BI et Data Mining
Cours 9
Master DAC Data Science
UPMC - LIP6

Ludovic Denoyer

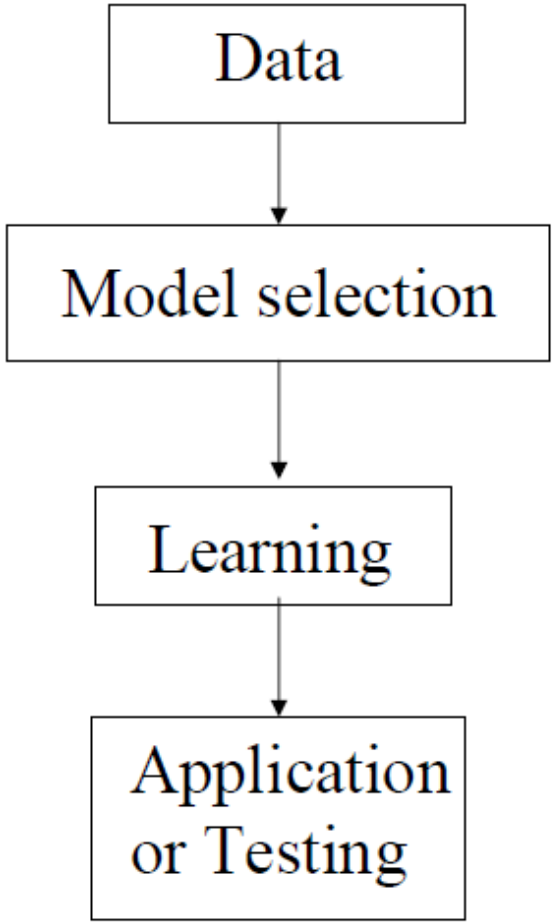
20 mars 2016

Brainstorming : Les modèles prédictifs

Les différentes étapes

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

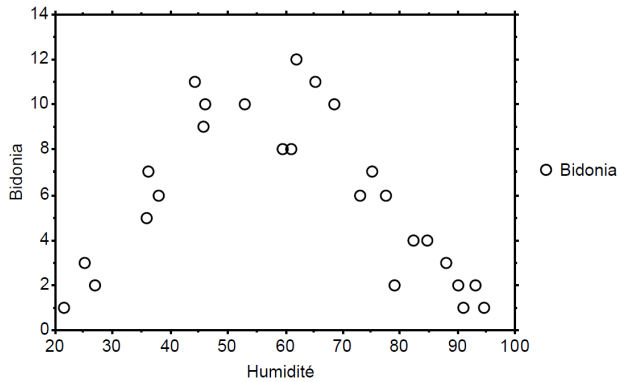
Les différentes étapes



Exemple : la régression polynomiale

Collecte des données

- Acquisition de données **vectorielles** à l'aide de capteurs, wrappers, etc...
- Inputs= $\mathcal{X} = (x_1, ..., x_n) \in \mathbb{R}$
- Acquisition d'une **vérité terrain** - valeurs à prédire - sur $\mathcal{X}$
- $y_1, ..., y_n$ avec $y_i \in \mathbb{R}$



N.B : La collecte de données implique un nettoyage important des données, ainsi qu'une sélection/engineering des caractéristiques

Exemple : la régression polynomiale

Choix du modèle

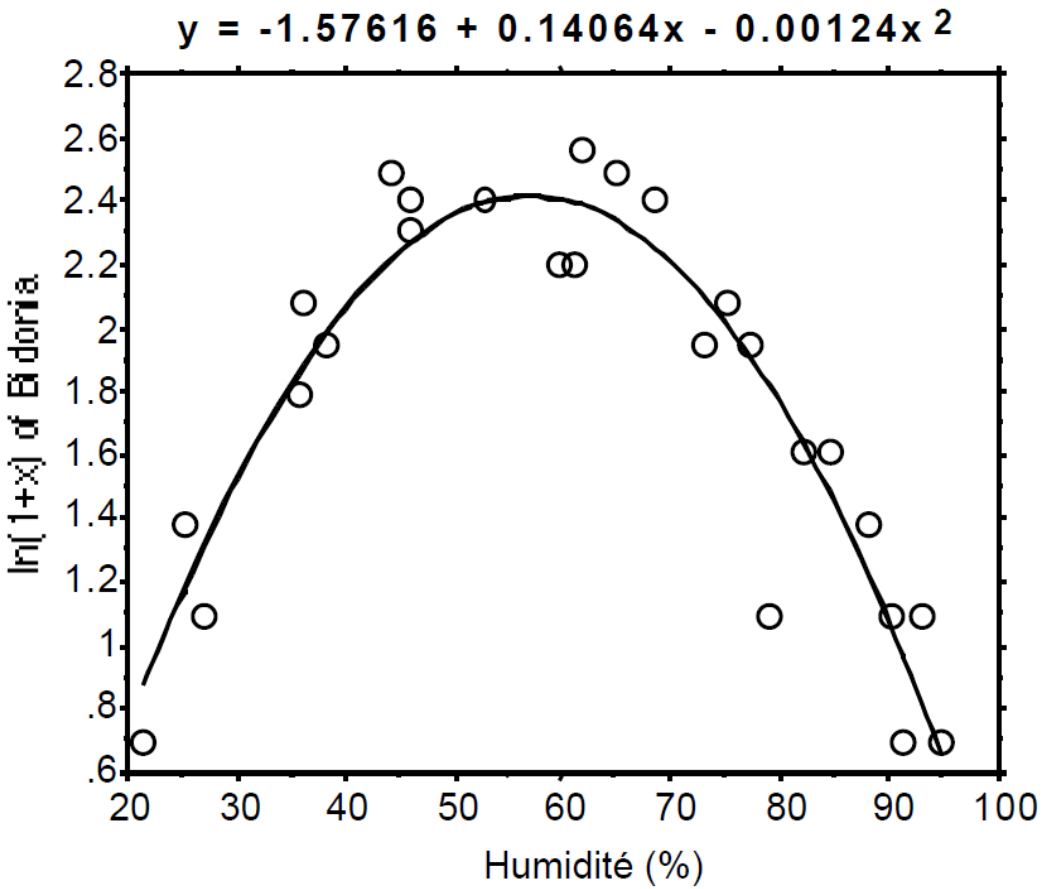
- Choisir un modèle paramétrique $f_\theta(x) \rightarrow y$
  - Catalogue : Régression, arbres, forêts, réseaux de neurones, SVM, CRFs, HMMs, Réseaux bayésiens, Machine de Boltzmann, ...
  - Choix de la topologie/hyper-paramètres du modèle
    - Topologie du réseau de neurone, type de régressions, type d'arbres, liens entre variables, ...
- $\theta = (\theta_1, ..., \theta_P)$
- $P =$ degré du polynôme
- Modèle : $y = \sum_{k=1}^{P} \theta_k x^k$

Exemple : la régression polynomiale

Apprentissage du modèle

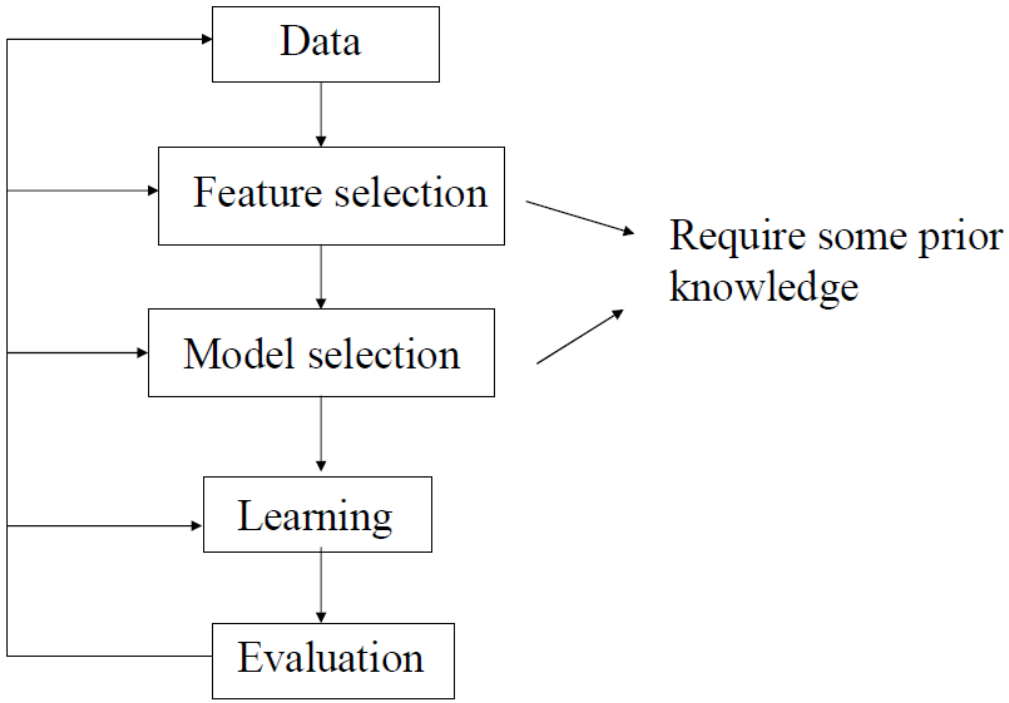
- Choix du critère d'apprentissage —- (cf AS)
  - Critère des moindres carrés : $\theta^* argmin_\theta \sum_i^n (x_i - f_\theta(x_i))^2$
- Choix de l'algorithme d'optimisation
  - Descente de gradient, SMO, Résolution analytique, ...
  - $\Rightarrow$ choix des paramètres de l'algorithme d'optimisation
- Optimisation effective des paramètres $\Rightarrow \theta^* \Rightarrow f_\theta^*$

Exemple : la régression polynomiale



- **Le modèle est prêt à être utilisé sur de nouvelles données ! ! !**
- Mais, beaucoup de choses à choisir. Comment faire ?

Designer un modèle



QQes mots sur le nettoyage de données

Acquisition des données

- A travers des "capteurs"
  - Capteurs réels, logiciels, wrappers
  - Nécessite une connaissance experte
- A travers l'usage de systèmes d'ETL
- ...

Petit retour en arrière : Intégration de données

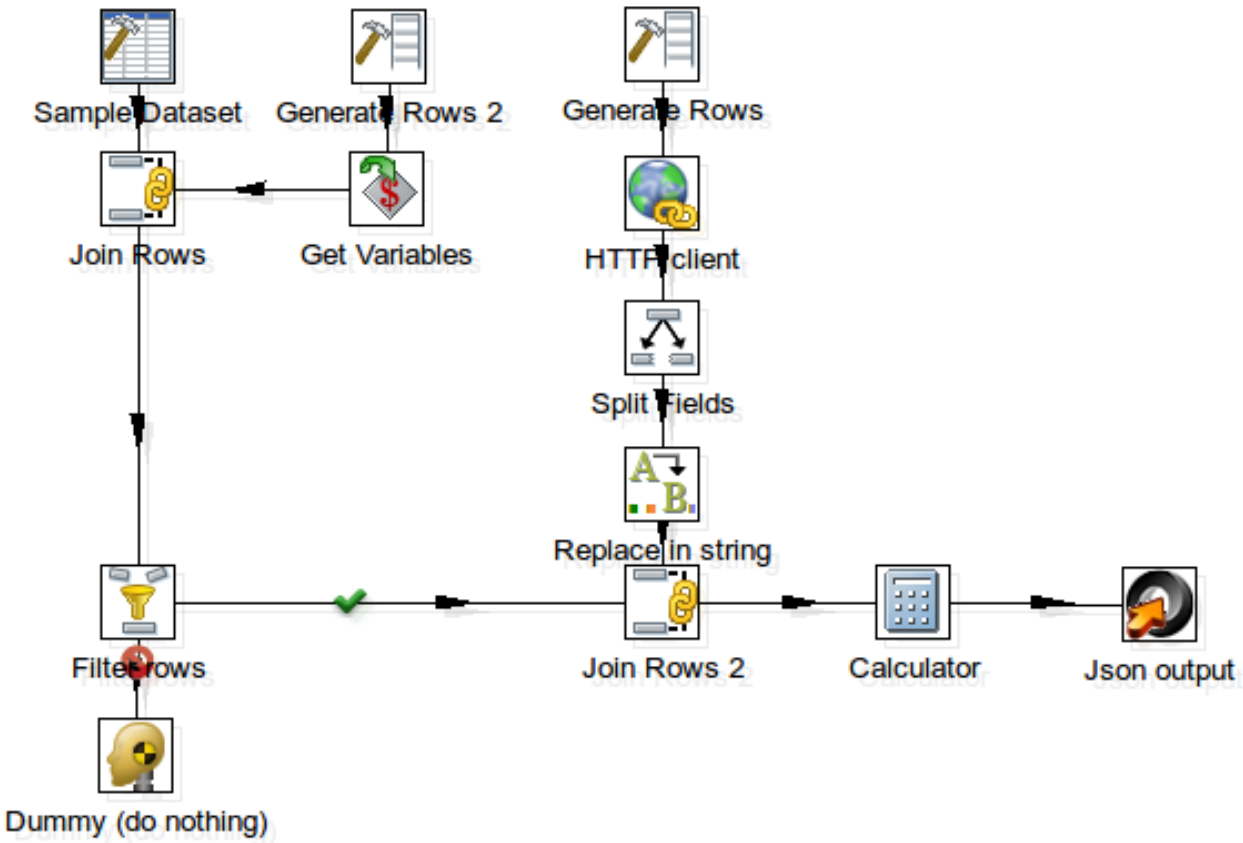
Définition

L'intégration de données regroupe les processus par lesquels les données provenant de différentes parties du système d'information sont déplacées, combinées et consolidées. Ces processus consistent habituellement à extraire des données de différentes sources (bases de données, fichiers, applications, Services Web, emails, etc.), à leur appliquer des transformations (jointures, lookups, déduplication, calculs, etc.), et à envoyer les données résultantes vers les systèmes cibles.
Source : wikiversity.org

Il existe plusieurs système d'intégration de données :

- La médiation au service de l'intégration de données d'entreprise (EII).
- L'intégration de données via les applications (EAI).
- L'intégration de données via les services Web (ESB, SOA).
- L'intégration de données en nuage (Data Cloud).
- L'ETL (Extract - Transform - Load)

Petit retour en arrière : Intégration de données



QQes mots sur le nettoyage de données

Acquisition des données

- A travers des "capteurs"
  - Capteurs réels, logiciels, wrappers
  - Nécessite une connaissance experte
- A travers l'usage de systèmes d'ETL
- A travers l'usage de systèmes d'apprentissage
  - **Apprendre à acquérir de l'information**
  - Apprendre à dialoguer
  - Apprendre à voir
  - .... ma recherche depuis une dizaine d'années.....

QQes mots sur le nettoyage de données

Acquisition des données

...

Preprocessing

- Rennomage
- Normalisation
- Discrétisation
- Abstraction
- Aggregation
- **Sélection d'attributs - Features sélection**
- Création d'attributs

QQes mots sur le nettoyage de données

Acquisition des données

...

Preprocessing

...

Biais dans les données

- Nécessité de comprendre la source des données sous peine d'obtenir des résultats 'inattendus'
- Les résultats obtenus à partir de données pré-sélectionnées sont rarement les meilleurs ! Attention à l'intuition ! !

Features Selection/Examples Selection

Features Selection

- Classification de texte : un document = un vecteur fréquentiel de mots
- Quel est le problème ?

Features Selection/Examples Selection

Features Selection

- Classification de texte : un document = un vecteur fréquentiel de mots
- Plusieurs millions de termes possibles $\Rightarrow f_\theta : \mathbb{R}^{1000000} \rightarrow \mathbb{R}$

Fléau de la dimension

Le Fléau de la dimension ou Malédiction de la dimension (Curse of dimensionality) est un terme inventé par Richard Bellman pour désigner divers phénomènes qui ont lieu lorsque l'on cherche à analyser ou organiser des données dans des espaces de grande dimension alors qu'ils n'ont pas lieu dans des espaces de dimension moindre.
Plusieurs domaines sont concernés et notamment l'apprentissage automatique, la fouille de données, les bases de données, l'analyse numérique ou encore l'échantillonnage. L'idée générale est que lorsque le nombre de dimensions augmente, le volume de l'espace croît rapidement si bien que les données se retrouvent " isolées " et deviennent éparses. Cela est problématique pour les méthodes nécessitant un nombre significatif de données pour être valides, les rendant alors peu efficaces voire inopérantes.
Source : Wikipedia

Features Selection/Examples Selection

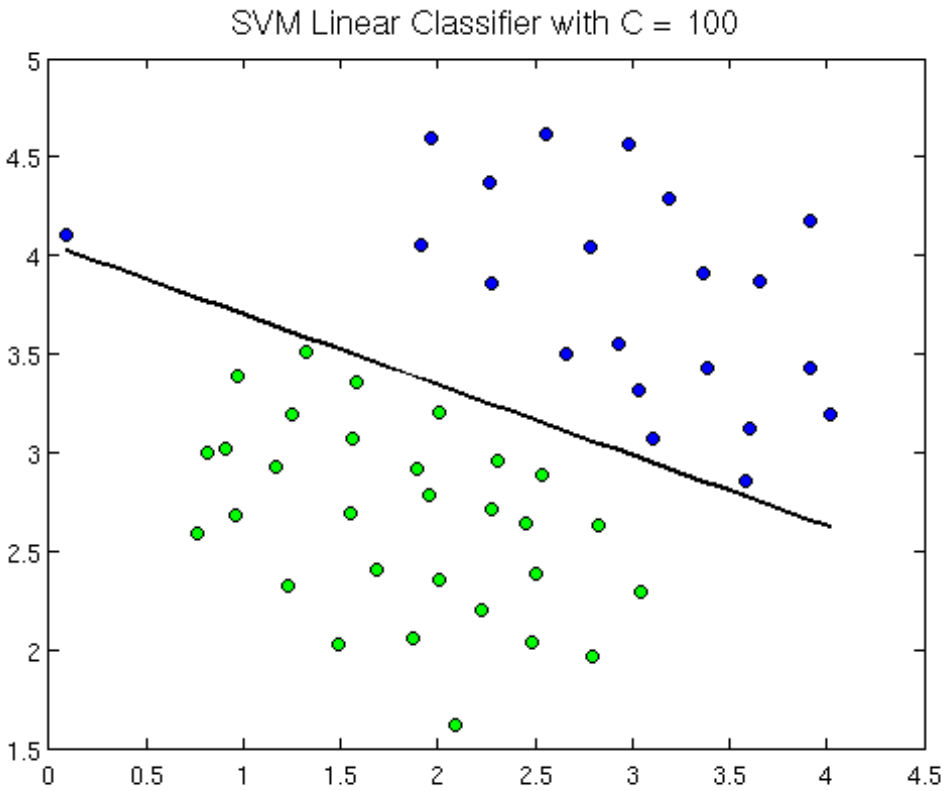
Features Selection

- Classification de texte : un document = un vecteur fréquentiel de mots
- Plusieurs millions de termes possibles $\Rightarrow f_\theta : \mathbb{R}^{1000000} \rightarrow \mathbb{R}$
- **Mais aussi : plusieurs millions de paramètres à apprendre...**

Réduire la dimension

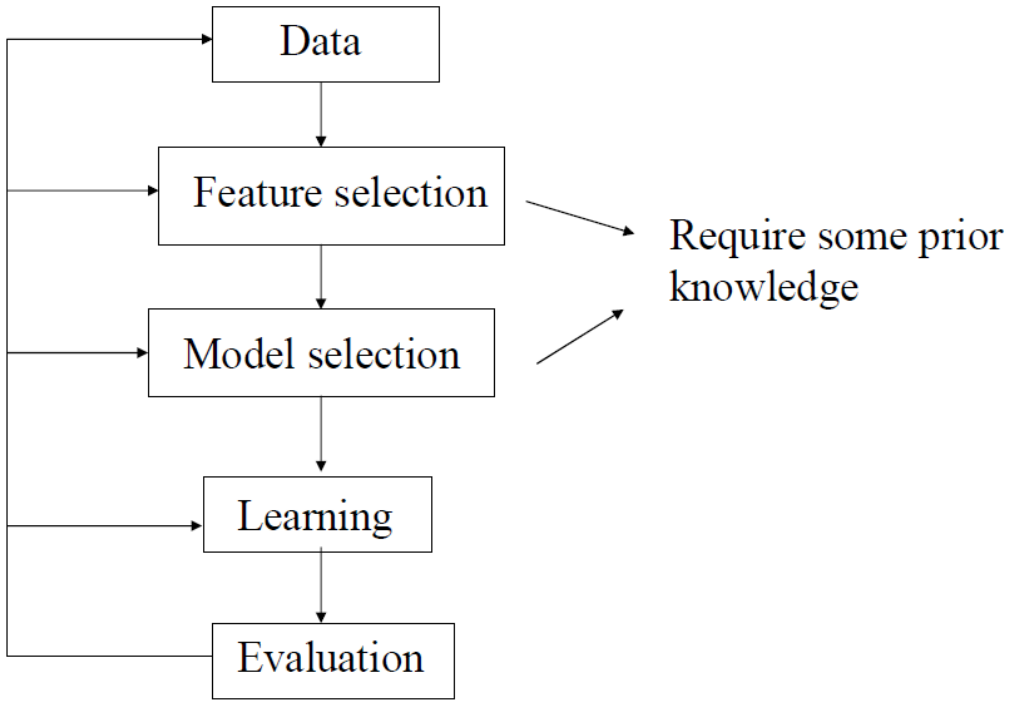
(voir autre cours)

Outliers

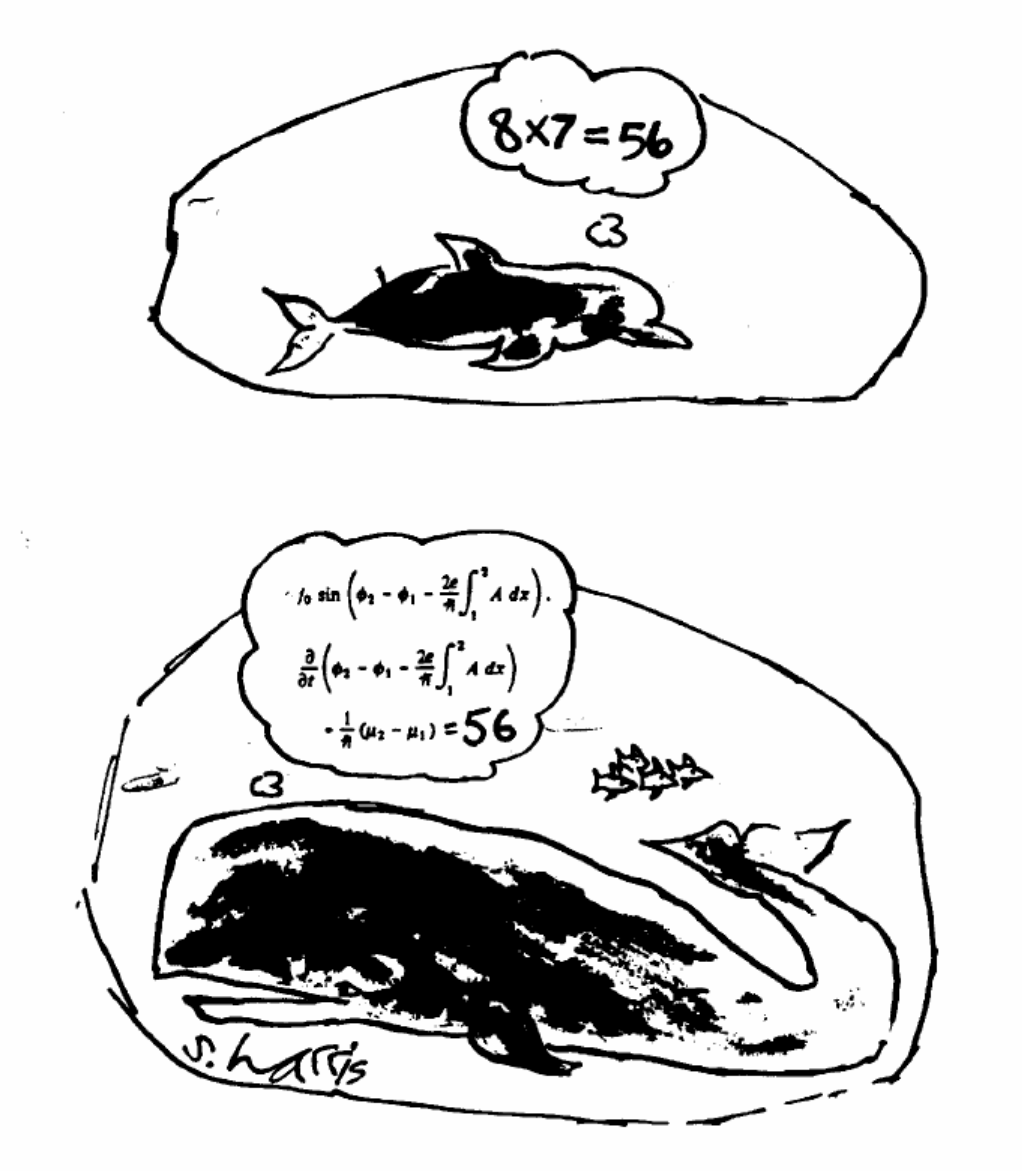


- Il faut supprimer les outliers....
- ...mais c'est pas simple $\rightarrow$ connaissances expertes.

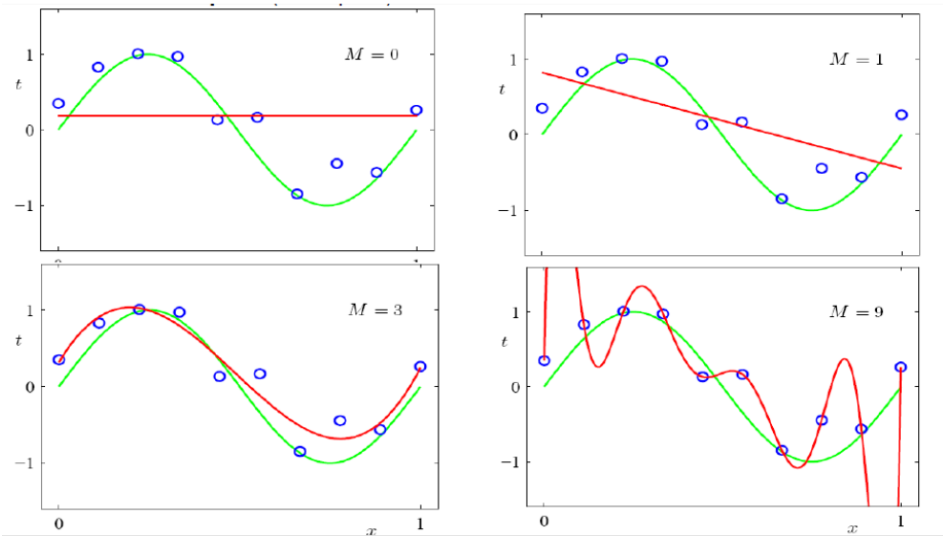
Designer un modèle



Sélection de modèles

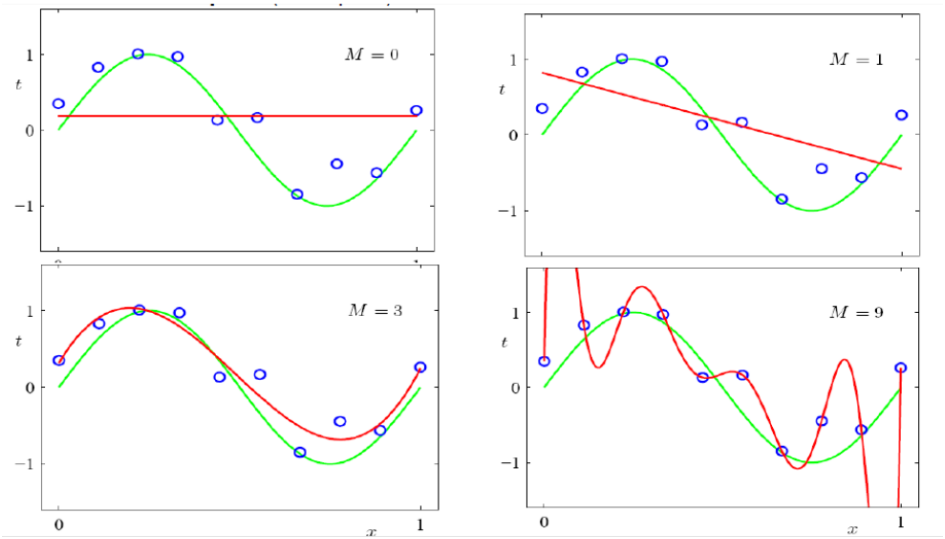


Sélection de modèles



Quel est le meilleur modèle ?

Sélection de modèles



Conclusion : On ne doit pas choisir le modèle qui correspond le mieux aux données, mais celui qui **généralise** le mieux

Sélection de Modèles

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

Bruit et Régularités

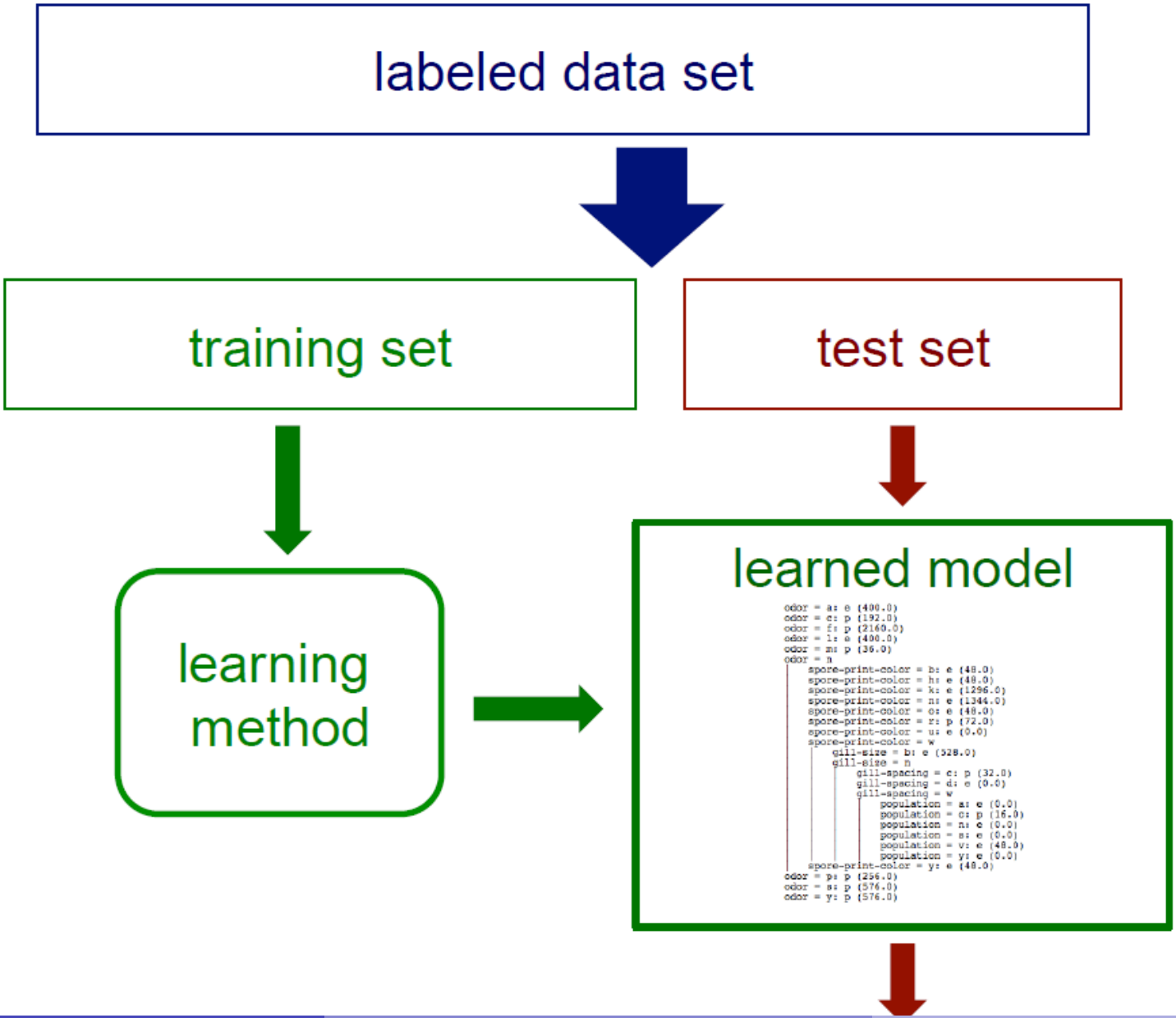
Données = **Bruit** + **Régularités**

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

Modèle final= **Capture du bruit** + **Modèle des régularités** $\Rightarrow$ on peut améliorer la "qualité" d'un modèle uniquement en augmentant sa capacité à capturer le bruit

Sur-apprentissage / Overfitting

Sélection par échantillonage



Sélection de modèles par échantillonage

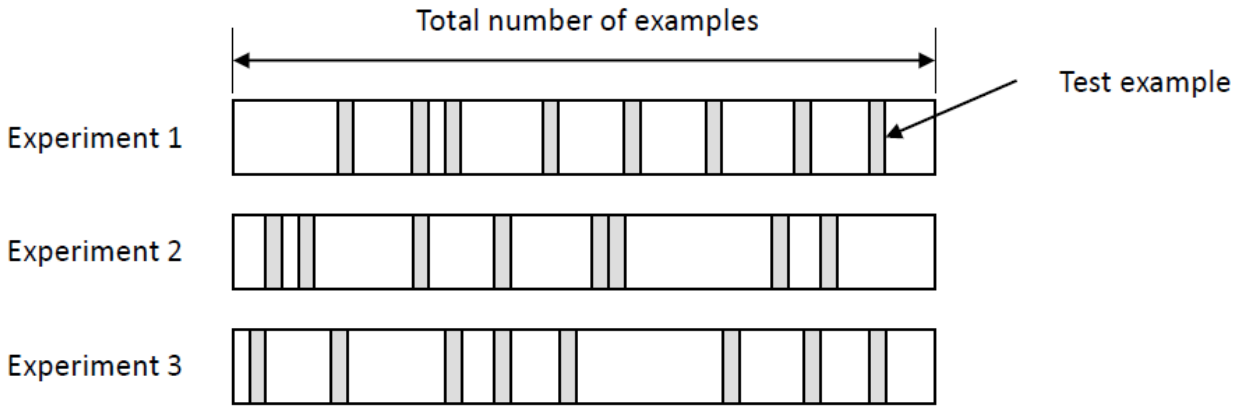
Problèmes

- As-t-on assez de données pour consituter ces différents ensembles ?
- L'utilisation d'un unique ensemble d'apprentissage ne nous permet pas de savoir si le modèle est sensibles aux données d'apprentissage

Plusieurs solutions :

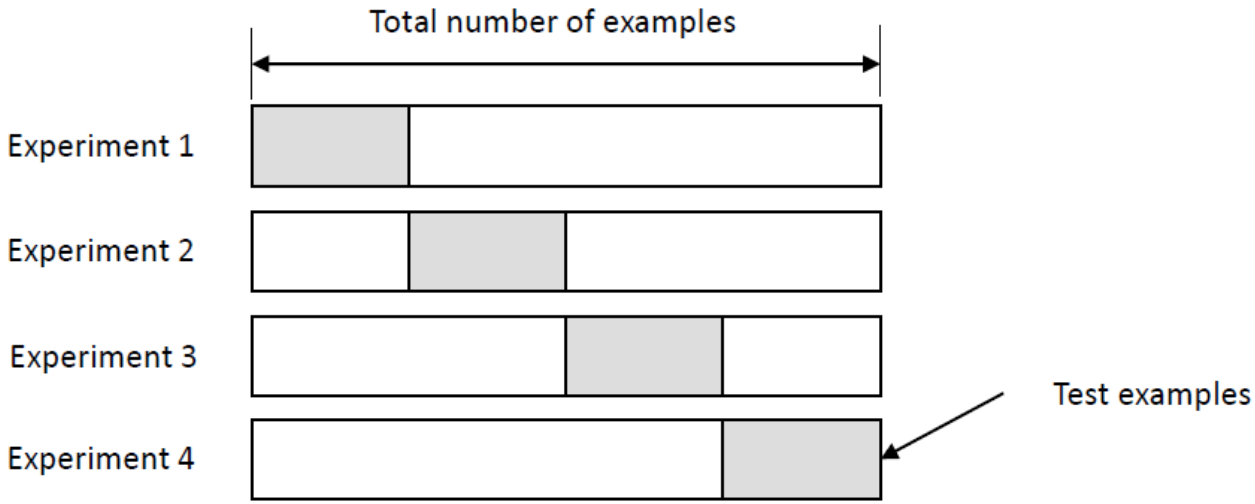
- Rééchantillonage aléatoire
- Cross-Validation

Rééchantillonage aléatoire



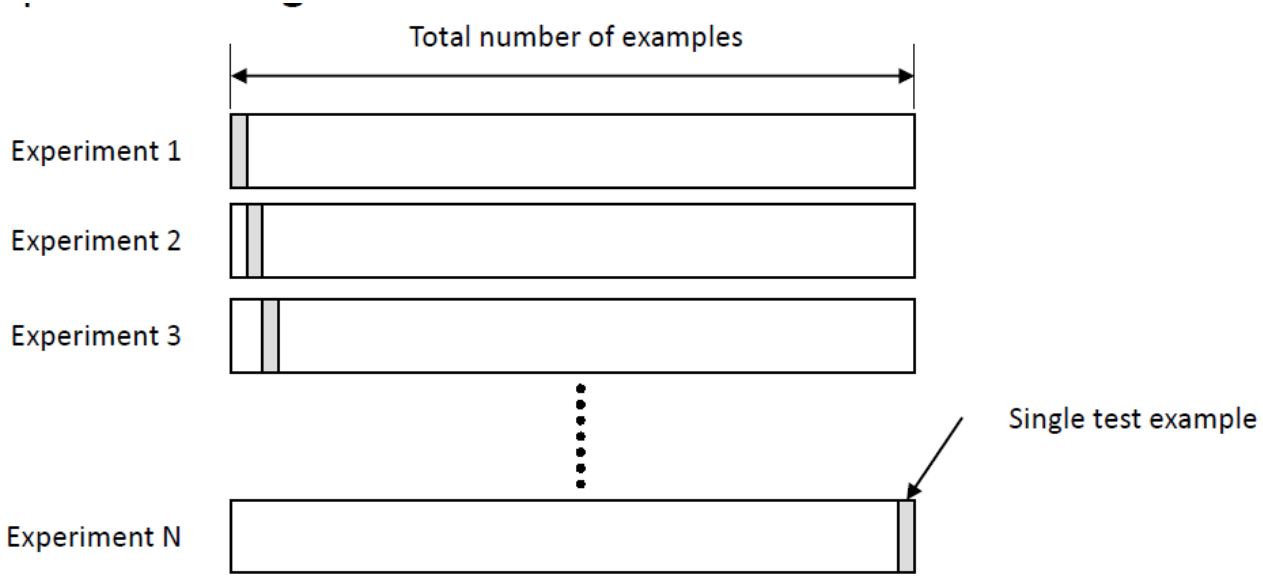
- L'estimation de l'erreur du modèle est obtenue en moyennant les erreurs obtenus sur les différentes expériences
- Cette estimation est significativement meilleure que celle obtenue précédemment, si le nombre d'expériences est suffisant

Cross-Validation



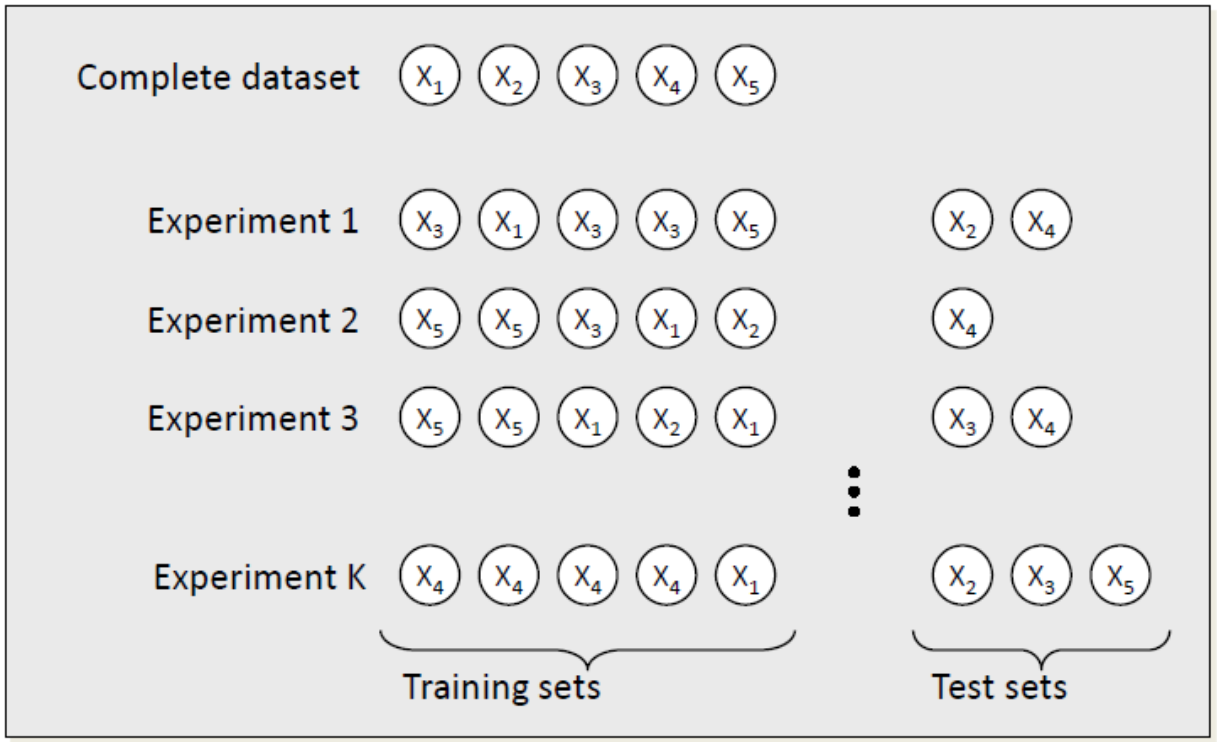
- L'estimation de l'erreur du modèle est obtenue en moyennant les erreurs obtenus sur les différentes expériences
- Tous les exemples sont utilisés pour apprendre au moins un modèle

Leave-one-out



- L'estimation de l'erreur du modèle est obtenue en moyennant les erreurs obtenus sur les différentes expériences
- Cas dégénéré de CV $\rightarrow$ plus robuste, meilleurs pour les petits jeux de données

Bootstrap



- Plus grande variance dans les différents "folds"
- Mais effet désirable car plus réaliste (c.f classification)

Train/Test/Validation

On considère le cas particulier où l'on veut **à la fois** trouver le meilleur modèle **mais aussi** estimer sa performance.

Solution

Il faut découper en trois :

- Train set
- Validation set : pour découvrir le meilleur modèle
- Test set : pour évaluer la performance

Conclusion

Protocole expérimental classique :

- Diviser les données en trois ensembles
- Entrainer un modèle sur *train*
- Evaluer le modèle sur *validation*
- Recommencer jusqu'à obtenir le meilleur modèle et les meilleurs hyper-paramètres
- Evaluer la qualité finale du modèle sur l'ensemble de test

Sources :

- CSCE 666 - Ricardo Guttierez-Osuna - CSE@TAMU
- CS2750 - Milos Hauskrecht - University of Pittsburgh
- `www.biostat.wisc.edu/~dpage/cs760/`