CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Feb 19

- Notes on Lab 2
- Phylogenetic Trees
- UPGMA algorithm
- Ultrametrics

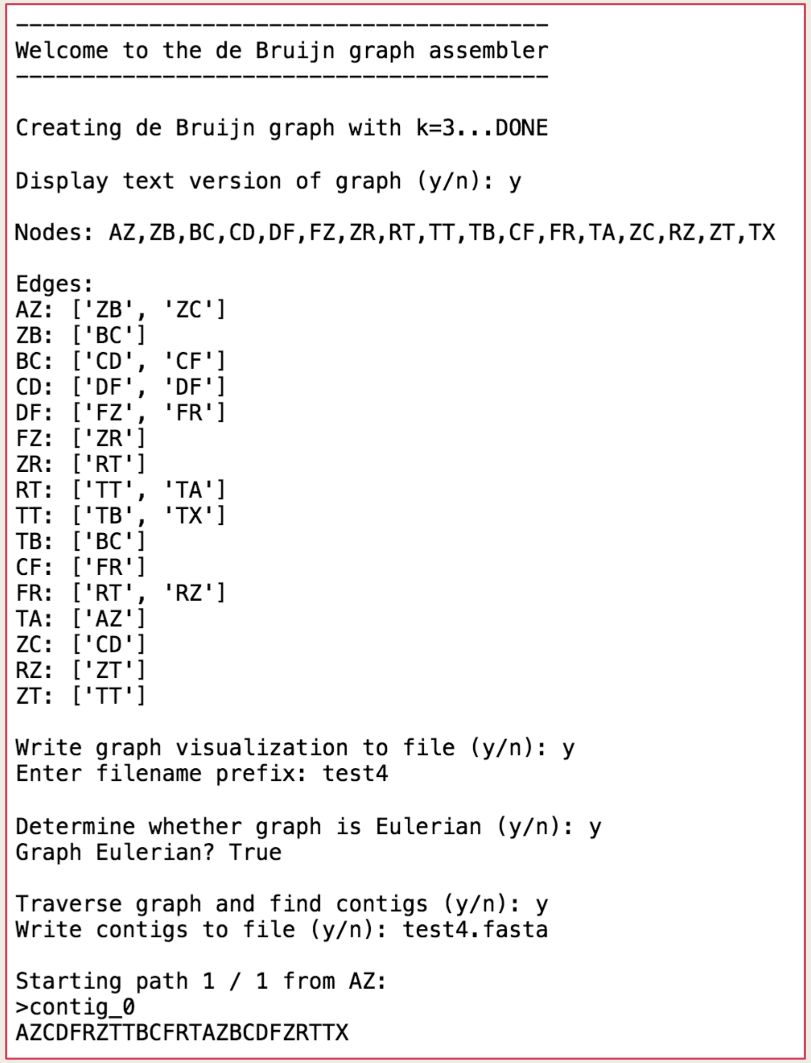
Notes:

- Office hours today: 3-5pm
- No office hours on Wednesday, moved to Tuesday 1-3pm
- Lab this week: practice midterm
- No class on Friday
- Reading posted (Durbin Chap. 7)

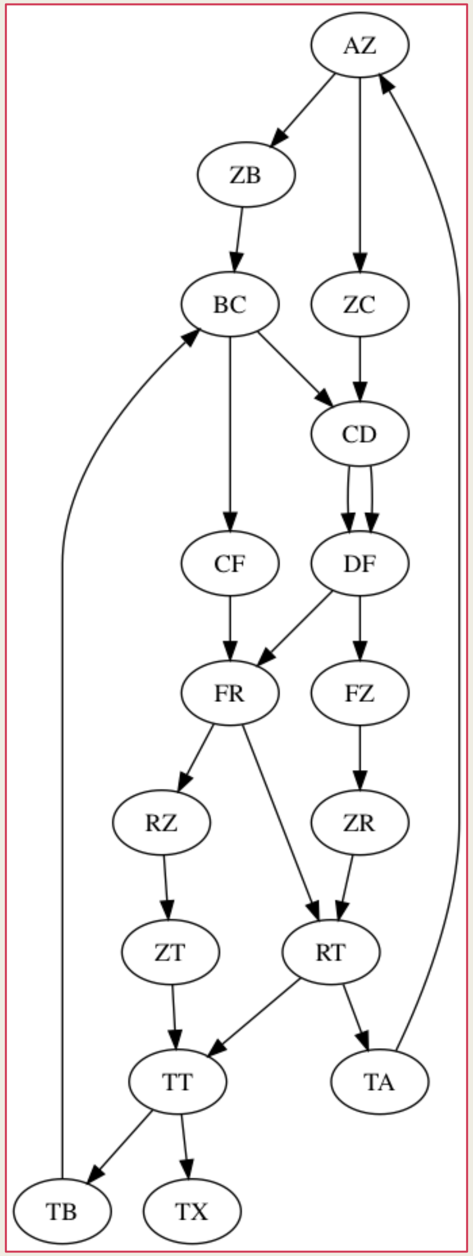
Lab 2 Notes

Additional test case

```
>read1
AZBCDFZR
>read2
ZRTTBCFRTAZ
>read3
AZCDFRZTTX
```

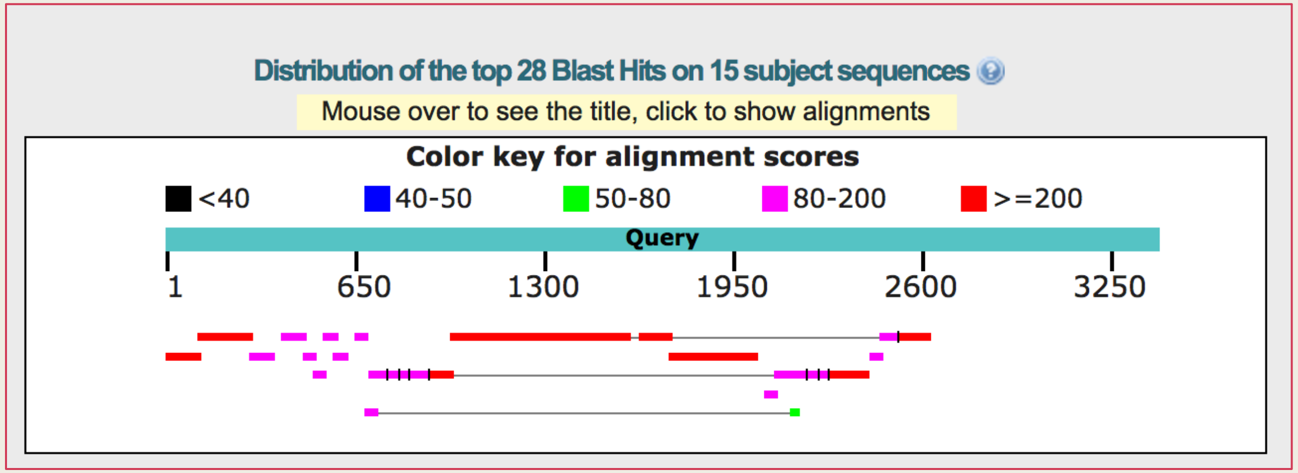


```
----------------------------------------
Welcome to the de Bruijn graph assembler
----------------------------------------

Creating de Bruijn graph with k=3...DONE

Display text version of graph (y/n): y

Nodes: AZ,ZB,BC,CD,DF,FZ,ZR,RT,TT,TB,CF,FR,TA,ZC,RZ,ZT,TX

Edges:
AZ: ['ZB', 'ZC']
ZB: ['BC']
BC: ['CD', 'CF']
CD: ['DF', 'DF']
DF: ['FZ', 'FR']
FZ: ['ZR']
ZR: ['RT']
RT: ['TT', 'TA']
TT: ['TB', 'TX']
TB: ['BC']
CF: ['FR']
FR: ['RT', 'RZ']
TA: ['AZ']
ZC: ['CD']
RZ: ['ZT']
ZT: ['TT']

Write graph visualization to file (y/n): y
Enter filename prefix: test4

Determine whether graph is Eulerian (y/n): y
Graph Eulerian? True

Traverse graph and find contigs (y/n): y
Write contigs to file (y/n): test4.fasta

Starting path 1 / 1 from AZ:
>contig_0
AZCDFRZTTBCFRTAZBCDFZRTTX
```



Genome X examples

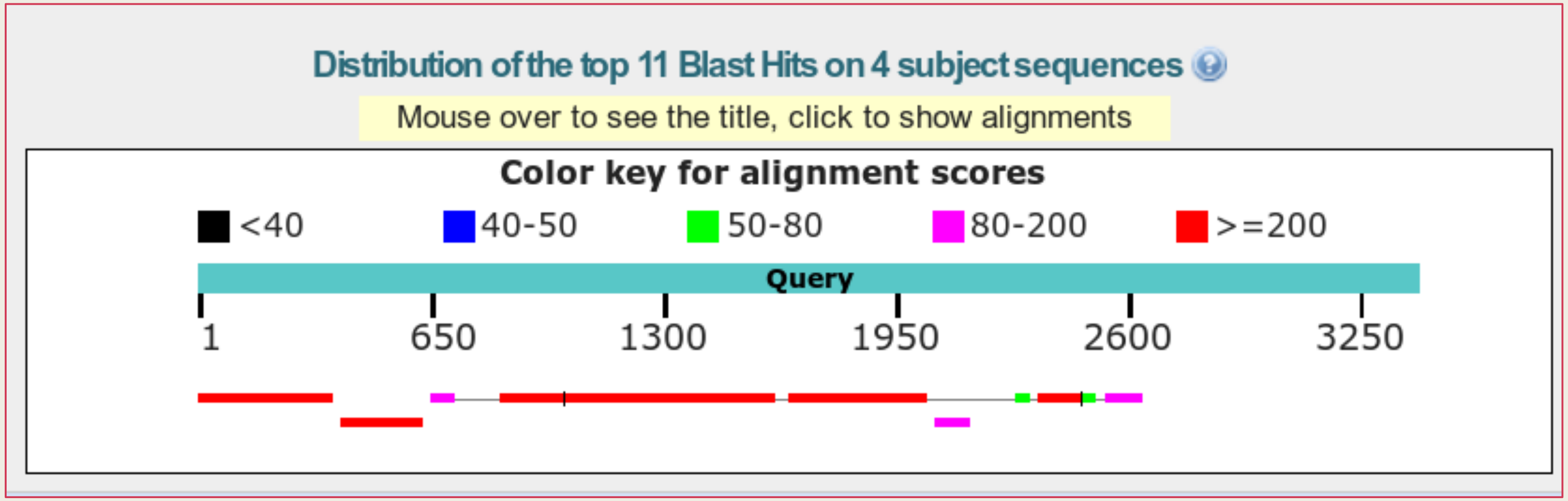
k=12



Distribution of the top 28 Blast Hits on 15 subject sequences

Mouse over to see the title, click to show alignments

k=15



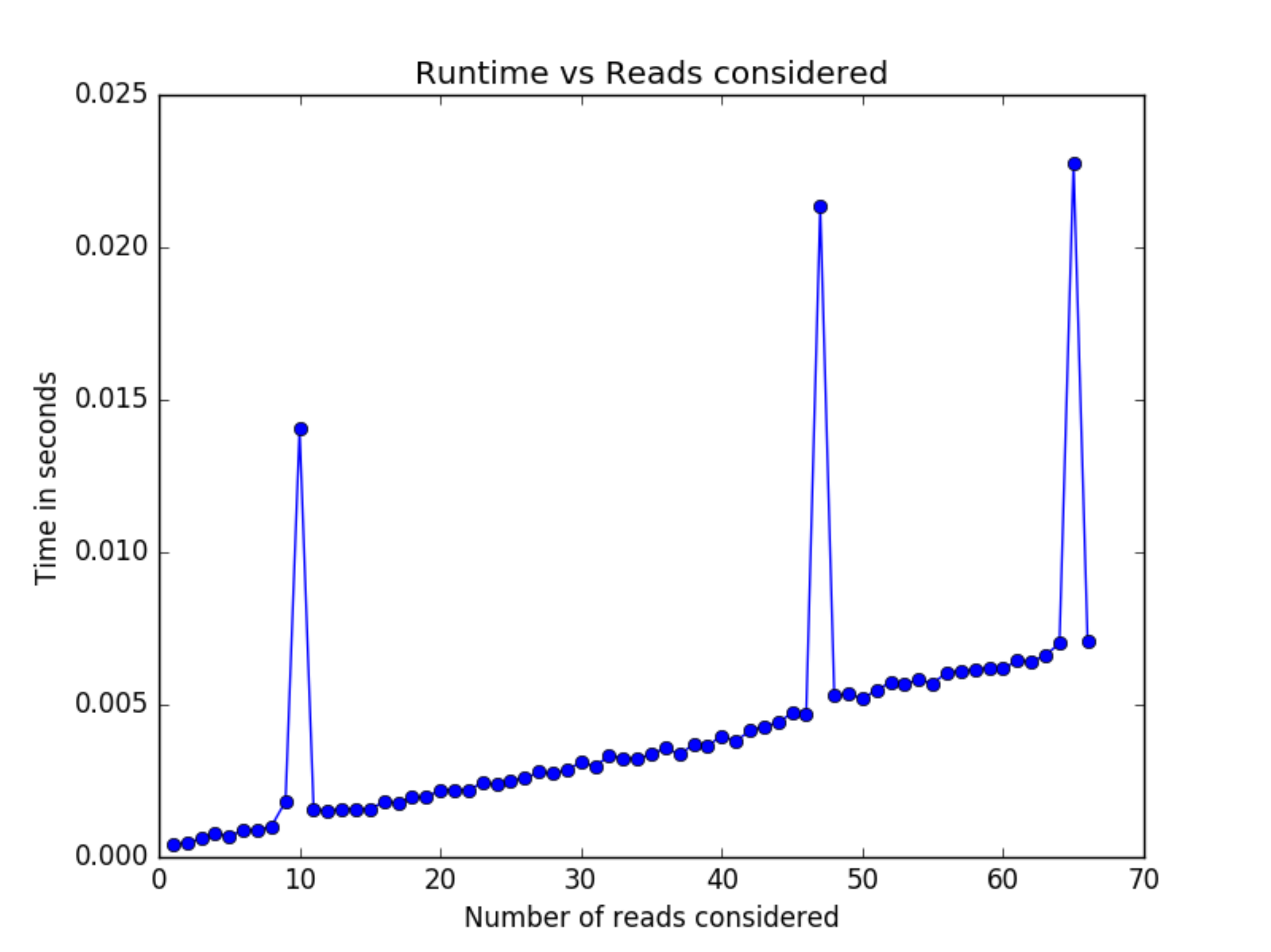
Distribution of the top 11 Blast Hits on 4 subject sequences

Mouse over to see the title, click to show alignments

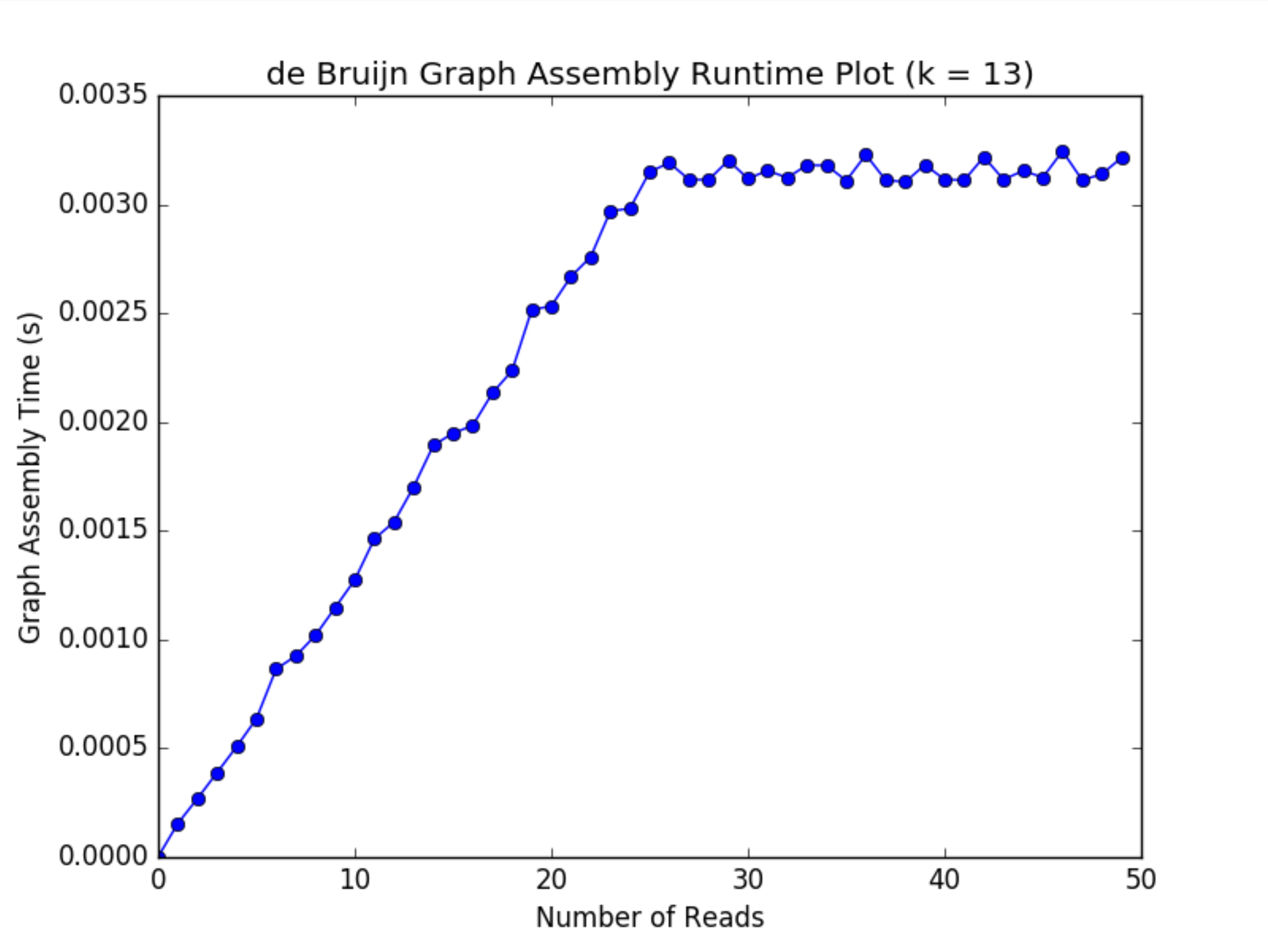
Lab 2: genome X, k=15

- Balance between low and high k
- If k is too small, will end up with overlaps that are not truly in the reads
- Graph becomes too connected and coverage differences create many non-balanced nodes
- If k is too large, graph becomes very disconnected and we end up building back only the reads

Runtime plot: Daniel and Sayed



Runtime plot: Nathan and Tyler



Phylogenetic Trees: UPGMA

(Unweighted Pair Group Method with Arithmetic mean)

Phylogenetic trees

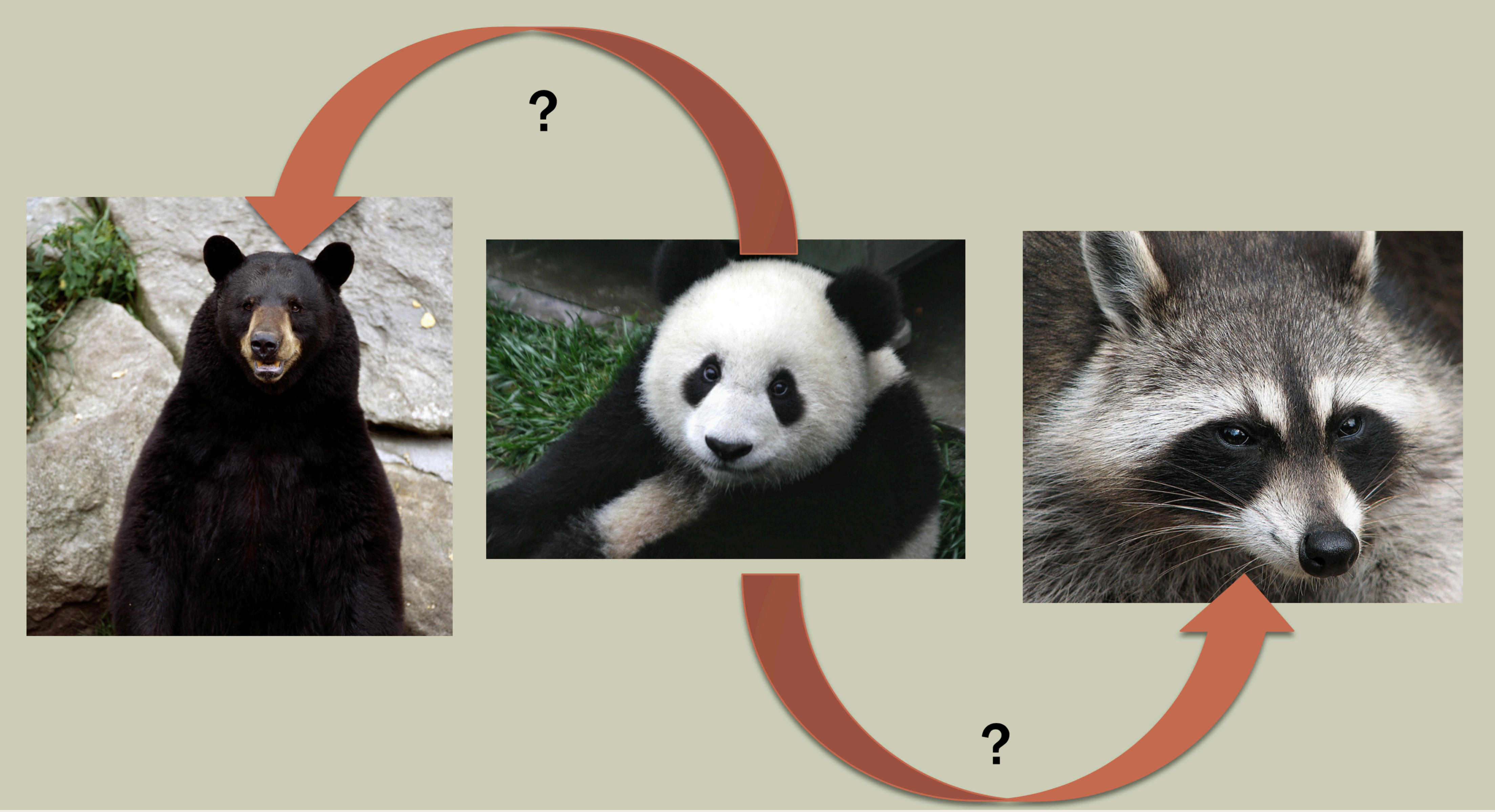
- **Definition:** diagram of inferred evolutionary relationships between samples (species, genes, individuals, etc)

- **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

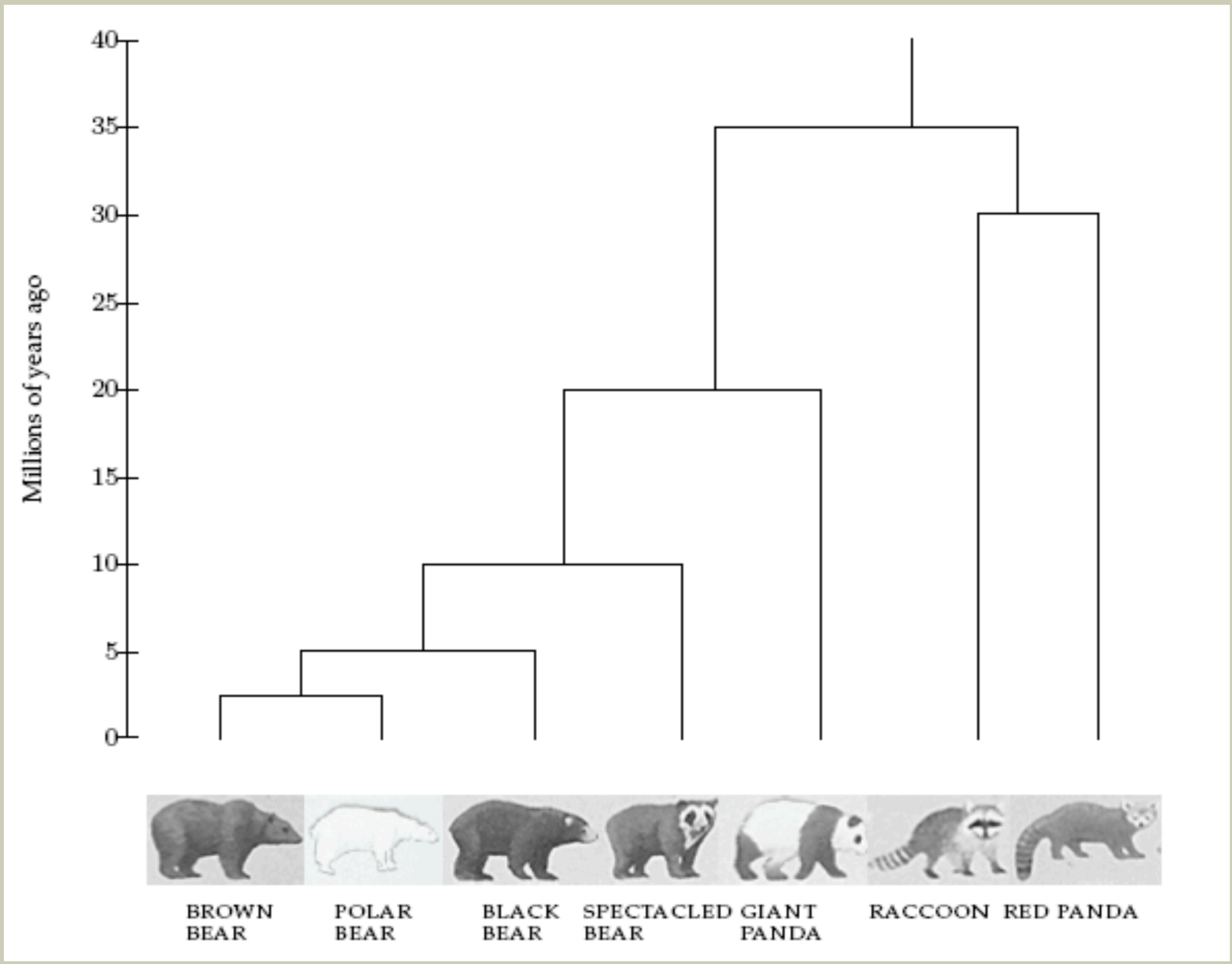
- **We can learn:** evolutionary history! Sequence of speciation events, function and evolution of common traits and genes, biology of common ancestors, tempo and mode of mutation, natural selection, recombination, migration, population size changes

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)

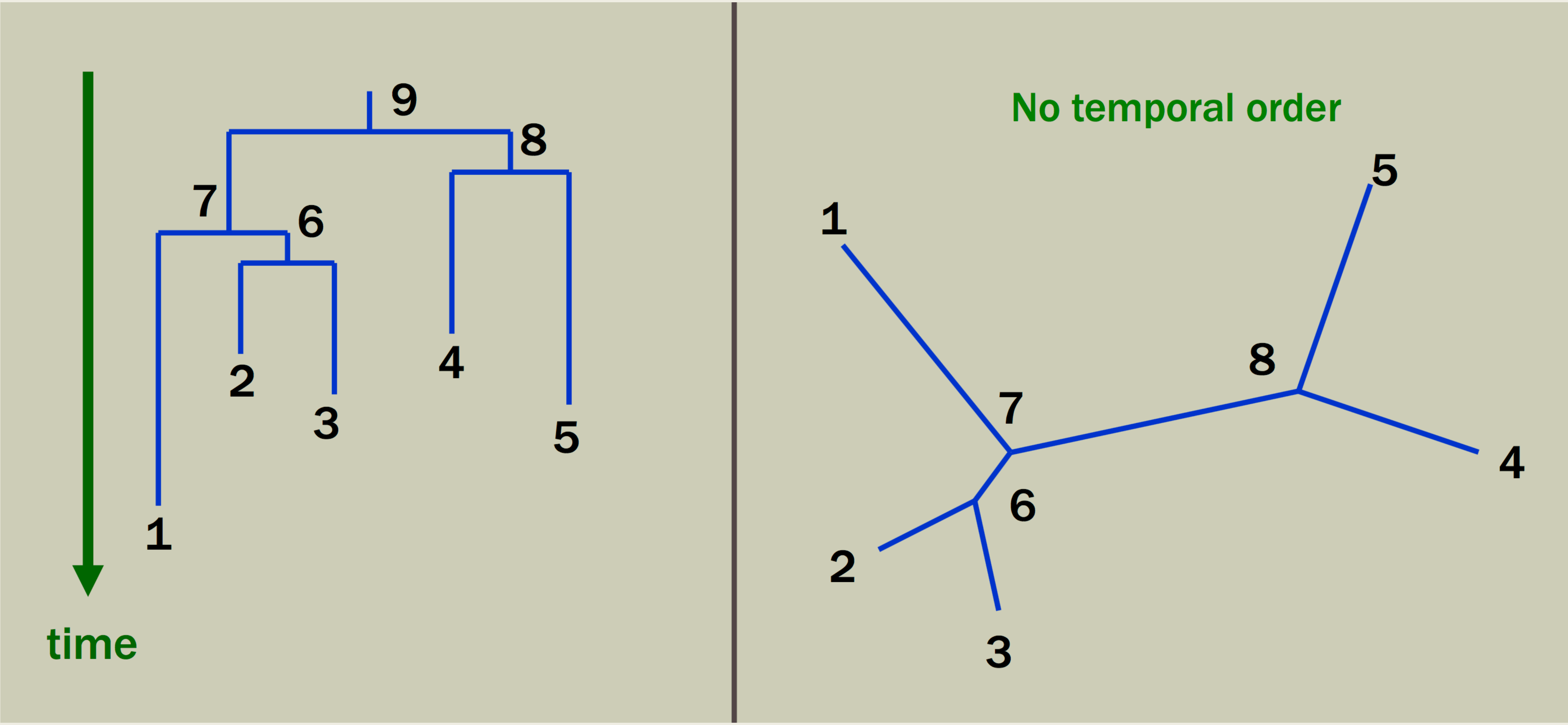
2) Why do we make the assumption that alleles at each site can be encoded as 0's and 1's?

   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.

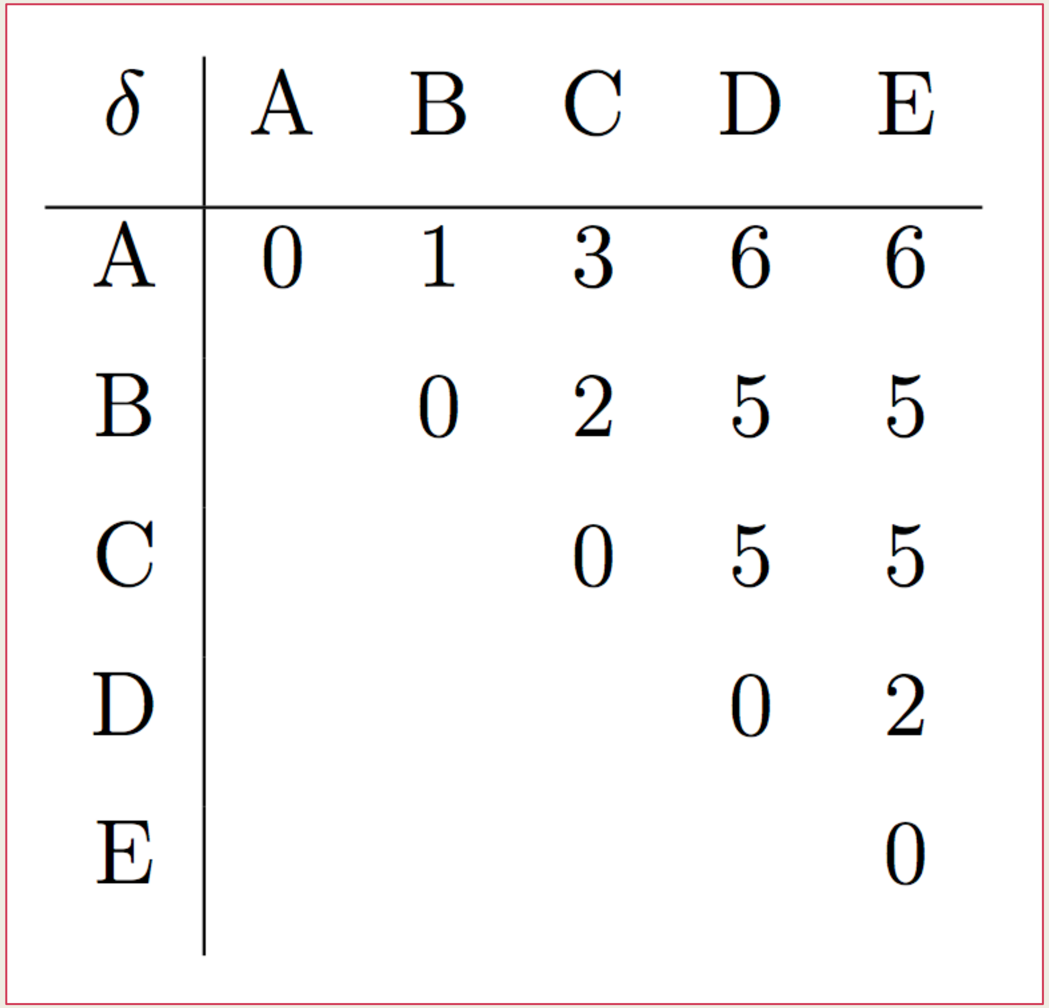
Rooted vs. unrooted trees



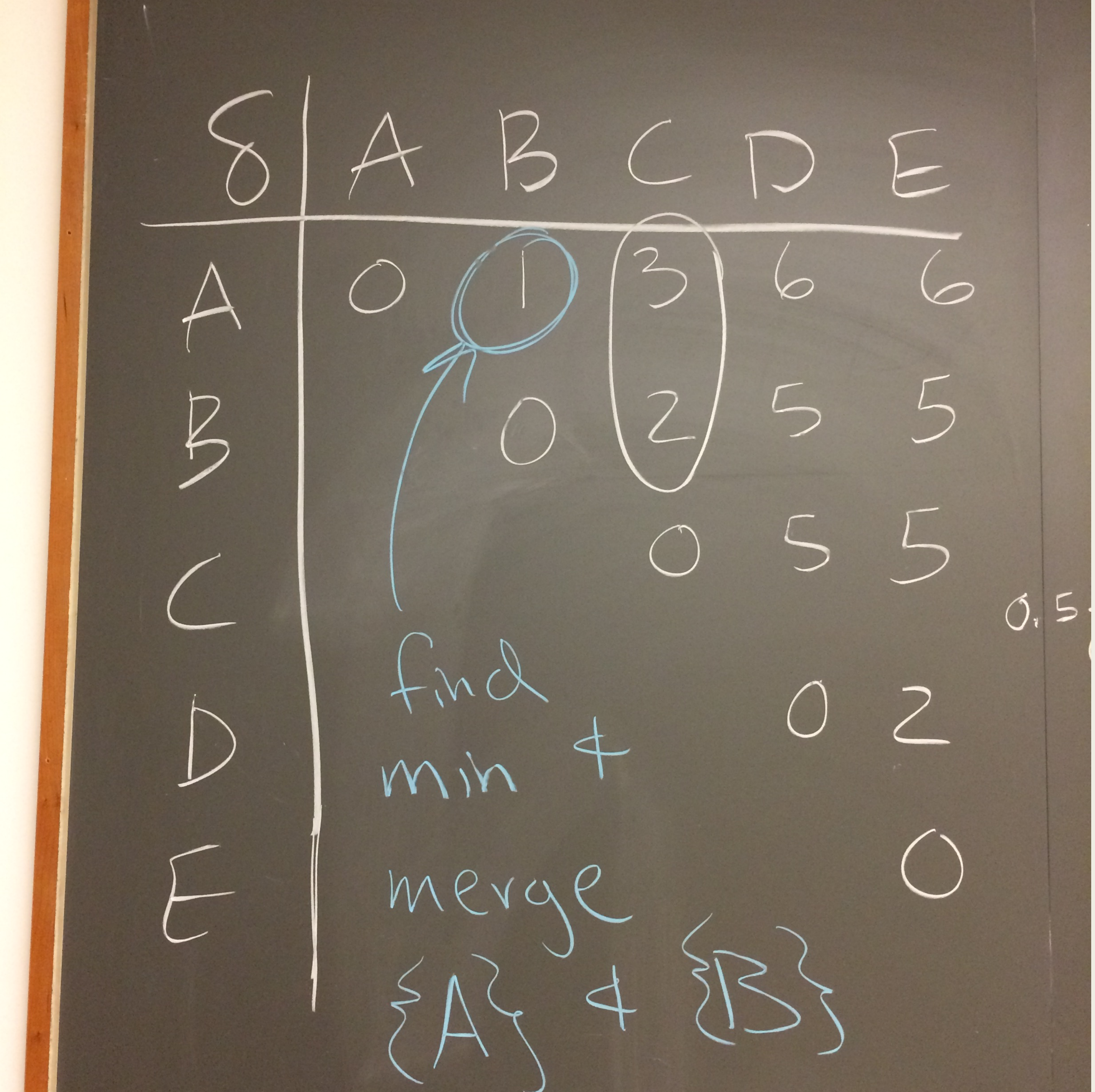
Credit: Ameet Soni

Example from last time

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



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



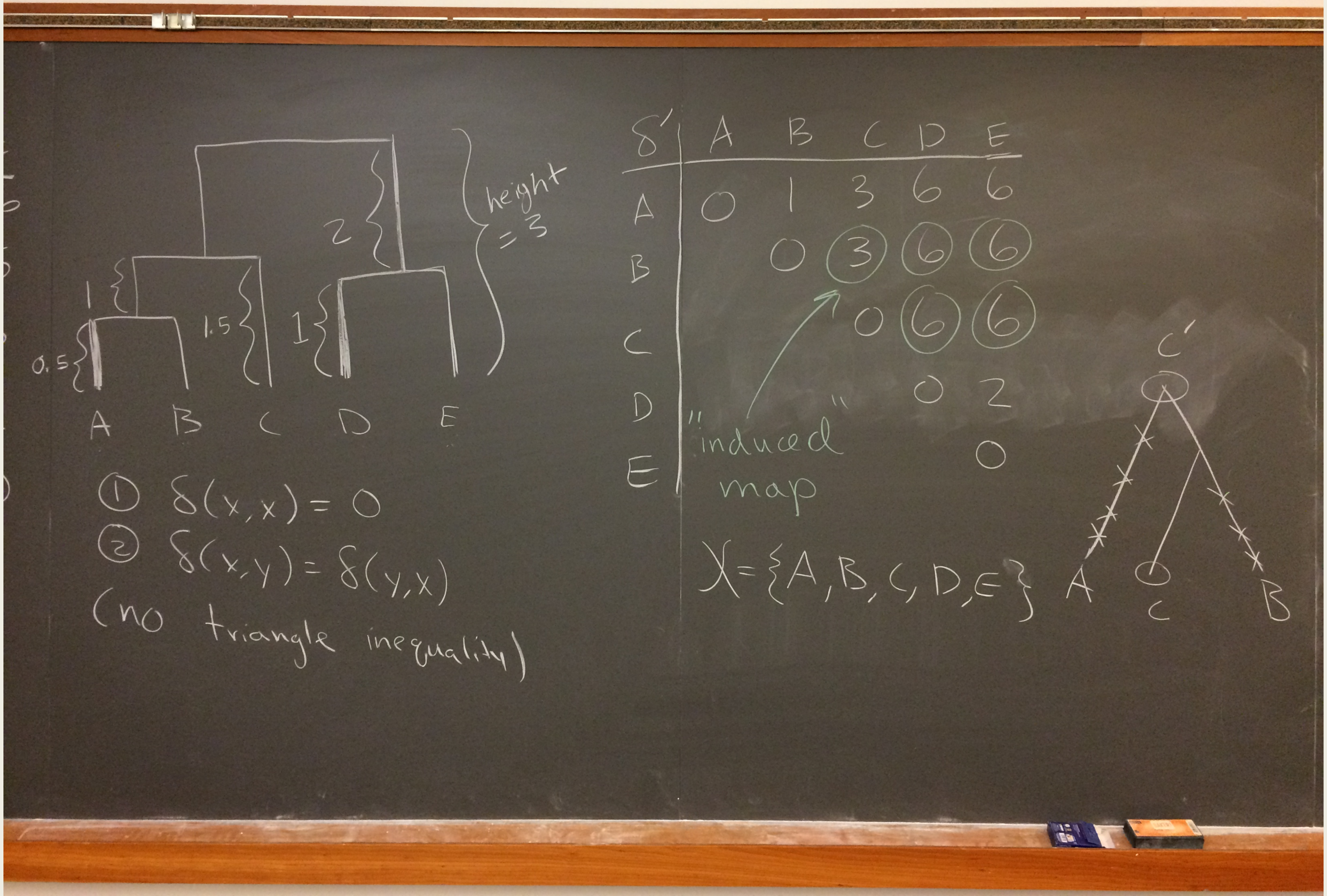
δ	A	B	C	D	E
A	0	1	3	6	6
B		0	2	5	5
C			0	5	5
D				0	2
E					0

find min &

merge {A} & {B}

0.5

“Guess” tree, example of the type of output we want



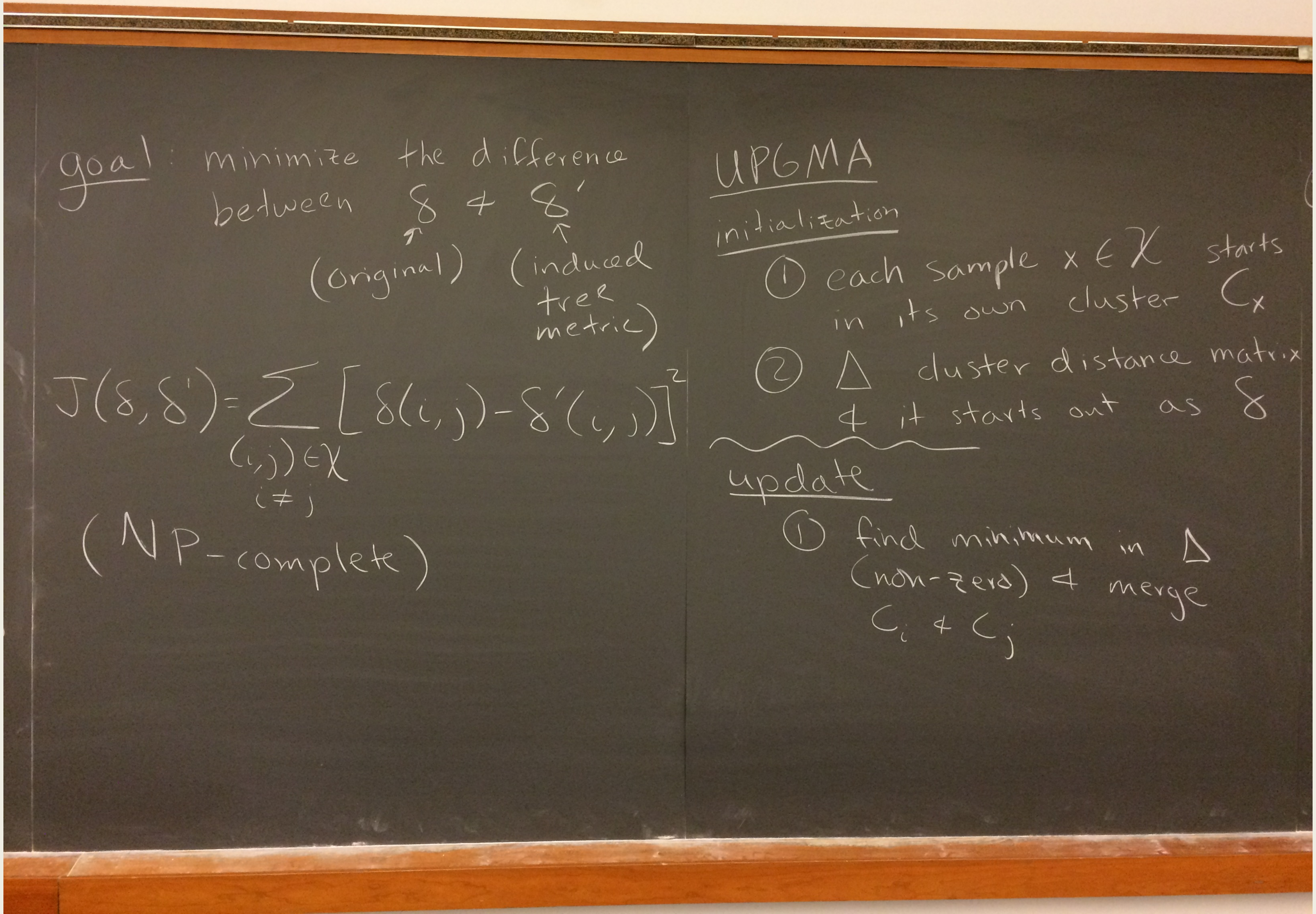
① $\delta(x,x) = 0$

② $\delta(x,y) = \delta(y,x)$

(no triangle inequality)

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

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



goal: minimize the difference between $\underset{\substack{\uparrow \\ \text{(original)}}}{\delta}$ & $\underset{\substack{\uparrow \\ \text{(induced tree metric)}}}{\delta'}$

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

(NP-complete)

UPGMA

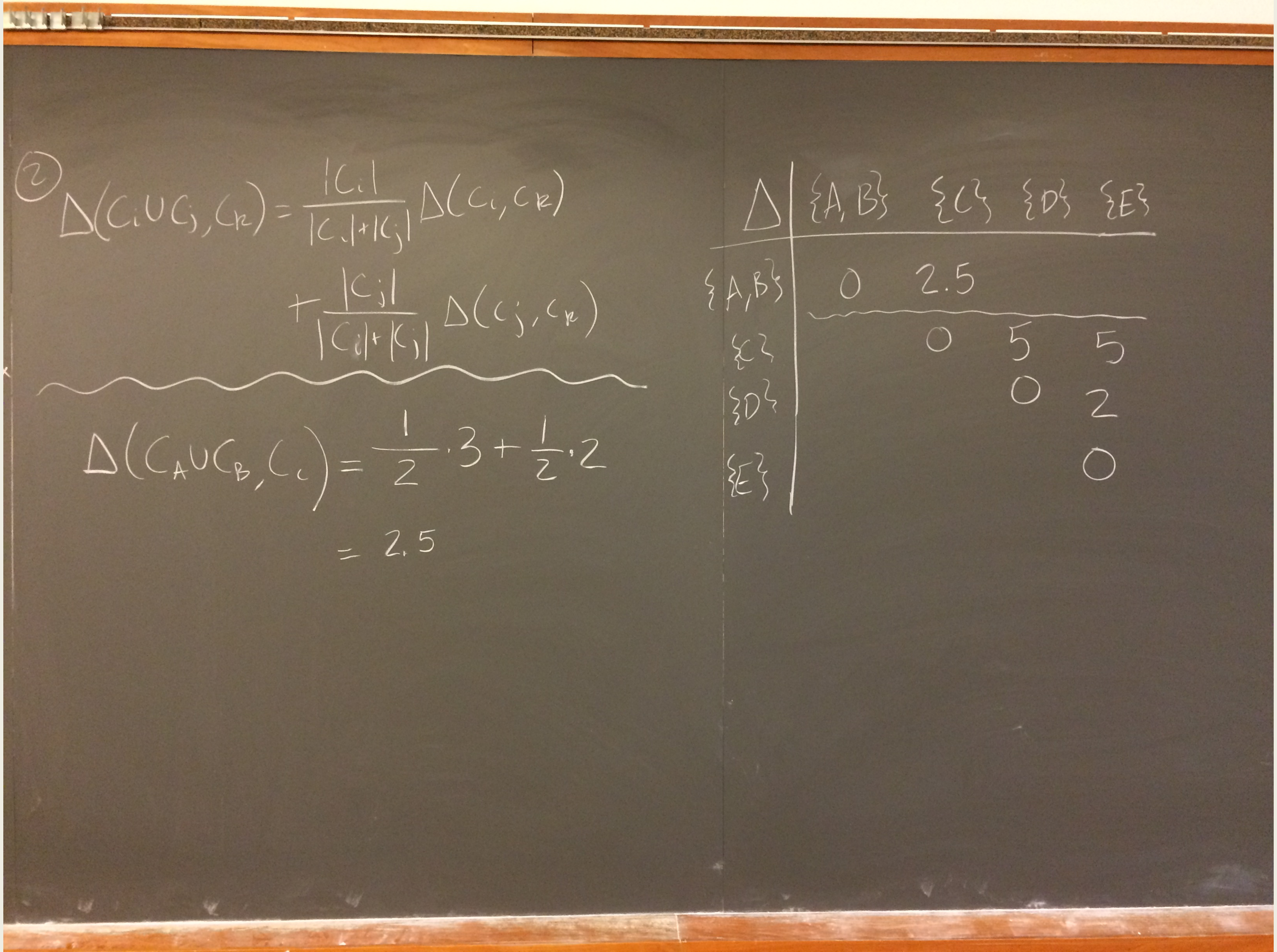
initialization

① each sample $x \in \chi$ starts in its own cluster C_x

② Δ cluster distance matrix & it starts out as δ

update

① find minimum in Δ (non-zero) & merge C_i & C_j



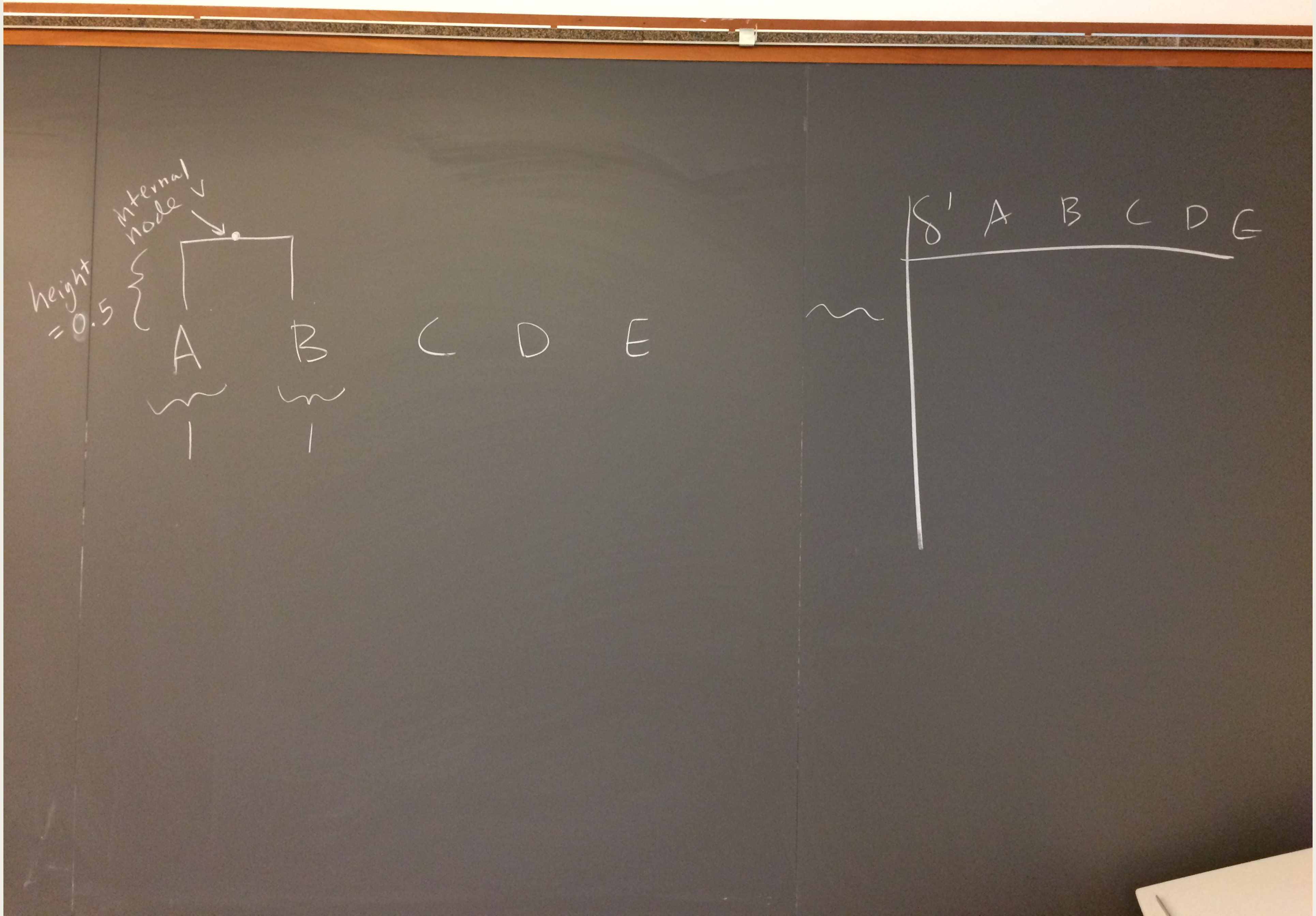
②

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

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

$$= 2.5$$

Δ	{A, B}	{C}	{D}	{E}
{A, B}	0	2.5		
{C}		0	5	5
{D}			0	2
{E}				0



δ'	A	B	C	D	E