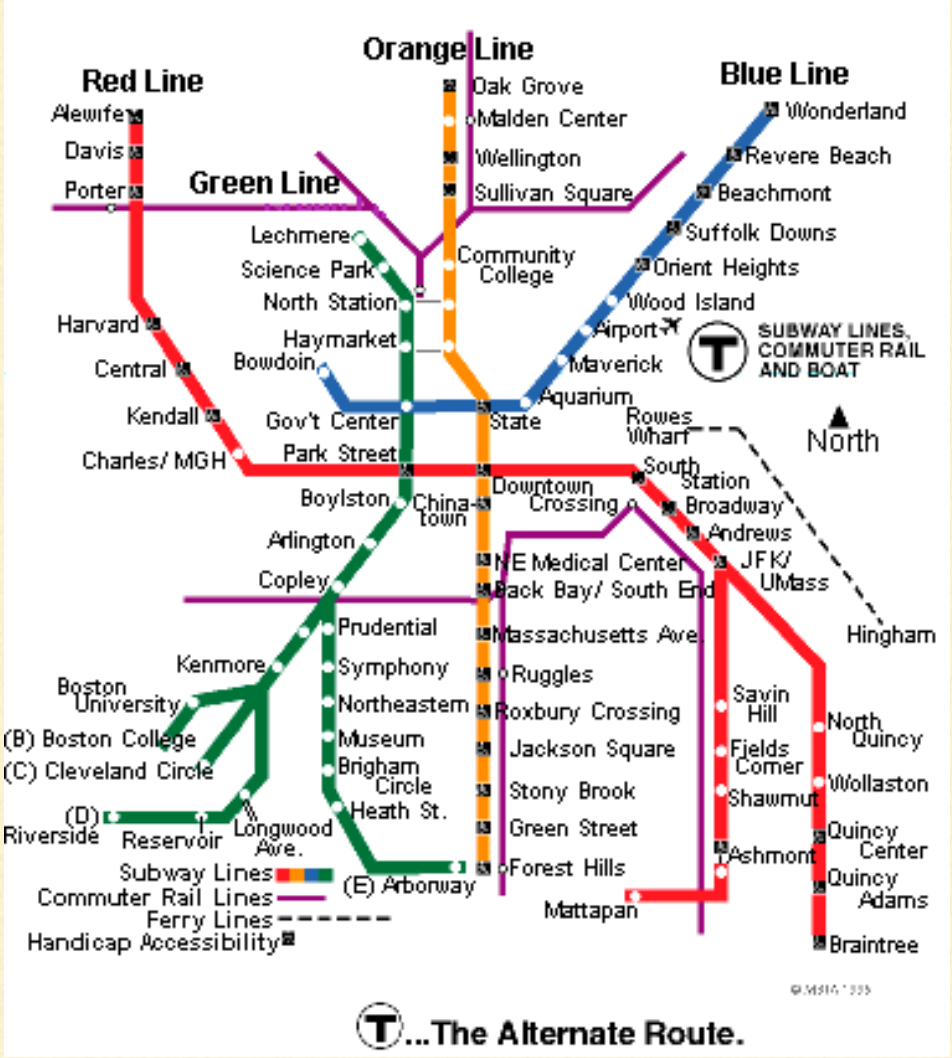
Models

This talk will be presented today by Paul Lewis, but most slides and examples are from the talk that David normally gives at this stage of the workshop.

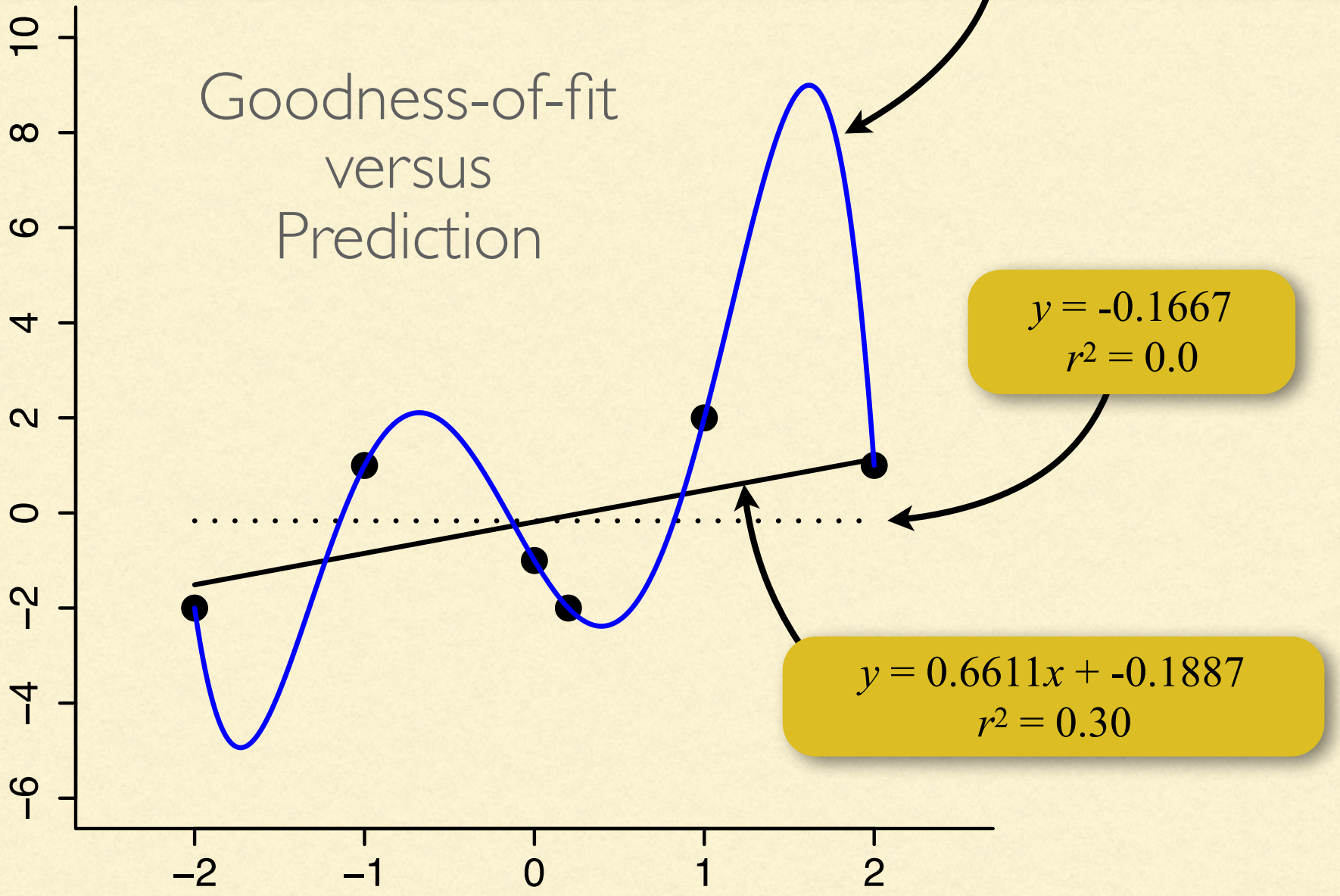
A very *realistic* MBTA subway map

A very *practical* MBTA subway map



Which is more useful?

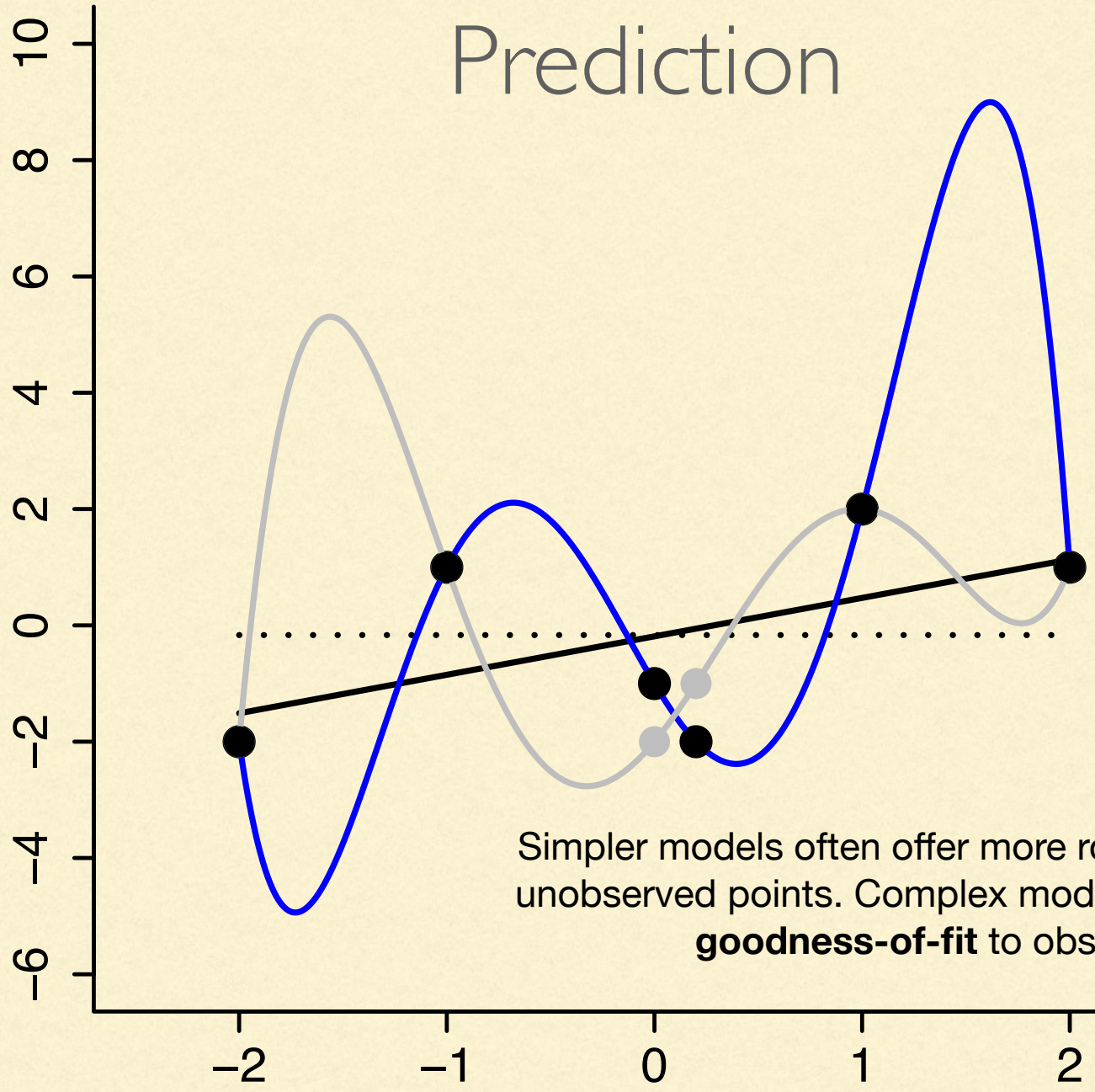
$y = -1.5972\,x^5 + -0.7917\,x^4 + 8.0694\,x^3 + 3.2917\,x^2 + -5.9722\,x + -1.0$
$r^2 = 1.0$



$y = -0.1667$
$r^2 = 0.0$

$y = 0.6611x + -0.1887$
$r^2 = 0.30$

Goodness-of-fit versus Prediction

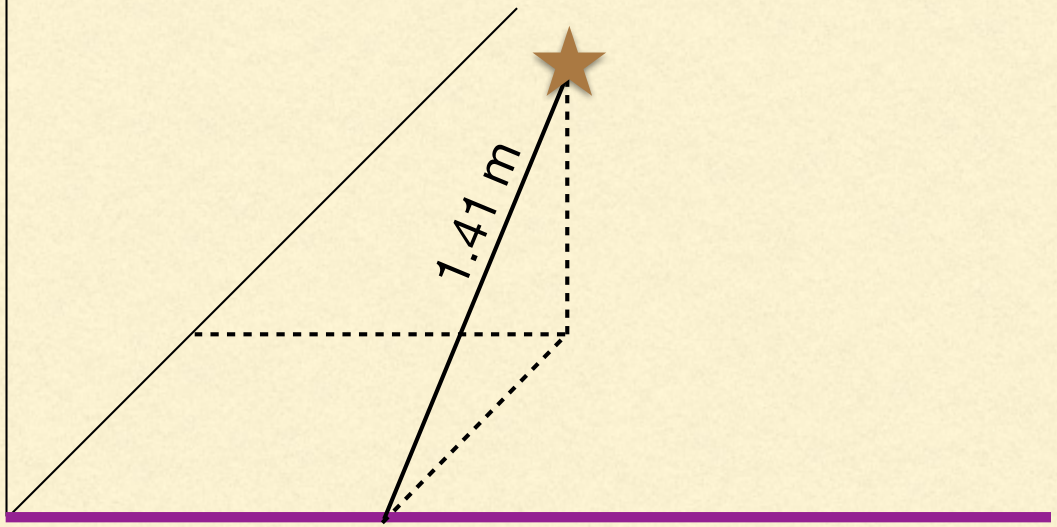


Shifting two data points slightly yields essentially the same linear regression line, but drastically alters predictions from the polynomial regression!

Simpler models often offer more robust **prediction** for unobserved points. Complex models often offer better **goodness-of-fit** to observations.

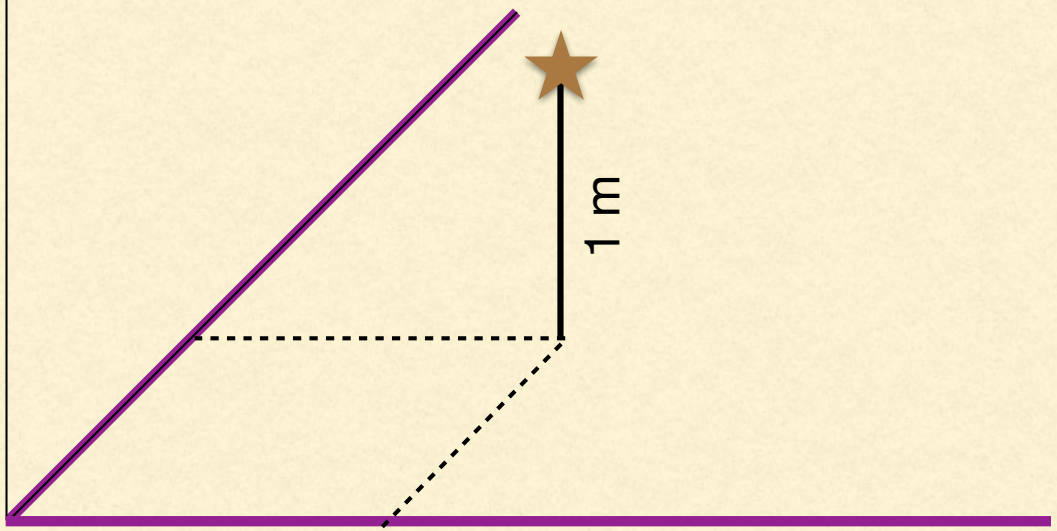
One parameter, one dimension

In this 1-parameter model, the closest we can get to the data point is 1.41 meters

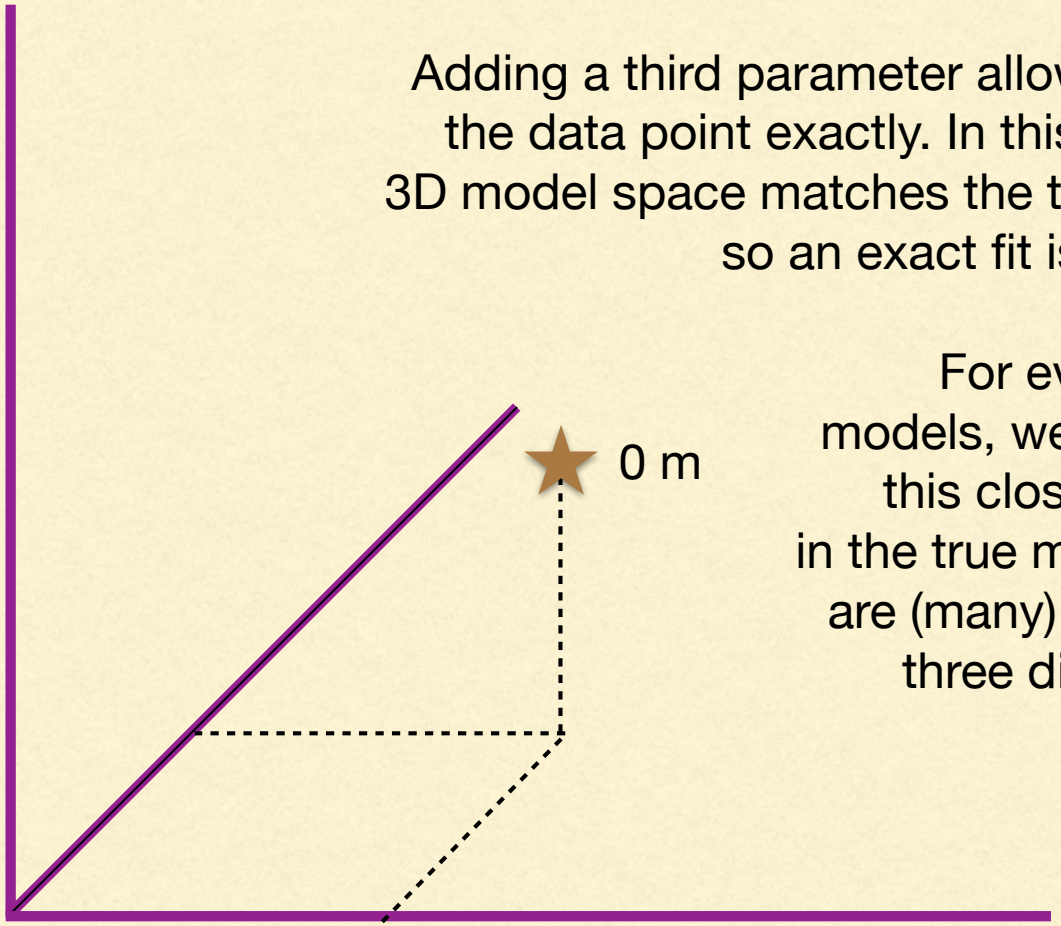


Two parameters, better fit

Adding a second parameter allows us to get a little closer to the data point. The model is now a 2-dimensional plane (i.e. the floor of the room)



Three parameters, perfect fit

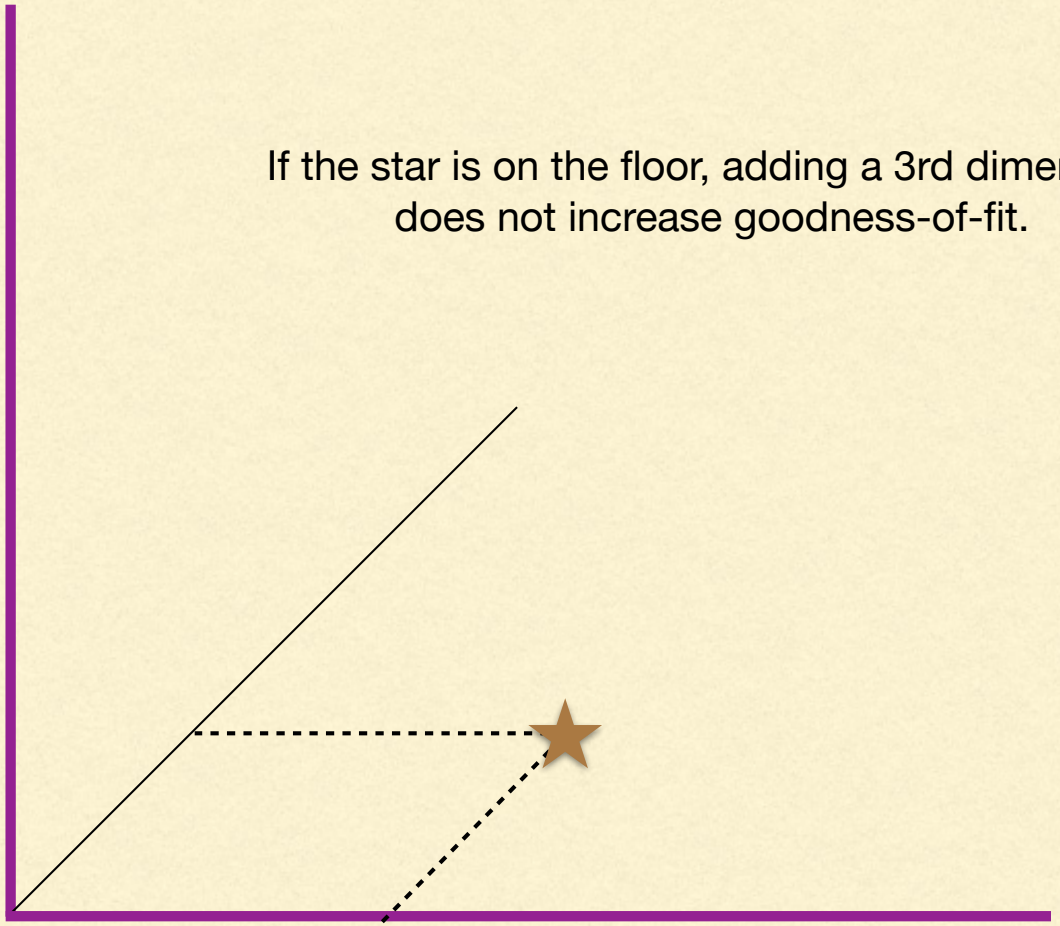


Adding a third parameter allows us to fit the data point exactly. In this case, the 3D model space matches the true space, so an exact fit is possible.

For evolutionary models, we never get this close because in the true model there are (many) more than three dimensions.

Useless parameters (dimensions)

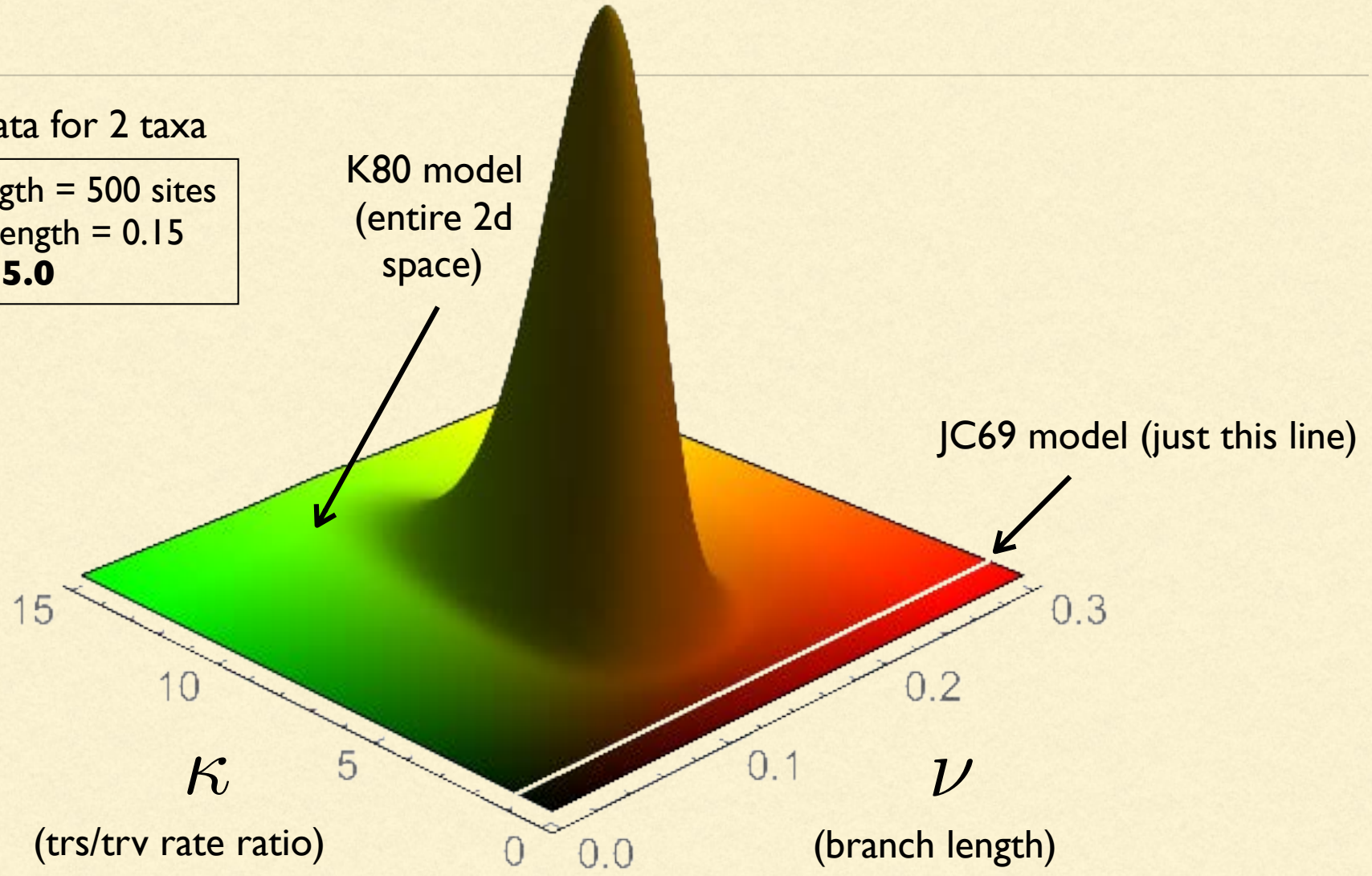
If the star is on the floor, adding a 3rd dimension does not increase goodness-of-fit.



Likelihood surface when K80 true

Simulated data for 2 taxa

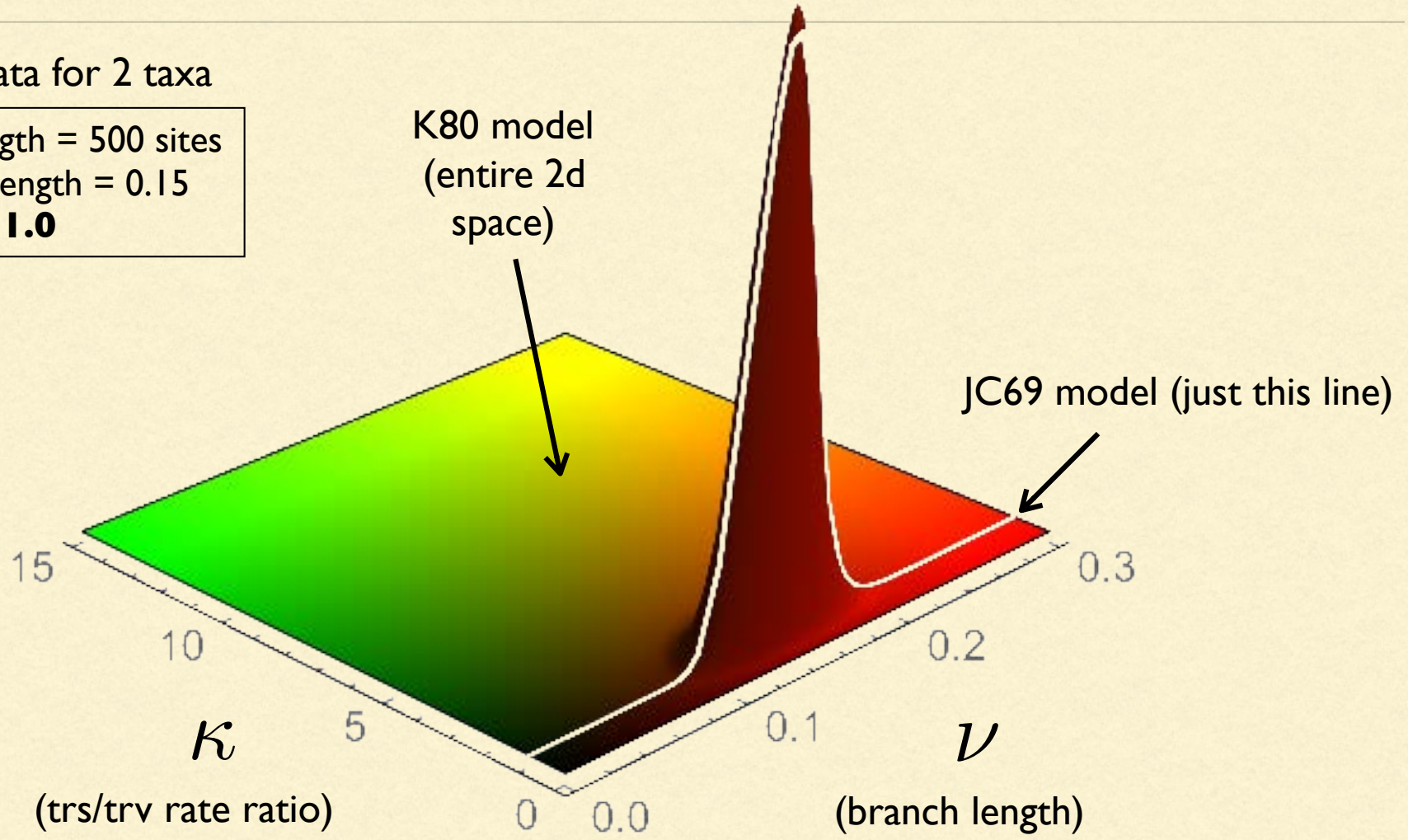
sequence length = 500 sites
true branch length = 0.15
true kappa = **5.0**



Likelihood surface when JC69 true

Simulated data for 2 taxa

sequence length = 500 sites
true branch length = 0.15
true kappa = **1.0**



Why do models matter?

Statistical consistency

Consistency means adding more data (i.e. sites) yields MLEs closer to the truth

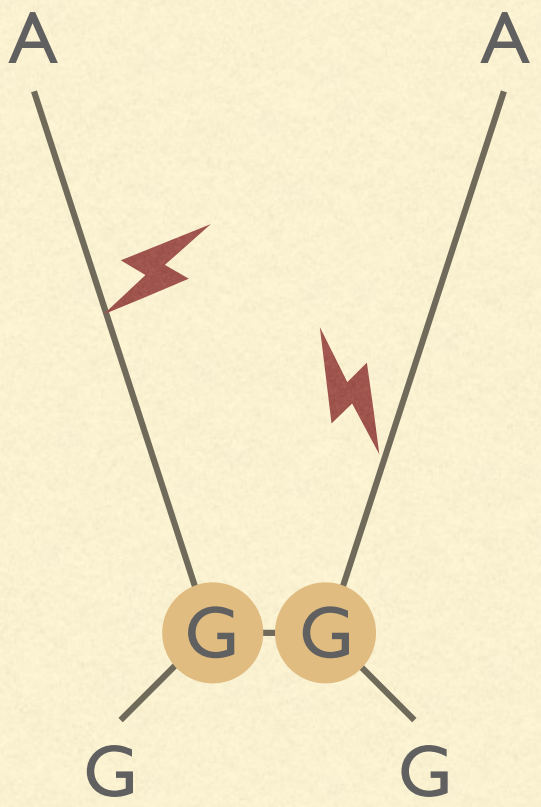
Statistical efficiency

Efficient estimators get close to the truth with fewer sites

A poor choice of model may be statistically inconsistent (positively misleading) or statistically inefficient

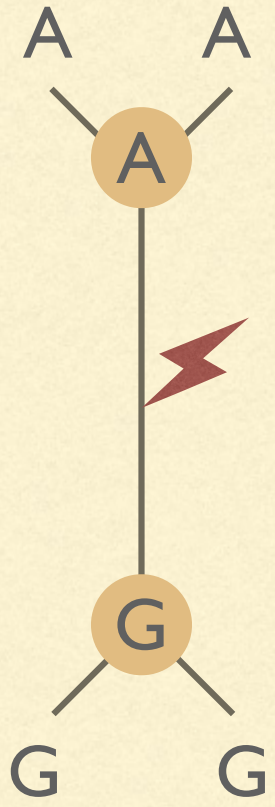
The "Felsenstein Zone"

Convergence explanation



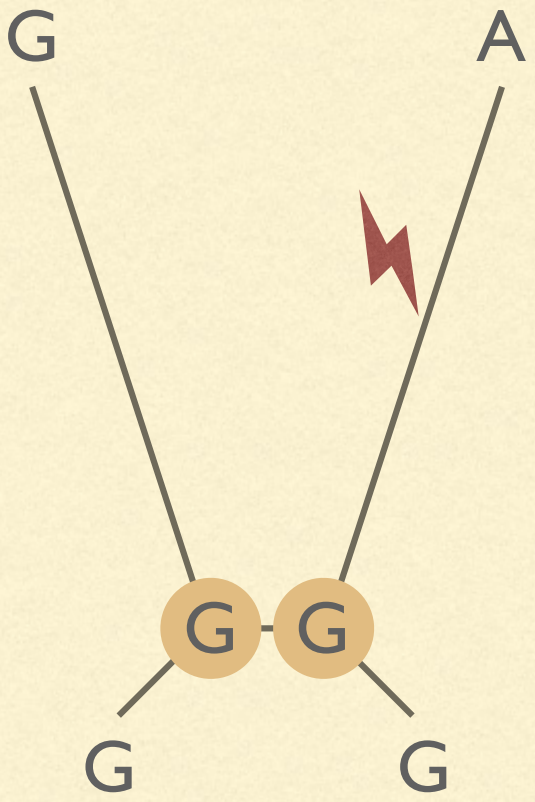
Parsimony methods prefer inheritance even if convergence explanation is the truth

Inheritance explanation



Felsenstein 1978

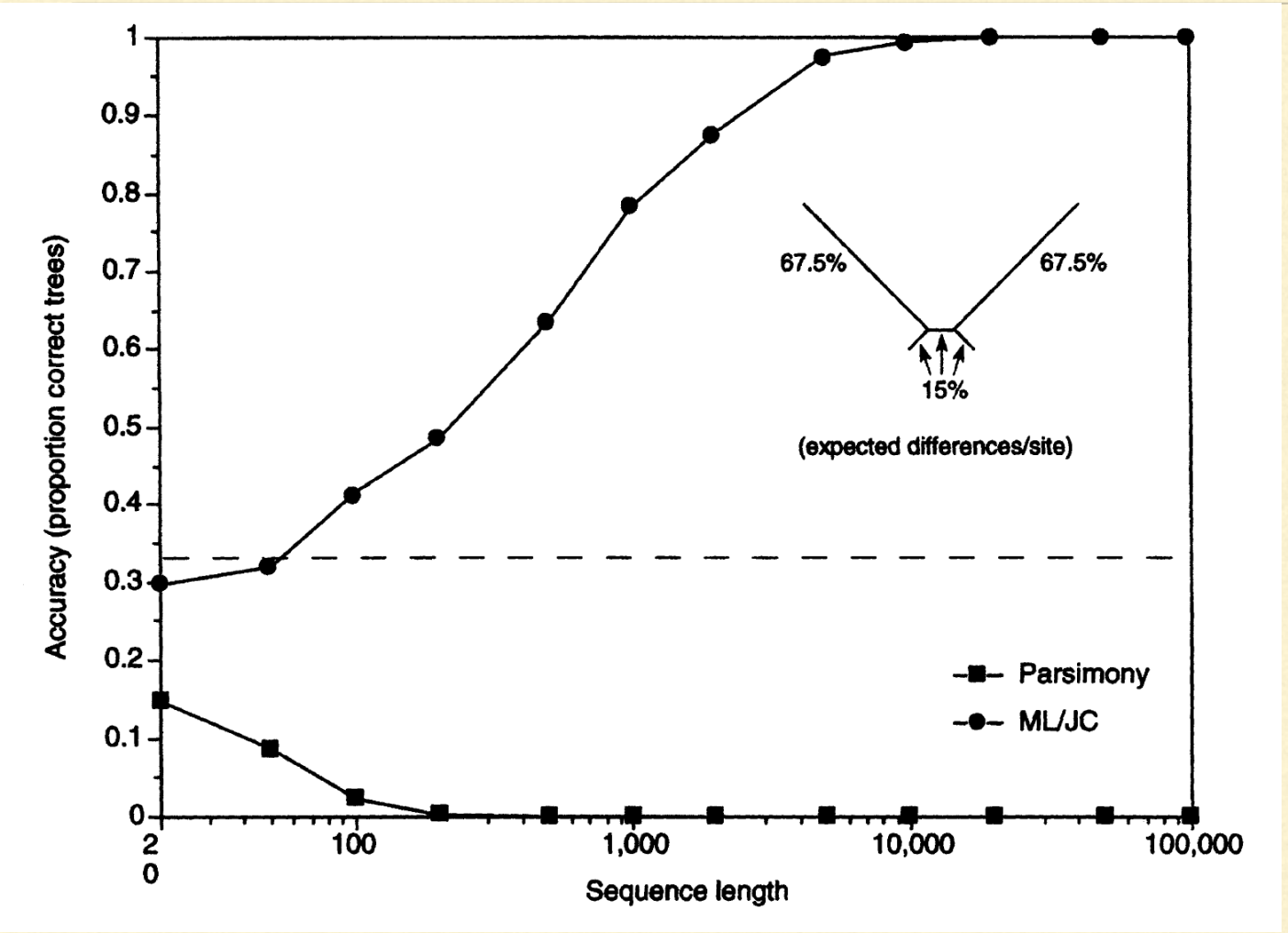
The "Felsenstein Zone"



Autapomorphous sites (only one leaf has a different state) provide evidence that these two edges are longer than the other three.

Felsenstein 1978

Statistical (in)consistency

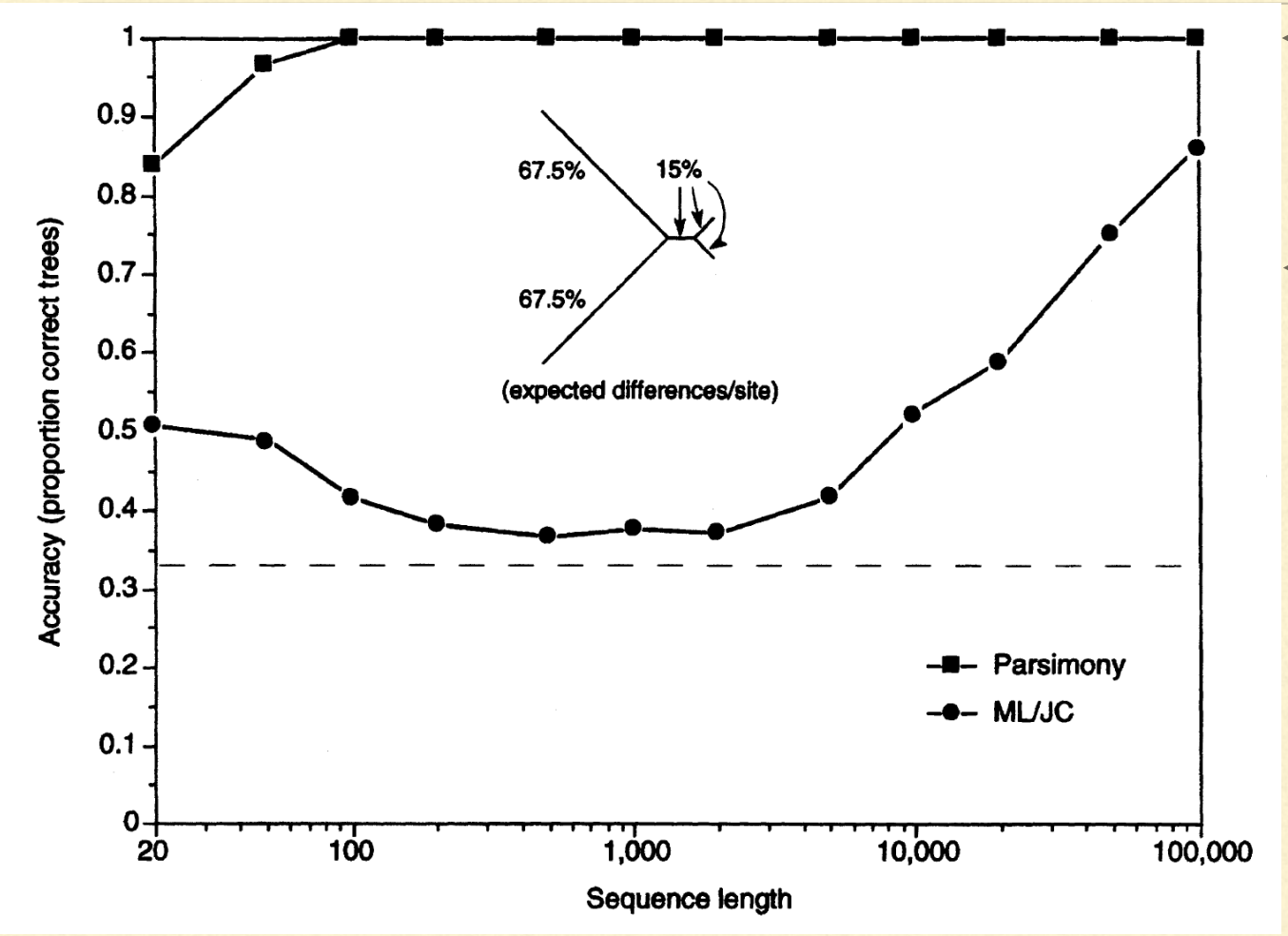


← consistency

← inconsistency

Swofford et al. 2001

What if long edges are together?



← consistent and very efficient!

← consistent but less efficient

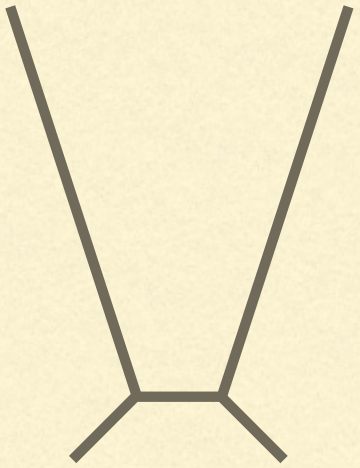
both are **biased**

should not be getting the tree correct over 80% of the time with only 20 sites given substitutional near-saturation on the 2 long edges

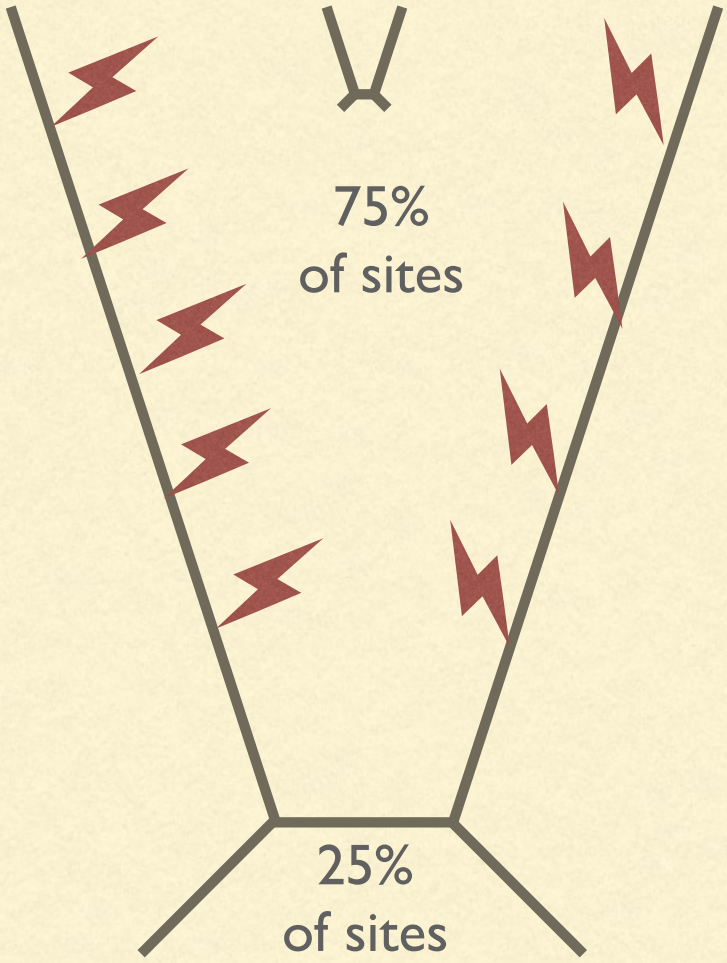
Swofford et al. 2001

Models are important!

Likelihood methods are statistically consistent (but possibly biased) when the model is correct. But what if the model misses an important feature of the data?



Equal Rates

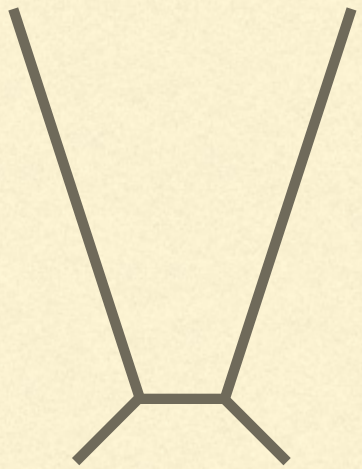


Rate Heterogeneity

Models are important!

If the rate heterogeneity model is the true model...

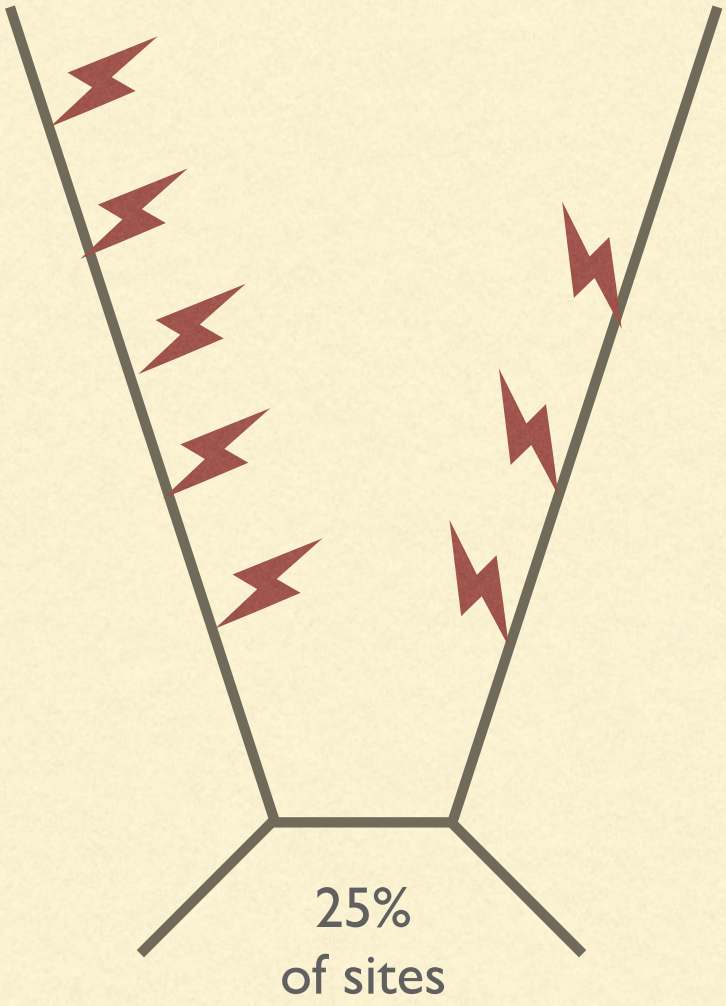
75%
of sites

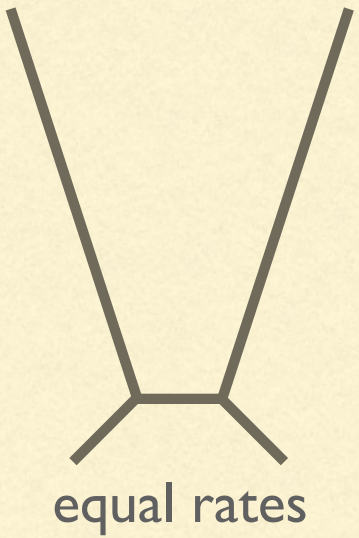


equal rates

...then for 75% of sites, the equal rates model assumes there was more evolution along all edges than really occurred...

Models are important!



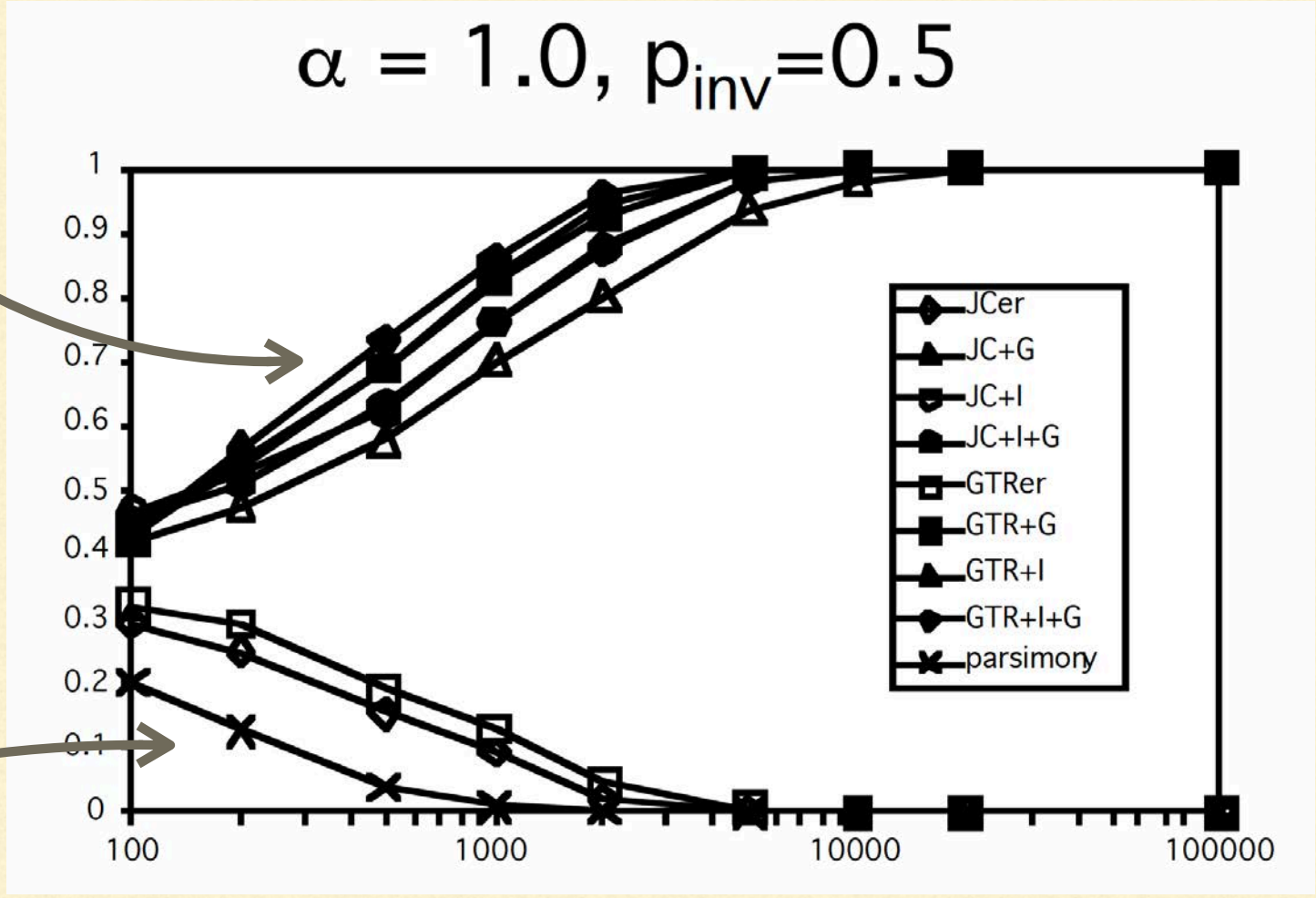


...and for 25% of sites, the equal rates model assumes there was less evolution along all edges than really occurred

This favors the inheritance explanation

Models that allow rates to vary across sites are statistically consistent

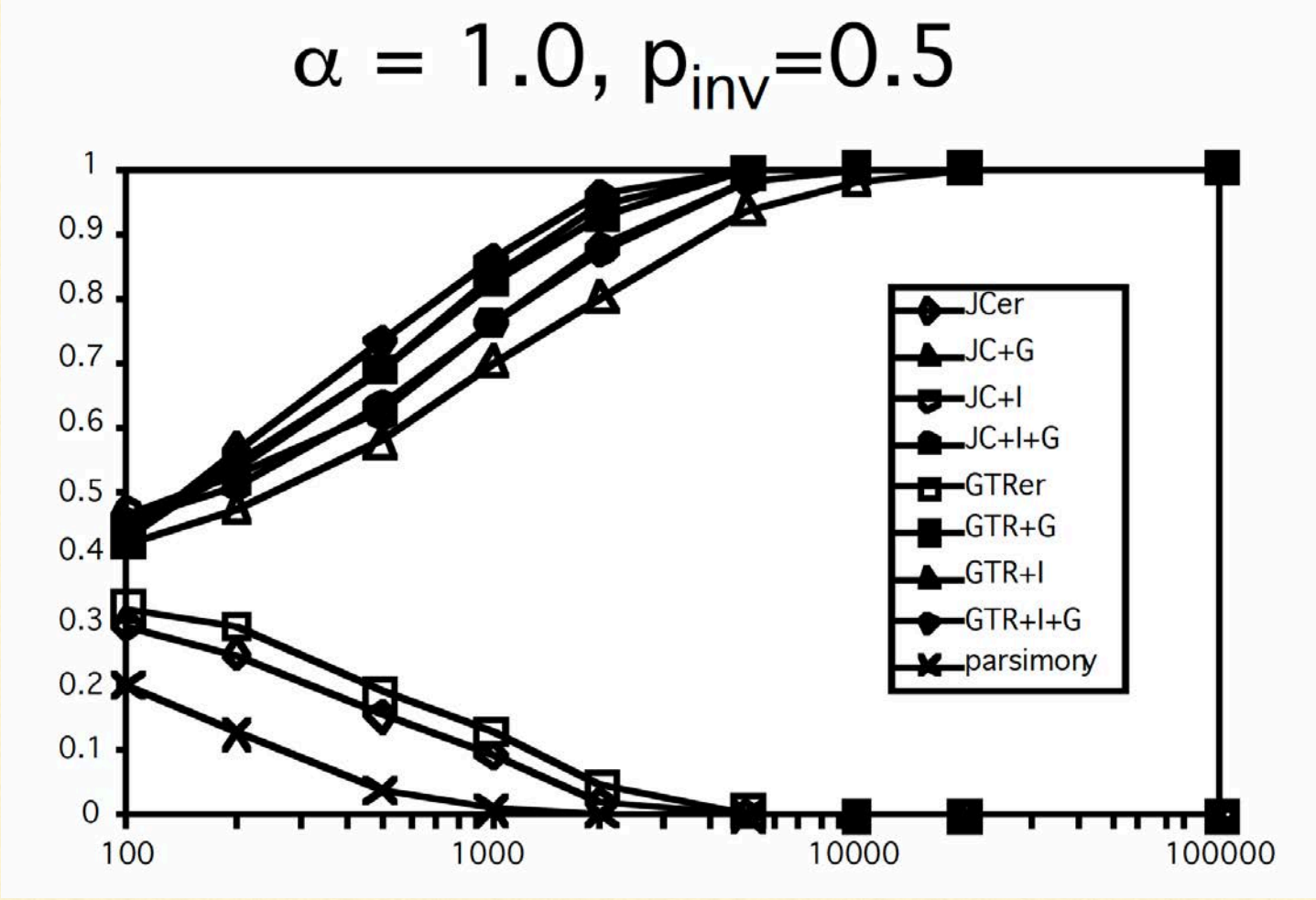
Data simulated with rate heterogeneity



Models that assume sites all evolve at the same rate are statistically inconsistent

Sullivan and Swofford 2001

Take-home message: models need not be perfect, but critical factors affecting sequence evolution must be accommodated



Sullivan and Swofford 2001

Assessing model fit if models nested

Likelihood ratio test statistic:

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

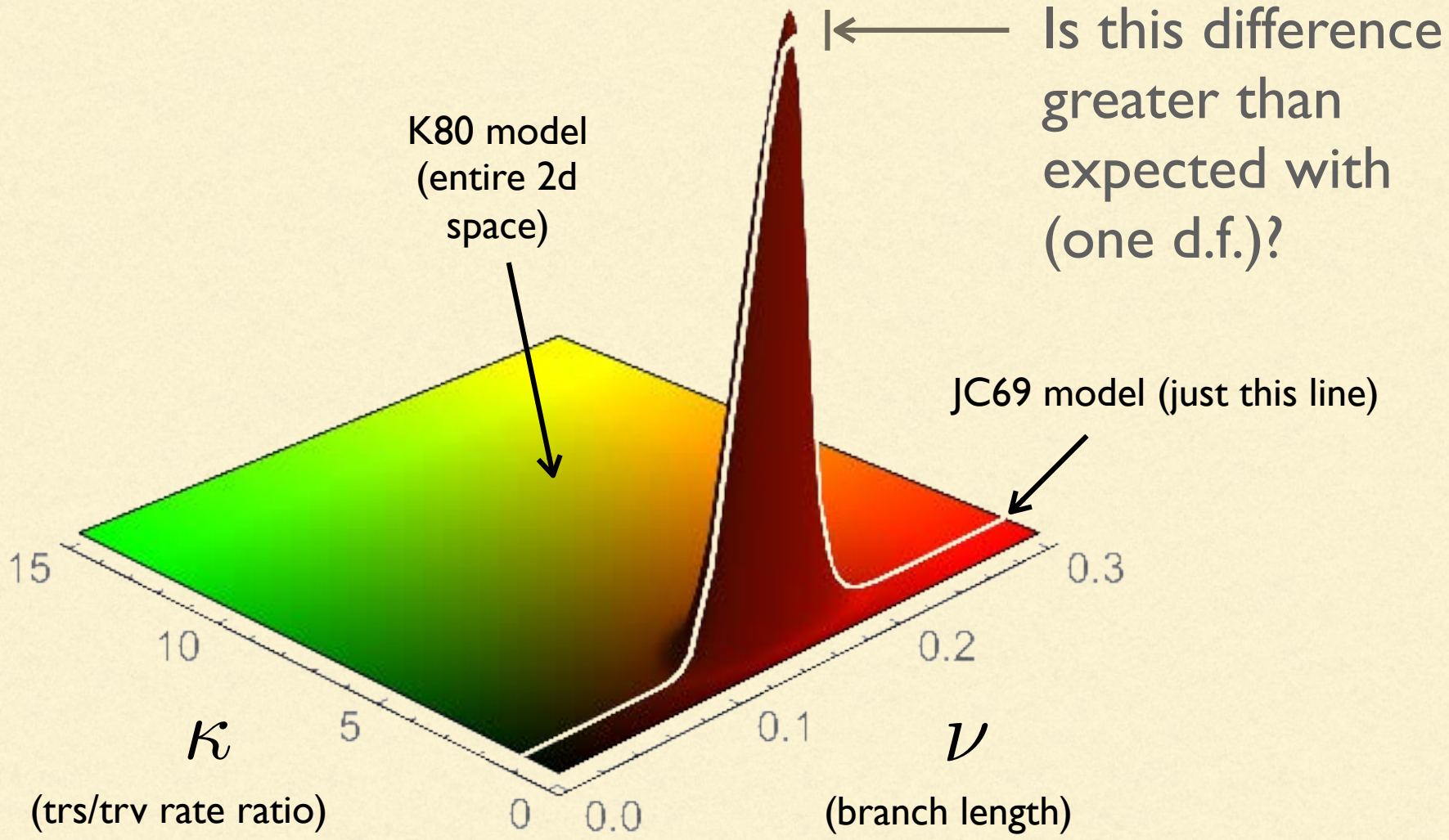
constrained model

unconstrained model

maximized log-likelihood

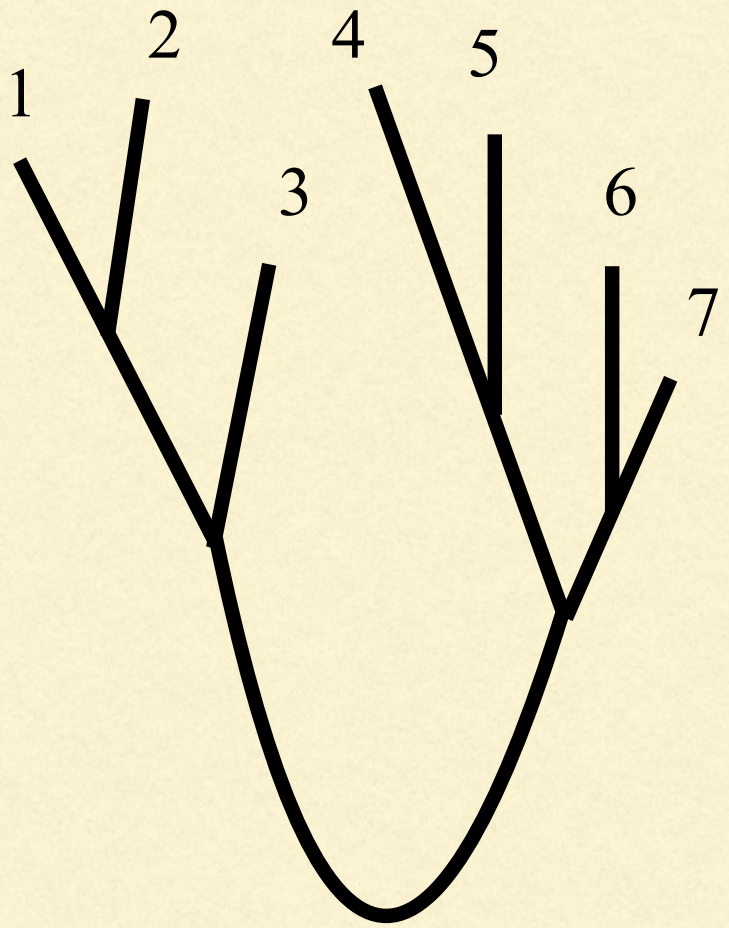
chi-squared with d.f. equal to difference in number of free parameters

Likelihood surface when JC69 true



Testing the molecular clock

Unconstrained model: need to estimate $2n$-3 = 11 edge lengths

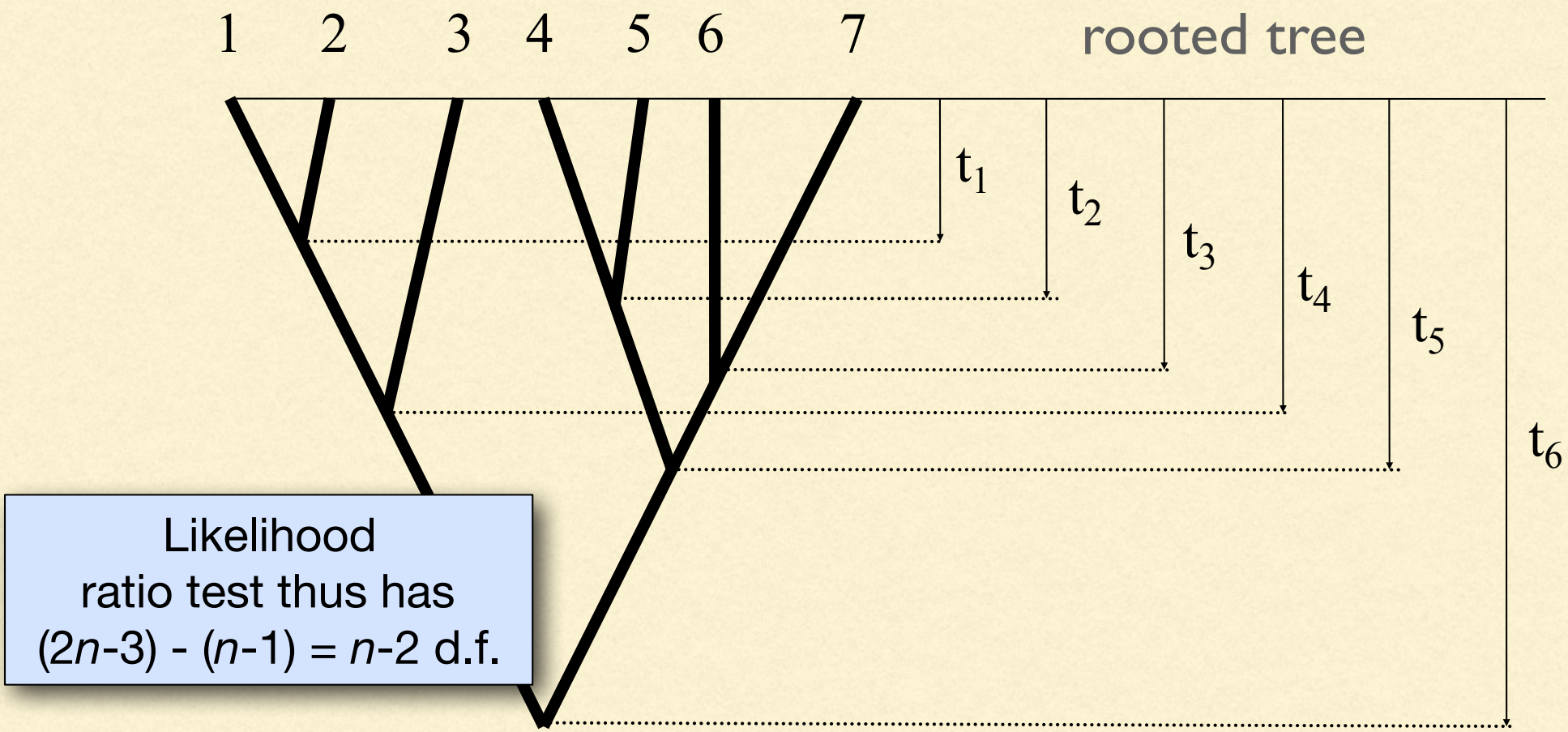


unrooted tree

Felsenstein 1983

Testing the molecular clock

Unconstrained model: need to estimate $2n$-3 = 11 edge lengths
Constrained model: need to estimate n-1 = 6 node depths



Likelihood ratio test thus has $(2n\text{-}3) - (n\text{-}1) = n\text{-}2$ d.f.

Felsenstein 1983

Likelihood Ratio Test

First 32 nucleotides of the ψη-globin gene of gorilla:

GAAGTCCTTGAGAAATAAACTGCACACACTGG

Find *maximum* logL under F81 (unconstrained) model:

$$\begin{aligned}\log L &= 12\log(\pi_A) + 7\log(\pi_C) + 7\log(\pi_G) + 6\log(\pi_T) \\ &= 12\log(0.375) + 7\log(0.219) + 7\log(0.219) + 6\log(0.187) \\ &= -43.1\end{aligned}$$

Find *maximum* logL under JC69 (constrained) model:

$$\begin{aligned}\log L &= 12\log(\pi_A) + 7\log(\pi_C) + 7\log(\pi_G) + 6\log(\pi_T) \\ &= 12\log(0.25) + 7\log(0.25) + 7\log(0.25) + 6\log(0.25) \\ &= -44.4\end{aligned}$$

F81 fits better (-43.1 > -44.4), but not significantly better (P = 0.457, chi-squared with 3 d.f.)

Akaike Information Criterion (AIC)

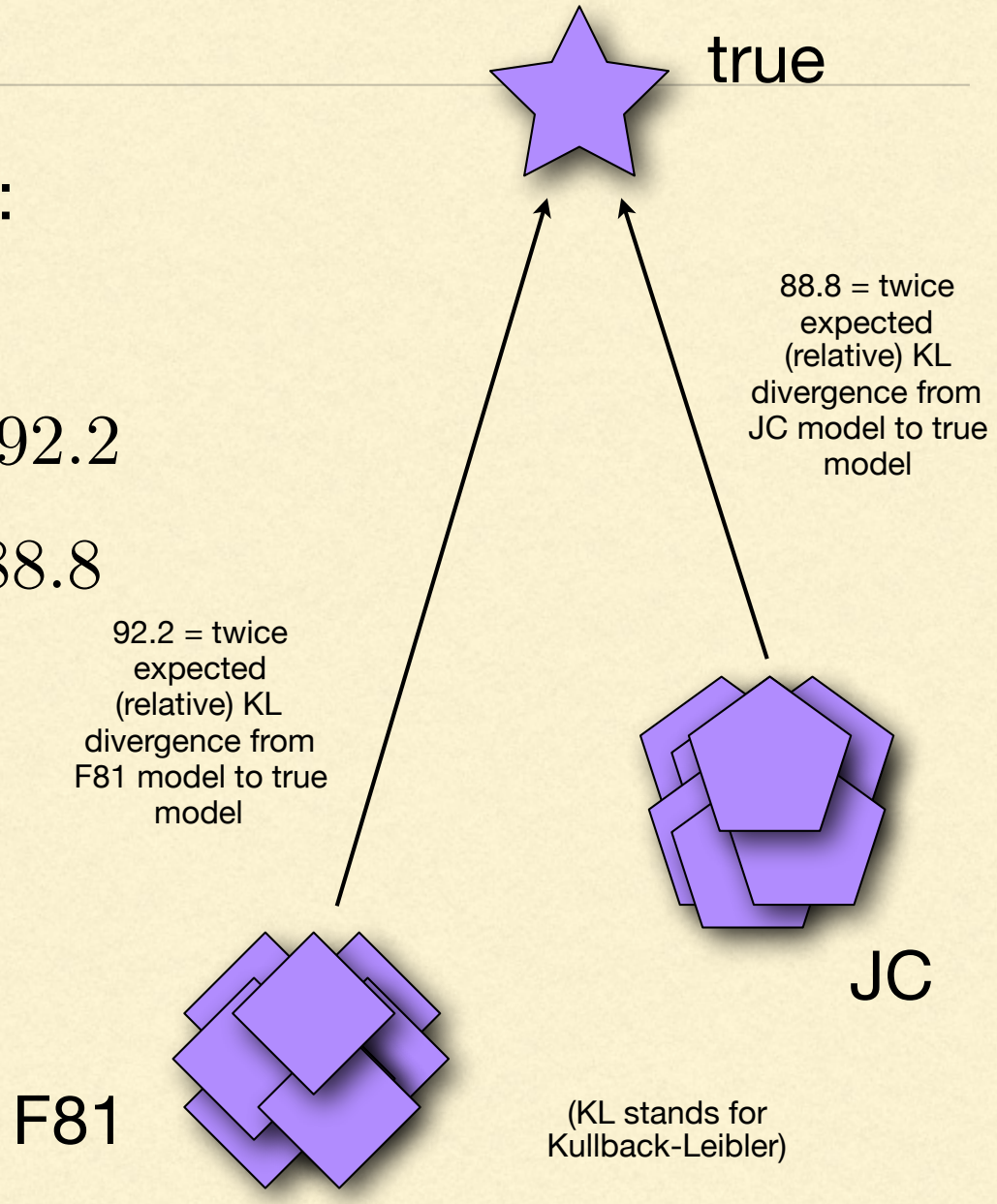
Calculate AIC for each model:

$$AIC = 2k - 2\log L_{\max}$$

$$AIC_{\mathrm{F81}} = 2(3) - 2(-43.1) = 92.2$$

$$AIC_{\mathrm{JC}} = 2(0) - 2(-44.4) = 88.8$$

The constrained model (JC) is a better choice than the unconstrained model (F81) according to AIC



AIC vs. AICc

AIC:

$$AIC = 2k - 2\log L_{\max}$$

AIC corrected for small sample size:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Usually makes little difference, but it doesn't hurt to always use AICc

AIC vs. BIC

AIC:

$$AIC = 2k - 2\log L_{\max}$$

BIC:

$$BIC = k\log n - 2\log L_{\max}$$

BIC penalizes complex models more than AIC

$$\log n > 2$$

$$n > e^2 = 7.389$$

If true model is included, BIC performs better than AIC at choosing the true model.

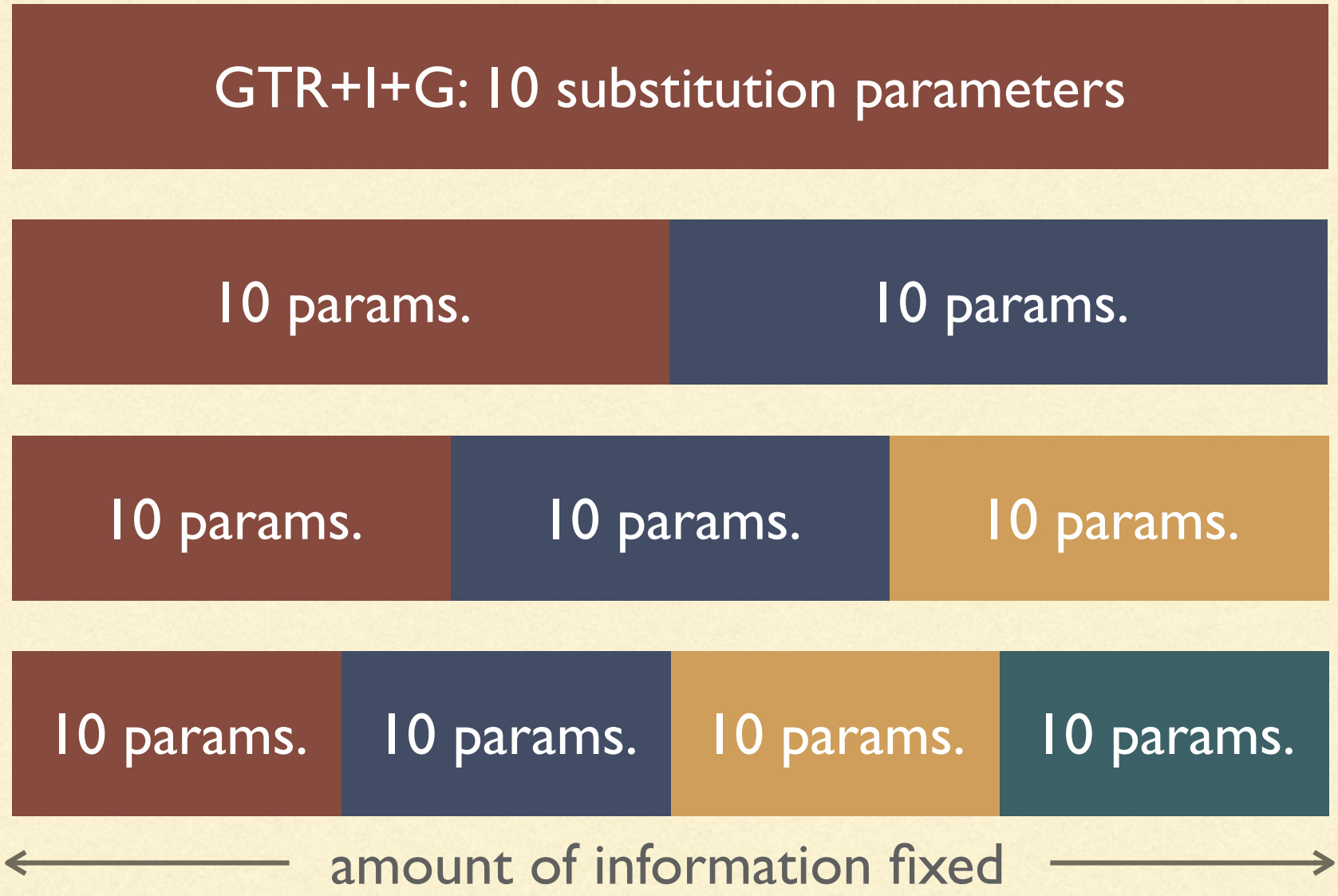
That said, the true model is never included (except in simulation experiments)

Partitioned models

Data can be divided into subsets that each have their own model of evolution. Here are some common ways to partition data:

- by codon position (usually quite effective)
- by gene (less effective)
- by gene and codon (risks overpartitioning)
- codon vs. noncoding
- stems vs. loops (but see Simon et al. 2006)
- by substitution rate (probably never a good idea: yields subsets having essentially zero information)

Overpartitioning



Naïve approach

Estimate model independently for each subset

Subset 1: K80 model kappa: 4.0

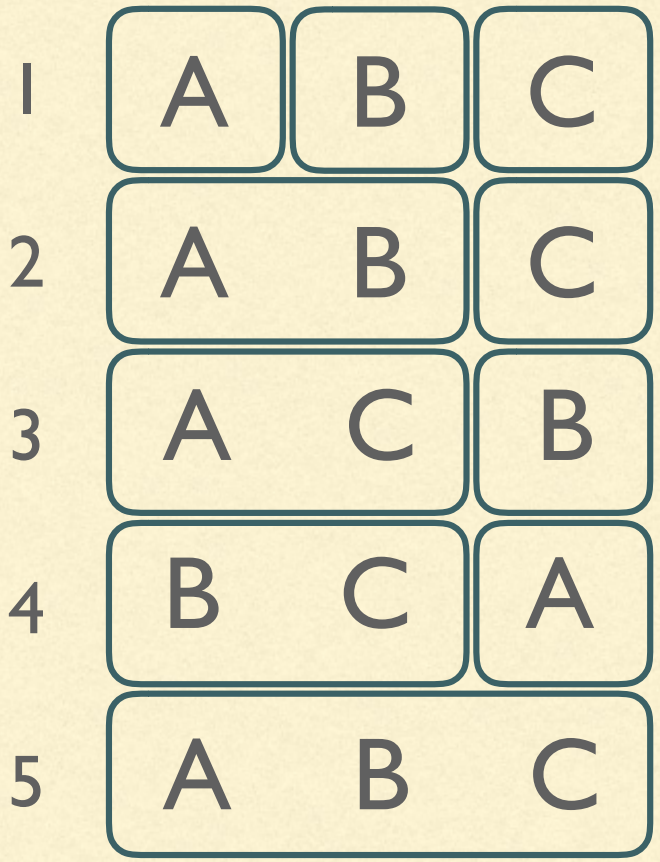
Subset 2: GTR model base freqs: 0.25, 0.249, 0.251, 0.25
rmatrix: 0.99, 4.05, 1.05, 0.95, 3.90, 1.10

These models are nearly identical - arguably better to not partition and use the K80 model for both

Enumerating partitionings

There are 5 distinct partition schemes for 3 subsets (A, B, C):

A suite of (e.g. 56) models is evaluated for each subset of each partition scheme



each has own model

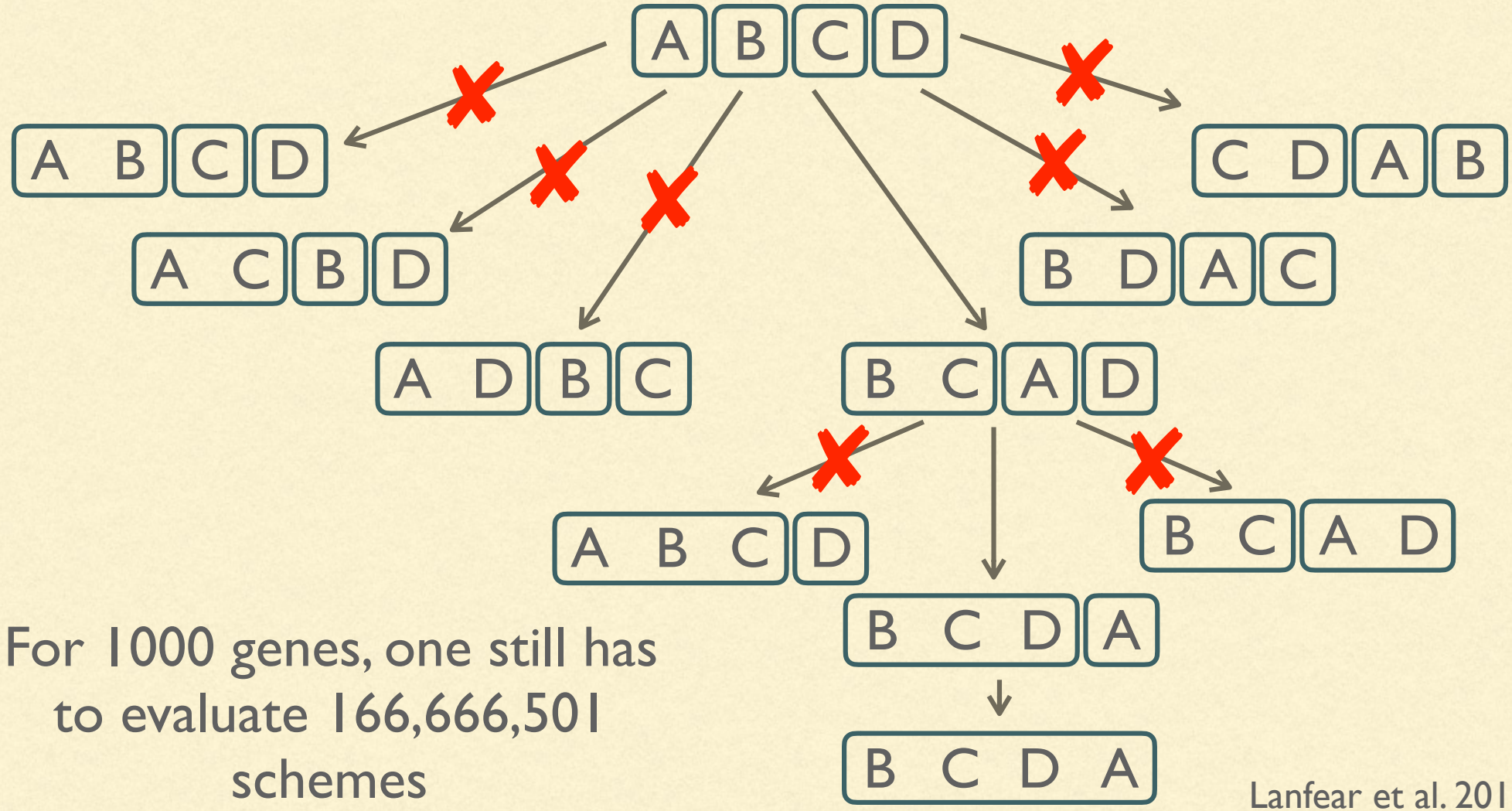
two of the three share a model

all share one model

Can get out of hand quickly

SUBSETS	PARTITIONINGS
2	2
3	5
4	15
5	52
10	115975
20	51724158235372
30	846749014511809332450147
40	1574505883912049312893243447025310 67

Greedy partitioning algorithm



For 1000 genes, one still has to evaluate 166,666,501 schemes

Lanfear et al. 2012

Clustering (for large problems)

- Li, Lu, and Ortí (2008)

  Estimate model parameters on a shared model; similar subsets have similar parameter estimates and cluster together (must use same model for all subsets)

- Frandsen et al. (2015)
- Lanfear et al. (2016): PartitionFinder 2

  Hierarchical (or non-hierarchical k-means) clustering using the same idea as Li et al. (very efficient implementation)

Literature cited

- U Bergthorsson, KL Adams, B Thomason, and JD Palmer. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature 424:197-201. (https://doi.org/10.1038/nature01743)
- J Felsenstein. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology* 27(4):401-410. (https://doi.org/10.1093/sysbio/27.4.401)
- J Felsenstein. 1983. Statistical inference of phylogenies. Journal of the Royal Statistical Society A 146:246-272. (https://doi.org/10.2307/2981654)
- PB Frandsen, B Calcott, C Mayer, and R Lanfear. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evolutionary Biology 15:13. (https://doi.org/10.1186/s12862-015-0283-7)
- R Lanfear, B Calcott, SYW Ho, and S Guindon. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution 29(6):1695-1701 (https://doi.org/10.1093/molbev/mss020)
- R Lanfear, PB Frandsen, AM Wright, T Senfeld, and B Calcott. 2016. Partitionfinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular Biololgy and Evolution 34(3):772-773. (https://doi.org/10.1093/molbev/msw260)
- PO Lewis, MH Chen, L Kuo, LA Lewis, K Fučíková, S Neupane, YB Wang, and D Shi. 2016. Estimating Bayesian phylogenetic information content. *Systematic Biology* 65(6):1009-1023. (https://doi.org/10.1093/sysbio/syw042)
- C Li, G Lu, and G Ortí. 2008. Optimal data partitioning and a test case for ray finned fishes (Actinopterygii) based on ten nuclear loci. Systematic Biology 57(4):519-539. (https://doi.org/10.1080/10635150802206883)

Literature cited (continued)

- C Simon, TR Buckley, F Frati, JB Stewart, AT Beckenbach. 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. Annual Review of Ecology and Systematics 37:545-579. (https://doi.org/10.1146/annurev.ecolsys.37.091305.110018)
- J Sullivan and DL Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology* 50(5):723–729. (https://doi.org/10.1080/106351501753328848)
- DL Swofford, PJ Waddell, JP Huelsenbeck, PG Foster, PO Lewis, and JS Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology* 50(4):525-539. (https://doi.org/10.1080/10635150117959)