Outils pour le Data Mining

Comment Choisir un outil ?

Systèmes commerciaux de data mining possèdent peu de propriétés communes :

- Différentes méthodologies et fonctionalités de data mining
- Différents types d'ensembles de données

Pour la sélection d'un outil, on a besoin d'une analyse multi-critère des systèmes existants

Critères de choix d'un logiciel

Variété des algorithmes de data mining, de statistique et de préparation des données

- + simple d'avoir tout dans un seul outil

Qualité des algorithmes implémentés

- documentation éditeur pas toujours accessible

Capacité à traiter de grands volumes de données

- peut être cruciale à partir de plusieurs centaines de milliers d'individus à traiter

Types de données gérés

- exemple : choix influencé si l'entreprise possède déjà un infocentre SAS…

Existence d'un langage de programmation évolué

- Convivialité du logiciel et facilités à produire des rapports

Prix !

Ce que l'on peut attendre d'un logiciel

Algorithmes de statistique et de data mining :

- classement (analyse discriminante linéaire, régression logistique binaire ou polytomique, modèle linéaire généralisé, régression logistique PLS, arbres de décision, réseaux de neurones, k-plus proches voisins...)
- prédiction (régression linéaire, modèle linéaire général, régression robuste, régression non-linéaire, régression PLS, arbres de décision, réseaux de neurones, + proches voisins…)
- Clustering (centres mobiles, nuées dynamiques, k-means, classification hiérarchique, méthode mixte, réseaux de Kohonen…)
- analyse des séries temporelles
- détection des associations

Ce que l'on peut attendre d'un logiciel

Fonctions de préparation des données

- manipulation de fichiers (fusion, agrégation, transposition…)
- visualisation des individus, coloriage selon critère
- détection, filtrage et winsorisation des extrêmes
- analyse et imputation des valeurs manquantes
- transformation de variables (recodage, standardisation, normalisation automatique, discrétisation…)
- création de nouvelles variables (fonctions logiques, chaînes, statistiques, mathématiques…)
- sélection des discrétisations, des interactions et des
- variables les plus explicatives

Ce que l'on peut attendre d'un logiciel

Présentation des résultats

- visualisation des résultats
- manipulation des tableaux
- bibliothèque de graphiques (2D, 3D, interactifs…)
- navigation dans les arbres de décision
- affichage des courbes de performances (ROC, lift, gain…)
- indice de Gini, aire sous la courbe ROC
- facilité d'incorporation de ces éléments dans un rapport
- Gestion des métadonnées
- variables définies identiquement pour tous les fichiers du projet (identifiant, cible, exclusions…)
- définition de groupes de variables

Ce que l'on peut attendre d'un logiciel

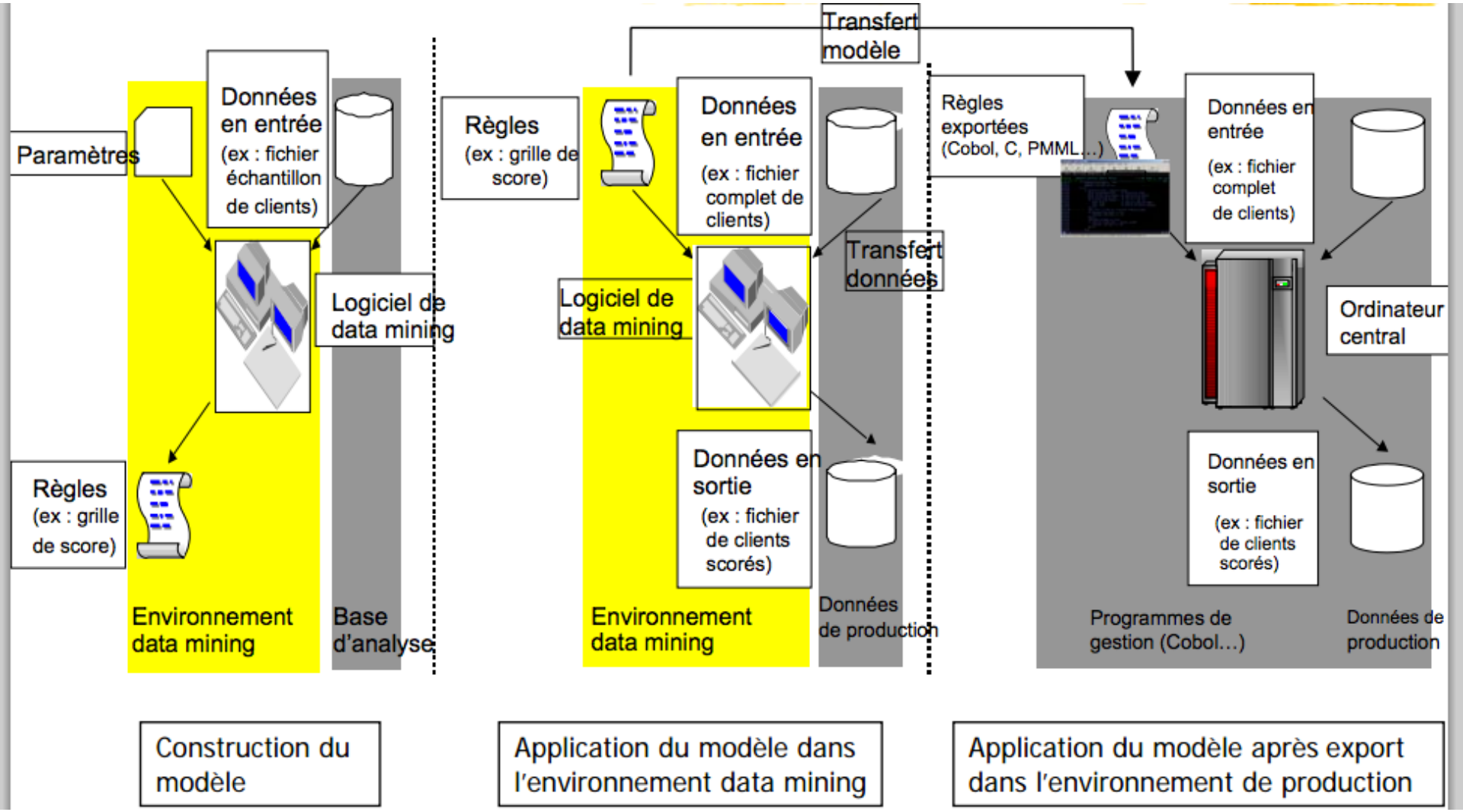
- Plates-formes supportées (Windows, Unix, Sun, IBM MVS…)
- Formats d'entrée/sortie des données gérés :
  - tables Oracle, Sybase, DB2, SAS, fichiers Excel, à plat...
- Enchaînements programmés de plusieurs algorithmes
- Volume de données pouvant être raisonnablement traité

Pour plus de puissance

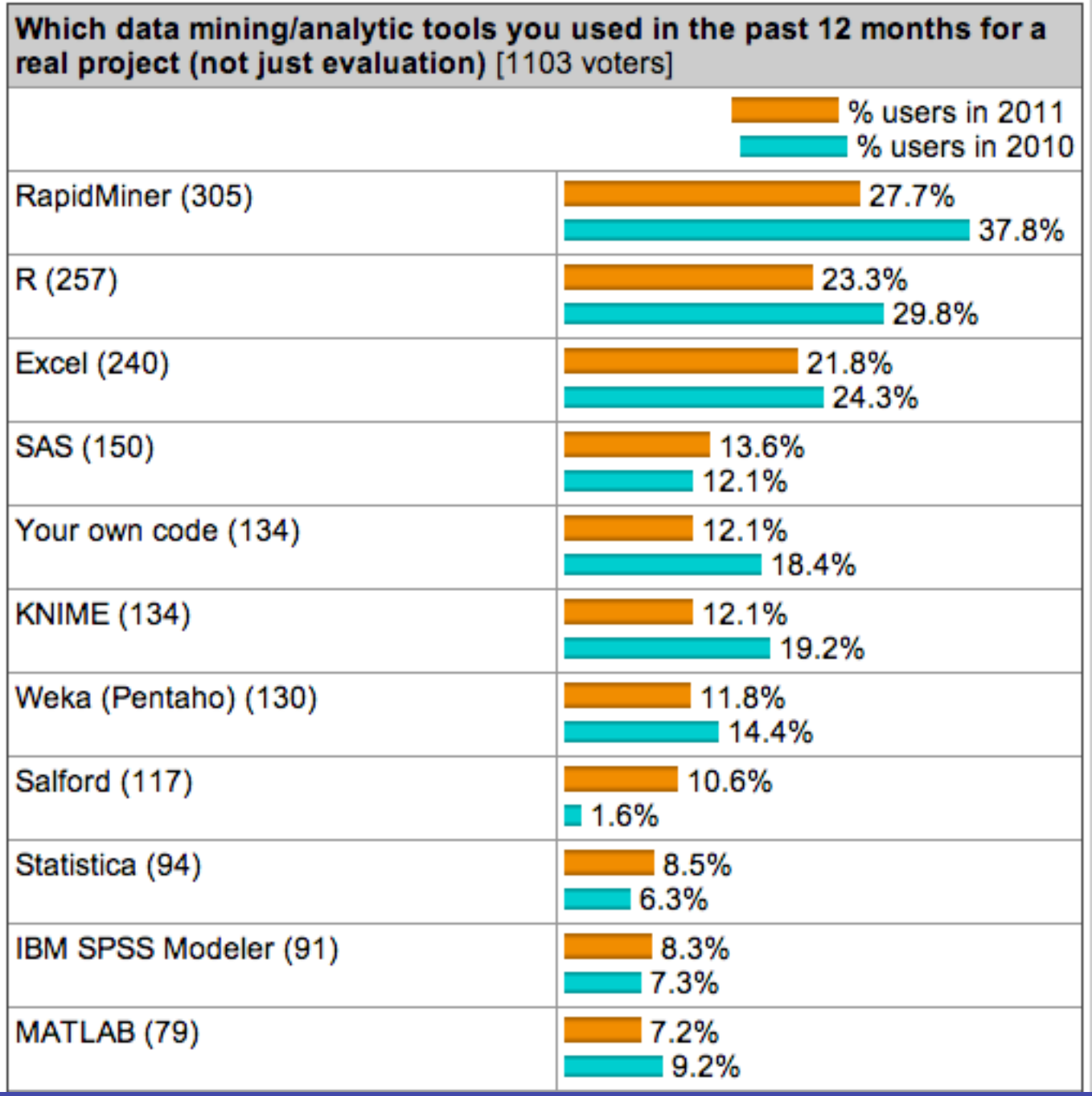
- architecture client-serveur : calculs sur le serveur et visualisation des résultats sur le client
- algorithmes parallélisés

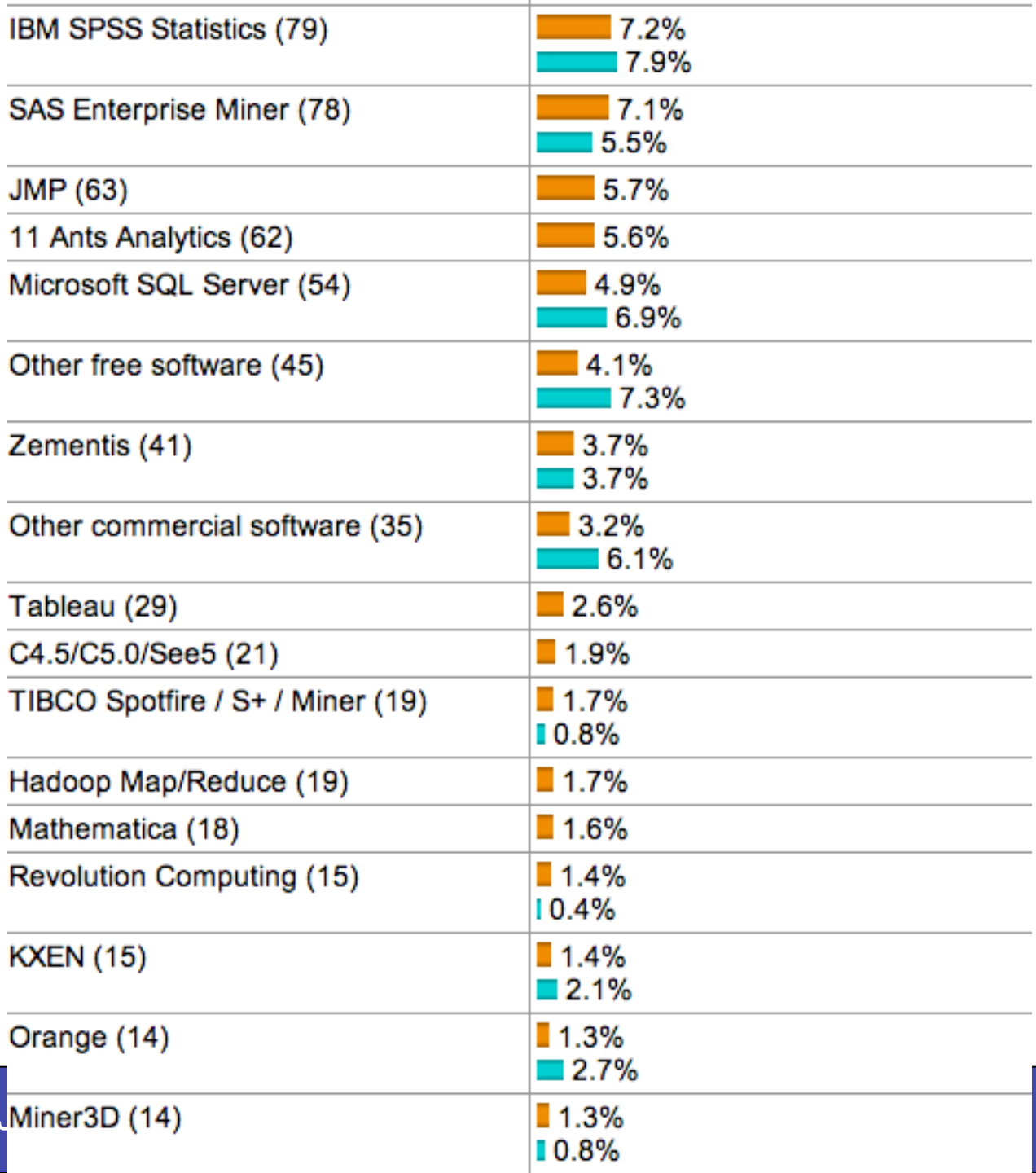
Exécution en mode interactif ou différé

Portabilité des modèles construits (C, XML, Java, SQL...)


Paramètres
Données en entrée (ex : fichier échantillon de clients)
Logiciel de data mining
Règles (ex : grille de score)
Environnement data mining
Base d'analyse
Construction du modèle
Transfert modèle
Règles (ex : grille de score)
Données en entrée (ex : fichier complet de clients)
Logiciel de data mining
Transfert données
Données en sortie (ex : fichier de clients scorés)
Environnement data mining
Données de production
Application du modèle dans l'environnement data mining
Règles exportées (Cobol, C, PMML…)
Données en entrée (ex : fichier complet de clients)
Ordinateur central
Données en sortie (ex : fichier de clients scorés)
Programmes de gestion (Cobol…)
Données de production
Application du modèle après export dans l'environnement de production

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [1103 voters]




IBM SPSS Statistics (79)
7.2%
7.9%
SAS Enterprise Miner (78)
7.1%
5.5%
JMP (63)
5.7%
11 Ants Analytics (62)
5.6%
Microsoft SQL Server (54)
4.9%
6.9%
Other free software (45)
4.1%
7.3%
Zementis (41)
3.7%
3.7%
Other commercial software (35)
3.2%
6.1%
Tableau (29)
2.6%
C4.5/C5.0/See5 (21)
1.9%
TIBCO Spotfire / S+ / Miner (19)
1.7%
0.8%
Hadoop Map/Reduce (19)
1.7%
Mathematica (18)
1.6%
Revolution Computing (15)
1.4%
0.4%
KXEN (15)
1.4%
2.1%
Orange (14)
1.3%
2.7%
Miner3D (14)
1.3%
0.8%

XLSTAT (10)	0.9%
NoSQL databases (10)	0.9%
Stata (9)	0.8%
Other cloud-based tools (9)	0.8%
Bayesia (9)	0.8% 0.1%
Angoss (9)	0.8% 0.9%
Oracle Data Miner (8)	0.7% 2.1%
Predixion (6)	0.5%
WordStat (5)	0.5%
Megaputer Polyanalyst/TextAnalyst (4)	0.4% 0.3%
Portrait Software (3)	0.3% 0.2%
Grapheur (3)	0.3%
Clarabridge (3)	0.3%
Centrifuge (3)	0.3% 0.2%
Viscovery (1)	0.1% 1.1%
Data Applied (1)	0.1% 0.2%

Outils non présents dans le sondage

IBM Cognos (used by 6% of data miners in 2010)

Minitab (9%)

FICO Fair Isaac (3%)

SAP Business Objects/NetWeaver (4%)

Teradata Warehouse Miner (2%)

Unica Predictive Insight (2%)

Logiciels multi-techniques	Insight - S-PLUS R (libre) Weka (gratuit) Tanagra (gratuit)	SAS Entreprise Miner SPSS Clementine Statsoft - Statistica Data Miner Insight - Insightful Miner SPAD KXEN
Logiciels mono-techniques	Salford Systems CART Isoft Alice Neuralware Predict DataLab (spécialiste du prétraitement des données)	SPSS Answer Tree
	Logiciels micros	**Logiciels gros systèmes**

Logiciels de data mining
(poids légers : dizaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Éditeur
Stat Lab		SLP InfoWare (Gemplus)
StartMiner	Réseaux de neurones – Arbres de décision	Grimmersoft
Alice	Arbres de décision	Isoft
Predict	Réseaux de neurones	Neuralware
NeuroOne	Réseaux de neurones	Netral
Wizwhy	Associations	Wizsoft
WEKA		« open source » (logiciel gratuit)
R		« open source » (initialement développé à l'Université d'Auckland, Nelle-Zélande)
DATALAB	Prétraitement des données	Complex Systems

Logiciels de data mining
(poids moyens : centaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Éditeur
4Thought	Réseaux de neurones	Cognos
KnowledgeSEEKER	Arbres de décision	Angoss
KnowledgeSTUDIO		Angoss
C5.0 (Unix) See5 (Windows)	Arbres de décision	RuleQuest Research
Data Mining Suite		Salford Systems
CART	Arbres de décision	Salford Systems
Polyanalyst		Megaputer
S-PLUS		Insightful
TANAGRA		Laboratoire ERIC de l'Université de Lyon

Logiciels de data mining (poids lourds : millions de lignes)

Produit	Spécialité (le cas échéant)	Éditeur
KXEN	Théorie de l'apprentissage de Vapnik	KXEN
Intelligent Miner	Classification relationnelle – Réseaux de neurones	IBM
Microsoft Analysis Services	Arbres de décision – clustering	Microsoft
Oracle Data Mining		Oracle
SPAD		SPAD
SPSS		SPSS
Clementine		SPSS
Statistica Data Miner		Statsoft
Insightful Miner		Insightful
SAS/STAT		SAS
Entreprise Miner		SAS

Tool	Type	Link
ADAPA (Zementis)	DMS	www.zementis.com
Alice (d'Isoft)	DMS	www.alice-soft.com
Bayesia Lab	SPEC	www.bayesia.com
C5.0	SPEC	www.rulequest.com
CART	SPEC	www.salford-systems.com
Data Applied	DMS	data-applied.com
DataDetective	DMS	www.sentient.nl/?dden
DataEngine	DMS	www.dataengine.de
Datascope	DMS	www.cygron.hu
DB2 Data Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
DeltaMaster	BI	www.bissantz.com/deltamaster
Forecaster XL	EXT	www.alyuda.com
GhostMiner	DMS	www.fqs.pl/business_intelligence/products/ghostminer
IBM Cognos 8 BI	BI	www.ibm.com/software/data/cognos/data-mining-tools.html
IBM SPSS Modeler	DMS	www.spss.com/software/modeling/modeler
IBM SPSS Statistics	MAT	www.spss.com/software/statistics
iModel	DMS	www.biocompsystems.com/products/imodel
InfoSphere Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
JMP	DMS	www.jmpdiscovery.com
KnowledgeMiner	SPEC	www.knowledgeminer.net
KnowledgeStudio	DMS	www.angoss.com
KXEN	DMS	www.kxen.com
Magnum Opus	SPEC	www.giwebb.com
MATLAB	MAT	www.mathworks.com
MATLAB Neural Network Toolbox	EXT	www.mathworks.com
Model Builder	DMS	www.fico.com
ModelMAX	SOL	www.asacorp.com/products/mmxover.jsp

Tool	Type	Link
Molegro Data Modeler	SOL	www.molegro.com
NAG Data Mining Components	LIB	www.nag.co.uk/numeric/DR/DRdescription.asp
NeuralWorks Predict	SPEC	www.neuralware.com/products.jsp
Neurofusion	LIB	www.alyuda.com
Neuroshell	SPEC	www.neuroshell.com
Oracle Data Mining (ODM)	DMS	www.oracle.com/technology/products/bi/odm/index.html
Partek Discovery Suite	DMS	www.partek.com/software
Partek Genomics Suite	SOL	www.partek.com/software
PolyAnalyst	DMS	www.megaputer.com/polyanalyst.php
PolyVista	BI	www.polyvista.com
Random Forests	SPEC	www.salford-systems.com
RapAnalyst	SPEC	www.raptorinternational.com/rapanalyst.html
R-PLUS	MAT	www.experience-rplus.com
SAP Netweaver Business Warehouse (BW)	BI	www.sap.com/platform/netweaver/components/businesswarehouse
SAS Enterprise Miner	DMS	www.sas.com/products/miner
See5	SPEC	www.rulequest.com
SPAD Data Mining	DMS	eng.spadsoft.com
SQL Server Analysis Services	DMS	www.microsoft.com/sql
STATISTICA	DMS	www.statsoft.com/products/data-mining-solutions/G259
SuperQuery	DMS	www.azmy.com
Teradata Database	BI	www.teradata.com
Think Enterprise Data Miner (EDM)	DMS	www.thinkanalytics.com
TIBCO Spotfire	DMS	spotfire.tibco.com
Unica PredictiveInsight	DMS	www.unica.com
WizRule and WizWhy	SPEC	www.wizsoft.com
XAffinity	SPEC	www.exclusiveore.com

Tool	Type	Link
ADaM*	LIB	datamining.itsc.uah.edu/adam
CellProfilerAnalyst	SOL	www.cellprofiler.org/index.htm
D2K*	DMS	alg.ncsa.uiuc.edu
Gait-CAD	INT	sourceforge.net/projects/gait-cad
GATE	SOL	gate.ac.uk/download
GIFT	RES	www.gnu.org/software/gift
Gnome Data Mine Tools	DMS	www.togaware.com/datamining/gdatamine
Himalaya	RES	himalaya-tools.sourceforge.net
ImageJ	SOL	rsbweb.nih.gov/ij
ITK	SOL	www.itk.org
JAVA Data Mining Package	LIB	sourceforge.net/projects/jdmp
JavaNNS	SPEC	www.ra.cs.uni-tuebingen.de/software/JavaNNS/welcome_e.html
KEEL	INT	www.keel.es
Kepler	MAT	kepler-project.org
KNIME	INT	www.knime.org
LibSVM	LIB	www.csie.ntu.edu.tw/ cjlin/libsvm
MEGA	SOL	www.megasoftware.net/m_distance.html
MLC++	LIB	www.sgi.com/tech/mlc
Orange	LIB	www.ailab.si/orange
Pegasus	RES	www.cs.cmu.edu/ pegasus
Pentaho	BI	sourceforge.net/projects/pentaho
Proximity	SPEC	kdl.cs.umass.edu/proximity/index.html
PRTools	EXT	www.prtools.org
R	MAT	www.r-project.org
RapidMiner	DMS	www.rapidminer.com
Rattle	INT	rattle.togaware.com
ROOT	LIB	root.cern.ch/root
ROSETTA	SPEC	www.lcb.uu.se/tools/rosetta/index.php
Rseslibs	RES	logic.mimuw.edu.pl/ rses
Rule Discovery System*	SPEC	www.compumine.com
RWEKA	INT	cran.r-project.org/web/packages/RWeka/index.html
TANAGRA	INT	eric.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html
Waffles	LIB	waffles.sourceforge.net
WEKA	DMS, LIB	sourceforge.net/projects/weka
XELOPES Library*	LIB	www.prudsys.de/en/technology/xelopes
XLMiner*	EXT	www.resample.com/xlminer

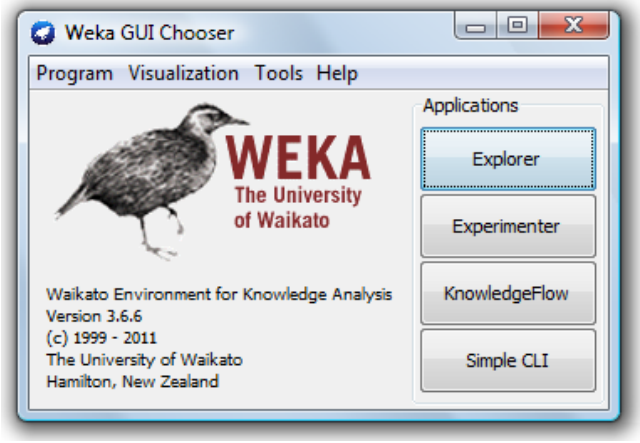
Exemple d'outils

- Intelligent Miner d'IBM
  - Intelligent Miner for Data (IMA)
  - Intelligent Miner for Text (IMT)
  - Tâches : groupage de données, classification, recherche d'associations, etc.
- Entreprise Miner de SAS
  - SAS : longue expérience en statistiques
  - Outil «complet» pour le DM

Weka

Logiciel d'apprentissage machine :

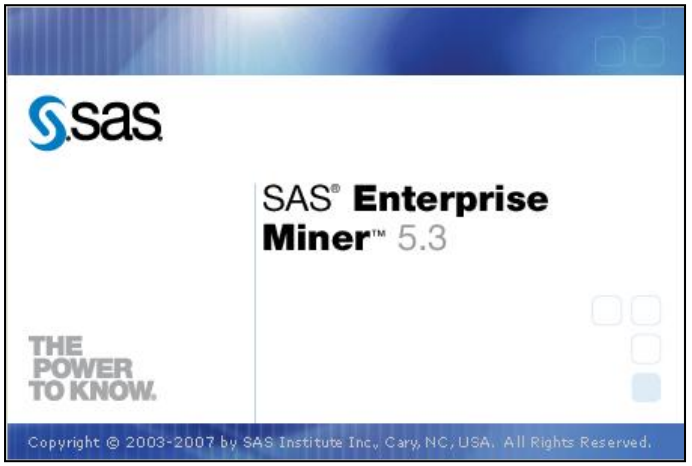
- Traitement de données
- Forage de données
- Comparaison d'algorithmes
- Etc.



Site web:

http://www.cs.waikato.ac.nz/ml/weka/index.html

SAS Entreprise Miner



Société : SAS Institute Inc.

Création : Mai 1998

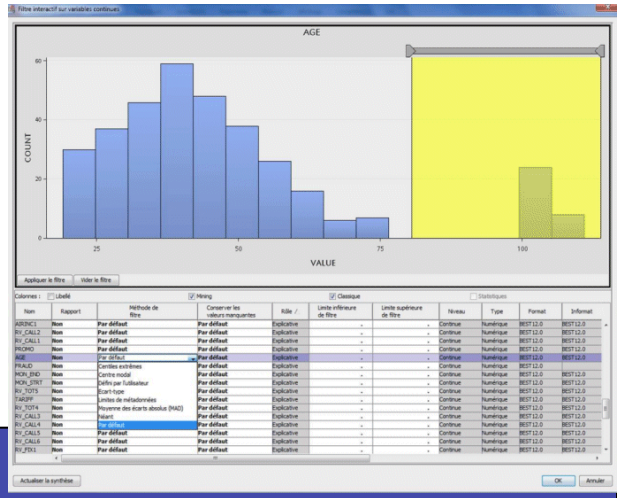
Plate-formes : Windows , Unix, Linux

Utilisation

- Réduction des coûts
- Maîtrise des risques
- Fidélisation
- Prospection

Outils de data warehouse

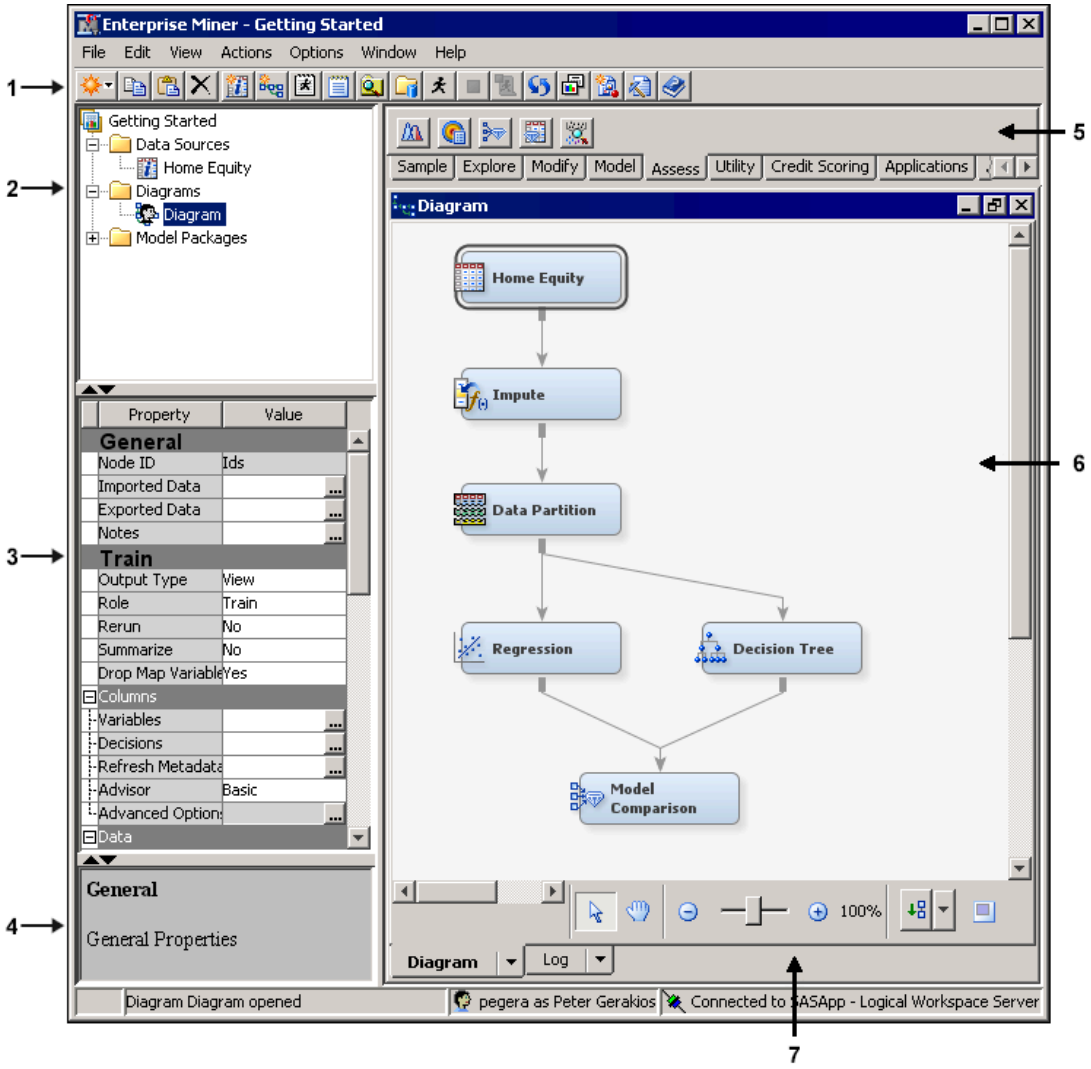
http://www.sas.com/offices/europe/france/software/technologies/datamining/datamining.html



SAS Entreprise Miner

Interface graphique (icônes)

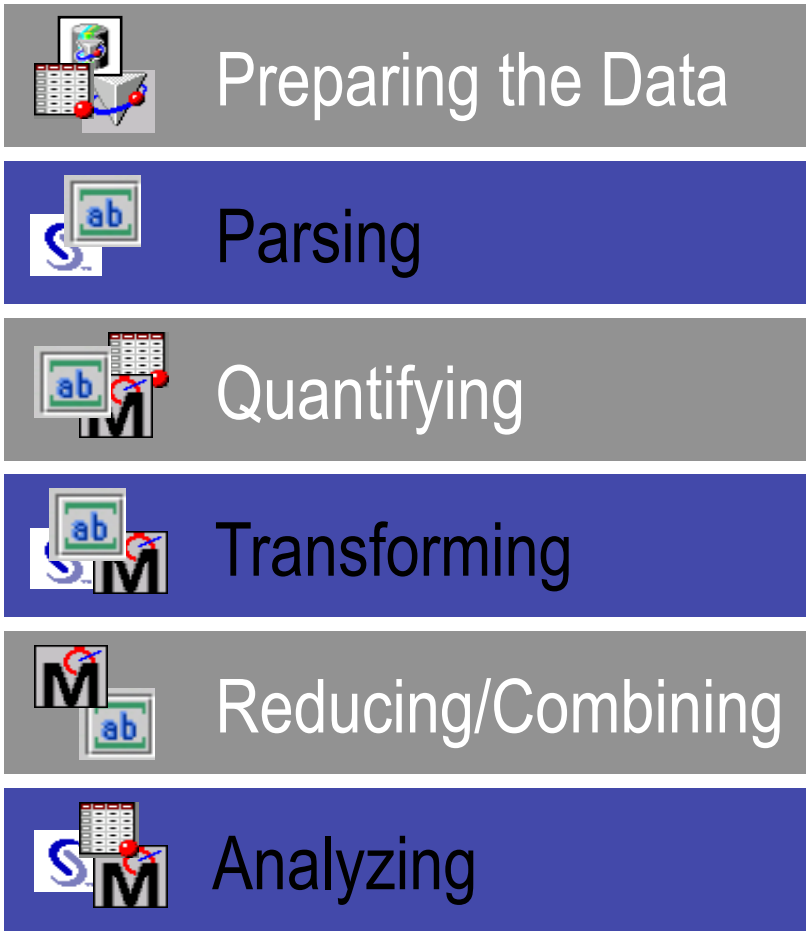
Construction d'un diagramm



SAS Entreprise Miner

Deux types d'utilisateurs
- Spécialistes en statistiques
- Spécialistes métiers (chef de projet, études…)

Techniques implémentées
- Arbres de décision
- Régression
- Réseaux de neurones



Alice

Société : ISoft

Création : 1988

Plate-formes :

Utilisation

- Marketing : études de marché, segmentation …
- Banque, Assurance : scoring, analyse de risques, détection de fraudes
- Industrie : contrôle qualité, diagnostic, segmentation, classification, construction de modèles, prédiction et simulation

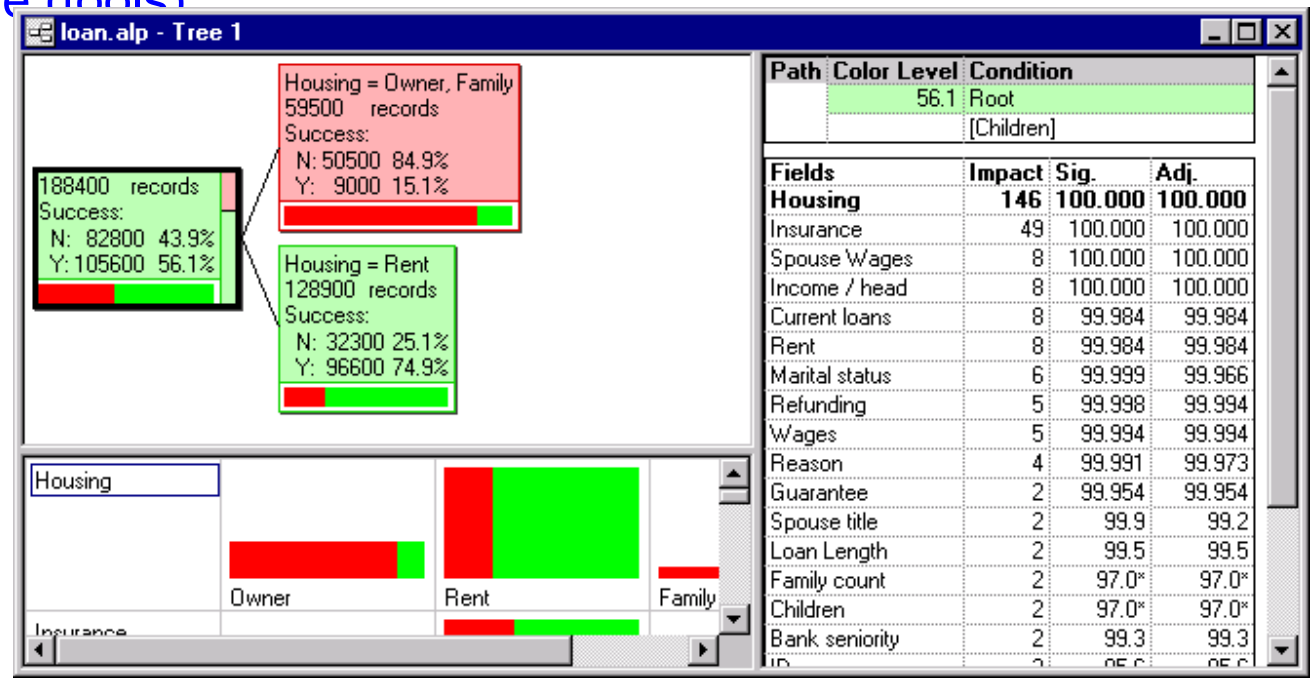
SITE WEB http://www.isoft.fr/html/prod_alice.htm

FONCTIONNALITES

- ALICE / Arbre de décision interactif
- ALICE / CLUSTERING: regroupe les individus semblables en classes

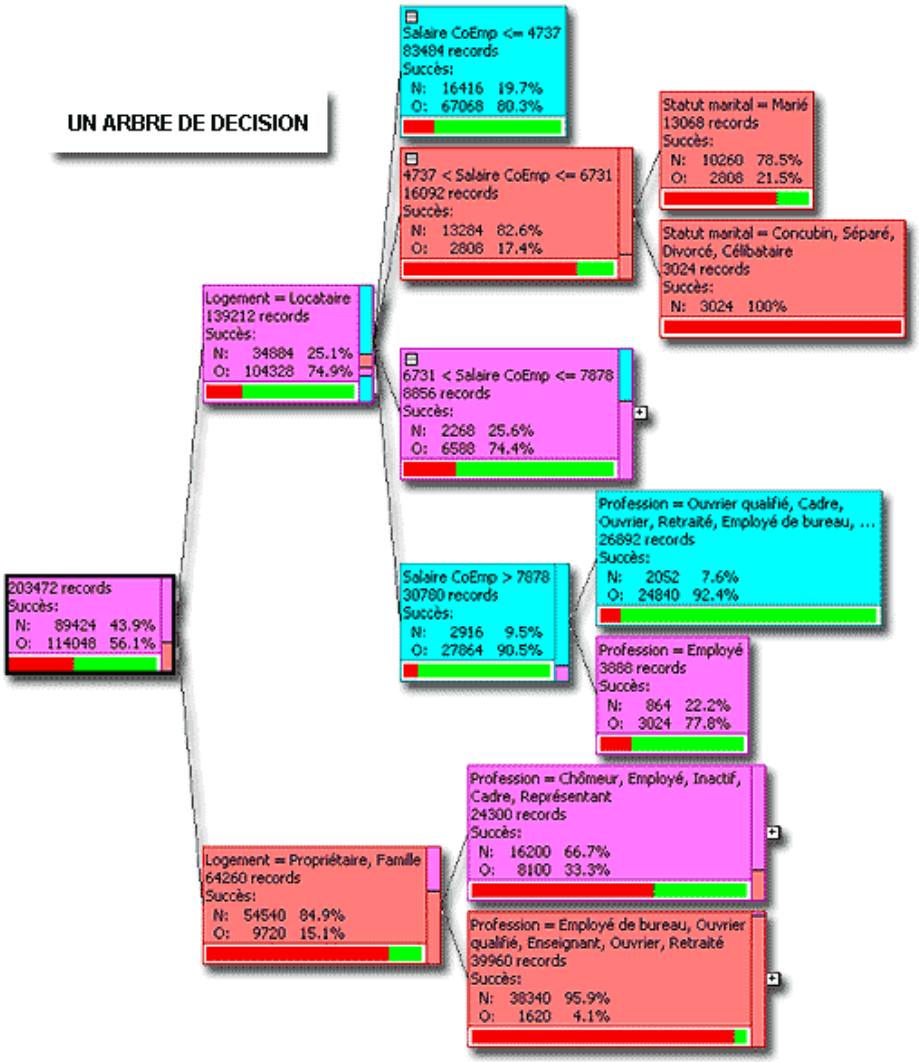
Alice

Interface graphique (tools)



Type d'utilisateur : responsables opérationnels

Alice


UN ARBRE DE DECISION
203472 records
Succès:
N: 89424 43.9%
O: 114048 56.1%
Logement = Locataire
139212 records
Succès:
N: 34884 25.1%
O: 104328 74.9%
Logement = Propriétaire, Famille
64260 records
Succès:
N: 54540 84.9%
O: 9720 15.1%
Salaire CoEmp <= 4737
83484 records
Succès:
N: 16416 19.7%
O: 67068 80.3%
4737 < Salaire CoEmp <= 6731
16092 records
Succès:
N: 13284 82.6%
O: 2808 17.4%
6731 < Salaire CoEmp <= 7878
8856 records
Succès:
N: 2268 25.6%
O: 6588 74.4%
Salaire CoEmp > 7878
30780 records
Succès:
N: 2916 9.5%
O: 27864 90.5%
Statut marital = Marié
13068 records
Succès:
N: 10260 78.5%
O: 2808 21.5%
Statut marital = Concubin, Séparé, Divorcé, Célibataire
3024 records
Succès:
N: 3024 100%
Profession = Ouvrier qualifié, Cadre, Ouvrier, Retraité, Employé de bureau, ...
26892 records
Succès:
N: 2052 7.6%
O: 24840 92.4%
Profession = Employé
3888 records
Succès:
N: 864 22.2%
O: 3024 77.8%
Profession = Chômeur, Employé, Inactif, Cadre, Représentant
24300 records
Succès:
N: 16200 66.7%
O: 8100 33.3%
Profession = Employé de bureau, Ouvrier qualifié, Enseignant, Ouvrier, Retraité
39960 records
Succès:
N: 38340 95.9%
O: 1620 4.1%

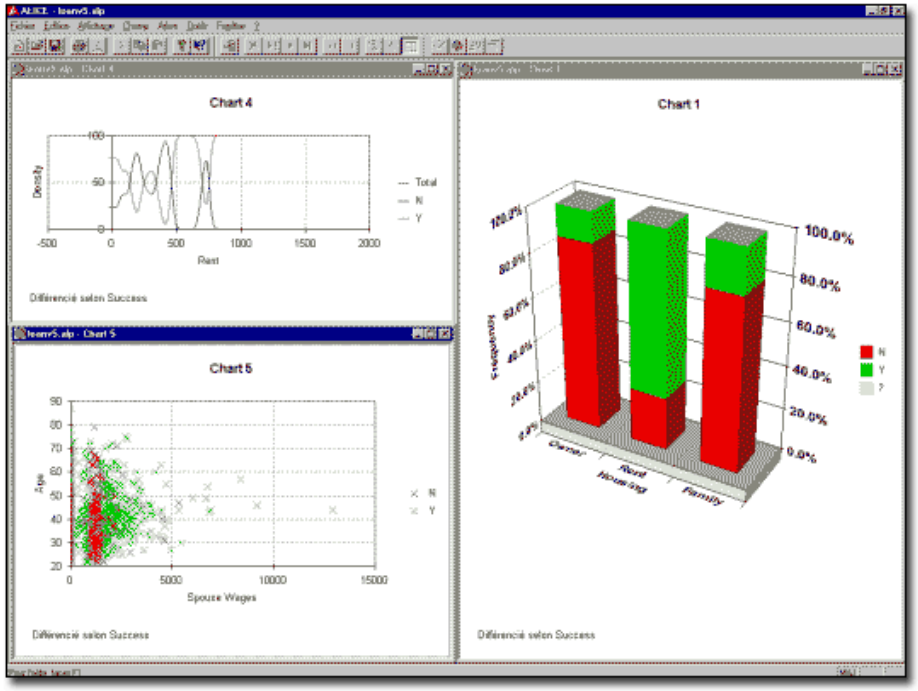


Chart 4
Chart 1
Chart 5
Différencié selon Success

Clémentine

Société : ISL (Integral Solutions Limited) racheté par SPSS et IBM

Création : 1994

Plate-formes : Windows NT, Unix

Utilisation

- Prévision de parts de marché
- Détection de fraudes
- Segmentation de marché
- Implantation de points de vente …

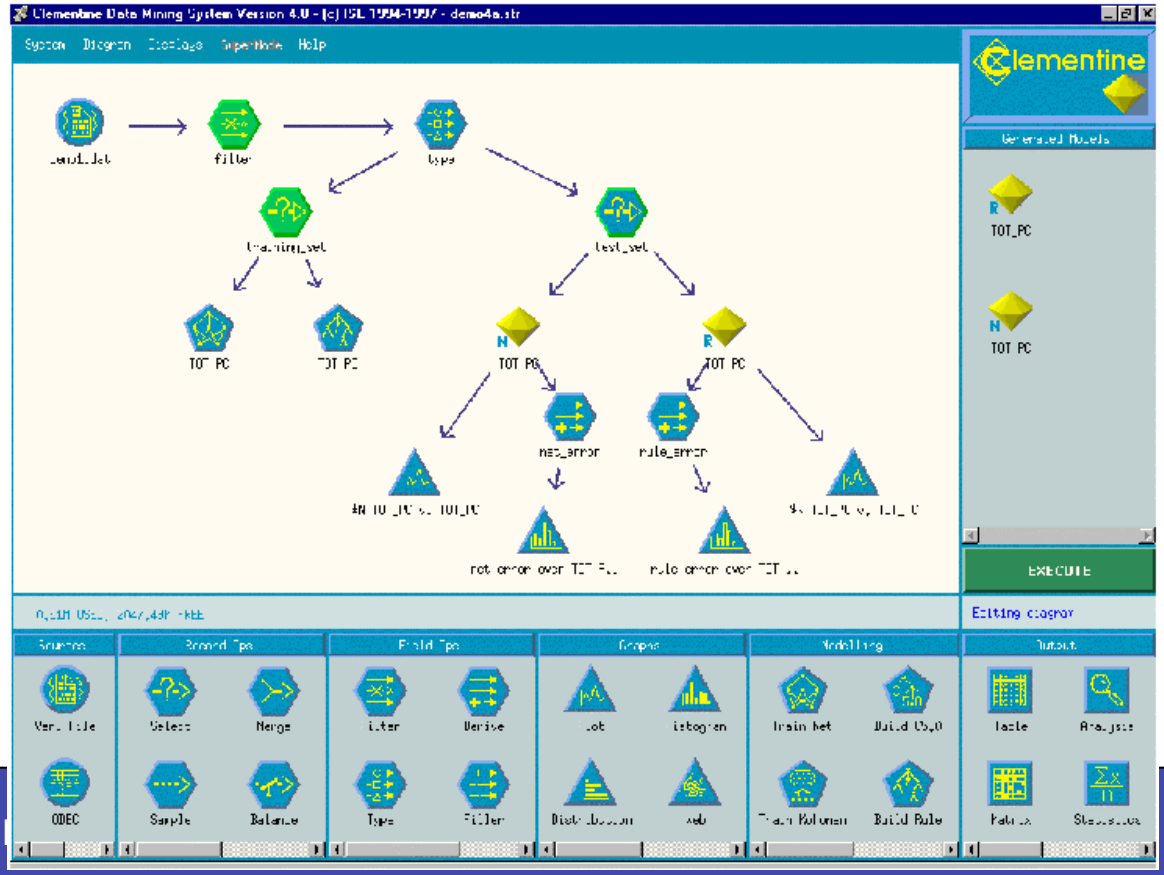
Techniques

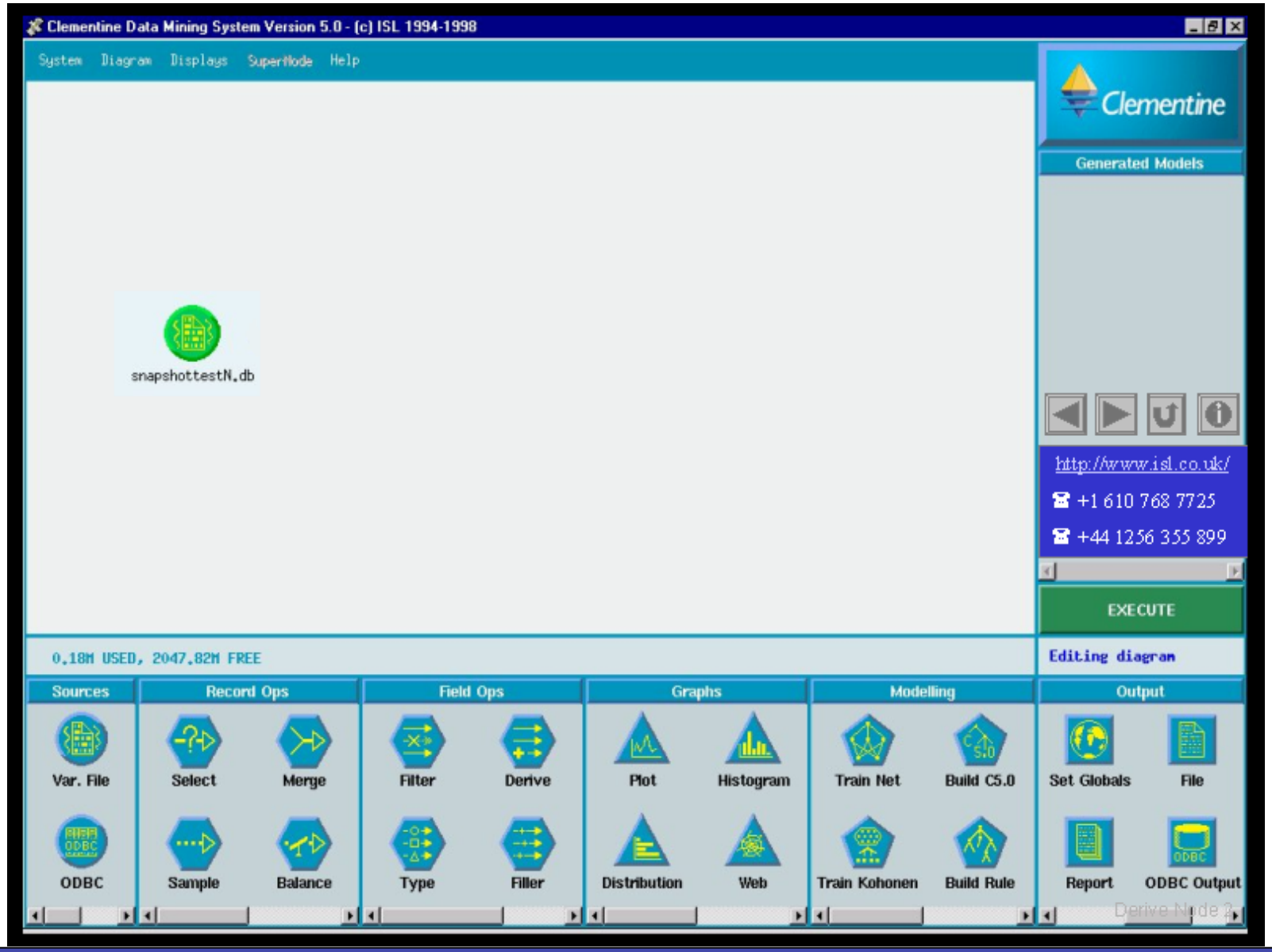
- Rule Induction
- Graph
- Clustering
- Association Rules
- Linear Regression
- - Neural Networks

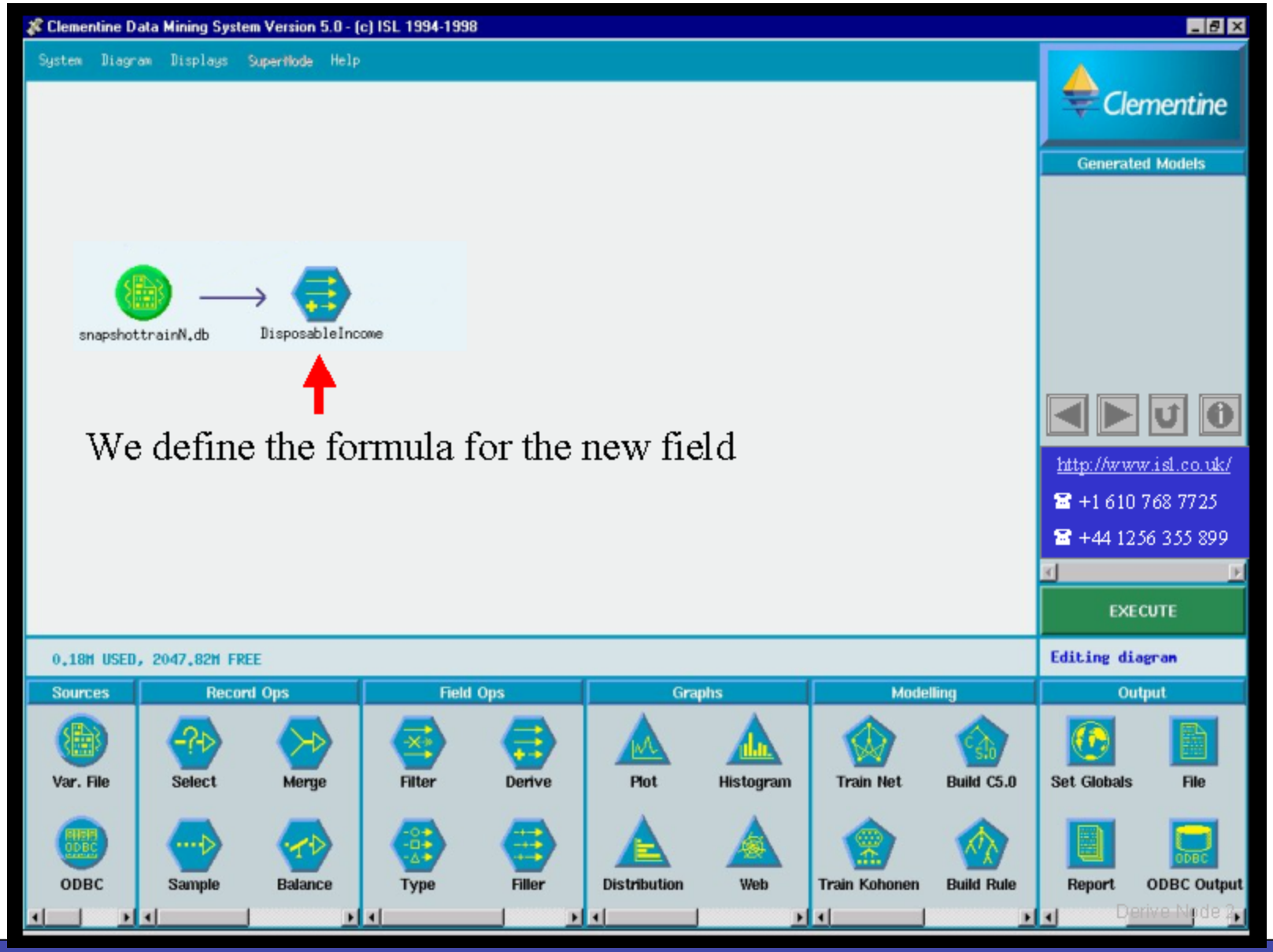
Clémentine

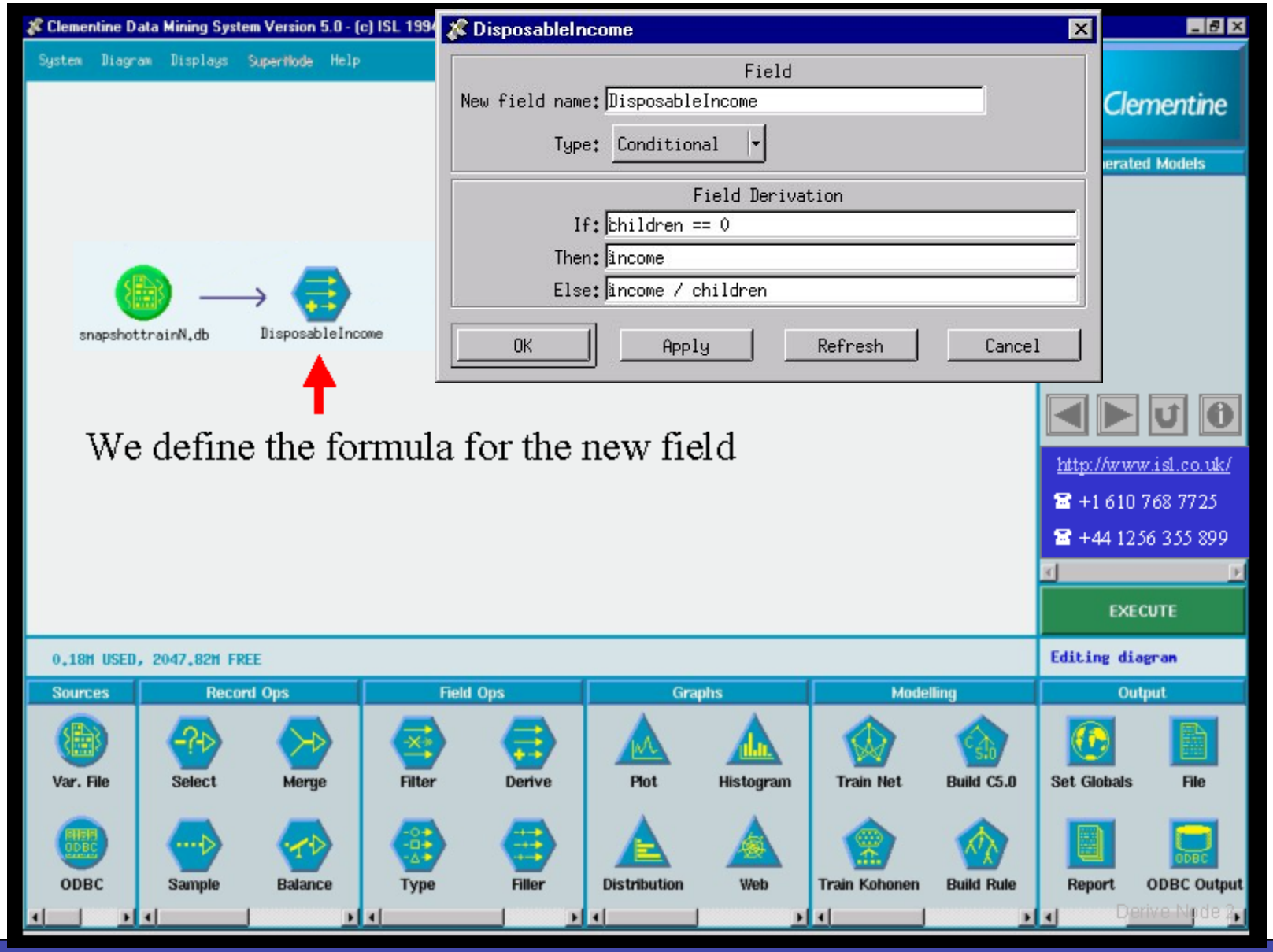
Interface simple, puissante et complète
interface conviviale

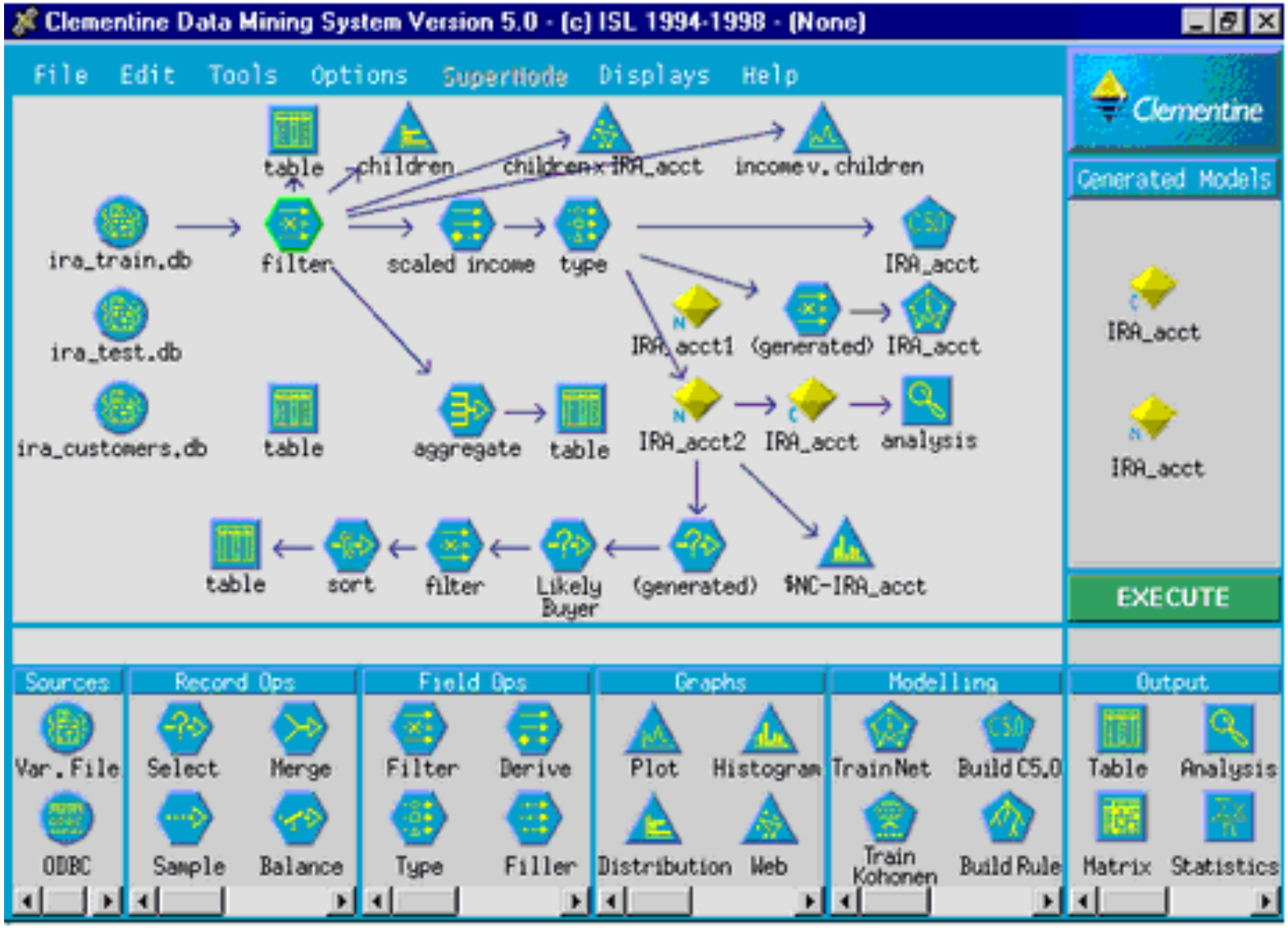
⇨




Clementine Data Mining System Version 5.0 - (c) ISL 1994-1998
System
Diagram
Displays
SuperNode
Help
Clementine
Generated Models
snapshottestN.db
http://www.isl.co.uk/
+1 610 768 7725
+44 1256 355 899
EXECUTE
0.18M USED, 2047.82M FREE
Editing diagram
Sources
Record Ops
Field Ops
Graphs
Modelling
Output
Var. File
Select
Merge
Filter
Derive
Plot
Histogram
Train Net
Build C5.0
Set Globals
File
ODBC
Sample
Balance
Type
Filler
Distribution
Web
Train Kohonen
Build Rule
Report
ODBC Output


Clementine Data Mining System Version 5.0 - (c) ISL 1994-1998
System Diagram Displays SuperNode Help
Clementine
Generated Models
snapshottrainN.db
DisposableIncome
We define the formula for the new field
http://www.isl.co.uk/
+1 610 768 7725
+44 1256 355 899
EXECUTE
0.18M USED, 2047.82M FREE
Editing diagram
Sources
Record Ops
Field Ops
Graphs
Modelling
Output
Var. File
Select
Merge
Filter
Derive
Plot
Histogram
Train Net
Build C5.0
Set Globals
File
ODBC
Sample
Balance
Type
Filler
Distribution
Web
Train Kohonen
Build Rule
Report
ODBC Output
Derive Node 2


Clementine Data Mining System Version 5.0 - (c) ISL 1994
System Diagram Displays SuperNode Help
DisposableIncome
Field
New field name: DisposableIncome
Type: Conditional
Field Derivation
If: children == 0
Then: income
Else: income / children
OK
Apply
Refresh
Cancel
snapshottrainN.db
DisposableIncome
We define the formula for the new field
Clementine
http://www.isl.co.uk/
+1 610 768 7725
+44 1256 355 899
EXECUTE
0.18M USED, 2047.82M FREE
Editing diagram
Sources
Record Ops
Field Ops
Graphs
Modelling
Output
Var. File
Select
Merge
Filter
Derive
Plot
Histogram
Train Net
Build C5.0
Set Globals
File
ODBC
Sample
Balance
Type
Filler
Distribution
Web
Train Kohonen
Build Rule
Report
ODBC Output


Clementine Data Mining System Version 5.0 - (c) ISL 1994-1998 - (None)
File Edit Tools Options SuperNode Displays Help
Clementine
Generated Models
table
children
children+IRA_acct
income v. children
ira_train.db
filter
scaled income
type
C5.0
IRA_acct
ira_test.db
IRA_acct1
(generated)
IRA_acct
IRA_acct
ira_customers.db
table
aggregate
table
IRA_acct2
IRA_acct
analysis
IRA_acct
table
sort
filter
Likely Buyer
(generated)
$NC-IRA_acct
EXECUTE
Sources
Record Ops
Field Ops
Graphs
Modelling
Output
Var. File
Select
Merge
Filter
Derive
Plot
Histogram
TrainNet
Build C5.0
Table
Analysis
ODBC
Sample
Balance
Type
Filler
Distribution
Web
Train Kohonen
Build Rule
Matrix
Statistics

Clémentine

Types d'utilisateurs : PME/PMI, administrations, consultants, universitaires, chefs de projets,…

Facilité d'utilisation (connaissances en statistiques non requises)

Vaste palette de choix graphiques

- Valeurs observées, prévisions, valeurs calculées sur l'historique, intervalles de confiance, diagnostics (erreurs)

Intelligent Miner

Société : IBM

Création : 1998

Plate-formes : AIX, OS/390, OS/400, Solaris, Windows 2000 & NT

Utilisation

- Domaines où l'aide à la décision est très importante (exemple : domaine médical)
- Analyse de textes

Fortement couplé avec DB2 (BD relationnel)

Intelligent Miner

Deux versions

- Intelligent Miner for Data (IMD)
- Intelligent Miner for Text (IMT)

Types d'utilisateurs : spécialistes ou professionnels expérimentés

Parallel Intelligent Miner

Intelligent Miner

L'IMD

- Sélection et codage des données à explorer
- Détermination des valeurs manquantes
- Agrégation de valeurs
- Diverses techniques pour la fouille de données
  - Règles d'association (Apriori), classification (Arbres de décision, réseaux de neurones), clustering, détection de déviation (analyse statistique & visualisation)
- Visualisation des résultats
- Algorithmes extensibles (scalability)

Intelligent Miner

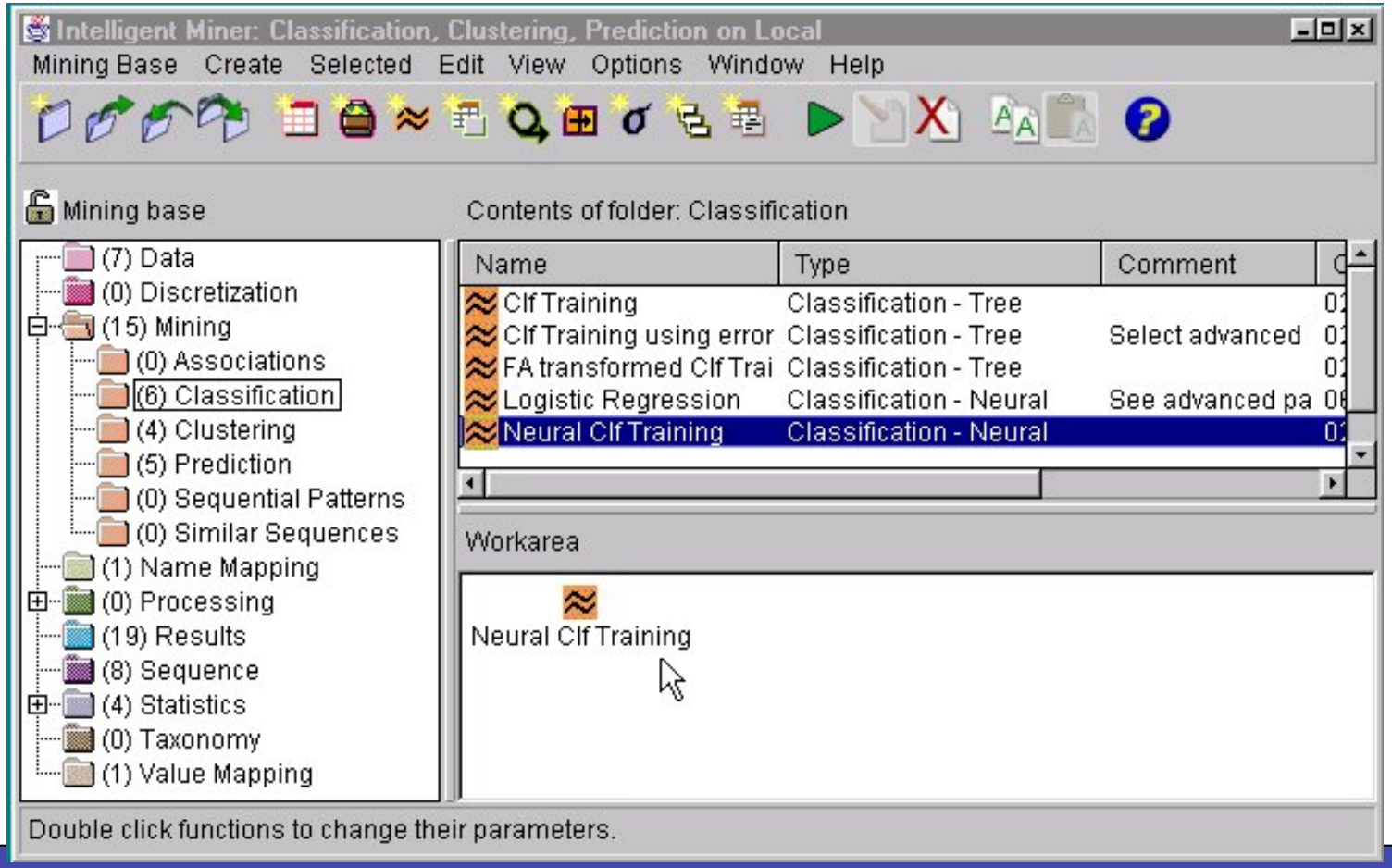
IMT = analyse de textes libres

Trois composants

- Moteur de recherche textuel avancé (TextMiner)
- Outil d'accès au Web (moteur de recherche NetQuestion et un méta-moteur)
- Outil d'analyse de textes (Text Analysis)

L'objectif général est de faciliter la compréhension des textes

Intelligent Miner



Comparatif de quelques solution

Characteristic	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Developer:	RapidMiner, Germany	worldwide development	Univ. of Waikato, New Zealand	Univ. of Ljubljana, Slovenia	KNIME.com AG,Switzerland	multiple; support: INRIA, Google
Programming language:	Java	C, Fortran, R	Java	C++, Python, Qt framew.	Java	Python+NumPy+ SciPy+matplotlib
License:	open s. (v.5 or lower); closed s., free Starter ed. (v.6)	free software, GNU GPL 2+	open source, GNU GPL 3	open source, GNU GPL 3	open source, GNU GPL 3	FreeBSD
Current version:	6	3.02	3.6.10	2.7	2.9.1	0.14.1
GUI / command line:	GUI	both; (GUI for DM = Rattle)	both	both	GUI	command line
Main purpose:	general data mining	sci. computation and statistics	general data mining	general data mining	general data mining	machine learning package add-on
Community support (est.):	large (~200 000 users)	very large (~ 2 M users)	large	moderate	moderate (~ 15 000 users)	moderate

Name	RapidMiner	R	Weka	Orange	KNIME	scikit-learn
Big data	S (not free: Radoop)	A (ff, ffbase)	S (CLI, knowl. flow, distributedWekaHadoop)	-	A	S
Link, graph mining	-	A (igraph, sna)	A	-	A	-
Spatial data analysis	-	A (ggmap)	-	-	A	S
Time-series analysis	A	+, A(forecast)	S (several time series filters)	-	+	S (timeseries module has bugs)
Semi-super-vised learning	S	A (upclass)	S	-	S	+ (label propagation)
Data streams	+	A (stream)	A (massiveOnlineAnalysis)	-	+	S
Text mining	A	A (tm, RTextTools, qdap)	S	A	A	+
Paralelization	S (enterprise ed.)	A (snow, multicore)	S	-	+	A (joblib)
Deep learning	-	S (darch: incomplete)	-	-	-	S (Restricted Boltzmann Mach.)

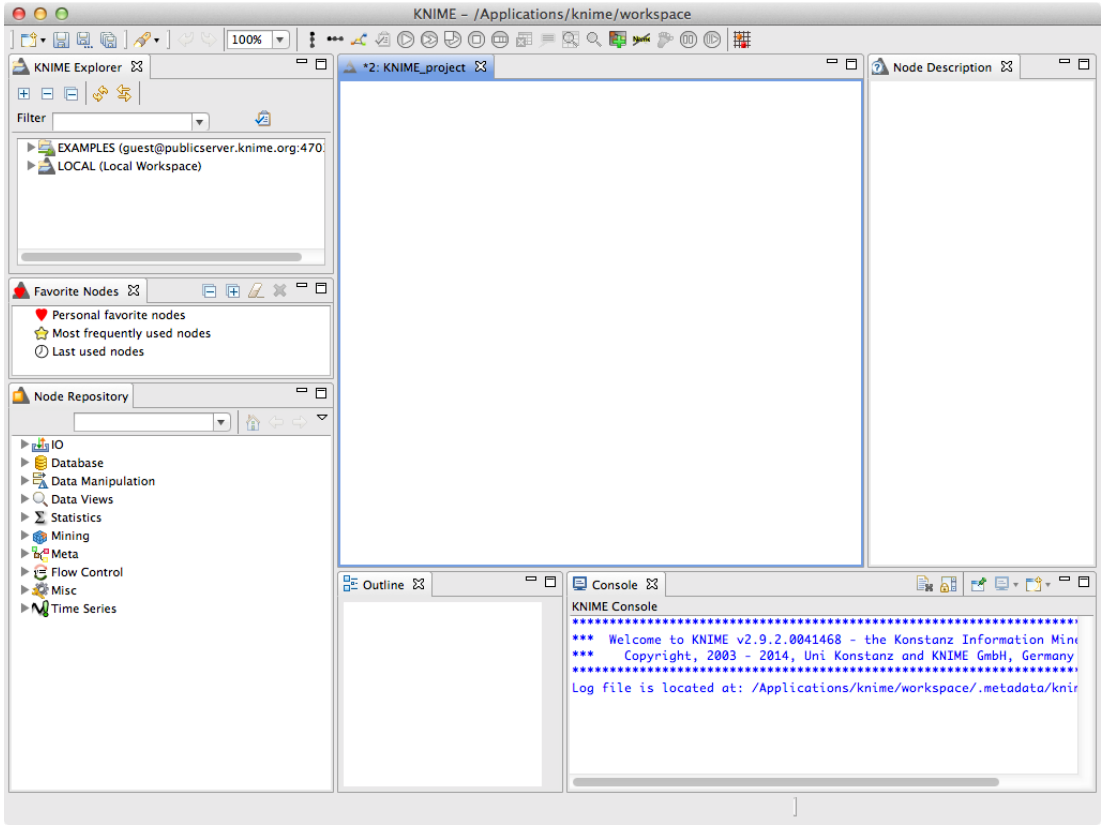
Outils Open Source

Quelques outils classiques

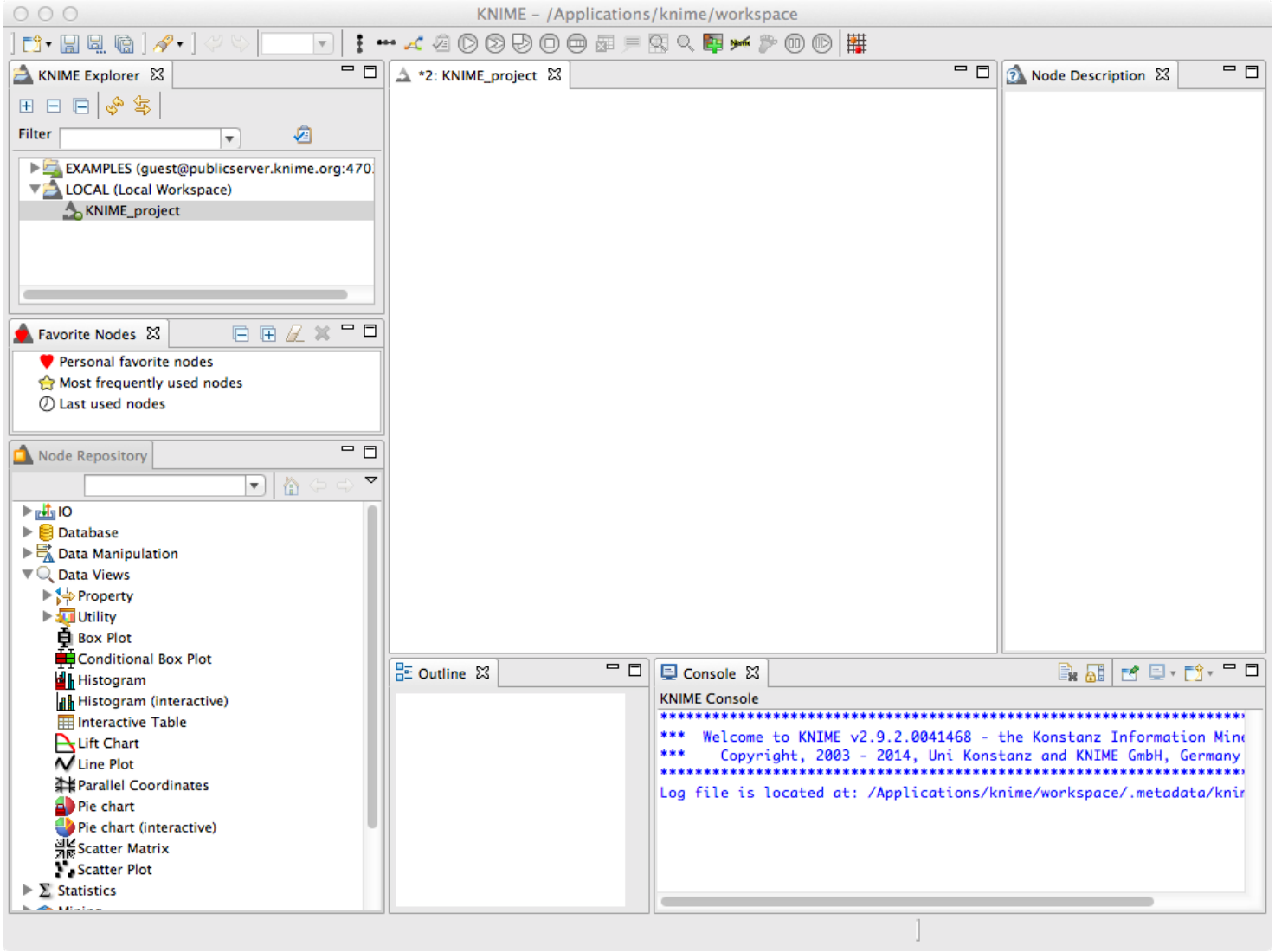
- Weka www.cs.waikato.ac.nz/ml/weka/
- Rapid Miner www.rapid-i.com
- Knime www.knime.org
- Orange orange.biolab.si

Knime

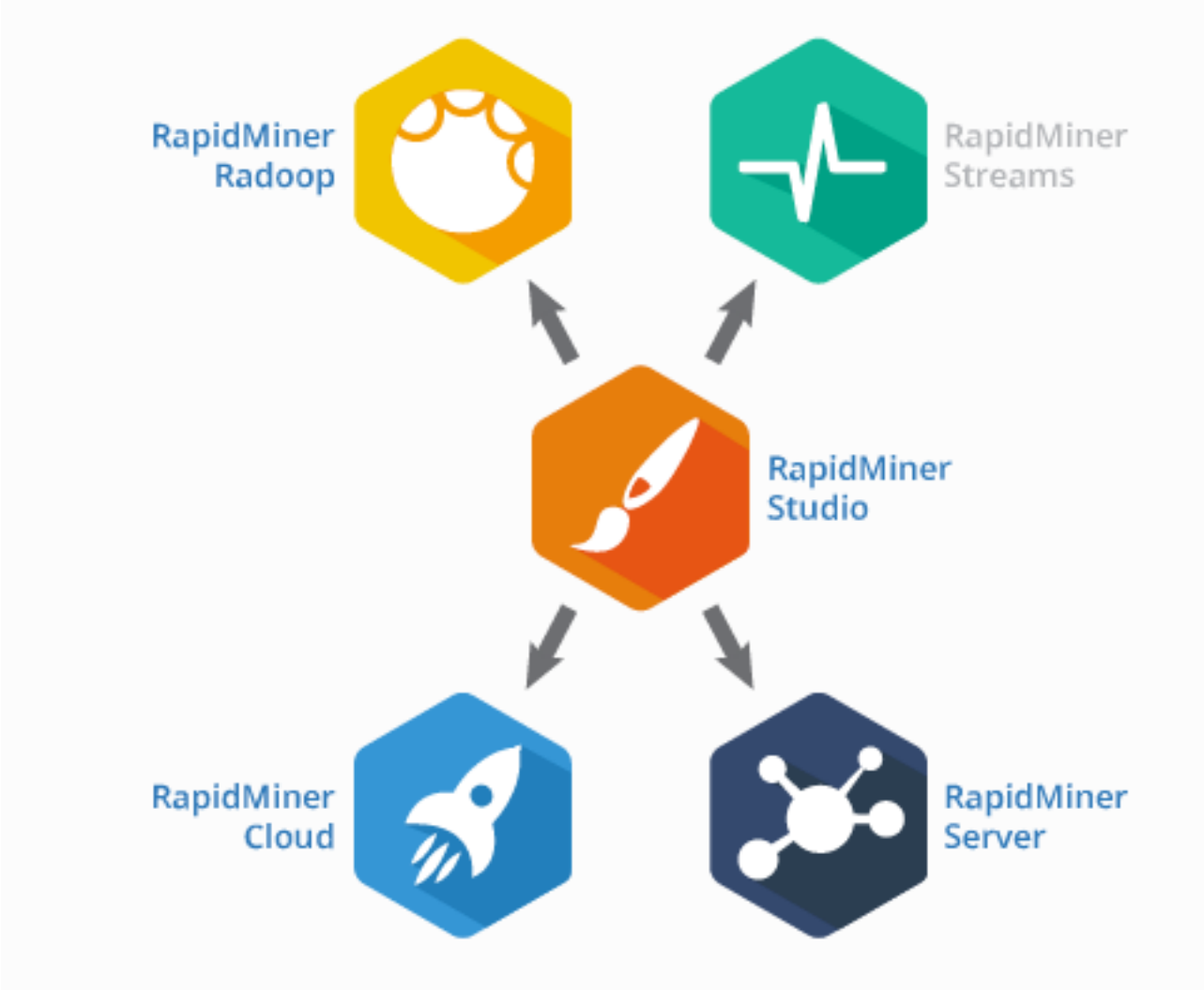
- Support Java



Knime

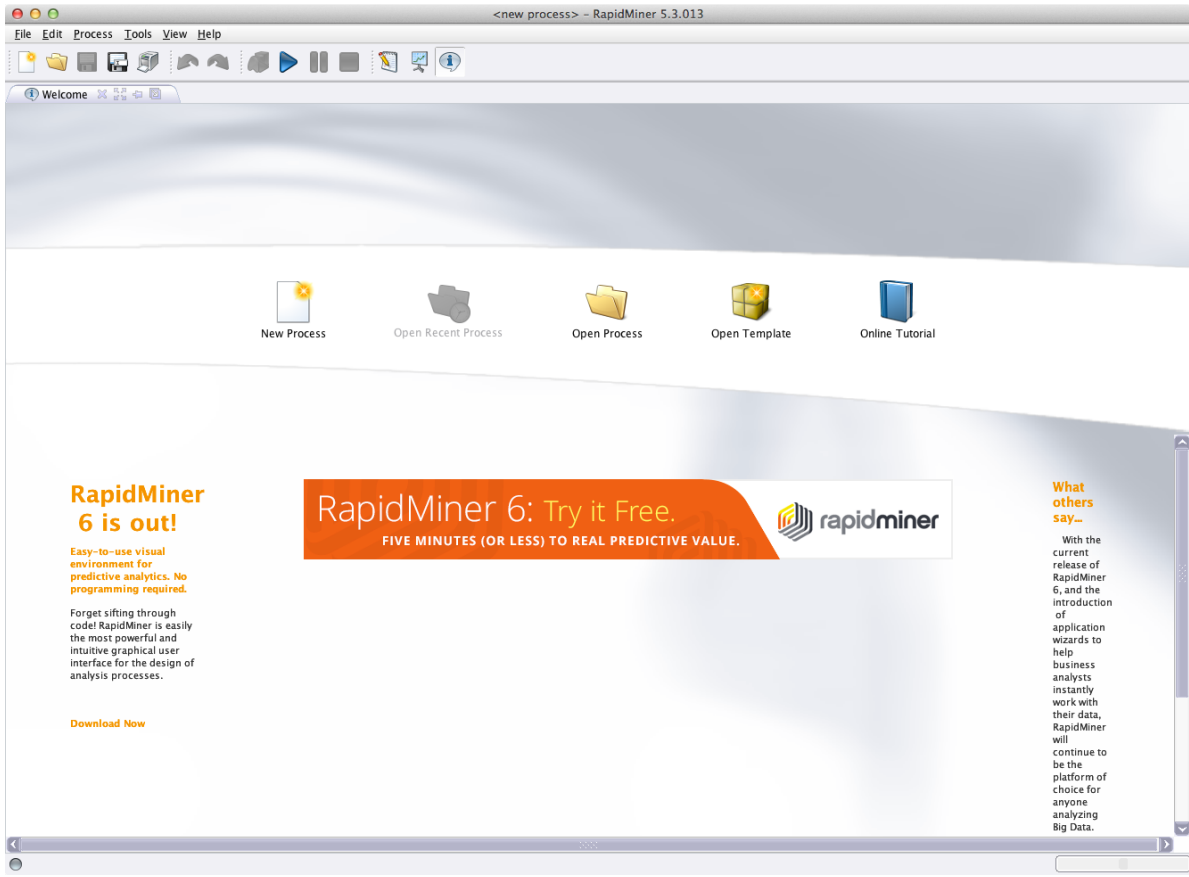

KNIME – /Applications/knime/workspace
KNIME Explorer
Filter
EXAMPLES (guest@publicserver.knime.org:470
LOCAL (Local Workspace)
KNIME_project
*2: KNIME_project
Node Description
Favorite Nodes
Personal favorite nodes
Most frequently used nodes
Last used nodes
Node Repository
IO
Database
Data Manipulation
Data Views
Property
Utility
Box Plot
Conditional Box Plot
Histogram
Histogram (interactive)
Interactive Table
Lift Chart
Line Plot
Parallel Coordinates
Pie chart
Pie chart (interactive)
Scatter Matrix
Scatter Plot
Statistics
Outline
Console
KNIME Console
*** Welcome to KNIME v2.9.2.0041468 - the Konstanz Information Mine
*** Copyright, 2003 - 2014, Uni Konstanz and KNIME GmbH, Germany
Log file is located at: /Applications/knime/workspace/.metadata/knir

Rapid Miner

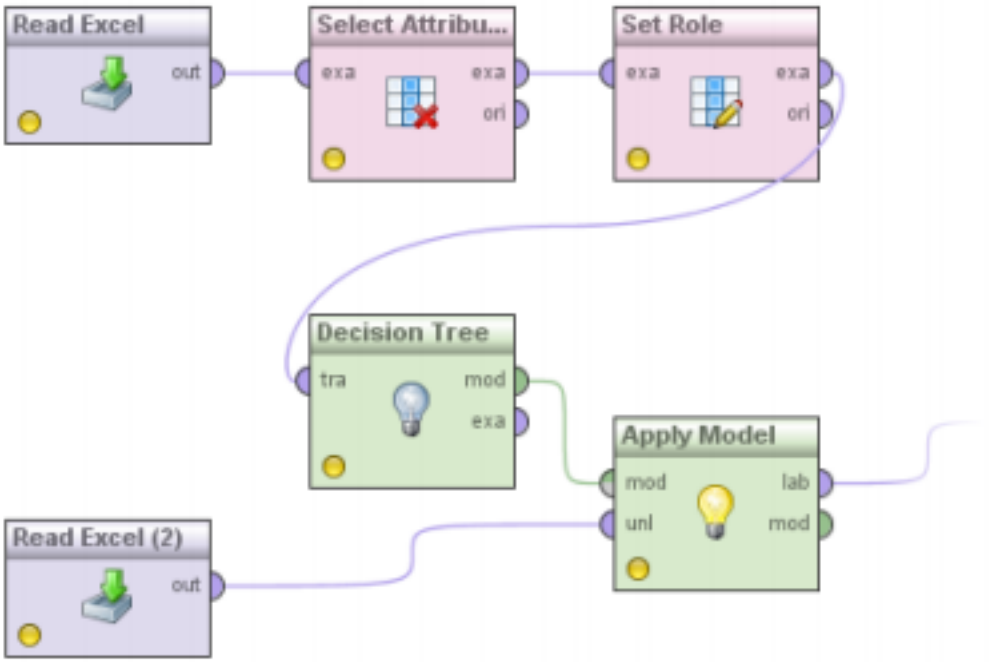

RapidMiner
Radoop
RapidMiner
Streams
RapidMiner
Studio
RapidMiner
Cloud
RapidMiner
Server

RapidMiner

Java

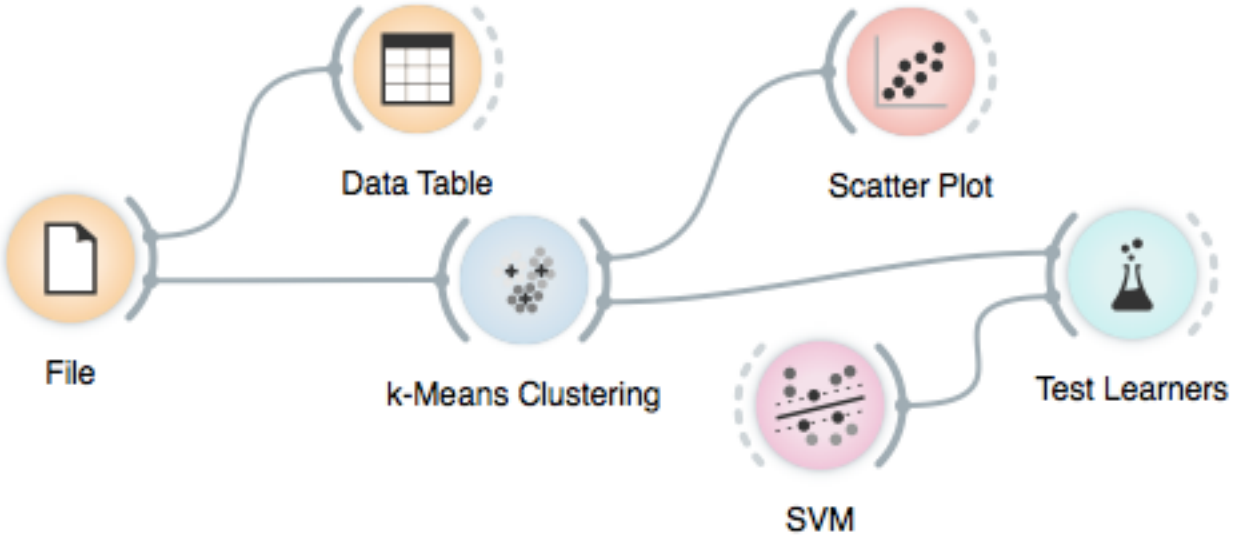


RapidMiner

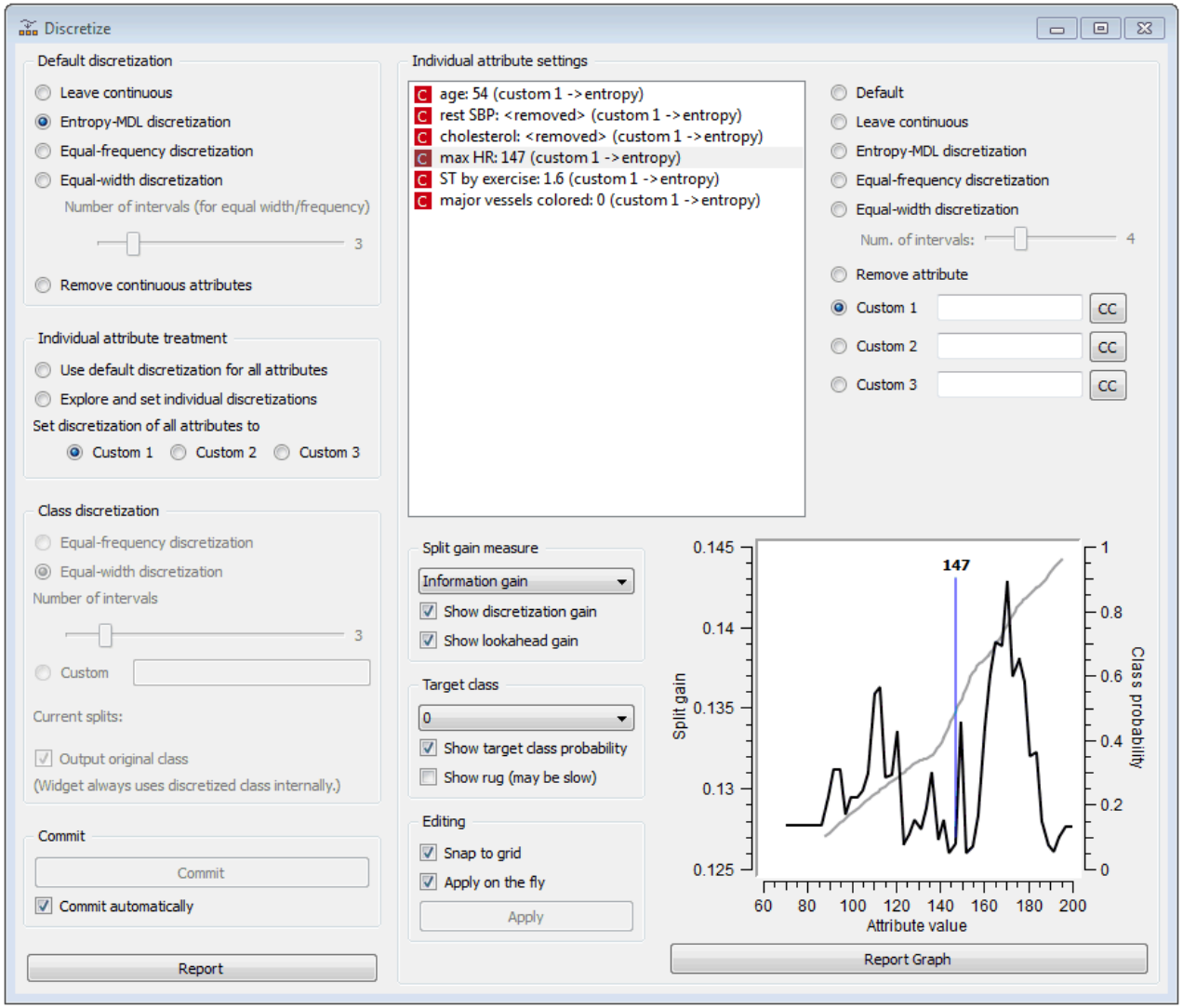

Read Excel
out
Select Attribu...
exa
exa
ori
Set Role
exa
exa
ori
Decision Tree
tra
mod
exa
Apply Model
mod
unl
lab
mod
Read Excel (2)
out

Orange

http://docs.orange.biolab.si/widgets/rst/index.html



Orange


Discretize
Default discretization
Leave continuous
Entropy-MDL discretization
Equal-frequency discretization
Equal-width discretization
Number of intervals (for equal width/frequency)
3
Remove continuous attributes
Individual attribute treatment
Use default discretization for all attributes
Explore and set individual discretizations
Set discretization of all attributes to
Custom 1
Custom 2
Custom 3
Class discretization
Equal-frequency discretization
Equal-width discretization
Number of intervals
3
Custom
Current splits:
Output original class
(Widget always uses discretized class internally.)
Commit
Commit
Commit automatically
Report
Individual attribute settings
age: 54 (custom 1 -> entropy)
rest SBP: <removed> (custom 1 -> entropy)
cholesterol: <removed> (custom 1 -> entropy)
max HR: 147 (custom 1 -> entropy)
ST by exercise: 1.6 (custom 1 -> entropy)
major vessels colored: 0 (custom 1 -> entropy)
Default
Leave continuous
Entropy-MDL discretization
Equal-frequency discretization
Equal-width discretization
Num. of intervals:
4
Remove attribute
Custom 1
Custom 2
Custom 3
CC
Split gain measure
Information gain
Show discretization gain
Show lookahead gain
Target class
0
Show target class probability
Show rug (may be slow)
Editing
Snap to grid
Apply on the fly
Apply
147
0.145
0.14
0.135
0.13
0.125
Split gain
1
0.8
0.6
0.4
0.2
0
Class probability
60
80
100
120
140
160
180
200
Attribute value
Report Graph

Autres tecnhniques

Web mining (contenu, usage, …)

Visual data mining (images)

Audio data mining (son, musique)

Data mining et requêtes d'interrogation "intelligentes"

Visualisation de données

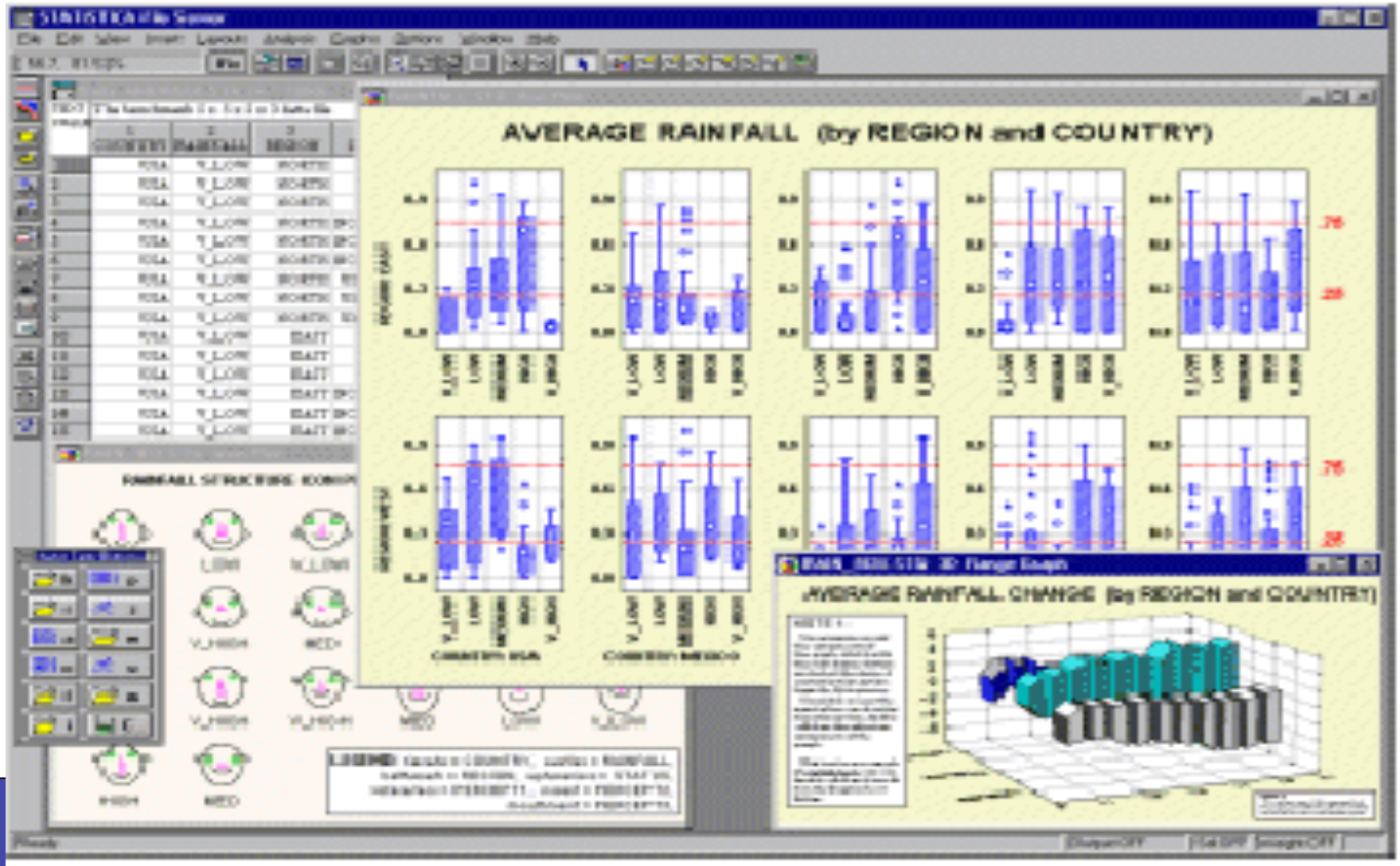
Données dans un base de données ou un entrepôt de données peuvent être visualisées :

À différents niveaux de granularité ou d'abstraction

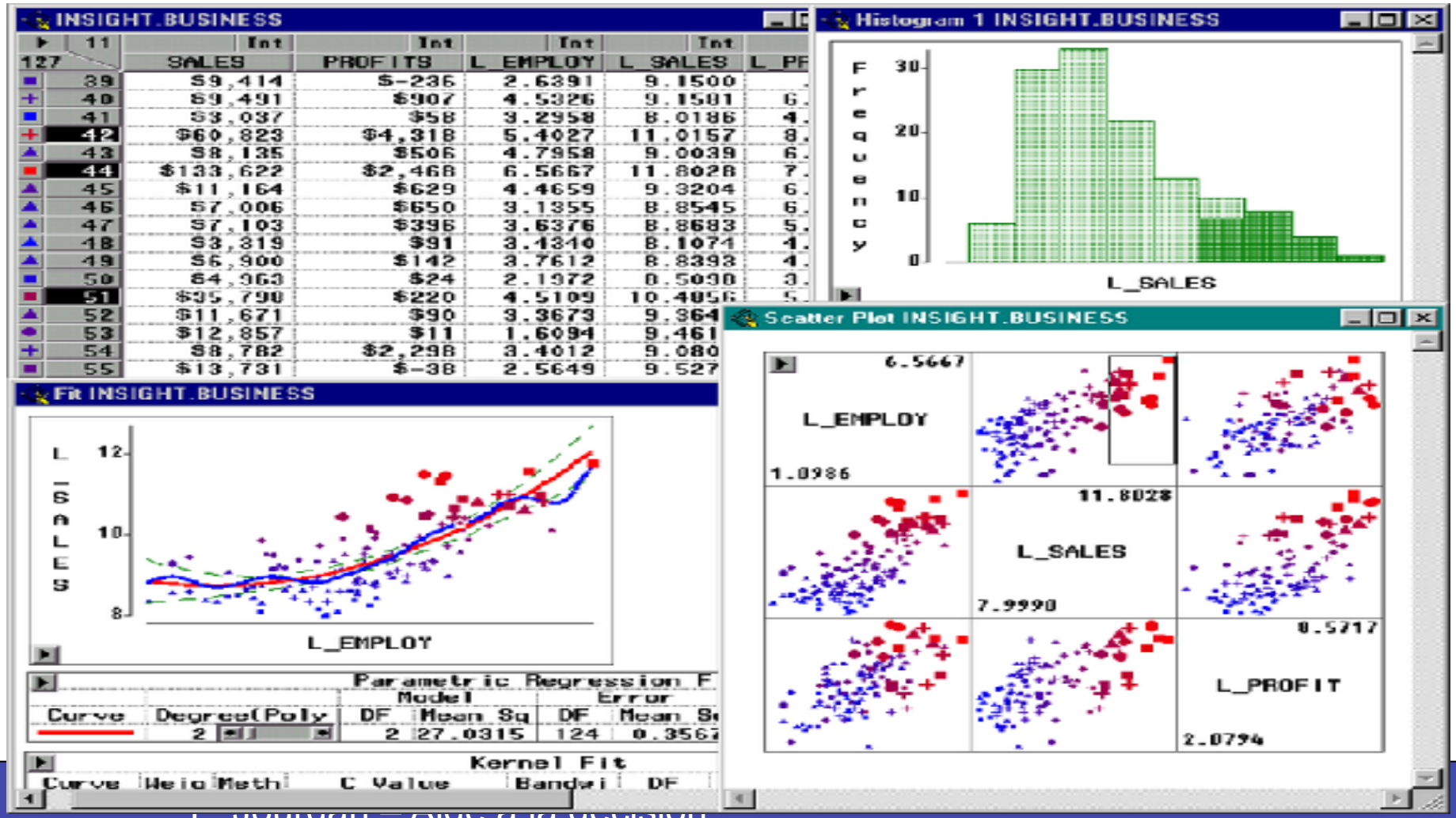
A l'aide de différentes combinaisons d'attributs ou dimensions

Résultats des outils de Data Mining peuvent être présentées sous diverses formes visuelles

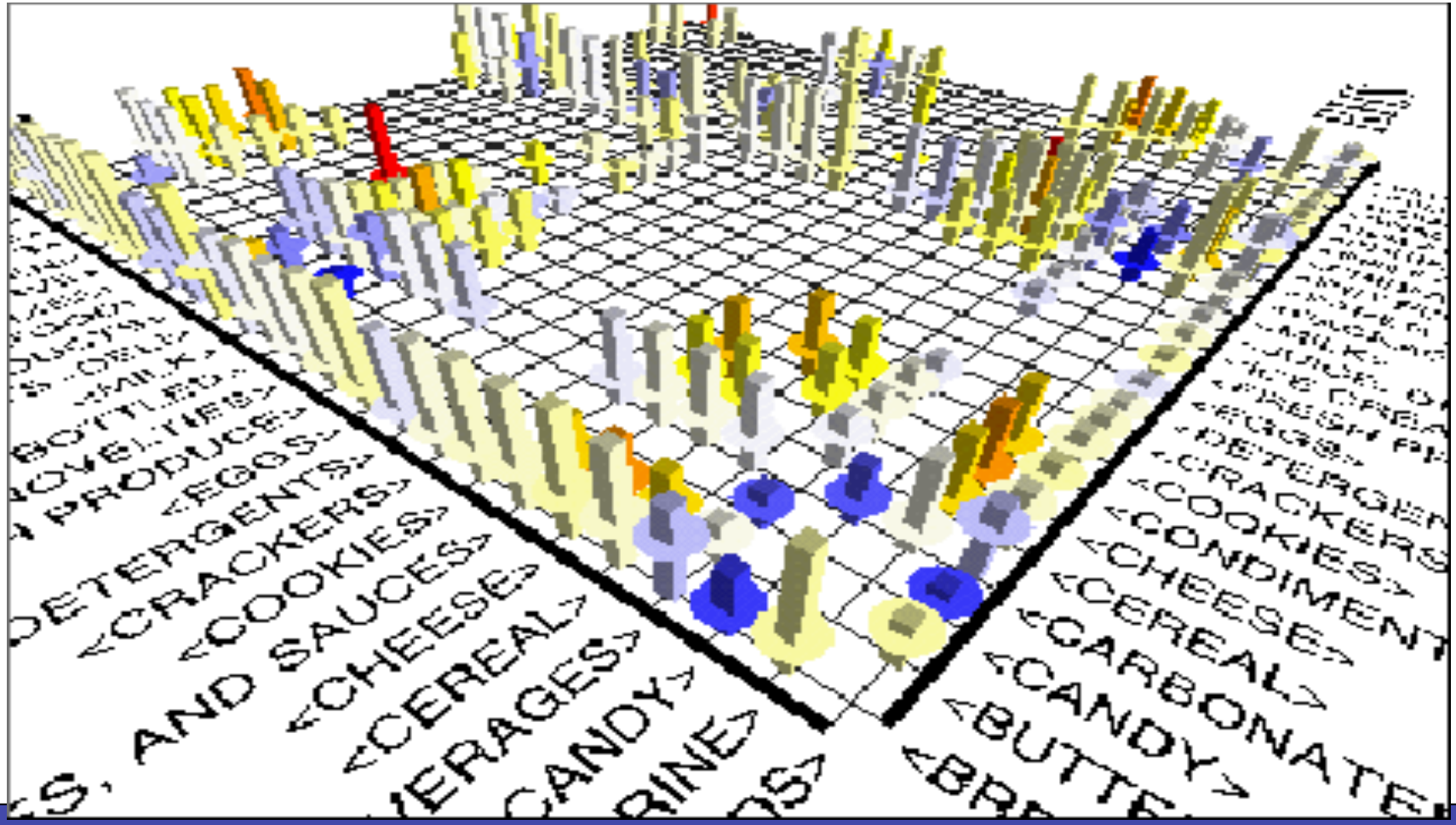
Box-plots dans StatSoft



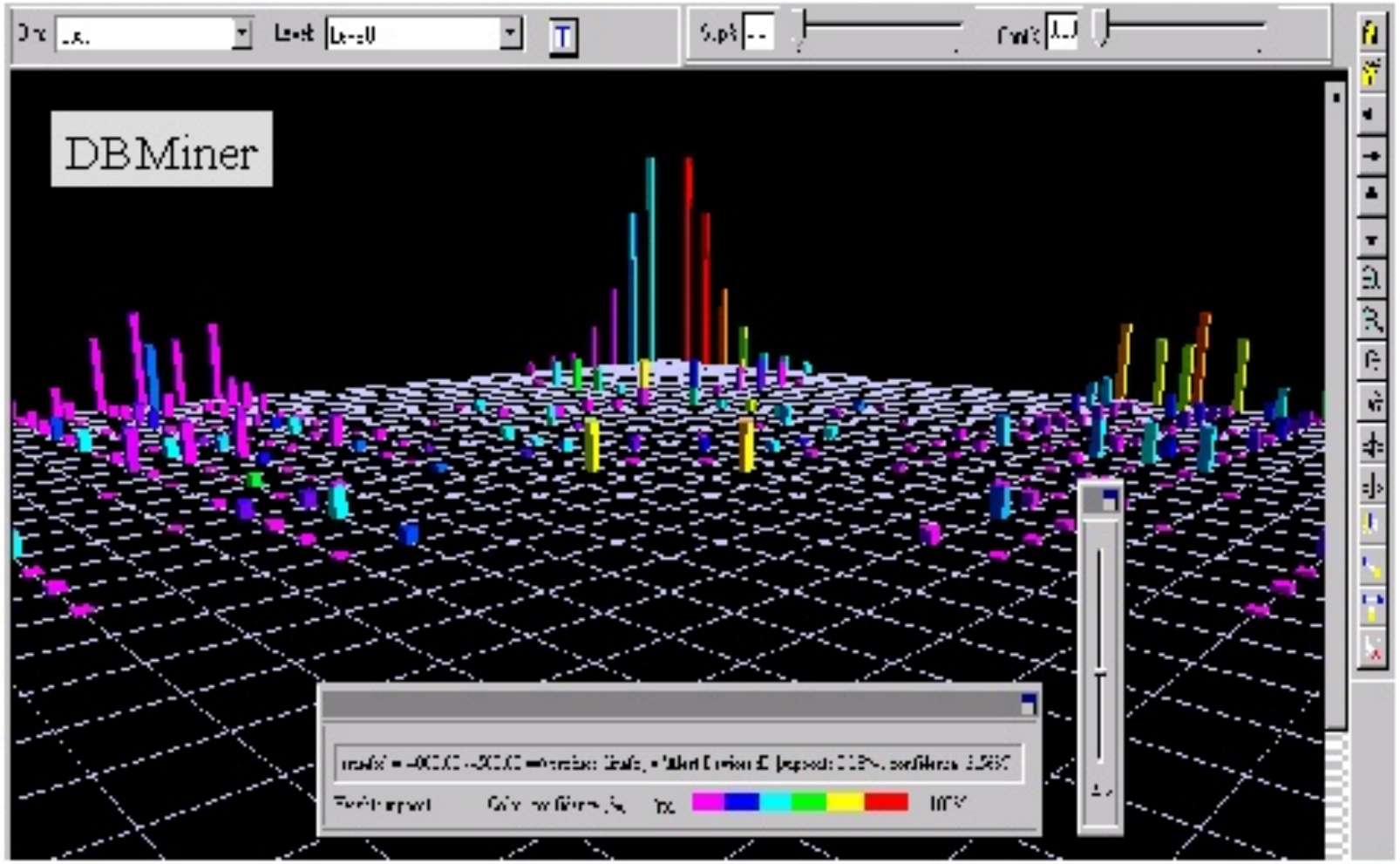
Scatter-plots dans SAS Enterprise Miner



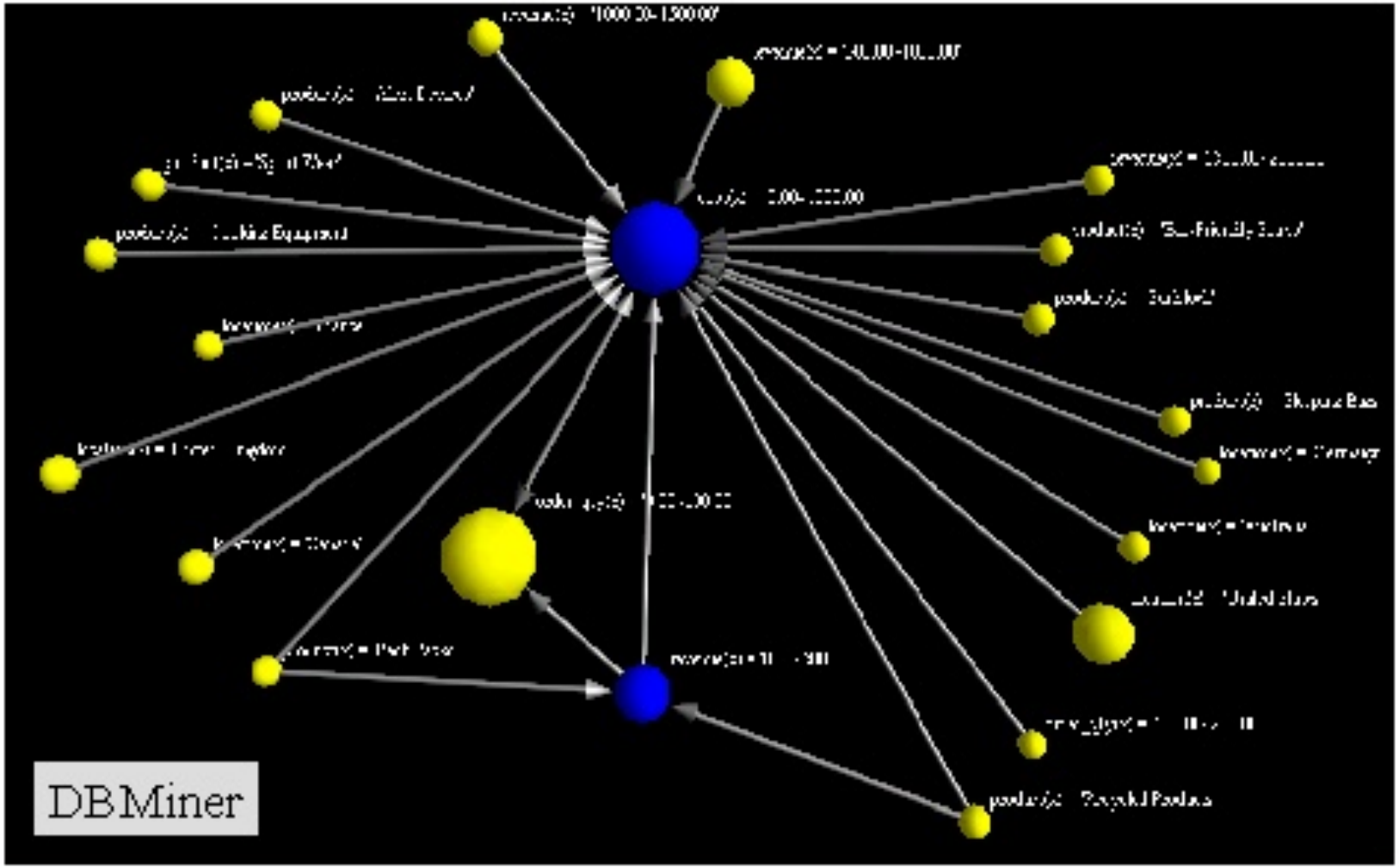
Règles d'association dans MineSet 3.0



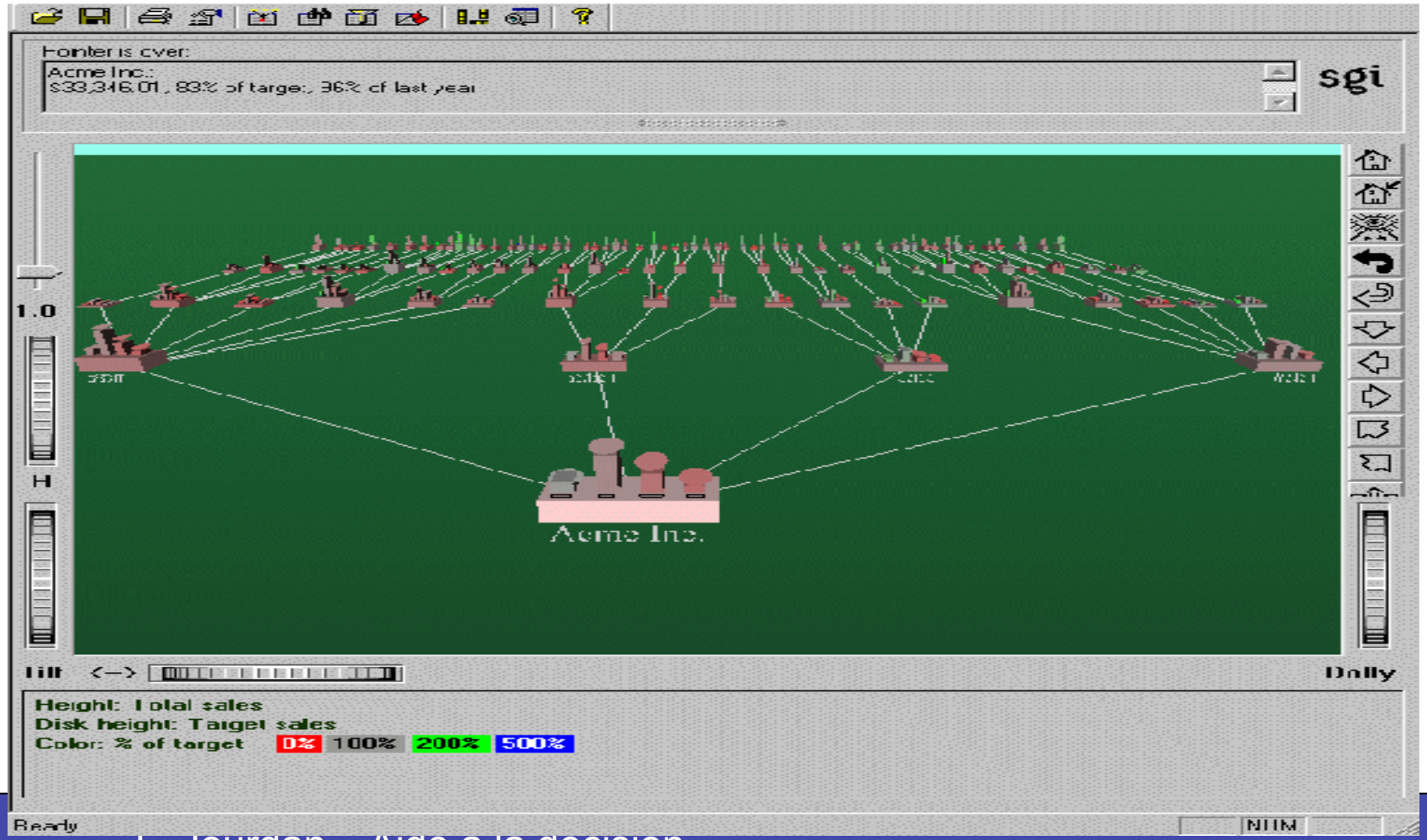
Visualization of Association Rule in Plane Form



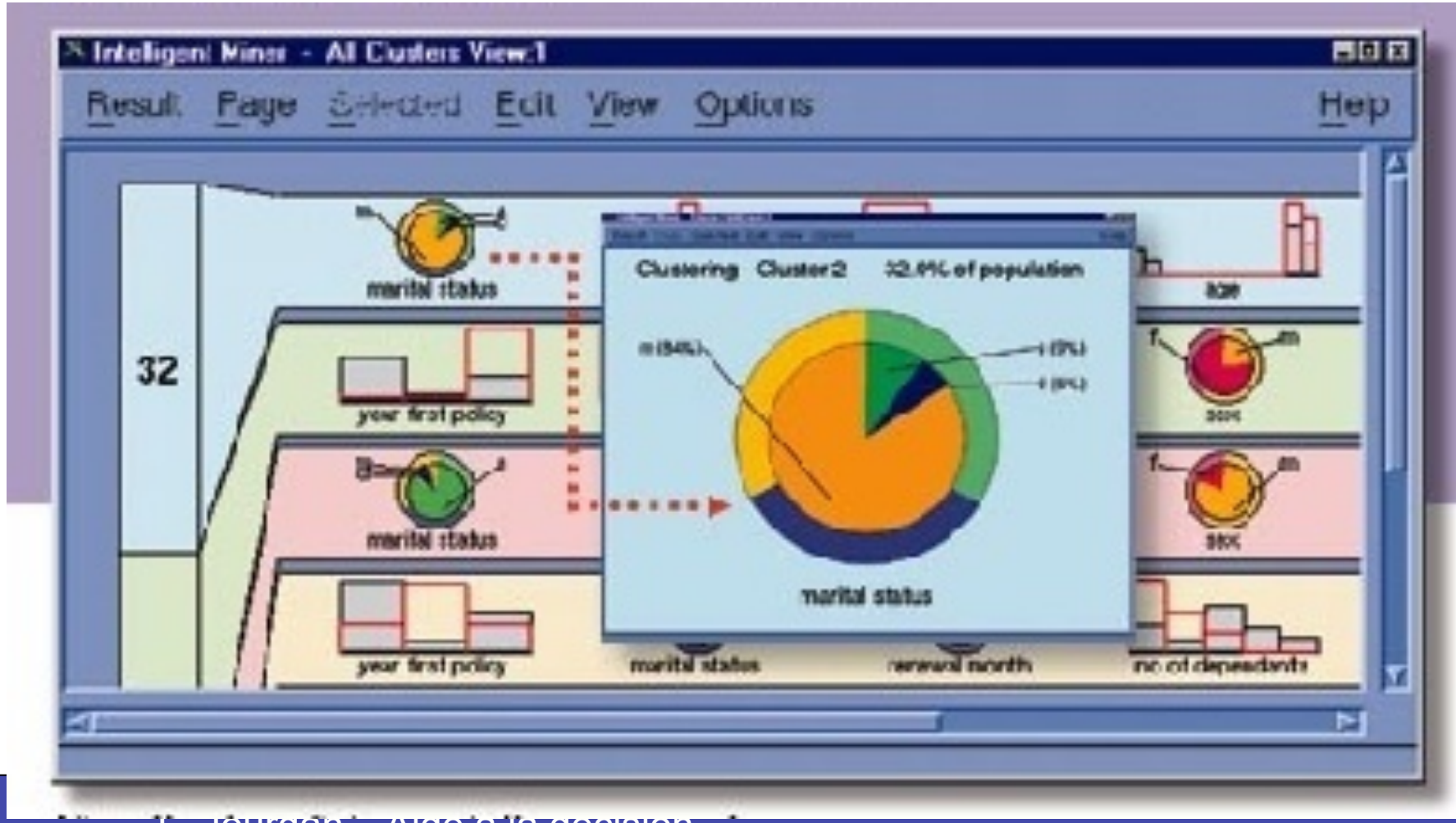
Visualization of Association Rule Using Rule Graph



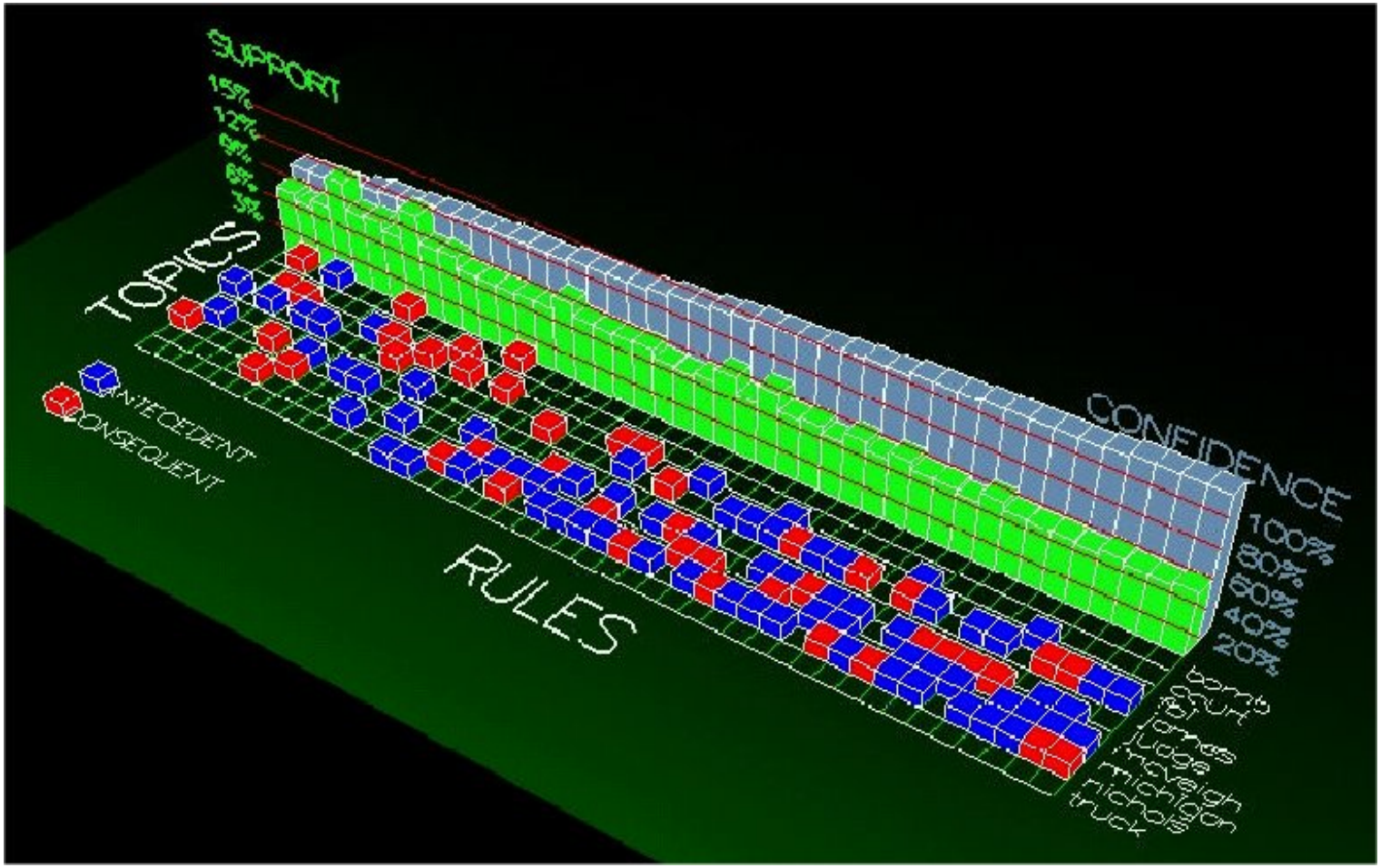
Arbres de décision dans MineSet 3.0



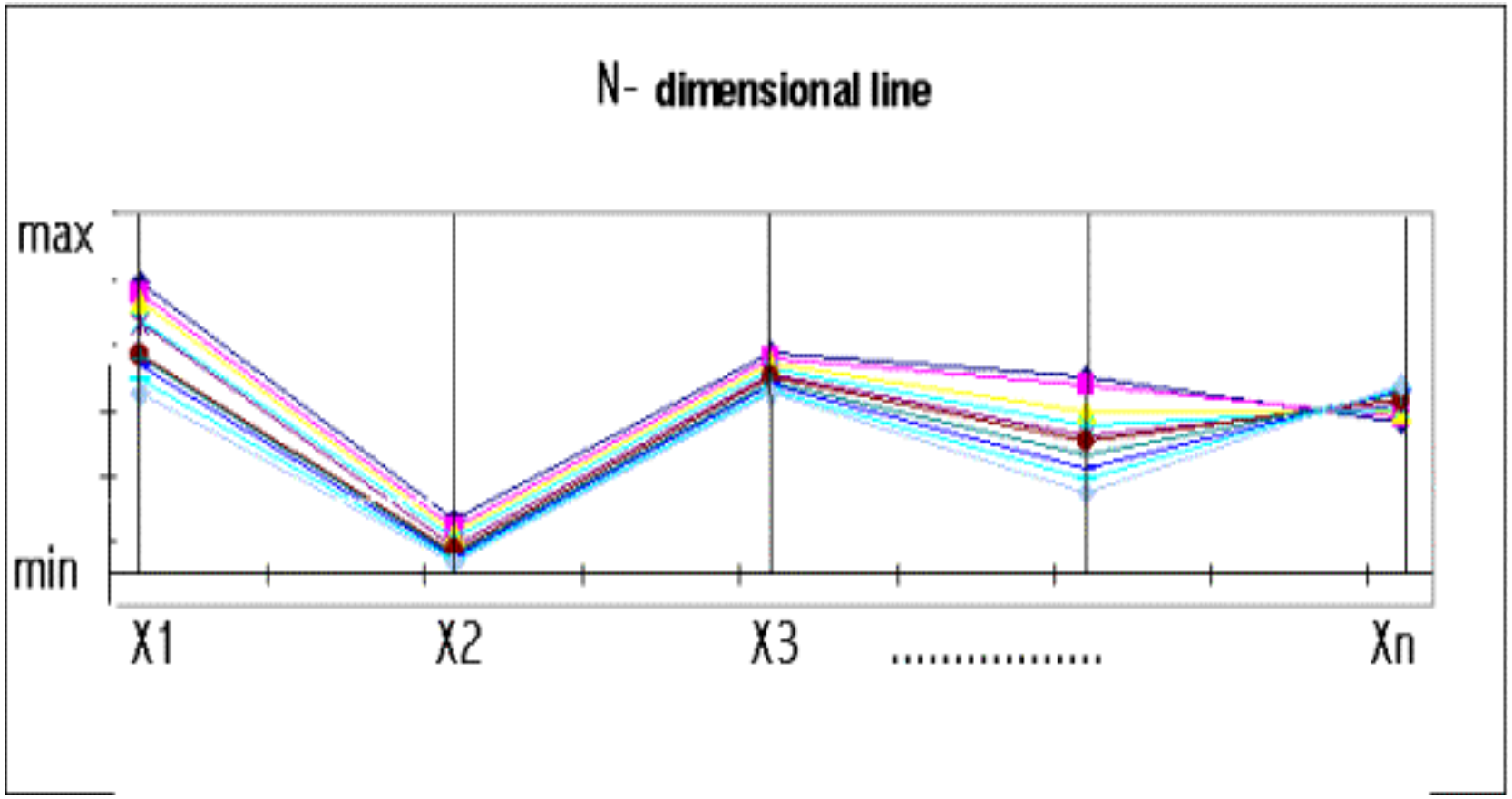
Clusters dans IBM Intelligent Miner



Règles d'association : La visualisation 3D



Règles d'association : Le N-Dimensional Line



Règles d'association : Le Double Decker Plot