BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

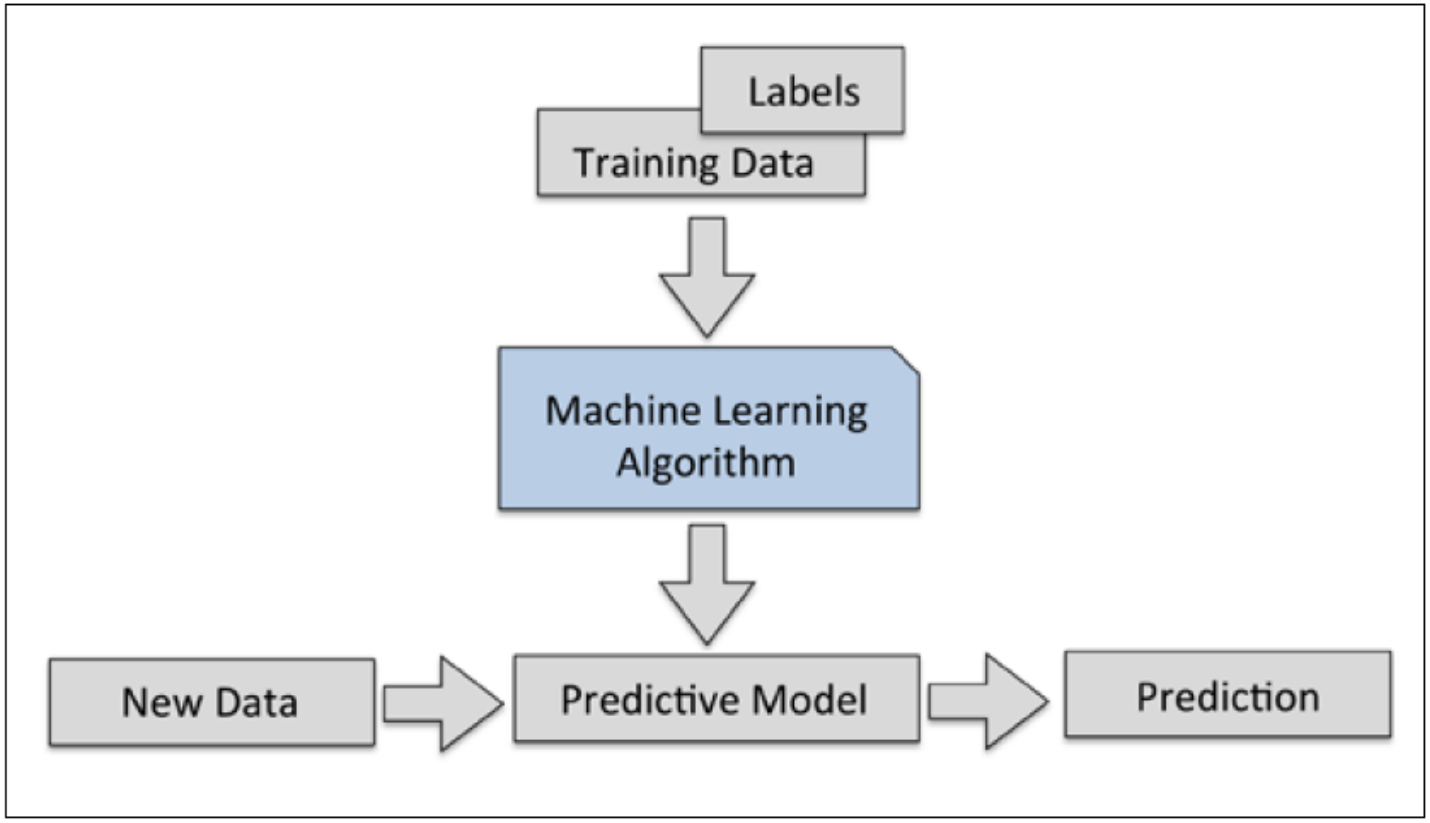
4 MACHINE LEARNING ALGORITHMS

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

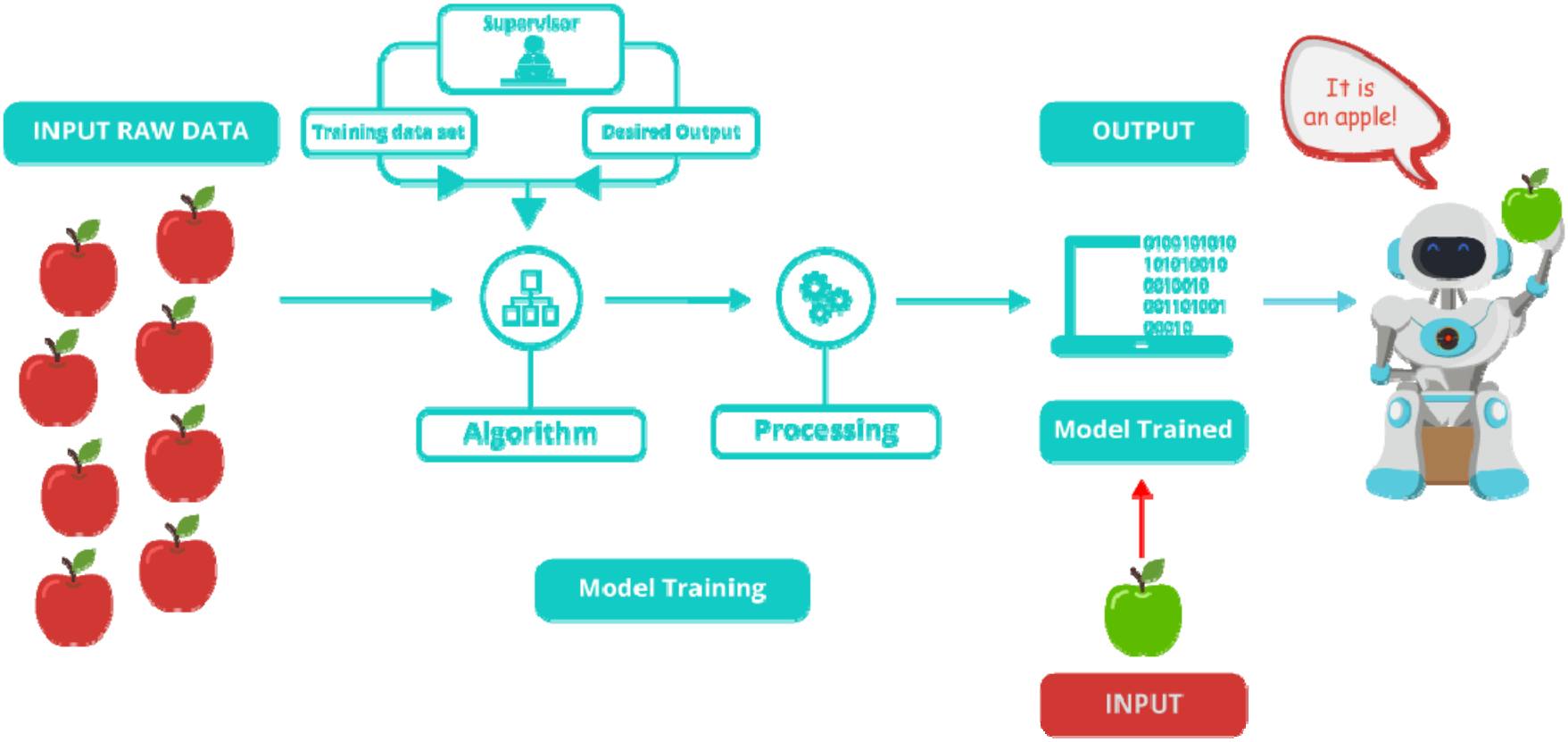
4 MACHINE LEARNING ALGORITHMS

1. What is machine learning?
2. Types of machine learning
3. How to choose the best algorithm in each case?
4. Supervised learning
    1. K-Nearest Neighbours
    2. Decision trees
    3. Linear Regression
    4. Decision trees for regression

Supervised learning



Example


INPUT RAW DATA
Supervisor
Training data set
Desired Output
Algorithm
Processing
Model Training
OUTPUT
Model Trained
INPUT
It is an apple!

BIG DATA Y MINERÍA DE DATOS GEOESPACIALES

4 MACHINE LEA

Christos Boulou
@ChristosBoulou

Seguir

I can't stop laughing..

Traducir Tweet

Om Pande
@Om_pande_

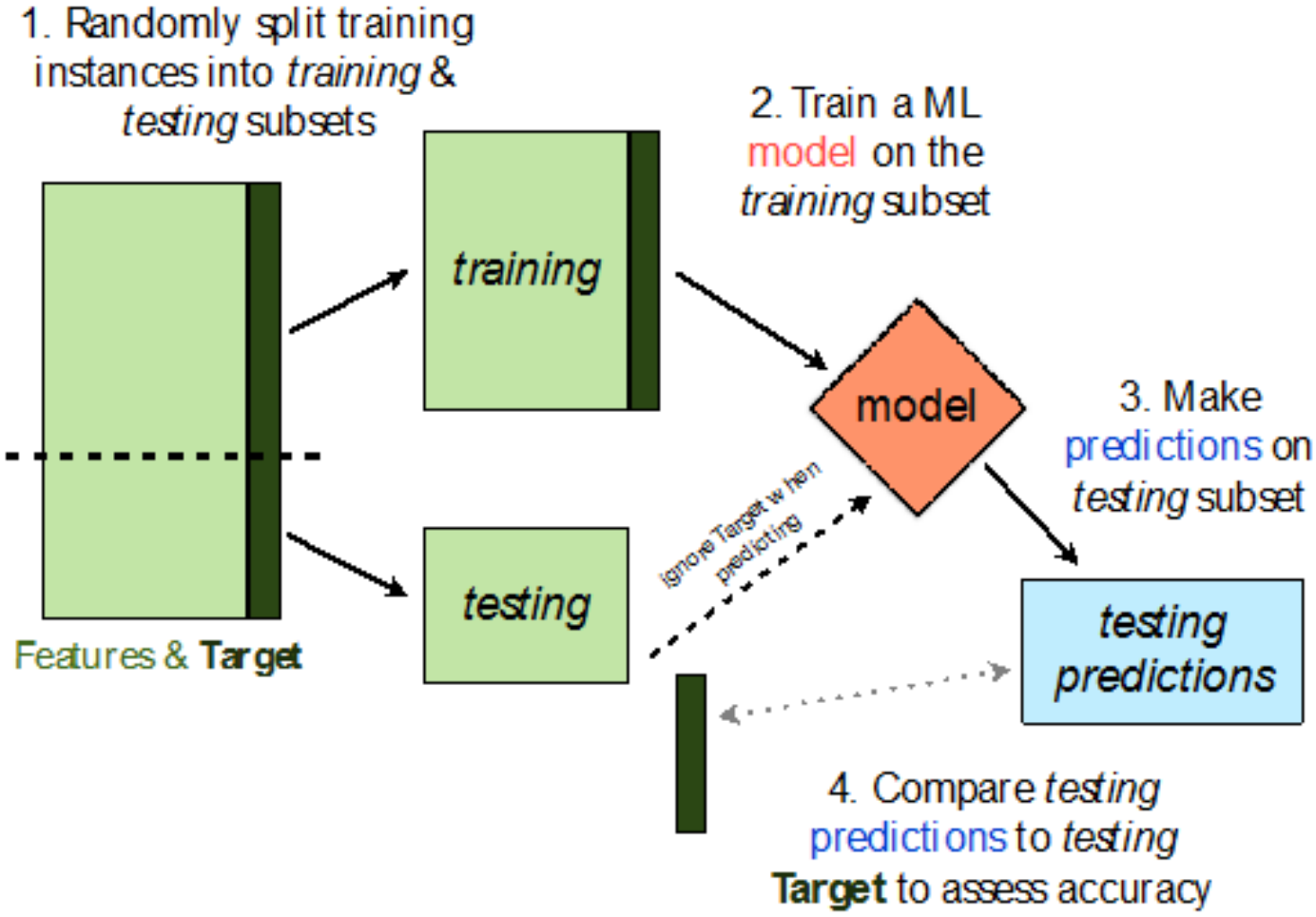
Parents: If all your friends jumped into the well, Will you ?
Kid: No !

Machine learning Algorithm: YES !

15:17 - 11 feb. 2019

100 Retweets **253** Me gusta

4 MACHINE LEARNING ALGORITHMS

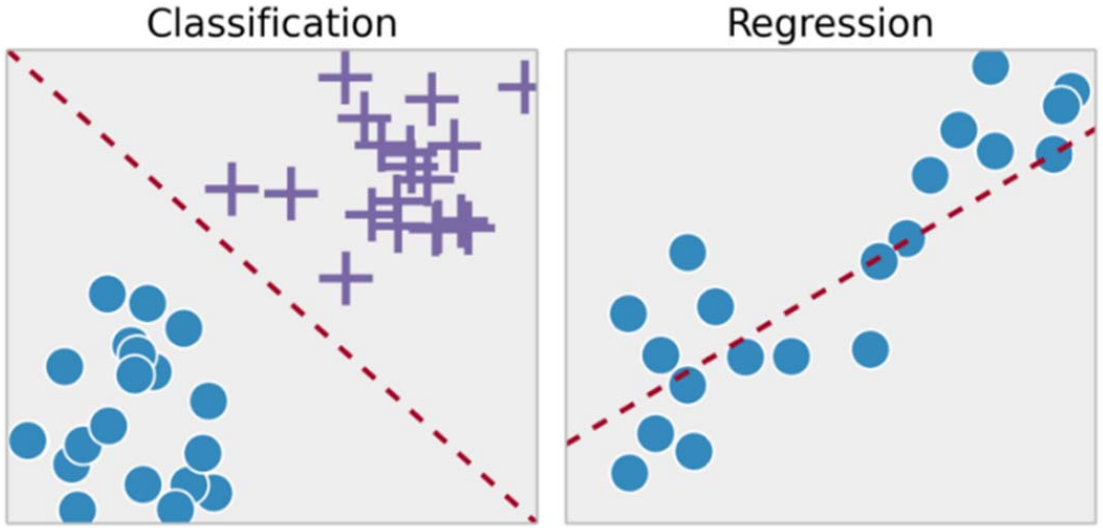


https://freecontent.manning.com/real-world-machine-learning-model-evaluation-optimization/

Definition

Learning a target function (f) that best maps input variables (X) to an output variable (Y):

Y = f(X)

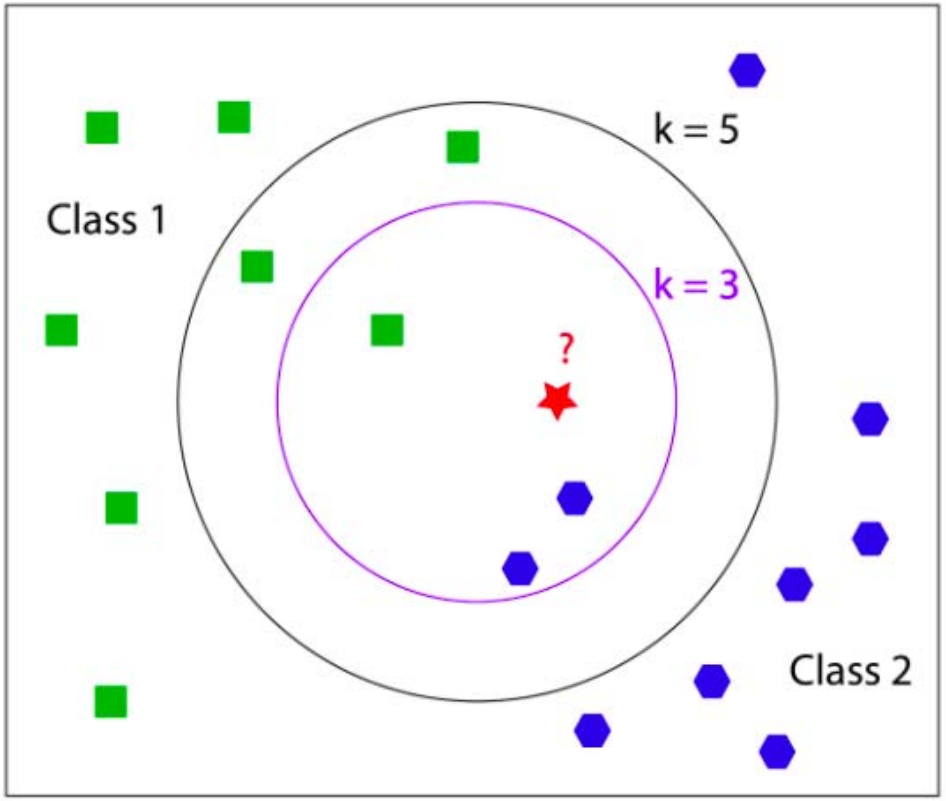


K-nearest neighbours
Decision trees

Linear regression

K-nearest neighbours

k-Nearest Neighbors identifies the k number of observations that are most proximate to the test sample, as defined by some distance metric



If k=3, the neighbors are {blue, blue, green} so we would classify the test sample as blue. If k=5, the neighbors are {blue, blue, green, green, green} and we would select green.

K-nearest neighbours. Characteristics

KNN searches the memorized training observations for the K instances that most closely resemble the new instance and assigns to it the their most common class.

Non-parametric: The model structure is determined from the data
Lazy: No explicit training phase

Therefore, it is computationally expensive and it has a high memory requirement

Based on feature similarity: different functions can be used

K-nearest neighbours. Elements to consider

Number of neighbours k

- When K is small, we are restraining the region of a given prediction and forcing our classifier to be “more blind” to the overall distribution
- A higher K averages more voters in each prediction and hence is more resilient to outliers, but prediction is more computationally costly

Distance

- Euclidean distance
- Cosine similarity
- Minkowski distance
- Haversine (for coordinates)
- …

K-nearest neighbours. Exercise

Given the following dataset, where X1 and X2 are two input variables and Y is the variable to predict. Which class would predict the k-NN algorithm, with k=3 using Euclidean distance, for the following input: X1=6, X2=2?

X1	X2	Y
7	7	*Bad*
7	4	*Bad*
3	4	*Good*
1	4	*Good*
4	2	*Good*
6	6	*Bad*

Confusion matrix

By definition a confusion matrix C is such that $C_{i,j}$ is equal to the number of observations known to be in group i but predicted to be in group j.

		Predicted (what our model says))			
Actual (what the data says)	CLASSES	A	B	C	Row totals
	A	5	2	3	10
	B	2	6	0	8
	C	3	2	2	7
	Column Totals	10	10	5	25

Diagonal numbers are rightly classified observations

Total number of observations/ records

Evaluation measures

Accuracy: percentage of correctly classified elements

Precision: number of times where the algorithm was correct out of all times where the algorithm predicted that category.

Recall: number of times where the algorithm was correct out of all of the cases where that category was the correct one.

F-beta score: weighted harmonic mean of the precision and recall;
best score=1, worst score= 0.

Support: number of occurrences of each class.

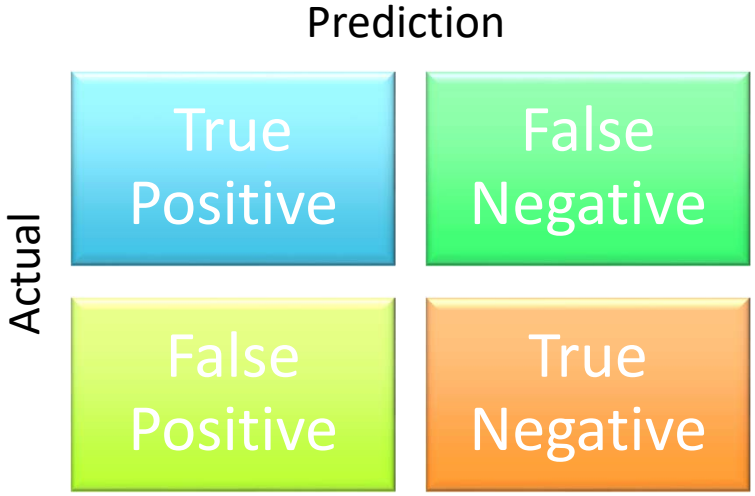
4 MACHINE LEARNING ALGORITHMS

$$\textbf{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

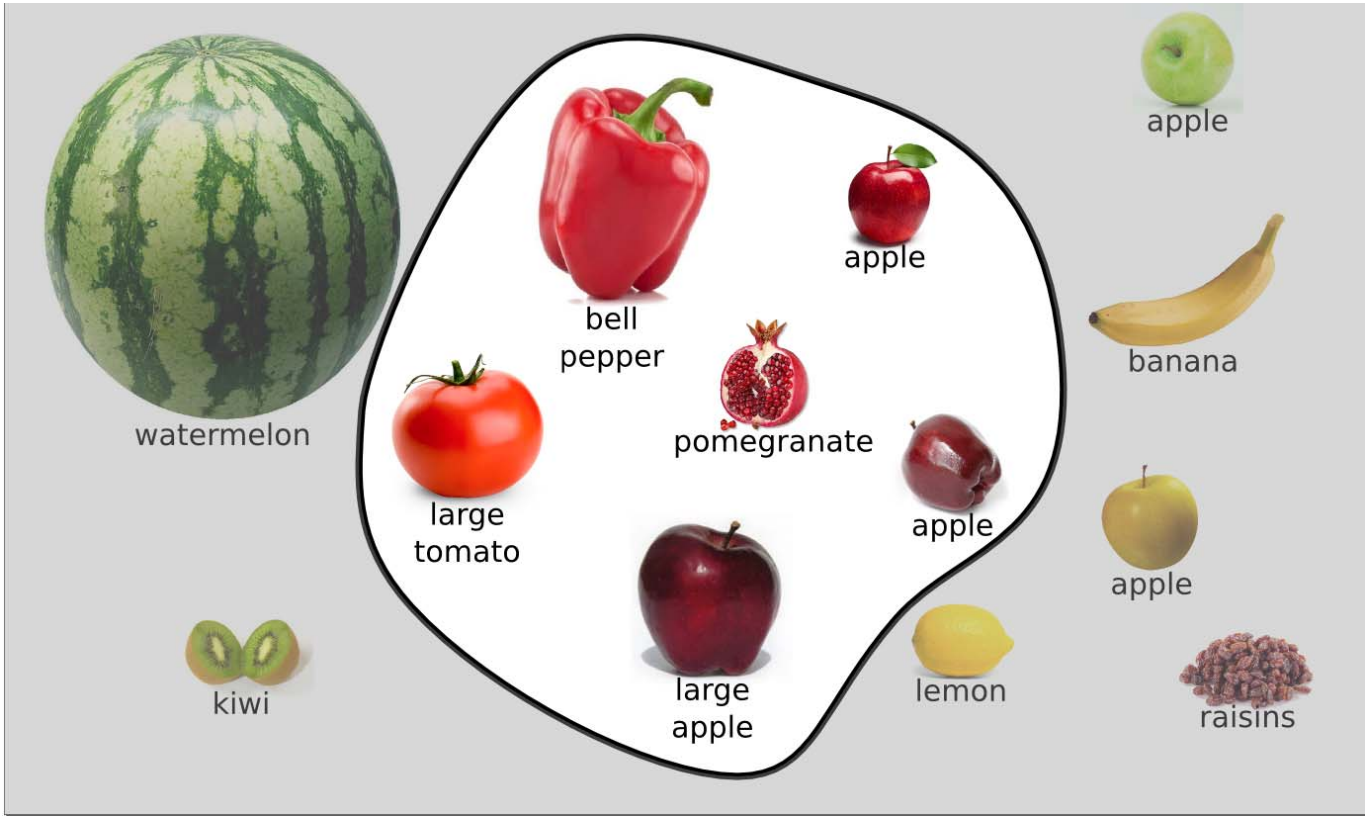
$$\textbf{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\textbf{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$



4 MACHINE LEARNING ALGORITHMS

Exercise: compute the confusion matrix of the following result, if the algorithm must classify these fruits into “Apple” or “No Apple”. From this confusion matrix, compute accuracy, precision and recall.



https://opensourceconnections.com/blog/2016/03/30/search-precision-and-recall-by-example/