

Introduction au Data Mining

Ricco Rakotomalala
Université Lumière Lyon 2



Ricco Rakotomalala

- ricco.rakotomalala@univ-lyon2.fr
- <http://chirouble.univ-lyon2.fr/~ricco/cours/>
Publications, ressources, liens, logiciels, ...



Plan

1. Qu'est ce que le Data Mining ?
2. Spécificités du Data Mining
3. Quelques exemples
4. Typologie des méthodes de Data Mining
5. Ressources - Sites web et bibliographie



Data Mining ?

Une démarche plus qu'une théorie !



Exemple introductif : demande de crédit bancaire



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



Expérience de l'entreprise : ses clients et leur comportement

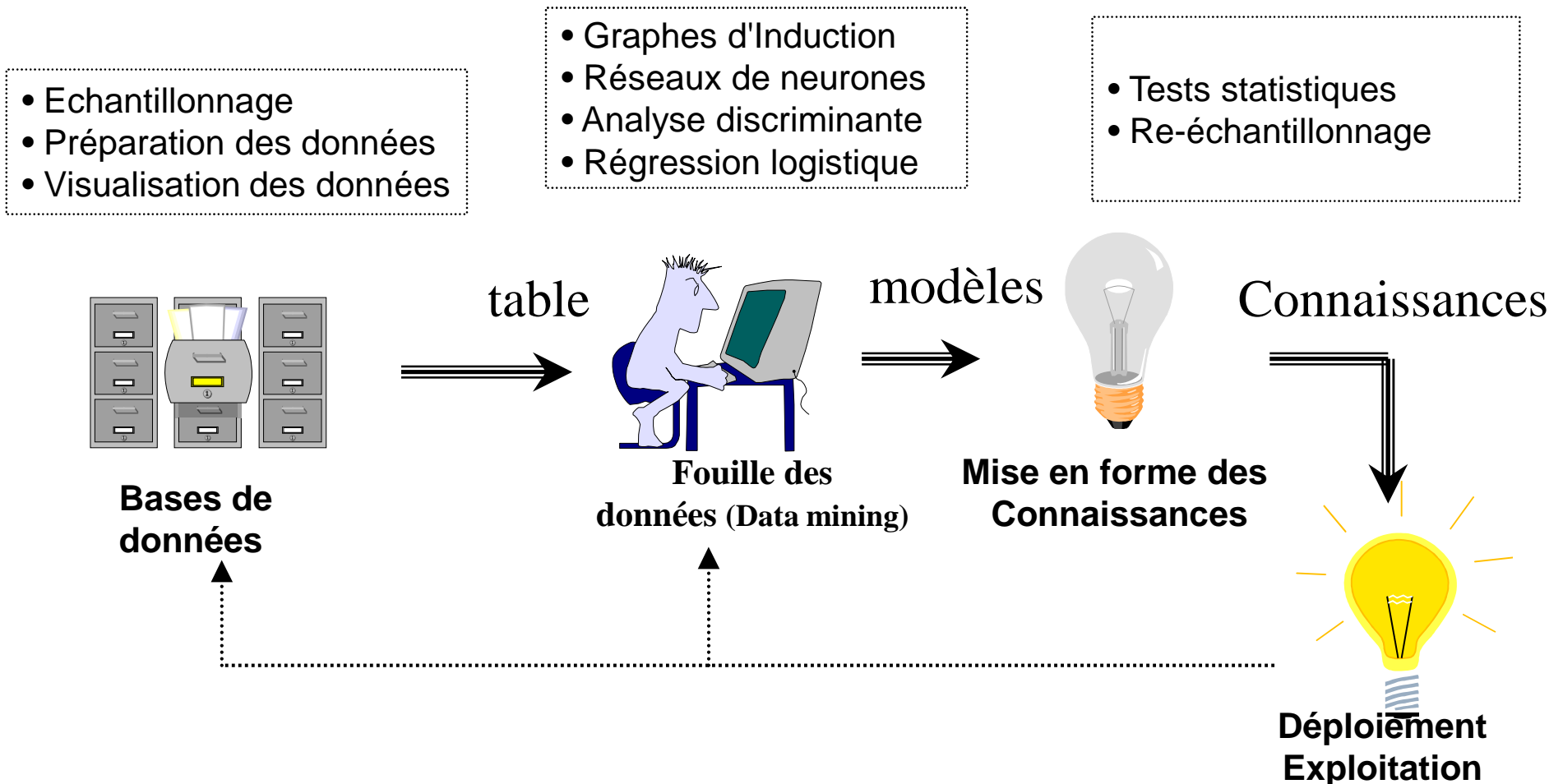


- coûteuse en stockage
- inexploitée

Comment et à quelles fins utiliser cette expérience accumulée



Le processus ECD (Extraction de connaissances à partir de données) KDD – Knowledge discovery in Databases

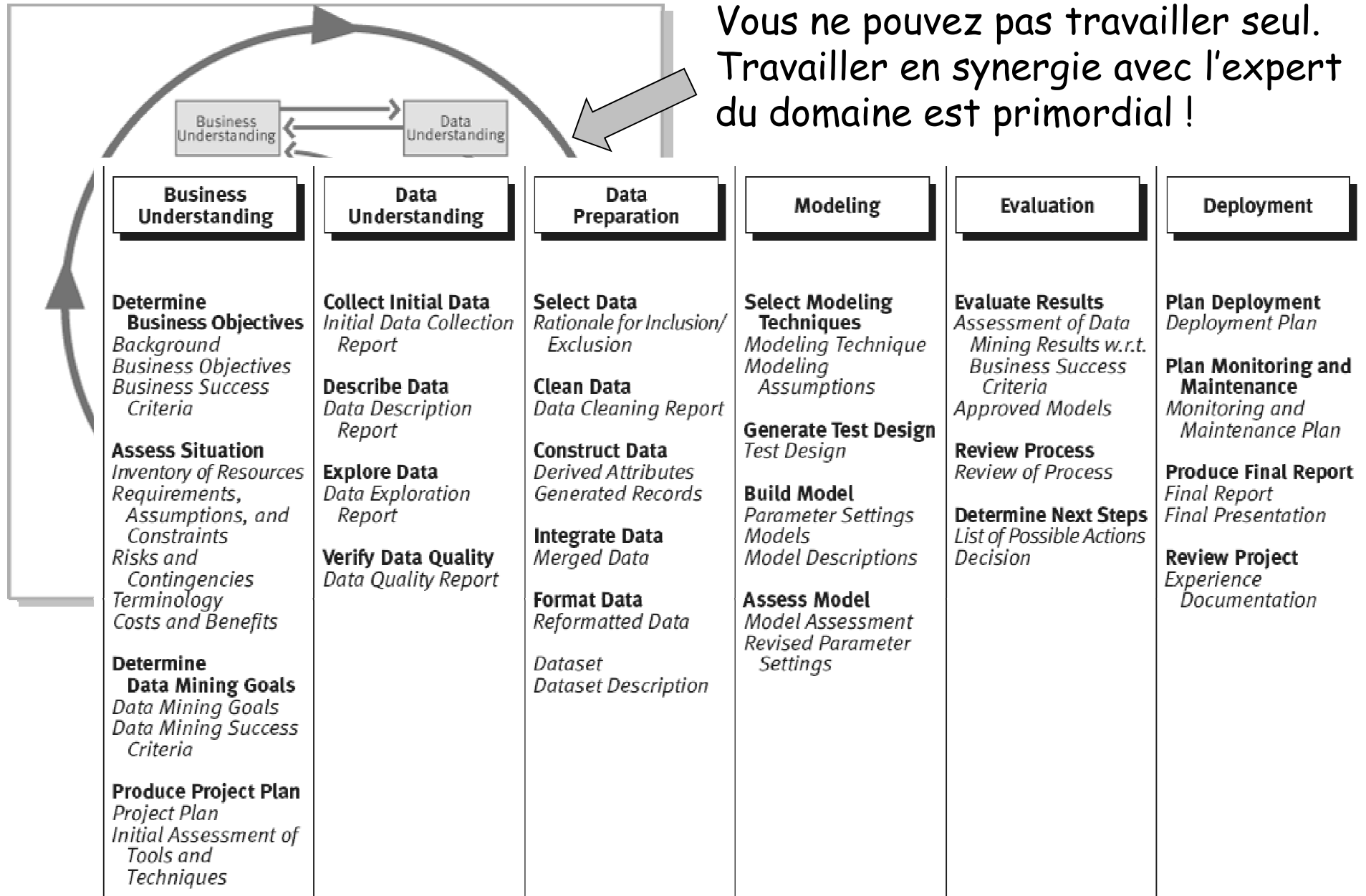


Définition : Processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données (Fayyad, 1996)



Source: CRISP-DM 1.0, Step-by-step Data Mining Guide, SPSS Publication

Vous ne pouvez pas travailler seul.
Travailler en synergie avec l'expert
du domaine est primordial !



Émergence de l'ECD : domaines d'applications

Domaine des assurances

- analyse des risques (caractérisation des clients à hauts risques, etc.)
- automatisation du traitement des demandes (diagnostic des dégâts et détermination automatique du montant des indemnités)

Services financiers

- consentements de prêts automatisés, support à la décision de crédit
- détection des fraudes

Grande distribution

- profils de consommateurs et modèles d'achats
- constitution des rayonnages
- marketing ciblé



Est-ce vraiment nouveau ?

Définition :

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>

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



Spécificités du Data Mining ?

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



Spécificités du Data Mining

Sources de données

- valoriser les fichiers de l'entreprise
- construire des entrepôts
- modifier le schéma organisationnel

Techniques utilisées

- Intégrer des techniques d'origines diverses

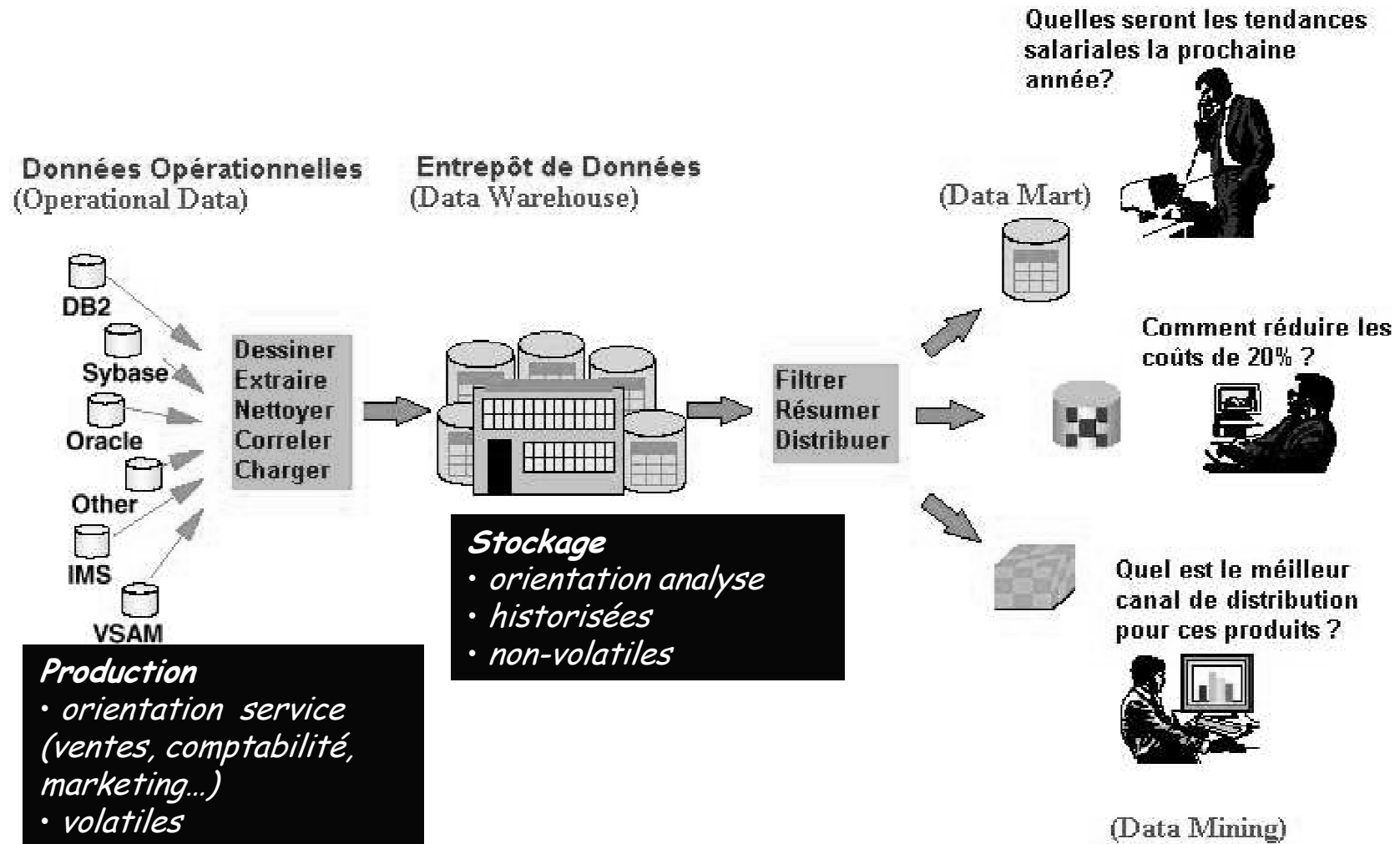
Élargissement des supports

- Text mining
- Image mining
- ... Multimédia mining

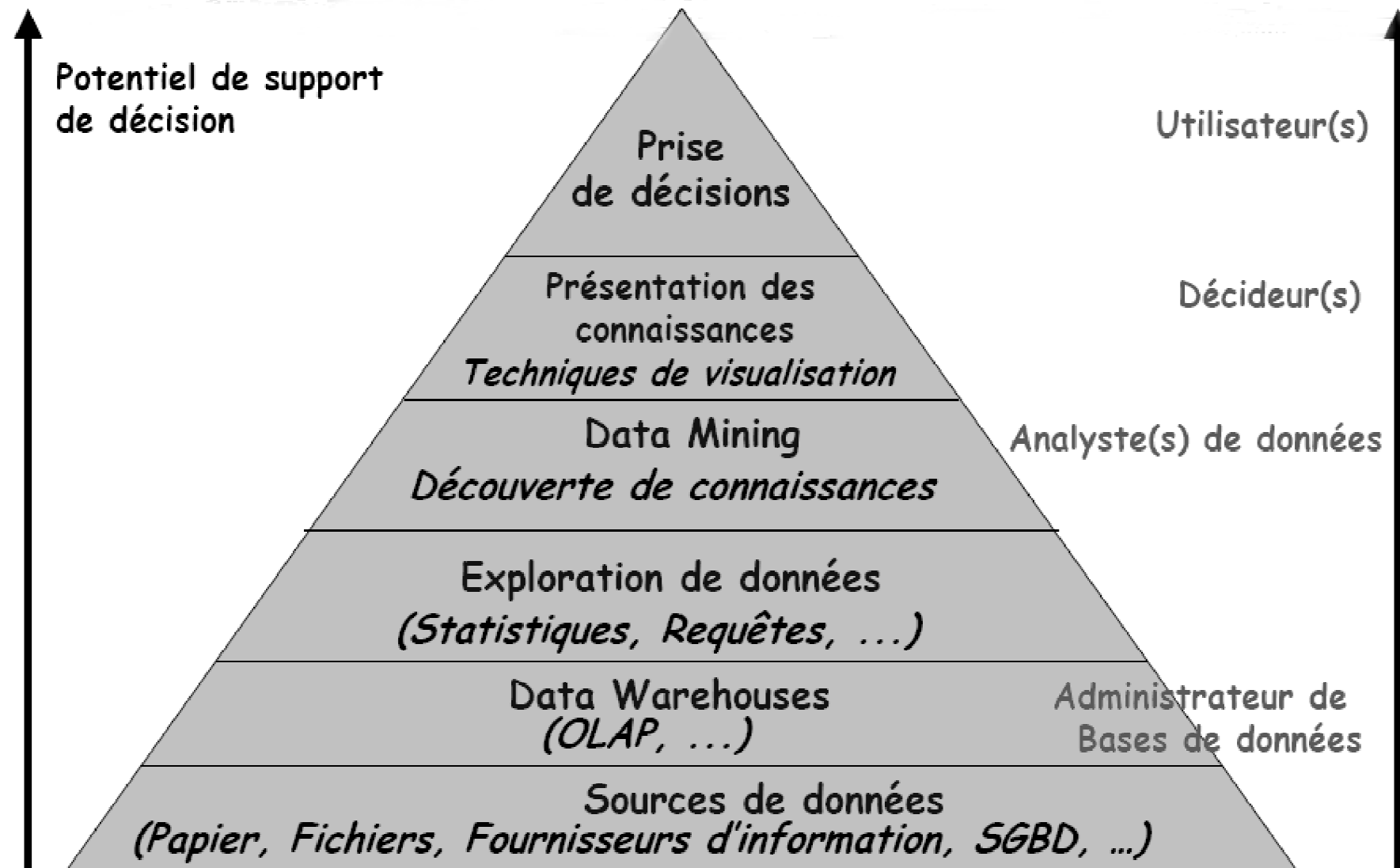


Les sources de données

Construire une Infrastructure d'Information Intelligente pour l'Entreprise



L'organisation du flux d'informations et les acteurs



Systèmes de gestion et systèmes décisionnels

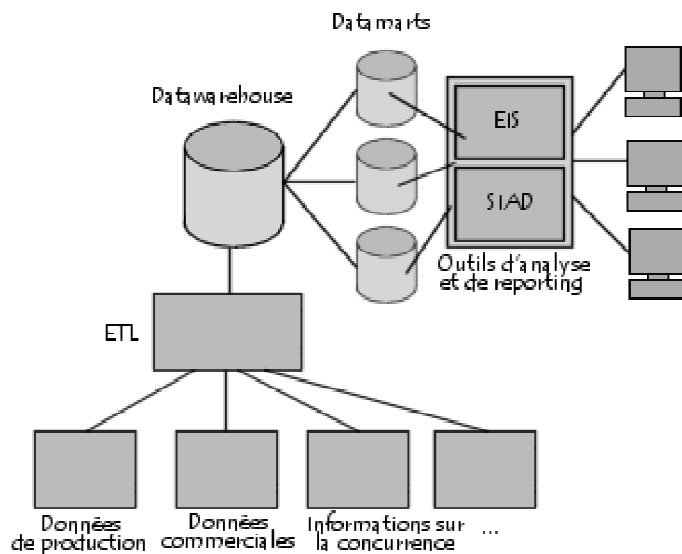
	Systèmes de gestion (opérationnel)	Systèmes décisionnels (analyse)
Objectif	dédié au métier et à la production ex: facturation, stock, personnel	dédié au management de l'entreprise (pilotage et prise de décision)
Volatilité (perennité)	données volatiles ex: le prix d'un produit évolue dans le temps	données historisées ex: garder la trace des évolutions des prix, introduction d'une information 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



Data Mining vs. Informatique Décisionnelle (Business Intelligence)

L'**informatique décisionnelle** (... BI pour *Business Intelligence*) désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données d'une entreprise en vue d'offrir une aide à la décision et de permettre aux responsables de la stratégie d'une entreprise d'avoir une vue d'ensemble de l'activité traitée.

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



- Sélectionner les données (par rapport à un sujet et/ou une période)
- Trier, regrouper ou répartir ces données selon certains critères
- Élaborer des calculs récapitulatifs « simples » (totaux, 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 est proche de ce cadre, mais elle introduit une dimension supplémentaire qui est la **modélisation « exploratoire »** (détection des liens de cause à effet, validation de leur reproductibilité) **!**



Spécificités du Data Mining

Techniques d'exploration de données

- Des techniques d'origines diverses, issues de cultures différentes
- ...mais qui traitent des problèmes similaires
- et qui partent toujours d'un tableau de données



Techniques utilisées selon leur « origine »

Statistiques

Théorie de l'estimation, tests
Économétrie

*Maximum de vraisemblance et moindres carrés
Régression logistique, ...*

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)

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, ...

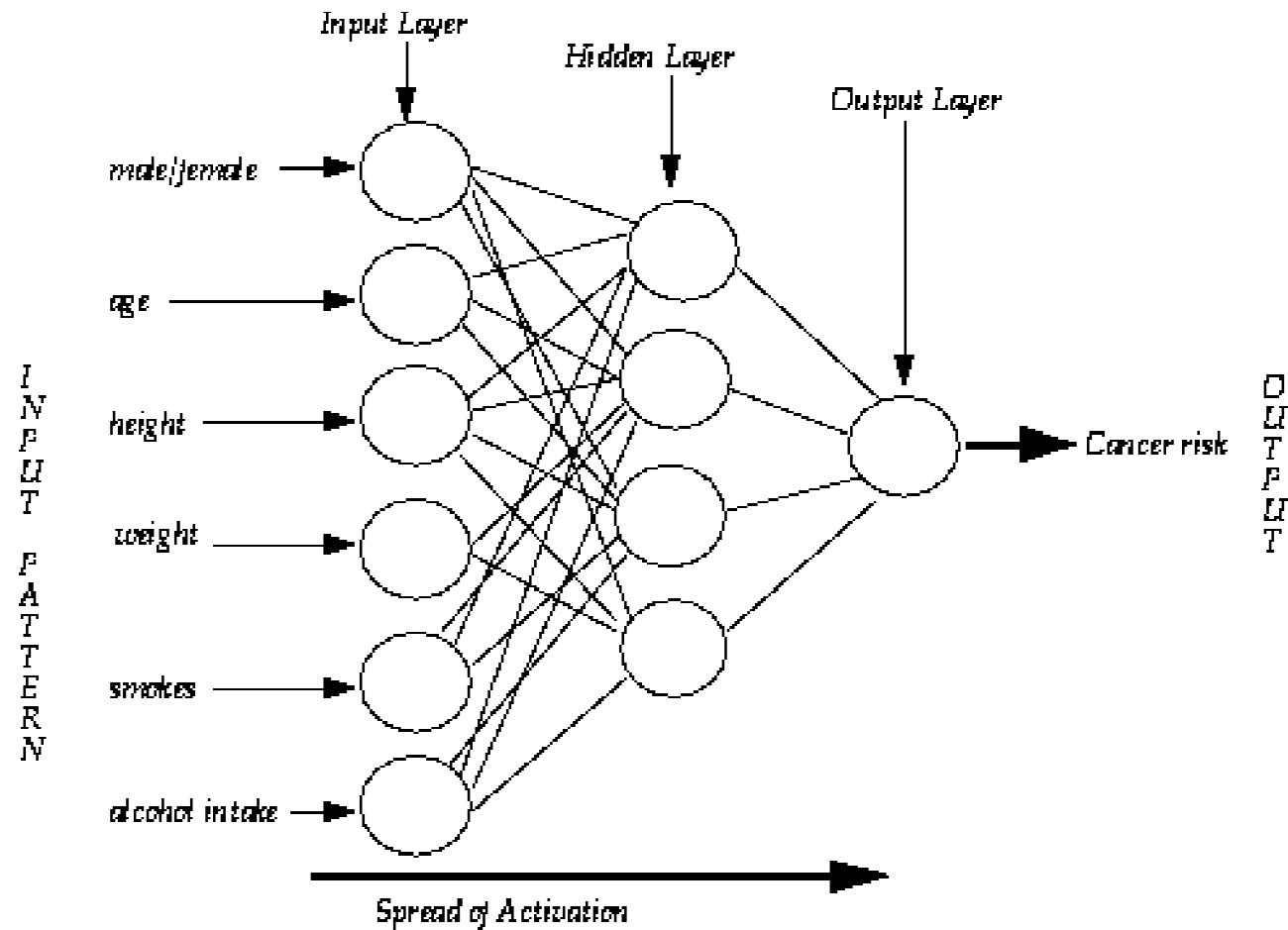


Très souvent, ces méthodes reviennent à optimiser les mêmes critères, mais avec des approches / formulations différentes



Techniques issues de l'Intelligence Artificielle

Les réseaux de neurones artificiels



- capacité d'apprentissage (universel)
- 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 type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	TITLE	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	SPOUSE_TIT	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	GUARANTEE	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	INSURANCE	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	HOUSING	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	HOUSING_TY	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9	JOB	Quality	<input type="checkbox"/>	<input type="checkbox"/>	<input 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_LENGT = 4.00**
Then
GUARANTEE *is* **No**
Rule's probability: 0.966
The rule exists in 28 records.

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

- traitement « omnibus »
- connaissance interprétable



Spécificités du Data Mining

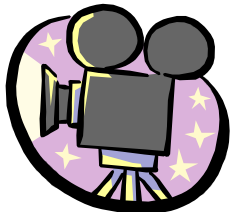
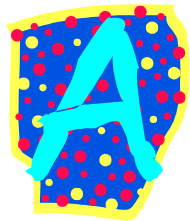
Élargissement des supports

- Text mining
- Image mining
- ...autres...

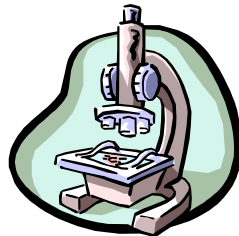
L'appréhension des sources multiples



Élargir les supports



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

Filtrage automatique des e-mails (spams, terrorisme,...)
Reconnaissance de la langue à une centrale téléphonique
Détection des images pornographiques sur le web
Analyse des mammographies
Etc.



DEFINITION

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

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.



Data Mining vs. Big Data (2/3) – Les 3 « V »

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, etc.).
Il faut pouvoir les (données) traiter !

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 simultanément !

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 rapidement !



Data Mining vs. Big Data (3/3)

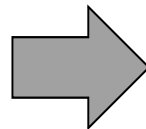
Big data vs. BI ?

(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... ;

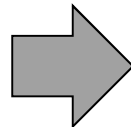
→ Big Data : utilisation de statistique inférentielle, sur des données à faible densité en information dont le grand volume permet d'inférer des lois (régressions...) donnant dès lors (avec les limites de l'inférence) au big data des capacités prédictives [modélisation, analyse prédictive,...].

Mais alors...



BIG DATA = DATA MINING ++

Avec de nouveaux défis technologiques / méthodologiques liés aux 3 « V »



- Cloud computing (ex. APACHE HADOOP / MAHOUT)
- Fouille de données complexes
- Data stream mining
- Etc.



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

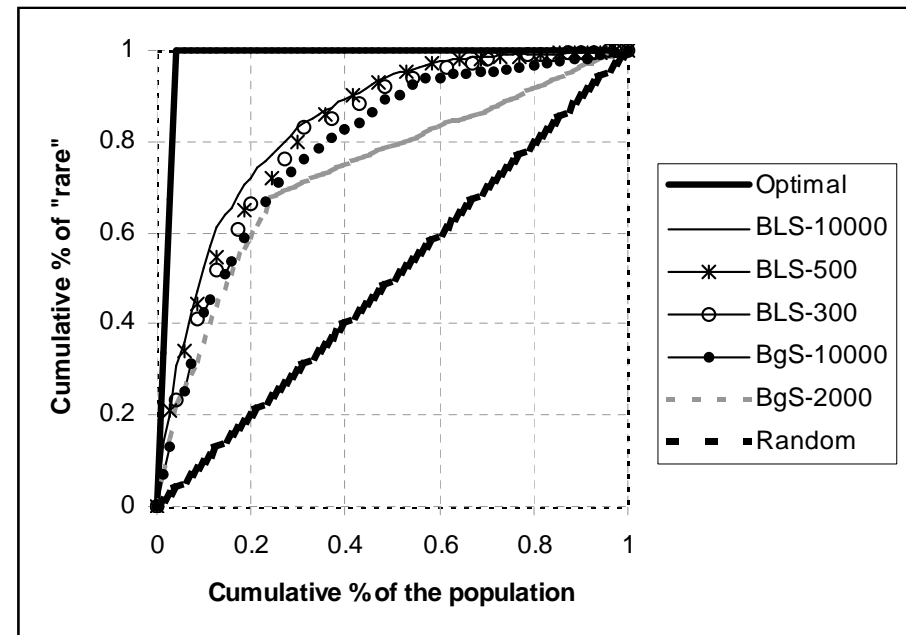
- les non-abonnés dans ces groupes seront (certainement ?) sensibles à une offre ciblée (hypothèse : s'ils ne sont pas abonnés, c'est qu'ils n'ont pas reçu l'information)
- technique : arbre de décision avec échantillonnage équilibré sur chaque noeud



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 shows a Reuters database interface with a table of news items and two preview windows. The table has columns: IDTEXTE, TEXTE, SEQ, CORN, CRUDE, and TRADE. The preview windows show the full text of selected news items.

IDTEXTE	TEXTE	SEQ	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	...

TEKSTE
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 - -
Ill. Track - - 3.16 - -
Toledo - 3.04 2.98 2.90 -

TEKSTE
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	0
2	1	1	1	0	1	1	0	1	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



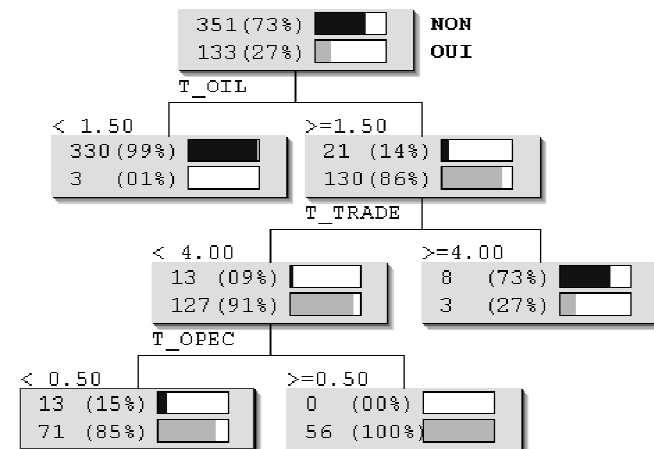
Text Mining – Catégorisation de nouvelles (3/3)

Visualiser les données

	IDTEXTE	TEXTE	ACQ	CORN	CRUDE	TRADE	T_OIL	T_CRUDE	T_BARRE	T_BARRE	T_OPEC	T_BPD	T_PETRO	T_PRICE	T_ENERG	T_GAS	T_EXPL
1	45	{MEMO}...	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
2	97	{MEMO}...	NON	OUI	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	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	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	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	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	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	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	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	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	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	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

58 Attributs 484 Individus Lecture seule

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



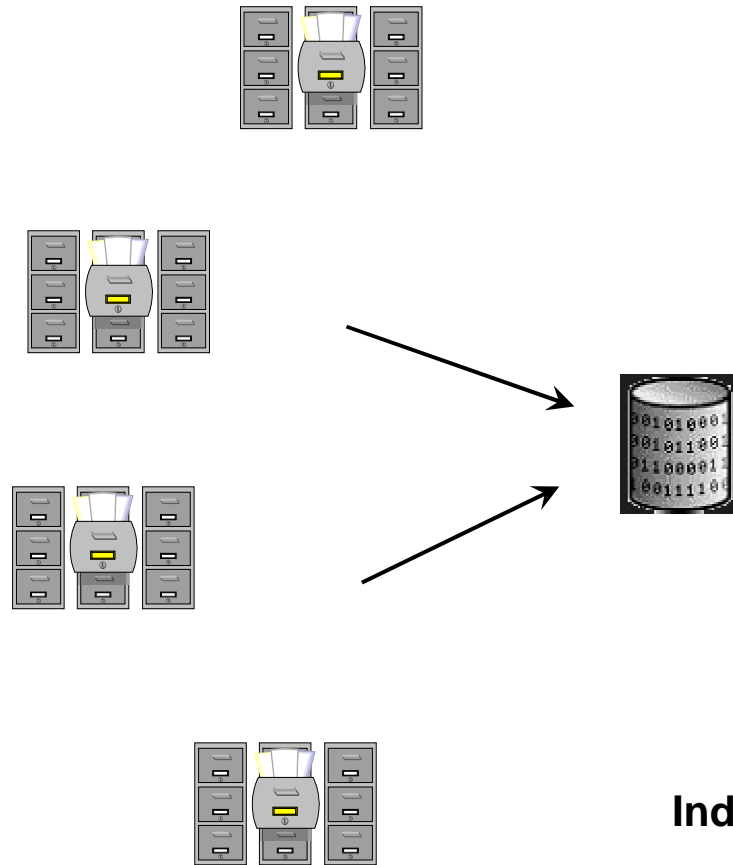
Typologies des méthodes

Quelle méthode utiliser par rapport :

- aux objectifs de l'étude ?
- aux données disponibles ?



Tableau de données



**Variables, caractères, attributs,
Descripteurs, champs, etc.**

Success	Wages	Job	Refunding
Y	0	Unemployed	Slow
N	2000	Skilled Worker	Slow
N	1400	Worker	Slow
N	1573	Retired	Slow
Y	2776	Skilled Worker	Slow
N	2439	Retired	Fast
N	862	Office employee	Slow
Y	1400	Salesman	Slow
N	1700	Skilled Worker	Slow
Y	785	Employee	Fast
Y	1274	Worker	Slow
N	960	Employee	Fast
N	1656	Worker	Fast
N	0	Unemployed	Slow

Individus, observations, objets, enregistrements, etc.



Types de variables

→ données nominales (ex. success, job...)

- nombre de cas dénombrables
- codés pour distinguer les *modalités*
- aucune relation d'ordre entre les codes
- opérateurs arithmétiques/mathématiques inapplicables

→ données ordinales (ex. Refunding...)

- nombre de cas dénombrables
- codés pour distinguer les modalités
- il existe une relation d'ordre entre les modalités
- les écarts ne sont pas quantifiables
- codés sous forme de rangs, on peut appliquer des calculs

→ données numériques ou continues (ex. Wages...)

- nombre de cas *théoriquement* infini
- il existe une relation d'ordre entre les valeurs
- les écarts sont quantifiables
- distinction entre échelle proportionnelle et non-proportionnelle
(ex. $20^{\circ}\text{C}/10^{\circ}\text{C} = 2$ et $68^{\circ}\text{F}/50^{\circ}\text{F} = 1.6$: non proportionnelle ; kg et livres : proportionnelle)
- calculs autorisés, algébriques



Distinguer les types de variables

On peut distinguer les différents types de données à partir de la définition de l'opérateur différence :

Nominale :
$$d_{AB} = \begin{cases} 0, & \text{si } x_a = x_b \\ 1, & \text{si } x_a \neq x_b \end{cases}$$

Ordinale :
$$d_{AB} = \begin{cases} +1, & \text{si } x_a > x_b \\ 0, & \text{si } x_a = x_b \\ -1, & \text{si } x_a < x_b \end{cases}$$

Continue :
$$d_{AB} = x_a - x_b$$




Qualitatives vers continues

Données qualitatives (nominales, ordinales) → Données continues

☞ *Codage disjonctif complet*

Refunding			
Fast			
Slow			
Fast			
Normal			
Slow			



Ref_Slow	Ref_Normal	Ref_Fast
0	0	1
1	0	0
0	0	1
0	1	0
1	0	0




! on perd l'information d'ordre sur les données ordinales

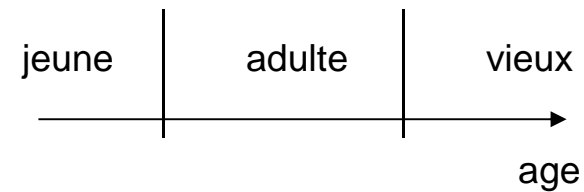
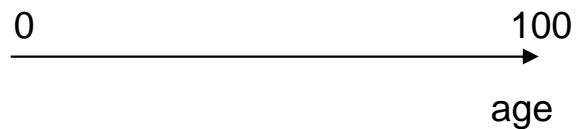


Continues vers ordinales

Données continues \longrightarrow Données ordinales

Discrétisation

-  par expert
-  automatique non-contextuelle
-  automatique contextuelle



- ! on perd l'information sur les écarts
- ! on peut traiter des relations non-linéaires



Continues vers continues

Données continues



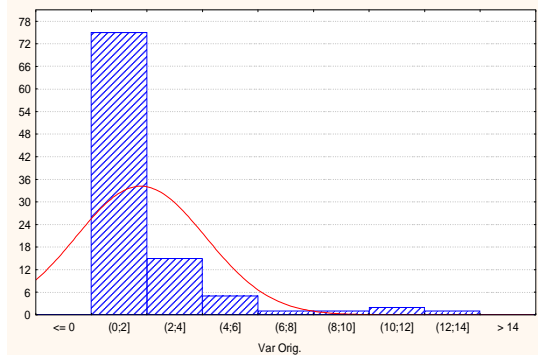
Données continues

➤ *Standardisation*

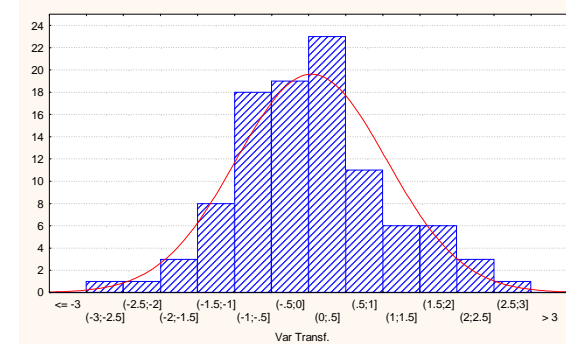
➤ centrage ex : taille = 2m20, taille = 0m50 au dessus de la moyenne

➤ réduction ex : taille = 0m50 ou taille = 50cm au dessus de la moyenne

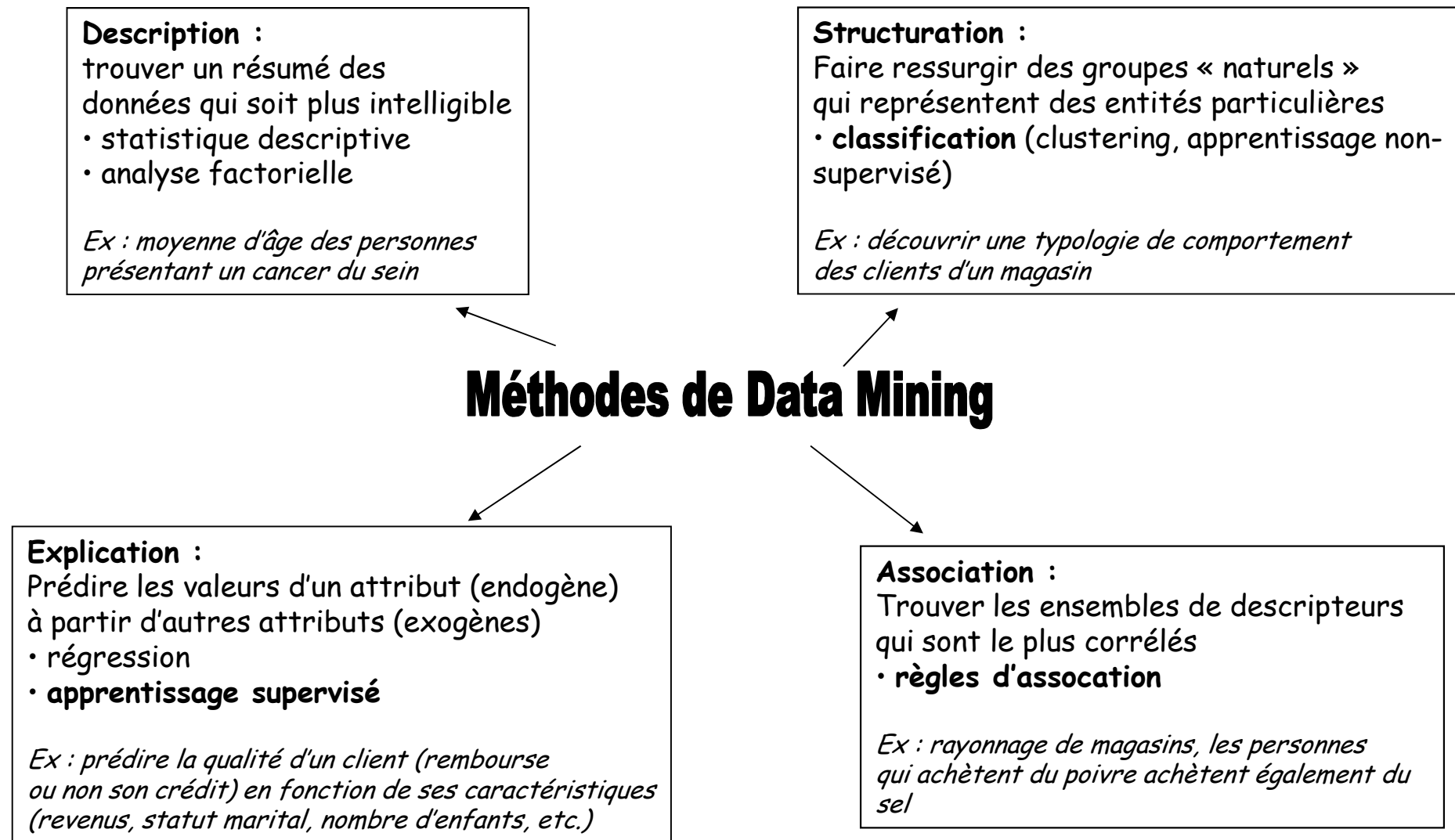
➤ *Transformation distributionnelle*



$$x_2 = \ln(x_1)$$



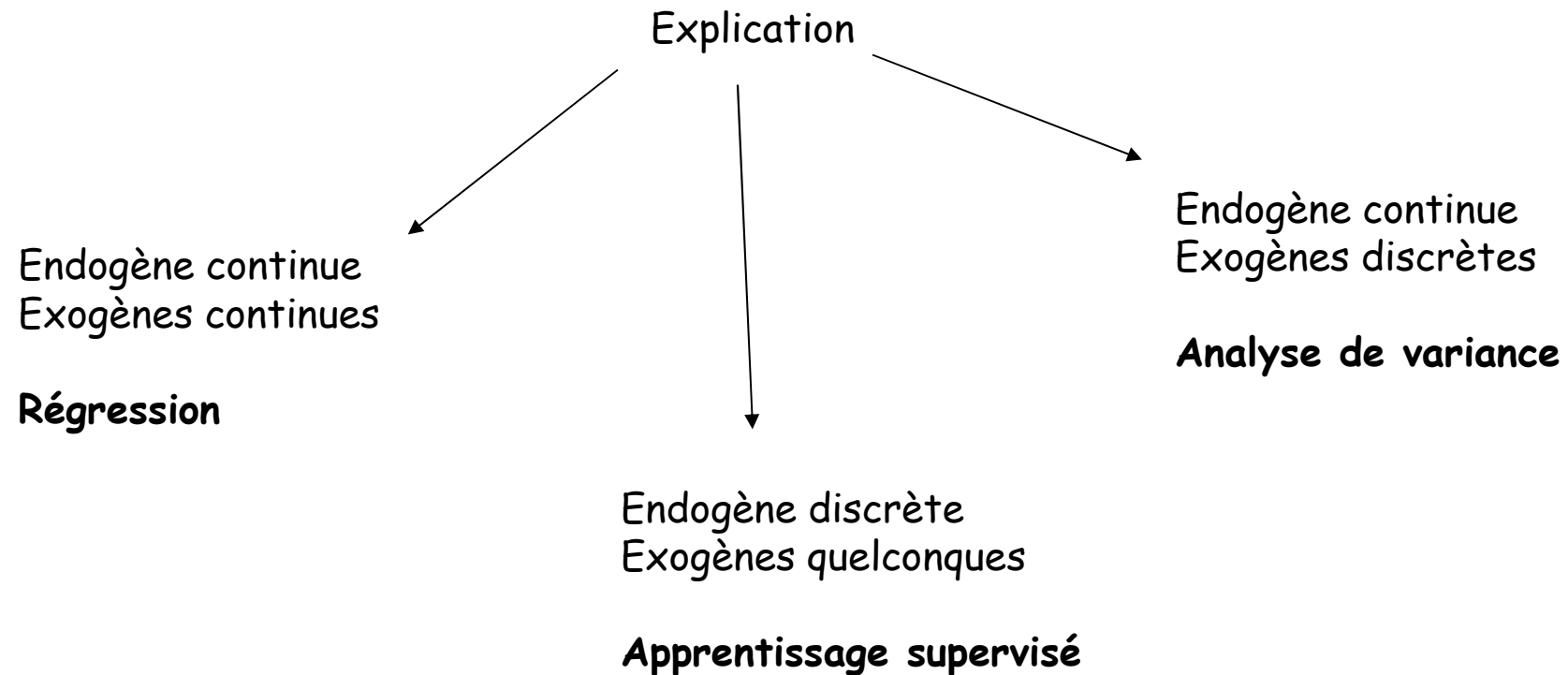
Typologie des méthodes selon les objectifs



les méthodes sont le plus souvent complémentaires !



Sous-typologie selon le type de données : la prédiction / explication



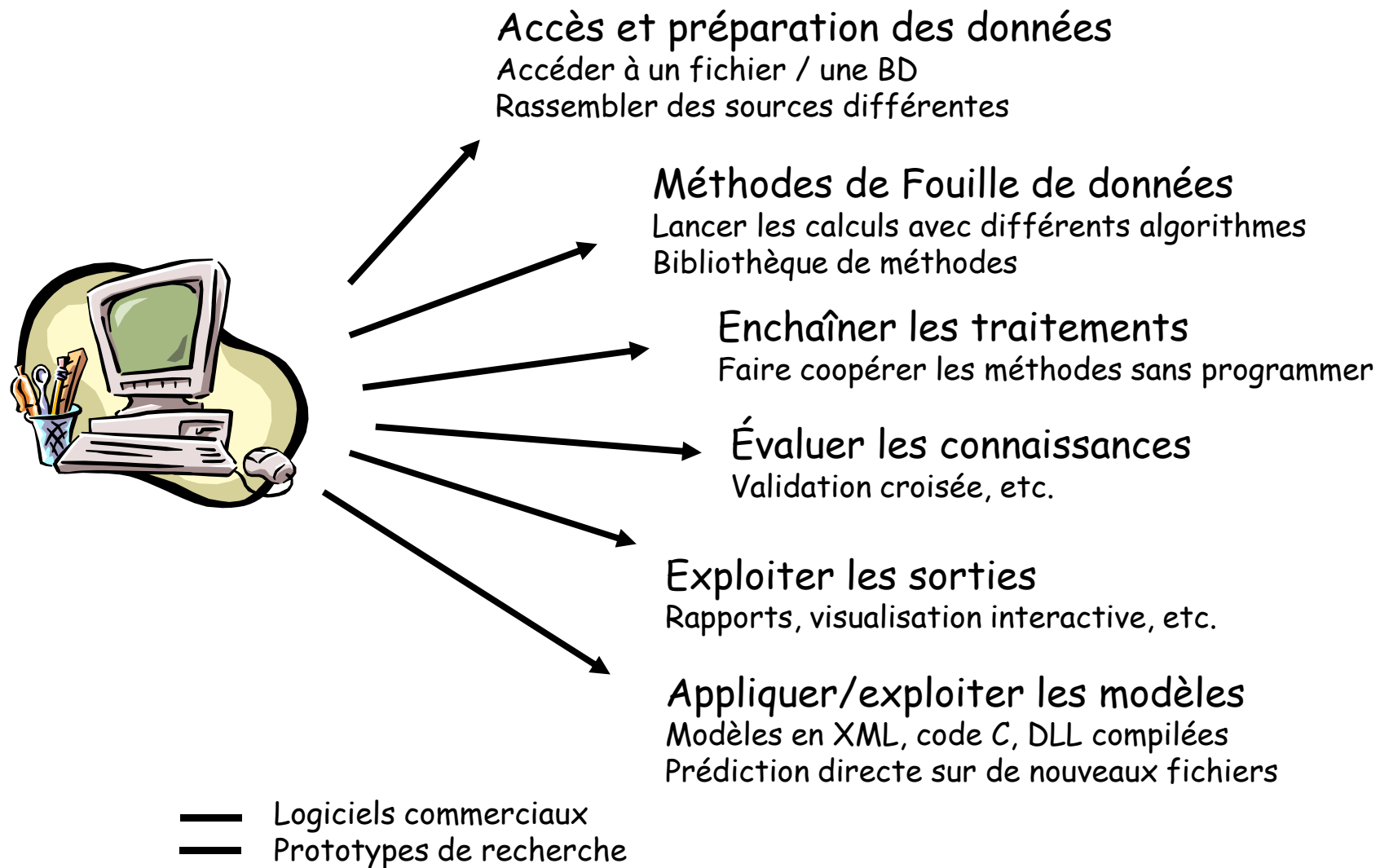
Ressources

(1) Logiciels

(2) Ouvrages et ressources en ligne



Logiciels de DATA MINING – Fonctionnalités



Logiciels de DATA MINING – Les logiciels disponibles

Commerciaux

SPAD

SAS Enterprise miner

SPSS Clementine

STATISTICA Data Miner

IBM Intelligent Miner

RAPIDMINER (*)

KNIME (*)

- Simplicité du pilotage (*filère - diagramme*)
- Techniques variées
- Déploiement
- Outils de « reporting »

Universitaires

R (*)

TANAGRA

SIPINA v2.5 & Recherche

WEKA (*)


ORANGE

- Spécifique à certaines techniques
- Techniques référencées - publiées
- Outils de validation



Conclusion

La démarche DATA MINING

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

Pas de miracle si :

Les objectifs sont mal définis

Les données disponibles ne conviennent pas

Les données sont mal « préparées »

On n'utilise pas les techniques appropriées



Bibliographie : pratique du Data Mining

- « Le Data mining », R. Lefebure et G. Venturi, ed. Eyrolles, 2001.
Peu technique, point de vue général, très bon recul, complet
- « Data Mining et statistique décisionnelle », S. Tufféry, ed. technique, 2006.
Plutôt guide pratique : repères pour les projets, opportunités, méthodes
- « Analyse discriminante - Application au risque et au scoring financier », M. Bardos, ed. Dunod, 2001.
Technique pratique, avec de bons repères théoriques, tourné vers les applications



Bibliographique : compréhension des méthodes

- « Data Mining : Practical machine learning tools and techniques with Java implementations », I. Witten and E. Frank, Morgan Kaufman Pub., 2000.
Très général et complet, logiciel libre accès, technique
- « The elements of statistical learning - Data Mining, Inference and Prediction », T. Hastie, R. Tibshirani, J. Friedman, Springer 2001.
Très technique, encyclopédique, indispensable pour la recherche, à lire plusieurs fois
- « Machine Learning », T. Mitchell, Mc Graw-Hill Editions, 1997.
Très très technique, surtout méthodes supervisées, encyclopédique



Ressources en ligne

Sites web et portails :

- <http://chirouble.univ-lyon2.fr/~ricco/data-mining>

Un portail pour la documentation : liens, supports de cours en ligne, logiciels, données

- Data Mining dicit Wikipédia : http://fr.wikipedia.org/wiki/Exploration_de_donn%C3%A9es

Site des tutoriels :

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

- <http://www.kdnuggets.com>

« Le » portail du DATA MINING, avec toute l'actualité du domaine

- Big data dicit SAS : <http://www.sas.com/big-data/>

