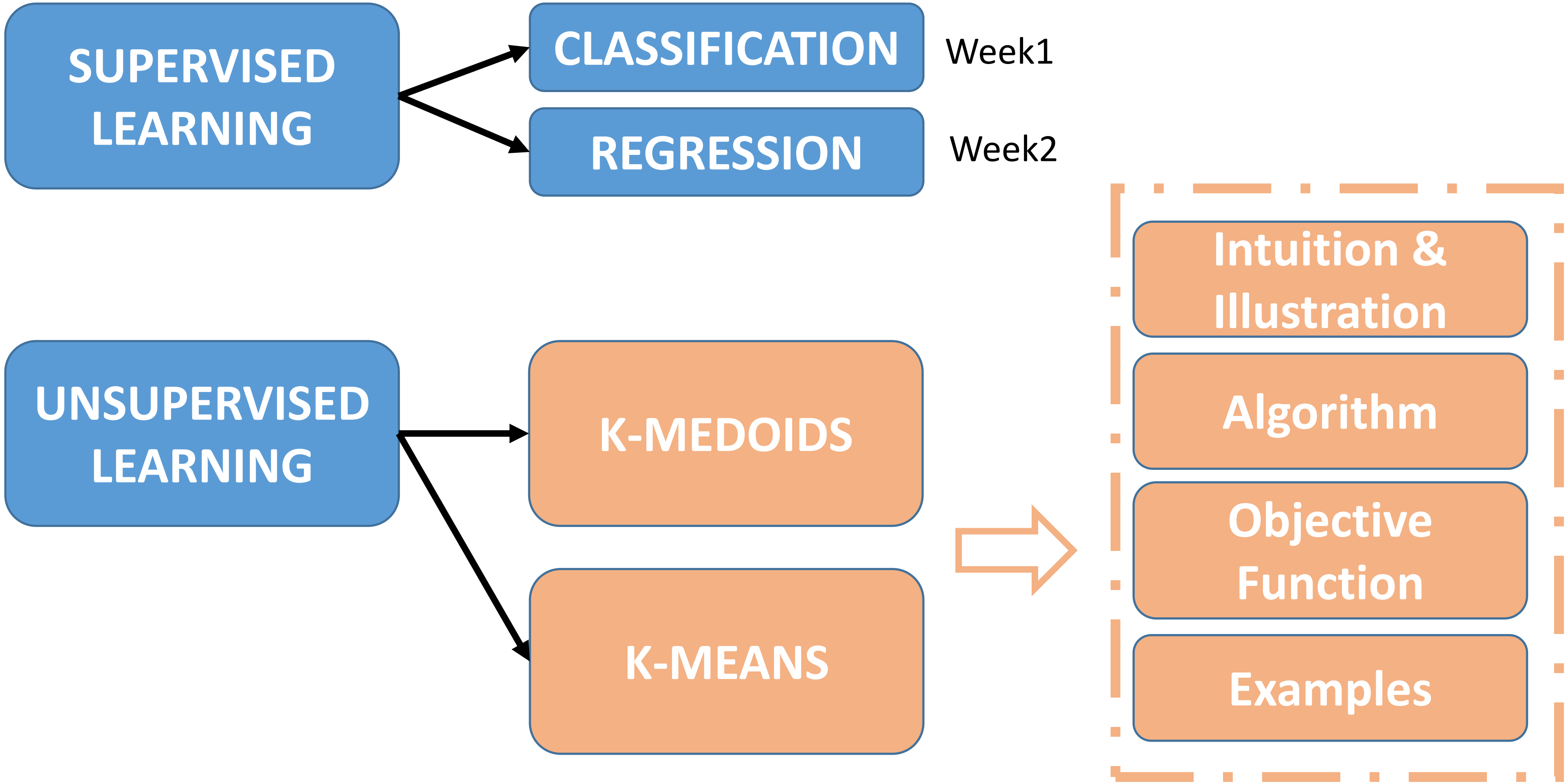
50.007 Machine Learning

K-Means & K-Medoids

Yixiao Wang

Assistant Professor, ISTD/DAI, SUTD

BIG PICTURE



SUPERVISED LEARNING

CLASSIFICATION

Week1

REGRESSION

Week2

UNSUPERVISED LEARNING

K-MEDOIDS

K-MEANS

Intuition & Illustration

Algorithm

Objective Function

Examples

You will learn all these by the end of this week :)

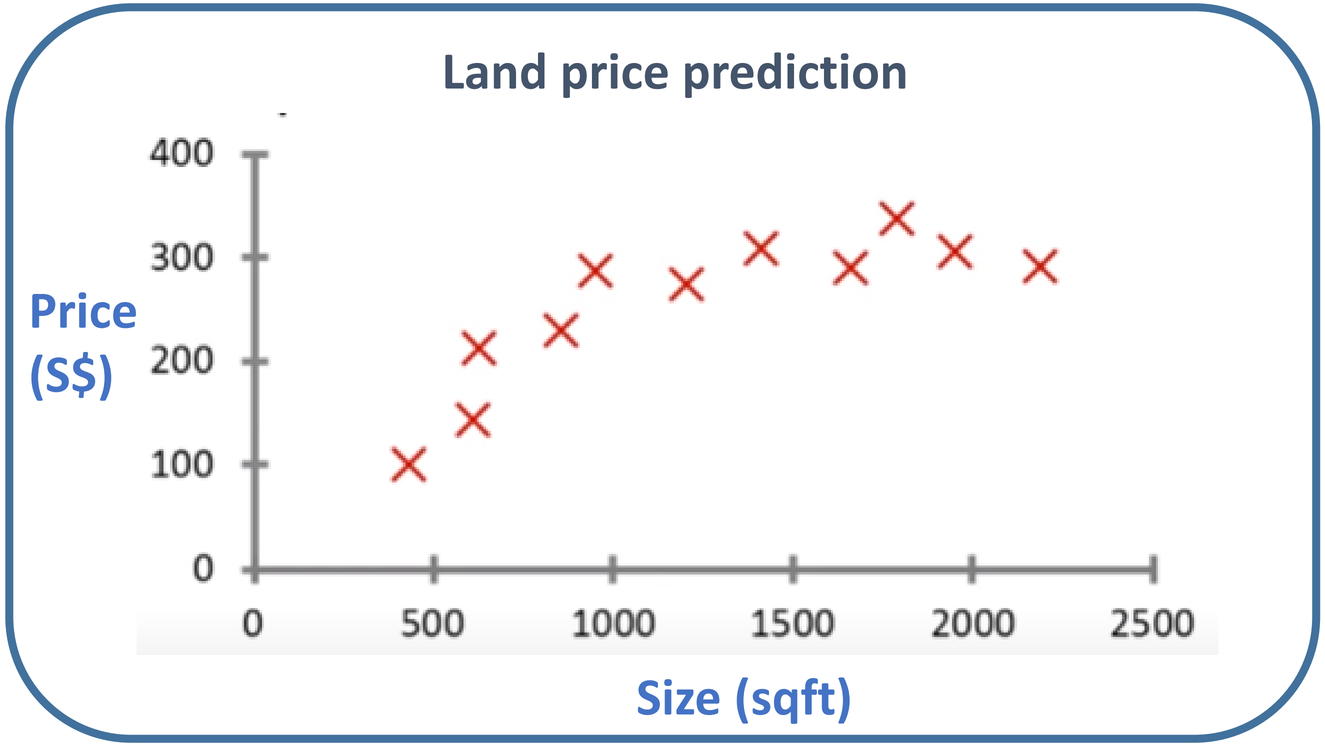
Supervised Learning

Supervised Learning: Regression

Recap of Supervised Learning

- In supervised learning, the machine learning model is trained on a **labeled dataset**. After training, we can predict the outcome of a new data point.
- Very common in machine learning! We'll study some examples together.

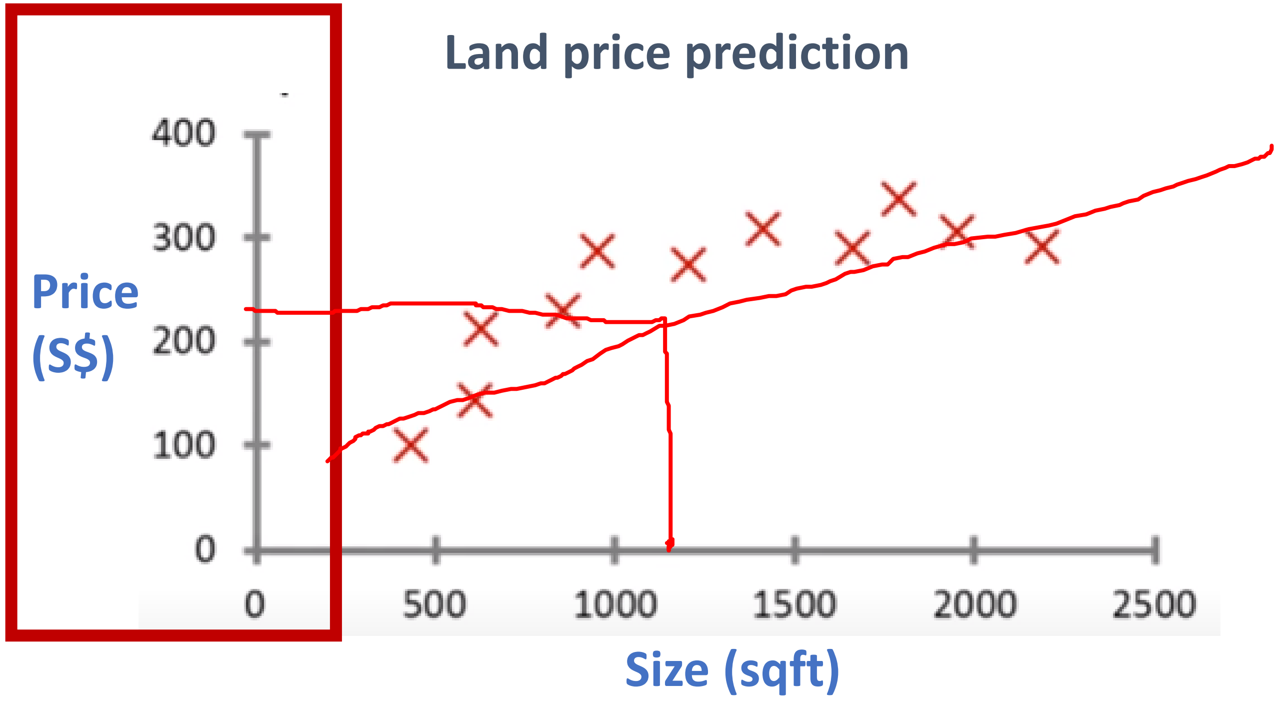
This is labeled dataset



TRAINING DATA

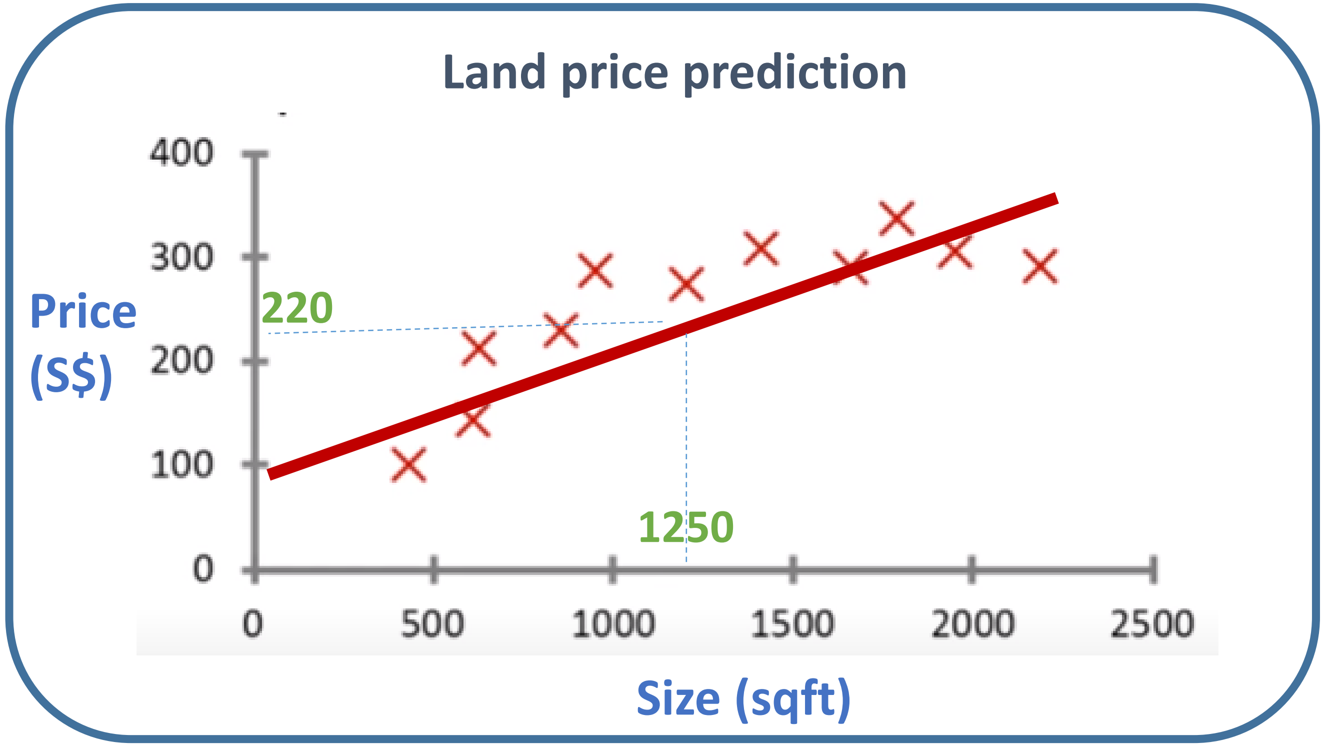
Recap of Supervised Learning

In supervised learning, **right answers** are given during training.



We have a dataset for lands, and in this dataset we have the price of each land.

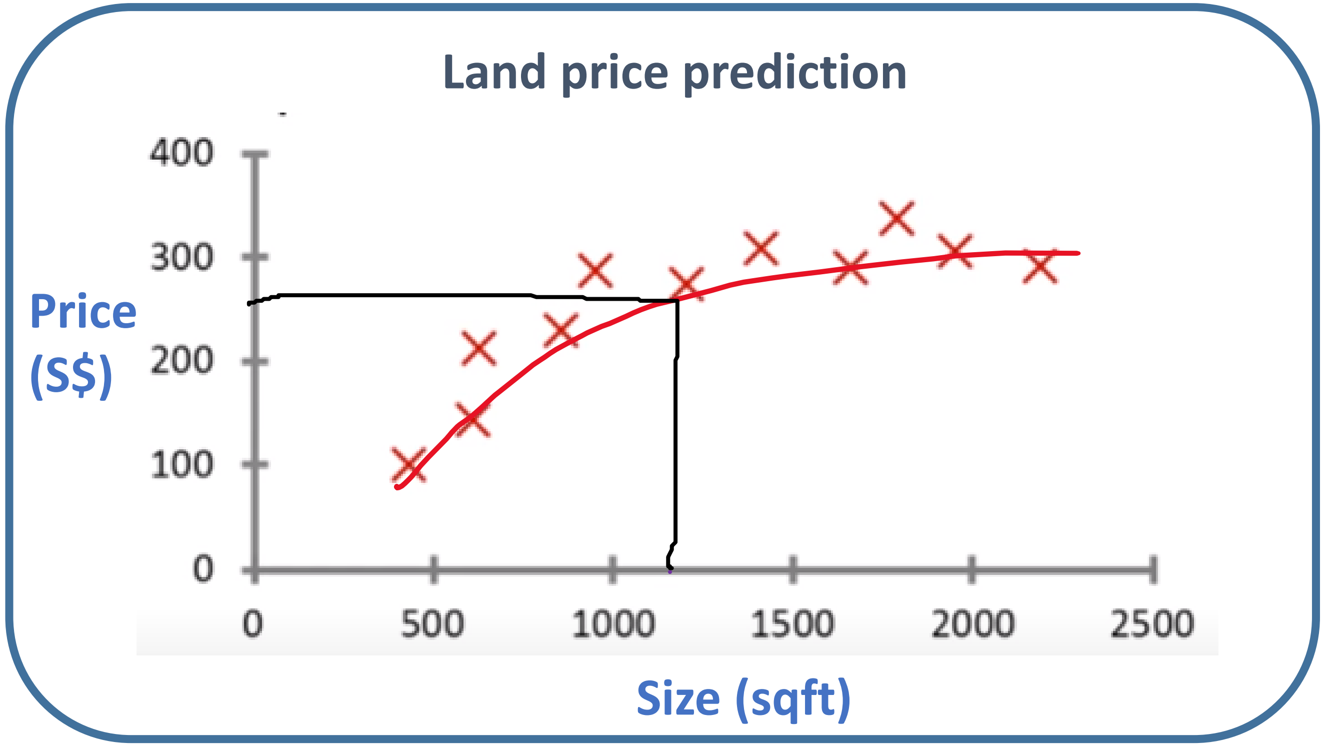
Introduction to Supervised Learning



Imagine that you own a land which is **1250 sqft**, and hoping to sell this land. How much will you get?

An easy way: try to fit a straight line to your data…

Introduction to Supervised Learning



We'll study more sophisticated algorithms to perform prediction.

F

Imagine that you own a land which is **1250 sqft**, and hoping to sell this land. How much will you get?

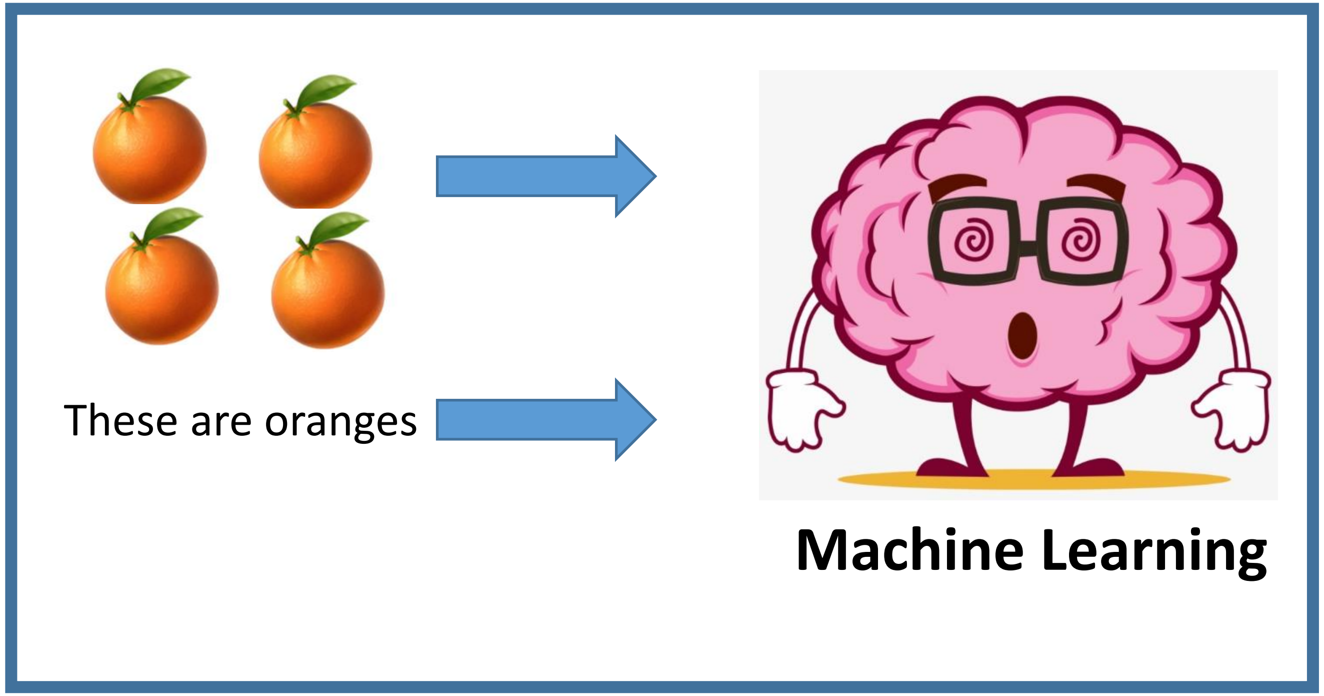
An easy way: try to fit a straight line to your data…

Supervised Learning: Classification

Introduction to Supervised Learning

Another simple example:

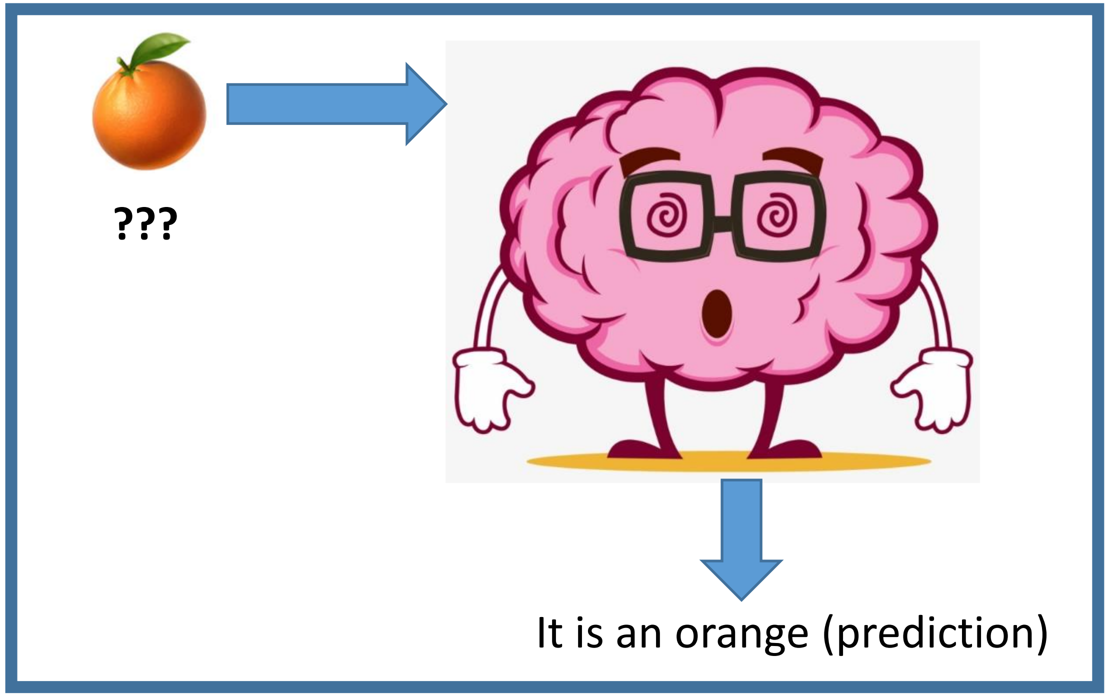
During training, we have the labels!



These are oranges

Machine Learning

TRAINING



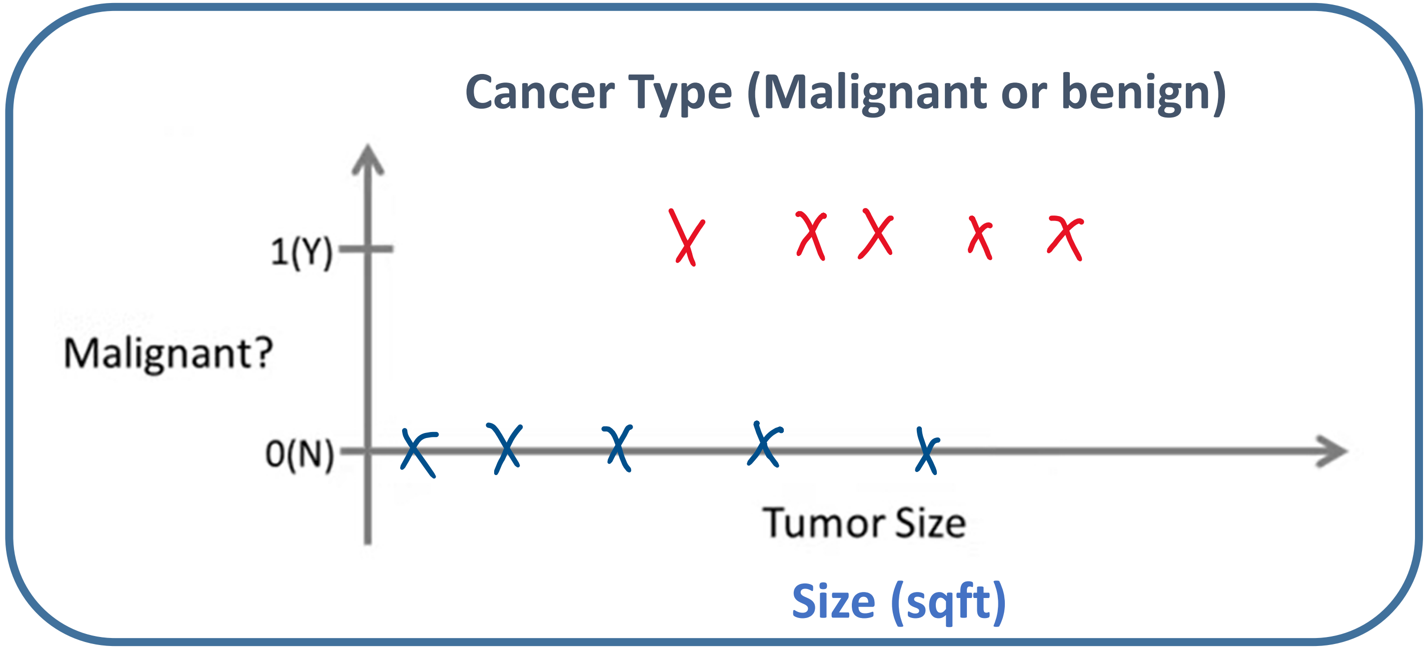
???

It is an orange (prediction)

TESTING

Supervised learning: regression, Naïve Bayes theorem, SVM (we'll study later)

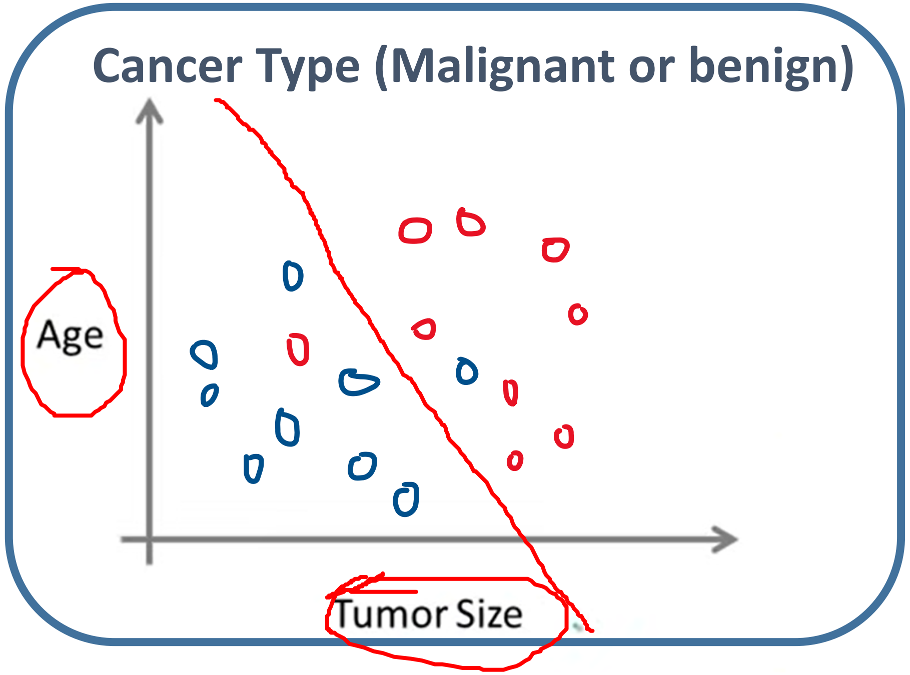
Introduction to Supervised Learning



There could be multiple groups.

Imagine someone got breast cancer, we want to know whether it’s a malignant or benign tumor based on tumor size.

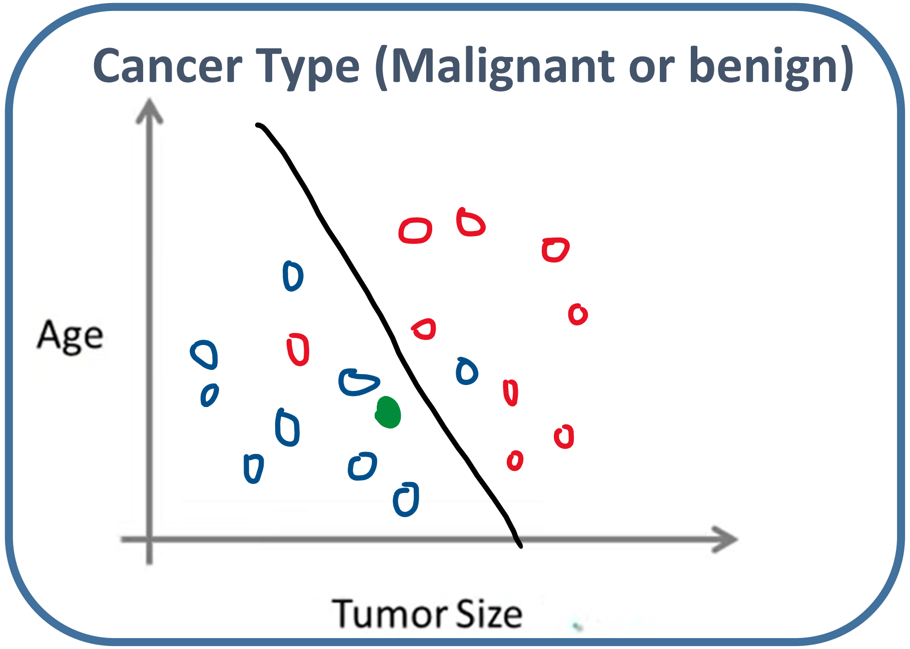
Introduction to Supervised Learning



There could be multiple features.

We could have multiple features/attributes for this classification problem to improve prediction accuracy.

Introduction to Supervised Learning



For instance, SVM could help us to deal with infinite number of features for a classification problem.

We could have multiple features/attributes for this classification problem to improve prediction accuracy.

Quick Question:

- Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months
- Problem 2: You would like software to examine individual customer accounts and for each account decide if it has been hacked or compromised.

What kind of supervised learning problems are problem 1 and problem 2? (regression or clustering?)

Quick Question:

Of the following examples, which would you address using the unsupervised learning algorithm?

1. Given email labeled as spam/not spam: learning a spam filter.
2. Given a set of new articles found on the web, group them into sets of articles about the same story.
3. Given a data set of customer data, automatically discover market segments and group customers into different market segments.
4. Given a data set of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

Unsupervised Learning

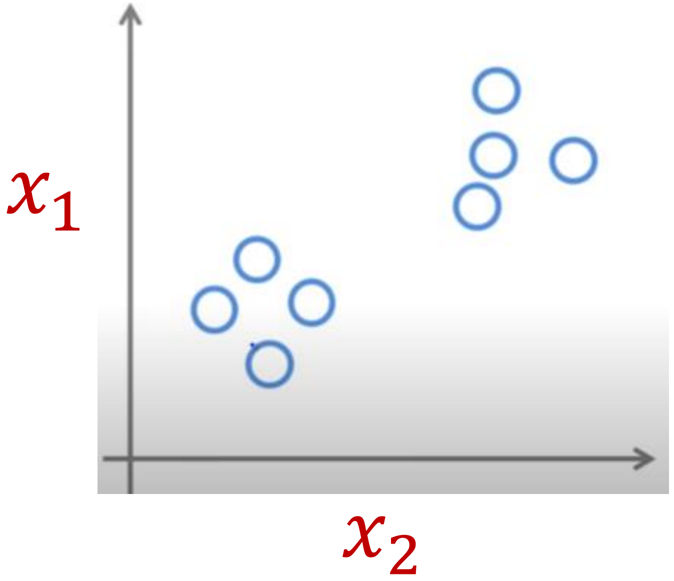
Introduction to Unsupervised Learning

- In unsupervised learning, the machine learning model is trained on **unlabeled dataset**.
- In supervised learning, we know what kind of data we are dealing (because we have the labels of data).
- Unsupervised learning is often more difficult than supervised learning as we know little about the training data.
- With unsupervised learning, we are trying to find groups and clusters; density estimation, dimensionality reduction.

Unsupervised Learning - Example

- Data without any labels (or right answers)

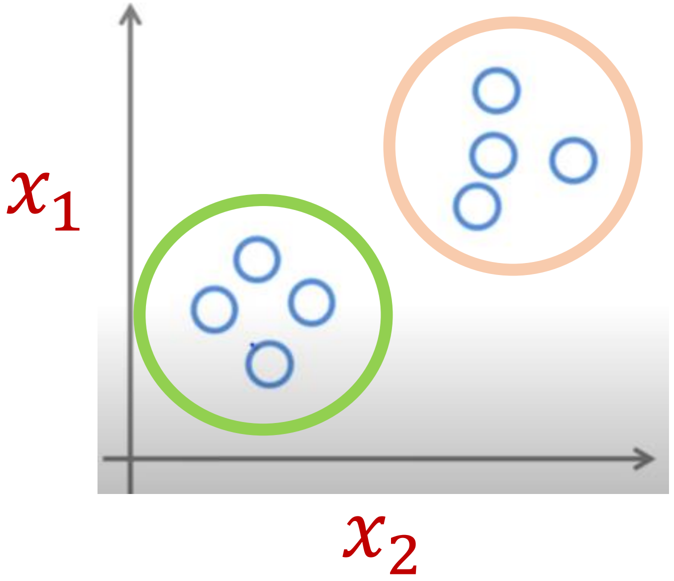
- **Example:** here is the dataset, can you find some structure?



Unsupervised Learning - Example

- Data without any labels (or right answers)

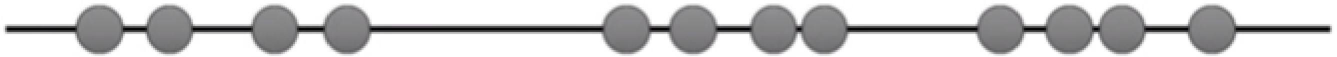
- **Example:** here is the dataset, can you find some structure?



Given this dataset, an unsupervised learning algorithm may decide that the data has 2 different clusters!

Unsupervised Learning - Example

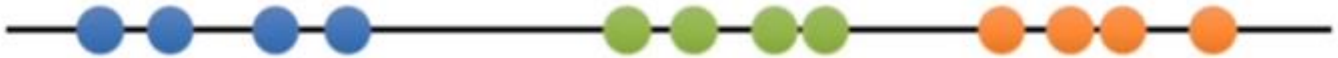
Imagine you have some data that you can plot on a line.



You also know that there will be 3 clusters.

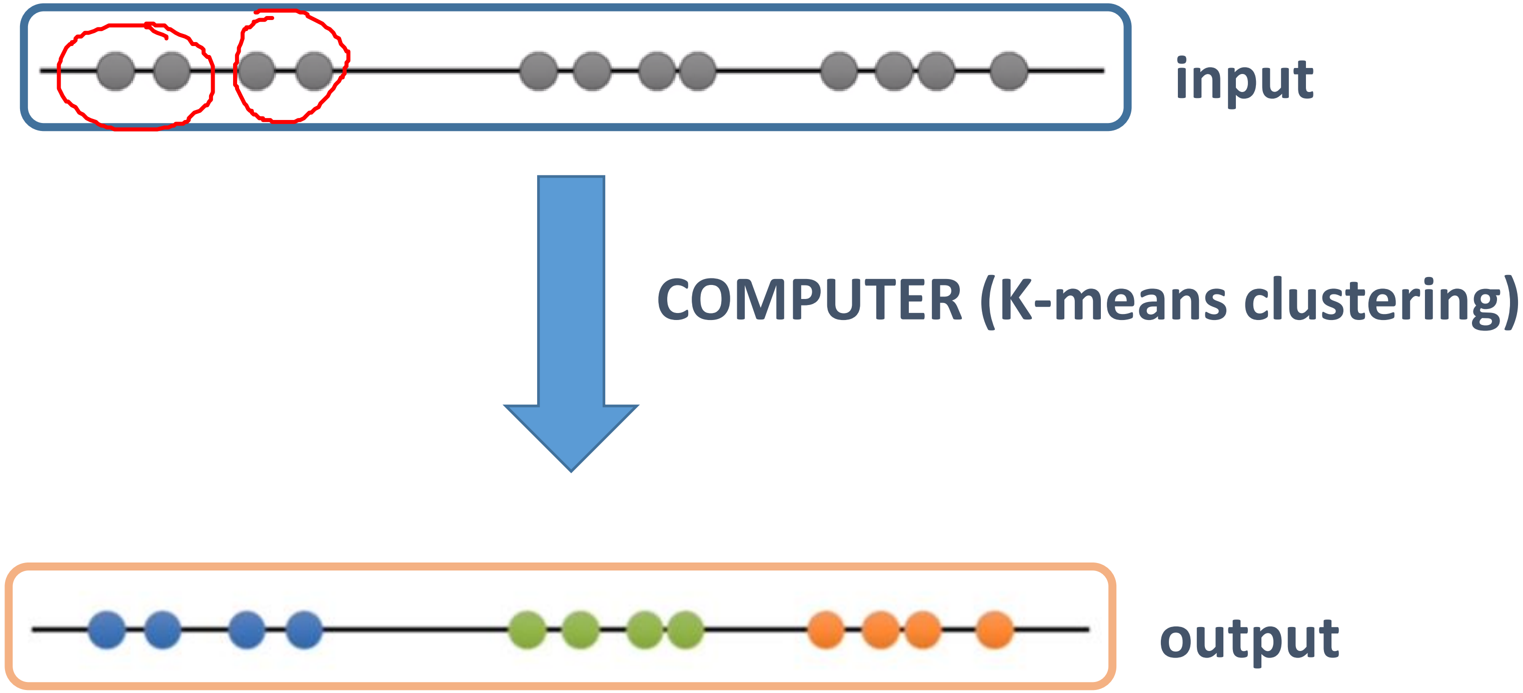
Imagine that these data points are from 3 different tumor types.

Very obvious (to human eye) clusters:



Unsupervised Learning - Example

Can computer identify the same 3 clusters?



input

COMPUTER (K-means clustering)

output

Supervised Learning vs Unsupervised Learning

Supervised Learning vs Unsupervised Learning

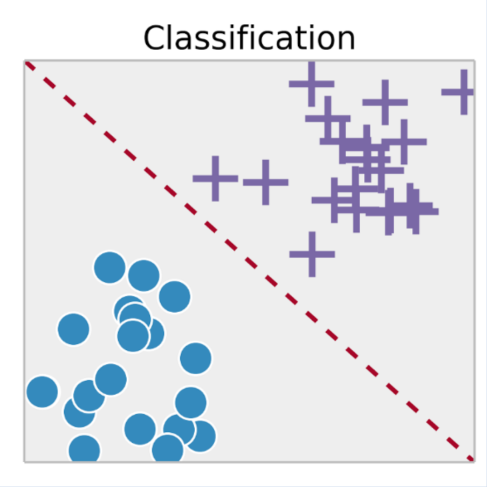
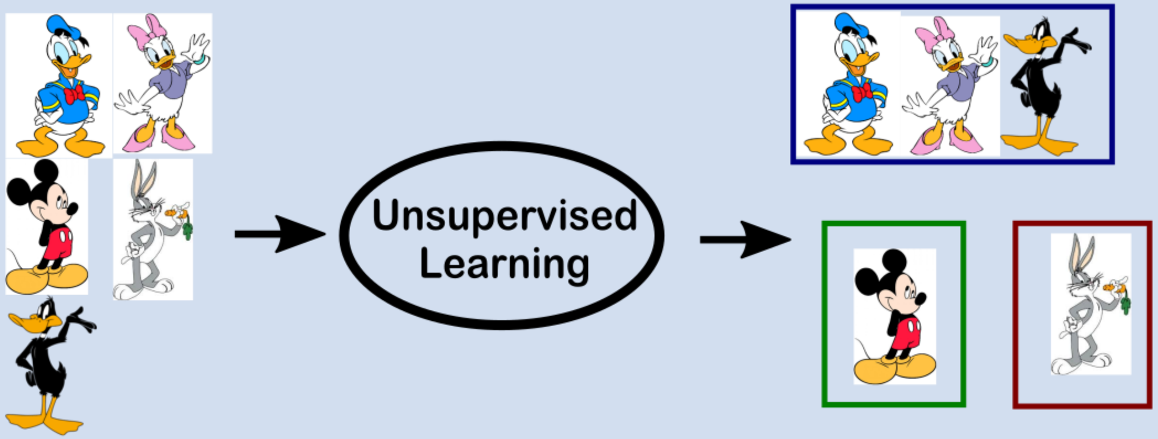
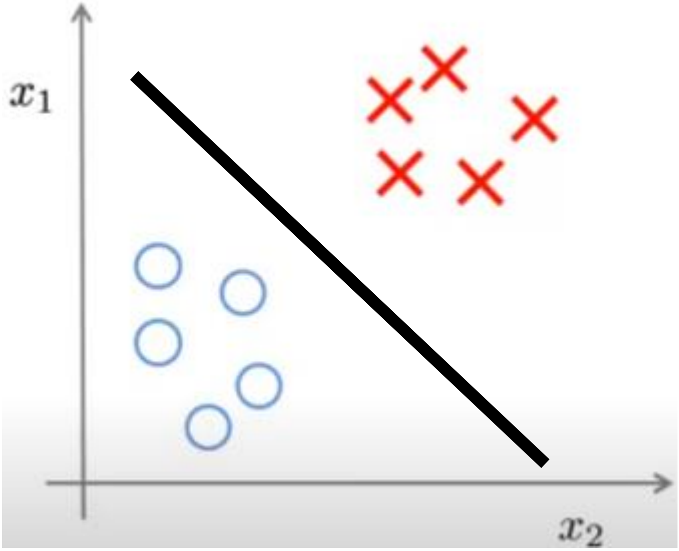
Supervised Learning	Unsupervised Learning
• Classification: classifying labeled data • Regression: Predicting trends using previous labeled data  	• Clustering: Finding patterns and groupings from unlabeled data  *Labels are not given*

Figure taken from towardsdatascience.com

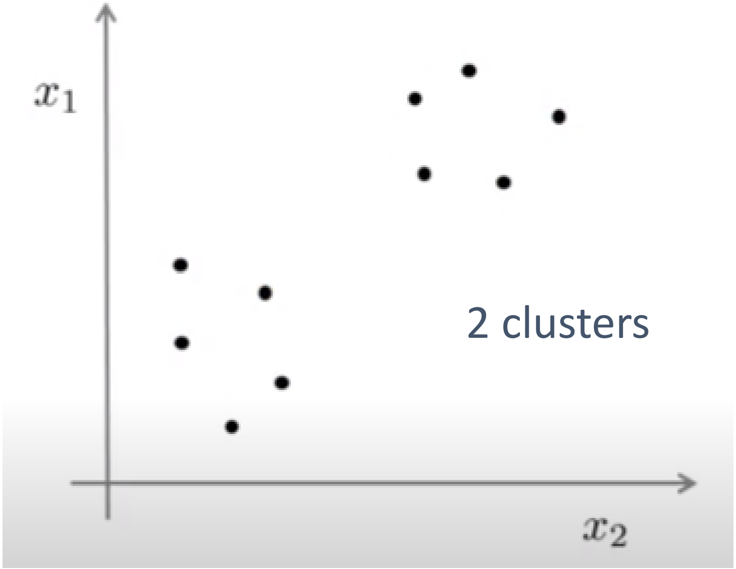
Supervised Learning vs Unsupervised Learning

A typical supervised learning problem



We are given dataset and labels, and
we are trying to find the decision boundary that
separates 2 classes.

A typical unsupervised learning problem

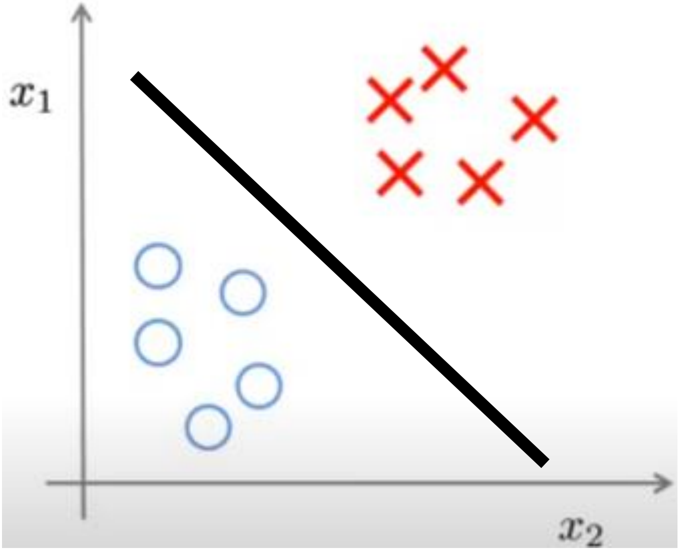


2 clusters

We are given this unlabeled dataset to an
algorithm, and ask algorithm to find some
structure in the data for us.

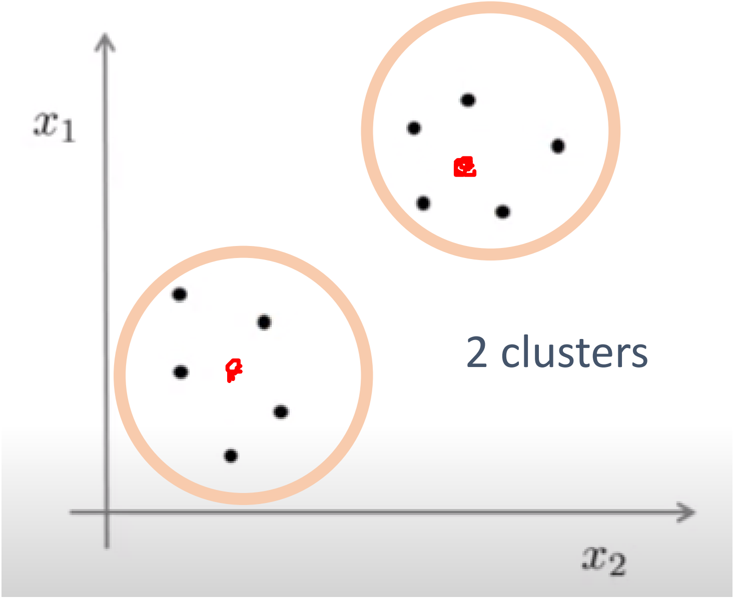
Supervised Learning vs Unsupervised Learning

A typical supervised learning problem



We are given dataset and labels, and
we are trying to find the decision boundary that separates 2 classes.

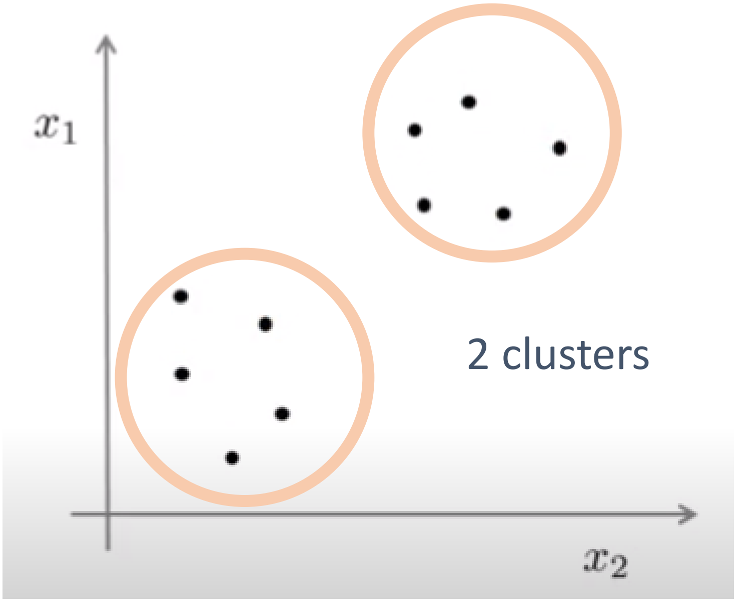
A typical unsupervised learning problem



We are given this unlabeled dataset to an algorithm, and ask algorithm to find some structure in the data for us.

An algorithm that finds clusters is a clustering algorithm.

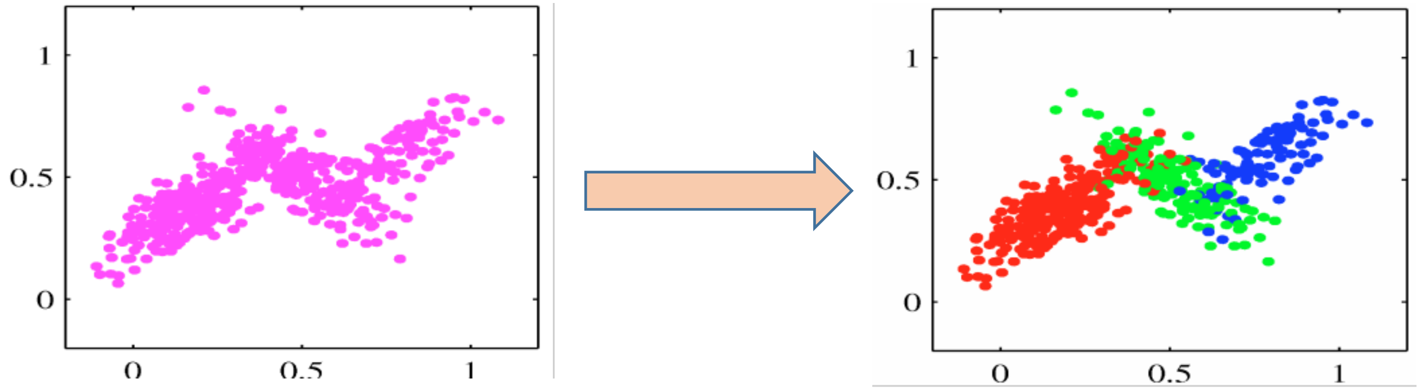
A typical unsupervised learning problem



We are given this unlabeled dataset to an algorithm, and ask algorithm to find some structure in the data for us.

Clustering

Goal: Segment data into groups of similar points



- Segment pixels in an image by object
- Group network participants into communities
- Identify cancer subtypes from gene expression patterns

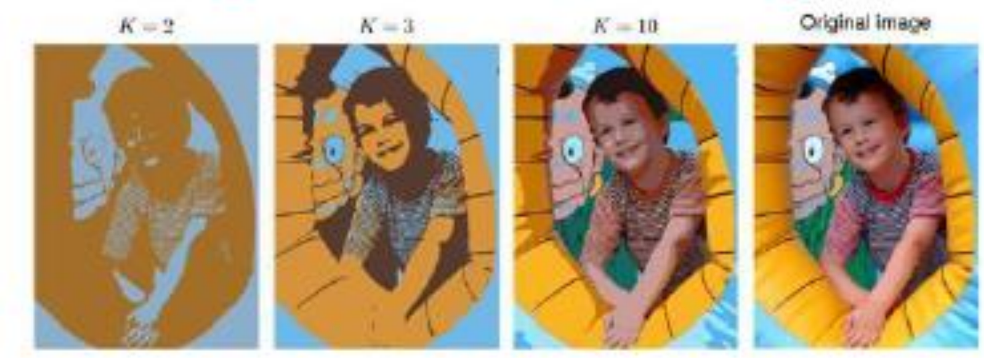
MANY MORE APPLICATIONS ☺

Begin with one of the simplest and most popular clustering algorithm: k-means

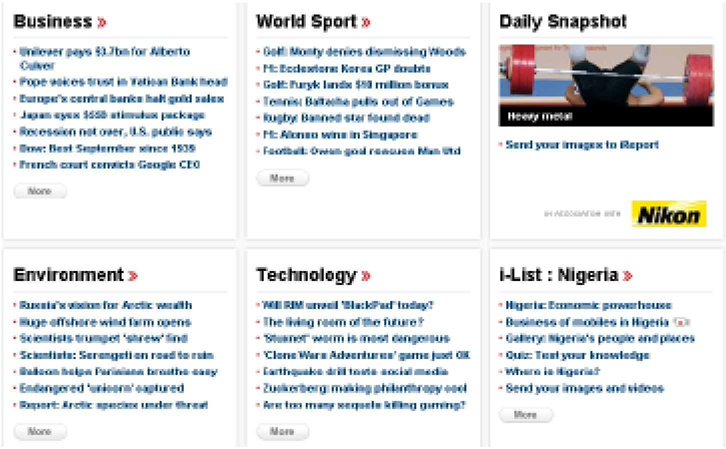
Clustering

Finding groups of objects such that the objects in a group will be **similar** (or related) to one another and **different** from (or unrelated to) the objects in other groups.

Segment image into regions



Better understanding and search



Business »

- Unilever pays $3.7bn for Alberto Culver
- Pope voices trust in Vatican Bank head
- Europe's central banks halt gold sales
- Japan eyes $55B stimulus package
- Recession not over, U.S. public says
- Dow: Best September since 1939
- French court convicts Google CEO

More

World Sport »

- Golf: Monty denies dismissing Woods
- F1: Ecclestone: Korea GP doubts
- Golf: Furyk lands $10 million bonus
- Tennis: Bartoli pulls out of Games
- Rugby: Banned star found dead
- F1: Alonso wins in Singapore
- Football: Owen goal rescues Man Utd

More

Daily Snapshot

Heavy metal

- Send your images to iReport

IN ASSOCIATION WITH Nikon

Environment »

- Russia's vision for Arctic wealth
- Huge offshore wind farm opens
- Scientists trumpet 'shrew' find
- Scientists: Serengeti on road to ruin
- Balloon helps Parisians breathe easy
- Endangered 'unicorn' captured
- Report: Arctic species under threat

More

Technology »

- Will RIM unveil 'BlackPad' today?
- The living room of the future?
- 'Stuxnet' worm is most dangerous
- 'Clone Wars Adventures' game just OK
- Earthquake drill tests social media
- Zuckerberg: making philanthropy cool
- Are too many sequels killing gaming?

More

i-List : Nigeria »

- Nigeria: Economic powerhouse
- Business of mobiles in Nigeria
- Gallery: Nigeria's people and places
- Quiz: Test your knowledge
- Where is Nigeria?
- Send your images and videos

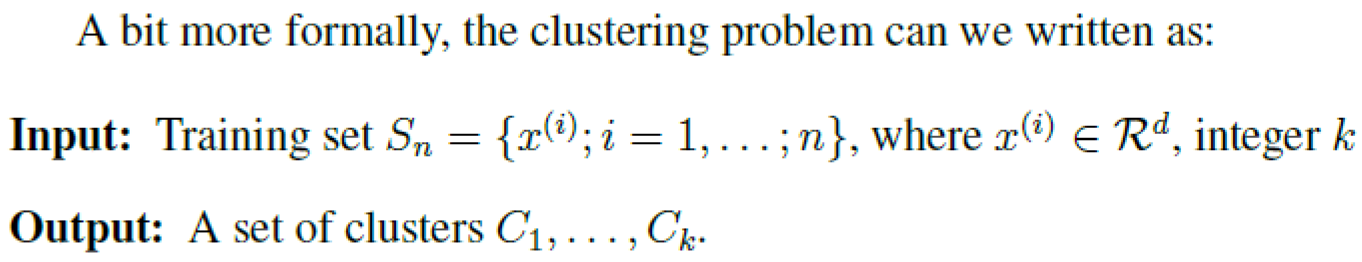
More

Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

Clustering & Unsupervised Learning

- The goal of unsupervised learning is to uncover useful structure in the data such as identify groups or clusters of similar examples.
- Clustering is one of the key problems in exploratory data analysis.
- Examples of clustering applications include:
  - mining customer purchase patterns,
  - market segmentation,
  - modeling language families,
  - grouping search results according to topics, and
  - data compression.

Clustering (mathematical aspect)



A bit more formally, the clustering problem can we written as:

Input: Training set $S_n = \{x^{(i)}; i = 1, \ldots ; n\}$, where $x^{(i)} \in \mathcal{R}^d$, integer k

Output: A set of clusters $C_1, \ldots, C_k$.

How to find these clusters?

Clustering - Distance Metric

- Note that we have yet to specify any criterion for selecting clusters or cluster representatives.
- In clustering algorithms, we should be able to compare pairs of points to determine whether:
  - they are indeed similar (should be in the same cluster), or
  - not (should be in a different cluster).
- The comparison can be either in terms of similarity such as **cosine similarity** or dissimilarity as in **Euclidean distance**.

Clustering - Distance Metric

- **Cosine similarity** is simply the angle between two vectors (elements):

$$\cos(x^{(i)}, x^{(j)}) = \frac{x^{(i)} \cdot x^{(j)}}{\|x^{(i)}\| \, \|x^{(j)}\|} = \frac{\sum_{l=1}^{d} x_l^{(i)} x_l^{(j)}}{\sqrt{\sum_{l=1}^{d} (x_l^{(i)})^2} \sqrt{\sum_{l=1}^{d} (x_l^{(j)})^2}}$$

- Cosine similarity measures the **similarity** between two vectors of an inner product space.
- It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.
- The **smaller** the angle, **higher** the cosine similarity.

Clustering - Distance Metric

- In this lecture, we will primarily use **squared Euclidian distance**

$$\mathrm{dist}(x^{(i)}, x^{(j)}) = \left\|x^{(i)} - x^{(j)}\right\|^2 = \sum_{l=1}^{d}(x_l^{(i)} - x_l^{(j)})^2$$

- In clustering, the choice of which distance metric to use is important as it will determine the type of clusters we will find.
- Once we have the distance metric, we can specify an objective function for clustering. In other words, we specify the cost of choosing any particular set of clusters or their representatives (a.k.a. centroids).
- The "optimal" clustering is then obtained by minimizing this cost.

TO SUM UP

You are given a data set where each observed example has a set of features, but has no labels. You need to find clusters.

Remember that labels are an essential ingredient to a supervised algorithm.

So we can't run supervised learning.

What can we do?

TO SUM UP

You are given a data set where each observed example has a set of features, but has no labels. You need to find clusters.

Remember that labels are an essential ingredient to a supervised algorithms.

So we can't run supervised learning.

K-Means algorithm

K-Means

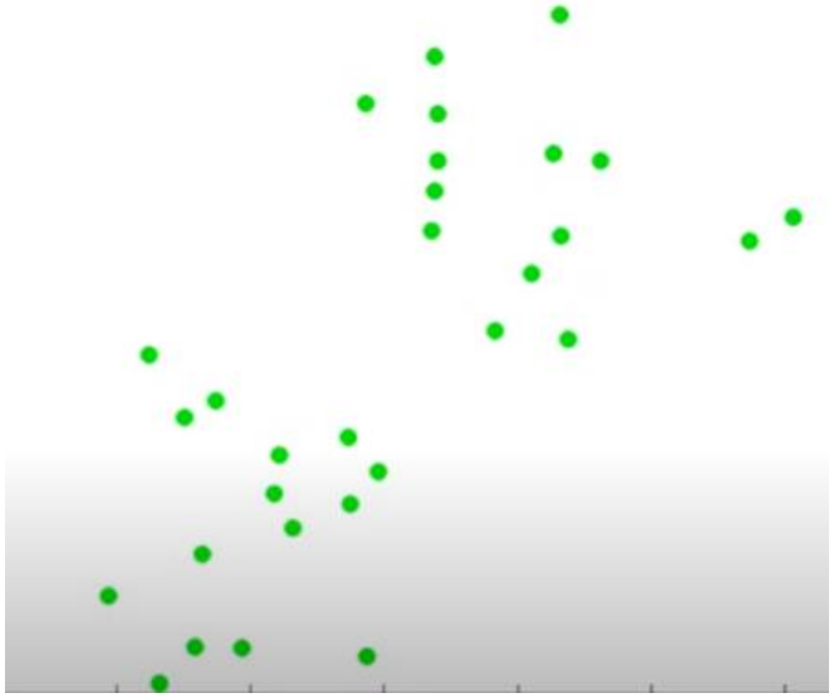
K-Means

- In clustering problem, we are given unlabelled dataset and we would like to have an algorithm that groups the data into subsets.
- K-Means algorithm is the most popular & widely-used clustering algorithm.
- Let’s study with pictures to understand how k-means algorithm works. (believe me, you’ll all understand!)

- We’ll then write the objective function & mathematical formulation.

K-Means (illustration)

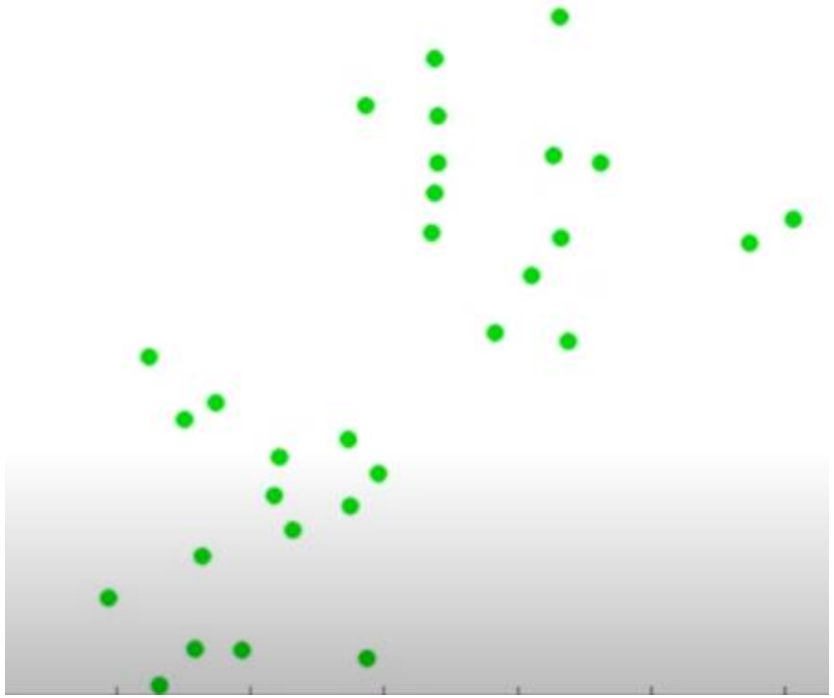
- Let’s go through the steps together.
- We would like to group the data into 2 clusters:



Unlabeled!

K-Means (illustration)

- Let’s go through the steps together.
- We would like to group the data into 2 clusters:

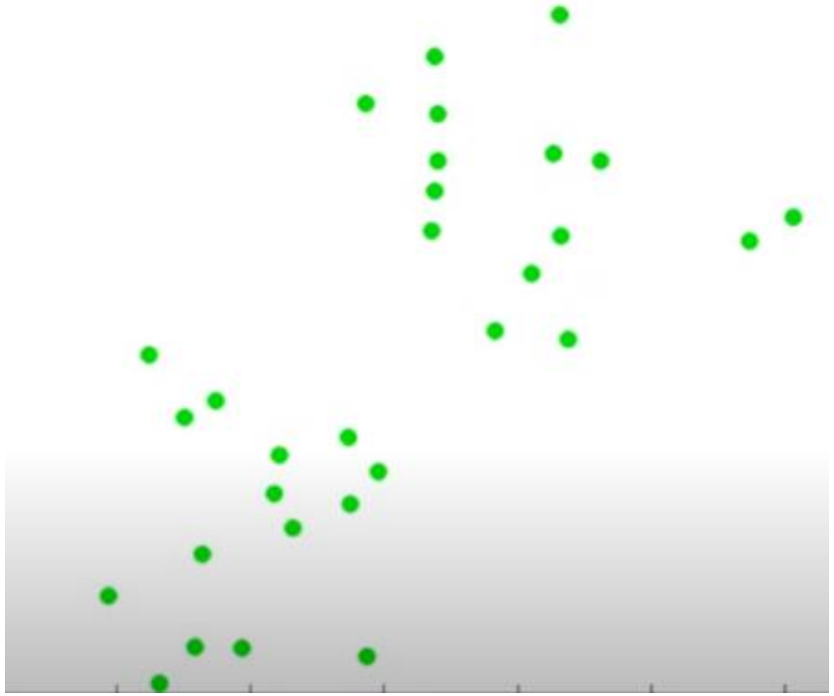


Unlabeled!

Randomly initialize by selecting 2 centroids

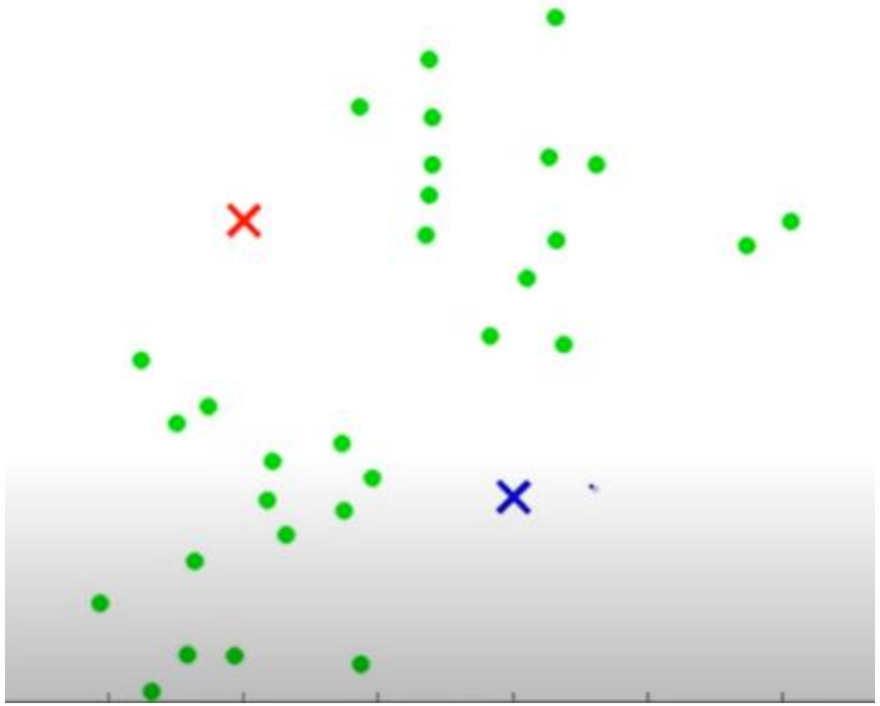
K-Means (illustration)

- Let’s go through the steps together.
- We would like to group the data into 2 clusters:



Unlabeled!

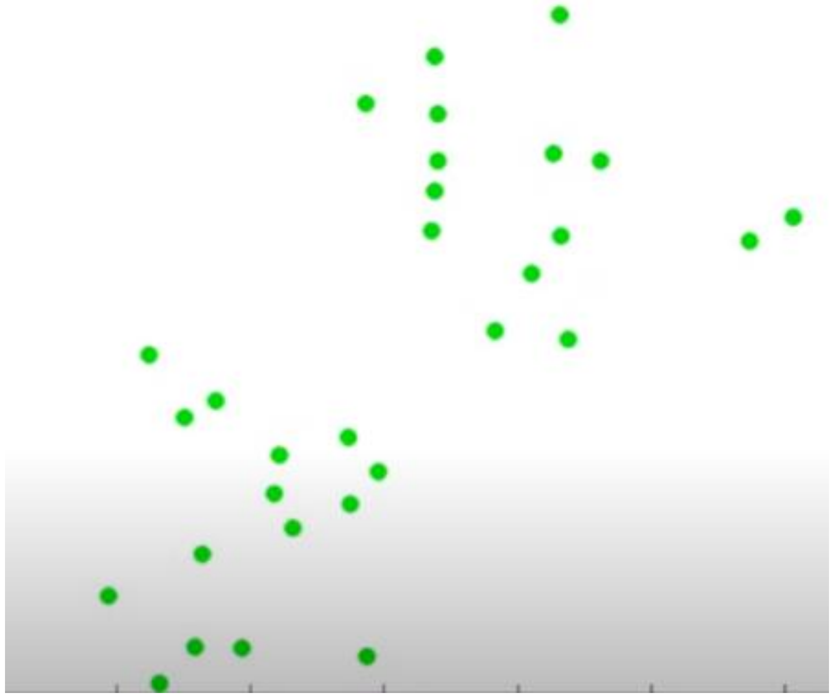
Randomly initialize by selecting 2 centroids



K-Means (illustration)

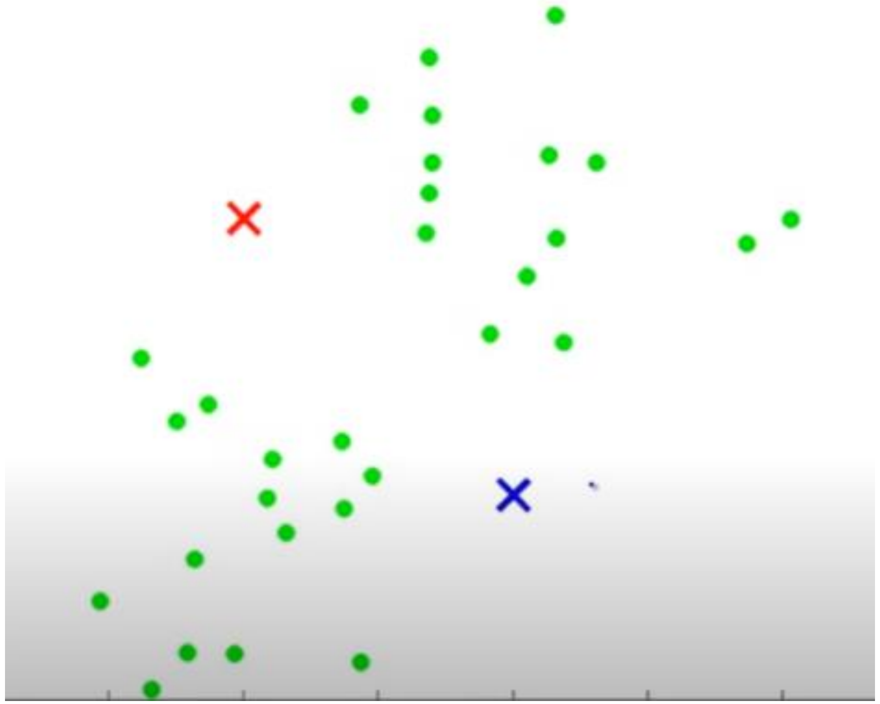
How to begin?

- Let’s go through the steps together.
- We would like to group the data into 2 clusters:



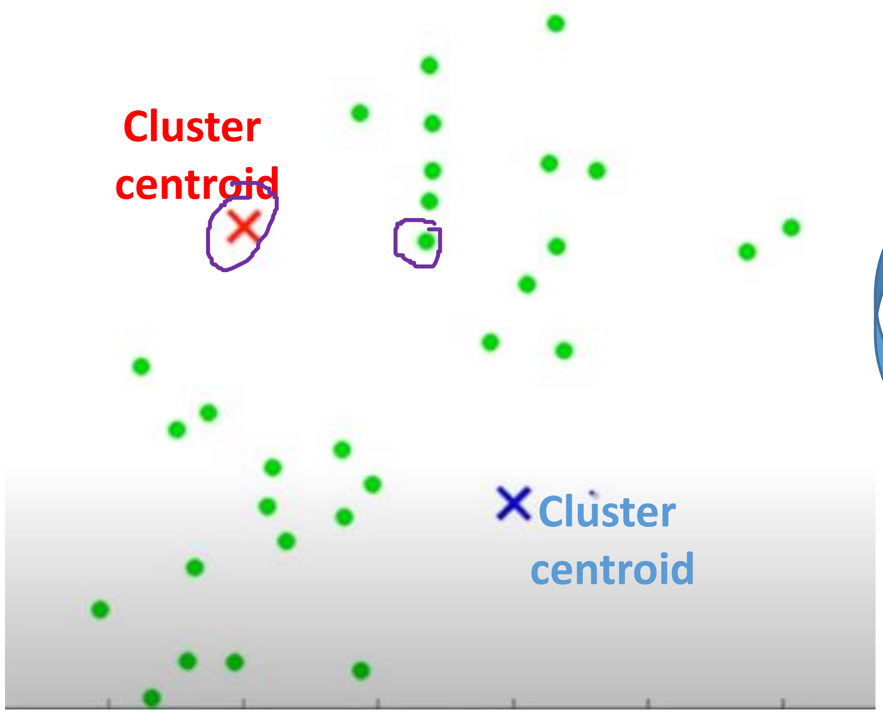
Unlabeled!

Randomly initialize by selecting 2 centroids



K-Means (illustration)

- We have 2 cluster centroids, because we would like to group the data into 2 clusters.



K-Means is an iterative algorithm and it does 2 things:

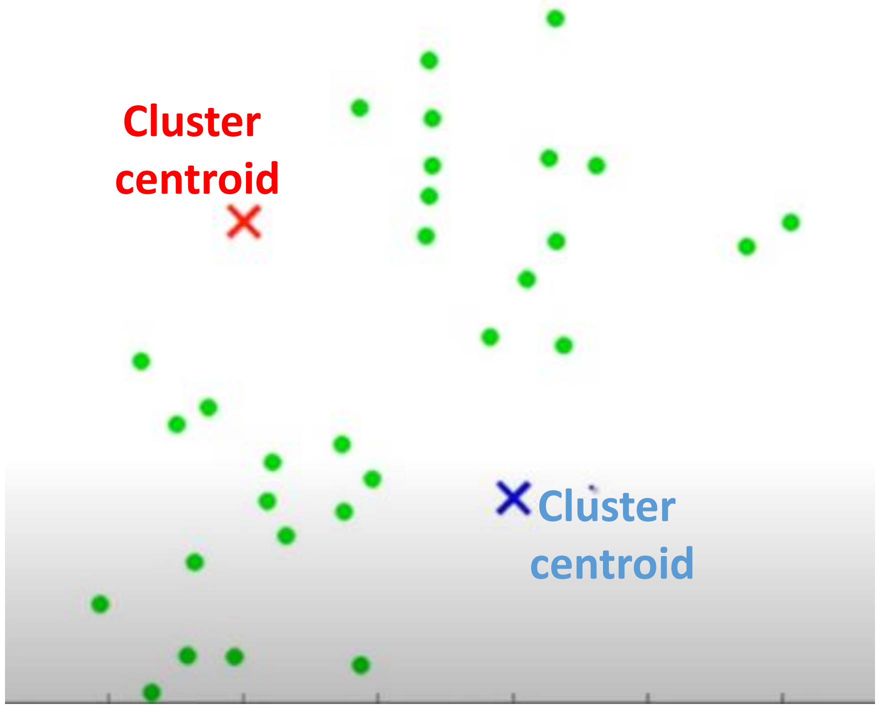
1. **Cluster assignment step:** algorithm will go through each of the examples (green dots) and depending on whether it is closer to red cluster centroid, or blue; algorithm will assign each of the data points in blue or red cluster.

2. **Move centroid step:** calculate the mean of the new clusters, and move the cluster centroids accordingly.

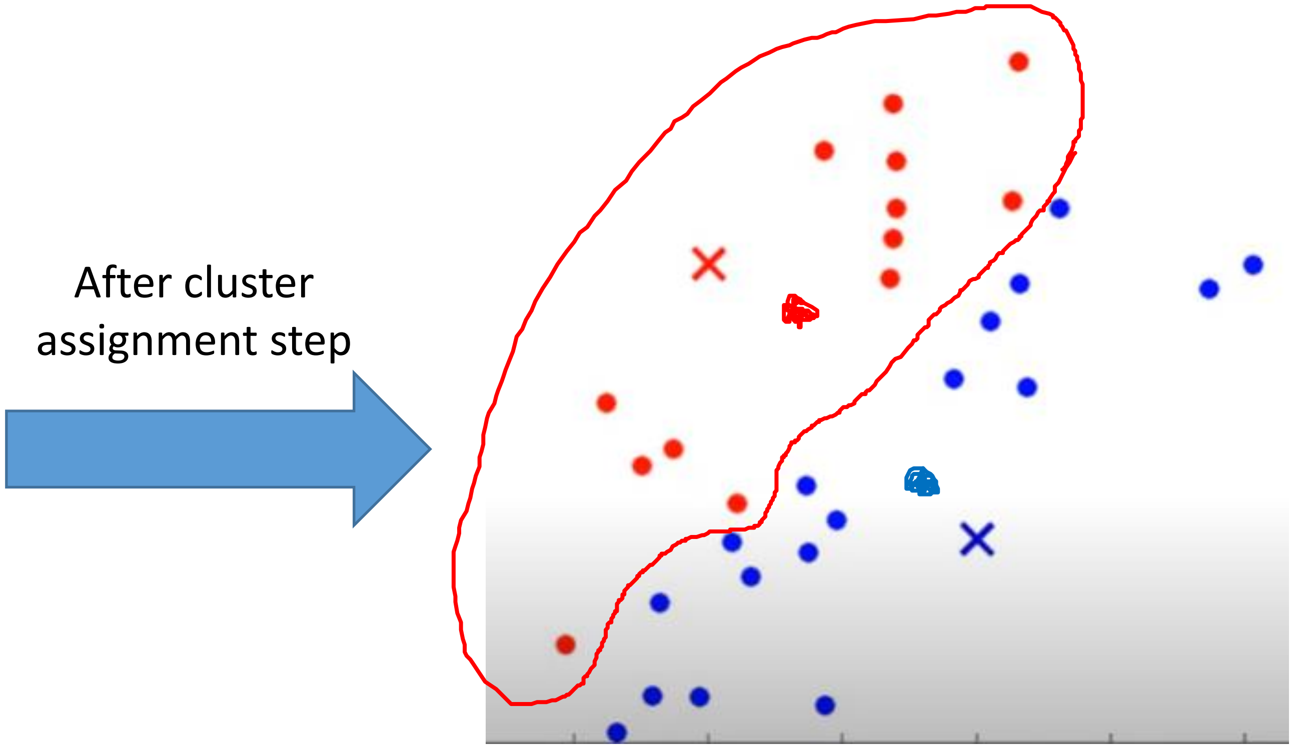
Step 1 (round1)

K-Means (illustration)

K-Means is an iterative algorithm and it does 2 things:
Step 1. Cluster assignment step: algorithm will go through each of the examples (green dots) and depending on whether it is closer to red cluster centroid, or blue; algorithm will assign each of the data points in blue or red cluster.



After cluster assignment step

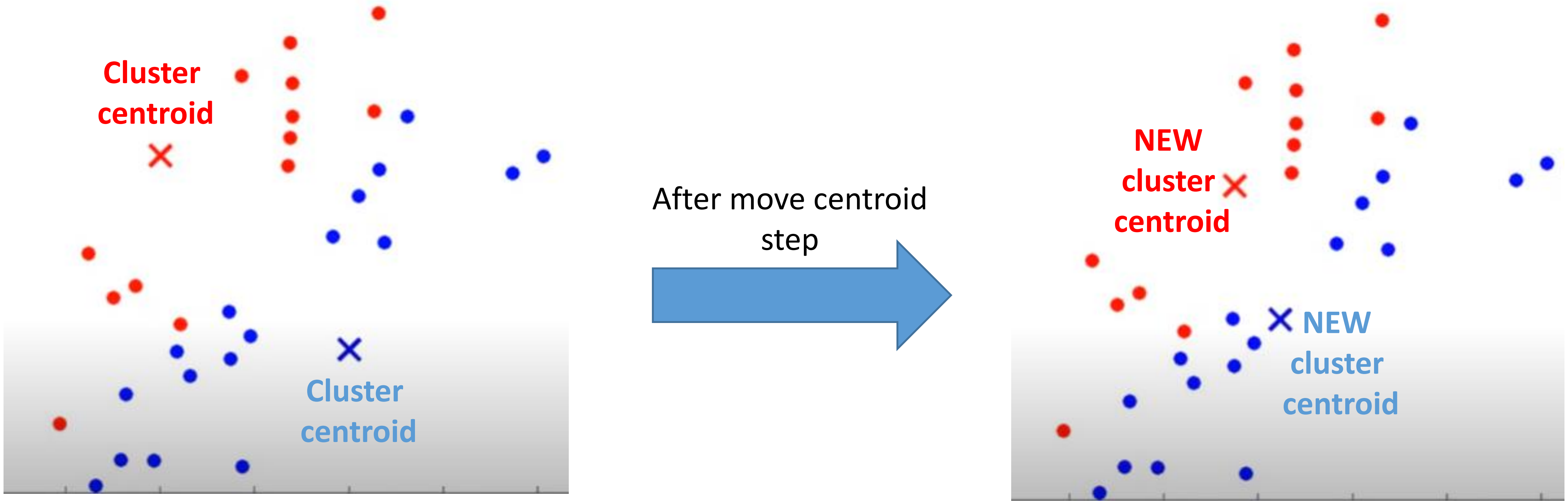


cluster centroids do NOT change in this step!

Step 2 (round1)

K-Means (illustration)

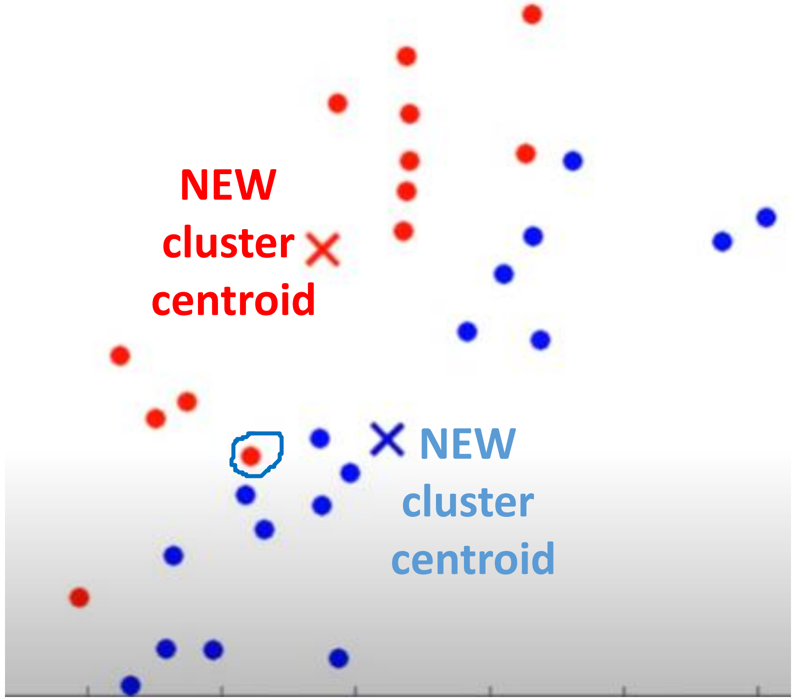
K-Means is an iterative algorithm and it does 2 things:
Step 2. Move centroid step: calculate the mean of the new clusters, and move the cluster centroids accordingly.

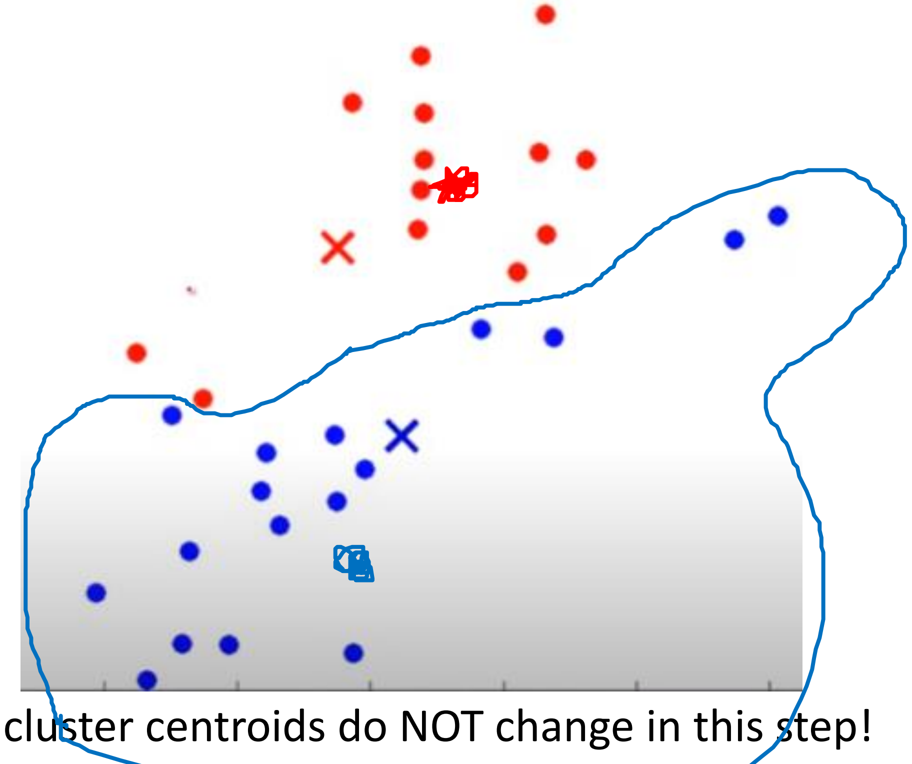


Step 1 (round 2)

K-Means (illustration)

K-Means is an iterative algorithm and it does 2 things:
Step 1. Cluster assignment step: algorithm will go through each of the examples and depending on whether it is closer to red cluster centroid, or blue; algorithm will assign each of the data points in blue or red cluster.



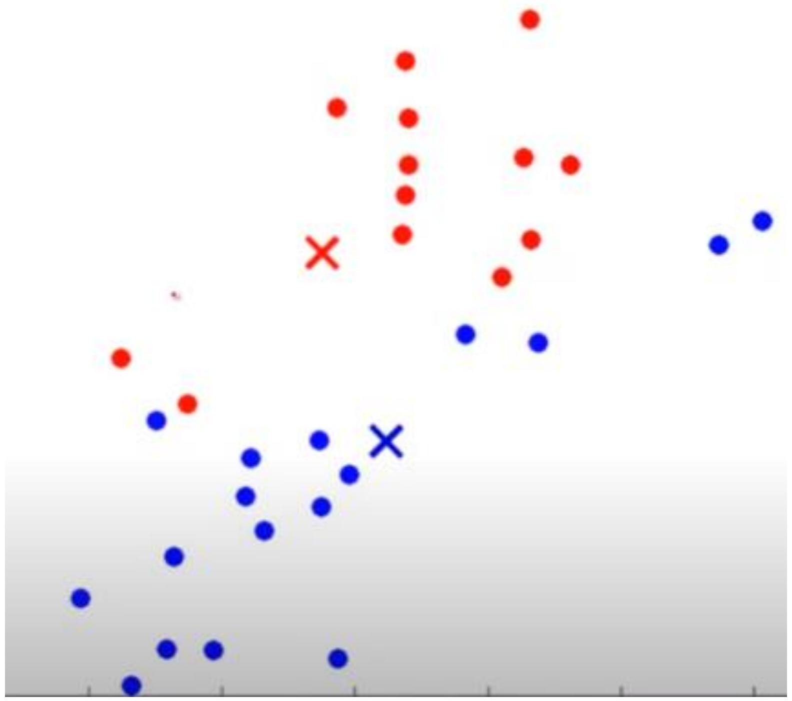


cluster centroids do NOT change in this step!

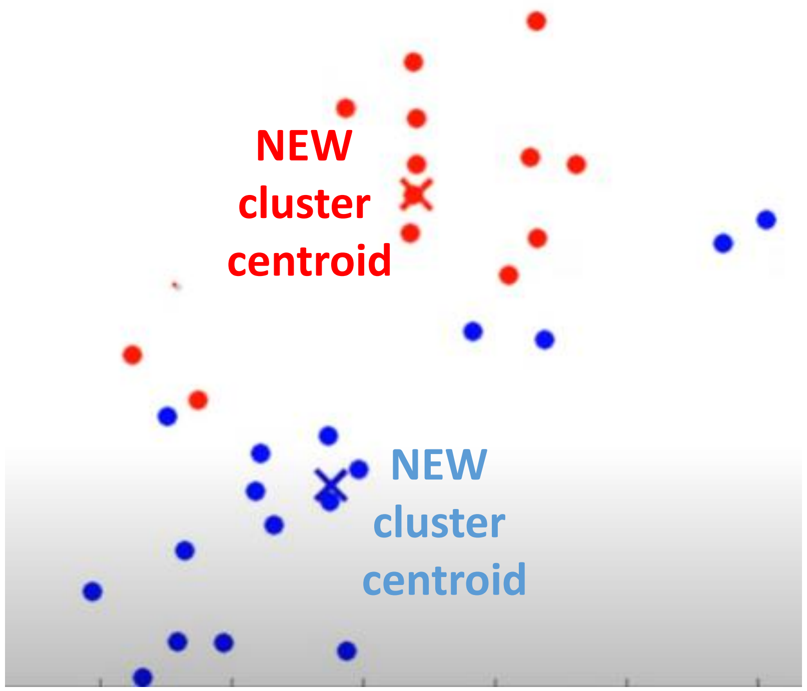
Step 2 (round 2)

K-Means (illustration)

K-Means is an iterative algorithm and it does 2 things:
Step 2. Move centroid step: calculate the mean of the new clusters (for blue and red points separately), and move the cluster centroids accordingly.



After move centroid step

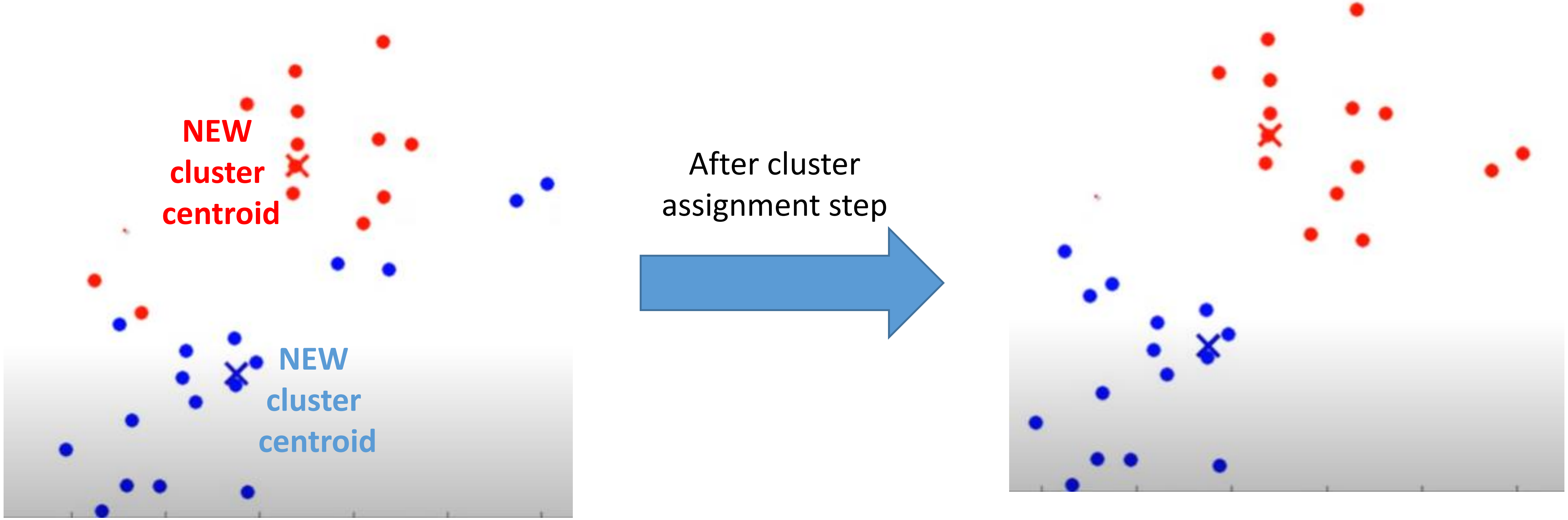


K-Means (illustration)

Step 1 (round 3)

K-Means is an iterative algorithm and it does 2 things:

Step 1. Cluster assignment step: algorithm will go through each of the examples and depending on whether it is closer to red cluster centroid, or blue; algorithm will assign each of the data points in blue or red cluster.



NEW cluster centroid

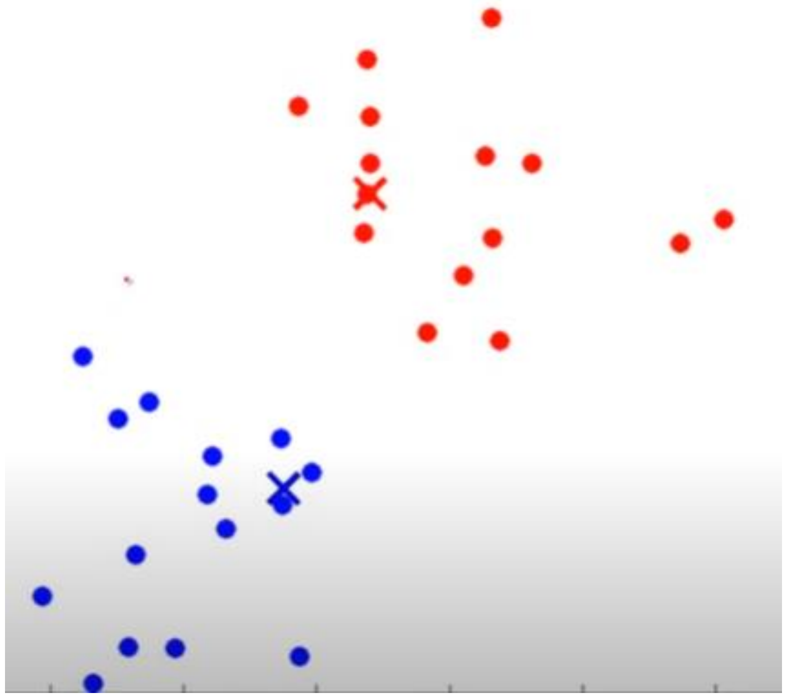
NEW cluster centroid

After cluster assignment step

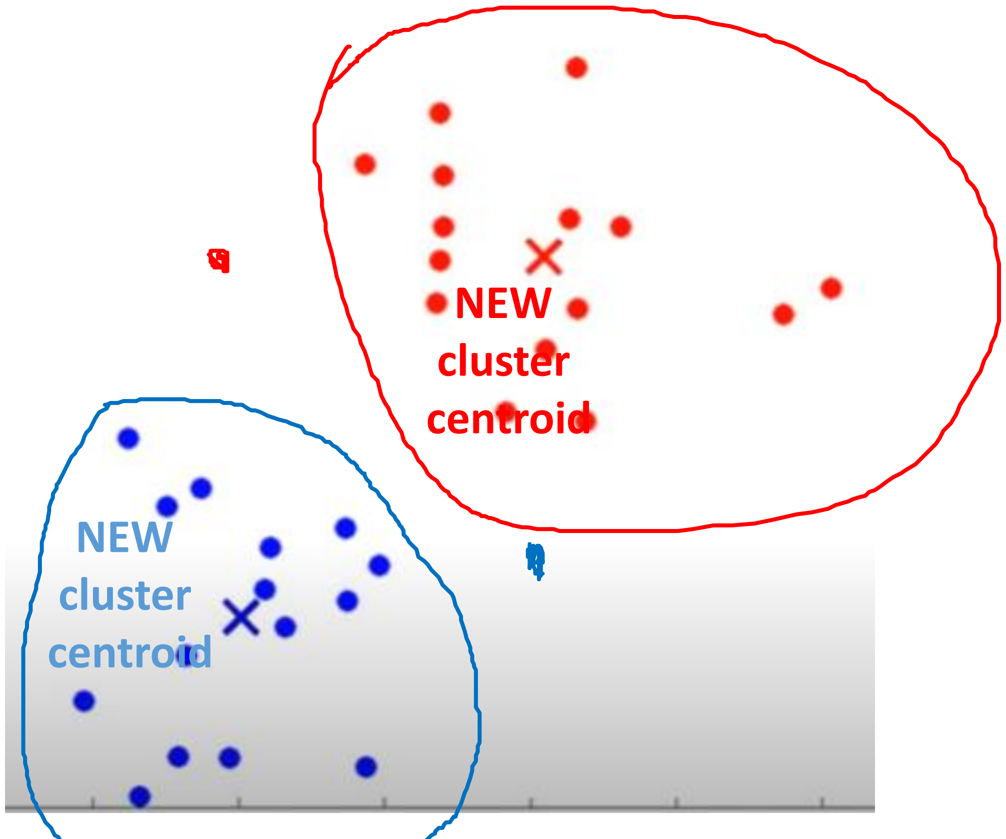
Step 2 (round 3)

K-Means (illustration)

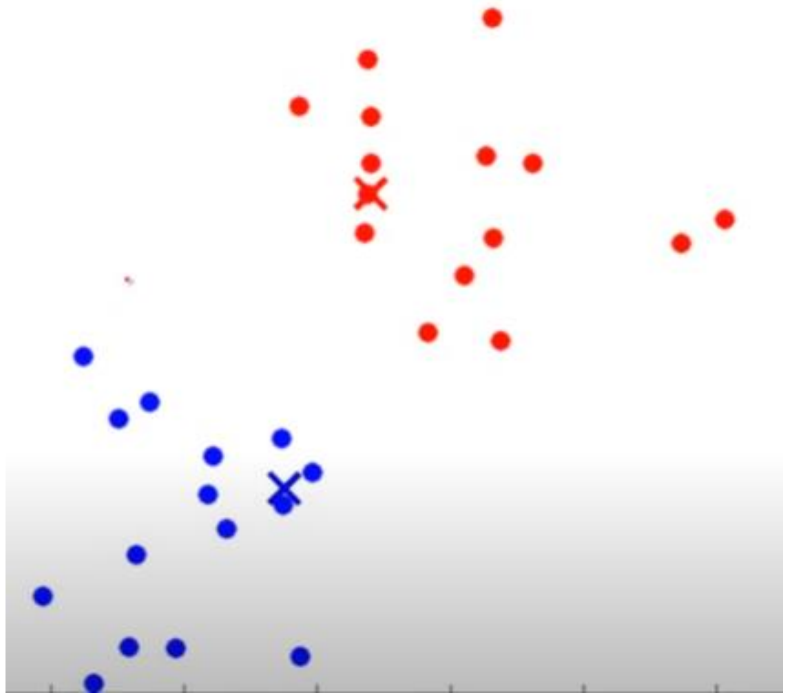
K-Means is an iterative algorithm and it does 2 things:
Step 2. Move centroid step: calculate the mean of the new clusters (for blue and red points separately), and move the cluster centroids accordingly.



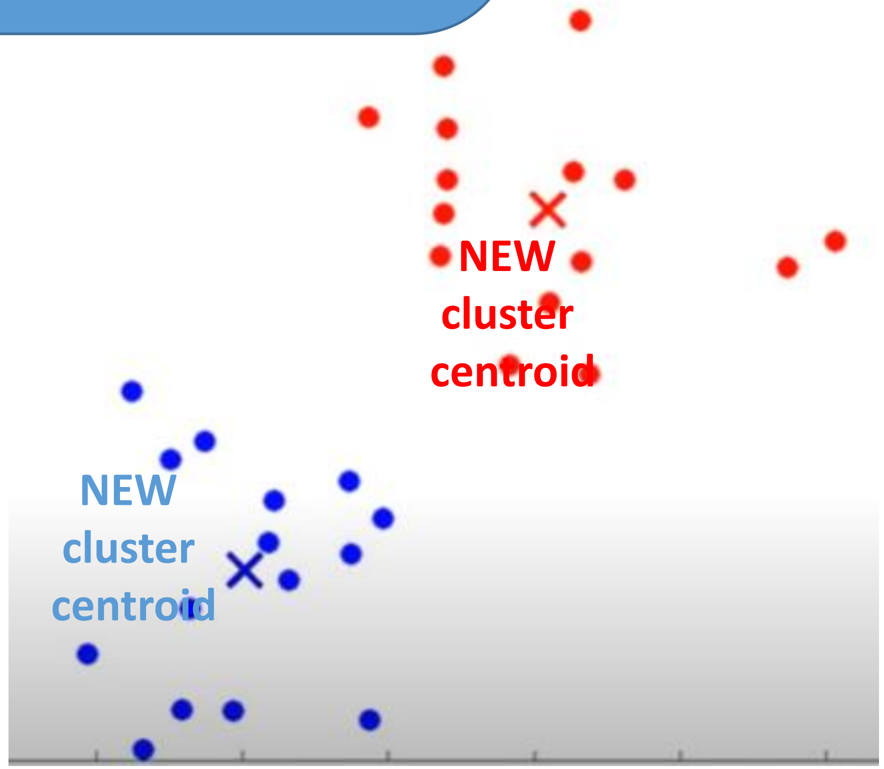
After move centroid step



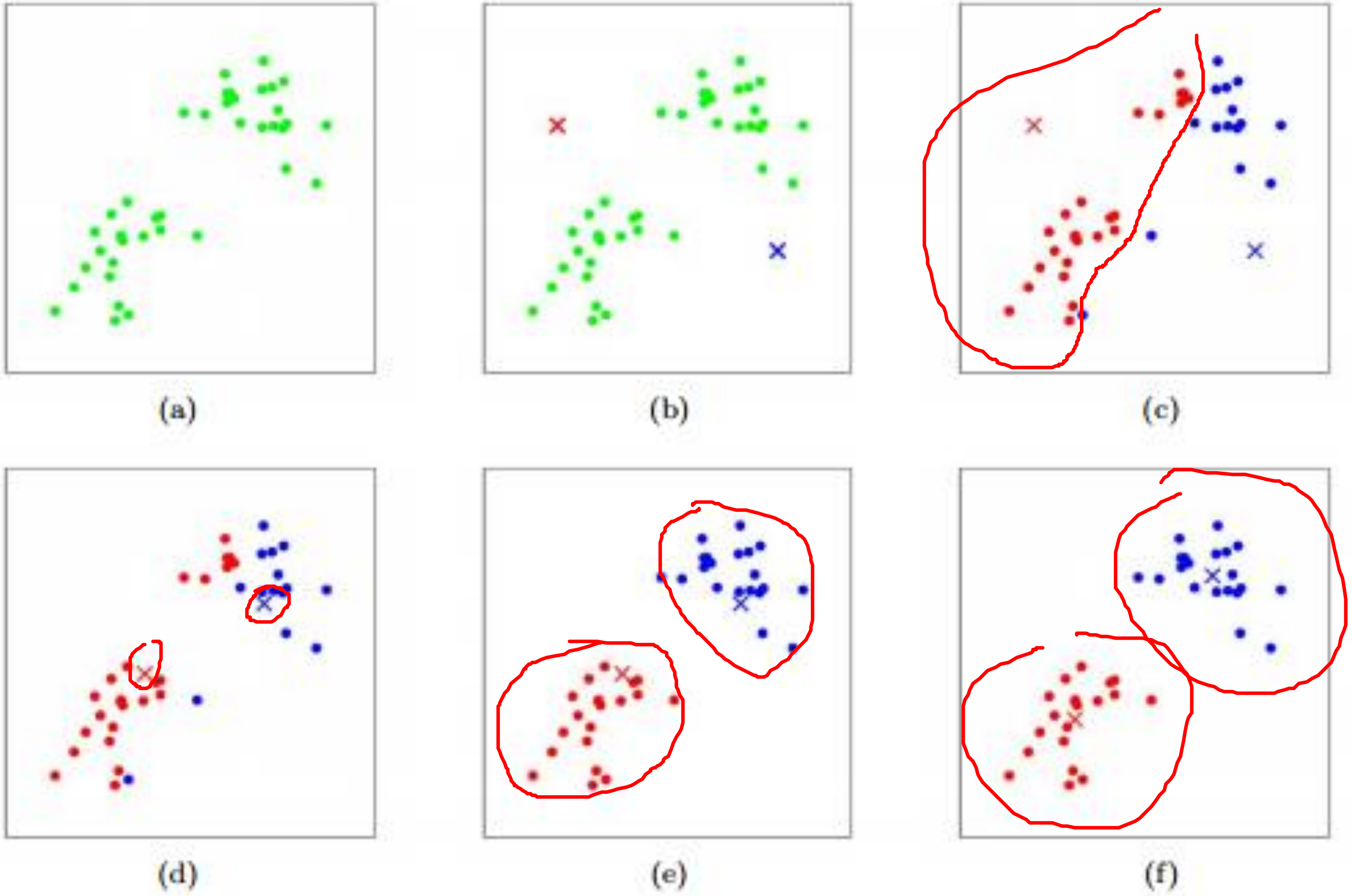
AND WE ARE DONE… IF YOU KEEP RUNNING K-MEANS INTERATIONS, CENTROIDS AND CLUSTER ASSIGNMENTS WILL NOT CHANGE.



After move centroid step

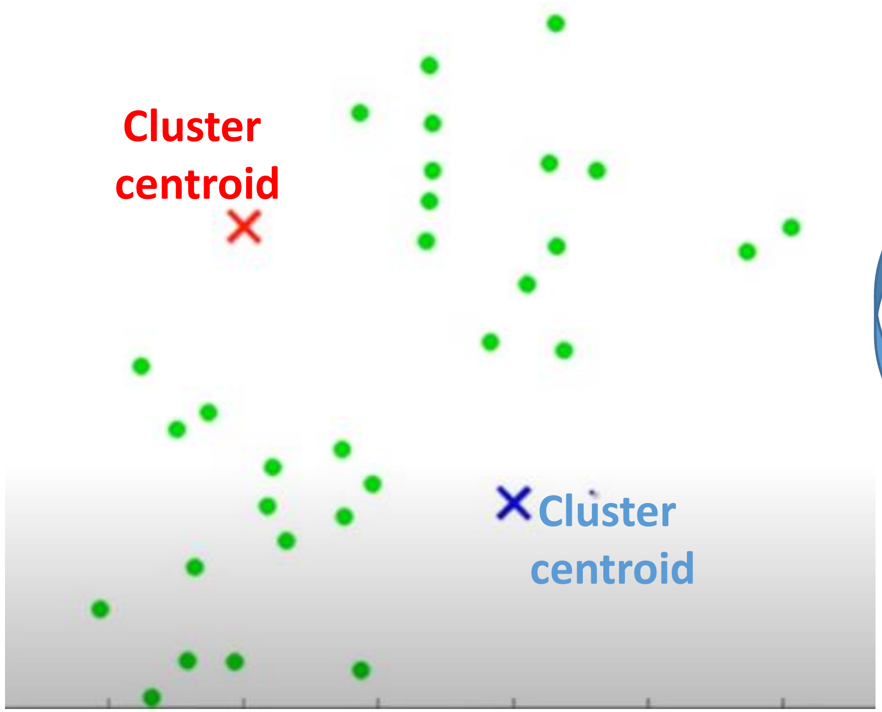


K-Means finds the best centroids by alternating between (1) assigning data points to clusters based on the current centroids (2) choosing centroids (points which are the center of a cluster) based on the current assignment of data points to clusters.



K-Means (illustration)

- We have 2 cluster centroids, because we would like to group the data into 2 clusters.



K-Means is an **iterative algorithm** and it does 2 things:

1. **Cluster assignment step:** algorithm will go through each of the examples (green dots) and depending on whether it is closer to red cluster centroid, or blue; algorithm will assign each of the data points in blue or red cluster.

2. **Move centroid step:** calculate the mean of the new clusters, and move the cluster centroids accordingly.