# Méthodes d'ensemble: Boosting, Bagging, Random Forest


## Résumé

*Nous explorons dans cette séance les classifieurs* construits à partir d'autres classifieurs *plus rudimentaires.*

— *La première partie du TP consiste à essayer de mieux cerner le comportement de l'algorithme de boosting grâce à une expérience simulée simple, mais pour laquelle tout est reprogrammé à la main (le code est fourni).*
— *Ensuite, on expérimente les méthodes vues en cours sur le jeu de données simulées pour lequel on connaît la frontière de Bayes, afin de pouvoir observer leur efficacité visuellement.*
— *Enfin, on expérimente sur données réelles dans le but de mettre en œvre les techniques désormais maîtrisées.*


## 1 Comment marche le boosting ?

Récupérer à l'adresse
https://www.math.univ-toulouse.fr/~agarivie/?q=node/178
le fichier R intitulé TEST_BOOSTING.R (le lien est appelé "Boosting experiments with R").

Il s'agit de comprendre :
— quel est l'objectif de ce code,
— comment il fonctionne (dans le détail),
— en quoi il permet de visualiser comment le boosting fonctionne,
— comment l'adapter à d'autres classifieurs faibles.

On cherchera en particulier à voir comment l'agrégation permet de s'affranchir des contraintes géométriques des classifieurs faibles, mais aussi comment le sur-apprentissage finit par apparaître.

## 2 Expérimentation sur données simulées

Sur les mêmes données simulées *matchmaker* que dans les séances précédentes, expérimenter la performance
— du boosting (en boostant les classfieurs de votre choix),
— de Random Forest.

Dans les deux cas, on essayera de visualiser la règle de classification obtenue et de la comparer à la règle de Bayes.

Quels sont les avantages et les inconvénients de ces deux méthodes ?

## 3 Expérimentations sur données réelles

Comparer l'efficacité de *toutes les méthodes de classification supervisée* vues jusque là sur le jeu de données *IRIS* présenté précédemment.

Mettre en œuvre une validation croisée 5-fold pour le choix des paramètres.

Donner pour chaque méthode un intervalle de confiance pour la probabilité de mal classifier.

*Si le travail est terminé, recommencer avec un jeu de données de votre choix tiré d'UCI, comportant au moins 10000 exemples de chaque classe.*