

Introduction à la DATA SCIENCE

Du DATA MINING au BIG DATA

Enjeux et opportunités

Ricco RAKOTOMALALA

Université Lumière Lyon 2



Plan

1. Data Science - Définition
2. Une première étape importante : le Data Mining
3. Spécificités du Data Mining – Applications
4. Big Data – Nouveauté, virage, évolution ?
5. Enjeux et opportunités
6. Les outils de data science
7. Bibliographie



DATA SCIENCE

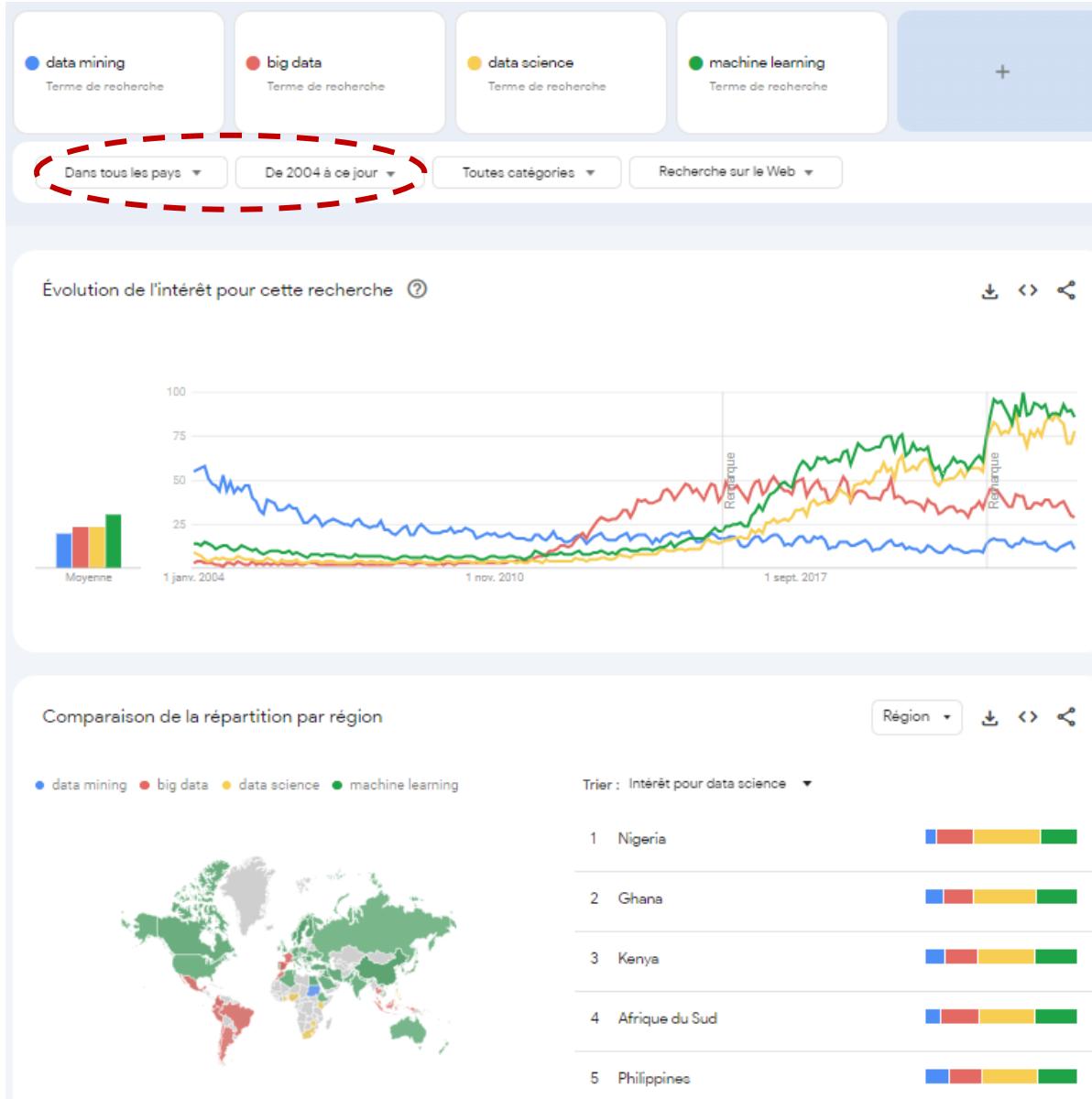
Science des données ? De quoi il retourne ?

(La notion est très en vogue)



Data Science, Data Mining, Machine learning, Big Data Analytics... quid ?

Des notions très « à la mode », comment les discerner ?



Mais finalement, l'idée maîtresse est la valorisation des données via des techniques de traitement statistique / data mining / machine learning (au sens large : description, exploration, modélisation, apprentissage,).

Cf. [Rapport Lauvergeon](#) ; cf. [The Elements of Statistical Learning](#)



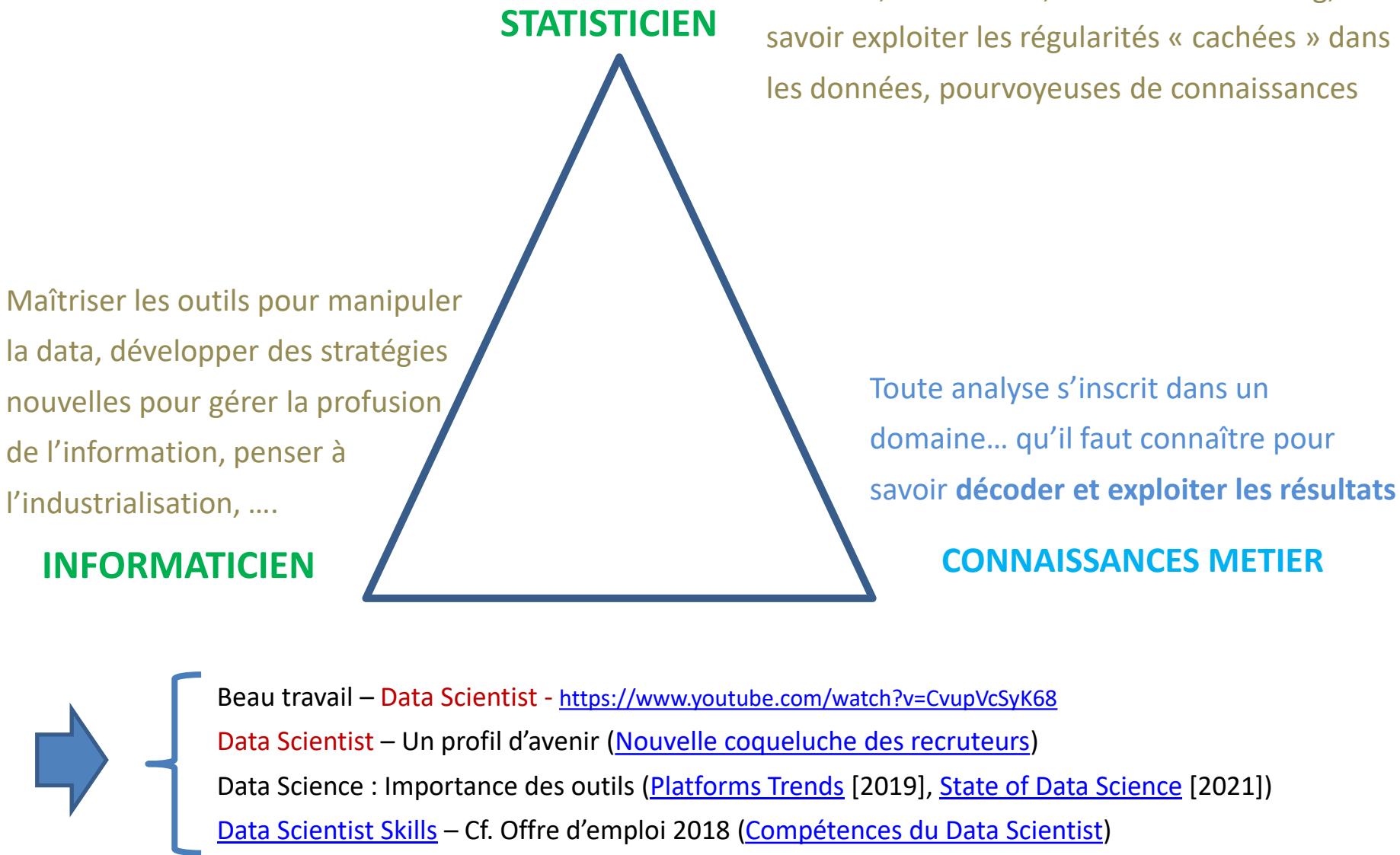
Data science is the study of the generalizable extraction of knowledge from data (**objet**), yet the key word is science. It incorporates varying elements and builds on techniques and theories from many fields, including signal processing, mathematics, probability models, machine learning, statistical learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modeling, data warehousing, and high performance computing...

(Double compétence : statistique et informatique) ([Wikipédia](#)).



Although use of the term data science has exploded in business environments, many academics and journalists see no distinction between data science and statistics. Writing in Forbes, Gil Press argues that data science is a **buzzword** without a clear definition and has simply replaced “business analytics” in contexts such as graduate degree programs... ([Wikipédia](#)).

Data science – A la croisée de 3 profils



Data science – Pourquoi une telle effervescence aujourd’hui ?

- 1 Nous sommes à l’heure des « data » ... qui arrivent **de partout** et que l’on sait **collecter** et **conserver** à moindre coût
- 2 Prise de conscience collective... **surtout des entreprises**... de la valeur ajoutée que l’on peut en tirer
- 3 La construction des modèles pré-entraînés sur des grands corpus est devenue la norme, modèles directement exploitables via les IA génératives par ex.

Statistique /
Analyse de données



Machine Learning
Data Mining



Data Science
Big Data Analytics

La progression s’accompagne d’une **évolution des techniques / technologies** et des **sources d’information**. !

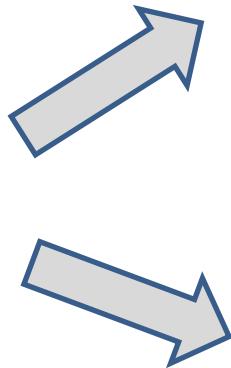


Statistique

Traitement statistique des données



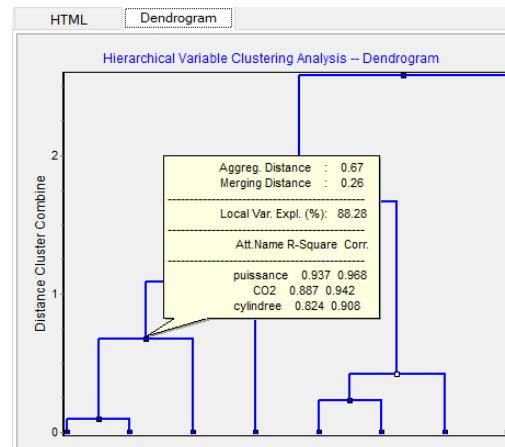
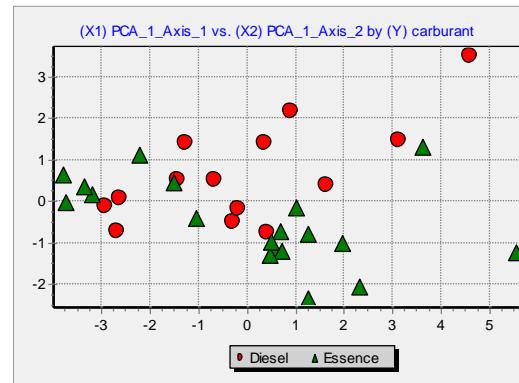
Application des techniques de modélisation et de statistique



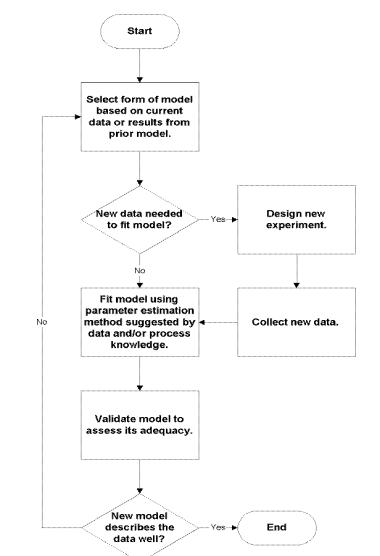
Les données sont spécifiquement recueillies à des fins d'étude (ex. enquête, expérimentations, etc.)

- Bonne qualité souvent
- Faible volumétrie (coût du recueil)

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



[Modeling Steps \(NIST – e-Handbook of Statistical Methods\)](#)



DATA MINING

La démarche Knowledge Discovery in Databases (KDD)



Exemple introductif : demande de crédit bancaire



L'expert se fonde sur son « **expérience** » pour prendre la bonne décision

- divorcé
- 5 enfants à charge
- chômeur en fin de droit
- compte à découvert



L'entreprise d'une « expérience » supplémentaire : « l'expérience numérique ». Les différentes bases qui lui permettent de fonctionner, et qui permettent de retracer son activité... Elles constituent une « mémoire » de l'entreprise.



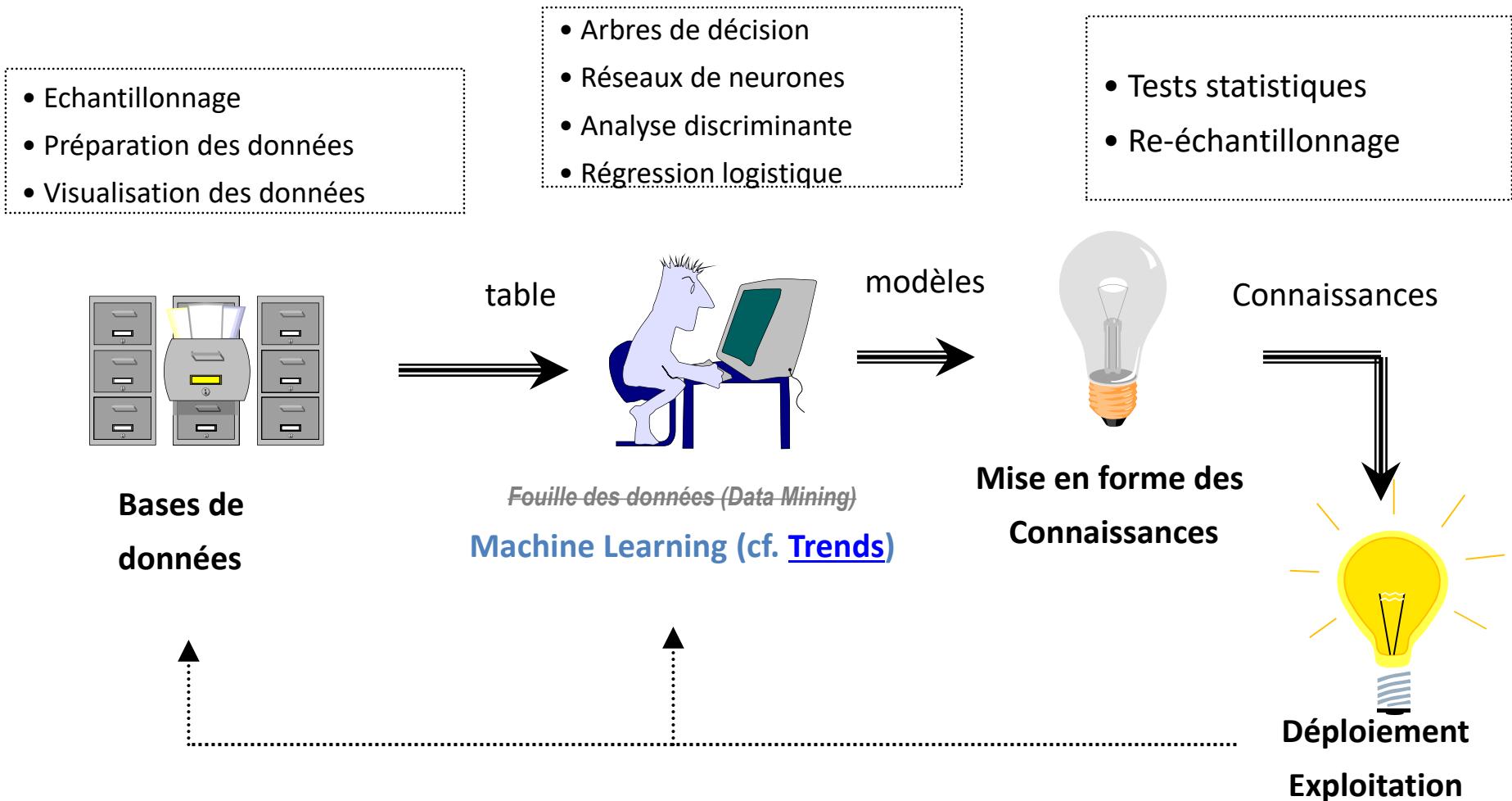
- coûteuse en stockage
- inexploitée pendant longtemps

Comment et à quelles fins utiliser cette expérience
accumulée



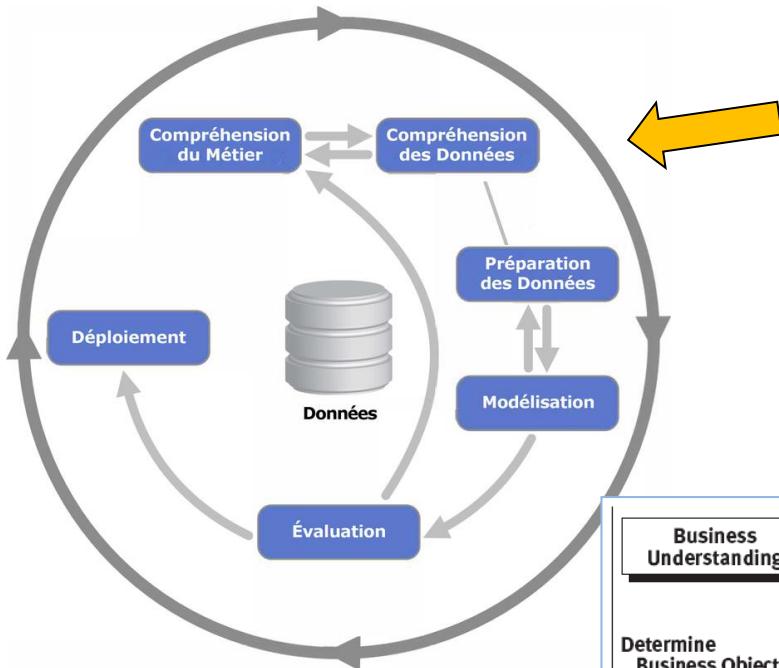
Le processus ECD (Extraction de connaissances à partir de données)

KDD – Knowledge discovery in Databases (<http://www.kdnuggets.com/>)



Définition : Processus non-trivial d 'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données (Fayyad, 1996)





Méthodologie CRISP-DM

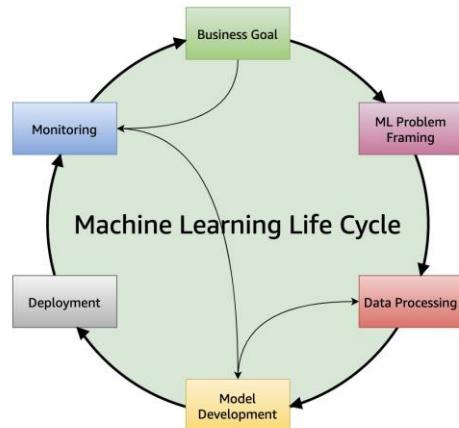
Travailler en synergie avec l'expert du domaine est primordial !

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i></p> <p>Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i></p> <p>Determine Data Mining Goals <i>Data Mining Goals Data Mining Success Criteria</i></p> <p>Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i></p>	<p>Collect Initial Data <i>Initial Data Collection Report</i></p> <p>Describe Data <i>Data Description Report</i></p> <p>Explore Data <i>Data Exploration Report</i></p> <p>Verify Data Quality <i>Data Quality Report</i></p>	<p>Select Data <i>Rationale for Inclusion/Exclusion</i></p> <p>Clean Data <i>Data Cleaning Report</i></p> <p>Construct Data <i>Derived Attributes Generated Records</i></p> <p>Integrate Data <i>Merged Data</i></p> <p>Format Data <i>Reformatted Data</i></p> <p>Dataset <i>Dataset Description</i></p>	<p>Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i></p> <p>Generate Test Design <i>Test Design</i></p> <p>Build Model <i>Parameter Settings Models Model Descriptions</i></p> <p>Assess Model <i>Model Assessment Revised Parameter Settings</i></p>	<p>Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i></p> <p>Review Process <i>Review of Process</i></p> <p>Determine Next Steps <i>List of Possible Actions Decision</i></p>	<p>Plan Deployment <i>Deployment Plan</i></p> <p>Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i></p> <p>Produce Final Report <i>Final Report Final Presentation</i></p> <p>Review Project <i>Experience Documentation</i></p>



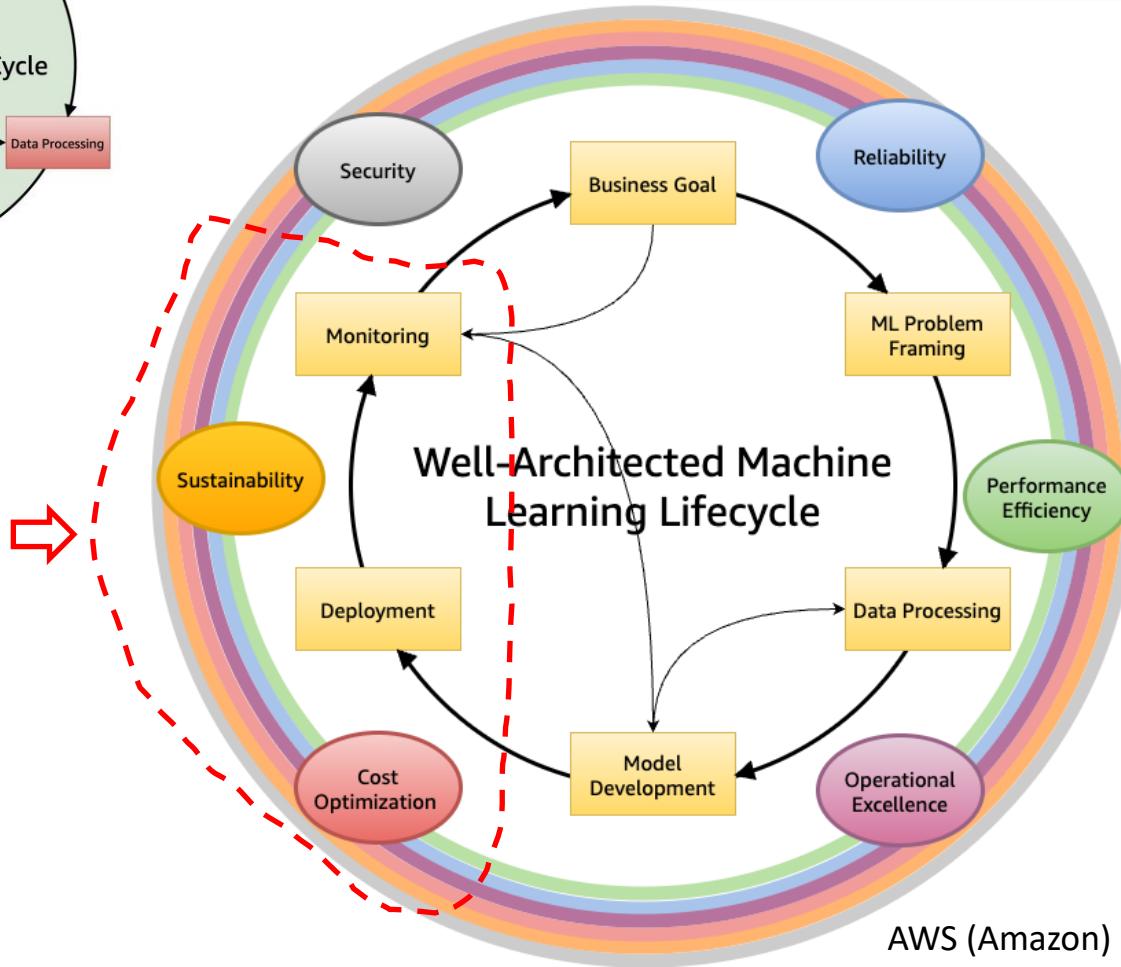
20 ans plus tard

« Cycle de vie des projets de machine learning »



Remarque : De nouveaux paradigmes arrivent aujourd'hui avec MLOps par ex., bonnes pratiques et outils pour développer et maintenir efficacement un projet machine learning, inspiré de DevOps; en tenant compte des problématiques de coûts, sécurité, efficacité, ...

Fait partie intégrante du processus aujourd'hui, avec de nouvelles compétences et de nouveaux outils informatiques ([MLFLOW](#), [AIRFLOW](#), [DOCKER](#), etc.)



AWS (Amazon) : [Well Architected machine learning lifecycle](#)

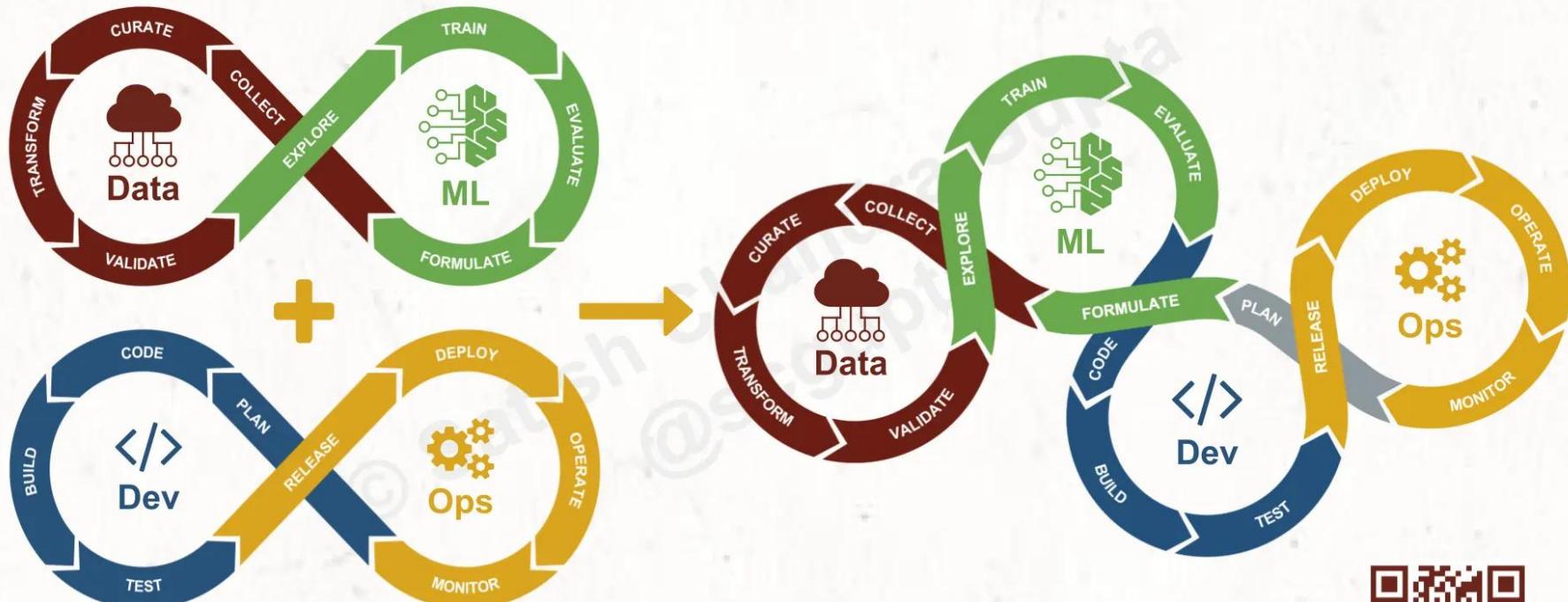


Machine Learning + Ingénierie informatique = Machine learning engineer

<https://www.datacamp.com/blog/machine-learning-lifecycle-explained>

MLOps = DataML + DevOps

ml4devs.com/mlops-lifecycle 



© 2022 Satish Chandra Gupta



CC BY-NC-ND 4.0 International License
creativecommons.org/licenses/by-nc-nd/4.0/

scgupta.me 

twitter.com/scgupta 

linkedin.com/in/scgupta 

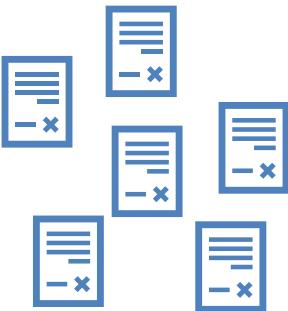


Ricco Rakotomalala

Tutoriels Data Science - <https://tutoriels-data-science.blogspot.com/>

Un exemple : Application de NLP basé sur du web scraping

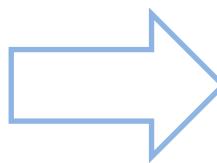
Différents canaux
(scraping, API). Il faut
harmoniser, nettoyer.



Offres d'emploi accessibles
sur différents sites (APEC,
Pôle Emploi, Indeed, ...).

Données textuelles.

ETL

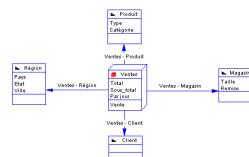


SGBD à instancier
(MySQL, PostgreSQL, ...) ou SQLite ou ...



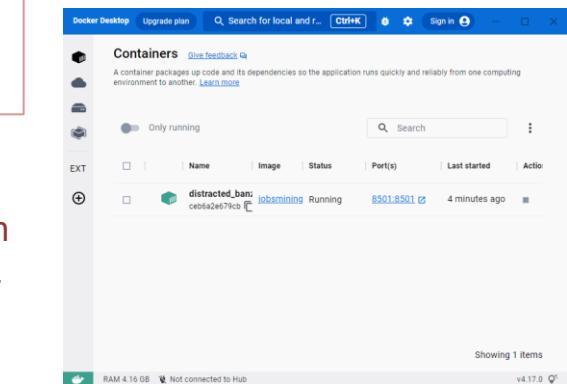
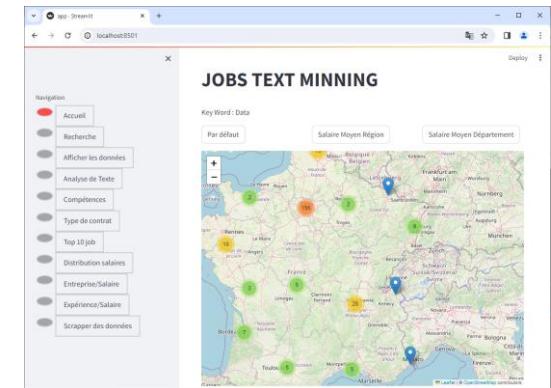
Modélisation

multidimensionnelle



Encapsulé dans un
conteneur Docker

Web App Python
(Streamlit, Dash, ...).
Analyses des offres.



Spécificités du Data Mining ?

- (1) Sources de données
- (2) Techniques utilisées
- (3) Multiplicité des supports



« Data Mining » était vraiment nouveau à l'époque ?

KDD (Data Mining) - <http://www.kdnuggets.com/>

Processus non-trivial d 'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données (Fayyad, 1996)

Data Mining : Une nouvelle façon de faire de la statistique ?

<http://cedric.cnam.fr/~saporta/DM.pdf>

L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.» (J.P.Benzécri, 1973)

The basic steps for developing an effective process model ?

<http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd41.htm>

A comparer avec Data Mining Concepts ([Microsoft](#)) ou Data Mining as process ([IBM](#))



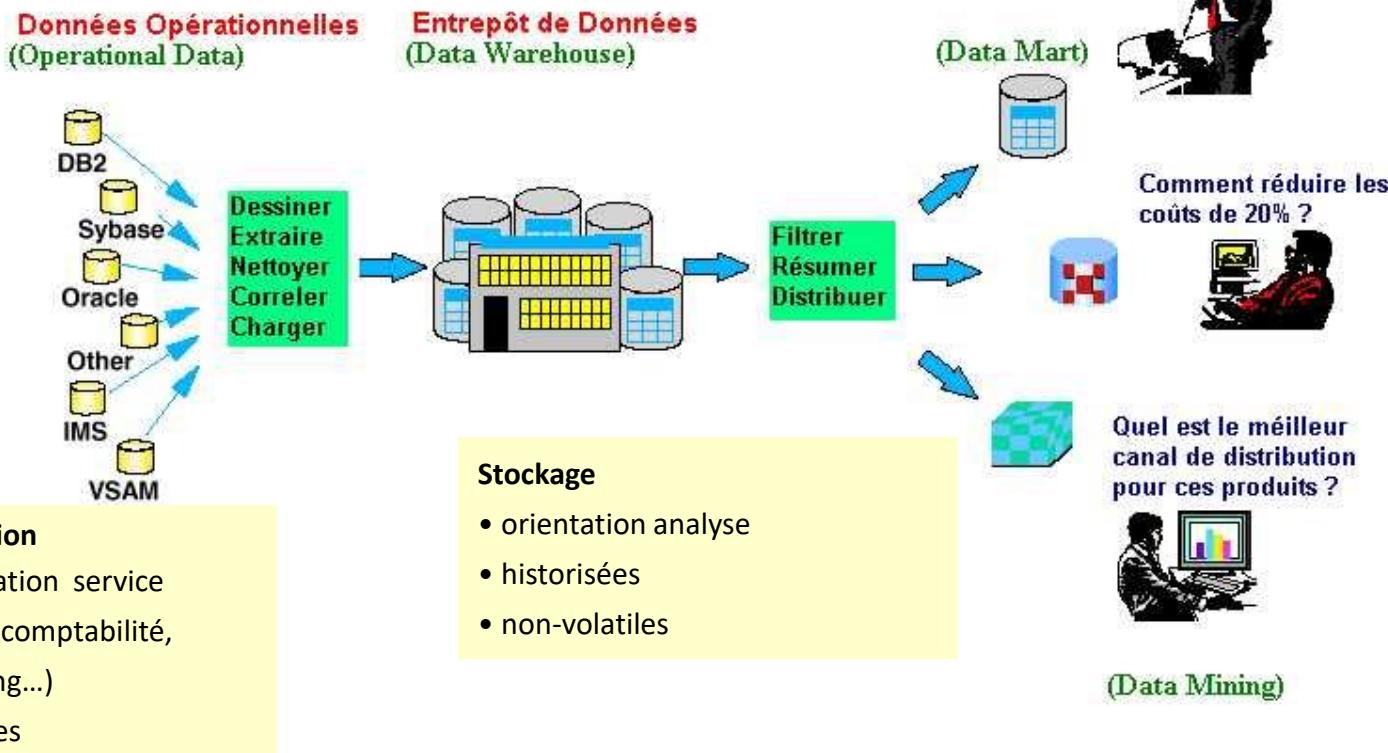
Spécif.1 - Les sources de données

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

~~Construire une Infrastructure d'Information Intelligente pour l'Entreprise~~

Bases décisionnelles

Quelles seront les tendances salariales la prochaine année?



LES MÉTIERS DE LA DATA



La maîtrise de la donnée crée de la valeur qui permet d'améliorer la performance de nos produits et services ainsi que la performance opérationnelle de l'ensemble des équipes.



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).

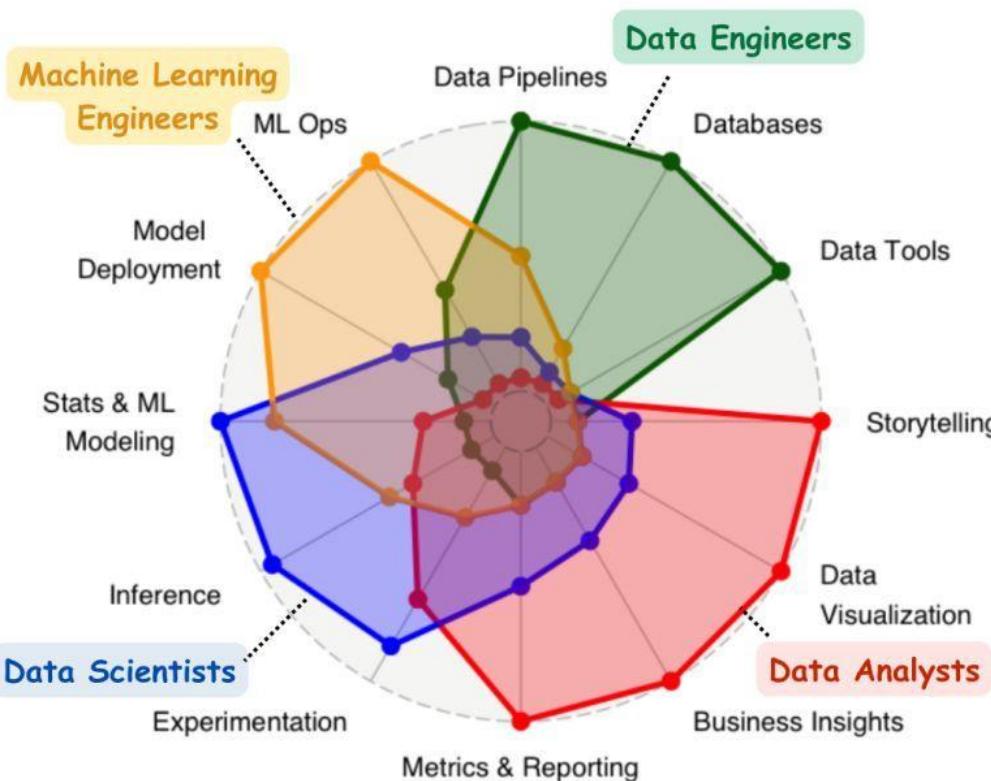
DATA SCIENTIST F/H

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 deviennent indissociables du métier de statisticien / machine learner, en particulier en amont (accès, préparation, pipeline des données) et en aval (visualisation, diffusion, déploiement, « dockerisation »)



Types of Data Roles - Where are you?



Pourquoi choisir ?
Ces compétences
sont indissociables !

[Kevin Rosamont Prombo \(2023\)](#)

Spécif.2 - Brassage des cultures et des techniques

Statistiques

Théorie de l'estimation, tests

Économétrie

Maximum de vraisemblance et moindres carrés

Régression linéaire, régression logistique, anova...

Analyse des données

(Statistique exploratoire)

Description factorielle

Discrimination

Clustering

Méthodes géométriques, probabilités

ACP, ACM, Analyse discriminante, CAH, ...

	var 1	var 2	...	var J
individu 1				
individu 2				
...			valeurs	
individu n				

Informatique

(Intelligence artificielle) - Machine learning

Apprentissage symbolique

Reconnaissance de formes

Une étape de l'intelligence artificielle

Réseaux de neurones, algorithmes génétiques...

Informatique

(Base de données)

Exploration des bases de données

Volumétrie

Règles d'association, motifs fréquents, ...

Très souvent, ces méthodes se rejoignent, mais avec des philosophies / approches / formulations différentes



Les méthodes selon les finalités

Description :

trouver un résumé des données qui soit plus intelligible

- statistique descriptive
- analyse factorielle

Ex : moyennes conditionnelles, etc.

Structuration :

Faire ressurgir des groupes « naturels » qui représentent des entités particulières

- **classification** (clustering, apprentissage non-supervisé)

Ex : découvrir une typologie de comportement des clients d'un magasin

Les méthodes sont le plus souvent complémentaires

Méthodes de « Machine Learning »

Explication :

Prédire les valeurs d'un attribut (endogène) à partir d'autres attributs (exogènes)

- régression
- **apprentissage supervisé**

Ex : prédire la qualité d'un client (rembourse ou non son crédit) en fonction de ses caractéristiques (revenus, statut marital, nombre d'enfants, etc.)

Association :

Trouver les ensembles de descripteurs qui sont le plus corrélés

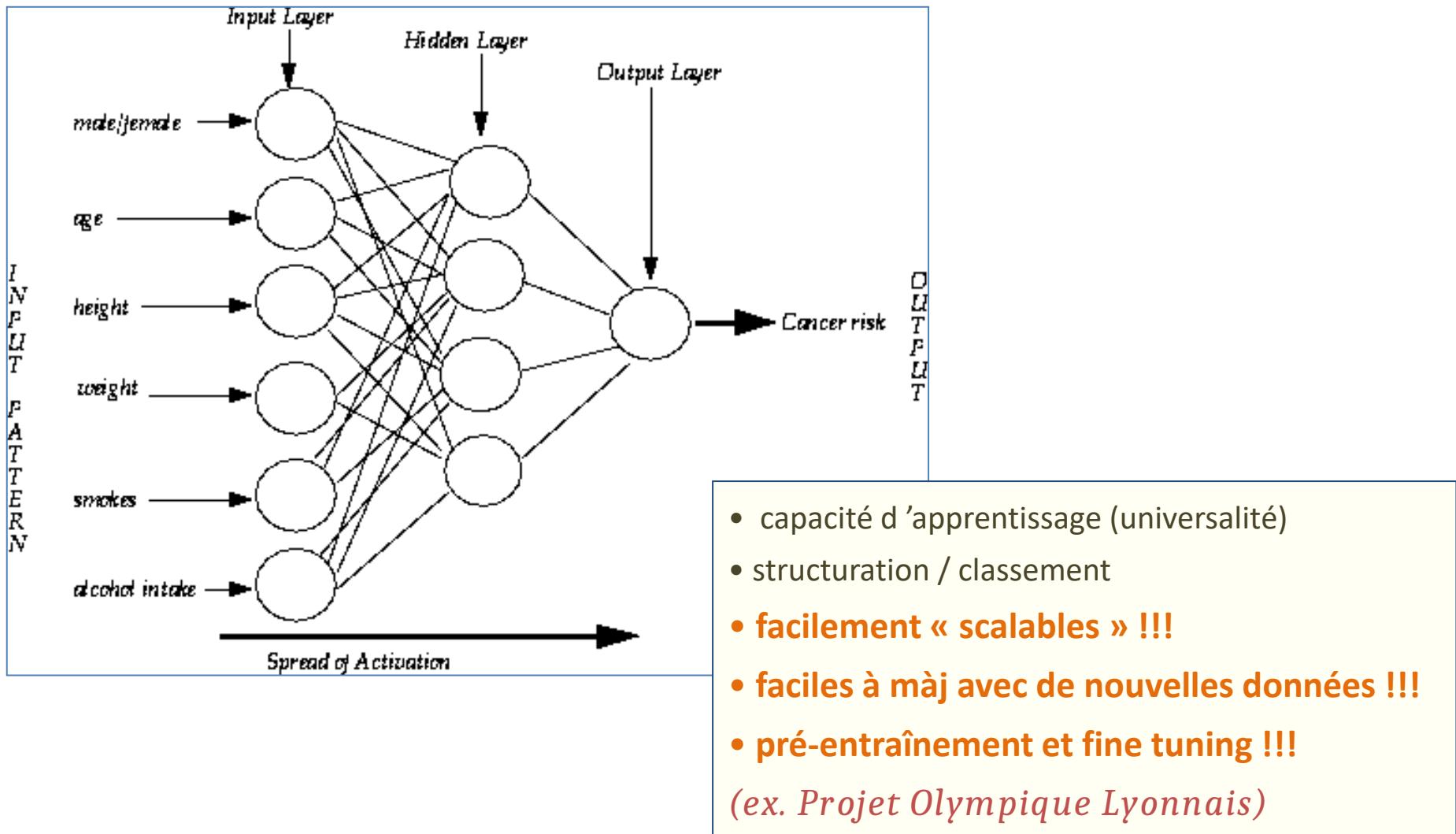
- **règles d'association**

Ex : rayonnement de magasins, les personnes qui achètent du poivre achètent également du sel

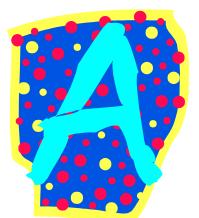
[10 Machine Learning Algorithms to Know in 2024 \(Coursera\)](https://www.coursera.org/learn/machine-learning-algorithms)



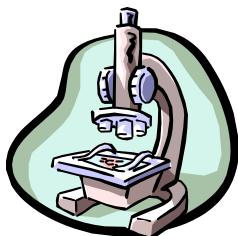
Les réseaux de neurones artificiels



Spécif.3 - Multiplicité des supports et des sources



Rôle fondamental de la préparation des données



Les applications

NLP ([SAS Text miner](#), [Sentiment analysis](#), ...)

Image mining (ex. Recherche par le contenu [Google](#), ...)

Mais déjà ancien finalement... ([JADT](#), [Zooplancton](#), etc...)



	var 1	var 2	...	var J
individu 1				
individu 2				
...			valeurs	
individu n				



Prédiction
Structuration
Description
Association

L'affaire devient particulièrement difficile lorsqu'il faut intégrer les différentes informations (nature, format, source,...) pour produire un modèle synthétique : **fouille de données complexes...**

Condition du succès d'un projet Data Mining

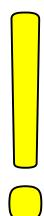
Démarche data mining





La démarche « valorisation de la data »

- formalisation des objectifs
- acquisition des données
- préparation des données
- apprentissage – application des méthodes
- interprétation – explication
- évaluation et validation
- déploiement



Ca ne marchera jamais si :

Le « métier » n'adhère pas à ce que vous faites

Les objectifs sont mal définis

Les données disponibles ne conviennent pas

Les données sont mal « préparées »

On n'utilise pas les techniques appropriées



BIG DATA

Tout le monde en parle ([Google trends](#))... c'est le terme à la mode

Tout le monde est persuadé que c'est très important

... mais de quoi il retourne exactement ?

... quel rapport avec les méthodes statistiques et de machine learning ?



BIG DATA – C'est important

Anne Lauvergeon et al., « Ambition 7 : La **valorisation** des données massives (Big Data) », in « [Un principe et sept ambitions pour l'innovation - Rapport de la commission Innovation 2030](#) », Octobre 2013 [[Rapport annoté](#)].

M.P. Hamel D. Marguerite, « **Analyse** des big data – Quels usages, quels défis », in [La note d'analyse](#), Commissariat Général à la Stratégie et à la Prospective, Département Questions Sociales, N°8, Novembre 2013 [[Rapport annoté](#)].

OCDE, « [Data-driven innovation for growth and well-being](#) », 2015.



C. Villani, « [Donner du sens à l'intelligence artificielle : pour une stratégie nationale et européenne](#) », 28 Mars 2018 [[Rapport annoté](#)].

BIG DATA – C'est dans l'air du temps

(tout le monde veut en être...)

Blog spécialisé sur « lemonde.fr »

<http://data.blog.lemonde.fr/>

Les acteurs du data mining (et des statistiques) investissent les lieux

[SAS](#), [IBM-SPSS](#), [STATISTICA](#), etc.

De nouvelles formations émergent, certaines à des tarifs qui arrachent la tête

[EM-Grenoble](#), [Telecom ParisTech](#), [ENSAI](#), [ENSAE ParisTech](#), [Ecole Centrale Paris](#), ...

Des instituts sur le Big Data se créent pour stimuler l'activité

[Canada](#), [New York](#), ...

Les « data » instaurent de nouvelles approches dans d'autres domaines

[Data journalism](#), etc., y compris [les autres domaines scientifiques](#) (astronomie, archéologie, etc.)



Quels métiers ?

Top 6 des métiers du Big Data recherché par les entreprises

<https://www.lebigdata.fr/emplois-big-data>

Les nouveaux horizons des ingénieurs

<http://etudiant.lefigaro.fr/orientation/actus-et-conseils/detail/article/les-nouveaux-horizons-des-ingenieurs-1066/>

Le **Big Data**, générateur d'emplois

<http://www.letudiant.fr/educpros/actualite/big-data-les-nouveaux-aventuriers-de-la-donnee.html>

L'APEC explique les métiers émergents de l'IT (Information technology)

<http://pro.clubic.com/emploi-informatique.clubic.com/actualite-562252-emploi-apec-metiers-emergents-it.html>



Spécificités du Big Data ?

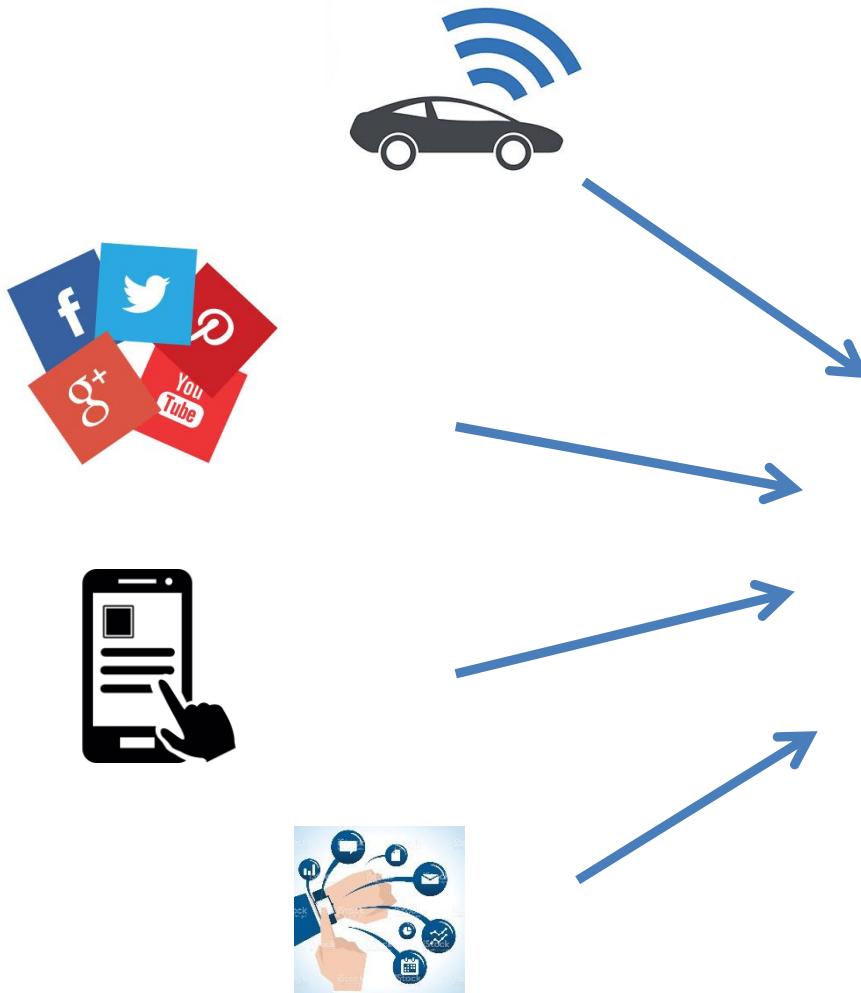
Nouvelles caractéristiques des données :

Volume – Variété – Vélocité

Parce que...

- (1) Nouvelles sources de données, nouveaux contenus ;**
- (2) Y compris les sources externes à l'entreprise.**





- (1) Enjeux de stockage (technologique)
- (2) Enjeux d'analyse (valorisation)

Variété des sources d'information, du type, des formats, fréquence des mises à jour, énorme volumétrie.

Définition ([Wikipedia](#))

DEFINITION

(Cadre)

Les big data, littéralement les grosses données, est une expression anglophone utilisée pour désigner des ensembles de données qui deviennent **tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion de l'information.**

ENJEUX

(Mobilise les énergies)

Le Big Data s'accompagne du **développement d'applications à visée analytique, qui traitent les données pour en tirer du sens**. Ces analyses sont appelées Big Analytics ou “Broyage de données”. Elles portent sur des données quantitatives complexes avec des méthodes de calcul distribué.

En 2001, un rapport de recherche du META Group (devenu Gartner) définit les enjeux inhérents à la croissance des données comme étant tri-dimensionnels : les analyses complexes répondent en effet à la règle dite des « 3V », **volume, vitesse et variété**. Ce modèle est encore largement utilisé aujourd'hui pour décrire ce phénomène.



Volume – Variété – Vélocité

VOLUME

Outils de recueil de données de plus en plus présents, dans les installations scientifiques, mais aussi et surtout dans notre vie de tous les jours (ex. cookies, GPS, réseaux sociaux [ex. lien « like » - « profils »], cartes de fidélité, les simulations en ligne sur certains sites de prêts ou d'assurance, etc.).

Il faut pouvoir les stocker et pouvoir les traiter (rapidement, efficacement) !

VARIETE

Sources, formes et des formats très différents, structurées ou non-structurées : on parle également de données complexes (ex. texte en provenance du web, images, liste d'achats, données de géolocalisation, etc.).

Il faut les traiter conjointement !

VELOCITE

Mises à jour fréquentes, données arrivant en flux, obsolescence rapide de certaines données... nécessité d'analyses en quasi temps réel (ex. détection / prévention des défaillances, gestion de file d'attente)

Il faut les traiter fréquemment (et/ou tenir compte du facteur d'obsolescence) !



Défis technologiques – Technologies Big Data

Cloud computing

Le cloud computing ... est l'exploitation de la puissance de calcul ou de stockage de serveurs informatiques distants par l'intermédiaire d'un réseau, généralement internet. Ces serveurs sont loués à la demande, le plus souvent par tranche d'utilisation selon des critères techniques (puissance, bande passante, etc.) mais également au forfait ([Wikipédia](#)). Ex. Amazon Web Services, Microsoft Azure,... [Azure Machine Learning](#).

Plateformes big data

L'architecture d'un environnement informatique ou d'un réseau est dite distribuée quand toutes les ressources ne se trouvent pas au même endroit ou sur la même machine.... Les architectures distribuées reposent sur la possibilité d'utiliser des objets qui s'exécutent sur des machines réparties sur le réseau et communiquent par messages au travers du réseau ([Wikipédia](#)). (Ex. Hadoop, Spark). Savoir programmer sous ces environnements devient un enjeu fort (cf. [tutoriels](#)).

Bases NOSQL

En informatique et en bases de données, NoSQL désigne une famille de systèmes de gestion de base de données (SGBD) qui s'écarte du paradigme classique des bases relationnelles. L'explicitation du terme la plus populaire de l'acronyme est Not Only SQL ([Wikipédia](#)). L'idée est d'acquérir plus de souplesse pour gérer notamment la variété des données (ex. [MongoDB](#), orienté document ; [Neo4j](#), orienté graphe, etc.). **Nouveau concept : [data lake](#).**



Big Data Analytics

Les Big Data Analytics désignent le processus de collecte, d'organisation et d'analyse de grands ensembles de données (Big Data) afin de découvrir de nouveaux modèles et **en tirer des informations utiles**. Les Big Data Analytics veulent fondamentalement découvrir la connaissance provenant de l'analyse des données ([Le Big Data](#)).

Aujourd'hui une priorité

Anne Lauvergeon et al., « Ambition 7 : La **valorisation** des données massives (Big Data) », in « [Un principe et sept ambitions pour l'innovation - Rapport de la commission Innovation 2030](#) », Octobre 2013 [[Rapport annoté](#)].



Les acteurs traditionnels de la statistique s'en approprient

SAS

Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analyses... may lead to more confident decision making.

http://www.sas.com/en_us/insights/big-data/what-is-big-data.html (voir les études de cas)

IBM

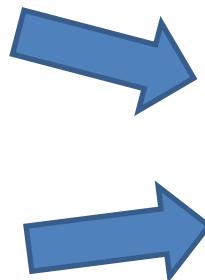
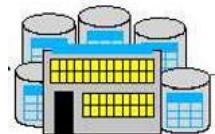
Chaque jour, nous générons 2,5 trillions d'octets de données. ... Ces données proviennent de partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels d'achats en ligne et de signaux GPS de téléphones mobiles, pour ne citer que quelques sources. Ces données sont appelées **Big Data**.... Le Big Data va bien au-delà de la seule notion de volume : il constitue une opportunité d'obtenir des connaissances sur des types de données et de contenus nouveaux...

<http://www-01.ibm.com/software/fr/data/bigdata/>



BIG DATA ANALYTICS

Données internes à l'entreprise



Données externes à l'entreprise

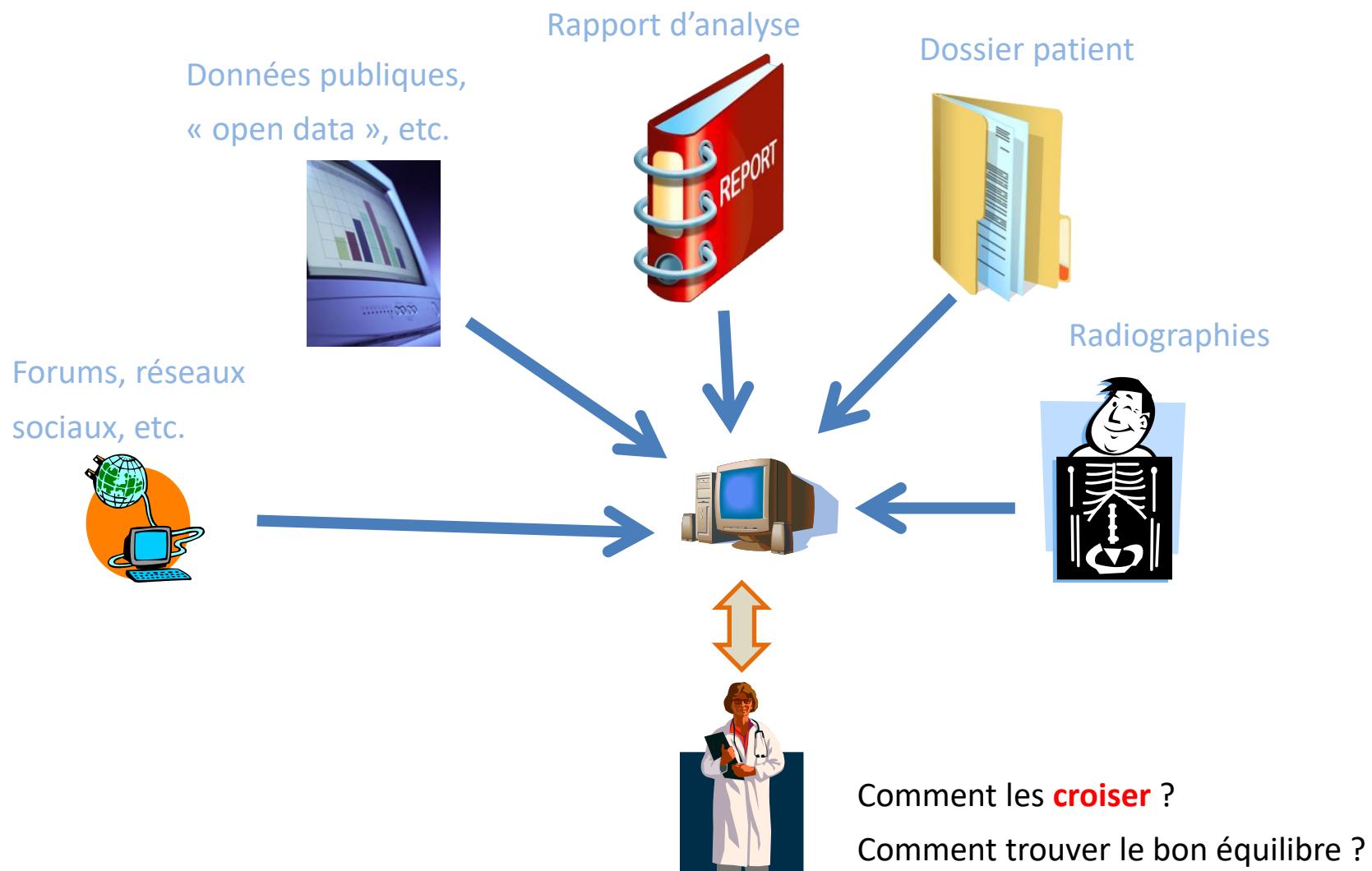
Pour rendre les analyses
plus performantes



La vague « **OPEN DATA** » va amplifier le déluge (des données)... et les attentes en termes d'analyse ([Enjeux de l'Open Data](#))

Améliorer l'intégration des données de différentes natures

Fouille de données complexes, « variété » plus et encore...



Nouvelles opportunités d'analyse

Text mining, Web mining, etc.



Services financiers

Scoring de l'emprunteur - <http://www.cbanque.com/credit/scoring-etude-dossier.php#>

« Crédit score » régit notre vie – Le « diktat de la solvabilité »

Y compris notre vie amoureuse

Grande distribution

Nous reste-t-il encore des secrets ?

Petite histoire du père américain

Cartes de fidélité - Renouvellement des informations au fil des années

Assurances

Scoring – Détermination des primes d'assurance (Amaguiz, Direct Assurances, etc.)

Assurance auto : les conductrices payeront plus cher

Sport

Dossier du Journal l'Equipe – La « data révolution » (<http://www.lequipe.fr/explore/la-data-revolution/>)

Tous les sports s'y mettent : le foot, le tennis, etc.

Autres

Les constructeurs automobiles s'y mettent (Carburant de demain, analyse prédictive, ...)

Fraude aux allocs (cibler les contrôles...), fraude à la carte bancaire (transactions suspectes...)

Présidentielles USA (cibler les électeurs et les donateurs...)

Recrutement et gestion des ressources humaines (programmes informatiques, drh, ...)



Avec de nouveaux usages (1)

Filtrage collaboratif et systèmes de recommandation

The screenshot shows the Amazon product page for 'Gil Jourdan : L'Intégrale 1'. The main product image is on the left, showing a man sitting in a chair. The title and author information are at the top. A red box highlights the 'Produits fréquemment achetés ensemble' section, which lists three related products: 'Gil Jourdan : L'Intégrale 2', 'Gil Jourdan - L'Intégrale - tome 2 - Gil Jourdan 2 (intégrale) 1960 - 1963', and 'Gil Jourdan : L'Intégrale 3'. The total price for these three items is EUR 72,00. An arrow points from this section to a callout box on the right.

Recommandation
basée sur les
transactions.

Recommandation
basée sur les
utilisateurs (clients).

Les clients ayant acheté cet article ont également acheté



Evaluations des produits



Avec de nouveaux usages et problématiques (2)

Analyse des opinions (sentiments, approbation, désapprobation, etc.). Ex. Twitter

Ex. « Sentiment Viz » - Tweet Sentiment Visualization - https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/tweet_app/

(0 : négative, 1 : neutre, 2 : positive)

	A	
1	opinion	message
2	2	Gas by my house hit \$3.39!!!! I'm going to Chapel Hill on Sat. :)
3	0	Theo Walcott is still shit, watch Rafa and Johnny deal with him on Saturday.
4	0	its not that I'm a GSP fan, i just hate Nick Diaz. can't wait for february.
5	0	Iranian general says Israel's Iron Dome can't deal with their missiles (keep talking like that and we r
6	1	Tehran, Mon Amour: Obama Tried to Establish Ties with the Mullahs http://pjmedia.com/tatler/2013/02/04/iranians-want-obama-to-establish-ties-with-the-mullahs
7	1	I sat through this whole movie just for Harry and Ron at christmas. ohlawd
8	2	with J Davlar 11th. Main rivals are team Poland. Hopefully we an make it a successful end to a toug
9	0	Talking about ACT's && SAT's, deciding where I want to go to college, applying to colleges and ever
10	1	Why is "Happy Valentines Day" trending? It's on the 14th of February not 12th of June smh..
11	0	They may have a SuperBowl in Dallas, but Dallas ain't winning a SuperBowl. Not with that quarterb
12	1	Apple software, retail chiefs out in overhaul: SAN FRANCISCO Apple Inc CEO Tim Cook on Monday
13	2	Watching English Vinglish!
14	2	One of my best 8th graders Kory was excited after his touchdown today!! He did the victor cruz!!lol
15	1	#Livewire Nadal confirmed for Mexican Open in February: Rafael Nadal is set to play at the Me... h
16	2	Men get mad at each other, fight, and get over it. Women get mad, and will hold a grudge that lasts
17	1	So Halev's next single is coming out in November..... I hope it's not true.



C'est du NLP (natural language processing) avec un cadre et des finalités particulières !!!

(longueurs des textes contraintes et homogènes, mises à jour très fréquentes, etc.)

Ah... les 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**. Projet avec Disneyland Paris (TripAdvisor, Avis Google, Booking)

The screenshot shows a comment section from the website **L'ÉQUIPE**. On the left, there is a sidebar with various navigation links: menu, accueil, directs, explore, live, le journal, and la chaîne. The main content area displays two comments:

Eder11
le 24 septembre 2022 à 09h30
105 GC hier en salle, 22 rafa, 21 Djoko, 20 Fed., 11 Borg et Laver, 7 mc enr, 6 Edberg, 4 courier et 3 murray ouf...
répondre 36 1 ?

vilarpurin
le 24 septembre 2022 à 09h39
Sport de nantis, dentistes et évadés fiscaux. L'archétype de ce qui pourrit notre monde.
répondre 16 154 ?

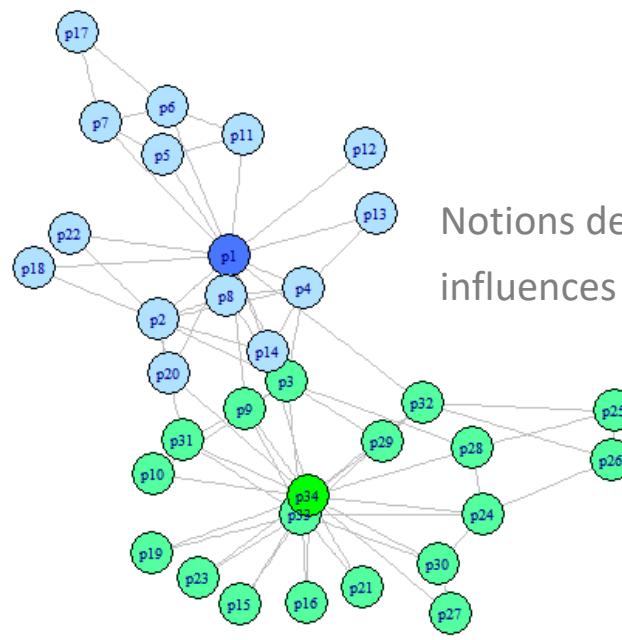
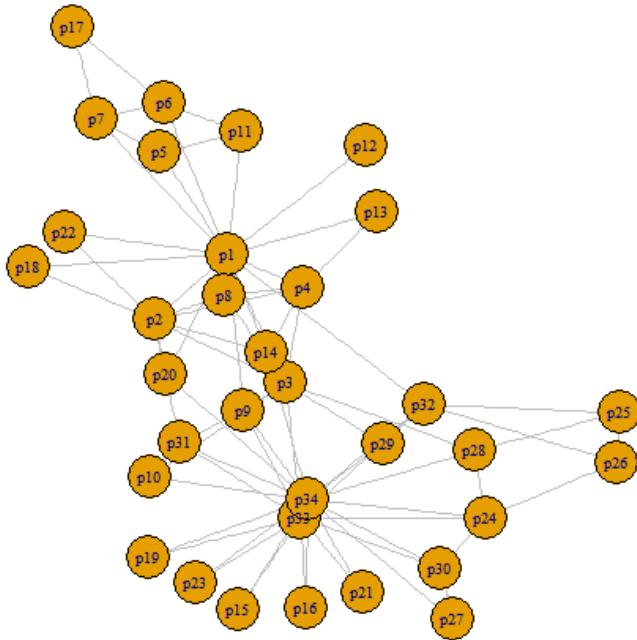
At the bottom of the comment section, there is a yellow button with the text "je m'abonne pour accéder à ce contenu >".

The screenshot shows a social media post with several comments. The post itself is from **frandu01** on September 10, 2020, at 23h04, stating: "Sergent Garcia peine à monter en grade...". Below it, **isidor31** responds on the same day at 23h51: "toi en intelligence! oups tu connais pas le mot ... cherche sur google . Mais tu l auras pas dommage." The comments have the following interaction counts:

- frandu01: 20 likes, 6 dislikes
- isidor31: 2 likes, 10 dislikes
- Alonso_Bistro: 8 likes, 1 dislike
- mauresque42: 12 likes, 6 dislikes
- antiquons: 6 likes, 3 dislikes



Détection de communautés dans les réseaux sociaux



Notions de centralité,
influences et communautés

Les idées sont anciennes mais ont connu un regain d'intérêt extraordinaire avec
l'apparition des médias sociaux ([Fergusson](#), [Paris Plage](#)).

Avec de nouveaux usages avec les modèles pré-entraînés (4)

Algorithmes de « Deep Learning »

Détection et reconnaissance des objets (la voiture grimée) – Voir aussi VGG16 et les applications des étudiants (Pokémon, etc.)



Reconnaissance faciale et détection de l'âge (<https://howold.doyoulook.com/> ; <https://age.toolpie.com/> ; ...) – Voir aussi le traitement temps réel des vidéos webcam des étudiants.



Avec de nouveaux usages, modèles pré-entraînés et IA générative (5)

Peux-tu me fournir une fonction Python qui programme l'analyse discriminante prédictive ?

?

(18 janv. 2024)

- ✓ Recherche de Python fonction analyse discriminante prédictive
 - ✓ Nous générerons des réponses pour vous...

En réalité, la question est traduite en requête ...

Bien sûr! Voici une fonction Python qui utilise le package `scikit-learn` pour effectuer une analyse discriminante linéaire prédictive:

Python

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

def predictive_discriminant_analysis(X_train, y_train, X_test):
    lda = LinearDiscriminantAnalysis()
    lda.fit(X_train, y_train)
    return lda.predict(X_test)
```

Code généré par l'IA. Examinez et utilisez soigneusement. Plus d'informations sur la [FAC](#).

Cette fonction prend en entrée un ensemble d'entraînement `X_train` et `y_train` ainsi qu'un ensemble de test `X_test`. Elle renvoie les prédictions de l'analyse discriminante linéaire sur l'ensemble de test.

J'espère que cela vous sera utile! N'hésitez pas à me poser d'autres questions si vous en avez.

[En savoir plus](#)

 eric.univ-lyon2.fr

2 statology.org

3 youtube.com



*La réponse est construite, « générée »...
(Benoît Saqot, Collège de France)*

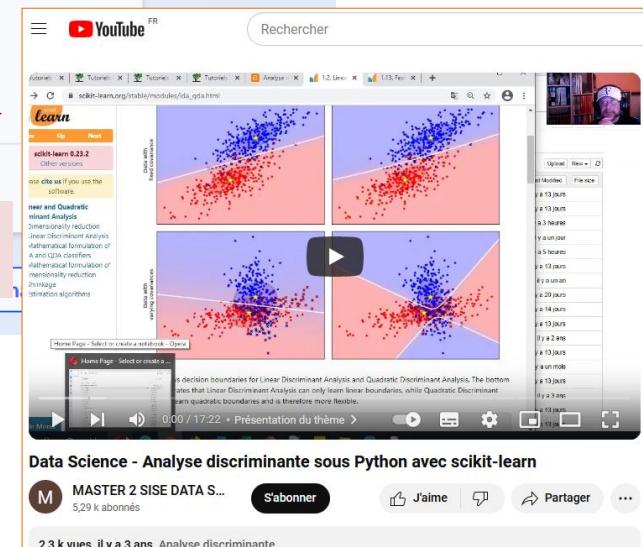
Tutoriel Tapogra

1 Introduction

Pratique de l'analyse discriminante sous Python avec le package « [scikit-learn](#) ». Comparaison des résultats avec ceux de SAS (PROC DISCRIM) et TANAGRA.

L'analyse discriminante linéaire est une méthode prédictive où le modèle s'exprime sous la forme d'un système d'équations linéaires des variables explicatives. On parle de « fonctions de classement ». Je l'affectionne particulièrement : parce qu'elle est simple et robuste ; parce qu'elle sait directement appréhender les problèmes multi-classes (nombre de modalités de la variable cible > 2) ; parce qu'elle se prête à des variétés d'interprétations qui éclairent sous autant d'angles les mécanismes de l'apprentissage supervisé (probabiliste, géométrique, ...), en ce sens elle fait le pont entre les cultures statistiques, machine learning et pattern recognition (reconnaissance de formes) ; parce qu'il existe de nombreuses bibliothèques de calcul efficaces (voir [TUTO](#), juillet 2012), ainsi que beaucoup moi-même travaillé sur l'optimisation de son implémentation (ex. « *Multithreading équilibré pour la discriminante* », juin 2013), etc. Le temps que

Quid des sources utilisées ?



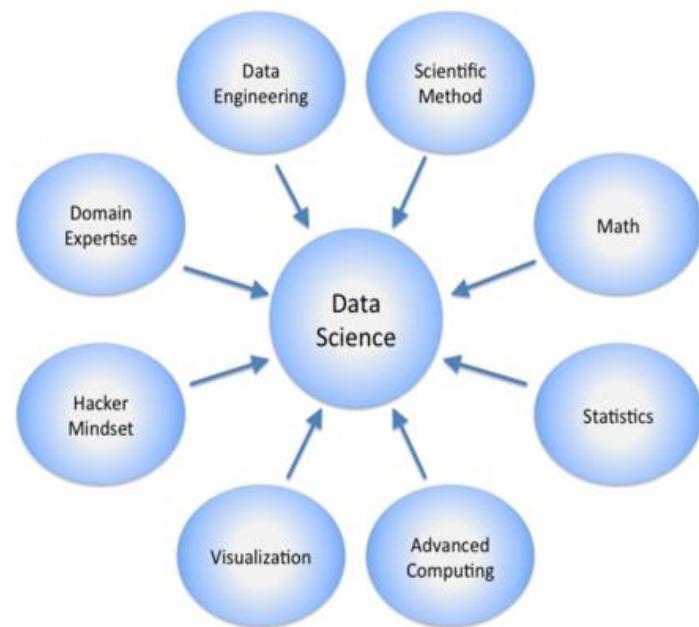
DATA SCIENCE

Finalement, de quoi il retourne ?



- Elle s'inscrit dans un contexte de profusion des données, internes aux entreprises, mais aussi externes aux entreprises. Volumétrie devient une composante clé et implique l'émergence de nouvelles technologies (technologies big data).
- Multiplicité des supports et des formats de données (BD classiques, entrepôts de données, web [texte/images/vidéo], capteurs, etc.).
- Multiplicité des domaines d'application. L'expertise du domaine est indispensable pour transformer la « relation » statistique en (1) connaissances et en (2) décisions stratégiques (attention à ne pas conclure n'importe quoi)
- Cela induit de nouvelles pratiques / démarches méthodologiques dans ces domaines.
- Importance des nouvelles technologies (ex. technologies big data, cloud, etc.).

Il s'agit bien d'extraire de la connaissance à partir de données



https://fr.wikipedia.org/wiki/Science_des_données



Synergie forte entre l'informatique et les statistiques / mathématiques.

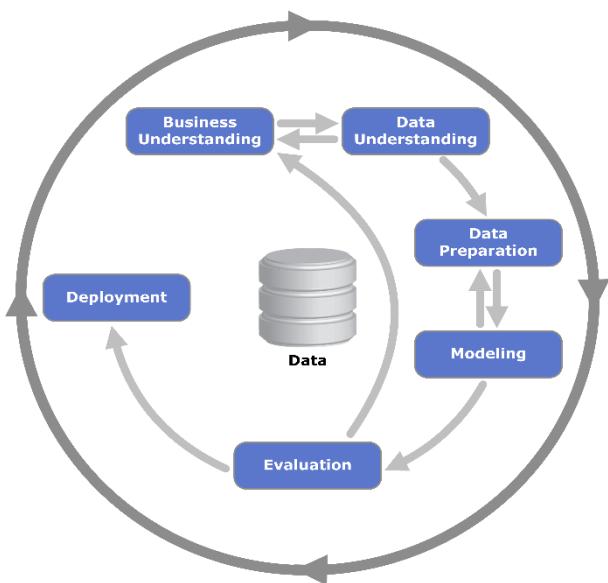


DATA MINING vs. DATA SCIENCE

Tout devient source de données

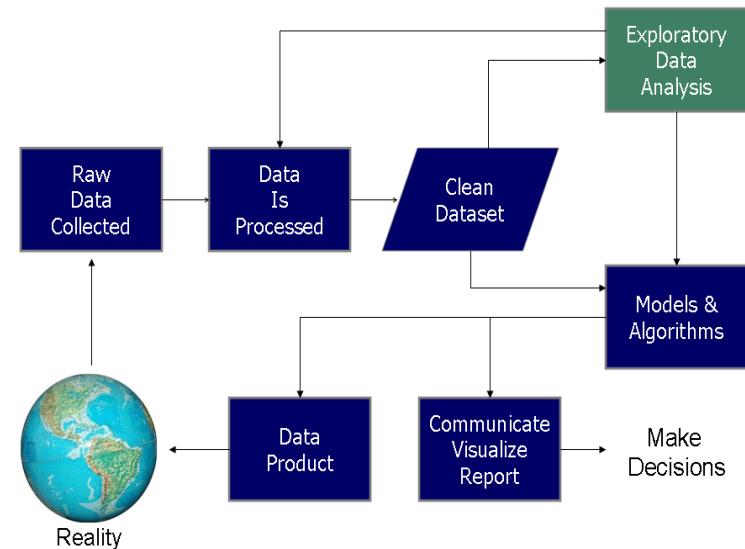


IBM CRISP DM



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Data Science Process



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



Compétences du Data Scientist

Soyons concrets...

Offre d'emploi AI LAB – BNP PARIBAS (Sept. 2018)

The screenshot shows a job listing on the BNP Paribas website. The header features the BNP Paribas logo and the tagline "La banque d'un monde qui change". The main heading is "NOUS RECHERCHONS UN DATA SCIENTIST / AI LAB". Below the heading, there's a brief description: "Type de contrat CDI", "Niveau d'études BAC+4/5", "Expérience 2 à 5 ANS", and a map showing the location in Paris.

Etes-vous notre prochain Data Scientist ?

Oui, si vous êtes diplômé(e) d'un Bac+5 en Ecole d'Ingénieur ou équivalent universitaire avec une spécialité en **Data Science, Big data, Machine Learning** et vous justifiez de deux années d'expérience dans l'un de ses domaines.

Les compétences techniques :

- | Vous maîtrisez un/ plusieurs langages de programmation (ex: Python, JavaScript, Go, Java, C++ ...)
- | Vous avez de fortes compétences en analyse statistique et quantitative
- | Vous avez une connaissance des bases de données, et avez une expérience avec les outils ETL (Dataiku, Alteryx ...)
- | Vous êtes sensibles aux enjeux de la BI, et connaissez des outils de visualisation (Tableau software, Qlikview)
- | Vous avez des compétences poussées en Machine Learning : SVM, Boosting, Hidden Markov Models, analyses de séries temporelles, réseaux de neurones (CNN, LSTM, GRU ...)
- | Toutes expériences en Data Mining, Text Mining, utilisation de NLP et technologies sémantiques sont également les bienvenues.
- | Vous parlez couramment anglais (le français et/ou le portugais sont un plus !)



Logiciels

Outils « classiques » d'obéissance statistique et machine learning (informatique)

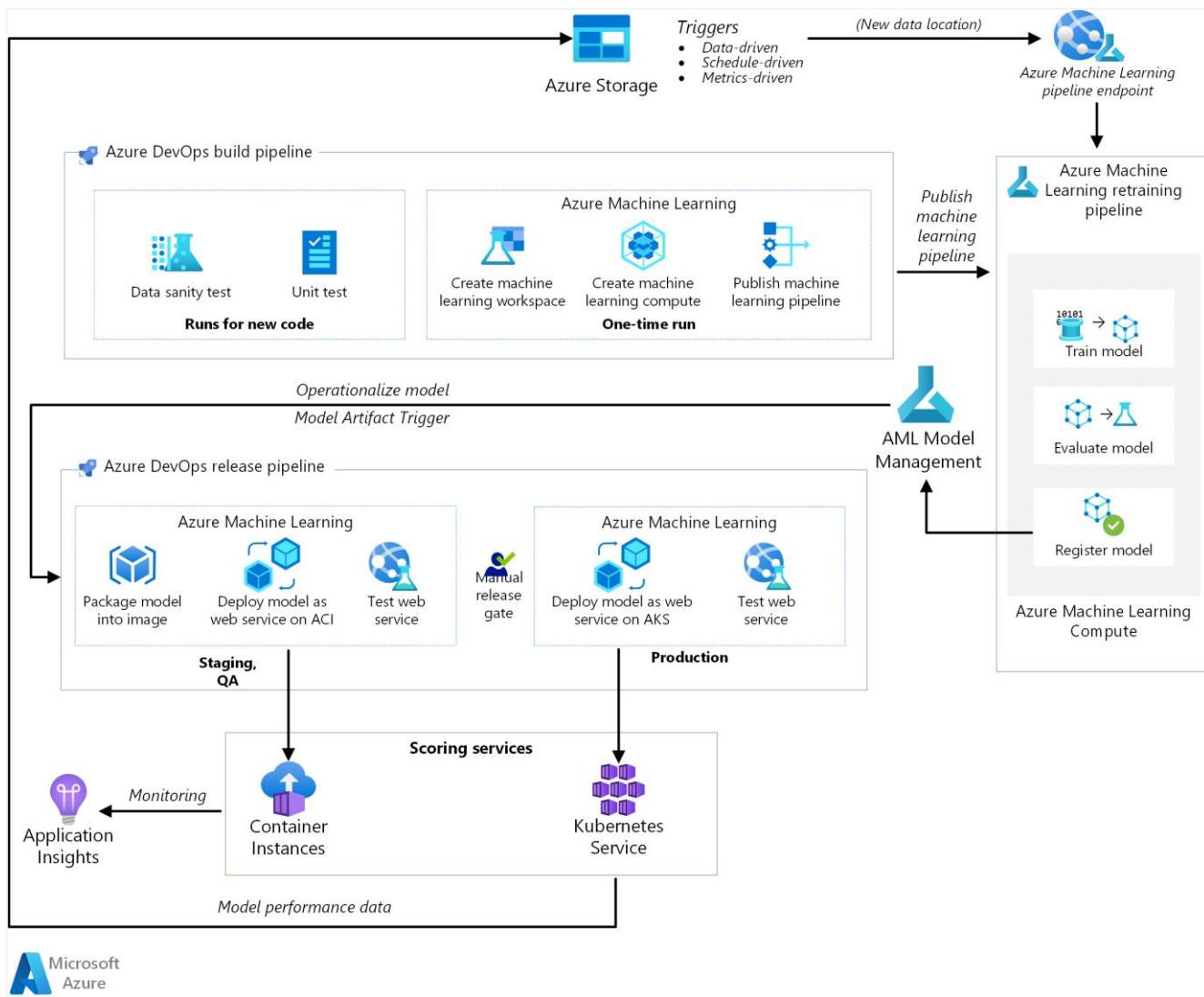
Excel, Python, R, SAS – OK (cf. offres d'emploi APEC)

→ Mais, aujourd'hui, il faut être vigilant, le **cloud** devient incontournable. (cf. Azure – Lyon 2 page suivante)

Travailler dans le cloud

Tendance lourde aujourd'hui (2023), les entreprises transfèrent leurs systèmes dans le cloud, sur toute la chaîne de la data (ex. acquisition, stockage → déploiement des modèles prédictifs dans le cloud par dockérisation... cf. Snowflake, Azure, AWS, Google...).

Ex. [Ateliers -- Pipelines sous Azure \(cf. Workspaces\)](#)
Ex. [Tutoriel](#), déjà en 2014 !
Ex. [Atelier des étudiants](#) (2016)



EXCEL (le tableau en général)

Tout le monde sait (ou croit savoir) le manipuler – Simple à utiliser

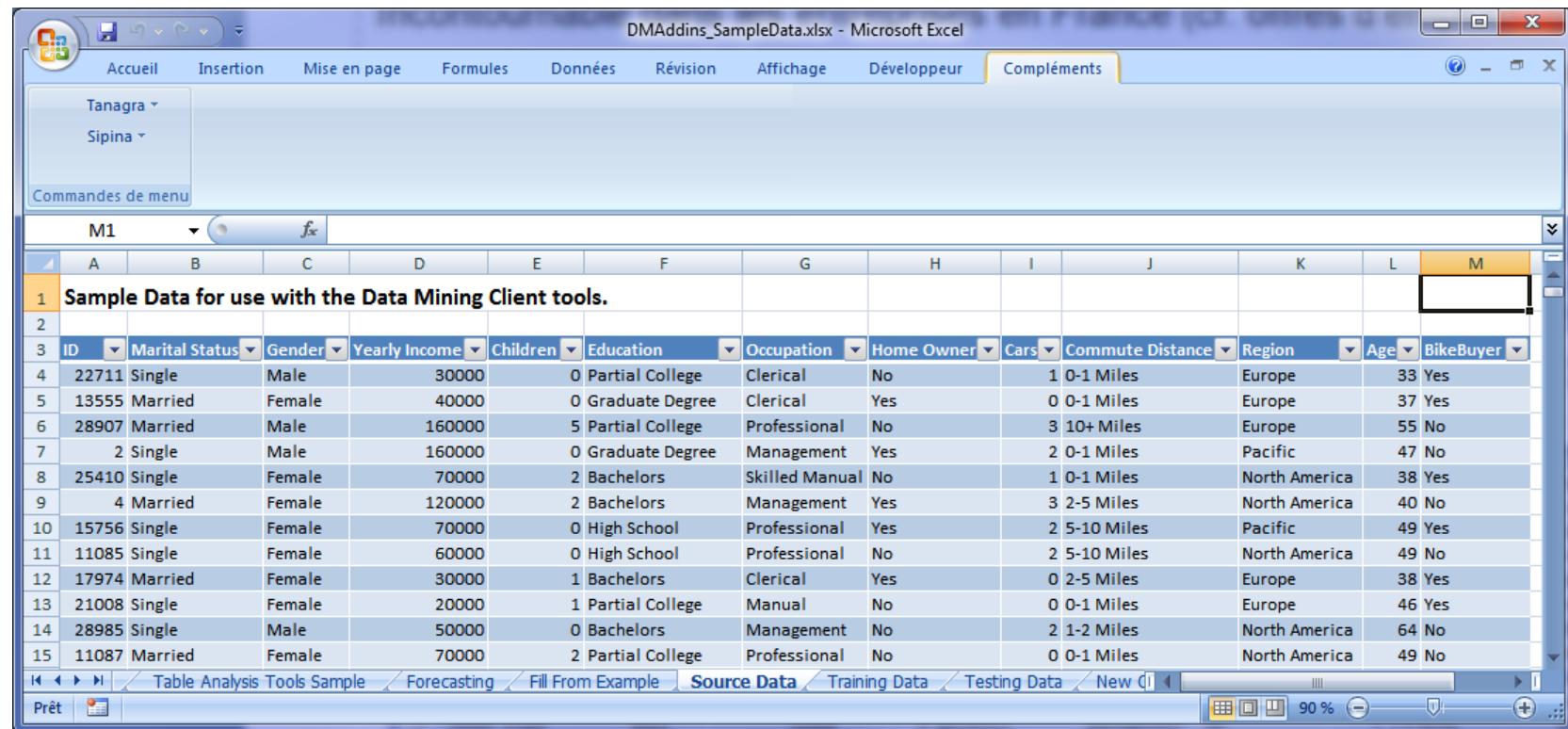
Fonctionnalités de manipulation et de préparation de données

Possibilité d'aller plus loin avec la programmation (VBA)

Possibilité d'extension via les add-ins (ex. [SQL Server](#), [SAS](#), [Real Statistics](#), [Tanagra](#), etc.)

Incontournable dans les entreprises en France (cf. offres d'emploi sur le site de l'[APEC](#))

Incontournable au niveau mondial (cf. Sondage annuel KD Nuggets)



The screenshot shows a Microsoft Excel window titled "DMAAddins_SampleData.xlsx - Microsoft Excel". The ribbon menu is visible with tabs like Accueil, Insertion, Mise en page, Formules, Données, Révision, Affichage, Développeur, and Compléments. The "Compléments" tab is selected. On the left, there's a ribbon of add-ins including Tanagra and Sipina. The main area displays a table titled "Sample Data for use with the Data Mining Client tools." The table has 15 rows of data and 14 columns. The columns are labeled: ID, Marital Status, Gender, Yearly Income, Children, Education, Occupation, Home Owner, Cars, Commute Distance, Region, Age, and BikeBuyer. The data includes various demographic and socioeconomic information. At the bottom, there are tabs for Table Analysis Tools Sample, Forecasting, Fill From Example, Source Data, Training Data, Testing Data, New CI, and a Prêt tab.

	ID	Marital Status	Gender	Yearly Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	BikeBuyer
1	22711	Single	Male	30000	0	Partial College	Clerical	No	1	0-1 Miles	Europe	33	Yes
2	13555	Married	Female	40000	0	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	37	Yes
3	28907	Married	Male	160000	5	Partial College	Professional	No	3	10+ Miles	Europe	55	No
4	2	Single	Male	160000	0	Graduate Degree	Management	Yes	2	0-1 Miles	Pacific	47	No
5	25410	Single	Female	70000	2	Bachelors	Skilled Manual	No	1	0-1 Miles	North America	38	Yes
6	4	Married	Female	120000	2	Bachelors	Management	Yes	3	2-5 Miles	North America	40	No
7	15756	Single	Female	70000	0	High School	Professional	Yes	2	5-10 Miles	Pacific	49	Yes
8	11085	Single	Female	60000	0	High School	Professional	No	2	5-10 Miles	North America	49	No
9	17974	Married	Female	30000	1	Bachelors	Clerical	Yes	0	2-5 Miles	Europe	38	Yes
10	21008	Single	Female	20000	1	Partial College	Manual	No	0	0-1 Miles	Europe	46	Yes
11	28985	Single	Male	50000	0	Bachelors	Management	No	2	1-2 Miles	North America	64	No
12	11087	Married	Female	70000	2	Partial College	Professional	No	0	0-1 Miles	North America	49	No



```
*D:\DataMining\Datasets_for_mining\dataset_for_mining
Fichier Édition Recherche Affichage Encodage
pipeline.py
7 #librairie pandas
8 import pandas
9 #chargement de la feuille de c
10 #version des données à 4 varia
11 vote_subset = pandas.read_excel('dataset_for_mining.xlsx')
12 print(vote_subset.info())
13 #importation de la librairie
14 from fanalysis.mca import MCA
15 #instanciation
16 acm = MCA(var_labels=vote_subset.columns[:4])
17 #apprentissage
18 coord = acm.fit_transform(vote_subset.iloc[:, :4].values)
19 #affichage des valeurs propres
20 print(acm.eig_)
21 #valeurs propres - graphique
22 print(acm.plot_eigenvalues())
23 #coordonnées des colonnes
24 print(acm.col_topandas())
25 #nombre var. actives
26 p = vote_subset.shape[1]-1
27 print(p)
28 #calcul des fonctions de projection
29 import numpy
30 fproj = numpy.zeros(acm.col_coord_.shape)
31 #pour chaque colonne
32 for j in range(fproj.shape[1]):
33     ... fproj[:,j] = acm.col_coord_[:,j]/(p*numpy.sqrt(acm.eig_[0,j]))
34 #affichage fonction
35 print(fproj)
36 #affichage plus avenant des deux premiers facteurs
37 print(pandas.DataFrame(fproj, index=acm.col_labels_))
38 #taille du tableau de données présenté à l'ADL
39 print(coord.shape)
40 #10 premières lignes
41 print(coord[:10,:])
42 #classe pour l'analyse discriminante
43 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
44 #instanciation
45 adl = LinearDiscriminantAnalysis()
46 #apprentissage
47 adl.fit(coord, vote_subset.group)
48 #affichage des coefficients des fonctions de classement
49 print(adl.coef_)
50 #la constante
51 print(adl.intercept_)
```

Ligne de commande + langage de programmation

Multi-paradigme (... dont statistique, analyse de données, machine learning)

Extensible à l'infini avec le système des librairies

Une des références avec R ([Top Software for Analytics, 2019](#))

Documentation très abondante (*trop parfois, il faut savoir chercher*)

Python

(<https://www.anaconda.com/download/>)

Exemple : la méthode DISQUAL

R

Ligne de commande + langage de programmation

Multi-paradigme (statistique, analyse de données, machine learning)

Extensible à l'infini avec le système des packages

Une des références avec Python (Top Software for Analytics, 2018)

Documentation très abondante (*trop parfois, il faut savoir chercher*)

The screenshot shows the official R Project website at <http://www.r-project.org/> and the RGui (64-bit) application window.

R Project Website:

- Left sidebar: About R, What is R?, Contributors, Screenshots, What's new?, Download, Packages, CRAN.
- Middle section: R Project Foundation, Members & Donors, Mailing Lists, Bug Tracking, Developer Page, Conferences, Search.
- Bottom sections: Documentation (Manuals, FAQs, The R Journal, Wiki, Books, Certification, Other), Misc (Bioconductor, Related Projects, User Groups, Links).

RGui Application:

- Top menu: File, Edit, View, Misc, Packages, Windows, Help.
- Toolbar: File, Edit, View, Misc, Packages, Windows, Help.
- Console window:

```
R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

During startup - Warning message:
Setting LC_CTYPE=en_US gnumeric & failed
> |
```

Annotations:

- A yellow box highlights the text: "Des éditeurs de code spécialisés existent : R-Studio, StatET : Plug-in pour Eclipse, etc... Des versions payantes sont apparues (ex. Revolution R pour le big data..., etc.)"

This server is hosted by the Institute for Statistics and Mathematics of WU

Diagramme de traitements (sur les standards des outils commerciaux, cf. [IBM SPSS](#)

[Modeler](#), [SAS Enterprise Miner](#), [SPAD](#), [STATISTICA](#), ...)

« Programmation » visuelle (boucles, programmation modulaire / meta nodes, ...)

Extensible avec des [plug-ins](#) (Weka, bibliothèques spécialisées ex. text mining, ...)

Multithread et possibilité de [swap](#) sur disque (armé pour les [gros volumes](#) ?)

Le logiciel est gratuit mais ... versions 'desktop' et 'professional'...

The screenshot displays the KNIME graphical user interface. On the left, there's a sidebar with links to 'PRODUCTS', 'APPLICATIONS', 'PARTNERS', 'SERVICES', 'RESOURCES', and 'COMPANY'. Below this, a section titled 'Why I KNIME...' features a screenshot of the software and the text 'KNIME makes mining fun.' A 'Data Analytics' section is also present. The main workspace shows a complex workflow diagram with nodes like 'XML Reader', 'XPath', 'Ungroup', 'String Manipulation', and 'Interactive Table'. A 'Download' button is visible at the top right of the workspace. On the right side, there are several panels: 'Node Description', 'XML Reader' (describing how to read XML documents), 'Dialog Options', 'Selected File' (specifying the XML file to read), 'XPath Filter' (a checkbox for filtering nodes based on XPath queries), and 'XPath Query' (a dropdown menu for selecting the query). At the bottom, a 'Console' window shows the startup message: 'Welcome to KNIME v2.9.1.0041089 - the Konstanz Information Miner *** Copyright, 2003 - 2013, Uni Konstanz and KNIME GmbH, Germany ***' and 'Log file is located at: C:\Program Files\knime_2.4.2\workspace\metadata\knime.log'. A footer at the bottom left includes icons for 'Modeling', 'Visualization', 'Professional', 'Integration', 'Modularity', and 'User Interface'.



Ancien, piloté par menu

Se plugge dans Excel ([KDnuggets Polls](#), May 2013)

Spécialisé dans les arbres de décision ([Kdnuggets Polls](#), [Algorithms](#), Nov 2011)

Sipina - Arbres de décision

Un logiciel gratuit de data mining pour l'induction des arbres de décision

Accueil **Versions** **Méthodes** **Capacités** **Références**

Liens

- [Page d'accueil](#)
- [Téléchargement](#)
- [Tutoriels pour Sipina](#)

Diaporama

Sipina

SIPINA est un logiciel gratuit de Data Mining spécialisé dans l'induction des arbres de décision. Curieusement, c'est un des très rares outils en libre accès intégrant des fonctionnalités interactives lors de la construction d'un arbre de décision. Fonctionnalités qui, pourtant, font tout le sel de cette méthode dans une activité de fouille de données.

SIPINA implémente également d'autres méthodes supervisées. Mais son intérêt est moindre dans ce contexte. Depuis le développement et la diffusion de TANAGRA (Janvier 2004), je conseille systématiquement d'utiliser ce dernier. Il comporte non seulement les méthodes supervisées mais également une grande majorité des techniques de statistique et d'analyse de données telles que les analyses factorielles, la classification automatique, etc., et la possibilité de les faire coopérer entre elles.

Les différentes versions de SIPINA sont disponibles sur le web depuis 1995. La version actuelle n'a guère évolué depuis 2000. Elle est néanmoins distribuée car, comme je le disais plus haut, il y a très peu d'équivalents gratuits au monde. Le site de distribution anglais est régulièrement consulté encore à ce jour, et le logiciel téléchargé. Il doit bien avoir une raison à cela. J'ai donc décidé de la documenter un peu plus, aspect totalement négligé à l'époque de son développement. Je redécouvre d'ailleurs ainsi de très nombreuses fonctionnalités imaginées, expérimentées, et finalement connues de moi seul... autant que tout le monde en profite.

Configuré judicieusement, SIPINA peut traiter de très gros volumes (plusieurs millions d'observations, plusieurs milliers de variables) tout en conservant ses fonctionnalités interactives.

Ce site rassemble tout le matériel concernant SIPINA. Autre évolution notable, il est entièrement en français, le site initial ayant toujours été exclusivement en anglais. Le logiciel reste en anglais, mais les mots clés sont relativement simples à appréhender.

SIPINA est totalement gratuit, quel que soit le contexte d'utilisation.

Ricco Rakotomalala.

Méthodologie des arbres

Prise en main de Sipina

Add-ins pour tableurs

Tutoriel : diagnostic d'une maladie cardio-vasculaire

Solutions grandes bases

- [Swap - Traitements sur disque](#)
- [Multithreading](#)
- [Echantillonnage](#)
- [Formats de fichiers spécifiques](#)

L'unique outil gratuit au monde proposant les fonctionnalités interactives des logiciels commerciaux.

Homologues commerciaux

[SAS](#), [SPAD](#), [STATISTICA](#), [IBM/SPSS](#), etc.

Diagramme de traitements (standard actuel), arborescent

Se plugge dans Excel – Les résultats sont directement récupérables

Multi-paradigme (statistique, analyse de données, machine learning)

Simplicité, facilité d'utilisation, documentation très abondante (FR et EN)

The screenshot shows two browser windows. The main window displays the Tanagra homepage with a navigation bar featuring icons for Présentation, Galerie, Caractéristiques, Didacticiels, Téléchargement, and Sipina. A yellow callout box on the right states: "Aujourd'hui, essentiellement un outil pédagogique." Below the homepage, a sub-section titled "Le projet TANAGRA" is visible. A second browser window is overlaid, showing a blog post titled "Tutoriels Tanagra pour le Data Mining". This post discusses didactic cases for Tanagra, including regression analysis for ranking. Arrows point from the "Didacticiels" link on the main page to the "Tutoriels" section of the blog post.

Aujourd'hui, essentiellement un outil pédagogique.

Présentation Galerie Caractéristiques Didacticiels Téléchargement Sipina

Le projet TANAGRA

Arbres de décision – Introduction

L'apprentissage par partitionnement. Construction d'un arbre de décision sur un jeu de données synthétique.

Principaux points à considérer lors de l'induction d'un arbre de décision à partir de données.

Arbres de décision – CHAID, C4.5 et CART

Les trois principales méthodes d'induction d'arbres proposées dans les logiciels. Les différences et les points communs.

Arbres de régression

La régression par arbres. Une méthode non-linéaire de régression. Rapprochement avec les arbres de décision.

Analyse discriminante prédictive

Un modèle paramétrique de discrimination. Analyse discriminante de Fisher. Évaluation globale du modèle, évaluation individuelle des

Tutoriels Tanagra pour le Data Mining

MARDI 29 AVRIL 2014

Régression linéaire pour le classement

Ces slides font suite au tutoriel consacré à l'utilisation de la régression linéaire multiple dans un problème de classement. La trame est la même. On montre (1) que la démarche est tout à fait licite lorsque la variable cible est binaire, il existe une passerelle avec l'analyse discriminante linéaire ; (2) les tests d'évaluation globale du modèle et de pertinence des variables sont applicables.

Mots clés : apprentissage supervisé, discrimination, classement, régression linéaire multiple, analyse discriminante linéaire, significativité globale du modèle, significativité individuelle des coefficients, R2, lambda de wilks

Lien : [regression_multiple_pour_le_classement.pdf](#)

Références :

Tutoriel Tanagra, "Analyse discriminante et régression linéaire", avril 2014.

Publié par Tanagra le 29.4.14

Libellés : Analyse discriminante, App. Supervisé - Scoring, Régression

TANAGRA – Classement automatique de planctons (Image mining)

Image originelle fournie par le scanner

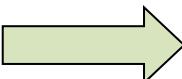
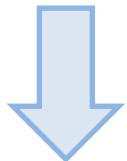


Image traitée en niveau de gris, à partir de laquelle sont calculés les paramètres

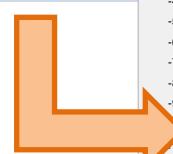


Avec l'outil
ImageJ

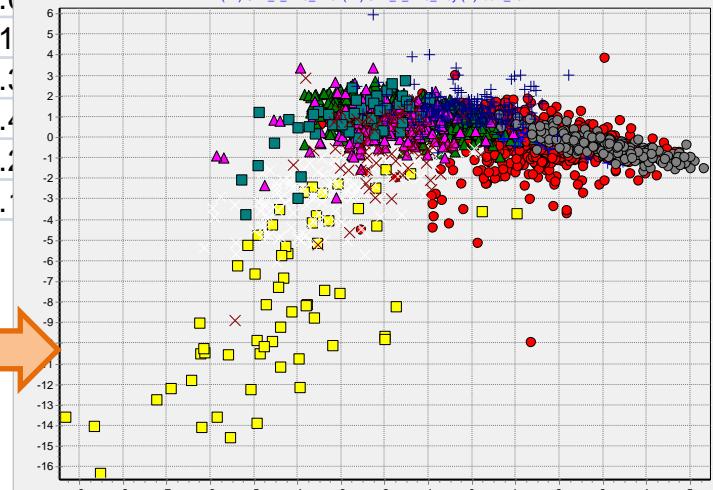


L'expert étiquette manuellement les objets

Ident_10	IntDen	Mean	StdDev	Mode	Min	Max
_detritus	276356	246.97	2.35	248	237	255
Copepods	568486	166.42	65.2	247	81	249
_detritus	173151	191.33	34.91	248	111	249
_detritus	858671	237.53	10.0	248	111	249
Copepods	403737	185.29	51.1	248	111	249
Copepods	921755	150.98	75.1	248	111	249
Chaetognaths	1017831	194.28	39.4	248	111	249
_Noctiluca	648439	226.49	35.2	248	111	249
Appendicularians	1564533	199.23	47.1	248	111	249



(X1) CDA_1_Axis_1 vs. (X2) CDA_1_Axis_2 by (Y) ident_10



Ex. de traitement :
description factorielle



RAPIDMINER

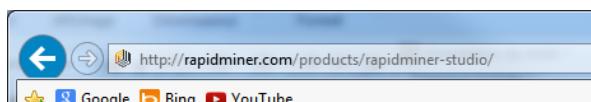
(<http://rapidminer.com/>)

Diagramme de traitements (sur les standards des outils commerciaux)

« Programmation » modulaire (meta nodes, ...)

Extensible avec des **plug-ins** (Weka, bibliothèques ex. text mining, ...)

La version gratuite est maintenant bridée...



The screenshot shows the RapidMiner Studio interface. On the left, there's a sidebar with links like 'PRODUCTS', 'SOLUTIONS', 'PRICING', 'LEARNING', 'DOWNLOAD', and 'ABOUT'. A prominent orange 'FREE TRIAL' button is visible. The main area features a 'Process' tab with a 'Main Process' diagram. The diagram consists of several nodes: 'Read Excel' (with 'fil' and 'out' ports), 'Process Docu...' (with 'wor' and 'exa' ports), and 'W-J48' (with 'tra' and 'mod' ports). Arrows connect these nodes. To the right of the process diagram are several panels: 'Context', 'Parameters', 'Process' (showing parameters like 'verbosity', 'logfile', etc.), 'Problems' (showing 'No problems found'), 'Log' (showing 'The root operator'), and 'Comment' (with a 'Help' link). At the bottom, there's a 'Synopsis' section with the text 'The root operator'.

RapidMiner Studio

Easy-to-use visual environment for predictive analytics. No programming required.

Forget sifting through code! RapidMiner is easily the most powerful and intuitive graphical user interface for the design of analysis processes. You can also choose to run in batch mode. Whatever you prefer, RapidMiner has it all.

[Compare Editions](#)



Ricco Rakotomalala

Tutoriels Data Science - <https://tutoriels-data-science.blogspot.com/>

RAPIDMINER – Discrimination de protéines

http://fr.wikipedia.org/wiki/Structure_des_protéines

http://fr.wikipedia.org/wiki/Famille_de_protéines



protéines.xls [Mode de compatibilité]

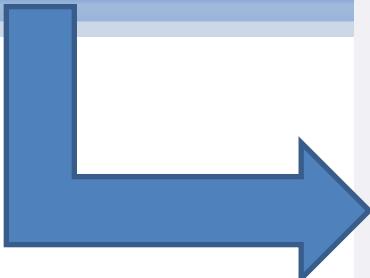
A	famille	description	B
1	F1	SQFRVSPLRTWNLGETVELKCQVLLSNPTSGCSWLFQPRGAAASPTFLLYSQNPKAAEGLDTQRFSKRLGDTFVLTLSDFRRNEGYFCALSNSIMYFSHFVPVFLPA	
2	F2	AVSKVYARSVYDSRGNPTEVELTTEKGVFVRSIVPSGASTGVHEALEMRDGDKSKWMGKGVHAVKNVNDVIAPAFVKANIDVKDQKAVDDFLISLDGTANKSKLGANAILGVSLAA	
3	F1	EPKFTKCRSPERETFSCHWTDEVHHGPIQLFYTRRNTEWTQEWEKCPDYVSAGENSCYFNSSFTSIWIPYCILTSNGGTDEKCFSVDEIVQP	
4	F1	LQQPTIQSFEQVGTKVNVTVEDERTLVRRNNNTFLSLRDVGFKDLIYTLYWKSSGKKTAKTNTEFLIDVDKGENYCFSVQAIPSRTVNRKSTDSPVECMG	
5	F1	SRCTHLENRDFVTGTQGTTRTLVLGGCVTITAEGKPSMDVWLDAIYQENKIVYTVKVEPHTGDYVAANETHSGRKTAFTISSEKTILTMGEYGDVSLLCRVASGPVAHIEGTKYHLKS	
6	F1	GSDWVIPPINLPENSRGPFQELVRIRSGRDKNLSRLYSVTGPGADQPPGIFIIINPISGQLSVTKPLDRELIARFHRLRAHAVDINGNQVENPIDIVINVIMNDNRPEF	
7	F1	ISGMSGRKASGSSPTSPINANKVENEDAFLLEEVAEEKPHVKPVFTKTILDMDVVEGSAARFDCKVEGYPDPEVMWFKDNPVKESRHFQIDYDEEGNCSLTISEVCGDDDAKYTCKAV	
8	F2	AVSKVYARSVYDSRGNPTEVELTTEKGVFVRSIVPSGASTGVHE	GTANKSKLGANAILGVSLAA
9	F2	MKIDAIIEAVIVDVPTKRPPIQMSITTVHQQSIVRVYSEGLGV	AKAAVEMALLDLKARALGV
10	F2	MERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGA	'ANLVFNNGIDAFYARCEQV
11	F2	APAPVKQGPSTVAYVEVNNNSMLNVGKYTLADGGGNNAFDVA	AGFANFPSQQAASAFAKQL
12	F2	SKIFDFVKPGVITGDDVQKVFQVAKENNfalPAVNCVGTDSIN	EHYGVPVILHTDHCAKKLLPV
13	F2	MNSNLRGVMAALLTPFDQQQALDKASLRLVQFNIIQQGIDGL	FDAVSAVTPFYYPFSFEEHC
14	F2	VQPTPADHFTFGLWTVGWTGADPFGVATRANLDPVEAVHKL	DGGFTSNDRSIRRFLAKVL
15	F2	KKTKVWCTGPKTESEEMIAKMLDAGCMNIVMPINESCHGQYAEW	WIEDEUJKAUCCGEMUWICKI
16	F2		

W-J48

J48 pruned tree

GVF <= 0

```
| AIA <= 0
|   | AES <= 0
|   |   | AKA <= 0
|   |   |   | AEA <= 0
|   |   |   |   | AGQ <= 0
|   |   |   |   |   | KVA <= 0
|   |   |   |   |   |   | NNG <= 0
|   |   |   |   |   |   |   | DIP <= 0: F1 (46.0)
|   |   |   |   |   |   |   |   | DIP > 0: F2 (3.0/1.0)
|   |   |   |   |   |   |   |   | NNG > 0: F2 (3.0)
|   |   |   |   |   |   |   |   | KVA > 0: F2 (4.0)
|   |   |   |   |   |   |   |   | AGQ > 0: F2 (3.0)
|   |   |   |   |   |   |   |   | AEA > 0: F2 (6.0)
|   |   |   |   |   |   |   |   | AKA > 0: F2 (4.0)
|   |   |   |   |   |   |   |   | AES > 0: F2 (6.0)
|   |   |   |   |   |   |   |   | AIA > 0: F2 (10.0)
|   |   |   |   |   |   |   |   | GVF > 0: F2 (15.0)
```



Autres outils

ORANGE



The screenshot shows the Orange Data Mining software interface. At the top, there's a navigation bar with links to Features, Download, Add-ons, Documentation, Development, Forum, and Blog. Below the navigation bar, there's a section titled "orange" with a logo. A central window displays a "CN2 Rules Viewer" showing a list of rules for classifying data. The rules are listed with their length, quality, coverage, and distribution. The interface also includes a "Attribute Statistics" panel, a "Value sorting" panel, and a "Orange Canvas" workspace at the bottom.

Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Components for machine learning. Add-ons for bioinformatics and text mining. Packed with features for data analytics.



(Downloads for other systems and versions)

Latest Blog Entries

WEKA / PENTaho



The screenshot shows the "Data Mining - Weka" page from the Pentaho website. The page features a large image of a computer monitor displaying green data points. Below the image, there's a heading "Explore and understand your data". A paragraph explains that Weka is an open-source software for data mining. On the left, there are tabs for "Description", "Main concepts", and "Contribute". On the right, there's a "QUICK LINKS" sidebar with links to Community Documentation, FAQ, Mailing List, Data Sets, and Screenshots.



Ricco Rakotomalala

Tutoriels Data Science - <https://tutoriels-data-science>

Sans oublier les outils commerciaux

Qui se distinguent souvent par :

- Performances (rapidité, traitement des grandes bases)
- Qualité et rigueur
- Utilisabilité (efficacité, ergonomie)
- **Existence d'un support professionnel !!!**

Quasiment tous
maintenant proposent
un mode opératoire
client/serveur

Quelques grands acteurs historiques des statistiques :

- SPAD (via [COHERIS Analytics Spad](#))
- SAS (via [SAS EM](#))
- IBM SPSS (via [Modeler](#))
- STATISTICA [Data Miner](#)

Mais aussi des acteurs des bases de données :

- Microsoft SQL SERVER [Data Mining](#)
- ORACLE [Data Mining](#)
- Microsoft [AZURE MACHINE LEARNING](#)... cloud, l'avenir...



Bibliographie



Wikipédia, « [Exploration des données](#) ».

IBM, « [CRISP-DM – Cross Industry Standard Process for Data Mining](#) », 2012.

M.P. Hamel D. Marguerite, « [Analyse des big data – Quels usages, quels défis](#) », in La note d'analyse, Commissariat Général à la Stratégie et à la Prospective, Département Questions Sociales, N°8, Novembre 2013.

Anne Lauvergeon et al., « [Ambition 7 : La valorisation des données massives \(Big Data\)](#) », in « Un principe et sept ambitions pour l'innovation - Rapport de la commission Innovation 2030 », Octobre 2013.

C. Villani, « [Donner du sens à l'intelligence artificielle : pour une stratégie nationale et européenne](#) », 28 Mars 2018.

