CS4054
Bioinformatics

Spring 2025

Rushda Muneer

An Idea for Distance-Based Phylogeny

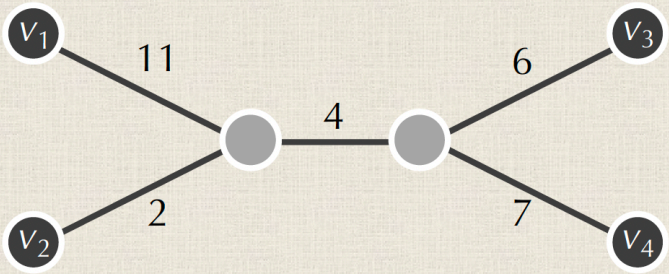
	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	12	13
v_3	21	12	0	13
v_4	22	13	13	0

Exercise: Apply our recursive approach to this additive distance matrix.

What Was Wrong With Our Algorithm?

	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	12	13
v_3	21	12	0	13
v_4	22	13	13	0

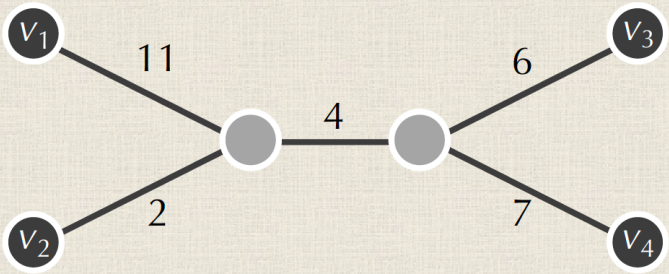
Here is the tree fitting the matrix...



What Was Wrong With Our Algorithm?

minimum
element is $D_{2,3}$

	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	**12**	13
v_3	21	12	0	13
v_4	22	13	13	0

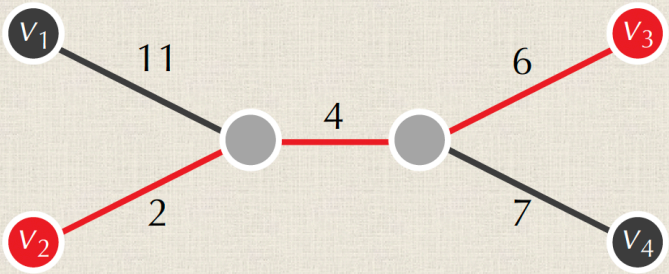


What Was Wrong With Our Algorithm?

minimum element is $D_{2,3}$

v_2 and v_3 are not neighbors!

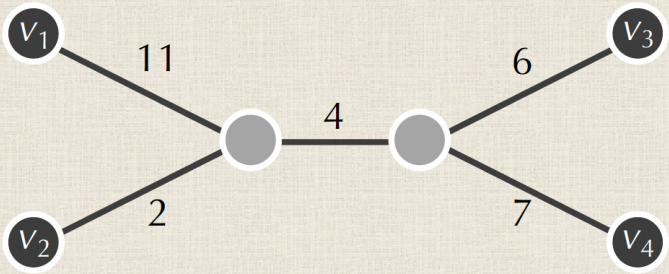
	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	**12**	13
v_3	21	12	0	13
v_4	22	13	13	0



What Was Wrong With Our Algorithm?

STOP: What about the tree prevents the minimum matrix element from corresponding to neighbors?

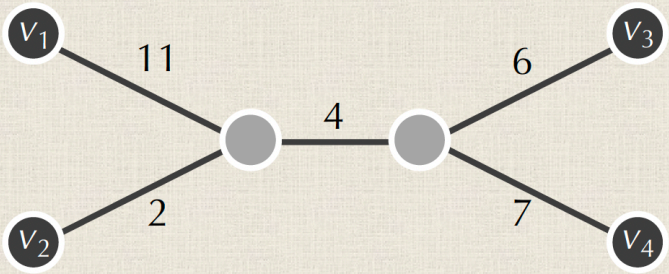
	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	12	13
v_3	21	12	0	13
v_4	22	13	13	0



What Was Wrong With Our Algorithm?

STOP: What about the tree prevents the minimum matrix element from corresponding to neighbors?

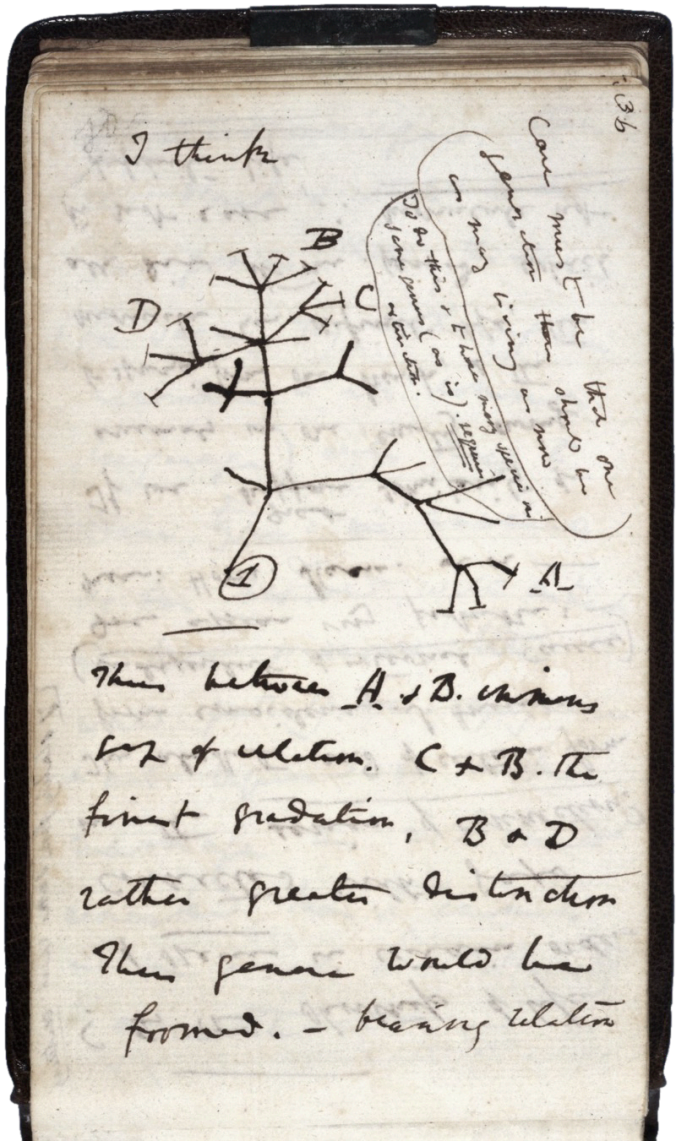
	v_1	v_2	v_3	v_4
v_1	0	13	21	22
v_2	13	0	12	13
v_3	21	12	0	13
v_4	22	13	13	0



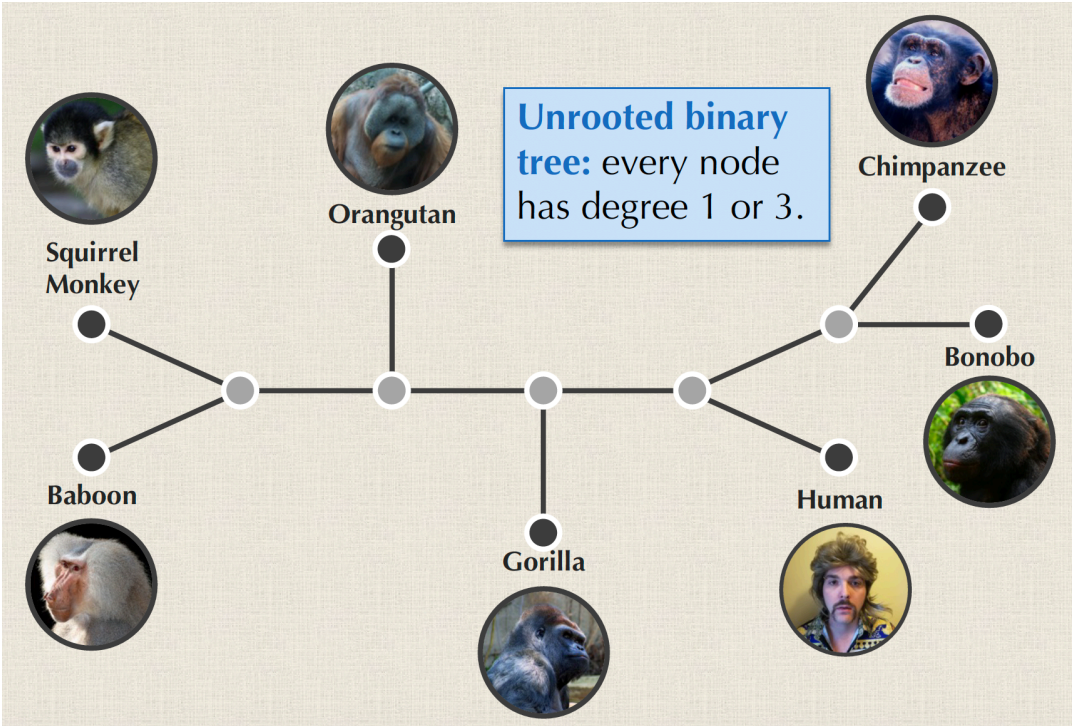
Answer: v_1 is an outlier, so the distance from v_1 to any other node is higher than average.

Modeling Speciations

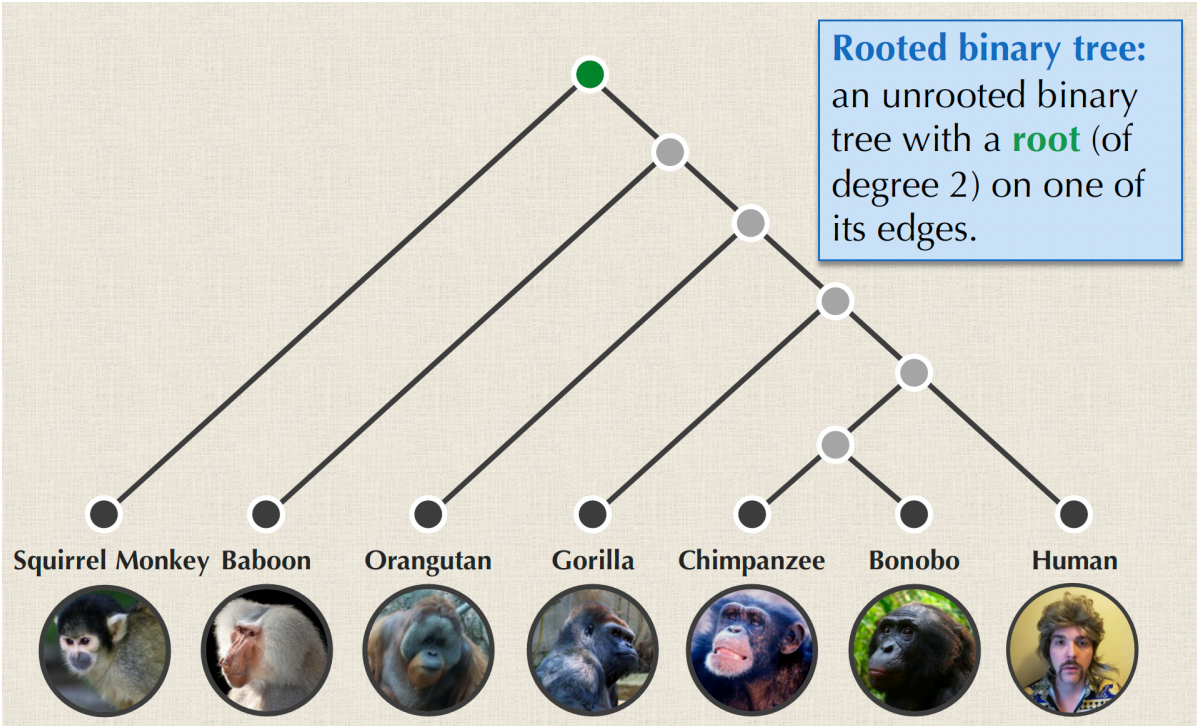
- Even though the minimum element of D doesn't necessarily correspond to neighbors, this assumption powers one of the most famous evolutionary tree heuristics.
- Researchers often assume that all internal nodes correspond to speciations, where one species splits into two.



We've Thus Far Seen Unrooted Trees

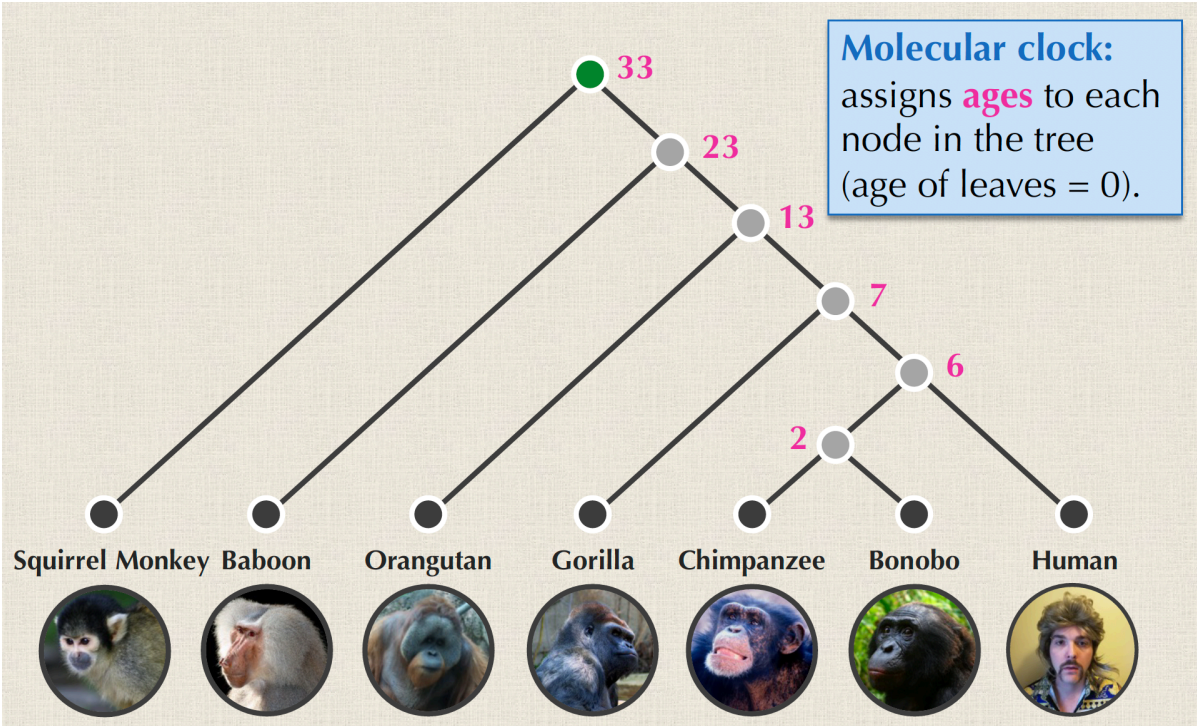


From Unrooted to Rooted Trees



From Unrooted to Rooted Trees

Our goal is to design a heuristic that models a “molecular clock” that assigns each leaf an age

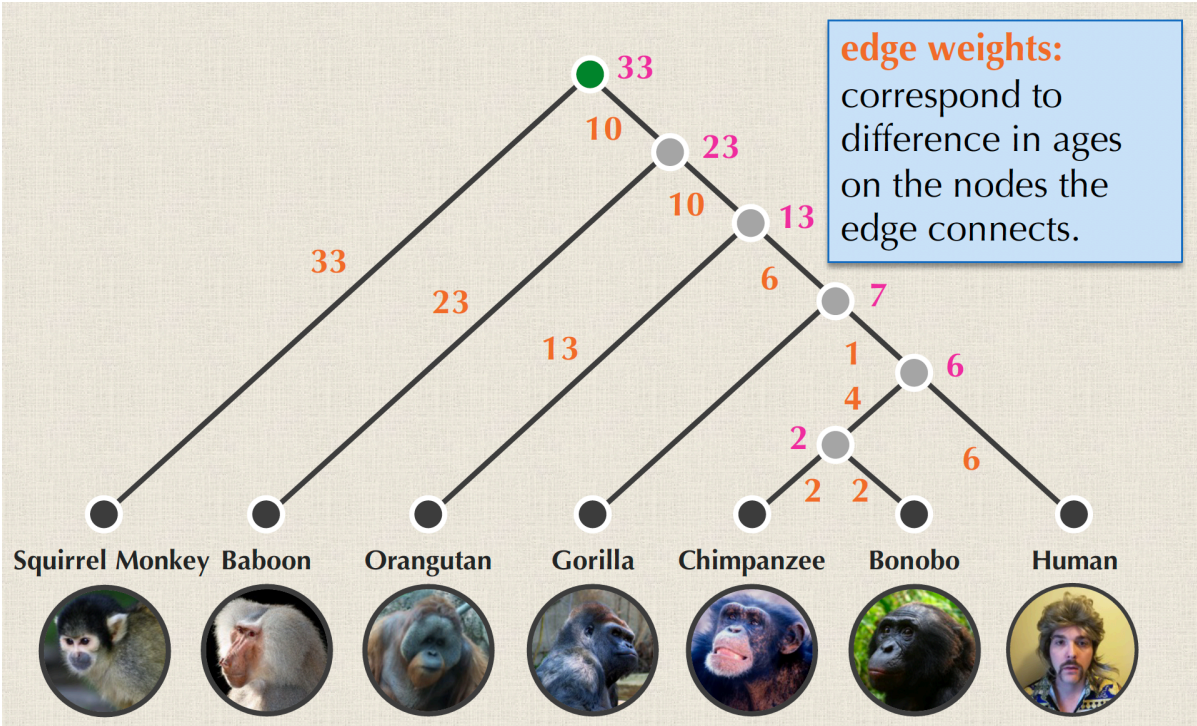


Afterwards we need to assign the ages to the internal nodes,

Which represents how long ago the speciation of that event at the internal node occurred

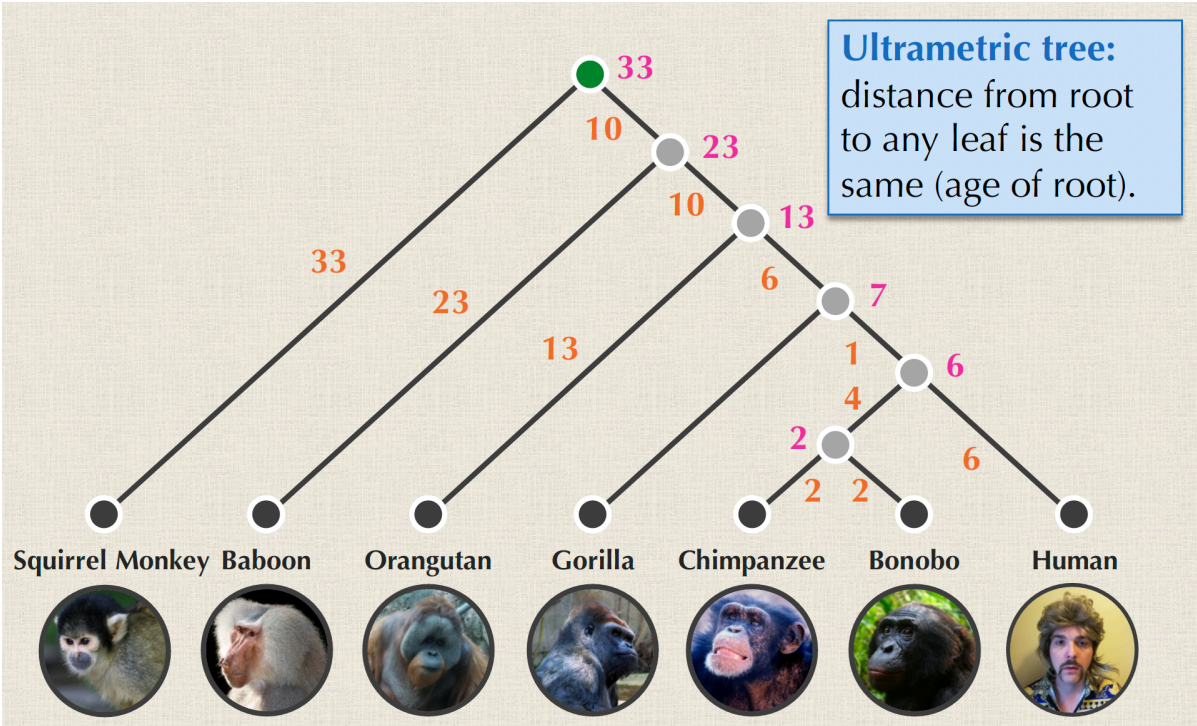
For this phylogeny, the ages are in millions of years

From Unrooted to Rooted Trees



The edge weights are computed by simply taking the distance between it's corresponding nodes

From Unrooted to Rooted Trees



This means that the length of every path from the leaf to the root is the same (33 million in the example)

UPGMA: A Clustering Heuristic

The heuristic used to construct an ultrametric, rooted binary tree is called, Unweighted Pair Group Method with Arithmetic Mean

1. Form a cluster for each present-day species, each containing a single leaf.

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0

UPGMA is the metric used to cluster the species from the distance matric into larger and larger cluster to form the evolutionary tree

UPGMA: A Clustering Heuristic

2. Find the two closest clusters C_1 and C_2 according to the minimum value in the current matrix.

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0

i **0** *j* **0** *k* **0** *l* **0**

UPGMA: A Clustering Heuristic

3. Merge C_1 and C_2 into a single cluster C.

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0

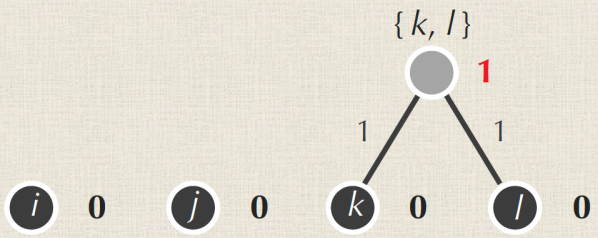
{ *k*, *l* }

i **0** *j* **0** *k* **0** *l* **0**

UPGMA: A Clustering Heuristic

4. Form a new node for C and connect to C_1 and C_2 by an edge. Set age of C as $D(C_1, C_2)/2$.

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0



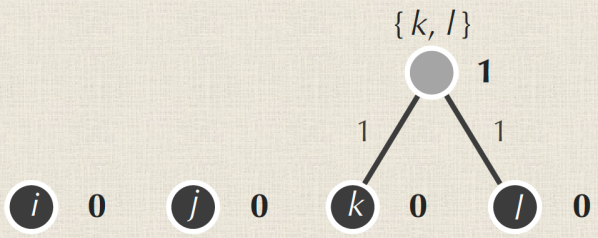
We set the age of the node as half of the distance between the two clusters

UPGMA: A Clustering Heuristic

It pairs up two closest clusters based upon the minimum average distance taken between all pairs of elements of the two clusters

5. Update the distance matrix by computing the average distance between each pair of clusters.

	i	j	$\{k, l\}$
i	0	3	3.5
j	3	0	4.5
$\{k, l\}$	3.5	4.5	0

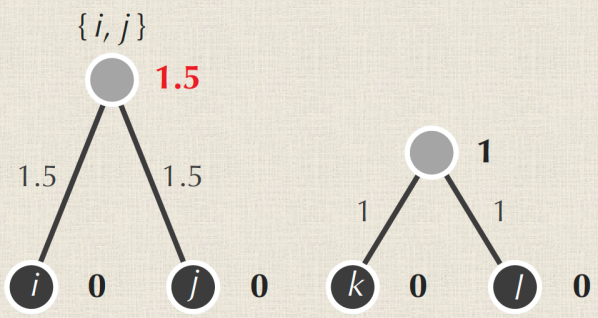


UPGMA: A Clustering Heuristic

Choose the minimum distance for the next iteration and compute the weights

6. Iterate steps 2-5 until a single cluster contains all species.

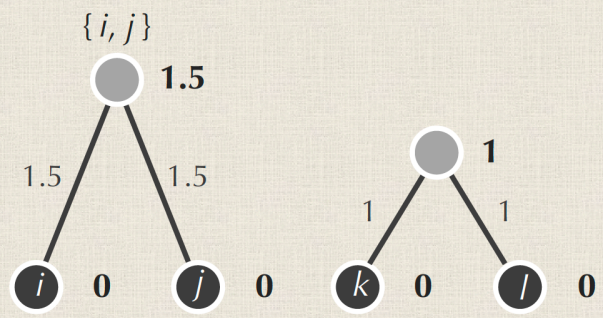
	i	j	$\{k, l\}$
i	0	**3**	3.5
j	**3**	0	4.5
$\{k, l\}$	3.5	4.5	0



UPGMA: A Clustering Heuristic

6. Iterate steps 2-5 until a single cluster contains all species.

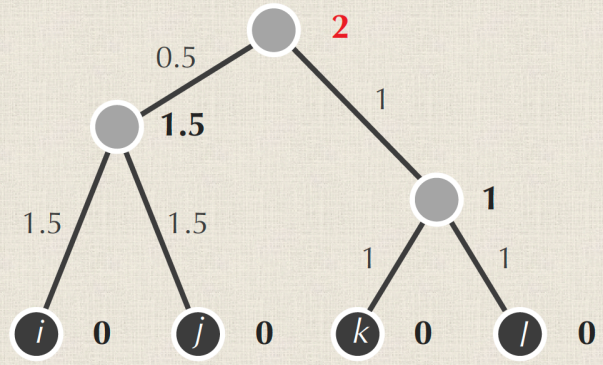
	$\{i, j\}$	$\{k, l\}$
$\{i, j\}$	0	4
$\{k, l\}$	4	0



UPGMA: A Clustering Heuristic

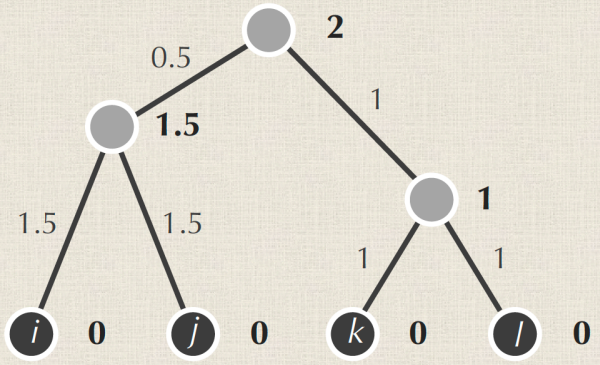
6. Iterate steps 2-5 until a single cluster contains all species.

	$\{i, j\}$	$\{k, l\}$
$\{i, j\}$	0	4
$\{k, l\}$	4	0



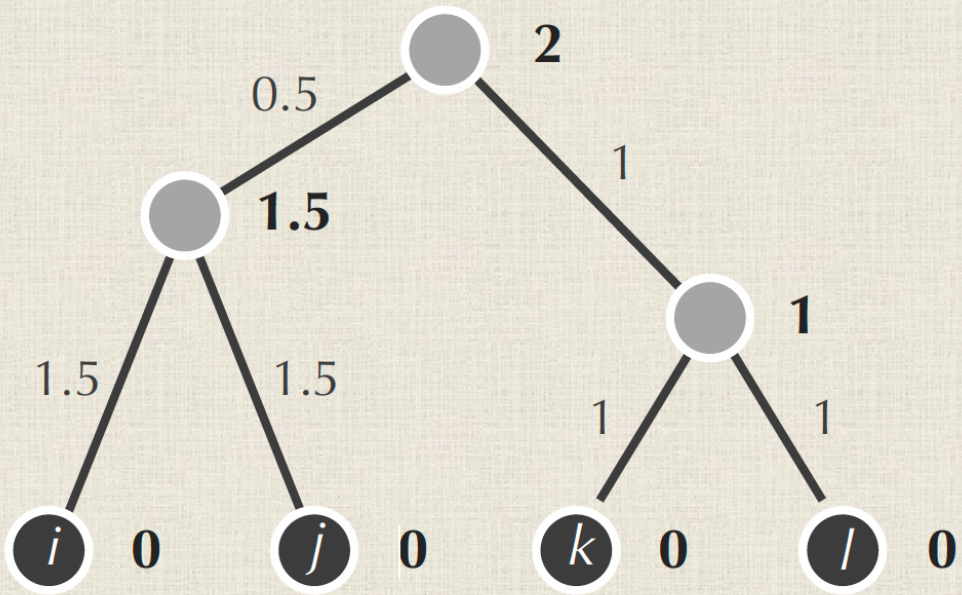
UPGMA: A Clustering Heuristic

6. Iterate steps 2-5 until a single cluster contains all species.



UPGMA Doesn't "Fit" a Tree to a Matrix

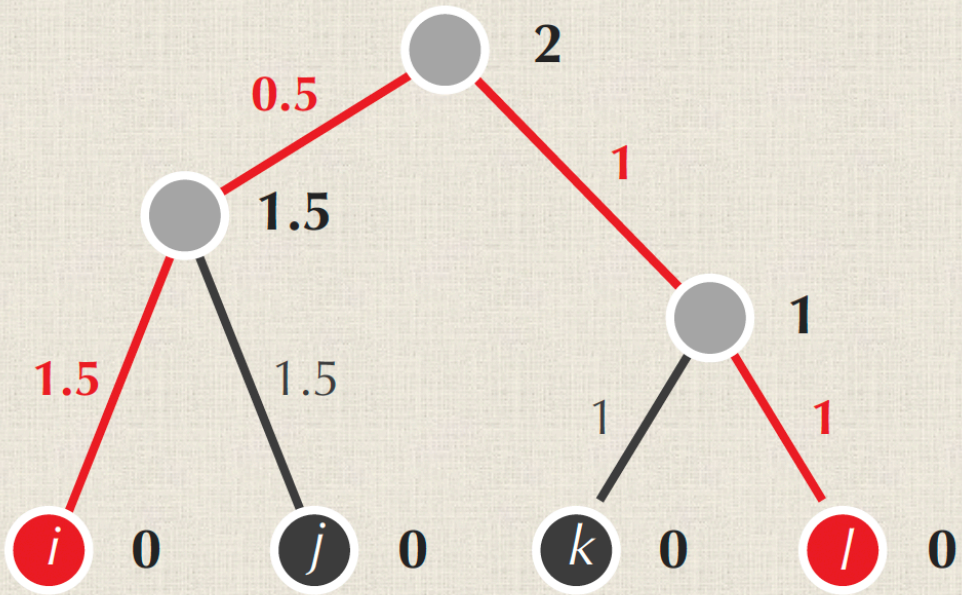
	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0



The goal of UPGMA is to construct an ultrametric evolutionary tree regardless of the distance matrix inputs

UPGMA Doesn't "Fit" a Tree to a Matrix

	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0



The distance between two species doesn't correspond to the distance in the distance matrix

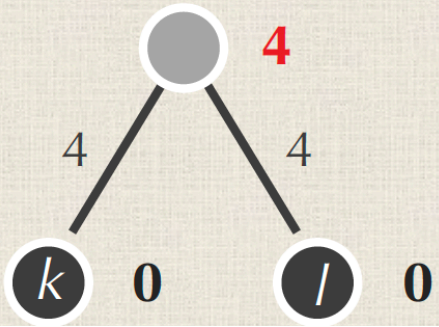
Class Activity

- Apply UPGMA to the following matrix.

	i	*j*	*k*	*l*
i	0	20	9	11
j	20	0	17	11
k	9	17	0	8
l	11	11	8	0

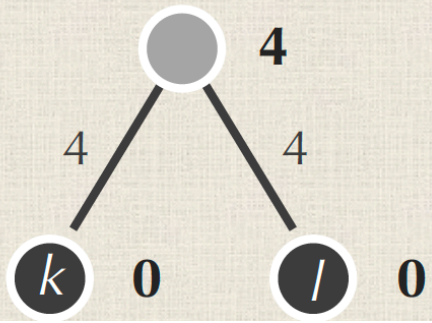
Class Activity

	i	*j*	*k*	*l*
i	0	20	9	11
j	20	0	17	11
k	9	17	0	**8**
l	11	11	**8**	0



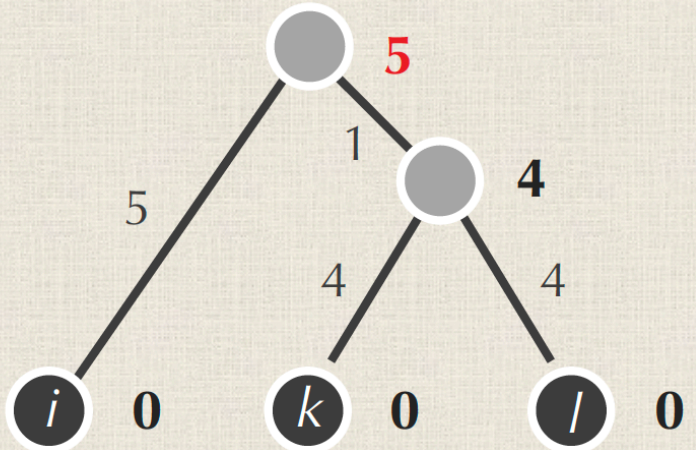
Class Activity

	i	j	$\{k, l\}$
i	0	20	10
j	20	0	14
$\{k, l\}$	10	14	0



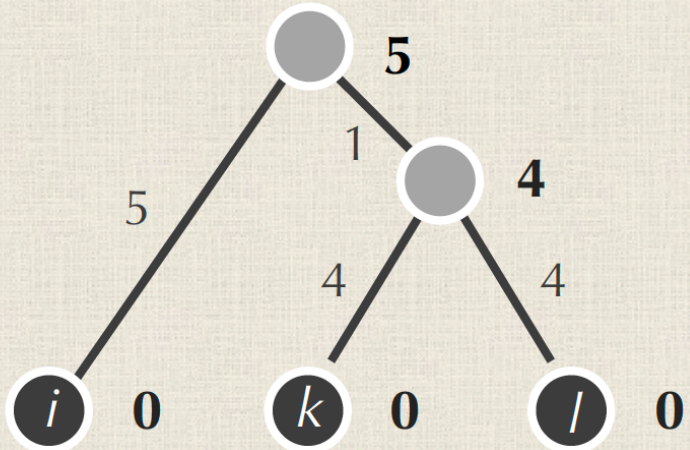
Class Activity

	i	j	$\{k, l\}$
i	0	20	**10**
j	20	0	14
$\{k, l\}$	**10**	14	0



Class Activity

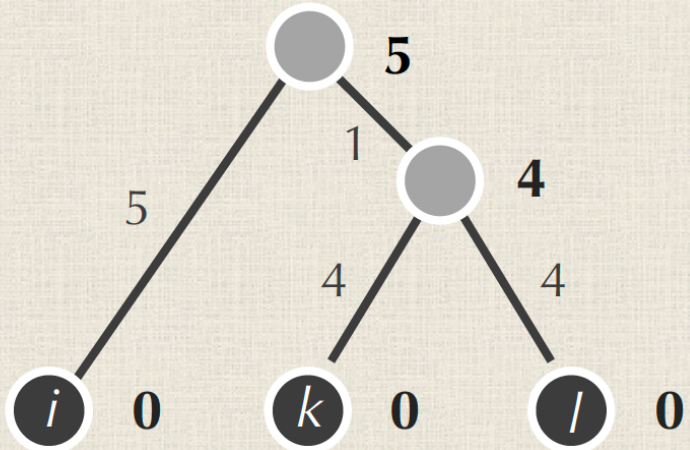
	j	$\{i, k, l\}$
j	0	17
$\{i, k, l\}$	17	0



Class Activity

This is wrong. Why?

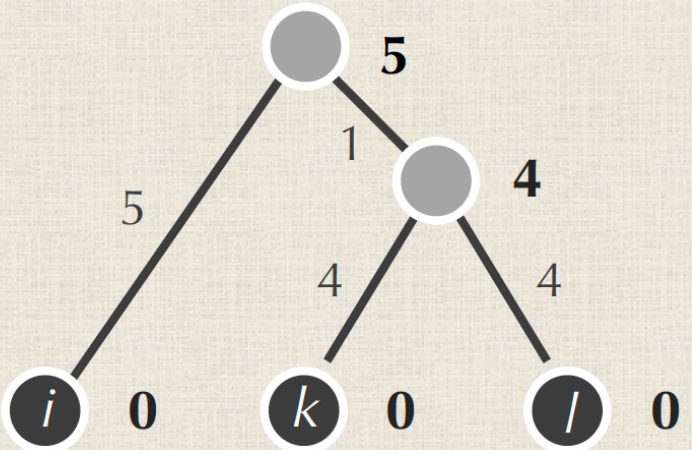
	j	$\{i, k, l\}$
j	0	17
$\{i, k, l\}$	17	0



Class Activity

Answer: The average distance from j to i, k, and l is (20+17+11)/3 = 16, not 17.

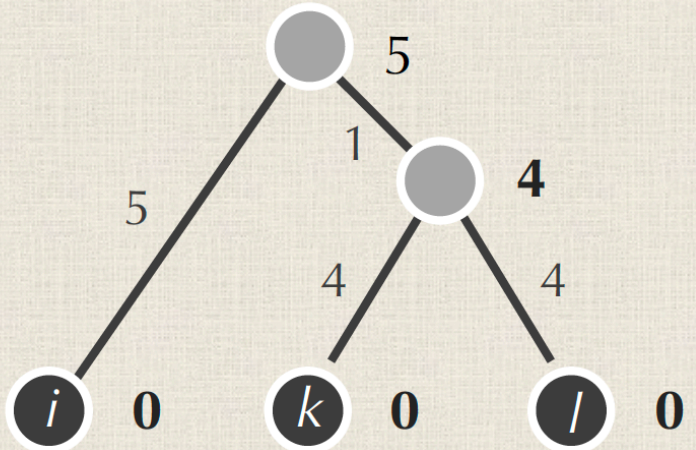
	i	*j*	*k*	*l*
i	0	**20**	9	11
j	20	0	17	11
k	9	**17**	0	8
l	11	**11**	8	0



Class Activity

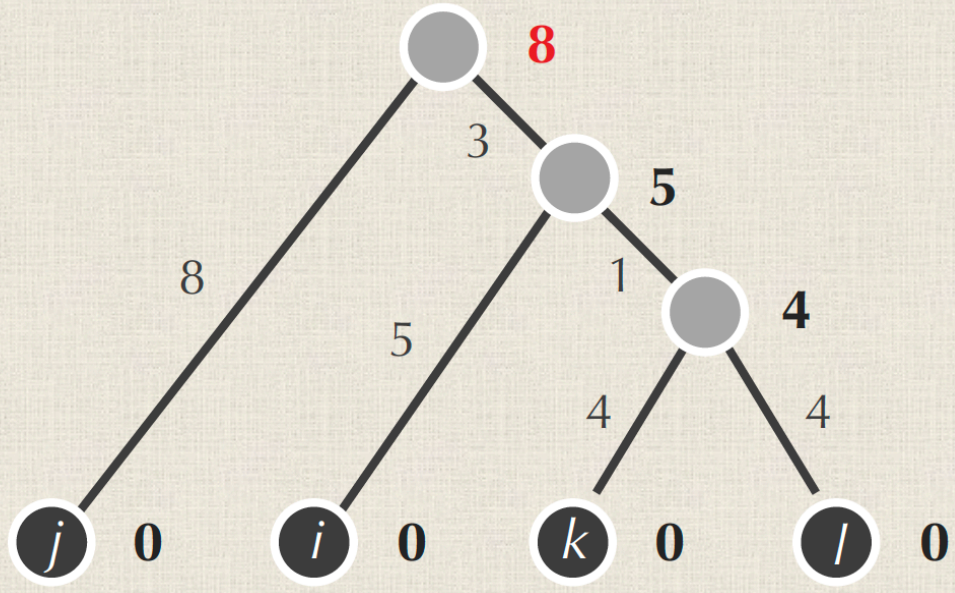
Answer: The average distance from j to i, k, and l is (20+17+11)/3 = 16, not 17.

	i	j	$\{k, l\}$
i	0	**20**	10
j	20	0	14
$\{k, l\}$	10	**14**	0

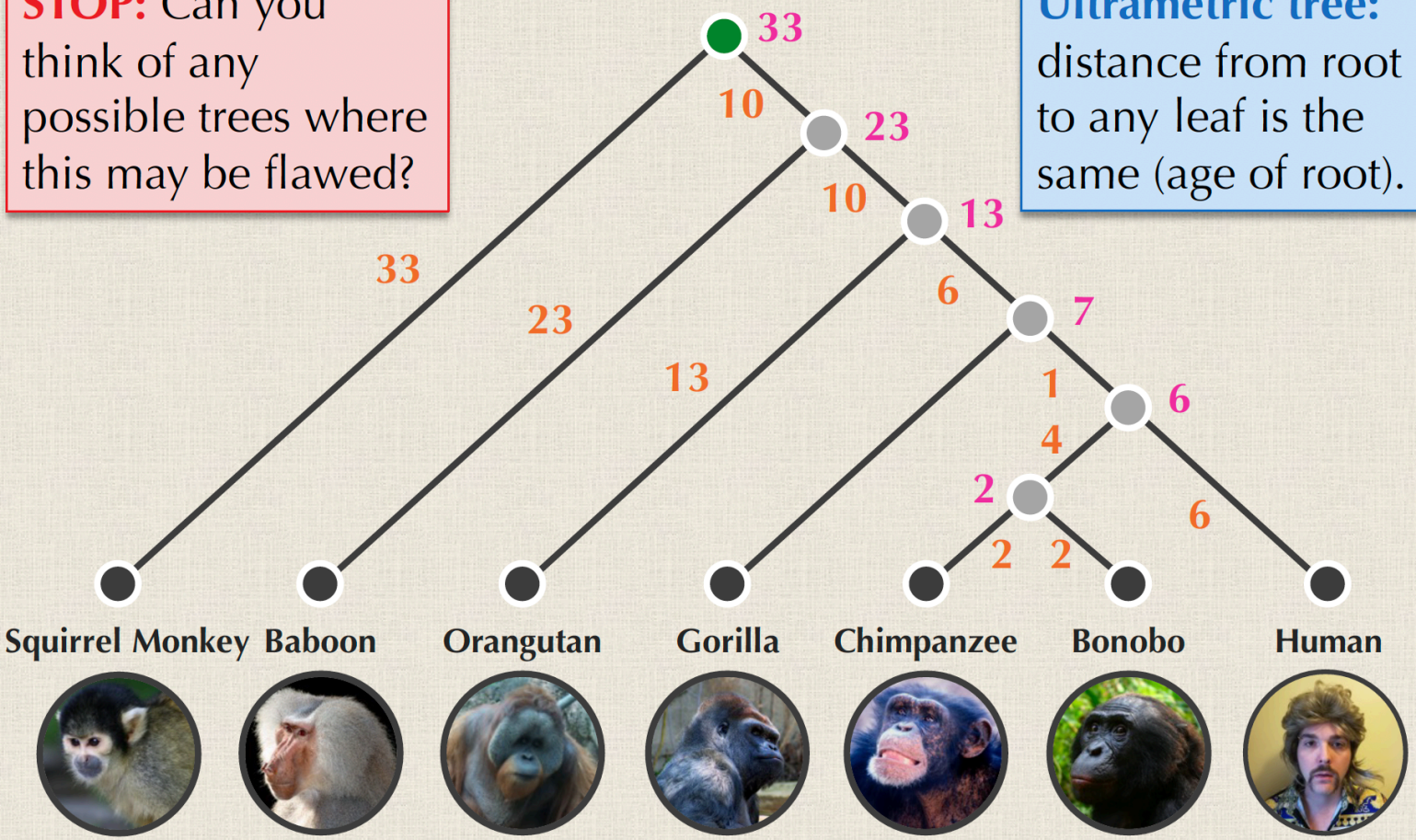


Average Distance Must be Weighted

	j	$\{i, k, l\}$
j	0	16
$\{i, k, l\}$	16	0



The Ultrametric Assumption is Very Strong

STOP: Can you think of any possible trees where this may be flawed?

Ultrametric tree: distance from root to any leaf is the same (age of root).