

Professeur D.A. ZIGHED
Directeur de l'Équipe de Recherche en Ingénierie des
Connaissances
Université Lumière Lyon 2

zighed@univ-lyon2.fr
<http://eric.univ-lyon2.fr>

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 1 -*

Organisation du cours

1. Lieux (salles de cours, TP,...)
2. Accès aux ressources : machines, bibliothèque, photocopie,...
3. Contacts enseignants, secrétariat,
4. Séminaires (ERIC, DEA ECD)

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 2 -*

Plan du cours

1. **Introduction à l'ECD**
2. **Apprentissage automatique**
3. **Les graphes d'induction**
4. **La discrétisation d'attributs**
5. **Analyse typologique**
6. **Les mesures d'association (Pr. G. Ritschard)**
7. **Recherche : Modèles et applications**

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 3 -*

Introduction à l'ECD

L'ingénierie des connaissances

Du stockage des masses de données à la
découverte des connaissances dans les bases
de données

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 4 -*

« C'est du volume des données dont elle dispose que notre espèce tire un sentiment immérité de supériorité alors que le véritable critère porte sur le degré auquel elle sait pétrir et maîtriser les informations dont elle dispose »

Goethe

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 5 -

Plan De L'exposé

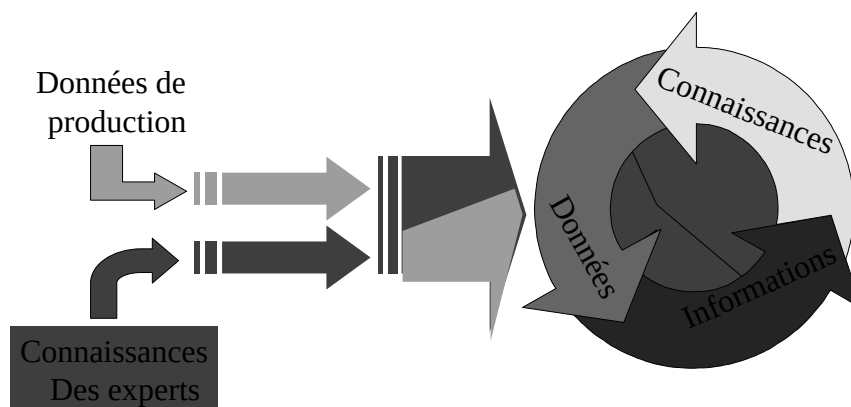
1. L'ingénierie des connaissances
2. L'Extraction des Connaissances à partir des Données
3. Le contexte d'émergence de l'ECD
4. Quelques exemples de problèmes d'ECD
5. Le processus d'ECD
6. Un panorama sur les techniques de fouille des données
7. Illustration par les graphes d'induction
8. Spécification d'un logiciel d'ECD
9. Quelques Infos et pointeurs pour approfondir
10. Les maladies de l'ECD

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 6 -

1. L'ingénierie des Connaissances

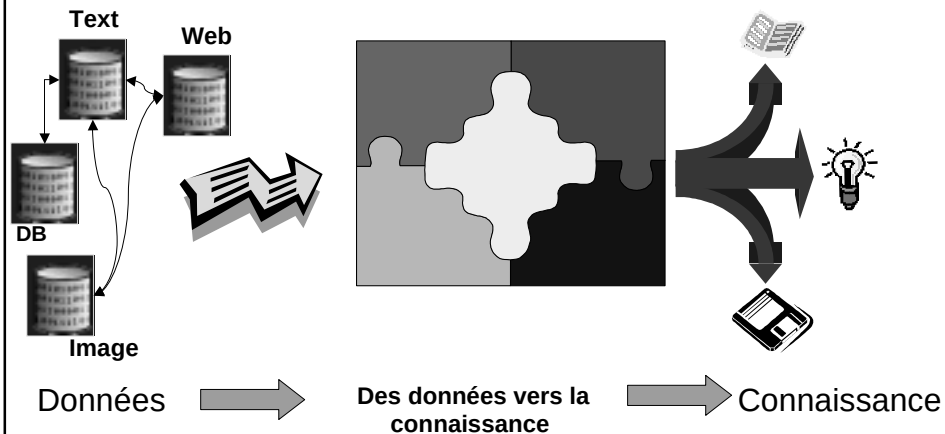
L'ingénierie des connaissances se propose d'offrir un ensemble de méthodes et de techniques informatiques permettant l'acquisition, le partage et l'exploitation des connaissances d'une organisation.

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 7 -



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 8 -

2. L'E.C.D. Knowledge Extraction Software SIPINA-PRO



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 9 -

3. Contexte d'émergence de l'E.C.D.

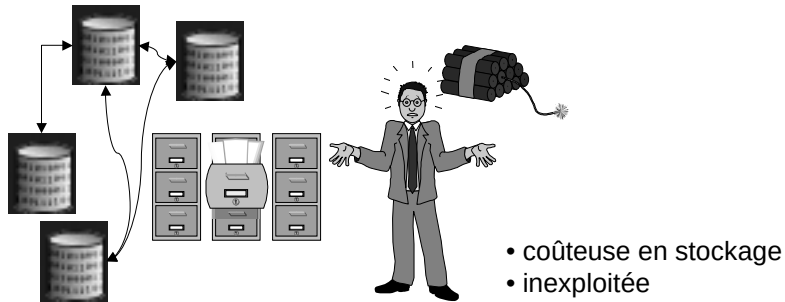
1. Exigences de l'environnement

- Environnement fortement concurrentiel
- Temps de réaction court
- Gain de productivité importants

☞ substitution d'une politique orientée produit
par une politique orientée client

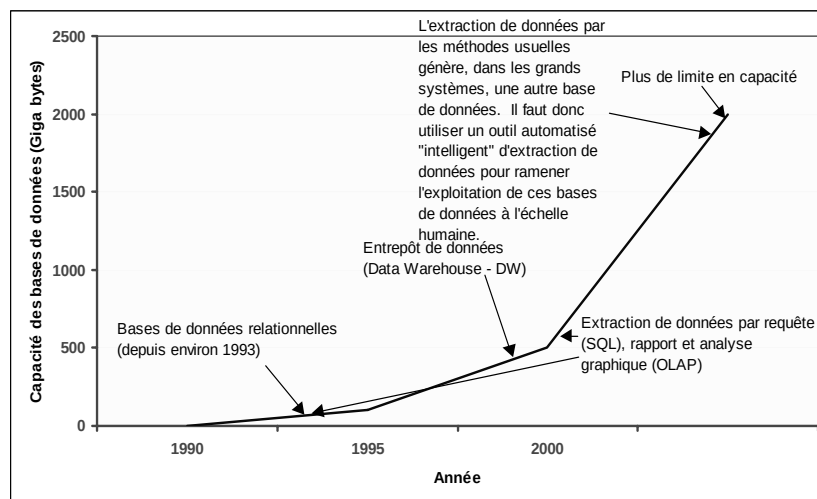
Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 10 -

2. L'état du Système d'information des organisations



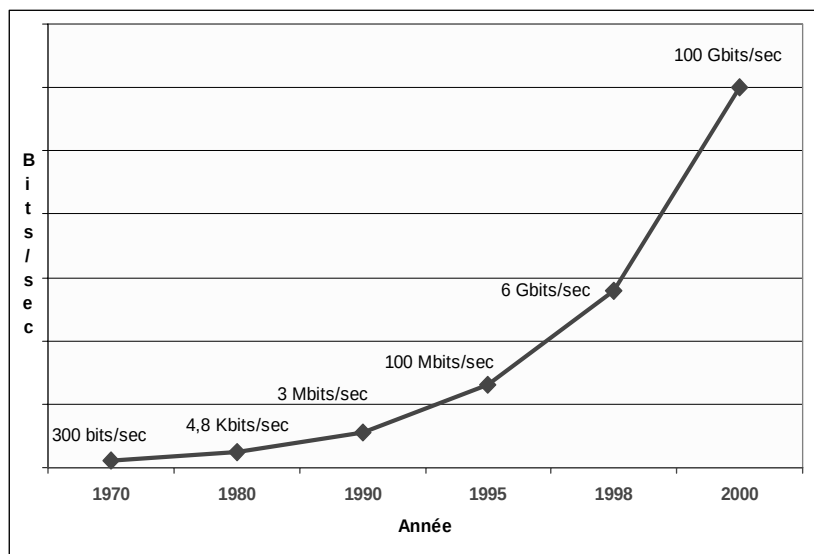
Comment et à quelles fins utiliser cette expérience accumulée ?

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 11 -



La taille des bases de données augmentent...
(META Group - Trends Technology, and Implementation Imperatives, 1999)

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 12 -



Les capacités de transfert réseaux s'améliorent...
(META Group - Trends in Technology, and Implementation Imperatives, 1999)

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 13 -*

4. Exemples de problèmes d'E.C.D.

- **Médecine** : Dépistage automatique du cancer du sein, diagnostic automatique des nodules thyroïdien froid,... Evaluation du risque infectieux chez le brûlé
- **Industrie** : Cimentation des puits de pétrole,
- **Banques** : Révélation des critères d'attribution de crédits aux particuliers,
- **Chimie** : Reconnaissance automatique des alcools,
- **Neuropsychologie** : Reconnaissance automatique des odeurs,
- **Archéologie** : Identification automatique des stèles funéraires
- **Biomécanique** : Reconnaissance automatique du type de boiterie humaine,
- **Urbanisme** : Mise à jour automatique de cadastre
- ...

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 14 -*

1. Médecine

Evolution clinique des grands brûlés

Hôpital Edouard Herriot Lyon, 1985

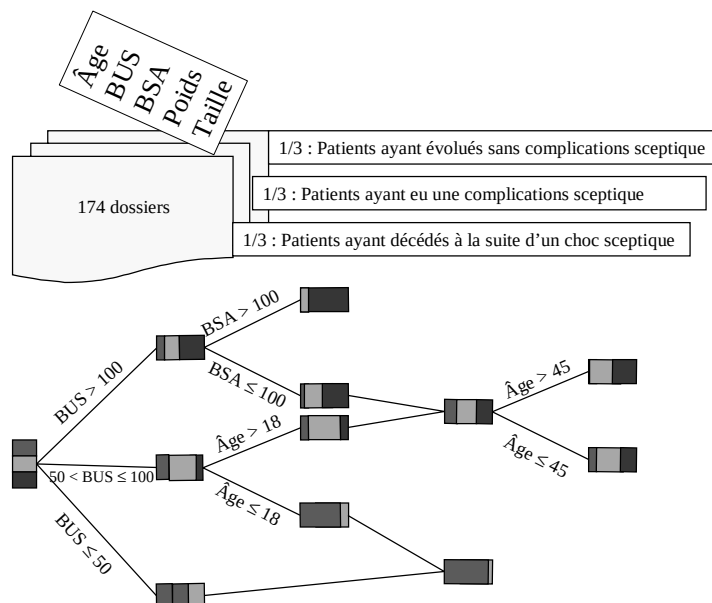
Problématique :

- Pathologie de crise
- Prise en charge lourde

Obectifs :

- Produire un modèle de pronostic fiable et intelligible
- Utiliser des paramètres faciles d'accès

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 15 -

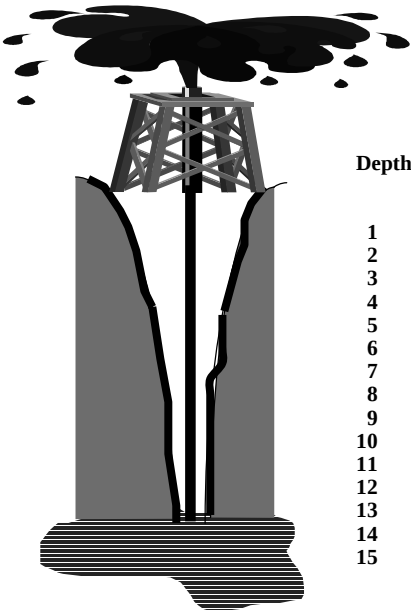


Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 16 -

2. Industrie

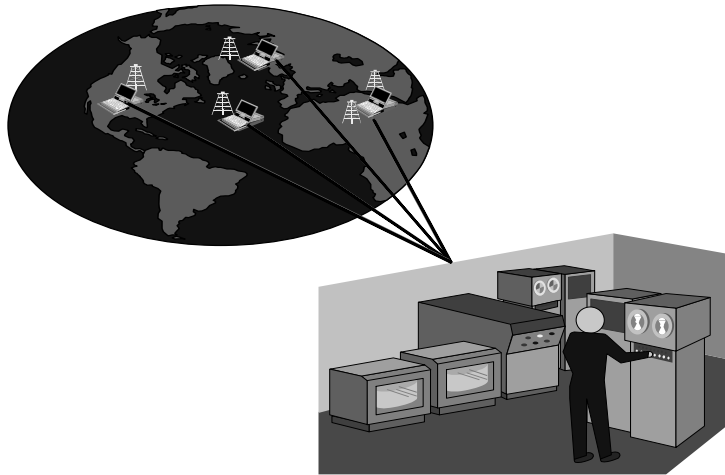


Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 17 -



Depth	Ciment.	SFH1	SSP1	Pt5
1	1	2	1	1
2	1	2	2	2
3	1	1	2	2
4	1	1	2	2
5	1	1	2	2
6	2	1	1	1
7	2	1	1	2
8	2	1	1	1
9	2	1	1	2
10	2	1	1	1
11	2	1	2	2
12	2	1	2	1
13	2	2	1	1
14	2	2	2	1
15	2	2	1	1

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 18 -



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 19 -

3. Urbanisme

Mise à jour automatique de cadastre



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 20 -

Objets à reconnaître

- Toits
- Routes
- arbres
- voitures
- Terrasses
- Terrain
-



Paramètres de texture : $X_1, X_2, \dots, X_p$

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 21 -

4. Texte Mining



Tri
Automatique
des messages

```
Return-Path: <oytcsywa@fsinet.fsid.cvut.cz>
Received: from relay15.mx.aol.com (relay15.mail.aol.com [172.31.106.74]) by
air13.mail.aol.com (v40.16) with SMTP; Sun, 22 Mar 1998 00:04:48
-0500
Received: from mail2.theultimatesexsite.com (lash4-16.flash.net
[209.30.75.16])
by relay15.mx.aol.com (8.8.5/8.8.5/AOL-4.0.0)
with SMTP id AAA22788;
Sun, 22 Mar 1998 00:04:12 -0500 (EST)
Date: Sun, 22 Mar 1998 00:04:12 -0500 (EST)
From: oytcsywa@fsinet.fsid.cvut.cz
Message-Id: <199803220504.AAA22788@relay15.mx.aol.com>
Subject: would you
Mime-Version: 1.0
Content-type: text/plain; charset=US-ASCII
Content-transfer-encoding: 7bit
```

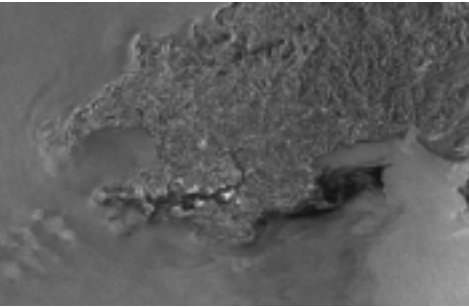
CONGRATULATIONS!!!

YOU HAVE JUST BEEN SELECTED TO RECEIVE A 100 PERCENT
FREE, FULL 7 DAYS MEMBERSHIP, TO THE NET'S ~~HOTTEST AND~~
~~LARGEST PREMIUM ADULT ENTERTAINMENT SITE!!~~

THAT'S RIGHT...TOTALLY 100 PERCENT FREE!

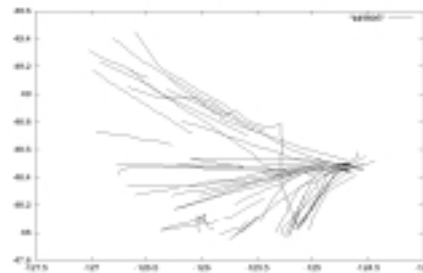
TO GET YOUR FREE MEMBERSHIP
CLICK HERE

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 22 -



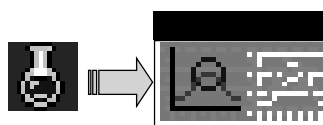
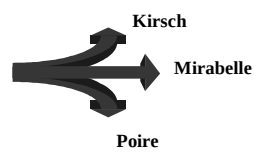
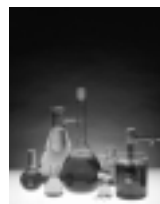
Identification du
type de trajectoire

Bateau de pêche,
Pétroliers,
...

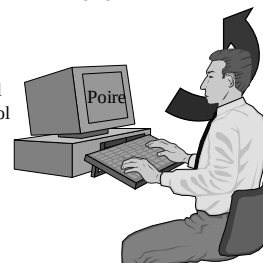


Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 23 -

Reconnaissance des alcools

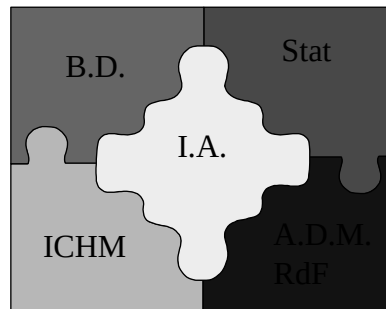


Buthanol
Acethanol
.
Propa



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 24 -

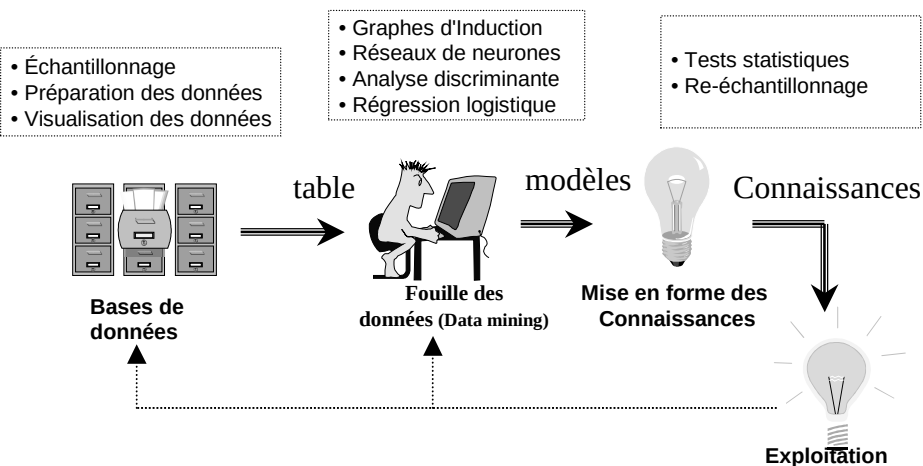
5. Le processus d'E.C.D.



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 25 -

Le processus ECD

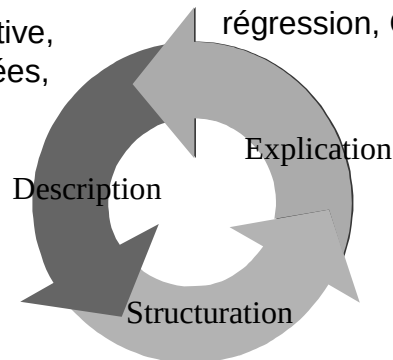
(Knowledge discovery in Databases)



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 26 -

6. Panorama sur les techniques de F.D.

Statistique descriptive,
Analyse des données,

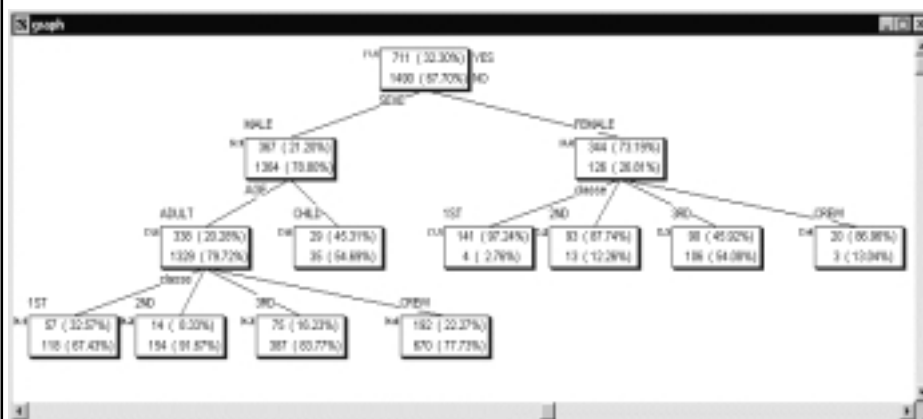


Modèles de prédiction :
régression, G.I., A.D., R.N.,...

Classification automatique,
Apprentissage non supervisé,...
CAH, NHD, RN, Autoclass, Cobweb

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 27 -

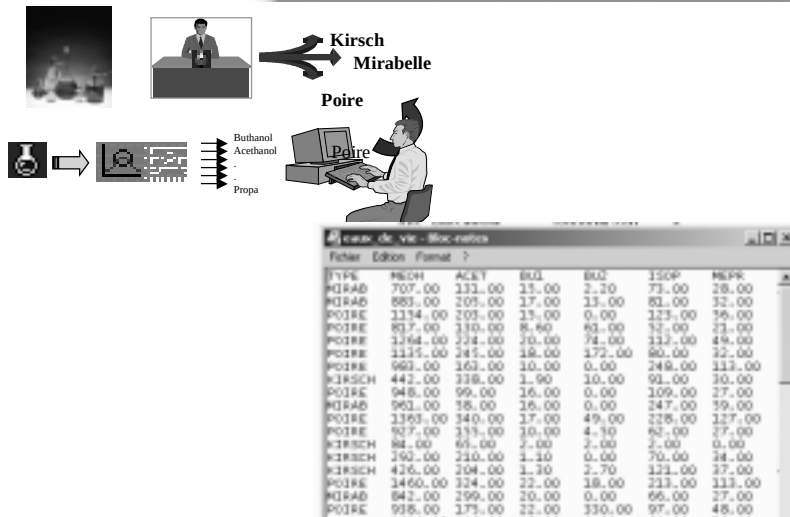
Méthode d'explication Graphes d'Induction



- rapide en apprentissage
- capacité à traiter de grandes bases
- possibilité de mise à jour
- classifieur « interprétable »

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 28 -

7. Illustration par les G.I.

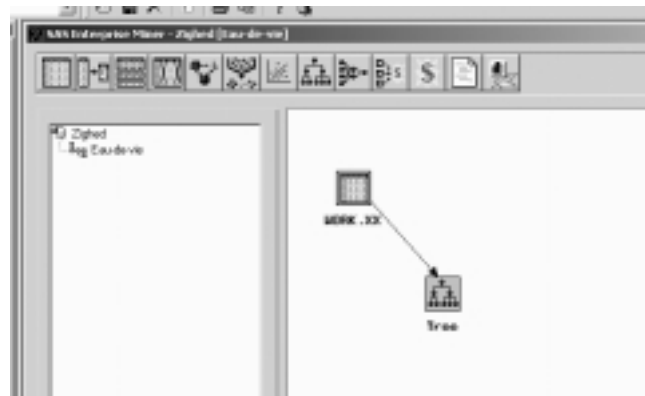


Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 29 -

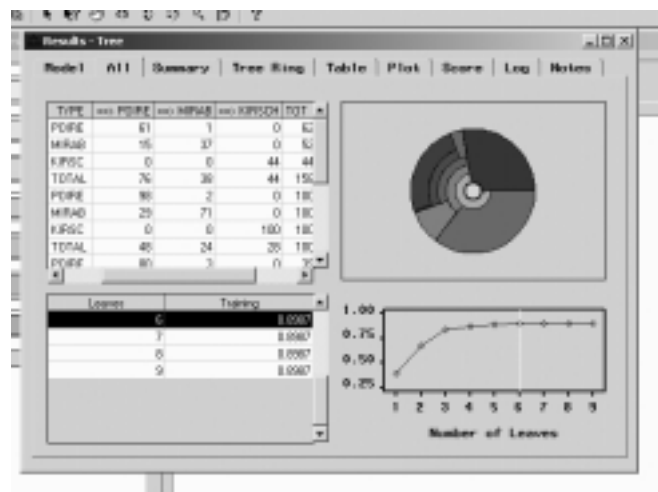
The screenshot shows the SAS software interface with the 'Input Data Source' window open. The window displays a list of variables with their names, roles, measurement types, and data types.

Name	Model Role	Measurement	Type	Pos
TYPE	Target	nominal	char	17
NECH	input	interval	num	BE
ACET	input	interval	num	BE
BUT	input	interval	num	BE
BUT	input	interval	num	BE
ISOP	input	interval	num	BE
NEPR	input	interval	num	BE
PROI	input	interval	num	BE
NECH	input	interval	num	BE

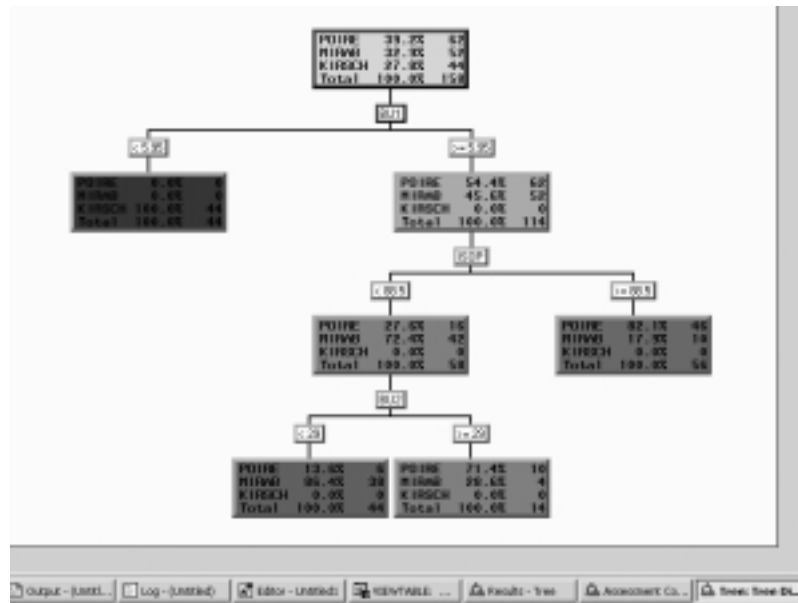
Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 30 -



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 31 -



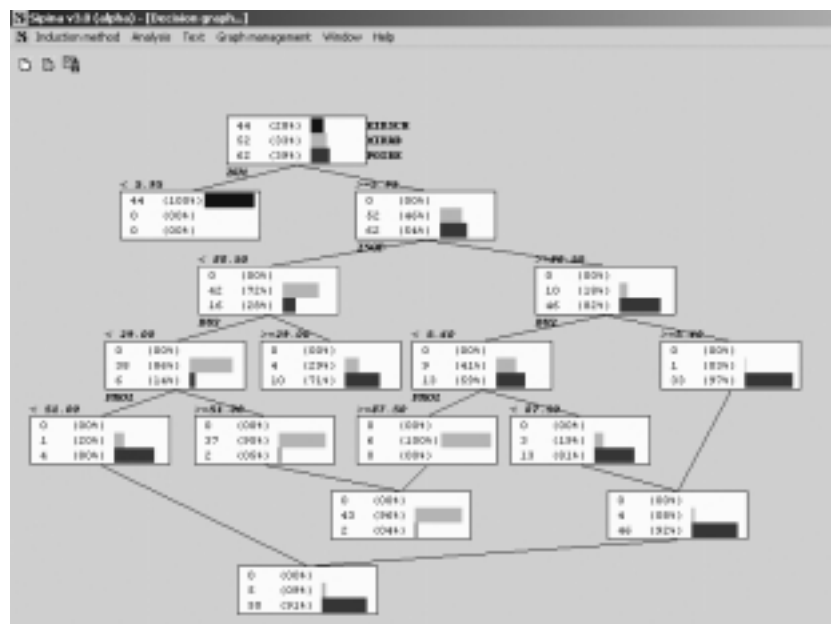
Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 32 -



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 33 -

Sapana v3.0 (Alpha)										
File Edit Data Statistics Induction/Method Analysis Text Window Help										
Learning set editor										
	TYPE	AGE	SEX	R1	R2	R3	R4	R5	R6	R7
76	MIRAD	767.00	131.80	15.00	2.00	73.80	28.00	131.80	9.80	
83	MIRAD	983.00	206.80	17.00	13.00	84.80	32.00	206.80	13.80	
97	POIPE	1154.80	200.80	15.00	8.00	123.00	56.00	50.80	14.80	
140	POIPE	877.00	130.80	0.00	61.00	52.80	25.00	212.80	0.80	
143	POIPE	1264.80	224.80	28.00	74.00	112.00	48.00	508.80	17.80	
151	POIPE	1135.80	245.80	18.00	172.80	90.80	32.00	910.80	21.80	
157	POIPE	983.00	160.80	18.00	8.00	248.00	113.00	36.80	6.80	
26	KIPECH	843.00	338.80	1.80	16.00	94.80	38.00	952.80	34.80	
103	POIPE	940.00	99.80	16.00	8.00	109.00	27.00	23.80	37.80	
83	MIRAD	961.00	96.80	16.00	8.00	247.00	58.00	86.80	6.80	
137	POIPE	1363.80	340.80	17.00	48.00	228.00	127.00	54.80	18.80	
171	POIPE	327.00	795.80	18.00	4.90	62.80	27.00	62.80	18.80	
12	KIPECH	84.00	45.80	2.80	2.00	2.80	0.80	288.80	6.80	
6	KIPECH	282.00	210.80	1.10	8.00	70.80	34.00	59.80	9.80	
14	KIPECH	626.00	208.80	1.30	2.70	121.00	37.00	432.80	9.40	
130	POIPE	1460.80	324.80	22.00	18.00	213.00	113.00	93.80	15.80	
60	MIRAD	842.00	298.80	28.00	8.00	86.80	27.00	52.80	11.80	
157	POIPE	938.00	175.80	22.00	330.80	97.80	48.00	281.00	12.80	
46	MIRAD	1890.80	706.80	18.00	8.00	60.80	26.00	99.80	32.80	
36	KIPECH	373.00	256.80	0.80	28.00	60.80	27.00	814.80	11.80	
50	MIRAD	1869.80	112.80	17.00	8.00	77.80	24.00	69.80	9.80	
110	POIPE	1121.80	164.80	16.00	7.80	72.80	32.00	50.80	17.80	
42	KIPECH	726.00	87.80	1.80	46.00	95.80	33.00	1407.00	26.80	
28	KIPECH	583.00	226.80	2.30	18.00	128.00	48.00	896.80	11.80	
21	KIPECH	0.80	26.80	0.10	8.00	0.80	6.80	253.80	7.80	
145	POIPE	620.00	123.80	28.00	128.80	95.80	26.00	536.80	18.80	
175	POIPE	1540.80	254.80	16.00	8.00	103.00	48.00	704.80	28.80	

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 34 -



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 35 -

9. Spécifications d'un logiciel d'ECD

Importation de fichiers, connexion bases de données...
Echantillonnage, Recodage, détection et traitement des données manquantes et/ou aberrantes,
Outils stat numériques, graphiques,...

Acquisition et préparation des données

- Graphes et Arbres de Décision
- Analyse discriminante
- Réseaux de Neurones
- Induction de règles
- et ensembles des techniques non sup.

Fouille des données

Graphes, règles
Distributions conditionnelles

Visualisation des Connaissances

Tests statistiques
Techniques de ré-échantillonnage
(Validation croisée, bootstrap...)

Validation des connaissances

- Sélection des règles
- Stratégie de votes

Généralisation



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 36 -

Les logiciels

☞ Commerciaux

SAS

SPSS / Clementine

Cognos

IBM Intelligent Miner

Angoss - Knowledge Studio

- Simplicité du pilotage (*wizard*)
- Outils de « reporting »

☞ Universitaire

SIPINA

dbMiner

IND & MLC++

- Richesse technique
- Techniques référencées
- Outils de validation

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 37 -*

10. Quelques pointeurs

Quelques ouvrages...

- **Advances in Knowledge Discovery and Data Mining**

Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy editors

(1996)

- **Le Data Mining**

René LEFEBURE, Gilles VENTURI

(Eyrolles, 1998)

- **Data Warehousing and Data Mining for Telecommunications**

Robert Matisson

(1998)

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 38 -*

Quelques sites web...

☞ <http://www.kdnuggets.com>

Site de référence

Liens vers : Logiciels, données, liens, ouvrages...

☞ <http://eric.univ-lyon2.fr>

Logiciel SIPINA en ligne et à télécharger, papiers...

☞ <http://clubs.yahoo.com/clubs/datamining>

Forums de discussion, mailing-list avec modérateur

☞ <http://www.cs.bham.ac.uk/~anp/papers.html>

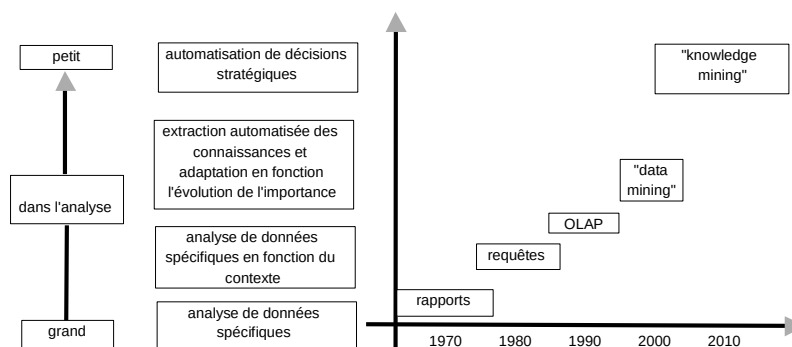
Collection d 'articles sur le Data Mining

☞ <http://bashful.qub.ac.uk/tec/courses/>

Séries de support de cours, y compris sur le Data Mining

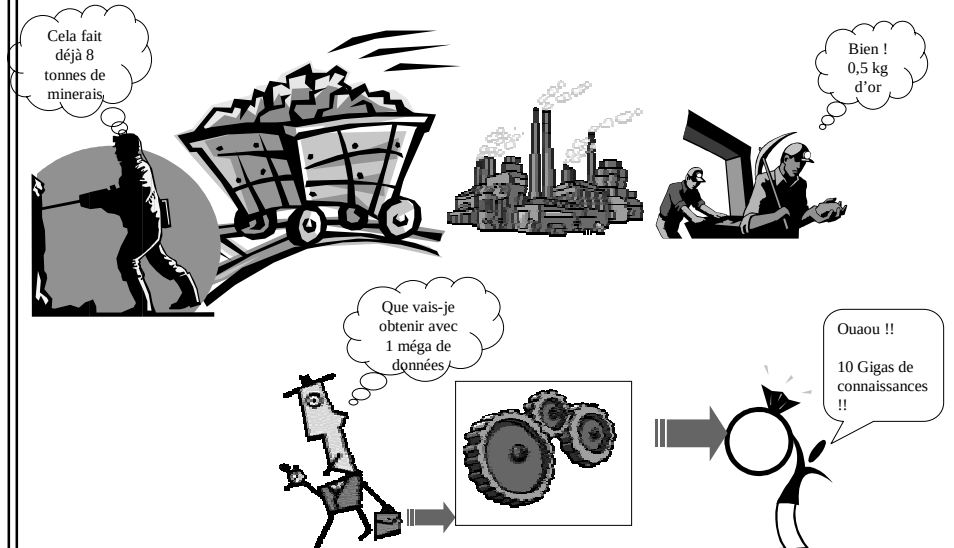
Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 39 -

8. Les maladies de l'ECD



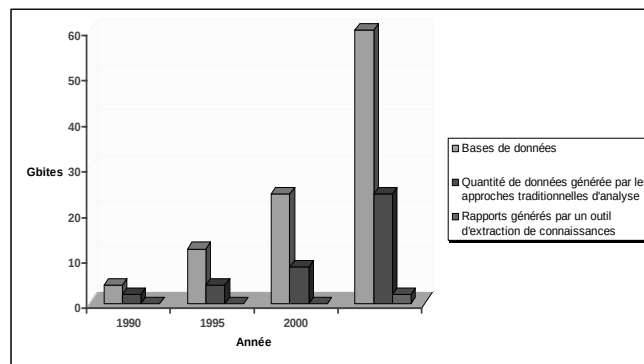
Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 40 -

Le Paradoxe du mineur



Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 41 -

Le syndrome de la boulimie



Faut-il traiter toutes les données ?

Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 42 -

Sujets compléments de formation

Chaque étudiant devra préparer un dossier sur l'un des sujets suivants. Le travail en binôme est autorisé et même recommandé

1. **Le CRM**
2. **La bio-informatique**
3. **Le Web mining**
4. **Les Télécoms**
5. **Le Knowledge Management**
6. **Observatoire des logiciels de DM**
7. **Accès aux bases documentaires**
8. **Data mining et BD chimie**
9. **Data mining dans les séquences vidéo**
10. **Modèles géométriques et SVM**
11. **Agrégation de prédicteurs**
12. **Règles d'association et analyse typologique**

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 43 -*

Professeur D.A. ZIGHED

Directeur de l'Équipe de Recherche en Ingénierie des
Connaissances
Université Lumière Lyon 2

zighed@univ-lyon2.fr

<http://eric.univ-lyon2.fr>

Derniers livres :

Graphes d'induction, apprentissage et data mining
D.A. Zighed & R. Rakotomalala
Hermès 2000,

Principles of Data Mining and Knowledge Discovery
Lecture Notes in Artificial Intelligence, 1910
D.A.Zighed, J. Komorowski, J. Zytkow (Eds.)
Springer Verlag, 2000

*Pr. D. A. ZIGHED, laboratoire ERIC, Université Lyon 2
DEA Extraction des Connaissances à partir des Données, support de cours - 44 -*