

# MEthodology in Data Science

Le monde n'est pas linéairement séparable

Olivier Schwander

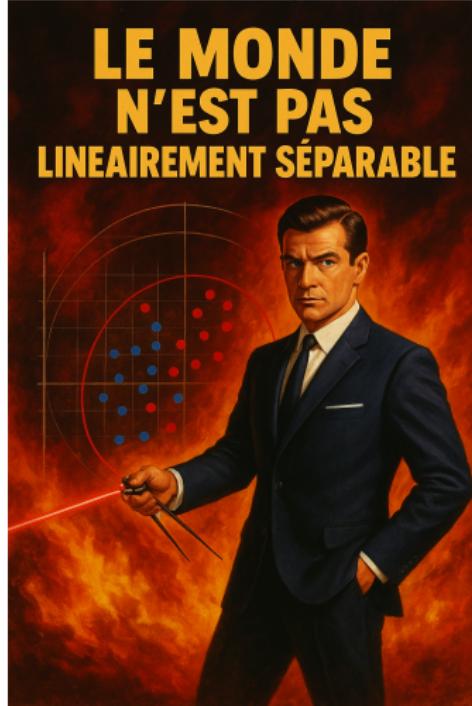
<olivier.schwander@sorbonne-universite.fr>

Master MIND  
Sorbonne Université



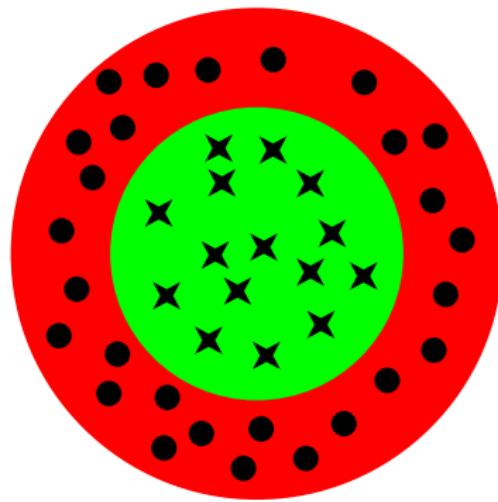
2025-2026

Le monde n'est pas linéairement séparable



[Copyright James Bond × ChatGPT]

# Le monde n'est pas linéairement séparable



## Stratégies

### Frontière de décision plus complexe

- ▶ Modèle quadratique, capable de tracer un cercle
- ▶ Quel modèle choisir ?

### Transformation intelligente des données

- ▶ Modèle linéaire mais
- ▶  $(x, y) \rightarrow \rho = x^2 + y^2$
- ▶ Comment la trouver ?

### Transformation mécanique

- ▶ Modèle linéaire mais
- ▶  $(x, y) \rightarrow (x, y, x^2, y^2)$
- ▶ Où s'arrêter ?

# Points de vue

## Différentes stratégies

- ▶ Travail sur le modèle
- ▶ Travail manuel sur les données
- ▶ Travail automatique sur les données

## Au final

- ▶ Frontière  $ax + by + cx^2 + dy^2 = 0$
- ▶ Éventuellement: on fixe  $a = b = 0$  et  $c = d$

## Objectif

- ▶ Trouver  $a = b = 0$  et  $c = d$
- ▶ Sans être intelligent
- ▶ Comprendre la tâche, à partir des données

## Connaissances a priori

### Expert · e · s du domaine

- ▶ Telle variable est utile pour la prédiction
- ▶ Telle variable est inutile, remplie d'erreur, etc
- ▶ Telle et telle variable il faut les multiplier, etc

### Difficultés

- ▶ Souvent critique pour faire marcher une application
- ▶ Besoin de cette expertise
- ▶ Besoin d'expliquer ce qu'on veut aux experts

### Souvent pas suffisant

- ▶ Linguistes et le traitement automatique des langues

## Bonne représentation

### Objectifs

- ▶ Petite dimension
- ▶ Éviter les redondances
- ▶ Décrire les données
- ▶ Utile pour la tâche

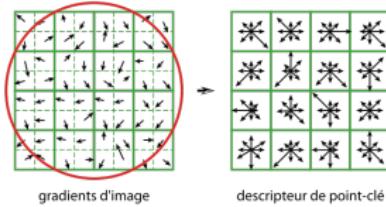
### Bonne représentation

- ▶ Smooth: un petit changement dans l'espace d'origine donne un petit changement dans l'espace d'arrivée
- ▶ Invariance: la variabilité intra-classe ne change pas trop la représentation
- ▶ Erreur de reconstruction ? Retrouver les points d'origine

## Vision avant le deep

### Descripteurs SIFT (Lowe 2004)

- ▶ Points-clés: recherche multi-échelle de points d'intérêt
- ▶ Description locale selon un mécanisme construit à la main



### Scattering transform (Mallat 2010)

- ▶ Version courte: CNN mais avec des filtres choisis à la main
- ▶ Mallat et al 2010 et suivantes

## Points communs

- ▶ Filtres

# Sélection de variables

## Sélection de caractéristiques

- ▶ Sous-ensemble des variables
- ▶ Besoin d'une heuristique: trop de sous-ensembles

## Extraction de caractéristiques

- ▶ Combinaison des variables
- ▶ Analyse en composante principale (éventuellement avec noyau)
- ▶ Auto-encodeurs
- ▶ Descripteurs d'images
- ▶ Spectrogrammes et autres représentations temps-fréquence

# Méthodes de filtrage

## Sélection *a priori*

- ▶ Estimation du pouvoir prédictif de chaque caractéristique
- ▶ Sélection de celles avec le pouvoir prédictif le plus élevé
- ▶ Rapide

## Pouvoir prédictif

- ▶ Dépendance linéaire: corrélation entre entrée et sortie
- ▶ Non-linéarités et combinaisons de variables ?
- ▶ Indirectement guidé par la tâche

# Méthodes de wrappers

## Sélection *a posteriori*

- ▶ Choix basé sur la qualité du modèle obtenu
- ▶ Guidé par la tâche, même protocole d'évaluation

## Recherche gloutonne

- ▶ Départ avec un petit sous-ensemble
- ▶ Ajout graduel de caractéristiques
- ▶ Décider si l'ajout est pertinent en fonction du score

# Analyse en composantes principales

[Dessin au tableau: nuage de points et ellipse]

## Version statistique

- ▶ Recherche des axes principaux qui maximisent la variance

## Version reconstruction

- ▶ Minimiser l'erreur de reconstruction

$$\arg \max \sum_i \|x_i - \hat{x}_i\|_2^2$$

- ▶ Points projetés  $\hat{x} \in \mathbb{R}^d$ , avec  $d$  petit et choisi à l'avance

# Objectif de l'ACP

[Dessins au tableau: reconstruction d'une image]

Pas le même objectif que l'apprentissage

- ▶ Apprentissage: erreur de classification (ou autre)
- ▶ Ici, erreur de reconstruction: pas forcément très intéressant

## Limites

- ▶ Impact sur les performances à évaluer, pas évident
- ▶ Limité par la linéarité: variantes non linéaires
- ▶ De temps en temps intéressant pour la visualisation (mais il y a mieux !)

# Boosting

## Classifieur faible

- ▶ Performance strictement supérieure à un classifieur aléatoire
- ▶ Pas besoin de plus

## Boosting

- ▶ Terme générique pour la combinaison de classifieur faibles
- ▶ Combinaison: classifieur très performant

## Idée

- ▶ Apprentissage successif de modèles
- ▶ Pondération des exemples d'apprentissage:
  - ▶ Points bien prédits  $\Rightarrow$  poids faible
  - ▶ Points mal prédits  $\Rightarrow$  poids fort
- ▶ Focalisation sur les parties de l'espace mal prédits.

# Interprétation du boosting

## Règle de décision

$$H = \text{signe} \left( \sum_{t=1}^T \alpha_t h_t(x) \right)$$

## Géométriquement

- ▶ Hyperplan en dimension T

## Bonne représentation

- ▶ Classifieur simple suffisant
- ▶ Choix de la représentation guidée par la tâche

# Réseaux convolutionnels

## Première interprétation

- ▶ Grosse fonction très non-linéaire

## Deuxième interprétation: deux blocs

- ▶ Début: convolutions
- ▶ Fin: couches denses

## Convolutions

- ▶ Transformations des données
- ▶ Représentation guidée par la tâche

## Couches denses

- ▶ Perceptron multi-couches
- ▶ Classifieur simple
- ▶ Mais sur les données transformées

# Réseaux convolutionnels

[Dessin CNN=Conv+MLP au tableau]

# Approche end-to-end

## Données d'entrée

- ▶ Image brute: tas de pixel
- ▶ Révolutionnaire à l'époque
- ▶ Pas besoin de pré-traitement

## Inconvénient

- ▶ Besoin de beaucoup de données
- ▶ Très coûteux
- ▶ Refaire sans cesse un entraînement similaire

## Fine-tuning

### Sortie de la partie convulsive

- ▶ Très bonne représentation: un MLP suffit (voire juste une couche)
- ▶ Générique (en fait c'est pas si éloigné de SIFT)

### Juste apprendre un nouveau classifieur

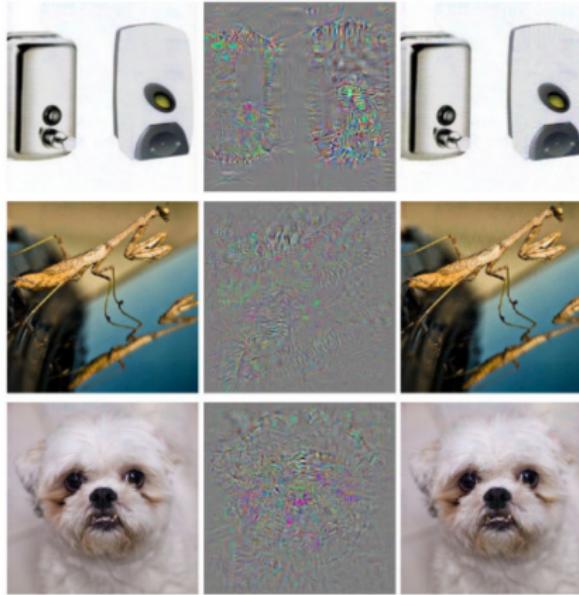
- ▶ Un autre MLP
- ▶ Ou n'importe quoi

### Guidé par une tâche similaire

- ▶ Domaines différents (images naturelles ou médicales)
- ▶ Tâche différente (classification ou segmentation)
- ▶ Représentation pré-entraînée

# Convolutions et smoothness

Contre-exemple: attaques adverses



# Modèles de mots

## Word2vec & co

- ▶ Non-supervisé
- ▶ Entraînement: prédiction d'un mot à partir de son contexte

## Pré-entraîné

- ▶ Indépendamment de la tâche
- ▶ Bonne représentation: capture la sémantique des mots

[Dessin au tableau: king - man + woman = queen]

# Modèles de langue

## Pré-entraîné

- ▶ Auto-supervisé: prédiction du mot suivant ou du mot masqué
- ▶ Pré-entraîné indépendamment de la tâche

## Chatbot

- ▶ SFT: Supervised fine-tuning avec des dialogues
- ▶ RLHF: Reinforcement Learning with Human Feedback

## Tâches précises

- ▶ Prompt sans entraînement
- ▶ Zero/few-shots

# Multimodal

## Cohérence sémantique

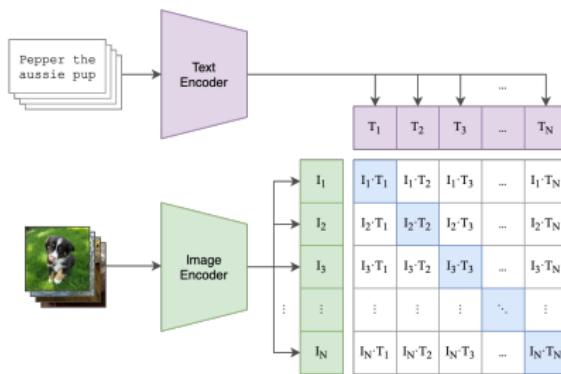
- ▶ Proximité (texte, texte) (traduction)
- ▶ Proximité (texte, image) (vision-langage)

## Apprentissage contrastif

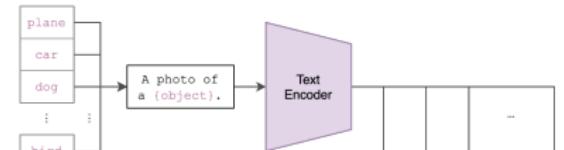
- ▶ Exemples négatifs: paires sémantiquement différentes (nombreuses !)
- ▶ Exemples positifs: paires sémantiquement proches (plus rares, souvent besoin d'augmentation)

# CLIP: Contrastive Language-Image Pre-training

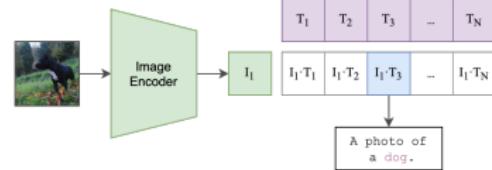
(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction



## VLM: Vision-Language Models

- ▶ Exploiter l'information textuelle pour mieux comprendre l'autre modalité
- ▶ Zero/few-shots et prompting

# Vers les modèles de fondation

	Supervision	Guidé par la tâche	État de l'art
Sélection	oui	Pas vraiment	Non
PCA	non	Non	Non
Boosting	oui	Oui	Souvent
CNN	oui	Oui	Non
Fine-tuning	oui	Oui	Souvent
Word2vec	non	Non	Non
LLM	auto	Non	Oui
CLIP	contrastif	Pas vraiment	Oui

# Conclusion

## Représentation

- ▶ Capital: peu importe la tâche, les données, les modèles
- ▶ Apprise ou pas

## Modèles de fondation

- ▶ Base des projets d'IA moderne
- ▶ Académique ou industriel

## Adaptation à une tâche

- ▶ Prompting, zero/few-shots
- ▶ Fine-tuning supervisé
- ▶ Renforcement et préférences humaines
- ▶ LORA: LOw-Rank Adaptation (diminution du nombre de paramètres à mettre à jour)