



CES Data Scientist

Introduction aux modules d'informatique



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 distribué**, 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

9 mars 2015





Par le téléchargement ou la consultation de ce document, l'utilisateur accepte la licence d'utilisation qui y est attachée, telle que détaillée dans les dispositions suivantes, et s'engage à la respecter intégralement.

La licence confère à l'utilisateur un droit d'usage sur le document consulté ou téléchargé, totalement ou en partie, dans les conditions définies ci-après et à l'exclusion expresse de toute utilisation commerciale.

Le droit d'usage défini par la licence autorise un usage à destination de tout public qui comprend :

- le droit de reproduire tout ou partie du document sur support informatique ou papier,
- le droit de diffuser tout ou partie du document au public sur support papier ou informatique, y compris par la mise à la disposition du public sur un réseau numérique,
- le droit de modifier la forme ou la présentation du document,
- le droit d'intégrer tout ou partie du document dans un document composite et de le diffuser dans ce nouveau document, à condition que :
 - L'auteur soit informé.

Les mentions relatives à la source du document et/ou à son auteur doivent être conservées dans leur intégralité.

Le droit d'usage défini par la licence est personnel et non exclusif.

Tout autre usage que ceux prévus par la licence est soumis à autorisation préalable et expresse de l'auteur : sitepedago@telecom-paristech.fr

