CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

Outline for April 19

- Finish: Gaussian Mixture Models
- Hierarchical clustering algorithms
- Dimensionality reduction

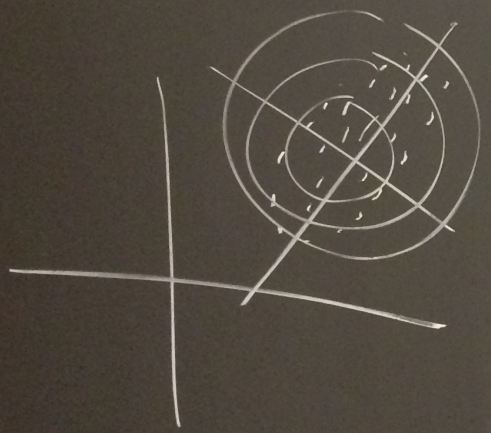
- Submitted proposal Wed: feedback and repo
- Submitted proposal Thurs: repo (will send feedback today)
- Everyone else: submit today!
- Mon/Wed next week: Midterm 2 review (midterm in lab)
- Fri next week: Guest lecture by Prof. Matt Zucker
- Office hours today: 1-3pm

Outline for April 19

- Finish: Gaussian Mixture Models
- Hierarchical clustering algorithms
- Dimensionality reduction

GMM / EM Algorithm

- π_k = cluster "size", prob of cluster k
- $\vec{\mu}_k$ = mean of cluster k
- σ_k^2 = variance for cluster k (Σ if $p > 1$)



initialize

$$\pi_k = \frac{1}{K}$$

$\vec{\mu}_k$ = random data point

$\vec{\sigma}_k$ = sample variance for all points closest to $\vec{\mu}_k$

ster 'size', prob of cluster k

of cluster k

ance for cluster k

(Σ if $p > 1$)

initialize

random data point

ample variance for all points closest to $\vec{\mu}_k$

E-step soft-assignment

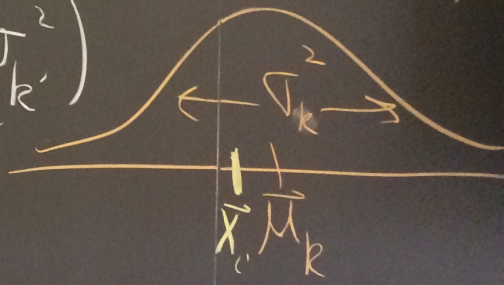
w_{ik} = prob $\vec{x}_i$ came from cluster k

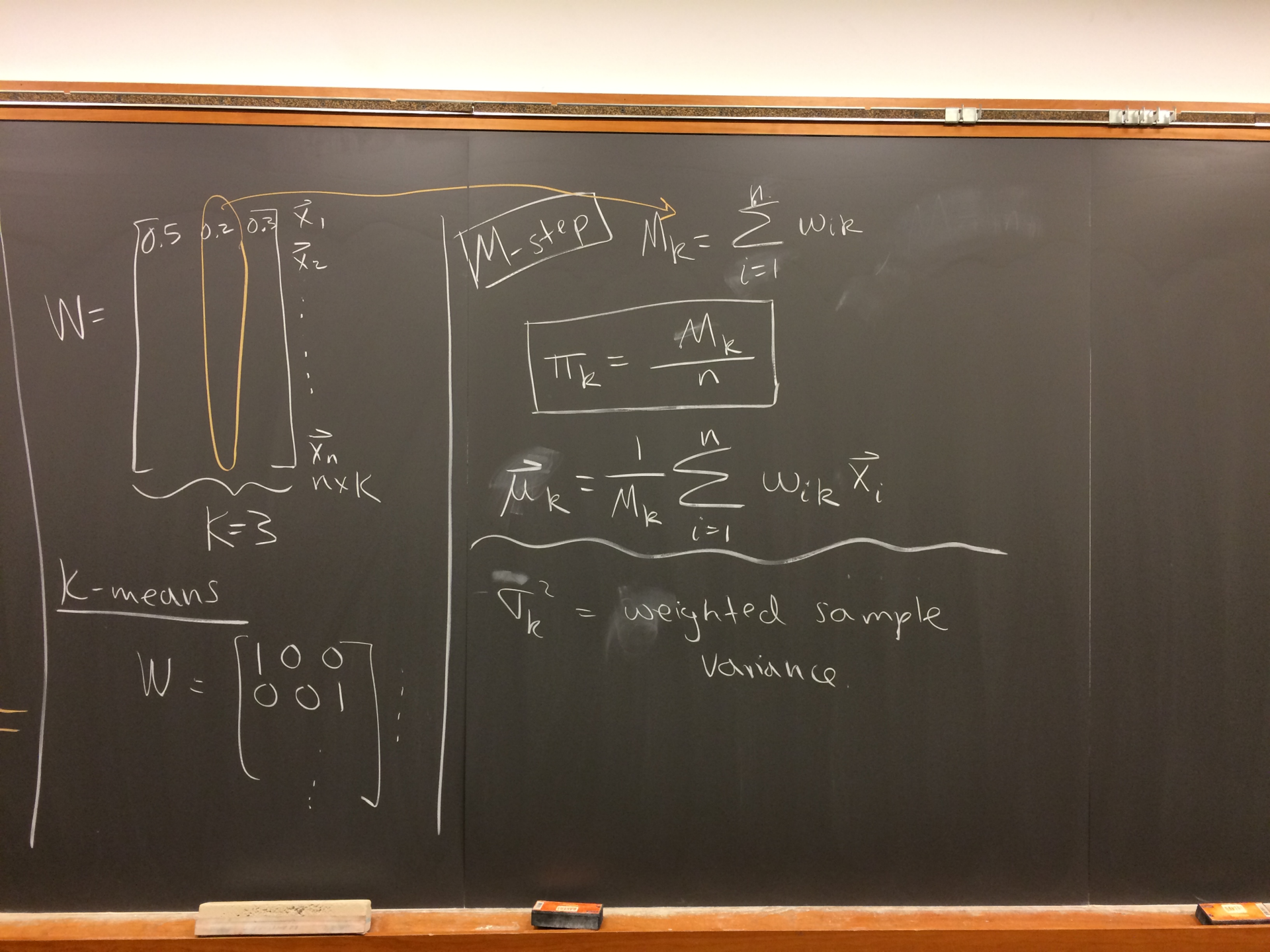
$$w_{ik} = p(k \mid \vec{x}_i) = \frac{p(k)\,p(\vec{x}_i \mid k)}{p(\vec{x}_i)}$$

$$= \frac{\pi_k N(\vec{x}_i ; \vec{\mu}_k, \sigma_k^2)}{\sum_{k'} \pi_{k'} N(\vec{x}_i ; \vec{\mu}_{k'}, \sigma_{k'}^2)}$$

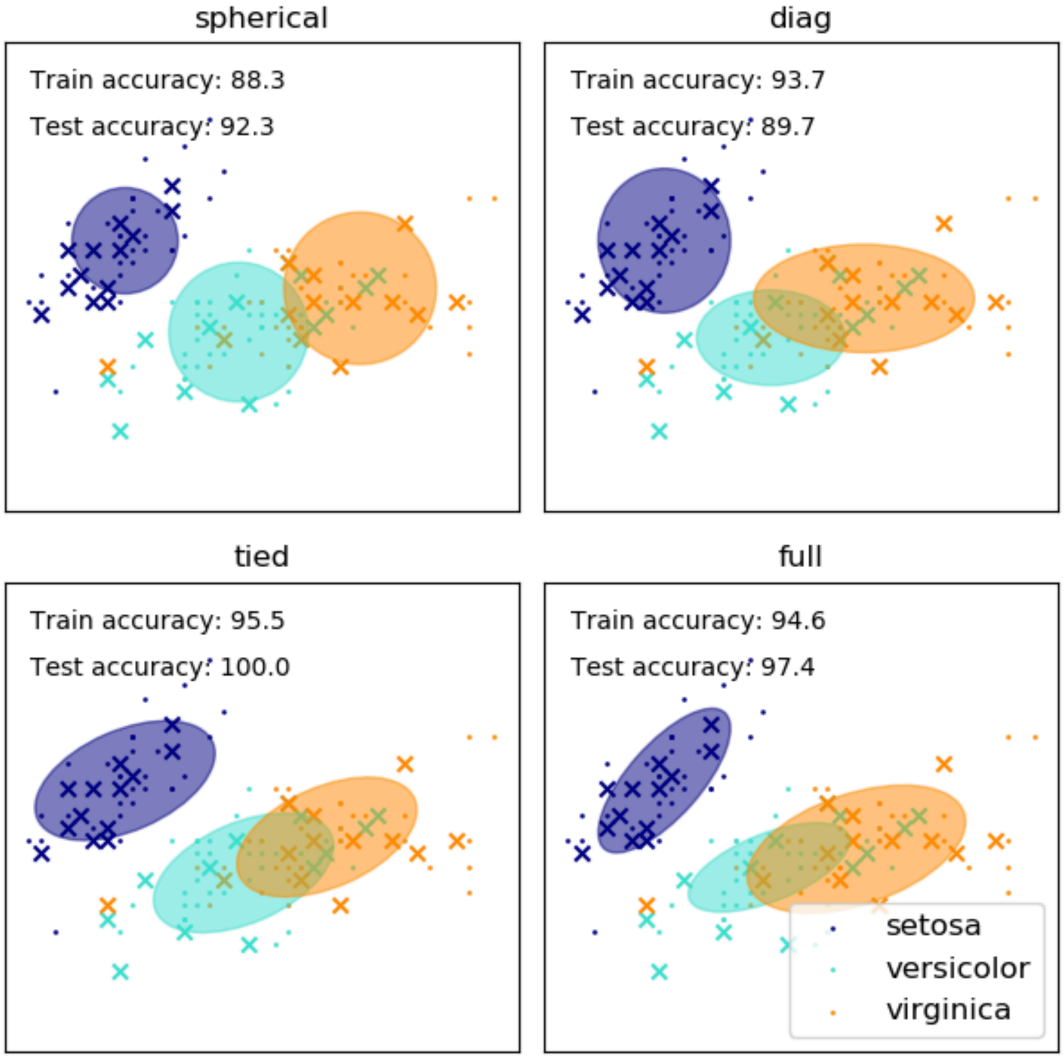
normalize

normal distribution (Gaussian)




W=
0.5
0.2
0.3
$\vec{x}_1$
$\vec{x}_2$
$\vec{x}_n$
n x K
K=3
K-means
$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$
M-step
$N_k = \sum_{i=1}^{n} w_{ik}$
$\pi_k = \frac{N_k}{n}$
$\vec{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} w_{ik} \vec{x}_i$
σ_k^2 = weighted sample variance

Example of different covariance constraints on the Iris flower data



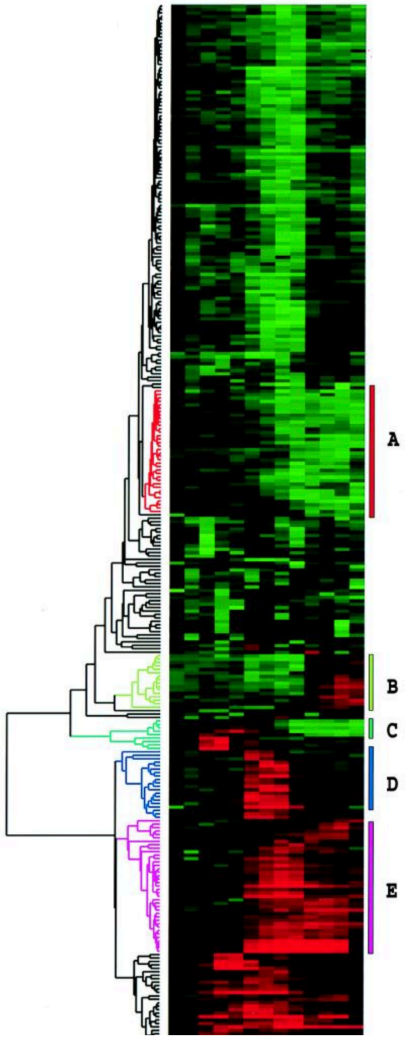
https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html#sphx-glr-auto-examples-mixture-plot-gmm-covariances-py

Outline for April 19

- Finish: Gaussian Mixture Models
- Hierarchical clustering algorithms
- Dimensionality reduction

Applications of clustering in ML

- Cluster genes with similar expression patterns



Cluster analysis and display of genome-wide expression patterns

Michael B. Eisen,* Paul T. Spellman,* Patrick O. Brown,† and David Botstein*‡

Applications of clustering in ML

- Group face images of the same person together

- Unsupervised since we don’t have any labeled faces, and don’t know how many people there are

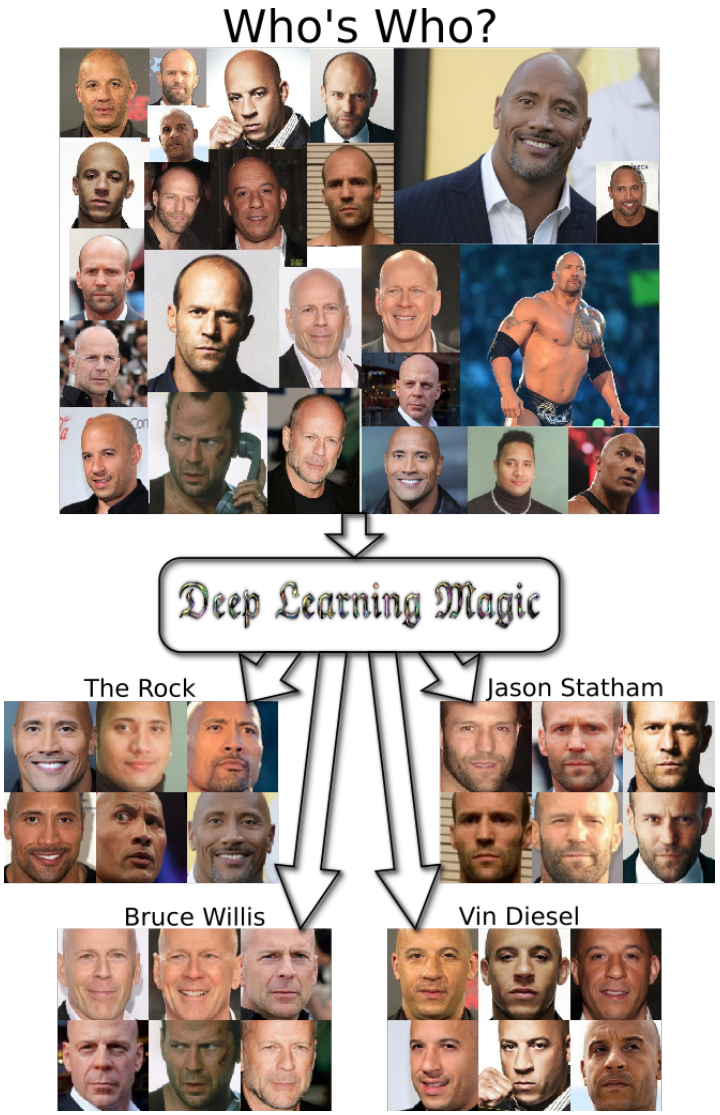


Image: http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html

Applications of clustering in ML

- Image segmentation: cluster similar regions of an image

Image: Jessica Wu

Applications of clustering in ML

- Clustering in social graphs

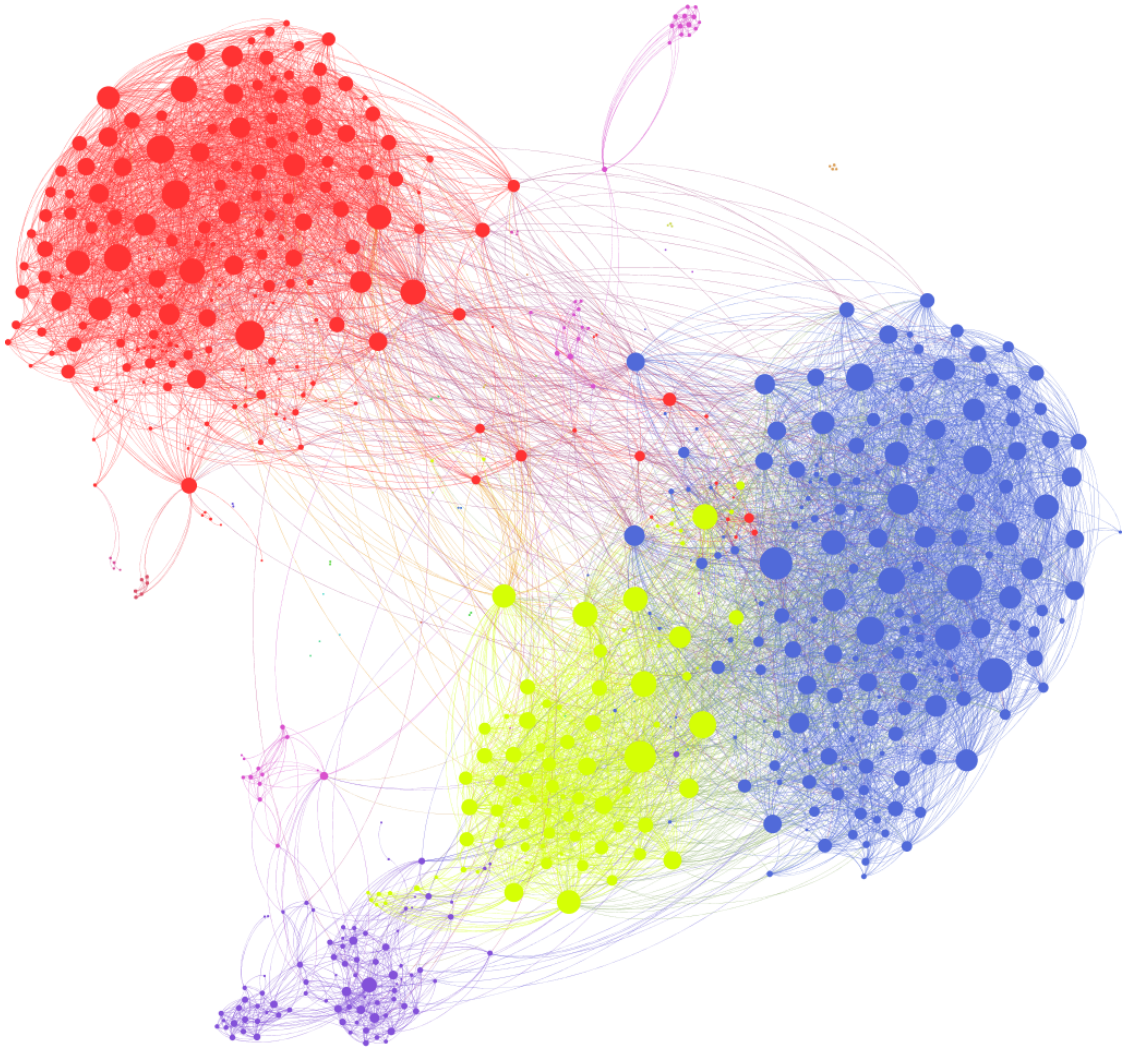
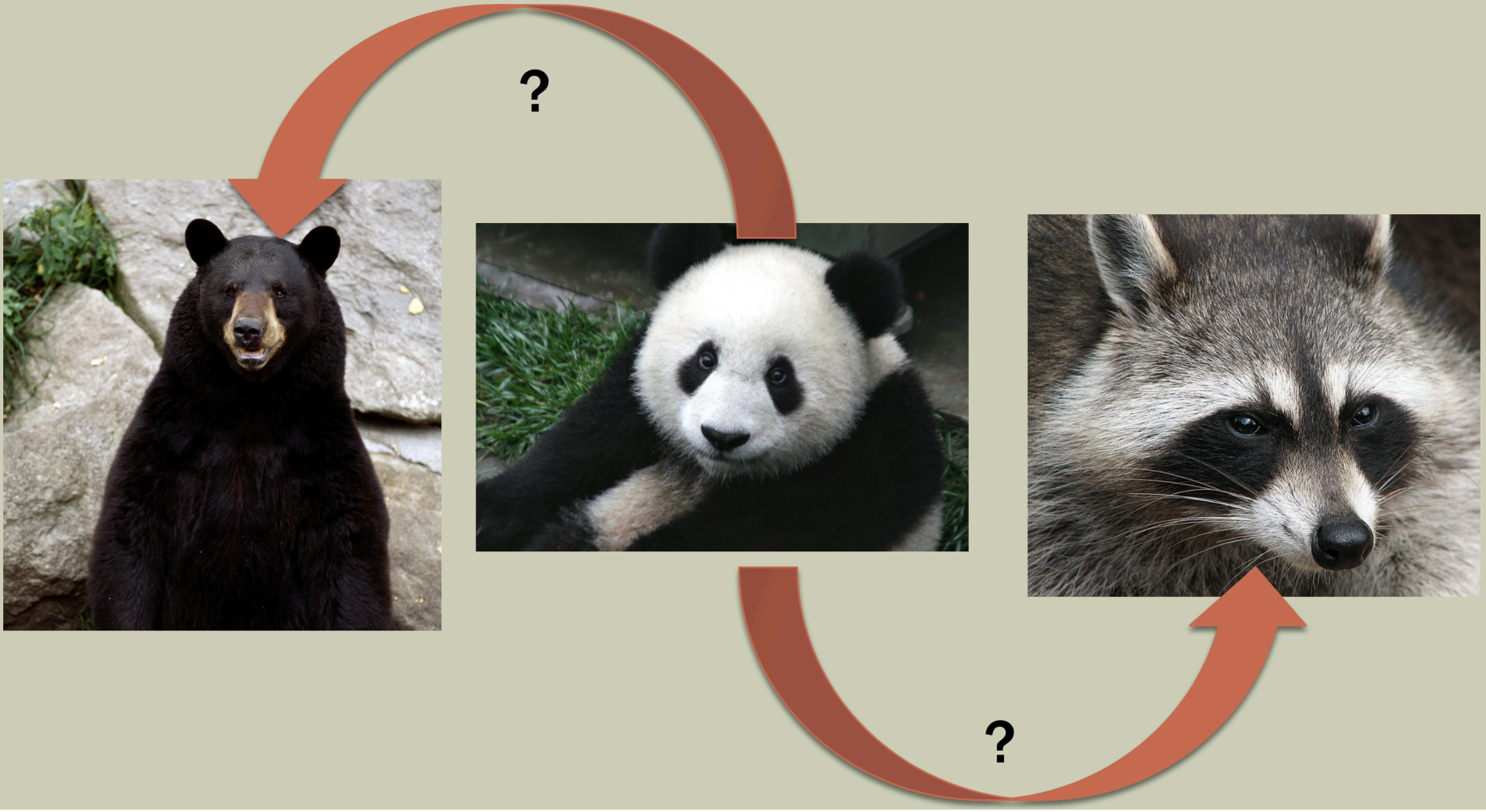


Image: https://griffsgraphs.wordpress.com/2012/07/02/a-facebook-network/

Two main types of clustering

- Flat/Partitional:
  - K-means
  - Gaussian mixture models
- Hierarchical:
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and Neighbor Joining

Are pandas more closely related to bears or raccoons?



Credit: Ameet Soni

UPGMA and Neighbor Joining

- Start with a dissimilarity map between examples (symmetric matrix)
- Say our examples are: 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

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

UNIVERSITY OF Southampton
School of Biological Sciences

A D B F G C E

Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

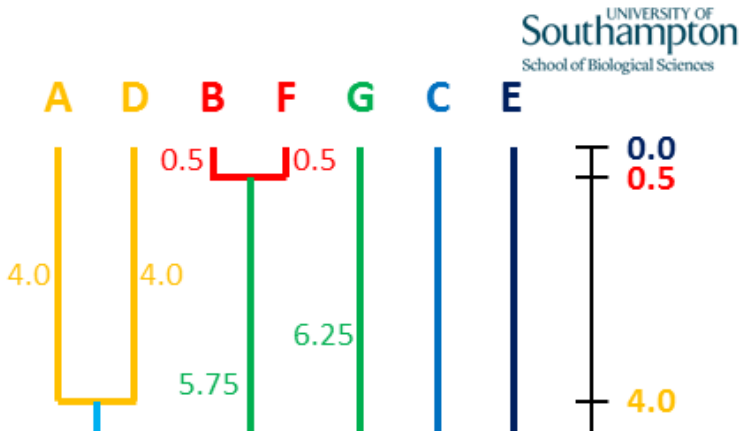


Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

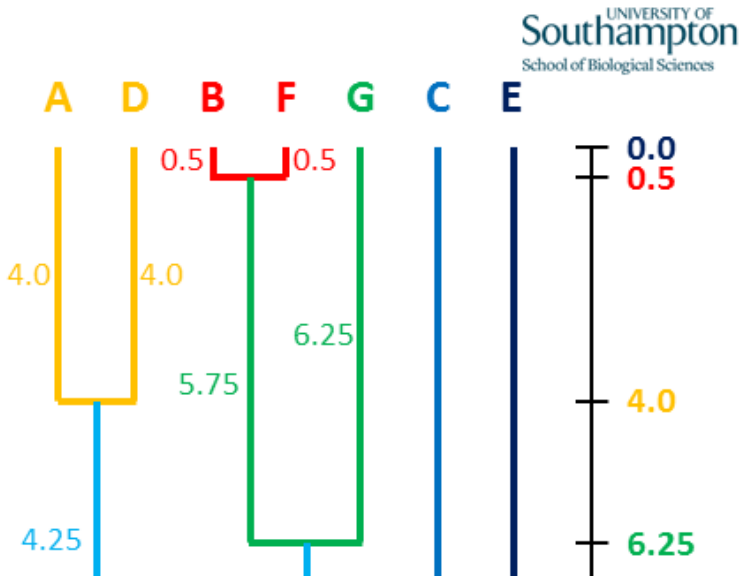


Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

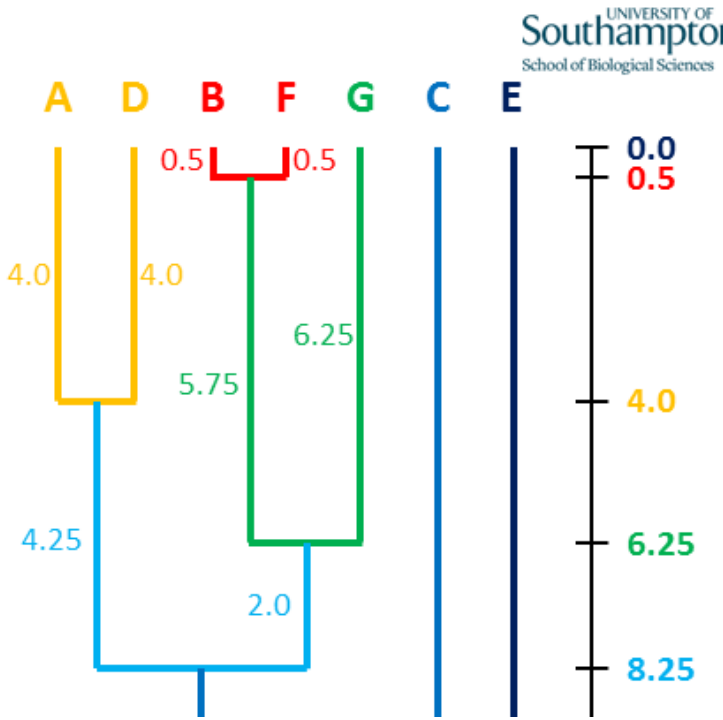


Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**

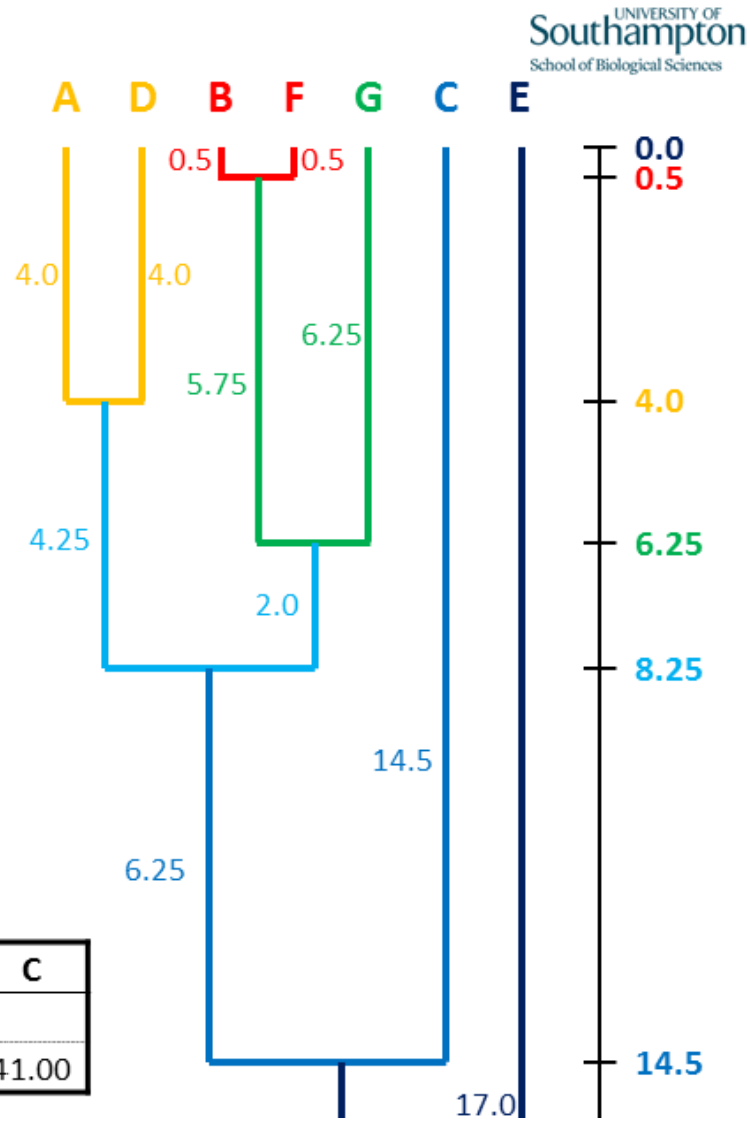


Figure: Dr. Richard Edwards

UPGMA example

	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**

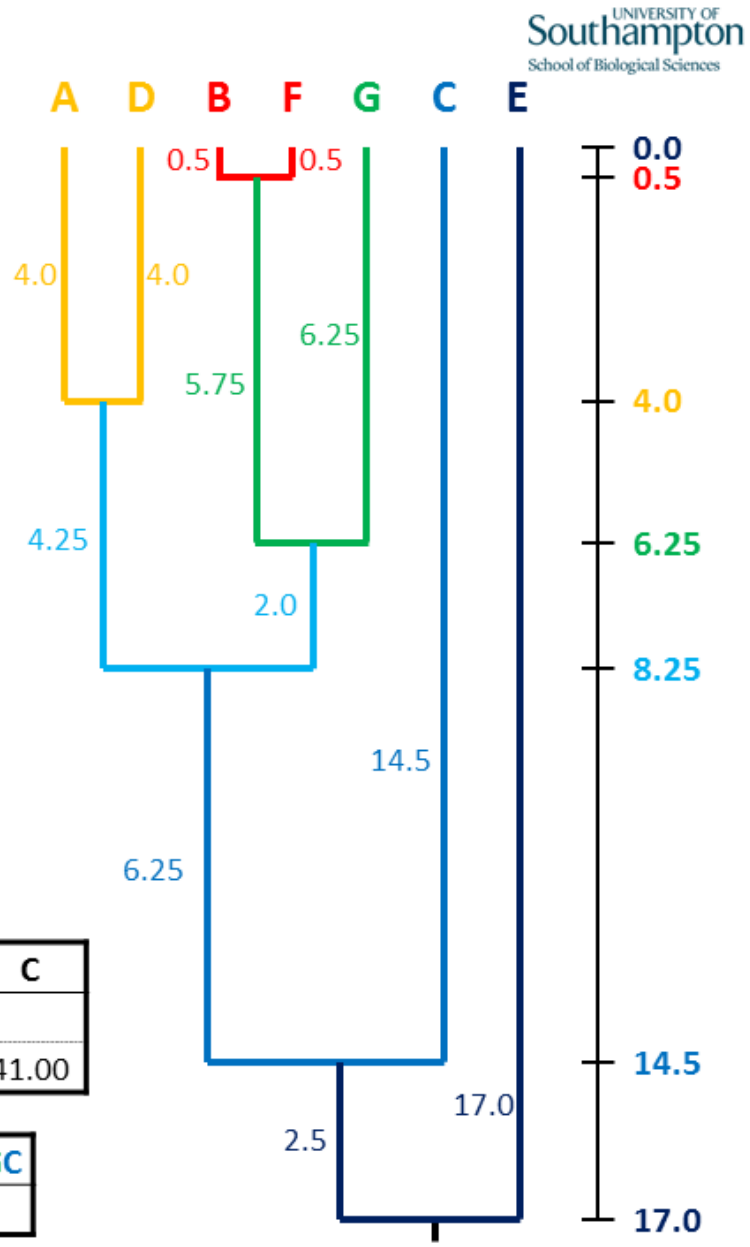
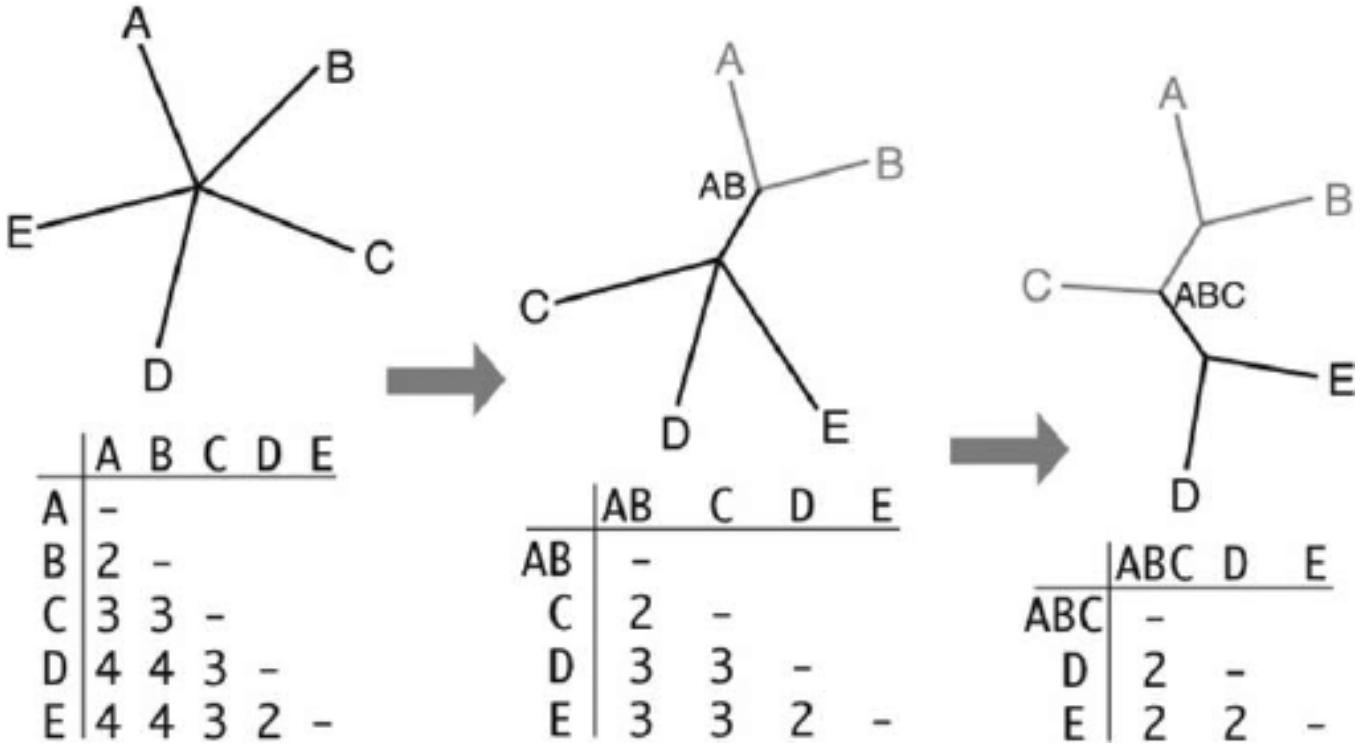


Figure: Dr. Richard Edwards

Neighbor Joining



	A	B	C	D	E
A	-				
B	2	-			
C	3	3	-		
D	4	4	3	-	
E	4	4	3	2	-

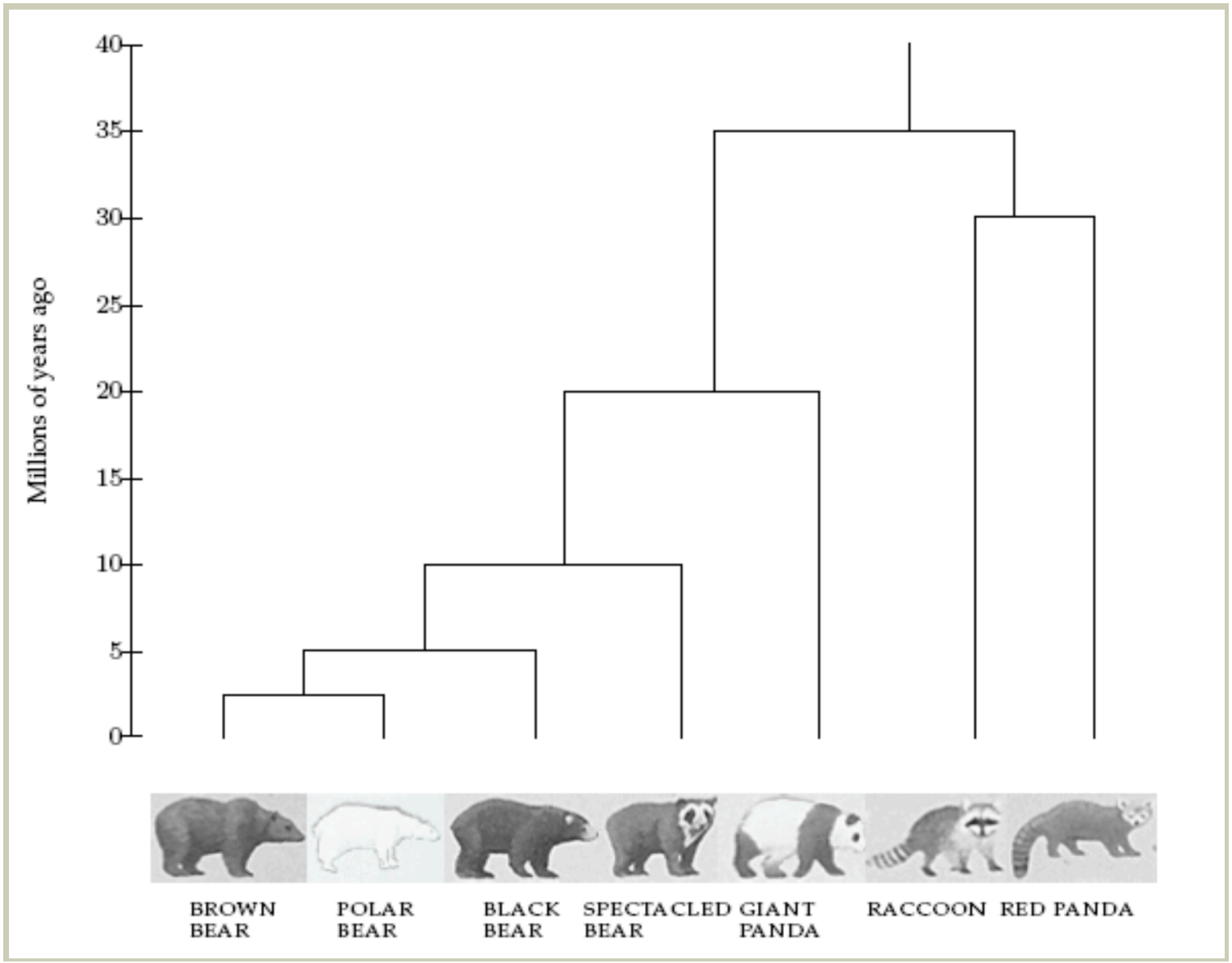
	AB	C	D	E
AB	-			
C	2	-		
D	3	3	-	
E	3	3	2	-

	ABC	D	E
ABC	-		
D	2	-	
E	2	2	-

Image: Michael Dunn

Back to the pandas….

Back to the pandas….



Credit: Ameet Soni

Outline for April 19

- Finish: Gaussian Mixture Models
- Hierarchical clustering algorithms
- Dimensionality reduction

Principal Components Analysis (PCA)

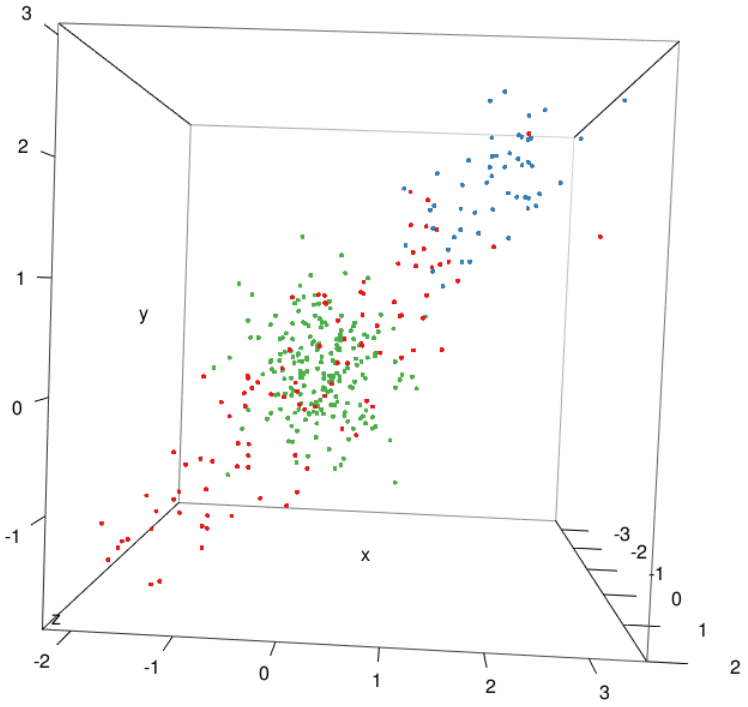
- Transforms *p*-dimensional data so that the new first dimension explains as much of the variation as possible, the new second explains as much of the remaining variation as possible, and so on

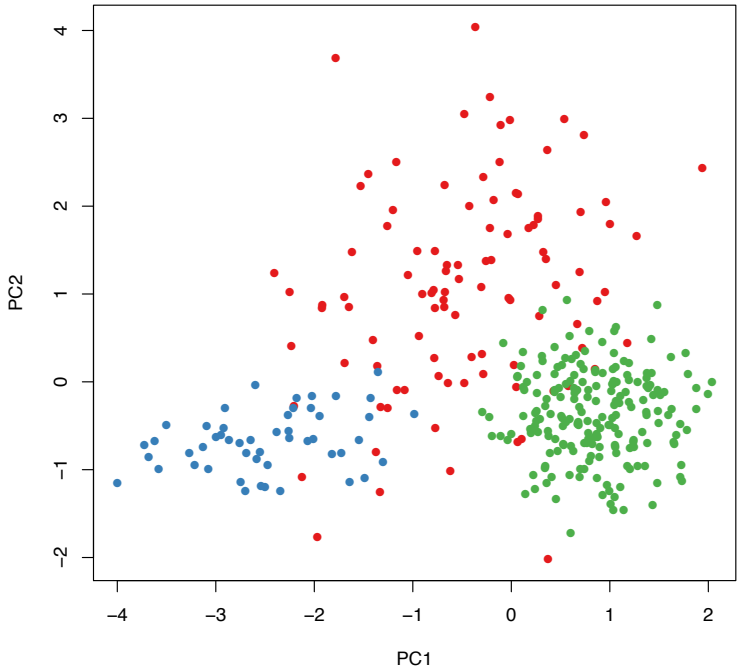
- Typically, we look at the first few dimensions of the transformed data and use as a means of dimensionality reduction

- PCA is a linear transformation

- PCA is often used for:
  - Data visualization
  - Infer qualitative relationships between groups

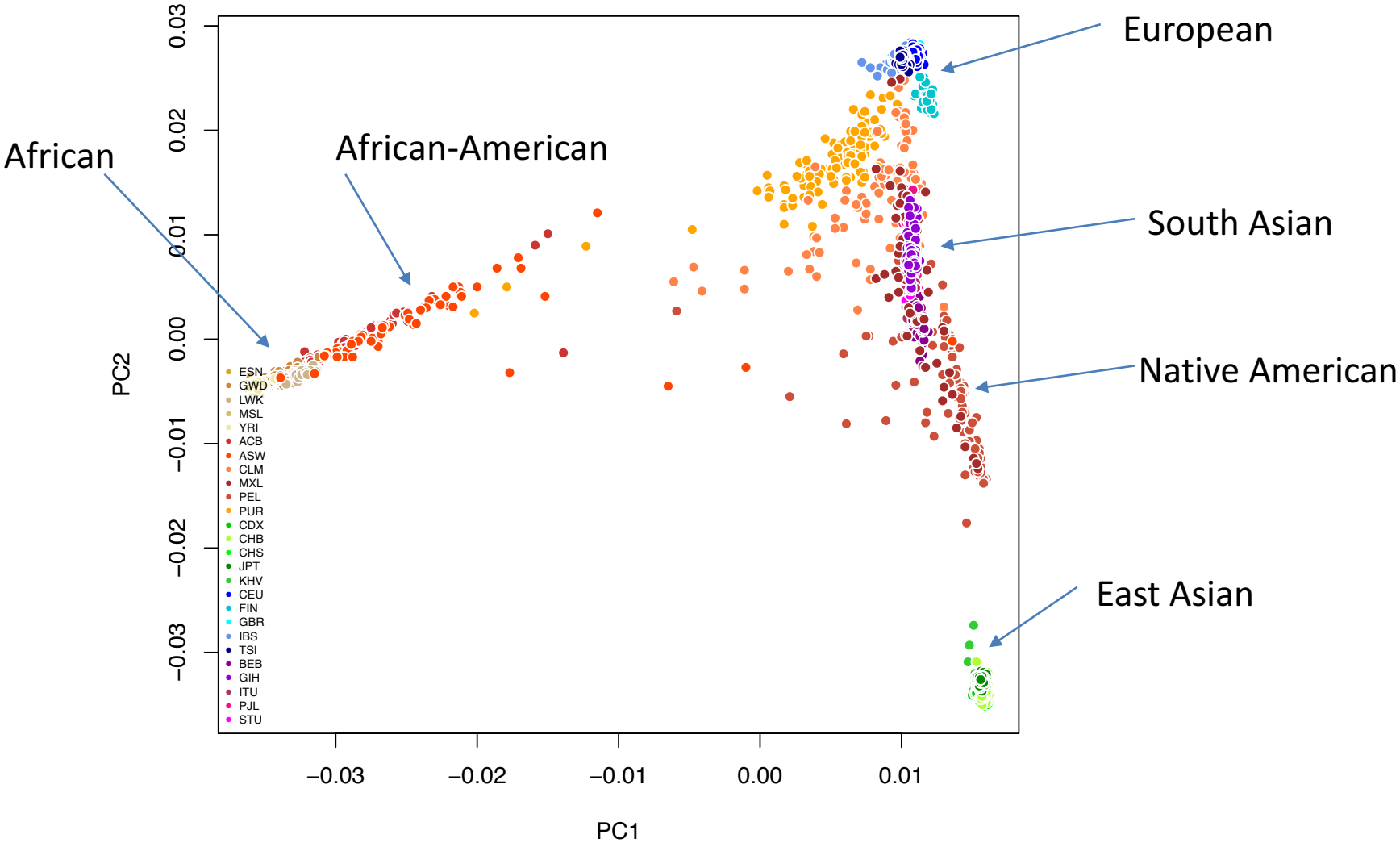
Principal component analysis





Slide: Iain Mathieson

Global population structure

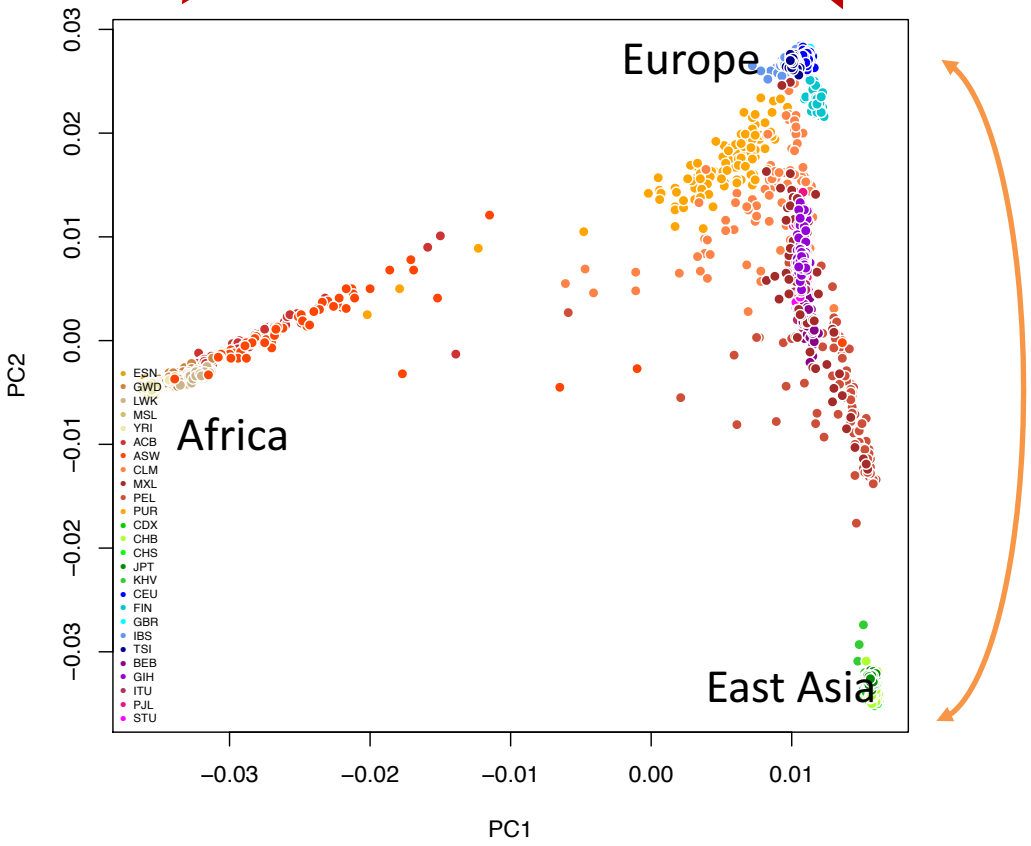


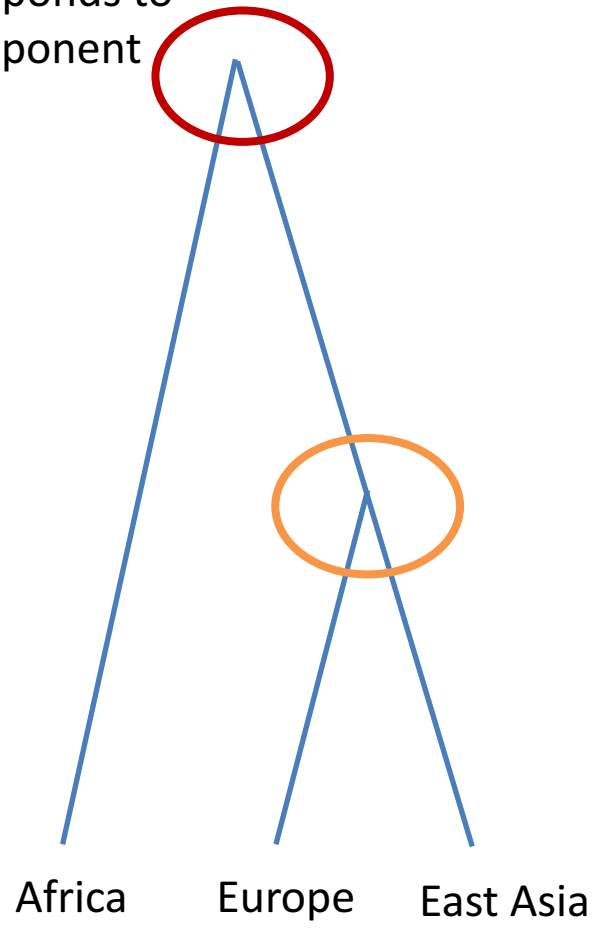
Slide: Iain Mathieson

What causes these patterns?

1. Populations splits separate populations

Oldest split corresponds to first principal component



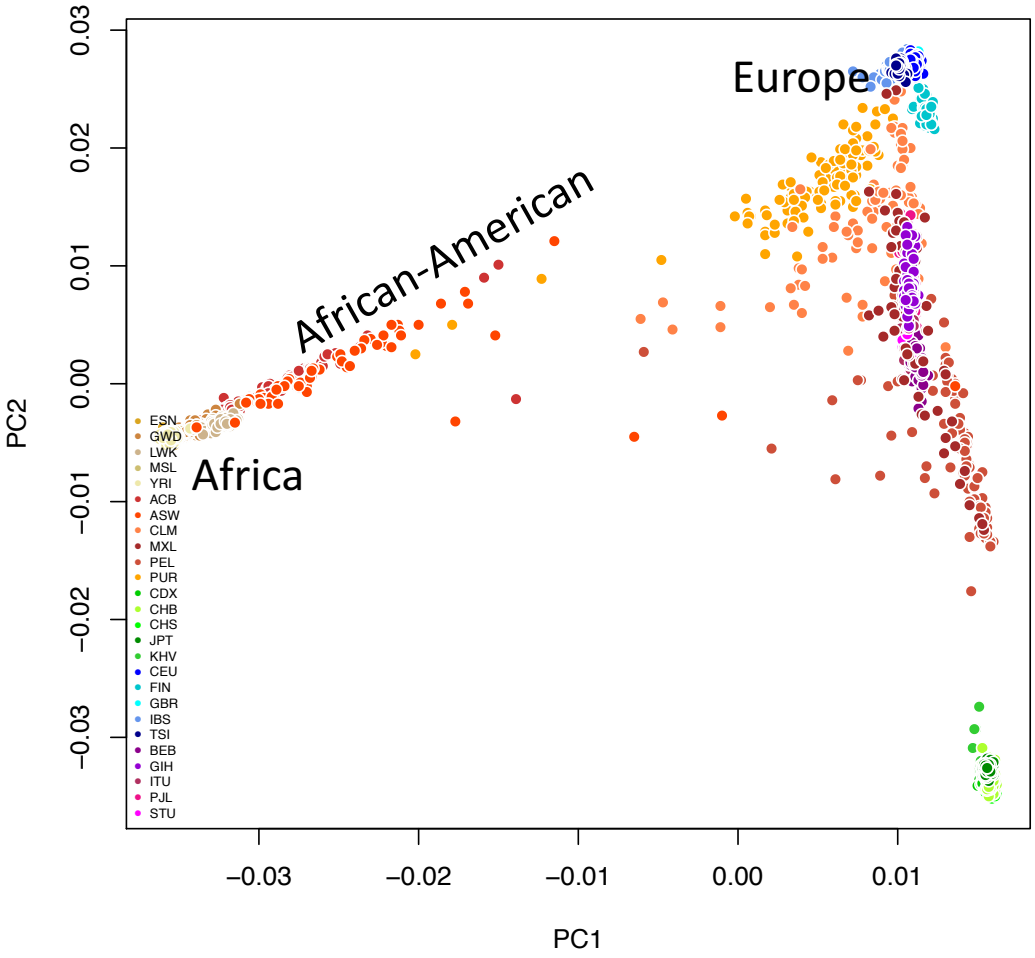


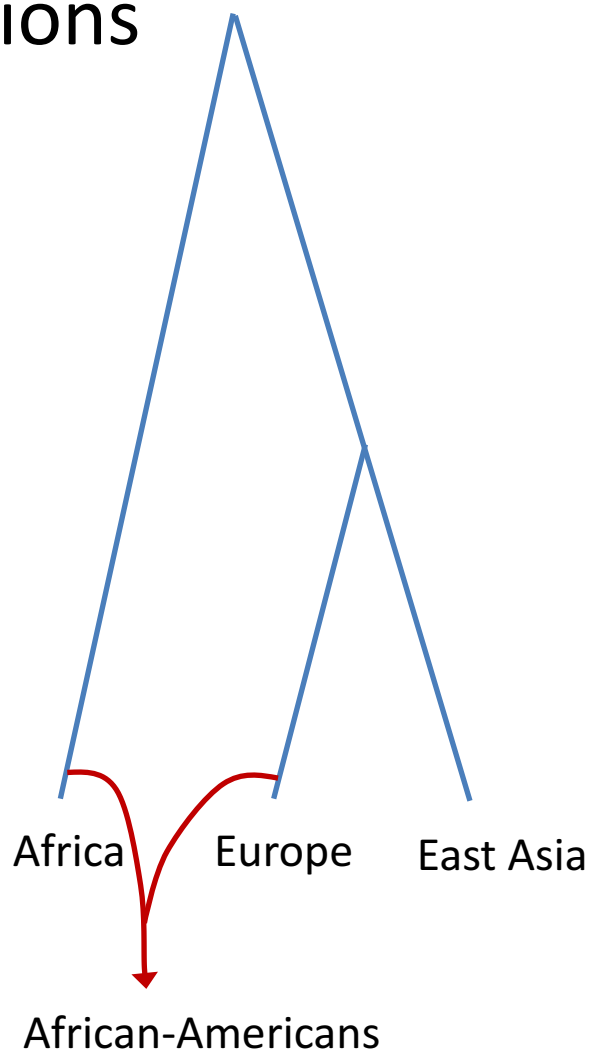
Slide: Iain Mathieson

McVean 2009: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000686

What causes these patterns?

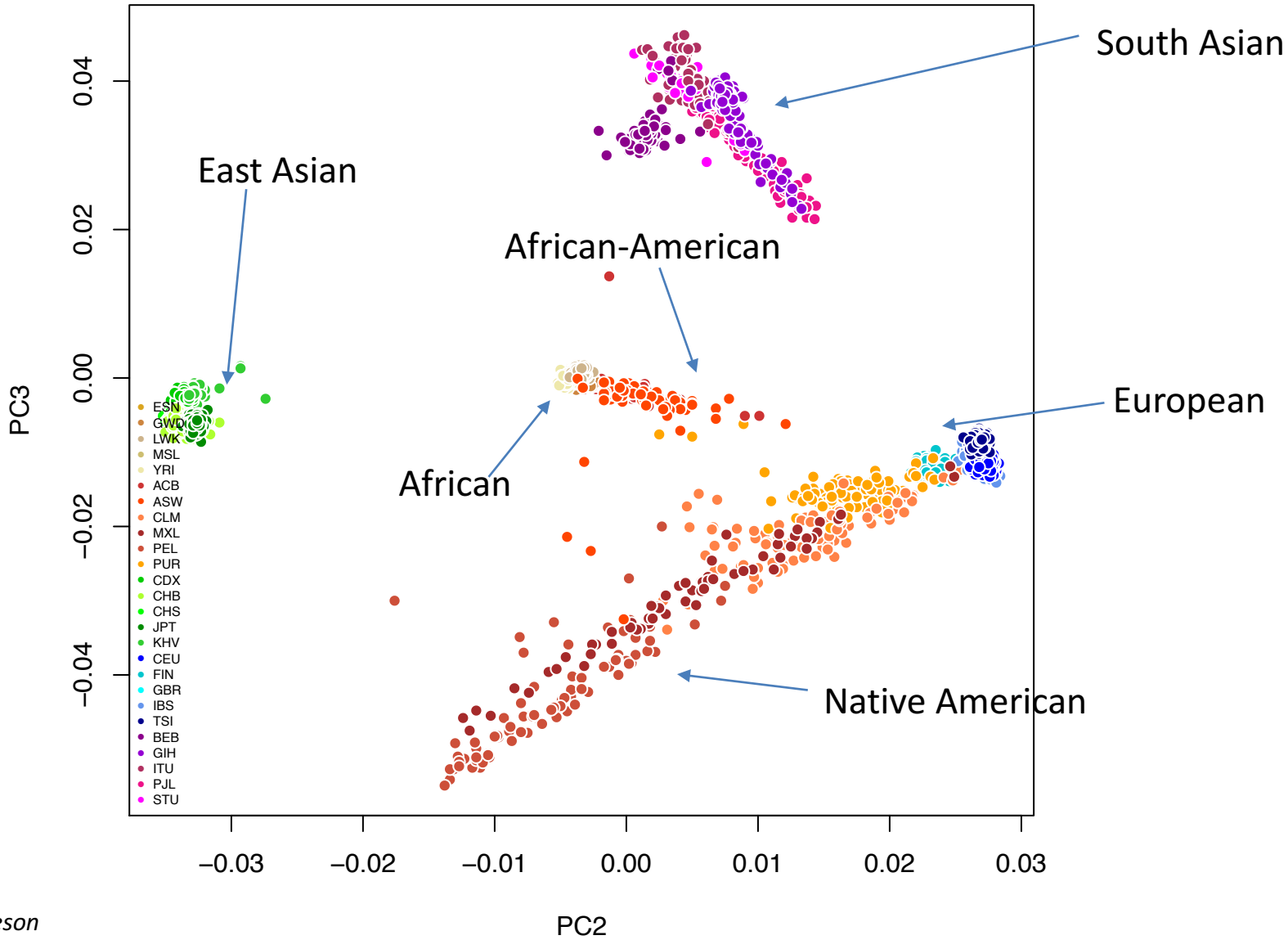
2. Admixture merges populations





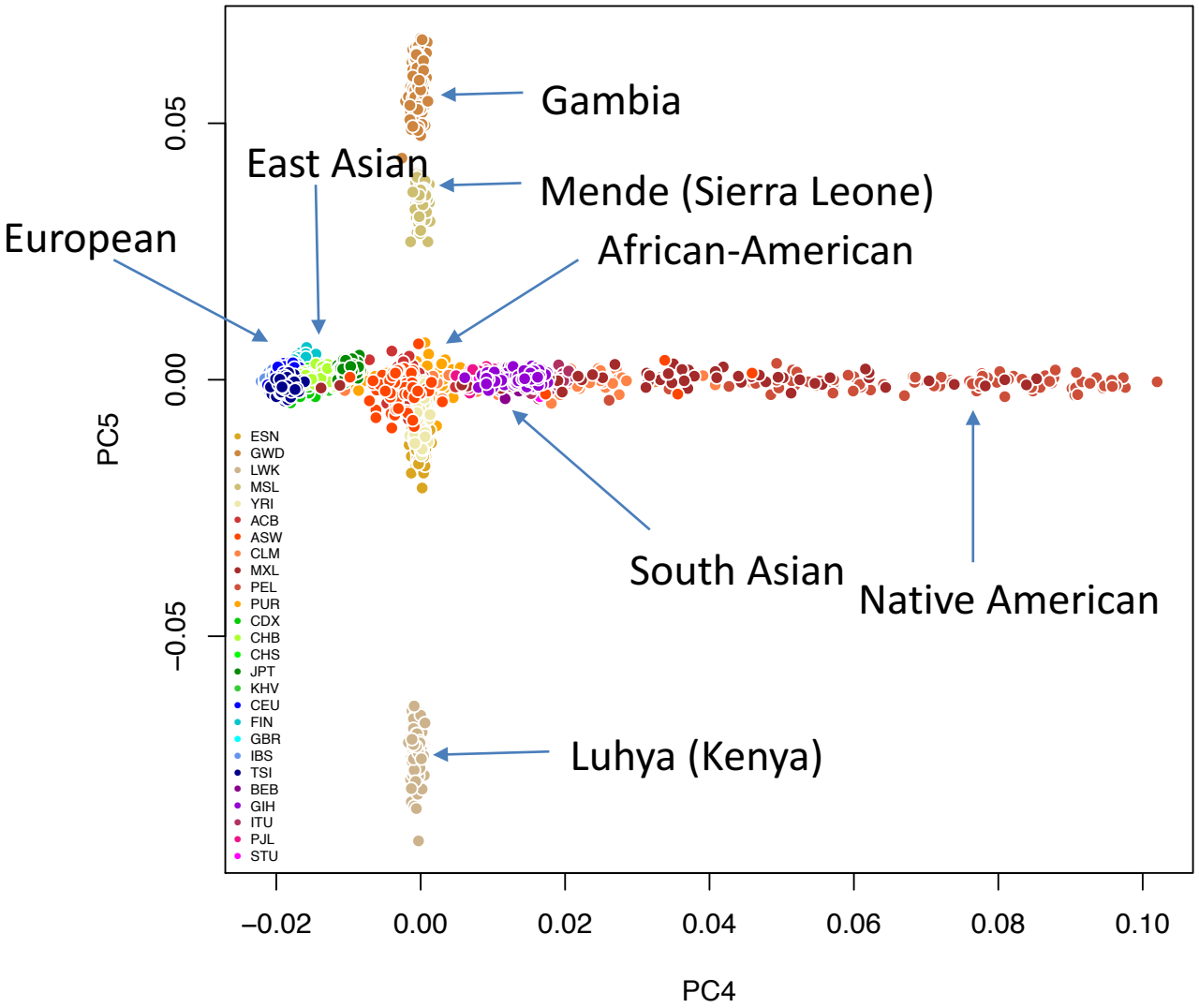
Slide: Iain Mathieson

Global population structure



Slide: Iain Mathieson

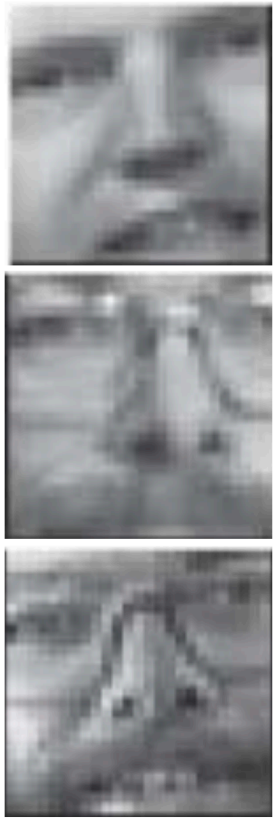
Global population structure



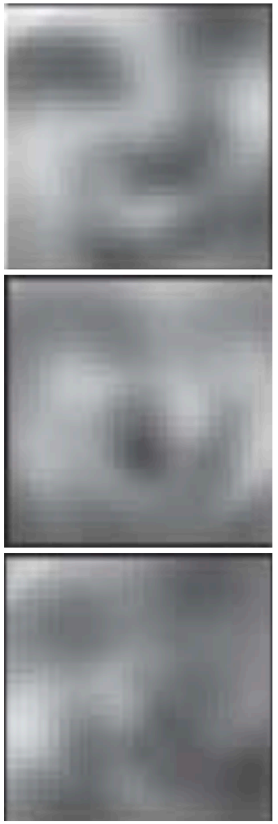
Slide: Iain Mathieson

PCA (linear) vs. Autoencoder (non-linear)

Original image



PCA reconstruction



Autoencoder reconstruction