



Statistique et Informatique pour les Big Data



Philippe Besse et Béatrice Laurent-Bonneau (INSA)



Aurélien Garivier et Jean-Michel Loubes (UT3 – Paul Sabatier)

Objectif :

mutualisation d'**actions** et de **moyens pédagogiques pluridisciplinaires** pour le développement des **Big Data** dans les formations

Projet 2014-2016
Financement : 15k€



Manifestations



Journée Big Data 2014



Session Enseignement
"big data" des **Journées
de la Société
Française de
Statistique**, juin 2015



Data Science |
Game



3 équipes au **Challenge de Science des
données 2015** mis en place par L'ENSAE-
ParisTech



Journées Big Data 2015



Challenge Big Data – Toulouse
(11/15 - 02/16)
Recommandation par Filtrage
Collaboratif

en lien avec le trimestre
thématique « Machine Learning »



IMAT, CMI-SID, INSA, ISAE, TSE

Projets & Formations

- (INSA) Parcours Transversal Pluridisciplinaire (GEI-GMM) "léger" : comparaison d'implémentations (temps, précision) d'algorithmes d'apprentissage statistique
- (INSA) projet tuteuré (Lyra-Networks) : détection de fraudes dans les paiements en ligne
- (CMI-SID) projets  + 
- (CMI-SID) « big data » MongoDB
- conférence « virtual box Oracle – Hadoop – MongoDB – Rhadoop »
- Achat d'un petit serveur dédié à l'expérimentation
- Formation pySpark/Hupi (12 participants IMT + IRIT, UPS + INSA)
- **Articles :**
 - Besse P., Laurent B. *De Statisticien à Data Scientist; développements pédagogiques à l'INSA de Toulouse*, Statistique et Enseignement, à paraître.
 - Louède J., Chevalier M., Garivier A., Mothe J. *Systèmes de recommandations : algorithmes de bandits et évaluation expérimentale*, Journées de statistiques Jun. 2015
 - Besse P., Gadat S., Garivier A., Loubes J-M., Simatos F. *Expérimentation d'un défi « big data » entre masters toulousains*, preprint (à suivre l'an prochain)



Ressources pédagogiques :



= centre de ressources en ligne pour l'enseignement de la statistique

Création de l'atelier Science des Données :

Introduction et compléments méthodologiques

- De Statisticien à Data Scientist
- Collaborative filtering
- Sequential and reinforcement learning. Stochastic Optimization (1) (2)

Compléments "technologiques"

- Programmation fonctionnelle et objet avec Python
- Introduction à MapReduce avec Rhadoop
- Apprentissage profond avec H2O
- Manipulation de RDDs et Hadoop avec pySpark
- Apprentissage sur données massives avec Mllib

Ateliers (Les calepins .ipynb sont accessibles à partir du descriptif de l'atelier)

- SD1: Reconnaissance de caractères (MNIST) avec R, Python, Spark
- SD2: Recommandation de films (MovieLens) avec Spark
- SD3: Catégorisation de produits (CDiscount) avec Python, Spark



À venir

- Nouveaux Challenge Big Data – Toulouse annuels (avec nouveaux participants dont Bordeaux)

2016-2017 : avec openbikes.co



- Journées Big Data 2016 @ UT1-TSE
- Nouvelles ressources pédagogiques sur wikistat
- Elargissement actif aux autres formations toulousaines potentiellement concernées
- Difficulté : synchronisation des calendriers des formations