DAC-4: CLUSTERING

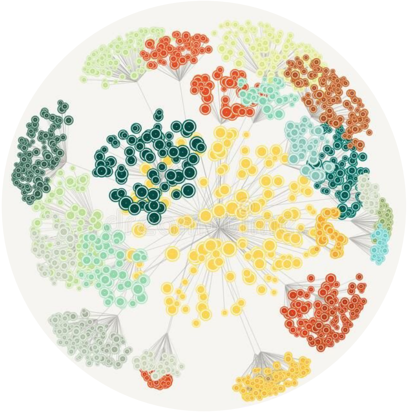
Duc-Trong Le (PhD), VNU-UET

November 2022

Outline

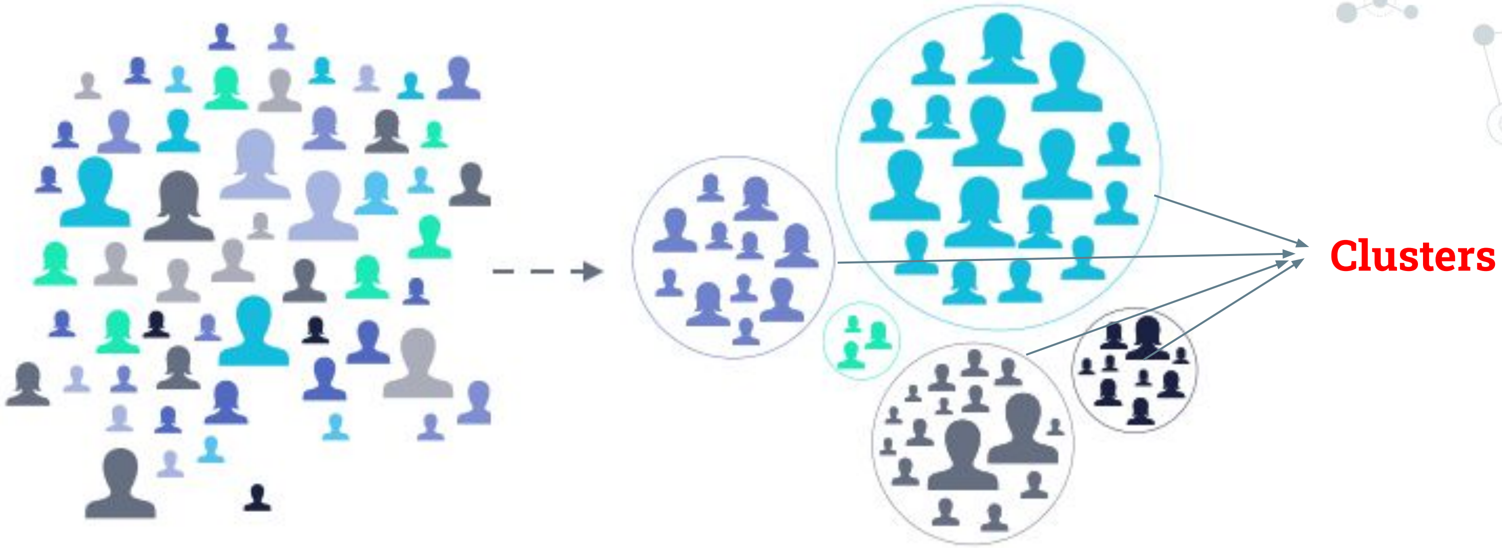
- Part I: Clustering - General Concepts
  - Real-life Applications
  - Types of Clusterings
- Part II: Typical Clustering Algorithms

1.
Clustering - General Concepts



Main idea, real-life applications, types

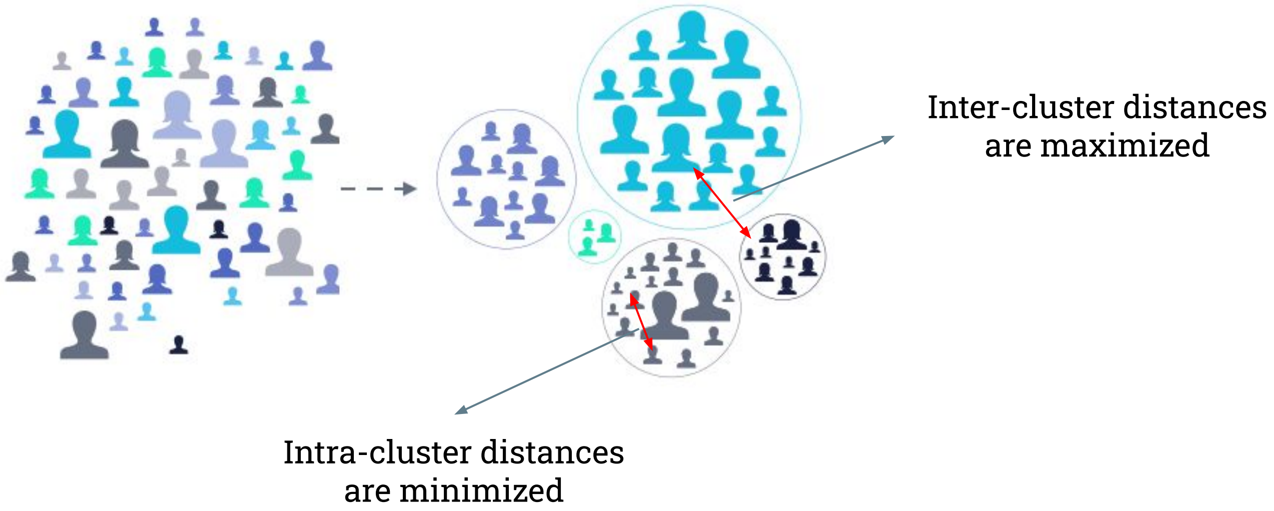
Motivating Example: Customer Segmentation



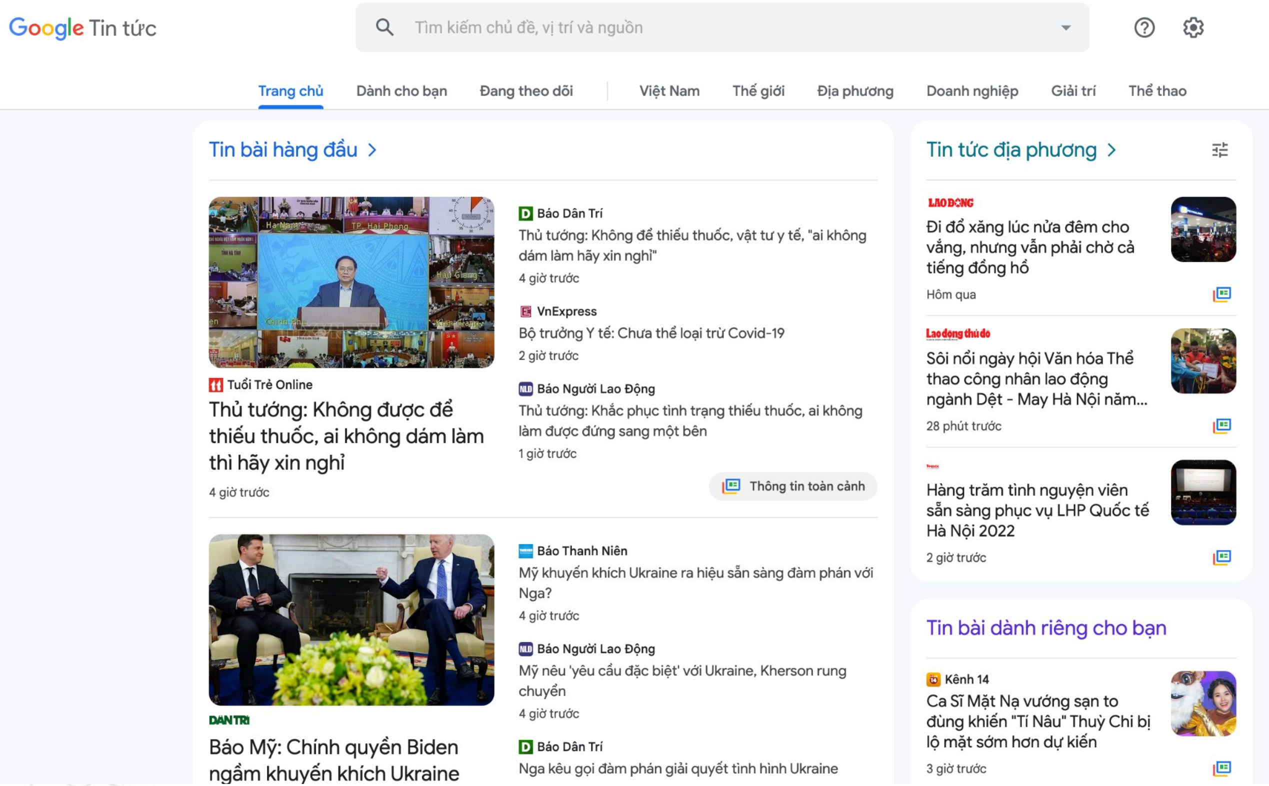
- Demographic
- Behavioral
- Geographic
- Psychographic

What is Cluster Analysis or Clustering?

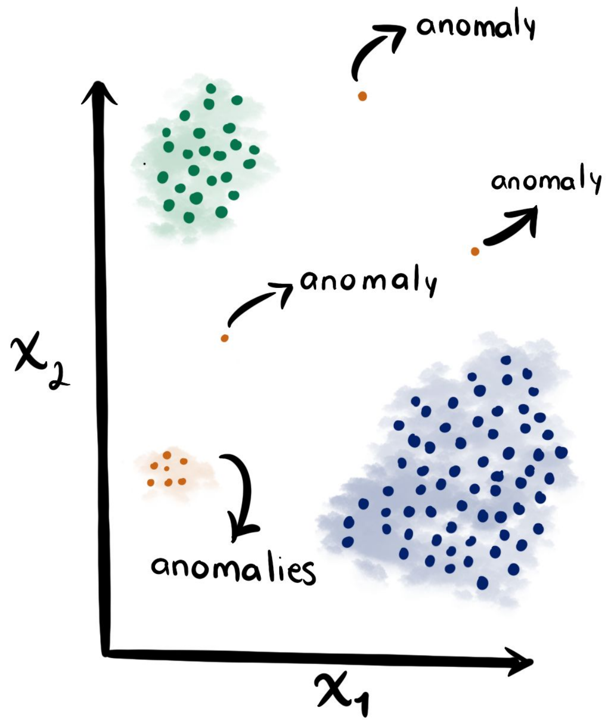
Given a set of objects, place them in **groups** such that the **objects in a group are similar** (or related) to **one another and different from (or unrelated to) the objects in other groups**



Real-life Applications: Google News



Real-life Applications: Anomaly Detection

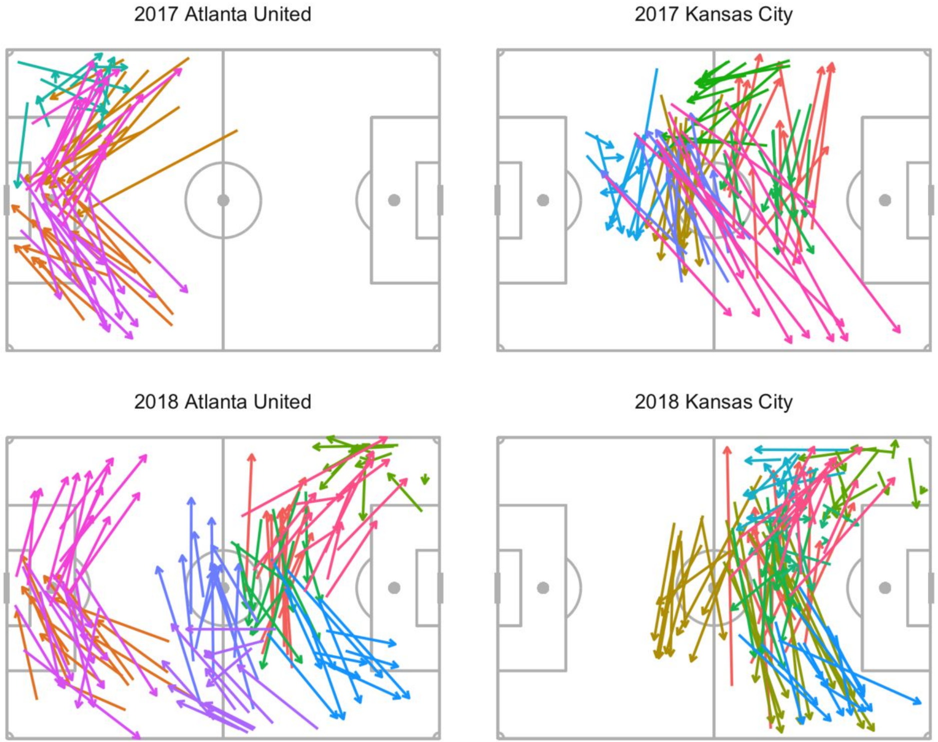


- Fake News Detection
- Fraud Detection
- Spam Email Detection

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

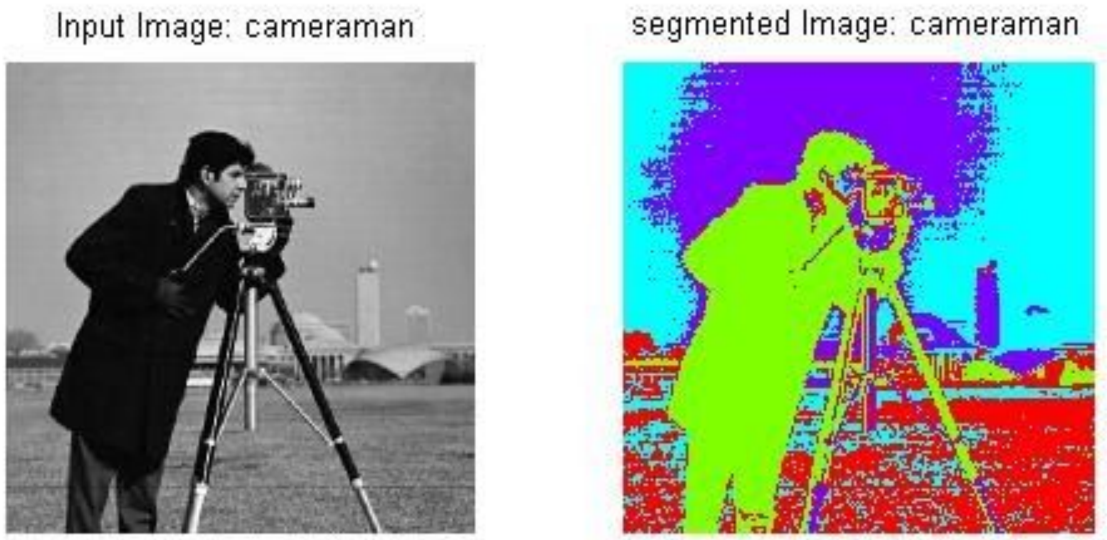
Real-life Applications: Sport Science

Find players with similar styles



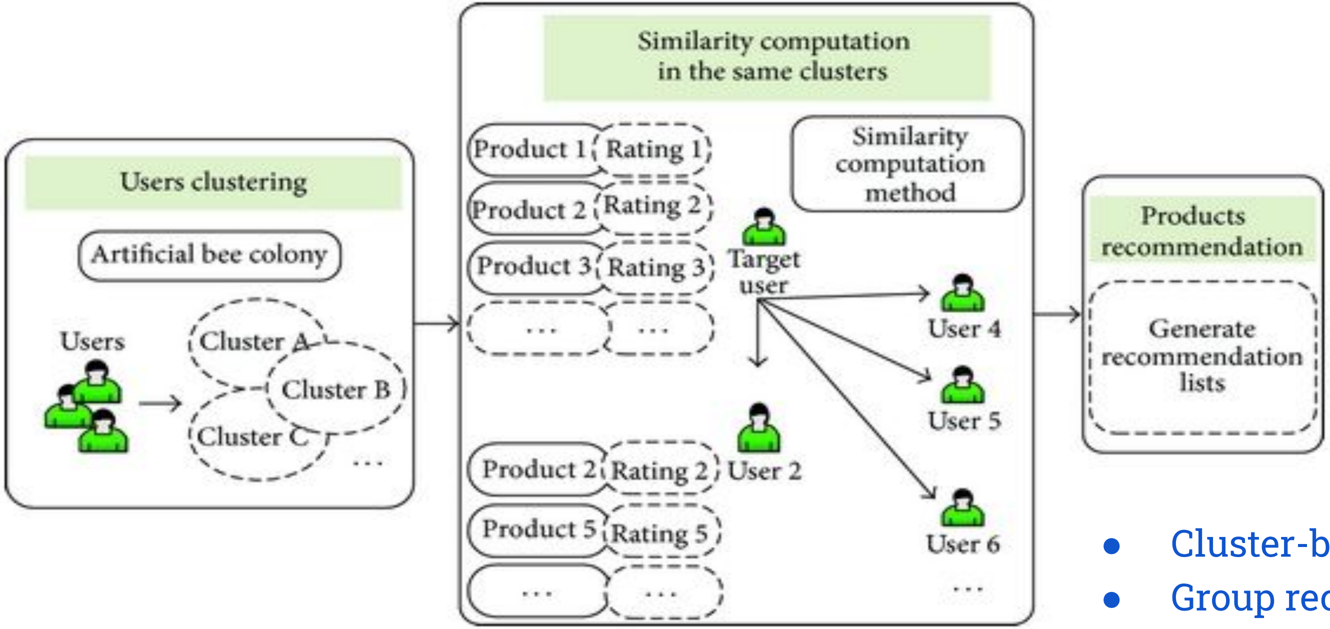
Source:
https://www.americansocceranalysis.com/home/2019/3/11/using-k-means-to-learn-what-soccer-passing-tells-us-about-playing-styles

Real-life Applications: Image Segmentation



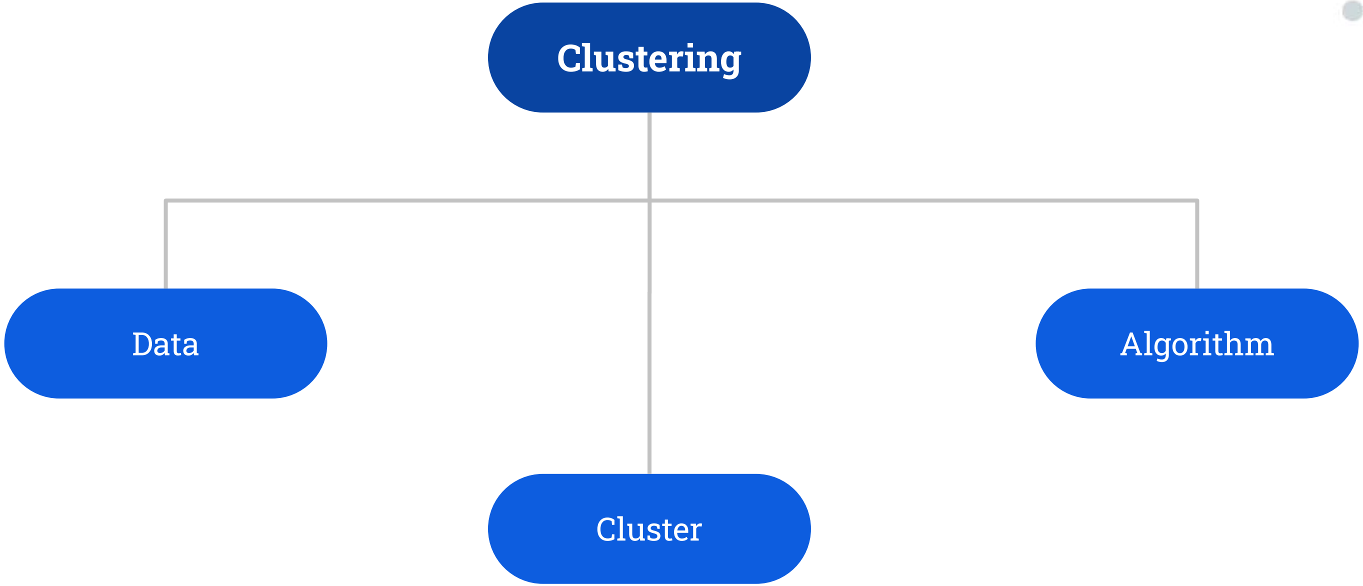
Source: http://pixelsciences.blogspot.com/2017/07/image-segmentation-k-means-clustering.html

Real-life Applications: Recommendation



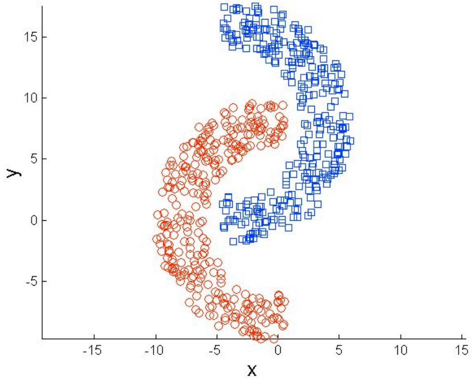
- Cluster-based ranking
- Group recommendation
- ...

What do affect on Cluster Analysis?



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**



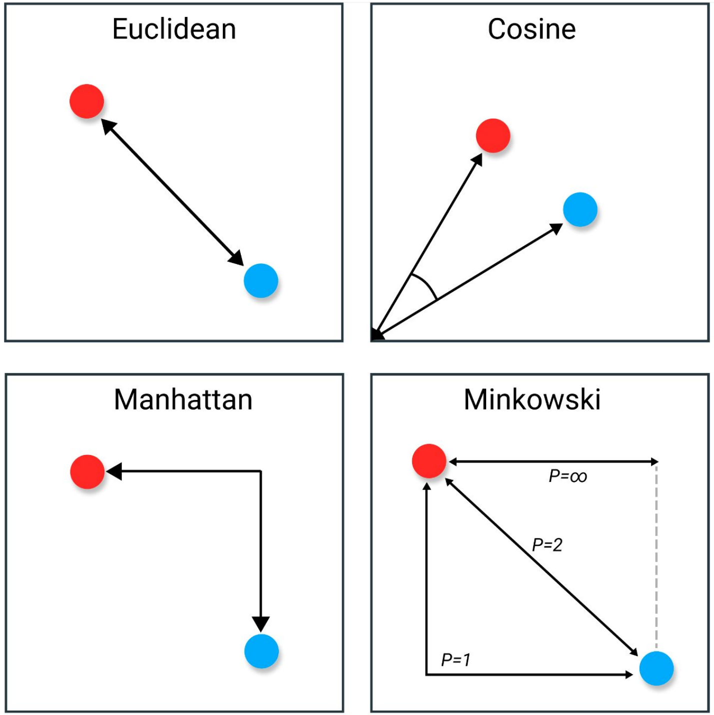
Characteristics of Cluster

- **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?

6 Clusters

2 Clusters

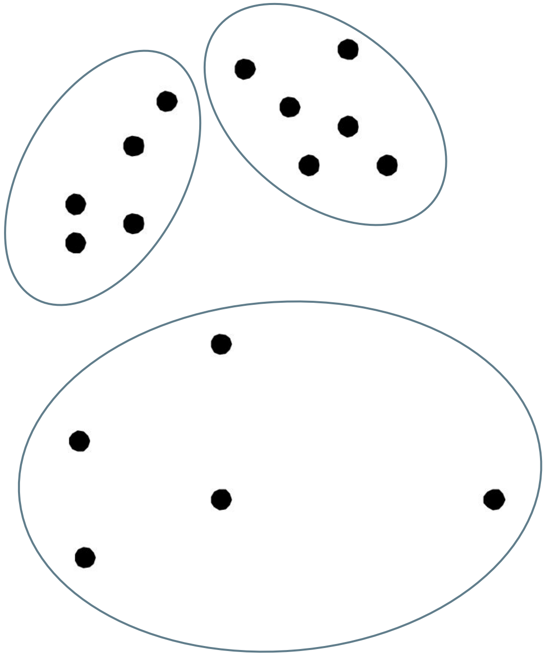
4 Clusters

Types of Clusterings

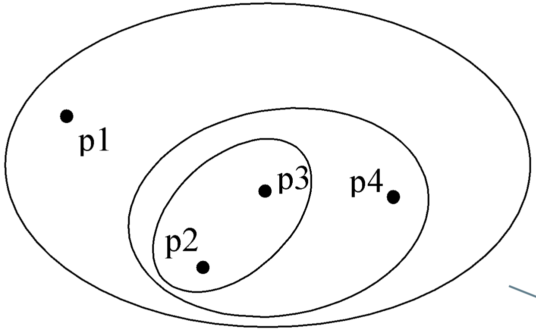
Source: https://www.datanovia.com/en/blog/types-of-clustering-methods-overview-and-quick-start-r-code/

Partitional Clustering

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

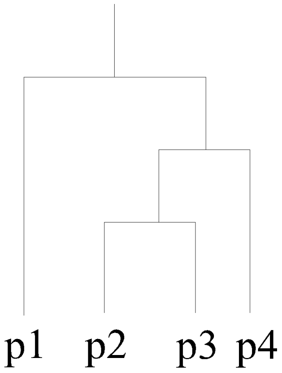


Hierarchical Clustering



Hierarchical Clustering

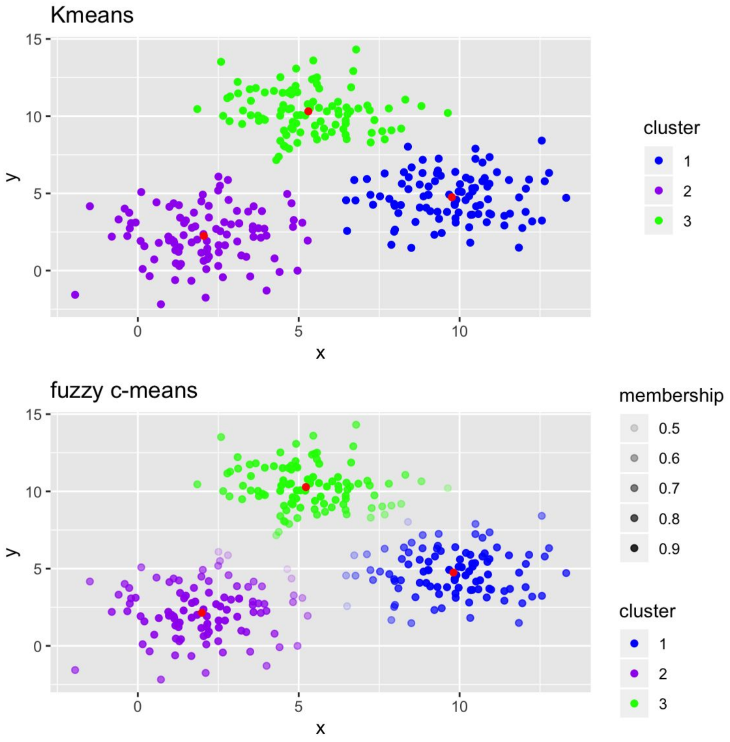
Data objects are separated into **nested clusters as a hierarchical tree**



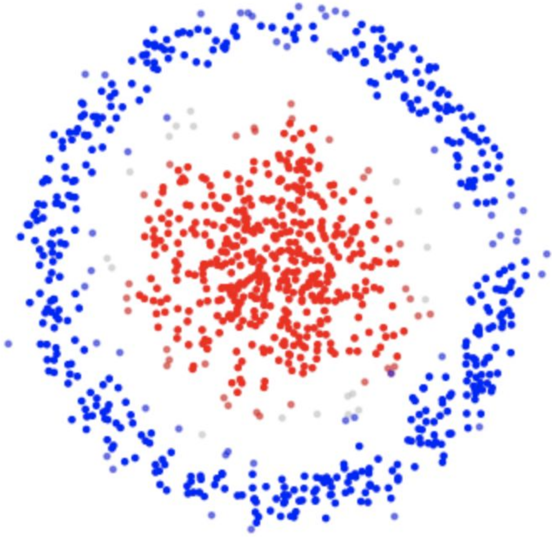
Clustering dendrogram

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**



Density-based Clustering

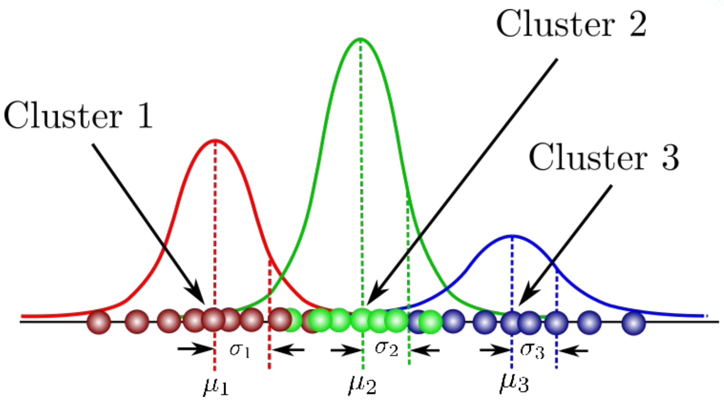


Non-linear separation

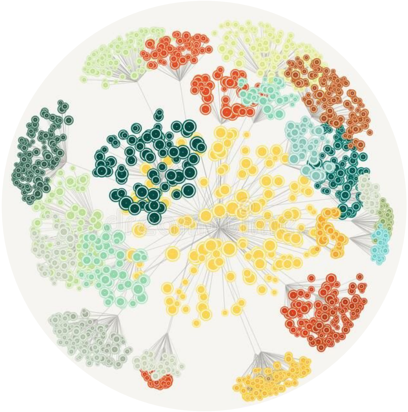
A cluster is **a dense region of points**, which is separated by **low-density regions**, from **other regions of high density**.

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



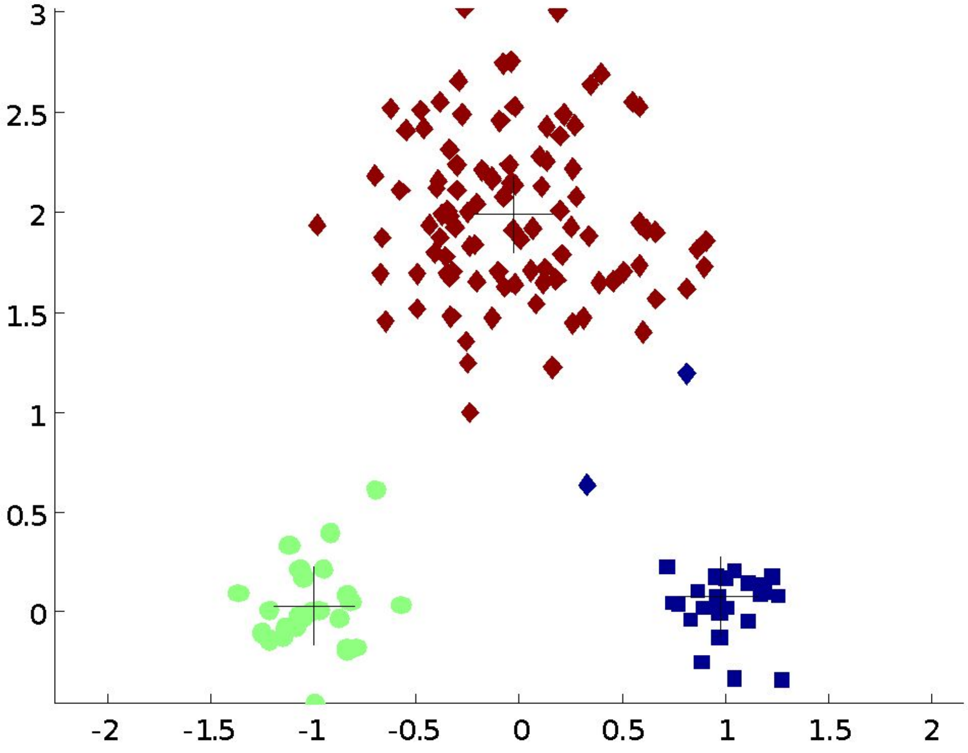
2.
Typical Clustering Algorithms

Intuition, Main Idea, Limitation

Typical Clustering Algorithms

- Partitional Clustering
  - K-Means & Variants
- Hierarchical Clustering
  - HAC
- Density-based Clustering
  - DBSCAN

K-Means Clustering: An Example



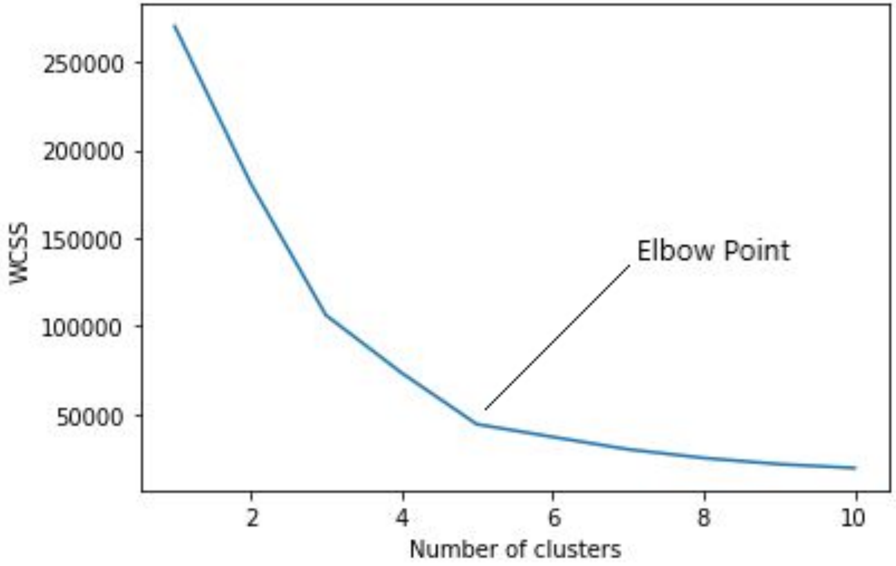
K-Means Clustering

Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change

- **Main idea**: Each point is assigned to the cluster **with the closest centroid**
- Number of clusters, **K, must be specified**
- Sum of Squared Error (SSE)
- Complexity: O(n * K * I * d)
    - n = number of points, K = number of clusters,
    - I = number of iterations, d = number of attributes

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

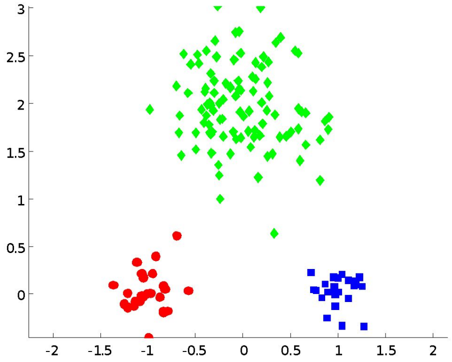
Elbow Method for Optimal Value of K

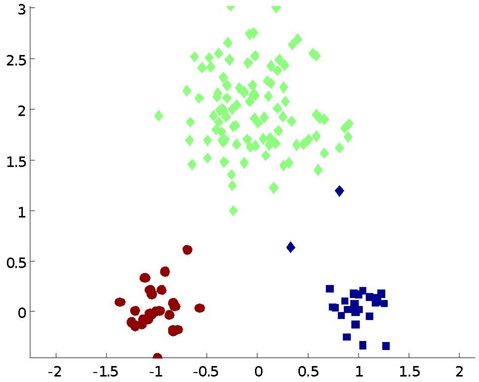


WCSS is the sum of squared distance between each point and the centroid in a cluster

The graph will rapidly change at a point named **Elbow Point**

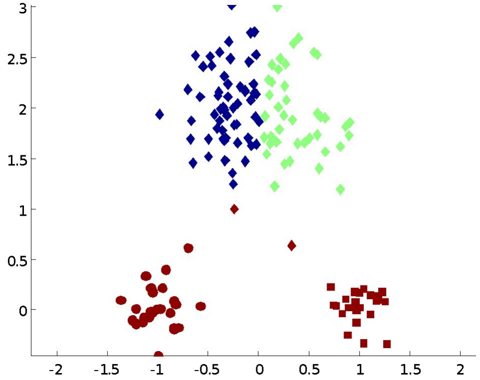
Two different K-means Clusterings





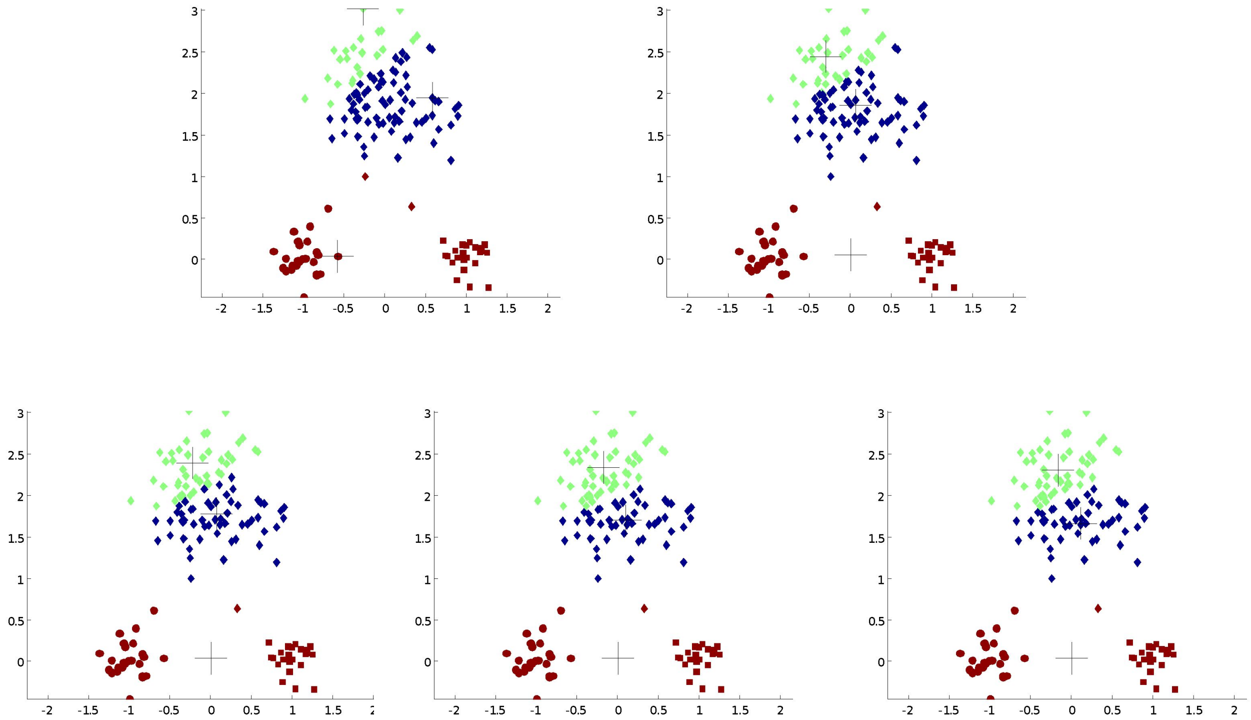
Optimal Clustering

Original Points



Sub-optimal Clustering

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**

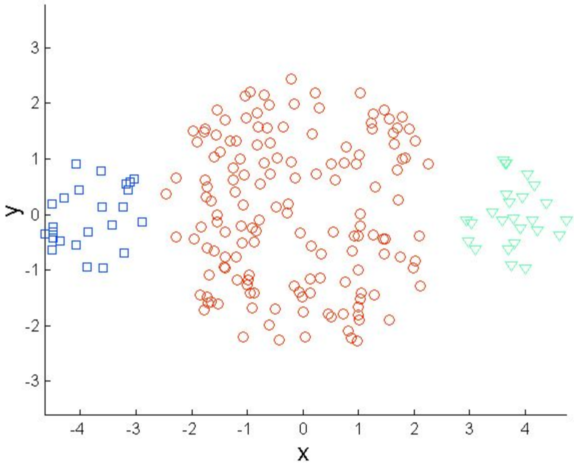
$$\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$$

Bisecting K-Means

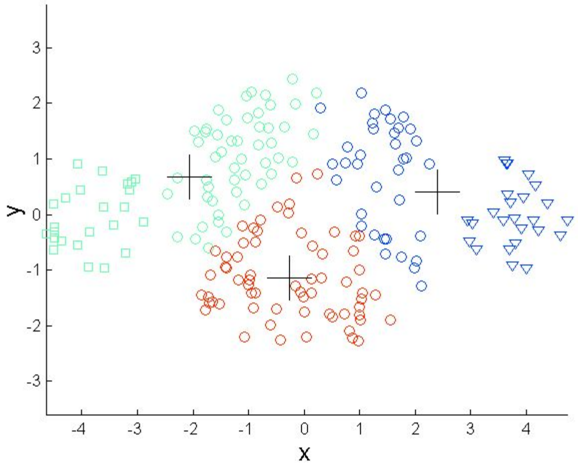
Initialize the list of clusters to contain the cluster containing all points.
repeat
    Select a cluster from the list of clusters
    for $i = 1$ to *number_of_iterations* **do**
        Bisect the selected cluster using basic K-means
    end for
    Add the two clusters from the bisection with the lowest SSE to the list of clusters.
until Until the list of clusters contains K clusters

It is a variant of K-means that can produce a partitional or a hierarchical clustering

Limitations of K-means: Differing Sizes

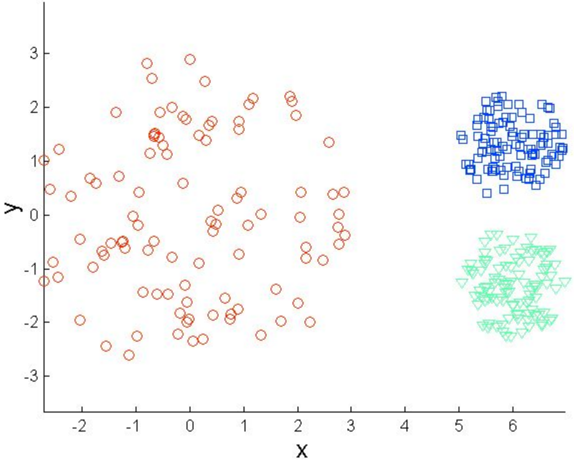


Original Points

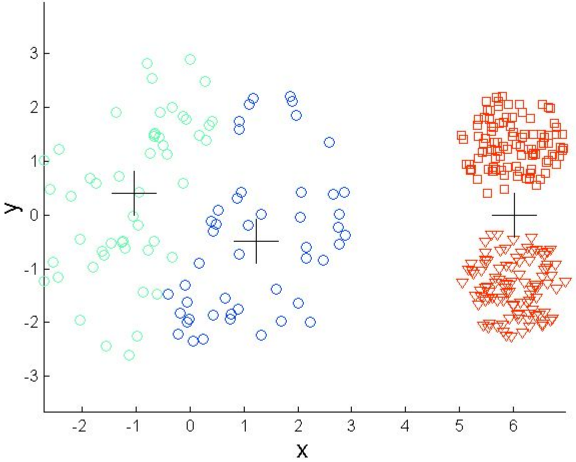


K-means (3 Clusters)

Limitations of K-means: Differing Density

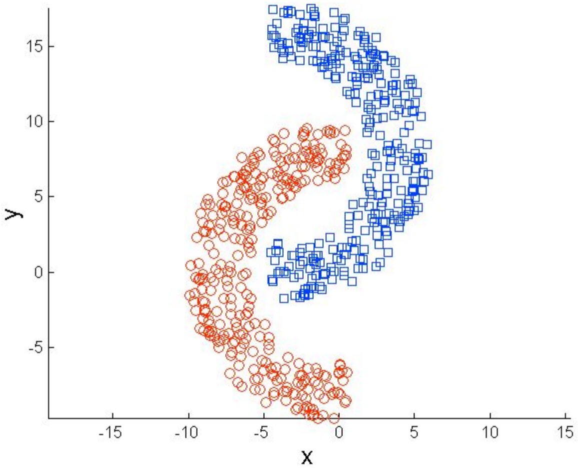


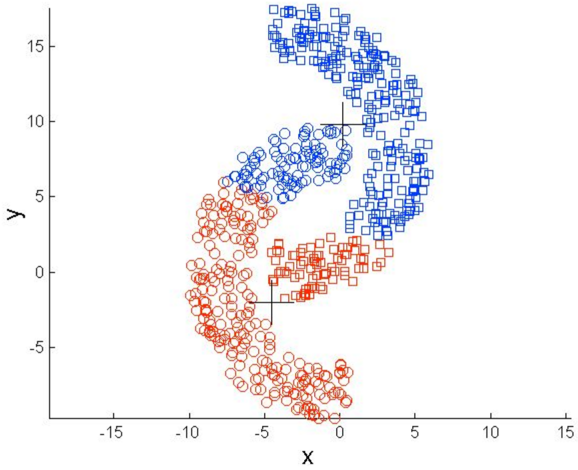
Original Points



K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

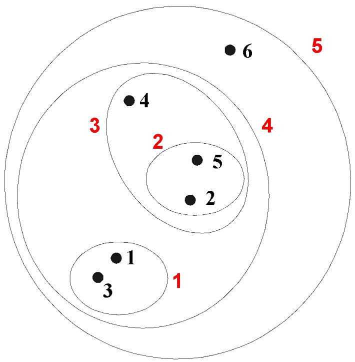


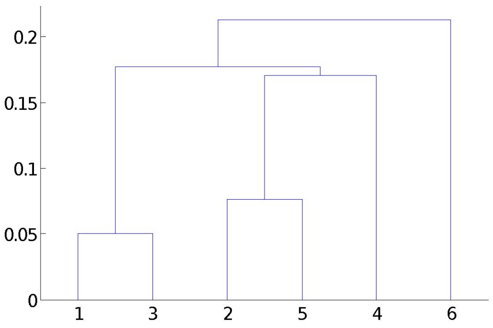


Original Points

K-means (2 Clusters)

Hierarchical Agglomerative Clustering

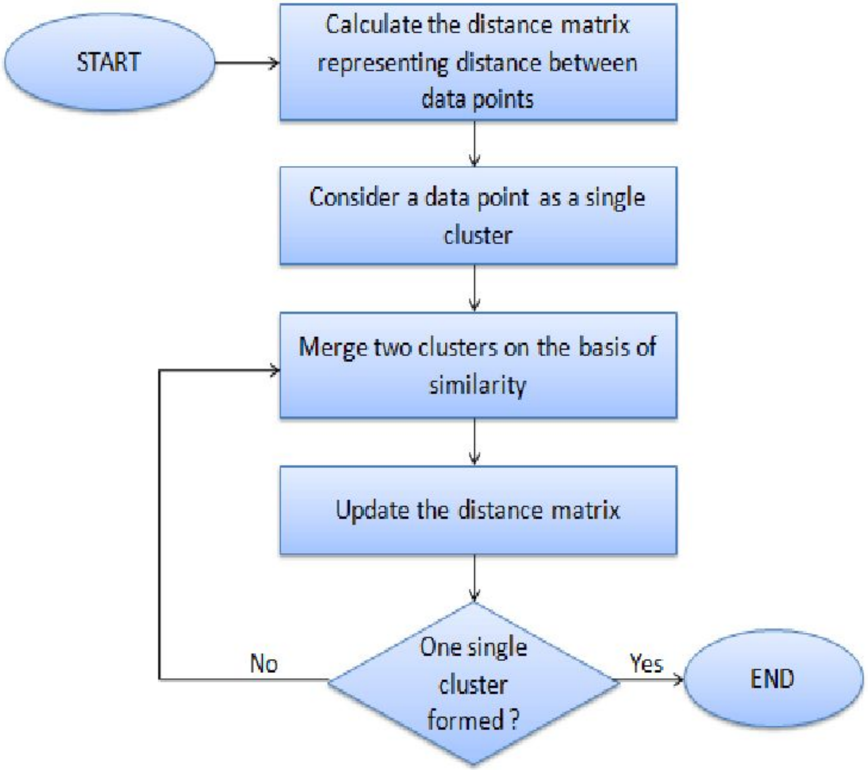




dendrogram

- **Main Idea**:
  - Start with the points as **individual clusters**
  - At each step, merge the **closest pair** of clusters until **only one cluster** (or K clusters) left
- Key operation is the computation of the proximity of two clusters
  - Worst-case Complexity: $O(N^3)$

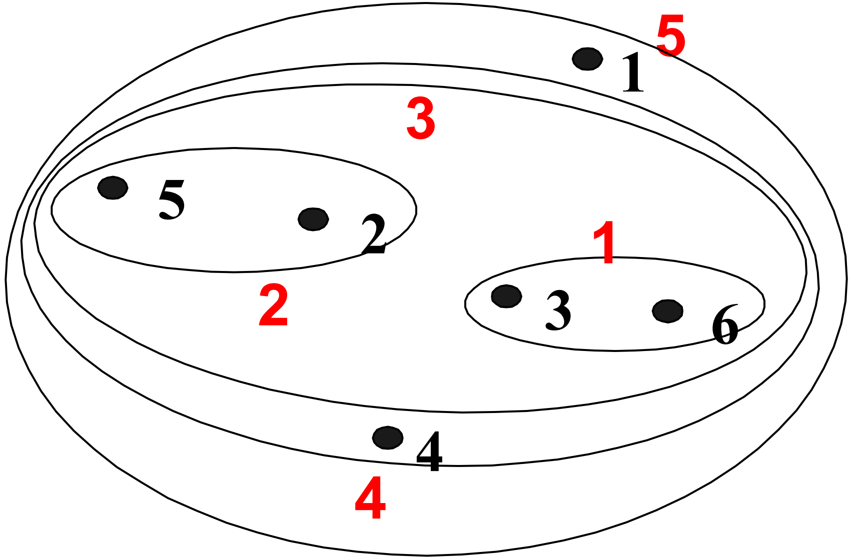
HAC: Algorithm



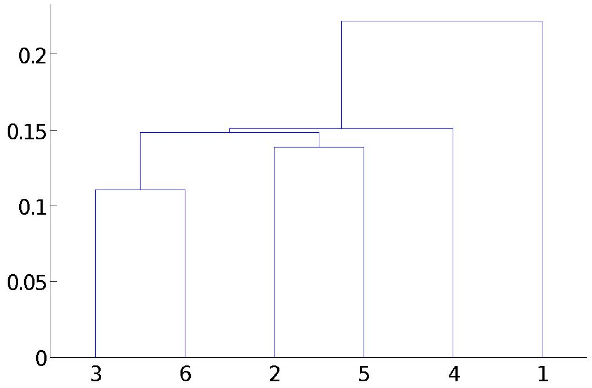
Closest Pair of Clusters

- 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

HAC - Single-link (MIN)

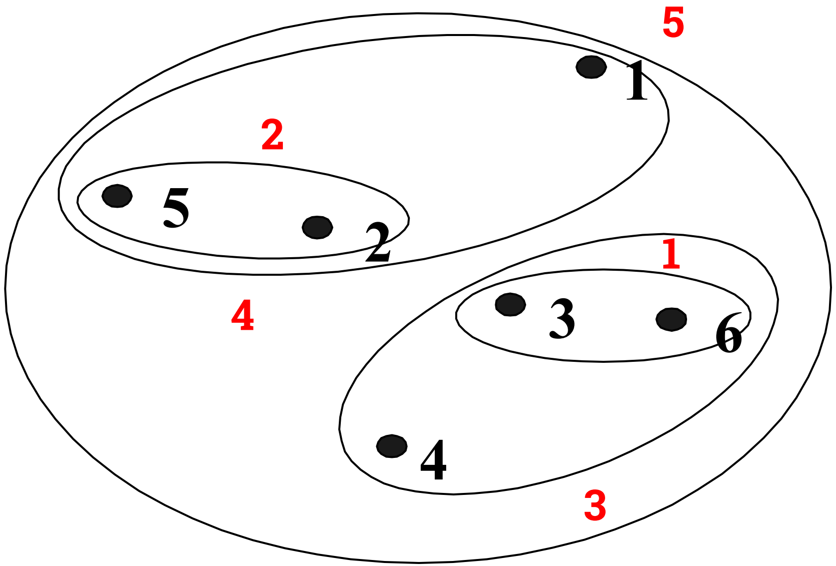


Nested Clusters

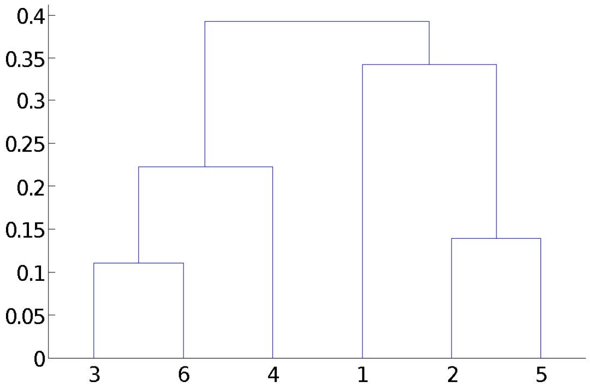


Dendrogram

HAC - Complete-link (MAX)

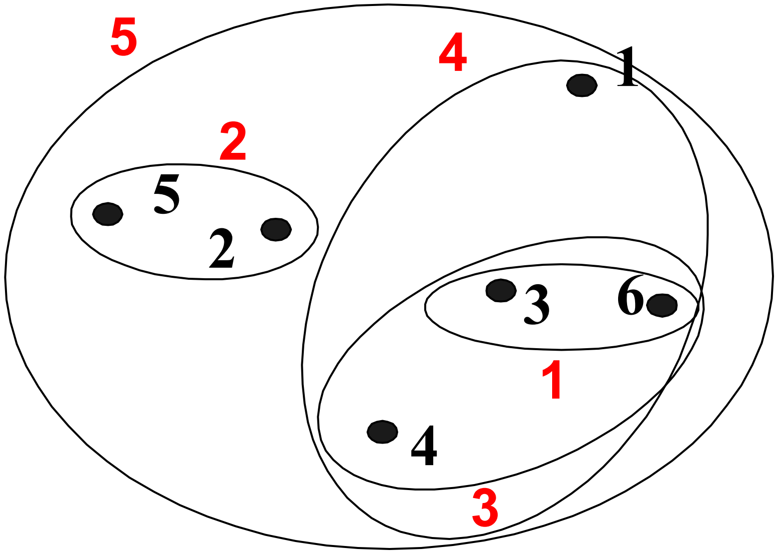


Nested Clusters

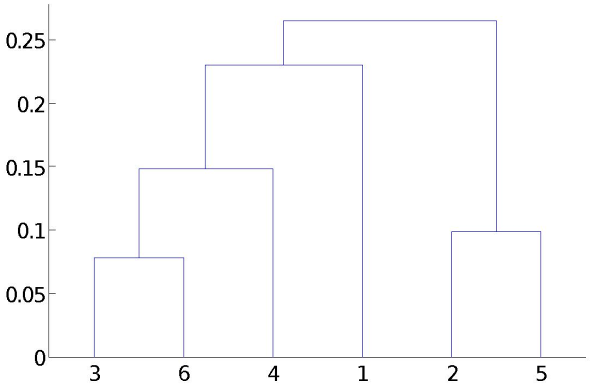


Dendrogram

HAC - Average-link



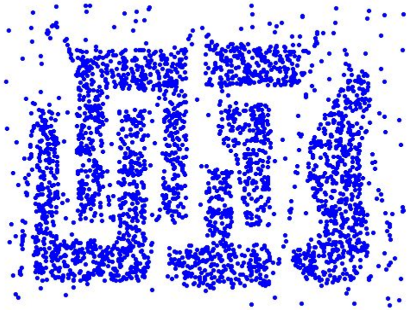
Nested Clusters



Dendrogram

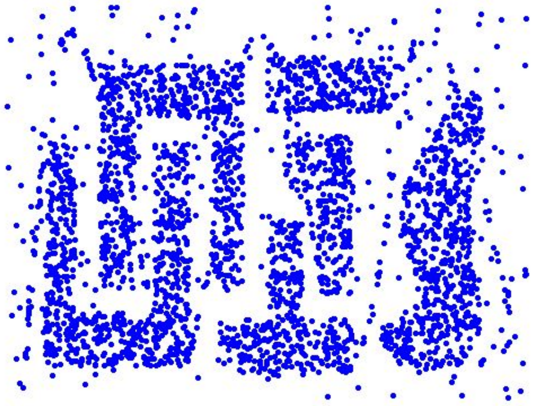
HAC: Limitations

- Once two clusters are combined, it **cannot be undone**
- No **global objective function** is directly minimized
- Typical Problems:
  - Sensitivity to noise
  - Difficulty handling clusters of different sizes and non-globular shapes
  - Breaking large clusters



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

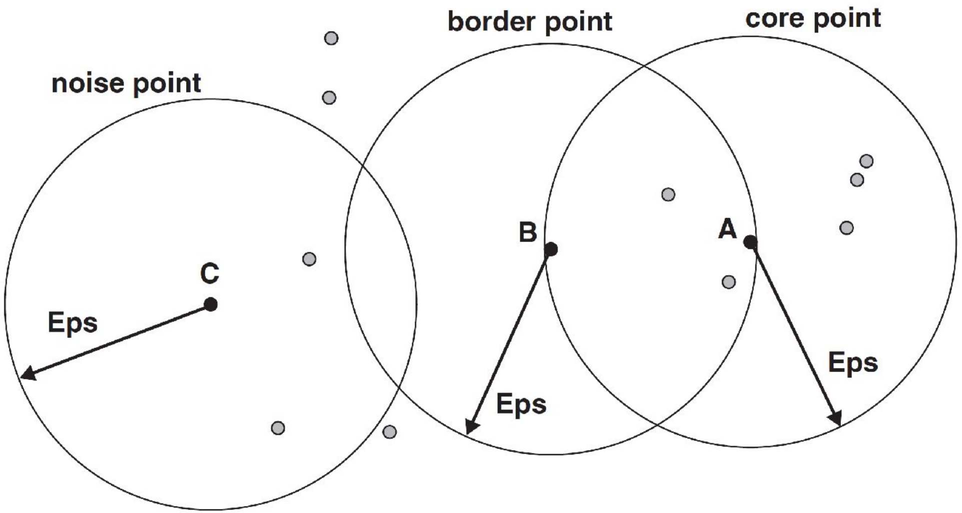


DBSCAN: Algorithm

DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters with its associated core points.

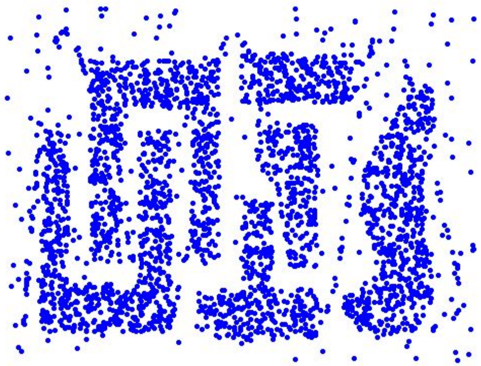
How to Determine Points?



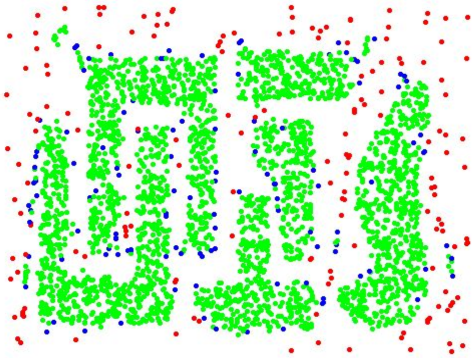
MinPts = 7

- **Core point:** Has at least a specified number of points (MinPts) within Eps
- **Border point**: not a core point, but is in the neighborhood of a core point
- **Noise point**: any point that is not a core point or a border point

DBSCAN: Core, Border and Noise Points



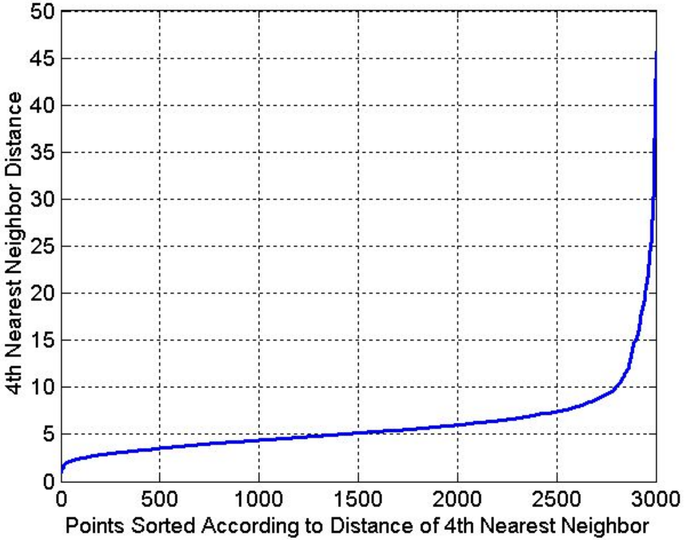
Original Points



Point types: core, border and noise

Eps = 10, MinPts = 4

DBSCAN: How to Determine Eps, MinPts?

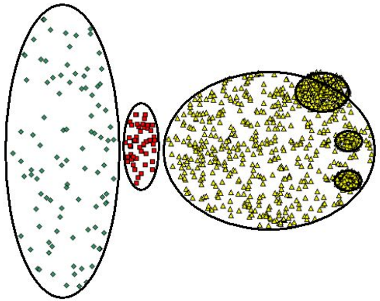


Intuition:

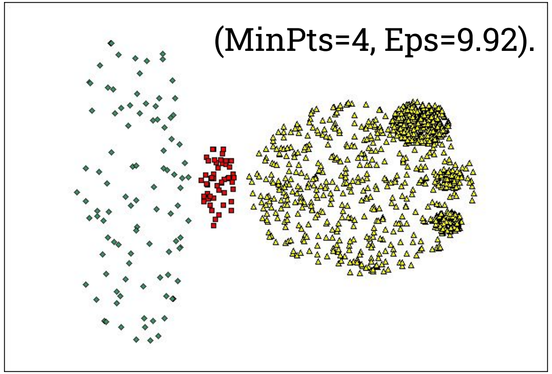
- Core point: the k-th nearest neighbors are at a close distance.
- Noise point: the k-*th* nearest neighbors are at a far distance.

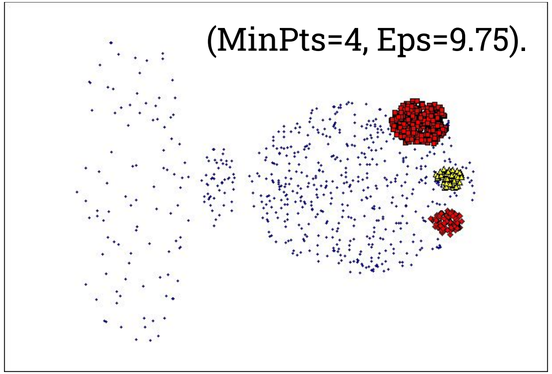
Plot sorted distance of every point to its k-th nearest neighbor

DBSCAN: Limitations



Original Points

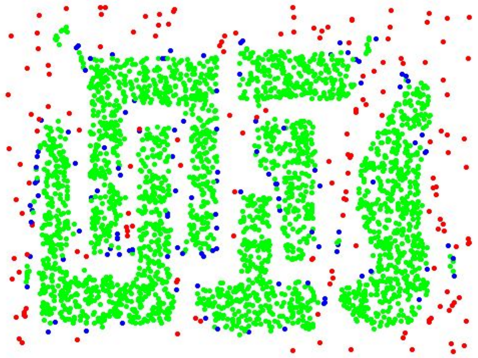




- **Varying densities**
- **High-dimensional data**

Which Clustering Algorithm?

- Type of Clustering
- Type of Cluster
  - Prototype vs connected regions vs density-based
- Characteristics of Clusters
- Characteristics of Data Sets and Attributes
- Noise and Outliers
- Number of Data Objects
- Number of Attributes
- Algorithmic Considerations



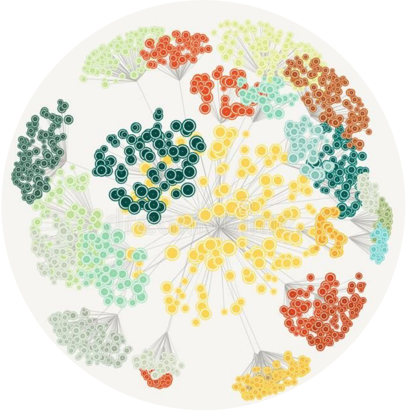
A Comparison on Clustering Algorithms

Criteria	Hierarchal clustering	K-mean	K-mediod	DBSCAN
Initial condition	No	Yes	Yes	Yes
Termination condition	Not precise	Precise	Precise	Precise
Arbitrary value	No requirement	Numeric attribute	Numeric attribute	Numeric attribute
Effect on Size of data sets	Not good	Good	Not good	Not good
Shape of data set	Arbitrary	Convex	Convex	Arbitrary
Granularity	Flexible	K and initial point	K and initial point	Threshold
Result optimization	Optimization	Rebuild optimization	Rebuild optimization	Rebuild optimization
Handling dynamic data	No	Yes	Yes	Yes
Behavior on noisy data	No influences	Influences	Influences	Not much influences
Distance measurement	Any	Distance at normal space	Distance at normal space	Density
Implementation	Simple	Simple	Complicated	Simple

Source: Text Clustering Algorithms: A Review

Summary

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



Thanks!

trongld@vnu.edu.vn