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?

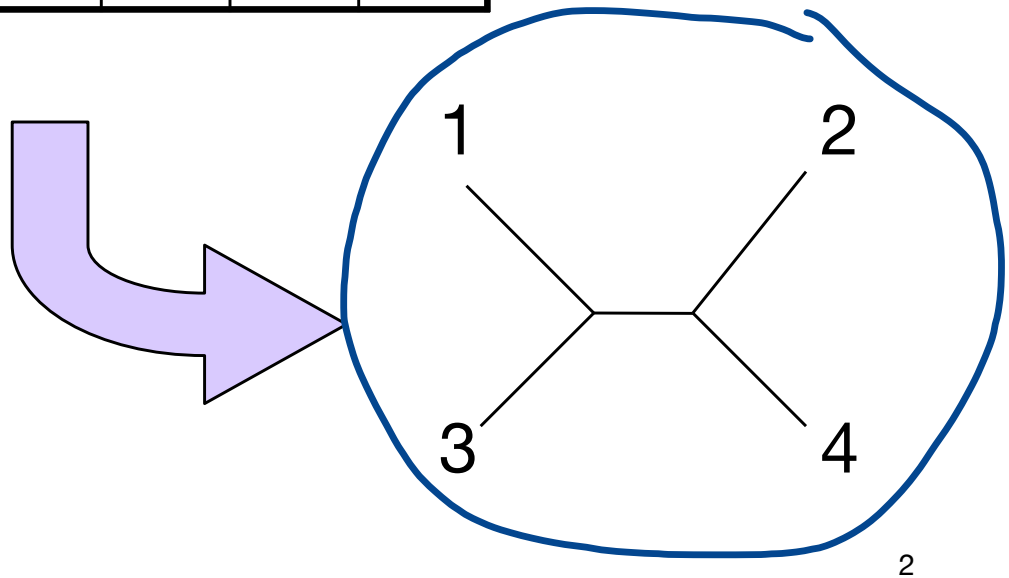
50 000

Felsenstein (1985)

Bootstrapping: first step

	1	2	3	4	5	6	7	...	*k*
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)

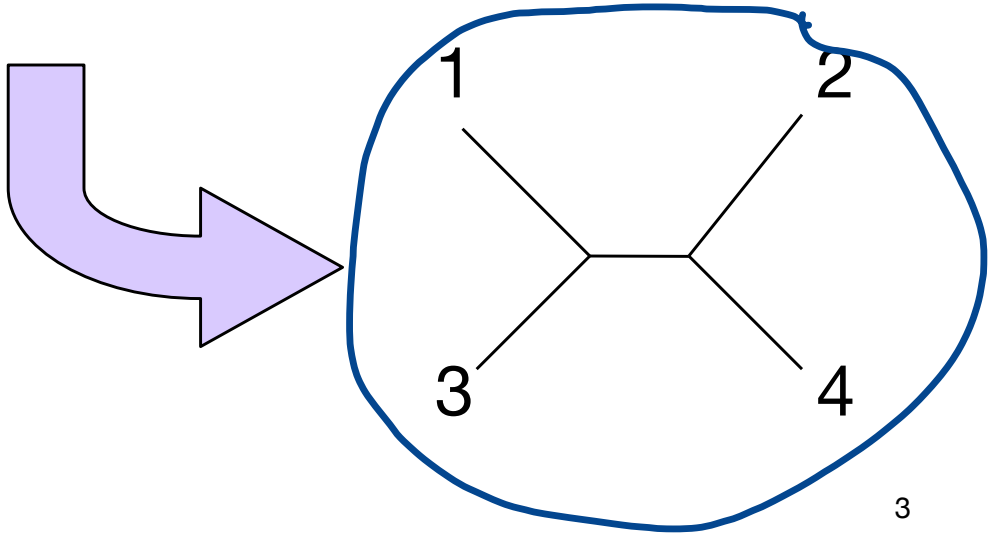


Bootstrapping: first replicate

	1	2	3	4	5	6	7	...	*k*
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 *k* (i.e., 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

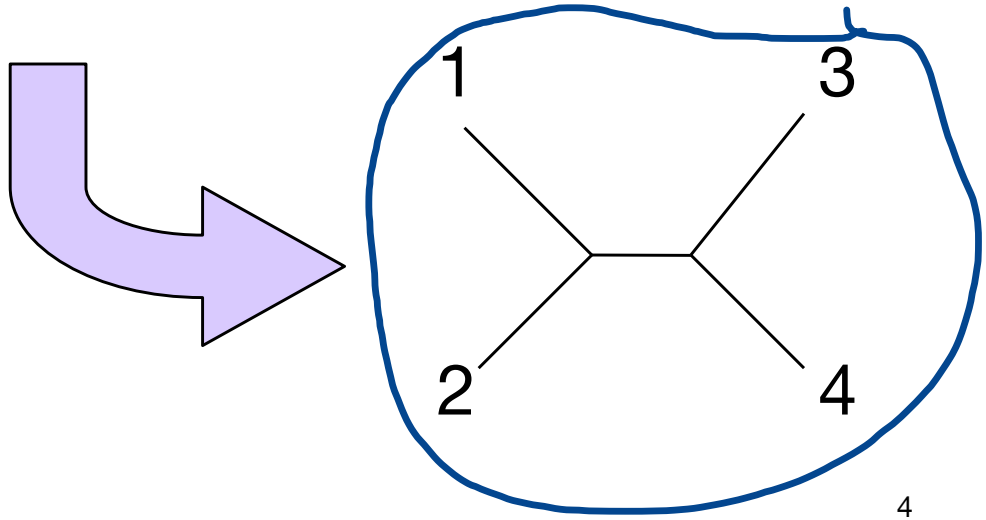


Bootstrapping: second replicate

	1	2	3	4	5	6	7	...	*k*
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

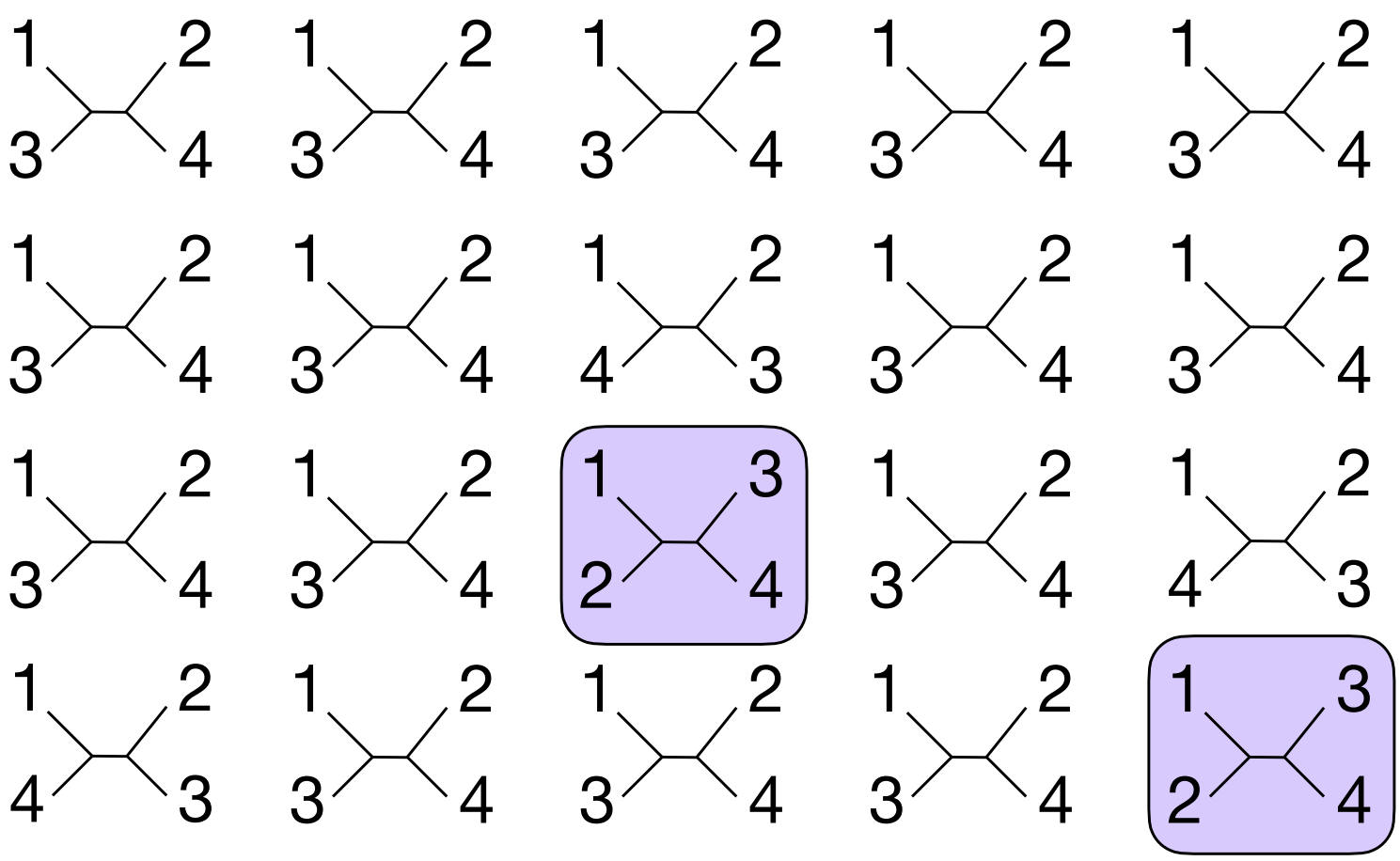
Note that weights are different this time, reflecting the random sampling with replacement used to generate the weights

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 "bootstrap"