Du DATA MINING au BIG DATA

Enjeux et opportunités

Ricco RAKOTOMALALA

Université Lumière Lyon 2

Ricco ?

- Enseignant chercheur – Université Lyon 2 – CNU 27
- Econométrie - Machine Learning – Data Mining - Applications
- Logiciels gratuits (SIPINA, TANAGRA)
- Ouvrages, supports de cours, tutoriels gratuits

677 visites par jour ces 5 dernières années (01/01/2009 – 31/12/2013)

Plan

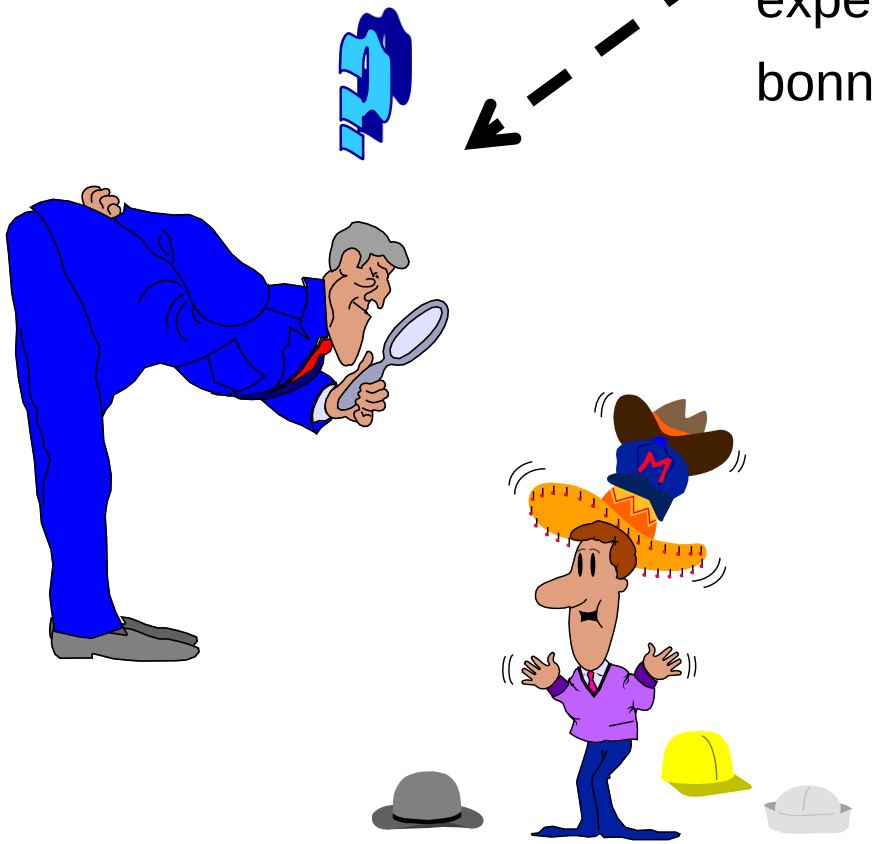
1. Data Mining – Définition
2. Spécificités du Data Mining – Applications
3. Big Data – Nouveauté, virage, évolution ?
4. Enjeux et opportunités
5. Etudes de cas avec des outils gratuits
6. Bibliographie

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

- coûteuse en stockage
- inexploitée

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/)

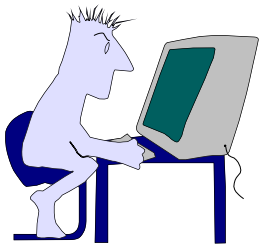
- Echantillonnage
- Préparation des données
- Visualisation des données

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

- Tests statistiques
- Re-échantillonnage

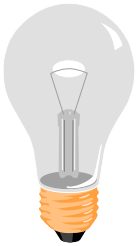
Bases de données

table



Fouille des données (Data Mining)

modèles

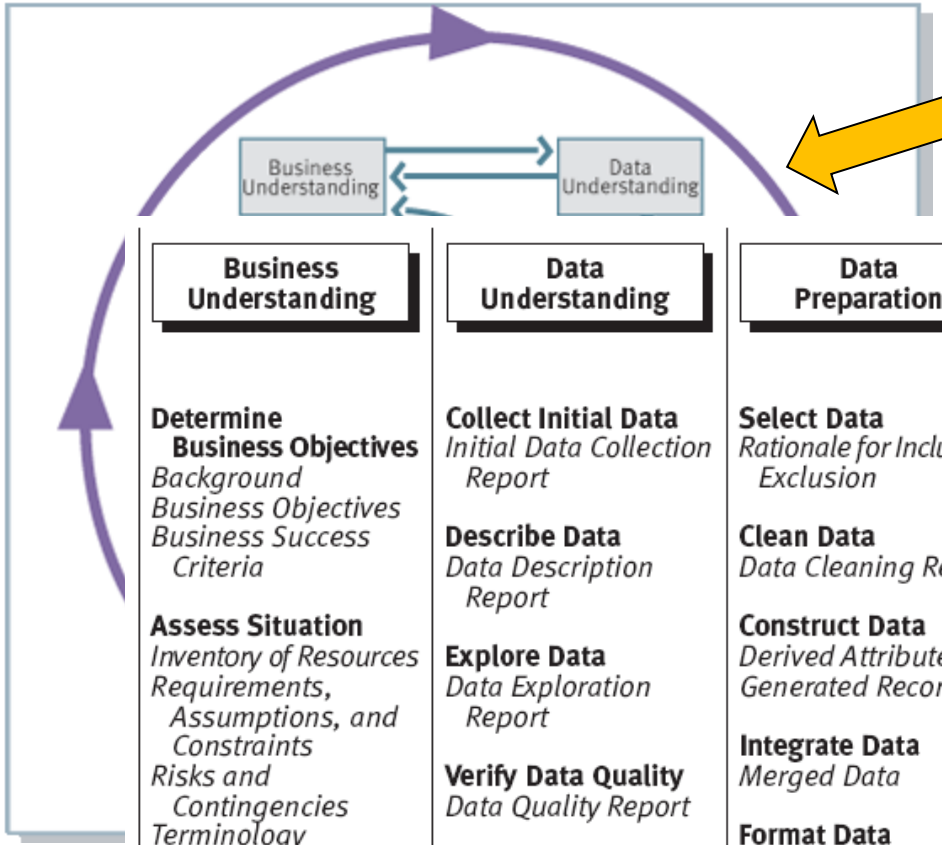


Mise en forme des Connaissances

Connaissances

Déploiement Exploitation

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



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

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.Benzécri1973)

The basic steps for developing an effective process model ?

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

http://www.theanalysisfactor.com/13-steps-regression-anova/

1. Model selection
2. Model fitting
3. Model validation

Domaines d'application

Partout où il y a des données disponibles, et où il y a des comportements ou des appétences à comprendre et à exploiter… Mots clés : **scoring** et **ciblage**.

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

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…)

Quels métiers ?

Data scientist – Un profil d'avenir

http://pro.clubic.com/it-business/actualite-693592-data-scientist-mouton-5-pattes-coeur-donnees.html

Le commerce en ligne s'arrache les **data miners**

http://www.lesechos.fr/15/07/2012/lesechos.fr/0202173368914_le-commerce-en-ligne-francais-s-arrache-les---data-miners--.htm

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

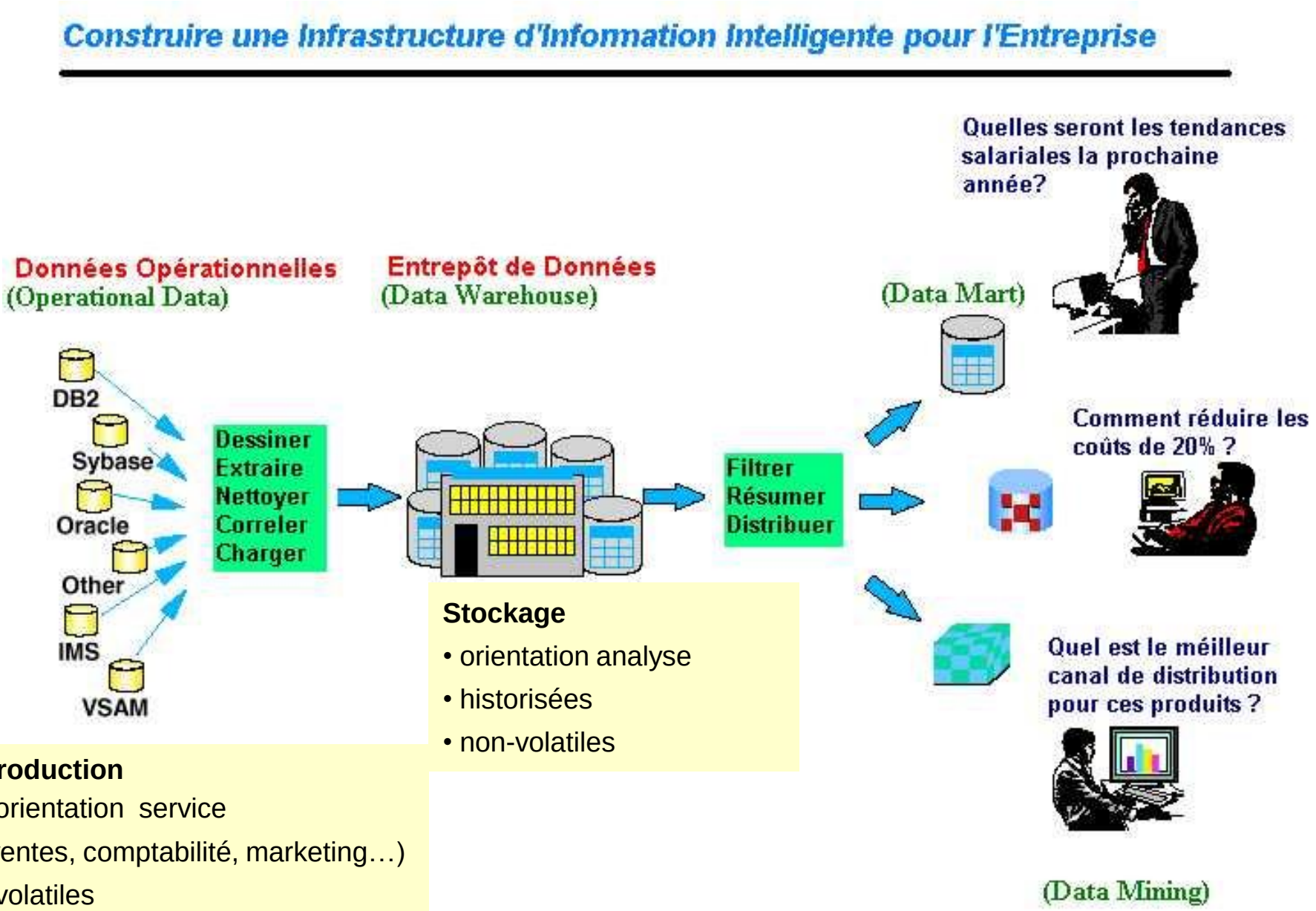
Spécificités du Data Mining ?

(1) Sources de données

(2) Techniques utilisées

(3) Multiplicité des supports

Spécif.1 - Les sources de données



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

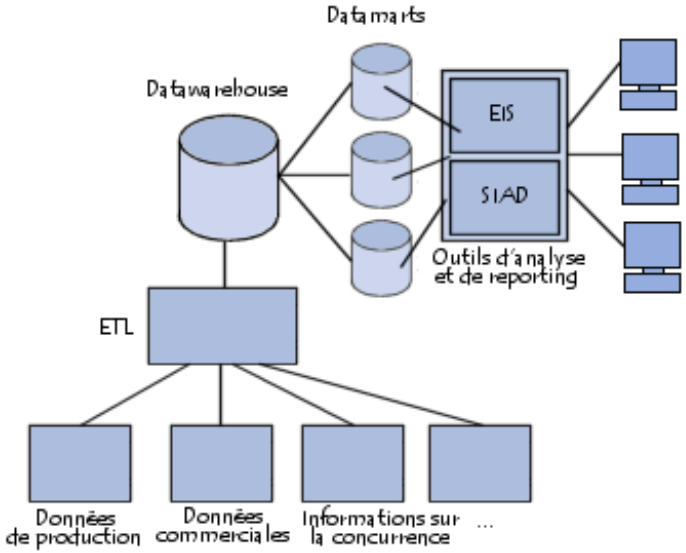
Entrepôts / Datamarts : Sources de données pour l'analyse

Conséquence : la volumétrie devient un élément important !!!

➔ **Découverte de connaissances à partir de données volumineuses (Data Mining)**

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)

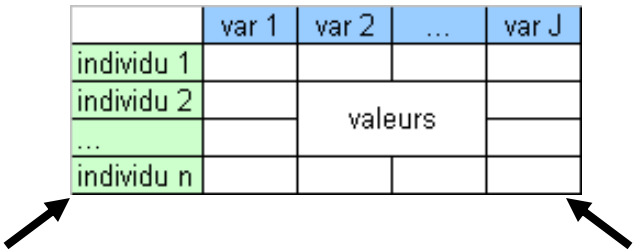
Description factorielle

Discrimination

Clustering

Méthodes géométriques, probabilités

ACP, ACM, Analyse discriminante, CAH, …



Informatique

(Intelligence artificielle)

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

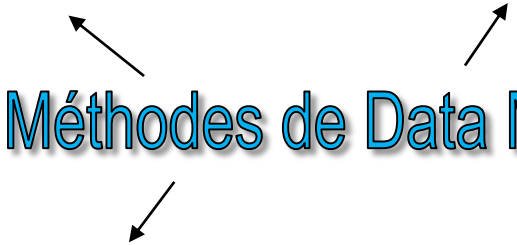
Ex : moyennes conditionnelles, etc.

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



(cf. Tenenhaus)

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 :

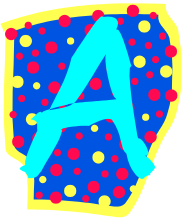
Trouver les ensembles de descripteurs
qui sont le plus corrélés

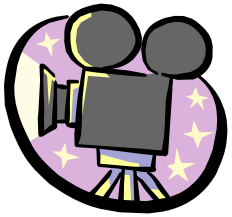
- **règles d'association**

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

Les méthodes sont le plus souvent complémentaires !

Spécif.3 - Multiplicité des supports et des sources





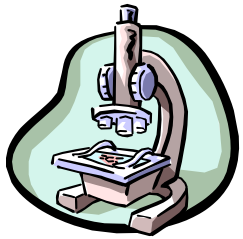
Les applications

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

Image mining (ex. Analyse des mammographies, …)

Etc.

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**…(ex. N° RNTI)

BIG DATA

Tout le monde en parle… 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].

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

EM-Grenoble, ENSAI, Telecom ParisTech, …

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.)

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.

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.

Définition (autres acteurs) – Big data = opportunités d'analyses

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/

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) !

Big Data vs. Data Mining

BIG DATA = DATA MINING ++

➔ Du moins dans sa partie BIG ANALYTICS (il y a d'autres aspects, ex. stockage, accès, …)

➔ Les nouvelles caractéristiques des données (3 V) deviennent partie intégrante du problème (cf. KDnuggets Polls) ou instillent de nouveaux défis technologiques et méthodologiques (ex. Mahout / Hadoop, …)

➔ **Internet** est un acteur / vecteur important de profusion des données (cf. https://www.youtube.com/watch?v=3aoxbaVm11E ; 6mn16)

➔ La vague « open data » va amplifier le déluge (des données)… et les attentes en termes d'analyse (http://lecubevert.fr/open-data-definitions-enjeux-et-perspectives/)

Big Data vs. Data Mining (Ex. text mining)

Analyse des opinions (sentiments, approbation, désapprobation, etc.) sur twitter

http://www.journaldunet.com/solutions/saas-logiciel/ordre-de-bourse-et-analyse-d-opinion-avec-twitter.shtml

https://asunews.asu.edu/20130920-kambhampati-twitter-analysis/

http://www.latribune.fr/opinions/tribunes/20140213trib000815265/les-cles-d-une-veritable-analyse-semantique-sur-twitter.html

(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..
	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.)

Big Data vs. Business Intelligence

(Wikipédia) ...la maturation du sujet fait apparaitre un autre critère plus fondamental de différence d'avec le Business Intelligence et concernant les données et leur utilisation :

➔ Business Intelligence : utilisation de statistique descriptive [reporting, tableaux de bord,…], sur des données à forte densité en information afin de mesurer des phénomènes, détecter des tendances… ;

➔ Big Data : utilisation de statistique inférentielle, sur des données à faible densité en information dont le grand volume permet d'inférer des lois (régressions….) donnant dès lors (avec les limites de l'inférence) au big data des capacités prédictives [modélisation, analyse prédictive,…].

Certains annoncent même la mort de la BI « traditionnelle »…(Les Echos, juin 2013)… *on n'en est pas là encore*…

Big Data

(Quelques) Enjeux et nouveaux développements

Une innovation guidée surtout par **l'évolution technologique et la volumétrie**

Données rares, enquêtes planifiées → Data Warehouse → Big Data

Statistique / Analyse de données → Data Mining → Big Analytics

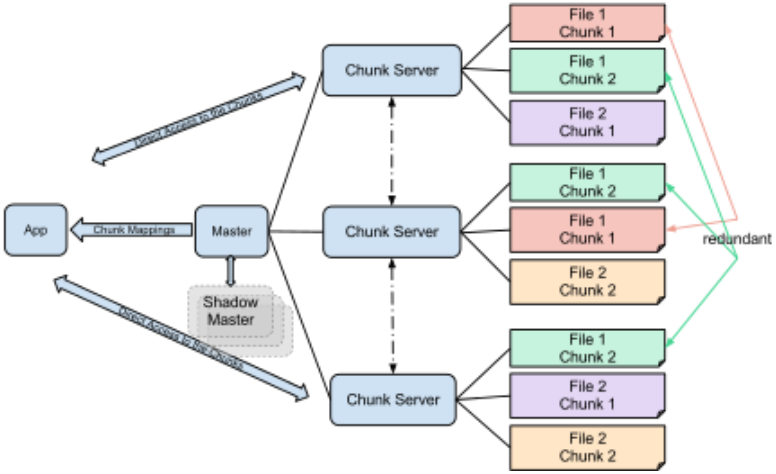
Repenser le stockage des données

Google File System (http://research.google.com/archive/gfs.html, 2003)

Le modèle MapReduce

La solution HADOOP qui implémente ces 2 principes (Framework Java)

(http://fr.wikipedia.org/wiki/Google_File_System)

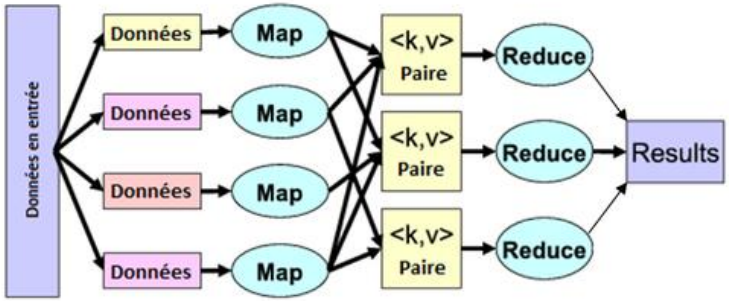


Les fichiers sont distribués sur des serveurs (clusters) qui peuvent être des machines « banales »

Les capacités de stockage sont démultipliées

Le coût de stockage est fortement réduit

(http://fr.wikipedia.org/wiki/MapReduce)



MAP : découpage d'un problème en sous-problèmes, traitements sur les nœuds (clusters)

REDUCE : consolidation des résultats, en s'appuyant sur la paire < clé , valeur > renvoyée par les nœuds

Repenser les implémentations des algorithmes

(1) S'appuyer sur le principe MapReduce dans l'implémentation des algorithmes de data mining (ex. Mahout)

(2) Un exemple sous R (MapReduce in R)

Le code usuel…

```
groups = rbinom(10, n = 50, prob = 0.4)
tapply(groups, groups, length)
```

…devient

```
groups = to.dfs(groups)
from.dfs
(
        mapreduce
        ( input = groups,
         map = function(., v) {keyval(v, 1)},
         reduce = function(k, vv) {keyval(k, length(vv))}
        )
)
```

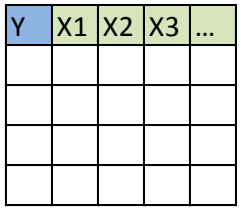
(3) D'autres pistes de parallélisation existent (ex. Programmation parallèle sous R)

Repenser les méthodes elles-mêmes ?

Au tout début du data mining

« p » modéré

« n » augmente



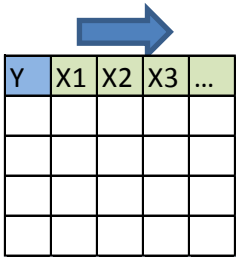
Méthodes classiques (statistique, machine learning)

Augmentation de la puissance de calcul

Arrivée des données non structurées

« n » faible

« p » augmente

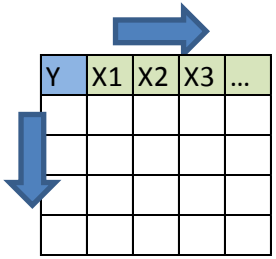


Avènement des méthodes fortement régularisées (support vector machine, regression ridge, random forest, … régression PLS…)

Contexte BIG DATA

« n » augmente

« p » augmente



Parallélisation, calcul distribué, OK.

Mais au niveau des méthodes ? Inventer des méthodes avec des temps de calcul contraints ?

Améliorer l'intégration des données de différentes natures

Fouille de données complexes, **« variété »** plus et encore…

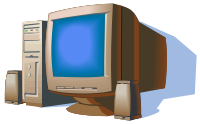
Rapport d'analyse

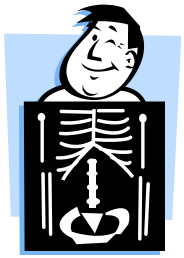
Dossier patient

Données publiques,
« open data », etc.

Radiographies

Forums, réseaux
sociaux, etc.





Comment les **croiser** ?

Comment trouver le bon équilibre ?

Etudes de cas avec des outils gratuits

(Outils accessibles « librement » sur le net)

Cahier des charges – Logiciel de Data Mining

Logiciels commerciaux

Prototypes de recherche

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

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…)

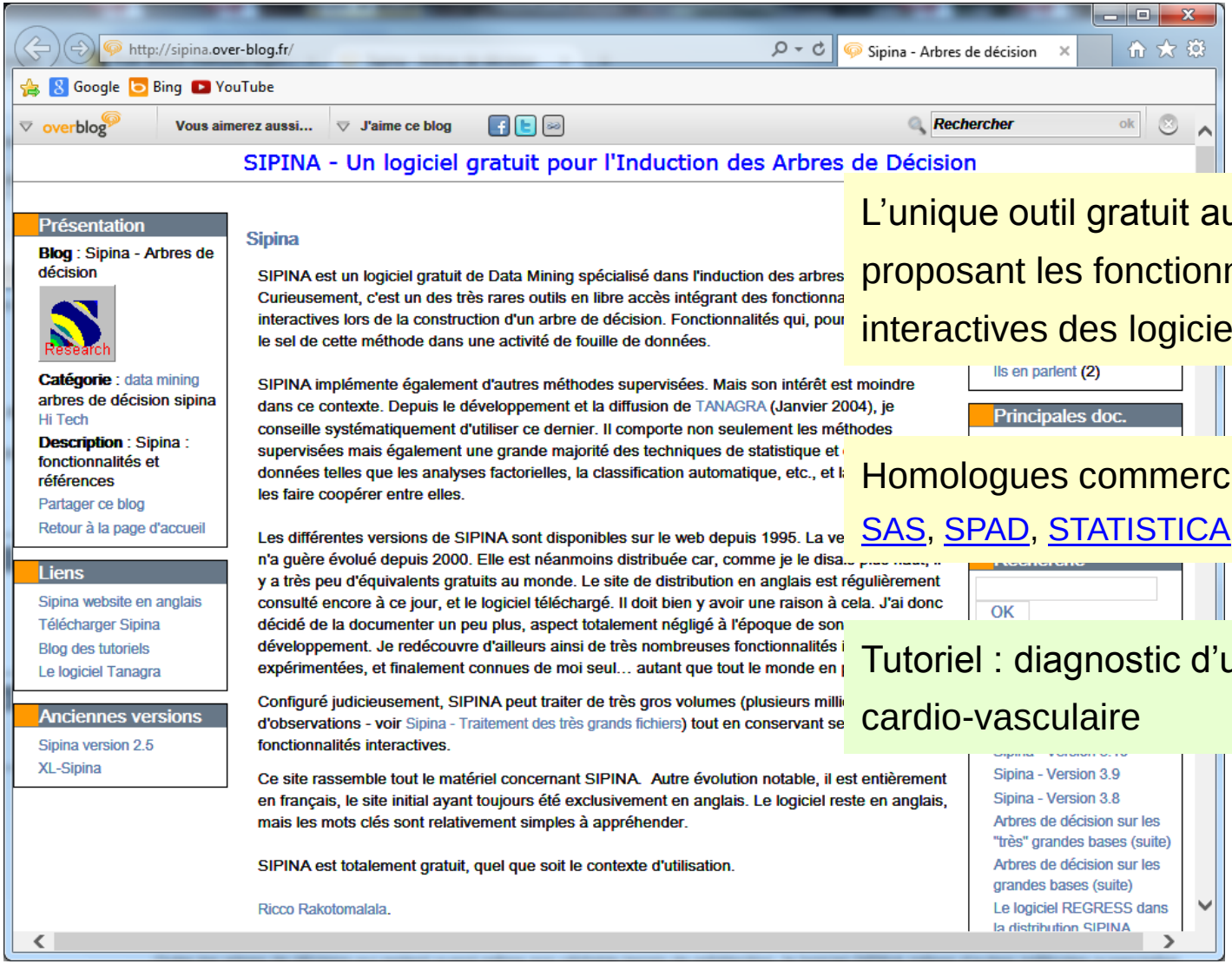
L'évolution porte sur les technologies, pas sur les méthodes analytiques

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)

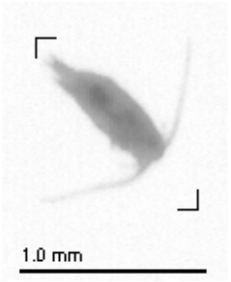
Simplicité, **facilité d'utilisation**, **documentation très abondante (FR et EN)**

Aujourd'hui, essentiellement un outil pédagogique.

TANAGRA – Classement automatique de planctons (Image mining)

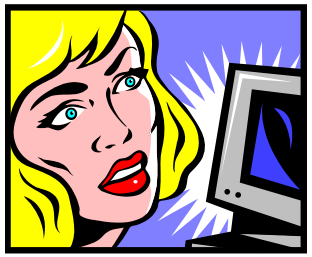
Image originelle fournie par le scanner

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



Avec l'outil

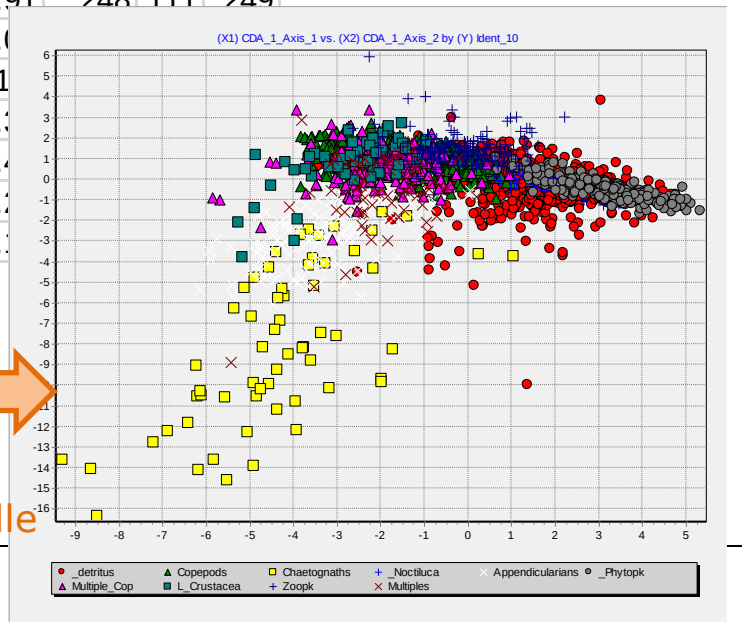
ImageJ



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/)

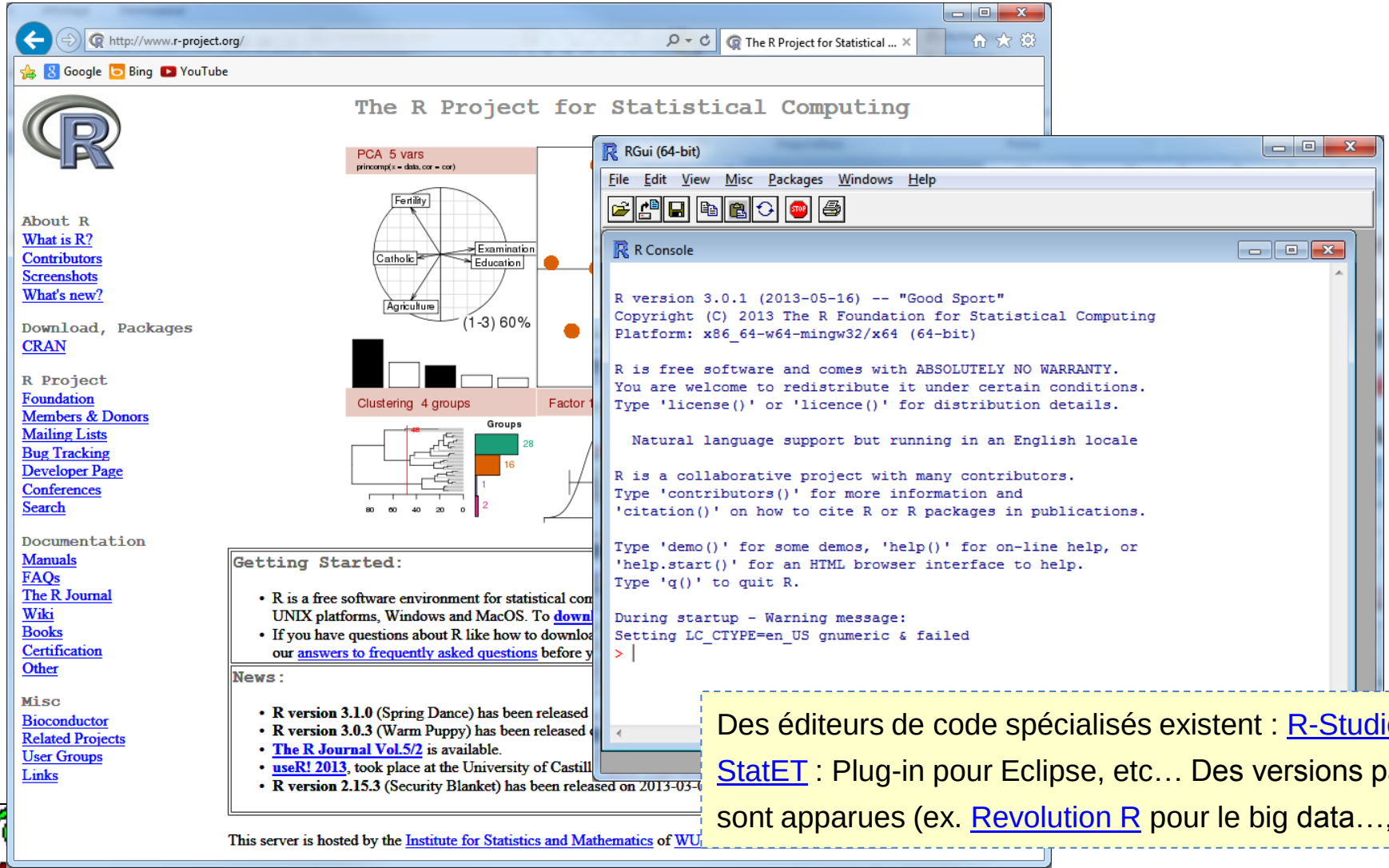
Ligne de commande + langage de programmation

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

Extensible à l'infini avec le système des **packages**

La référence actuelle et future (outil + langage)

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.)

KNIME

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

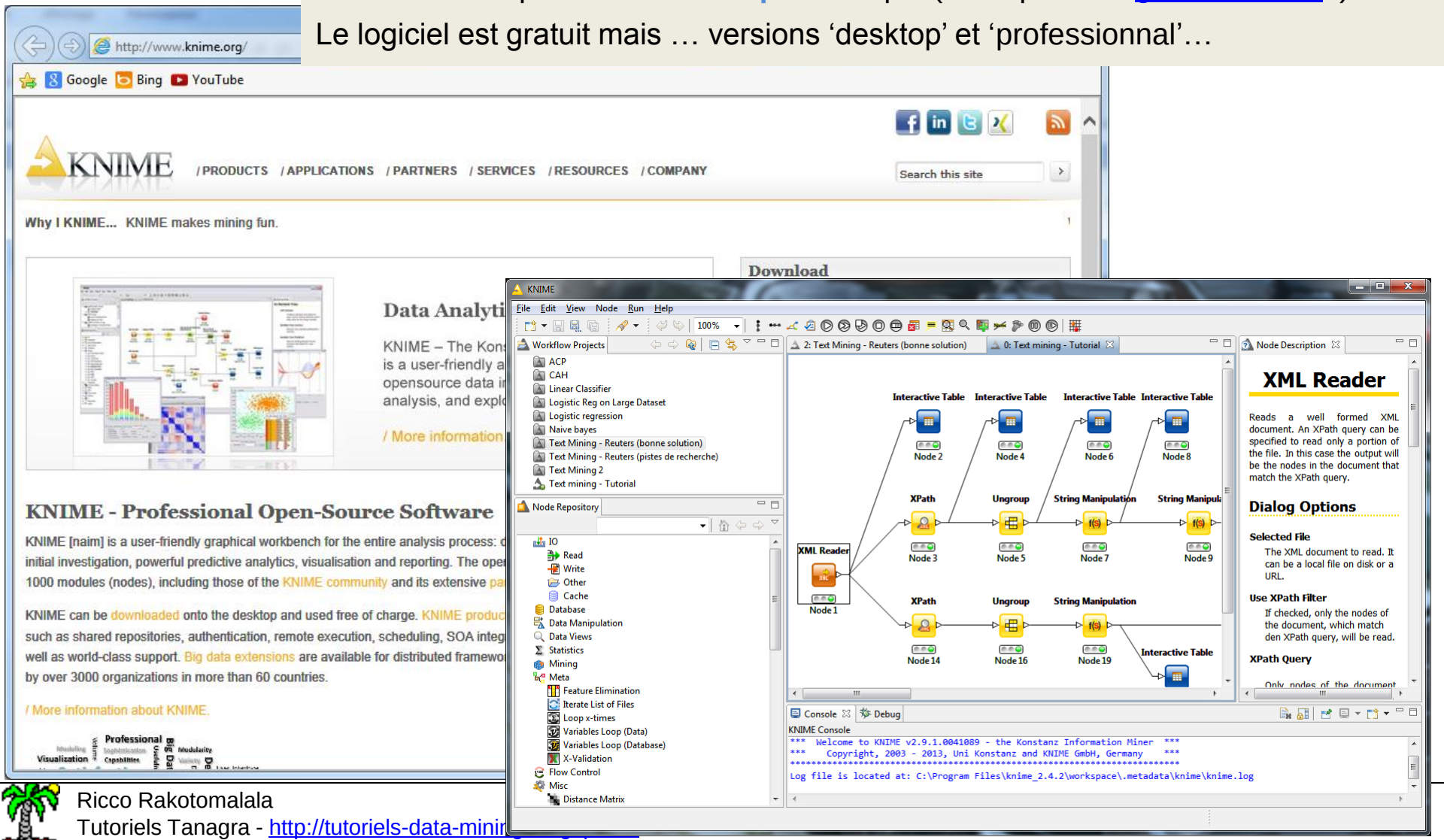
Diagramme de traitements (sur les standards des outils commerciaux, cf. IBM SPSS Modeler, SAS Enterprise Miner, SPAD, STATISTICA, …)

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

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

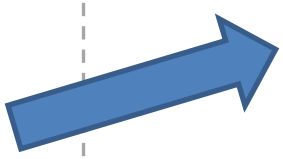
Multithread et possibilité de **swap** sur disque (armé pour les gros volumes ?)

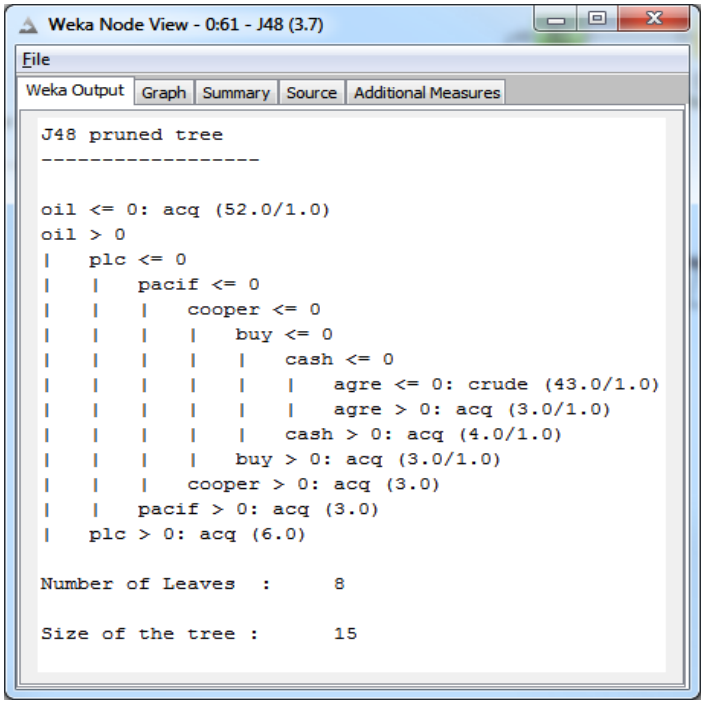
Le logiciel est gratuit mais … versions 'desktop' et 'professionnal'…

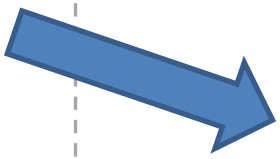


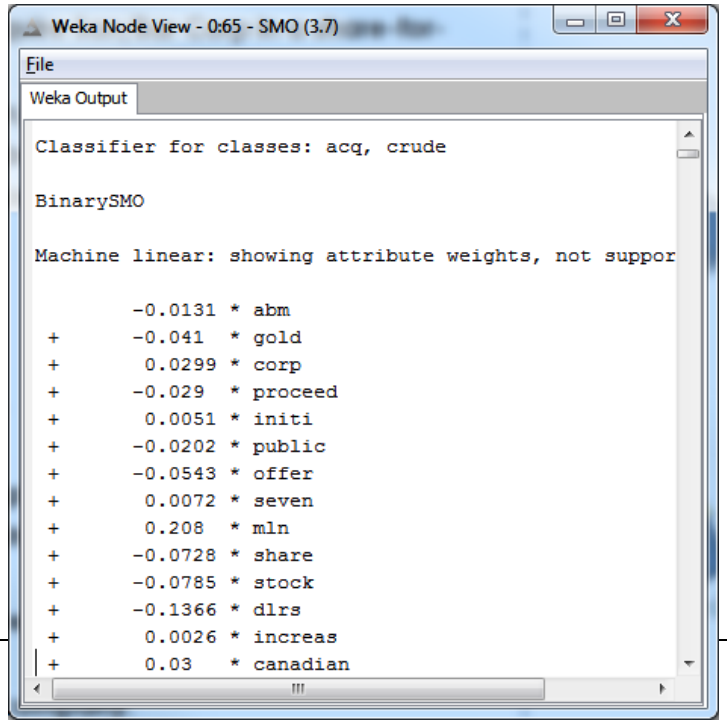
KNIME – Catégorisation de nouvelles (Reuters)

```
<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>
```









RAPIDMINER

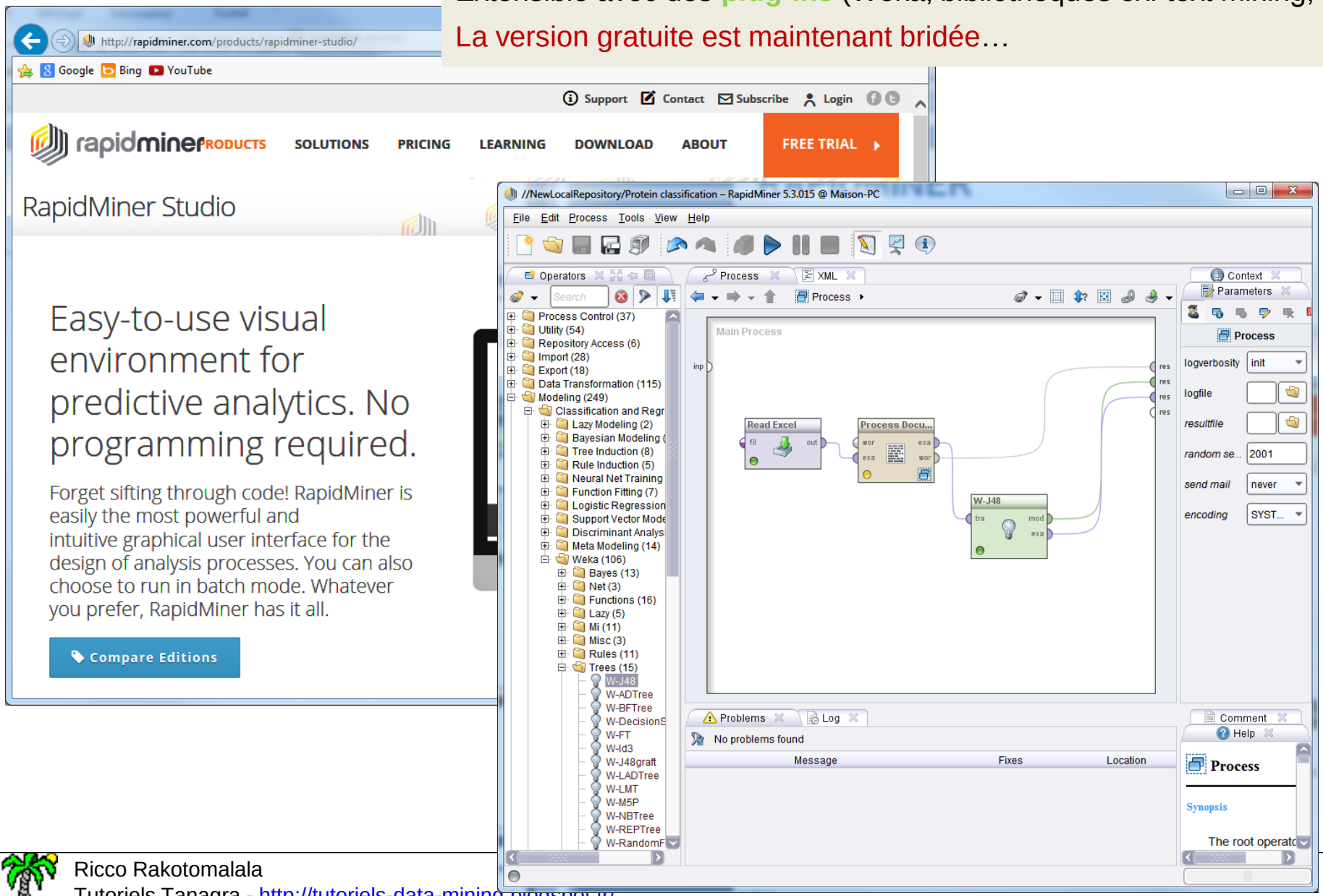
(http://rapidminer.com/)

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…KAAVEMALLDLKARALGVS
11	F2	MERYENLFAQLNDRREGAFVPFVTLGDPGIEQSLKIIDTLIDAGA…ANLVFNNGIDAFYARCEQV
12	F2	APAPVKQGPTSVAYVEVNNNSMLNVGKYTLADGGGNAFDVA…AGFANFPSQQAASAFAKQL
13	F2	SKIFDFVKPGVITGDDVQKVFQVAKENNFALPAVNCVGTDSIN…EHYGVPVILHTDHCAKKLLPV
14	F2	MNSNLRGVMAALLTPFDQQQALDKASLRRLVQFNIQQGIDGL…FDAVSAVTPFYYPFSFEEHCD
15	F2	VQPTPADHFTFGLWTVGWTGADPFGVATRANLDPVEAVHKL…DGGFTSNDRSIRRFALAKVL

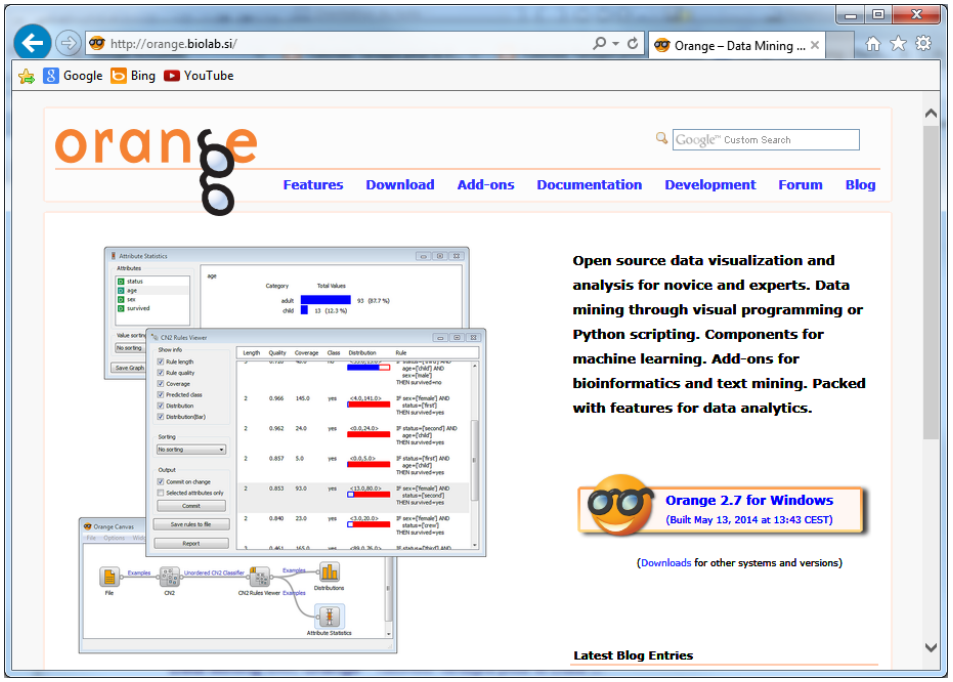
proteines

W-J48

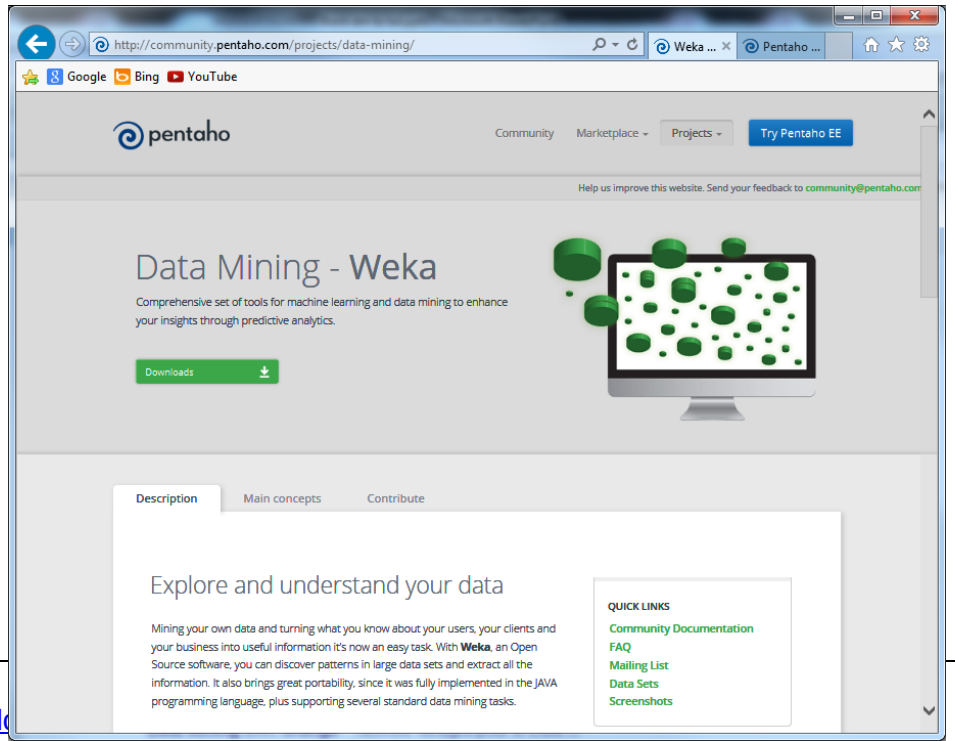
```
J48 pruned tree
------------------

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



Bibliographie

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.