REsearch and methodology in Data Science

Cours 1 – Traitement de données et protocole expérimental

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

Master DAC
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

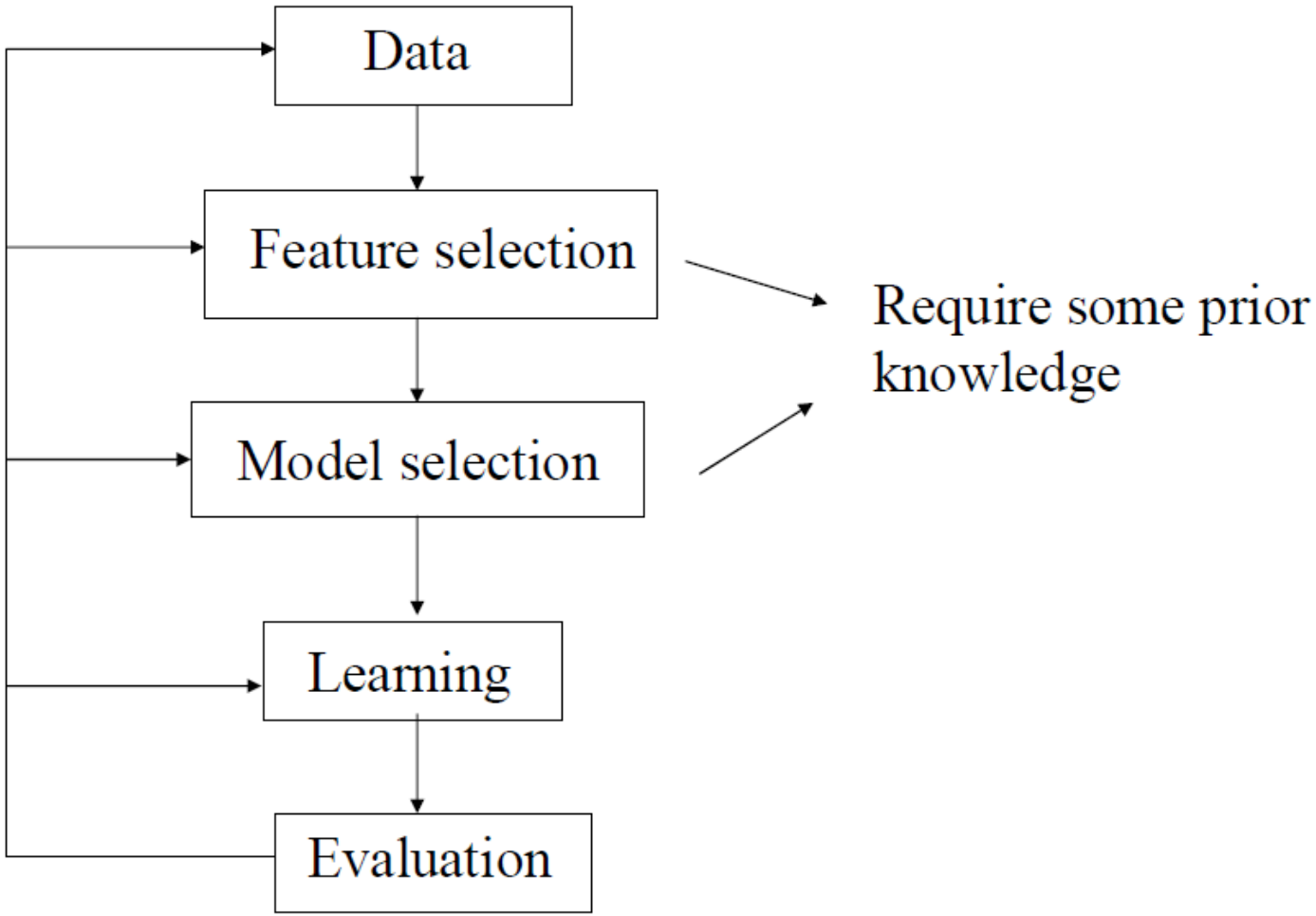
Différentes tâches

- 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

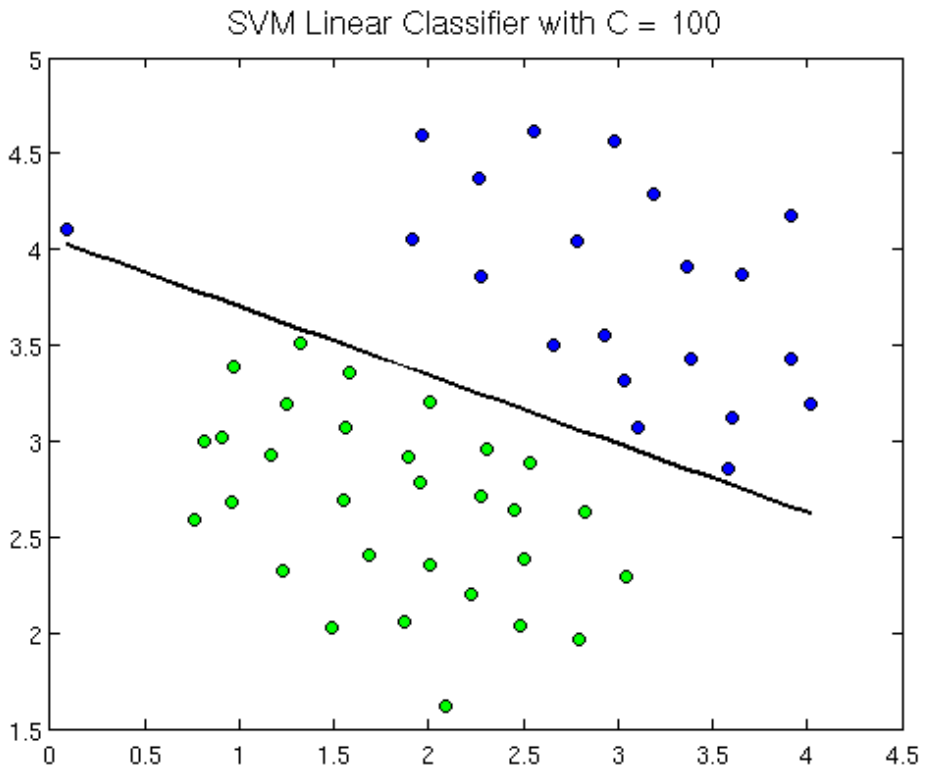
Heureusement

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

Solutions

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

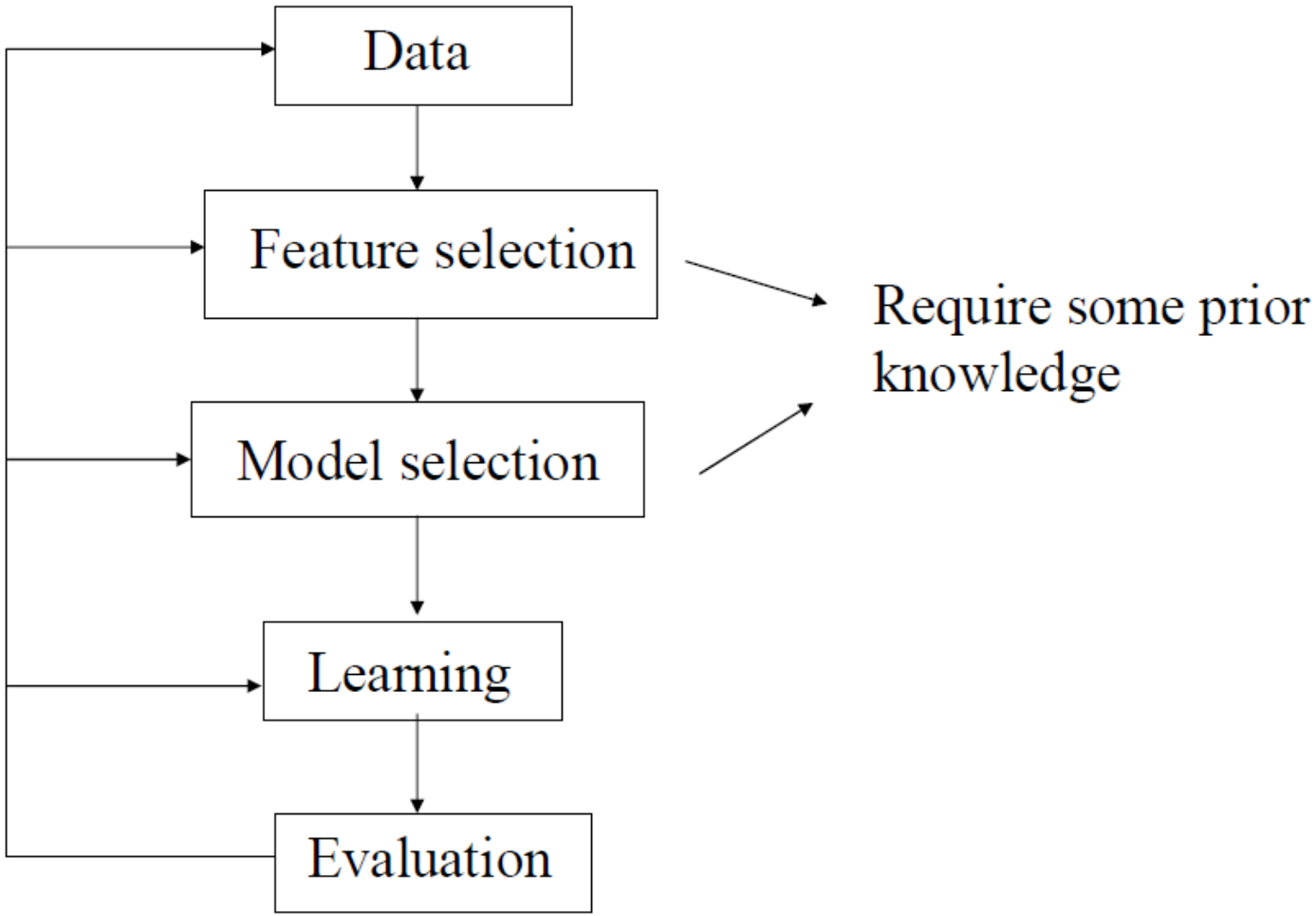
Outliers



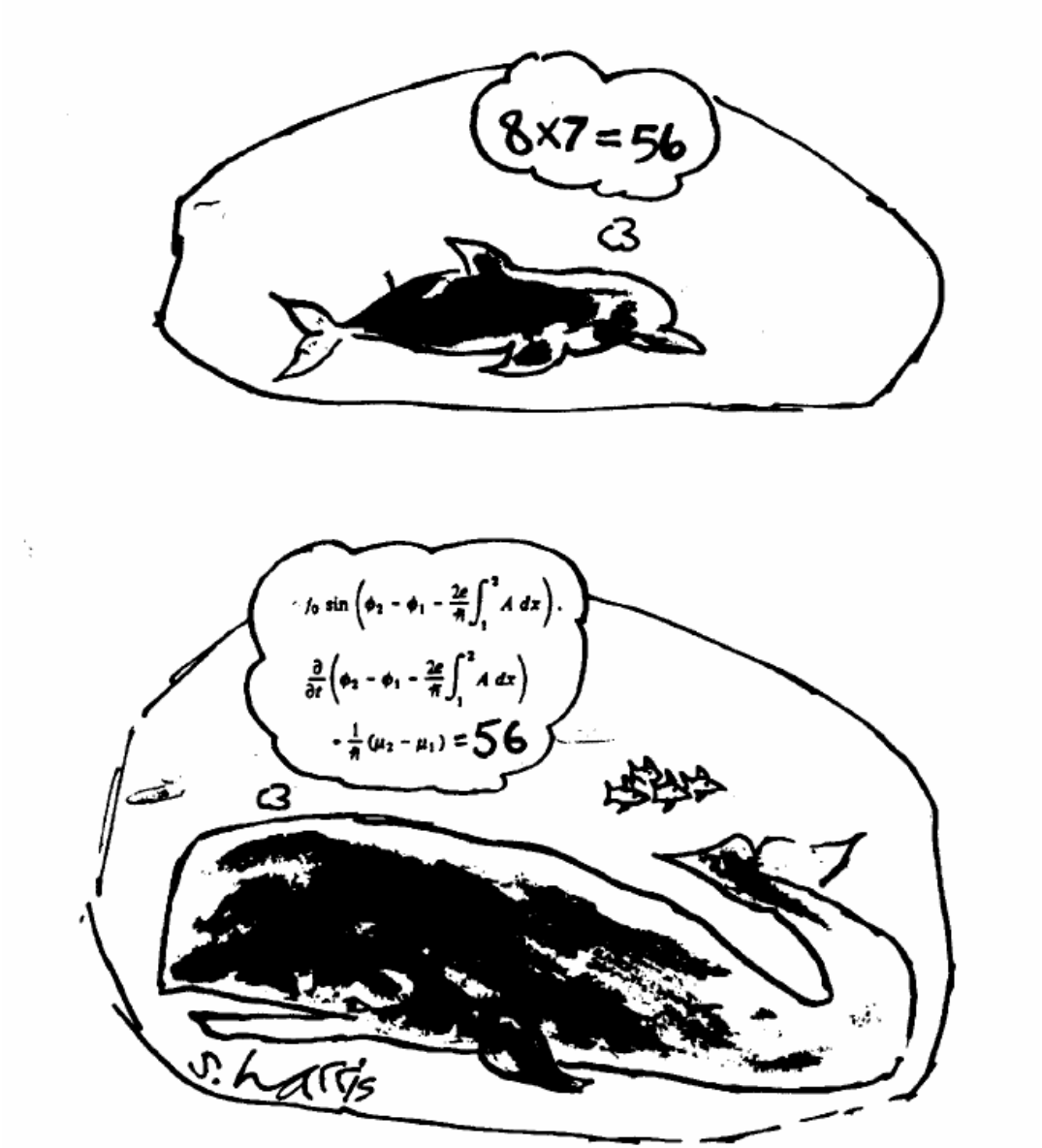
Suppresion des outliers

- Connaissances métier
- Méthodes statistique
- Difficile...

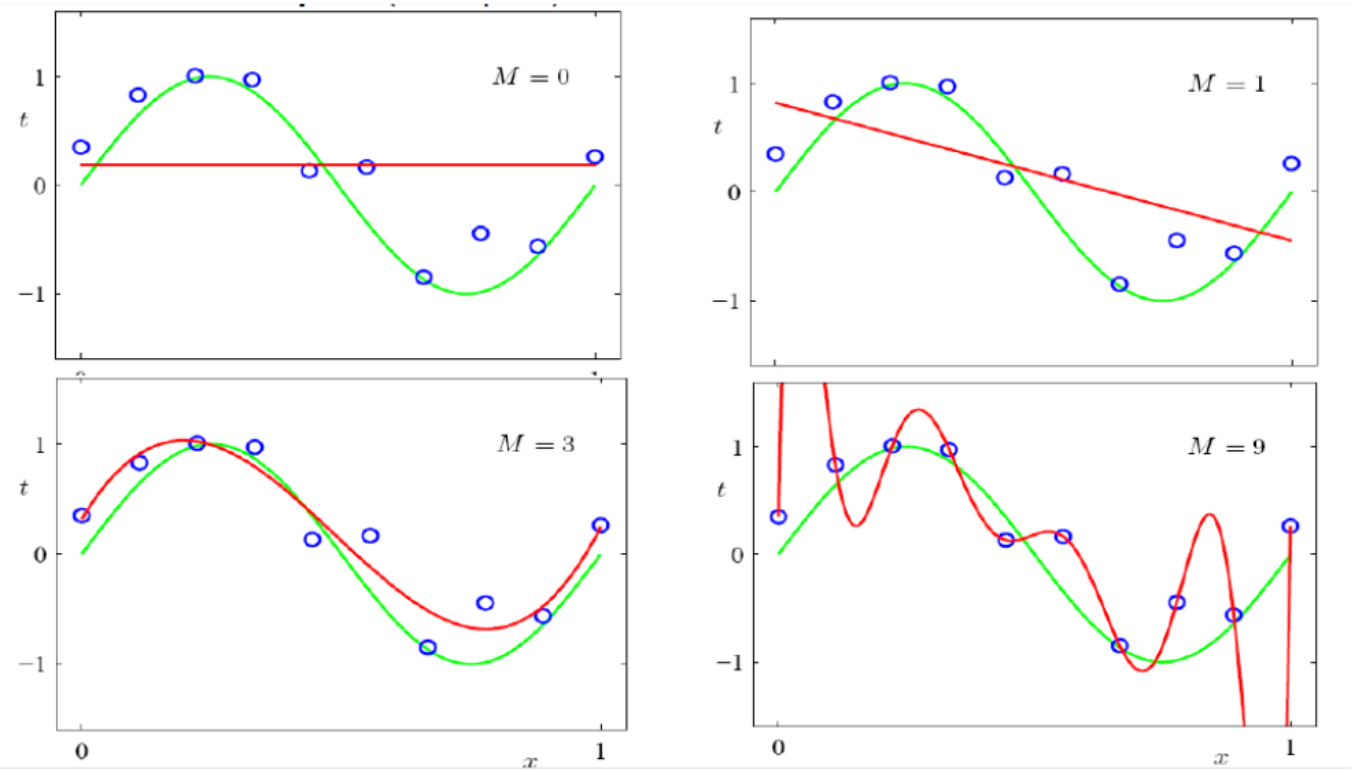
Concevoir un modèle



Sélection de modèle

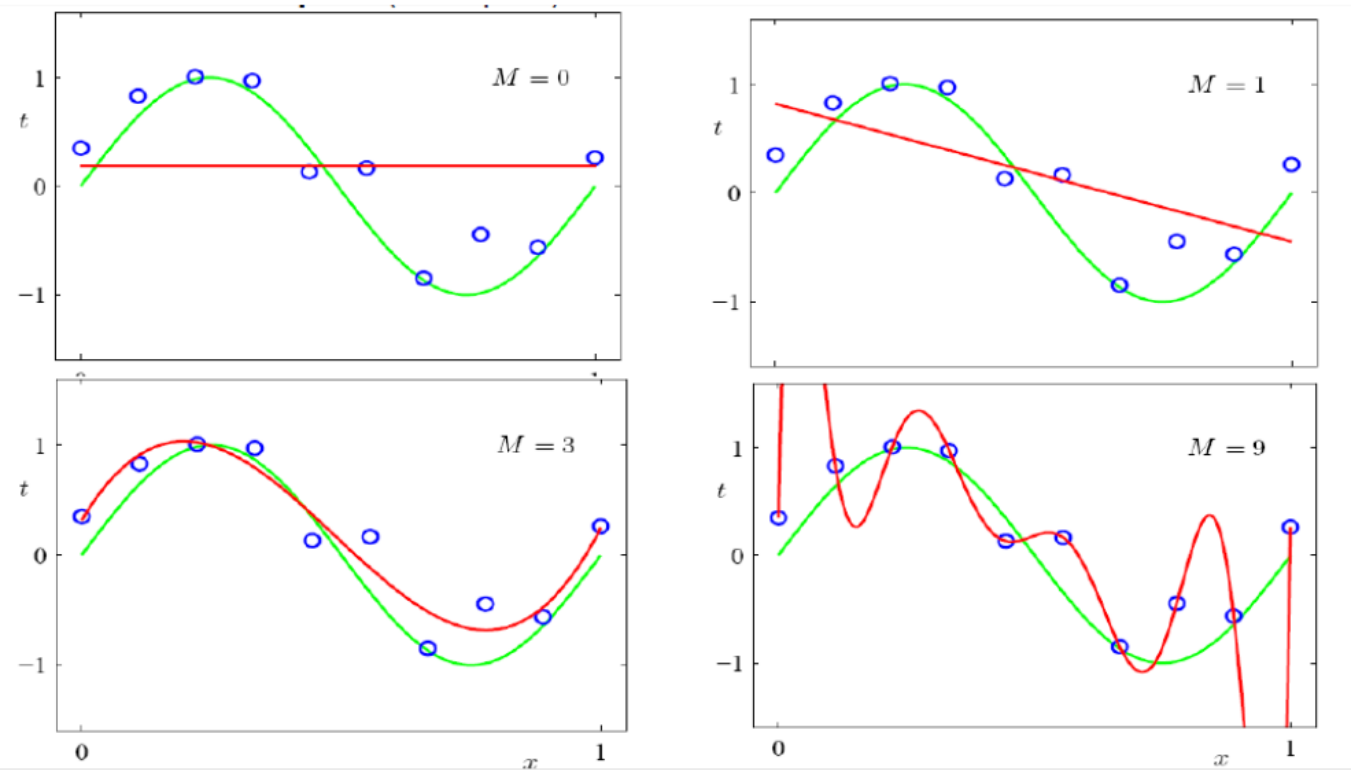


Sélection de modèle



Quel est le meilleur modèle ?

Sélection de modèle



Objectif: généralisation