Machine Learning – Outils pour l'enseignement

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 et informatique, data mining et ses applications - Data Science

- « Père » des logiciels gratuits SIPINA v.3 et TANAGRA (open source)
- Auteur d'une dizaine d'ouvrages libres
- Auteur de **plus de 500** supports de cours, de tutoriels en français et en anglais
- 630 sessions par jour sur plus de 10 années (depuis 01.02.2008 - Compteur Google Analytics)

- Responsable du Master SISE (**S**tatistique et **I**nformatique pour la **S**cience des donné**E**s)

Interrogations…

Peut-on « se contenter » de logiciels libres pour l'enseignement du machine learning dans nos formations ?

Pédagogie

Praticabilité dans les enseignements

Richesse fonctionnelle

Insertion professionnelle des étudiants

Si oui, lesquels de logiciels libres ? Et pourquoi ?

R et Python

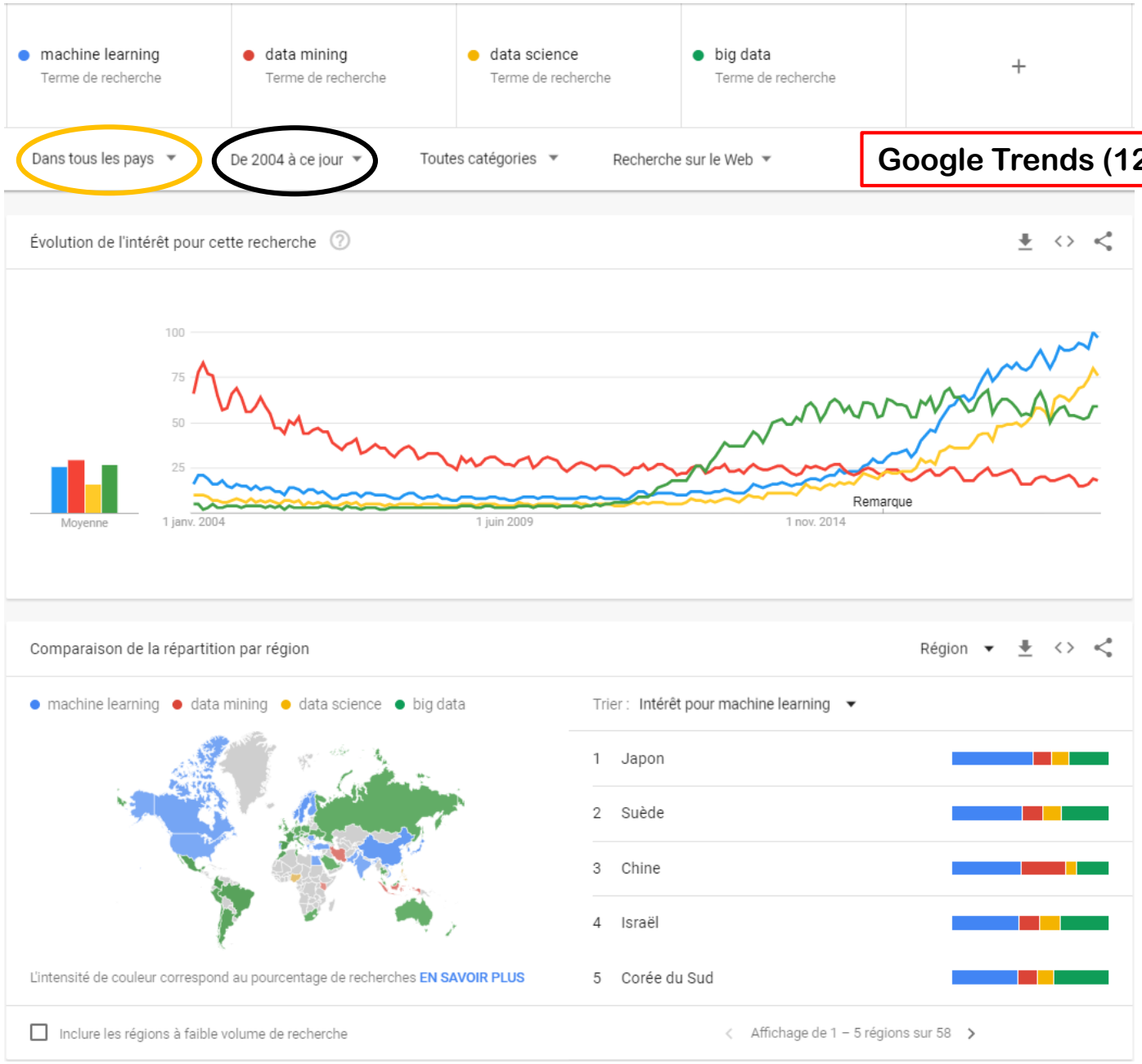
Plan

1. Machine Learning, Data Mining, Data Science…
2. Logiciels – Attentes pédagogiques
3. Réalisations étudiantes sous R et Python
4. Conclusion

Machine Learning

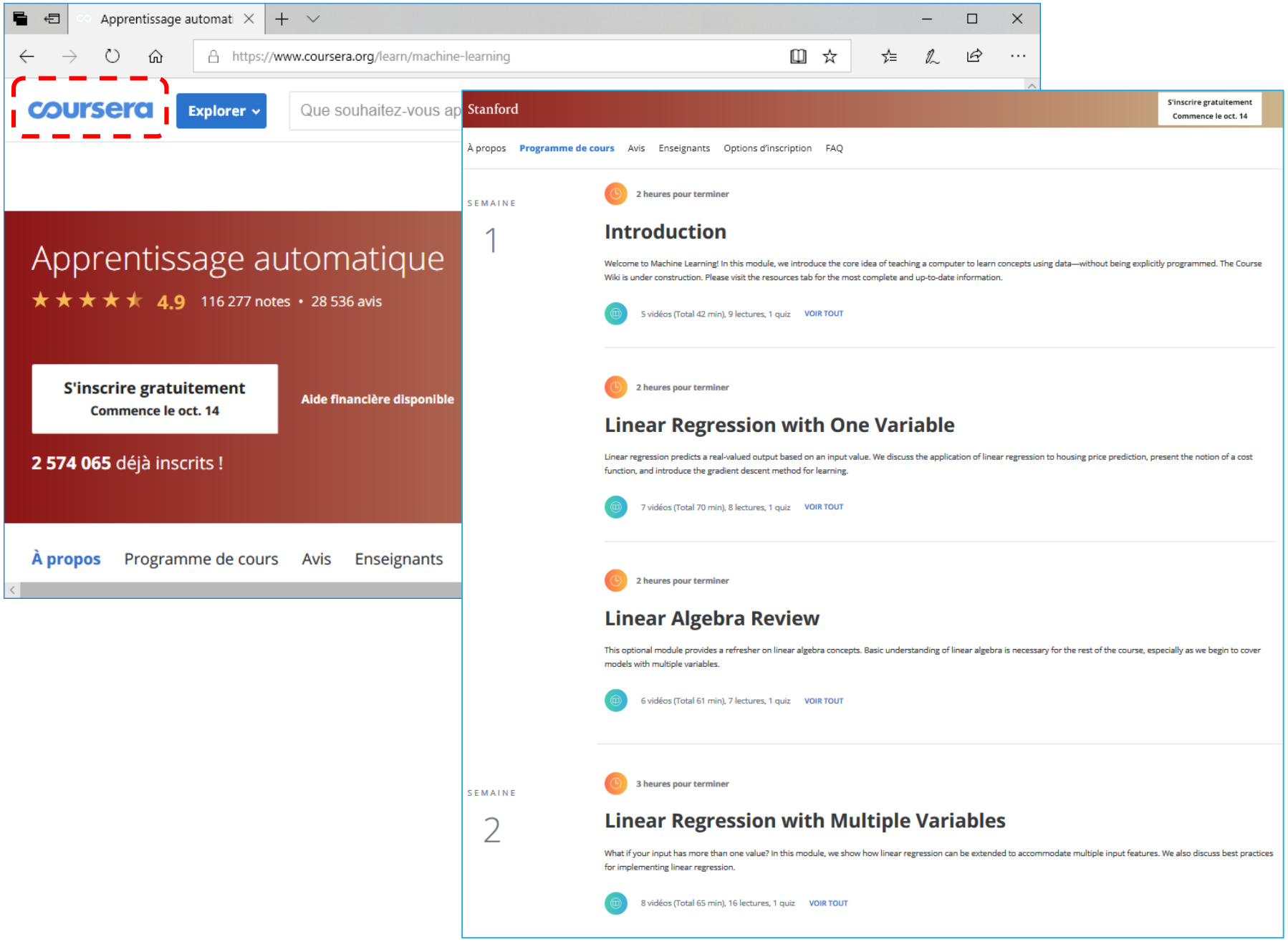
Faire du neuf avec du vieux ? Big Data Analytics

Machine Learning – Regain de popularité récent



« Machine Learning » (apprentissage automatique, apprentissage machine, apprentissage statistique), branche de l'intelligence artificielle, qui trouve ses origines dans les années 40 - 50 (Wikipédia).

Machine Learning vs. Statistique



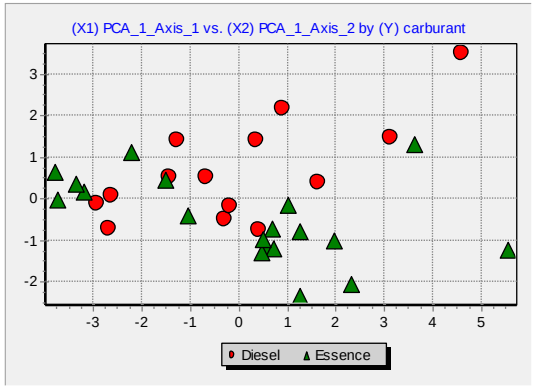
« On aimerait faire de l'IA et du Machine Learning… où il est question de convolutions et de gradient… avec du traitement d'images…»

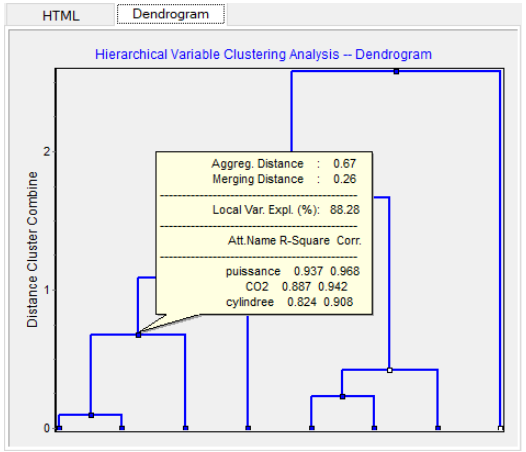
Dépasser les chapelles de naguère : statistique, économétrie, analyse de données, apprentissage automatique….

Statistique – Statistique exploratoire

Les données sont spécifiquement recueillies à des fins d'étude (ex. enquête, expérimentations, etc.)

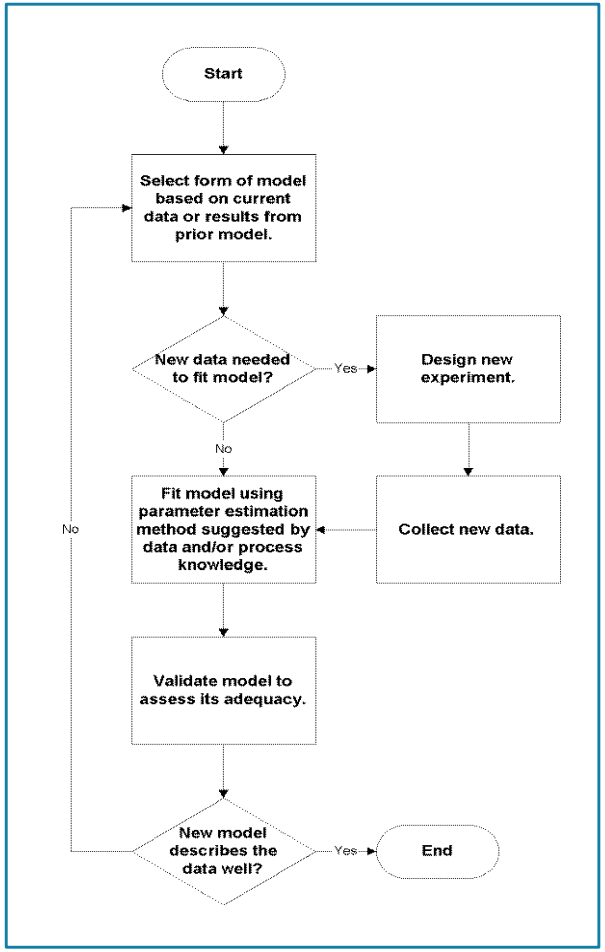
- Bonne qualité souvent
- Faible volumétrie (rareté)





Volume de traitements – de toute manière – limité par les capacités des outils informatiques disponibles (à l'époque).

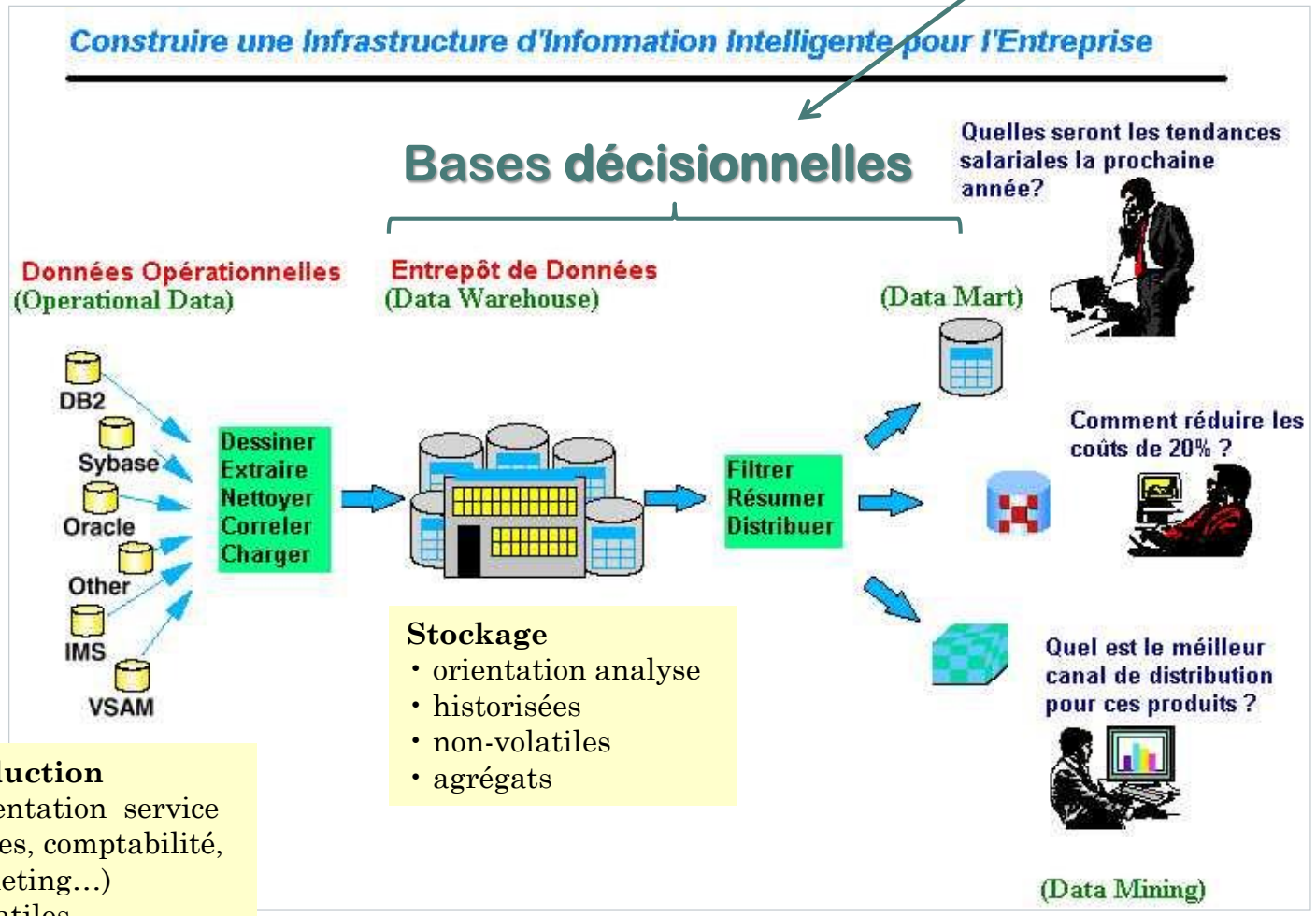
Application des techniques de modélisation et de statistique



NIST – e-Handbook of Statistical Methods

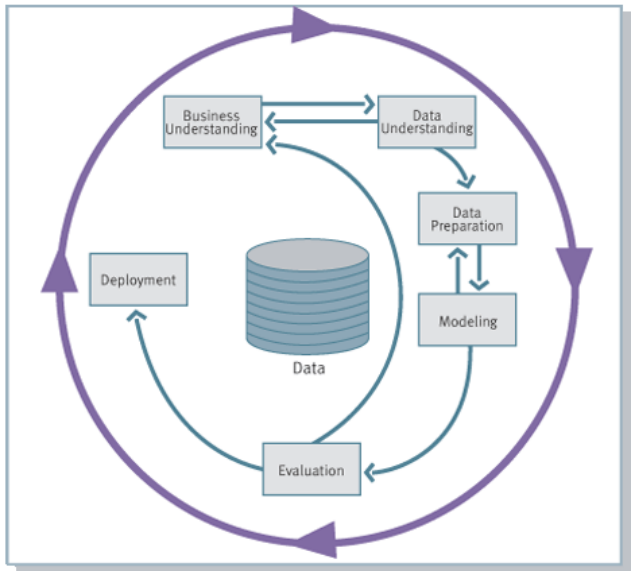
Vague « Data Mining » à partir de la fin des années 90

Les données sont organisées et stockées de manière à ce que nous puissions mener des analyses.



Production
- orientation service (ventes, comptabilité, marketing…)
- volatiles

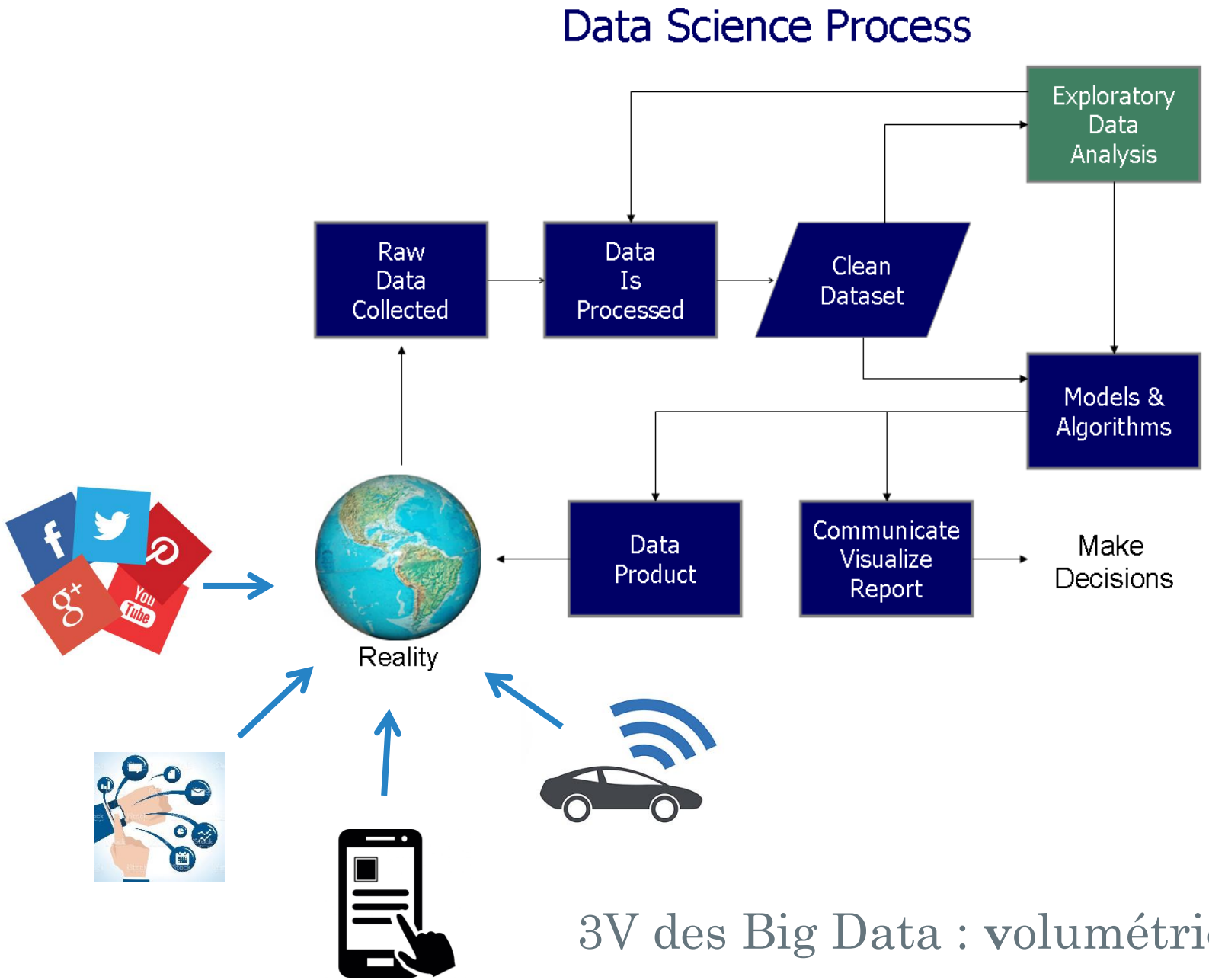
Stockage
- orientation analyse
- historisées
- non-volatiles
- agrégats



CRISP-DM

Les sources d'information et les technologies évoluent. Elément clé : l'entrepôt de données.

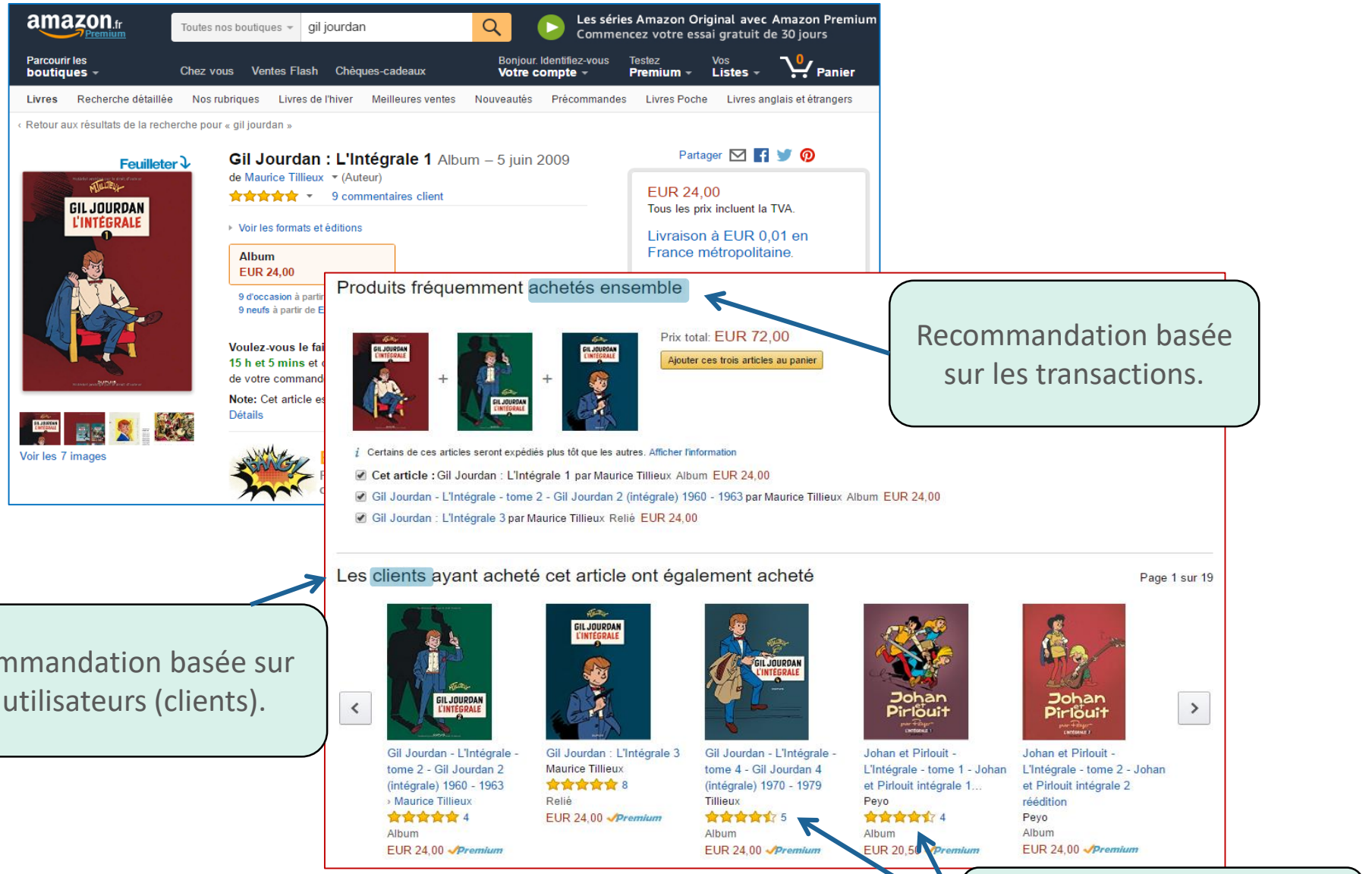
Vague « Big Data Analytics » (Data Science) actuelle



Sources additionnelles d'information externes à l'entreprise, multiplicité des formats, nouveaux enjeux technologiques pour le stockage et le traitement (stockage NoSQL, data lake, tech. cloud, info. distribuée [Hadoop, Spark]...) et nouvelles opportunités d' analyse ! (cf. Rapports Lauvergeon, Villani)

3V des Big Data : volumétrie, variété, vélocité.

Exemple. Filtrage collaboratif, Systèmes de recommandation.



Attentes vis-à-vis des logiciels de data science / machine learning

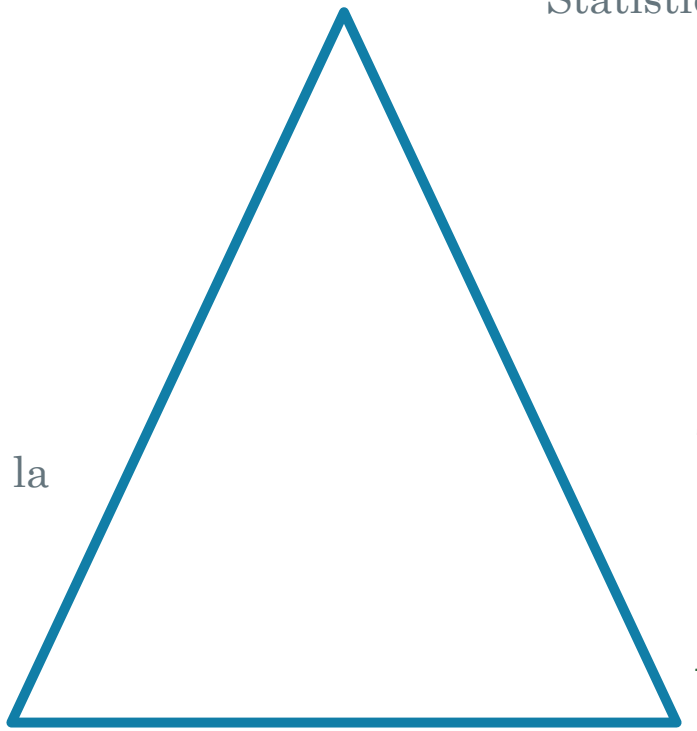
Qu'attendre aujourd'hui des logiciels pour l'enseignement ?

Enseignement de la data science

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. Statistique, 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 d'application : données de sécurité, données du web, analyse des réseaux sociaux, etc.

APPLICATIONS

Le logiciel joue un rôle très important !

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 (hadoop, spark,…)
6. Accès aux données non structurées et primitives de traitements (texte, image, …)
7. Interfaçage avec les API du web (ex. Twitter, Google Analytics, OpenStreetMap, …)
8. Manipulation des données : transformations, recodage,…
9. Exploration graphique : représentations, visualisations, interactions,…
10. Bibliothèques de techniques de machine learning : supervisées, …, deep learning,…
11. Evaluation et comparaisons : comparaison des approches, benchmarking…
12. Reporting et solutions pour le déploiement (PMML,…)

Logiciels pour l'enseignement – Objectif pédagogique des TD sur machine

L'utilisation du logiciel doit s'inscrire dans une démarche pédagogique. Deux extrêmes à éviter : Qu'est-ce qu'il faut faire là ? vs. On recopie sans rien comprendre… (cf. Fiches de TD en Master)

Objectifs d'une séance de TD sur machine :

- Mettre en œuvre une technique de machine learning vue en cours c.-à-d.
  - Charger les données
  - Lancer l'algorithme en respectant un schéma prédéfini (ex. app-test en prédictif)
  - Savoir lire et inspecter / expertiser les résultats
  - Déployer les modèles
- Plus loin : apprendre à manipuler les paramètres des algorithmes
- Plus loin encore : s'intéresser aux problèmes pratiques (recodage, données manquantes, comparaison des approches, solutions pour volumétrie, etc.)

Logiciels pour l'enseignement – Attente vis-à-vis des outils utilisés par nos étudiants

Nous formons des étudiants qui vont en entreprise. Il ne faut pas que nos choix les impactent négativement en les emmenant sur des voies de garage.

Une utilisation conforme aux standards du domaine et répondant aux objectifs pédagogiques :

- S'attacher au fond et non la forme (la base : charger les données [aux formats usuels], appliquer les méthodes, restituer les résultats. Cf. les objectifs d'une séance de TD).
- Pas de mode opératoire spécifique, nécessitant des compétences supplémentaires pouvant parasiter le discours (Programmation vs. Diagramme de traitement).
- Démarche harmonisée pour un large spectre d'applications (même environnement pour le machine learning « classique », le text mining, image mining, etc.). Ne pas multiplier les outils en fonction de la matière étudiée. Intérêt des dispositifs à packages.

Logiciels libres pour l'enseignement - La question de R et Python

Question ancienne : opportunité d'utiliser des logiciels libres pour l'enseignement du data mining (Déc. 2005).

OUI pour les aspects méthodologiques, MAIS attention aux aspects opérationnels (ex. reporting, déploiement) [WEKA, ORANGE ML, TANAGRA], et *pas vocation à être utilisés en entreprise.*

Aujourd'hui, R et Python s'imposent dans nos formations parce que :

- Gratuits, totalement, ce n'est pas moindre de leurs qualités.
- Forte pénétration de ces outils dans les entreprises (cf. les offres d'emploi sur le site de l'APEC, avec les mots-clés « statistique », « data science », etc.).
- Le niveau en programmation des étudiants a considérablement évolué. Et on peut s'appuyer sur ces outils pour l'améliorer encore.
- Outils polyvalents. Très large spectre d'utilisation.

Compétence sur ces outils devient un marqueur fort dans les CV.

Sondage annuel KDnuggets (2006, 2013, 2019)

Les appellations changent au fil des années…

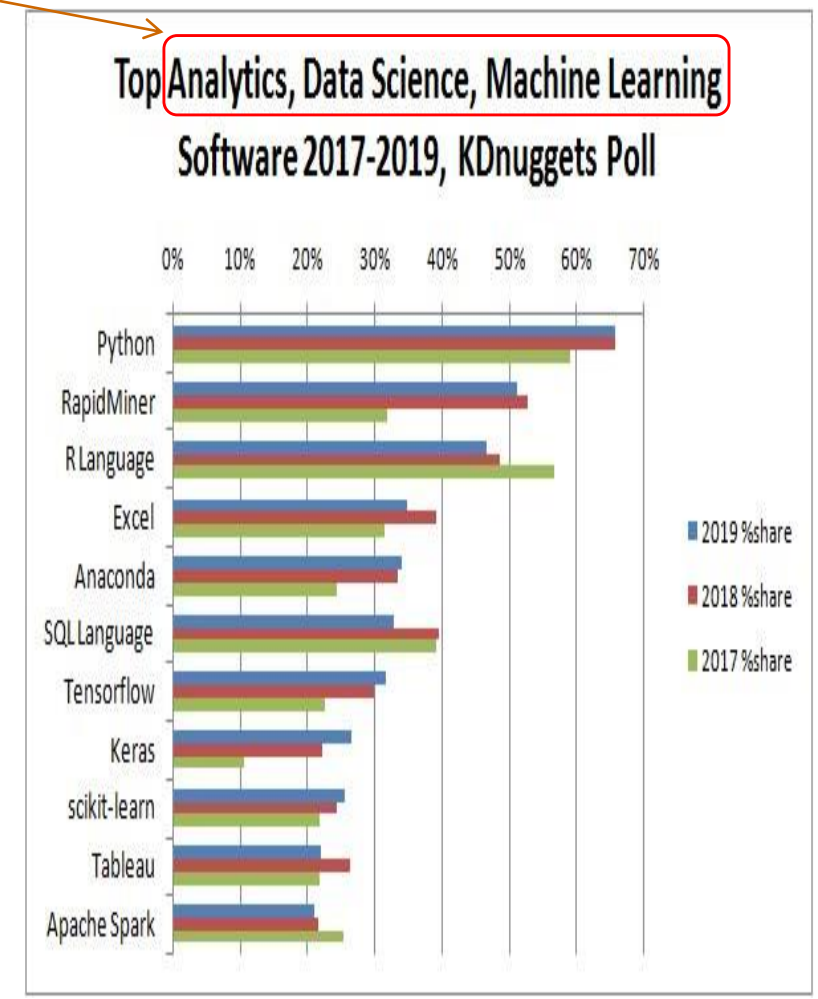
Poll

Data mining/analytic tools you used in 2006: [561 voters]

CART/MARS/TreeNet/RF	159 (72 alone)
SPSS Clementine	127 (47 alone, 46 with SPSS)
SPSS	100 (5 alone, 46 with Clementine)
Excel	100 (3 alone)
KXEN	90 (75 alone)
your own code	77 (1 alone)
SAS	72 (3 alone, 13 with E-Miner)
Weka	62 (7 alone)
R	53 (5 alone)
MATLAB	41 (5 alone)
other free tools	39 (3 alone)
SAS E-Miner	37 (9 alone, 13 with SAS)
SQL Server	32 (3 alone)
other commercial tools	31 (7 alone)
Oracle Data Mining	20 (13 alone)
Insightful Miner/ S-Plus	20 (0 alone)
C4.5/C5.0/See5	18 (0 alone)
Megaputer	16 (14 alone)
Statsoft Statistica	13 (2 alone)
Angoss	8 (2 alone)
Mineset (PurpleInsight)	5 (3 alone)
Eudaptics Viscovery	5 (3 alone)
Xelopes	4 (4 alone)
Visumap	3 (2 alone)
IBM I-miner	3 (0 alone)
Equbits	3 (3 alone)

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [1880 voters]

Legend: Red: Free/Open Source tools Green: Commercial tools	% users in 2013 % users in 2012
Rapid-I RapidMiner/RapidAnalytics free edition(737), 30.9% alone	39.2% 26.7%
R(704), 6.5% alone	37.4% 30.7%
Excel(527), 0.9% alone	28.0% 29.8%
Weka / Pentaho(269), 5.6% alone	14.3% 14.8%
Python with any of numpy/scipy/pandas/iPython... packages(250), 0% alone	13.3% 14.9%
Rapid-I RapidAnalytics/RapidMiner Commercial Edition(225), 52.4% alone	12.0%
SAS(202), 2.0% alone	10.7% 12.7%
MATLAB(186), 1.6% alone	9.9% 10.0%
StatSoft Statistica(170), 45.9% alone	9.0% 14.0%
IBM SPSS Statistics(164), 1.8% alone	8.7% 7.8%



Il y a matière à réflexion quand on regarde cette évolution

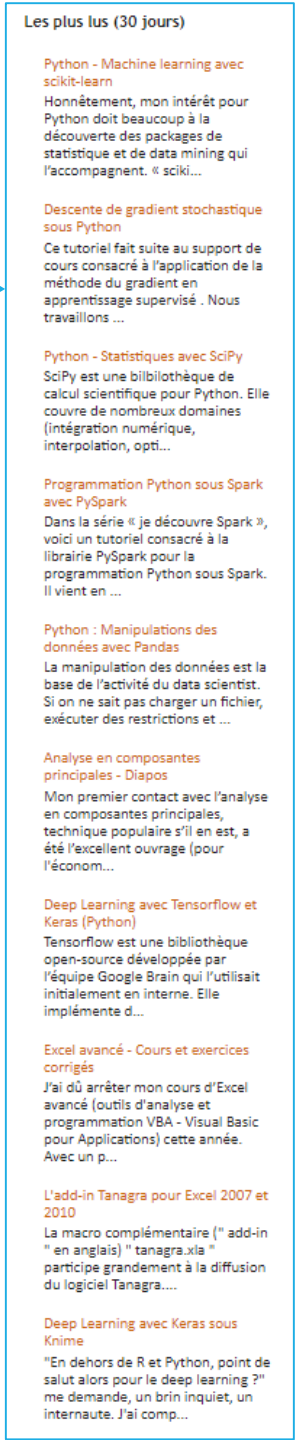
Site de Tutoriel Tanagra
au 13 oct. 2019

Constat à ce jour :

- L'intérêt exacerbé pour Python (cf. mon site des tutoriels)
- Dans l'esprit des entreprises en France : Machine Learning = Python (cf. démo de l'étude sous R-Shiny, les stages encadrés, enfin les offres d'emploi sur l'APEC)
- La « lenteur » de R est souvent décriée (mais est-ce vraiment justifié…)

Ce que je pense :

- Selon le thème, le contexte, les objectifs… et les packages disponibles, on peut préférer l'un ou l'autre
- On peut facilement développer une compétence conjointe forte pour R et Python



UE Informatique appliquée (9 ECTS)

- Programmation Statistique sous R (R. Rakotomalala)

Apprentissage de la programmation sous R. Structures avancées. Programmation des algorithmes de statistique et de data mining sous R. Modèle objet sous R. Programmation big data (map reduce) sous hadoop. Programmation R sous spark. Création de packages.

M2 SISE 2019-2020

- Machine Learning sous Python (N. Sawadogo)

Bases de la programmation python, structures vectorielles et matricielles. Algorithmes de machine learning d'apprentissage supervisé et non supervisé (svm - support vector machine, dbscan, birch,…). Image mining, traitement des données images. Projets de ces dernières années : reconnaissance faciale, reconnaissance et recommandation musicale, programmation d'un chatbot.

Et les outils payants ?

Positionnement des outils analytiques qui proposent (au moins) une licence commerciale - Gartner Magic Quadrant for Data Science and Machine Learning Platforms, KDnuggets, Février 2019.

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.

Parlons-en de KNIME…

Etudes de cas

Sous R et Python – Projets POC effectués par les étudiants

Traitements amusants sur le web

Reconnaissance d'objets
avec Google Image

https://www.google.fr/imghp

Détection de l'âge avec
la technologie Microsoft

https://www.how-old.net/#

Analyse des offres d'emploi

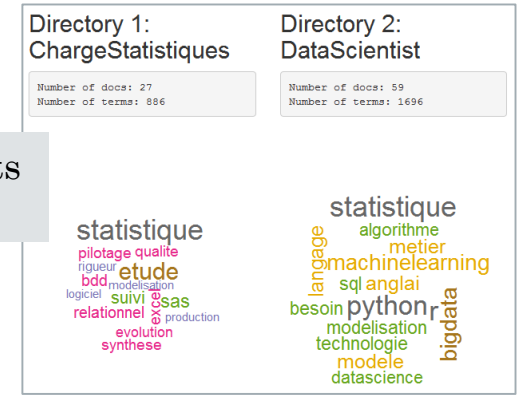
Analyse de documents textuels (text mining) et classement / classification. Projet sous R (Shiny)

Offres d'emploi qui ont été étiquetées manuellement.

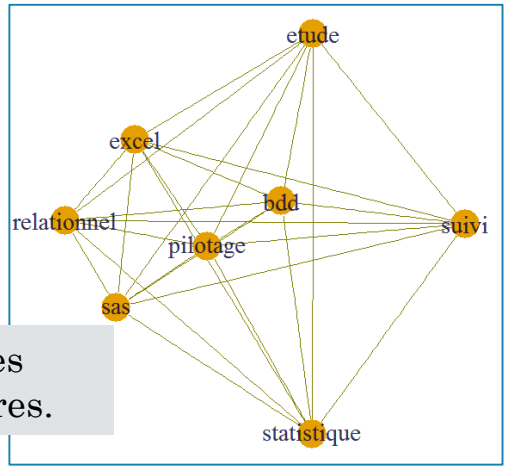
Métiers : Chargés d'études statistique, consultant BI, data analyst, data engineer, data manager, data miner, data scientist, data visualisation

Analyse et développement d'un application Shiny

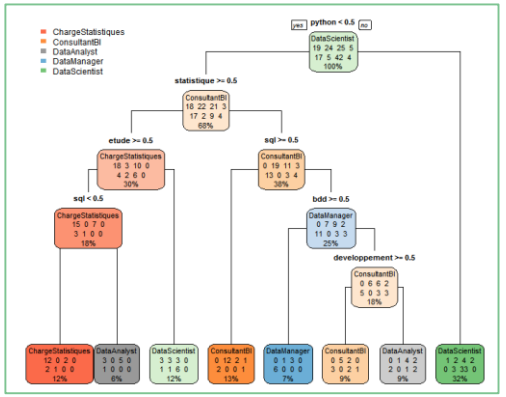
Mots clés fréquents selon les métiers



Association entre les termes dans les offres.



Identification des métiers selon les termes de l'offre.



Reconnaissance faciale

Démarche de recherche d'information par le contenu. Projet en Python (2016, 2019).

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

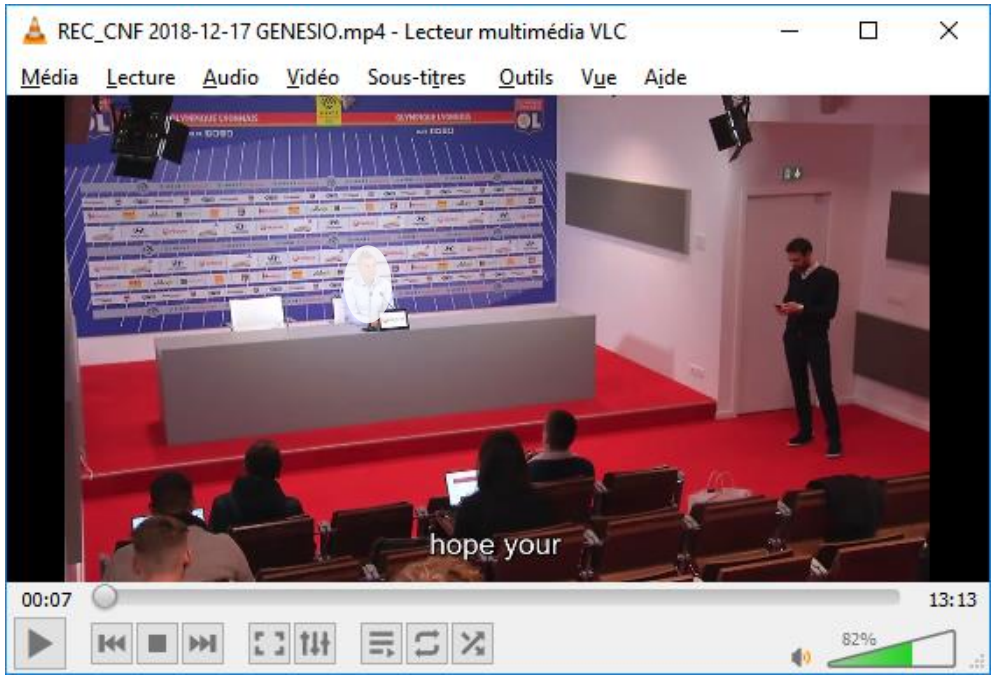
X1	X2	X3	X4	X5

Image « requête »

Vecteur de description de l'individu « requête »

Identification avec degré de fiabilité.

Traduction automatique de vidéos



Traduction automatique de vidéos

FR pour sous-titrage en anglais

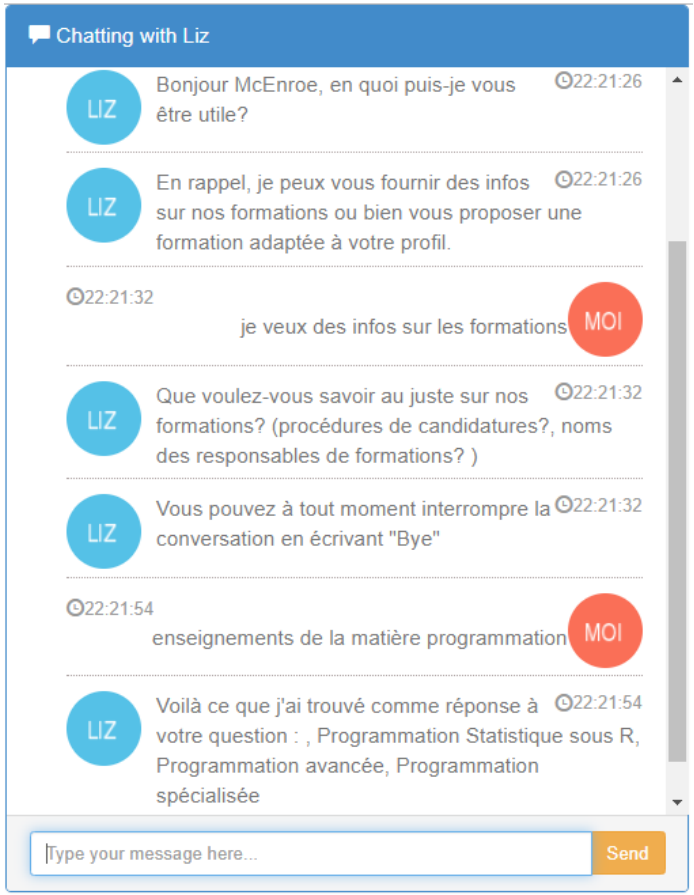
Speech-to-text

- Gestion du niveau de bruit
- Découpage de la bande son pour extraction des phrases

Utilisation d'outils de traduction

- Choix du plus performant
- Apprentissage du vocabulaire spécifique au domaine

Programmation d'un chatbot



Un chatbot pour aiguiller les étudiants dans leurs choix de formation

Traduire les questions formulées en langage naturel sous forme de requêtes SQL, avec identification des mots-clés, gestion des synonymes, des stopwords, etc.

Utilisation du framework Flask pour une architecture C/S.

Conclusion et bibliographie

Conclusion

R et Python jouent actuellement - tous deux - un rôle essentiel dans l'enseignement du Machine Learning.

Parce que

- Répondent parfaitement aux objectifs pédagogiques
- Permettent de développer des compétences de programmation
- Répondent aux attentes en matière de traitements – IA, Machine Learning – dans les entreprises (cf. les stages SISE)
- Et sont (de ce fait) explicitement spécifiées dans les offres d'emploi

Et…

Développer une expertise élevée dans les deux outils / langages n'implique pas un coût pédagogique additionnel fort.

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.

Ateliers du Master SISE, « Logiciels de Data Science », oct. 2016 ; « Outils de la Data Science », oct. 2017 ; « Outils de la Dataviz », oct. 2018.

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

Piatetsky G., « Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis », KDnuggets 2019 Software Poll Results, May 2019.

Enquête annuelle, évolutions, comparaisons avec les années précédentes.

Piatetsky G., « Gainers, Losers, and Trends in Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms », KDnuggets, February 2019.

Enquête annuelle, évolutions, comparaisons avec les années précédentes.