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 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…

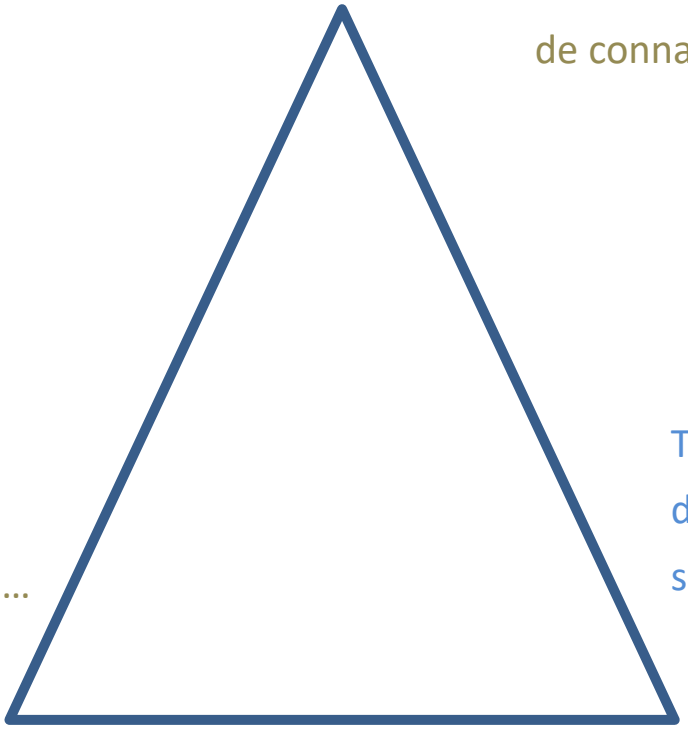
(**Double compétence : statistique et informatique**) (Wikipédia).

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 – A la croisée de 3 profils

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

STATISTICIEN



Maîtriser les outils pour manipuler les données, développer des stratégies nouvelles pour gérer la profusion de l'information,…

INFORMATICIEN

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

CONNAISSANCES METIER

- Beau travail – Data Scientist - https://www.youtube.com/watch?v=CvupVcSyK68
- Data Scientist – Un profil d'avenir (Nouvelle coqueluche des recruteurs)
- Data Science : Importance des outils (Platforms Trends [2019], State of Data Science [2021])
- Data Scientist Skills – Cf. Offre d'emploi 2018 (Compétences du Data Scientist)

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

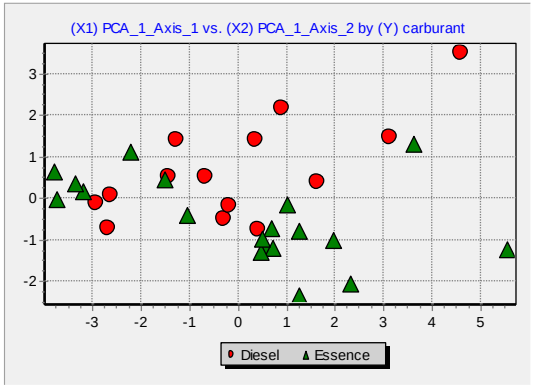
Statistique – Statistique exploratoire

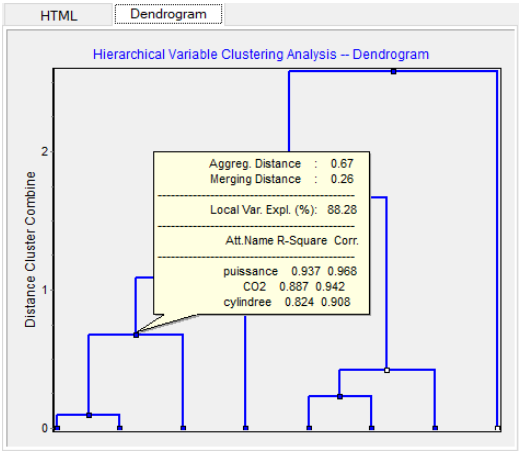
Les données sont spécifiquement recueillies à des fins d'étude (ex. enquête, expérimentations, etc.)

- Bonne qualité souvent
- Faible volumétrie

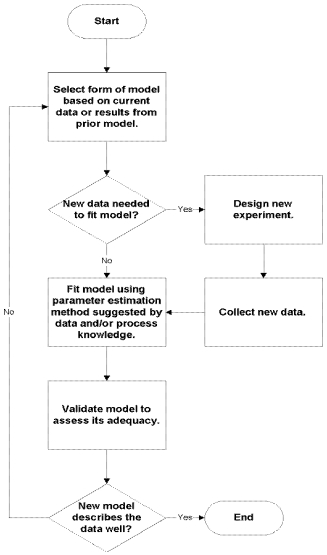
Volume de traitements – de toute manière – limité par les capacités des outils informatiques disponibles (à l'époque).

Application des techniques de modélisation et de statistique





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

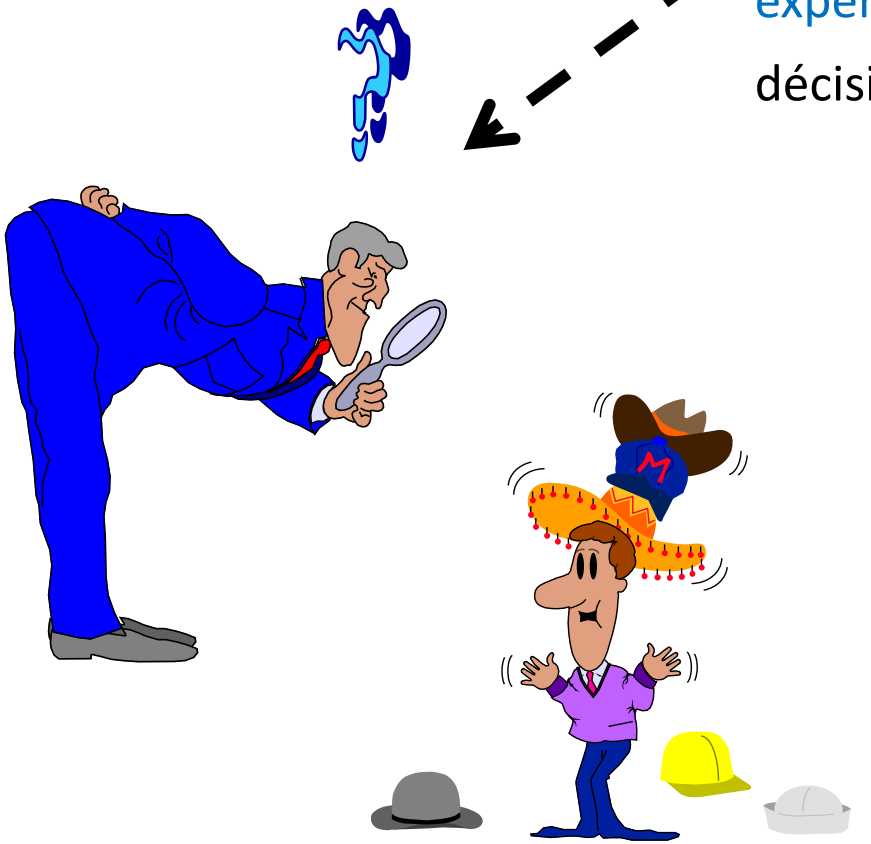


DATA MINING

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/)

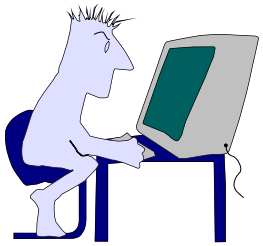
- Echantillonnage
- Préparation des données
- Visualisation des données

- Arbres de décision
- Réseaux de neurones
- Analyse discriminante
- Régression logistique

- Tests statistiques
- Re-échantillonnage

Bases de données

table

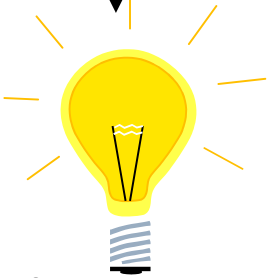


Fouille des données (Data Mining)

modèles

Mise en forme des Connaissances

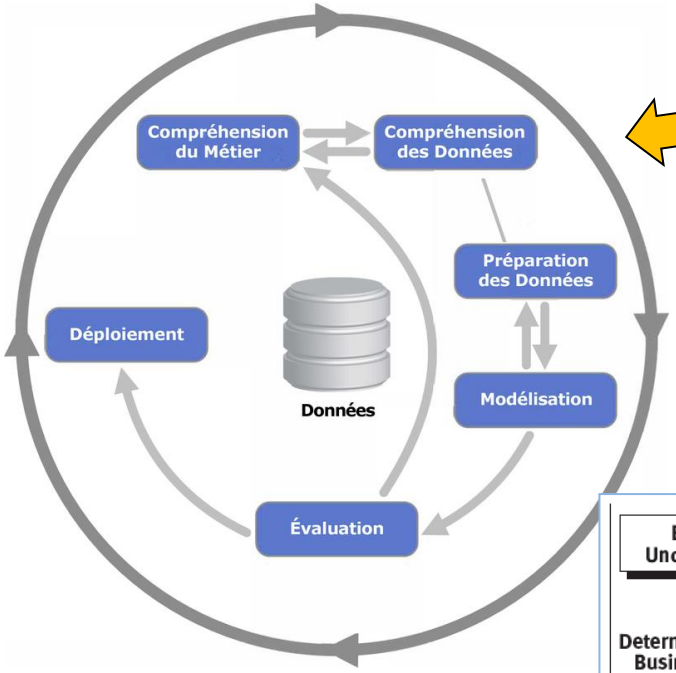
Connaissances



Déploiement

Exploitation

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

Remarque : De nouveau paradigmes arrivent aujourd'hui avec MLOps par ex., une recommandation de bonnes pratiques pour développer et maintenir efficacement un projet machine learning, inspiré de DevOps.

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 ?

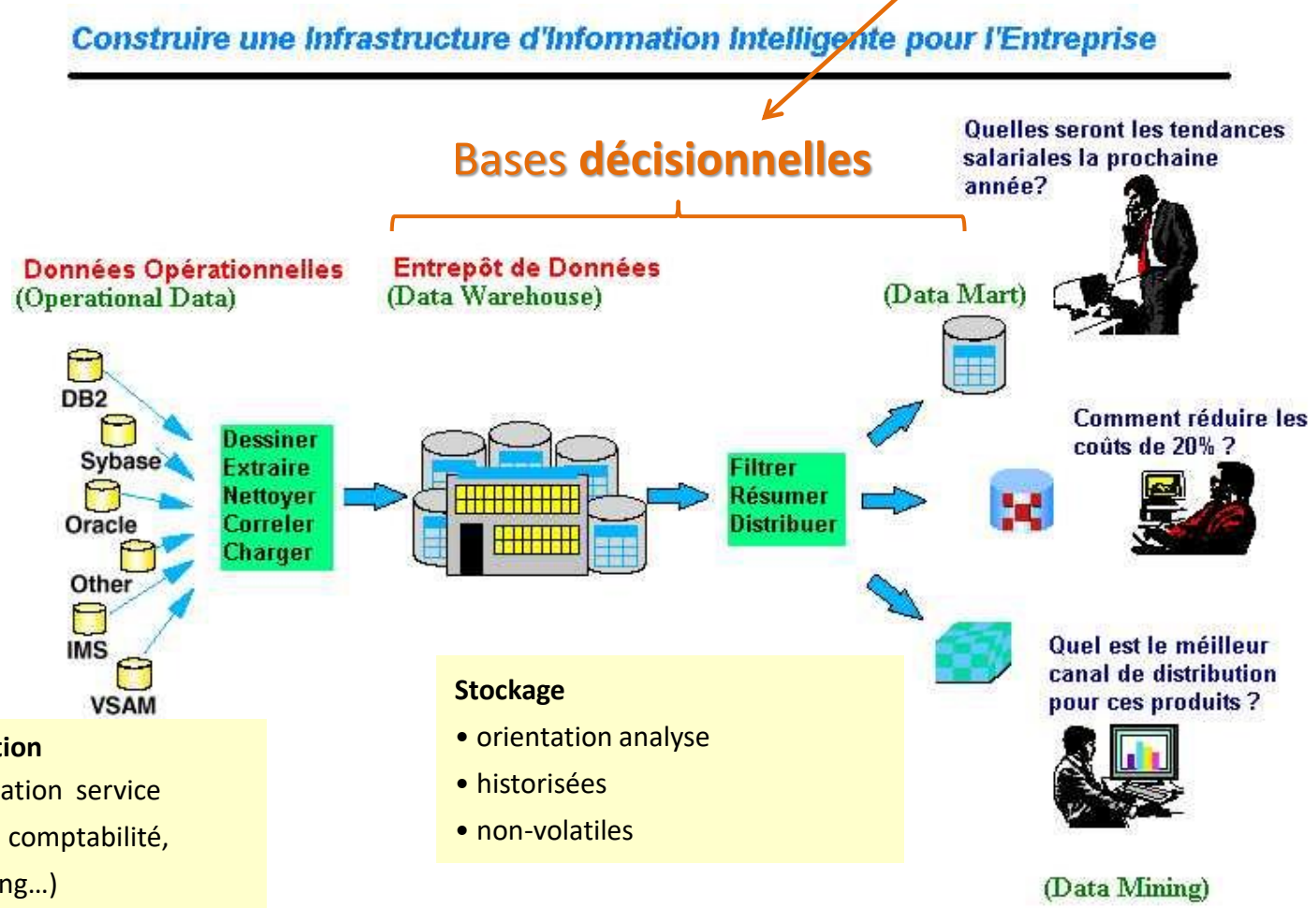
(1) Sources de données

(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.



Production

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

Stockage

- orientation analyse
- historisées
- non-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 proiduit évolue dans le temps	données historisées ex: garder la trace des évolutions des prix, introduction d'une information daté
Optimisation	pour les opérations associées ex: passage en caisse (lecture de code barre)	pour l'analyse et la récapitulation ex: quels les produits achetés ensembles
Granularité des données	totale, on accède directement aux informations atomiques	agrégats, niveau de synthèse selon les besoins de l'analyse

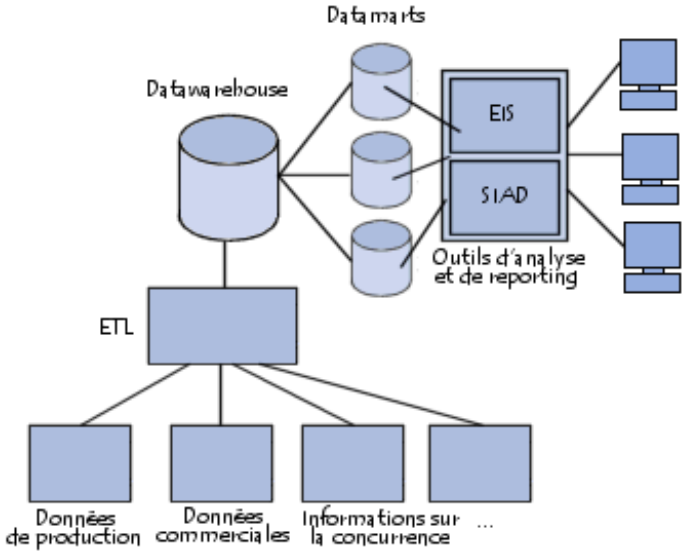
Entrepôts / Datamarts : Sources de données pour l'analyse

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).



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

- 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**

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)

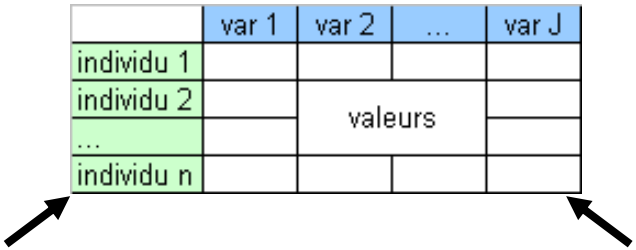
Description factorielle

Discrimination

Clustering

Méthodes géométriques, probabilités

ACP, ACM, Analyse discriminante, CAH, …



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 :

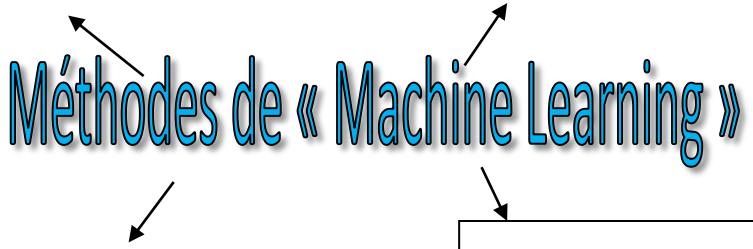
Faire ressurgir des groupes « naturels »

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



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

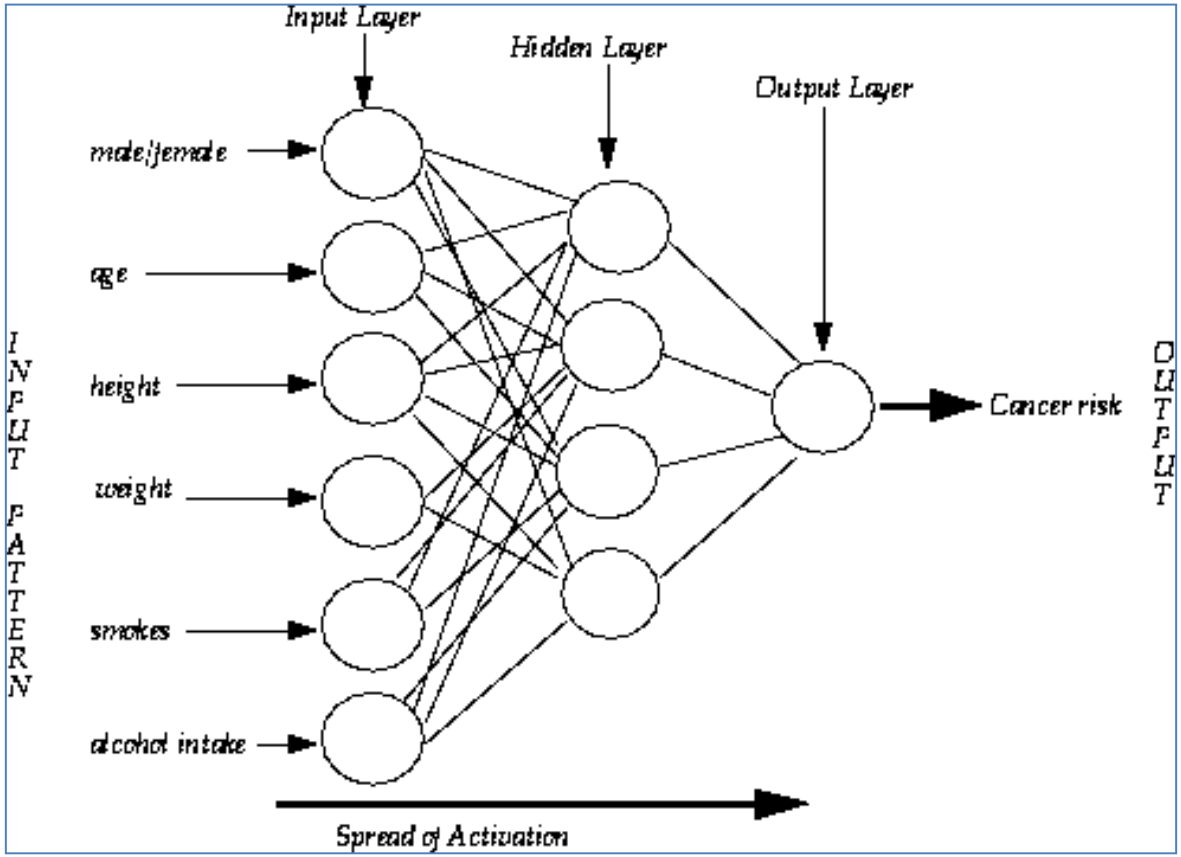
(Book, 2020) Machine Learning from scratch

Coursera Machine Learning - Stanford

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

Techniques issues de l'Intelligence Artificielle

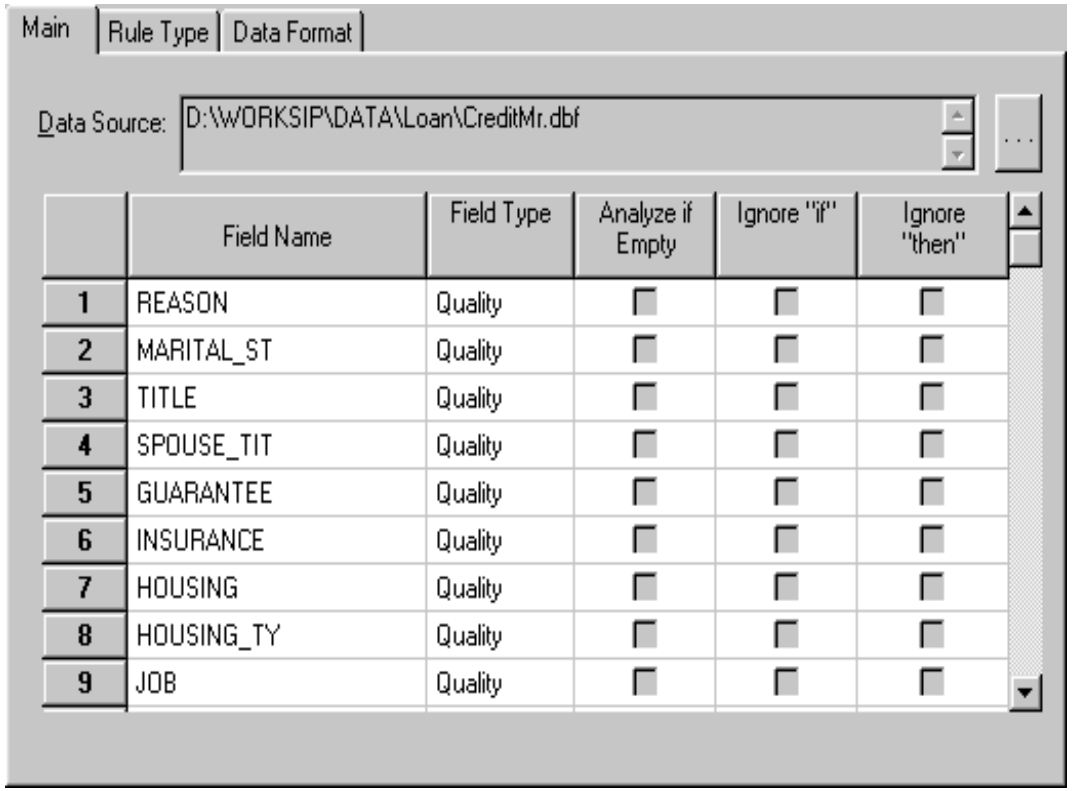
Les réseaux de neurones artificiels



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

Techniques en provenance des BD

Les règles d'association



	Field Name	Field Type	Analyze if Empty	Ignore "if"	Ignore "then"
1	REASON	Quality	☐	☐	☐
2	MARITAL_ST	Quality	☐	☐	☐
3	TITLE	Quality	☐	☐	☐
4	SPOUSE_TIT	Quality	☐	☐	☐
5	GUARANTEE	Quality	☐	☐	☐
6	INSURANCE	Quality	☐	☐	☐
7	HOUSING	Quality	☐	☐	☐
8	HOUSING_TY	Quality	☐	☐	☐
9	JOB	Quality	☐	☐	☐

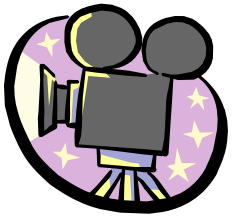
If **MARITAL_ST** *is* **Divorced**
Then
SPOUSE_TIT *is* **None**
Rule's probability: **0.952**
The rule exists in **40** *records.*

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.*

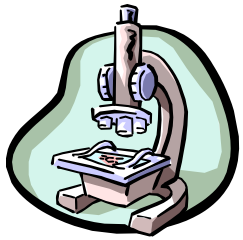
A = **B + 2.00**
where: **A** = **FAMILY_COUNT**
B = **CHILDREN**
Accuracy level : **0.96**
The rule exists in **397** *records.*

- 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

Les applications

Text mining (SAS Text miner, Sentiment analysis, ...)

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

Mais déjà ancien finalement... (JADT, Zooplancton, etc...)

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

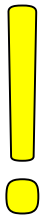
Démarche data mining

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

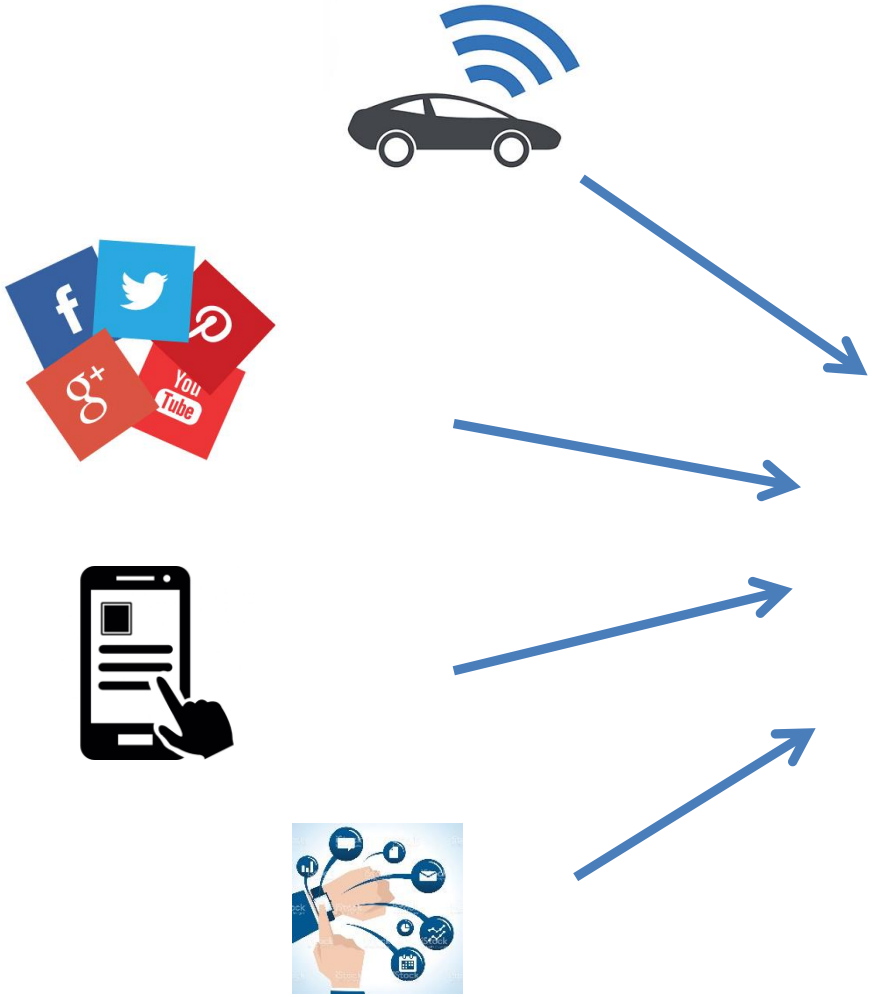
Parce que…

(1) Nouvelles sources de données, nouveaux contenus ;

(2) Y compris les sources externes à l'entreprise.

Big Data – Règne des objets connectés

(Internet des objets – IoT)



(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 (Wikipédia)

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**, **vélocité** 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

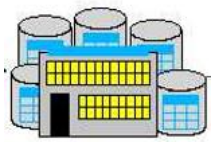
Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analyses… may lead to more confident decision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

Comment les **croiser** ?

Comment trouver le bon équilibre ?

Nouvelles opportunités d'analyse

Text mining, Web mining, etc.

L'analyse des données est rentrée dans nos mœurs

L'ÉQUIPE

« Les performances offensives et défensives des clubs de Ligue 1 en trois graphiques »

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

Recommandation basée sur les transactions.

Recommandation basée sur les utilisateurs (clients).

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)

learn.txt

	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/20
	1	I sat through this whole movie just for Harry and Ron at christmas. ohlawd
	2	with J Davlar 11th. Main rivals are team Poland. Hopefully we an make it a successful end to a toug
	0	Talking about ACT's && SAT's, deciding where I want to go to college, applying to colleges and ever
	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!!lo
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 Haley's next single is coming out in November..... I hope it's not true.

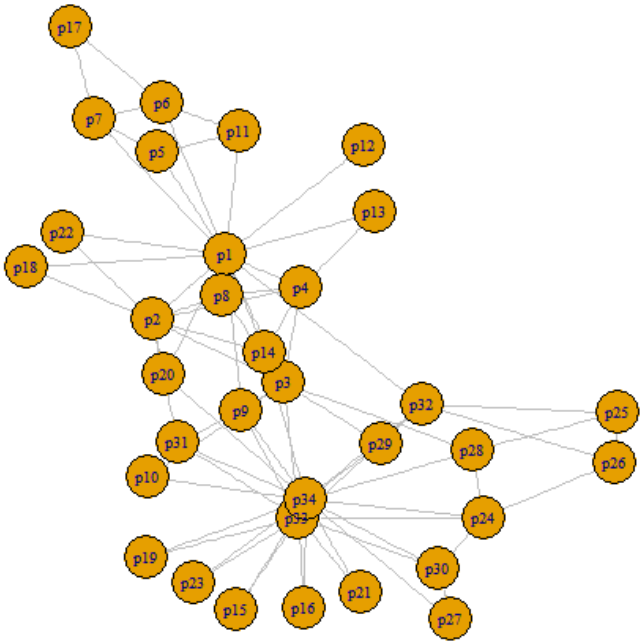
learn

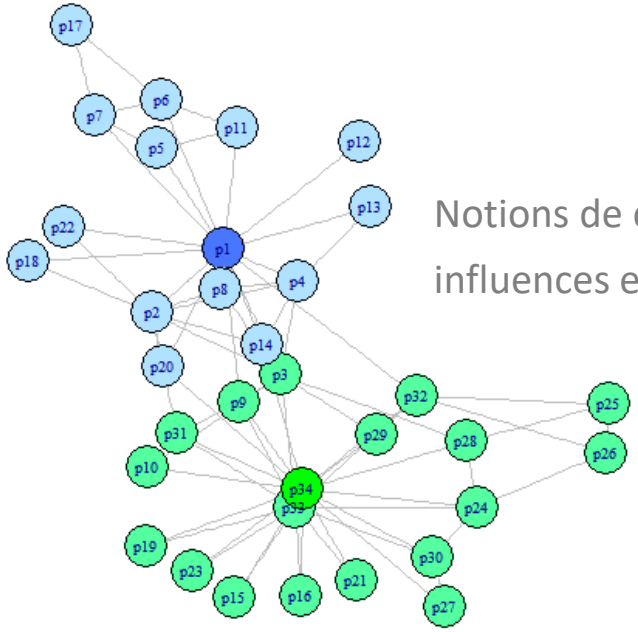
C'est du text mining avec un cadre et des finalités particulières !!!

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

Avec de nouveaux usages (3)

Détection de communautés dans les réseaux sociaux





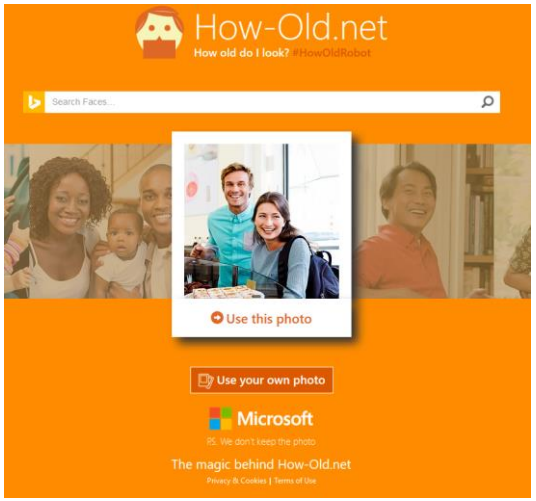
Notions de centralité, influences et communautés

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, parfois surprenantes (4)

Détection et reconnaissance des objets (la voiture grimée)

Reconnaissance faciale et détection de l'âge

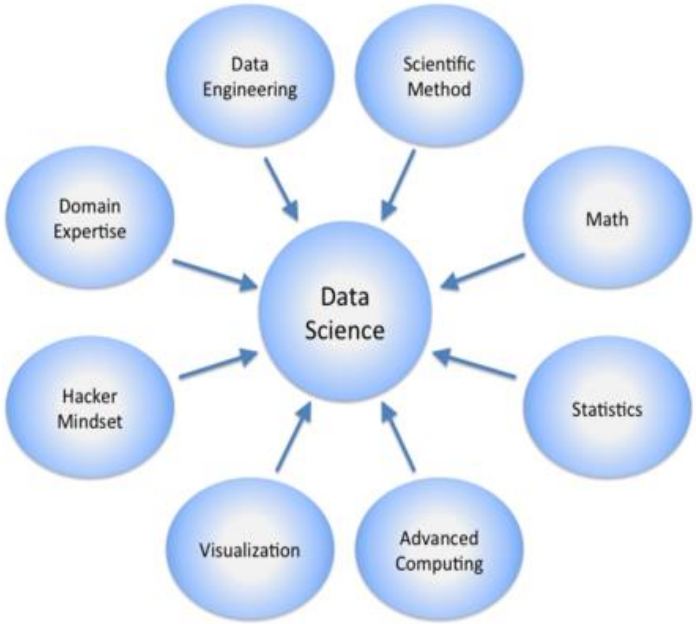


DATA SCIENCE

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.

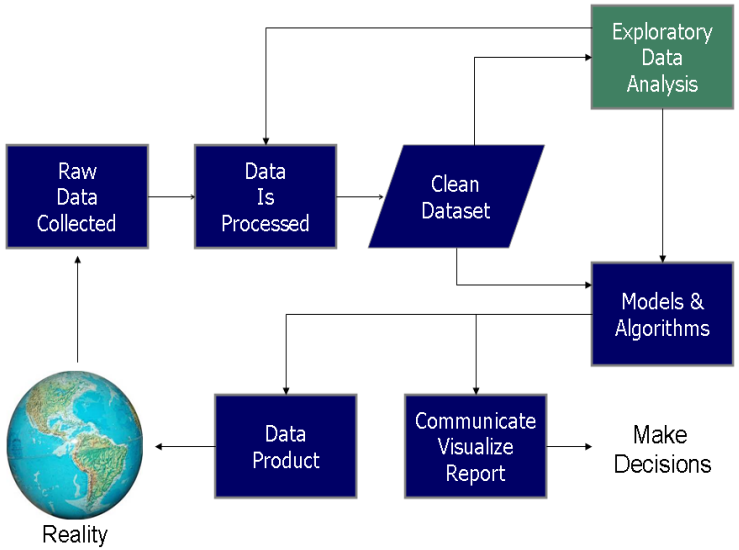
DATA MINING vs. DATA SCIENCE

Tout devient source de données

IBM CRISP DM

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

Data Science Process

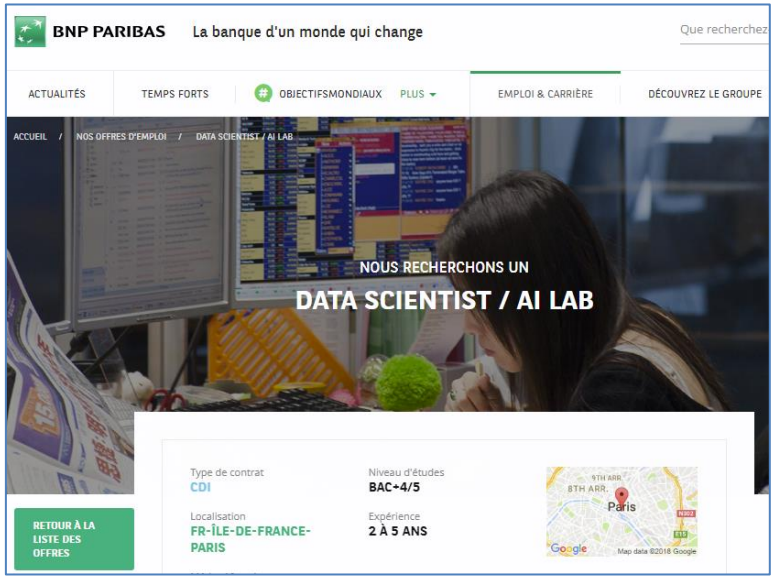


https://en.wikipedia.org/wiki/Data_science

Compétences du Data Scientist

Soyons concrets…

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



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 maitrisez 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

Python leads the 11 top Data Science, Machine Learning platforms : Trends and analysis (May 2019)

(https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html)

Data Science Tools Popularity, animated (Juin 2020)

(https://www.kdnuggets.com/2020/06/data-science-tools-popularity-animated.html)

Cloud, l'avenir ? A critical comparison of ML Platforms in an Evolving Market (Février 2021)

(https://www.kdnuggets.com/2021/02/critical-comparison-machine-learning-platforms-evolving-market.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**, peu sur les méthodes analytiques

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

EXCEL (le tableur en général)

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

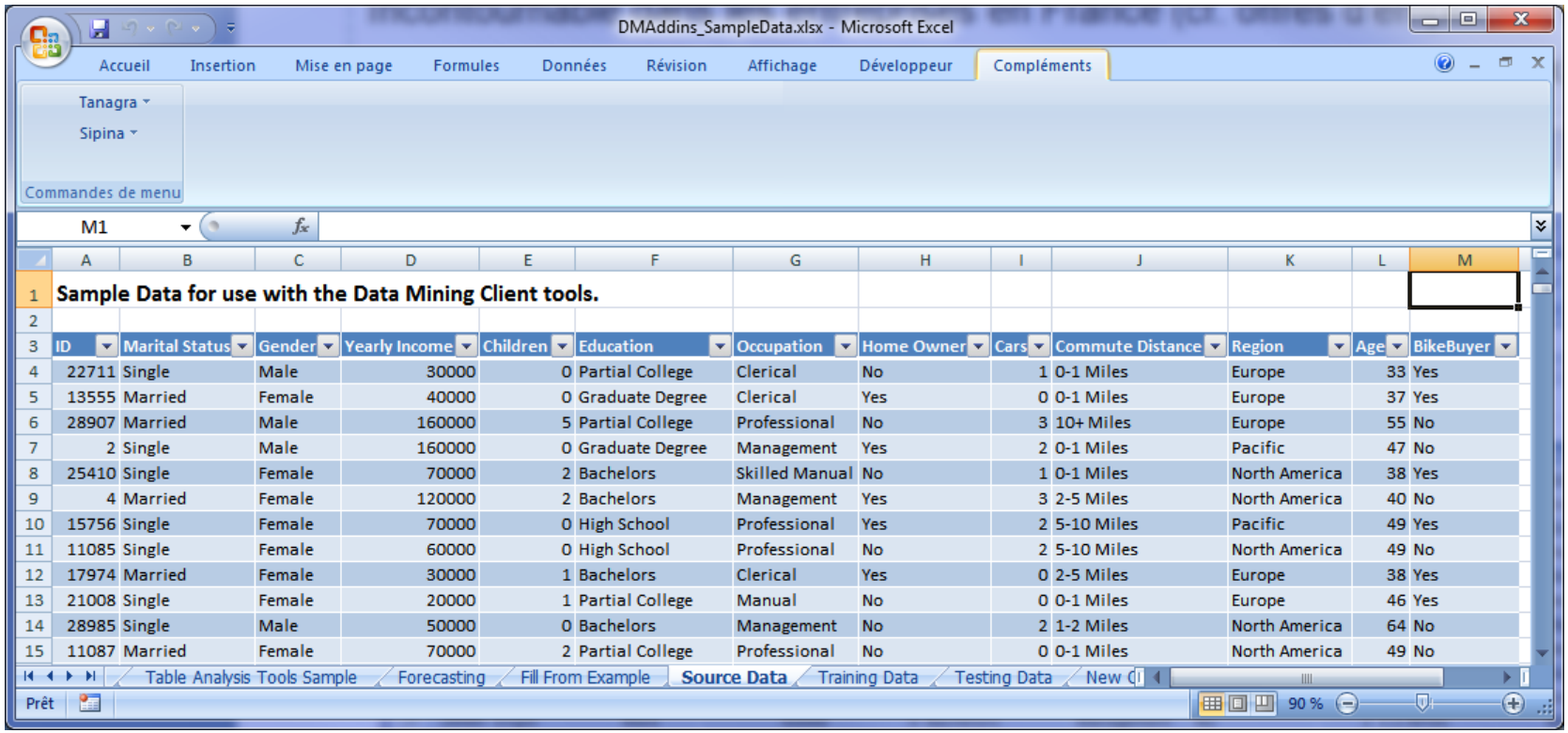
Fonctionnalités de manipulation et de préparation de données

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 KDNuggets)



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

SIPINA

Ancien, piloté par menu

Se plugge dans Excel (KDnuggets Polls, May 2013)

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

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

Homologues commerciaux

SAS, SPAD, STATISTICA, IBM/SPSS, etc.

Tutoriel : diagnostic d'une maladie cardio-vasculaire

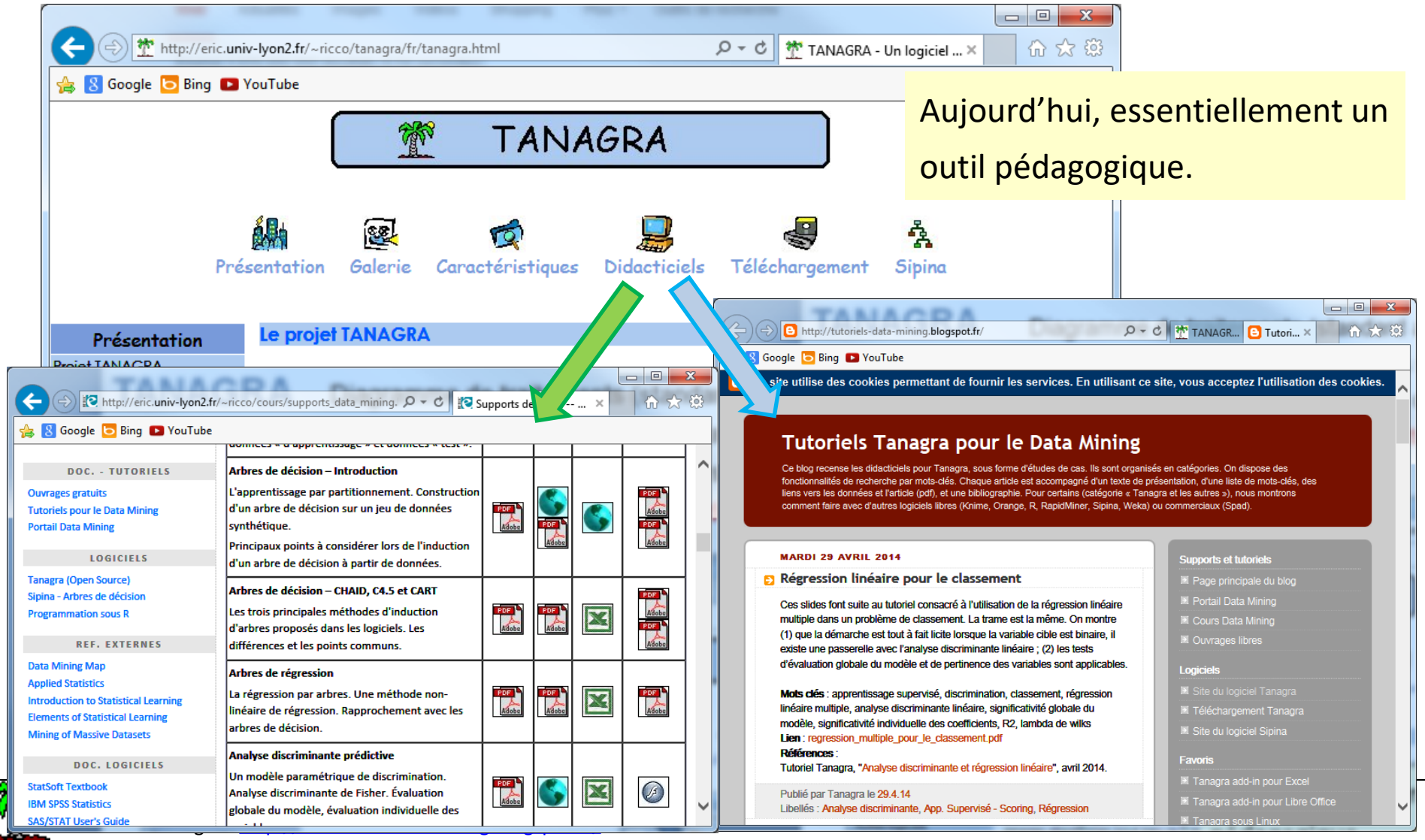
TANAGRA

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)**

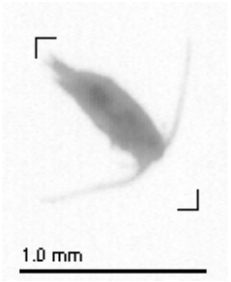


Aujourd'hui, essentiellement un outil pédagogique.

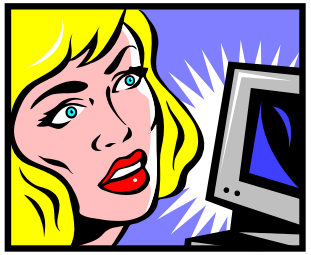
TANAGRA – Classement automatique de planctons (Image mining)

Image originelle fournie par le scanner

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



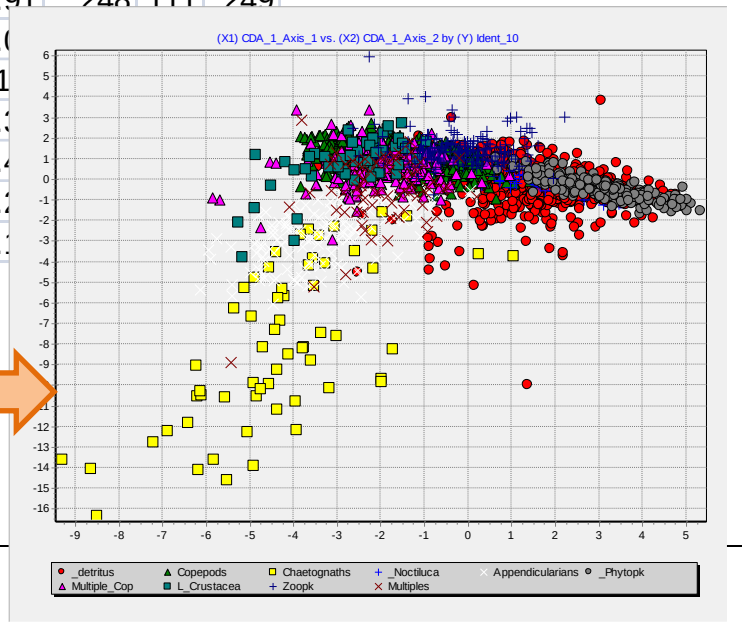
Avec l'outil ImageJ



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.(			
Copepods	403737	185.29	51			
Copepods	921755	150.98	75.3			
Chaetognaths	1017831	194.28	39.4			
_Noctiluca	648439	226.49	35.2			
Appendicularians	1564533	199.23	47.1			

Ex. de traitement : description factorielle



R

(http://www.r-project.org/)

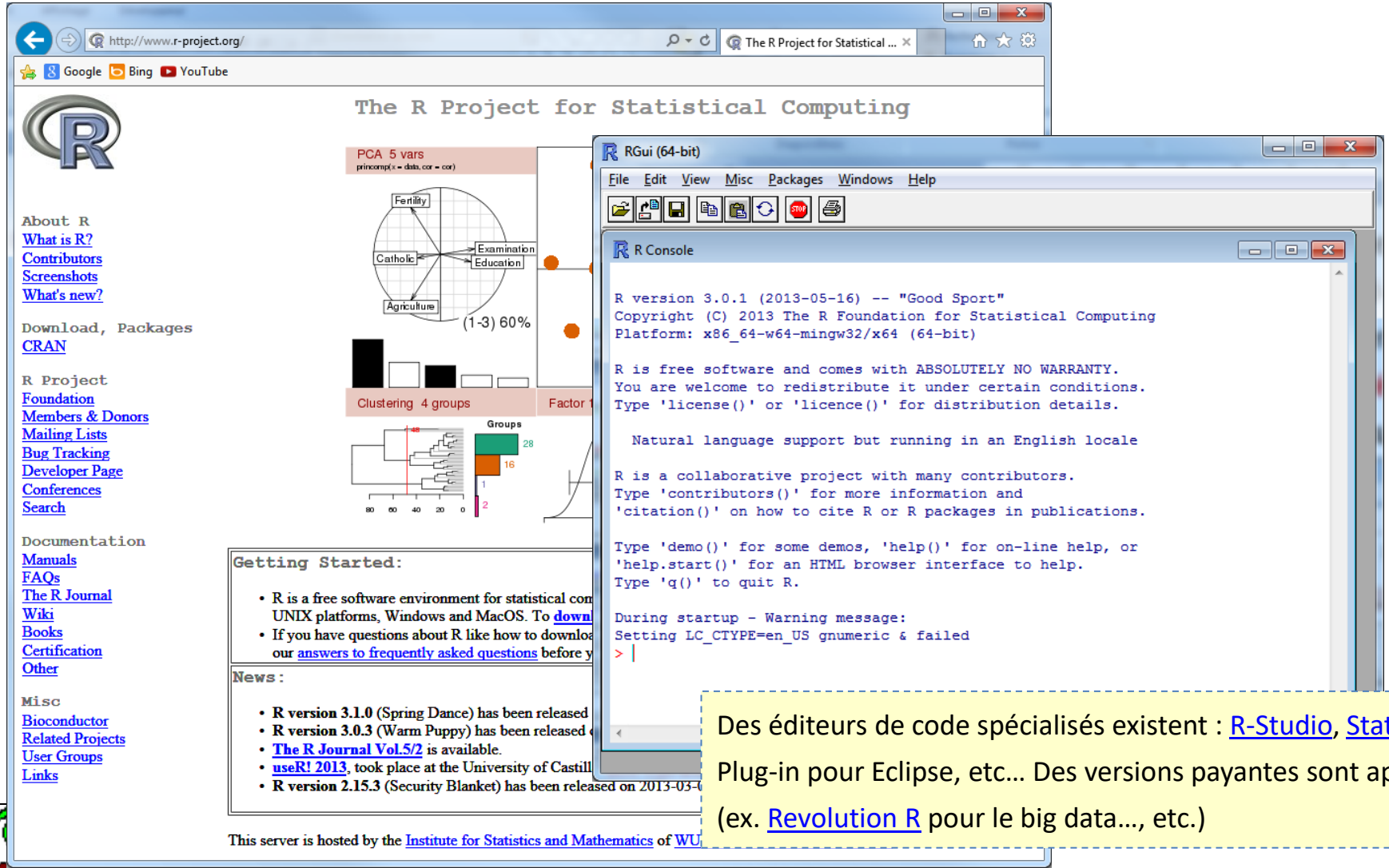
Ligne de commande + langage de programmation

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*)



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.)

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*)

*D:\DataMining\Databases_for_mining\dataset_for_s

Fichier Édition Recherche Affichage Encodage

pipeline.py

```
#librairie pandas
import pandas
#chargement de la feuille de c
#version des données à 4 varia
vote_subset = pandas.read_exce
print(vote_subset.info())
#importation de la librairie
from fanalysis.mca import MCA
#instanciation
acm = MCA(var_labels=vote_subset.columns[:4])
#apprentissage
coord = acm.fit_transform(vote_subset.iloc[:,:4].values)
#affichage des valeurs propres
print(acm.eig_)
#valeurs propres - graphique
print(acm.plot_eigenvalues())
#coordonnées des colonnes
print(acm.col_topandas())
#nomnbre var. actives
p = vote_subset.shape[1]-1
print(p)
#calcul des fonctions de projection
import numpy
fproj = numpy.zeros(acm.col_coord_.shape)
#pour chaque colonne
for j in range(fproj.shape[1]):
    fproj[:,j] = acm.col_coord_[:,j]/(p*numpy.sqrt(acm.eig_[0,j]))
#affichage fonction
print(fproj)
#affichage plus avenant des deux premiers facteurs
print(pandas.DataFrame(fproj,index=acm.col_labels_))
#taille du tableau de données présenté à l'ADL
print(coord.shape)
#10 premières lignes
print(coord[:10,:])
#classe pour l'analyse discriminante
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
#instanciation
adl = LinearDiscriminantAnalysis()
#apprentissage
adl.fit(coord,vote_subset.group)
#affichage des coefficients des fonctions de classement
print(adl.coef_)
#la constante
print(adl.intercept_)
```

Python file | length : 5 175 lines : 162 | Ln : 57 Col : 1 Sel : 0 | 0 | Windows (CR LF) | UTF-8 | INS

Python

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

Exemple : la méthode DISQUAL

KNIME

(http://www.knime.org/)

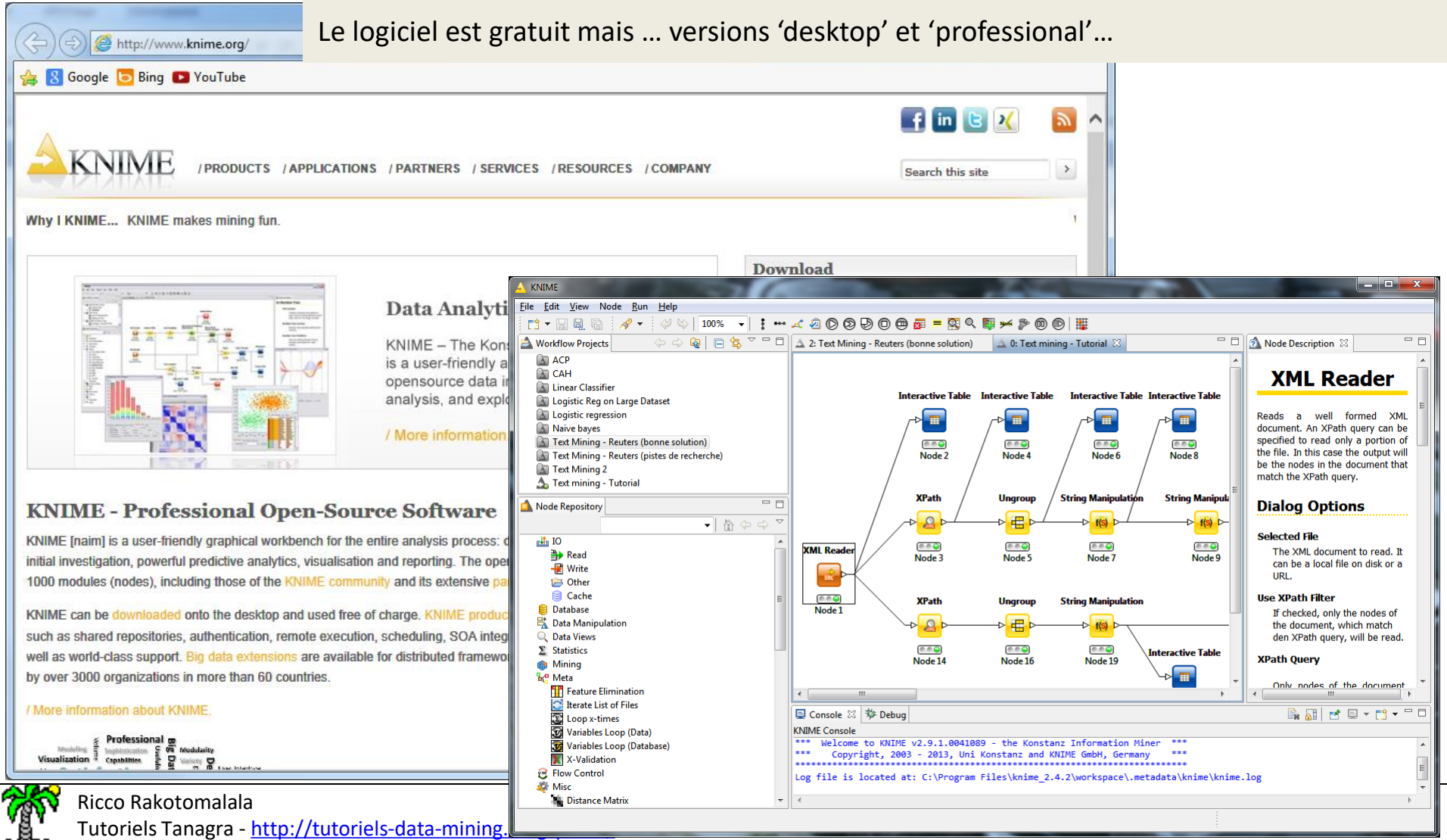
Diagramme de traitements (sur les standards des outils commerciaux, cf. IBM SPSS Modeler, SAS Enterprise Miner, SPAD, STATISTICA, …)

« 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'…



RAPIDMINER

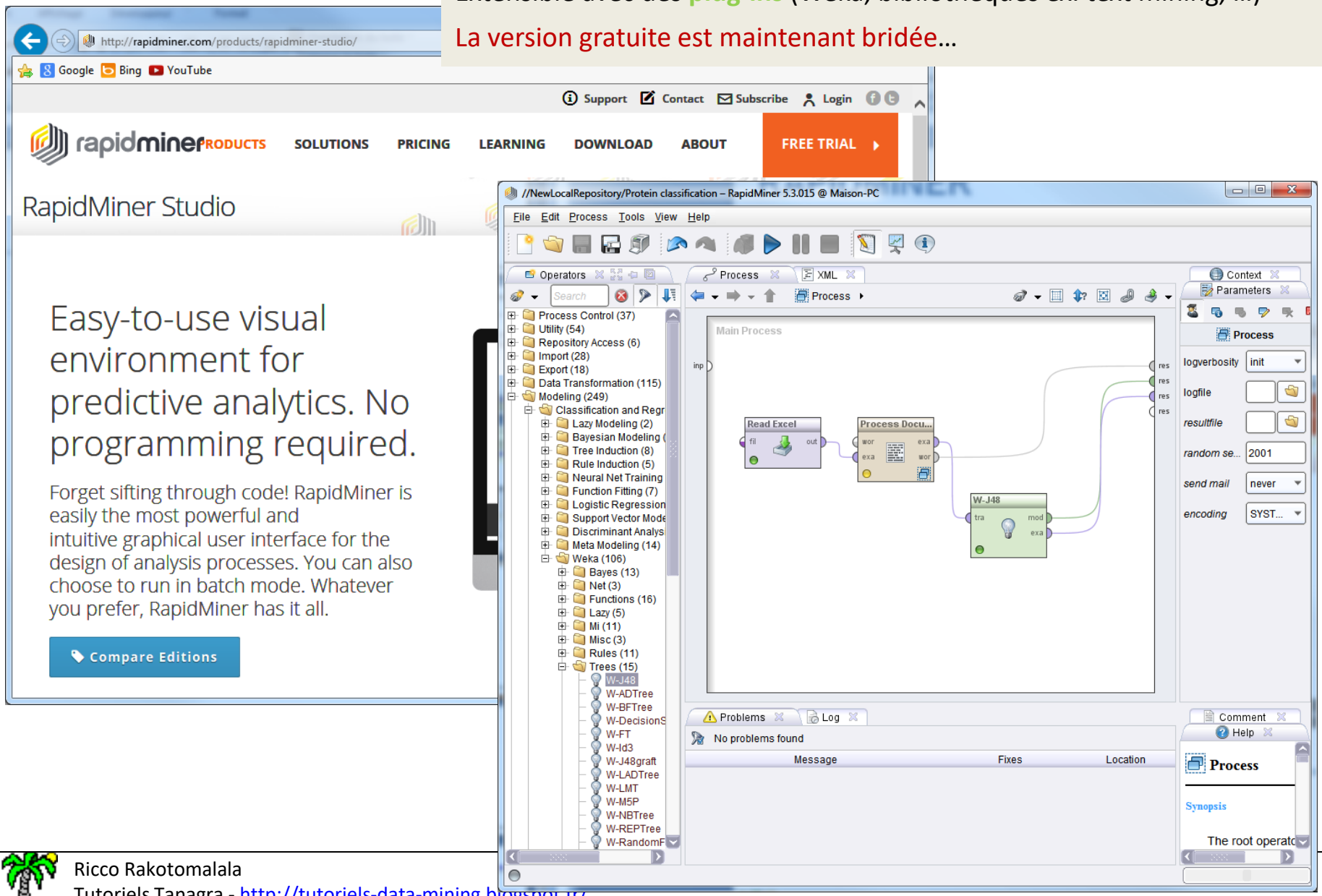
(http://rapidminer.com/)

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…



RAPIDMINER – Discrimination de protéines

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

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

proteines.xls [Mode de compatibilité]

	A	B
1	famille	description
2	F1	SQFRVSPLDRTWNLGETVELKCQVLLSNPTSGCSWLFQPRGAAASPTFLLYLSQNKPKAAEGLDTQRFSGKRLGDTFVLTLSDFRRENEGYYFCSALSNSIMYFSHFVPVFLPA
3	F2	AVSKVYARSVYDSRGNPTVEVELTTEKGVFRSIVPSGASTGVHEALEMRDGDKSKWMGKGVLHAVKNVNDVIAPAFVKANIDVKDQKAVDDFLISLDGTANKSKLGANAILGVSLAA
4	F1	EPKFTKCRSPERETFSCHWTDEVHHGPIQLFYTRRNTEWTQEWKECPDYVSAGENSCYFNSSFTSIWIPYCIKLTSNGGTVDEKCFSVDEIVQP
5	F1	LGQPTIQSFEQVGTKVNVTVEDERTLVRRNNTFLSLRDVFGKDLIYTLYYWKSSGKKTAKTNTNEFLIDVDKGENYCFSVQAVIPSRTVNRKSTDSPVECMG
6	F1	SRCTHLENRDFVTGTQGTTRVTLVLELGGCVTITAEGKPSMDVWLDAIYQENKIVYTVKVEPHTGDYVAANETHSGRKTASFTISSEKTILTMGEYGDVSLLCRVASGPVAHIEGTKYHLKS
7	F1	GSDWVIPPINLPENSRGPFPQELVRIRSGRDKNLSLRYSVTGPGADQPPTGIFIINPISGQLSVTKPLDRELIARFHLRAHAVDINGNQVENPIDIVINVIDMNDNRPEF
8	F1	ISGMSGRKASGSSPTSPINANKVENEDAFLEEVAEEKPHVKPYFTKTILDMDVVEGSAARFDCKVEGYPDPEVMWFKDDNPVKESRHFQIDYDEEGNCSLTISEVCGDDDAKYTCKAVN
9	F2	AVSKVYARSVYDSRGNPTVEVELTTEKGVFRSIVPSGASTGVHE…GTANKSKLGANAILGVSLAA
10	F2	MKIDAIEAVIVDVPTKRPIQMSITTVHQQSYVIVRVYSEGLVGV…AKAAVEMALLDLKARALGVS
11	F2	MERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGA…ANLVFNNGIDAFYARCEQV
12	F2	APAPVKQGPTSVAYVEVNNNSMLNVGKYTLADGGGNAFDVA…AGFANFPSQQAASAFAKQL
13	F2	SKIFDFVKPGVITGDDVQKVFQVAKENNFALPAVNCVGTDSIN…EHYGVPVILHTDHCAKKLLPV
14	F2	MNSNLRGVMAALLTPFDQQQALDKASLRRLVQFNIQQGIDGL…FDAVSAVTPFYYPFSFEEHCD
15	F2	VQPTPADHFTFGLWTVGWTGADPFGVATRANLDPVEAVHKL…DGGFTSNDRSIRRFALAKVL

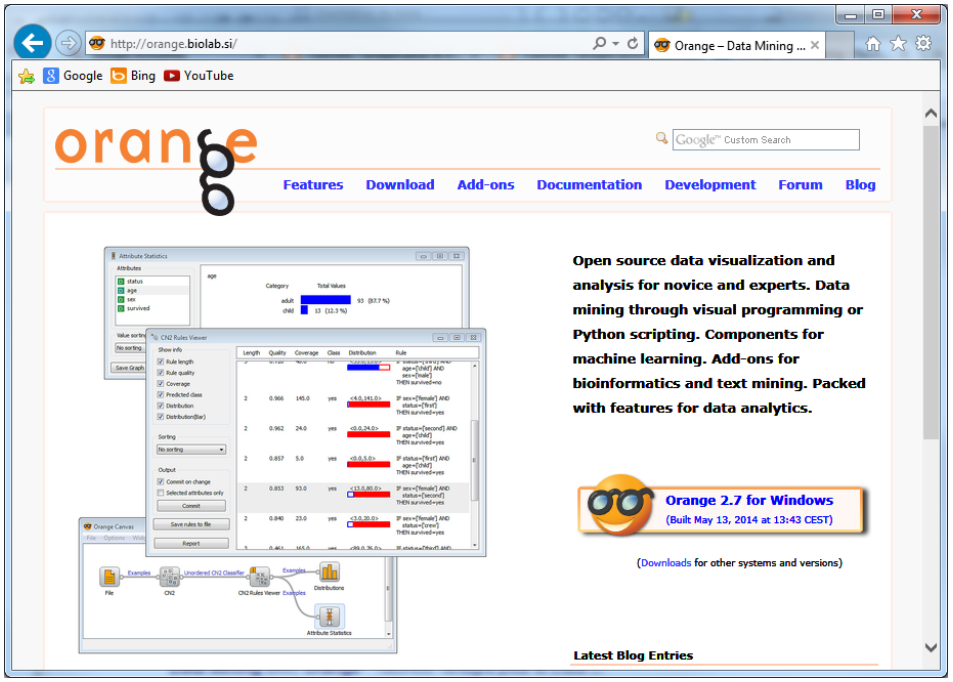
proteines

W-J48

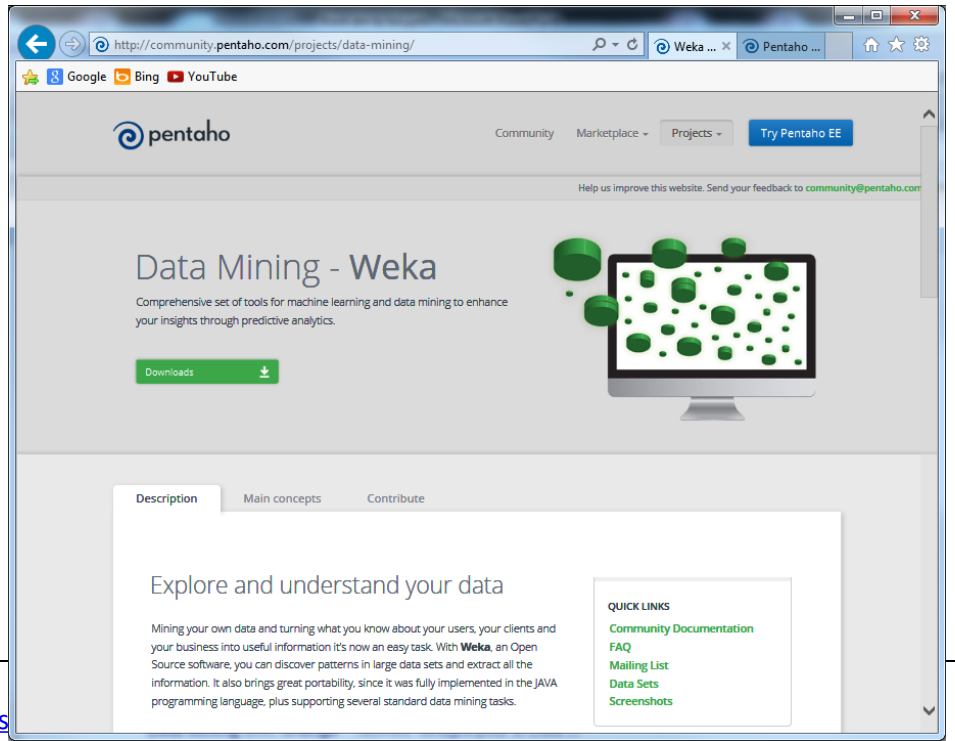
```
J48 pruned tree
------------------

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



WEKA / PENTAHO



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 !!!**

Quelques grands acteurs historiques des statistiques :

Quasiment tous maintenant proposent un mode opératoire client/serveur

- 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.