Dr. Azadeh Mohammadi

Lecturer in Data Science

University of Salford MANCHESTER

SCHOOL OF SCIENCE, ENGINEERING & ENVIRONMENT

Clustering

Dr. Azadeh Mohammadi

School of Science, Engineering & Environment
University of Salford

Outline

- Basic Concepts
- Partitioning Methods
- Hierarchical Methods
- Cluster Evaluation & Interpretation
- Summary

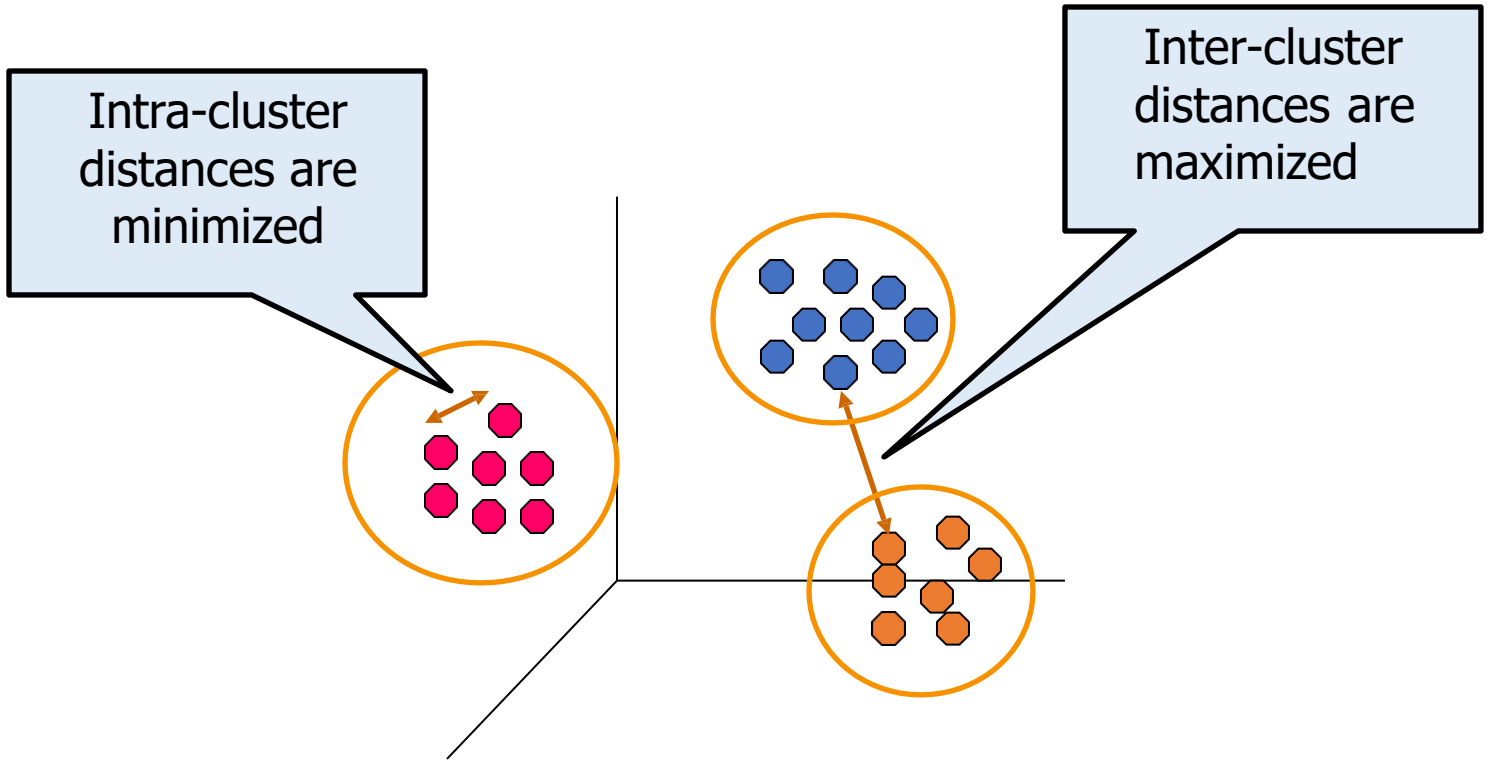
What is clustering?

- It is a process of dividing a set of data into a set of meaningful groups.
  - Objects (samples) in a cluster are similar into each other and dissimilar from objects in other clusters

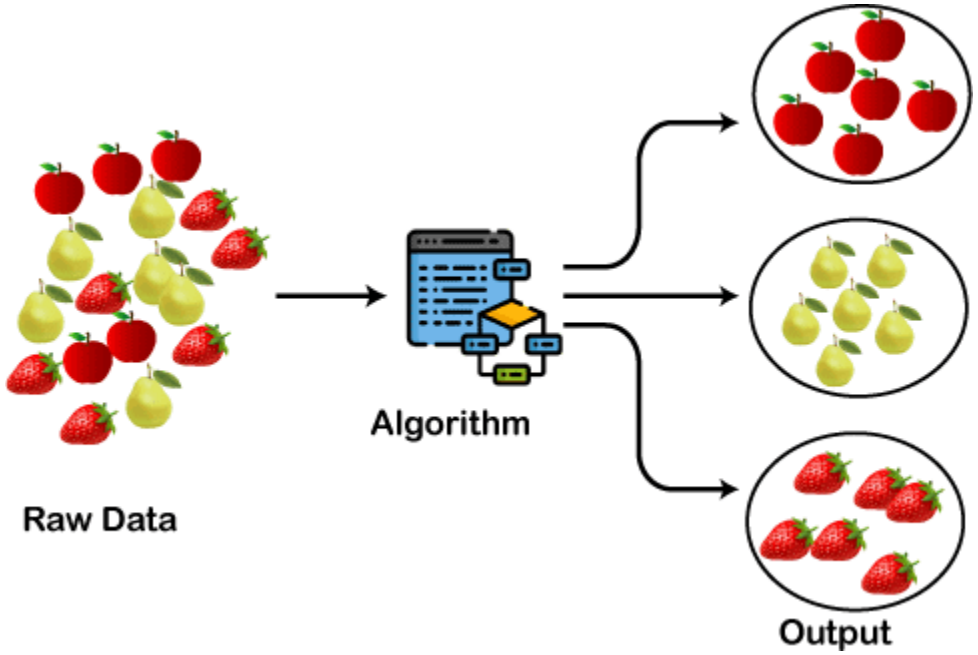
- Unsupervised Learning
  - No training data unlike classification

What is clustering?

- Clustering : given a collection of data objects group them so that
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters



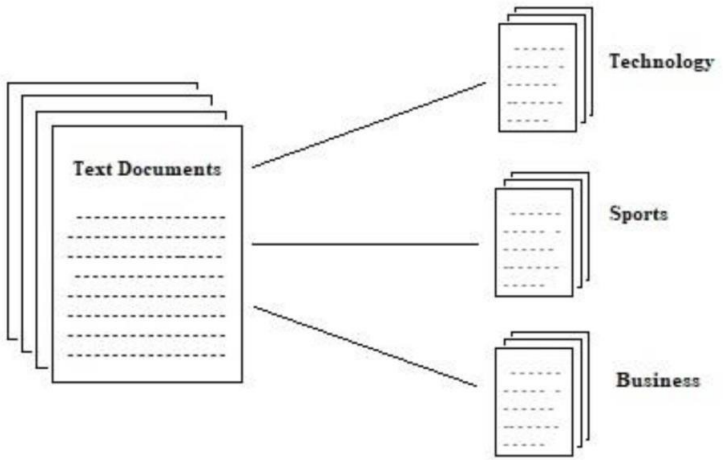
What is clustering?



Recourse: https://www.datacamp.com/

What is clustering?

- Clustering helps the user to understand the natural grouping or structure in a data set
  - e.g., Documents that share same properties are categorized into same clusters.



Notion of a Cluster can be Ambiguous

- We should decide about:
  - Group definition
    - Measure of similarity/ dissimilarity (distance)
  - Number of clusters/ Cluster size

What is the “natural grouping”?

Clustering is very subjective!

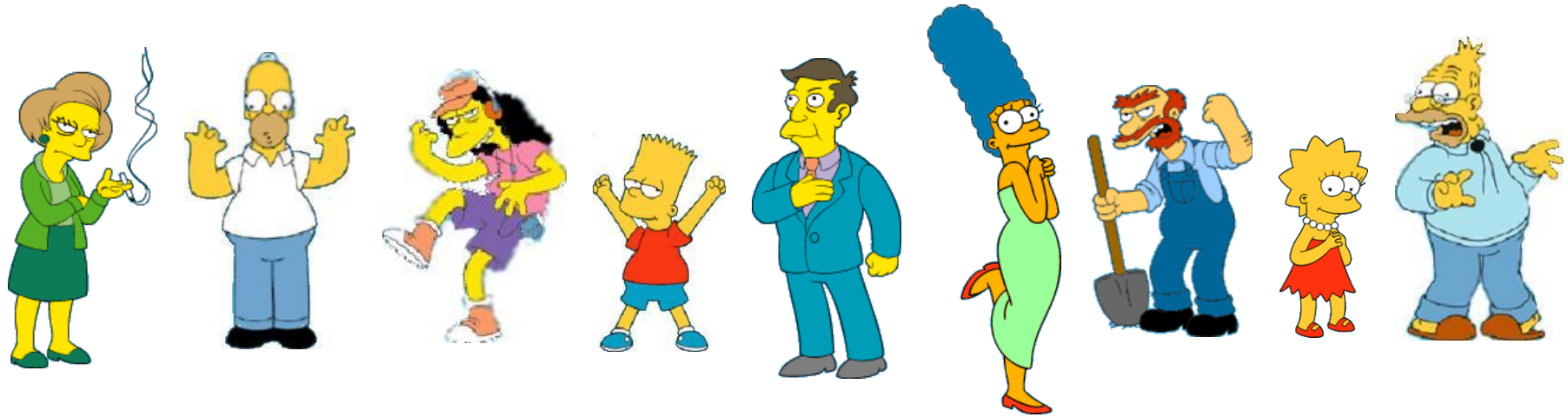
Distance metric is important!

group by gender

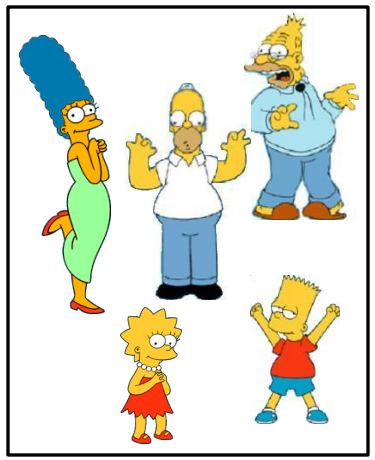
group by source of ability

group by costume

What is the “natural grouping”?

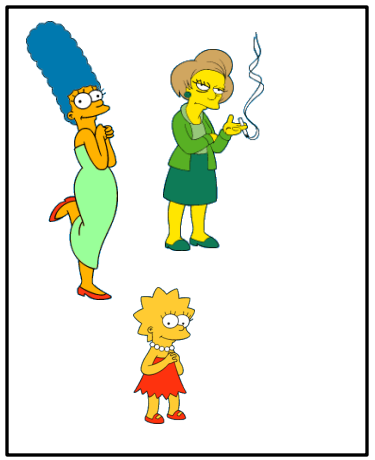


Clustering is subjective!

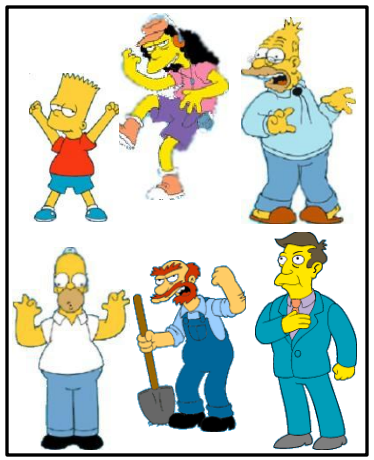


Simpson's Family

School Employees



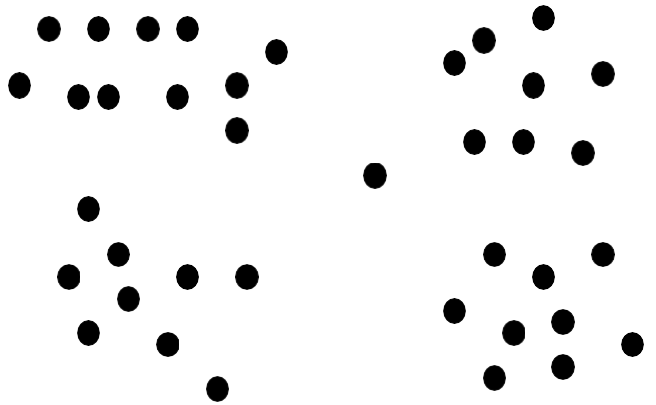
Females

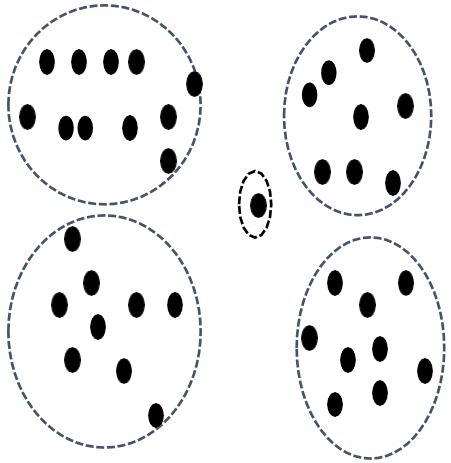


Males

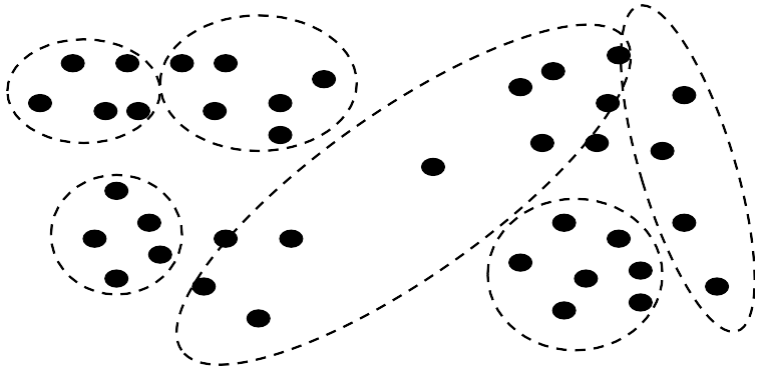
What is the “natural grouping”?

Properties
(houses)



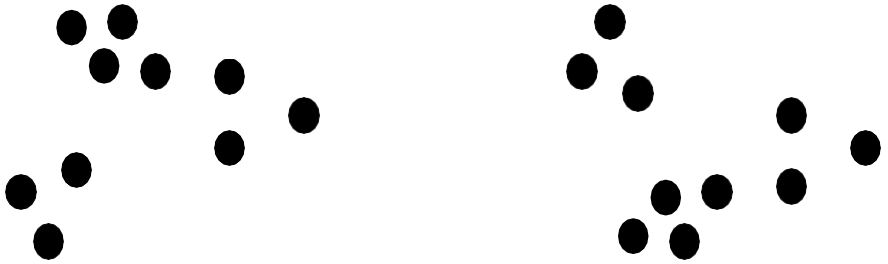


Geographic Distance Based

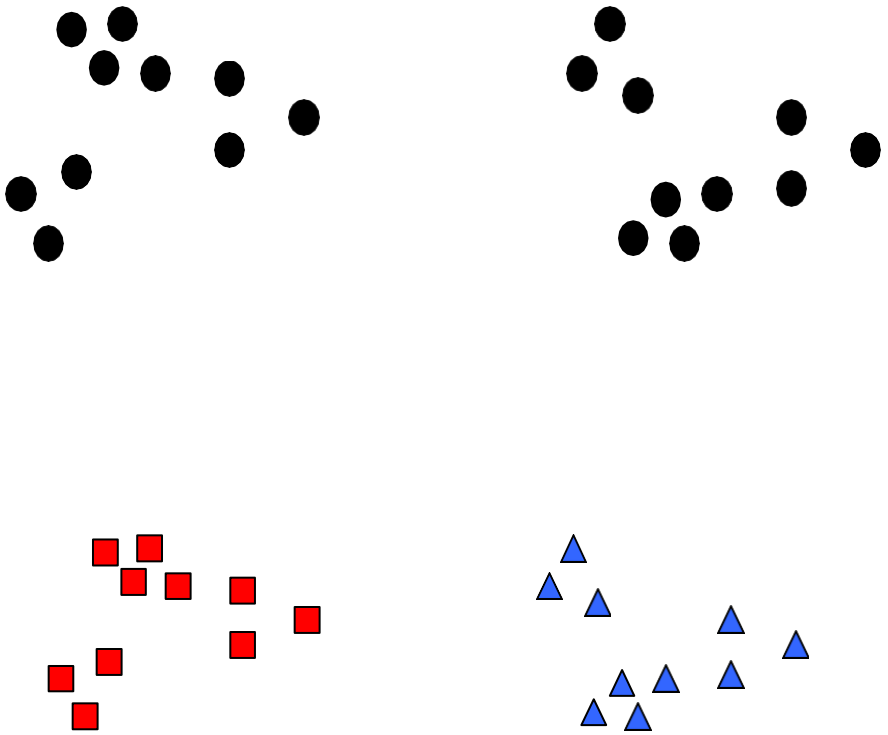


Size Based

How many clusters?

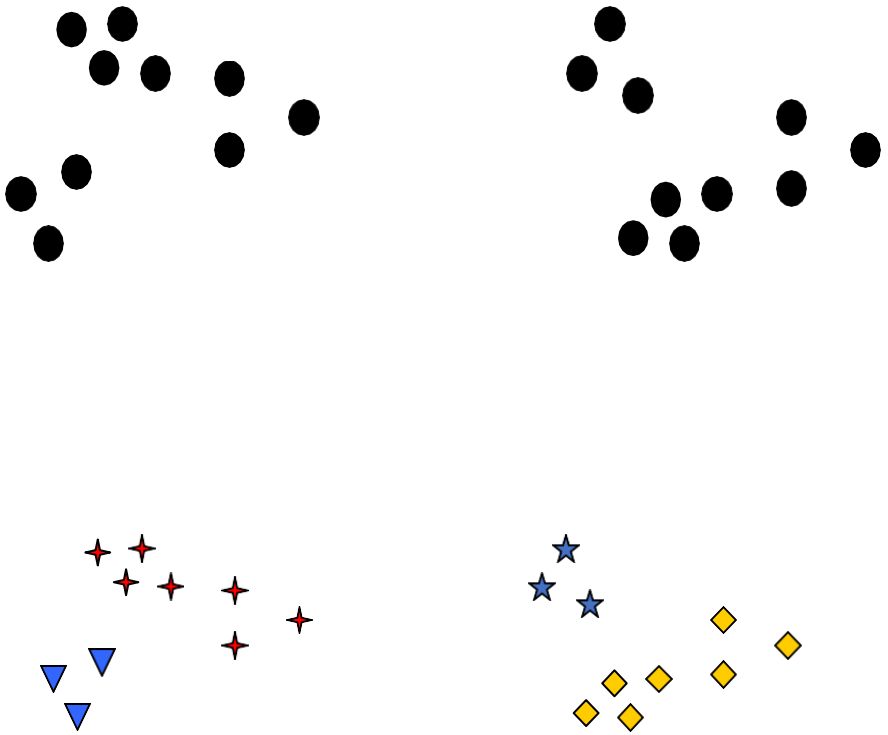


How many clusters?



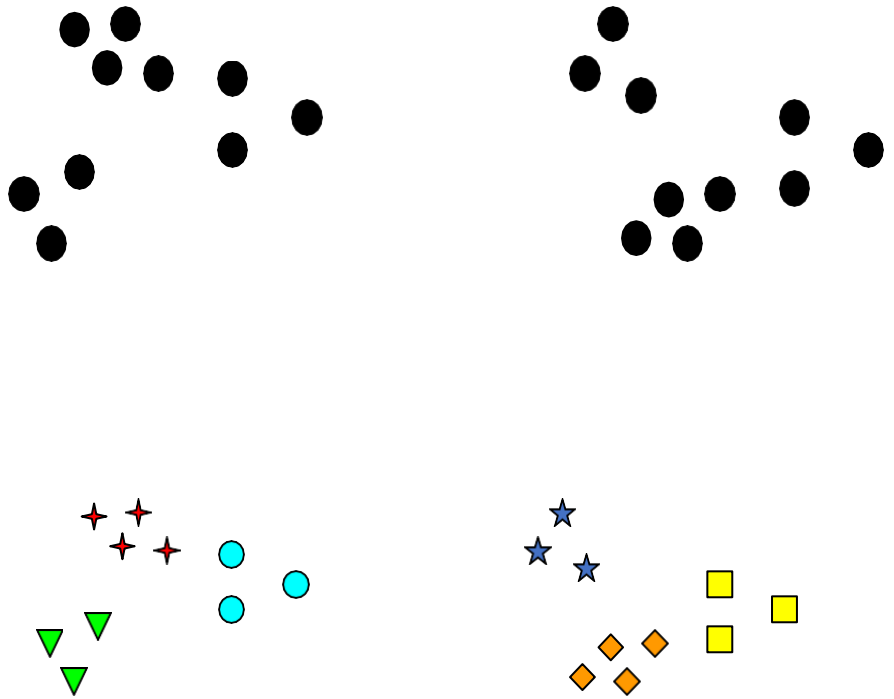
Two Clusters

How many clusters?



Four Clusters

How many clusters?



Six Clusters

Why do we cluster?

- Clustering results are used:
  - As a stand-alone tool to get insight into data distribution
    - Visualization of clusters may unveil important information
  - As a pre-processing step for other algorithms
    - Efficient indexing or compression often relies on clustering
    - Instead of dealing with each item separately, we can deal with a cluster

Applications of clustering?

- Image Processing: Cluster images based on their visual content
- Web: Cluster groups of users based on their access patterns on webpages / Cluster webpages based on their content
- Bioinformatics: Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- Marketing: Find groups of customers with similar behavior given a large database of customer data containing their properties and past buying records
- Biology: Categorization of plants and animals given their features;
- City-planning: Identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: Clustering observed earthquake epicentres to identify dangerous zones;
- …many more

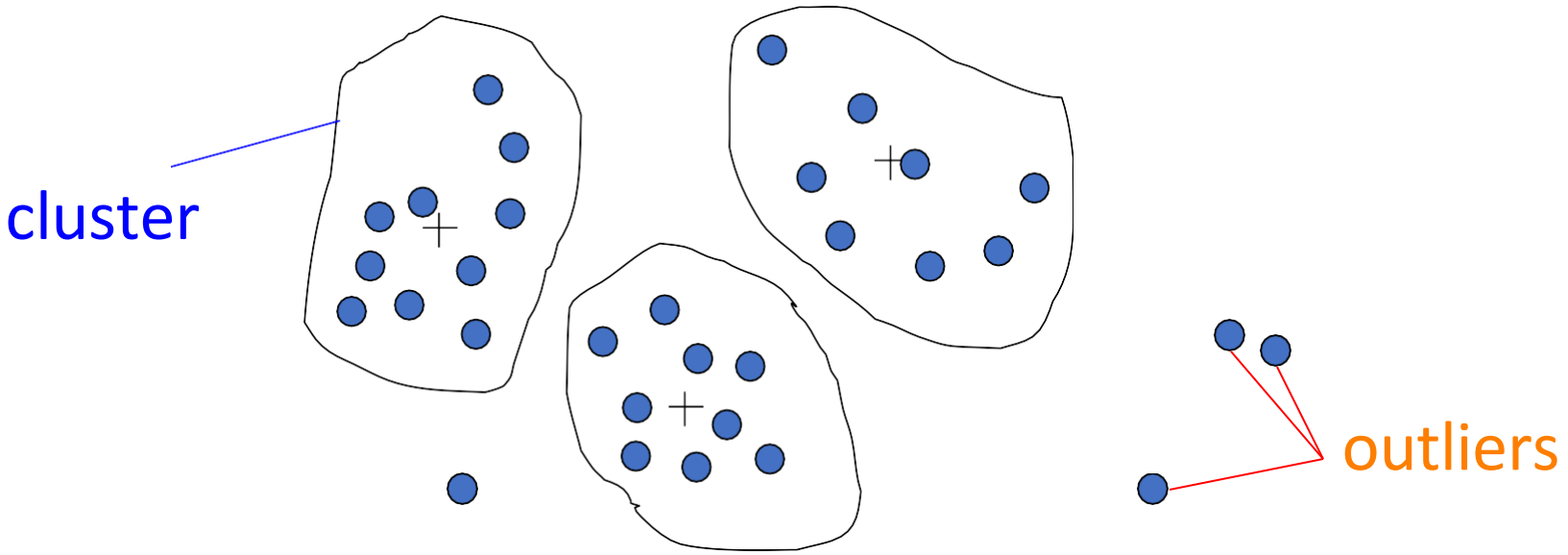
Question

- As a data scientist, explore how clustering can provide advantages to an online retail store

  - Name some applications of clustering in retail store
  - How the clusters can be used?

Outliers

- **Outliers** are **objects that do not belong to any cluster** or form clusters of very small cardinality



Applications of outlier detection

- In some applications we are interested in discovering outliers, not clusters (outlier analysis)
  - Medical diagnosis (Rare disease)
  - Fault detection
  - Fraud detection
  - …

The clustering task

Clustering: Cluster observations into groups so that the observations belonging in the same group are similar, whereas observations in different groups are different

- Basic questions:
  - What does "similar" mean
  - What is a good division of the objects? i.e., how is the quality of a solution measured
  - How to find a good division of the observations

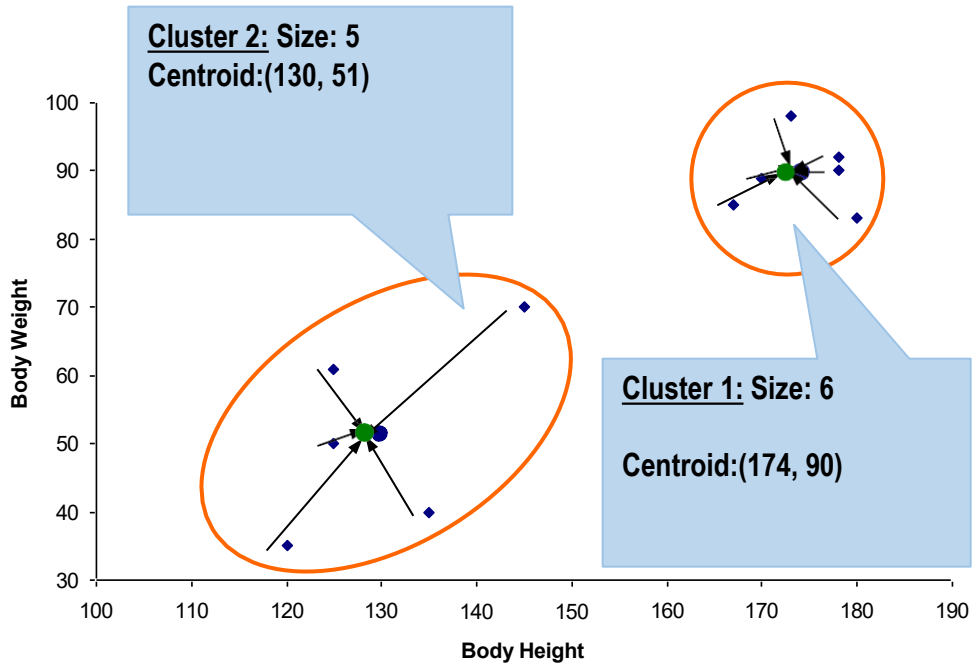
What is good clustering?

- A good clustering method will produce high quality clusters with:
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is measured by its ability to discover some or all of the hidden patterns.

Cluster Detection Problem

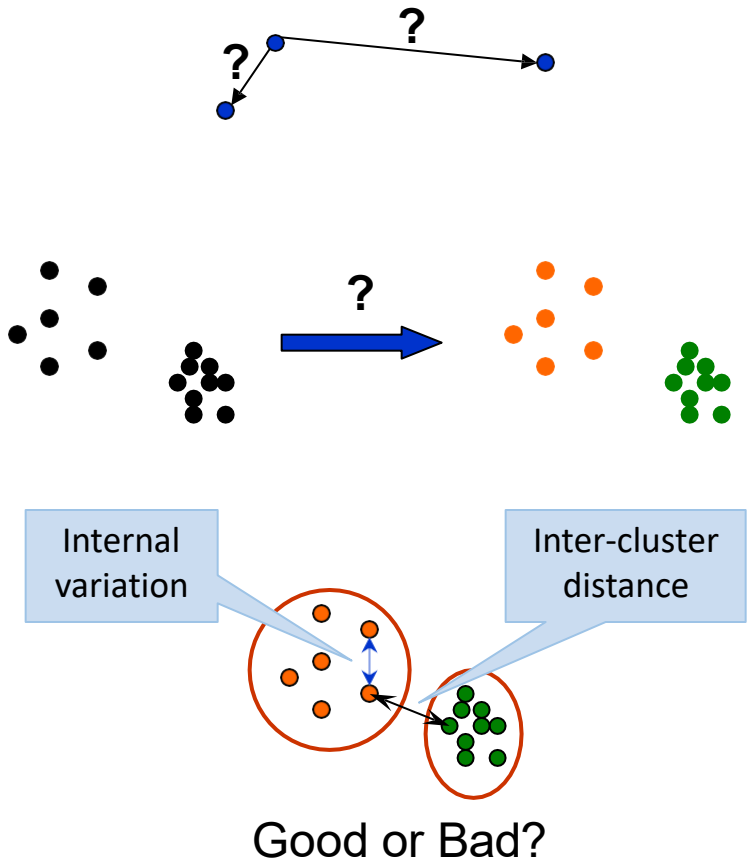
- Outputs of cluster detection process:
  - Assigned cluster tag for each sample
  - Cluster summary: size, centroid, variations, etc.

Subject ID	Body Height	Body Weight	Cluster Tag
s1	125	61	2
s2	178	90	1
s3	178	92	1
s4	180	83	1
s5	167	85	1
s6	170	89	1
s7	173	98	1
s8	135	40	2
s9	120	35	2
s10	145	70	2
s11	125	50	2



Cluster Detection Problem

- Basic elements of a clustering solution
  - 1) A sensible measure for similarity, e.g. Euclidean
  - 2) An effective and efficient clustering algorithm, e.g. K-means
  - 3) A goodness-of-fit function for evaluating the quality of resulting clusters, e.g. SSE(Sum of Squares Due to Error)



Measures of Proximity

- Proximity between two data objects is represented by either similarity or dissimilarity
  - Similarity: a numeric measure of the degree of alikeness
  - dissimilarity: numeric measure of the degree of difference between two objects

- Similarity measure and dissimilarity measure are often convertible; normally dissimilarity is preferred

- Measure of dissimilarity:
  - Measuring the difference between values of the corresponding attributes
  - Combining the measures of the differences

Measures of Proximity

- Distance function
  - Metric properties of function d:
    - $d(x, y) \geq 0$ and $d(x, x) = 0$, for all data objects x and y
    - $d(x, y) = d(y, x)$, for all data objects x and y
    - $d(x, y) \leq d(x, z) + d(z, y)$, for all data objects x, y and z

- Difference of values for a single attribute is directly related to the domain type of the attribute.

- It is important to consider which operations are applicable

Measures of Proximity

- Distance between two data objects:
  - Numerical value (interval/ratio attributes)
    - Minkowski function

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+...+|x_{ip}-x_{jp}|^q)}$$

    - Special cases:

Manhattan distance (q =1) $d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+...+|x_{ip}-x_{jp}|$

Euclidean distance (q =2) $d(i,j)=\sqrt{(|x_{i_1 j_1}-|^2+|x_{i2}-x_{j2}|^2+...+|x_{ip}-x_{jp}|^2)}$

Supremum/Chebyshev (q =∞) $d(i,j)=\max_t |x_{it}-xj_t.|$

Measures of Proximity

- Distance between two data objects:
  - Binary or Nominal features:
    - *Ratio of mismatched features:* Given two data objects i and j of p binary/nominal attributes. Let m represent the number of attributes where the values of the two objects match, their distance is:

$$d(i, j) = \frac{p - m}{p}$$

Measures of Proximity

- e.g.:

Body Weight	Body Height	Blood Pressure	Blood Sugar	Habit	Class
heavy	short	high	low	smoker	P
heavy	short	high	high	nonsmoker	P
normal	tall	normal	low	nonsmoker	N
heavy	tall	normal	high	smoker	N
low	medium	normal	low	nonsmoker	N
low	tall	normal	low	nonsmoker	P
normal	medium	high	high	smoker	P
low	short	high	low	smoker	P
heavy	tall	high	low	nonsmoker	P
low	medium	normal	high	smoker	P
heavy	medium	normal	low	nonsmoker	N

$$d(row1, row2) = \frac{6-4}{6} = \frac{1}{3}$$

$$d(row1, row3) = \frac{6-1}{6} = \frac{5}{6}$$

Measures of Proximity

- Distance between two data objects:
  - Ordinal features:
    - Converting ordinal values to consecutive integers:
      - e.g., A, B, C, D, E

        A: 5, B: 4, C: 3, D: 2, E:1.

        A – B $\Rightarrow$ 1 and A – D $\Rightarrow$ 3

    Then we can use distance function for numerical values

Types of Clustering

- There are different types of clustering methods
- Important distinction between hierarchical and partitional sets of clusters
  - Partitional Clustering
    - Divide data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
  - Hierarchical clustering
    - Divide data objects into a set of nested clusters organized as a hierarchical tree

Types of Clustering

- Some of the most important types of clustering:
  - Centroid-based:
    - In this approach, each cluster is represented by a central vector, and the objects are assigned to the clusters such that the distance from the central vector is minimized
    - Typical algorithms: k-means, k-medoids
  - Hierarchical:
    - In this approach, a hierarchical decomposition of the set of data (or objects) using some criteria is created
    - Typical algorithms : Agglomerative, Divisive
  - Density based:
    - This approach is based on the notion that clusters are dense regions in the data space, separated by regions of the lower density of points
    - Typical algorithms: DBSCAN
  - There are other types as well

Centroid-based methods

- Centroid-based methods:
  - Given *k*, find a partition of *k* clusters that optimizes the chosen partitioning criterion, e.g., partitioning a database *D* of *n* objects into a set of *k* clusters, such that the sum of squared distances is minimized (where *ci* is the centroid or medoid of cluster *Ci*)

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

  - Main centroid-based algorithms:
    - k-means (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
    - k-medoids (PAM (Partition Around Medoids) Algorithm) (Kaufman & Rousseeuw'87): Each cluster is represented by medoid which is an object in the cluster whose average dissimilarity to all the objects in the cluster is minimal

Centroid-based methods

- Non-hierarchical
- Creates clusters in one step as opposed to several steps
- Since only one set of clusters is output, the user normally has to input the desired number of clusters, k
- Usually deals with static sets.

Centroid-based methods

- K-means: Given *k*, the k-means algorithm is implemented in four steps:

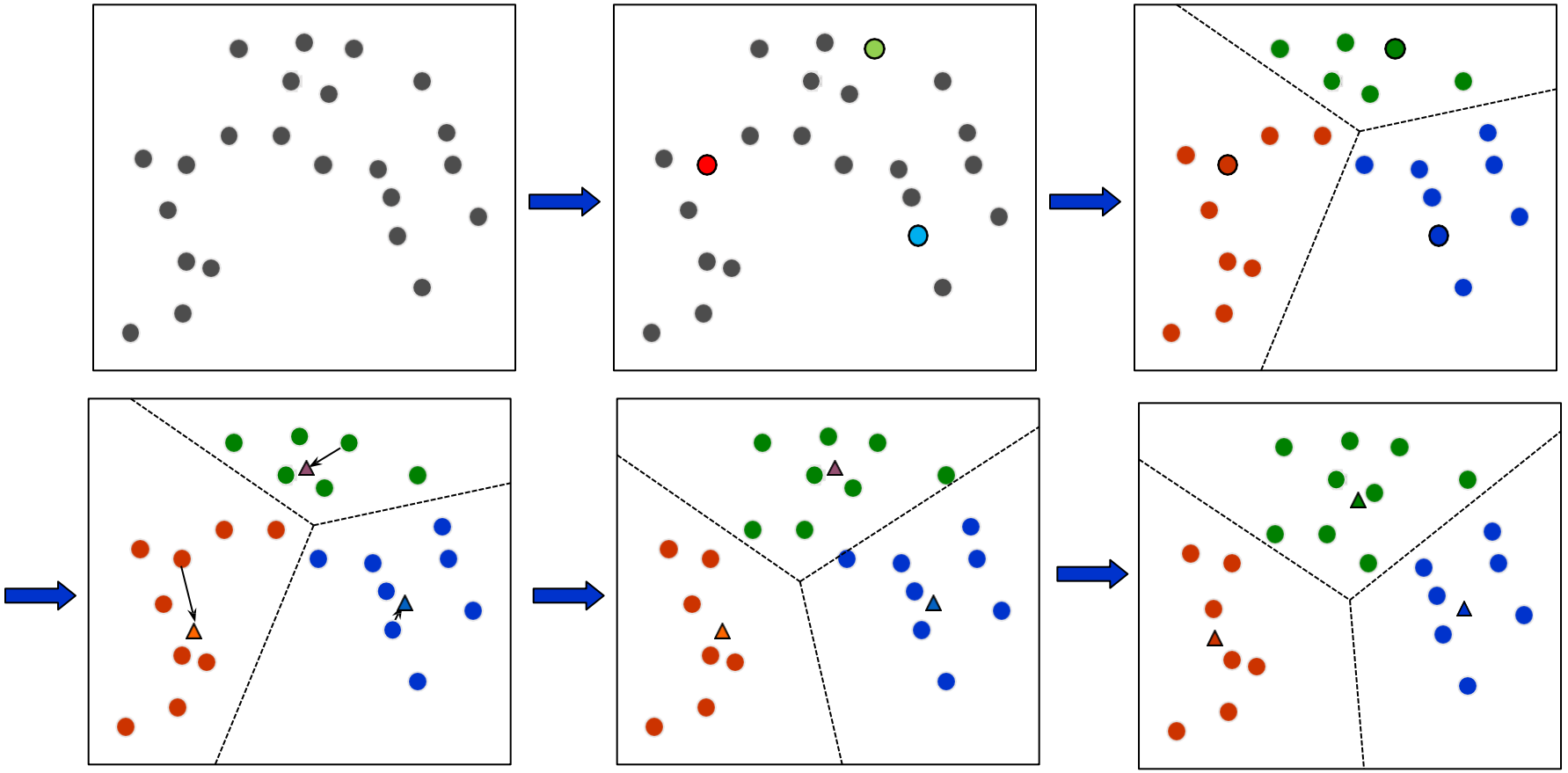
1. Choose k data objects randomly to serve as the initial centroids for the *k* clusters
2. Calculate the distance of each data point to all cluster centers and assign each data object to the nearest cluster represented by its centroid
3. Find a new centroid for each cluster by calculating the mean vector of its members

   Cluster Ki = {ti1,ti2,…,tim}, the cluster mean is mi = (1/m)(ti1 + … + tim)

4. Go back to Step 2 and repeat the process until cluster membership no longer changes or a maximum number of iterations is reached.

Centroid-based methods

- Example:



Centroid-based methods

- Example of k-means:
  - Given: {2,4,10,12,3,20,30,11,25}, k=2
  - Randomly assign means (centres): $m_1=1$, $m_2=6$
  - $K_1=\{2,3\}$, $K_2=\{4,10,12,20,30,11,25\}$, $m_1=2.5, m_2=16$
  - $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$
  - $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$
  - $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$, $m_1=7, m_2=25$
  - Stop as the clusters with these means are the same.

Centroid-based methods

- **Strengths & weaknesses of k-mean**

  - Strengths
    - Simple and easy to implement
    - Quite efficient

  - Weaknesses
    - Need to specify the value of k, but we may not know what the value should be beforehand
    - Sensitive to noise

Hierarchical Clustering

- Clusters are created in levels, actually creating sets of clusters at each level

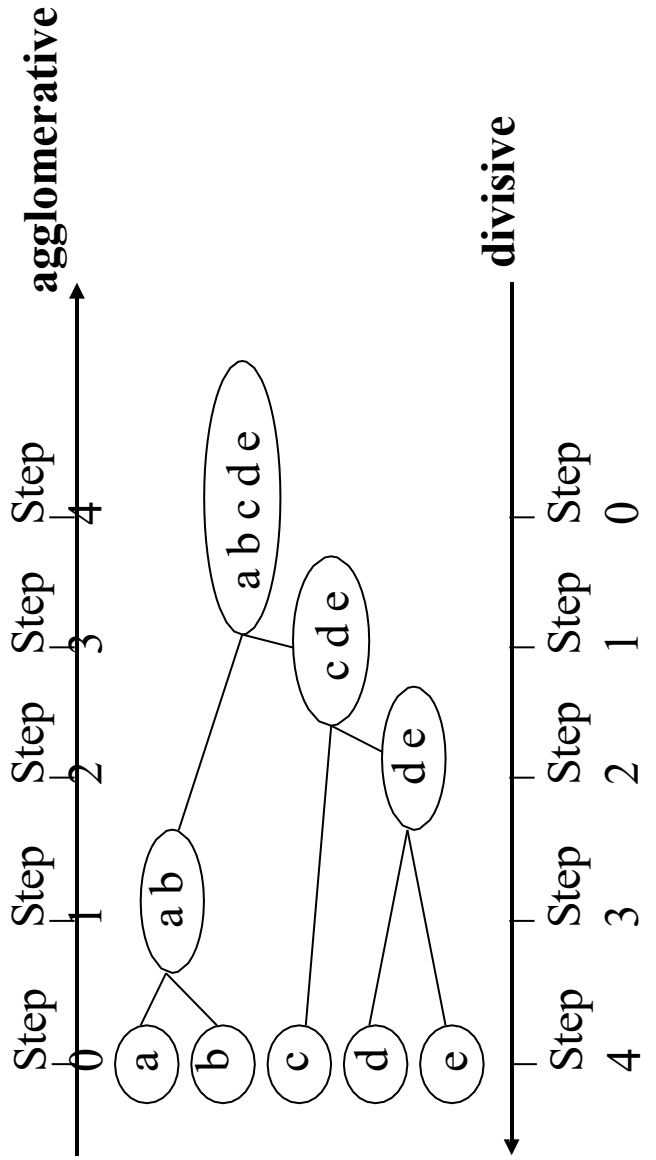
- Algorithms:

  - Agglomerative
    - Bottom Up
    - Initially each item in its own cluster
    - Iteratively clusters are merged together

  - Divisive
    - Top Down
    - Initially all items in one cluster
    - Large clusters are successively divided

Hierarchical Clustering

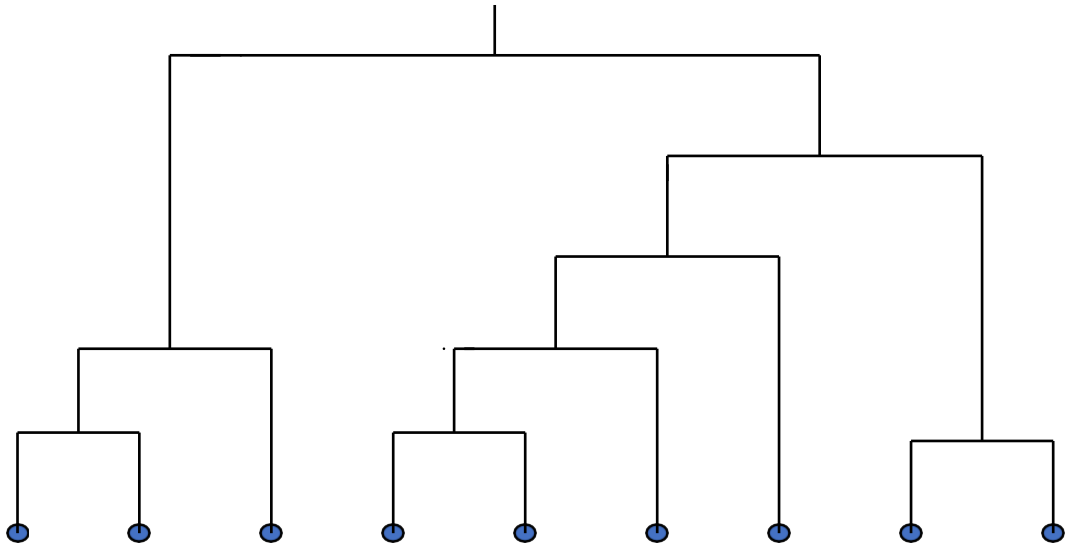
- Use distance matrix as clustering criteria.
- This method does not require the number of clusters k as an input, but needs a termination condition



Hierarchical Clustering

- Hierarchical Clustering decomposes data objects into several levels of nested partitioning (tree of clusters), called a dendrogram

- Dendrogram is a tree data structure which illustrates hierarchical clustering output

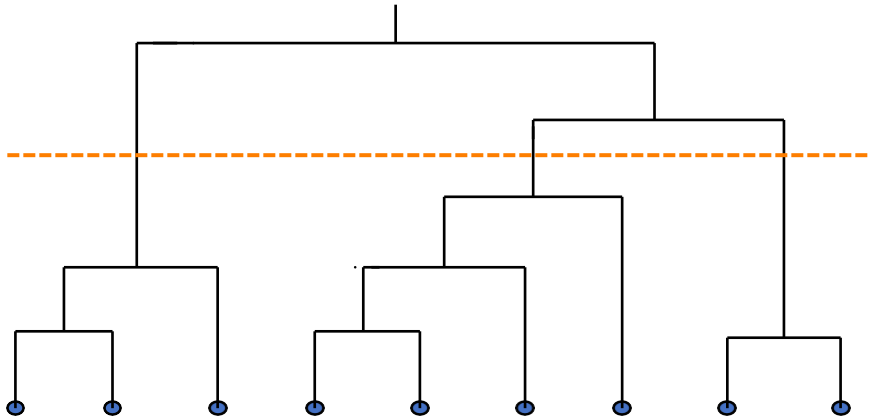


Hierarchical Clustering

- Each level shows clusters for that level.
  - Leaf : individual clusters
  - Root : one cluster

- A cluster at level i is the union of its children clusters at level i+1

- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster



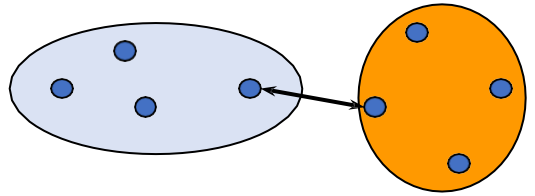
Hierarchical Clustering

- When using hierarchical clustering it is necessary to specify both the *distance metric* and the *linkage criteria*
  - *Distance metric: How to define distance of two samples*
    - *Euclidean distance*
    - *Manhattan*
    - *…*
  - *Linkage criteria: How to define distance of two clusters*
    - *Single link*
    - *Complete link*
    - *Average link*

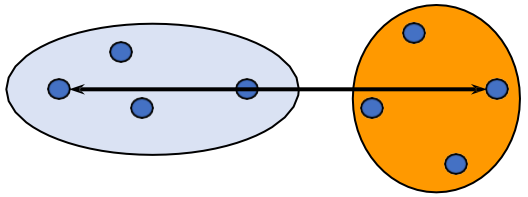
Hierarchical Clustering

- Distance between clusters:

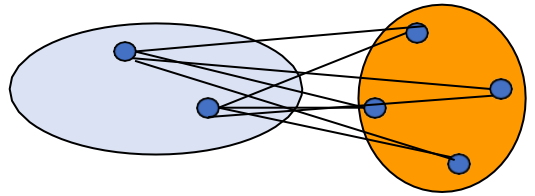
- Single link: smallest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = min(t_{ip}, t_{jq})$



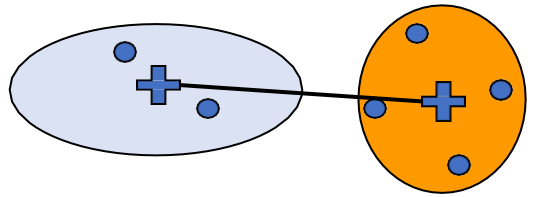
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = max(t_{ip}, t_{jq})$



- Average link: Avg distance between elements in one cluster and elements in the other, i.e., $dist(K_i, K_j) = avg(t_{ip}, t_{jq})$



- Centroid: distance between the centroids of two clusters, i.e., $dist(K_i, K_j) = dist(C_i, C_j)$



Hierarchical Clustering

- Centroid: the "middle" of a cluster

$$C_i = \frac{\Sigma_{p=1}^{N} (t_{ip})}{N}$$

- Radius: square root of average distance from any point of the cluster to its centroid

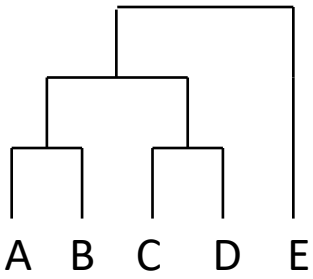
$$R_i = \sqrt{\frac{\Sigma_{p=1}^{N} (t_{ip} - c_i)^2}{N}}$$

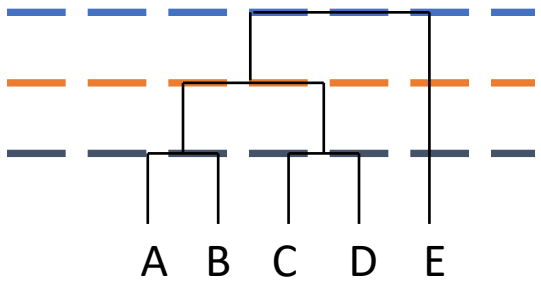
Hierarchical Clustering

- Agglomerative:
  1. Each data object is a cluster
  2. Combine two most similar clusters
  3. Repeat step 2 until all objects are in one cluster

Distance matrix

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



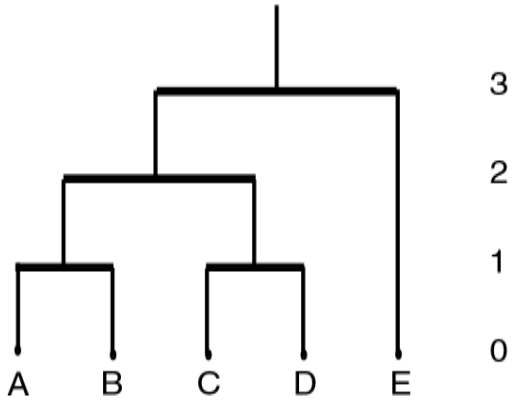


Threshold of

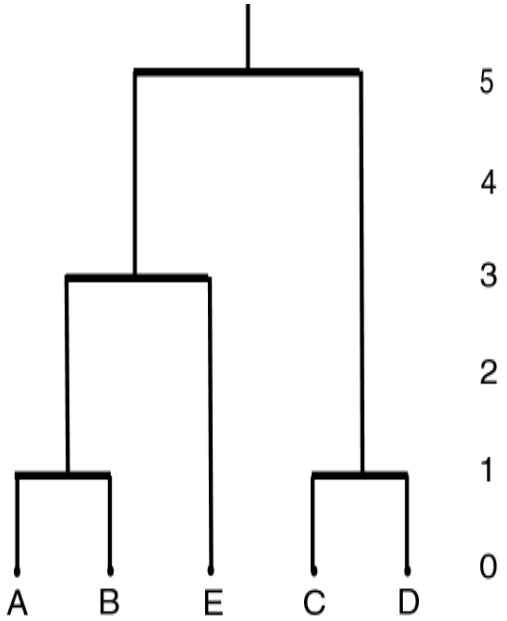
1 2 3

Hierarchical Clustering

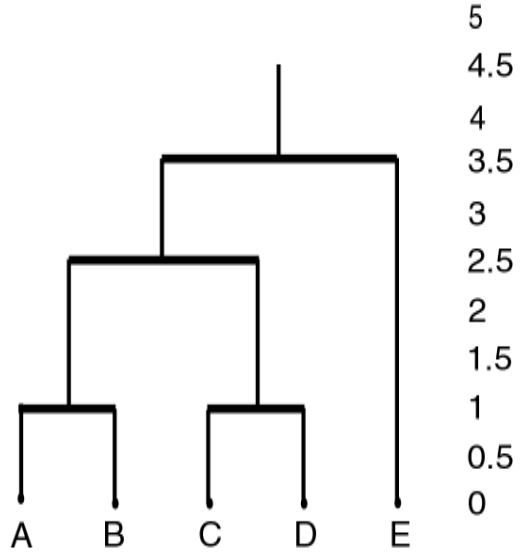
	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



a) Single Link



b) Complete Link



c) Average Link

Hierarchical Clustering

- **Strengths and weaknesses of Agglomerative algorithm**

  - Strengths
    - Deterministic results
    - Multiple possible versions of clustering
    - No need to specify the value of a k beforehand (We can decide about number of clusters based on dendrogram)

  - Weaknesses
    - Does not scale up for large data sets
    - Cannot undo membership like the K-means

Question

- Use complete link agglomerative clustering to group the data described by the following distance matrix. Which samples should be merged in the first step? What about next step? Imagine complete link is the linkage criteria

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Cluster Evaluation & Interpretation

- Cluster quality:
  - Principle:
    - High-level similarity(low-level dissimilarity) within a cluster
    - Low-level similarity (High-level dissimilarity) between clusters
  - The measures
    - Cohesion (Compactness): is measured by the within cluster sum of squares of distance (the sum of SSEs for all clusters) (WC). Lower level dissimilarity shows higher cohesion
    - Separation (Isolation): is measured by sum of distances between clusters (BC)
    - Combining the cohesion and separation, the ratio BC/WC is a good indicator of overall quality.

C_k: cluster k
c_k: centroid of C_k

$$SSE(C_k) = \sum_{x \in C_k} d(x, c_k)^2$$

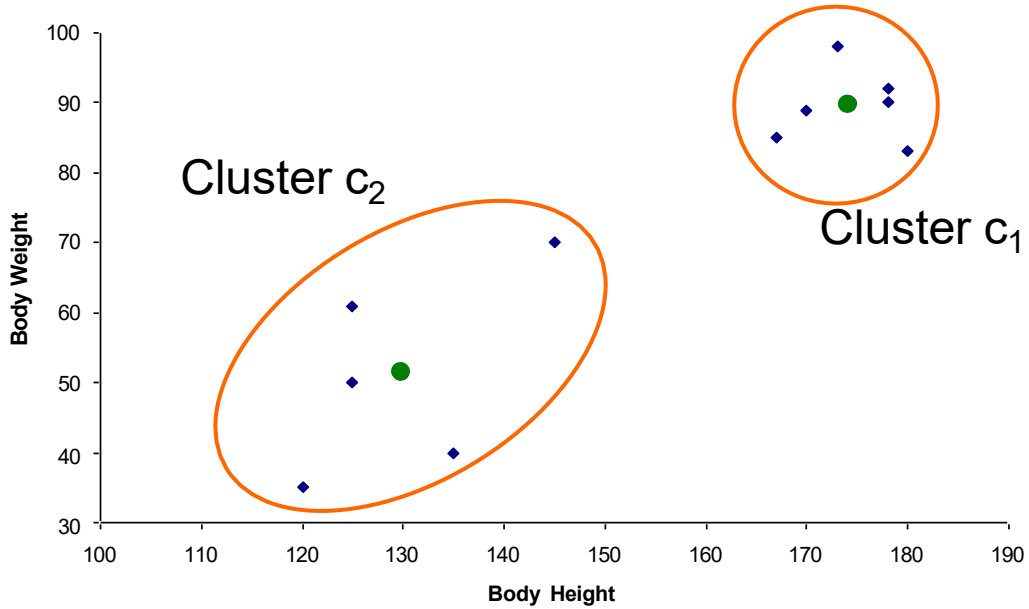
$$WC = \sum_{k=1}^{K} SSE(C_k)$$

$$BC = \sum_{1 \le j < k \le K} d(c_j, c_k)^2$$

$$Q = \frac{BC}{WC}$$

Cluster Evaluation & Interpretation

SubjectID	Body Height	Body Weight	Cluster Tag
s1	125	61	2
s2	178	90	1
s3	178	92	1
s4	180	83	1
s5	167	85	1
s6	170	89	1
s7	173	98	1
s8	135	40	2
s9	120	35	2
s10	145	70	2
s11	125	50	2



$SSE(C_1) \approx 277.00$

$SSE(C_2) = 1238.8$

$WC = 277.00 + 1238.8 = 1515.80$

$BC \approx 3432.3$

$\therefore$ C1 is a higher-quality cluster than C2.

$\therefore Q = \frac{3432.3}{1515.80} \approx 2.268$

Cluster Evaluation & Interpretation

- Silhouette score:
  - For an individual point i, the silhouette coefficient is:

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

- Where
  - ai : is the intra cluster distance defined as the average distance to all other points in the cluster to which i belongs
  - bi : is the inter cluster distance defined as the average distance to closest cluster of point i (the lowest average distance between point *i*i and points in any other cluster)

- Overall Silhouette score for the complete dataset can be calculated as the mean of silhouette score for all data points in the dataset

- The silhouette score of 1 means that the clusters are very dense and nicely separated