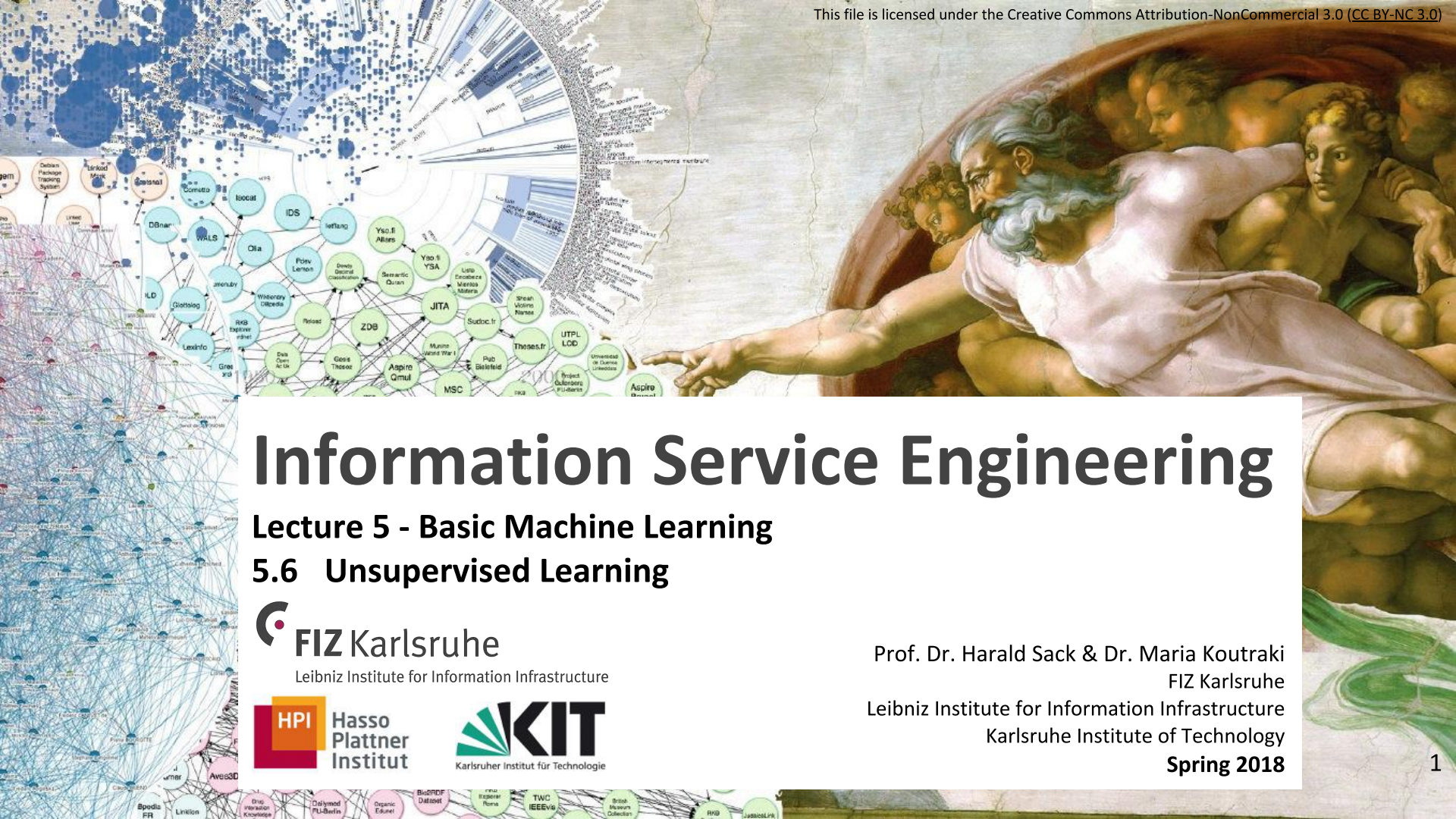




Information Service Engineering

Lecture 5 - Basic Machine Learning

5.6 Unsupervised Learning

 **FIZ Karlsruhe**

Leibniz Institute for Information Infrastructure

 **Hasso Plattner Institut**

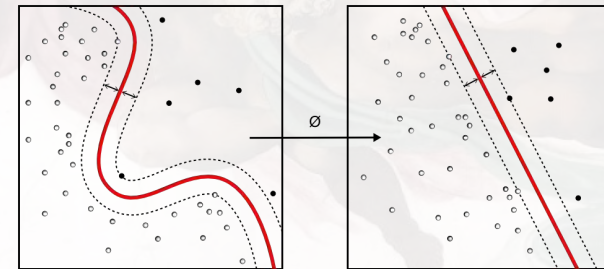
 **KIT**
Karlsruher Institut für Technologie

Prof. Dr. Harald Sack & Dr. Maria Koutraki
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure
Karlsruhe Institute of Technology
Spring 2018

Information Service Engineering

Lecture 5: Basic Machine Learning

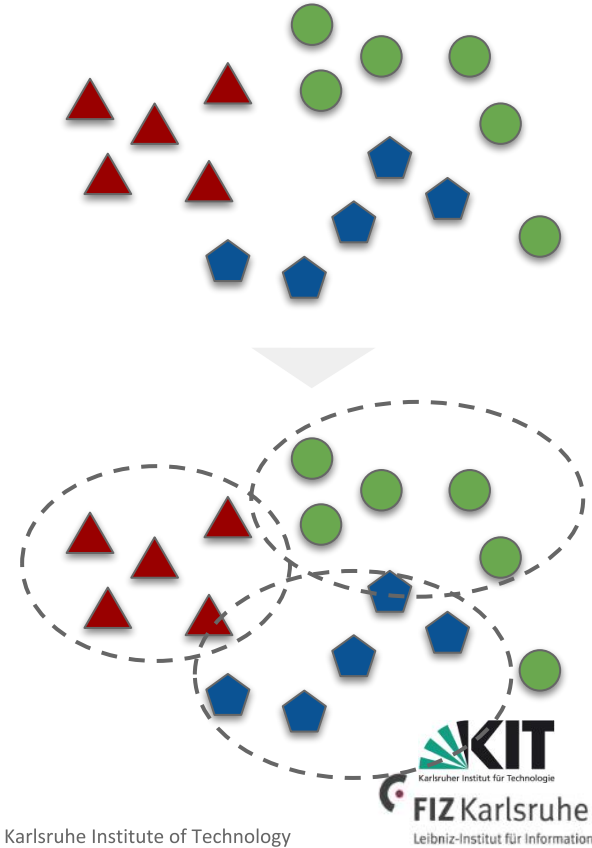
- 5.1 A Brief History of AI
- 5.2 Introduction to Machine Learning
- 5.3 Evaluation and Generalization Problems
- 5.4 Basic ML Algorithms
- 5.5 Basic ML Algorithms II
- 5.6 Unsupervised Learning**
- 5.7 Deep Learning



Unsupervised Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

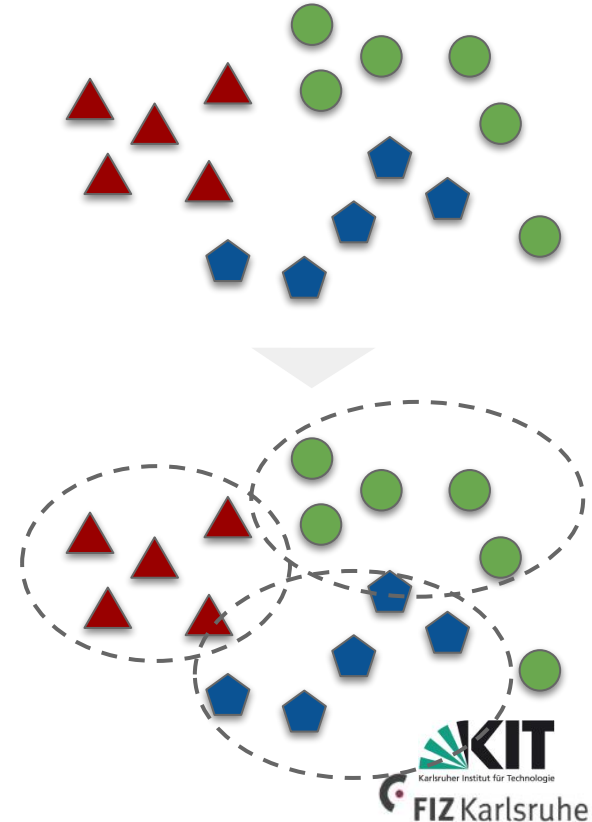
- In **supervised learning** aka **predictive analytics**, data consisted of observations (x_i, y_i) , $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$
- such data is called **labelled**, and the y_i are thought of as the labels for the data



Unsupervised Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

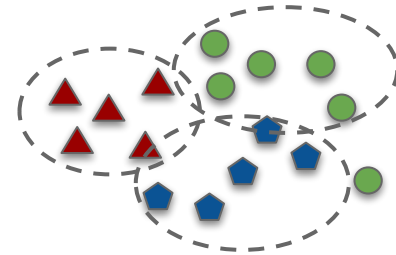
- In **supervised learning** aka **predictive analytics**, data consisted of observations (x_i, y_i) , $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$
- such data is called **labelled**, and the y_i are thought of as the labels for the data
- In **unsupervised learning**, we just look at data $x_i \in \mathbb{R}^p$, $i = 1, \dots, n$
- This is called **unlabelled data**
- Even if we have labels y_i , we may still wish to temporarily ignore the y_i and conduct unsupervised learning on the inputs x_i



Examples for Unsupervised Learning Tasks

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

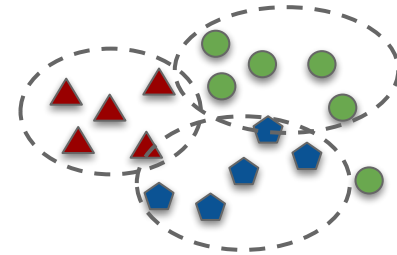
- Identify similar groups of **online shoppers** based on their browsing and purchasing history
- Identify similar groups of **music listeners** or **movie viewers** based on their ratings or recent listening/viewing patterns
- Cluster input variables based on their correlations to remove redundant predictors from consideration
- Identify similar groups of patients based on their medical records
- Determine how to **place sensors, broadcasting towers, law enforcement, or emergency-care centers** to guarantee that desired coverage criteria are met



Unsupervised Learning

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

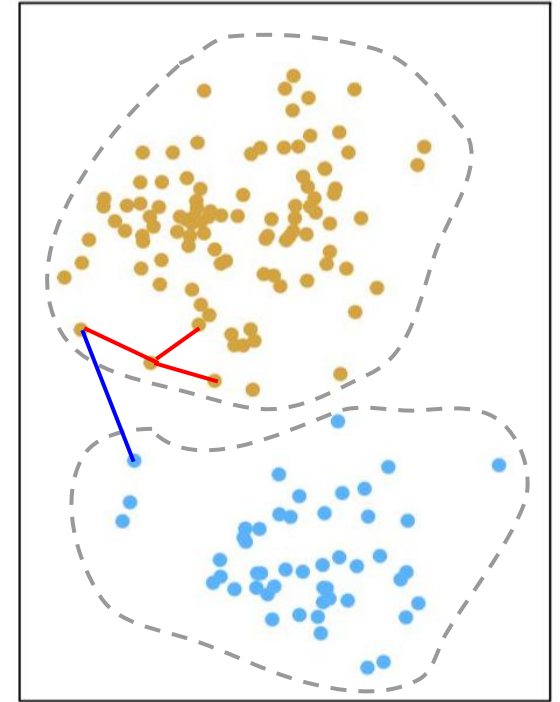
- If there are no labels, how do we know if results are meaningful?
 - Experts might interpret the result (external evaluation)
- However, we need it, because:
 - Labeling large datasets is very costly
 - We may have no idea what/how many classes there are (data mining)
 - We may want to use **clustering** to gain some insight into the structure of the data before designing a classifier



Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

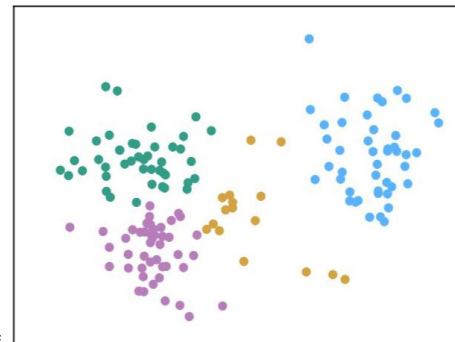
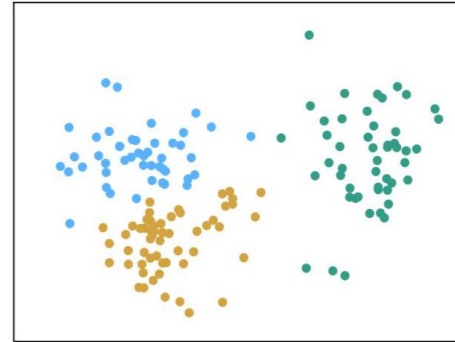
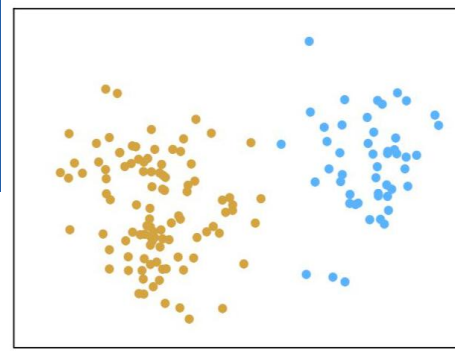
- What is a good clustering?
 - internal distances (**within the cluster**) should be small
 - External distances (**inter-cluster**) should be large
- Clustering is a way to discover new categories



Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

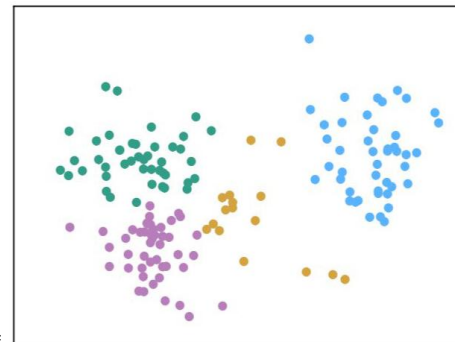
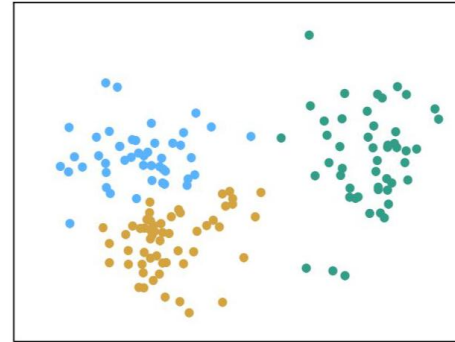
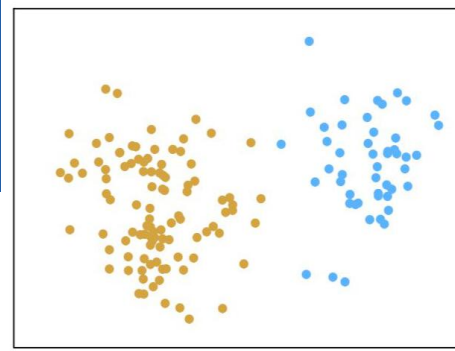
- A **clustering** is a partition $\{C_1, \dots, C_k\}$, where each C_k denotes a subset of the observations.
- Each observation belongs to one and only one of the clusters
- To denote that the i^{th} observation is in the k^{th} cluster, we write $i \in C_k$



Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

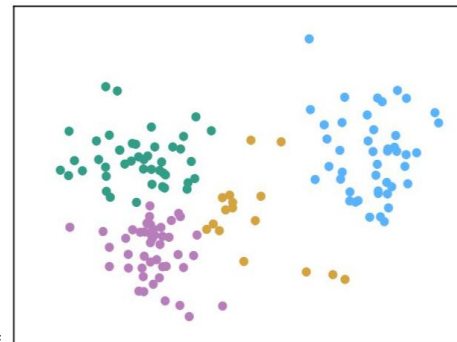
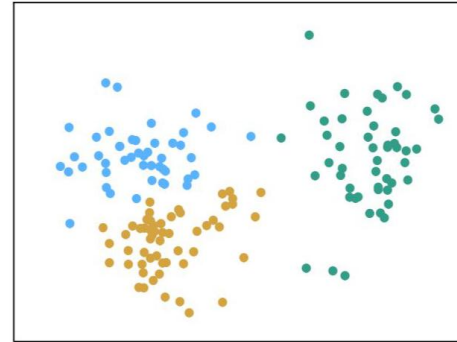
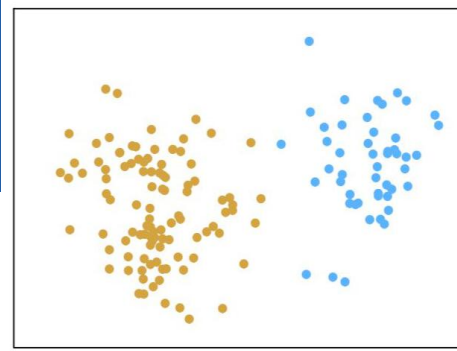
- To start with, we need a **proximity measure**



Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

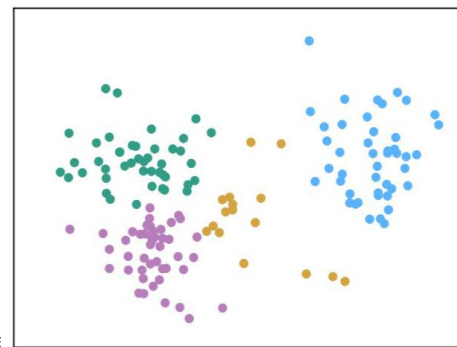
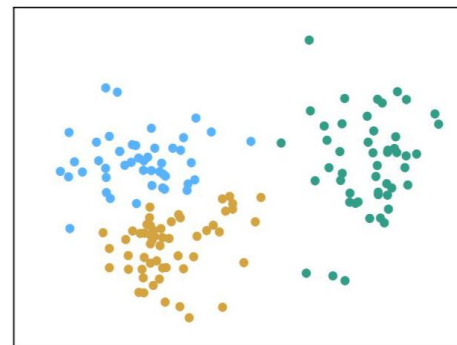
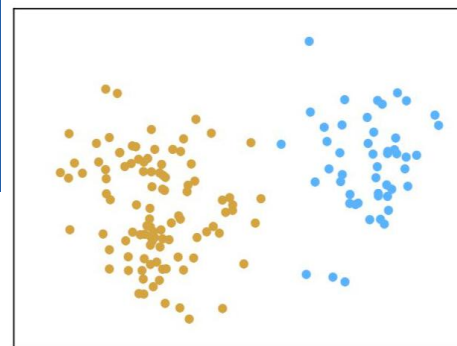
- To start with, we need a **proximity measure**
 - **similarity measure**
 $s(x_i, x_k)$: large, if x_i and x_k are similar



Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

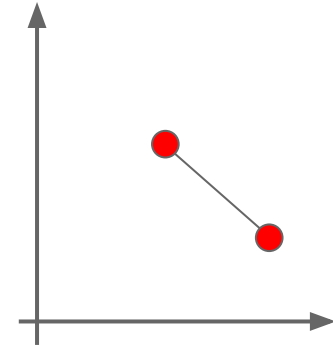
- To start with, we need a **proximity measure**
 - **similarity measure**
 $s(x_i, x_k)$: large, if x_i and x_k are similar
 - **dissimilarity measure** (distance)
 $d(x_i, x_k)$: small, if x_i and x_k are similar



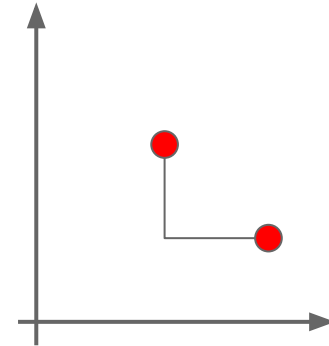
Distance Measures

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- **Euclidean distance:**
$$d(x_i, x_j) = \sqrt{\sum_{k=1}^d (x_i^{(k)} - x_j^{(k)})^2}$$
 - translation invariant



- **Manhattan distance:**
$$d(x_i, x_j) = \sum_{k=1}^d |x_i^{(k)} - x_j^{(k)}|$$
 - less complex to compute



K-Means Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- **Main idea:**
A good clustering is one for which the **within-cluster variation** is as small as possible.
- The **within-cluster variation** $WCV(C_k)$ for cluster C_k is some measure of the amount by which the observations within each class differ from one another
- **Goal:** Find $C_1, \dots, C_K$ that minimize $\sum_{k=1}^K WCV(C_k)$

“Partition the observations into K clusters such that the WCV summed up over all K clusters is as small as possible.”

K-Means Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- **Determine within-cluster variation**

- **Goal:** Find $C_1, \dots, C_K$ that minimize $\sum_{k=1}^K WCV(C_k)$
- Use Euclidean distance:
$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

where $|C_k|$ denotes the number of observations in cluster K

- The total number of clusters K is a fixed parameter.

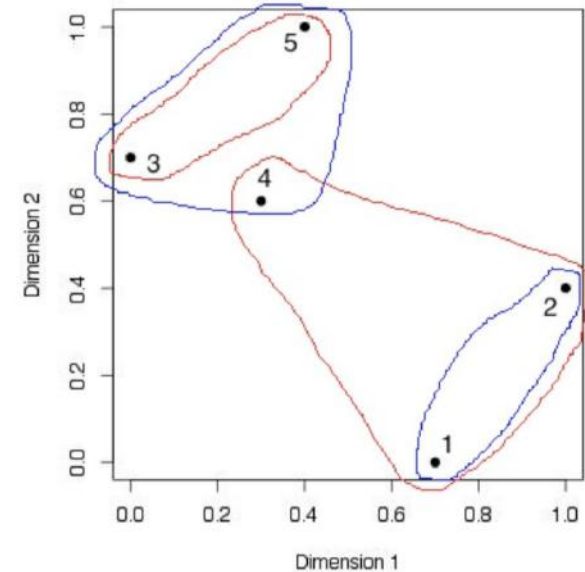
K-Means Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- **Simple Example**

- $n=5$ and $K=2$
- distance matrix

	1	2	3	4	5
1	0	0.25	0.98	0.52	1.09
2	0.25	0	1.09	0.53	0.72
3	0.98	1.09	0	0.10	0.25
4	0.52	0.53	0.10	0	0.17
5	1.09	0.72	0.25	0.17	0



- **Red clustering:** $\sum WCV_k = (0.25 + 0.53 + 0.52)/3 + 0.25/2 = 0.56$
- **Blue clustering:** $\sum WCV_k = 0.25/2 + (0.10 + 0.17 + 0.25)/3 = 0.30$

K-Means Clustering

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- $WCV(C_k)$ can be rewritten:
$$\begin{aligned} WCV(C_k) &= \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|(x_i - x_{i'})\|_2^2 \\ &= 2 \sum_{i \in C_k} \|x_i - \bar{x}_k\|^2 \end{aligned}$$
- With $\bar{x}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$ is just the average of all the points in Cluster C_k

K-Means Algorithm

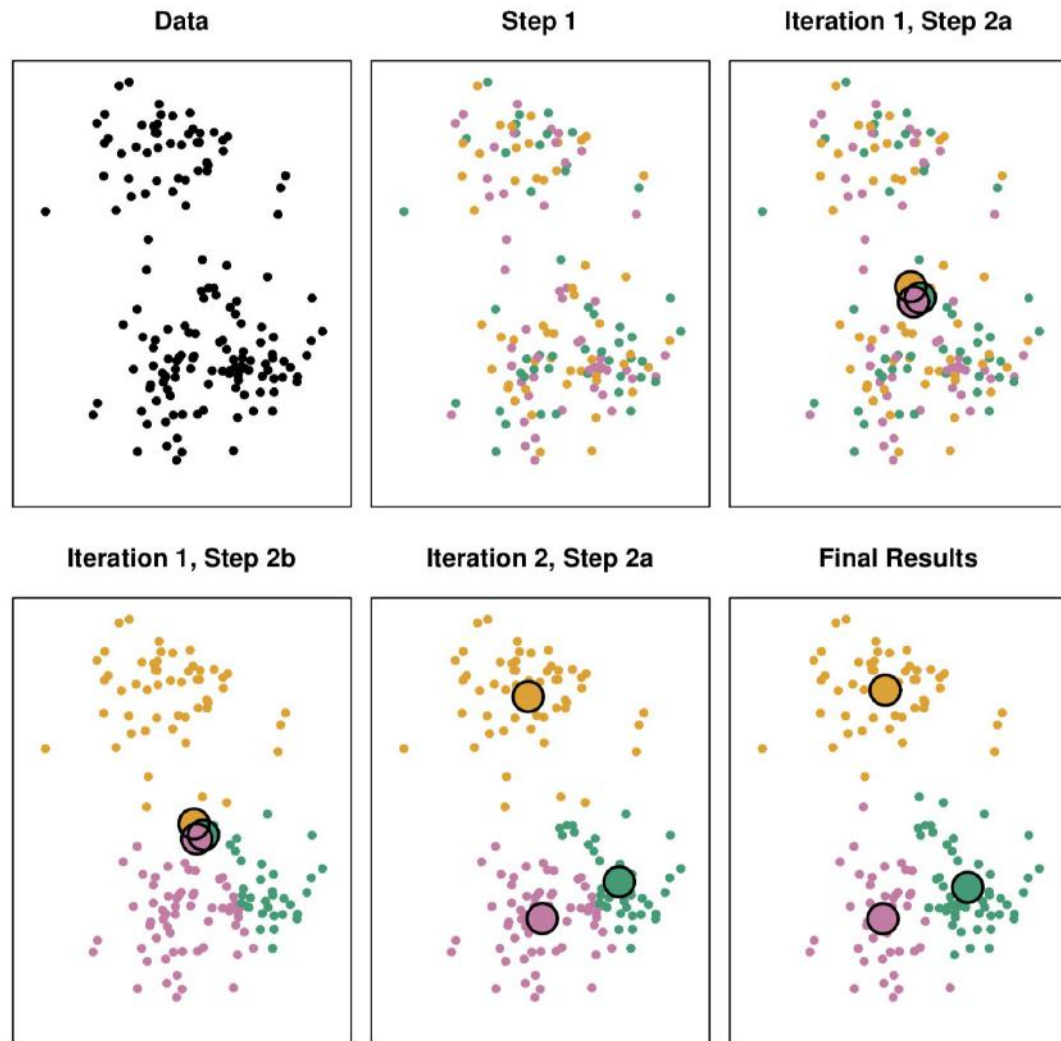
Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

1. Start by **randomly** partitioning the observations into **K clusters**
2. Until the clusters stop changing, repeat:
 - a. For each cluster, compute the **cluster centroid**
 - b. Assign each observation to the cluster whose centroid is the closest

K-Means Algorithm

Information Service Engineering - Lecture 5 Basic Machine

- Example: $K=3$



K-Means Summary

Information Service Engineering - Lecture 5 Basic Machine Learning - 5.6 Unsupervised Learning

- It is infeasible to actually optimize $WCV(C_k)$ in practice, but K-means provides a so-called **local optimum** of this objective
- The achieved result depends both on K, and also on the random initialization
- It is a good idea to try different random starts and pick the best result among them
- There is a method called **K-means++** that improves how the clusters are initialized



Picture References:

- P.17: Utagawa Kuniyoshi (1798-1861), The Actor, https://commons.wikimedia.org/wiki/File:Kuniyoshi_Utagawa,_The_actor_17.jpg