CLUSTERING

Anand Paul

Clustering - Definition

- **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).

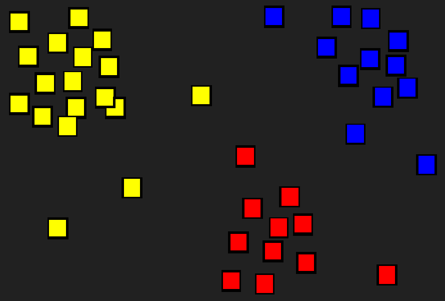
Goal of Clustering

- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes
- As a stand-alone tool to get insight into data distribution
- As a preprocessing step for other algorithms

Goal of clustering

- group data points that are close (or **similar**) to each other
- identify such groupings (or clusters) in an **unsupervised** manner
  - Unsupervised: no information is provided to the algorithm on which data points belong to which clusters

- Example



The result of a cluster analysis shown as the coloring of the squares into three clusters.

1– Nearest neighbor

- Input: Query article A1
- Output: *Most* similar article (out of all articles available in the corpus.)
    - Algorithm:

○ - Search over each article $X_{1,}X_2.X_3$… in the corpus

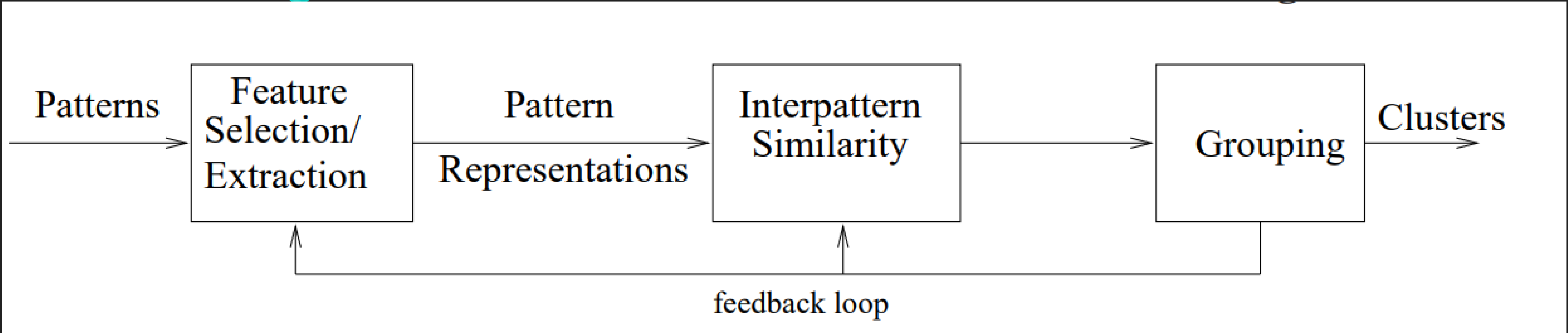
- Compute x = similarity(A_1, X_1)
- If X_1 has similarity store X1 then compare with other X_n in the corpus
- If $X_{21} > _{X1}$, store and set = X_{21}

○ - Return X_{21}

K Nearest Neighbor

- Input: Query article
- Output: *List of k* similar articles

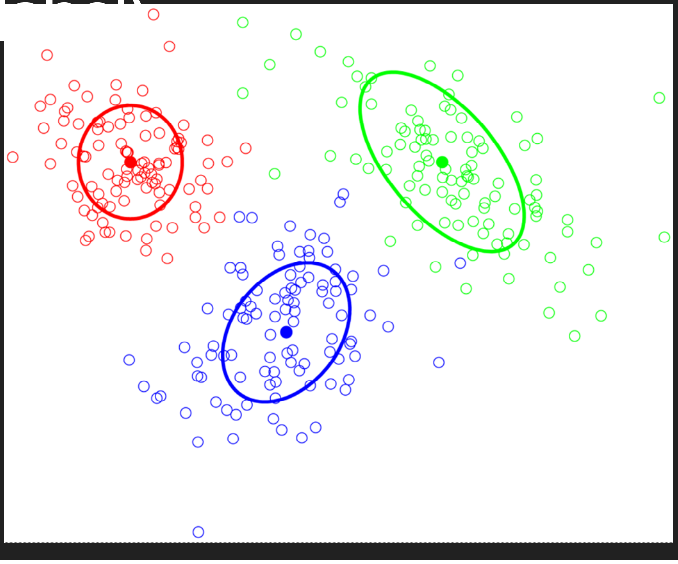
Stages of Clustering



A.K. Jain and et al, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999

- Cluster defined by center & shape/spread

- Assign observation (doc) to cluster (topic label)
  - Score under cluster is higher than others
  - Often, just more similar to assigned cluster center cluster centers



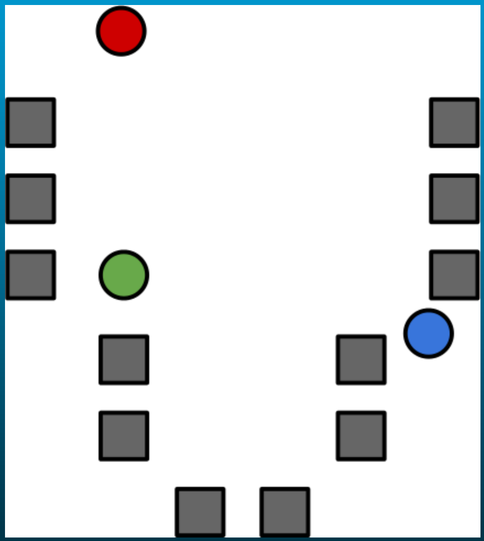
K – Means

k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

k-means

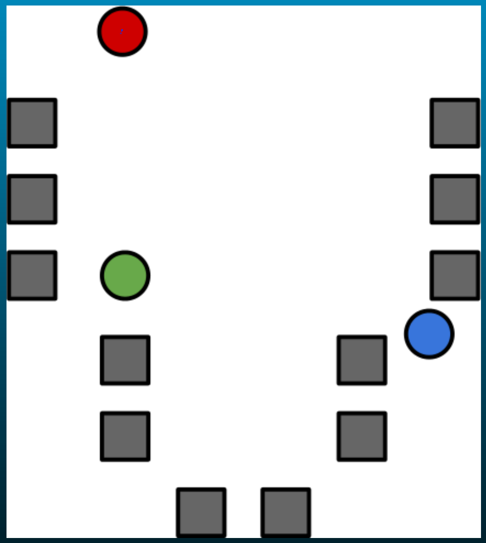
- Assumption

  -Similarity metric = distance to cluster center (smaller better)



k-means algorithm

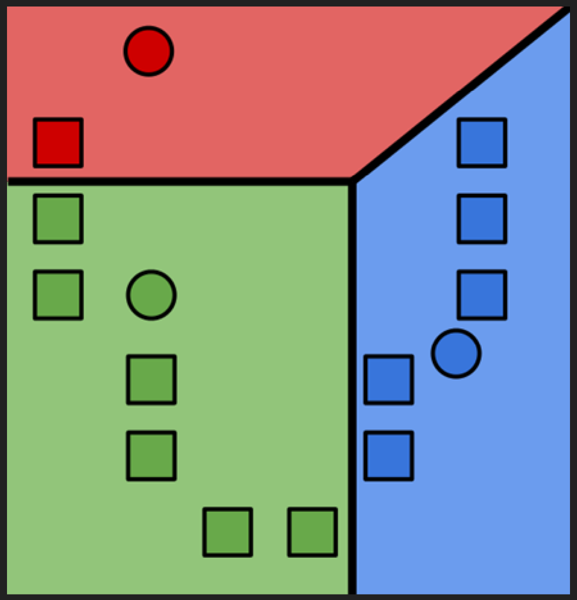
0. Initialize cluster centers



k-means algorithm

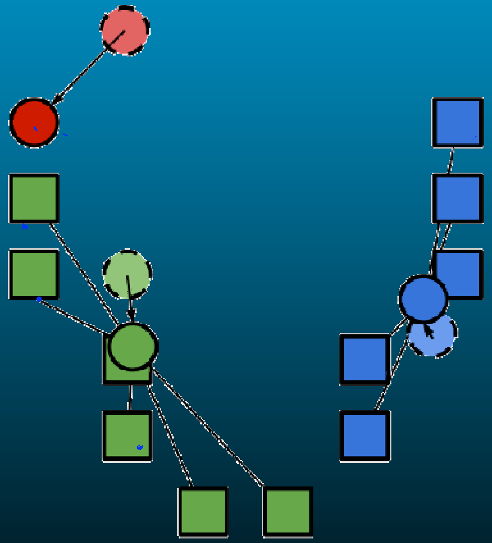
Initialize cluster centers

Assign observations to closest cluster center



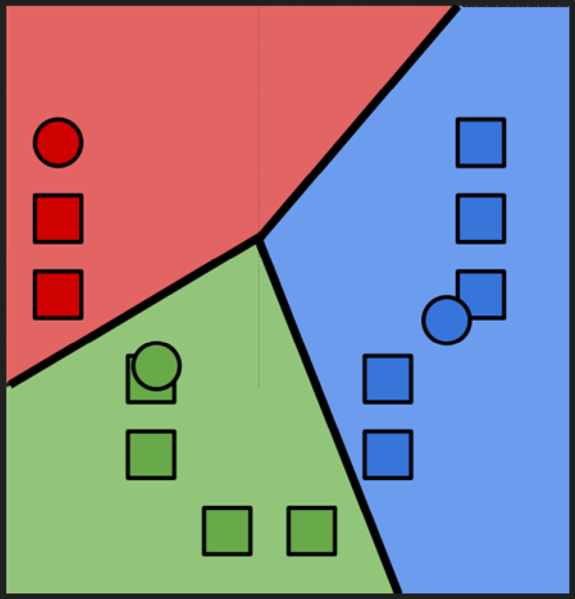
k-means algorithm

1. Initialize cluster centers
2. Assign observations to closest cluster center
3. Revise cluster centers as mean of assigned observations



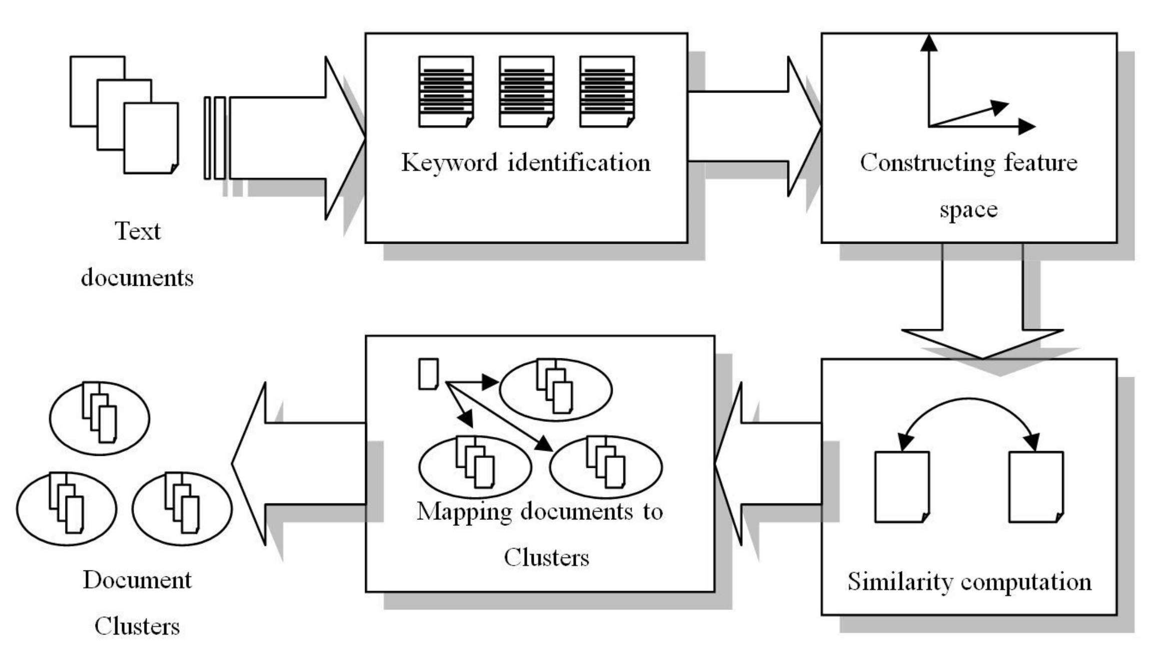
k-means algorithm

1. Initialize cluster centers
2. Assign observations to closest cluster center
3. Revise cluster centers as mean of assigned observations
4. Repeat 1.+2. until convergence

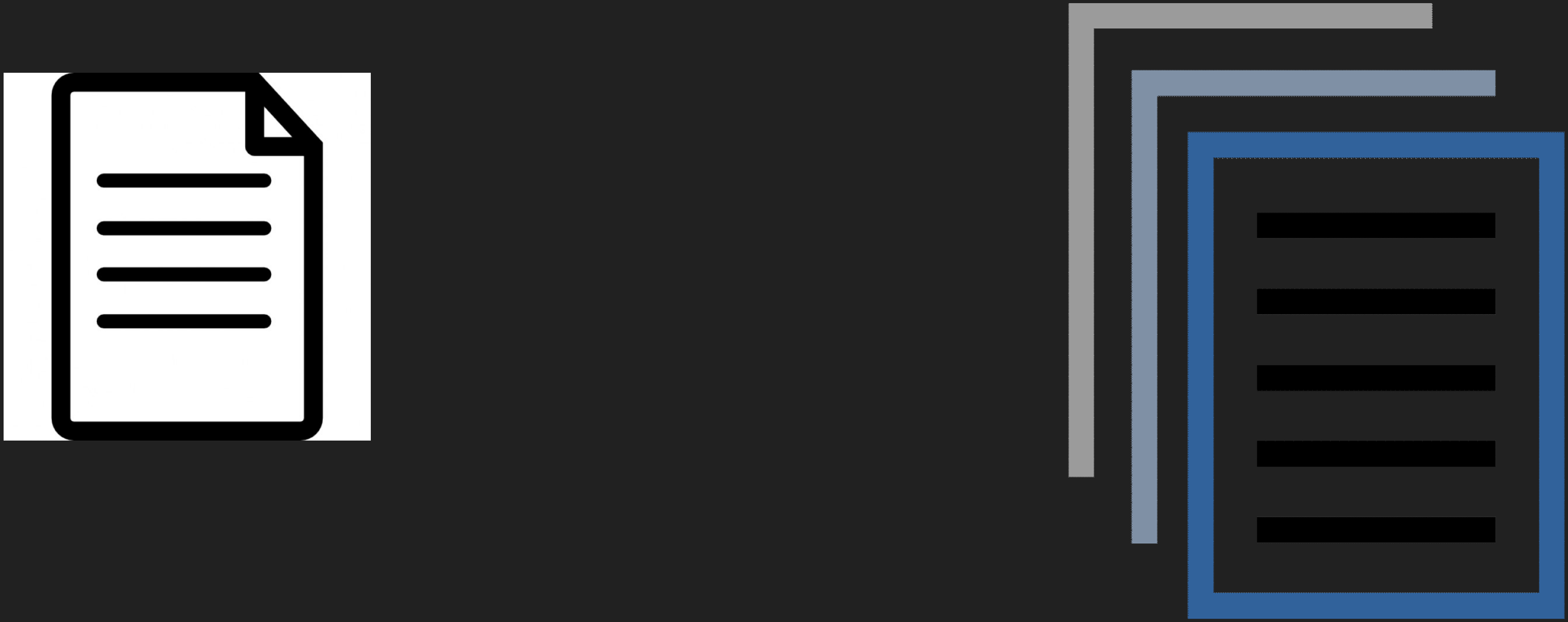


Example : Document Retrieval

- **Document retrieval** is defined as the matching of some stated user query against a set of free-text records. These records could be any type of mainly unstructured text, such as newspaper articles, real estate records or paragraphs in a manual.

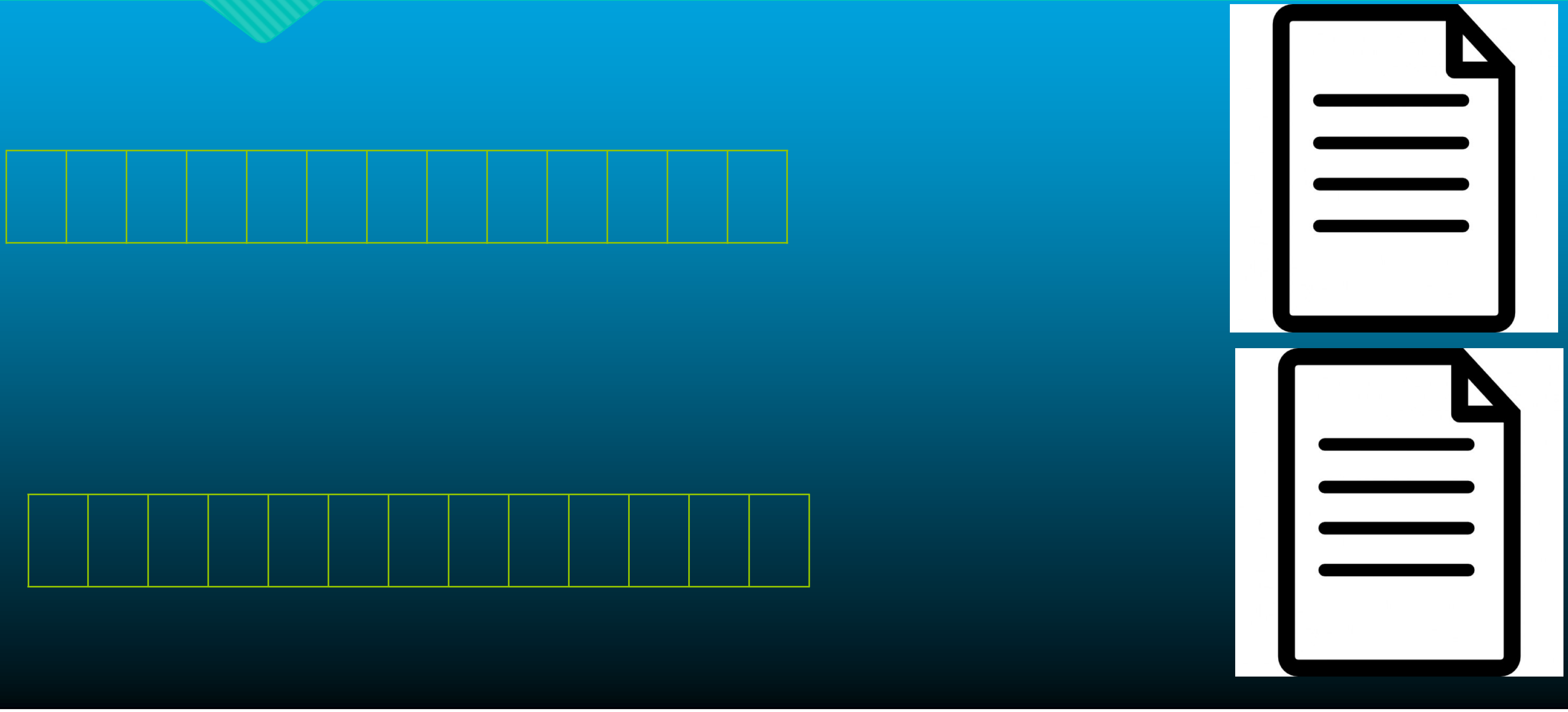


Yuan-Chao Liu, Ming Liu and Xiao-Long Wang, "Application of Self-Organizing Maps in Text Clustering: A Review", Chapter 9 from the book Applications of Self-Organizing Maps



Objective is to find a similar document

Word Count Similarity



Summary of Clustering

- **Clustering** is the process of grouping similar objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset according to some defined distance measure.
- Application of Clustering: Data Mining
- Pattern recognition
- Document Retrival
- Bioinformatics
- Machine Learning
- Text mining
- City Planning

Application of Clustering

- Data Mining
- Pattern recognition
- Document Retrieval
- Bioinformatics
- Machine Learning
- Text mining
- City Planning