machinepulse

Machine Learning & Real-world Applications

Ajay Ramaseshan

MachinePulse.

Agenda

❑ **Why Machine Learning**?

❑ **Supervised Vs Unsupervised learning**.

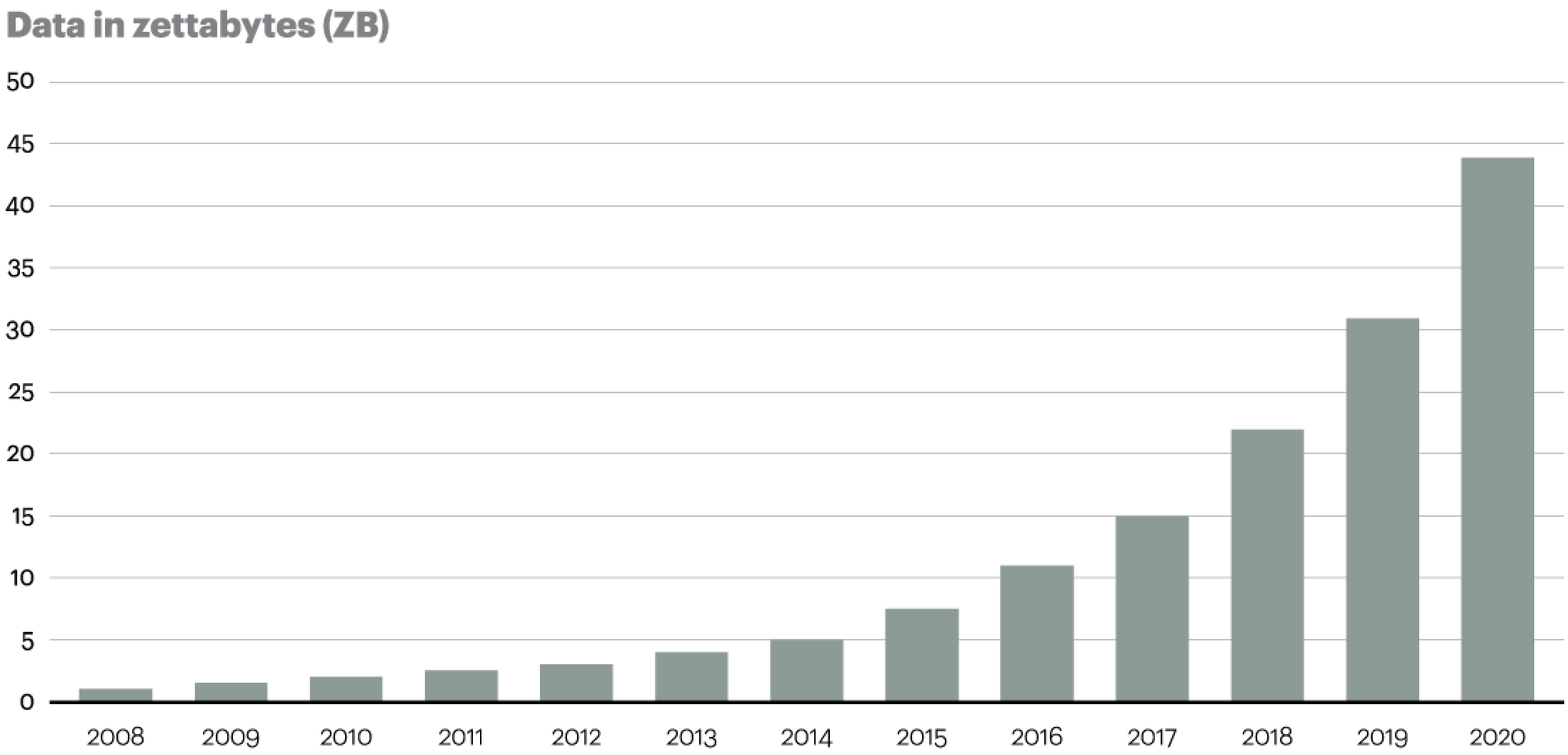
❑ **Machine learning methods.**

❑ **Real world applications.**

Why Machine Learning?

Figure 1

Data is growing at a 40 percent compound annual rate, reaching nearly 45 ZB by 2020



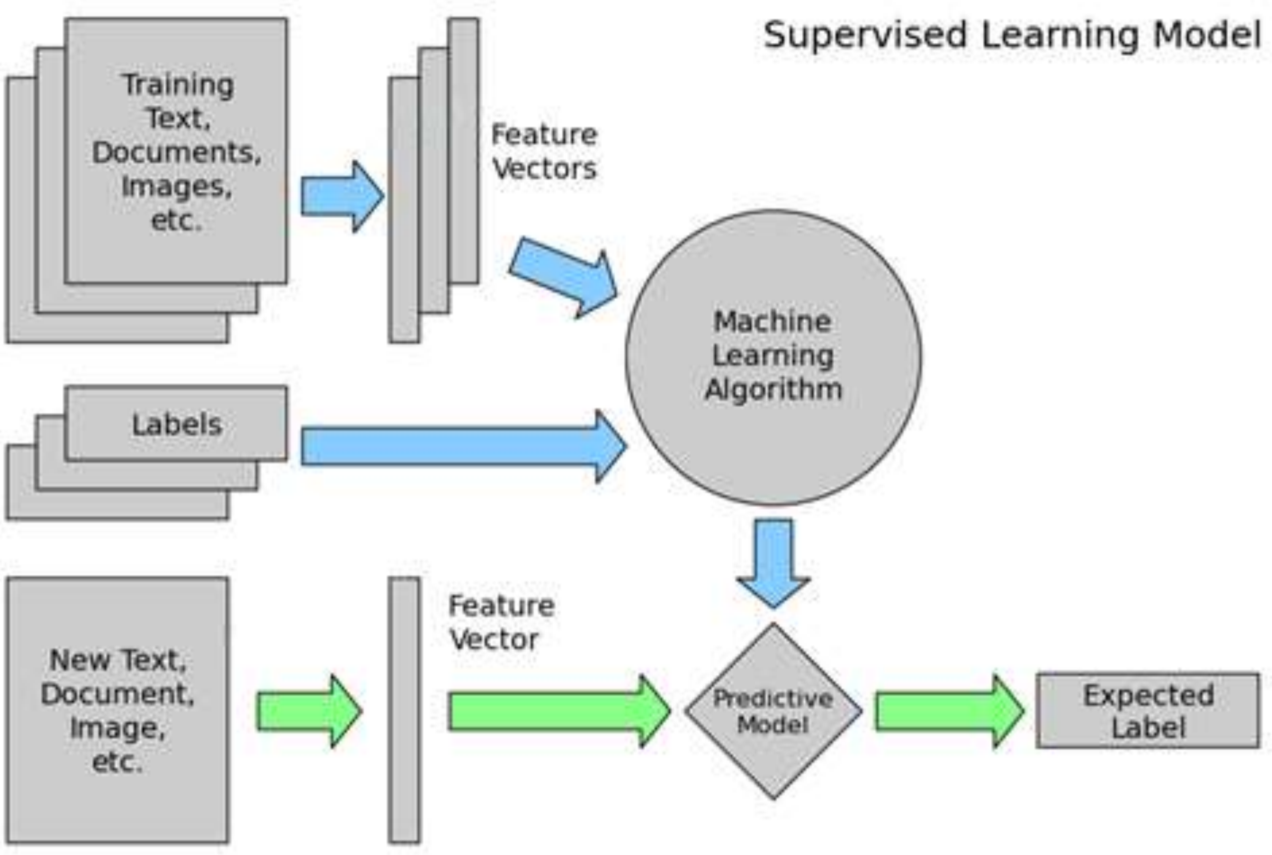
Source: Oracle, 2012

Why Machine Learning?

- ❑ Volume of data collected growing day by day.
- ❑ Data production will be 44 times greater in 2020 than in 2009.
- ❑ Every day, 2.5 quintillion bytes of data are created, with 90 percent of the world's data created in the past two years.
- ❑ Very little data will ever be looked at by a human.
- ❑ Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- ❑ Knowledge Discovery is **NEEDED** to make sense and use of data.
- ❑ Machine Learning is a technique in which computers learn from data to obtain insight and help in knowledge discovery.

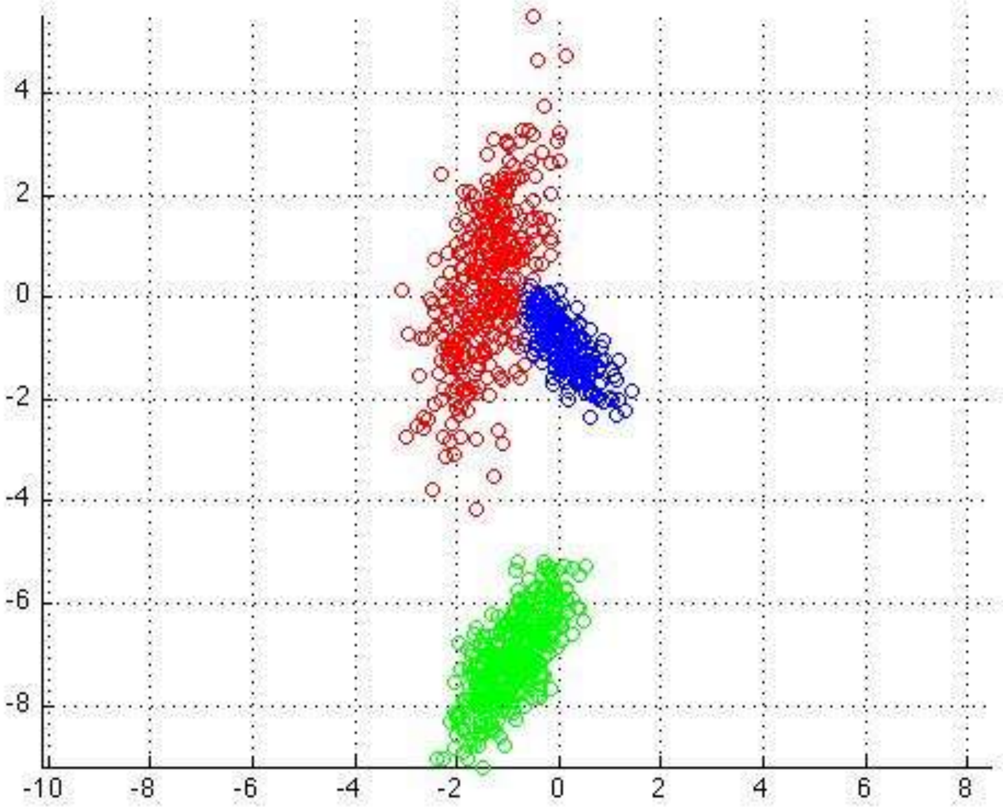
Overview of Machine Learning Methods

- ❑ Supervised learning – class labels/ target variable known


Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Overview (Contd)

- Unsupervised learning – no class labels provided, need to detect clusters of similar items in the data.



Parametric Vs Nonparametric

- ❑ Parametric Models:

  Assumes prob. distribution for data, and learn parameters from data

  E.g. Naïve Bayes classifier, linear regression etc.

- ❑ Non-parametric Models:

  No fixed number of parameters.

  E.g. K-NN, histograms etc.

Generative Vs Discriminative

❑ Generative model – learns model for generating data, given some hidden parameters.

❑ Learns the joint probability distribution p(x,y).

e.g. HMM, GMM, Naïve Bayes etc.

❑ Discriminative model – learns dependence of unobserved variable y on observed variable x.

❑ Tries to model the separation between classes.

❑ Learns the conditional probability distribution p(y|x).

e.g. Logistic Regression, SVM, Neural networks etc.

Classification

- ❑ Classification – Supervised learning.

- ❑ Commonly used Methods for Classification –

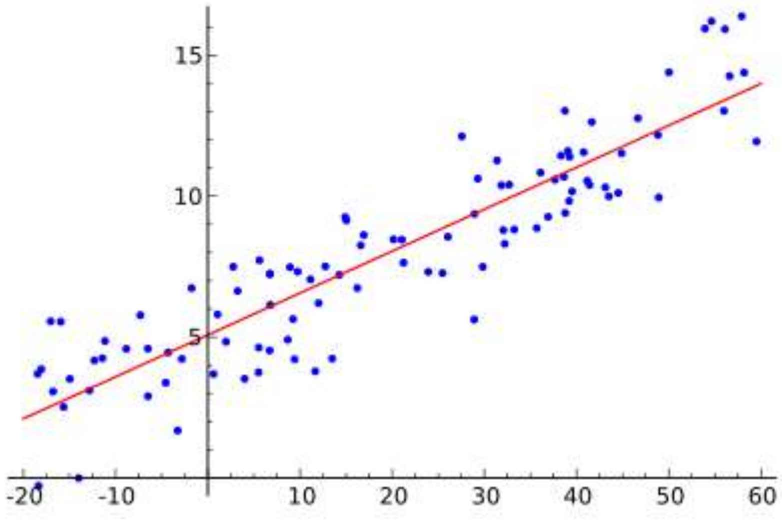
  - ➢Naïve Bayes
  - ➢Decision tree
  - ➢K nearest neighbors
  - ➢Neural Networks
  - ➢Support Vector Machines.

Regression

❑ Regression – Predicting an output variable given input variables.

❑ Algorithms used –

- ➢Ordinary least squares
- ➢ Partial least squares
- ➢ Logistic Regression
- ➢ Stepwise Regression
- ➢ Support Vector Regression
- ➢Neural Networks



Clustering

❑ Clustering:

Group data into clusters using similarity measures.

❑ Algorithms:

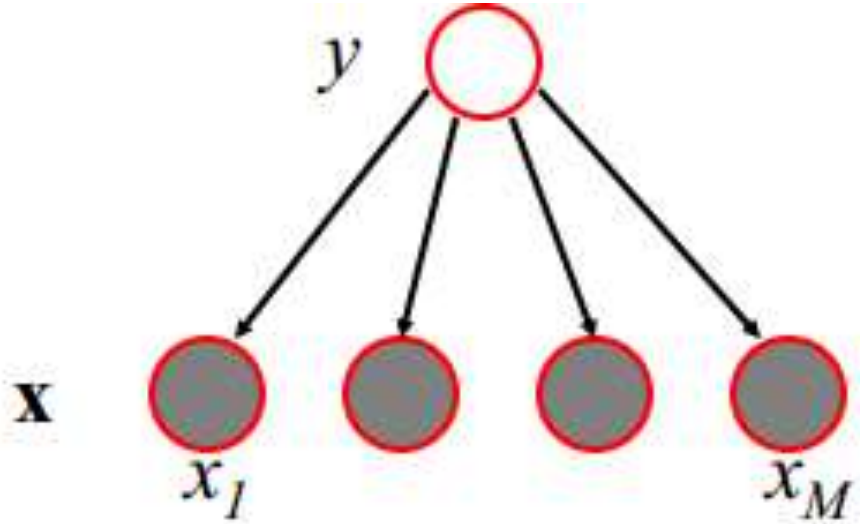
- ➢K-means clustering
- ➢ Density based EM algorithm
- ➢ Hierarchical clustering.
- ➢ Spectral Clustering

Naïve Bayes

- ❑ Naïve Bayes classifier

  Assumes conditional independence among features.

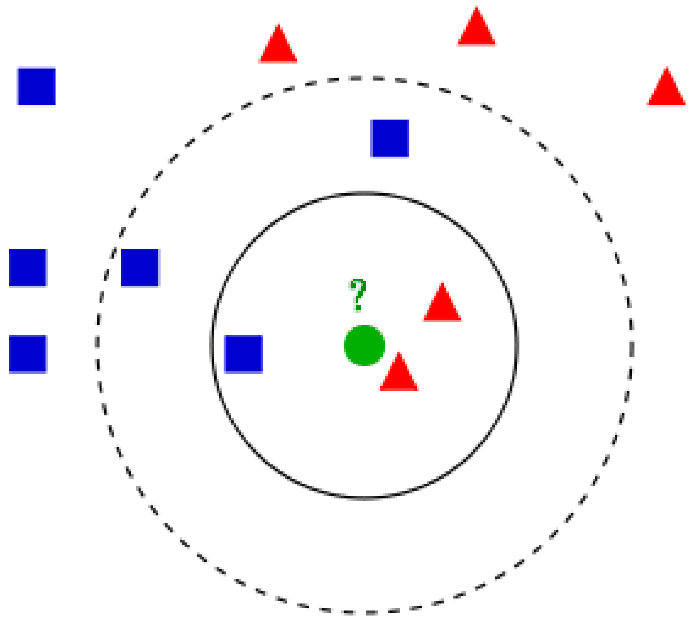
- ❑ $P(x_i,x_j,x_k|C) = P(x_i|C)P(x_j|C)P(x_k|C)$



K-Nearest Neighbors

- K-nearest neighbors: Classifies the data point with the class of the k nearest neighbors.

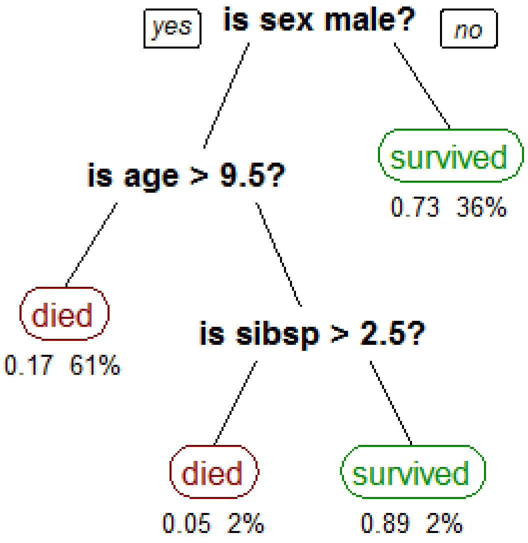
- Value of k decided using cross validation



Decision trees

- ❑ Leaves indicate classes.
- ❑ Non-terminal nodes – decisions on attribute values

- ❑ Algorithms used for decision tree learning
  - ➢ C4.5
  - ➢ ID3
  - ➢ CART.

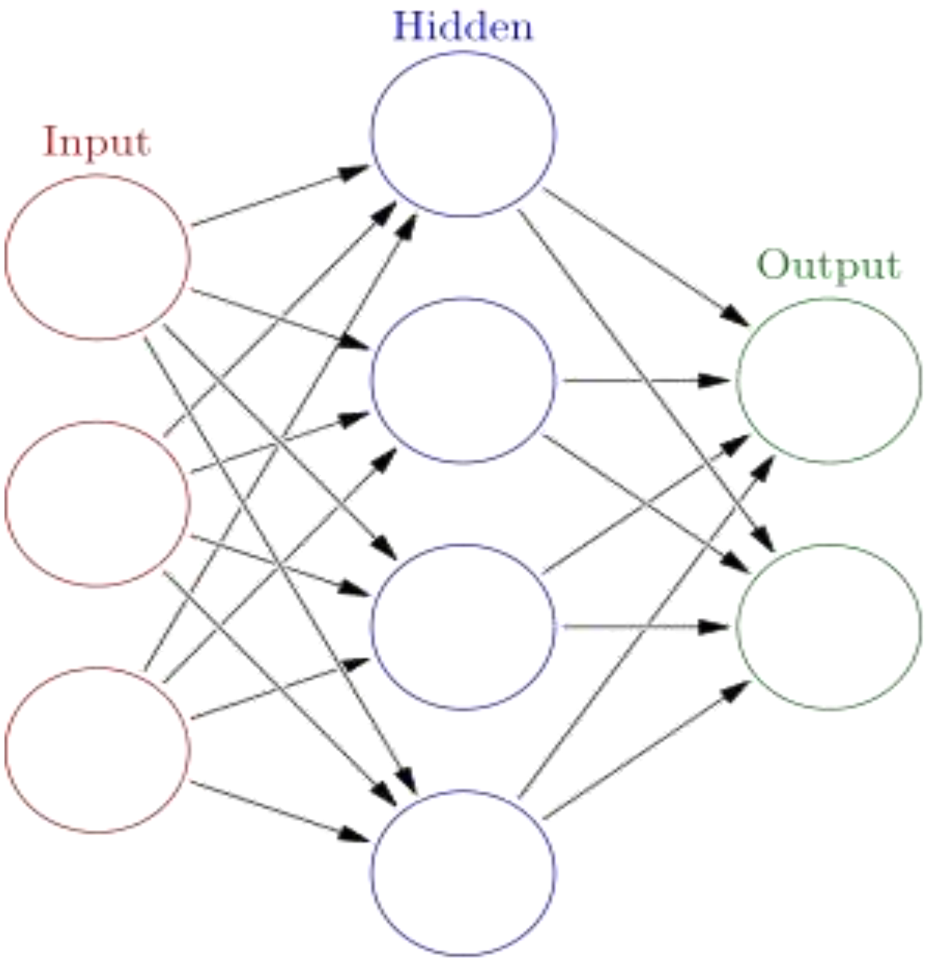


Neural Networks

- ❑ Artificial Neural Networks Modeled after the human brain

- ❑ Consists of an input layer, many hidden layers, and an output layer.

- ❑ Multi-Layer Perceptrons, Radial Basis Functions, Kohonen Networks etc.



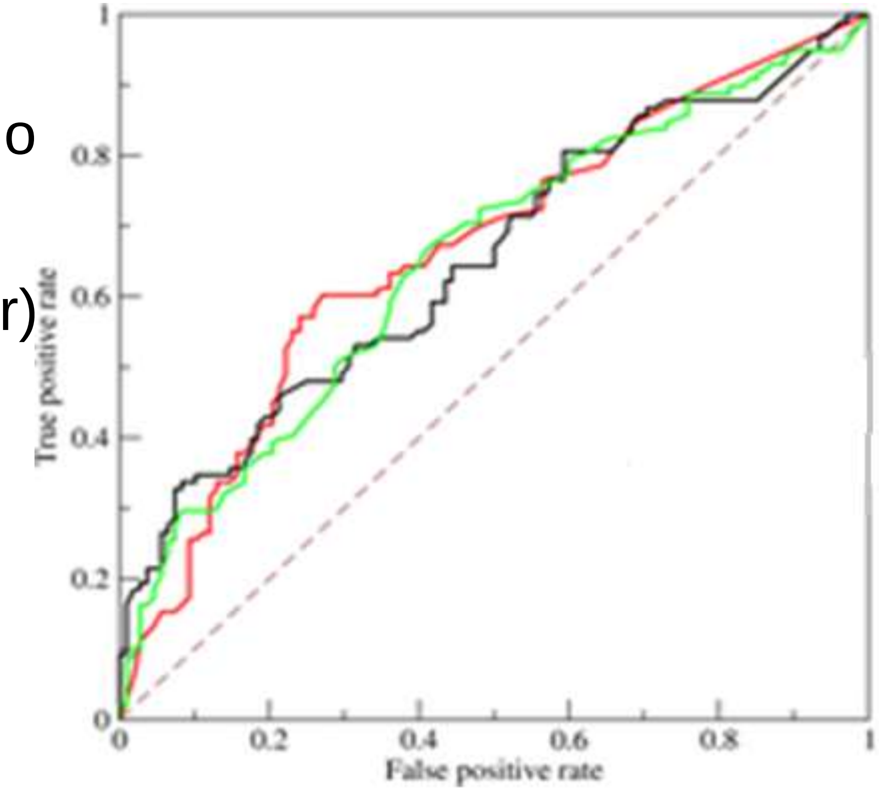
Evaluation of Machine Learning Methods

❖Validation methods

- ❑ Cross validation techniques
  - ➢K-fold cross validation
  - ➢Leave one out Cross Validatio

❑ ROC curve (for binary classifier)

❑ Confusion Matrix



Real life applications

Some real life applications of machine learning:

❑ Recommender systems – suggesting similar people on Facebook/LinkedIn, similar movies/ books etc. on Amazon,

❑ Business applications – Customer segmentation, Customer retention, Targeted Marketing etc.

❑ Medical applications – Disease diagnosis,

❑ Banking – Credit card issue, fraud detection etc.

❑ Language translation, text to speech or vice versa.

Breast Cancer Dataset and k-NN

Wisconsin Breast Cancer dataset:

- ❑ Instances : 569
- ❑ Features : 32
- ❑ Class variable : malignant, or benign.
- ❑ Steps to classify using k-NN
  - ➢ Load data into R
  - ➢ Data normalization
  - ➢ Split into training and test datasets
  - ➢ Train model
  - ➢ Evaluate model performance

Thank you!