Chapter 06

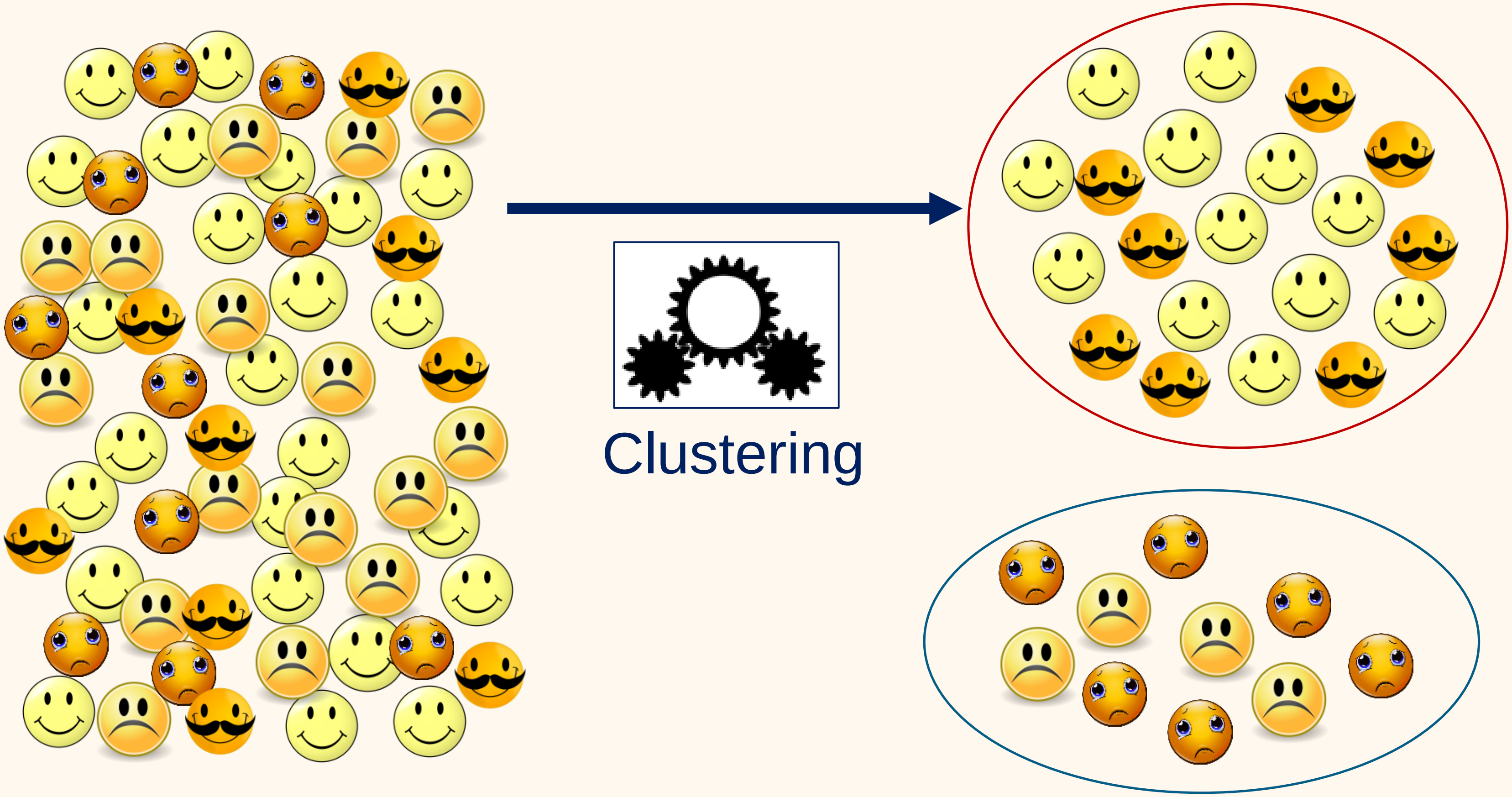
Clustering

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

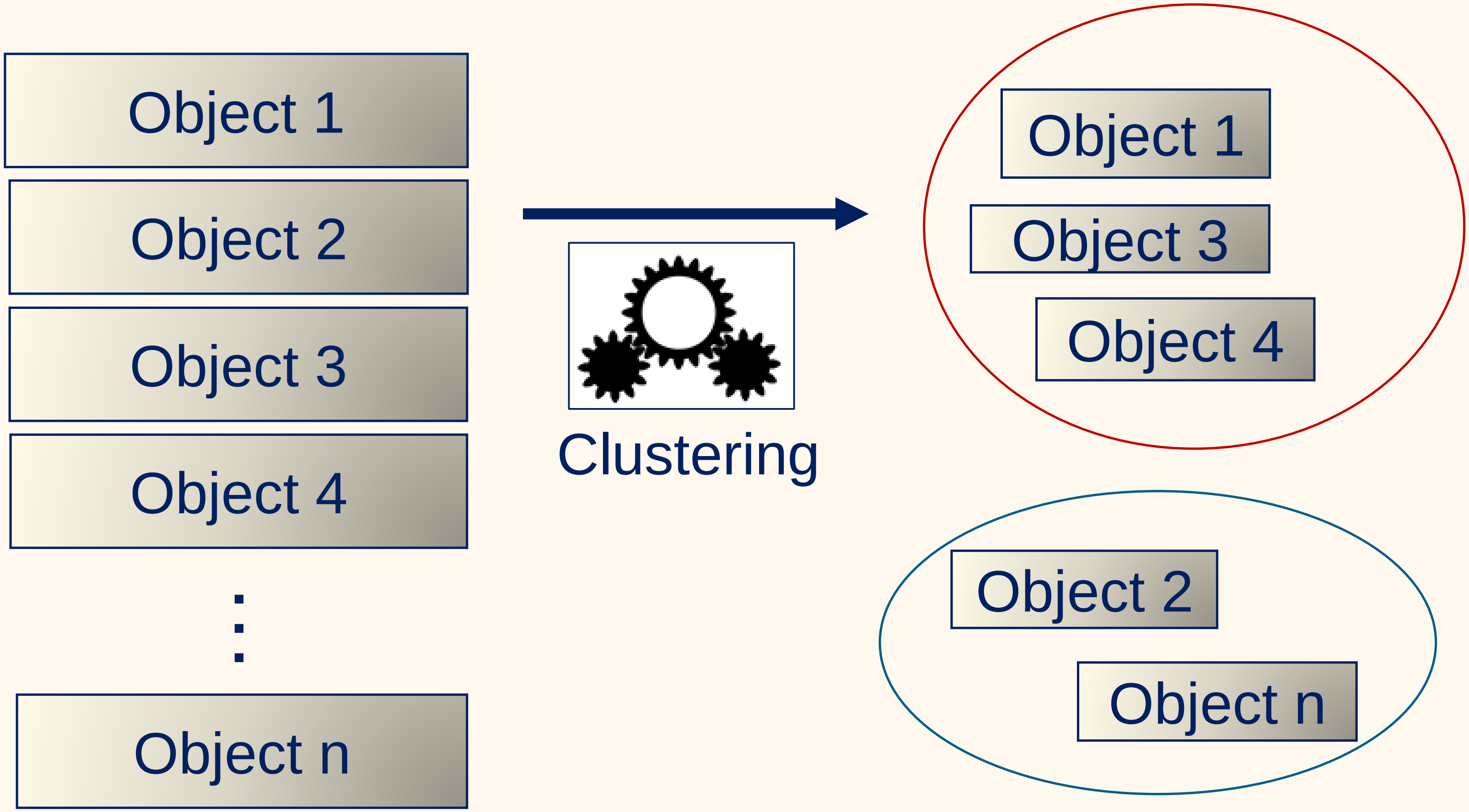
Outline

- Overview
- Clustering algorithms
  - k-means Clustering
  - EM Clustering
  - DBSCAN Clustering
  - Single Linkage Clustering
- Comparison of the Clustering Algorithms
- Summary

Example of Clustering

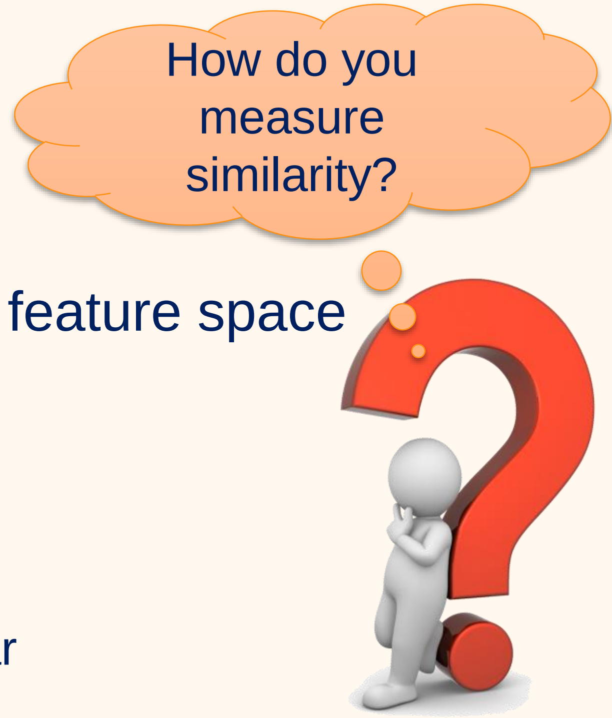


The General Problem



The Formal Problem

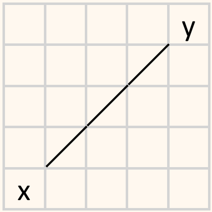
- Object space
  - $O = \{object_1, object_2, \dots\}$
  - Often infinite

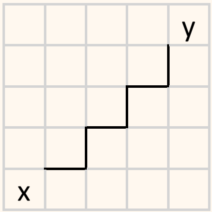


- Representations of the objects in a (numeric) feature space
  - $\mathcal{F} = \{\phi(o), o \in O\}$
- Clustering
  - Grouping of the objects
  - Objects in the same group $g \in G$ should be similar
  - $c: \mathcal{F} \rightarrow G$

Measuring ~~Similarity~~ Distances

- Small distance = similar
- Euclidean Distance
  - Based the eucledian norm $||x||_2$
  - $d(x,y) = ||y-x||_2 = \sqrt{(y_1-x_1)^2 + \cdots + (y_n-x_n)^2}$
- Manhatten Distance
  - Based on the Manhatten norm $||x||_1$
  - $d(x,y) = ||y-x||_1 = |y_1-x_1| + \cdots + |y_n-x_n|$
- Chebyshev Distance
  - Based on the maximum norm $||x||_\infty$
  - $d(x,y) = ||y-x||_\infty = \max_{i=1..n} |y_i - x_i|$





2	2	2	2	2
2	1	1	1	2
2	1	0	1	2
2	1	1	1	2
2	2	2	2	2

Evaluation of Clustering Results

- No general metrics, depends on algorithms
  - Low variance for k-Means
  - High density for DBSCAN
  - Good fit in comparison to model variables for EM clustering
  - …

- Often manual checks
  - Do the clusters make sense?
  - Can be difficult
    - Very large data
    - Many clusters
    - High dimensional data

Outline

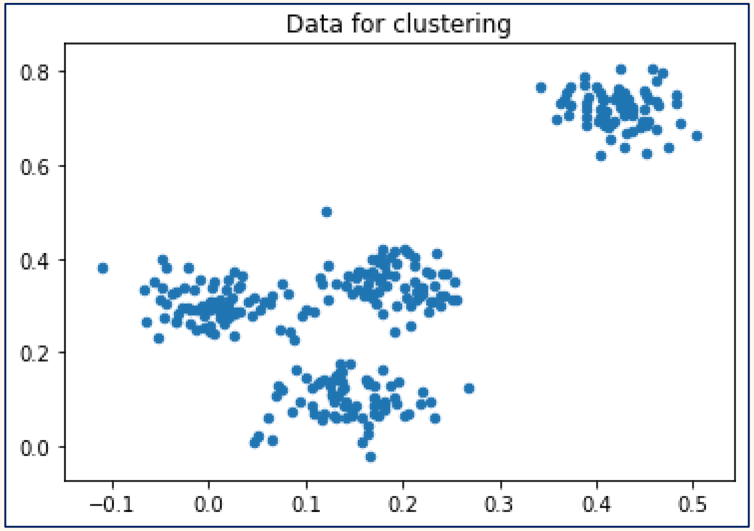
- Overview
- **Clustering algorithms**
  - **k-means Clustering**
  - EM Clustering
  - DBSCAN Clustering
  - Single Linkage Clustering
- Comparison of the Clustering Algorithms
- Summary

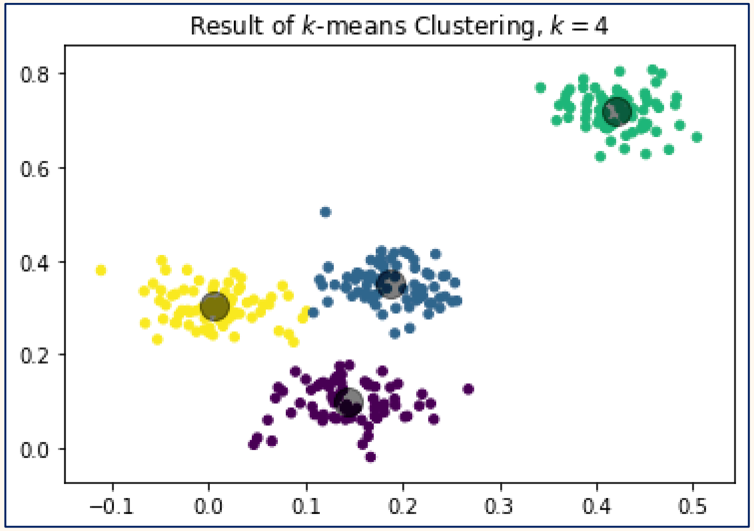
Idea Behind k–means Clustering

How do you get the centroids?

- Clusters are described by their center
  - The centers are called *centroid*
  - Centroid-based clustering

- Objects are assigned to the closests centroid

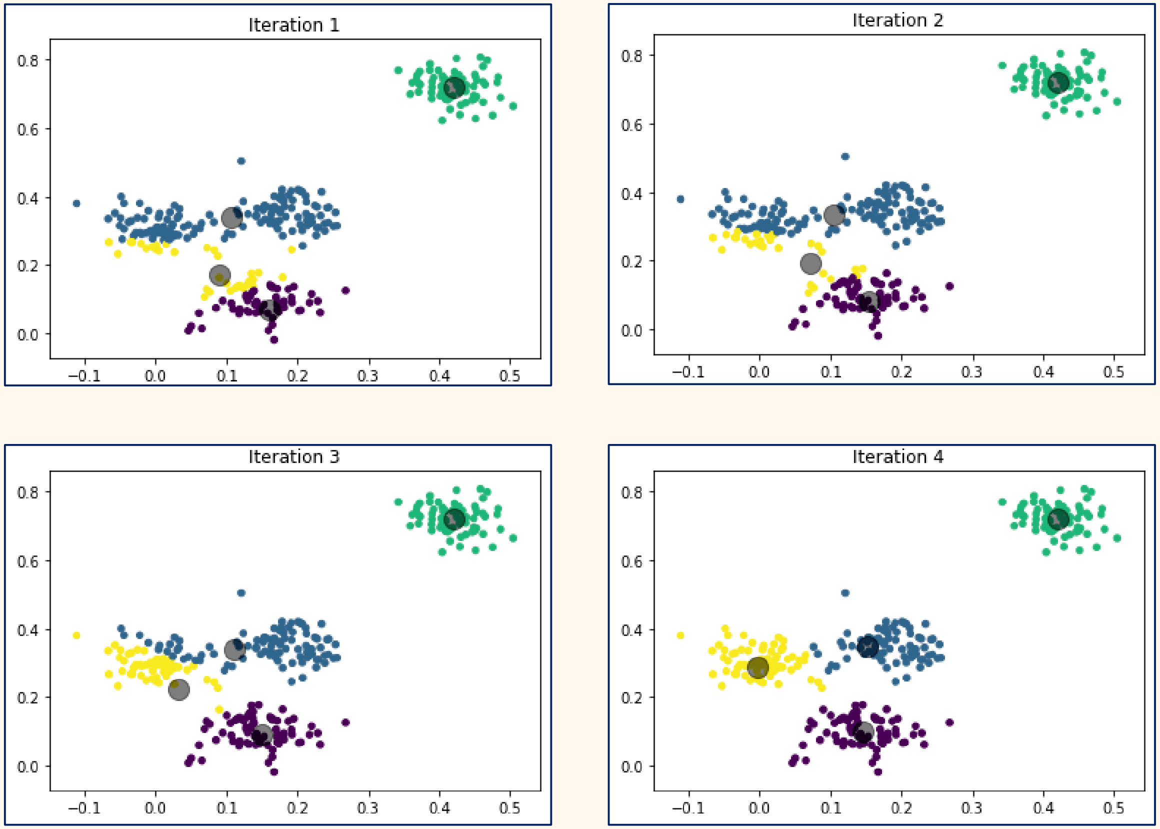




Simple Algorithm

- Select initial centroids $C_1, \dots, C_k$
  - Randomized
- Assign each object to closest centroid
  - $c(x) = \operatorname{argmin}_{i=1..k} d(x, C_i)$
- Update centroid
  - Arithmetic mean of assigned objects
  - $C_i = \frac{1}{|\{x : c(x)=i\}|} \sum_{x:c(x)=i} x_i$
- Repeat update and assignment
  - Until convergence, or
  - Until maximum number of iterations

Visualization of the k–means Algorithm



Selecting k

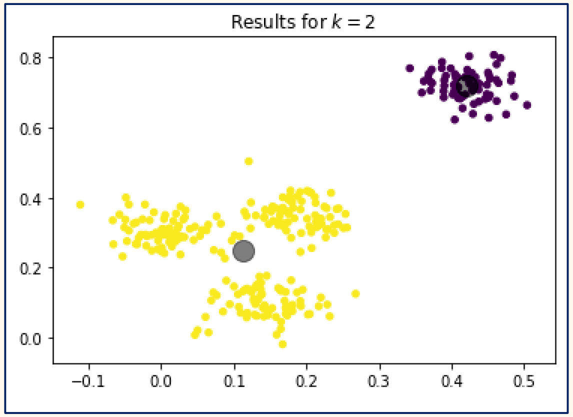
- Intuition and knowledge about data
  - Based on looking at plots
  - Based on domain knowledge

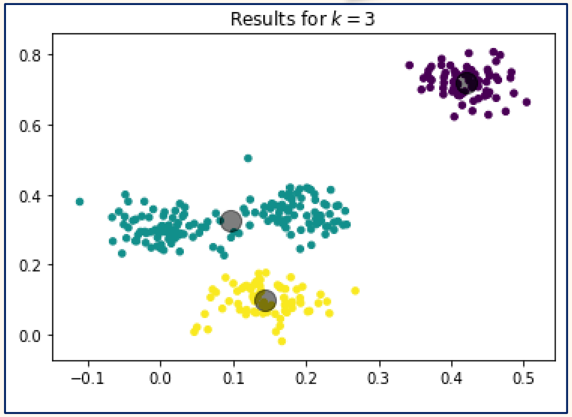
- Due to goal
  - Fixed number of groups desired

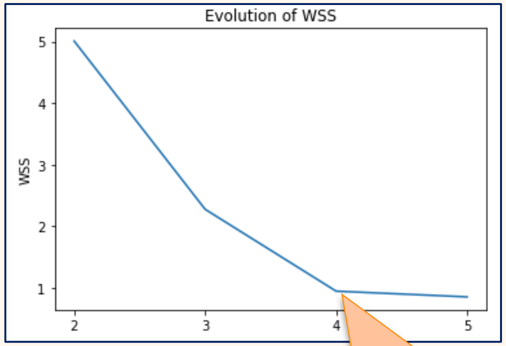
- Based on best fit
  - Within-sum-of-squares
  - $WSS = \sum_{i=1}^{k} \sum_{x:\ c(x)=i} d(x, C_i)^2$

Results for $k = 2, \ldots, 5$

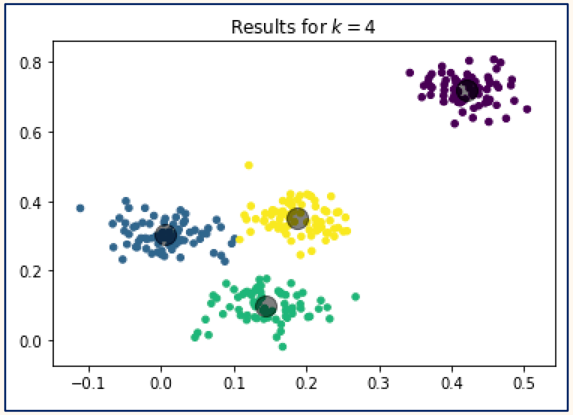
2, 3, and 4 all okay
→ use domain knowledge to decide

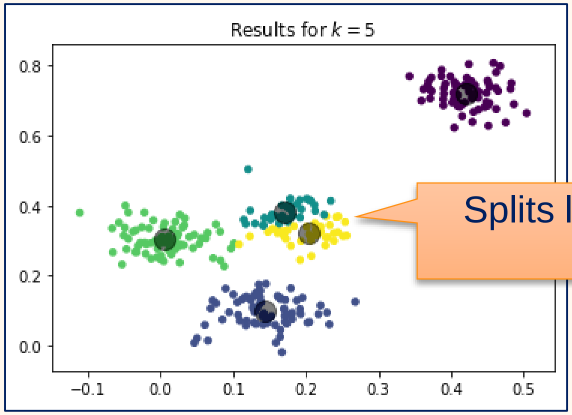






Big changes in slope (elbows) indicate potentially good values for k

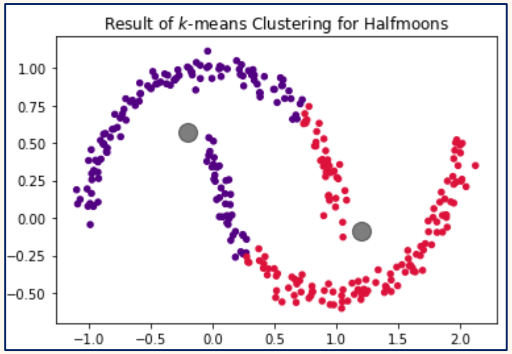




Splits like these indicate too many clusters

Problems of k-Means

- Depends on initial clusters
  - Results may be unstable
- Wrong k can lead to bad results
- All features must have a similar scale
  - Differences in scale introduce artificial weights between features
  - Large scales dominate small scales
- Only works well for “round“ clusters



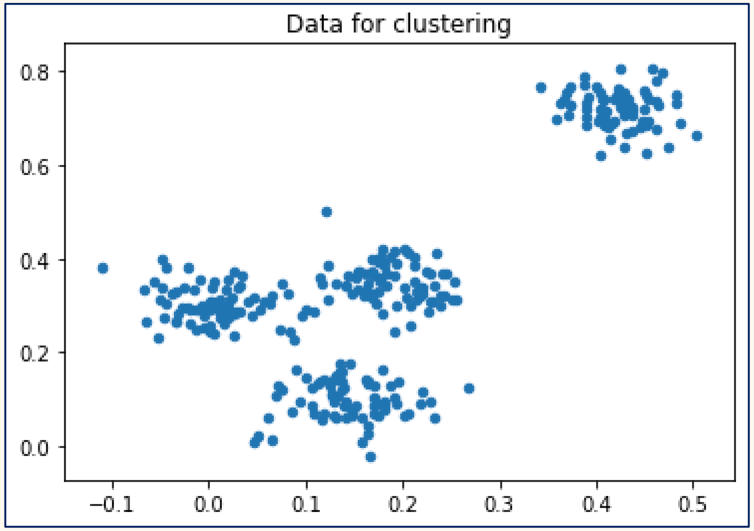
Outline

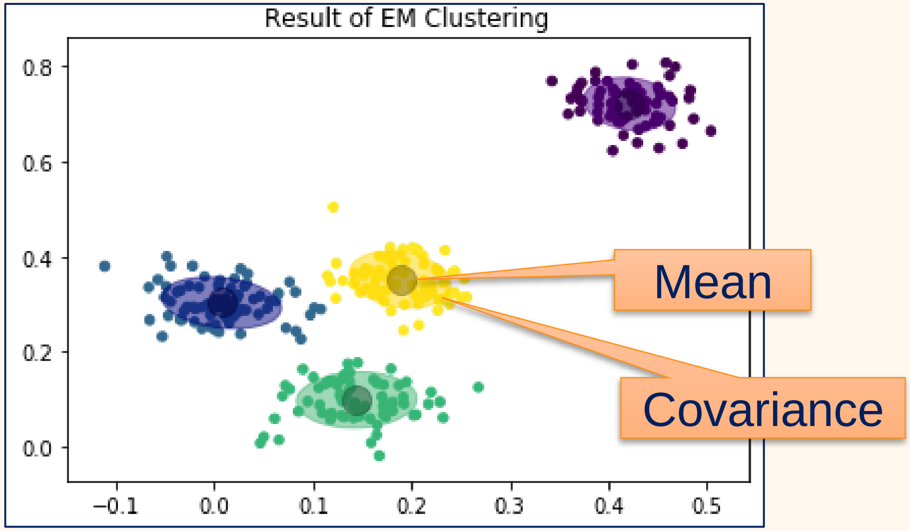
- Overview
- **Clustering algorithms**
  - k-means Clustering
  - **EM Clustering**
  - DBSCAN Clustering
  - Single Linkage Clustering
- Comparison of the Clustering Algorithms
- Summary

Idea Behind EM Clustering

How do you get the distributions?

- Clusters are described by probability distributions
  - Usually normal distribution („Gaussian Mixture Model“)
  - Distribution-based clustering
- Objects are assigned to the "most likely“ cluster



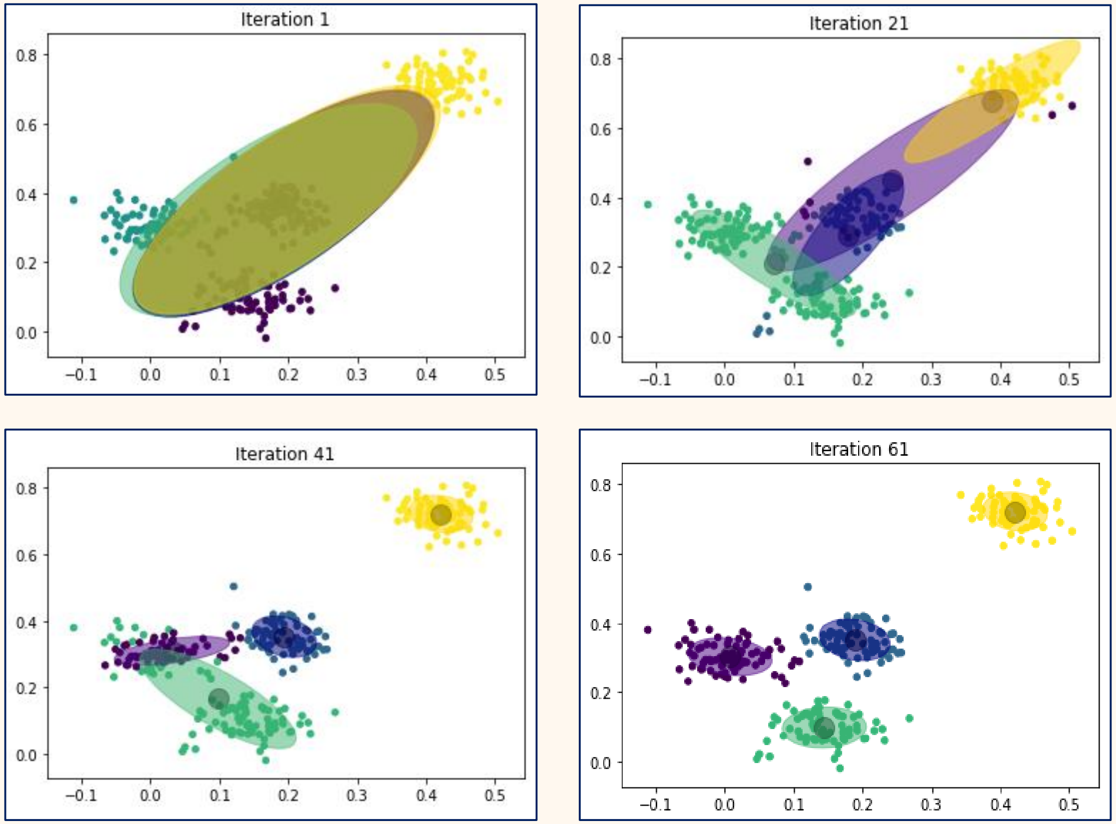


(Simplified!) EM Algorithm

- Task: Determine k normal distributions that "fit" the data well
  - $C_1 \sim (\mu_1, \sigma_1), \ldots, C_k \sim (\mu_k, \sigma_k)$,
  - Estimate start values similar to k-means
- **E**xpectation step
  - Calculate weights of objects
  - Weights define the likelihood that an object belongs to a cluster
  - $w_j(x) = \frac{p(x|\mu_j,\sigma_j)}{\sum_{i=1}^{k} p(x|\mu_i,\sigma_i)}$ for all objects $x \in X$
- **M**aximization step
  - Update mean values
  - $\mu_j = \frac{1}{|X|} \sum_{x \in X} w_j(x) \cdot x$

WARNING:
This is a correct, but simplified version of the algorithm that ignores the update of the (co)variance.

Visualization of the EM Algorithm



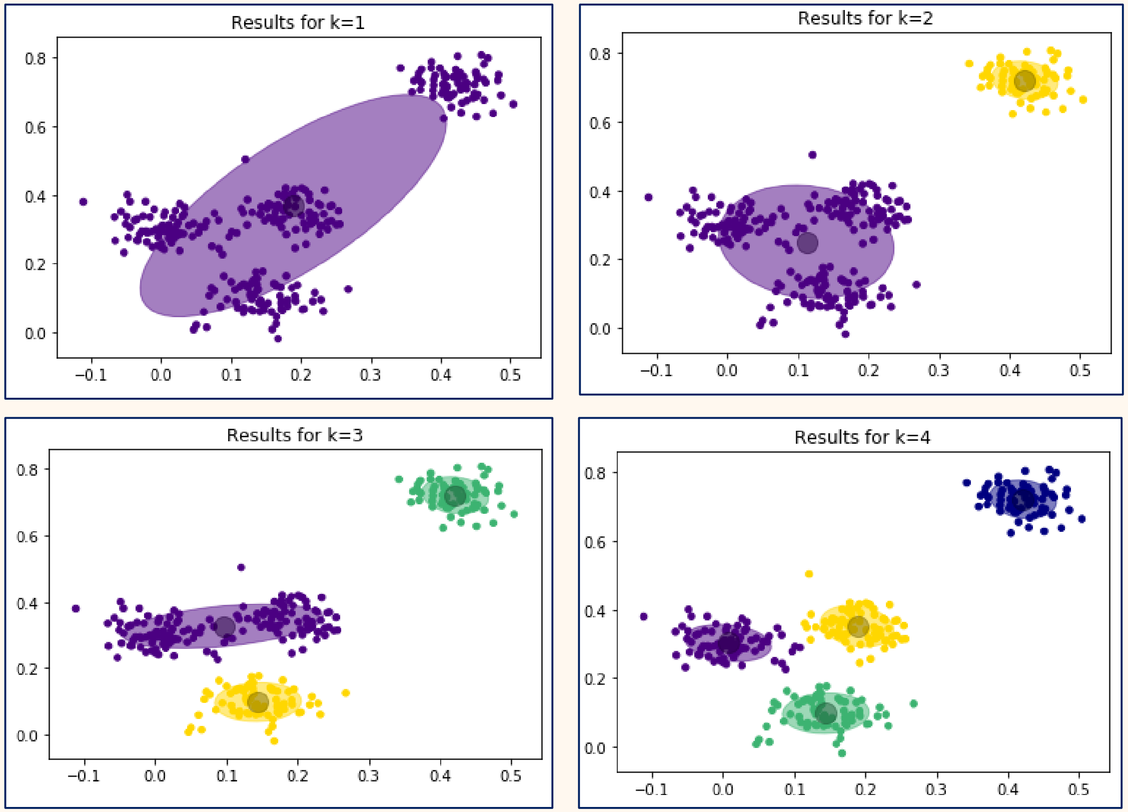
Selecting k

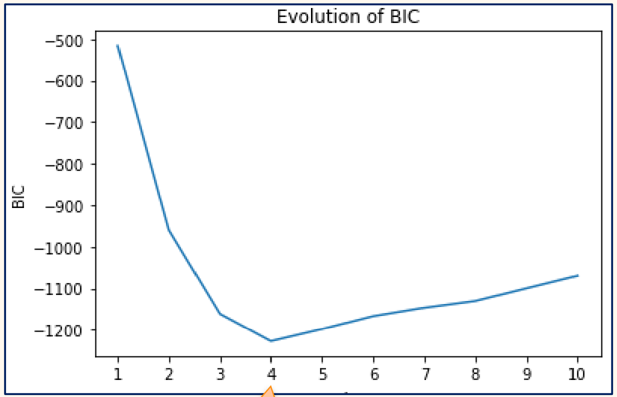
- Same as k-means: Intuition, knowledge, goal

- Bayesian Information Criterion (BIC)
  - Difference between the model complexity and the likelihood of the clusters
  - $BIC = \ln(|\mathrm{X}|)\mathrm{k}' - \hat{L}(C_1, \dots, C_k; X)$
    - k' is the number of model parameters (i.e., mean values, covariances)
    - $\hat{L}(C_1, \dots, C_k; X) = p(C_1, \dots C_k | X)$ is the likelihood function

  - The lower the better
    - Decreases with less complex models
    - Decreases with better likelihood

Results for $k = 1, \dots, 4$



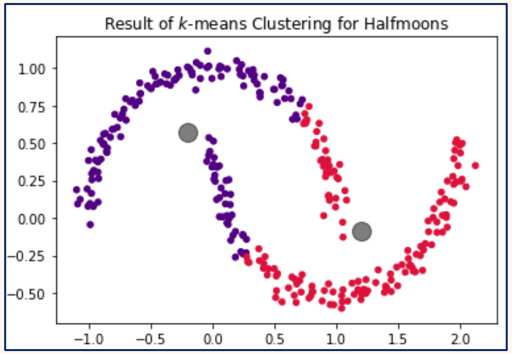


Minimum = optimal ratio between model complexity and goodness of fit

Problems of EM Clustering

- Depends on initial clusters
  - Results may be unstable
- Wrong k can lead to bad results
- May not converge

- Only works well with normally distributed clusters

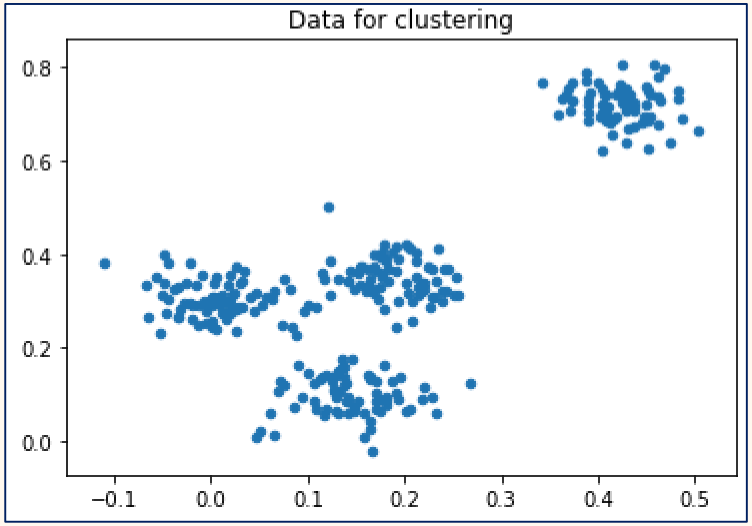


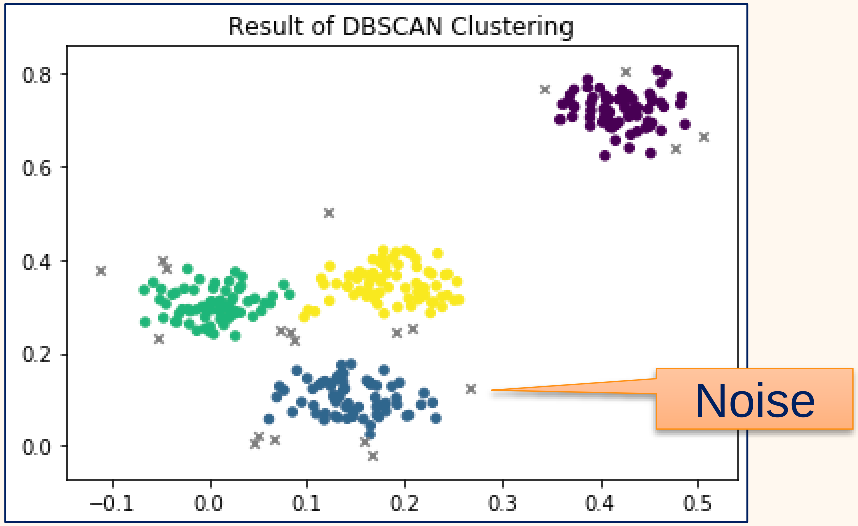
Outline

- Overview
- **Clustering algorithms**
  - k-means Clustering
  - EM Clustering
  - **DBSCAN Clustering**
  - Single Linkage Clustering
- Comparison of the Clustering Algorithms
- Summary

Idea behind DBSCAN

- Clusters are described by other objects close by
  - Density-based clustering
- Scan area around an object for other objects
  - If objects are found, they probably belong to the same group
  - If no objects are found, the object is probably noise

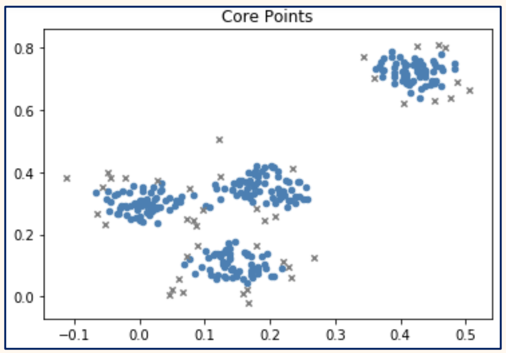


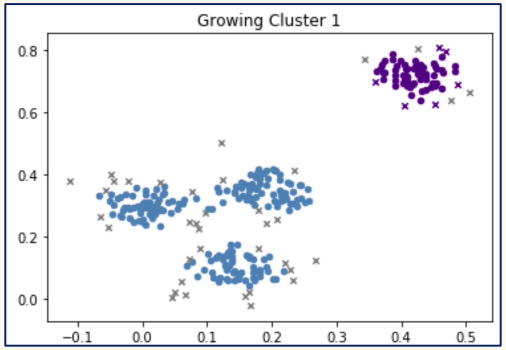


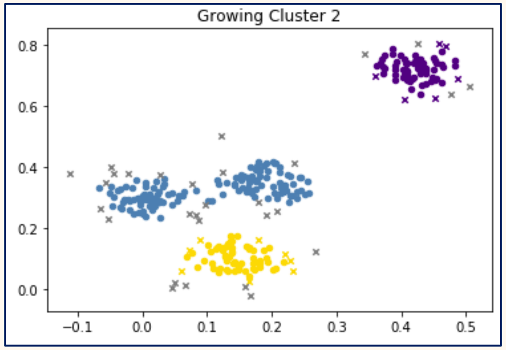
(Relatively) Simple Algorithm

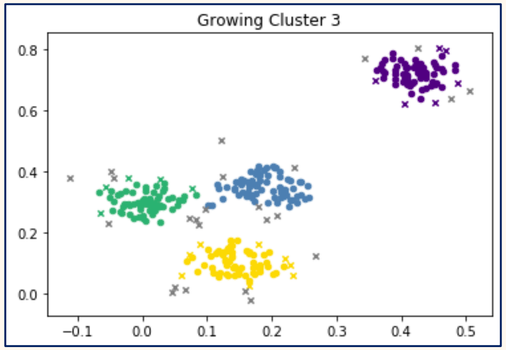
- Two parameters
  - Neighborhood size ϵ
  - Minimal number of points to be considered dense $minPts$
- Determine all objects with dense neighborhoods (core points)
  - $x \in X \text{ such that } |\{x' \in X : d(x, x') \leq \epsilon\}| \geq minPts$
- Grow clusters by assigning all points that share a neighborhood to the same cluster
- All points that are neither core points nor in the neighborhood of a core point are noise

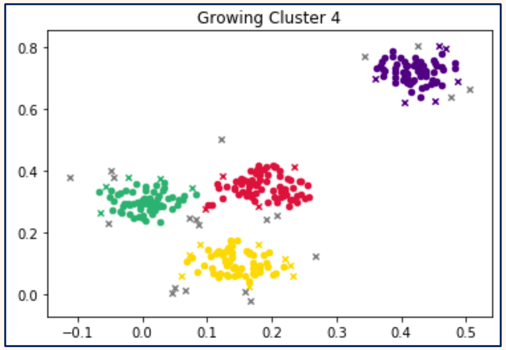
Visualization of the DBSCAN Algorithm





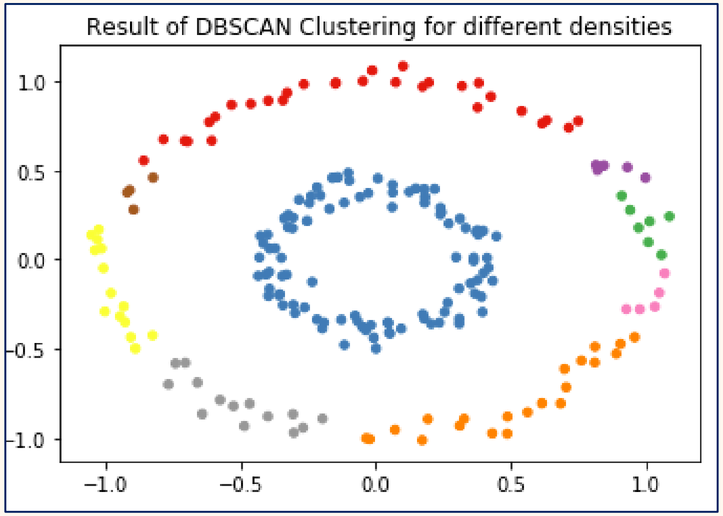






Problems of DBSCAN

- All features must be in the same range
- What if different clusters have different densities?
  → Main problem of DBSCAN!



- This is also related to the size of the data
  → DBSCAN is very sensitive to sampling

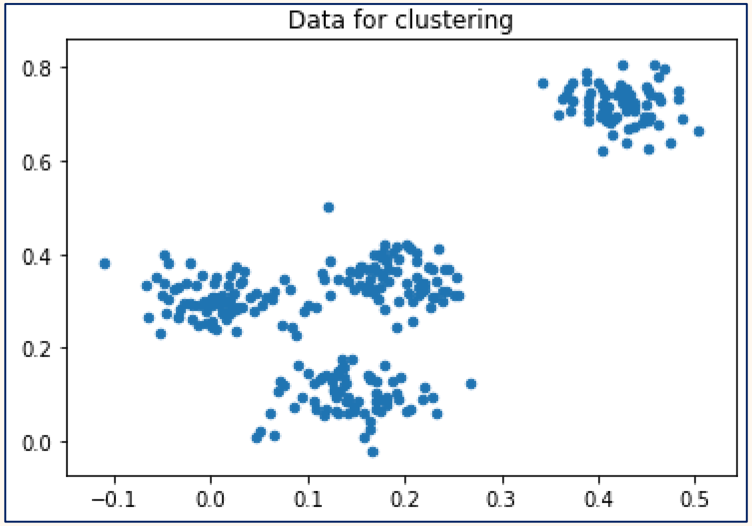
Outline

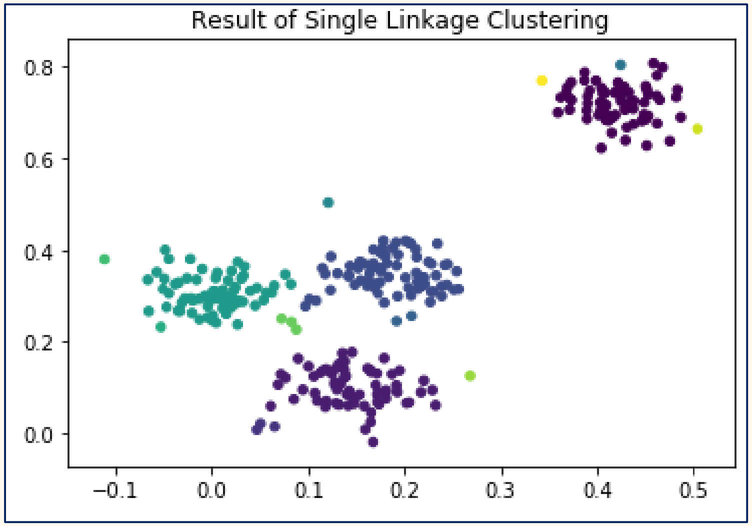
- Overview
- **Clustering algorithms**
  - k-means Clustering
  - EM Clustering
  - DBSCAN Clustering
  - **Single Linkage Clustering**
- Comparison of the Clustering Algorithms
- Summary

Idea behind Hierarchical Clustering

- Clusters are described by hierachies of similarity
  - Hierarchical clustering (also called connectivity-based clustering)

- Find most similar pair of objects and establish link
  - “Nearest Neighbor Clustering“





~~Simple~~ Single Linkage Algorithm (SLINK)

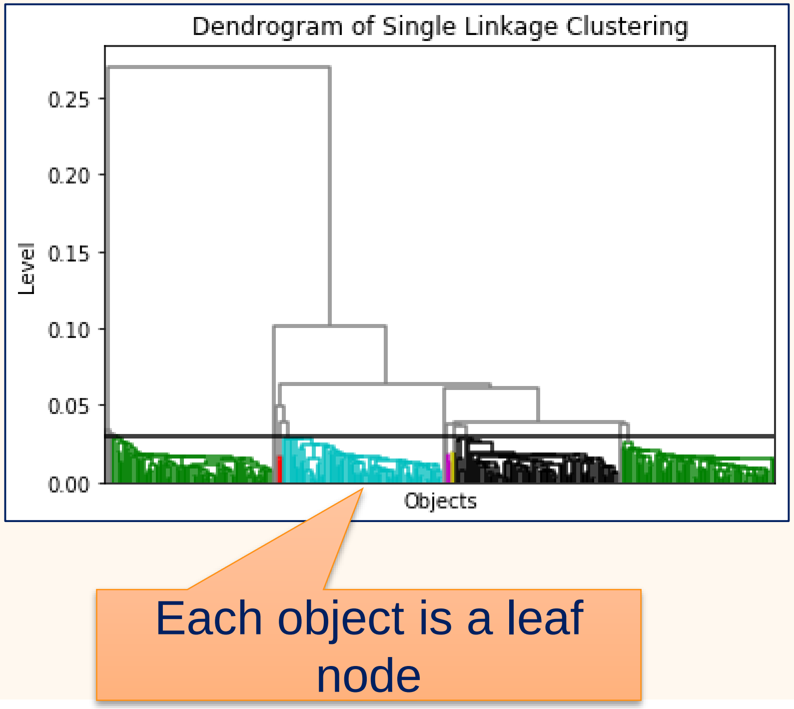
- Every object has its own cluster at the beginning
- The *level* of all these basic clusters is 0
  - $L(C) = 0$ for all $C = \{x\}$ with $x \in X$

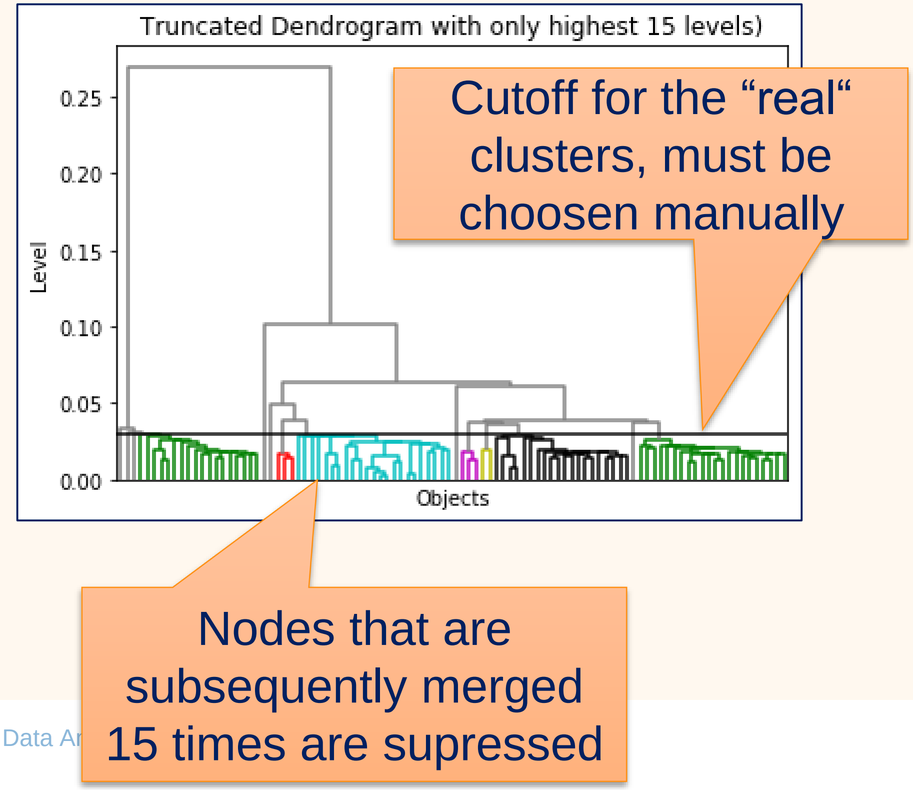
- Find two closest clusters
  - $C, C' = \operatorname{argmin}_{C,C' \in Clusters} d(C, C')$
  - $d(C, C') = \min_{x \in C, x' \in C'} d(x, x')$

- Merge C, C' into a new cluster $C_{new} = C \cup C'$
- The level is the distance between the initial clusters
  - $L(C_{new}) = d(C, C')$

Dendrograms of Clustering

- Visualizes clustering as a tree
  - Horizontal line: Merging of two clusters
  - Vertical line: Increase of the level due to merge





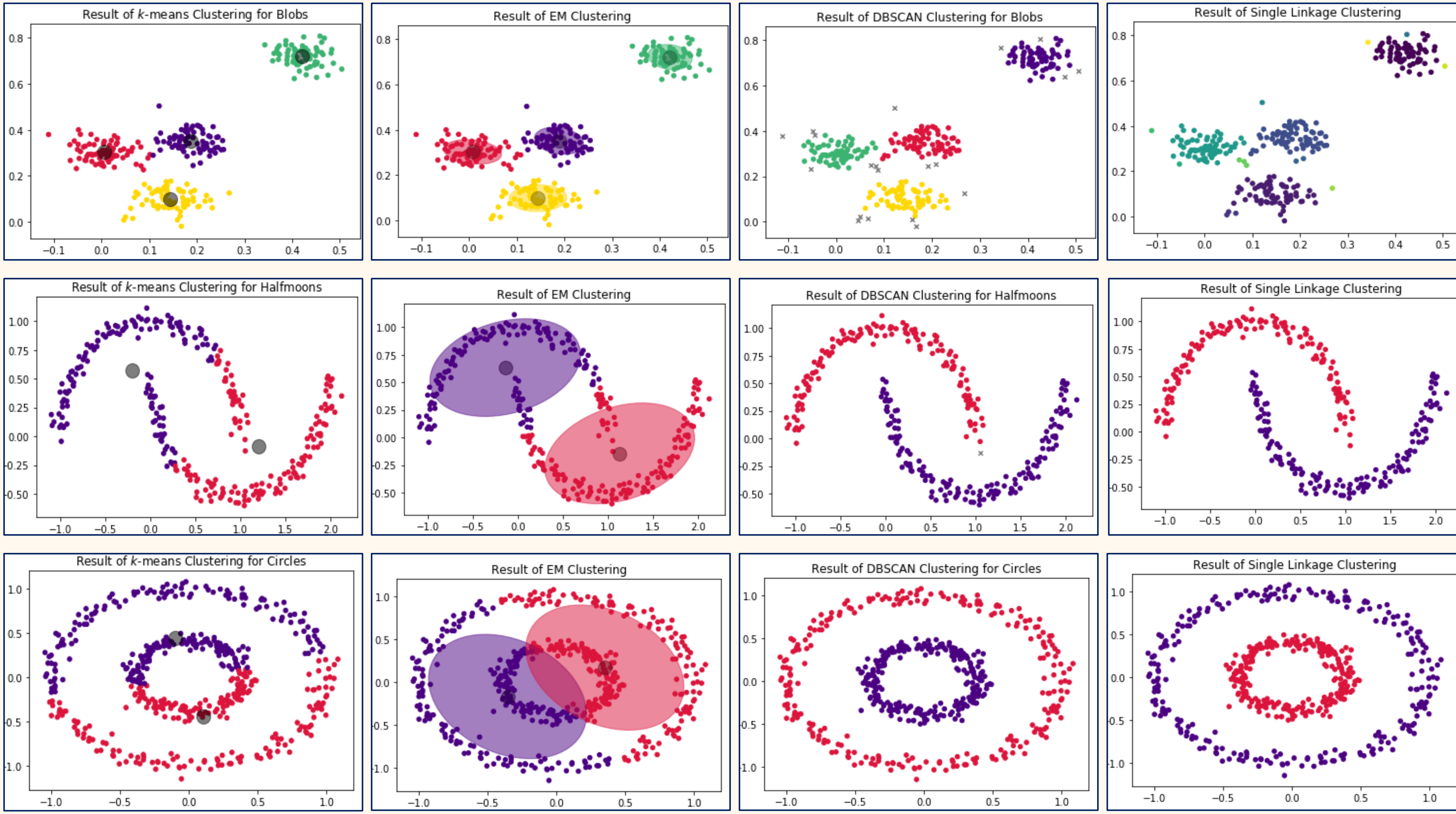
Problems with Hierarchical Clustering

- Often scales badly in terms of memory consumption
  - Standard algorithm requires square matrix of distances between all objects
- All features must be in the same range
- Different densities in different clusters may be problematic
  - Hard to find single cut-off
  - Can be solved by visual analysis of of the dendrogram

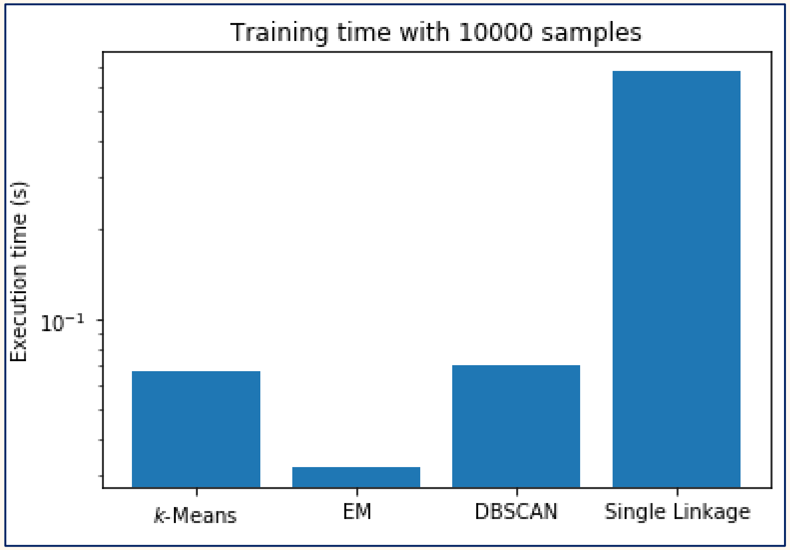
Outline

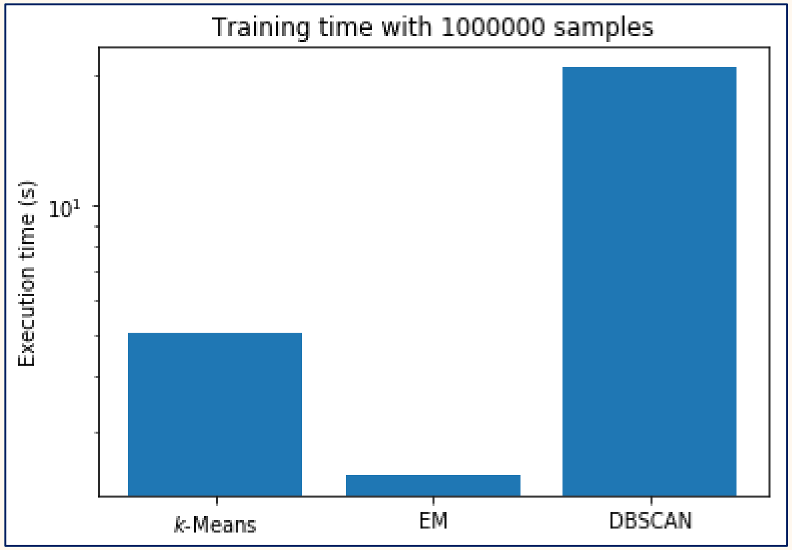
- Overview
- Clustering algorithms
  - k-means Clustering
  - EM Clustering
  - DBSCAN Clustering
  - Single Linkage Clustering
- **Comparison of the Clustering Algorithms**
- Summary

Comparison of Clusters



Comparison of Execution Times





- Single linkage requires to much memory for larger clusters

Strengths and Weaknesses

	Cluster number	Explanatory value	Consise representation	Categorical features	Missing features	Correlated features
k-means	-	+	+	-	-	0
EM	0	+	+	-	-	0
DBSCAN	+	-	-	-	-	0
SLINK	0	+	-	-	-	0

There are clustering algorithms for categorical data, e.g., k-modes

Summary

- Clustering is concerned with the inference of groups for objects
- Works well for numeric data but is often not well suited for categorical data
  - Scales are very important for most clustering algorithms
- Different types of clustering algorithms
  - Centroid-based
  - Distribution-based
  - Density-based
  - Hierarchical / connectivity-based
- Evaluation often difficult and requires manual intervention