CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- Lab 6 graded
- Presentation schedule up
- Lab today
  - Final project check-ins with all groups
  - Try to come to the same lab as your partner
- Candidate talk at 4:15pm TODAY
  - Tea at 4pm
  - Student lunch Wednesday 12:30-1:30pm

Outline for November 28

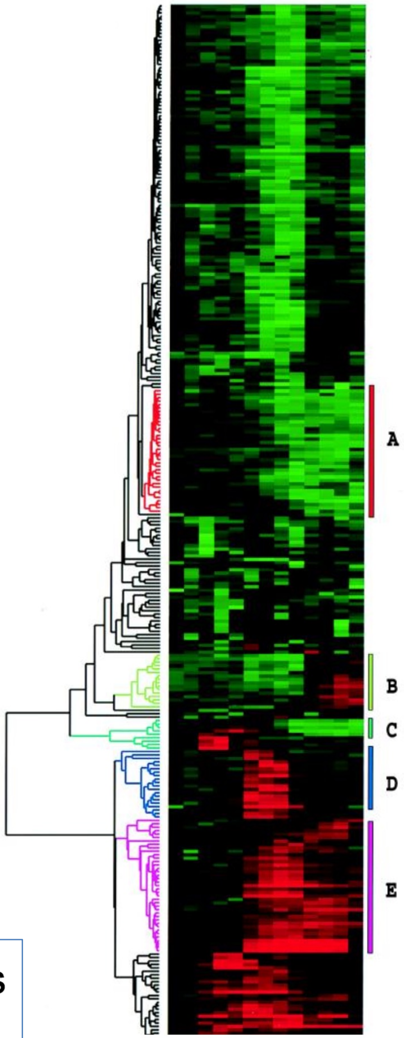
- Clustering overview
- K-means
- Gaussian Mixture Models (GMMs)

Outline for November 28

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

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

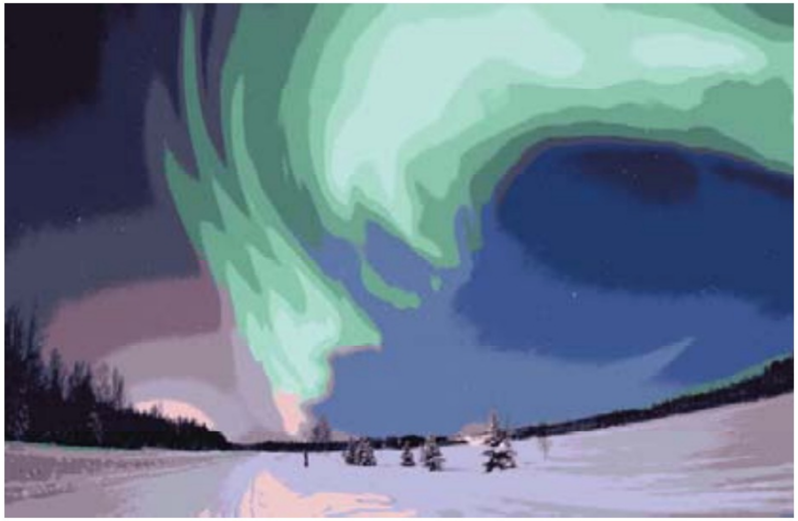


Image: Jessica Wu

Applications of clustering

- Clustering in social graphs

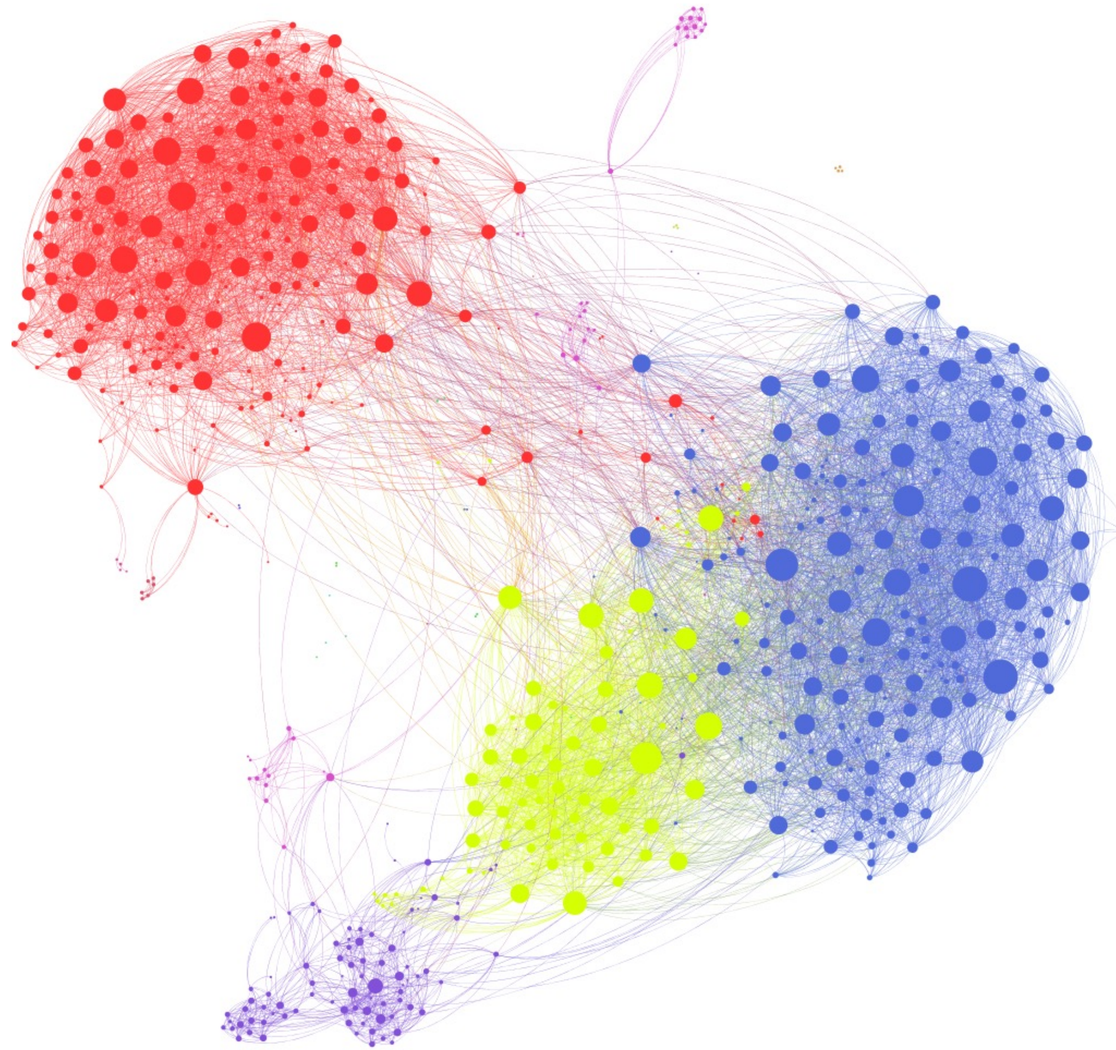
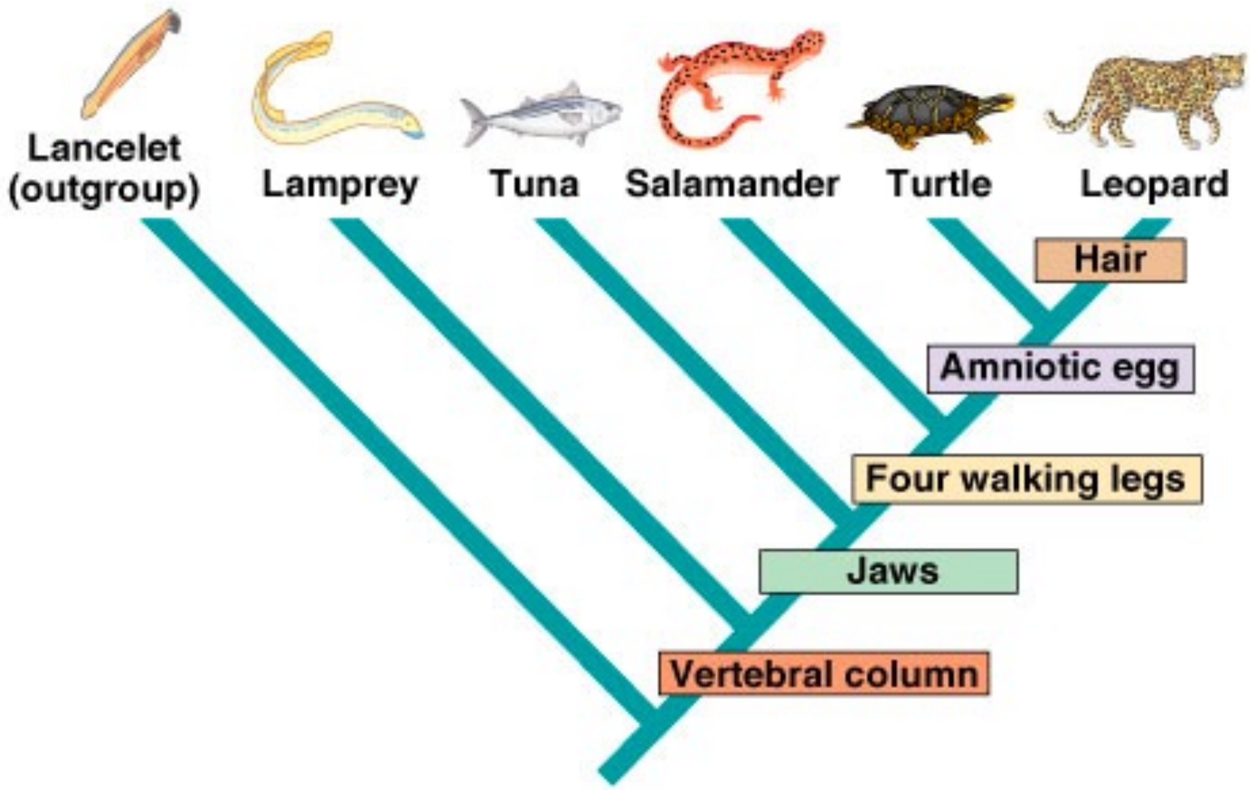


Image: https://griffsgraphs.wordpress.com/2012/07/02/a-facebook-network/

Two main types of clustering

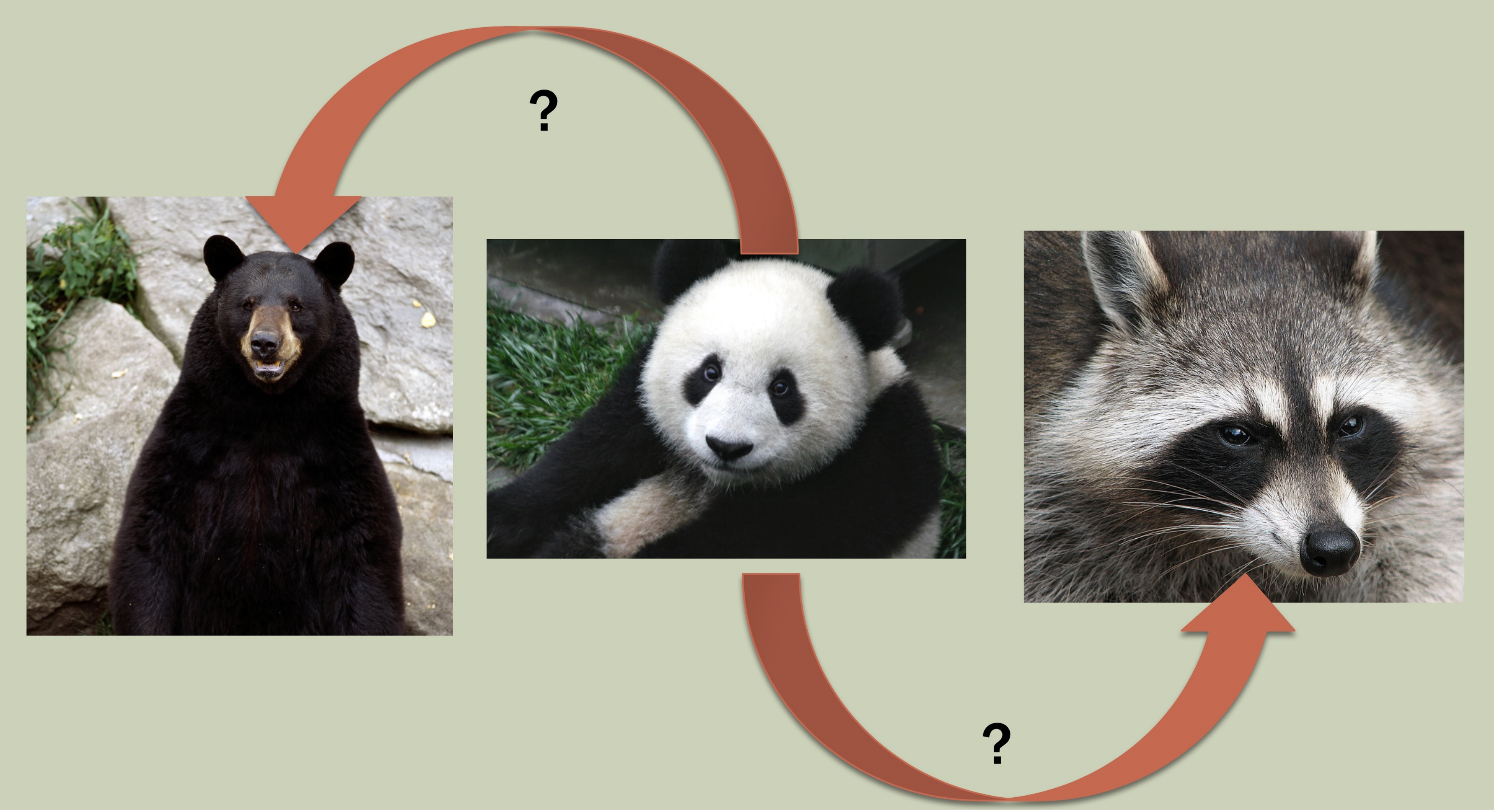
- Flat/Partitional:
  - K-means
  - Gaussian mixture models
- Hierarchical:
  - Agglomerative: bottom-up
  - Divisive: top-down
  - Examples: UPGMA and Neighbor Joining

Hierarchical clustering example: trees



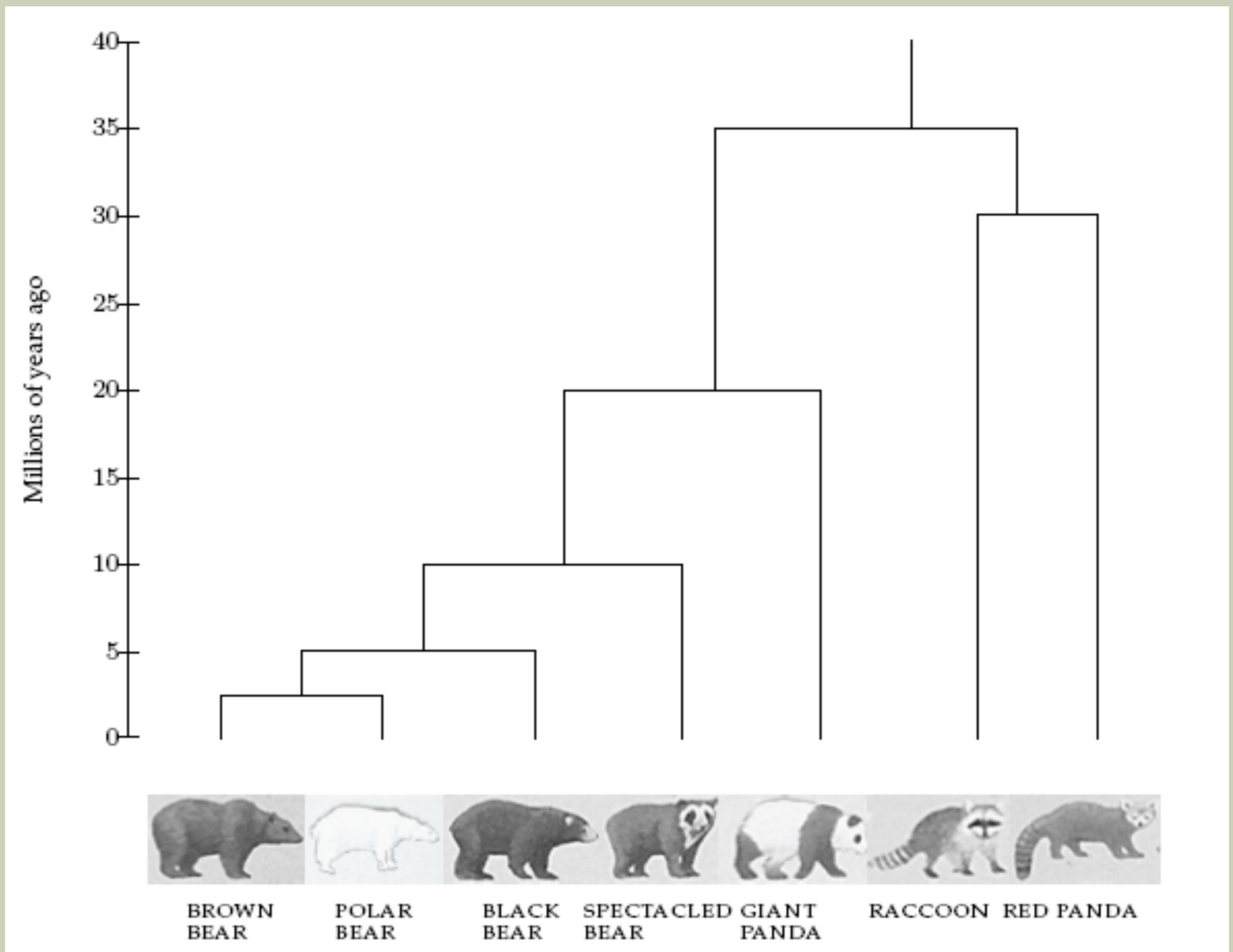
Credit: Pearson Education, Benjamin Cummings

Are pandas more closely related to bears or raccoons?



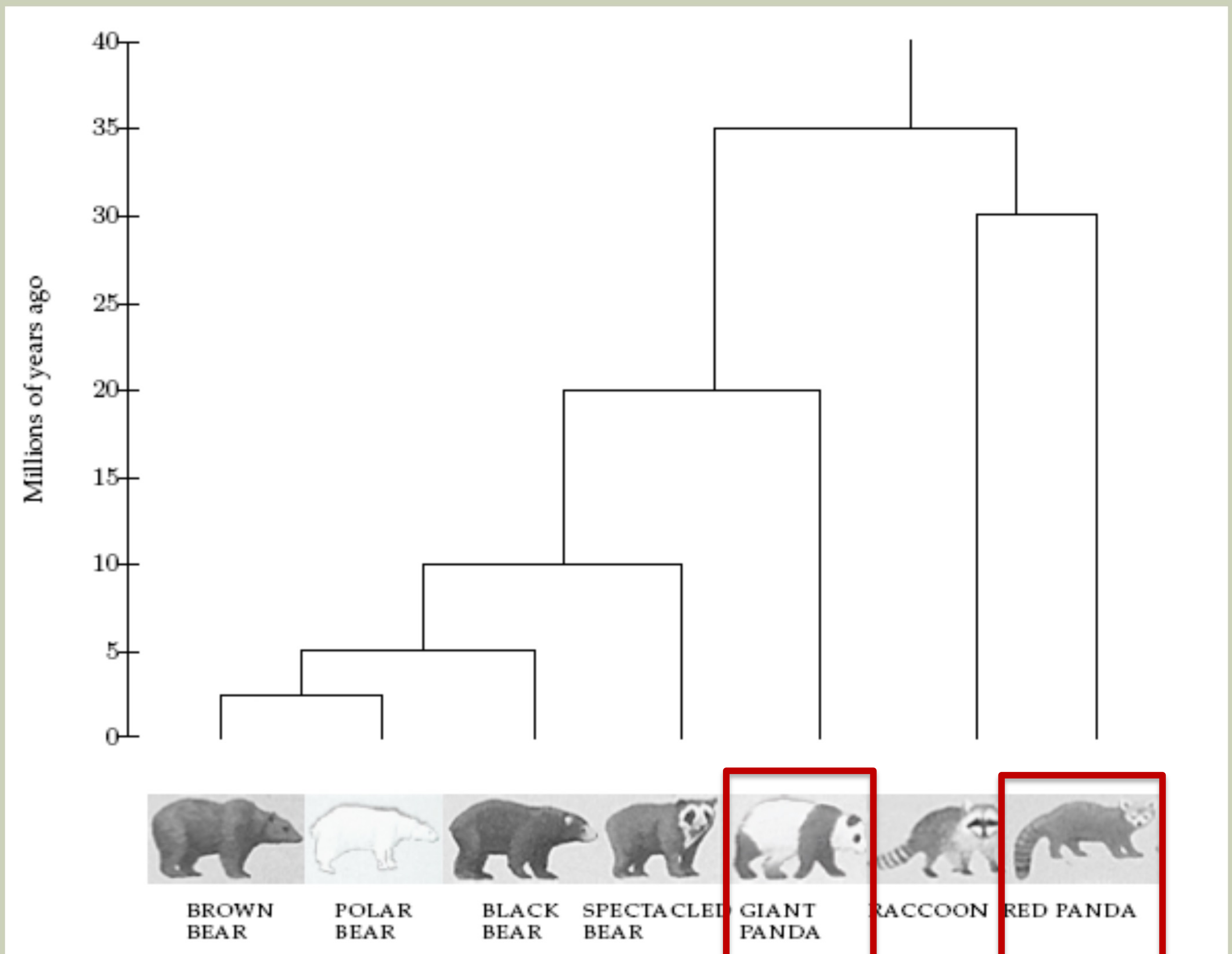
Credit: Ameet Soni

Are pandas more closely related to bears or raccoons?



Credit: Ameet Soni

Are pandas more closely related to bears or raccoons?



Credit: Ameet Soni

Outline for November 28

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

Discriminative vs. Generative

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

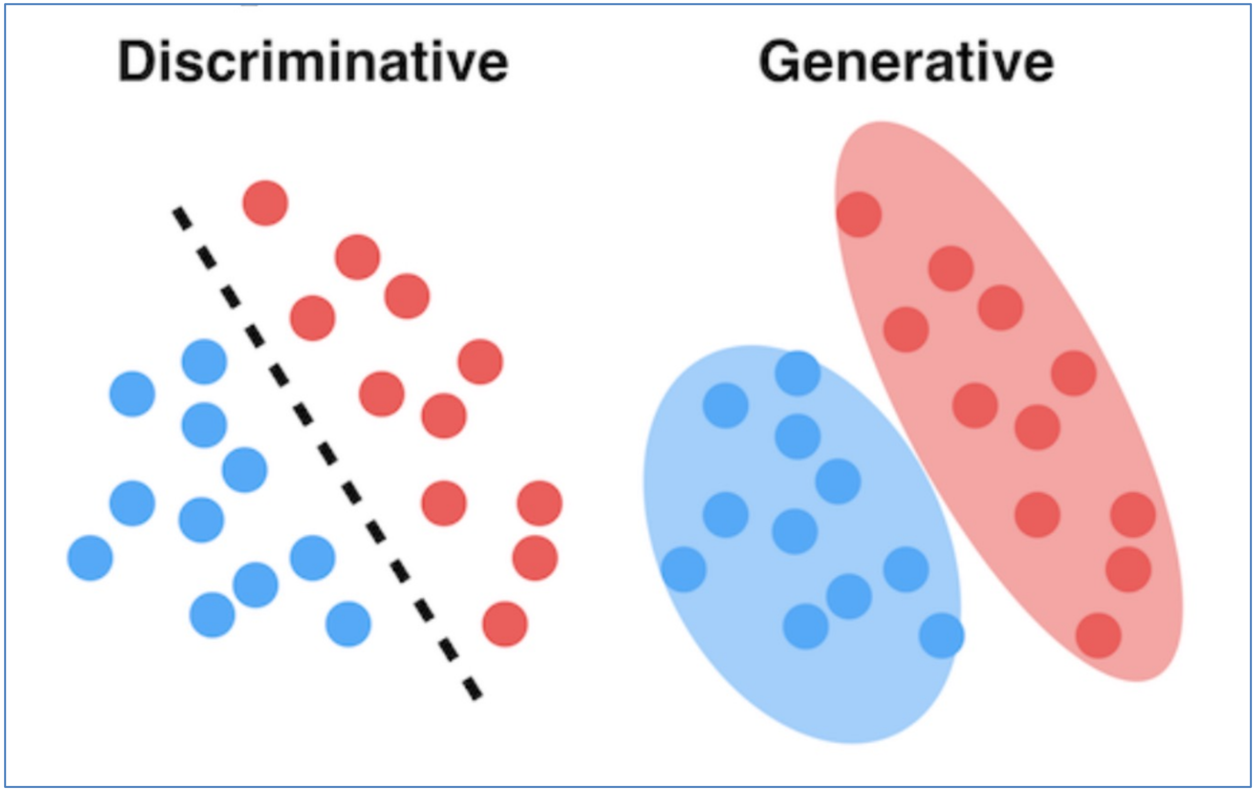


Figure: Ameet Soni

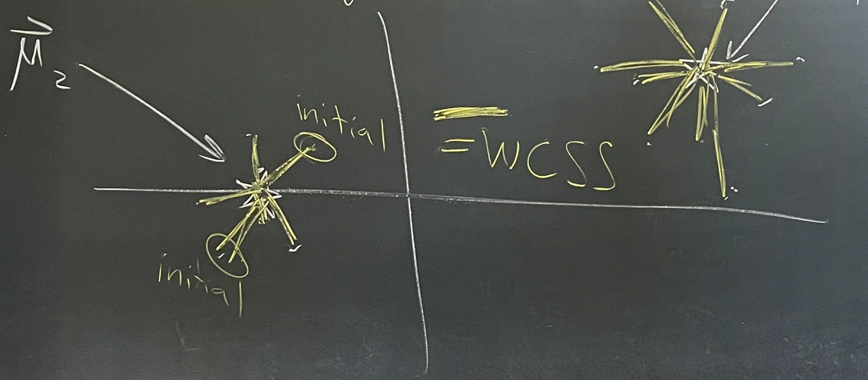
K-means

K clusters

minimize

$$\underbrace{WCSS}_{} = \sum_{k=1}^{K} \sum_{\vec{x}_i \underset{\text{"in"}}{\in} C_k} \|\vec{x}_i - \vec{\mu}_k\|^2$$

within-cluster
sum of squares



EM algorithm

$\boxed{\text{E-step}}$ assignment

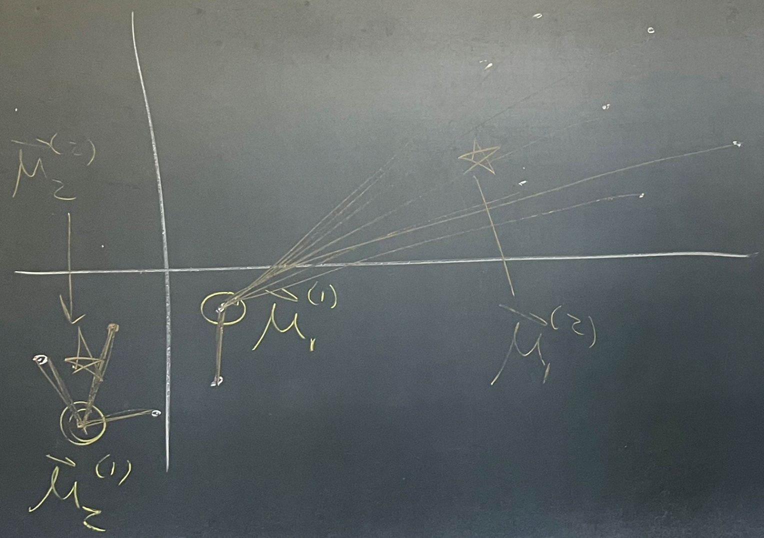
for $\vec{x}_i$, find closest cluster
mean $\Rightarrow$ $\vec{x}_i \in C_k^{(t)}$

$\boxed{\text{M-step}}$ parameter update

$$\mu_k^{(t+1)} = \frac{1}{\underbrace{|C_k^{(t)}|}_{\text{size of cluster } k}} \sum_{\vec{x}_i \in C_k^{(t)}} \vec{x}_i$$

initialization

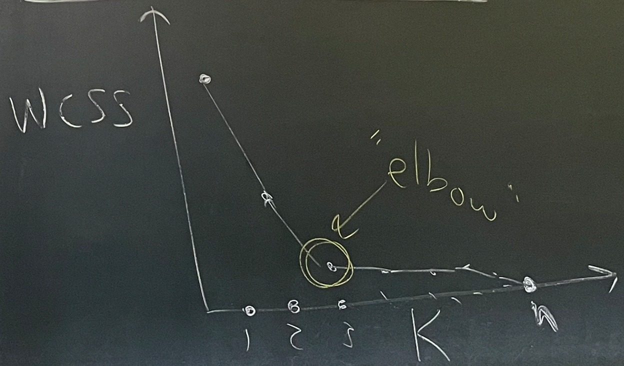
$\vec{\mu}$'s? choose randomly from our data points



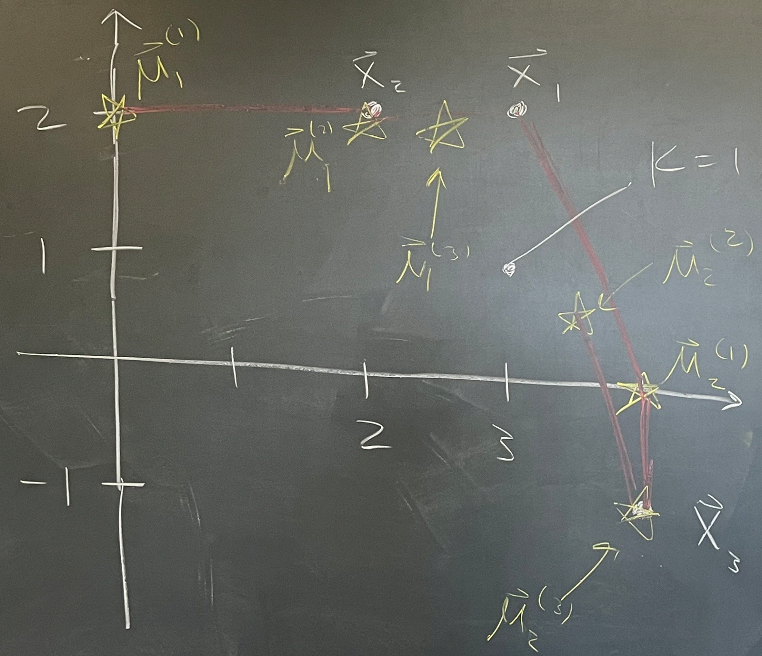
Stopping criteria

- no change in cluster membership
- max # iters exceeded (T)
- configuration you've see before

how to choose K?



Handout 22



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

(b) $\boxed{\text{M-step}}$ $\vec{\mu}_1^{(2)} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, $\vec{\mu}_2^{(2)} = \begin{bmatrix} 3.5 \\ 0.5 \end{bmatrix}$

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

$\boxed{\text{M-step}}$

$\mu_1^{(3)} = \begin{bmatrix} 2.5 \\ 2 \end{bmatrix}$, $\vec{\mu}_2^{(3)} = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$

(2) yes (monotonic)

(3) $K=3, \quad WCSS = 0$

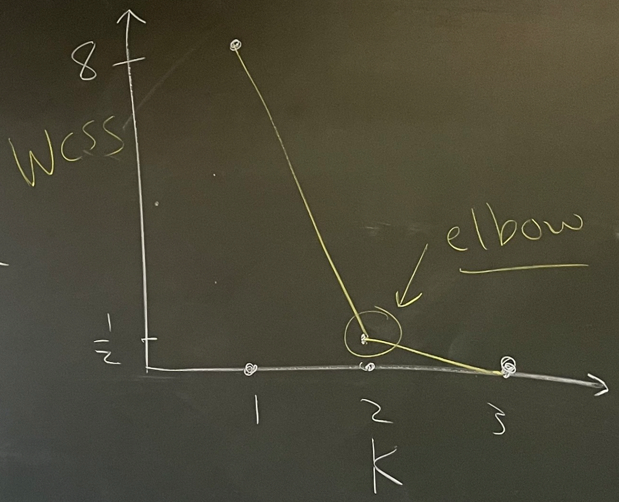
$K=2, \quad WCSS = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + 0^2$

$= \frac{1}{2}$

$K=1, \quad \bar{\mu}_1 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

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

$WCSS = 1^2 + (\sqrt{2})^2 + (\sqrt{5})^2 = 8$



Outline for November 28

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

Problems with K-means

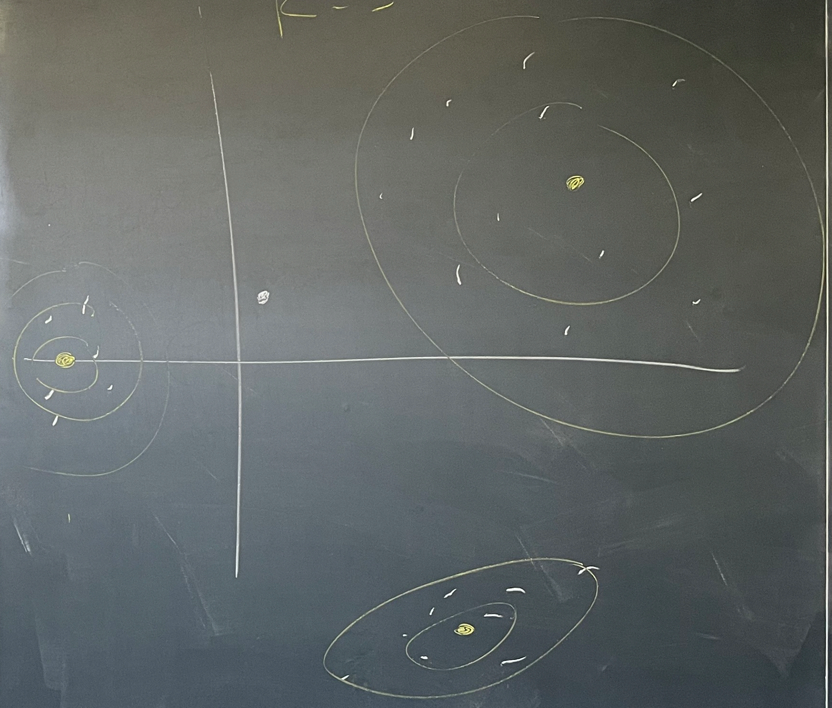
* not generative (could not create a new data-point)

* does not account for different cluster sizes & variances

  $\uparrow$ π_k

* does not allow points to belong to multiple clusters

K=3



Gaussian Mixture Models (GMM)

Likelihood ← maximize

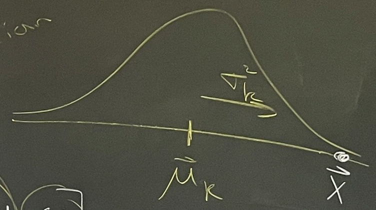
$$p(\vec{x}) = \sum_{k=1}^{K} p(\vec{x}, z=k) = \sum_{k=1}^{K} \pi_k \underbrace{p(\vec{x} \mid z=k)}_{\text{Gaussian with } \vec{\mu}_k,\ \sigma_k^2}$$

one pt

cluster membership

BAYES

prior over cluster sizes



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

all data

don't know

know

w_{11} w_{12} w_{13}

$$W = \begin{bmatrix} 0.2 & 0.1 & 0.7 \\ & & \\ & & \end{bmatrix}$$

π_k

$n \times K$

EM for GMM

Initialization

- $\pi_k = \dfrac{1}{K}$ (prior) (uniform)
- $\vec{\mu}_k$ = choose random points
- σ_k^2 = sample variance of all points closest to each mean

E-step

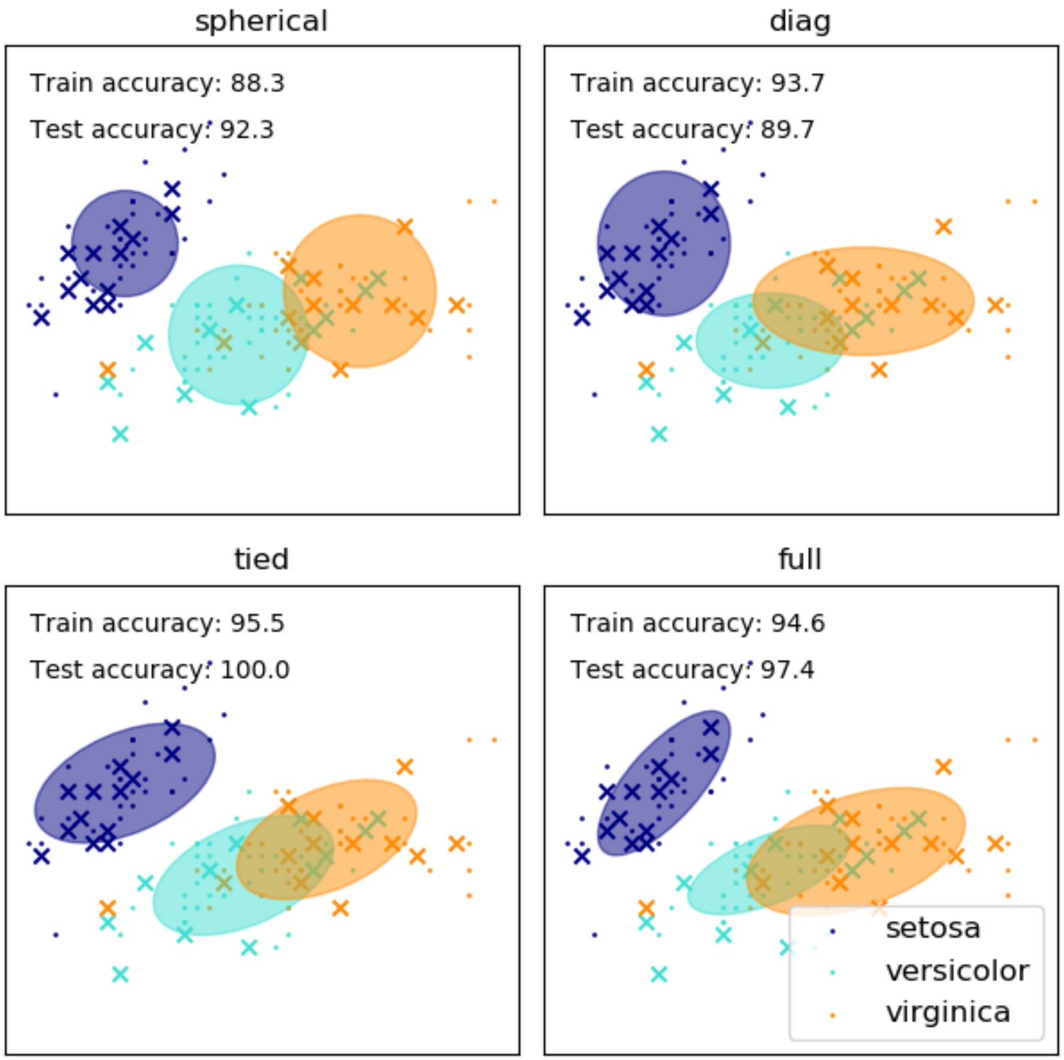
"Soft" assignment

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

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

$$= \frac{\pi_k N(\vec{x}_i, \vec{\mu}_k, \sigma_k^2)}{\sum_{k'=1}^{K} \pi_{k'} N(\vec{x}_i, \vec{\mu}_{k'}, \sigma_{k'}^2)}$$

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