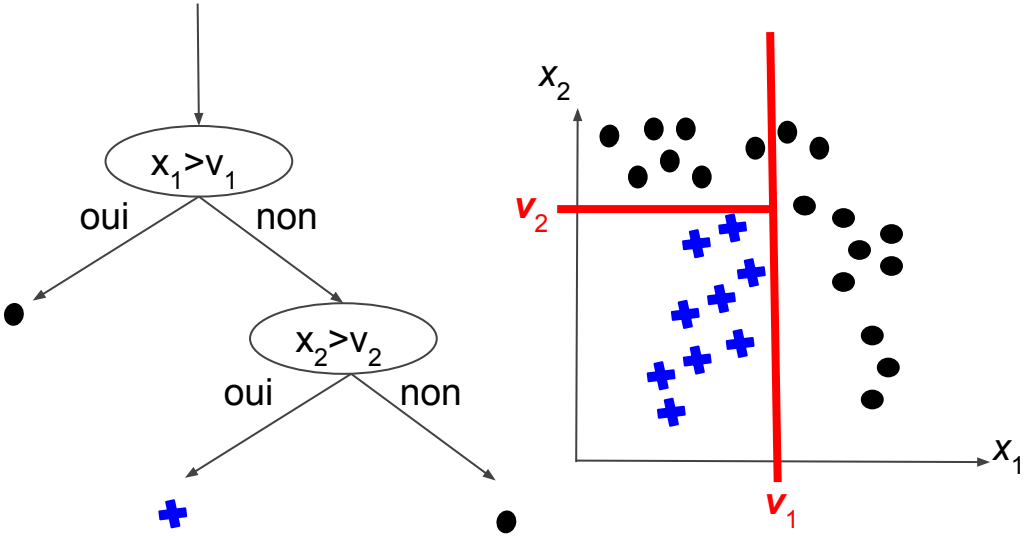
XGBoost vs Random Forest

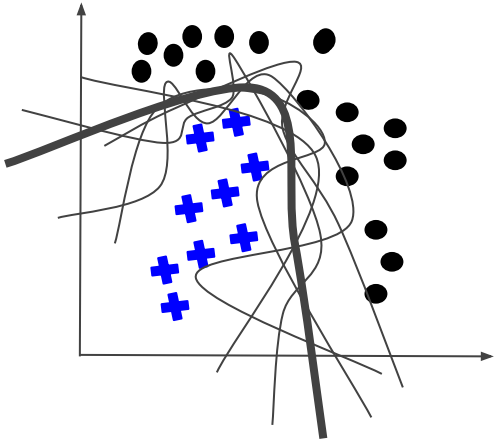
...

Riwal LEFORT
Groupe Arkéa
Service DataLabs

- **Un arbre :**



- **Une forêt :**



Chaque arbre de la forêt est issu d'un échantillonnage aléatoire des exemples d'apprentissage et/ou des composantes de l'espace vectoriel.

- Le résultat final est obtenu en combinant les résultats de tous les arbres :

$$p(y=1 \mid x) \propto \sum_t h_t(x),$$

  - $y \in \{0,1\}$: variable à prédire,
  - $x \in \mathbb{R}^{Nx1}$: variable observée,
  - $h_t \in \{0,1\}$: variable prédite par l'arbre à l'itération t.

- **Le boosting repose sur deux idées fondamentales :**
  - Le résultat final est obtenu en combinant les résultats **pondérés** de tous les arbres :

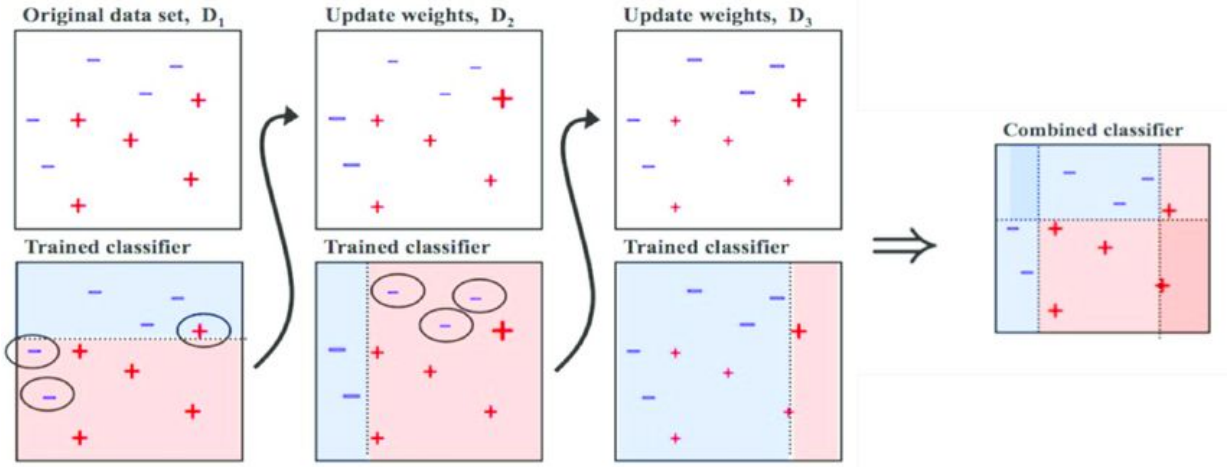
$$p(y=1 \mid x) \propto \sum_t \alpha_t h_t(x),$$

    - $\alpha_t \in [0,1]$: poids associé à la prédiction de l'itération t.

- **Le boosting repose sur deux idées fondamentales :**
  - Le résultat final est obtenu en combinant les résultats **pondérés** de tous les arbres :

$$p(y=1 \mid x) \propto \sum_t \alpha_t h_t(x),$$

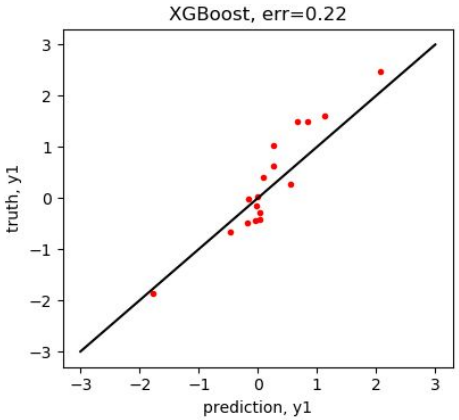
    - $\alpha_t \in [0,1]$: poids associé à la prédiction de l'itération t.
  - Les poids des exemples d'apprentissage sont mis à jours au fure et à mesure des itérations :

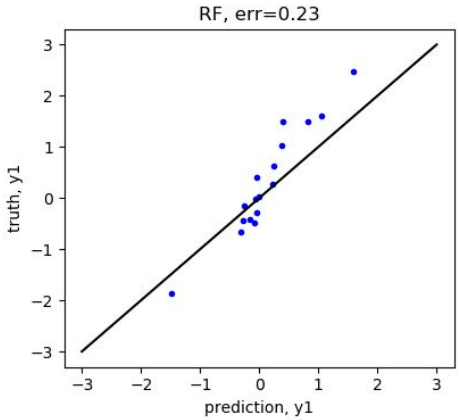


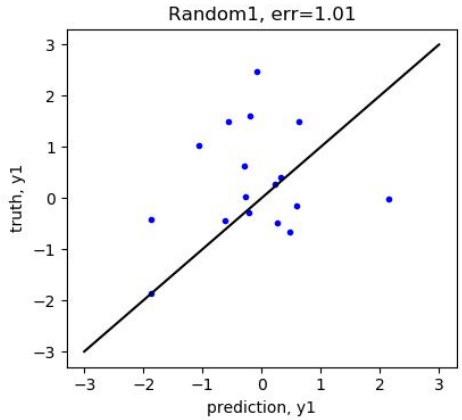
- **Pseudo code :**
  - cf. presentation “gradient_boosting.pdf”, maitre de conférence, univ Lyon.
- **Librairie :**
  - T. Chen and C. Guestrin, “XGboost: a scalable tree boosting system”, KDD, 2016.

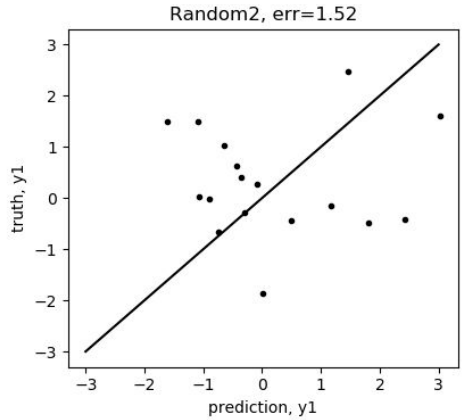
- **Base de données d'IFREMER de l'atelier en Machine Learning du 12/12/2018 :**
  - régression,
  - on essaie de prédire les caractéristiques des sédiments :
    - 2 variables continues (ratio de classes granulométriques),
  - en fonction de divers paramètres,
    - 92 variables,
      - analytiques, qui décrivent les échantillons,
      - bathymétriques, qui sont dérivée de la bathymétrie,
      - issue de la réflectivité (acoustique ?),
  - train : 50, test: 17.

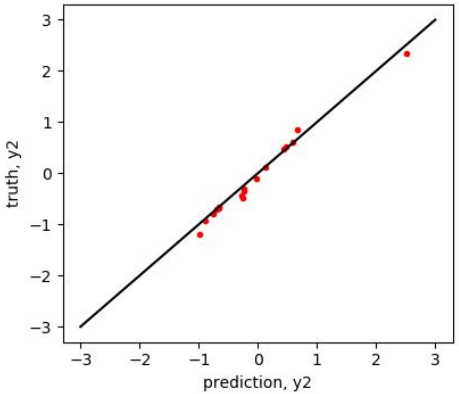
Random Forest Boosting **Résultats**

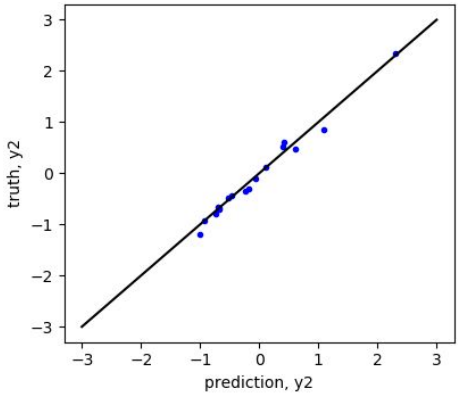


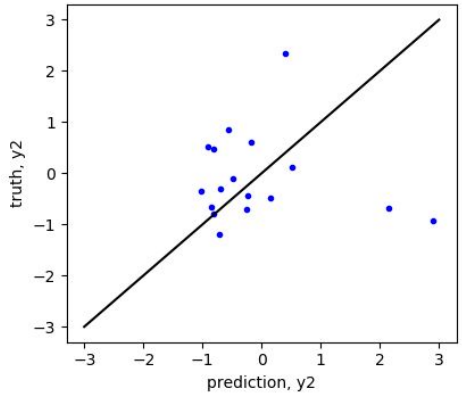


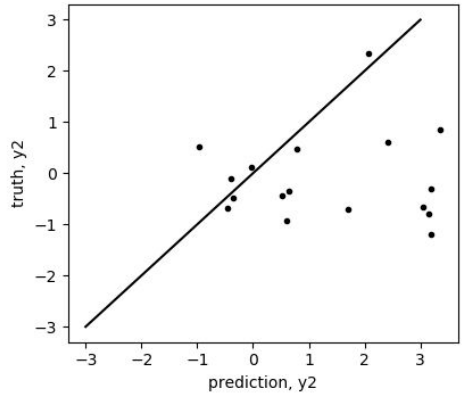












- **Base de données “breast_cancer” :**
  - classification (212 negatifs / 357 positifs),
  - on essaie de prédire un cancer du sein, en fonction de 30 paramètres qui décrivent le prélèvement cellulaire :
    - dimension, symétrie, aire, texture, *etc.*
  - train : 40/40, test : 172/317,

- **Résultats :**

---------- XGBoost ----------
error: 0.067
Confusion Matrix:
[[154 18]
[15 302]]

---------- RF ----------
error: 0.073
Confusion Matrix:
[[152 20]
[16 301]]