# Introduction

Jérémie Cabessa

Laboratoire DAVID, UVSQ

# Intelligence Artificielle (IA)

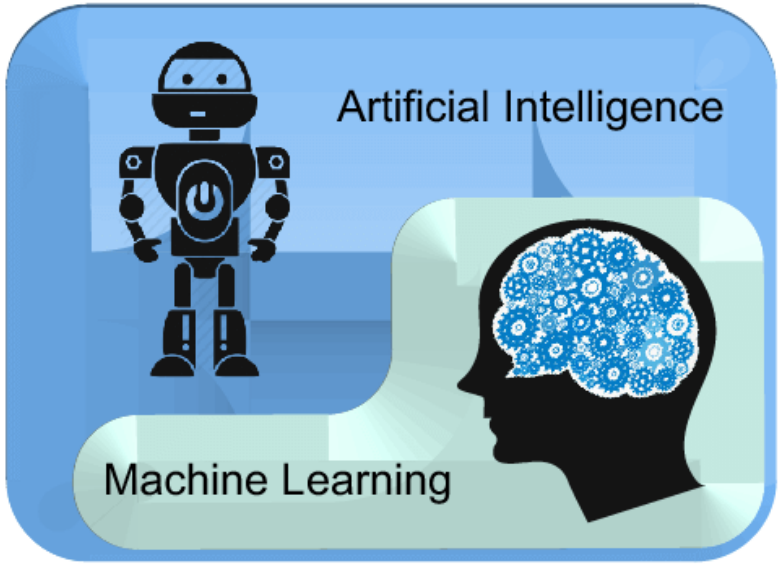
- L'**intelligence artificielle (IA)** désigne l'ensemble des théories et techniques visant à simuler certains aspects de l'intelligence humaine par des systèmes informatiques.
- Ces aspects de l'intelligence comprennent entre autres: le raisonnement, l'apprentissage, la perception, etc.

# Intelligence Artificielle (IA)

- L'**intelligence artificielle (IA)** désigne l'ensemble des théories et techniques visant à simuler certains aspects de l'intelligence humaine par des systèmes informatiques.
- Ces aspects de l'intelligence comprennent entre autres: le raisonnement, l'apprentissage, la perception, etc.

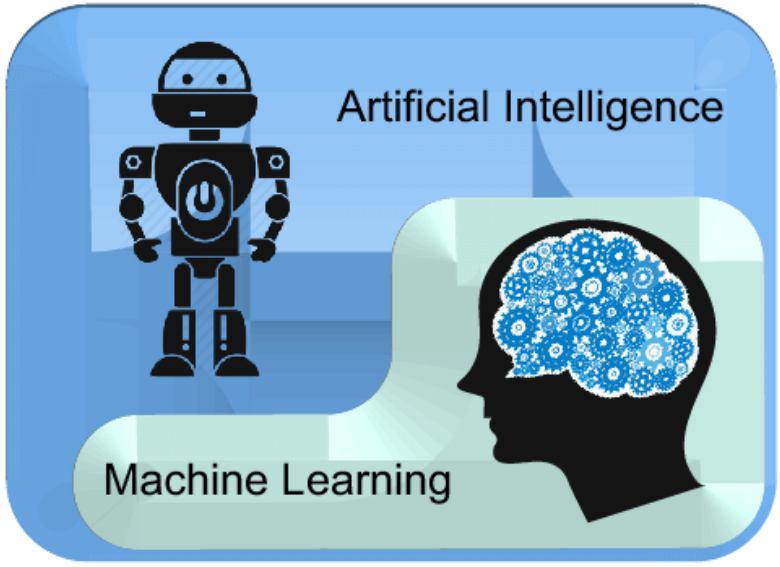
# Machine Learning (ML)

- Le **machine learning (ML)** désigne l'ensemble des théories et techniques algorithmiques qui permettent d'apprendre à des systèmes informatiques à résoudre des problèmes, exécuter des tâches, etc., en se basant sur des données (data).
- Le machine learning (ML) est un sous-domaine de l'intelligence artificielle (IA).



# Machine Learning (ML)

- Le **machine learning (ML)** désigne l'ensemble des théories et techniques algorithmiques qui permettent d'apprendre à des systèmes informatiques à résoudre des problèmes, exécuter des tâches, etc., en se basant sur des données (data).
- Le machine learning (ML) est un sous-domaine de l'intelligence artificielle (IA).



# Machine Learning (ML)

- Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
  - **Approche "hard-coded":** se base sur notre compréhension du concept de "fleur", programmé en dur.

    On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
  - **Approche "machine learning":** se base sur des exemples, des images de "fleurs" et de "non-fleurs", utilisés comme exemples pour

    On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

# Machine Learning (ML)

- Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
  - **Approche "hard-coded":** se base sur notre compréhension du concept de "fleur", programmé en dur.

    On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
  - **Approche "machine learning":** se base sur des data, i.e., des images de "fleurs" et de "non-fleurs" utilisées comme exemples.

    On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

# Machine Learning (ML)

- Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
  - **Approche "hard-coded":** se base sur notre compréhension du concept de "fleur", programmé en dur.

    On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
  - **Approche "machine learning":** se base sur des data, i.e., des images de "fleurs" et de "non-fleurs" utilisées comme exemples.

    On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

# Machine Learning (ML)

- Le machine learning constitue un **changement de paradigme** dans la résolution de problèmes.
- **Exemple:** on aimerait créer un algorithme de reconnaissance d'images qui détecte les images contenant des fleurs.
  - **Approche "hard-coded":** se base sur notre compréhension du concept de "fleur", programmé en dur.

    On code les formes, couleurs, etc., que peuvent avoir les différentes fleurs, et on crée un algorithme qui détecte les fleurs à partir de ces caractéristiques.
  - **Approche "machine learning":** se base sur des data, i.e., des images de "fleurs" et de "non-fleurs" utilisées comme exemples.

    On crée un algorithme qui apprend par lui-même à détecter les fleurs à partir des data.

# Apprentissage supervisé et non-supervisé

- Deux types de méthodes d'apprentissage:
  - **apprentissage supervisé (supervised learning)**
  - **apprentissage non-supervisé (unsupervised learning)**
- **Apprentissage supervisé**

  Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).
- **Apprentissage non-supervisé**

  Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

# Apprentissage supervisé et non-supervisé

- Deux types de méthodes d'apprentissage:
  - **apprentissage supervisé (supervised learning)**
  - **apprentissage non-supervisé (unsupervised learning)**
- **Apprentissage supervisé**

  Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).
- **Apprentissage non-supervisé**

  Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

# Apprentissage supervisé et non-supervisé

- Deux types de méthodes d'apprentissage:
  - **apprentissage supervisé (supervised learning)**
  - **apprentissage non-supervisé (unsupervised learning)**
- **Apprentissage supervisé**

  Les données contiennent des variables d'input et des variables d'output. Il s'agit de modéliser la relation entre les variables d'input et la variable d'output (prediction, classification).
- **Apprentissage non-supervisé**

  Les données ne contiennent que des variables d'input, et pas de variable d'output. Il s'agit d'extraire des classes ou groupes de données présentant des caractéristiques communes (clustering).

# Apprentissage supervisé

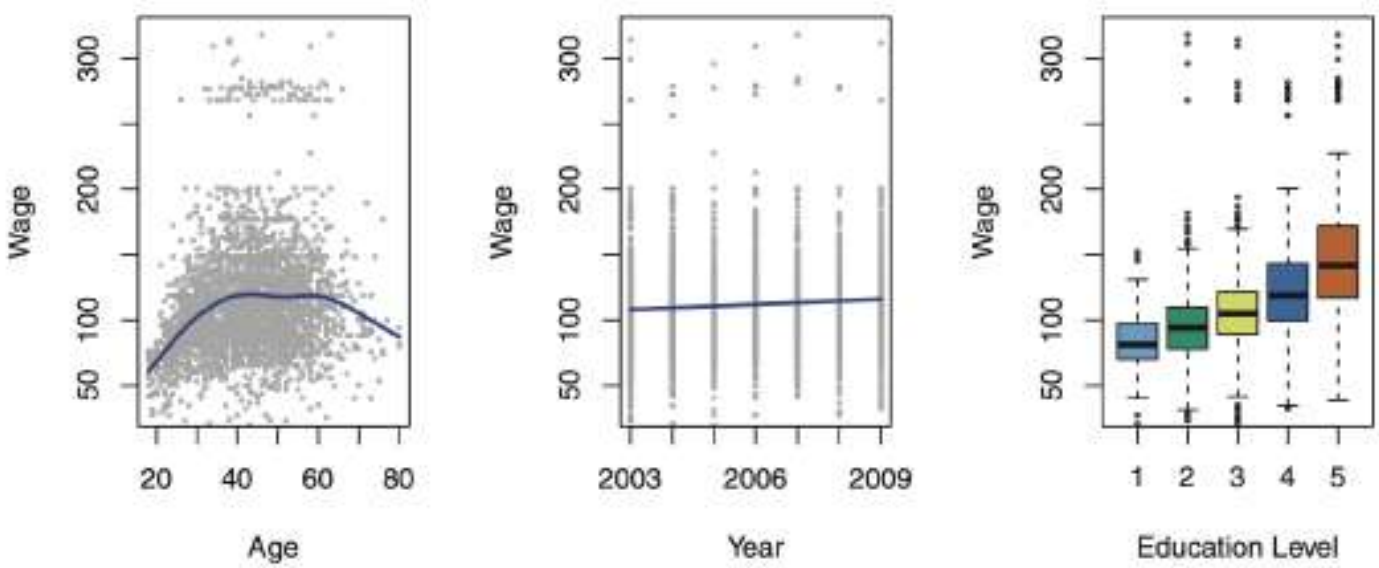



**FIGURE 1.1.** *Wage data, which contains income survey information for males from the central Atlantic region of the United States.* Left: *wage as a function of age. On average, wage increases with age until about 60 years of age, at which point it begins to decline.* Center: *wage as a function of year. There is a slow but steady increase of approximately* \$10,000 *in the average wage between* 2003 *and* 2009. Right: *Boxplots displaying wage as a function of education, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average, wage increases with the level of education.*

Figure taken from [James et al., 2013].

# Apprentissage supervisé

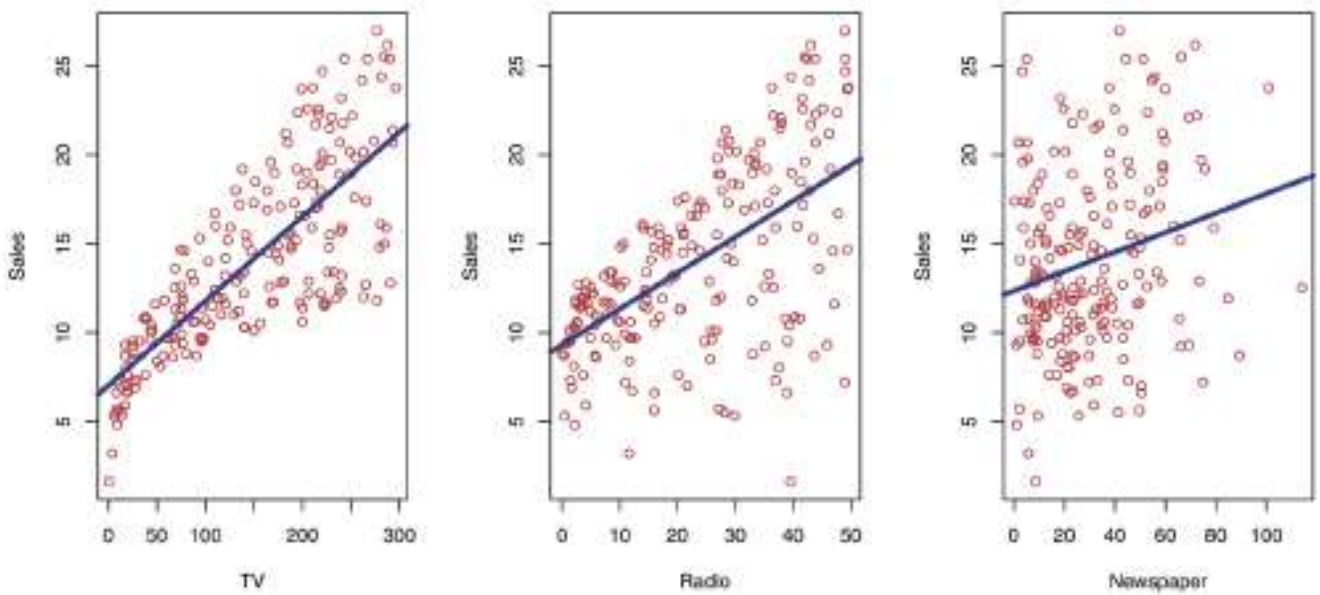



**FIGURE 2.1.** *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

Figure taken from [James et al., 2013].

# Apprentissage supervisé

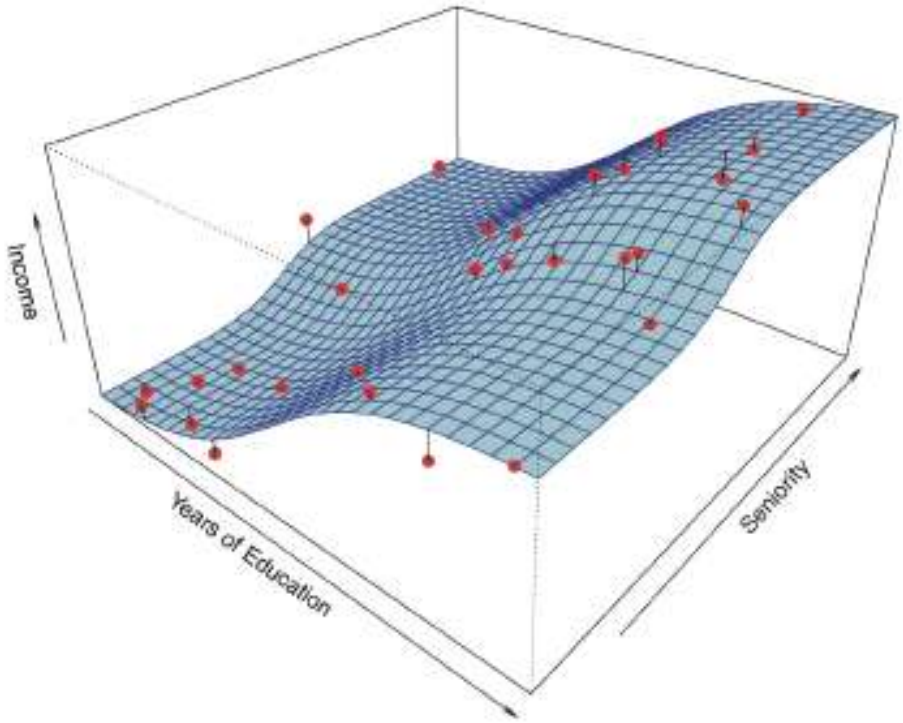



**FIGURE 2.3.** *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

Figure taken from [James et al., 2013].

# Apprentissage non-supervisé

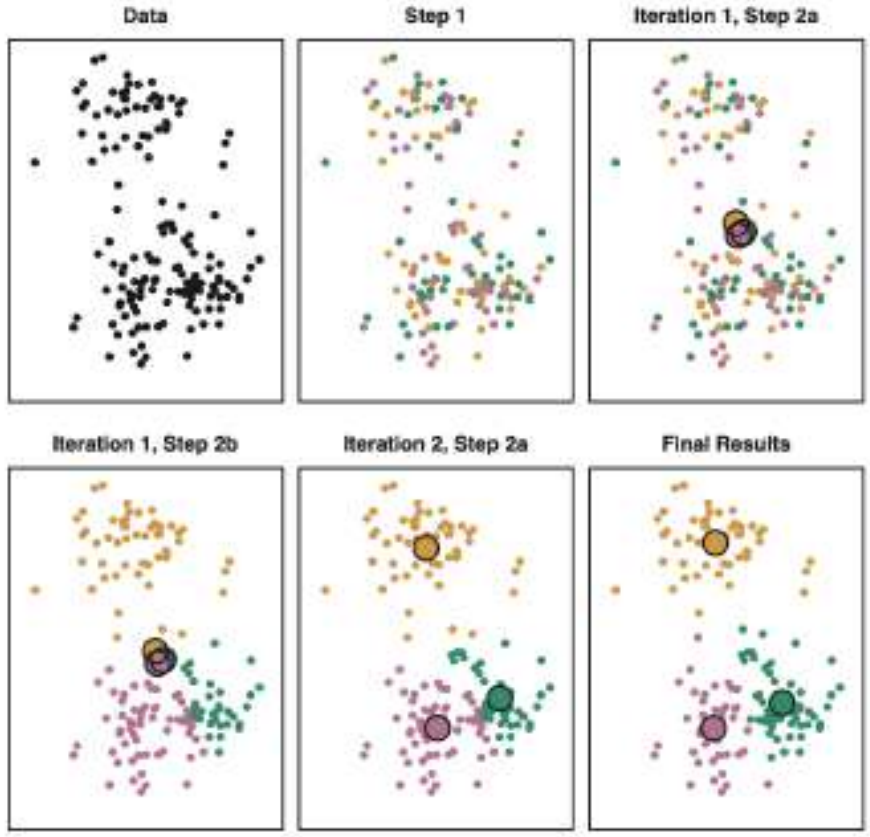



**FIGURE 10.6.** *The progress of the K-means algorithm on the example of Figure 10.5 with K=3.* Top left: *the observations are shown.* Top center: *in Step 1 of the algorithm, each observation is randomly assigned to a cluster.* Top right: *in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.* Bottom left: *in Step 2(b), each observation is assigned to the nearest centroid.* Bottom center: *Step 2(a) is once again performed, leading to new cluster centroids.* Bottom right: *the results obtained after ten iterations.*

Figure taken from [James et al., 2013].

# Apprentissage non-supervisé

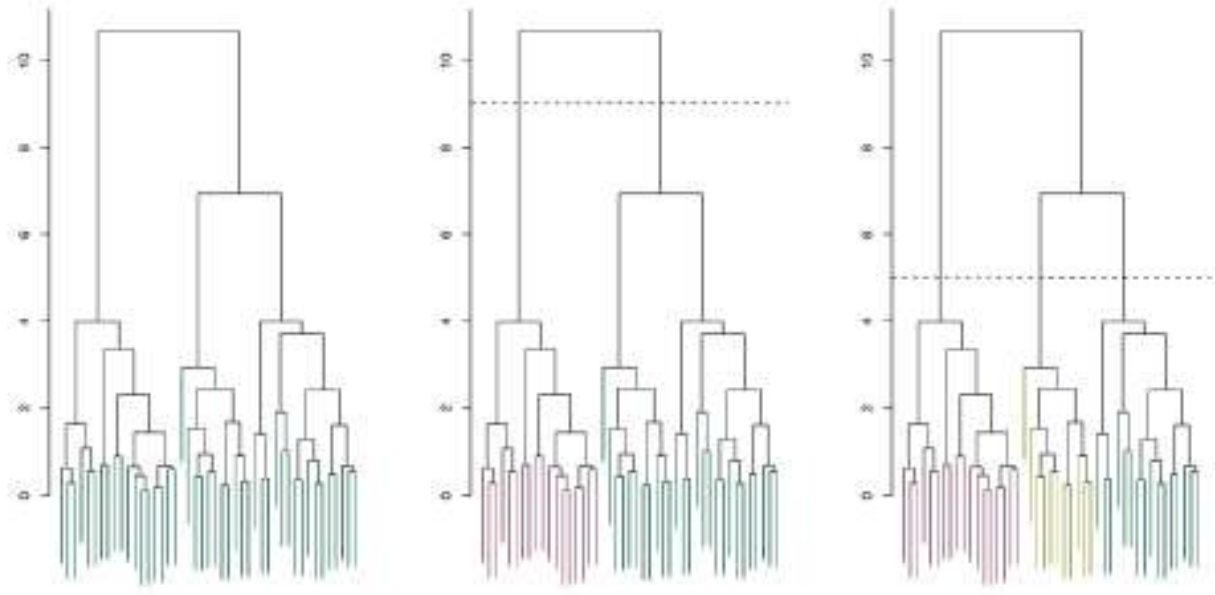


**FIGURE 10.9.** Left: *dendrogram obtained from hierarchically clustering the data from Figure 10.8 with complete linkage and Euclidean distance.* Center: *the dendrogram from the left-hand panel, cut at a height of nine (indicated by the dashed line). This cut results in two distinct clusters, shown in different colors.* Right: *the dendrogram from the left-hand panel, now cut at a height of five. This cut results in three distinct clusters, shown in different colors. Note that the colors were not used in clustering, but are simply used for display purposes in this figure.*

Figure taken from [James et al., 2013].

# Régression et Classification

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**

  La variable d'output est **quantitative**.
- **Méthodes de classification**

  La variable d'output est **qualitative**.

# Régression et Classification

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**

  La variable d'output est **quantitative**.
- **Méthodes de classification**

  La variable d'output est **qualitative**.

# Régression et Classification

- Dans le cadre de l'**apprentissage supervisé**, on distingue deux types de méthodes:
- **Méthodes de régression**

  La variable d'output est **quantitative**.
- **Méthodes de classification**

  La variable d'output est **qualitative**.

# Régression

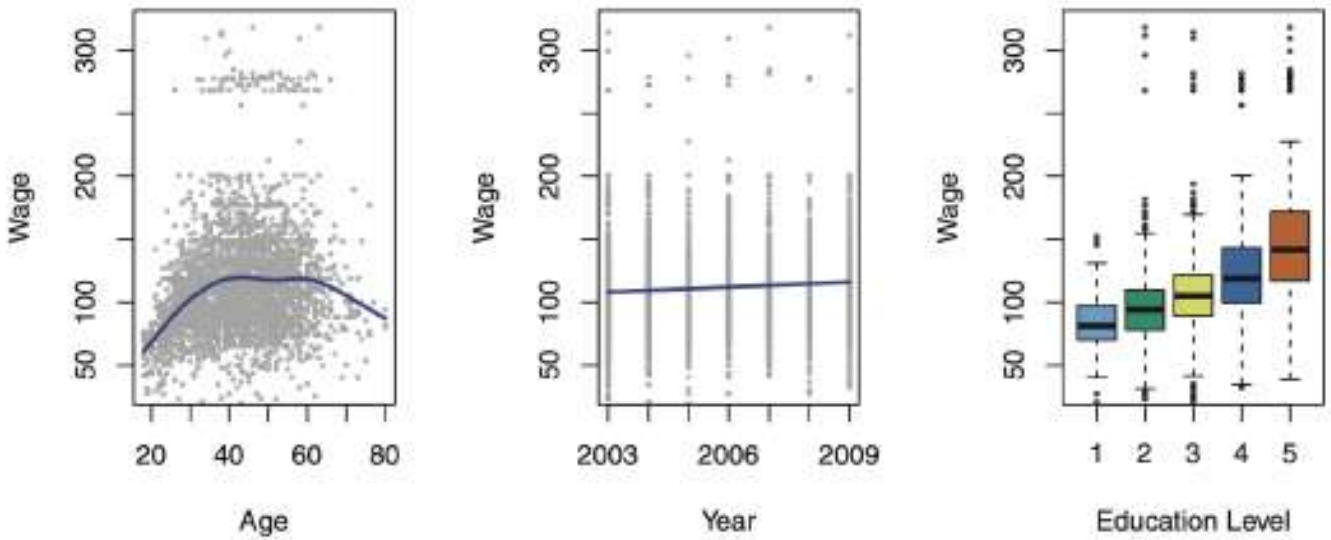



**FIGURE 1.1.** *Wage data, which contains income survey information for males from the central Atlantic region of the United States.* Left: *wage as a function of age. On average, wage increases with age until about* 60 *years of age, at which point it begins to decline.* Center: *wage as a function of year. There is a slow but steady increase of approximately* $10,000 *in the average wage between* 2003 *and* 2009. Right: *Boxplots displaying wage as a function of education, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average, wage increases with the level of education.*

Figure taken from [James et al., 2013].

# Régression

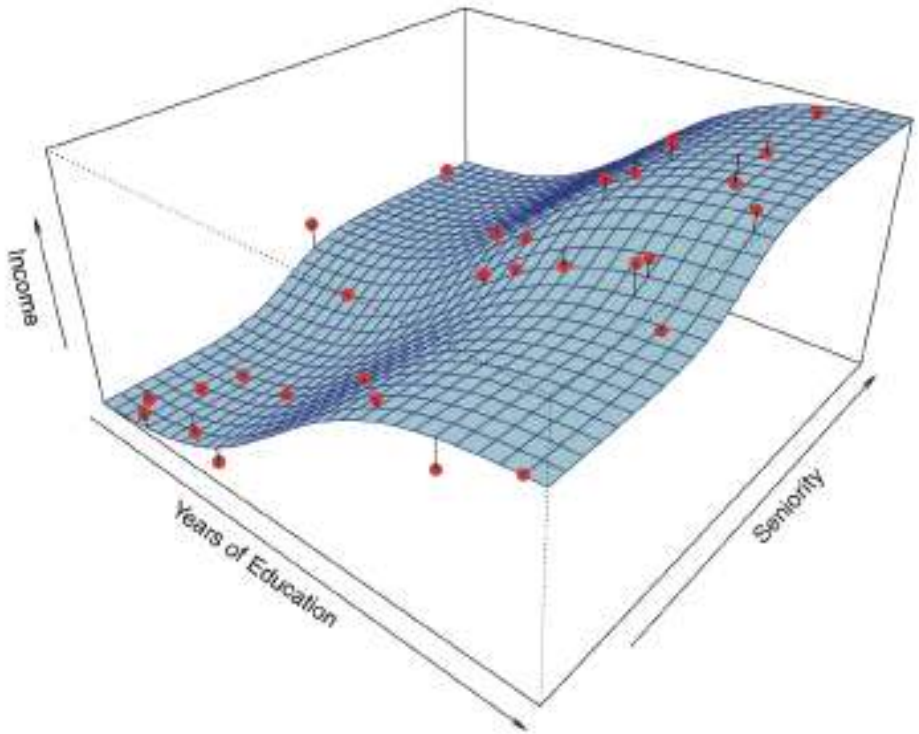



**FIGURE 2.3.** *The plot displays* income *as a function of* years of education *and* seniority *in the* Income *data set. The blue surface represents the true underlying relationship between* income *and* years of education *and* seniority*, which is known since the data are simulated. The red dots indicate the observed values of these quantities for* 30 *individuals.*

Figure taken from [James et al., 2013].

# Classification

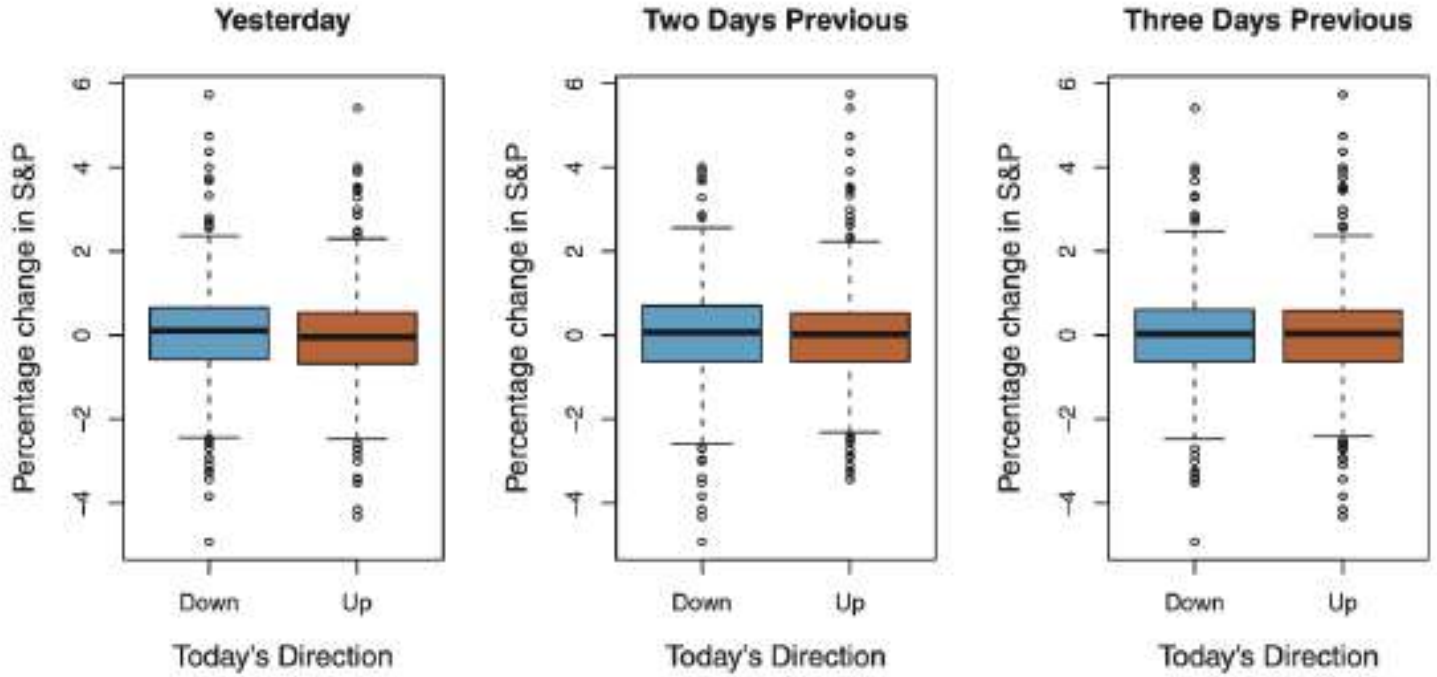



**FIGURE 1.2.** Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Figure taken from [James et al., 2013].

# Classification

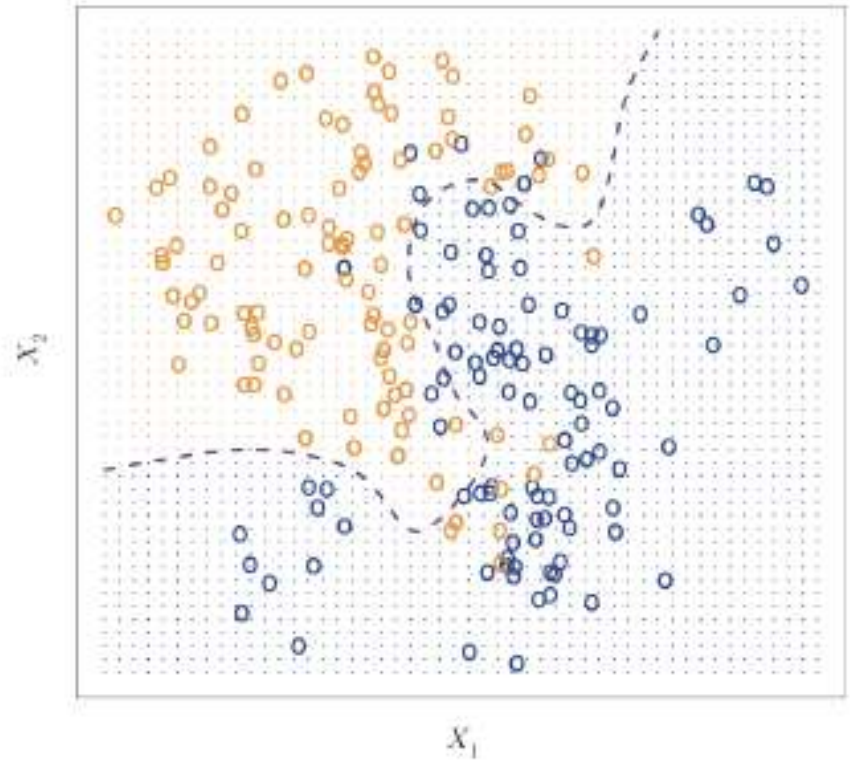



**FIGURE 2.13.** *A simulated data set consisting of* 100 *observations in each of two groups, indicated in blue and in orange. The purple dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.*

Figure taken from [James et al., 2013].

# Apprentissage supervisé: formulation du problème

- Soient $X_1, \ldots, X_p$ et $Y$ des variables aléatoires.
- $X_1, \ldots, X_p$ sont appelées variables d'inputs, variables indépendantes, variables explicatives, prédicteurs, (features).
- $Y$ est appelée variable d'output, variable dépendante, réponse, (response, target).

# Apprentissage supervisé: formulation du problème

- Soient $X_1, \ldots, X_p$ et $Y$ des variables aléatoires.
- $X_1, \ldots, X_p$ sont appelées **variables d'inputs**, **variables indépendantes**, **variables explicatives**, **prédicteurs**, **(features)**.
- $Y$ est appelée variable d'output, variable dépendante, réponse, (response, target).

# Apprentissage supervisé: formulation du problème

- Soient $X_1, \ldots, X_p$ et $Y$ des variables aléatoires.
- $X_1, \ldots, X_p$ sont appelées **variables d'inputs**, **variables indépendantes**, **variables explicatives**, **prédicteurs**, **(features)**.
- $Y$ est appelée **variable d'output**, **variable dépendante**, **réponse**, **(response, target)**.

# Apprentissage supervisé: formulation du problème

- On suppose qu'il existe une (vraie) **relation** $f$ entre $X_1, \ldots, X_p$ et $Y$ de la forme

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

où $f$ est une fonction inconnue et $\epsilon$ est une variable aléatoire indépendante de $X_1, \ldots, X_p$ et de moyenne $0$, le bruit.

- On aimerait apprendre une (bonne) **estimation** $\hat{f}$ de $f$. On aura alors

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p)$$

où $\hat{f}$ est l'**estimation** de $f$ et $\hat{Y}$ est la **prediction** de $Y$.

# Apprentissage supervisé: formulation du problème

- On suppose qu'il existe une (vraie) **relation** $f$ entre $X_1, \ldots, X_p$ et $Y$ de la forme

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

où $f$ est une fonction inconnue et $\epsilon$ est une variable aléatoire indépendante de $X_1, \ldots, X_p$ et de moyenne $0$, le bruit.

- On aimerait apprendre une (bonne) **estimation** $\hat{f}$ de $f$. On aura alors

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p)$$

où $\hat{f}$ est l'**estimation** de $f$ et $\hat{Y}$ est la **prediction** de $Y$.

# Apprentissage supervisé: formulation du problème

- Pour apprendre l'estimation $\hat{f}$ de $f$, on dispose de données (data)

$$S_{\text{train}} = \left\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\right\}$$

où $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{i_p})$ pour tout $i = 1, \ldots, n$.

- Ces données constituent le "**training set**" (training data).

# Apprentissage supervisé: formulation du problème

- Pour apprendre l'estimation $\hat{f}$ de $f$, on dispose de données (data)

$$S_{\text{train}} = \left\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\right\}$$

où $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{i_p})$ pour tout $i = 1, \ldots, n$.

- Ces données constituent le **"training set"** (training data).

# Apprentissage supervisé: formulation du problème

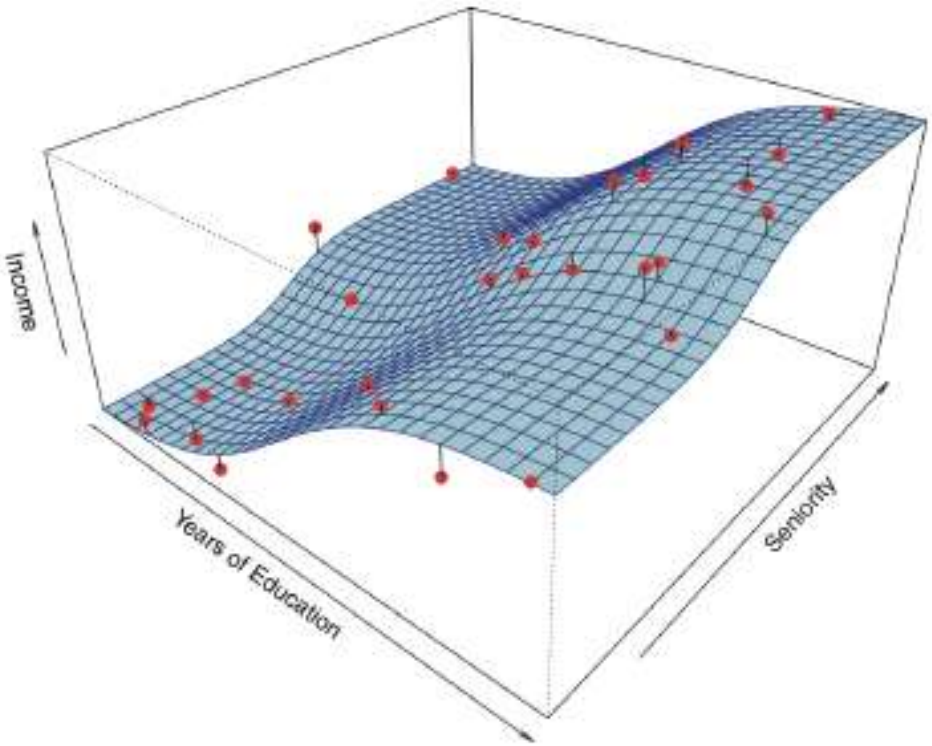



Figure taken from [James et al., 2013]

# Erreur réductible et irréductible

- On a donc

$$\begin{aligned} Y &= f(X_1, \ldots, X_p) + \epsilon \quad \text{vraie relation} \\ \hat{Y} &= \hat{f}(X_1, \ldots, X_p) \quad \text{estimation} \end{aligned}$$

- On peut facilement montrer que

$$\begin{aligned} \mathrm{E}[Y - \hat{Y}]^2 &= \mathrm{E}[f(X_1, \ldots, X_p) + \epsilon - \hat{f}(X_1, \ldots, X_p)]^2 \\ &= \mathrm{E}[f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})]^2 \\ &= \underbrace{\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2}_{\text{erreur réductible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{erreur irréductible}} \end{aligned}$$

où $\boldsymbol{X} = (X_1, \ldots, X_p)$.

# Erreur réductible et irréductible

- On a donc

$$\begin{aligned} Y &= f(X_1, \ldots, X_p) + \epsilon \quad \text{vraie relation} \\ \hat{Y} &= \hat{f}(X_1, \ldots, X_p) \quad \text{estimation} \end{aligned}$$

- On peut facilement montrer que

$$\begin{aligned} \mathrm{E}[Y - \hat{Y}]^2 &= \mathrm{E}[f(X_1, \ldots, X_p) + \epsilon - \hat{f}(X_1, \ldots, X_p)]^2 \\ &= \mathrm{E}[f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})]^2 \\ &= \underbrace{\left(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right)^2}_{\text{erreur réductible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{erreur irréductible}} \end{aligned}$$

où $\boldsymbol{X} = (X_1, \ldots, X_p)$.

# Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de $f$, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" réel intrinsèque au modèle.

# Erreur réductible et irréductible

- **Erreur réductible (reducible error):** peut être réduite; plus $\hat{f}$ est une bonne estimation de $f$, plus cette erreur sera faible.
- **Erreur irréductible (irreducible error):** ne peut être réduite; par le bais de notre estimation $\hat{f}$, nous n'avons aucune prise sur le "bruit" réel intrinsèque au modèle.

# Erreur réductible et irréductible

- Exemple de deux estimations $\hat{f}$. La deuxième estimation est meilleure car elle est associée à une erreur réductible plus faible.

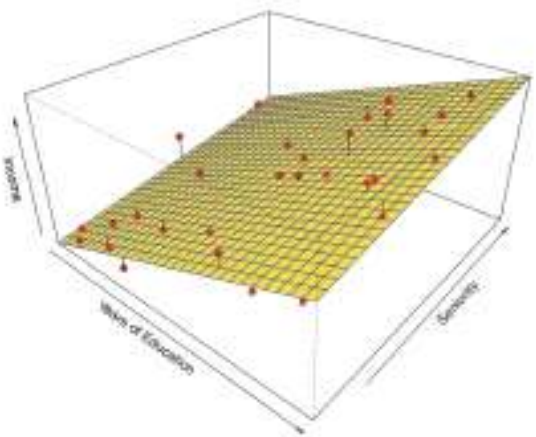



FIGURE 2.4. *A linear model fit by least squares to the* Income *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

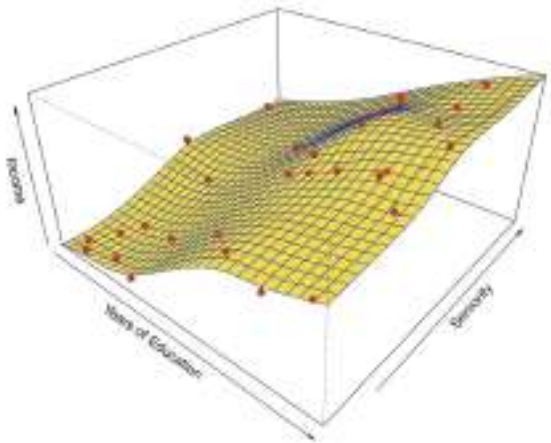



FIGURE 2.5. *A smooth thin-plate spline fit to the* Income *data from Figure 2.3 is shown in yellow; the observations are displayed in red. Splines are discussed in Chapter 7.*

Figure taken from [James et al., 2013]

# Ensembles de train et de test (train/test sets)

- Le machine learning se base sur des données (data). L'ensemble de toutes les données dont on dispose au départ d'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent **ensemble de train** ou **train set**.
- Pour évaluer le modèle, on utilise un ensemble de données distinct du train set, des données que le modèle n'a jamais vues auparavant, qu'on appelle **ensemble de test** ou **test set**.

# Ensembles de train et de test (train/test sets)

- Le machine learning se base sur des données (data). L'ensemble de toutes les données dont on dispose au départ d'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent **ensemble de train** ou **train set**.
- Pour évaluer le modèle, on utilise un ensemble de données distinct du train set, des données que le modèle n'a jamais vues auparavant, qu'on appelle **ensemble de test** ou **test set**.

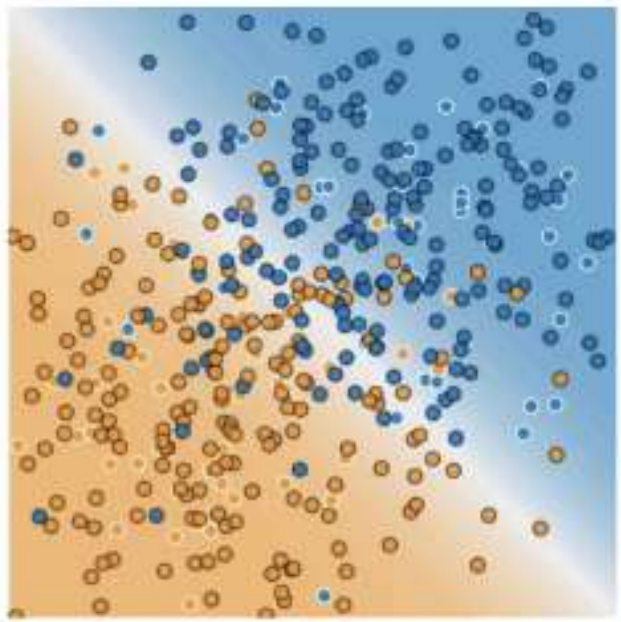
# Ensembles de train et de test (train/test sets)

- Le machine learning se base sur des données (data). L'ensemble de toutes les données dont on dispose au départ d'appelle un **dataset**
- Les données utilisées pour entraîner le modèle s'appellent **ensemble de train** ou **train set**.
- Pour évaluer le modèle, on utilise un ensemble de données distinct du train set, des données que le modèle n'a jamais vues auparavant, qu'on appelle **ensemble de test** ou **test set**.

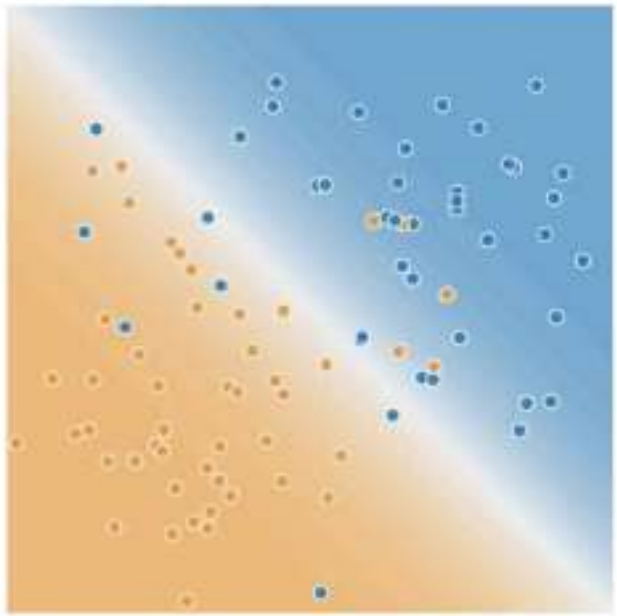
# Ensembles de train et de test (train/test sets)

- En général, on sépare de **dataset** de départ en prenant 80% des données pour le **train set** et 20% pour le **test set** (train-test split).

# Ensembles de train et de test (train/test sets)



Training Data



Test Data

# Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2$$

# Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2$$

# Fonction de coût (cost function)

- Comment mesurer la qualité d'un modèle $\hat{f}$?
- On utilise une **fonction de coût (cost or loss function)**.
- La plus célèbre est **l'erreur des moindre carrés (mean squared error) MSE**. Étant donné un training set

$$S_{\text{train}} = \left\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_n}, y_n)\right\}$$

on définit

$$\text{MSE}_{\text{train}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\boldsymbol{x_i})\right)^2$$

# Overfitting

- **Problème:** Le modèle $\hat{f}$ a été construit sur la base du training set $S_{\text{train}}$. Ainsi, il peut être très performant lorsqu'il est évalué sur le training set, mais bien moins performant lorsqu'il est évalué sur des données qu'il n'a jamais vues, le **test set**.
- Lorsque le modèle est performant sur le **training set** ($\text{MSE}_{\text{train}}$ basse), mais qu'il est bien moins performant sur sur le **test set** ($\text{MSE}_{\text{test}}$ plus élevée), on est dans une situation d'**overfitting**.

# Overfitting

- **Problème:** Le modèle $\hat{f}$ a été construit sur la base du training set $S_{\text{train}}$. Ainsi, il peut être très performant lorsqu'il est évalué sur le training set, mais bien moins performant lorsqu'il est évalué sur des données qu'il n'a jamais vues, le **test set**.
- Lorsque le modèle est performant sur le **training set** ($\text{MSE}_{\text{train}}$ basse), mais qu'il est bien moins performant sur sur le **test set** ($\text{MSE}_{\text{test}}$ plus élevée), on est dans une situation d'**overfitting**.

# Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{test}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), de sorte que ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

# Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{test}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), de sorte que ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

# Overfitting

- **Overfitting:** Le modèle est très performant sur le **training set** (i.e., $\mathrm{MSE}_{\mathrm{train}}$ basse) alors qu'il est bien moins performant sur sur le **test set** (i.e., $\mathrm{MSE}_{\mathrm{test}}$ élevée).
- Le modèle a donc "sur-appris" (overfit) les données d'apprentissage (train set), de sorte que ses performances ne se généralisent pas bien sur des données inconnues (test set).
- En fait, le modèle a appris le bruit des données d'apprentissage, au lieu de l'ignorer.

# Overfitting

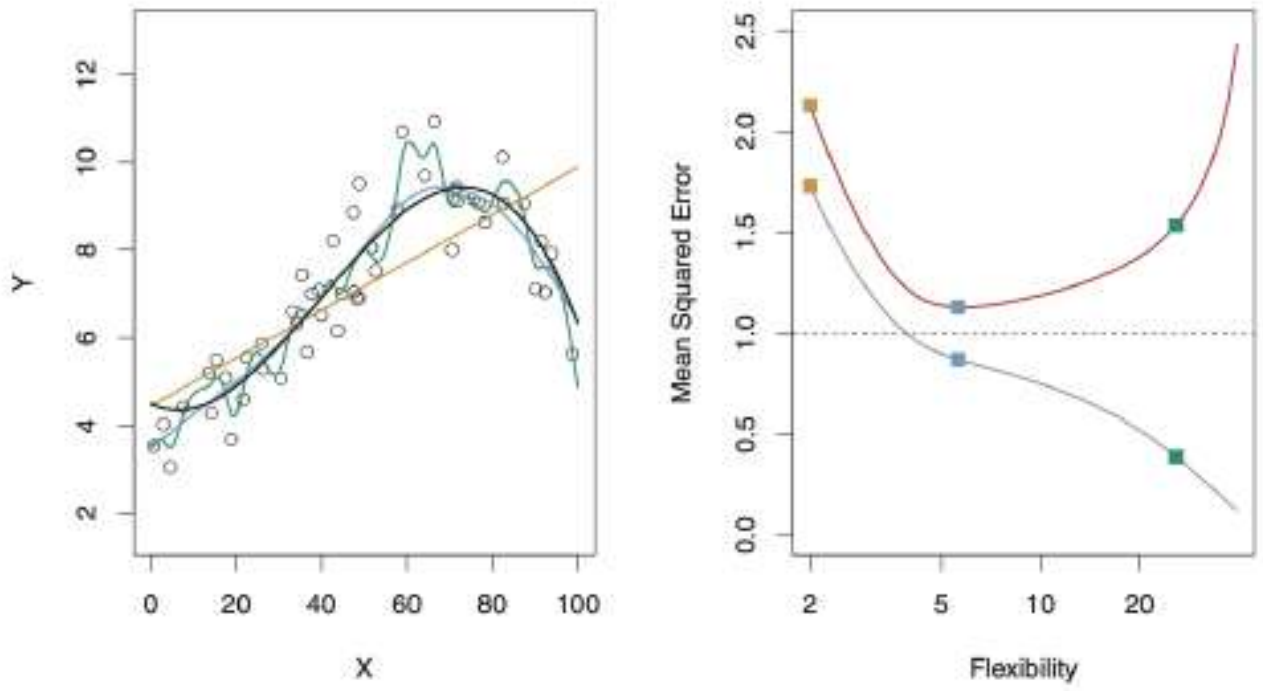



**FIGURE 2.9.** Left: *Data simulated from $f$, shown in black. Three estimates of $f$ are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Figure taken from [James et al., 2013]

# Overfitting

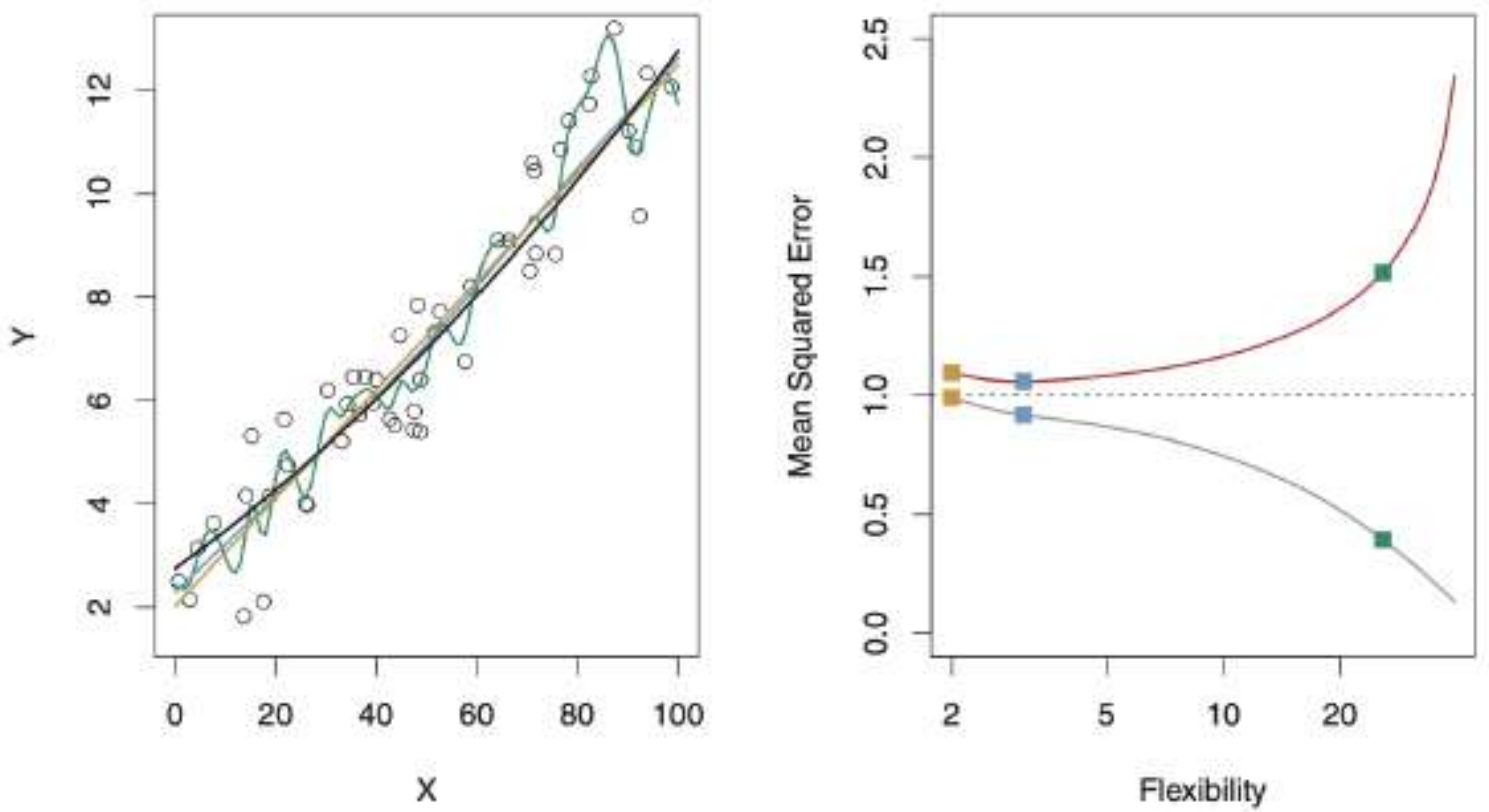



**FIGURE 2.10.** *Details are as in Figure 2.9, using a different true $f$ that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

Figure taken from [James et al., 2013]

# Overfitting

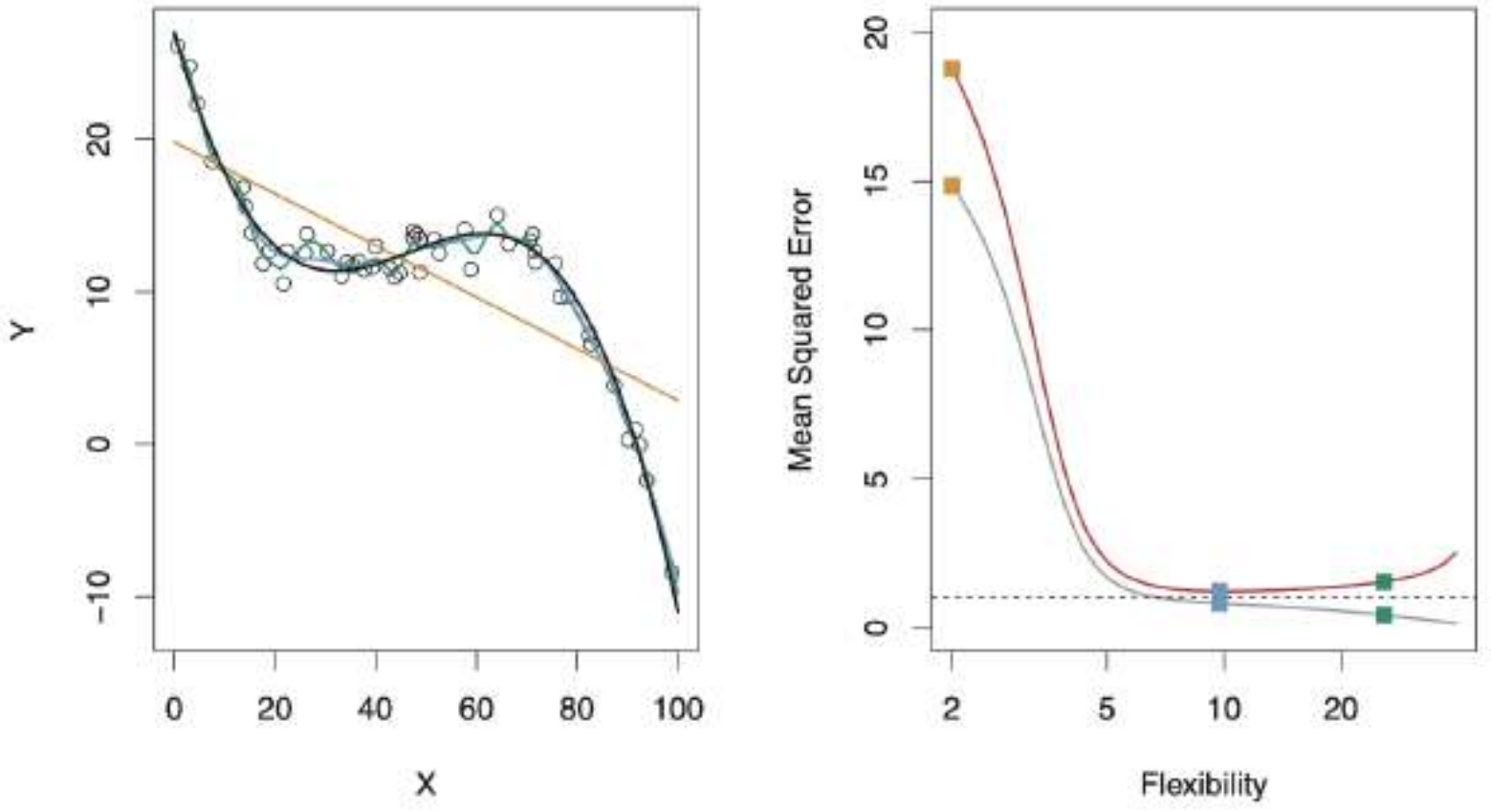



**FIGURE 2.11.** *Details are as in Figure 2.9, using a different $f$ that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

Figure taken from [James et al., 2013]

# Bibliographie

Most images taken from [James et al., 2013].

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
*An Introduction to Statistical Learning: with Applications in R*, volume 103 of *Springer Texts in Statistics*.
Springer, New York.