# Machine Learning

## 1. Introduction

Mathieu Emily

Département de statistique et informatique
Agrocampus Ouest, Rennes

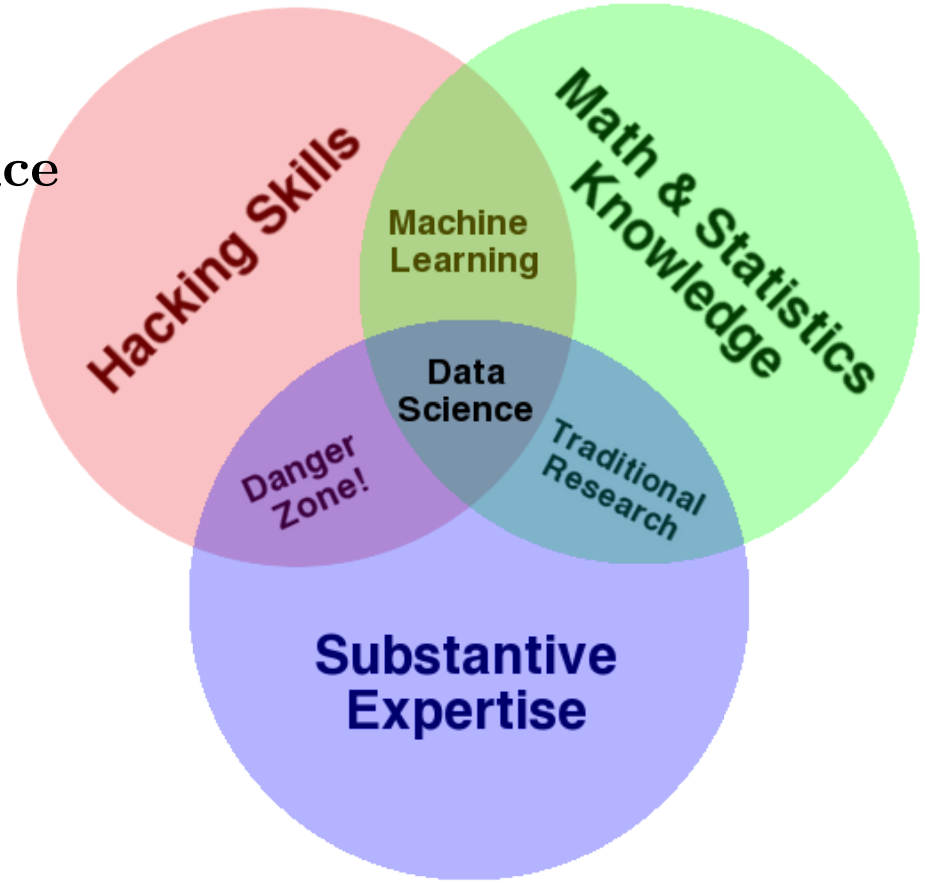
## Outline

## Outline

## Data science

Computer science

Statistics




Domain science

*Drew Conway*

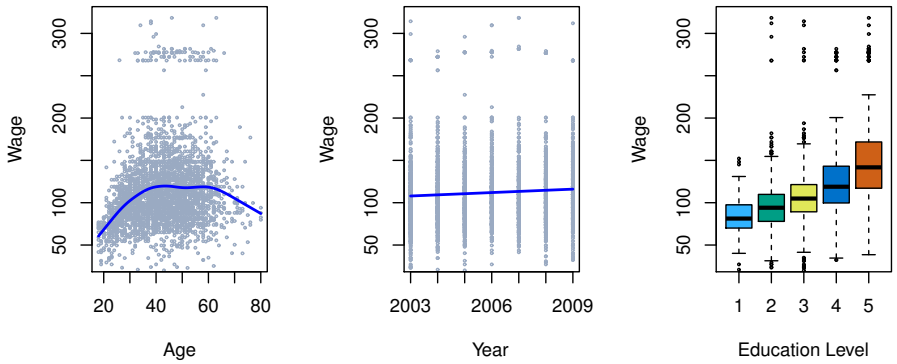
## What is machine learning ?

> Machine learning refers to a vast set of tools for understanding data
>
> *G. James et al. (2013)*

- Illustration of *understanding data* through 3 examples :
  - Wage
  - Stock Market
  - Auto

## What is machine learning ? - Wage illustration

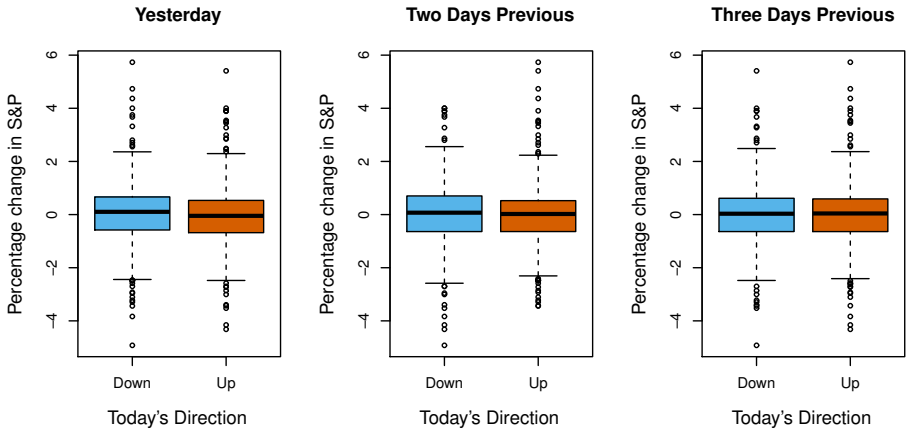
In this application, we examine 3 factors (`age`, `education` and `year`) that relate to `wages` for a group of males from Atlantic region of the US.




- Given an employee's age, can we use this curve to predict his wage ?
  - Yes : the trend is clear with the blue line
  - No : variability associated with the average value
- A good prediction should account for :
  - the most informative variables
  - the shape (linear, non-linear, ...) of the relationship between wage and the variables

## What is machine learning ? - Stock Market illustration

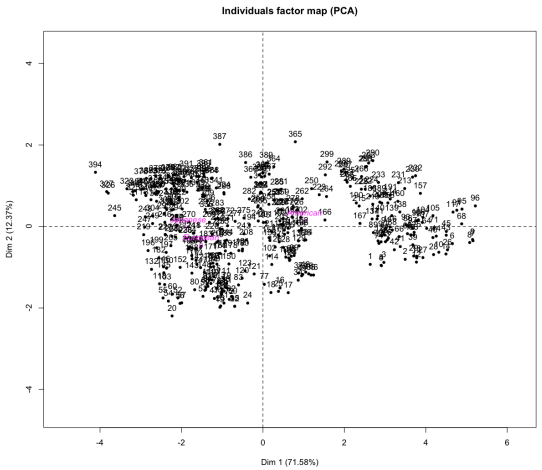
In this application, we examine daily movements in Standard and Poor's (S & P) stock index between 2001 and 2005. The goal is to predict whether the index will increase or decrease on a given day.

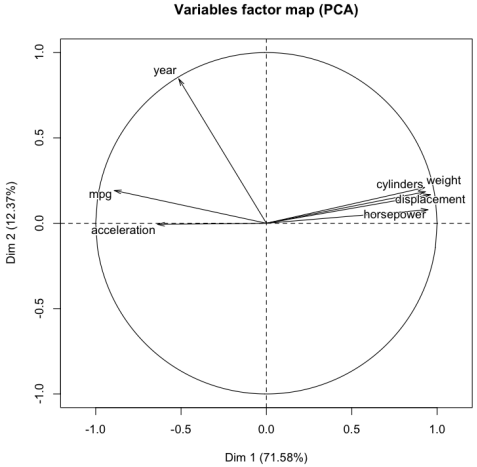


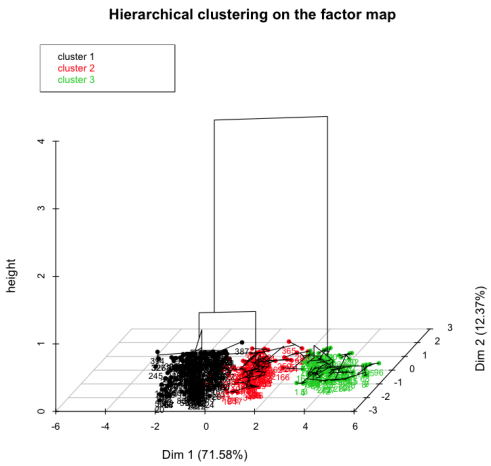

- The lack of pattern is expected : in the presence of strong correlations between successive days' returns, one could adopt a simple trading strategy to generate profits from the market
- Can we use the data to build a model that can predict the direction of movement in the market 60% of the time ?

## What is machine learning ? - Auto illustration

In this application, we examine the relationship between 7 characteristics of 392 vehicles (American, European and Japenese).



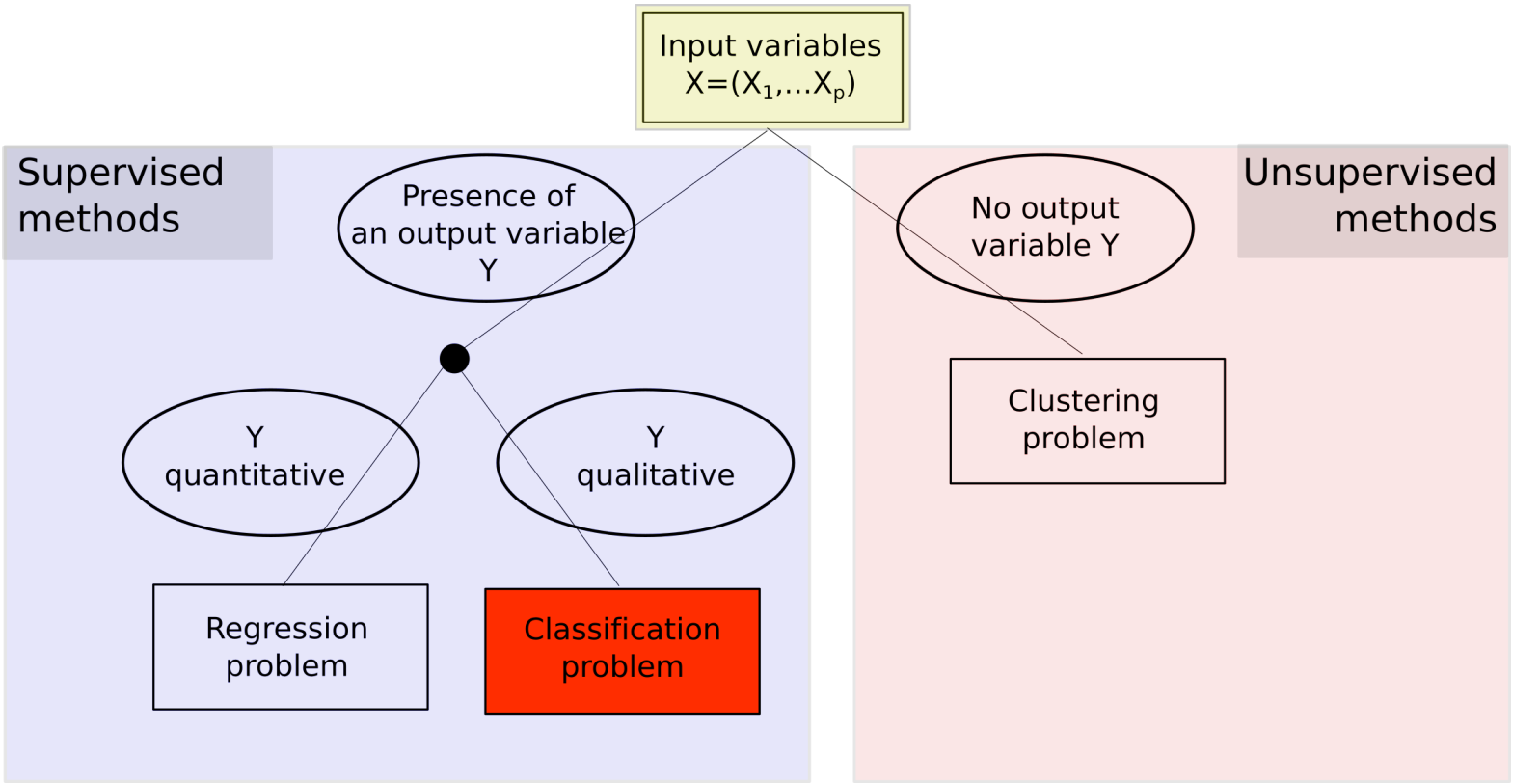




- Interpretation of the main variability between vehicles
- Clustering of the vehicles and interpretation with supplementary variables

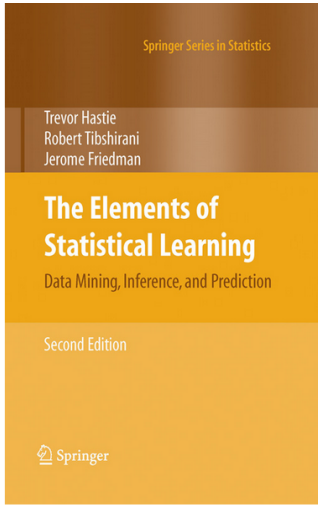
## Focus of this course

What are the differences between these three examples ?

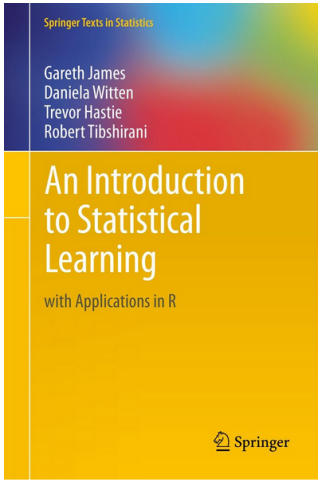
# Focus of this course



## References : two books (free pdf)




- ESL : The Elements of Statistical Learning *(Hastie, Tibshirani, Friedman, 2009)*




- ISLR : An Introduction to Statistical Learning *(James, Witten, Hastie, Tibshirani, 2013)*

## Outline

## Notations for the observed data (1)

- $n$ individuals
- $p$ input variables (or explanatory variables, independent variables, predictors, features, ...)

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

  $x_{ij}$ represent the value of the $j$th variable for the $i$th observation.
- 1 output (or response, dependent) variable

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

## Notations for the observed data (2)

- Let $x_i$ the $p$ variable measurements for the $i$th observation

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- Let $\mathrm{x}_i$ the $p$ variable measurements for the $i$th observation

$$\mathrm{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- Then

$$\mathbf{X} = (\mathrm{x}_1 \;\; \mathrm{x}_2 \;\; \ldots \;\; \mathrm{x}_p) = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

## Notations for the observed data (3)

- Observed data consists of

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}$$

  where $x_i$ is a vector of length $p$.
- Observations are realizations of random variables :

$$Y, X_1, \ldots, X_p$$

  with :
  - $Card\,(Y(\Omega)) = K$ (the cardinal of the support of $Y$ is $K$, *i.e.* $Y$ has $K$ categories).
  - $X = (X_1, X_2, \ldots, X_p)$ a $p-$dimensional random vector.

## Outline

## Classification : example ♯1 - default on credit card debt

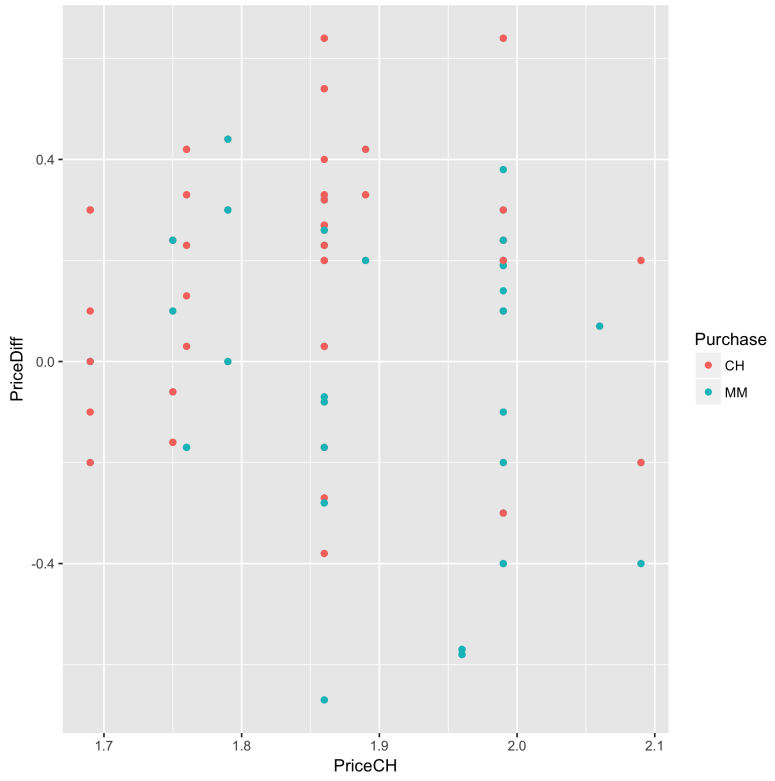
- $n = 10000$ individuals
- $p = 2$ variables : `balance` [a] and `income`
- Objective
  - Identification of defaults ($K = 2$)


a. The average balance that the customer has remaining o card after making their monthly payment

Explanation and prediction

## Classification : example ♯2 - Purchase of orange juice

- $n = 1070$ Observations
- $p = 2$ variables : `PriceCH` [a] and `PriceDiff`
- Objective
  - Compare the purchase of $K = 2$ orange juice




a. Price charged for Citrus Hill

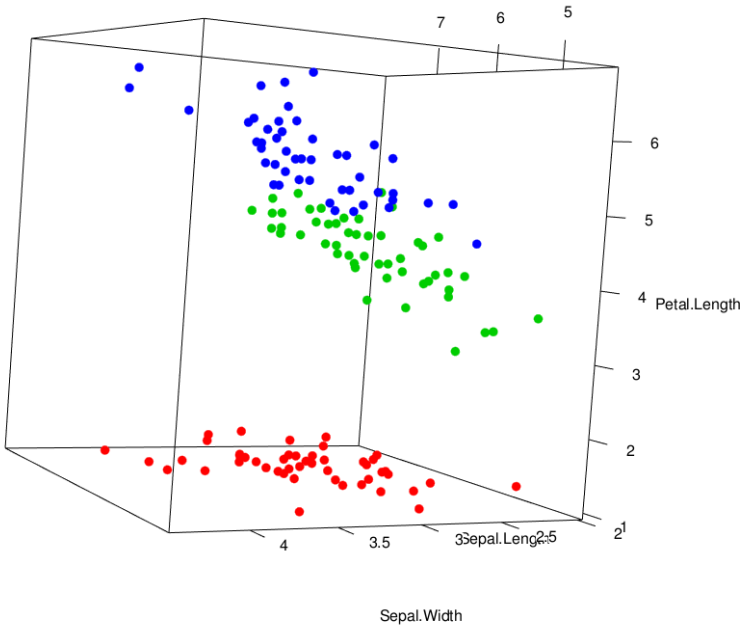
b. Sale price of Minute Maid less sale price of Citrus Hill

Explanation

## Classification : example ♯3 - The famous (Fisher's or Anderson's) iris data

- $n = 150$ Observations
- $p = 4$ biometrical variables
- Objective
  - Identification of $K = 3$ iris species

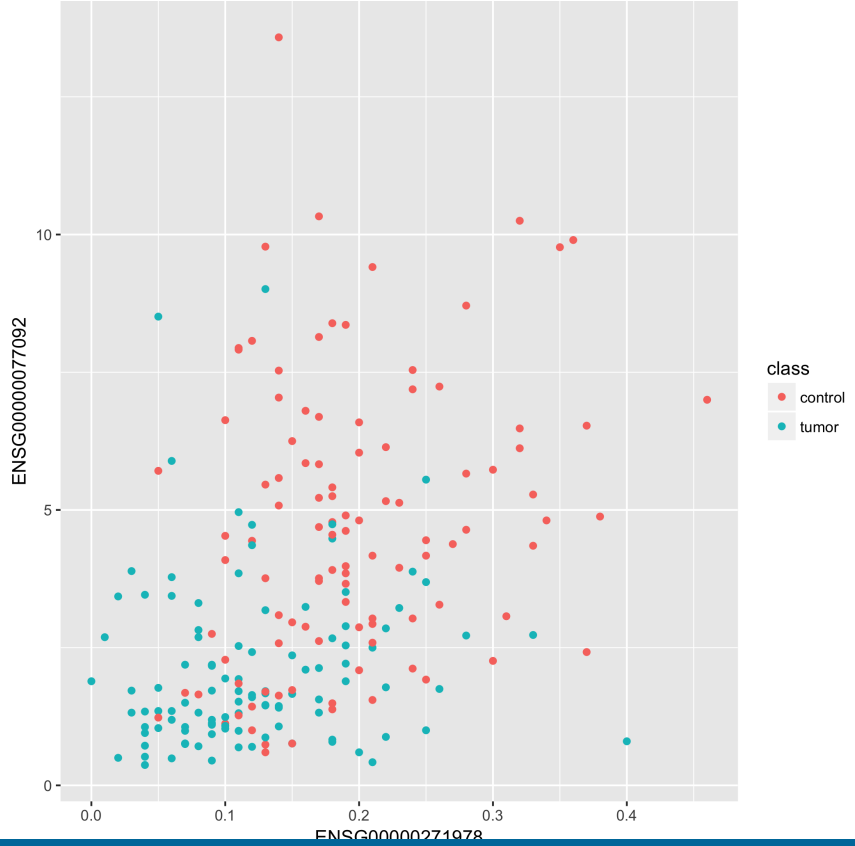
Explanation and prediction



## Classification : example ♯4 - cancer diagnosis

- n=212 individuals
- Expression level of $p > 30000$ genes
- Objective
  - Identification of biomarkers for developing cancer ($K = 2$)

Exploration, explanation, selection and prediction



## Statistical characteristics

- $\mathbf{Y}$ can have $K = 2$ or $K > 2$ categories
- Low ($n >> p$) or high ($n \leq p$) dimensional problem
- Objective of the statistical analysis :
  - Inference
  - Prediction
  - Both

## The “no free lunch theorem"

**Theorem**

*No one method dominates all others over all possible data sets*

- Many methods have been introduced and published with the argument :

  “My method outperforms competitors on my particular data set”

## The main objective of the course

1. Introduction of a series of learning methods
   - kNN
   - Logistic Regression
   - Discriminant analysis
   - Neural network
   - Classification tree (CART)
   - Random forests
   - Bagging and boosting
   - Support Vector Machine
2. Dessign of a methodology to answer the question :

   "How do I select the best approach for a given data set"

## The main directions of the course

1. Many statistical learning methods are relevant and useful in a wide range of academic and non academic disciplines.
   - Rather than considering every possible approach, I present the methods that I believe are most widely applicable
2. Machine learning should not be viewed as a series of black boxes
   - It is important to understand the cogs inside the box and the interaction between those cogs to select the best box
   - I attempt to describe the model, intuition, assumptions and trade-offs behind each method
3. While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside each box
   - I have minimized the technical details related to fitting procedures and theoretical properties
4. The interest is in applying machine learning methods to real-world problems
   - R is used throughout this course

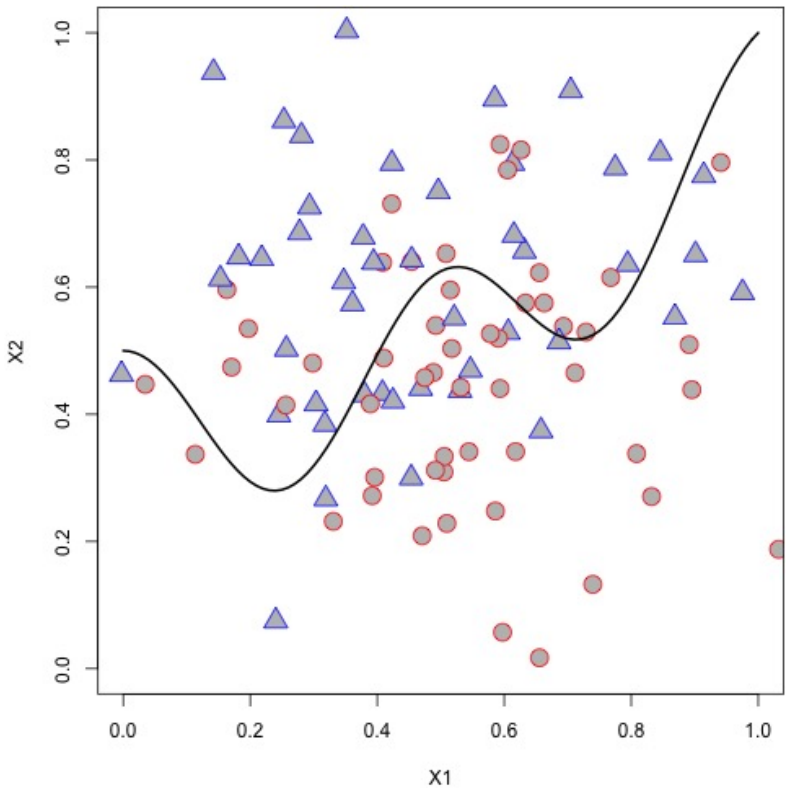
## Outline

## Model

- $Y = f(X) + \varepsilon$
  with $X = (X_1, \ldots, X_p)$
  - $f$ is some fixed but unknown function. $f$ is the systematic information that $X$ provides about $Y$
  - $\varepsilon$ is an error term

Machine learning refers to a set of approaches for estimating $f$

- In practice, estimating $f$ is performed to reach two main goals :
  - Prediction
  - Inference



## Objective 1 : Prediction

- $\widehat{Y} = \widehat{f}(X)$
- $\widehat{f}$ can be treated as a black box
  - As long as $\widehat{f}$ yields accurate prediction, we don't care about its form.
- Example :
  - Diagnosis of a patient based on blood sample characteristics

## Objective 2 : Inference

- The goal here is to understand the relationship between $Y$ and $X = (X_1, \ldots, X_p)$.
- $\widehat{f}$ cannot be treated as a black box : we need to know its exact form.
- Inference is used to answer the following questions :
  - Which predictors are associated with the response ?
  - What is the relationship between the response and each predictor ?
  - Can the relationship between $Y$ and each predictor be adequately summarized using a linear equation, or is the relationship more complicated ?

## Challenges in machine learning

Choosing the appropriate method for estimating $f$ depends on our ultimate (the so-called "statistical" question). Are we interested in

- prediction
- inference
- both : prediction and inference

For example

- Linear models allow for simple and interpretable inference but may not yield accurate predictions
- Highly non-linear approaches can provide accurate predictions for $Y$, but this comes at the expense of a less interpretable model for which inference is not possible.

## Up to you

TP1

## Outline

## Outline

## Bayes rule

- The main idea is to use the *posterior* probability

$$\forall Y \in [1, \ldots, K] \quad : \quad \mathbb{P}[Y = j | X = x_0]$$

- the *posterior* probability refers to the *conditional* probability of $Y$ given $X$

$$\mathbb{P}[Y = j | X = x_0] = \frac{\mathbb{P}[X = x_0 | Y = j]\mathbb{P}[Y = j]}{\mathbb{P}[X = x_0]}$$

  - $\mathbb{P}[Y = j]$ is the *prior* probability
  - $\mathbb{P}[X = x_0 | Y = j]$ is the emission law or the likelihood

In practice, the *posterior* probabilities are unknown. Most of the Machine Learning methods relies on the estimation of the *posterior* probabilities.

## Bayes classifier

**Definition**

*The Bayes classifier assigns each observation to the most likely class, given its predictor values*

- When $K = 2$ (a two-class problem) the Bayes classifier is given by :

  If $\mathbb{P}[Y = 1|X = x_0] > 0.5$ then assign class 1 and 2 otherwise

- When $\mathbb{P}[Y = j|X = x_0]$ is **known**, the Bayes classifier has optimal statistical properties

## Outline

## kNN classifier

### Definition

*Let $x_0 = (x_1^0, \ldots, x_p^0)$ be a point in the observed space and $K$ an integer. The kNN first identifies the $K$ observed points that are closest to $x_0$, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class $j$ as the fraction of points in $\mathcal{N}_0$ whose response values equal $j$ :*

$$\mathbb{P}[Y = j | X = x_0] = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{I}(y_i = j)$$

*Finally, kNN applies the Bayes rule and classifies the test observation $x_0$ to the class with the largest probability.*

## kNN classifier : generalities

- The kNN is a non-parametric method that aim at estimating the conditional distribution of $Y$ given $X = (x_1, \ldots, x_p)$
- Despite it is very simple, kNN can often produce classifiers close to the optimal Bayes classifier
- The kNN has only one (hyper-)parameter $K$ :
  - The choice of $K$ has drastic effect
  - Choosing the correct $K$ is a critical task

## Outline

## Outline

## Two types of errors (1)

- The accuracy of a classifier depends on two quantities :
  - The reducible error
  - The irreducible error
- With $Y = f(X) + \varepsilon$ and $\widehat{Y} = \widehat{f}(X)$ :

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

## Two types of errors (2)

- The reducible error
  - $\widehat{f}$ is not a perfect estimate for $f$
  - The error is reducible by using the appropriate classifier
  - The error cannot be reduced to zero !
- The irreducible error : $\varepsilon$
  - $\varepsilon$ may contain unmeasured variables that are useful to predict $Y$
  - $\varepsilon$ may contain unmeasurable variation : (e.g. individual emotional state)
    - See TP1

- A good classifier has managed to reduce the reducible error
- The best classifier have the lowest possible reducible error

## Outline

## Accuracy

- The accuracy can be measured by the predictive error :

$$\mathbb{E}[\mathbb{I}(\widehat{(Y)} = Y)]$$

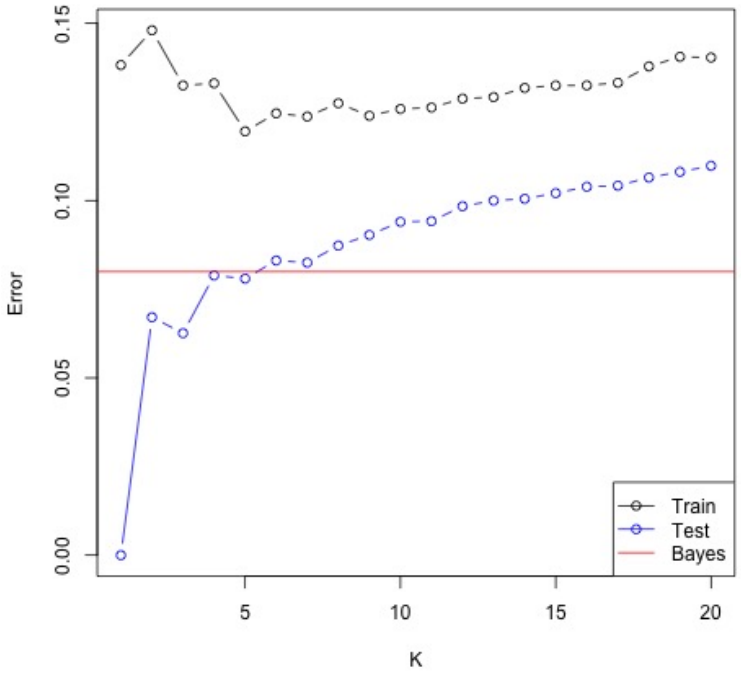
  - The predictive error is estimated using a set of $n_0$ observations $\{(y_1, x_1), \ldots, (y_{n_0}, x_{n_0})\}$

$$\frac{1}{n_0} \sum_{i=1}^{n_0} \mathbb{I}(y_i \neq \widehat{y_i})$$

- Based on the training set $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ the training error is $\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_i \neq \widehat{y_i})$
  - The training error is biased and should be used to summarized model accuracy
- The test error is associated with a test set of observations independent from the training set of observations

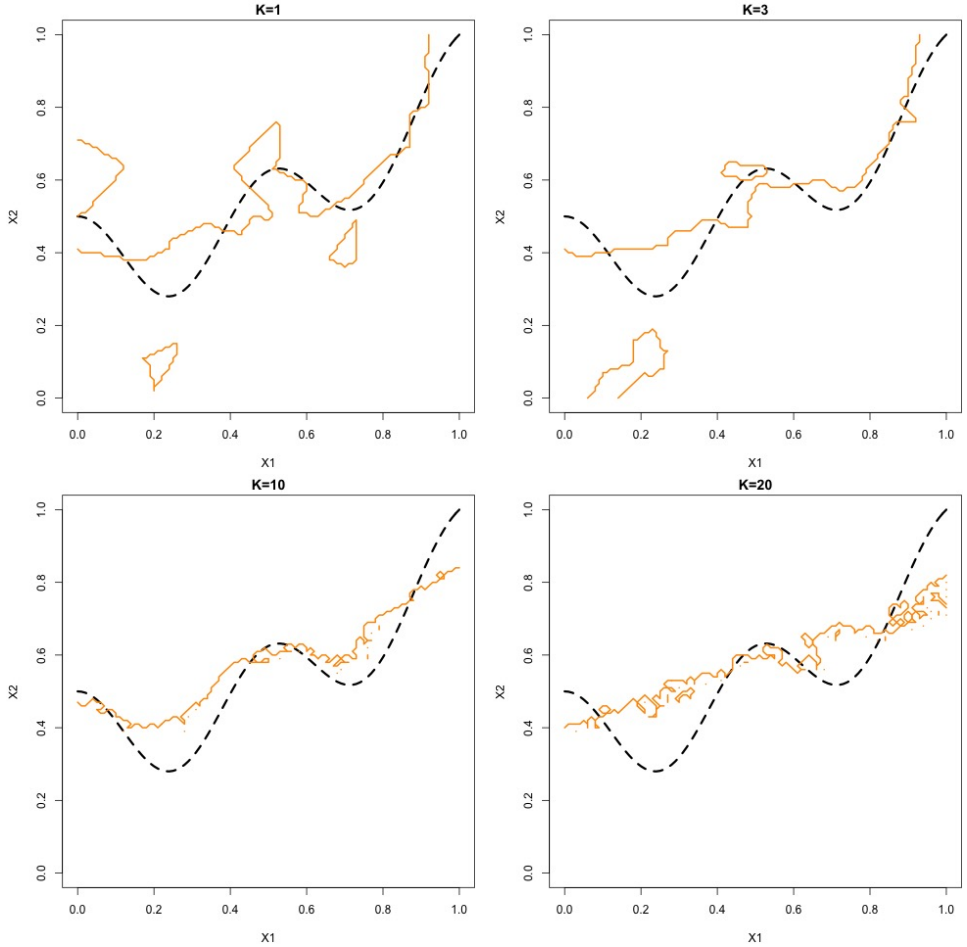
A good classifier is one for which the **test error is smallest**

## UShape




- $K$ is a measure of the flexibility of the method
  - The lower $K$ is, the more flexible the method is.

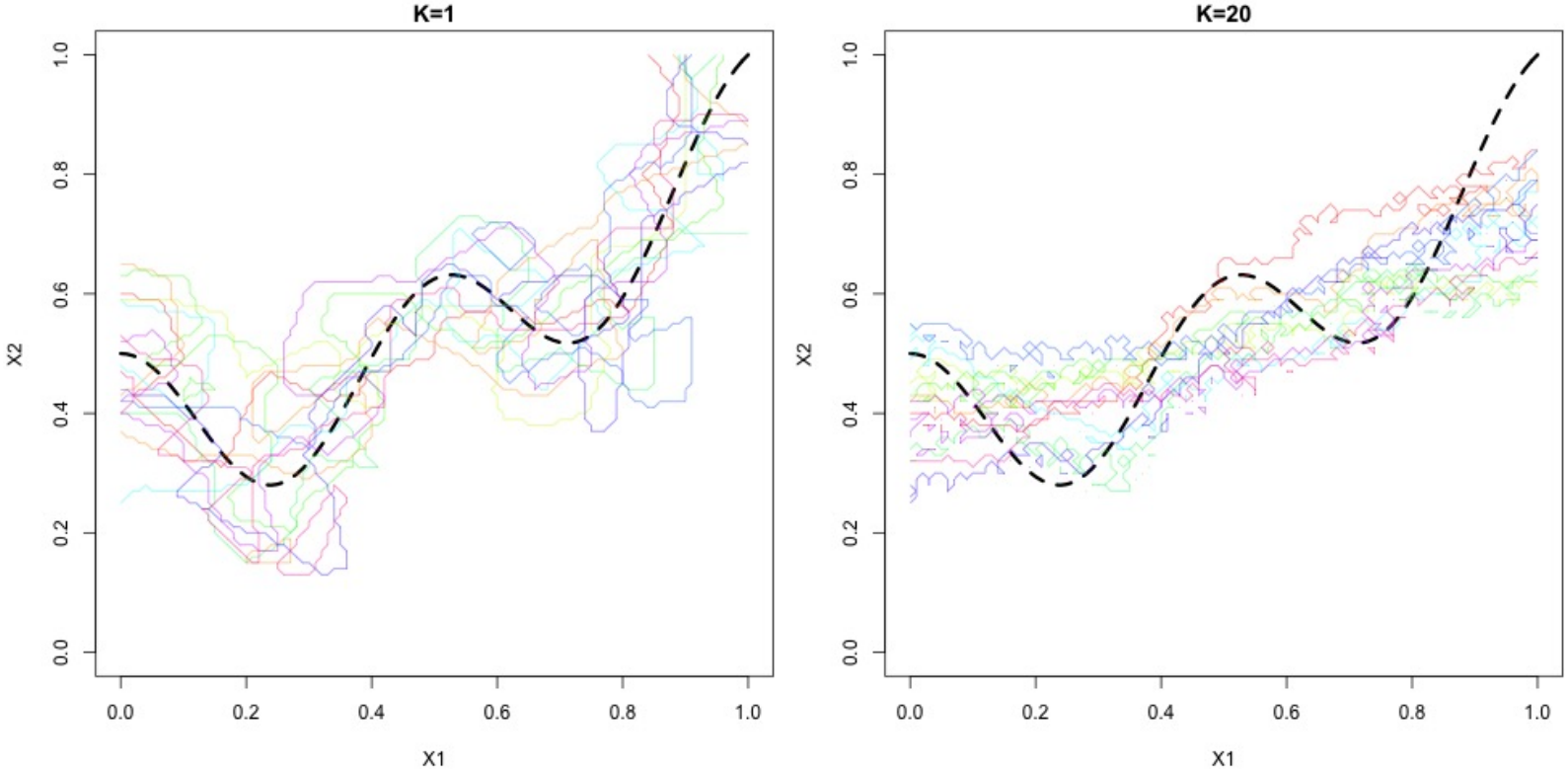
## Contour - Biais




- Flexibility can be seen as the capacity of the method to draw smooth boundaries.
- Flexibility is related to the bias of the method

## Contour - Variance

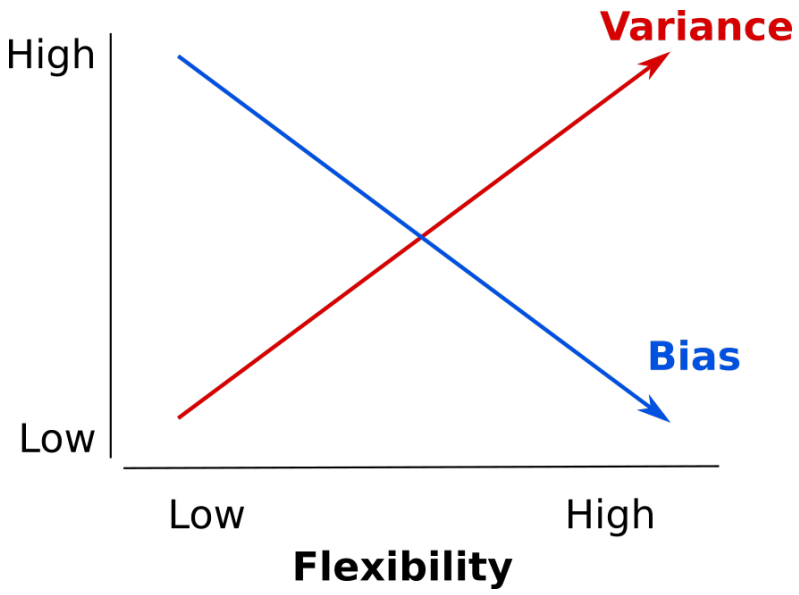
- Estimation of the boundaries with 10 different simulated datasets




- Flexibility is related to the variance of the estimated boundaries.

## The Biais-Variance trade-off

- Machine Learning aims at proposing methods with
  - low bias
  - low variance




- How to choose the best model in a class of models ?
- How to estimate the right level of flexibility ?

## The Biais-Variance trade-off

- High flexibility ⇔ Low bias and high variance
  - Very good performance on the training dataset
  - Risk : Overfitting !
- Low flexibility ⇔ High bias and low variance
  - Reliability of the prediction
  - Risk : the shape of the class of models might not be adapted to the classification problem (linear vs non-linear for example).

## Outline

## Back to the kNN

- The only parameter $K$ can be tuned by minimizing the test error
- kNN is hardly interpretable : kNN doesn't know which attributes are more important.
  - When computing distance between data points (usually Euclidean distance or other generalisations of it), each attribute normally weighs the same to the total distance.
  - This means that attributes which are not so important will have the same influence on the distance compared to more important attributes.

## General scheme

For a given dataset :
For a class of model,

- Choose a measure of accuracy (A)
- Choose a sampling scheme (SS)
- With (A) and (SS), choose the best model