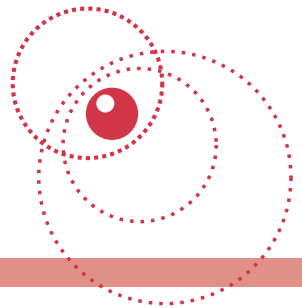

CES Data Scientist

Introduction aux modules d'informatique

9 mars 2015

Pierre Senellart

TELECOM ParisTech

La datamasse, le Big Data, les masses de données :

- Données collectées pour certaines applications, par certaines entreprises, données librement disponibles, etc.
- Très grande valeur à analyser (fouille, prédiction)
- Significativement plus complexe que les données traditionnelles :
  - Volume : ordres de grandeur au-dessus de ce qu'on peut traiter de manière centralisé
  - Variété : types de données (texte, multimédia, graphes, structurées) variés, organisation des données variées
  - Vélocité : données changeantes ou produites à grande vitesse (LHC : 100 millions de collision par seconde), parfois plus que ce qu'on est capable de stocker
  - Véracité : qualité de l'information très variable ; imprécision dans l'information

La datamasse, le Big Data, les masses de données :

- Données collectées pour certaines applications, par certaines entreprises, données librement disponibles, etc.
- Très grande valeur à analyser (fouille, prédiction)
- Significativement plus complexe que les données traditionnelles :
  - **Volume :** ordres de grandeur au-dessus de ce qu'on peut traiter de manière centralisé
  - **Variété :** types de données (texte, multimédia, graphes, structurées) variés, organisation des données variées
  - **Vélocité :** données changeantes ou produites à grande vitesse (LHC : 100 millions de collision par seconde), parfois plus que ce qu'on est capable de stocker
  - **Véracité :** qualité de l'information très variable ; imprécision dans l'information

30 et 31 mars : Données structurées et numériques, Raja Chiky

Bases de données NoSQL (Cassandra, MongoDB, CouchDB, bases de données graphes), flux de données

TP : MongoDB

20 et 21 avril : Données textuelles et Web, Fabian Suchanek

Représentation de connaissances, contenu Web semi-structuré et non structuré, Web sémantique, extraction d'informations depuis le Web (entités nommées, extraction de faits)

TP : Extraction d'informations depuis Wikipedia

28 et 29 septembre : Visu. de données massives, James Eagan

Principes de la visualisation de données, le langage et environnement de développement Processing, interaction

TP : Visualisation de données géographiques interactive avec Processing

19 et 20 octobre : Stockage distribué, Pierre Senellart

Introduction aux systèmes distribués, systèmes de fichiers distribués (HDFS/GFS), bases de données NoSQL (BigTable/HBase, Chord, Dynamo/Voldemort), index inversé

TP : installation d'Hadoop ; index inversé en HBase

9 et 10 novembre : Calcul distrubué, Pierre Senellart

Bases du calcul distribué, Hadoop MapReduce, Spark, Storm, Giraph

TP : calcul distribué d'index inversé

30 novembre et 1er décembre : Apprentissage distribué et graph mining, Mauro Sozio

Distribution d'algorithmes d'apprentissage, factorisation de matrices, PageRank, découverte de communautés

TP : PageRank distribué