Open Source et Data Science

Ricco Rakotomalala

Université Lumière Lyon 2 – Data, informatique et statistique

http://dis.univ-lyon2.fr/

Curriculum vitae

- Formation en économétrie (statistique, économie mathématique)
- Thèse de doctorat en Machine Learning (Apprentissage statistique)

- Enseignant chercheur, en poste à l'Université Lumière Lyon 2
- Spécialité : statistique, data mining et ses applications, informatique - Data Science
- Responsable du Master SISE (**S**tatistique et **I**nformatique pour la **S**cience des donné**E**s)

- « Père » des logiciels gratuits SIPINA v.3 et TANAGRA (open source)
- Auteur d'une dizaine d'ouvrages libres
- Auteur de près de 500 supports de cours et tutoriels en français et en anglais

- 650 visites par jour depuis le 1er février 2008 (Compteur Google Analytics)

Plan

1. Fonctionnalités des logiciels de data science
2. Panorama des logiciels open source
3. Projets POC sous R et Python
4. Conclusion

Fonctionnalités des logiciels de data science

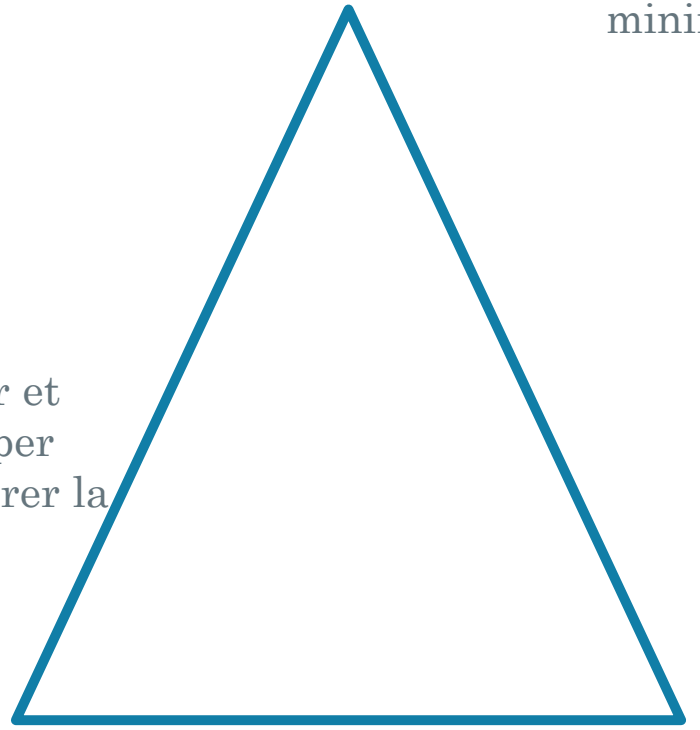
Qu'attendre aujourd'hui des logiciels de data science ?

Compétences du data scientist

STATISTIQUE

DATA MINING

Connaître et comprendre les techniques de modélisation, d'analyse de données, d'inférence… savoir exploiter les régularités « cachées » dans les données, pourvoyeuses de connaissances. Data mining, Machine Learning.



Maîtriser les outils pour accéder et manipuler les données, développer des stratégies nouvelles pour gérer la profusion de l'information,… Technologies big data

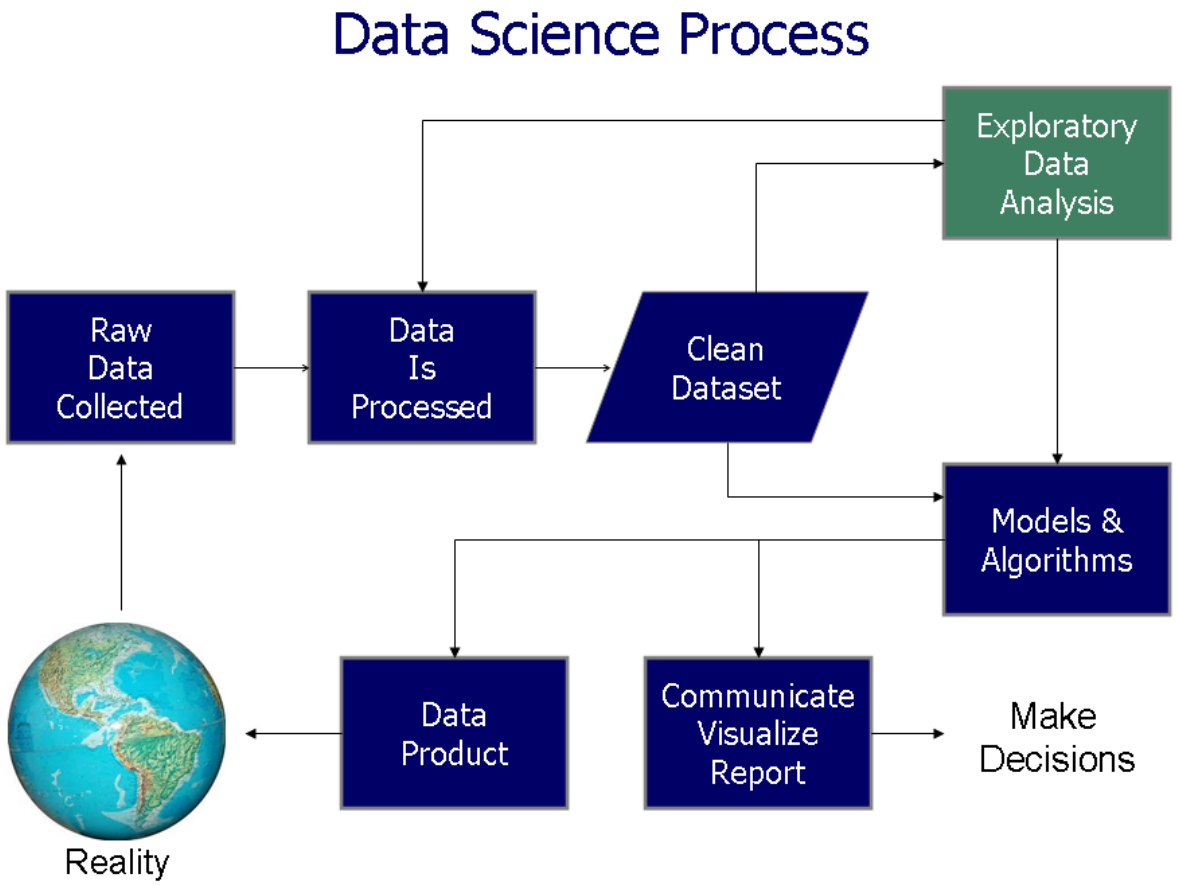
INFORMATIQUE

Toute analyse s'inscrit dans un domaine… qu'il faut connaître pour savoir décoder et exploiter les résultats

CONNAISSANCES METIER

Le logiciel joue un rôle très important !

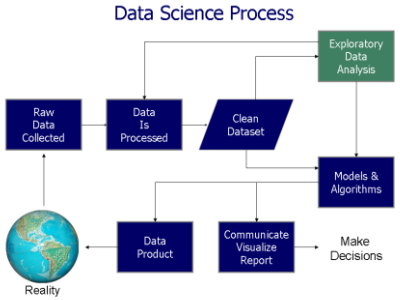
Démarche data science



(https://en.wikipedia.org/wiki/Data_science)

A chaque étape sont associées des tâches spécifiques que doivent assurer les outils / logiciels de data mining. Préparation des données, modélisation et présentation sont au cœur du métier de statisticien.

Critères pour les logiciels de data science

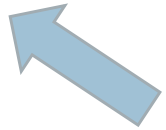


1. Architecture : stand-alone, client-serveur, via un navigateur, …
2. Mode opératoire : diagramme de traitements, langage de script, pilotage par menu,…
3. Performances, capacités de traitement, temps de calcul
4. Accès aux données : fichiers textes, Excel, accès aux bases de données,…
5. Solutions pour la volumétrie, technologies big data
6. Accès aux données non structurées et primitives de traitements (texte, image, …)
7. Interfaçage avec les API du web (ex. Twitter, Google+, OpenStreetMap, …)
8. Manipulation des données : transformations, recodage,…
9. Exploration graphique : représentations, visualisations, interactions,…
10. Bibliothèques de techniques de machine learning : supervisées, non-supervisées, …
11. Evaluation et comparaisons : comparaison des approches, benchmarking…
12. Reporting et solutions pour le déploiement (PMML,…)

Logiciels libres pour l'enseignement

Opportunité d'utiliser des logiciels libres pour l'enseignement du data mining (Déc. 2005).

- S'attacher au fond et non à la forme (cours de data mining)
- Utilisation ne nécessitant pas des compétences additionnelles spécifiques (ex. programmation)
- Former des étudiants qui vont sur le marché du travail



Cet aspect est très important, il ne faut pas nos choix impactent négativement les étudiants.

- La réponse à l'époque était OUI (pour les aspects méthodologiques), MAIS attention aux aspects opérationnels (ex. reporting, déploiement) – Logiciels testés : WEKA, ORANGE ML, TANAGRA
- Les conclusions ont évolué aujourd'hui, notamment avec R et Python.

Panorama des logiciels open source

Les études de KDnuggets et Gartner

Evaluation des logiciels

Comment étudier les outils de data science ?

- Faire un travail de recensement des outils
- Evolution de la popularité au fil du temps

Etude annuelle KDnuggets + Gartner

- Construire une grille d'évaluation (cf. critères)
- Positionner les outils sur la grille
- Réaliser une synthèse

Etudes de cas en faisant un focus sur R et Python

L'article de Goebel & Le Gruenwald (1999) a posé une trame maintes fois reprise.

Panorama des logiciels libres

On trouve ici ou là sur le web un travail de recensement sur les outils, mais l'exploration des caractéristiques reste superficielle.

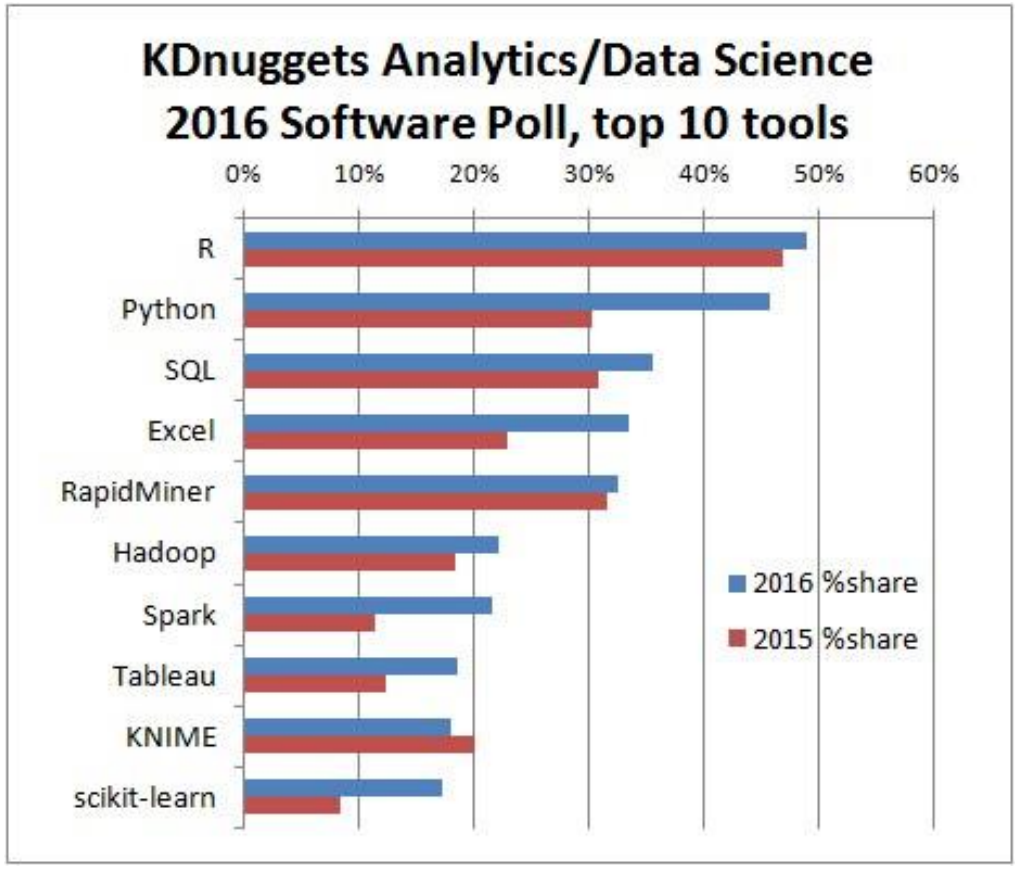
- Analytics Vidhya Content Team, « 18 Free Exploratory Data Analysis Tools for People who don't code so well », Septembre 2016.
- Hassani Parina, « Best 19 Free Data Mining Tools », Mars 2017.
- Kdnuggets, « Software Suites/Platforms for Analytics, Data Mining & Data Science ».
- Predictive Analytics Today, « 50 Top Free Data Mining Software ».
- Etc.

Les outils existent. Mais en connaître précisément les fonctionnalités passe par une étude approfondie qui n'est pas/peu disponible.

Une étude récente réalisée par les étudiants « Logiciels de Data Science », octobre 2016.

Enquête annuelle KDnuggets

La popularité peut être un indicateur de qualité (*pas toujours…*). On demande aux internautes d'indiquer les logiciels qu'ils utilisent (Mai 2016, 2895 votants choisissant parmi 102 outils différents) (2016 Software Polls Results)



R fait toujours figure de leader.

Python est en train de le rattraper

A voir l'enquête 2017…

Les résultats mélangent plusieurs concepts, mais l'information importante est le rôle prépondérant des deux outils leader que l'on retrouve dans les offres d'emploi en France (APEC - https://www.apec.fr/)

Gartner Magic Quadrant (2017)

Evaluation de 16 outils analytiques commerciaux (*pas tous*) sur la base de 15 critères. (Février 2017, Data Science Platforms: gainers and losers).

Codes de lecture (Gartner.com) :

Leaders execute well against their current vision and are well positioned for tomorrow.

Visionaries understand where the market is going or have a vision for changing market rules, but do not yet execute well.

Niche Players focus successfully on a small segment, or are unfocused and do not out-innovate or outperform others.

Challengers execute well today or may dominate a large segment, but do not demonstrate an understanding of market direction.

Focus sur R et Python

Etudes de cas – Projets étudiants POC (proof-of-concept) sur une période de 1 mois à 1 mois 1/2

Reconnaissance faciale

Démarche de recherche d'information par le contenu. Projet en Python.

Disposer d'une banque d'images

Extraction de caractéristiques

Matrice de description, ligne : individus, colonnes : caractéristiques.

X1	X2	X3	X4	X5

Recherche de similarités.

Extraction de caractéristiques

X1	X2	X3	X4	X5

Image « requête »

Vecteur de description de l'individu « requête »

Identification avec degré de fiabilité.

Analyse des offres d'emploi

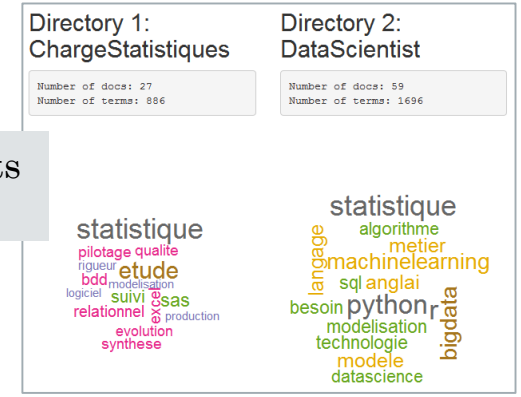
Analyse de documents textuels (text mining) et classement / classification. Projet sous R (Shiny)

Offres d'emploi qui ont été étiquetées manuellement.

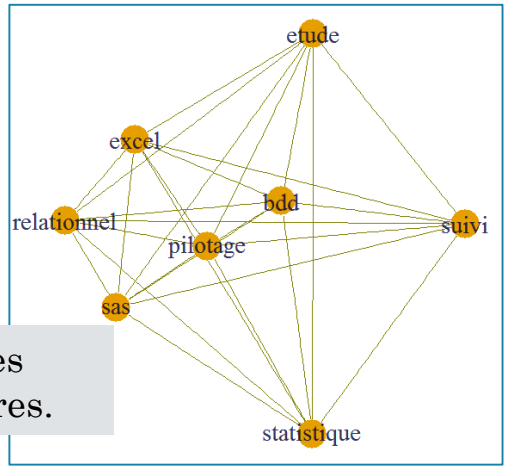
Analyse et développement d'un application Shiny

Métiers : Chargés d'études statistique, consultant BI, data analyst, data engineer, data manager, data miner, data scientist, data visualisation

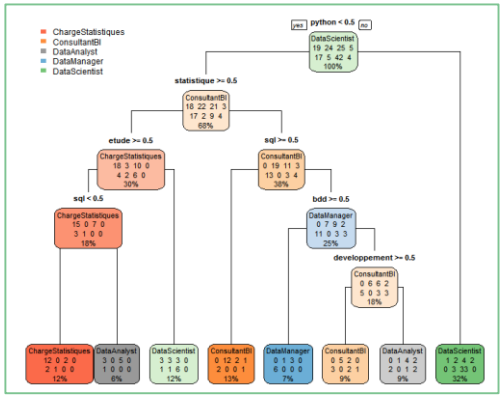
Mots clés fréquents selon les métiers



Association entre les termes dans les offres.



Identification des métiers selon les termes de l'offre.



Reconnaissance musicale

Démarche de recherche d'information par le contenu. Projet en Python.

Disposer d'une banque de musiques

Extraction de caractéristiques

Matrice de description, ligne : chansons, colonnes : caractéristiques.

X1	X2	X3	X4	X5

Matching + recherche de similarités.

Extraction de caractéristiques

X1	X2	X3	X4	X5

Chanson « requête »

Vecteur de description de la chanson « requête »

Identification et recommandation

Conclusion et bibliographie

Conclusion

Les logiciels « open source » de Data Science, en particulier R et Python, proposent des fonctionnalités et des performances opérationnelles. Le système des packages permet de les enrichir à l'infini.

Mais

- La facilité d'utilisation (utilisabilité) n'est pas toujours au rendez-vous. Il faut une formation spécifique pour savoir réellement en tirer parti.
- L'absence de support (direct) peut être source de difficultés. Il y a bien les forums, etc., mais discerner ce qui nous est vraiment utile prend du temps.

Bibliographie - Webographie

Goebel M., Gruenwald L., « A survey of data mining and knowledge discovery software tools », ACM SIGKDD Explorations, 1(1), June 1999.

Un des premiers articles populaires ayant posé les bases de la comparaison de logiciels de data mining.

Master SISE, « Etude des logiciels de data science », octobre 2016.

Présentation, études de cas sous forme de scénarios de TD, tutoriels sur Youtube.

Piatetsky G., « R, Python Duel As Top Analytics, Data Science Software », KDnuggets 2016 Software Poll Results, June 2016.

Enquête annuelle, évolutions, comparaisons avec les années précédentes.

Piatetsky G., « Gartner 2017 Magic Quadrant for Data Science Platforms: gainers and losers », KDnuggets, February 2017.

Enquête annuelle, évolutions, comparaisons avec les années précédentes.