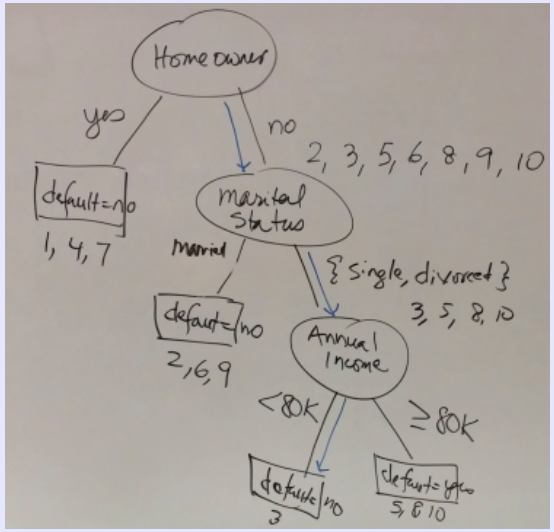
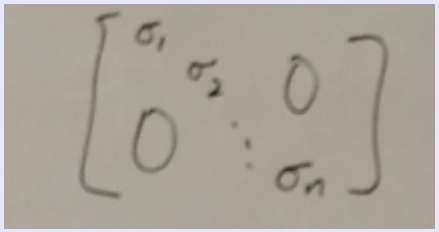
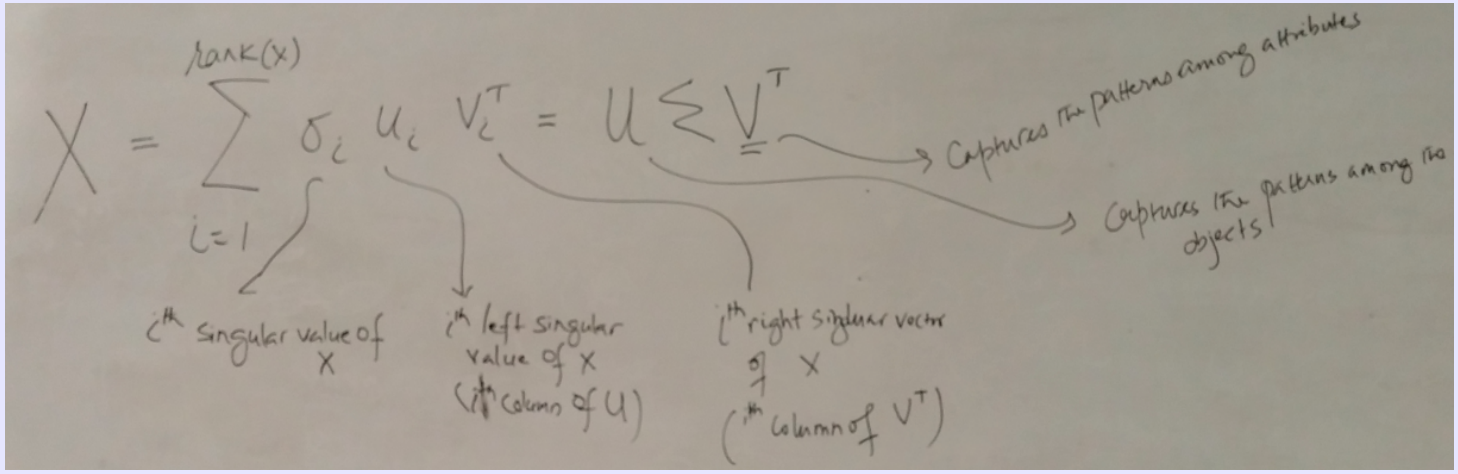
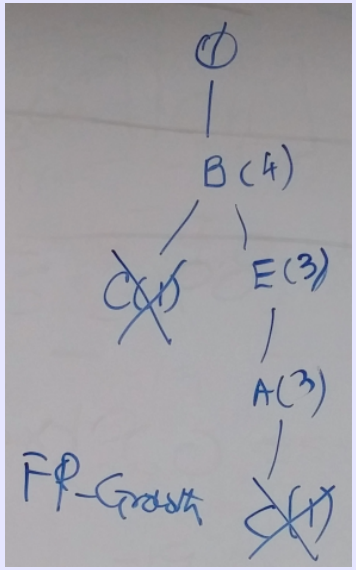








CS 422: Data Mining
Vijay K. Gurbani, Ph.D.,
Illinois Institute of Technology

Clustering I

Clustering: Introduction

- Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both.
  - Objects in a cluster share the same characteristics.
- Used in a variety of fields: health and medicine, business, …
  - Health and medicine: Cluster patients according to symptoms presented upon evaluation.
  - Business: Cluster stores according to sales / customer satisfaction / …
  - Computer networking: Cluster traffic according to application type
    - Encrypted traffic identification.
    - Anomaly detection.
    - ...

Clustering: Introduction

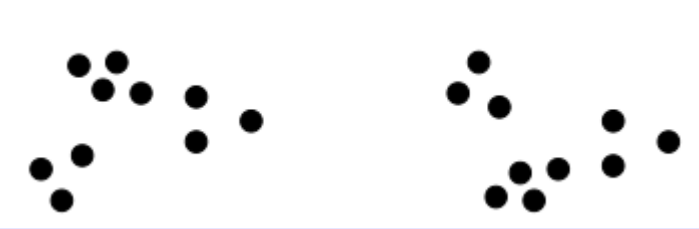
- Cluster analysis divides data into groups (clusters) that are meaningful, useful, or both.
  - Meaningful: clusters should capture the natural structure of the data.
  - Useful: *cluster prototype*.
    - Using the cluster prototype, clusters can be used as a starting point for data summarization, compression (vector quantization), classification (nearest neighbour).

Clustering: Introduction

- Clustering groups data objects based on information found in the data that describes the objects and their relationships.
  - Goal: Objects within a cluster be similar to one another, but different from objects in other clusters.
- Notion of a cluster is not well defined.

Clustering: Introduction

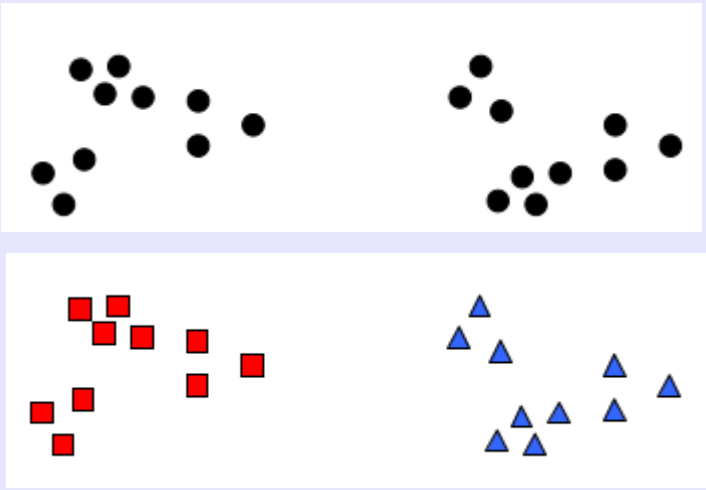
- Notion of a cluster is not well defined.
- Consider: Original data points.



How many clusters?

Clustering: Introduction

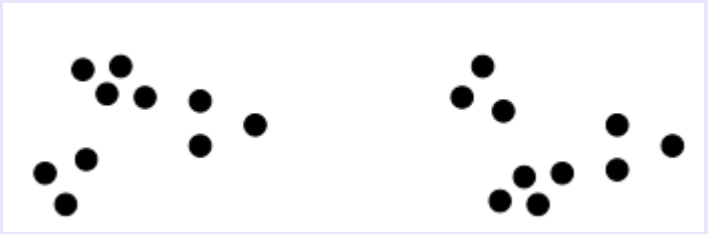
- Notion of a cluster is not well defined.
- Consider: Original data points.



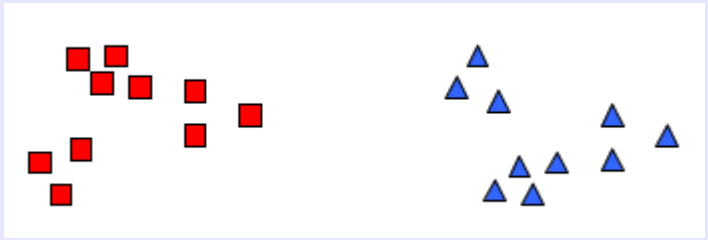
How many clusters?

Clustering: Introduction

- Notion of a cluster is not well defined.
- Consider: Original data points.

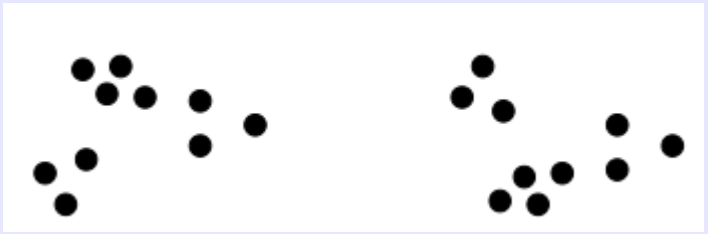


How many clusters?

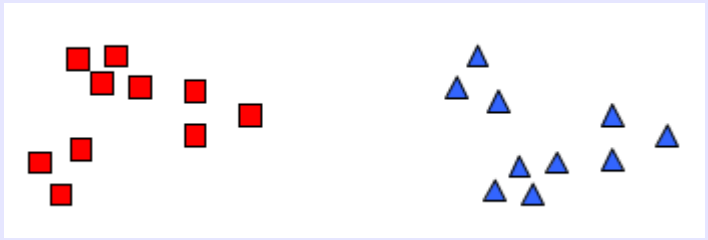


Clustering: Introduction

- Notion of a cluster is not well defined.
- Consider: Original data points.

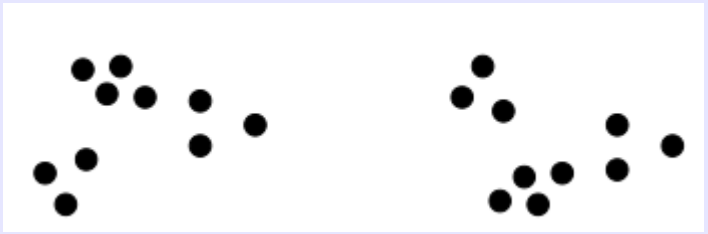


How many clusters?

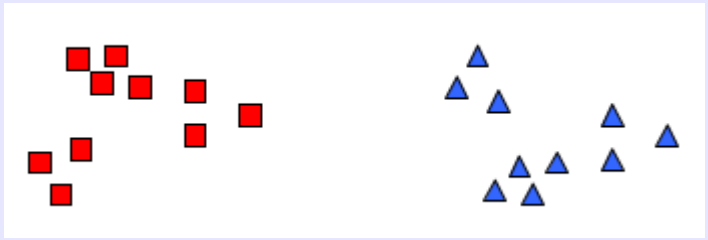


Clustering: Introduction

- Notion of a cluster is not well defined.
- Consider: Original data points.



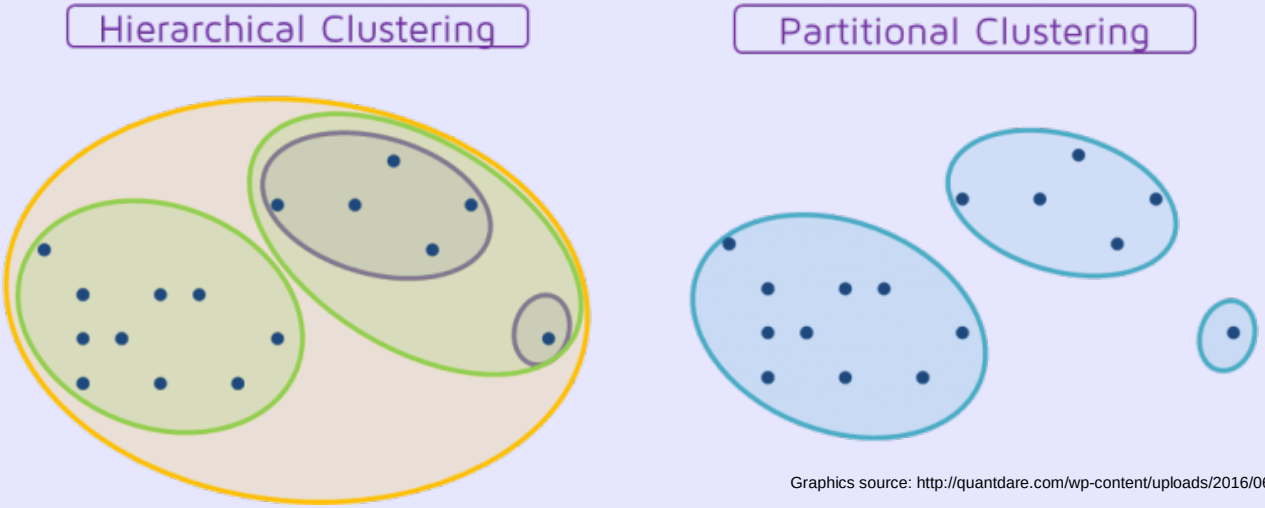
How many clusters?



Where do we draw cluster boundaries?

Clustering: Introduction

- Different type of clustering
  - Partitional vs. hierarchical:
    - Partitional: divide objects into non-overlapping clusters such that each object belongs to exactly one cluster.
    - Hierarchical: Clusters can have subclusters.



Graphics source: http://quantdare.com/wp-content/uploads/2016/06/HierarPartClustering-800x306.png

Clustering: Introduction

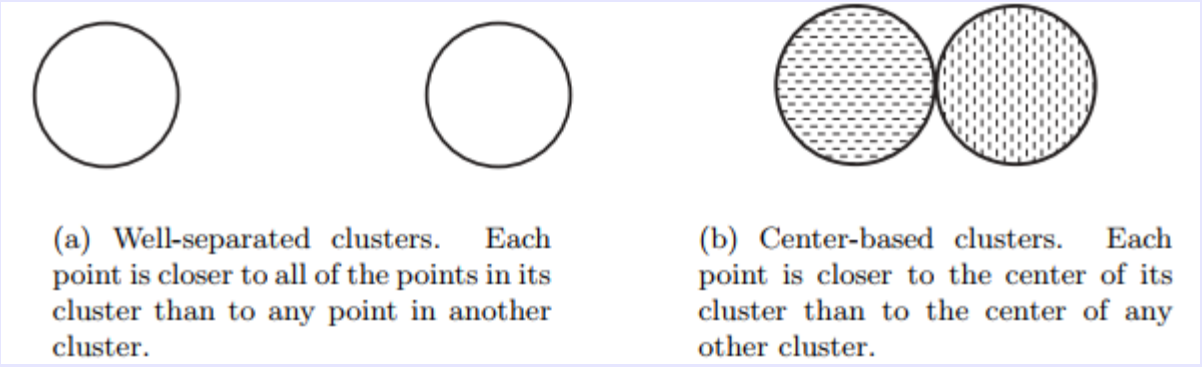
- Different type of clustering
  - Exclusive vs. overlapping vs. fuzzy:
    - Exclusive: 1:1 relationship between object and cluster.
    - Overlapping: 1:n relationship between object and cluster; an object can belong to > 1 cluster.
    - Fuzzy: n:n relationship, all objects belong to all clusters with a certain probability (or *membership weight*). Each object's probability of belonging to all clusters should sum up to 1.0.

Clustering: Introduction

- Different type of clustering
  - Complete vs. partial:
    - Complete: Assign every point to at least one cluster.
    - Partial: Some objects may not be assigned to any cluster. Such objects may represent “noise”, or “outliers”.
      - Outlier detection, or anomaly detection, is a rich area of current research.

Clustering: Introduction

- Different type of clusters

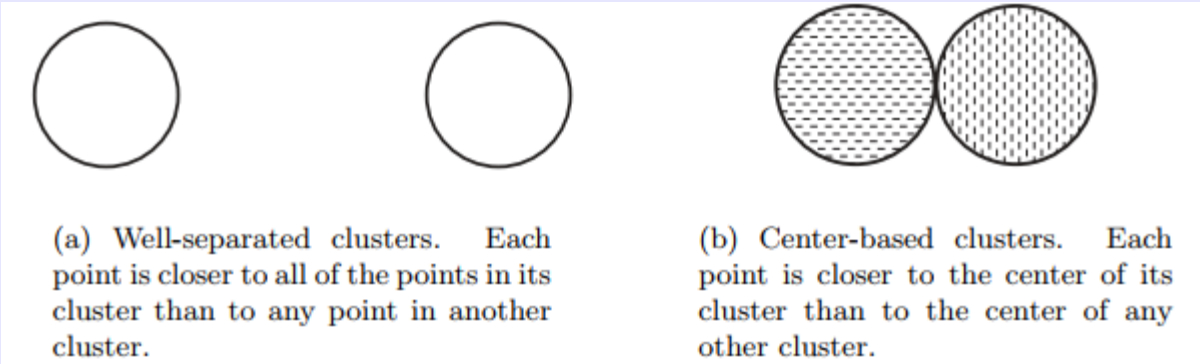


(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

Clustering: Introduction

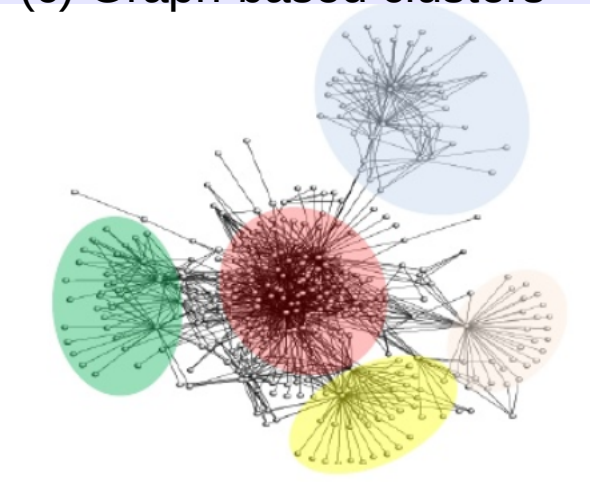
- Different type of clusters



(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

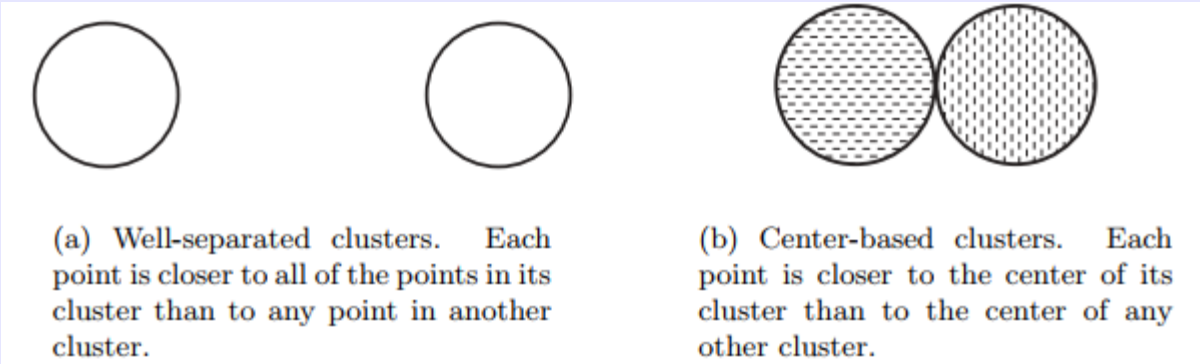
(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

(c) Graph-based clusters



Clustering: Introduction

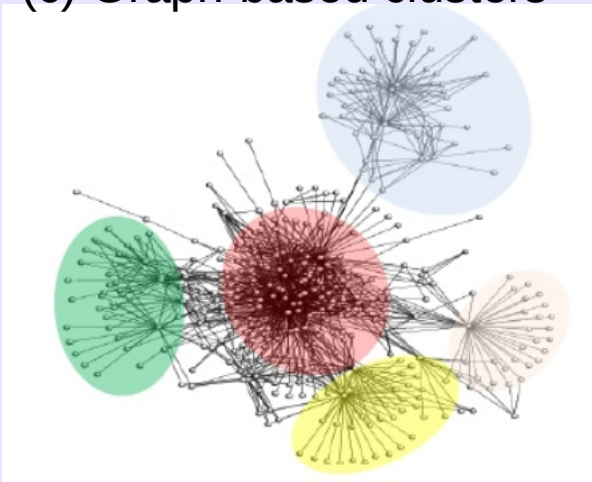
- Different type of clusters



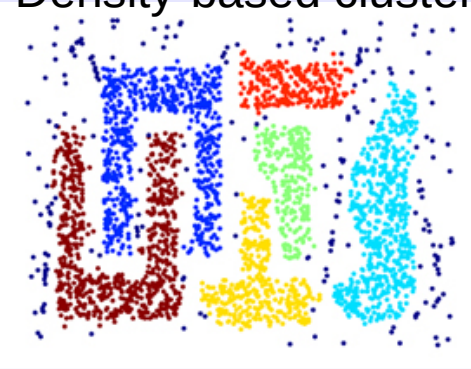
(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.

(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.

(c) Graph-based clusters



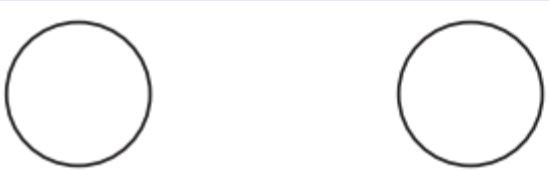
(d) Density-based clusters



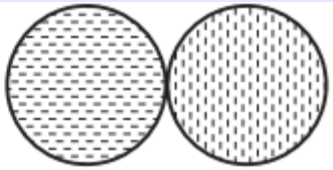
Used when clusters are irregular or intertwined, or when noise and outliers are present.

Clustering: Introduction

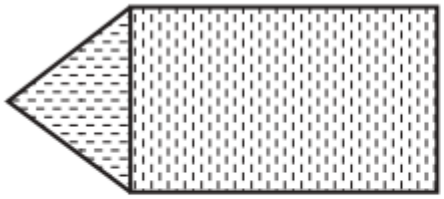
- Different type of clusters

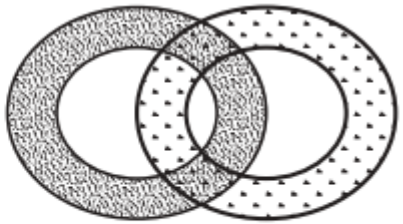


(a) Well-separated clusters. Each point is closer to all of the points in its cluster than to any point in another cluster.



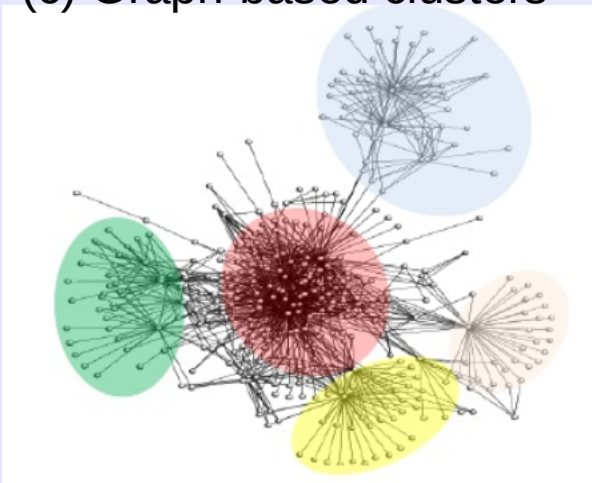
(b) Center-based clusters. Each point is closer to the center of its cluster than to the center of any other cluster.



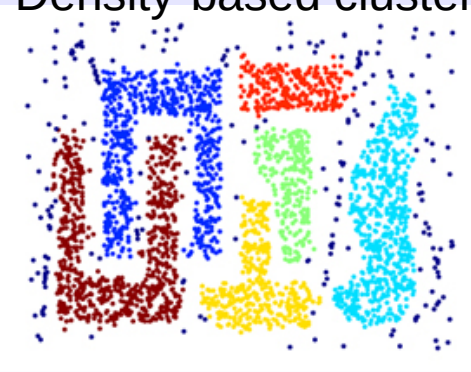


(e) Conceptual clusters. Points in a cluster share some general property that derives from the entire set of points. (Points in the intersection of the circles belong to both.)

(c) Graph-based clusters



(d) Density-based clusters

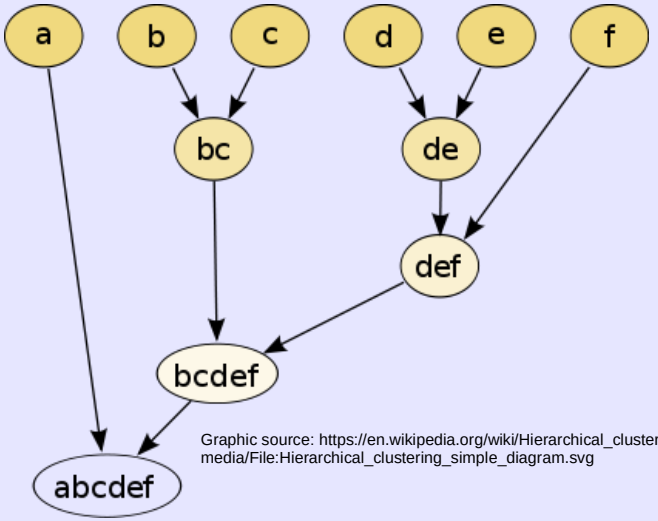


Used when clusters are irregular or intertwined, or when noise and outliers are present.

Clustering: Algorithms

- What we will study:
  - K-means clustering: a prototype-based, partitional clustering techniques to find patterns in the data to create *k* clusters.
  - Hierarchical clustering: Build a hierarchy of clusters starting from singleton clusters.
  - DBSCAN: Density-based clustering.

Graphic source: http://yaikhom.com/res/dbscan.png

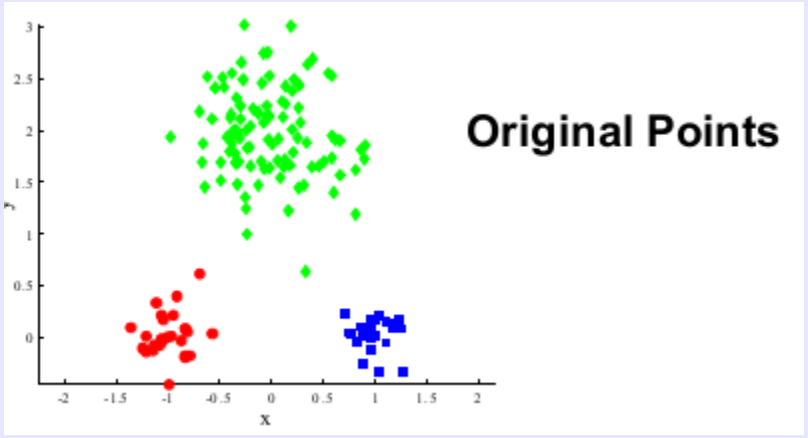


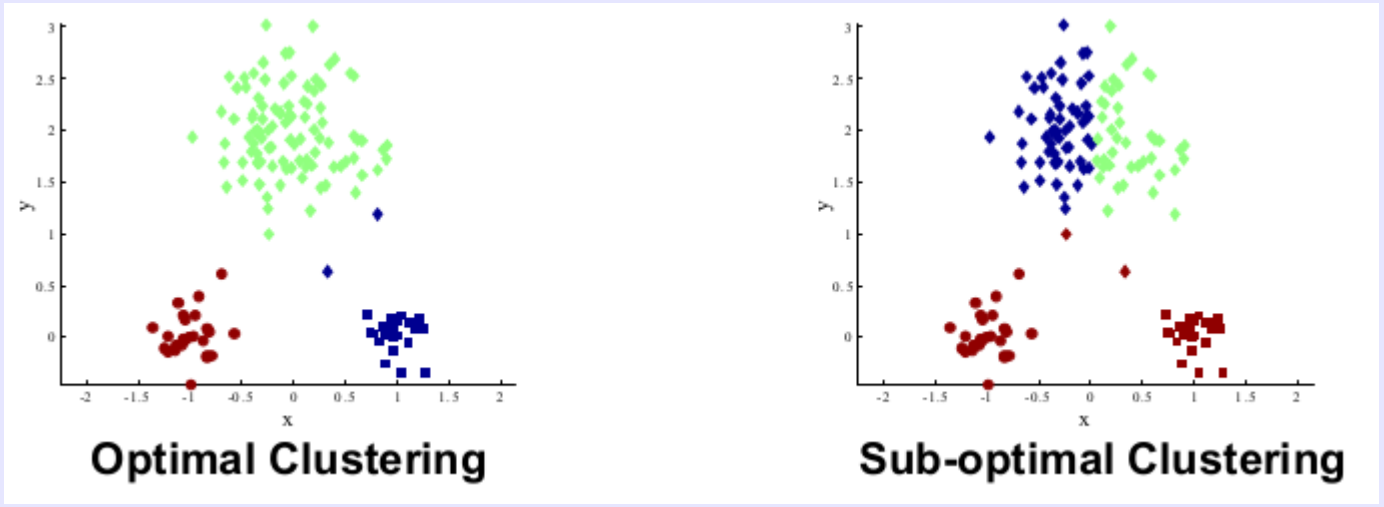
Graphic source: https://en.wikipedia.org/wiki/Hierarchical_clustering#/media/File:Hierarchical_clustering_simple_diagram.svg

Clustering: Algorithms

- K-Means clustering: Takes *n* observations and partitions them into *k* clusters (*k* << *n*).
  - Prototype-based clustering scheme.
  - Computationally difficult (NP-hard); greedy algorithms exist that converge quickly to a local optimum.

Clustering: K-Means





Clustering: K-Means

- Prerequisite:
  - Attributes must be numeric
  - Attributes must be standardized (scaled).

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

- Issues:
  - How do we compute distances to centroids?
  - How do we choose *K*?
  - When do we know when to stop?
  - How do we choose the initial centroids?