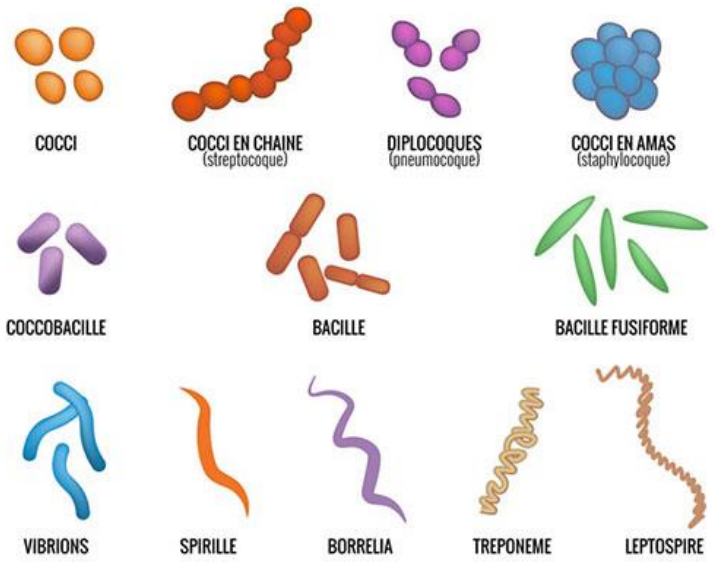
Data Challenge

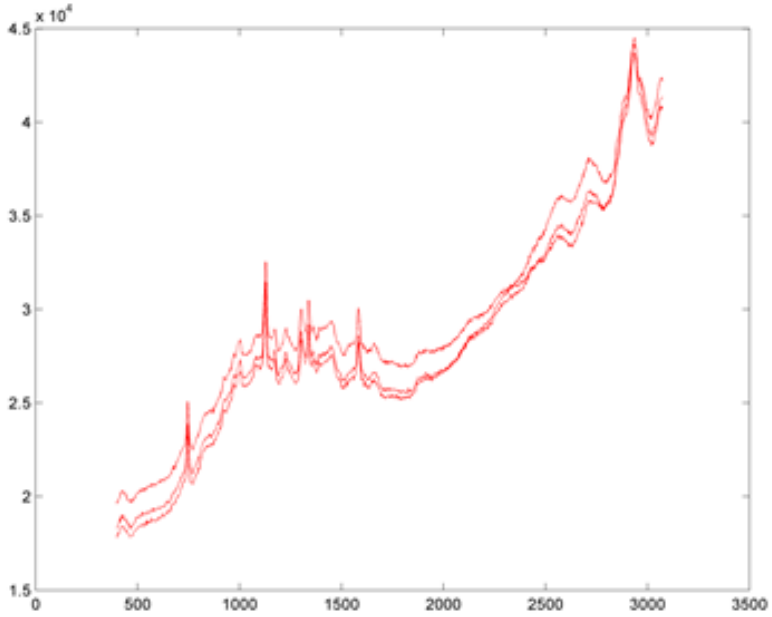
UE Apprentissage-Stat II – M2 SSD

17/01/2022

Objectif

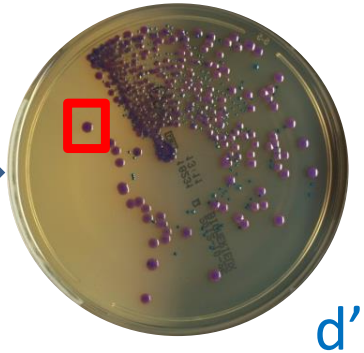
- <u>Identification bactérienne et spectroscopie</u> : reconnaitre l'espèce d'une souche bactérienne à partir d'un spectre (RAMAN).



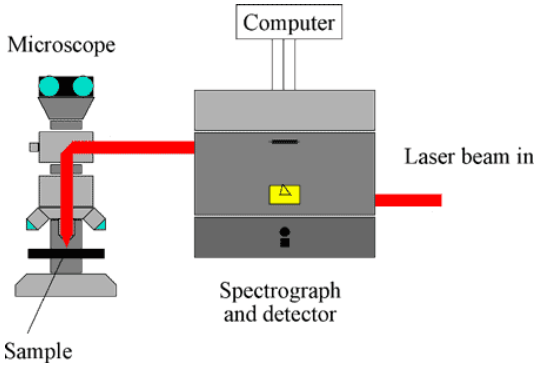


Comment ça marche ?

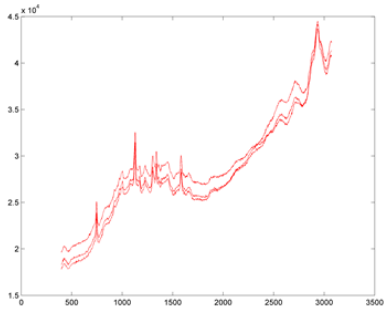
mise en culture



sélection d'une colonie



acquisition d'un spectre



algorithme de classification

Jeu de données disponible

- Apprentissage :
  - 346 souches bactériennes (les individus) provenant de 42 espèces bactériennes (les classes)
  - 9 spectres acquis par souche → 3114 spectres

- Test :
  - 133 nouvelles souches
  - 9 spectres acquis par souche → 1197 spectres

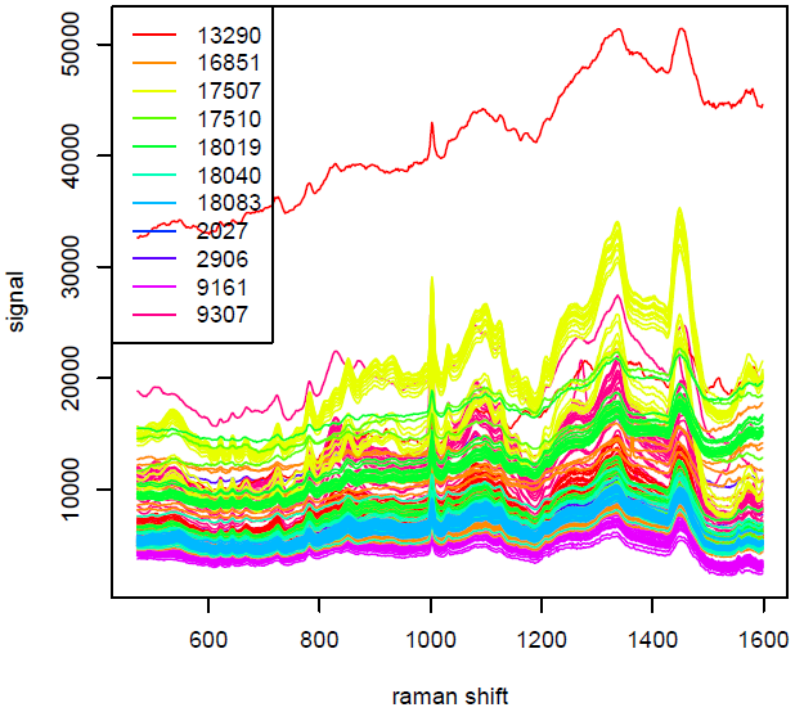
Jeu de données disponible

- Apprentissage :
  - Fichier spectra-train.csv : 3114 lignes x 627 colonnes
  - Fichier meta-train.csv : 3115 lignes x 2 colonnes
    - Colonne no1 : espèces (dans {sp_1, …, sp_42})
    - Colonne no2 : souche (entre 1 et 346)
- Test :
  - Fichier spectra-test.csv : 1197 x 627 colonnes
  - Fichier meta-test.csv : 1198 lignes x 1 colonne
    - Colonne no1 : souche (entre 1 et 133)

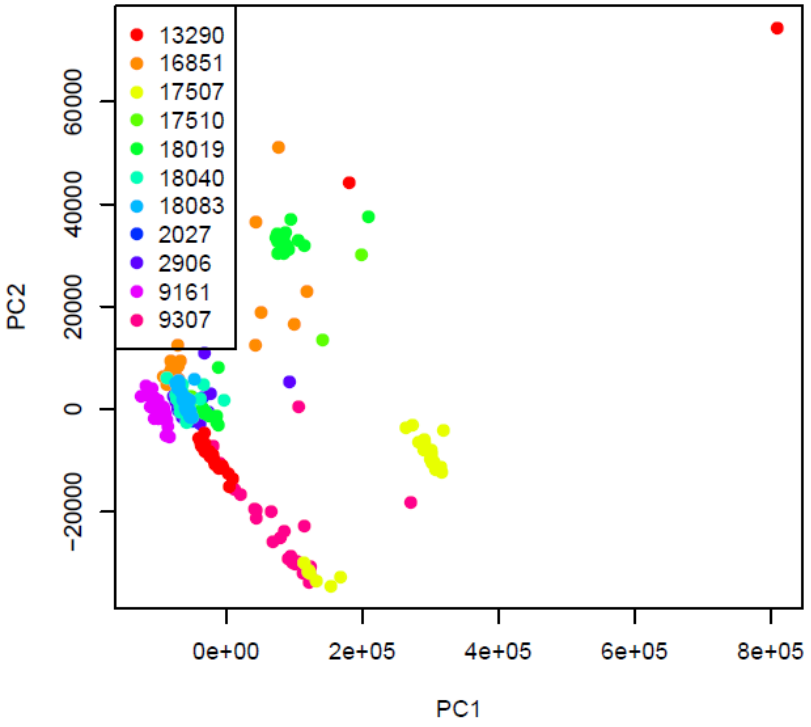
Jeu de données disponible

- Illustration : 1 espèce, plusieurs souches

spectres bruts

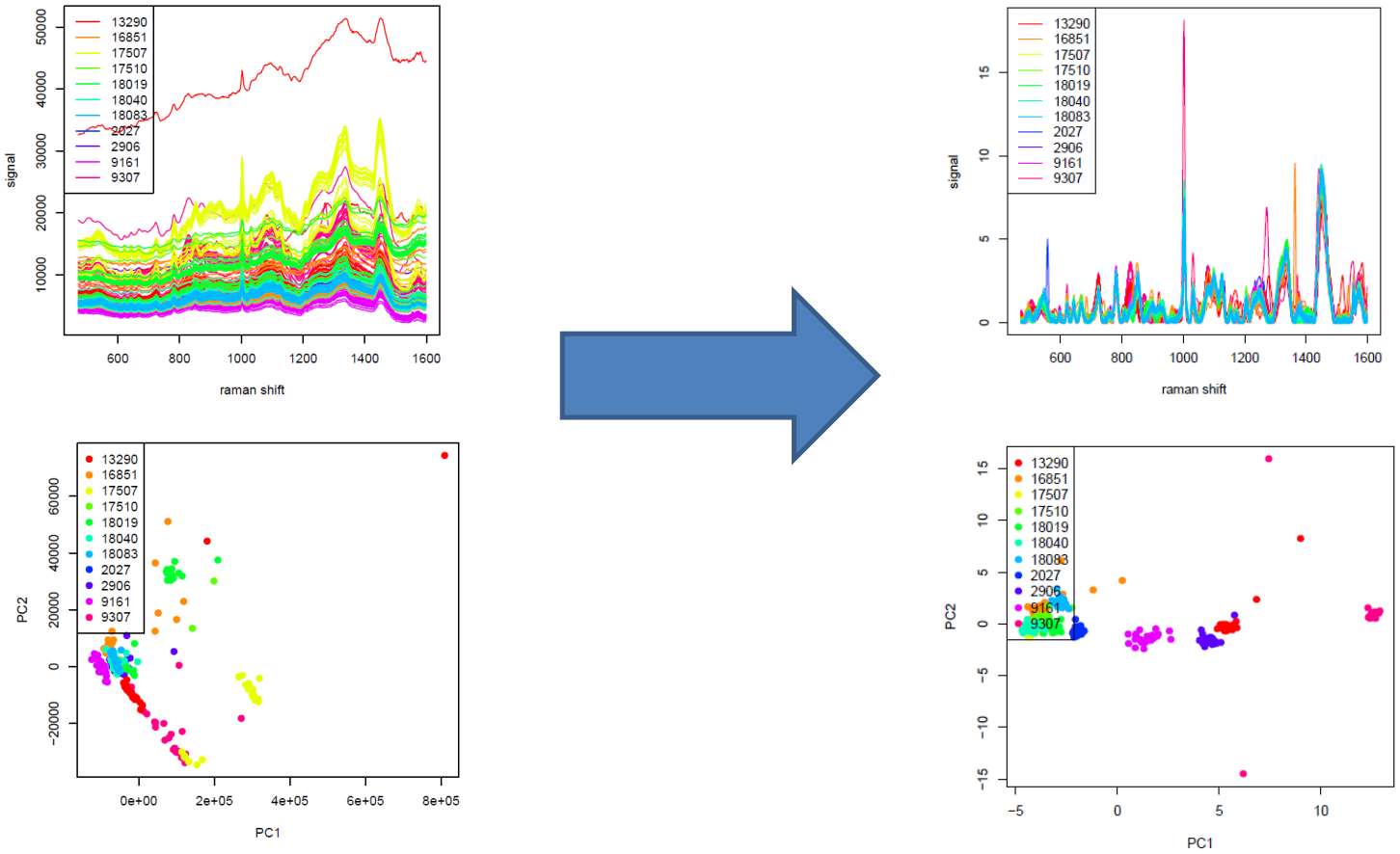


ACP



Jeu de données disponible

- Données disponibles = spectres pré-traités



Objectifs – 1/2

- Objectif #1: prédire l'espèce de chacune des 133 souches de test

➔ NB : 9 spectres par souche, mais 1 prédiction par souche

➔Critère de performance = taux de bonne classification (accuracy)

Objectifs – 2/2

- Objectif #2: évaluer l'intérêt des méthodes de « deep learning » dans ce contexte

➔ Construire une « baseline » au moyen d'algorithmes « standard » (SVM, RF, ….)

➔ Essayer de faire mieux avec différentes architectures : MLP, CNN, réseaux récurrents (type GRU / LSTM)…ou autres !

Remarques

- Questions ouvertes
  - Pré-traitement des spectres ?
  - Gestion des réplicats au sein d'une même souche ?
  - Données « fonctionnelles » & réseaux de neurones ?
  - …

- Attention aux outliers et au déséquilibre entre les classes…

Evaluation

- A rendre pour le 17/01, 20h (via moodle):
  - prédictions sur jeu de test
    - Fichier texte de 133 lignes.
    - Ligne #i = prédiction de la souche #i (dans {sp_1, …, sp_42})
  - rapport d'analyse
    - 8 pages maxi
    - **Expliciter les différences entre les différentes architectures neuronales considérées et leurs intérêts dans ce contexte**.

- Critères d'évaluation
  - 1) performance
  - 2) clarté du rapport
  - 3) créativité et exhaustivité (surprenez moi !)