

Boots



Bootstrap

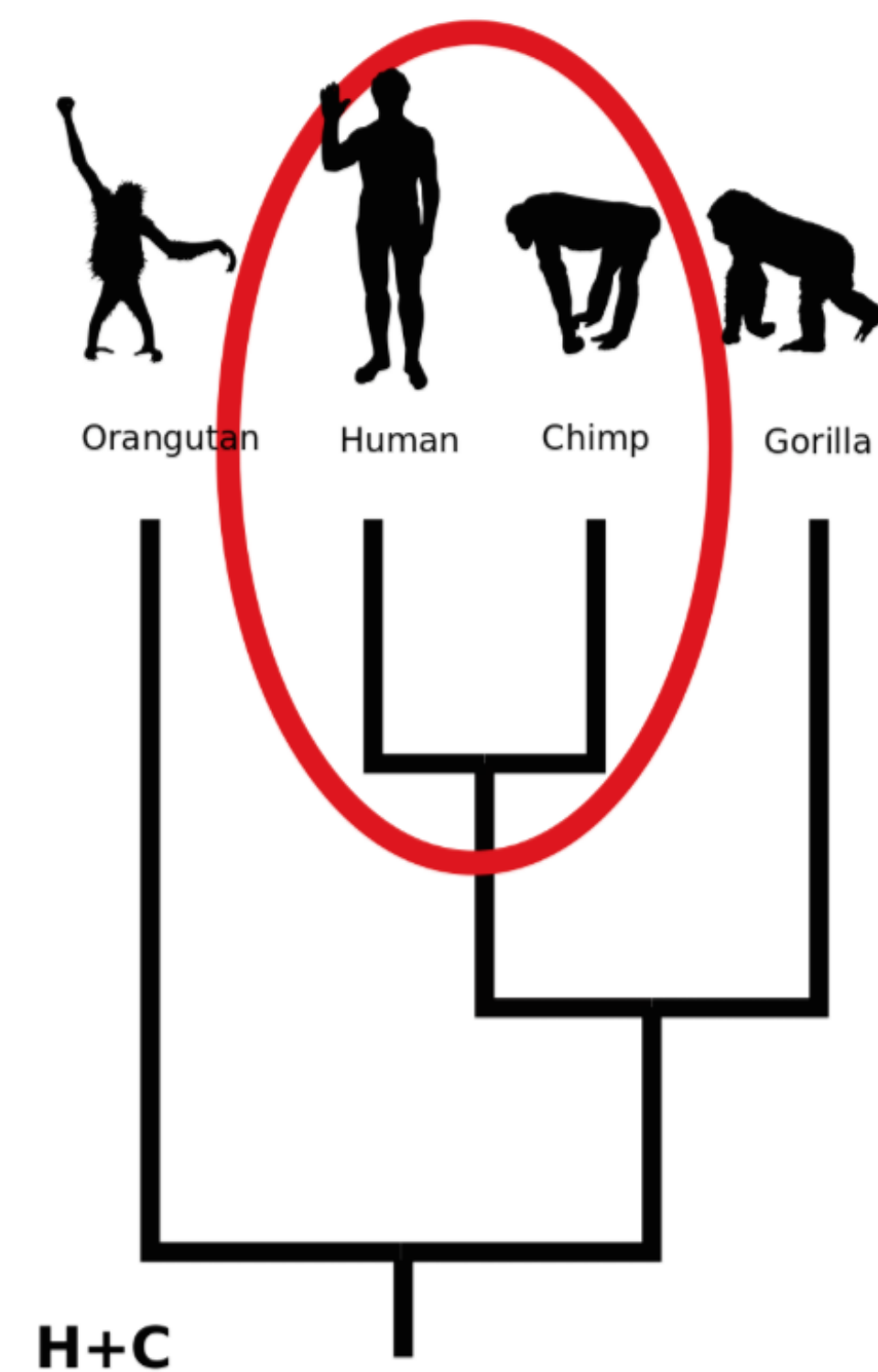
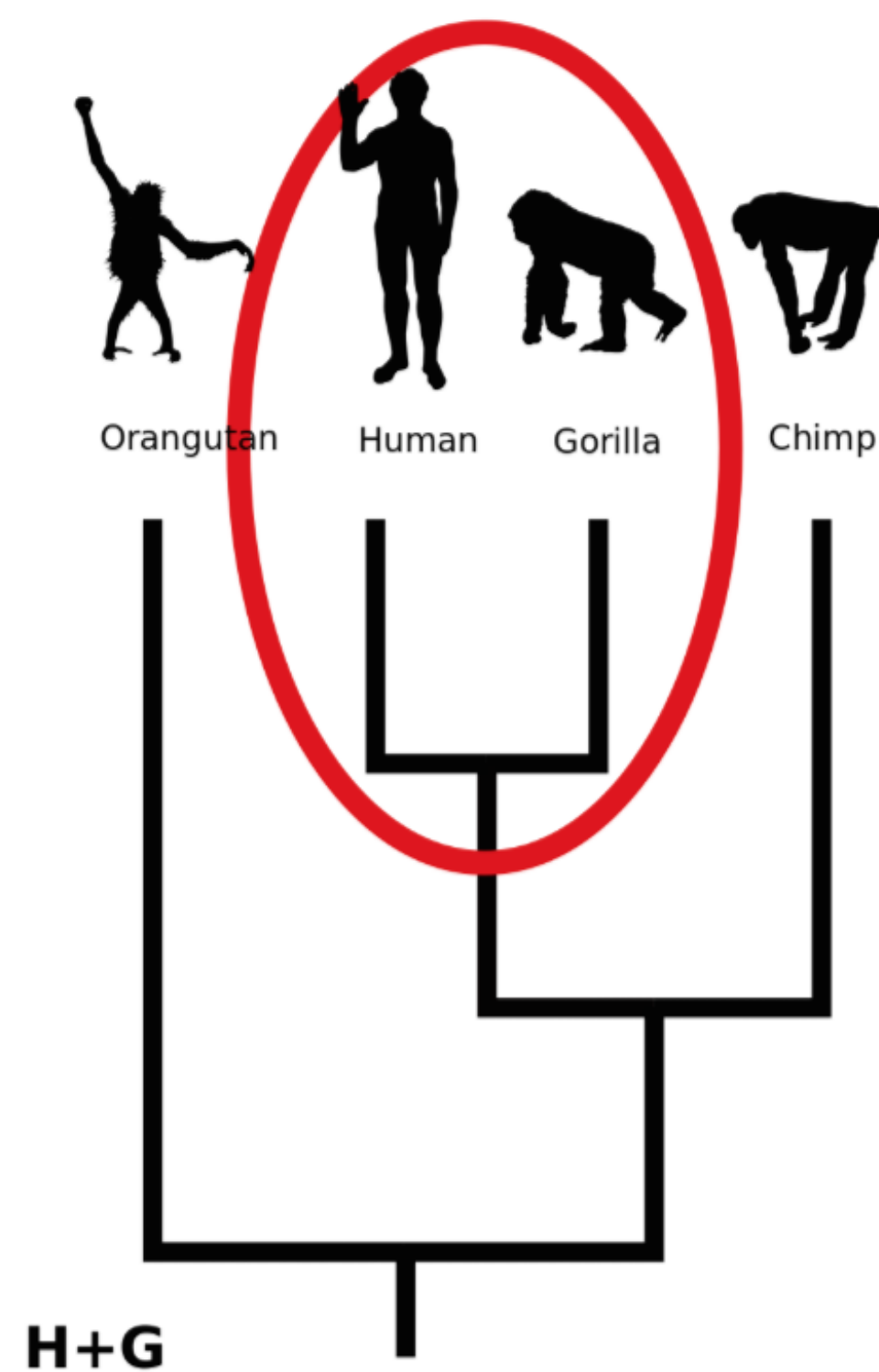
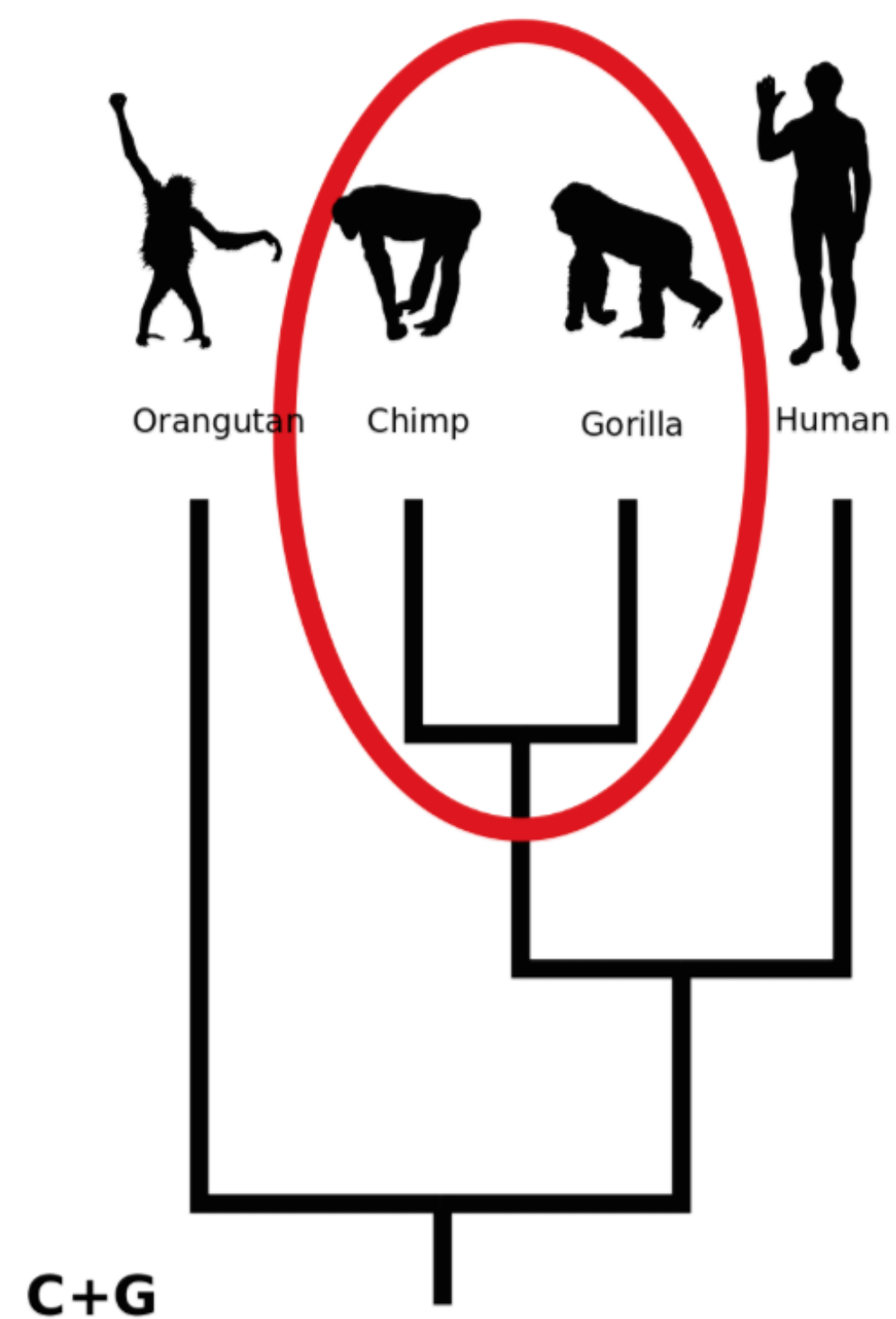
and phylogenetic trees



What does it mean

Tall **boots** may have a **tab, loop or handle** at the top known as a bootstrap, allowing one to use fingers or a boot hook tool to help pulling the boots on. The saying "**to pull oneself up by one's bootstraps**" was already in use during the 19th century as an example of an impossible task. The idiom dates at least to 1834, when it appeared in the Workingman's Advocate: "It is conjectured that Mr. Murphee will now be enabled to hand himself over the Cumberland river or a barn yard fence by the straps of his boots." In 1860 it appeared in a comment on metaphysical philosophy: "The attempt of the mind to analyze itself [is] an effort analogous to one who would lift himself by his own bootstraps." **Bootstrap as a metaphor, meaning to better oneself by one's own unaided efforts, was in use in 1922.**

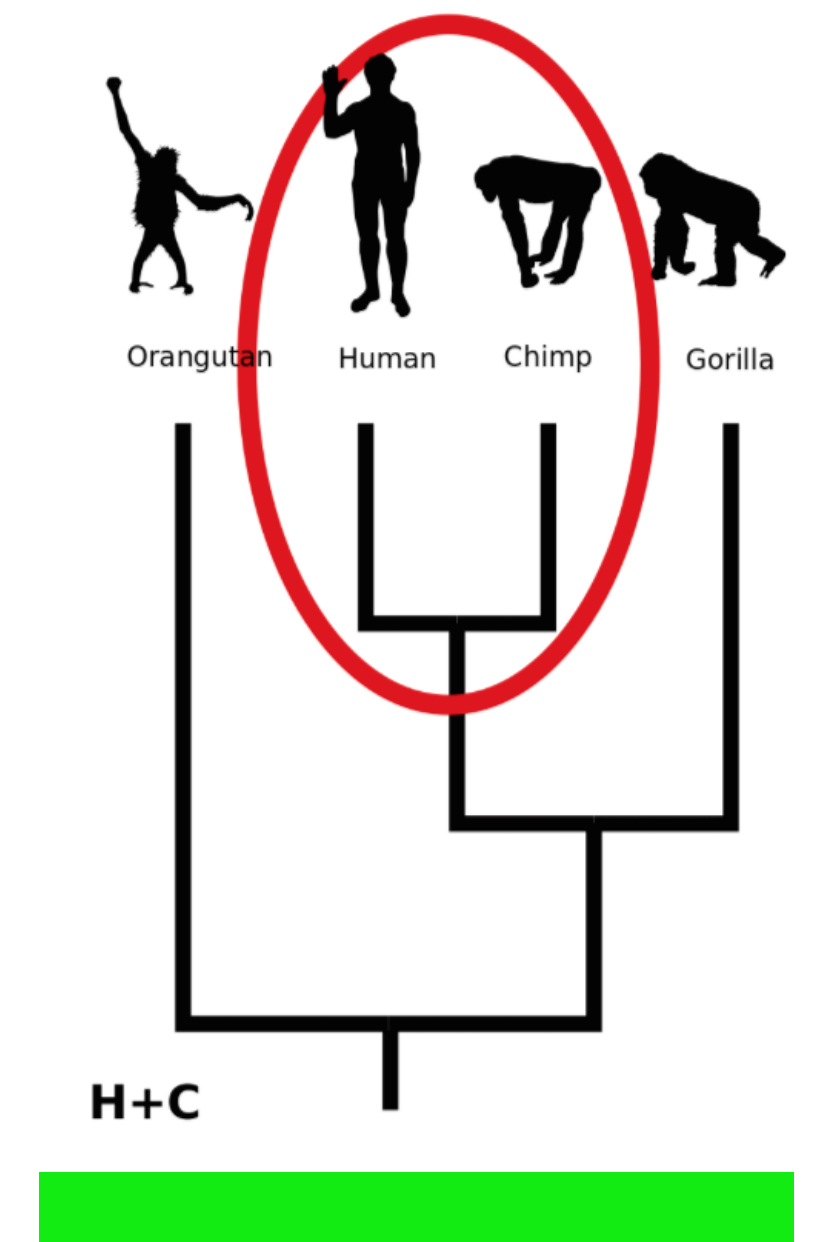
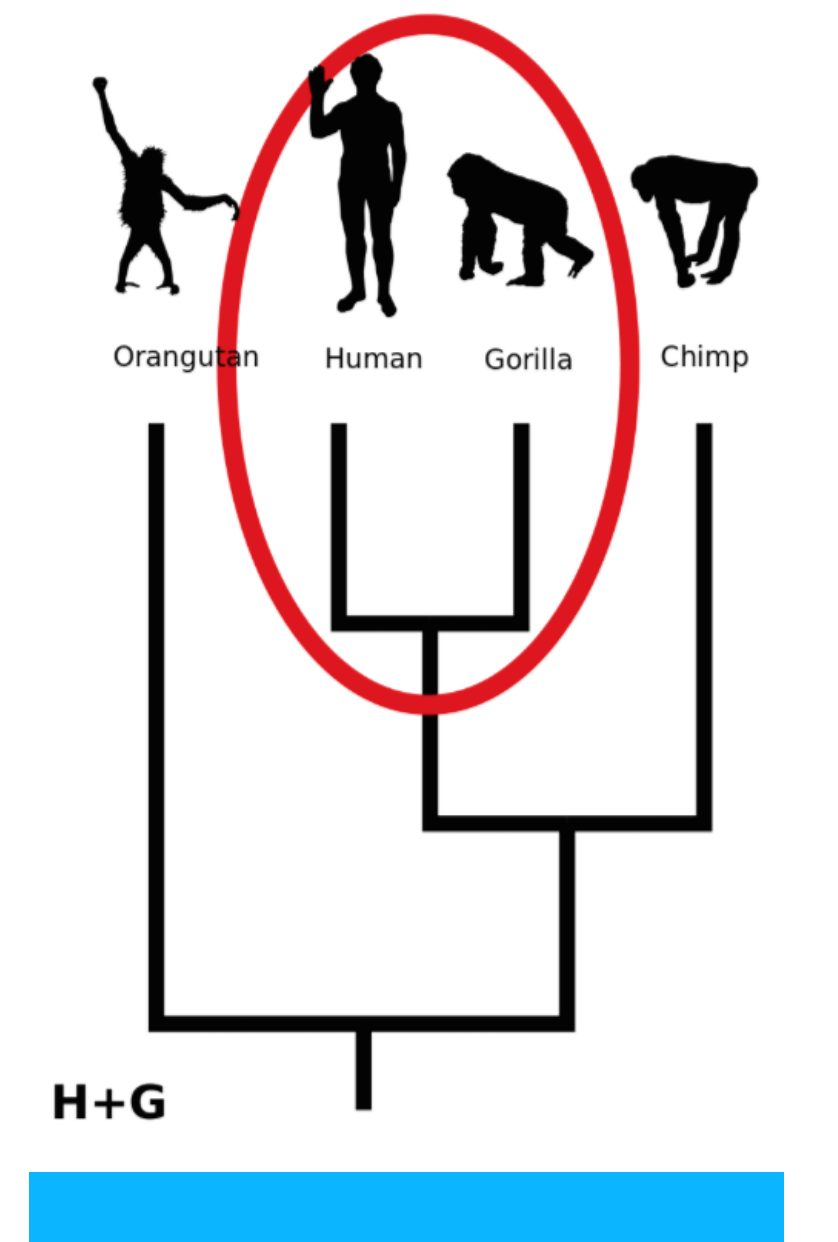
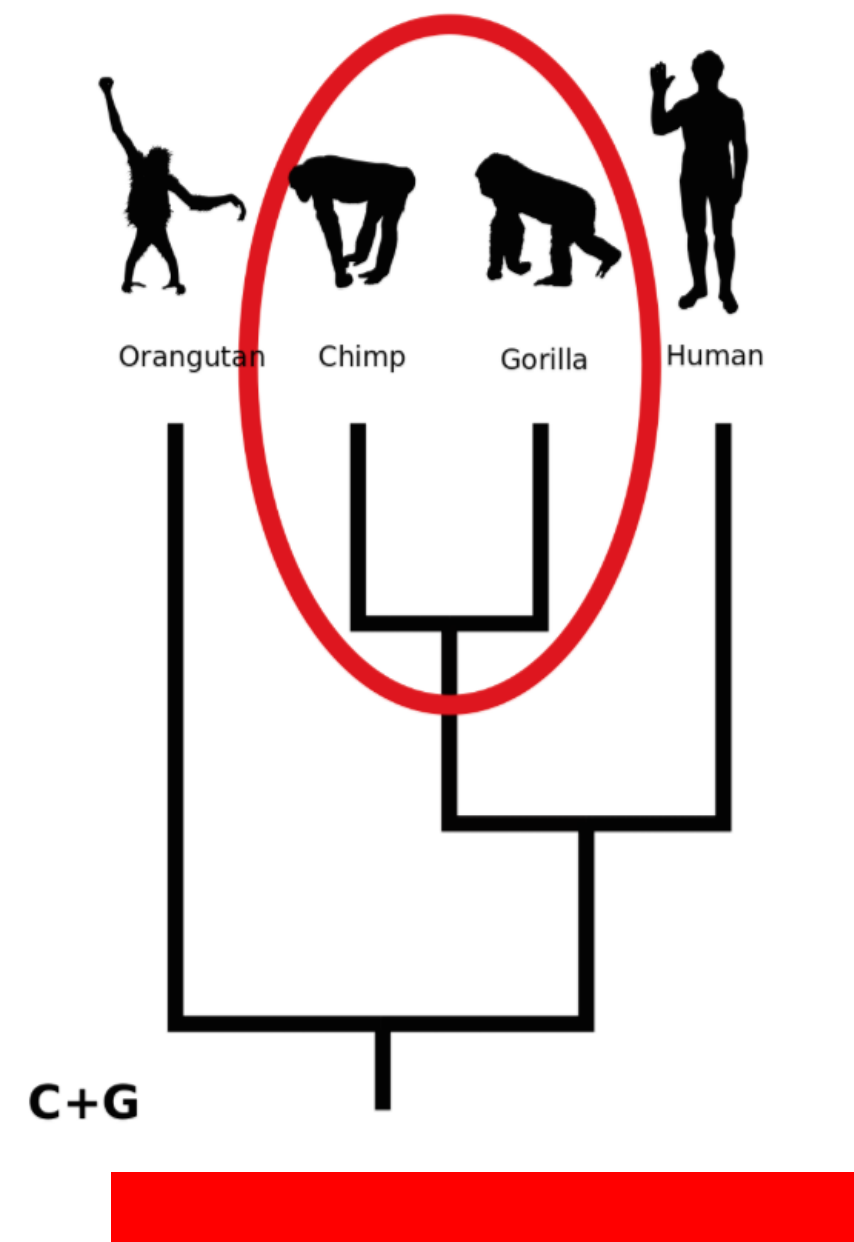
Bootstrapping phylogenies



from Mark Holder

Bootstrapping phylogenies

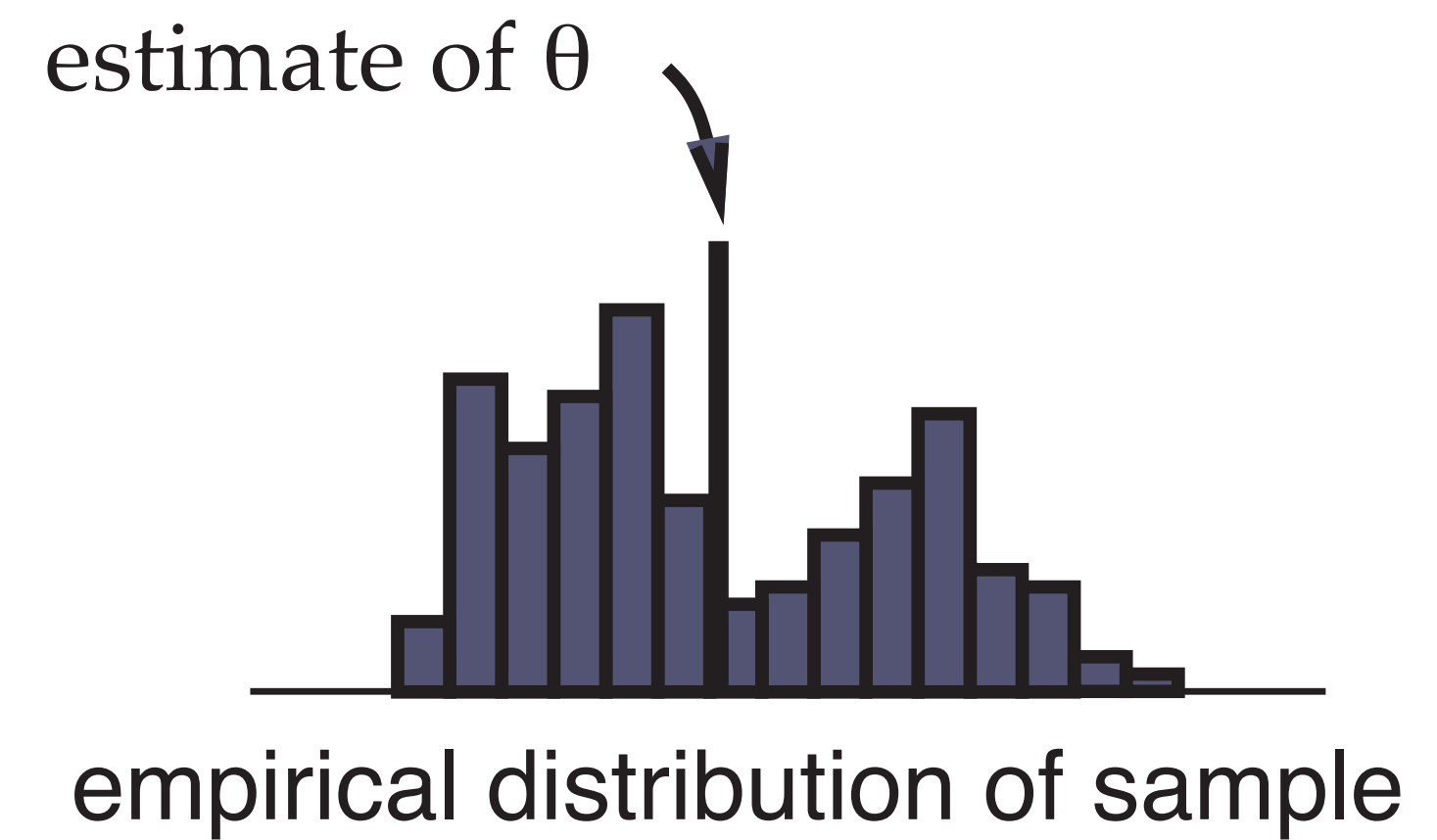
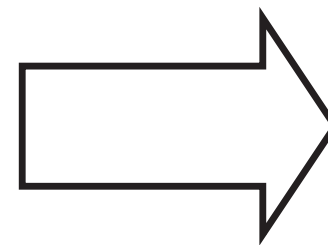
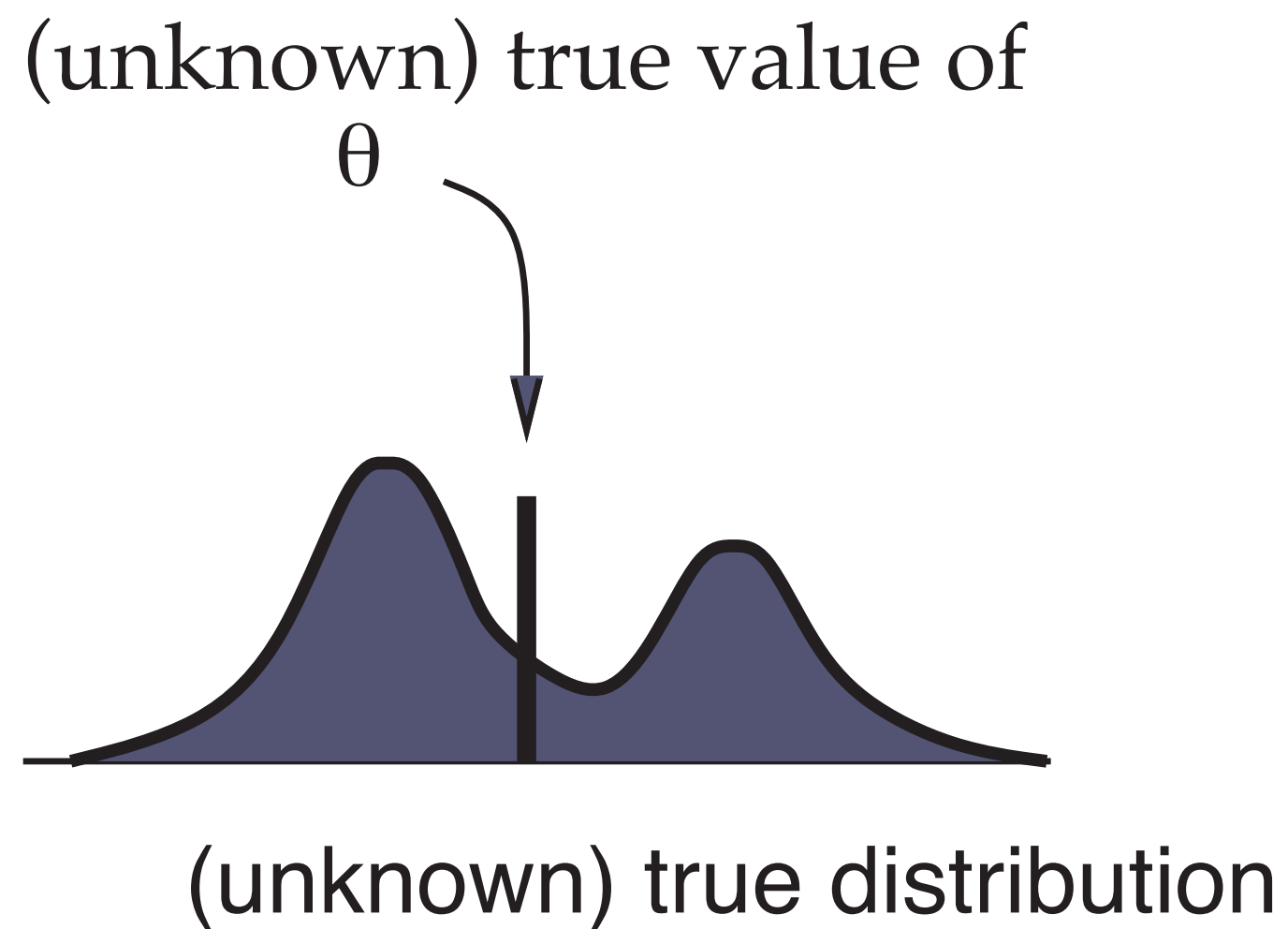
Human	G	C	C	A	G	C	T	G	C	T
Chimp	G	C	C	G	G	G	T	A	T	T
Gorilla	G	T	C	A	A	C	T	A	T	G
Orang	C	T	C	A	A	G	T	G	C	G



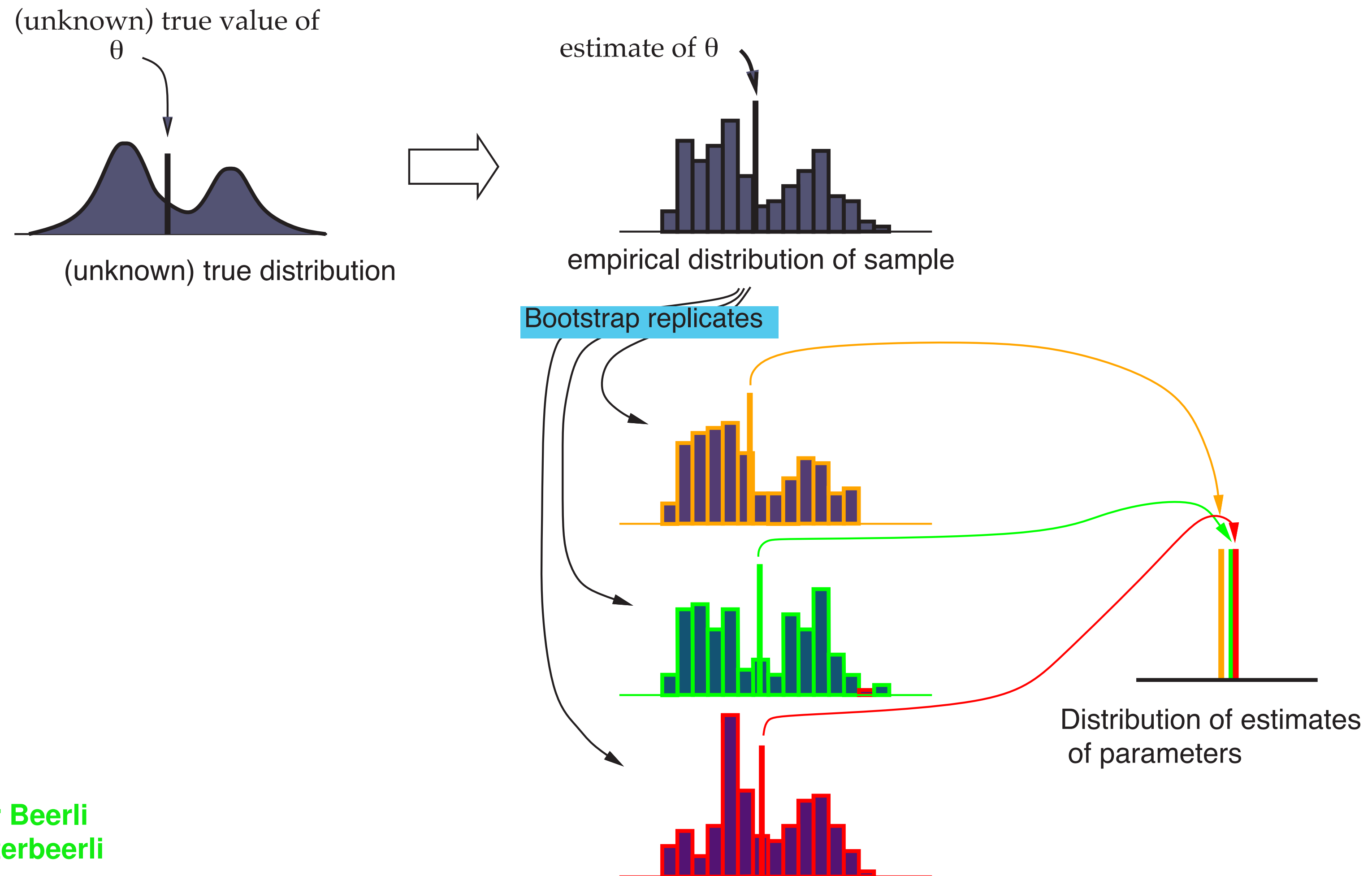
<http://phylo.bio.ku.edu/mephytis/boot-sample.html>

Bootstrapping phylogenies

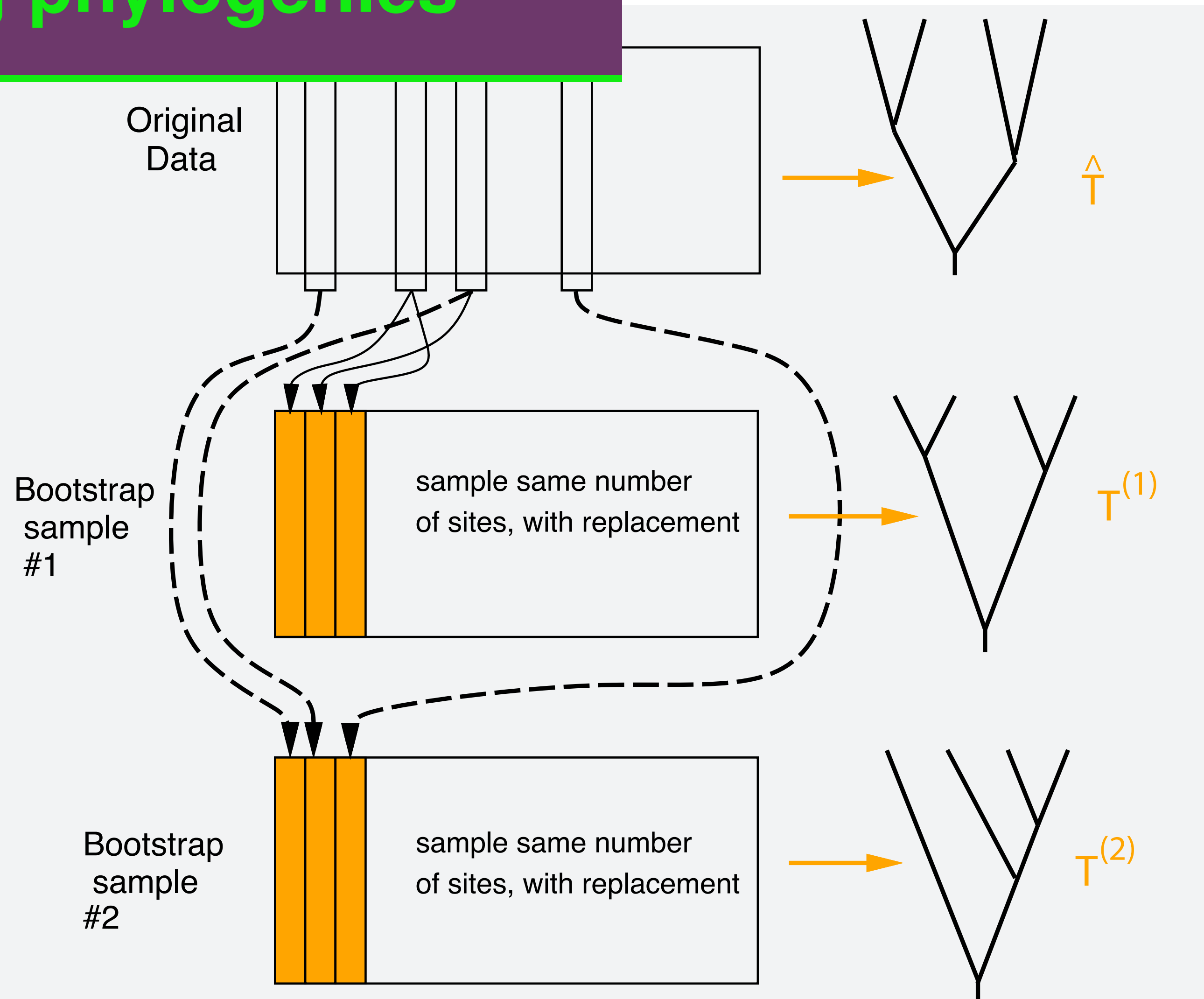
The bootstrap



Bootstrapping phylogenies



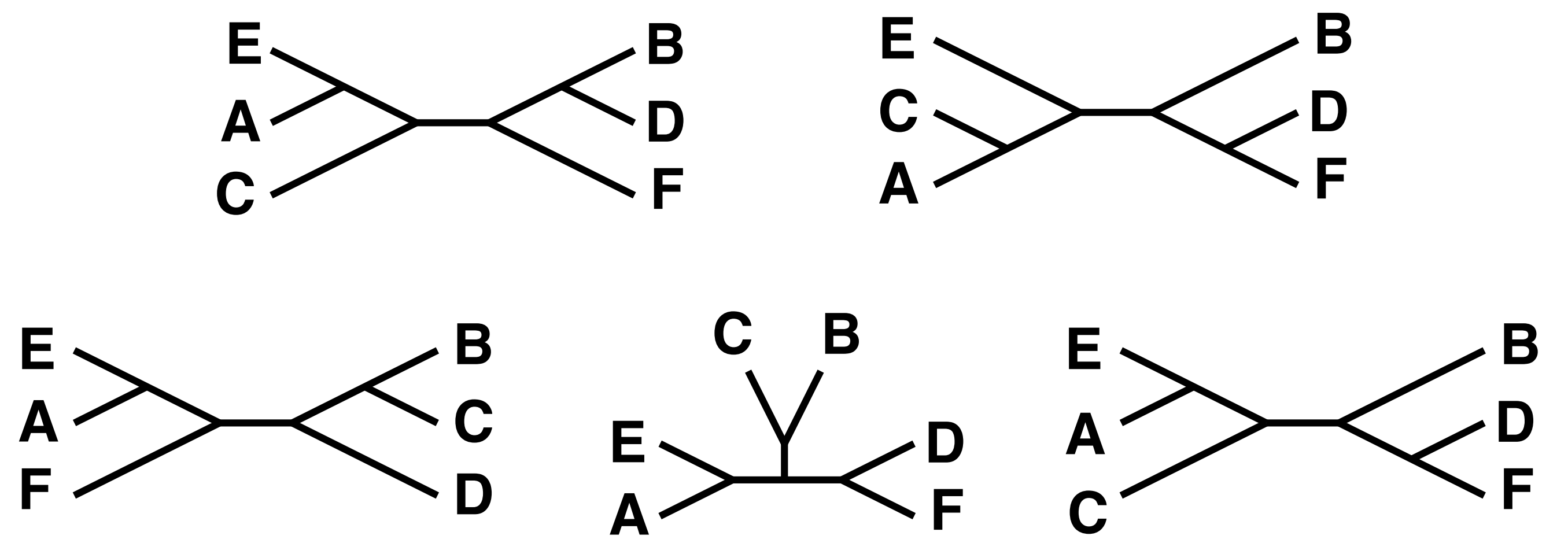
Bootstrapping phylogenies



Bootstrapping

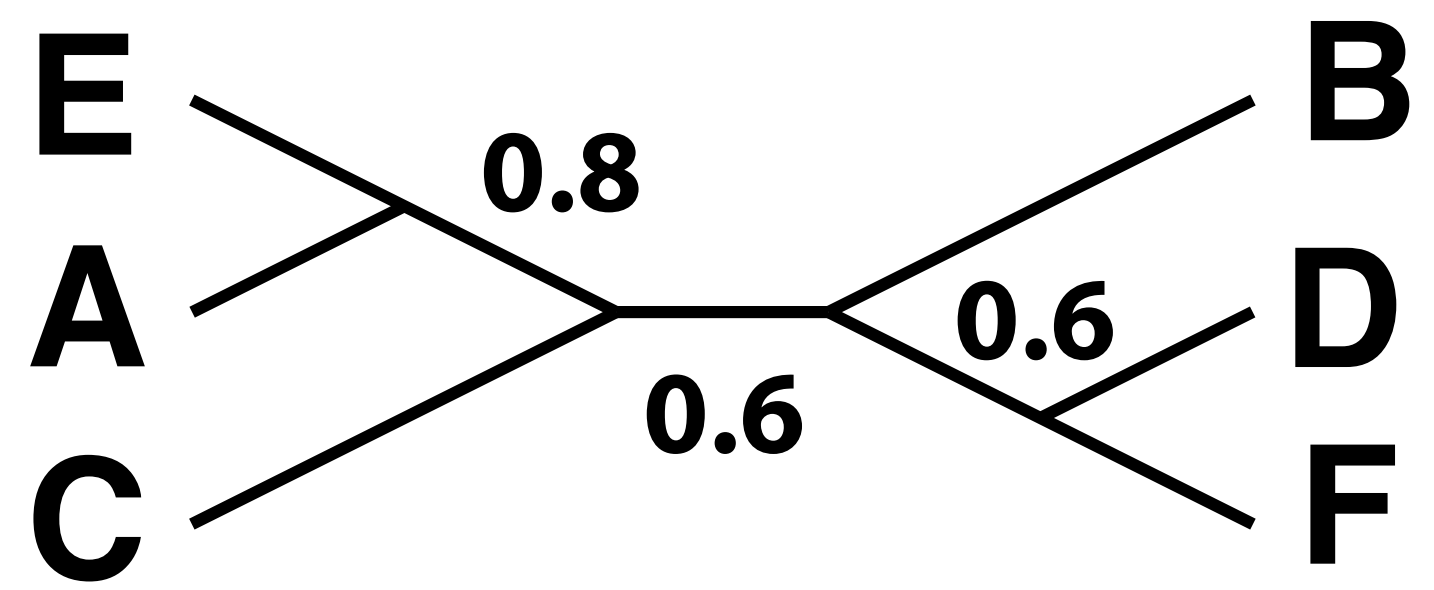
The majority-rule consensus tree

Trees:

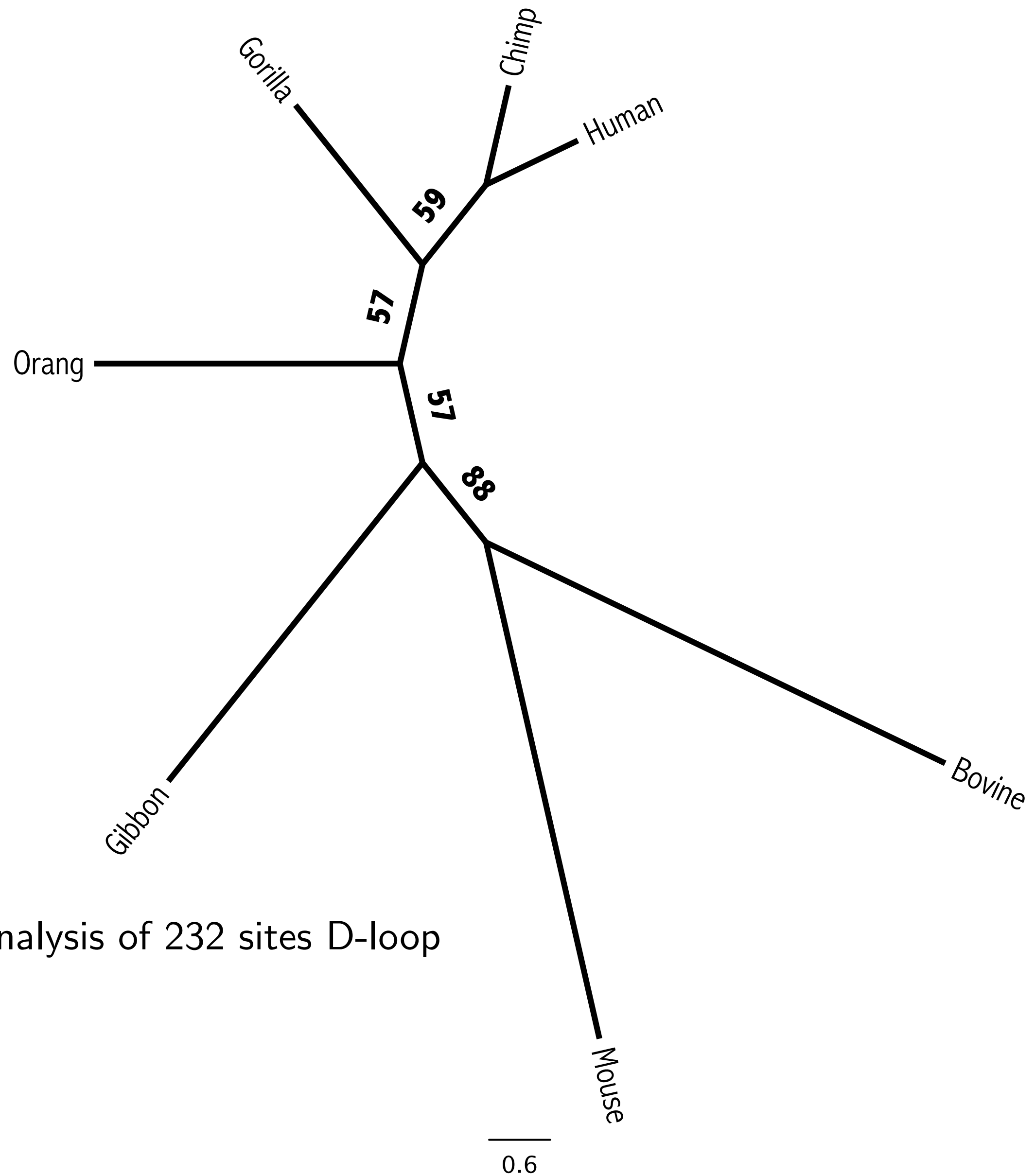


How many times each partition of species is found:

AE BCDF	4
ACE BDF	3
ACEF BD	1
AC BDEF	1
AEF BCD	1
ADEF BC	2
ABCE DF	3



Bootstrapping

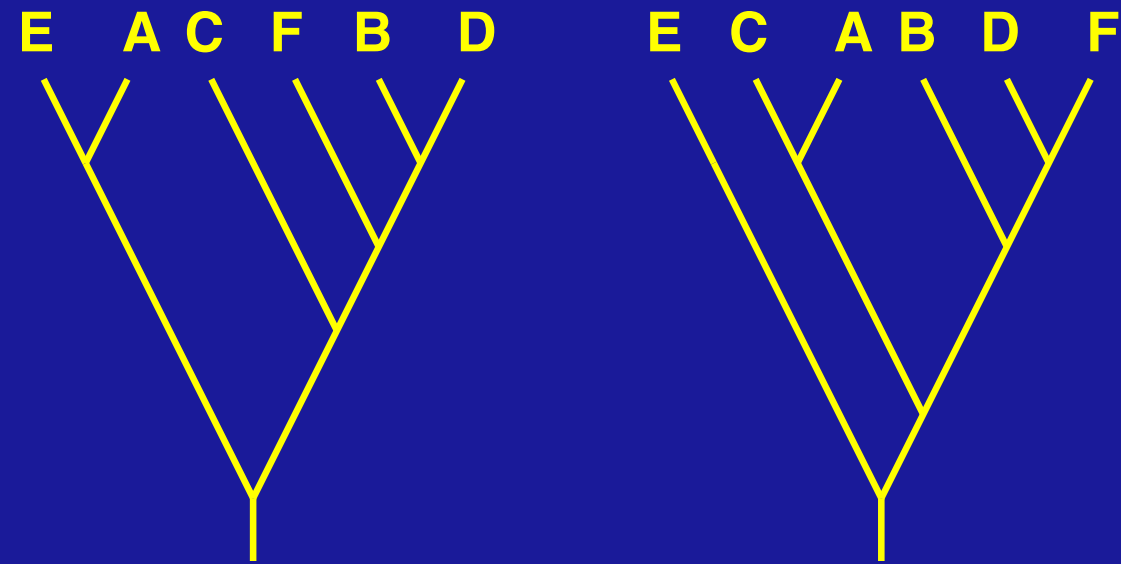


From Hasegawa's analysis of 232 sites D-loop

Summary

- ◆ The bootstrap allows you to generate a distribution based on your sample. This allows to take into account unknown correlation structure among the data entries.
- ◆ The bootstrap has also problem in that we may need to block sample to consider correlations among sampling entries (for example sites in DNA sequences are correlated, any scheme that draws bootstrap samples independently for each site may be flawed.)
- ◆ A great tool to handle sampling that deviates from normal distributions.

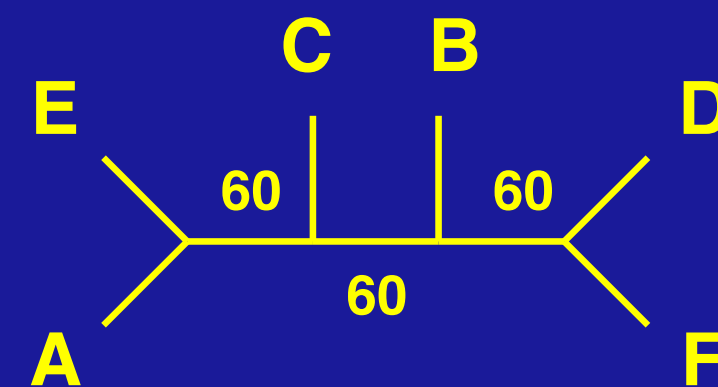
Trees:

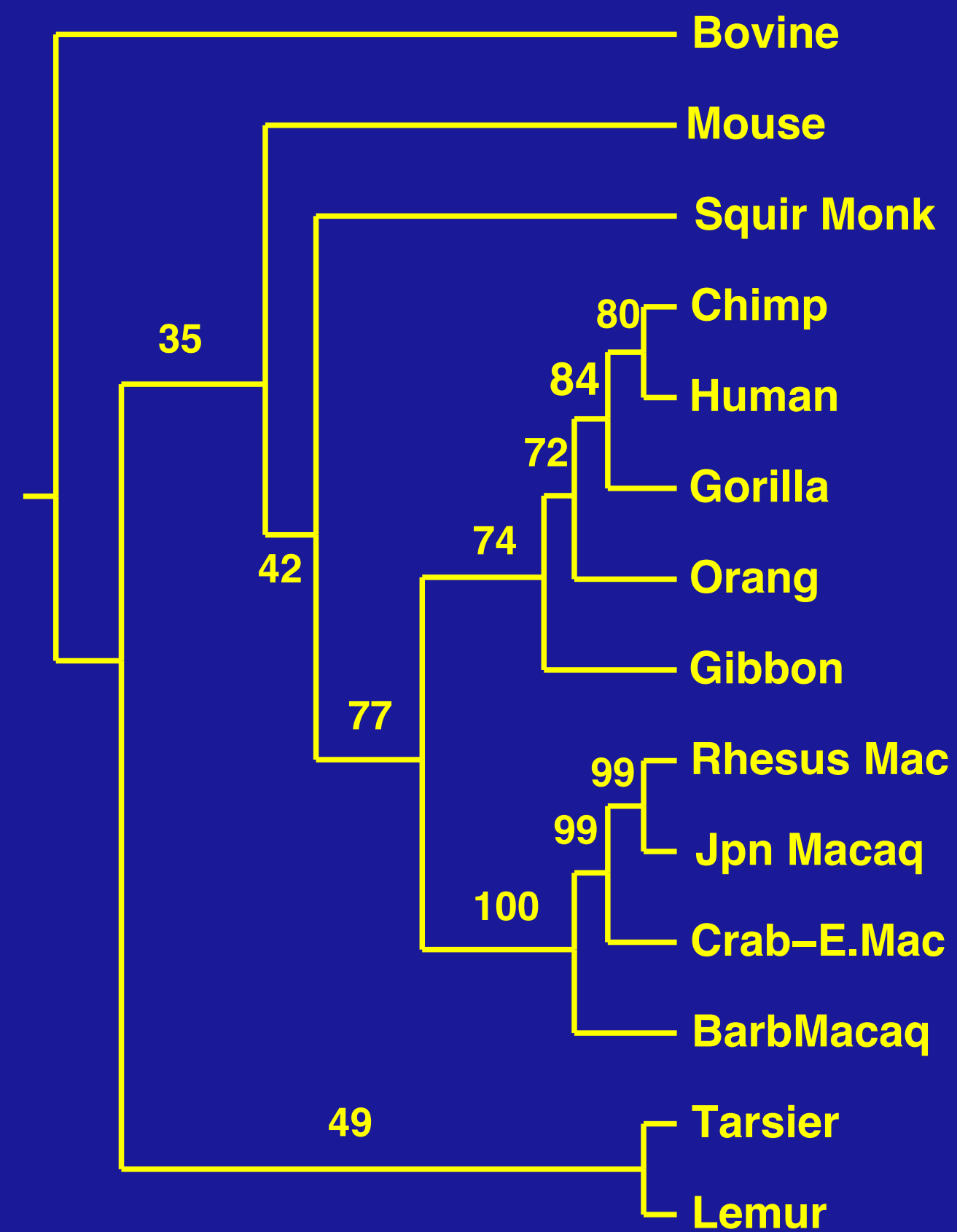


How many times each partition of species is found:

AE BCDF	3
ACE BDF	3
ACEF BD	1
AC BDEF	1
AEF BCD	1
ADEF BC	2
ABDF EC	1
ABCE DF	3

Majority-rule consensus tree of the unrooted trees:

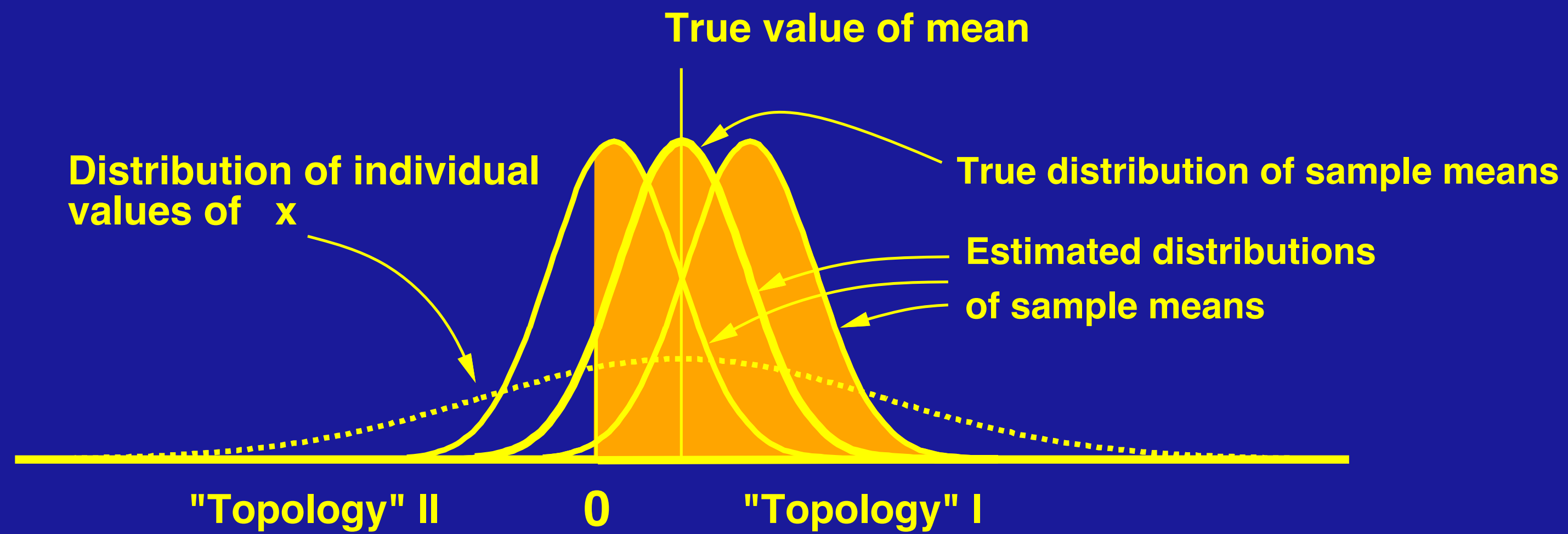




An example of bootstrap sampling of trees
232 nucleotide, 14-species mitochondrial D-loop data set
Analyzed by parsimony, 100 bootstrap replicates

Potential problems with the bootstrap

1. Sites may not evolve independently
2. Sites may not come from a common distribution (but can consider them sampled from a mixture of possible distributions)
3. If do not know which branch is of interest at the outset, a “multiple-tests” problem means P values are overstated
4. P values are biased (too conservative)
5. Bootstrapping does not correct biases in phylogeny methods



A model showing the bias in bootstrap P vales

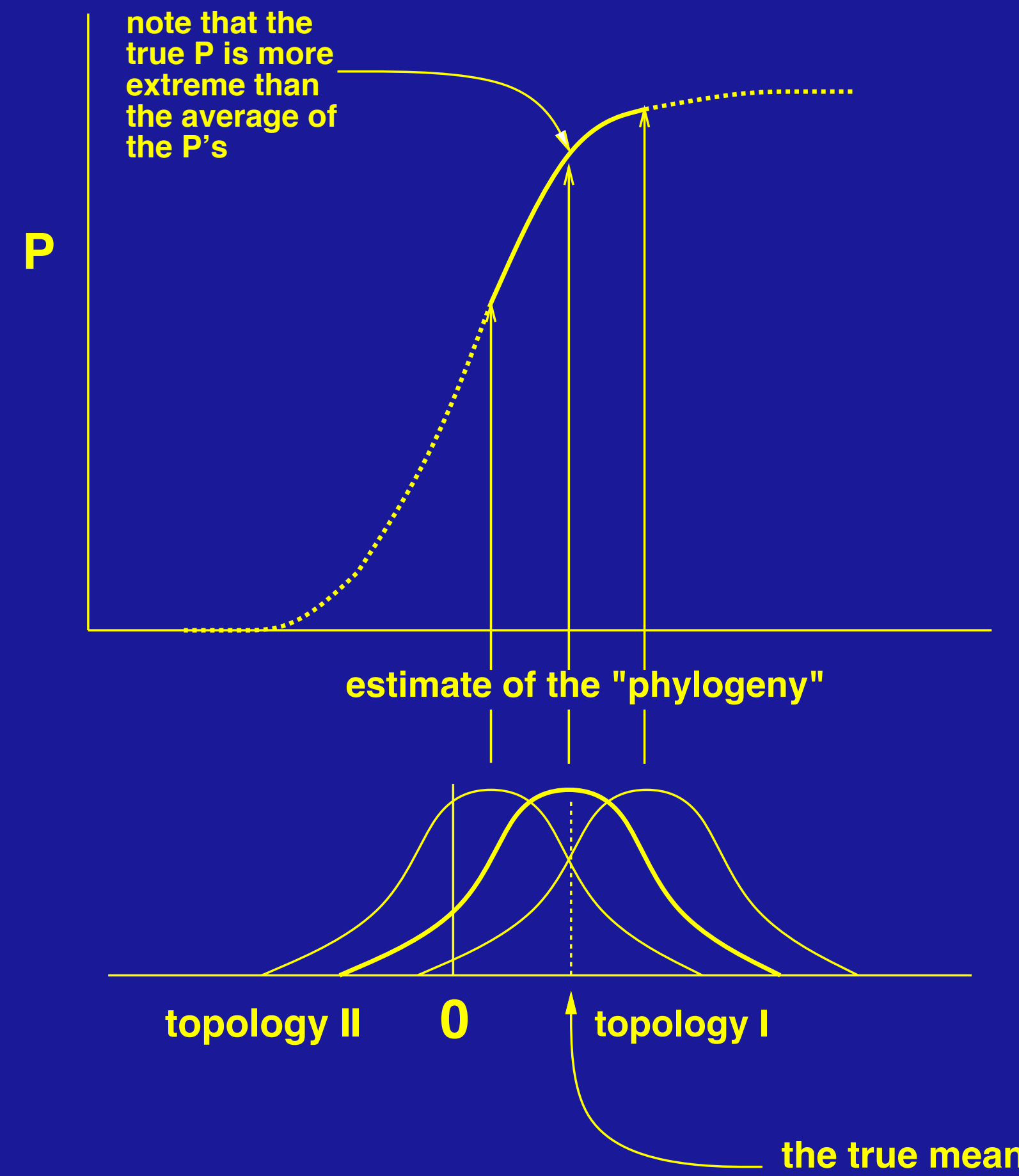
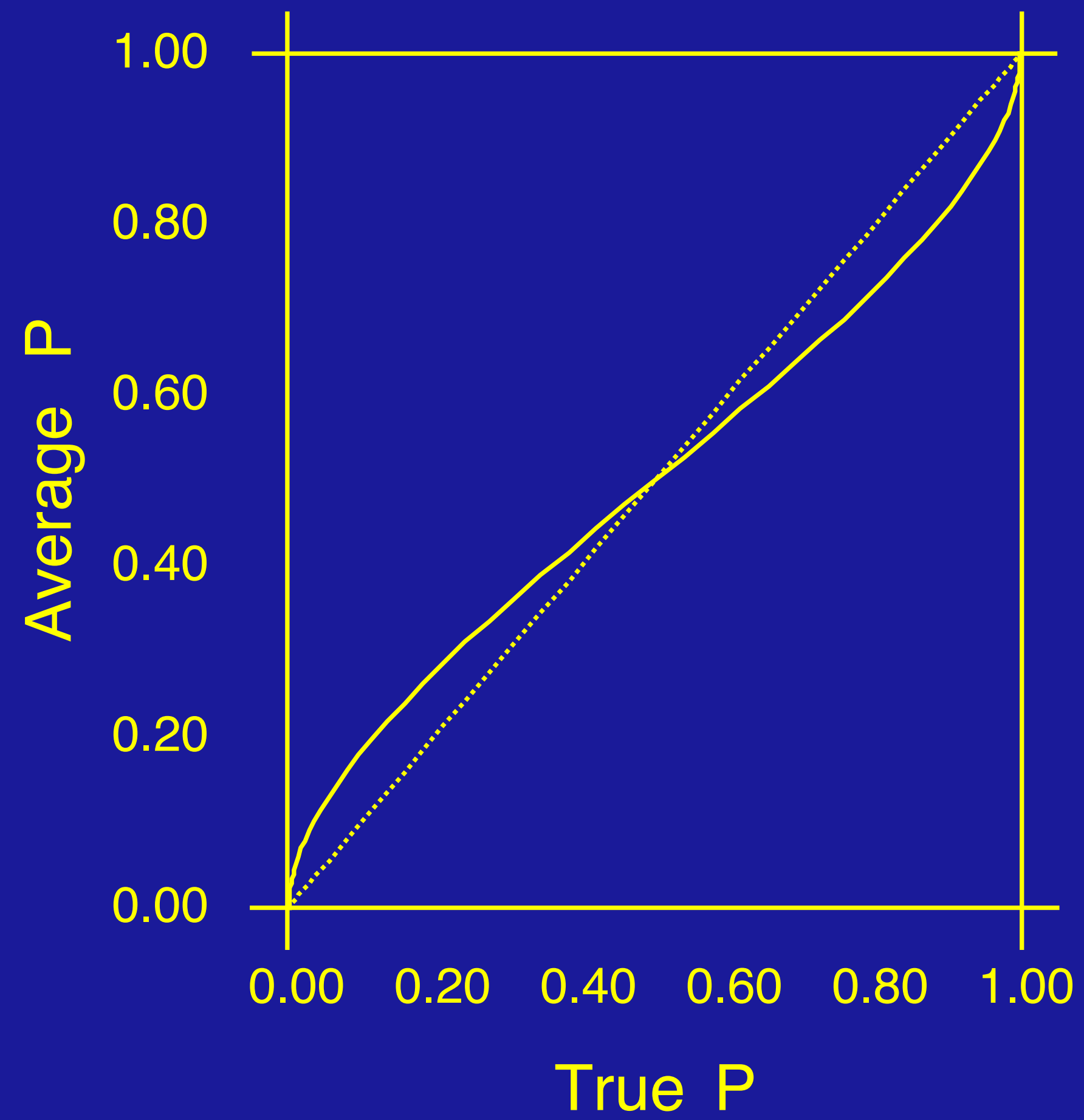
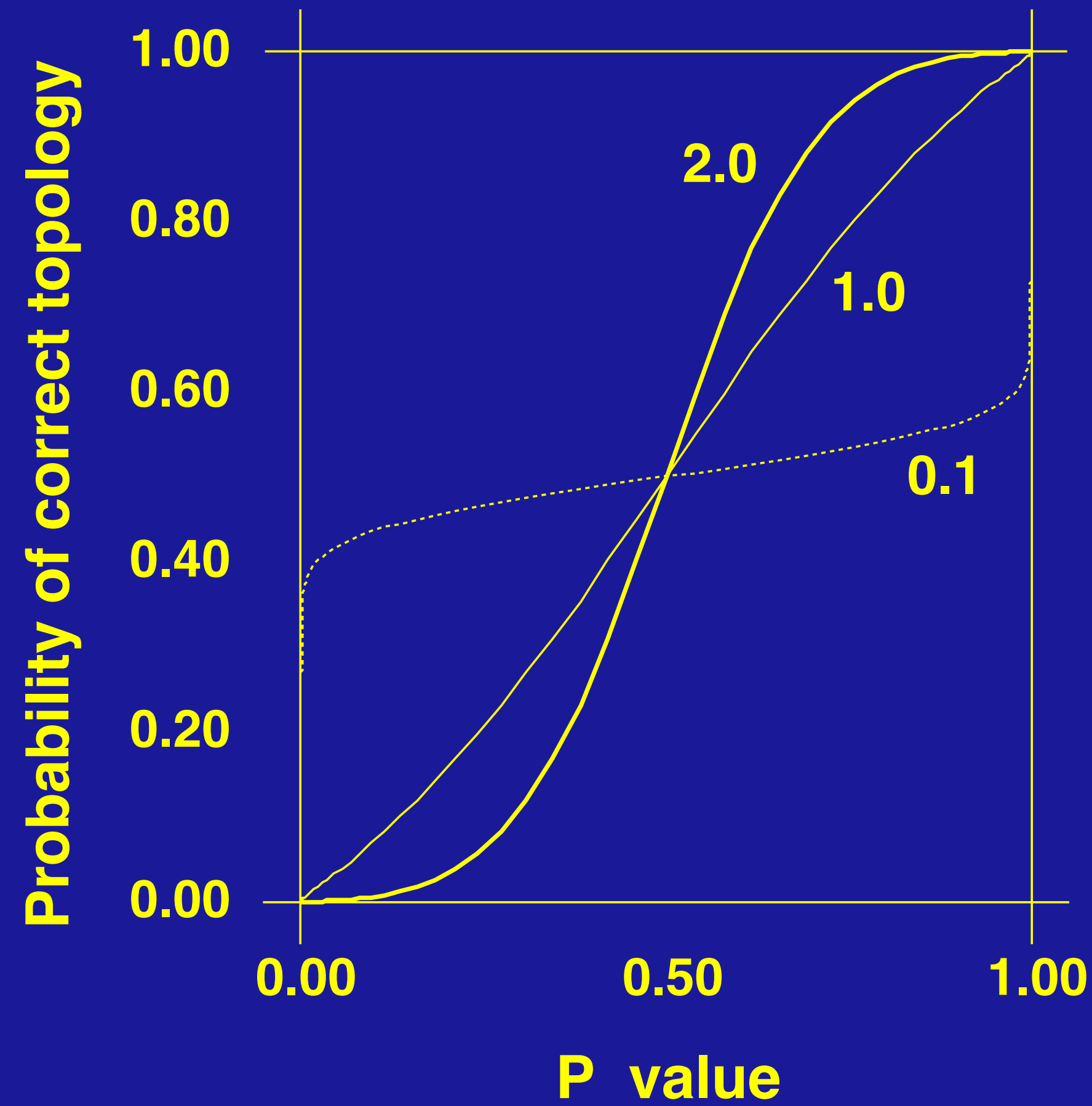


Illustration of the source of the bias



Extent of the bias in the example



Probability of being correct and variance of prior

Other resampling methods

- Delete-half jackknife. Sample a random 50% of the sites, *without* replacement.
- Delete- $1/e$ jackknife (Farris et. al. 1996) (too little deletion from a statistical viewpoint).
- Reweighting characters by choosing weights from an exponential distribution.
- In fact, reweighting them by any exchangeable weights having coefficient of variation of 1
- Parametric bootstrap – simulate data sets of this size assuming the estimate of the tree is the truth
- (to correct for correlation among adjacent sites) (Künsch, 1989) Block-bootstrapping – sample n/b blocks of b adjacent sites.