CS 364 COMPUTATIONAL BIOLOGY

Sara Mathieson
Haverford College

Outline

- Phylogenetic Trees
- UPGMA algorithm

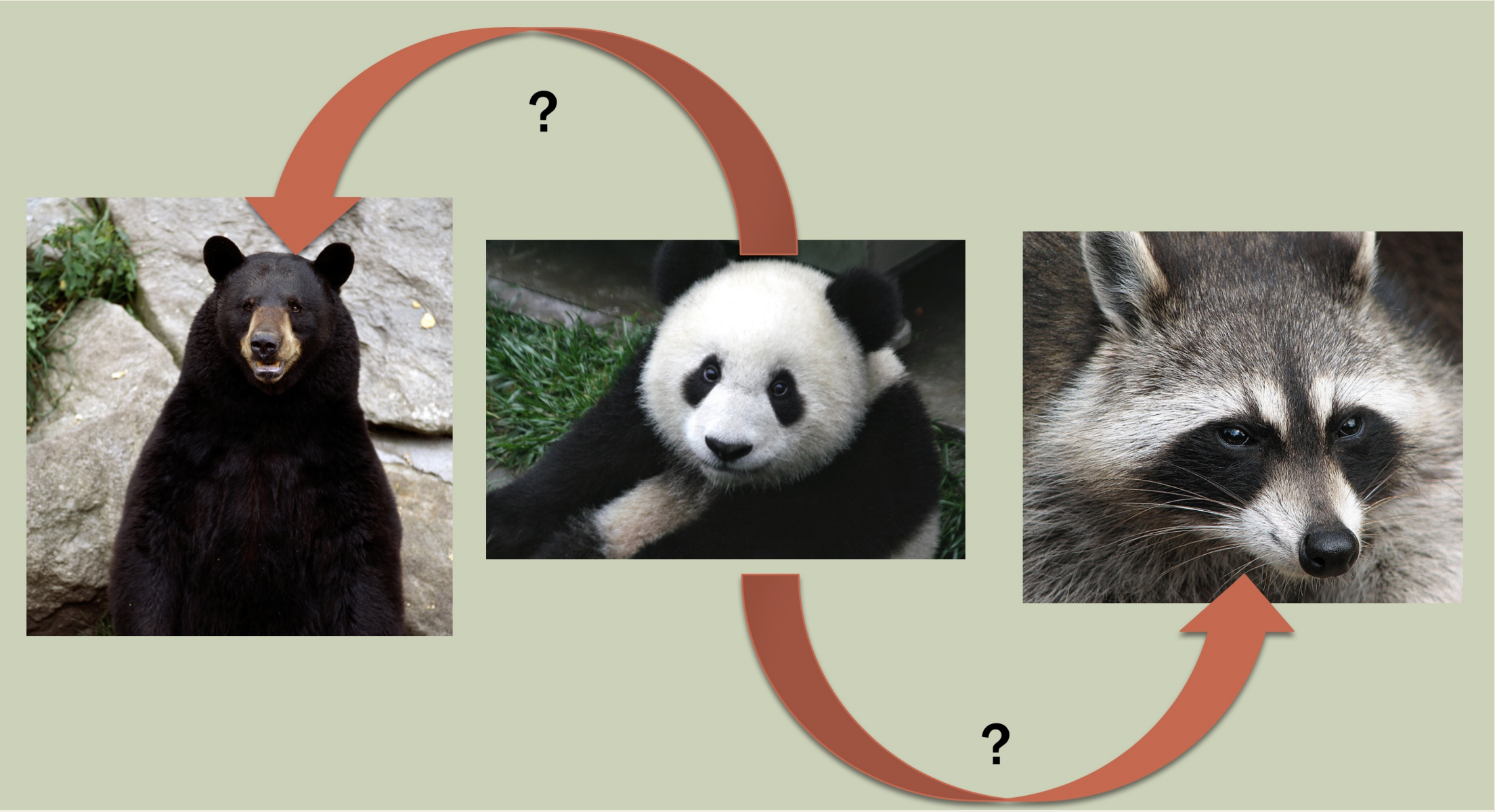
Lab 4 due tonight!
This week: phylogenetics
Next week:
* Tues: review
* Thurs: exam (in-class)

Phylogenetic Trees

Phylogenetic trees

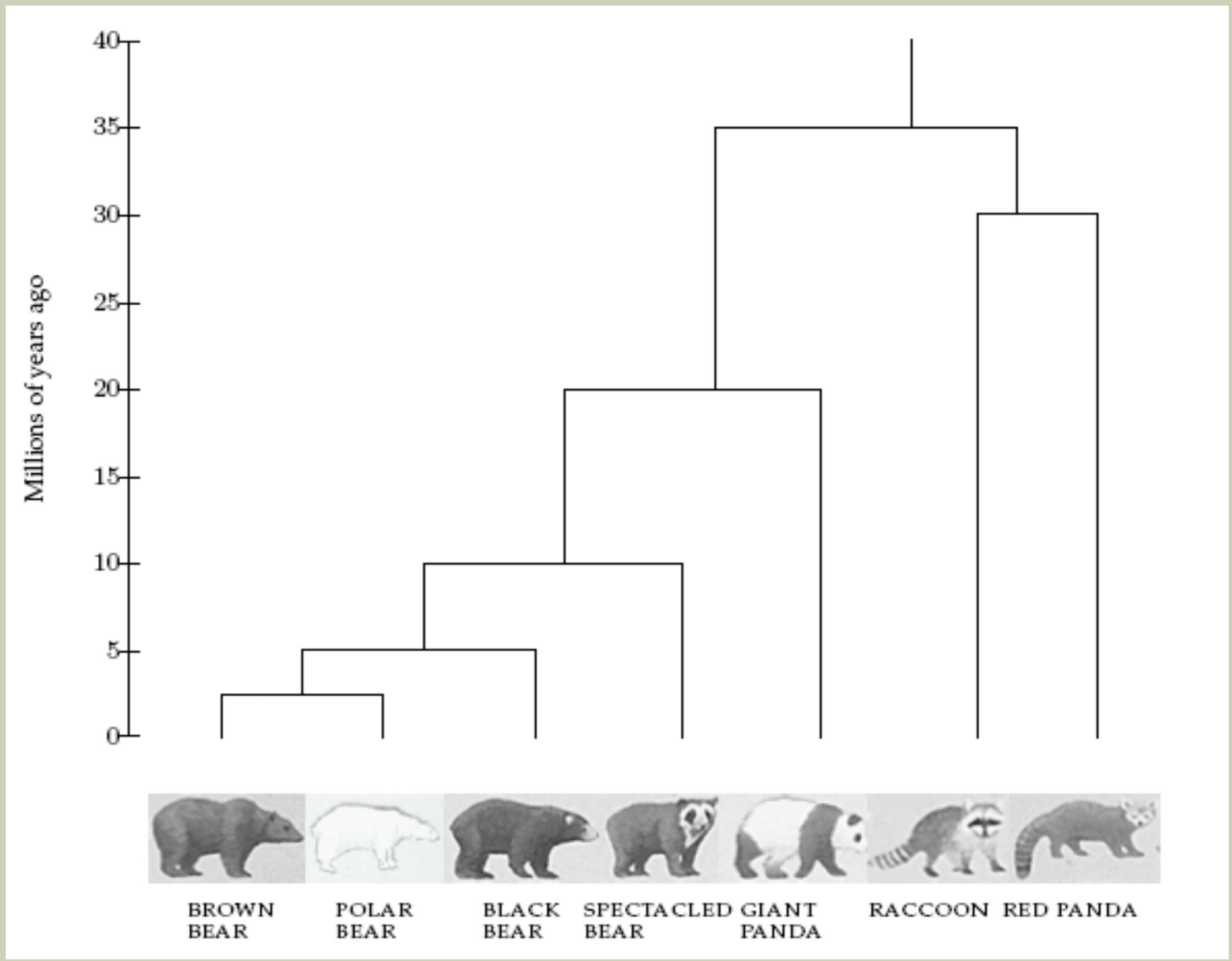
- **Definition:** diagram of inferred evolutionary relationships between samples (species, genes, individuals, etc)
- **Input:** usually genetic data, although it could be from the fossil record. Preprocessing usually involves alignment (either pairwise or multiple sequence). Then process the alignments to obtain the number of pairwise differences or another form of “dissimilarity”
- **Output:** tree structure PLUS branch lengths which represent time
- **We can learn:** evolutionary history! Sequence of speciation events, function and evolution of common traits and genes, biology of common ancestors, tempo and mode of mutation, natural selection, recombination, migration, population size changes

Great Panda Mystery



Credit: Ameet Soni

Phylogenetic tree of bears and raccoons



Credit: Ameet Soni

Recap + extensions (discuss with a partner)

1) In phylogenetic trees, observed sequences usually exist at the: (a) root of the tree, (b) internal nodes of the tree, (c) leaves of the tree, (d) all of the above

2) Why do we make the assumption that alleles at each site can be encoded as 0's and 1's?

3) Is the reference sequence always the ancestral sequence?

Recap + extensions (discuss with a partner)

1) In phylogenetic trees, observed sequences usually exist at the: (a) root of the tree, (b) internal nodes of the tree, (c) leaves of the tree, (d) all of the above

Answer: Usually at the leaves (*sometimes* we can get ancient DNA)

2) Why do we make the assumption that alleles at each site can be encoded as 0's and 1's?

3) Is the reference sequence always the ancestral sequence?

Recap + extensions (discuss with a partner)

1) In phylogenetic trees, observed sequences usually exist at the: (a) root of the tree, (b) internal nodes of the tree, (c) leaves of the tree, (d) all of the above

   Answer: Usually at the leaves (*sometimes* we can get ancient DNA)

2) Why do we make the assumption that alleles at each site can be encoded as 0's and 1's?

   Answer: multiple mutations at the same site are very rare (most sites are therefore biallelic, not triallelic)

3) Is the reference sequence always the ancestral sequence?

Recap + extensions (discuss with a partner)

1) In phylogenetic trees, observed sequences usually exist at the: (a) root of the tree, (b) internal nodes of the tree, (c) leaves of the tree, (d) all of the above

   Answer: Usually at the leaves (*sometimes* we can get ancient DNA)

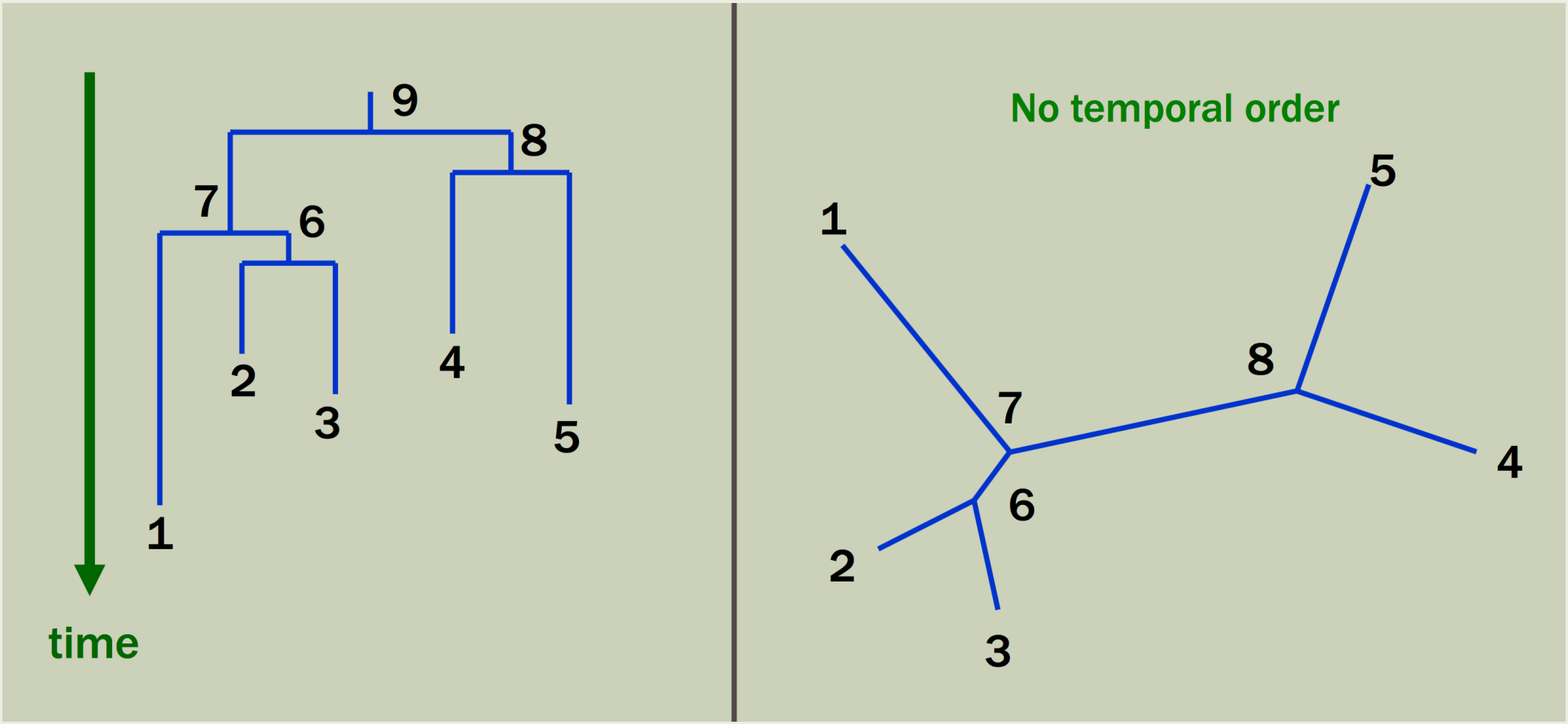
2) Why do we make the assumption that alleles at each site can be encoded as 0's and 1's?

   Answer: multiple mutations at the same site are very rare (most sites are therefore biallelic, not triallelic)

3) Is the reference sequence always the ancestral sequence?

   Answer: No! usually not. The reference happened to be sequenced first. Most of the time we don't know the ancestral sequence, but phylogenetic trees can help us reconstruct it.

Rooted vs. unrooted trees



Credit: Ameet Soni

Dissimilarity maps

How to measure relationships between taxa?

Define a dissimilarity map $\boldsymbol{\delta}$(x,y) between any two taxa

e.g. for DNA sequence, we might just count the number of differences after alignment:

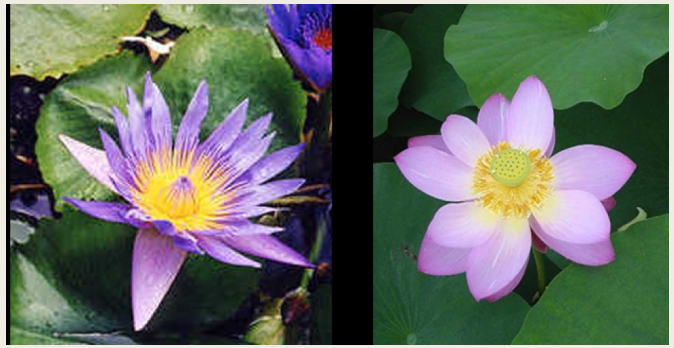
```
x=AAGTAGATATAGATAGATATTACATTA
y=AAGTAGTTATAGATACCTATTAGATTA
```

$\boldsymbol{\delta}$(x,y)=4

Represent as a symmetric matrix

	A	B	C	D
A	0	3	1	1
B	3	0	2	1
C	1	2	0	3
D	1	1	3	0

Anything else, that's really up to you (i.e. image metric?)



Dissimilarity maps

- Record pairwise differences (which could be obtained from a pairwise sequence alignment)
- We will use a dissimilarity map as input to our phylogenetic tree algorithms

A *dissimilarity map* δ is a function mapping pairs of samples from a set $\mathcal{X}$ to distances. It has the following two properties, but not necessarily the triangle inequality.

1. $\delta(x, x) = 0$
2. $\delta(x, y) = \delta(y, x)$

Example:

δ	A	B	C	D	E
A	0	1	3	6	6
B		0	2	5	5
C			0	5	5
D				0	2
E					0

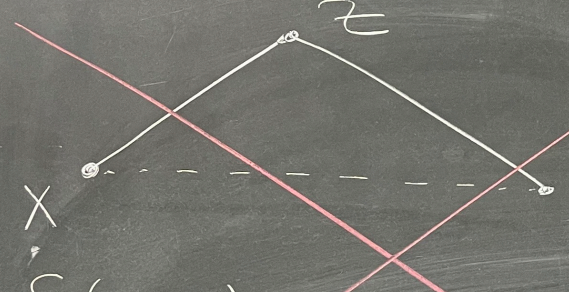
$X = \{A, B, C, D, E\}$

Δ	$\{A\}$	$\{B\}$	$\{C\}$	$\{D\}$	$\{E\}$
$\{A\}$	0	1	3	6	6
$\{B\}$		0	2	5	5
$\{C\}$			0	5	5
$\{D\}$				0	2
$\{E\}$					0

dissimilarity map

① $\delta(x,x) = 0$

② $\delta(x,y) = \delta(y,x)$



triangle inequality

$$\cancel{\delta(x,y) \leq \delta(x,z) + \delta(z,y)}$$

not necessary for dissimilarity map

ultrametric

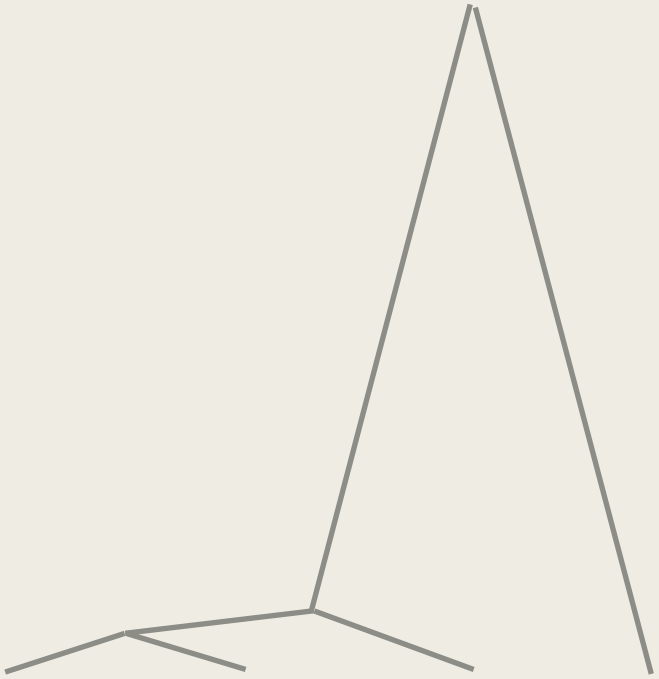
"stronger" triangle inequality

$$\delta(a,b) \leq \max\{\delta(a,c), \delta(b,c)\}$$

Ultrametric trees

Rooted Trees where all the leaves are the same distance from the root

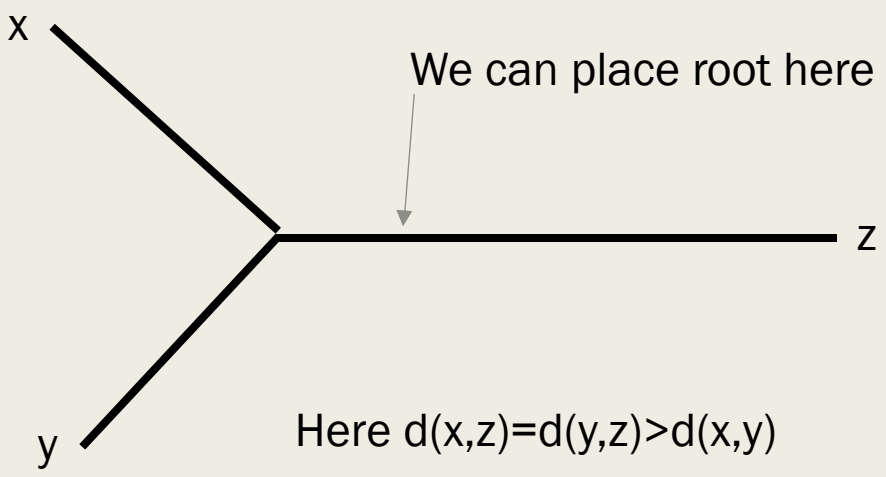
Tree with edge weights

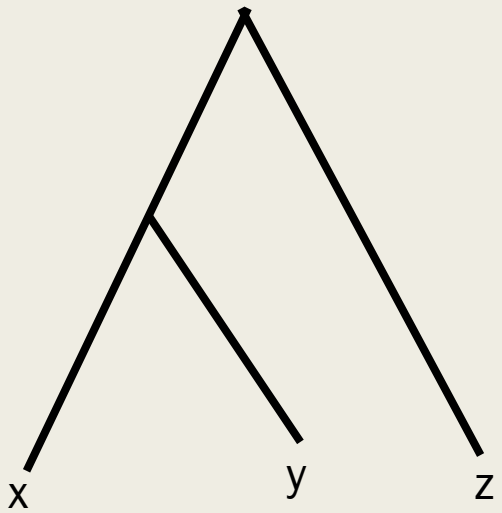


Ultrametric (“molecular clock”) tree

“3-point” condition

The ultrametric condition is that, for every set of three taxa {x,y,z}, of the three pairwise distances {d(x,y) d(y,x), d(y,z)}, two are equal and one is less than or equal to the other two.





1. For every unrooted ultrametric tree, there is a unique place to put the root
2. For every ultrametric distance matrix, there is a unique rooted tree

Biological interpretation of ultrametric trees

- Evolution happens at the same rate on every branch
- Is this plausible?

Biological interpretation of ultrametric trees

- Evolution happens at the same rate on every branch
- Is this plausible?
- Maybe approximately on short scales (within species)
- Not really on longer scales (e.g. apes)
- Or in other contexts (linguistics)

UPGMA

(Unweighted Pair Group Method with Arithmetic mean)

Sokal & Michener 1958

UPGMA

Greedy, bottom-up clustering method

Given a dissimilarity map $\boldsymbol{\delta}$, produces a rooted, ultrametric tree

If the dissimilarity map $\boldsymbol{\delta}$ is ultrametric, then UPGMA is guaranteed to reconstruct the correct tree

[i.e. if $\boldsymbol{\delta}$ is not ultrametric, UPGMA will still produce an ultrametric tree, but it might have the wrong topology]

UPGMA

δ	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	6	6	6	-		
E	6	6	6	4	-	
F	8	8	8	8	8	-

A B C D E F

Set Δ equal to δ to start

UPGMA

Δ	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	6	6	6	-		
E	6	6	6	4	-	
F	8	8	8	8	8	-

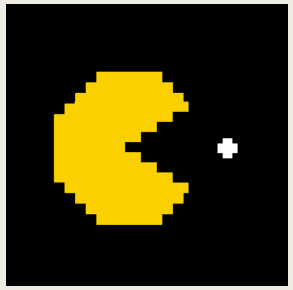
A B C D E F

Set Δ equal to $\boldsymbol{\delta}$ to start

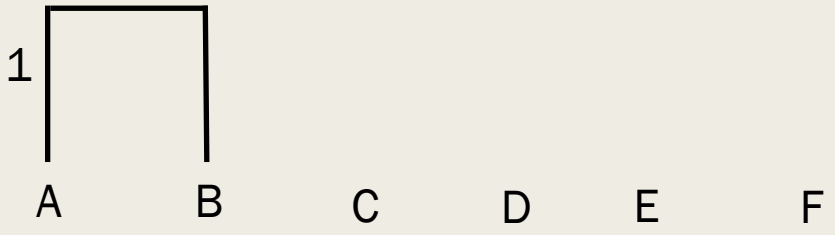
UPGMA

1. Pick the closest taxa x,y and join them at height Δ(x,y)/2

	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	6	6	6	-		
E	6	6	6	4	-	
F	8	8	8	8	8	-



This is the greedy bit!



UPGMA

2. Calculate new distance matrix by averaging over distances

Δ	A	B	C	D	E	F
A	-					
B	2	-				
C	4	4	-			
D	6	6	6	-		
E	6	6	6	4	-	
F	8	8	8	8	8	-

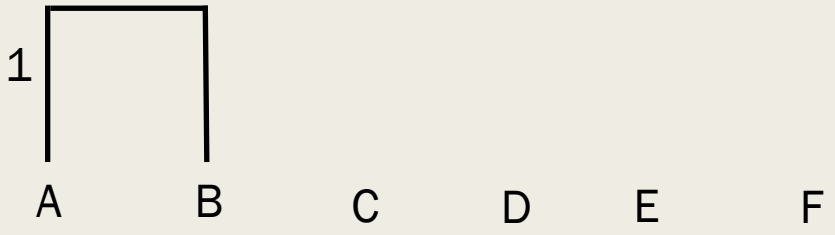
$\Delta(AB,C)=\{\Delta(A,C)+\Delta(B,C)\}/2$

$\Delta(AB,D)=\{\Delta(A,D)+\Delta(B,D)\}/2$

$\Delta(AB,E)=\{\Delta(A,E)+\Delta(B,E)\}/2$

$\Delta(AB,F)=\{\Delta(A,F)+\Delta(B,F)\}/2$

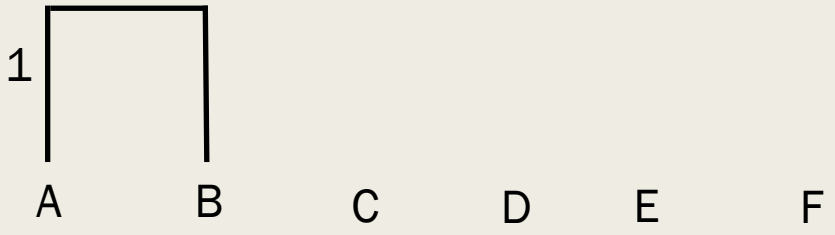
Δ	AB	C	D	E	F
AB	-				
C	4	-			
D	6	6	-		
E	6	6	4	-	
F	8	8	8	8	-



UPGMA

2. Calculate new distance matrix by averaging over distances

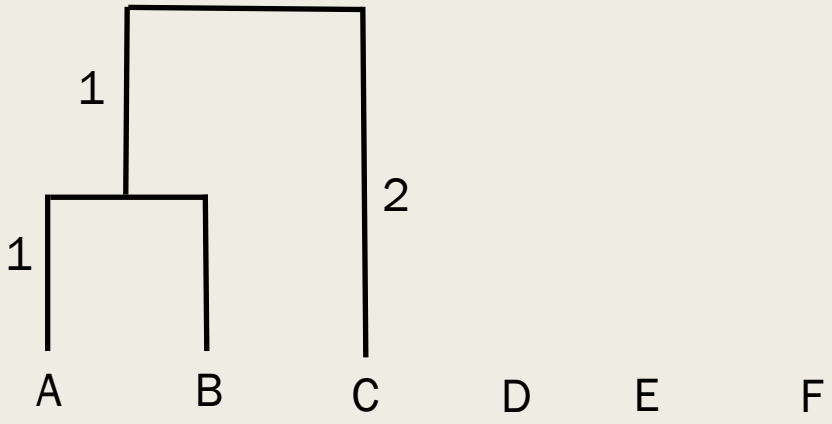
Δ	AB	C	D	E	F
AB	-				
C	4	-			
D	6	6	-		
E	6	6	4	-	
F	8	8	8	8	-



UPGMA

1. Pick the closest taxa x,y and join them at height Δ(x,y)/2

	AB	C	D	E	F
AB	-				
C	4	-			
D	6	6	-		
E	6	6	4	-	
F	8	8	8	8	-



UPGMA

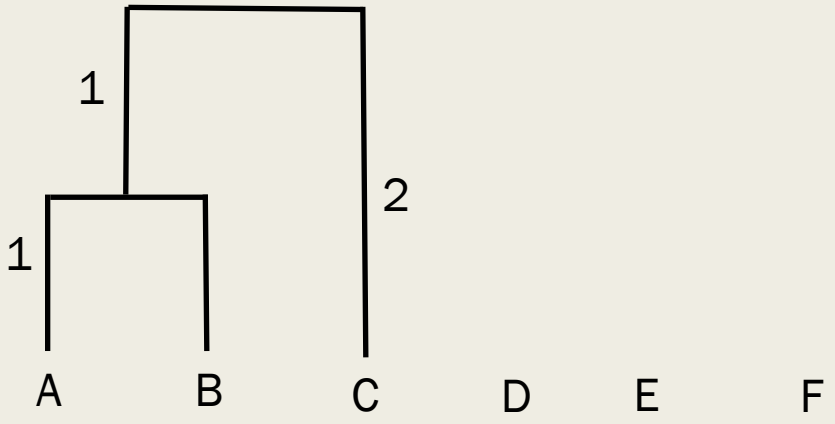
1. Pick the closest taxa x,y and join them at height Δ(x,y)/2

	AB	C	D	E	F
AB	-				
C	4	-			
D	6	6	-		
E	6	6	4	-	
F	8	8	8	8	-

	ABC	D	E	F
ABC	-			
D	6	-		
E	6	4	-	
F	8	8	8	-

Δ(ABC, D)={2 Δ(AB,D)+ Δ(C,D)}/3

Etc..



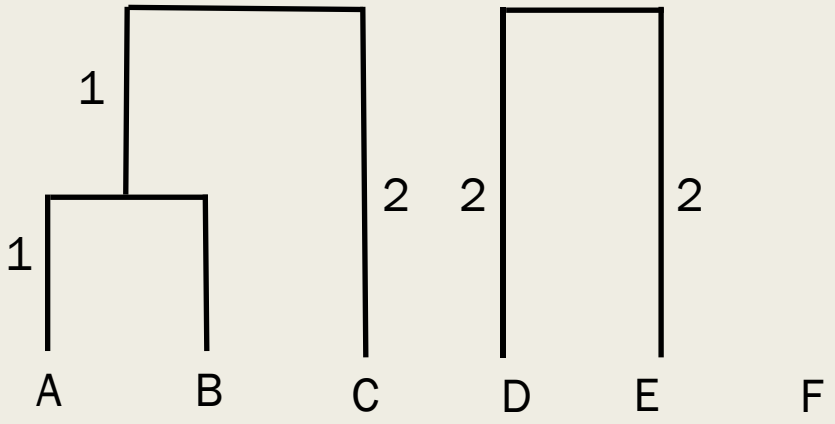
UPGMA

1. Pick the closest taxa x,y and join them at height Δ(x,y)/2
2. Calculate new distance matrix by averaging over distances

	ABC	DE	F
ABC	-		
DE	6	-	
F	8	8	-

Δ(ABC,DE)={Δ(ABC,D)+ Δ(ABC,E)}/2

Δ(F,DE)={Δ(F,D)+ Δ(F,E)}/2



UPGMA

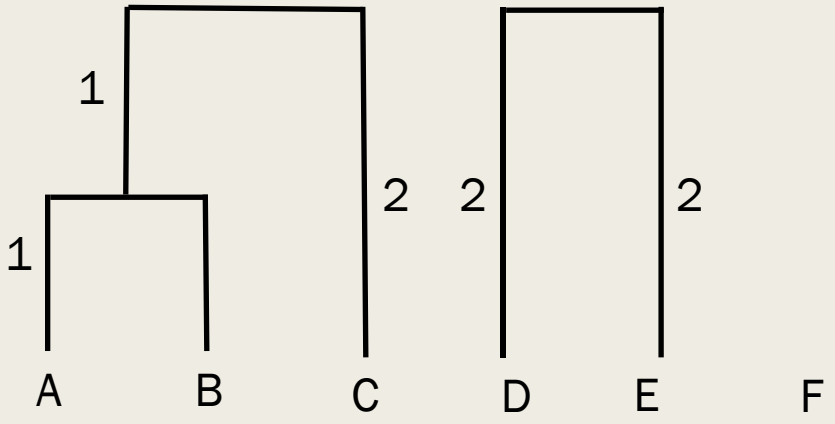
1. Pick the closest taxa x,y and join them at height Δ(x,y)/2
2. Calculate new distance matrix by averaging over distances

	ABC	D	E	F
ABC	-			
D	6	-		
E	6	4	-	
F	8	8	8	-

	ABC	DE	F
ABC	-		
DE	6	-	
F	8	8	-

Δ(ABC,DE)={Δ(ABC,D)+ Δ(ABC,E)}/2

Δ(F,DE)={Δ(F,D)+ Δ(F,E)}/2

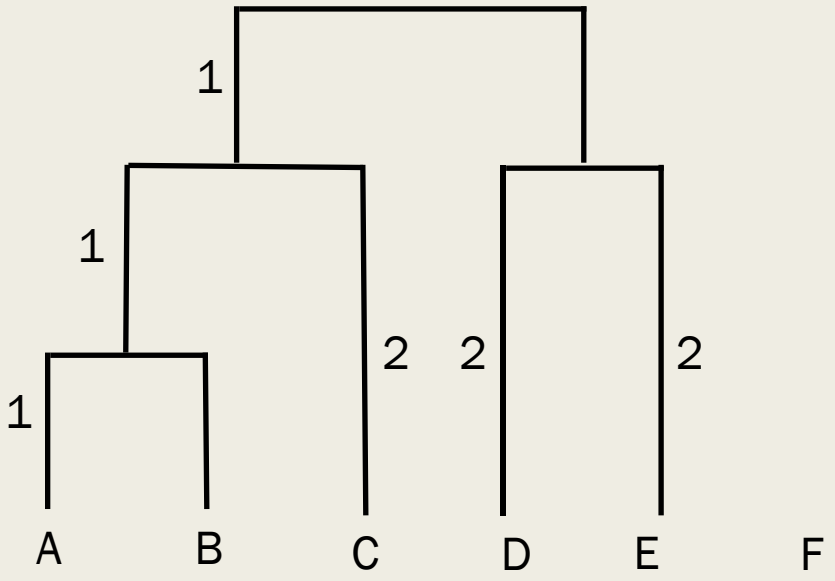


UPGMA

1. Pick the closest taxa x,y and join them at height Δ(x,y)/2
2. Calculate new distance matrix by averaging over distances

	ABC	DE	F
ABC	-		
DE	6	-	
F	8	8	-

	ABCDE	F
ABCDE	-	
F	8	-

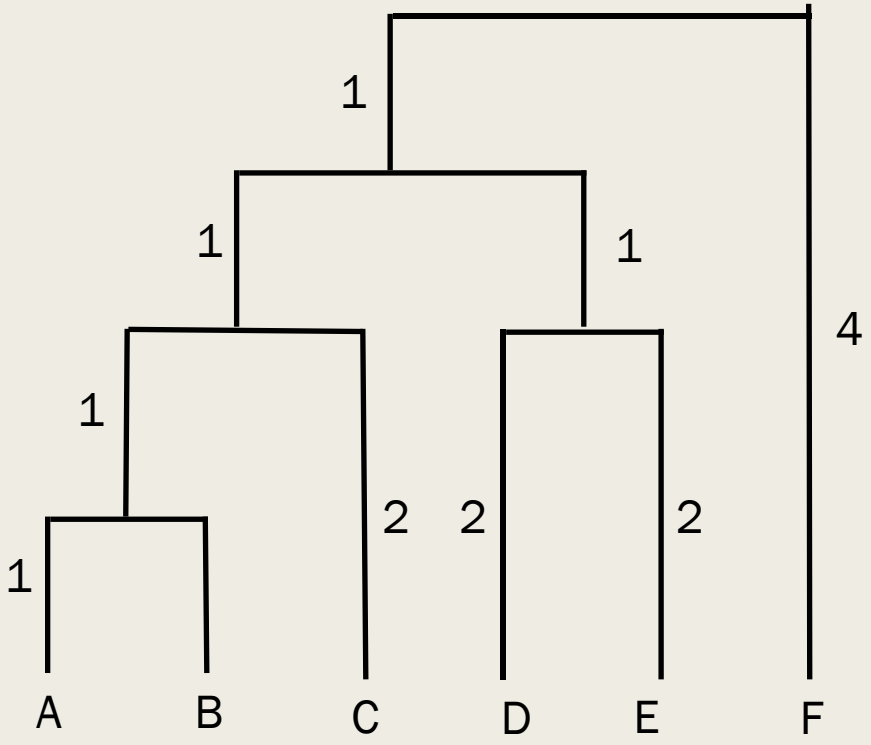


Δ(ABCDE, F)={3 Δ(ABC,F)+2 Δ(DE,F)}/5

UPGMA

1. Pick the closest taxa x,y and join them at height Δ(x,y)/2
2. Calculate new distance matrix by averaging over distances

	ABCDE	F
ABCDE	-	
F	8	-



UPGMA algorithm

$X = \{$set of taxa$\}$

initialization:

- each sample $x \in X$ is its own cluster $C_x = \{x\}$
- map $\Delta(C_i, C_j) = \delta(i, j)$

update rule (iteration)

1. find C_i & C_j that minimize $\Delta(C_i, C_j)$ & merge to create

$$\boxed{C_{ij} = C_i \cup C_j}$$

2. $\Delta($

3. **terminat**

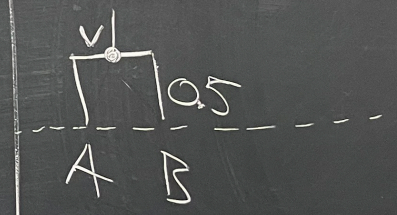
taxa}

(2) set distance from C_{ij} to C_k

$$\Delta(C_{ij}, C_k) = \frac{|C_i|}{|C_i|+|C_j|}\Delta(C_i, C_k) + \frac{C_j}{|C_i|+|C_j|}\Delta(C_j, C_k)$$

(3) join C_i & C_j with vertex v,
set height of v equal to

$$\frac{\Delta(C_i, C_j)}{2}$$



<u>termination</u>

Stop when you have one cluster

UPGMA algorithm

UPGMA initialization:

1. Each sample $x \in \mathcal{X}$ starts in it's own cluster $C_x = \{x\}$
2. Set cluster distances $\Delta(C_i, C_j) = \delta(i, j)$ for all i, j

UPGMA algorithm

UPGMA initialization:

1. Each sample $x \in \mathcal{X}$ starts in it's own cluster $C_x = \{x\}$
2. Set cluster distances $\Delta(C_i, C_j) = \delta(i, j)$ for all i, j

UPGMA update:

1. Find C_i and C_j (where $i \neq j$) that minimize $\Delta(C_i, C_j)$, and merge to create $C_{ij} = C_i \cup C_j$
2. Set the distances from C_{ij} to every other cluster C_k using the update rule:

$$\Delta(C_i \cup C_j, C_k) = \frac{|C_i|}{|C_i| + |C_j|}\Delta(C_i, C_k) + \frac{|C_j|}{|C_i| + |C_j|}\Delta(C_j, C_k)$$

3. Join C_i and C_j with interior vertex v; set the height of v equal to $\Delta(C_i, C_j)/2$

Tree metric

- UPGMA induces not only a *tree metric* but also an *ultrametric* on the samples in ***X***

A dissimilarity map δ is a *tree metric* if $\exists$ a tree topology and edges weights such that $\forall\ x, y \in \mathcal{X}$,

$$\delta(x, y) = \sum \text{all edge weights in the path from } x \text{ to } y.$$

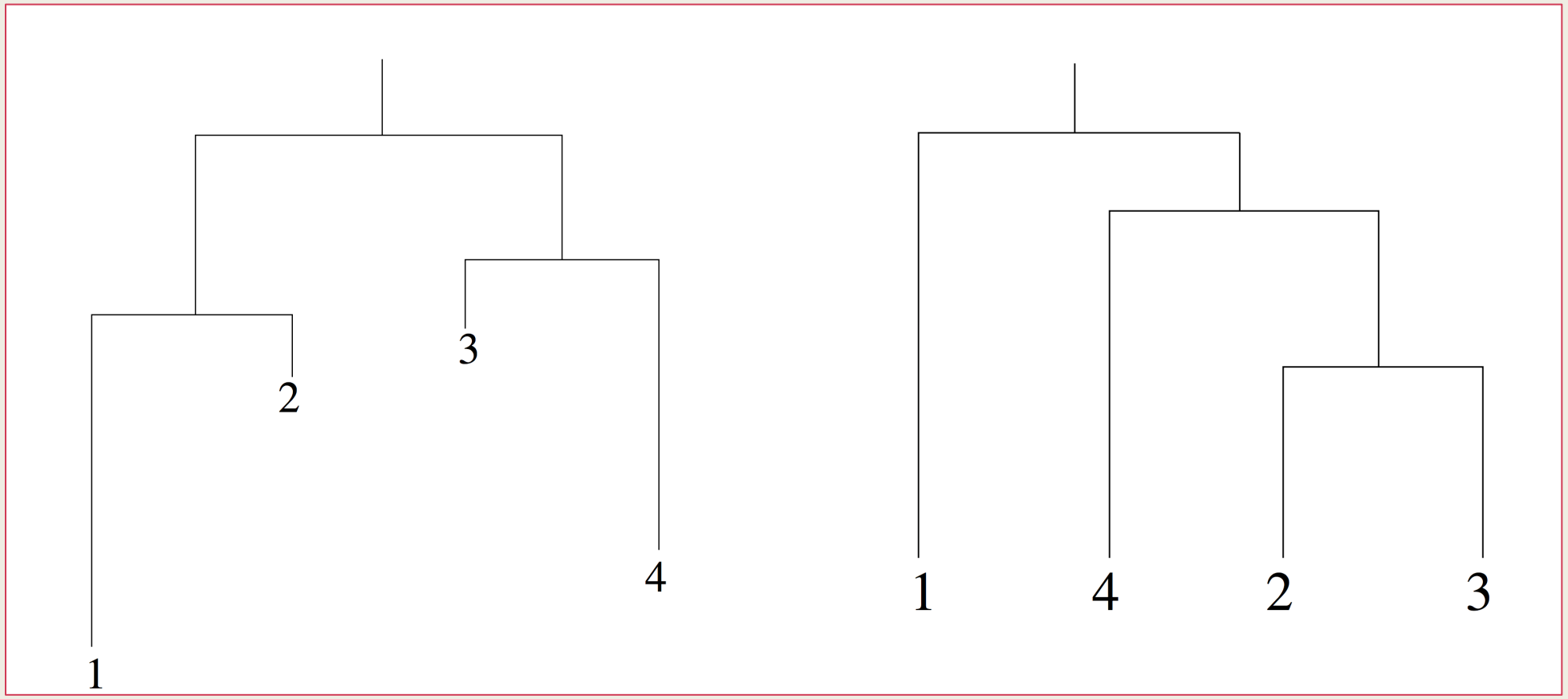
Goal minimize the difference between δ (orig dissimilarity map) and δ' (induced " ")

$$\min_{\delta'} J(\delta, \delta') = \sum_{\substack{\{i,j\} \in X \\ i \neq j}} [\delta(i,j) - \delta'(i,j)]^2$$

$\Rightarrow$ $\boxed{\text{NP-complete!}}$

UPGMA & NJ are heuristic methods!

UPGMA can produce unrealistic trees

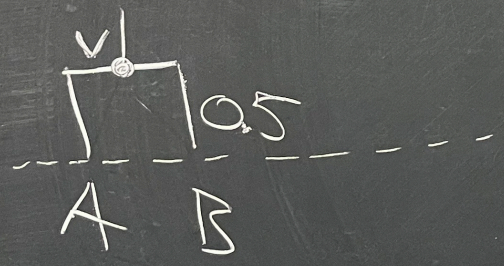


Closest tree to input data

UPGMA tree

$\Delta(C_j, C_k)$

Δ	$\{A,B\}$	$\{C\}$	$\{D\}$	$\{E\}$
$\{A,B\}$	0	$\boxed{2.5}$	$\boxed{5.5}$	$\boxed{5.5}$
$\{C\}$		0	5	5
$\{D\}$			0	2
$\{E\}$				0



Handout 9, page 1

$$\Delta(C_A \cup C_B, C_C) = \frac{1}{1+1}\underbrace{\Delta(C_A, C_C)}_{3} + \frac{1}{1+1}\underbrace{\Delta(C_B, C_C)}_{2}$$

$$= \frac{1}{2}\cdot 3 + \frac{1}{2}\cdot 2 = 2.5$$

$$\Delta(C_A \cup C_B, C_D) = \frac{1}{2}6 + \frac{1}{2}5 = 5.5$$

$$\Delta(C_A \cup C_B, C_E) = \frac{1}{2}6 + \frac{1}{2}5 = 5.5$$

	$\{A,B\}$	$\{C\}$	$\{D,E\}$
$\{A,B\}$	0	2.5	$\boxed{5.5}$
$\{C\}$		0	$\boxed{5}$
$\{D,E\}$			0

$$\Delta(C_D \cup C_E, C_{AB}) = \frac{1}{2}(5.5) + \frac{1}{2}(5.5) = 5.5$$

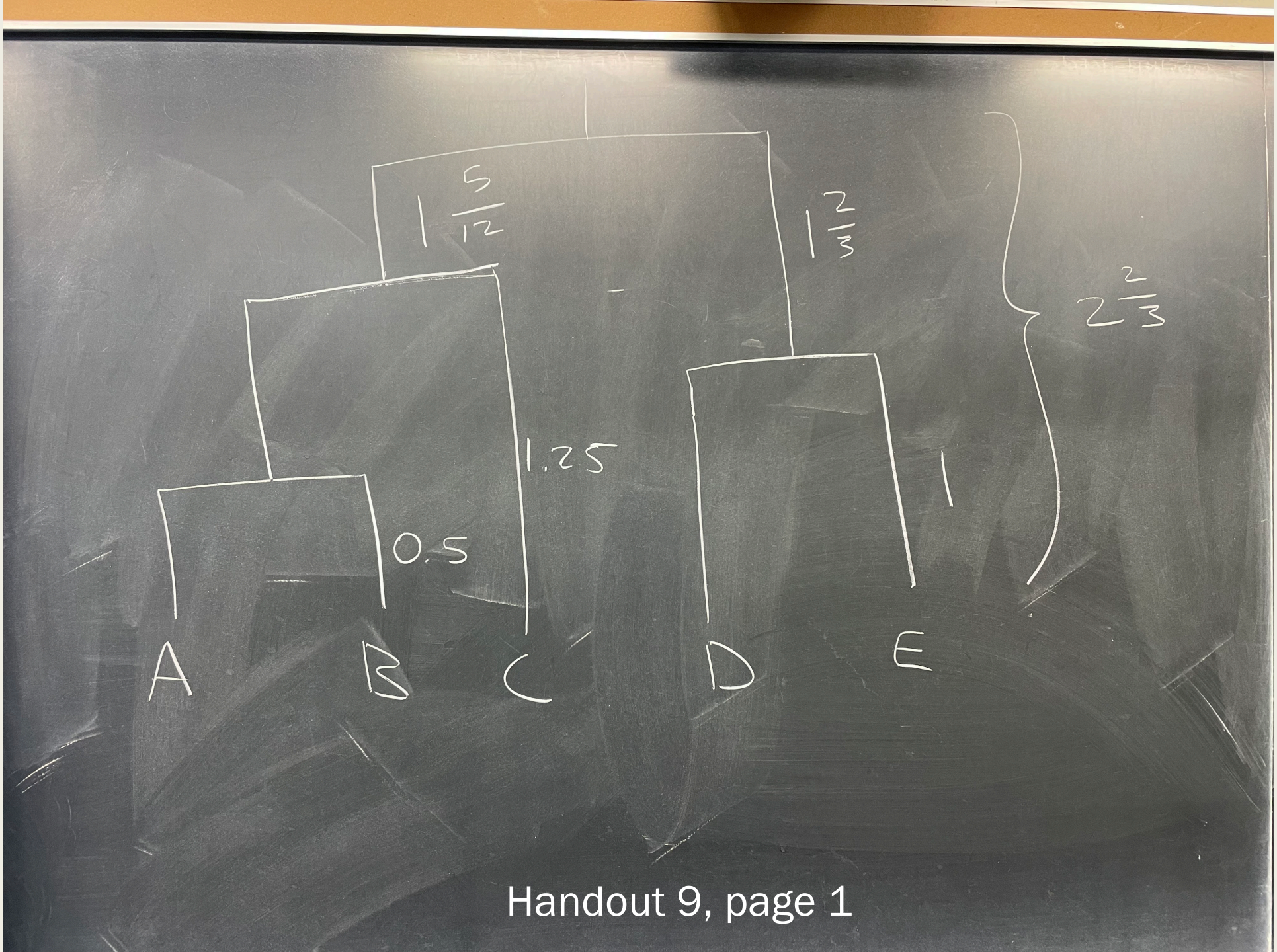
$$\Delta(C_D \cup C_E, C_C) = \frac{1}{2}5 + \frac{1}{2}5 = 5$$

	$\{A,B,C\}$	$\{DE\}$
$\{A,B,C\}$	0	$\boxed{5\frac{1}{3}}$
$\{D\}$		0

$$\Delta(C_{AB} \cup C_C, C_{DE})$$

$$= \frac{2}{3}\Delta(C_{AB}, C_{DE}) + \frac{1}{3}\Delta(C_C, C_{DE})$$

$$= \frac{2}{3}(5.5) + \frac{1}{3}5 = \frac{16}{3} = 5\frac{1}{3}$$

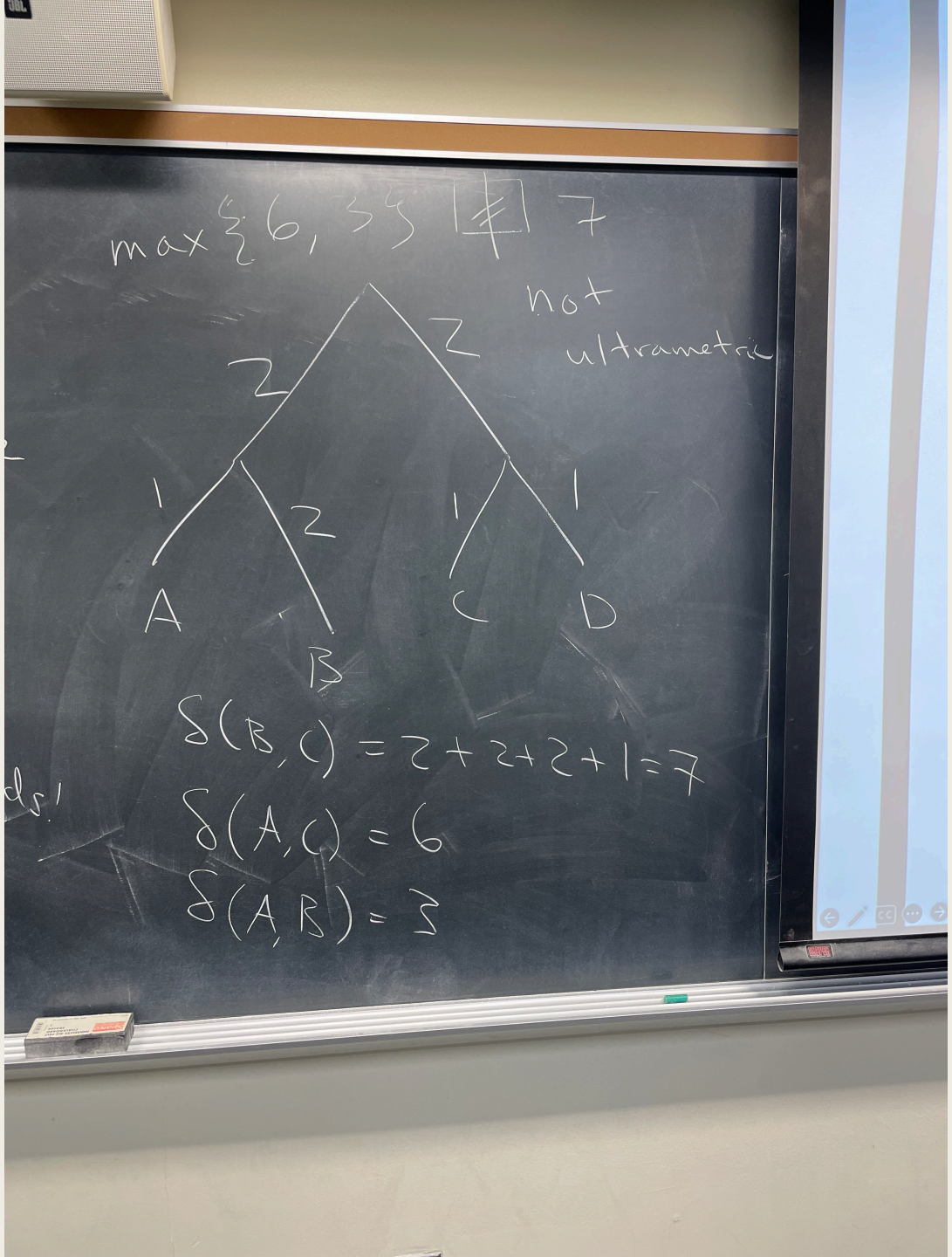


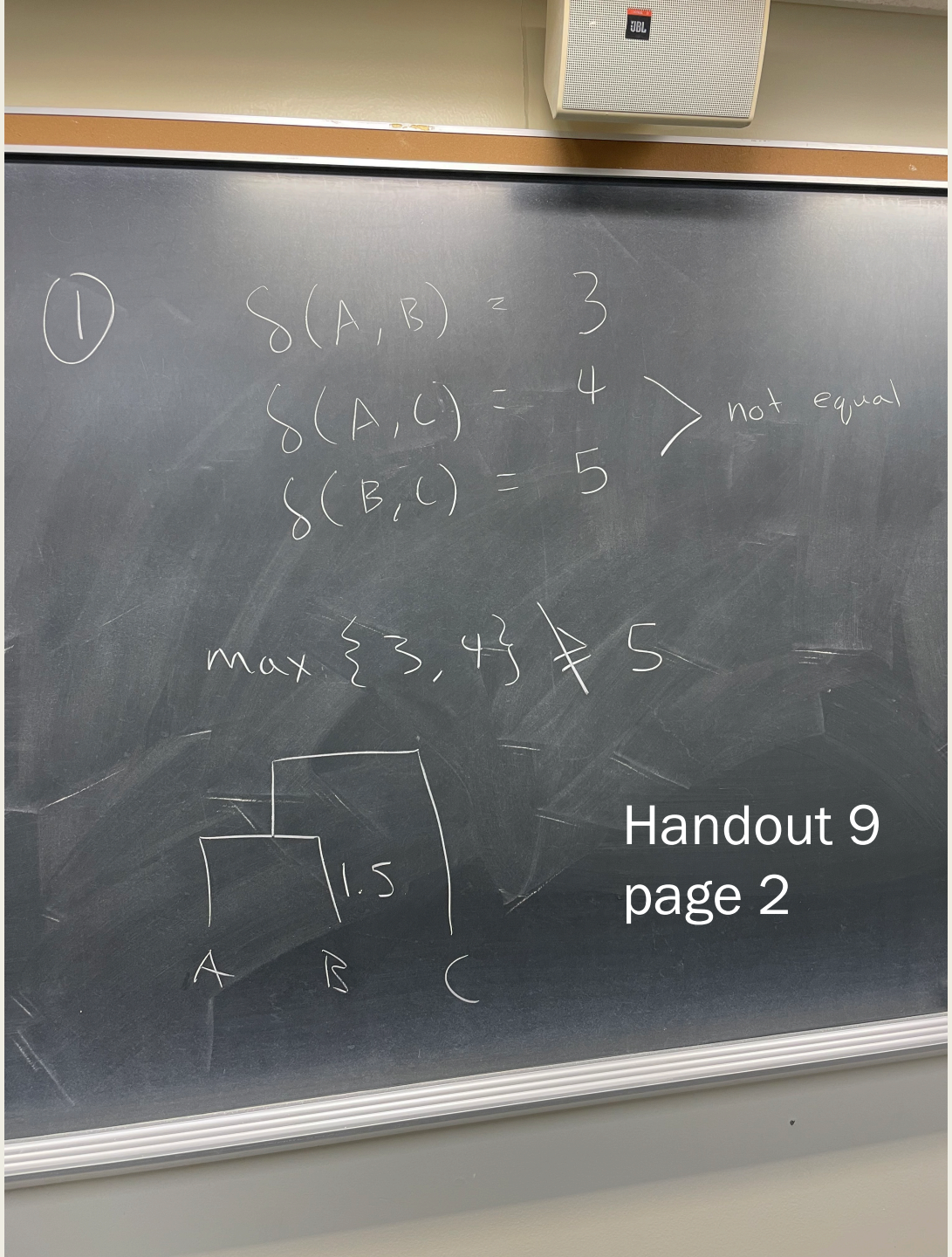
Handout 9, page 1

δ' induced metric

	A	B	C	D	E
A	0	1	2.5	$5\frac{1}{3}$	$5\frac{1}{3}$
B		0	2.5	$5\frac{1}{3}$	$5\frac{1}{3}$
C			0	$5\frac{1}{3}$	$5\frac{1}{3}$
D				0	2
E					0

not equal to original metric!





Handout 9
page 2