Machine Learning

DATA SCIENCE CLUB

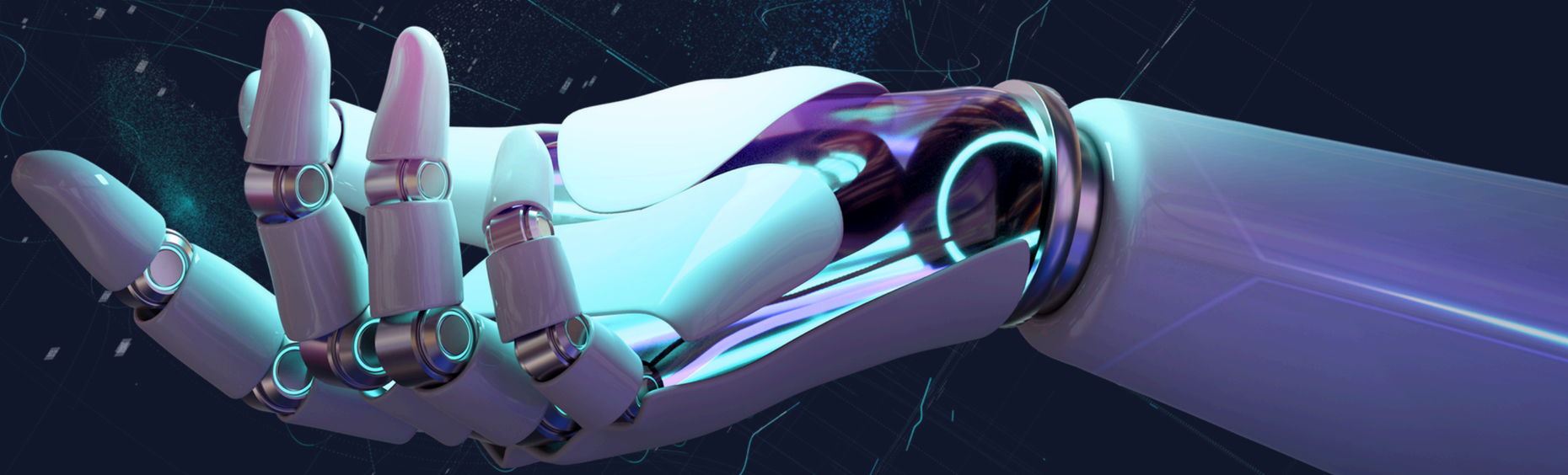
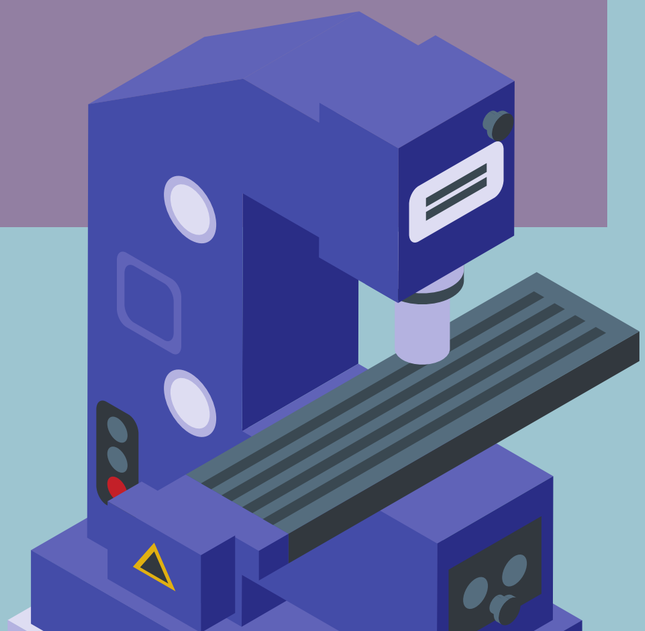


Table of contents



Supervised and unsupervised learning is based on the type of data being used

Supervised

- Labelled data
- Uses an algorithm to train itself to minimize error
- Has a baseline of what the “correct” accepted value is

Unsupervised

- Data not labelled
- Finds relation in the data’s structure without guidance

Supervised

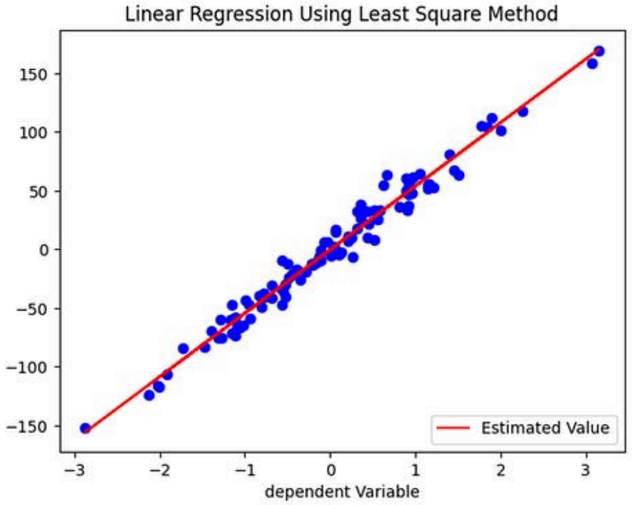
- Goal: predict the relationship between the input and output data
- Uses: classification, regression, etc.
- Example: sentiment analysis, pricing changes, etc.

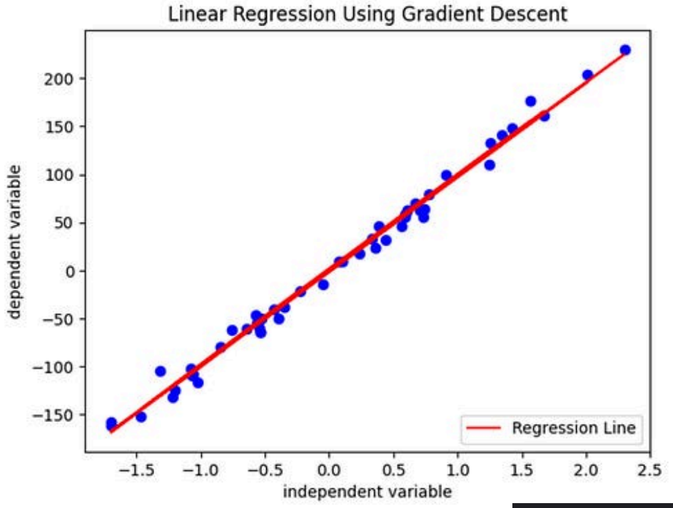
Unsupervised

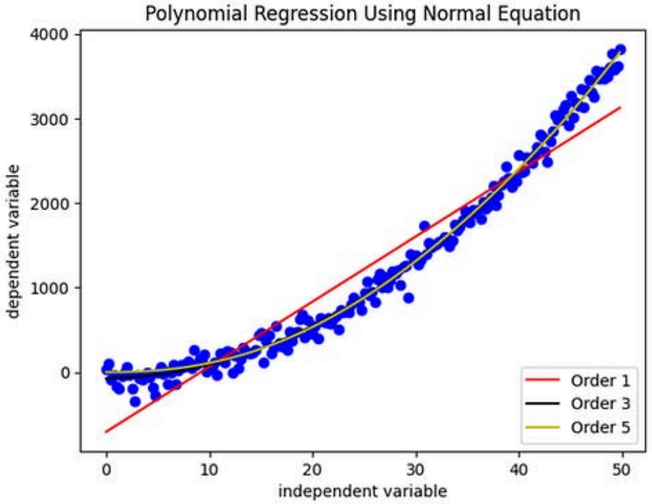
- Goal: find relationships between given data
- Uses: Exploratory data analysis, clustering, etc.
- Example: customer segmentation, natural language processing

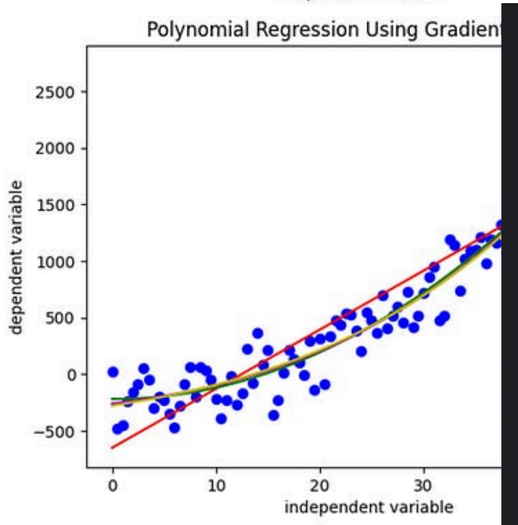
Regression Analysis

Supervised









Formula

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable

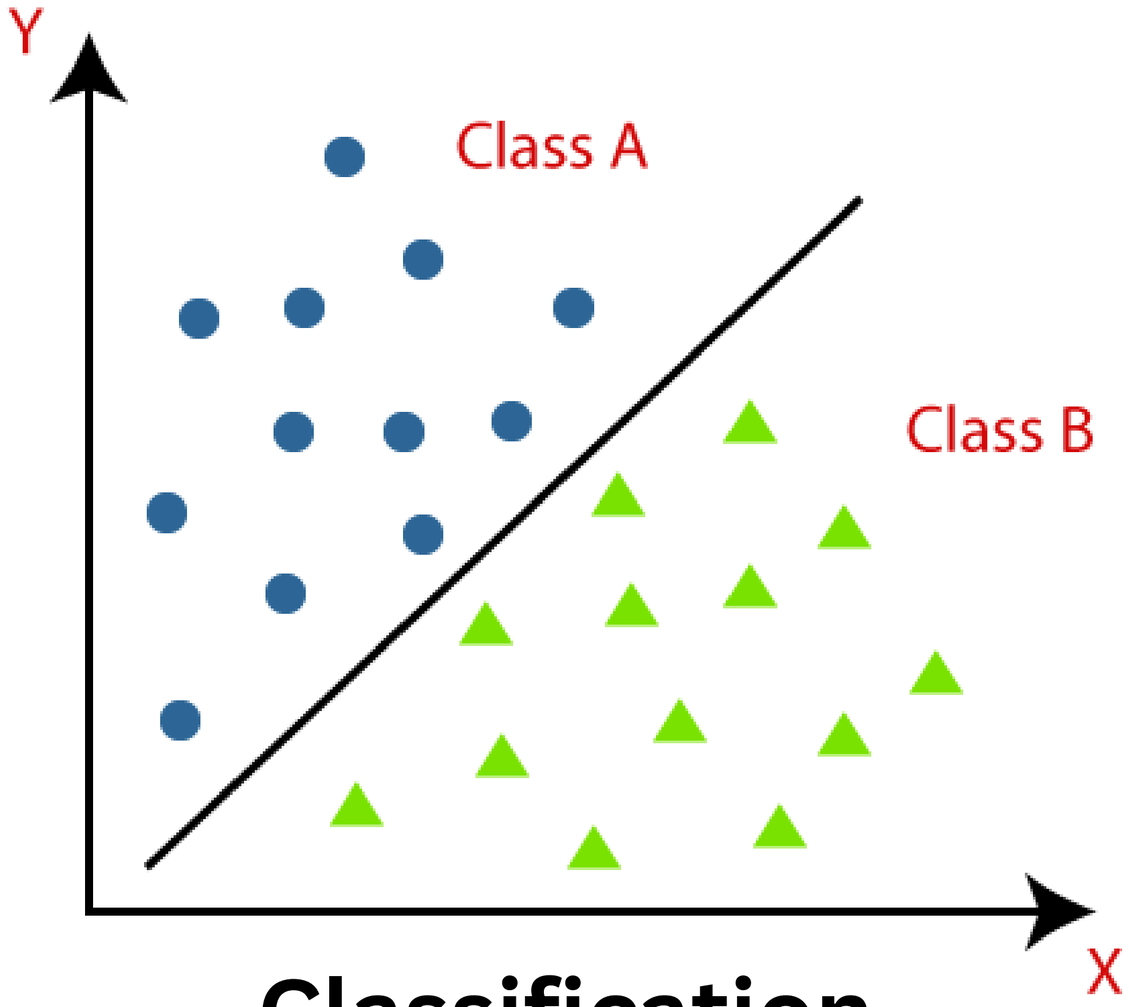
f = function

X_i = independent variable

β = unknown parameters

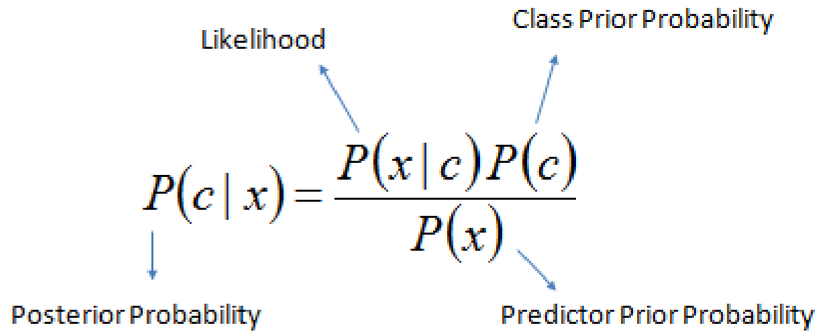
e_i = error terms

Supervised



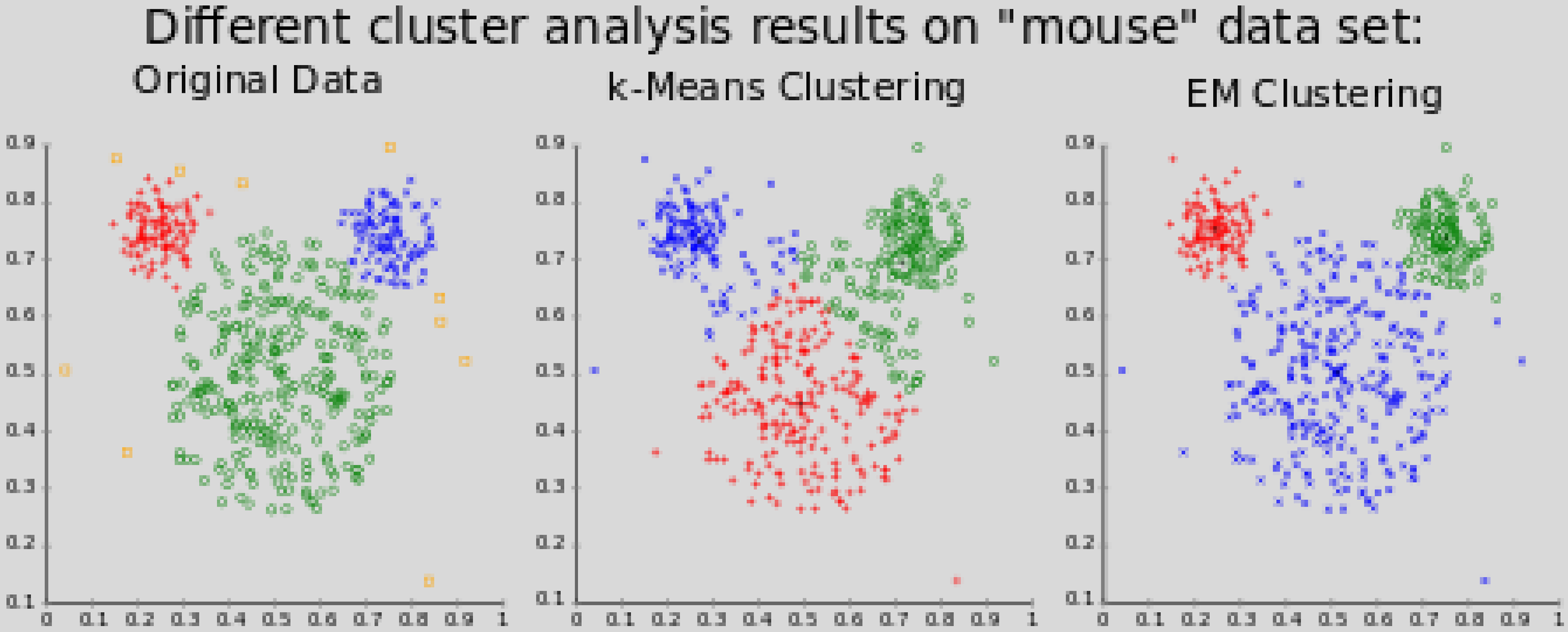
Classification

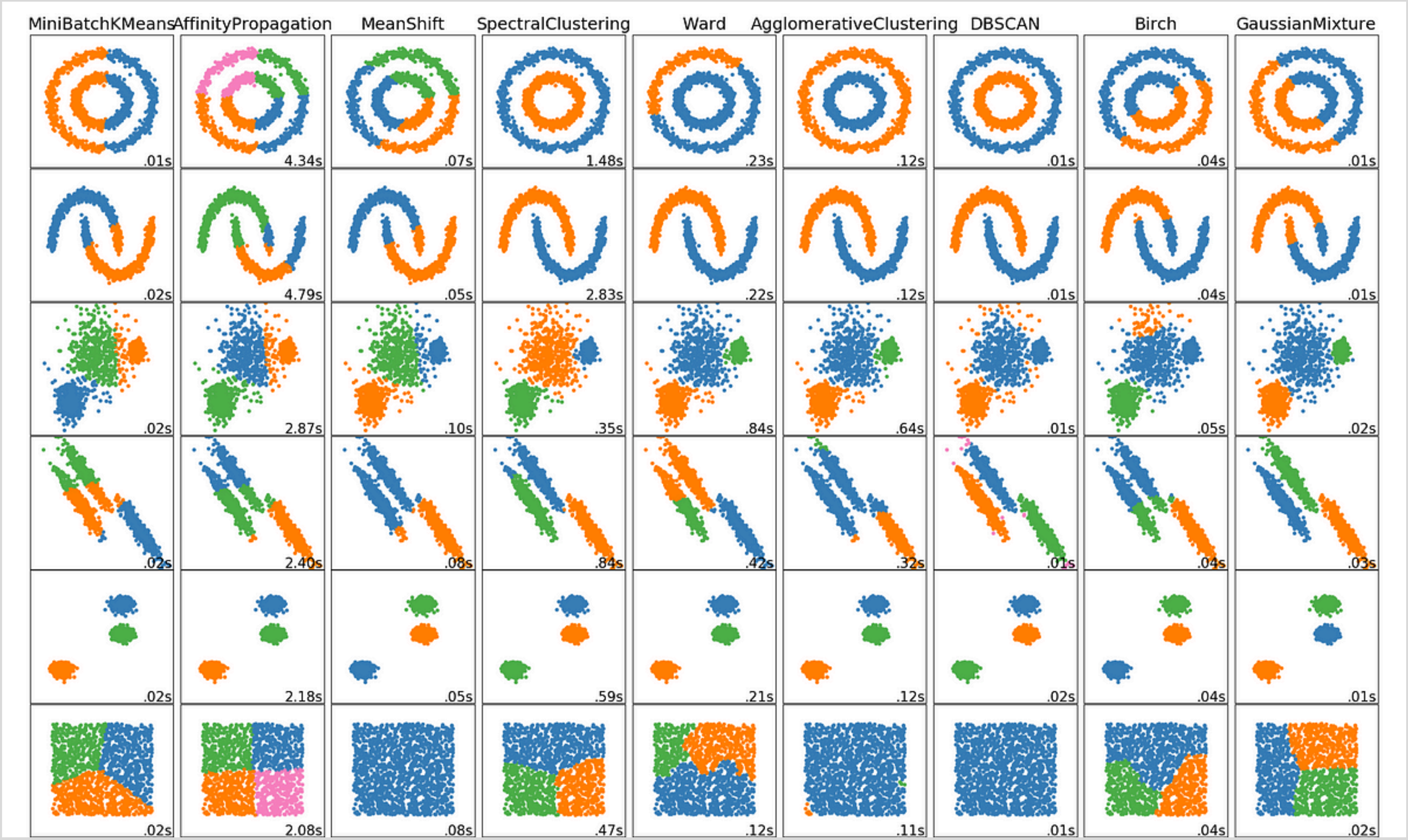
$$\mathrm{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{r=1}^{d} |x_r - z_r|^p \right)^{1/p}.$$



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

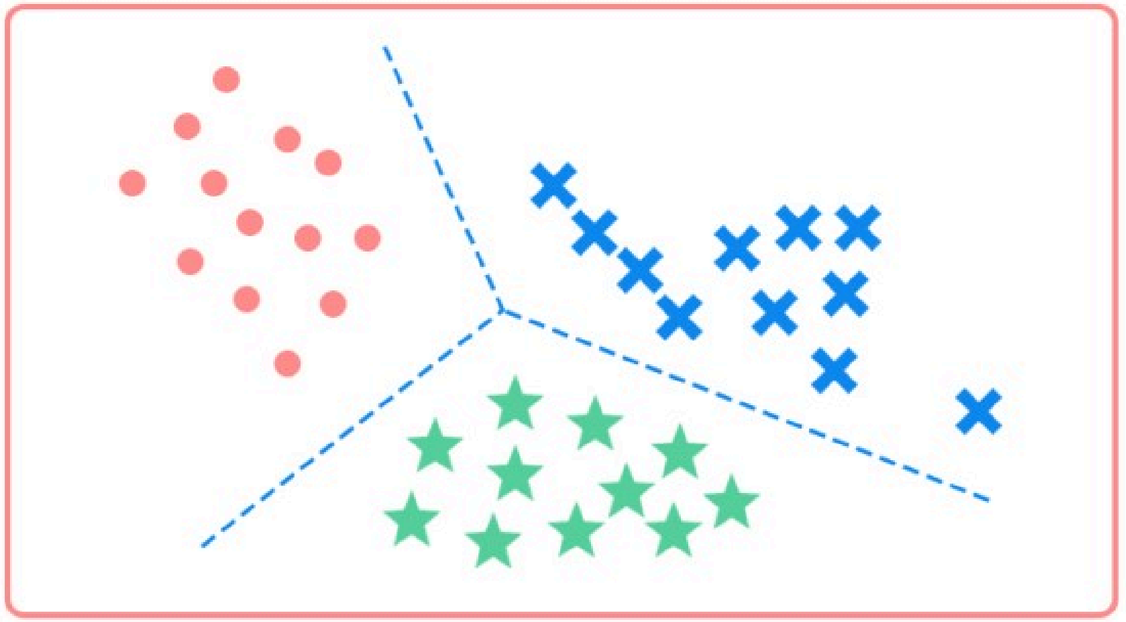
Unsupervised




MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s

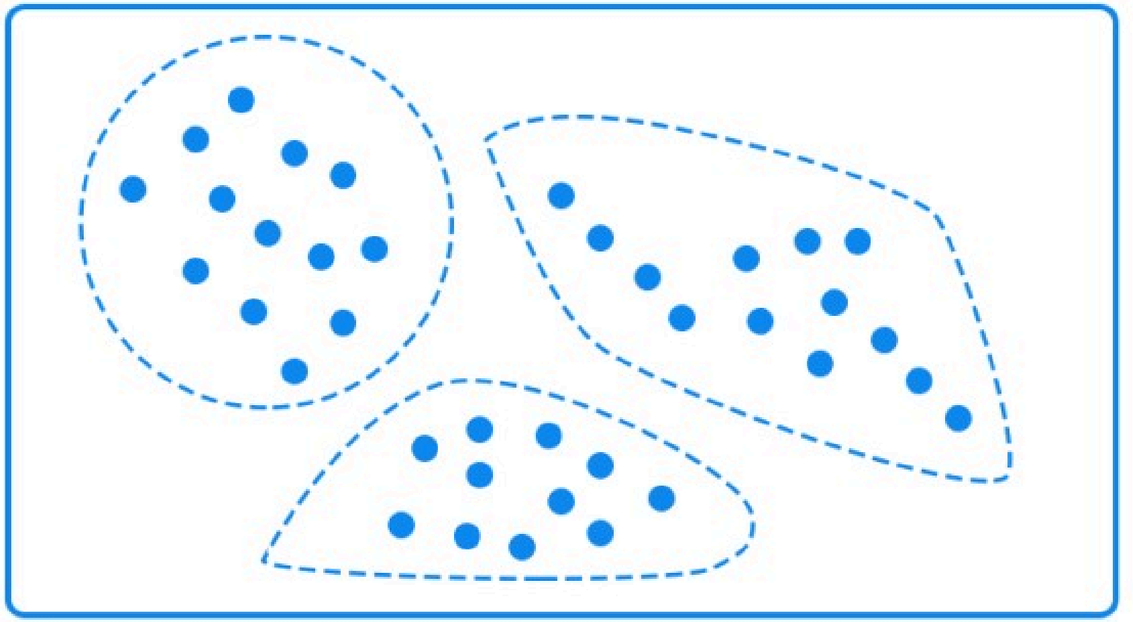
Supervised vs. Unsupervised Learning

Classification



Supervised learning

Clustering



Unsupervised learning

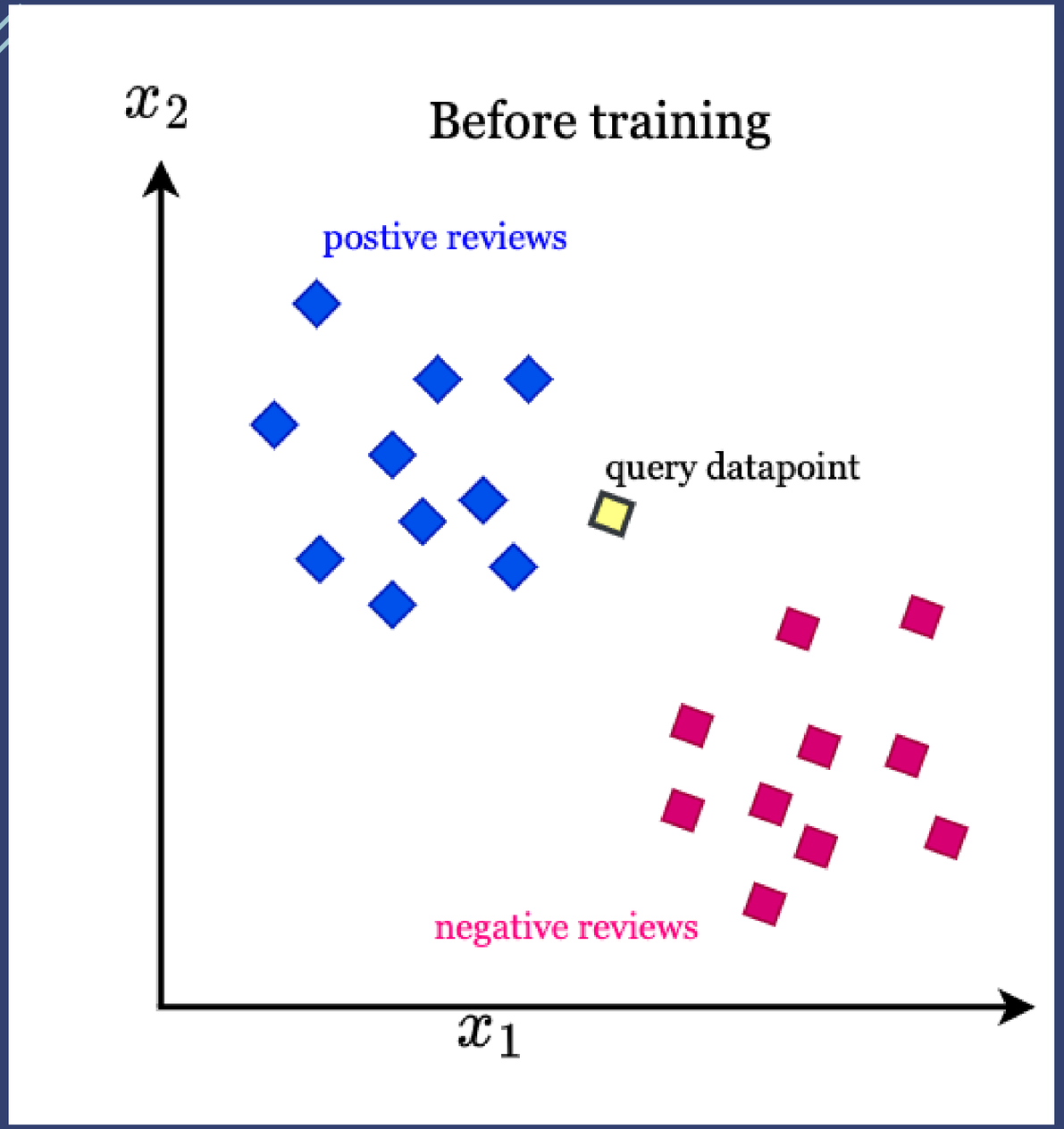
K Nearest Neighbours - Overview

It is a supervised ML algorithm that uses classification.

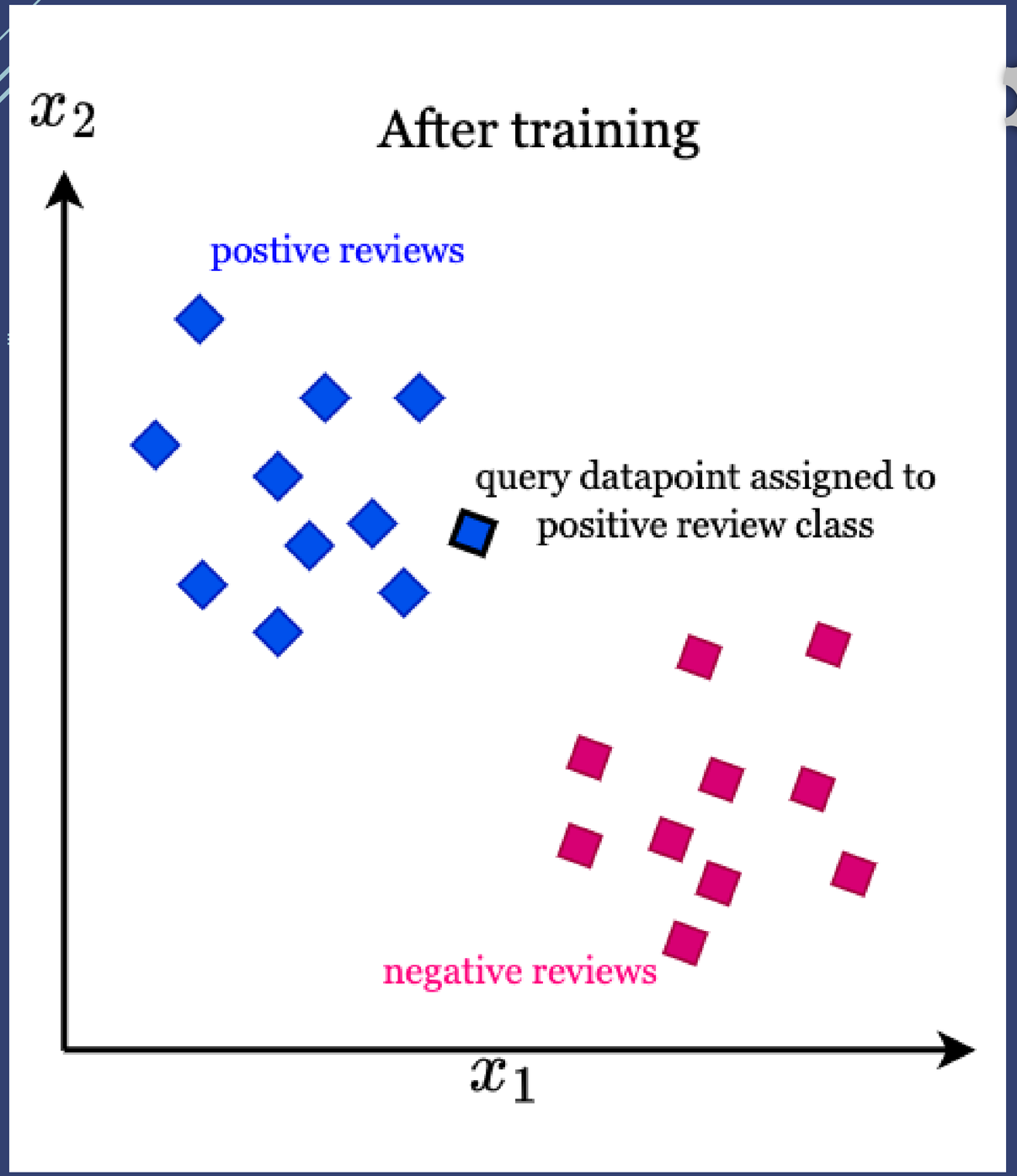
You are given a data set → **In the dataset, the data points are classified into groups** → **Now, a new data point in inserted** → **KNN uses its k-value and finds how many data points of each category fall into the circle of k values**

Let's See a Visual Demo!

Data Set Before KNN



Data Set After KNN



So.... How does the math work?

KNN has 3 main steps

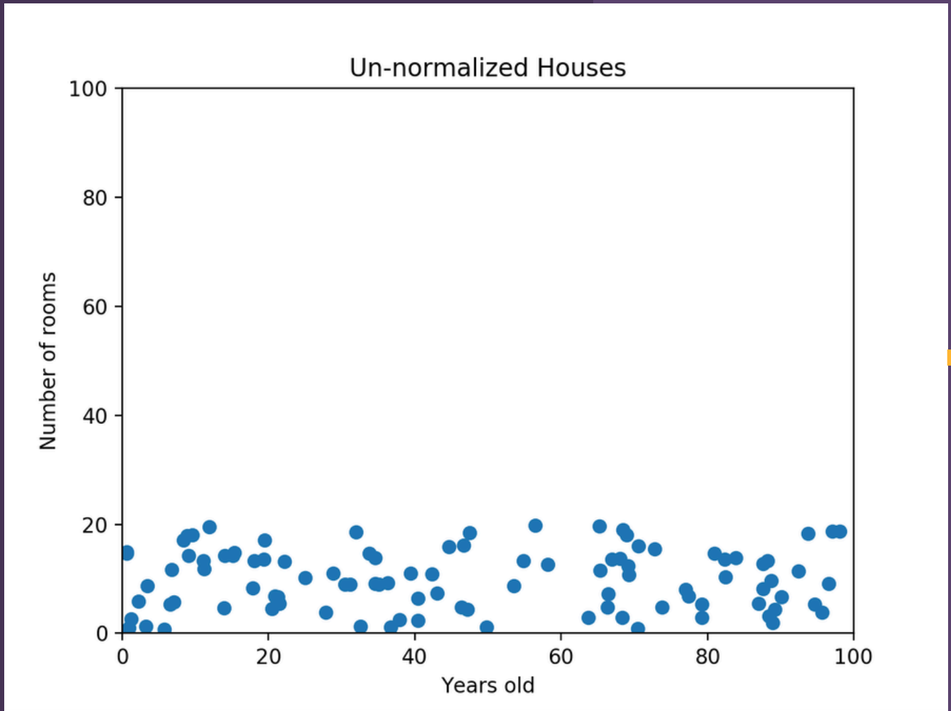
Step 1 Data Normalization

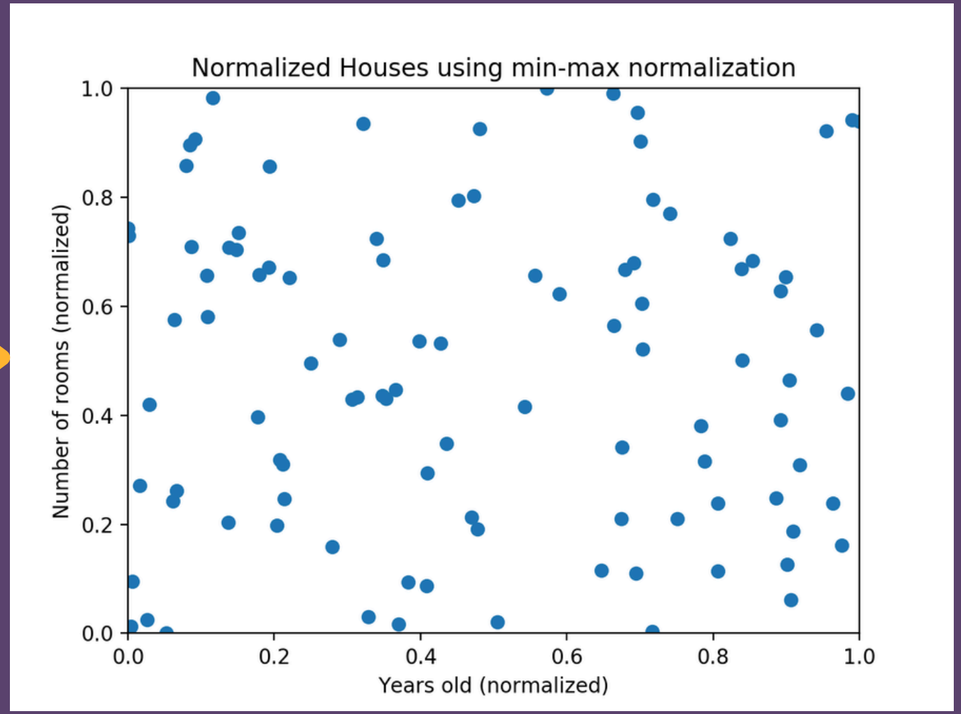
Step 2 Distance Formula

Step 3 Categorize using top k neighbours

Data Normalization

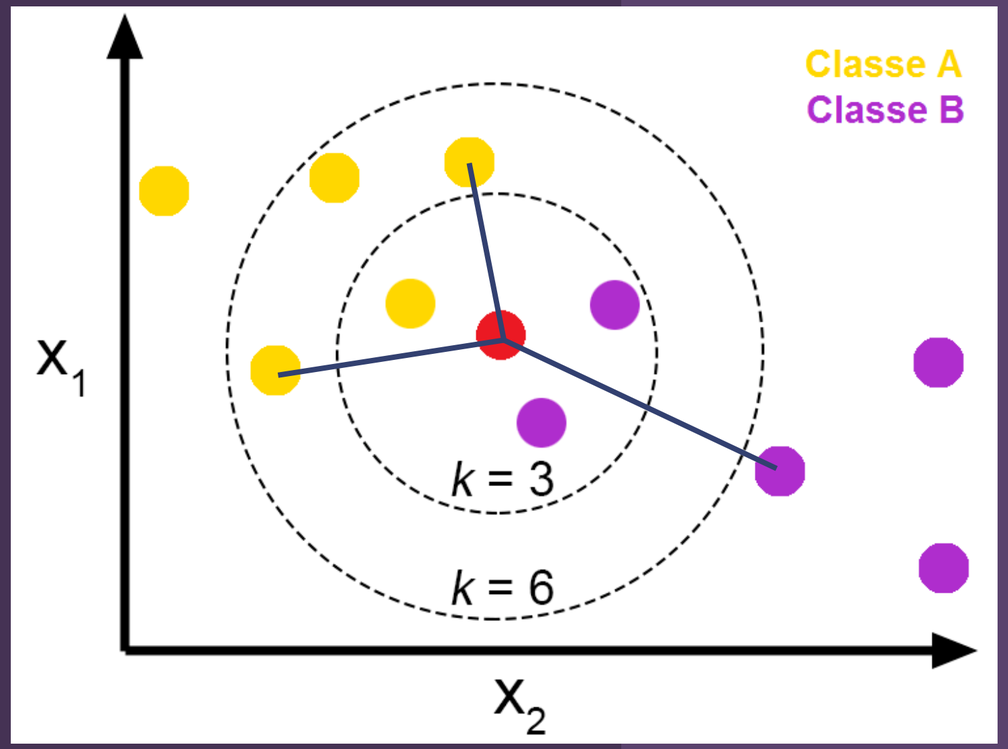
Sometimes, the data set has really different scales. This would affect our distance formula in step 2.





Using the Distance Formula

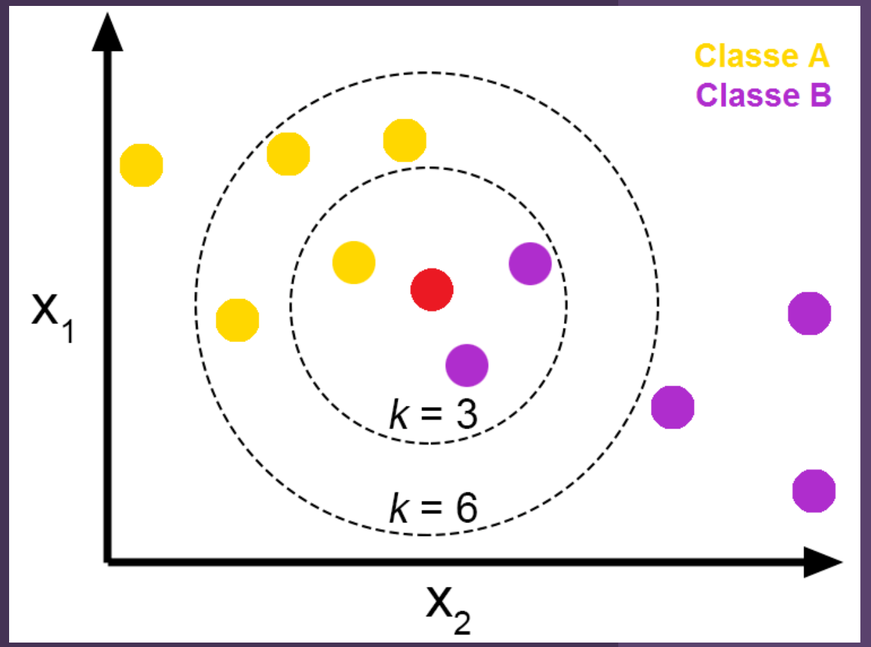
Now, the new data point is introduced. The distance from each point to the new data point is measured and arranged in ascending order.



What do you think the KNN would do for this graph?

Using K Value and Classifying

The k-value is now used. The top k distances are separated. In these top k values, we see how many data points of each group exist. The new data point is classified into the group with more data points.



If k=3, then the top three closest data will be chosen

Then, in those there are 2 PURPLE and 1 YELLOW data points

So, the RED (unknown) point will be classified as PURPLE