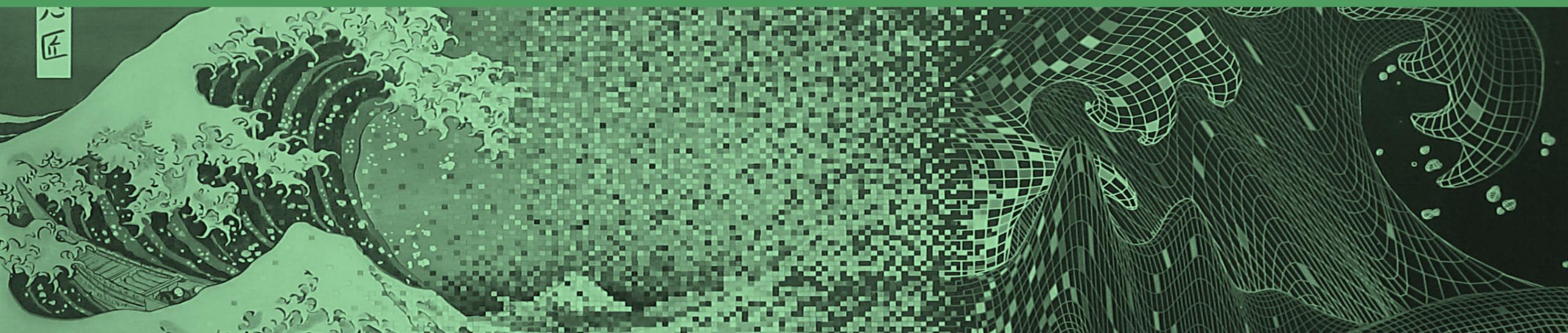
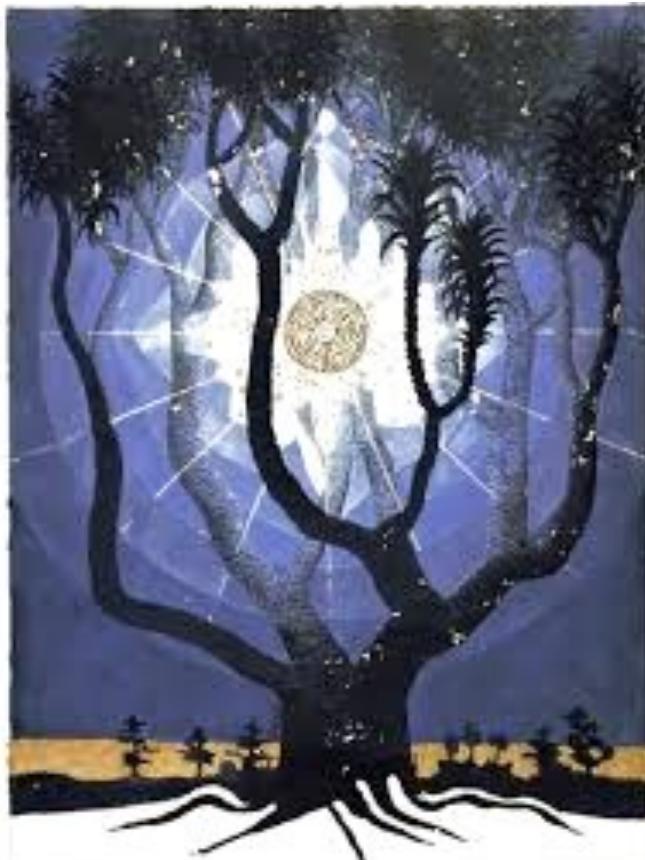


# Clustering

Clustering approaches and evaluation



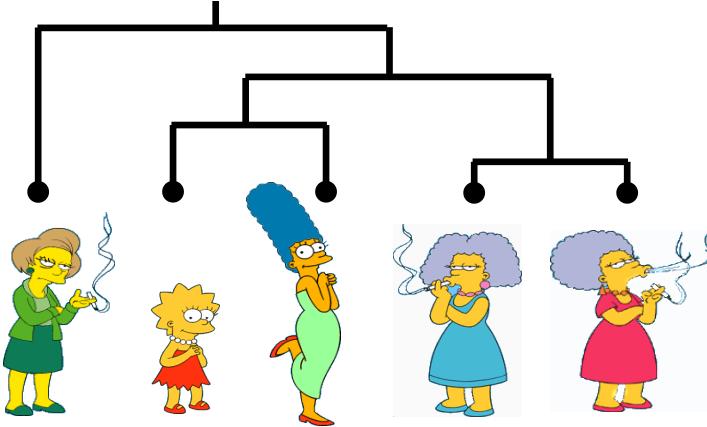
# Outline



- Introduction to clustering
- Partition-based clustering:  $k$ -means
- Model-based clustering: EM
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# Motivation

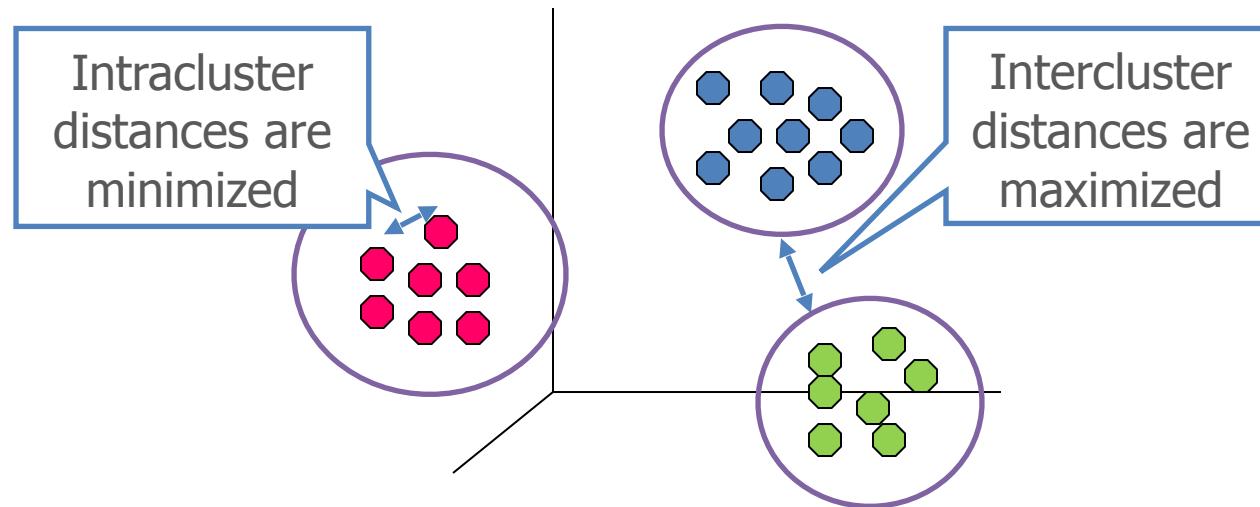
- **Patients** with a shared clinical condition:  
how to **understand disease**?
  - stratified diagnostics and therapeutics
  - cancer types, dementia progression, risk groups
- **Customers**: how to segment their profile for **personalized marketing**?
- **Webpages**, shopping products, **media**, **documents**:  
how to categorize them for **recommendations**?
- **Genes**, proteins and metabolites with different expression and concentration profile: how to understand their correlated behavior (biological functions)?
- **Students**, researchers, professors: how to improve science and **education**?



# Clustering

**Cluster:** group of data observations

**Cluster analysis:** group observations into clusters according to their (dis)similarity:  
observations in the same cluster are more similar than those in different clusters



Notes on notation: when all attributes are numeric: data observations = data point

# ML applications

- Stand-alone tool to get **insight into data structures**
- **Unsupervised data labeling**
- **Preprocessing step** to facilitate other tasks
  - **prediction**: e.g. learn one predictor per cluster to mitigate challenges arising from data heterogeneity
  - **outlier analysis**: identify observations that deviate from expectations
- **Data compression**
  - reduce the number of data points: centers of the clusters seen as the reference/prototype observations
  - highly common when handling very large data (e.g. webpage retrieval, e-commerce)

# Clustering: distance and approach

Two major factors impact solutions: ***distance + approach***

- **Distance metrics** depend on the:

- input variables:
  - *numeric* distances, e.g. Euclidean 
$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{j=1}^m (a_j - b_j)^2}$$
  - *nominal* distances, e.g. Hamming
  - *ordinal* encodings
  - *non-iid attributes*
- data structure: multivariate, time series, image, spatiotemporal data, events...

- **Clustering approach:** partitioning, hierarchical, density-based, model-based

Recall the distances introduced for kNN!

# Distances in clustering

- Given a cluster of observations  $C_j$

- **mean center** of a cluster:  $\mathbf{c}_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$
  - **cluster**:  $C_j = \{\mathbf{x} \mid d(\mathbf{x}, \mathbf{c}_j) = \min_i d(\mathbf{x}, \mathbf{c}_i)\}$

- Distances can be applied between:

- two observations  $d(\mathbf{x}_i, \mathbf{x}_j)$
  - one observation and a cluster  $d(\mathbf{x}_i, \mathbf{c}_j)$
  - two clusters  $d(\mathbf{c}_i, \mathbf{c}_j)$

- (Squared) **error of clustering solution**:  $E = \sum_{k=1}^K \sum_{\mathbf{x} \in C_k} d(\mathbf{x}, \mathbf{c}_k)^2$

# Approaches

## Partitioning: ⇐

- Create partitions and iteratively update them (e.g.  $k$ -means,  $k$ -modes,  $k$ -medoids)

## Hierarchical:

- Create hierarchical decomposition of data points

## Density-based:

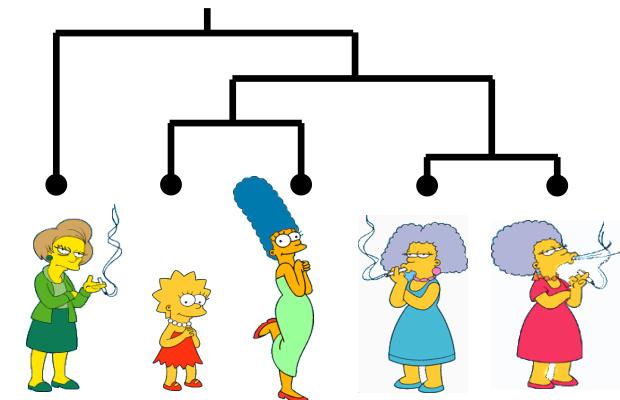
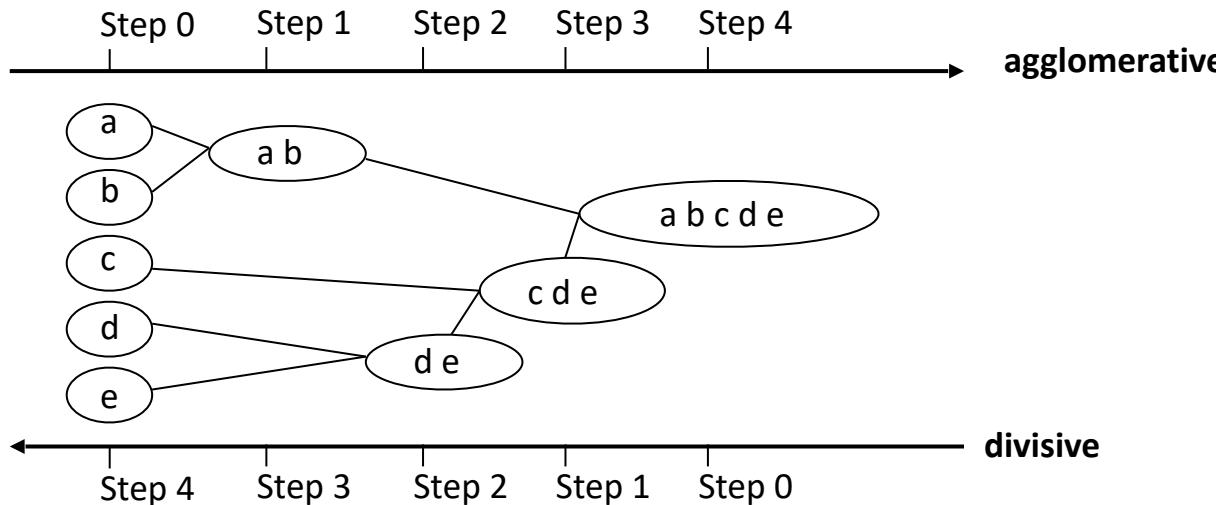
- Group points based on connectivity and density functions

## Model-based: ⇐

- Data are seen as a mixture of distributions (e.g. EM)

# Tree/hierarchical clustering

- Tree clustering algorithm allow us to reveal the internal similarities of a given pattern set
  - similarities structured hierarchically
  - for  $n$  patterns this algorithm generates a sequence of 1 to  $n$  clusters
- Tree (dendrogram) can be structured bottom up or top down

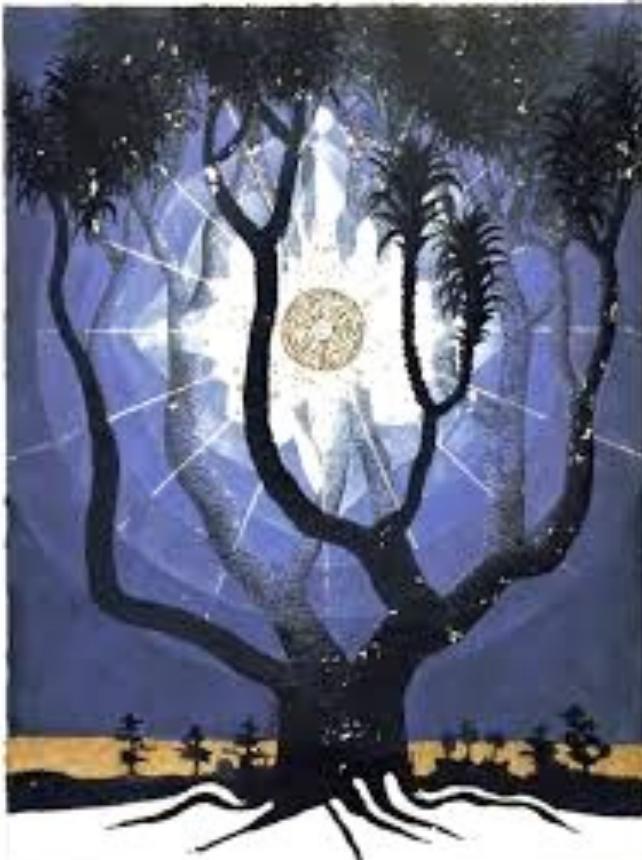


# Goal for today

- Goal

- given a set of  $n$  observations (sample)  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$
- partition the data set into  $K$  clusters (we can assume for now that  $K$  is given) where  $K \ll n$
- Firstly, we approach clustering with a non-probabilistic technique called  **$k$ -means** [Lloyd, 1982]
- Then, we introduce the latent variable view of mixture distributions
  - discrete latent variables defining assignments of data points to specific components of the mixture
  - a general technique to find maximum likelihood estimators in latent variable models is the **expectation-maximization (EM)** algorithm

# Outline



- Introduction to clustering
- **Partition-based clustering:  $k$ -means**
- Model-based clustering: EM
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# Centers and distances

- Cluster as a group of data points with...
  - small inter-point distances when compared with the distances to points outside of the cluster
- The cluster centers (also called **centroids**) represent the compressed data set
  - **mean center** of a cluster:  $c_k = \frac{1}{|C_k|} \sum_{x \in C_k} x$
- Data points are grouped into clusters according to a distance function, applicable to...
  - two observations  $d(x_i, x_j)$
  - one observation and a cluster  $d(x_i, c_k)$

# Clustering objective function

- In this context, a cluster is defined as

$$C_k = \{x \mid d(x, c_k) = \min_i d(x, c_i)\}$$

- We can then define an objective function to assess the error of a clustering solution
  - squared error

$$E = \sum_{k=1}^K \sum_{x \in C_k} \|x - c_k\|^2 = \sum_{k=1}^K \sum_{x \in C_k} d(x, c_k)^2$$

# Partitioning algorithms

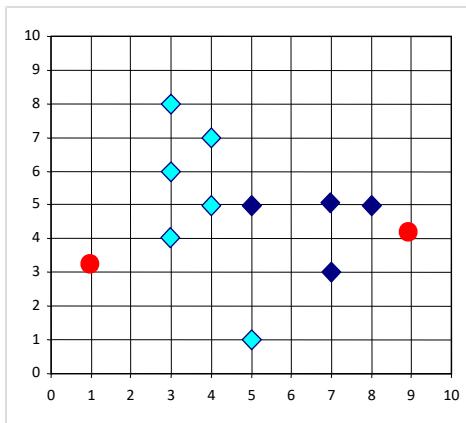
For  $k$  clusters:

1. partition objects into  $k$  nonempty subsets
2. compute the centroid  $c_k$  of each subset
  - centroid is the center of mass: e.g., mean or median centers
3. reassign each observation to the cluster with the nearest centroid
4. go to step 2, and stop when:
  - i)* assignment does not change, *ii)*  $|E^{new} - E^{old}| < \varepsilon$ , or *iii)* max iterations is reached

## Variants

- centroid calculus
- selection of the initial seeds
- adjustments for **batches** of observations (instead of all) for very large datasets

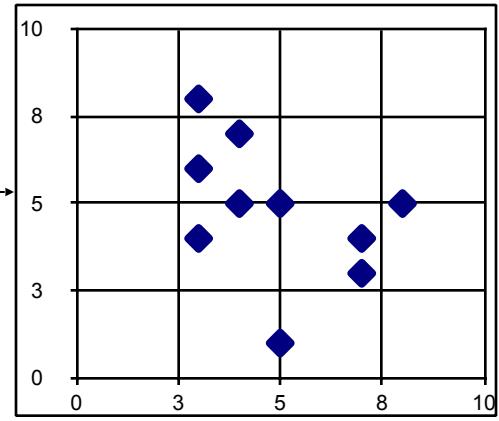
# *k*-means



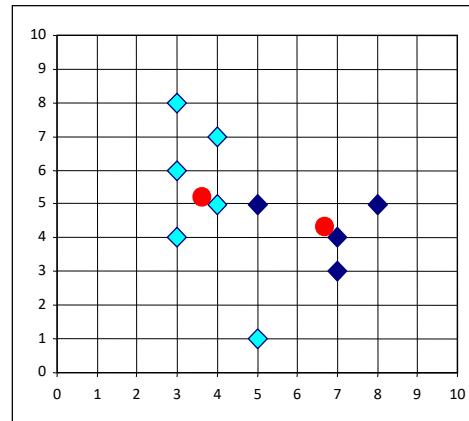
$k=2$

Arbitrarily choose  
k object as initial  
cluster center

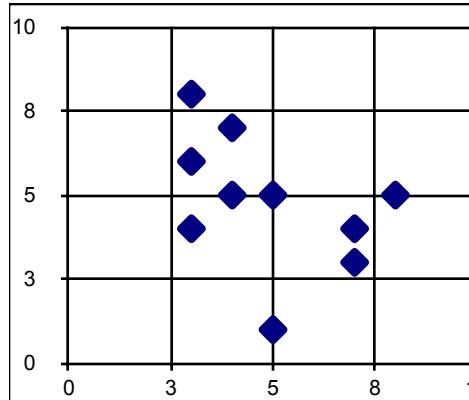
Assign  
each  
object  
to the  
most  
similar  
center



Update  
the  
cluster  
means



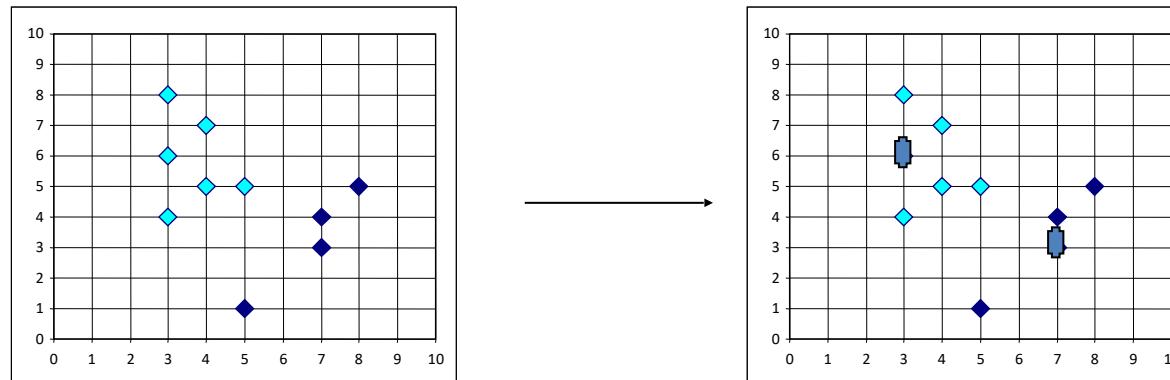
reassign



Update  
the  
cluster  
means

# *k*-medoids

1. numeric data: centroid as the mean or median
  2. categoric data: centroid as the mode – *k*-modes [Huang'98]
    - frequency-based procedure to update modes of clusters
  3. mixed data: centroid combining mean and modes (*k*-prototype)
- ***k*-medoids**: the most centrally located observation in a cluster is the centroid
    - observation with minimum average distance to all observations in the cluster
  - What is the algorithm most robust to outliers: *k*-means or *k*-medoids?

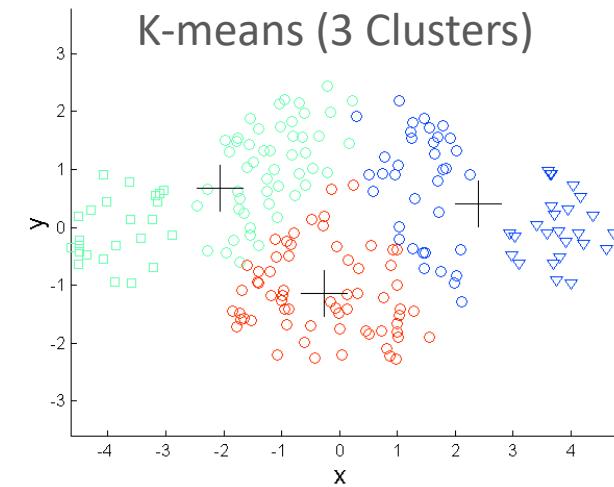
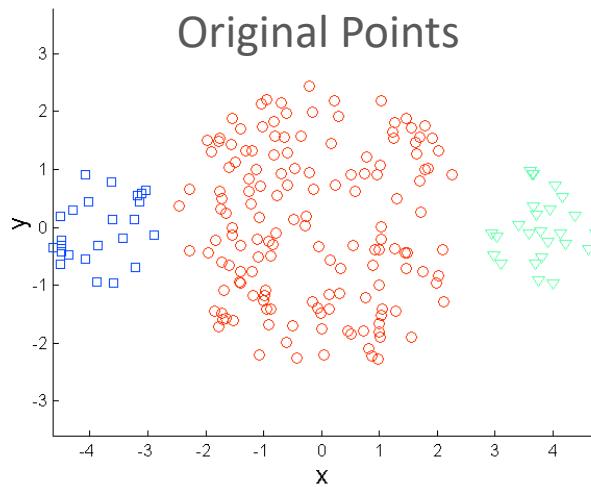


# *k*-means: challenges

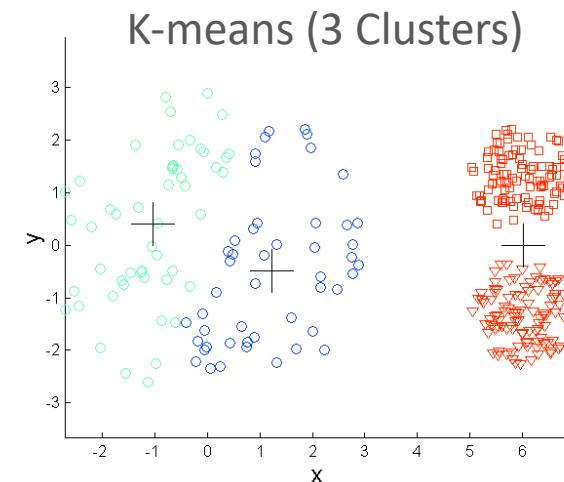
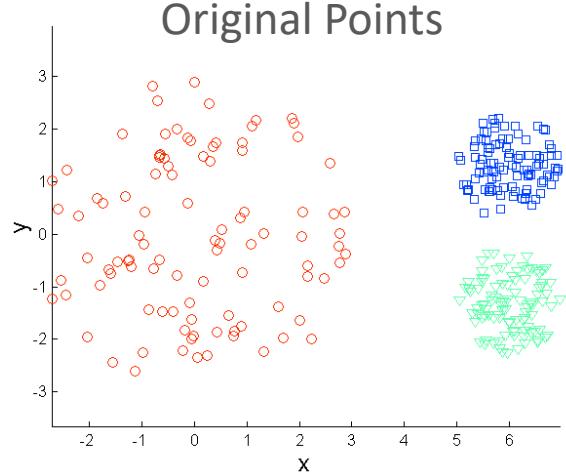
- **Efficiency:**  $O(tkn)$   $n=\#$ observations,  $k=\#$ clusters,  $t=\#$ iterations, usually  $k, t \ll n$
- **Problems**
  - dependent on initialization
  - sensitive to outliers
  - sensitive noisy data
    - noise can substantially distort centroids
    - not suitable to discover clusters with *non-convex shapes*
      - deal only with clusters with spherical symmetrical point distribution
    - need to specify  $k$ , the *number* of clusters, in advance
    - convergence?

# Limitations of $k$ -means

Different sizes

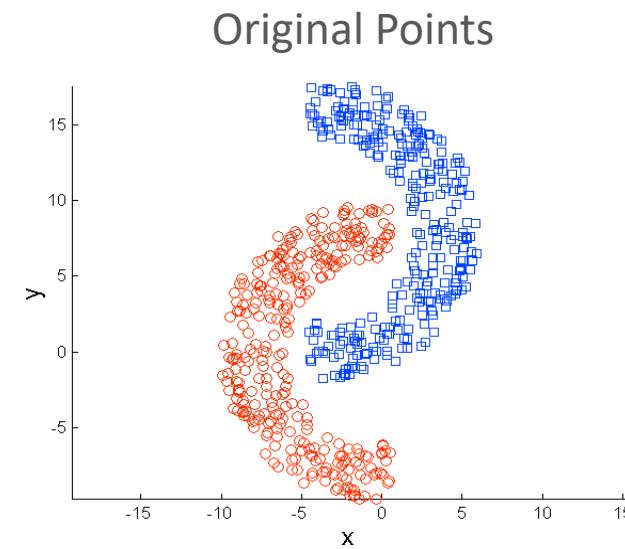


Different densities

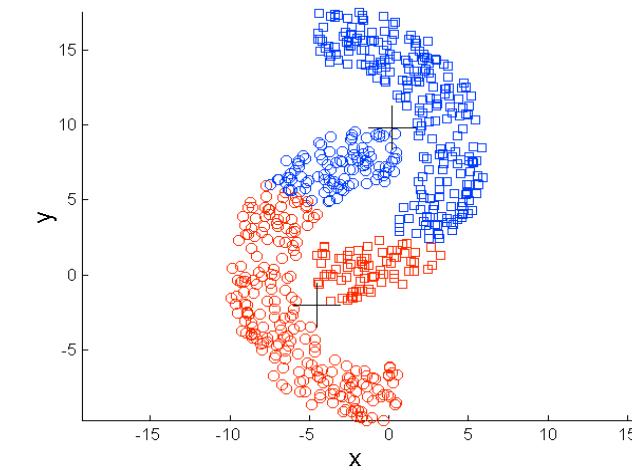


# Limitations of $k$ -means

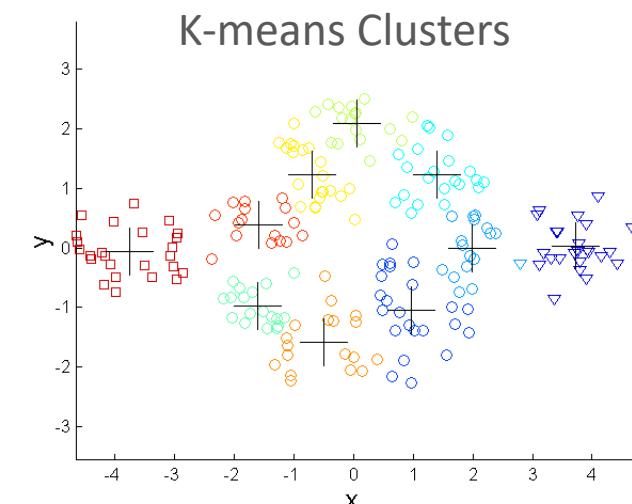
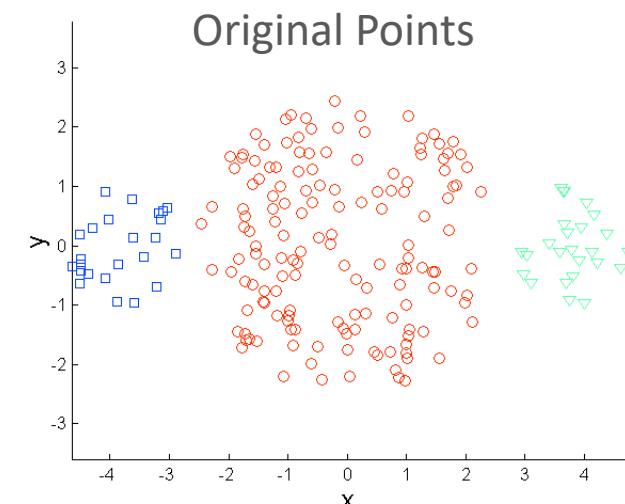
Non-globular shapes



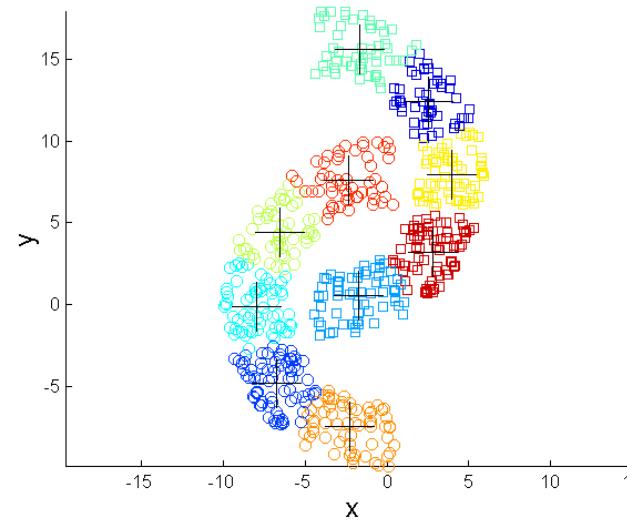
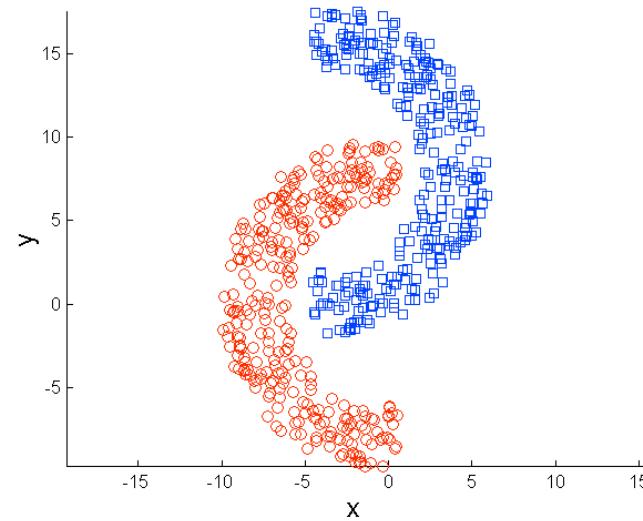
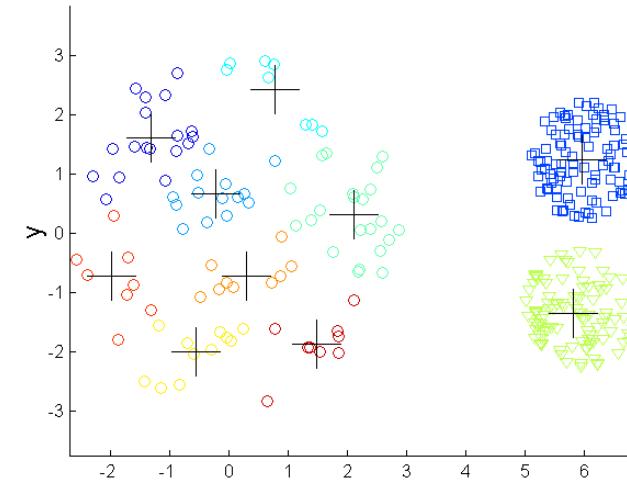
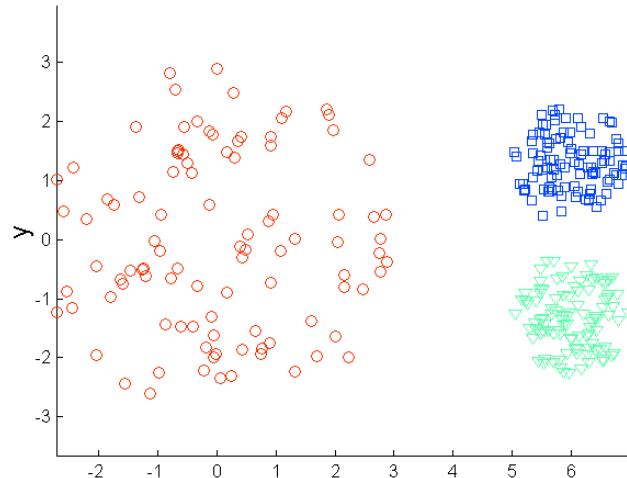
K-means (2 Clusters)



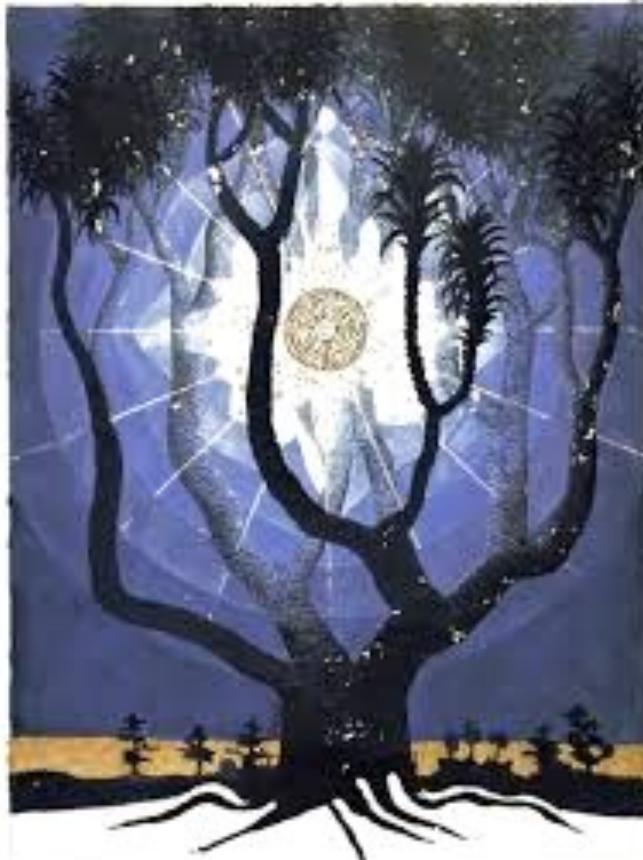
Fixed number of clusters



# Overcoming $k$ -means limitations



# Outline



- Introduction to clustering
- Partition-based clustering:  $k$ -means
- **Model-based clustering: EM**
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# EM clustering

- Expectation maximization (EM) clustering
  - cluster described by a multivariate distribution
  - observations have a probability of belonging to each cluster
    - **soft** clustering (in contrast with hard clustering)
- Recovering essentials
  - **posterior** probability (after the evidence is obtained):  $P(\text{cluster} = k \mid \mathbf{x})$
  - **Baye's rule**  $P(\text{cluster} = k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \text{cluster} = k)P(\text{cluster} = k)}{P(\mathbf{x})}$
  - **covariance**: direction and strength on how two variables  $y_i$  and  $y_j$  vary
  - **univariate Gaussian** probability density function,  $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$

# Multivariate Gaussian

- A Gaussian distribution in a  $m$ -dimensional space is defined as:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{u})^T \Sigma^{-1} (\mathbf{x} - \mathbf{u})}$$

- where  $\mathbf{u}$  is the  $m$ -dimensional mean vector
- $\Sigma$  is the  $m \times m$  covariance matrix
- $|\Sigma|$  is the determinant of  $\Sigma$

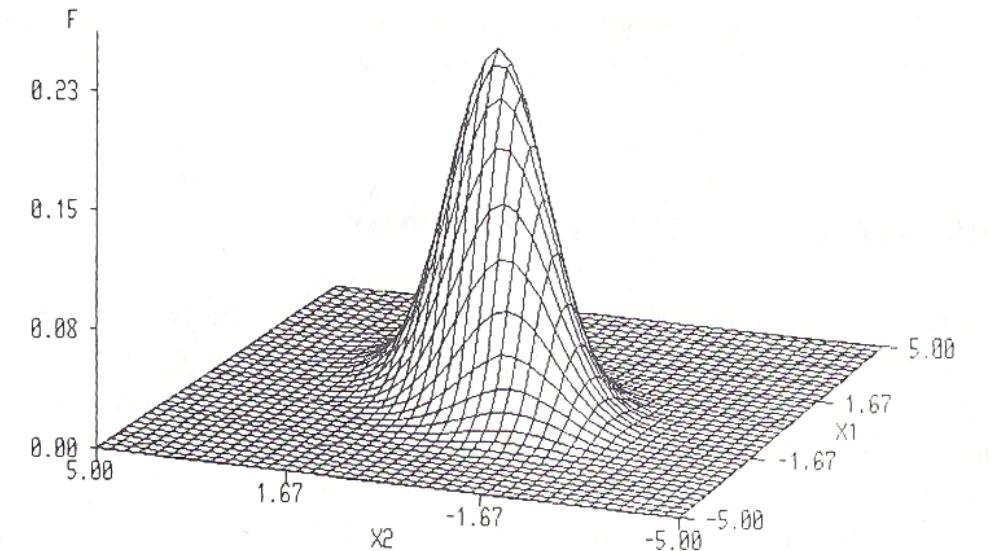
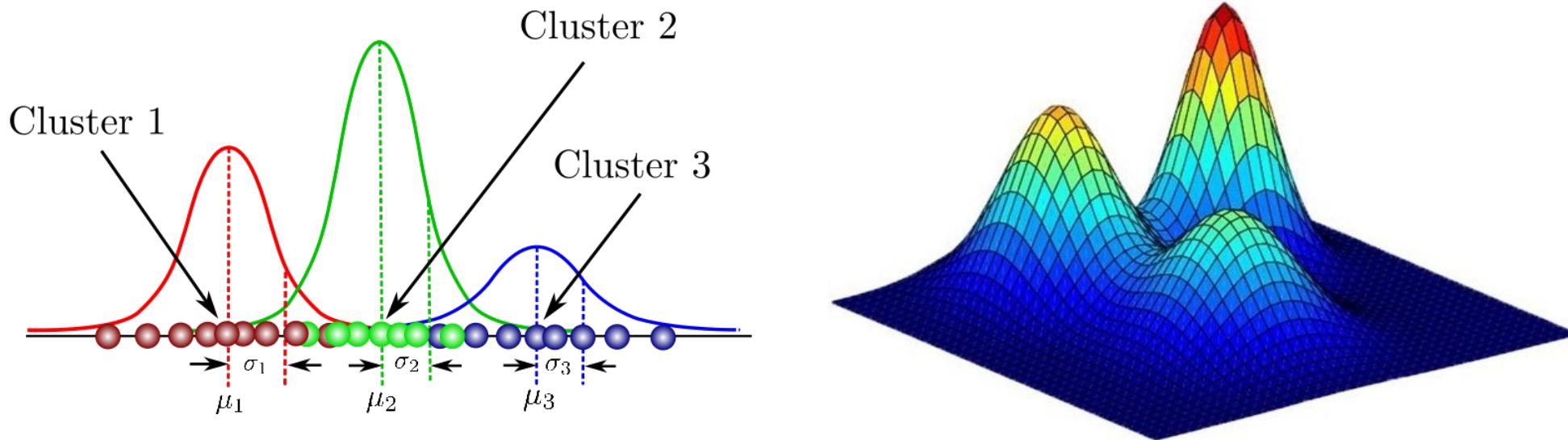


FIG. 7.3. The bivariate normal distribution.

# Gaussian mixture

- What kind of probability distribution might have generated the data?
  - clustering presumes that the data are generated from a mixture of distributions



$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \cdot N(\mathbf{x} | \mathbf{u}_k, \Sigma_k)$$

with

$$\sum_{k=1}^K \pi_k = 1$$

# Gaussian of cluster $c_k$

- Univariate density

- cluster  $c_k$  generates  $x$

$$p(x | c_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-u_k}{\sigma_k}\right)^2}$$

- where  $u_k$  and  $\sigma_k$  are the mean and standard deviation of component  $c_k$

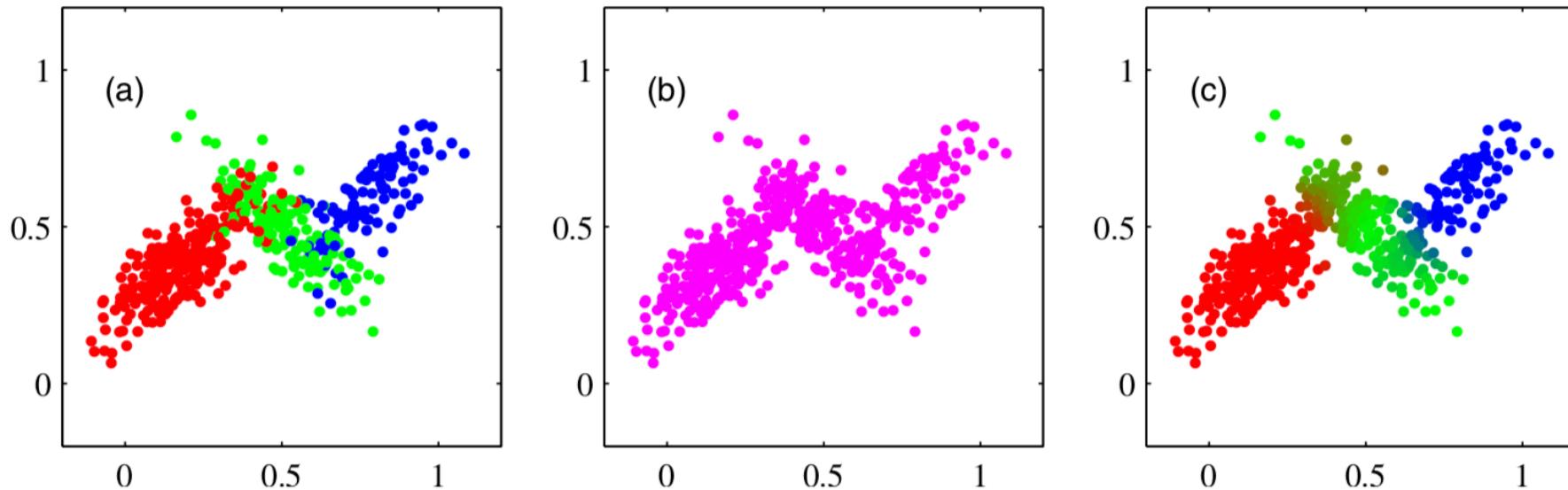
- Multivariate density

- $\mathbf{x}$  generated by cluster  $c_k$

$$p(\mathbf{x} | c_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{u}_k)^T \Sigma_k^{-1} (\mathbf{x} - \mathbf{u}_k)}$$

- where  $\mathbf{u}_k$  and  $\Sigma_k$  are the mean vector and covariance matrix of component  $c_k$

# Example



- (a) three components of the mixture depicted in red, green, and blue
- (b) the corresponding samples from the marginal distribution  $p(\mathbf{x})$
- (c) the same samples in which the colors represent the value of the responsibilities

$$p(c_k \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid c_k)p(c_k)}{p(\mathbf{x})}$$

# Maximum likelihood

- Training set consists on  $n$  observations (samples)

$$X = (\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)$$

- We can represent the dataset as a design matrix  $X$  of size  $n \times m$  as before
- The log of the likelihood function is given by

$$\log p(X | \boldsymbol{\pi}, \mathbf{u}, \boldsymbol{\Sigma}) = \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \cdot N(\mathbf{x} | \mathbf{u}_k, \boldsymbol{\Sigma}_k) \right)$$

# EM algorithm

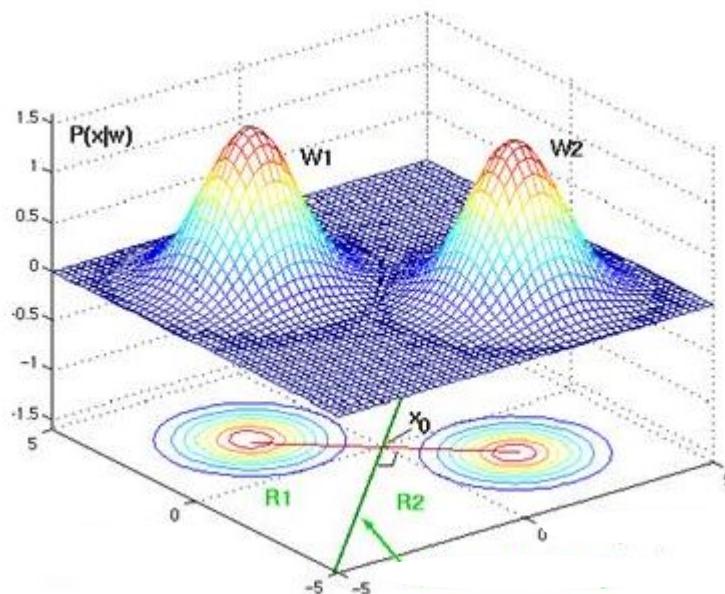
- Given a Gaussian mixture model, the goal is to maximize the likelihood function
  - ... with respect to the parameters: means, covariances and the mixing coefficients
- Four major steps
  - 1. **Initialization**
  - 2. **Expectation**
  - 3. **Maximization**
  - 4. **Evaluate** the log likelihood

$$\log p(X | \boldsymbol{\pi}, \mathbf{u}, \boldsymbol{\Sigma}) = \sum_{i=1}^n \log \left( \sum_{k=1}^K \pi_k \cdot N(\mathbf{x} | \mathbf{u}_k, \boldsymbol{\Sigma}_k) \right)$$

- if convergence criterion is not satisfied return to step 2

# 1. Initialization

- We initialize the mixture parameters arbitrarily
  - mean vectors, usually correspond to random points
  - covariance matrices, generally correspond to identity matrix
  - $\pi_k$  mixing coefficients, generally initialized as  $1/K$  (equiprobability)



## 2. Expectation (E-step)

- Compute for each data point  $\mathbf{x}_i$  and each cluster  $c_k$

$$\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \frac{p(\mathbf{x}_i \mid c_k)p(c_k)}{p(\mathbf{x}_i)} = \frac{N(\mathbf{x}_i \mid \mathbf{u}_k, \Sigma_k) \cdot \pi_k}{\sum_k \pi_k N(\mathbf{x}_i \mid \mathbf{u}_k, \Sigma_k)}$$

- As  $p(\mathbf{x}_i)$  is invariant to the components, we can compute...

$$p(c_k, \mathbf{x}_i) = p(\mathbf{x}_i \mid c_k)p(c_k) = N(\mathbf{x}_i \mid \mathbf{u}_k, \Sigma_k) \cdot \pi_k$$

and then normalize,  $\gamma_{ki} = p(c_k \mid \mathbf{x}_i) = \frac{p(c_k, \mathbf{x}_i)}{\sum_j p(c_j, \mathbf{x}_i)}$

### 3. Maximization (M-step)

- Each observation  $\mathbf{x}_i$  will contribute to update cluster  $\mathbf{c}_k$  with weight  $\gamma_{ki}$
- Accordingly, recalculate the components of the mixture. For each  $\mathbf{c}_k$ :

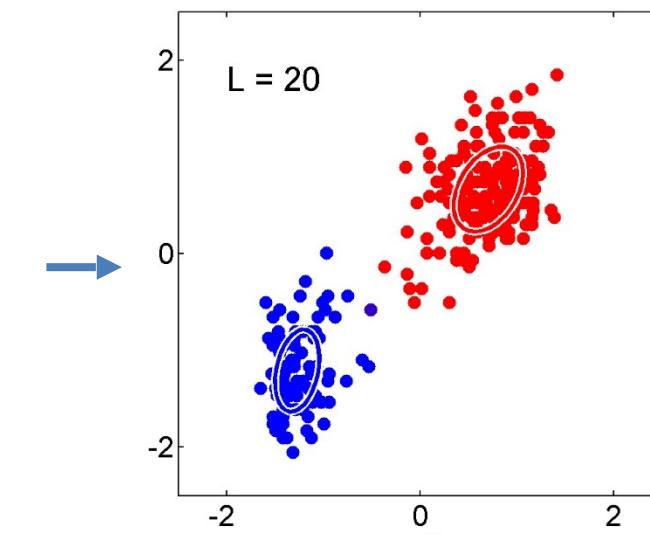
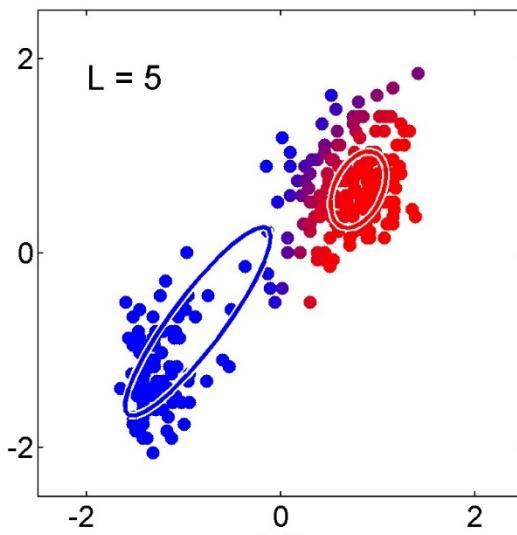
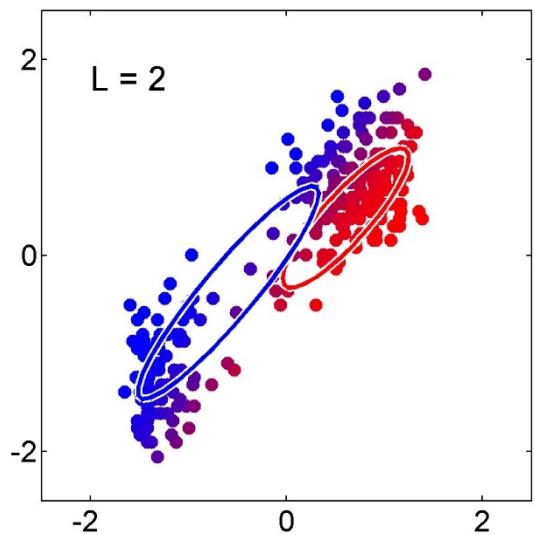
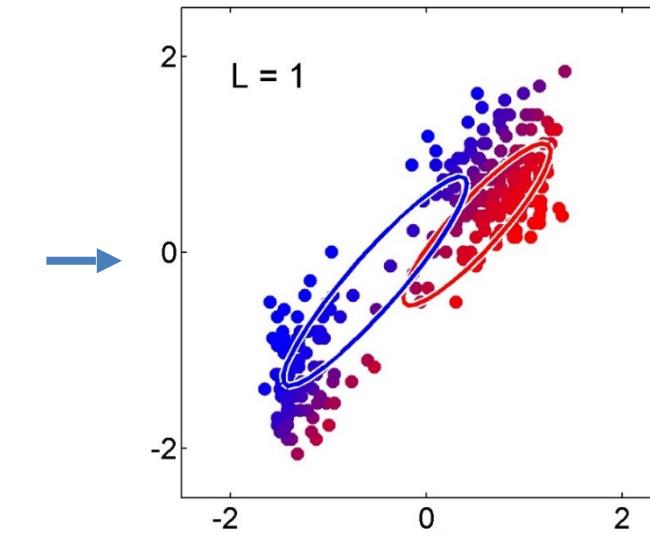
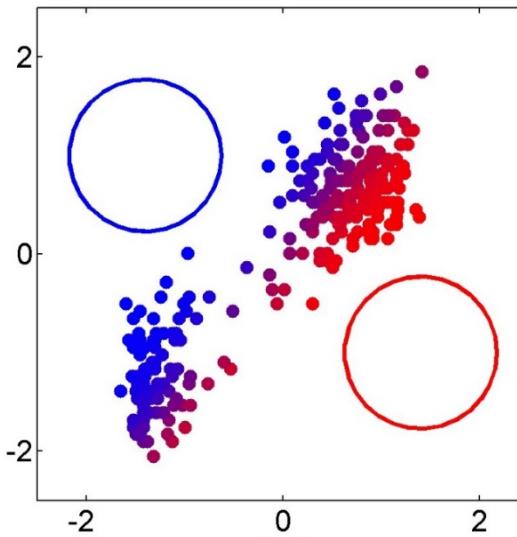
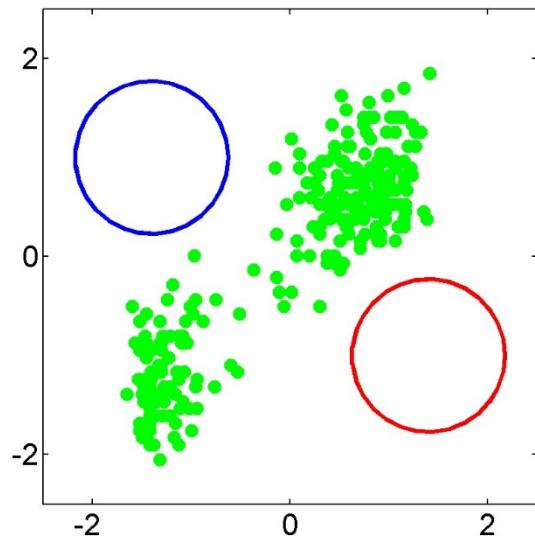
$$N_k = \sum_{i=1}^n \gamma_{ki}$$

$$\mathbf{u}_k = \frac{1}{N_k} \sum_{i=1}^n \gamma_{ki} \bullet \mathbf{x}_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^n \gamma_{ki} \bullet (\mathbf{x}_i - \mathbf{u}_k) \bullet (\mathbf{x}_i - \mathbf{u}_k)^T$$

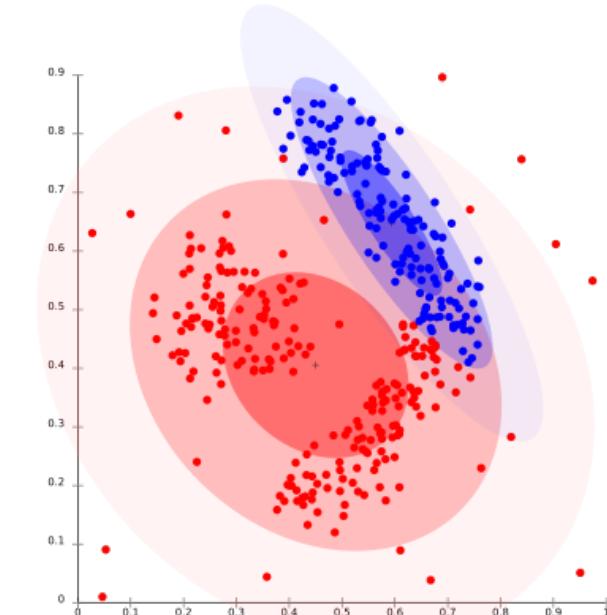
$$\pi_k = p(\mathbf{c}_k) = \frac{N_k}{N}$$

# EM example



# Challenges

- Gaussian components can shrink to cover just a single point
  - When variance goes to zero, likelihood will go to infinity
  - Two components can “merge” acquiring identical means and variances and sharing their data points
  - Other serious problems, especially in high dimensions
- 
- Extension of EM **beyond Gaussian mixture** assumptions
    - exactly as we work with Bayesian approaches:  
replacing  $p(c_k, \mathbf{x}_i)$  by alternative probabilistic views



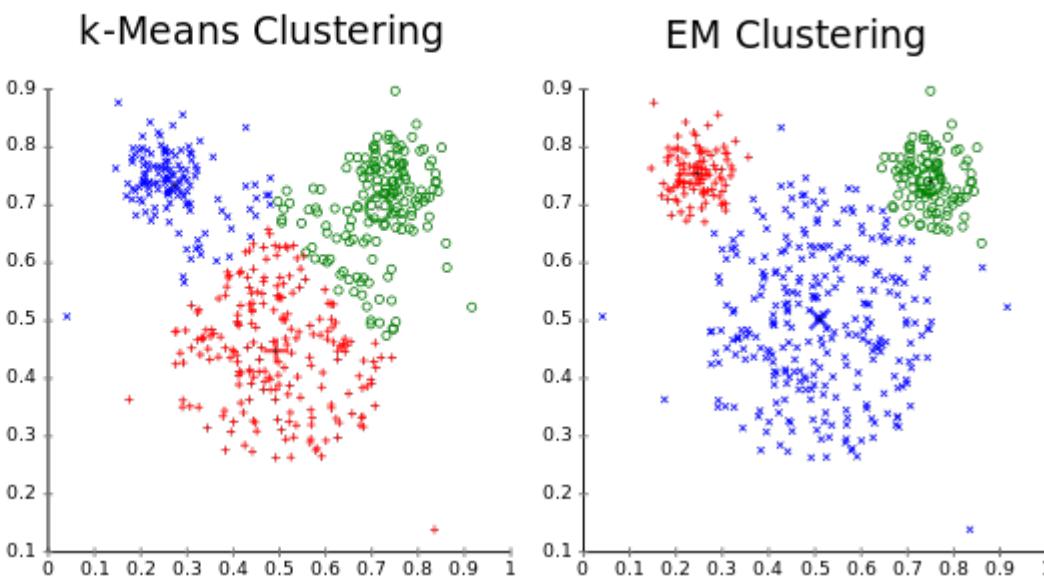
# EM exercise

- Given:
  - \_ data  $\mathbf{x} = \left\{ \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$
  - two clusters with priors  $p(c_1) = 0.6$  and  $p(c_2) = 0.4$
- *Exercise:* perform one step of the EM under real-valued data given by the **Gaussian mixture**

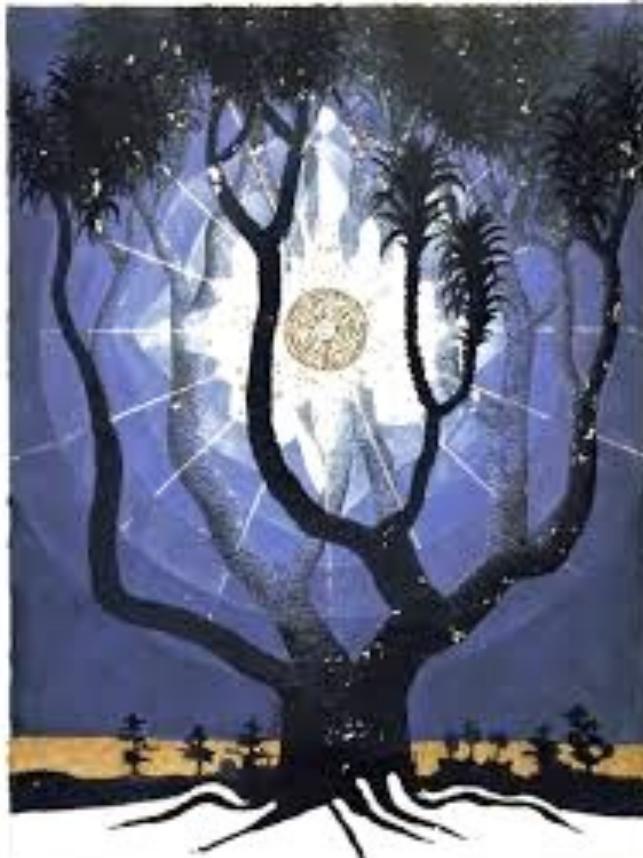
$$\mathbf{u}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

# EM and K-means Clustering

- Practical comparison shows close similarity, yet...
  - $k$ -means algorithm performs a **hard assignment** of data points to clusters, in which each data point is associated uniquely with one cluster
    - in  $k$ -means the shape of the cluster is described by Euclidean distance function...
  - EM algorithm makes a **soft assignment** based on the posterior probabilities



# Outline



- Introduction to clustering
- Partition-based clustering:  $k$ -means
- Model-based clustering: EM
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# Evaluation: clustering validation

- 3 kinds of validity measures: external, internal and relative indexes
- **External criteria** (supervised): extent to which cluster labels match **pre-specified structure**
  - requires prior or expert knowledge
- **Internal criteria** (unsupervised): goodness without external information
  - how well they are separated (e.g. silhouette)
  - should be independent from algorithm-specific functions (unbiased)
- **Relative**: compare different cluster structures (different parameters or algorithms)

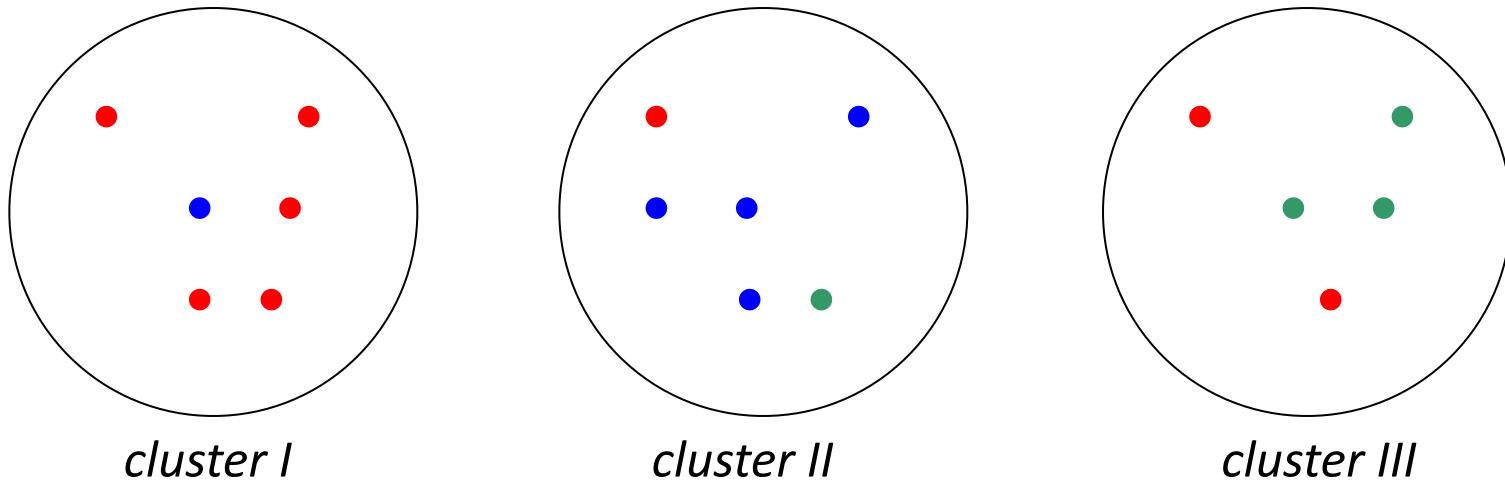
# External measures: purity

- $\{C_1, \dots, C_K\}$  is the set of clusters  
 $\{L_1, \dots, L_G\}$  is the set of reference classes

$$purity = \frac{1}{n} \sum_{k=1}^K \max_j (|C_k \cap Lj|)$$

- **Problem:** biased  $\Rightarrow K = n$  clusters minimize error
- **Alternative:** entropy of classes in clusters (or mutual information between classes and clusters)

## External measures: purity

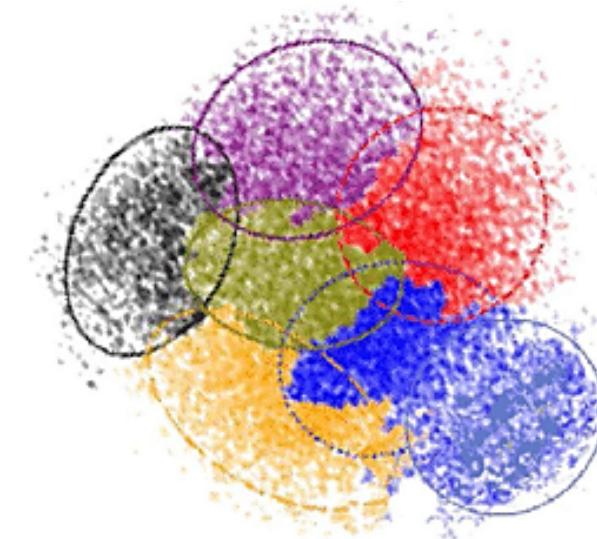


$$\text{purity} = \frac{1}{17}(\max(1,0,5) + \max(4,1,1) + \max(0,3,2)) = \frac{12}{17}$$

$$\text{purity}(C_1) = \frac{5}{6}, \quad \text{purity}(C_2) = \frac{4}{6}, \quad \text{purity}(C_3) = \frac{3}{5}$$

# Internal criteria

- There are two major internal criteria [Berry and Linoff, 1996]
  - **compactness** (cohesion)
    - the members of each cluster should be as close to each other as possible
    - a common measure of compactness is the variance, which should be minimized
  - **separation**
    - the clusters themselves should be widely spaced



# Internal measures: silhouette

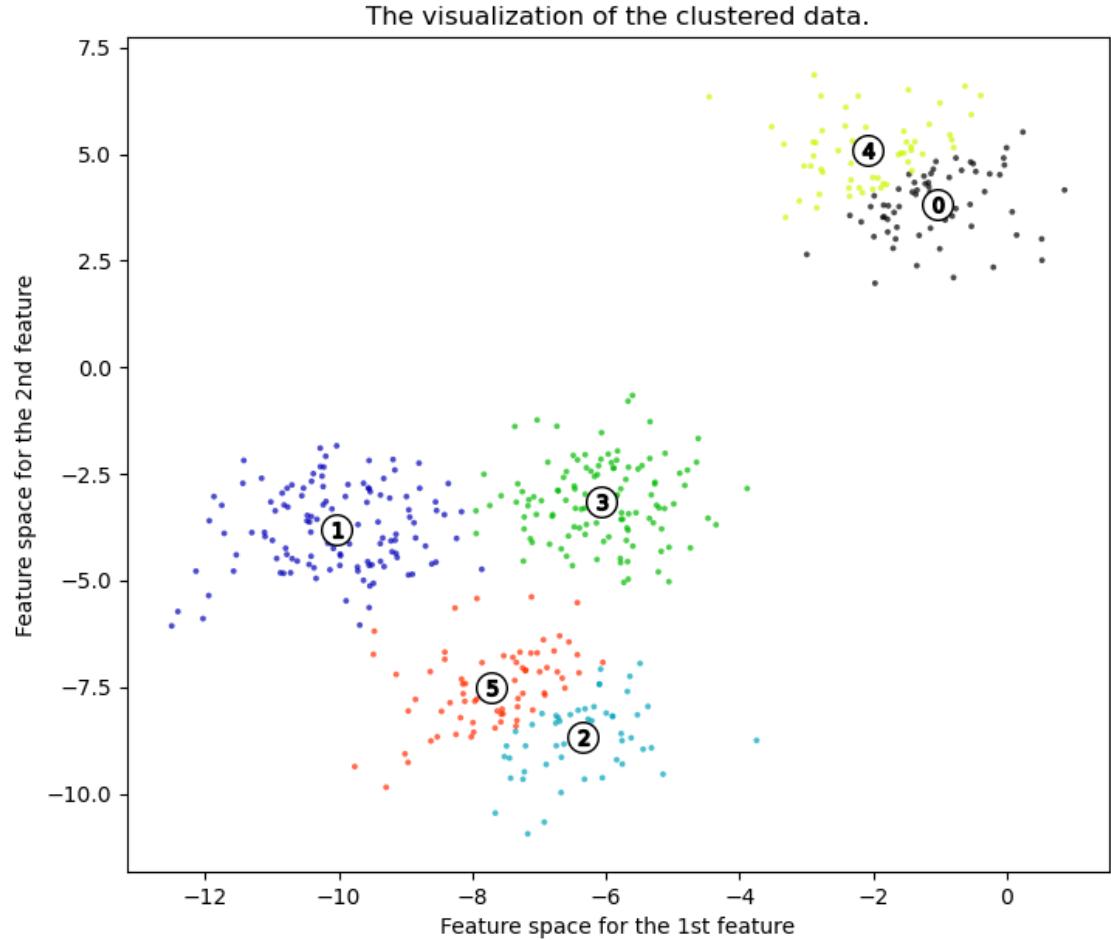
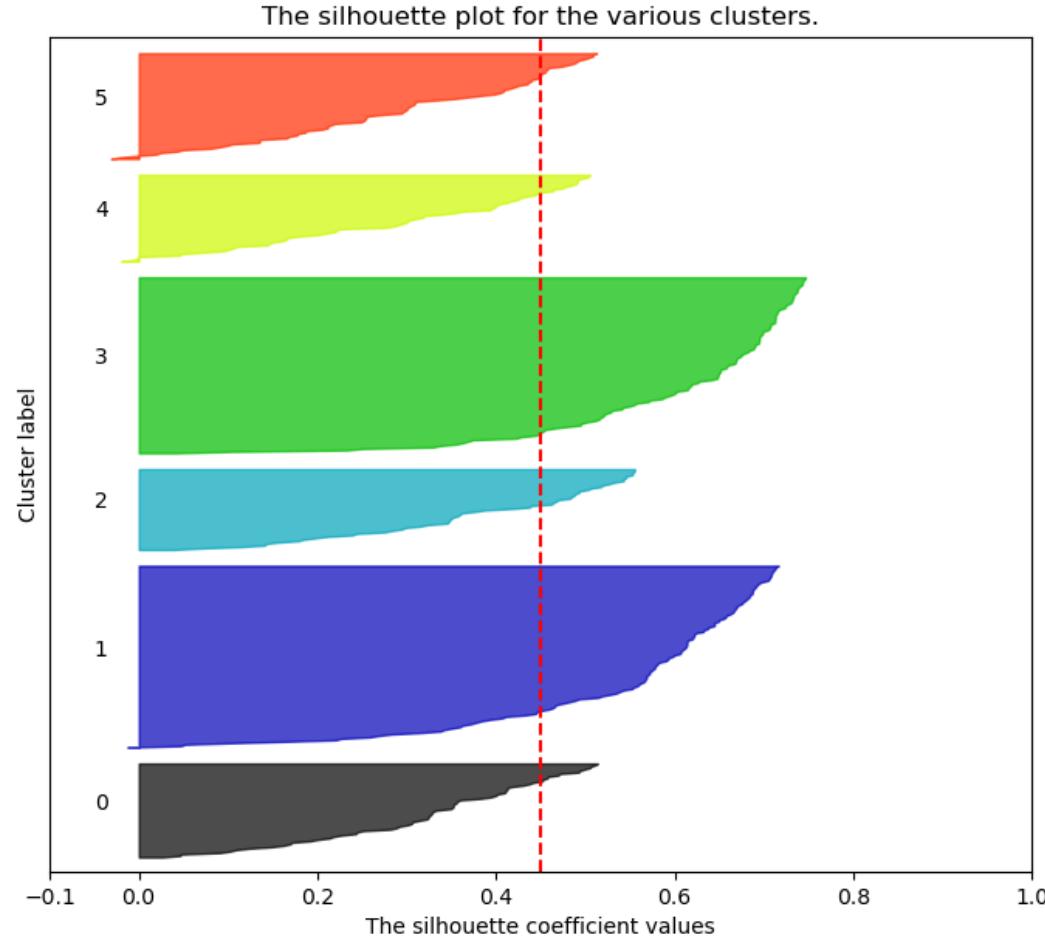
- Silhouette combines both **cohesion** and **separation**
- Calculated for a specific object  $\mathbf{x}_i$ 
  - $a(\mathbf{x}_i)$ = average distance of  $\mathbf{x}_i$  to the points in its cluster
  - $b(\mathbf{x}_i) = \min$  (average distance of  $\mathbf{x}_i$  to points in another cluster)
  - the silhouette coefficient for a point is then given by

$$s(\mathbf{x}_i) = 1 - \frac{a(\mathbf{x}_i)}{b(\mathbf{x}_i)} \text{ if } a < b, \quad (\text{or } s = b/a - 1 \text{ if } a \geq b, \text{ not the usual case})$$

between -1 and 1 (the closer to 1 the better)

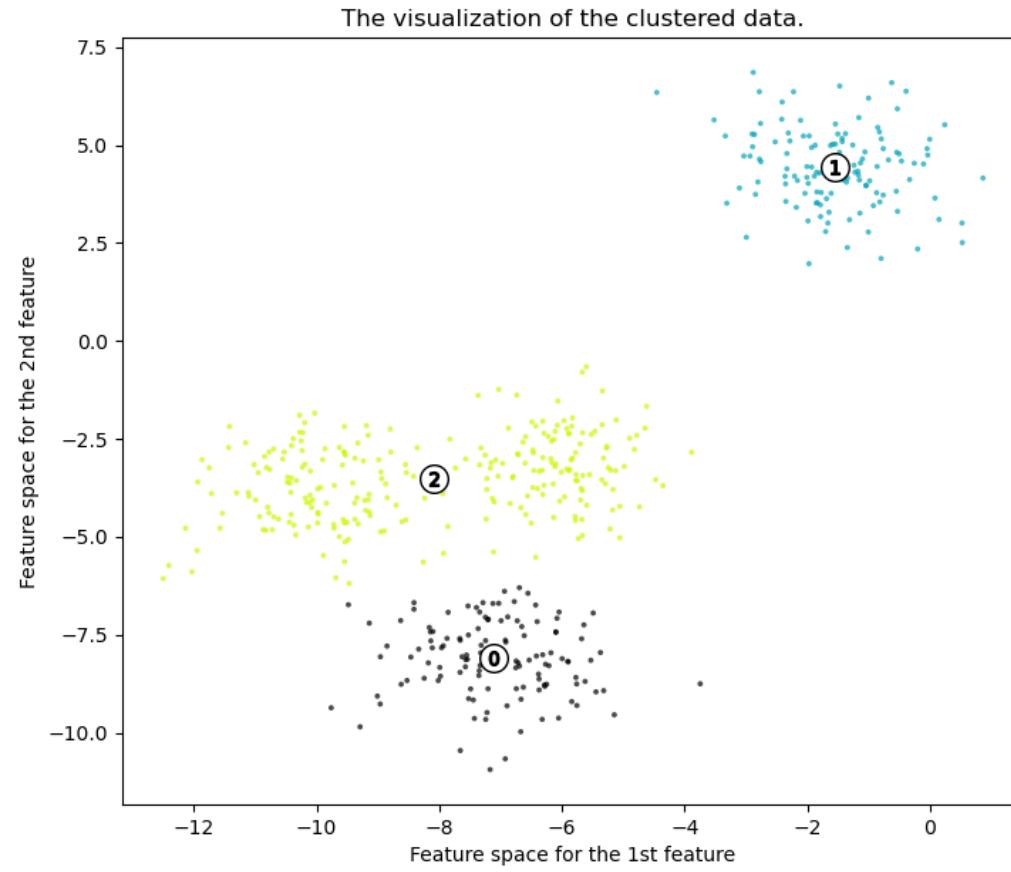
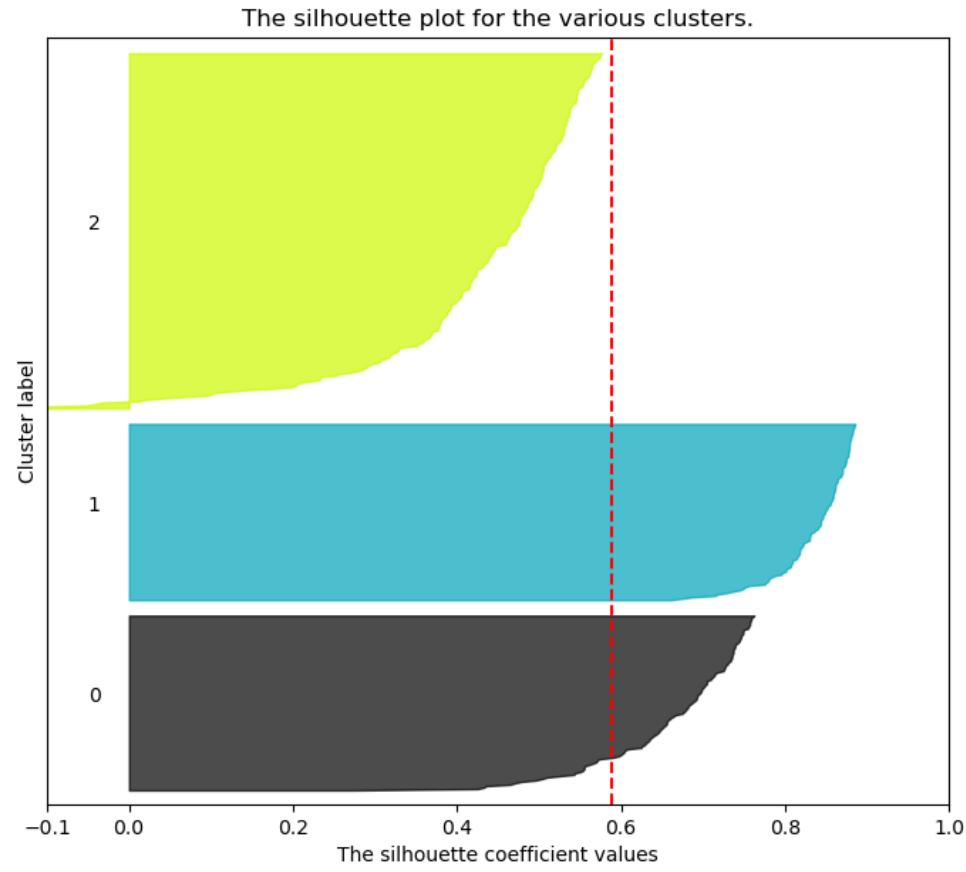
- Silhouette of **cluster**: average of observation silhouettes
- Silhouette of **clustering solution**: average of cluster silhouettes

# Internal measures: silhouette



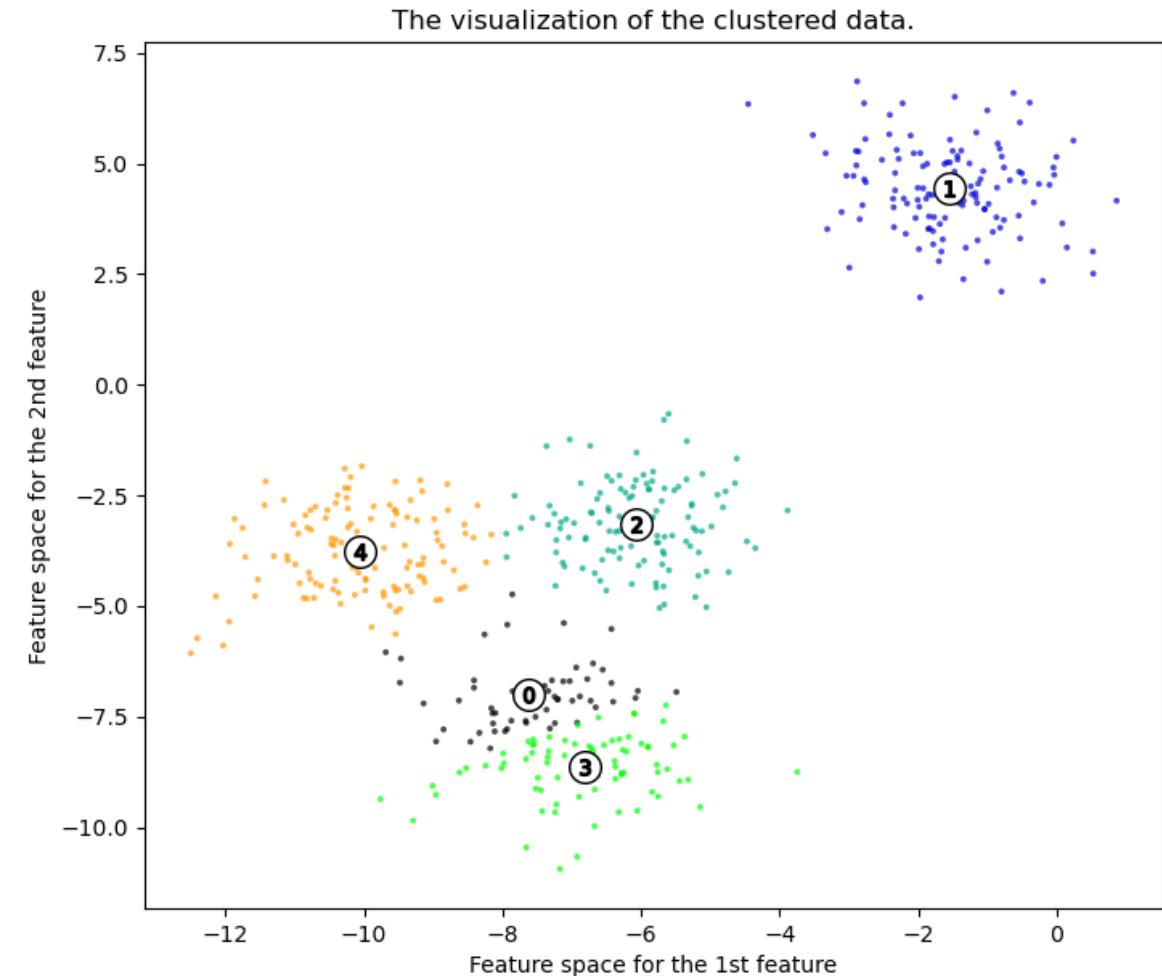
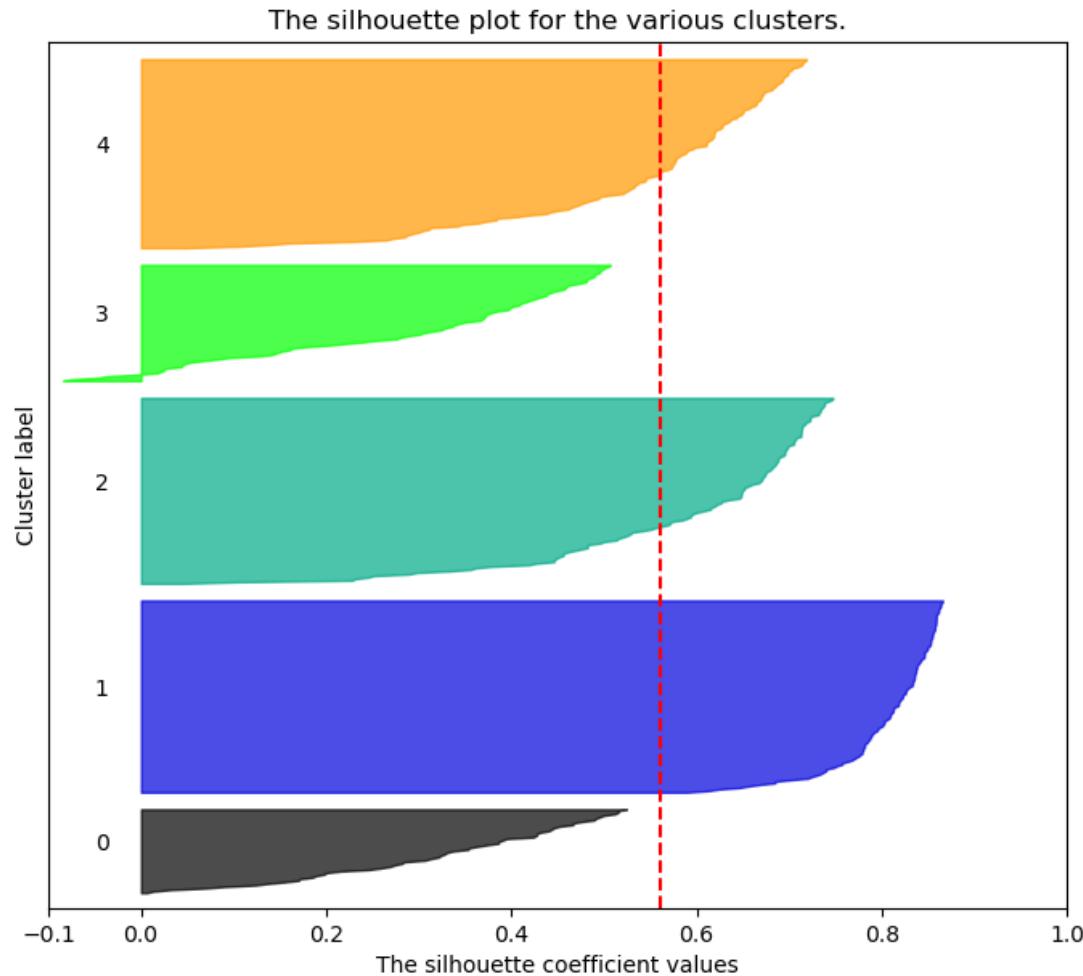
# Internal measures: silhouette

Silhouette analysis for KMeans clustering on sample data with  $n_{clusters} = 3$



# Internal measures: silhouette

Silhouette analysis for KMeans clustering on sample data with `n_clusters = 5`



## Dunn index (*optional*)

- Dunn index  $D_K$ , a cluster validity index originally proposed for  $k$ -means [Dunn, 1974]
  - ***compact*** (dense) and ***well-separated*** clusters

$$d(C_1, C_2) = \min_{\mathbf{x}_i \in C_1, \mathbf{x}_j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$$

$$diam(C) = \max_{\mathbf{x}_i, \mathbf{x}_j \in C} d(\mathbf{x}_i, \mathbf{x}_j)$$

$$D_K = \frac{\min_{1 \leq i < j \leq K} d(C_i, C_j)}{\max_{1 \leq i \leq K} \{ diam(C_i) \}}$$

# Dunn index (*optional*)

- If the dataset contains compact and well-separated clusters:
  - the ***distance*** between the clusters is expected to be *large*
  - the ***diameter*** of the clusters is expected to be *small*
- Large values of the index indicate the presence of compact and well-separated clusters
- $D_K$  can be **normalized** to more easily compare solutions
- Silhouette and  $D_K$  do not exhibit any trend with respect to number of clusters
  - the maximum in a plot with the indices against the number of clusters  $K$  offer an indication of the best number of clusters that fits the data
- Challenges:
  - considerable amount of time to compute silhouette and  $D_K$
  - sensitive to the presence of noise and outliers in data

## Davies-Bouldin (*optional*)

- Revision: Davies-Bouldin (DB) index (1979)

$$DB_K = \frac{1}{K} \sum_{i=1}^K \max_{(i,j), \quad i \neq j} \left\{ \frac{diam(C_i) + diam(C_j)}{d(c_i, c_j)} \right\}$$

- in contrast to  $D_K$ , small  $DB_K$  indices correspond to good clusters, i.e. compact with centers far away
- similarly to  $D_K$ ,  $DB_K$  index exhibits no trend with respect to the number of clusters

# Distances and diameters (*optional*)

- Different distances

- \_ **single** linkage:  $d(C_1, C_2) = \min_{\mathbf{x}_i \in C_1, \mathbf{x}_j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$

- \_ **complete** linkage:  $d(C_1, C_2) = \max_{\mathbf{x}_i \in C_1, \mathbf{x}_j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$

- comparison of **centroids**:  $d(C_1, C_1) = d(c_1, c_2)$

- \_ **average** linkage:  $d(C_1, C_2) = \frac{1}{|C_1| \times |C_2|} \sum_{\mathbf{x}_i \in C_1} \sum_{\mathbf{x}_j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$

(default in silhouette)

# Distances and diameters (cont.)

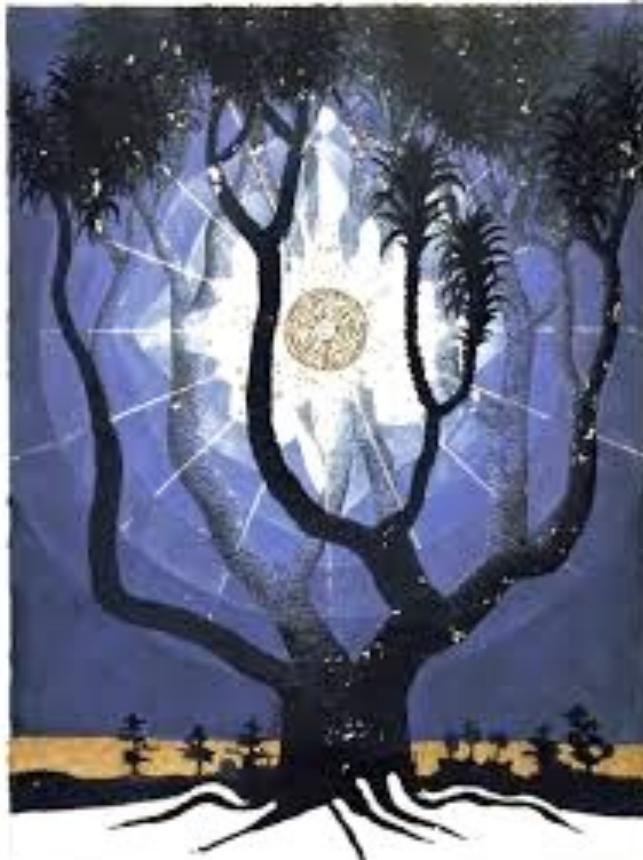
- Different diameters

\_ **maximum:**  $diam(C) = \max_{\mathbf{x}_i, \mathbf{x}_j \in C} d(\mathbf{x}_i, \mathbf{x}_j)$

\_ **radius:**  $diam(C) = \max_{\mathbf{x}_i \in C} d(\mathbf{x}_i, \mathbf{c})$

\_ **average distance:**  $diam(C) = \frac{1}{\binom{|C|}{2}} \sum_{\mathbf{x}_i, \mathbf{x}_j \in C, i < j} d(\mathbf{x}_i, \mathbf{x}_j)$

# Outline

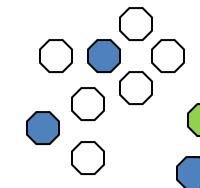


- Introduction to clustering
- Partition-based clustering:  $k$ -means
- Model-based clustering: EM
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# Clustering variants

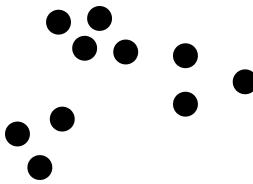
- **Semi-supervised clustering**

- cluster observations when:
  - labels of some observations may be known **or**
  - pairs of observations are known to belong to the same cluster

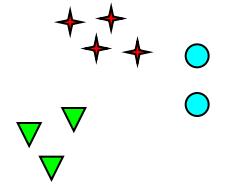
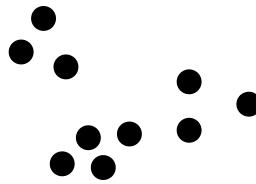


- Separation of clusters: exclusive versus **non-exclusive** (overlapping clusters)
- Hard versus **soft clustering** (each object has a membership for every cluster)
- Complete versus **partial clustering** (observations may not belong to clusters)
- Uniform versus **weighted attributes**

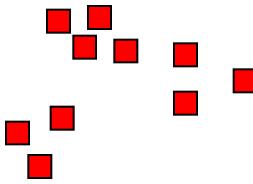
# Number of clusters?



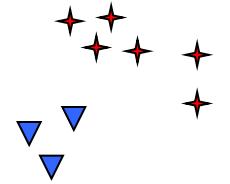
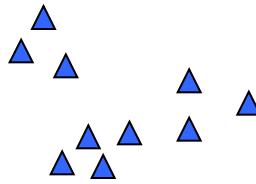
How many clusters?



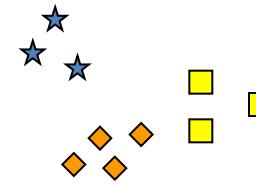
Six Clusters



Two Clusters



Four Clusters

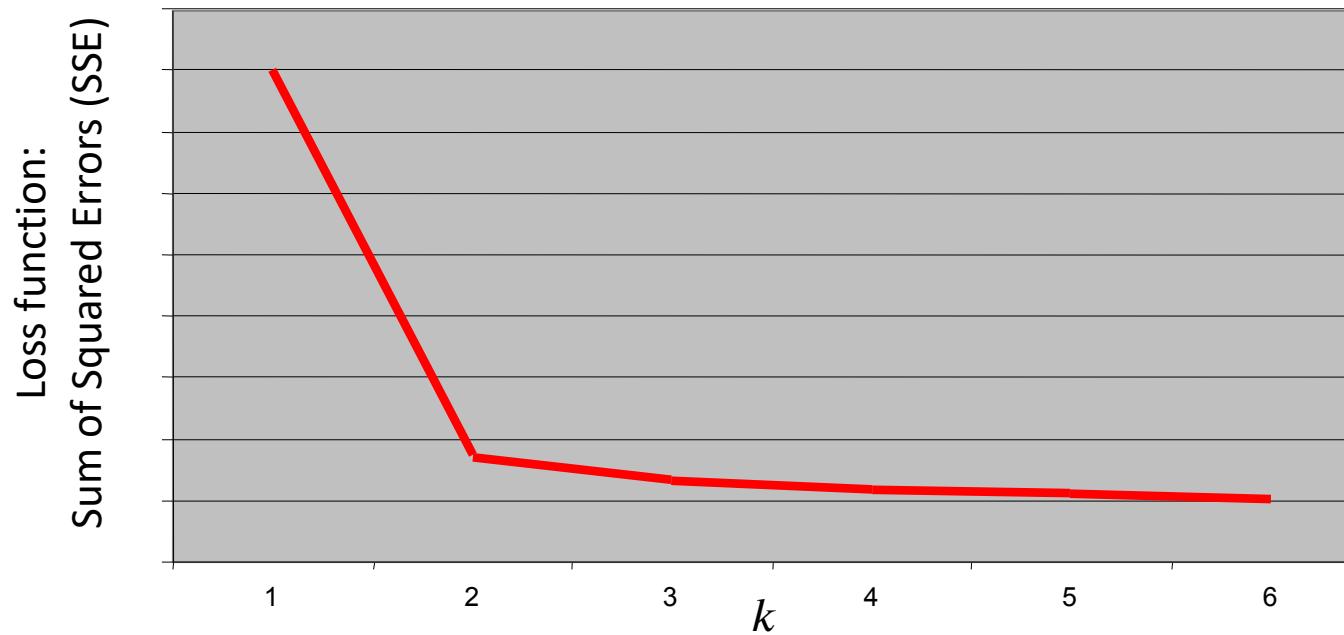


# Number of clusters

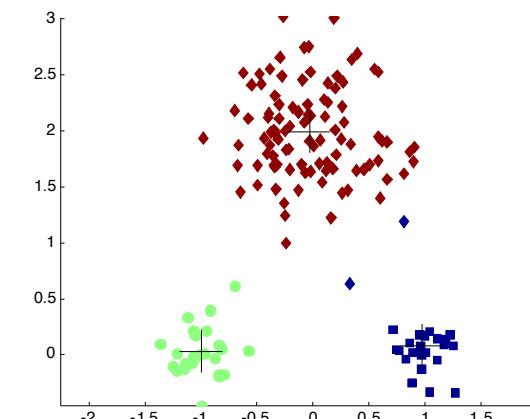
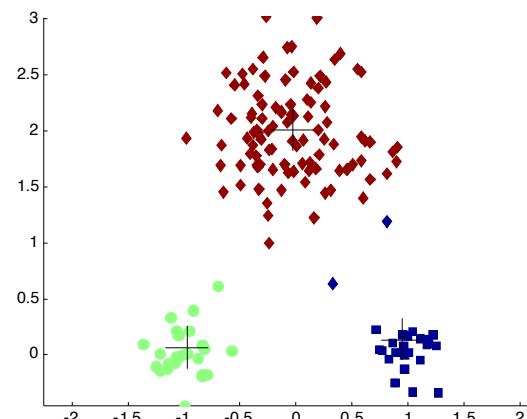
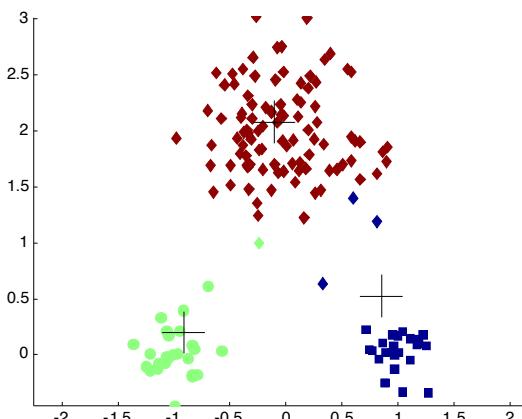
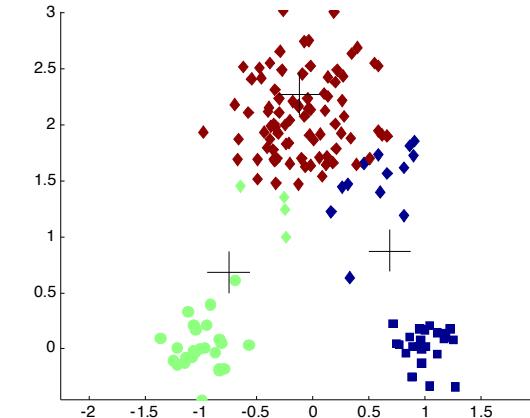
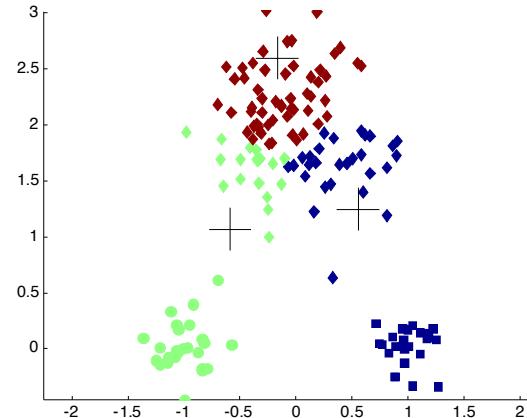
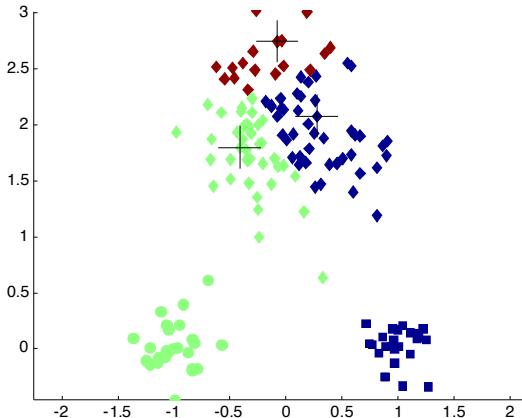
**Knee/elbow finding:**

- plot the squared loss function values for different  $k$
- abrupt changes are highly suggestive =>  $k = 2$  clusters in the example below

**Alternatives:** many others including  $k$  that maximizes internal indices (e.g. silhouette)



# Importance of seeding initial centroids



# Initialization

- Many alternatives for seeding
  - run clustering multiple times with varying seeds and select the best solution
  - use the centroids of an alternative clustering algorithm as seeds of the target clustering algorithm
  - adaptive initialization
    - choose a maximum *radius* and seed spheres (clusters) with the given radius
    - a data point becomes a new cluster seed if not covered by the existing spheres
      - K-MAI clustering (Wichert et al. 2003)
  - alternatives?



# Pre- and post-processing

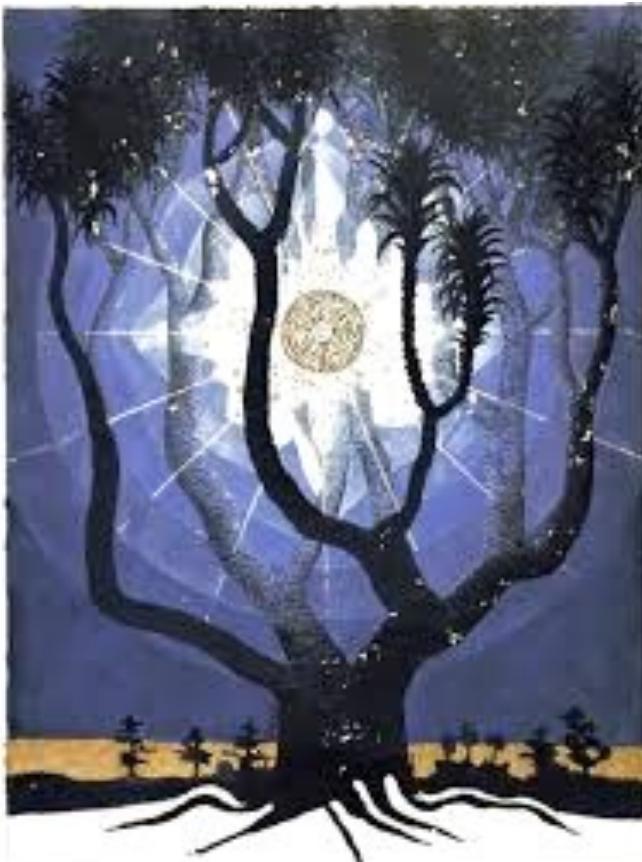
- **Pre-processing**

- normalize data
- data reduction and transformation
- remove outliers

- **Post-processing**

- eliminate small clusters that may represent outliers
- split '*loose*' clusters (clusters with relatively high SSE)
- merge clusters that are '*close*' (clusters with relatively low SSE)
- these steps can be integrated within the clustering process

# Outline



- Introduction to clustering
- Partition-based clustering:  $k$ -means
- Model-based clustering: EM
- Evaluation
  - external measures: purity
  - internal measures: silhouette
- Advanced aspects

# Thank You



[miguel.j.couceiro@tecnico.ulisboa.pt](mailto:miguel.j.couceiro@tecnico.ulisboa.pt)  
[andreas.wichert@tecnico.ulisboa.pt](mailto:andreas.wichert@tecnico.ulisboa.pt)