REsearch and methodology in Data Science

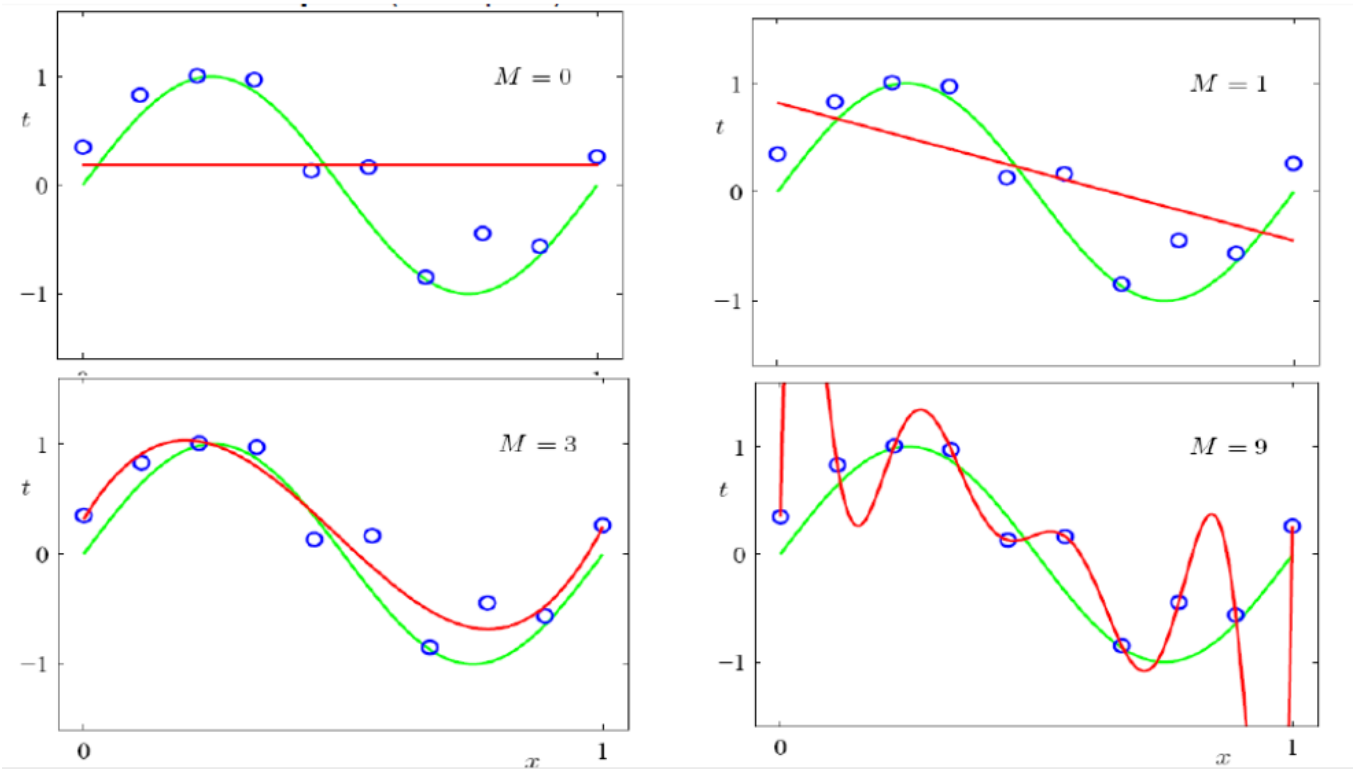
Cours 2 – Protocole expérimental

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

Master DAC
Sorbonne Université

2023-2024

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**

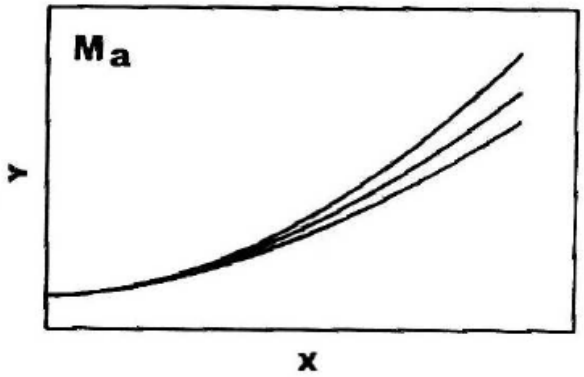
Meilleur modèle:

- Meilleur modèle des régularité
- Meilleure capture du bruit

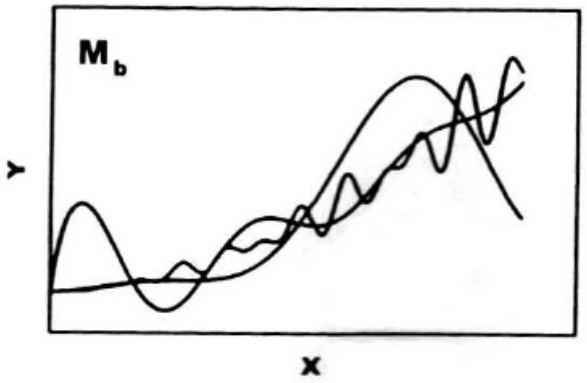
Généraliser: éviter le sur-apprentissage

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

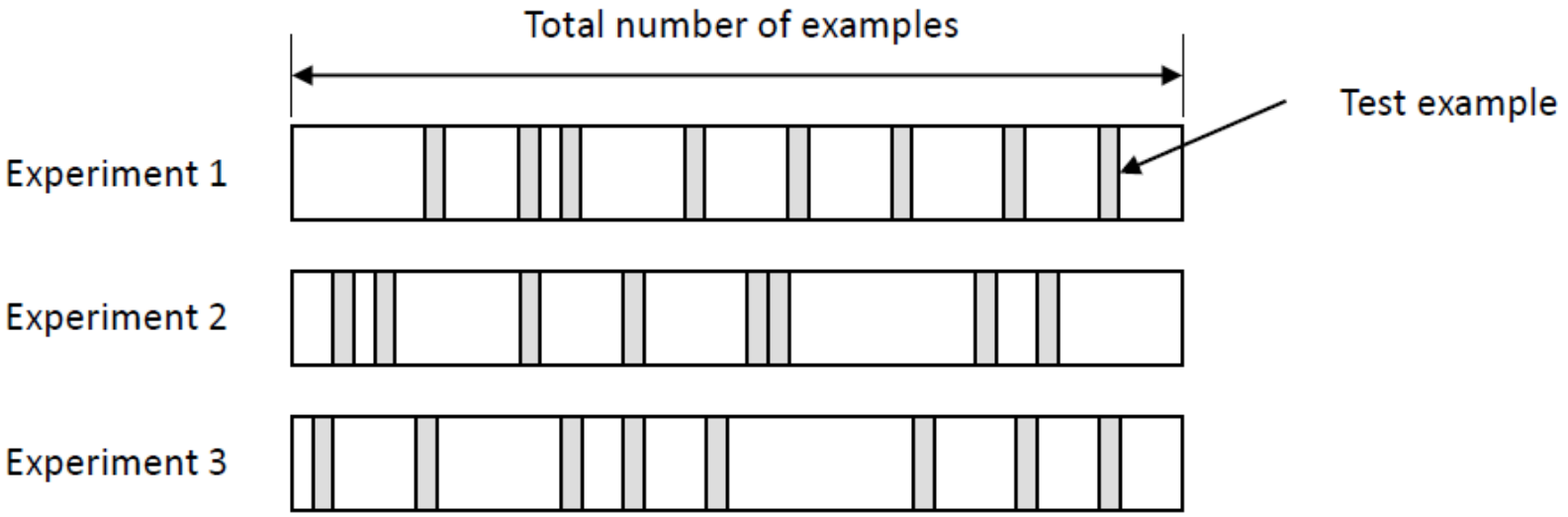
Problèmes

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

Plusieurs solutions

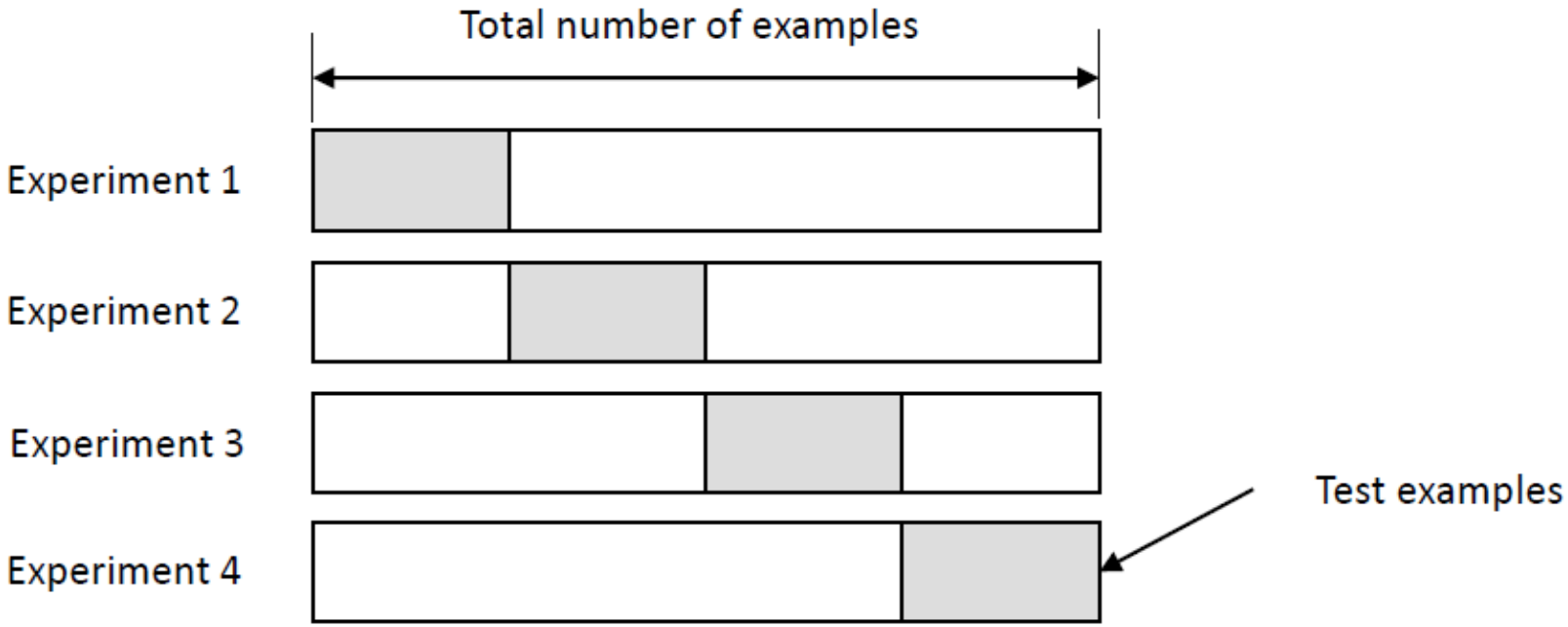
- Rééchantillonage aléatoire
- Cross-validation

Rééchantillonage aléatoire



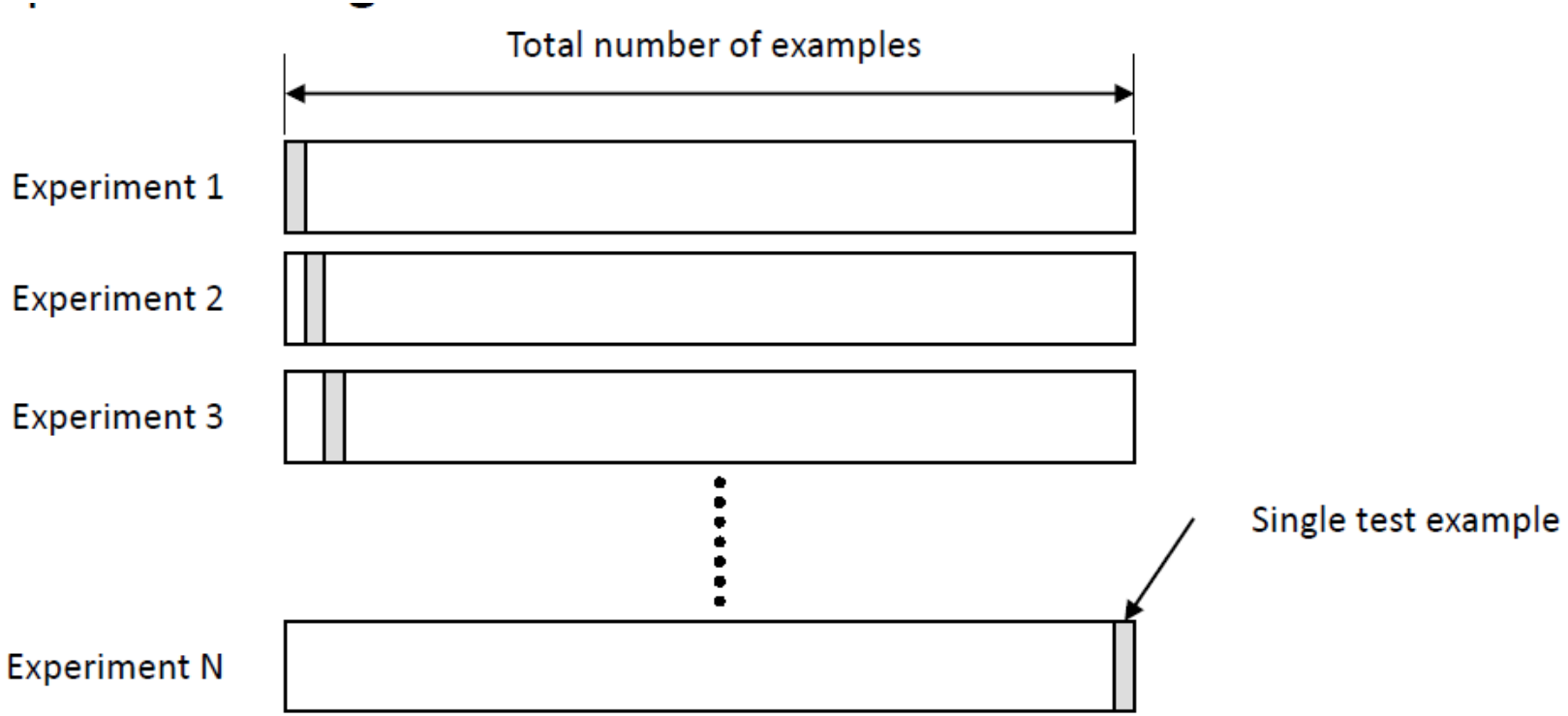
- Erreur du modèle: moyenne sur les différentes expériences
- Estimation significativement meilleure (avec assez de tirages)

Cross-Validation



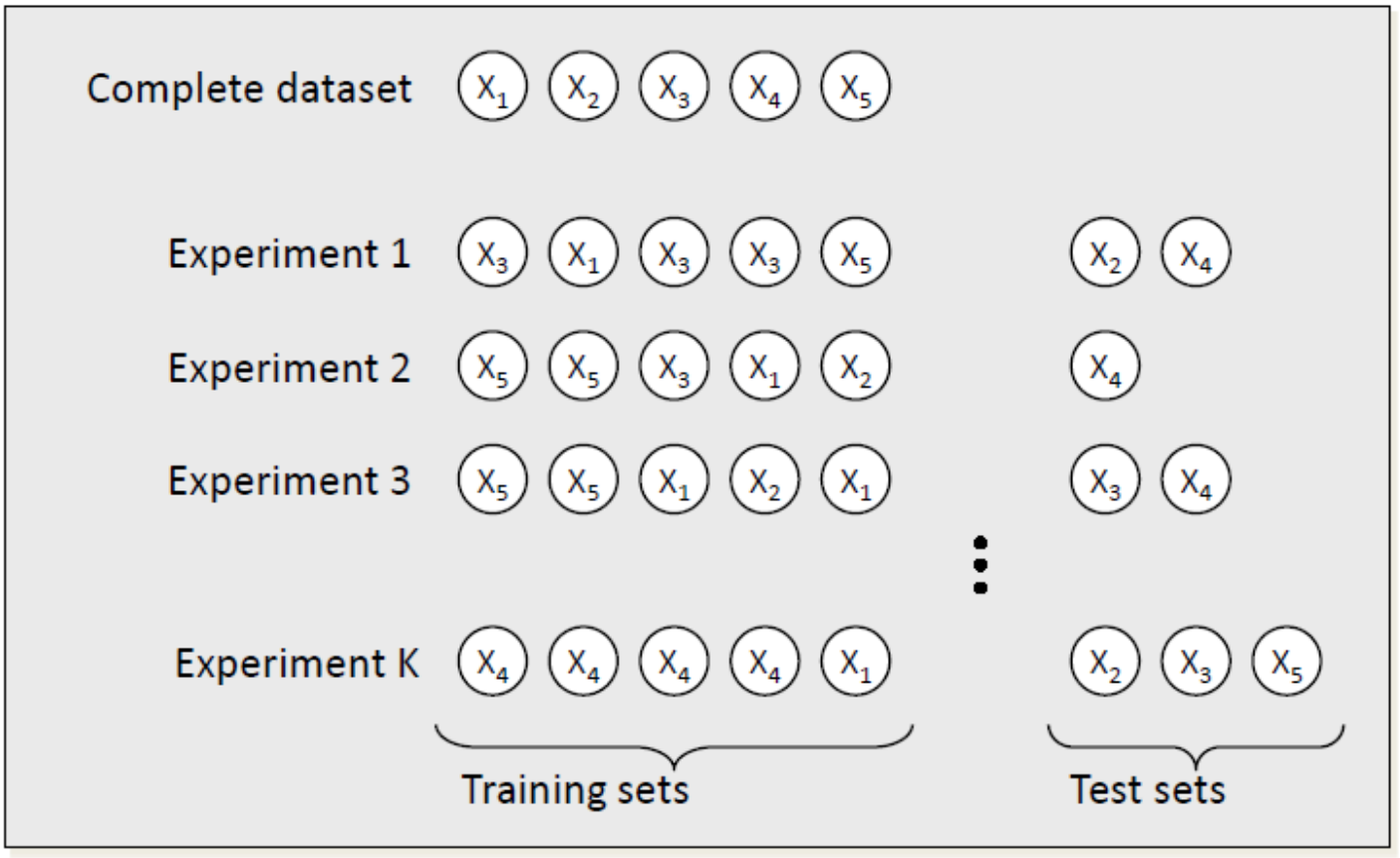
- Erreur du modèle: moyenne sur les différentes expériences
- Tous les exemples sont utilisés au moins une fois en train

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.