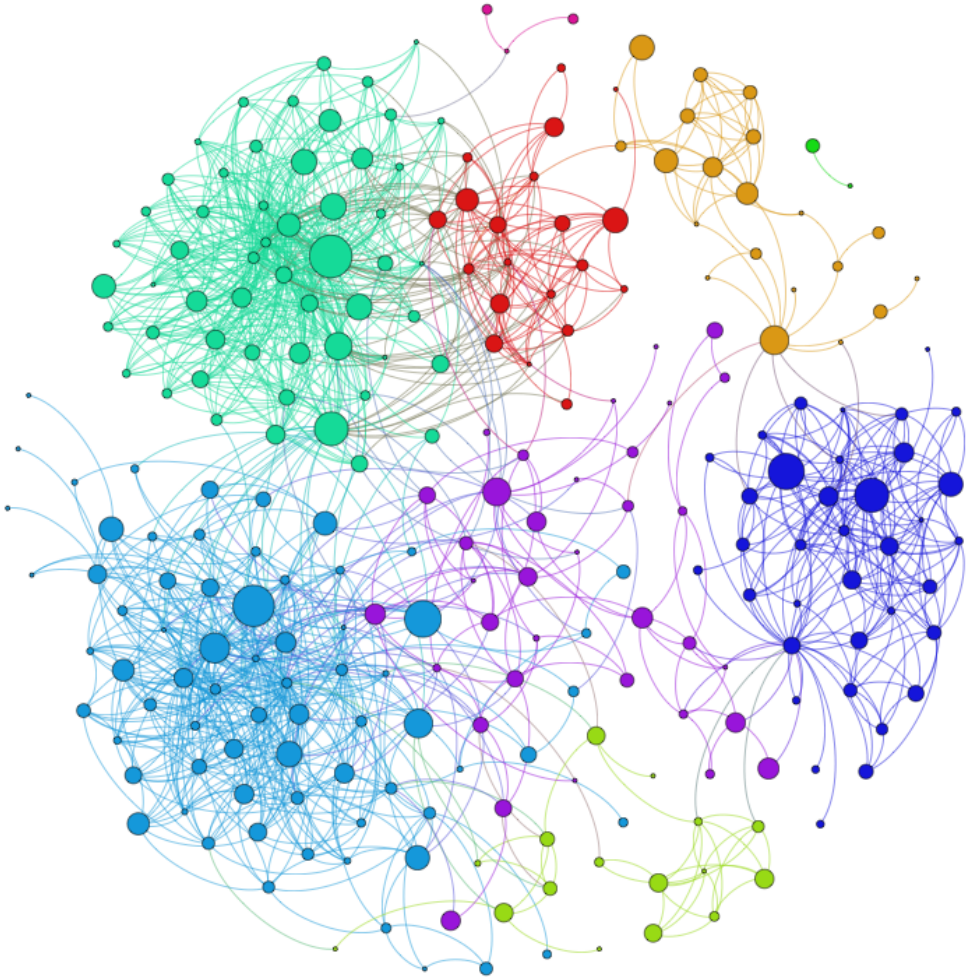


Introduction to Cluster Analysis

Instructor

Mousum Dutta
Chief Data Scientist, Spotle.ai
Computer Science, IIT KGP

Unsupervised learning

Unsupervised learning:

- Data with no target attribute.
- Need to find hidden structures from unlabeled data.
- Explore the data to find some natural patterns.

Cluster Analysis:

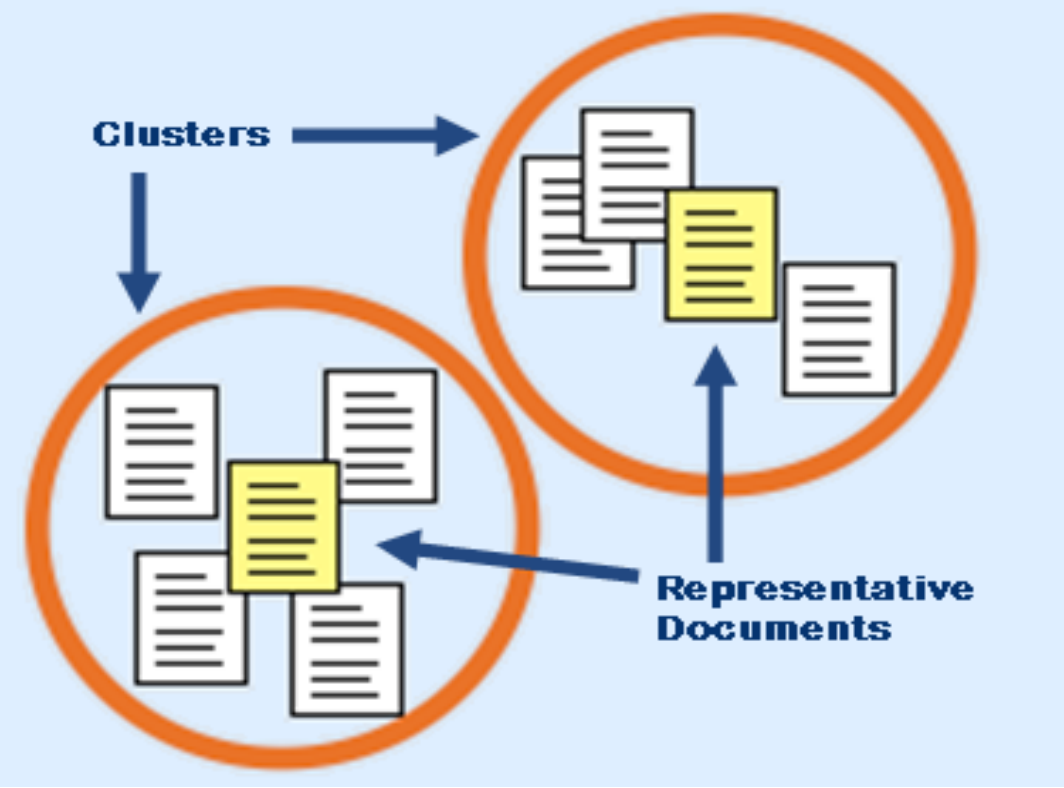
- Most popular type of unsupervised learning.
- Perform the task of grouping (called clusters) observations in such a way that most similar observations are in same groups and observations from different groups are relatively different.

Usefulness

- Divide large heterogeneous data to some homogeneous groups or clusters.
- Understanding hidden structures of data.
- Preprocessing for further analysis.

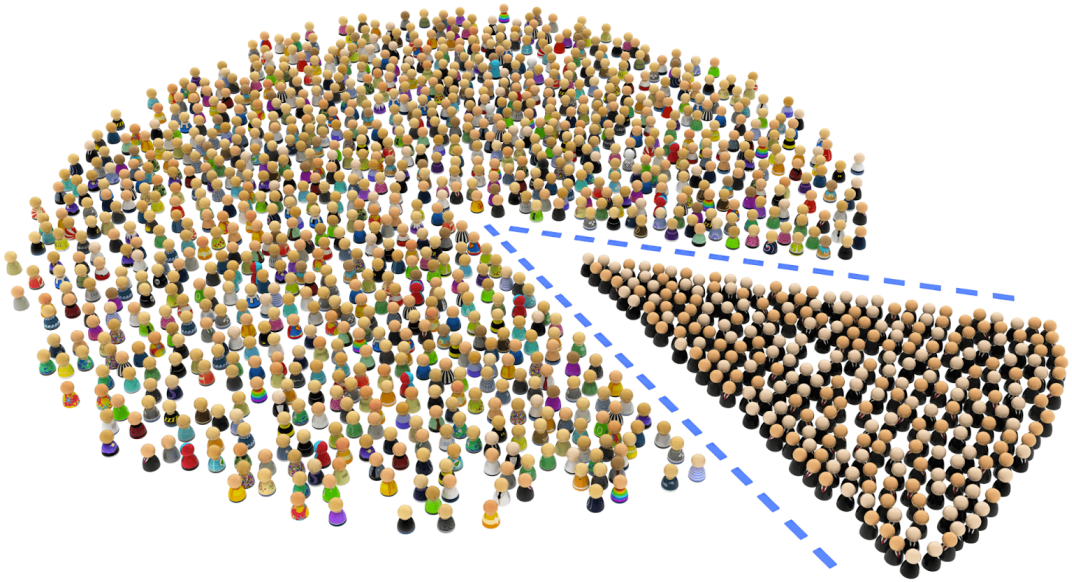
Example: Document clustering

Companies like pharmaceutical companies have thousands and thousands of similar reports from past few decades.
Can save many man hours if there are standards templates.



Example: Customer segmentation

The concept of segmentation relies on grouping the customers based on common demands and behaviours.
Clustering helps to divide whole heterogeneous market as the sum of smaller homogeneous markets.
Customer segmentation helps target campaigning, up-sell, cross-sell etc.

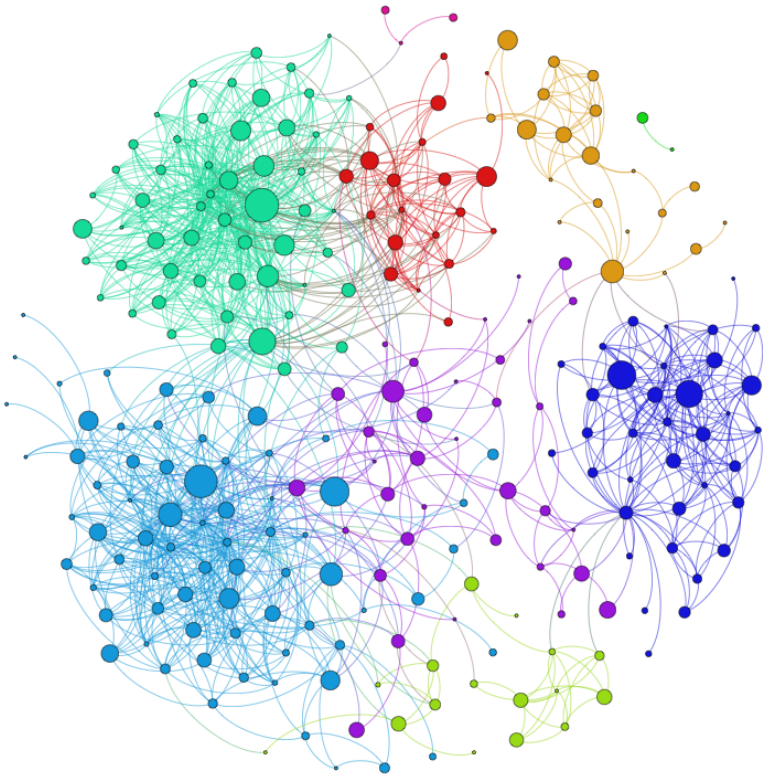


Example: Social network analysis

Social network is relationships and flows between entities like people, groups, organizations, countries, computers, URLs etc.

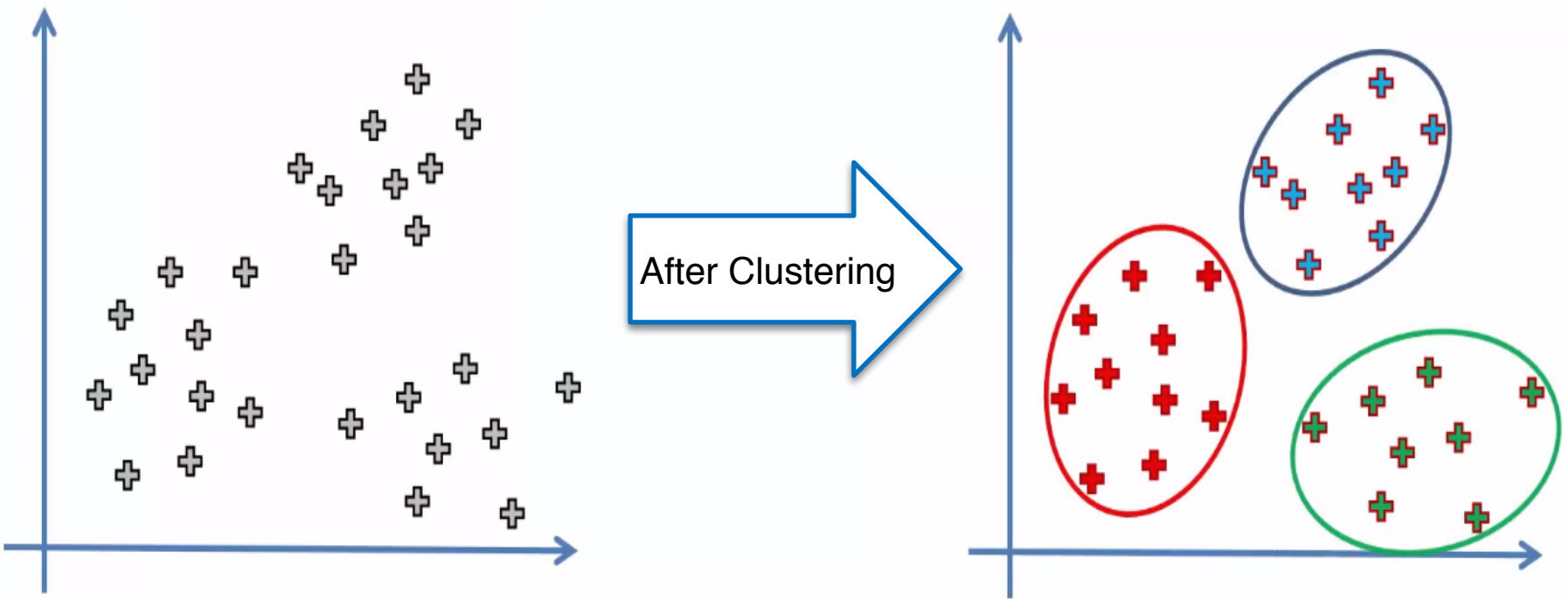
SNA helps to find out hidden groups or communities.

SNA is useful to prevent organized crime like terrorism or fraud.



Visualization

Visually that there are 3 natural clusters.
Minimum expectation from clustering algorithm is to form 3 clusters.



Features of cluster analysis

Clustering algorithms:

- Partitioned based clustering – k-means clustering
- Tree based clustering - Hierarchical clustering

Measure of similarities:

- Euclidean distance
- Minkowski family
- Cosine distance
- Correlation

Validity indices

- Optimal number of clusters.
- Quality of the clusters formed.

Upcoming topics

- K-means clustering in details
- Hierarchical clustering in details
- Popular similarity measures
- Popular cluster validity indices