Apprentissage statistique

Cours 1 – Traitement de données et protocole expérimental

Olivier Schwander
<olivier.schwander@sorbonne-universite.fr>

Master Probabilités et Finance
Sorbonne Université

2023-2024

Les différentes étapes

Quelles sont les différentes étapes effectuées par un système de traitement de données ?

Conception d'un système

Les questions à se poser en premier

- Quel type de données ?
- Quel type de tâche ?
- Quelle quantité de données ?
- Quelle qualité des données ?
- Quels objectifs ?

Ensuite

- Quel prétraitement des données ?
- Quelles méthodes ?
- Comment choisir les paramètres ?
- Comment les évaluer ?
- Comment présenter les résultats ?
- Comment les interpréter ?

Données et tâches

Types de données

- Vectorielles
- Temporelles
- Graphes
- Texte

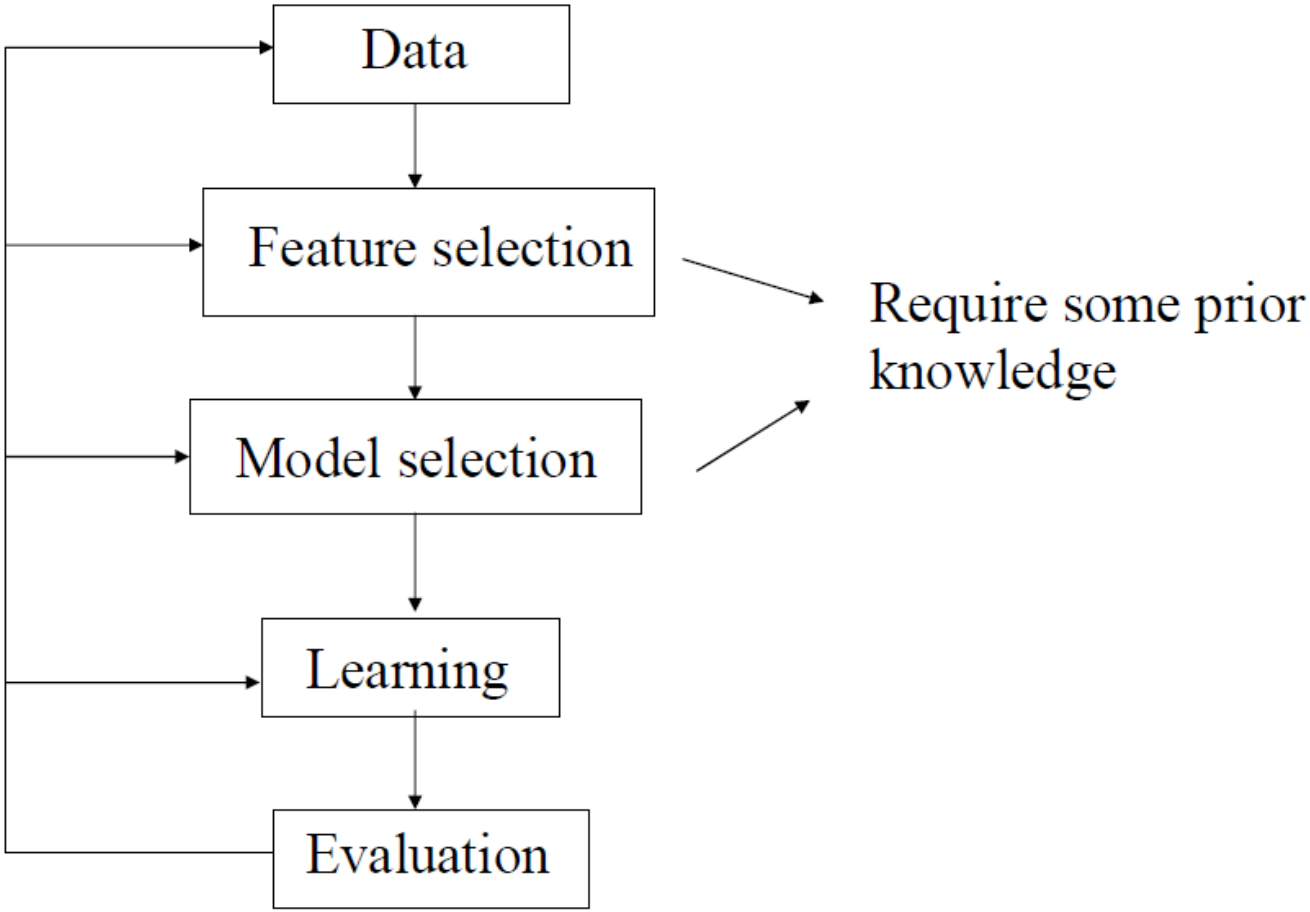
Différentes tâches

- Classification
- Régression
- Détection d'évènements
- Segmentation
- Recherche d'information
- Recommandation

Chaîne de traitement des données

1. **Données**
   - Charger
   - Analyser
   - Transformer
2. **Méthodes**
   - Choisir
   - Paramétrer
   - Apprendre
3. **Évaluation**
   - Mesurer
   - Présenter
   - Interpréter

Concevoir un modèle



Acquisition des données

Capteurs

- Données physiques (erreurs intrinsèques, position du capteur)
- Températures, humidité, pression, etc

Indicateurs

- Calculés d'une façon ou d'un autre
- Rentrés à la mainn

Extract / Transform / Load - Business Intelligence

Systèmes d'apprentissage

- Vision, Texte, Voix
- pour guider un autre système d'IA

Pré-traitement

- Renommage
- Normalisation
- Discrétisation
- Abstraction
- Aggrégation
- *Sélection d'attributs - Features sélection*
- Création d'attributs

Biais dans les données

- Comprendre la source des données
- Éviter des choix a priori basé sur l'intuition
- Connaissance experte souvent utile

Malédiction de la dimension

- Dimension du problème trop élevée
- Trop de variables d'entrées
- Trop de paramètres du modèle

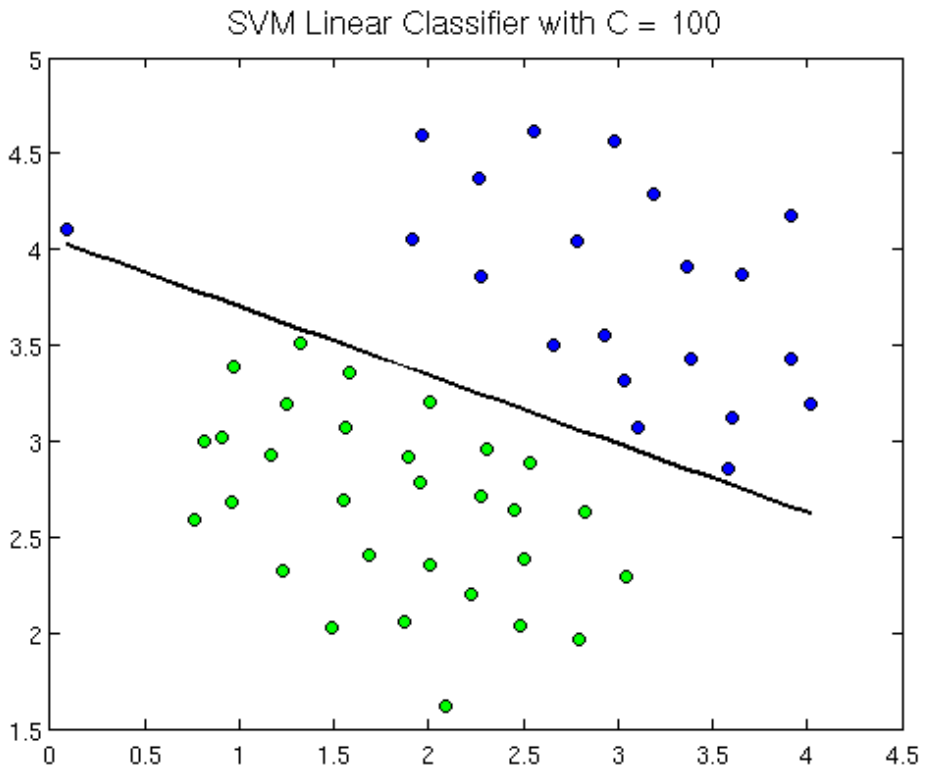
Heureusement

- Données concentrées dans un petit sous-espace
- Structure dans les données

Solutions

- Réduire la dimension
- Transformation manuelle (expert métier)
- Apprendre la transformation

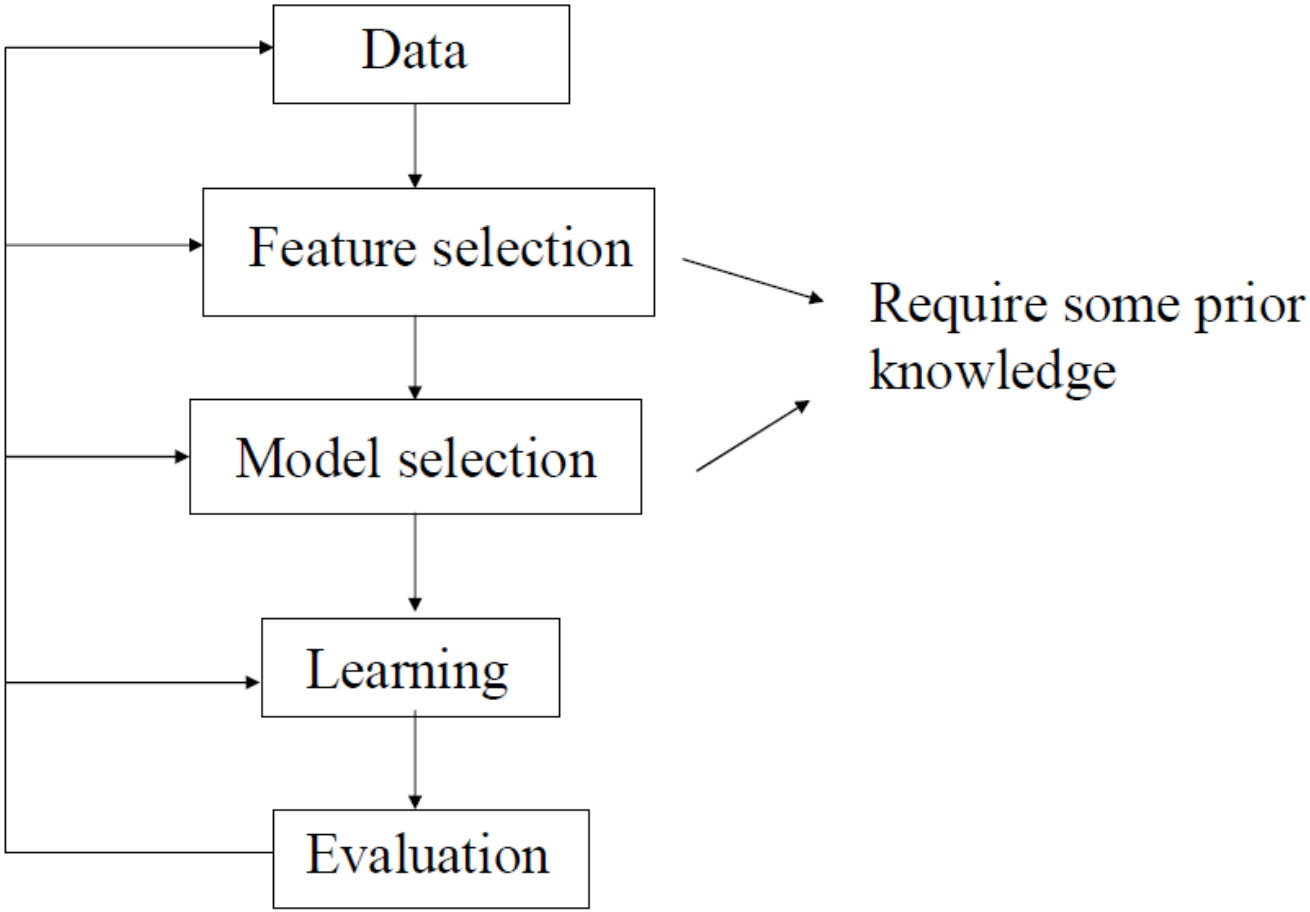
Outliers



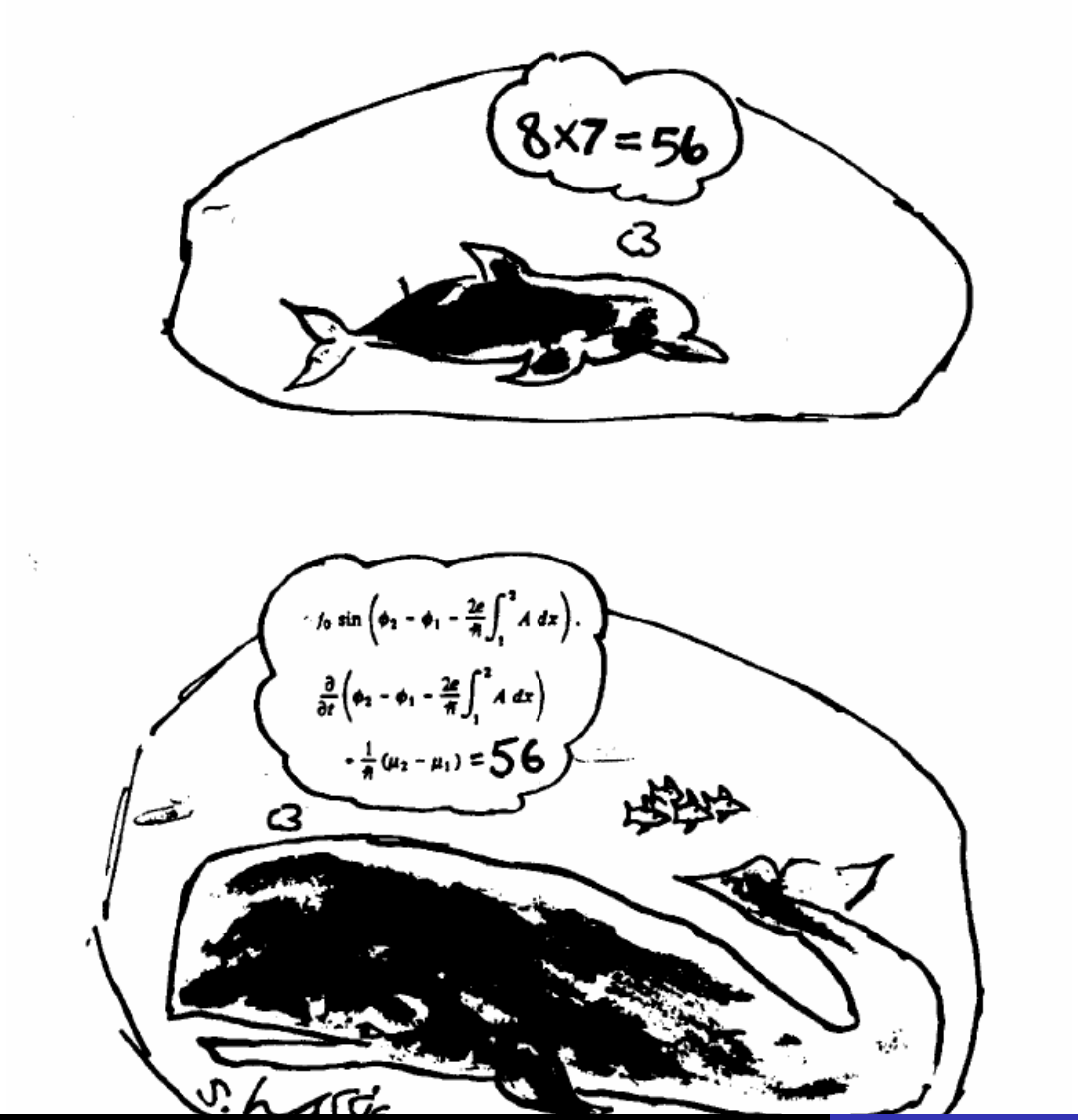
Suppresion des outliers

- Connaissances métier
- Méthodes statistique

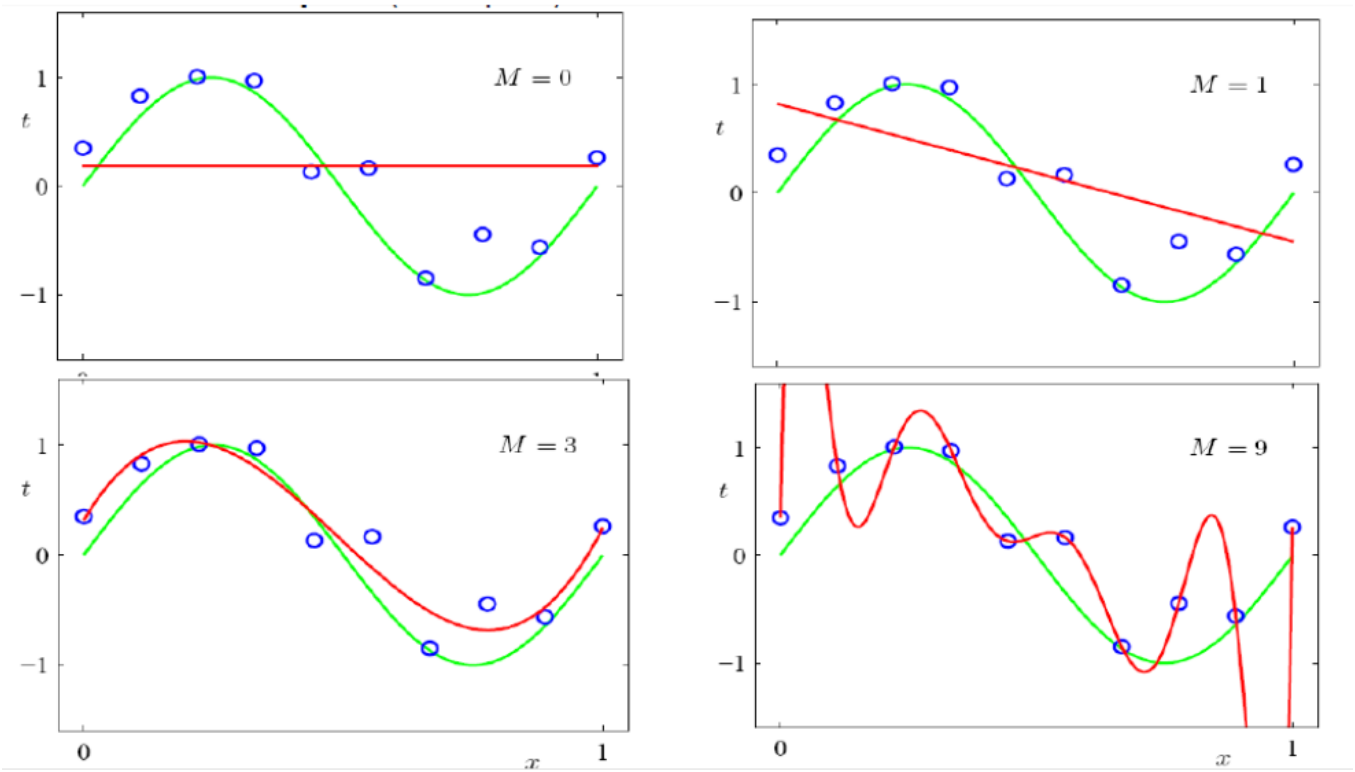
Concevoir un modèle



Sélection de modèle

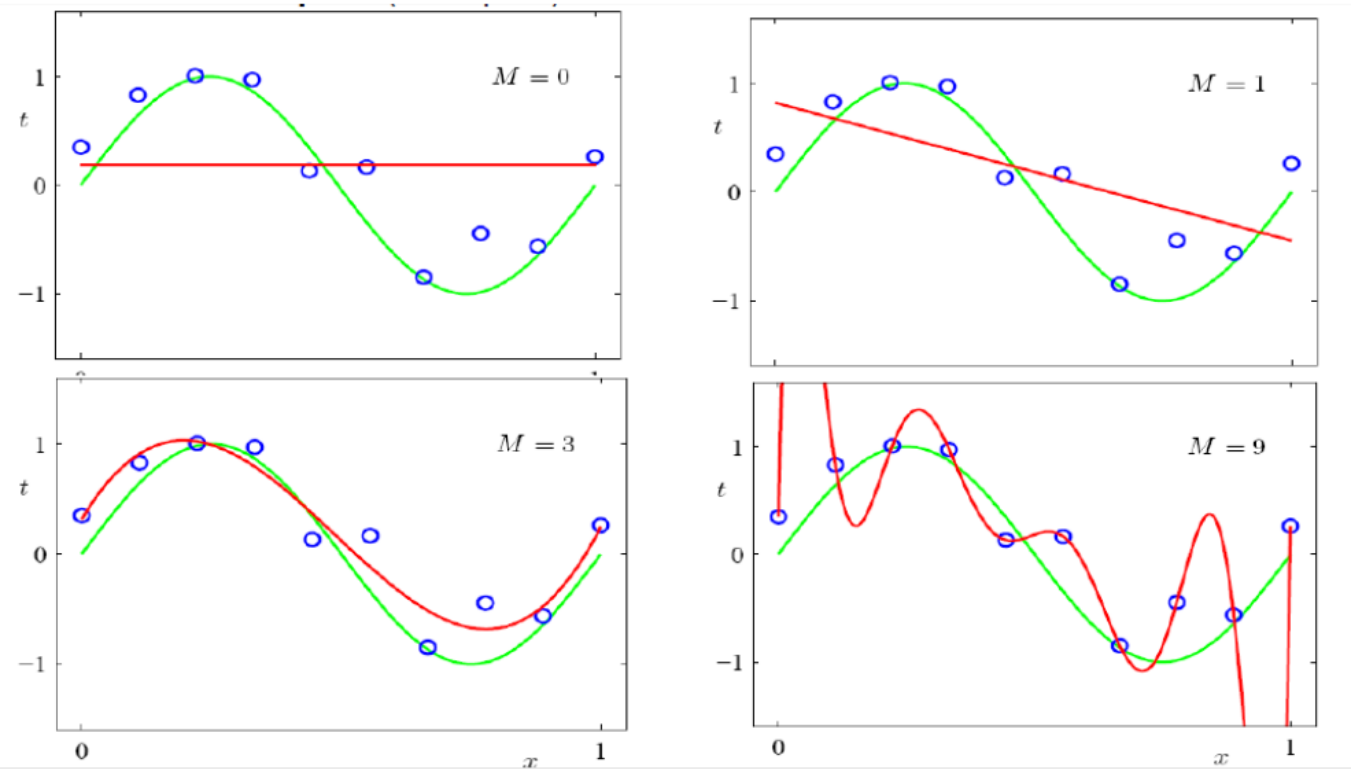


Sélection de modèle



Quel est le meilleur modèle ?

Sélection de modèle



Objectif: généralisation

Sélection de modèle

On cherche des moyens de sélectionner le “meilleur” modèle parmi un ensemble de modèles possibles

Bruit et Régularités **Données** = **Bruit** + **Régularités**

- Bruit: Erreurs dans l'acquisition
- Régularités: Processus de génération sous jacent

Objectif: **Modèle final** = **Capture du bruit** + **Modèle des régularités**

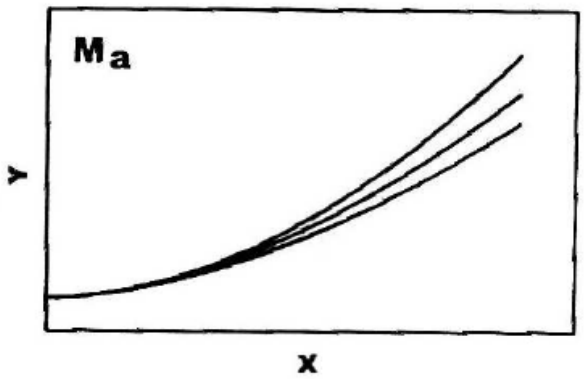
Meilleur modèle:

- Meilleur modèle des régularité
- Meilleure capture du bruit

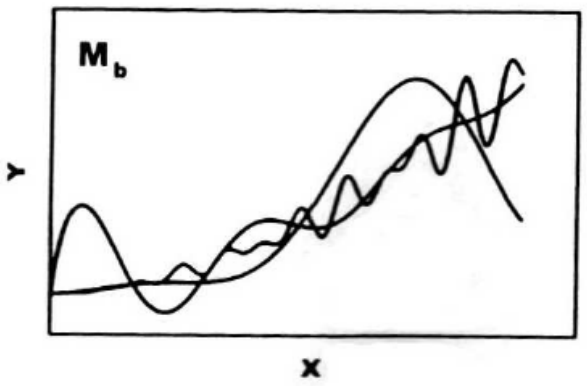
Généraliser: éviter le sur-apprentissage

Complexité d'un modèle

Simple Model



Complex Model



- Nombre de paramètre
- Classe de fonction choisie

Critère d'information d'Akaike - 1973

$$AIC = -2 \ln \hat{L} + 2k$$

- $\hat{L}$ est la vraisemblance du modèle sur les données $= P(x|\theta^*, f)$
- k est le nombre de paramètres du modèle

Méthodologie

- Pas de découpage train/test
- Entraîner plusieurs modèles
- Calculer leur AIC
- Prendre le modèle avec le meilleur AIC (le plus faible)

Critère d'information d'Akaike - 1973

Divergence de Kullback-Leibler (KL)

- On suppose que les données sont générées par un processus p
- Soit des modèles f_i
- $KL(p||f_i)$ mesure l'information perdue en approchant p par f_i
- Le meilleur modèle est celui qui minimise cette divergence
- **Problème:** on ne connait pas p

Estimateur **asymptotique**

- l'AIC permet de comparer des modèles

Variante pour petits jeux de données:

- $AICc = AIC + \frac{2k(k+1)}{n-k-1}$

Autres critères

- Critère d'information Bayésien - 1978: $BIC = -2 \ln \hat{L} + k \ln n$
- Minumum Description Length - 1978: *learning as data compression*

Principe général à retenir: rasoir d'Occam

- *Pluralitas non est ponenda sine necessitate*
- *Les multiples ne doivent pas être utilisés sans nécessité*
- Sélectionner le modèle le plus simple qui modélise les données *suffisamment* bien

Sélection de modèles par échantillonage

Deux grandes familles de méthodes pour se faire une idée de l'erreur de généralisation..

- La loi des grands nombres: l'utilisation de bornes statistiques permettant de borner la différence entre l'erreur empirique et l'erreur théorique (sous certaines hypothèses)

$$\forall f \in \mathcal{F}, \quad \mathcal{R}_P(f) \leqslant \widehat{\mathcal{R}}_n(f) + \frac{1}{\sqrt{2n}} \sqrt{\ln(2) \underbrace{|f|_\pi}_{\text{complexite}} + \ln \frac{1}{\delta}}.$$

- L'utilisation d'échantillons différents pour l'évaluation de l'erreur

Découpage train/test

Deux sous-ensembles

Base d'apprentissage

- Utilisé pour l'entraînement
- Sous-apprentissage: mauvaise performances en train
- Besoin d'une performance correcte

Base de test

- **Distinct du train**
- Quelle taille ?
- Choix des exemples ?
- Ojectif: bien ce comporter sur ce dataset

Sélection de modèles par échantillonage

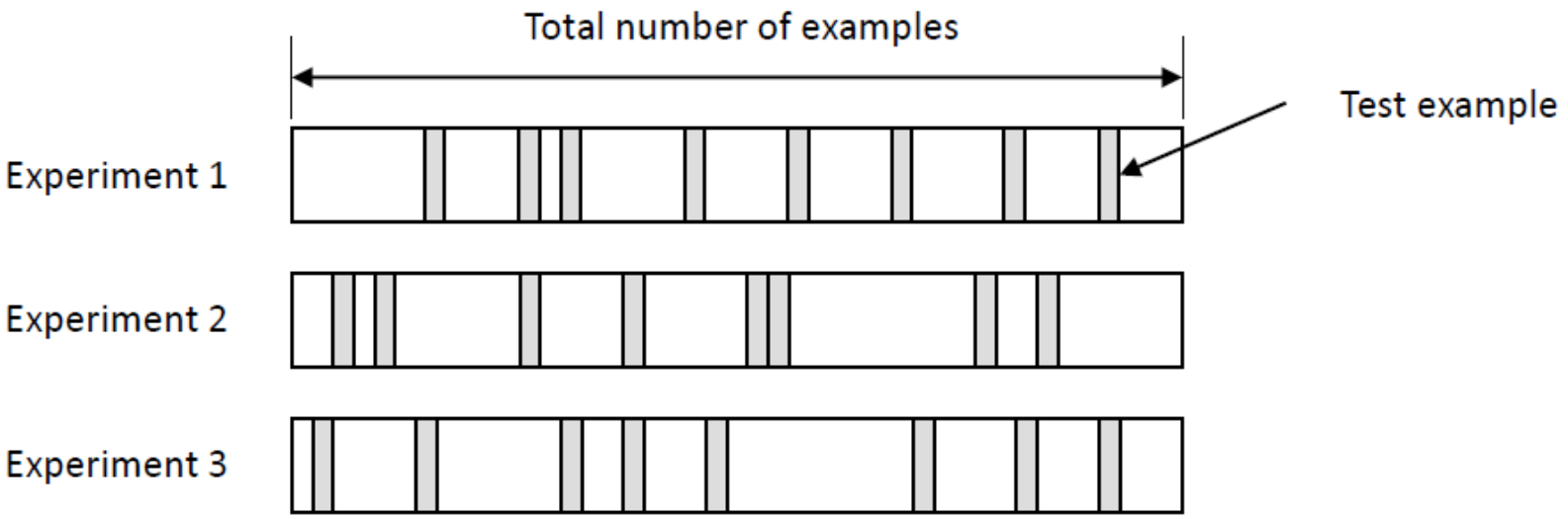
Problèmes

- Pas assez de données qui restent en train ?
- Sous-ensemble facile ? difficile ?
- Sensibilité aux données d'apprentissage

Plusieurs solutions

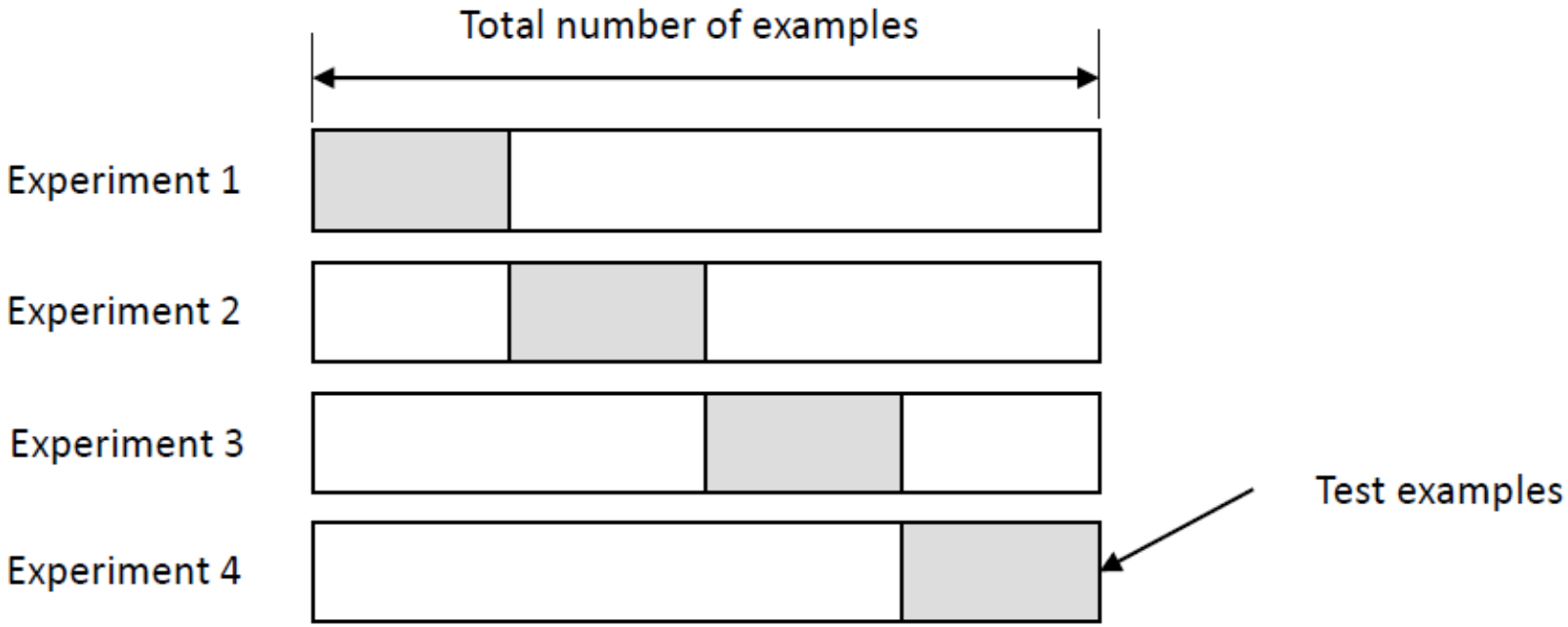
- Rééchantillonage aléatoire
- Cross-validation

Rééchantillonage aléatoire



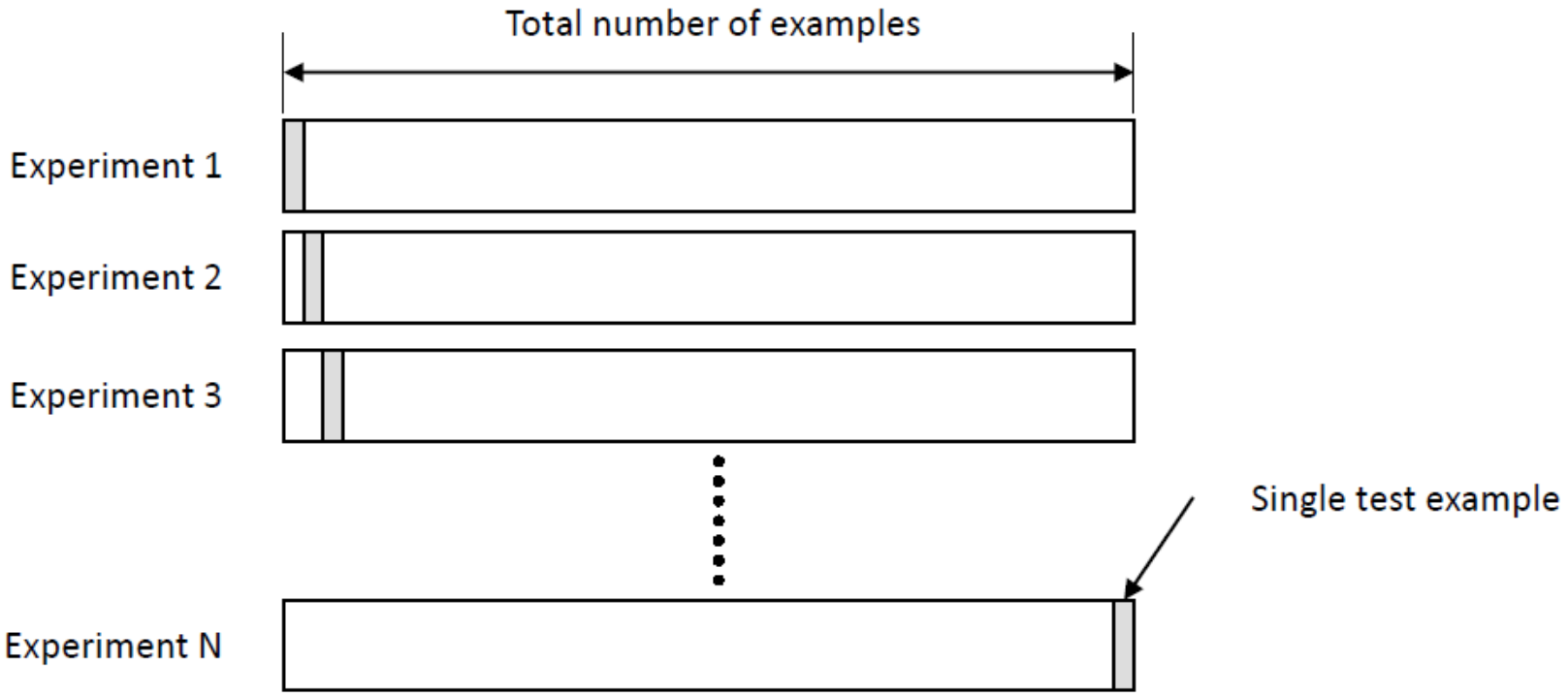
- Erreur du modèle: moyenne sur les différentes expériences
- Estimation significativement meilleure (avec assez de tirages)

Cross-Validation



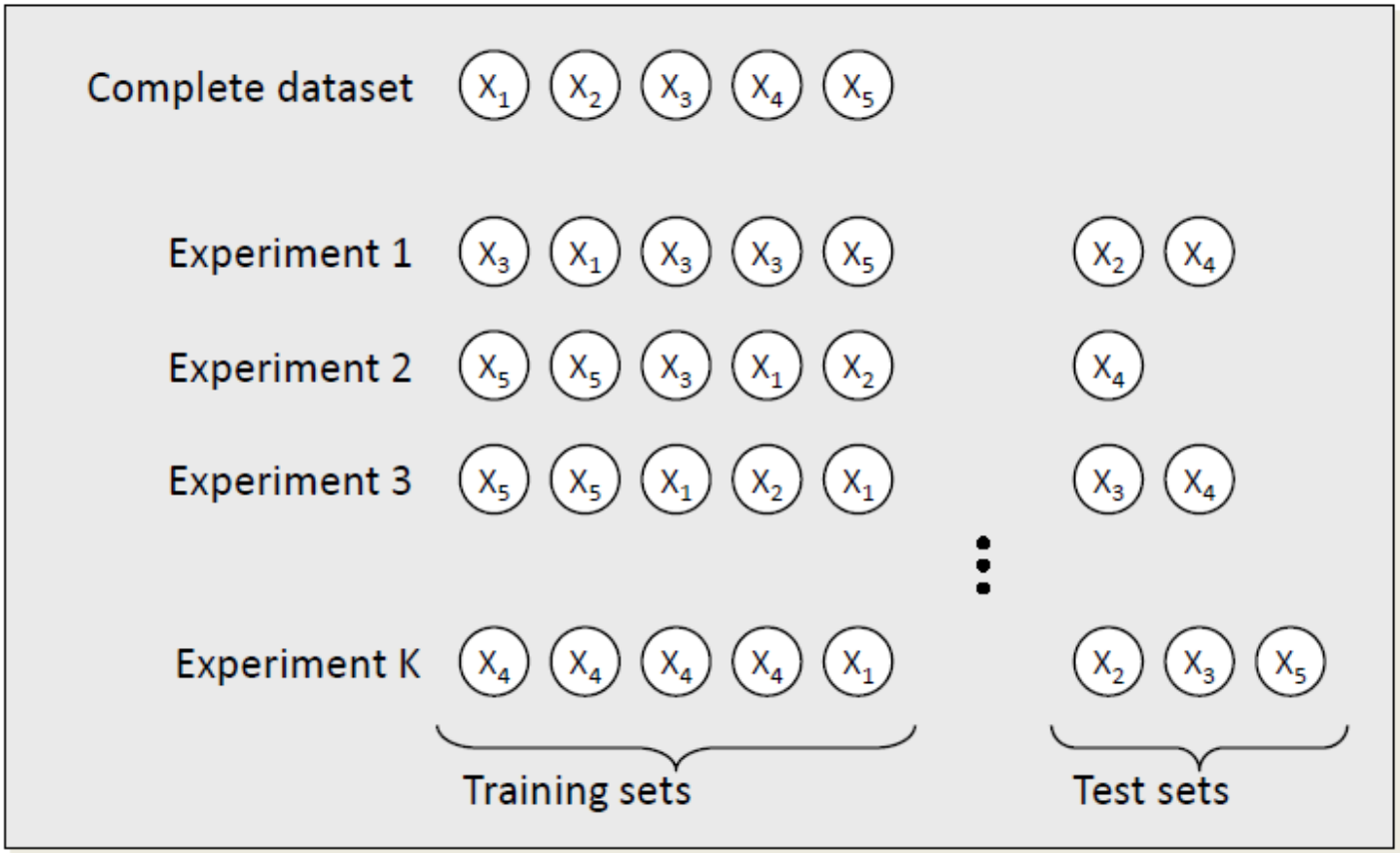
- Erreur du modèle: moyenne sur les différentes expériences
- Tous les exemples sont utilisés au moins une fois en train

Leave-one-out



- Erreur du modèle: moyenne sur les différentes expériences
- Cas dégénéré de CV: plus robuste, meilleurs pour les petits jeux de données

Bootstrap



- Plus grande variance dans les différents “folds”
- Mais effet désirable car plus réaliste

Ensemble de validation

En même temps

- Trouver le meilleur modèle
- Estimer la performance en généralisation

3 sous-ensembles:

- *Train*
- *Validation* pour la séléction
- *Test* pour l'évaluation

Courbes d'apprentissage

(dessin au tableau)

Protocole expérimental

Ensemble des choix faits précédemment

- Dataset
- Découpage train/val/test, avec cross-val ou non, etc
- Méthode de mesure du score

Comparer des modèles

- **Même protocole expérimental**
- Doit rester identique au cours du projet
- Doit être documenté **précisément** pour le futur

Documentation

En lisant un rapport, ou un article, on doit pouvoir mettre en œuvre le même protocole expérimental, pour pouvoir se comparer aux scores présentés.

Rapport technique

Contenu exploitable

- Mise en production
- Poursuite du travail par quelqu'un d'autre
- **Que fait-on et comment**

Comparaison des résultats

- Description du protocole expérimental

Contenu intéressant

- **Objectif**: information à faire passer
- Intéressant pour le lecteur, pas pour l'auteurice
- Surtout pas de cours de machine learning (pitié !!!!!!!!!!!!!!!)

Figures

Utilité

- Pas seulement pour faire joli
- Commentaire dans le texte
- Objectif: information à faire passer

Lisibilité

- Lisible sans zoomer, penser à l'impression noir&blanc
- Vectoriel (donc *pdf* et pas *png* ou *jpg*)
- Titres sur les axes, légendes

Conseils

- Enregistrer les données brutes, pratique pour refaire la figure
- Noms de fichiers clairs

Article de recherche

Pareil

Subtilités

- Style d'écriture du domaine
- Description des contributions: **insister lourdement**
- Motivation des contributions
- Bibliographie
- Protocole expérimental (j'en ai peut-être déjà parlé, non ?)

Slides

Pareil

Subtilités

- Éviter les phrases dans les slides
- Ce qui est affiché est là pour être lu, pas pour faire joli
- Cibler plus, peut-être en parlant de moins de choses
- **Adapter au public**

Séléction de caractéristique

Sélection de caractéristiques sélectionner un sous-ensemble des caractéristiques existantes:

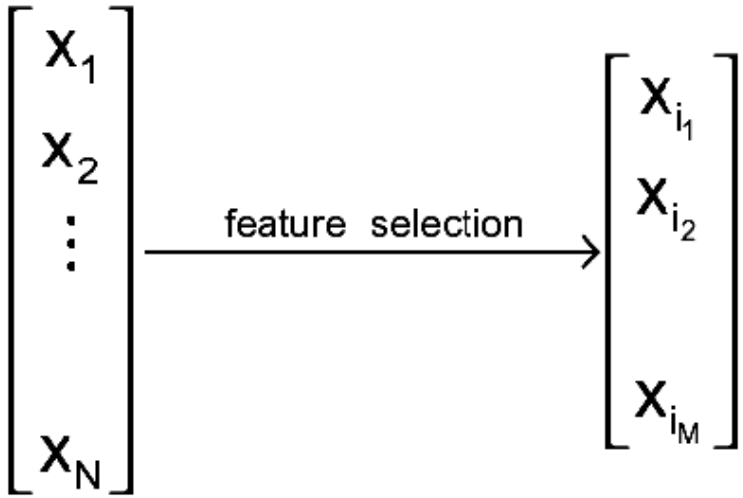
- Approches de type **Filtering**
- Approches de type **Wrappers**

Extraction de caractéristiques combiner des caractéristiques existantes pour obtenir un (petit nombre) de caractéristiques pertinentes:

- Approches de type **PCA**
- Approches de type **Auto-Encodage**
- Approches de type **Representation Learning (Deep Learning)**

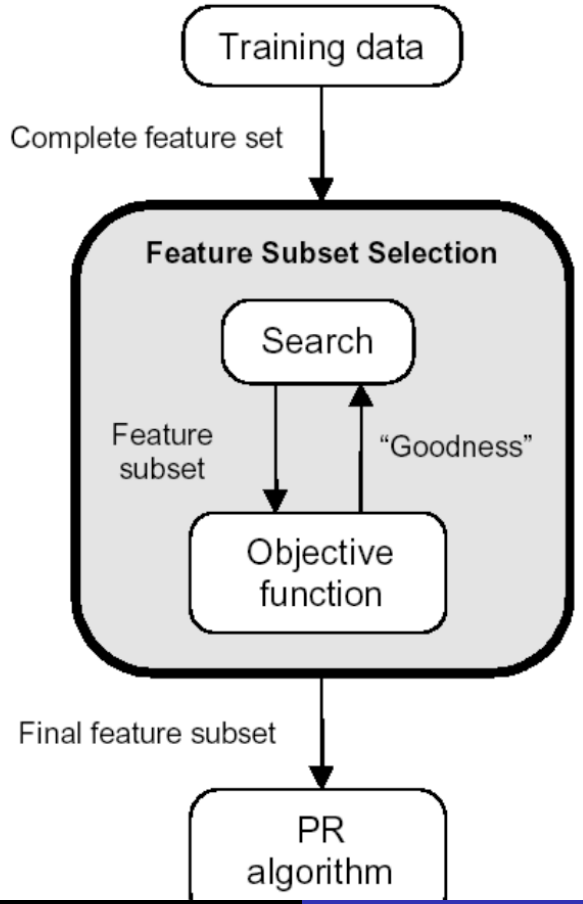
Séléction de caratéristiques

- Soit un ensemble d'entrée $\mathcal{X} = \mathbb{R}^n$ tel que $x = (x_1, x_2, ..., x_n)$
- On cherche à trouver un sous-ensemble de dimensions caractérisé par un ensemble $\mathcal{I}$ d'index dans $[1; n]$
- Etant donné $\mathcal{I} = (i_1, ...i_M)$, le nouvel espace d'entrée sera caractérisé par $x = (x_{i_1}, x_{i_2}, ..., x_{i_M})$



Sélection de caratéristiques

- Très grand espace de recherche
- Besoin de méthodes approchées



Deux approches

Méthodes de filtrage: sélection *a priori*

- Estimation du pouvoir prédictif de chaque caractéristique
- Étude mono-dimensionnelle de chaque caractéristique
- Sélection de celles avec le pouvoir prédictif le plus élevé

Méthodes de wrappers: sélection *a posteriori*

- Choix basé sur la qualité du modèle obtenu

Corrélation

Mesure de l'intensité de la liaison entre deux variables

Corrélation linéaire Soit la variable X_i (caractéristique) et la variable Y (étiquette):

$$Corr(X_i, Y) = \frac{Cov(X_i, Y)}{\sqrt{Var(X_i)Var(Y)}}$$

- $Cov(X_i, Y) = E[X_i Y] - E[X_i]E[Y] = E\left[(X_i - E[X_i])(Y - E[Y])\right]$
- $Cov(X_i, Y) = 0$ ssi X_i et Y sont indépendantes

Corrélation empirique

Comme d'habitude lois inconnues pour X_i et Y

Estimateur

$$R(i) = \frac{\sum_{k=1}^{N} (x_i^k - \bar{x}_i)(y^k - \bar{y})}{\sqrt{\sum_{k=1}^{N} (x_i^k - \bar{x}_i)^2 \sum_{k=1}^{N} (y^k - \bar{y})^2}}$$

- Dépendance linéaire
- Versions non-linéaires
- **Corrélation n'est pas causalité**

Méthodes de filtrage

- Tri des variable par ordre de pertinence
- Conservation des caractéristiques les plus pertinentes

Avantage

- Chaque caractéristique est analysée **indépendamment des autres**.
- Rapide

Limite

- Chaque caractéristique est analysée **indépendamment des autres**.
- Une variable pourrait être utile en combinaison avec une autre

Méthodes de wrappers

- Choisir un sous-ensemble de caractéristiques
- Entraîner un modèle et l'évaluer
- Choisir le sous-ensemble qui donne les meilleurs performances

Coûteux:

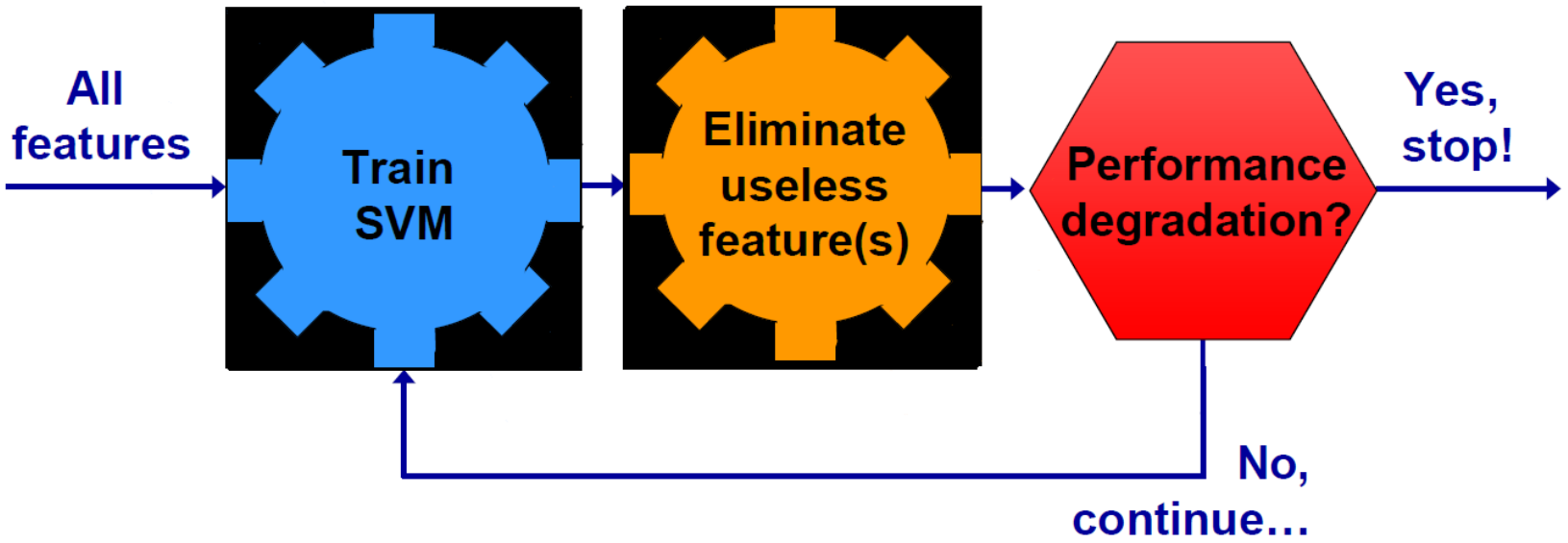
- Nombre exponentiel de sous-ensembles
- Entraînement des modèles

Recherche gloutonne: ajout graduel de caractéristiques basé sur un score à chaque pas de l'algorithme

Attention: le score doit refléter la performance du système (en généralisation)

Méthodes embarquées

De moins en moins de caractéristiques



Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*

Conclusion

Grandes lignes d'un projet ML

- Définir la tâche et y réfléchir
- Analyser les données (statistiques descriptives, feature engineering)
- Établir un protocole expérimental
- Choisir un modèle et évaluer ses performances
- Présenter le travail