

Introduction à la DATA SCIENCE

Du DATA MINING au BIG DATA

Enjeux et opportunités



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 ?

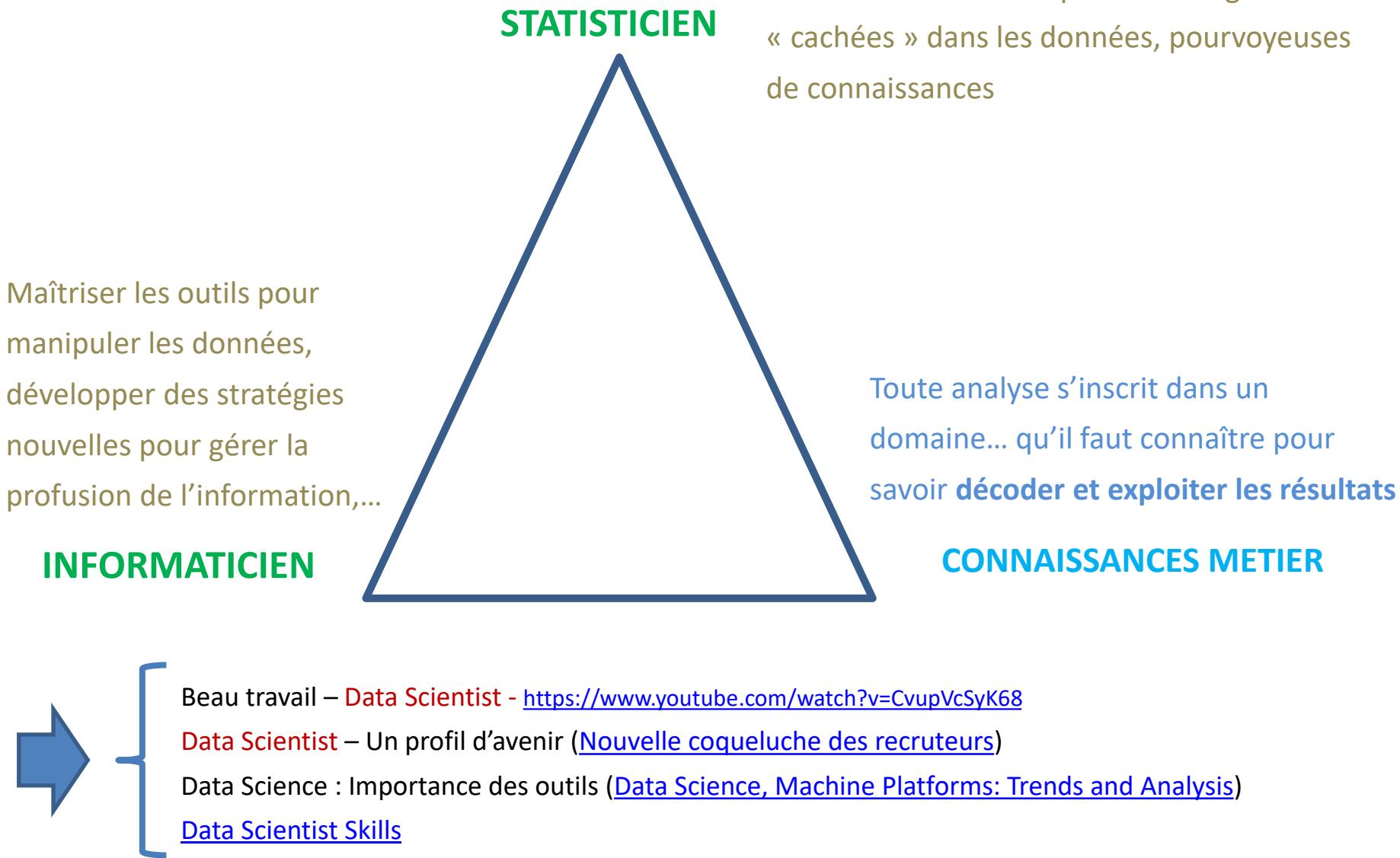
(La notion est très en vogue cf. [Google Trends](#))



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 – 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**

2 Prise de conscience collective... **surtout des entreprises**... de la valeur ajoutée que l’on peut en tirer

3 Indéniablement, il y a un effet de mode. Les éditeurs de solutions informatiques n’y sont pas étrangers.

Statistique /
Analyse de données



Data Mining



Data Science
Big Data Analytics

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

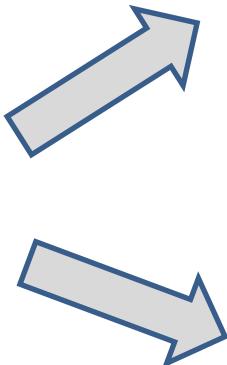


Statistique

Traitement statistique des données



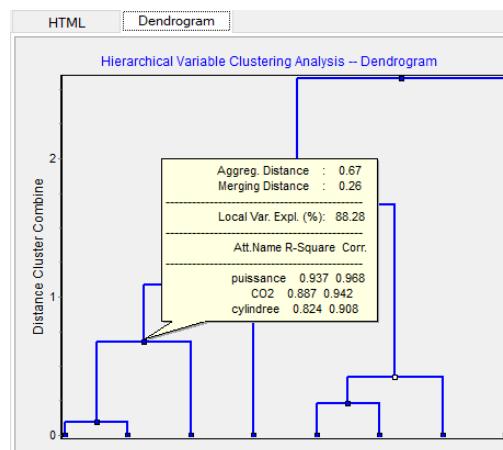
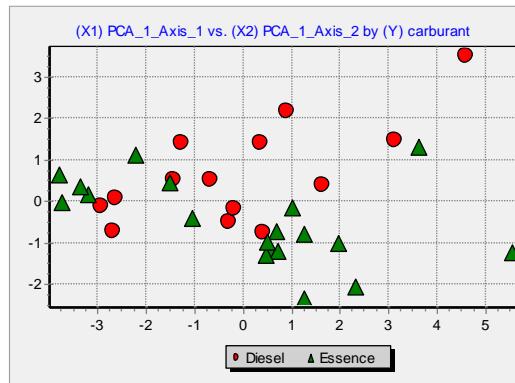
Application des techniques de modélisation et de statistique



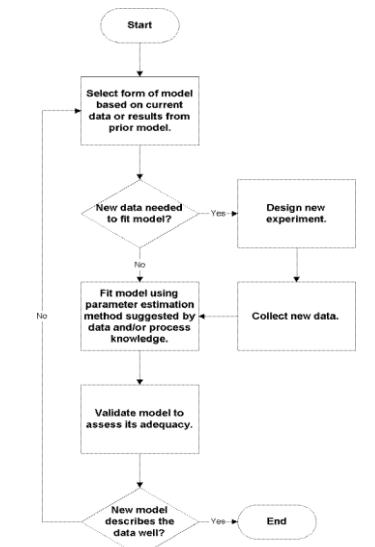
Les données sont spécifiquement recueillies à des fins d'étude (ex. enquête, expérimentations, etc.)

- Bonne qualité souvent
- Faible volumétrie

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

Expérience de l'entreprise : ses clients et leur comportement



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.



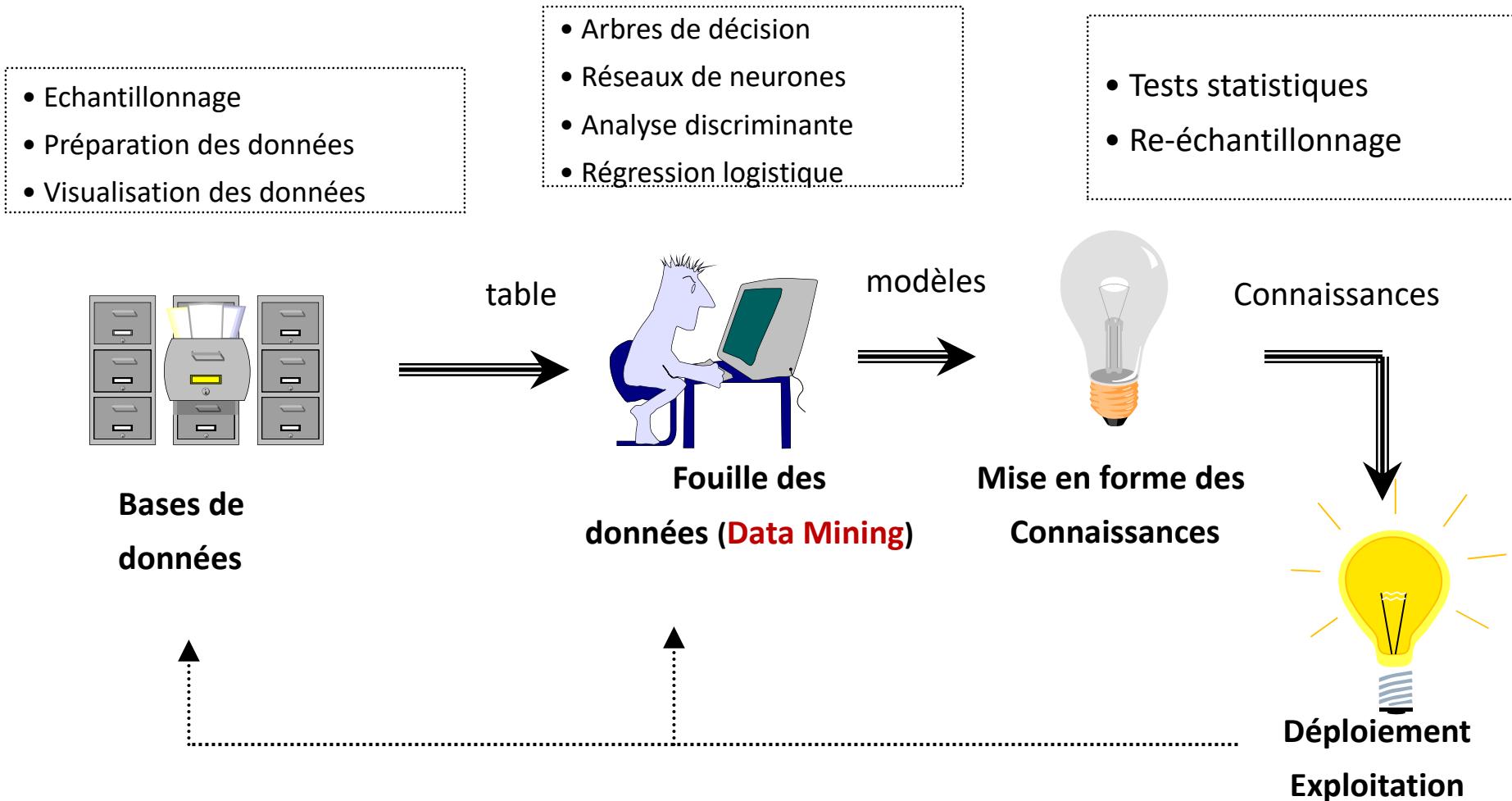
- coûteuse en stockage
- inexploitée pendant longtemps

Comment et à quelles fins utiliser cette expérience
accumulée



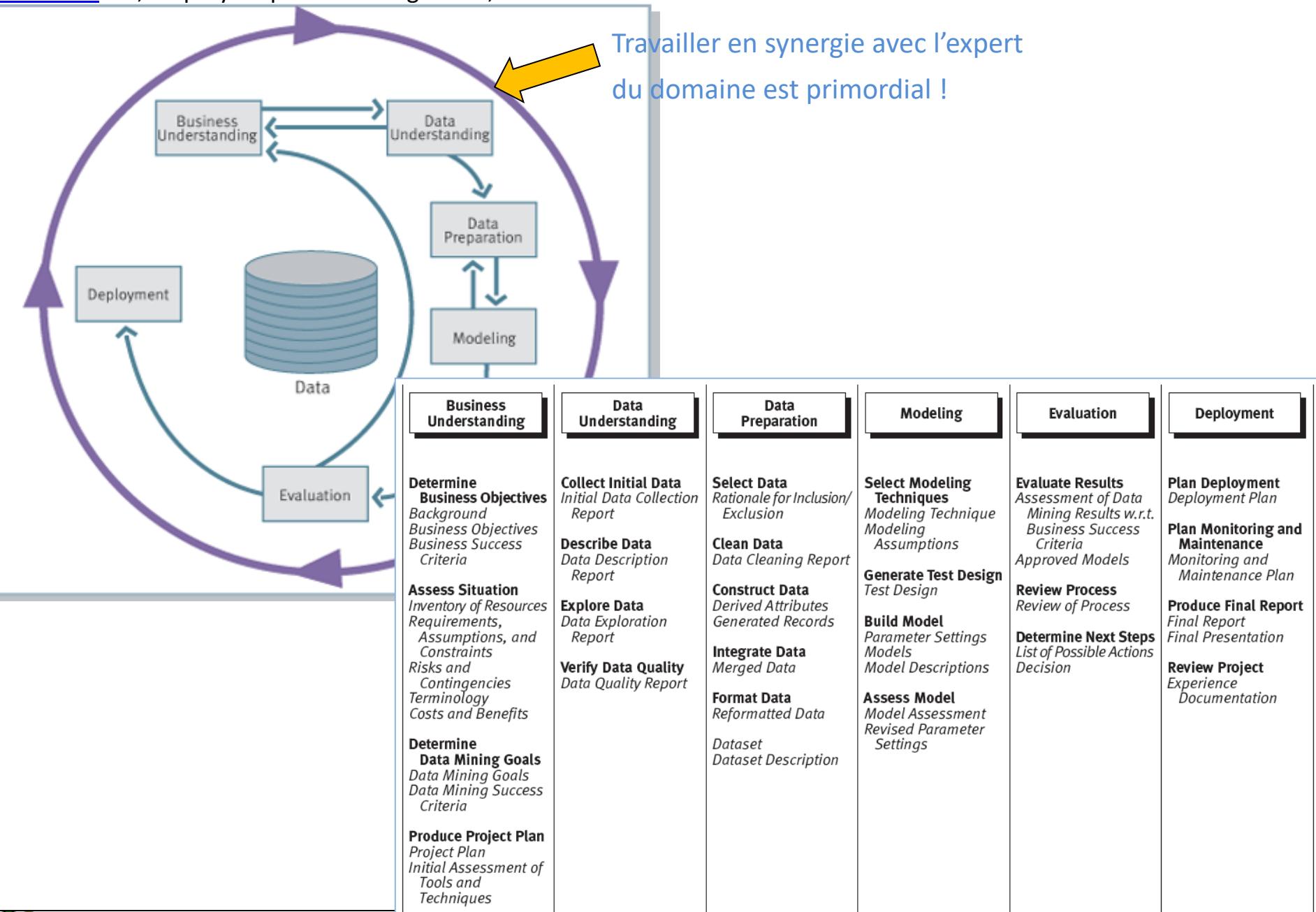
Le processus ECD (Extraction de connaissances à partir de données)

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)





Est-ce vraiment nouveau ?

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](#))



Spécificités du Data Mining ?

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



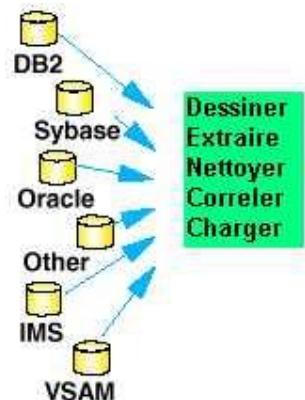
Spécif.1 - Les sources de données

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

Données Opérationnelles (Operational Data)

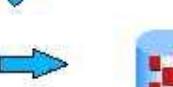


Entrepôt de Données (Data Warehouse)

Stockage

- orientation analyse
- historisées
- non-volatiles

(Data Mart)



Quelles seront les tendances salariales la prochaine année?



Comment réduire les coûts de 20% ?



Quel est le meilleur canal de distribution pour ces produits ?



(Data Mining)

Production

- orientation service (ventes, comptabilité, marketing...)
- volatiles



B.D. de gestion vs. B.D. décisionnelles

	Systèmes de gestion (opérationnel)	Systèmes décisionnels (analyse)
Objectif	dédié au métier et à la production ex: facturation, stock, personnel	dédié au management de l'entreprise (pilotage et prise de décision)
Volatilité (perennité)	données volatiles ex: le prix d'un produit évolue dans le temps	données historisées ex: garder la trace des évolutions des prix, introduction d'une information datée
Optimisation	pour les opérations associées ex: passage en caisse (lecture de code barre)	pour l'analyse et la récapitulation ex: quels produits achetés ensemble
Granularité des données	totale, on accède directement aux informations atomiques	agrégats, niveau de synthèse selon les besoins de l'analyse



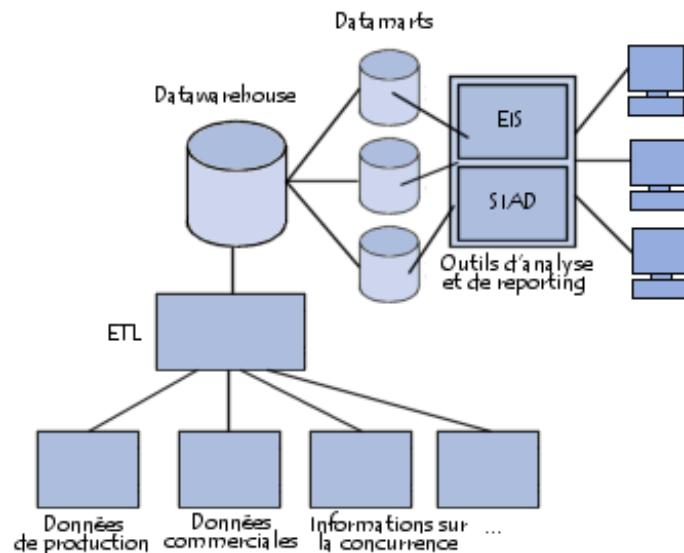
Entrepôts / Datamarts : Sources de données pour l'analyse

Conséquence : la volumétrie devient un élément important !!!

→ Découverte de connaissances à partir de données volumineuses

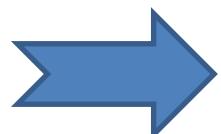
Data Mining vs. Informatique Décisionnelle (Business Intelligence)

Business intelligence (BI) is a set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business purposes. ... BI, in simple words, makes interpreting voluminous data friendly (http://en.wikipedia.org/wiki/Business_intelligence).



- Sélectionner les données (vs. un sujet et/ou une période)
- Trier, regrouper ou répartir ces données selon certains critères
- Élaborer des **calculs récapitulatifs « simples »** (proportions, moyennes conditionnelles, etc.)
- Présenter les résultats de manière synthétique (graphique et/ou tableaux de bord) → **REPORTING**

<http://www.commentcamarche.net/entreprise/business-intelligence.php3>



Le **Data Mining** introduit une dimension supplémentaire qui est la **modélisation « exploratoire »** (détection des liens de cause à effet, validation de leur reproductibilité)
➔ Un autre terme consacré est « **analytics** ».
(http://en.wikipedia.org/wiki/Business_analytics)



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 de 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.

Les méthodes sont le plus souvent complémentaires

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.)

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

Méthodes de « Machine Learning »

Association :

Trouver les ensembles de descripteurs qui sont le plus corrélés

- **règles d'association**

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

(Book, 2020) [Machine Learning from scratch](#)

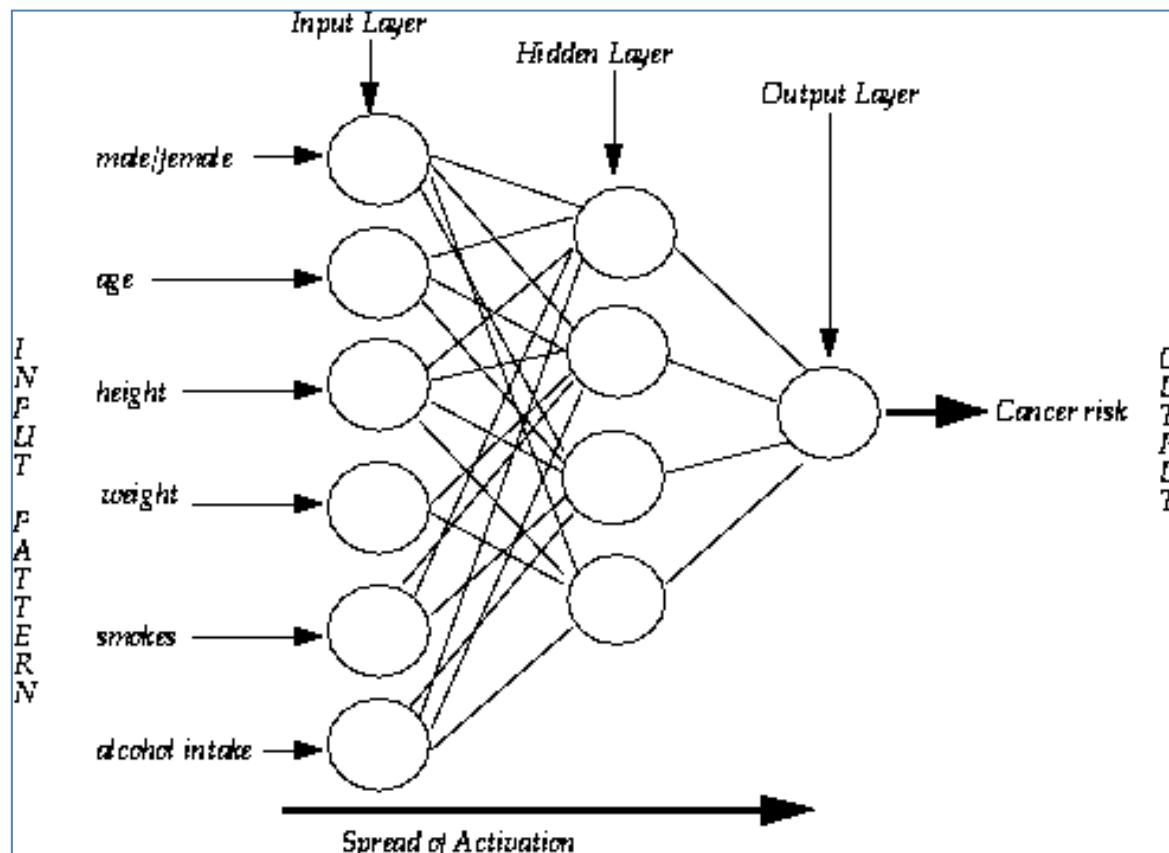
[Coursera Machine Learning - Stanford](#)

[Top Data Science and Machine Learning Used](#) (2018, 2019)



Techniques issues de l'Intelligence Artificielle

Les réseaux de neurones artificiels



- capacité d 'apprentissage (universalité)
- structuration / classement



Techniques en provenance des BD

Les règles d'association

Main Rule Type Data Format

Data Source: D:\WORKSIP\DATA\Loan\CreditMr.dbf

	Field Name	Field Type	Analyze if Empty	Ignore "if"	Ignore "then"
1	REASON	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	MARITAL_ST	Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
3	TITLE	Quality	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
4	SPOUSE_TIT	Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
5	GUARANTEE	Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
6	INSURANCE	Quality	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
7	HOUSING	Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
8	HOUSING_TY	Quality	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
9	JOB	Quality	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

If MARITAL_ST is Divorced

Then

SPOUSE_TIT is None

Rule's probability: 0.952

The rule exists in 40 records.

If MARITAL_ST is Divorced

and LOAN_LENGTH = 4.00

Then

GUARANTEE is No

Rule's probability: 0.966

The rule exists in 28 records.

A = B + 2.00

where: A = FAMILY_COUNT

B = CHILDREN

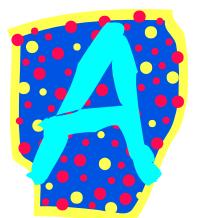
Accuracy level : 0.96

The rule exists in 397 records.

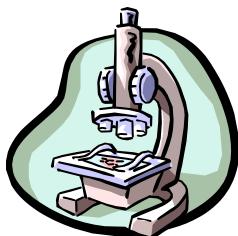
- traitement « omnibus »
- connaissance interprétable



Spécif.3 - Multiplicité des supports et des sources



Rôle fondamental de la préparation des données



	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 DATA MINING

- 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 le Data Mining ?



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, certains à des tarifs qui arrachent

[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>



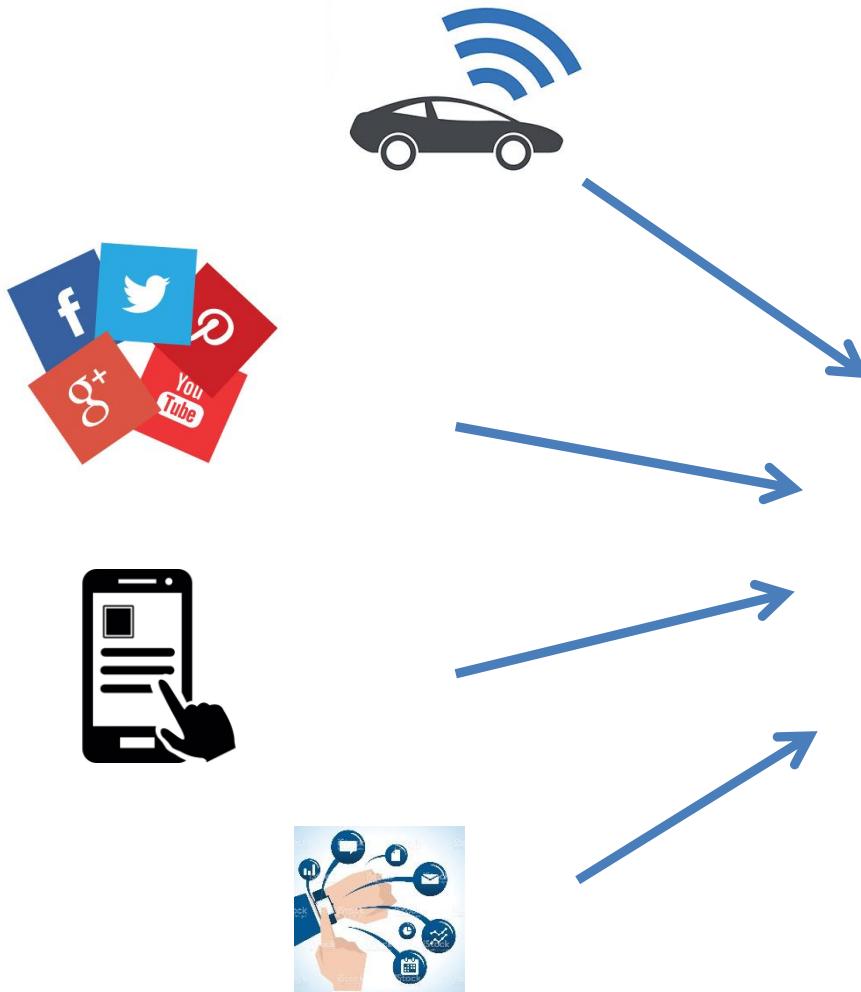
Spécificités du Big Data ?

Nouvelles caractéristiques des données :

Volume – Variété – Vélocité

Parce que...

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



Enjeux de stockage (technologique)
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

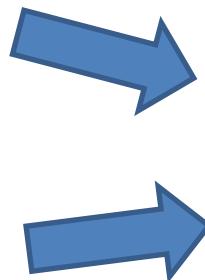
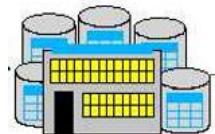
Chaque jour, nous générions 2,5 trillions d'octets de données. ... Ces données proviennent de partout : de capteurs utilisés pour collecter les informations climatiques, de messages sur les sites de médias sociaux, d'images numériques et de vidéos publiées en ligne, d'enregistrements transactionnels 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

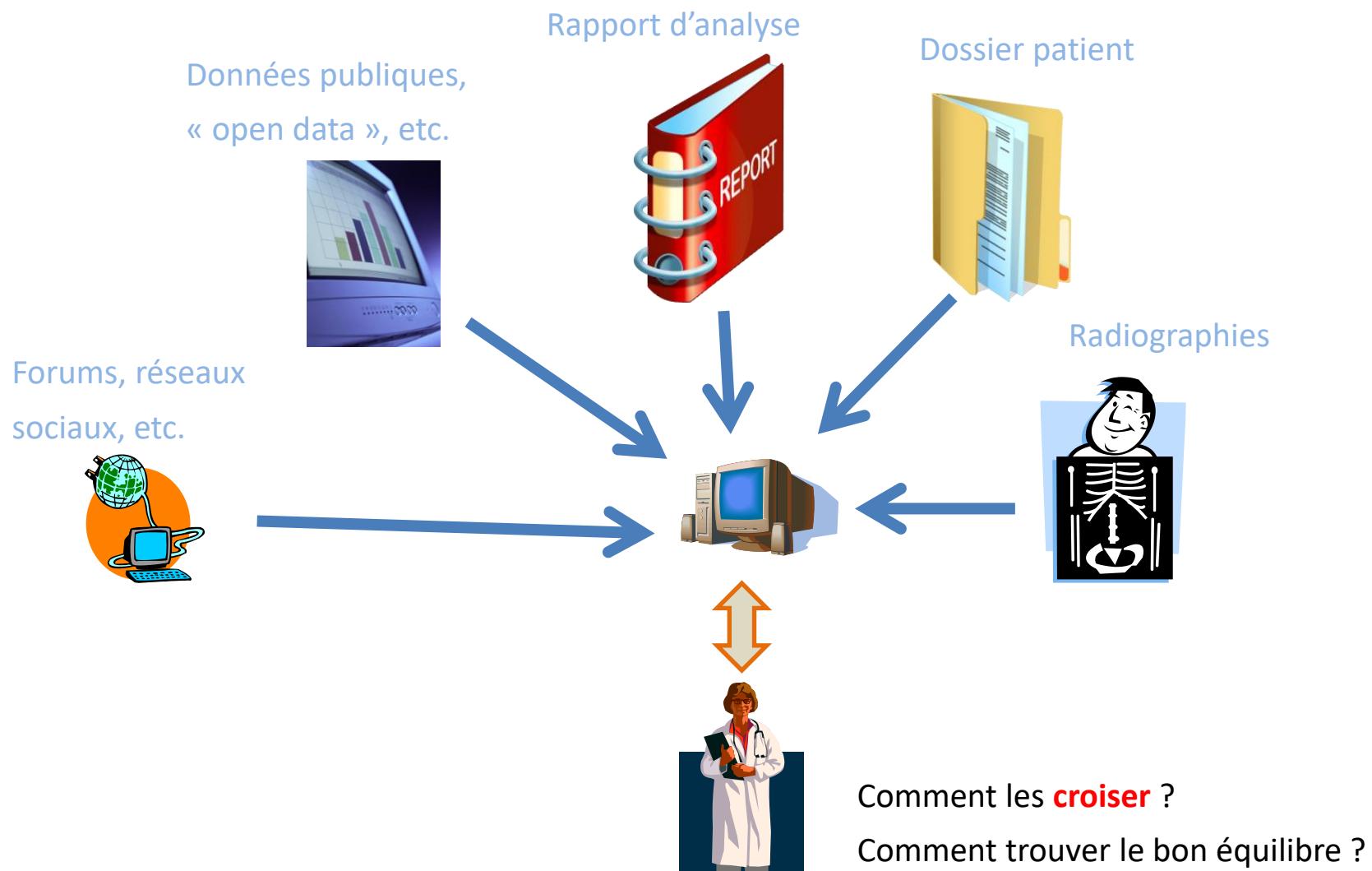
Pour rendre les analyses
plus performantes



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 (Amaguz, 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 product page for 'Gil Jourdan : L'Intégrale 1'. The main product image is a black and white illustration of Gil Jourdan sitting in a chair. The title is 'Gil Jourdan : L'Intégrale 1' by Maurice Tillieux. The price is listed as EUR 24,00. Below the main product, there is a section titled 'Produits fréquemment achetés ensemble' (Products frequently bought together) which shows 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'. A red 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)

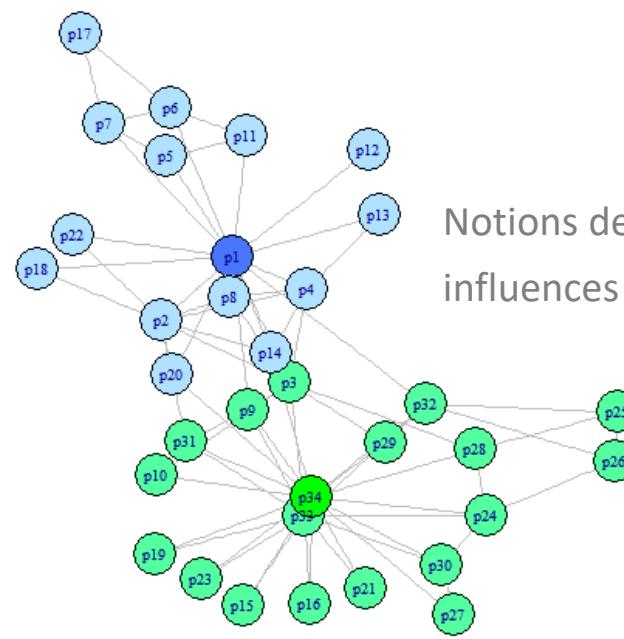
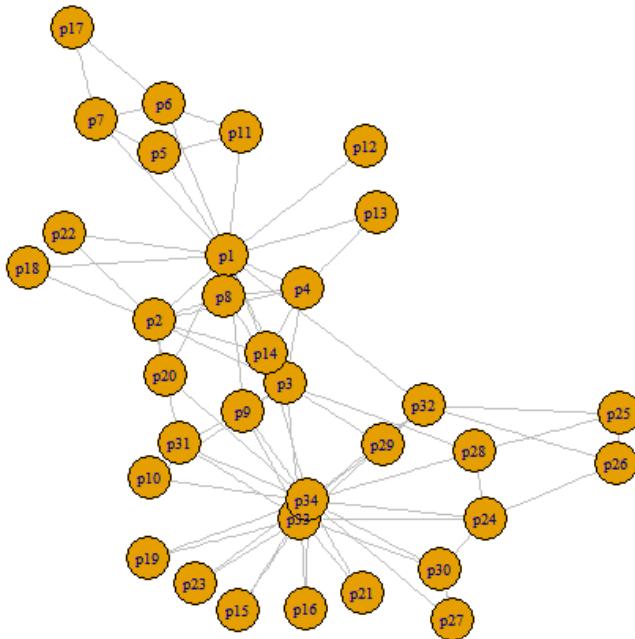
	A	
1	opinion	message
2	2	Gas by my house hit \$3.39!!!! I'm going to Chapel Hill on Sat. :)
3	0	Theo Walcott is still shit, watch Rafa and Johnny deal with him on Saturday.
4	0	its not that I'm a GSP fan, i just hate Nick Diaz. can't wait for february.
5	0	Iranian general says Israel's Iron Dome can't 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 text mining avec un cadre et des finalités particulières !!!

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

Détection de communautés dans les réseaux sociaux



Avec de nouveaux usages, parfois surprenantes (4)

Détection et reconnaissance des objets (la voiture grimée)



Reconnaissance faciale et détection de l'âge



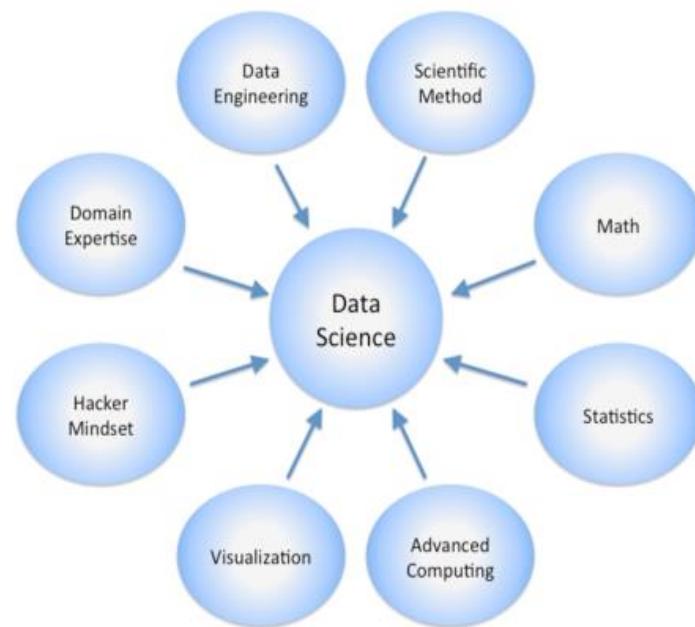
DATA SCIENCE

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.

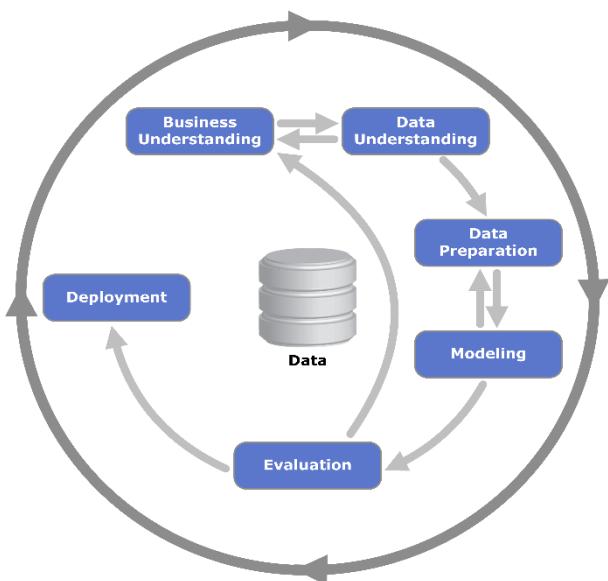


DATA MINING vs. DATA SCIENCE

Tout devient source de données

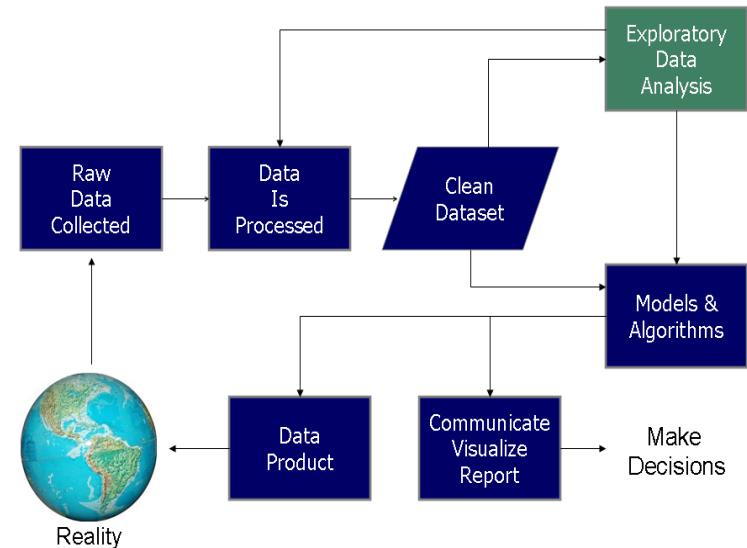


IBM CRISP DM



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 reads "BNP PARIBAS La banque d'un monde qui change". The main heading is "NOUS RECHERCHONS UN DATA SCIENTIST / AI LAB". Below it, 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 !)



Les logiciels de data science

“ Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis “ (May 2018)

(<https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html>)



Cahier des charges – Logiciel de Data Mining



Accès et préparation des données

Accéder à un fichier / une BD

Rassembler des sources différentes

Méthodes de Fouille de données

Lancer les calculs avec différents algorithmes
Bibliothèque de méthodes

Enchaîner les traitements

Faire coopérer les méthodes sans programmer

Évaluer les connaissances

Validation croisée, etc.

Exploiter les sorties

Rapports, visualisation interactive, etc.

Appliquer/exploiter les modèles

Modèles en XML (PMML), code C, DLL compilées
Prédiction directe sur de nouveaux fichiers



Piloté par menu, langage de commande + script, diagramme de traitements

Traitement local, traitement distribué

L'estampille Big Analytics Platforms – Quoi de plus ?



D. Hensen, « [16 Top Big Data Analytics Platforms](#) », InformationWeek, Janv 2014.

Besoin de plus de puissance, de plus de rapidité (ex. [analyse en mémoire revisitée](#), en 64 bits, environnement distribué)

Synergie encore plus forte avec les bases de données ([SQL Server Decision Tree](#), Oracle [Decision Tree](#), ...)

Architecture distribuée encore et toujours plus (à chacun sa solution autour de Hadoop...)



L'évolution porte sur les technologies, pas sur les méthodes analytiques



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

EXCEL (le tableau en général)

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

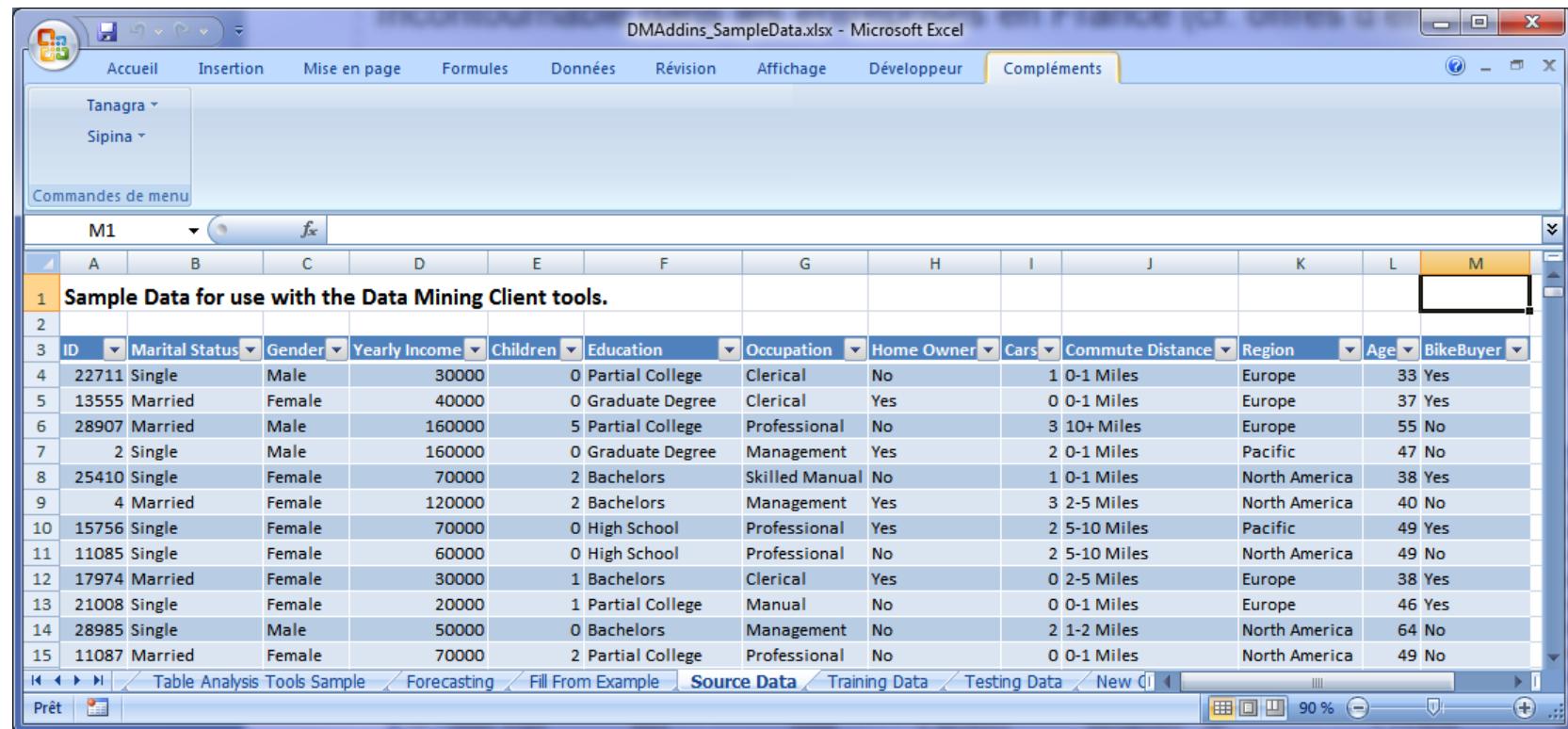
Fonctionnalités de manipulation et de préparation de données

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 bar with "Tanagra" and "Sipina" icons. The main area displays a data table with 15 rows 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 first row contains the header "Sample Data for use with the Data Mining Client tools.". 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.

1	Sample Data for use with the Data Mining Client tools.												
2	ID	Marital Status	Gender	Yearly Income	Children	Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age	BikeBuyer
3	22711	Single	Male	30000	0	Partial College	Clerical	No	1	0-1 Miles	Europe	33	Yes
4	13555	Married	Female	40000	0	Graduate Degree	Clerical	Yes	0	0-1 Miles	Europe	37	Yes
5	28907	Married	Male	160000	5	Partial College	Professional	No	3	10+ Miles	Europe	55	No
6	2	Single	Male	160000	0	Graduate Degree	Management	Yes	2	0-1 Miles	Pacific	47	No
7	25410	Single	Female	70000	2	Bachelors	Skilled Manual	No	1	0-1 Miles	North America	38	Yes
8	4	Married	Female	120000	2	Bachelors	Management	Yes	3	2-5 Miles	North America	40	No
9	15756	Single	Female	70000	0	High School	Professional	Yes	2	5-10 Miles	Pacific	49	Yes
10	11085	Single	Female	60000	0	High School	Professional	No	2	5-10 Miles	North America	49	No
11	17974	Married	Female	30000	1	Bachelors	Clerical	Yes	0	2-5 Miles	Europe	38	Yes
12	21008	Single	Female	20000	1	Partial College	Manual	No	0	0-1 Miles	Europe	46	Yes
13	28985	Single	Male	50000	0	Bachelors	Management	No	2	1-2 Miles	North America	64	No
14	11087	Married	Female	70000	2	Partial College	Professional	No	0	0-1 Miles	North America	49	No



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

sipina-arbres-de-decision.blogspot.fr

DiskStation

Accueil Versions Méthodes Capacités Références

Sipina - Arbres de décision - Data Mining

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

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

Homologues commerciaux

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

Add-ins pour tableurs

Tutoriel : diagnostic d'une maladie cardio-vasculaire

Solutions grandes bases

[Swap - Traitements sur disque](#)

[Multithreading](#)

[Echantillonnage](#)

[Formats de fichiers spécifiques](#)



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

Ricco Rakotomalala.

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)

Aujourd'hui, essentiellement un outil pédagogique.

The screenshot shows the Tanagra software interface. At the top, there's a banner with a palm tree icon and the word "TANAGRA". Below it, a navigation bar has icons for Présentation, Galerie, Caractéristiques, Didacticiels, Téléchargement, and Sipina. A green arrow points from the "Didacticiels" button to a separate window titled "Tutoriels Tanagra pour le Data Mining". This window contains a list of tutorials with PDF icons and a detailed description of the Data Mining Tutorials section. Another blue arrow points from the "Tutoriels" button to the same "Tutoriels Tanagra pour le Data Mining" window. The main window also shows a decision tree diagram and various menu options like "DOC. - TUTORIELS", "LOGICIELS", "REF. EXTERNES", and "DOC. LOGICIELS".

TANAGRA – Classement automatique de planctons (Image mining)

Image originelle fournie par le scanner

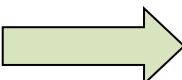
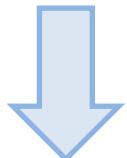


Image traitée en niveau de gris, à partir de laquelle sont calculés les paramètres

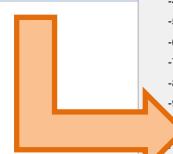


Avec l'outil
ImageJ

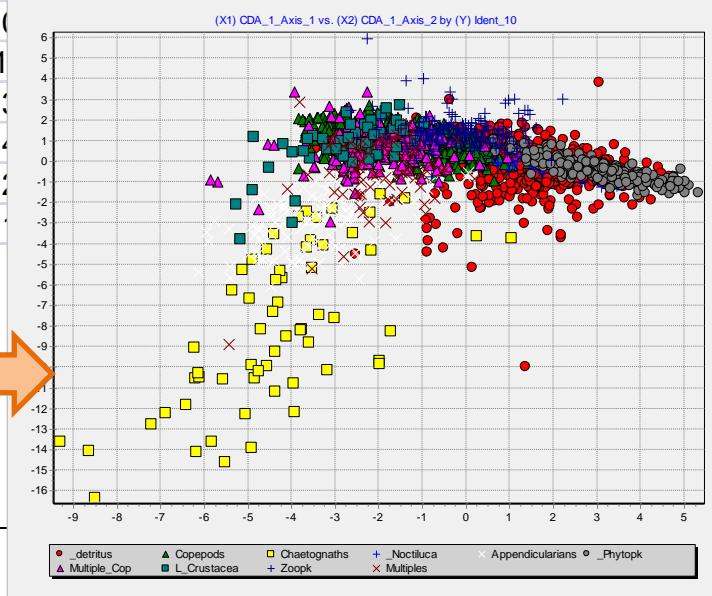


L'expert étiquette manuellement les objets

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



Ex. de traitement :
description factorielle



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 application running in the background.

R Project Website:

- Left Sidebar:** Links to About R, What is R?, Contributors, Screenshots, What's new?, Download, Packages, CRAN, R Project Foundation, Members & Donors, Mailing Lists, Bug Tracking, Developer Page, Conferences, Search, Documentation, Manuals, FAQs, The R Journal, Wiki, Books, Certification, Other, Misc, Bioconductor, Related Projects, User Groups, and Links.
- Content Area:** Features a PCA plot titled "PCA 5 vars" with axes Fertility, Examination, Education, Catholic, Agriculture, and a Factor 1 score of (1-3) 60%. Below it is a dendrogram titled "Clustering 4 groups". A "Getting Started" section provides an overview of R, and a "News" section lists recent releases.

RGui Application:

- File Menu:** File, Edit, View, Misc, Packages, Windows, Help.
- Toolbar:** Includes icons for file operations like Open, Save, Print, and Stop.
- R Console Window:** Displays the R version information and its license terms. It also shows the natural language support message, the collaborative nature of the project, and instructions for demos and help.

Text Overlay:

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](#)

*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_exce
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_)

```

Python file length: 5175 lines: 162 Ln: 57 Col: 1 Sel: 0|0 Windows (CR LF) UTF-8 INS

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

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 'professionnal'...

The screenshot shows the KNIME graphical user interface. On the left, there's a sidebar with links to 'PRODUCTS', 'APPLICATIONS', 'PARTNERS', 'SERVICES', 'RESOURCES', and 'COMPANY'. Below this is a section titled 'Why I KNIME...' with the tagline 'KNIME makes mining fun.' and a link to 'More information'. The main workspace displays a complex workflow diagram. At the top of the diagram, there are four 'Interactive Table' nodes. Below them, several 'XPath' nodes are connected to 'Ungroup' nodes. These are then connected to 'String Manipulation' nodes, which finally connect to another set of 'Interactive Table' nodes. A 'Node Repository' panel on the left lists categories like 'IO', 'Database', 'Mining', and 'Meta'. A 'Workflow Projects' panel on the right shows a list of projects including 'ACP', 'CAH', 'Linear Classifier', etc. A 'Download' button is visible at the top 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' (describing how to filter nodes based on XPath queries), and 'XPath Query' (allowing users to enter XPath queries). The bottom of the interface shows a 'Console' window displaying the welcome message for KNIME v2.9.1.0041089.

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 'professionnal'...

Why I KNIME... KNIME makes mining fun.

KNIME - Professional Open-Source Software

KNIME [naim] is a user-friendly graphical workbench for the entire analysis process: from initial investigation, powerful predictive analytics, visualisation and reporting. The open source KNIME platform provides over 1000 modules (nodes), including those of the [KNIME community](#) and its extensive partners.

KNIME can be [downloaded](#) onto the desktop and used free of charge. KNIME products include a wide range of features such as shared repositories, authentication, remote execution, scheduling, SOA integration, and more. KNIME is used by over 3000 organizations in more than 60 countries.

/ More information about KNIME.

Modeling Visualization Professional Segmentation Capabilities Reporting Modularity

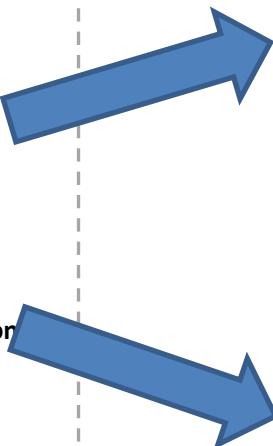
Ricco Rakotomalala
Tutoriels Tanagra - <http://tutoriels-data-mining>

KNIME – Catégorisation de nouvelles (Reuters)

```
<xml>
<document>
<sujet>acq</sujet>
<texte>
Resdel Industries Inc said
it has agreed to acquire San/Bar Corp in a share-for-share
exchange, after San/Bar distributes all shgares of its
Break-Free Corp subsidiary to San/Bar shareholders on a
share-for-share basis.

...
</texte>
</document>
<document>
<sujet>acq</sujet>
<texte>
Warburg, Pincus Capital Co L.P., an
investment partnership, said it told representatives of Symbion
Inc it would not increase the 3.50-dlr-per-share cash price it
has offered for the company.

...
</texte>
</document>
...
</xml>
```



Weka Node View - 0:61 - J48 (3.7)

File

Weka Output Graph Summary Source Additional Measures

```
J48 pruned tree
-----
oil <= 0: acq (52.0/1.0)
oil > 0
|   plc <= 0
|   |   pacif <= 0
|   |   |   cooper <= 0
|   |   |   |   buy <= 0
|   |   |   |   |   cash <= 0
|   |   |   |   |   |   agre <= 0: crude (43.0/1.0)
|   |   |   |   |   |   agre > 0: acq (3.0/1.0)
|   |   |   |   |   cash > 0: acq (4.0/1.0)
|   |   |   |   |   buy > 0: acq (3.0/1.0)
|   |   |   |   cooper > 0: acq (3.0)
|   |   |   pacif > 0: acq (3.0)
|   plc > 0: acq (6.0)

Number of Leaves : 8
Size of the tree : 15
```

Weka Node View - 0:65 - SMO (3.7)

File

Weka Output

```
Classifier for classes: acq, crude
BinarySMO

Machine linear: showing attribute weights, not support vectors

-0.0131 * abm
+ -0.041 * gold
+ 0.0299 * corp
+ -0.029 * proceed
+ 0.0051 * initi
+ -0.0202 * public
+ -0.0543 * offer
+ 0.0072 * seven
+ 0.208 * mln
+ -0.0728 * share
+ -0.0785 * stock
+ -0.1366 * dlers
+ 0.0026 * increas
+ 0.03 * canadian
```



RAPIDMINER

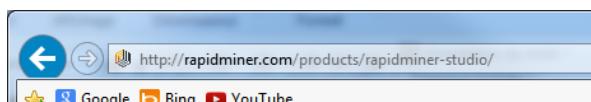
(<http://rapidminer.com/>)

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...



Screenshot of the RapidMiner Studio software interface:

The interface shows a graphical workflow editor with the following components:

- Operators View:** A tree view of available operators categorized by type:
 - Process Control (37)
 - Utility (54)
 - Repository Access (6)
 - Import (28)
 - Export (18)
 - Data Transformation (115)
 - Modeling (249)
 - Classification and Regression
 - Lazy Modeling (2)
 - Bayesian Modeling (1)
 - Tree Induction (8)
 - Rule Induction (5)
 - Neural Net Training (1)
 - Function Fitting (7)
 - Logistic Regression (1)
 - Support Vector Machine (1)
 - Discriminant Analysis (1)
 - Meta Modeling (14)
 - Weka (106)
 - Bayes (13)
 - Net (3)
 - Functions (16)
 - Lazy (5)
 - Mi (11)
 - Misc (3)
 - Rules (11)
 - Trees (15)
 - W-J48
 - W-ADTree
 - W-BFTree
 - W-Decisions
 - W-FT
 - W-Id3
 - W-J48graft
 - W-LADTree
 - W-LMT
 - W-M5P
 - W-NBTree
 - W-REPTree
 - W-RandomF
- Process View:** Displays the "Main Process" which includes:
 - A "Read Excel" operator with "fil" input and "out" output.
 - A "Process Document" operator with "wor" input and "exa" output.
 - A "W-J48" operator with "tra" input and "mod" output.
 - Connections between "out" and "wor", "exa" and "mod", and "mod" and "res".
- Parameters View:** Shows configuration for the process:
 - verbosity: init
 - logfile: (empty)
 - resultfile: (empty)
 - random seed: 2001
 - send mail: never
 - encoding: SYSTEM
- Problems View:** Shows "No problems found".
- Log View:** Shows "The root operator".
- Comment View:** Shows "Process Synopsis".

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

Tutoriel Tanagra - <http://tutoriels-data-mining.blogspot.fr/>

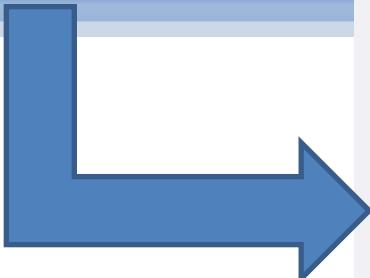
http://fr.wikipedia.org/wiki/Structure_des_protéines

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



protéines.xls [Mode de compatibilité]

A	famille	description
1	F1	SQFRVSPLRTWNLGETVELKCQVLLSNPTSGCSWLFQPRGAAASPTFLLYSQNPKAAEGLDTQRFSKRLGDTFVLTLSDFRRENEGYFCALSNSIMYFSHFVPVFLPA
2	F2	AVSKVYARSVYDSRGNPTEVELTTEKGVFVRSIVPSGASTGVHEALEMRDGDKSKWMGKGVLHAVKNVNDVIAPAFVKANIDVKDQKAVDDFLISLDGTANKSKLGANAILGVSLAA
3	F1	EPKFTKCRSPERETFSCHWTDEVHHGPIQLFYTRRNTEWTQEWEKCPDYVSAGENSCYFNSSFTSIWIPYCILTSNGGTDEKCFSVDEIVQP
4	F1	LQQPTIQSFEQVGTKVNVTVEDERTLVRRNNNTFLSLRDVGFKDLIYTLYWKSSGKKTAKTNTEFLIDVDKGENYCFSVQAIPSRTVNRKSTDSPVECMG
5	F1	SRCTHLENRDFVTGTQGTTRTLVLGGCVTITAEGKPSMDVWLDAIYQENKIVYTVKVEPHTGDYVAANETHSGRKTAFTISSEKTILTMGEYGDVSLLCRVASGPVAHIEGTKYHLKS
6	F1	GSDWVIPPINLPENSRGPFQELVRIRSGRDKNLSLRYSVTGPGADQPPGIFIIINPISGQLSVTKPLDRELIARFHRLRAHAVDINGNQVENPIDIVINVIMNDNRPEF
7	F1	ISGMGSGRKASGSPTSPINANKVENEDAFLLEEVAEEKPHVKPVFTKTILDMDVVEGSAARFDCKVEGYPDPEVMWFKDNPVKESRHFQIDYDEEGNCSLTISEVCGDDDAKYTCKAV
8	F2	AVSKVYARSVYDSRGNPTEVELTTEKGVFVRSIVPSGASTGVHE
9	F2	MKIDAIIEAVIVDVPTKRPPIQMSITTVHQQSIVRVYSEGLGV
10	F2	MERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGA
11	F2	APAPVKQGPSTVAYVEVNNNSMLNVGKYTLADGGGNNAFDVA
12	F2	SKIFDFVKPGVITGDDVQKVFQVAKENNfalPAVNCVGTDSIN
13	F2	MNSNLRGVMAALLTPFDQQQALDKASLRLVQFNIIQQGIDGL
14	F2	VQPTPADHFTFGLWTVGWTGADPFGVATRANLDPVEAVHKL
15	F2	KKTKVWCTGPKTESEEMIAKMLDACMNNVMPNNECHGPKYAEK



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)
```

Number of Leaves : 10



Autres outils

ORANGE



The screenshot shows the Orange Data Mining software interface. On the left, there's a canvas with nodes like 'File', 'Examples', 'CN2', 'Distributions', and 'Attribute Statistics'. A 'CN2 Rules Viewer' window is open, displaying a list of rules with columns for Length, Quality, Coverage, Class, Distribution, and Rule. One rule is highlighted: 'IF age <= 30.0 AND sex = female THEN survived=yes'. The right side of the interface has a sidebar with 'Features', 'Download', 'Add-ons', 'Documentation', 'Development', 'Forum', and 'Blog'. Below the sidebar, a text block reads: '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.' There's also a download section for 'Orange 2.7 for Windows' (Built May 13, 2014 at 13:43 CEST) and links for 'Downloads for other systems and versions' and 'Latest Blog Entries'.

WEKA / PENTaho



The screenshot shows the Weka Data Mining software website. At the top, there's a navigation bar with links for 'Community', 'Marketplace', 'Projects', and 'Try Pentaho EE'. The main content area features a heading 'Data Mining - Weka' with the subtext 'Comprehensive set of tools for machine learning and data mining to enhance your insights through predictive analytics.' Below this is a 'Downloads' button with a download icon. To the right, there's a graphic of a computer monitor displaying green data points. At the bottom, there's a 'Description' tab, a 'Main concepts' link, and a 'Contribute' link. A section titled 'Explore and understand your data' with the subtext 'Mining your own data and turning what you know about your users, your clients and your business into useful information it's now an easy task. With Weka, an Open Source Software, you can discover patterns in large data sets and extract all the information. It also brings great portability, since it was fully implemented in the JAVA programming language, plus supporting several standard data mining tasks.' On the right side, there's a 'QUICK LINKS' box containing links to 'Community Documentation', 'FAQ', 'Mailing List', 'Data Sets', and 'Screenshots'.



Ricco Rakotomalala

Tutoriels Tanagra - <http://tutoriels-data-mining.blogspot.com>

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](#)... l'avenir...



Quelques exemples

- (1) Ciblage de clientèle : le scoring
- (2) Étiquetage automatique de « nouvelles »



Ciblage de clientèle par publipostage (1/2)

Banque française

Objectif : Augmenter l'adhésion à un service en ligne (taux d'abonnement actuel 4%)

Base marketing : plusieurs centaines de milliers de clients,
~200 variables (95% sont quantitatives)

Méthode : isoler des groupes d'individus se ressemblant dans lequel
le taux d'abonnement est élevé

- les non-abonnés dans ces groupes seront (certainement ?) sensibles à une offre ciblée
(hypothèse : s'ils ne sont pas abonnés, c'est qu'ils n'ont pas reçu l'information)
- technique : arbre de décision avec échantillonnage équilibré sur chaque noeud

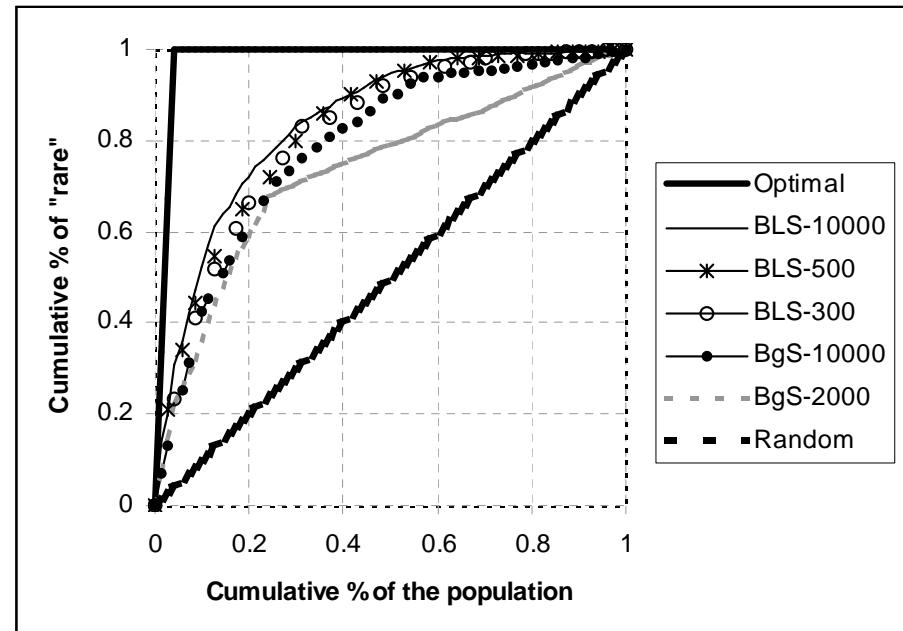


Ciblage de clientèle par publipostage (2/2)

Évaluation : dépasser le taux (coût) d'erreur, mesurer la qualité du ciblage

➤ meilleur ciblage : toutes les personnes contactées ont souscrit un contrat

Individu	Probabilité de souscrire	Pourc. Ind. cumul	Pourc. Ciblés Cumul	Pourc. Ciblés:
4	0.95	10%	19%	0.19
9	0.9	20%	37%	0.18
10	0.8	30%	53%	0.16
6	0.65	40%	66%	0.13
3	0.6	50%	78%	0.12
7	0.5	60%	88%	0.1
2	0.35	70%	95%	0.07
5	0.25	80%	100%	0.05
8	0	90%	100%	0
1	0	100%	100%	0
5.00				



Text Mining – Catégorisation de nouvelles (1/3)

The screenshot illustrates a text mining workflow. At the bottom, a 'REUTERS' interface displays a database table with columns: IDTEXTE, TEXTE, ACQ, CORN, CRUDE, and TRADE. The table lists various news items, with row 237 highlighted. An arrow points from this row to a detailed news article window above it. Another arrow points from the 'TRADE' column of row 237 to a second news article window at the bottom right.

IDTEXTE	TEXTE	ACQ	CORN	CRUDE	TRADE
45	(MEMO)	OUI	NON	NON	NON
97	(MEMO)	NON	OUI	NON	NON
110	(MEMO)	OUI	NON	NON	NON
144	(MEMO)	NON	NON	OUI	NON
235	(MEMO)	NON	OUI	NON	NON
236	(MEMO)	NON	NON	OUI	NON
237	(MEMO)	NON	NON	OUI	
246	(MEMO)	NON	NON	OUI	
248	(MEMO)	NON	NON	OUI	
273	(MEMO)	NON	NON	OUI	
302	(MEMO)	OUI	NON	NON	
342	(MEMO)	NON	NON	NON	

ASCS TERMINAL MARKET VALUES FOR PIK GRAIN
KANSAS CITY, Feb 26 - The Agricultural Stabilization and Conservation Service (ASCS) has established these unit values for commodities offered from government stocks through redemption of Commodity Credit Corporation commodity certificates, effective through the next business day.
Price per bushel is in U.S. dollars. Sorghum is priced per CWT, com yellow grade only.

	WHEAT	HRW	HRS	SRW	SWW	DURUM
Chicago	-	3.04	2.98	-	-	
III. Track	-	-	3.16	-	-	
Toledo	-	3.04	2.98	2.90	-	

INDONESIA SEEN AT CROSSROADS OVER ECONOMIC CHANGE
By Jeremy Clift, Reuters
JAKARTA, March 1 - Indonesia appears to be nearing a political crossroads over measures to deregulate its protected economy, the U.S. Embassy says in a new report.
To counter falling oil revenues, the government has launched a series of measures over the past nine months to boost exports outside the oil sector and attract new investment.
Indonesia, the only Asian member of OPEC and a leading primary commodity producer, has been severely hit by last



Text Mining – Catégorisation de nouvelles (2/3)

Codage de texte en tableau de données

*Les chercheurs qui cherchent, on en trouve
Mais les chercheurs qui trouvent, on en cherche*

Mots clés

- lemmatisation
- stopwords

Phrase	Les	Chercheurs	Qui	Cherchent	On	En	Trouve	Mais	Trouvent	Cherche
1	1		1	1	1	1	1	1	0	0
2	1		1	1	0	1	1	0	1	1

3-grams

- corresp. avec les mots
- problème du sens

Phrase	Les	es	s	c	ch	che	her	rch	eur
1	1		1	1	2	4	2	2	1
2	1		1	1	1	4	2	2	1



Text Mining – Catégorisation de nouvelles (3/3)

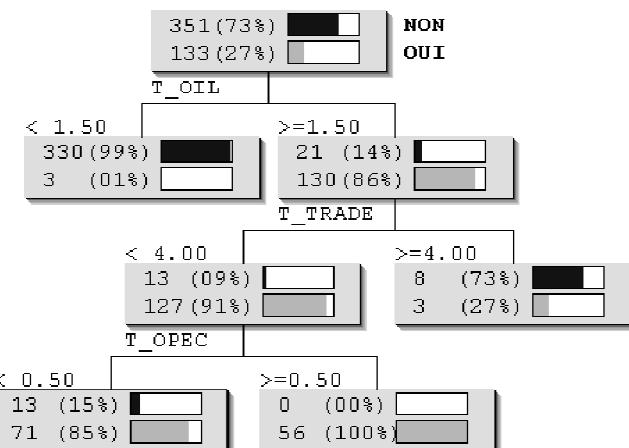
Visualiser les données

	IDTEXTE	TEXTE	ACQ	COPN	CRUDE	TRADE	T_OIL	T_CRUDE	T_BARRET	T_BARRET	T_OPEC	T_BPD	T_PETRO	T_PRICES	T_ENERG	T_GAS	T_EXPL
1	45	{MEMO}... OUI	NON	NON	NON	NON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	97	{MEMO}... NON	OUI	NON	NON	NON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	110	{MEMO}... OUI	NON	NON	NON	NON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	144	{MEMO}... NON	NON	OUI	NON	NON	12.00	0.00	0.00	0.00	16.00	4.00	1.00	6.00	2.00	0.00	0.00
5	235	{MEMO}... NON	OUI	NON	NON	NON	5.00	1.00	0.00	0.00	0.00	0.00	0.00	2.00	0.00	0.00	0.00
6	236	{MEMO}... NON	NON	OUI	NON	NON	7.00	2.00	1.00	3.00	9.00	7.00	0.00	5.00	1.00	0.00	0.00
7	237	{MEMO}... NON	NON	OUI	NON	NON	4.00	0.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00	0.00	0.00
8	246	{MEMO}... NON	NON	OUI	NON	NON	5.00	0.00	1.00	0.00	2.00	0.00	1.00	1.00	0.00	0.00	0.00
9	248	{MEMO}... NON	NON	OUI	NON	NON	9.00	0.00	1.00	2.00	7.00	2.00	0.00	9.00	0.00	0.00	0.00
10	273	{MEMO}... NON	NON	OUI	NON	NON	5.00	6.00	1.00	2.00	5.00	9.00	1.00	5.00	0.00	0.00	0.00
11	302	{MEMO}... OUI	NON	NON	NON	NON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	342	{MEMO}... NON	NON	NON	OUI	NON	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

56 Attributs

484 Individus

Lecture seule



Exemple : appartenance au sujet « crude »
(pétrole brut)



Ricco Rakotomalala

Tutoriels Tanagra - <http://tutoriels-data-mining.blogspot.fr/>

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.

