Bootstrapping

Suppose you sequence the 18S rRNA gene and estimate the tree.

What tree would you have estimated had you chosen a different gene to sequence?

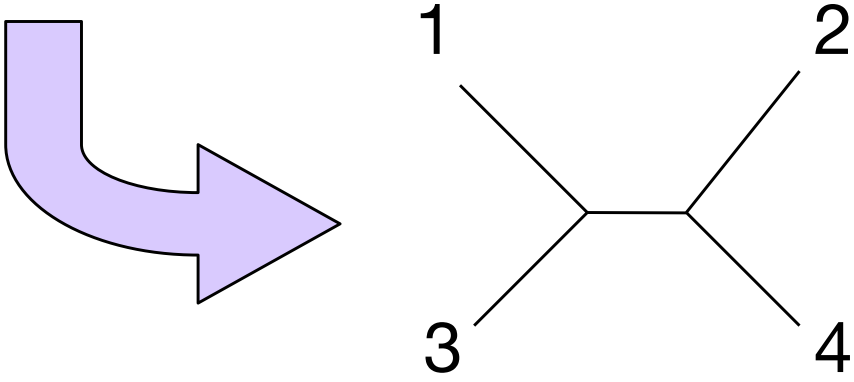
Which parts of the tree (i.e. splits) would you expect to be present in trees estimated from genes that evolved in a way similar to the one you sampled?

Felsenstein (1985)

Bootstrapping: first step

	1	2	3	4	5	6	7	...	*N*
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

From the original data, estimate a tree using, say, maximum likelihood (could use parsimony or distance methods, however)

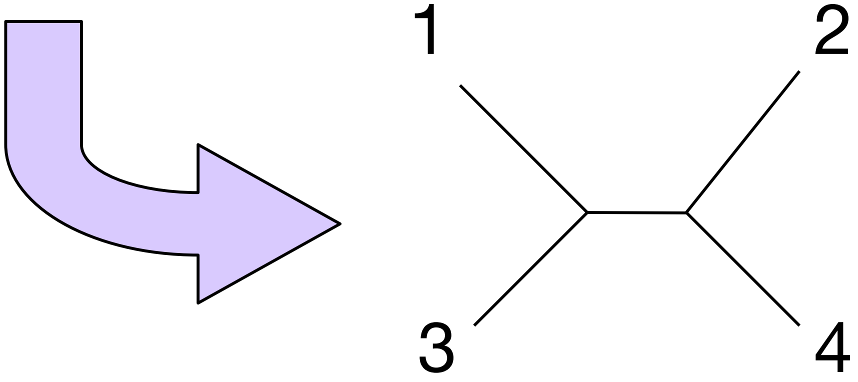


Bootstrapping: first replicate

	1	2	3	4	5	6	7	...	*N*
weights	1	2	0	0	1	3	1	...	2
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

Sum of weights equals *N* (each bootstrap dataset has same number of sites as the original)

From the bootstrap dataset, estimate the tree using the same method you used for the original dataset

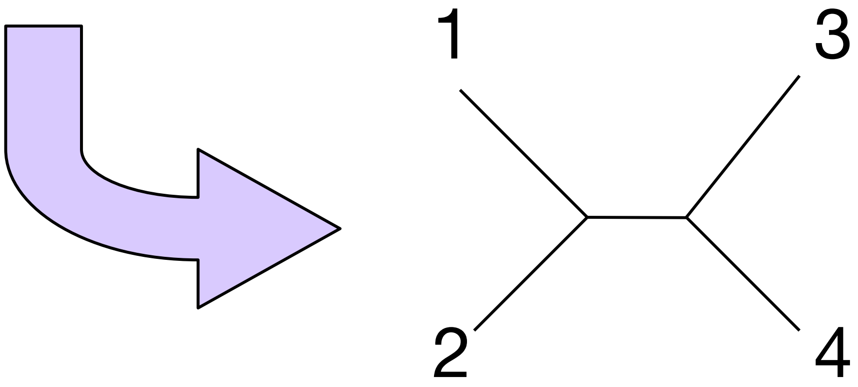


Bootstrapping: second replicate

	1	2	3	4	5	6	7	...	*N*
weights	0	1	1	1	1	3	0	...	0
1	T	A	G	T	C	G	T	...	A
2	T	C	A	T	C	G	T	...	G
3	A	T	G	T	C	A	C	...	G
4	A	T	A	T	C	G	C	...	G

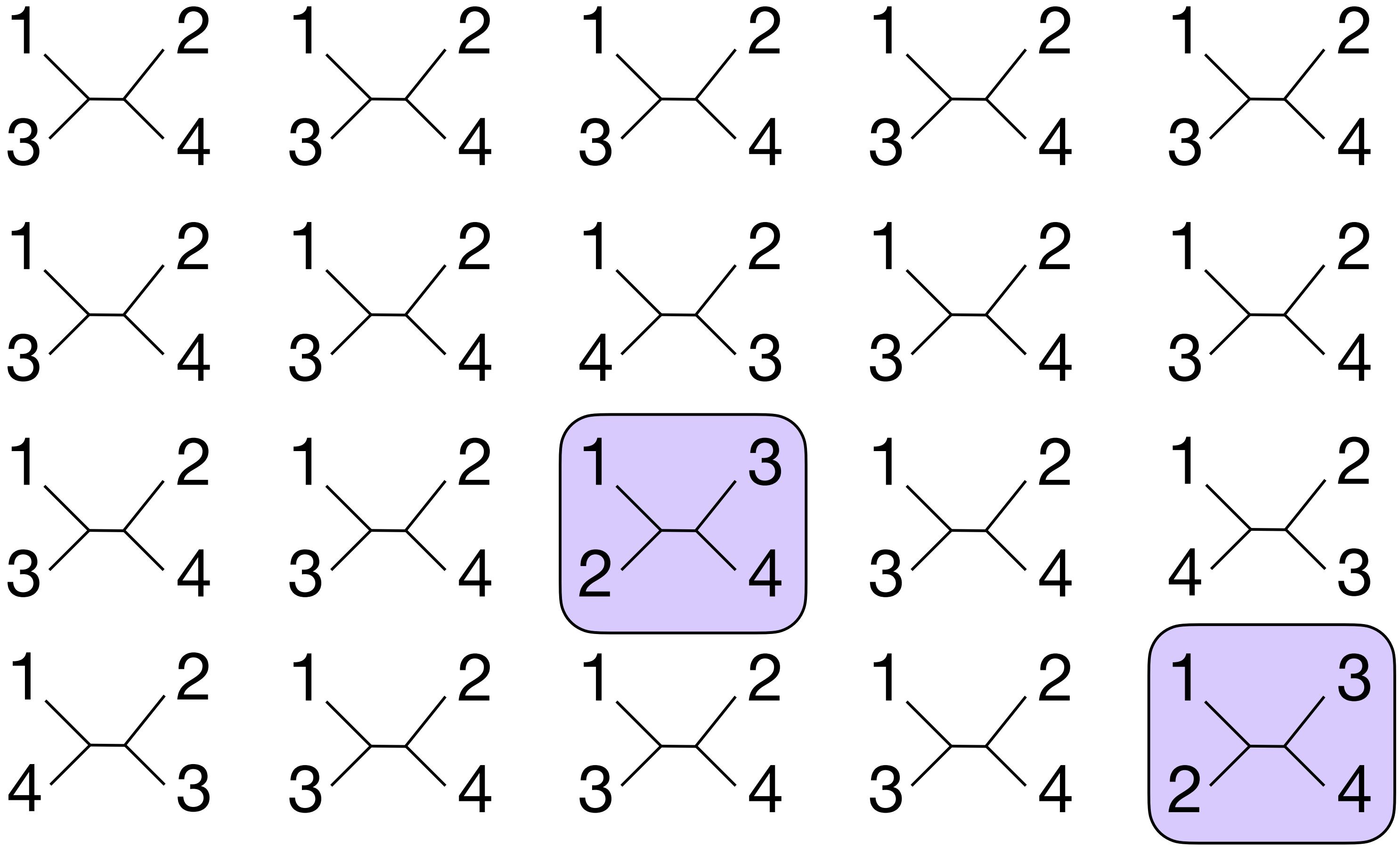
Note that weights are different this time, reflecting the random sampling with replacement used to generate the weights

This time the tree that is estimated is different than the one estimated using the original dataset.



Bootstrapping: 20 replicates

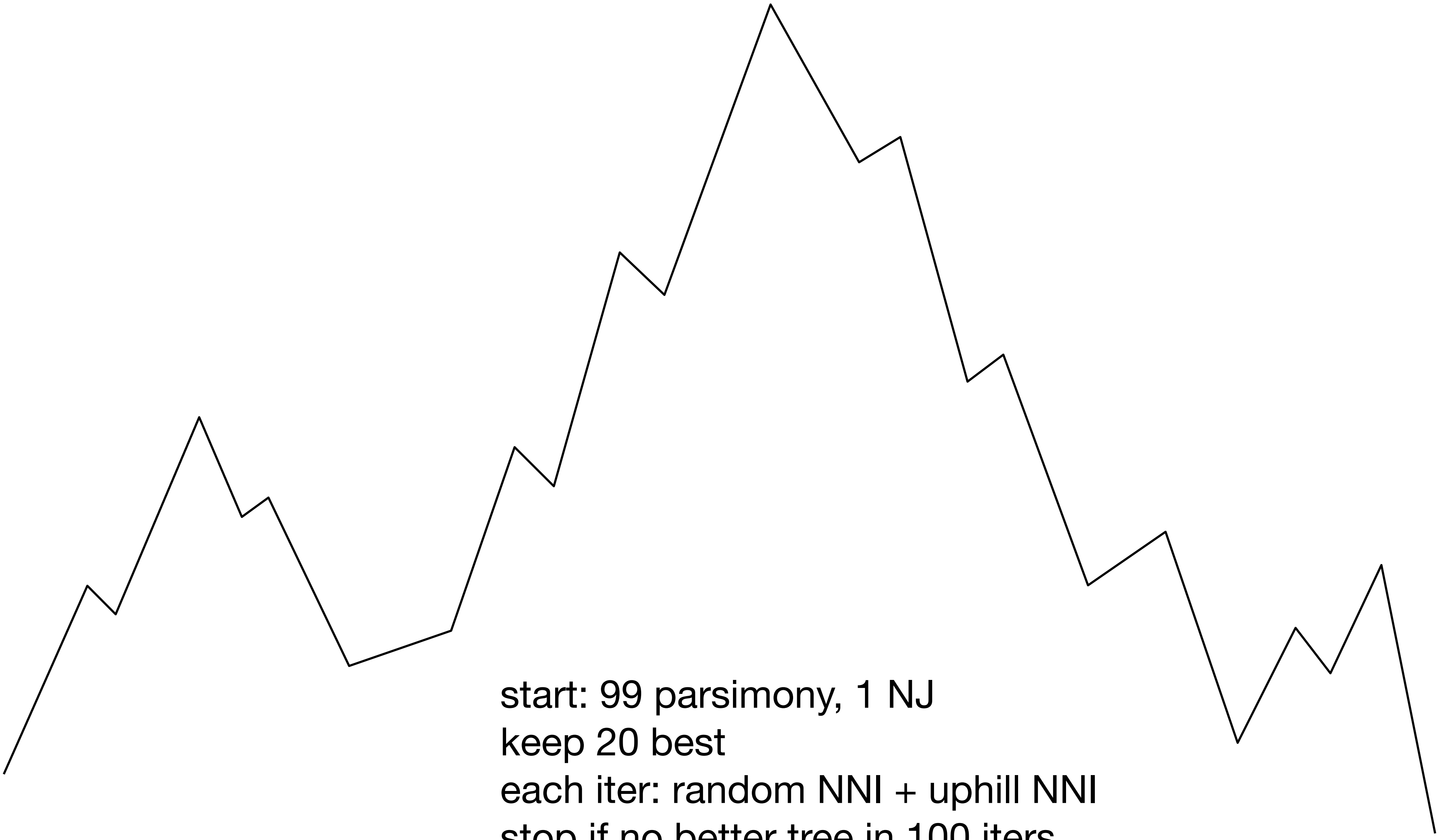
```
  Freq
----------
-*-*  75.0
-**-  15.0
--**  10.0
```



Note: usually at least 100 replicates are performed, and 500 is better

e.g. 2/20, or 10%, have 3 and 4 together

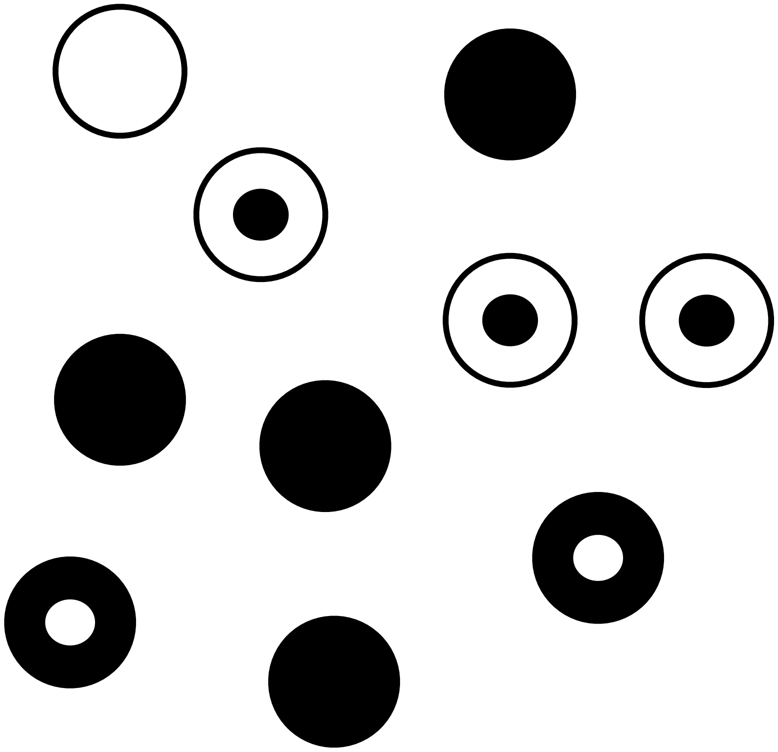
IQ-TREE searching and ultrafast "bootstraps"



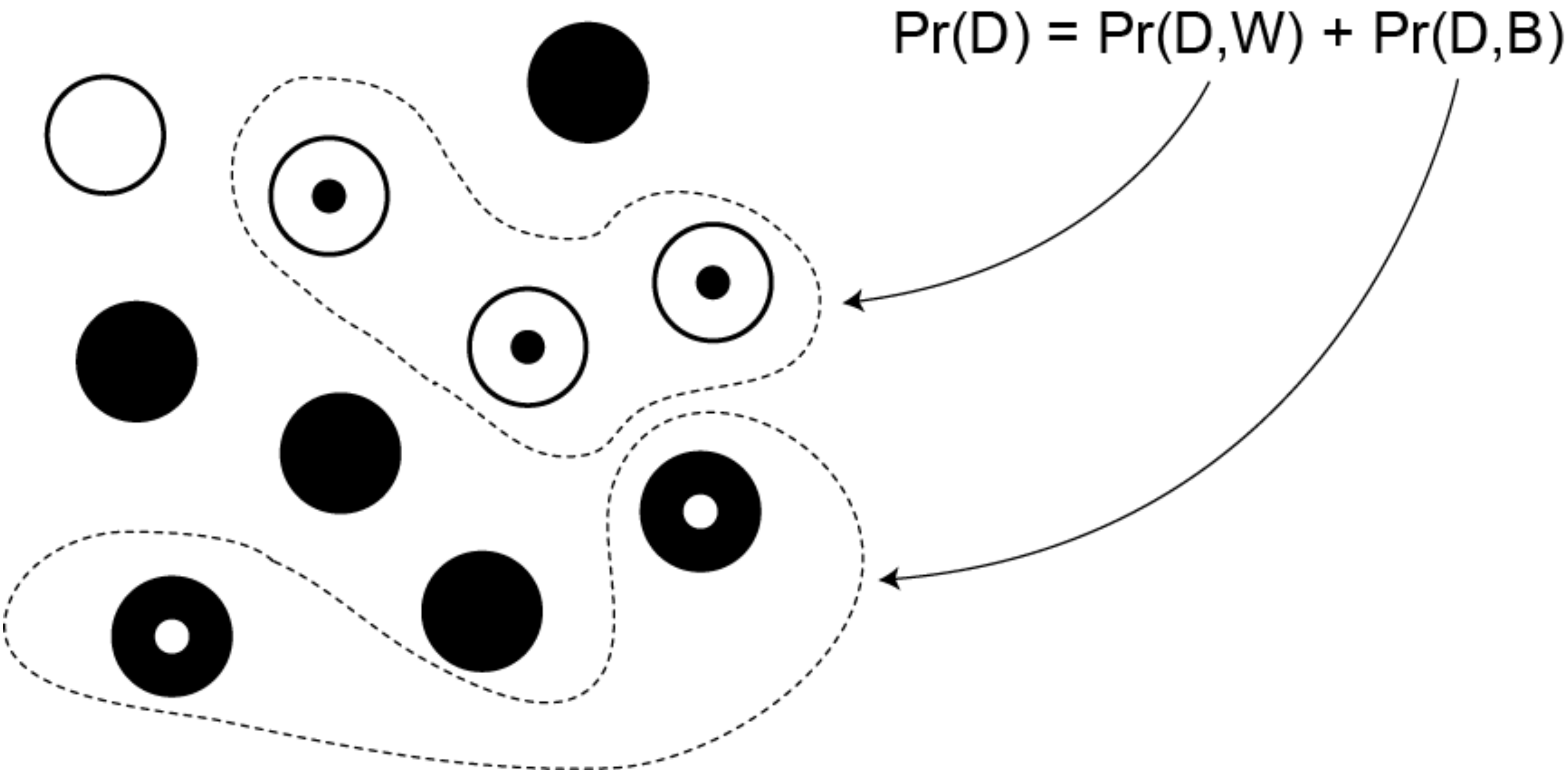
start: 99 parsimony, 1 NJ
keep 20 best
each iter: random NNI + uphill NNI
stop if no better tree in 100 iters

Bayes' rule

Pr(B,D)



Probability of "Dotted"



$$Pr(D) = Pr(D,W) + Pr(D,B)$$

Bayes' rule (cont.)

$$\Pr(B|D) = \frac{\Pr(B)\Pr(D|B)}{\Pr(D)}$$

$$= \frac{\Pr(D,B)}{\Pr(D,B) + \Pr(D,W)}$$

Pr(*D*) is the **marginal probability** of being dotted
To compute it, we **marginalize over colors**

Marginal (total) probabilities

	B	W
D	Pr(D,B)	Pr(D,W)
S	Pr(S,B)	Pr(S,W)

Bayes' rule (cont.)

$$\Pr(B|D) = \frac{\Pr(B)\Pr(D|B)}{\Pr(D,B) + \Pr(D,W)}$$

$$= \frac{\Pr(B)\Pr(D|B)}{\Pr(B)\Pr(D|B) + \Pr(W)\Pr(D|W)}$$

$$= \frac{\Pr(B)\Pr(D|B)}{\sum_{\theta \in \{B,W\}} \Pr(\theta)\Pr(D|\theta)}$$

Bayes’ rule in statistics

Likelihood of hypothesis *θ*

Prior probability of hypothesis *θ*

$$\Pr(\theta|D) = \frac{\Pr(D|\theta)\Pr(\theta)}{\sum_\theta \Pr(D|\theta)\Pr(\theta)}$$

Posterior probability of hypothesis *θ*

Marginal probability of the data (marginalizing over hypotheses)

Simplest paternity example

child's genotype: **Aa**

mother's genotype: **aa**

possible fathers

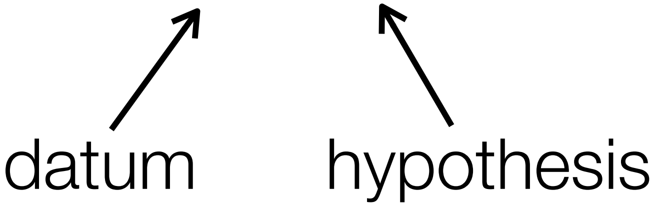
Possibilities	θ_1	θ_2	Row sum
Genotypes	AA	Aa	---
Prior			
Likelihood			
Likelihood × Prior			
Posterior			

The prior can be your friend

Suppose the test for a **rare** disease has the following true and false positive probabilities:

$$\Pr(+ \mid \text{disease}) = 1.00$$
$$\Pr(+ \mid \text{healthy}) = 0.01$$

(Note that we do not need to consider the case of a negative test result.)



Suppose further I **test positive** for the disease. How worried should I be?

It is very tempting to (mis)interpret the likelihood as a posterior probability and conclude “There is a 100% chance that I have the disease.”

The prior can be your friend

$$\Pr(\text{disease}|+) = \frac{(1.0)(\frac{1}{1000000})}{(1.0)(\frac{1}{1000000}) + (0.01)(\frac{999999}{1000000})}$$

1 person out of a million has a true positive result

10,000 people out a million will have a false positive result

Thus, the odds *against* having the disease are actually 10000 to 1!

Bayes' rule: continuous case

Likelihood

Prior probability density

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{\int p(D|\theta')\, p(\theta') d\theta'}$$

Posterior probability density

Marginal probability of the data

(a.k.a. marginal likelihood)