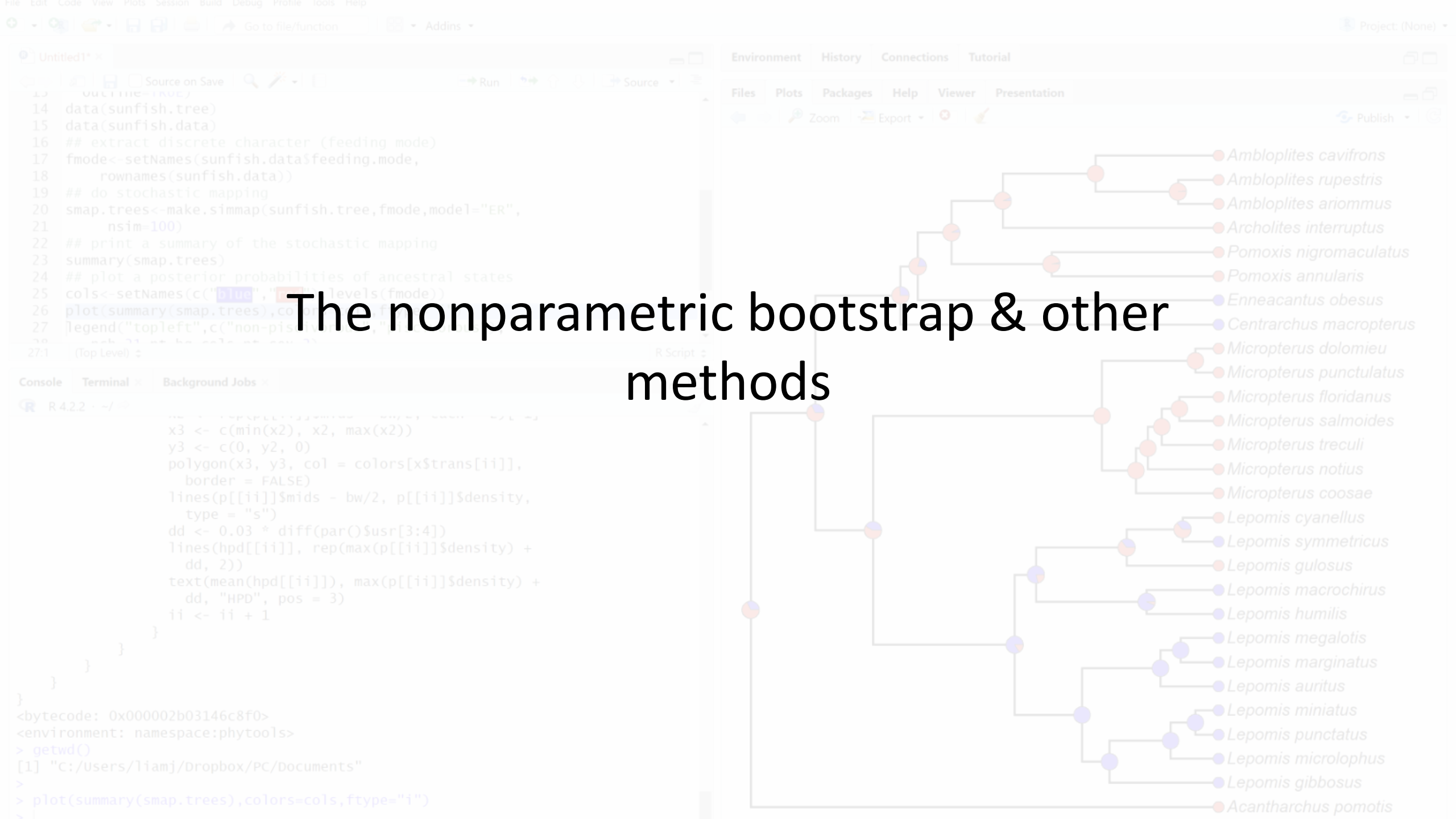
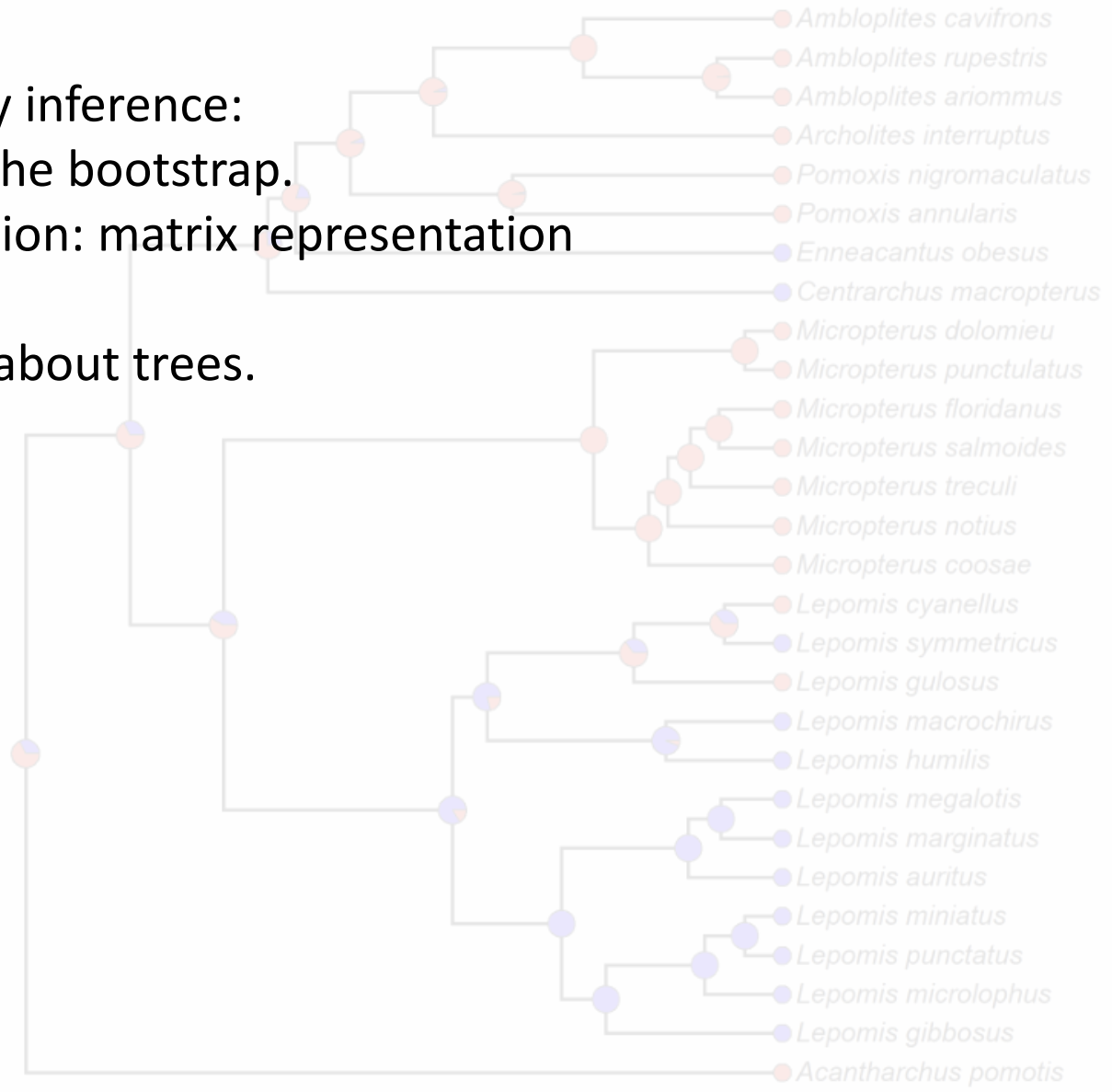
The nonparametric bootstrap & other methods



Agenda

1. Other topics in phylogeny inference:
   a) Assessing support: the bootstrap.
   b) “Supertree” estimation: matrix representation parsimony (MRP).
   c) Testing hypotheses about trees.



Evolution, 39(4), 1985, pp. 783–791

CONFIDENCE LIMITS ON PHYLOGENIES: AN APPROACH USING THE BOOTSTRAP

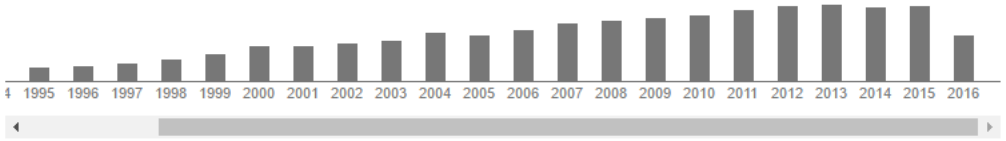
JOSEPH FELSENSTEIN
Department of Genetics SK-50, University of Washington, Seattle, WA 98195

Joe Felsenstein

Confidence limits on phylogenies: an approach using the bootstrap

[PDF] from ed.ac.uk

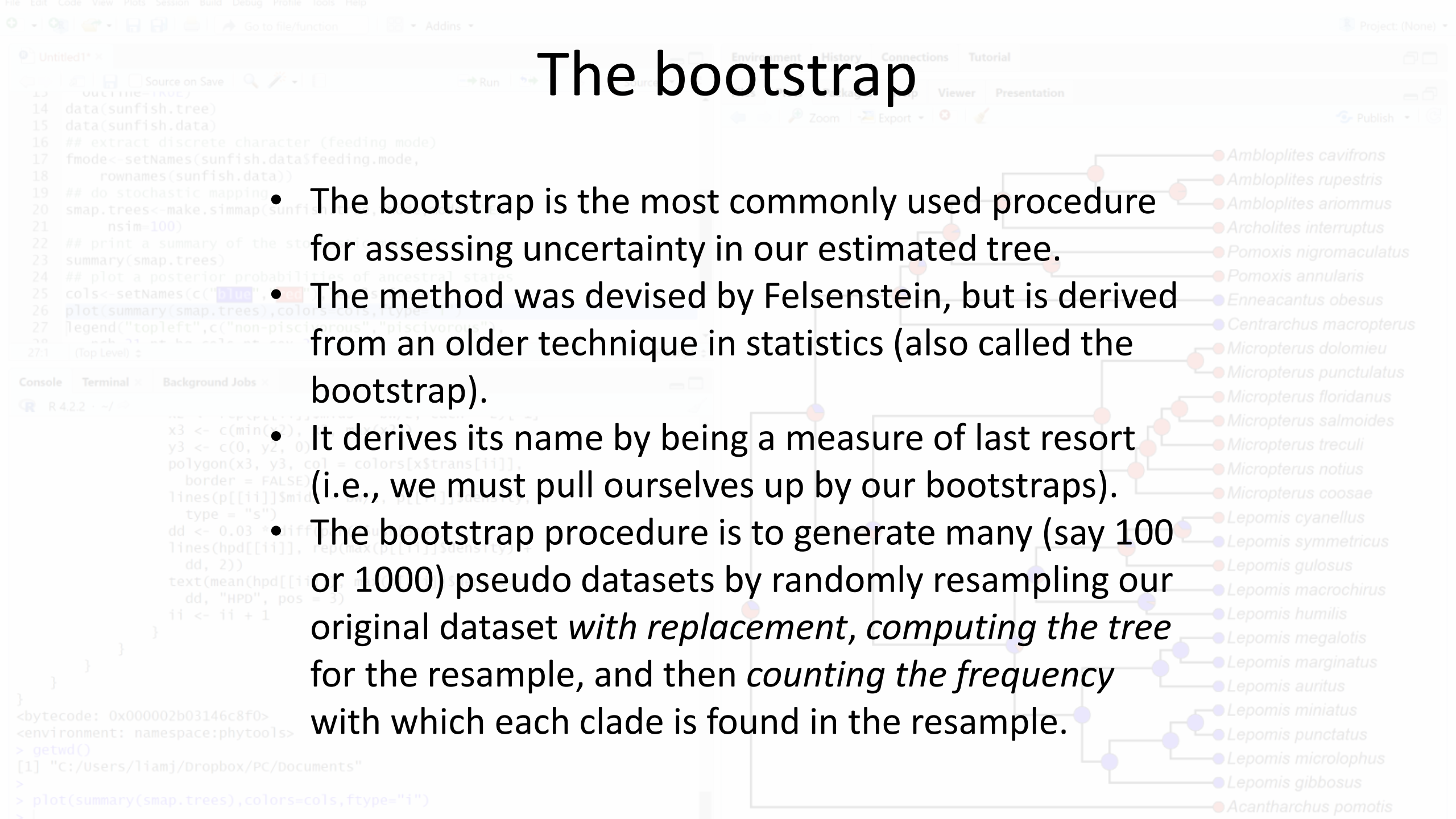
Authors	Joseph Felsenstein
Publication date	1985/7/1
Journal	Evolution
Pages	783-791
Publisher	Society for the Study of Evolution
Description	The recently-developed statistical method known as the" bootstrap" can be used to place confidence intervals on phylogenies. It involves resampling points from one's own data, with replacement, to create a series of bootstrap samples of the same size as the original data. Each of these is analyzed, and the variation among the resulting estimates taken to indicate the size of the error involved in making estimates from the original data. In the case of phylogenies, it is argued that the proper method of resampling is to keep all of the original ...
Total citations	Cited by 30906



Scholar articles
Confidence limits on phylogenies: an approach using the bootstrap
J Felsenstein - Evolution, 1985
Cited by 30906 - Related articles - All 20 versions

The bootstrap

- The bootstrap is the most commonly used procedure for assessing uncertainty in our estimated tree.
- The method was devised by Felsenstein, but is derived from an older technique in statistics (also called the bootstrap).
- It derives its name by being a measure of last resort (i.e., we must pull ourselves up by our bootstraps).
- The bootstrap procedure is to generate many (say 100 or 1000) pseudo datasets by randomly resampling our original dataset *with replacement*, *computing the tree* for the resample, and then *counting the frequency* with which each clade is found in the resample.



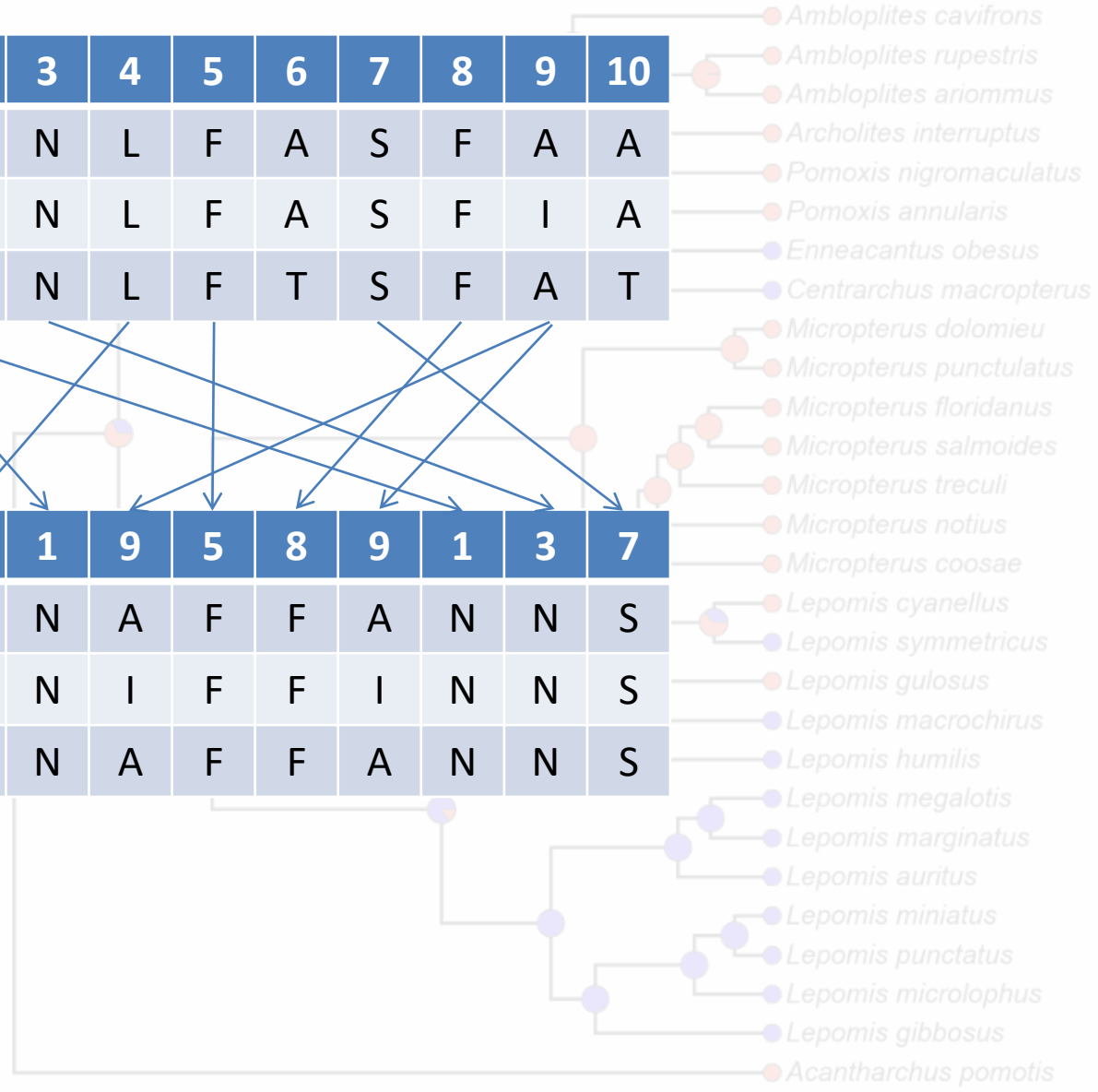
Bootstrap resample

Original alignment

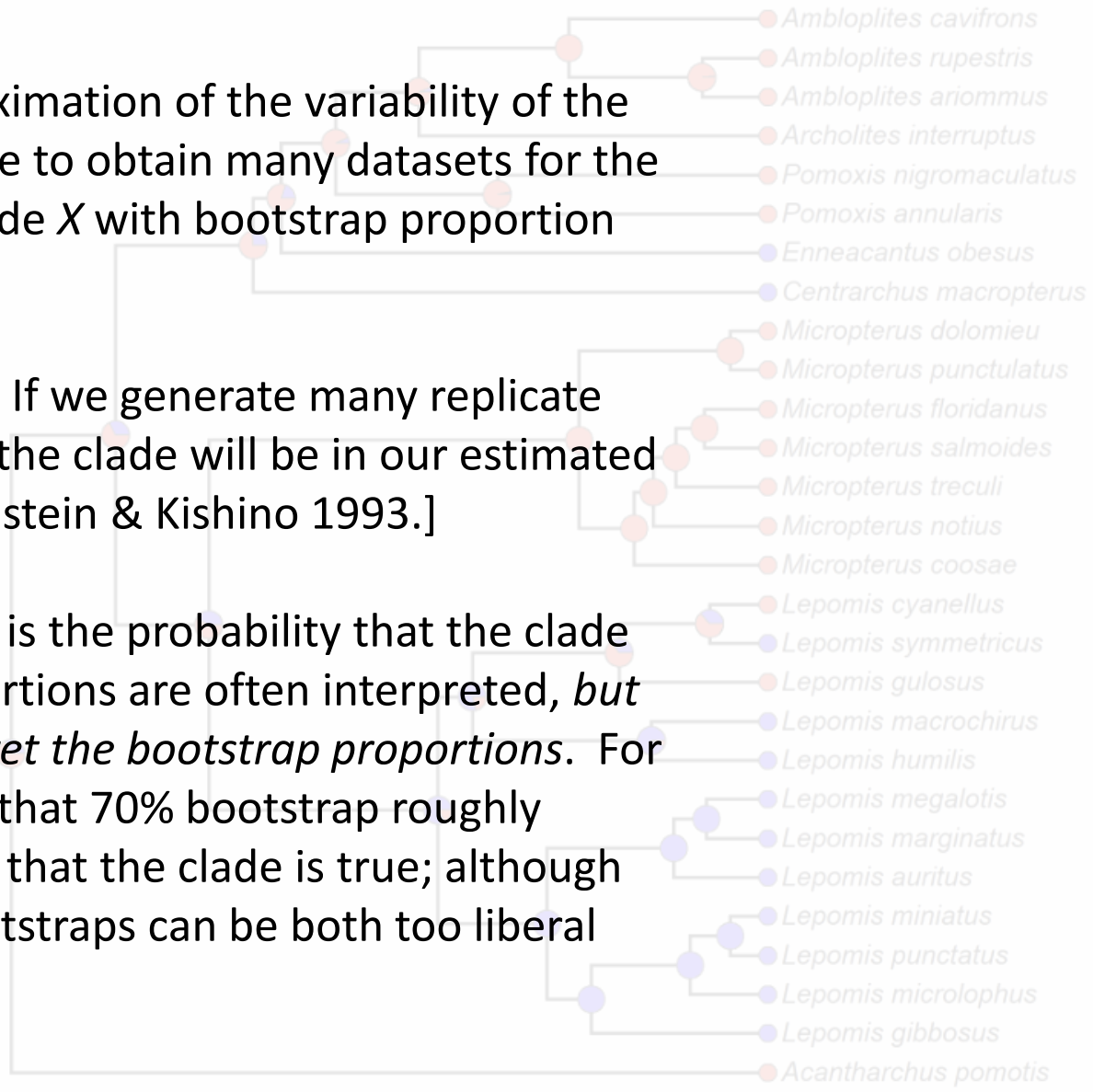
Site	1	2	3	4	5	6	7	8	9	10
Chimp	N	E	N	L	F	A	S	F	A	A
Gorilla	N	E	N	L	F	A	S	F	I	A
Gibbon	N	E	N	L	F	T	S	F	A	T

Bootstrap alignment

Site	2	4	1	9	5	8	9	1	3	7
Chimp	E	L	N	A	F	F	A	N	N	S
Gorilla	E	L	N	I	F	F	I	N	N	S
Gibbon	E	L	N	A	F	F	A	N	N	S



Interpretations of the bootstrap



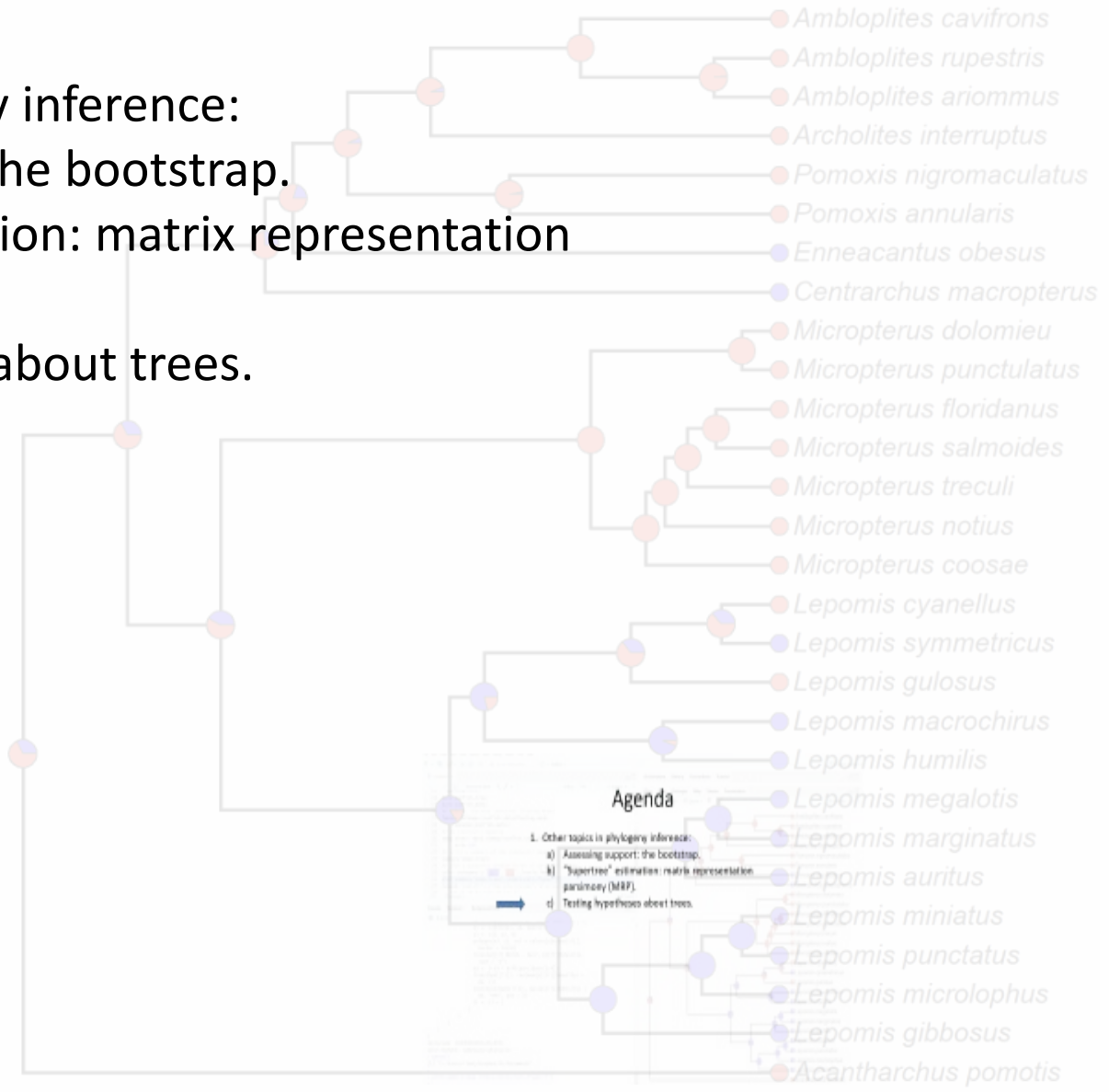
Repeatability: Bootstraps are approximation of the variability of the generating process. Thus, if we were to obtain many datasets for the same set of taxa, we would infer clade *X* with bootstrap proportion *P*(*X*). [But see Hillis & Bull 1993.]

Type-I error rate/False positive rate: If we generate many replicate datasets in which clade *X* is *absent*, the clade will be in our estimated tree with probability 1-*P*(*X*). [Felsenstein & Kishino 1993.]

Accuracy: The bootstrap proportion is the probability that the clade is true. This is how bootstrap proportions are often interpreted, *but this is probably a bad way to interpret the bootstrap proportions*. For instance, some studies have shown that 70% bootstrap roughly corresponds to a probability of 0.95 that the clade is true; although this may not be true in general (bootstraps can be both too liberal and too conservative).

Agenda

1. Other topics in phylogeny inference:
   a) Assessing support: the bootstrap.
   b) “Supertree” estimation: matrix representation parsimony (MRP).
   c) Testing hypotheses about trees.



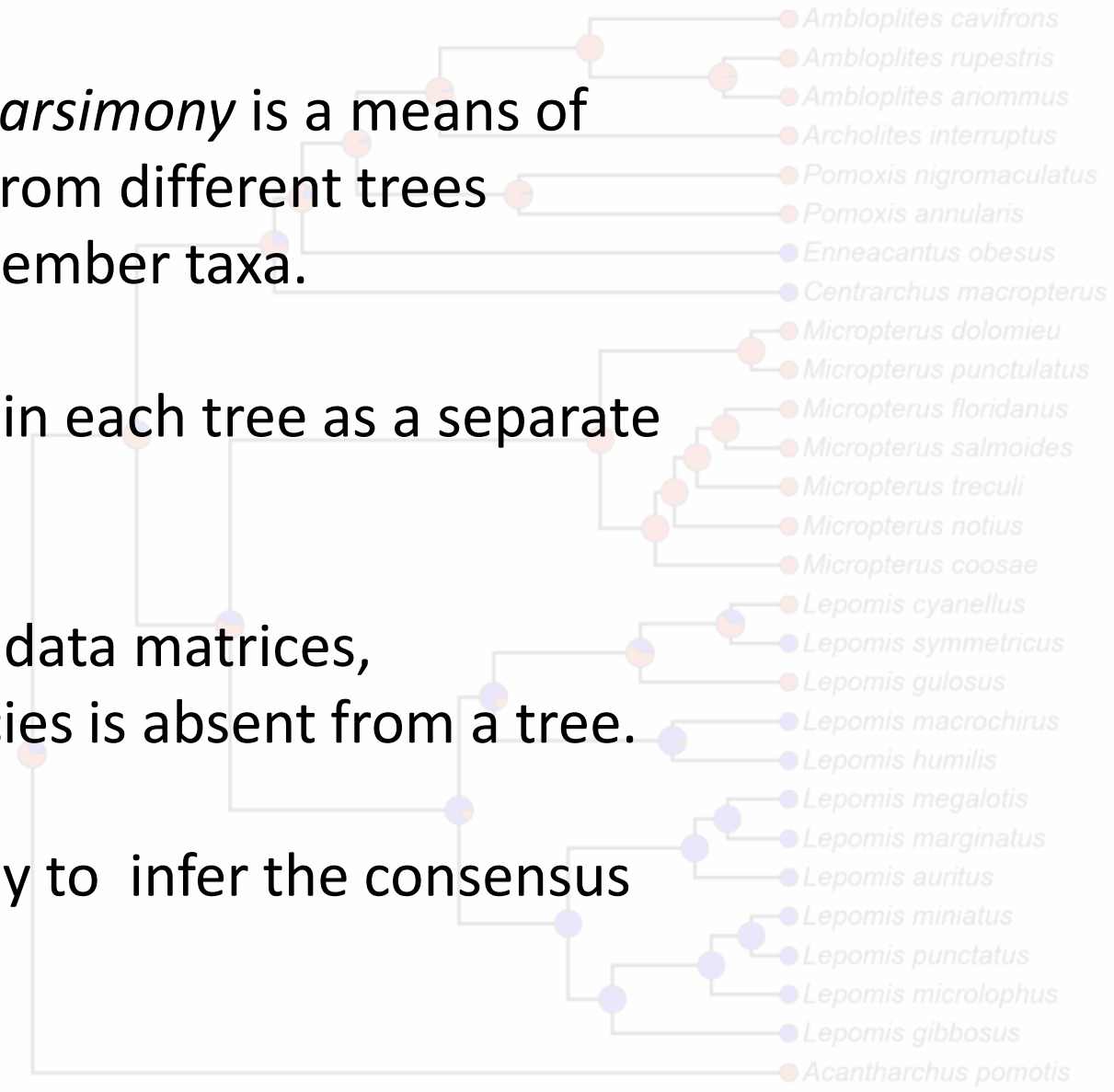
Supertree via MRP

- *Matrix Representation Parsimony* is a means of combining information from different trees consisting of different member taxa.

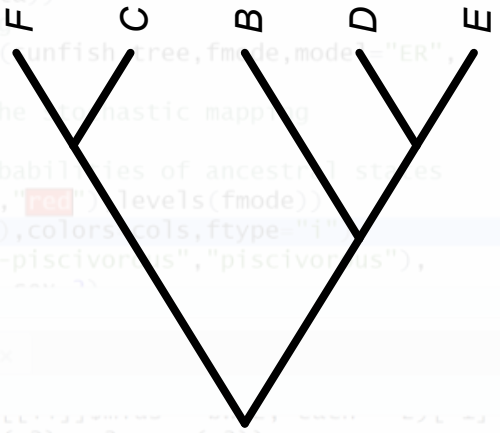
- We first score each split in each tree as a separate binary character.

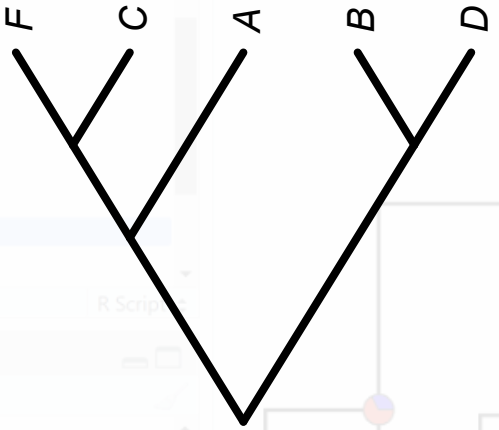
- We then combine these data matrices, substituting “?” if a species is absent from a tree.

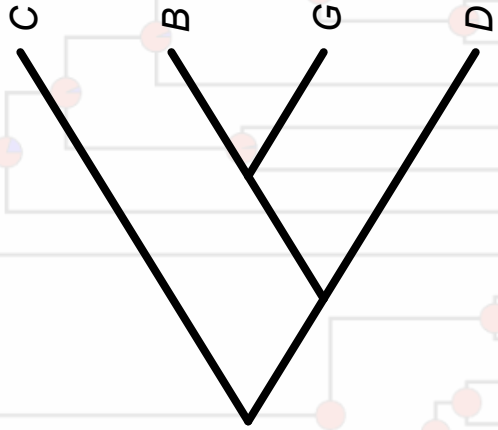
- Finally, we use parsimony to infer the consensus tree from this matrix.



Supertree via MRP







	[,1]	[,2]	[,3]
F	1	0	0
C	1	0	0
B	0	1	0
D	0	1	1
E	0	1	1

	[,1]	[,2]	[,3]
F	1	1	0
C	1	1	0
A	1	0	0
B	0	0	1
D	0	0	1

	[,1]	[,2]
C	0	0
B	1	1
G	1	1
D	1	0

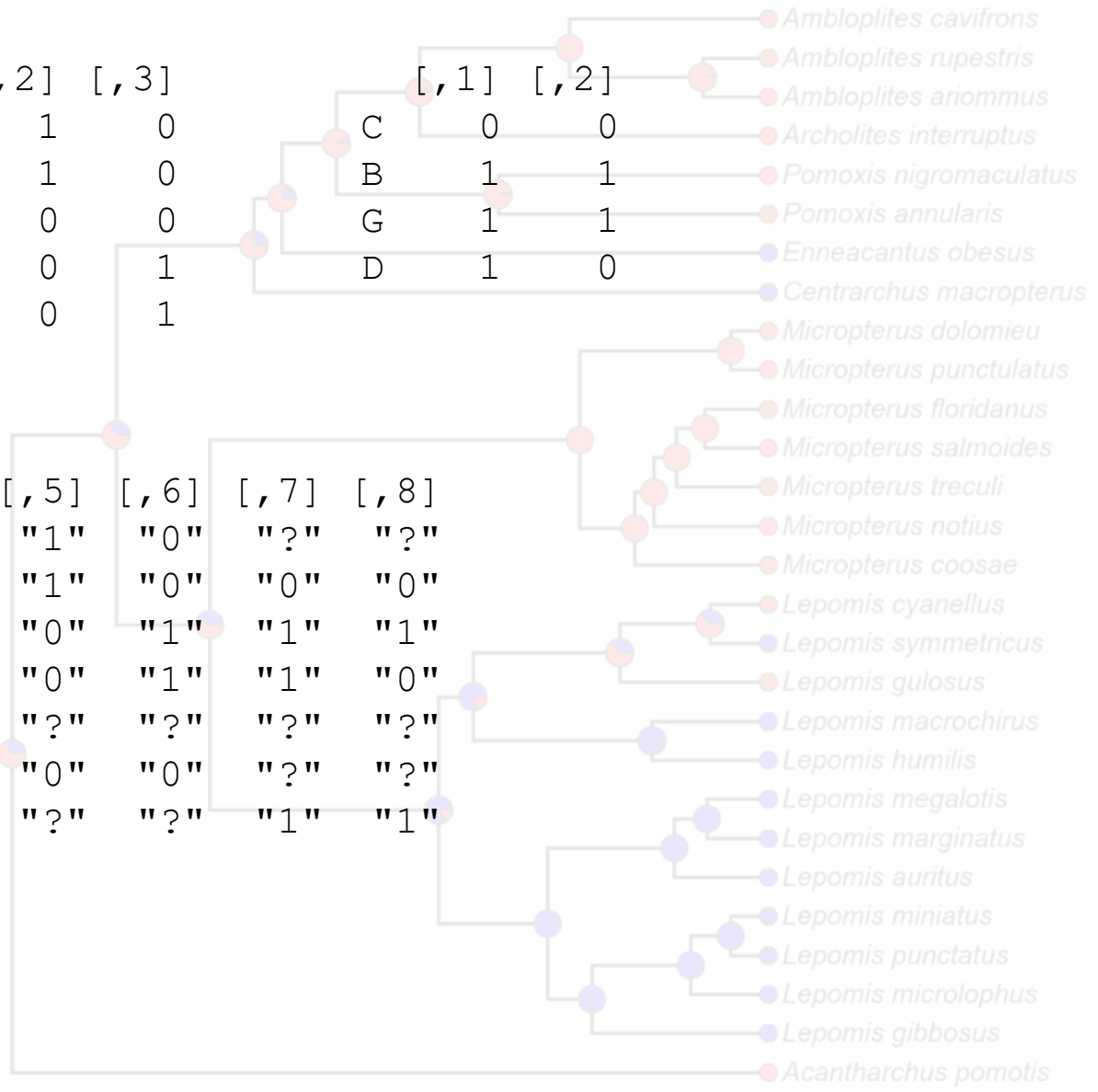
Supertree via MRP

	[,1]	[,2]	[,3]
F	1	0	0
C	1	0	0
B	0	1	0
D	0	1	1
E	0	1	1

	[,1]	[,2]	[,3]
F	1	1	0
C	1	1	0
A	1	0	0
B	0	0	1
D	0	0	1

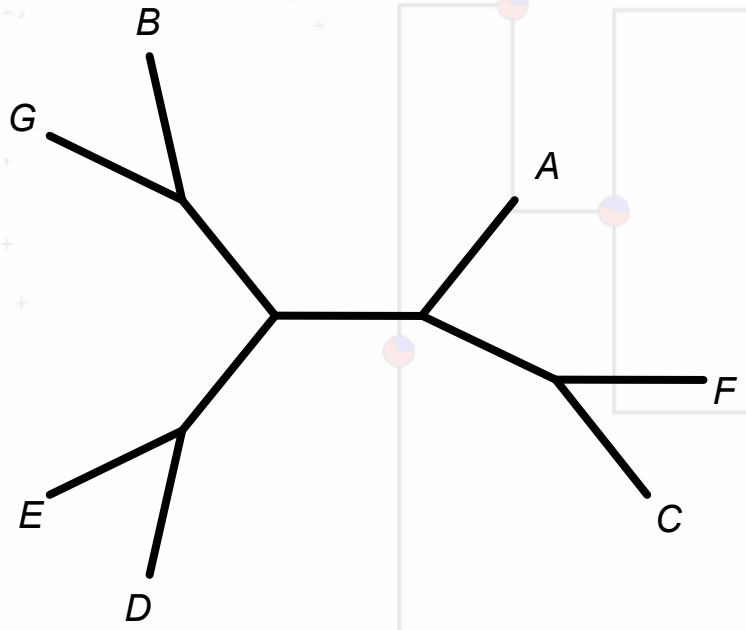
	[,1]	[,2]
C	0	0
B	1	1
G	1	1
D	1	0

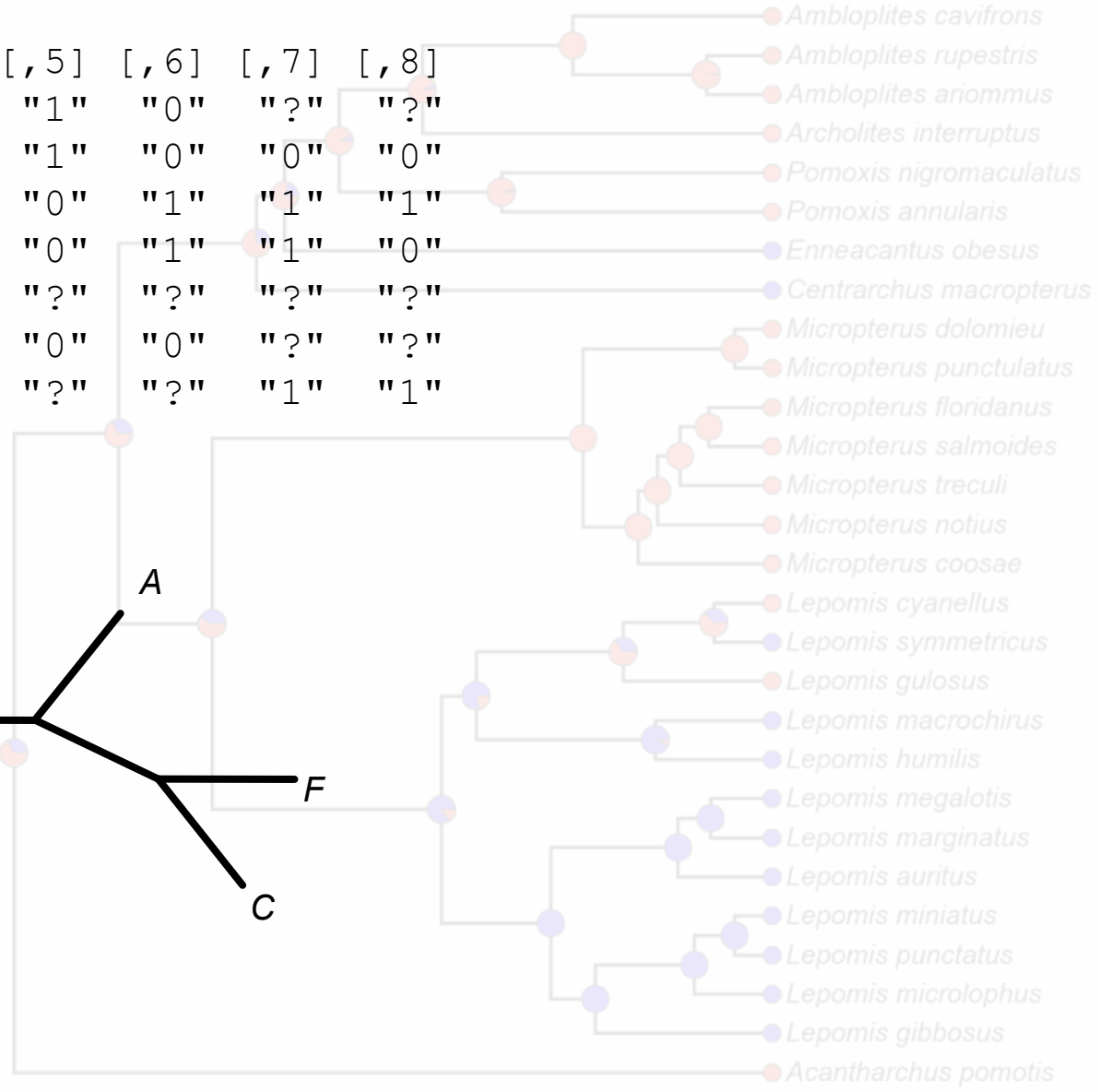
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
F	"1"	"0"	"0"	"1"	"1"	"0"	"?"	"?"
C	"1"	"0"	"0"	"1"	"1"	"0"	"0"	"0"
B	"0"	"1"	"0"	"0"	"0"	"1"	"1"	"1"
D	"0"	"1"	"1"	"0"	"0"	"1"	"1"	"0"
E	"0"	"1"	"1"	"?"	"?"	"?"	"?"	"?"
A	"?"	"?"	"?"	"1"	"0"	"0"	"?"	"?"
G	"?"	"?"	"?"	"?"	"?"	"?"	"1"	"1"



Supertree via MRP

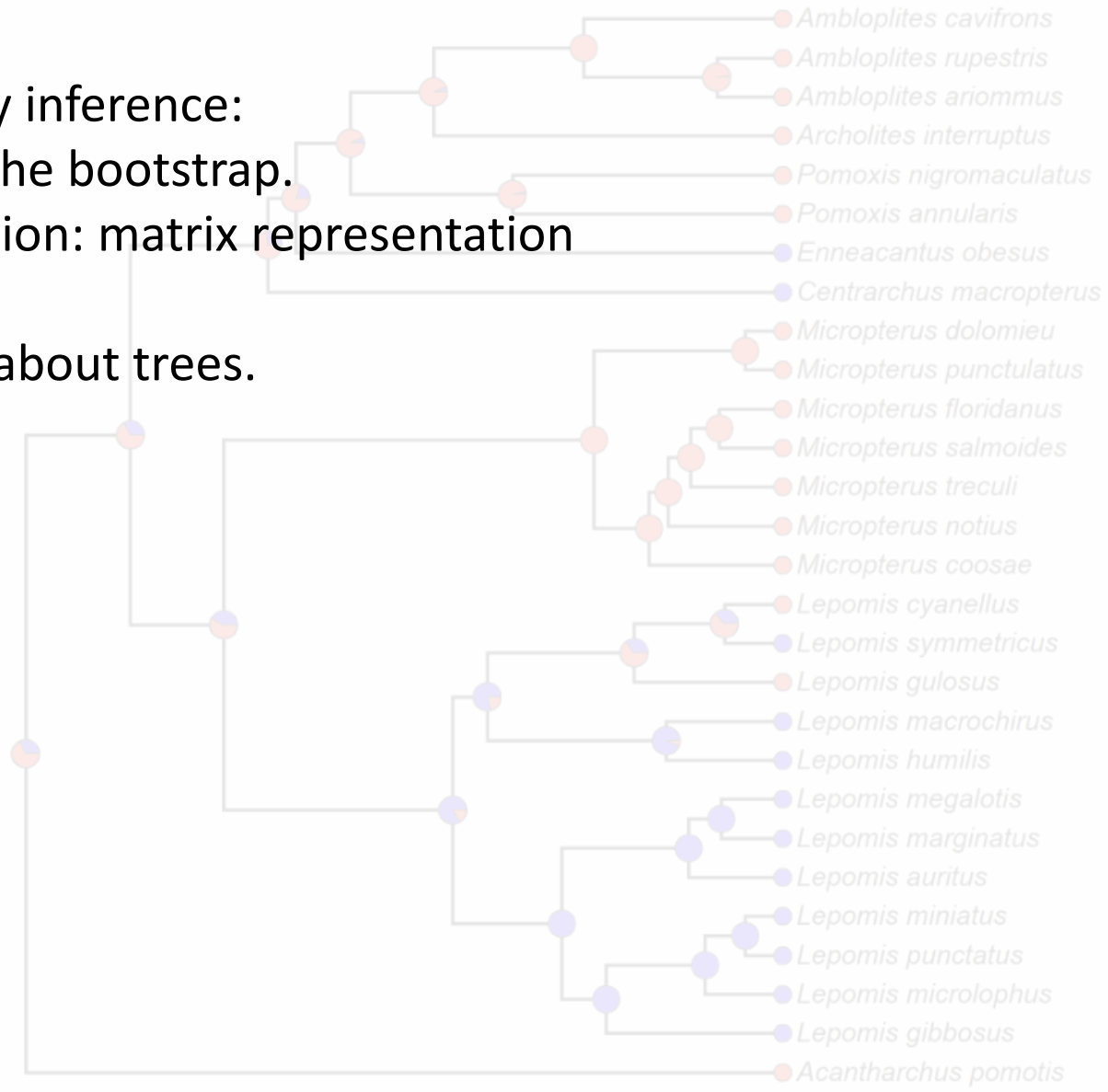
	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]
F	"1"	"0"	"0"	"1"	"1"	"0"	"?"	"?"
C	"1"	"0"	"0"	"1"	"1"	"0"	"0"	"0"
B	"0"	"1"	"0"	"0"	"0"	"1"	"1"	"1"
D	"0"	"1"	"1"	"0"	"0"	"1"	"1"	"0"
E	"0"	"1"	"1"	"?"	"?"	"?"	"?"	"?"
A	"?"	"?"	"?"	"1"	"0"	"0"	"?"	"?"
G	"?"	"?"	"?"	"?"	"?"	"?"	"1"	"1"





Agenda

1. Other topics in phylogeny inference:
   a) Assessing support: the bootstrap.
   b) “Supertree” estimation: matrix representation parsimony (MRP).
   c) Testing hypotheses about trees.

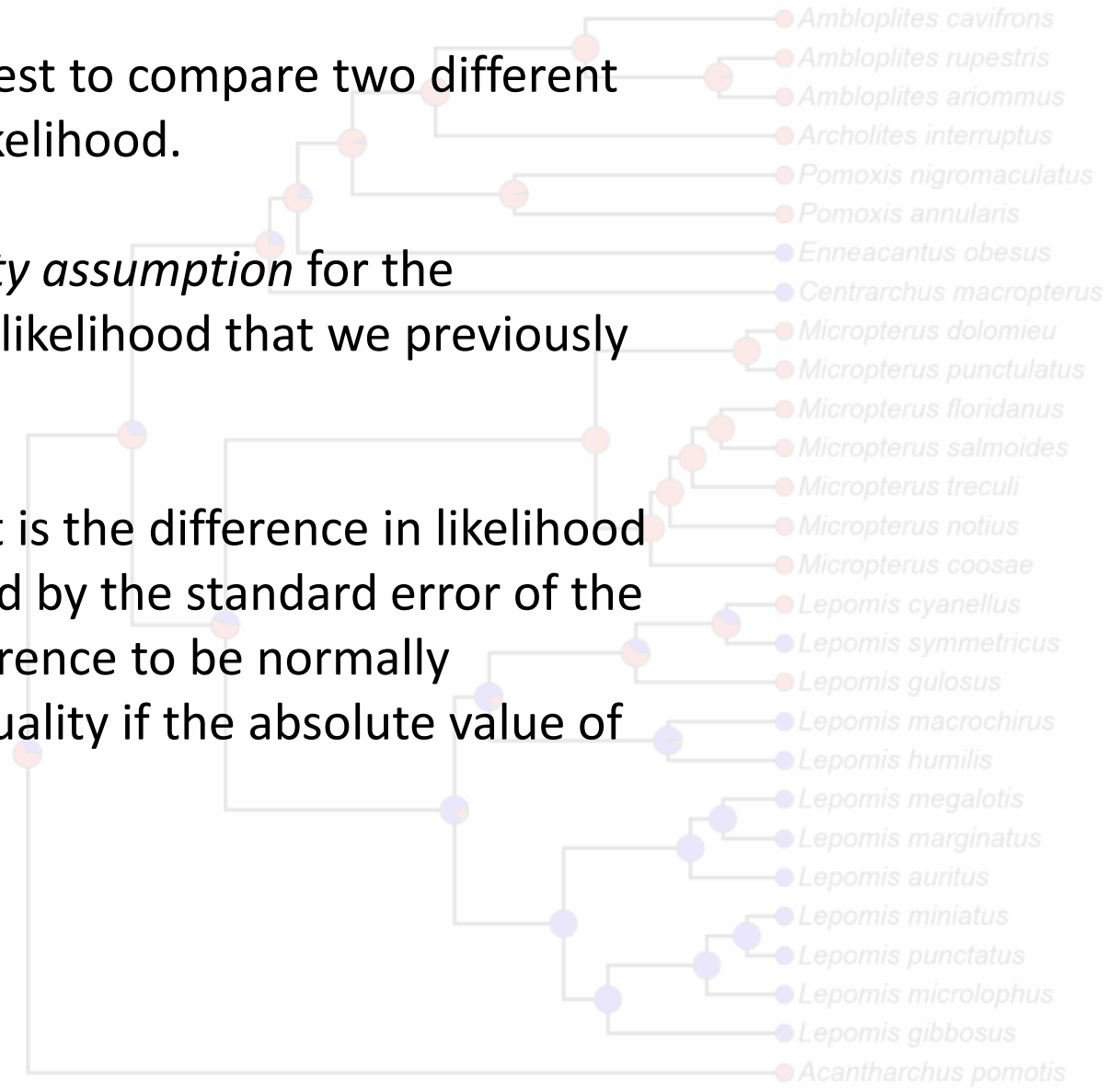


Comparing two trees: the K-H test

Kishino & Hasegawa devised a test to compare two different trees, specified *a priori*, using likelihood.

The K-H test relies on a *normality assumption* for the likelihoods (this is a property of likelihood that we previously discussed).

The test statistic for the K-H test is the difference in likelihood between the two models divided by the standard error of the difference. We expect this difference to be normally distributed, so we can reject equality if the absolute value of this ratio is greater than 1.96.



Comparing two trees: the K-H test

In particular, we compute the difference in the log-likelihoods of the two trees (Δ) as:

$$\Delta = l_1 - l_2$$

where the following is the log-likelihood of tree *i*, **x** is the data, and our dataset consists of *n* sites:

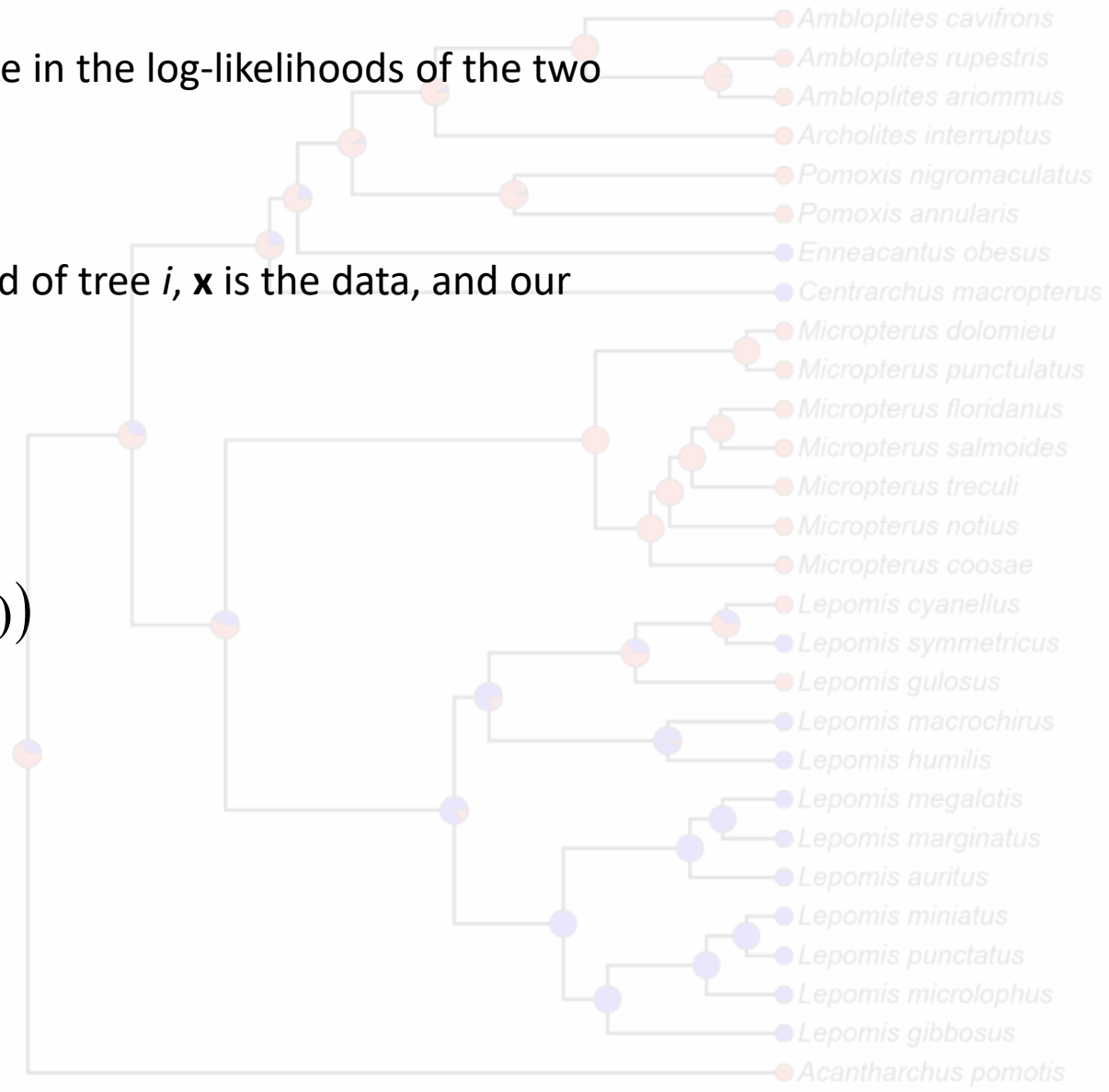
$$l_i = \sum_{h=1}^{n} \log\left(L(x_h \mid \tau_i)\right)$$

We can then compute:

$$d_h = \log\left(L(x_h \mid \tau_1)\right) - \log\left(L(x_h \mid \tau_2)\right)$$

and then the variance in Δ is just:

$$\mathrm{var}(\Delta) = \frac{n}{n-1} \sum_{h=1}^{n} (d_h - \Delta/n)^2$$



Comparing two trees

The K-H test was designed to compare two trees that had been specified *a priori*; however it has often been used to compare a hypothetical tree to the ML tree.

This is an incorrect procedure and will lead to tend to reject the null hypothesis that the trees are equally likely. (This could probably be ameliorated by using a two-tailed test, but I am not aware that this has been explored.)

Shimodaira & Hasegawa (1999) developed a more conservative test to correct for the bias of the K-H test. The details are beyond the scope of this class, but it is implemented in the {phangorn} R package.