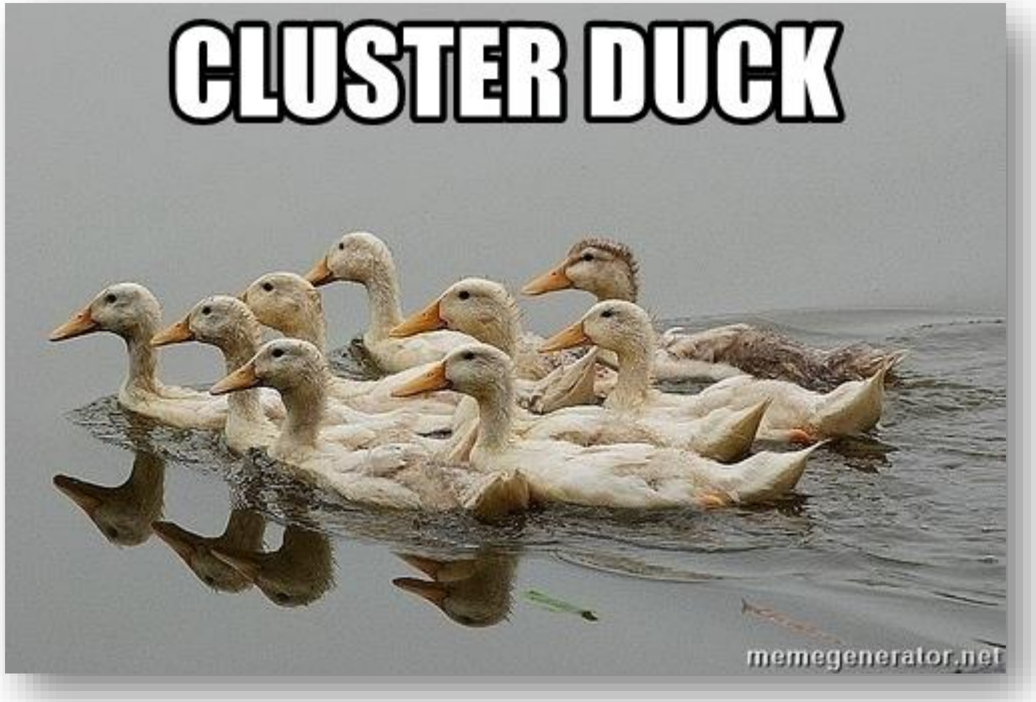


Clustering

# MACHINE LEARNING

# Topics

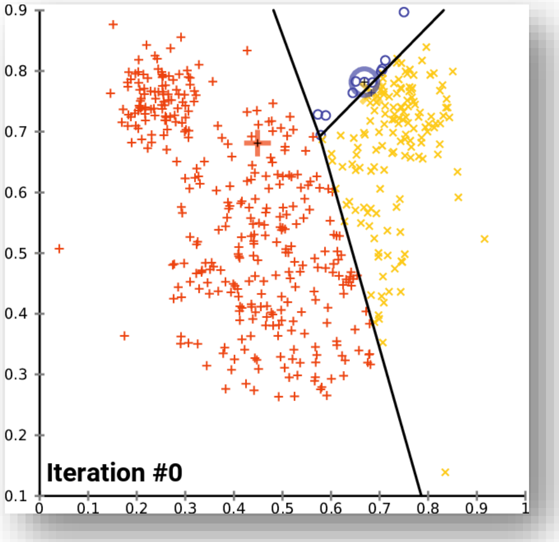
- Introduction: GUI and basic calculations
- Coding 1: Scripts, style, and variable classes
- Coding 2: Control statements and loops
- Visualization 1: Basics, subplots, get and set
- Coding 3: Functions
- Visualization 2: Descriptive plots
- Coding 4: Basic input and output
- Visualization 3: Distribution and 3D plots
- Coding 5: Input and output specials – last lecture before holidays
- Machine Learning 1: Introduction and dimension reduction
- Machine Learning 2: Clustering
- Machine Learning 3: Classification
- Coding 6: Efficiency and debugging basics
- Coding 7: Advanced functions and debugging

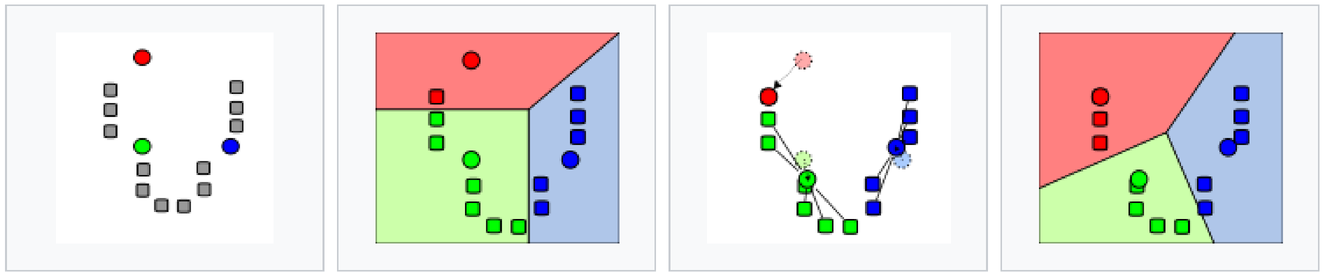
# Cluster Analysis

- Group a **set of data points** (each with a number of features) in a way that objects in the **same group** (cluster) are **more similar to each other than those in other groups**
- Common tool for exploratory data analysis and statistics
- Distance in n-D is generally the used measure
  - Distance in n-D can be defined in various ways, not just euclidean
  - Another measure can be density
- Invented in 1932 (Driver & Kroeber)
  - Famously used for personality analysis (Cattell, 1943)

# K-Means

- A number of clusters (k) are represented by a centroid (artificial data point)
- The centroids move around and „collect“ the closest data points
- Distance within the cluster (spread) is minimized while distance to other clusters is maximized
- Random initialization, iterative approach
  - Not converging to the same solution all the time!

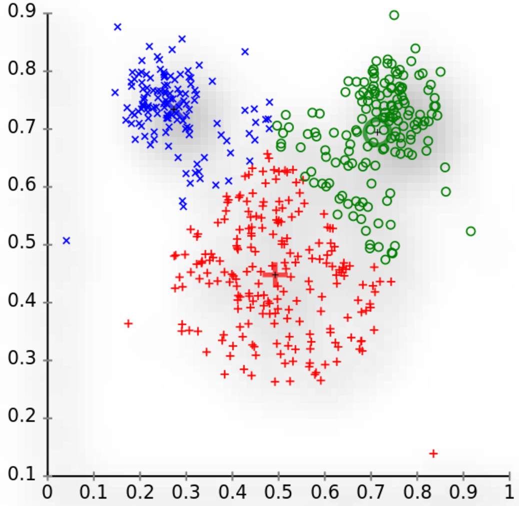




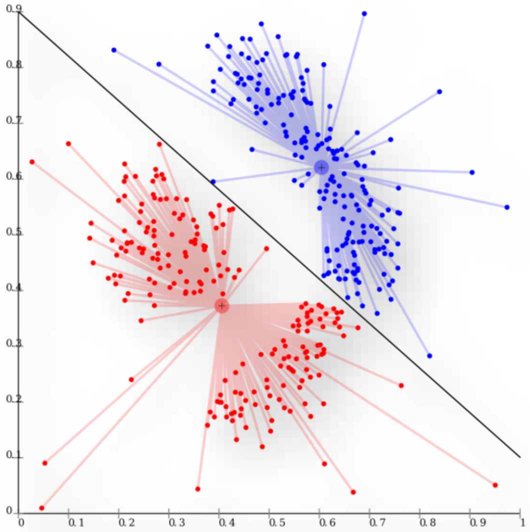
https://en.wikipedia.org/wiki/K-means_clustering

# K-Means Limitations

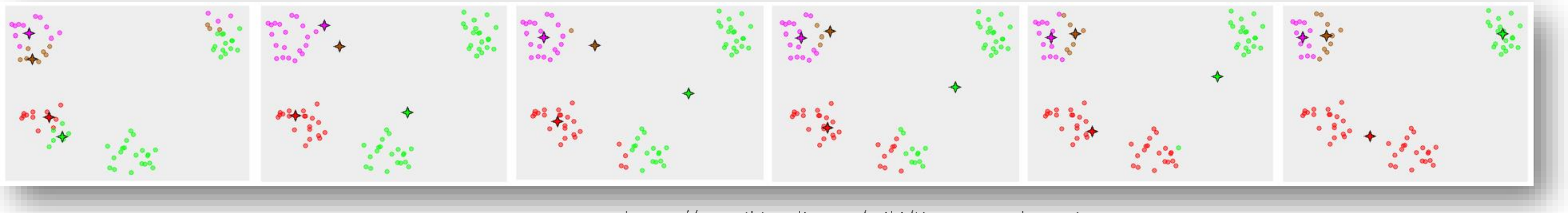
- K needs to be specified!
- Works well only with spherical distributions
- Assumes same-sized clusters
- Can run into local minima



https://en.wikipedia.org/wiki/K-means_clustering



https://de.wikipedia.org/wiki/Clusteranalyse



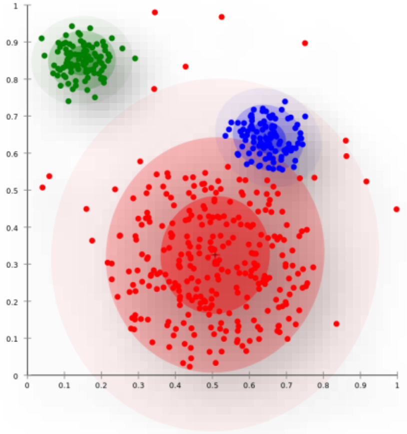
https://en.wikipedia.org/wiki/K-means_clustering

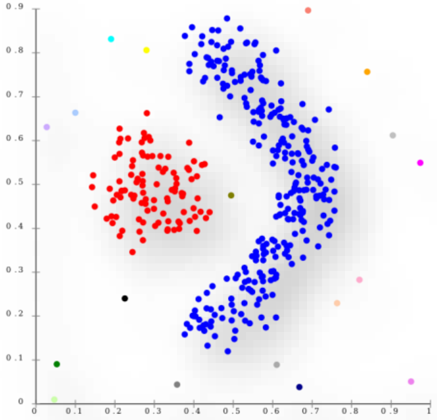
# Other Algorithms

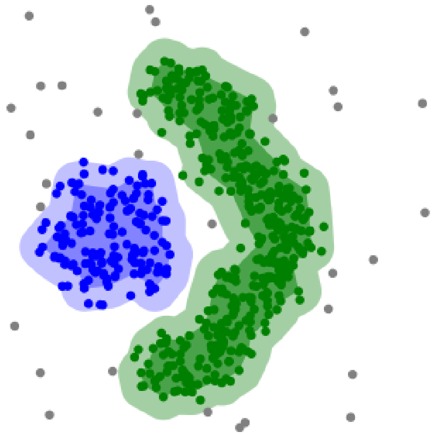
- Distribution-based
  - E.g. gaussian mixture models
- Connectivity-based
  - Clusters based on linkage of data points
  - Nice for swiss roll

-https://en.wikipedia.org/wiki/Swiss_roll

- Density-based
  - Arbitrarily shaped clusters of dense regions
  - Sparse regions are „noise“/border-points
  - Have problems with overlapping boundaries







https://de.wikipedia.org/wiki/Clusteranalyse