CLUSTERING METHODS AND ALGORITHMS IN GENOMICS AND EVOLUTION

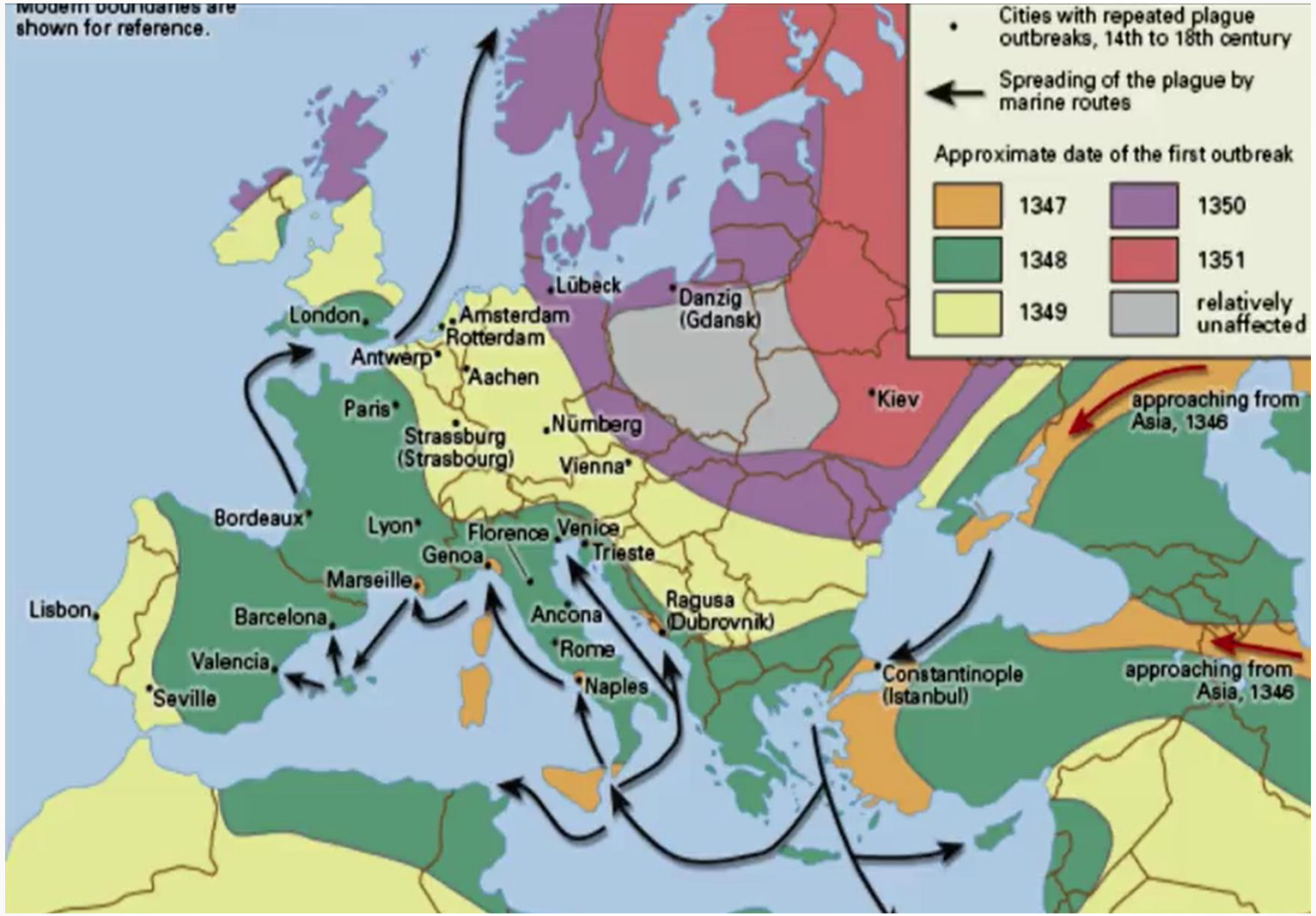
Session 7

Distance based methods for tree inference

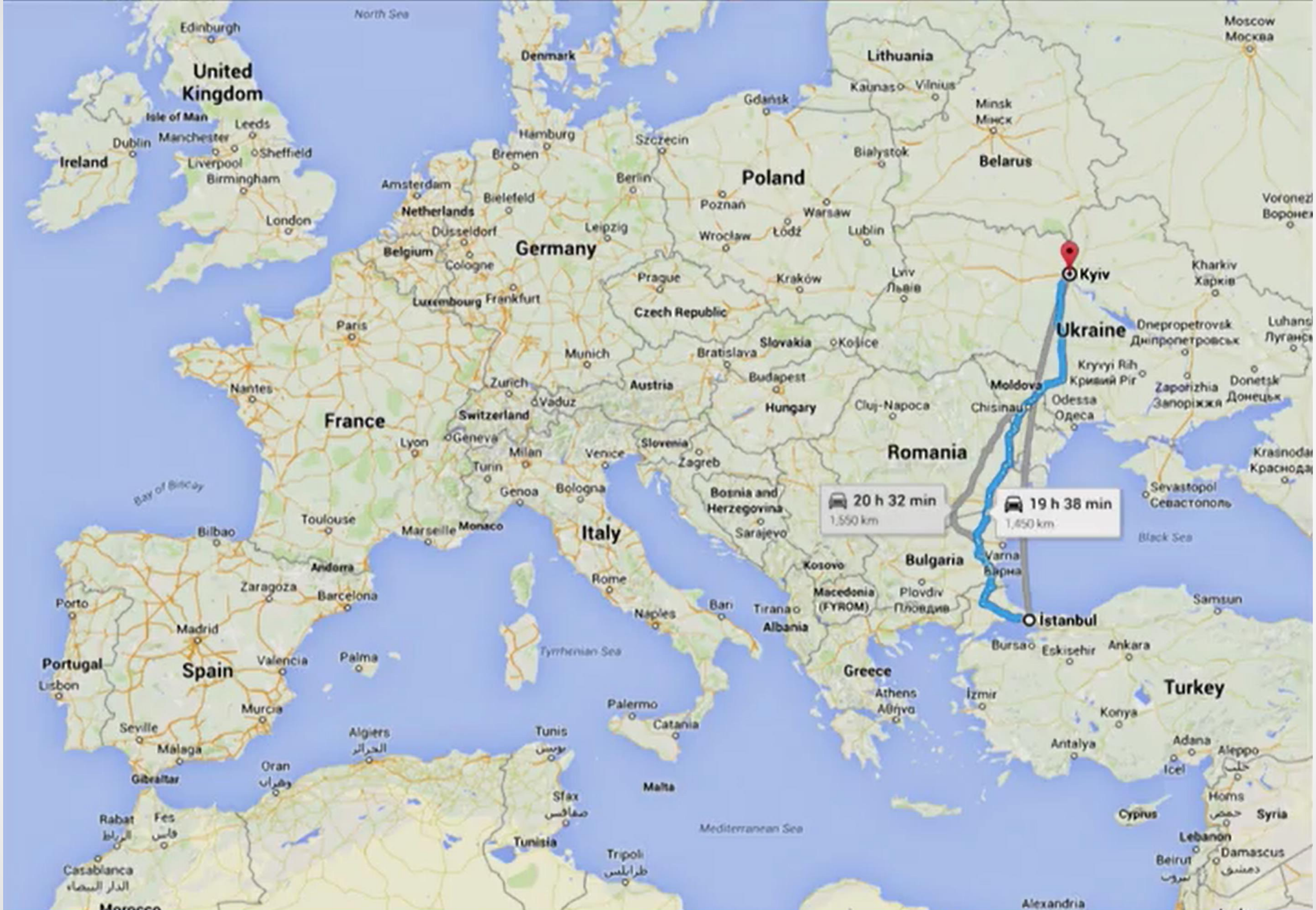
Trouble at the Metropole Hotel

- Hong Kong
- February 21, 2003

Outbreak of Black Death in 14th Century



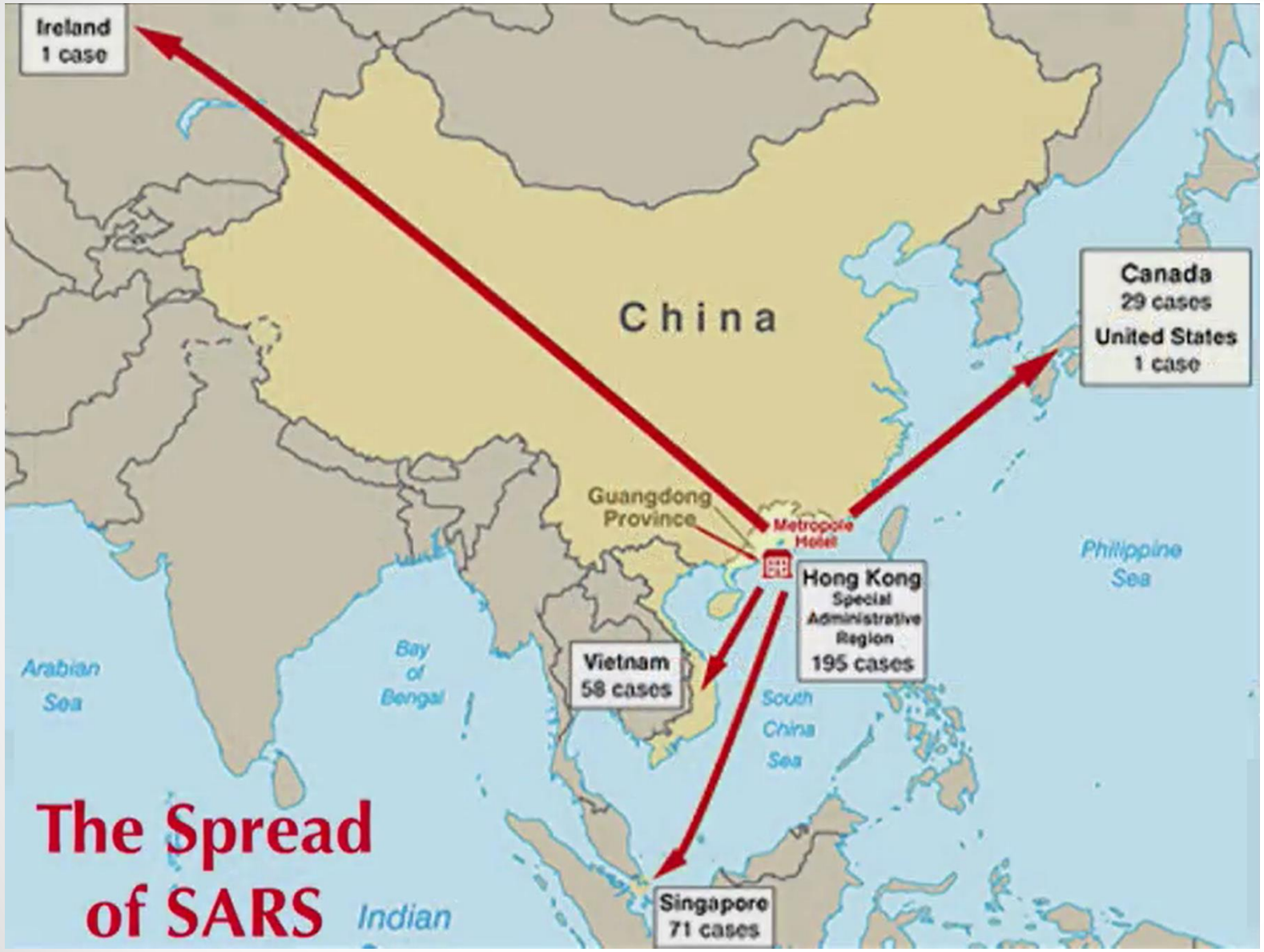
- Black Death killed a third of all Europeans. It took four years for the Black Death to travel from Constantinople (Istanbul) to Kiev.

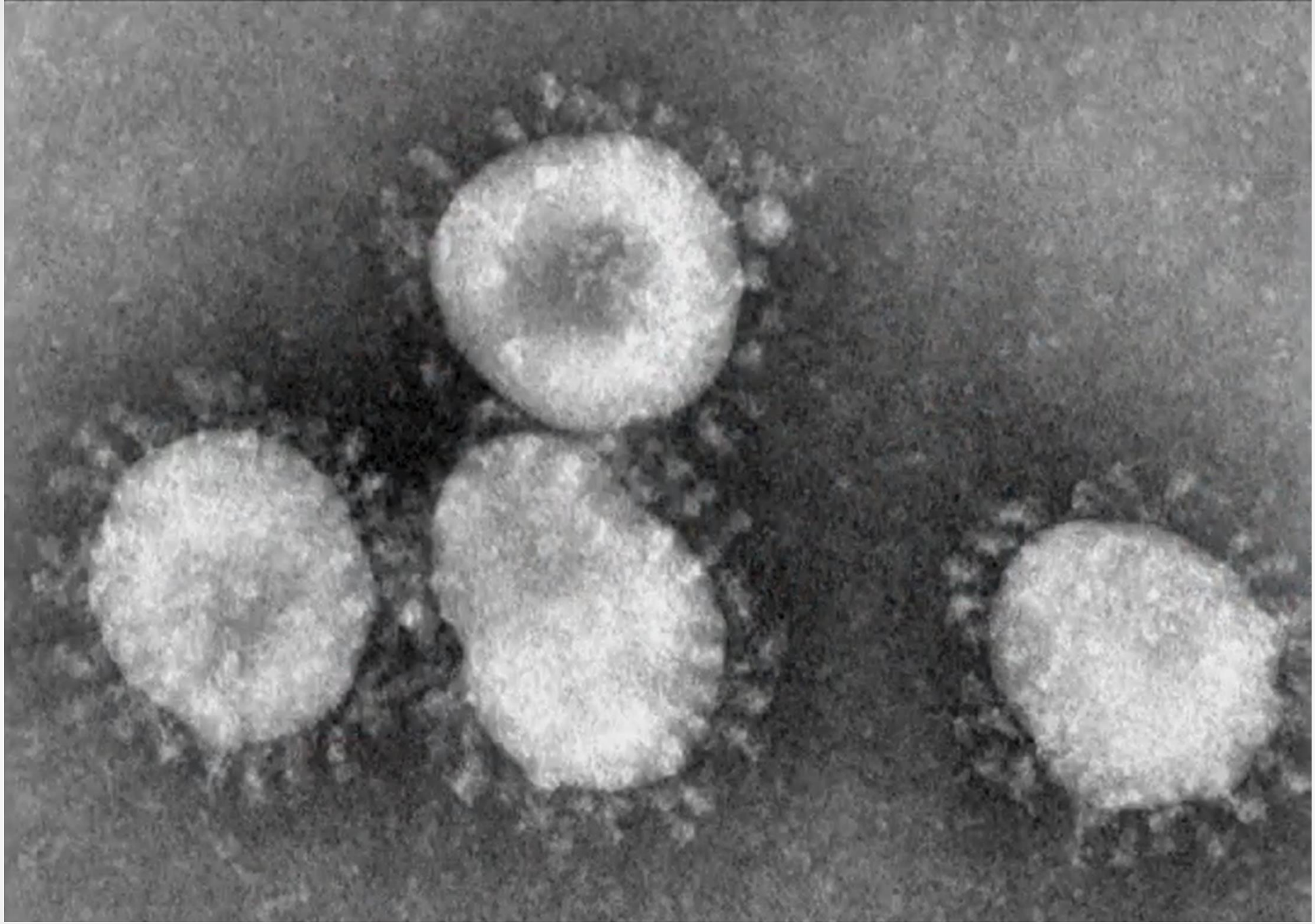

North Sea
Edinburgh
United Kingdom
Isle of Man
Leeds
Manchester
Sheffield
Dublin
Ireland
Liverpool
Birmingham
London
Denmark
Hamburg
Bremen
Amsterdam
Netherlands
Bielefeld
Düsseldorf
Belgium
Cologne
Germany
Berlin
Leipzig
Luxembourg
Frankfurt
Szczecin
Gdańsk
Poland
Poznań
Warsaw
Wrocław
Łódź
Lublin
Białystok
Lithuania
Kaunas
Vilnius
Minsk
Минск
Belarus
Moscow
Москва
Voronezh
Kyiv
Lviv
Львів
Kharkiv
Харків
Ukraine
Dnepropetrovsk
Дніпропетровськ
Prague
Czech Republic
Kraków
Slovakia
Košice
Bratislava
Budapest
Hungary
Munich
Austria
Zurich
Vaduz
Switzerland
Paris
Nantes
France
Lyon
Geneva
Milan
Turin
Venice
Slovenia
Zagreb
Genoa
Bologna
Moldova
Chisinau
Kryvyi Rih
Кривий Ріг
Zaporizhia
Запоріжжя
Donetsk
Донецьк
Odessa
Одеса
Cluj-Napoca
Romania
Krasnodar
Краснодар
Sevastopol
Севастополь
Bay of Biscay
Bosnia and Herzegovina
Sarajevo
20 h 32 min
1,550 km
19 h 38 min
1,450 km
Black Sea
Toulouse
Marseille
Monaco
Italy
Bilbao
Andorra
Bulgaria
Varna
Варна
Kosovo
Rome
Zaragoza
Barcelona
Porto
Macedonia (FYROM)
Plovdiv
Пловдив
Istanbul
Samsun
Naples
Bari
Tirana
Albania
Madrid
Bursa
Eskişehir
Ankara
Tyrrhenian Sea
Portugal
Lisbon
Spain
Valencia
Palma
Greece
Athens
Αθήνα
İzmir
Turkey
Konya
Murcia
Palermo
Catania
Seville
Malaga
Algiers
الجزائر
Tunis
تونس
Antalya
Adana
Aleppo
Icel
Oran
وهران
Gibraltar
Malta
Sfax
صفاقس
Homs
حمص
Syria
Cyprus
Rabat
الرباط
Fes
فاس
Mediterranean Sea
Tunisia
Lebanon
Beirut
بيروت
Damascus
دمشق
Tripoli
طرابلس
Casablanca
الدار البيضاء
Morocco
Alexandria

Globalization: Diseases travel fast in 21st Century

- 13 people infected in the hotel
- Two days later disease was in Hanoi
- Three days after it crossed the Pacific to Toronto
- In five days it entered Singapore
- Being still unknown the disease converted into global epidemic

The Fastest Outbreak: Severe Acute Respiratory Syndrome





Coronavirus particles. Photo: Luc Viatour.

A solar eclipse with the sun’s corona visible. Photo: Luc Viatour.

Coronovirus genome: RNA

```
28261 aauaauacug cgucuugguu cacagcucuc acucagcaug gcaaggagga acuuagauuc
28321 ccucgaggcc agggcguucc aaucaacacc aauaguggguc cagaugacca aauuggcuac
28381 uaccgaagag cuacccgacg aguucguggu ggugacggca aaaugaaaga gcucagcccc
28441 agaugguacu ucuauuaccu aggaacuggc ccagaagcuu cacuucccua cggcgcuaac
28501 aaagaaggca ucguaugggu ugcaacugag ggagccuuga auacacccaa agaccacauu
28561 ggcacccgca auccuaauaa caaugcugcc accgugcuac aacuuccuca aggaacaaca
28621 uugccaaaag gcuucuacgc agagggaagc agaggcggca gucaagccuc uucucgcucc
28681 ucaucacgua gucgcgguaa uucaagaaau ucaacuccug gcagcaguag gggaaauucu
28741 ccugcucgaa uggcuagcgg agguggugaa acugcccucg cgcuauugcu gcuagacaga
28801 uugaaccagc uugagagcaa aguuucuggu aaaggccaac aacaacaagg ccaaacuguc
28861 acuaagaaau cugcugcuga ggcaucuaaa aagccucgcc aaaaacguac ugccacaaaa
28921 caguacaacg ucacucaagc auuuggggaga cgugguccag aacaaaccca aggaaauuuc
28981 ggggaccaag accuaaucag acaaggaacu gauuacaaac auuggccgca aauugcacaa
29041 uuugcuccaa gugccucugc auucuuugga augucacgca uuggcaugga agucacaccu
29101 ucgggaacau ggcugacuua ucauggagcc auuaaauugg augacaaaga uccacaauuc
29161 aaagacaacg ucauacugcu gaacaagcac auugacgcau acaaaacauu cccaccaaca
29221 gagccuaaaa aggacaaaaa gaaaaagacu gaugaagcuc agccuuugcc gcagagacaa
29281 aagaagcagc ccacugugac ucuucuuccu gcggcugaca uggaugauuu cuccagacaa
29341 cuucaaaauu ccaugagugg agcuucugcu gauucaacuc aggcauaaac acucaugaug
29401 accacacaag gcagaugggc uauguaaacg uuuucgcaau uccguuuacg auacauaguc
29461 uacucuugug cagaaugaau ucucguaacu aaacagcaca aguagguuua guuaacuuua
29521 aucucacaua gcaaucuuua aucaaugugu aacauuaggg aggacuugaa agagccacca
29581 cauuuucauc gaggccacgc ggaguacgau cgaggguaca gugaauaaug cuagggagag
29641 cugccuauau ggaagagccc uaauguguaa aauuaauuuu aguagugcua uccccaugug
29701 auuuuaauag cuucuuagga gaaugacaaa aaaaaaaaaa aaaaaaaaaa a
```

A short snippet from the end of the RNA genome.

In 2003 SARS coronovirus was quickly sequenced, which makes up a genome with 29,751 nucleotides.

Coronovirus genome: RNA

RNA replication has a much higher error rate than DNA. RNA viruses mutate fast.

- That is why the flu shot changes yearly and there are so many types of HIV and no vaccine.

Coronovirus genome: RNA

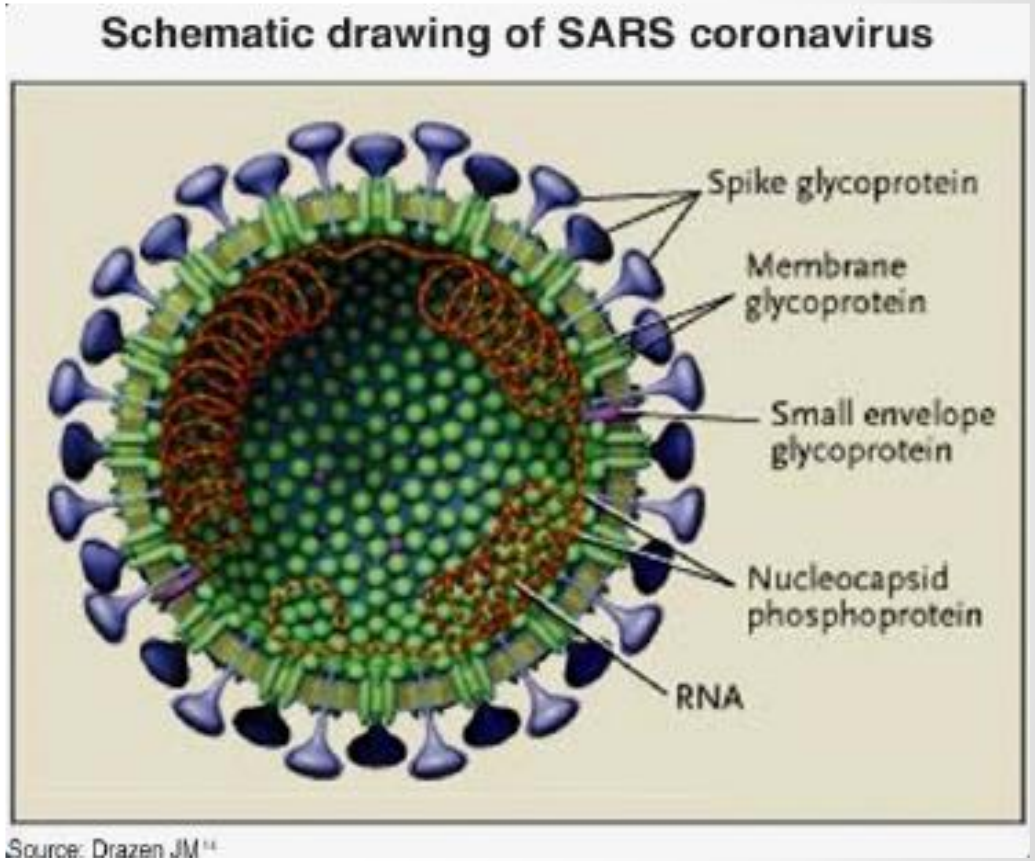
Researchers Assumptions

- Like HIV and influenza, SARS had jumped from animals to humans.
- Blamed birds: similarity of SARS with bird flu (influenza originated in chickens and crossed the species barrier to humans in 1997).

Spike Protein

Function: Identifies the receptor sites on the host's cell membrane.

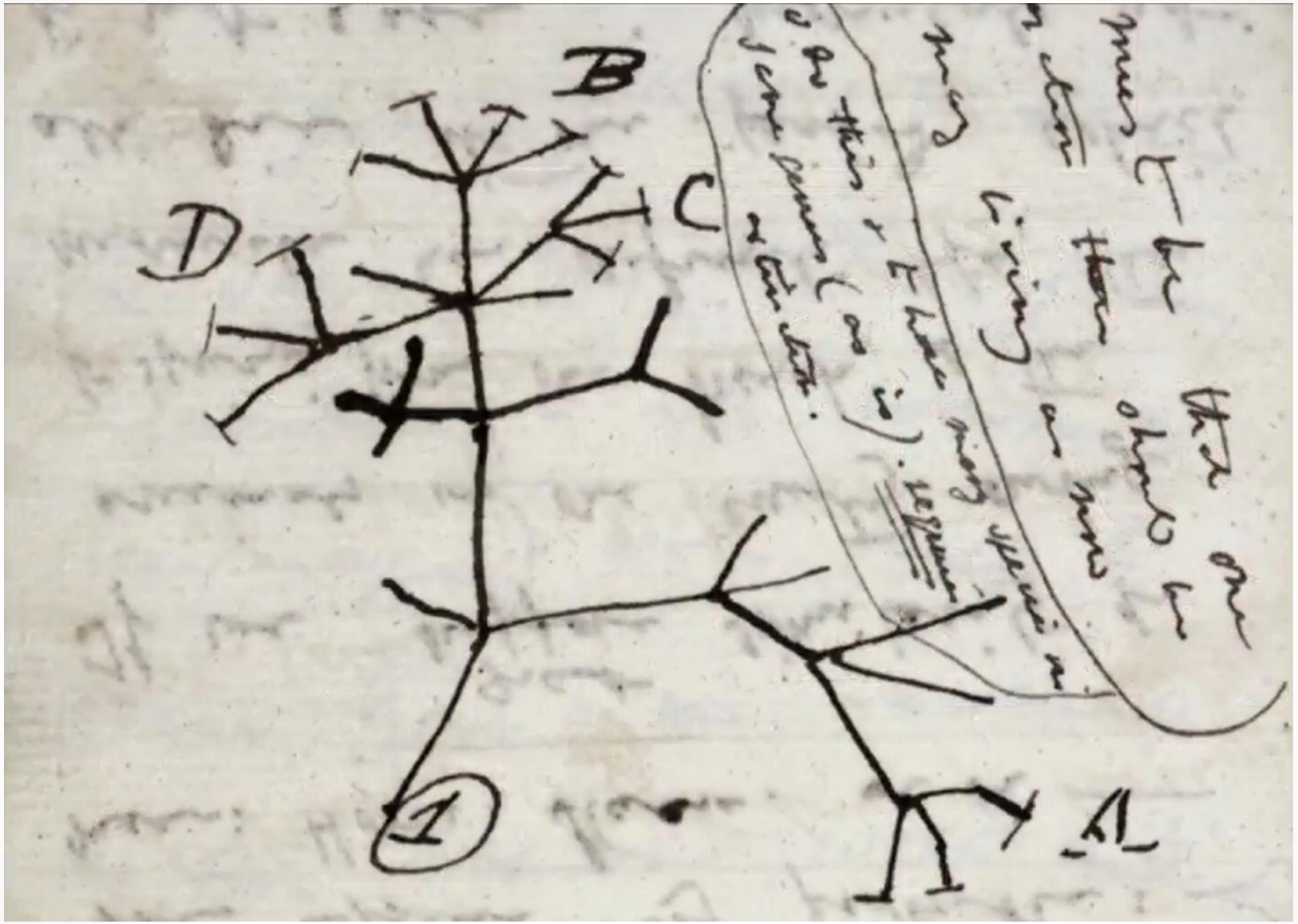
```
   1 mkvlivllcl glvtaqdgcg histkpqplm dkfsssrrgv yynddifrsd vlhltqdyfl
  61 pfdtnltryl sfnmdsatkv yfdnptlpfg dgiyfaatek snvvrgwifg stmdnttqsa
 121 iivnnsthii irvcyfnlck epmyaisneq hykswvyqna ynctydrveq sfqldtapqt
 181 gnfkdlreyv fknkdgflsv ynayspidip rglpvgfsvl kpilklpisi nitsfkvvms
 241 mfsrttsnfl pevaayfvgn lkystfmlnf nengtitdai dcaqnplsel kctiknfnvs
 301 kgiyqtsnfr vspthevirf pnitnrcpfd kvfnasrfpn vyawertkis dcvadytvly
 361 nstsfstfkc ygvspsklid lcftsvyadt flirssevrq vapgetgvia dynyklpddf
 421 tgcviawnta kqdqgqyyyr ssrktklkpf erdltsdeng vrtlstydfy pnvpieyqat
 481 rvvvlsfell napatvcgpk lstglvknqc vnfnfnglkg tgvltdsskr fqsfqqfgrd
 541 tsdftdsvrd pqtlqildit pcsfggvsvi tpgtnassev avlyqdvnct dvptairadq
 601 ltpawrvyst gvnvfqtqag cligaehvna syecdipiga gicasyhtas tlrsvgqksi
 661 vaytmslgae nsiayannsi aiptnfsisv ttevmpvsma ktsvdctmyi cgdsqecsnl
 721 llqygsfctq lnraltgval eqdkntqevf aqvkqmyktp aikdfggfnf sqilpdpskp
 781 tkrsfiedll fnkvtladag fmkqygeclg disardlica qkfngltvlp plltdemiaa
 841 ytaalvsgta tagwtfgaga alqipfamqm ayrfngigvt qnvlyenqkq ianqfnkais
 901 qiqeslttts talgklqdvv nqnaqalntl vkqlssnfga issvlndils rldkveaevq
 961 idrlitgrlq slqtyvtqql iraaeirasa nlaatkmsec vpgqskrvdf cgrgyhlmsf
1021 pqaaphgvvf lhvtyvpsqe knfttapaic hegkayfpre gvfvsngtsw fitqrnfysp
1081 qiittdntfv agncdvvigi inntvydplq peldsfkeel dkyfknhtsp dvdlgdisgi
1141 nasvvniqke idrlnevakn lneslidlqe lgkyeqyikw pwyvwlgfia gliaivmati
1201 llccmtsccs clkgacscgs cckfdeddse pvlkgvklhy t
```



How did the Spike protein got his name? Viruses form spherical particles, and their viral envelopes are studded with many little "spikes" formed by viral proteins. For example, HIV particles are embedded in the viral envelope with 72 spikes, formed by gp120 and gp41 proteins.

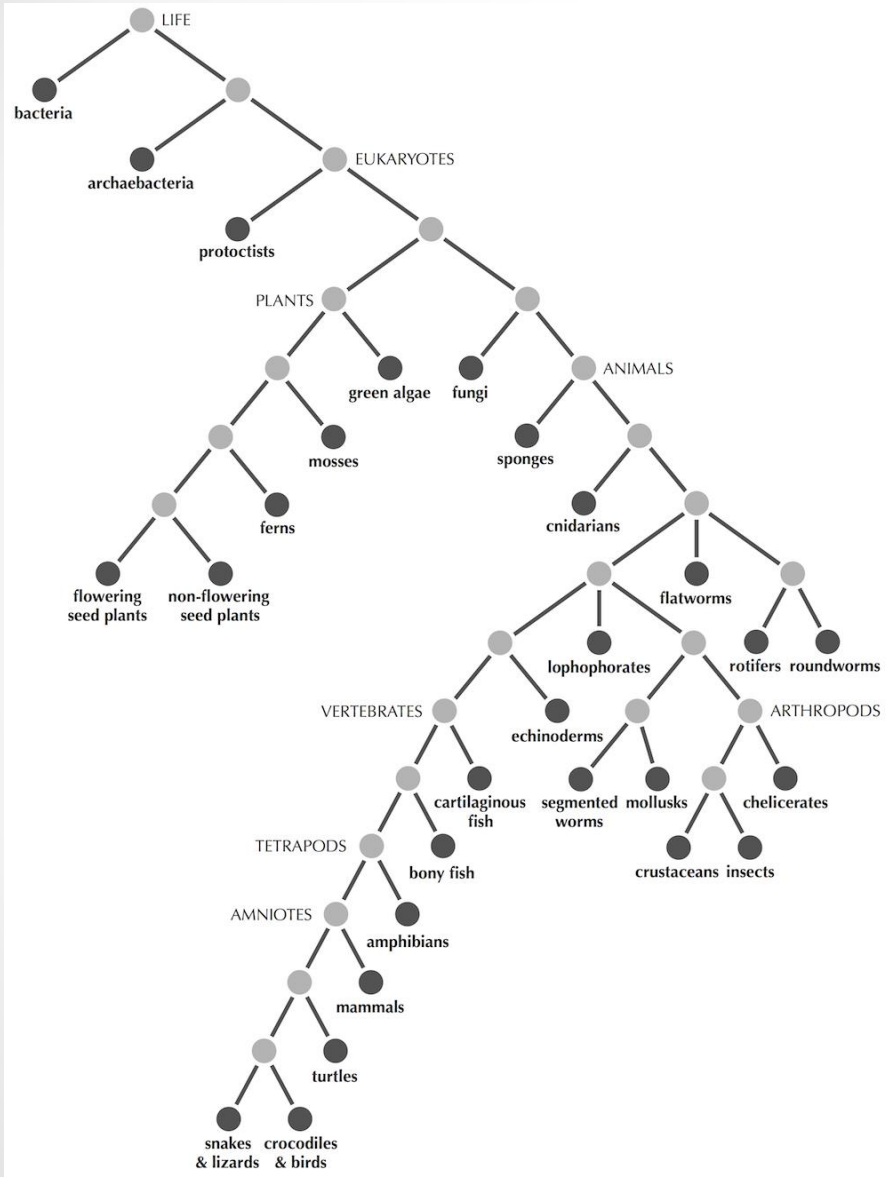
Questions about SARS

- Which animal gave us SARS?
- How were we first infected?
- How did SARS spread around the world?
- All these questions relate to constructing **evolutionary trees** (a.k.a. phylogenies).



First evolutionary tree drawn by Charles Darwin in 1837.

Trees



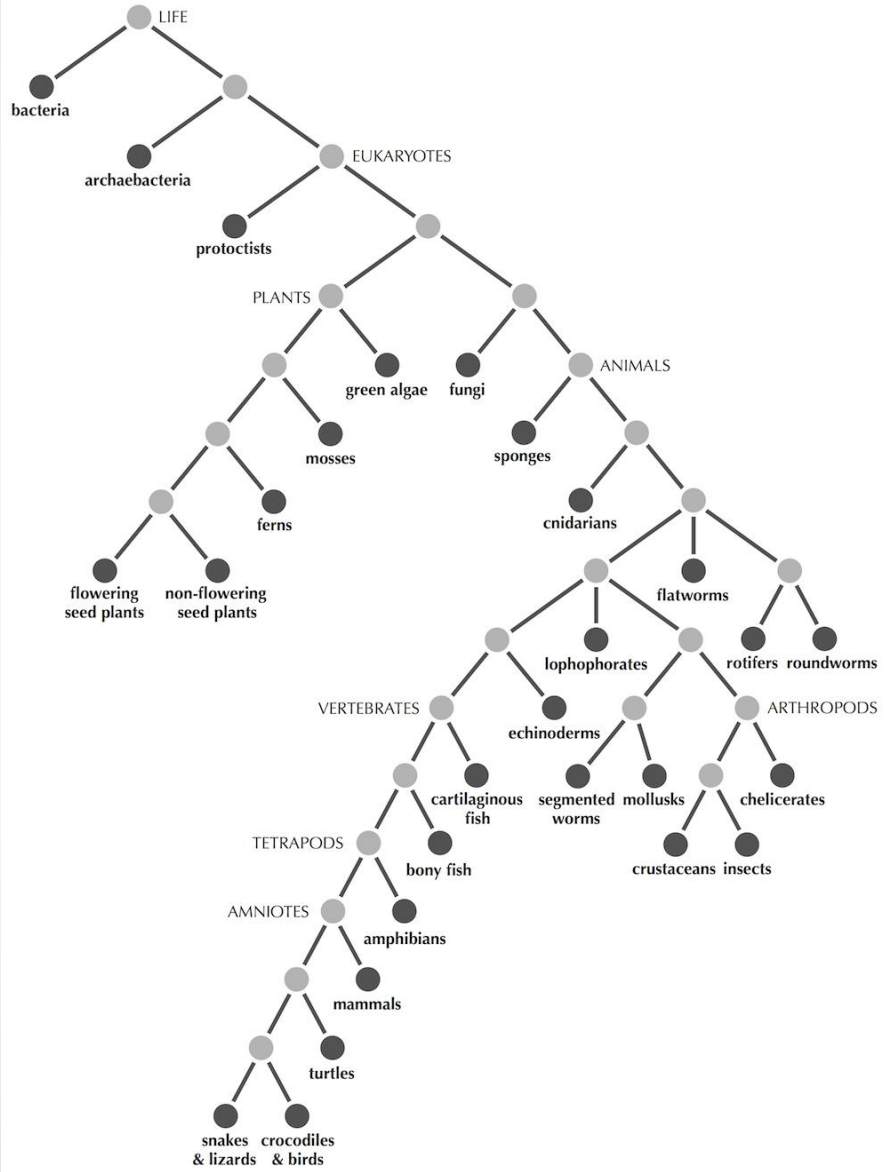
Tree: *Connected graph containing no cycles.*

Leaves (degree = 1): *present-day species.*

- Should be at the ending nodes of the tree.
- **Degree**- a number of edges connecting a node to other nodes.

A connected graph without cycles that models an evolutionary tree of life on Earth. Present-day species are shown as darker nodes.

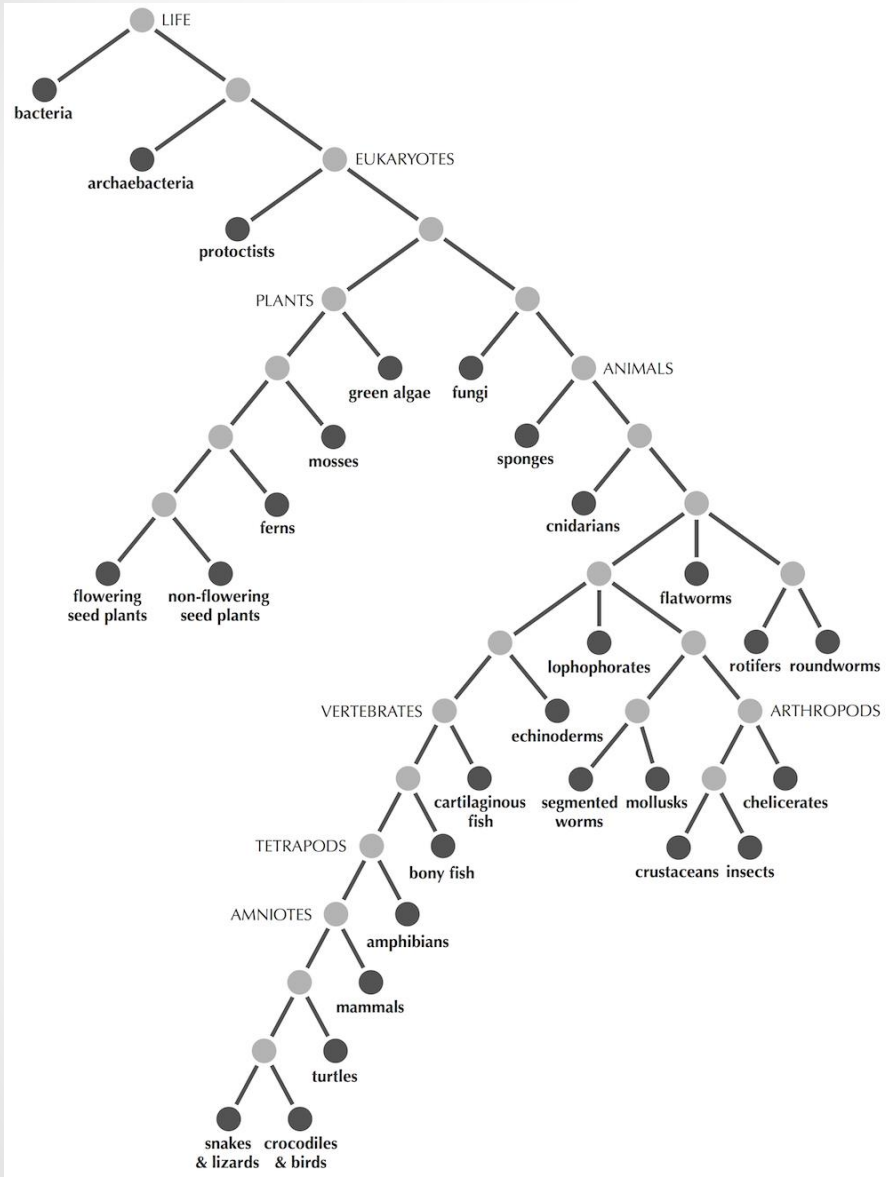
Trees



What degrees have the internal nodes of this tree

A connected graph without cycles that models an evolutionary tree of life on Earth. Present-day species are shown as darker nodes.

Trees



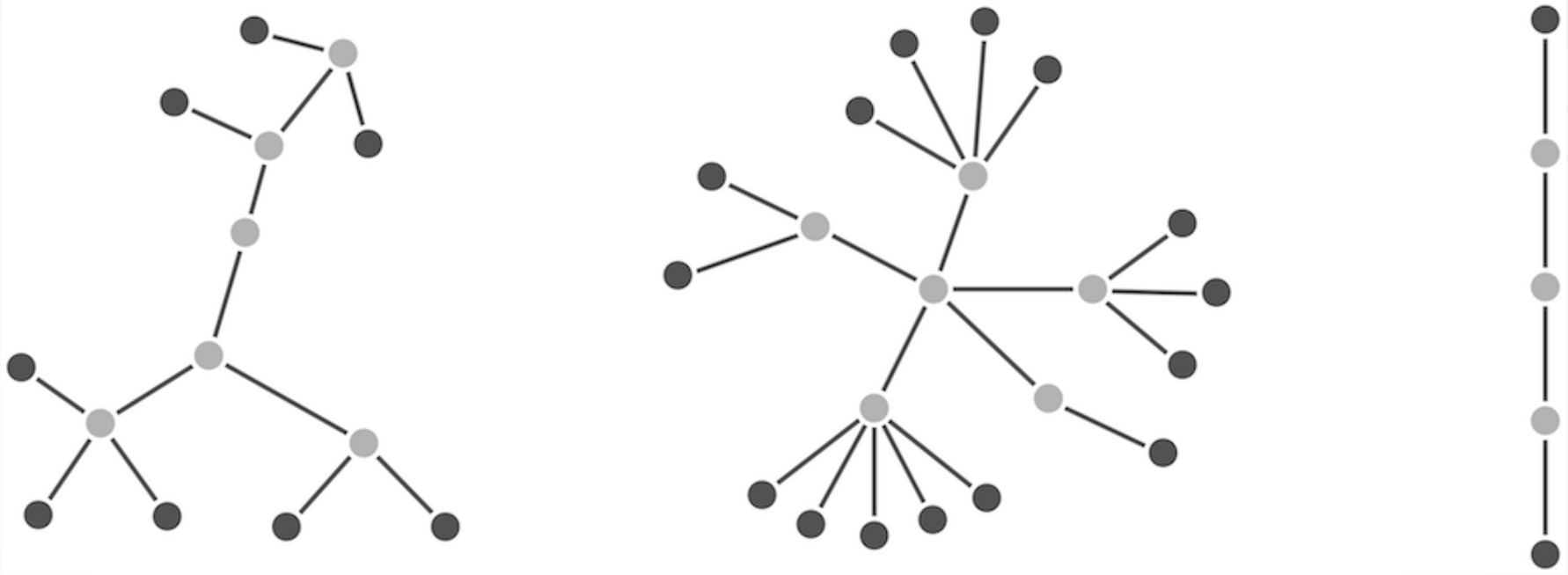
Tree: *Connected graph containing no cycles.*

Leaves (degree = 1): *present-day species.*

Internal nodes (degree>1): *ancestral species.*

A connected graph without cycles that models an evolutionary tree of life on Earth. Present-day species are shown as darker nodes.

Trees



Trees come in a variety of different shapes. In each of the three trees shown, leaves have been drawn darker than internal nodes.

Regardless the shape of the tree, the fact that it is connected graph leads to some common properties:

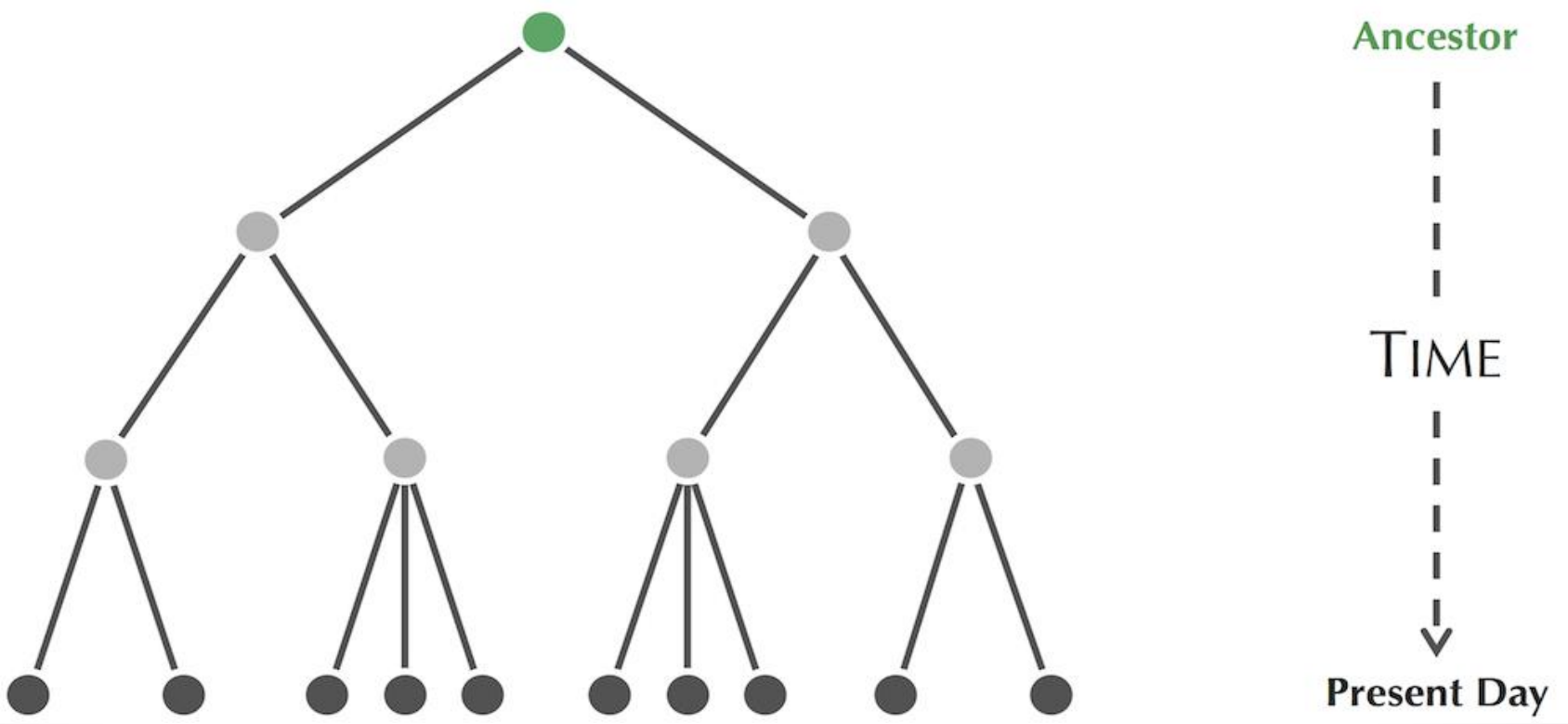
1. *Every tree with at least two nodes has at least two leaves.*
2. *Every tree with at least n nodes has exactly n-1 edges.*

Prove these two statements.

Exam Question:

How many nodes must a tree with 1167 edges have?

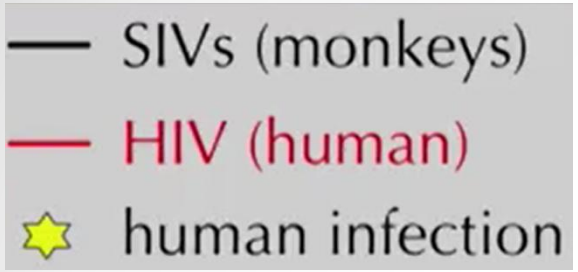
Trees



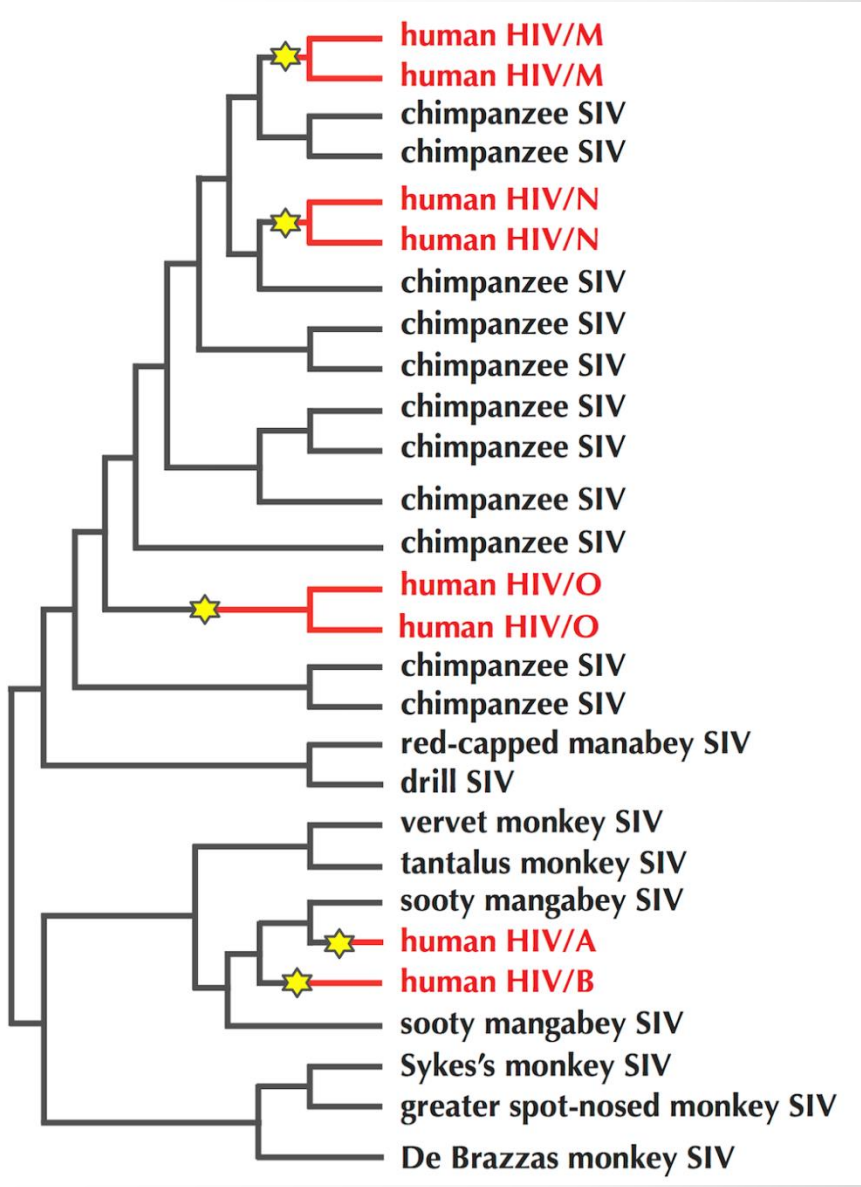
A rooted tree, with the root (representing an ancestor of all species in the tree) indicated in green at the top of the tree. The presence of the root implies an orientation of edges in the tree away from the root such that time flows downward from the root to the leaves in the sense that each edge of the tree connects an older species to a more recent species.

Rooted tree: one node is designated as the **root** (most recent common ancestor)

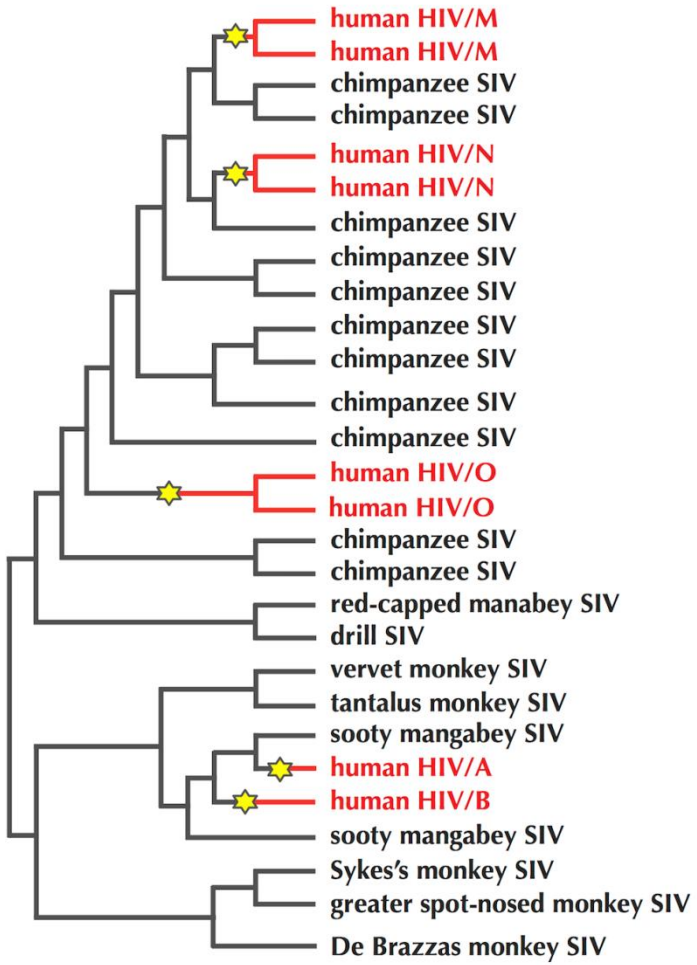
Example: HIV Evolutionary Tree



- HIV comprises five different viral families, denoted as A, B, M, N, and O, with the M family responsible for 95% of all HIV infections.
- The five families are different offshoots of the evolutionary tree for Simian Immunodeficiency Virus (SIV), which infects primates.
- Stars indicate viruses transitioning from primates to humans. The A and B families originated in sooty mangabey monkeys, whereas the M, N, and O families originated in chimpanzees.

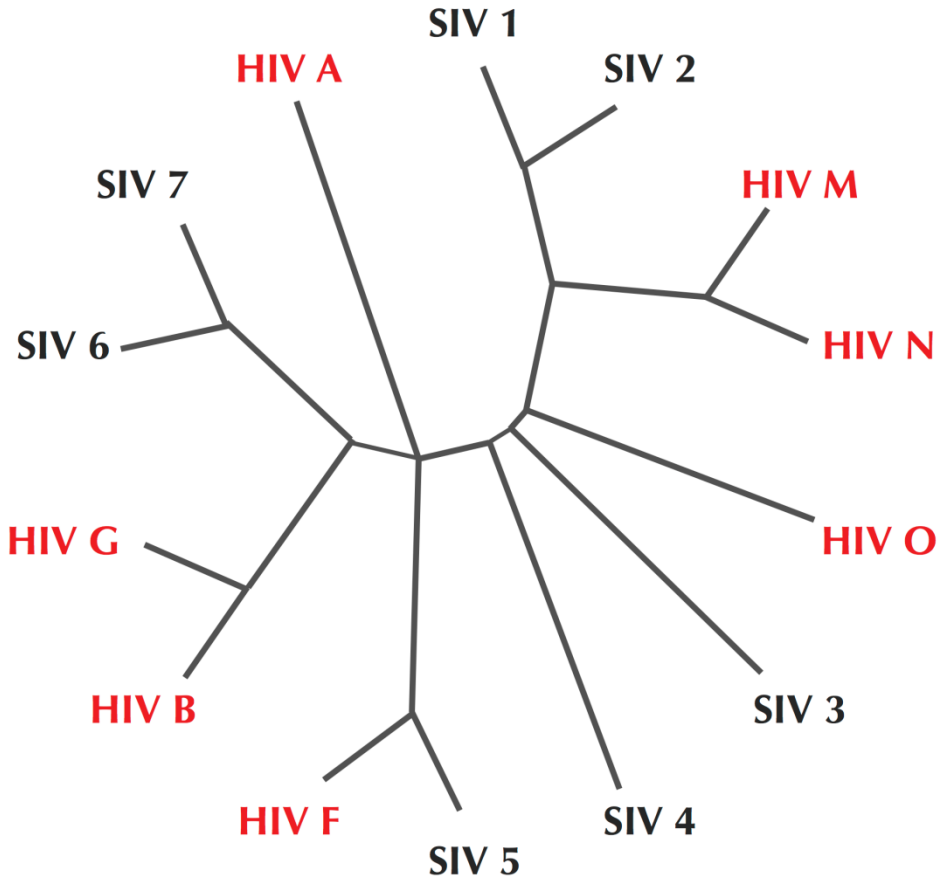


Trees



STOP and Think: Where would you place the root in this phylogeny?

Trees



An unrooted tree of HIV and SIV viruses produced from different dataset that suggests additional viral families F and G in addition to the viral families A, B, M, N, and O shown in the figure on the previous step.

What algorithms do we need to construct evolutionary trees

Outline

- **Transforming Distance Matrices into Evolutionary Trees**
- Toward an Algorithm for Distance-Based Phylogeny Construction
- Additive Phylogeny
- Ultrametric Evolutionary Trees (UPGMA reconstruction)
- The Neibour-Joining Algorithm
- Using Least-Squares to construct Distance-Based Phylogenies

Constructing a Distance Matrix

SPECIES	ALIGNMENT
Chimp	ACGTAGGCCT
Human	ATGTAAGACT
Seal	TCGAGAGCAC
Whale	TCGAAAGCAT

A toy multiple alignment of hypothetical DNA sequences from four species…

Constructing a Distance Matrix

$D_{i,j}$ = number of differing symbols between *i*-th and *j*-th rows of a multiple alignment.

Species	Alignment	Distance Matrix (*D*)			
		Chimp	Human	Seal	Whale
Chimp	ACGTAGGCCT	0	3	6	4
Human	ATGTAAGACT	3	0	7	5
Seal	TCGAGAGCAC	6	7	0	2
Whale	TCGAAAGCAT	4	5	2	0

A multiple alignment of hypothetical DNA sequences from four species, along with the distance matrix produced by counting the number of differing symbols between each pair of rows in this multiple alignment.

Constructing a Distance Matrix

$D_{i,j}$ = number of differing symbols between *i*-th and *j*-th rows of a multiple alignment.

Species	Alignment	Distance Matrix (*D*)				
		Chimp	Human	Seal	Whale	
Chimp	ACGTAGGCCT	0	**3** ($D_{2,1}$)	6	4	$D_{1,2} = D_{2,1} = 3$
Human	ATGTAAGACT	**3** ($D_{1,2}$)	0	7	5	
Seal	TCGAGAGCAC	6	7	0	2	
Whale	TCGAAAGCAT	4	5	2	0	

A multiple alignment of hypothetical DNA sequences from four species, along with the distance matrix produced by counting the number of differing symbols between each pair of rows in this multiple alignment.

Constructing a Distance Matrix

$D_{i,j}$ = number of differing symbols between *i*-th and *j*-th rows of a multiple alignment.

Species	Alignment	Distance Matrix (*D*)			
		Chimp	Human	Seal	Whale
Chimp	ACGTAGGCCT	0	3	6	4
Human	ATGTAAGACT	3	0	7	5
Seal	TCGAGAGCAC	6	7	0	2
Whale	TCGAAAGCAT	4	5	2	0

Regardless of which distance function we use, in order to be a distance matrix, *D* must satisfy three properties. It must be **symmetric** (for all *i* and *j*, $D_{i,j} = D_{j,i}$), **non-negative** (for all *i* and *j*, $D_{i,j} \geq 0$) and satisfy the **triangle inequality** (for all *i*, *j*, and *k*, $D_{i,j} + D_{j,k} \geq D_{i,k}$), where k is any third species.

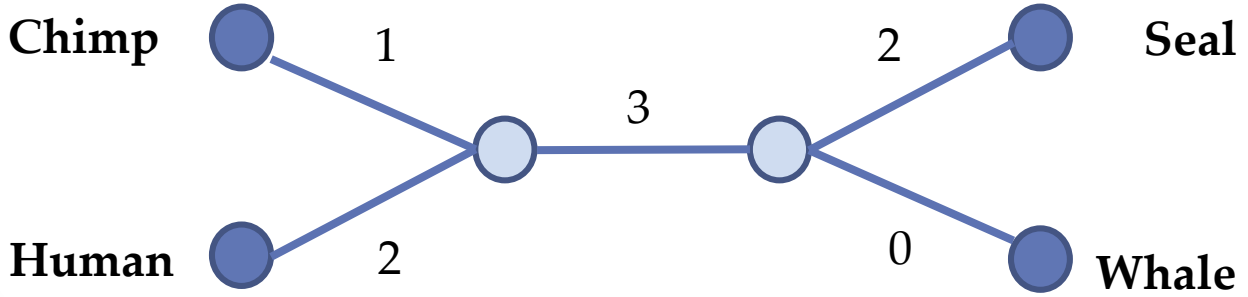
Fitting a Tree to a Matrix

	Chimp	**Human**	**Seal**	**Whale**
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0

The toy distance matrix constructed from a multiple alignment.

Fitting a Tree to a Matrix

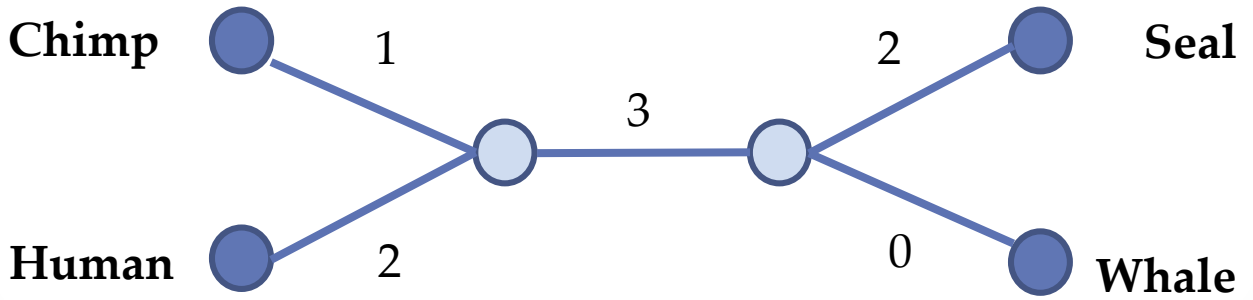
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

Fitting a Tree to a Matrix

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

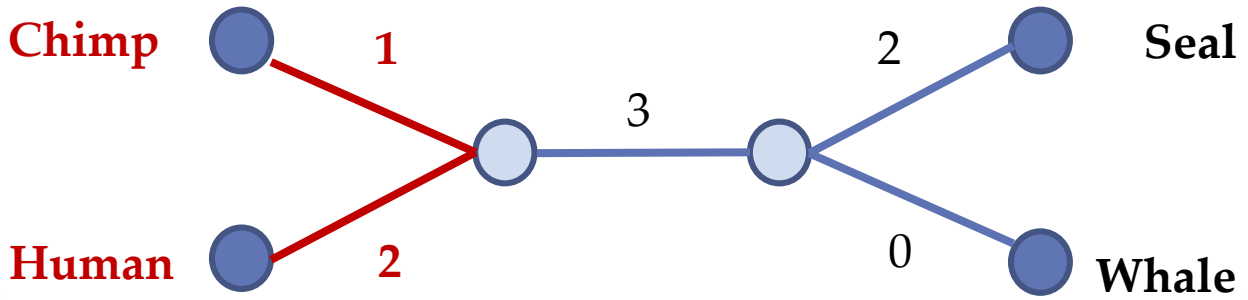
Deriving an unrooted tree from a distance matrix:

1) The leaves of this tree should correspond to the species represented by the matrix.

2) Assign each edge a non-negative length representing the evolutionary distance between the organisms that the edge connects.

Fitting a Tree to a Matrix

	Chimp	Human	Seal	Whale
Chimp	0	**3**	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0

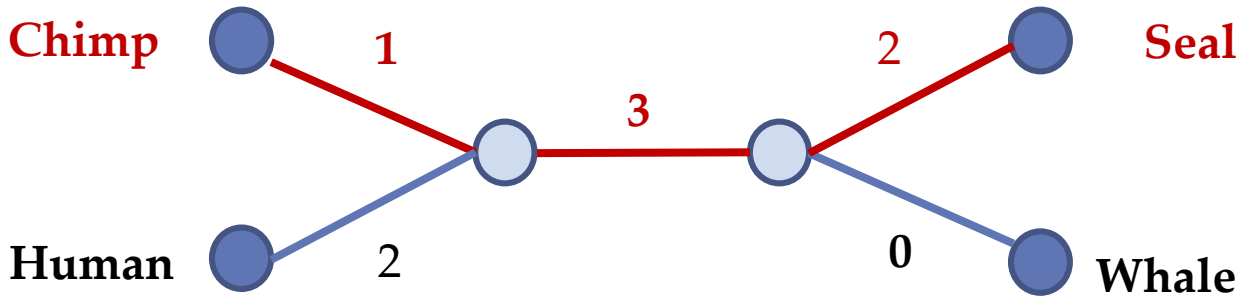


Unrooted tree fitting the distance matrix.

We will need to assign weights to the edges of this tree so that the sum of weights along a path that connects two leaves corresponds to the distance matrix value for those two leaves.

Fitting a Tree to a Matrix

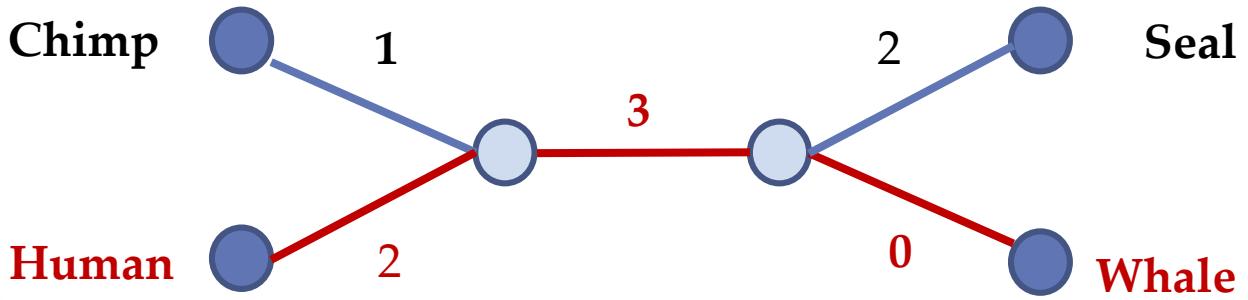
	Chimp	Human	Seal	Whale
Chimp	0	3	**6**	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

Fitting a Tree to a Matrix

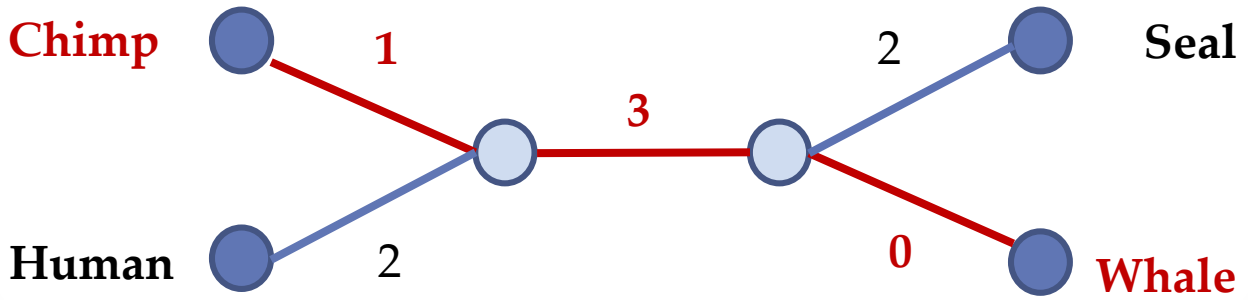
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	**5**
Seal	6	7	0	2
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

Fitting a Tree to a Matrix

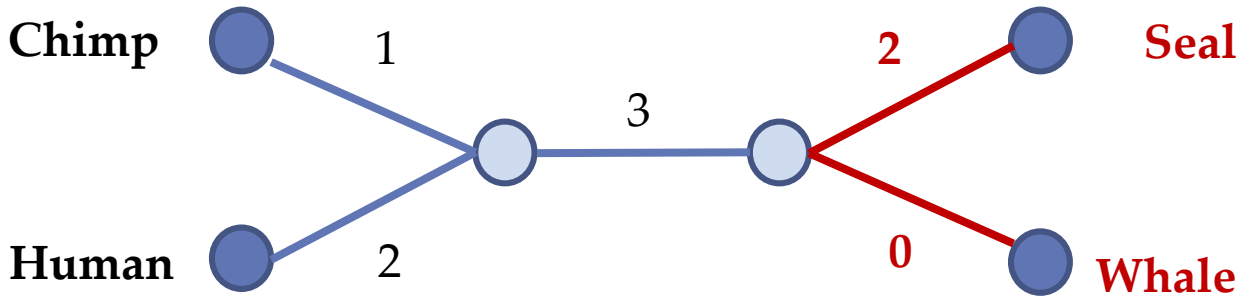
	Chimp	Human	Seal	Whale
Chimp	0	3	6	**4**
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

Fitting a Tree to a Matrix

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0



Unrooted tree fitting the distance matrix.

Distance Between Leaves

Distance Between Leaves Problem: *Compute the distances between leaves in a weighted tree.*

- **Input:** A weighted tree with *n* leaves.
- **Output:** An *n* x *n* matrix $(d_{i,j})$, where $d_{i,j}$, is the length of the path between leaves *i* and *j*.

If a tree whose edges are weighted is given, it is pretty straightforward problem to just compute the distances between each pair of leaves.

Distance-Based Phylogeny Problem

Distance-Based Phylogeny Problem: *Construct an evolutionary tree from a distance matrix.*

- **Input:** A distance matrix.
- **Output:** The unrooted tree "fitting" this distance matrix.

STOP and Think: Does the Distance-Based Phylogeny Problem always have a solution?

Return to Distance-Based Phylogeny

Exercise Break: Try fitting a tree to the following 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

No Tree fits this matrix

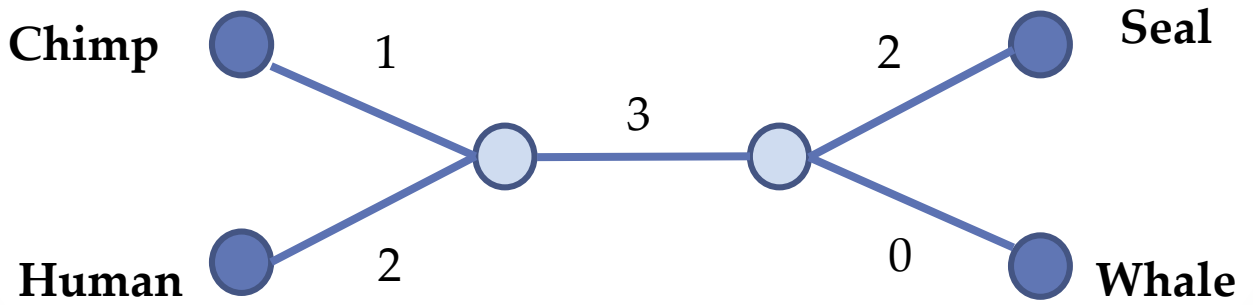
Exercise Break: Try fitting a tree to the following matrix.

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

Additive matrix: *distance matrix such that there exists an unrooted tree fitting it.*

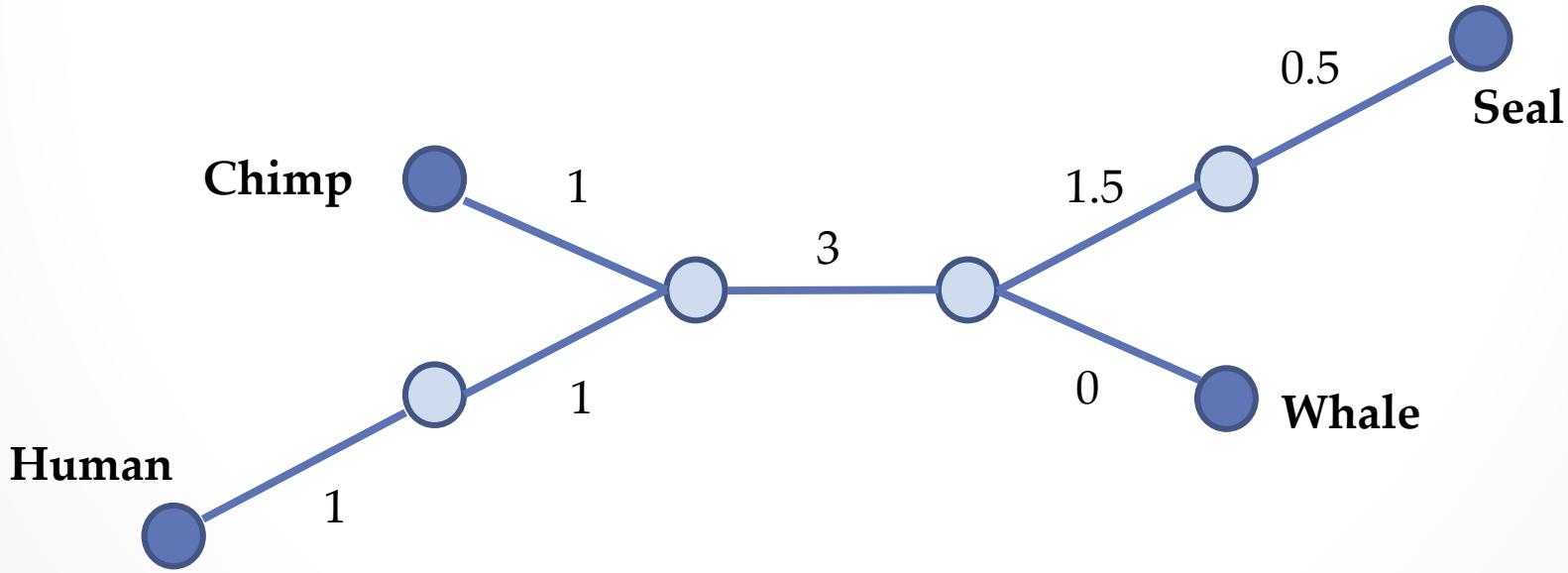
More than one Tree fits a matrix

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



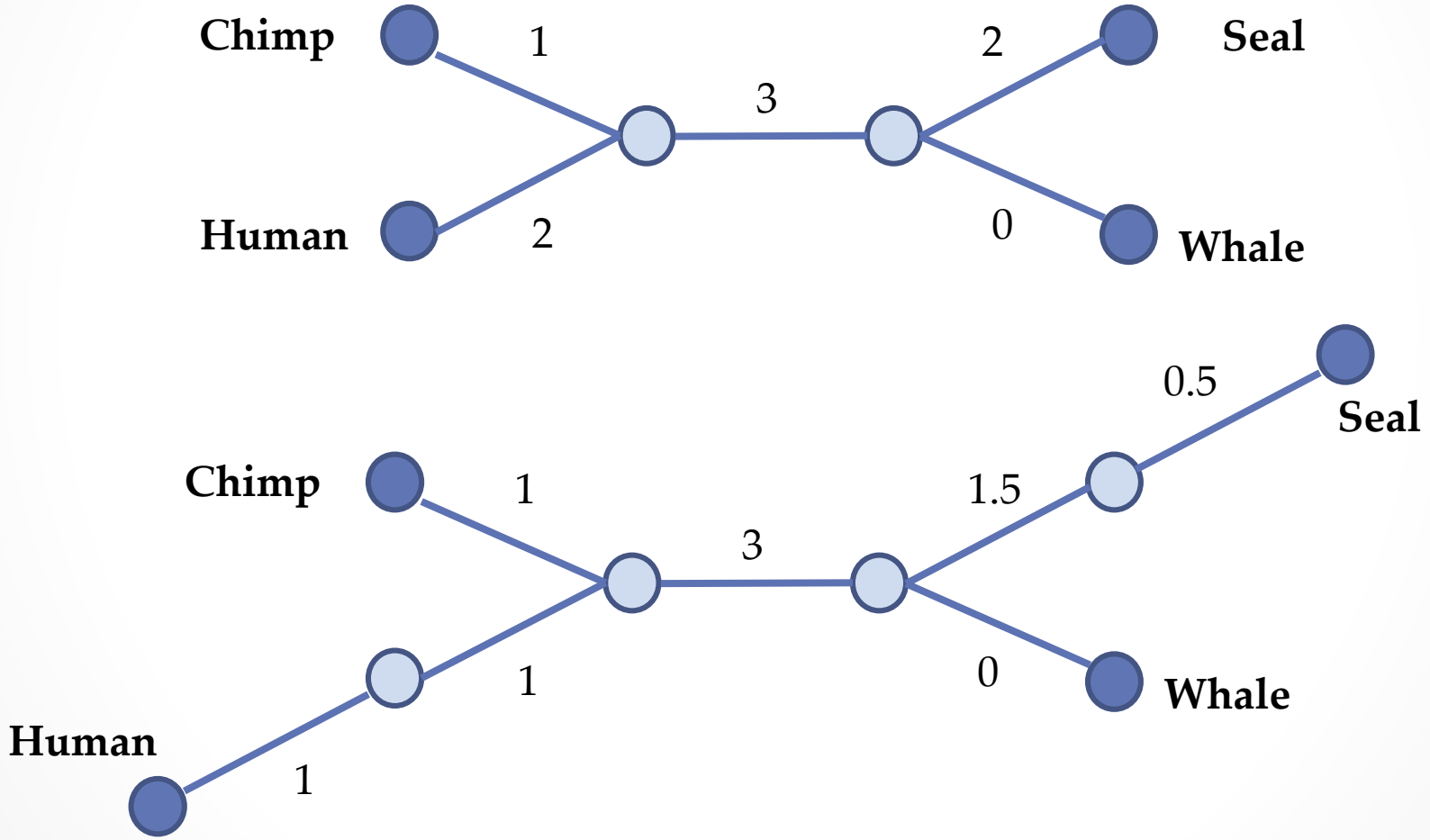
More than one Tree fits a matrix

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0

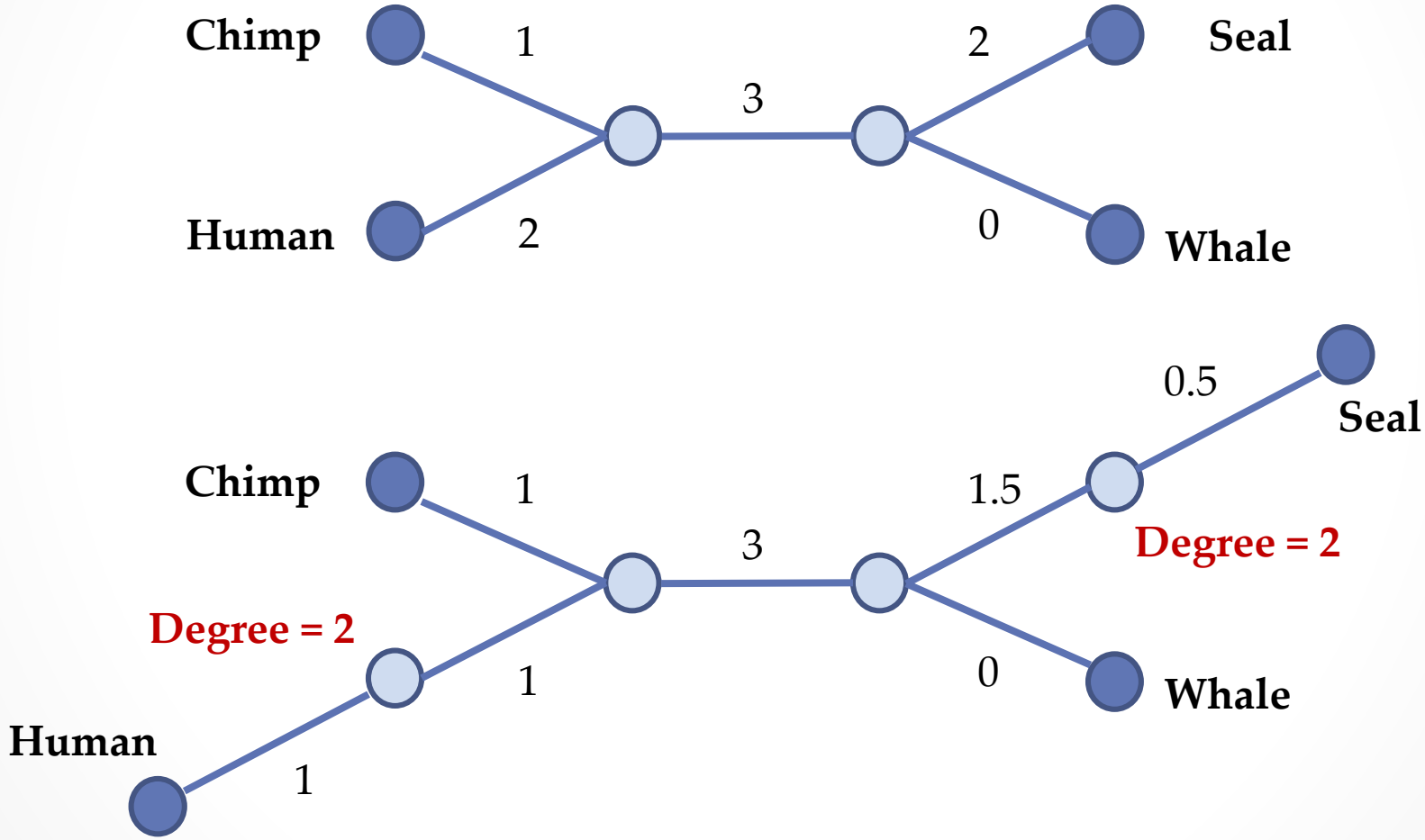


We simply stretch out the edges of the tree we had before into longer paths and still have a tree that fits the distance matrix.

Which Tree is "Better"?

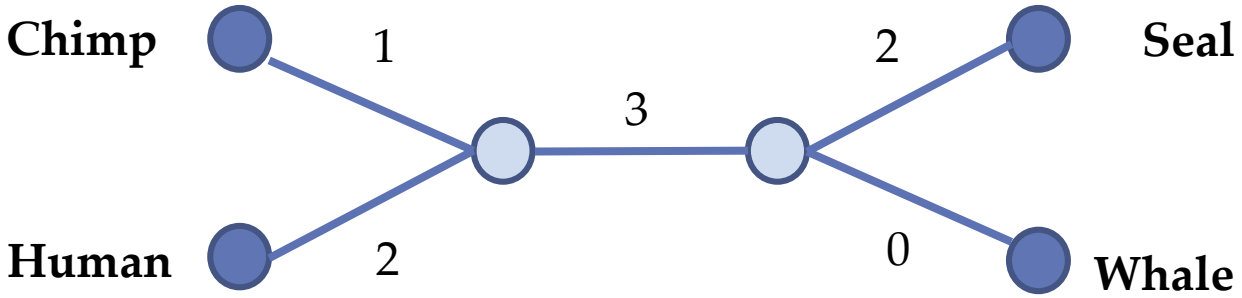


Which Tree is "Better"?



- We can get the graph on the top from the graph on the bottom by simply compressing all paths that contain degree 2 nodes.

Which Tree is "Better"?



Simple tree: *tree with no nodes of degree 2.*

Theorem: There is a *unique* simple tree fitting an *additive* matrix.

Every simple tree with *n* leaves has at most *n*-2 internal nodes.

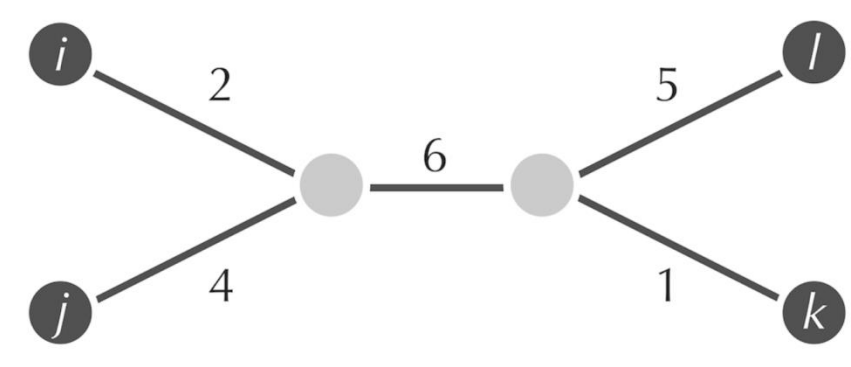
Reformulating Distance-Based Phylogeny

Distance-Based Phylogeny Problem: *Construct an evolutionary tree from a distance matrix.*

- **Input:** A distance matrix.
- **Output:** The **simple tree** "fitting" this distance matrix (if this matrix is **additive**).

Exam Question:

Which of the following matrices fit to the tree shown below?



a)

	i	*j*	*k*	*l*
i	0	6	9	13
j	6	0	11	15
k	9	11	0	6
l	13	15	6	0

b)

	i	*j*	*k*	*l*
i	0	6	10	13
j	6	0	12	15
k	10	12	0	7
l	13	15	7	0

c)

	i	*j*	*k*	*l*
i	0	6	9	14
j	6	0	11	16
k	9	11	0	7
l	14	16	7	0

d)

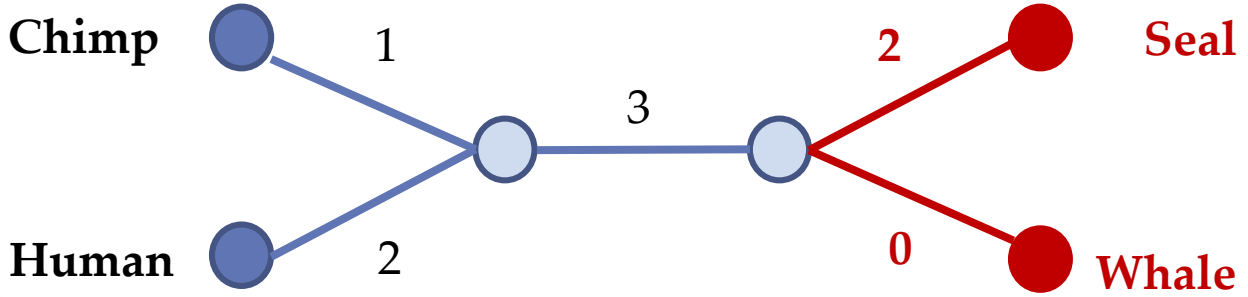
	i	*j*	*k*	*l*
i	0	6	10	14
j	6	0	12	16
k	10	12	0	6
l	14	16	6	0

Outline

- Transforming Distance Matrices into Evolutionary Trees
- **Toward an Algorithm for Distance-Based Phylogeny Construction**
- Additive Phylogeny
- Ultrametric Evolutionary Trees (UPGMA reconstruction)
- The Neibour-Joining Algorithm
- Using Least-Squares to construct Distance-Based Phylogenies

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0

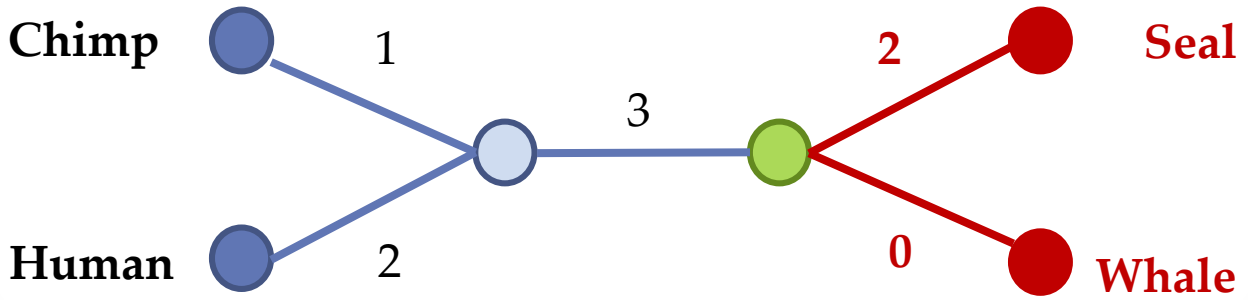


The minimum element of this matrix corresponds to two leaves that are next to each other on the tree.

An Idea of Distance-Based Phylogeny

Seal and whale are **neighbors** (meaning they share the same **parent**).

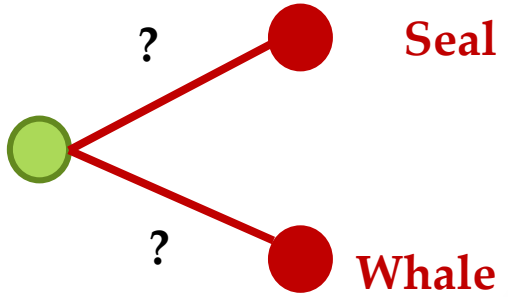
Theorem: Every simple tree with at least two nodes has at least one pair of neighboring leaves.



More than two leaves can share the same parent.

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0

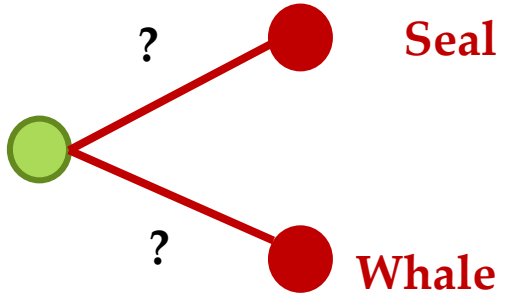


Let's pretend that we don't know the tree that fits the distance matrix, and see if we can use the fact, that seal and whale are neighbors in order to reconstruct the tree.

An Idea of Distance-Based Phylogeny

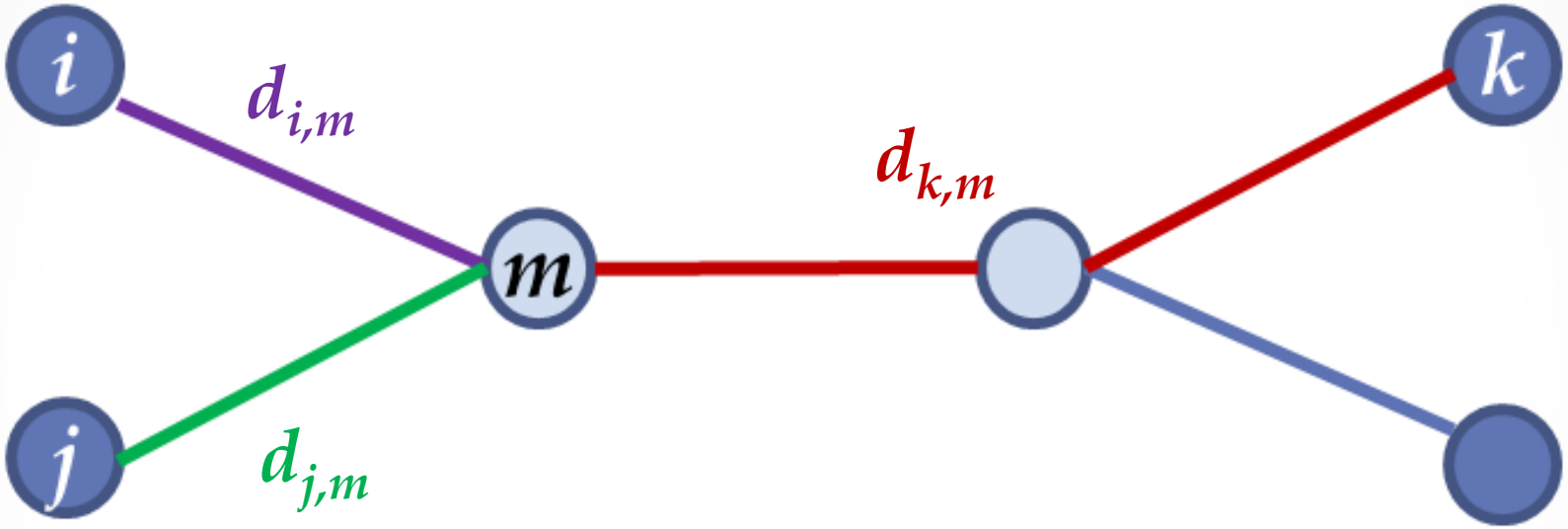
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0

STOP and Think: How do we compute the unknown distances?



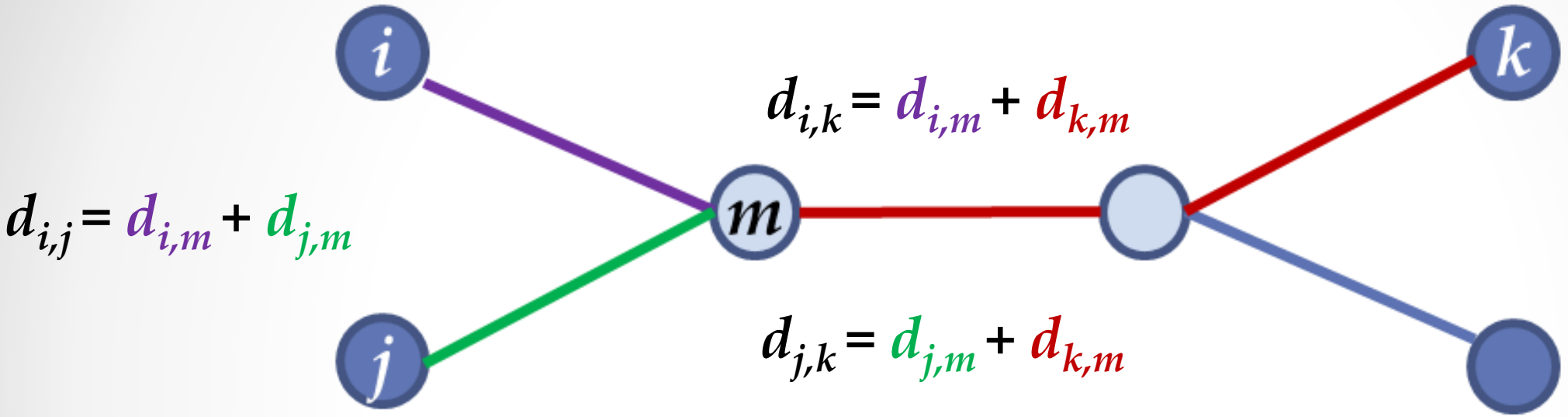
The sum of the question marks is equal to 2.

Toward a Recursive Algorithm

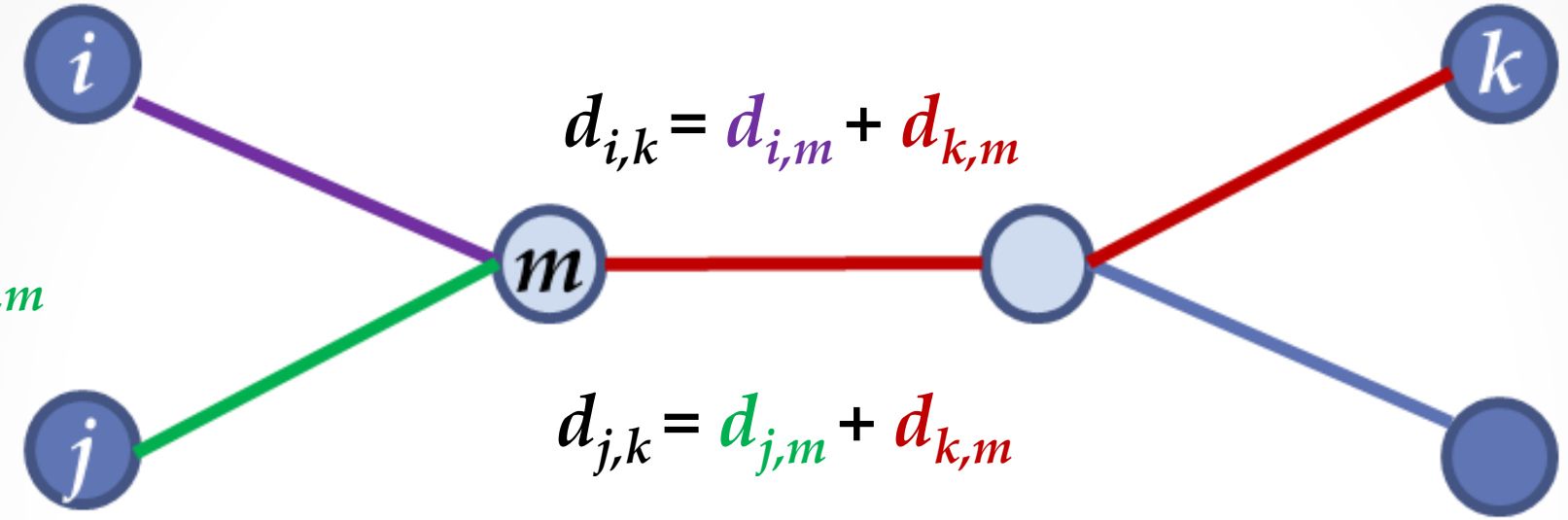


A tree with neighboring leaves ***i*** and ***j*** that share a parent ***m***. We try to reconstruct the ***green*** and ***purple*** distances. But if ***k*** is some other leaf in the tree, the ***red*** distance will help us out.

Toward a Recursive Algorithm



Toward a Recursive Algorithm



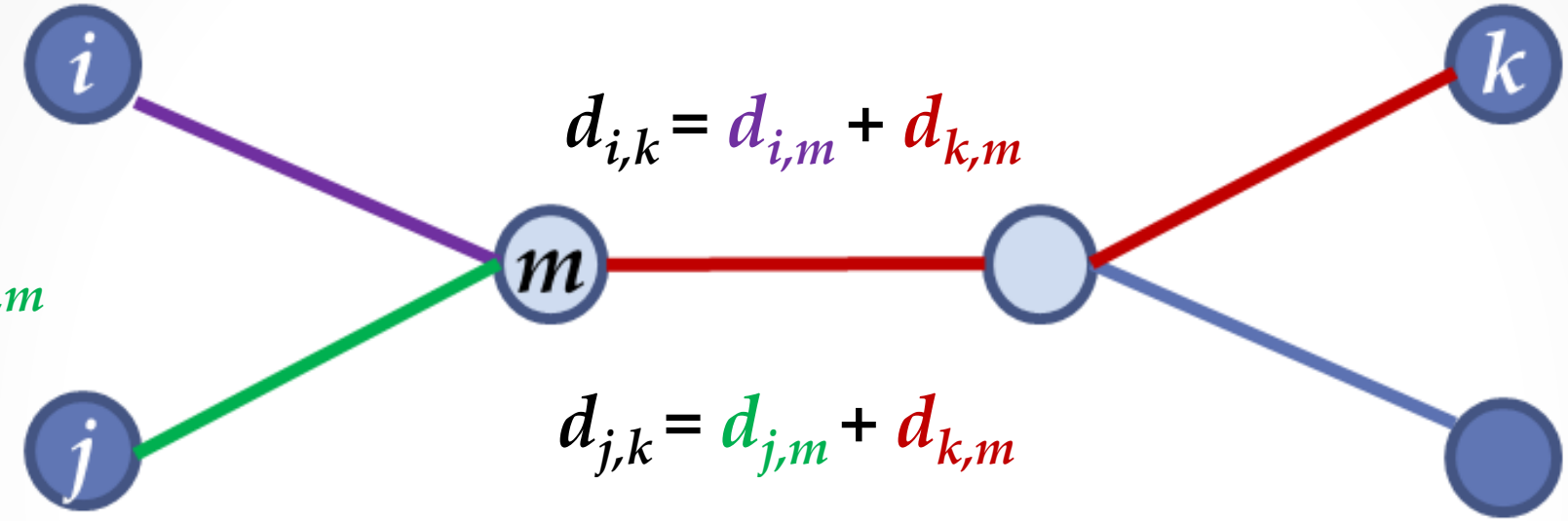
$d_{i,j} = d_{i,m} + d_{j,m}$

$d_{i,k} = d_{i,m} + d_{k,m}$

$d_{j,k} = d_{j,m} + d_{k,m}$

$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$

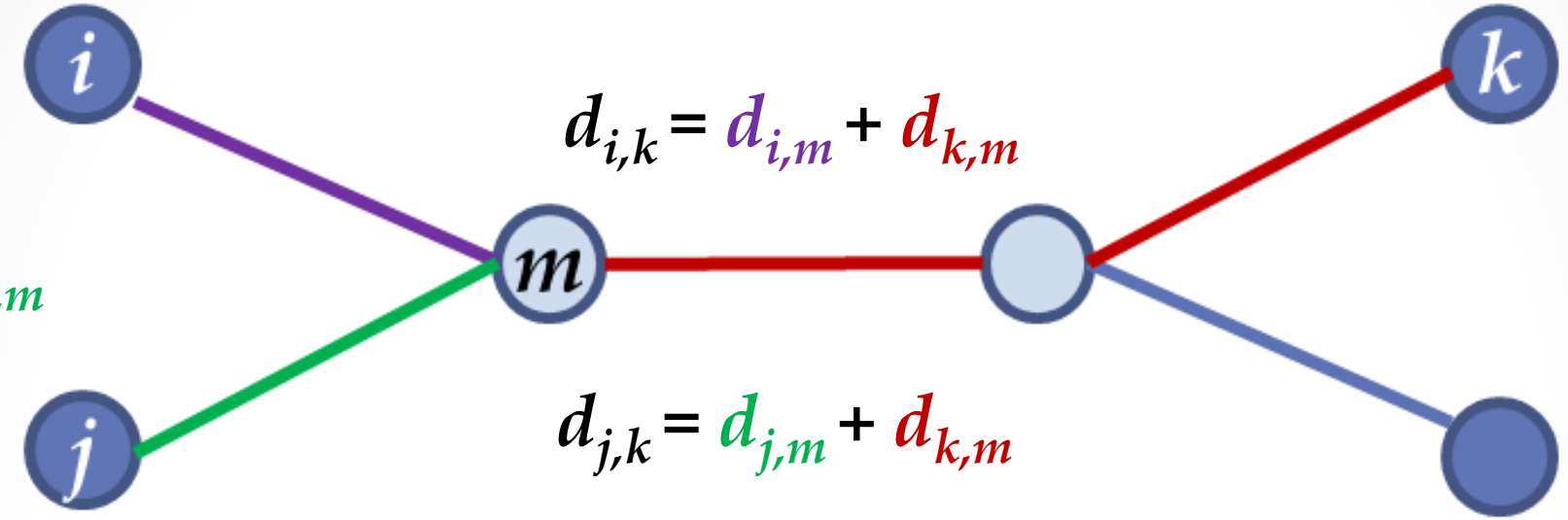
Toward a Recursive Algorithm



$d_{i,j} = d_{i,m} + d_{j,m}$

$d_{k,m} = [(\cancel{d_{i,m}} + d_{k,m}) + (\cancel{d_{j,m}} + d_{k,m}) - (\cancel{d_{i,j}} + \cancel{d_{j,m}})] / 2$

Toward a Recursive Algorithm

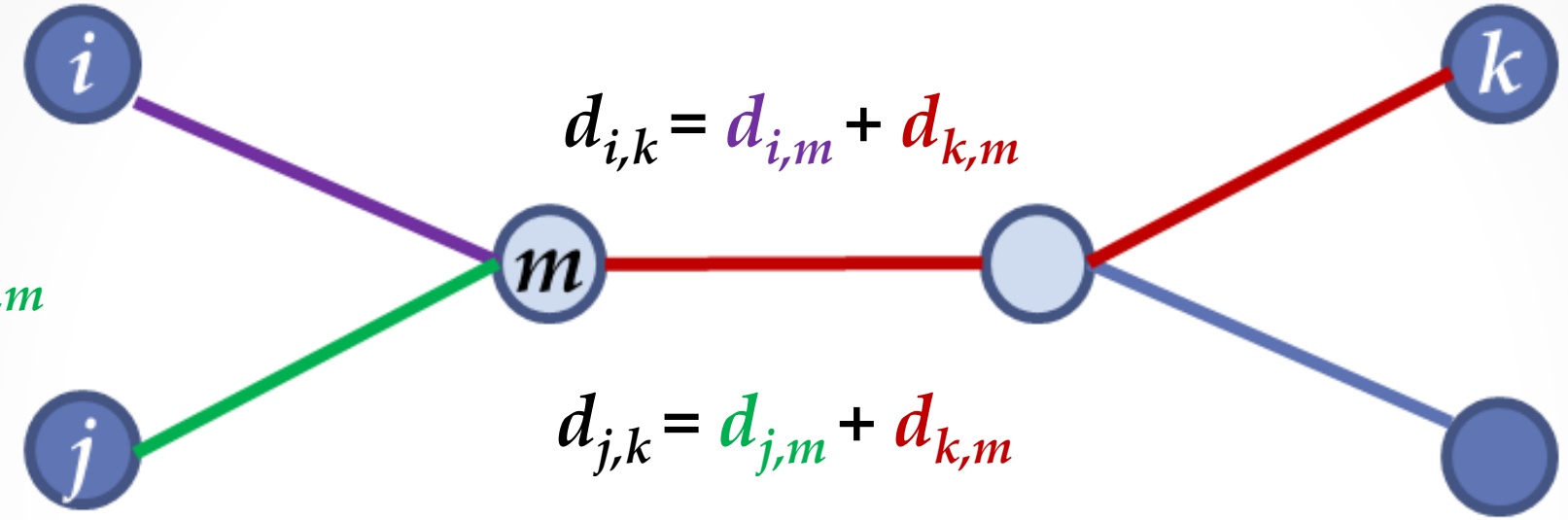


$d_{i,j} = d_{i,m} + d_{j,m}$

$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$

$d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j}) / 2$

Toward a Recursive Algorithm



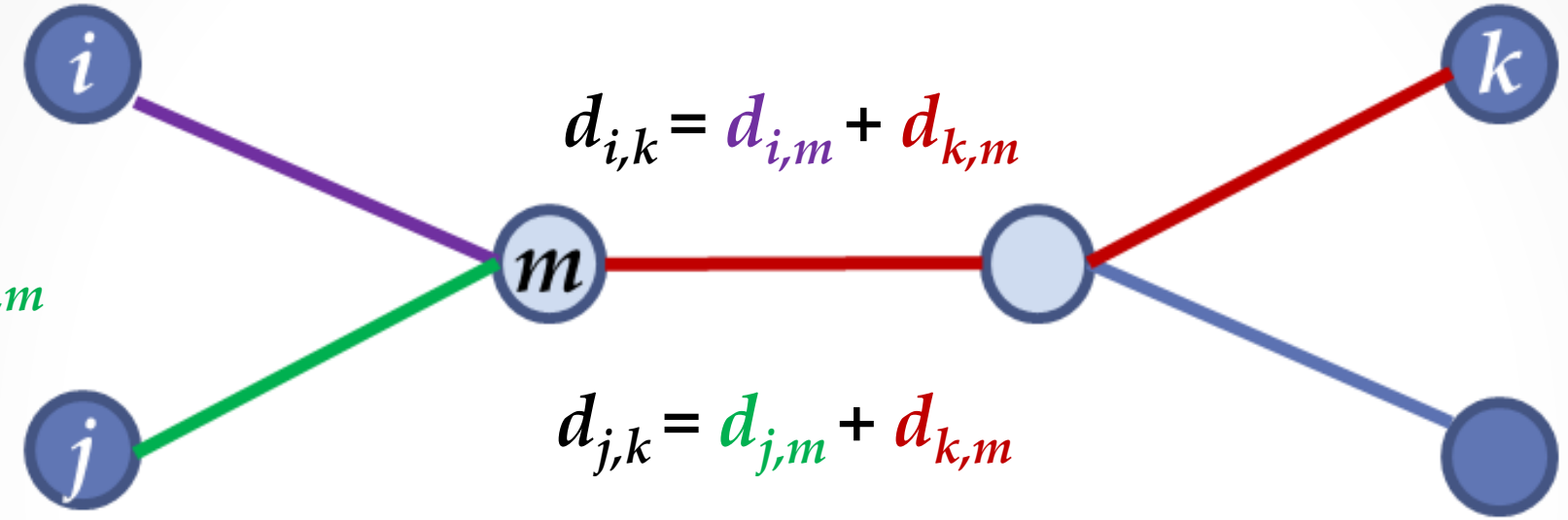
$$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$$

$$d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j})/ 2$$

$$d_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})/ 2$$

Important: It is not known a priori, what weights of the internal edges are, only distances between leaves. The expression on the right side is now written in terms of distances between leaves, so ***d*** could be substituted with ***D***.

Toward a Recursive Algorithm



$d_{i,j} = d_{i,m} + d_{j,m}$

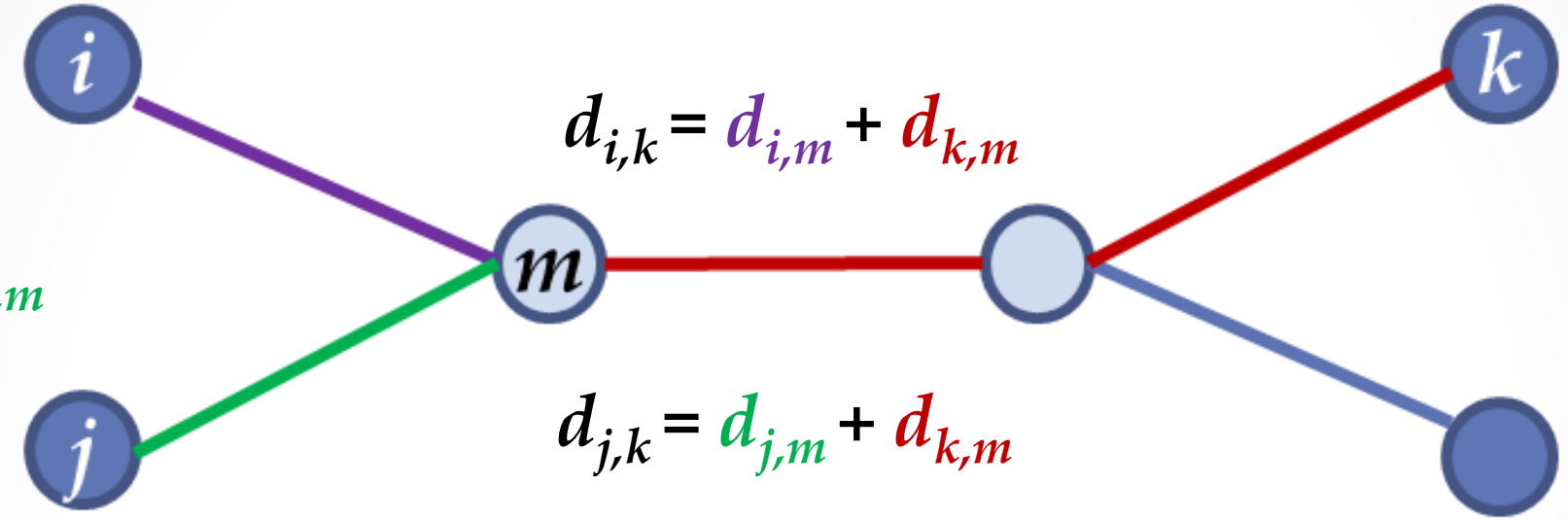
$$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$$

$$d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j})/ 2$$

$$d_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})/ 2$$

$$d_{i,m} = (D_{i,k} + D_{i,j} - D_{j,k})/ 2$$

Toward a Recursive Algorithm



$$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$$

$$d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j})/ 2$$

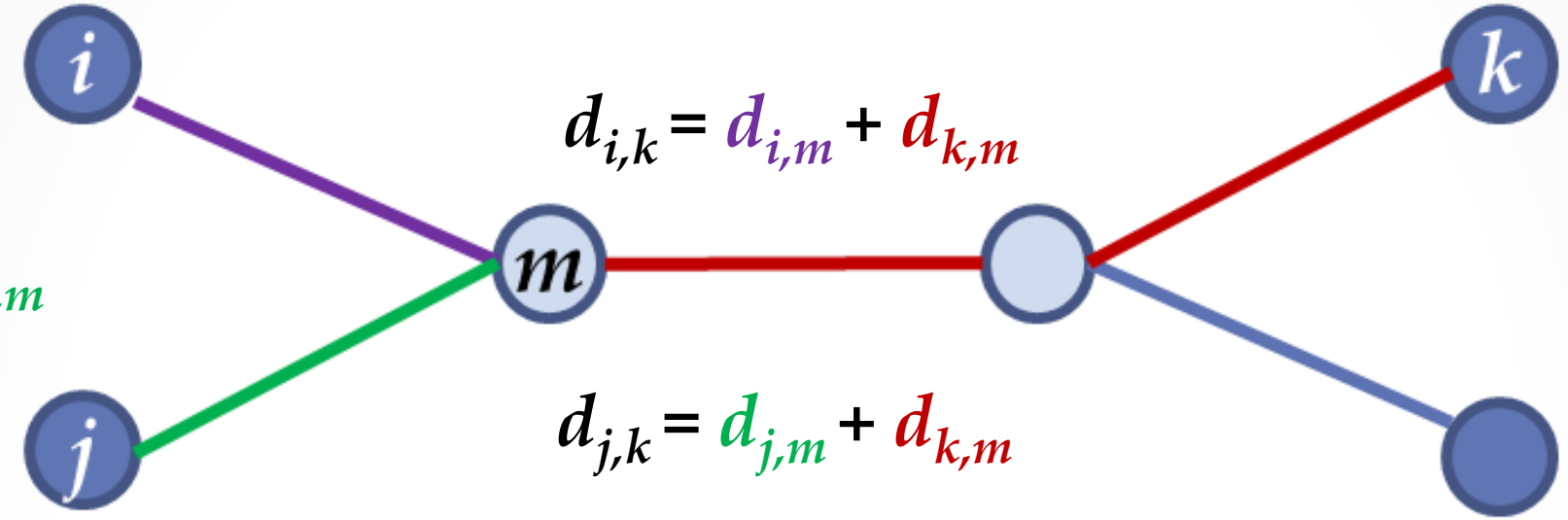
$$d_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})/ 2$$

$$d_{i,m} = (D_{i,k} + D_{i,j} - D_{j,k})/ 2$$

$$d_{j,m} = (D_{j,k} + D_{i,j} - D_{i,k})/ 2$$

Remember that ***k*** was just an **arbitrary leaf** other than the **neighboring leaves *i* and *j***.

Toward a Recursive Algorithm



$$d_{k,m} = [(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})] / 2$$

$$d_{k,m} = (d_{i,k} + d_{j,k} - d_{i,j}) / 2$$

$$d_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j}) / 2$$

$$d_{i,m} = (D_{i,k} + D_{i,j} - D_{j,k}) / 2$$

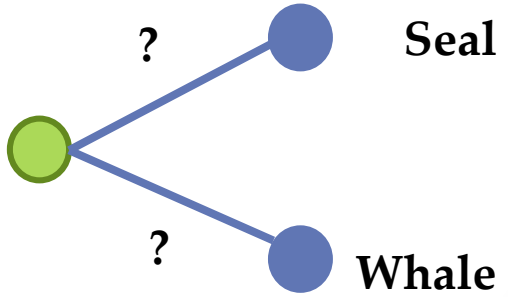
$$d_{j,m} = (D_{j,k} + D_{i,j} - D_{i,k}) / 2$$

If it is known that *i* and *j* are neighbors, we can compute the distance from them to their parent, just from the distance matrix alone.

An Idea of Distance-Based Phylogeny

The formula for the distance from a leaf to its parent

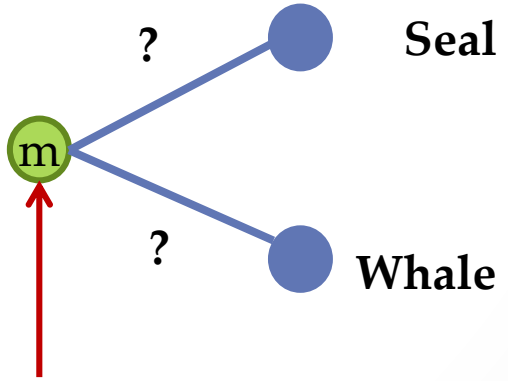
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0



$$d_{i,m} = (D_{i,k} + D_{i,j} - D_{j,k})/\,2$$

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0

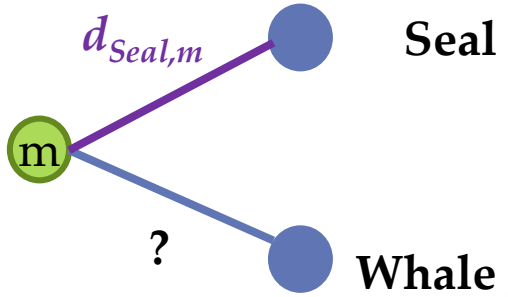


the parent of neighbors seal and whale

$$d_{i,m} = (D_{i,k} + D_{i,j} - D_{j,k})/ 2$$

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0

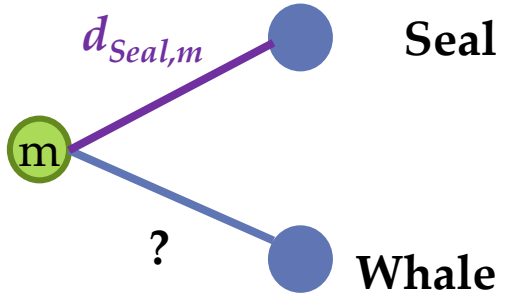


$$d_{Seal,m} = (D_{Seal,k} + D_{Seal,j} - D_{j,k})/\,2$$

i replaced with ***Seal***

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



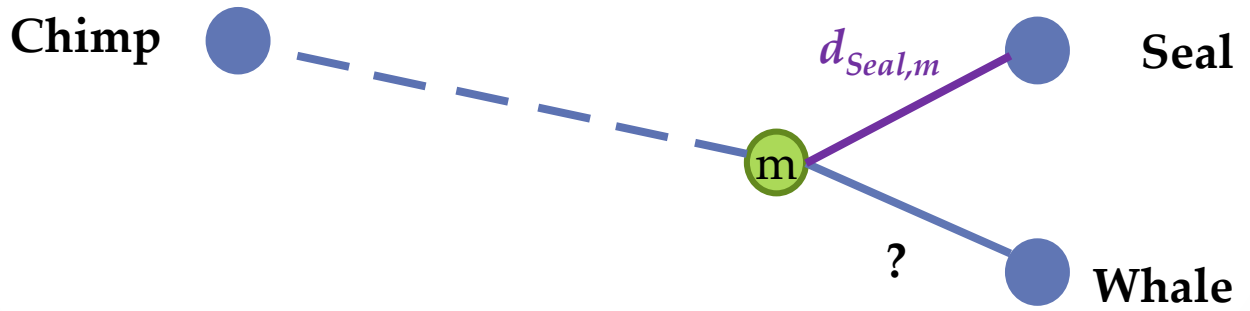
$$d_{Seal,m} = (D_{Seal,k} + D_{Seal,Whale} - D_{Whale,k})/ 2$$

i replaced with ***Seal***

j replaced with ***Whale***

An Idea of Distance-Based Phylogeny

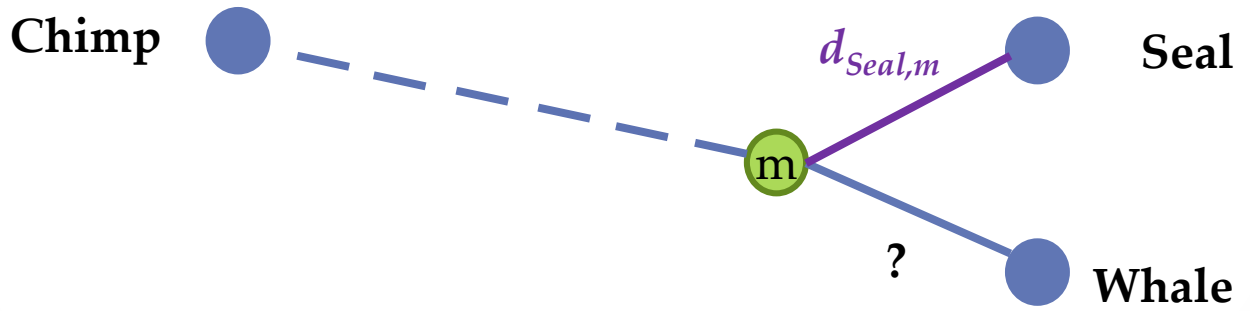
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



$$d_{Seal,m} = (D_{Seal,Chimp} + D_{Seal,Whale} - D_{Whale,Chimp})/ 2$$

An Idea of Distance-Based Phylogeny

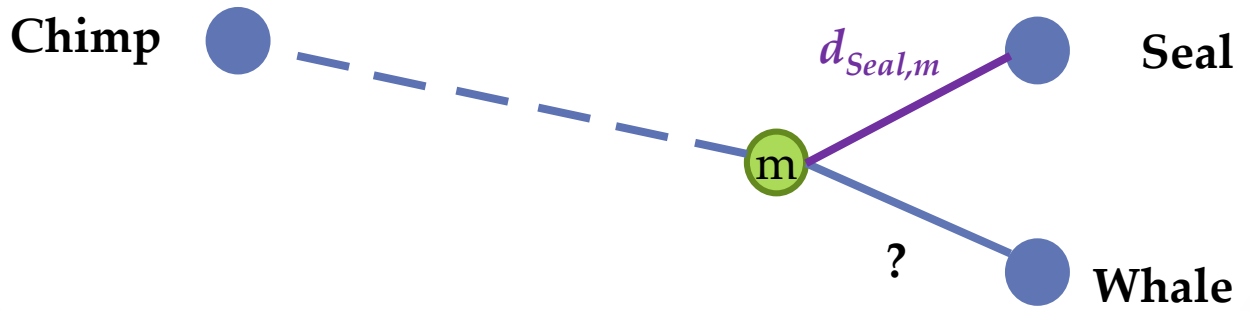
	Chimp	Human	Seal	Whale
Chimp	0	3	**6**	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



$$d_{Seal,m} = (\ 6\ + D_{Seal,Whale} - D_{Whale,Chimp})/\ 2$$

An Idea of Distance-Based Phylogeny

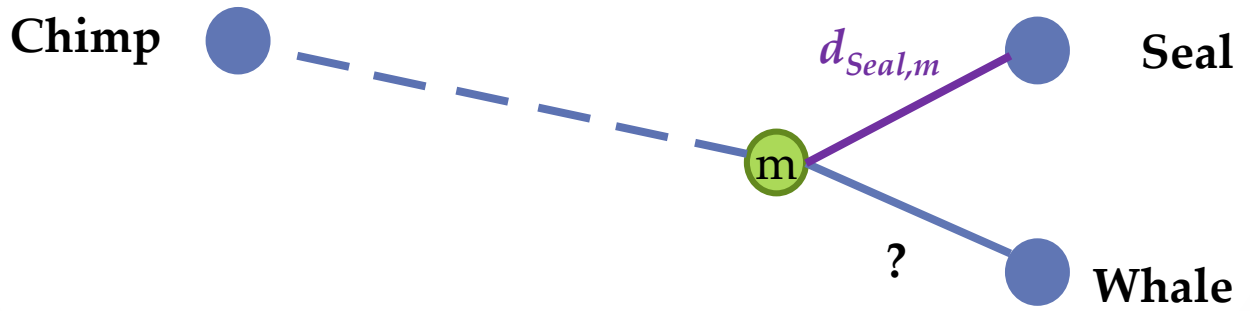
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	**2**	0



$$d_{Seal,m} = (\ 6\ +\ 2\ - D_{Whale,Chimp})/\ 2$$

An Idea of Distance-Based Phylogeny

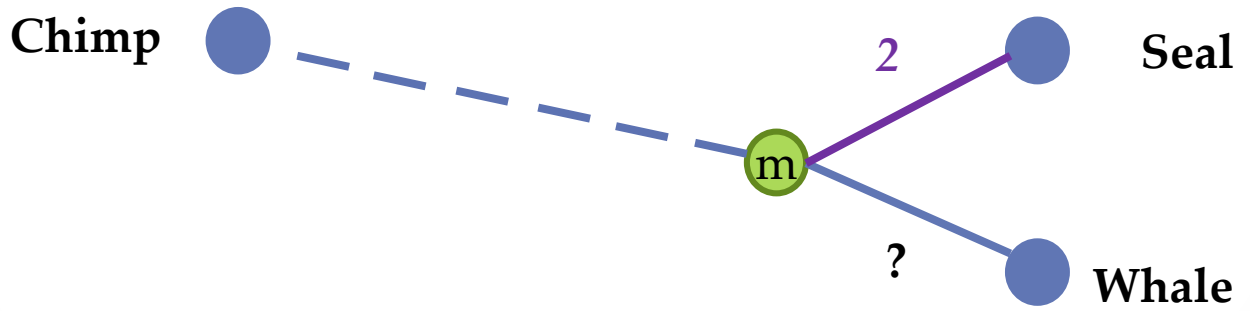
	Chimp	Human	Seal	Whale
Chimp	0	3	6	**4**
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



$$d_{Seal,m} = (\ 6\ +\ 2\ -\ 4\ )/\ 2$$

An Idea of Distance-Based Phylogeny

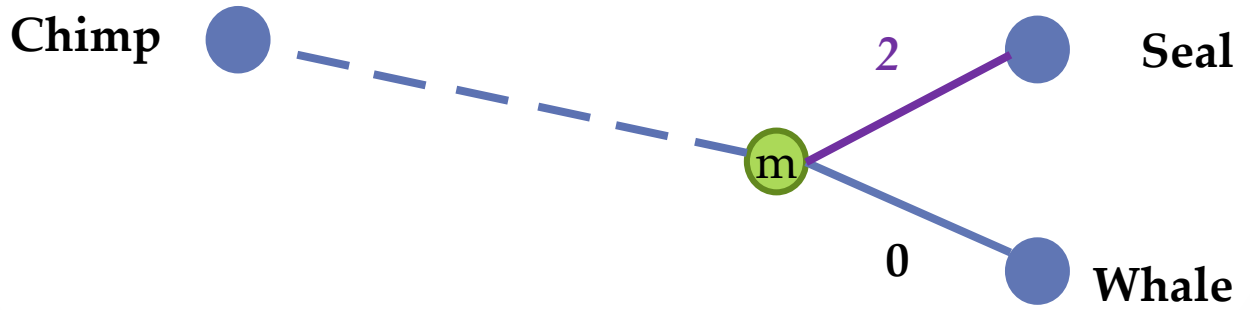
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



$d_{Seal,m} = 2$

An Idea of Distance-Based Phylogeny

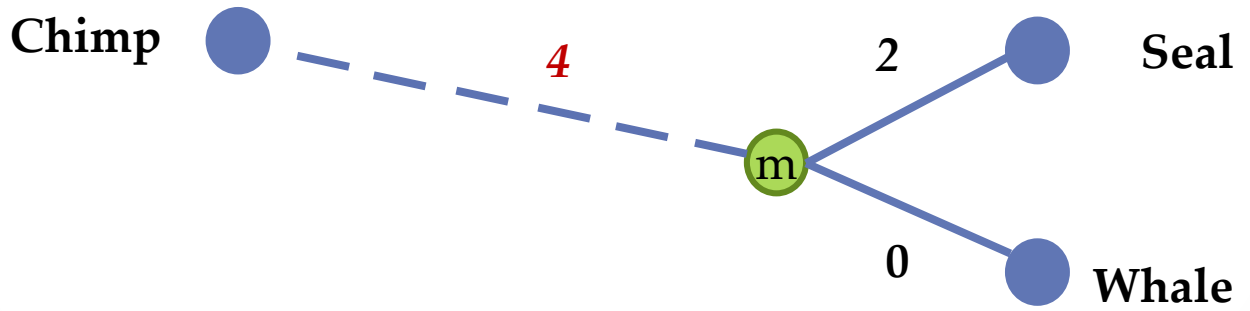
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	**2**
Whale	4	5	2	0



$d_{Seal,m} = 2$

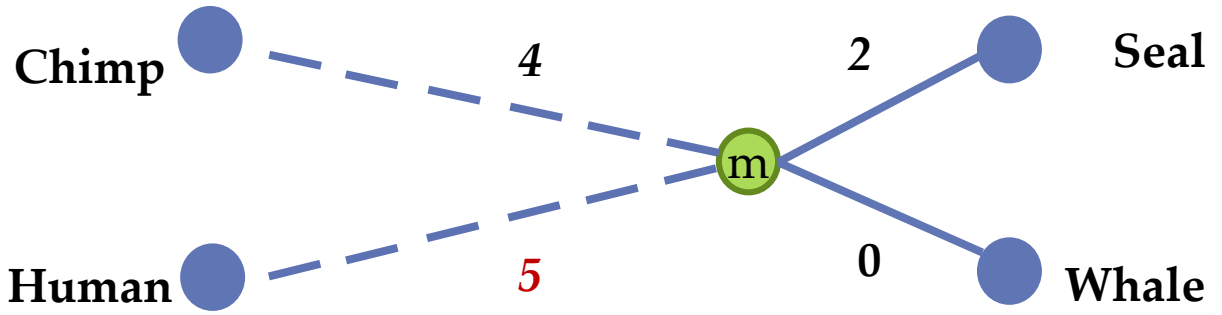
An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	**6**	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



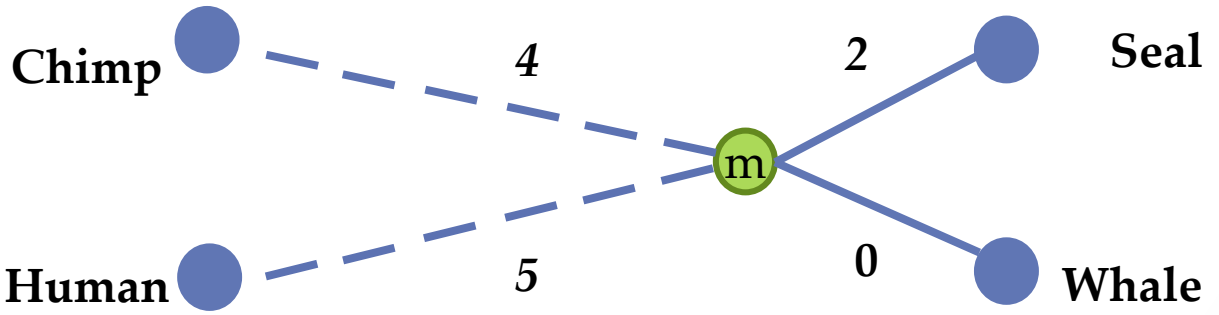
An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	**7**	5
Seal	6	7	0	2
Whale	4	5	2	0



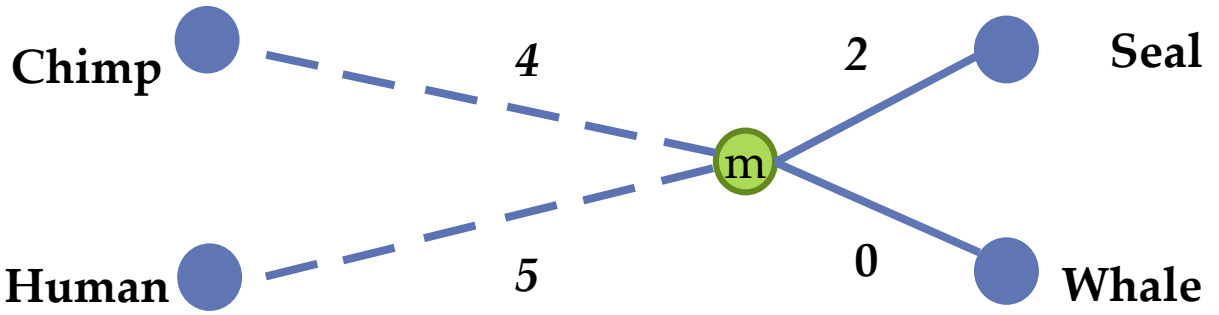
An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale	m
Chimp	0	3	6	4	4
Human	3	0	7	5	5
Seal	6	7	0	2	2
Whale	4	5	2	0	0
m	4	5	2	0	0



An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale	m
Chimp	0	3	6	4	4
Human	3	0	7	5	5
Seal	6	7	0	2	2
Whale	4	5	2	0	0
m	4	5	2	0	0

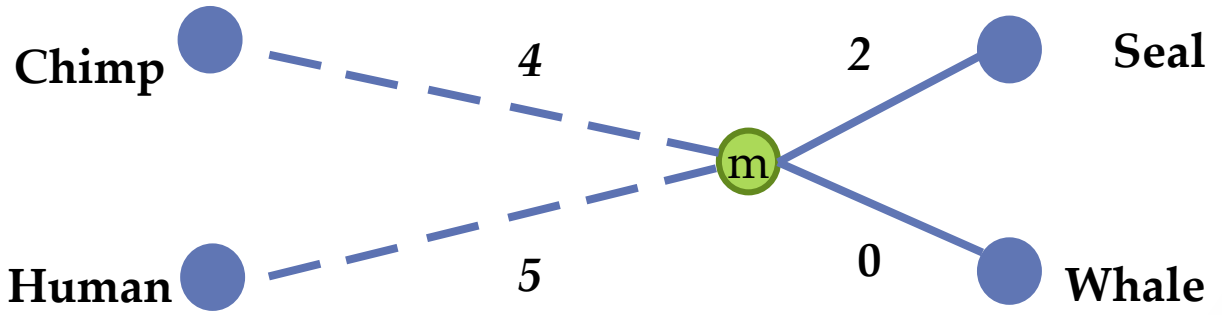


As we already added ***Seal*** and ***Whale*** to the tree, their columns could be ignored.

An Idea of Distance-Based Phylogeny

	Chimp	Human	m
Chimp	0	3	4
Human	3	0	5
m	4	5	0

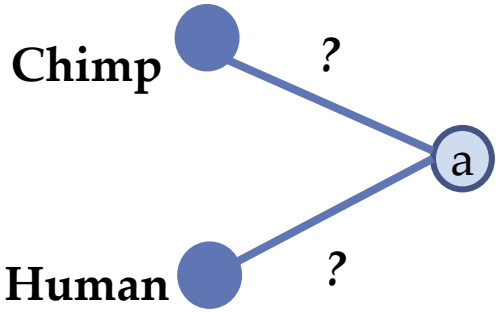
Getting rid of Seal and Whale entirely yields a smaller 3x3 matrix.

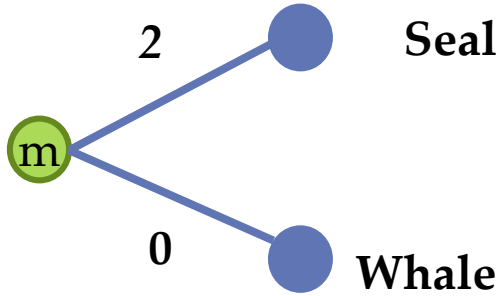


An Idea of Distance-Based Phylogeny

	Chimp	Human	m
Chimp	0	**3**	4
Human	**3**	0	5
m	4	5	0

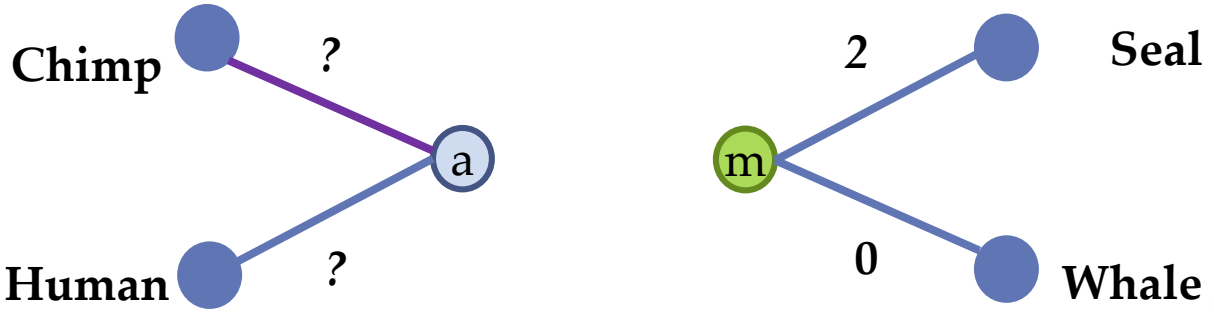
Now we just apply the same rules recursively. As min value of the matrix is three, so ***Chimp*** and ***Human*** must be neighbors with the parent ***a***.





An Idea of Distance-Based Phylogeny

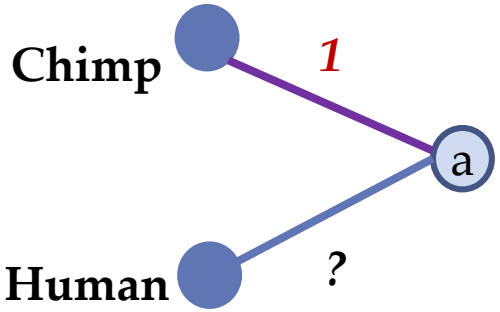
	Chimp	Human	m
Chimp	0	**3**	4
Human	**3**	0	5
m	4	5	0

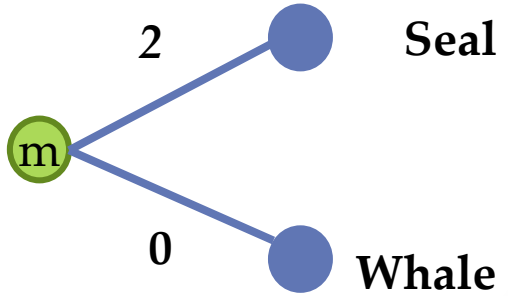


$$d_{Chimp,a} = (D_{Chimp,m} + D_{Chimp,Human} - D_{Human,m})/ 2$$

An Idea of Distance-Based Phylogeny

	Chimp	Human	m
Chimp	0	**3**	4
Human	**3**	0	5
m	4	5	0

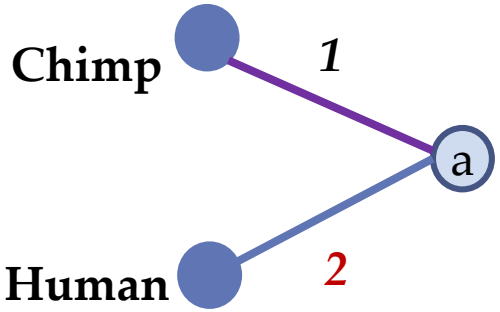


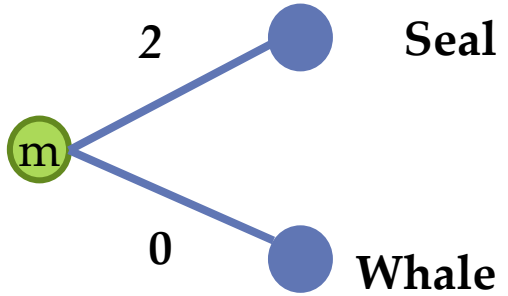


$d_{Chimp,a} = 1$

An Idea of Distance-Based Phylogeny

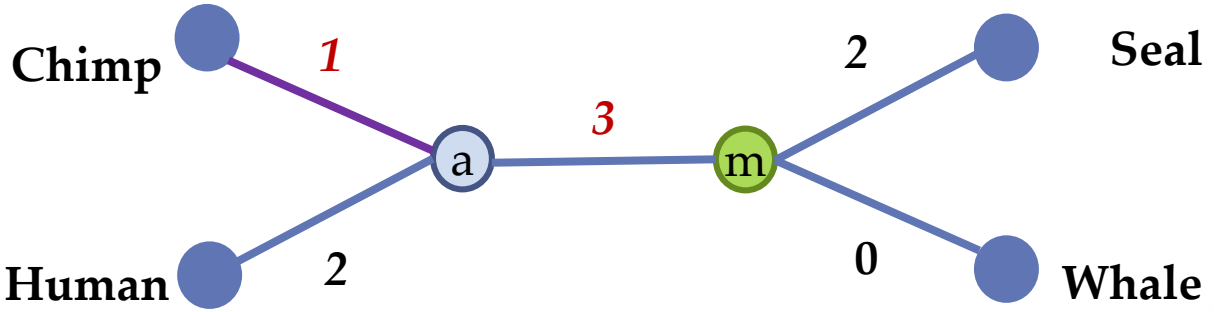
	Chimp	Human	m
Chimp	0	**3**	4
Human	3	0	5
m	4	5	0





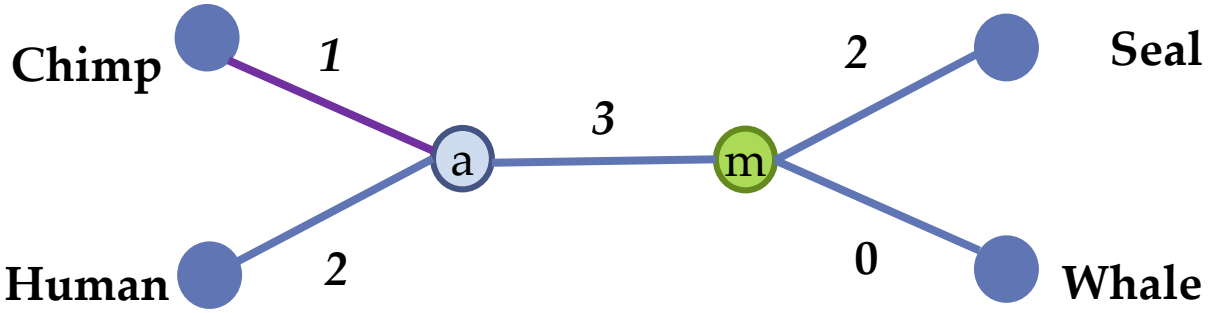
An Idea of Distance-Based Phylogeny

	Chimp	Human	m
Chimp	0	3	**4**
Human	3	0	5
m	4	5	0



An Idea of Distance-Based Phylogeny

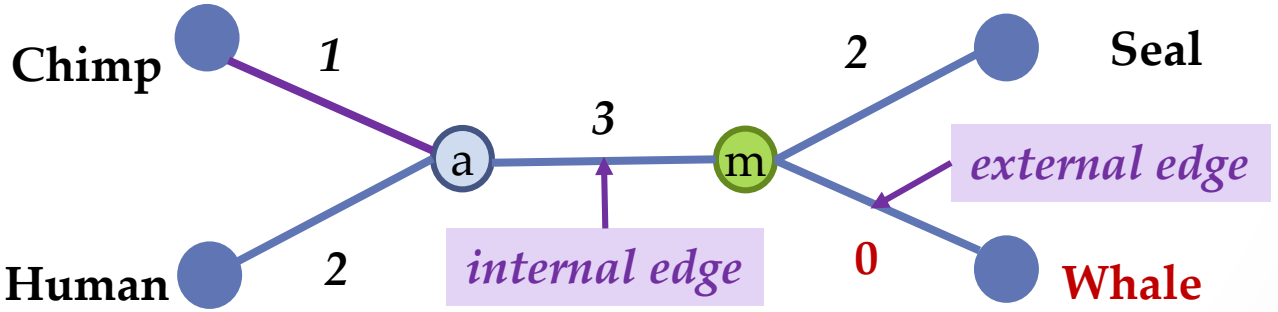
	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



The ***simple tree*** that fits to the original matrix.

An Idea of Distance-Based Phylogeny

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



An Idea of Distance-Based Phylogeny

Exercise: Apply this recursive approach to distance matrix below.

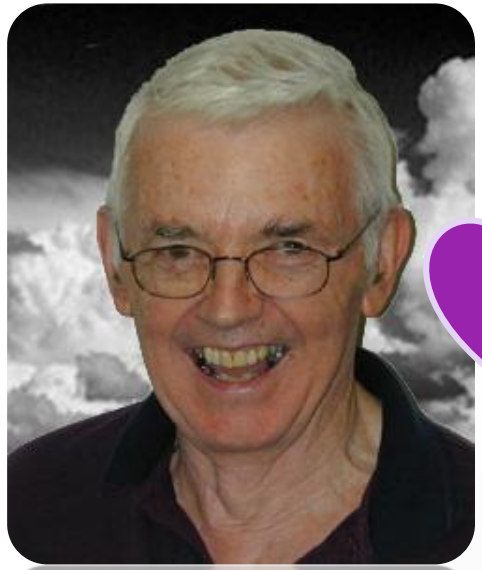
	i	j	k	l
i	0	13	21	22
j	13	0	12	13
k	21	12	0	13
l	22	13	13	0

Distance Matrix Methods

Luidgi L. Cavalli-Sforza

Anthony W.F. Edwards

1967



Walter M. Fitch

Emanuel Margoliash

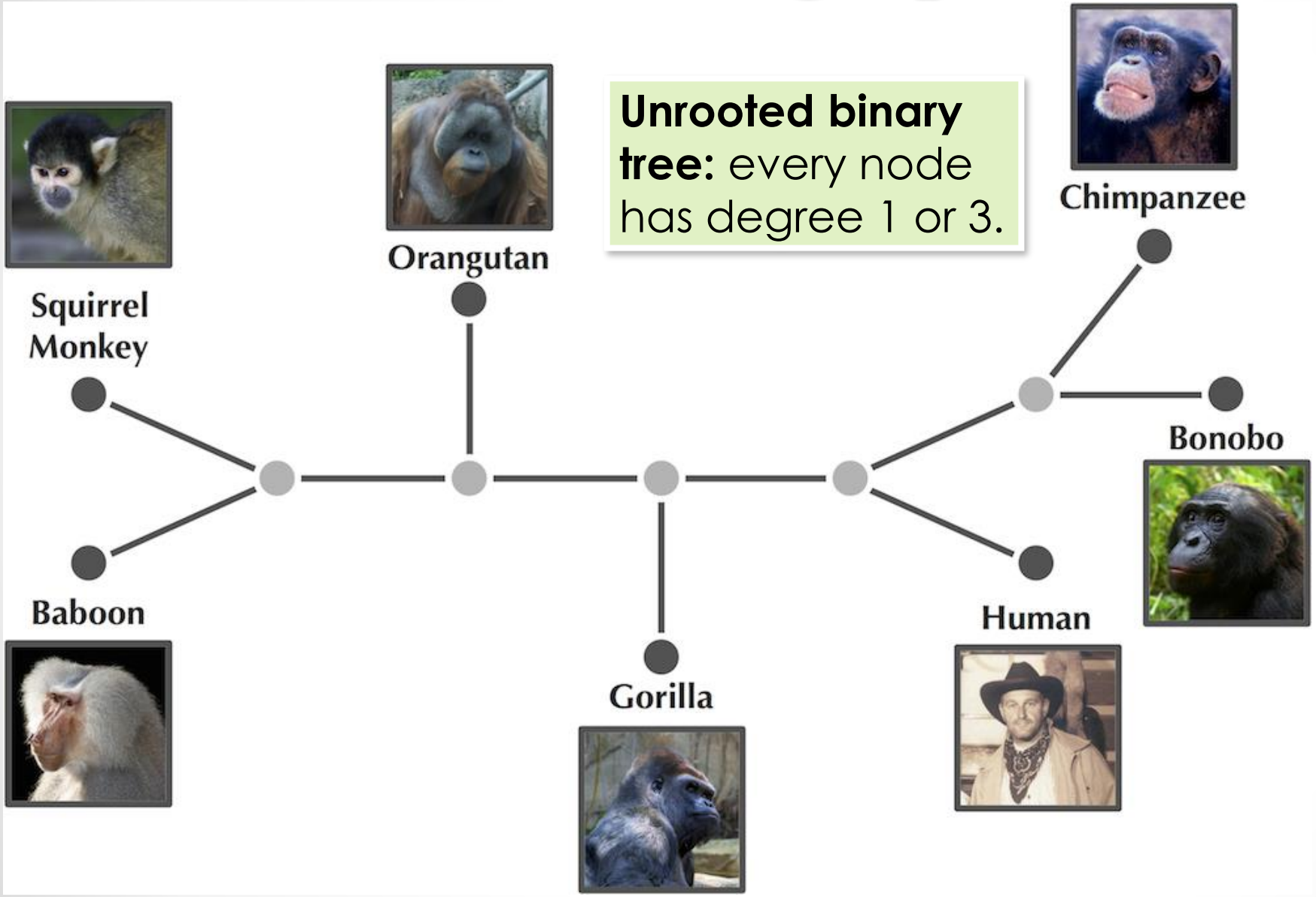
Outline

- Transforming Distance Matrices into Evolutionary Trees
- Toward an Algorithm for Distance-Based Phylogeny Construction
- Additive Phylogeny
- **Ultrametric Evolutionary Trees (UPGMA reconstruction)**
- The Neibour-Joining Algorithm
- Using Least-Squares to construct Distance-Based Phylogenies

Modeling Speciations

Researches often assume that all internal nodes correspond to **speciations**, where one species splits into two.

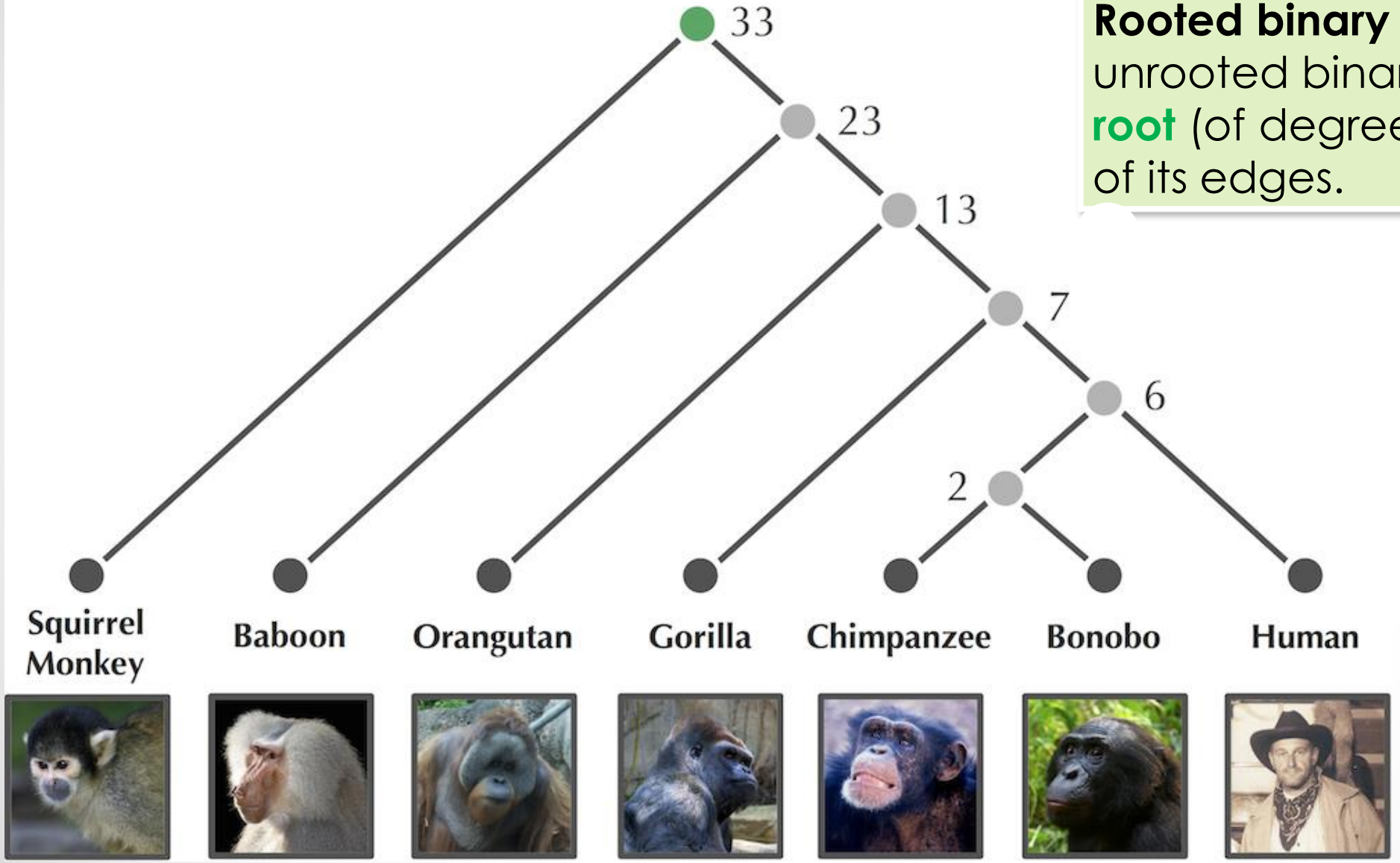
Modeling Speciations



In computer science, a **binary tree** is a tree data structure in which each node has at most two children, which are referred to as the left child and the right child.

We need to place limits on the internal nodes of the tree: every internal node needs to have degree 3.
Progressing from the root to a leaf, every time we encounter an internal node, the tree splits into two pieces.

Modeling Speciations

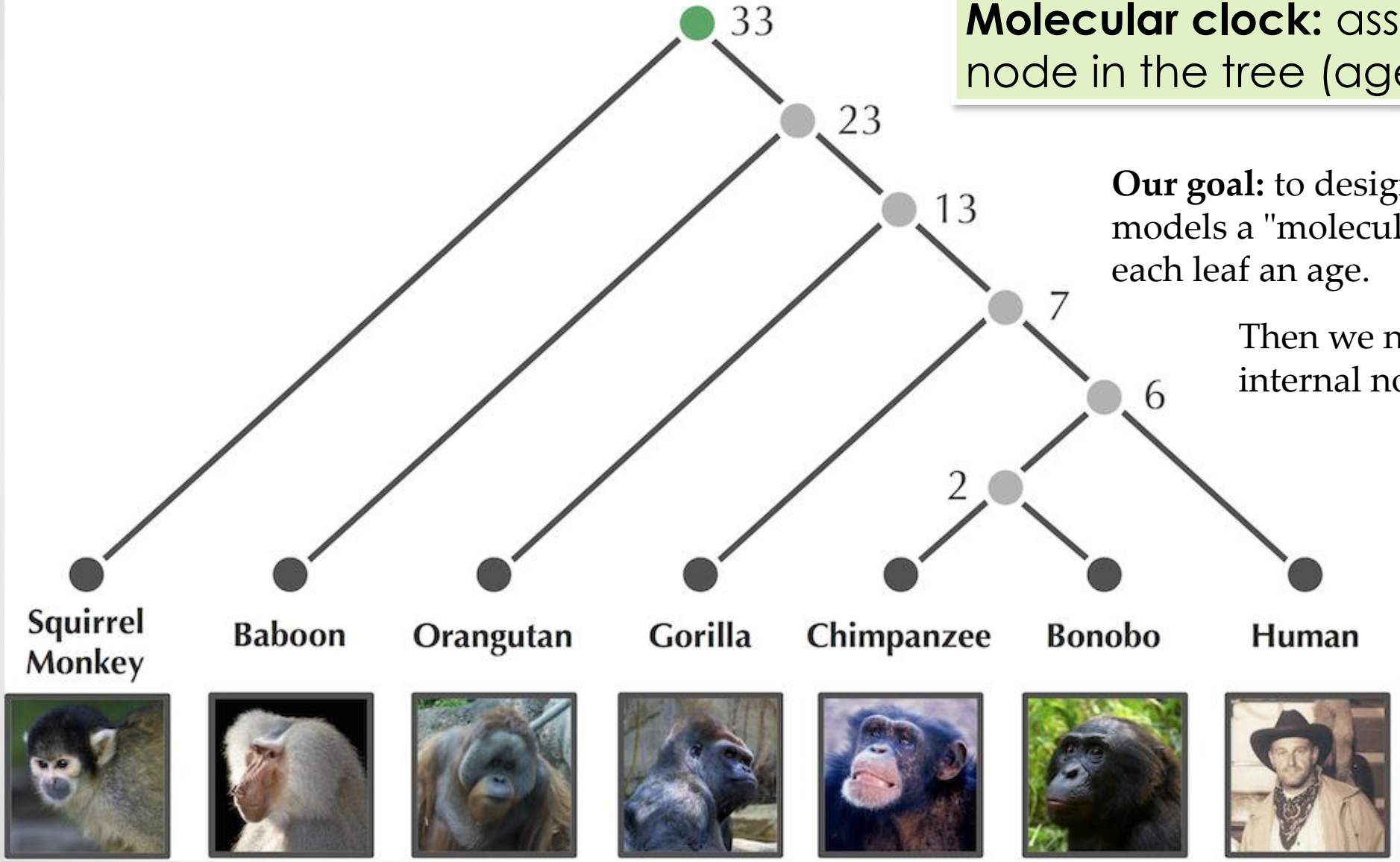


Rooted binary tree: an unrooted binary with a **root** (of degree 2) on 1 of its edges.

Placing a root on the squirrel monkey's limb results in a rooted binary tree.

The number at each node corresponds to the number of million years ago that the divergence at this node occurred.

Ultrametric Trees



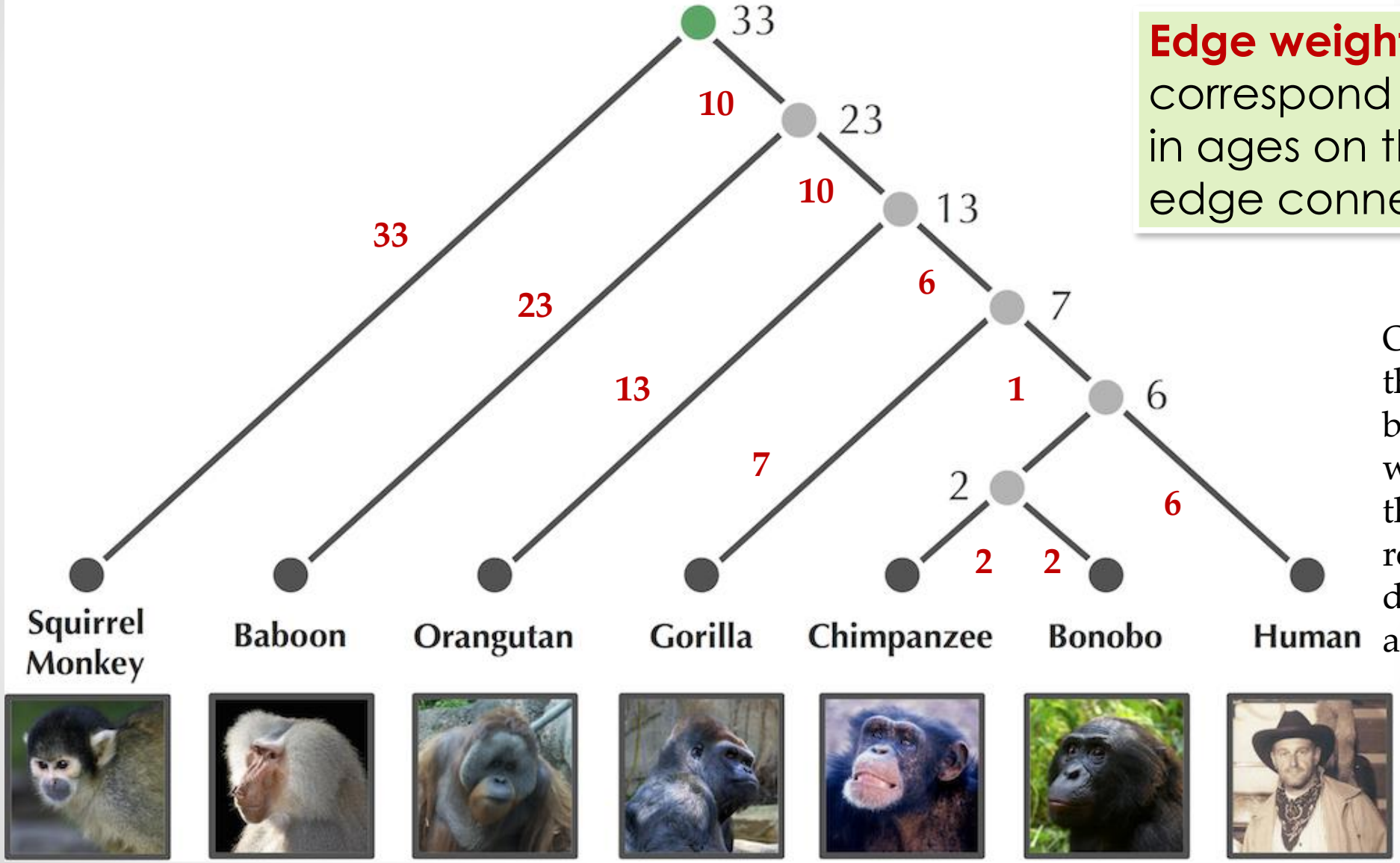
Molecular clock: assigns **ages** to each node in the tree (age of leaves=0)

Our goal: to design a heuristic that models a "molecular clock" that assigns each leaf an age.

Then we need to assign ages to the internal nodes.

The age of an internal node corresponds to how long ago the speciation event represented at that internal node occurred.

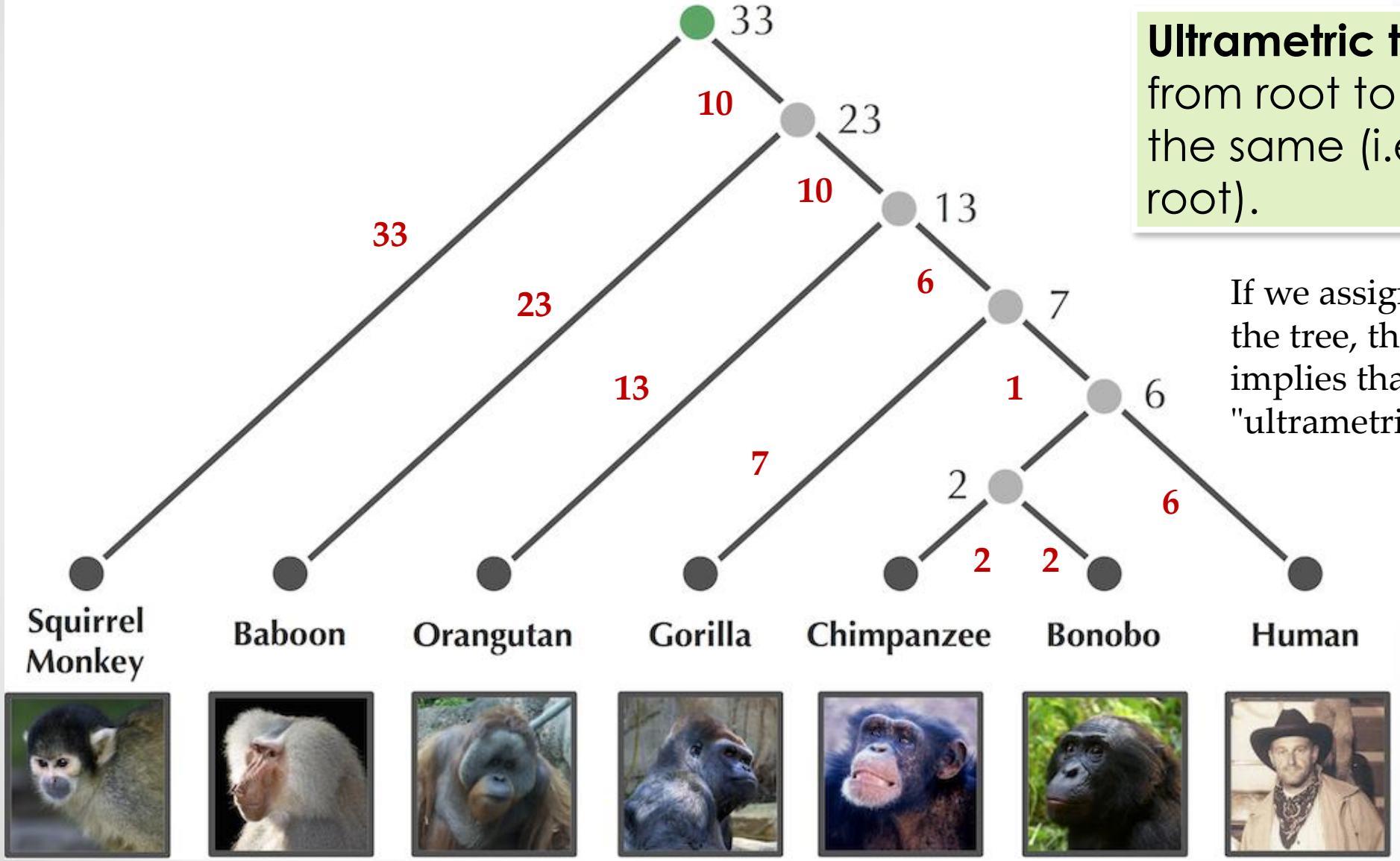
Ultrametric Trees



Edge weights: correspond to difference in ages on the nodes the edge connects.

Once we assign ages to the nodes of a rooted binary tree, we get the weights of the edges of this tree (here shown in red) simply by taking the difference between the ages.

Ultrametric Trees

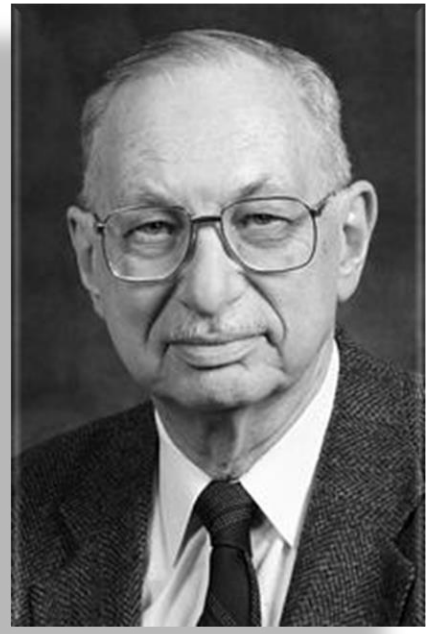


Ultrametric tree: distance from root to any leaf is the same (i.e., age of root).

If we assign ages to the nodes of the tree, then it automatically implies that the tree is "ultrametric".

UPGMA: A Clustering Heuristic

- **UPGMA** (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean) is a simple agglomerative (bottom-up) hierarchical clustering method.



Robert A. Sokal, biostatistician

&

Charles D. Michener, entomologist

*Sokal R and Michener C (1958). "A statistical method for evaluating systematic relationships". University of Kansas Science Bulletin. **38**: 1409–1438.*

UPGMA: A Clustering Heuristic

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 constructs an evolutionary tree by clustering the species from the distance matrix into larger and larger clusters, beginning with single element clusters.

UPGMA: A Clustering Heuristic

2. Find the two closest clusters C1 and C2 according to the average distance

$$D_{avg}(C_1, C_2) = \sum_{i \text{ in } C1, j \text{ in } C2} D_{i,j} / |C_1| \cdot |C_2|$$

Where $|C|$ denotes the number of elements in C.

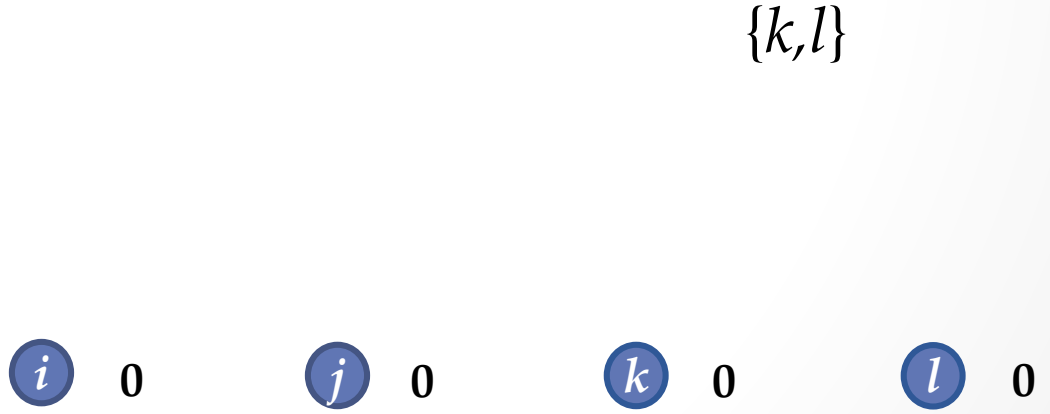
	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0

At each step it looks for the two closest clusters, according to the average distance among all pairs of elements taken from the two clusters. At this stage of the algorithm, we're dealing with single element clusters, so if we're looking for the closest clusters, that's just the smallest element of the distance matrix, which corresponds to k and l.

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

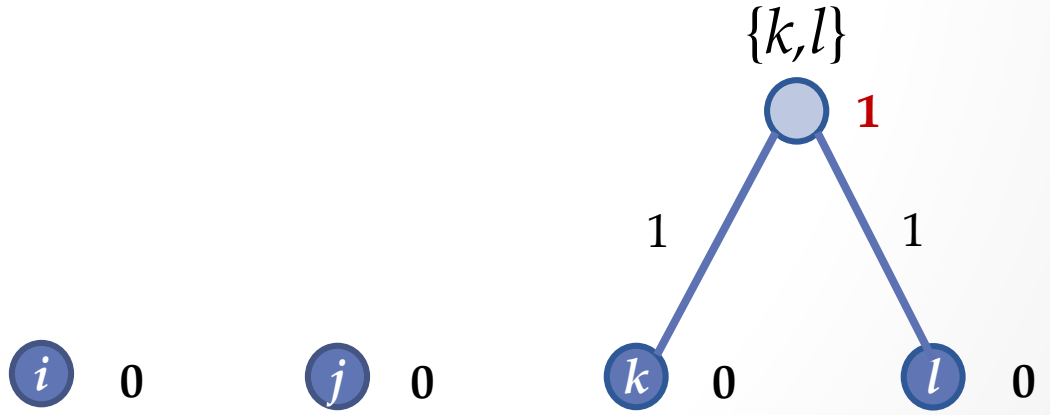


Once the two closest clusters C1 and C2 are found, we can merge them into a single cluster, C. Here we put k and l into a cluster together because they're the closest.

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 $D_{avg}(C_1, C_2)/2$.

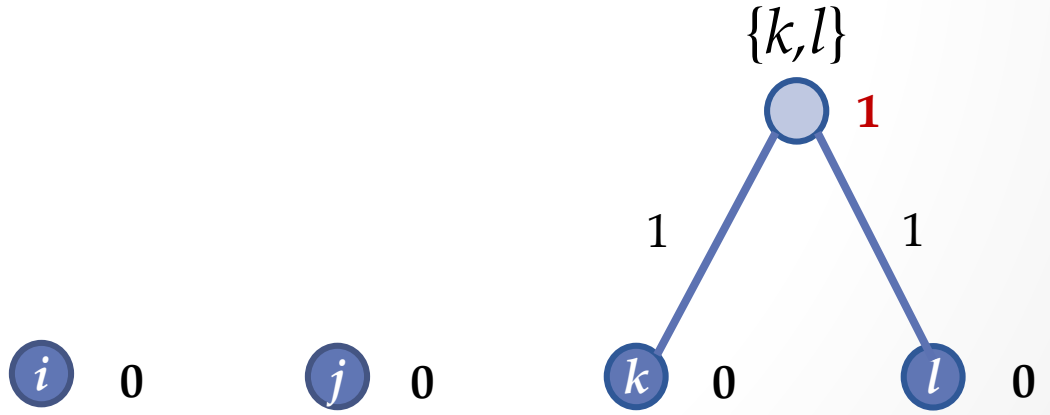
	i	j	k	l
i	0	3	4	3
j	3	0	4	5
k	4	4	0	**2**
l	3	5	**2**	0



UPGMA: A Clustering Heuristic

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

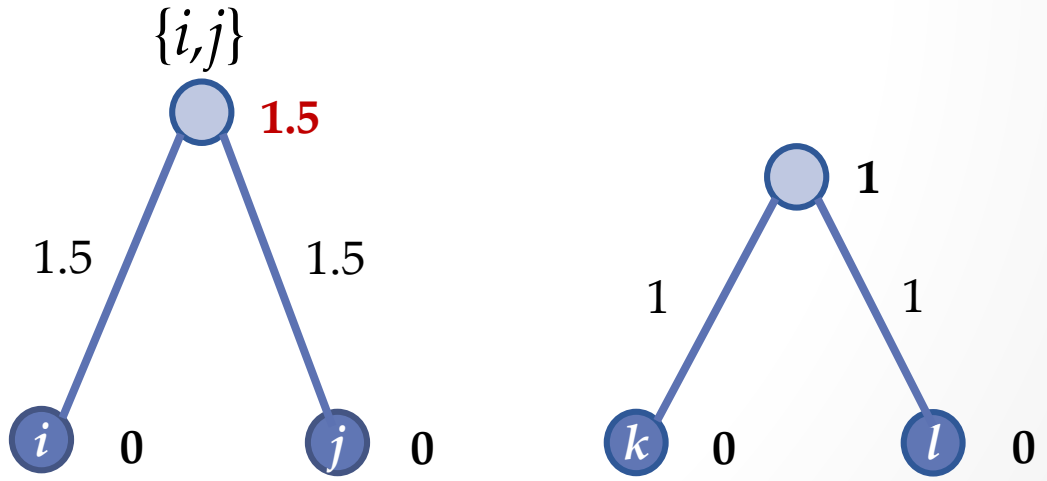
	i	j	$\{k,l\}$
i	0	3	3.5
j	3	0	4.5
$\{k,l\}$	3.5	4.5	0



UPGMA: A Clustering Heuristic

6. Iterate until a single cluster contains all species.

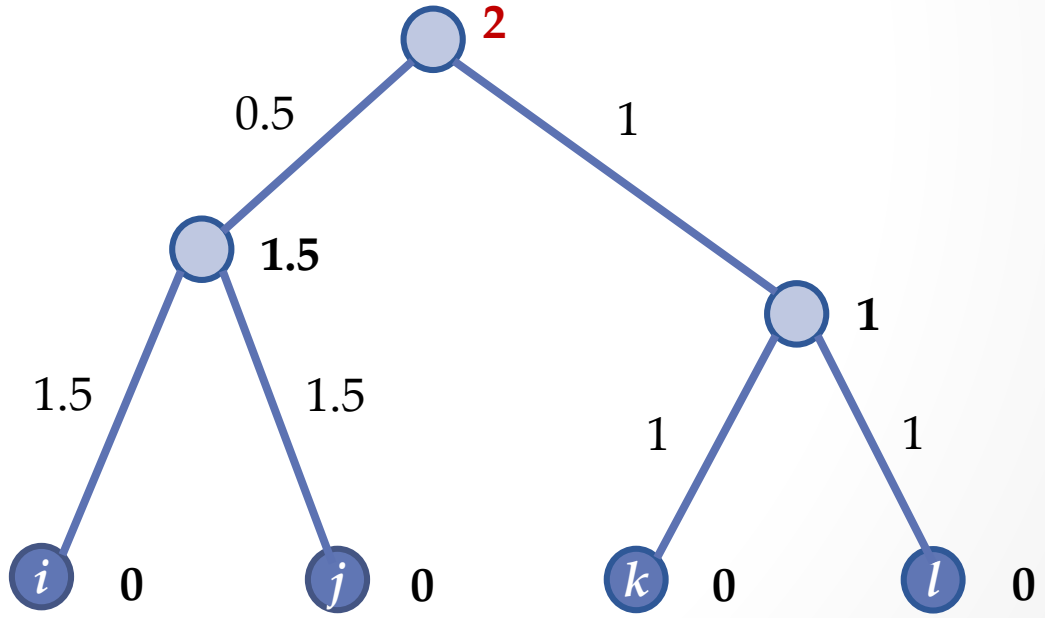
	i	j	$\{k,l\}$
i	0	**3**	3.5
j	**3**	0	4.5
$\{k,l\}$	3.5	4.5	0



UPGMA: A Clustering Heuristic

6. Iterate until a single cluster contains all species.

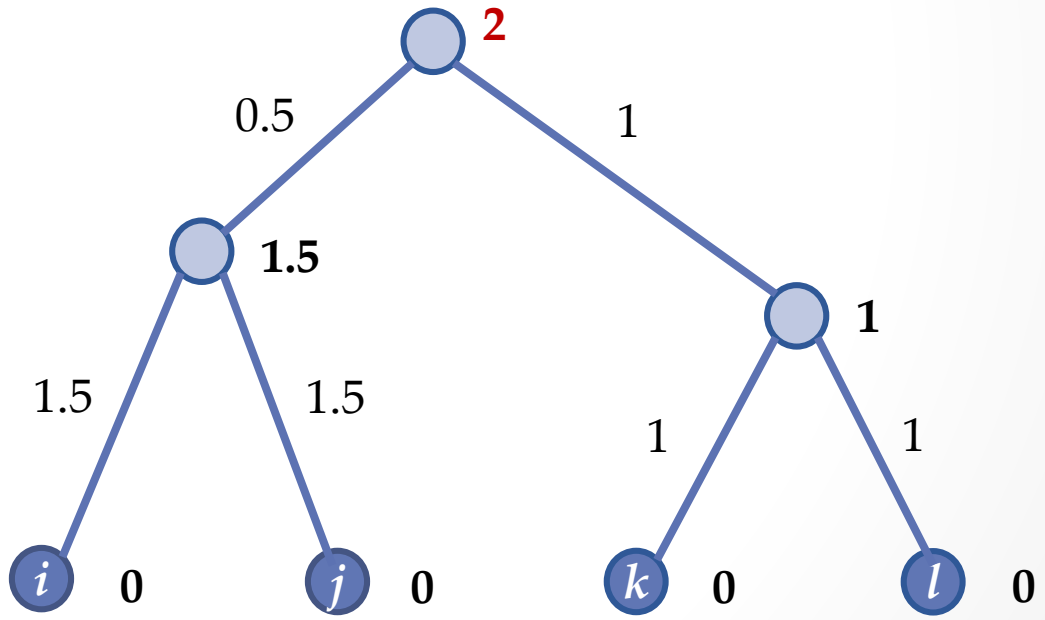
	$\{i,j\}$	$\{k,l\}$
$\{i,j\}$	0	4
$\{k,l\}$	4	0



UPGMA: A Clustering Heuristic

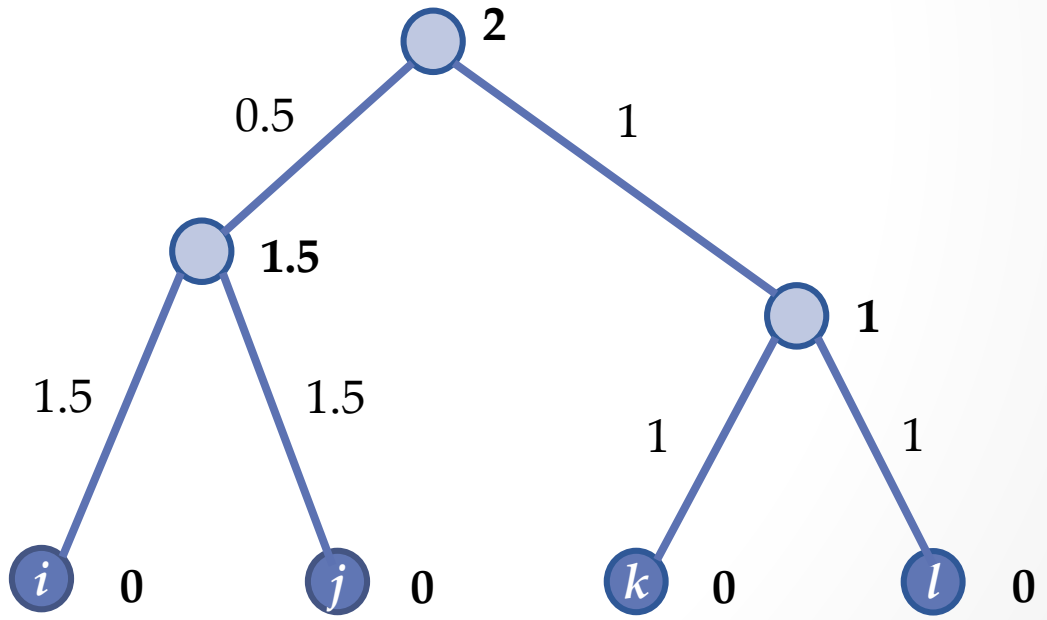
6. Iterate until a single cluster contains all species.

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



UPGMA: A Clustering Heuristic

6. Iterate until a single cluster contains all species.



UPGMA: A Clustering Heuristic

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

2. Find the two closest clusters C1 and C2 according to the average distance
   $D_{avg}(C_1, C_2) = \sum_{i \text{ in } C1, j \text{ in } C2} D_{i,j} / |C_1| \cdot |C_2|$
   Where $|C|$ denotes the number of elements in C.

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

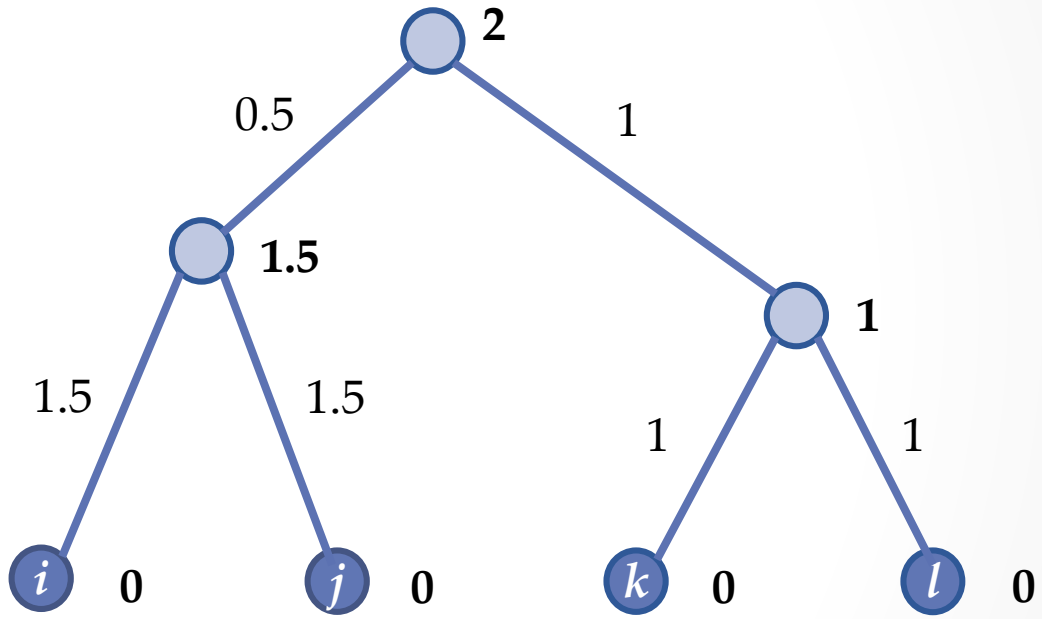
4. Form a new node for C and connect to C_1 and C_2 by an edge. Set age of C as $D_{avg}(C_1, C_2)/2$.

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

6. Iterate until a single cluster contains all species.

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

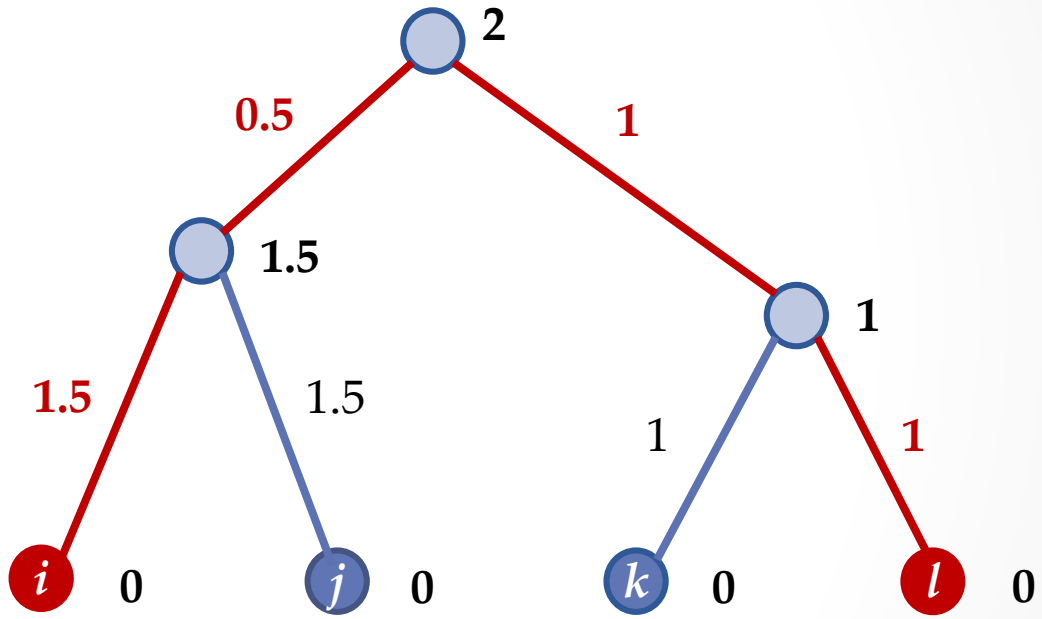
	i	*j*	*k*	*l*
i	0	3	4	3
j	3	0	4	5
k	4	4	0	2
l	3	5	2	0



The goal of UPGMA is not to fit a distance matrix, but to provide a reliable method that always can construct an ultrametric tree, regardless of what the input data is.

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 tree produced by UPGMA can't possibly fit this matrix, because the matrix is non-additive.
For example, the distance from i to l is 3 in the distance matrix, but 4 according to the UPGMA tree.

Exercise Break

- Reconstruct phylogenetic tree from the following distance matrix using UPGMA approach:

OTUS	A	B	(CD)	E
B	6			
(CD)	29	31		
E	24	26	32	
F	30	28	15	30

- What would be the topology of this tree? (Parentheses indicate the order of grouping):

a) ((EAB(CDF)); b) (EA(B))(CD)F); c) ((AB)(CD)F)E; d) (E(AB))((CD)F)

- If in the previous exercise dF(CD) – d(CD)=9, what is the distace of the both taxons C and D to their most recent common ancestor?

a) 3,0 ; b) 1,5 ; c) 1,0 ; d) 2,0

Exercise Break

- Below is a distance matrix D. If C1 is the cluster containing i and k, and C2 is the cluster containing j and l, compute D(C1, C2).

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