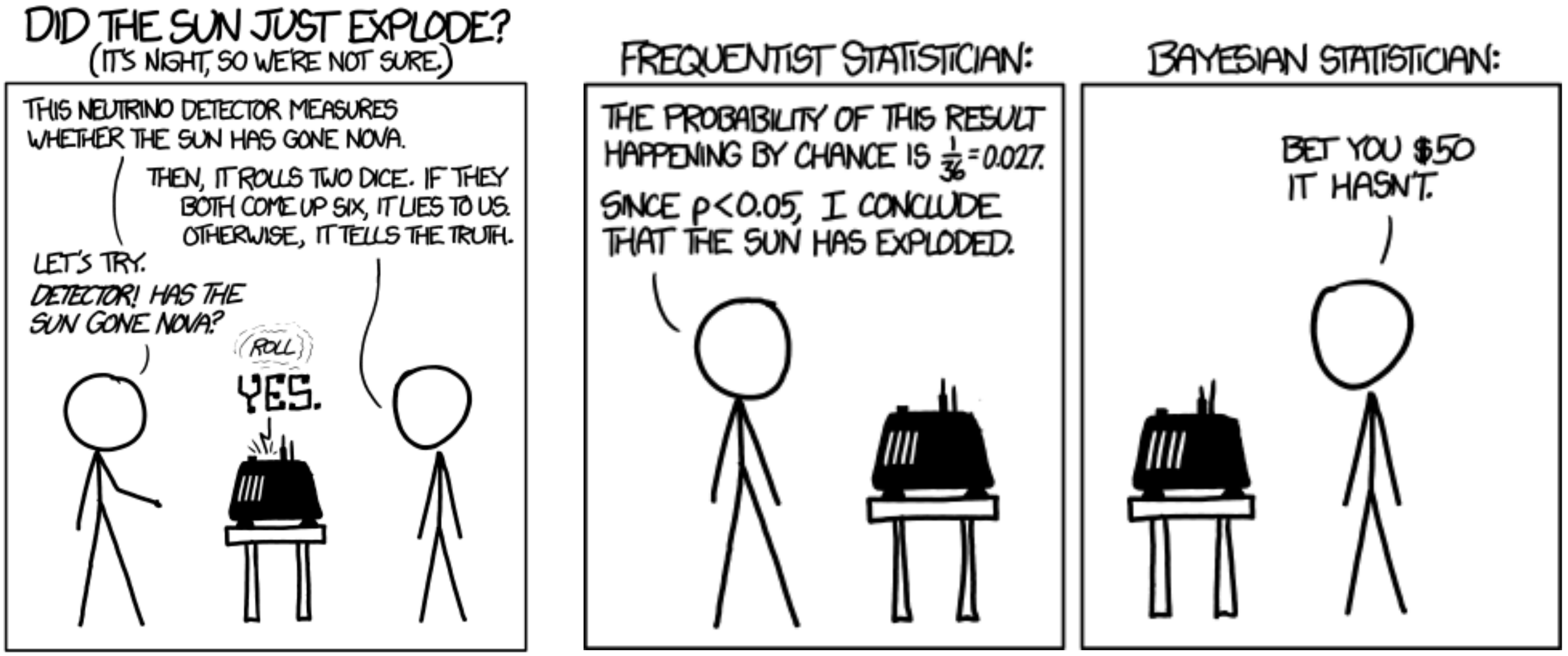
Likelihood and Bayes



http://xkcd.com/1132/

Brian O' Meara
http://www.brianomeara.info

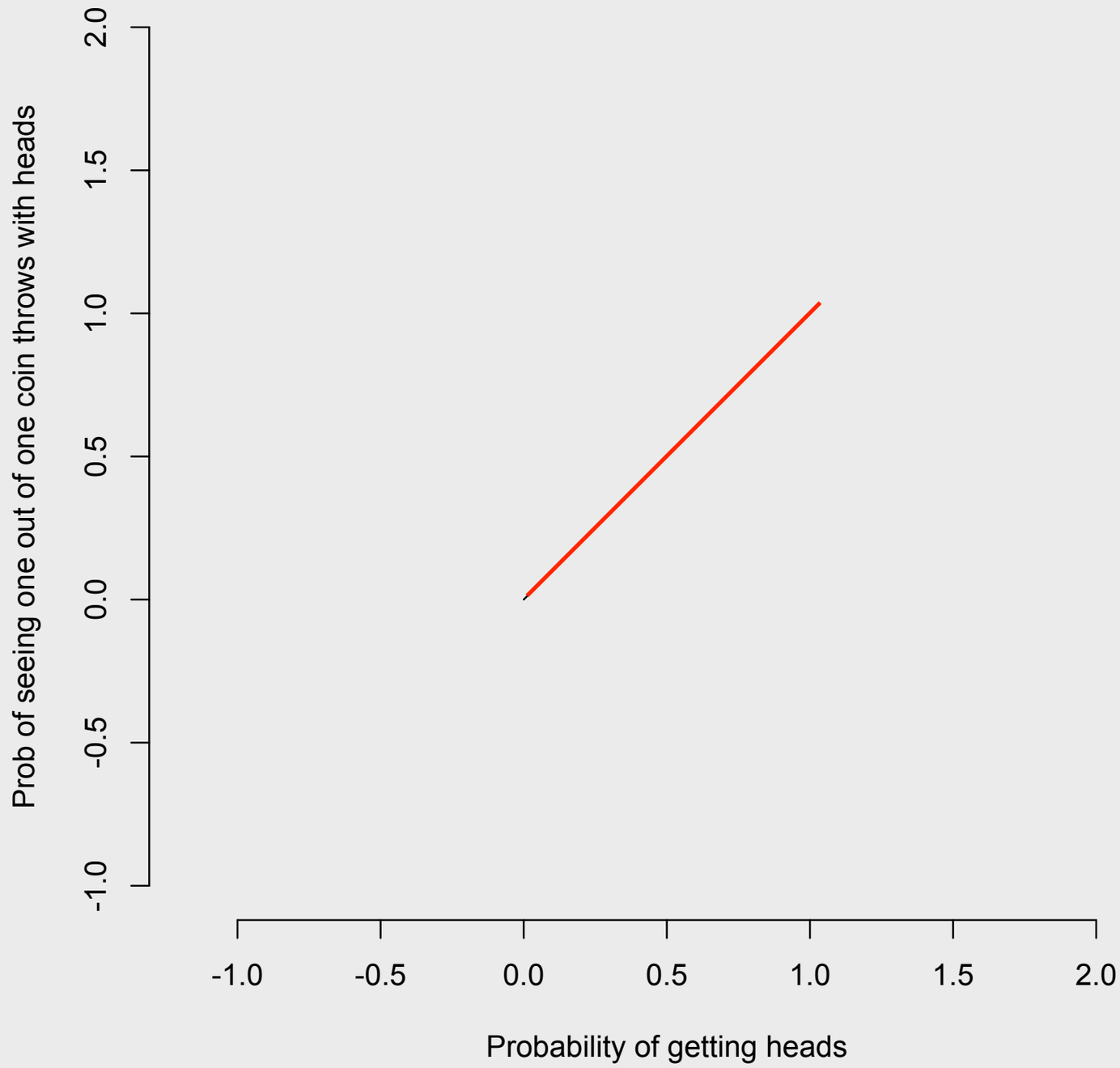
Predict the number of heads in the next 18 flips

http://tinyurl.com/gamblingforcandy

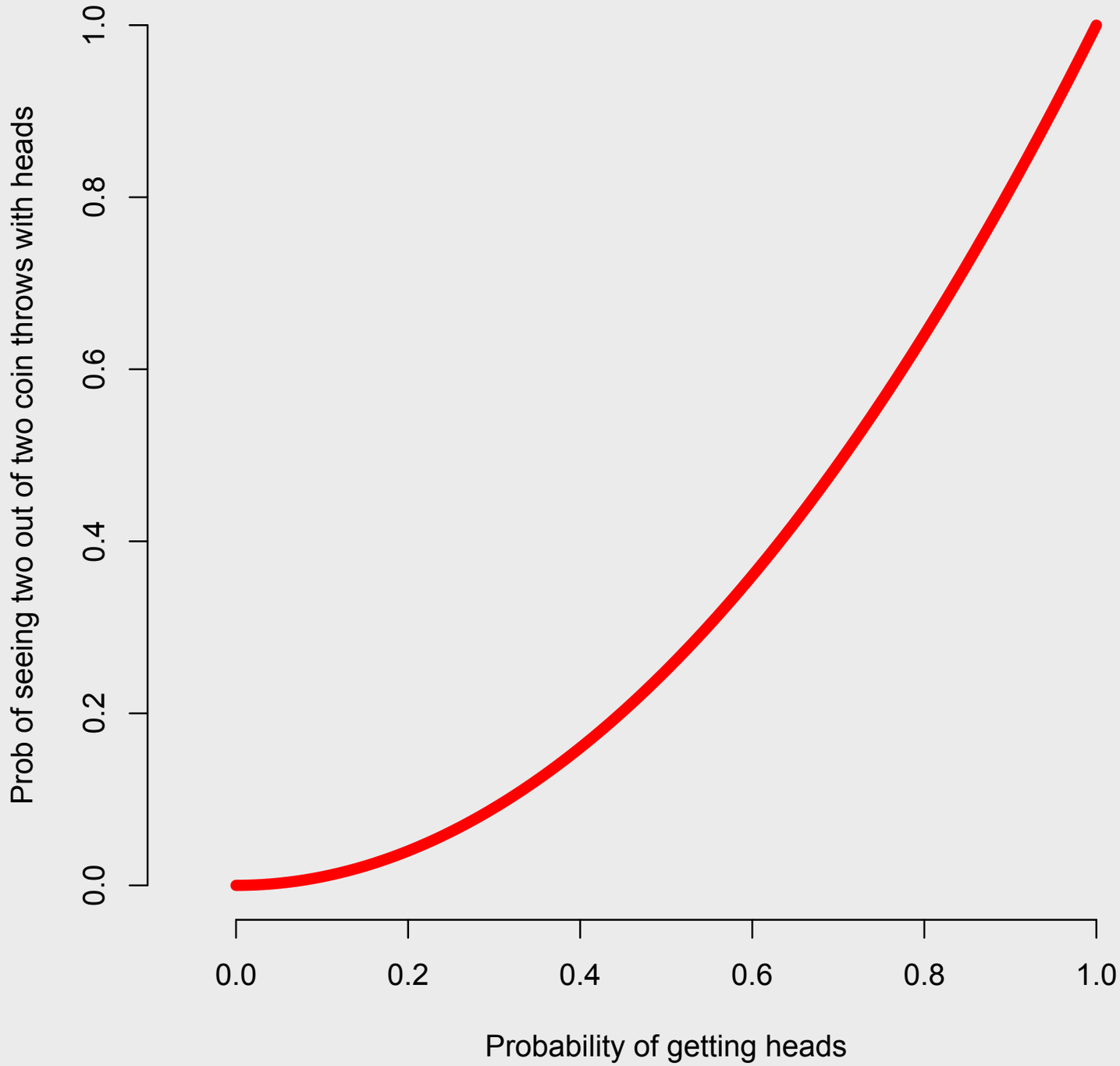
Guess right and get a candy

[regardless of how many others get it right]

Probability of getting a single heads given $p = p$



Probability of getting two heads given $p = p^2$



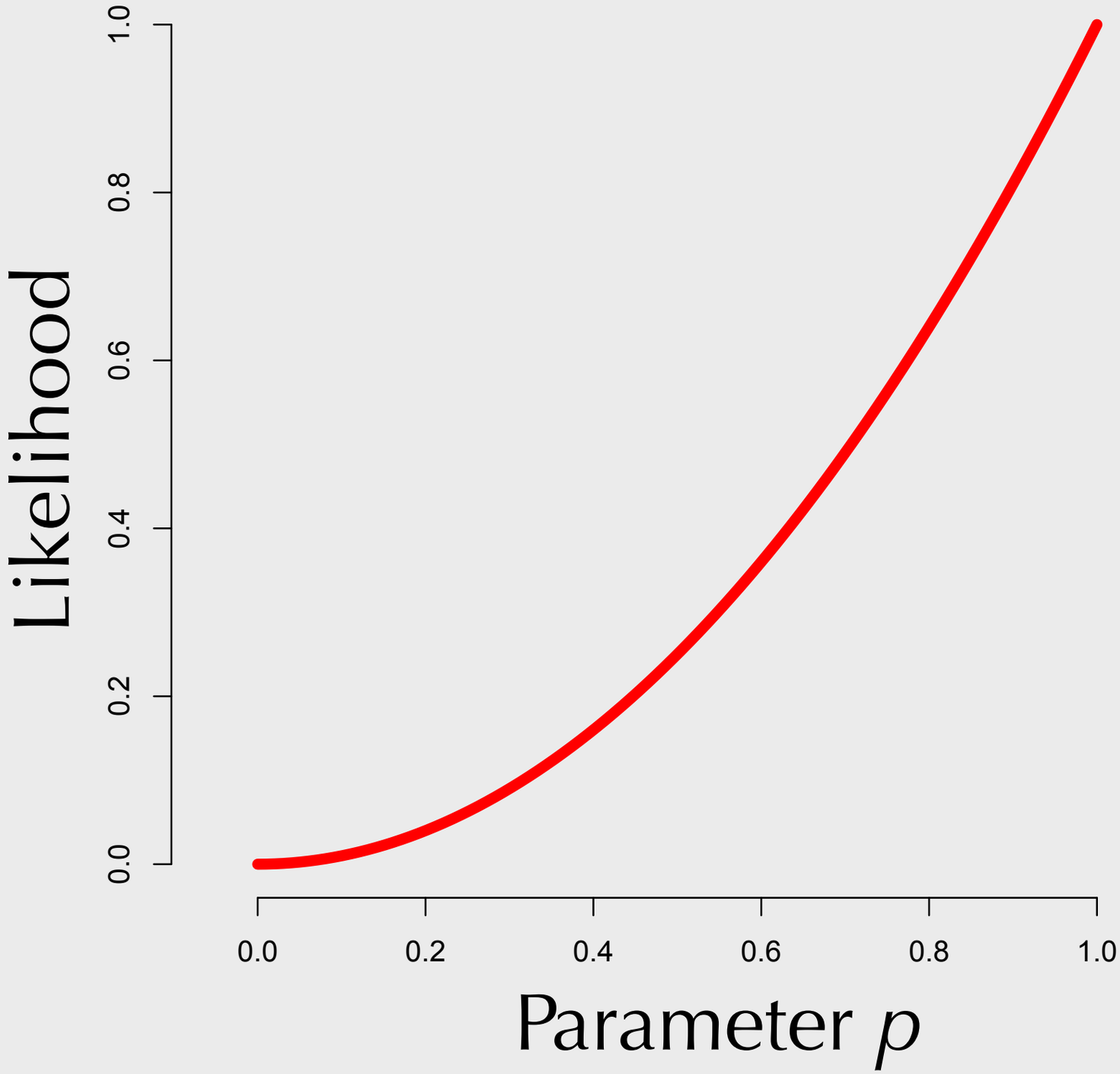
Probability of getting two heads given $p = p^2$

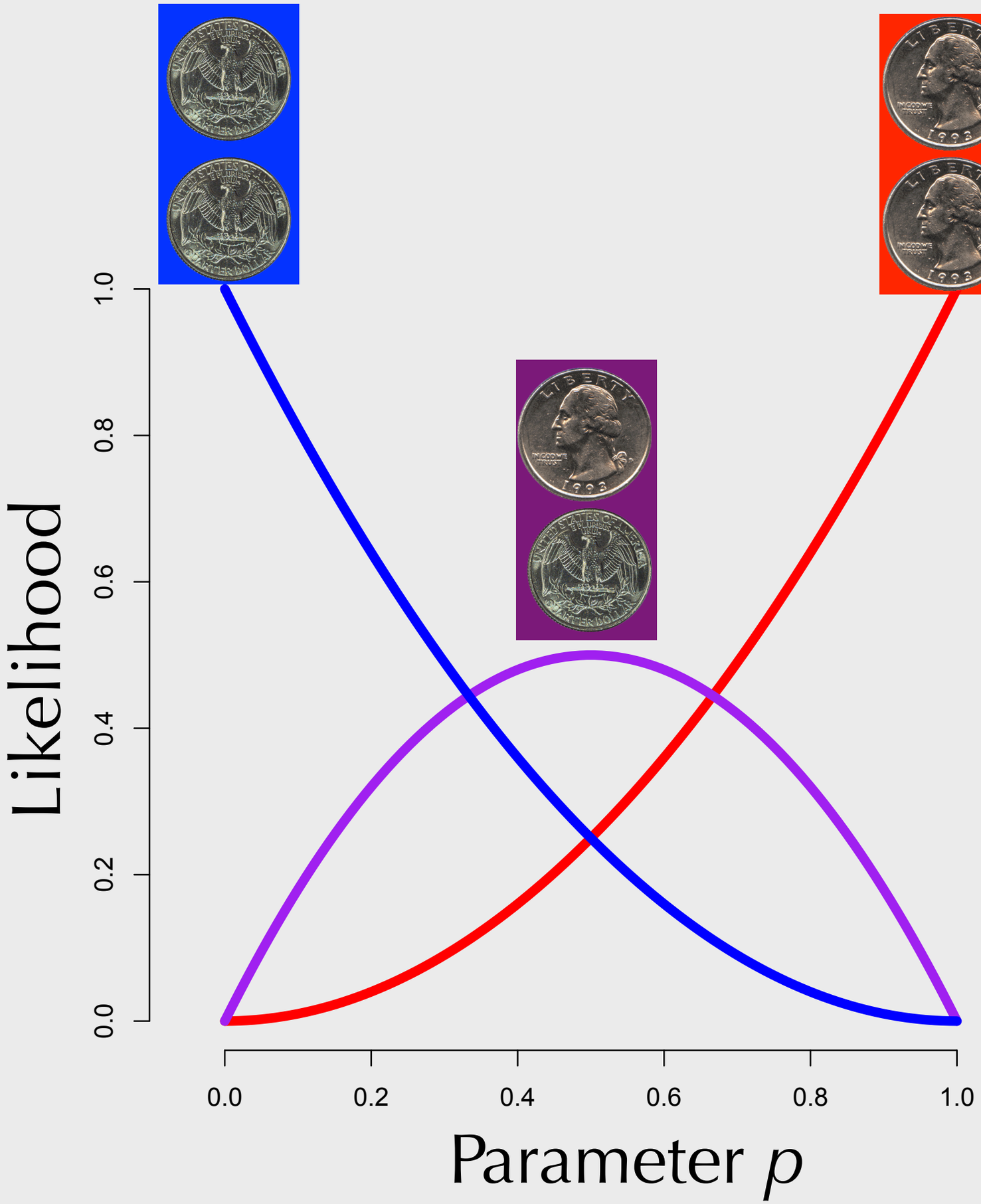
Probability(data given p) = p^2

Probability(data | p) = p^2

Likelihood(p | data) = Probability(data | p) = p^2

Probability of getting two heads given $p = p^2$



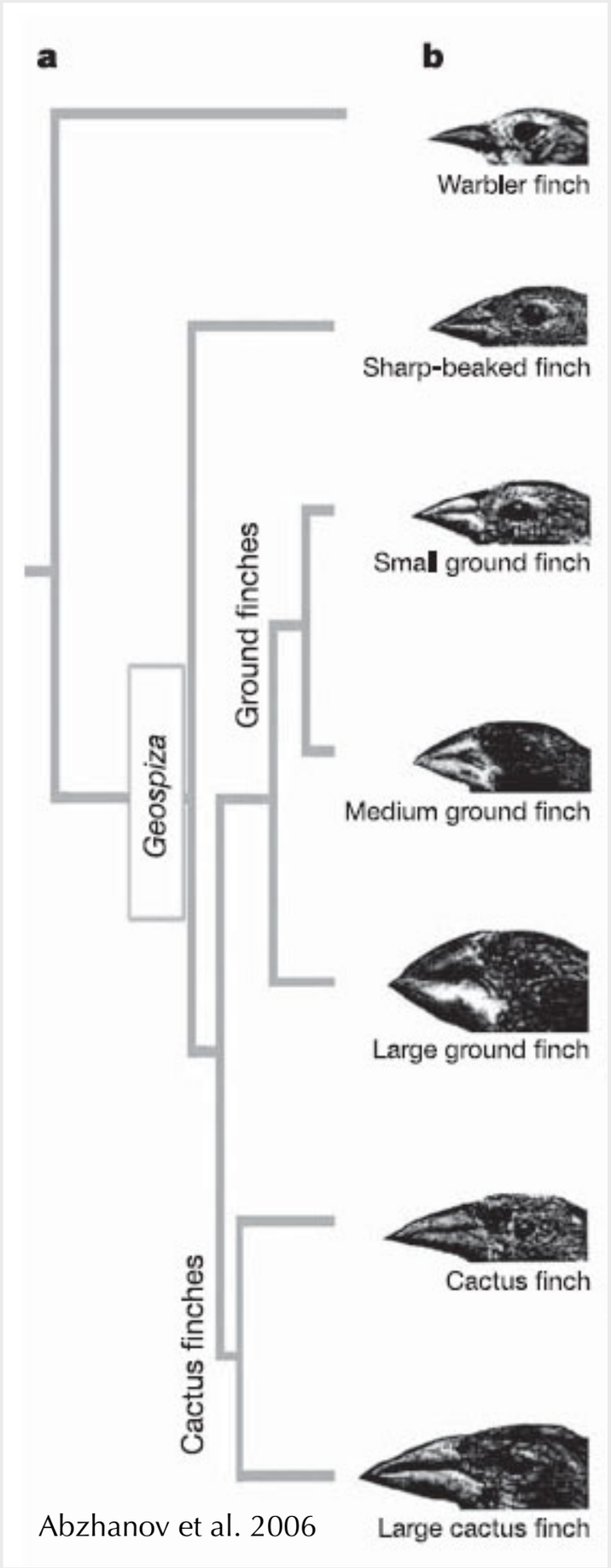

Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Parameter p

Likelihood

0.0
0.2
0.4
0.6
0.8
1.0

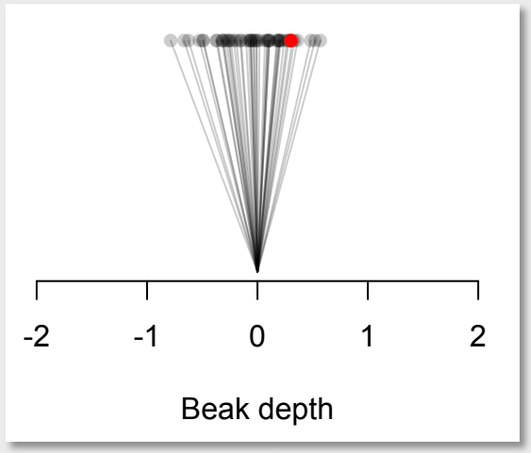
0.0
0.2
0.4
0.6
0.8
1.0

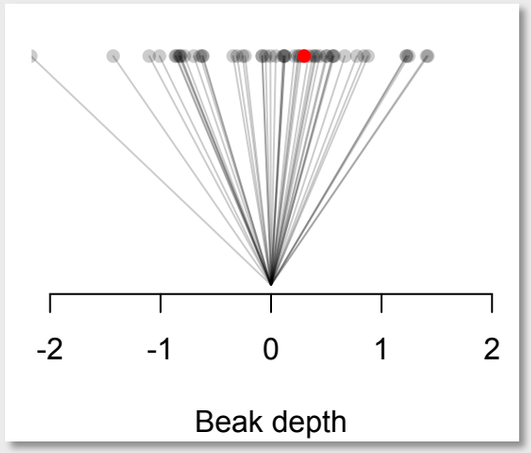
Parameter p



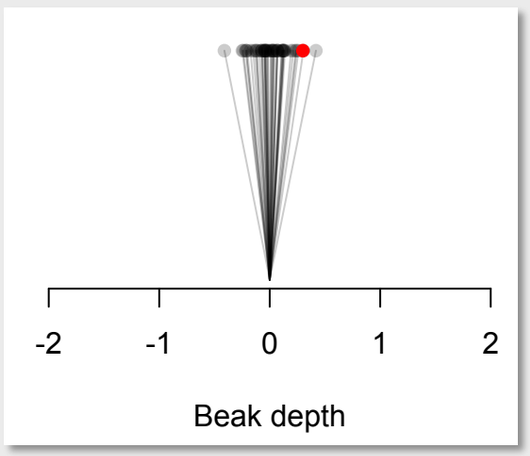
Abzhanov et al. 2006


Large cactus finch


-2
-1
0
1
2
Beak depth


-2
-1
0
1
2
Beak depth

Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Evolutionary rate


-2
-1
0
1
2
Beak depth

Properties of likelihood as a method

Consistent (given correct model)

Efficient (no other estimator has a lower mean squared error as dataset size approaches infinity)

Often biased, bias decreases with sample size

Likelihoods can get really small, even with simple fair coin

Number of throws	Number of heads	Likelihood
1	1	0.5
5	2	0.3125
100	25	0.00000019314
500	150	0.00000000000 0000000005279

Limits of your machine's precision in R:

```
noquote(unlist(format(.Machine)))
```

Likelihood ratio test

Models must be nested
(one must be a restriction of the other)

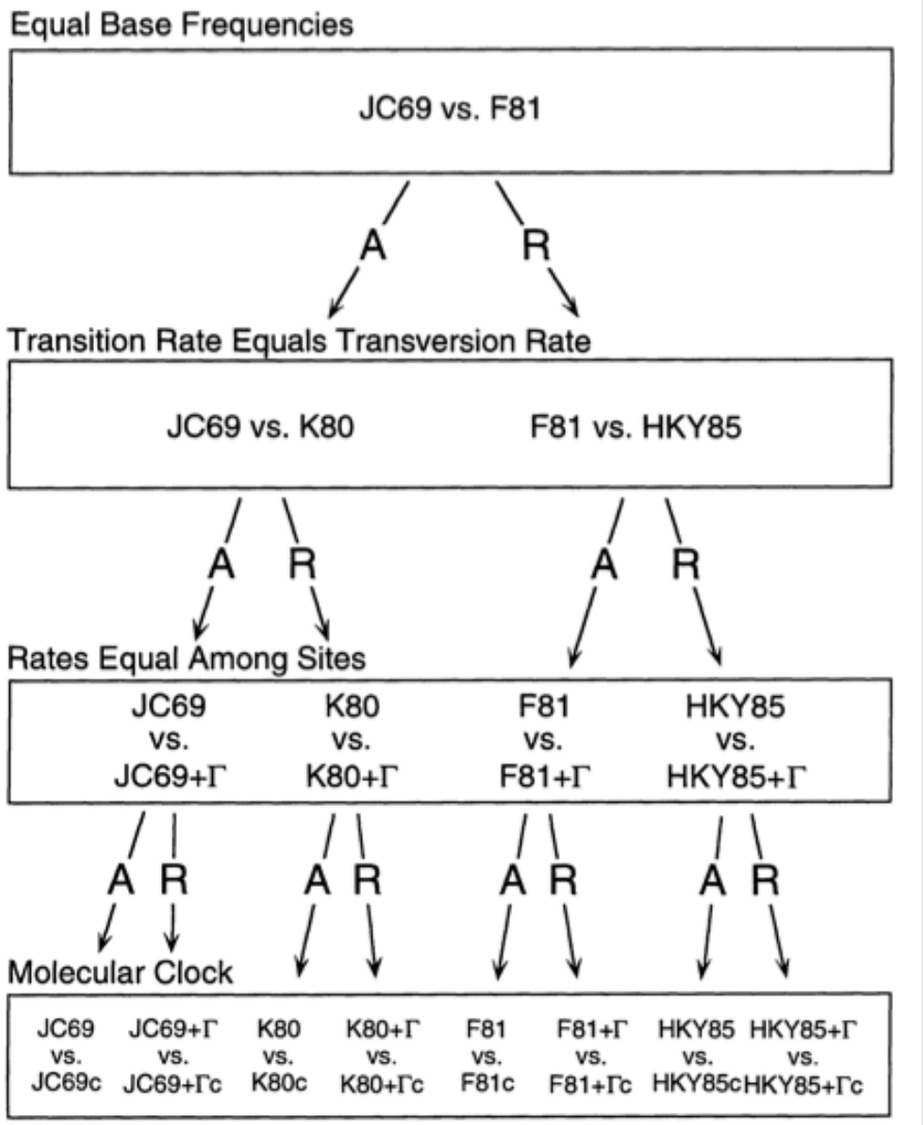


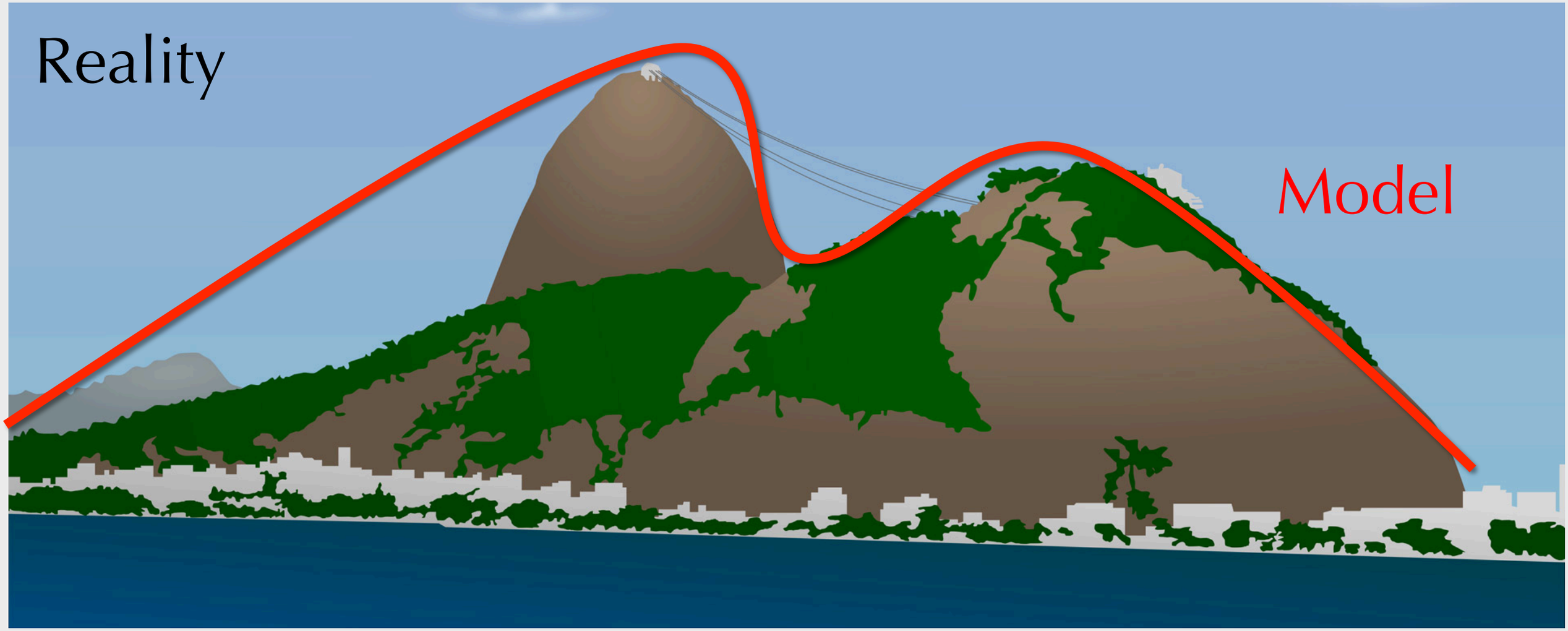
Figure 4 The hierarchy of hypotheses examined for the albumin data from five vertebrates. The parameters of the models are explained in Table 1. At each level, the null hypothesis is either accepted, "A," or rejected, "R."

Table 2 The results of likelihood ratio tests performed on the albumin DNA data from five vertebrates

Null hypothesis	Models compared	$\log L_0$	$\log L_1$	$-2\log\Lambda$	d.f.	P
Equal base frequencies	H_0: JC69 H_1: F81	−7675.86	−7667.08	17.56	3	2.78×10^{-5}
Transition rate equals transversion rate	H_0: F81 H_1: HKY85	−7667.08	−7628.03	78.10	1	9.75×10^{-19}
Equal rates among sites	H_0: HKY85 H_1: HKY85+Γ	−7628.03	−7568.56	118.94	1	0
Molecular clock	H_0: HKY85+Γc H_1: HKY85+Γ	−7573.81	−7568.56	10.5	3	1.47×10^{-2}

L_0 and L_1 denote the likelihoods under the null (H_0) and alternative (H_1) hypotheses, respectively. P represents the probability of obtaining the observed value of the likelihood ratio test statistic ($-2\log\Lambda$) if the null hypothesis were true. Because multiple tests are performed, the significance value for rejection of the null hypothesis should be adjusted using a Bonferroni correction (hence, the significance level for rejection of the null hypothesis is set to 1.25×10^{-2}).

Huelsenbeck& Crandall 1997

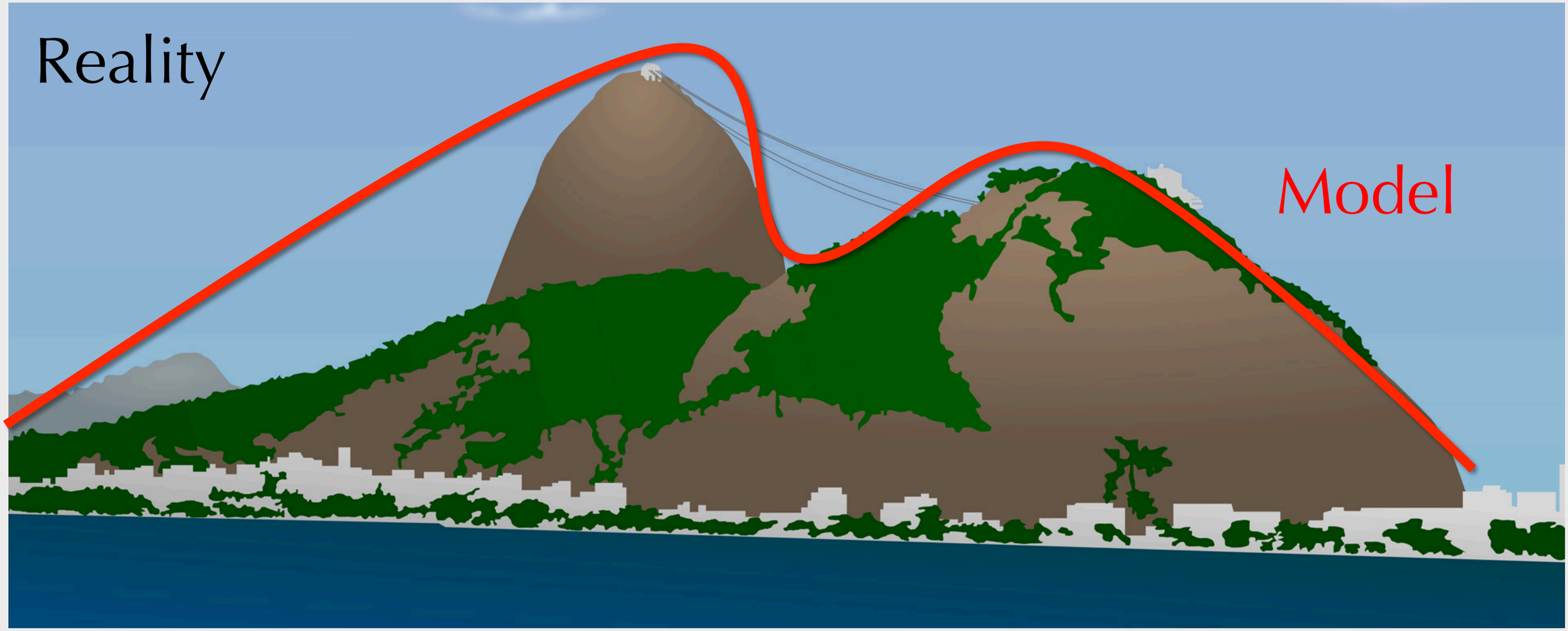


Reality ≠ Model

Q: How much information about reality does the model lose?

A1: Kullback-Leibler distance

A2: Akaike Information Criterion estimates KL distance



A2: Akaike Information Criterion estimates KL distance

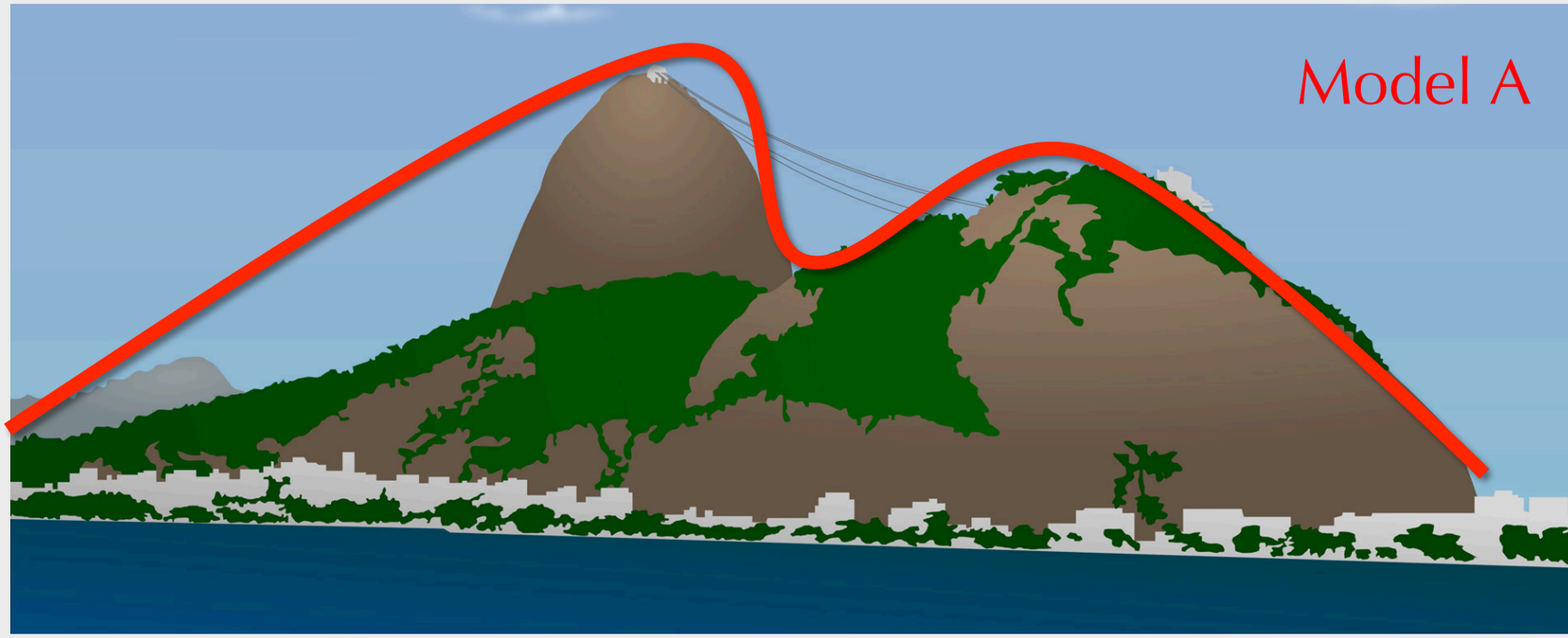
AIC: -2 log(likelihood) + 2 (# free parameters)

Two models

$AIC_A = 402$

$\Delta AIC_A = 5$

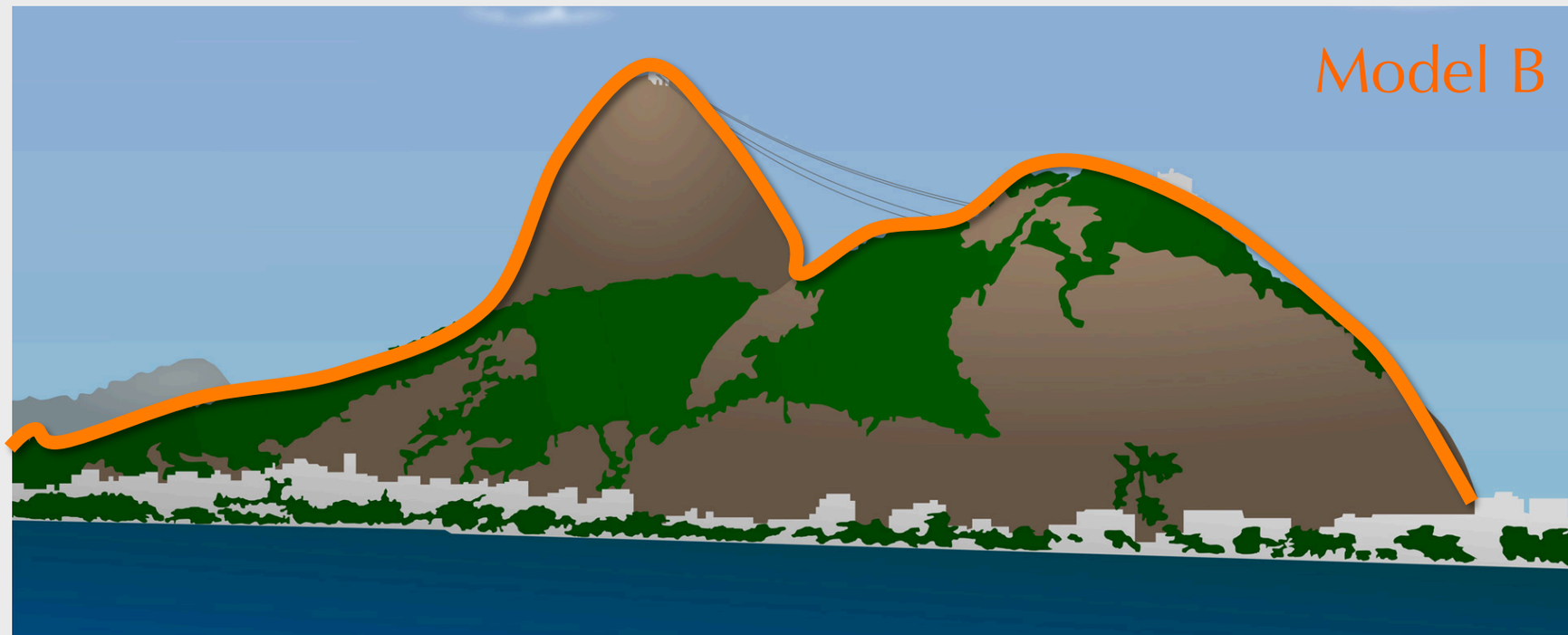
$weight_A = 0.08$



$AIC_B = 397$

$\Delta AIC_B = 0$

$weight_B = 0.92$



Akaike Information Criterion: Want the model that minimizes the information lost. Complex idea, simple to calculate ($2 \times$ #parameters - $2\ \ln(L)$). Generally, one subtracts the smallest AIC from a set of models from the values for all models to get ΔAIC (so best, lowest AIC, model has $\Delta AIC \equiv 0$).

Advantages: doesn't require nested models. Based on information theory. Allows model weighting.

AIC

ΔAIC	Level of empirical support for model
0 – 2	Substantial
4 – 7	Considerably less
10+	Essentially none

Burnham & Anderson 2004

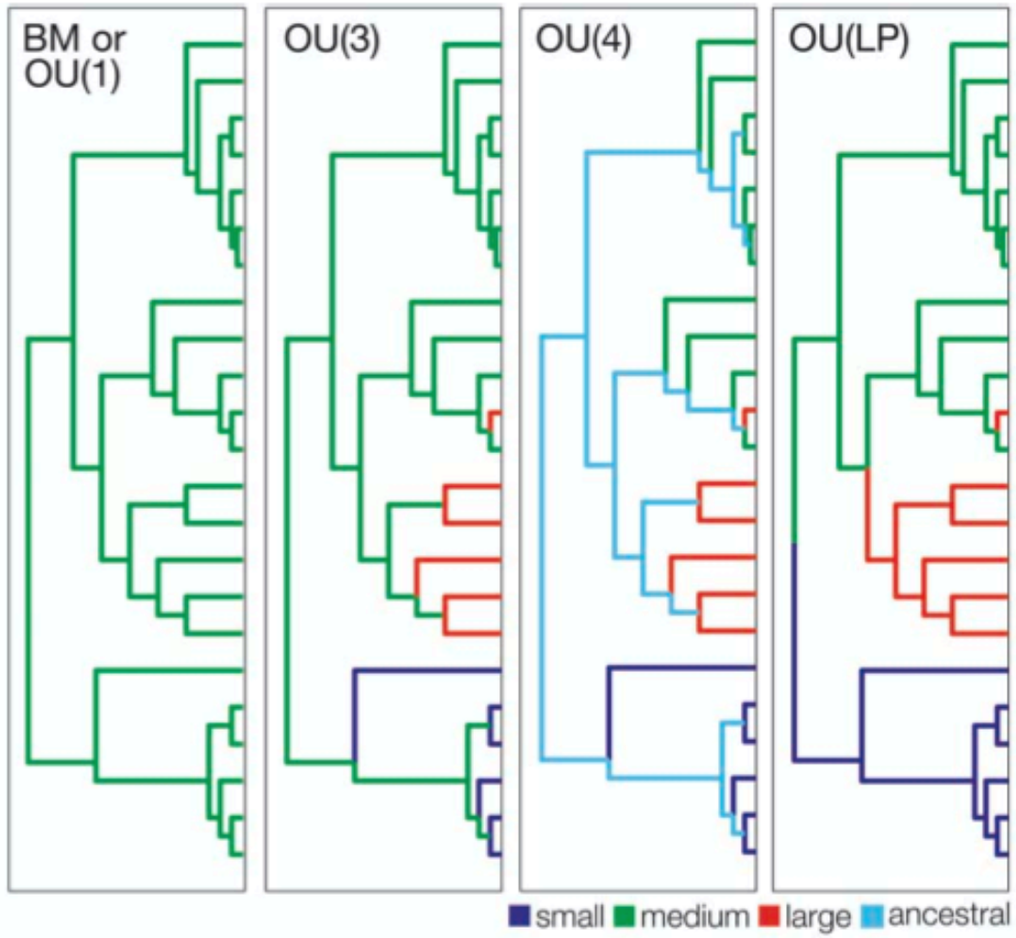


Table 2: Parameters estimated for the five models comparing character displacement with alternative hypotheses

	BM	OU(1)	OU(3)	OU(4)	OU(LP)
α		0	.32	14.67	2.49
σ	.21	.21	.20	.47	.22
θ_0	2.95	2.95	3.99	…[a]	.86
θ_{small}		…[a]	−1.40	2.58	2.75
θ_{medium}			.18	3.11	3.24
θ_{large}			2.71	3.30	3.56
$\theta_{ancestral}$				2.83	
ΔAICc	7.03	11.03	9.48	4.47	0.00
relative likelihood	0.030	0.004	0.000	0.107	1
weight	0.03	0.00	0.08	0.09	0.87

Model averaged sigma:

0.21 x 0.03 + 0.21 x 0 + 0.20 x 0.08 + 0.47 x 0.09 + 0.22 x 0.87 = 0.26

Butler & King 2004

Things folk tend to misunderstand with AIC

- It is not a significance measure.
  - No, really. Not even $\Delta 2$. Just stop.
- The best model isn't always the true model.
- The true model doesn't have to be in the set of models.

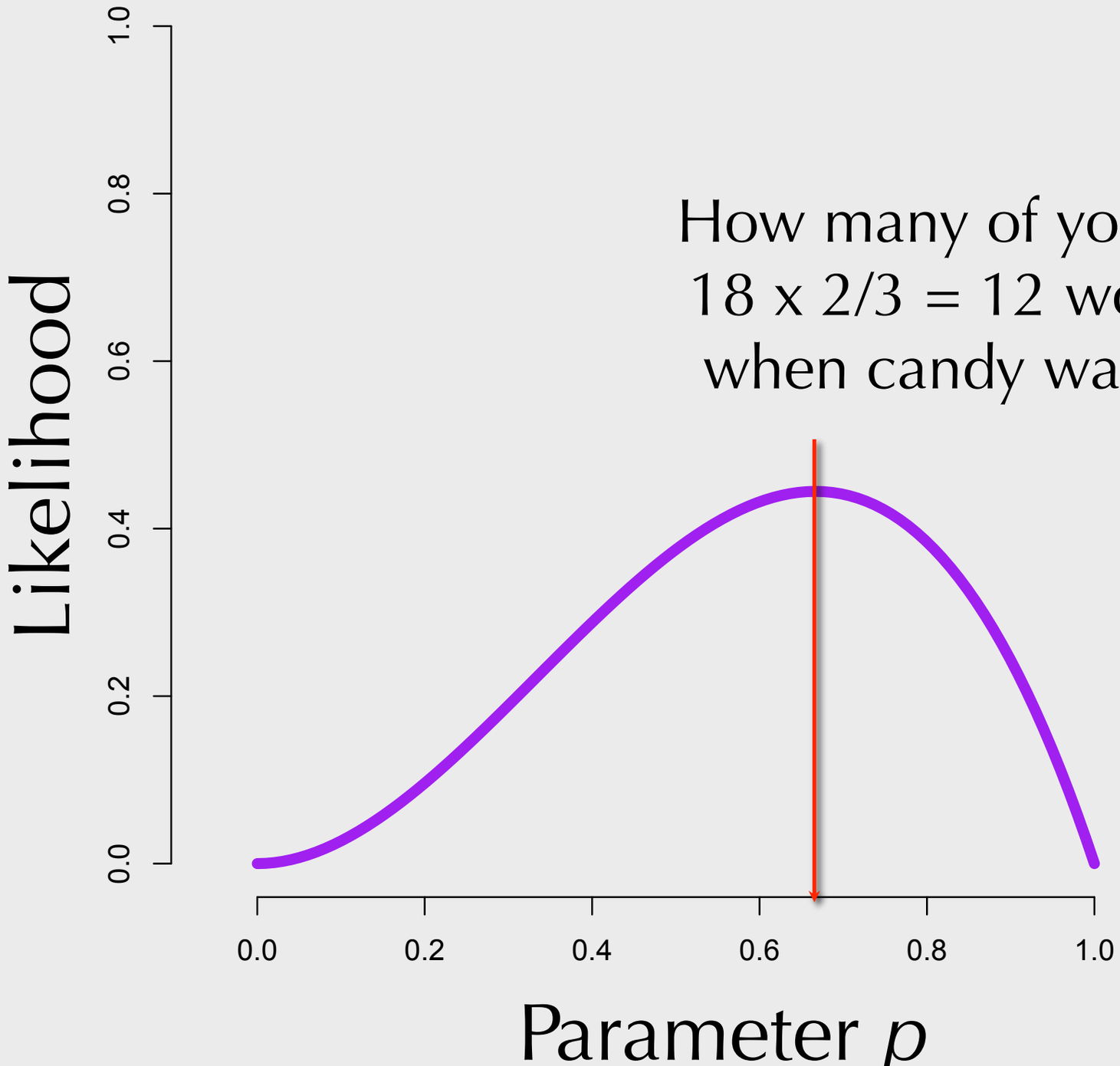
k = # free parameters
n = number of data points

AIC: -2 log(likelihood) + 2 k
AICc: -2 log(likelihood) + 2 k * (n / (n – k – 1))
BIC: -2 log(likelihood) + log(n) * k

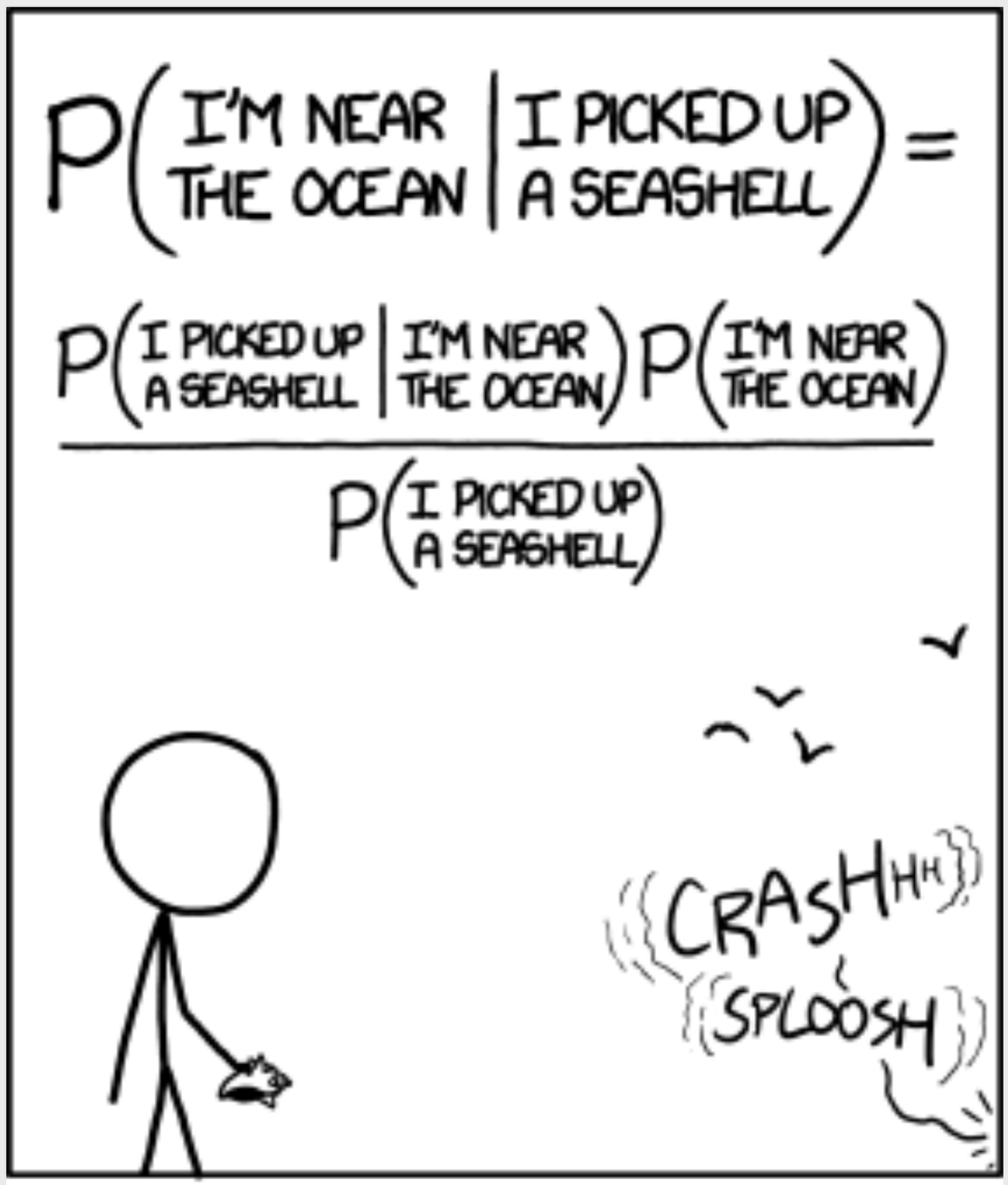
AICc has a correction for small sample size.
BIC is Bayesian Information Criterion (but easiest to think of as another information criterion). Regular AIC tends to overfit.

What is n?

How many of you guessed that 18 x 2/3 = 12 would be heads when candy was on the line?

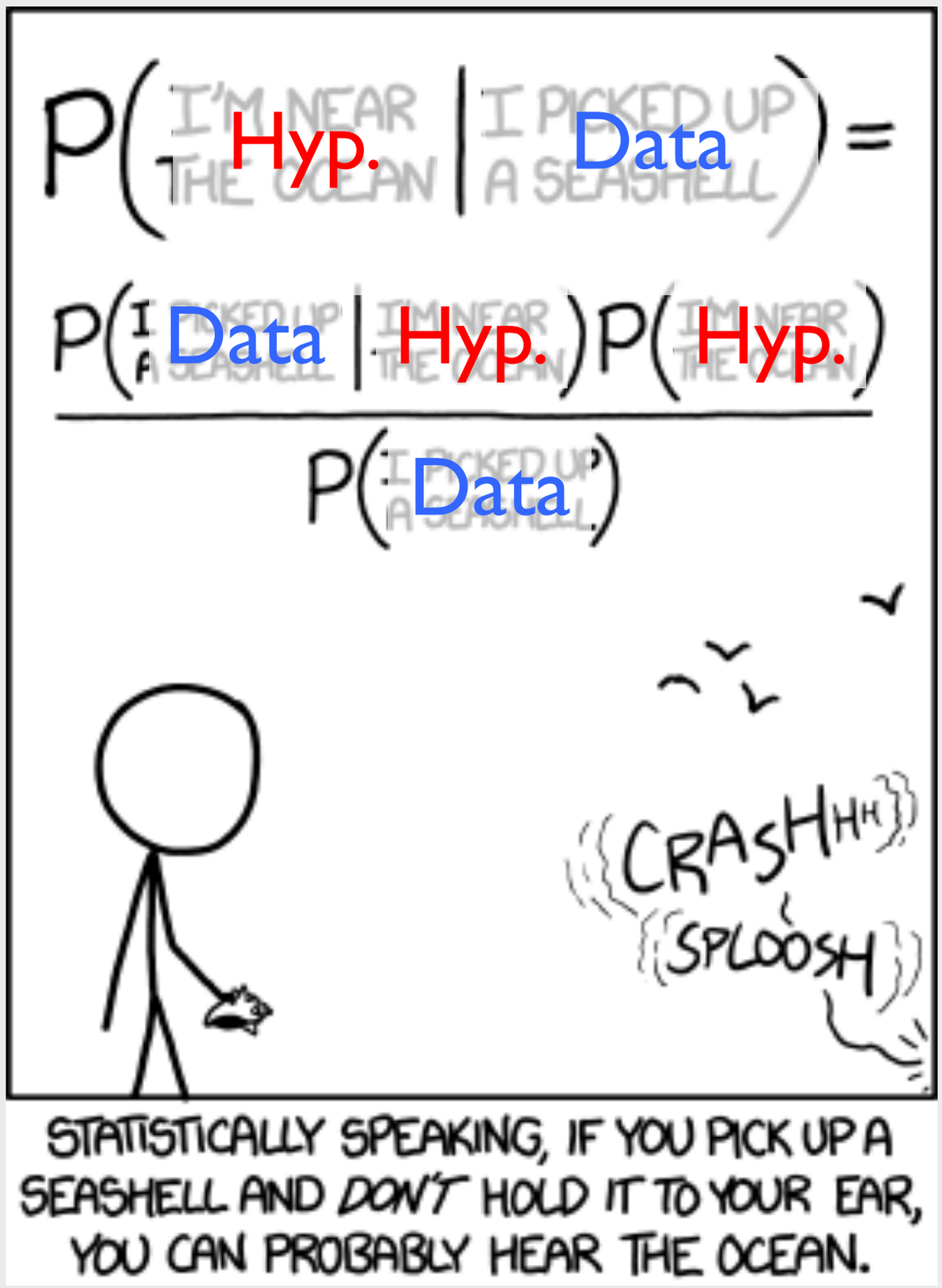


Conditional probability



http://xkcd.com/1236/

Bayesian statistics



http://xkcd.com/1236/

Bayesian statistics

$$P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}$$

$P(H \mid D)$ = Posterior

$P(D \mid H)$ = Likelihood

$P(H)$ = Prior

$P(D)$ = Prob of the data, over any hypothesis

Bayesian statistics

$$P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}$$

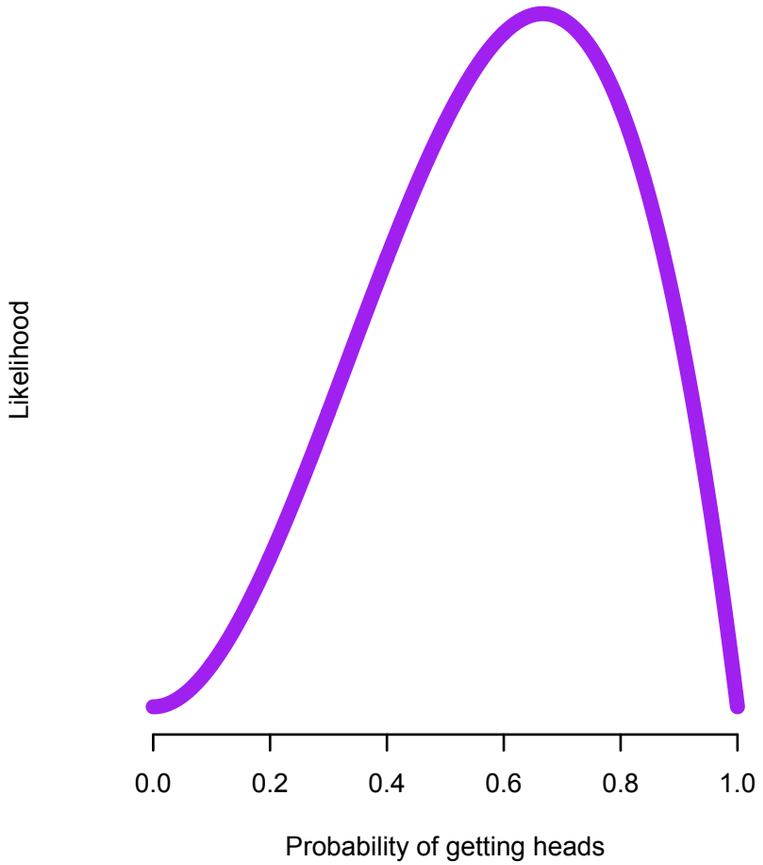
P(H | D) = Posterior

P(D | H) = Likelihood

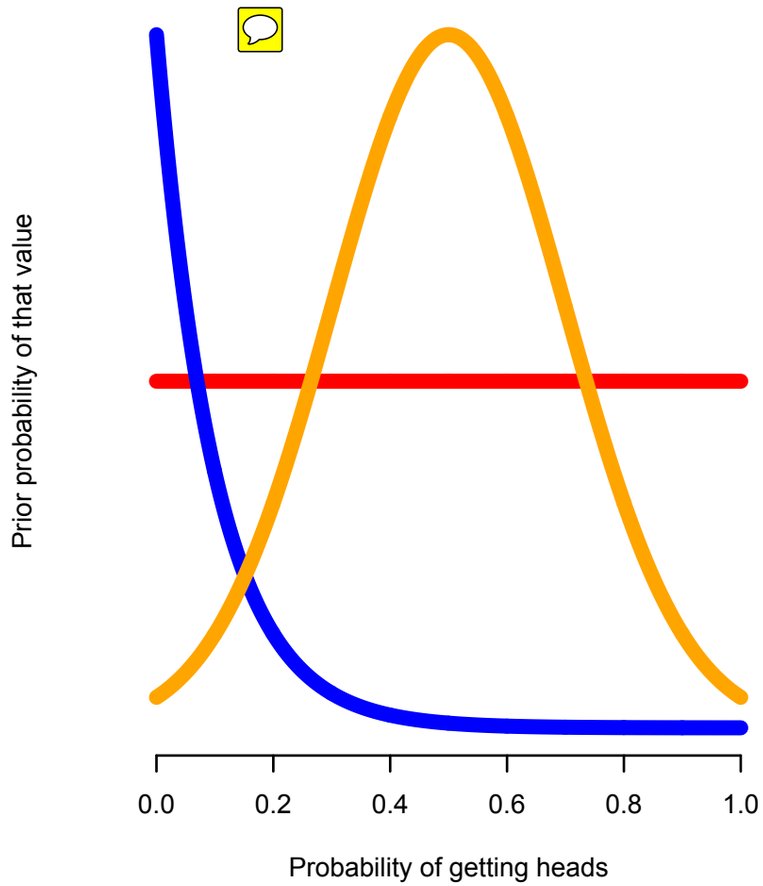
P(H) = Prior

P(D) = Prob of the data, over any hypothesis

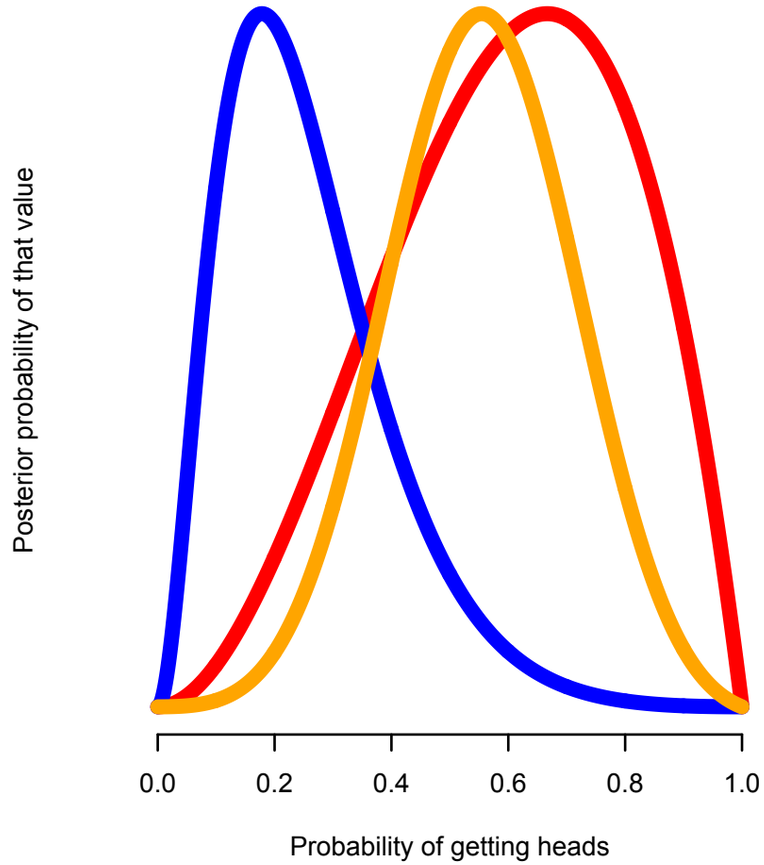
Likelihood



Priors



Posteriors



Bayesian statistics

$$P(H \mid D) = \frac{P(D \mid H) \times P(H)}{P(D)}$$

P(H | D) = Posterior

P(D | H) = Likelihood

P(H) = Prior

P(D) = Prob of the data, over any hypothesis

Markov Chain Monte Carlo

Markov Chain: series of steps, each step ONLY depends on current state, not states further in the past

Monte Carlo: repeated sampling from distribution. Think Las Vegas


http://salmagundiboston.blogspot.com/

Bayes Factors

$$K = \frac{\Pr(D|M_1)}{\Pr(D|M_2)} = \frac{\int \Pr(\theta_1|M_1)\Pr(D|\theta_1, M_1)\, d\theta_1}{\int \Pr(\theta_2|M_2)\Pr(D|\theta_2, M_2)\, d\theta_2}$$

scale of evidence for Bayes factors	
Bayes factor	Interpretation
B.F. < 1/10	Strong evidence for Model 2
1/10 < B.F. < 1/3	Moderate evidence for Model 2
1/3 < B.F. < 1	Weak evidence for Model 2
1 < B.F. < 3	Weak evidence for Model 1
3 < B.F < 10	Moderate evidence for Model 1
B.F. > 10	Strong evidence for Model 1

Reversible jump MCMC