# Unsupervised ML

Institute of Technology of Cambodia

October 30, 2023

# Objective

- Discover information from data without labeled examples
- Extract some hidden organisation, patterns, relation between element
- There are three cases in unsupervised learning:
  - Clustering
  - Dimensional reduction
  - Association rule

# Unsupervised learning

# Unsupervised learning

- Unsupervised learning is a paradigm in machine learning where, in contrast to supervised learning and semi-supervised learning, algorithms learn patterns exclusively from unlabeled data.
- `https://en.wikipedia.org/wiki/Unsupervised_learning`

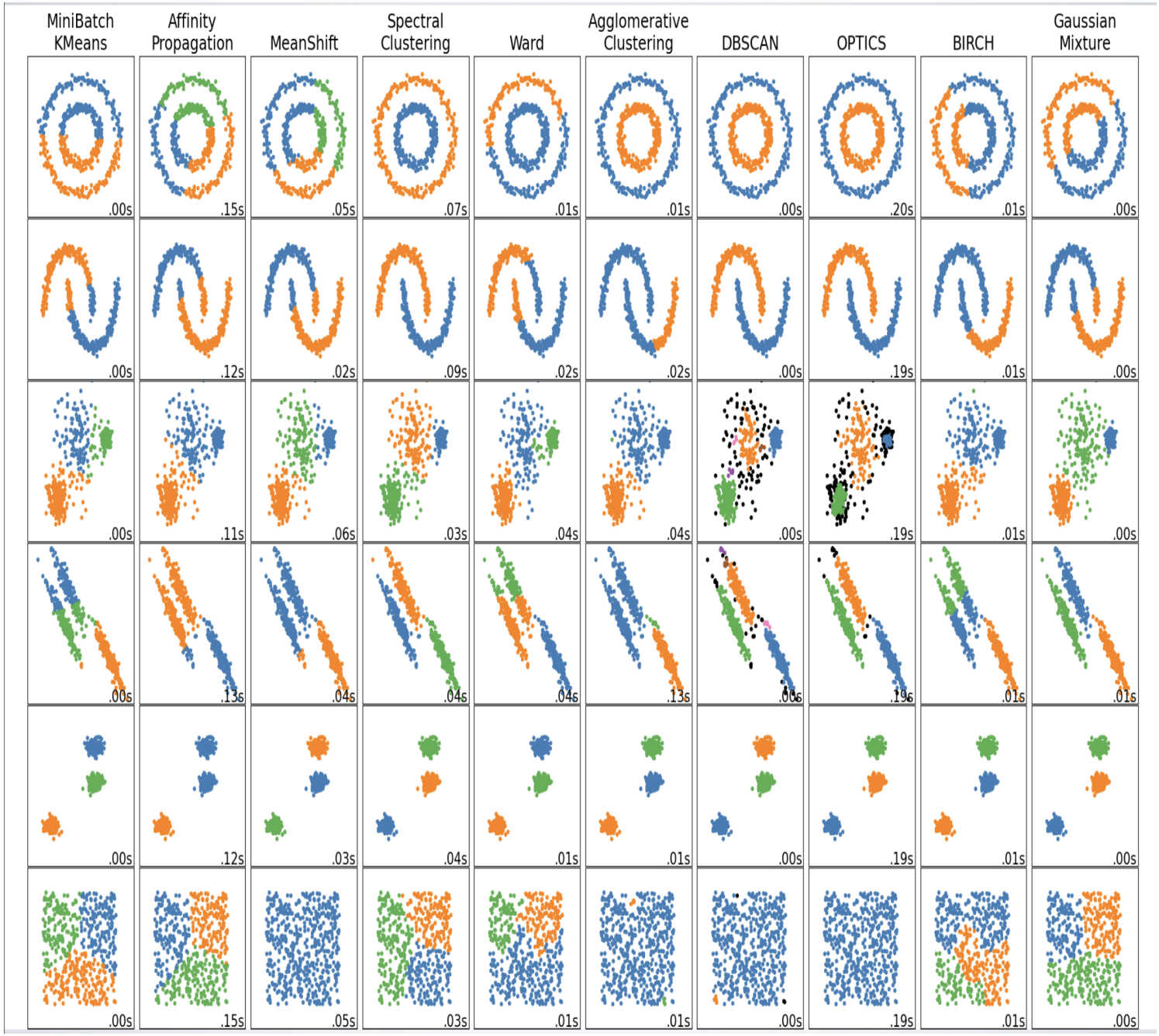
# CLUSTERING

# Clustering

- The most famous unsupervised ML problem
- Most people use "good old" methods: k-means(1967), DBSCAN(1996)
- Part of the problem: clustering is not well defined

# Clustering

- How would you define a good cluster?
- A good partition in clusters?


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .15s .05s .07s .01s .01s .00s .20s .01s .00s
.00s .12s .02s .09s .02s .02s .00s .19s .01s .00s
.00s .11s .06s .03s .04s .04s .00s .19s .01s .00s
.00s .13s .04s .04s .04s .13s .00s .19s .01s .01s
.00s .12s .03s .04s .01s .01s .00s .19s .01s .00s
.00s .15s .05s .03s .01s .01s .00s .19s .01s .00s

# K-Means

Definition

- For a target number of cluster K
- Find the item assignment minimizing
  - The inter-cluster variance (weighted by cluster size)
  - The squared distance from points to their cluster center
  - The squared distance between cluster elements

# K-Means

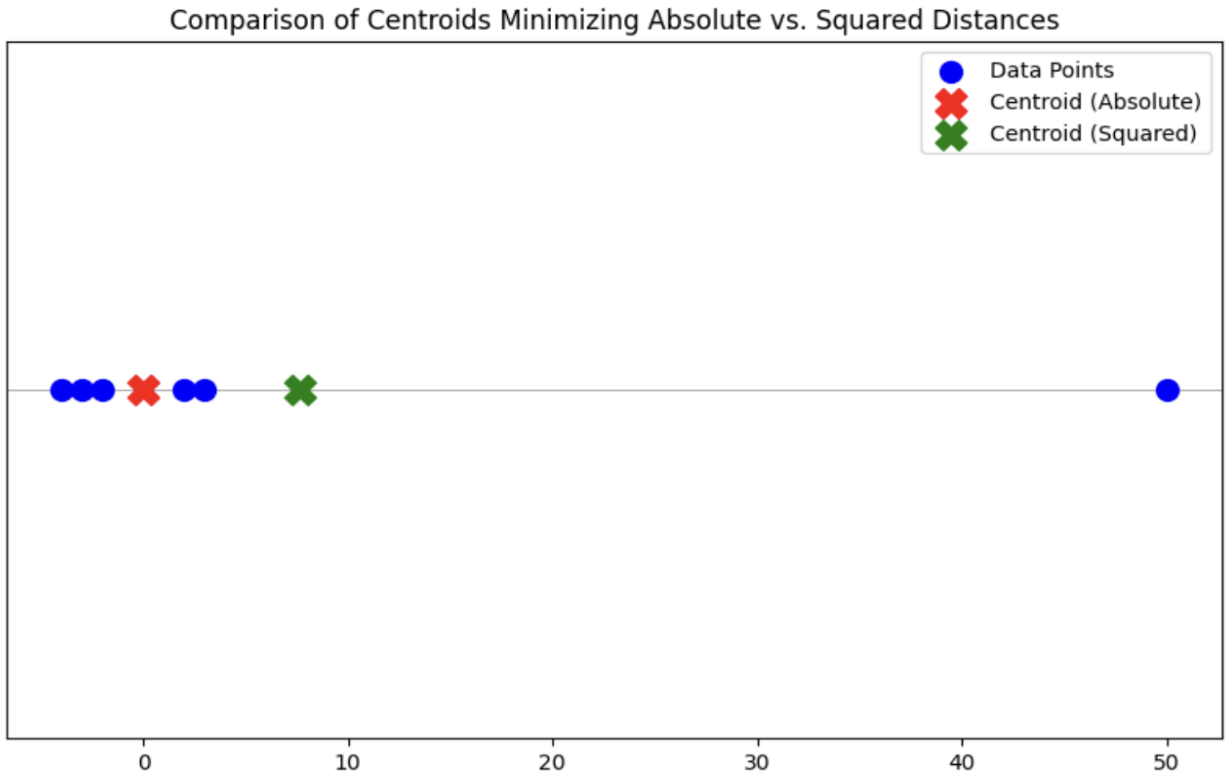
$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \arg\min_{S} \sum_{i=1}^{k} \| S_i Var(S_i)$$

- S a cluster assignment
- k a number of cluster
- x a dimensional item

# K-Means

Consequence: outliers penalized more (pros and cons)

- Squared distance minimized by the **mean**
- Absolute distance minimized by the **median**



# K-Means

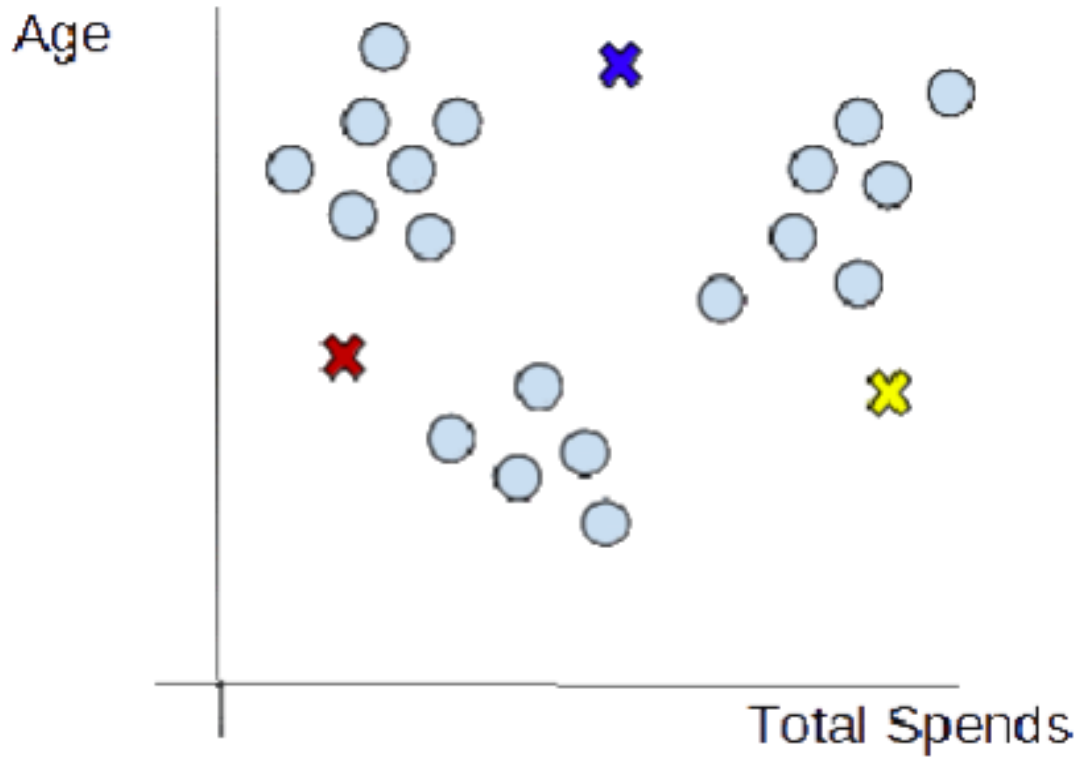
$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \arg\min_{S} \sum_{i=1}^{k} \| S_i Var(S_i)$$

Note that without fixing k, there is a trivial solution with each item alone in its own cluster.

# K-Means

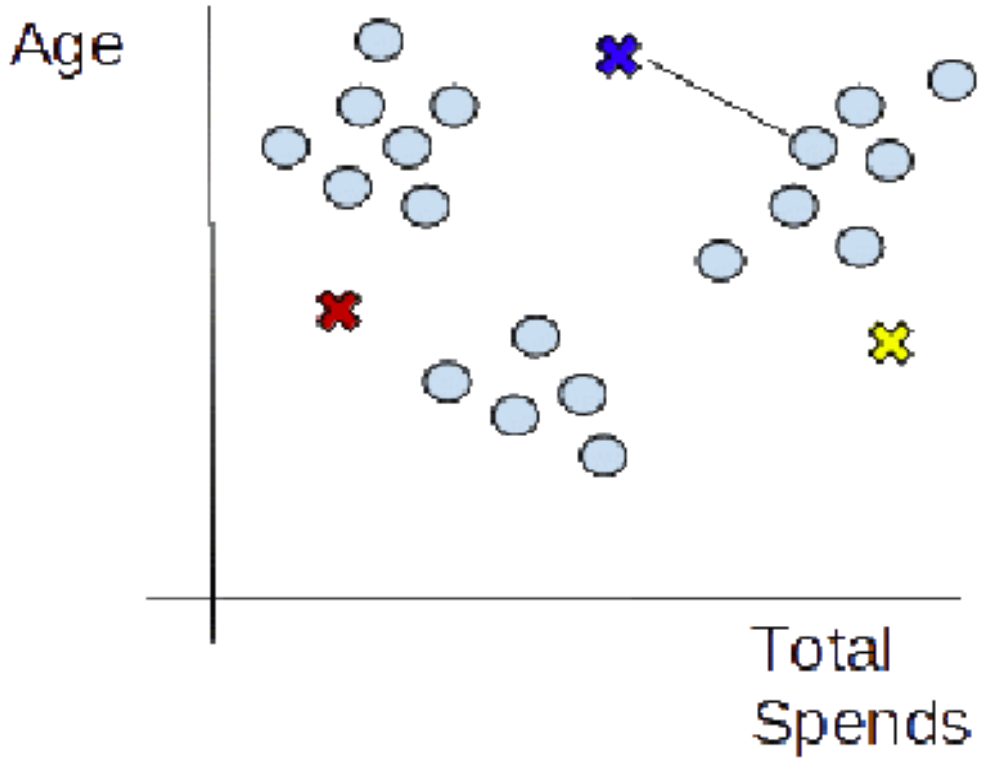
How does K-Means work?

1. Choosing the number of clusters
2. Initializing centroids (the center of a cluster)



# K-Means

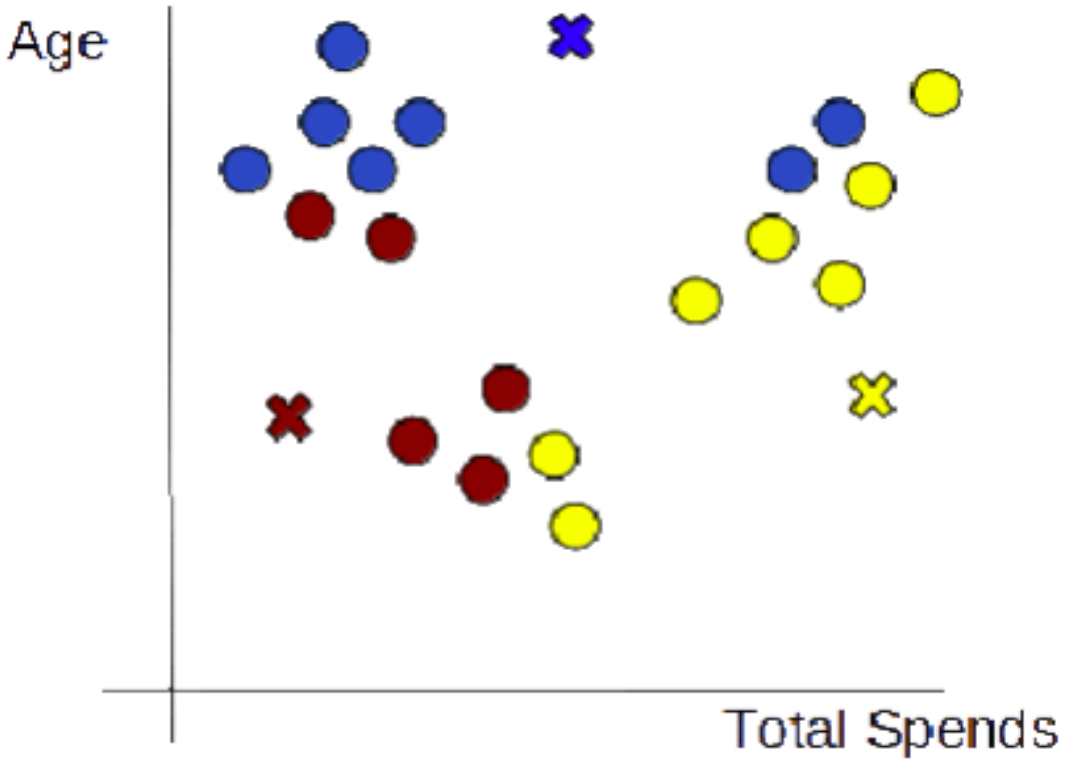
3. Assign data points to the nearest cluster




In this step, we will first calculate the distance between data point X and centroid C using Euclidean Distance metric.

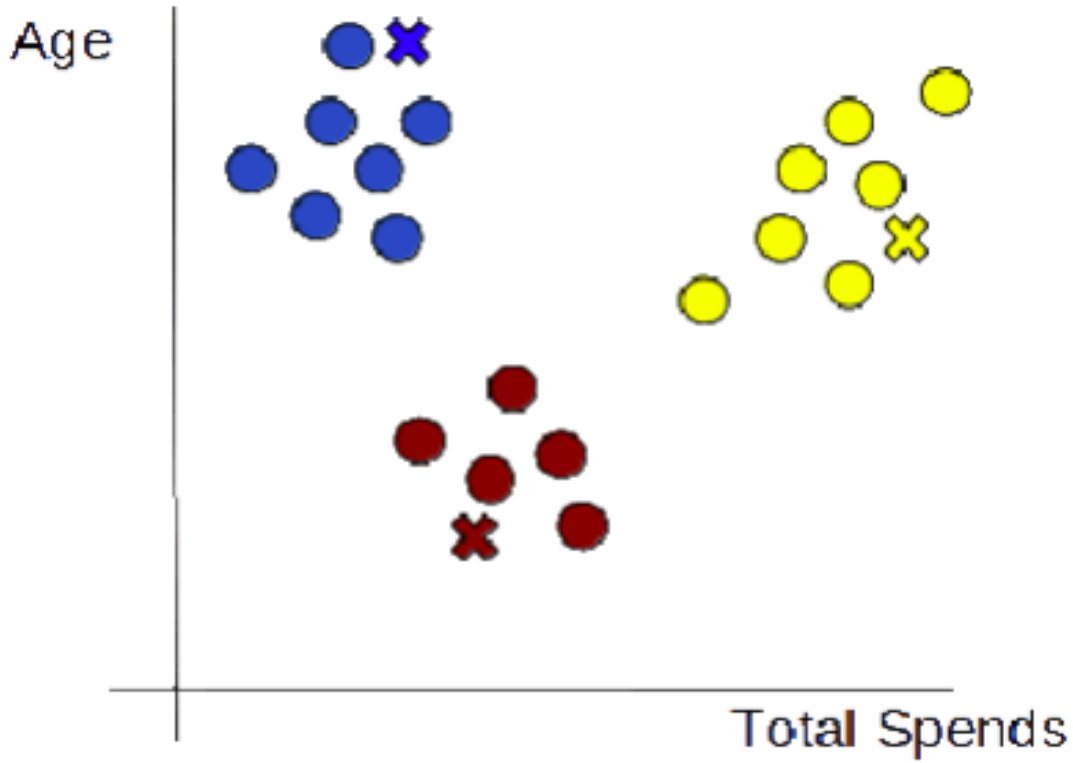
# K-Means

Then choose the cluster for data points where the distance between the data point and the centroid is minimum.



# K-Means

4. Re-initialize centroids
5. Repeat steps 3 and 4



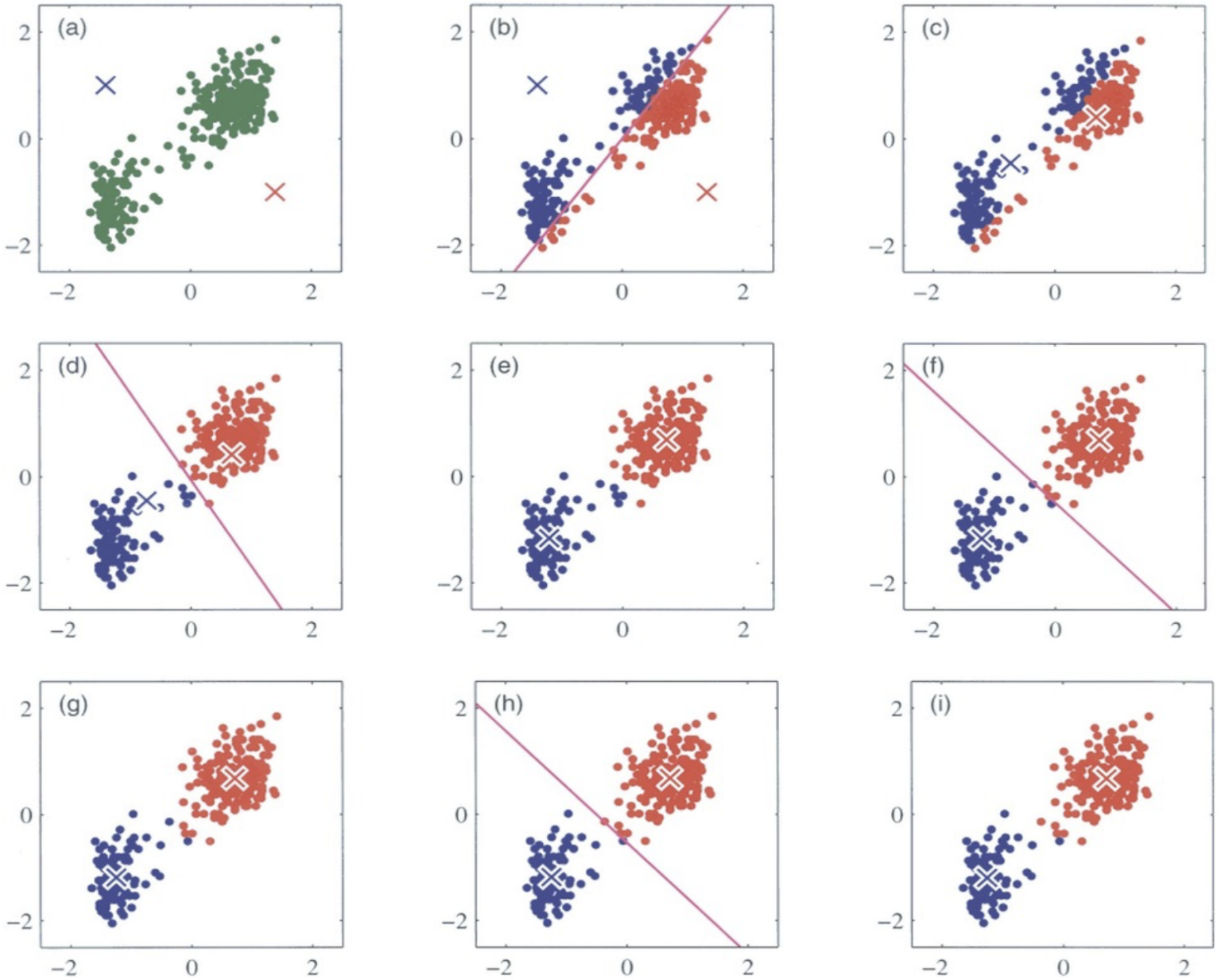
# K-Means

- Discovering the global minimum is NP-hard
- How to find quickly a good solution?
  - Naive k-means
  - K-means ++ (used in most current implementations)
  - Use optimized data structure

# Naive K-Means

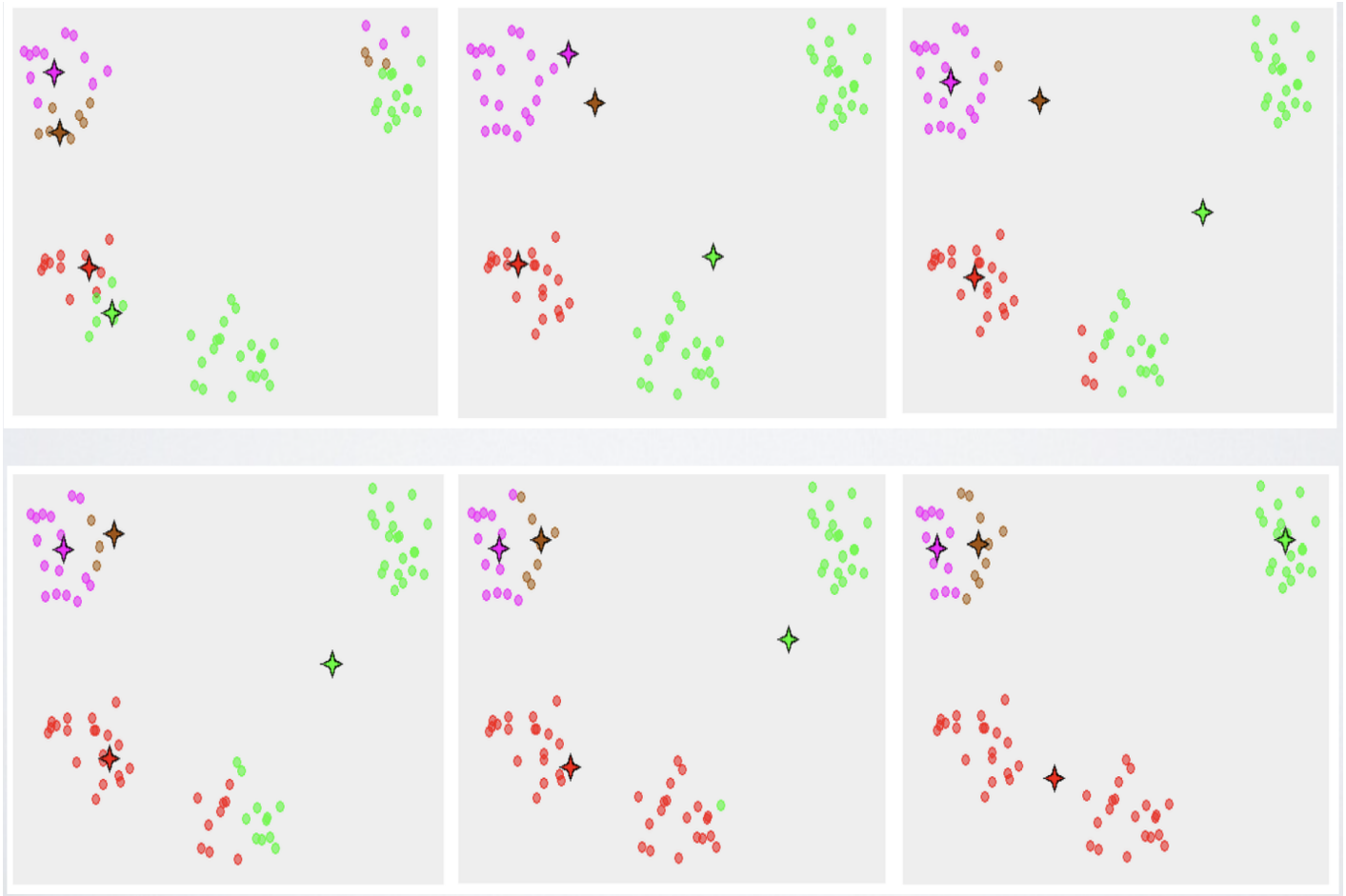
- Assignment: assign each item to its closest cluster center
- Update: recompute the center of each cluster as the mean (centroid) of items that compose that cluster
- Start with random centroid

# Naive K-Means



# Naive K-Means

Known limit: convergence to poor local minimum if poor initial centroids



# K-Means++

- Several variants to choose wisely the initial centroids
- K-means++ is proven to improve the results, statistically.
  - Not always, but improves more often than deteriorate the results.

# K-Means++

1. Choose one center uniformly at random among the data points
2. For each data point x not chosen yet, computer 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.

# Weakness

We can identify some clear weaknesses:

- K-means has a tendency to search for clusters of equal sizes (minimize overall cluster variance)
- Clusters tend to be circular, since all directions are worth the same

# Normalization

Important point: k-means is based on Euclidean distance

- We minimize the inter-cluster Eclidean distance between points
- We could adapt the method to other distances

Data nees to be normalized/standardized

- Clustering based on age in years
- Remember: normalization/standardization are not fixing magically problems (outlier).

# Experiments

- Go to the webpage of the class and do today's experiments
- The "advance" section is not mandatory, you can do it if you have time