Machine Learning

Leonard Wee

orcid.org/0000-0003-1612-9055

Place a letter next to the topic that matches the set diagram ….

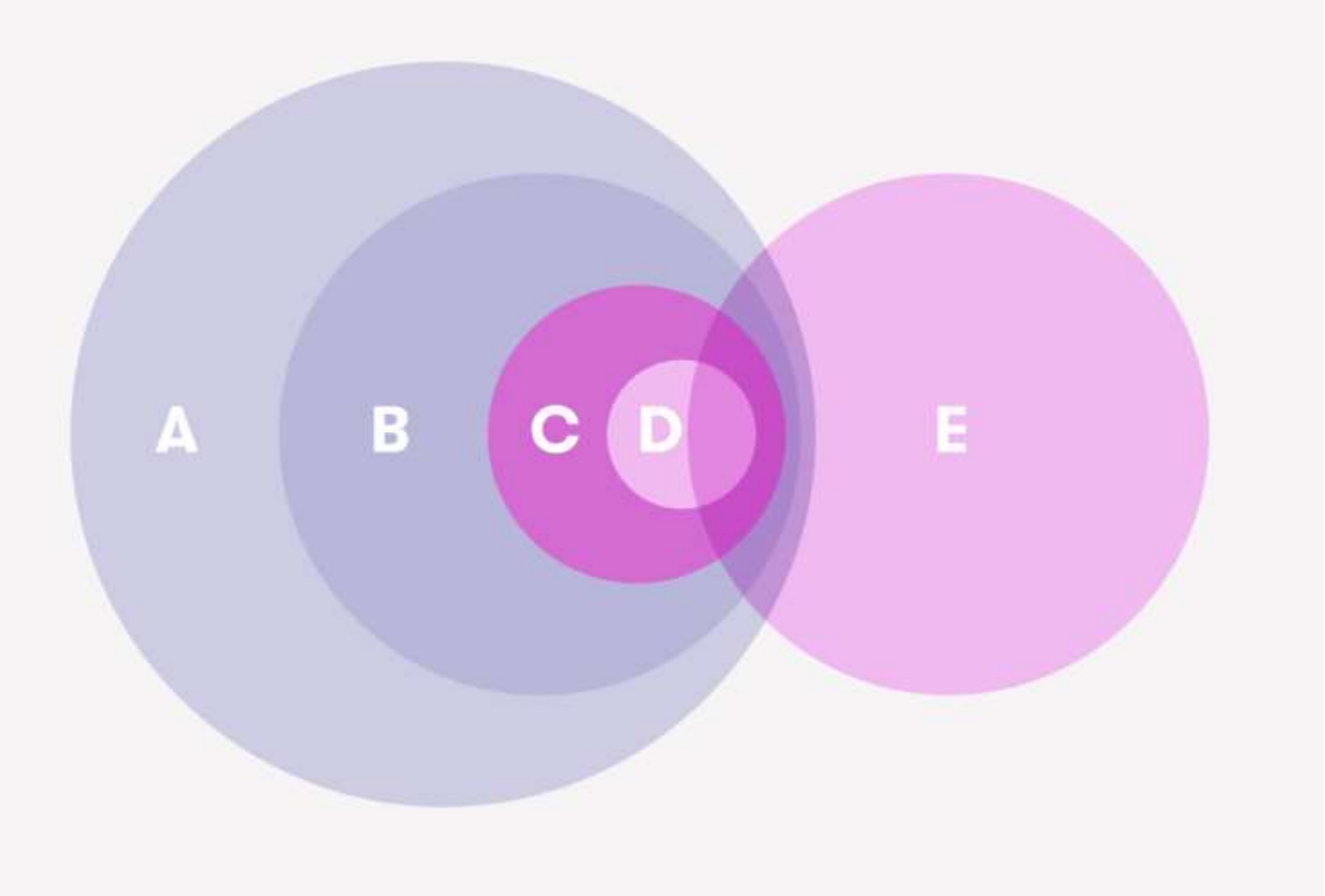
Machine learning : ____

Artificial intelligence : ____

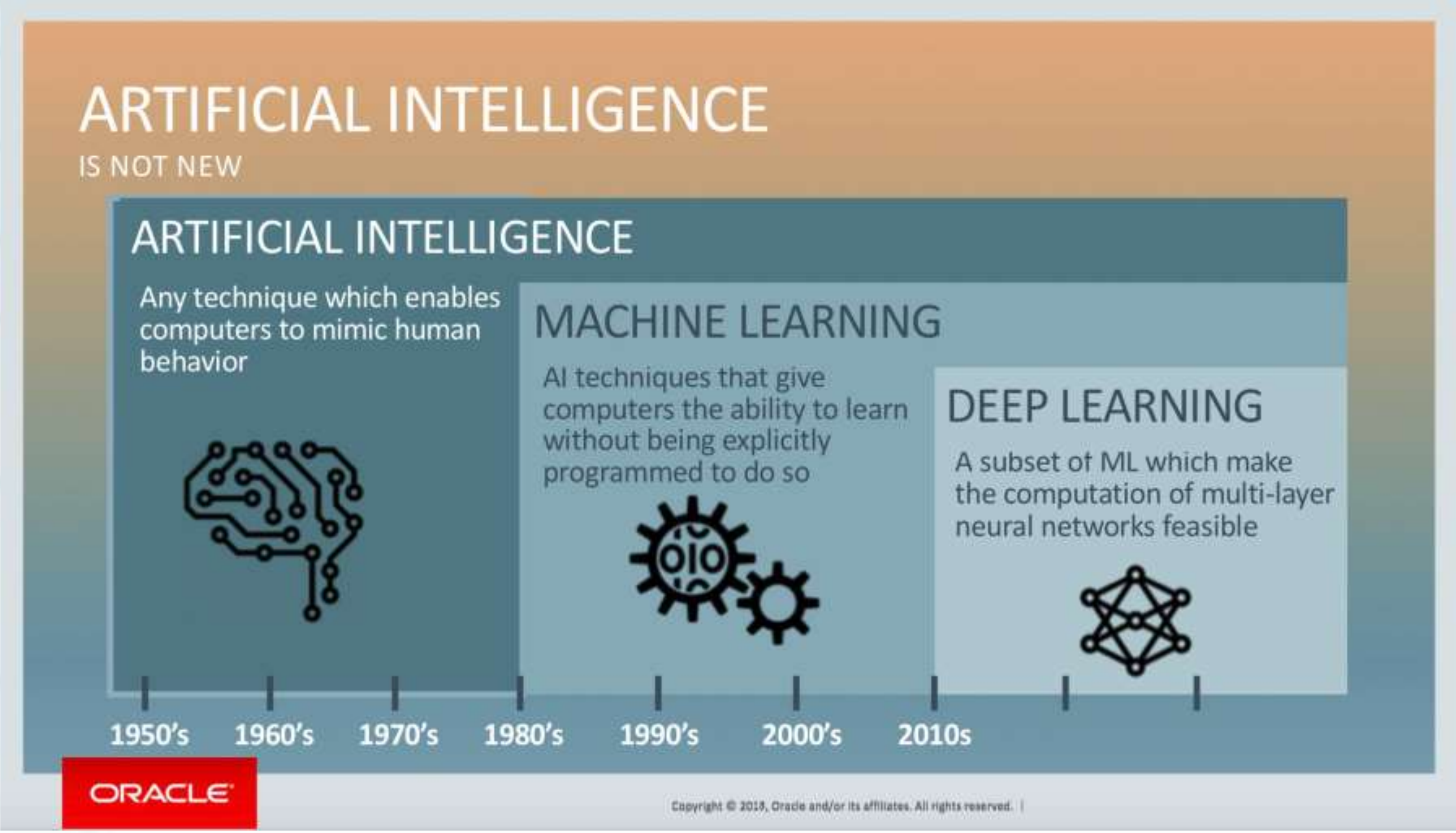
Deep learning : ____

Data science : ____

Computer science : ____



AI, ML and DL are - by today's measure – quite "old" ideas!



Learning Objectives

What is ML

Understand the difference in approach and guiding philosophy, rather than any specific tool or situation

Classification

Trees and Forests

Vector spaces

Artificial neurons

Using R

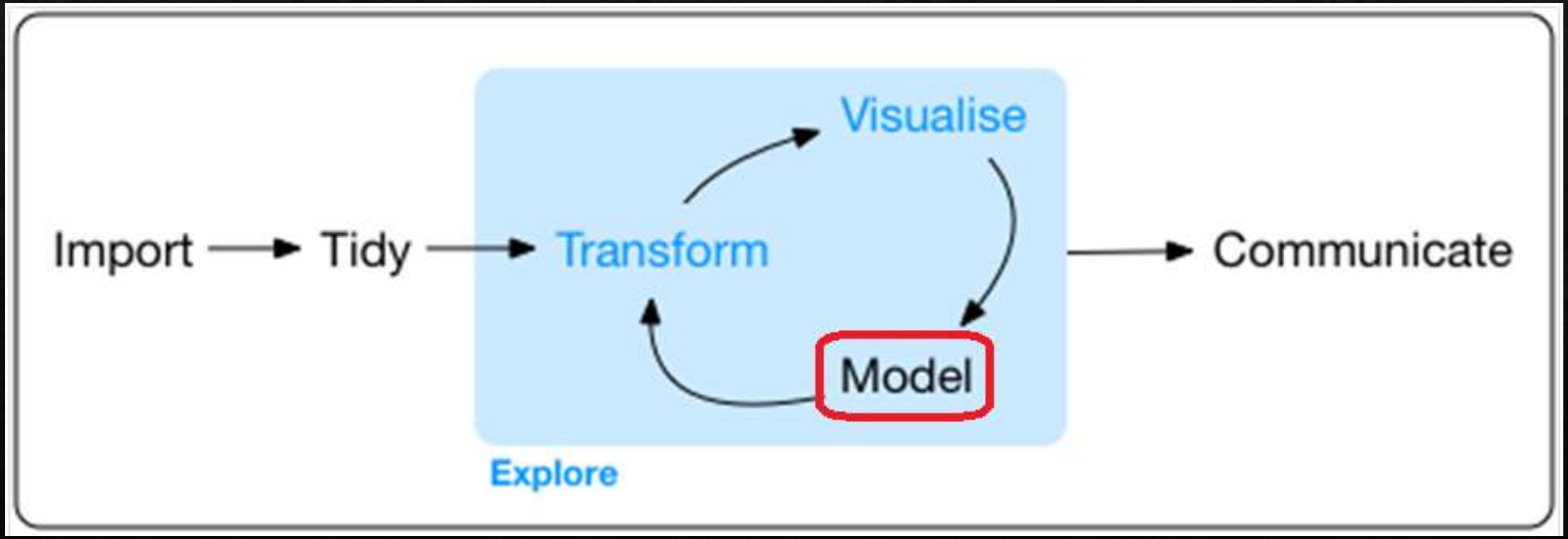
Caret package

Train simple ML models

Compare multiple models

Online Resources

- R for Data Science : https://r4ds.had.co.nz/ (Hadley Wickham)
- The caret package : https://topepo.github.io/caret/ (Max Kuhn)



Here is one way to solve a problem …

ALICE : 2 ■ + 3 ▲ = 5

$$\begin{bmatrix} 2 & 3 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} S \\ T \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \end{bmatrix}$$

Alice and Bob are BOTH able to compute the solution for S and T independently, eg by calculating the inverse matrix. Thus, S = 1 and T = 1.

BOB : ■ - 2 ▲ = -1

And a different way of solving the same problem …

ALICE : 2 ■ + 3 ▲ = 5

BOB : ■ - 2 ▲ = -1

Alice proposes [4,-1] (error 0)

Alice counter-proposes [3,0] (error 1)

Alice accepts [1,1] (error 0)

“Reconciliation chamber”

Counter-proposes [2,2]

Counter-proposes [1,1]

The global model has to be [1,1]!

Bob proposes [5,3] (error 0)

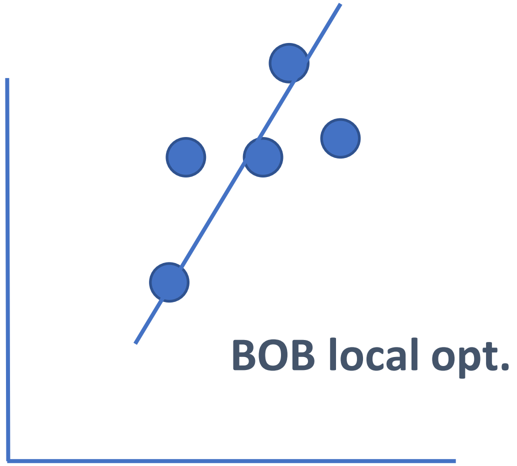
Bob counter-proposes [2,1] (error 1)

Bob accepts [1,1] (error 0)

Both approaches end up being valid

Because in the presence of RANDOM NOISE, both approaches change this into **an optimization problem because the perfect solution will not exist!**

IE : ORDINARY "LEAST-SQUARES-OF-ERRORS" STATISTICAL REGRESSION



$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

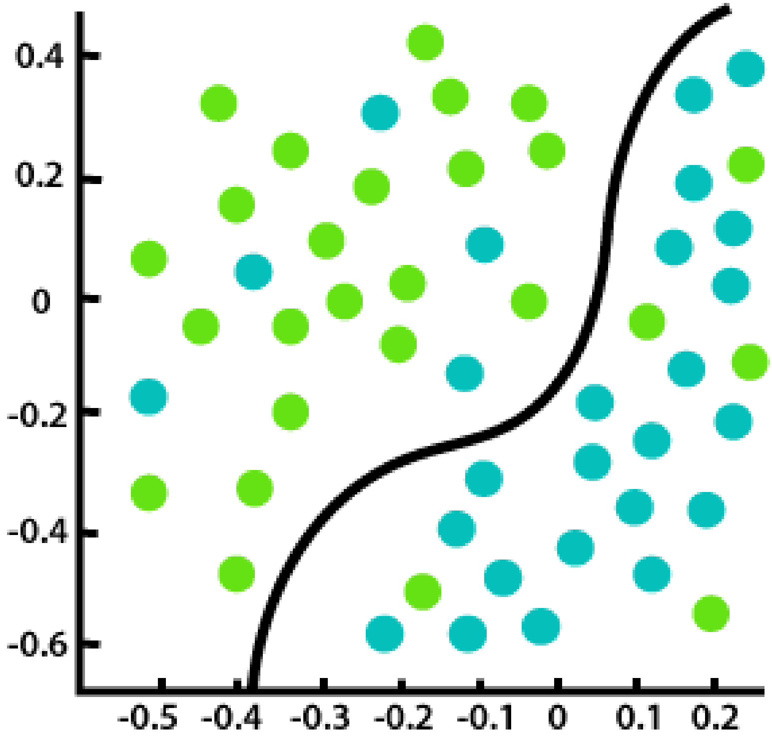
$$Y = X\beta + \varepsilon$$

Observations **Outcomes**

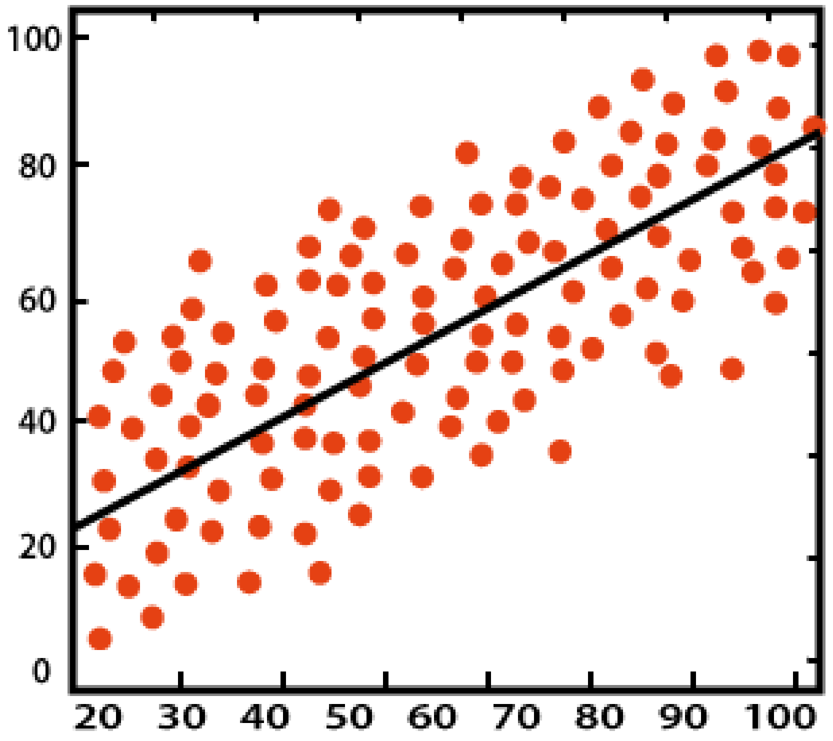
$$\hat{\beta} = (X^T.X)^{-1}X^T.Y$$

*** MODEL**

Important : Regression and classification



Classification



Regression

Important : Parameter estimation and prediction

One of the key aims of Statistical Modelling :

"Obtain the most accurate estimation of the parameters describing the true relationship between the control variable(s) and the response variable(s)".

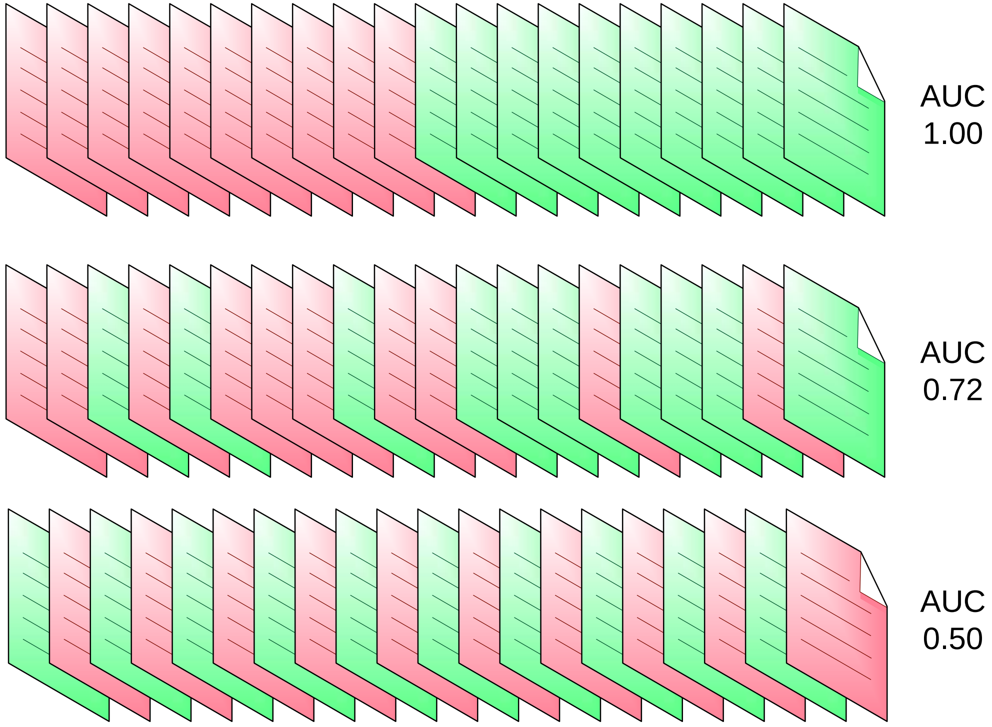
e.g. the true Hazard Ratio / Odds Ratio of serum cholesterol for heart attack

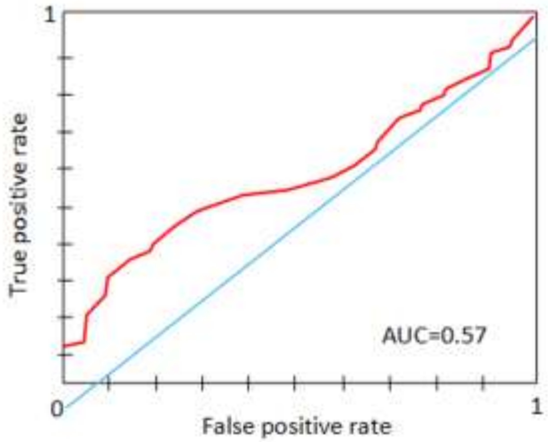
One of the key aims of Machine Learning :

"Obtain the most accurate prediction of the expected outcome for an individual having certain initial characteristics."

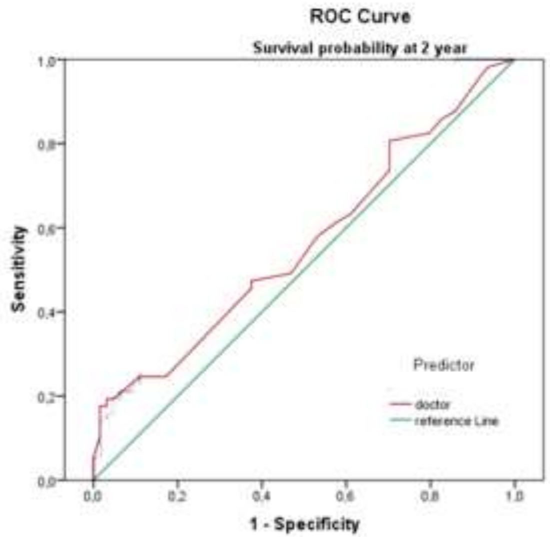
e.g. whose cancer is going to return within 2 years after having surgery

Statistic for measuring predictive performance





NSCLC (Lung Cancer)
2 year survival
30 patients
8 MDs
Retrospective
AUC: 0.57



NSCLC (Lung Cancer)
2 year survival
158 patients
5 MDs
Prospective
AUC: 0.56

A prospective study comparing the predictions of doctors versus models for treatment outcome of lung cancer patients:
A step toward individualized care and shared decision making.
C Oberije et al., https://doi.org/10.1016/j.radonc.2014.04.012

Classical modelling	Machine learning
Most accurate parameter estimation possible	Most accurate individual-level prediction possible
Hypothesis driven	Data driven
Mechanistic / knowledge basis	Mostly pattern recognition and phenomenological
Mostly (not all) parametric models	Parameterization is potentially useful but not essential
Assumptions can be strong	Assumptions are generally weaker
Prominent (pre-eminent) role of pre-existing knowledge and domain expertise	Vastly reduced role of pre-existing knowledge and domain expertise
Usually easily intuitive / interpretable	Can be explainable but generally less interpretable
Hypothesis testing	External validation

Examples of research questions

Classical modelling

- If I reduce the number of hospital beds by 10%, what is the expected increase in mortality?
- After correcting for age and sex, is there a relationship between serum cholesterol and heart disease?
- If I wish to raise my market share by 10%, should I spend money on advertising the existing credit card or launch a new credit card?
- Will a social-media based public health campaign reduce smoking rates among young women by 25%?

Machine learning

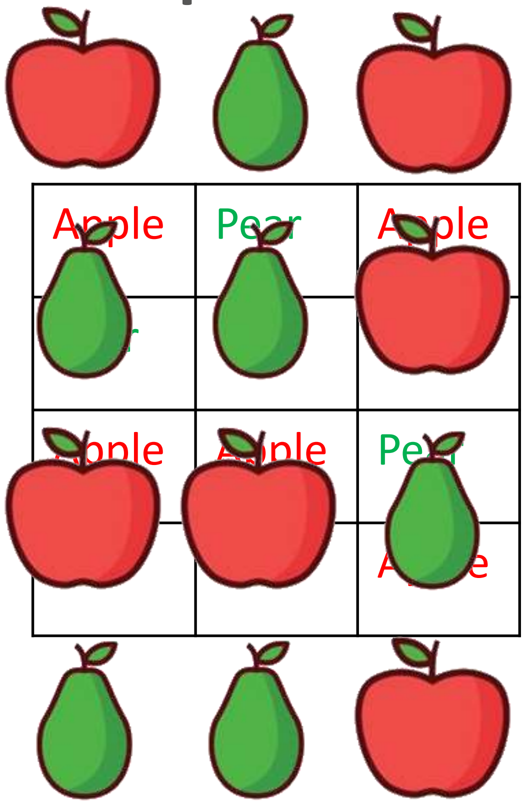
- If I give this particular Br Ca patient in front of me a higher radiation dose to the breast, does her cancer recur within the next 2 years or not?
- Should I put this person with these specific blood level readings on a “surveillance” plan for heart attack?
- This customer is right now asking for a home loan of $200k, shall I approve it or not?
- Which order of ads should I present on the smartphone of this young woman to encourage her to quit smoking?

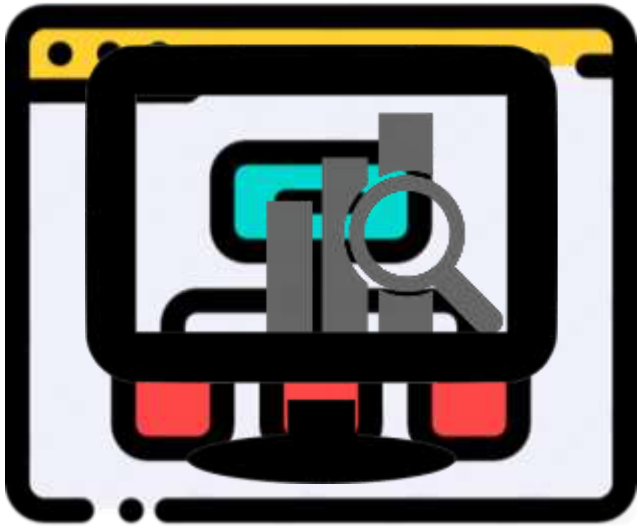
Two distinct flavour of “models”

- **Predictive**
  - Generates independently testable “predictions” using **LABELLED** data
  - Sometimes referred to as “supervised learning”
- **Discovery**
  - Searches for relationships among variables using **UNLABELLED** data
  - Sometimes referred to as “<u>**un**</u>supervised learning”

Supervised machine learning

Input





Machine learning algorithm

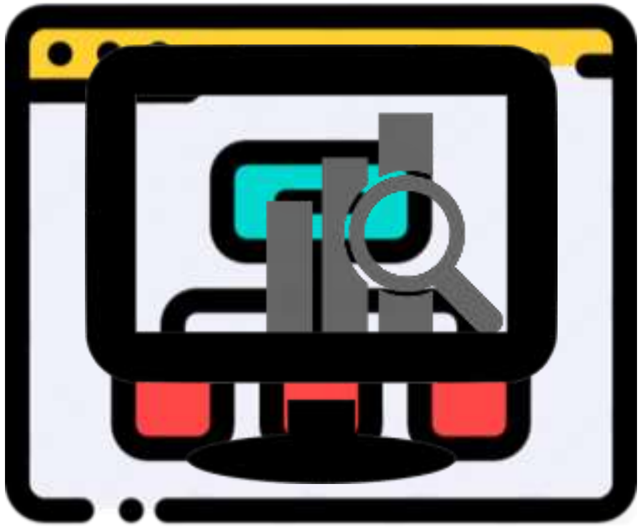
"A <u>deterministic</u> <u>process</u> that <u>adapts its internal state through repetitions</u> to produce a particular response <u>without being specifically programmed to do so</u>."

Supervised machine learning

Input

Output

?



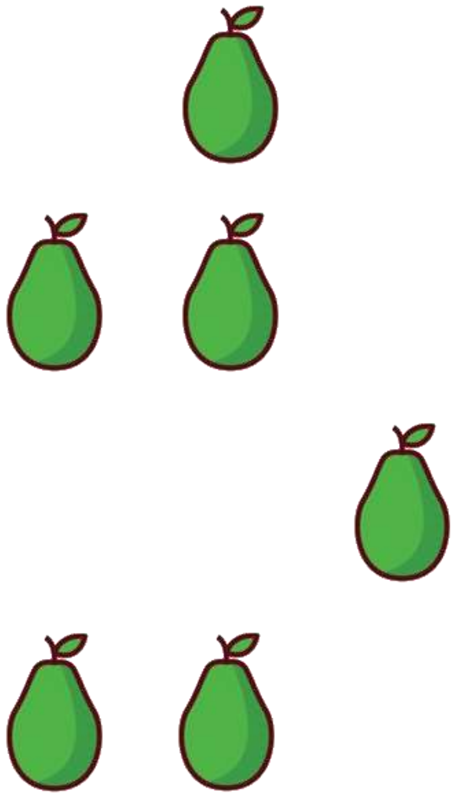
Machine learning algorithm

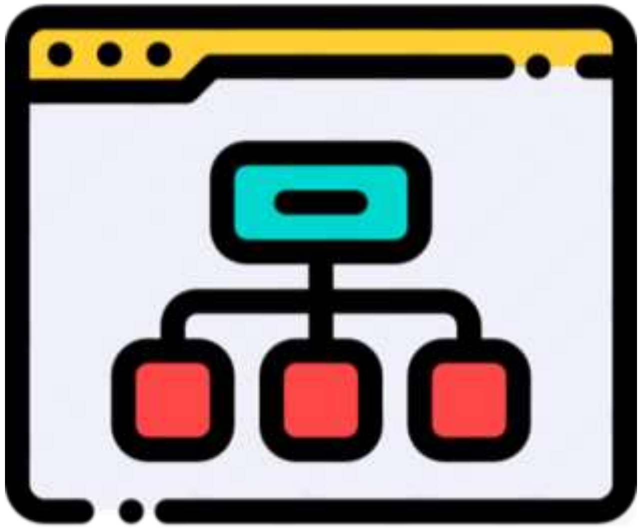
!Pear

“A deterministic process that adapts its internal state through repetitions to produce a particular response without being specifically programmed to do so.”

Supervised machine learning

Input





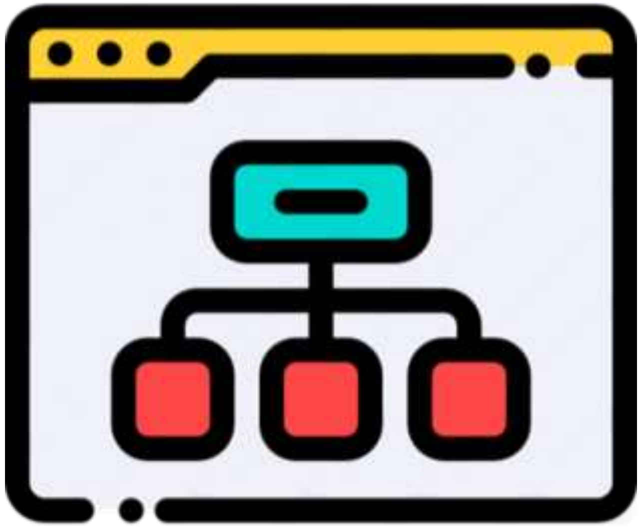
Machine learning algorithm

"A deterministic process that adapts its internal state through repetitions to produce a particular response without being specifically programmed to do so."

Supervised machine learning

Input

?



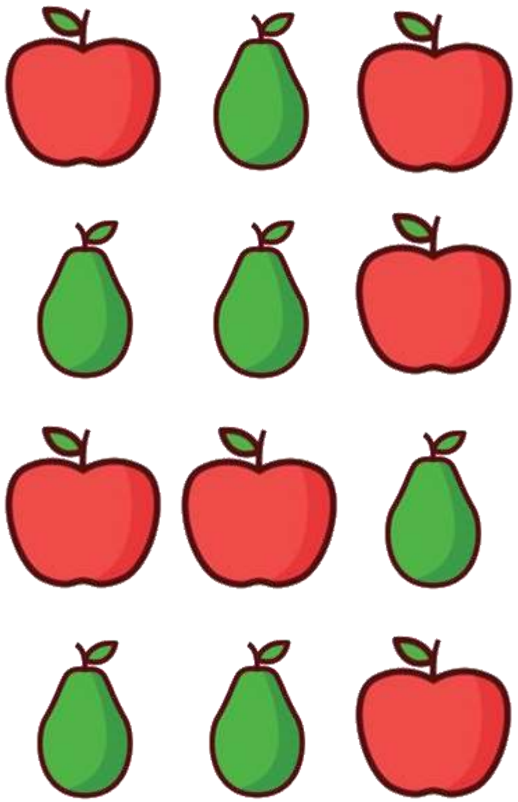
Machine learning algorithm

Output

! Pear

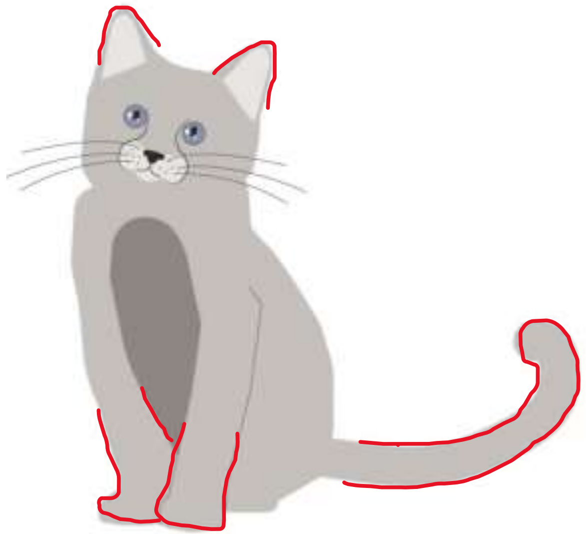
“A deterministic process that adapts its internal state through repetitions to produce a particular response without being specifically programmed to do so.”

What <u>feature</u> is this machine using to "diagnose" apple?



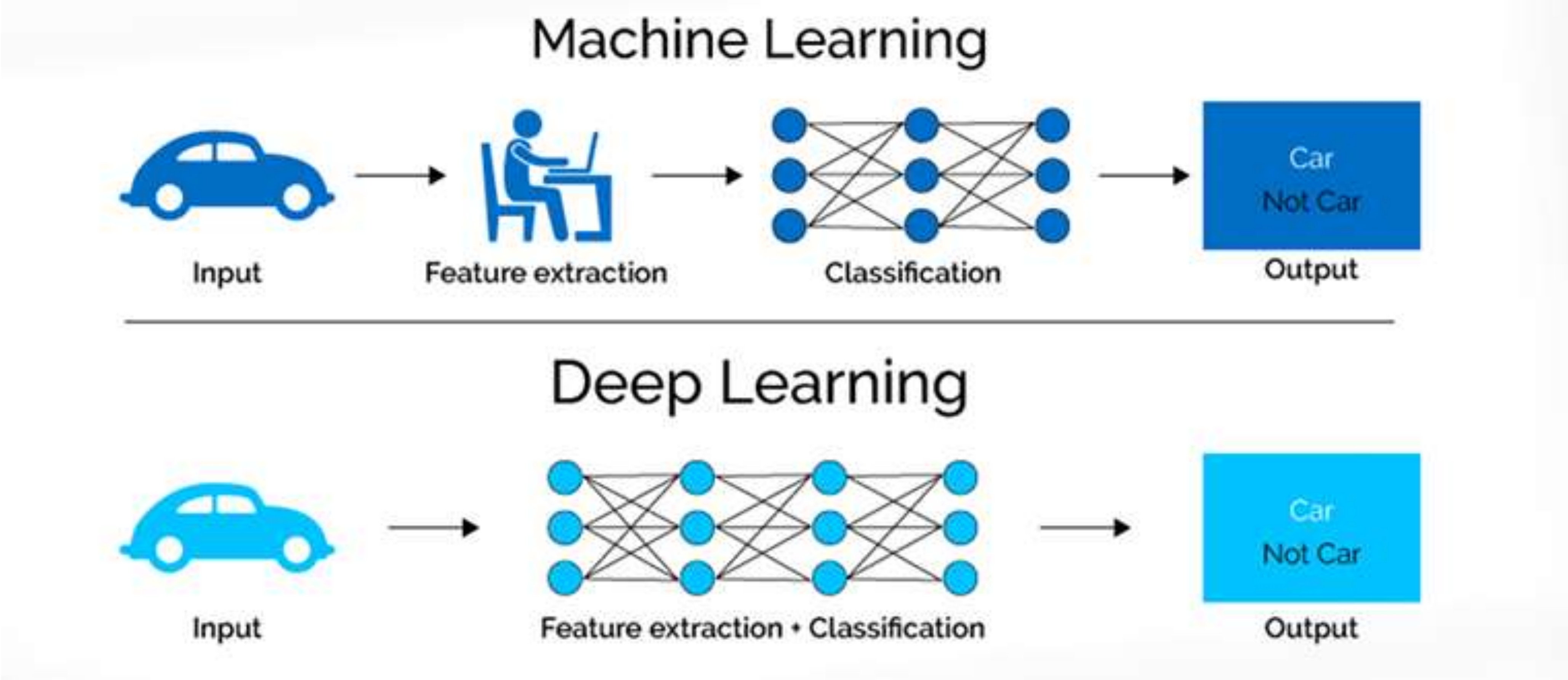
- Redness?
- Sphericity?
- Left-sided leaf?

What features to define “CAT”?

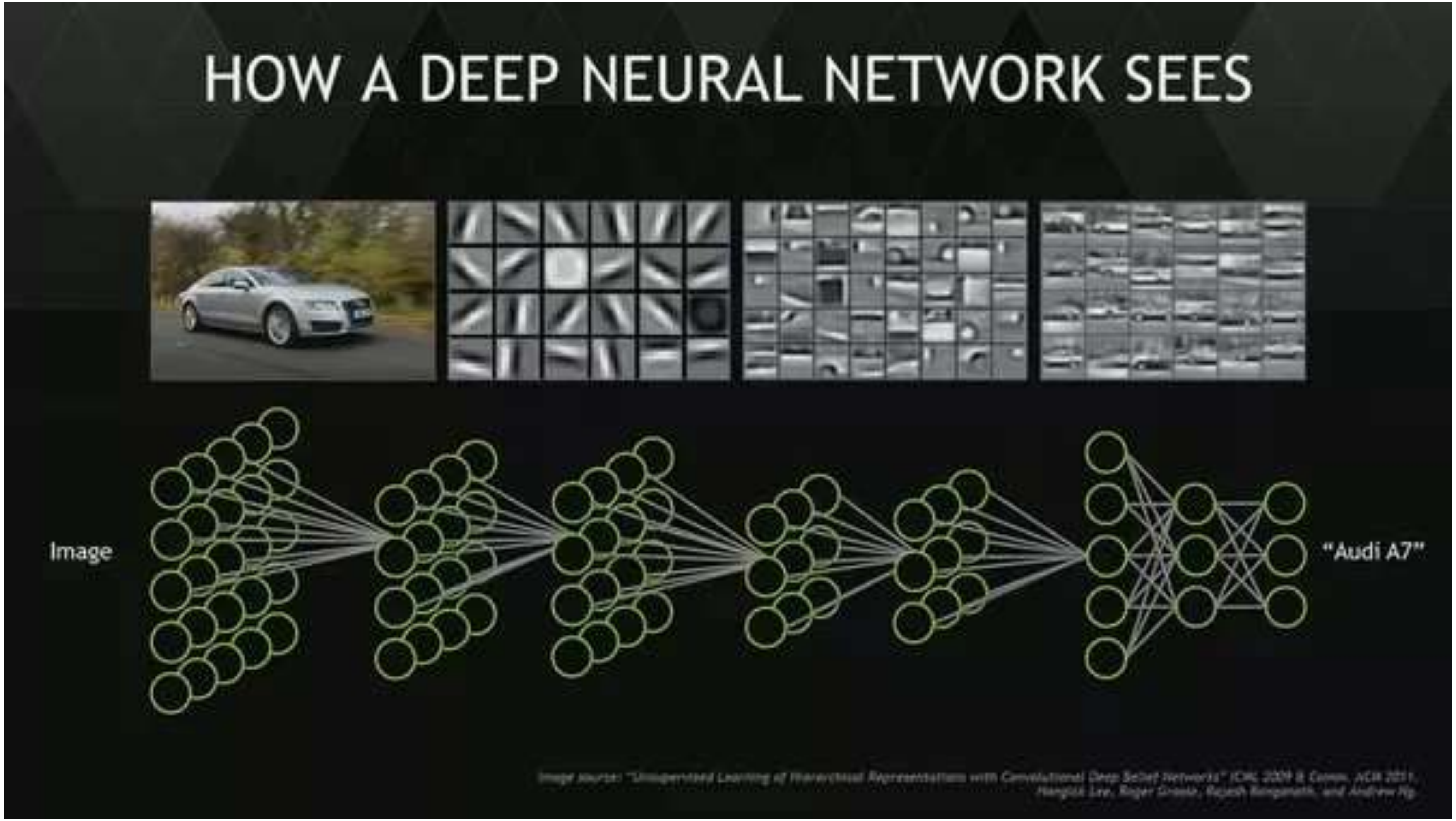


What features needed to define “DOG”?

Features can be human-defined or derived spontaneously from the data



Can you explain the features needed to classify a “car”?

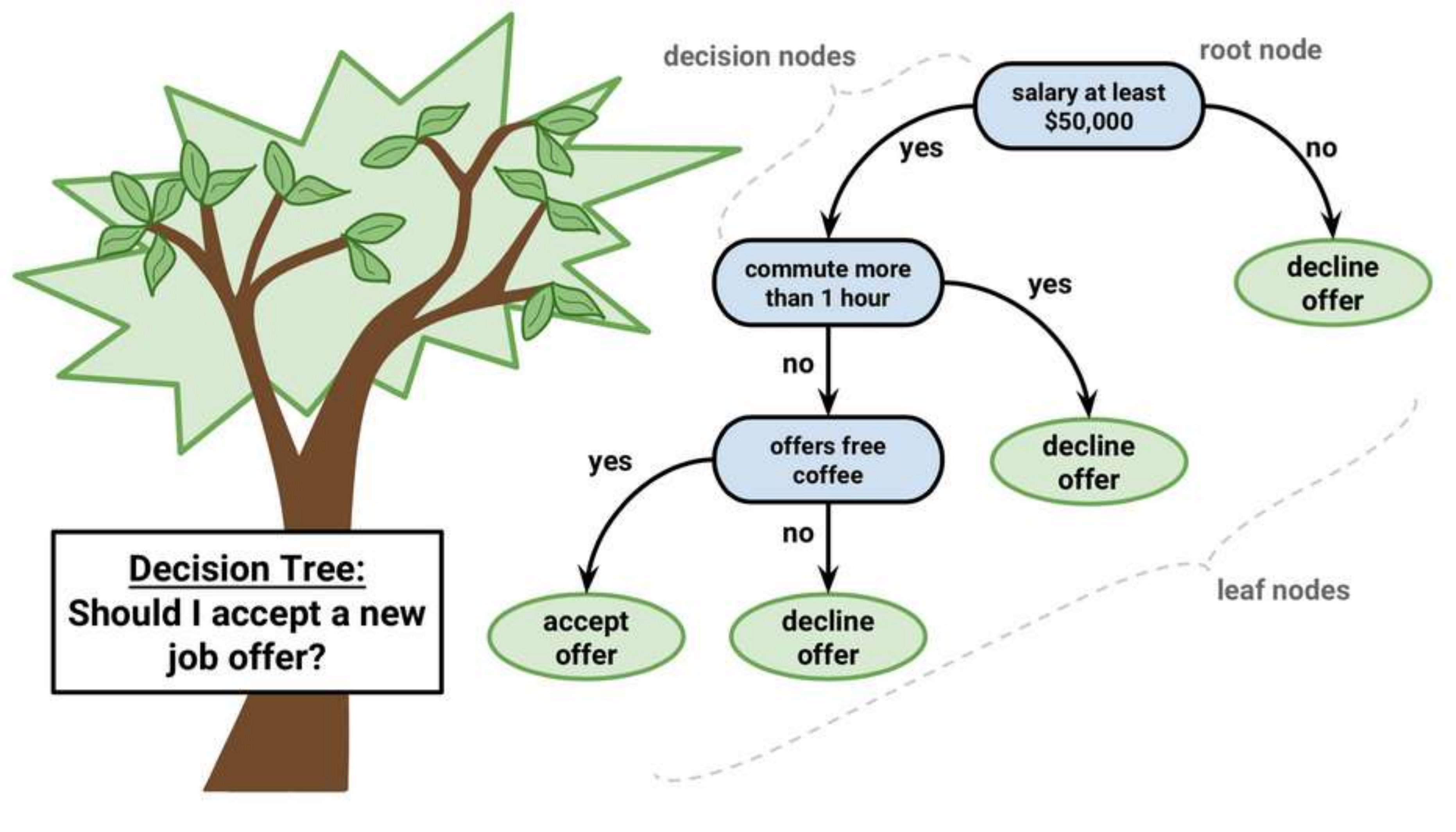


Supervised machine learning

- One logistic regression classifier, machine learning style, as example
- Trees and forests in the R practice session
- Vector (higher dimensional) spaces in the R practice session
- Very simple artificial neural net in the R practice session
- **Install “caret” R library in advance of the practice session**

Brief change to markdown and html

Tree-based classification

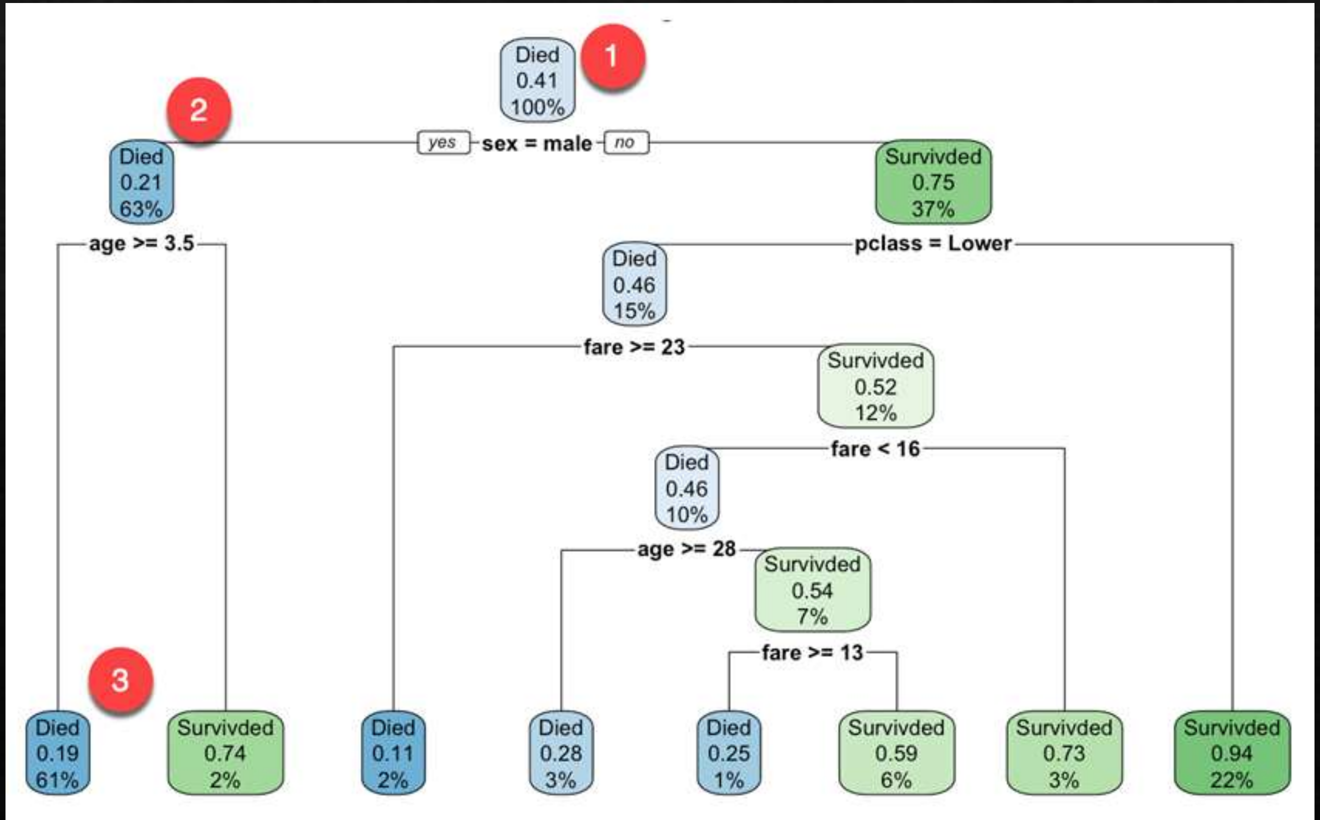


Tree-based classification

- Tree-based models use a series of **consecutive if-then rules** to divide up the initial input group into multiple smaller subgroups.
- Machine-learning based tree models **do not require humans to manually select** the thresholds of the if-then rules, or even which sequence of consecutive if-then rules to apply.
- Tree-based models can be constructed for either regression (predicting numerical values) or **classification** (predicting categorical values).

Tree-based classification

- Assuming the Titanic is representative of ship sinking events, **predict** - given some basic passenger characteristics - whether he/she will survive.



Tree-based classification

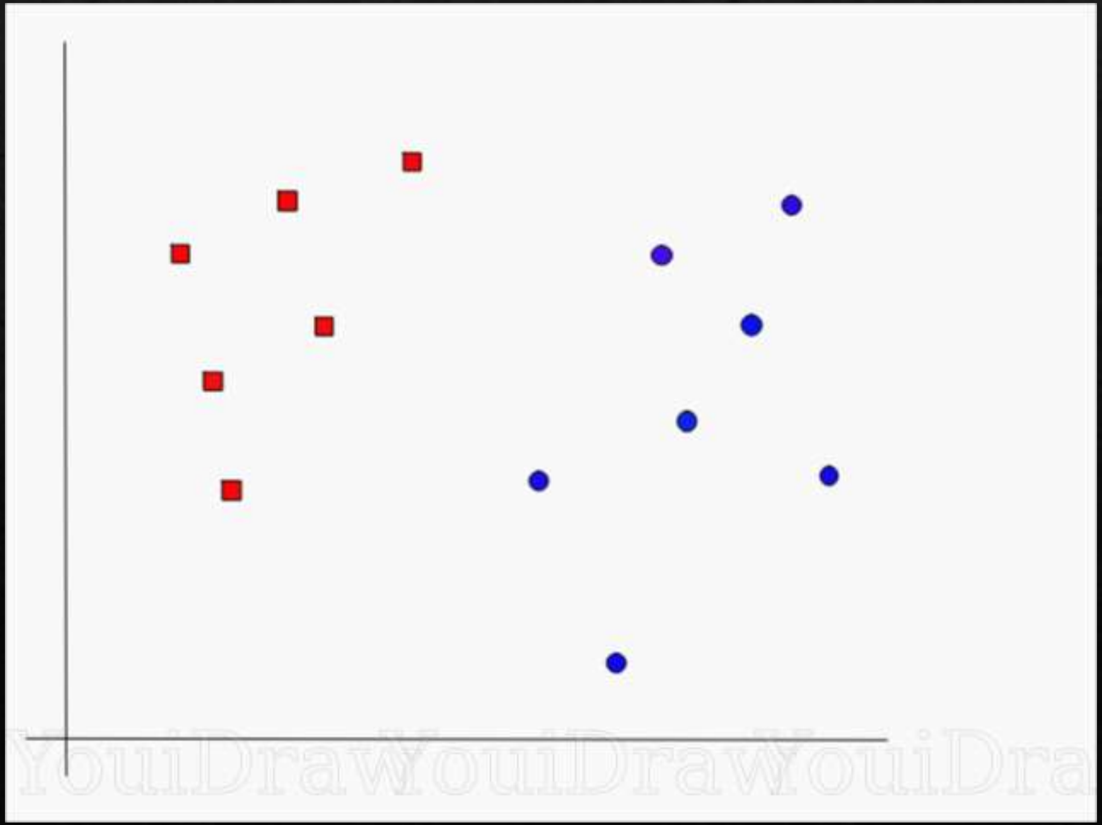
- Note similarities with UNSUPERVISED CLUSTERING from Sander's module?
- Main principle here : in any given tree and each of its sub-tree(s),
  - **Minimize entropy** within subgroups, **and**
  - **Maximize information gain** ie biggest Δ(entropy) by choosing a threshold to split to subgroups
- **We need an entropy metric** → Gini (class heterogeneity) index (happens to be default in the R package rpart)
- Human guidance need to decide where to **PRUNE THE TREE** to avoid fitting model on singular (or very few) events
- The main problem with trees is generally poor predictive performance in an unseen sample (external validity)

Random forests

- The best way to think of Random Forests (RFs) is a very large ensemble of tree models.
- RF has the flexibility to train on many random sub-samples of the dataset, one tree each
- RF has the flexibility to train on many randomly picked subsets of the features / parameters, one tree each
- Usually use the combination of both kinds of random subsets above (hence random forest)
- The prediction by the random forest model is therefore the consensus of a very large number of unique individual trees

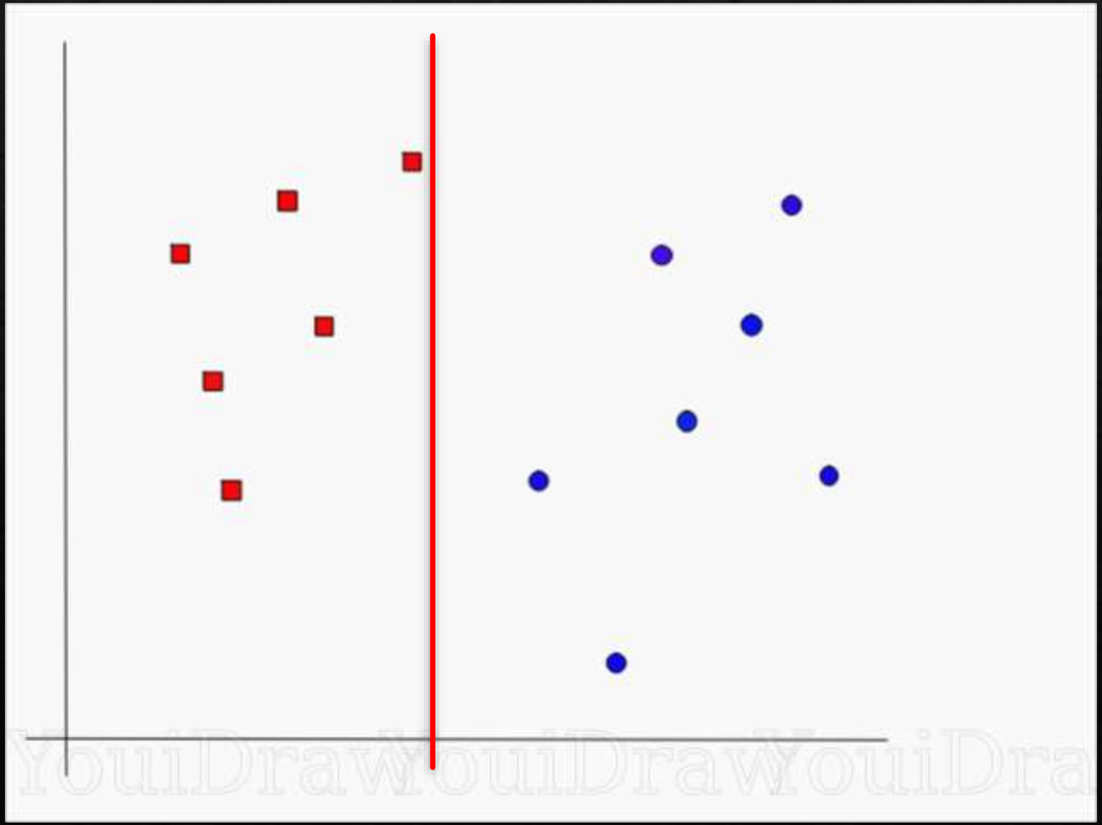
Support vector models

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



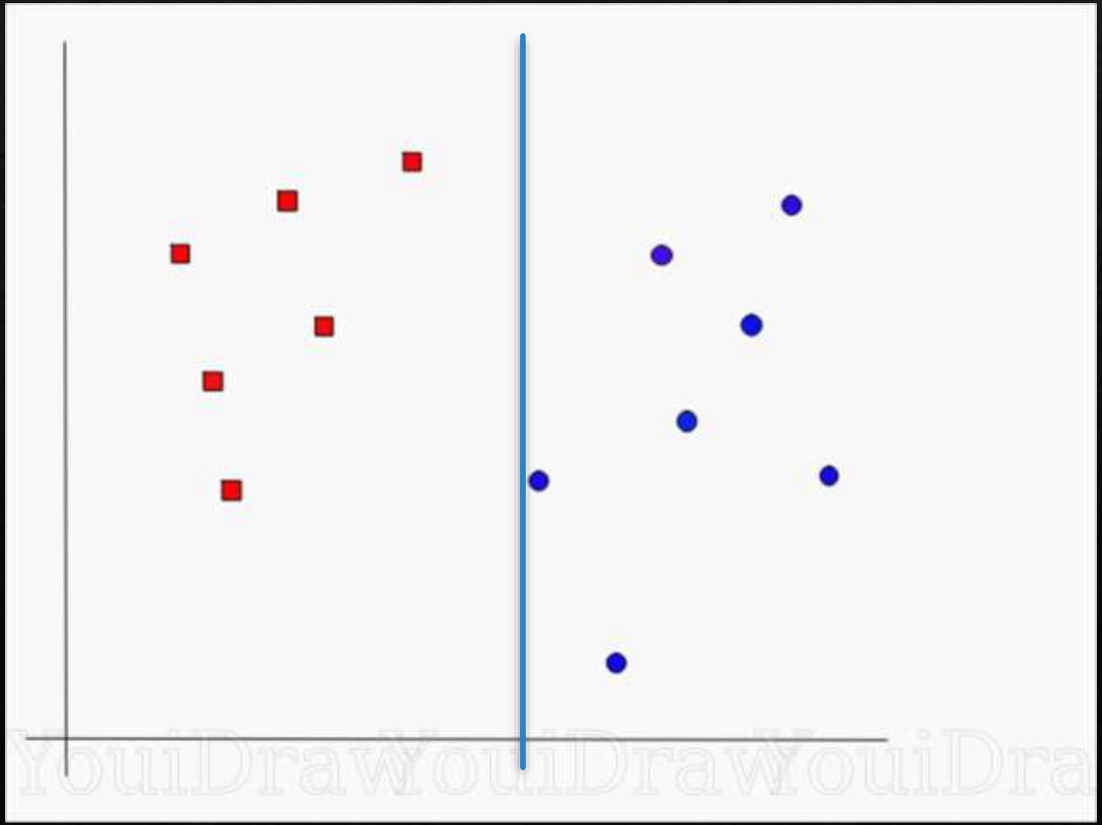
Support vector models

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



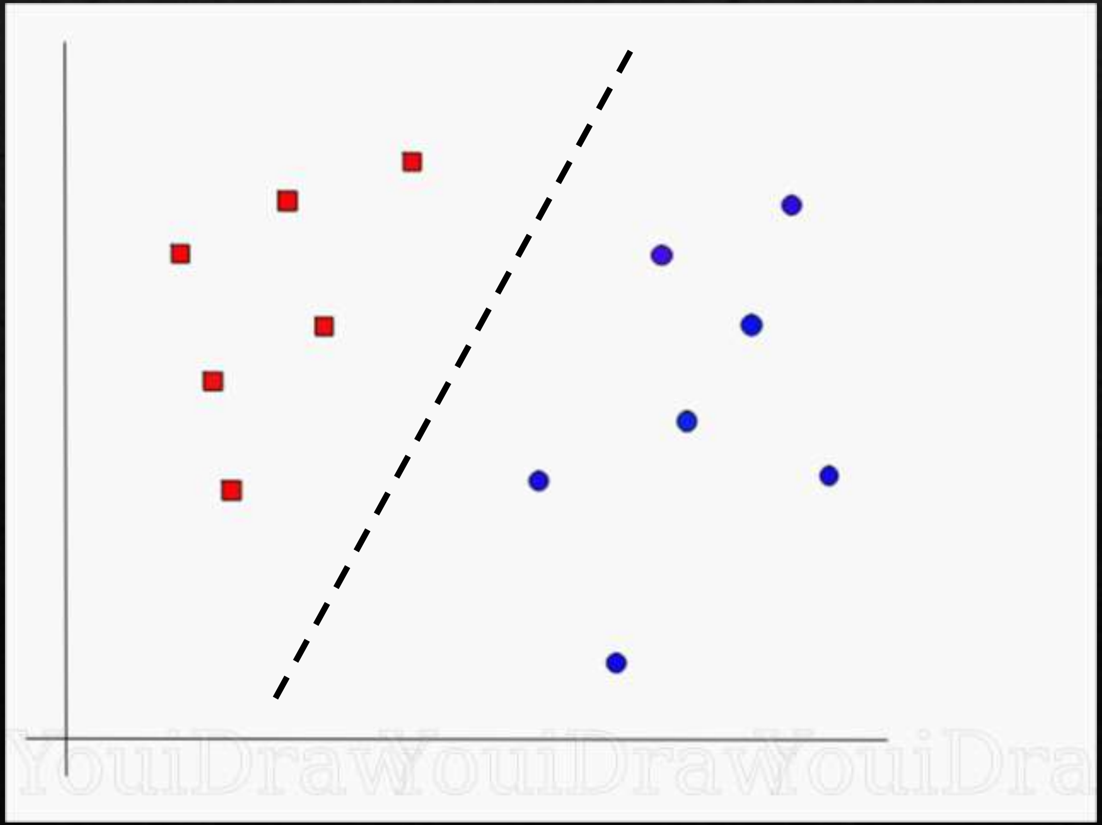
Support vector models

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



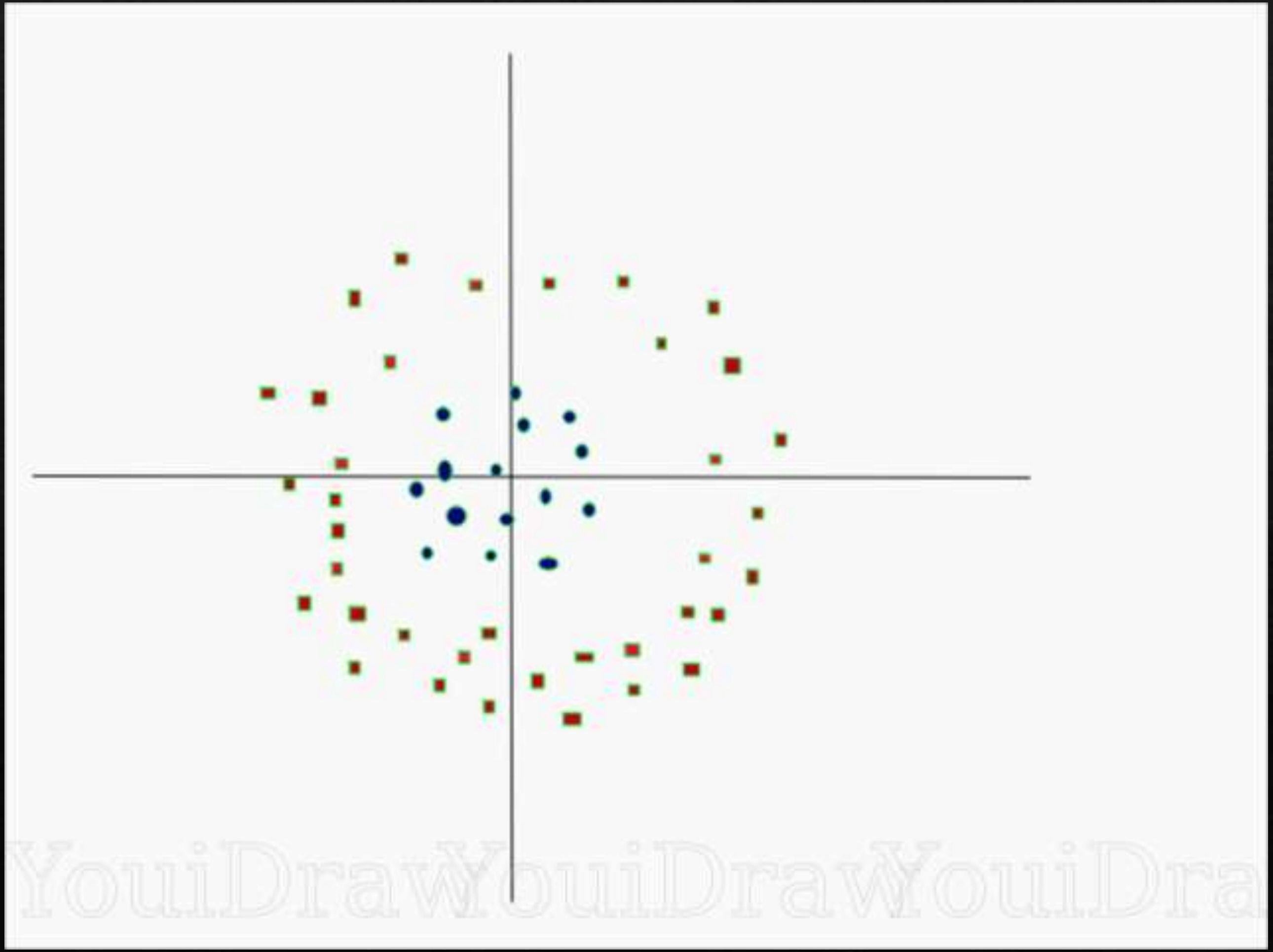
Support vector models

- Example : Draw a classification line to divide the “red” outcomes from the “blue” outcomes.



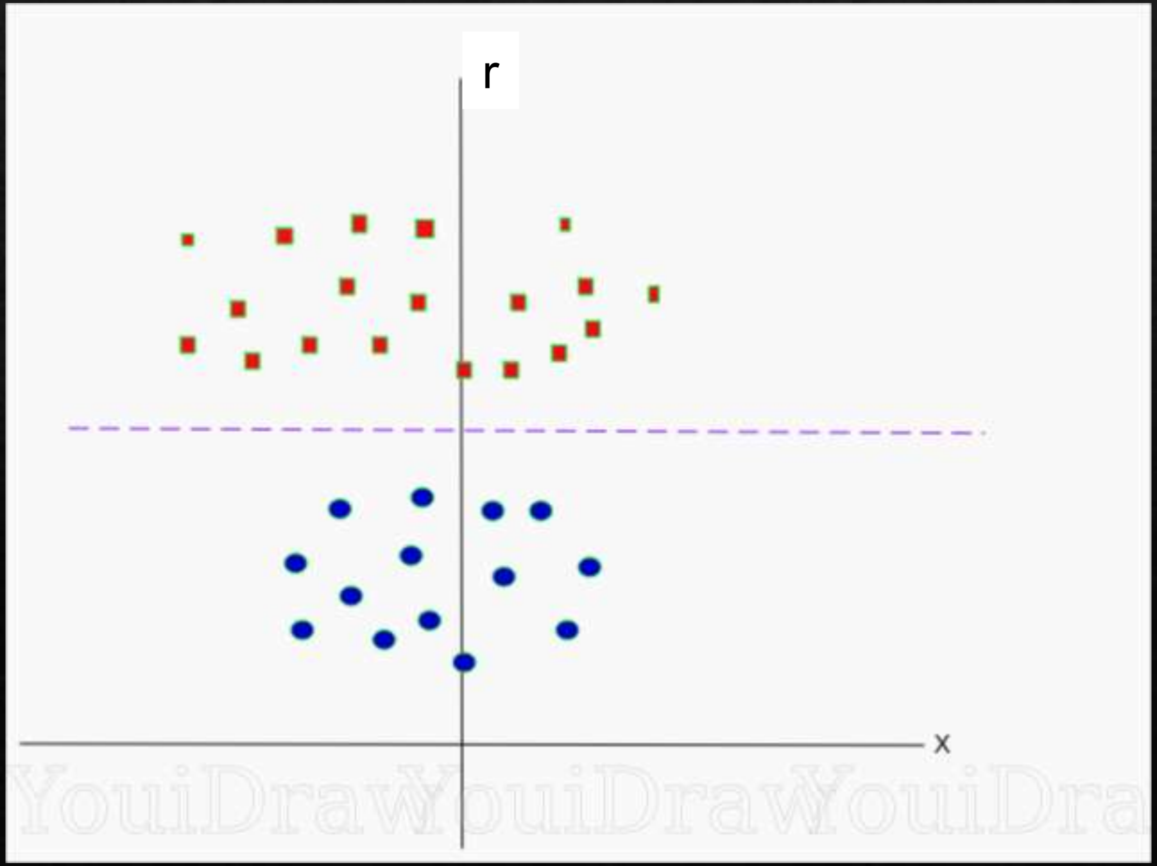
Support vector models

- Now try this one …



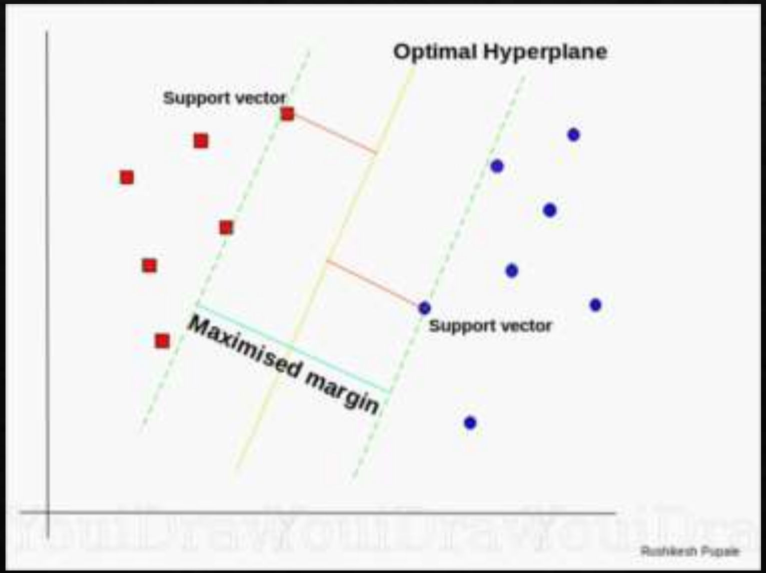
Support vector models

- Effect of a transforming from a Cartesian to a Radial coordinate system …

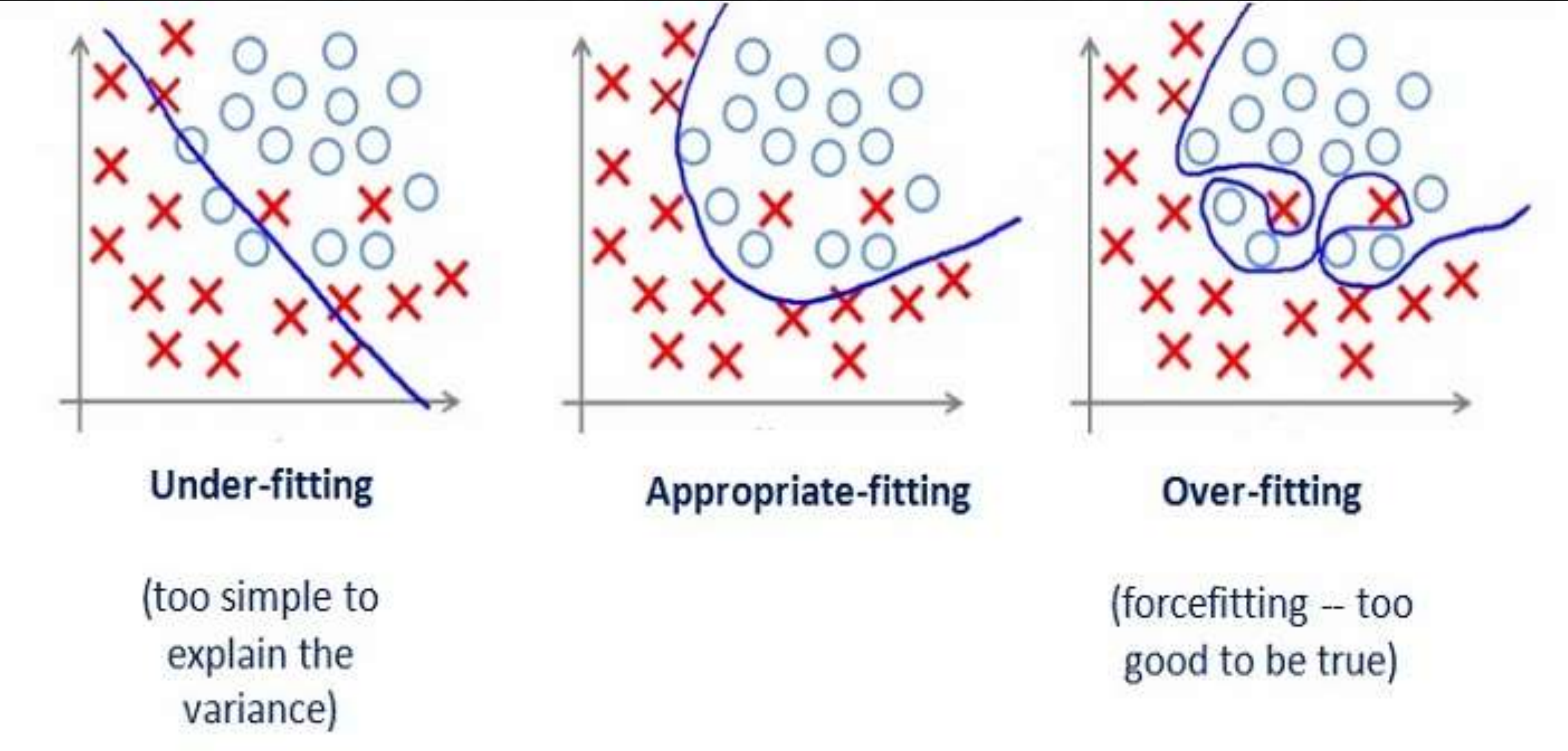


Support vector models

- Hyperplane : “In n-dimensional space, the hyperplane is a (n-1) dimensional surface that cuts the space into two distinct regions.”
- An SVM **optimal hyperplane** is the one that divides the data points so that the labels have minimum overlap and maximum distance from the hyperplane.
- **Linear SVM** = n-dimensional Cartesian space
- **Radial SVM** = n-dimensional Spherical space



Support vector models - caution

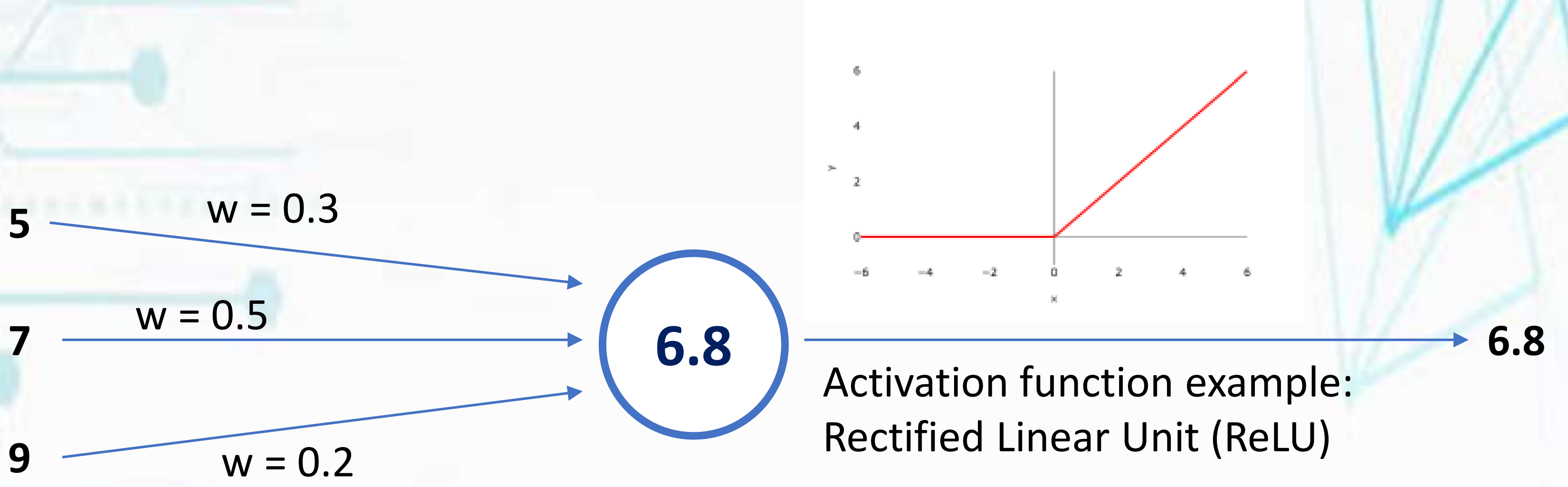


Introduction to neural networks

Artificial neuron

** Idea was inspired by a simplified mathematical model of how animal brains work, but it is NOT supposed to explain how real brain cells actually work.*

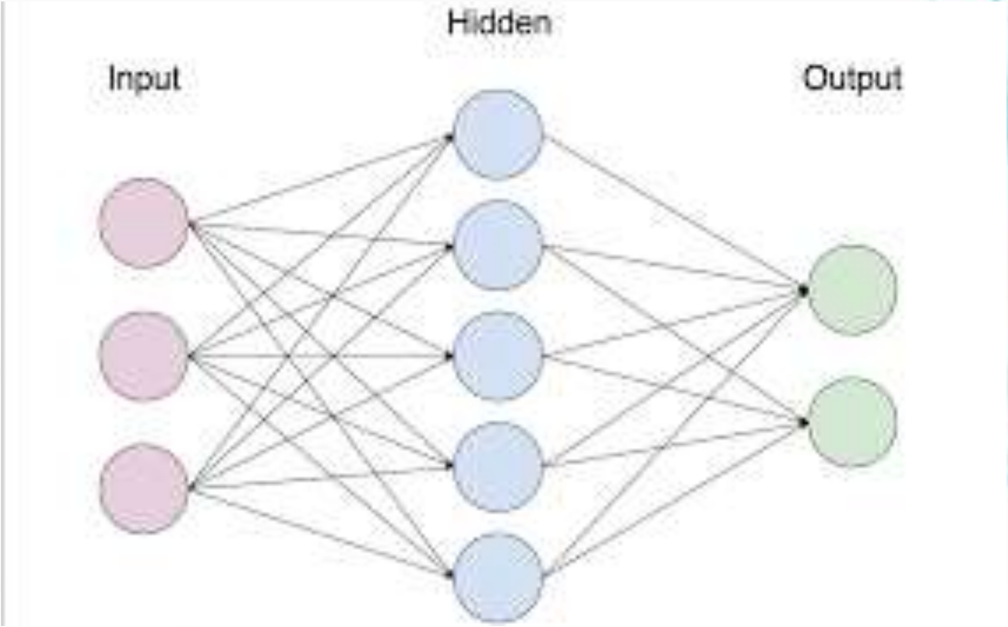
An **artificial neuron** is a mathematical operator that takes the **weighted sum** of all its input values, and calculates the output value using an **activation function**.



Activation function example:
Rectified Linear Unit (ReLU)

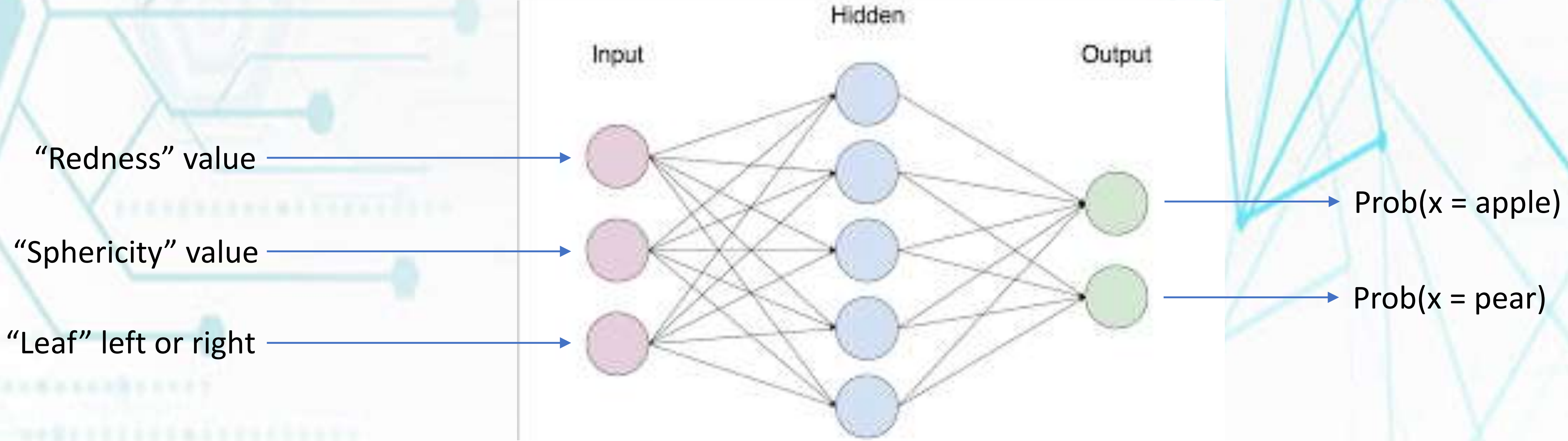
Putting neurons together = neural network

An **artificial neural network (ANN)** consists of layers of artificial neurons that do not interact with its neighbours but only **feeds forward** its own output to the next layer.



The universal function approximator theorem

- <u>Any</u> continuous real-number mathematical function in <u>any</u> finite number of dimensions, no matter how complicated …
- can be **<u>approximated</u>** by a single feed-forward neural network with only a <u>finite number of artificial neurons in a single “hidden” layer</u> between the input layer and the output layer.



Coming up in practice session :

- Trees and random forests
- Support vector model
- Very simple artificial neural network
- **Please install “caret” R library in advance of the practice session**
- **Keep caret documentation nearby : https://topepo.github.io/caret/**

Reminder : Learning Objectives

What is ML

Understand the difference in approach and guiding philosophy, rather than any specific tool or situation

Classification

Trees and Forests

Vector spaces

Artificial neurons

Comparison

Compare multiple models