

# Big Data Analytics

Random Forest

---

- les arbres de décision overfit facilement
- ils sont rapides à apprendre
- en combiner beaucoup est faisable et réduit la variance

→ création d'une forêt (ensemble d'arbres) aléatoire

# But

Produire des arbres décorrélés et moyenner leurs prédictions pour réduire la variance.

## Outil 1 — bagging (row sampling)

**Bootstrap aggregating** (Bagging) :

- tirer un échantillon du dataset avec replacement
- entraîner un arbre sur cet échantillon
- répéter  $B$  fois

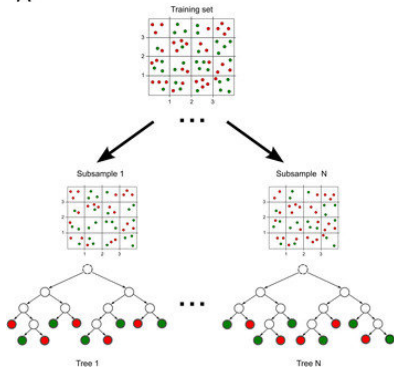
Le bagging s'appelle aussi row sampling.

## Outil 2 — random subspace method (column sampling)

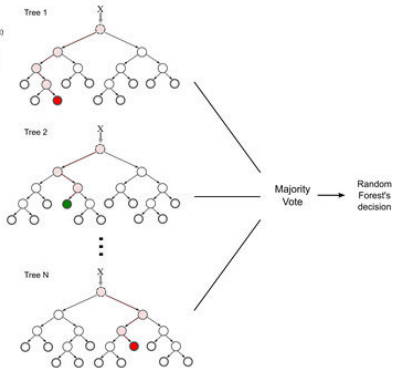
- à chaque split, considérer seulement un sous-ensemble des features
- valeurs conseillées :
  - classification :  $\lfloor \sqrt{m} \rfloor$  features par split
  - régression :  $\lfloor \frac{m}{3} \rfloor$  features par split, 5 exemples par node minimum

# Random Forest

A



B



- Pas de sur-apprentissage en augmentant le nombre d'arbres
- Une fois appris, le modèle est très rapide

## Conclusion

---



- les arbres sont interprétables, rapides à entraîner, combinables.
- random forest combine des arbres faibles en un prédicteur versatile