Tree formats and vocabulary

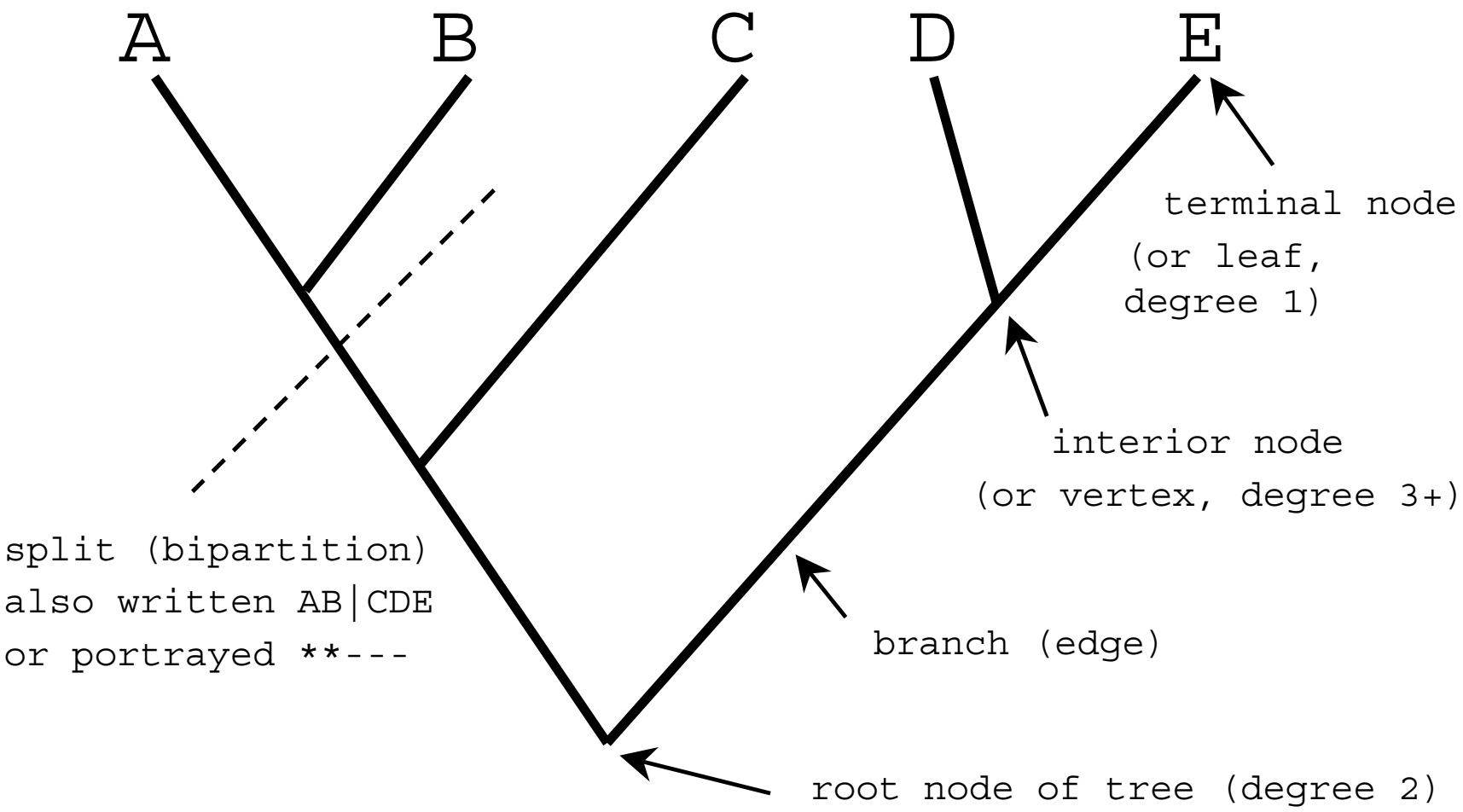
Emily Jane McTavish

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

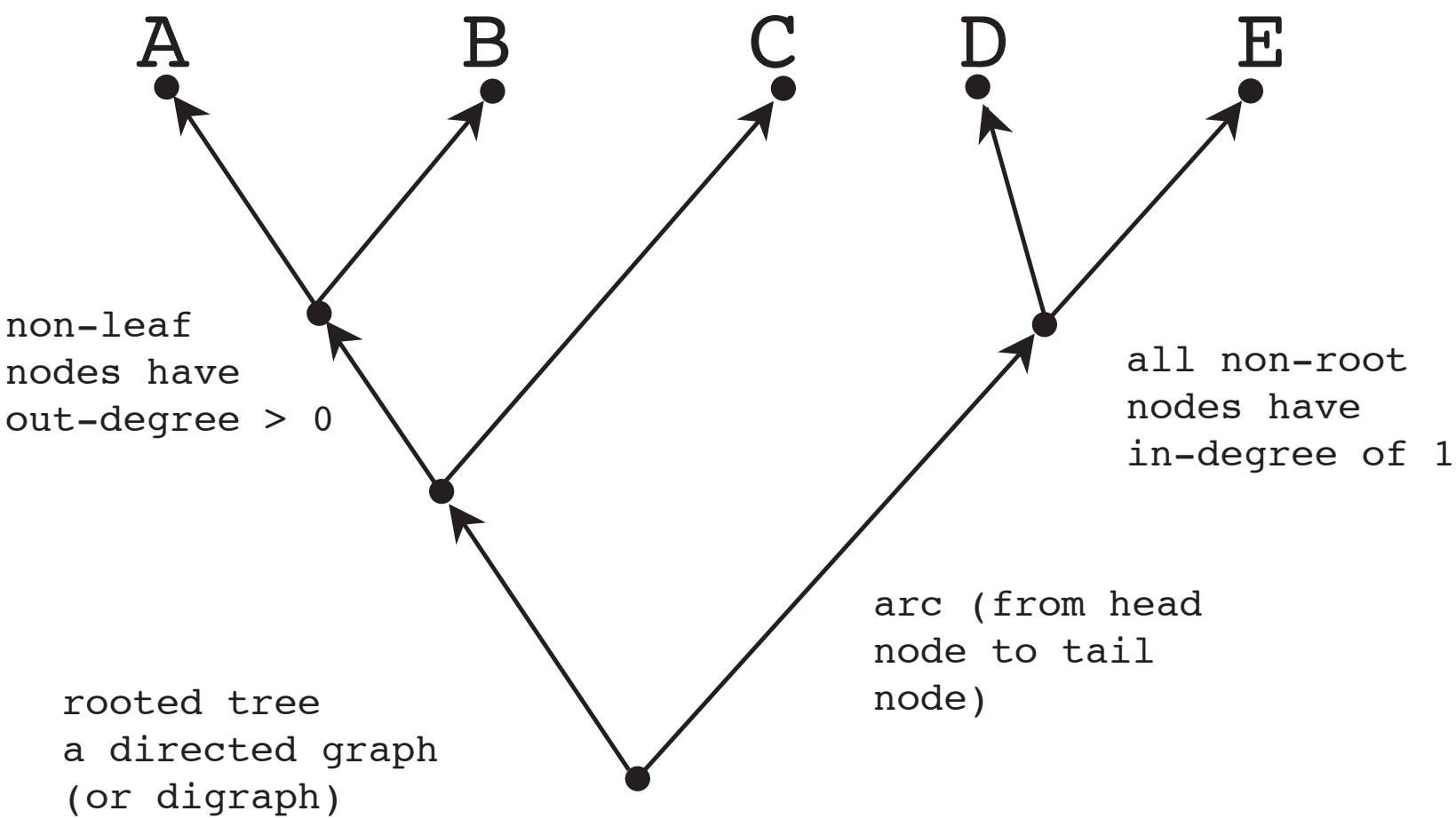
(With thanks to Mark Holder, Paul Lewis, Joe Felsenstein, and David Hillis for slides)

Phylogenies describe shared ancestry
and
inform our understanding of evolutionary processes

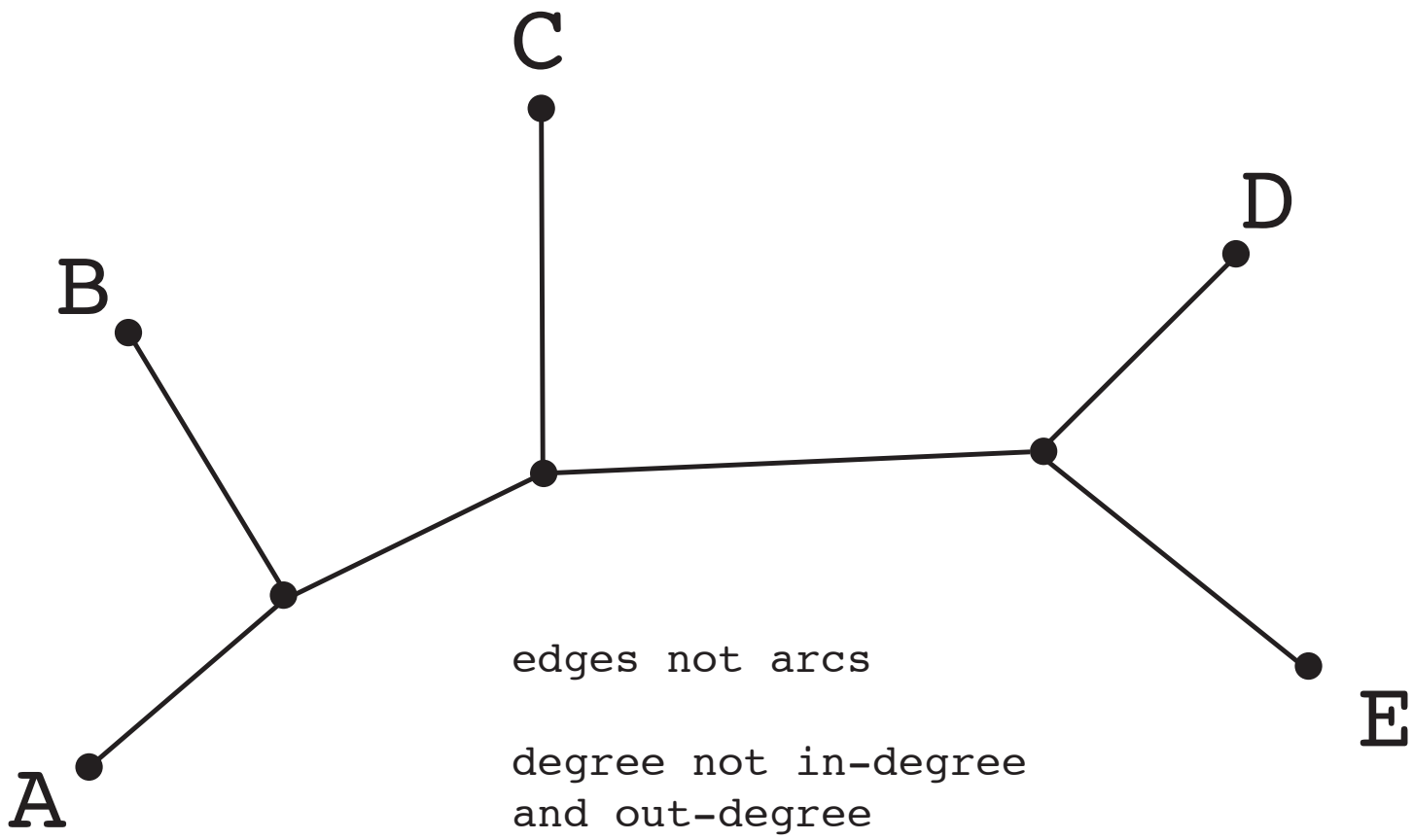
Tree terminology



Rooted tree terminology



Rooted tree terminology



Tree terms

A tree is a connected, acyclic graph.

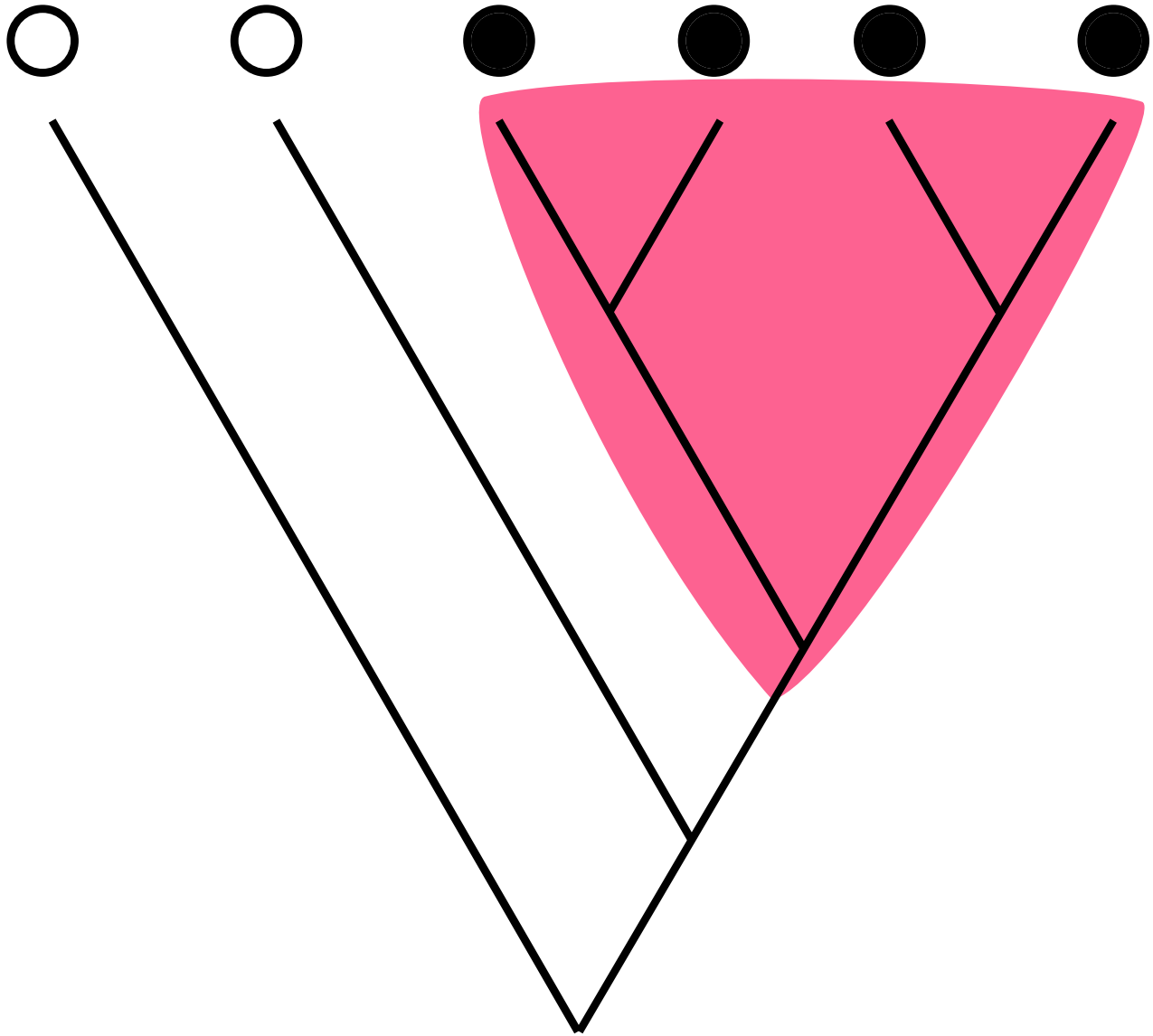
A rooted tree is a connected, acyclic directed graph.

A polytomy or multifurcation is a node with a degree > 3 (in an unrooted tree), or a node with an out-degree > 2 (in a rooted tree).

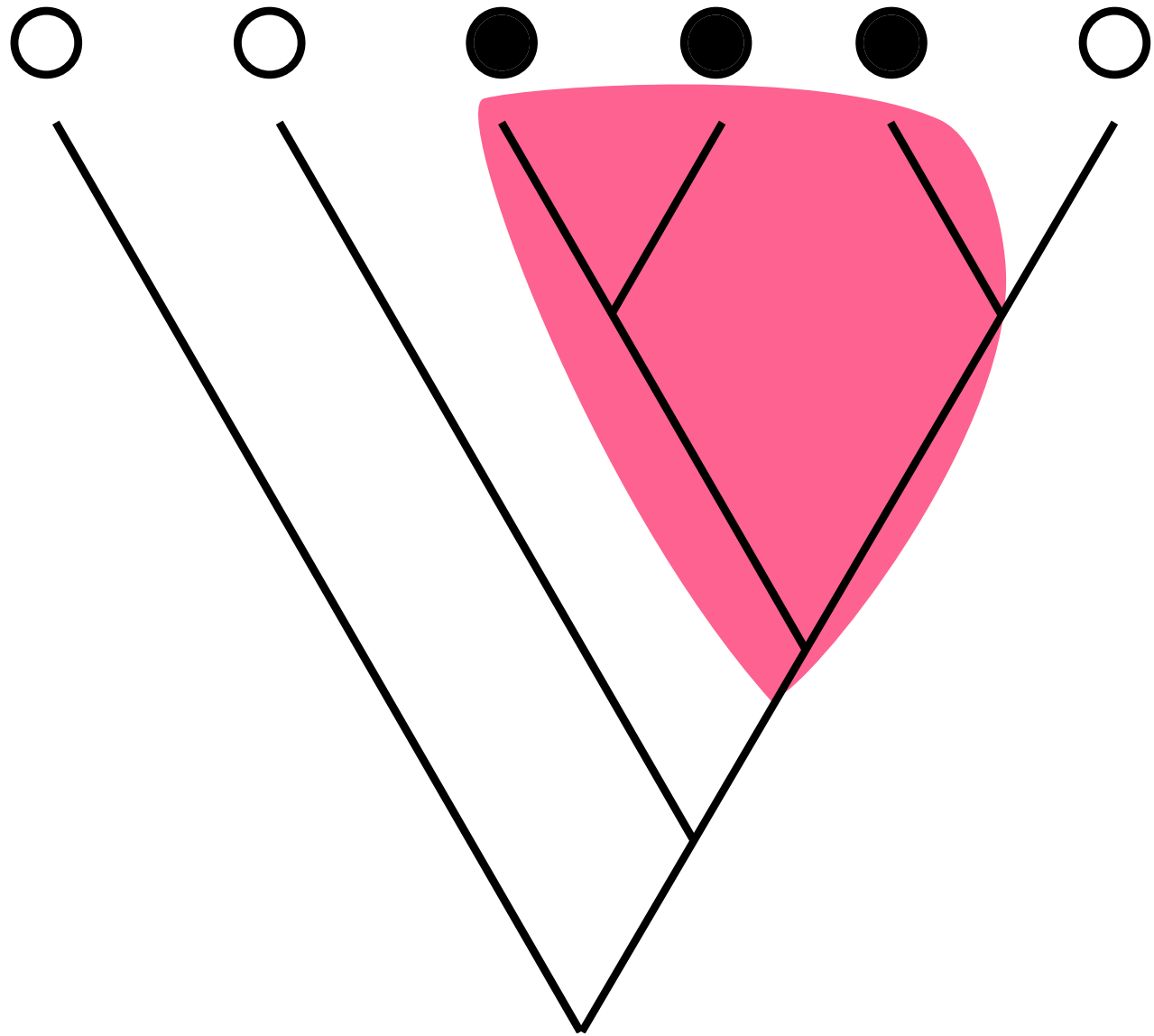
Collapsing an edge means to merge the nodes at the end of the branch (resulting in a polytomy in most cases).

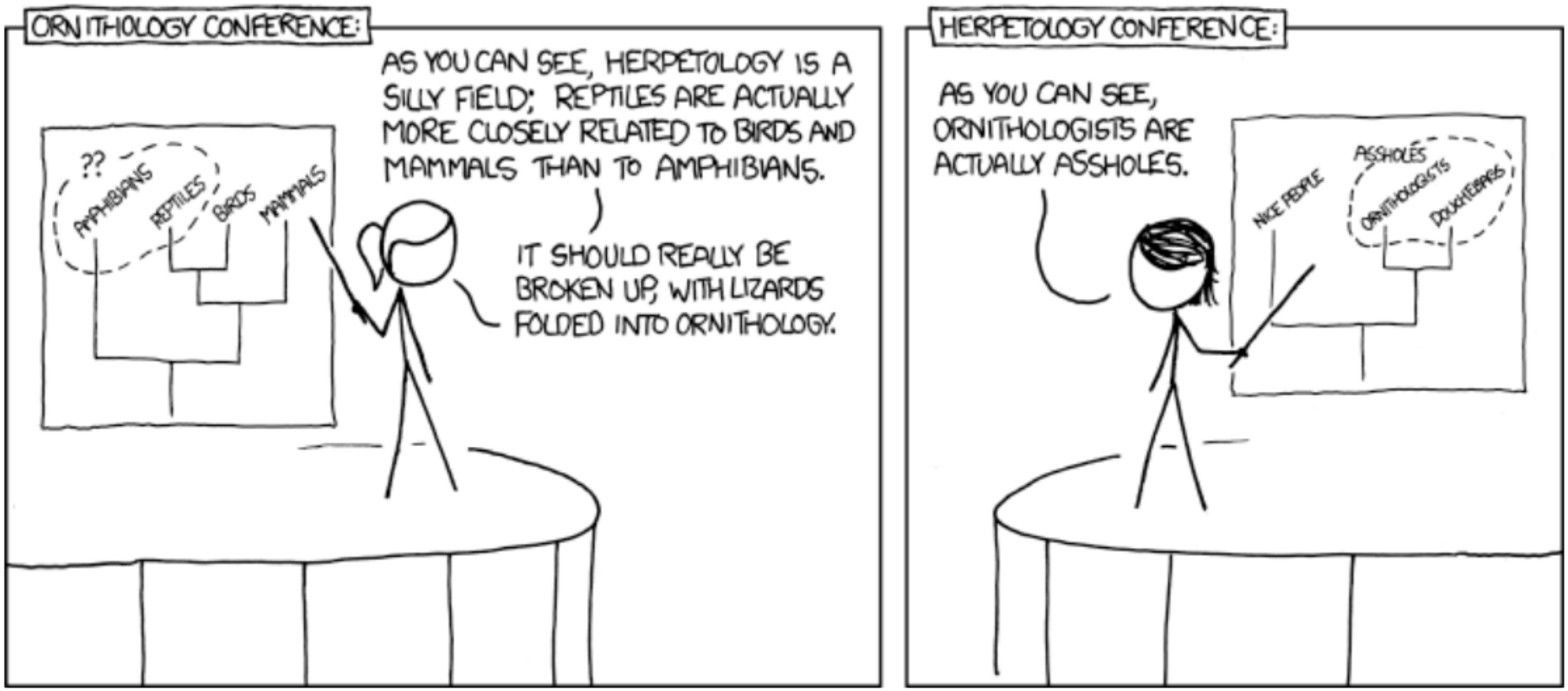
Refining a polytomy means to "break" the node into two nodes that are connected by an edge.

Monophyletic groups ("clades"): the basis of phylogenetic classification



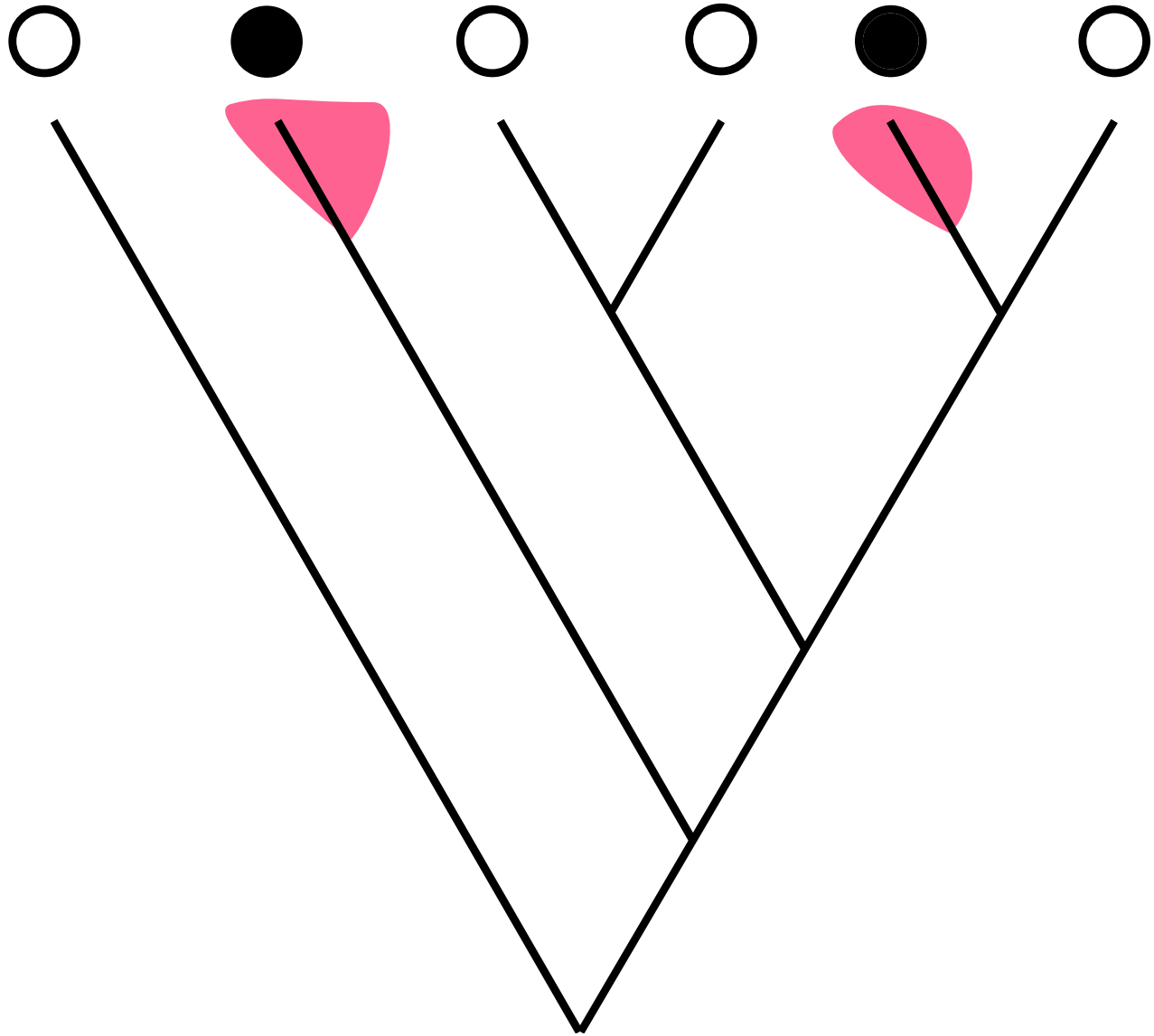
Paraphyletic groups: error of omitting some species

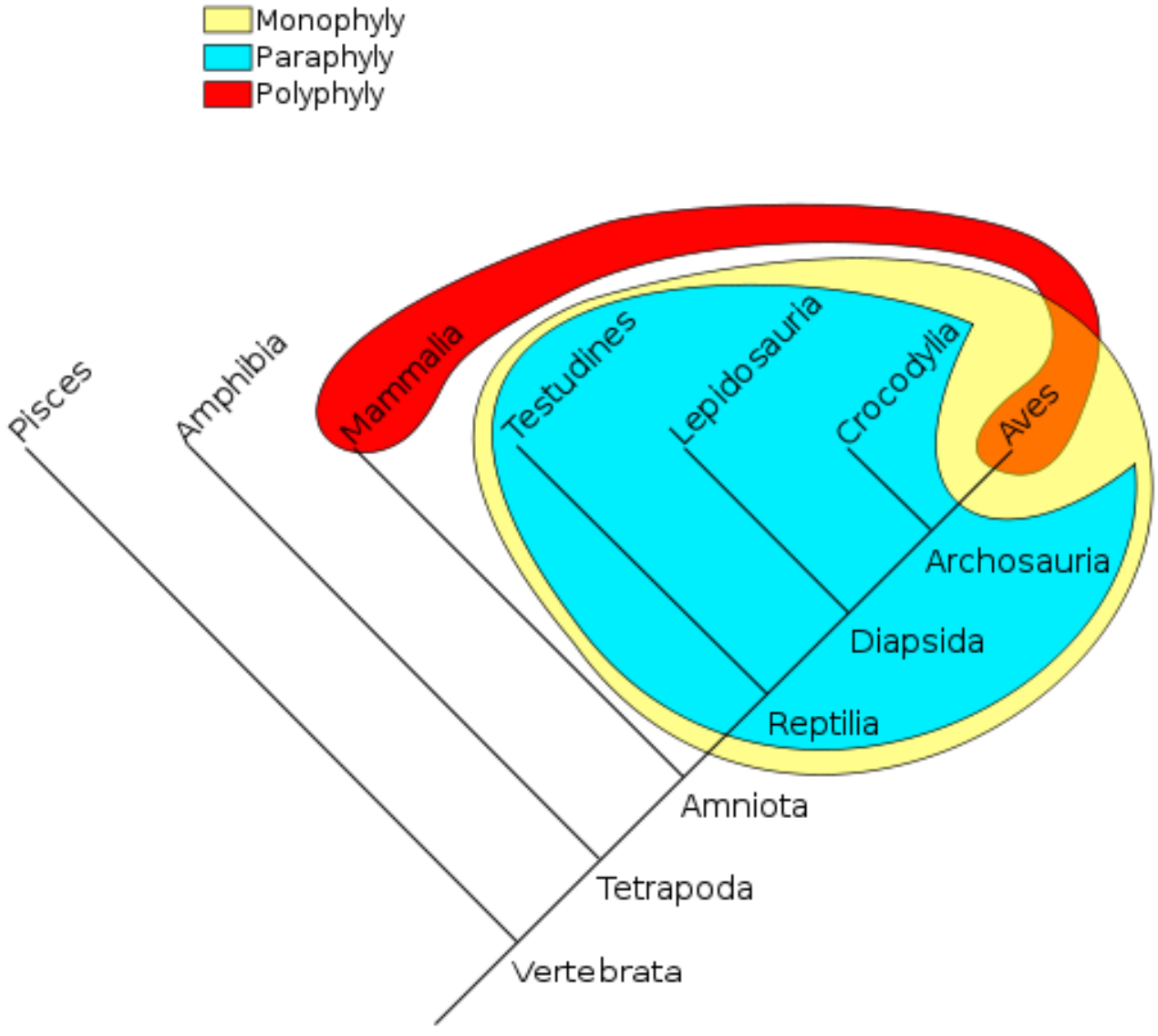




https://xkcd.com/867/

Polyphyletic groups: error of grouping “unrelated” species



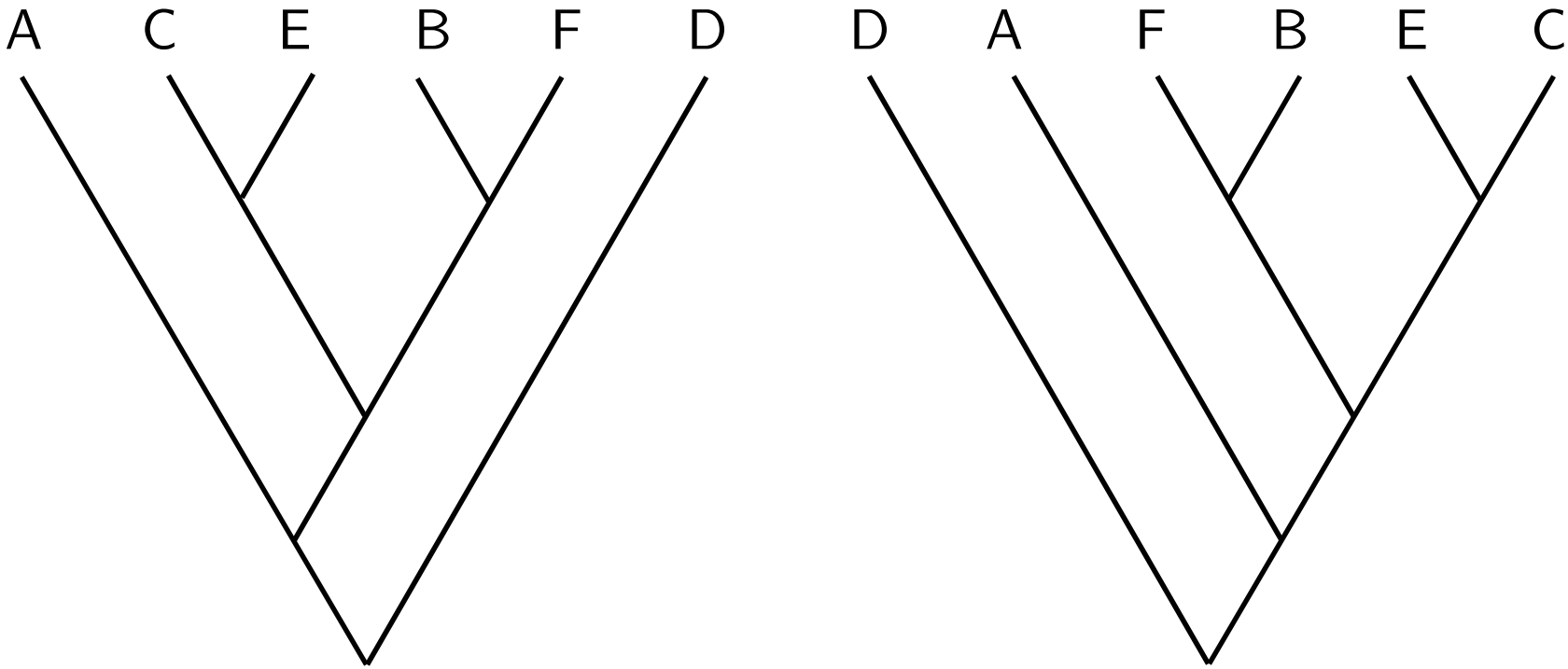


from wikipedia

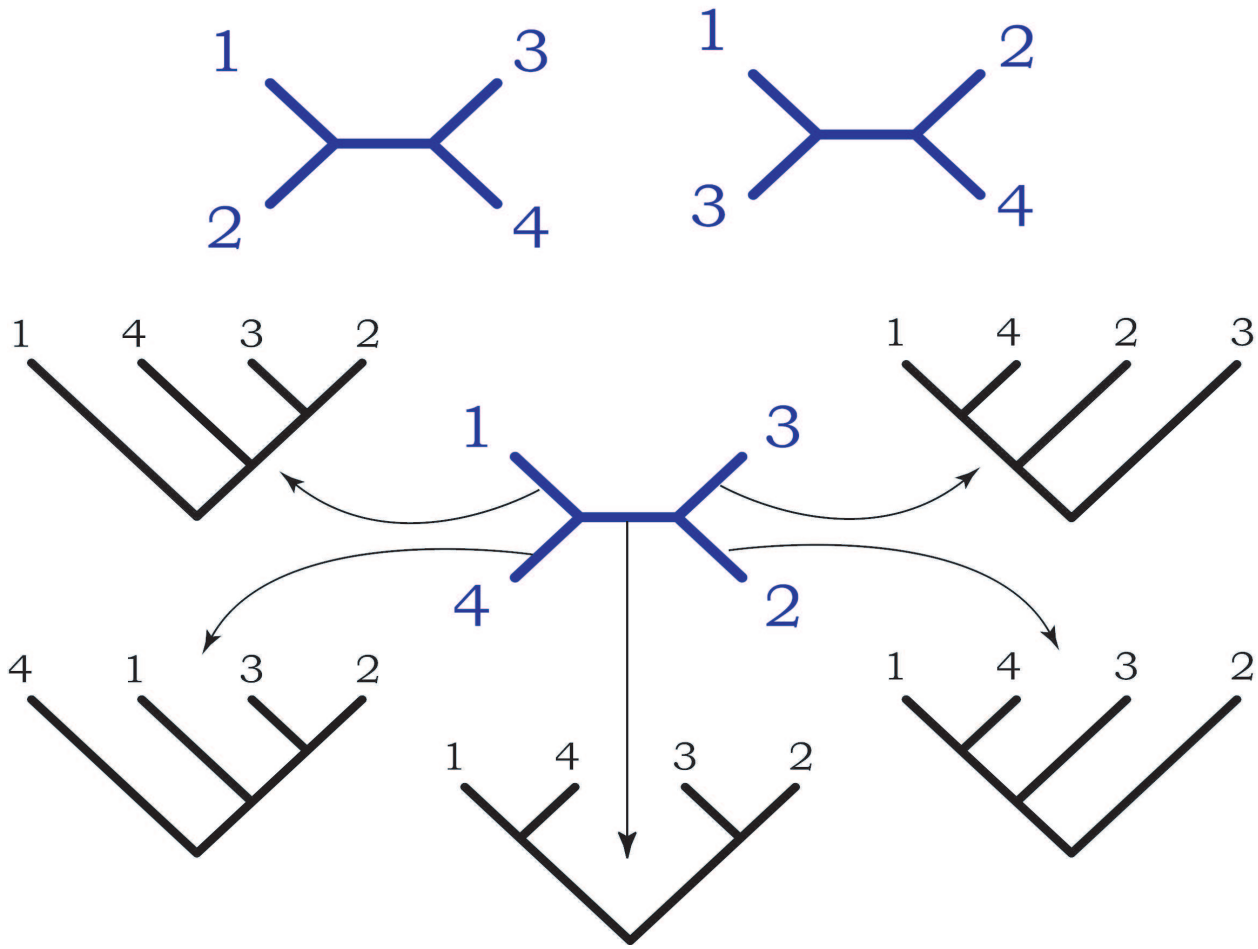
more terms:

- sister taxa: taxa or monophyletic groups which share a most recent common ancestor
- outgroup: taxon that is determined *a priori* to be sister to all other taxa in the analysis. Used for rooting tree

Branch rotation does not matter



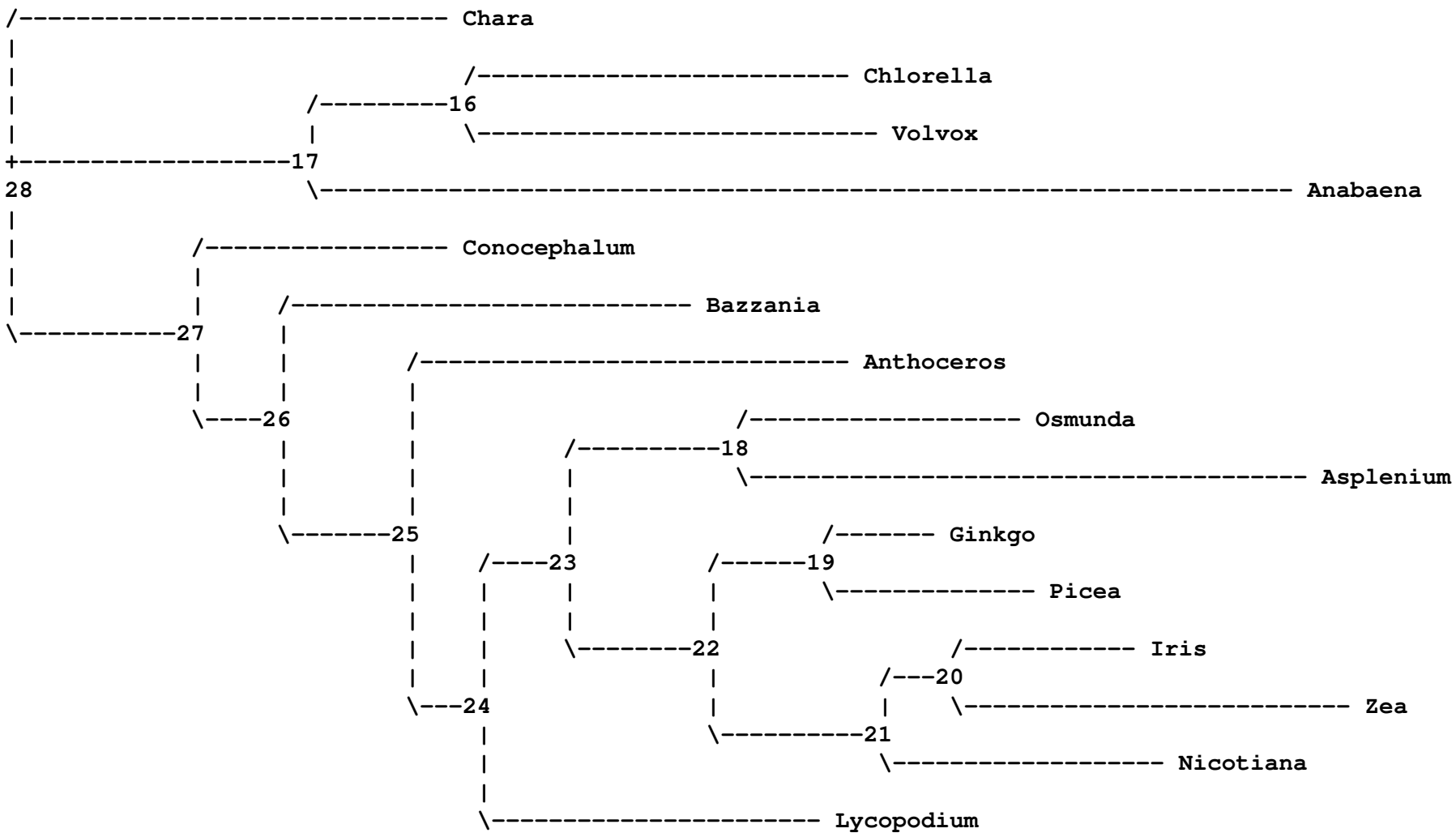
Rooted vs unrooted trees



Splits

- It is useful to think of unrooted trees in terms of 'splits'
- Each branch in an unrootred tree splits the taxa into two groups.
- Membership in those groups can be denoted by ** vs ..
- e.g. a split between 1+2 and 3+4 can be summarized as
- 1234
- ** ..

Warning: software often displays unrooted trees like this:



a brief digression into newick tree file format

Newick's Lobster House was the site of an historic 1986 meeting at which a standard was devised for storing descriptions of phylogenetic trees as strings. (Photo from Paul Lewis)

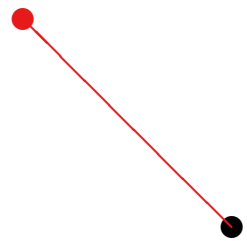
Note: ((1,2),3,4) is referred to as Newick or New Hampshire notation for the tree.

You can read it by following the rules:

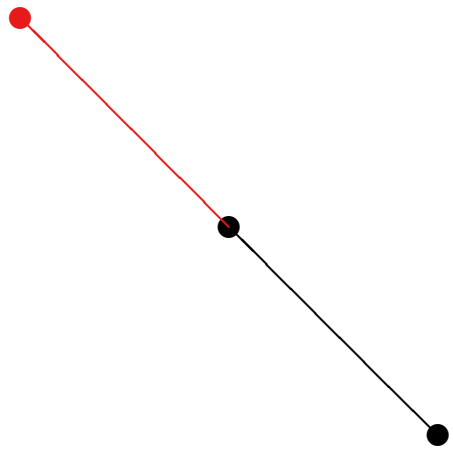
- start at a node,
- if the next symbol is '(' then add a child to the current node and move to this child,
- if the next symbol is a label, then label the node that you are at,
- if the next symbol is a comma, then move back to the current node's parent and add another child,
- if the next symbol is a ')', then move back to the current node's parent.

((1,2),3,4)

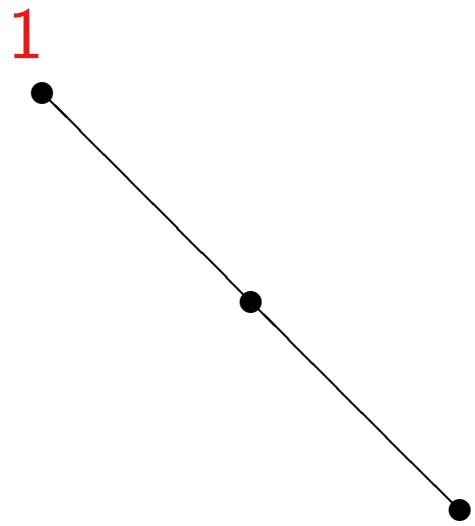
((1,2),3,4)



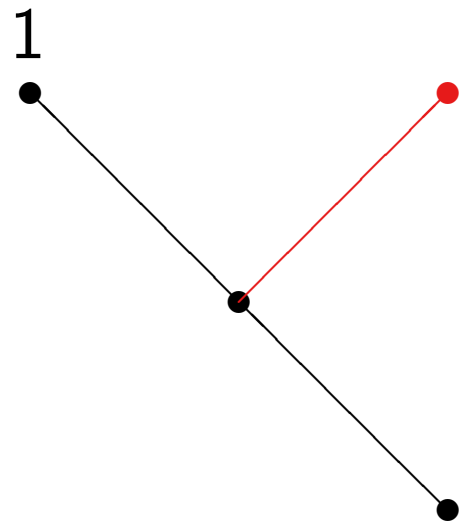
((1,2),3,4)



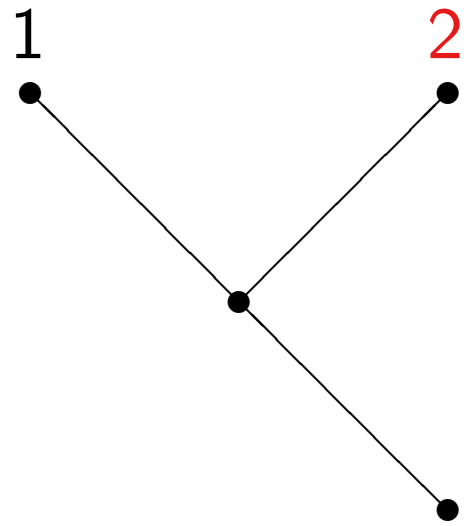
((1,2),3,4)



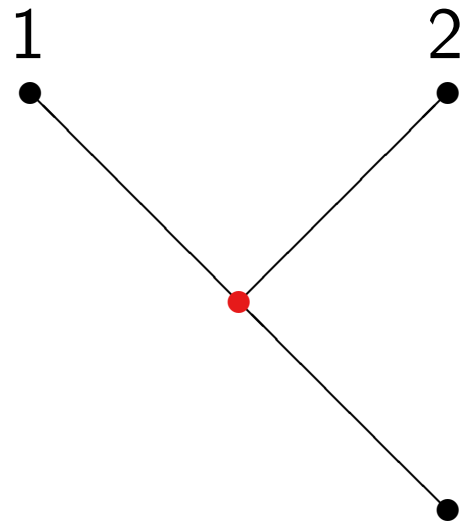
((1,2),3,4)



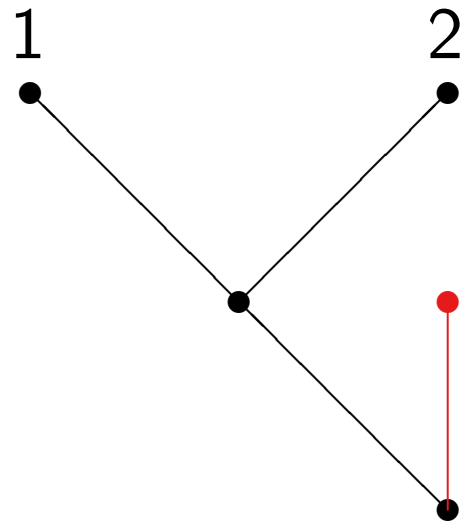
((1,2),3,4)



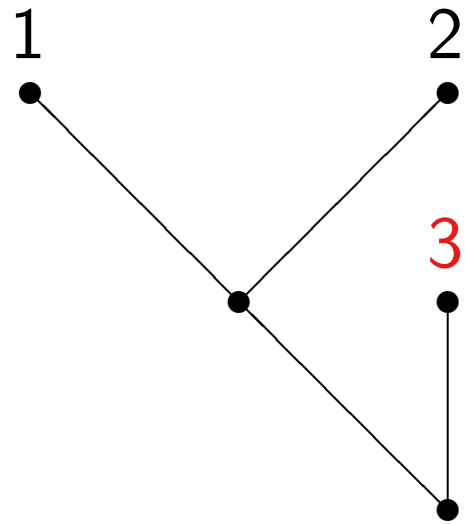
((1,2),3,4)



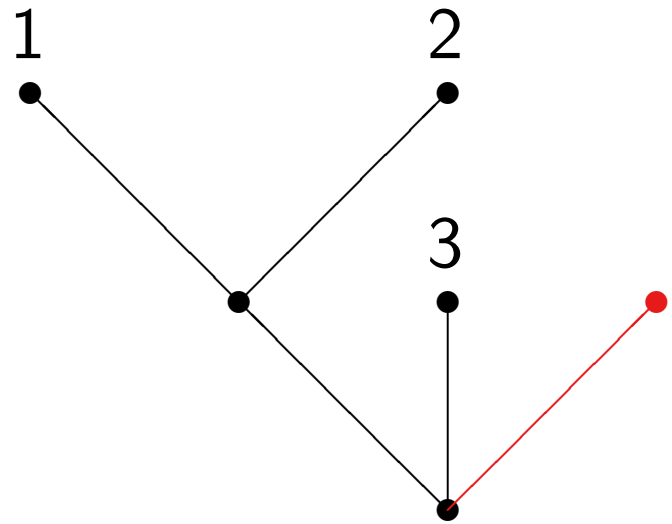
((1,2),3,4)



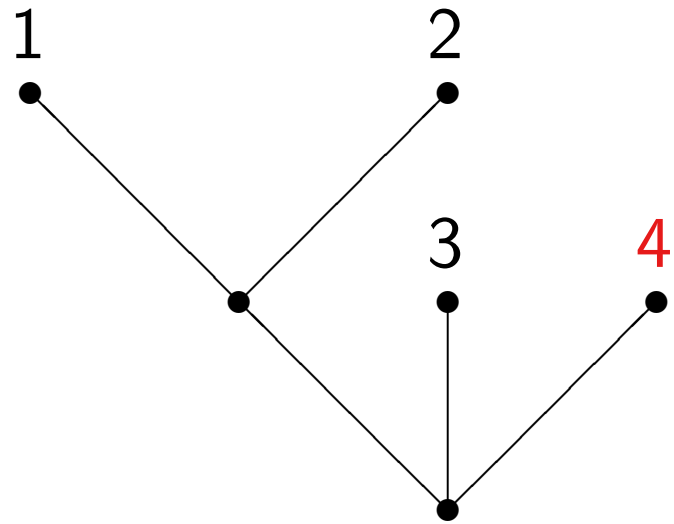
((1,2),3,4)



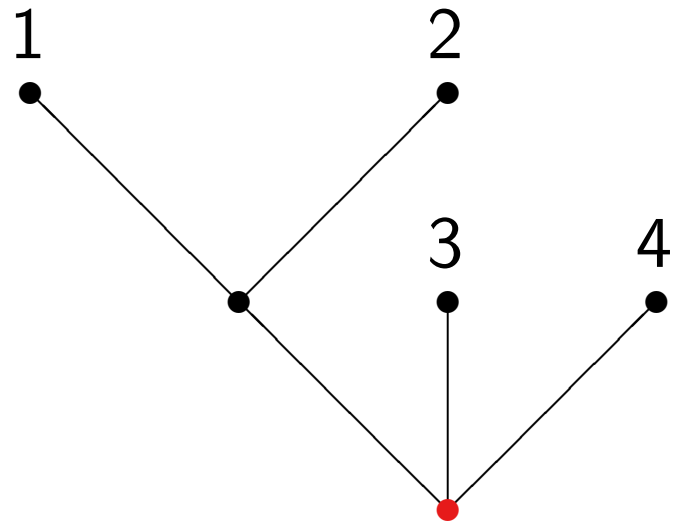
((1,2),3,4)



((1,2),3,4)



((1,2),3,4)



Newick

- Parenthetical tree format
- Rooted vs. unrooted trees are not differentiated
- Some programs interpret polytomy at root as ‘unrooted
- Branches and nodes not well differentiated
- A name can contain and characters except blanks, colons, semicolons, parentheses, and square brackets

Nexus

- Starts with #nexus
- Can contain blocks of alignments, trees, commands, and more!
- Blocks between ‘begin’ and ‘end’
- Trees in Newick format, prepended with [&U] unrooted or [&R] rooted

Nexus

```
#nexus
...
begin taxa;
  dimensions ntax=5;
  taxlabels
    Giardia
    Thermus
    Deinococcus
    Sulfolobus
    Haobacterium
  ;
end;
```

```
#nexus
...
begin data;
  dimensions ntax=5 nchar=54;
  format datatype=dna missing=? gap=-;
  matrix
    Ephedra       TTAAGCCATGCATGTCTAAGTATGAACTAATTCCAAACGGTGAAACTGCGGATG
    Gnetum        TTAAGCCATGCATGTCTATGTACGAACTAATC-AGAACGGTGAAACTGCGGATG
    Welwitschia   TTAAGCCATGCACGTGTAAGTATGAACTAGTC-GAAACGGTGAAACTGCGGATG
    Ginkgo        TTAAGCCATGCATGTGTAAGTATGAACTCTTTACAGACTGTGAAACTGCGAATG
    Pinus         TTAAGCCATGCATGTCTAAGTATGAACTAATTGCAGACTGTGAAACTGCGGATG
                 [----+--10|----+--20|----+--30|----+--40|----+--50|----]
  ;
end;
```

http://hydrodictyon.eeb.uconn.edu/eebedia/index.php/Phylogenetics:_NEXUS_Format

Nexus

```
#nexus
...
begin trees;
  translate
    1 Ephedra,
    2 Gnetum,
    3 Welwitschia,
    4 Ginkgo,
    5 Pinus
  ;
  tree one = [&U] (1,2,(3,(4,5));
  tree two = [&U] (1,3,(5,(2,4));
end;
```

```
#nexus
...
begin sets;
  charset trnL_intron = 562-4226;
  taxset gnetales = Ephedra Gnetum Welwitschia;
end;
```

http://hydrodictyon.eeb.uconn.edu/eebedia/index.php/Phylogenetics:_NEXUS_Format

NeXML

- Phylogenetic data as XML
- Can capture all information from Nexus
- Full semantic annotation
- Easily extensible

NeXML

Computer readable, but not very human readable

```
 <otu about="#otu99" id="otu99" label="Parupeneus barberinoides">
  <meta datatype="xsd:string" property="ot:originalLabel" xsi:type="nex:LiteralMeta">Parupeneus
  <meta datatype="xsd:int" property="ot:ottId" xsi:type="nex:LiteralMeta">758968</meta>
  <meta datatype="xsd:string" property="ot:ottTaxonName" xsi:type="nex:LiteralMeta">Parupeneus b
 </otu>
</otus>
<trees about="#trees1" id="trees1" otus="otus1">
 <tree about="#tree1" id="tree1" label="Untitled (tree1)" xsi:type="nex:FloatTree">
  <meta datatype="xsd:string" property="ot:branchLengthDescription" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:branchLengthMode" xsi:type="nex:LiteralMeta">ot:undef
  <meta datatype="xsd:string" property="ot:curatedType" xsi:type="nex:LiteralMeta">Bayesian infe
  <meta datatype="xsd:string" property="ot:inGroupClade" xsi:type="nex:LiteralMeta">node2</meta>
  <meta datatype="xsd:string" property="ot:nodeLabelMode" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:nodeLabelTimeUnit" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:outGroupEdge" xsi:type="nex:LiteralMeta"/>
  <meta datatype="xsd:string" property="ot:specifiedRoot" xsi:type="nex:LiteralMeta">node1</meta
  <meta datatype="xsd:boolean" property="ot:unrootedTree" xsi:type="nex:LiteralMeta">false</meta
  <node about="#node1" id="node1" root="true"/>
  <node about="#node2" id="node2"/>
  <node about="#node144" id="node144"/>
  <node about="#node145" id="node145"/>
  <node about="#node146" id="node146"/>
  <node about="#node147" id="node147"/>
  <node about="#node148" id="node148"/>
  <node about="#node149" id="node149"/>
  <node about="#node150" id="node150"/>
  <node about="#node151" id="node151"/>
  <node about="#node152" id="node152"/>
  <node about="#node153" id="node153"/>
  <node about="#node154" id="node154"/>
  <node about="#node155" id="node155" otu="otu72">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node156" id="node156" otu="otu73">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node157" id="node157" otu="otu74">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
  <node about="#node158" id="node158"/>
  <node about="#node159" id="node159" otu="otu75">
   <meta datatype="xsd:boolean" property="ot:isLeaf" xsi:type="nex:LiteralMeta">true</meta>
  </node>
```

Phylip (sequence data format)

- First line must be two integers: <number of taxa> <number of sites>
- Sequence ID followed by spaces up to 10 char.
- No duplicate names
- Relaxed phylip up to 250 characters followed by a space

```
      5    42
Turkey     AAGCTNGGGC ATTTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT ACAGGTTGGC CGTTCAGGGT AA
Chimp      AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla    AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA
```

Phylip interleaved

```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
Salmo gairAAGCCTTGGC AGTGCAGGGT
H. SapiensACCGGTTGGC CGTTCAGGGT
Chimp     AAACCCTTGC CGTTACGCTT
Gorilla   AAACCCTTGC CGGTACGCTT

GAGCCCGGGC AATACAGGGT AT
GAGCCGTGGC CGGGCACGGT AT
ACAGGTTGGC CGTTCAGGGT AA
AAACCGAGGC CGGGACACTC AT
AAACCATTGC CGGTACGCTT AA
```

Phylip sequential

```
  5    42
Turkey    AAGCTNGGGC ATTTCAGGGT
GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT
GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT
ACAGGTTGGC CGTTCAGGGT AA
Chimp     AAACCCTTGC CGTTACGCTT
AAACCGAGGC CGGGACACTC AT
Gorilla   AAACCCTTGC CGGTACGCTT
AAACCATTGC CGGTACGCTT AA
```

Fasta (sequence data format)

- Description line before each sequence starts with (">") symbol in the first column

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

An Exercise: Writing and Viewing a Tree by Hand

1. Make sure you have:
   - A *good* text editor installed.
   - *FigTree* (or some other tree visualizer of choice).
2. Create a subdirectory for all our labs, e.g. "`$HOME/projects/GradPhylo`".
3. Create a work subdirectory within this directory for this particular lab, e.g. "`lab-01`".
4. Using a command line text editor (nano or or personal favorite) create a simple Newick tree file using a text editor, e.g. "simple.newick".
5. This tree should reflect the real topological relationships of at least 4 species.
6. Visualize it in *FigTree*.
7. Export as a NEXUS file.
8. Open this second file in a text editor and edit the tip labels.
9. Visualize it again.

DIY

Create a newick tree file in your text editor with the content:
`(((C,(D,E)),(F,G),A),B);`
Save it as 'example.tre'.

1. Draw the tree by hand
2. Write down all the splits in "AB|CD" format.
3. Load the tree in a tree viewer (e.g. phylo.io, figtree). Re-root the tree. What rootings make the following true? Which cannot be true?
   - 4a. A is more closely related to G than it is to C
   - b (C,D,E) is sister to (A,B,F,G)
   - c (C,D) is sister to (A,B,E,F,G)
   - d (C,D,E) is a paraphyletic group
   - e (C,D,E) is a monophyletic group
   - f (A,B,C) is a monophyletic group