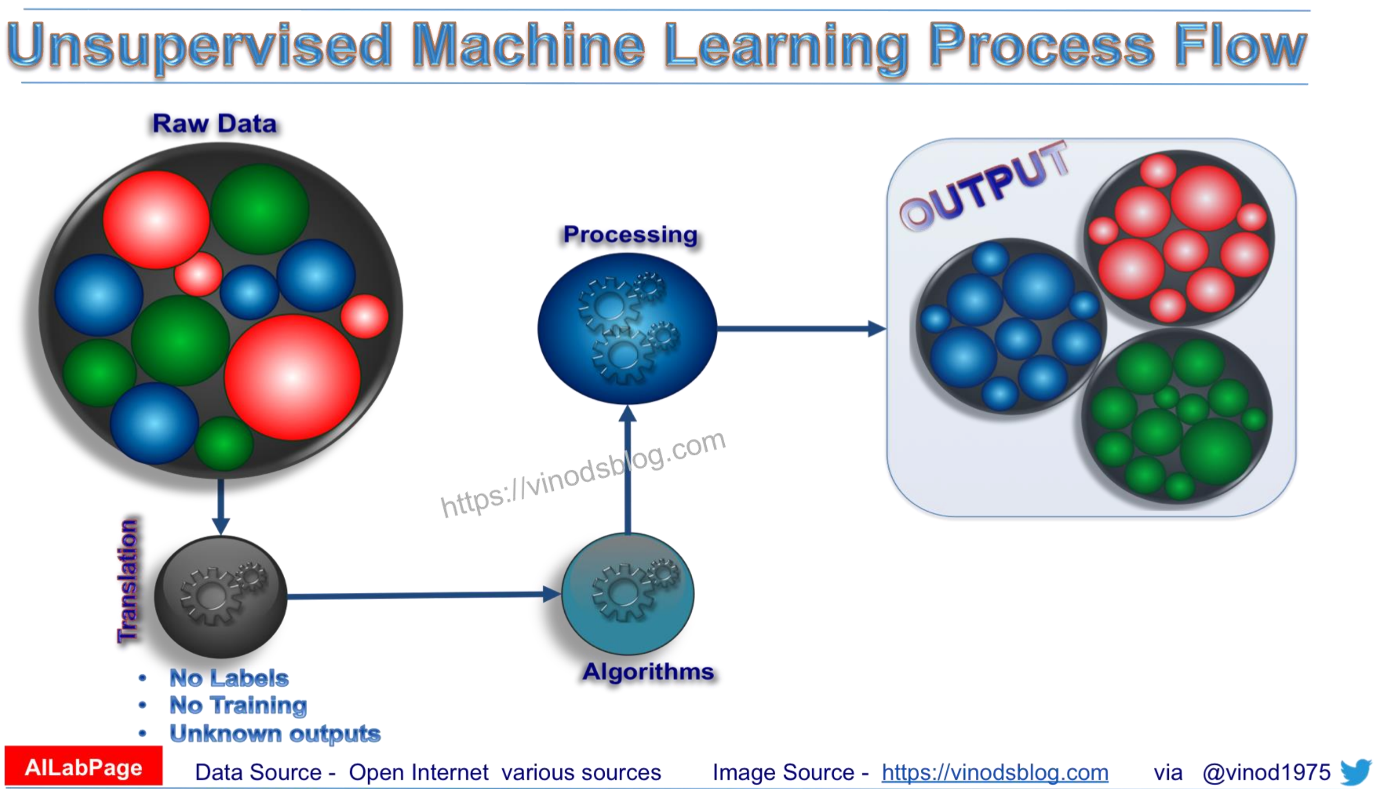
Clustering/Unsupervised Learning

Dr. Ujwala Bharambe



Model Building : Unsupervised Learning

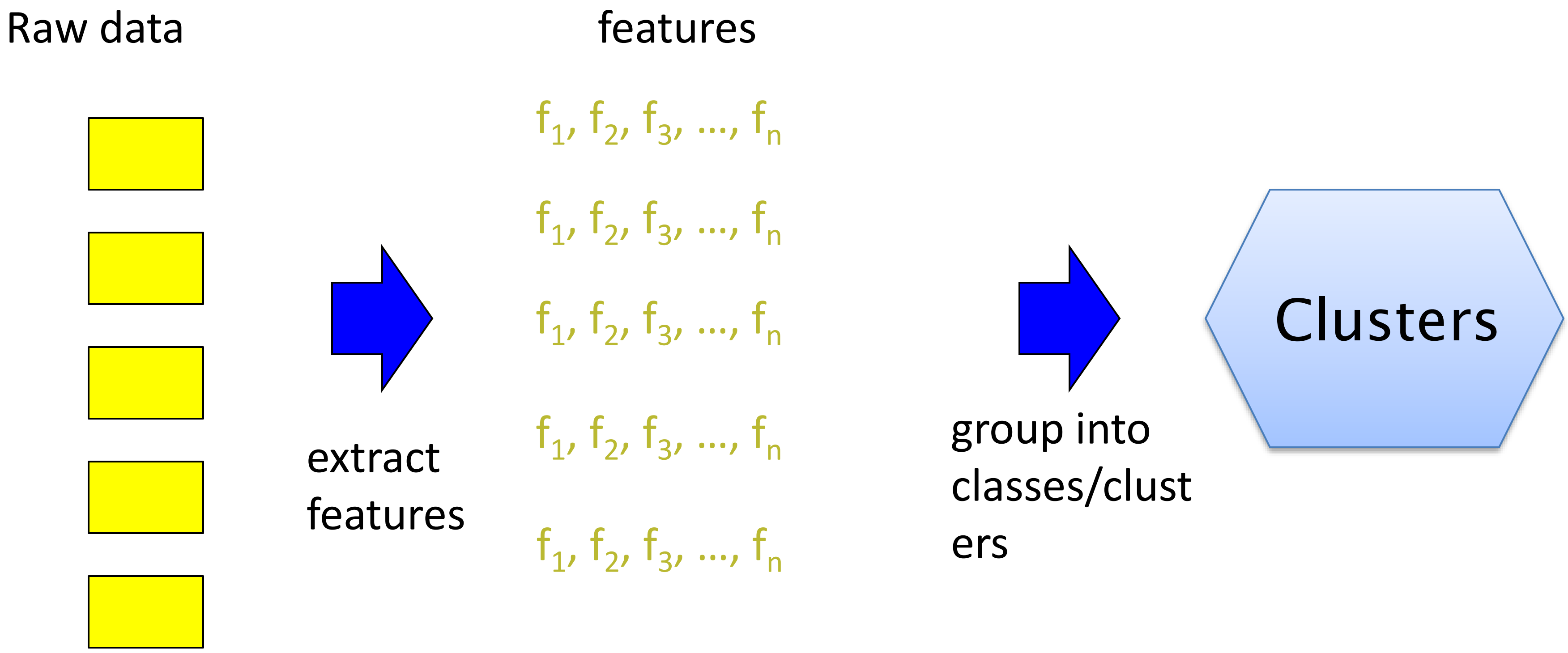




Unsupervised learning: given data, i.e. examples, but no labels

Model Building : Unsupervised Learning

Clustering



No "supervision", we're only given data and want to find natural groupings

Unsupervised learning: clustering

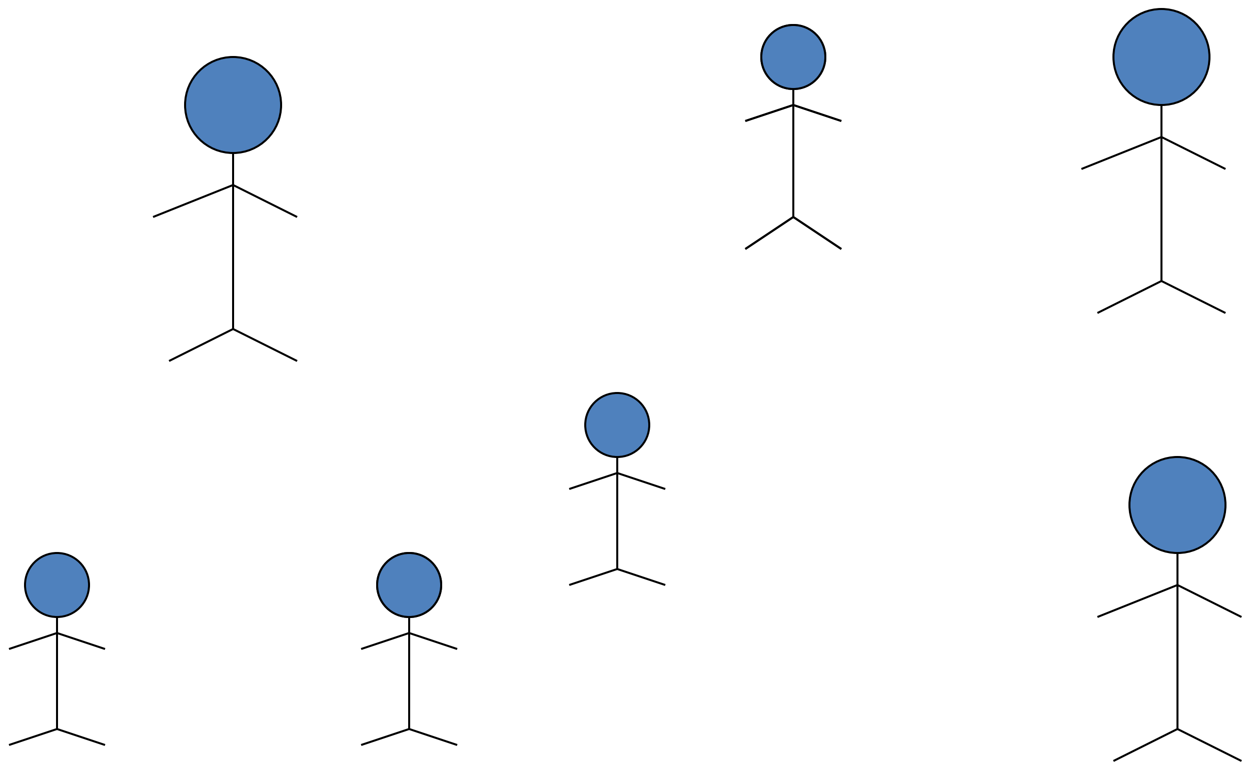
Model Building : Unsupervised Learning

What is Clustering ?

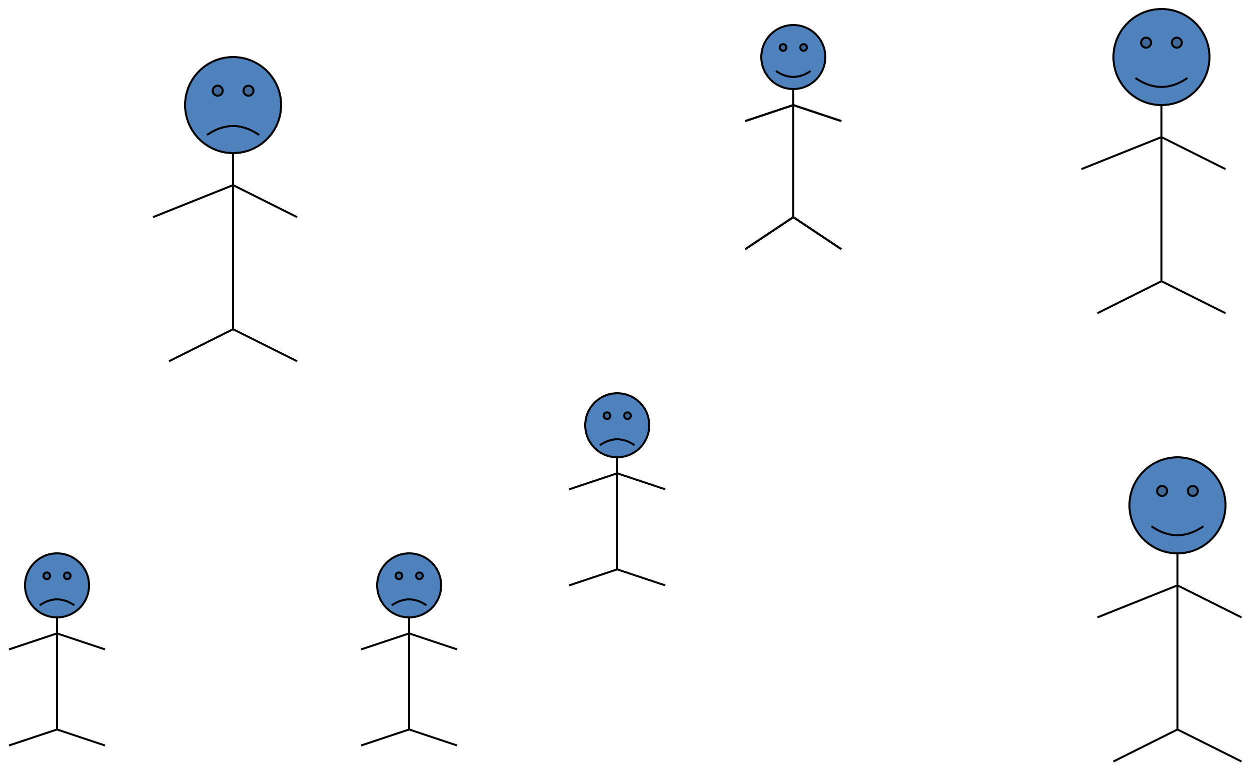
- Discovering similarities in a set of objects
- Decrease the amount of information and express it in a concise way
- Clusters structure is not known in advance (classification), however similarity/ measure is usually given a priority.
- **The obvious: structure of clusters is highly subjective to object features / criteria considered in an algorithm**

What is clustering about?

Discovering similarities in a set of objects

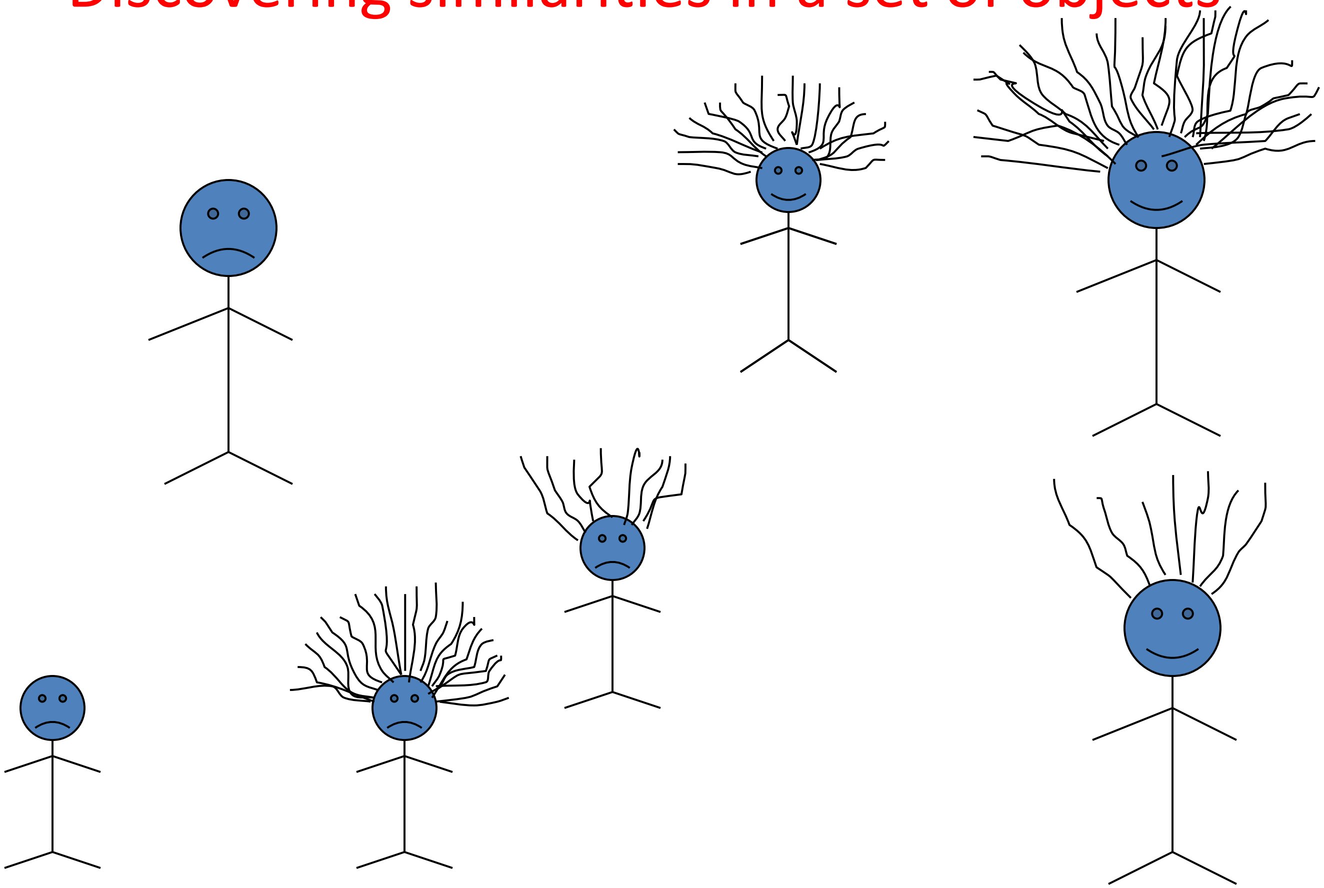


What is clustering about?
Discovering similarities in a set of objects



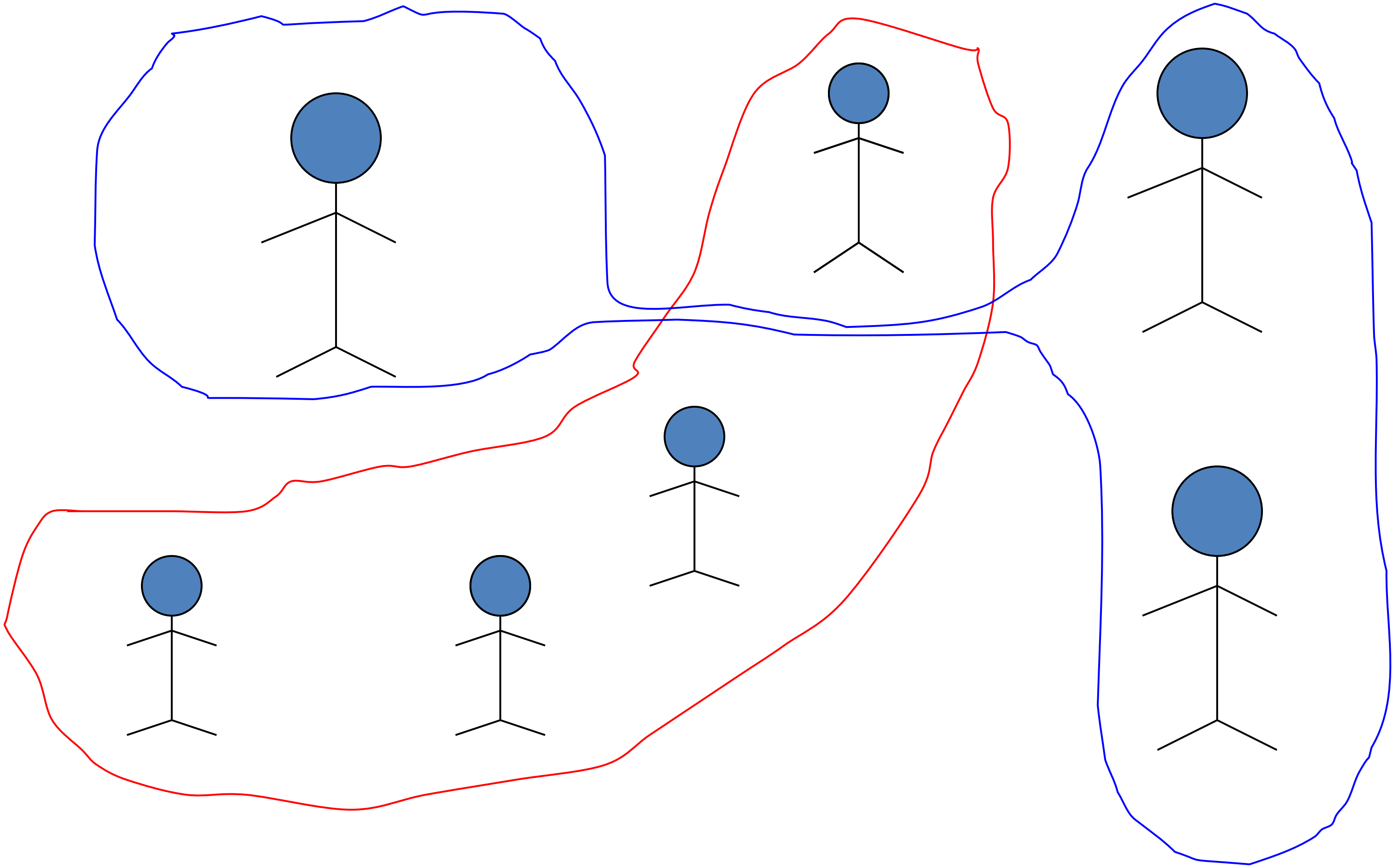
What is clustering about?

Discovering similarities in a set of objects



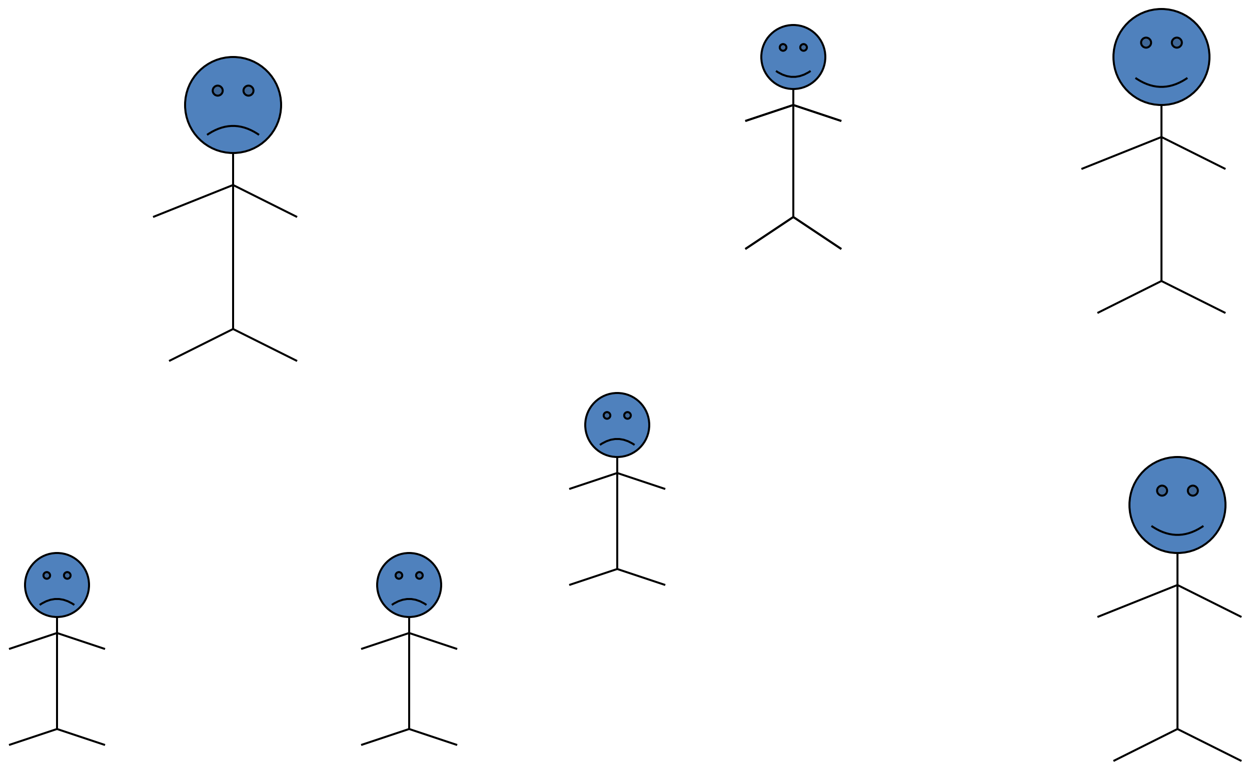
What is clustering about?

Discovering similarities in a set of objects



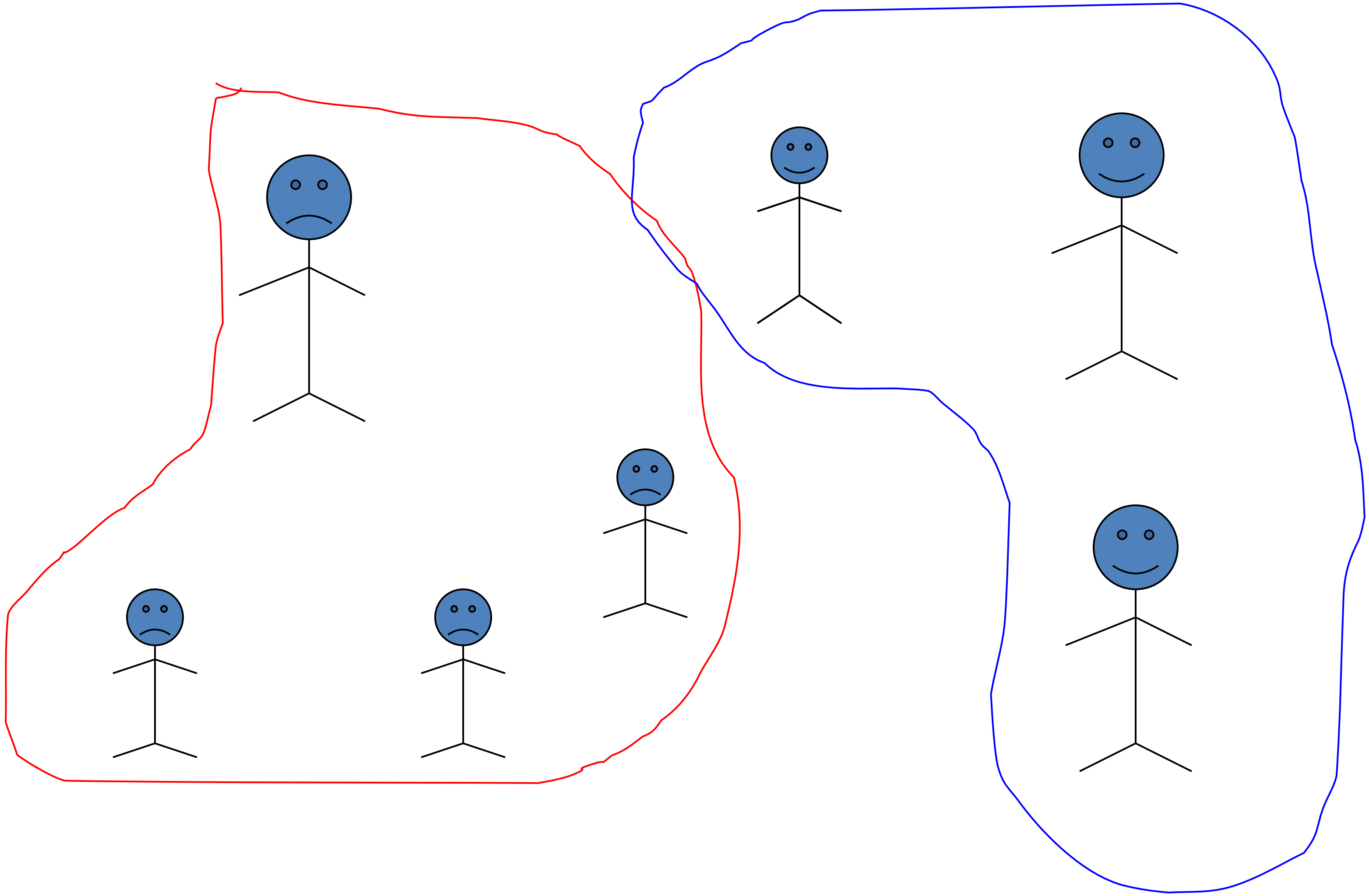
What is clustering about?

Discovering similarities in a set of objects



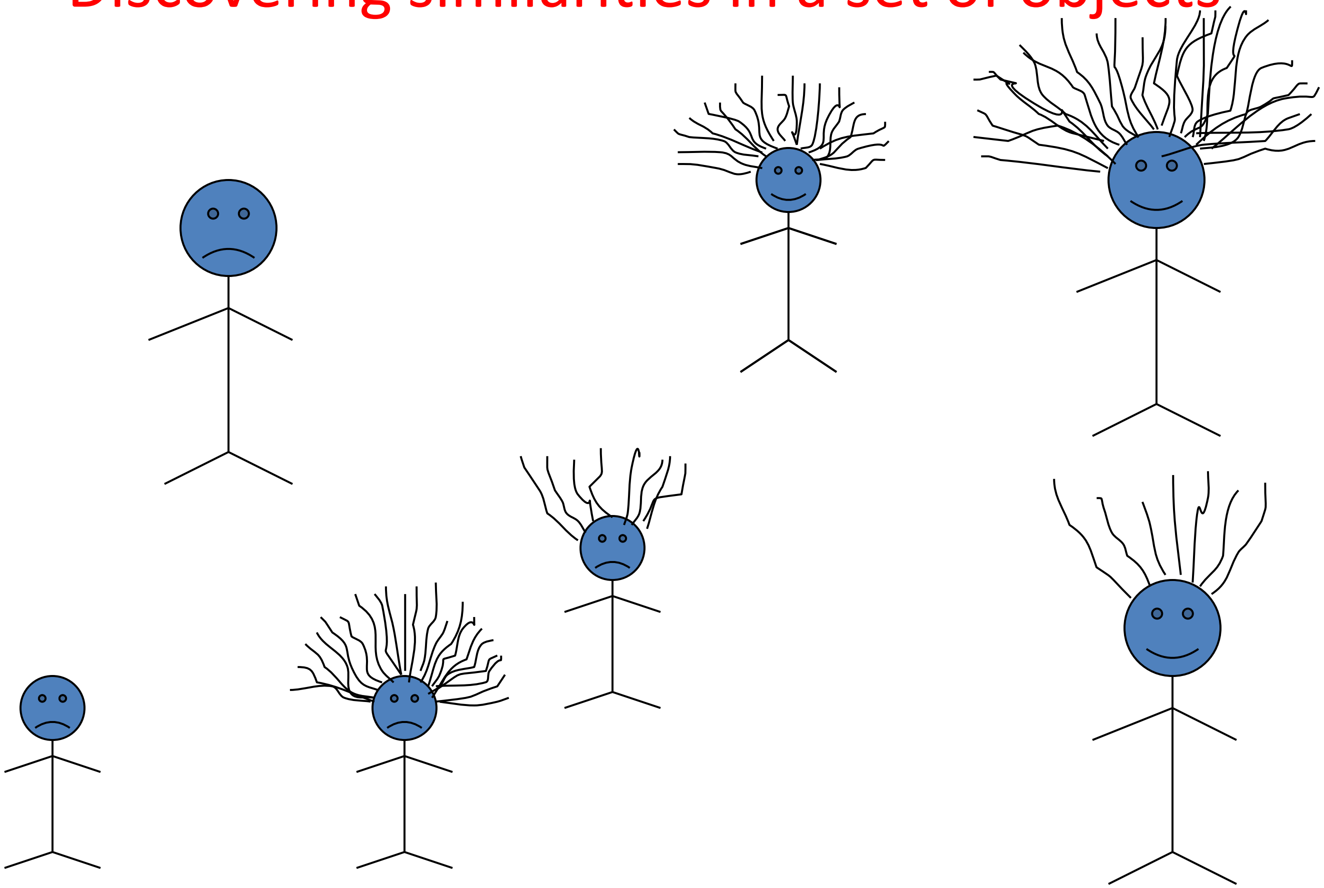
What is clustering about?

Discovering similarities in a set of objects



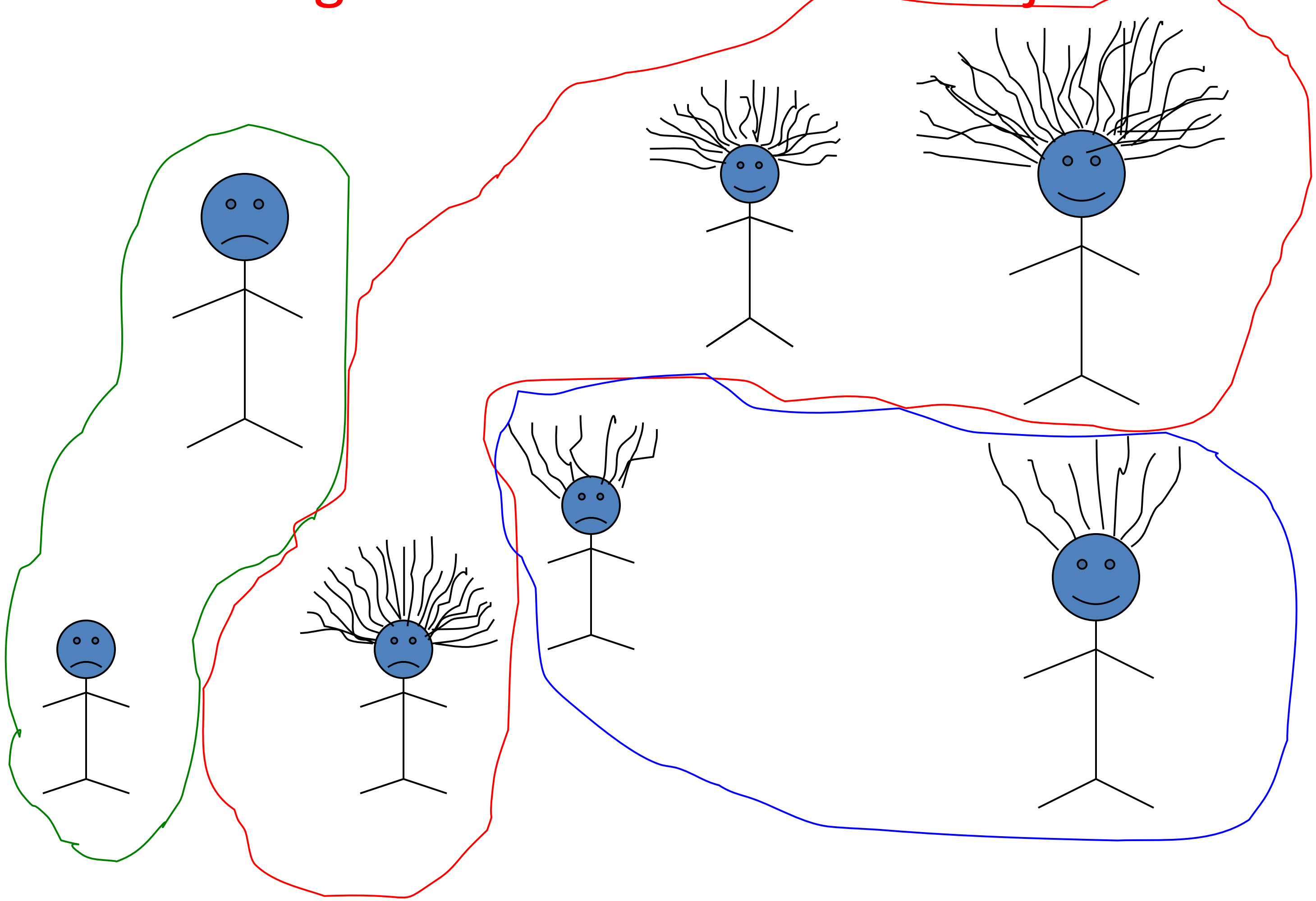
What is clustering about?

Discovering similarities in a set of objects



What is clustering about?

Discovering similarities in a set of objects

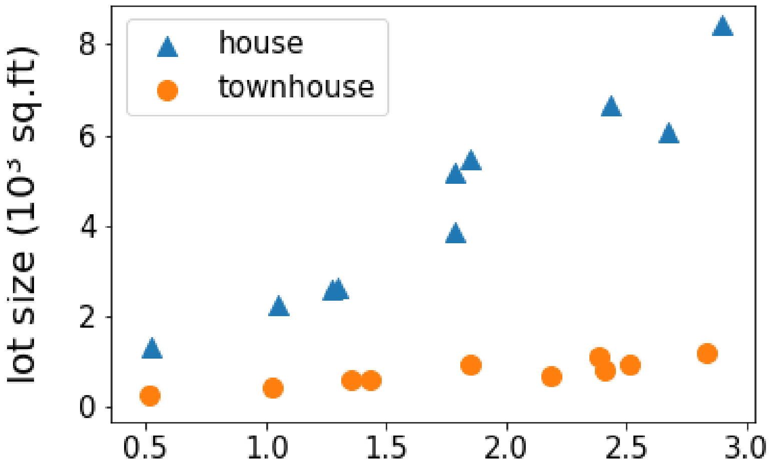


Model Building : Unsupervised Learning

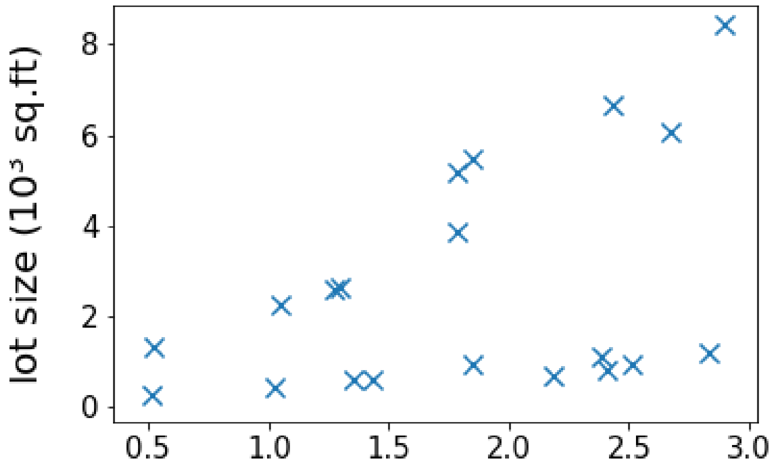
Clustering

- Dataset contains no labels: $x^{(1)}, \ldots x^{(n)}$
- Goal (vaguely-posed): to find interesting structures in the data

supervised



unsupervised



Model Building : Unsupervised Learning

Web Search Results Clustering

Search Results for: Jaguar 1 – 6 of 70,000,000

Clusters

1. Car
2. Animal
3. Mac OS
4. Other

1. Jaguar
Official worldwide web site of Jaguar Cars.

2. Apple - Mac OS X
The Apple Mac OS X product page.

3. Jaguar UK - R is for Racing
The essence of the Jaguar breed

4. Jaguar
General information from Big Cats Online.

5. Jaguar AU - Jaguar Cars
Services and news

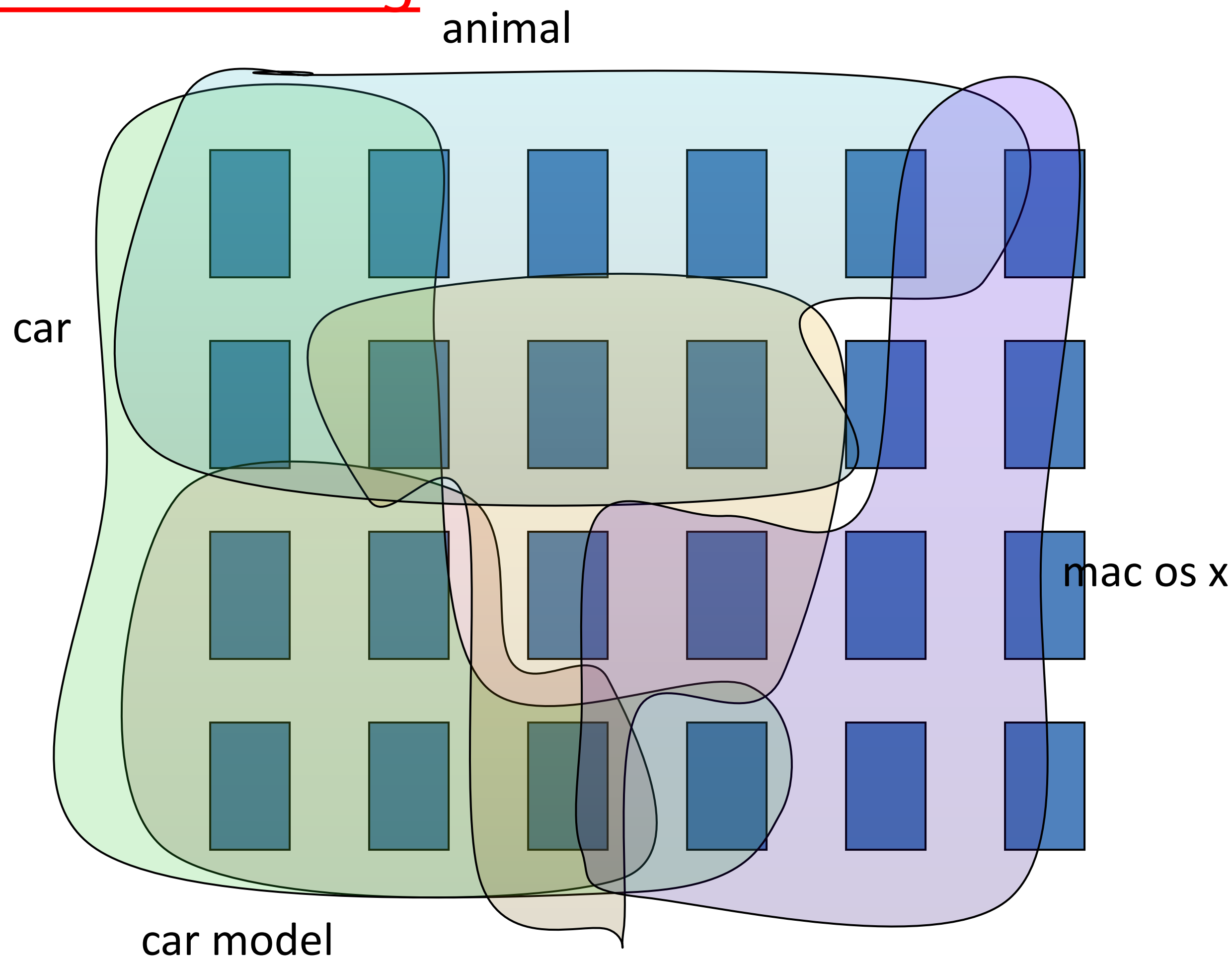
6. Jaguar -- Defenders of Wildlife
Size, appearance, life span and diet.

Model Building : Unsupervised Learning

Web Search Results Clustering

- Query

"jaguar"



Model Building : Unsupervised Learning

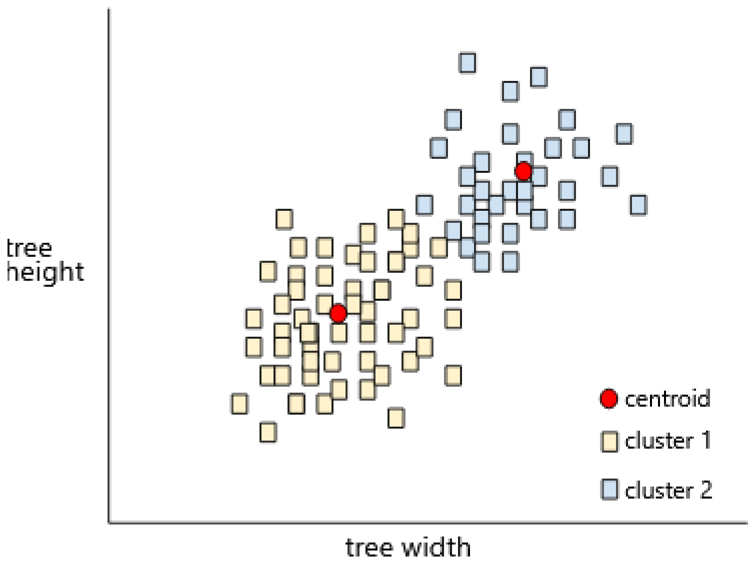
Traditional Clustering Algorithm

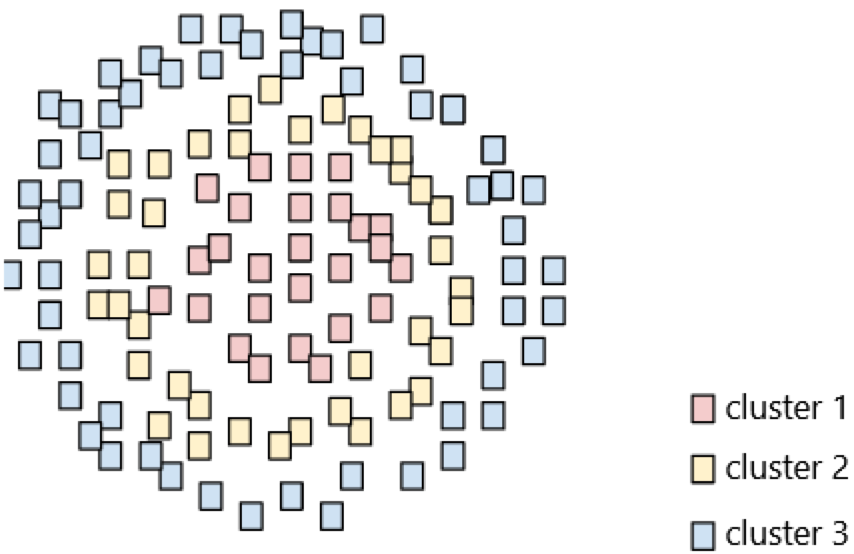
- Distance-based
- Hierarchical
    - Agglomerative Hierarchical Clustering (AHC)
- Flat
    - *K*-means (can be fuzzy)
    - Single-pass (incremental)

Model Building : Unsupervised Learning

K-Means Clustering

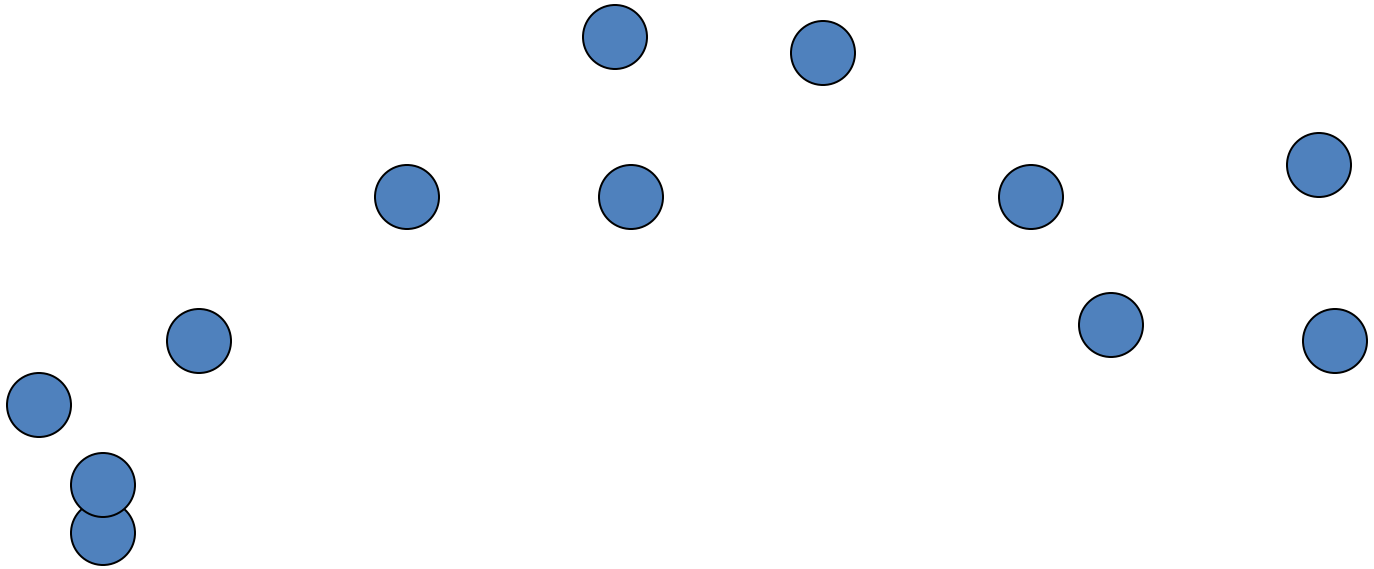
- For example, the **k-means** algorithm clusters examples based on their proximity to a **centroid**, as in the following diagram:





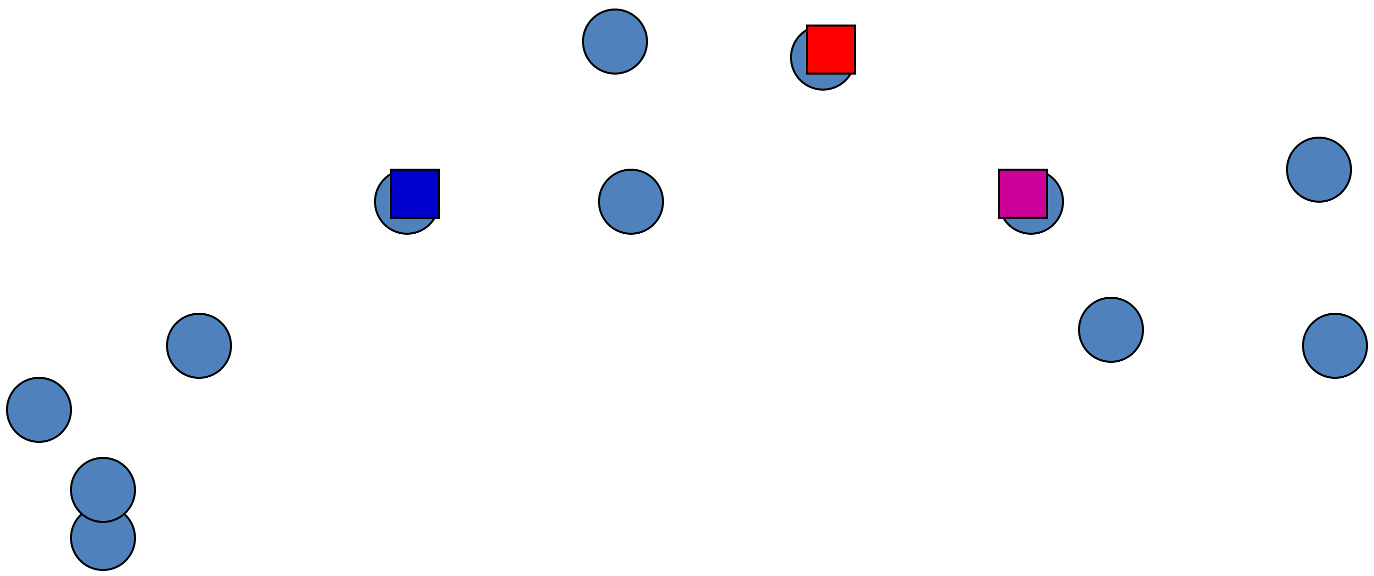
Model Building : Unsupervised Learning

K-Means: an example



Model Building : Unsupervised Learning

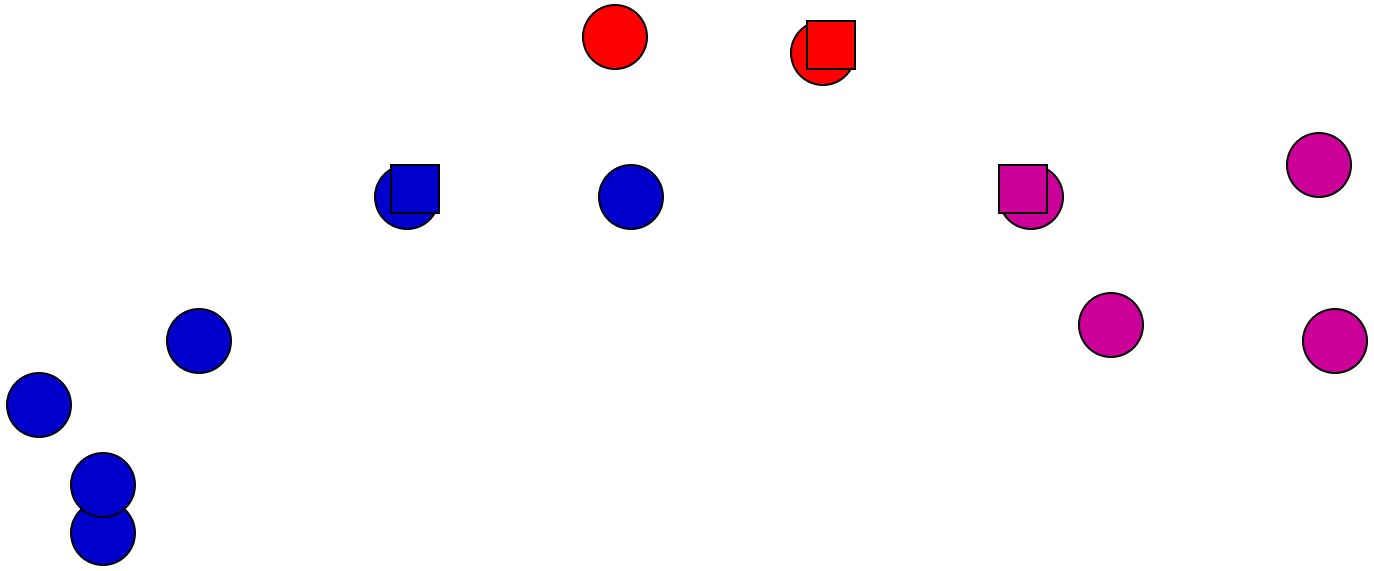
K-Means: an example



K-means: Initialize centers randomly

Model Building : Unsupervised Learning

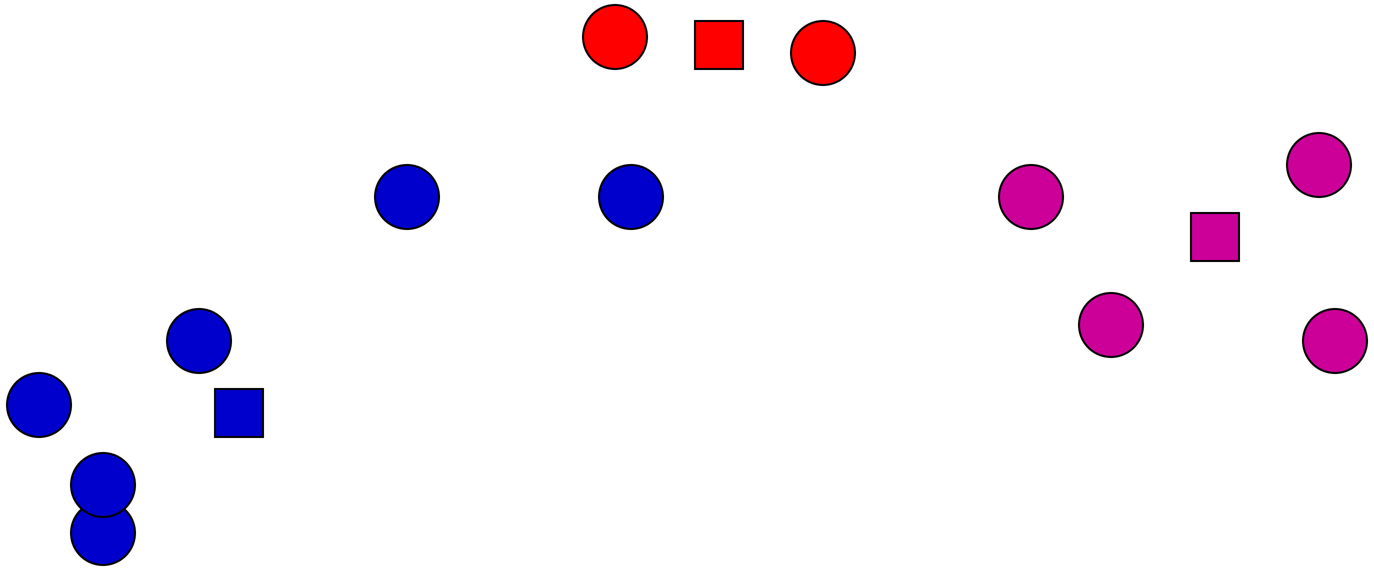
K-Means: an example



K-means: assign points to nearest center

Model Building : Unsupervised Learning

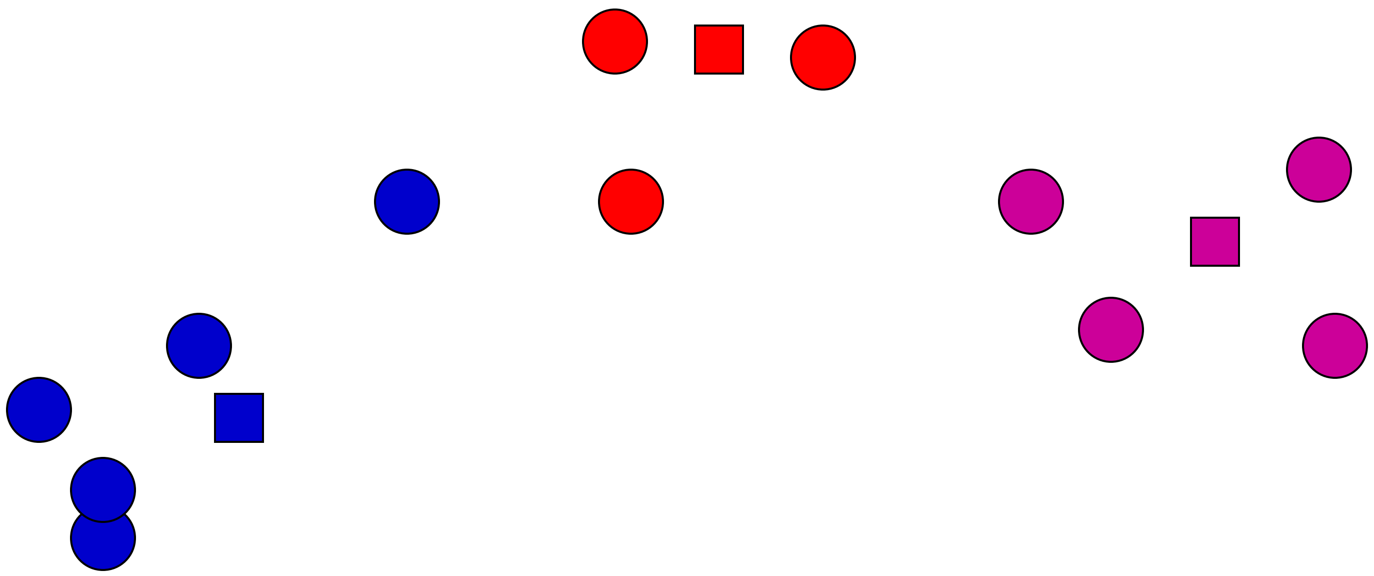
K-Means: an example



K-means: readjust centers

Model Building : Unsupervised Learning

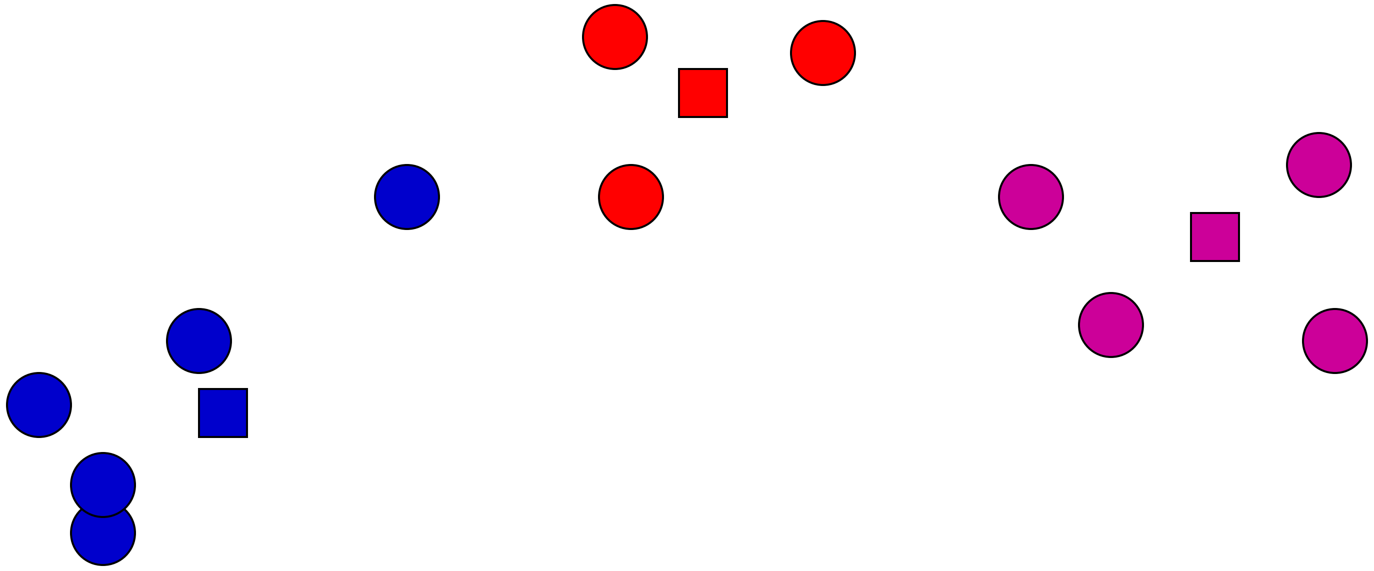
K-Means: an example



K-means: assign points to nearest center

Model Building : Unsupervised Learning

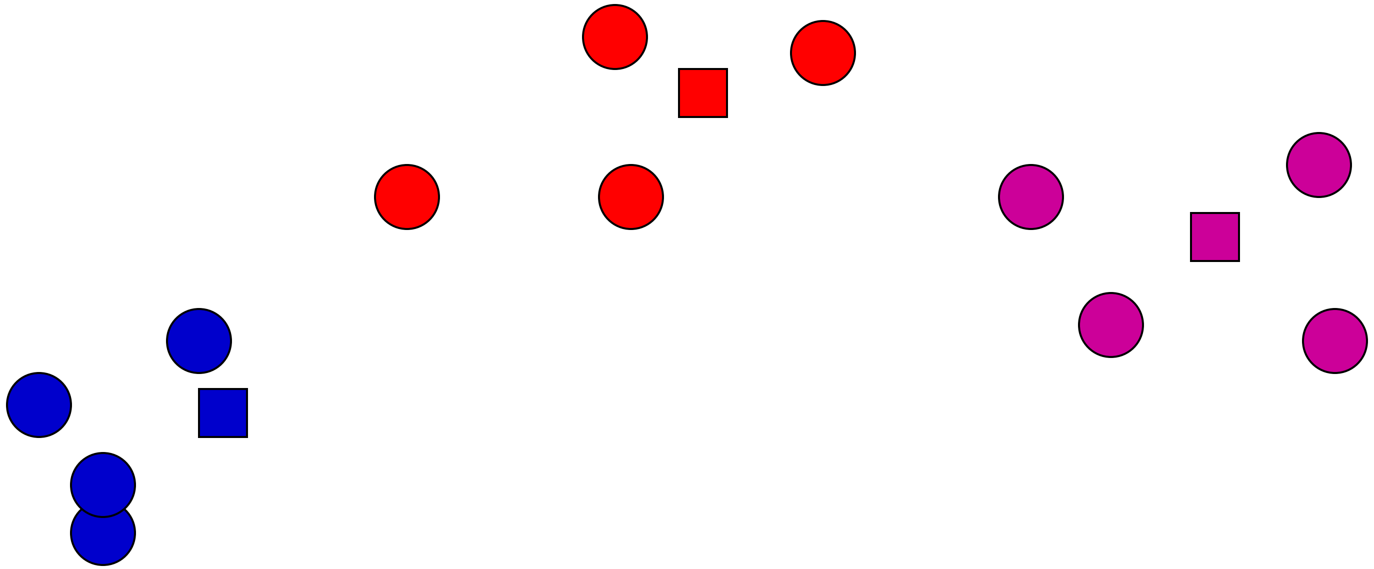
K-Means: an example



K-means: readjust centers

Model Building : Unsupervised Learning

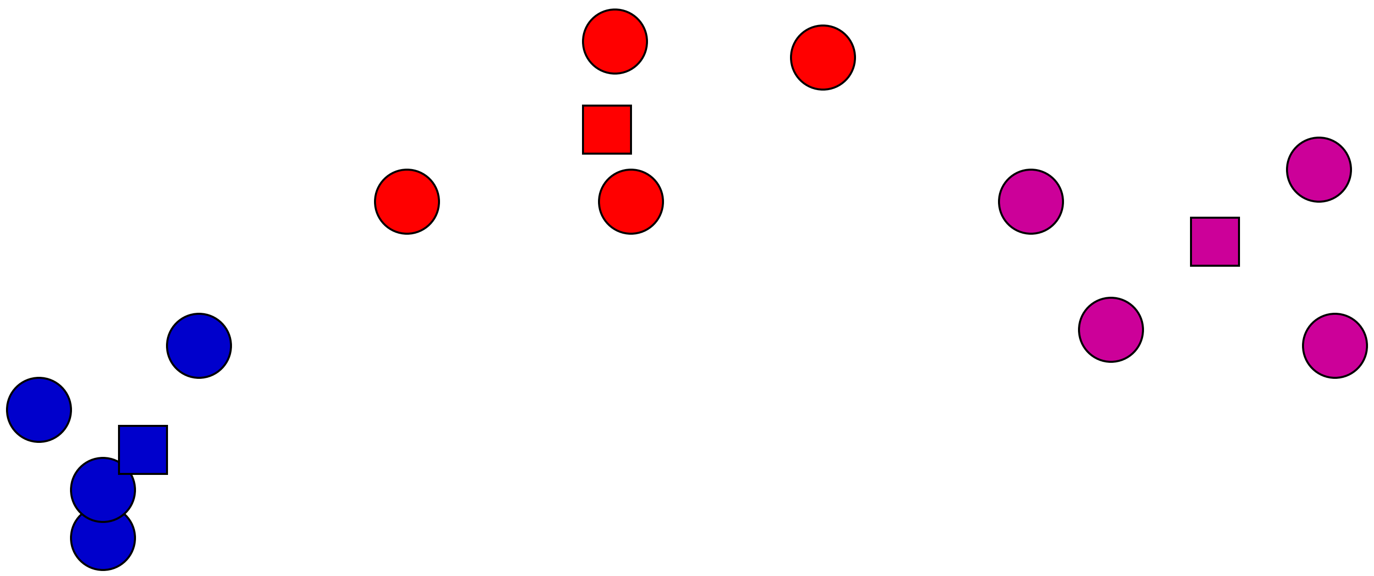
K-Means: an example



K-means: assign points to nearest center

Model Building : Unsupervised Learning

K-Means: an example

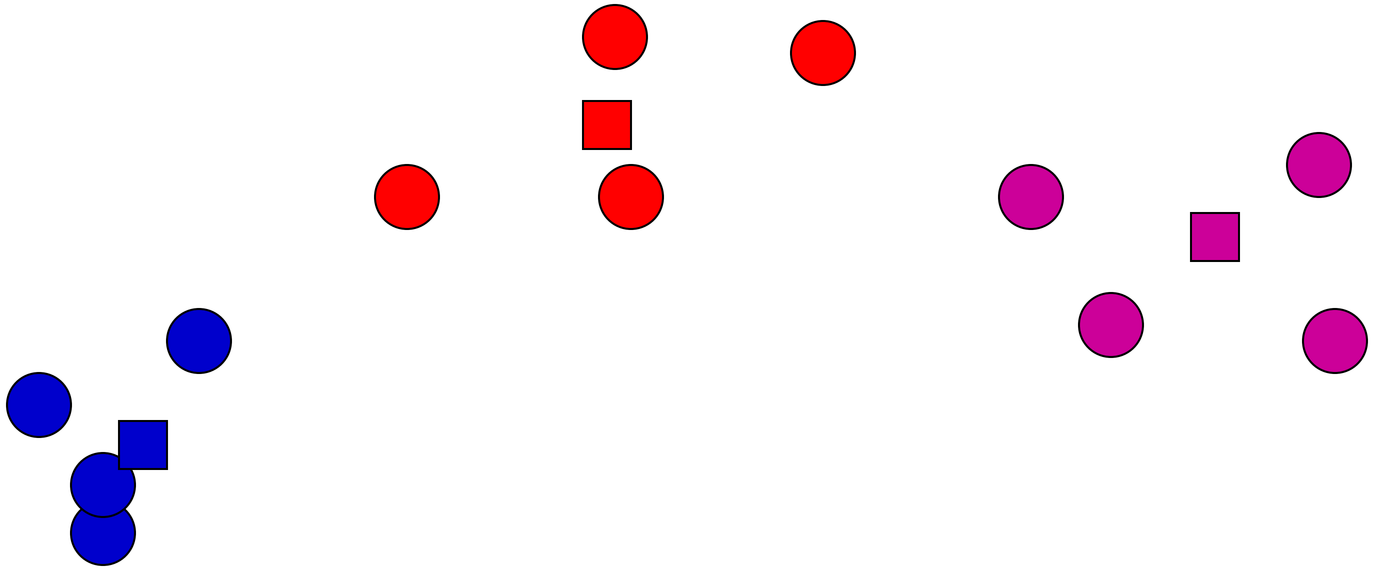


K-means: readjust centers

Model Building : Unsupervised Learning

K-Means: an example

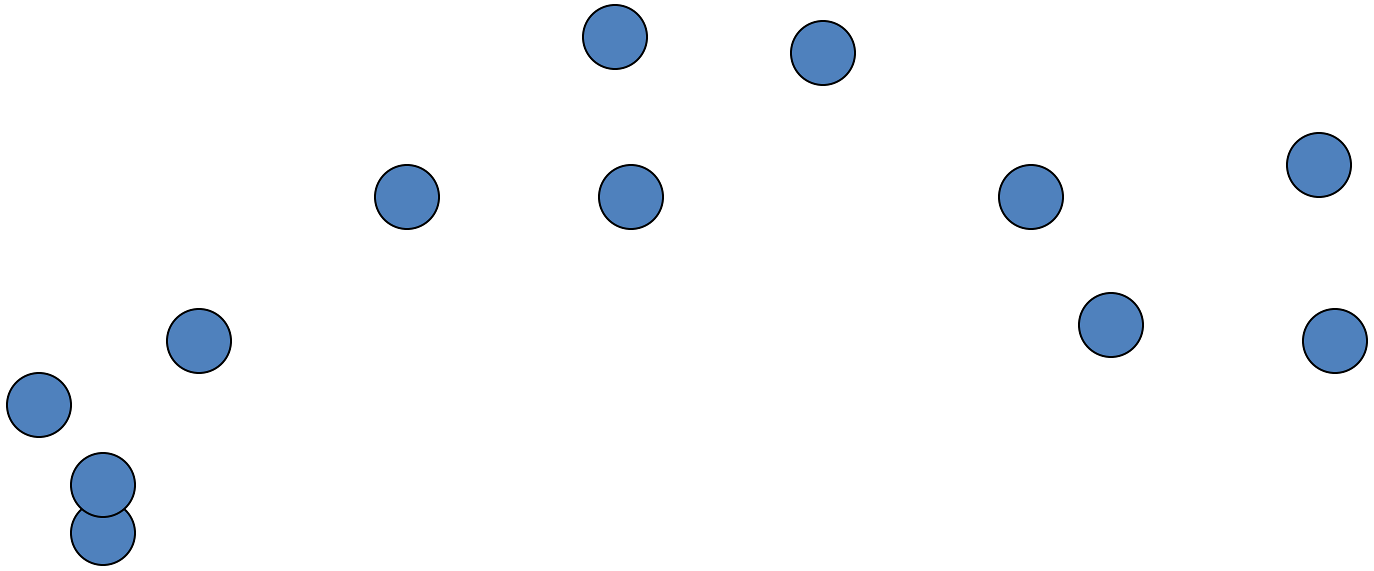
K-means: assign points to nearest center



No changes: Done

Model Building : Unsupervised Learning

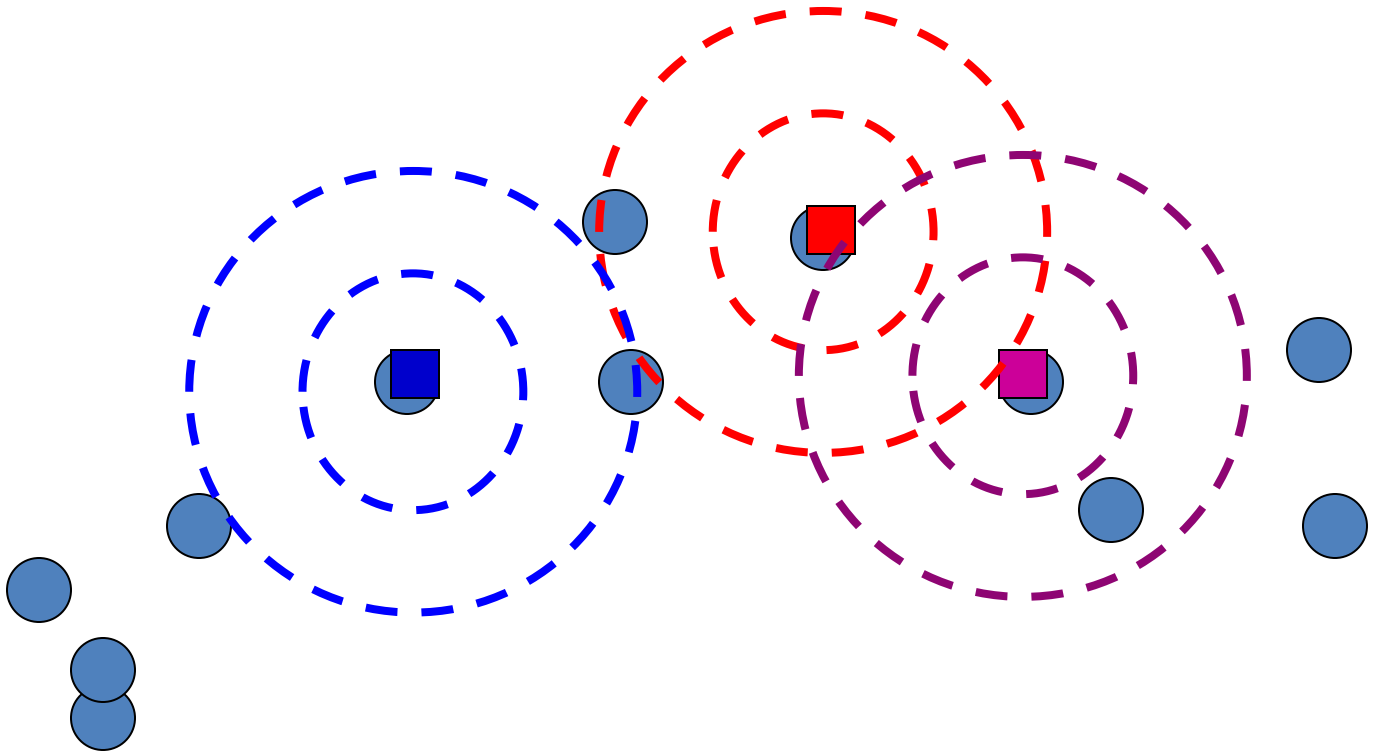
K-Means: an example



K-means: another view

Model Building : Unsupervised Learning

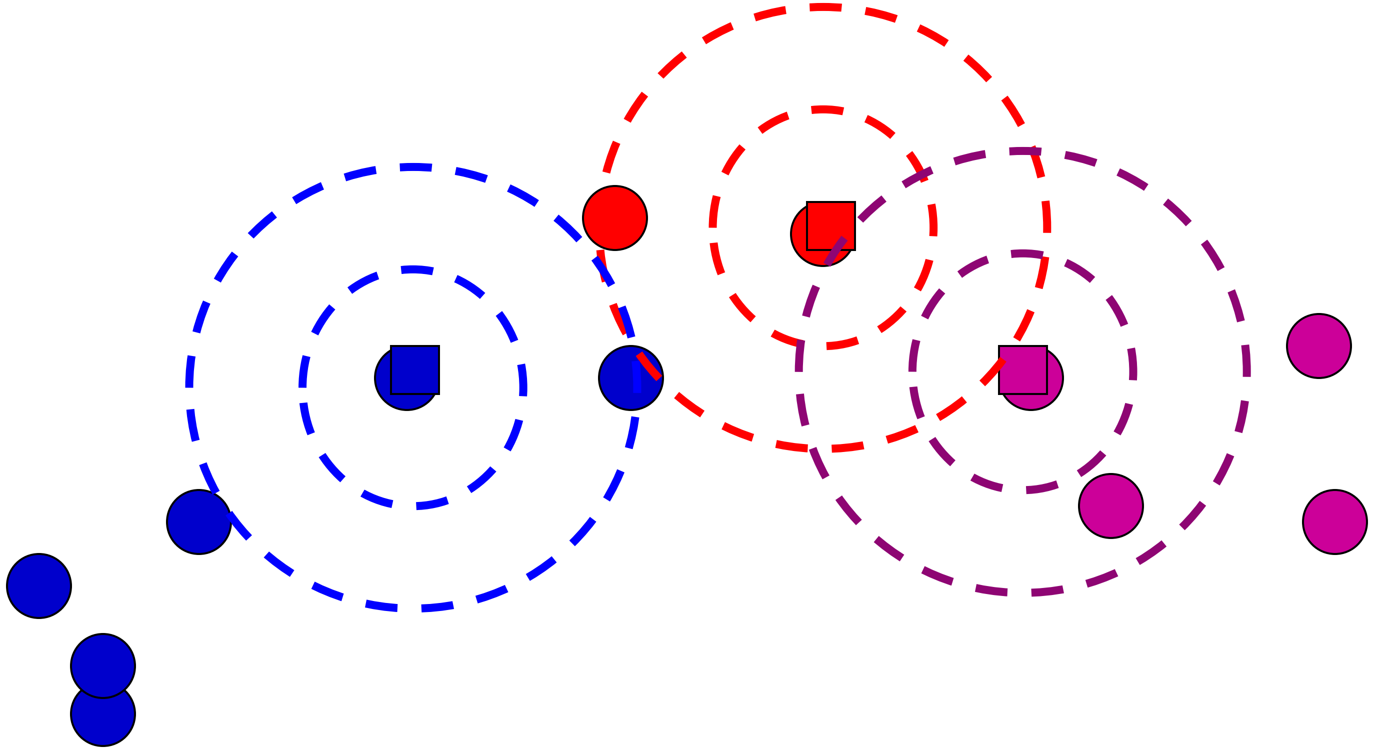
K-Means: an example



K-means: another view

Model Building : Unsupervised Learning

K-Means: an example



K-means: assign points to nearest center

K-Means

- The MacQueen k-means algorithm (MacQueen, 1967) aims to separate *n* objects in *k* non-overlapping groups as to minimize the sum of squared errors (i.e. the sum of distances between the points and the center of their group).

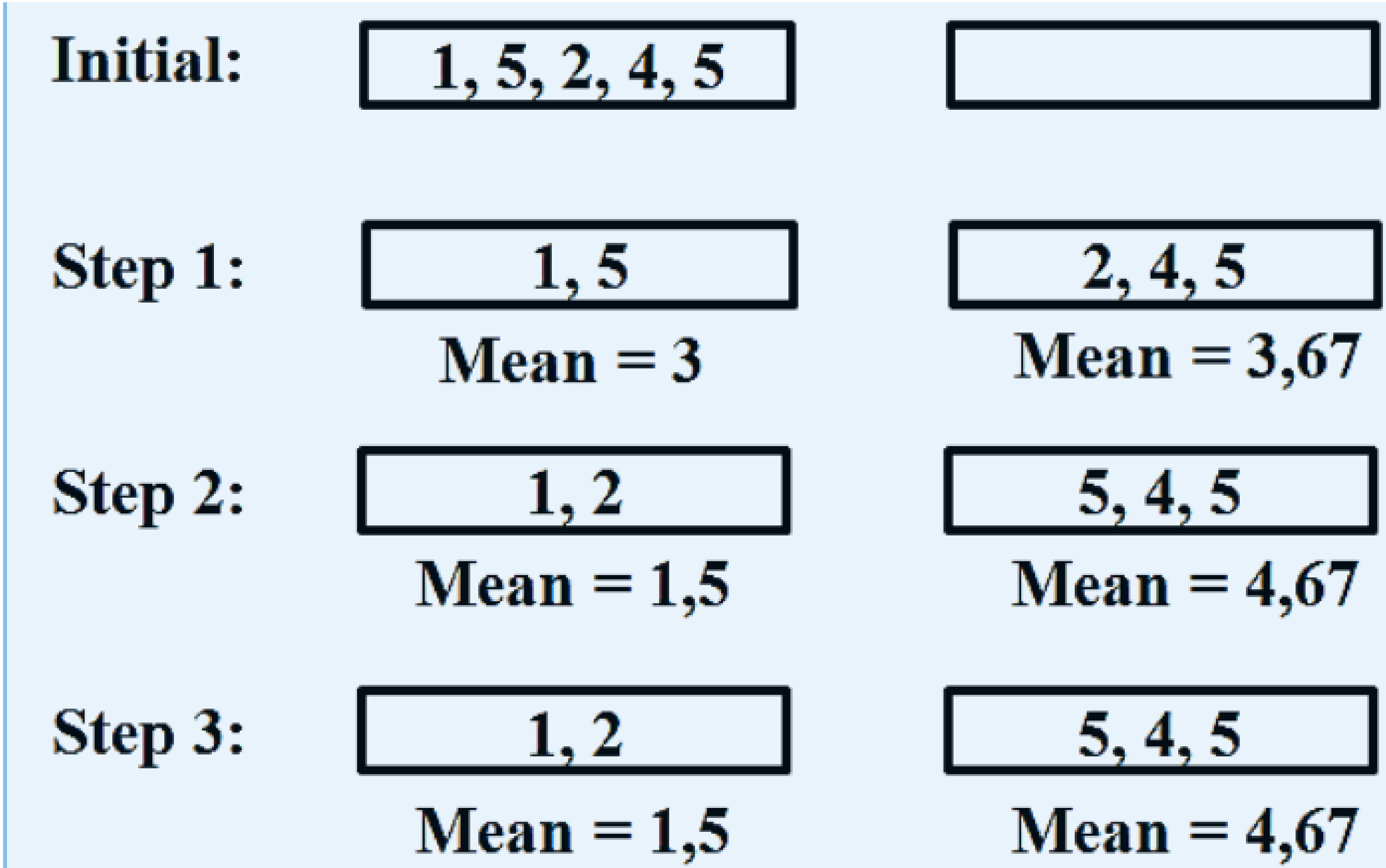
Model Building : Unsupervised Learning

K-Means Clustering

Input : K number of clusters, D: data set containing object

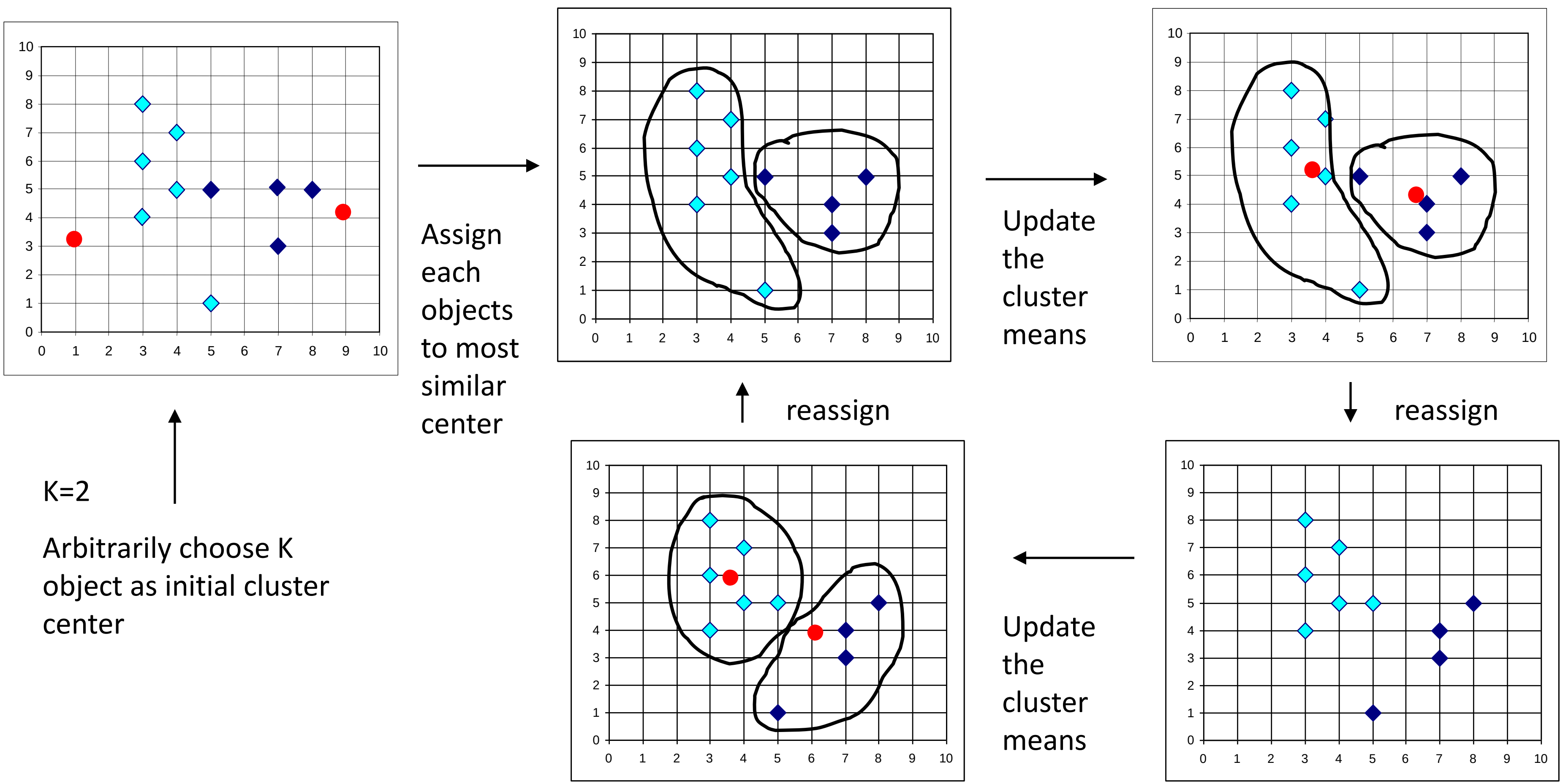
Output: Dataset containing n object

- Given *k*, the *k-means* algorithm is implemented in four steps:
  - **Partition** objects into *k* nonempty subsets
  - **Compute seed points** as the centroids of the clusters of the current partition (the centroid is the center, i.e., ***mean*** *point*, of the cluster)
  - Assign each object to the cluster with the **nearest** seed point
  - Go back to Step 2, stop when no more new assignment


Initial:
1, 5, 2, 4, 5
Step 1:
1, 5
2, 4, 5
Mean = 3
Mean = 3,67
Step 2:
1, 2
5, 4, 5
Mean = 1,5
Mean = 4,67
Step 3:
1, 2
5, 4, 5
Mean = 1,5
Mean = 4,67

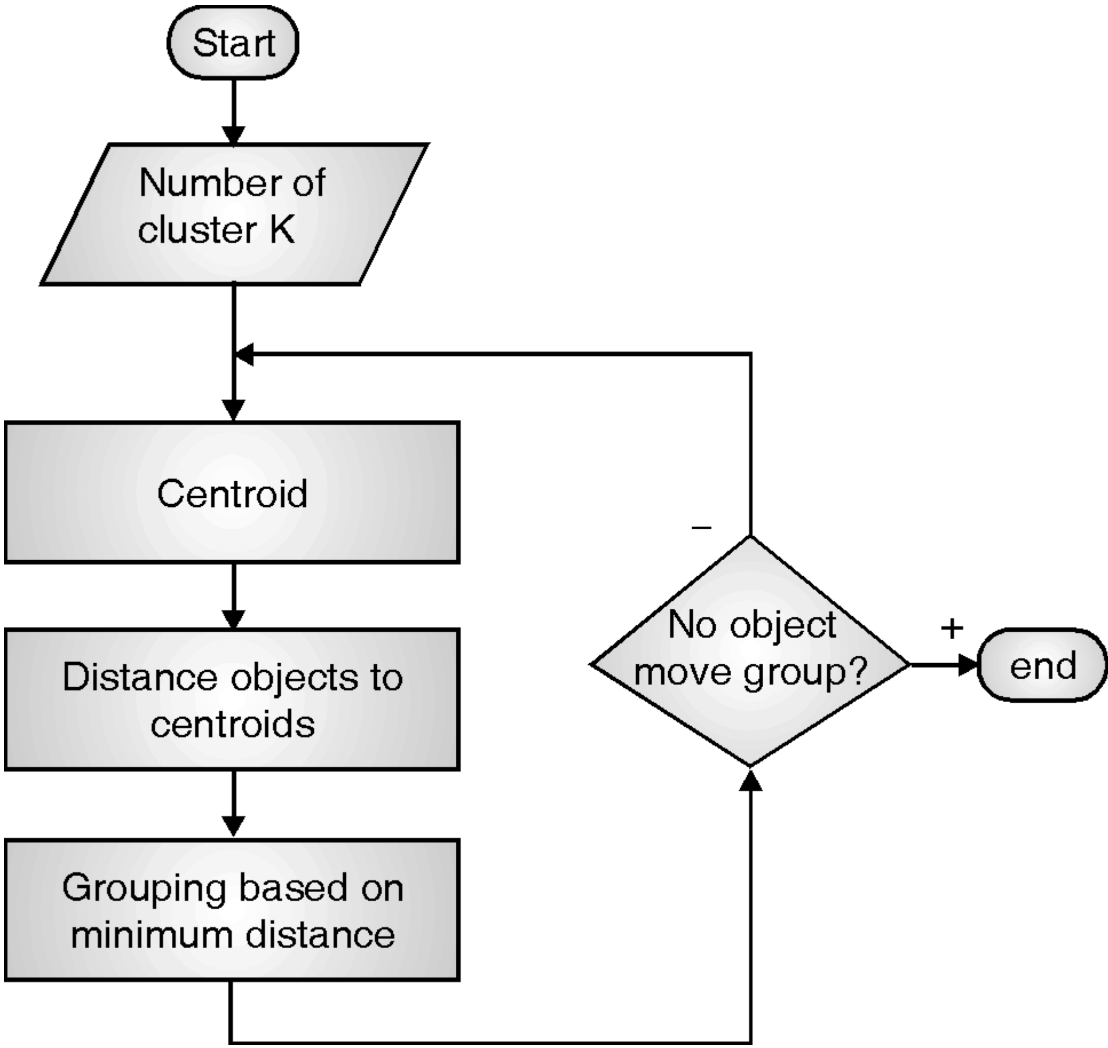
Model Building : Unsupervised Learning

K-Means: an example



Model Building : Unsupervised Learning

K-Means: an example



Model Building : Unsupervised Learning

Variation of the K-Means Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data:
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

mean

The mean is the average or norm.

- Add up all of the values to find a total.
- Divide the total by the number of values you added together.

$\underbrace{2 + 2 + 3 + 5 + 5 + 7 + 8}_{\text{There are 7 values}} = \boxed{32}$ Divide the total by 7

$32 \div 7 = 4.57$

The mean is 4.57

median

The median is the middle value.

- Put all of the values into order.
- The median is the middle value.
- If there are two values in the middle, find the mean of these two.

2, 2, 3, **5,** 5, 7, 8

The median is 5

mode

The mode is the most frequent value.

- Count how many of each value appears.
- The mode is the value that appears the most.
- You can have more than one mode.

2, 2, 3, 5, 5, 7, 8

2 5

Model Building : Unsupervised Learning

Given : {2,4,10,12,3,20,30,11,25}

Assume number of cluster i.e. k = 2.

Randomly assign means: m_1= 3, m_2 = 4

- 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
- K_1={2,3,4,10,11,12},K_2={20,30,25}

Model Building : Unsupervised Learning

Randomly assign alternative values to each cluster

Number of cluster = 2, therefore

K1 = {2,10,3,30, 25}, Mean = 14

K2 = {4,12, 20, 11}, Mean = 11.75

Re-assign

K1 = {20, 30, 25}, Mean = 25

K2 = {2,4, 10, 12, 3, 11}, Mean= 7

Re-assign

K1 = {20, 30, 25}, Mean = 25

K2 = {2,4, 10, 12, 3, 11}, Mean= 7

So the final answer is K_1={2,3,4,10,11,12},K_2={20,30,25}

Model Building : Unsupervised Learning

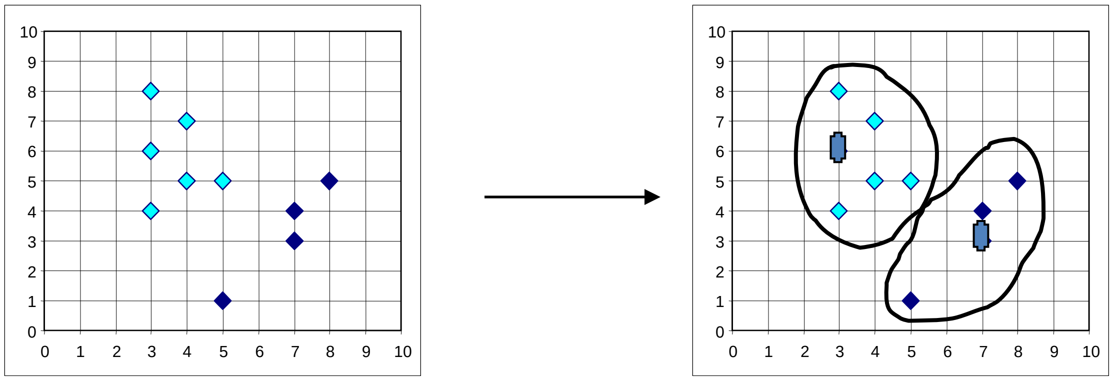
K-Means

- Strength: *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.
- Comment: Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
  – Applicable only when *mean* is defined, then what about categorical data?
  – Need to specify *k,* the *number* of clusters, in advance
  – Unable to handle noisy data and *outliers*
  – Not suitable to discover clusters with *non-convex shapes*

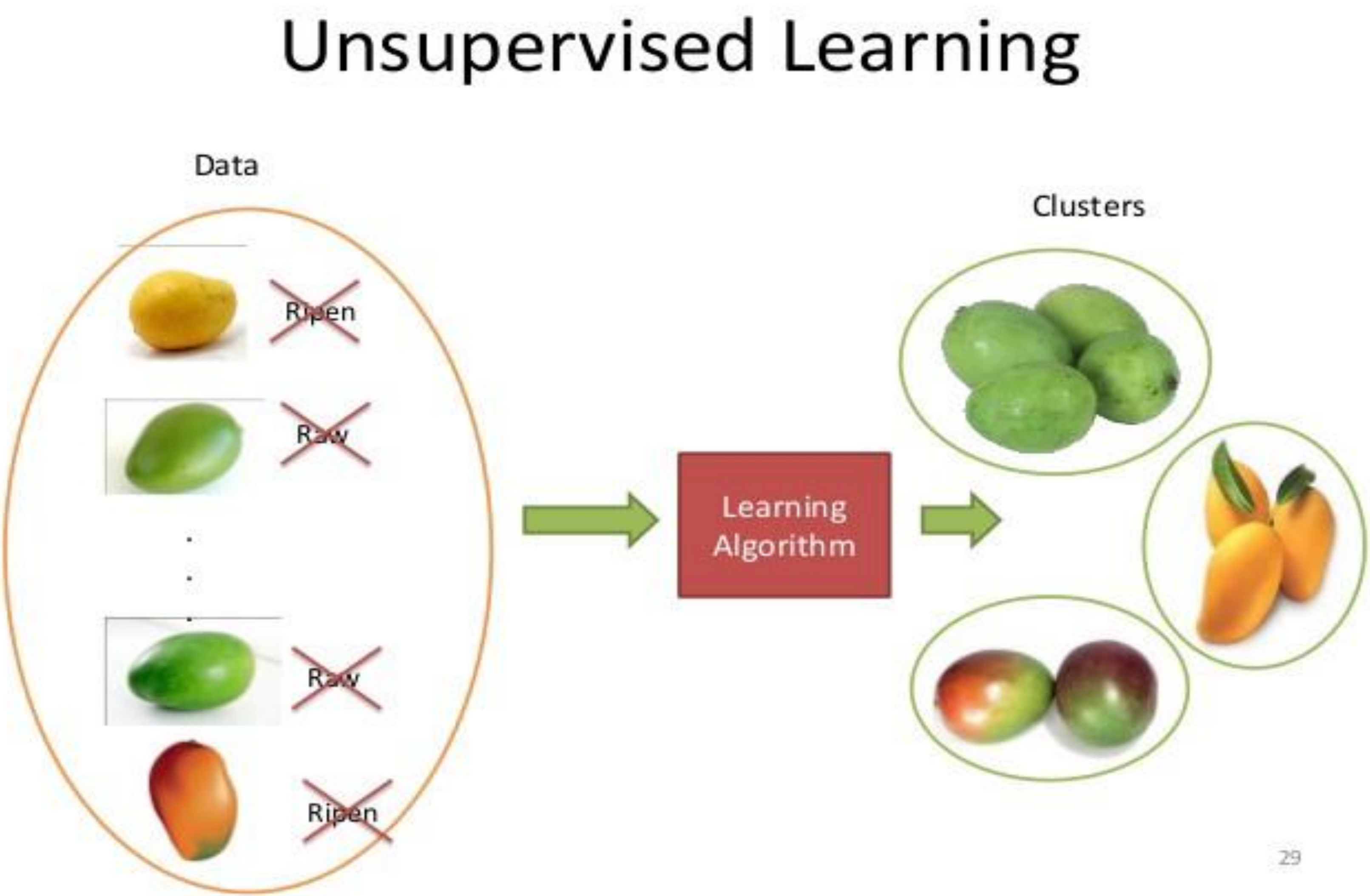
Comments on the *K-Means* Method

What Is the Problem of the K-Means Method?

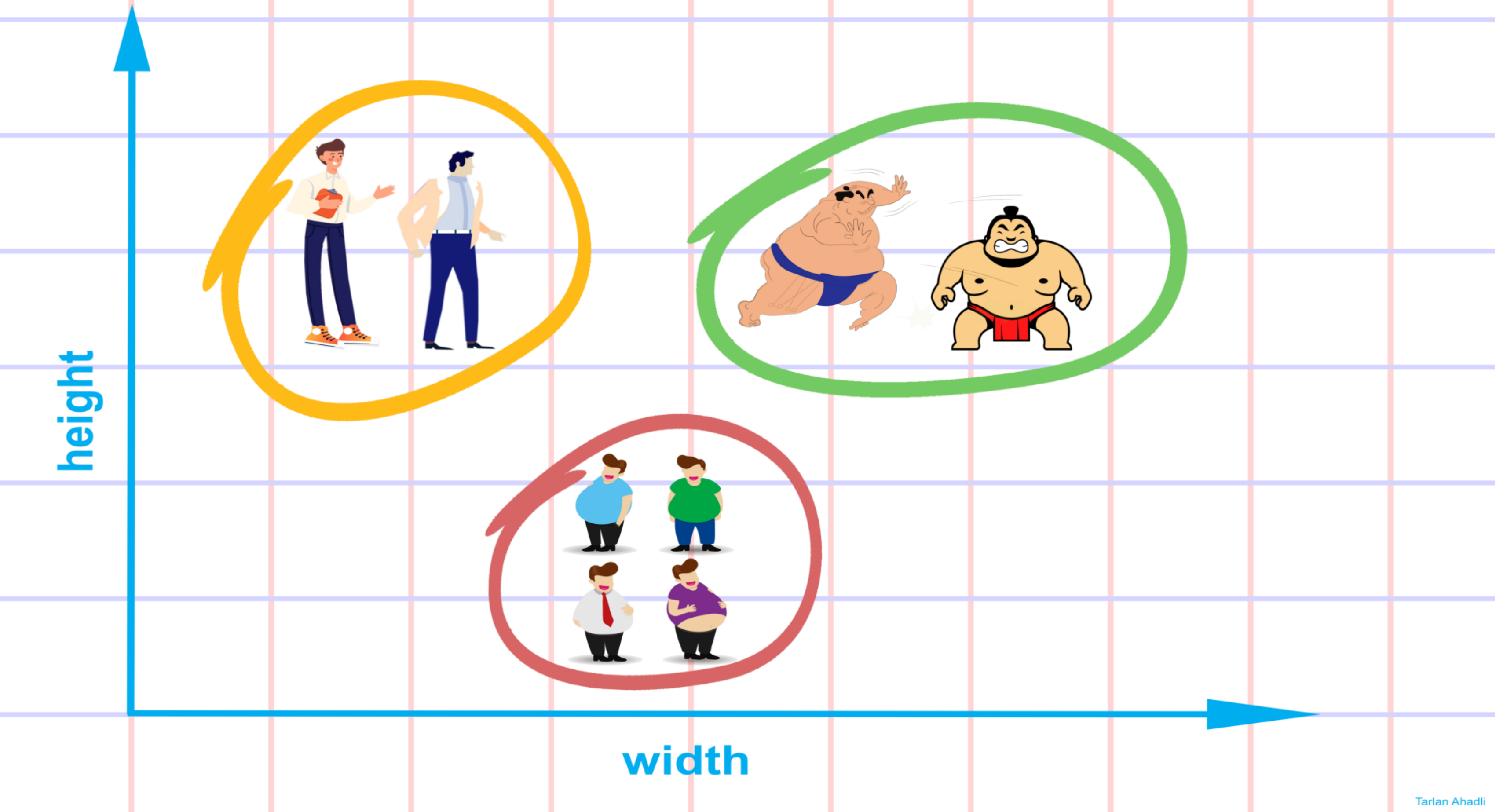
- The k-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data.
- K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



Clustering


Unsupervised Learning
Data
Ripen
Raw
Raw
Ripen
Learning Algorithm
Clusters

K-MEANS CLUSTERING ALGORITHM



Clustering

- Pros
  - It can detect what human eyes can not understand
  - The potential of hidden patterns can be very powerful for the business or even detect extremely amazing facts, fraud detection etc.
  - Output can determine the un explored territories and new ventures for businesses. Exploratory analytics can be applied to understand the financial, business and operational drivers behind what happened.
- Cons
  - As seen in above explanation unsupervised learning is harder as compared to supervised learning.
  - It can be a costly affair, as we might need external expert look at the results for some time.
  - Usefulness of the results; are of any value or not is difficult to confirm since no answer labels are available.