Conférences Axe RSME

Méthodes de Machine Learning pour les données de multi-omiques

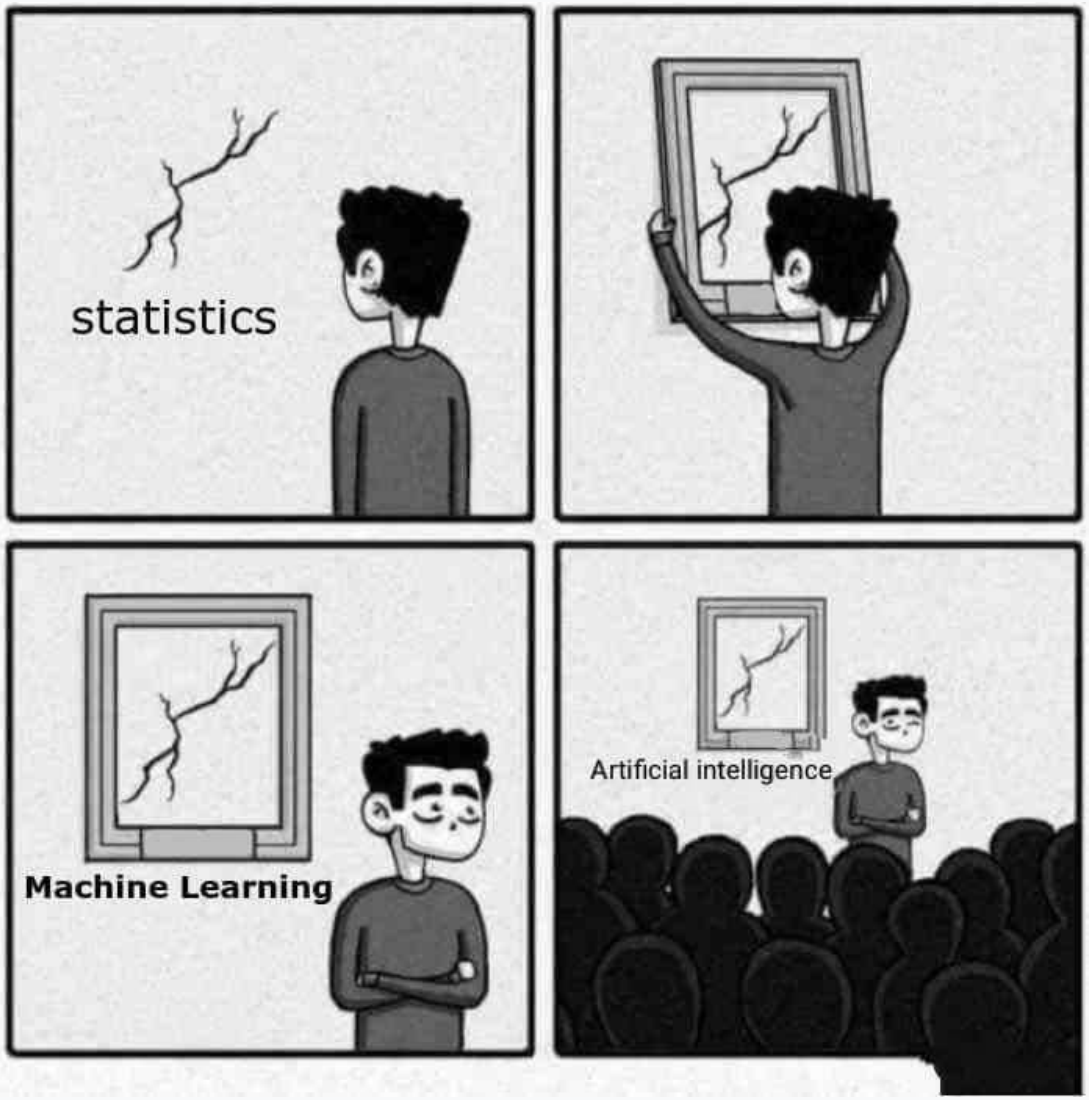
Loïc Mangnier, PhD

Arnaud Droit Lab

loic.mangnier@gmail.com

https://statsxomics.blog/

github.com/lmangnier


statistics
Machine Learning
Artificial intelligence

Inférence ou Prédiction ? 3 questions existentielles

Si l'on cherche à déterminer l'impact d'une maladie sur l'expression d'un gène. Inférence ? Prédiction ?

Si l'on cherche à prédire l'impact d'un médicament sur le profil métabolique d'un patient. Inférence ? Prédiction ?

Si l'on ne cherche qu'à identifier les biomarqueurs les plus prédictifs pour l'apparition d'une maladie. Inférence ? Prédiction ?

En conclusion

- **Inférence**: Cherche l'association entre plusieurs variables → Significativité statistique (valeurs-p et intervalles de confiance)
- **Prédiction**: Cherche à prédire de nouvelles valeurs sur la base d'un **modèle** et de **valeurs passées**

Cependant dans bien des cas les chercheurs vont combiner des approches du type **Machine Learning** et de l'**Inférence** pour répondre à des questions complexes

Exemple: Réduction de dimension + prédiction

Construction d'un modèle de Machine Learning

- **Point de départ**: Question de recherche impliquant de la prédiction
- **Étape 2** Analyse exploratoire des données
- **Étape 3** Sélection de 2-3 modèles **maximum** et choix des métriques de performance dépendants des données et de la question de recherche.
- **Étape 4** Application des modèles, fine-tuning et comparaison
- **Étape 5** Selection du meilleur modèle et validation

Types de Machine Learning

- **Apprentissage supervisé**: la réponse est connue (régression, classification)
- **Apprentissage non-supervisé**: la réponse est inconnue (clustering)
- **Apprentissage semi-supervisé**: présence de données avec réponse et sans réponse

Exemple: Réduction de dimension sur données multi-omics (PCA)

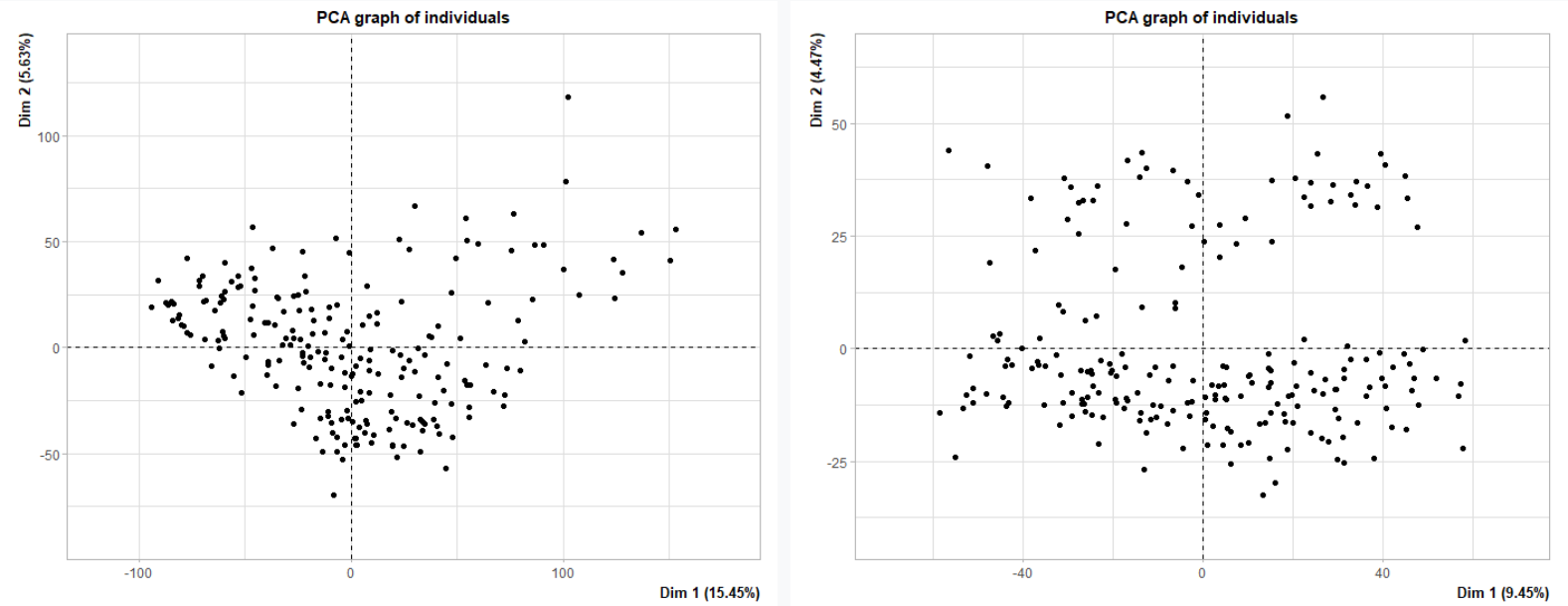


Figure: PCA sur données de Microbiome et de Métabolome

Exemple: Clustering sur données multi-omics (Clustering Hiérarchique)

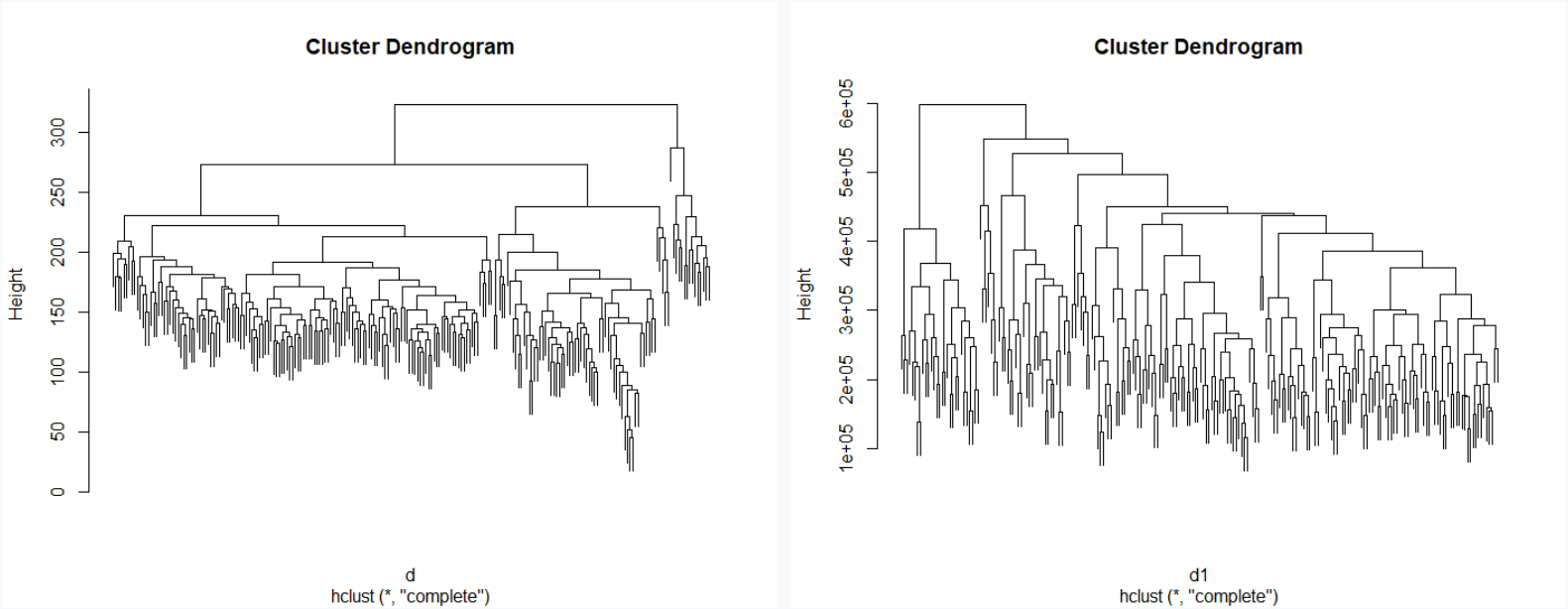
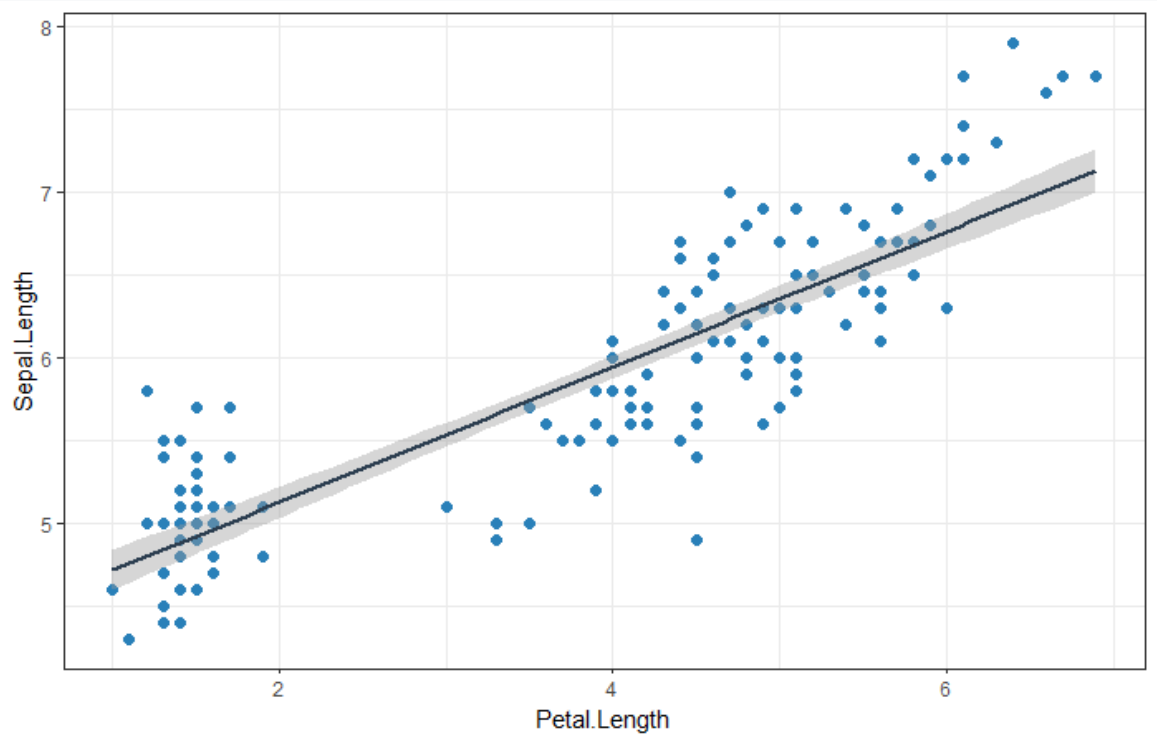
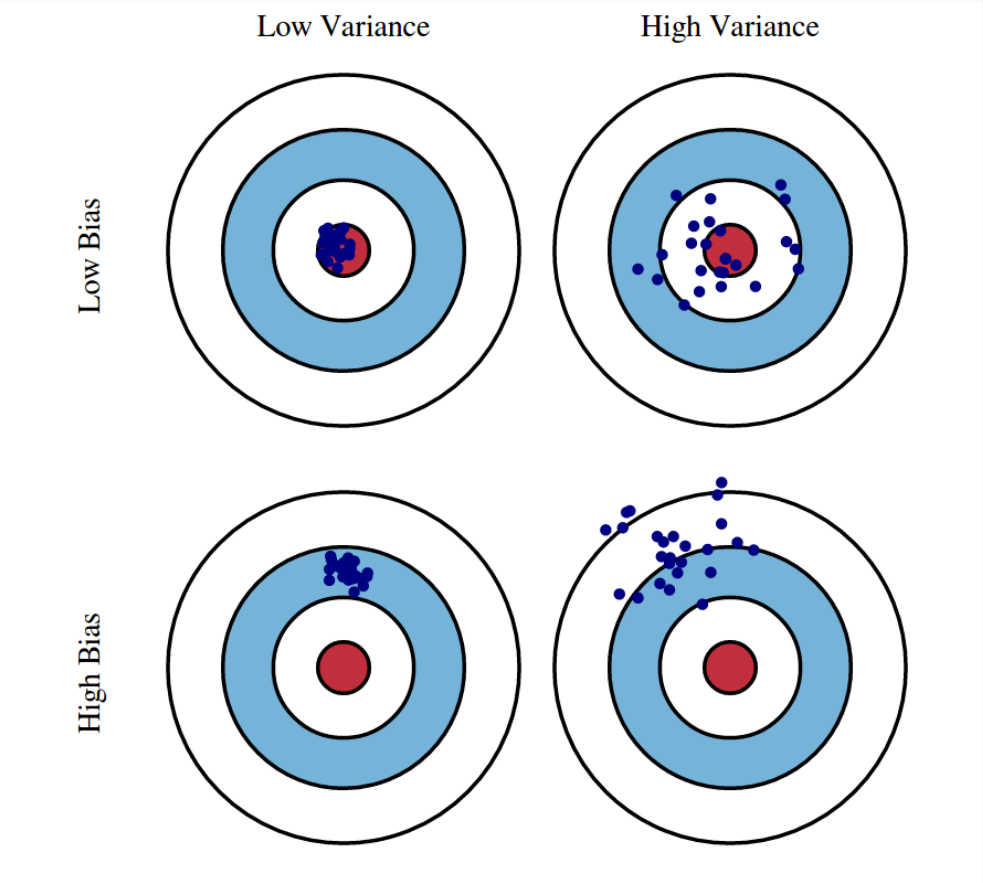


Figure: Clustering Hiérarchique sur données de Microbiome et de Métabolome

Exemple: Regression (Régression linéaire)



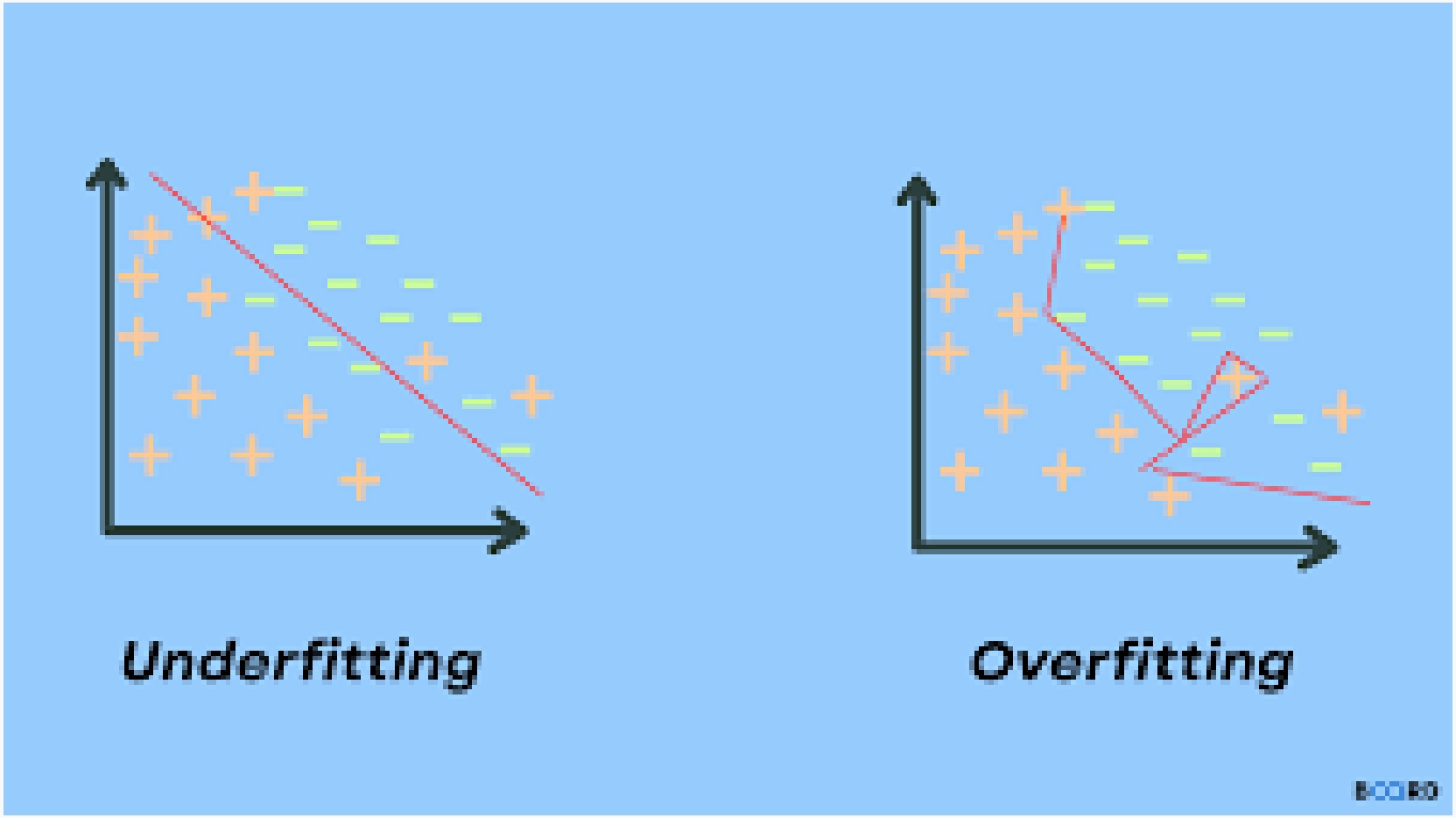
Balance Biais-Variance



Sous-apprentissage et Sur-apprentissage

- **Sous-apprentissage**: Les structures apprises par le modèle sont trop grossières : la variance est **faible**, le biais est **fort**
- **Sur-apprentissage**: Le modèle capture tout le bruit des données au lieu d'apprendre les structures générales: la variance est **forte**, le biais est **faible**
- Un modèle idéal généralisable est alors un bon compromis entre biais et variance

Sous-apprentissage et Sur-apprentissage



Machine Learning pour les données de multi-omics: Stratégies intégratives

- **Fusion précoce**: Les omics sont combinés dans un seul jeu de données avant la modélisation → **dépendance**
- **Fusion tardive**: Les modèles sont ajustés afin d'identifier les features centrales, un modèle final est utilisé → **indépendance**
- **Approches multi-vues**: Un modèle est ajusté en tenant compte d'un certain niveau d'**entente** entre les omics

Exemple d'application pour données de Métagénomique et de Métabolomique pour les maladies chroniques intestinales

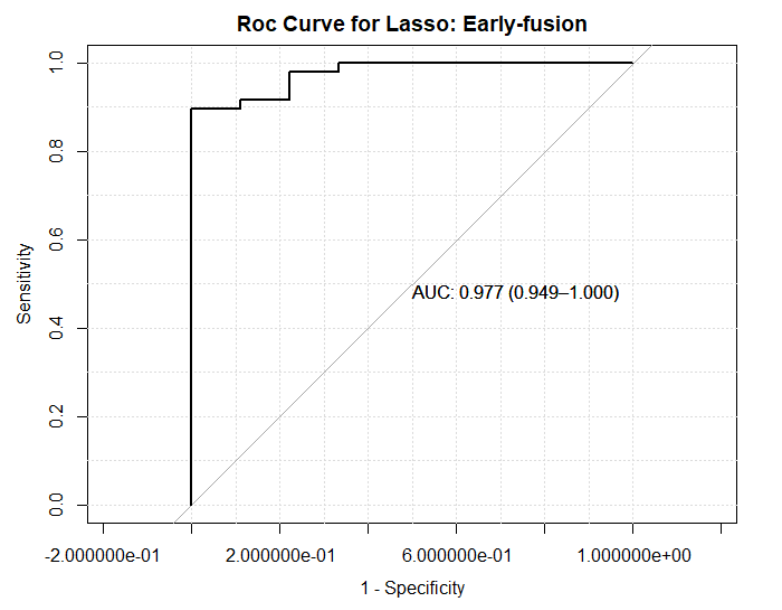
Les données de Métagénomique et de Métabolomique en bref

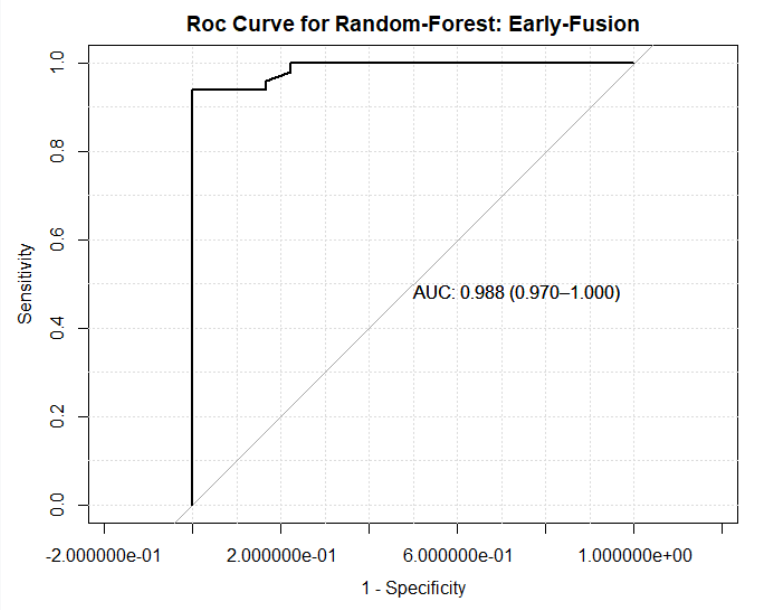
- **Métagénomique**: données de **comptage**, **compositionnelles** et **surdispersées**
- **Métabolomique**: données de **concentration surdispersées**
- 220 individus: 56 sains + 164 malades (CD + IBD)
- 55,882 espèces
- 8,848 métabolites

Modèles et méthodologie

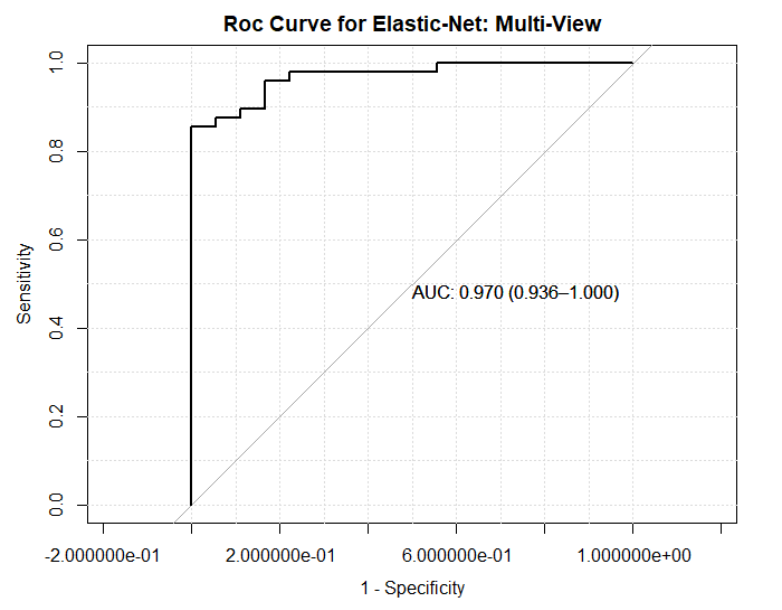
- **Stratégies d'intégration**: Fusion précoce, approche multi-vues
- **Modèles avec Fusion précoce**: Régression Elastic-Net, Random Forest
- **Modèles avec Multi-vues**: Régression Elastic-Net, sPLS-DA par blocs
- **Métriques de performance**: AUC
- 70% entrainement 30% test
- **Validation**: Courbe de Calibration

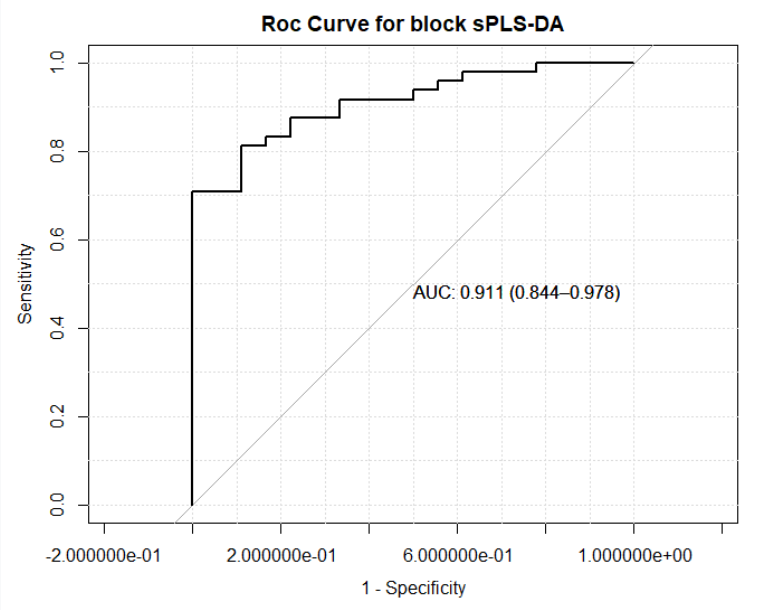
Comparaison des modèles: Fusion Précoce





Comparaison des modèles: Multi-vues



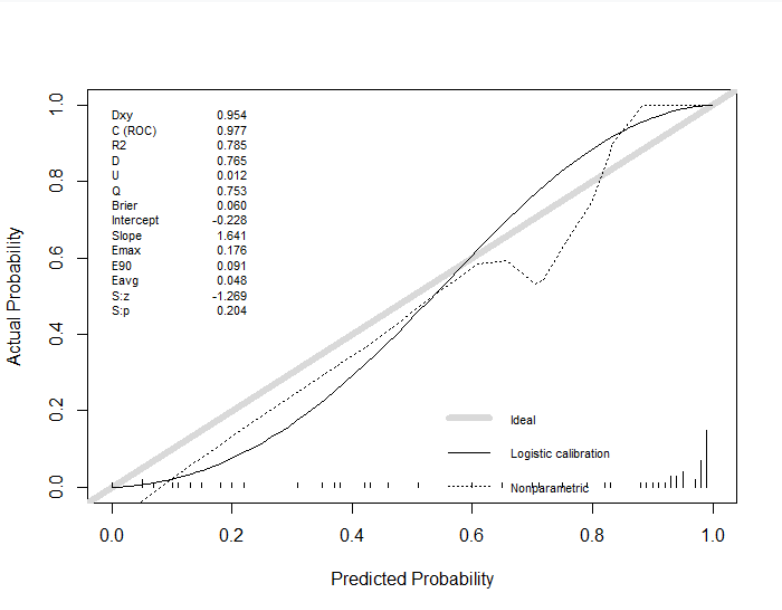


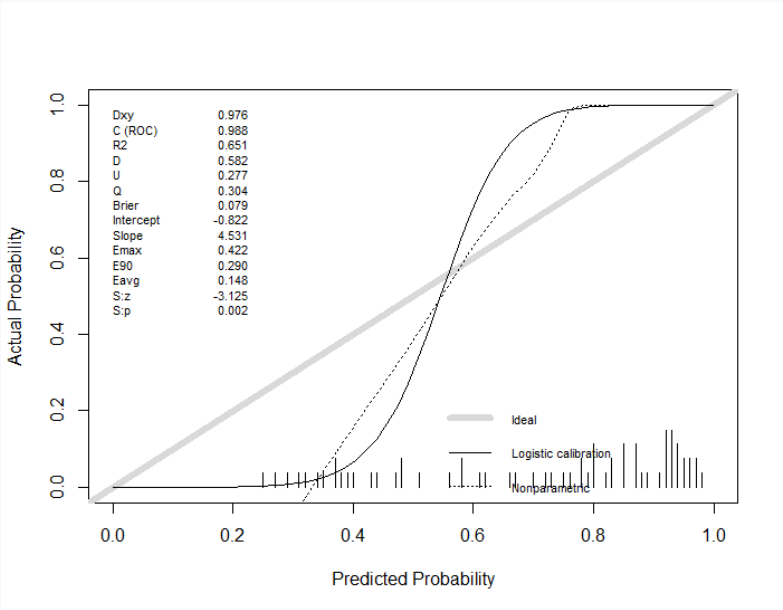
Quelques mots sur la calibration !

Calibration

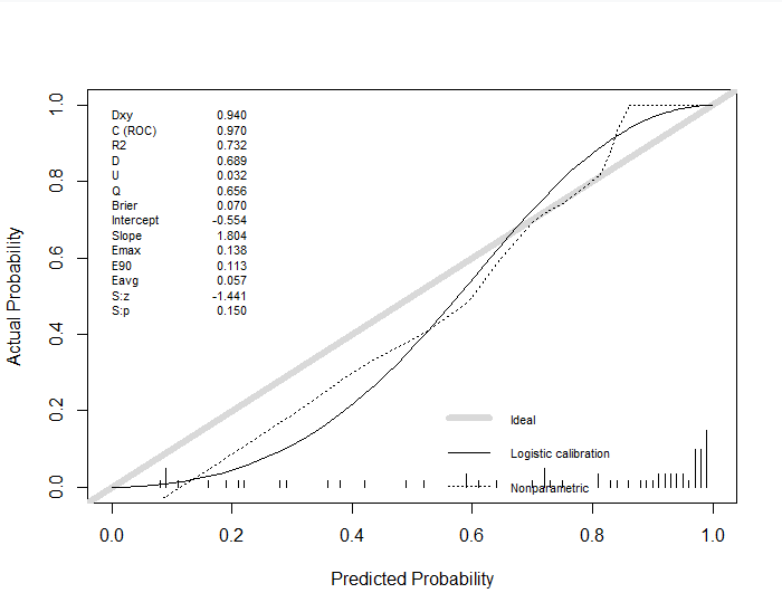
- **Calibration**: Permet d'interpréter la probabilité renvoyée par le modèle comme un risque de développer la maladie
- Si mauvaise calibration: il existe des corrections post-hoc

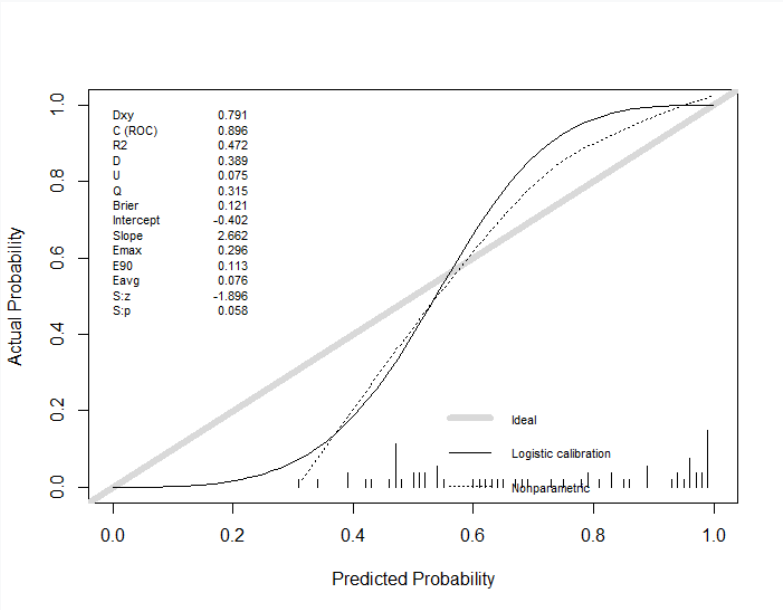
Calibration





Calibration





Conclusion

- *Principe de compréhension*: Comprendre les données avec un modèle simple surpassera toujours une utilisation aveugle d'un modèle complexe
- *Principe de parcimonie*: Nombre de modèles limité, Balance Biais-Variance, Nombre restreint de variables
- *Principe de validité externe*: Calibration
- Cependant le plus important de tous: *Principe de reproductibilité*

Pour aller plus loin

- Extension à des tâches de régression
- Emploi de 3 ensembles Entrainement-Validation-Test lorsque présence d'hyperparamètres
- Certains modèles nécessitent une normalisation des variables
- D'autres approches de séparation des données sont disponibles pour favoriser la généralisation (Validation croisée)

Pas d'approche systématique. Comprenez vos données avant tout !

Merci pour votre attention !
Questions ? Commentaires ?

Suivez moi sur mes réseaux sociaux

Matériel supplémentaire

- *Ressources d'apprentissage*: https://www.coursera.org/specializations/machine-learning-introduction; https://www.statlearning.com/; https://www.fun-mooc.fr/en/courses/machine-learning-python-scikit-learn/
- *Métriques de classification*: Area Under the Curve (AUC), Score-F1, Matthews Correlation Coefficient (MCC)
- *Métriques de régression*: Mean Squared error (MSE), R2
- *Métriques de calibration*: Courbe de calibration (modèle proche de la droite diagonale), Brier Score (0 = modèle parfaitement calibré; 1 = modèle pas calibré)