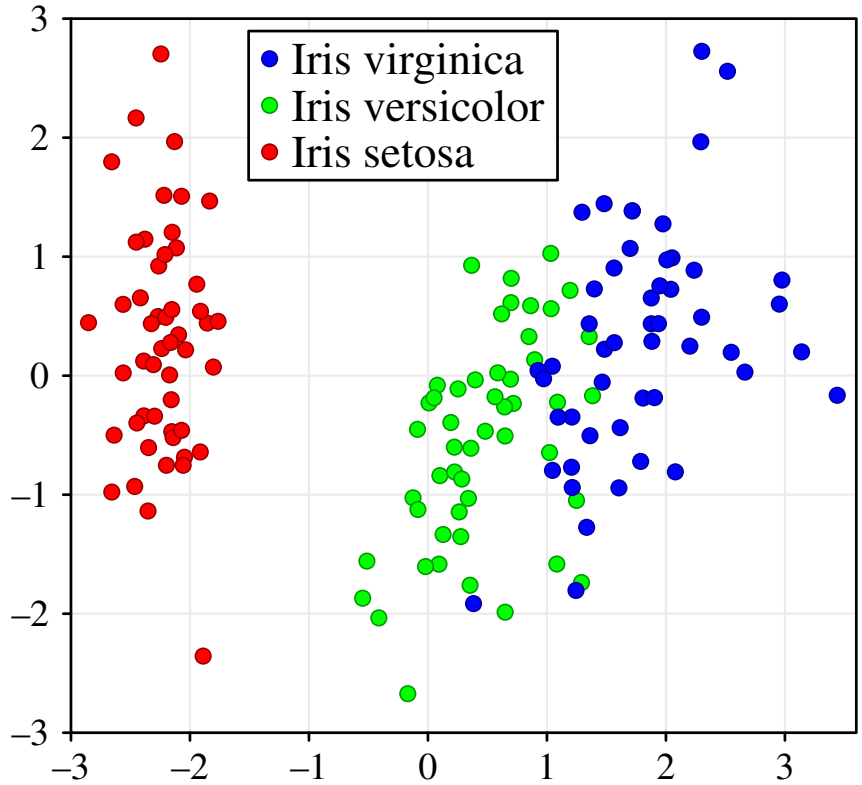
Finding Clusters

Types of Clustering Approaches:

- Linkage Based, e.g. Hierarchical Clustering
- Clustering by Partitioning, e.g. k-Means
- Density Based Clustering, e.g. DBScan
- Grid Based Clustering

Hierarchical Clustering

Hierarchical clustering



In the two-dimensional MDS (Sammon mapping) representation of the Iris data set, two clusters can be identified. (The colours, indicating the species of the flowers, are ignored here.)

Hierarchical clustering

- Hierarchical clustering builds clusters step by step.

Hierarchical clustering

- Hierarchical clustering builds clusters step by step.
- Usually a bottom up strategy is applied by first considering each data object as a separate cluster and then step by step joining clusters together that are close to each other. This approach is called agglomerative hierarchical clustering.

Hierarchical clustering

- Hierarchical clustering builds clusters step by step.
- Usually a bottom up strategy is applied by first considering each data object as a separate cluster and then step by step joining clusters together that are close to each other. This approach is called agglomerative hierarchical clustering.
- In contrast to agglomerative hierarchical clustering, divisive hierarchical clustering starts with the whole data set as a single cluster and then divides clusters step by step into smaller clusters.

Hierarchical clustering

- Hierarchical clustering builds clusters step by step.
- Usually a bottom up strategy is applied by first considering each data object as a separate cluster and then step by step joining clusters together that are close to each other. This approach is called agglomerative hierarchical clustering.
- In contrast to agglomerative hierarchical clustering, divisive hierarchical clustering starts with the whole data set as a single cluster and then divides clusters step by step into smaller clusters.
- In order to decide which data objects should belong to the same cluster, a (dis-)similarity measure is needed.

Hierarchical clustering

- Hierarchical clustering builds clusters step by step.
- Usually a bottom up strategy is applied by first considering each data object as a separate cluster and then step by step joining clusters together that are close to each other. This approach is called agglomerative hierarchical clustering.
- In contrast to agglomerative hierarchical clustering, divisive hierarchical clustering starts with the whole data set as a single cluster and then divides clusters step by step into smaller clusters.
- In order to decide which data objects should belong to the same cluster, a (dis-)similarity measure is needed.
- Note: We do need to have access to features, all that is needed for hierarchical clustering is an $n \times n$-matrix $[d_{i,j}]$, where $d_{i,j}$ is the (dis-)similarity of data objects i and j. (n is the number of data objects.)

Hierarchical clustering: Dissimilarity matrix

The dissimilarity matrix $[d_{i,j}]$ should at least satisfy the following conditions.

- $d_{i,j} \geq 0$, i.e. dissimilarity cannot be negative.
- $d_{i,i} = 0$, i.e. each data object is completely similar to itself.
- $d_{i,j} = d_{j,i}$, i.e. data object i is (dis-)similar to data object j to the same degree as data object j is (dis-)similar to data object i.

It is often useful if the dissimilarity is a (pseudo-)metric, satisfying also the

- triangle inequality $d_{i,k} \leq d_{i,j} + d_{j,k}$.

Agglomerative hierarchical clustering: Algorithm

Input: $n \times n$ dissimilarity matrix $[d_{i,j}]$.

1. Start with n clusters, each data objects forms a single cluster.
2. Reduce the number of clusters by joining those two clusters that are most similar (least dissimilar).
3. Repeat step 3 until there is only one cluster left containing all data objects.

Measuring dissimilarity between clusters

- The dissimilarity between two clusters containing only one data object each is simply the dissimilarity of the two data objects specified in the dissimilarity matrix $[d_{i,j}]$.

Measuring dissimilarity between clusters

- The dissimilarity between two clusters containing only one data object each is simply the dissimilarity of the two data objects specified in the dissimilarity matrix $[d_{i,j}]$.

- But how do we compute the dissimilarity between clusters that contain more than one data object?

Measuring dissimilarity between clusters

Measuring dissimilarity between clusters

- **Centroid**
  Distance between the centroids (mean value vectors) of the two clusters[1]

[1] Requires that we can compute the mean vector!

Measuring dissimilarity between clusters

- **Centroid**
  Distance between the centroids (mean value vectors) of the two clusters[1]
- **Average Linkage**
  Average dissimilarity between all pairs of points of the two clusters.

[1] Requires that we can compute the mean vector!

Measuring dissimilarity between clusters

- **Centroid**
  Distance between the centroids (mean value vectors) of the two clusters[1]
- **Average Linkage**
  Average dissimilarity between all pairs of points of the two clusters.
- **Single Linkage**
  Dissimilarity between the two most similar data objects of the two clusters.

[1] Requires that we can compute the mean vector!

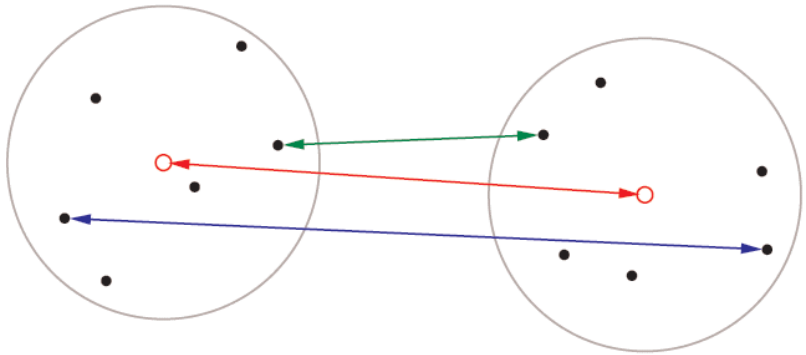
Measuring dissimilarity between clusters

- **Centroid**
  Distance between the centroids (mean value vectors) of the two clusters[1]
- **Average Linkage**
  Average dissimilarity between all pairs of points of the two clusters.
- **Single Linkage**
  Dissimilarity between the two most similar data objects of the two clusters.
- **Complete Linkage**
  Dissimilarity between the two most dissimilar data objects of the two clusters.

[1] Requires that we can compute the mean vector!

Measuring dissimilarity between clusters

- **Centroid**
  Distance between the centroids (mean value vectors) of the two clusters[1]
- **Average Linkage**
  Average dissimilarity between all pairs of points of the two clusters.
- **Single Linkage**
  Dissimilarity between the two most similar data objects of the two clusters.
- **Complete Linkage**
  Dissimilarity between the two most dissimilar data objects of the two clusters.

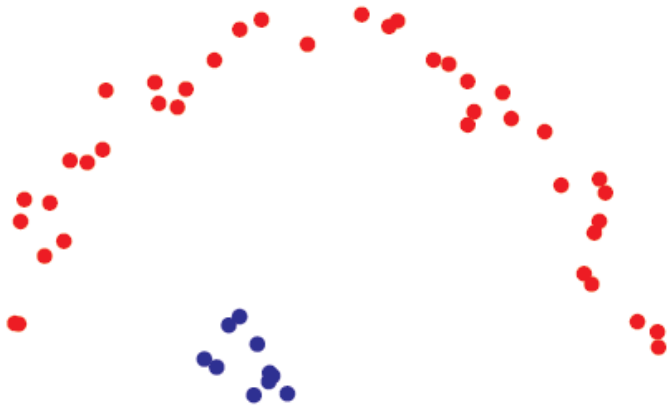


[1] Requires that we can compute the mean vector!

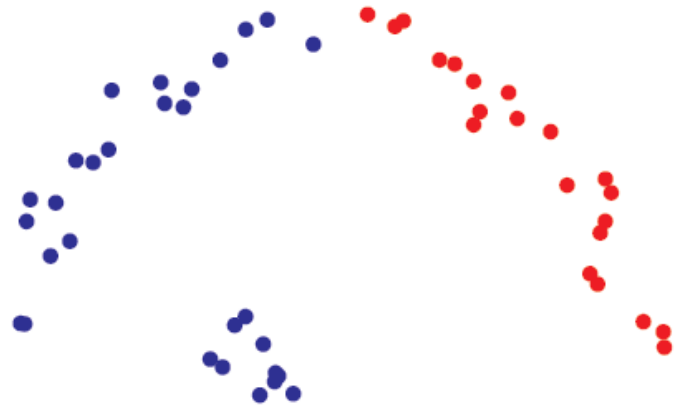
Measuring dissimilarity between clusters

- **Single linkage** can "follow chains" in the data (may be desirable in certain applications).
- **Complete linkage** leads to very compact clusters.
- **Average linkage** also tends clearly towards compact clusters.

Measuring dissimilarity between clusters



Single linkage



Complete linkage

Measuring dissimilarity between clusters

Ward's method

- another strategy for merging clusters
- In contrast to single, complete or average linkage, it takes the number of data objects in each cluster into account.

Measuring dissimilarity between clusters

The updated dissimilarity between the newly formed cluster $\{\mathcal{C} \cup \mathcal{C}'\}$ and the cluster $\mathcal{C}''$ is computed in the follwing way.

$$d'(\{\mathcal{C} \cup \mathcal{C}'\}, \mathcal{C}'') = ...$$

$$\begin{aligned}
\text{single linkage} &= \min\{d'(\mathcal{C}, \mathcal{C}''), d'(\mathcal{C}', \mathcal{C}'')\} \\
\text{complete linkage} &= \max\{d'(\mathcal{C}, \mathcal{C}''), d'(\mathcal{C}', \mathcal{C}'')\} \\
\text{average linkage} &= \frac{|\mathcal{C}| d'(\mathcal{C}, \mathcal{C}'') + |\mathcal{C}'| d'(\mathcal{C}', \mathcal{C}'')}{|\mathcal{C}| + |\mathcal{C}'|} \\
\text{Ward} &= \frac{(|\mathcal{C}| + |\mathcal{C}''|) d'(\mathcal{C}, \mathcal{C}'') + (|\mathcal{C}'| + |\mathcal{C}''|) d'(\mathcal{C}', \mathcal{C}'') - |\mathcal{C}''| d'(\mathcal{C}, \mathcal{C}')}{|\mathcal{C}| + |\mathcal{C}'| + |\mathcal{C}''|} \\
\text{centroid}^2 &= \frac{1}{|\mathcal{C} \cup \mathcal{C}'||\mathcal{C}''|} \sum_{\mathbf{x} \in \mathcal{C} \cup \mathcal{C}'} \sum_{\mathbf{y} \in \mathcal{C}''} d(\mathbf{x}, \mathbf{y})
\end{aligned}$$

[2] If metric, usually mean vector needs to be computed!

Dendrograms

- The cluster merging process arranges the data points in a binary tree.
- Draw the data tuples at the bottom or on the left
  (equally spaced if they are multi-dimensional).
- Draw a connection between clusters that are merged, with the distance to the data points representing the distance between the clusters.

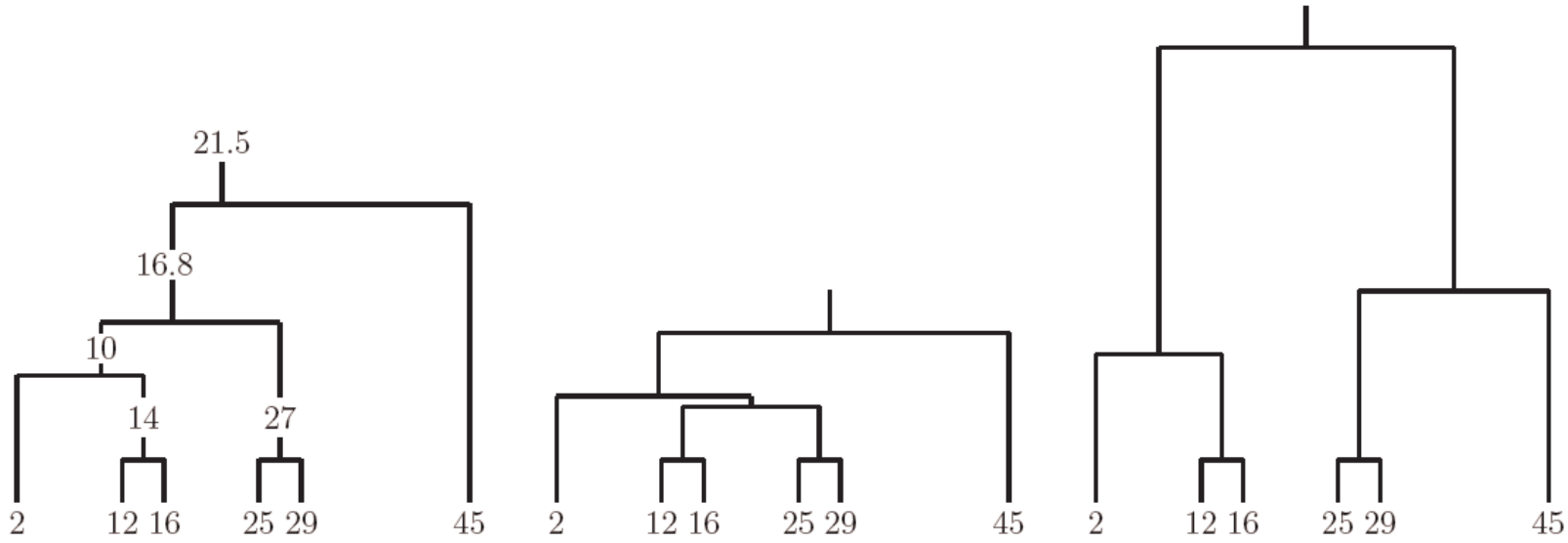
Hierarchical clustering

Example

Clustering of the 1-dimensional data set $\{2, 12, 16, 25, 29, 45\}$.

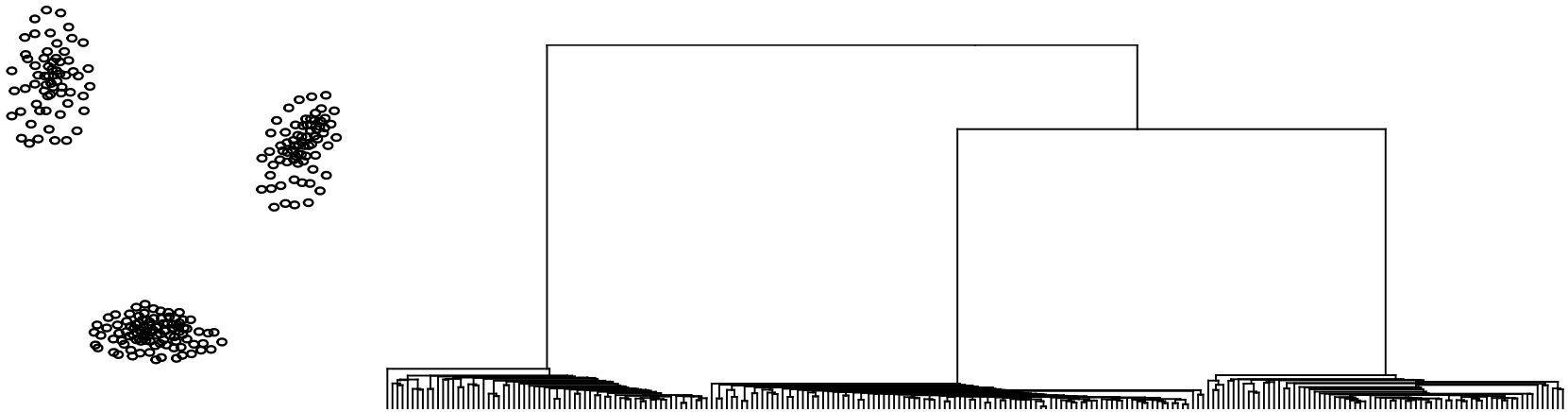
All three approaches to measure the distance between clusters lead to different dendrograms.

Hierarchical clustering

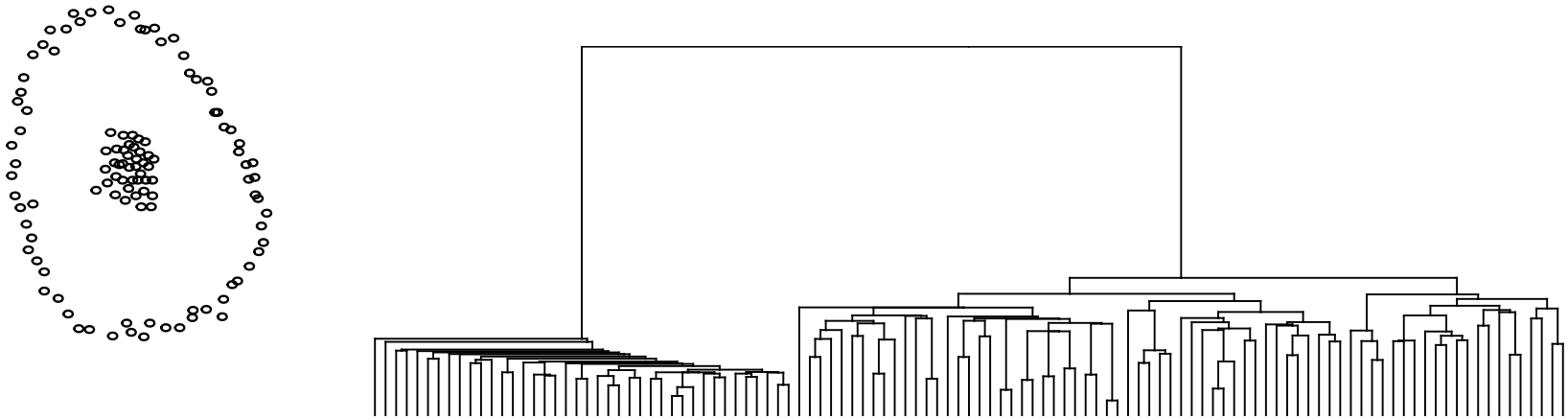


Centroid Single linkage Complete linkage

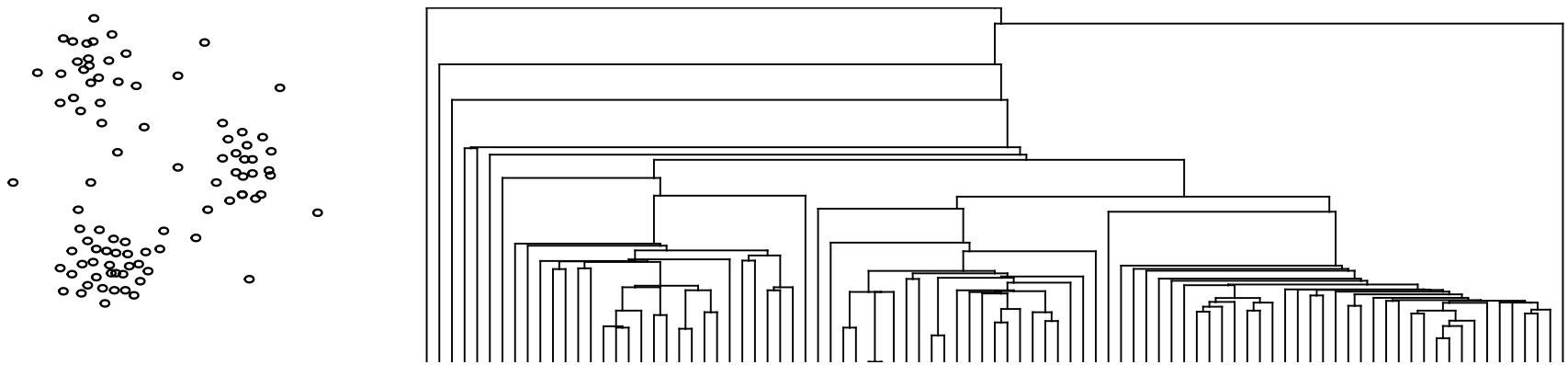
Dendrograms



Dendrograms



Dendrograms



Choosing the right clusters

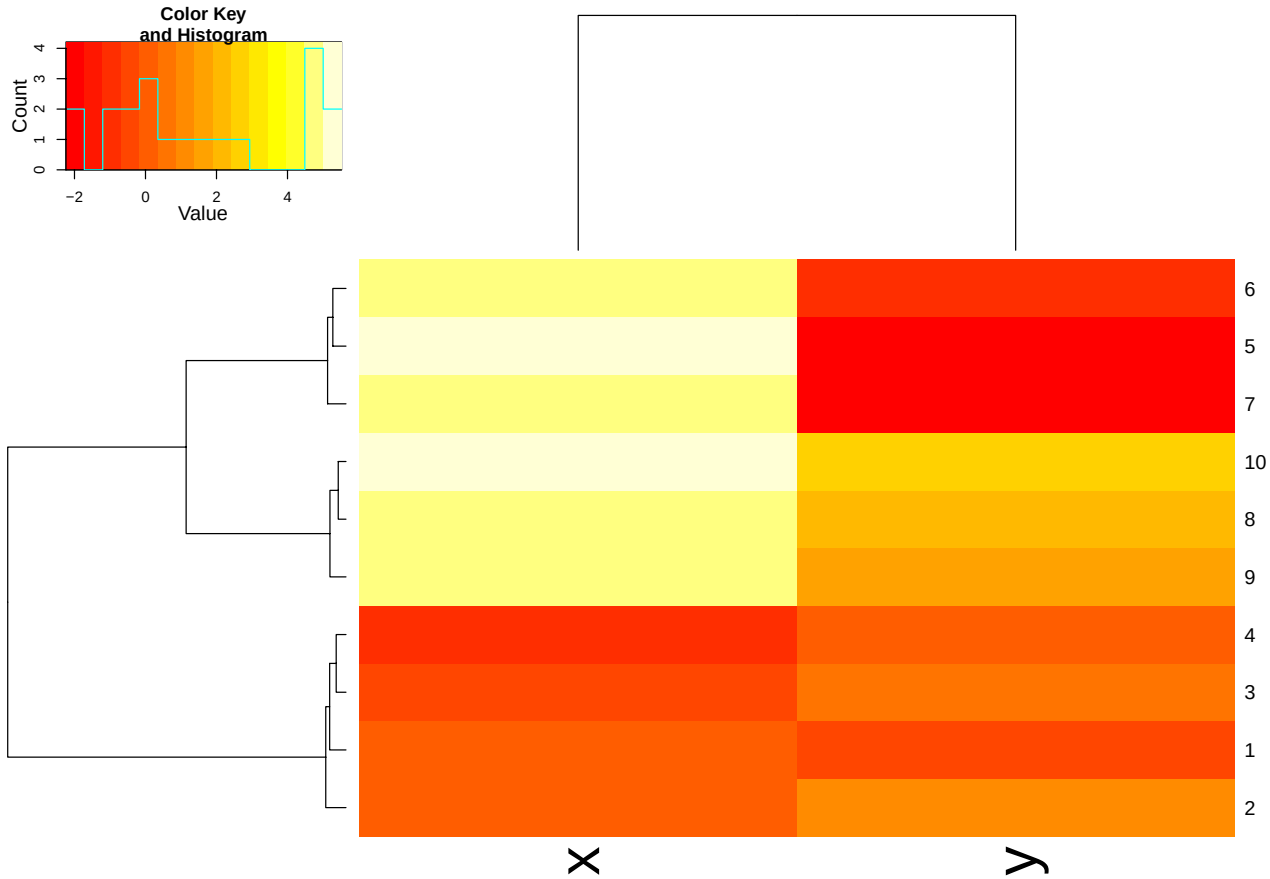
- **Simplest Approach:**
  - Specify a minimum desired distance between clusters.
  - Stop merging clusters if the closest two clusters are farther apart than this distance.
- **Visual Approach:**
  - Merge clusters until all data points are combined into one cluster.
  - Draw the dendrogram and find a good cut level.
  - Advantage: Cut needs not be strictly horizontal.
- **More Sophisticated Approaches:**
  - Analyze the sequence of distances in the merging process.
  - Try to find a step in which the distance between the two clusters merged is considerably larger than the distance of the previous step.
  - Several heuristic criteria exist for this step selection.

Heatmaps

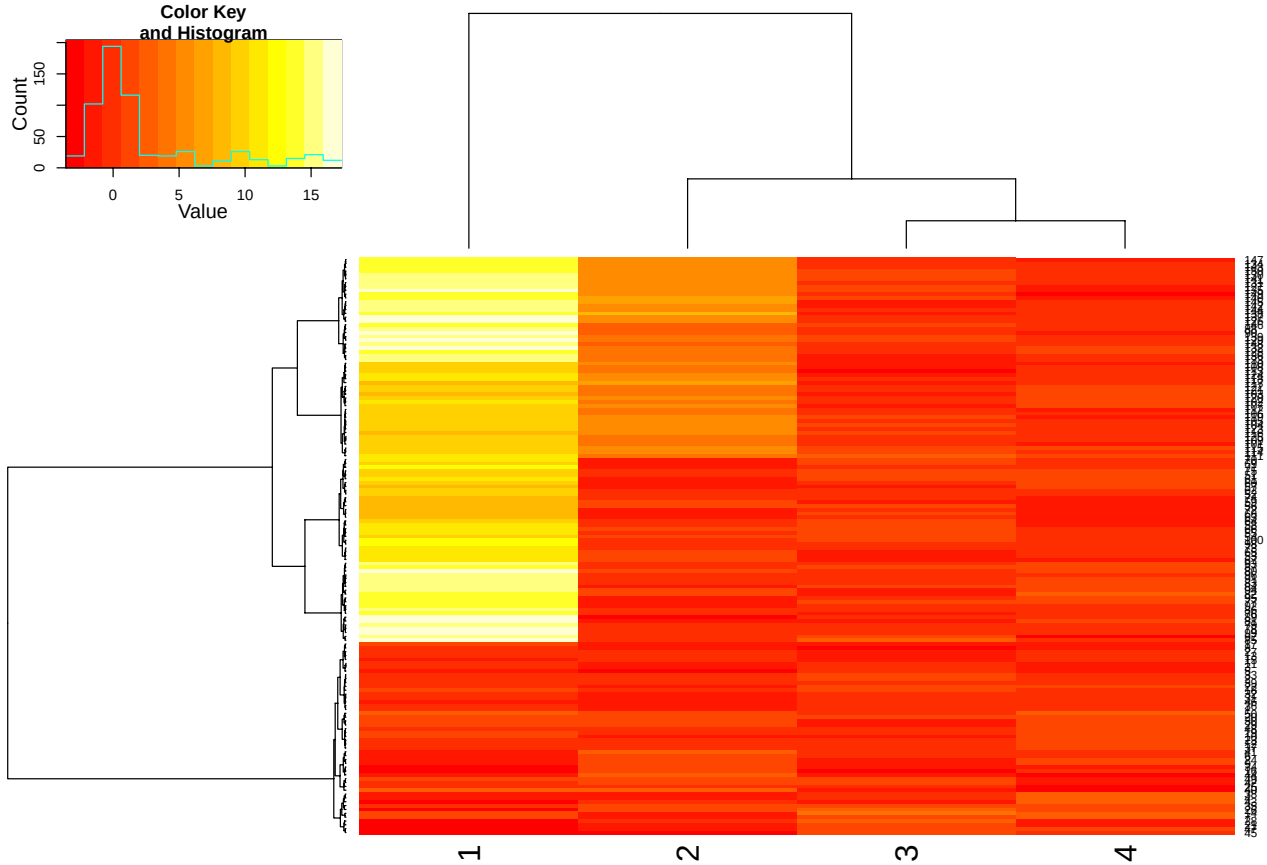
A heatmap combines

- a dendrogram resulting from clustering the data,
- a dendrogram resulting from clustering the attributes and
- colours to indicate the values of the attributes.

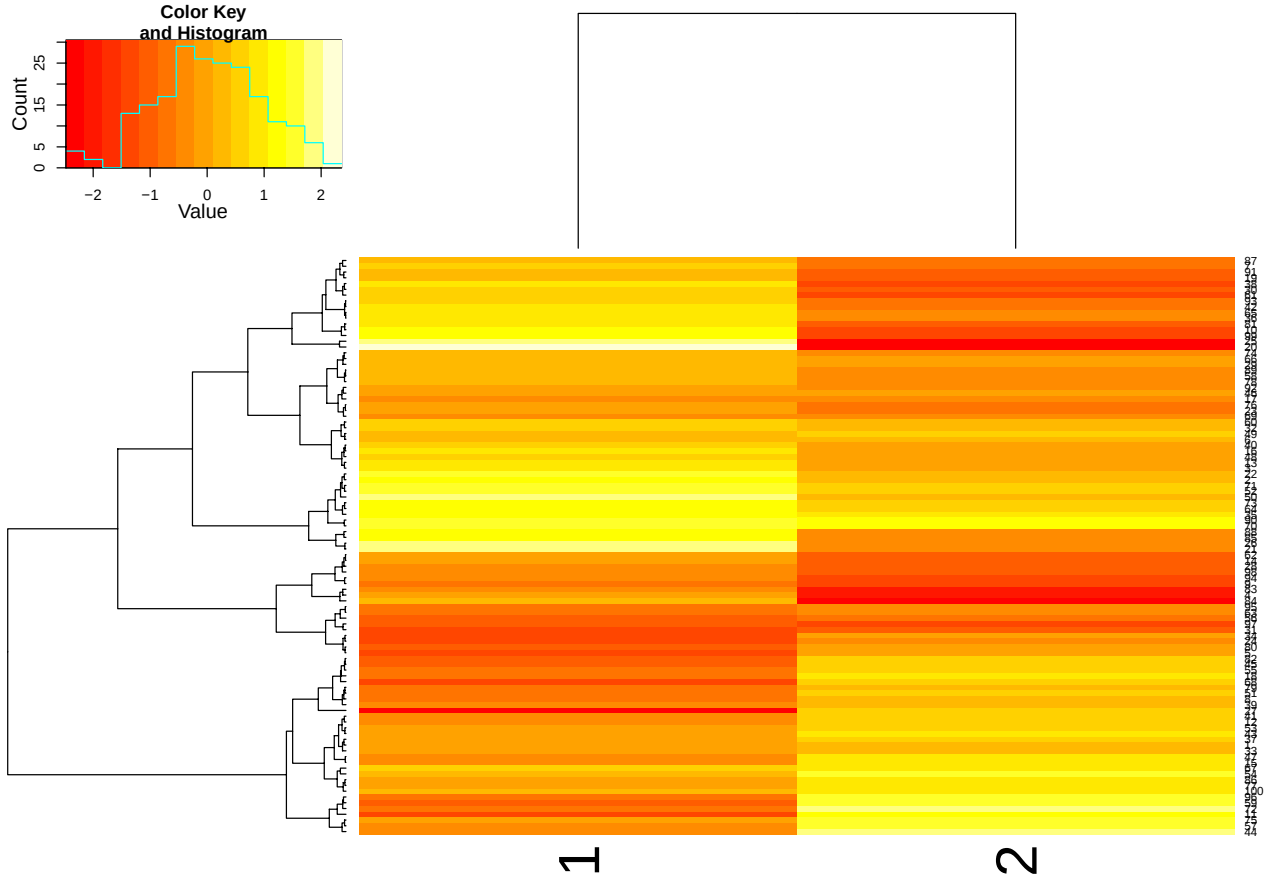
Example: Heatmap and dendrogram



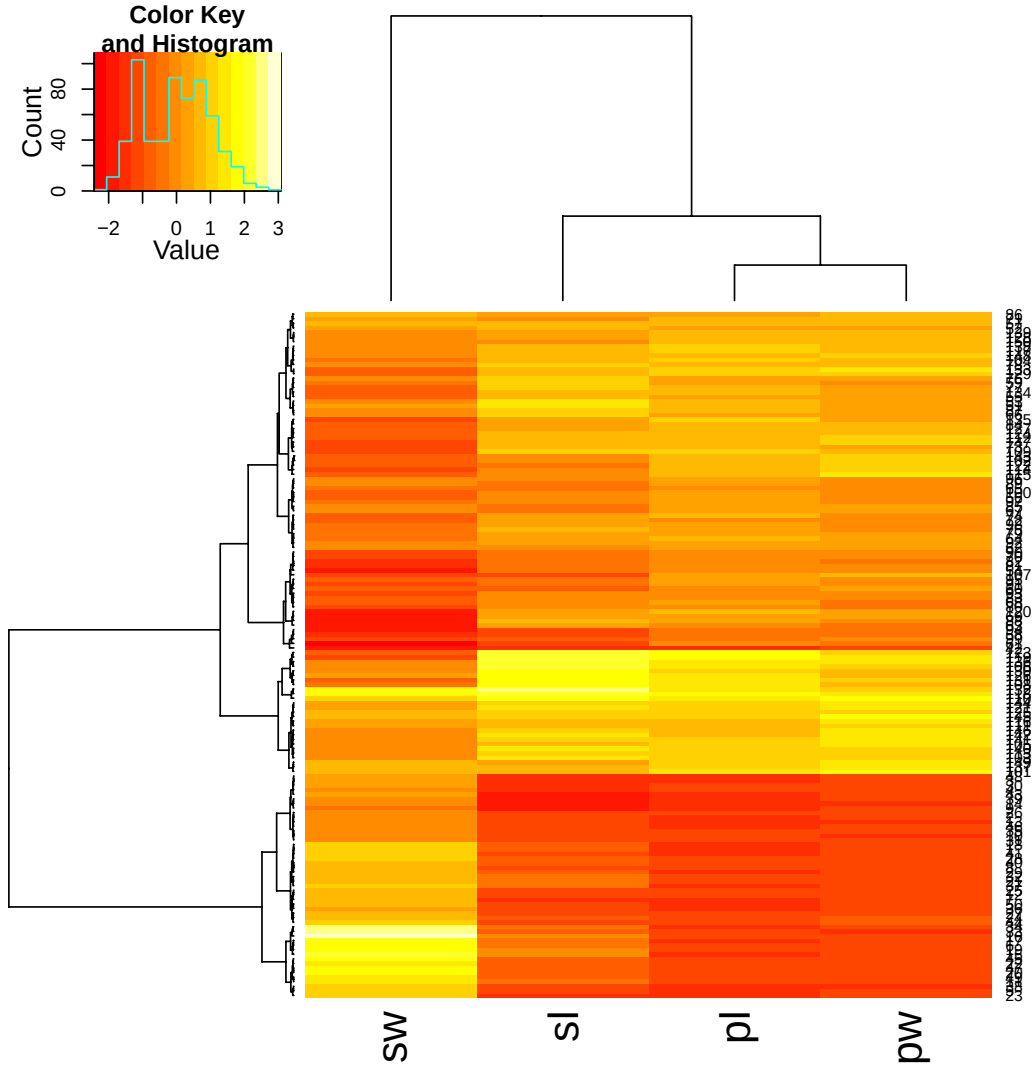
Example: Heatmap and dendrogram



Example: Heatmap and dendrogram



Iris Data: Heatmap and dendrogram



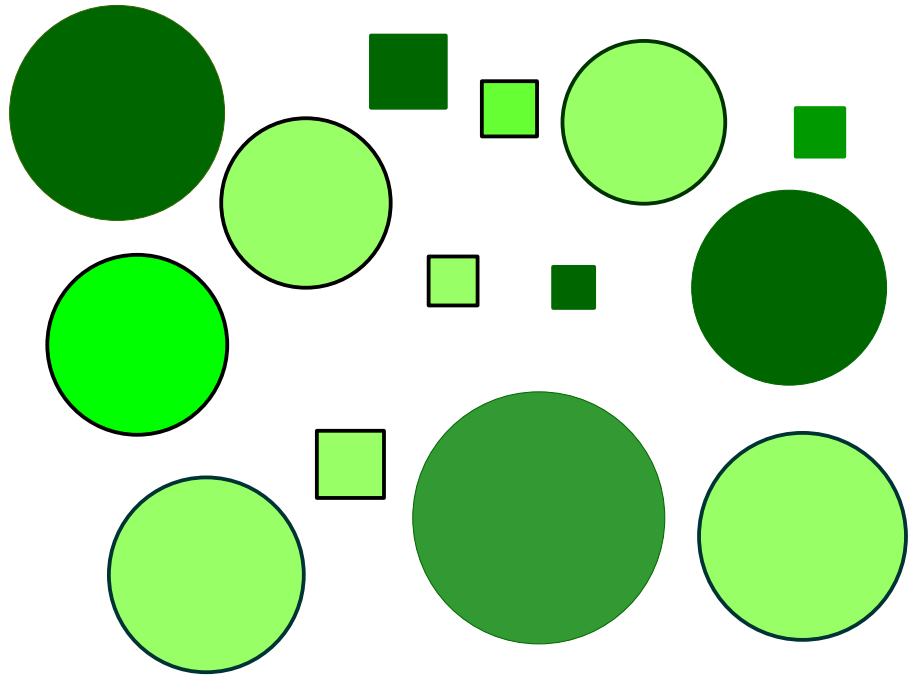
Divisive hierarchical clustering

The top-down approach of divisive hierarchical clustering is rarely used.

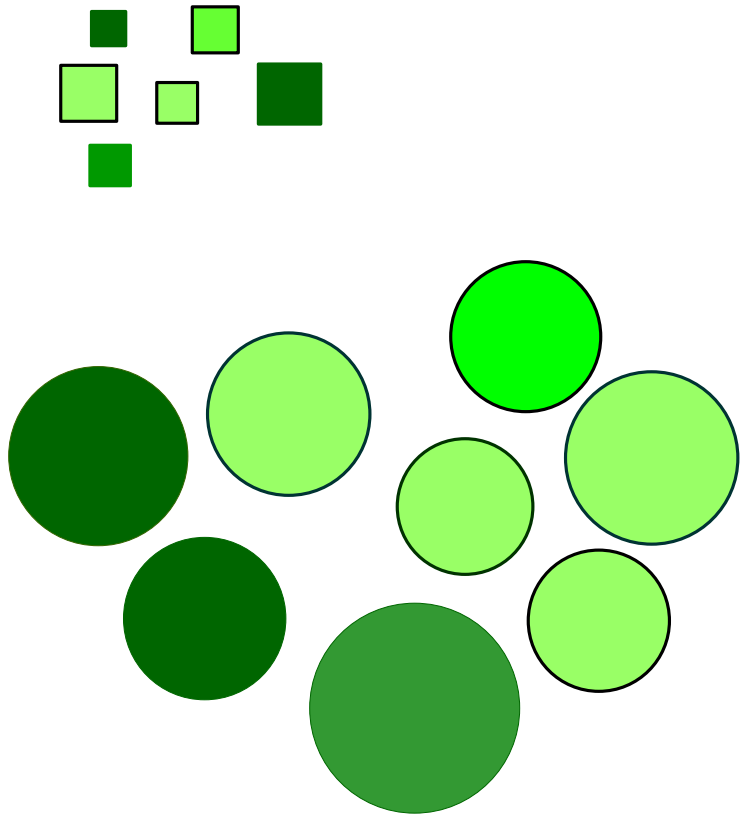
- In **agglomerative** clustering the minimum of the pairwise dissimilarities has to be determined, leading to a quadratic complexity in each step (quadratic in the number of clusters still present in the corresponding step).
- In **divisive** clustering for each cluster all possible splits would have to be considered.
- In the first step, there are $2^{n-1} - 1$ possible splits, where n is the number of data objects.

What is Similarity?

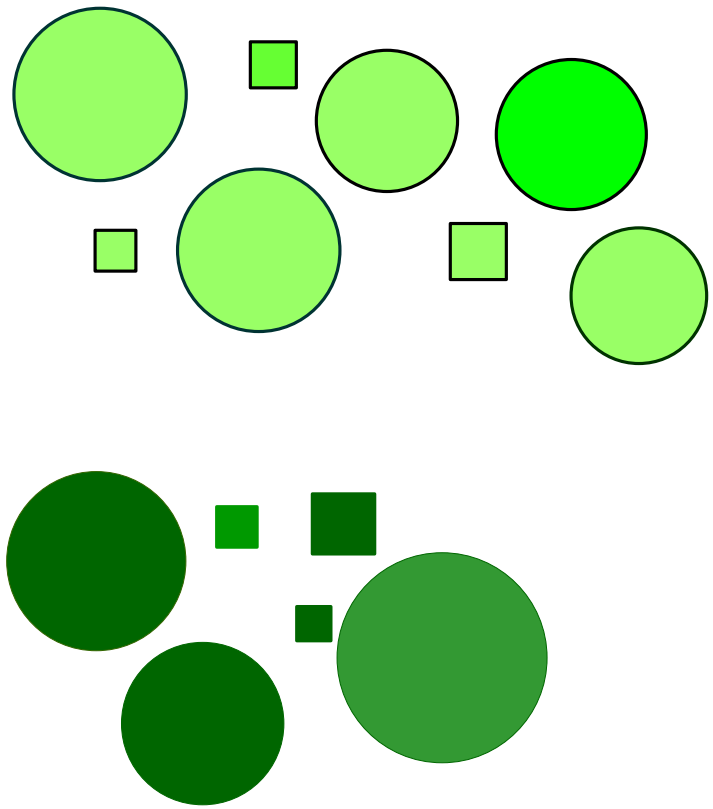
How to cluster these objects?



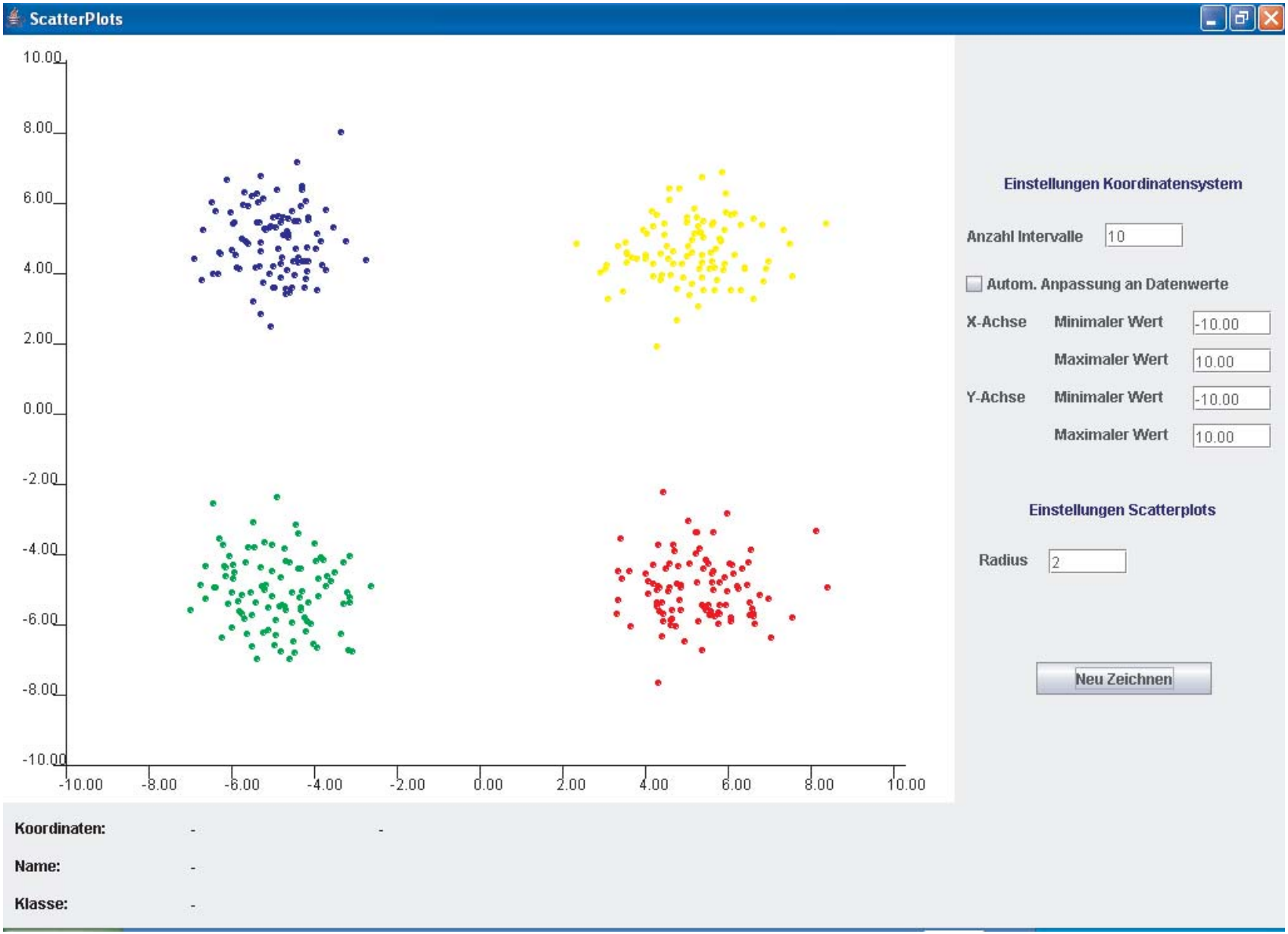
How to cluster these objects?



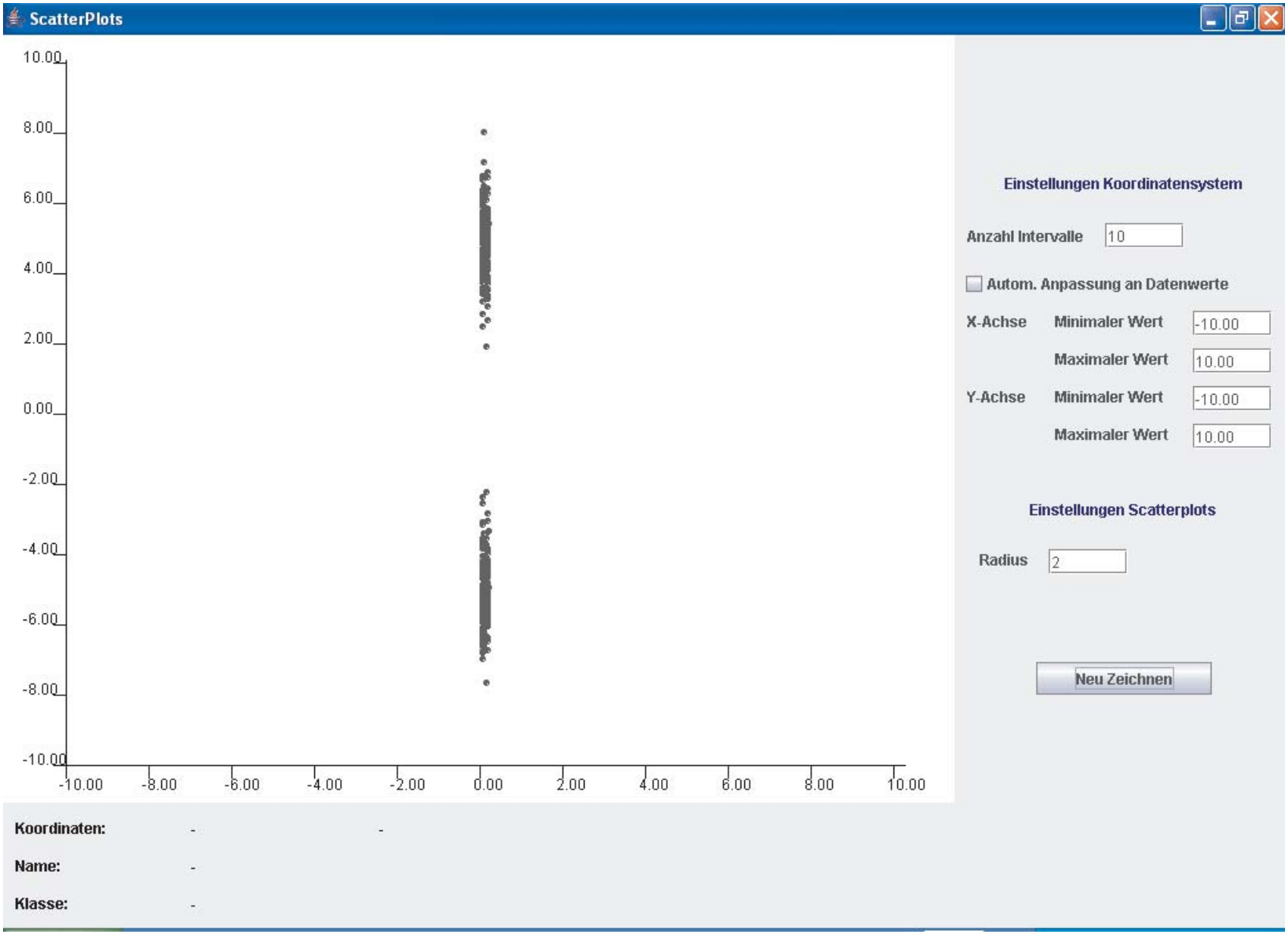
How to cluster these objects?



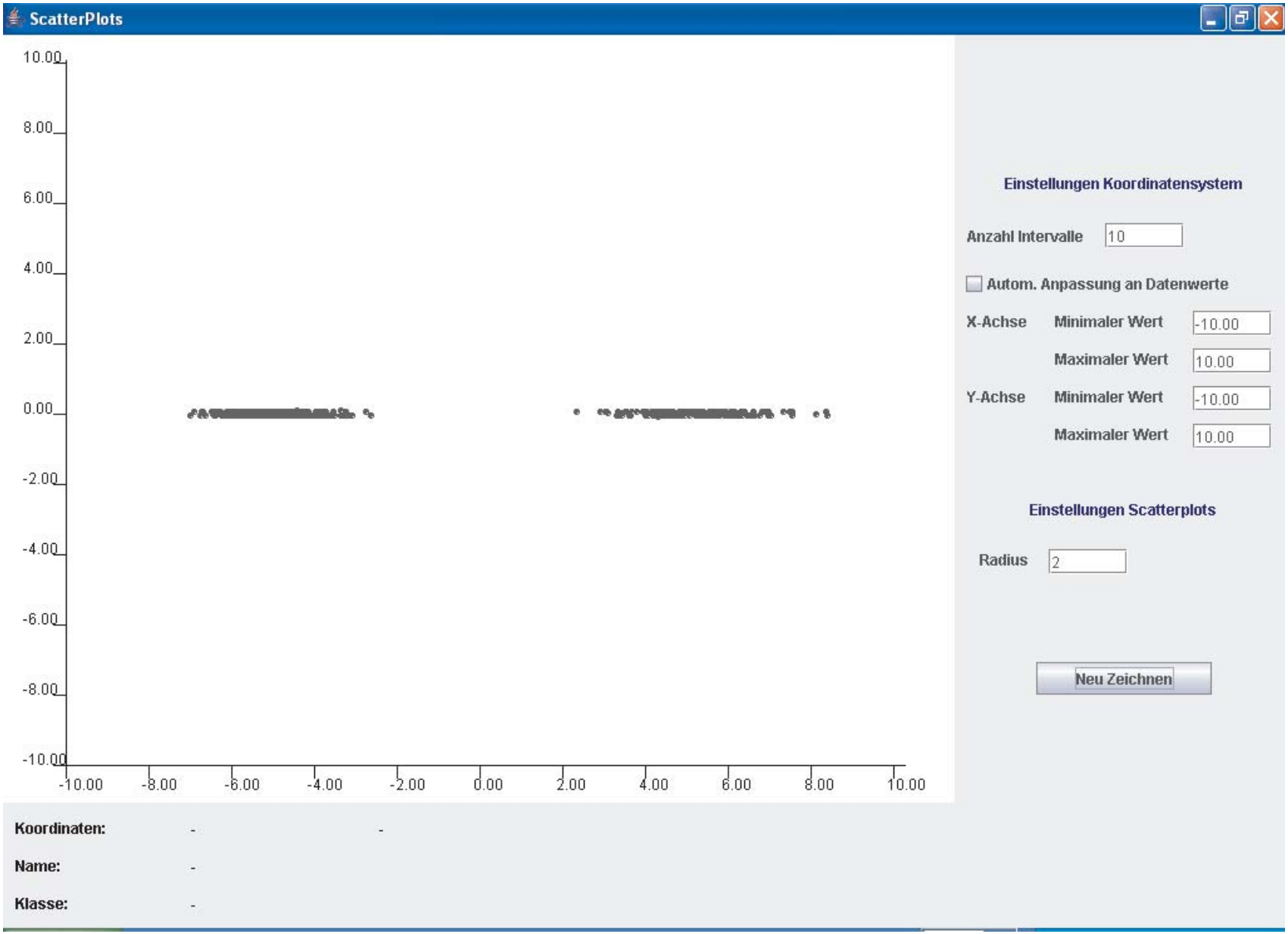
Clustering example



Clustering example



Clustering example



Scaling

The previous three slides show the same data set.

- In the second slide, the unit on the x-axis was changed to centi-units.
- In the third slide, the unit on the y-axis was changed to centi-units.

Scaling

The previous three slides show the same data set.

- In the second slide, the unit on the x-axis was changed to centi-units.
- In the third slide, the unit on the y-axis was changed to centi-units.

Clusters should not depend on the measurement unit!

Scaling

The previous three slides show the same data set.

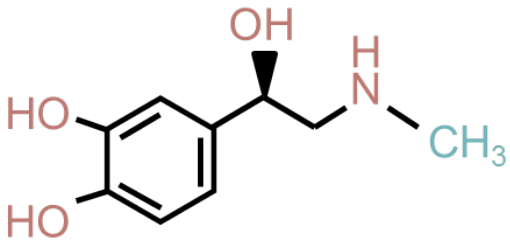
- In the second slide, the unit on the x-axis was changed to centi-units.
- In the third slide, the unit on the y-axis was changed to centi-units.

Clusters should not depend on the measurement unit!

Therefore, some kind of normalisation (see the chapter on data preparation) should be carried out before clustering.

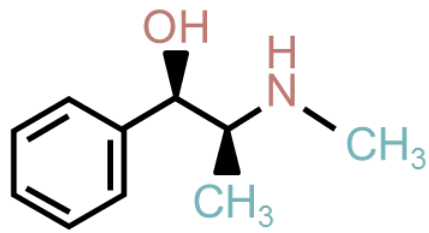
Complex Similarities: An Example

A few Adrenalin-like drug candidates:

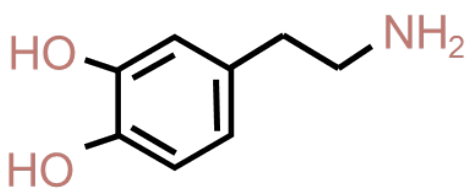


Adrenalin

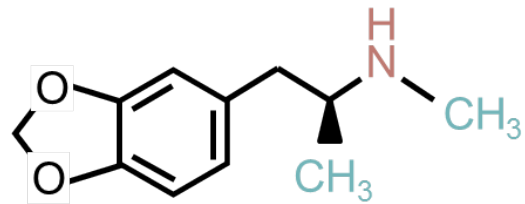
(D)



(C)



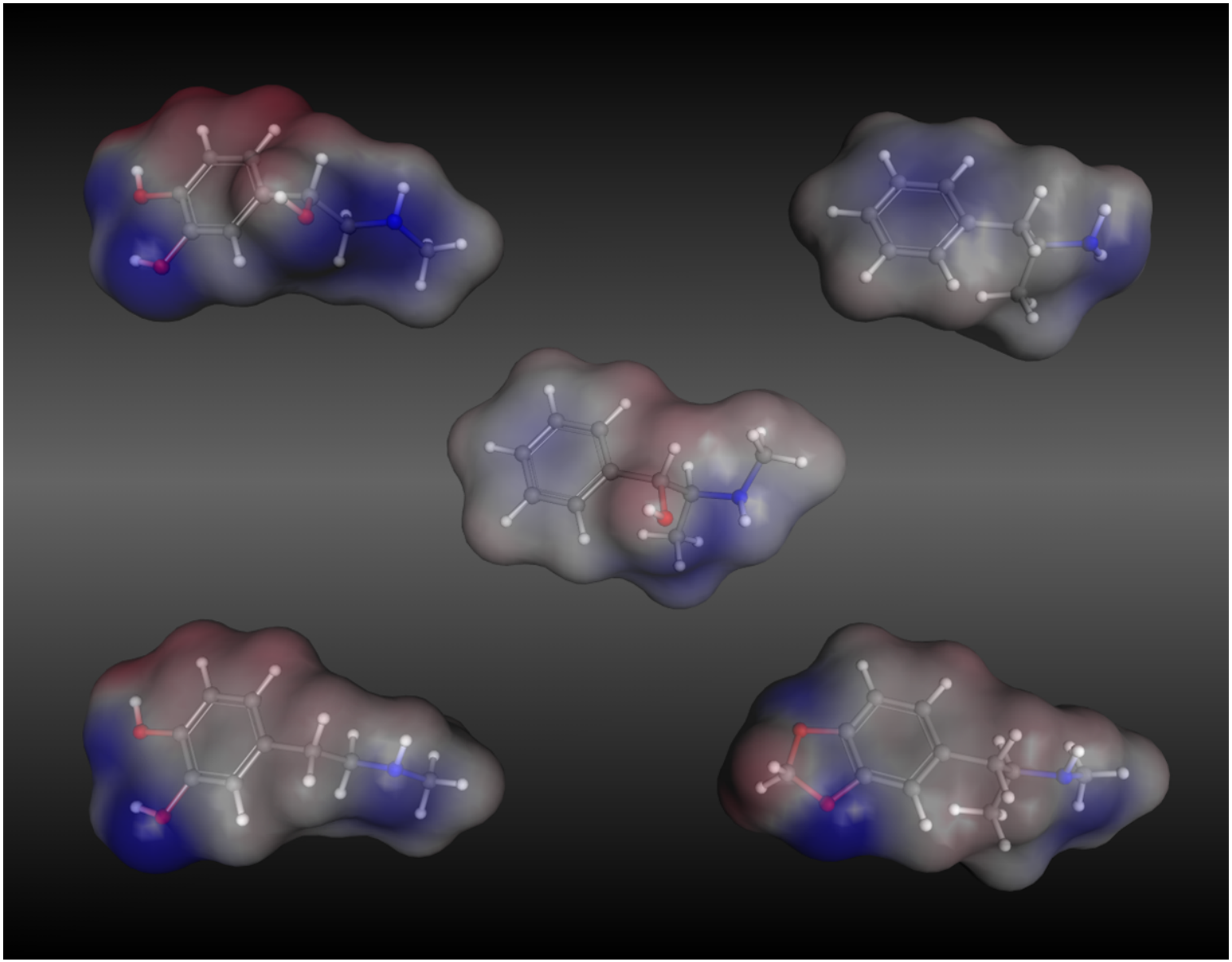
(B)



(E)

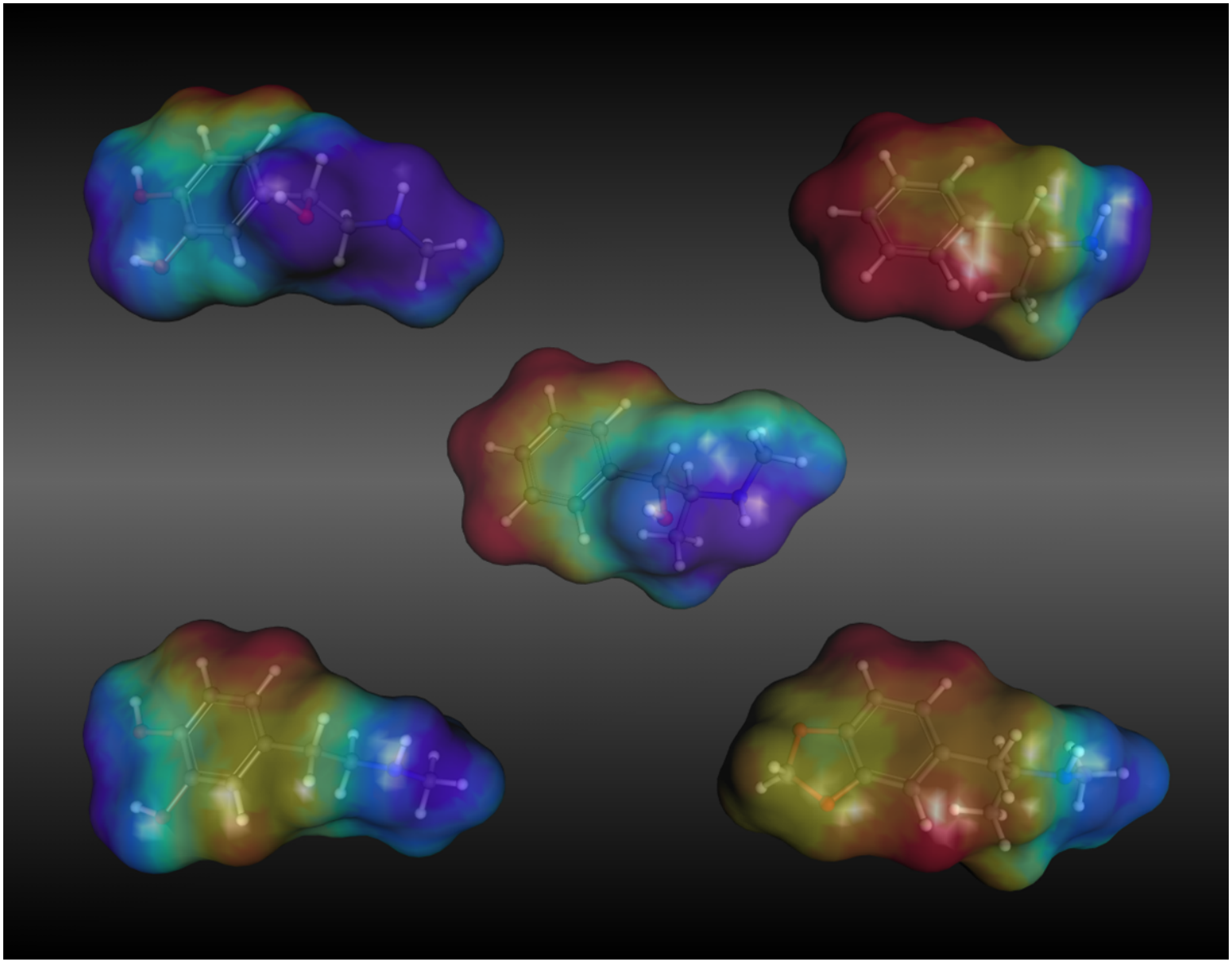
Complex Similarities: An Example

Similarity: Polarity



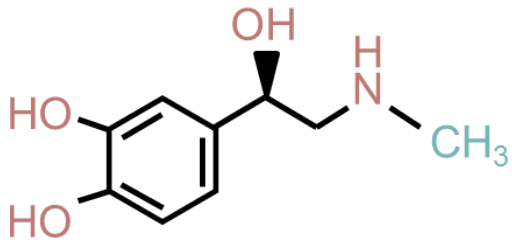
Complex Similarities: An Example

Dissimilarity: Hydrophobic / Hydrophilic

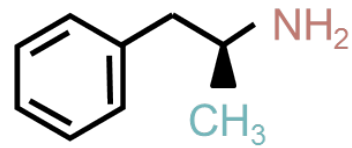


Complex Similarities: An Example

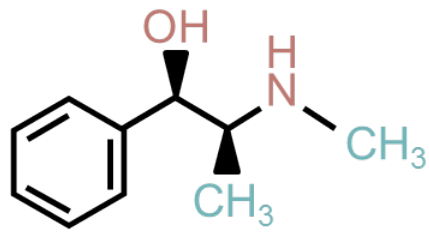
Similar to Adrenalin...



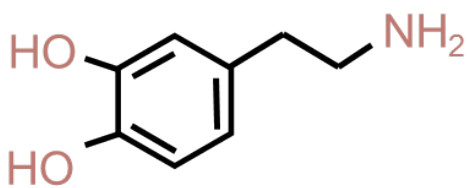
Adrenalin



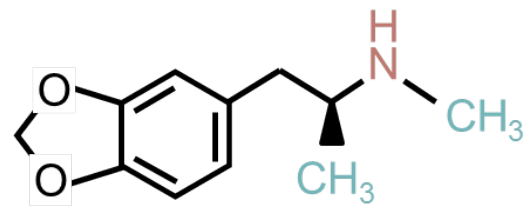
Amphetamin



Ephedrin



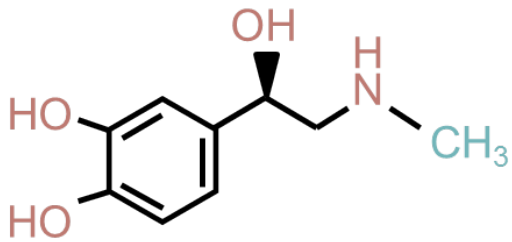
Dopamin



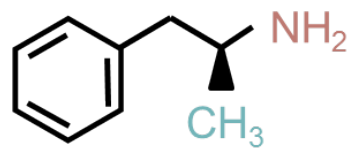
MDMA

Complex Similarities: An Example

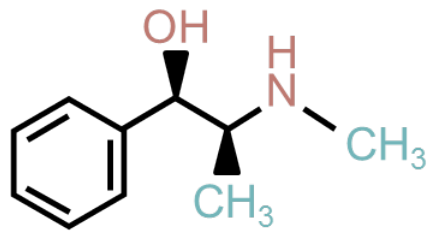
Similar to Adrenalin...but some cross the blood-brain barrier



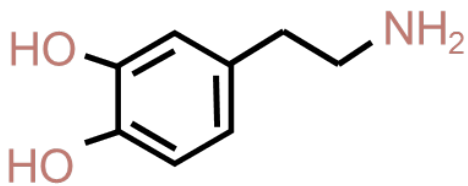
Adrenalin



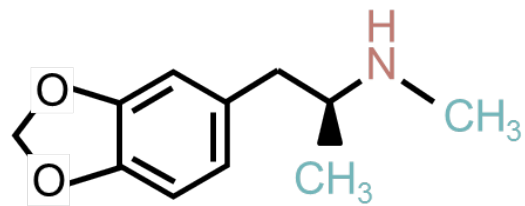
Amphetamin (Speed)



Ephedrin



Dopamin

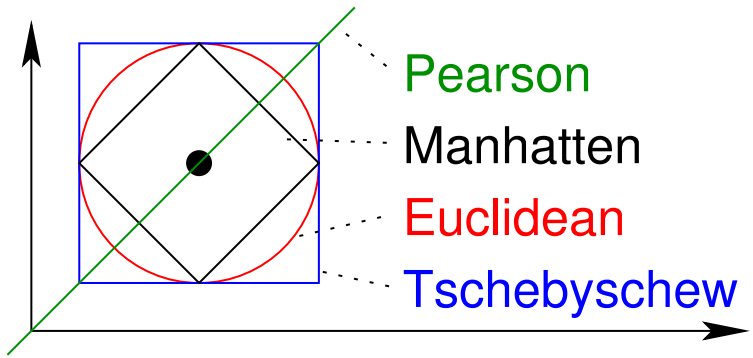


MDMA (Ecstasy)

Similarity Measures

Notion of (dis-)similarity: Numerical attributes

Various choices for dissimilarities between two numerical vectors:



Minkowksi	L_p	$d_p(x,y) = \sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$
Euclidean	L_2	$d_E(x,y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$
Manhattan	L_1	$d_M(x,y) = \lvert x_1 - y_1 \rvert + \ldots + \lvert x_n - y_n \rvert$
Tschebyschew	L_∞	$d_\infty(x,y) = \max\{\lvert x_1 - y_1 \rvert, \ldots, \lvert x_n - y_n \rvert\}$
Cosine		$d_C(x,y) = 1 - \frac{x^\top y}{\lVert x \rVert \lVert y \rVert}$
Tanimoto		$d_T(x,y) = \frac{x^\top y}{\lVert x \rVert^2 + \lVert y \rVert^2 - x^\top y}$
Pearson		Euclidean of z-score transformed $\mathbf{x}$, $\mathbf{y}$

Notion of (dis-)similarity: Binary attributes

The two values (e.g. 0 and 1) of a binary attribute can be interpreted as some property being absent (0) or present (1).

In this sense, a vector of binary attribute can be interpreted as a set of properties that the corresponding object has.

Example

- The binary vector $(0, 1, 1, 0, 1)$ corresponds to the set of properties $\{a_2, a_3, a_5\}$.
- The binary vector $(0, 0, 0, 0, 0)$ corresponds to the empty set.
- The binary vector $(1, 1, 1, 1, 1)$ corresponds to the set $\{a_1, a_2, a_3, a_4, a_5\}$.

Notion of (dis-)similarity: Binary attributes

Dissimilarity measures for two vectors of binary attributes.
Each data object is represent by the corresponding set of properties that are present.

	binary attributes	sets of properties
simple match	$d_S = 1 - \frac{b+n}{b+n+x}$	
Russel & Rao	$d_R = 1 - \frac{b}{b+n+x}$	$1 - \frac{\lvert X \cap Y \rvert}{\lvert \Omega \rvert}$
Jaccard	$d_J = 1 - \frac{b}{b+x}$	$1 - \frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Dice	$d_D = 1 - \frac{2b}{2b+x}$	$1 - \frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$

no. of predicates that...	
$b =$	...hold in both records
$n =$	...do not hold in both records
$x =$	...hold in only one of both records

x	y	set X	set Y	b	n	x	d_M	d_R	d_J	d_D
101000	111000	$\{a_1, a_3\}$	$\{a_1, a_2, a_3\}$	2	3	1	$0.1\bar{6}$	$0.6\bar{6}$	$0.3\bar{3}$	0.20

Notion of (dis-)similarity: Nominal attributes

Nominal attributes may be transformed into a set of binary attributes, each of them indicating one particular feature of the attribute (1-of-n coding).

Notion of (dis-)similarity: Nominal attributes

Nominal attributes may be transformed into a set of binary attributes, each of them indicating one particular feature of the attribute (1-of-n coding).

Example

Attribute *Manufacturer* with the values *BMW*, *Chrysler*, *Dacia*, *Ford*, *Volkswagen*.

manufacturer	...
Volkswagen	...
Dacia	...
Ford	...

$\rightarrow$

binary vector
00001
01000
00100

Then one of the dissimilarity measures for binary attribute can be applied.

Notion of (dis-)similarity: Nominal attributes

Nominal attributes may be transformed into a set of binary attributes, each of them indicating one particular feature of the attribute (1-of-n coding).

Example

Attribute *Manufacturer* with the values *BMW*, *Chrysler*, *Dacia*, *Ford*, *Volkswagen*.

manufacturer	...
Volkswagen	...
Dacia	...
Ford	...

$\rightarrow$

binary vector
00001
01000
00100

Then one of the dissimilarity measures for binary attribute can be applied.

Another way to measure similarity between two vectors of nominal attributes is to compute the proportion of attributes where both vectors have the same value, leading to the Russel & Rao dissimilarity measure.

Prototype Based Clustering

Prototype Based Clustering

- **given:** dataset of size n
- **return:** set of typical examples of size $k << n$.

k-Means clustering

- Choose a number k of clusters to be found (user input).
- Initialize the cluster centres randomly
  (for instance, by randomly selecting k data points).
- **Data point assignment**:
  Assign each data point to the cluster centre that is closest to it (i.e. closer than any other cluster centre).
- **Cluster centre update**:
  Compute new cluster centres as the mean vectors of the assigned data points. (Intuitively: centre of gravity if each data point has unit weight.)

k-Means clustering

- Repeat these two steps (data point assignment and cluster centre update) until the clusters centres do not change anymore.
- It can be shown that this scheme must converge,
  i.e., the update of the cluster centres cannot go on forever.

k-Means clustering

Aim: Minimize the objective function

$$f = \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} d_{ij}$$

under the constraints $u_{ij} \in \{0, 1\}$ and

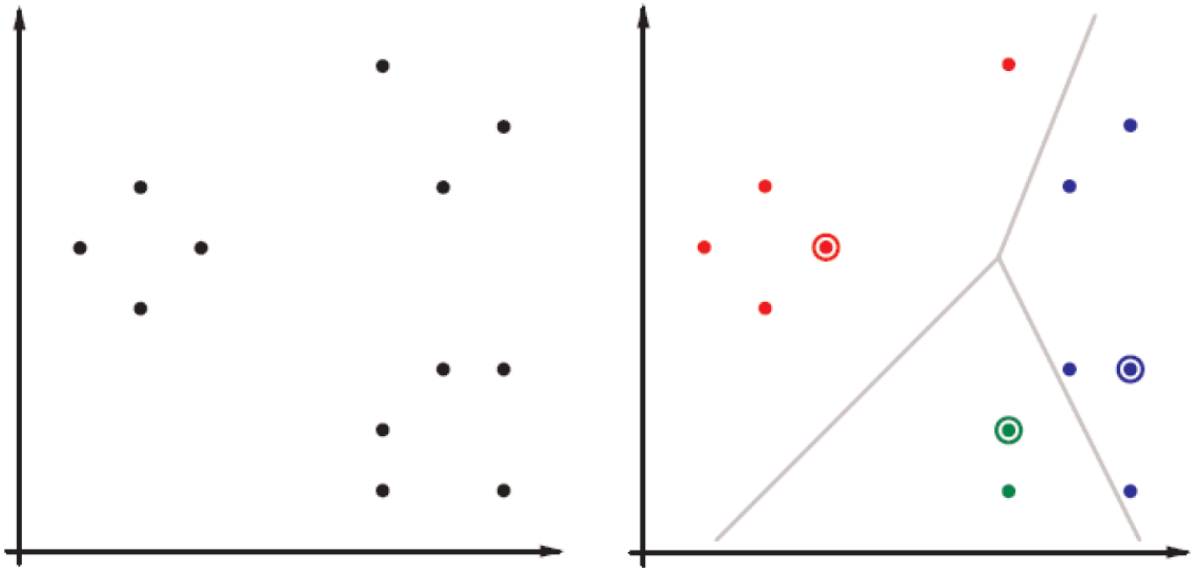
$$\sum_{i=1}^{k} u_{ij} = 1 \quad \text{for all } j = 1, \ldots, n.$$

Alternating optimization

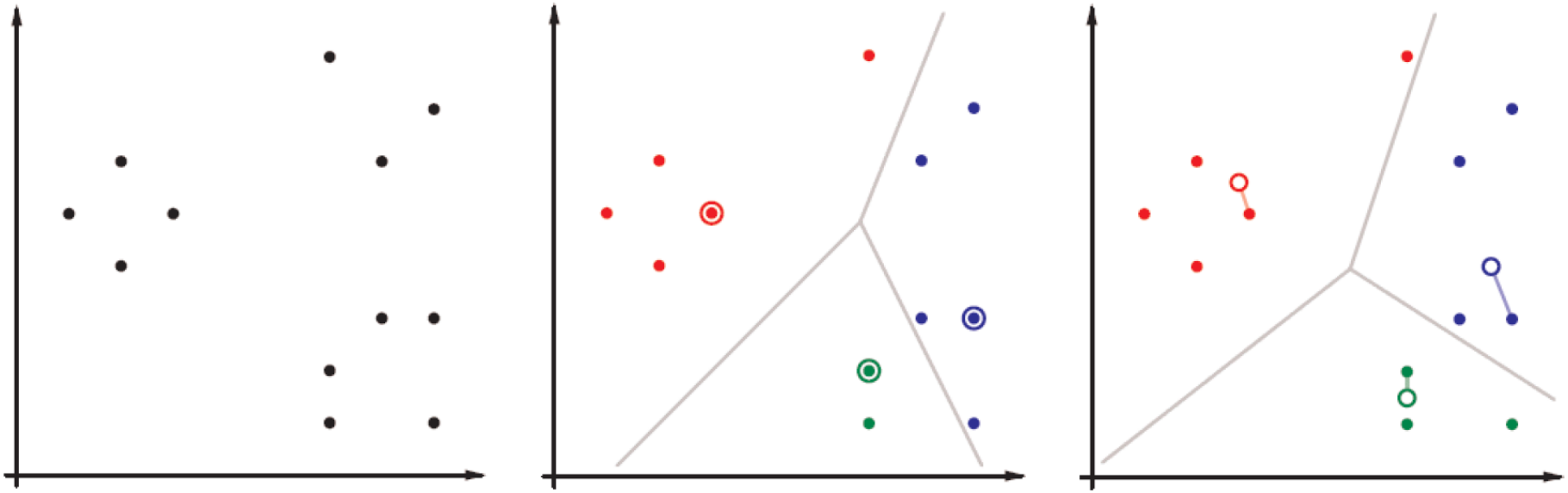
- Assuming the cluster centres to be fixed, $u_{ij} = 1$ should be chosen for the cluster i to which data object x_j has the smallest distance in order to minimize the objective function.
- Assuming the assignments to the clusters to be fixed, each cluster centre should be chosen as the mean vector of the data objects assigned to the cluster in order to minimize the objective function.

k-Means clustering: Example

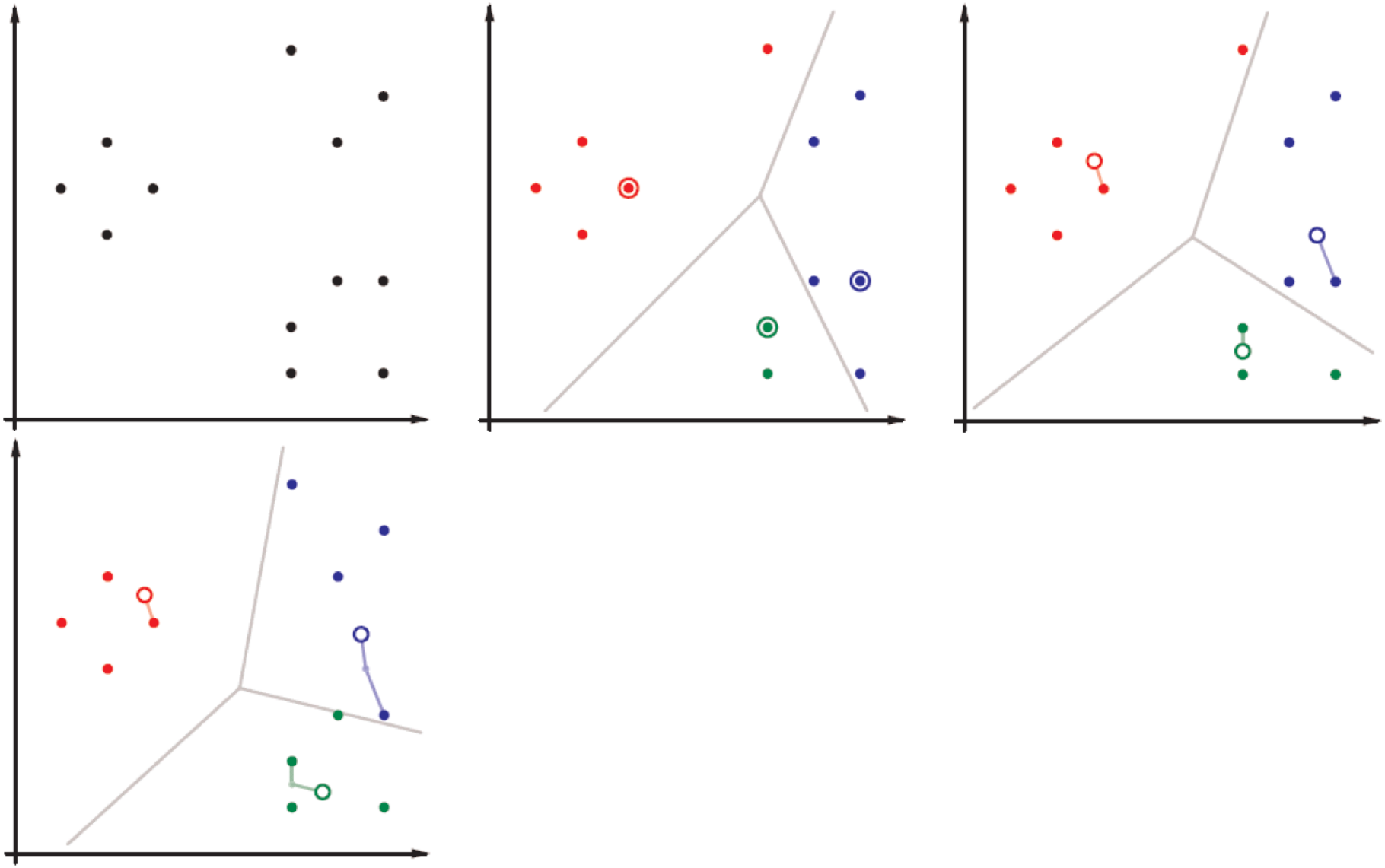
k-Means clustering: Example



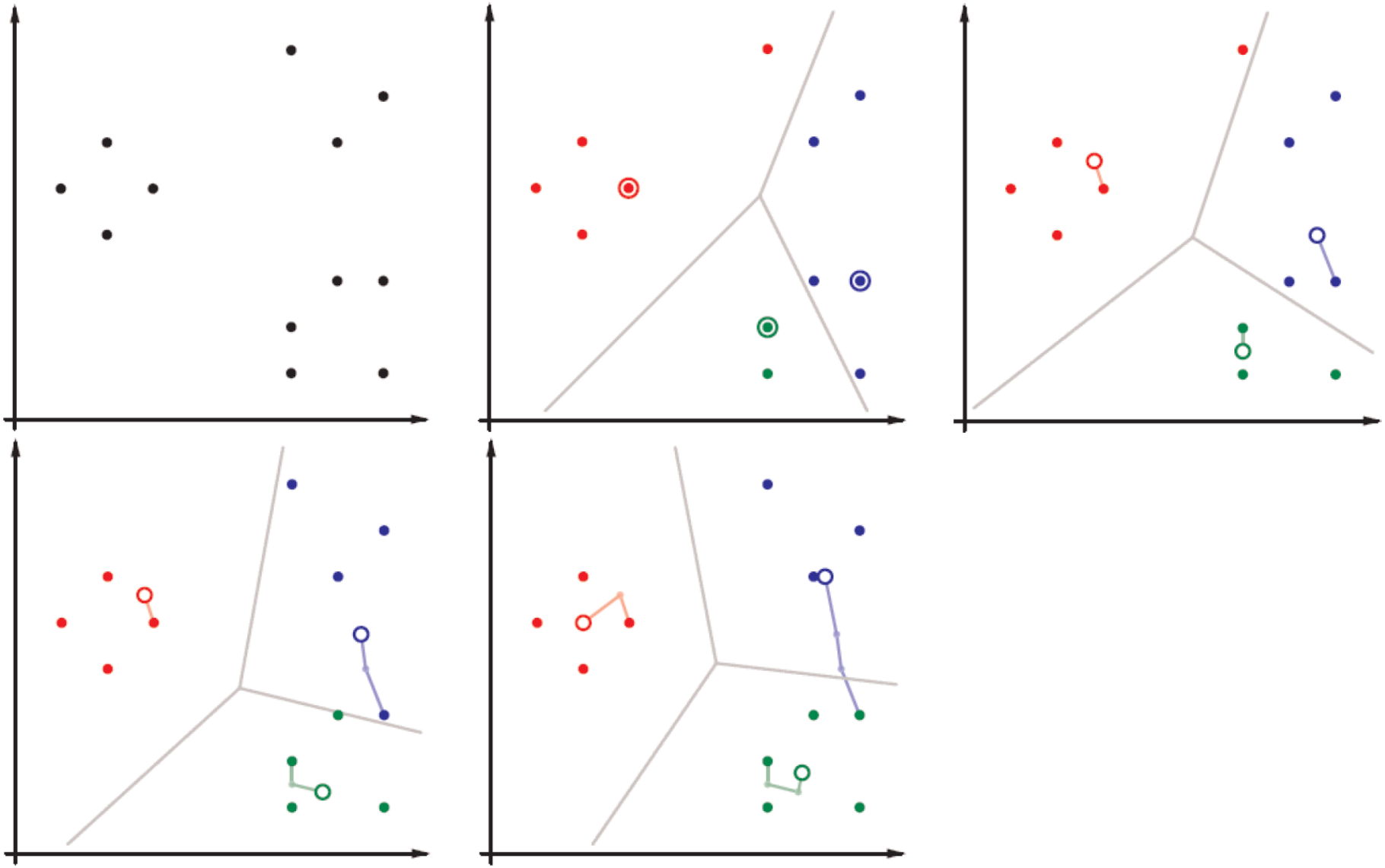
k-Means clustering: Example



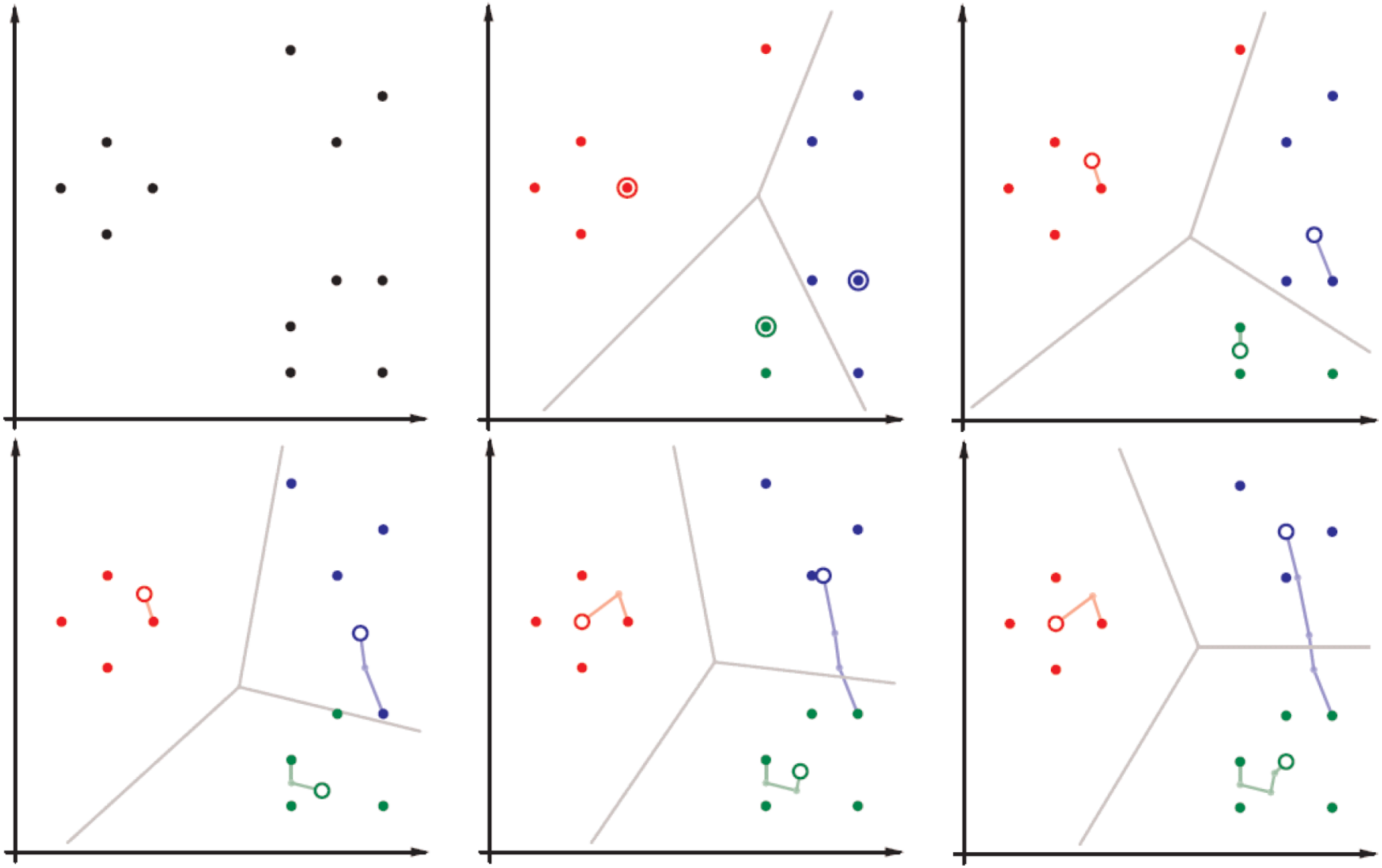
k-Means clustering: Example



k-Means clustering: Example



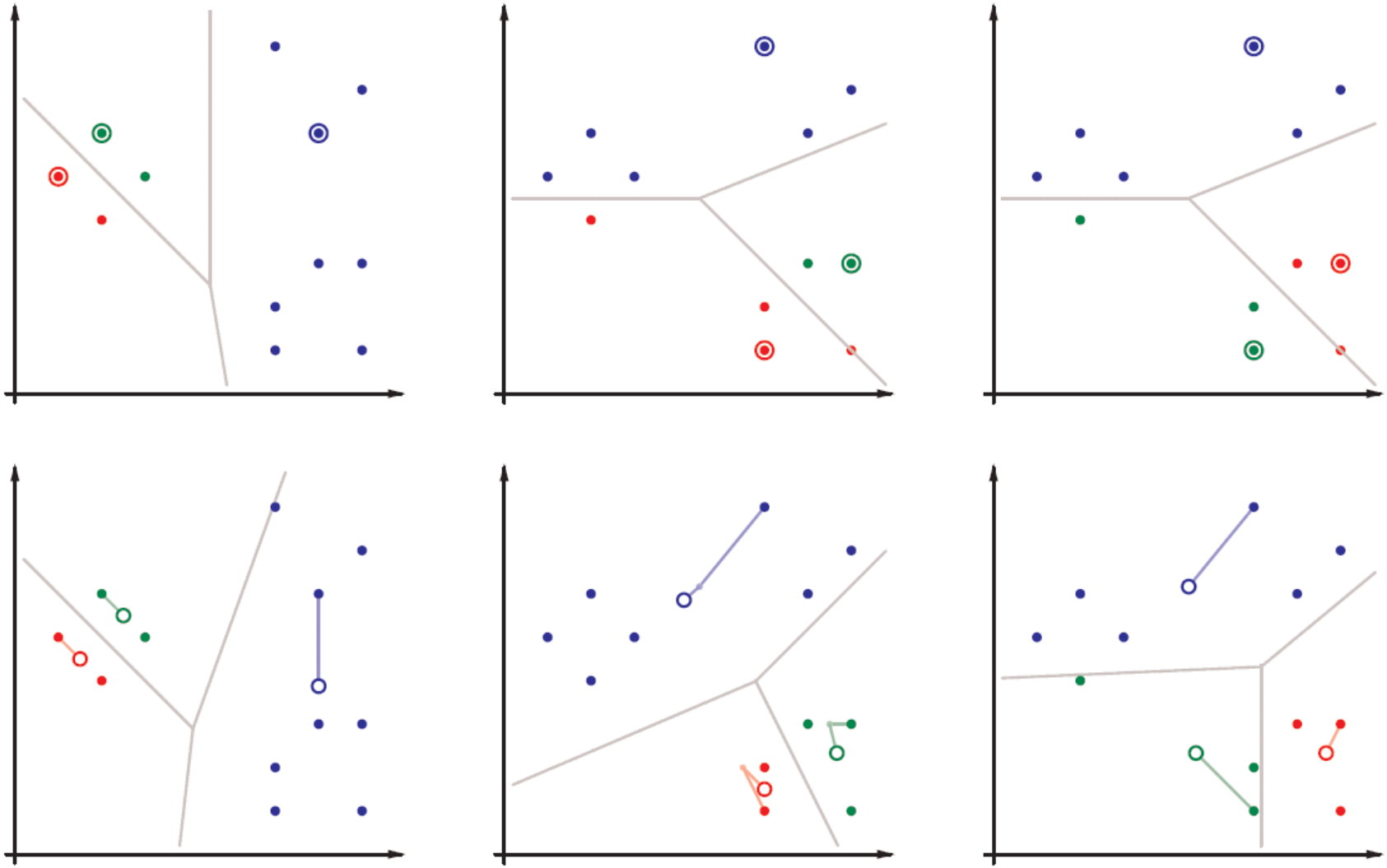
k-Means clustering: Example



k-Means clustering: Local minima

- Clustering is successful in this example:
  The clusters found are those that would have been formed intuitively.
- Convergence is achieved after only 5 steps.
  (This is typical: convergence is usually very fast.)
- However: The clustering result is fairly **sensitive to the initial positions** of the cluster centres.
- With a bad initialisation clustering may fail
  (the alternating update process gets stuck in a local minimum).

k-Means clustering: Local minima

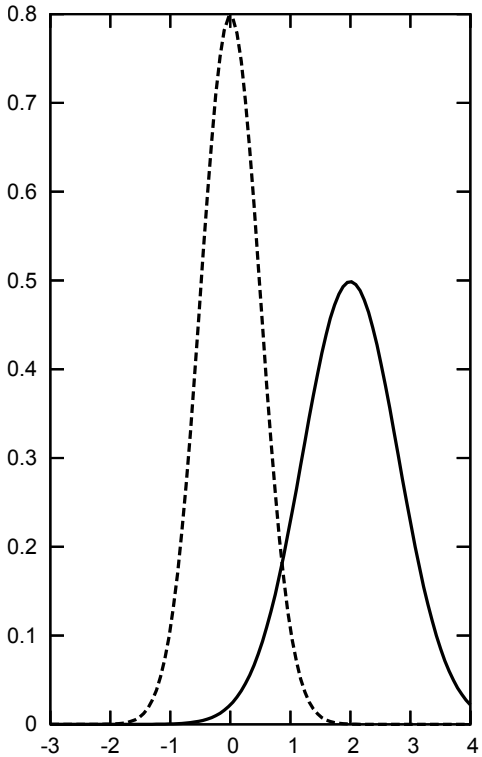


Gaussian Mixture Models

Gaussian mixture models – EM clustering

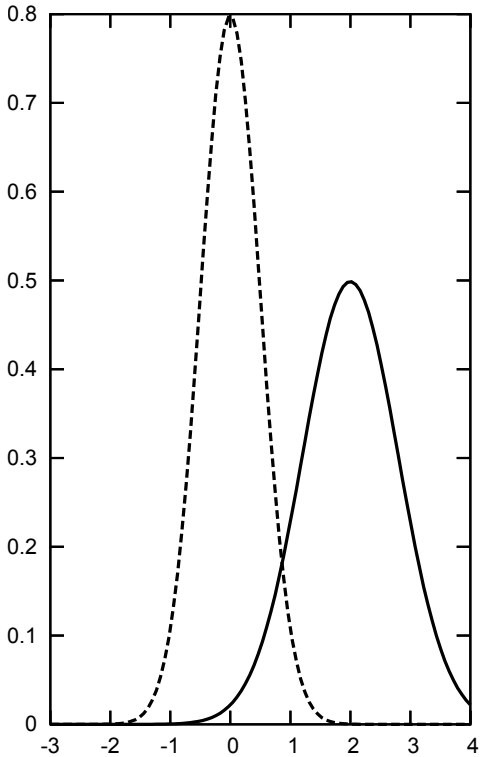
- **Assumption:** Data was generated by sampling a set of normal distributions.
  (The probability density is a mixture of normal distributions.)
- **Aim:** Find the parameters for the normal distributions and how much each normal distribution contributes to the data.

Gaussian mixture models

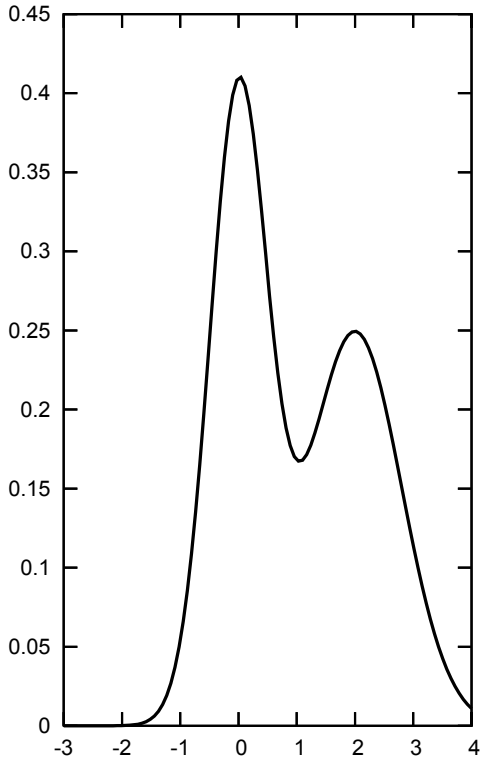


Two normal distributions.

Gaussian mixture models

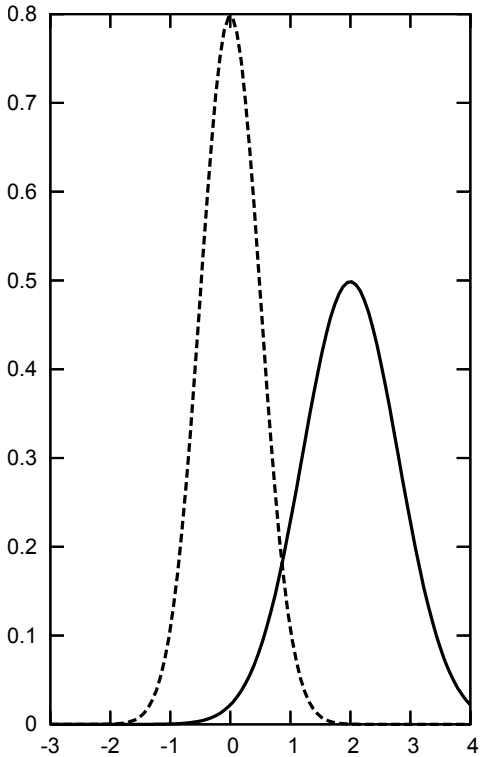


Two normal distributions.

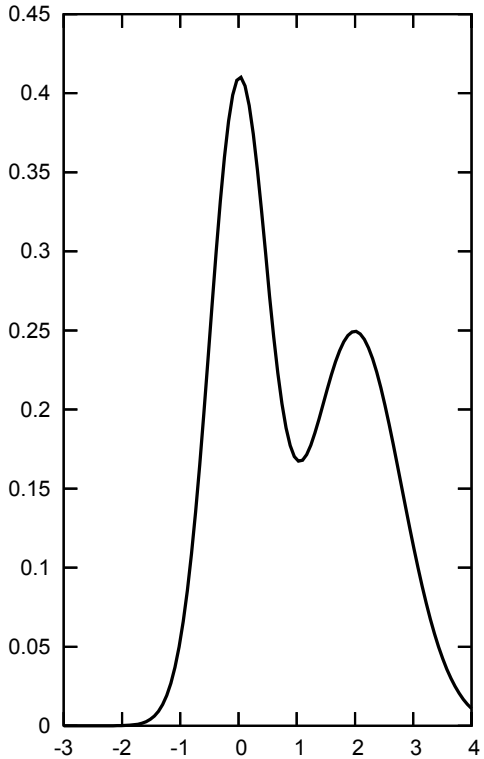


Mixture model (both normal distrubutions contribute 50%).

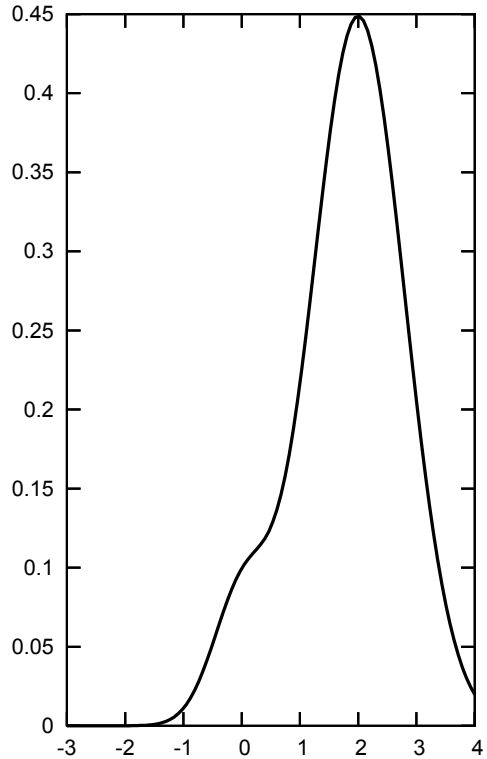
Gaussian mixture models



Two normal distributions.

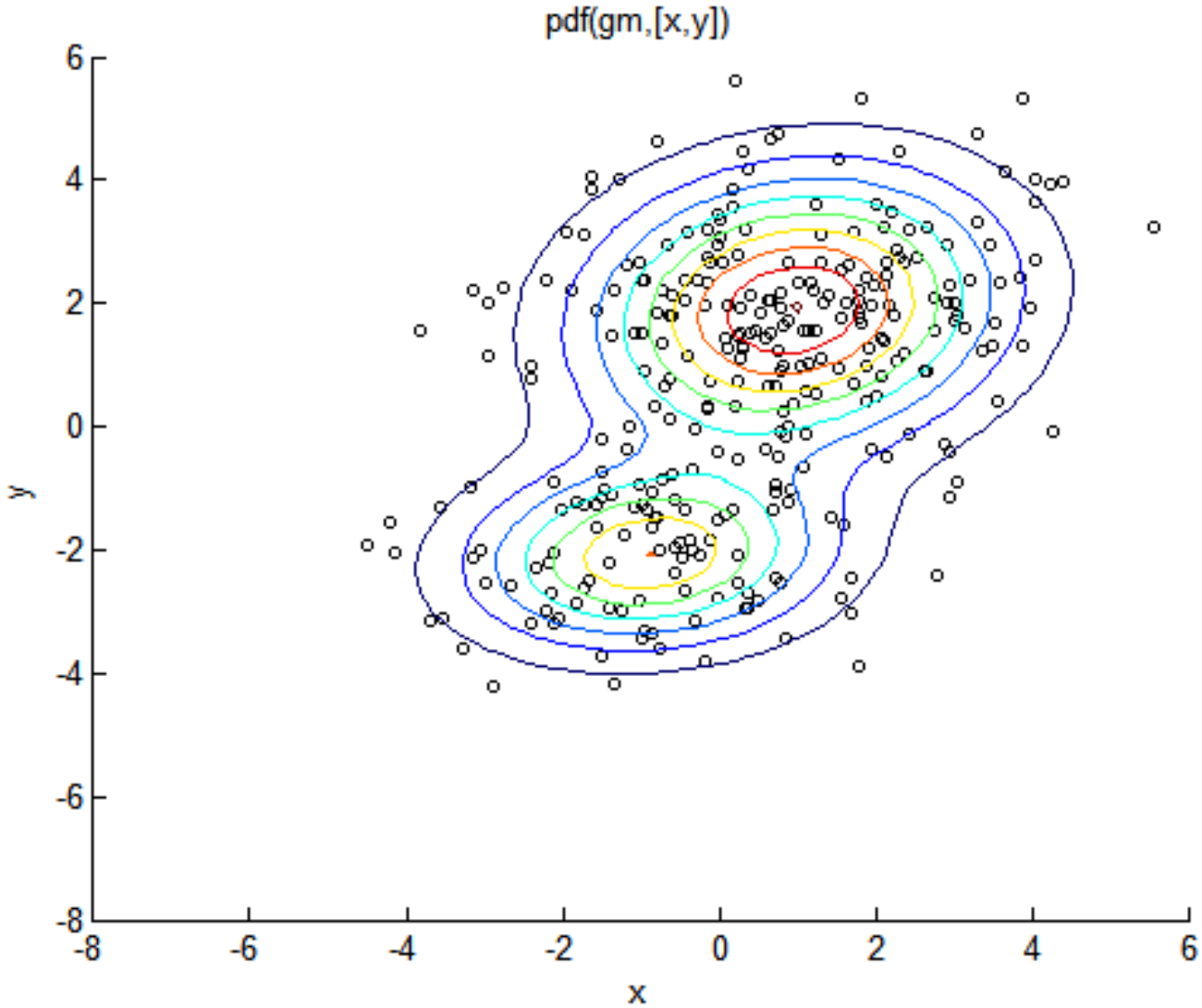


Mixture model (both normal distrubutions contribute 50%).



Mixture model (one normal distrubutions contributes 10%, the other 90%).

Gaussian mixture models



Gaussian mixture models – EM clustering

- **Assumption:** Data were generated by sampling a set of normal distributions.
  (The probability density is a mixture of normal distributions.)
- **Aim:** Find the parameters for the normal distributions and how much each normal distribution contributes to the data.
- **Algorithm:** EM clustering (expectation maximisation). Alternating scheme in which the parameters of the normal distributions and the likelihoods of the data points to be generated by the corresponding normal distributions are estimated.

Density Based Clustering

Density-based clustering

For numerical data, density-based clustering algorithm often yield the best results.

Principle: A connected region with high data density corresponds to one cluster.

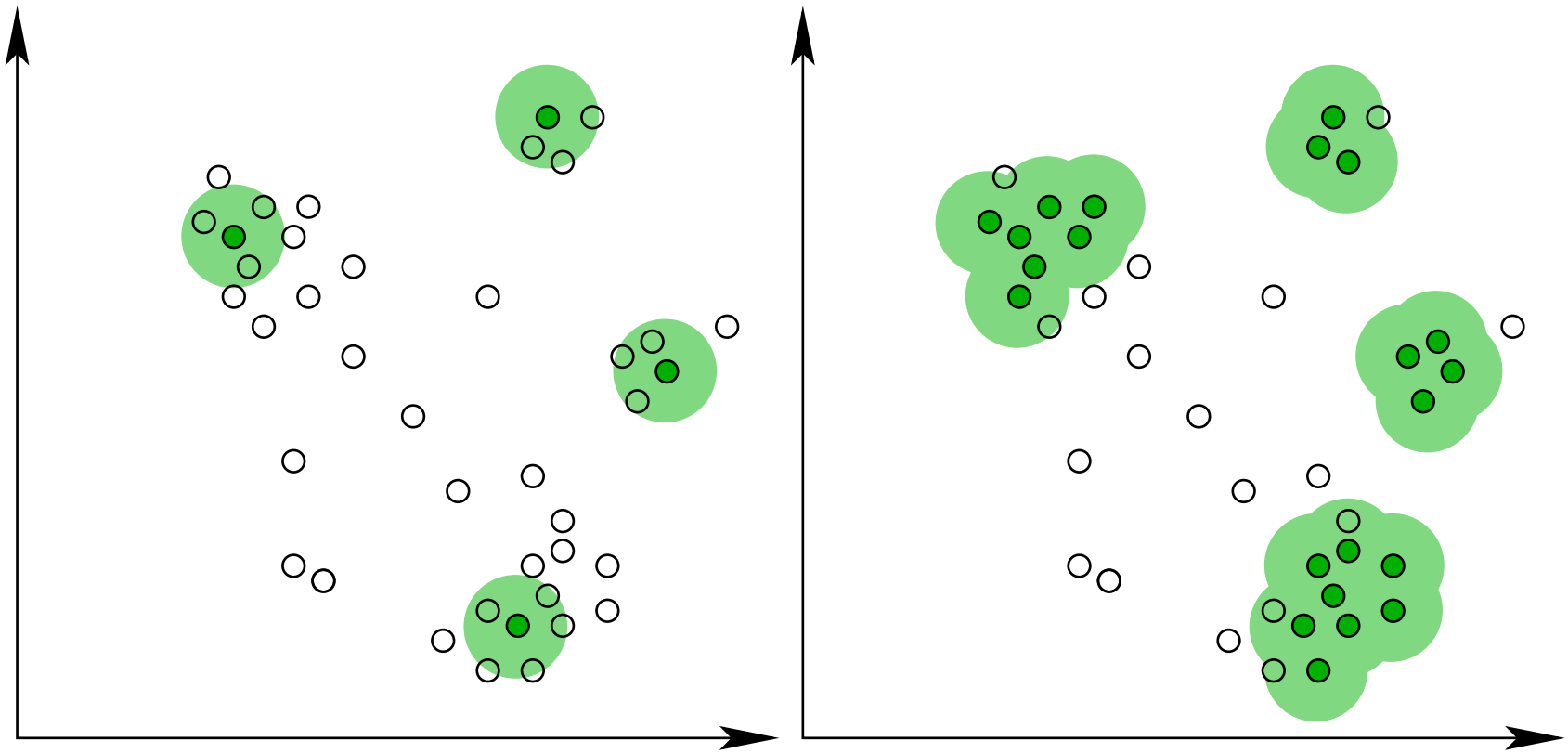
DBScan is one of the most popular density-based clustering algorithms.

Density-based clustering: DBScan

Principle idea of DBScan:

1. Find a data point where the data density is high, i.e. in whose ε-neighbourhood are at least ℓ other points. (ε and ℓ are parameters of the algorithm to be chosen by the user.)
2. All the points in the ε-neighbourhood are considered to belong to one cluster.
3. Expand this ε-neighbourhood (the cluster) as long as the high density criterion is satisfied.
4. Remove the cluster (all data points assigned to the cluster) from the data set and continue with 1. as long as data points with a high data density around them can be found.

Density-based clustering: DBScan



neighbourhood cell with at least 3 hits

neighbourhood cell with at least 3 hits