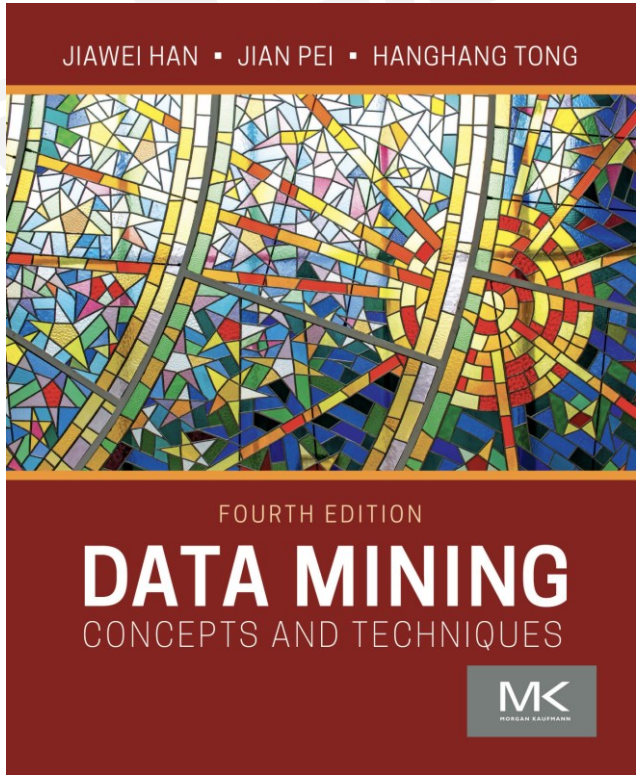
viettel

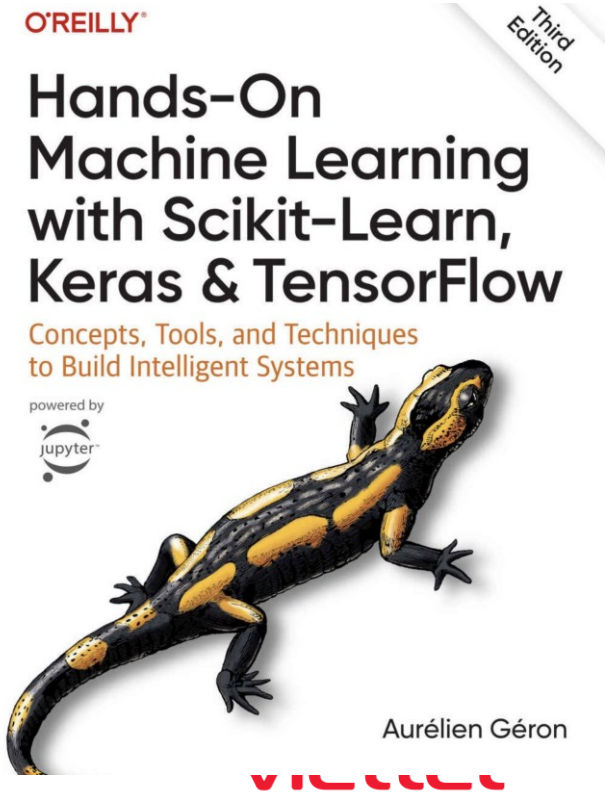
Clustering Analysis

Hoàng Anh Dũng
Trung tâm Phân tích Dữ liệu - Khối CNTT
Tổng Công ty Viễn thông Viettel

Reading

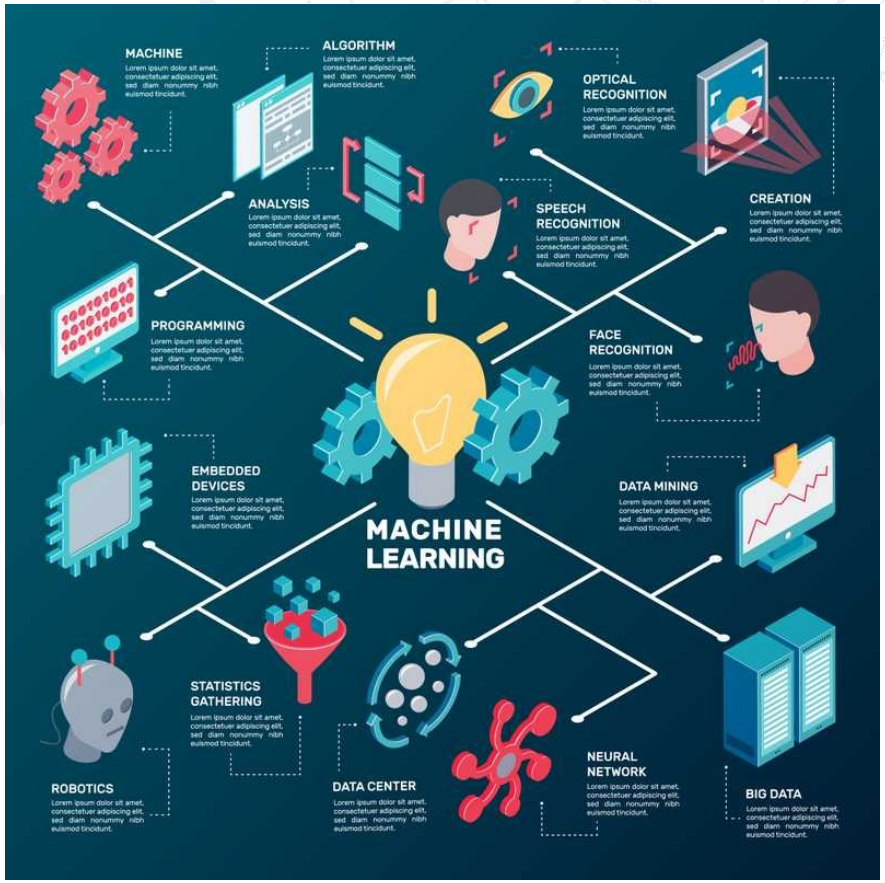
Chapter 8, 9 - Data Mining - Concepts and techniques - Morgan Kaufmann
Chapter 9 - Hands-On Machine Learning. - Aurélien Géron





1. Clustering - General Concepts

Main idea, real-life applications, types



Part I: Clustering – General Concepts

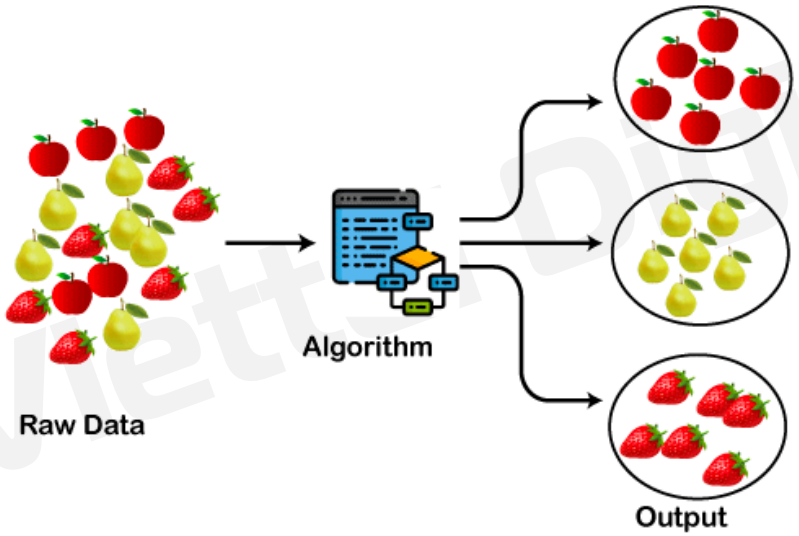
Real-life Applucations

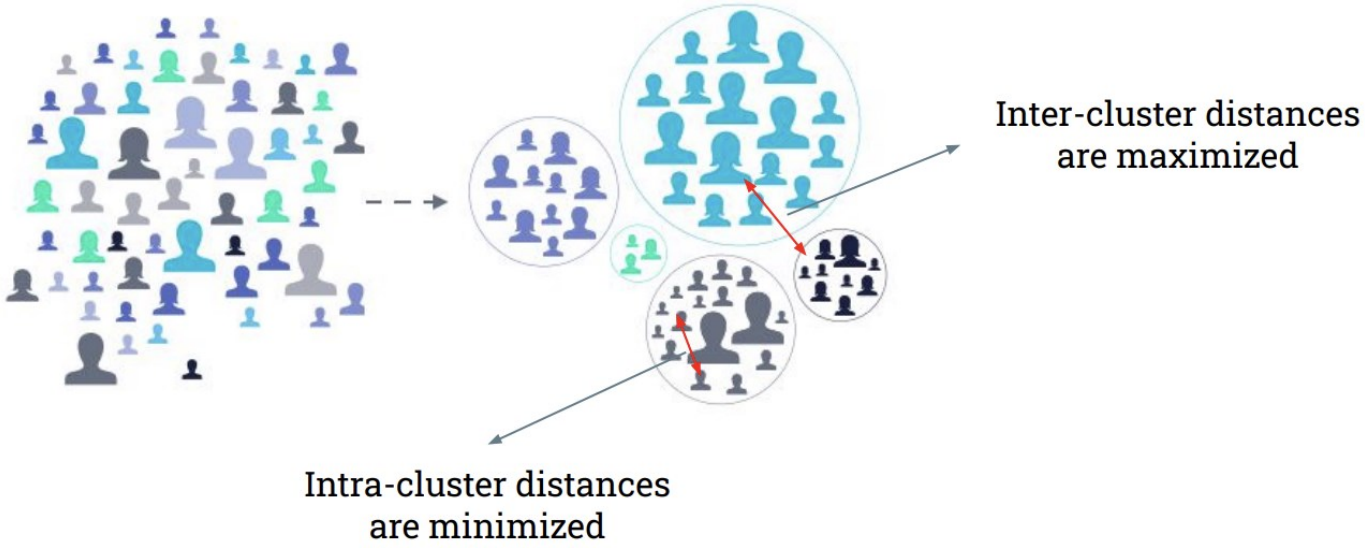
Types of Clusterings

Part II: Typical Clustering Algorithms

What is Cluster Analysis or Clustering?

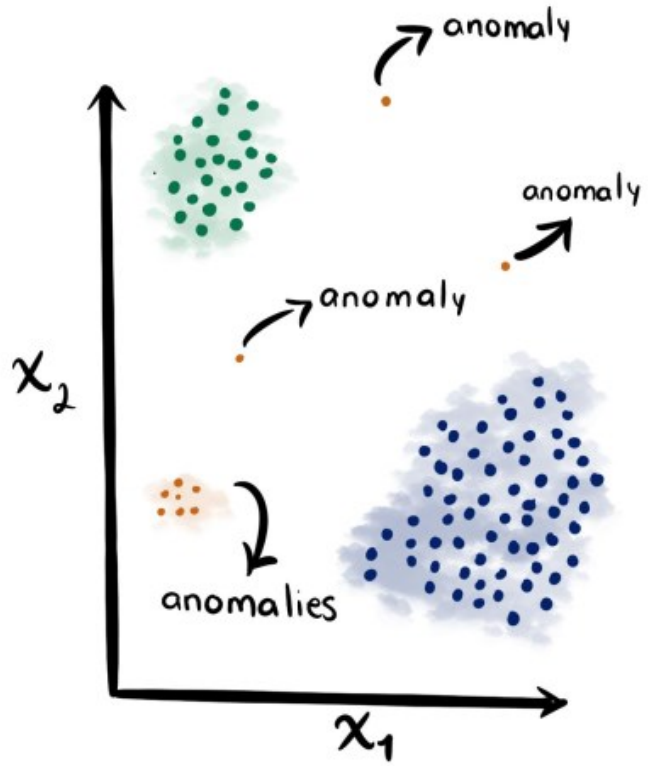
Given a set of objects, place them in **groups** such that the **objects in a group are similar** (or related) to **one another and different from (or unrelated to) the objects in other groups**





Real-life Applications:

Google News
Anomaly Detection

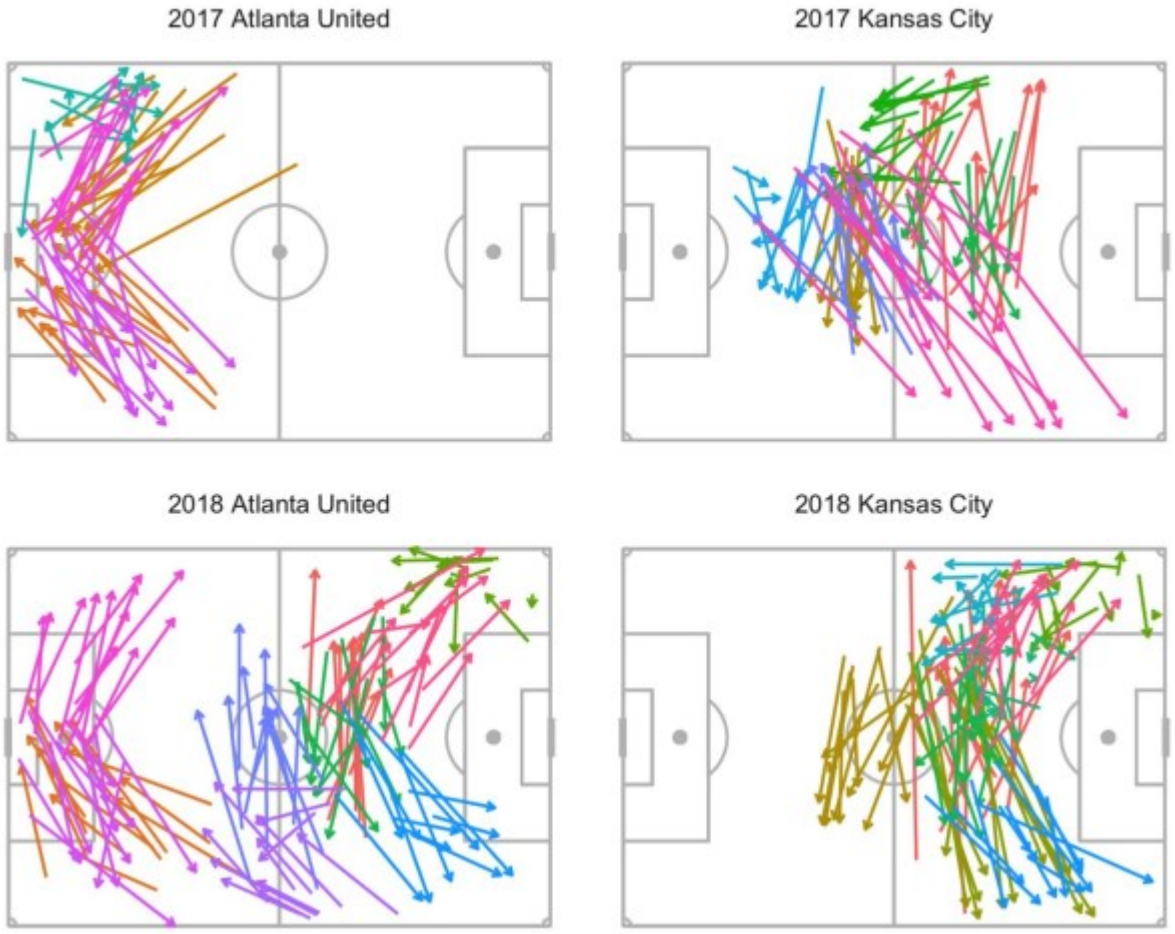


Fake News Detection
Fraud Detection
Spam Email Detection

Source:
https://towardsdatascience.com/unsupervised-anomaly-detection-on-spotify-data-k-means-vs-local-outlier-factor-f96ae783d7a7

Real-life Applications:

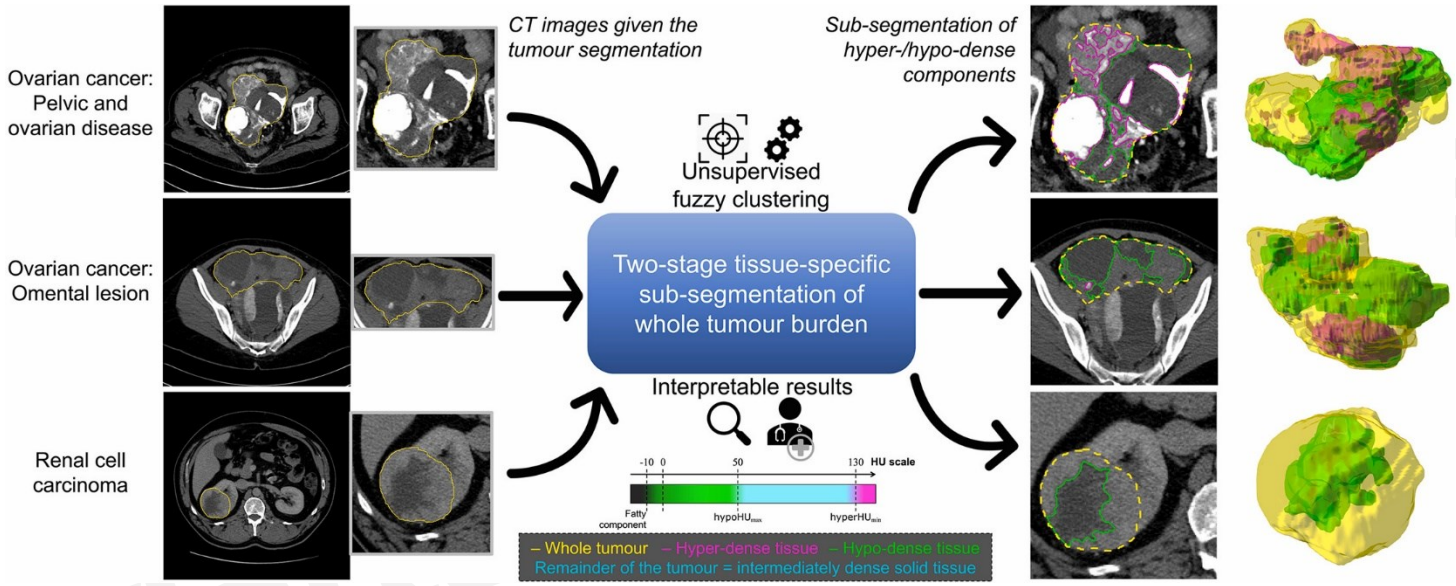
Sport Science Find players with similar styles

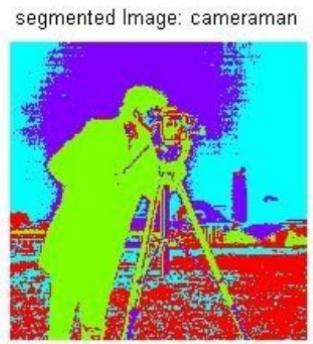


Source:
https://www.americansocceranalysis.com/home/2019/3/11/using-k-means-to-learn-what-soccer-passing-tells-us-about-playing-styles

Real-life Applications:

Image Segmentation

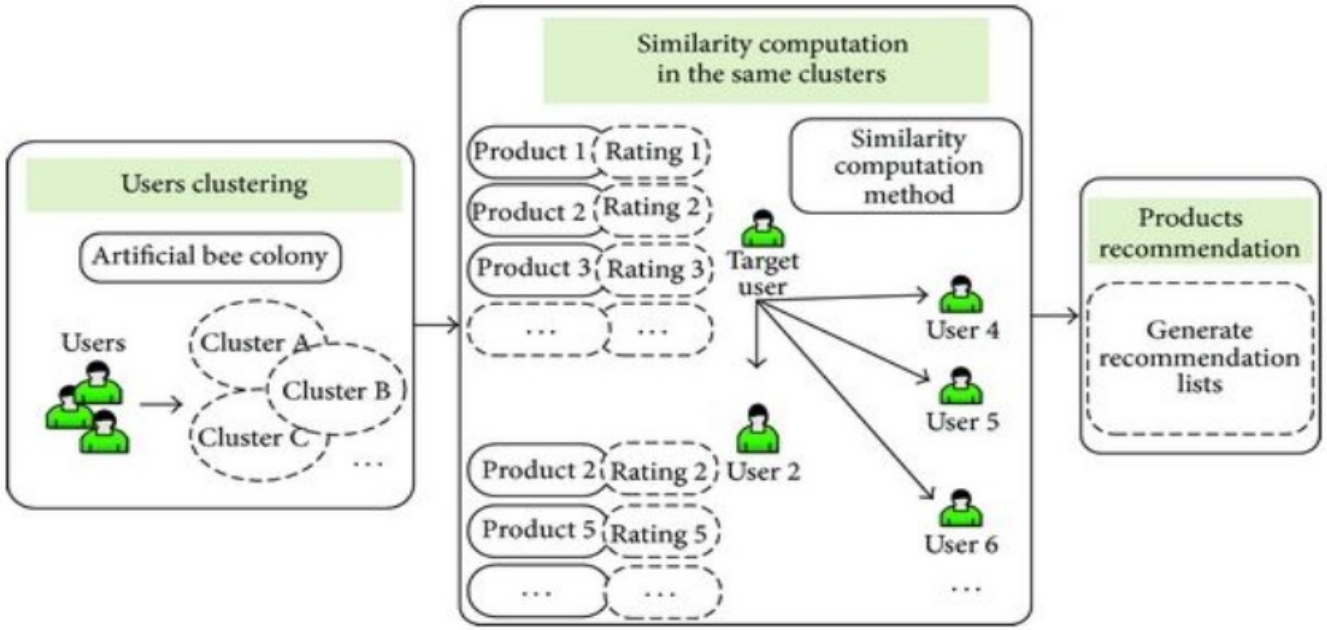




Source:
https://www.sciencedirect.com/science/article/pii/S0010482520301293
https://www.americansoccernalysis.com/home/2019/9/11/using-k-means-to-learn-what-soccer-passing-tells-us-about-playing-styles

Real-life Applications:

Recommendations



Cluster-based ranking
Group Recommendation

What do affect on Cluster Analysis?

Clustering

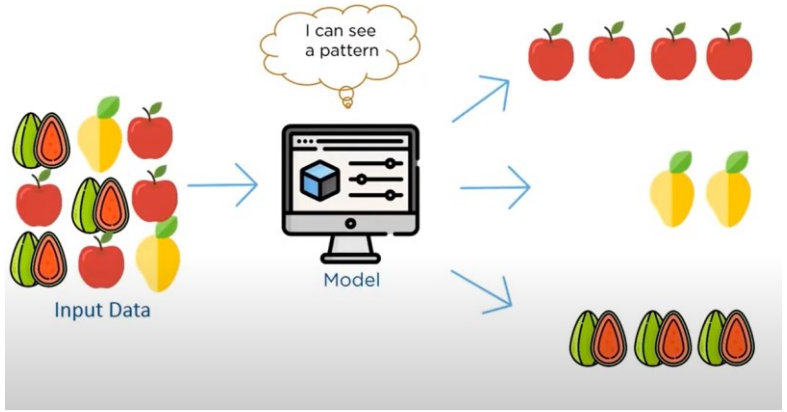
Data

Cluster

Non-labeled training data
No feedback
Find hidden structure in data

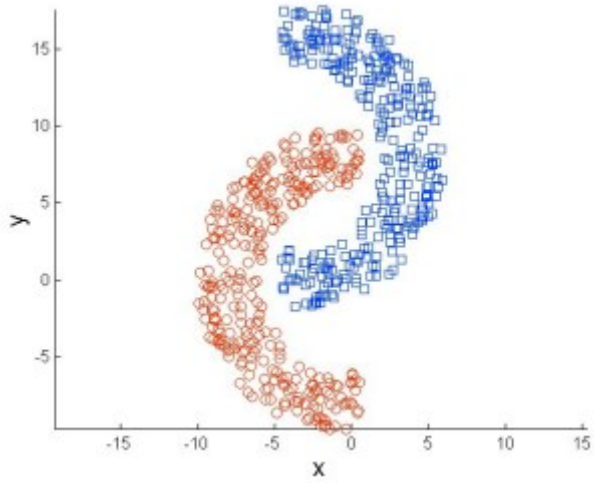
Algorithm

The machine learns



Characteristics of the Input Data Are Important

High dimensionality
- Dimensionality reduction
- **Types of attributes**
- Binary, discrete, continuous, asymmetric
- Mixed attribute types, e.g., continuous & nominal)
- **Differences in attribute scales**
- Normalization techniques
- **Size of data set**
- **Noise and Outliers**
- **Properties of the data space**



Characteristics of the Input Data Are Important

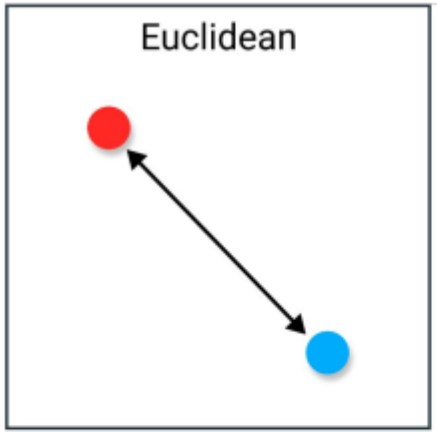
Data distribution

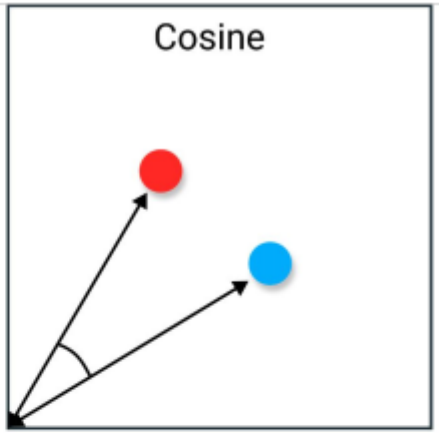
- Parametric models
- **Shape**
- Globular or arbitrary shape
- **Differing sizes**
- **Differing densities**
- **Level of separation among clusters**
- **Relationship among clusters**
- **Subspace clusters**

⇨ **Distance Metrics**

How to Measure the Similarity/Distance?

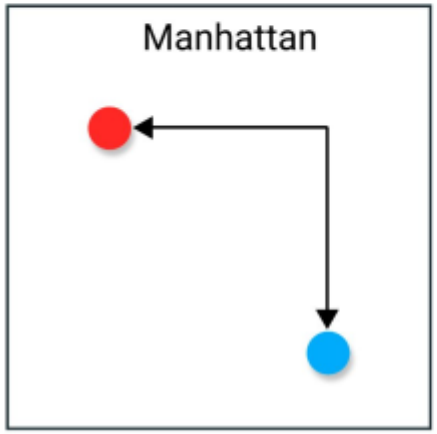
$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$

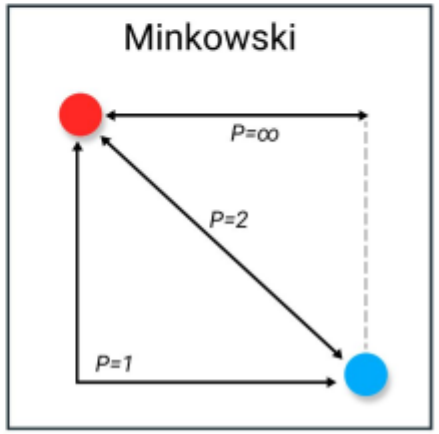




$$D(x,y) = cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

$$D(x,y) = \sum_{i=1}^{k} |x_i - y_i|$$





$$D(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Source: https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Notion of a Cluster can be Ambiguous

How many clusters?

4 Clusters

2 Clusters

6 Clusters

Types of Clusterings

01

Partitioning Methods

02

Hierachical Clustering

03

Fuzzy Clustering

04

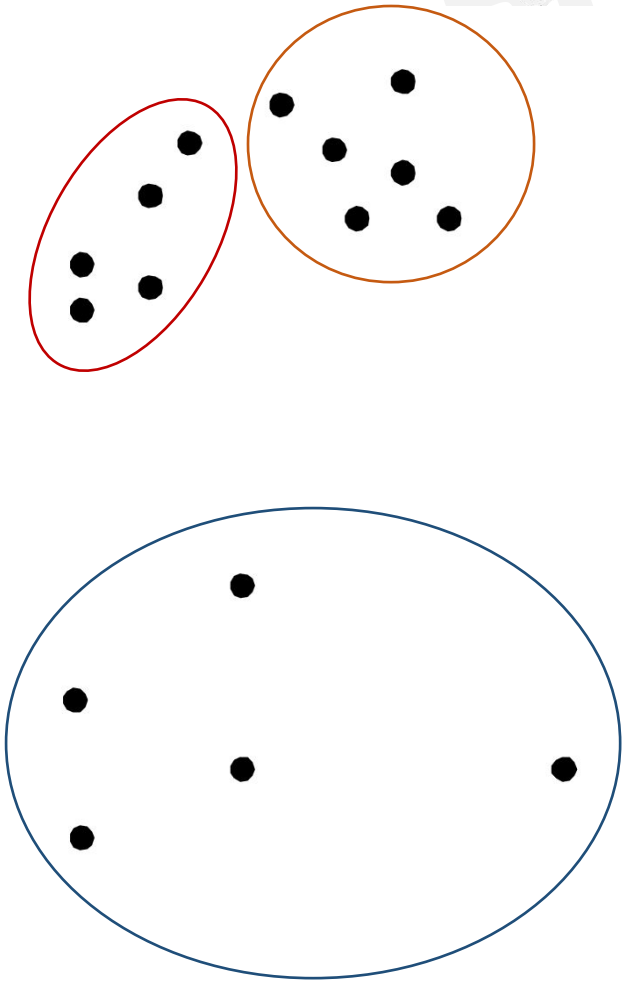
Density Based Clustering

05

Model Based Clustering

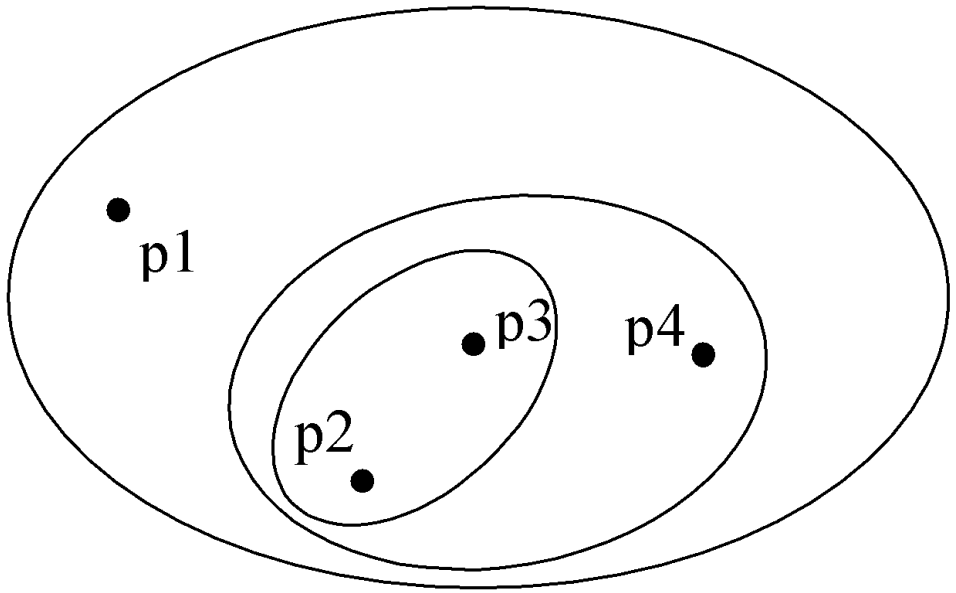
Partitional Clustering

Data objects are separated into **non-overlapping subsets, i.e., clusters**

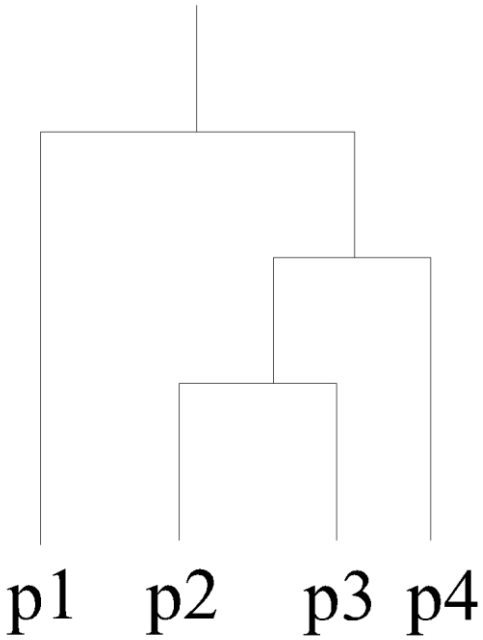


Hierarchical Clustering

Data objects are separated into
nested clusters as a hierarchical tree



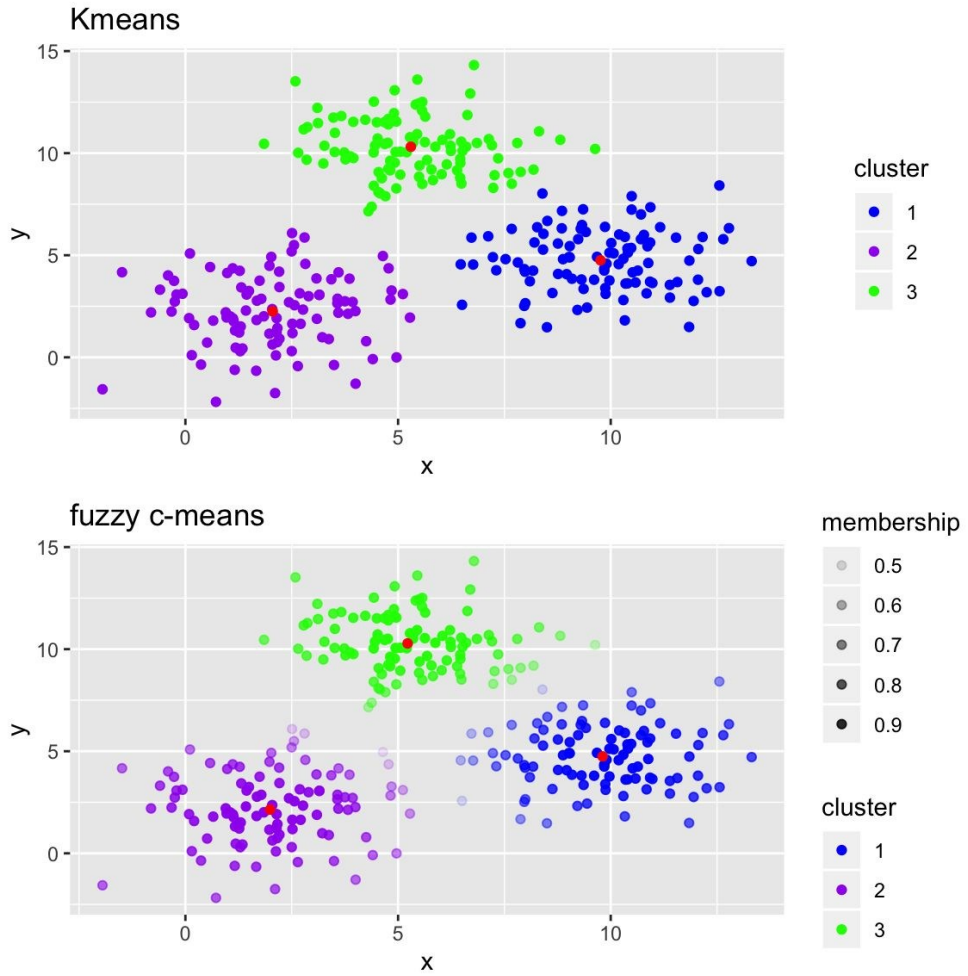
Hierarchical Clustering



Clustering dendrogram

Fuzzy Clustering

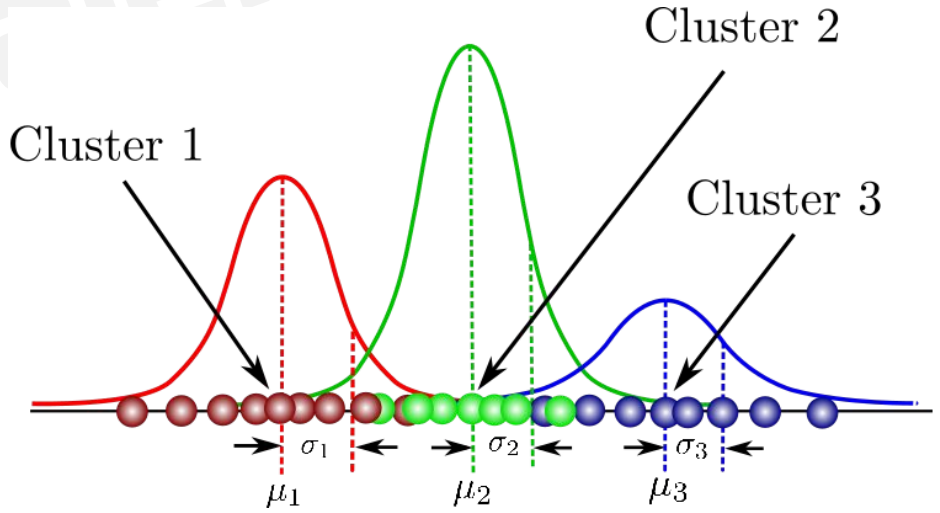
Fuzzy clustering, i.e., soft clustering, is a form of clustering in which **each data point can belong to more than one cluster with weights**



viettel

Model-based Clustering

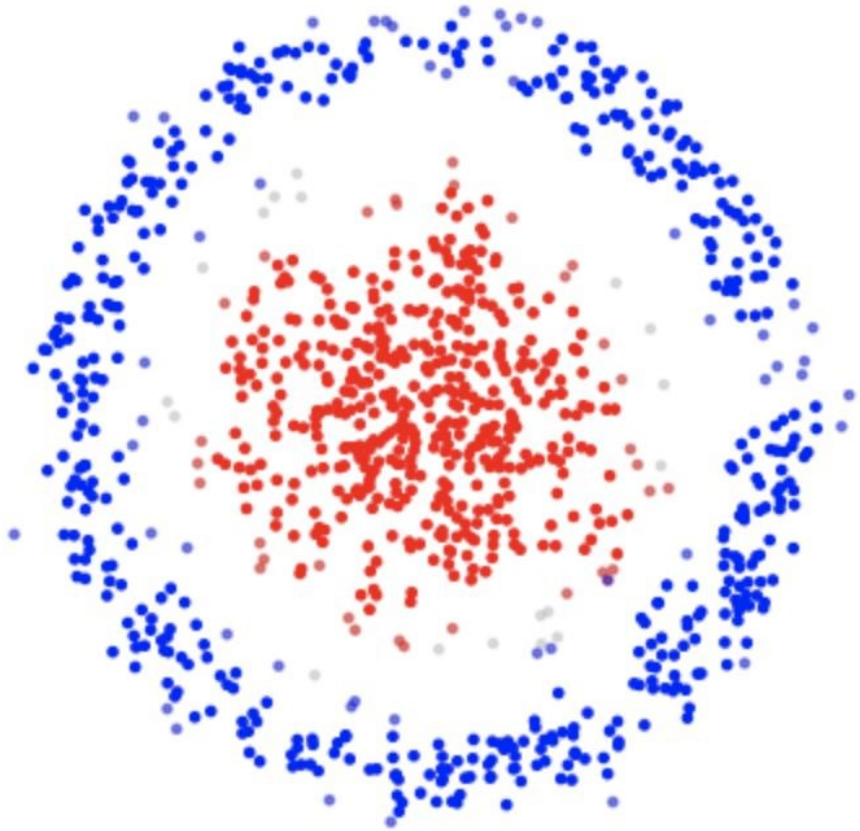
Model-based clustering assumes that **the data were generated by a model** and **tries to recover the original model from the data**.



Gaussian Mixture Model

Density-based Clustering

A cluster is **a dense region of points**, which is separated by **low-density regions**, from **other regions of high density**.



Non-linear separation

viettel

2.
Typical Clustering Algorithms

Intuition, Main Idea, Limitation

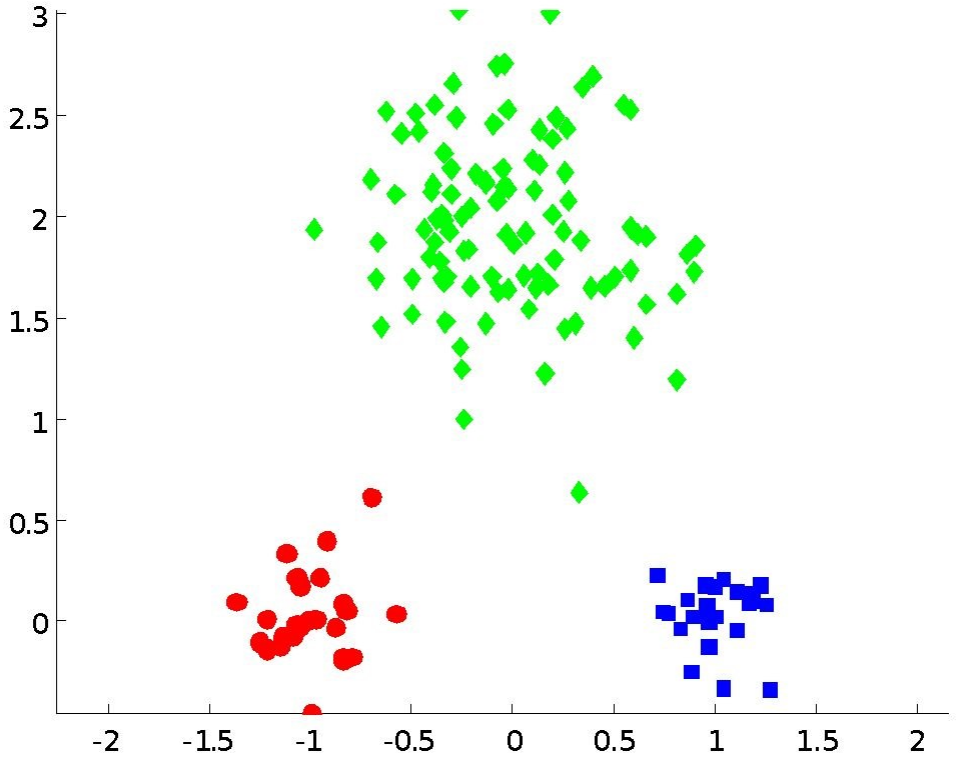
Typical Clustering Algorithms

◎ **Partitional Clustering**
- K-Means & Variants

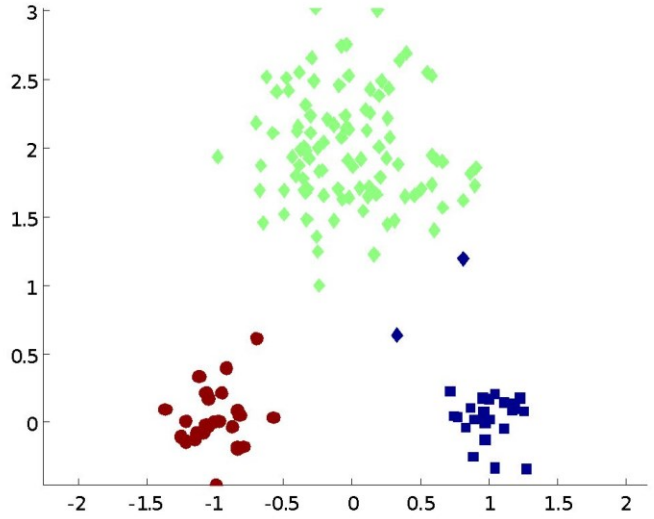
◎ **Hierarchical Clustering**
- HAC

◎ **Density-based Clustering**
- DBSCAN

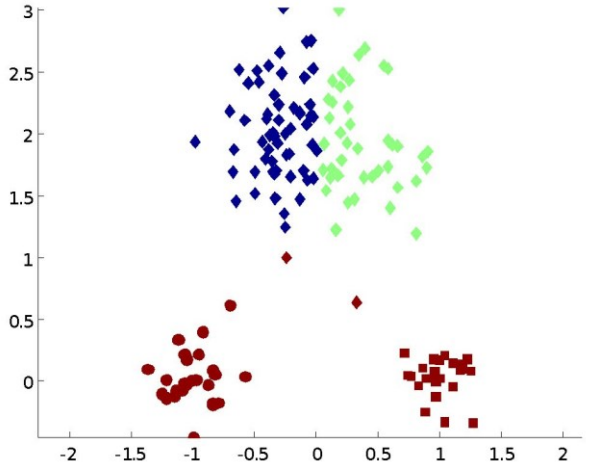
Two different K-means Clusterings



Original Points



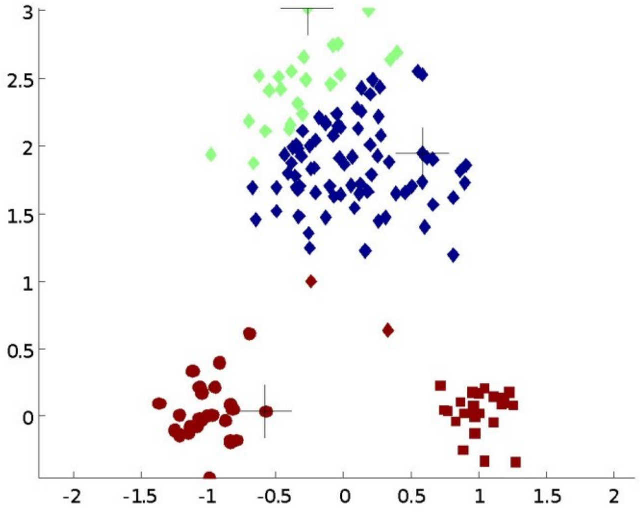
Optimal Clustering

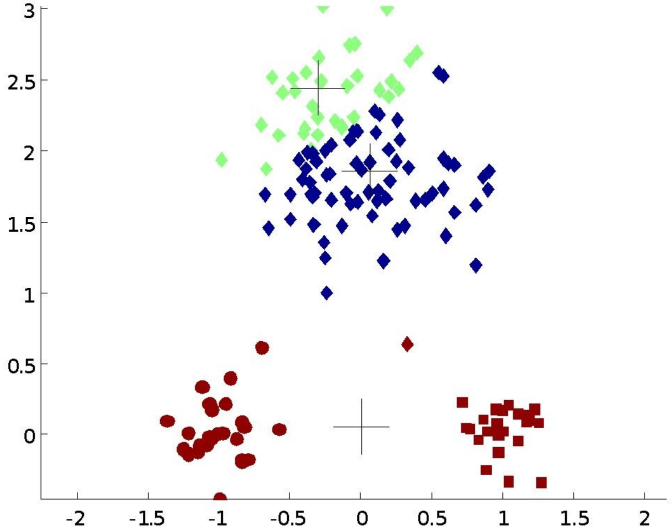


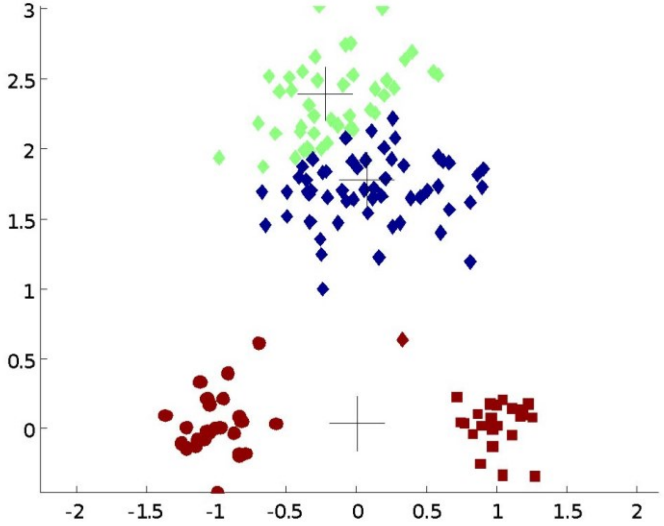
Sub-optimal Clustering

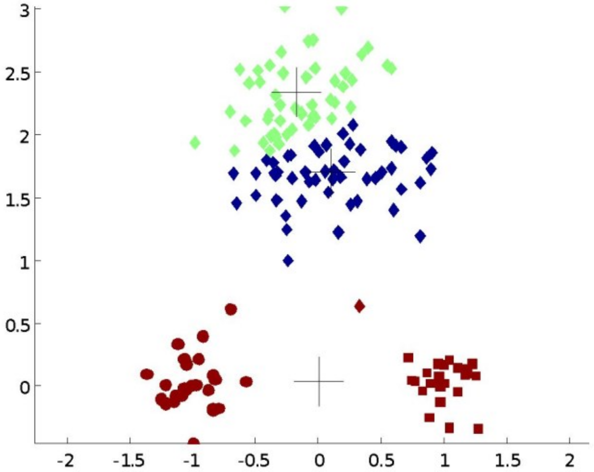
viettel

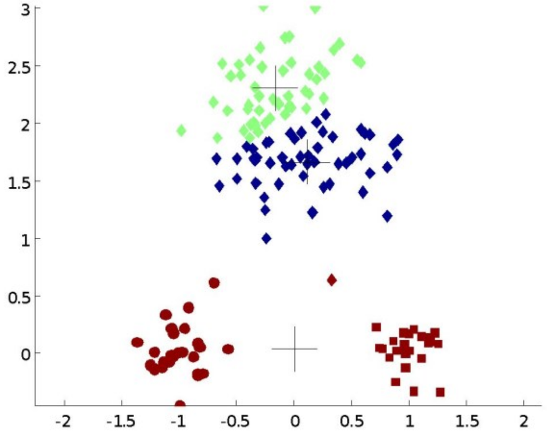
Importance of Choosing Initial Centroids











Solutions to Initial Centroids Problem

Multiple runs

- Helps, but probability is not on your side

Use some strategies to select the k initial centroids and then select among these initial centroids

- Select most widely separated, e.g., K-means++
- Use hierarchical clustering to determine initial centroids

Bisecting K-Means

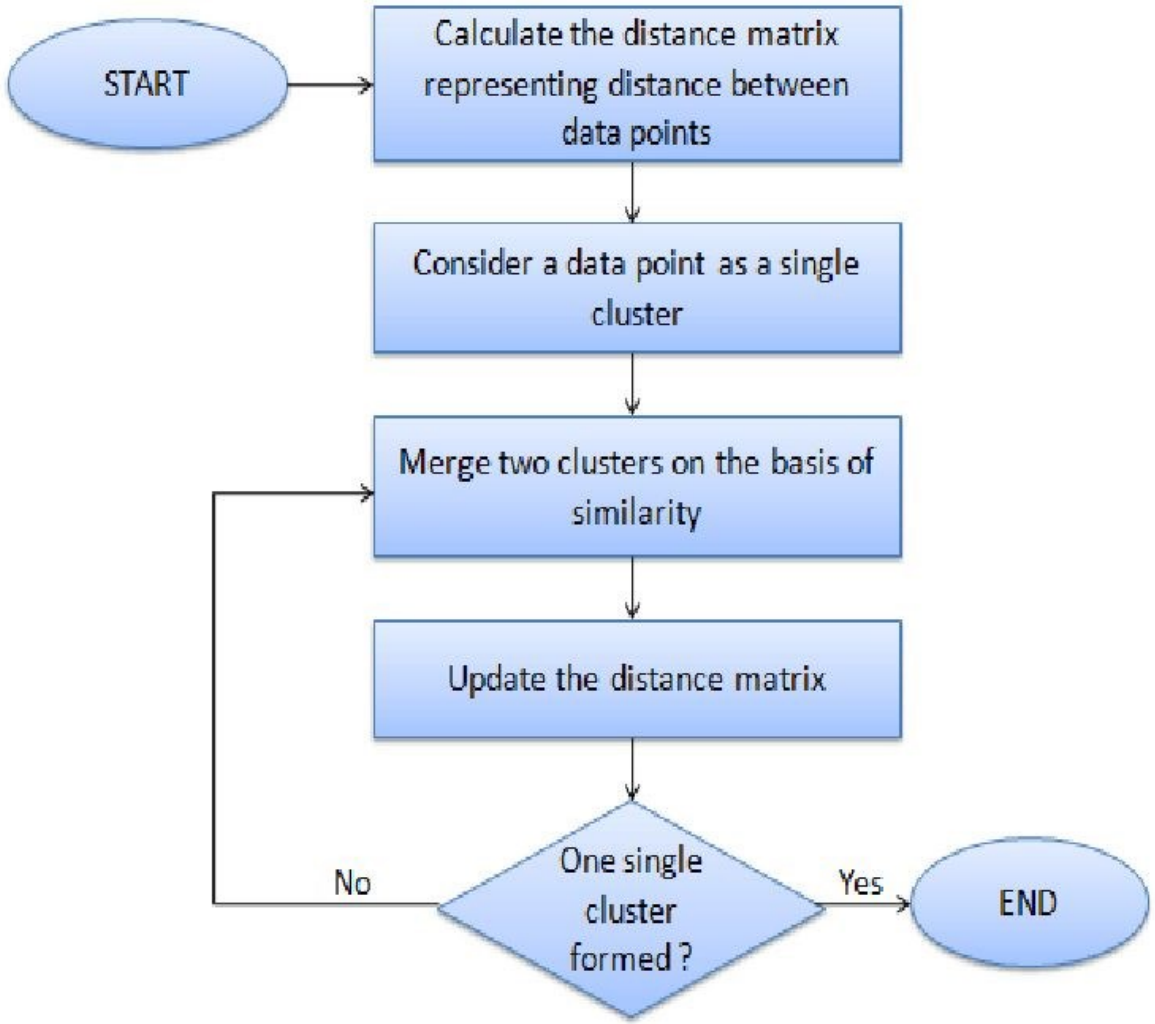
- Not as susceptible to initialization issues

K-Means++

1. **Choose one center uniformly** at random among the data points.
2. For each data point **x** not chosen yet, compute **D(x)**, the distance between x and the nearest center that has already been chosen.
3. Choose **one new data point at random** as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)_2$.
4. Repeat Steps 2 and 3 until k centers have been chosen.
5. Now that the initial centers have been chosen, proceed using **standard K-Means clustering**

$$\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$$

HAC: Algorithm

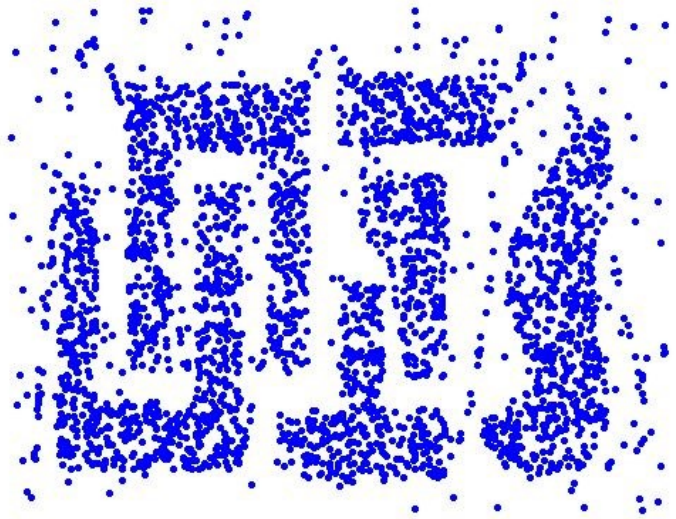


Closest Pair of Clusters

- Many variants to defining closest pair of clusters
- **Single-link**
  - Similarity of the closet elements
- **Complete-link**
  - Similarity of the “furthest” points
- **Average-link**
  - Average cosine between pairs of elements
- **Ward’s Method**
  - The increase in squared error when two clusters are merged

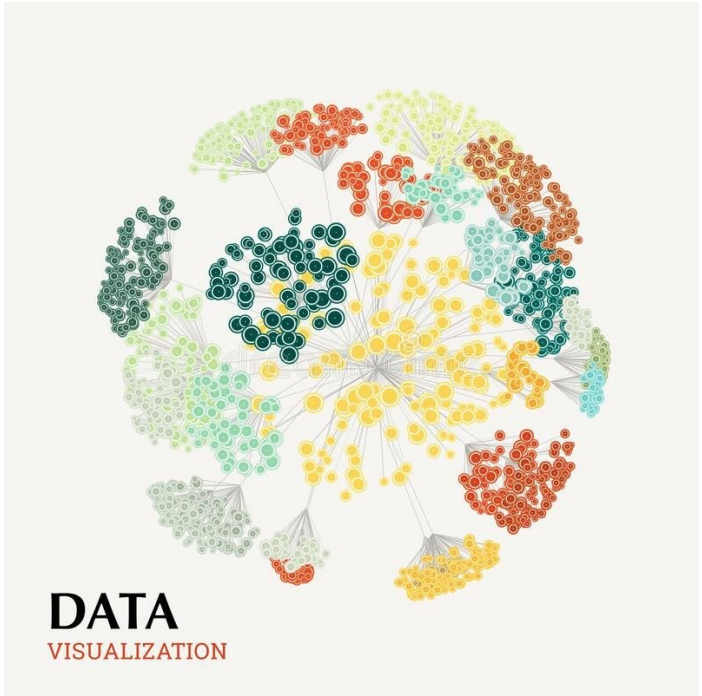
Density-based Clustering - DBSCAN

- **Main Idea**: Clusters are **regions of high density** that are **separated** from one another by **regions on low density**.
- **Density** = number of points within **a specified radius (Eps)**
  - Core point
  - Border point
  - Noise point



Summary

- **General Concepts of Clustering**
  - Definition
  - Real-life Applications
  - Types of Clustering
- **Typical Clustering Algorithms**
  - K-Means
  - HAC
  - DBSCAN



viettel

Thanks for your attention!