

Introduction data science

Data science
Master 2 ISIDIS

SÉBASTIEN VEREL

verel@lisic.univ-littoral.fr

<http://www-lisic.univ-littoral.fr/~verel>

Université du Littoral Côte d'Opale

Laboratoire LISIC

Equipe OSMOSE

Information

But, évaluation, objectifs, support de cours, bibliographie :

cf. siteweb

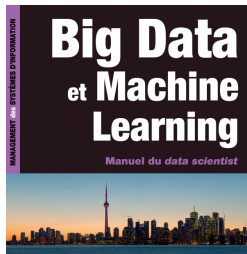
Bibliographie

Le cours et les supports reposent principalement sur ces sources bibliographiques :



Data Science : fondamentaux et études de cas
Machine Learning avec Python et R
Eric Biernat, Michel Lutz, 2015.

Bibliographie

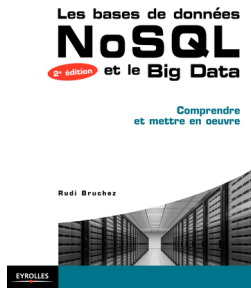


Pirmin Lemberger, Marc Batty
Médéric Morel, Jean-Luc Raffaelli
Préface de Michel Delattre

DUNOD

Big Data et Machine Learning
Manuel du data scientist Pirmin Lemberger, Marc Batty, 2015.

Bibliographie



Que les auteurs en soient remerciés chaleureusement !

Contenu Général

A Machine learning :

- Bases du data scientist
regression linéaire, logistique, bayésien naïf, etc.
- Les outils avancés
random forest, gradient boosting, SVM, etc.
- Concepts généraux
grandes dimensions, évaluation de modèles, etc.

B Hadoop avec Map-reduce :

- Système HDFS
- Map-reduce : exemples de bases
- Map-reduce : exemples avancés
- Framework pig

C NoSQL pour le big data :

- Présentation, différence SQP/noSQL
- Quelques implémentations : Hbase, Sqoop, Hive, etc.
- Machine learning en big data (mahout, Mllib)

Résoudre des problèmes

ah ! résoudre des problèmes...

- Panne d'une voiture
- Connaitre l'opinion sur un sujet dans les réseaux sociaux
- Prévoir la consommation électrique

Une définition

Data science

"Démarche empirique qui se base sur des données pour apporter une réponse à des problèmes"

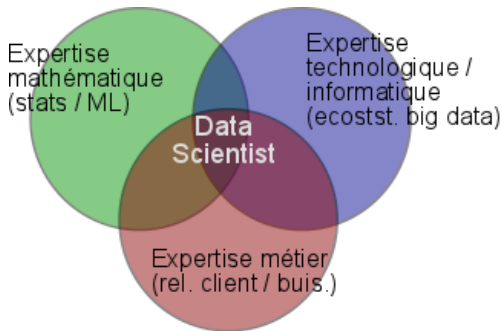
Data science : fondamentaux et études de cas, E. Biernat, M. Lutz, Eyrolles, 2015.

Le métier de Data scientist

Data scientist

- Apparue en 2008, DJ. Patil et Jeff Hammerbacher de Facebook et LinkedIn, ce sont appelés "data scientist"
- Généralisé à partir de 2012 :
"Data scientist : The sexiest Job of the 21th Century", T.H. Davenport, DJ. Patil, Harvard Business Review, oct. 2012.
- Rôle du data scientist gagne en importance dans les entreprises :
 - Augmentation (explosion !) du volume des données non structurées (big data)
- Dans les 10 prochaines années, profil data scientist sera très recherché

Les compétences



On peut aussi aller lire un post de Alex Woodie :
[http://www.datanami.com/2015/01/07/
9-must-skills-land-top-big-data-jobs-2015/](http://www.datanami.com/2015/01/07/9-must-skills-land-top-big-data-jobs-2015/)

Les jobs

Exercices

- Sur google trends : observer l'usage de "data scientist"
- Rechercher des offres d'emploi profils "big data", "business intelligence" ...
- Consulter le référentiel métier de l'apec "data scientist"

Big data

Un déluge de données

Source des données :

- Activité humaine
emails, photos, vidéo, logs, likes, etc.
- Activité des machines
capteurs en tout genre, compteurs en tout genre
(électrique, etc.), véhicules, électro-ménager
- Open data des institutions, des entreprises
horaires, statistiques sur les régions, géolocalisation, etc.
- open API de twitter, google, etc.
<http://www.programmableweb.com/>
- Le web !

Big data

Un déluge de données

Source des données :

- Activité humaine
emails, photos, vidéo, logs, likes, etc.
- Activité des machines
capteurs en tout genre, compteurs en tout genre (électrique, etc.), véhicules, électro-ménager
- Open data des institutions, des entreprises
horaires, statistiques sur les régions, géolocalisation, etc.
- open API de twitter, google, etc.
<http://www.programmableweb.com/>
- Le web !

Avertissement, data science ne se réduit pas au big data

Causes économiques

Les coûts baissent exponentiellement

- Capacité de stockage
- Capacité de calcul
- Bande passante

⇒ Emergence de data centers : Google, Amazon, LinkedIn, Yahoo!, OVH, etc.

How big ?

- Internet : $> 10 \text{ Po}$
- Data center : $> 100 \text{ To}$
- Disque dur : $\approx 10 \text{ To}$
- RAM : $< 100 \text{ Go}$

Frontière big data : lorsque les données ne peuvent être traitées en temps "raisonnable" ou "utile"

Calculer le temps nécessaire pour lire un disque dur de 1 To à 100 Mo/s ?

How big ?

- Internet : $> 10 \text{ Po}$
- Data center : $> 100 \text{ To}$
- Disque dur : $\approx 10 \text{ To}$
- RAM : $< 100 \text{ Go}$

Frontière big data : lorsque les données ne peuvent être traitées en temps "raisonnable" ou "utile"

Calculer le temps nécessaire pour lire un disque dur de 1 To à 100 Mo/s ?

Attention : données \neq information

Les fameux 3V (Gartner)

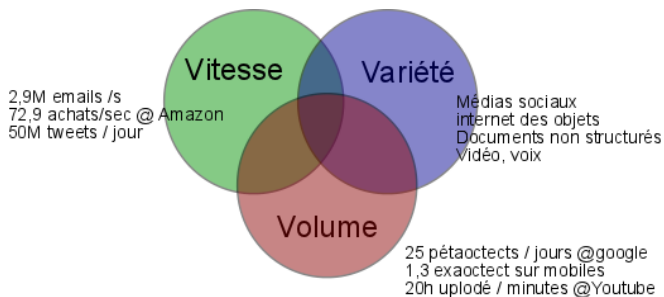


Schéma d'après "Big data et Machine Learning", Dunod, 2015.

Champs d'application

Nombreux champs d'applications actuels et futurs

- Tous les domaines de la science :
climat, physique, épidémiologie, médical, etc.
- En politique
Campagne Obama, etc.
- Secteur privé :
Relation clients, marketing ciblé, fréquentation, etc.
- Secteur public :
amélioration des services, adaptation aux besoins, etc.

Beaucoup de perspectives en vue !

Nouveaux besoins, nouveaux outils...

Quelques remarques éthiques

- Attention aux droits sur les données :
à qui appartient les données, leur exploitations, etc.
- Toutes les données ont un coût
- Une donnée peut être juste mais l'analyse fausse
- Une analyse de données n'est jamais neutre au sens objective !
- Une donnée n'est jamais neutre :
Une donnée est récoltée et exploitée dans un but précis

Démarche en data science

Démarche globale

- 1 Imaginer un produit, ou une question
- 2 Collecter les données
- 3 Préparer les données
- 4 Concevoir un modèle prédictif
- 5 Visualiser les résultats
- 6 Optimiser le modèle (calibration)
- 7 Déploiement, industrialisation

Le gros volume de données n'est pas une contrainte
mais une opportunité !

Contenu Général

A Machine learning :

- Bases du data scientist
regression linéaire, logistique, bayésien naïf, etc.
- Les outils avancés
random forest, gradient boosting, SVM, etc.
- Concepts généraux
grandes dimensions, évaluation de modèles, etc.

B Hadoop avec Map-reduce :

- Système HDFS
- Map-reduce : exemples de bases
- Map-reduce : exemples avancés
- Framework pig

C NoSQL pour le big data :

- Présentation, différence SQL/noSQL
- Quelques implémentations : Hbase, Sqoop, Hive, etc.
- Machine learning en big data (mahout, Mllib)