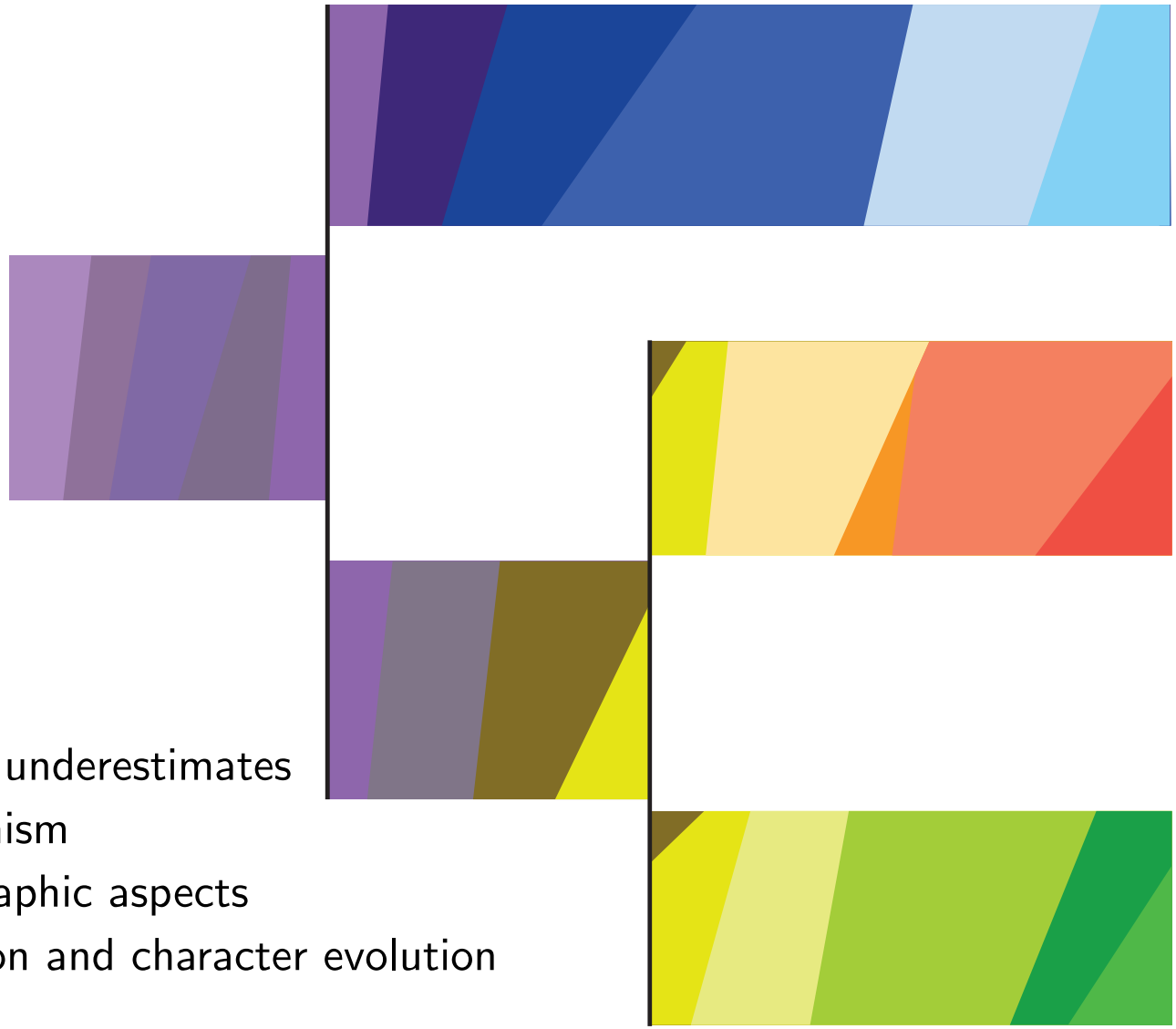
# Homology and inference

Emily Jane McTavish

Life and Environmental Sciences
University of California, Merced
ejmctavish@ucmerced.edu, twitter:snacktavish

(With thanks to Mark Holder for slides)

Phylogeny with complete genome + "phenome" as colors:

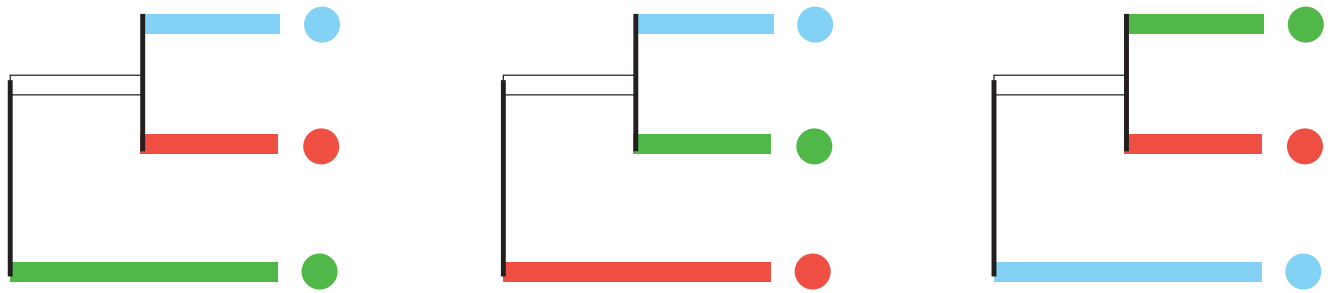


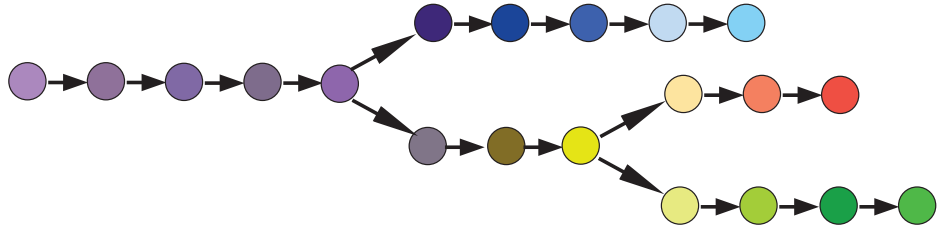
This figure:
dramatically underestimates
polymorphism
ignore geographic aspects
of speciation and character evolution

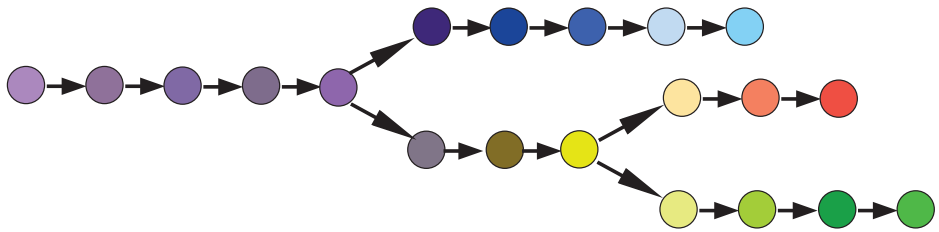
Extant species are just a thin slice of the phylogeny:

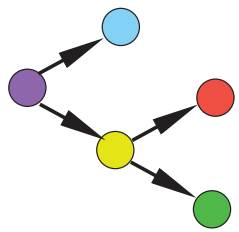
Our exemplar specimens are a subset of the current diversity:

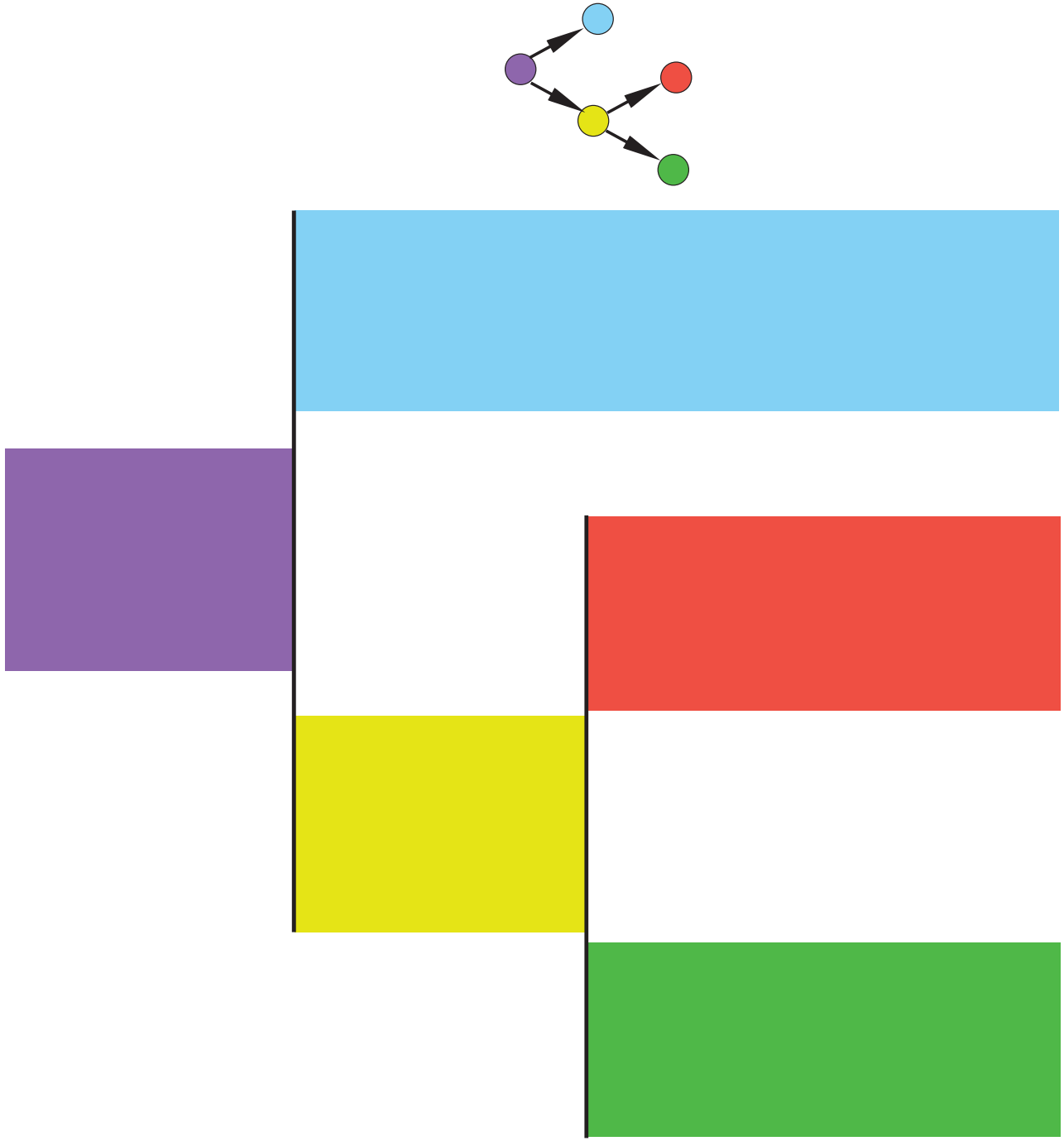
The phylogenetic inference problem:

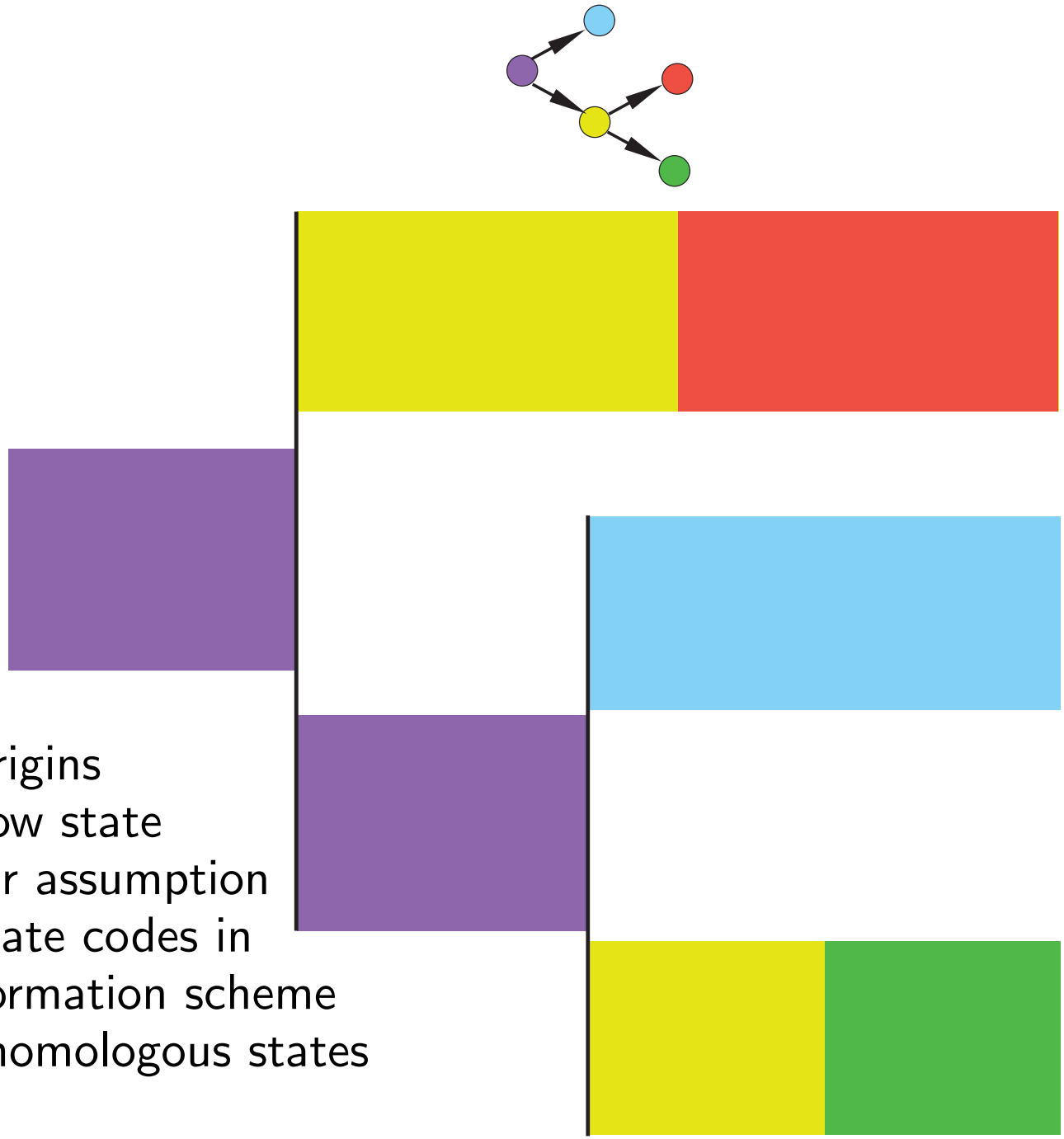












Multiple origins
of the yellow state
violates our assumption
that the state codes in
our transformation scheme
represent homologous states

Character matrices:

| | | Characters | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Taxa | *Homo sapiens* | 0.13 | A | A | rounded | 1 | 1610 - 1755 |
| | *Pan paniscus* | 0.34 | A | G | flat | 2 | 0621 - 0843 |
| | *Gorilla gorilla* | 0.46 | C | G | pointed | 1 | 795 - 1362 |

Characters (aka "transformation series") are the columns.
The values in the cells are character states (aka "characters").

| | | Characters | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Taxa | *Homo sapiens* | 0.13 | A | A | rounded | 1 | 1610 - 1755 |
| | *Pan paniscus* | 0.34 | A | G | flat | 2 | 0621 - 0843 |
| | *Gorilla gorilla* | 0.46 | C | G | pointed | 1 | 795 - 1362 |

Character coding:

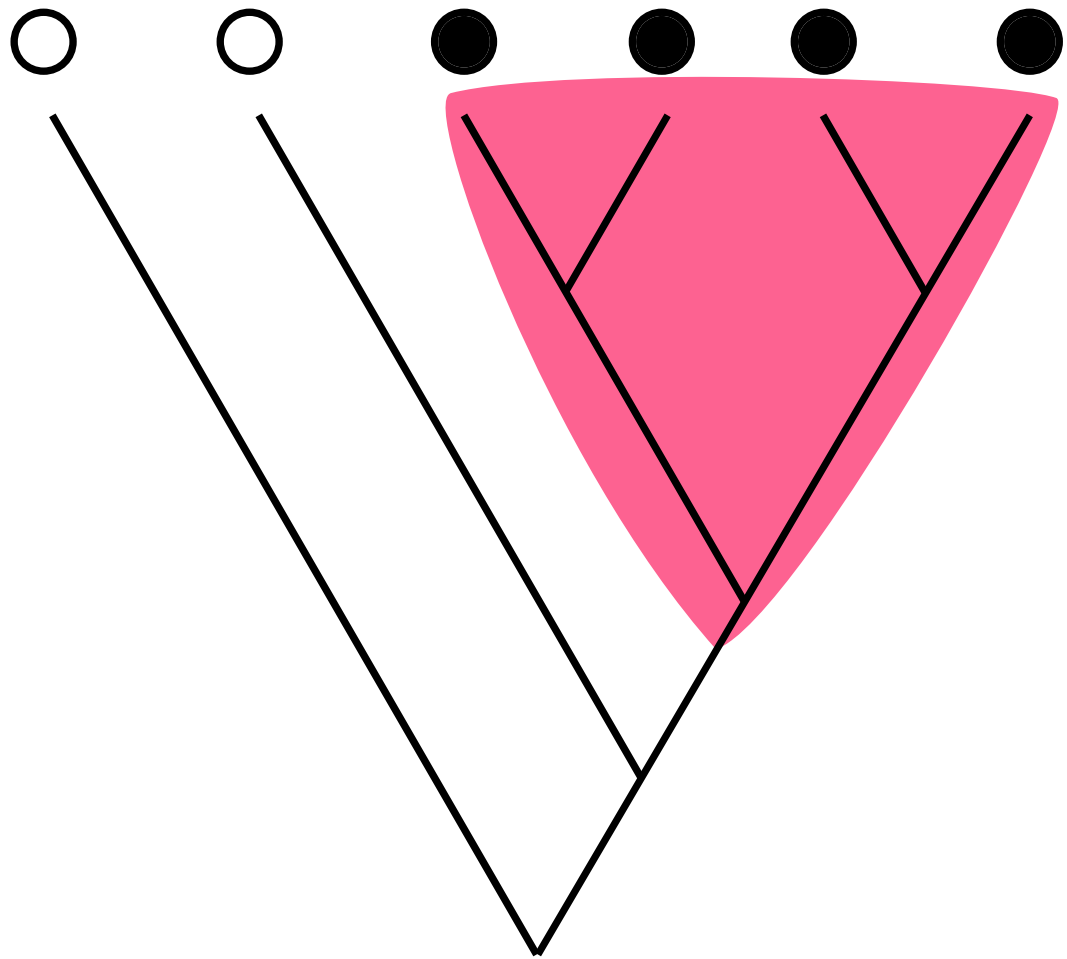
| | | Characters | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| Taxa | *Homo sapiens* | 0 | A | A | 0 | 1 | 4 |
| | *Pan paniscus* | 2 | A | G | 1 | 2 | 0,1 |
| | *Gorilla gorilla* | 3 | C | G | 2 | 1 | 1,2 |

The meaning of homology (**very roughly**):

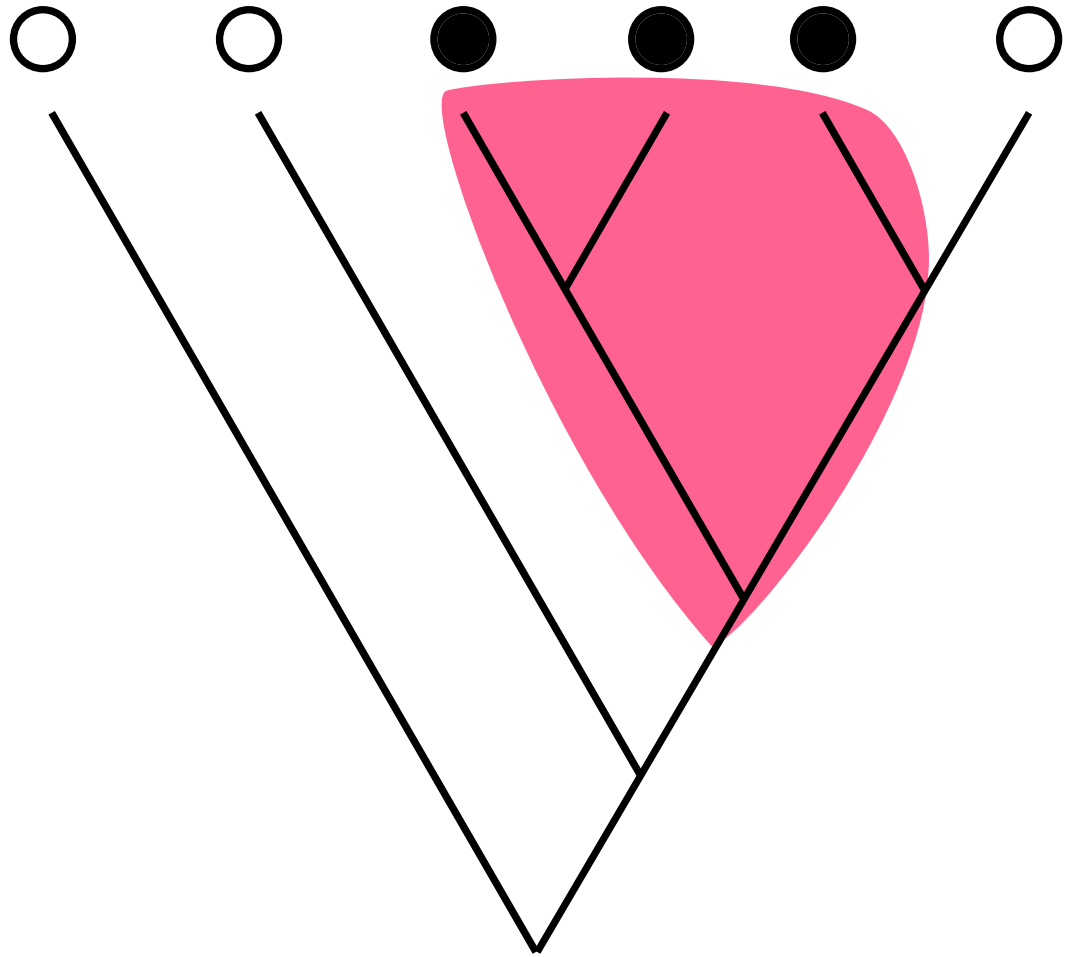
1. comparable (when applied to characters)
2. identical by descent (when applied to character states)

Ideally, each possible character state would arise once in the entire history of life on earth.

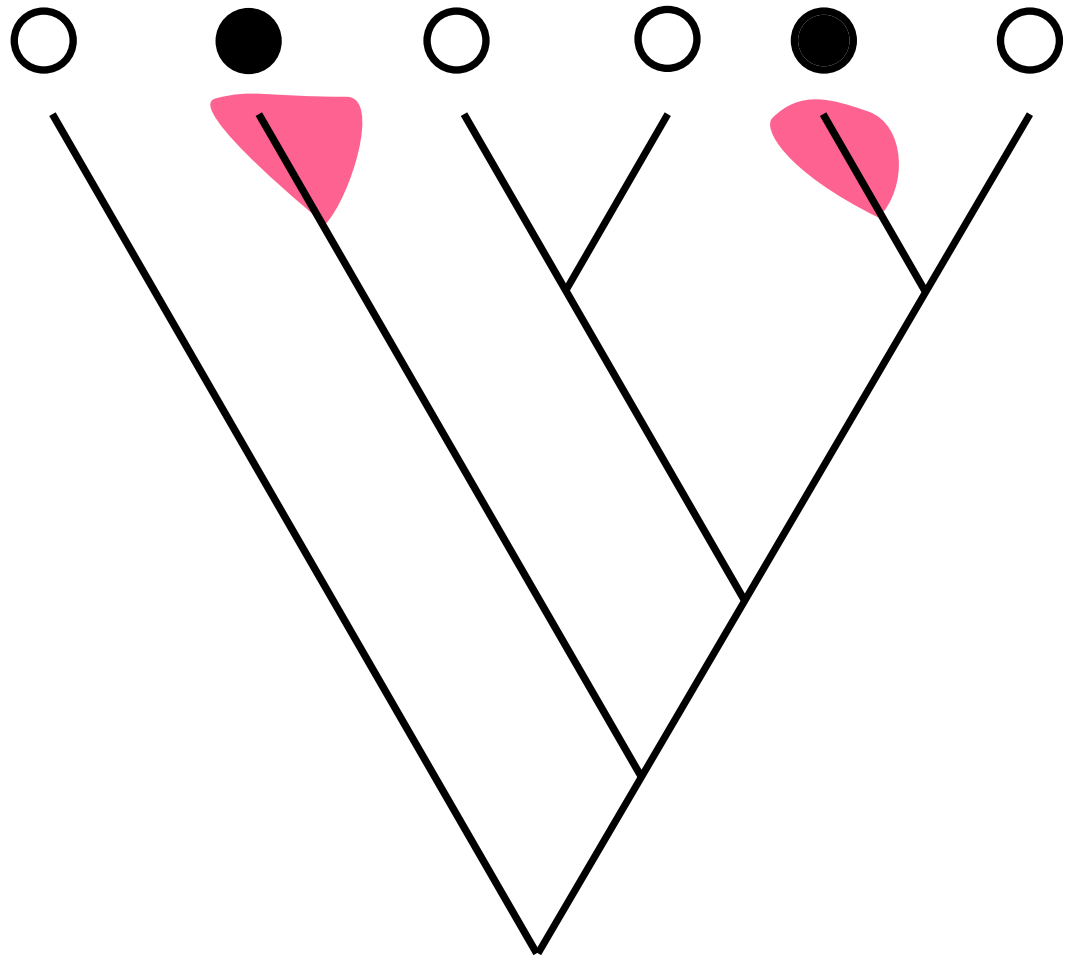
Instances of the filled character state are homologous
Instances of the hollow character state are homologous



Instances of the filled character state are homologous
Instances of the hollow character state are NOT homologous



Instances of the filled character state are NOT homologous
Instances of the hollow character state are homologous



## Inference

"deriving a conclusion based solely on what one already knows" [1]

- logical
- statistical

[1] definition from Wikipedia, so it must be correct!

## Logical Inference

Deductive reasoning:

1. start from premises
2. apply proper rules
3. arrive at statements that were not obviously contained in the premises.

If the rules are valid (logically sound) and the premises are true, then the conclusions are *guaranteed* to be true.

## Logical approach to phylogenetics

Premise: The following character matrix is correctly coded (character states are homologous in the strict sense):

| | 1 |
|---|---|
| taxon A | Z |
| taxon B | Y |
| taxon C | Y |

Is there a valid set of rules that will generate the tree as a conclusion?

## Logical approach to phylogenetics (cont)

Rule: Two taxa that share a character state must be more closely related to each other than either is to a taxon that displays a different state.

Is this a valid rule?

## Invalid rule

Here is an example in which we are confident that the homology statements are correct, but our rule implies two conflicting trees:

| | placenta | vertebra |
|---|---|---|
| *Homo sapiens* | Z | A |
| *Rana catesbiana* | Y | A |
| *Drosophila melanogaster* | Y | B |

Rule: Two taxa that share a character state must be more closely related to each other than either is to a taxon that displays a different state.(method suggested by Hennig)
*Is this a valid rule?*

## Hennigian logical analysis

The German entomologist Willi Hennig (in addition to providing strong arguments for phylogenetic classifications) clarified the logic of phylogenetic inference.

Hennig's correction to our rule: Two taxa that share a **derived** character state must be more closely related to each other than either is to a taxon that displays the **primitive** state.

## Hennig's logic is valid

Here we will use 0 for the primitive state, and 1 for the derived state.

| | placenta | vertebra |
|---|---|---|
| *Homo sapiens* | 1 | 1 |
| *Rana catesbiana* | 0 | 1 |
| *Drosophila melanogaster* | 0 | 0 |

Now the character "placenta" does not provide a grouping, but "vertebra" groups human and frog as sister taxa.

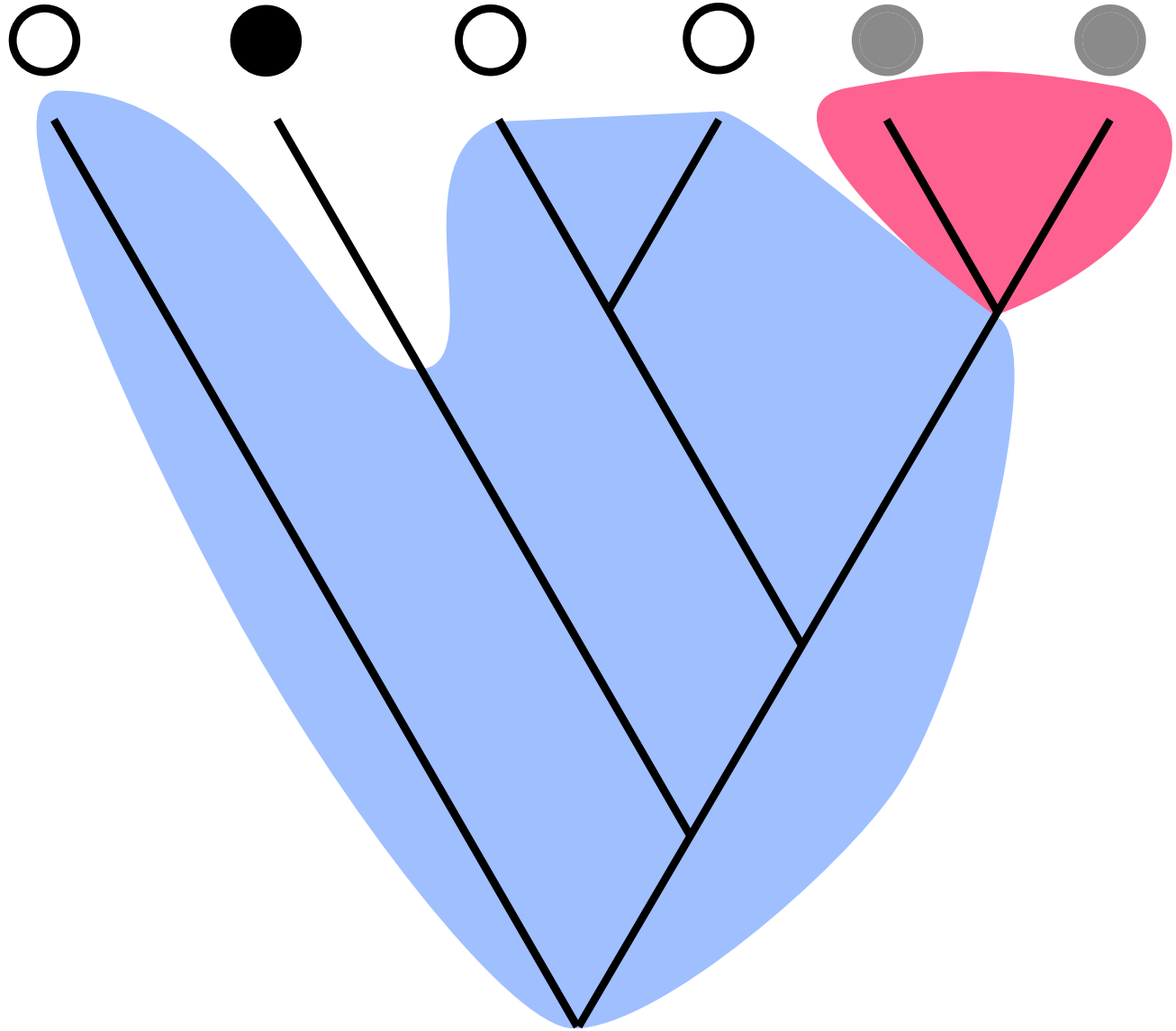
## Hennigian terminology

prefixes:

- “apo” - refers to the new or derived state
- “plesio” - refers to the primitive state
- “syn” or “sym” - used to indicate shared between taxa
- “aut” - used to indicate a state being unique to one taxon

## Hennigian rules

- synapomorphy - shared, derived states. Used to diagnose monophyletic groups.
- symplesiomorphy - shared, primitive states. Diagnose icky, unwanted paraphyletic groups.
- autapomorphy – a unique derived state. **No** evidence of phylogenetic relationships.
- constant characters – columns in a matrix with no variability between taxa. **No** evidence of phylogenetic relationships.



## Hennigian inference

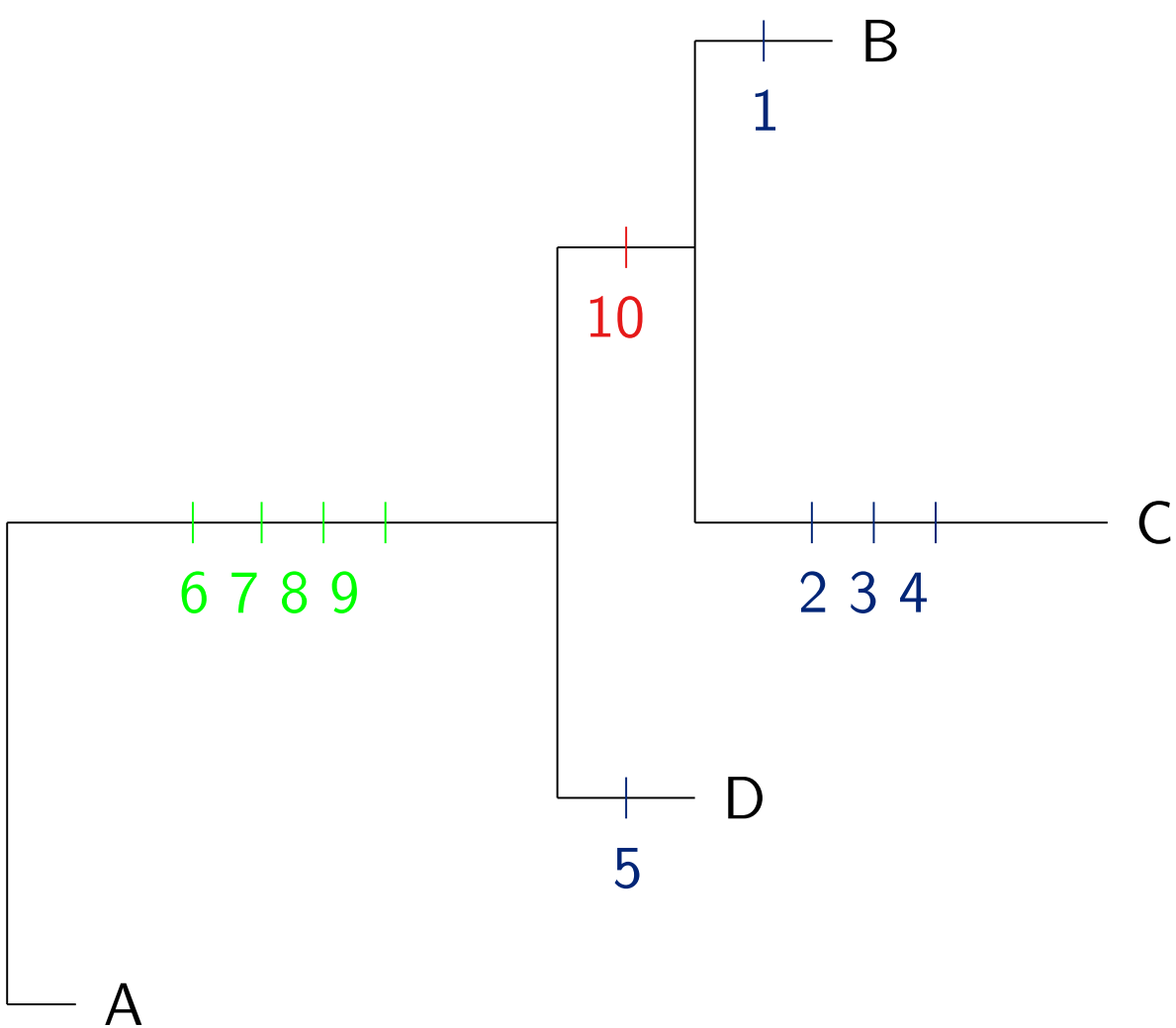
When we create a character matrix for Hennig's system, it is crucial that:

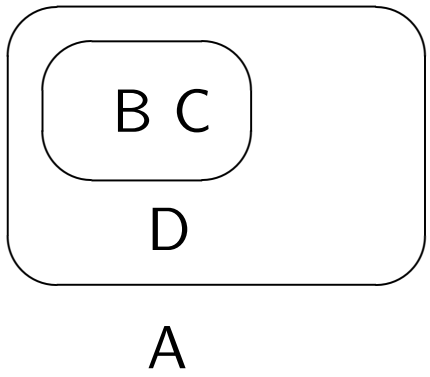
- traits assigned the same state represent homologous states (trace back to the MRCA)
- we correctly identify the directionality of the transformations (which state is plesiomorphic and which is apomorphic). The process of identifying the direction of change is called polarization.

Polarization could be done based on developmental considerations, paleontological evidence, or biogeographic considerations, but the most common technique is outgroup polarization.

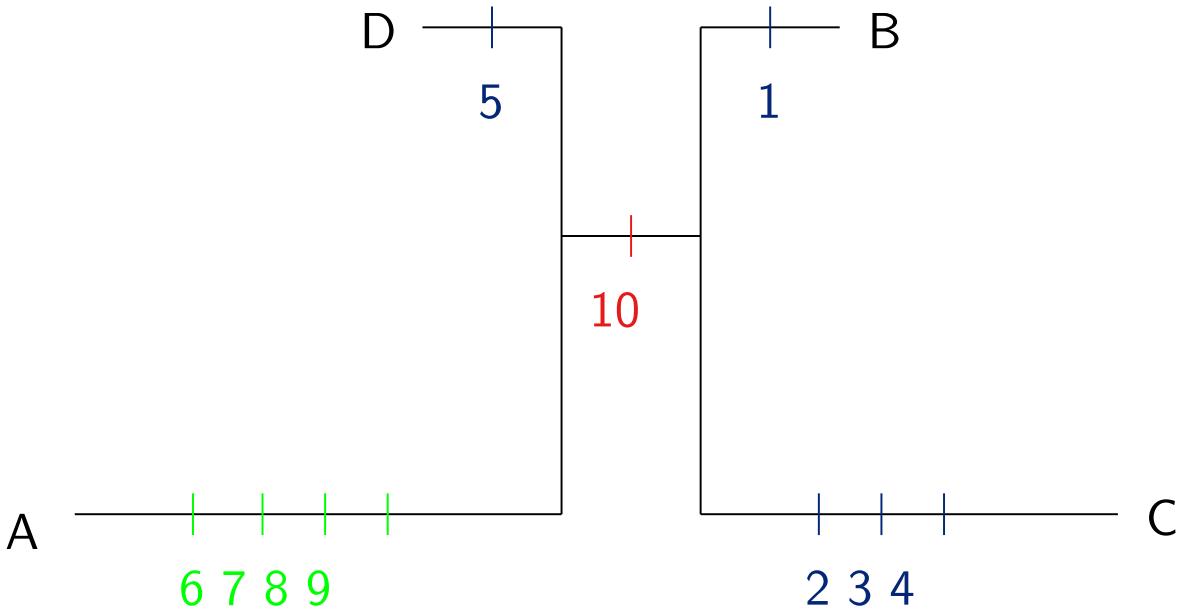
| Taxon | Character # | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |
| C | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| D | 0 | 0 | 0 | 0 | 1 | 1 |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

Draw the rooted tree and unrooted trees consistent with these data. Mark character state changes on the tree.


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


B C
D
A

- If characters are not polarized (ancestral and descendent states known) method can infer unrooted trees.
- We can infer 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.




B C

A D

**problems with this approach**

- We don't know polarization
- We observe character conflict in real data sets

## 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.

## 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 analyses which rely on counting and minimziing changes.

| 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 |

Draw the a tree consistent with these data. Mark character state changes on the tree.

| 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 |

How many trees can be drawn with the same number of character state changes for these data?

What factors would make each of those alternative trees seem more or less likely?

**Should we expect character conflict?**

- Data type?
- Evolutionary history?

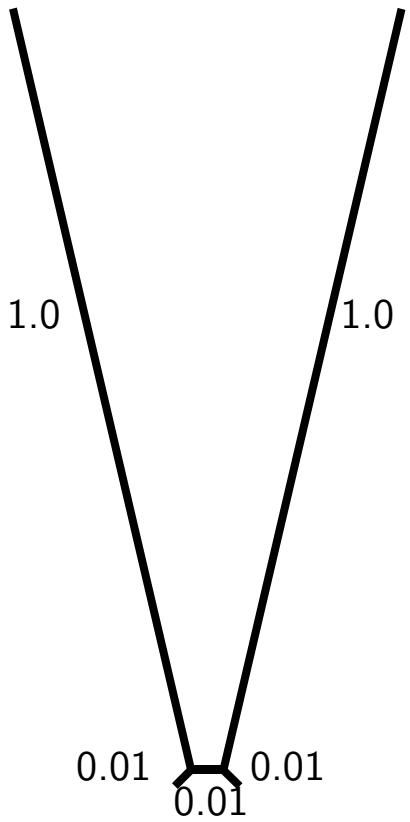
## Parsimony

- The simplest explanation is the most likely to be true
- Applied to phylogenetic inference, it is the inference metric that the tree with the fewest state changes is correct

## Qualitative description of parsimony

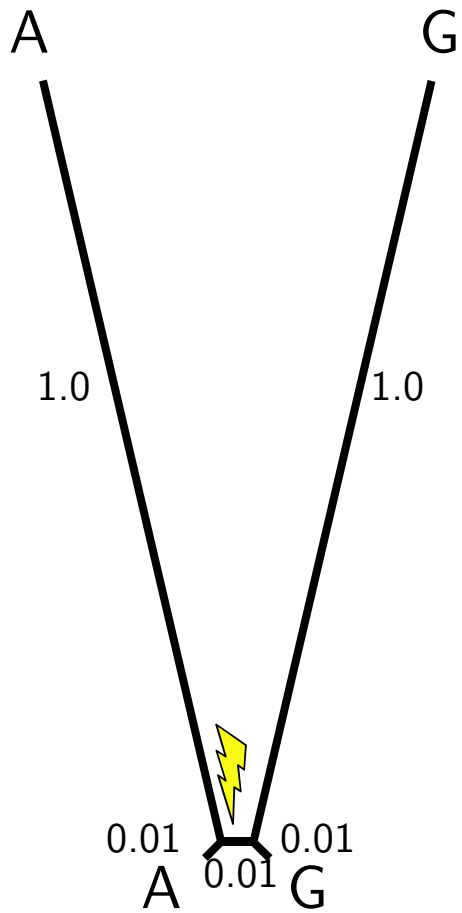
- Enables estimation of ancestral sequences.
- Even though parsimony always seeks to minimizes the number of changes, it can perform well even when changes are not rare.
- Does not "prefer" to put changes on one branch over another
- Hard to characterize statistically
  - the set of conditions in which parsimony is guaranteed to work well is very restrictive (low probability of change and not too much branch length heterogeneity);
  - Parsimony often performs well in simulation studies (even when outside the zones in which it is guaranteed to work);
  - Estimates of the tree can be extremely biased.

# Long branch attraction




Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

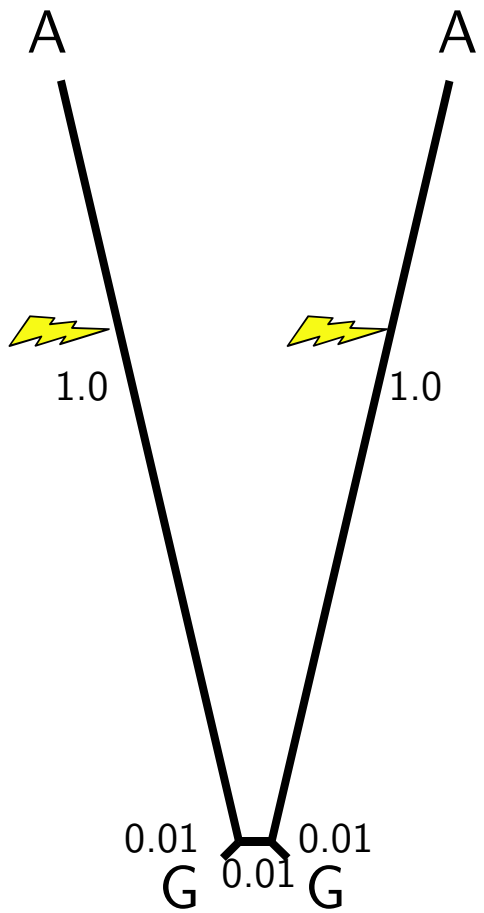
## Long branch attraction




Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

# Long branch attraction



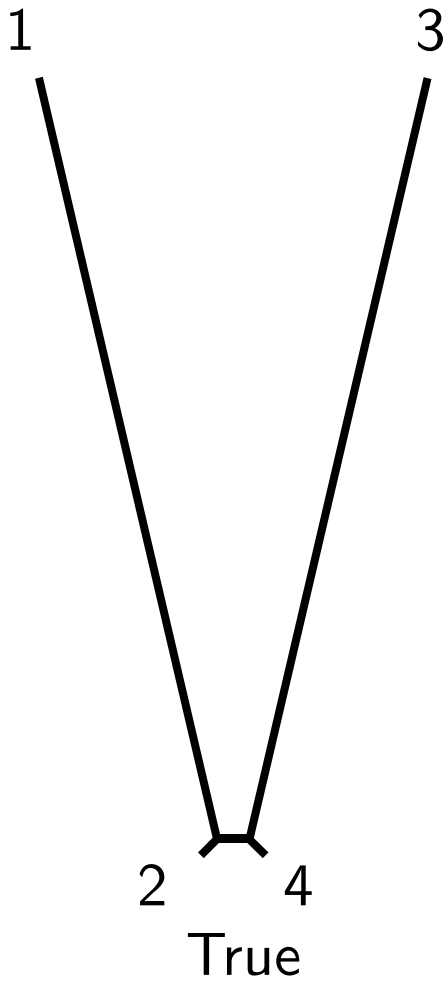

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**: 401-410.

The probability of a parsimony informative site due to inheritance is very low, (roughly 0.0003).

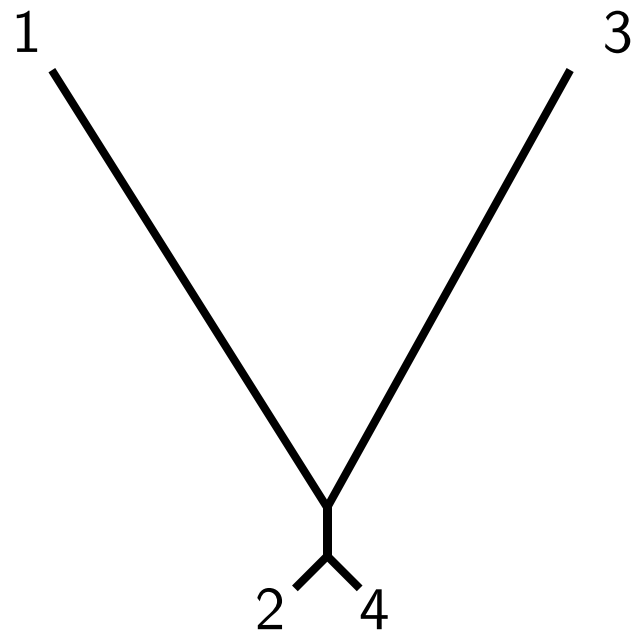
The probability of a misleading parsimony informative site due to parallelism is much higher (roughly 0.008).

## Long branch attraction

Parsimony is almost guaranteed to get this tree wrong.




True




Inferred

**How can we deal with character conflict?**

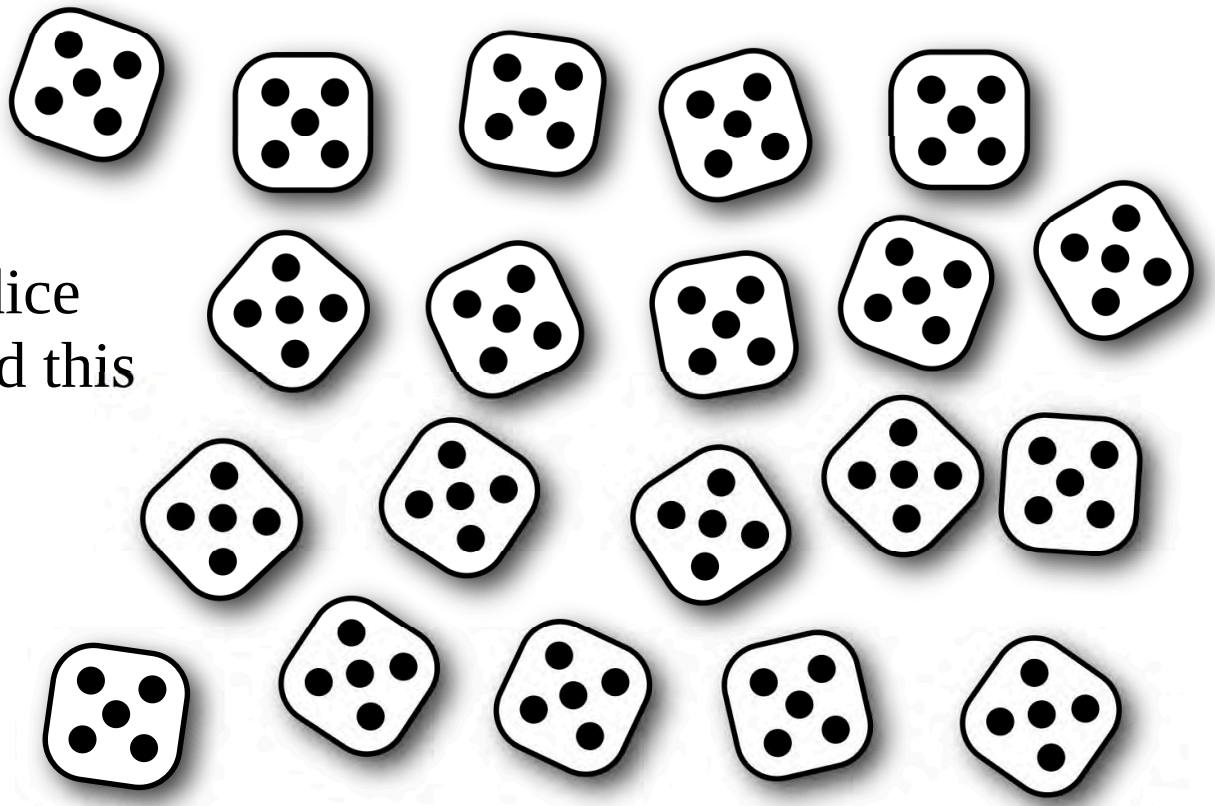
- We need to apply an error model
- Likelihood provides a measure of surprise under different models

# The Likelihood Criterion

The probability of the observations computed using a model tells us how surprised we should be.

*The preferred model is the one that surprises us least.*

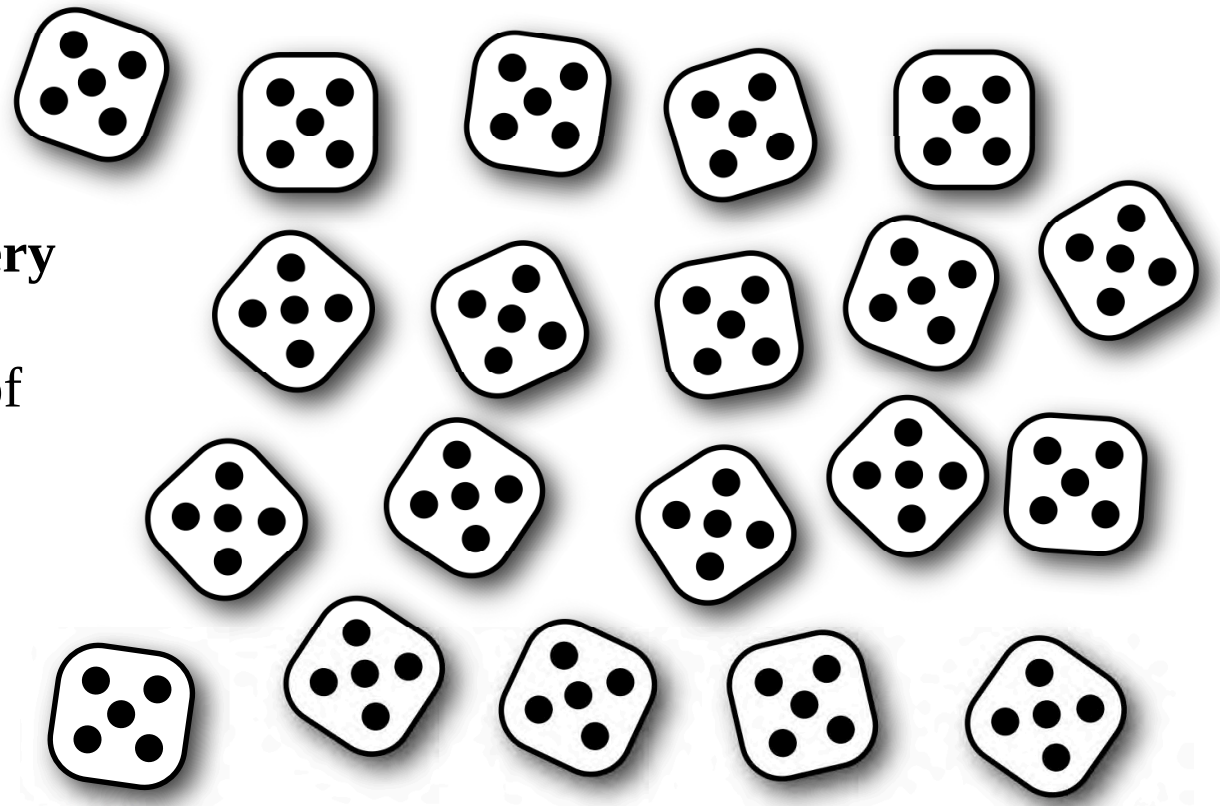
Suppose I threw 20 dice down on the table and this was the result...



# The Fair Dice model

$$\Pr(\text{obs.}|\text{fair dice model}) = \left(\frac{1}{6}\right)^{20} = \frac{1}{3,656,158,440,062,976}$$

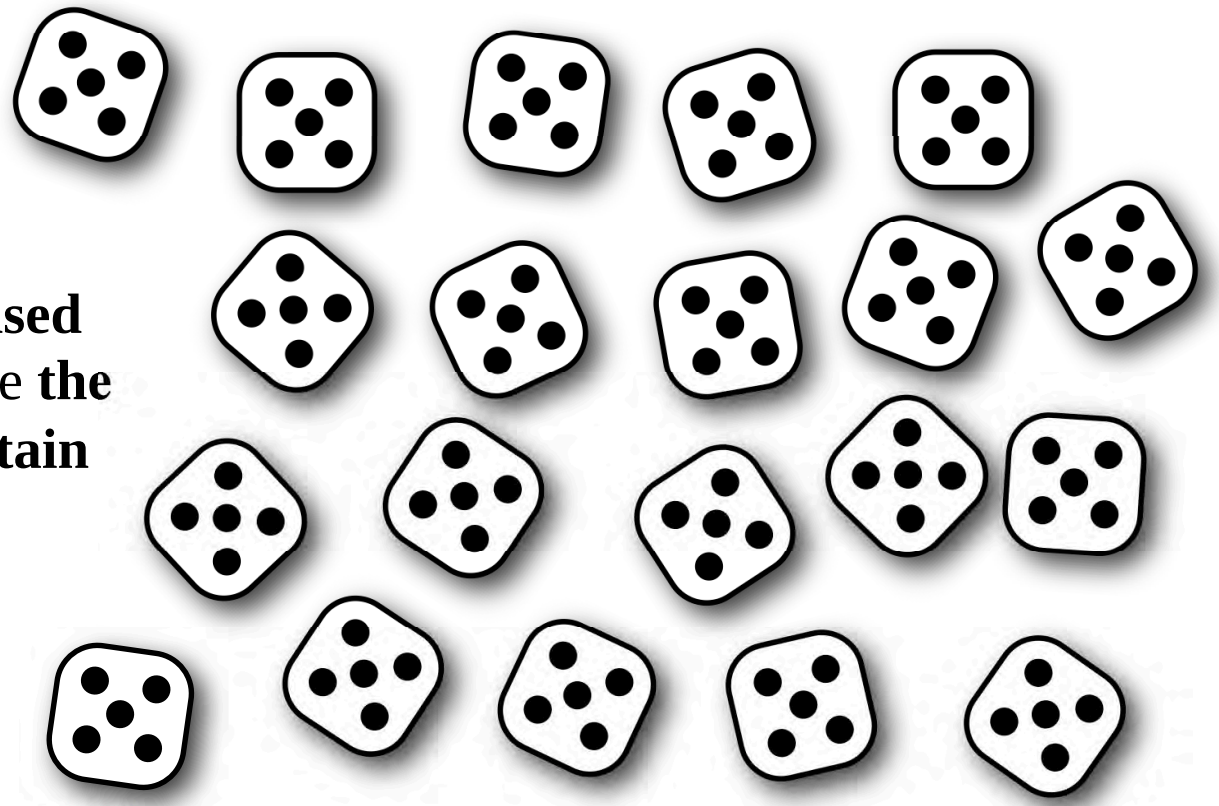
You should have been **very surprised** at this result because the probability of this event is **very small**: only 1 in 3.6 quadrillion!



# The Trick Dice model

(assumes dice each have 5 on every side)

$$\Pr(\text{obs.}|\text{trick dice model}) = 1^{20} = 1$$



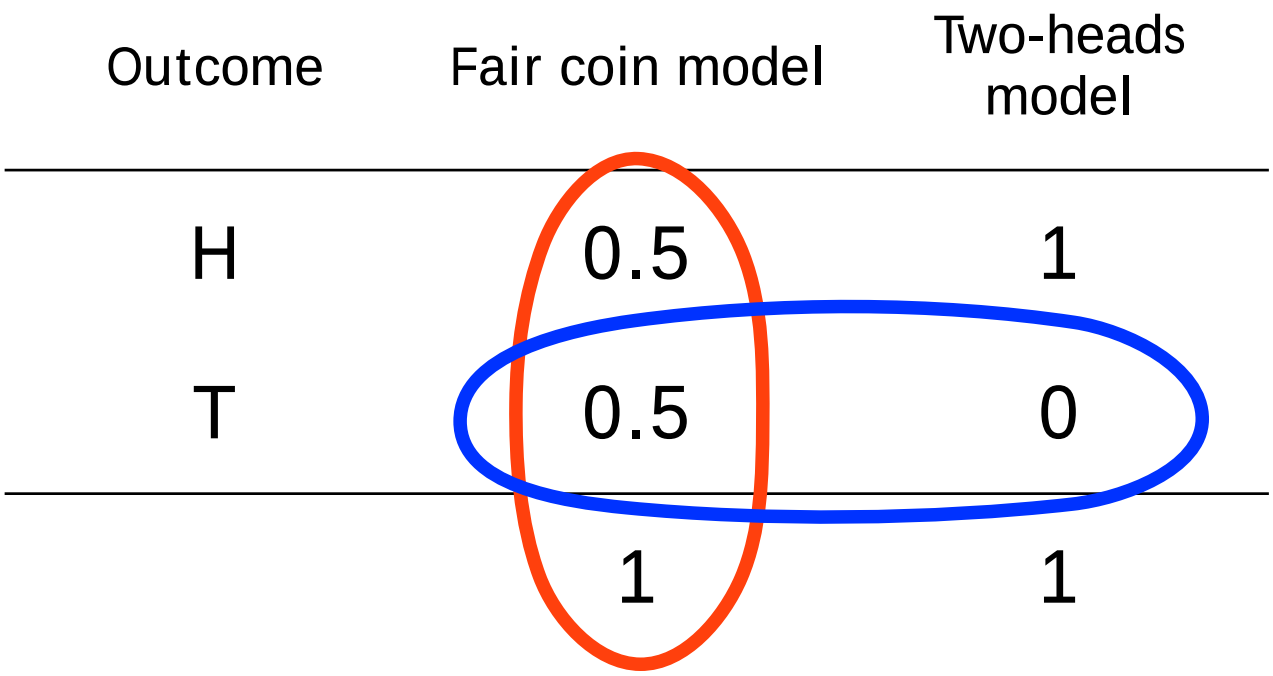
You should **not be surprised at all** at this result because **the observed outcome is certain** under this model

# Results

| Model | Likelihood | Surprise level |
|---|---|---|
| Fair Dice | $\frac{1}{3{,}656{,}158{,}440{,}062{,}976}$ | Very, *very*, ***very*** surprised |
| Trick Dice | 1 | Not surprised at all |

winning model maximizes likelihood (and thus minimizes surprise)

# Likelihood: why a new term?




**Likelihoods** of models given one particular data outcome are *not* expected to sum to 1.0

**Probabilities** of data outcomes given one particular model sum to 1.0

# Likelihood and model comparison

- Analyses using likelihoods ultimately involve **model comparison**
- The models compared can be **discrete** (as in the fair vs. trick dice example)
- More often the models compared differ **continuously**:
  - Model 1: branch length is 0.01
  - Model 2: branch length is 0.02
  - Model 3: branch length is 0.03

Rather than having an infinity of models, we instead think of the branch length as a **parameter** within one model