# Introduction à l'apprentissage automatique

## Application à la recherche et à l'extraction d'information

Anne-Laure Ligozat

2018/2019 [1]

1. librement inspiré des cours de Sylvain Chevallier, Benjamin Piwowarski, Vincent Guigue et Andrew Ng

# Plan

1. Apprentissage
   - Bases
   - Formalisme et modèles
   - Evaluation
   - Quelques modèles
2. Données textuelles
   - Modélisation du texte
   - Tâches de RI et d'EI

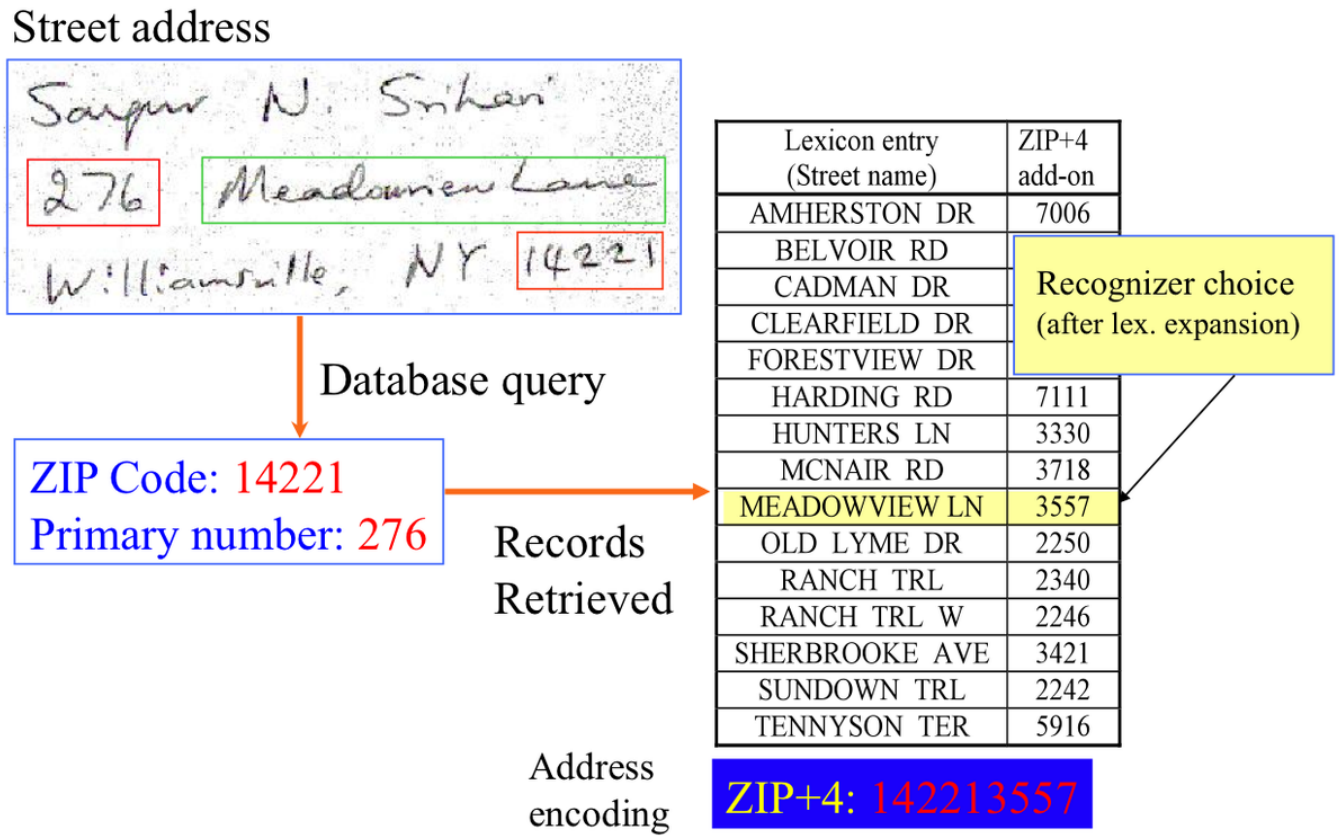
# Apprentissage artificiel (*machine learning* ou *ML*)

- apprentissage = donner la capacité aux ordinateurs d'apprendre sans avoir été programmés explicitement (Arthur Samuel, 1959)
  - apprentissage = changement dans un système qui lui permet d'améliorer sa performance sur la même tâche ou les mêmes données avec l'expérience (Herbert Simon, 1983)
  - nouvelles capacités : systèmes robustes et adaptatifs
  - programme apprend à partir d'exemples ou de son ≪expérience≫

Comprendre comment marche :

- son filtre anti-spam
- la reconnaissance de caractères
- la gestion des crédits bancaires
- la détection de fraudes téléphoniques & bancaires
- la reconnaissance des chiffres sur les chèques, les lettres
- les recommandations sur Amazon, Spotify, etc

# Quelques exemples d'applications




| Lexicon entry (Street name) | ZIP+4 add-on |
|---|---|
| AMHERSTON DR | 7006 |
| BELVOIR RD | |
| CADMAN DR | |
| CLEARFIELD DR | |
| FORESTVIEW DR | |
| HARDING RD | 7111 |
| HUNTERS LN | 3330 |
| MCNAIR RD | 3718 |
| MEADOWVIEW LN | 3557 |
| OLD LYME DR | 2250 |
| RANCH TRL | 2340 |
| RANCH TRL W | 2246 |
| SHERBROOKE AVE | 3421 |
| SUNDOWN TRL | 2242 |
| TENNYSON TER | 5916 |

reconnaissance d'adresse manuscrite

# Quelques exemples d'applications

détection de personnes

# Quelques exemples d'applications


suggestions de recherche

# Quelques exemples d'applications


jeu de go

# Quand utiliser l'apprentissage ?

**expertise** il n'y a pas d'expert humain, ou trop coûteux

**quantité** la quantité de données est telle qu'une analyse humaine est impossible

**adaptation** les modèles doivent être adaptés à l'utilisateur, les données évoluent rapidement dans le temps

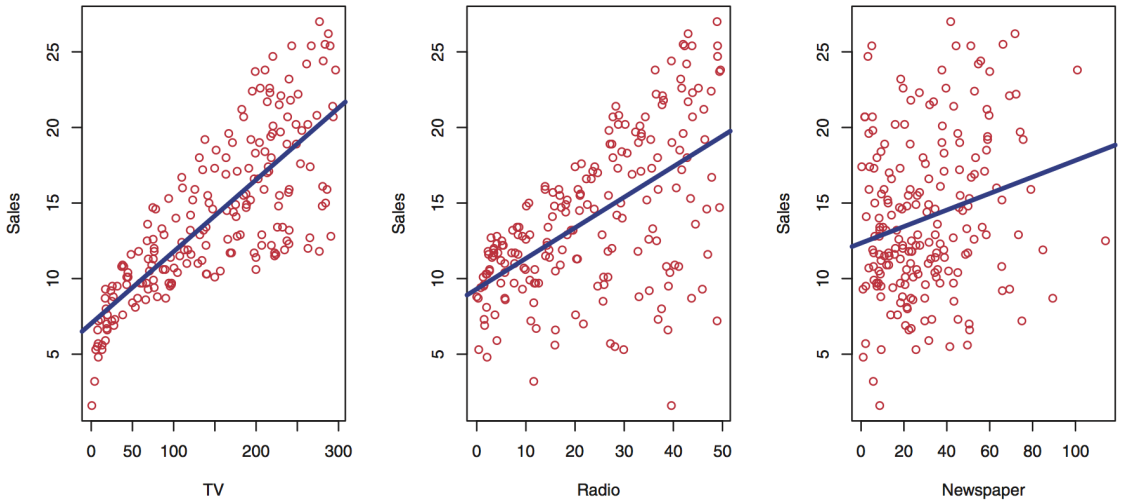
# Exemples
## Médias

entrées variables ou *features*

ici, budgets pub : $X_1 = \text{TV}$, $X_2 = \text{radio}$, $X_3 = \text{journaux}$

sortie $Y = \text{ventes}$

On cherche à prévoir les ventes $Y$ en fonction des entrées
$X = (X_1, X_2, X_3)$, $Y = f(X) + \epsilon$



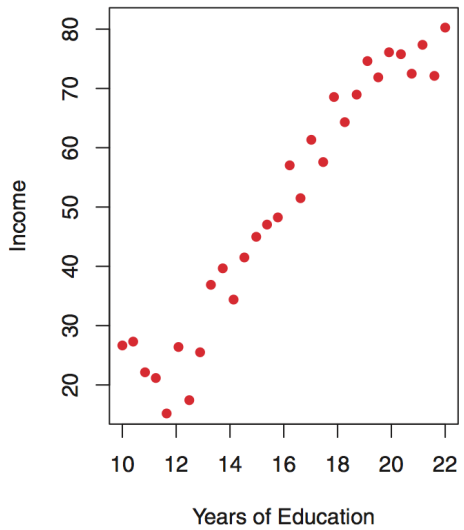

Régression des moindres carrés

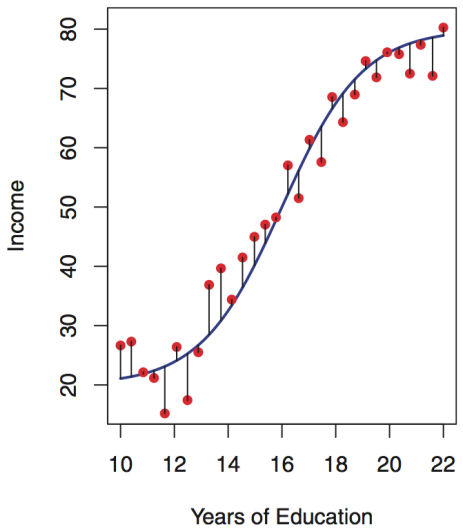
# Exemples
## Salaires

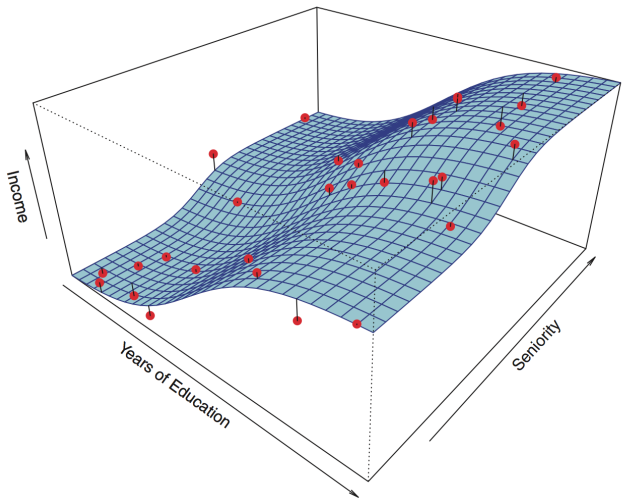
Un exemple simulé :

entrées années d'études et années d'expérience

sortie salaire




Données simulées, génératrices $f(X_1)$ et $f(X_1, X_2)$

# Objectifs de l'apprentissage

## Prédiction

- On connaît les entrées, on connaît certaines sorties
- $\hat{Y} = \hat{f}(X)$, $\hat{f}$ = estimation de $f$, $\hat{Y}$ = prediction pour $Y$
- $\hat{Y}$ dépend d'une erreur réductible et irréductible

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = [f(X) - \hat{f}(X)]^2 + \text{Var}(\epsilon)$$

- ex : risque d'effets secondaires pour un médicament donné en fonction d'analyse sanguine

## Inférence

- Liens qualitatifs entre $Y$ et $X$
  - Quels sont les meilleurs prédicteurs de $Y$ ?
  - Quelle est la relation entre $Y$ et chaque entrée ?
  - Quel modèle peut relier $Y$ et les prédicteurs ?
- ex : ventes en fonction du budget pub dans médias : quel média contribue le plus aux ventes ?

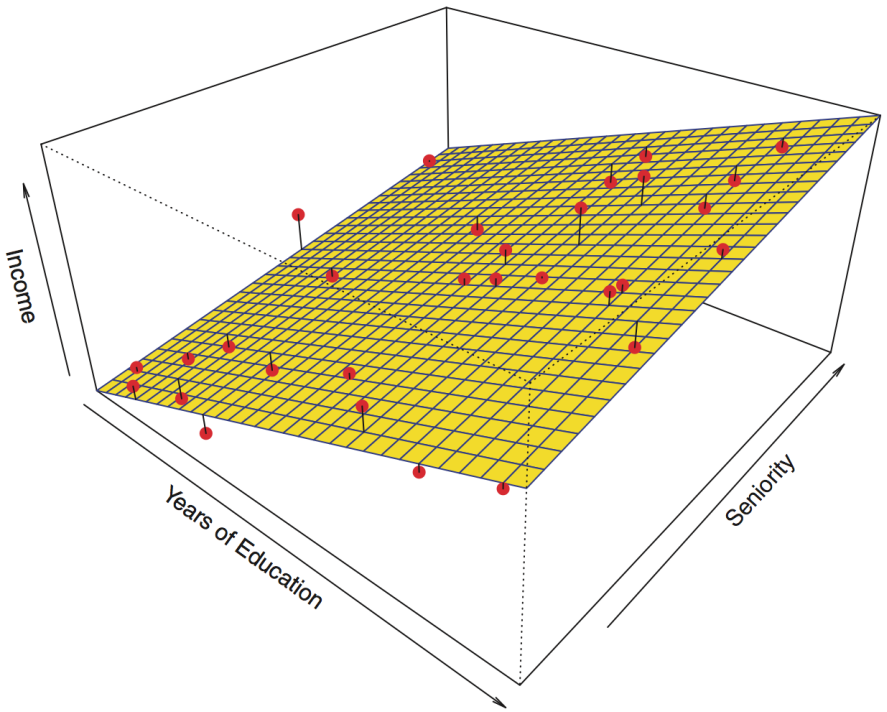
# Modèles paramétriques des données

1. Proposer un modèle, par exemple $f$ est linéaire en $X$ :

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

2. À partir des données, choix des meilleures valeurs de $\beta$ tel que

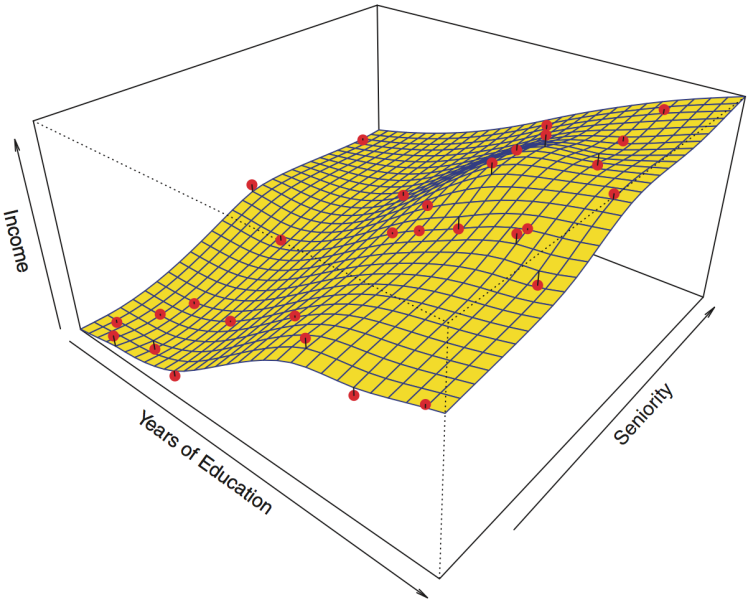
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$



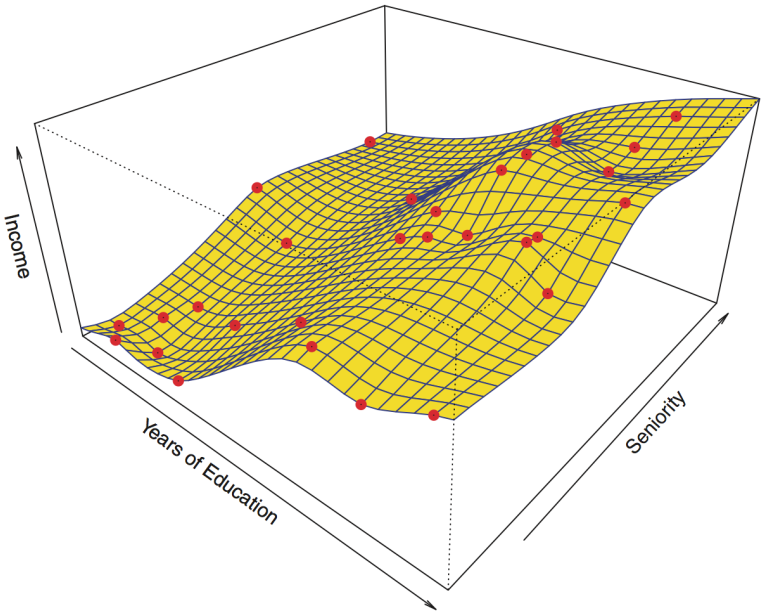

Estimation des moindres carrés pour le modèle linéaire

# Modèles non-paramétriques

- Pas d'hypothèses fonctionnelles sur la forme de $f$
- Besoin de plus de données
- Attention au surapprentissage (*overfitting*)




*thin-plate spline*
Lissage fort




Lissage faible
$\Rightarrow$ Overfit

# Prédiction vs interprétabilité

- Méthodes plus ou moins flexibles, plus ou moins restrictives
  Exemple : modèles linéaires vs. *thin-plate spline*
- Méthodes restrictives plus facilement interprétables
- Méthodes flexibles ont souvent une approche type boîte noire

Compromis à trouver

- Qualité des résultats (prédiction)
- Interprétabilité des résultats (inférence)

# Supervisé vs non-supervisé

## Supervisé

Étiquette ou classe :
pour chaque mesure $x_i$, $i = 1, \ldots, n$, une réponse est associée $y_i$

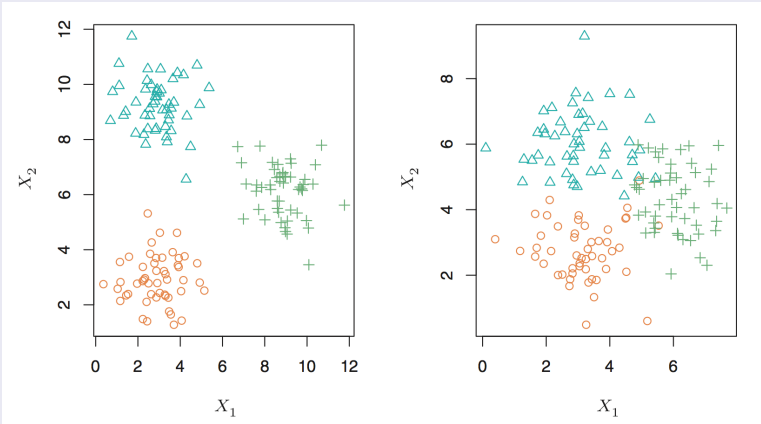
- Prédire la classe d'un échantillon

$\rightarrow$ Appliquer à des futurs échantillons

- ex : précédents

## Non-supervisé

- Pas d'information de classe
- Approche en aveugle
- Trouver une partition des données
- Séparation en *clusters*
- ex : profils clients



# Exemples d'apprentissage

## Supervisé ou non supervisé ?

- détection de spams
- regrouper les articles de presse qui parlent de la même histoire
- en partant d'une base de données d'achats, trouver les clients qui ont les mêmes habitudes d'achat
- en partant d'une base de données de patients diabétiques ou non, décider si de nouveaux patients le sont
- prédire le prix de logements à partir d'une base de données de ventes

Tâche ? Expérience ? Performance ?

# Régression vs classification

## Classification

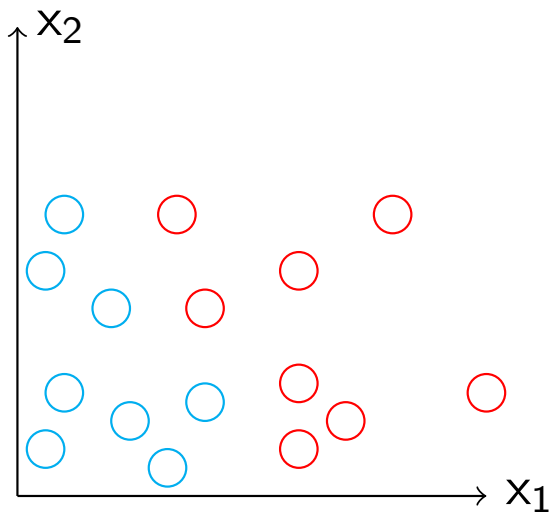
- Prédiction qualitative
- Binaire ou multi-classes
- Exemples : genre, groupes, diagnostics

## Regression

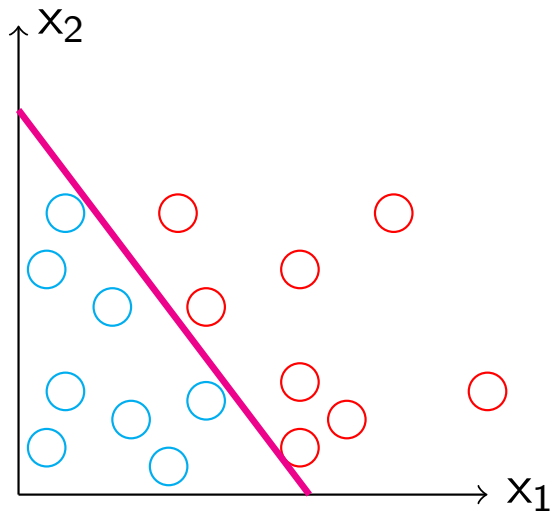
- Prédiction quantitative
- Exemples : taille, revenu, valeur, prix

- Beaucoup de méthodes font les deux
- Mais méthodes usuelles pour la classification diffèrent de la régression

# Ensemble d'apprentissage et frontière de décision

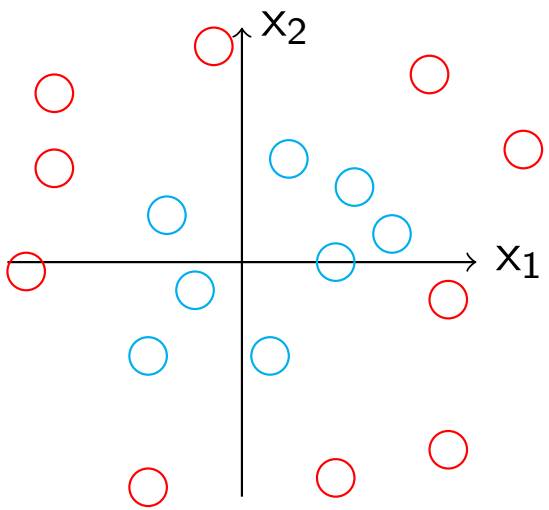


# Ensemble d'apprentissage et frontière de décision

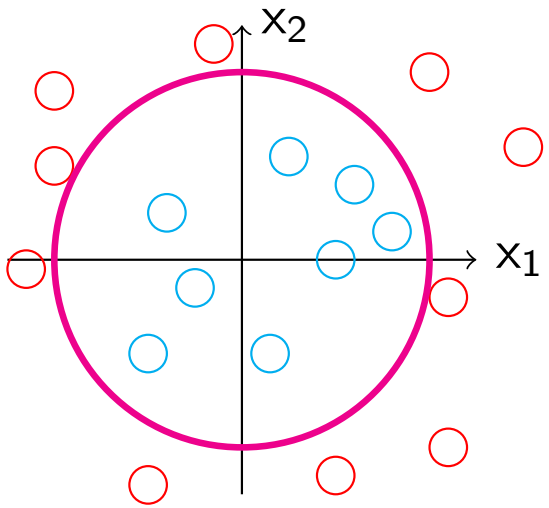



fontière linéaire

# Ensemble d'apprentissage et frontière de décision

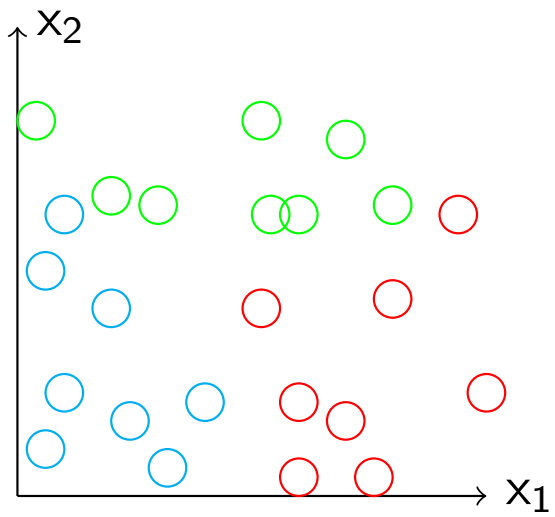


# Ensemble d'apprentissage et frontière de décision




fontière non linéaire

# Classification multi-classes

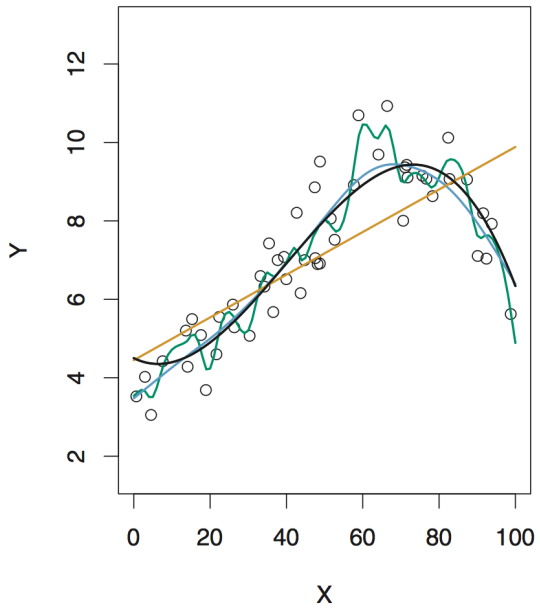


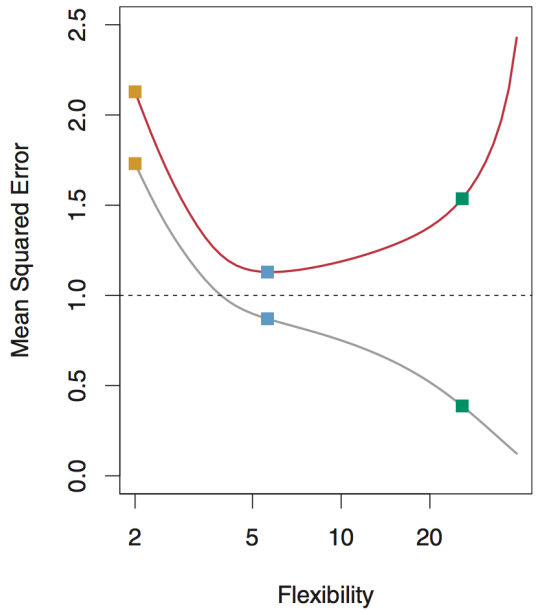
# Évaluation de l'erreur

Il n'y a pas de méthode parfaite, adaptée à tous les problèmes
$\Rightarrow$Théorème *no free-lunch*
Qualité de la solution, choix possible en régression :

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

- Exemples d'entraînement $\rightarrow$paramètres du modèle
- Quelle est l'erreur sur les données réelles ?



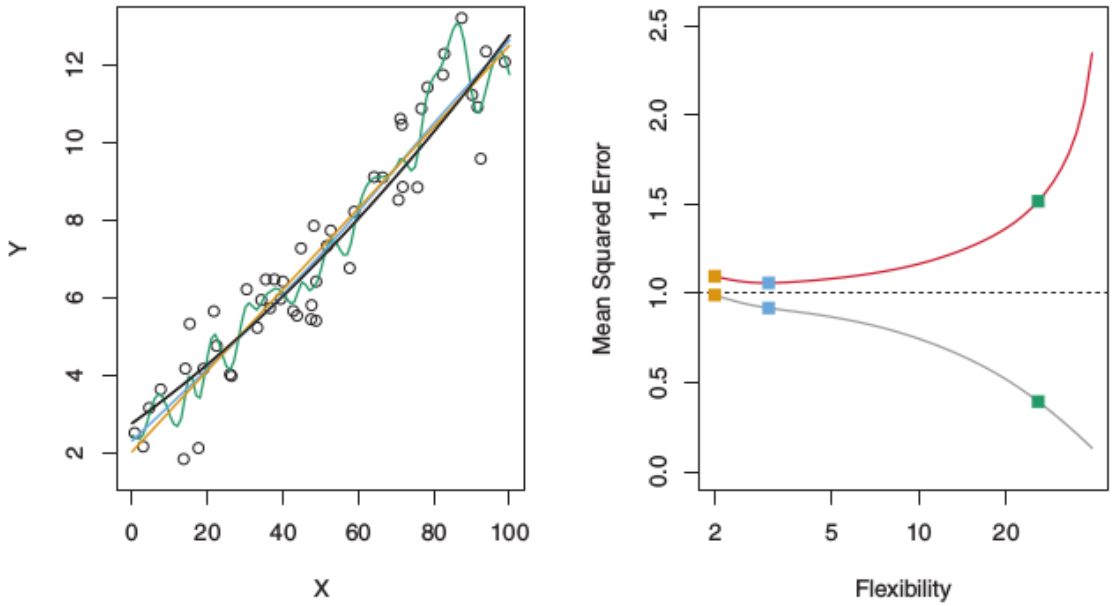



# Évaluation de l'erreur

Il n'y a pas de méthode parfaite, adaptée à tous les problèmes
⇒Théorème *no free-lunch*
Qualité de la solution, choix possible en régression :

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

- Exemples d'entraînement →paramètres du modèle
- Quelle est l'erreur sur les données réelles ?

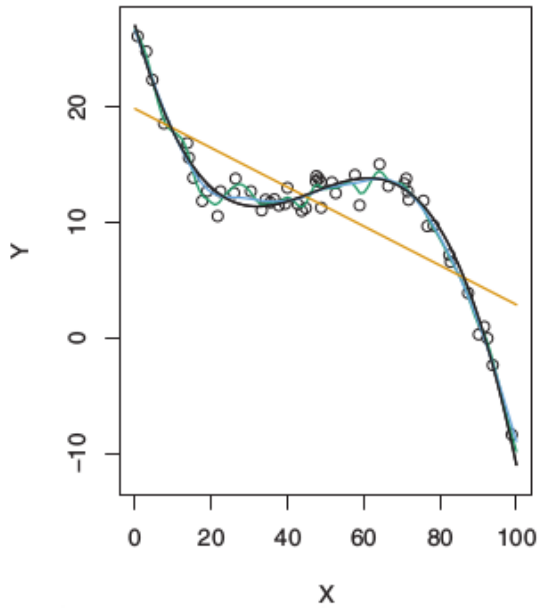


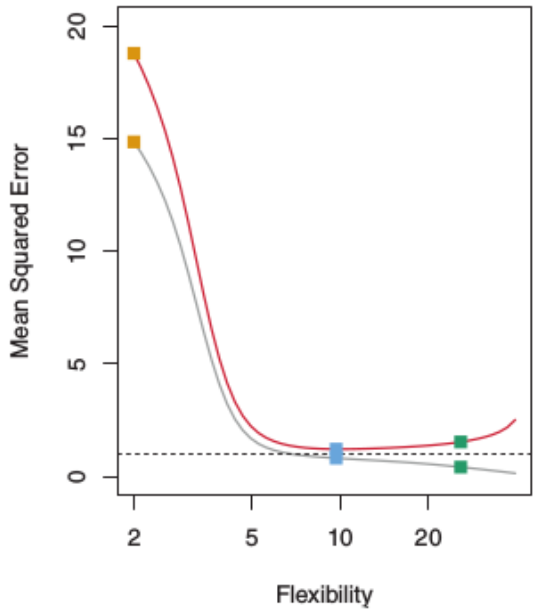
# Évaluation de l'erreur

Il n'y a pas de méthode parfaite, adaptée à tous les problèmes
$\Rightarrow$Théorème *no free-lunch*
Qualité de la solution, choix possible en régression :

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

- Exemples d'entraînement $\rightarrow$paramètres du modèle
- Quelle est l'erreur sur les données réelles ?



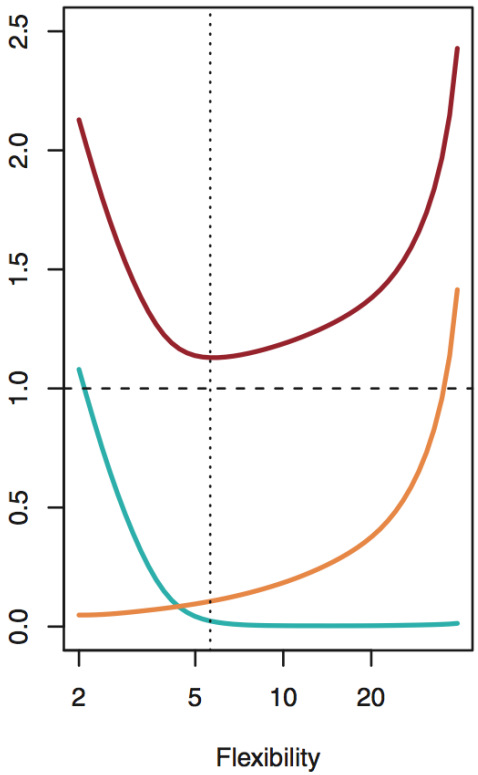
# Compromis biais-variance

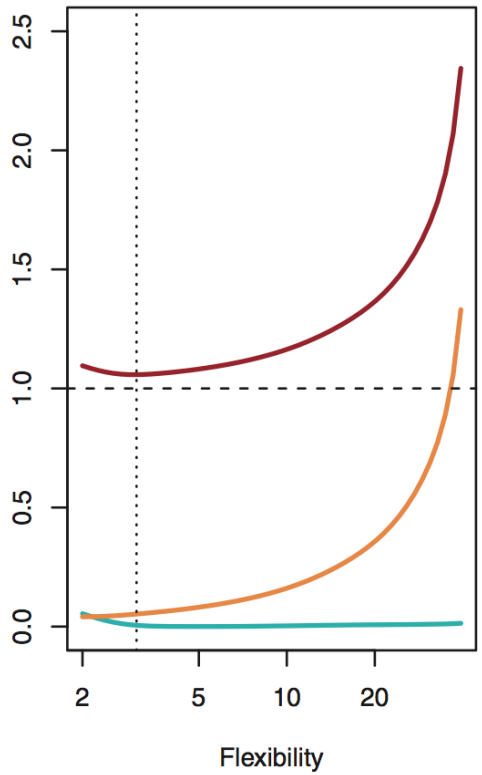
Les variations de l'erreur quadratique sont imputables à :

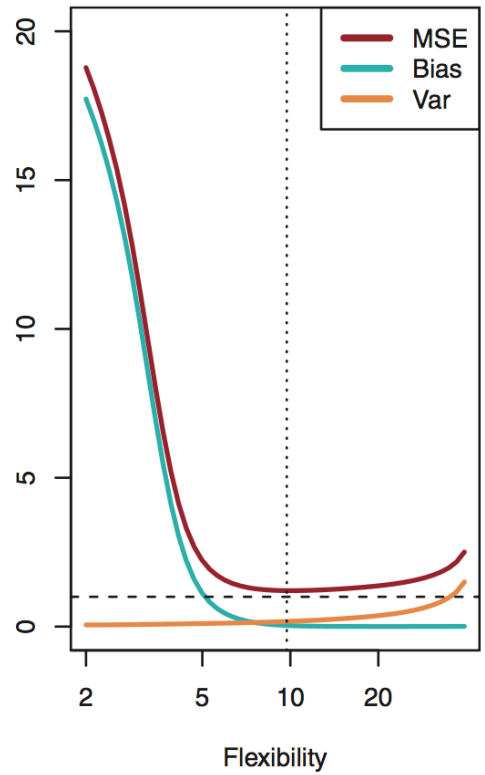
**variance** erreur de $\hat{f}(x_0)$ due à l'ensemble d'apprentissage

**biais** erreur $\hat{f}(x_0)$ due à la simplicité du modèle

**erreur** variations dues aux bruits de mesure, $\mathrm{Var}(\epsilon)$

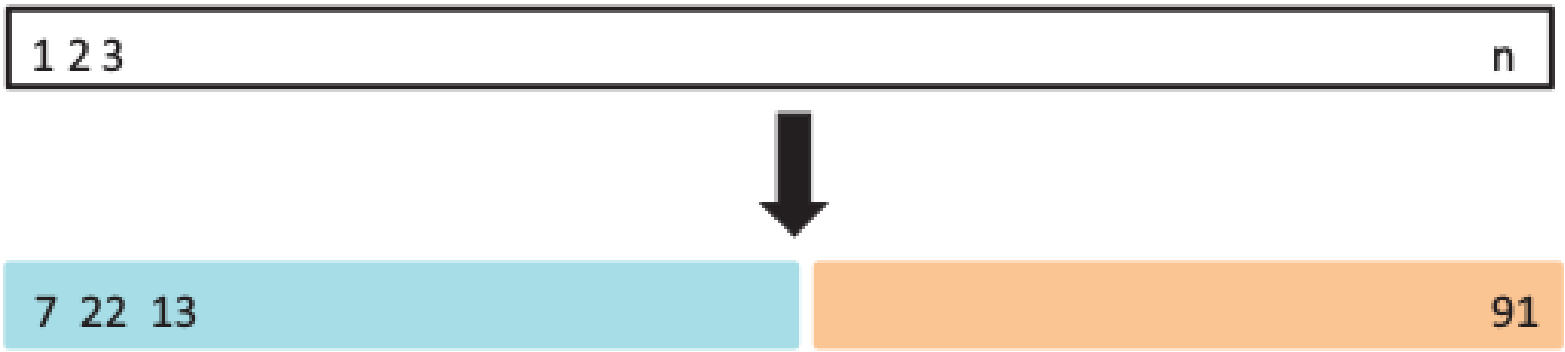


# Évaluation

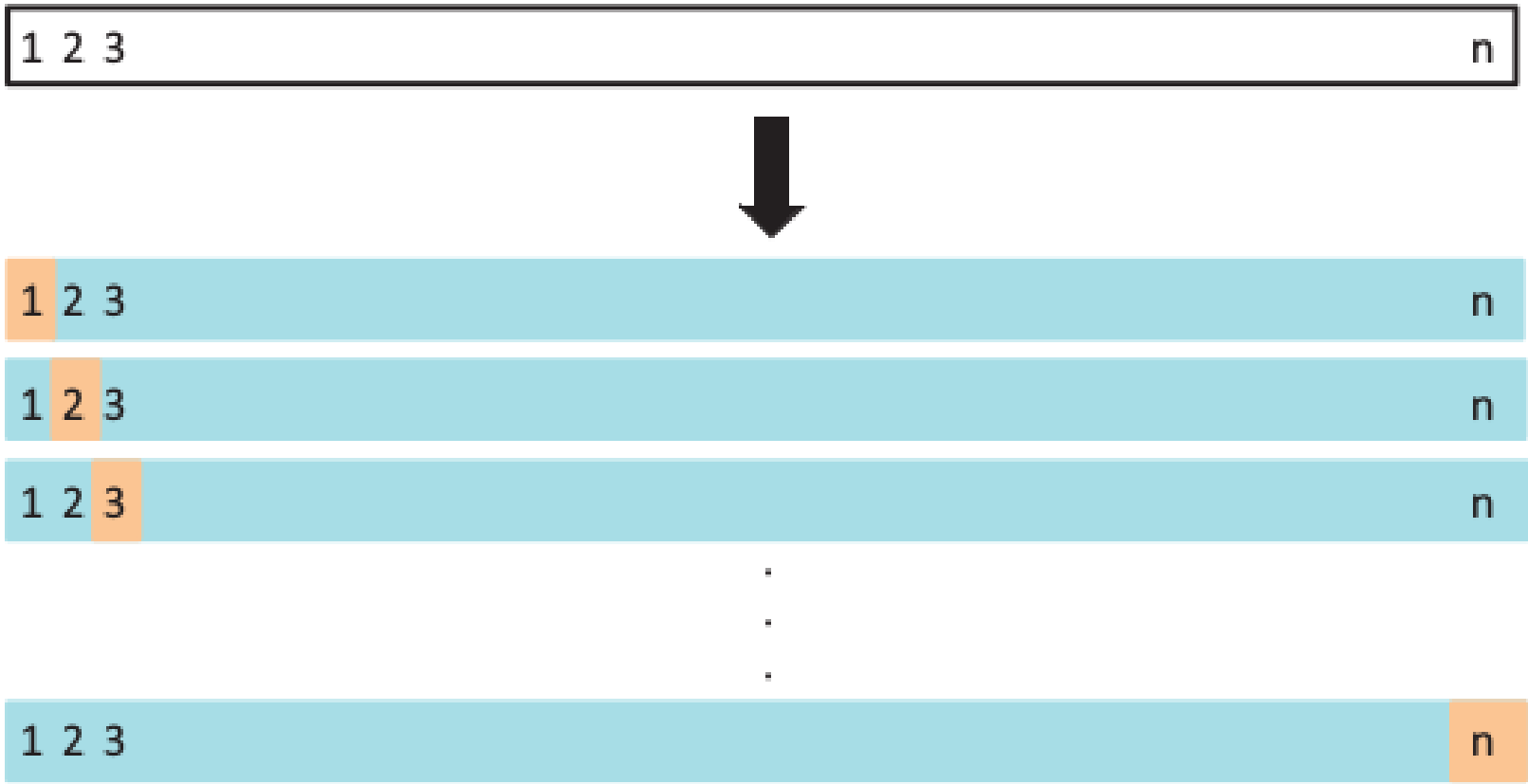
## Séparation entraînement/test




entraînement (*train*) et de test
si hyper-paramètres, + jeu de validation

# Évaluation

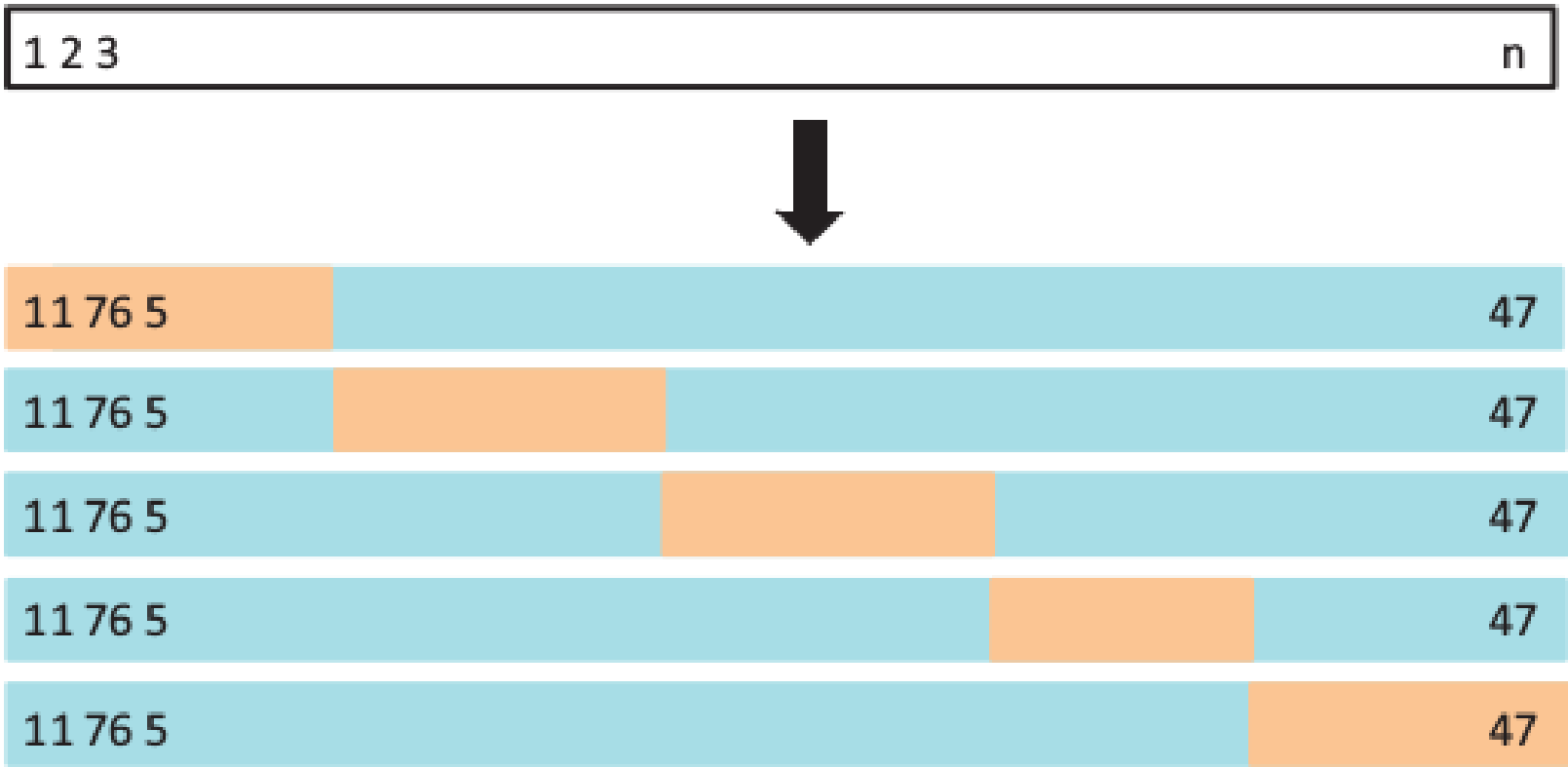
## Cross-validation leave-one-out




validation croisée (*cross validation*) en leave-one-out

# Évaluation

## Cross-validation ($k$-fold)




validation croisée en 5 fois

# Métrique d'évaluation

Taux d'erreur pas toujours pertinent quand jeu de données déséquilibré (cf. TP EN)

| | | classe réelle : 1 | classe réelle : 0 |
|---|---|---|---|
| classe prédite | 1 | vrai positif | faux positif |
| | 0 | faux négatif | vrai négatif |

Précision $p = \frac{vp}{vp+fp}$

Rappel $r = \frac{vp}{vp+fn}$

$F_1$ mesure $f = 2\frac{pr}{p+r}$

# Plein de modèles existent

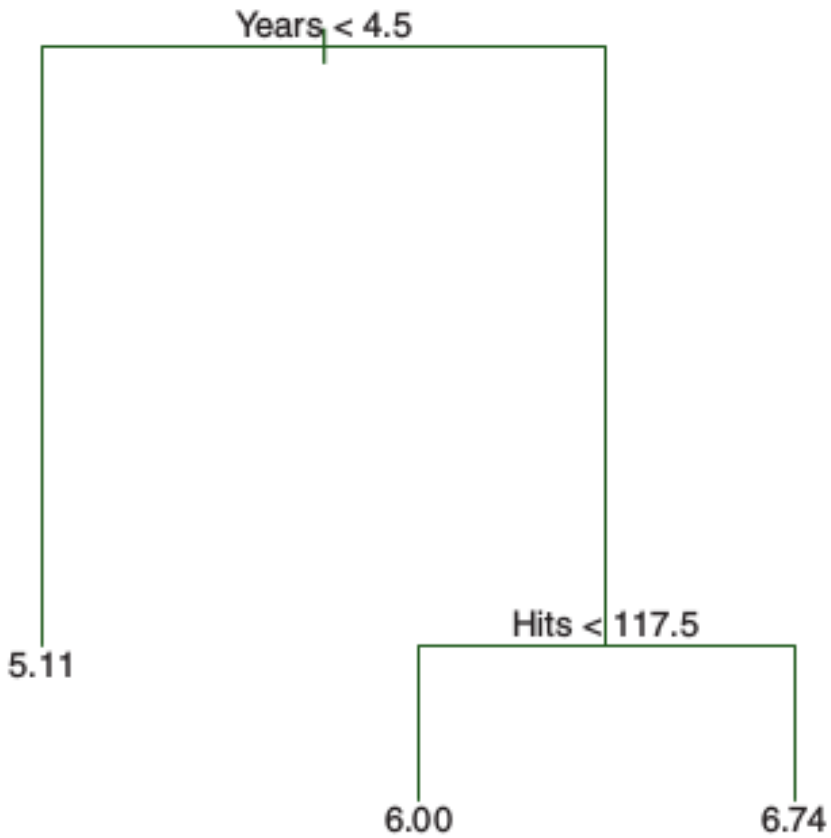
- arbres de décision
- réseaux de neurones
- SVM
- réseaux bayésiens

## différences

- complexité (mémoire/CPU)
- type d'entrée/sortie (vecteur, séquences, structures...)
- méthode d'optimisation sous-jacente (convexe, non convexe)
- performance

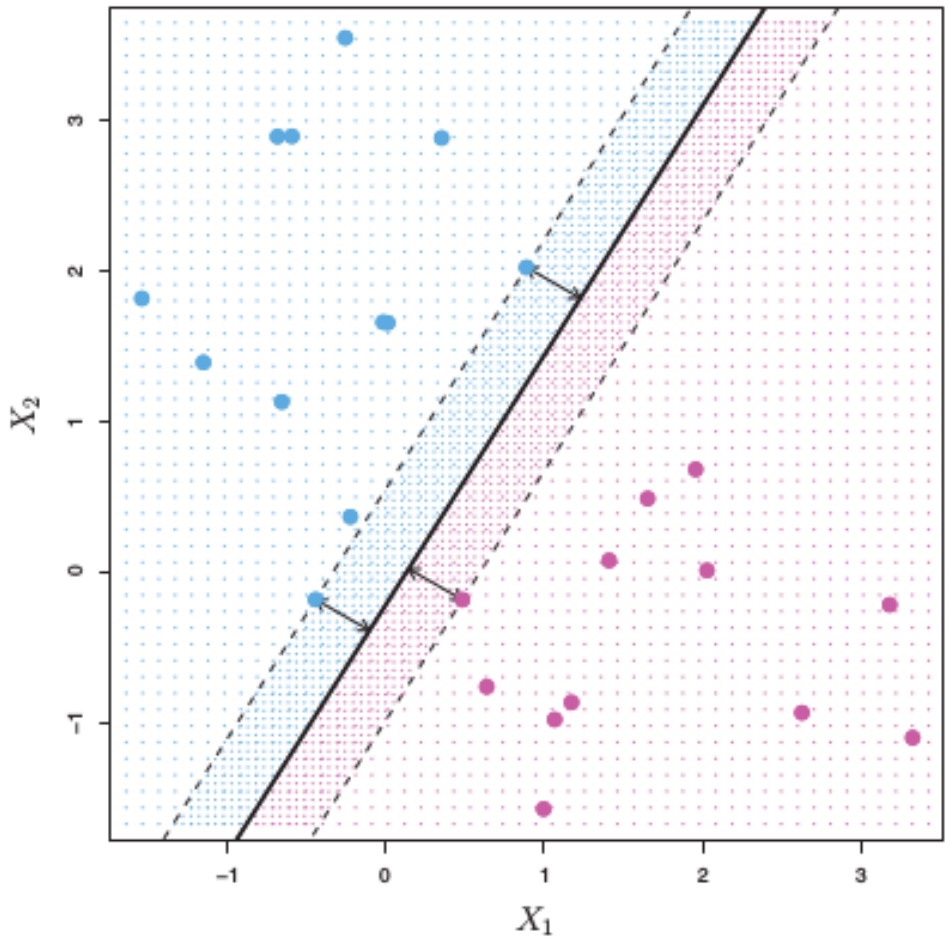
# Aperçu de quelques modèles

## Arbre de régression



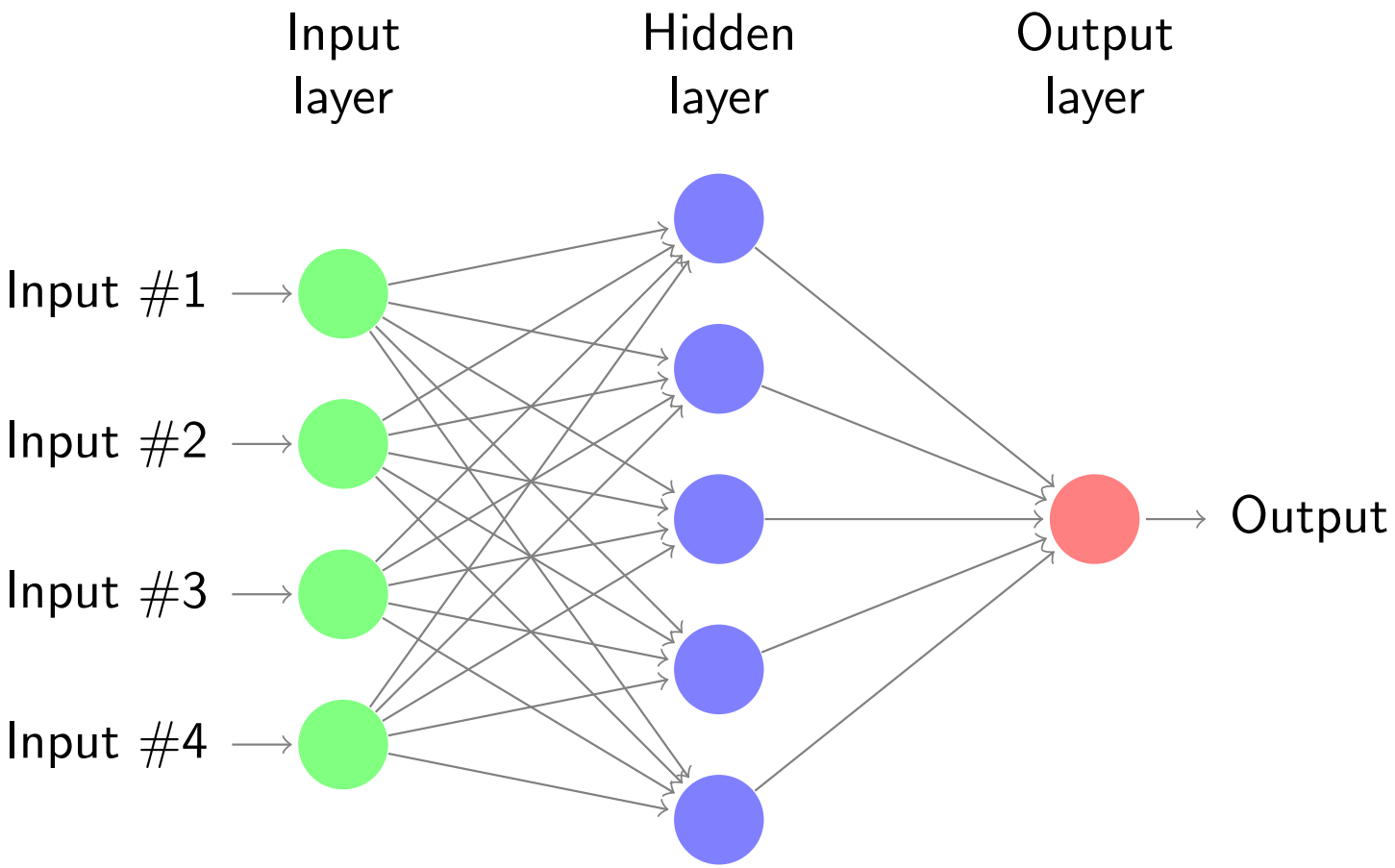

Salaire d'un joueur de baseball en fonction des années d'expériences et du nombre de hits

# Aperçu de quelques modèles
## SVM



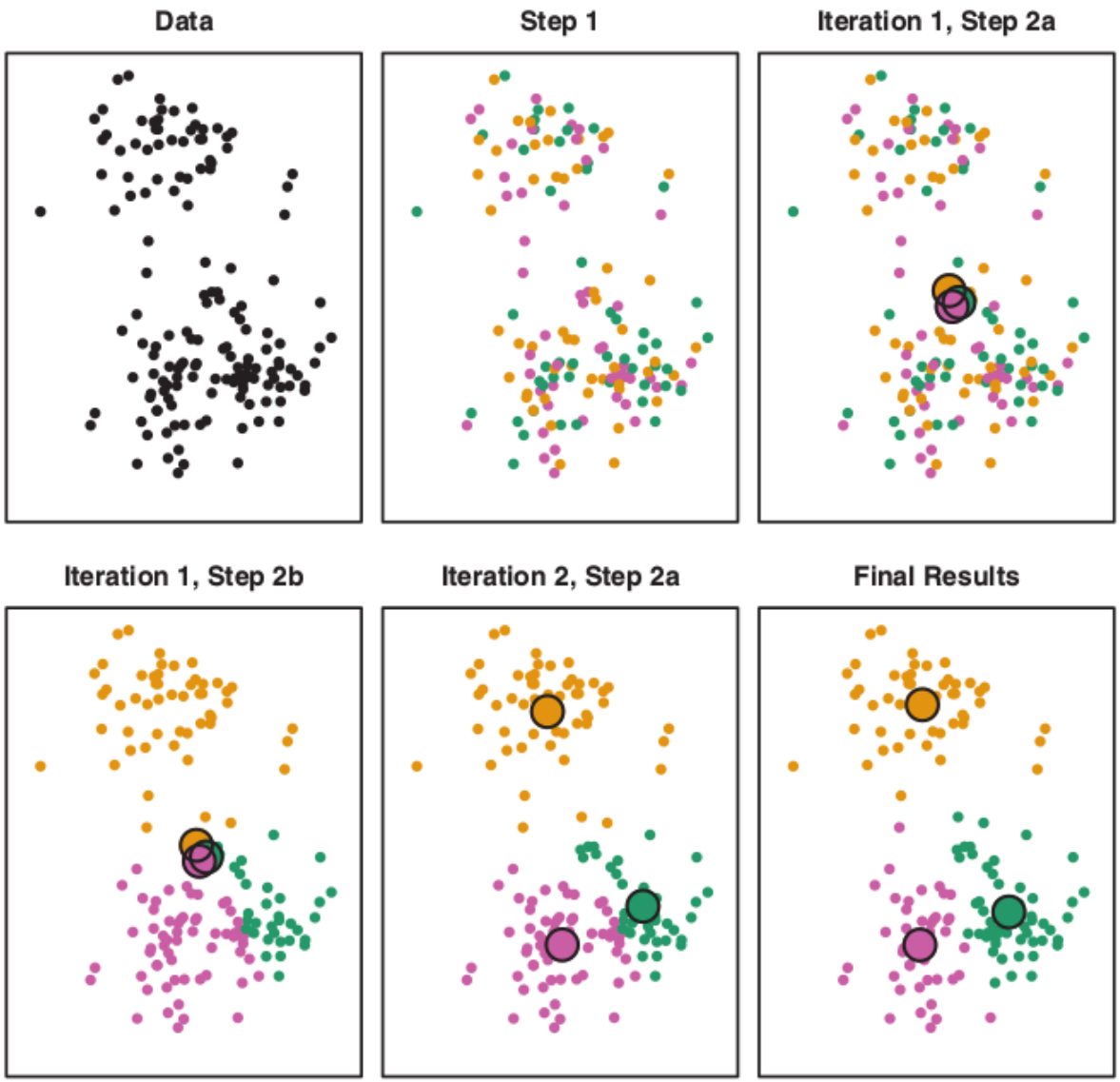
# Aperçu de quelques modèles

## Réseaux de neurones



# Aperçu de quelques modèles

## K-means



# Méthodologie

- Commencer par un algorithme simple et rapidement implémentable. Tester en cross validation
- Générer des courbes d'apprentissage pour savoir si plus de données, features etc. nécessaires
- Analyse d'erreur : examiner manuellement les exemples sur lesquels l'algorithme s'est trompé et essayer de détecter des tendances systématiques

# Plan

1. Apprentissage
   - Bases
   - Formalisme et modèles
   - Evaluation
   - Quelques modèles

2. Données textuelles
   - Modélisation du texte
   - Tâches de RI et d'EI

# Modélisations du texte

## Sac de mots (*bag-of-words ou BoW*)

Le bureau qui est sale est dans la pièce. $\Longrightarrow$

| | |
|---|---|
| bureau | 1 |
| est | 2 |
| la | 1 |
| le | 1 |
| les | 0 |
| pièce | 1 |
| table | 0 |
| ... | ... |

ensemble des mots du corpus = dictionnaire

### Avantages et inconvénients

+ simple et léger, rapide ; implémentations existances efficaces

  possibilités d'enrichissement

  bien adapté pour classification de documents

- perte de la structure des phrases/documents

  → pb pour certaines tâches : NER, POS tagging, génération de textes...

# Modélisations du texte

## Enrichissement de description vectorielle

- filtrage des mots vides (*stop list*), peu fréquents
- pondération : tfidf...
- racinisation, lemmatisation
- informations linguistiques : parties du discours (PoS)...
- n-grams de mots
- description du contexte des mots
- usage type : amélioration des tâches de classification au niveau du document

# Mieux gérer les séquences

- Modèles de Markov Cachés (Hidden Markov Models ou HMM)
- Conditional Random Fields ou CRF : approche discriminante
- Réseaux de Neurones Récurrents (Recurrent Neural Network ou RNN)

# Apprentissage et extraction d'information

- extraction d'entités $\rightarrow$ classification de mots
- extraction de relations $\rightarrow$ classification de phrases

# Apprentissage et recherche d'information

- classification thématique, classification de sentiments → classification de textes
- recherche d'information → ordonnancement (*ranking*) de textes

# Références
## ou Pour en savoir plus

- Introduction to statistical learning de Hastie Tibshirani
- Cours Machine Learning d'Andrew Ng