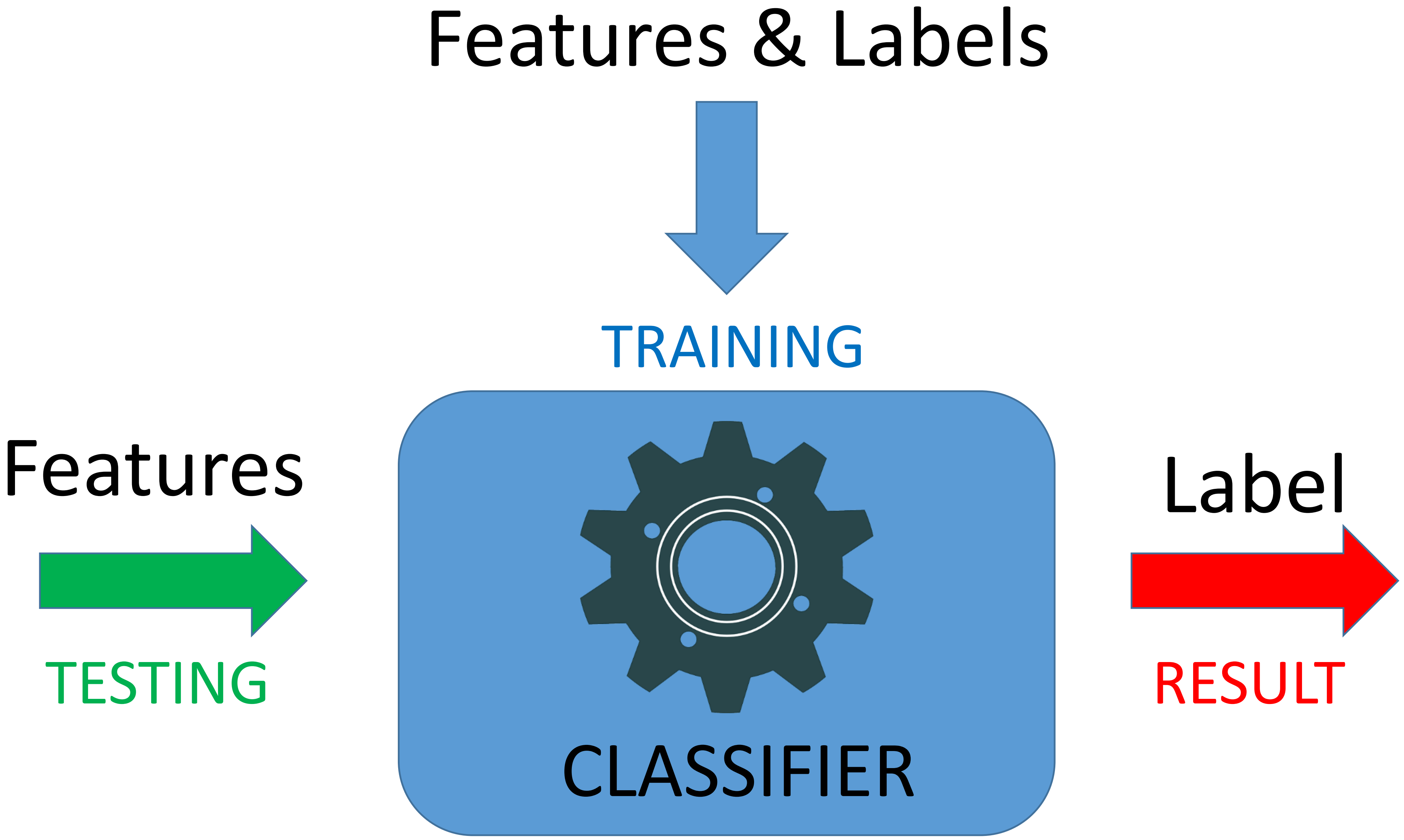
ACADEMY OF INNOVATIVE EDUCATION

DAY 03

INTRODUCTION TO ML CLASSIFIERS - II

KNN Classifier

Thakshila Dasun
BSc. Eng in Mechatronics Eng
CIMA (UK)



Features & Labels

TRAINING

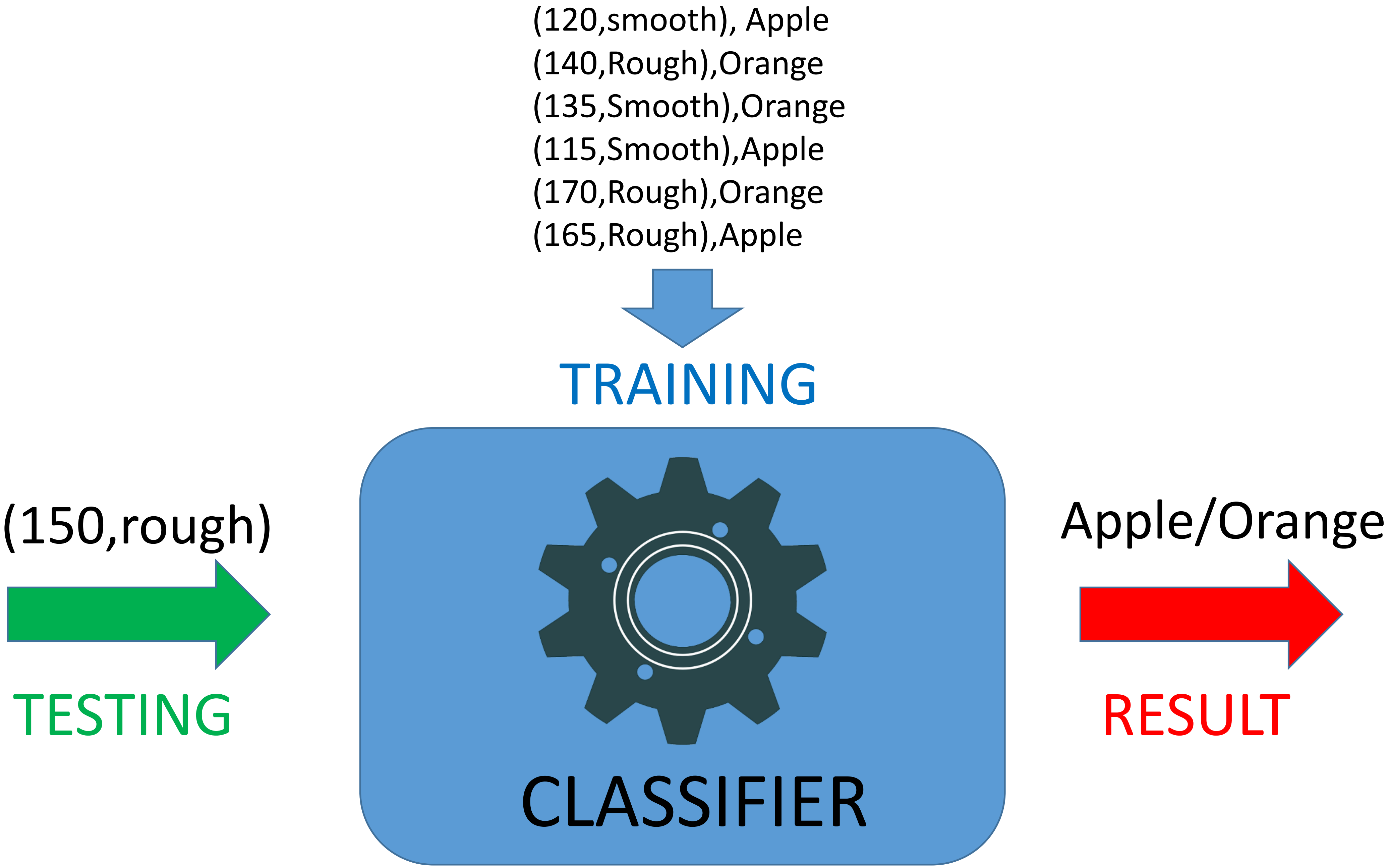
Features

TESTING

CLASSIFIER

Label

RESULT



(120,smooth), Apple
(140,Rough),Orange
(135,Smooth),Orange
(115,Smooth),Apple
(170,Rough),Orange
(165,Rough),Apple

TRAINING

(150,rough)

TESTING

CLASSIFIER

Apple/Orange

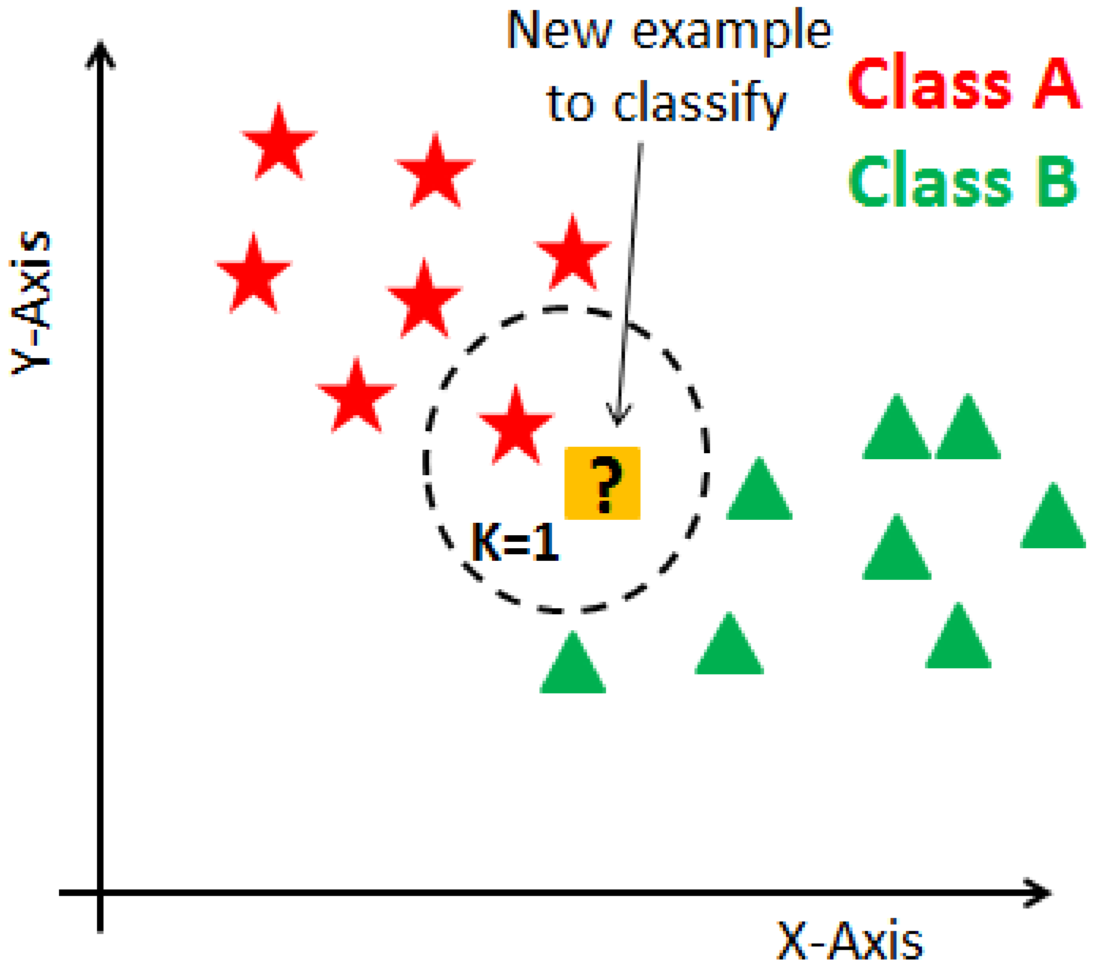
RESULT

KNN Classifier(1)

- K Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms.
- KNN used in the variety of applications such as finance, healthcare, political science, handwriting detection, image recognition and video recognition.
- In Credit ratings, financial institutes will predict the credit rating of customers
- In loan disbursement, banking institutes will predict whether the loan is safe or risky. In political science, classifying potential voters in two classes will vote or won’t vote.
- KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

KNN Classifier(2)

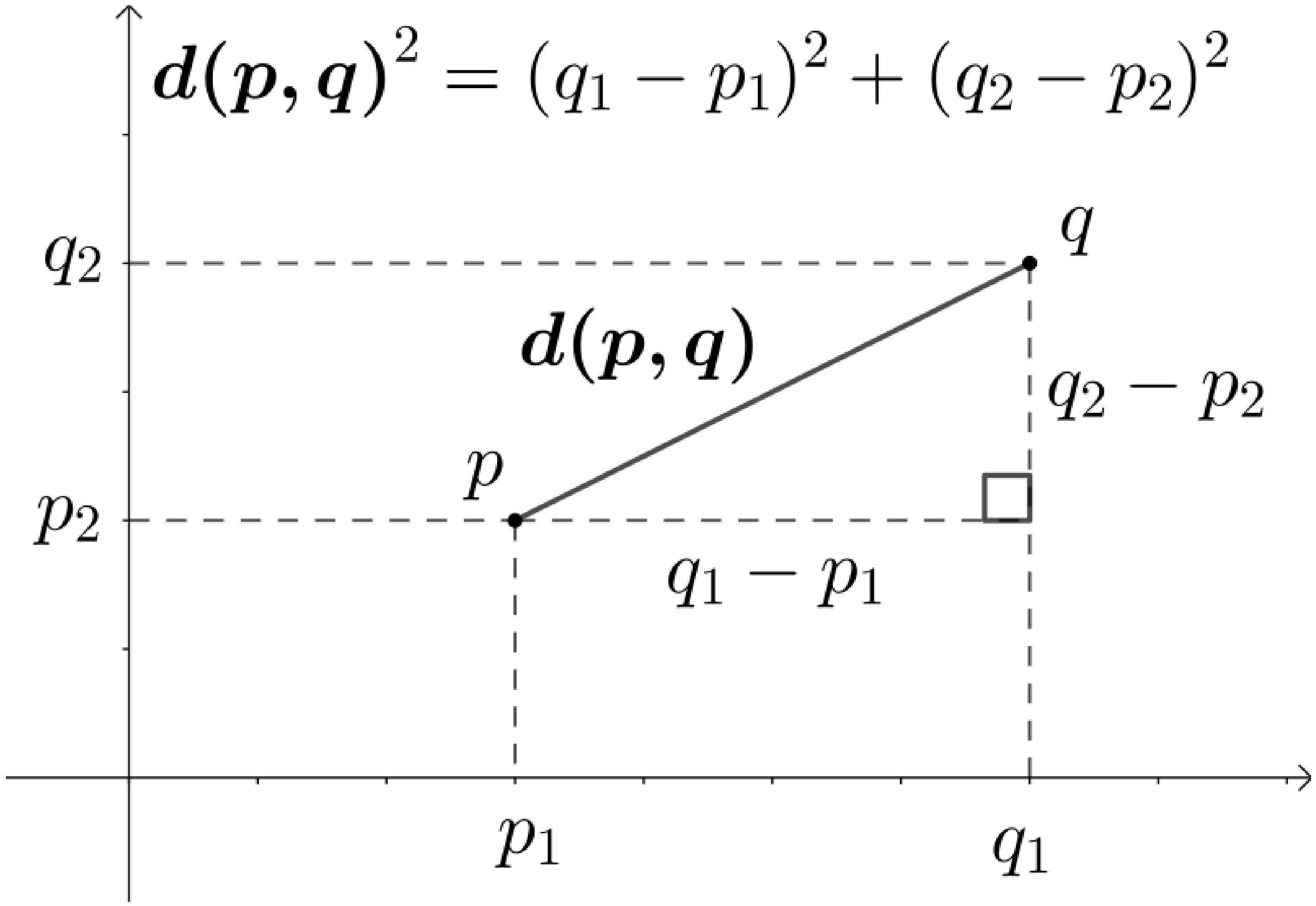
- In KNN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor.
- K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case.
- Suppose P1 is the point, for which label needs to predict. First, you find the one closest point to P1 and then the label of the nearest point assigned to P1.



KNN Classifier(3)

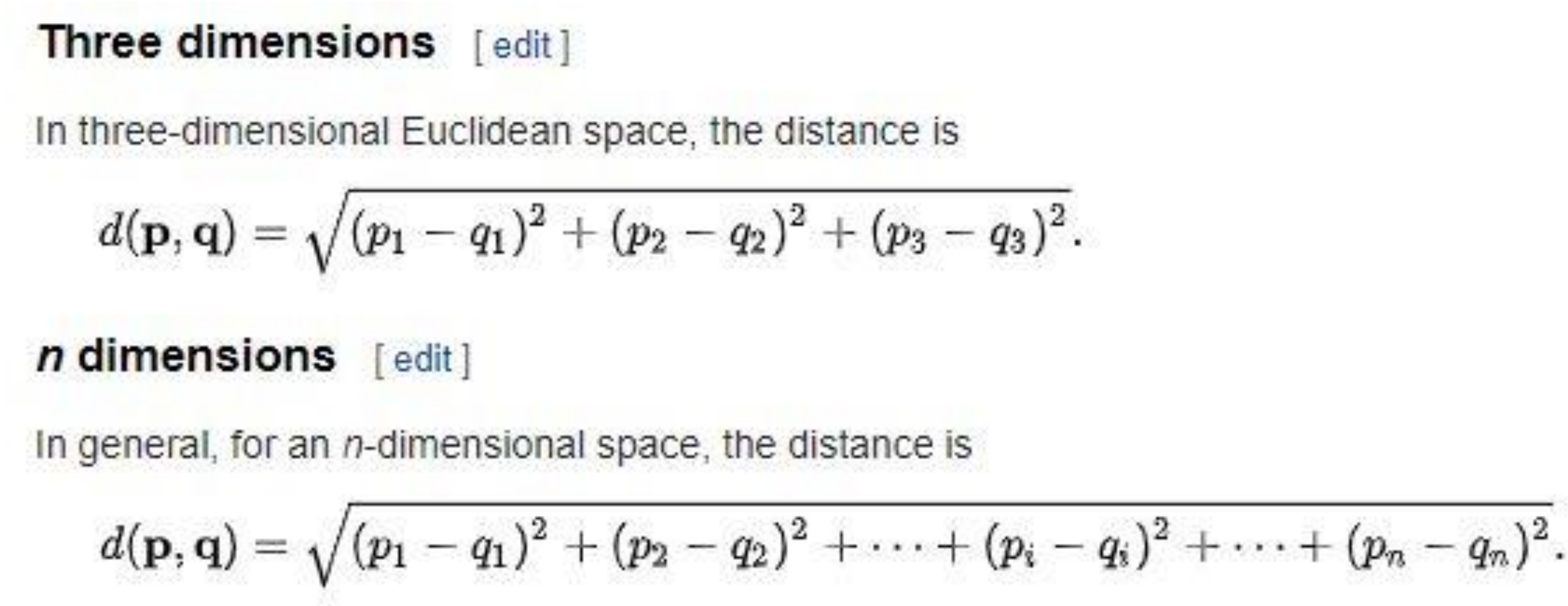
- Suppose P1 is the point, for which label needs to predict.
- First, you find the k closest point to P1 and then classify points by majority vote of its k neighbors.
- Each object votes for their class and the class with the most votes is taken as the prediction.
- For finding closest similar points, you find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance.

Euclidean Distance (1)



$$\boldsymbol{d(p,q)}^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$$

Euclidean Distance (2)



Three dimensions [edit]

In three-dimensional Euclidean space, the distance is

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

n dimensions [edit]

In general, for an *n*-dimensional space, the distance is

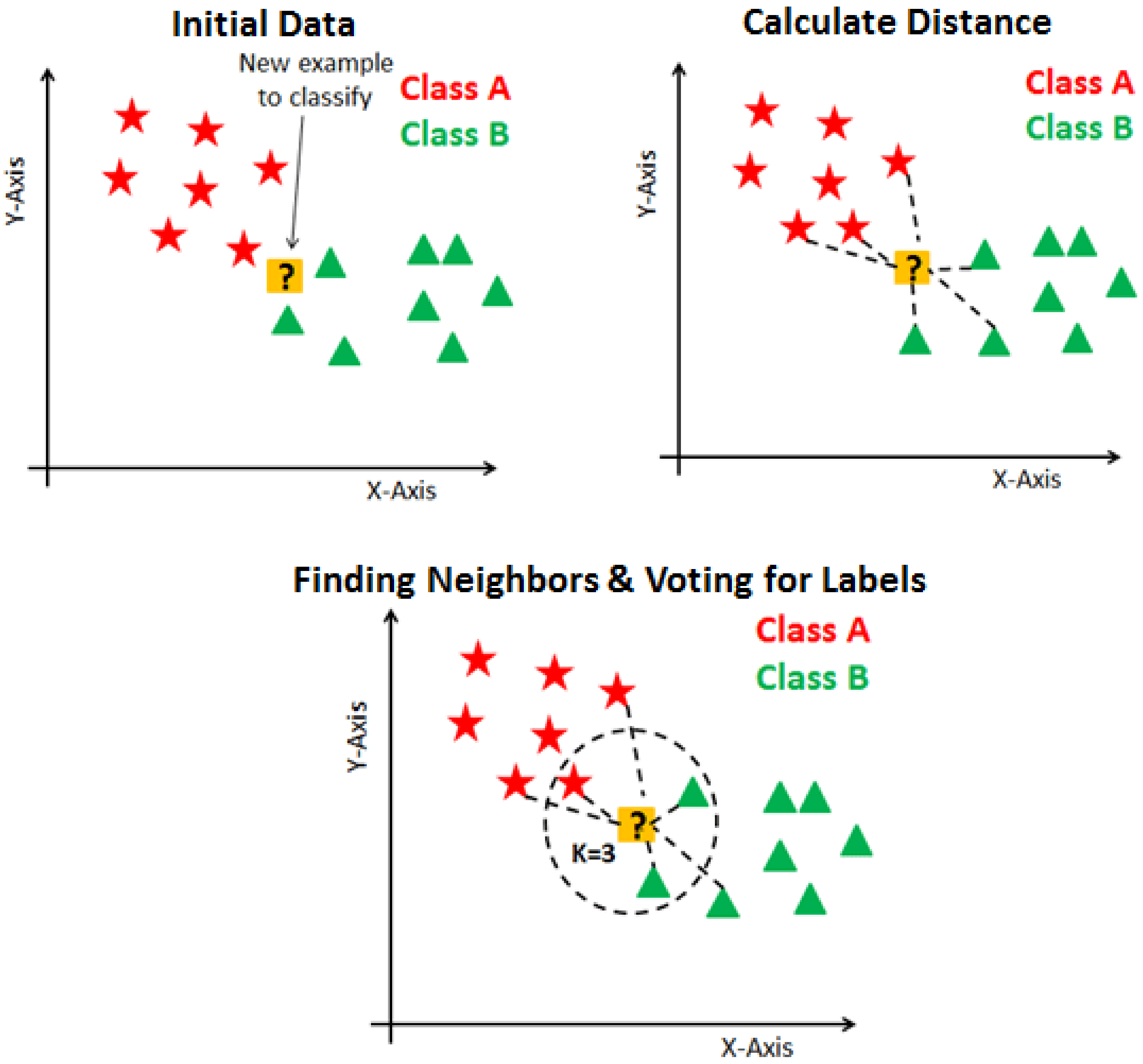
$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

KNN Classifier(4)

Calculate Distance

Find Closest Neighbors

Vote for Labels

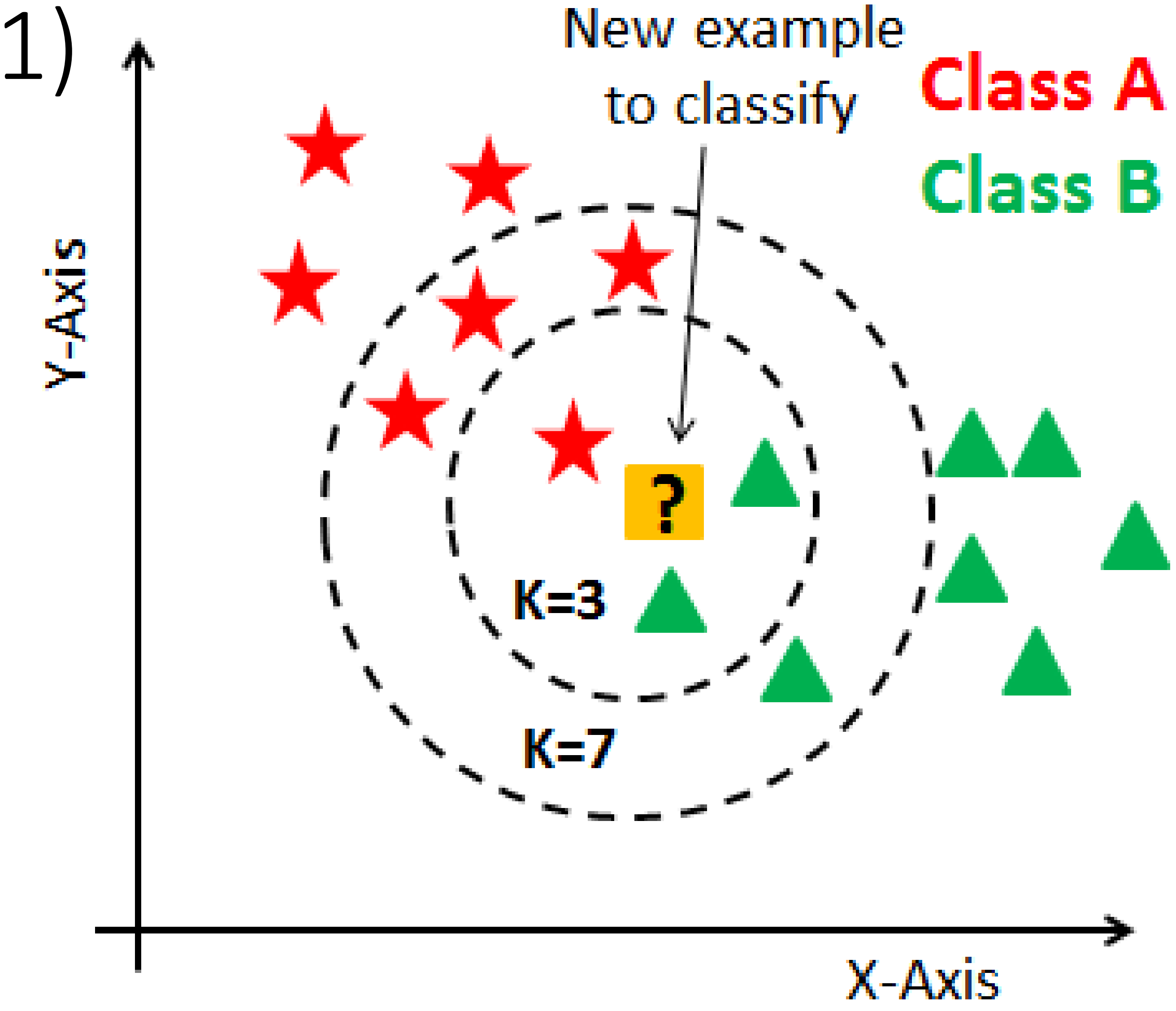


How to Decide K value(1)

- Research has shown that no optimal number of neighbors suits all kind of data sets. Each dataset has it's own requirements.
- In the case of a small number of neighbors, the noise will have a higher influence on the result, and a large number of neighbors make it computationally expensive.
- Research has also shown that a small amount of neighbors are most flexible fit which will have low bias but high variance and a large number of neighbors will have a smoother decision boundary which means lower variance but higher bias.

How to Decide K value(1)

- Generally, Data scientists choose as an odd number if the number of classes is even.
- You can also check by generating the model on different values of k and check their performance. You can also try Elbow method.



Points to Remember

- The training phase of K-nearest neighbor classification is much faster compared to other classification algorithms.
- KNN can be useful in case of nonlinear data. It can be used with the regression problem. Output value for the object is computed by the average of k closest neighbors value.
- The testing phase of K-nearest neighbor classification is slower and costlier in terms of time and memory.
- Euclidean distance is sensitive to magnitudes, therefore KNN also not suitable for large dimensional data.