Statistisches Data Mining (StDM)
Woche 5

Oliver Dürr

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

oliver.duerr@zhaw.ch

Winterthur, 18 Oktober 2016

No laptops, no phones, no problems

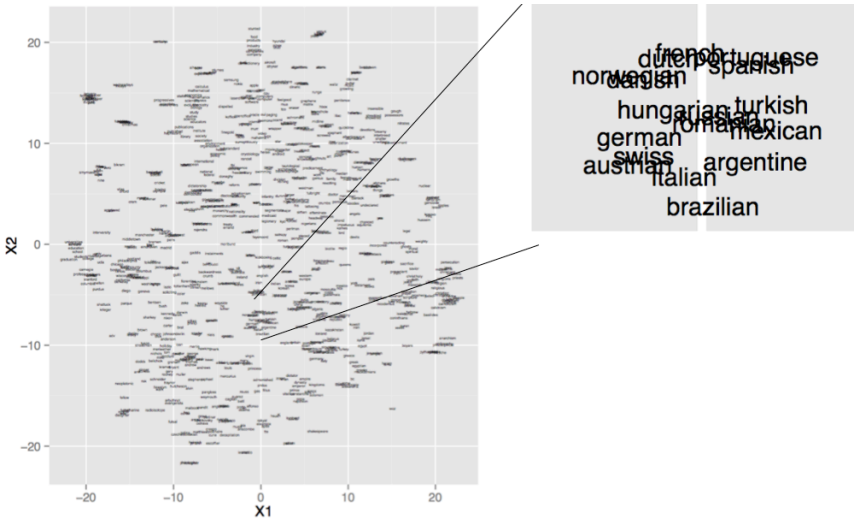
Multitasking senkt Lerneffizienz:

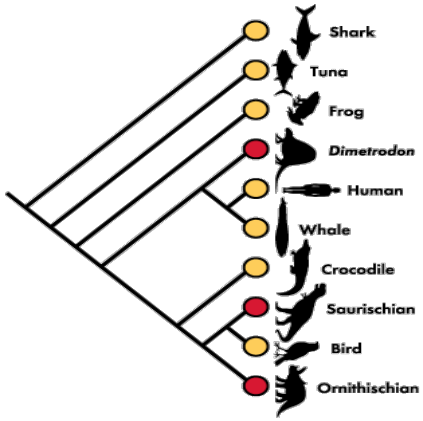
- **Keine Laptops im Theorie-Unterricht Deckel zu oder fast zu (Sleep modus)**

Overview of the semester

Part I (Unsupervised Learning)

- Dimension Reduction
  - PCA
- Similarities, Distance between objects
  - Euclidian, L-Norms, Gower,…
- Visualizing Similarities (in 2D)
  - MDS, t-SNE
- Clustering
  - K-Means
  - Hierarchical Clustering



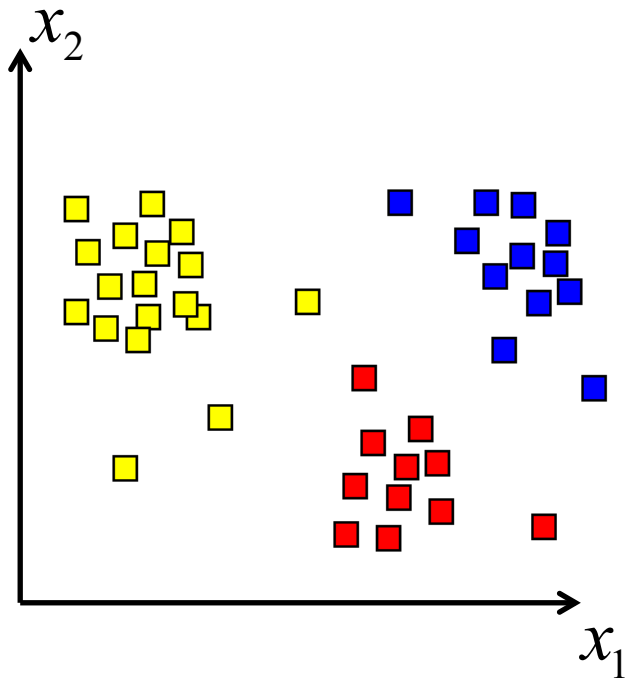


Part II (Supervised Learning)

- …

Clustering

10.3 Clustering Methods in ILSR



Now again in line with ILSR

- Inline again with ILSR
- See section 10.3 Clustering Methods in ILSR

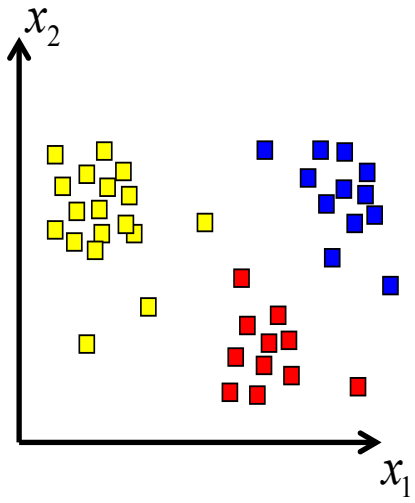
- Aims:
  - PCA (and other dimension reduction methods) look to find a low-dimensional representation of the observations
  - Clustering looks to find homogeneous subgroups among the observations.

- Examples of applications
  - Personalized medicine
    - Segment into subgroups needing different medication
  - Market segmentation
  - …

Descriptive and unsupervised: Cluster Analysis

Cluster analysis or clustering is the task of assigning a set of objects into groups.

Objects in the same cluster should be more similar to each other than to those in other clusters.



To perform clustering one must define a measure of similarity or distance based on the observed values describing different properties of the objects.

$$e.g.\ euclidean\ :\ dist(o_k, o_l) = \sum_{i=1}^{p} (x_{ki} - x_{li})^2$$

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

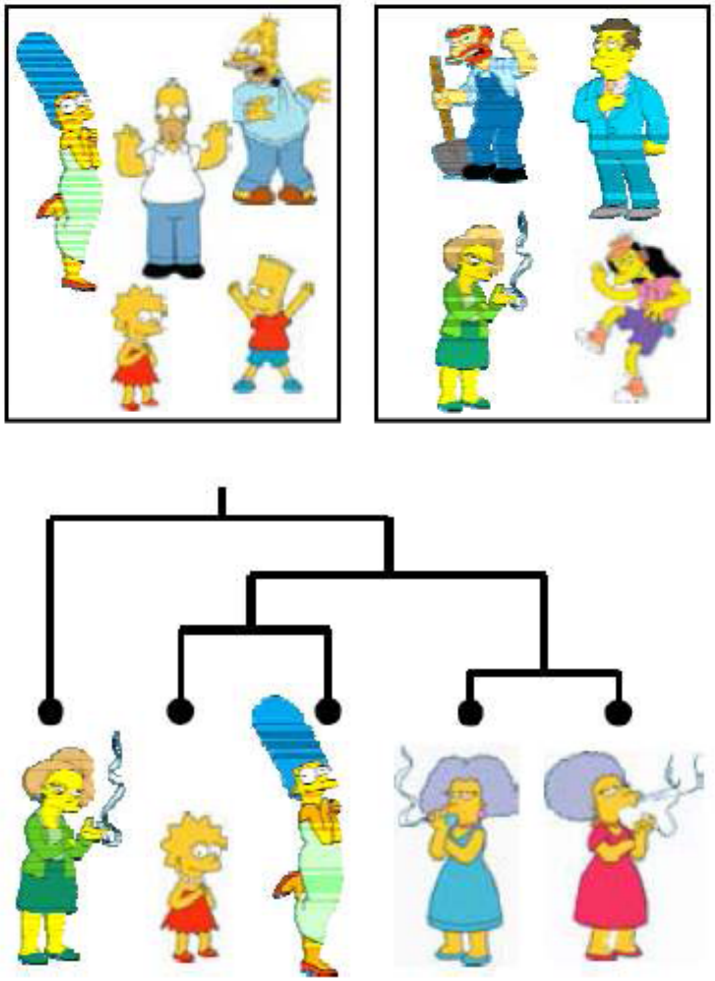
Four Clusters

Types of Clustering

- **Important distinction between hierarchical and partitional sets of clusters**
- **Partitional Clustering**

  A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering**

  A set of nested clusters organized as a hierarchical tree



Partitional Clustering (Recap)

What is optimized in K-means Clustering ?

The goal in k-means is to partition the observations into K clusters such that the total within-cluster variation (WCV), summed over all K clusters C_k, is as small as possible.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \text{WCV}(C_k) \right\}$$

WCV is often based on Euclidian distances

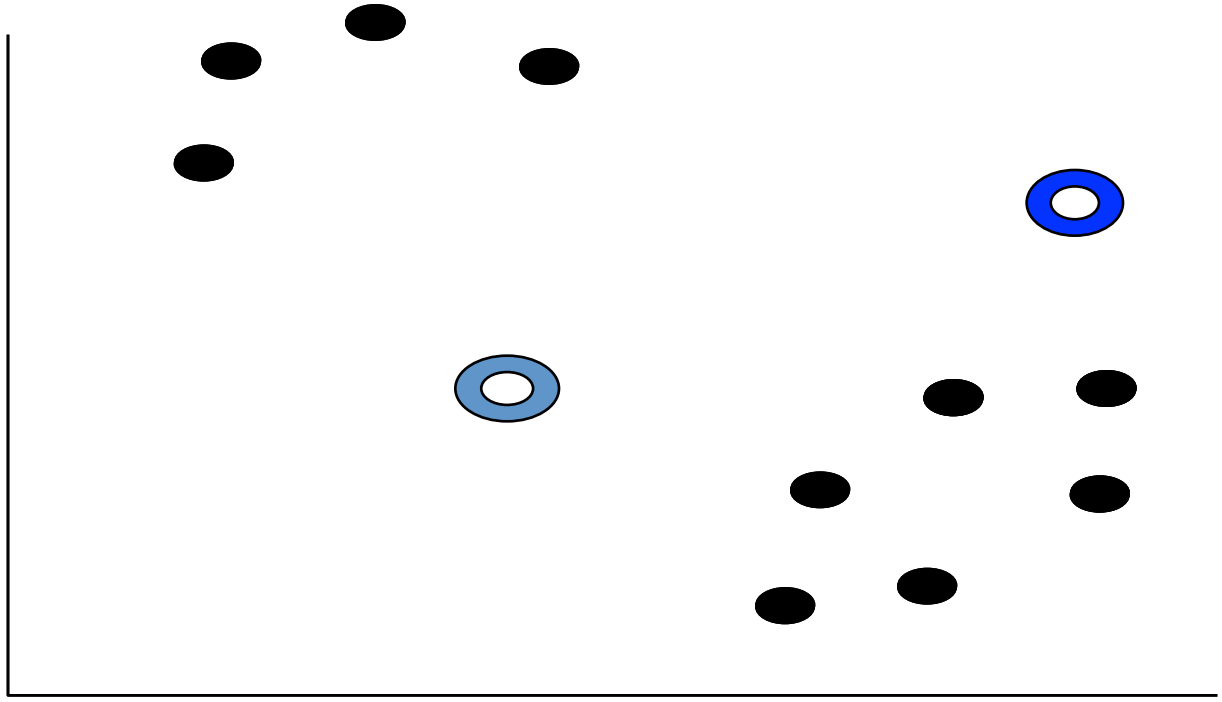
Squared Euclidian distance between data points i and i'

$$\text{WCV}(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

where $|C_k|$ denotes the number of observations in the kth cluster and p is the number of variables (dimensions).

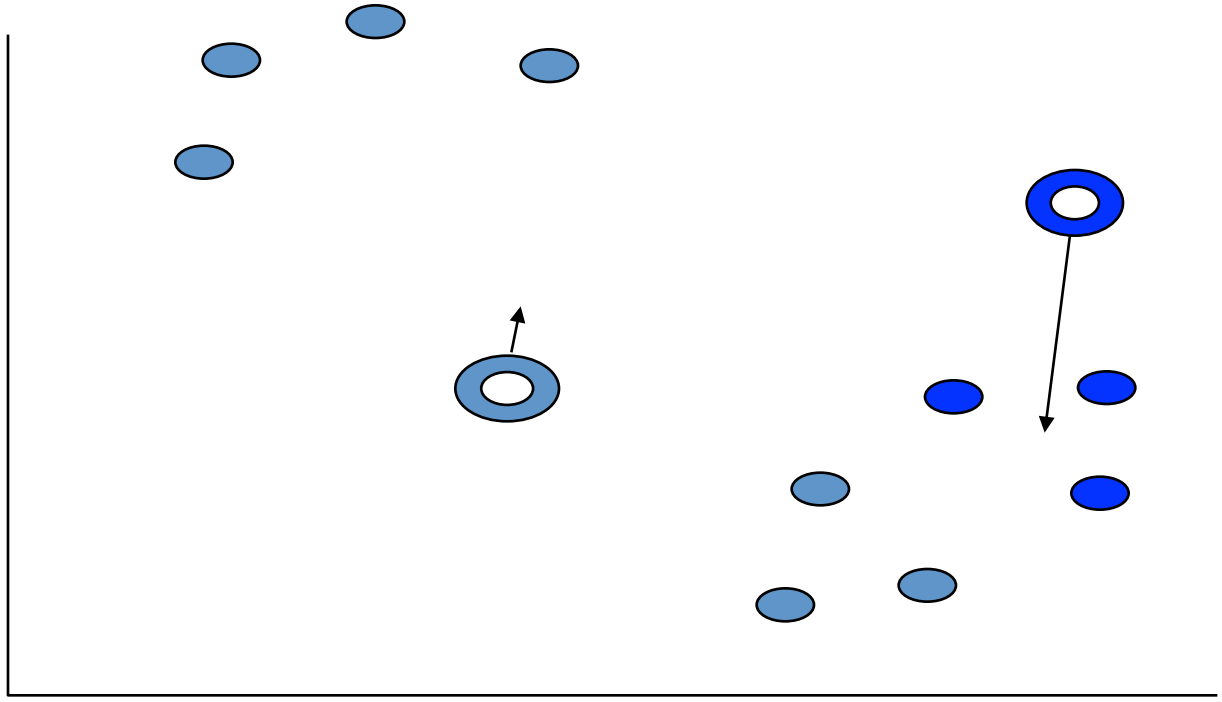
Partitioning Clustering: K-means

+ Given a number of objects and an initial (randomly choosen) set of cluster centers, assign each object to the closest cluster center



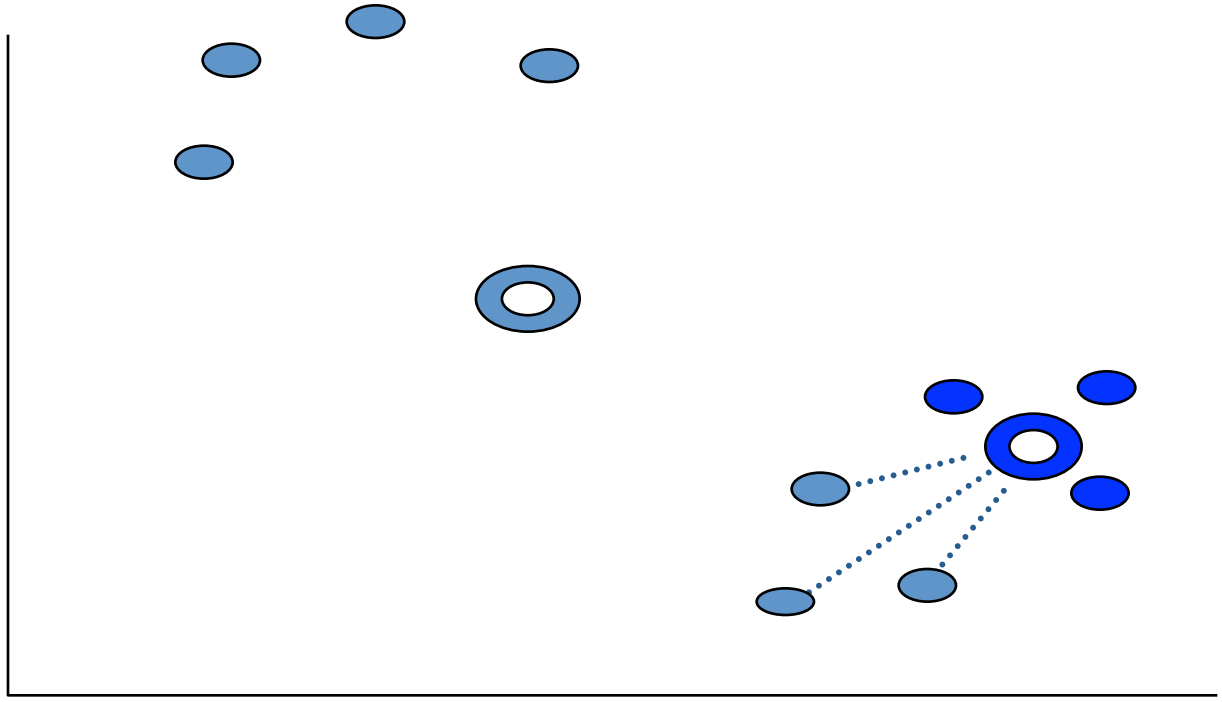
Partitioning Clustering: K-means

+ Update the coordinates of each cluster center to the average coordinate of the objects associated with it



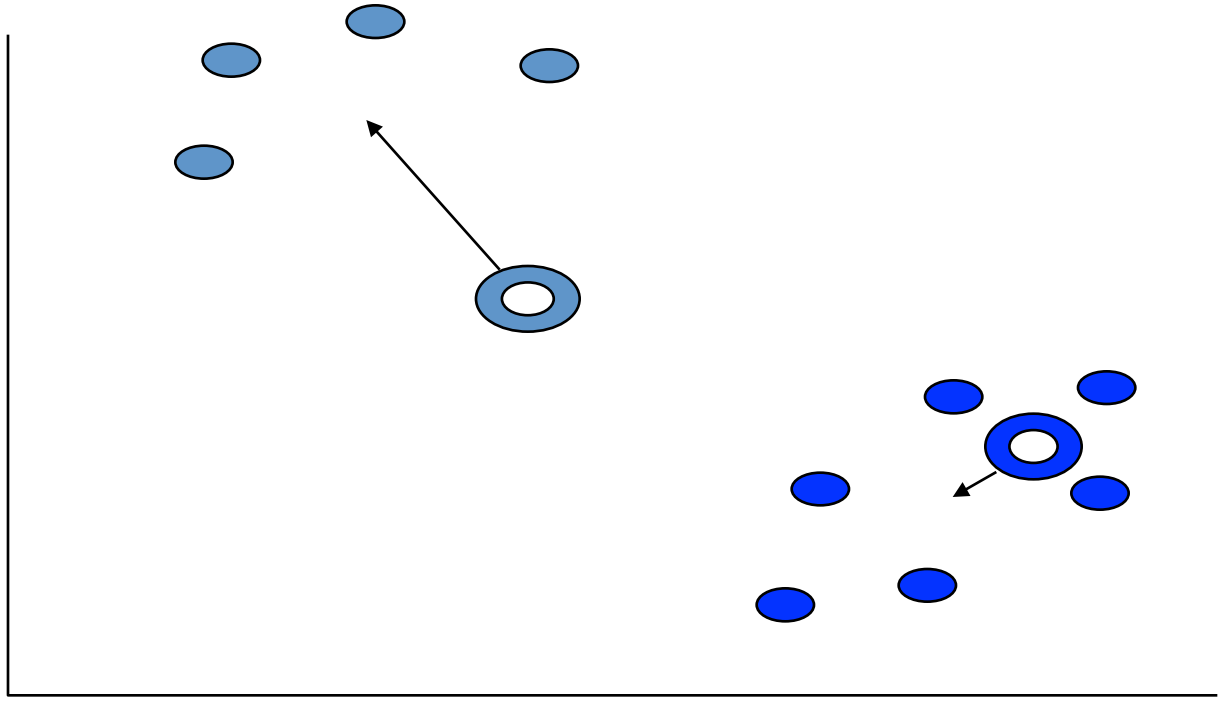
Partitioning Clustering: K-means

+ Re-assign each gene to the closest cluster centre



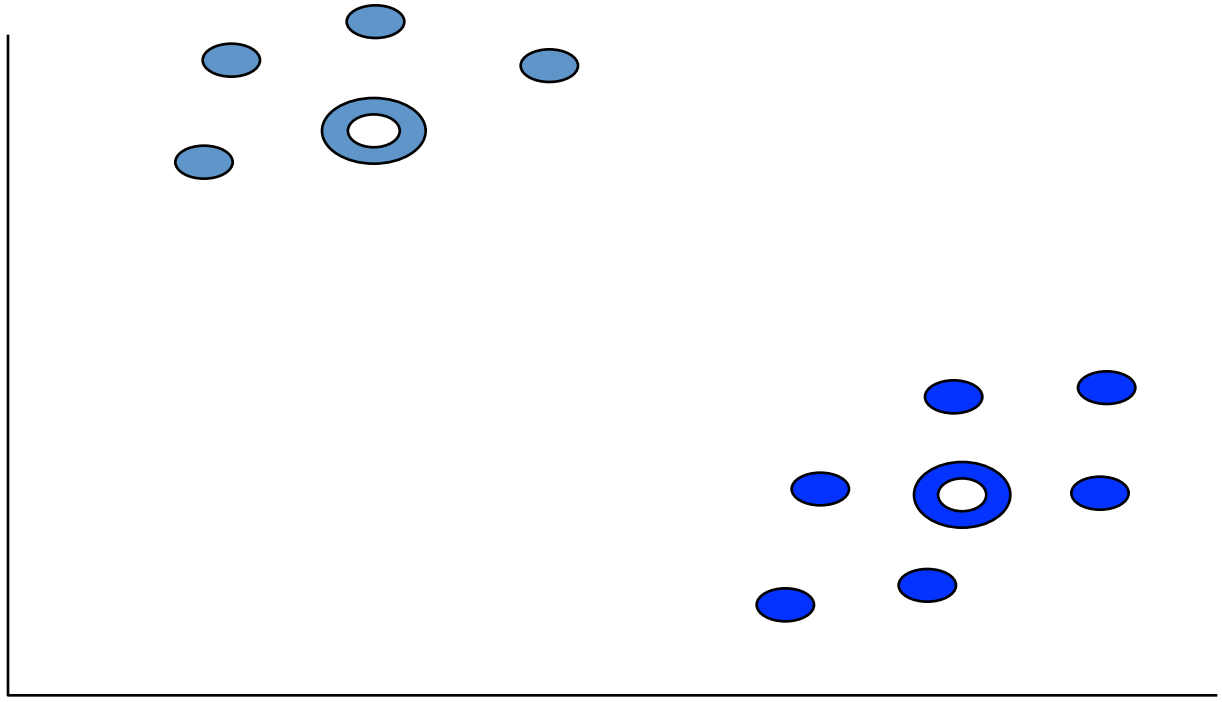
Partitioning Clustering: K-means

+ Recalculate the co-ordinates of each cluster centre according to the average co-ordinate of the genes associated with it



Partitioning Clustering: K-means

+ Repeat these steps until no reassignment is possible (or maximal number of iterations have been reached).



Hierarchical Clustering

Types of Clustering

- **Important distinction between hierarchical and partitional sets of clusters**

- **Partitional Clustering**

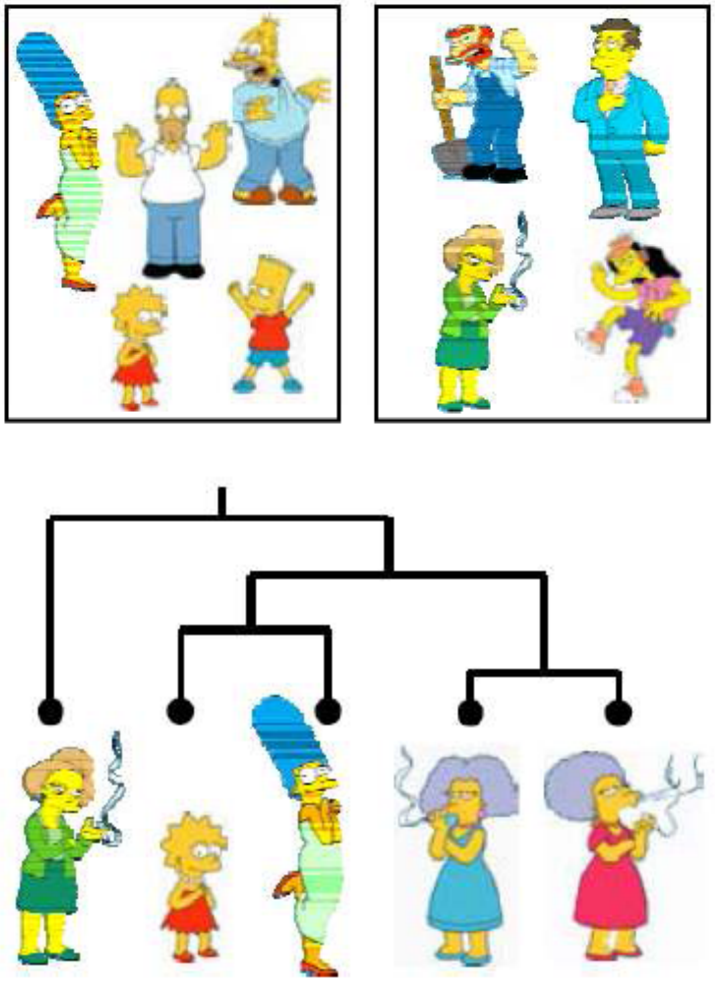
  A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

  – K-Means clustering needs K!

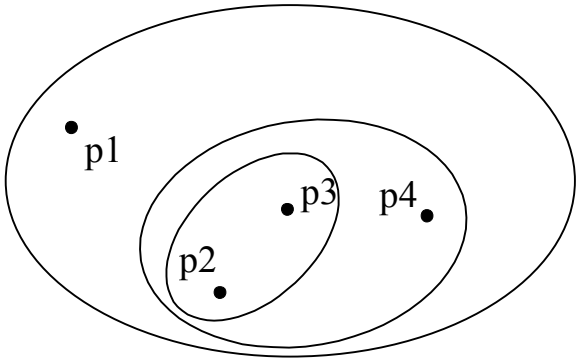
- **Hierarchical clustering**

  A set of nested clusters organized as a hierarchical tree

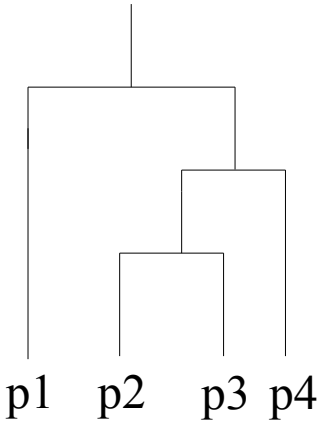
  – Constructs a complete tree of dependencies. No need specify K in advance



Hierarchical Clustering



Hierarchical Clustering



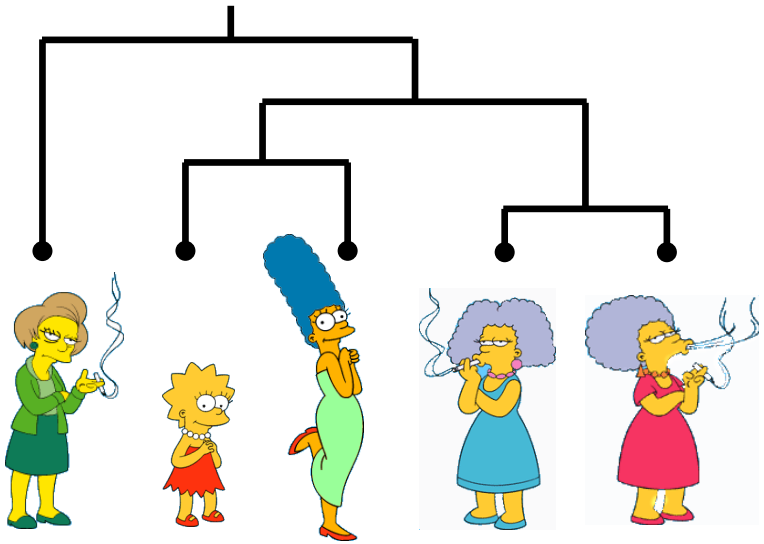
Dendrogram

- From bottom to top “agglomerativ” ← The usual way, done here
- From top to bottom “divisive”

How to do hierarchical Clustering?

Without proof: The number of dendrograms with n leafs:
$= (2n-3)!/[(2^{(n-2)})\,(n-2)!]$

Number of Leafs	Number of Possible Dendrograms
2	1
3	3
4	15
5	105
...	…
10	34,459,425



Since we cannot test all possible dendrograms we will have to heuristic search of all possible dendrograms. We could do this..

Bottom-Up (agglomerative):
Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

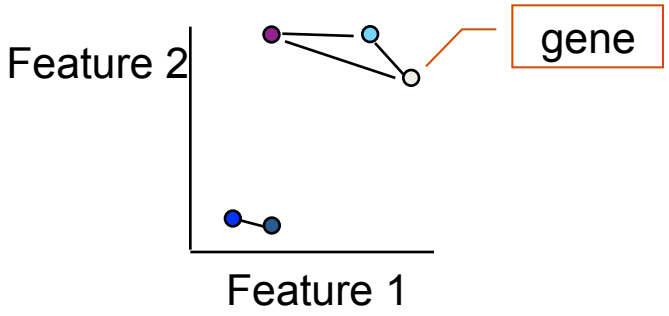
Top-Down (divisive):
Starting with all the data in a single cluster, consider every possible way to divide the cluster into two. Choose the best division and recursively operate on both sides.

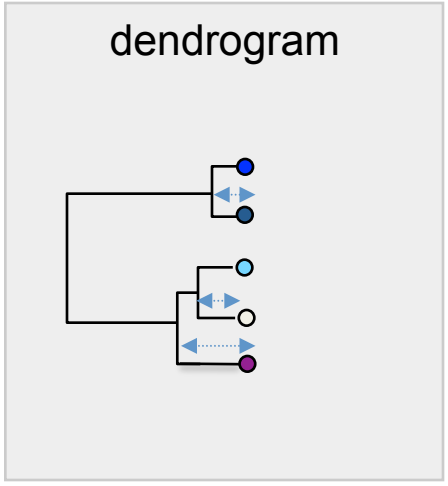
Dissimilarity between samples or observations

Any dissimilarity we have seen before can be used

- euclidean
- manhattan
- simple matching coefficent
- Jaccard dissimilarity
- Gower's dissimilarity
- etc.

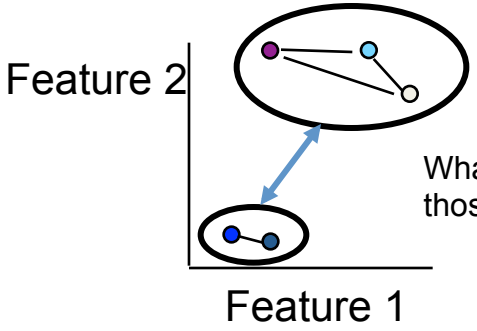
Aglomatrative Hierarchical Clustering





Problem:

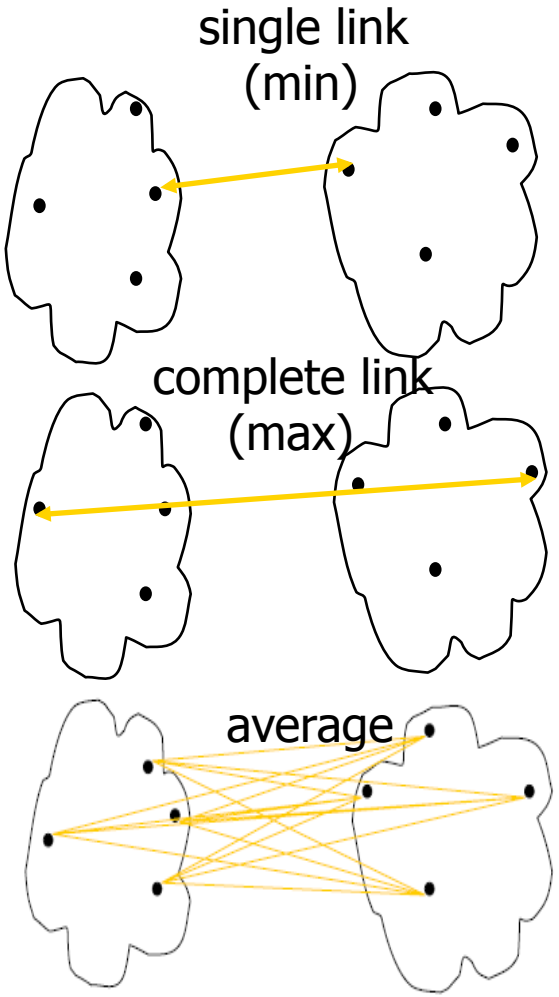
Need a generalization of the distance between the objects to compound of objects.

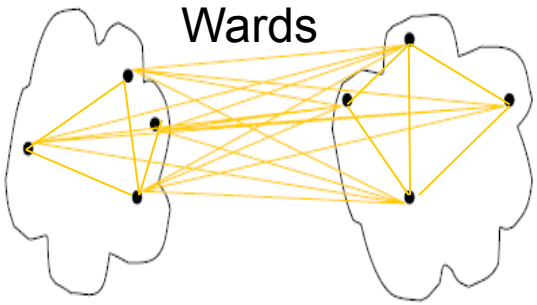


What is the distance between those objects. -> **Linkage**

Dissimilarity between clusters: Linkages

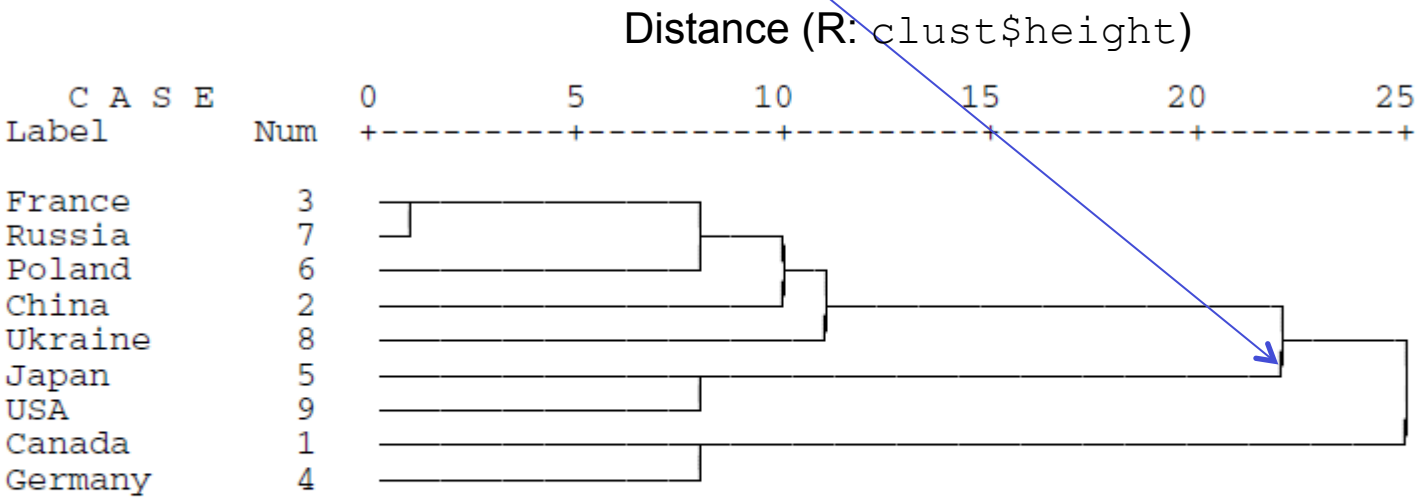
- Single link: smallest distance between point-pairs linking both clusters
- Complete link: largest distance
- Average: avg distance between
- Wards: In this method, we try to minimize the variance of the merged clusters





How to read a dendrogram

The position of the join node on the distance-scale indicates the distance between clusters (this distance depends on the linkage method). For example, if you see two clusters merged at a height 22, it means that the distance between those clusters was 22 .



When you read a dendrogram, you want to determine at what stage the distance between clusters that are combined is large.
You look for large distances between sequential join nodes (here vertical lines).

Simple example

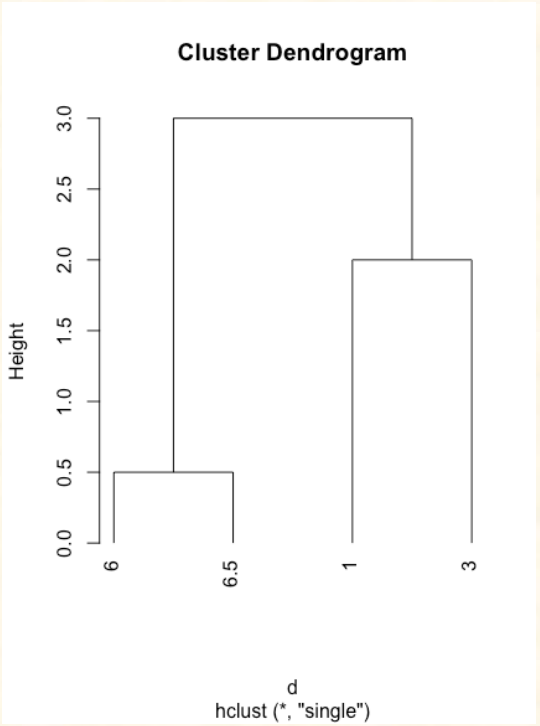
Zeichnen sie das Dendrogram für die eindimensionale Datenmatrix

feature
1
3
6
6.5

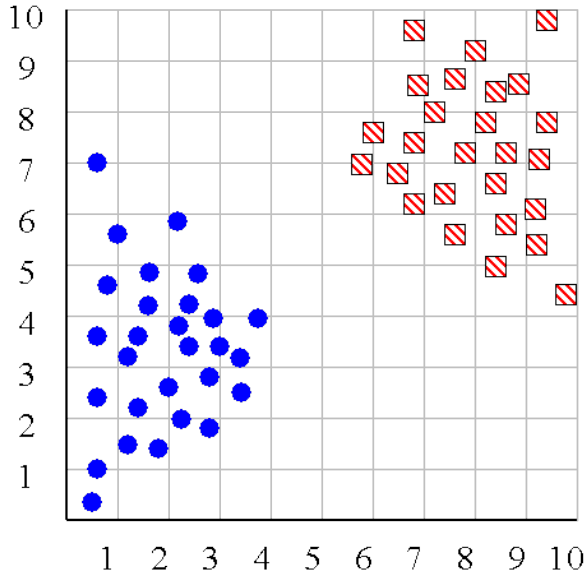
Verwenden Sie dazu die Euklidische Distanzen und die single-linkage

Simple example

```
x = c(1,3,6,6.5)
names(x) = c('1','3','6', '6.5')
d = dist(x)
cluster = hclust(d, method = 'single')
plot(cluster, hang=-10, axes = FALSE)
axis(2)
```

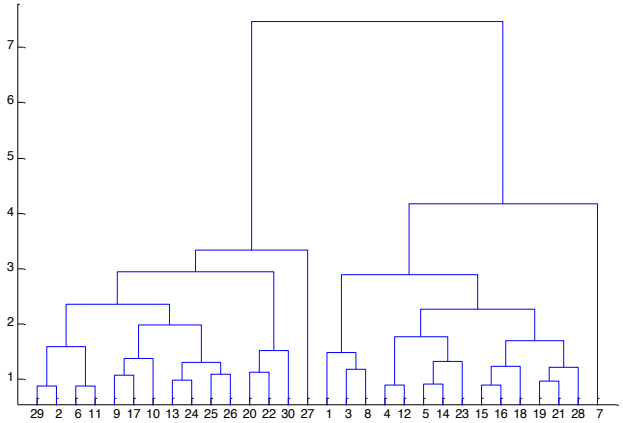


Compare linkage methods

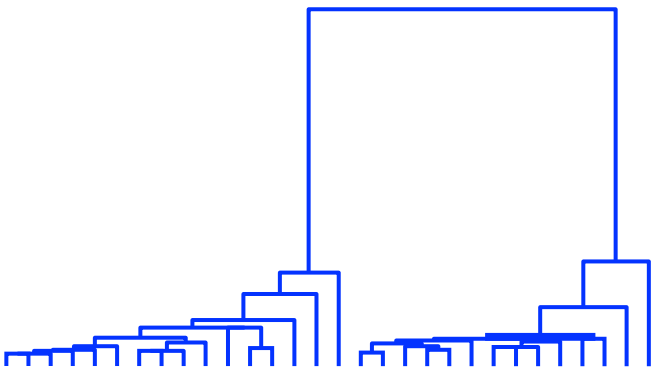


- Single-Linkage produce long and skinny clusters.
- Wards produce of ten very separated clusters
- Average linkage yield more round clusters

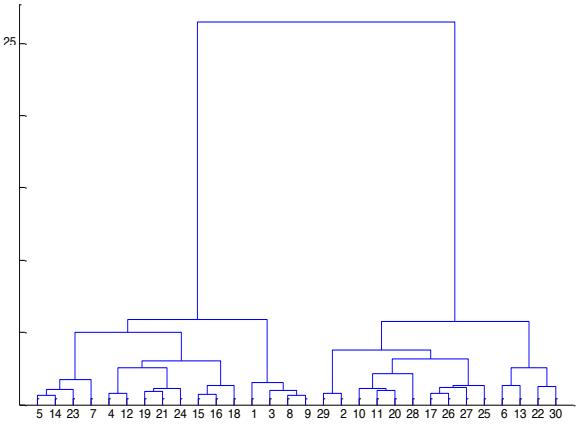
•Generally clustering is an exploratory tool.
Use the linkage which produces the “best” results.



Average linkage



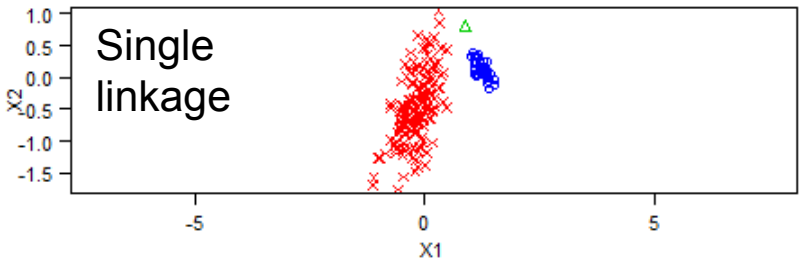
Single linkage

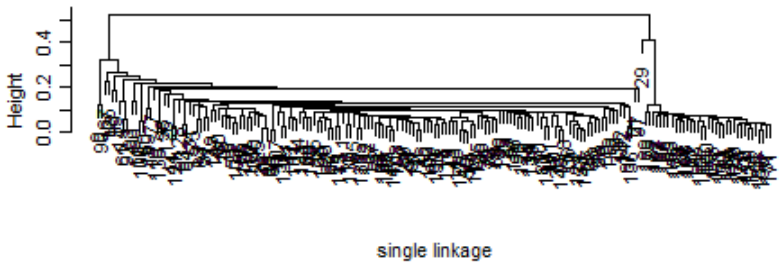


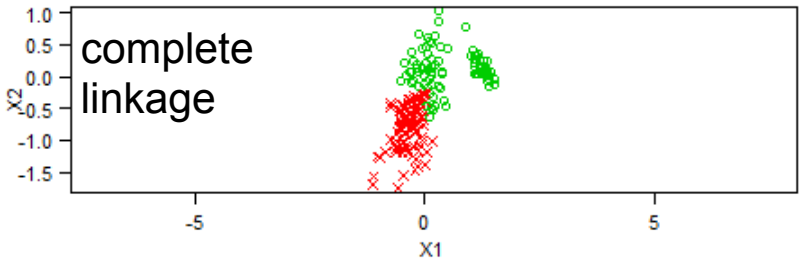
Wards linkage

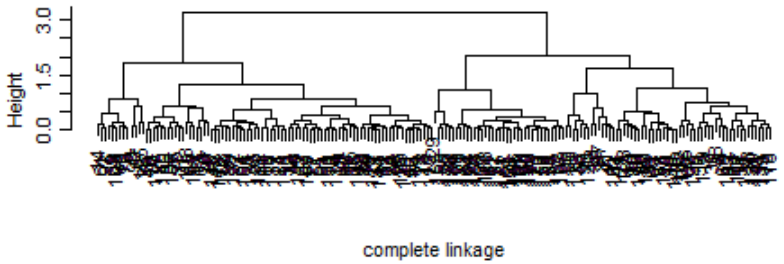
Cluster result depend on data structure, distances and linkage methods

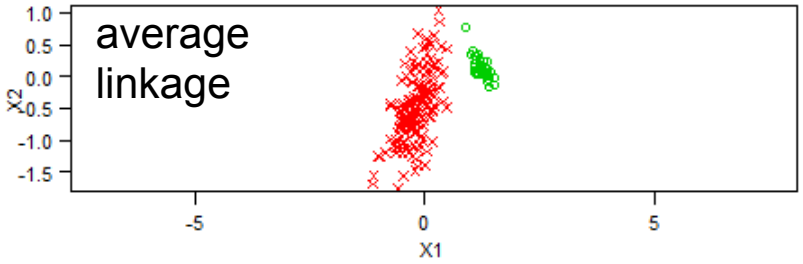
Data: we simulated 2 2D-Gaussian Clusters with very different sizes

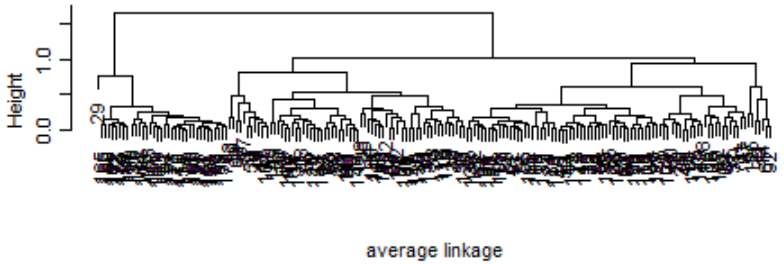


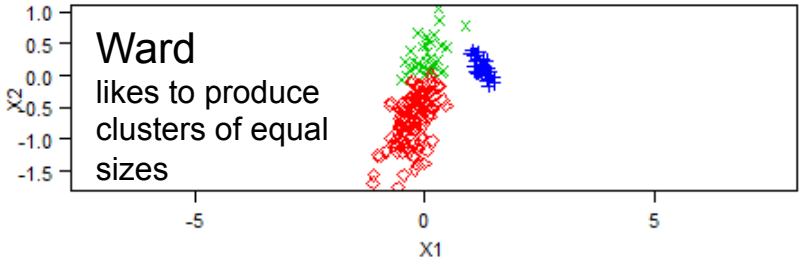


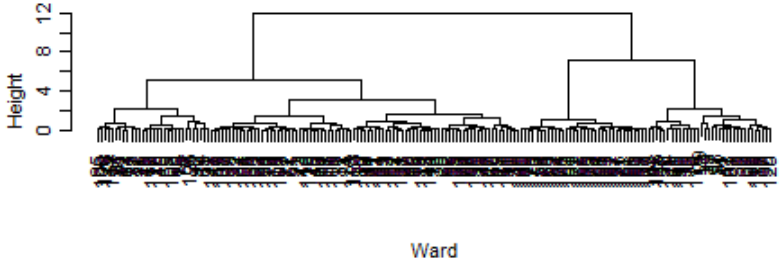












Heatmaps

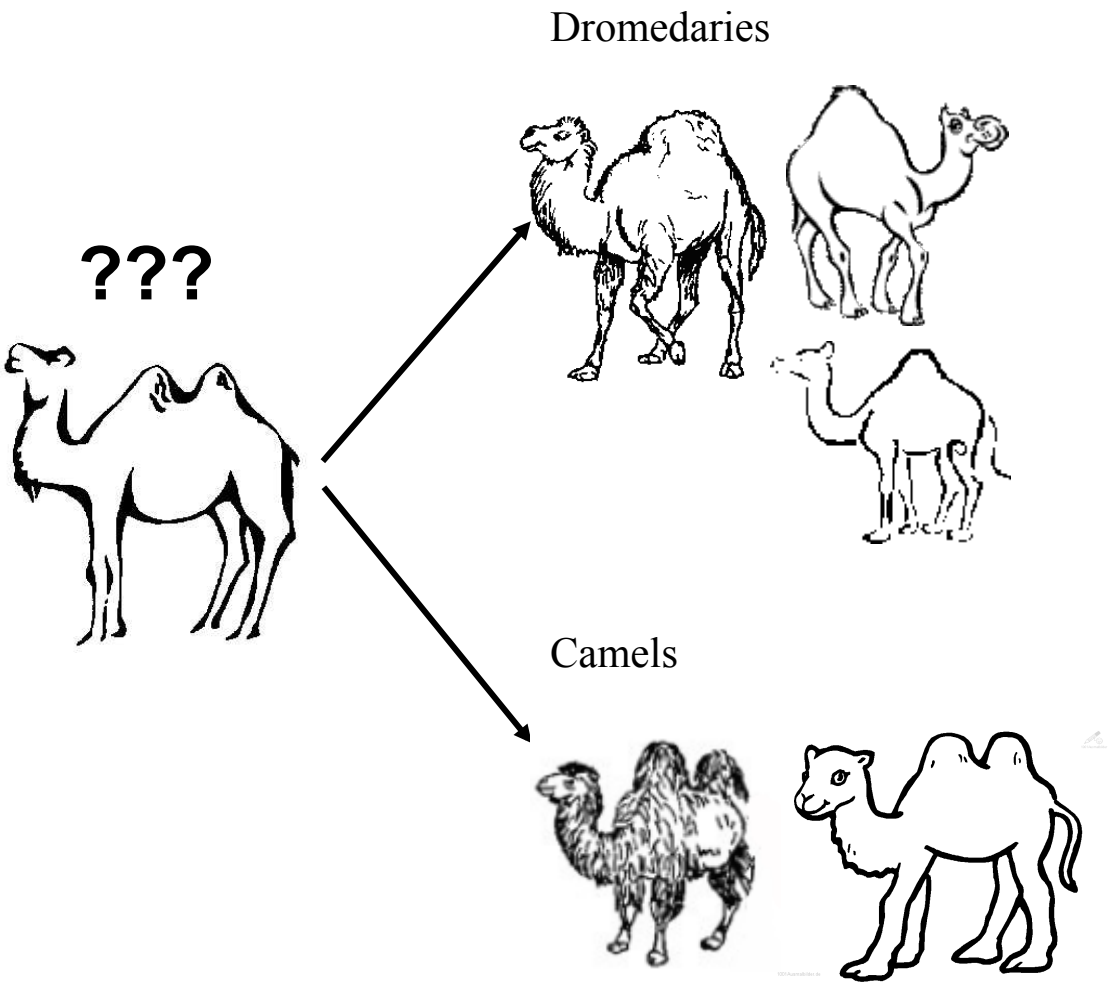
What is a classification task?

- Classification is a prediction method

- **Idea:**

  Train a classifier based on training data (examples with known class labels) and use the classifier to classify new test observations with unknown class label.

- Which feature should we use to describe an observation (animal)?



Feature extraction

ID of animal	**Class label**	Number of legs	Number of bumps	Length of legs [cm]
1	Dromedar	4	1	98
2	Kamel	3	2	87
…	…	…	…	…
150	Kamel	4	2	103

Defining appropriate features is essential for the success of the classification task!

It is not always as simple as it is in this example:

Features can be combined to new features or selected.

Data Matrix

Class label
Dromedar
Kamel
…
Kamel

Number of legs	Number of bumps	Length of legs [cm]
4	1	98
3	2	87
…	…	…
4	2	103

Y

Labels are categorical
Labels continuous ➔ regression

X

Data Matrix with several features (can also contain categorical values).
One row called ‘feature vector’

In classification aka supervised learning we try to predict the class labels using the features.

Principal Idea Classification

Klassifikatoren
•Neuronale Netze
•Entscheidungsbäum
•...

Training Data

id	Type	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	setosa	5.1	3.5	1.4	0.2
2	setosa	4.9	3	1.4	0.2
3	virinica	3.3	3.2	1.6	0.5
4	setosa	5.1	3.5	1.4	0.2
...	...	...	...	...	...
150	virinica	4.9	3	1.4	0.2

Learn a classifier

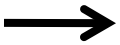
Classifier

Predict

Unknown data / Test data

d	Type	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	?	3.1	3.5	1.4	0.2
2	?	4.9	3	1.4	0.2
3	?	3.3	3.2	1.6	0.5
4	?	5.1	3.5	31.4	0.2

Classifier



Type

Note:
To evaluate the performance a part of the labelled data not used to train the classifier but left aside to check the performance of the classifier to new data.

Examples of Classification Task

- Is a given text e.g. tweet about a product positive, negative or neutral. *Sentiment Analysis*

  "The movie XXX actually neither that funny, nor super witty" → Negative

- Churn in Marketing: Predict which customer wants to quit and offer them a discount

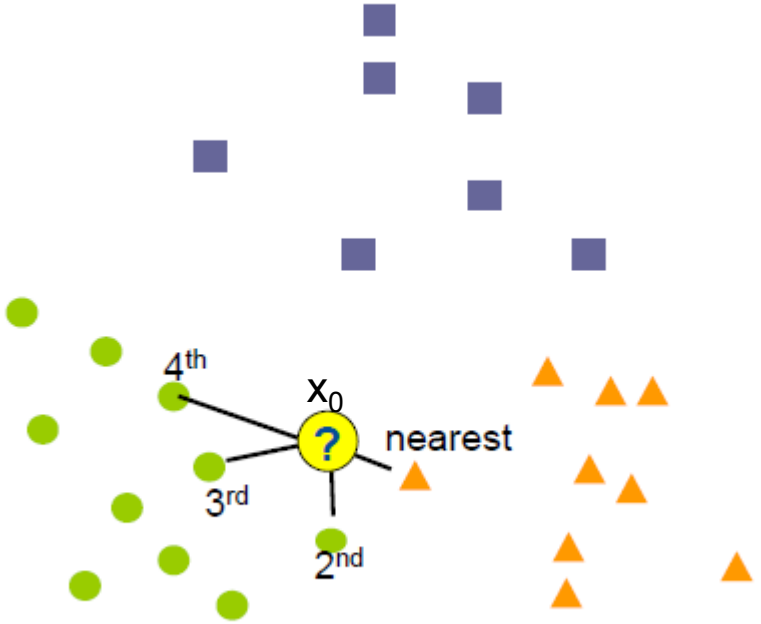
- Face detection. Image (array of pixels) ➔ John

…

K-Nearest-Neighbors in a nutshell

Idea of knn classification:

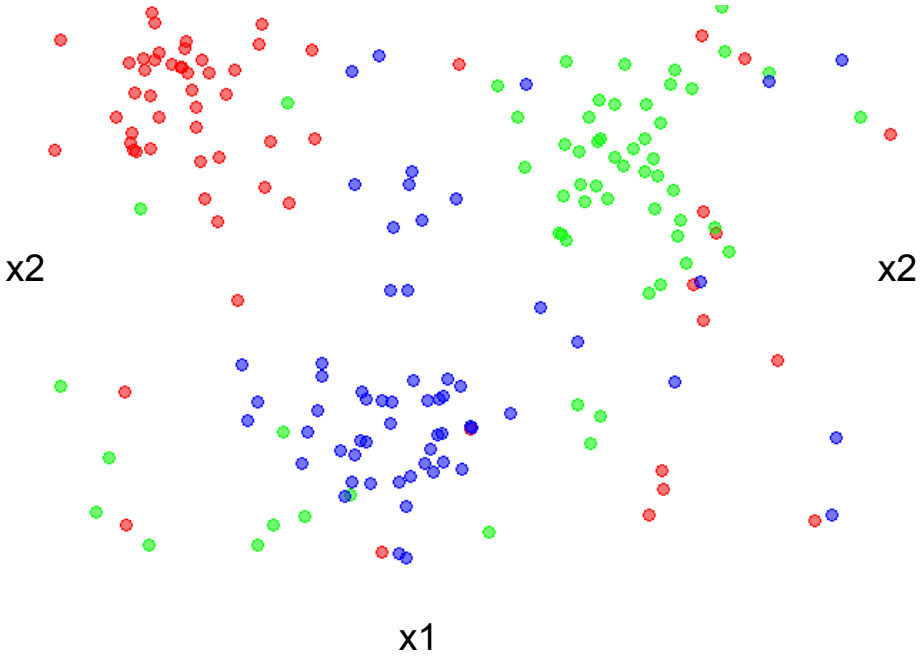
- Start with an observation x_0 with unknown class label
- Find the k training observations, that have the smallest distance to x_0
- Use the majority class among the k neighbors as class label for x_0



R functions to know
- From package “class”: “knn”

knn-classificator with 3 class-labels after training with k=1

data with true class label



Trained classificator (knn with k=1)