



---

**MOVIELENS RECOMMENDER :**  
Movie recommandation system

---

13 mars 2017

## Introduction

Il est arrivé à plus d'un de vouloir regarder un film chez soi le soir mais ne pas savoir quoi..

Cette question est une opportunité d'explorer un problème intéressant d'Apprentissage Machine (Machine Learning) en construisant un *système de recommandation* basé sur la base de données de films rassemblée par le groupe *Movielens*

## Méthodes d'Apprentissage Automatique

La section suivante a pour but de présenter sommairement chacune des méthodes de Machine Learning qui ont été utilisés dans la composition de notre système de recommandation.

### *KMeans (K-Moyennes)*

La méthode des k-means est un outil de classification classique qui permet de répartir un ensemble de données en  $k$  classes homogènes. Dans le cadre de la classification non supervisée, on cherche généralement à partitionner l'espace en classes concentrées et isolées les unes des autres. La méthode va essayer de séparer les données en classes de même variance, en cherchant à minimiser la variance intra-classe (carré de la distance de chaque élément au barycentre), ce qui a pour conséquence principale que les clusters seront supposés de forme sphériques.

### *DBSCAN (density-based spatial clustering of applications with noise)*

L'algorithme du DBSCAN perçoit les clusters comme des zones de hautes densités séparé par des zones de petites densités; du fait de cette approche les clusters trouvés par la méthode peuvent être de forme quelconque, au contraire de la méthode précédente qui suppose que les clusters sont de formes convexes.

### *Birch (balanced iterative reducing and clustering using hierarchies)*

Algorithme d'exploration de données non supervisée utilisé pour produire une segmentation hiérarchisée sur des volumes de données particulièrement importants. Elle construit ce qui est appelé CFT (Characteristic Feature Tree) basé sur une supposition d'existence de sous arbre..

### *Agglomerative Clustering (Grappes agglomérés)*

La classification hiérarchique (Hierarchical Clustering) est une famille de méthodes de classification qui construit des clusters emboîtés en les fusionnant ou les séparant, leur représentation se fait sous forme de dendrogramme. Le Agglomerative Clustering performe un clustering avec une approche descendante (de bas en haut), à savoir que chaque observation commence comme son propre cluster et les clusters sont progressivement fusionnés, c'est le critère de liaison qui détermine la métrique pour la stratégie de fusion. (ex : Ward  $\rightarrow$  minimisation de la somme des carrés, une minimisation de variance approche)

## Comparatif des méthodes

Méthode	MSE	Homogeneity
Kmeans	1.239770	0.246491
DBSCAN	1.182537	0.252728
Birch	1.203780	0.243844
AG	1.258531	0.229231

Au vu des performances nous choisirons de développer (bâtir) notre système de recommandation sur la méthode d'apprentissage du *Density-Based Spatial of Applications with Noise* les raisons de ce choix seront expliquées par la signification de chacune de ces métriques expliquées dans le notebook **load preprocessing**