Logiciels pour l'enseignement de la statistique

Ricco Rakotomalala

Master SISE (Statistique et Informatique pour la Science des donnéEs)

Université Lumière Lyon 2

https://www.youtube.com/@master2sisedatascience

Curriculum vitae

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

- 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'ouvrages (~10) et de supports de cours (~100) (https://cours-machine-learning.blogspot.com/)
- Près de 700 tutoriels en français et en anglais (https://tutoriels-data-science.blogspot.com/)
- Chaîne YouTube – Support pédagogique (https://www.youtube.com/@master2sisedatascience)

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

Interrogations…

Quels logiciels pour l'enseignement des statistiques (traitements statistiques des données, data mining, machine learning, data science, …) dans nos formations ?

Pédagogie

Praticabilité dans les enseignements

Richesse fonctionnelle

Insertion professionnelle des étudiants

La question des licences. Peut-on se contenter des logiciels libres ?

R et Python

Qu'attendent de nous les étudiants aujourd'hui ? (à l'ère des smartphones où tout doit arriver très vite, immédiatement…)

Capter leur intérêt, besoin de challenge, susciter de l'enthousiasme …

Ecueils d'une séance de TD-machine ratée…

Qu'est-ce qu'il faut faire là ?

On recopie le code mais on ne comprend rien là en fait…

C'est (vous êtes) sympa mais ça sert à quoi ce qu'on fait ?

Qui utilisent ces outils en entreprise ? (même R, j'ai dû me justifier… KDNuggets Polls, 2006)

Avec tel outil (tel package), on le fait très facilement… (TD sous Excel/Tableur… APEC)

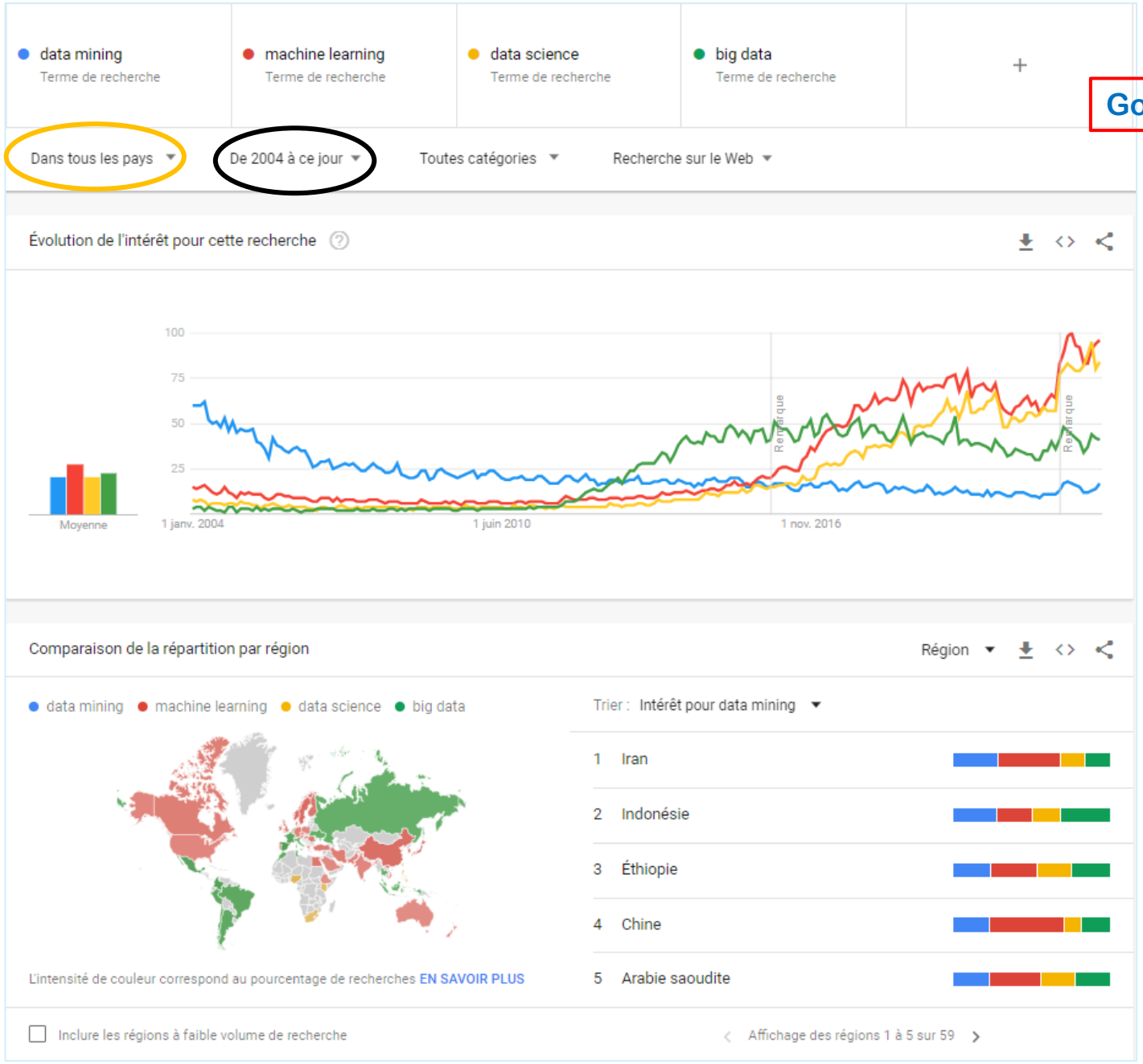
Plan

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

Statistique, Machine learning, Data Science

Faire du neuf avec du vieux ? Big Data Analytics

Phénomènes de « mode », même dans nos domaines

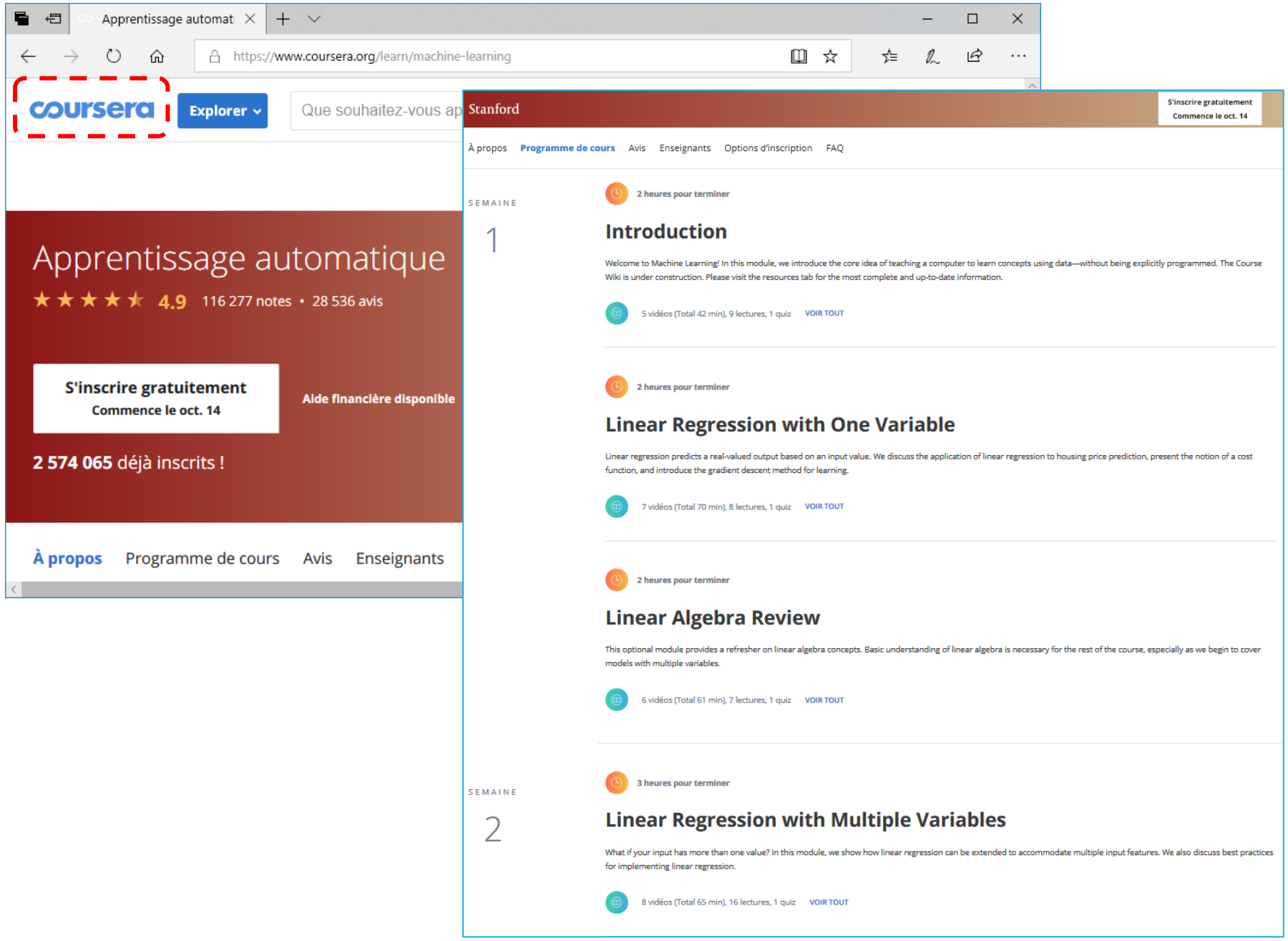


Google Trends (16 nov. 2022)

Mais finalement, l'idée maîtresse est la valorisation des données via des techniques de traitement statistique (au sens large : exploration, modélisation, apprentissage, ….).

Cf. Rapport Lauvergeon

Machine Learning vs. Statistique



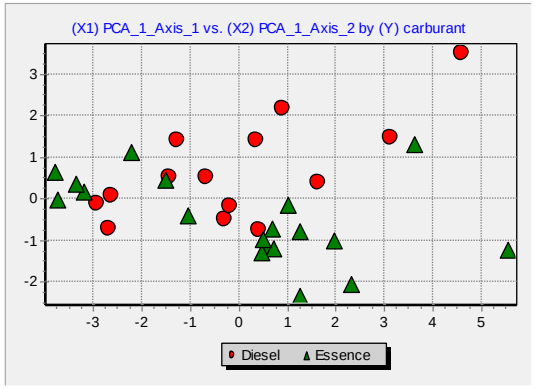
« On aimerait faire de l'IA et du Machine Learning… où il est question de convolutions et de gradient… avec du traitement d'images…»

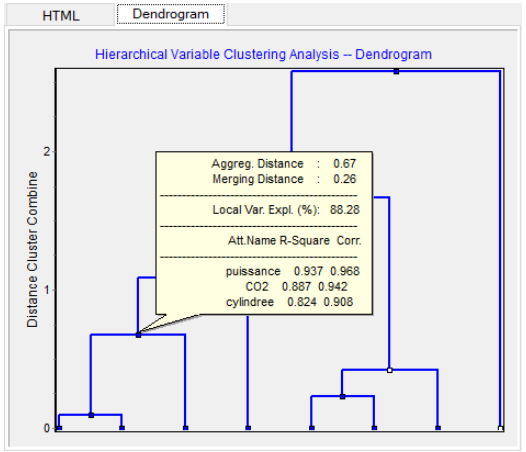
Dépasser les chapelles de naguère : statistique, économétrie, analyse de données, apprentissage automatique, traitement d'images….

Statistique – Statistique exploratoire – Analyse des données

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

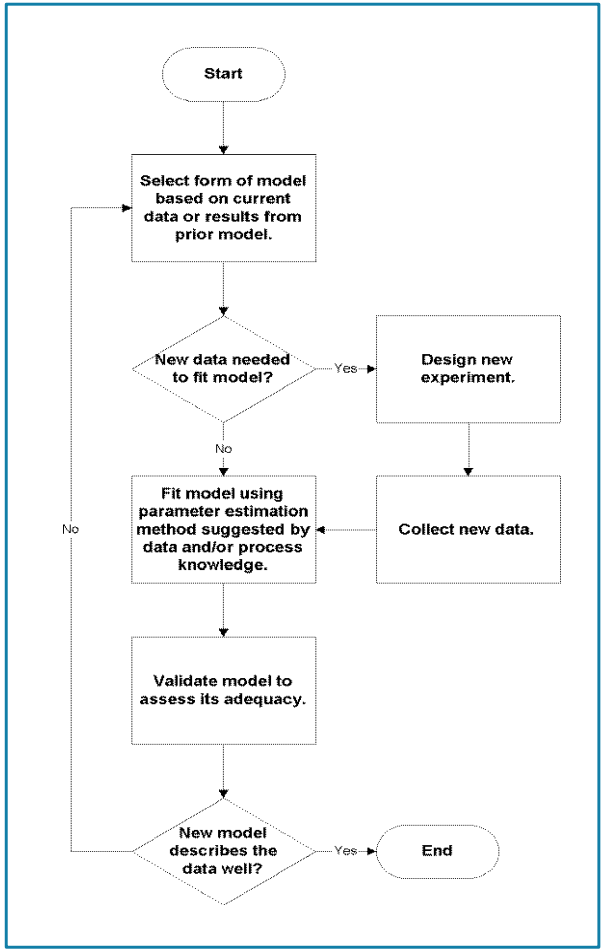
- Bonne qualité souvent
- Faible volumétrie (rareté)





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

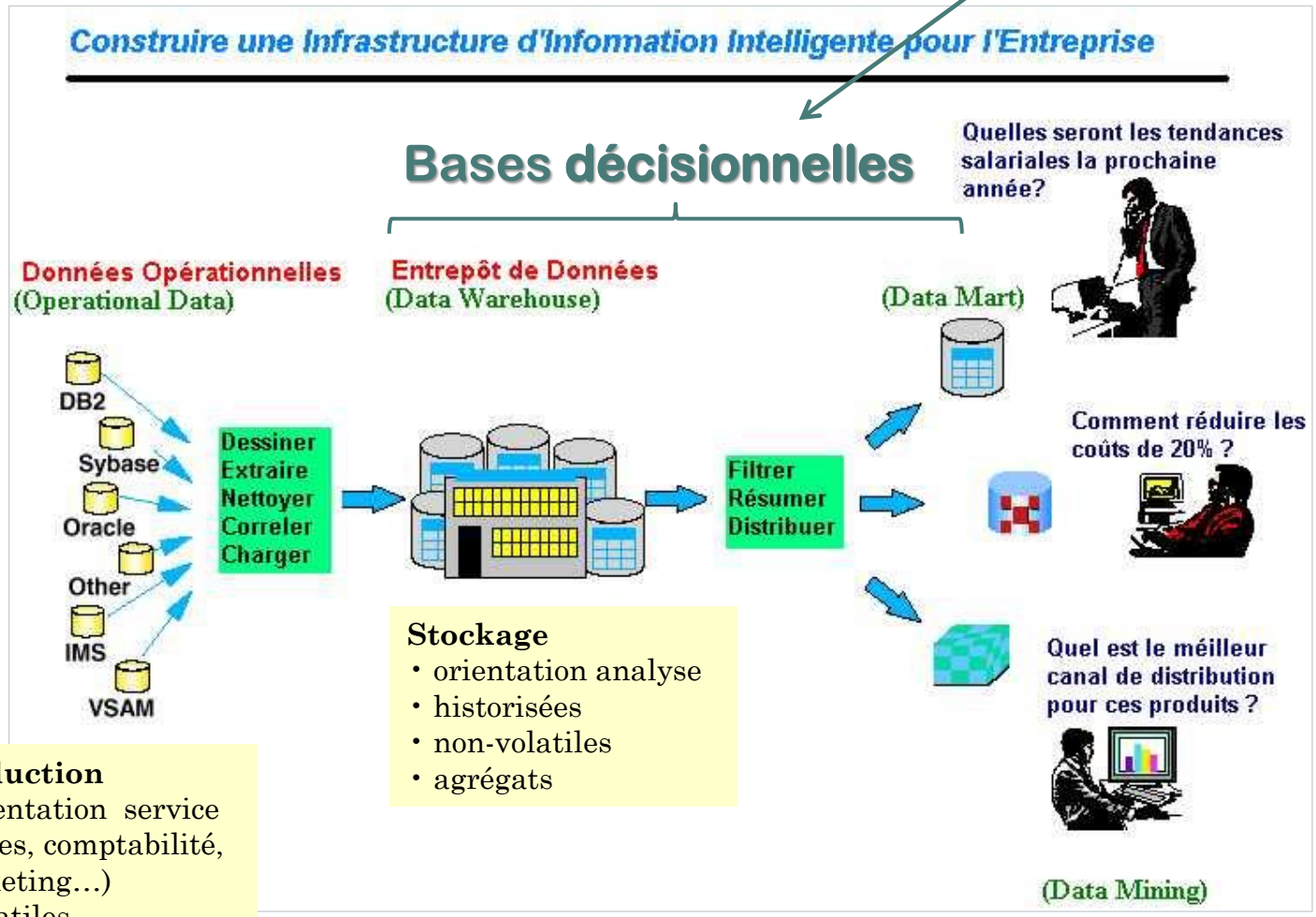
Application des techniques de modélisation et de statistique



NIST – e-Handbook of Statistical Methods

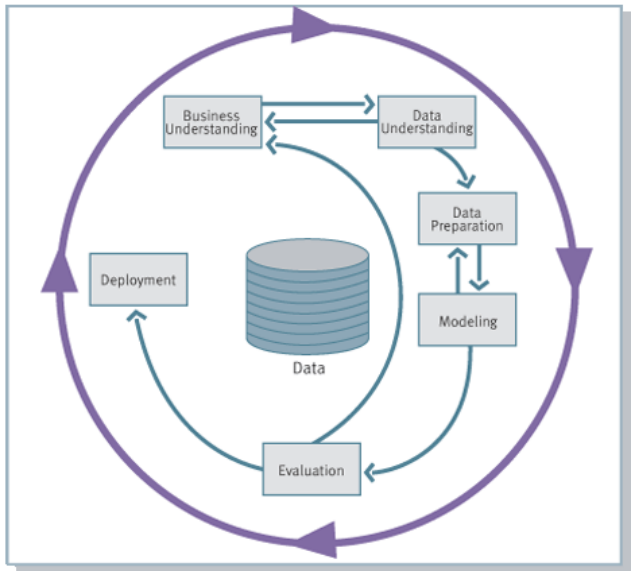
Vague « Data Mining » à partir de la fin des années 90

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.

Fiches métiers de la « data » - APEC

(https://www.apec.fr/tous-nos-metiers/informatique/)

DATA ENGINEER F/H

Le/la data engineer est un développeur informatique qui a pour mission de mettre en place la collecte et la mise à disposition des données au sein de l'entreprise. Il/elle est également en charge d'industrialiser et mettre en production des traitements sur les données (par exemple : mise à disposition de tableaux de bord, intégration de modèles statistiques) en lien avec les équipes métiers et les équipes qui les analysent.

DATA ANALYST F/H

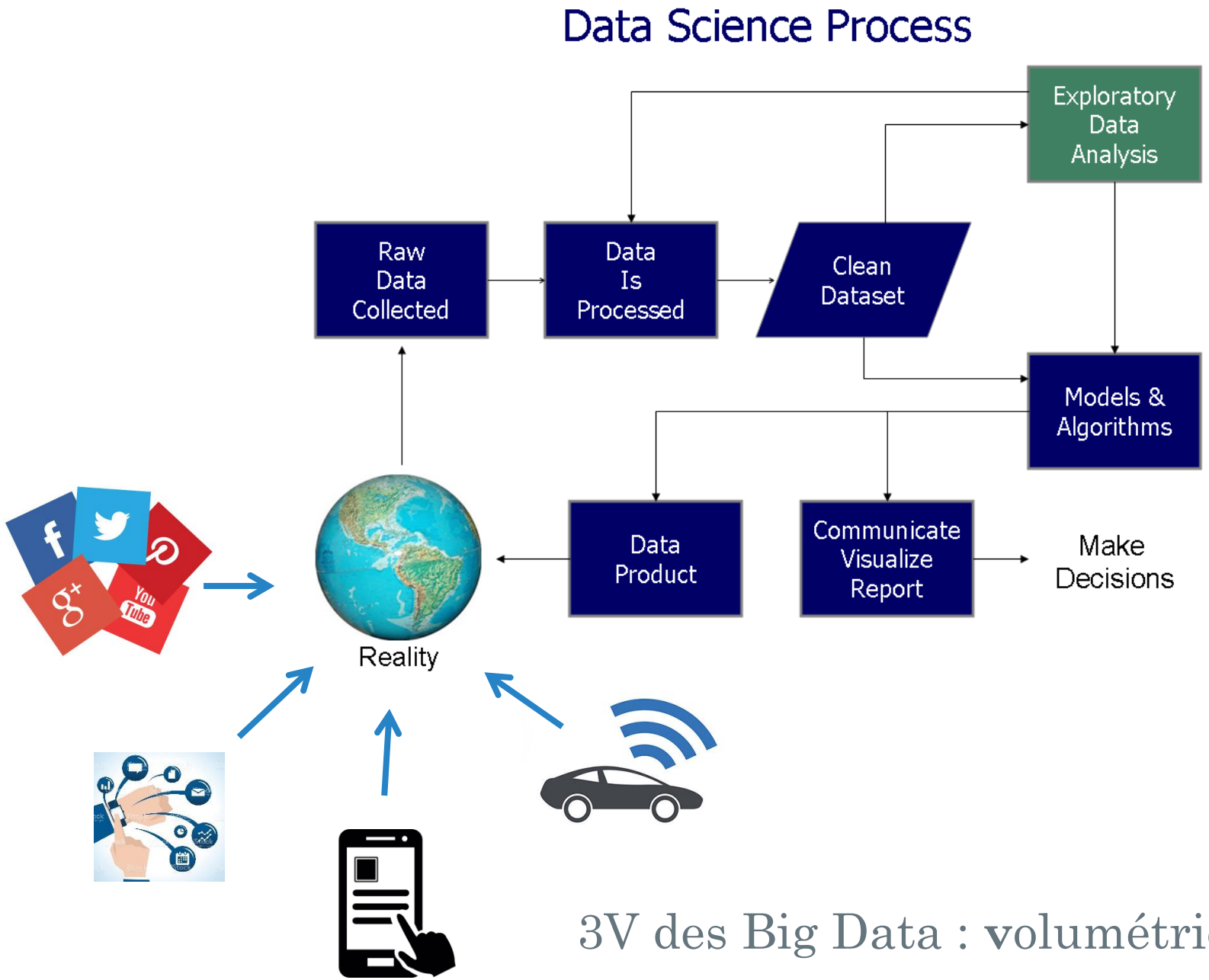
Le/la data analyst valorise l'ensemble des données d'une entreprise pour en faire un levier de création de valeur. Il/elle utilise notamment les données recueillies en masse (big data) pour réaliser les nombreux tableaux de bord nécessaires à différents services de l'entreprise (marketing, relations clients, production...). Il/elle est également en charge de construire des modèles statistiques pour éclairer les services opérationnels (segmentations clients ou analyses prédictives).

DATA SCIENTIST F/H

Le/la data scientist développe des algorithmes d'apprentissage automatique selon les besoins des équipes métiers. Ses compétences en statistiques lui permettent de construire des modèles de machine learning et ses connaissances en informatique l'aident à anticiper leur mise en production. En amont de ces deux missions, il/elle est également en charge de structurer et d'analyser les données qu'il/elle utilise.

On constate surtout que les compétences en informatique sont indissociables du métier de statisticien, en particulier en amont (accès, préparation, pipeline des données) et en aval (visualisation, diffusion, déploiement, « dockerisation »)

Vague « Big Data Analytics » (Data Science) actuelle



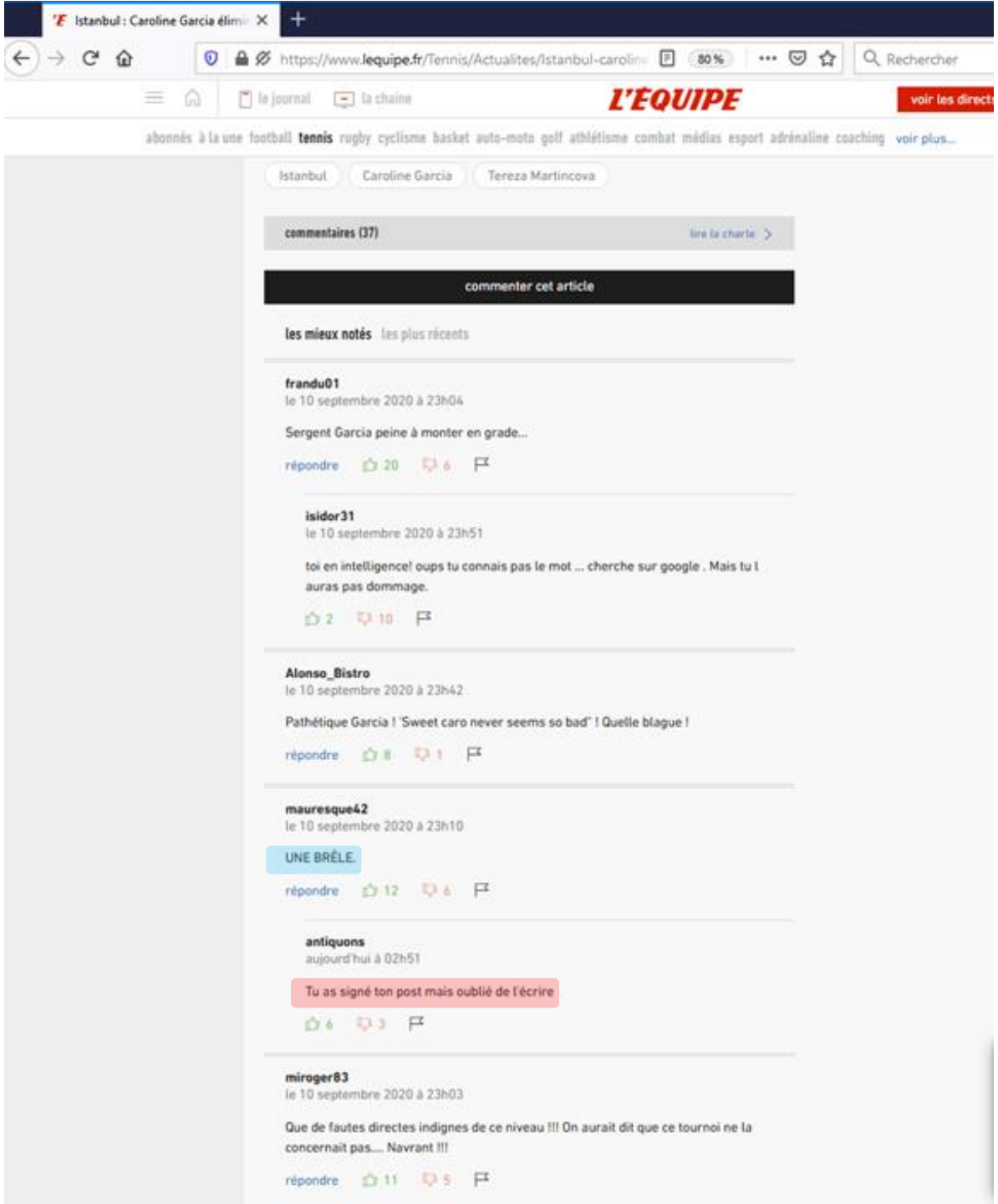
Sources additionnelles d'information externes à l'entreprise (web scrapping, API, …), multiplicité des formats, nouveaux enjeux technologiques pour le stockage et le traitement (stockage NoSQL, data lake, tech. CLOUD, info. distribuée [Hadoop, Spark]…) et nouvelles opportunités d' analyse ! (cf. Rapports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

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

Exemple 1. Filtrage collaboratif, Systèmes de recommandation.

Exemple 2. Analyse des réseaux sociaux.

On en voit des vertes et des pas mûres sur les réseaux sociaux. Toujours intéressant, souvent désespérant aussi. Ex. Commentaires sur le site web de l'Equipe.



Logiciels de statistique / data science / machine learning

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

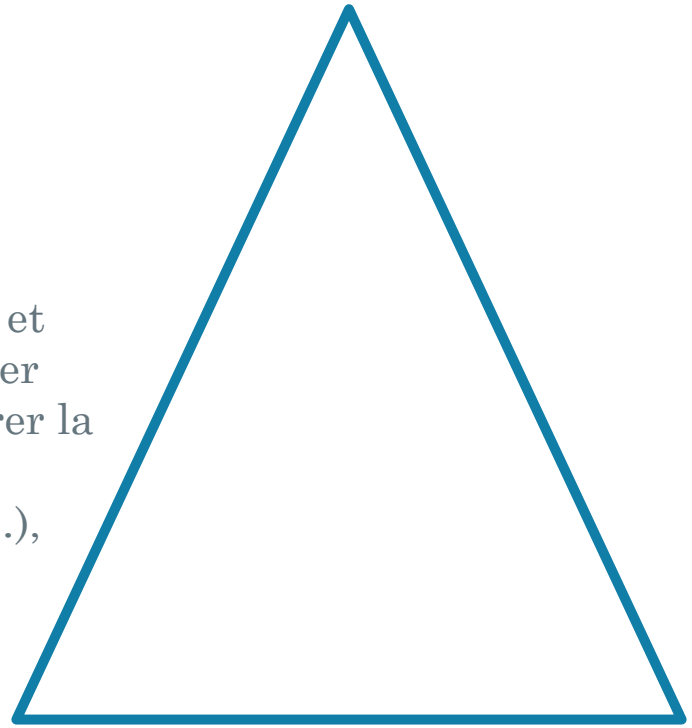
Enseignement de la data science

STATISTIQUE

DATA MINING

MACHINE LEARNING

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 (Hadoop,…), Cloud, MLOps

INFORMATIQUE

Toute analyse s'inscrit dans un domaine d'application : données de sécurité, données du web (web scraping), analyse des réseaux sociaux (API), etc.

APPLICATIONS

L'outil – le logiciel – joue un rôle très important !

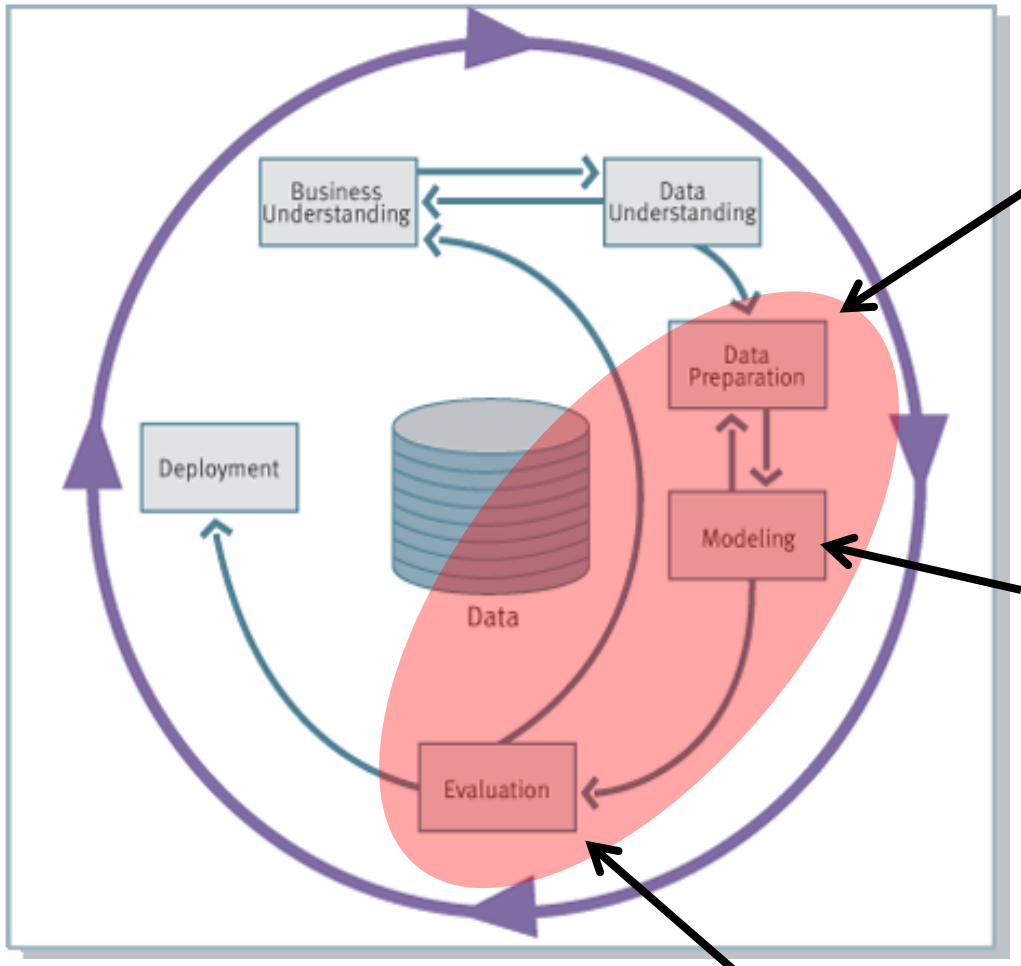
Logiciels pour l'enseignement – Objectif pédagogique des TD sur machine

L'utilisation du logiciel doit s'inscrire dans une démarche pédagogique.

Objectifs d'une séance « classique » de TD sur machine :

- Mettre en œuvre une technique de statistique 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, accès aux données / parsing, etc.)

TD sur machine - Positionnement dans la chaîne complète



Accès à des fichiers plats. Sachant que la partie préparation, à moins d'y consacrer des séances spécifiques (ex. text mining), est réduite à sa plus simple expression, ou pire à des solutions « automatiques » sans véritable recul (ex. données NA).

Beaucoup sur l'étude des méthodes, le paramétrage, l'optimisation, la lecture et l'interprétation des résultats. Association avec les représentations graphiques (en amont et en aval de la modélisation). Pour ma part, je débute toujours par le tableur dans mes cours d'initiation (ex. analyse discriminante ; arbre de décision ; …)

Se limite souvent à la mesure des performances et aux comparaisons. Parfois interprétation.

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 (cf. programmation vs. choix de Tanagra en 2005).
- Démarche harmonisée pour un large spectre d'applications (même environnement pour les statistiques, le machine learning, le text mining, image mining, etc.). Ne pas multiplier les outils en fonction de la matière étudiée ou de la tâche à réaliser. 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.*

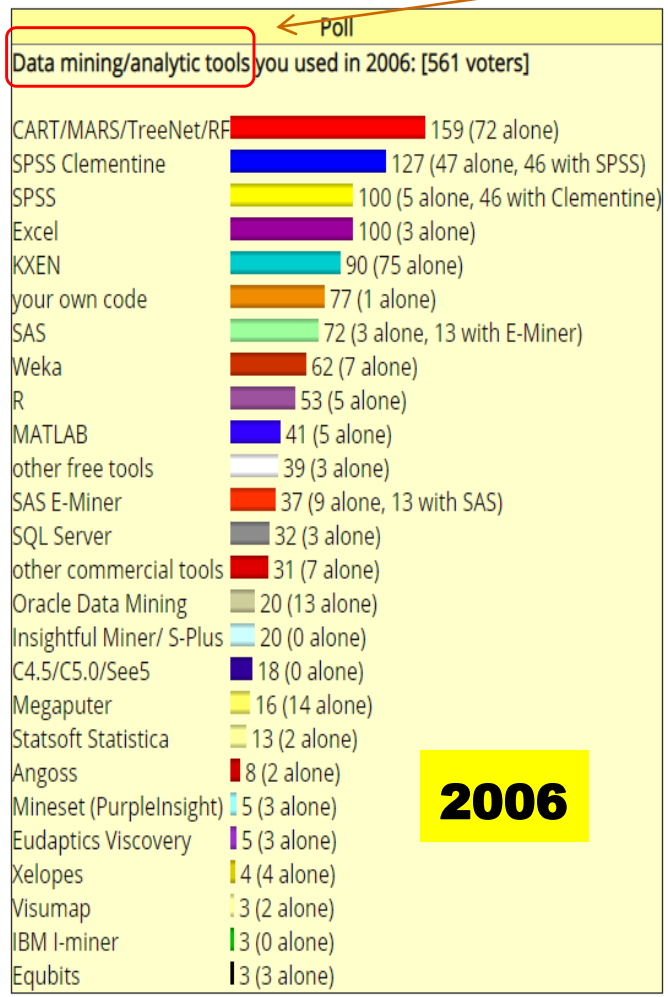
Aujourd'hui, R et Python s'imposent dans nos formations parce que :

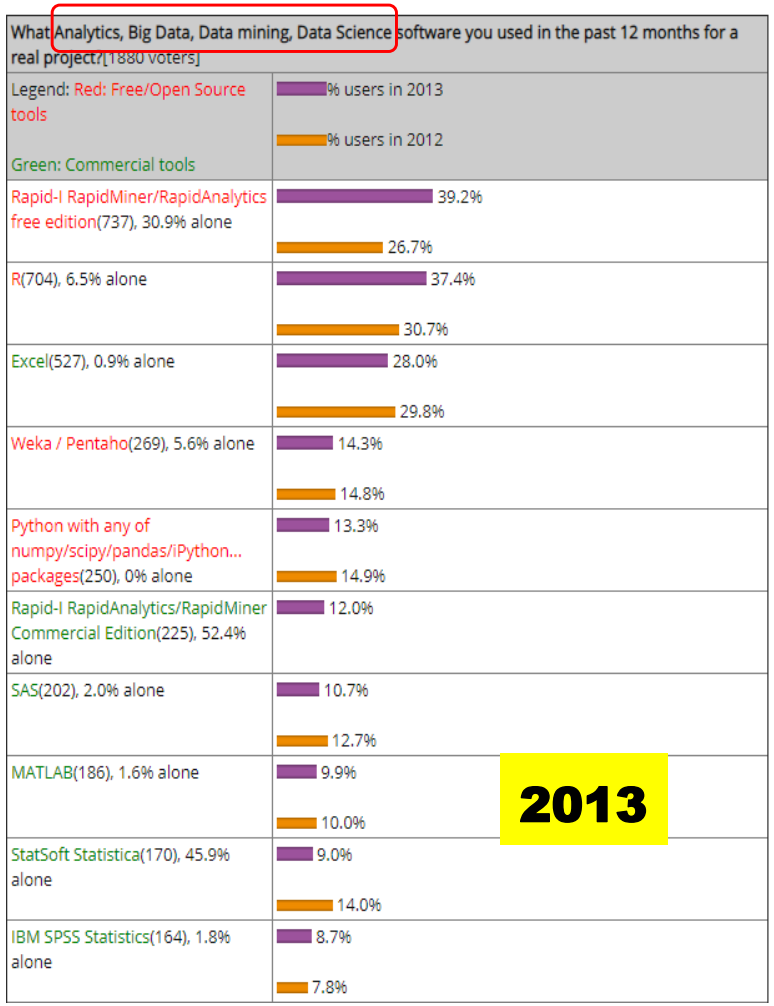
- Gratuits, totalement, ce n'est pas moindre de leurs qualités.
- Forte pénétration de ces outils dans les entreprises (cf. les offres d'emploi sur le site de l'APEC, avec les mots-clés « statistique », « data science », « machine learning », etc.).
- Le niveau en programmation des étudiants a considérablement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

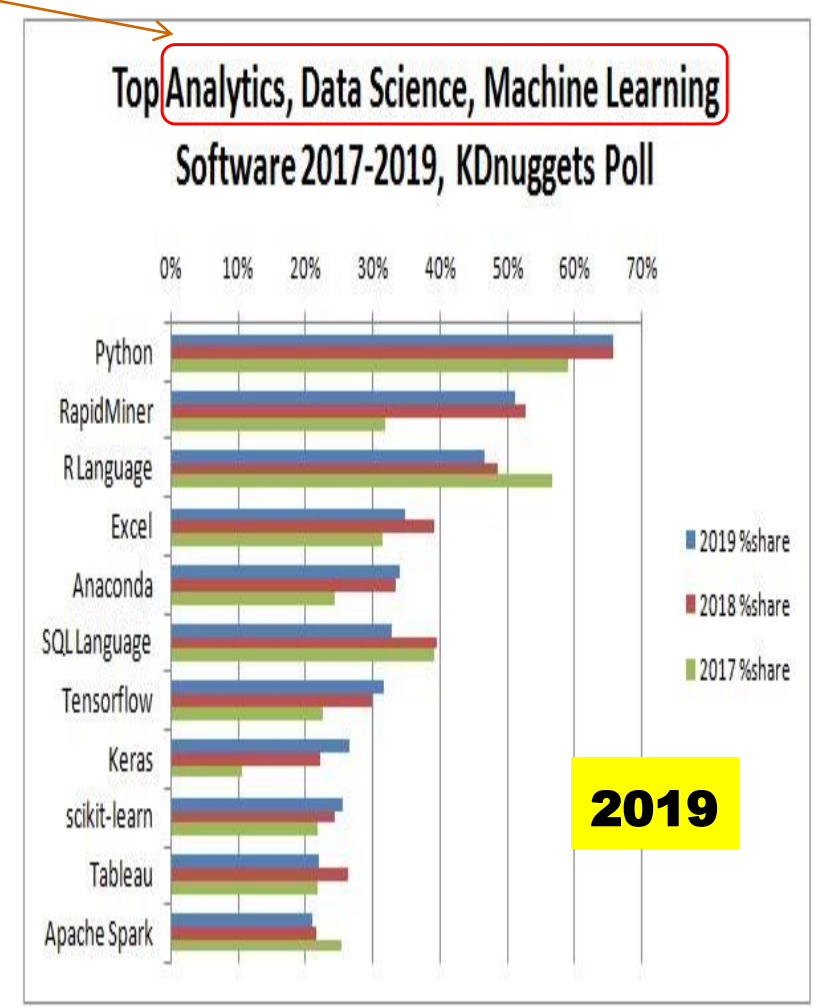
Compétence sur ces outils devient un marqueur fort dans les CV.

Sondage annuel KDnuggets (2006, 2013, 2019)

Les appellations changent au fil des années…





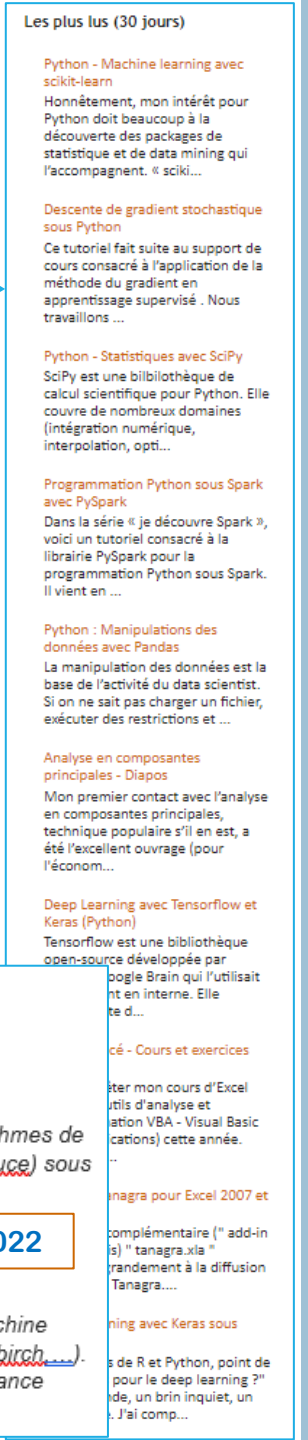


Il y a matière à réflexion quand on regarde cette évolution

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. APEC, les stages encadrés) ; chargé d'études statistiques (autres outils) ; etc.
- La « lenteur » de R est souvent décriée (mais est-ce vraiment justifié…)


Les plus lus (30 jours)

Python - Machine learning avec scikit-learn
Honnêtement, mon intérêt pour Python doit beaucoup à la découverte des packages de statistique et de data mining qui l'accompagnent. « sciki…

Descente de gradient stochastique sous Python
Ce tutoriel fait suite au support de cours consacré à l'application de la méthode du gradient en apprentissage supervisé . Nous travaillons …

Python - Statistiques avec SciPy
SciPy est une bilbilothèque de calcul scientifique pour Python. Elle couvre de nombreux domaines (intégration numérique, interpolation, opti…

Programmation Python sous Spark avec PySpark
Dans la série « je découvre Spark », voici un tutoriel consacré à la librairie PySpark pour la programmation Python sous Spark. Il vient en …

Python : Manipulations des données avec Pandas
La manipulation des données est la base de l'activité du data scientist. Si on ne sait pas charger un fichier, exécuter des restrictions et …

Analyse en composantes principales - Diapos
Mon premier contact avec l'analyse en composantes principales, technique populaire s'il en est, a été l'excellent ouvrage (pour l'économ…

Deep Learning avec Tensorflow et Keras (Python)
Tensorflow est une bibliothèque open-source développée par …ogle Brain qui l'utilisait …nt en interne. Elle …te d…

…cé - Cours et exercices

…ter mon cours d'Excel …utils d'analyse et …ation VBA - Visual Basic …cations) cette année.

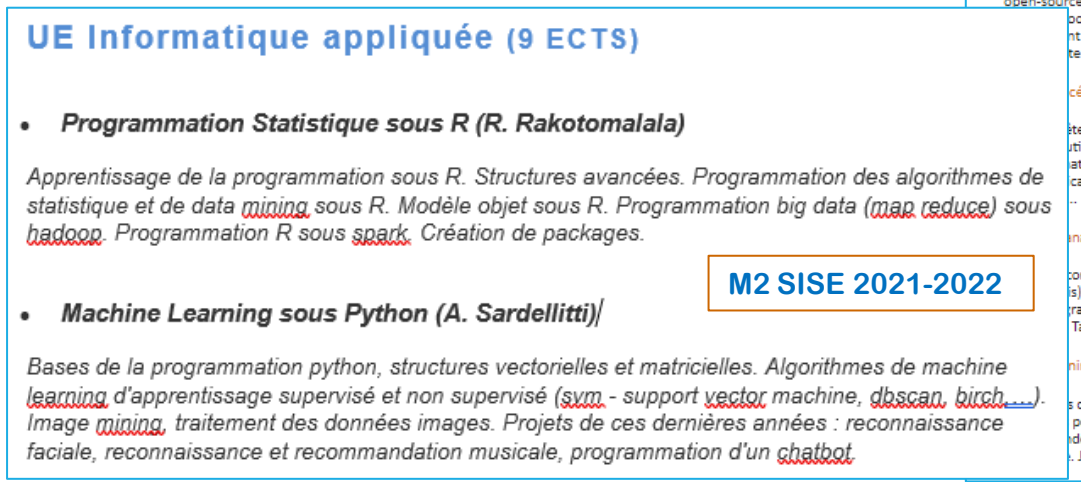
…nagra pour Excel 2007 et …complémentaire (" add-in …s) " tanagra.xla " …randement à la diffusion … Tanagra….

…ning avec Keras sous …

…s de R et Python, point de … pour le deep learning ?" …de, un brin inquiet, un …. J'ai comp…

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 (cf. TD ACP, rotation d'axes, clustering de variables)


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 2021-2022

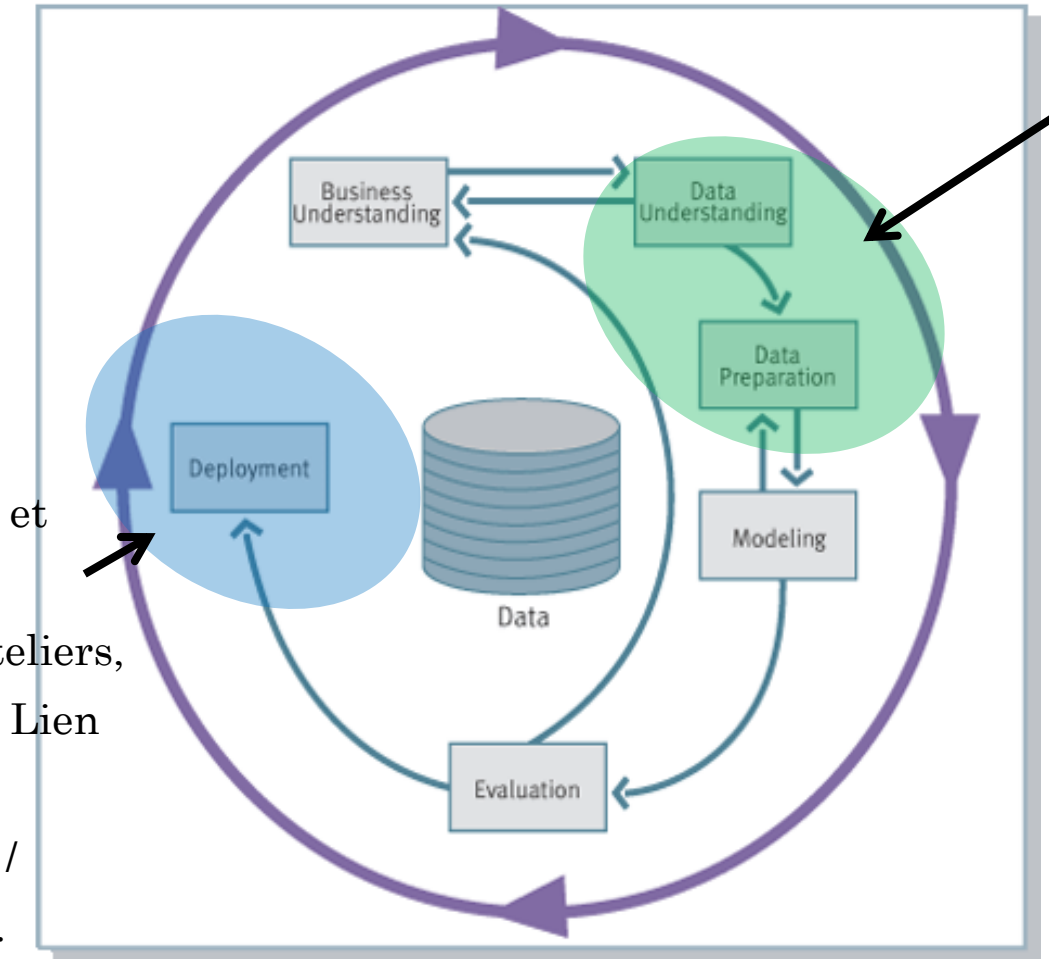
- ***Machine Learning sous Python (A. Sardellitti)***

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 autres étapes ?

Informatique. Accès aux données du web (web scraping), utilisation des API (ex. OpenStreetMap, Tweeter, …). Données textuelles, images, vidéos.

➔ Souvent sous forme d'ateliers, ou plus sûrement dans les projets.



Informatique. Technologies et solutions de déploiement.

➔ Ici aussi, soit cours, soit ateliers, soit intégrés dans les projets. Lien avec les applications web dynamiques (Rshiny, Python / Dash, …). Data visualisation.

Etudes de cas

Sous R et Python – Projets POC effectués par les étudiants

Reconnaissance faciale (1)

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 »

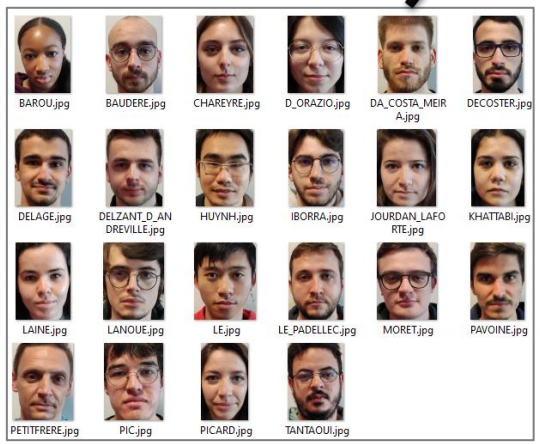
Vecteur de description de l'individu « requête »

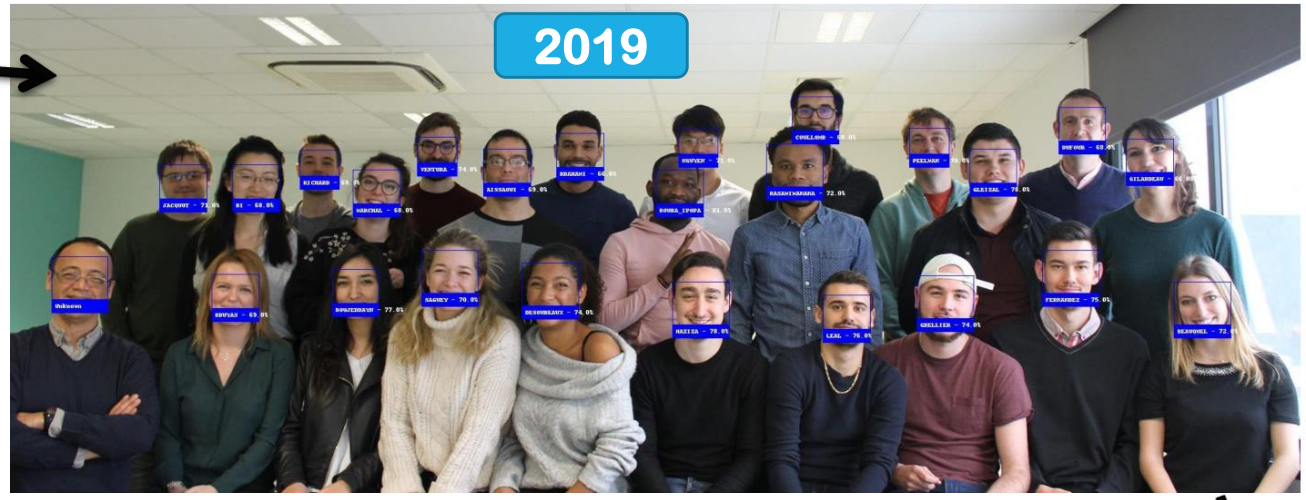
Identification avec degré de fiabilité.

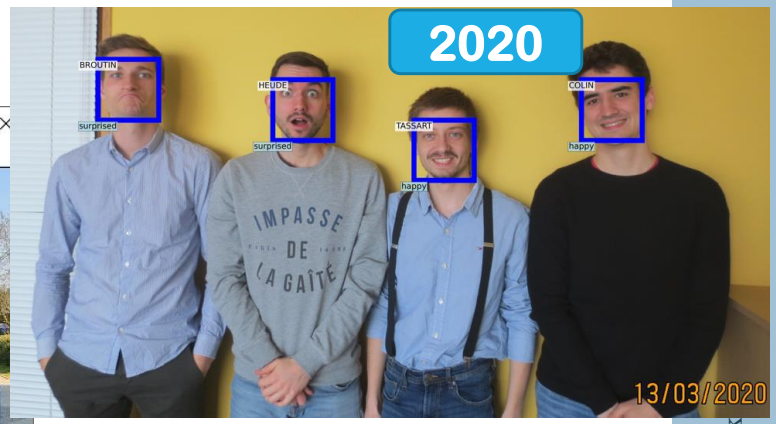
Reconnaissance faciale++ – **Rôle des modèles pré-entraînés** (2)

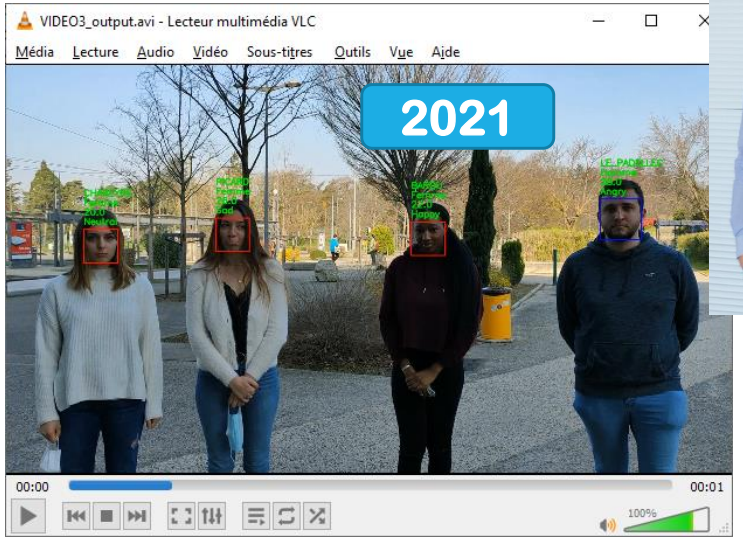
Photos Etudiants

Promotion M2 SISE









Jusqu'où je peux aller …

Analyse des offres d'emploi (1)

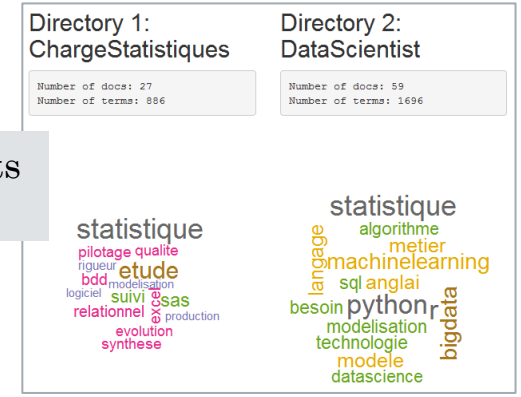
Analyse de documents textuels (text mining) et développement application visuelle. Projet sous R (Shiny)

Offres d'emploi qui ont été étiquetées manuellement.

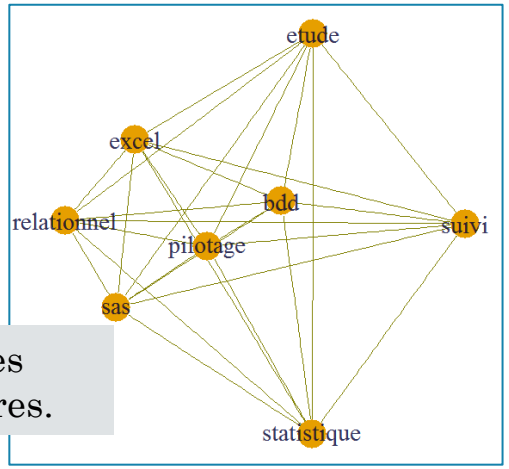
Analyse et développement d'un application Shiny

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

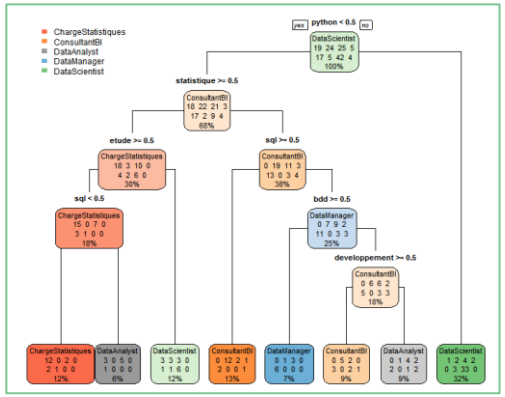
Mots clés fréquents selon les métiers



Association entre les termes dans les offres.



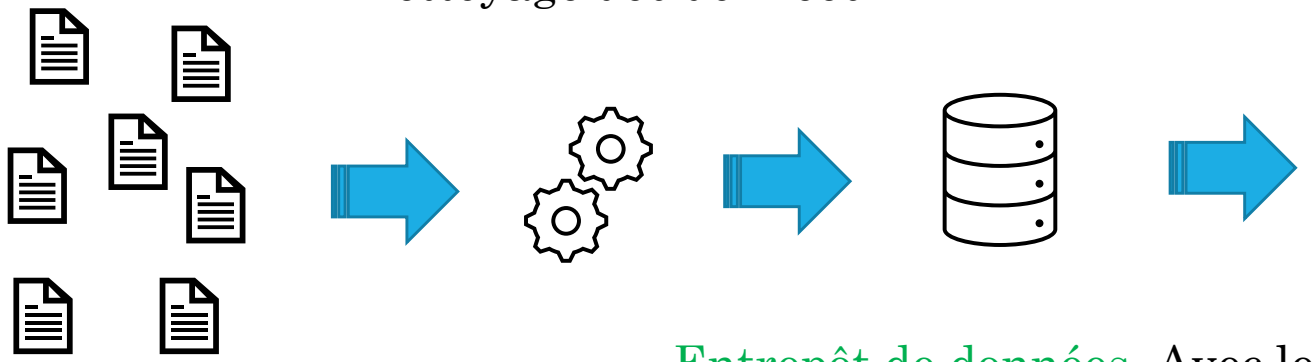
Identification des métiers selon les termes de l'offre.

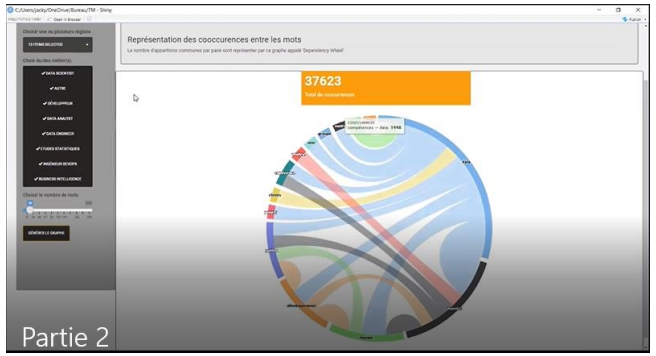


Analyse des offres d'emploi++ – **Chaîne complète et déploiement** (2)

Web scrapping ou utilisation des API. Récupération et nettoyage des données

Text Mining (NLP). Développement d'une application (programmation informatique) d'analyse et de visualisation des données textuelles (machine learning)





Description des offres d'emploi sur des sites spécialisés. APEC, INDEED, POLE EMPLOI, …

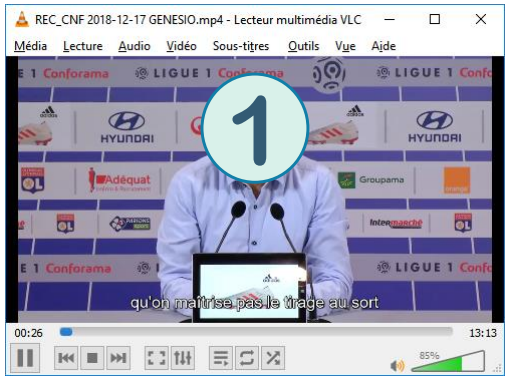
Entrepôt de données. Avec les « standards » du domaine, tables de faits et dimensions (axes d'analyse). Hébergement dans un cloud (Azure, Google, …).

Solliciter les multiples compétences (la polyvalence) des étudiants.
Plus qu'on ne le croit quand il a fallu passer au déploiement (conteneur docker) !!!

Traduction automatique de vidéos (1)

Projet de « service » IA.

Technologies cloud.



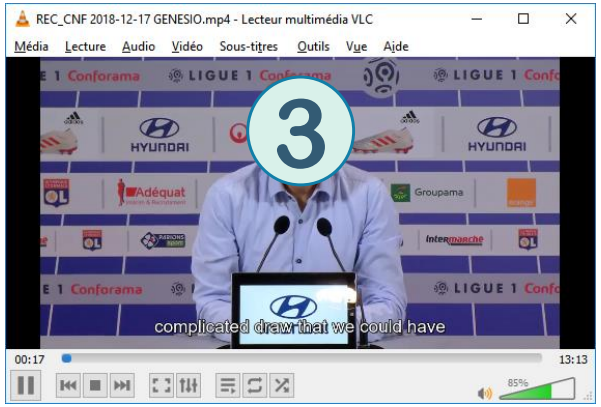
(1) Speech-to-text : extraction des bandes et transcription sous forme textuelle

Mise en concurrence de 3 technologies : Microsoft Azure, IBM Watson, Google API

0:0:0	0:0:13	Bonjour bonjour Bruno j'espère que ta première réaction sur ce tirage au sort	Hello hello Bruno I hope your first reaction on this draw
0:0:13	0:0:27	ma foi c'est certainement le tirage le plus un des plus compliqué qu'on pouvait avoir non maintenant comme je dis souvent et des choses qu'on maîtrise d'autres qu'on maîtrise pas le tirage au sort	my faith is certainly the most complicated draw that we could have not now as I say often and things we master others we do not master the draw

2

(2) Traduction automatique de texte, avec respect de l'horodatage. Avec une partie Machine Learning pour le vocabulaire spécifique au football.



(3) Insertion de sous-titres en anglais dans la vidéo

Traduction automatique de vidéos (2)

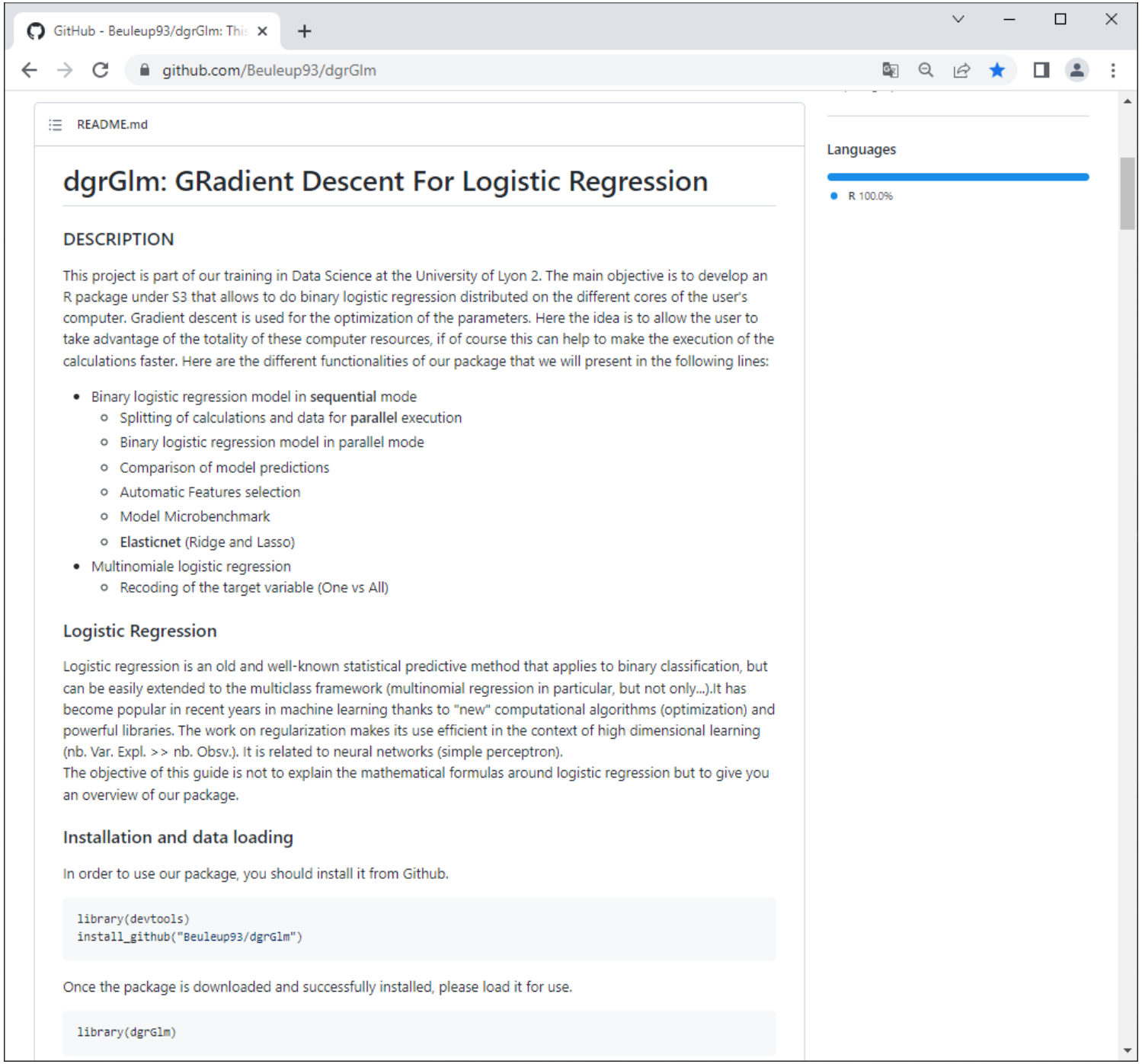
Développement de packages R

Implémentation d'un algorithme de machine learning. Diffusion sous forme de package R. Installable directement à partir de GitHub.

Ex.1. Régression logistique. Descente de gradient et descente de gradient stochastique (dgrGlm ; GradDesc). Parallélisation.

Ex.2. Analyse discriminante linéaire (discriminR).

Ex.3. Régression PLS en classement.



Conclusion et bibliographie

Conclusion

R et Python jouent actuellement - tous deux - un rôle essentiel dans l'enseignement des Statistiques, Machine Learning, Data Science.

Parce que

- Répondent parfaitement aux objectifs pédagogiques. D'autant plus que les compétences informatiques sont de plus en plus prégnantes dans nos métiers.
- Répondent aux attentes en matière de traitements – IA, Machine Learning – qui suscitent l'enthousiasme des étudiants.
- Sont explicitement demandés dans les offres d'emploi / stages
- Développer une expertise élevée dans les deux outils / langages (R et Python) n'implique pas un coût pédagogique additionnel fort.

Mais…

Il me paraît évident aujourd'hui que nous devons de plus en plus regarder du côté du cloud (cf. APEC) : Azure Cloud, AWS Cloud, Google Cloud, …. Sans pour autant renier nos connaissances (ex. Python sur Azure)

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.

Hamel G., « Data Science Tools Popularity, animated », June 2020.

Enquête annuelle, évolutions, comparaisons au fil des années précédentes.

Kaggle, « State of Data Science and Machine Learning 2022 ».

Quelques tendances intéressantes sur les outils et les usages.

La revue MODULAD, « La page Excel'Ense ».

Quoiqu'en en dise, le tableur est un très bon outil pédagogique, en plus d'être très présent dans certains métiers liés à la statistique.