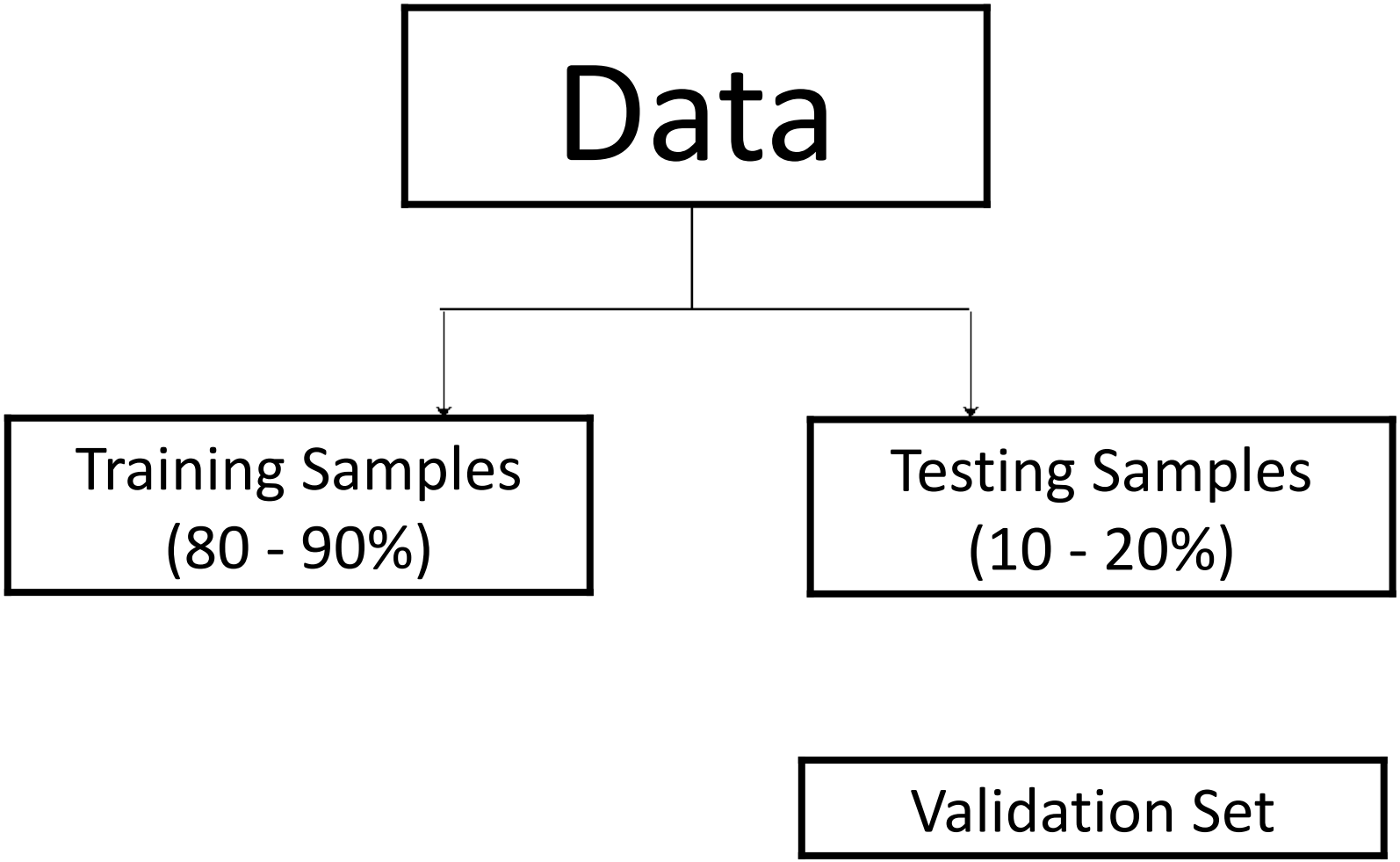
Machine Learning


Data
Training Samples
(80 - 90%)
Testing Samples
(10 - 20%)
Validation Set

Supervised Learning

- Supervised learning is a technique in which we teach or train the machine using data which is well labeled.

- It is where we have an input variable(X) and output variable(Y) and we use an algorithm to learn the mapping function from the input to the output.

$$Y=f(X)$$

SL algorithms:

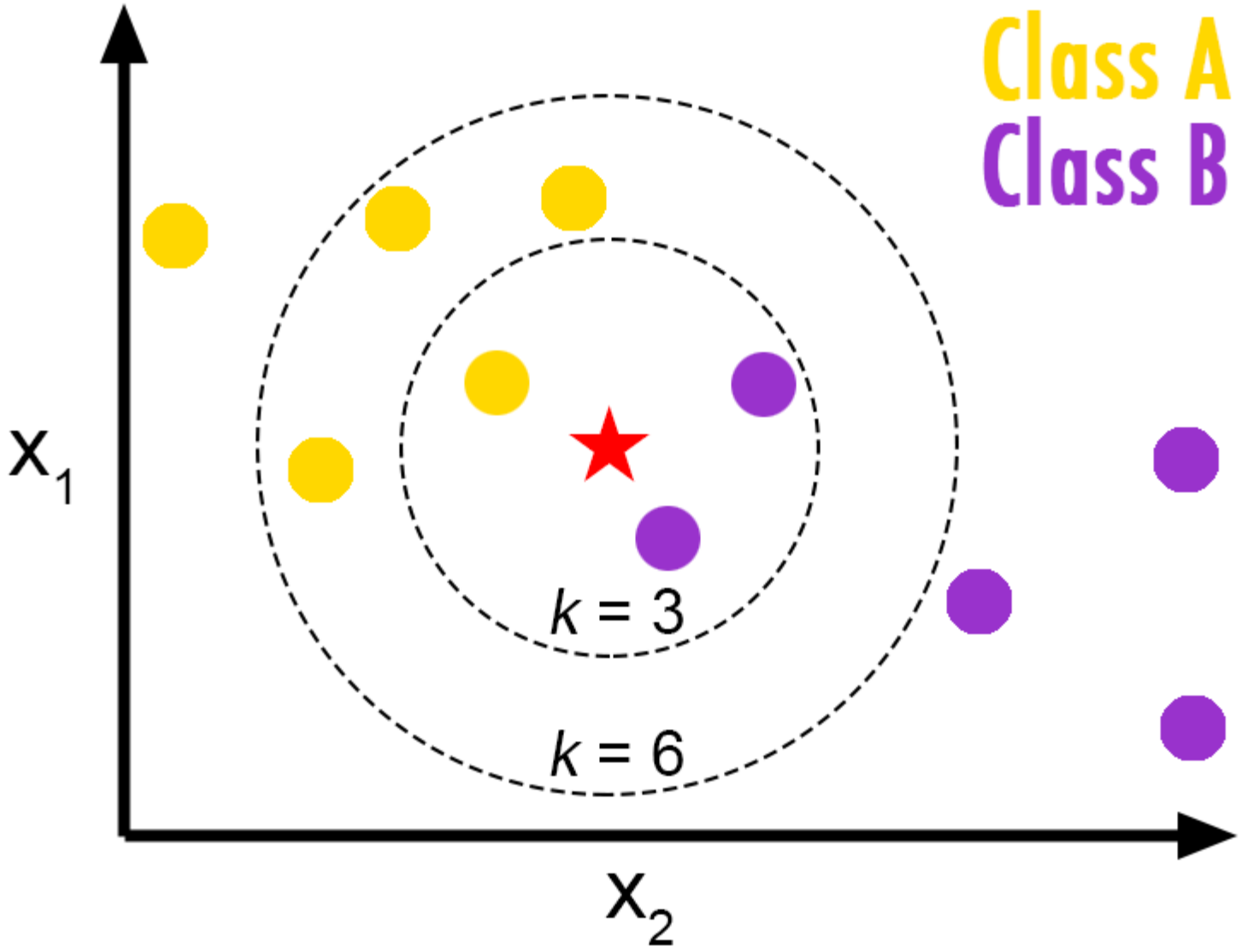
- kNN
- Naive Bayes Classifier
- Decision Tree
- Regression (Linear, Logistics)
- Random Forest

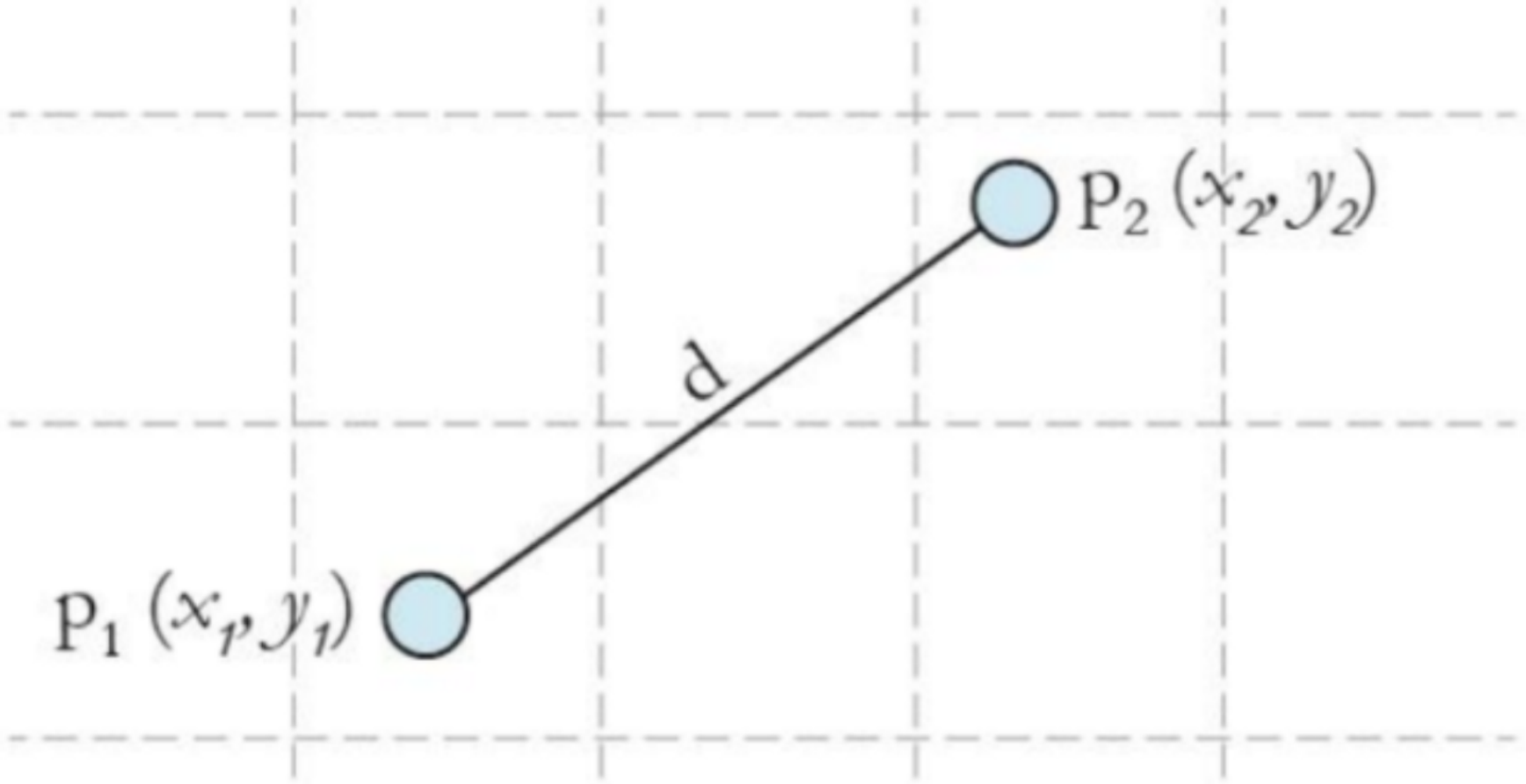
kNN(k- Nearest Neighbors)

- An object is classified based on a majority vote of its neighbors, or case based on a similarity measure.

- e.g. Search Engine

kNN





$$\text{Euclidean distance (d)} = \sqrt{(x_2\text{-}x_1)^2 + (y_2\text{-}y_1)^2}$$


Sepal
Petal
Versicolor
Virginica
Setosa

kNN - distance formula

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \ldots + (x_{ip} - x_{lp})^2}$$

Applications of KNN

- If you're searching for semantically similar documents (i.e., documents containing similar topics), this is referred to as Concept Search.

- The biggest use case of K-NN search might be Recommender Systems. If you know a user likes a particular item, then you can recommend similar items for them.

Example: Spam Filter

Input: email

Output: spam/ham

Setup:

- Get a large collection of example emails, each labeled "spam" or "ham"
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future emails

Features: The attributes used to make the ham / spam decision

- Words: FREE!
- Text Patterns: $dd, CAPS
- Non-text: SenderInContacts
- …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Slide from Mackassy

Example: Digit Recognition

Input: images / pixel grids

Output: a digit 0-9

Setup:

- Get a large collection of example images, each labeled with a digit
- Note: someone has to hand label all this data!
- Want to learn to predict labels of new, future digit images

0

1

2

1

??

Features: The attributes used to make the digit decision

- Pixels: (6,8)=ON
- Shape Patterns: NumComponents, AspectRatio, NumLoops
- ...

Slide from Mackassy

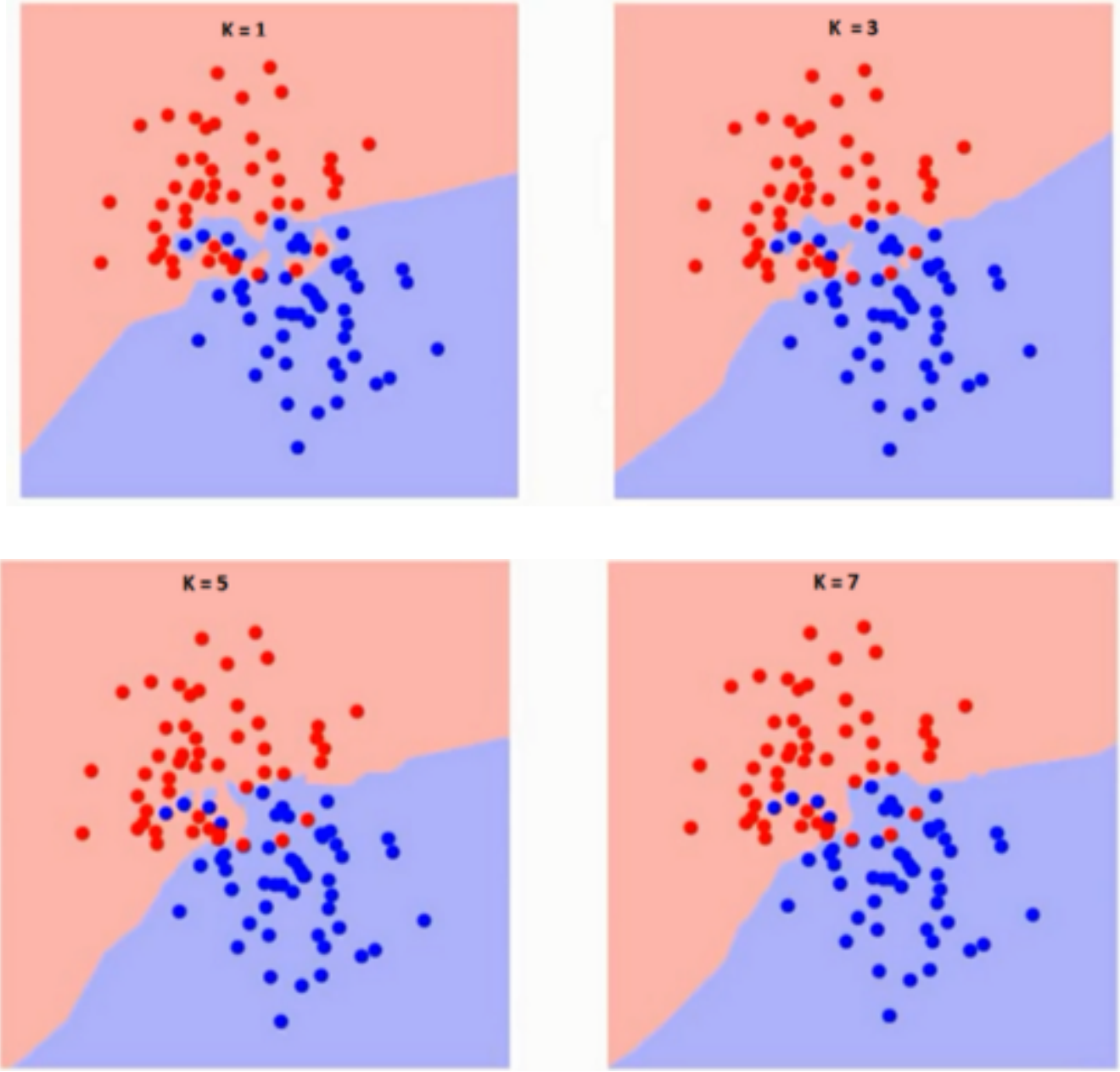
Advantages

- No assumptions about data — useful, for example, for nonlinear data
- Simple algorithm — to explain and understand/interpret
- High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models
- Versatile — useful for classification or regression

Disadvantages

- Computationally expensive — because the algorithm stores all of the training data
- High memory requirement
- Prediction stage might be slow (with big N)
- Sensitive to irrelevant features and the scale of the data.

kNN - Effect of k



kNN - Effect of k

- when k=1 decision surface is not smooth and the point's are separated more accurately. It is also called as Over fitting.

- when k=7 decision surface is lightly smooth and many point's are wrongly classified.It is also called as Under fitting.