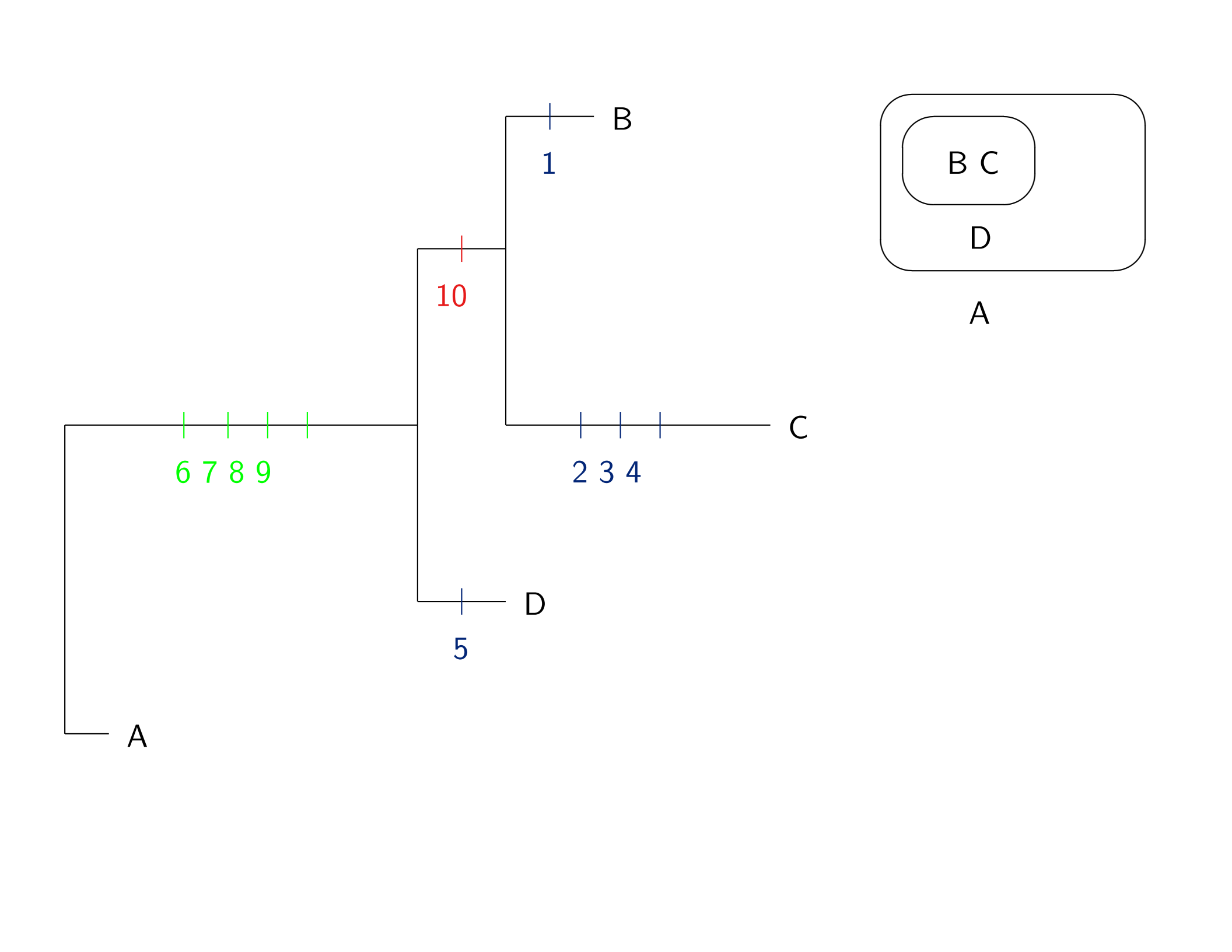
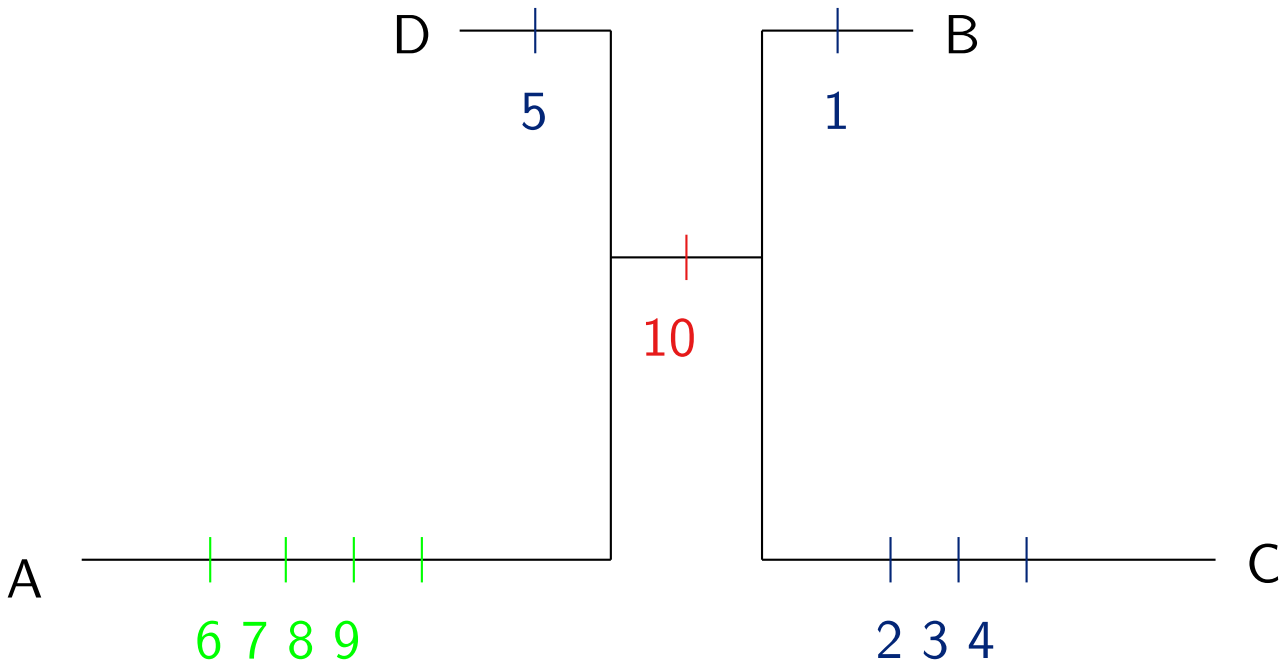
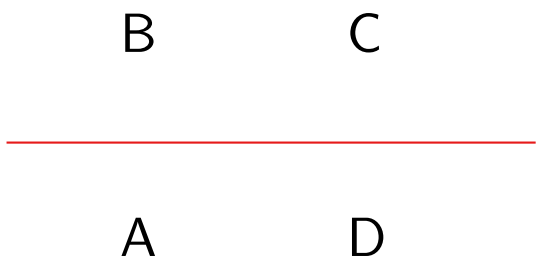
Taxon	Character #									
	1	2	3	4	5	6	7	8	9	10
A	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1
D	0	0	0	0	1	1	1	1	1	0


B
1
10
C
6 7 8 9
2 3 4
D
5
A
B C
D
A

Interestingly, without polarization Hennig's method can infer unrooted trees. We can get the tree topology, but be unable to tell paraphyletic from monophyletic groups.

The outgroup method amounts to inferring an unrooted tree and then rooting the tree on the branch that leads to an outgroup.


D
B
5
1
10
A
C
6 7 8 9
2 3 4


B
C
A
D

Inadequacy of logic

Unfortunately, though Hennigian logic is valid we quickly find that we do not have a reliable method of generating accurate homology statements.

The logic is valid, but we don't know that the premises are true.

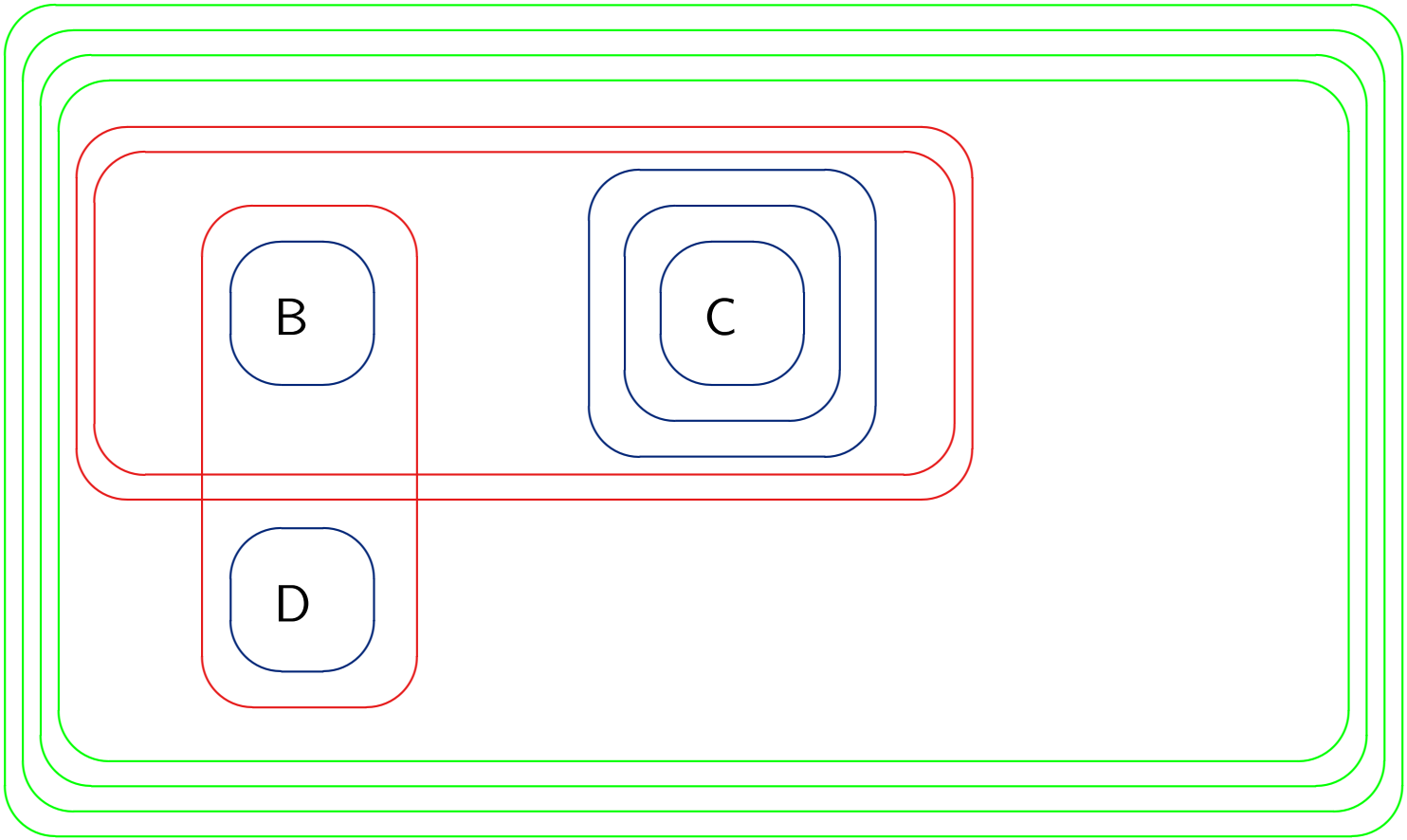
In fact, we almost always find that it is impossible for all of our premises to be true.

Character conflict

Homo sapiens	AGTTCAAGT
Rana catesbiana	AATTCAAGT
Drosophila melanogaster	AGTTCAAGC
C. elegans	AATTCAAGC

The red character implies that either (*Homo* + *Drosophila*) is a group (if G is derived) and/or (*Rana* + *C. elegans*) is a group.
The green character implies that either (*Homo* + *Rana*) is a group (if T is derived) and/or (*Drosophila* + *C. elegans*) is a group.
The green and red character cannot both be correct.

Taxon	Character #											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	1	1	1	1	1	1
C	0	1	1	1	0	1	1	1	1	1	1	0
D	0	0	0	0	1	1	1	1	1	0	0	1



A

Character conflict

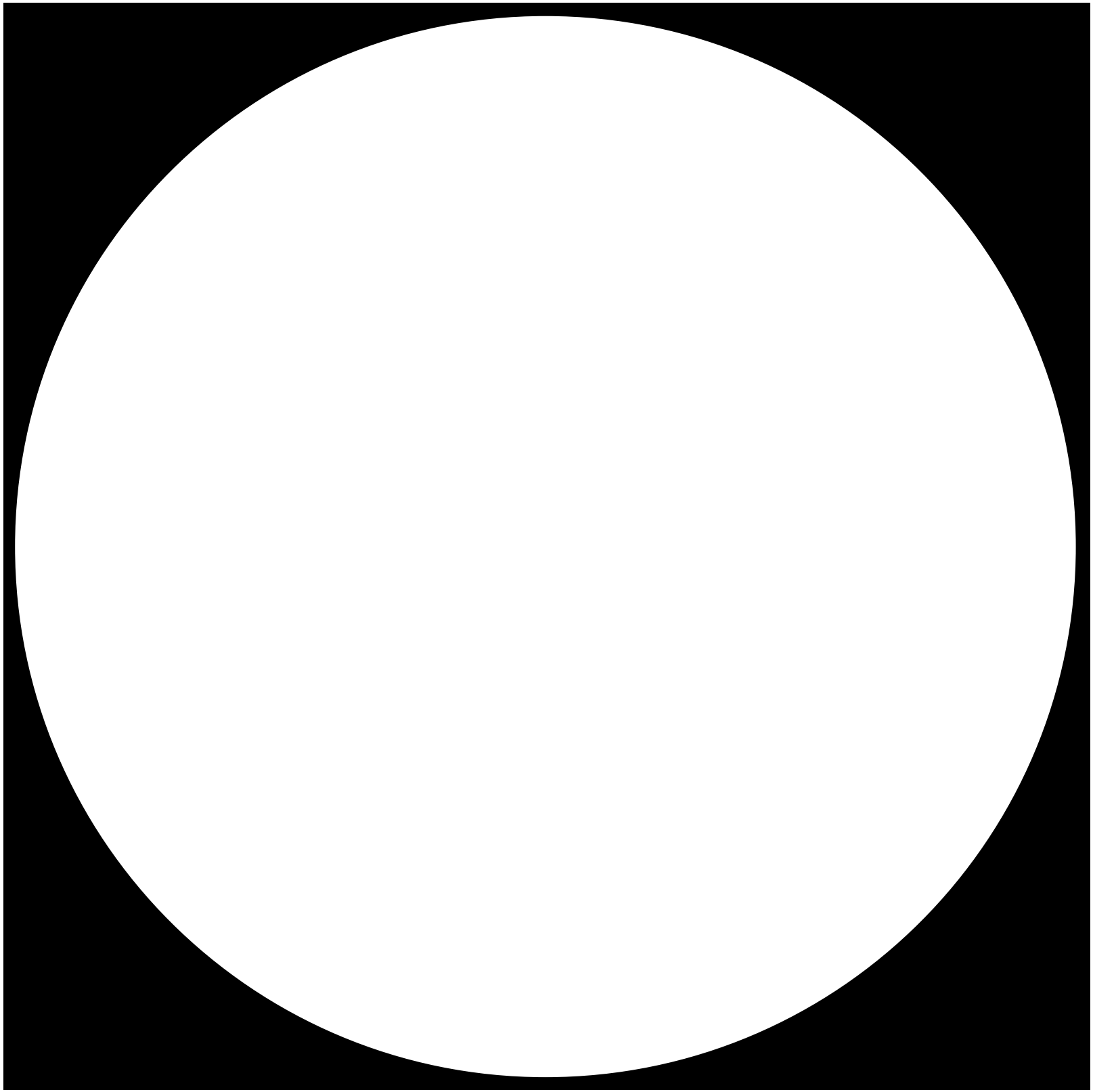
Two characters are compatible if they can both be mapped on the same tree so that all of the character states displayed could be homologous.

Incompatible characters are evidence of *homoplasy* in the data

Homoplasy literally means the "same change" has occurred more than once in the evolutionary history of the group.

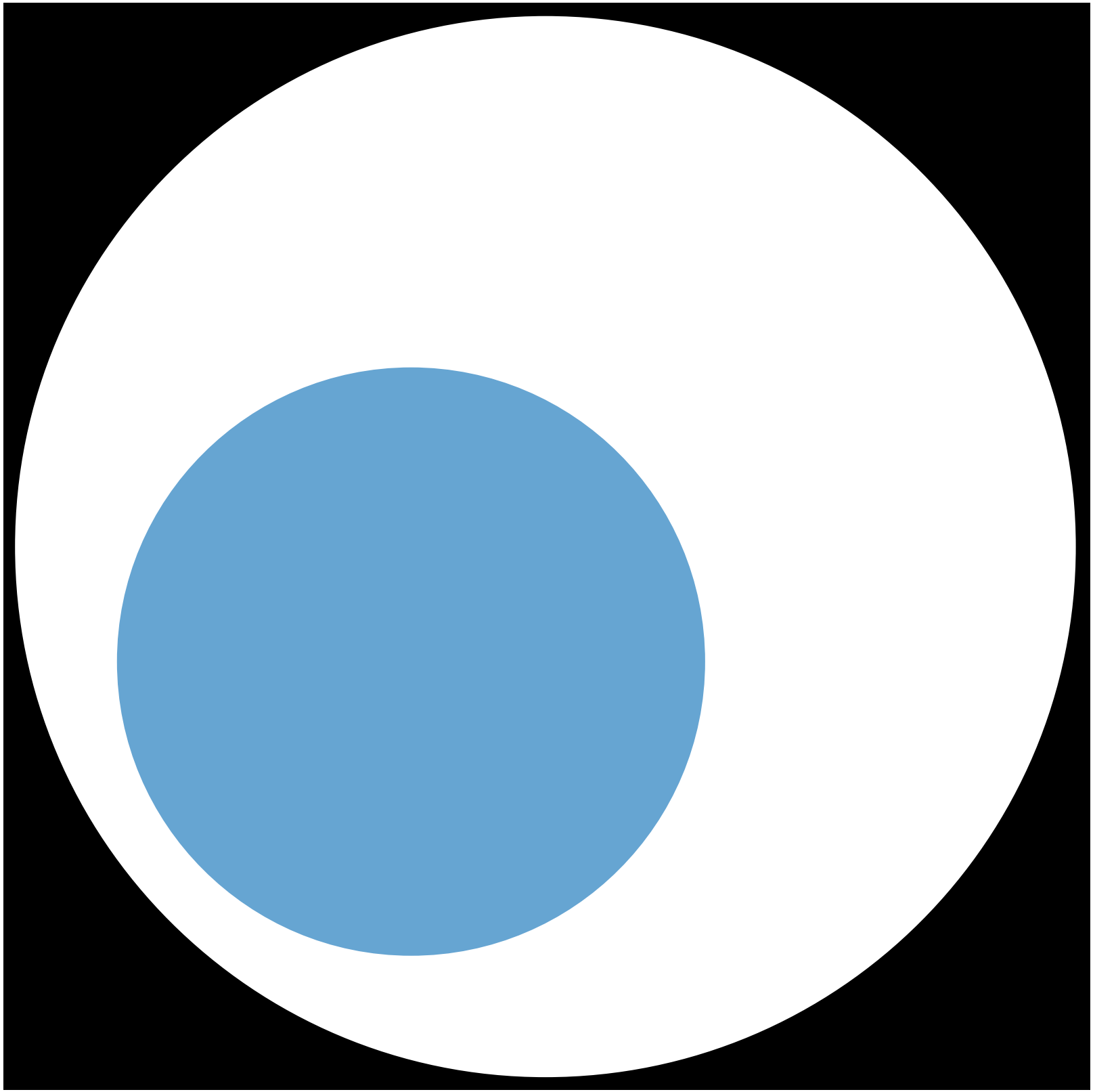
The presence of homoplasy undermines Hennigian analyses.

white = space
of all possible
matrices



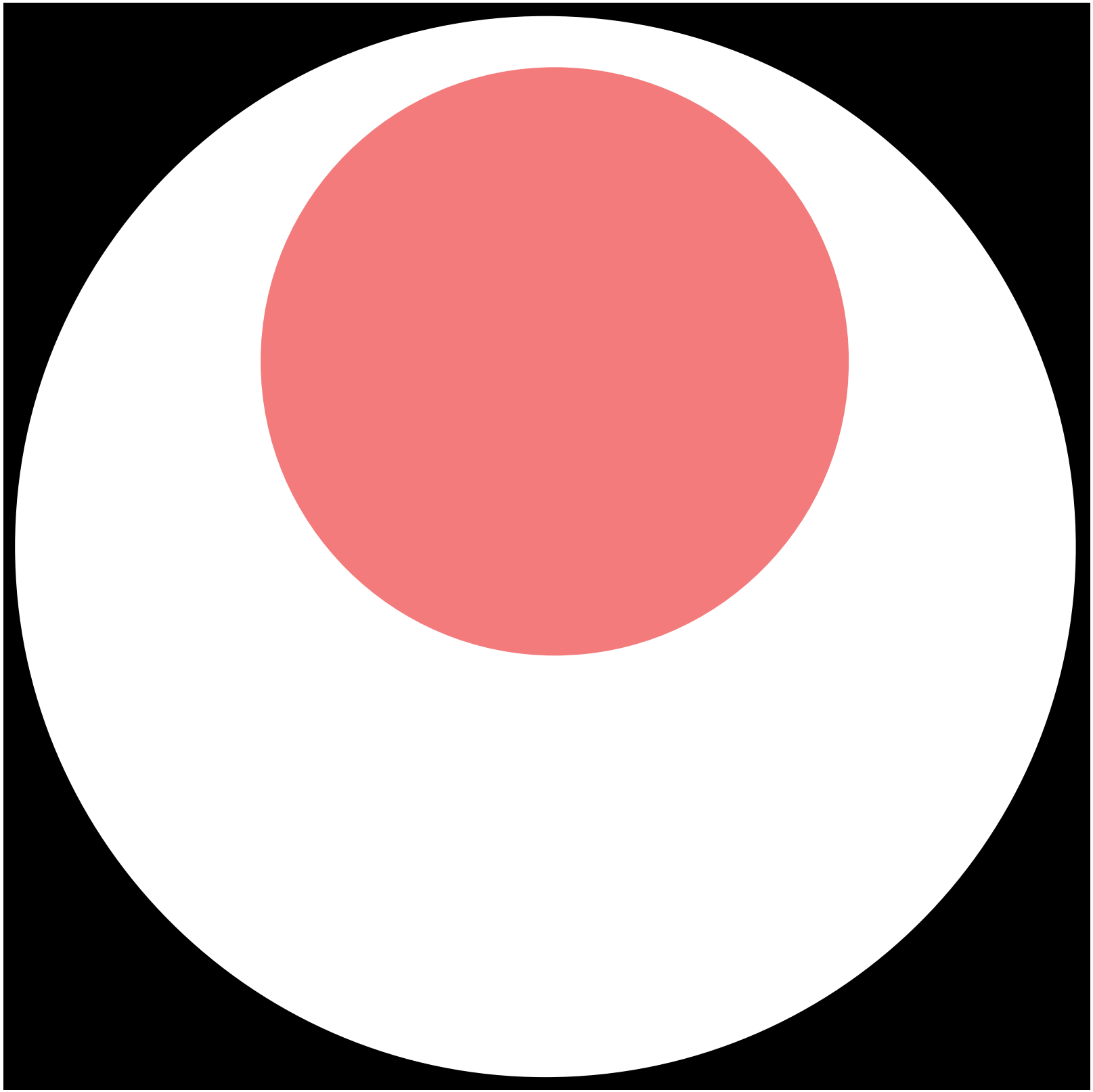
blue = space of matrices with the pattern:

```
A B C D
- * * -
```



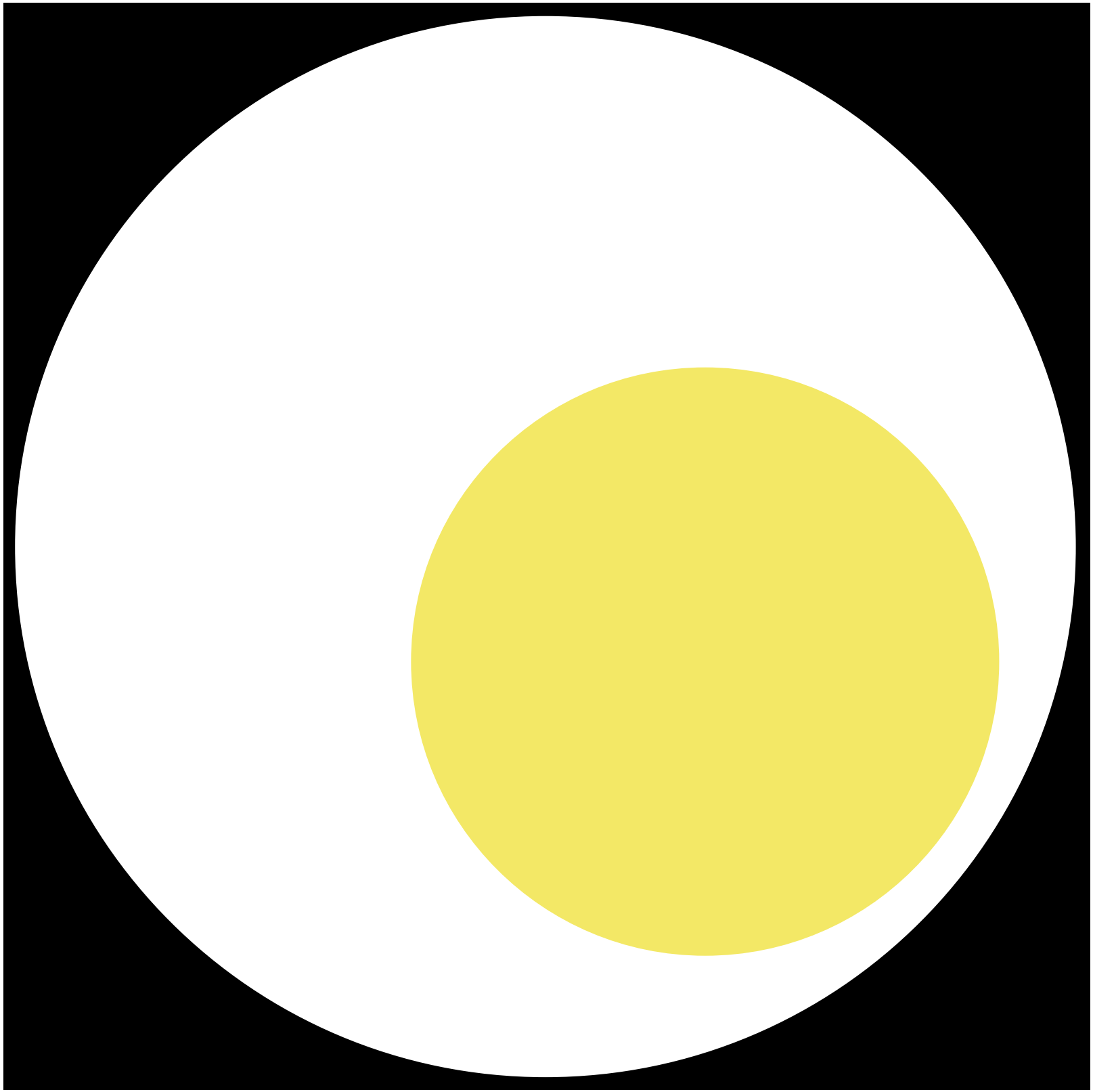
red = space
of matrices with
the pattern:

```
A B C D
- * - *
```



yellow = space of matrices with the pattern:

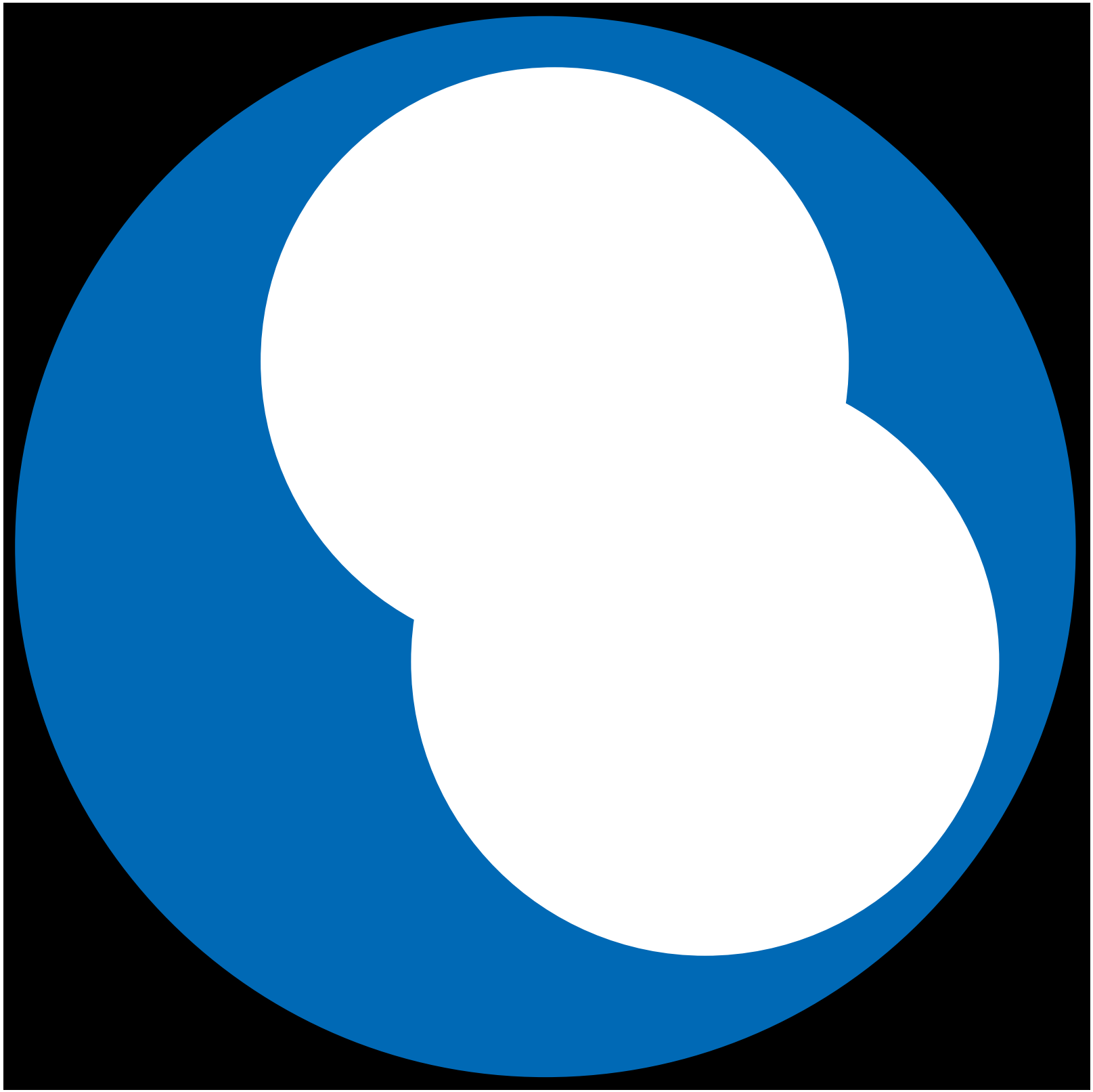
```
A B C D
- - * *
```



all eight
categories of
matrices

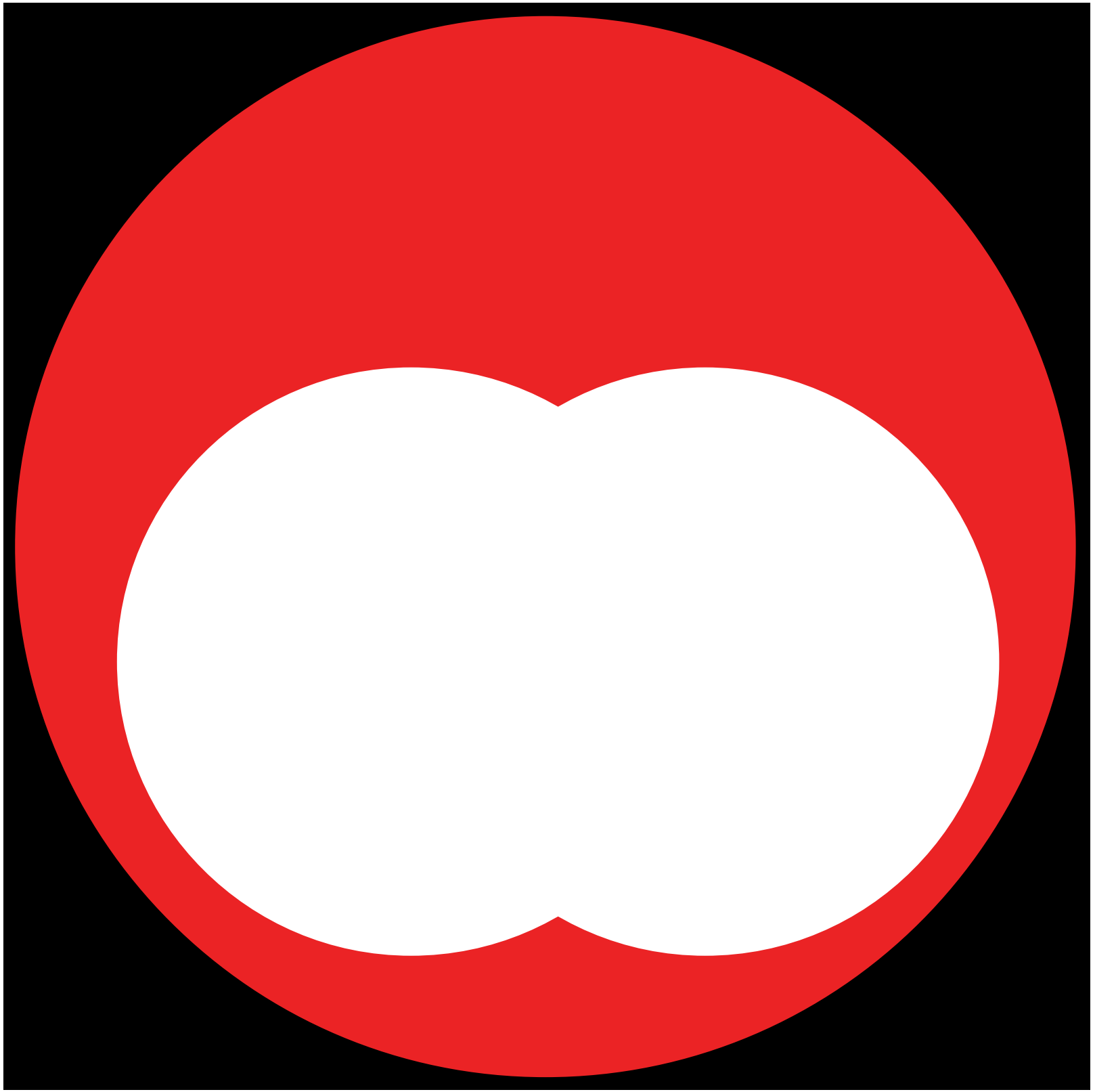
blue = space of matrices compatible with tree:

`(A,(B,C),D)`



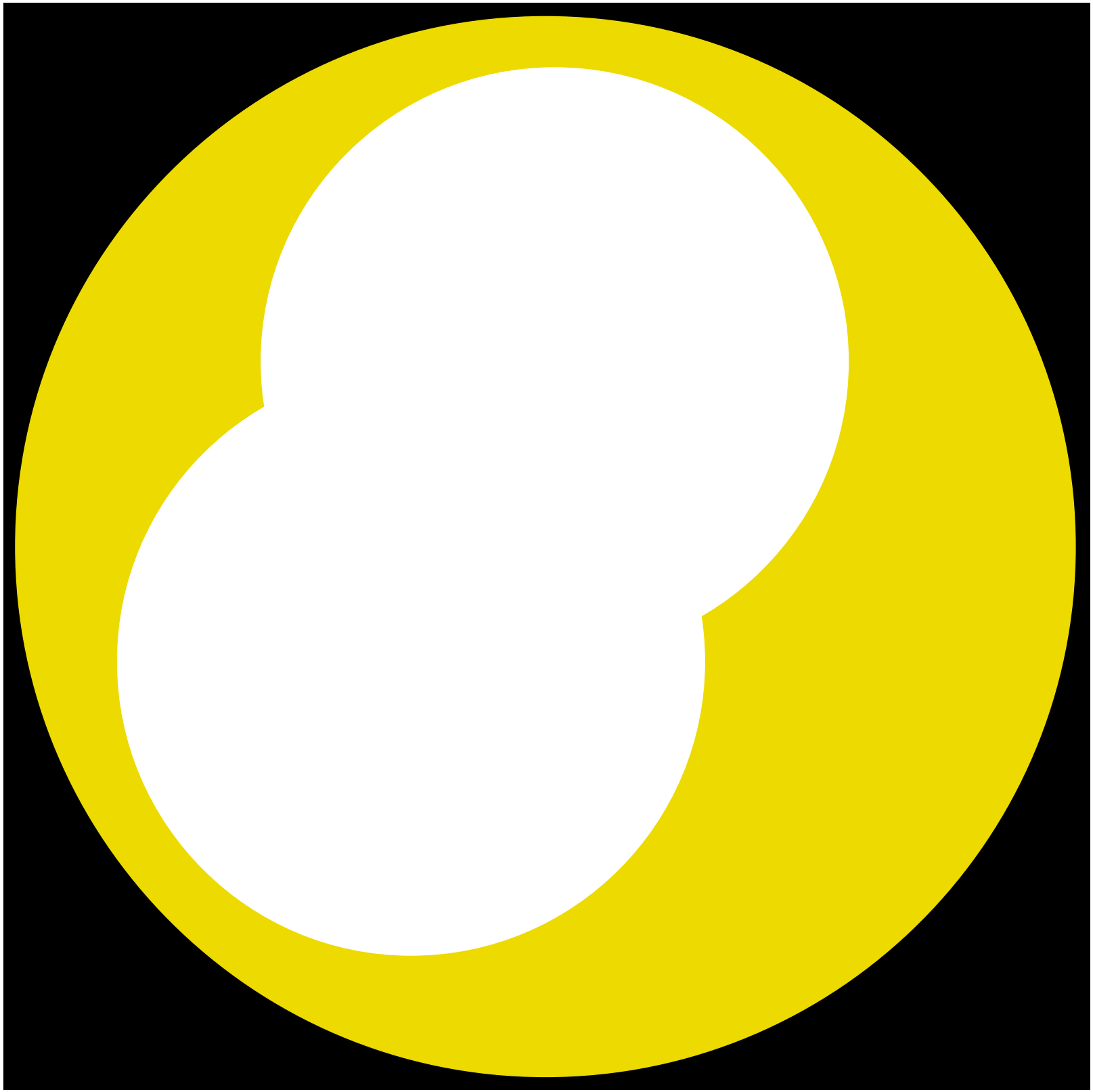
blue = space
of matrices
compatible
with tree:

```
(A,C,(B,D))
```

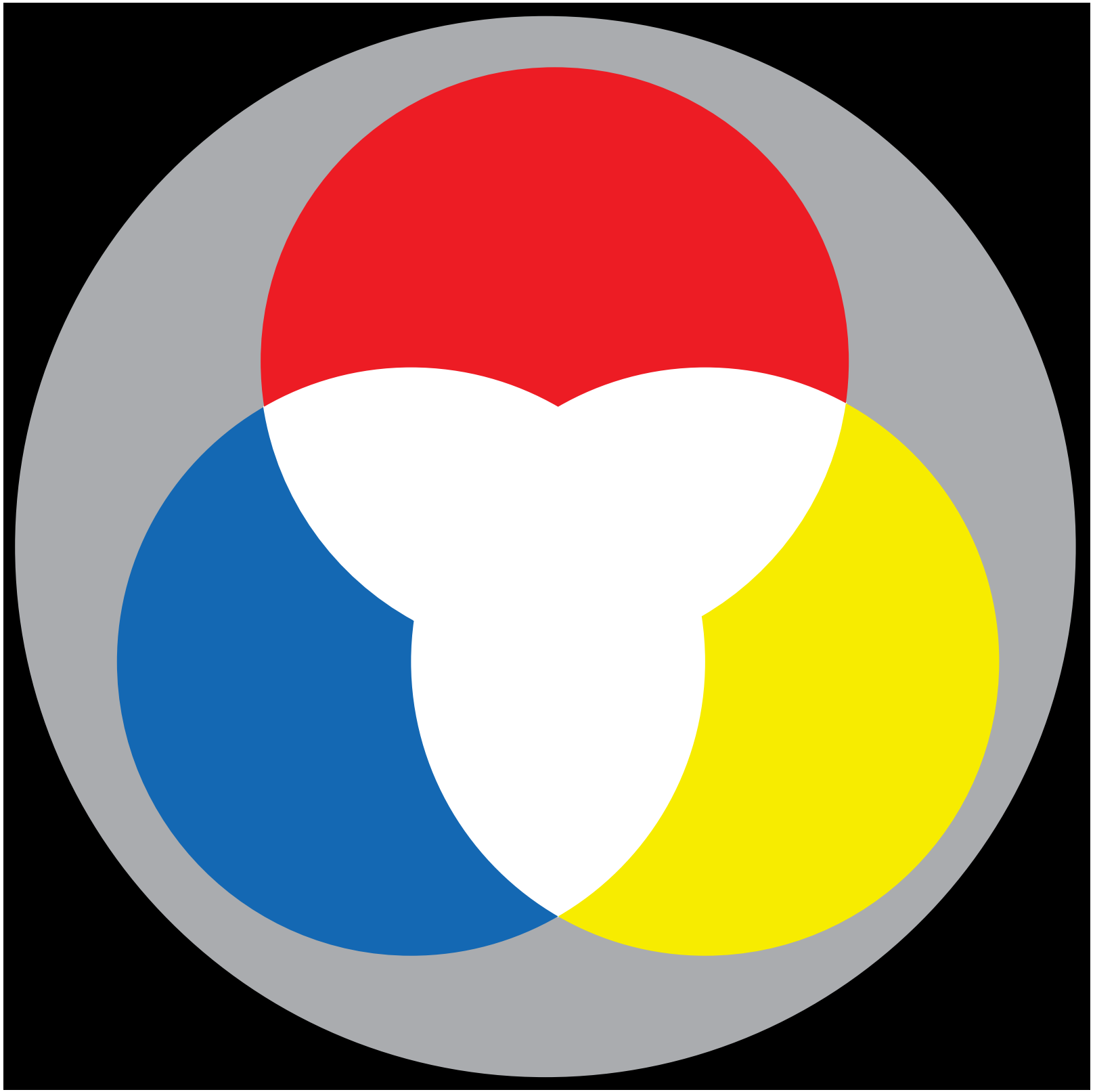


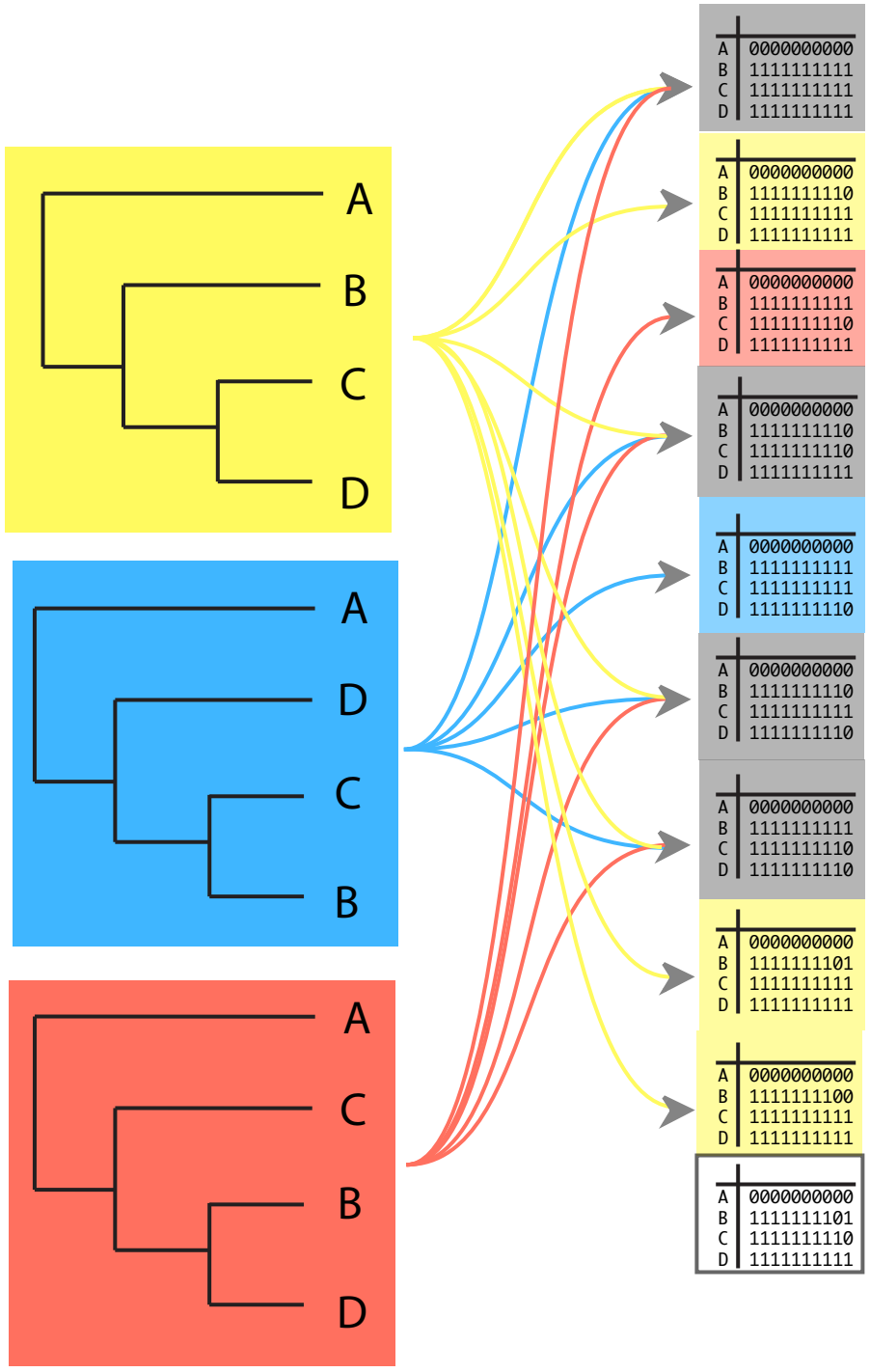
blue = space
of matrices
compatible
with tree:

```
(A,B,(C,D))
```



Hennigian:
grey = any tree
blue = B+C
red = B+D
yellow = C+D
white = no tree
(conflicting
characters)




A
B
C
D
A
D
C
B
A
C
B
D
A 00000000000
B 1111111111
C 1111111111
D 1111111111
A 00000000000
B 1111111110
C 1111111111
D 1111111111
A 00000000000
B 1111111111
C 1111111110
D 1111111111
A 00000000000
B 1111111110
C 1111111110
D 1111111111
A 00000000000
B 1111111111
C 1111111111
D 1111111110
A 00000000000
B 1111111110
C 1111111111
D 1111111110
A 00000000000
B 1111111111
C 1111111110
D 1111111110
A 00000000000
B 1111111101
C 1111111111
D 1111111111
A 00000000000
B 1111111100
C 1111111111
D 1111111111
A 00000000000
B 1111111101
C 1111111110
D 1111111111

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- can we detect character conflict?
- is there a logic-based solution to the problem of character conflict?

Detecting character conflict in binary characters

Consider the four possible combinations of states in a two-character matrix.

The characters are incompatible *iff* (when you look across all taxa) you see all four state combinations.

		Char 1	
		0	1
Char 2	0	×	×
	1	×	×

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

- Can we detect character conflict? Yes

- Is there a logic-based solution to the problem of character conflict?
  - recoding characters?
  - "reciprocal illumination"?

What can we do if our data end up in the white (character conflict) or grey (uninformative characters only) zone?

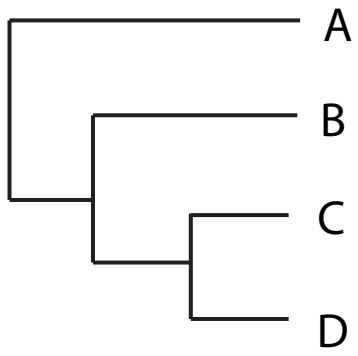
- Can we detect character conflict? Yes
- Is there a logic-based solution to the problem of character conflict? No, nothing purely based on logic (and the suggestions for culling data to make matrices suitable for logical inference can lead to unsatisfyingly subjecive analyses).
- What can we do?

We must have an "error model"

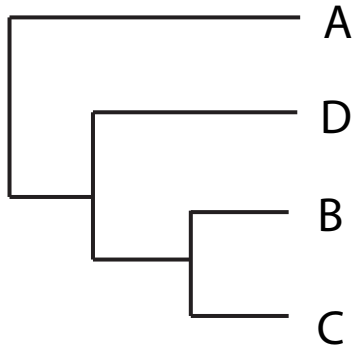
Statistical inference

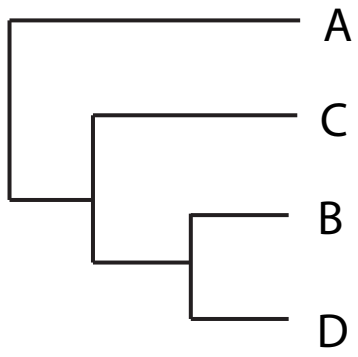
There are many ways to derive estimators, we are going to talk about maximum likelihood estimation:

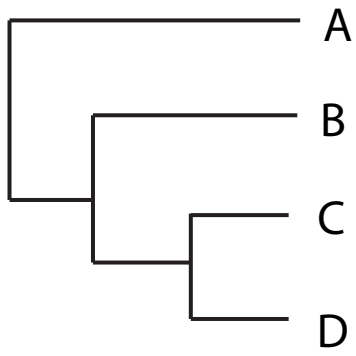
$$
\begin{aligned}
\theta &\in \Theta \\
X &\in \mathcal{X} \\
x &\sim \Pr(X = x|\theta) \\
\mathcal{L}(\theta) &= \Pr(X = x|\theta) \\
\hat{\theta} &= \arg\max \mathcal{L}(\theta)
\end{aligned}
$$



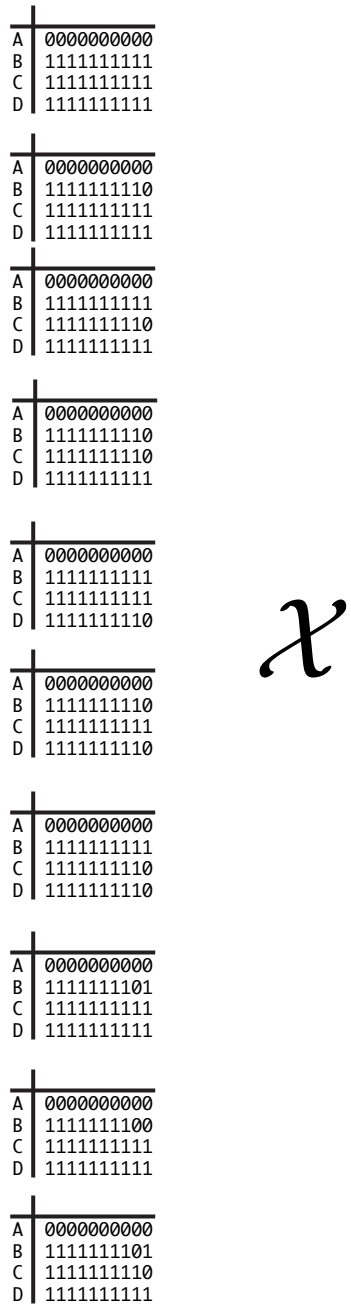
Θ





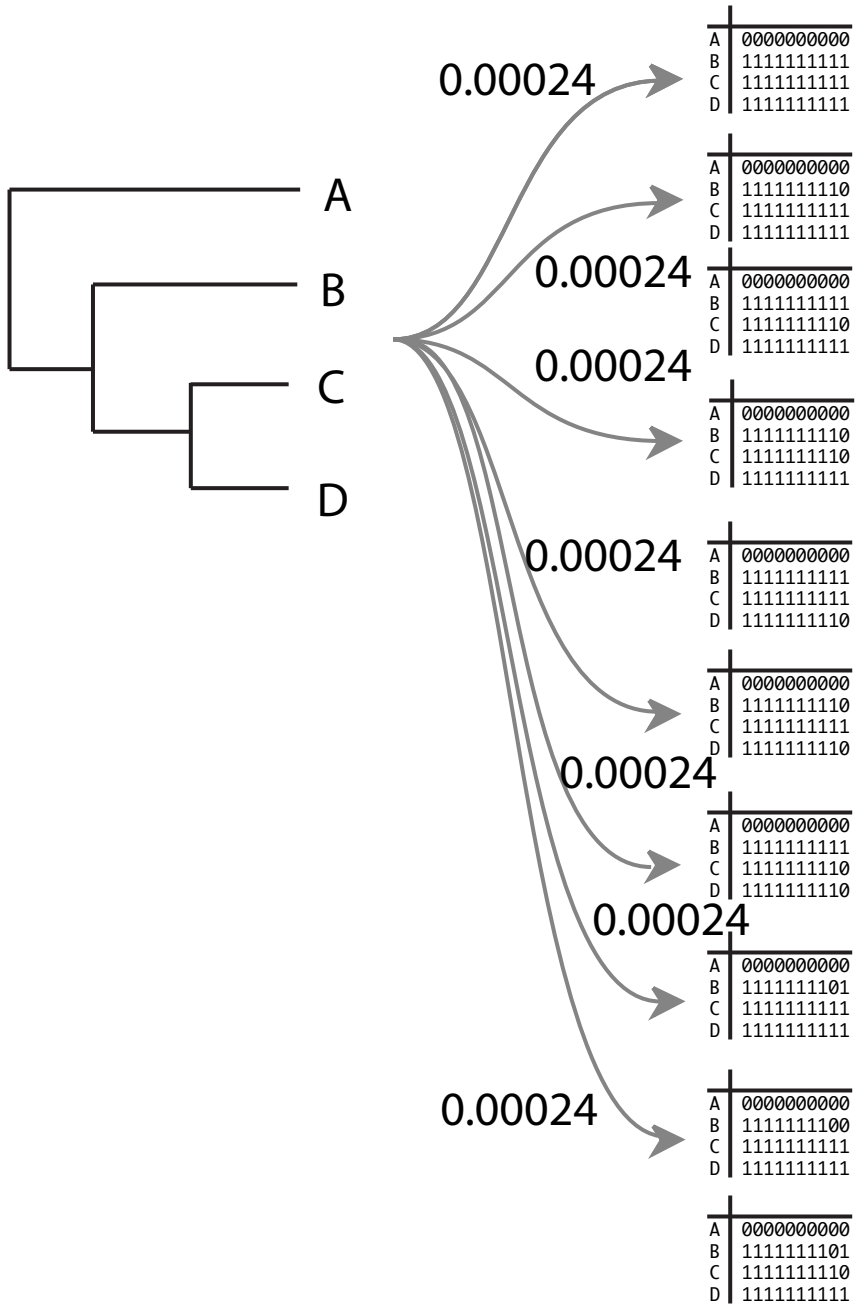


$$\theta \in \Theta$$

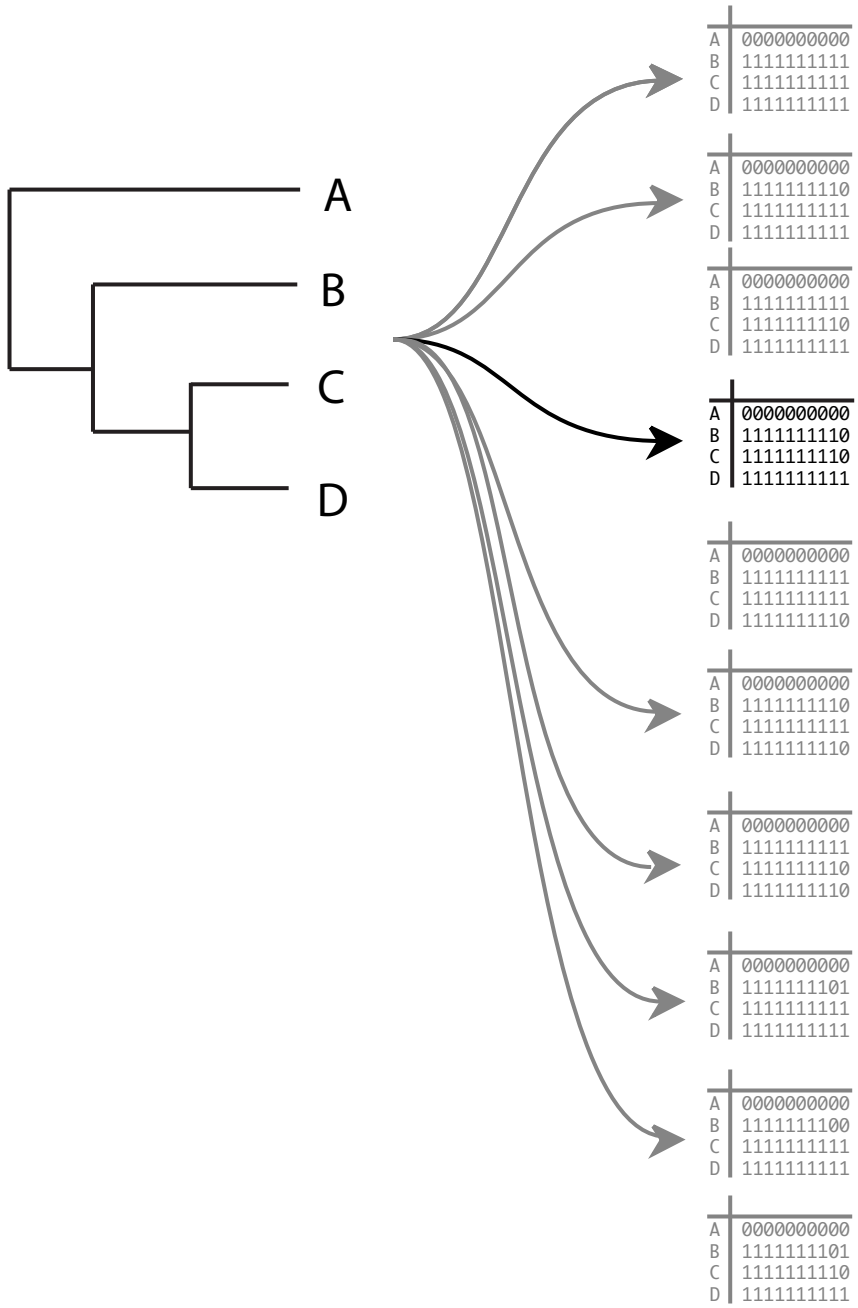

A 0000000000
B 1111111111
C 1111111111
D 1111111111
A 0000000000
B 1111111110
C 1111111111
D 1111111111
A 0000000000
B 1111111111
C 1111111110
D 1111111111
A 0000000000
B 1111111110
C 1111111110
D 1111111111
A 0000000000
B 1111111111
C 1111111111
D 1111111110
A 0000000000
B 1111111110
C 1111111111
D 1111111110
A 0000000000
B 1111111111
C 1111111110
D 1111111110
A 0000000000
B 1111111101
C 1111111111
D 1111111111
A 0000000000
B 1111111100
C 1111111111
D 1111111111
A 0000000000
B 1111111101
C 1111111110
D 1111111111

x

•••


A
B
C
D
0.00024
A 0000000000
B 1111111111
C 1111111111
D 1111111111
0.00024
A 0000000000
B 1111111110
C 1111111111
D 1111111111
0.00024
A 0000000000
B 1111111111
C 1111111110
D 1111111111
0.00024
A 0000000000
B 1111111110
C 1111111110
D 1111111111
0.00024
A 0000000000
B 1111111111
C 1111111111
D 1111111110
0.00024
A 0000000000
B 1111111110
C 1111111111
D 1111111110
0.00024
A 0000000000
B 1111111111
C 1111111110
D 1111111110
0.00024
A 0000000000
B 1111111101
C 1111111111
D 1111111111
0.00024
A 0000000000
B 1111111100
C 1111111111
D 1111111111
A 0000000000
B 1111111101
C 1111111110
D 1111111111

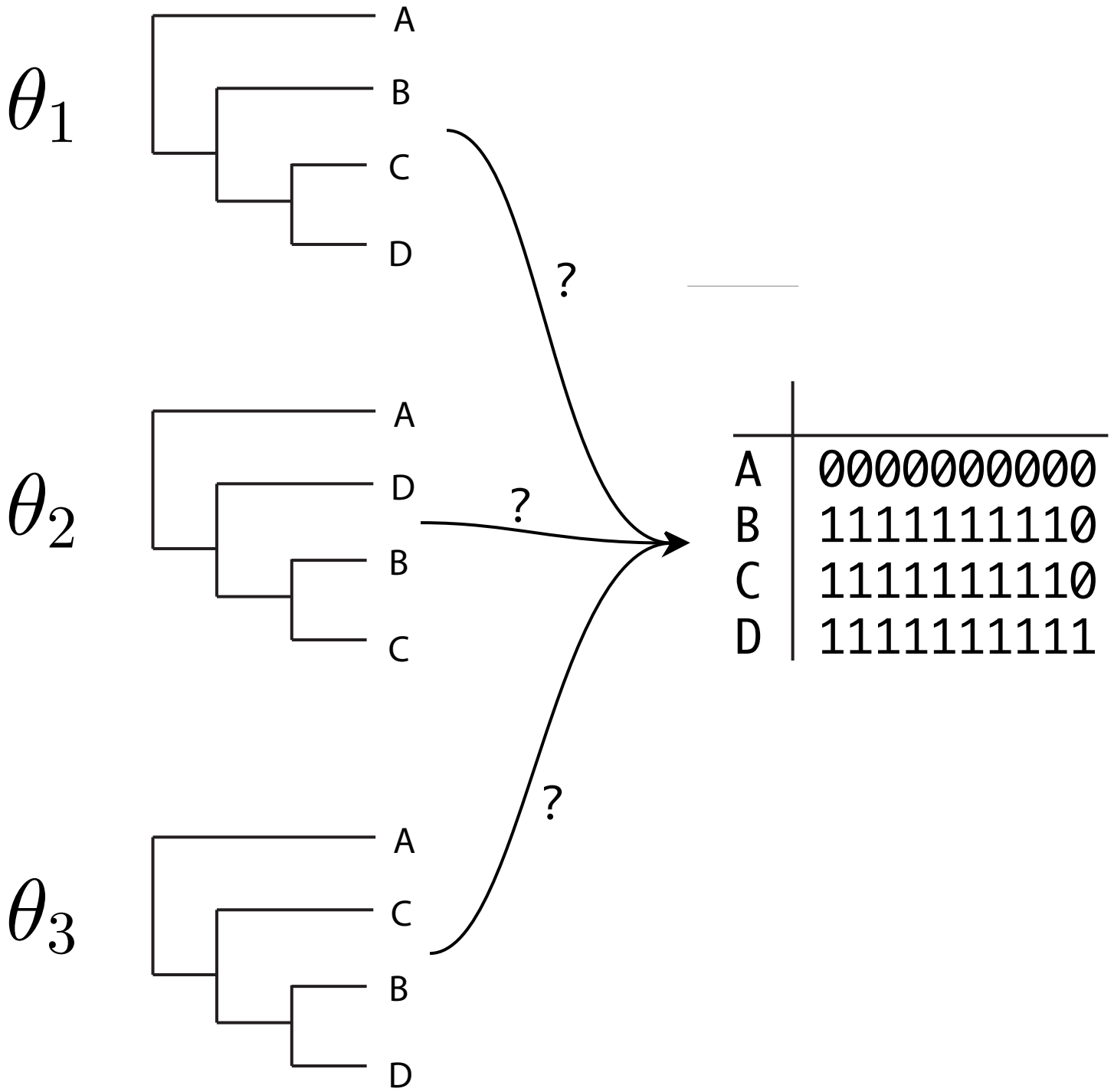
$$\Pr(X = x|\theta)$$

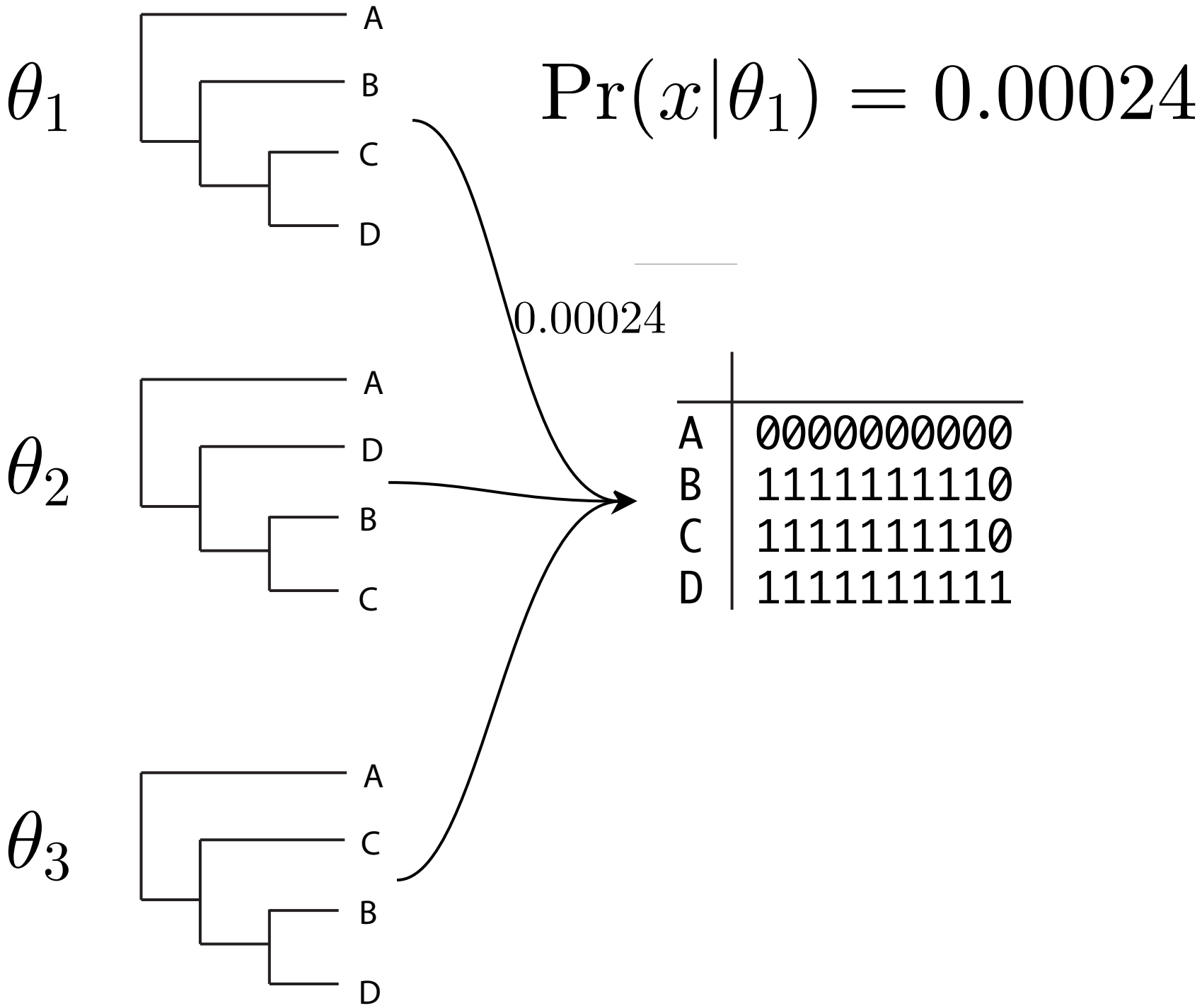


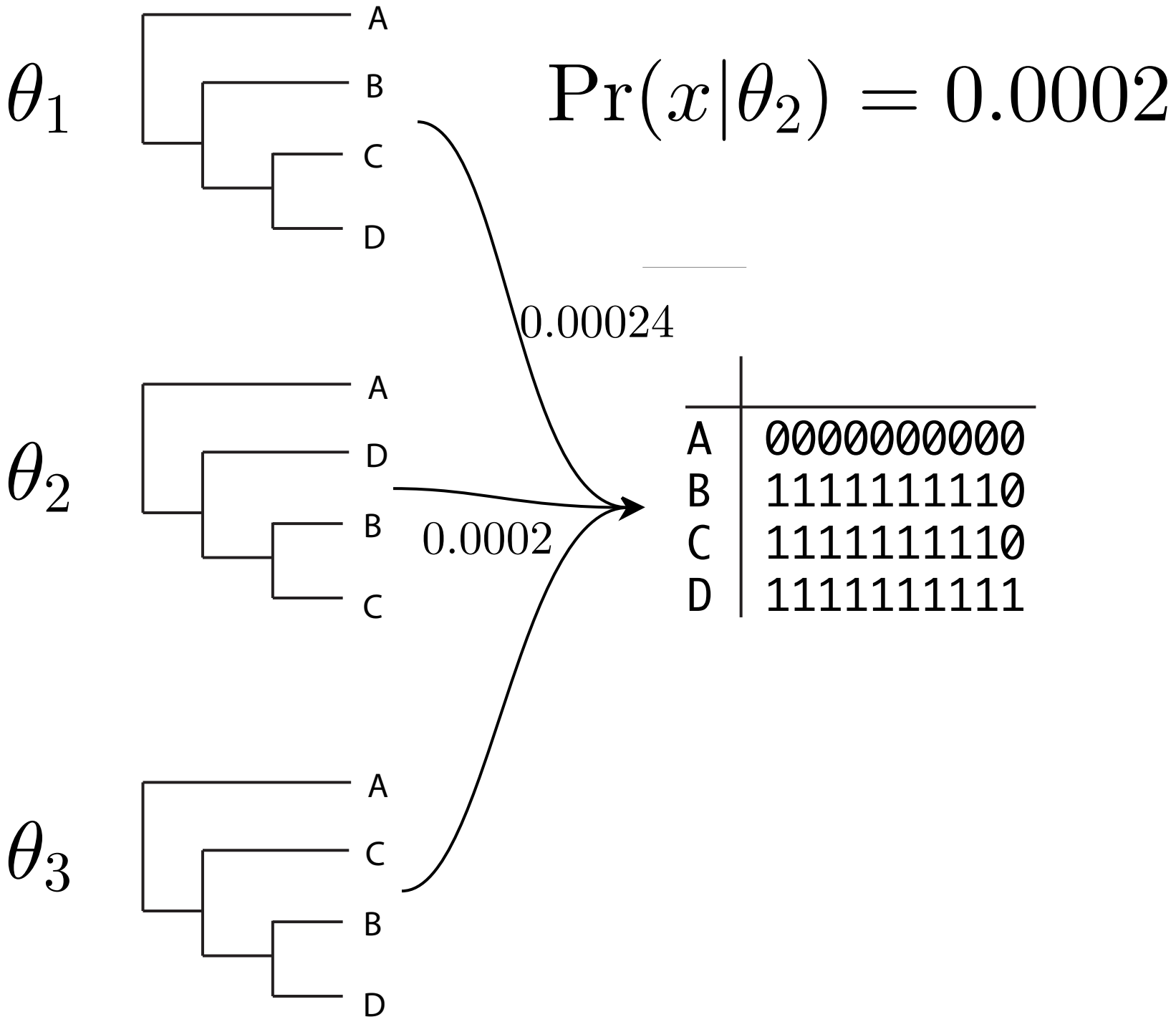
$$x \sim \Pr(X = x|\theta)$$

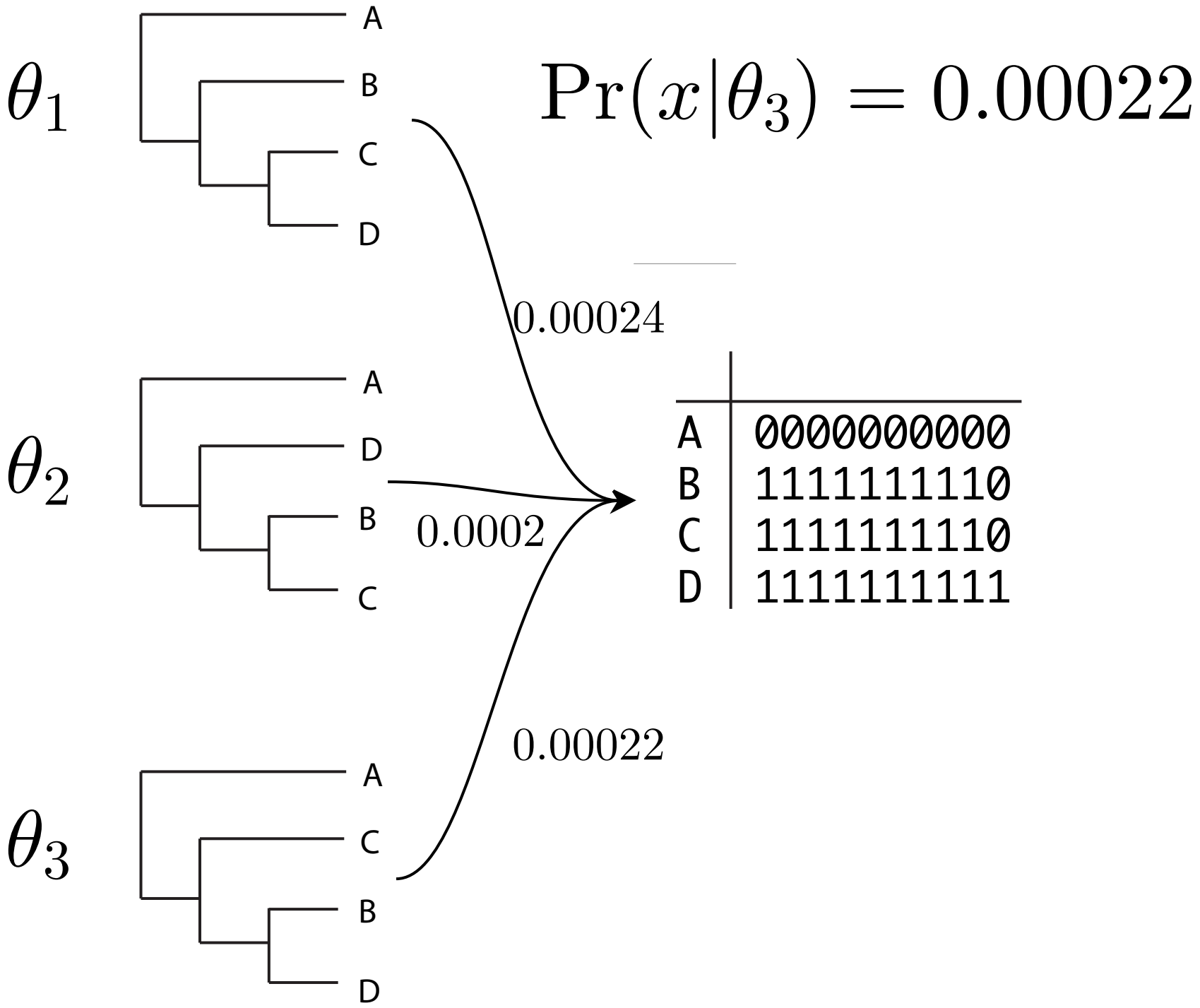
x represents

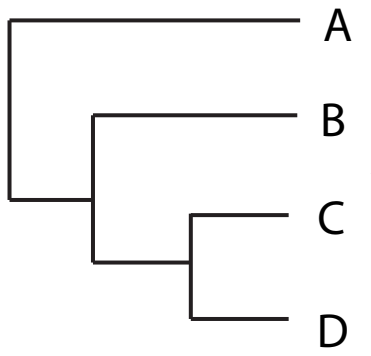
A	0000000000
B	1111111110
C	1111111110
D	1111111111


θ_1
A
B
C
D
?
θ_2
A
D
B
C
?
θ_3
A
C
B
D
?
A 0000000000
B 1111111110
C 1111111110
D 1111111111

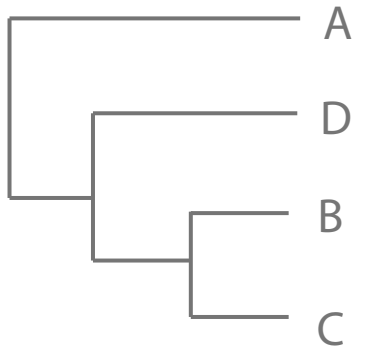

θ_1
A
B
C
D
$\Pr(x|\theta_1) = 0.00024$
0.00024
θ_2
A
D
B
C
A 0000000000
B 1111111110
C 1111111110
D 1111111111
θ_3
A
C
B
D


θ_1
A
B
C
D
θ_2
A
D
B
C
θ_3
A
C
B
D
$\Pr(x|\theta_2) = 0.0002$
0.00024
0.0002
A 0000000000
B 1111111110
C 1111111110
D 1111111111

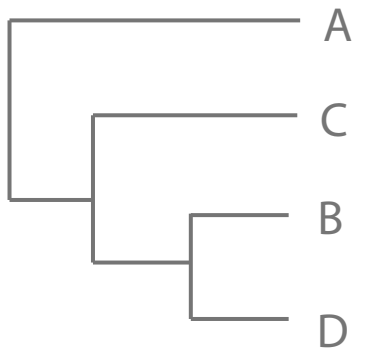

θ_1
A
B
C
D
θ_2
A
D
B
C
θ_3
A
C
B
D
0.00024
0.0002
0.00022
$\Pr(x|\theta_3) = 0.00022$
A 0000000000
B 1111111110
C 1111111110
D 1111111111



$$\hat{\theta} = \arg\max \mathcal{L}(\theta)$$



A	0000000000
B	1111111110
C	1111111110
D	1111111111



ML Estimation

- Flexible form of inference
- Requires a model: $\Pr(X = x|\theta)$

Under mild conditions, ML estimation is asymptotically:

- not very biased,
- efficient

How can we come up with a model?