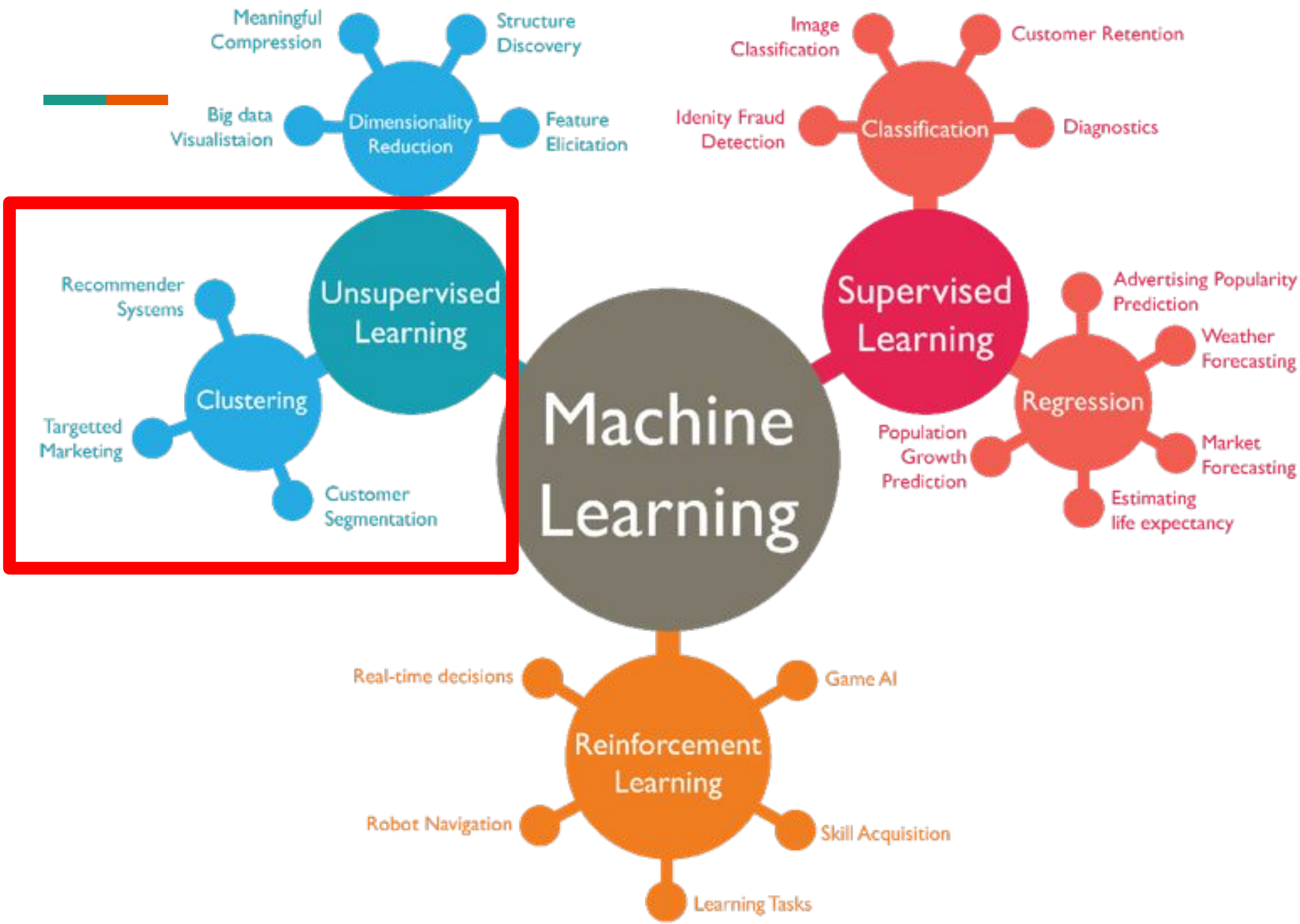
Lesson_12: Unsupervised Learning (Clustering)

Ali Aburas, PhD

Some material adapted from slides by Andrew Moore + Andrew Ng


Machine Learning
Unsupervised Learning
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

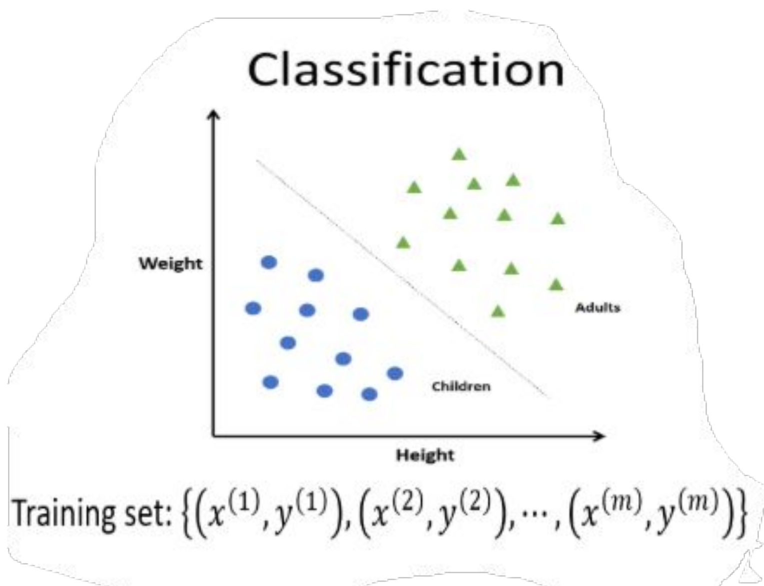
Unsupervised learning

Unsupervised learning methods can be categorized under the following broad areas of ML tasks relevant to unsupervised learning.

- **Clustering** intelligence is the capability of grouping similar objects.
    - Clustering groups “unlabeled” data into “clusters” of similar inputs.
- **Dimensionality reduction** is the process of reducing the number of features in a dataset while retaining as much information as possible.
    - This can be done for a variety of reasons, such as to reduce the complexity of a model, to improve the performance of a learning algorithm, or to make it easier to visualize the data.
    - There are multiple popular algorithms available for dimensionality reduction like Principal Component Analysis (PCA), nearest neighbors, and discriminant analysis.

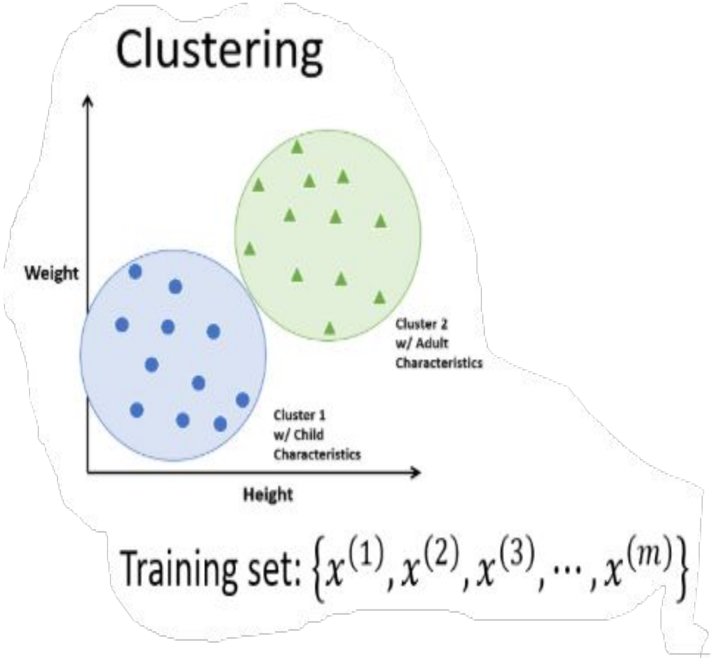
Recap: Classification

- Classification systems:
  - Supervised learning
  - Make a prediction given evidence
  - We’ve seen several methods for this
  - Useful when you have labeled data



Clustering

- Clustering systems:
  - Unsupervised learning
  - Detect patterns in unlabeled data
    - E.g. group emails or search results
    - E.g. find categories of customers
    - E.g. detect anomalous program executions
  - Useful when don’t know what you’re looking for
  - Requires data, but no labels



Clustering

- **Clustering**
  - Goal: split an unlabeled data set into groups or clusters of “similar” data points
    - What could “similar” mean?
      - One option: small (squared) Euclidean distance
  - Requires data, but no labels
  - Useful when don’t know what you’re looking for
    - Group emails or search results
    - Customer shopping patterns
    - Regions of images
      - Break up the image into meaningful or perceptually similar regions.

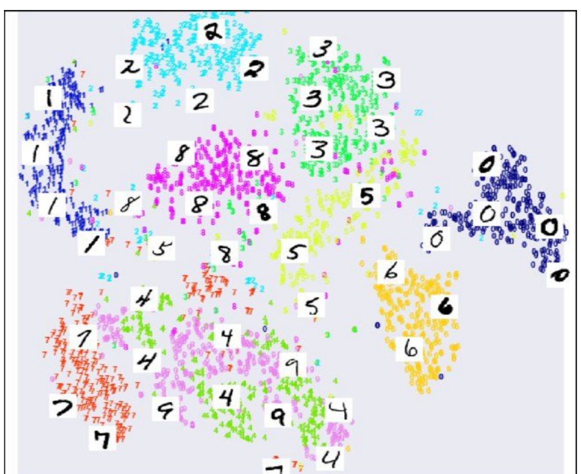


Image Segmentation

[Slide from James Hayes]

Clustering algorithms

- Many clustering algorithms
- Clustering, typically done using a **distance measure** defined between instances or points
- Distance defined by instance **feature space**, so it works with numeric features
  - Requires encoding of categorial values; may benefit from normalization
- We'll look at

1. Centroid-based clustering (e.g., Kmeans)

2. Hierarchical clustering

3. DBSCAN (density-based clustering algorithm)

Common Distance measures:

- Suppose two object *x* and *y* both have *p* features

$$x = (x_1, x_2, \cdots, x_p)$$

$$y = (y_1, y_2, \cdots, y_p)$$

1. The Euclidean distance (also called 2-norm distance) is given by:

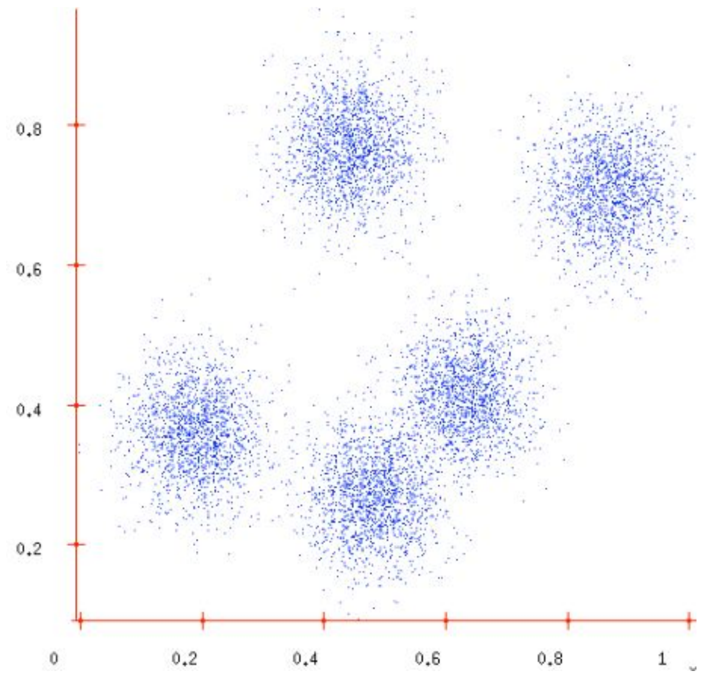
$$d(x, y) = \sqrt[2]{\sum_{i=1}^{p} | x_i - y_i |^2}$$

2. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

$$d(x, y) = \sum_{i=1}^{p} | x_i - y_i |$$

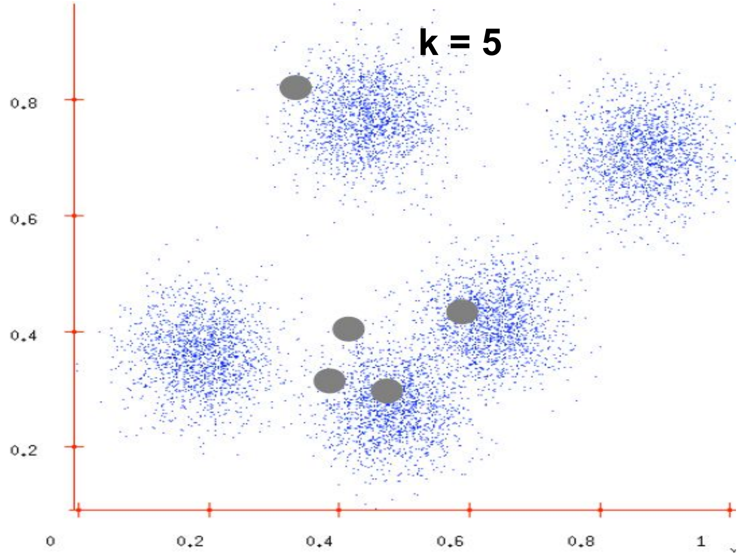
Clustering Data

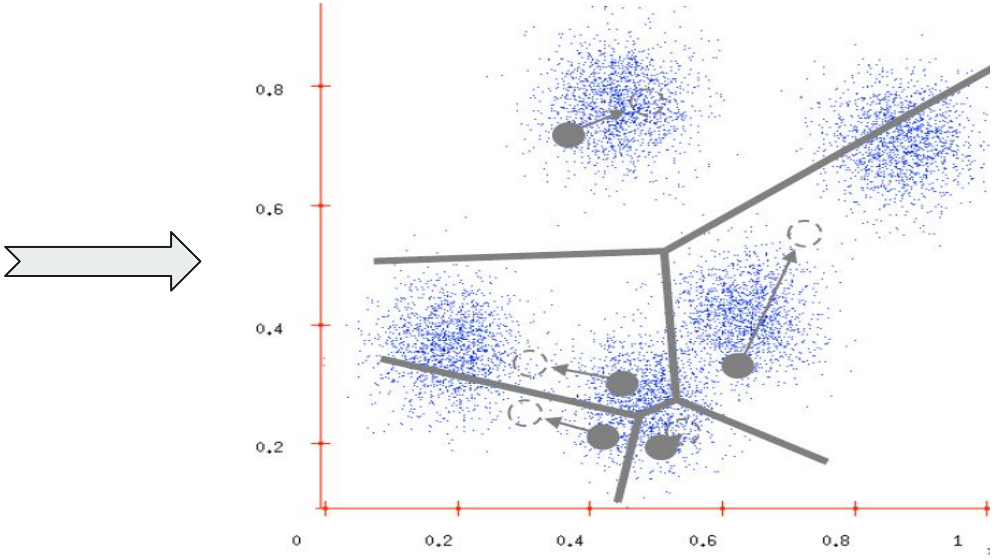
- Given a collection of points (x,y), group them into one or more clusters based on their distance from one another
- How many clusters are there?
- How can we find them?



K-Means Clustering

1. **Randomly** choose k cluster center locations, aka mean (or **centroids)**
2. **Loop until convergence**
   a. assign one point to cluster of the closest mean
   b. Assign each **mean** to the average of its assigned points
3. **Convergence**: no point is assigned to a different cluster





k-Means Clustering

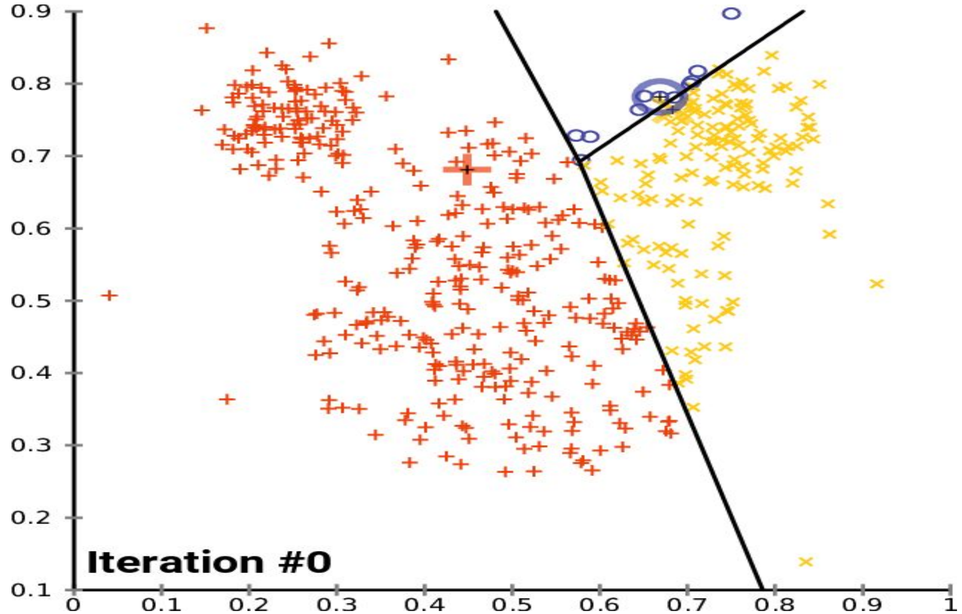
Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



https://en.wikipedia.org/wiki/K-means_clustering#/media/File:K-means_convergence.gif

Instance	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

Data Points

also, first two objects as initial centroids:

Centroid for first cluster c1 = (185, 72)

Centroid for second cluster c2 = (170, 56)

Iteration 1: Now calculating similarity by using *Euclidean distance* measure as:

d(c1, 3) = $\sqrt{(185-168)^2+(72-60)^2} = \sqrt{(17)^2+(12)^2} = \sqrt{289+144} = \sqrt{433}$

d(c2, 3) = $\sqrt{(170-168)^2+(56-60)^2} = \sqrt{(2)^2+(-4)^2} = \sqrt{4+16} = \sqrt{20}$

Here, d(c2, 3) < d(c1, 3)

So, data point 3 belongs to c2.

d(c1, 4) = $\sqrt{(185-179)^2+(72-68)^2} = \sqrt{(6)^2+(4)^2} = \sqrt{36+16} = \sqrt{52}$

d(c2, 4) = $\sqrt{(170-179)^2+(56-68)^2} = \sqrt{(-9)^2+(-12)^2} = \sqrt{81+144} = \sqrt{225}$

Here, d(c1, 4) < d(c2, 4)

So, data point 4 belongs to c1.

d(c1, 5) = $\sqrt{(185-182)^2+(72-72)^2} = \sqrt{(3)^2+(0)^2} = \sqrt{9}$

d(c2, 5) = $\sqrt{(170-182)^2+(56-72)^2} = \sqrt{(-12)^2+(-16)^2} = \sqrt{144+256} = \sqrt{400}$

Here, d(c1, 5) < d(c2, 5)

So, data point 5 belongs to c1.

d(c1, 6) = $\sqrt{(185-188)^2+(72-77)^2} = \sqrt{(-3)^2+(-5)^2} = \sqrt{9+25} = \sqrt{34}$

d(c2, 6) = $\sqrt{(170-188)^2+(56-77)^2} = \sqrt{(-18)^2+(-21)^2} = \sqrt{324+441} = \sqrt{765}$

Here, d(c1, 6) < d(c2, 6)

So, data point 6 belongs to c1.

Instance	X	Y	Distance(C1)	Distance(C2)	Cluster
1	185	72			c1
2	170	56			c2
3	168	60	√433	√20	c2
4	179	68	√52	√225	c1
5	182	72	√9	√400	c1
6	188	77	√34	√765	c1

Iteration 2: Now calculating centroid for each cluster:

$$Centroid\ for\ first\ cluster\ c1 = \left(\frac{185+179+182+188}{4}, \frac{72+68+72+77}{4}\right) = (183.5, 72.25)$$

$$Centroid\ for\ second\ cluster\ c2 = \left(\frac{170+168}{2}, \frac{56+60}{2}\right) = (169, 58)$$

Calculating centroid as mean of data points

Now, again calculating similarity:

$$d(c1,1) = \sqrt{(183.5 - 185)^2 + (72.25 - 72)^2} = 1.5207$$
$$d(c2,1) = \sqrt{(169 - 185)^2 + (58 - 72)^2} = 21.2603$$

Here, d (c1, 1) < d (c2, 1)
So, data point 1 belongs to c1.

$$d(c1,2) = \sqrt{(183.5 - 170)^2 + (72.25 - 56)^2} = 21.1261$$
$$d(c2,2) = \sqrt{(169 - 170)^2 + (58 - 56)^2} = 2.2361$$

Here, d (c2, 2) < d (c1, 2)
So, data point 2 belongs to c2.

$$d(c1,3) = \sqrt{(183.5 - 168)^2 + (72.25 - 60)^2} = 19.7563$$
$$d(c2,3) = \sqrt{(169 - 168)^2 + (58 - 60)^2} = 2.2361$$

Here, d (c2, 3) < d (c1, 3)
So, data point 3 belongs to c2.

$$d(c1,4) = \sqrt{(183.5 - 179)^2 + (72.25 - 68)^2} = 6.1897$$
$$d(c2,4) = \sqrt{(169 - 179)^2 + (58 - 68)^2} = 14.1421$$

Here, d (c1, 4) < d (c2, 4)
So, data point 4 belongs to c1.

$$d(c1,5) = \sqrt{(183.5 - 182)^2 + (72.25 - 72)^2} = 1.5207$$
$$d(c2,5) = \sqrt{(169 - 182)^2 + (58 - 72)^2} = 19.1050$$

Here, d (c1, 5) < d (c2, 5)
So, data point 5 belongs to c1.

$$d(c1,6) = \sqrt{(183.5 - 188)^2 + (72.25 - 77)^2} = 6.5431$$
$$d(c2,6) = \sqrt{(169 - 188)^2 + (58 - 77)^2} = 26.8701$$

Here, d (c1, 6) < d (c2, 6)
So, data point 6 belongs to c1.

Instance	X	Y	Distance(C1)	Distance(C2)	Cluster
1	185	72	1.5207	21.2603	c1
2	170	56	21.1261	2.2361	c2
3	168	60	19.7563	2.2361	c2
4	179	68	6.1897	14.1421	c1
5	182	72	1.5207	19.105	c1
6	188	77	6.5431	26.8701	c1

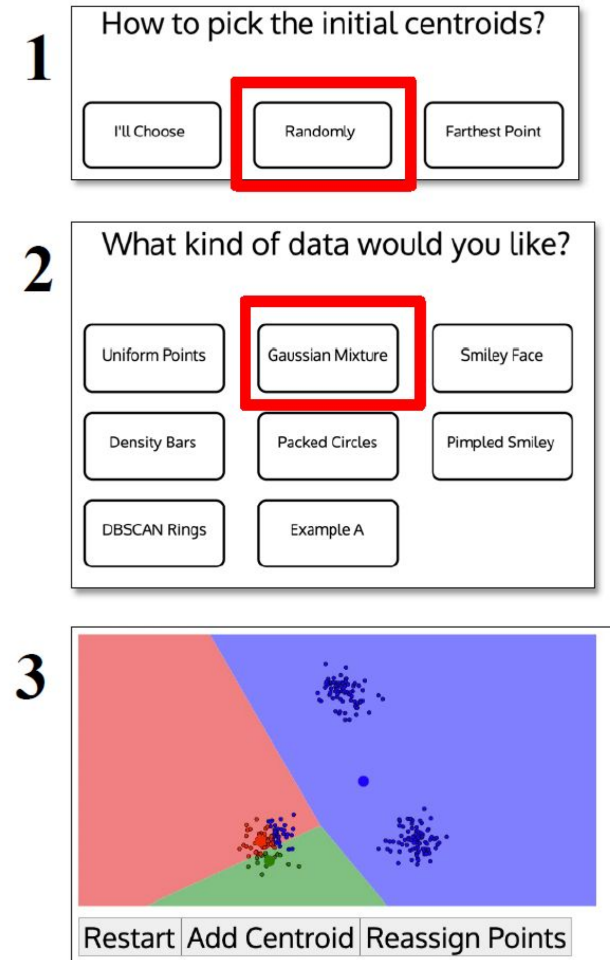
Visualizing k-means (CLICK ME)

Visualizing k-means

Interactively experiment with K-means clustering

1. Three ways to assign positions of initial centroids
2. Eight ways to generate data points to be clustered
3. You choose the value of k when adding centroids
4. Then walk through the iterations of the k-means algorithm

It can also demonstrate the DBSCAN clustering algorithm



K-means issues

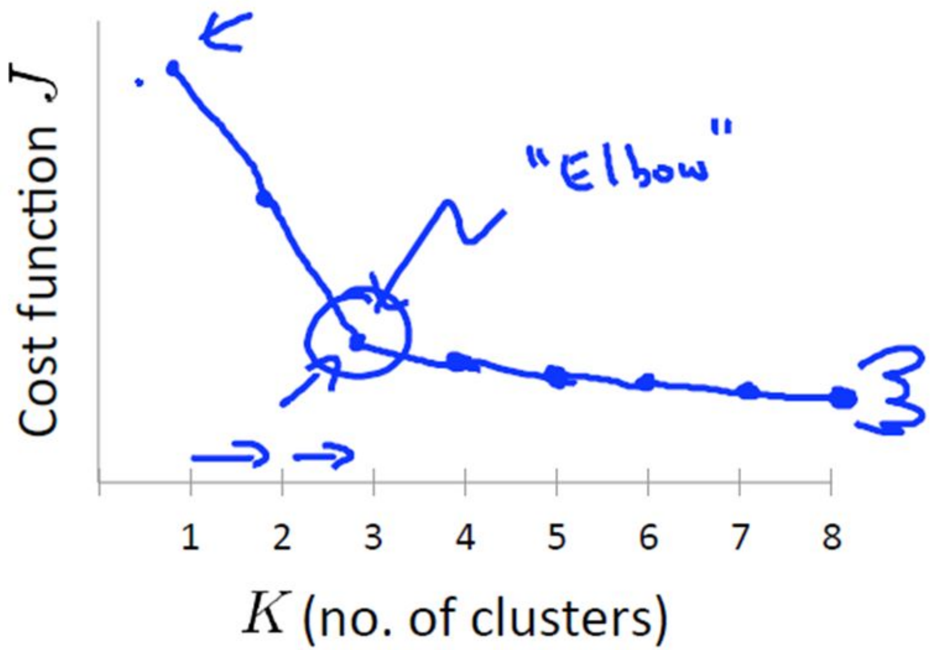
How to choose the number of clusters (K)?

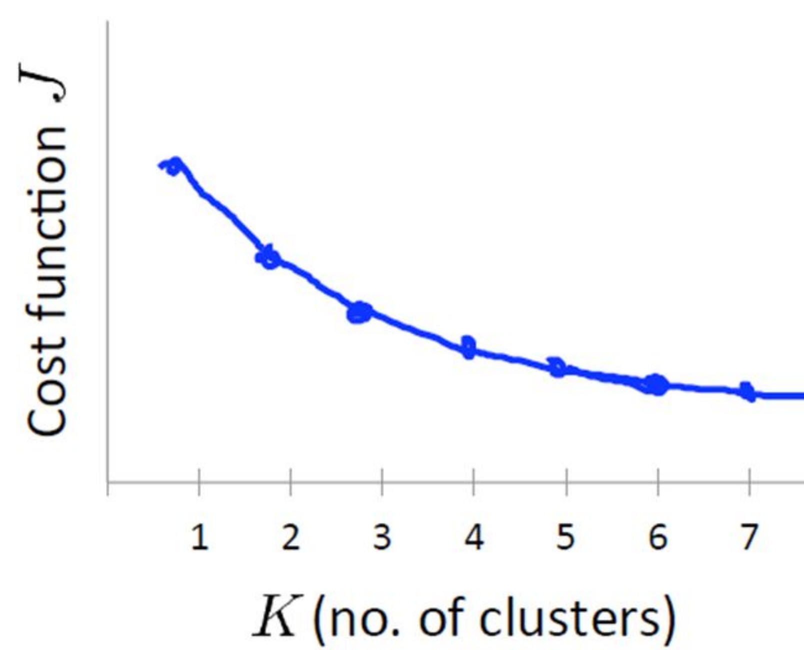
How to initialize K (Local optima)?

Hard vs. soft clustering

How to choose K?

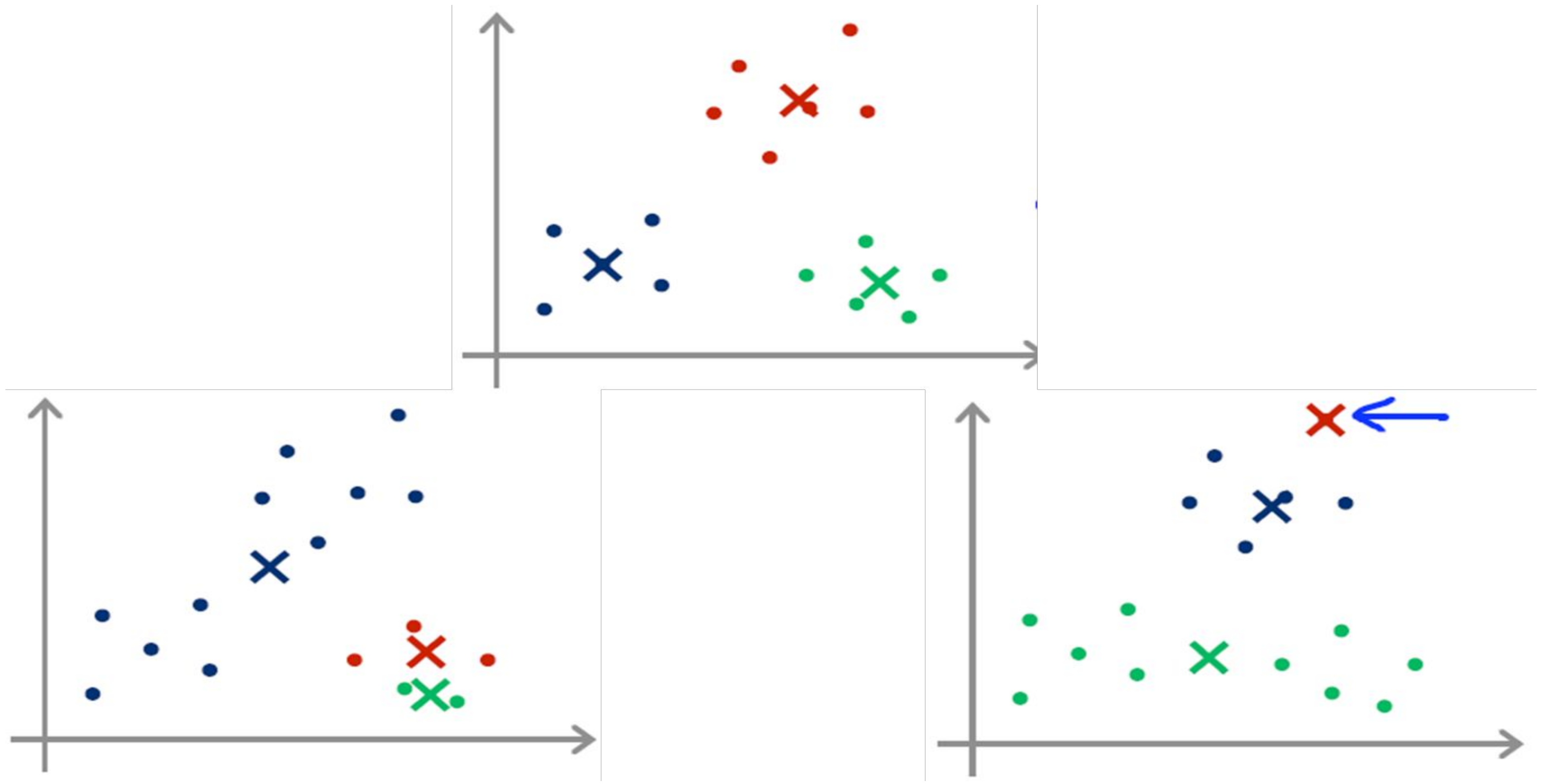
1) Elbow method





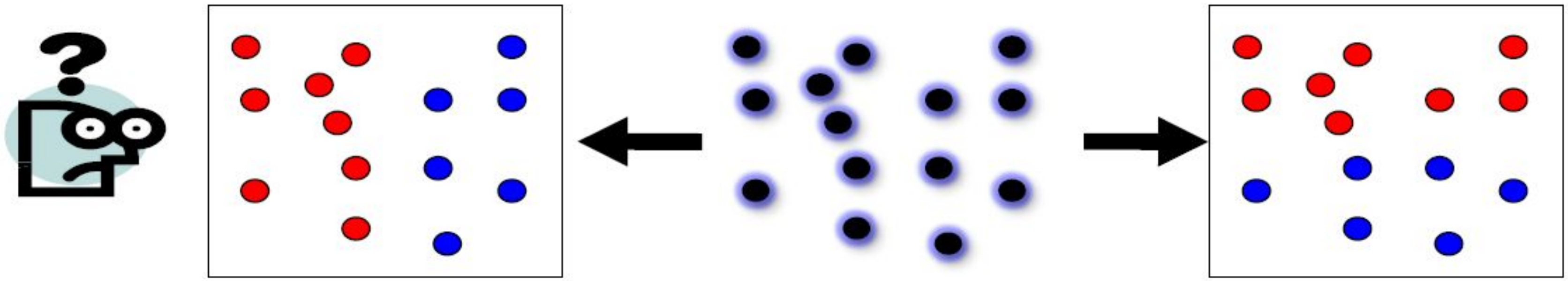
Local optima

- **Problem**: Converted solution may not be "**Optimal**"
- But it produced "**reasonable**" clusters in practice.

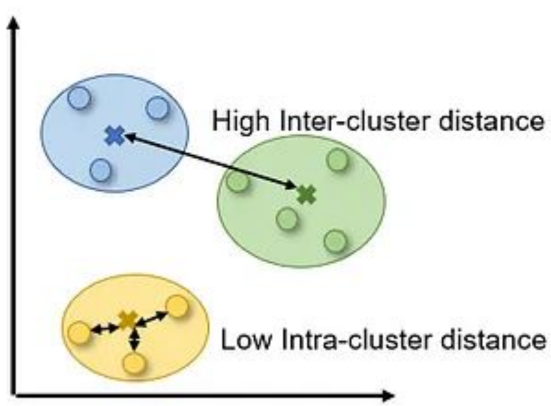


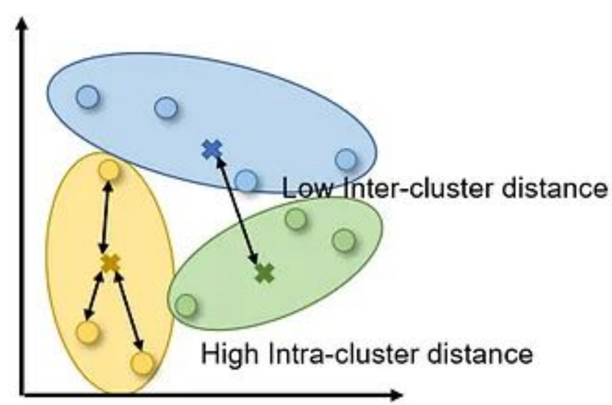
Local optima

How do you tell it which clustering you want?



We aim to reach a high intra-cluster (within-cluster) similarity and a low inter-cluster (between-cluster) similarity.





Hard vs. soft clustering

- **Hard clustering**: Each document belongs to exactly one cluster
  - More common and easier to do
- **Soft clustering**: A document can belong to more than one cluster.
  - Makes more sense for applications like creating browsable hierarchies
  - You may want to put a pair of **sneakers** in two clusters:
    - (i) sports apparel and (ii) shoes
  - You can only do that with a soft clustering approach
- We only covered hard clustering..

Recap: K-means Clustering

- **K-Means** is an iterative algorithm that assigns K clusters to a dataset where each cluster has a center that is the average of all the points situated in it, always referred to as the centroid.
  - K-means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters.
- **Advantages of K-Means**
  - **Simplicity:** The advantage of K-Means is that it is simple to use and has a rather uncomplicated algorithm.
  - **Efficiency:** It is effective in terms of time complexity and thus can easily work with large data sets.
  - **Speed:** generally converges quickly.
- **Disadvantages of K-Means**
  - **Sensitivity to Outliers**: K-Means is also susceptible to noise and outliers, chiefly because of its reliance on means as the critical measure in binning assignments.
  - **Initial Centroids**: The selection of the first K centroids may differ, and this may lead to different clustering and hence inaccurate clustering.

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/