

TECHNOLOGIES DE L'IA

INTRODUCTION

Philippe Besse et Brendan Guillouet

29 Octobre 2018

Institut National des Sciences Appliquées

INTRODUCTION

INTRODUCTION

DÉFINITION ET OBJECTIFS

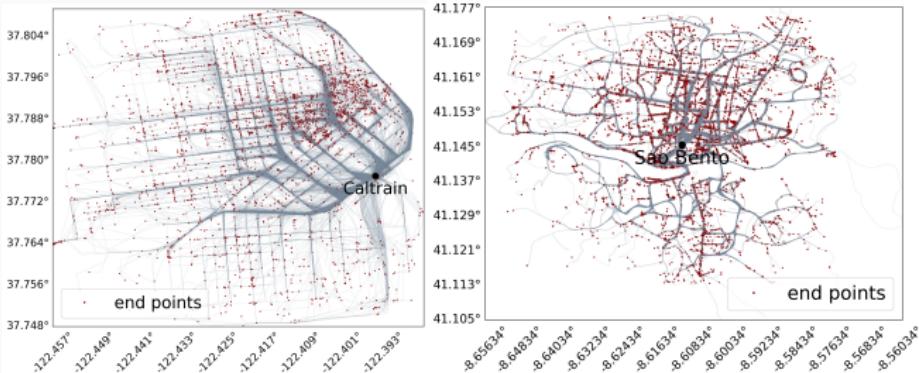
SUCCÈS DE L'IA : rencontre de

- Données massives
- Puissance de calcul
- Algorithme d'apprentissage statistique (*Deep Learning*).

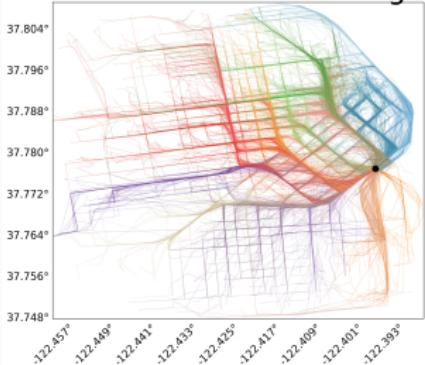
Définition des GROSSES DONNÉES

- Volume : croissance exponentielle
- Variété : signaux, images, graphes
- Vélocité : décision séquentielle, optimisation stochastique
- Validité, Valorisation → Prévision → Apprentissage

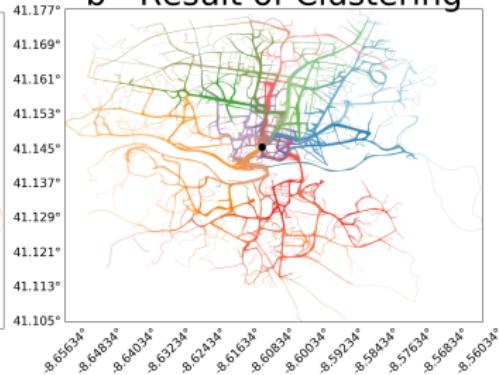
VARIÉTÉ : CLASSIFICATION DE TRAJECTOIRES GPS



b - Result of Clustering



b - Result of Clustering



Confusion de domaines 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)
- Données distribuées :
- *Hadoop, MapReduce, scalability*

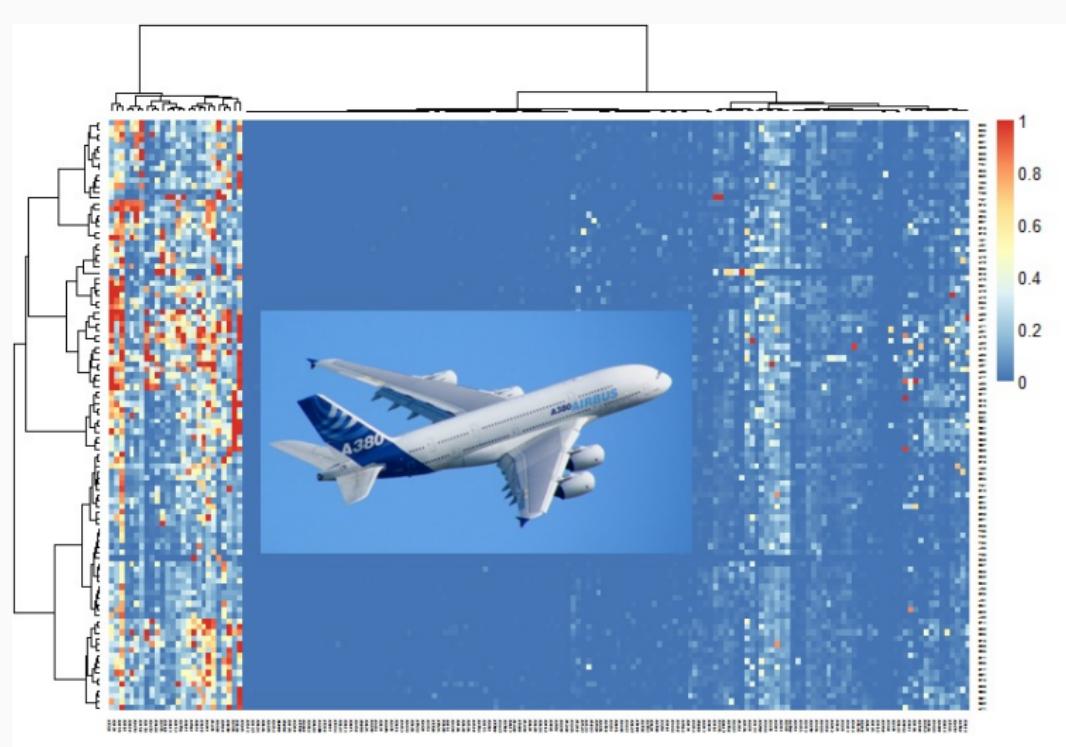


Ferme de données

OBJECTIFS

- Apprentissage sur données massives
- Question : Quelles (meilleures) technologies utiliser ?
- Matériel : Poste personnel, GPU, serveur(s), *Cloud*
- Données distribuées (*Hadoop*) ?
- Point de vue du "statisticien"
 - Prototype puis passage à l'échelle du même code
 - Cas d'usage : MovieLens, Cdiscount, MNIST, Cats/Dogs
 - Algorithmes : *munging*, NMF, Logit, RF, SVM, *Gradient boosting*, *Deep Learning*
 - environnements : R, Python, Spark, XGBoost, TensorFlow, Keras
- Comparaisons de difficiles à impossibles
- Scripts dans [▶ `http://github.com/wikistat`](http://github.com/wikistat)

AIRBUS : ANALYSE DES MESSAGES D'INCIDENTS EN VOL (700 000 EN 6 MOIS)



INTRODUCTION

SCIENCE DES DONNÉES MASSIVES

NOUVELLE (?) SCIENCE DES DONNÉES

- 1995 *Data Mining* : GRC & suites logicielles
- 2010 *Data Science* : Publicité en ligne, *cloud computing*
 - Données préalables (fouille), dimensions (omiques) $p \gg n$,
 - Pas de nouvelles méthodes, seules celles échelonnables
 - Data pas toujours Grosses mais *datification* du quotidien
 - Éthique et virtualisation / transparence de la décision
- "Science" des données : **nouveau** terme d'erreur
 - Erreur d'optimisation + Compromis biais / variance
 - Optimisation stochastique (Robbins Monro);
non différentiable (parcimonie)
 - Parallélisation et/ou distribution des calculs

NOUVEAU MODÈLE ÉCONOMIQUE

- Marges réduites sur matériels (IBM)
- Logiciels sous licence GNU, MIT, Apache... (Microsoft, SAS)
- Vendre du **service** :
 - Enthought (Canopy), Continuum analytics (Anaconda),
 - Horton Works, Cloudera (Hadoop, Spark...)
 - Databricks (Spark), Oxdata (H2O)
 - Revolution Analytics (RHadoop) – Microsoft
- Nouveau marché du *cloud computing*
 - Platform aaS, Software aaS, Service aaS...
 - Amazon Web Service
 - Microsoft Azure, Google Cloud Computing
 - IBM Analytics, SAS Advanced Analytics...
- Développement industriel vs. Recherche académique

ÉCOSSYSTÈME DES DONNÉES MASSIVES ET DE L'IA (TURCK, 2018)



V1 - Last updated 4/19/2018

© Matt Turck (@mattturck), Demilade Obayomi (@demilade_obyayomi), & FirstMark (@firstmarkcap) mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

LES ATELIERS

Objectifs de ces ateliers :

- Aborder les technologies récentes de l'IA
- Traiter des cas d'usage réalistes

CINQ ATELIERS

1. Compléments Python, introduction à Spark pour données distribuées
2. Analyse d'images (MNIST, cats vs. dogs) avec Keras & Tensor FFlow :
3. Calcul intensif avec la *Google cloud platform*
4. Système de recommandation (base movieLens) avec Spark/MLLib
5. NLP (base Cdiscount) traitement du langage naturel avec Spark, Python, Gensim

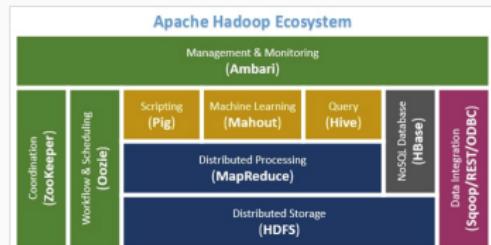
ÉVALUATION

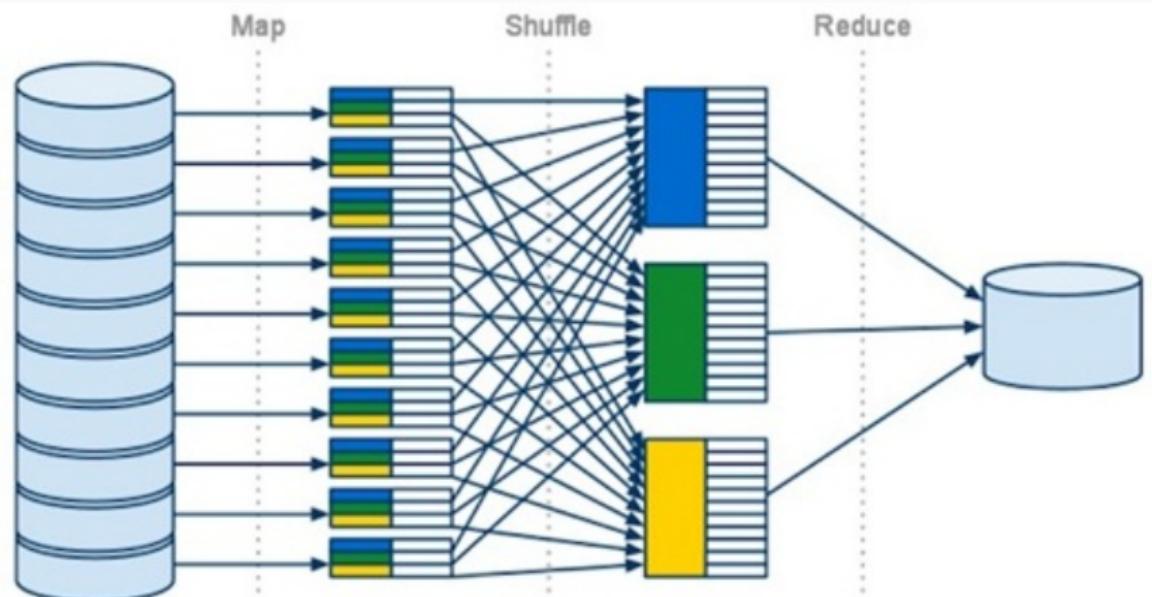
- Classement au DÉFI IA 2019
- Présentation orale

HADOOP, MAPREDUCE, SPARK

HADOOP, MAPREDUCE

- Environnement : *Google puis Apache* (2009)
- **Hadoop Distributed File System (HDFS)**
- Données hétérogènes distribuées
- Tolérance aux pannes matérielles
- Scalabilité (multiplier les noeuds)
- Parallélisation : *Map Reduce*
- Communication par (*clef, valeur*)
- Déplacer les algorithmes, pas les données
- Données *immuables*
- Lecture unique ou *streaming*





doop Distributed File System (HDFS) &
MapReduce

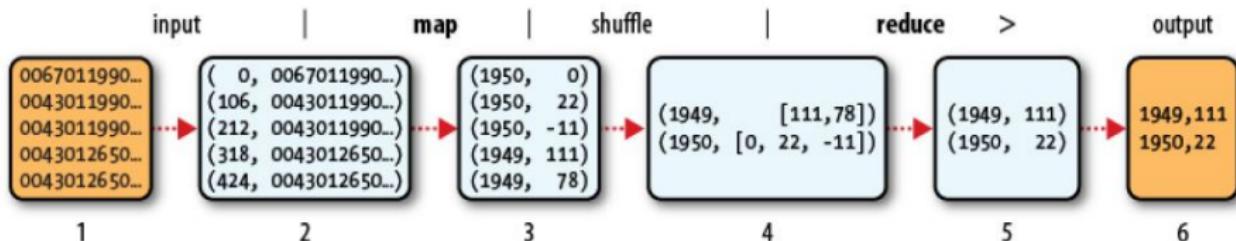
Ha-

MAPREDUCE

ALGORITHME :

1. Input \Rightarrow list (k1, v1)
2. Map() \Rightarrow list (k2, v2)
3. Shuffle, Combine \Rightarrow (k2, list(v2))
 - Implicite
 - Clefs identiques vers même nœud réducteur
4. Reduce() \Rightarrow list (k3, v3)

EXEMPLE :

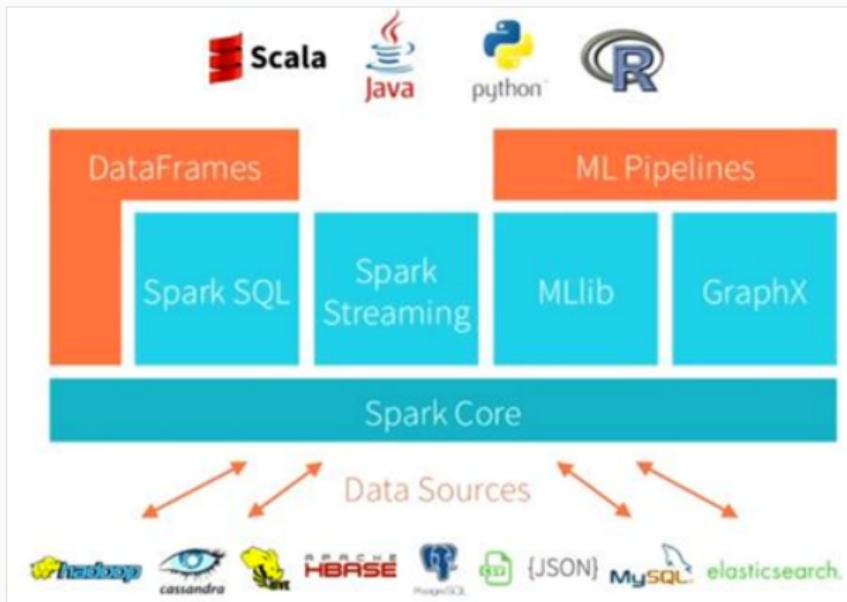


CLASSIFICATION PAR CENTRES MOBILES ($\approx k\text{-means}$)

- Définition d'une **distance euclidienne**
- Algorithme de **Forgy** (1965)
 - 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 de **Spark** : *Resilient Distributed Dataset*,
(Zaharia *et al.*, 2012)



LA TECHNOLOGIE Spark ET SON ÉCOSYSTÈME



- Spark 2.3
- **MLlib** (Resilient Distributed Dataset) : k -means, SVD, NMF (ALS), Régression linéaire et logistique (l_1 et l_2), SVM linéaires, Classifieur Bayésien Naïf, Arbre, Forêt Aléatoire, Boosting
- Évolution de **SparkML** : *DataFrame, pipeline*
- Peu de méthodes mais passage à l'échelle "Volume"

FONCTIONNEMENT DES ATELIERS

PROGRESSION "PÉDAGOGIQUE" :

- Statistique de **petites** données avec R
- Traitement et stat de données plus **grosses** : Python
- Traitement et stat de données **distribuées** : PySpark
- Autres **algorithmes** avec Python : *XGBoost, deep learning*

MATÉRIELS

- Ordinateurs personnels avec R, Python 3.6 (Anaconda)
- GMM salles TP avec 16 cartes GPU ; R, Python 3.6, Spark 2 (Hadoop virtuel)
- *Cloud : Google platform*

RAPPEL DE L'OBJECTIF PRINCIPAL

Apprendre à s'auto-former à des technologies
en perpétuelle (r)évolution