

CS 260: Foundations of Data Science

Prof. Sara Mathieson

Spring 2025



Haverford
COLLEGE

Admin

- **Project meetings** with all groups in lab today
- **Midterm** returned on Thursday
- Next week: **project presentations in-class**

Outline

- Introduction to unsupervised learning
- K-means clustering
- Gaussian mixture models

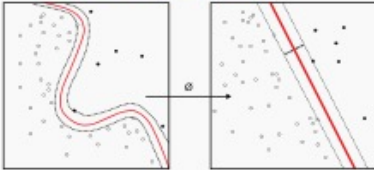
Outline

- Introduction to unsupervised learning
- K-means clustering
- Gaussian mixture models

Supervised Learning:

makes use of examples where we know the underlying “truth” (label/output)

Machine learning and data mining



Problems [show]

Supervised learning [hide]
(classification • regression)

Decision trees • Ensembles (Bagging, Boosting, Random forest) • *k*-NN • Linear regression • Naive Bayes • Neural networks • Logistic regression • Perceptron • Relevance vector machine (RVM) • Support vector machine (SVM)

Clustering [hide]
BIRCH • Hierarchical • *k*-means • Expectation-maximization (EM) • DBSCAN • OPTICS • Mean-shift

Dimensionality reduction [hide]
Factor analysis • CCA • ICA • LDA • NMF • PCA • t-SNE

Structured prediction [hide]
Graphical models (Bayes net, CRF, HMM)


Anomaly detection [hide]
k-NN • Local outlier factor

Neural nets [hide]
Autoencoder • Deep learning • Multilayer perceptron • RNN • Restricted Boltzmann machine • SOM • Convolutional neural network

Reinforcement Learning [hide]
Q-Learning • SARSA • Temporal Difference (TD)

Theory [show]

Machine learning venues [show]

 **Machine learning portal**

V • T • E

Unsupervised Learning:

Learn underlying structure or features without labeled training data

Unsupervised learning: 3 main areas

- 1) Clustering: group data points into clusters based on features only
- 2) Dimensionality reduction: remove feature correlation, compress data, visualize data
- 3) Structured prediction: model latent variables (example: Hidden Markov Models)

Unsupervised learning examples from biology: clustering

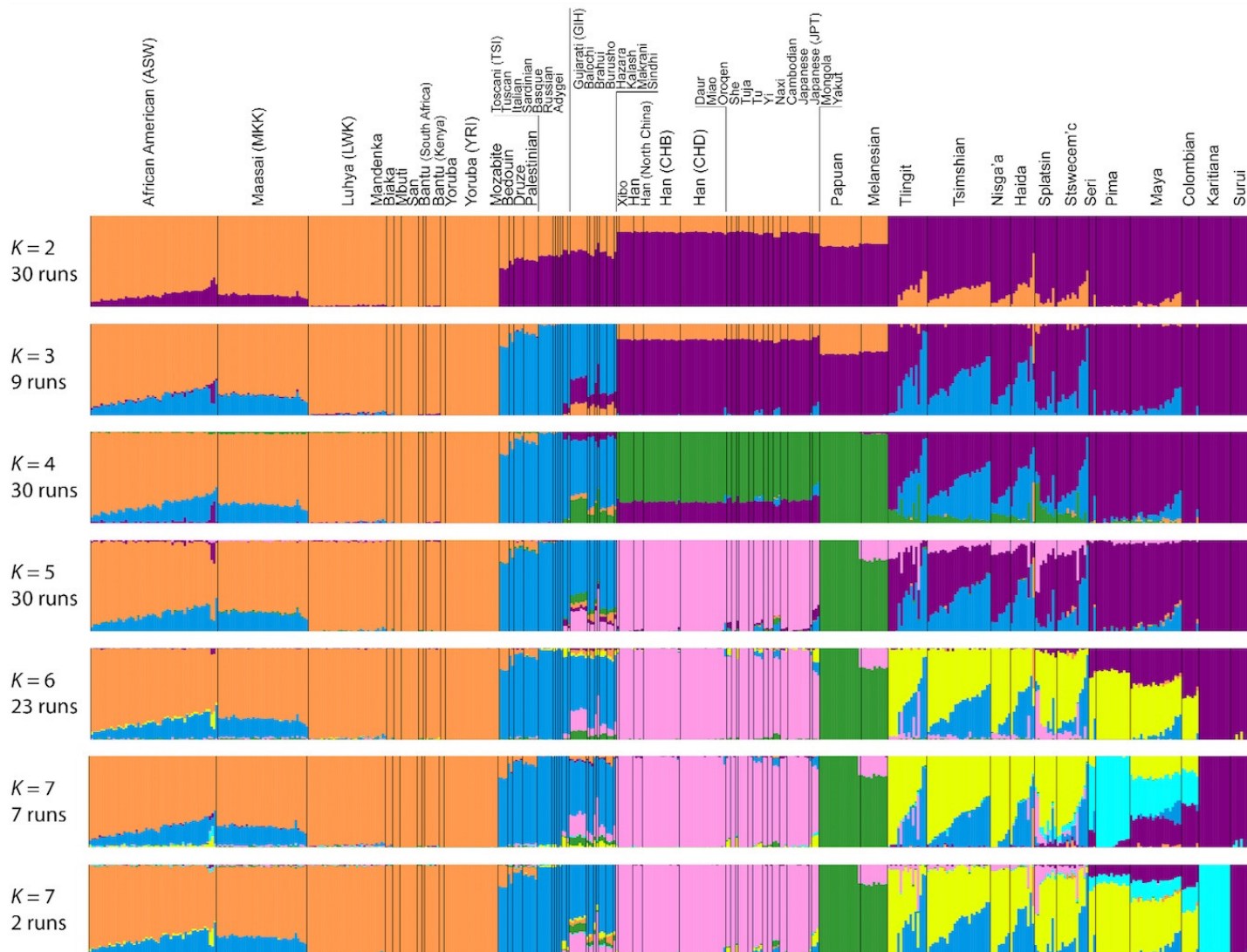
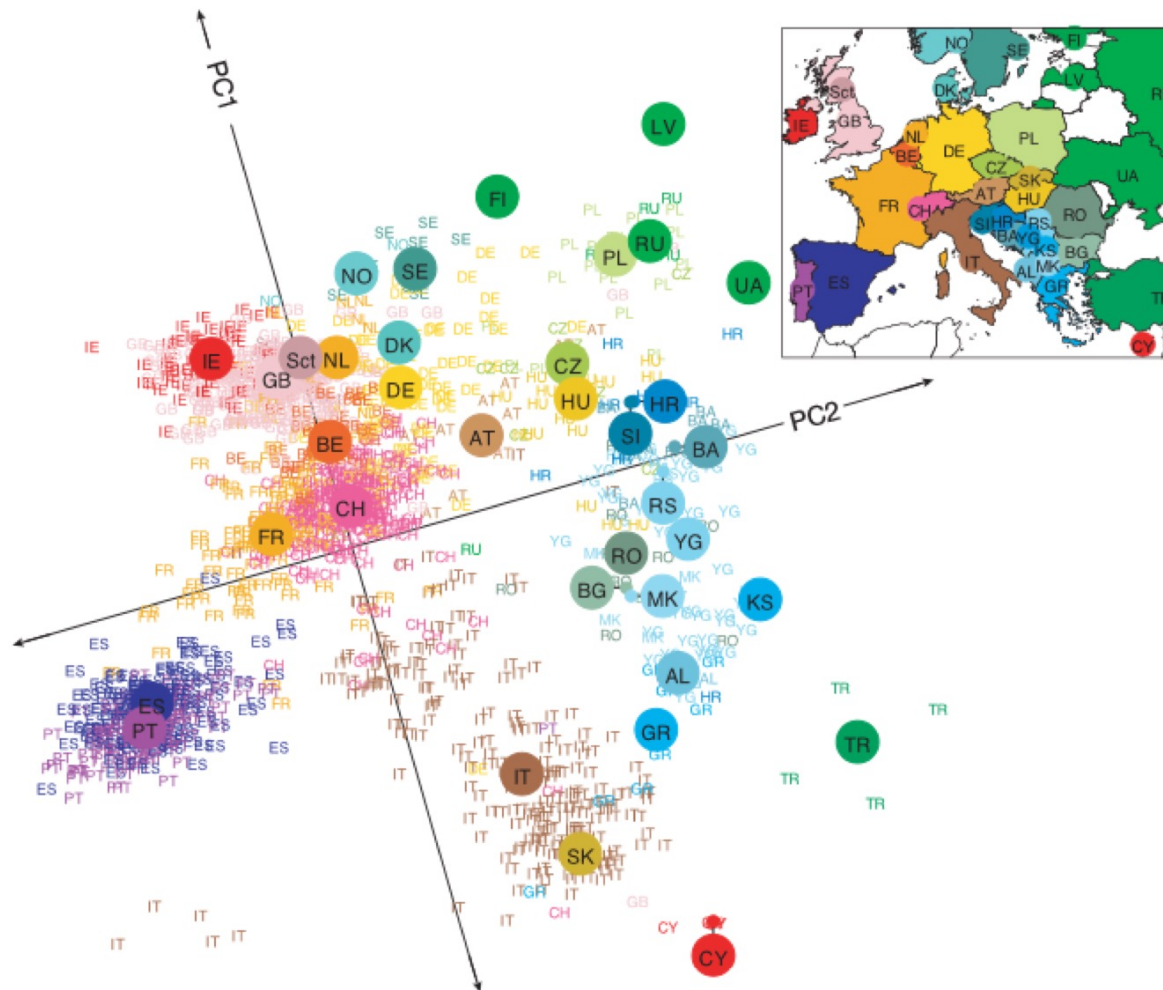
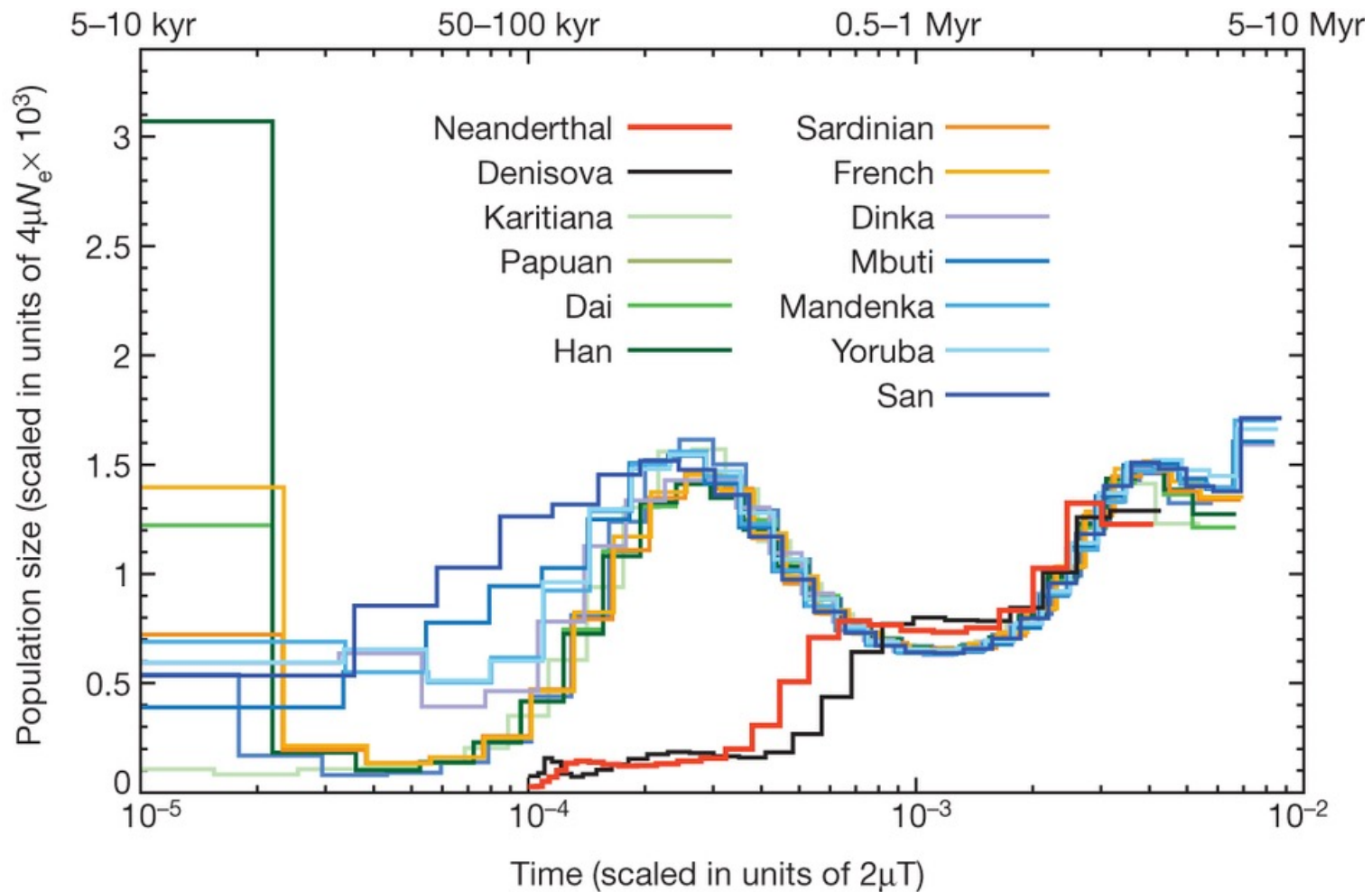


Figure: German Dziel

Unsupervised learning examples from biology: structured prediction



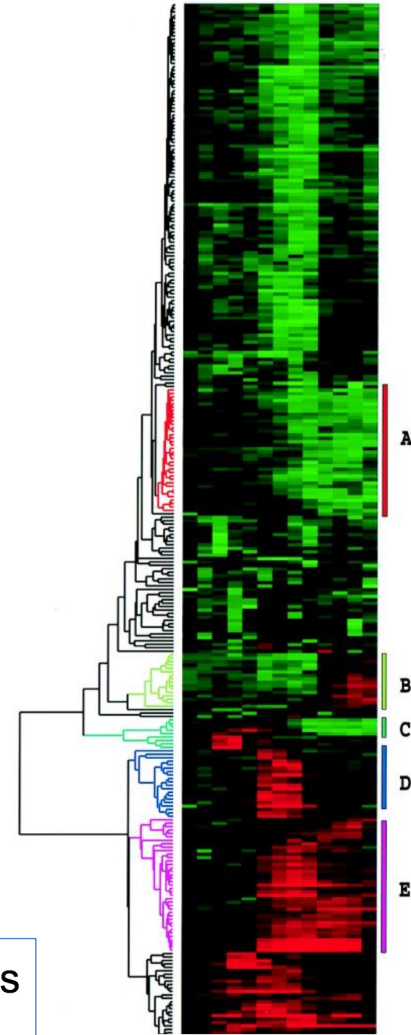
Unsupervised learning examples from biology: structured prediction



Clustering overview

Applications of clustering

- Cluster genes with similar expression patterns

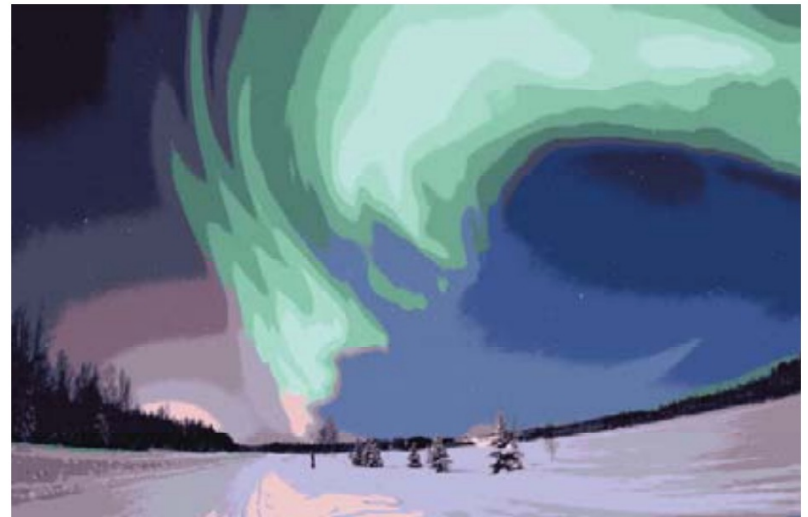


Cluster analysis and display of genome-wide expression patterns

[Michael B. Eisen](#),^{*} [Paul T. Spellman](#),^{*} [Patrick O. Brown](#),[†] and [David Botstein](#)^{*‡}

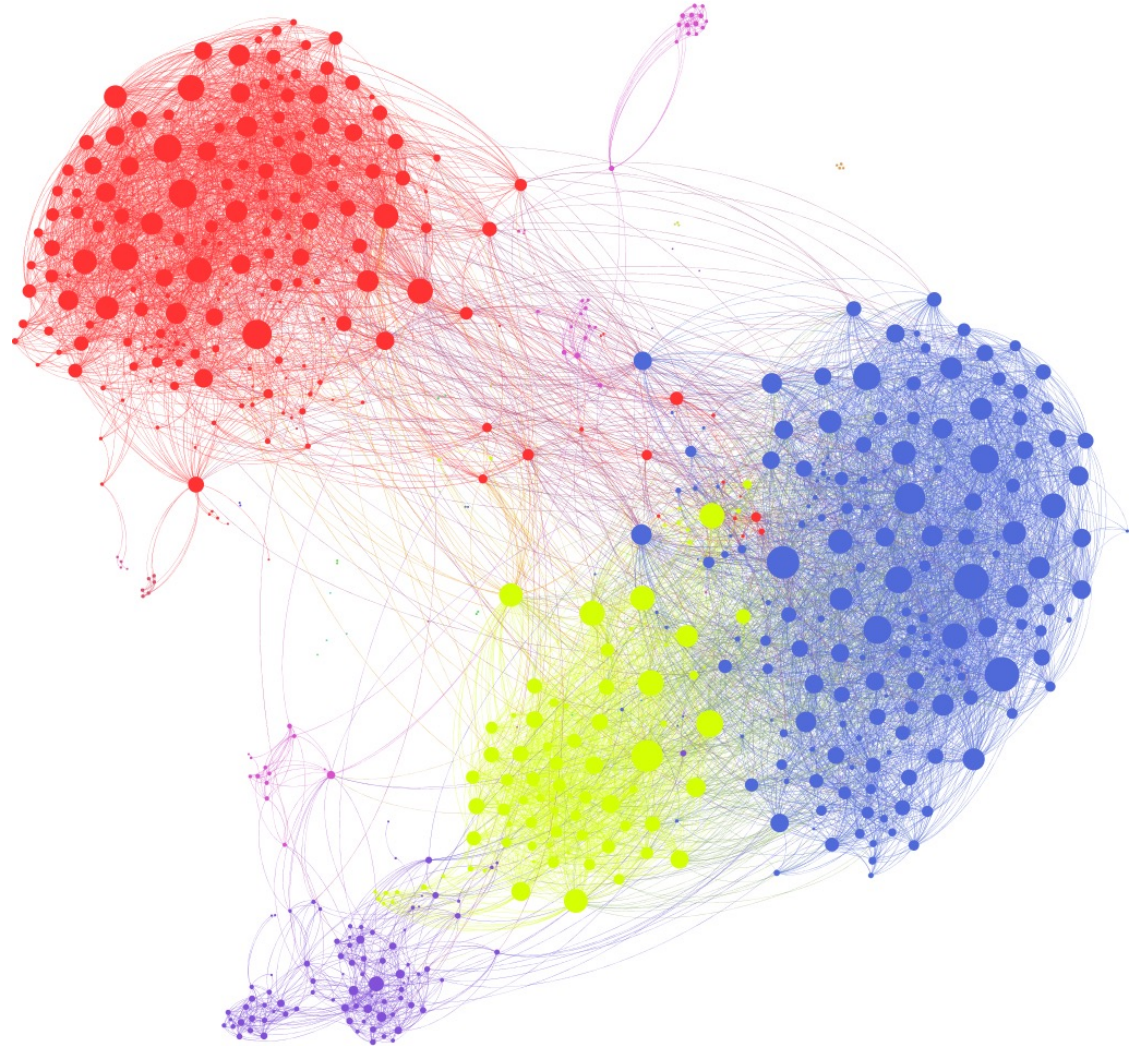
Applications of clustering

- Image segmentation: cluster similar regions of an image



Applications of clustering

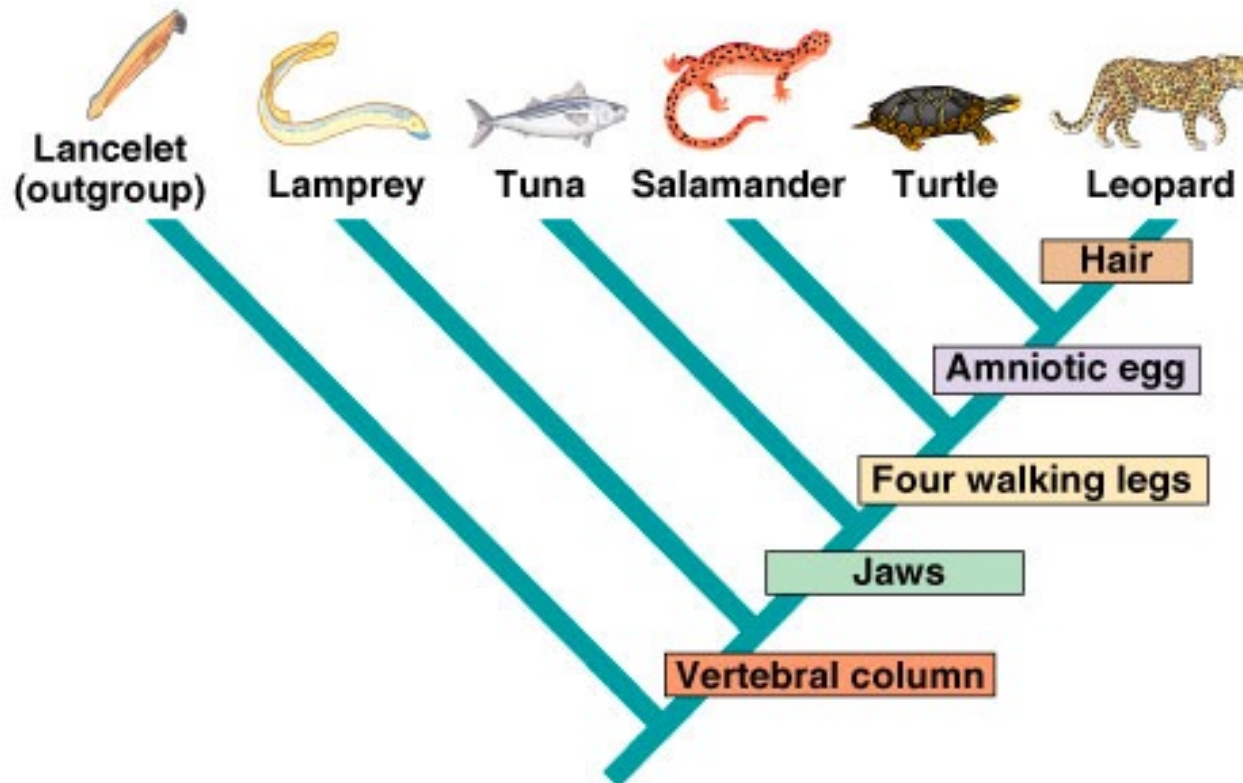
- Clustering in social graphs



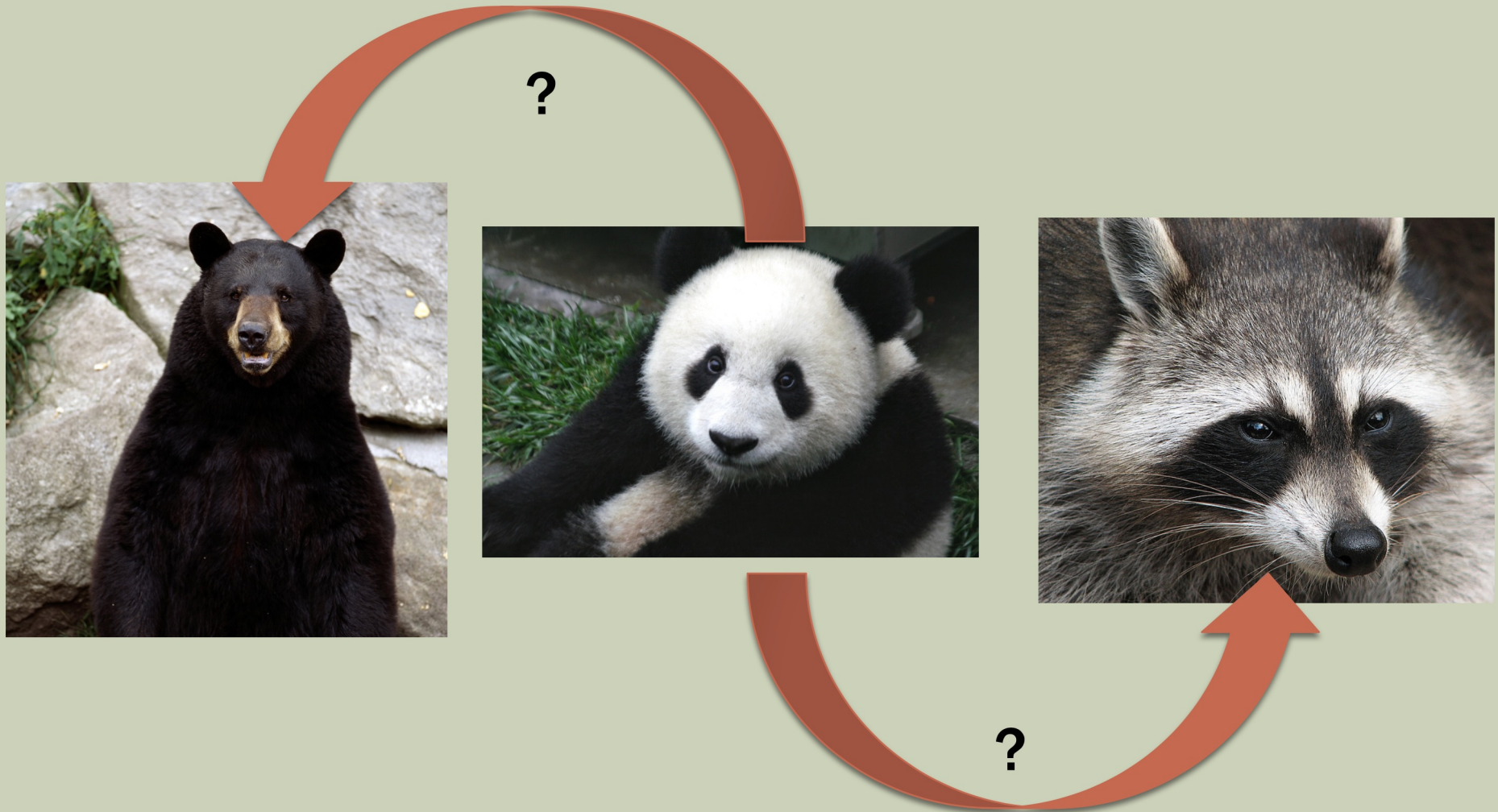
Two main types of clustering

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

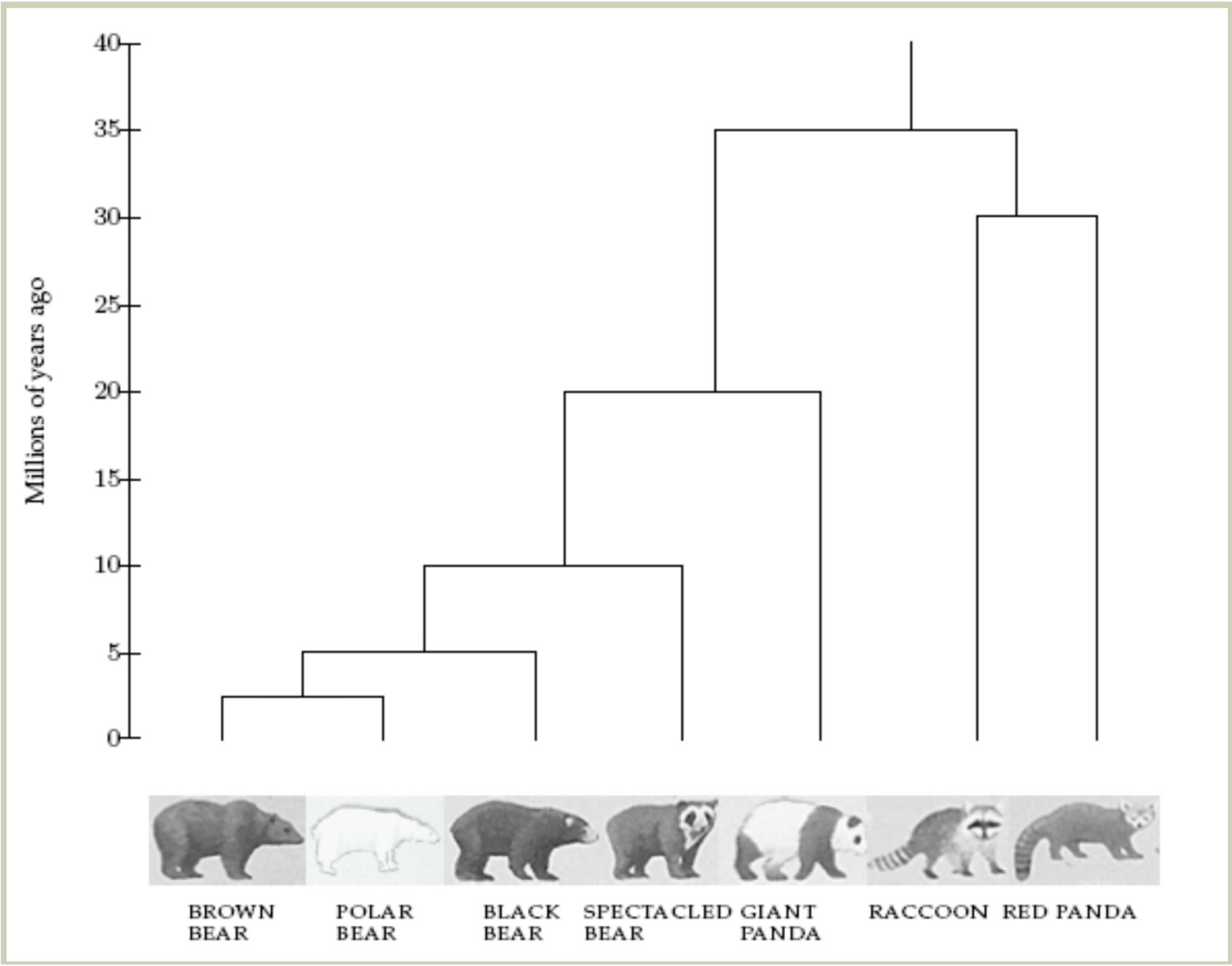
Hierarchical clustering example: trees



Are pandas more closely related to bears or raccoons?

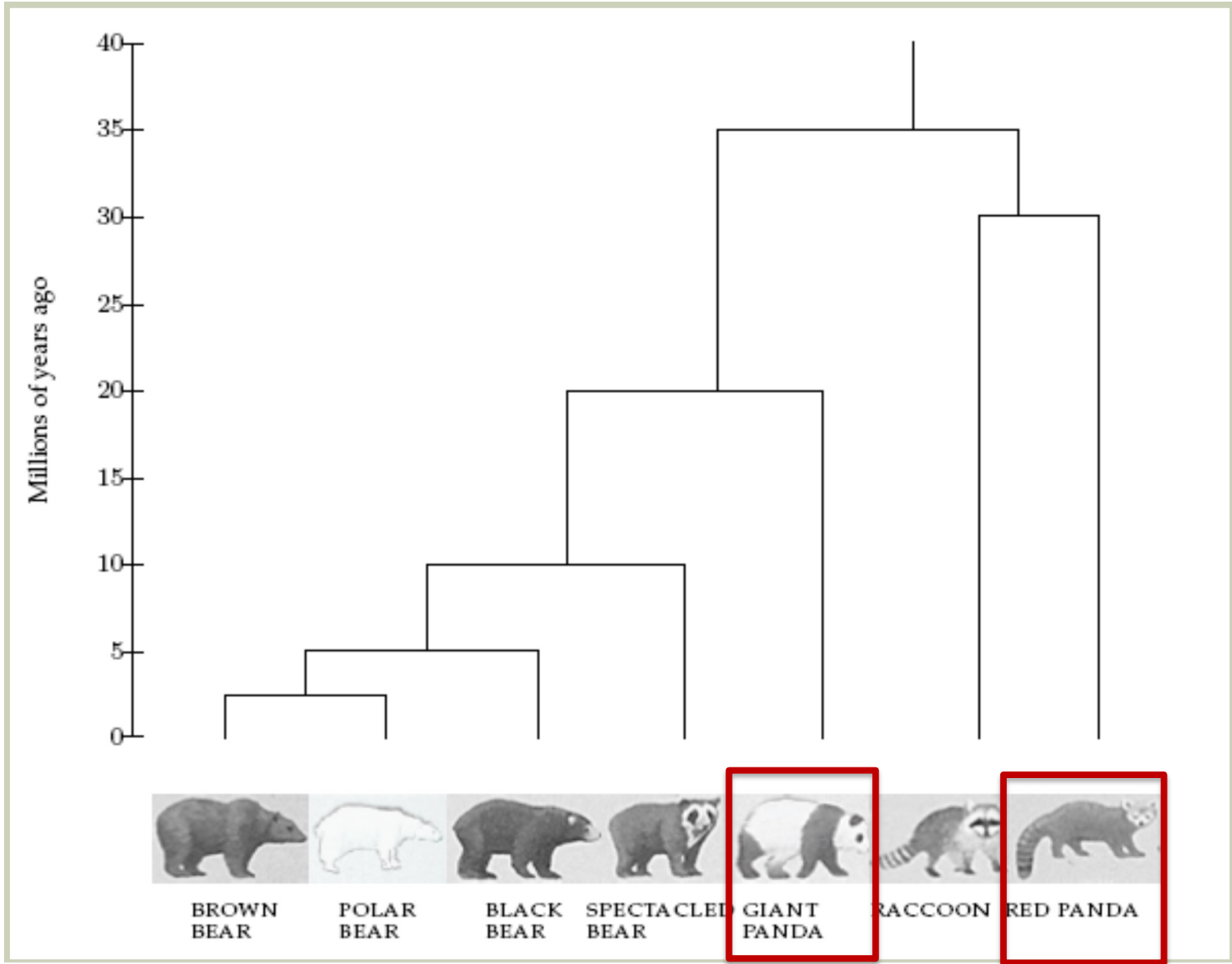


Are pandas more closely related to bears or raccoons?



Credit:
Ameet
Soni

Are pandas more closely related to bears or raccoons?



Credit:
Ameet
Soni

Outline

- Introduction to unsupervised learning
- **K-means clustering**
- Gaussian mixture models

K-means

given

$$X = \begin{bmatrix} - & \vec{x}_1^T & - \\ & \vdots & \\ - & \vec{x}_n^T & - \end{bmatrix}_{n \times p}$$

n points
p features/
(dimensions)

K : # clusters

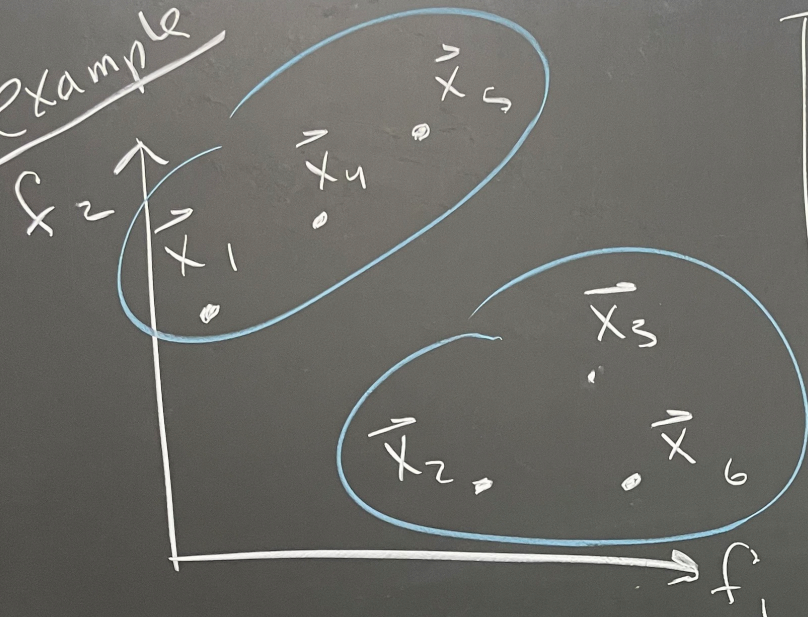
Goal

find

$$\mathcal{C} = \{c_1, c_2, \dots, c_K\}$$

that minimize WCSS

Example



$$\begin{array}{l} n = 6 \\ p = 2 \\ K = 2 \end{array}$$

$$e_1 = \{ \vec{x}_1, \vec{x}_4, \vec{x}_5 \}$$

$$e_2 = \{ \vec{x}_2, \vec{x}_3, \vec{x}_6 \}$$

desired
output

$$e = \{ e_1, e_2 \}$$

WCSS: within cluster sum of squares

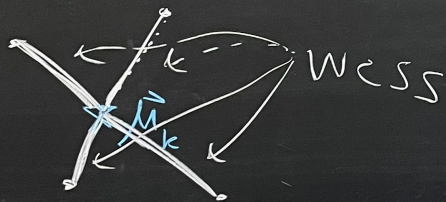
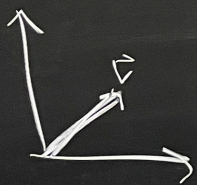
Cost function

$$\underbrace{J(\mathcal{C})}_{\text{minimize!!}} = \sum_{k=1}^K \sum_{\vec{x}_i \in \mathcal{C}_k} \|\vec{x}_i - \vec{\mu}_k\|^2$$

↑ all clusters ↑ pts within cluster k ↑ mean of cluster k

NP-hard

L^2 norm: $\|\vec{v}\|^2 = v_1^2 + v_2^2 + \dots + v_p^2$



K-means algorithm

① initialization: choose means (center) from among training data.

$$\vec{\mu}_1^{(1)}, \vec{\mu}_2^{(1)}, \dots, \vec{\mu}_K^{(1)}$$

iterate (iteration t)

① E-step assignment

assign each data point \vec{x} to the closest mean

$$\vec{x} \in \mathcal{C}_k^{(t)}$$

↑
is

② M-step update

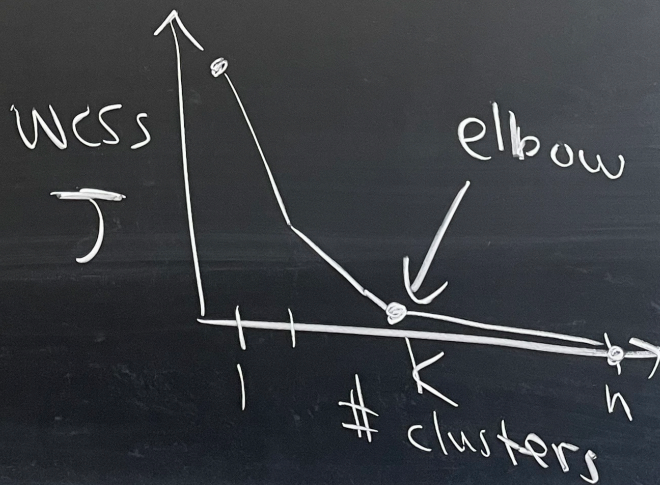
recompute each mean as the average of all cluster members

$$\vec{\mu}_k^{(t+1)} = \frac{1}{|e_k^{(t)}|} \sum_{\vec{x}_i \in e_k^{(t)}} \vec{x}_i$$

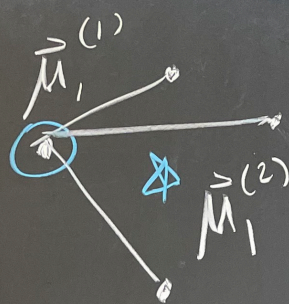
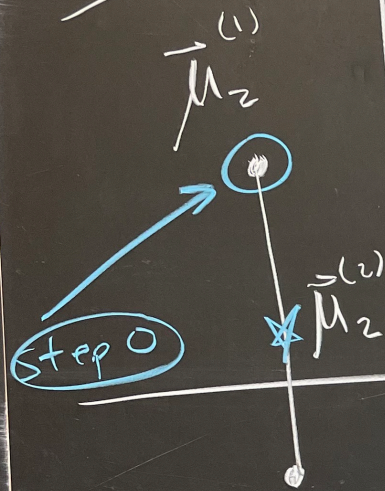
how to stop?

- max # iters T
- means / clusters not changing
- cycle

how to choose K ?

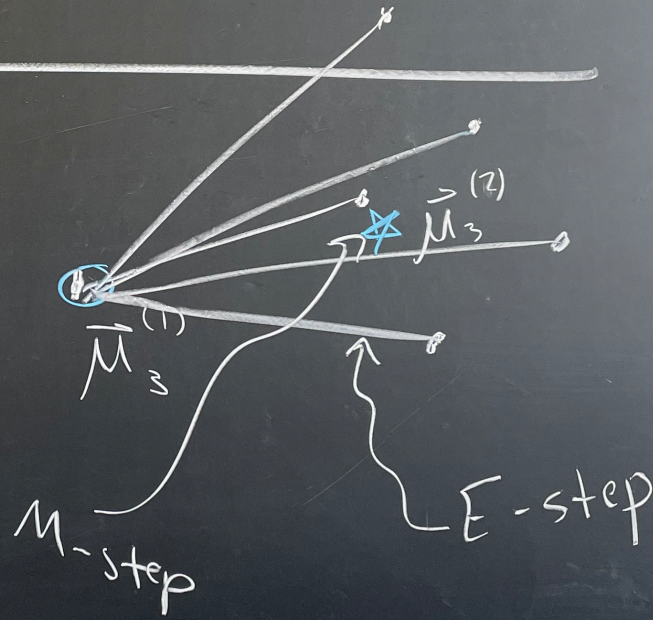


$K=3$



~~$K=n$~~
 ~~$J=0$~~

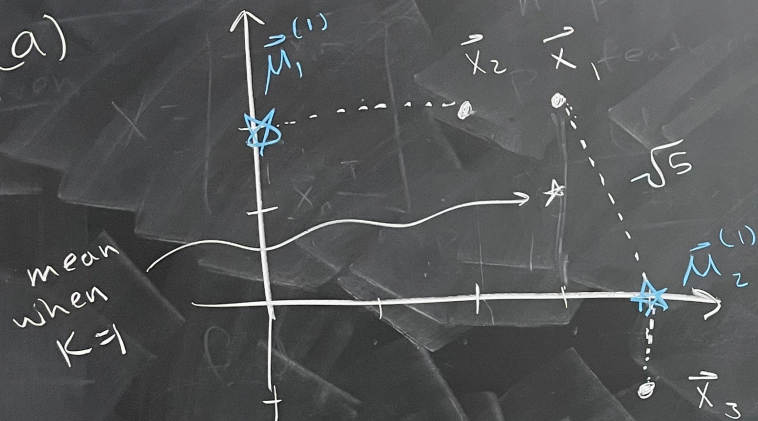
$K?$
1



Handout 23

①

(a)



$$X = \begin{pmatrix} 3 & 2 \\ 2 & 2 \\ 4 & -1 \end{pmatrix}$$

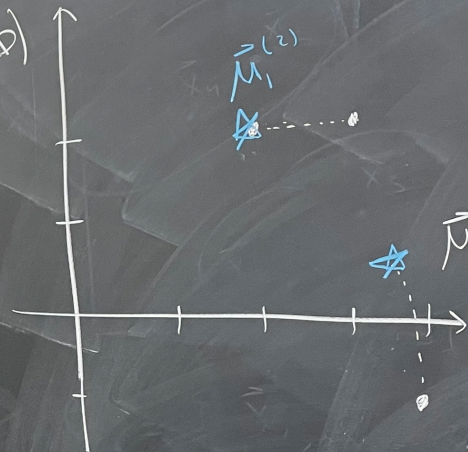
3 1

$$\vec{x}_1, \vec{x}_2, \vec{x}_3$$

$$e_1^{(1)} = \{\vec{x}_2\}$$

$$e_2^{(1)} = \{\vec{x}_1, \vec{x}_3\}$$

(b)



$$e_1^{(2)} = \{\vec{x}_1, \vec{x}_2\} \rightarrow \vec{\mu}_1^{(2)} = \begin{bmatrix} 2.5 \\ 2 \end{bmatrix}$$

$$e_2^{(2)} = \{\vec{x}_3\} \rightarrow \vec{\mu}_2^{(2)} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$$

② $WCSS$ should generally decrease as K increases

③ $K=1$

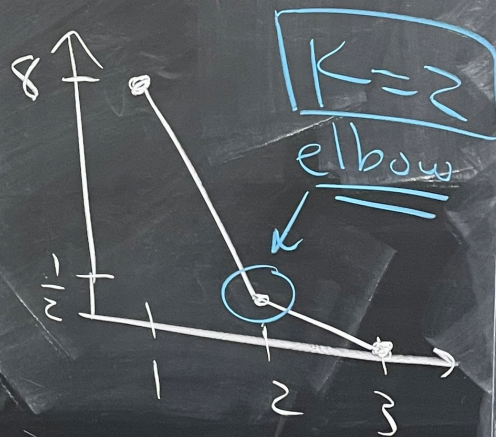
$$WCSS = 2 + 1 + 5 = \boxed{8}$$

$K=2$

$$WCSS = \boxed{\frac{1}{2}}$$

$K=3$

$$WCSS = \boxed{0}$$

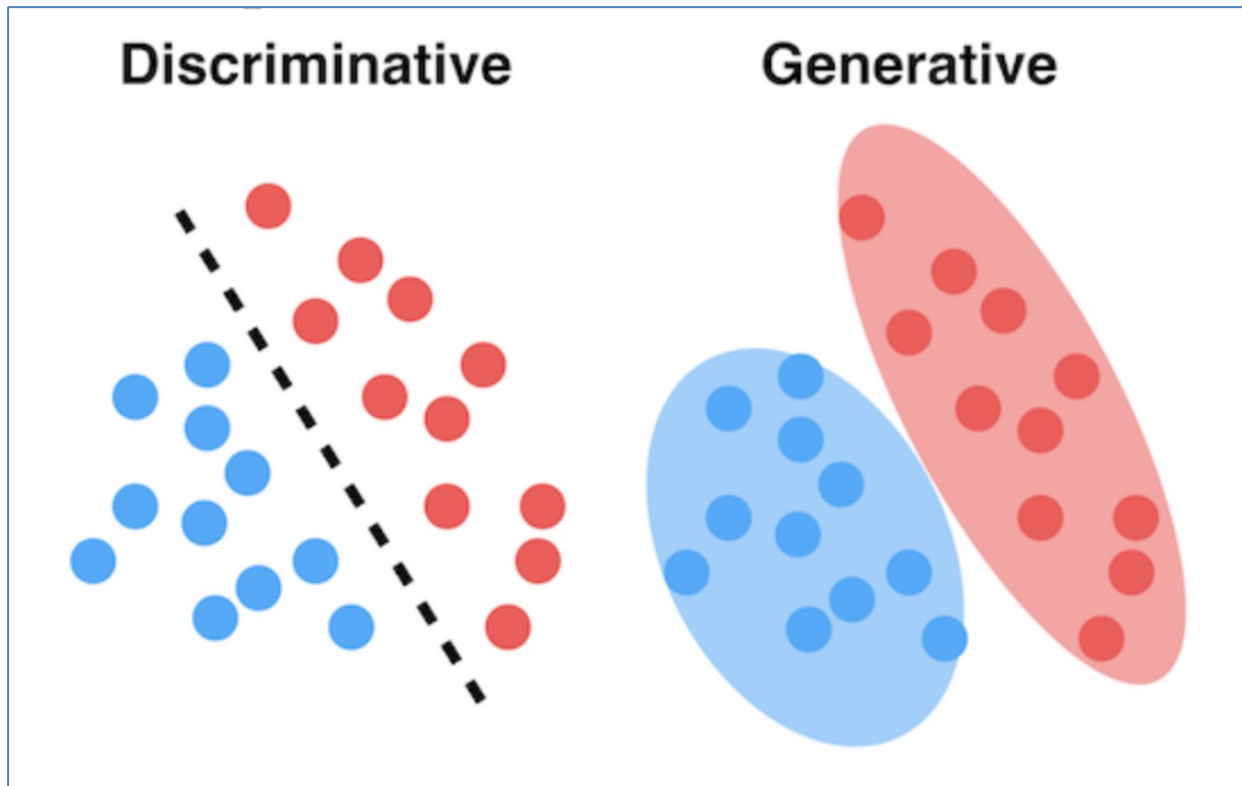


Outline

- Introduction to unsupervised learning
- K-means clustering
- Gaussian mixture models

Discriminative vs. Generative

- Discriminative: finds a decision boundary
 - Logistic regression, K-means
- Generative: estimates probability distributions
 - Naïve Bayes, Gaussian Mixture Models



Gaussian Mixture Models (GMMs)

Problems w/ K-means

- * not generative (could not create a new datapoint)
- * point cannot belong to more than one cluster
- * does not account for different cluster sizes & variances

Likelihood

$$P(\vec{x}) = \sum_{k=1}^K P(\vec{x}, z=k) = \sum_{k=1}^K p(z=k) p(\vec{x} | z=k)$$

one point

cluster membership

BAYES

size of cluster k (prior)

assume normal/gaussian distribution

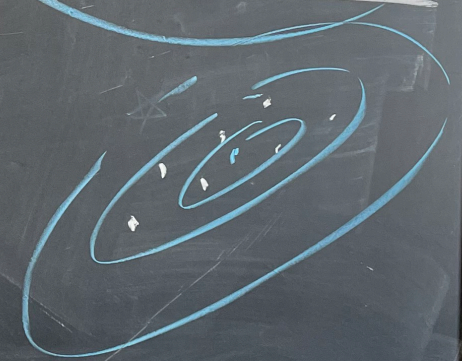
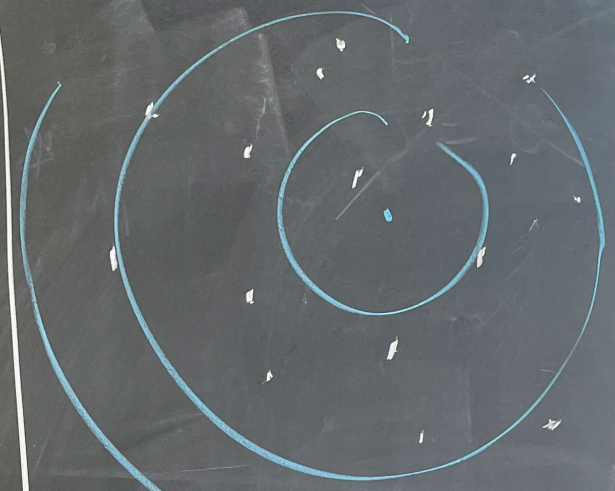
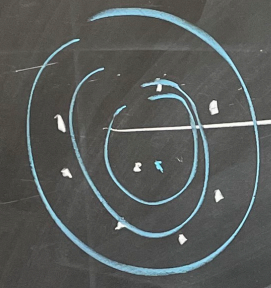
$$L(X) = \prod_{i=1}^n \sum_{k=1}^K \pi_k N(\vec{x}_i; \vec{\mu}_k, \Sigma_k^2)$$

Goal: find $\pi_k, \vec{\mu}_k, \Sigma_k^2$ for $k=1 \dots K$

$|z=k)$

assume
normal/gaussian
distribution

$\frac{1}{k}$



Example of GMMs with different covariance constraints on the Iris flower data

