CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2024

HAVERFORD COLLEGE

Materials by Sara Mathieson

Admin

- **Midterm 2** due today!

- **Tuesday + Wednesday:** work on final project
  - Try to come to the same lab session as your partner

Outline for today

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

Outline for today

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

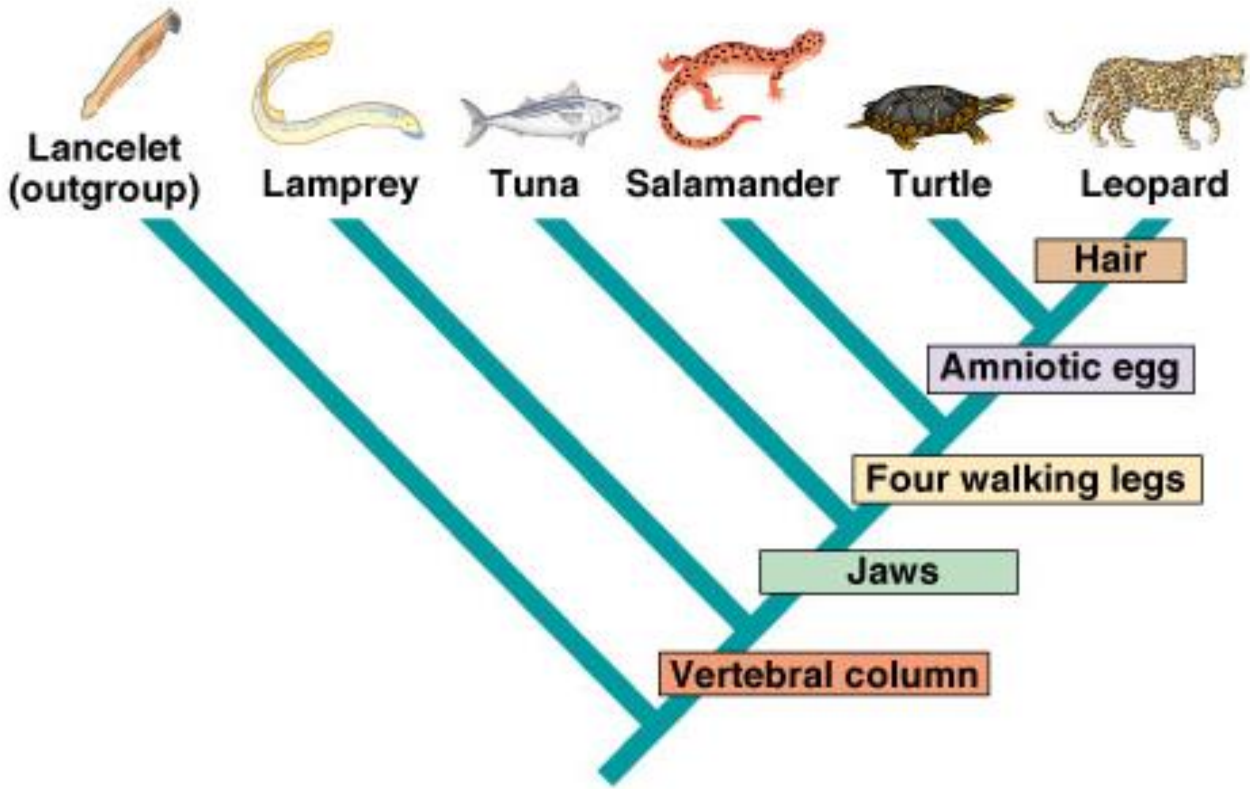
Clustering

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

Two main types of clustering

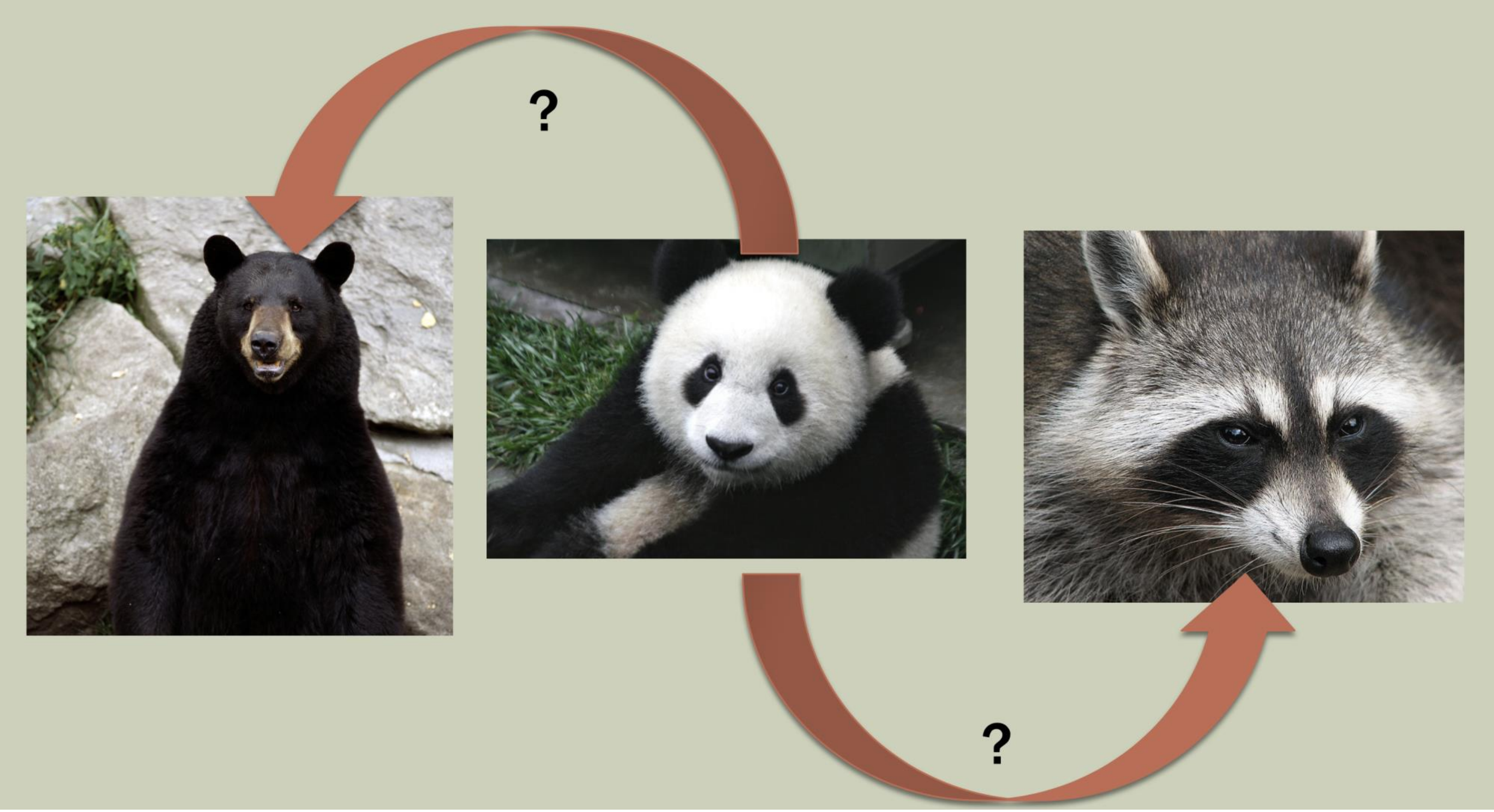
- Flat/Partitional:
  - K-means
  - Gaussian mixture models
- Hierarchical:
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and Neighbor Joining

Hierarchical clustering example: trees



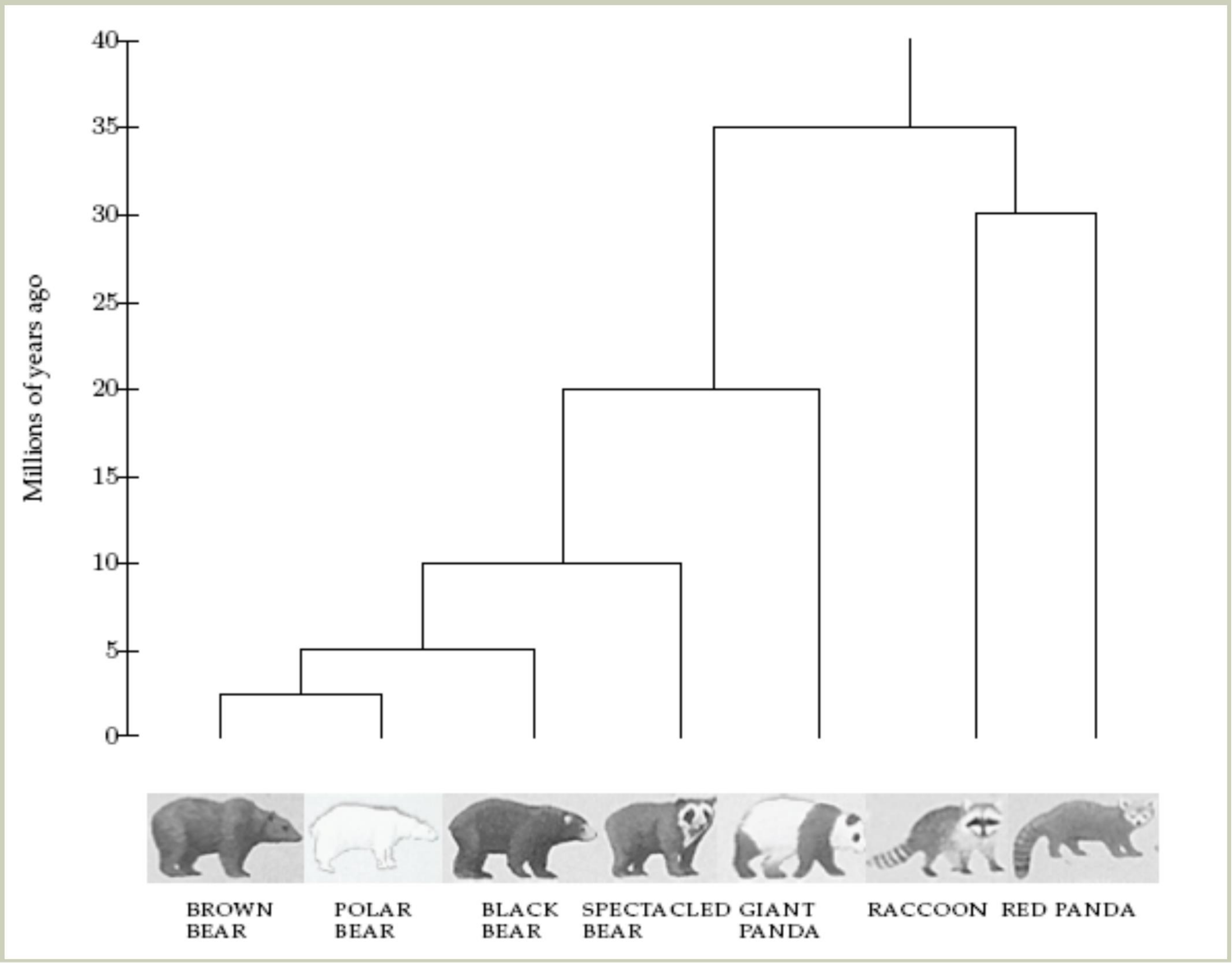
Credit: Pearson Education, Benjamin Cummings

Are pandas more closely related to bears or raccoons?



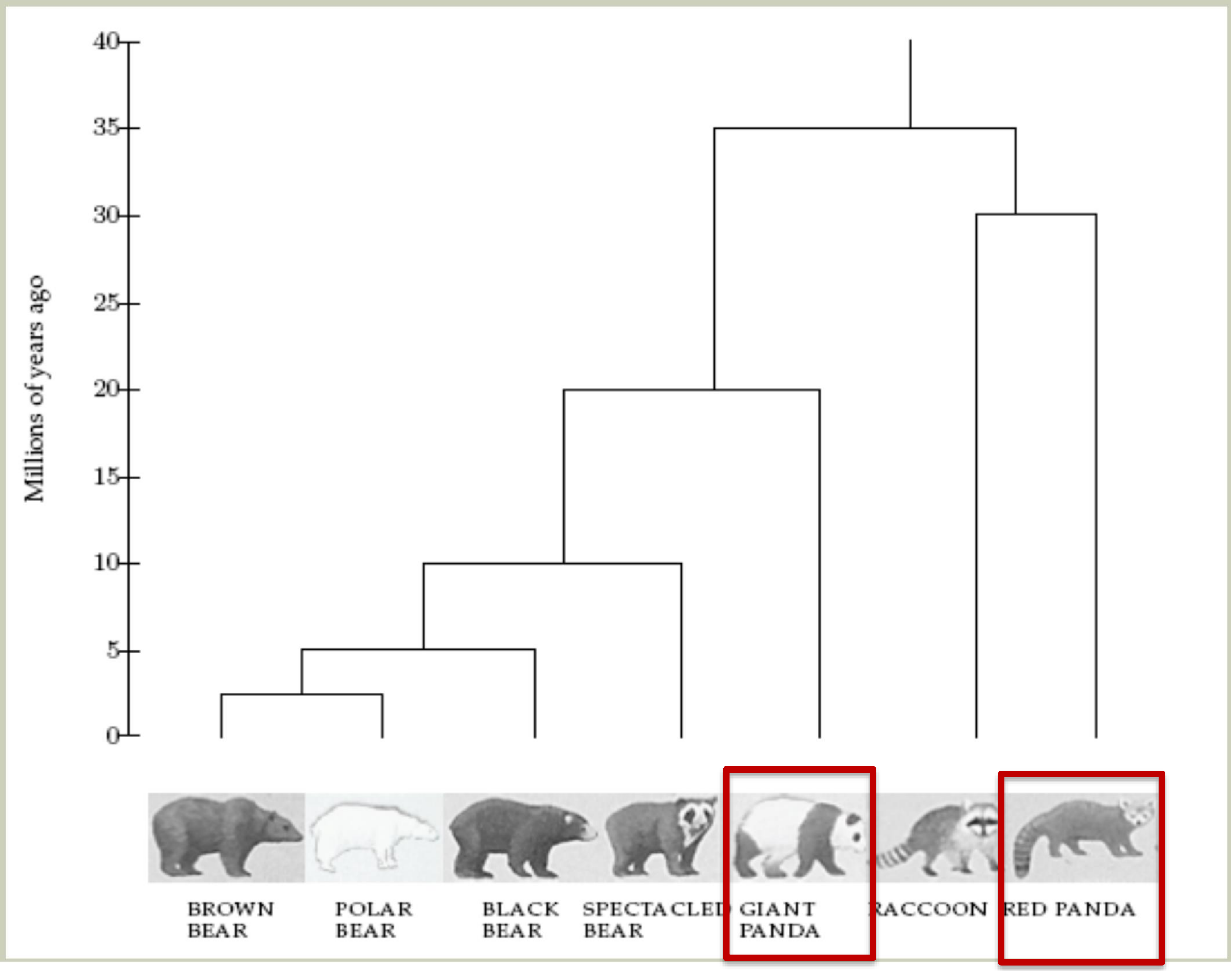
Credit: Ameet Soni

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 23

Handout 23

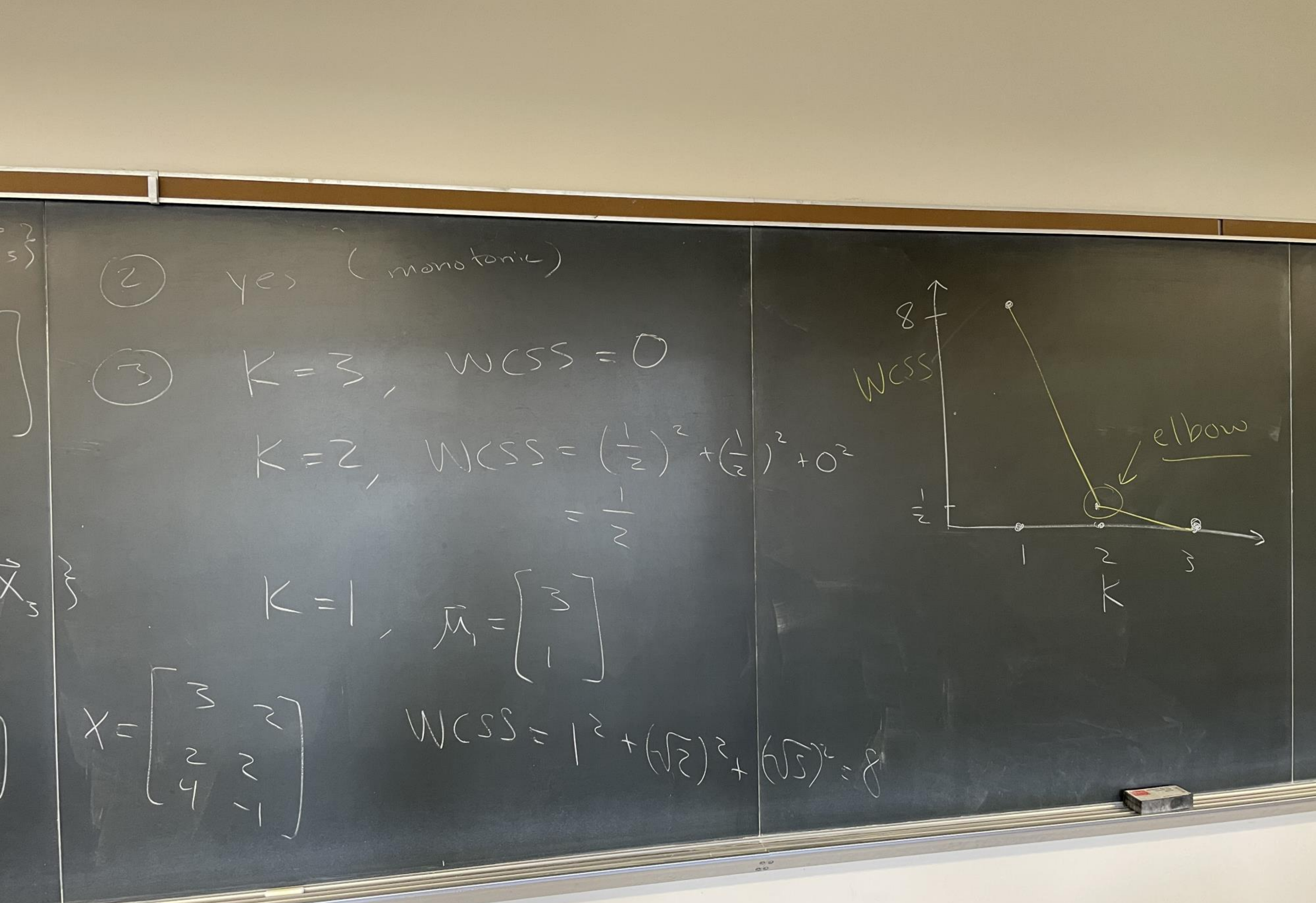
1.

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



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

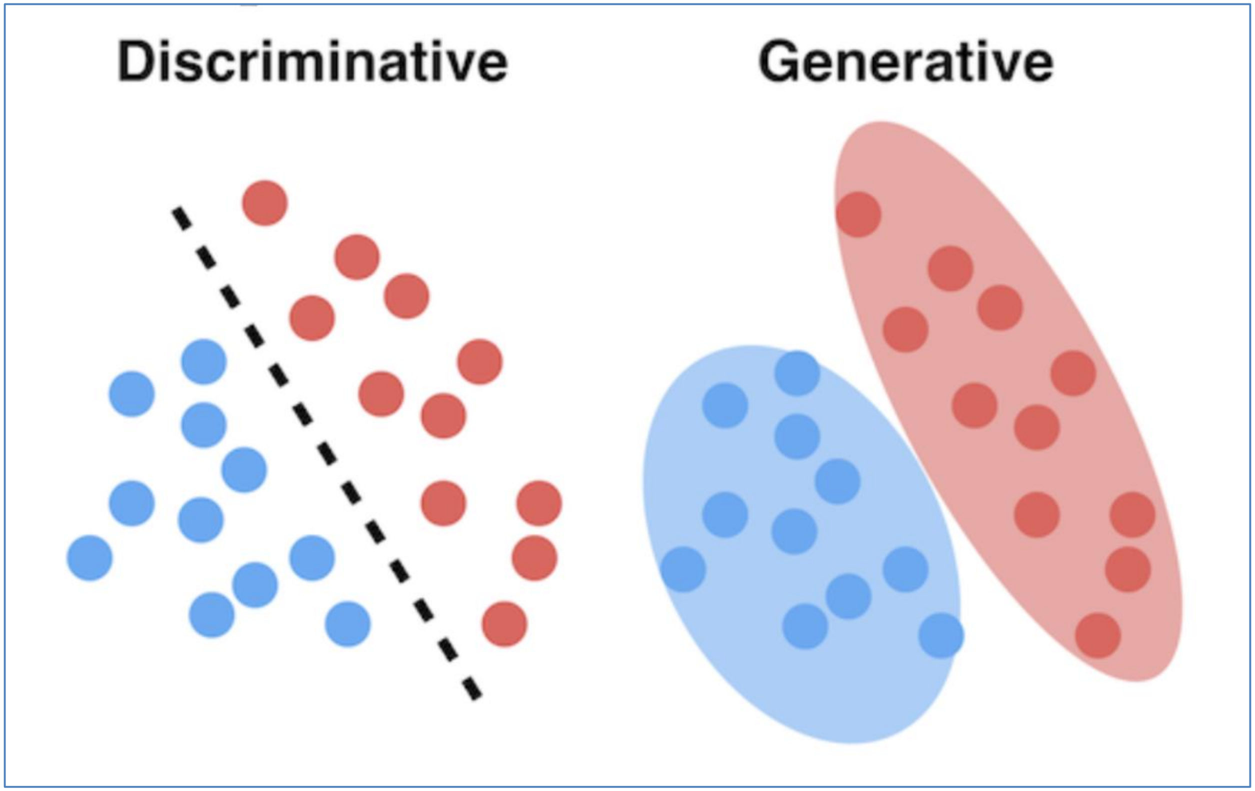


Figure: Ameet Soni

Gaussian Mixture Models (GMMs)

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

- Maximize likelihood:

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

Model parameters

Gaussian Mixture Models (GMMs)

- Initialization step: for each cluster
  - Probability $\pi_k = 1/K$ (uniform prior)
  - Mean $\vec{\mu}_k$ = choose random point
  - Variance σ_k^2 = sample variance

- E-step: "soft" assignment

$$w_{ik} = p(k|\vec{x}_i) = \frac{p(k)p(\vec{x}_i|k)}{p(\vec{x}_i)} = \frac{\pi_k N(\vec{x}_i|\vec{\mu}_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j N(\vec{x}_i|\vec{\mu}_j, \sigma_j^2)}$$

probability that $\vec{x}_i$
came from cluster k

Gaussian Mixture Models (GMMs)

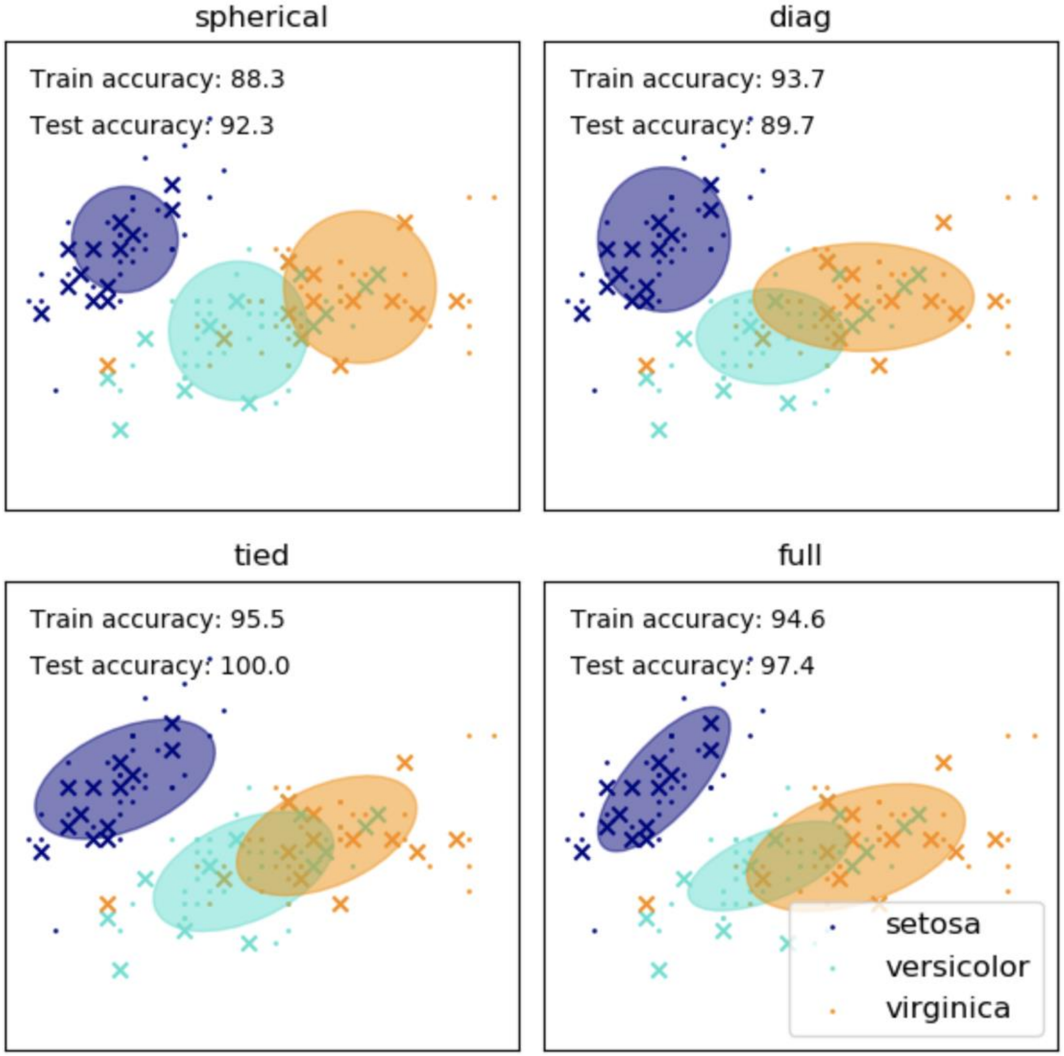
- M-step: parameter update

$N_k = \sum_{i=1}^{n} w_{ik}$ (# of points assigned to cluster k)

  - $\pi_k = \frac{N_k}{n}$
  - $\vec{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} w_{ik}\, \vec{x}_i$
  - $\sigma_k^2 = \frac{1}{N_k} \sum_{i=1}^{n} w_{ik} \left(\vec{x}_i - \vec{\mu}_k\right)^2$

use updated mean

Example of GMMs with 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