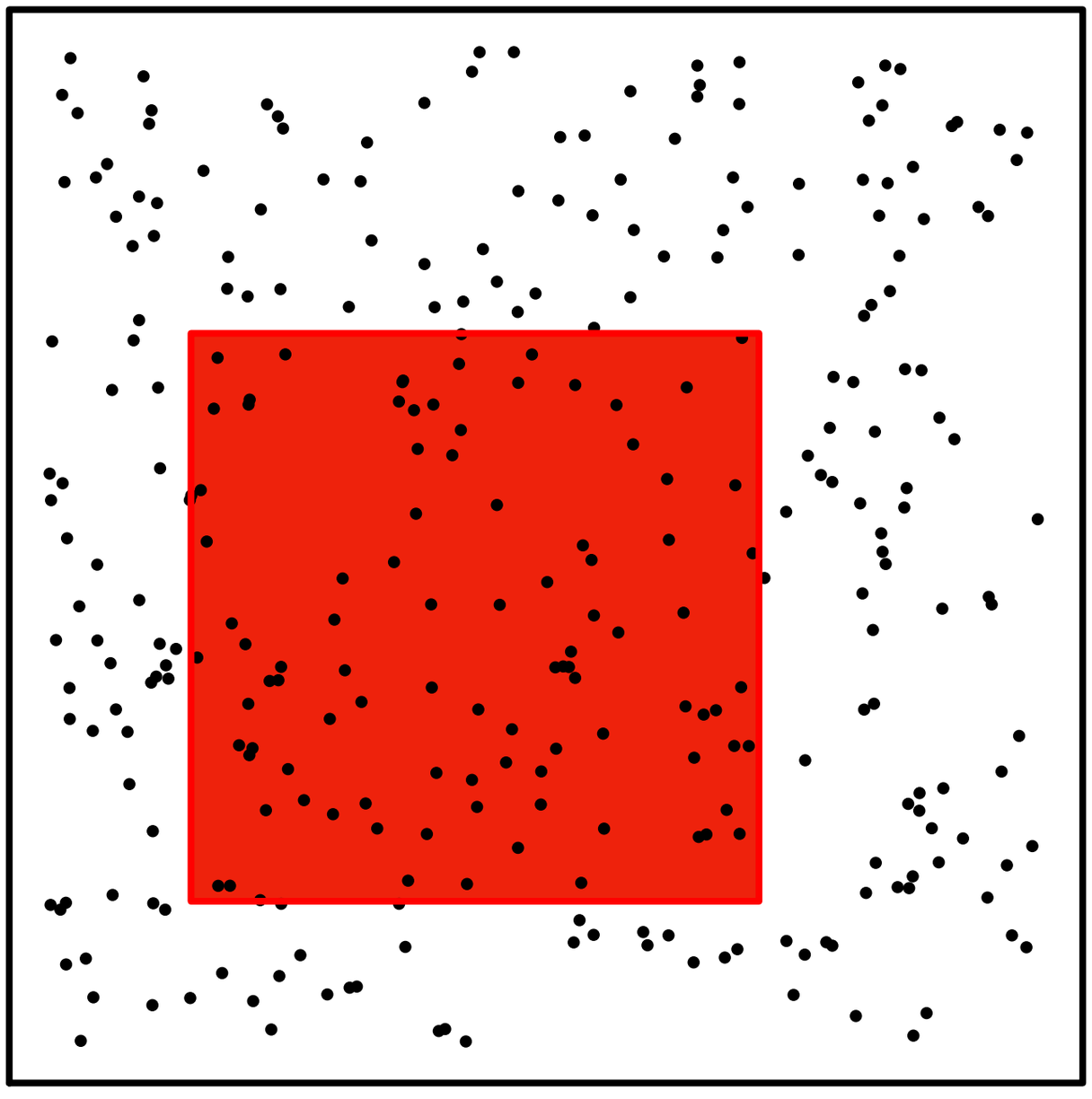
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Absolute Model Fit

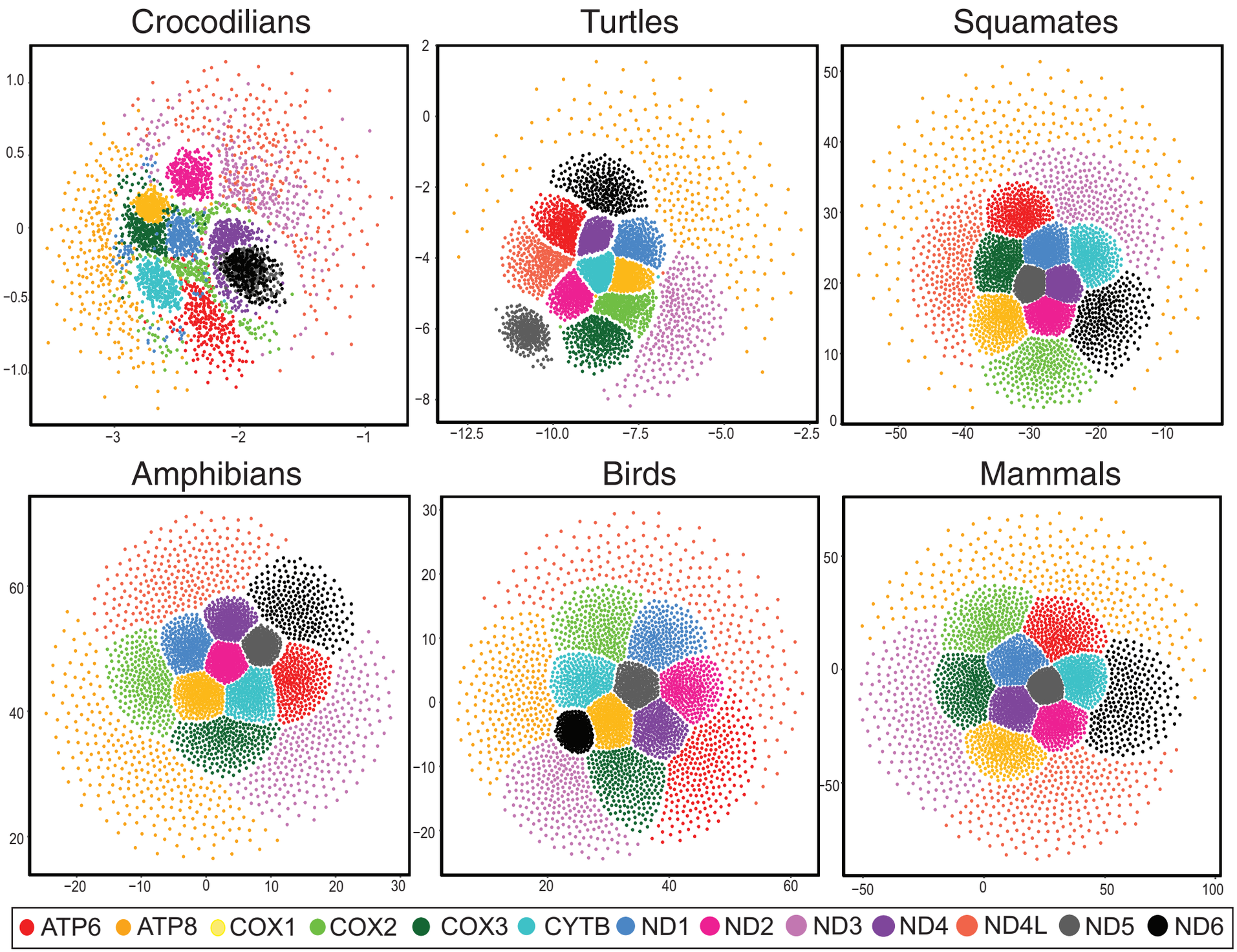
All Possible Ways Genes Could Evolve



The Types of Evolution That We Model Well

Conflict Within the Mitochondrion

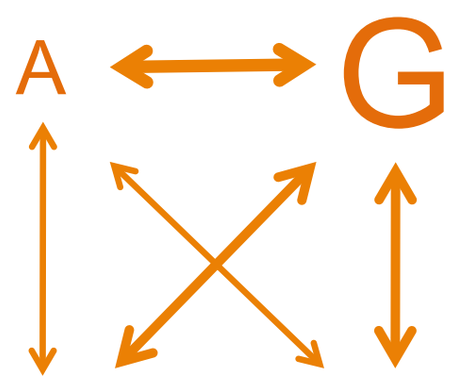
Because mt genes are linked, we **expect inferred gene trees to be the same.**

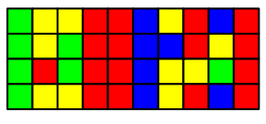
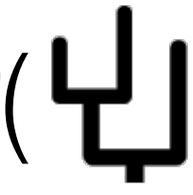
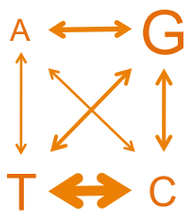
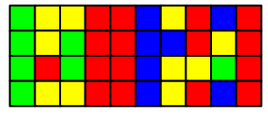


Richards, Brown, Barley, Chong, Thomson. 2018. Systematic Biology.

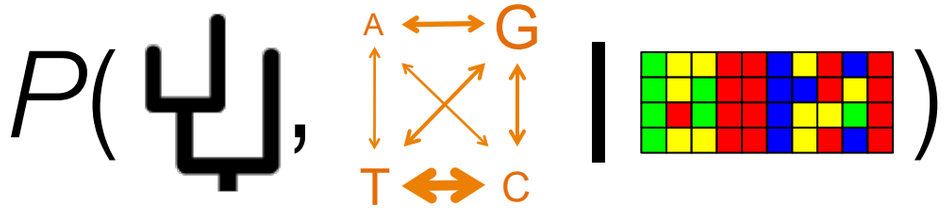
Phylogenetic Posterior Prediction

Could the model plausibly have generated the data?

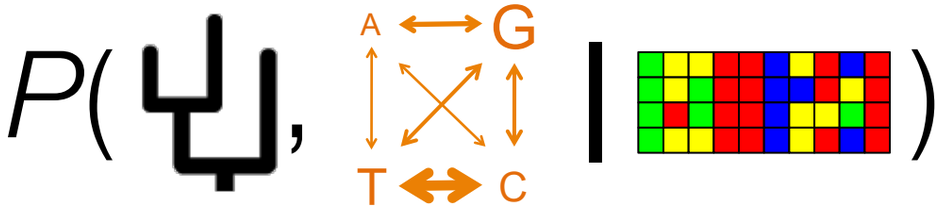


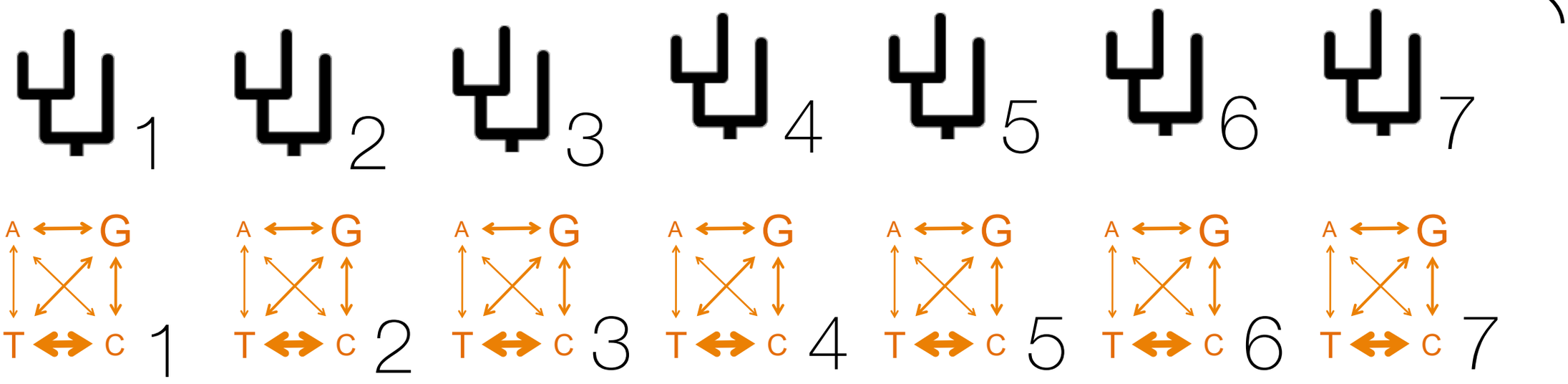
Could  have come from $P($  ,  |  $)$?

Posterior Predictive **Simulation**

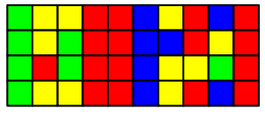


Posterior Predictive **Simulation**

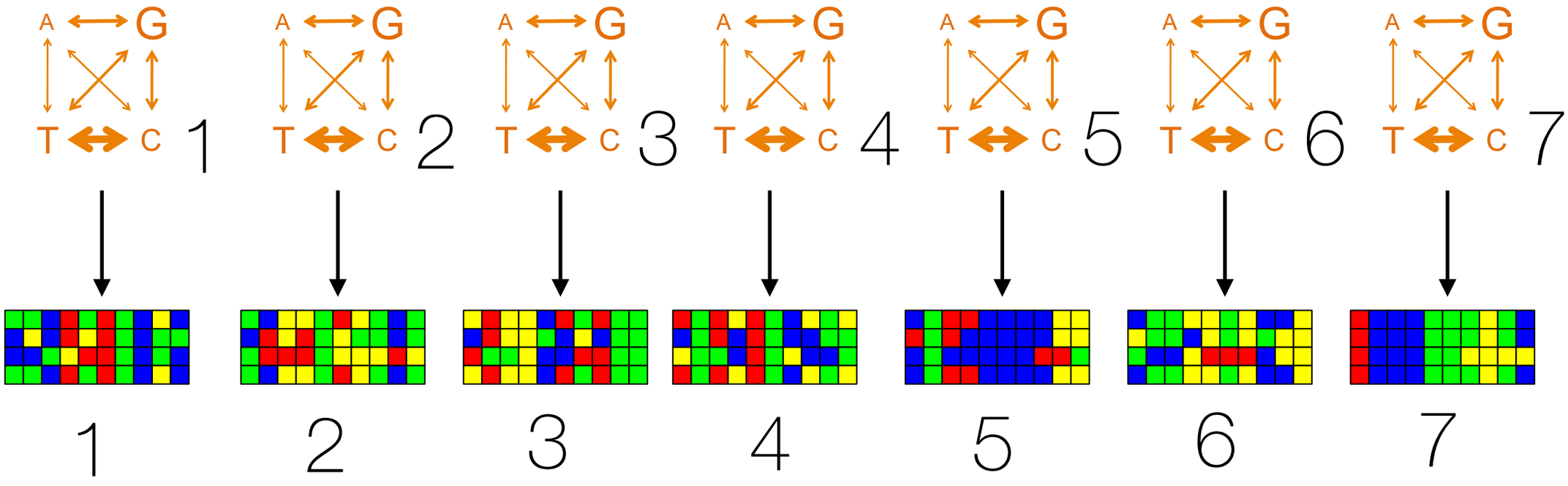




Posterior Predictive **Simulation**

$P(\Psi, \; \mathrm{A}\leftrightarrow\mathrm{G},\; \mathrm{T}\leftrightarrow\mathrm{C} \;|\; \text{  })$

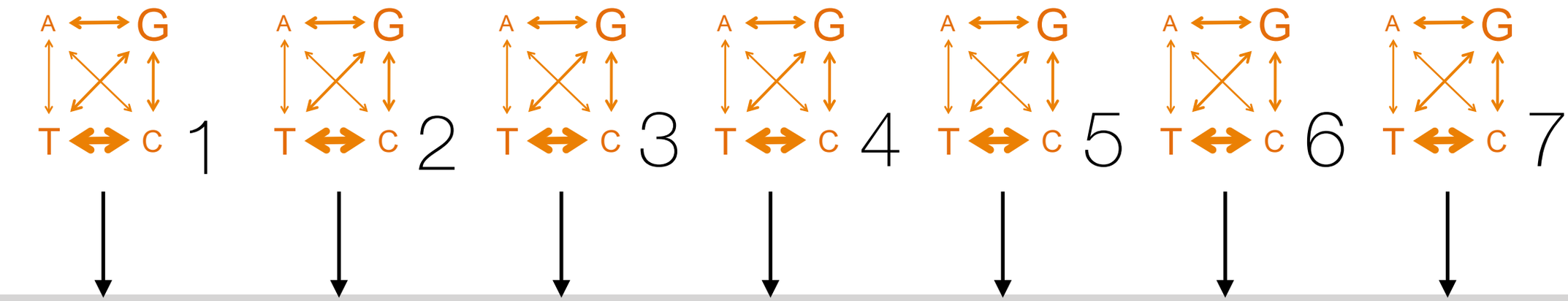
$\Psi_1 \; \Psi_2 \; \Psi_3 \; \Psi_4 \; \Psi_5 \; \Psi_6 \; \Psi_7$



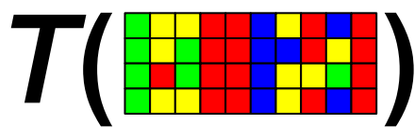
Posterior Predictive **Simulation**

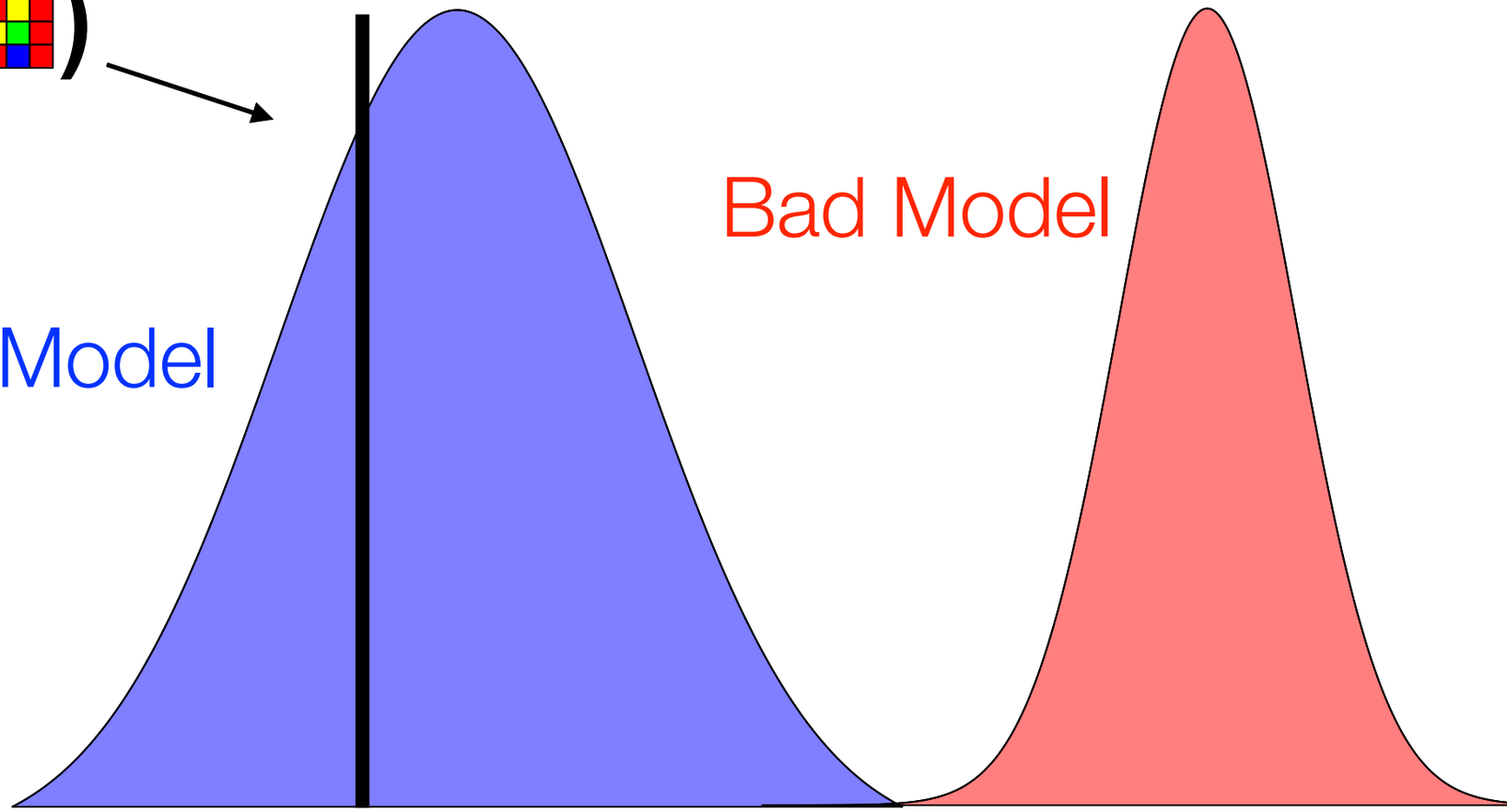
$P(\Psi, \theta \mid \text{data})$

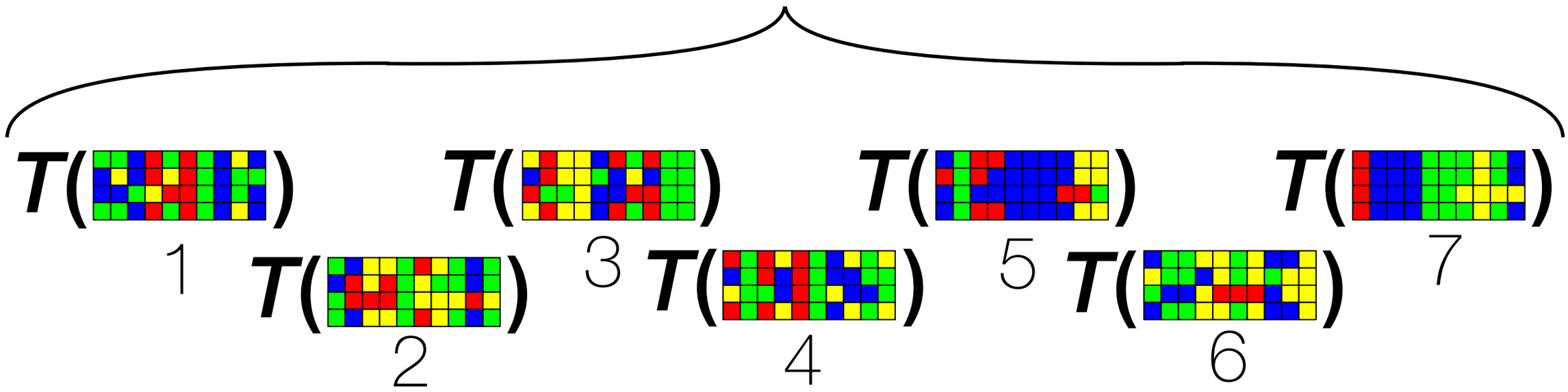
Ψ_1 Ψ_2 Ψ_3 Ψ_4 Ψ_5 Ψ_6 Ψ_7



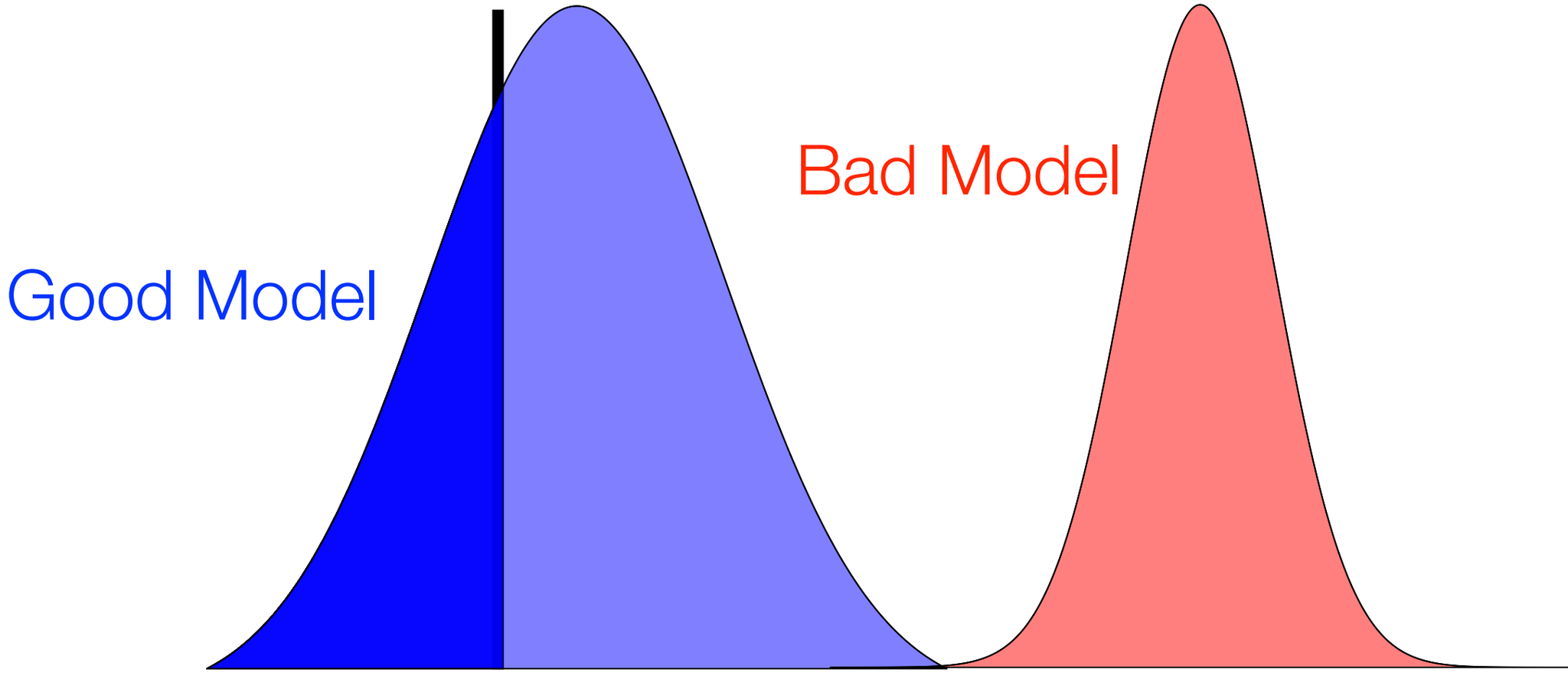
Posterior Predictive **Tests**





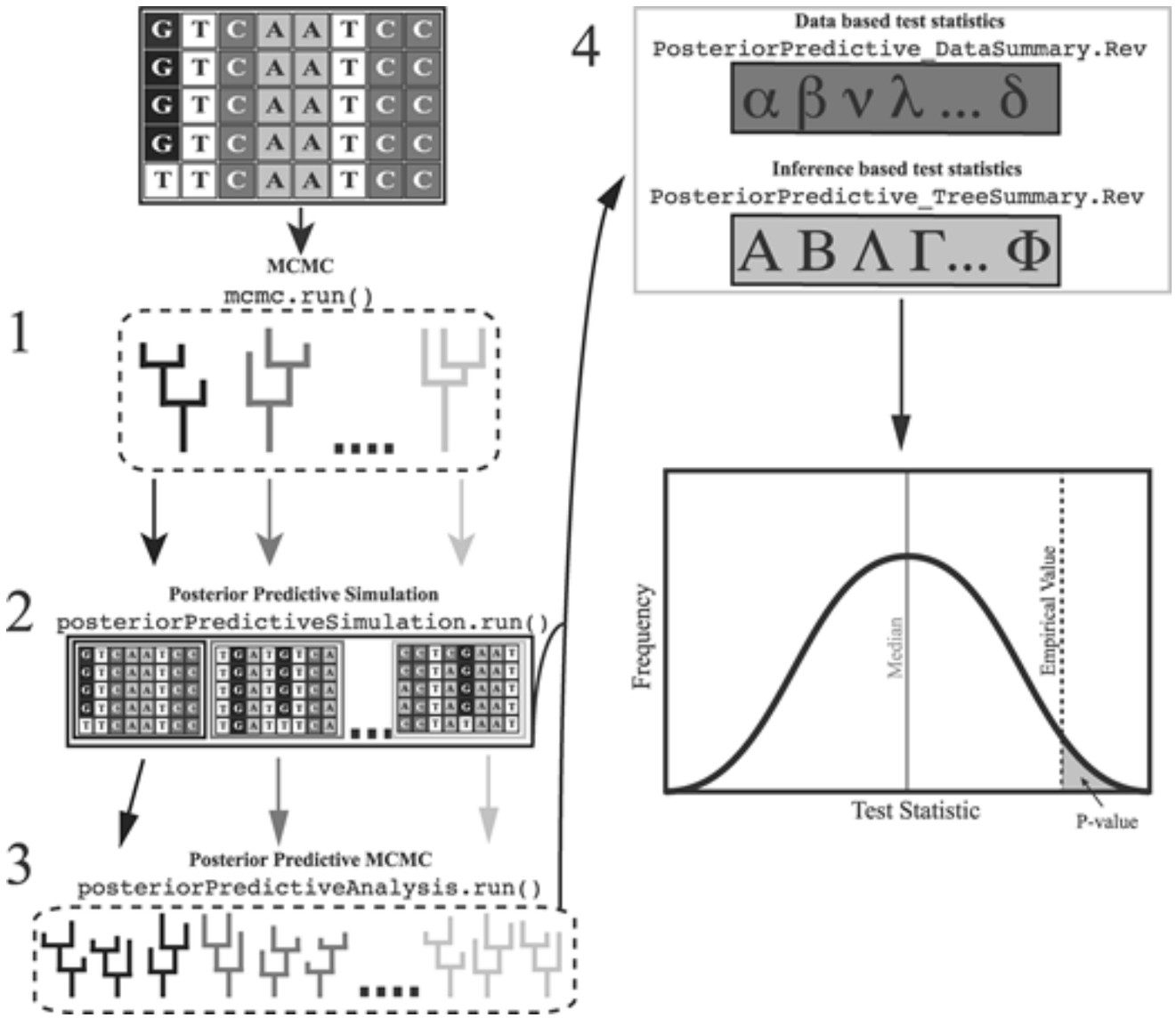


Posterior Predictive **P-Values**



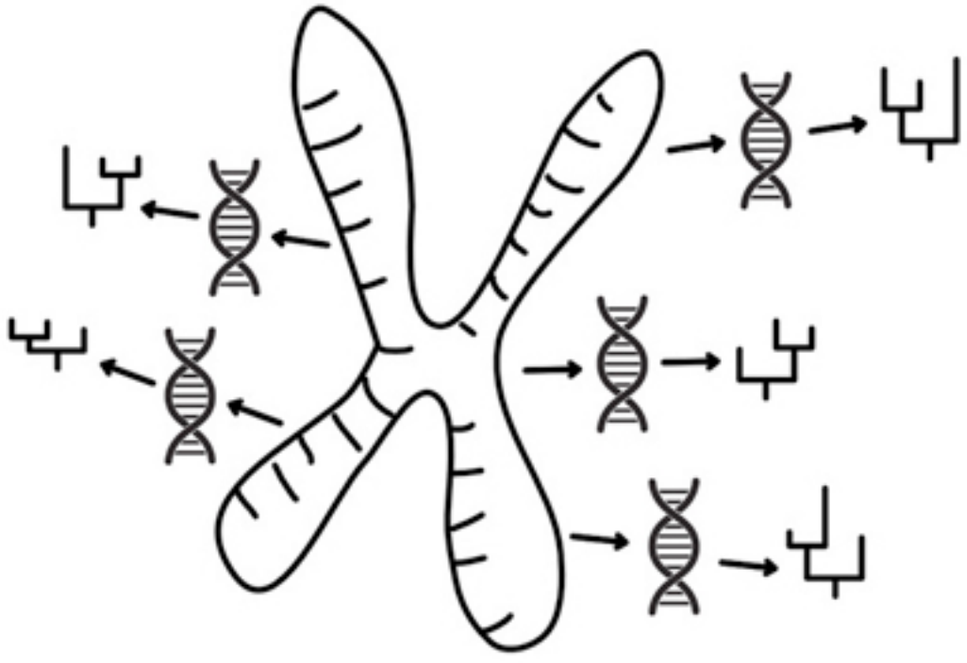
Phylogenetic Posterior Prediction (P^3) in **RevBayes**

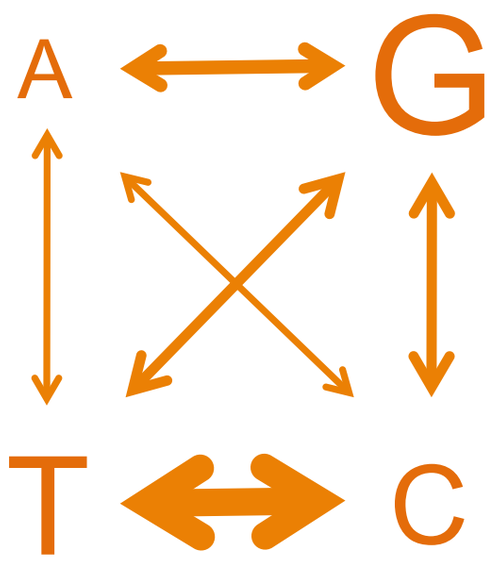
Pipeline **self-contained** and highly **customizable**.



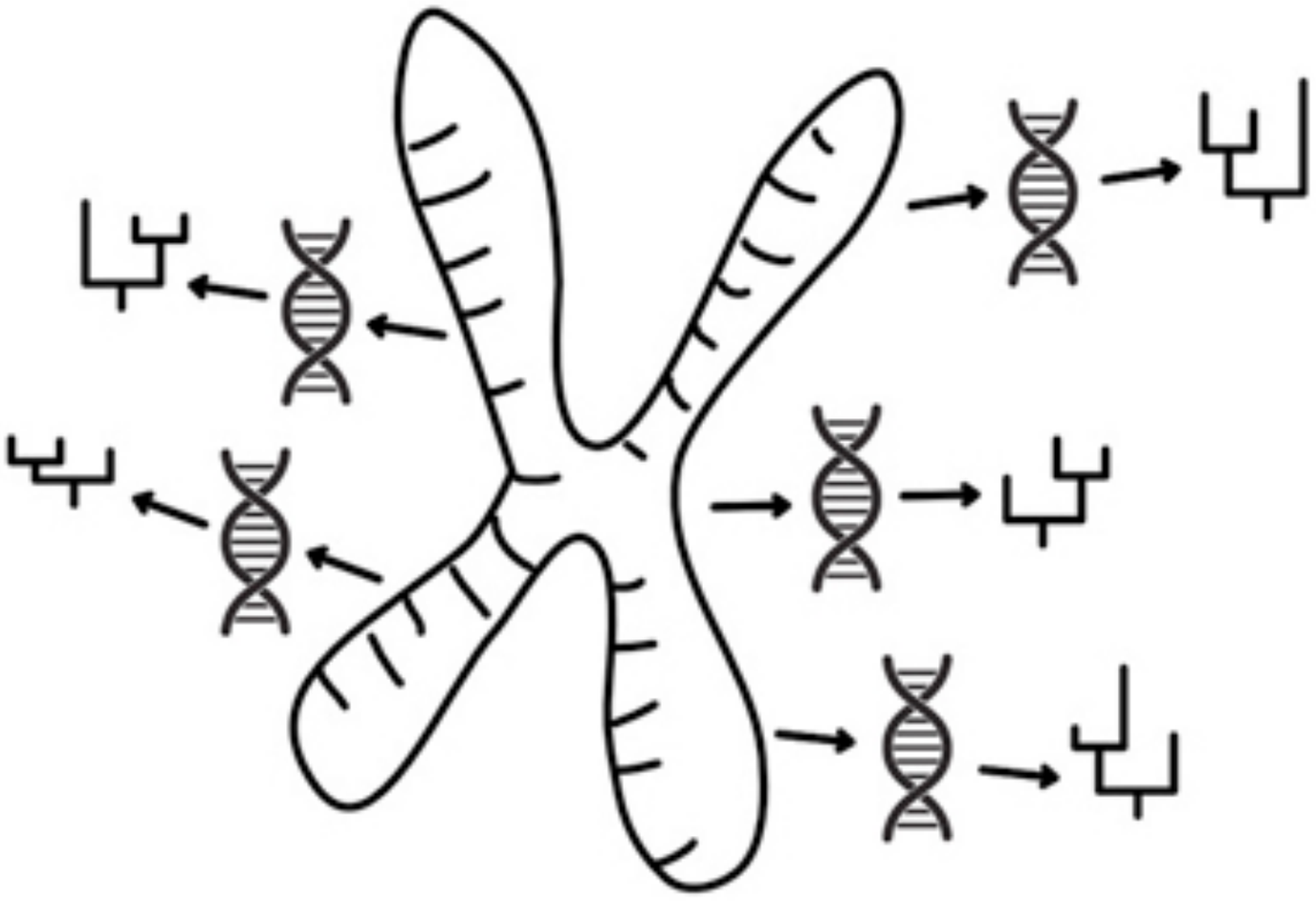
Höhna, Coghill, Mount, Thomson, Brown. 2018. Molecular Biology and Evolution.

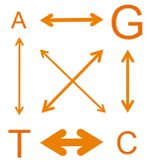
The pitfall of having genomic data and spending less time thinking about our models.





The pitfall of having genomic data and spending less time thinking about our models.





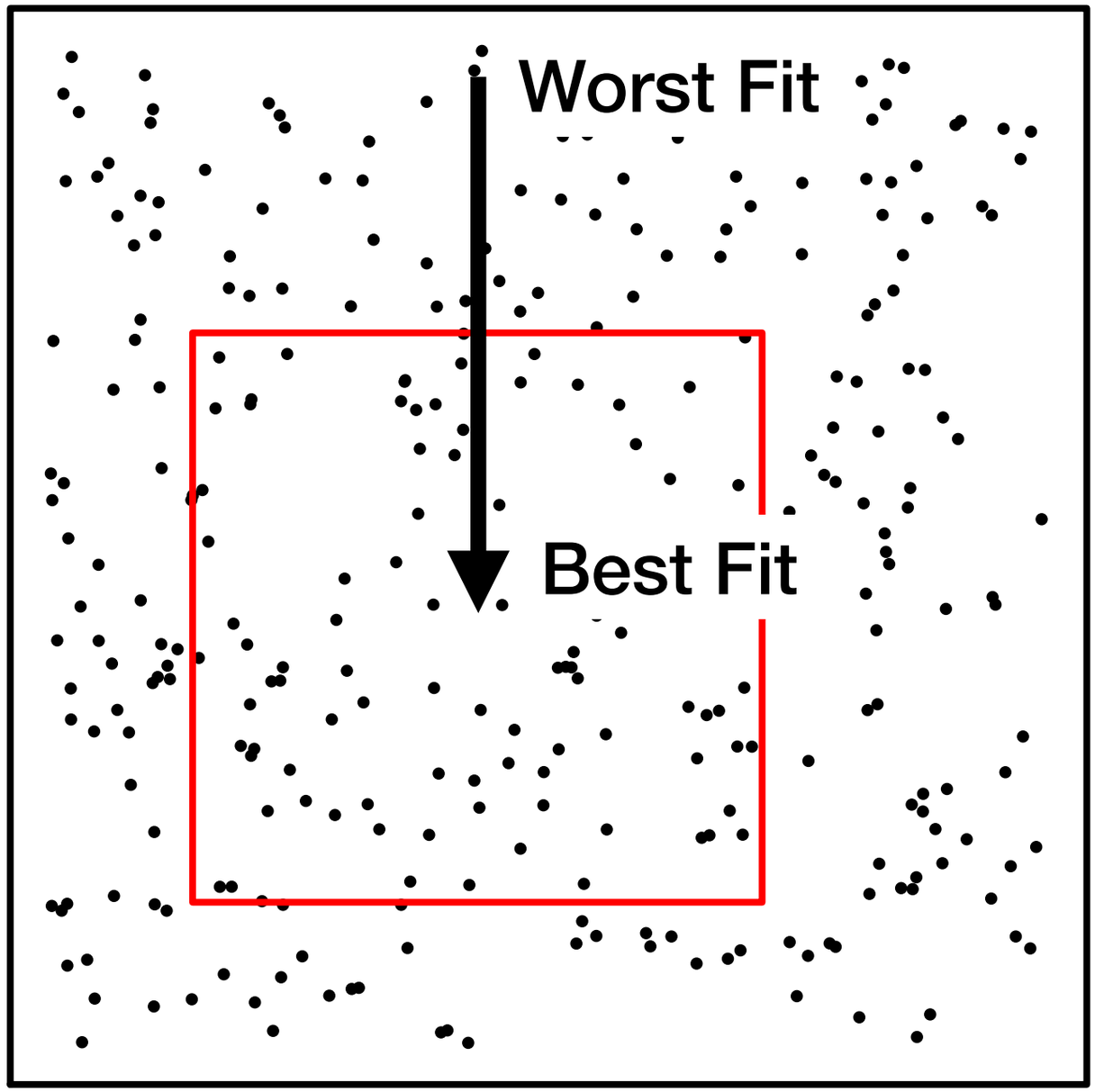
Posterior Prediction

microbewiki.kenyon.edu

Yeast
343 orthologs
18 taxa
Hess & Goldman (2011)

Amniotes
1,145 orthologs
10 taxa
Crawford et al. (2012)

phylopic.org



Doyle, Young, Naylor, Brown. 2015. Systematic Biology.

Posterior Prediction

Best Fit

Lowest
10%

DECILE 1
34 Genes

DECILE 2
34 Genes

DECILE 3
34 Genes

DECILE 4
34 Genes

DECILE 5
34 Genes

DECILE 6
34 Genes

DECILE 7
34 Genes

DECILE 8
34 Genes

DECILE 9
34 Genes

DECILE 10
34 Genes

Highest
10%

Worst Fit

Randee Young
Undergraduate Researcher

Vinson Doyle
Postdoctoral Researcher

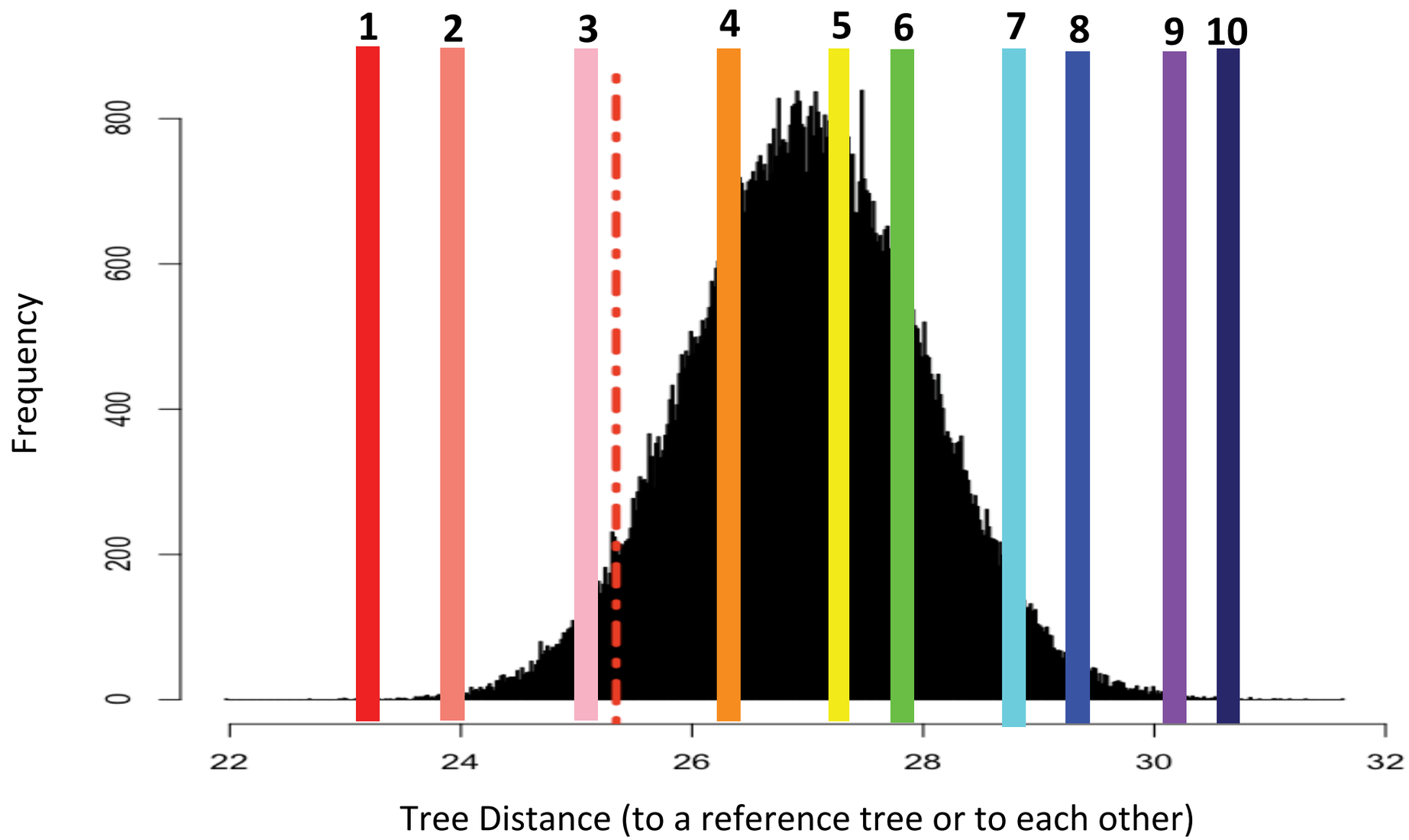
Doyle, Young, Naylor, Brown. 2015. Systematic Biology.

Posterior Prediction

What might we expect from ideal filtering approaches?

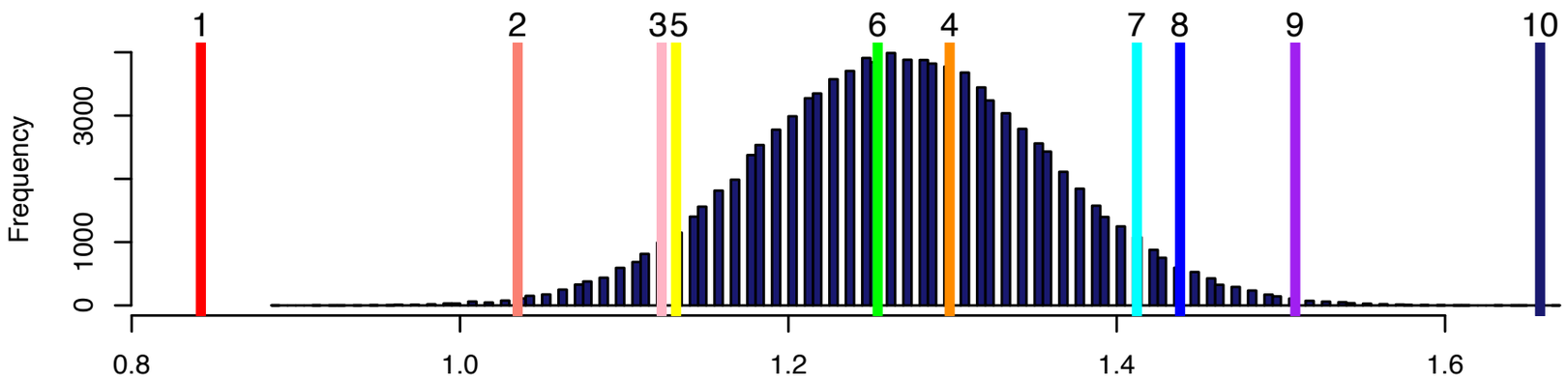
Perfect association between decile membership and tree distance

rho (r_s) = 1

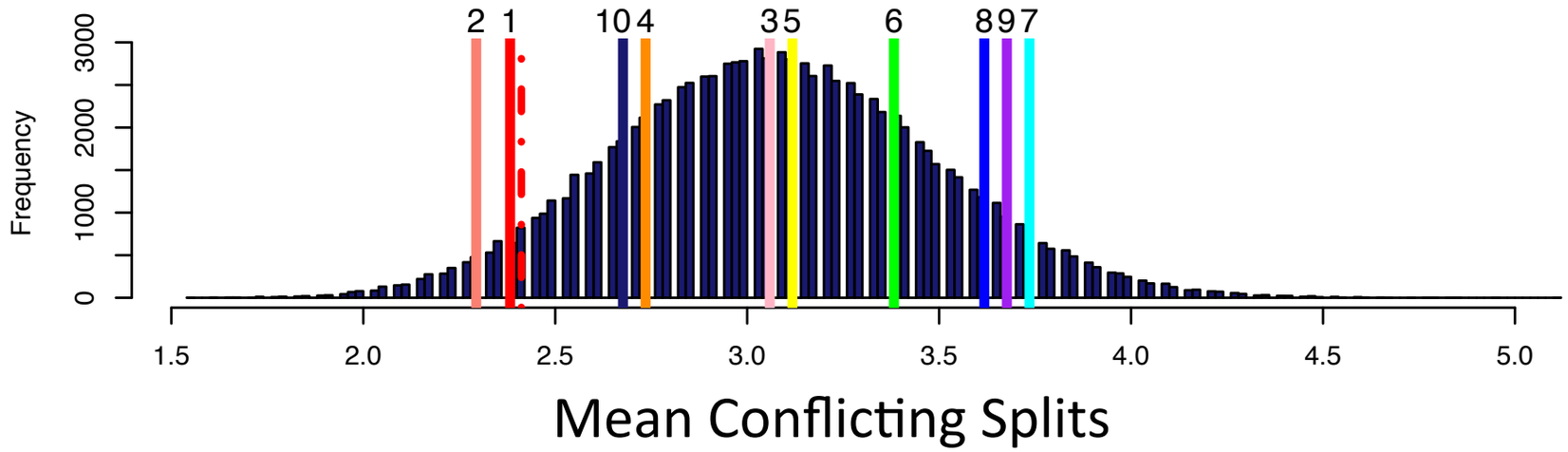


Doyle, Young, Naylor, Brown. 2015. Systematic Biology.

Posterior Prediction



r_s=0.964, P=2.2x10^{-16}



r_s=0.600, P=0.03656

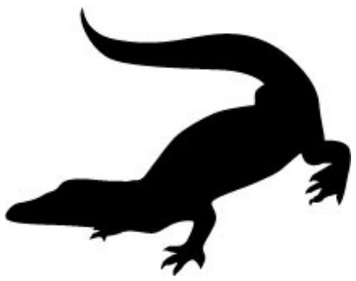
Doyle, Young, Naylor, Brown. 2015. Systematic Biology.

Posterior Prediction

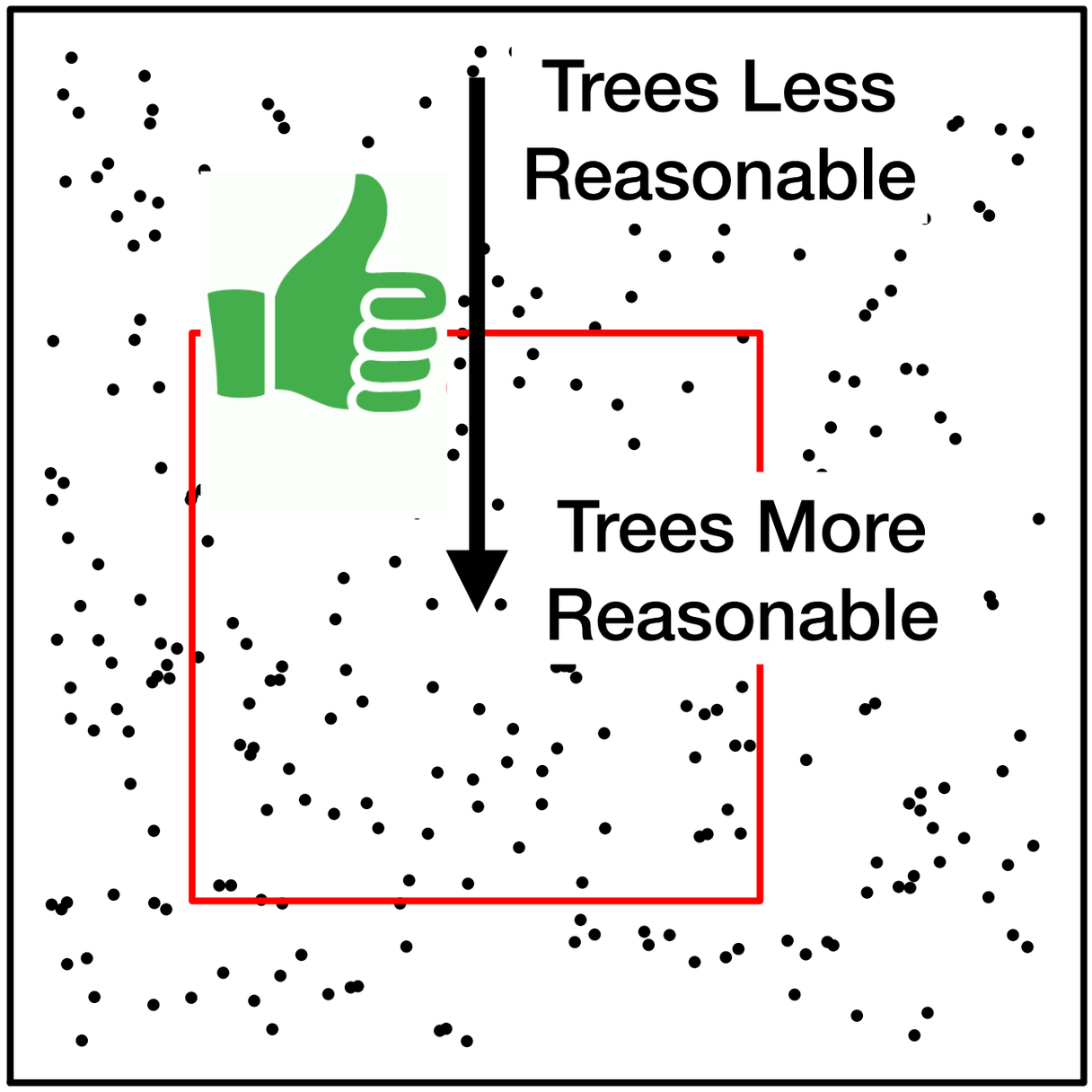
microbewiki.kenyon.edu

Yeast
343 orthologs
18 taxa
Hess & Goldman (2011)

Amniotes
1,145 orthologs
10 taxa
Crawford et al. (2012)



phylopic.org



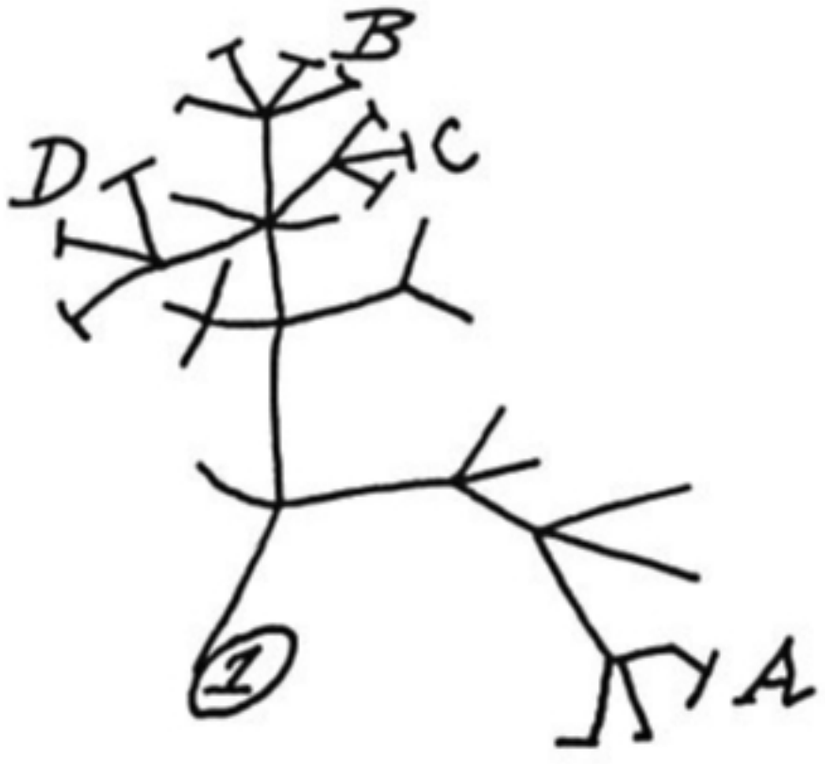
Doyle, Young, Naylor, Brown. 2015. Systematic Biology.

How consistent is poor model fit?

(And how common is poor model fit?)

Kylie Domangue

How consistent is poor model fit?



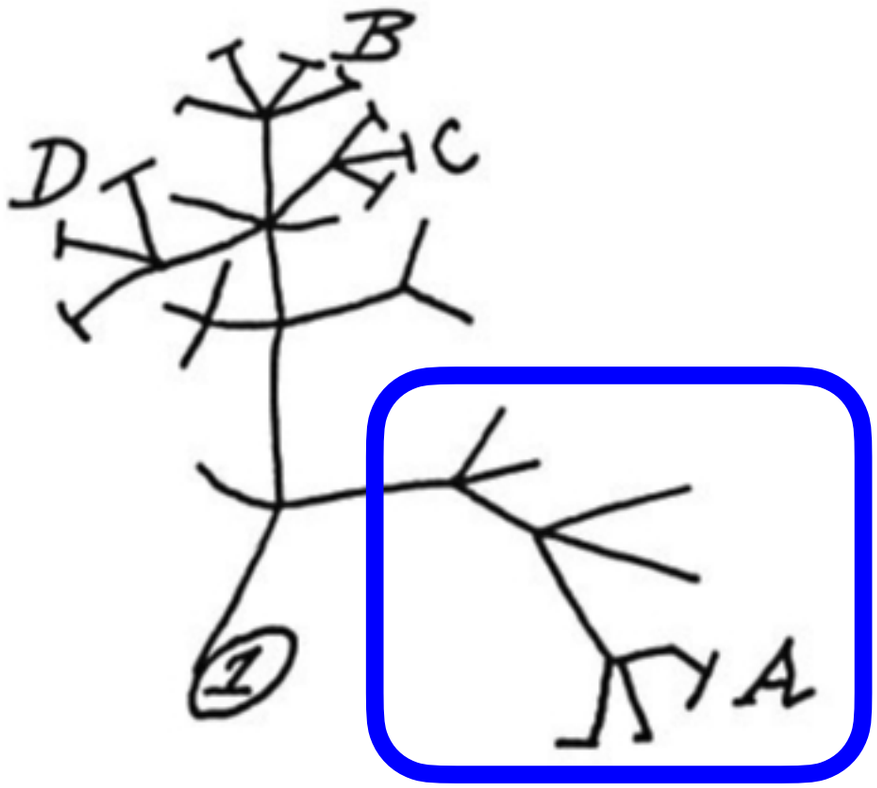
Gene 1

Gene 2

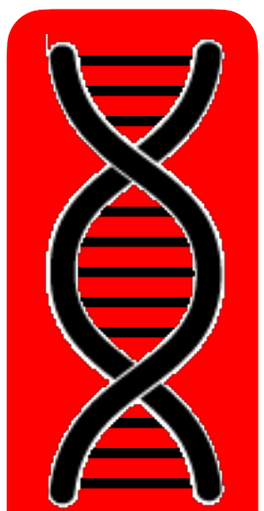
Gene 3

Gene 4

How consistent is poor model fit?



Gene 1

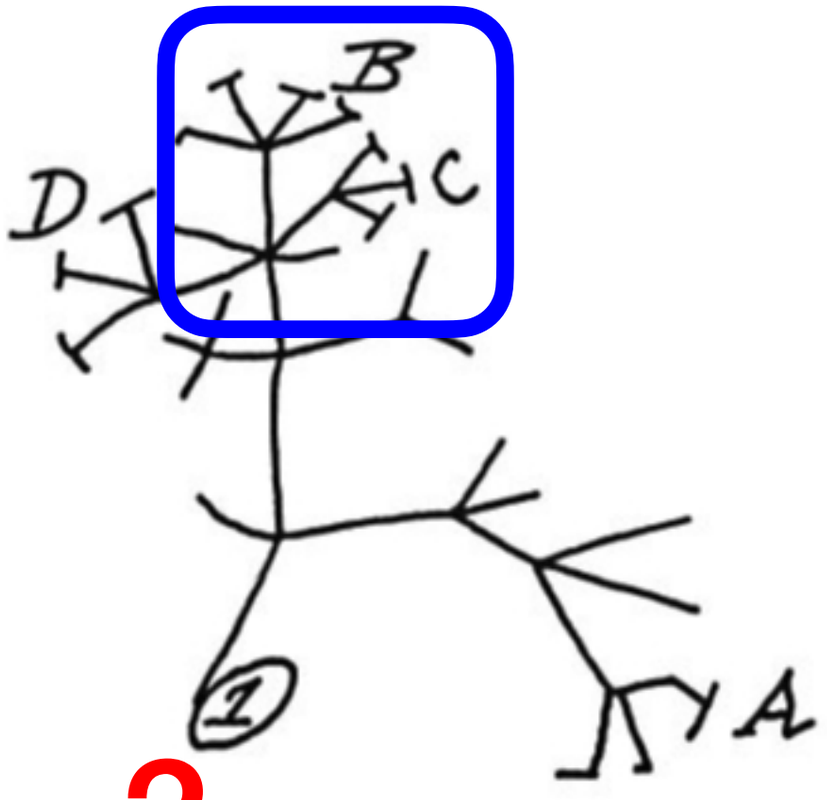


Gene 2

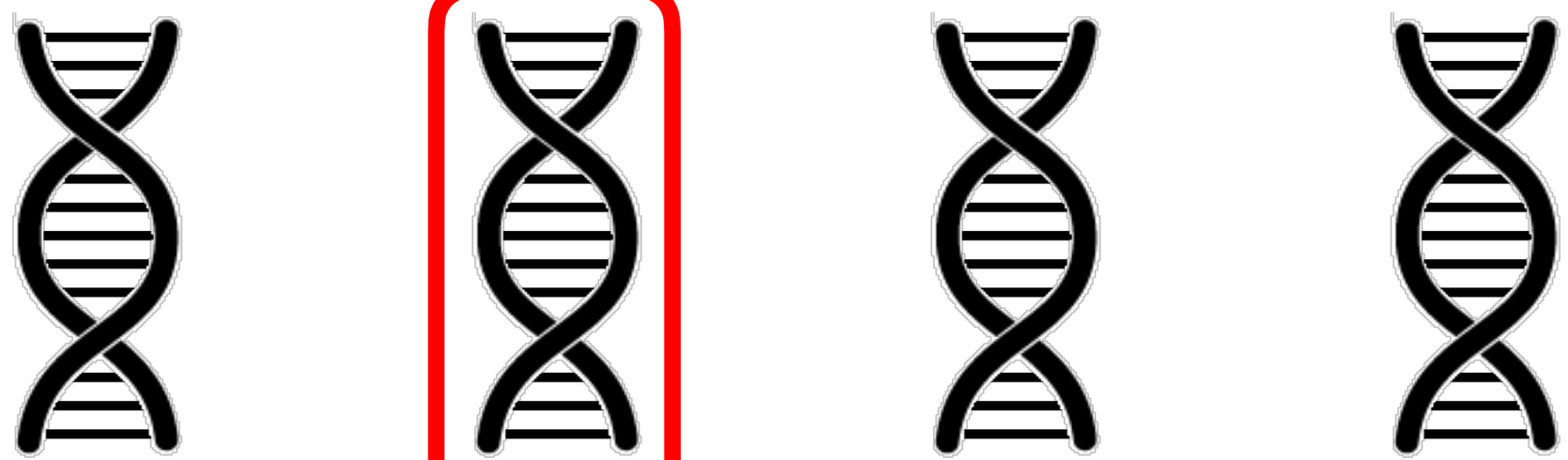
Gene 3

Gene 4

How consistent is poor model fit?



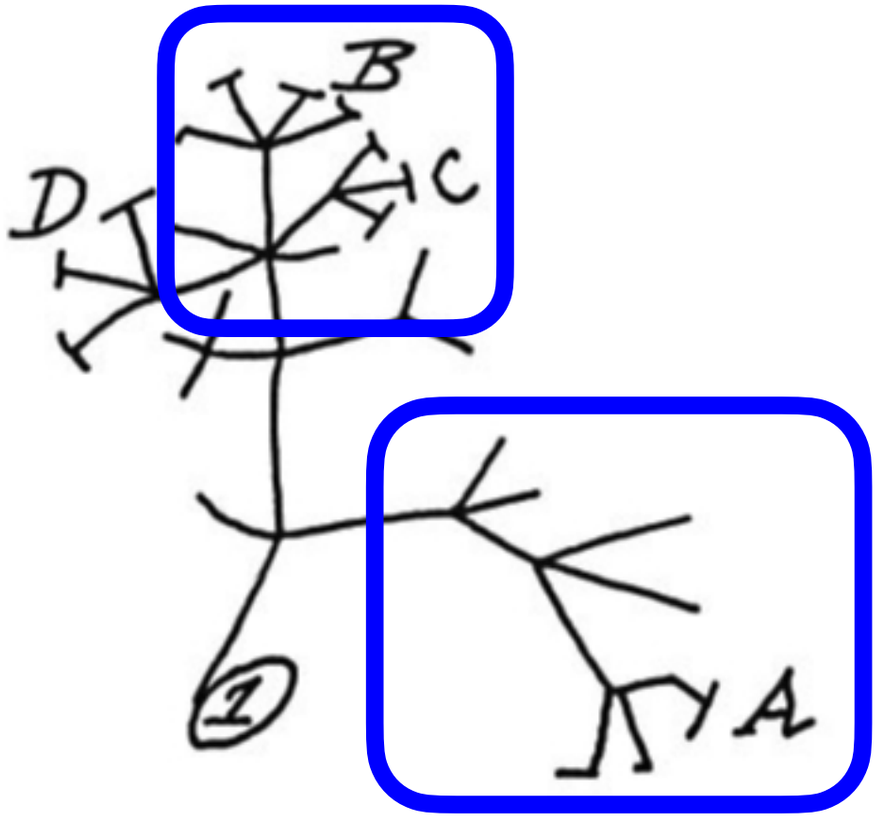
?



Gene 1 Gene 2 Gene 3 Gene 4

How consistent is poor model fit?

If the same genes always fit poorly, we can target others…and try to figure out what's going on.



Gene 1

Gene 2

Gene 3

Gene 4

How consistent is poor model fit?

213 highly curated **orthologs** from mammals (OrthoMam)

We divided taxa into 3 monophyletic groups:

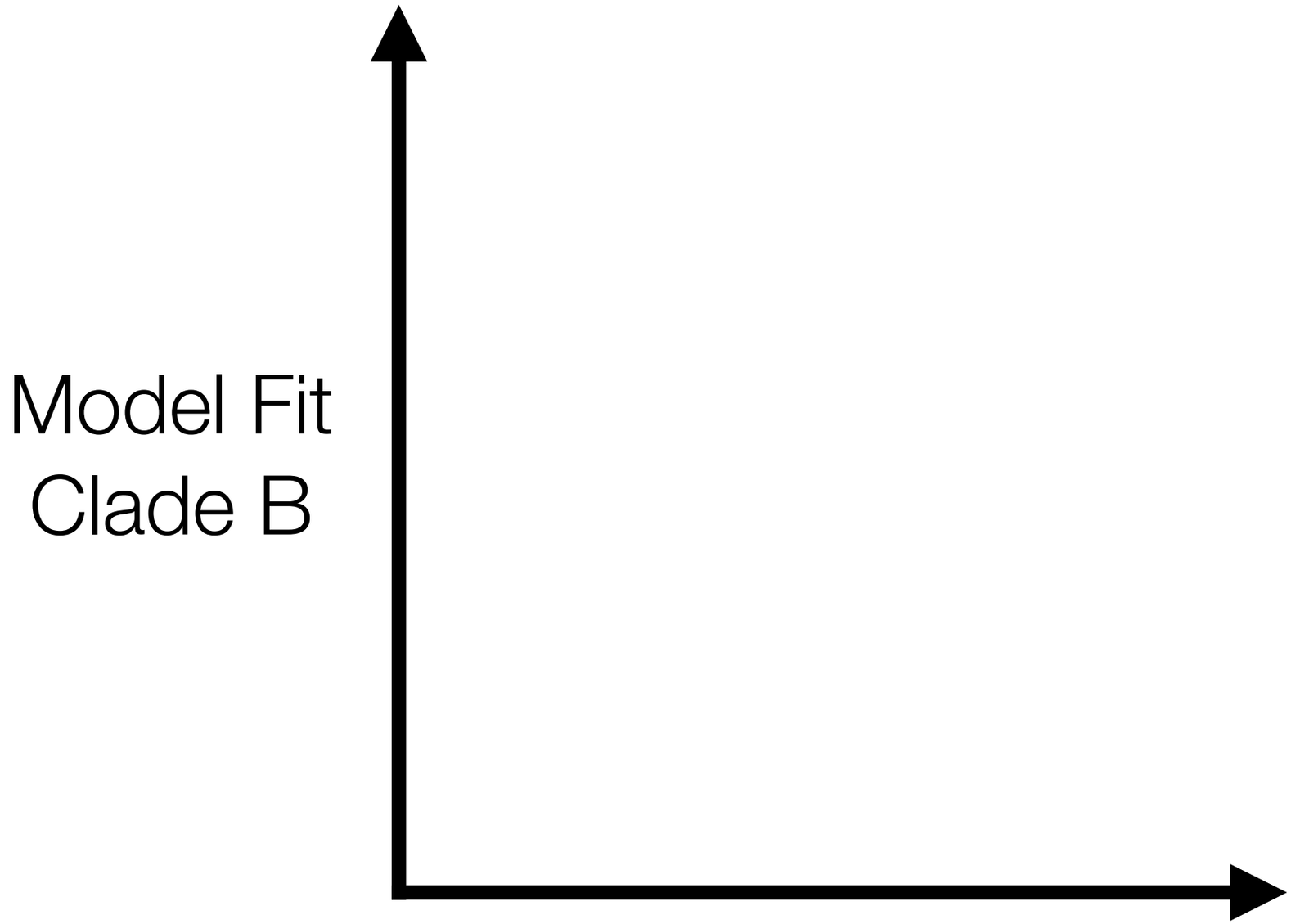
Carnivores
11 spp.

Rodents
23 spp.

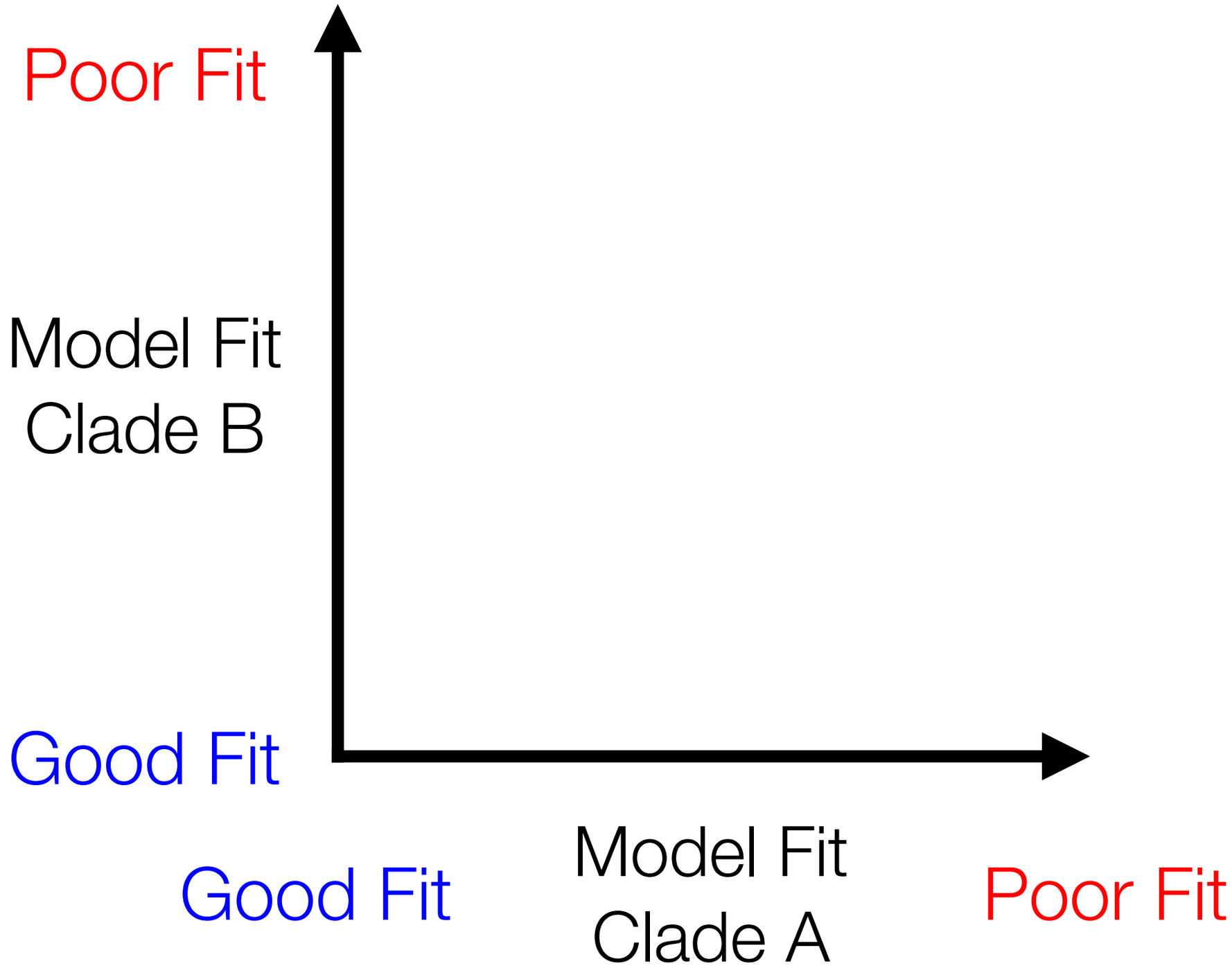
Primates
25 spp.

Analyzing and comparing many test statistics.

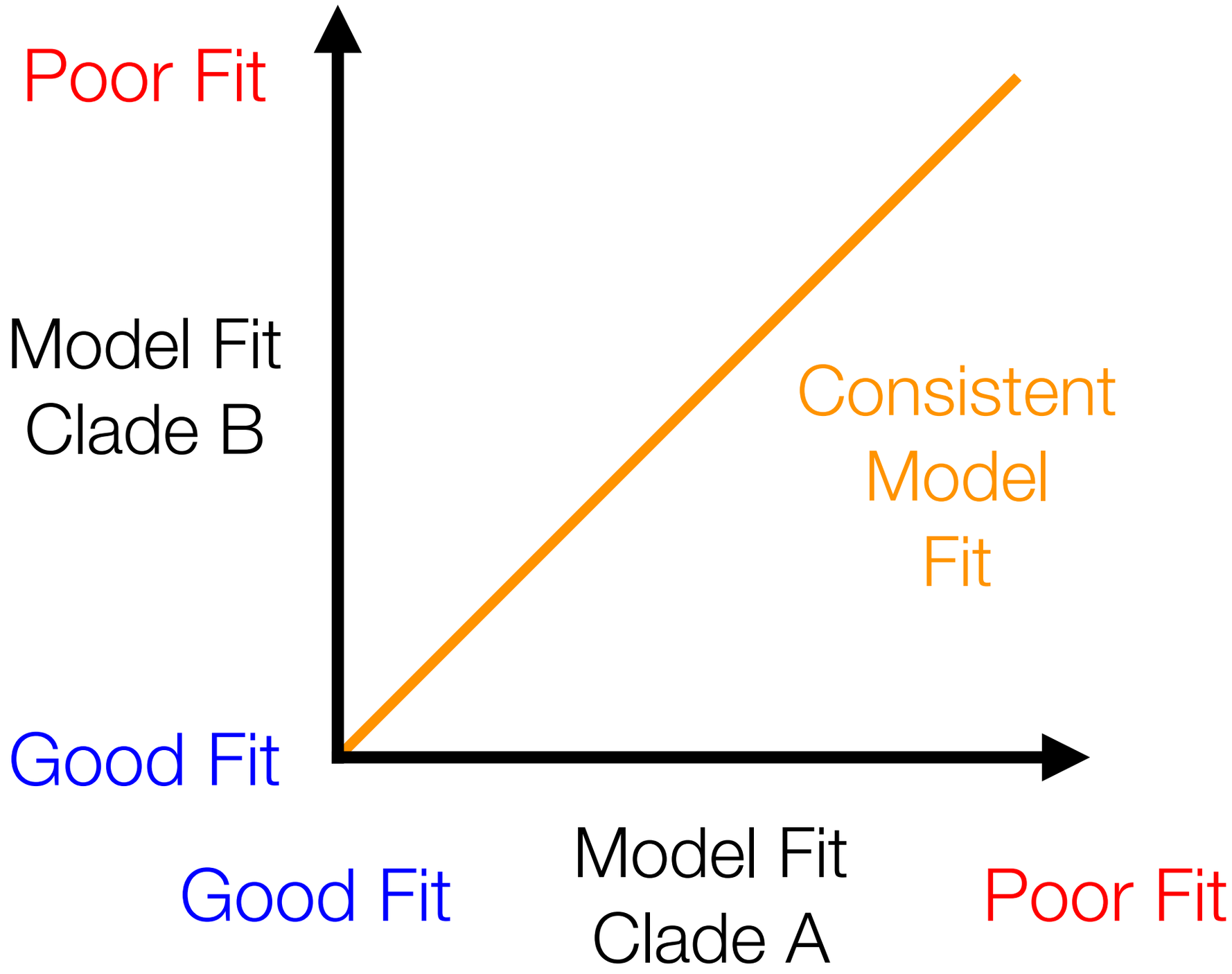
How consistent is poor model fit?



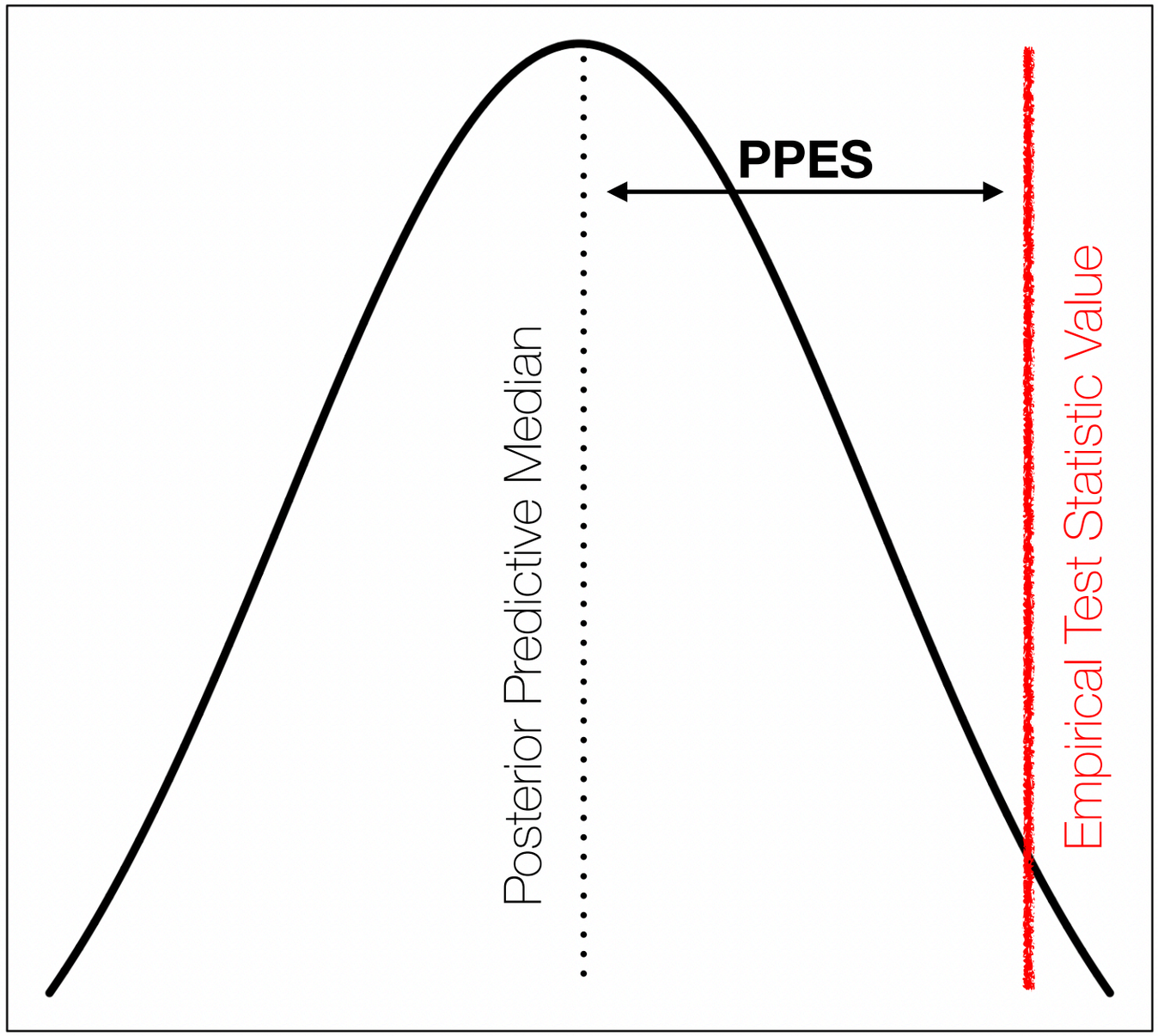
How consistent is poor model fit?



How consistent is poor model fit?

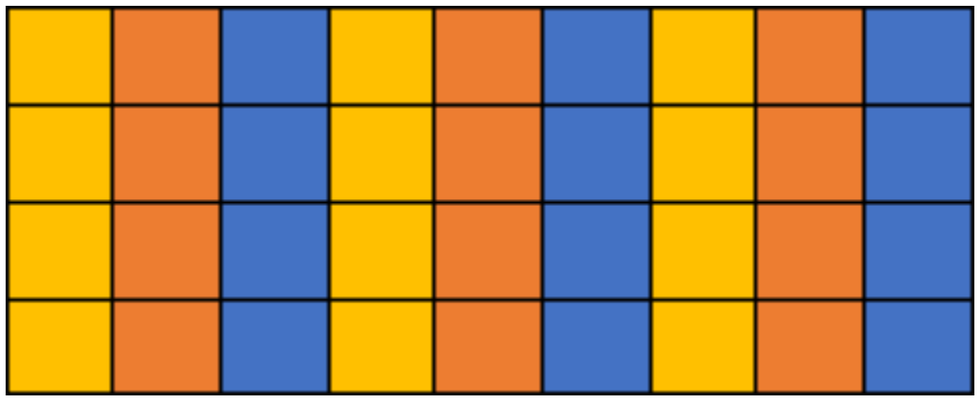


Posterior Predictive Effect Sizes



How consistent is poor model fit?

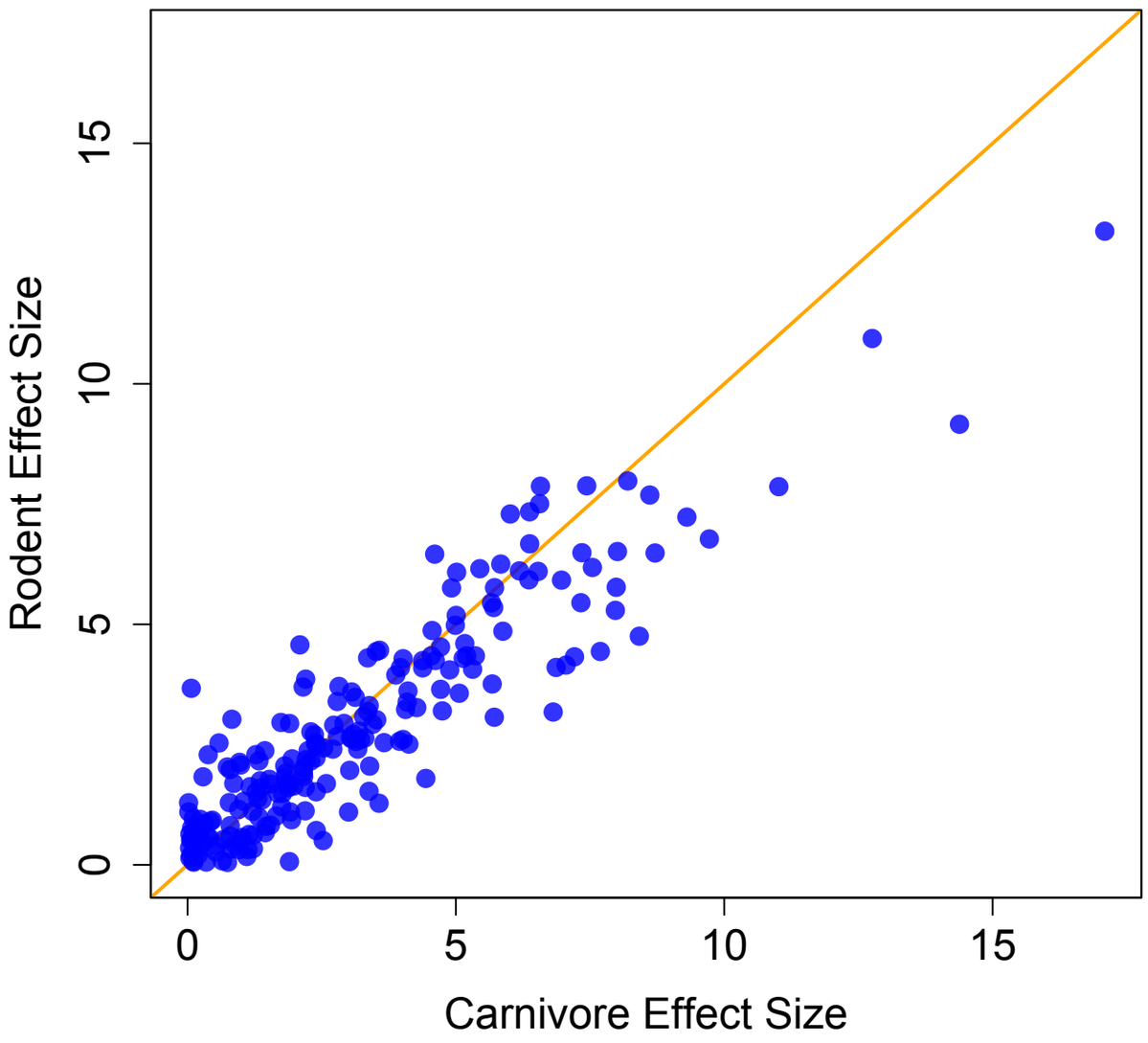
Mean GC content at 3rd codon positions

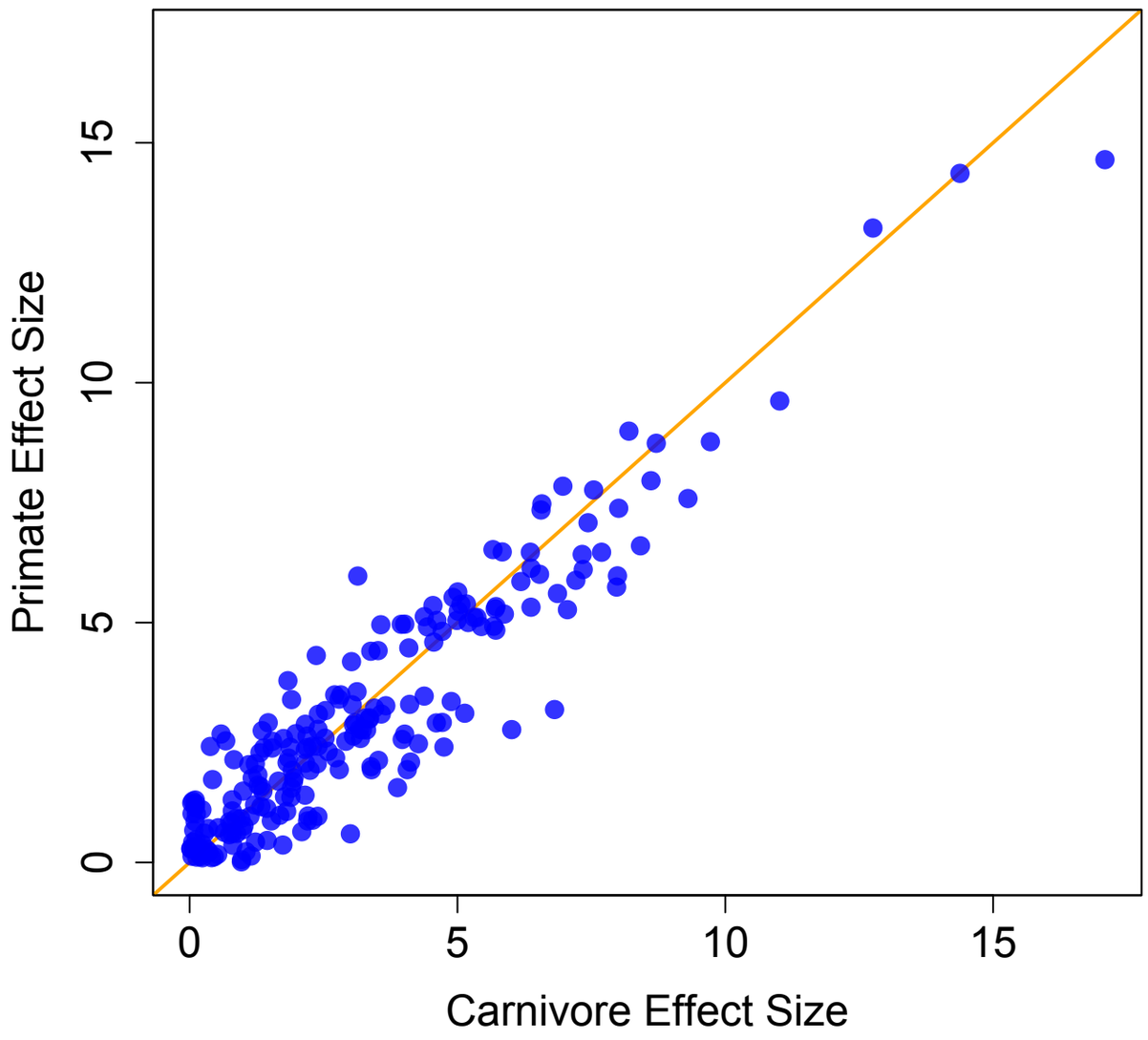


How consistent is poor model fit?

Mean GC content at 3rd codon positions

Consistent

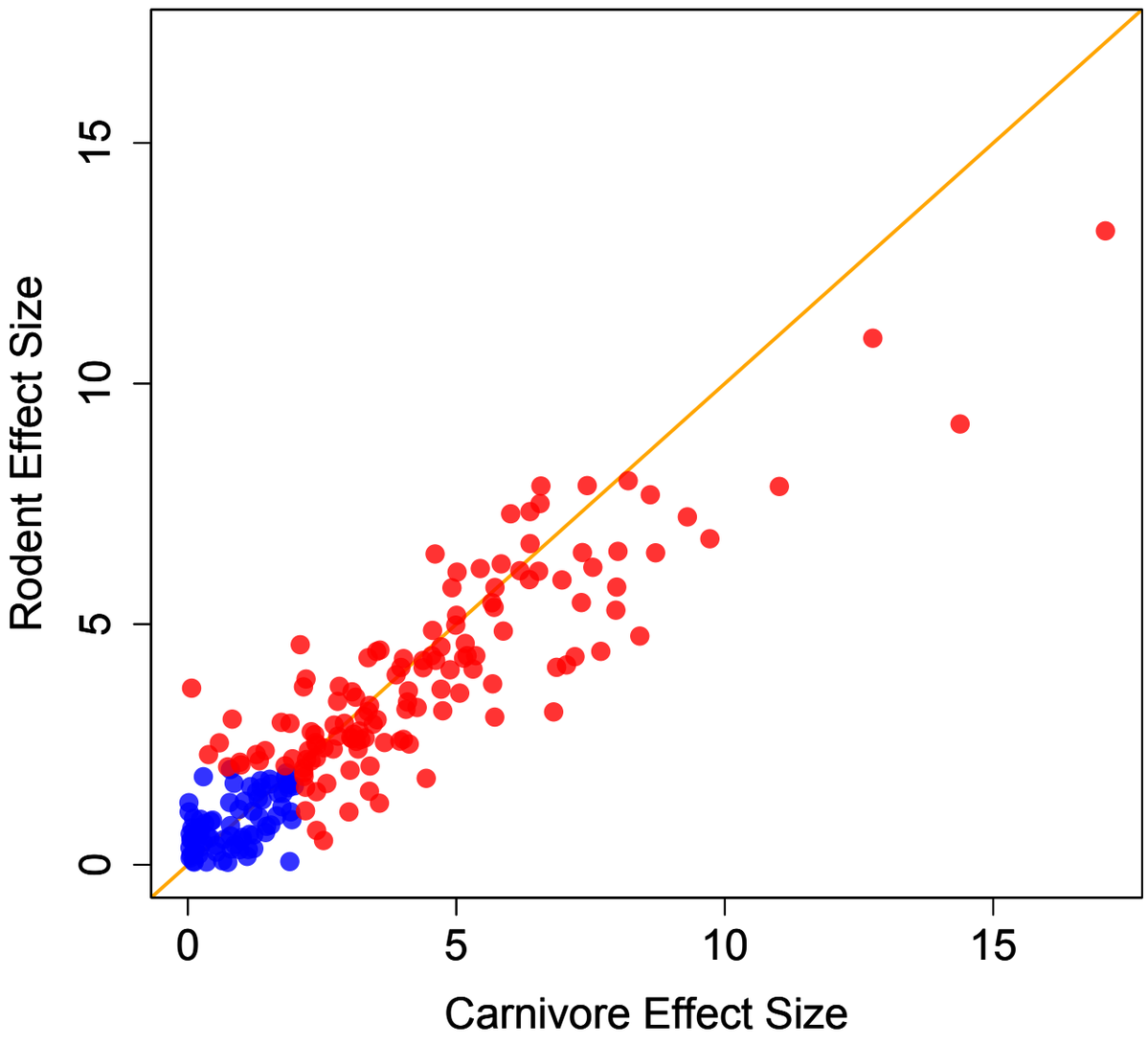


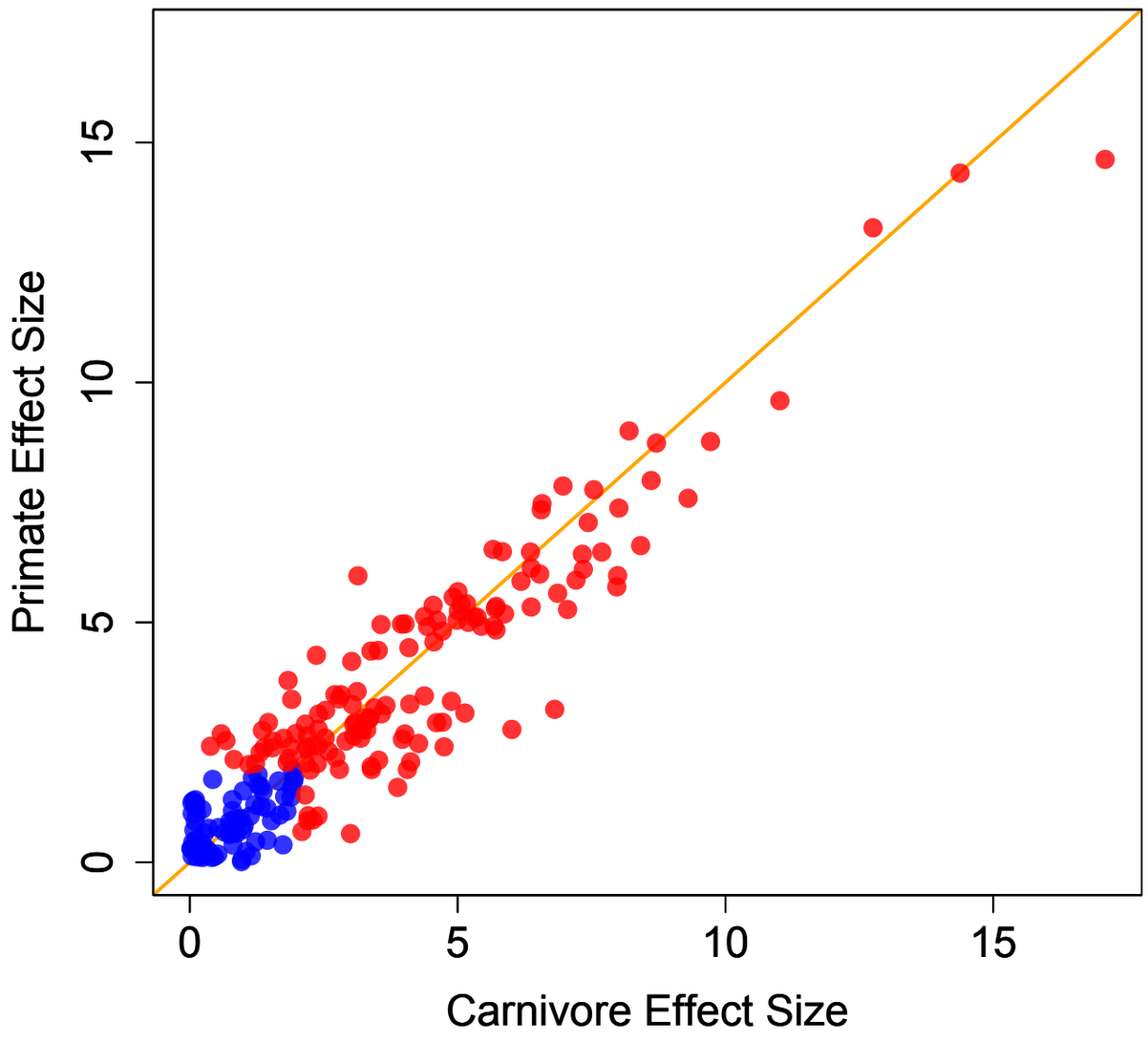


How consistent is poor model fit?

Mean GC content at 3rd codon positions

Consistent, but often poor

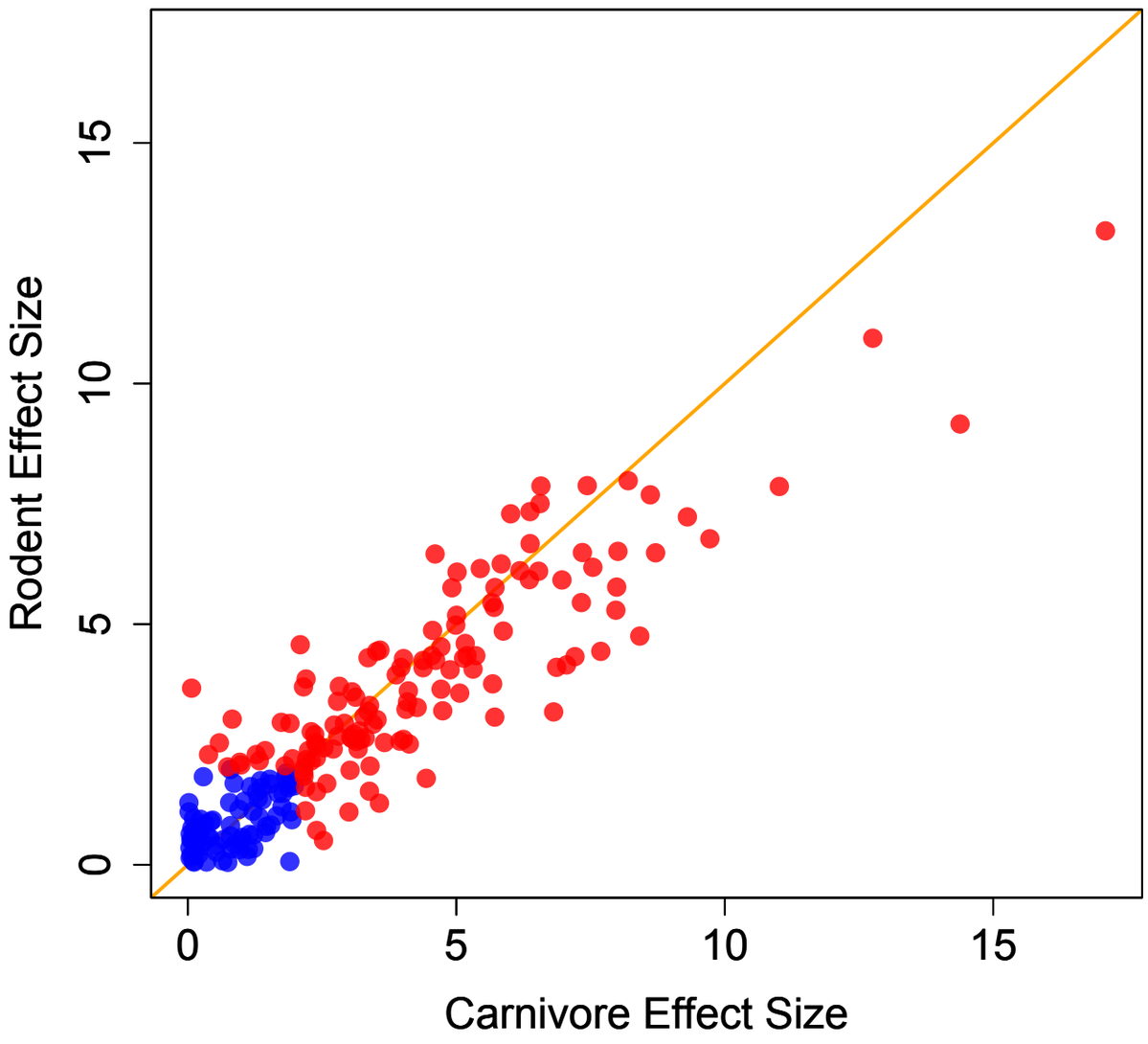


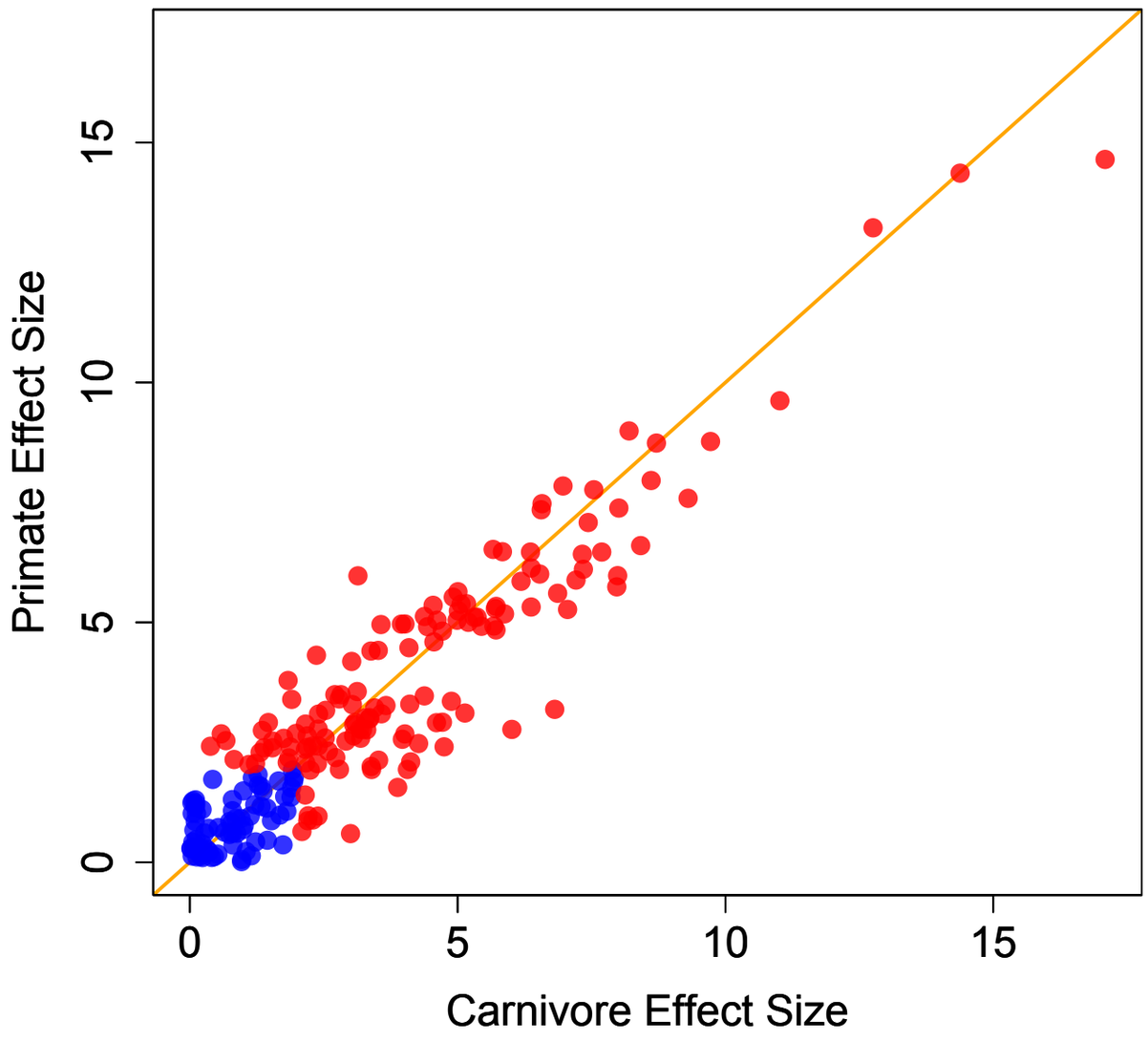


How consistent is poor model fit?

Mean GC content at 3rd codon positions

Poor fit could be improved with a **partitioned model**.





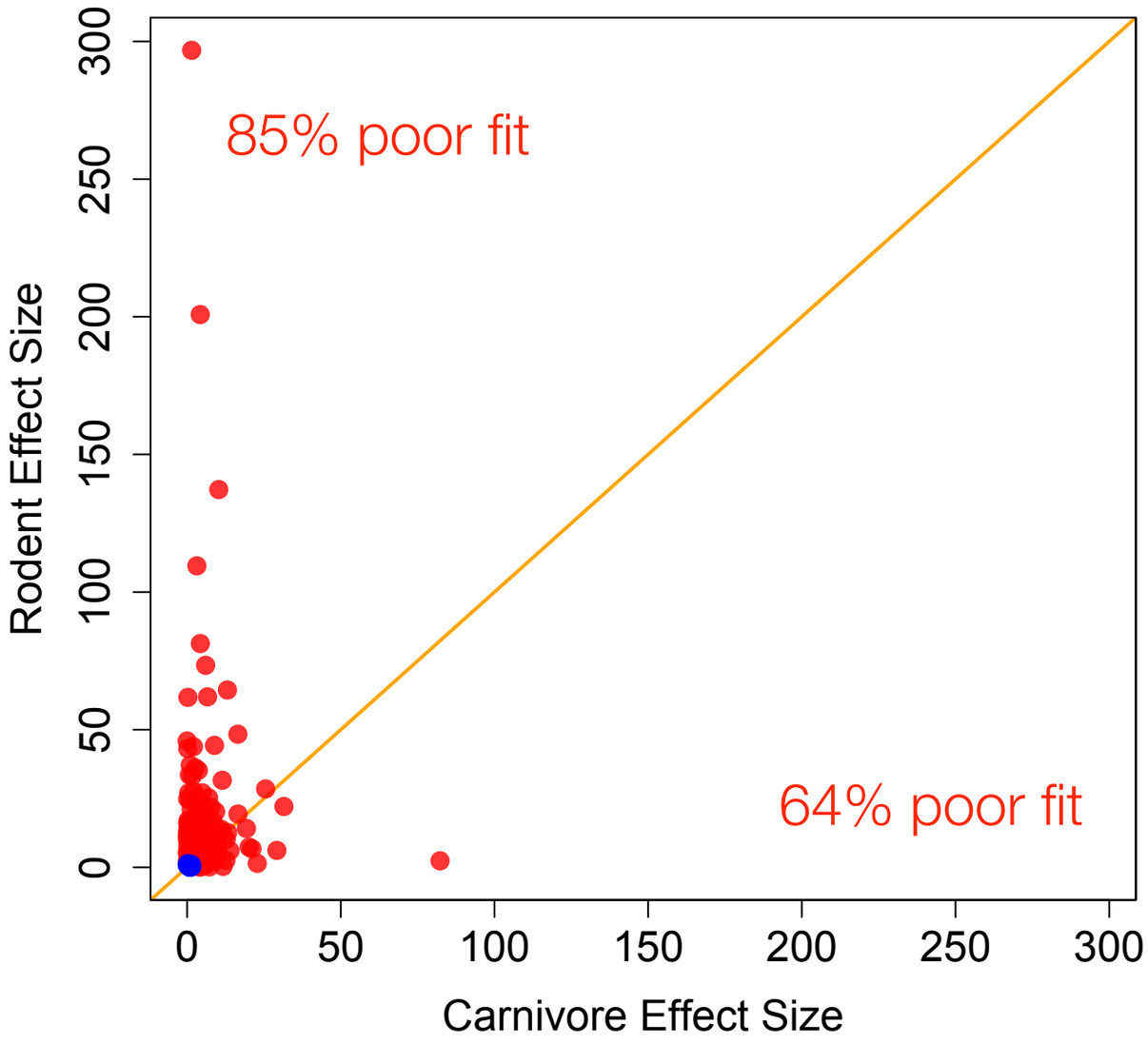
How consistent is poor model fit?

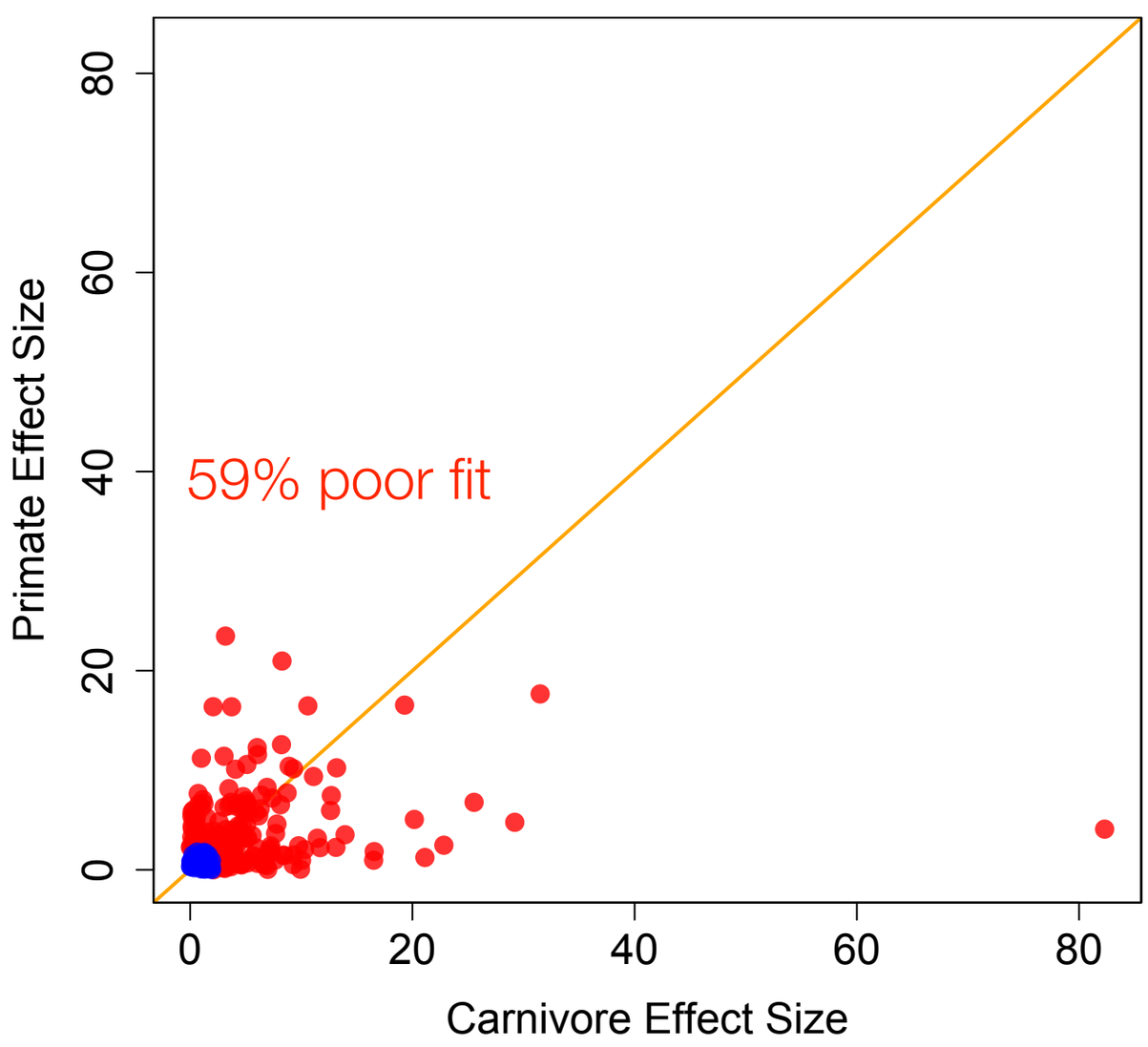
Variation **across taxa** in GC content at 3rd codon positions

How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

Very inconsistent, and very poor

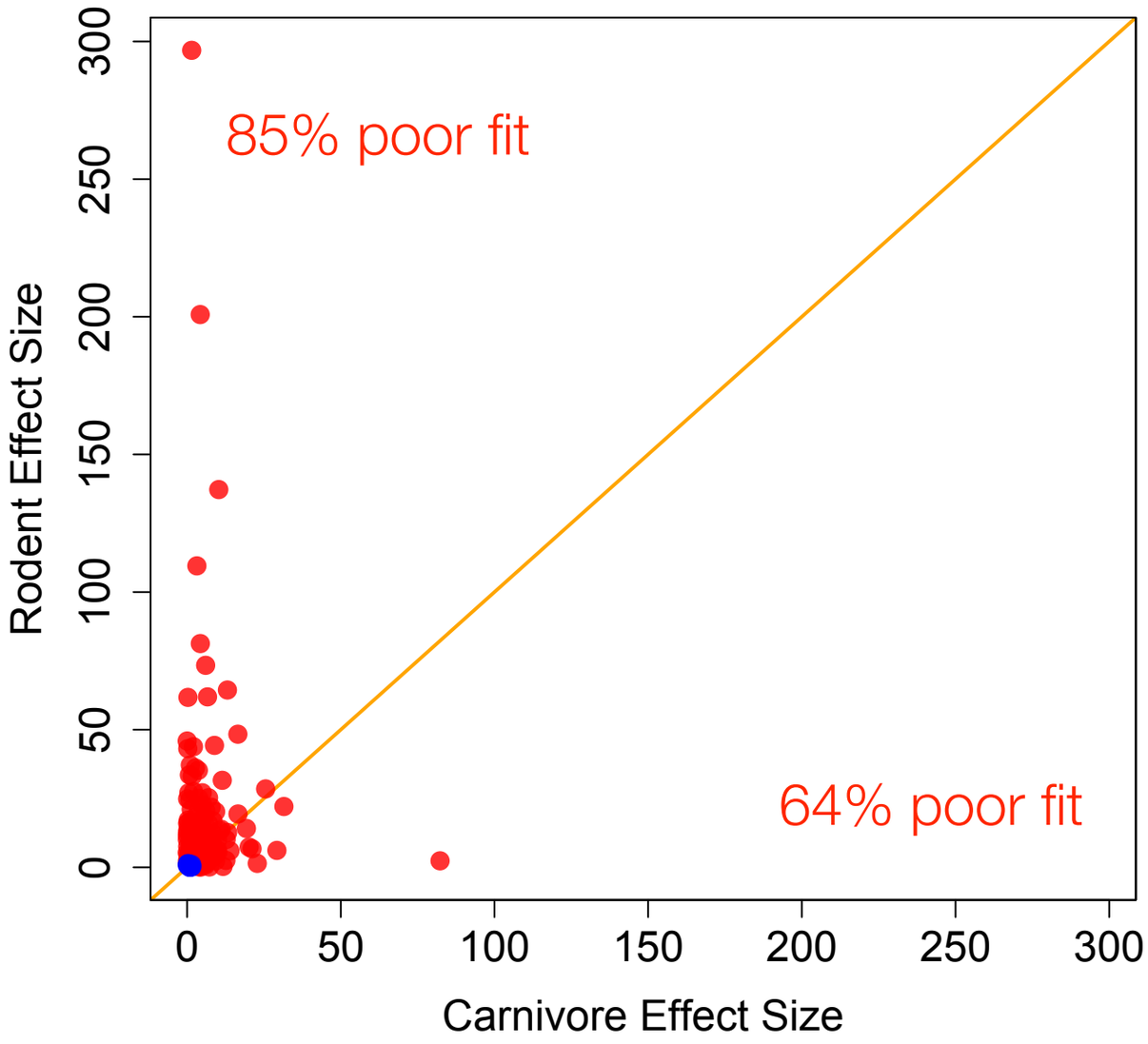


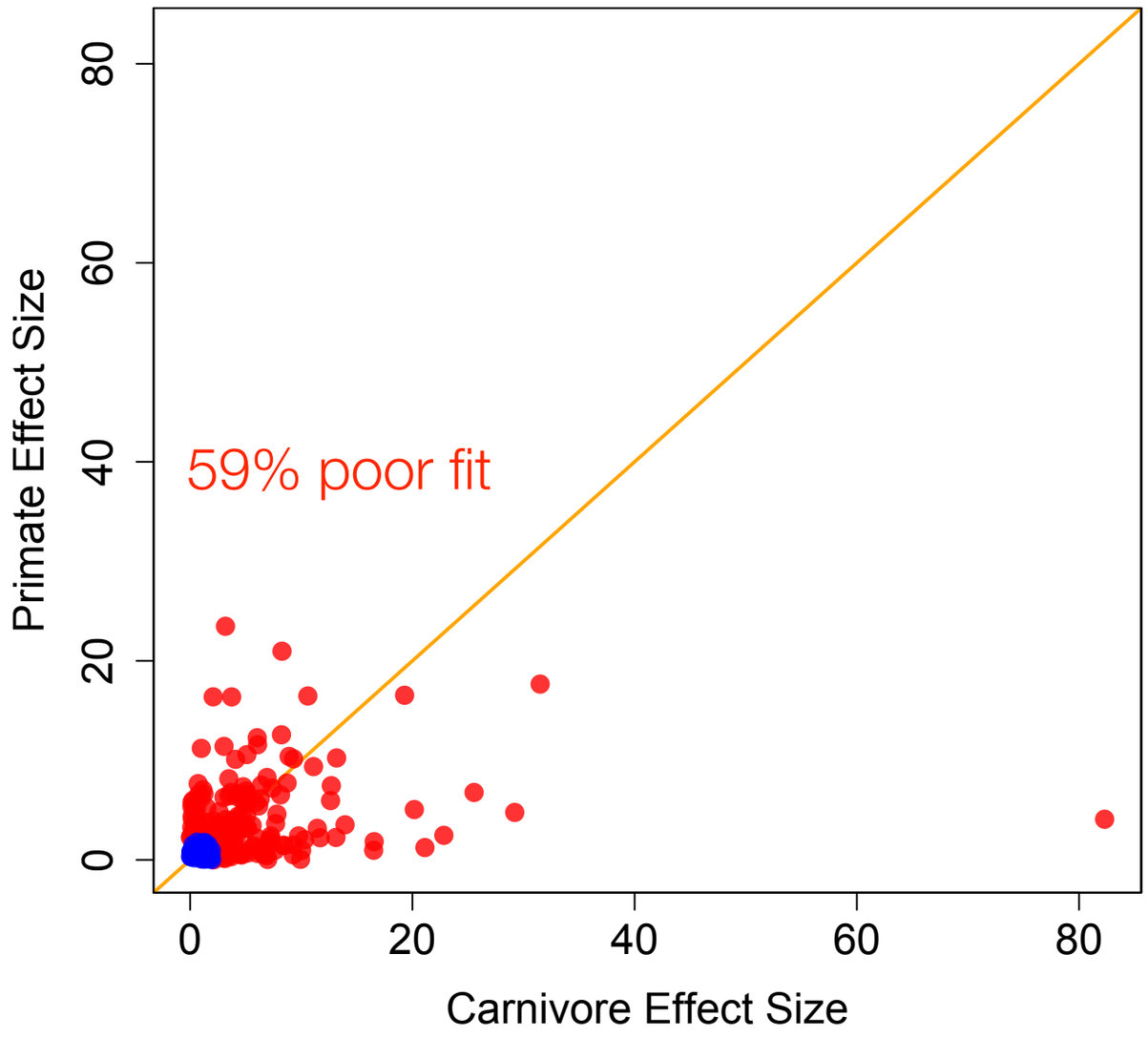


How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

No commonly used models account for this variation.

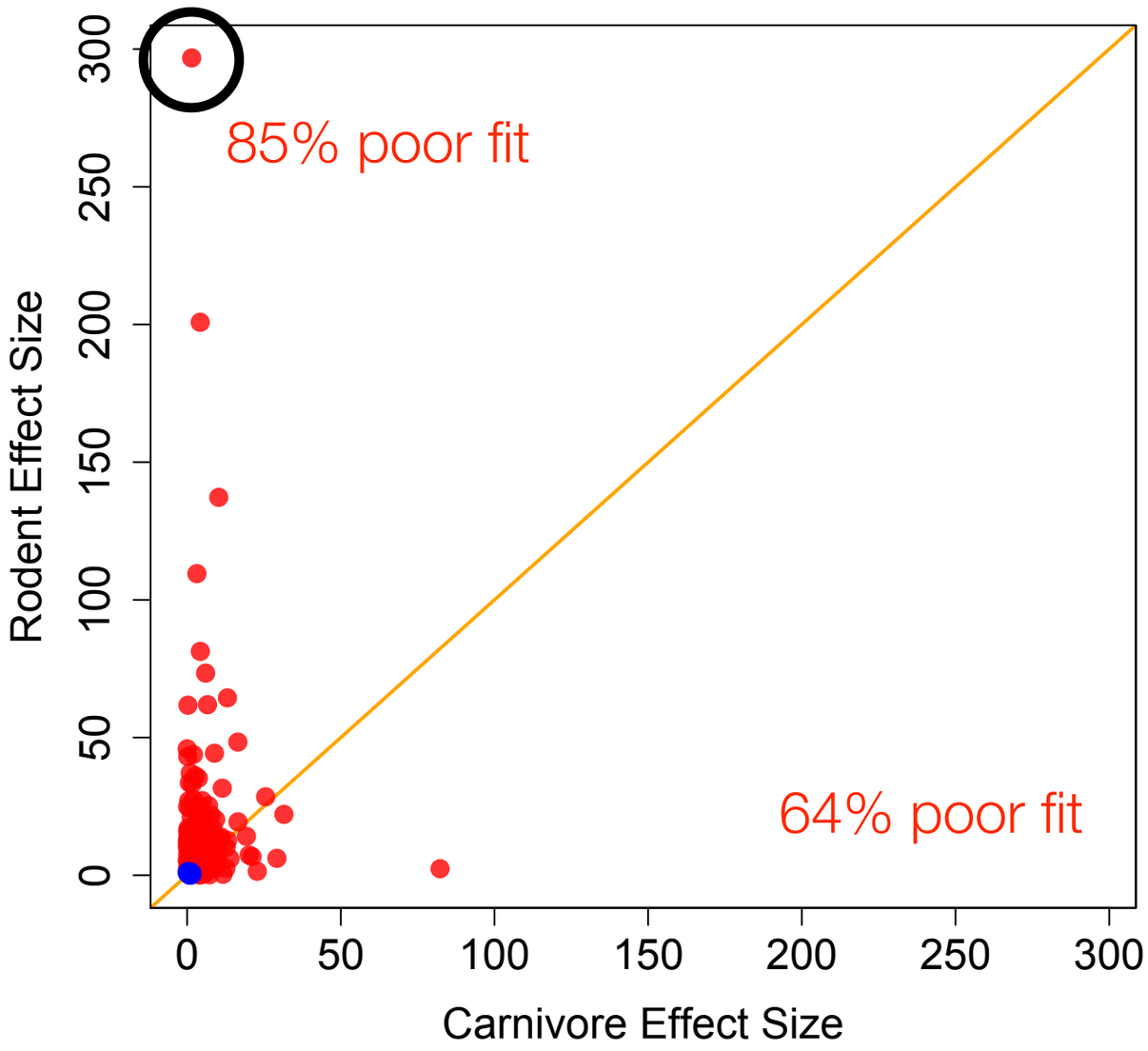




How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

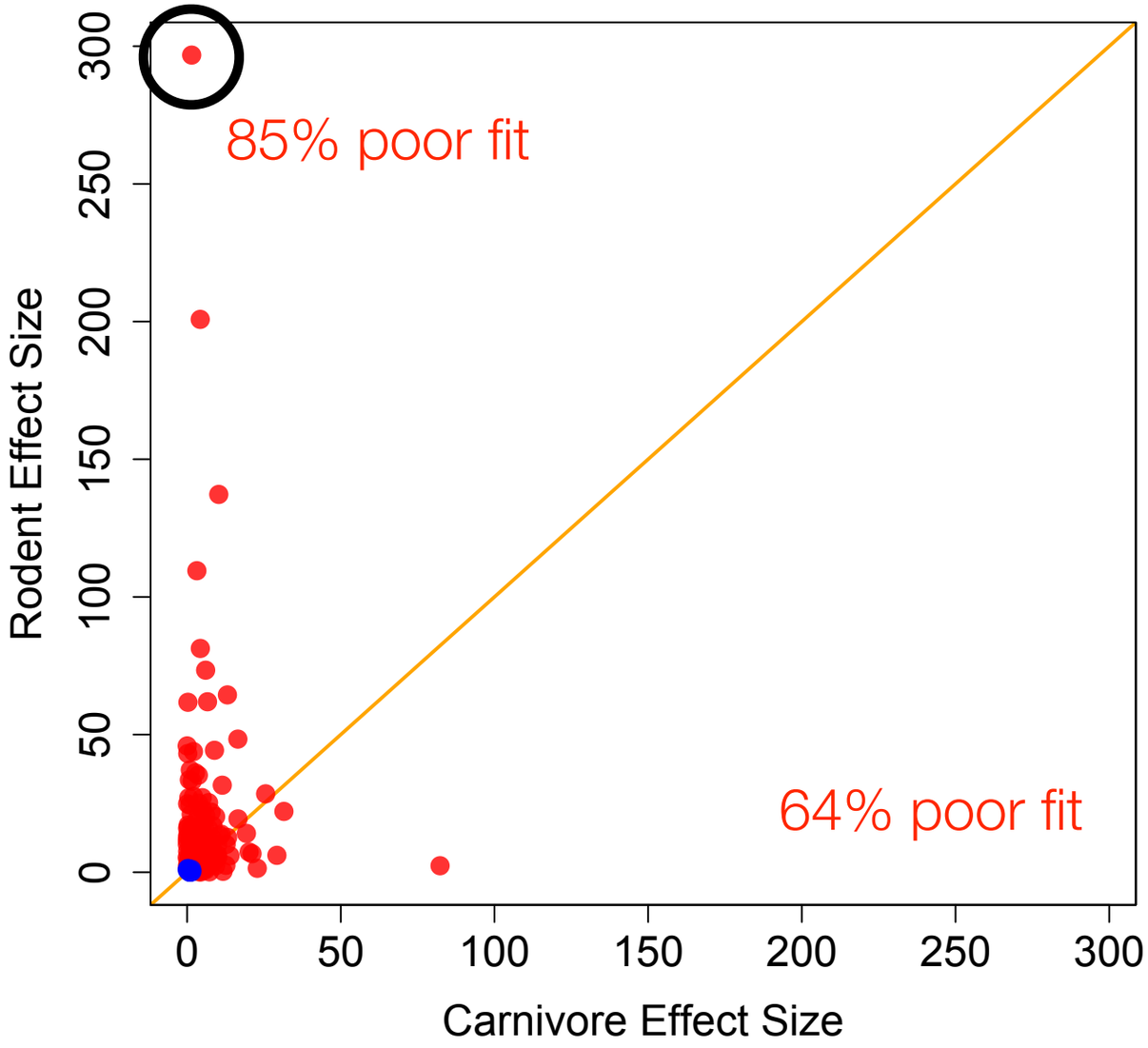
No commonly used models account for this variation.

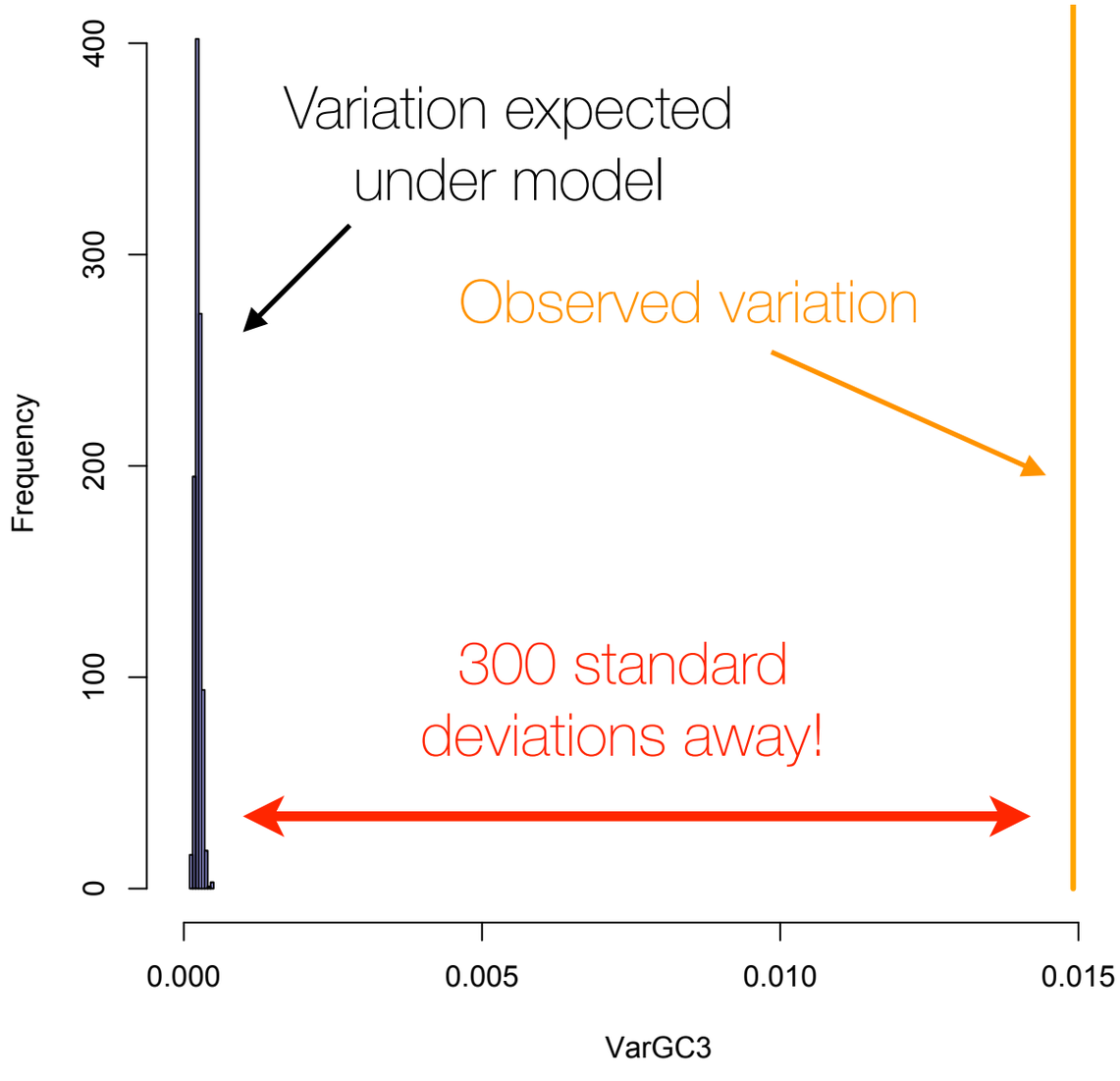


How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

No commonly used models account for this variation.

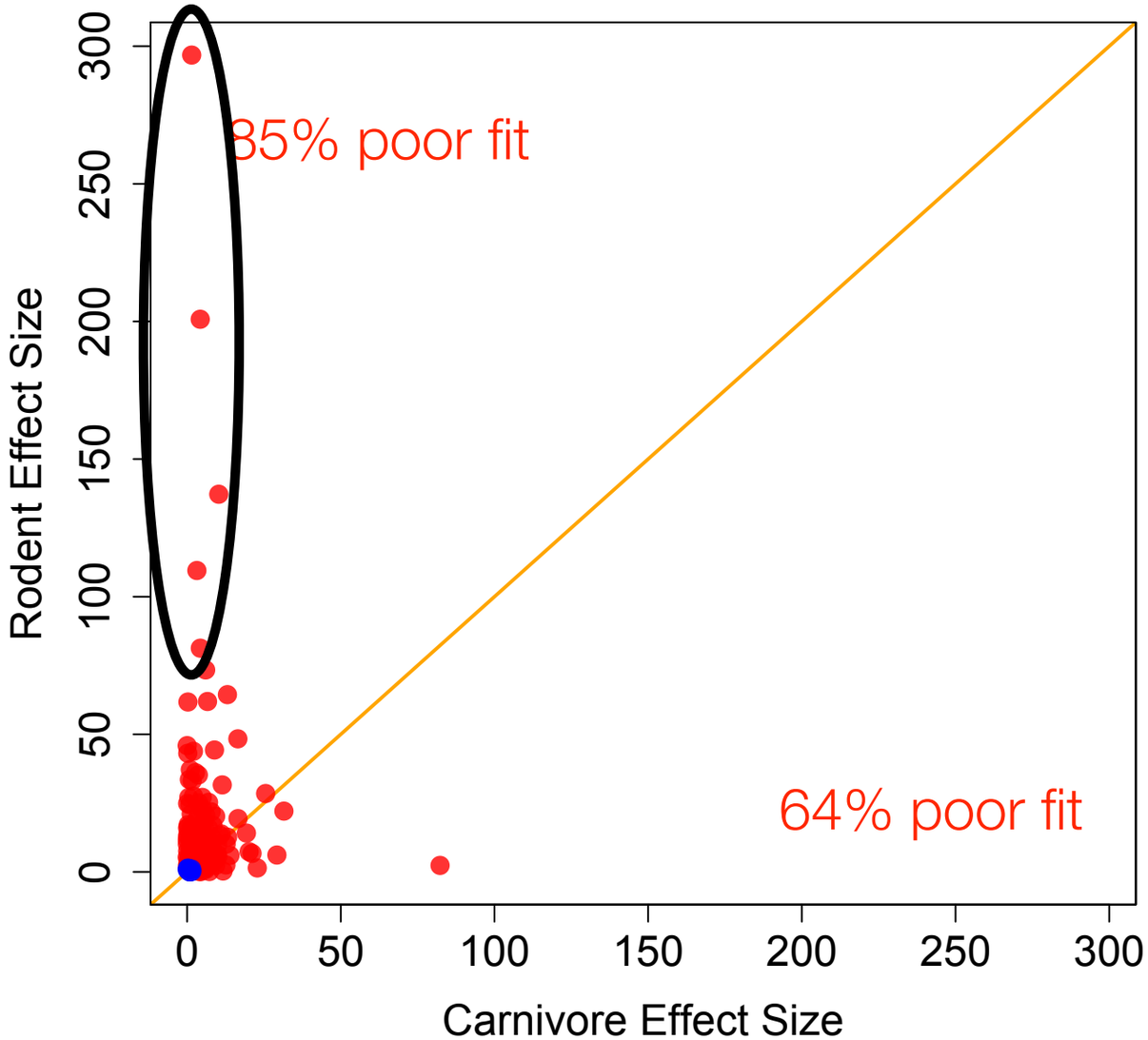




How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

No commonly used models account for this variation.



~30% increase in GC content relative to other rodents!

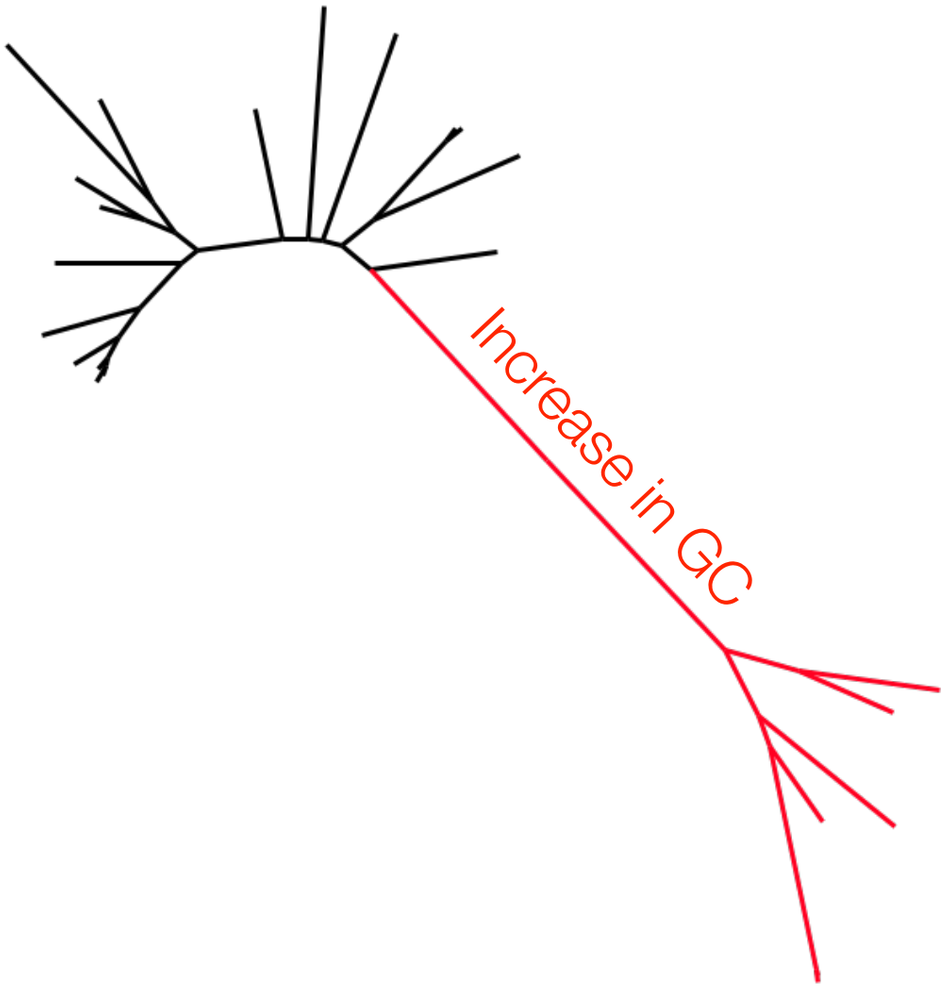
Increase in GC confined to 3rd codon positions.

Five genes seem subject to a strong mutational bias.

How consistent is poor model fit?

Variation **across taxa** in GC content at 3rd codon positions

No commonly used models account for this variation.



~30% increase in GC content relative to other rodents!

Increase in GC confined to 3rd codon positions.

Five genes seem subject to a strong mutational bias.