Apprentissage Statistique et Données Massives

Philippe Besse

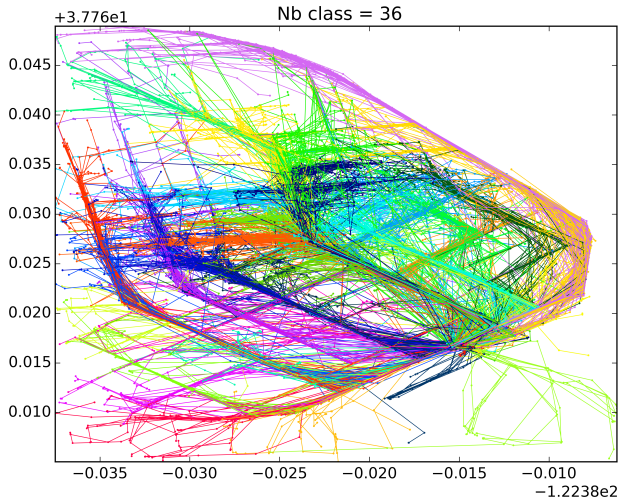
Université de Toulouse
INSA – Dpt GMM
Institut de Mathématiques – ESP
UMR CNRS 5219

Big Data

- Croissance exponentielle du Volume
- Variété, Vélocité
- Valorisation et analyse (ML)
- Passage à l'échelle Volume
- Méthodes d'apprentissage *vs.* Nouvelles technologies

Point de vue ”pédagogique”

- De Statisticien à *Data Scientist*
- Quelles compétences ?
- Quelques Exemples



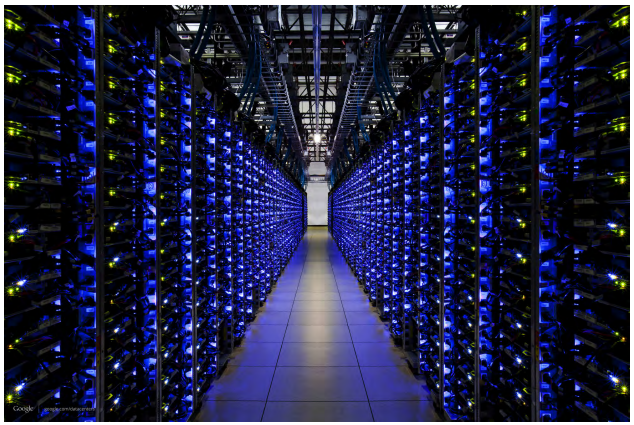
Trajectoires GPS

Domaines et objectifs très variés

- E-commerce : recommandations et Réseaux sociaux
- Publique : administrations, santé et (*open data*)
- Recherche Météo, Biologie, Astronomie...
- Industrie : défaillance, fraudes, maintenance...

Réellement massives ?

- Seuils technologiques (RAM, Disque)
- Préparation (*munging*) des données (Python-pandas)
- Données distribuées : *Hadoop distributed file system*



Ferme de données

Réalité ou confusion ?

- Aspects sociétaux et Datafication du quotidien
- big data *vs.* big brother (NSA)
  Information / prévision de comportement moyen / individuel
- Assurances et asymétrie d'information
  Segmentation *vs.* mutualisation des risques

Question

La Science des données est-elle une Nouvelle Science ?

Volume et ”nouveaux” paradigmes

1990s MO *Data Mining* & Expérimentation

2000s GO *Bioinformatique* & Parcimonie ($p >> n$)

2010s TO *Science des données* & Optimisation

Nouveau terme d'erreur

- Erreur d'approximation *vs.* d'estimation (biais / variance)
- Erreur d'*optimisation*
  - Contrainte de ressources (temps, RAM, nb processeurs)
  - Taille échantillon *vs.* temps d'exécution & mémoire
  - Méthodes disponibles pour données distribuées

Nouveaux modèles économiques

- Eldorado de la pub en ligne (*advertising*)
- *Cloud computing* : SaaS, IaaS, PaaS, DaaS, ITaaS...
  - Marges sur matériels et logiciels (*open source*)
  - Amazon (WS), Microsoft (Azure),
  - Google *cloud*, IBM (Analytics), ...
  - Python et *Enthought*, *Continuum analytics*
  - Spark (Databricks), H20 (Oxdata), RHadoop (Revolution Analytics – Microsoft)

Nouvelles Méthodes stat ou ML ?

- Hadoop et *MapReduce* pour paralléliser
- Retour vers le futur (SVD, k-means, logistique, RF...)
- Obligation de collaborer entre Maths, Info
  - Michael Jordan (SFdS 13/10/2015)
  - GDR MADICS du CNRS (juin 2015)

Nouveaux problèmes d'optimisation

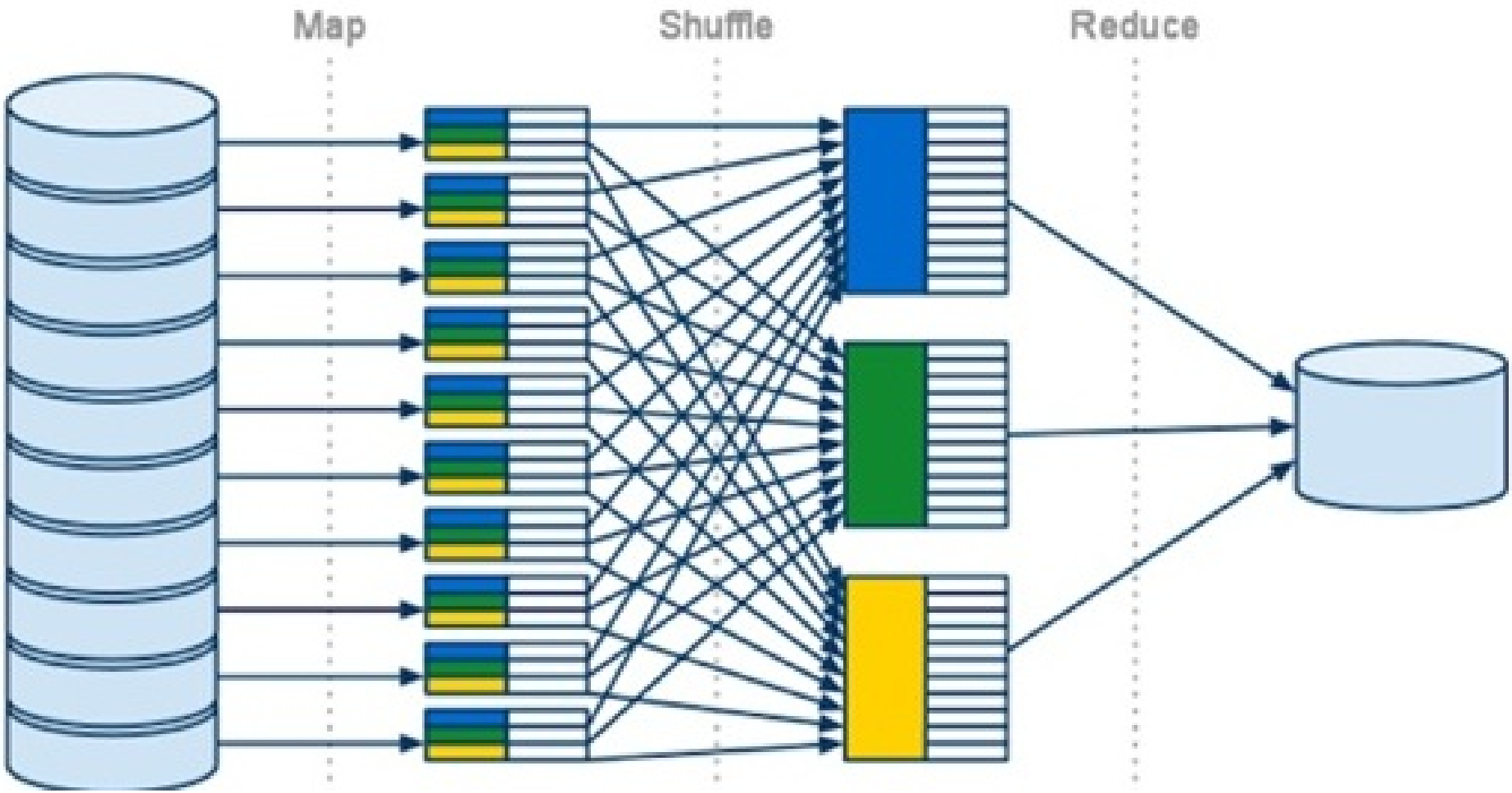
- Optimisation convexe et parcimonie (Candes & Tao, 2010)
- Gradient stochastique (données distribuées ou en flux)
- Librairies d'algorithmes parallélisés

En résumé

- Exemple d'*Amazon Web Service Machine Learning*
  - *Utiliser une technologie d'apprentissage-machine puissante sans avoir besoin de maîtriser les algorithmes et techniques de l'apprentissage-machine* (*sic !*)
  - *Cloud* mais qu'avec modèle linéaire ou logistic
  - Pénalisation *Lasso, Ridge*, mais ... manuelle
- Science des données
  - Nouveau *packaging*
  - Nouveaux enjeux
  - Explosion de nouvelles technologies
  - Problèmes d'optimisation

Hadoop

- Environnement : Google puis Apache
- *Hadoop Distributed File System* (HDFS)
- Données hétérogènes distribuées
- Distribution : Cloudera, Hortonworks, Oracle, IBM...
- Parallélisation : *Map Reduce*

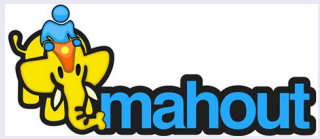


Hadoop Distributed File System (HDFS)

Classification par centres mobiles ($\approx k$-means)

- Définition d'une distance euclidienne (ou non : PAM)
- Algorithme de Forgy
  - Initialisation des k centres
  - Itération des étapes *MapReduce*
    - Map : Affectation de chaque individu `(valeur)` au centre `(clef)` le plus proche
    - Reduce : Calcul des centres des individus de même clef
    - Mise à jour des centres
- Problème : accès disques à chaque itération
- Solution actuelle : Spark (*Resilient Distributed Dataset*)

Logiciels de Statistique & Hadoop

-   (Scikit Learn)  , Knime, Weka...
- Bibliothèques R : `bigmemory, parallel, snow`...
- RHadoop (*Revolution Analytics*)
- Bibliothèque Java d'apprentissage (Apache) : 

Hadoop & Après

- Hive, pig...
- Spark Zaharia et al. (2012)
  Scala, Java, Python (pyspark),
  (SparkR) & MLlib,

Méthodes échelonnables (*scalable*)

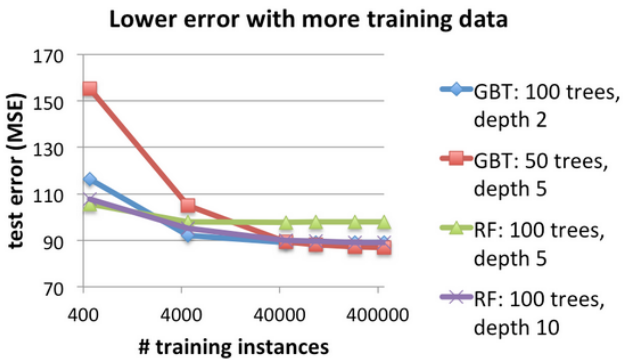
- *RHadoop* : k-*means*, régression, régression logistique...
- *MLlib* de Spark : k-*means*, SVD, NMF (ALS), régression linéaire et logistique avec pénalisations, SVM linéaires, classifieur bayésien naïf, Arbre, Forêt Aléatoire, Boosting
- Finalement peu de méthodes
- Mais passage direct à l’échelle ”volume”

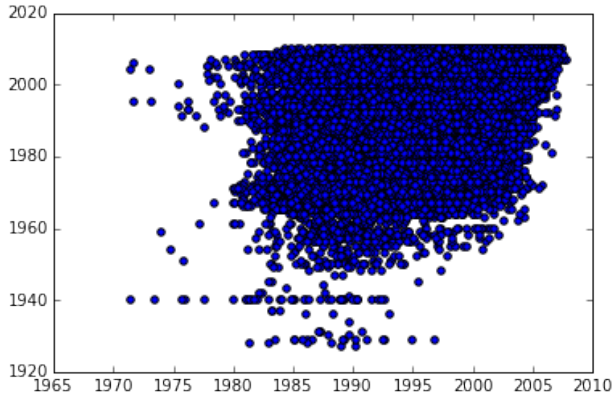
Implémentations des forêts aléatoires

- Python (`scikit-learn`) équivalente à R (`randomForest`)
- `ntree`, `mtry`
- MLlib de Spark
  - `maxDepth`, `minInstancesPerNode`
  - `featureSubsetStrategy`
  - `subsamplingRate`
  - `maxBins` (32) $< n$
  - `maxMemoryInMB`, `useNodeIdCache`, `checkpointDir`, `checkpointInterval`

Données massives et ”information”

- Compenser un signal faible (sismique inverse)
- Représentativité
  - Fiasco de *Google flu trend*
  - Exhaustivité et mesure d'audience (Médiamétrie)
    Philippe Tassi (SFdS 13/10/2015)
- Qualité des variables (*features*)
  - Million Song Dataset : benchmark de Databricks





Base UCI — Crédit : Databricks — Année observée vs. prédite

Méthode × Technologie

- Critéo (*advertising*) : régression logistique et *group lasso*
- Tinyclues (profilage) : factorisation non négative
- CDiscount (catégorisation) : régressions logistiques
- Deepki (bâtiments) : random forest et boosting
- Airbus (essais en vol) : archivage Hadoop (Oracle)
- IRT Saint-Exupéry (images satellites) : boosting (Spark)

Conclusion

- *Platform as a service* Amazon WS : +50% par an
- *Software as a Service* : Watson, AWS ML, tensor flow...
- Hadoop Spark MLlib
- Domaines en développement pérenne, d'autres pas...
- *Gartner Hype Cycle* : *Trough of Disillusionment of ML*
- Sélection ”naturelle” des technologies