

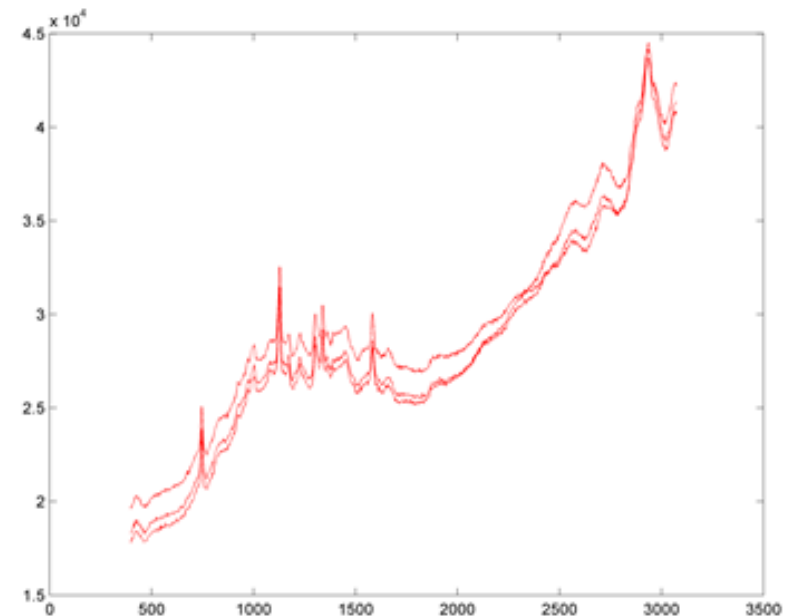
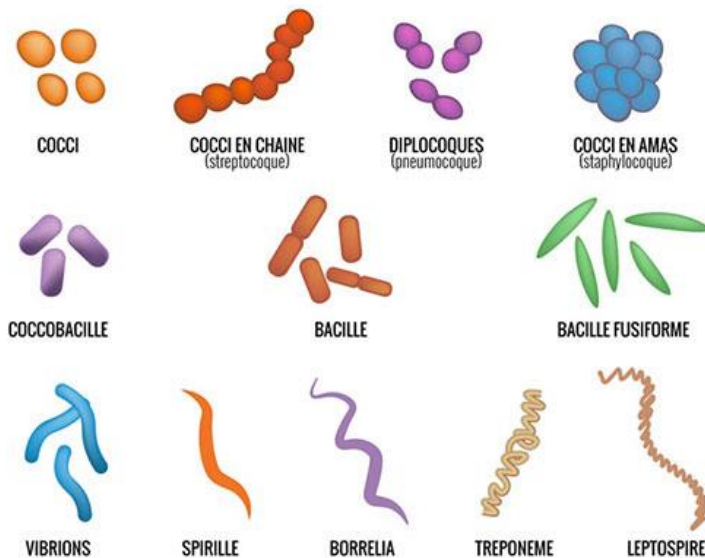
Data Challenge

UE Fouille de Données – M2 SSD

20/01/2020

Objectif

- Identification bactérienne et spectroscopie : reconnaître l'espèce d'une souche bactérienne à partir d'un spectre (RAMAN).



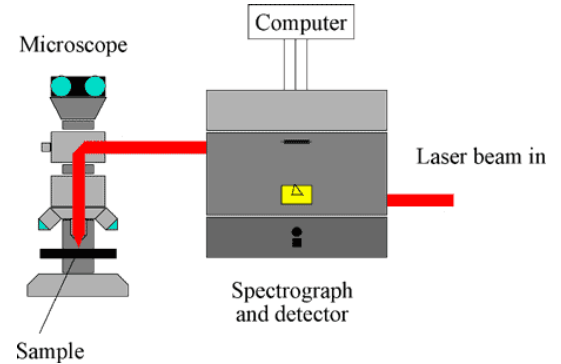
Comment ça marche ?



mise en culture



sélection d'une colonie



acquisition d'un spectre

algorithme de classification



Jeu de données disponible

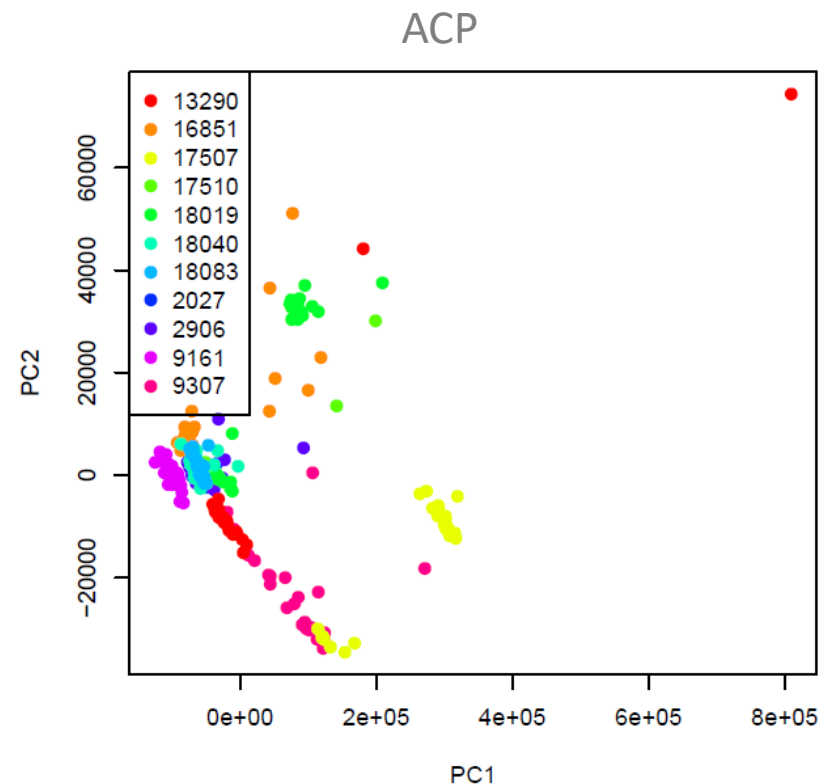
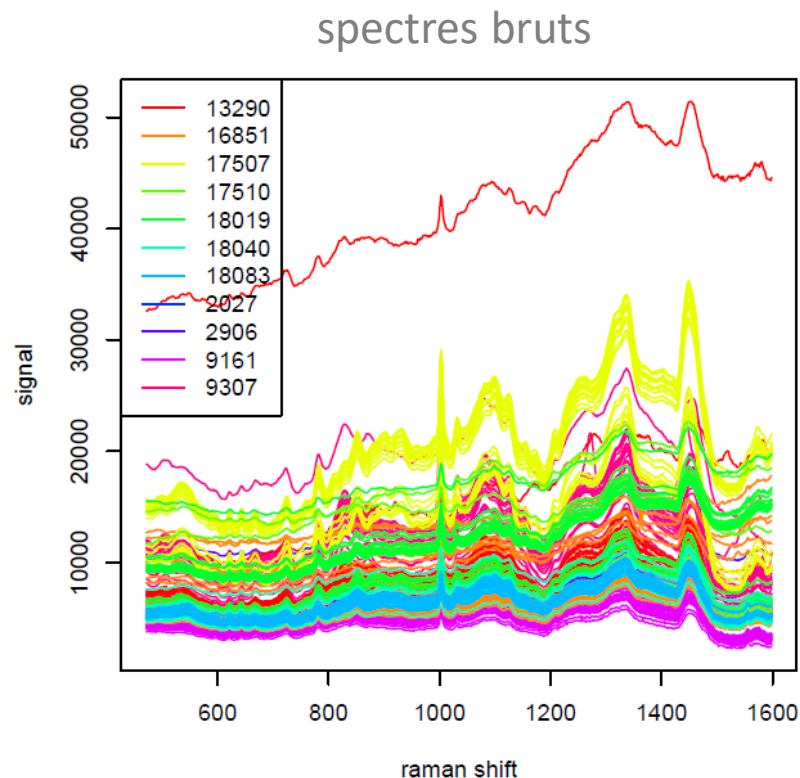
- Apprentissage :
 - 346 souches bactériennes (les individus) provenant de 42 espèces bactériennes (les classes)
 - 9 spectres acquis par souche ➔ 3114 spectres
- Test :
 - 133 nouvelles souches
 - 9 spectres acquis par souche ➔ 1197 spectres

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)

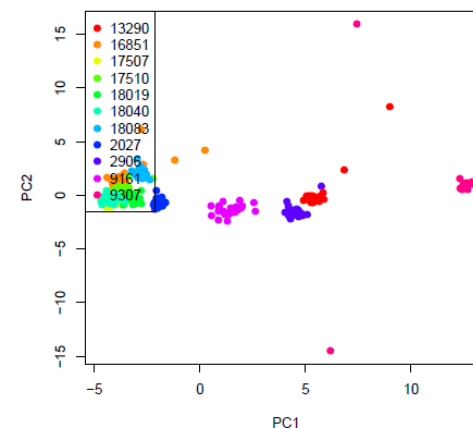
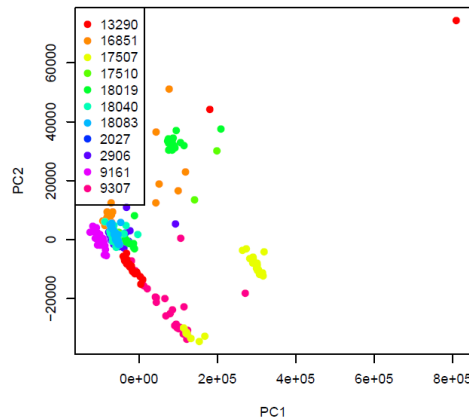
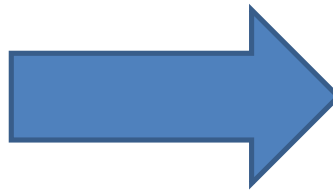
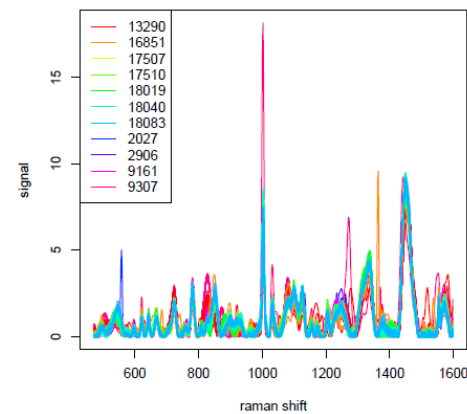
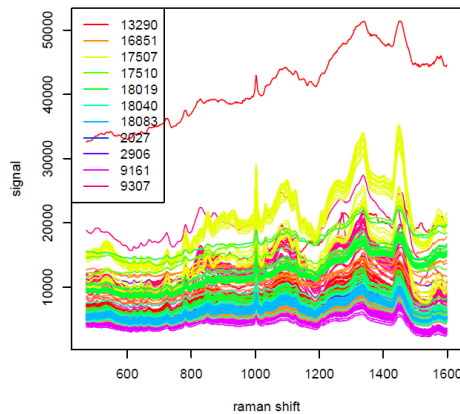
Jeu de données disponible

- **Illustration** : 1 espèce, plusieurs souches



Jeu de données disponible

- Données disponibles = spectres pré-traités



Objectif

- Objectif : 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)

Remarques

- Questions ouvertes
 - Pré-traitement des spectres ?
 - Gestion des réplicats au sein d'une même souche ?
 - Algorithme de classification à considérer ?
 - Données « fonctionnelles » ?
 - ...
- Attention aux outliers et au déséquilibre entre les classes...

Evaluation

- A rendre pour 18h (par email):
 - 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
 - 5 pages maxi
- Critères d'évaluation
 - 1) performance
 - 2) clarté du rapport
 - 3) créativité et exhaustivité