CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025

HAVERFORD COLLEGE

Materials by Sara Mathieson

Admin

- **Final project presentation** sign-up on Piazza
  - Presentation guidelines posted on course webpage
  - Email me pdfs of your slides the night before
  - Class attendance taken for Dec 9, 10, 11

Outline for today

- Clustering overview
- K-means
- Gaussian Mixture Models (GMMs)

Outline for today

- Clustering overview
- K-means
- Gaussian Mixture Models (GMMs)

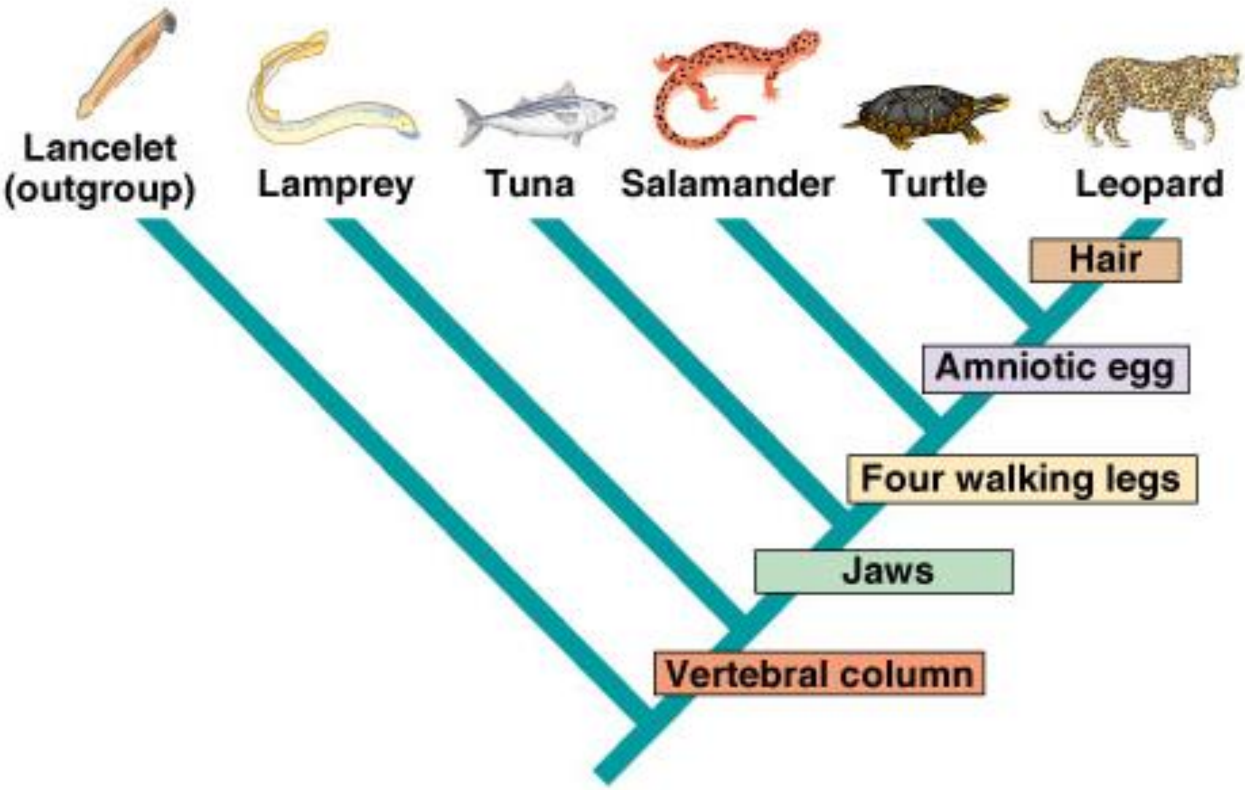
Clustering

- Learn about the structure in our data
- Cluster new data (prediction)
- Goal: $\mathrm{C} = \{C_1, C_2, \dots, C_k\}$ such that within cluster difference is minimized

Two main types of clustering

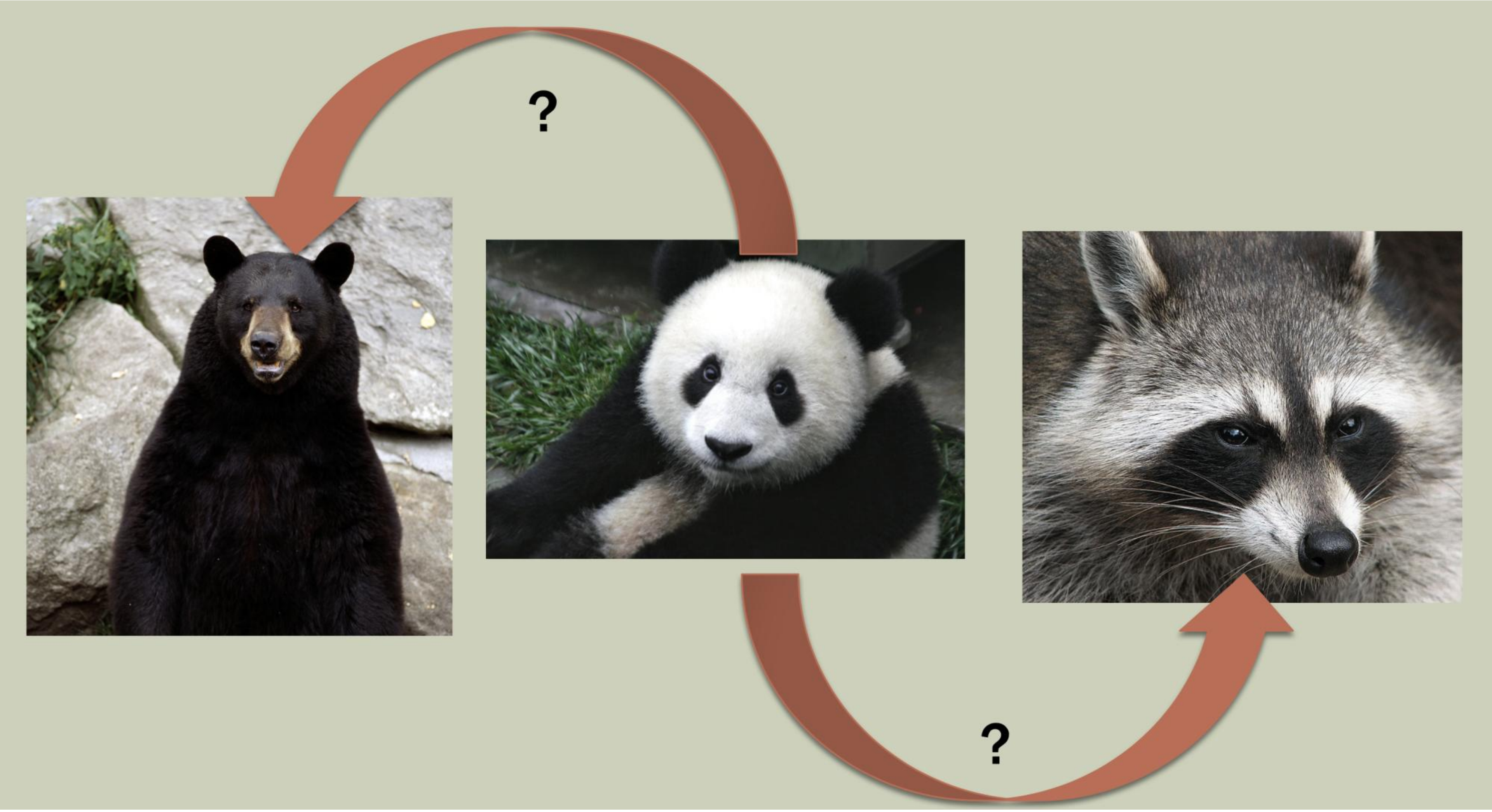
- Flat/Partitional:
  - K-means
  - Gaussian mixture models
- Hierarchical:
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and Neighbor Joining

Hierarchical clustering example: trees



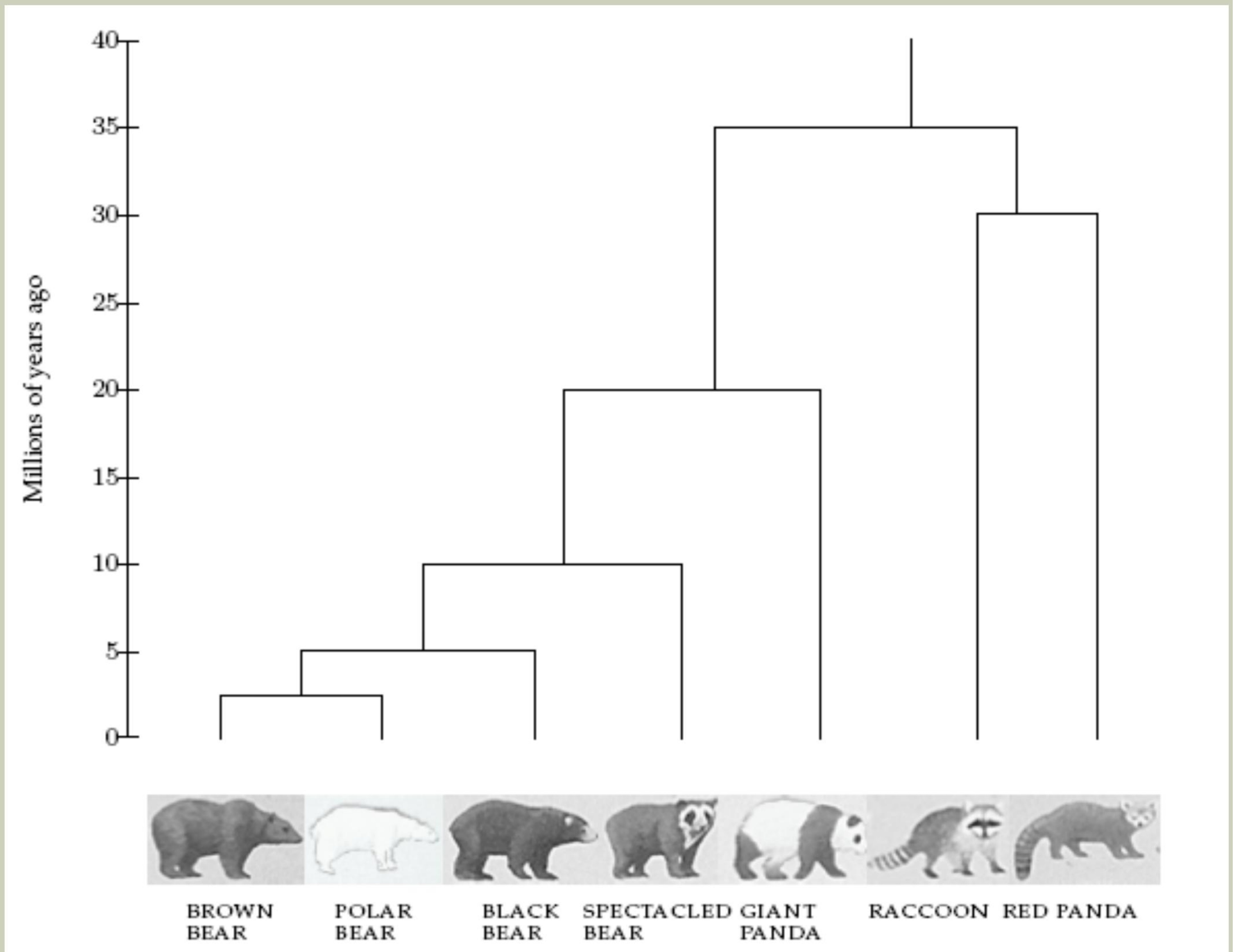
Credit: Pearson Education, Benjamin Cummings

Are pandas more closely related to bears or raccoons?



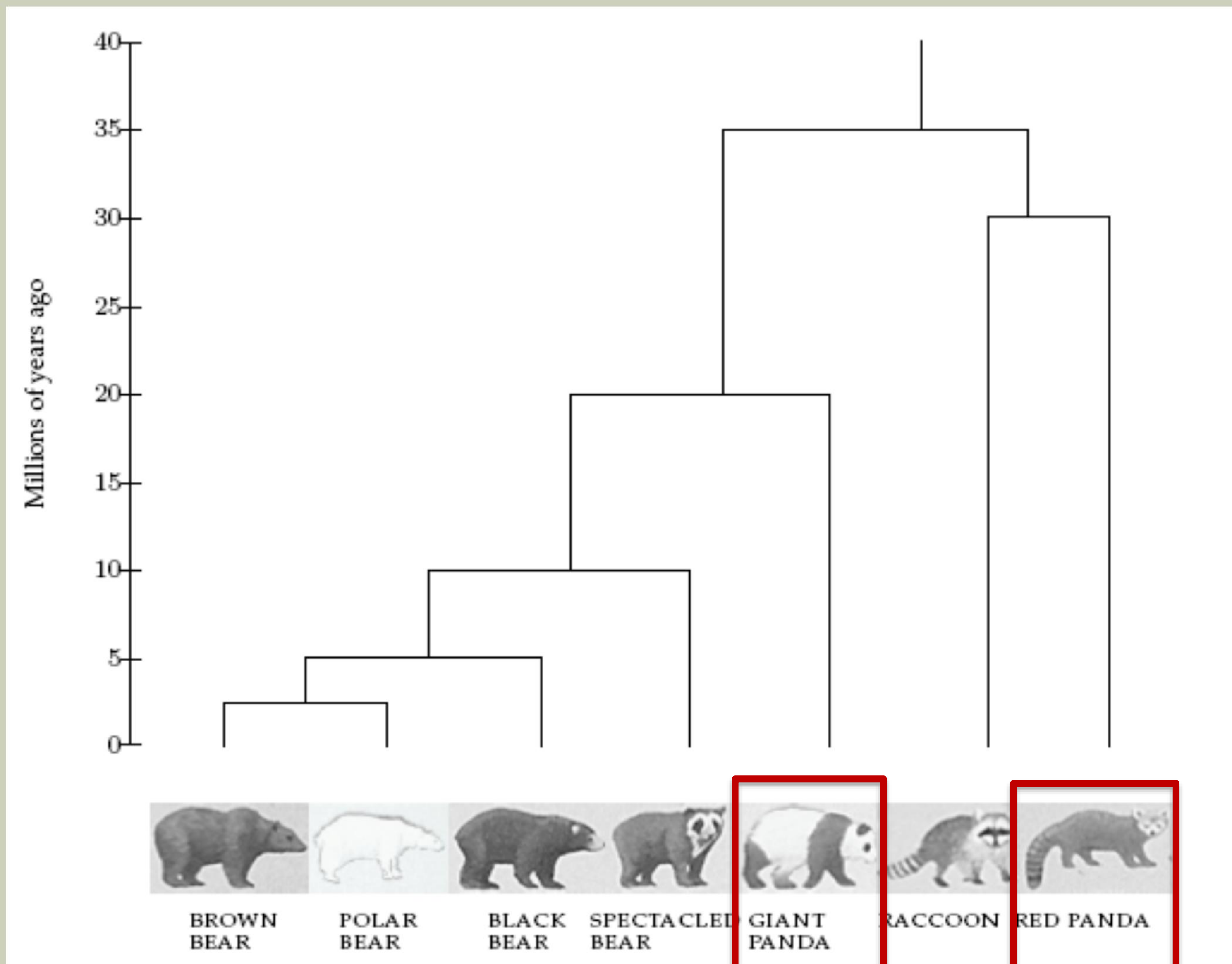
Credit: Ameet Soni

Are pandas more closely related to bears or raccoons?



Credit: Ameet Soni

What about red pandas?



Credit: Ameet Soni

Outline for today

- Clustering overview
- K-means
- Gaussian Mixture Models (GMMs)

K-means Algorithm

- Initialization step: Choose k means (cluster centers) randomly from the data

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

- Expectation-maximization (EM) algorithm
  - E-step: assign each datapoint to the closest mean

$$\vec{x}_i \in C_k^{(t)}$$

  - M-step: recompute means as the cluster average

iterate

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

K-means Algorithm

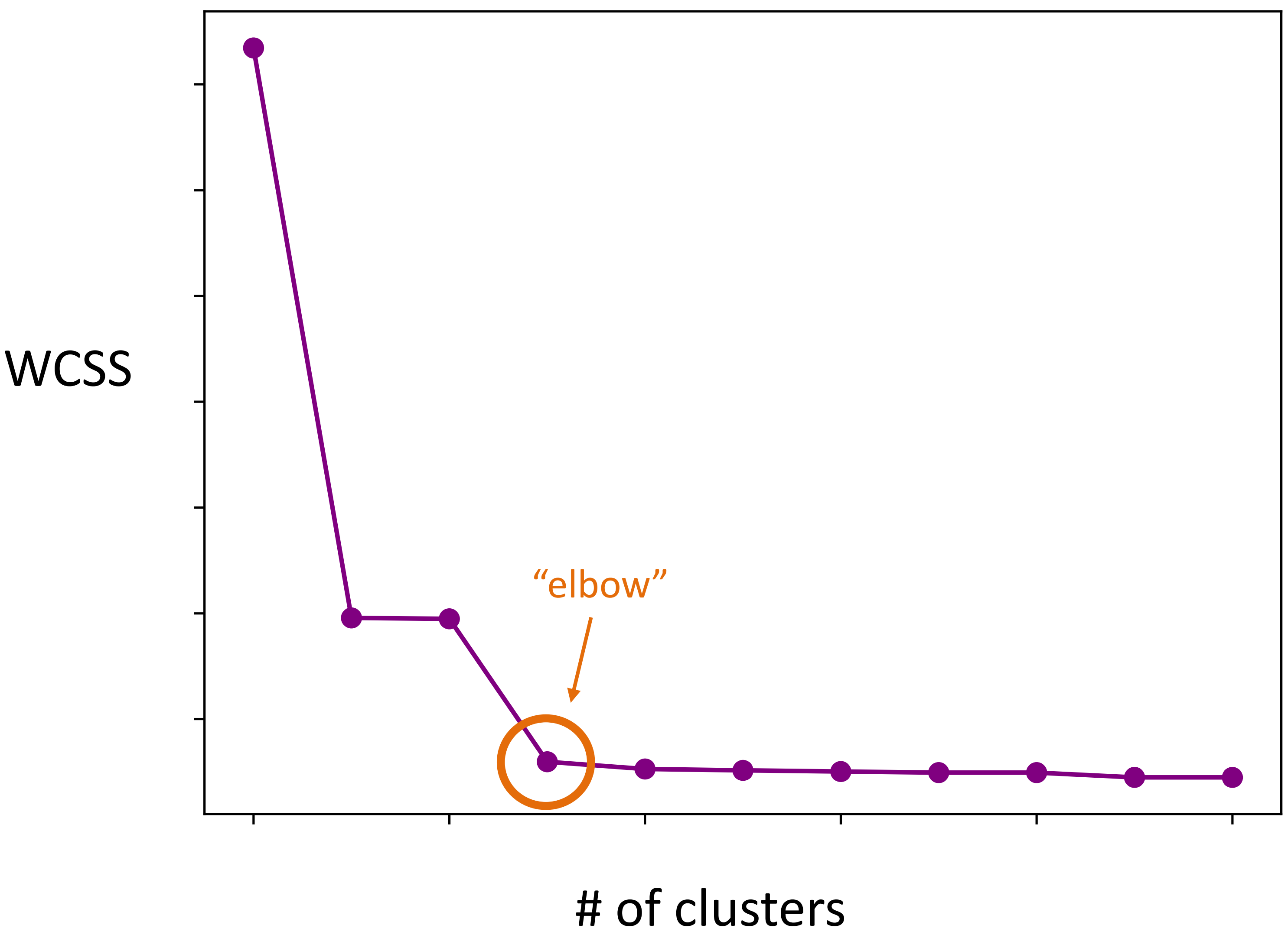
- Minimizes:

$$WCSS = \sum_{k=1}^{K} \sum_{\vec{x}_i \in C_k} \left\| \vec{x}_i - \vec{\mu}_k \right\|^2$$

within-cluster
sum of squares

- Stopping criteria:
  - No change in cluster membership
  - Max # of iterations exceeded
  - Configuration/pattern you've seen before

How to choose k?



Handout 21

Handout 21

1.

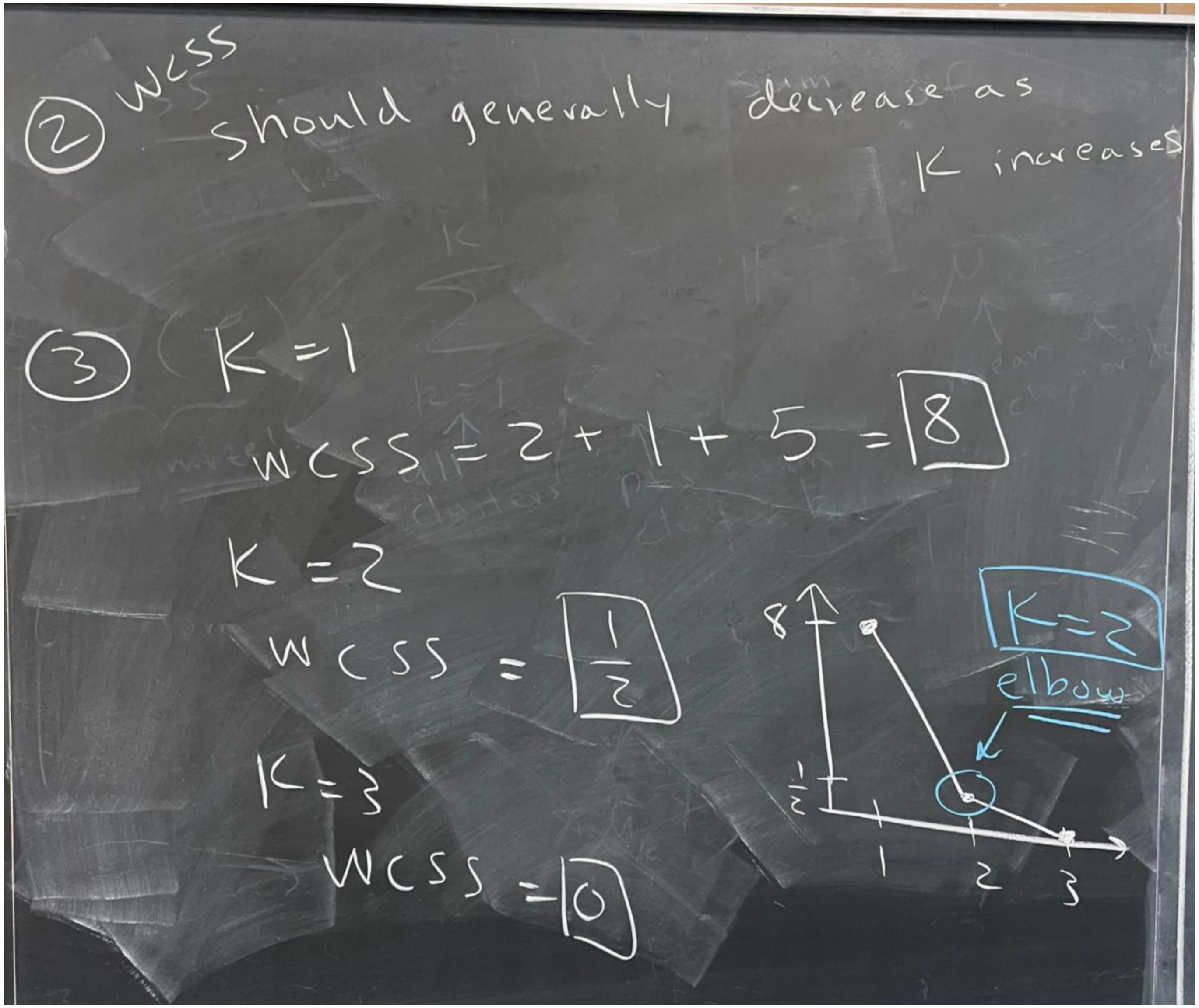
a) E-step: $C_1^{(1)} = \{\vec{x}_2\},\ C_2^{(1)} = \{\vec{x}_1, \vec{x}_3\}$

b) M-step: $\vec{\mu}_1^{(2)} = [2 \quad 2]^T, \vec{\mu}_2^{(2)} = [3.5 \quad 0.5]^T$

c) E-step: $C_1^{(2)} = \{\vec{x}_1, \vec{x}_2\},\ C_2^{(2)} = \{\vec{x}_3\}$

M-step: $\vec{\mu}_1^{(3)} = [2.5 \quad 2]^T, \vec{\mu}_2^{(3)} = [4 \quad -1]^T$

Handout 21



5. Runtime is O(npKT)

Outline for today

- Clustering overview
- K-means
- Gaussian Mixture Models (GMMs)

Problems with K-means

- Does not account for different cluster sizes, variances, and shapes
- Does not allow points to belong to multiple clusters
- Not generative (cannot create a new data point)

Discriminative vs. Generative Algorithms

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

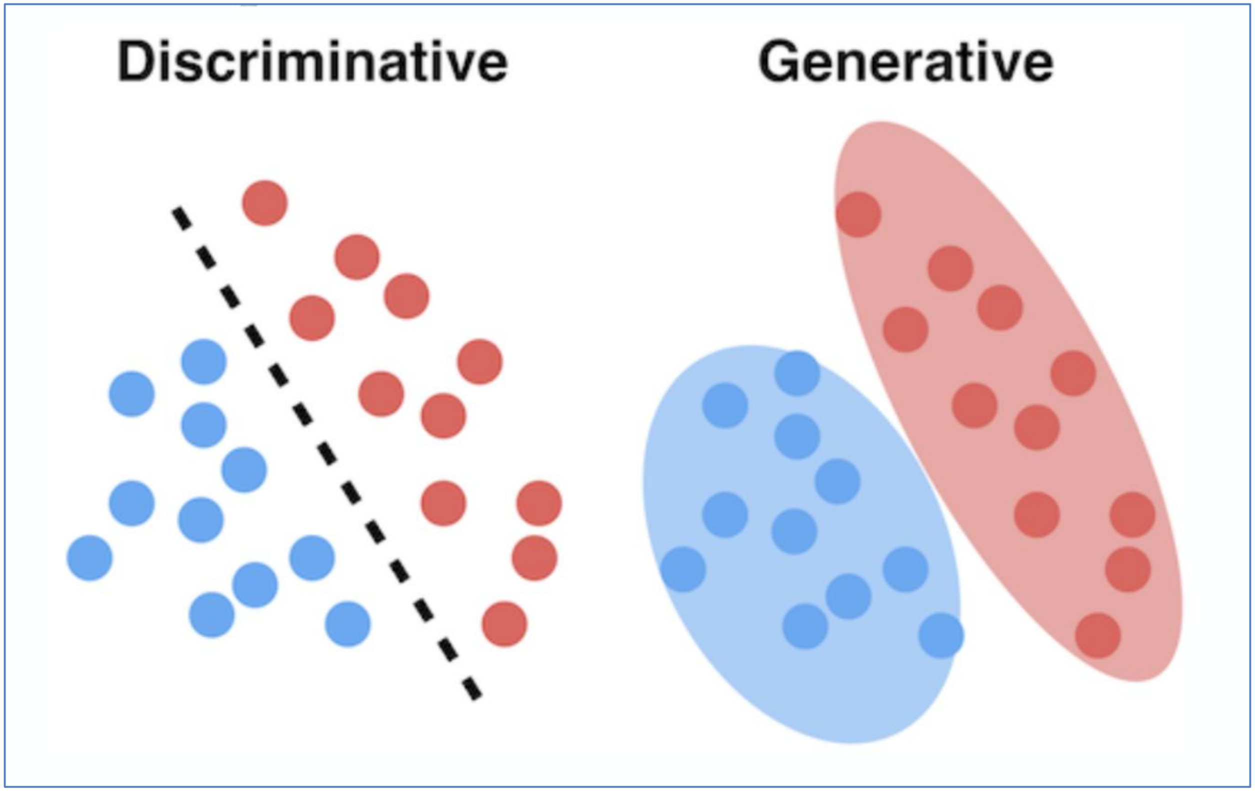


Figure: Ameet Soni

Gaussian Mixture Models (GMMs)

$$p(\vec{x}_i) = \sum_{k=1}^{K} \underbrace{p(\vec{x}_i, k)}_{\text{cluster membership}} = \sum_{k=1}^{K} p(k)p(\vec{x}_i|k) = \sum_{k=1}^{K} \overbrace{\pi_k}^{\text{prior over cluster sizes}} \underbrace{N\left(\vec{x}_i \middle| \vec{\mu}_k, \sigma_k^2\right)}_{\text{assume Gaussian distribution}}$$

- Maximize likelihood:

$$L(X) = \prod_{i=1}^{n} p(\vec{x}_i) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N\left(\vec{x}_i \middle| \vec{\mu}_k, \sigma_k^2\right)$$

Model parameters