

# Tanagra

Un logiciel de Data Mining gratuit pour  
l'enseignement et la recherche

Ricco RAKOTOMALALA

Université Lyon 2

Laboratoire ERIC

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

## Ricco – Qui est-il ? Que fait-il ?

Enseignant chercheur – CNU 27 – Informatique

Université Lumière Lyon 2

Culture Économétrie (Statistique)

Thèse Apprentissage automatique – Data Mining

- Arbres de décision, Sélection de variables, Échantillonnage, ...
- **Applications** (*classement de protéines, classement de planctons, reconnaissance de la langue, etc.*)

Développement et diffusion de logiciels libres (TANAGRA, *SIPINA*)

Rédaction et diffusion de didacticiels

Rédaction et diffusion de fascicules de cours

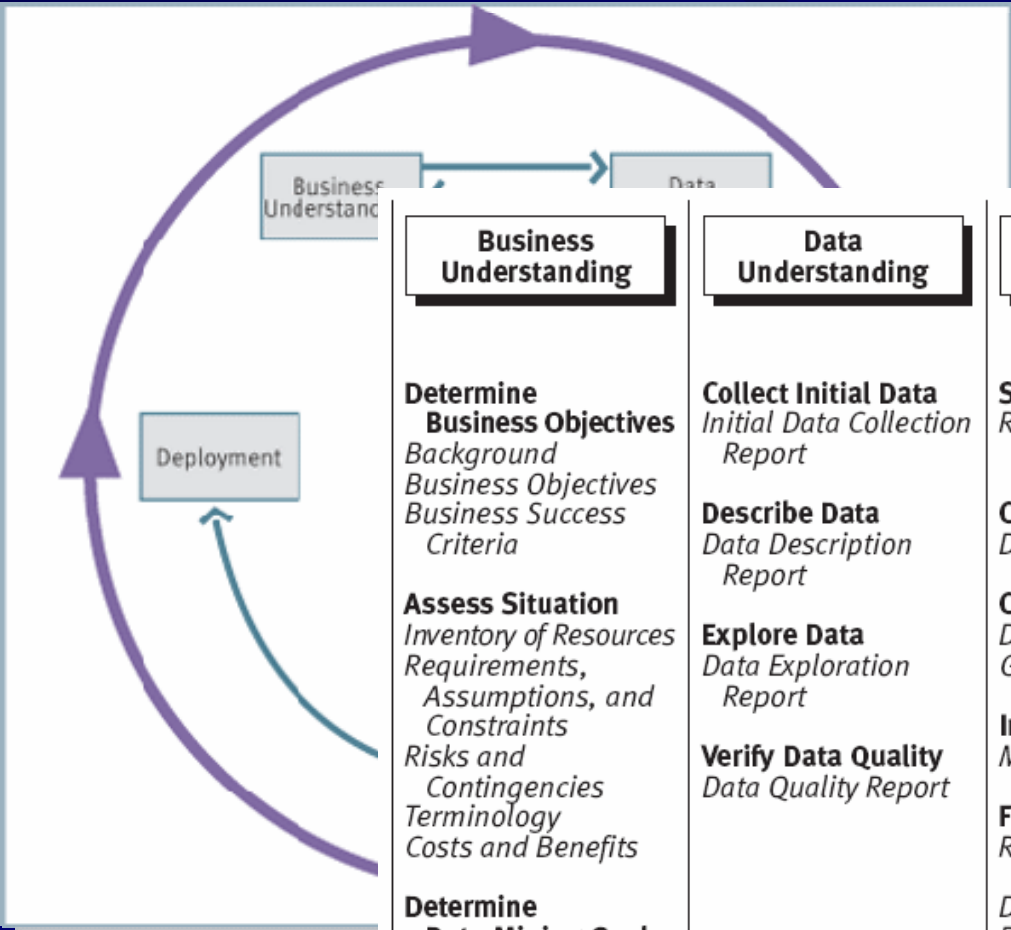
# Plan

1. Data Mining
2. Pourquoi le logiciel libre dans le data mining
3. Tanagra – Spécification, développement, promotion
4. Et les autres logiciels libres ?  
*Knime, Orange, R, RapidMiner, Weka,...*
5. Démonstration
6. Conclusion

# 1. Data Mining

ECD : Extraction de connaissances à partir de données  
 (Knowledge Discovery in Databases)

} Data Mining



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

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

## Data Mining – Est-ce vraiment novateur ?

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

Data Mining : Une nouvelle façon de faire de la statistique ? (G. Saporta)

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 ?

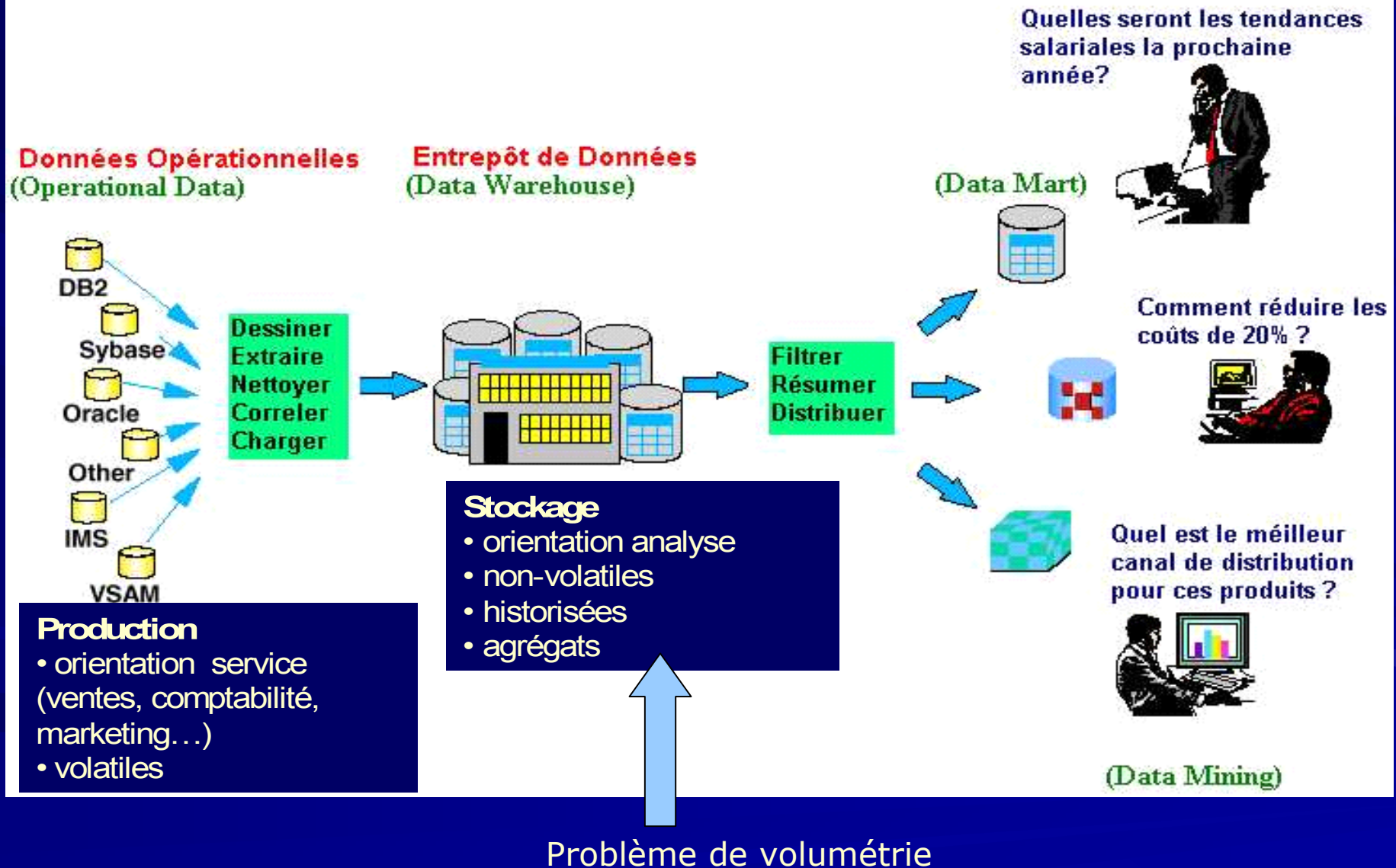
<http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd41.htm>

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

Travailler sur des entrepôts de données

Faire partie intégrante du flux d'informations dans l'entreprise

## Construire une Infrastructure d'Information Intelligente pour l'Entreprise





Mixer des techniques d'horizons différents  
Apprentissage automatique, Reconnaissance de formes,  
Statistique, Analyse de données, ...

**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				

Le point de ralliement est le tableau  
« attribut valeur ».  
Même si, de plus en plus, on essaie d'aller plus loin  
en traitant des tableaux de structure plus  
élaborés, plus adaptées au problème à traiter (ex.  
données transactionnelles, données séquentielles,  
données complexes en XML, etc.)

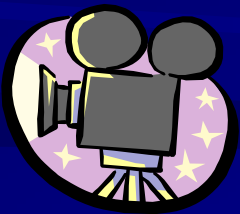
**Informatique  
(Intelligence artificielle)**  
Apprentissage symbolique  
Reconnaissance de formes  
  
Une étape de l'intelligence artificielle  
*Réseaux de neurones, algorithmes génétiques...*

**Informatique  
(Base de données)**  
Exploration des bases de données  
  
Volumétrie  
*Règles d'association, motifs fréquents, ...*

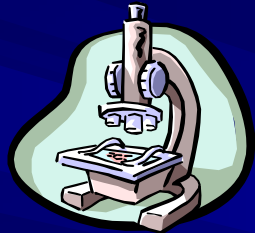


# Traitement des données non structurées

Textes, images, etc... autre que le simple « attribut-valeur »



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

Analyse des mammographies

Etc.

## 2. Data Mining et logiciel libre

*Attention, les informaticiens (et internet) arrivent...*

# Quel espace pour les logiciels libres ?

## Aspects du data mining prolifiques en développement

### Développer des méthodes au cœur des entrepôts de données

Les B.D. sont surtout intéressés par le développement des plate-formes B.I.

Proximité très (trop) forte avec les applications industrielles (ORACLE, SQL-Server...)

Développement lourds, peu valorisants pour l' « apprentissage automatique » (publications)

*Récupération d'outils existants. Ex. intégration de WEKA dans PENTAHO...*

### Traitement des données non structurées

Trop spécifique – Impossible de développer un outil générique

Proximité des applications industrielles

### Développer des outils génériques de traitement de données

Intégrer des méthodes avec des finalités (origines) différentes

Pouvoir les faire coopérer entre elles

Tester et diffuser une nouvelle méthode publiée

Développement de plate-forme peu onéreuse, c'est le développement des algorithmes de traitements qui est difficile (ex. *RAPIDMINER* et *KNIME* reposent en partie sur le moteur *WEKA*)



# Quel public pour le logiciel libre de data mining ?

Qui sont les utilisateurs, quels sont leurs besoins ?

## Un logiciel pour l'enseignement et le profil « utilisateur - praticien » du Data Mining

Les cours, explication des méthodes, outil pédagogique

Illustrer les techniques en cours, les mettre en oeuvre en TD

Sans connaissances spécifiques (langage de prog., etc.) - Former sur le fond et non la forme

Avec un niveau de qualité conforme aux « standards » du domaine

Les études « réelles » - les « dossiers » - les chercheurs des autres domaines (biologie, médecine, etc.)



## Une plate-forme pour la recherche

Plate-forme d'expérimentation pour tests à grande échelle

Implémenter ses méthodes (et les tester)

Les comparer (toutes choses égales par ailleurs i.e. dans le même environnement)

Les diffuser (pour d'autres, à des fins d'expérimentation, de comparaison)

Une publication n'est crédible que si reproductible (données, outils)

## Un outil pédagogique pour l'apprentissage de la programmation

Spécifications et conception de ce type de logiciel - Apprendre par l'exemple

Connaître les outils et les bibliothèques types

Sujets de stages pour les étudiants



## Protéger les chercheurs, protéger les utilisateurs

A qui appartient un logiciel développé par un enseignant-chercheur ?

Est-ce le même statut que pour les ouvrages ?

Pouvoir développer sans contraintes (chercheur)

Pouvoir utiliser sans mauvaises surprises (utilisateur)

## Diffusion du logiciel = valider les publications

Logiciels accessibles à tous → Comparaison et vérification des résultats

Reproduire « exactement » les expérimentations

## Comparer le code = comparer les implémentations

Comparer les interprétations d'un même problème (ex. Relieff WEKA)

Lecture du code par d'autres chercheurs (ex. Naive Bayes classifier)

Optimiser le code avec différentes versions

## Outil ouvert = Outil vivant

Introduire ses propres algorithmes

Discuter sur la base de prototypes et d'évolutions

Monter et partager des bibliothèques types (ex. générateurs aléatoires, fonctions de répartition, les fameux packages...)



# Logiciel de data mining

## Quelles fonctionnalités implémenter



**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

— Logiciels commerciaux  
— Prototypes de recherche



# Logiciel de data mining

Quel mode opératoire ?

## Logiciels pilotés par menu (STATISTICA, OPEN STAT, SIPINA, ...)

- (+) Organisation de type « tableur »
- (+) Rapidité de prise en main
- (-) Enchaînement « à la main » des traitements
- (-) Pas de trace des opérations effectuées
- (-) Et donc reproductibilité difficile des traitements

## Ligne de commande (SAS, S-PLUS, R, ...)

- (+) Souplesse et puissance de la programmation
- (+) Sauvegarde des traitements, reproductibilité
- (-) Apprentissage d'un langage

## Filière (diagramme de traitements) – Estampillé « Data Mining »

- (+) Programmation « visuelle » - Pas d'apprentissage
- (+) Enchaînement des traitements
- (+) Sauvegarde des traitements, reproductibilité
- (-) Pas la puissance d'un « vrai » langage de programmation

SPAD, SAS Enterprise Miner,  
SPSS Clementine, S-PLUS  
Insightfull Miner, STATISTICA  
Data Miner, ...

KNIME, ORANGE, RAPIDMINER,  
TANAGRA, WEKA

# Exemple de pilotage par menu Sipina

The screenshot displays the Sipina Research Version software interface. The main window is titled "Sipina Research Version" and features a menu bar with options: Induction method, Analysis, Tree management, View, Window, and Help. A "Learning set editor" table is visible, showing columns for sep\_length, sep\_width, pet\_length, pet\_width, and type. A context menu is open over the "Analysis" menu, listing options such as "Define class attribute...", "Set weight field...", "Learning...", "Classification", "Test...", "LIFT -- ROC curve...", "Error measurements", "Feature selection", and "Personal tests". The "Classification" option is selected, and a sub-menu is open, showing "generate scores" with options for "on same dataset" and "on other dataset". A decision tree diagram is displayed in the foreground, showing splits based on "pet\_length" and "pet\_width" attributes. The tree structure is as follows:

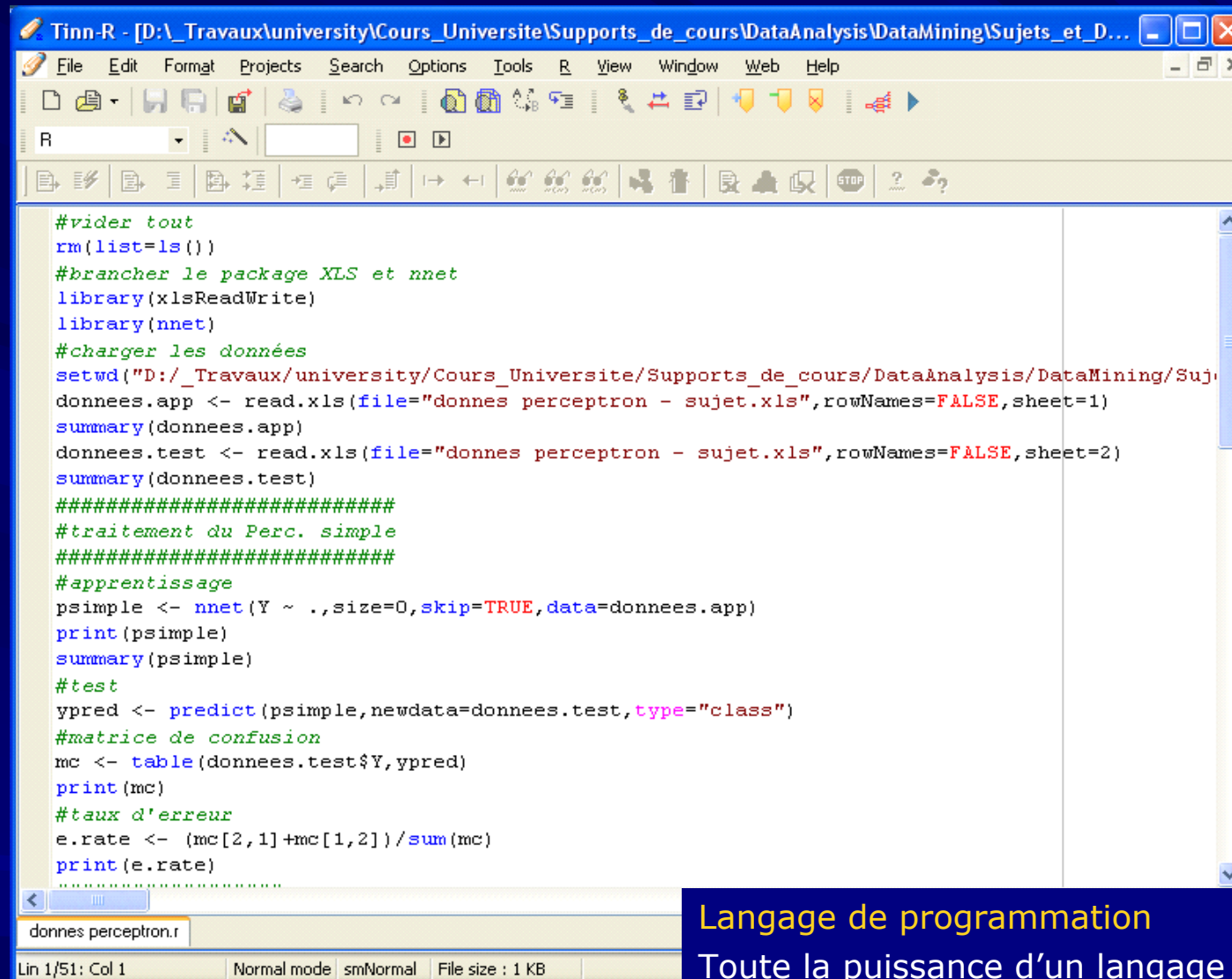
- Root node: pet\_length < 2.45 (50 (33%), Iris-setosa) and pet\_length >= 2.45 (50 (33%), Iris-versicolor and 50 (33%), Iris-virginica)
- Left child of root: pet\_width < 1.75 (0 (0%), 49 (91%), 5 (09%)) and pet\_width >= 1.75 (0 (0%), 1 (02%), 45 (98%))

The bottom status bar indicates the method used: Improved ChAID (Tschuprow Goodness of Split).

**Pilotage par menu**  
Simple au premier abord mais ingérable dès que le logiciel gagne en complexité  
Impossible de garder la trace d'une analyse complète et donc de la reproduire  
Exige une documentation complète et constamment à jour (Open Stat & Stat 4U sont dans la même situation)

# Exemple de ligne de commande + langage de programmation

R



```
Tinn-R - [D:\_Travaux\university\Cours_Universite\Supports_de_cours\DataAnalysis\DataMining\Sujets_et_D...
File Edit Format Projects Search Options Tools R View Window Web Help
R
#vider tout
rm(list=ls())
#brancher le package XLS et nnet
library(xlsReadWrite)
library(nnet)
#charger les données
setwd("D:/_Travaux/university/Cours_Universite/Supports_de_cours/DataAnalysis/DataMining/Sujets_et_D...
donnees.app <- read.xls(file="donnees perceptron - sujet.xls",rowNames=FALSE,sheet=1)
summary(donnees.app)
donnees.test <- read.xls(file="donnees perceptron - sujet.xls",rowNames=FALSE,sheet=2)
summary(donnees.test)
#####
#traitement du Perc. simple
#####
#apprentissage
psimple <- nnet(Y ~ .,size=0,skip=TRUE,data=donnees.app)
print(psimple)
summary(psimple)
#test
ypred <- predict(psimple,newdata=donnees.test,type="class")
#matrice de confusion
mc <- table(donnees.test$Y,ypred)
print(mc)
#taux d'erreur
e.rate <- (mc[2,1]+mc[1,2])/sum(mc)
print(e.rate)
.....
donnes perceptron.r
Lin 1/51: Col 1 Normal mode smNormal File size : 1 KB
```

Langage de programmation

Toute la puissance d'un langage de programmation

L'accès au langage est une barrière à l'entrée qui rebute certains

# Exemple de diagramme de traitements

Tanagra

The screenshot shows the TANAGRA 1.4.32 interface. On the left, a workflow diagram titled 'Default title' shows a sequence of components: Dataset (breast\_sorted\_on\_mitoses.txt), Define status 1, Supervised Learning 1 (Linear discriminant analysis), Cross-validation 1, Runs filtering 1, Supervised Learning 2 (Linear discriminant analysis), Cross-validation 2, Stepdisc 1, Supervised Learning 3 (Linear discriminant analysis), Cross-validation 3, Principal Component Analysis 1, Define status 2, and Supervised Learning 4 (Linear discriminant analysis), followed by Cross-validation 4. On the right, a table shows the overall cross-validation error rate: MIN 0.0420, MAX 0.0420, Trial 1 Err rate 0.0420. Below this is a table for 'Overall cross-validation error rate' with columns for Error rate (0.0420), Values prediction, and Confusion matrix. The Confusion matrix table is as follows:

Values prediction			Confusion matrix			
Value	Recall	1-Precision		begin	malignant	Sum
begin	0.9758	0.0390	begin	444	11	455
malignant	0.9234	0.0482	malignant	18	217	235
			Sum	462	228	690

Computation time : 905 ms.  
Created at 01/09/2009 12:50:45

At the bottom, a 'Components' panel lists various machine learning and statistical methods: Binary logistic regression, C4.5, C-PLS, C-RT, CS-CRT, CS-MC4, C-SVC, Decision List, ID3, K-NN, Linear discriminant analysis, Log-Reg TRIRLS, Multilayer perceptron, Multinomial Logistic Reg, and Naive bayes.

## Diagramme de traitements

« Programmation » visuelle – Enchaînement des traitements

*Mais pas toutes les fonctionnalités d'un langage de programmation*

Mise à jour facilitée par adjonction de composants

Garder une trace de l'analyse et pouvoir la sauvegarder

Possibilité de fragmenter la documentation par « composants »

C'est le **standard** actuel

# Exemple de diagramme de traitements

Knime

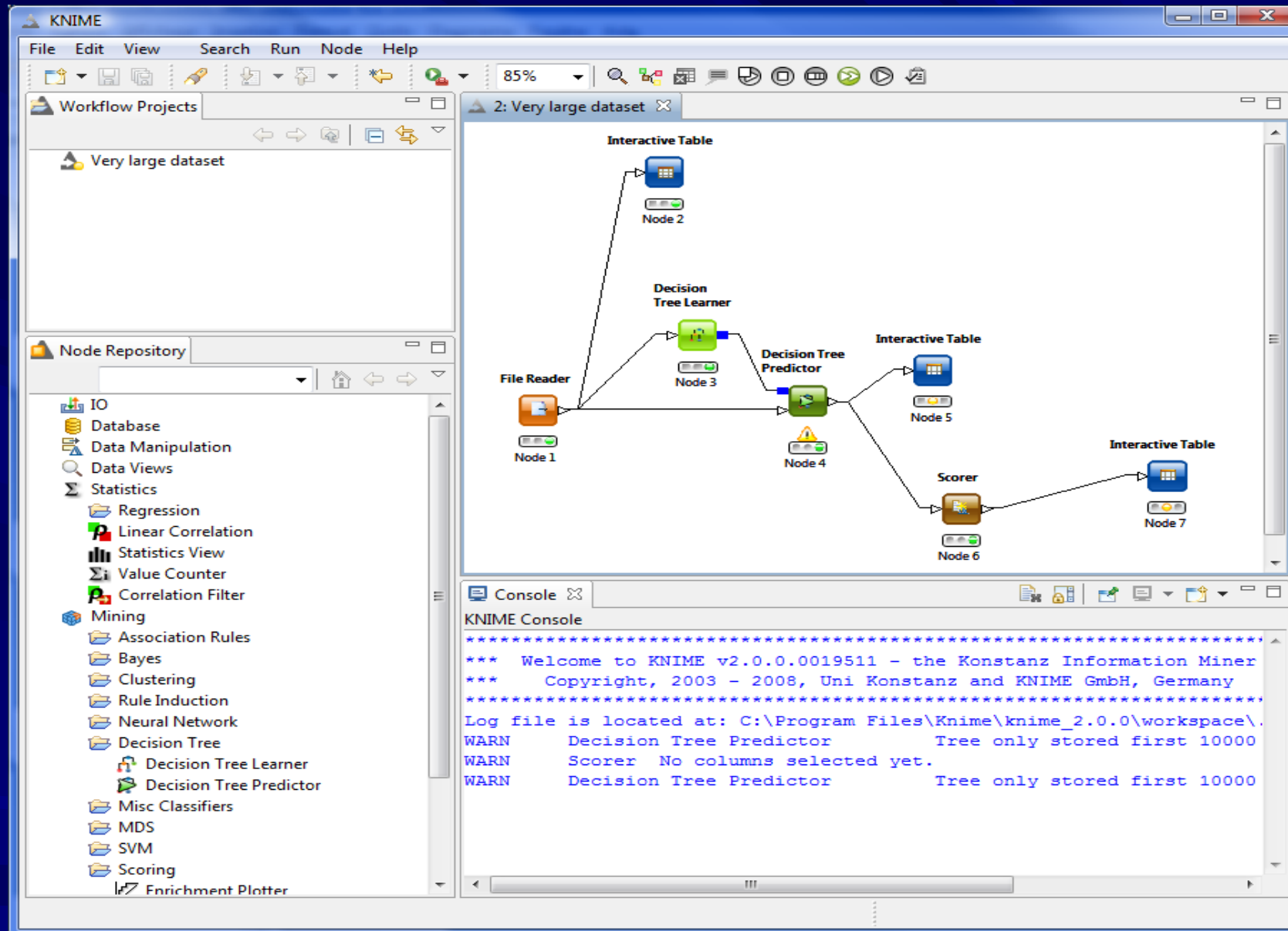


Diagramme de traitements

Une autre manière de présenter les diagrammes de traitements

KNIME est un des très rares à savoir représenter une boucle (notion de « méta composant »)



# 3. Tanagra



# Tanagra

Définir un cahier des charges aussi précis que possible

## Miser sur la simplicité d'utilisation

Installation simplifiée – Pas de serveurs lourds à installer

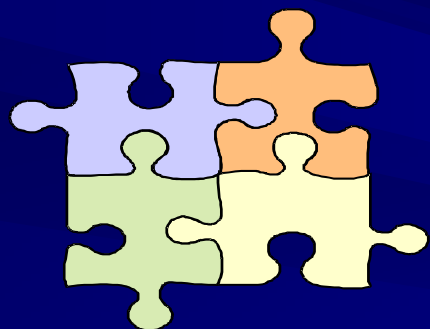
Gestion simplifiée des données - Format texte et accès au format tableur

Fonctionnement par diagramme de traitements

Couvrir les statistiques, l'analyse de données et le data mining. De manière unifiée.

Résultats lisibles, en adéquation avec les « standards »

Interfaçage avec les tableurs (Excel, Open Office Calc)



## Mettre définitivement de côté les aspects « professionnels »

Interfaçage fort avec les SGBD

Déploiement et mise en production des résultats

Reporting dynamique et performant

Exploration graphique évoluée et interactive des données

## Simplicité également pour le programmeur

Simplifier à l'extrême le code permettant d'ajouter une nouvelle méthode d'analyse

Minimiser le code dédié à la gestion des données et de l'interface

Pouvoir intégrer facilement n'importe quelle technique traitant des tableaux « individus x variables »

# Simplicité pour les utilisateurs

## Installation simplifiée et automatisée



### Tout doit être automatisé

L'utilisateur ne doit jamais avoir à intervenir à l'installation

Attention aux bibliothèques externes (SGBD, TCL/TK, PYTHON, etc.)

Choisir la configuration au pire cas

### Exclure les bibliothèques externes compilées / payantes

Bibliothèque externe compilée = dépendance accrue

Bibliothèque payante = pieds et poings liés (y compris sur les architectures)

Miser sur des versions stables et sources libres

Attention à la gestion des mises à jour

### Mettre des exemples de traitements (ex. RapidMiner)

L'utilisateur lance toujours « pour voir » sans lire la documentation

# Simplicité pour les utilisateurs

## Définir les traitements

The screenshot shows the TANAGRA 1.4.32 software interface. The main window displays a workflow diagram on the left and a results window on the right. The workflow diagram shows a sequence of steps: Dataset (breast\_sorted\_on\_mitoses.txt), Define status 1, Supervised Learning 1 (Linear discriminant analysis), Cross-validation 1, Runs filtering 1, Supervised Learning 2 (Linear discriminant analysis), Cross-validation 2, Stepdisc 1, Supervised Learning 3 (Linear discriminant analysis), Cross-validation 3, Principal Component Analysis 1, Define status 2, and Supervised Learning 4 (Linear discriminant analysis). The results window displays the overall cross-validation error rate and a confusion matrix.

**Enchaînement des traitements**

**Fenêtre de visualisation des résultats**

MIN	0.0420
MAX	0.0420
Trial	Err rate
1	0.0420

**Overall cross-validation error rate**

Error rate	0.0420					
Values prediction			Confusion matrix			
Value	Recall	1-Precision		begin	malignant	Sum
begin	0.9758	0.0390	begin	444	11	455
malignant	0.9234	0.0482	malignant	18	217	235
			Sum	462	228	690

Computation time : 905 ms.  
Created at 01/09/2009 12:50:45

**Components**

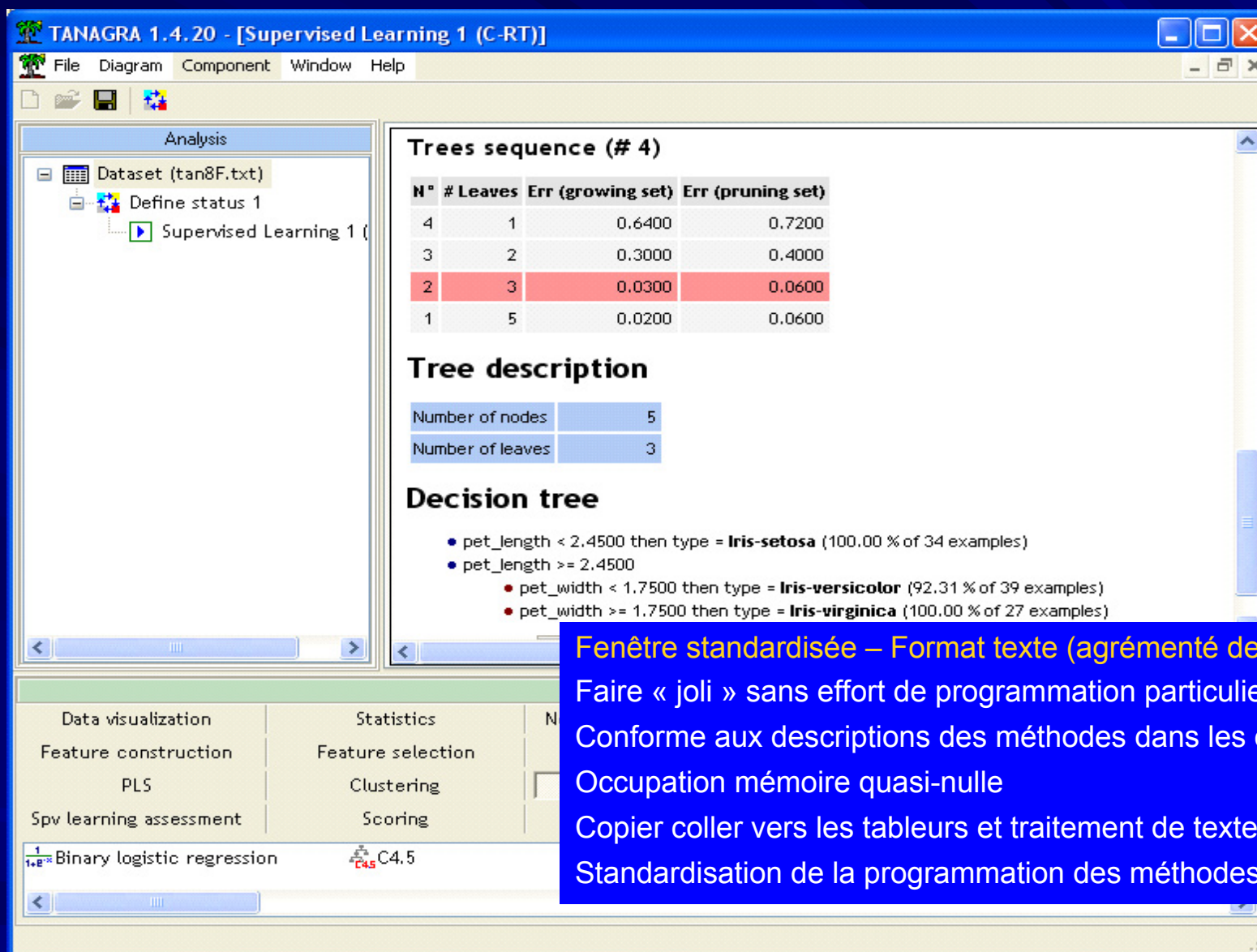
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construction	Feature selection
Regression	Factorial analysis	PLS	Clustering	Spv learning	Meta-spv learning
Spv learning assessment	Scoring	Association			

Binary logistic regression, C4.5, C-PLS, C-RT, CS-CRT, CS-MC4, C-SVC, Decision List, ID3, K-NN, Linear discriminant analysis, Log-Reg TRIRLS, Multilayer perceptron, Multinomial Logistic Reg, Naive bayes

Composants de calcul  
Méthodes de data mining

# Simplicité pour les utilisateurs

## Standardisation des affichages



The screenshot shows the TANAGRA 1.4.20 interface. The main window displays the results of a supervised learning analysis. The left pane shows the project structure: Dataset (tan8F.txt) -> Define status 1 -> Supervised Learning 1. The right pane shows the results for 'Trees sequence (# 4)'. A table lists the number of leaves, error on the growing set, and error on the pruning set for different tree sizes. The row for 3 leaves is highlighted in red. Below the table, the 'Tree description' shows the number of nodes (5) and leaves (3). The 'Decision tree' section shows the rules: pet\_length < 2.4500 then type = Iris-setosa (100.00 % of 34 examples); pet\_length >= 2.4500 then type = Iris-versicolor (92.31 % of 39 examples) or Iris-virginica (100.00 % of 27 examples).

N°	# Leaves	Err (growing set)	Err (pruning set)
4	1	0.6400	0.7200
3	2	0.3000	0.4000
2	3	0.0300	0.0600
1	5	0.0200	0.0600

**Tree description**

Number of nodes	5
Number of leaves	3

**Decision tree**

- pet\_length < 2.4500 then type = **Iris-setosa** (100.00 % of 34 examples)
- pet\_length >= 2.4500
  - pet\_width < 1.7500 then type = **Iris-versicolor** (92.31 % of 39 examples)
  - pet\_width >= 1.7500 then type = **Iris-virginica** (100.00 % of 27 examples)

The bottom of the interface shows a list of methods: Binary logistic regression, C4.5, etc.

Fenêtre standardisée – Format texte (agrémenté de HTML)

Faire « joli » sans effort de programmation particulier

Conforme aux descriptions des méthodes dans les ouvrages

Occupation mémoire quasi-nulle

Copier coller vers les tableurs et traitement de texte

Standardisation de la programmation des méthodes

# Simplicité pour les programmeurs

## Vive la programmation objet

### Classes de calcul

```
UCalcTreeStructureDefinition,  
UCalcSpvTreeDefinition;  
  
TYPE  
//feuille  
TSplitLeafSpvC45 = class(TSplitLeafSpv)  
    end;  
  
//split  
TSplitAttributSpvC45 = class(TSplitAttributSpv)  
    protected  
    function getClassSplitLeaf(): TClassSplitLeaf; override;  
    function ComputeGoodness(): double; override;  
    function ComputeAcceptSplit(): boolean; override;  
    end;  
  
//liste de splits  
TLstSplitAttSpvC45 = class(TLstSplitAttSpv)  
    protected  
    function getClassSplitAttribut(): TClassSplitAttribut; override;  
    end;  
  
//noeud de l'arbre  
TMLTreeNodeSpvC45 = class(TMLTreeNodeSpv)  
    private  
    //calcul de l'écart à l'erreur pour avoir la borne -- taille sommet, contre-  
    //extrait du livre de Quinlan, "Programs for Machine Learning..."  
    function addErrs(N,CE,CF: double): double;  
    protected  
    procedure AssignConclusion(); override;  
    function isNoSplitNeeded(): boolean; override;  
    function getClassLstSplitAttributes(): TClassLstSplitAttributes; override;  
    public  
    //erreur pessimiste (c'est le nombre d'erreur ici !!!) -- calculée lors de l'  
    FPessimisticErr: double;  
    //savoir si un noeud est prunable, i.e. a des enfants, et ce sont tous des fe  
    function isPrunable(): boolean;  
    //erreur pessimiste  
    property pessimistic: double read FPessimisticErr;  
    end;  
  
//structure d'arbre  
TMLTreeStructureSpvC45 = class(TMLTreeStructureSpv)  
    protected  
    function getClassMLTreeNode(): TClassMLTreeNode; override;  
    public  
    //ce qui est spécifique à C4.5  
    procedure PostPruning(); override;  
    end;
```

1: 1

Insertion

Code/

```
D:\Travaux\personal\Programmation\Tanagra\Source\Diagram\MLComponents\SpvLearning\SpvAlgorithm\DecisionT...
UFrmMainForm | UCalcSpvTreeCART | UCalcSpvTreeC45 | UCompSpvCSTreeC45
TYPE
{générateur de composant C4.5}
TMLGCompSpvCSTreeC45 = class(TMLGCompSpvTree)
public
function    GetClassMLComponent: TClassMLComponent; override;
end;

{le composant SpvTree C4.5}
TMLCompSpvCSTreeC45 = class(TMLCompSpvTree)
protected
function    getClassOperator: TClassOperator; override;
end;

{l'opérateur C4.5}
TOPSpvCSTreeC45 = class(TOpSpvTree)
protected
function    getClassParameter: TClassOperatorParameter; override;
function    getClassSpvLearning(): TClassCalcSpvLearning; override;
end;

{paramètre de l'algo CART}
TOPrmSpvCSTreeC45 = class(TOpPrmSpvTree)
private
//taille minimale pour les feuilles
FMinSizeLeaf: integer;
//lambda pour l'estimation laplacienne des distributions
FLambda: double;
//matrice de coût
FCostMatrix: TCostMatrix;
protected
procedure   SetDefaultParameters(); override;
function    CreateDlgParameters(): TForm; override;
public
{son propriétaire est passé en paramètre : l'opérateur}
constructor Create(prmOp: TOperator); override;
destructor  Destroy(); override;
function    getHTMLParameters(): string; override;
procedure   LoadFromStream(prmStream: TStream); override;
procedure   SaveToStream(prmStream: TStream); override;
procedure   LoadFromINI(prmSection: string; prmINI: TMemIniFile); override;
procedure   SaveToINI(prmSection: string; prmINI: TMemIniFile); override;
property    CostMatrix: TCostMatrix read FCostMatrix;
property    Lambda: double read FLambda write FLambda;
property    MinSizeLeaf: integer read FMinSizeLeaf write FMinSizeLeaf;
end;
1: 1 | Insertion | \Code/
```

Classes de  
gestion des  
composants



```
D:\Temp\Exe\tanagra_components.xml

<component class_name="TMLGCompSpvTreeC45">
  <name>
    C4.5
  </name>

  <bitmap>
    MLSpvTreeC45.bmp
  </bitmap>

  <description>
    Quinlan (1993), decision tree algorithm.
  </description>

  <precondition>
    At least one discrete attribute (TARGET) and one or more discrete/continuous attributes (INPUT) must be av
  </precondition>

  <target-attributes>
    Discrete class attribute.
  </target-attributes>

  <input-attributes>
    One or more discrete/continuous attributes.
  </input-attributes>

  <postcondition>
    A new column (discrete attribute) is added, it corresponds to the predicted values of the class attribute.
  </postcondition>

</component>
```

Fichier externe de gestion des composants pour les versions spécialisées [et aussi au cas où on passait par une gestion par plug-ins]

→ l'adjonction d'un composant est très peu contraignante

# Simplicité pour les programmeurs

## Encore plus loin dans la modularité : les plugins

### La solution idéale ?

L'application mère est une matrice qui gère et transmet les données

Les techniques sont des procédures programmées sous forme de bibliothèques externes

### Mais des contraintes fortes

Organisation ultra rigoureuse des protocoles

→ Passage des informations et des données

→ Affichage des résultats

→ Documentation (fichier d'aide)

### Bref...

Souvent rédhibitoire, alors que l'objectif était d'offrir un outil modulaire

Intéressant si plugins = procédures de calculs qui renvoient des objets standardisés

Et qu'une vraie équipe organise la vie autour du logiciel

→ **Le logiciel R est le seul à avoir su le faire convenablement**

# Implémentation

Quels outils pour la programmation ?

## Spécifications

Outil libre (*ça coûte moins cher*)

Largement diffusé (*pour avoir des programmeurs*)

Avec une large bibliothèque de classes (*calculs, conteneurs, etc.*)

Qui permet de faire des interfaces agréables, simplement, rapidement

## Pourquoi DELPHI pour Tanagra ?

[DELPHI 6.0 PERSO est gratuite](#)

Cours de DELPHI en L3 et M1 dans le département « Informatique – Statistique »

Accès aux anciennes bibliothèques de calculs, validées depuis longtemps déjà

Connaissance étendue des bibliothèques libres (Turbo Power, etc.)

Permet de faire des interfaces agréables, simplement, rapidement

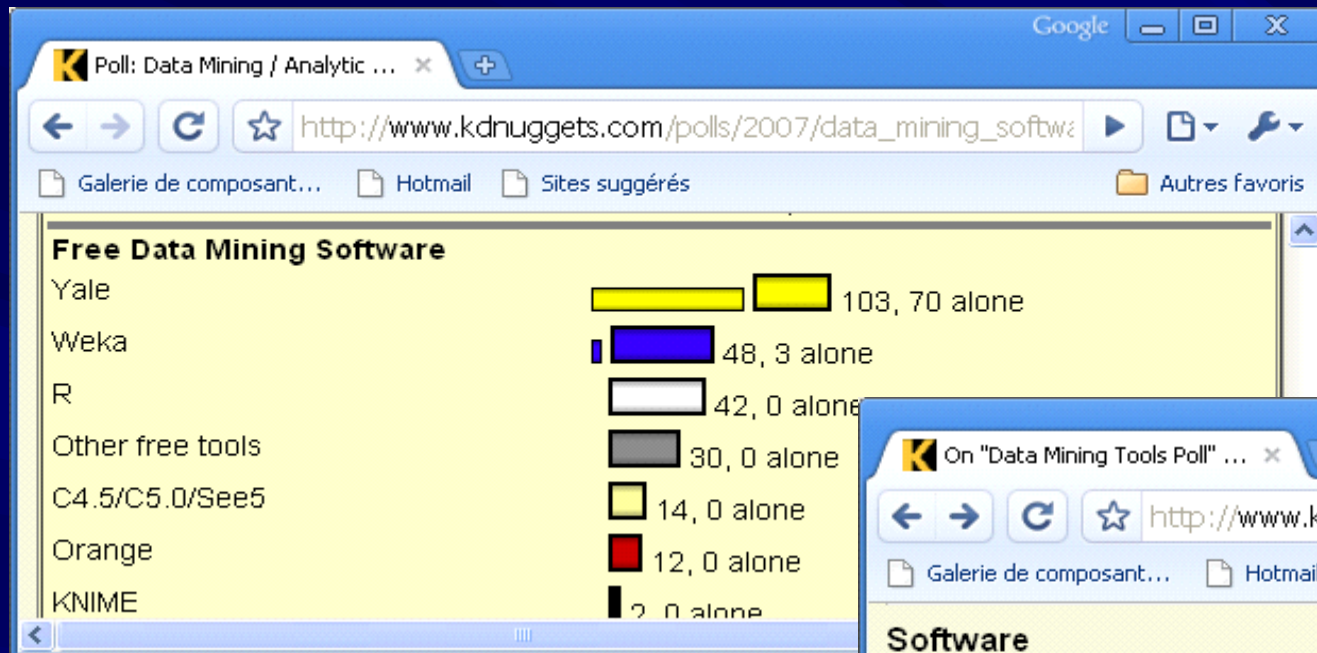
*Affinités personnelles...*

J'aurais du le faire en JAVA ? L'écueil WEKA →

# Implémentation

Pourquoi ne pas avoir intégré des bibliothèques de calcul existantes ?

Sondage : quel logiciel utilisez vous en 2007 ?



**Software**

**From:** Dr. Alexander K. Seewald  
**Date:** 27 Nov 2007  
**Subject:** On "Data Mining Tools Poll" - RapidMiner is a version of Weka

The free data mining software tool Yale (now named RapidMiner) is heavily based on Weka. I would classify it as WEKA with a nifty interface, even the code tree is quite similar at first glance. This does not seem to be widely known. On a more positive note, this means that WEKA is on first place in usage (103+48 = 151 :-) among free tools in

[KDnuggets 2007 Poll: Data Mining / Analytic Software Tools](#)

This should be accounted for in next years poll.

Dr. Alexander K. Seewald  
alexATseewaldDOTat

Bookmark

## Promotion

Comment faire connaître le logiciel... sachant qu'on n'a rien à vendre ?

### La promotion dans les conférences

*Ateliers, démonstrations, contacts chercheurs, mailing- list, etc.*



### Écrire un article de référence

Voilà toujours une publication de plus

Marquer le coup en annonçant le logiciel

C'est la référence que citeront les utilisateurs

### Monter un site web attrayant (attractif)

La visibilité internet est primordiale

Le téléchargement du logiciel n'est pas le seul enjeu

Système de « news - actualités » pour soutenir l'attention des utilisateurs

### Documenter le logiciel

Documenter les méthodes : description théorique

Documenter leur mise en oeuvre : les tutoriels

Facilitée par le découpage en « composants » du logiciel



### Mon principal cheval de bataille aujourd'hui

→ Travail de fond. A son rythme. Sans contraintes de temps. Sans pression.

→ Étendu aux autres logiciels. **Je documente les autres logiciels libres !!!**





# Promotion

## Le site web Tanagra



The screenshot shows a web browser window with the URL <http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>. The page features a navigation menu with icons and labels: Présentation, Galerie, Caractéristiques, Didacticiels, Téléchargement, and Sipina. A sidebar on the left contains a menu with items: Présentation, Projet TANAGRA, Nouveautés **NEW!**, Historique, Références, Autres logiciels, Portail Data Mining, and Contact. The main content area is titled "Le projet TANAGRA" and contains the following text:

TANAGRA est un logiciel gratuit de DATA MINING destiné à l'enseignement et à la recherche. Il implémente une série de méthodes de fouilles de données issues du domaine de la statistique exploratoire, de l'analyse de données, de l'apprentissage automatique et des bases de données.

TANAGRA est un projet ouvert au sens qu'il est possible à tout chercheur d'accéder au code et d'ajouter ses propres algorithmes pour peu qu'il respecte la licence de distribution du logiciel.

L'objectif principal du projet TANAGRA est d'offrir aux chercheurs et aux étudiants une **plate-forme de Data Mining** facile d'accès, respectant les standards des logiciels du domaine, notamment en matière d'interface et de mode de fonctionnement, et permettant de **mener des études** sur des données réelles et/ou synthétiques.

Le second objectif de TANAGRA est de proposer aux chercheurs une architecture leur permettant d'implémenter aisément les techniques qu'ils veulent étudier, de comparer les performances des algorithmes. TANAGRA se comporte plus comme une **plate-forme d'expérimentation** qui leur permettrait d'aller à l'essentiel en leur épargnant toute la partie ingrate de la programmation de ce type d'outil : la gestion des données.

Le troisième et dernier objectif, en destination des apprentis programmeurs, vise à **diffuser une méthodologie possible d'élaboration de ce type de logiciel**. L'accès au code leur permettra de voir comment se construit ce type de logiciel, quels sont les écueils à éviter, quelles sont les principales étapes d'un tel projet, et quels sont les outils et les bibliothèques qu'il faut préparer pour le mener à bien. En ce sens, TANAGRA est plus un outil d'apprentissage des techniques de programmation.

TANAGRA n'intègre pas en revanche, à l'heure actuelle, tout ce qui fait la puissance des outils commerciaux du marché : multiplicité des sources de données, accès direct aux entrepôts de données et autres datamarts, appréhension des données à problèmes (valeurs manquantes...), interactivité des traitements.




TANAGRA - A free DATA ... x


← → ↻ 🏠 ☆ http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html ▶ 📄 🔧


📁 e-mail 📧 Prévisions météo de... 📺 YouTube - Page Yo... 📁 Bmw Story 📁 Autres favoris




# TANAGRA


  
Presentation

  
Screenshots

  
Functionalities

  
Tutorials

  
Download

  
Sipina

Presentation	TANAGRA project
TANAGRA project	TANAGRA is a free DATA MINING software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.
Releases <b>NEW!</b>	
History	
References	This project is the successor of <a href="#">SIPINA</a> which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms...
Related softwares	TANAGRA is an "open source project" as every researcher can access to the source code, and add his own algorithms, as far as he agrees and conforms to the software distribution license.
Contact	The main purpose of Tanagra project is to give researchers and students an easy-to-use <b>data mining software</b> , conforming to the present norms of the software development in this domain (especially in the design of its GUI and the way to use it), and allowing to analyse either real or synthetic data.
	The second purpose of TANAGRA is to propose to researchers an architecture allowing them to easily

# Promotion

## Documentation des méthodes – Pointeurs vers les ressources

The screenshot shows a Windows Internet Explorer browser window titled "DATA MINING - Windows Internet Explorer". The address bar shows the URL "http://eric.univ-lyon2.fr/~ricco/data-mining/". The browser interface includes a menu bar with "Fichier", "Edition", "Affichage", "Favoris", and "Outils". Below the menu bar are search engines like "Yahoo! Search" and a "Favoris" section. The website content is displayed on a blue background. At the top, there is a navigation bar with "RESSOURCES DATA MINING" and several menu items: "Data Mining", "Documentation en ligne", "Données et logiciels", and "Logiciel TANAGRA". Below these are sub-items: "Préparation des données", "Apprentissage supervisé", "App. supervisé (suite)", and "Logiciel SIPINA". The main content area is divided into two columns. The left column is a sidebar with a "Data Mining" header and links to "Cours + TD + Données", "Vidéos", "Machine Learning (Andrew Ng - Stanford)", "Statistical Aspects of Data Mining (D. Mease)", "Conférence EGC-2009", "Glossaires", "Portails Data Mining", "Portails Machine Learning", and "Portails Statistiques". The right column features a section titled "Extraction de connaissances à partir de données (ECD)" with a detailed paragraph explaining the concept and its history, followed by a list of two specific characteristics of ECD.

**RESSOURCES DATA MINING**

- Data Mining
- Documentation en ligne
- Données et logiciels
- Logiciel TANAGRA

Préparation des données    Apprentissage supervisé    App. supervisé (suite)    Logiciel SIPINA

**Data Mining**

- Cours + TD + Données
- Vidéos
- Machine Learning (Andrew Ng - Stanford)
- Statistical Aspects of Data Mining (D. Mease)
- Conférence EGC-2009
- Glossaires
- Portails Data Mining
- Portails Machine Learning
- Portails Statistiques

### Extraction de connaissances à partir de données (ECD)

L'Extraction de Connaissances à partir de Données (ECD), communément appelée DATA MINING, est un domaine aujourd'hui très en vogue, pour ne pas dire à la mode. On la définit comme **"un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données (Fayyad, 1996)"**. Cette définition est une des premières qui traite explicitement de l'ECD (Knowledge Discovery in Databases en anglais), par la suite plusieurs tentatives de re-définition sont apparues pour mieux préciser le domaine mais aucune ne s'est réellement imposée. En tous les cas, à la lecture des différents documents qui traitent de l'ECD, on peut se dire que, finalement, cela fait plus de 30 ans qu'on le pratique avec ce qu'on appelle l'analyse de données et les statistiques exploratoires. Et on n'aurait pas complètement tort.

En réalité, ce n'est pas aussi simple, l'ECD possède des particularités qui sont loin d'être négligeables :

- (1) des techniques d'analyse qui ne sont pas dans la culture des statisticiens, en provenance de l'apprentissage automatique (Intelligence artificielle) et des bases de données ;
- (2) l'extraction de connaissances est intégrée dans le schéma organisationnel de l'entreprise. Ainsi, les données ne sont plus issues d'enquêtes ou de sondages mais proviennent d'entrepôts construits sciemment pour une exploitation aux fins d'analyse. Le DATAWAREHOUSE. D'une part, une

http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html



# Promotion

## Documentation des méthodes – Écrire et diffuser des supports libres

**NOS FORMATIONS**

[Département Info-Stat](#)  
[Diplômes Licence - Master](#)  
[Formation continue](#)

**TUTORIELS**

[Portail Data Mining](#)  
[Tutoriels pour le Data Mining](#)

**LOGICIELS**

[Tanagra \(Open Source\)](#)  
[Sipina - Arbres de décision](#)

**VIDÉOS**

[Machine Learning \(A. Ng\)](#)  
[Statistical Aspects \(D. Mease\)](#)

**REF. EXTERNES**

[Applied Statistics](#)  
[Data Mining tutorials](#)  
[DM and Analytic Technologies](#)  
[Statnotes : Topics in MVA](#)

**CHAPITRES DE COURS**

[Généralités - Data Mining](#)  
[App. supervisé - Scoring](#)  
[Règles d'association](#)  
[Analyse factorielle](#)  
[Classification - Clustering](#)  
[Régression - Économétrie](#)  
[Probabilités et Statistique](#)

**RECHERCHE**

**Test de normalité**  
Test statistique d'adéquation à la loi normale (normality test) : test de Shapiro Wilk, test de Lilliefors, test d'Anderson-Darling, test de D'Agostino, test de Jarque-Bera.  
Test de symétrie des distributions : test basé sur le coefficient d'asymétrie, test de Wilcoxon, test de Van der Waerden.

**Corrélation et corrélation partielle**  
Covariance, corrélation linéaire, corrélations croisées, tests de significativité. Corrélation bisériale ponctuelle, corrélation mutuelle, le coefficient phi, rho de Spearman, tau de Kendall, rapport de corrélation. Corrélations partielles et semi-partielles d'ordre p. Corrélation partielle de rangs.

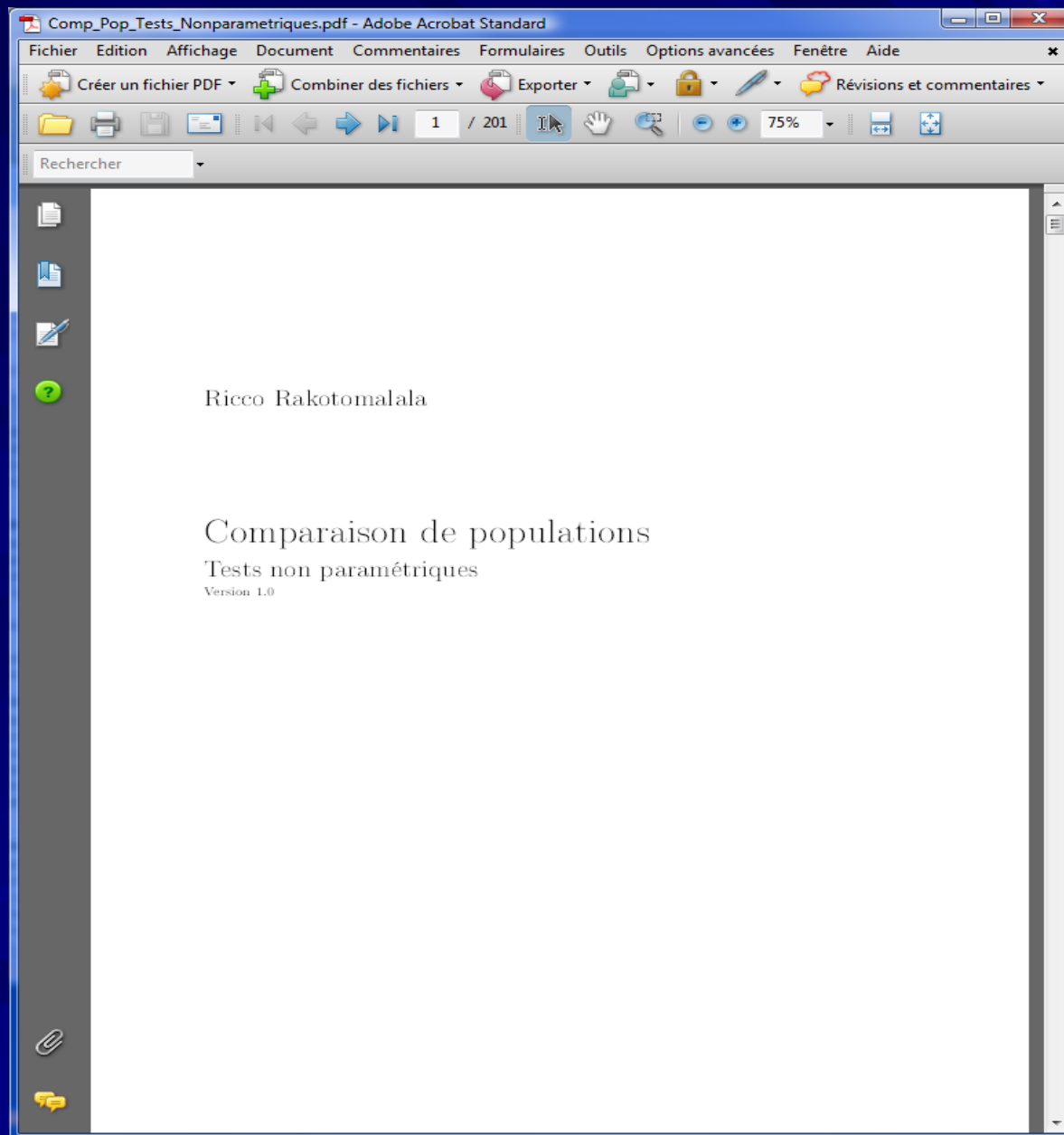
**Mesures d'association pour variables nominales**  
Test d'indépendance du KHI-2. Mesures dérivées du KHI-2 (T de Tschuprow, c de Cramer...). Mesures asymétriques d'association (PRE measures) : Lambda et Tau de Goodman & Kruskal, U de Theil. Éléments spécifiques aux tableaux 2 x 2 : Q de Yule, Odds-ratio, Risque relatif, correction de Yates.  
Coefficient de concordance pour variables nominales : Kappa de Cohen, Kappa de Fleiss, Kappa généralisé.  
Mesures d'association pour les variables ordinales (Gamma de Goodman et Kruskal, Tau-b et Tau-c de Kendall, d de Sommers).

**Comparaison de populations - Tests paramétriques**  
Comparaison de 2 moyennes, échantillons indépendants, variances égales et inégales. Comparaison de 2 moyennes, échantillons appariés. Comparaison de variances, échantillons indépendants et appariés. Comparaison de K moyennes, échantillons indépendants (ANOVA) et appariés (blocs aléatoires complets). Test multivariés : T2 de Hotelling, Lambda de Wilks, Trace de Pillai. Test de Bartlett pour comparaison des matrices de variance covariance.

**Comparaison de populations - Tests non paramétriques**  
Test de Kolmogorov-Smirnov, test de Kuiper, test de Cramer - von Mises, test de Wilcoxon-Mann-Whitney, test de Kruskal-Wallis, test de Mood, test de Klotz, test des signes, test des rangs signés de Wilcoxon pour échantillons appariés, anova de Friedman, test de Mc Nemar, test Q de Cochran, test de Jonckheere-Terpstra, test de Page

Ricco Rakotomalala – Université Lyon 2

Présenter à la fois les aspects pratiques et théoriques



http://eric.univ-lyon2.fr/... x

http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique\_regression\_logistique.pdf

e-mail Prévisions météo de... YouTube - Page Yo... Bmw Story Autres favoris

Rechercher 1 / 272

Présenter à la fois les aspects pratiques et théoriques

Ricco Rakotomalala

Pratique de la Régression Logistique  
Régression Logistique Binaire et Polytomique  
Version 2.0

# Promotion

## Documenter la mise en œuvre des méthodes – Les tutoriels

### Tutoriels Tanagra pour le Data Mining

Ce blog recense les didacticiels pour Tanagra. Ils sont organisés en catégories. On dispose des fonctionnalités mots-clés. Chaque article est accompagné d'un texte de présentation, d'une liste de mots-clés, du lien vers le didacticiel (pdf) et de la bibliographie. Dans certains cas (catégorie « Tanagra et les autres »), nous mentionnons d'autres logiciels libres (Krnime, Orange, R, RapidMiner, Sipina, Weka) ou commerciaux (Spad).

MERCREDI 3 FÉVRIER 2010

#### ↳ Discrétisation - Comparaison de logiciels

La discrétisation consiste à découper une variable quantitative en intervalles. Il s'agit d'une opération de recodage. De quantitative, la variable est transformée en qualitative ordinaire. Nous devons répondre à deux questions pour mener à bien l'opération : (1) comment déterminer le nombre d'intervalles à produire ; (2) comment calculer les bornes de discrétisation à partir des données. La résolution ne se fait pas forcément dans cet ordre.

J'ai l'habitude de dire que le découpage d'expert est le meilleur possible. En effet, lui seul peut fournir une discrétisation raisonnée tenant compte des connaissances du domaine, tenant compte de tout un tas de contraintes dont on n'a pas idée si on se base uniquement sur les données, et en adéquation avec les objectifs de l'étude. Malheureusement, la démarche s'avère délicate parce que : d'une part, les connaissances ne sont pas toujours au rendez-vous ou sont difficilement quantifiables ; d'autre part, elle n'est pas automatisable, le traitement d'une base comportant des centaines de variables se révèle rapidement ingérable. Souvent donc, nous sommes obligés de nous baser uniquement sur les données pour produire un découpage qui soit un tant soit peu pertinent.

Discrétisation comme prétraitement des variables en apprentissage supervisé. Tout d'abord, il faut situer le canevas dans lequel nous réalisons l'opération. Selon le cas, il est évident que la démarche et les critères utilisés ne seront pas les mêmes. Dans ce didacticiel, nous nous plaçons dans le cadre de l'apprentissage supervisé. Les variables quantitatives sont préalablement recodées avant d'être présentées à un algorithme d'apprentissage supervisé. La variable à prédire, elle, est naturellement qualitative. Lors de la discrétisation, il est par conséquent souhaitable que les groupes soient le plus purs possibles c.-à-d. les individus situés dans le même intervalle doivent appartenir majoritairement à l'une des modalités de la variable à prédire.

Dans ce didacticiel, nous comparerons le comportement des techniques supervisées et non supervisées implémentées dans les logiciels [Tanagra 1.4.35](#), [Sipina 3.3](#), [R 2.9.2](#) (package dprep), [Weka 3.6.0](#), [Krnime 2.1.1](#), [Orange 2.0b](#) et [RapidMiner 4.6.0](#). Comme nous pouvons le constater, tout logiciel de Data Mining se doit de proposer ce type d'outils. Nous mettrons en avant le paramétrage et la lecture des résultats.

**Mots clés** : mdlpc, discrétisation supervisée, discrétisation non supervisée, intervalles de largeurs égales, intervalles de fréquences égales

**Composants** : MDLPC, Supervised Learning, Decision List

**Lien** : [fr\\_Tanagra\\_Discretization\\_for\\_Supervised\\_Learning.pdf](#)

**Données** : [data-discretization.aff](#)

**Références** :

F. Mühlenbach, R. Rakotomalala, « Discretization of Continuous Attributes », in Encyclopedia of Data Warehousing and Mining, John Wang (Ed.), pp. 397-402, 2005

Support

Page

Porta

Cour

Ouvr

Logicie

Site c

Télé

Site c

Favoris

Tana

Tana

Comp

Tuto

Recher

Catégo

Anal

Anal

App.

Arbre

Class

Cons

Impo

Ouvr

Règl

Régn

Régn

Régn

Sipin

Stat

Séle

## Tanagra - Data Mining Tutoriels

This Web log maintains an alternative layout of the tutorials about Tanagra. Each entry describes shortly the subject, it is followed by the link to the tutorial (pdf) and the dataset. The technical references (book, papers, website,...) are also provided. In some tutorials, we compare the results of Tanagra with other free software such as Knime, Orange, R software, RapidMiner (Yale), Sipina or Weka.

THURSDAY, FEBRUARY 11, 2010

### ↳ Supervised rule induction - Software comparison

Supervised rule induction methods play an important role in the Data Mining framework. Indeed, it provides an easy to understand classifier. A rule uses the following representation: "IF premise THEN conclusion" (e.g. IF an account problem is reported on a client THEN the credit is not accepted).

Among the rule induction methods, the "separate and conquer" approaches are very popular during the 90's. Curiously, they are less present today into proceedings or journals. More troublesome still, they are not implemented in commercial software. They are only available in free tools from the Machine Learning community. However, they have several advantages compared to other techniques.

In this tutorial, we describe first two separate and conquer algorithms for the rule induction process. Then, we show the behavior of the classification rules algorithms implemented in various tools such as [Tanagra 1.4.34](#), [Sipina Research 3.3](#), [Weka 3.6.0](#), [R 2.9.2](#) with the RWeka package, [RapidMiner 4.6](#), or [Orange 2.0b](#).

**Keywords**: rule induction, separate and conquer, top-down, CN2, decision tree

**Composants** : SAMPLING, DECISION LIST, RULE INDUCTION, TEST

**Tutorial**: [en\\_Tanagra\\_Rule\\_Induction.pdf](#)

**Dataset**: [life\\_insurance.zip](#)

**References**:

J. Furnkranz, "Separate-and-conquer Rule Learning", Artificial Intelligence Review, Volume 13, Issue 1, pages 3-54, 1999.

P. Clark, T. Niblett, "The CN2 Rule Induction Algorithm", Machine Learning, 3(4):261-283, 1989.

P. Clark, R. Boswell, "Rule Induction with CN2: Some recent improvements", Machine Learning - EWSL-91, pages 151-163, Springer Verlag, 1991.

Posted by Tanagra at 2:04 AM

Labels: [Software Comparison](#), [Supervised Learning](#)

Data mining tutorials

↳ [Home page](#)

Downloads

↳ [Tanagra home page](#)

↳ [Setup download](#)

↳ [Sipina home page](#)

Favorites

↳ [Data file handling](#)

↳ [Software comparison](#)

↳ [Tanagra under Linux](#)

Subscribe To

Posts

All Comments

Search This Blog

powered by [Google™](#)

Categories

↳ [Association rules \(9\)](#)

↳ [Clustering \(12\)](#)

↳ [Data file handling \(8\)](#)

↳ [Decision tree \(18\)](#)

↳ [Diagram management \(2\)](#)

↳ [Exploratory Data Analysis \(9\)](#)

↳ [Feature Construction \(2\)](#)

↳ [...](#)

pratique\_regression\_....pdf

[Afficher tous les téléchargements...](#)

Avec un contenu rodé...





### Écriture du cahier des charges

Janvier 2003, plusieurs prototypes de janvier à juin 2003

### Début du développement

Juillet 2003

### Création du site web et mise en ligne

Janvier 2004 (~25 visiteurs par jour sur 2004)

### Techniques implémentées (version 1.4.36 – Mars 2010)

170 méthodes stat., analyse de données, data mining

### Documentation libre en ligne (Mars 2010)

8 ouvrages libres en PDF

30 « slides » en PDF

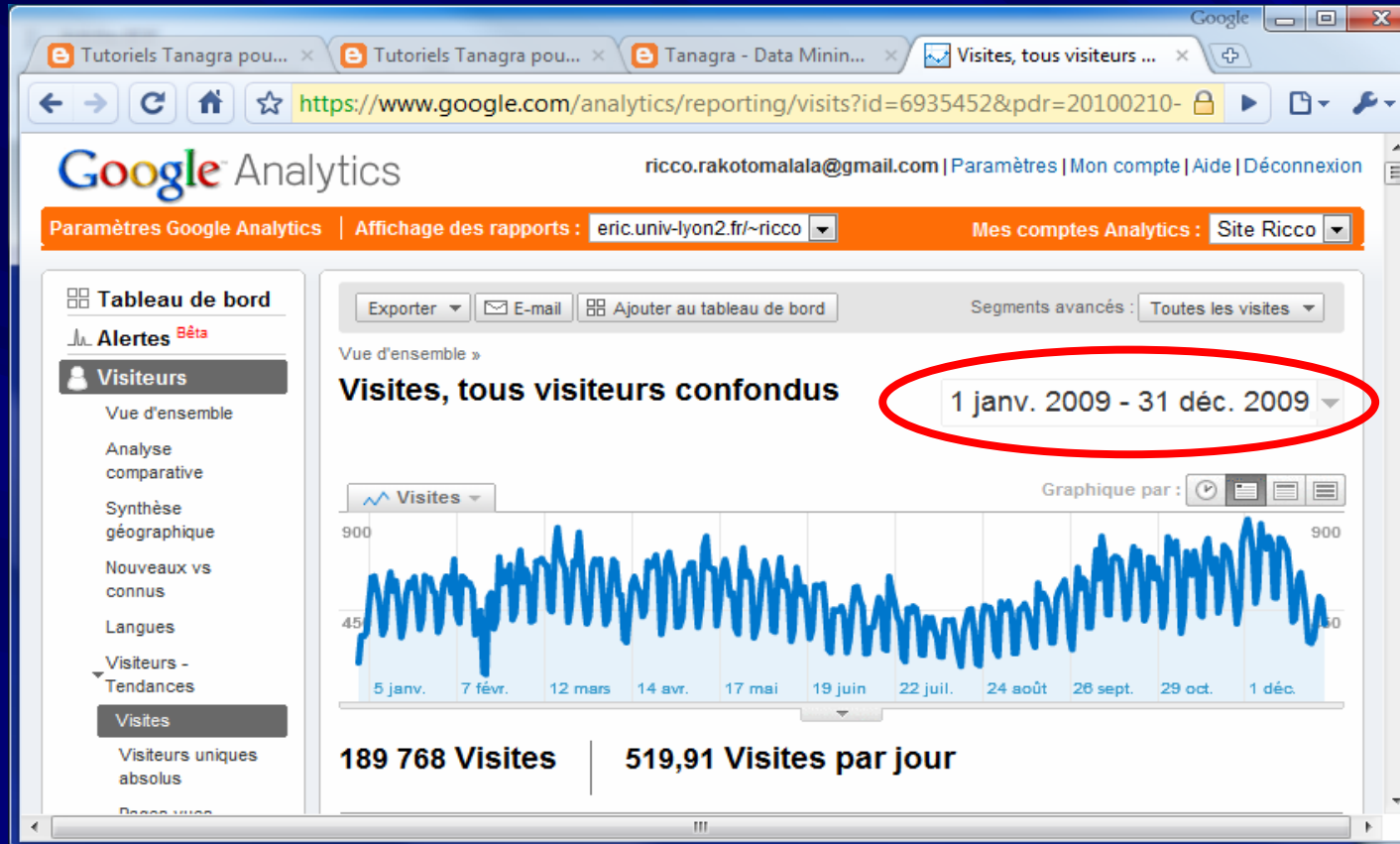
140 didacticiels en français

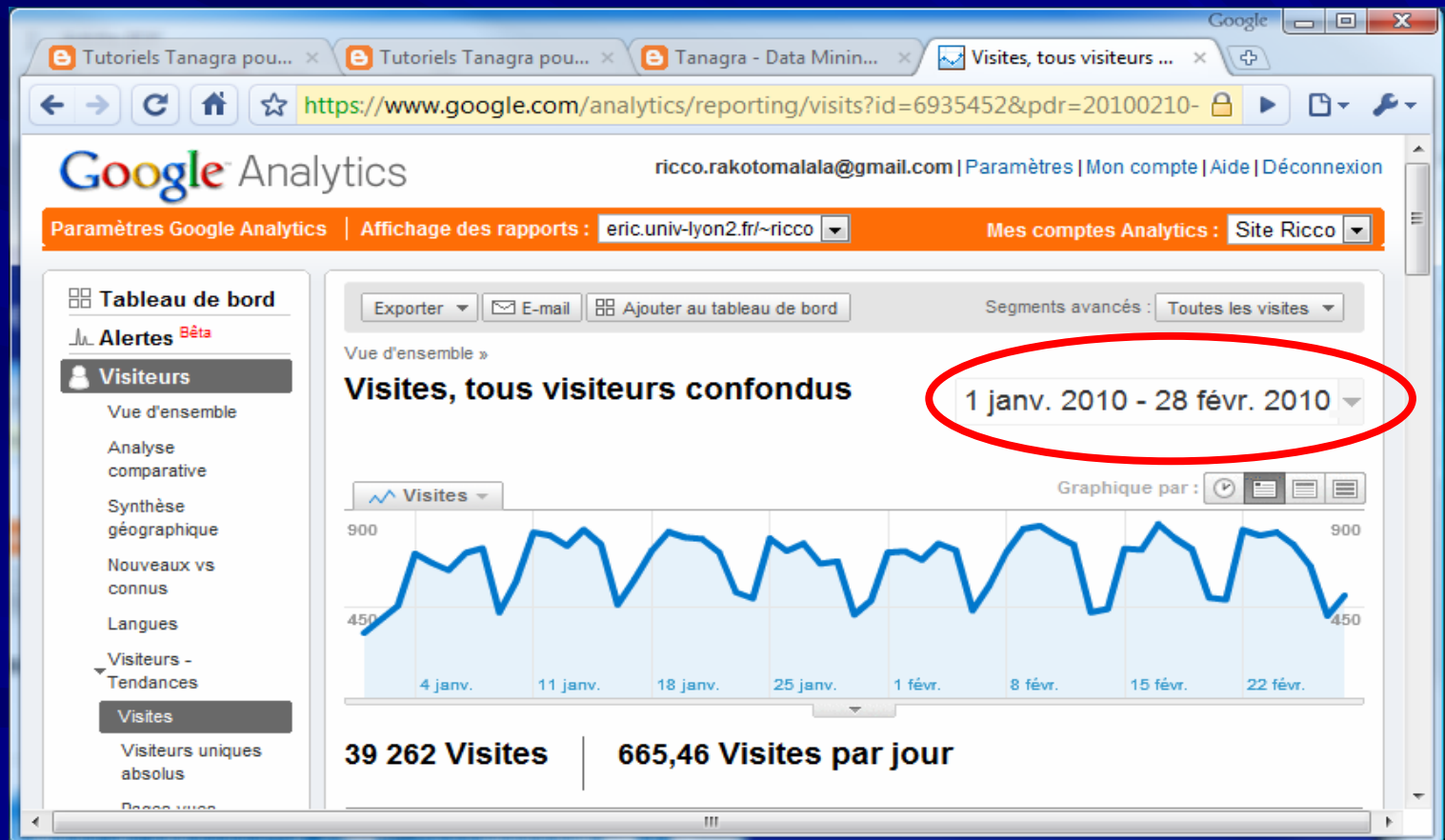
100 didacticiels en anglais

# Tanagra

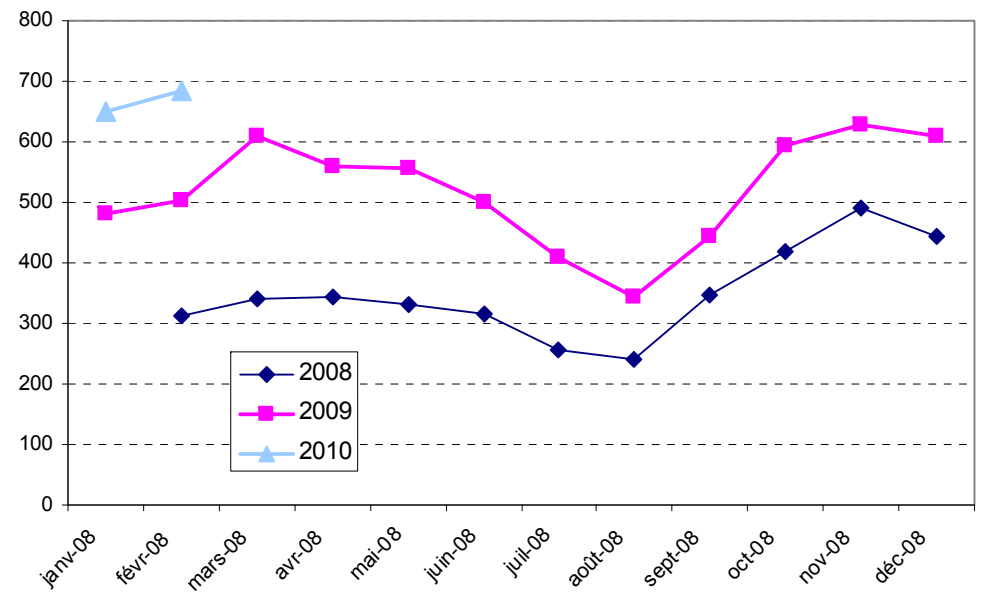
## Bilan (2) – Quelle fréquentation du site ?





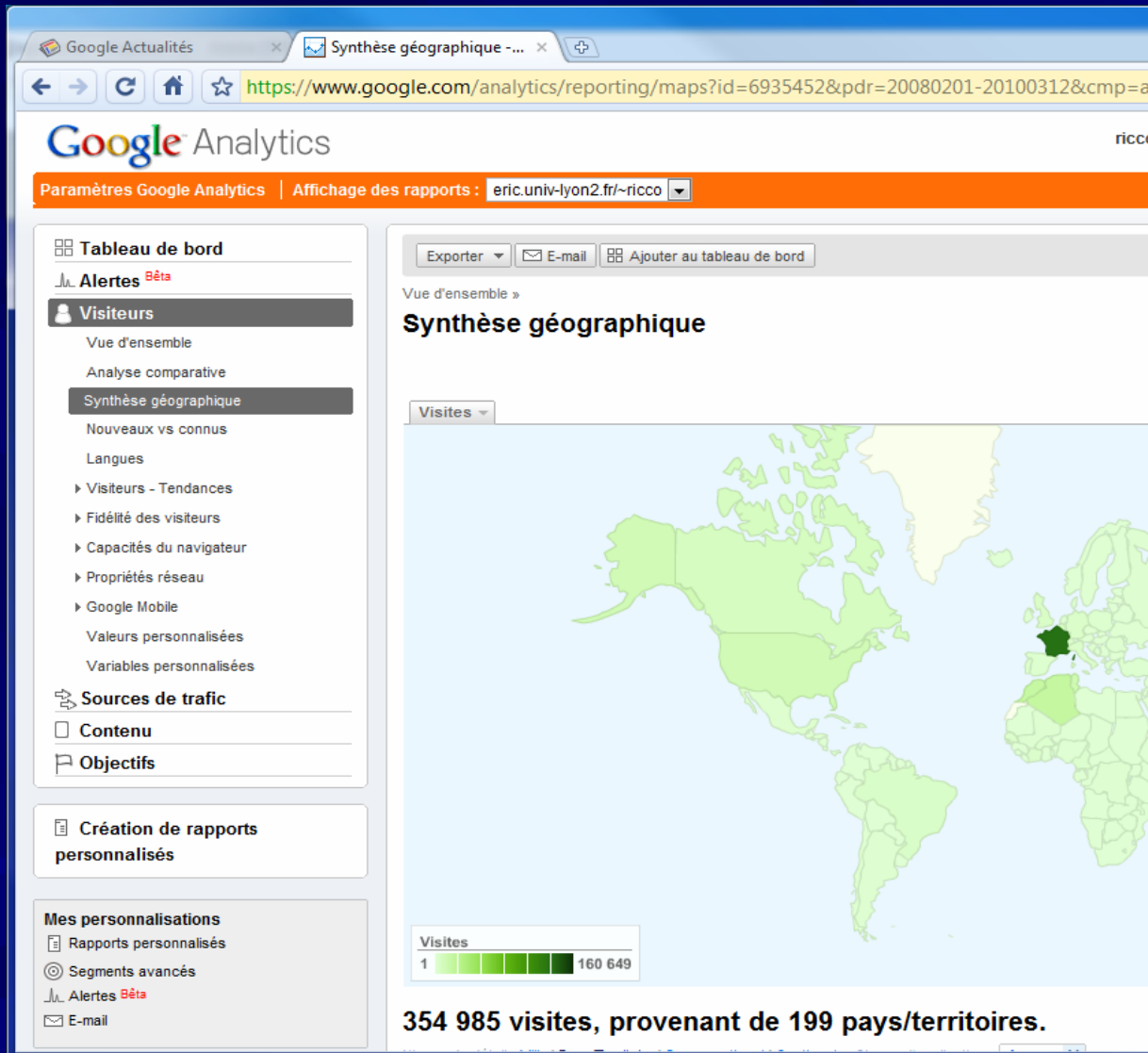


Nombre de visites par jour



# Tanagra

## Bilan (3) – Qui fréquente le site ?



	Niveau de détail : Pays/Territoire	Visites ↓
1.	France	160 649
2.	Algeria	23 853
3.	Morocco	22 943
4.	Tunisia	15 935
5.	United States	14 148
6.	Canada	8 585
7.	India	7 960
8.	Belgium	7 640
9.	Brazil	5 513
10.	United Kingdom	5 351

Quels pays ?

	Titre de la page	Pages vues ↓
1.	Portail DATA MINING	74 984
2.	Supports de cours -- Data Mining	62 583
3.	Tanagra EN	56 990
4.	Tanagra FR	56 129
5.	Cours econometrie	55 190
6.	Download Tanagra EN	37 456
7.	Download Tanagra FR	34 337
8.	Cours Excel	34 312
9.	Tutoriels Tanagra pour le Data Mining	31 985
10.	Sipina Main Menu	25 531

Quelles pages ?

Les sites des tutoriels sont récents (FR et EN)



4. Et les autres outils libres ?

# Knime

Estampillé « Intelligent Data Analysis »



KNIME | Konstanz Informat...

http://www.knime.org/

Galerie de composant... Hotmail Sites suggérés

Autres favoris

Interested in professional support for KNIME?

Introduction Download Documentation Developer About Supporters

**KNIME**  
Konstanz Information Miner  
a modular, extendable data exploration platform to visually create data pipelines.

Learn Get Use

http://www.knime.org/images/front\_page/views\_800.jpg

Université de Konstanz - Allemagne

Culture I.D.A

Code source libre C++ (méthodes)

Doc sous forme de fichier d'aide intégré

Mode diagramme

Avec des fonctionnalités avancées (boucles,...)

Les méthodes sont des plugins

Possibilité d'importer des classes Weka

Possibilité d'intégrer des packages R

Multi-thread et possibilité de swap pour certaines méthodes, le mieux armé pour les gros volumes

# Knime Interface

The screenshot displays the KNIME software interface with the following components:

- Workflow Projects:** Shows a project named "Zooplankton analysis".
- Node Repository:** Lists various nodes categorized by function, including IO (Read, Write), Database, Data Manipulation, and Data Views.
- Workflow Diagram:** A central workspace showing a workflow with seven nodes: File Reader (Node 1), Column Filter (Node 2), Partitioning (Node 3), Decision Tree Learner (Node 4), Decision Tree Predictor (Node 5), Scorer (Node 6), and Interactive Table (Node 7). The nodes are connected in a sequence from left to right.
- Node Description Panel:** A panel on the right titled "Scorer" that provides a detailed description of the Scorer node's functionality and its ports.
- Console Window:** A window at the bottom showing the KNIME Console output, including a welcome message and several warning messages.

```
graph LR; N1[File Reader Node 1] --> N2[Column Filter Node 2]; N2 --> N3[Partitioning Node 3]; N3 --> N4[Decision Tree Learner Node 4]; N3 --> N5[Decision Tree Predictor Node 5]; N4 --> N5; N5 --> N6[Scorer Node 6]; N6 --> N7[Interactive Table Node 7];
```

**Node Description: Scorer**

Compares two columns by their attribute value pairs and shows the confusion matrix, i.e. how many rows of which attribute and their classification match. Additionally, it is possible to highlight cells of this matrix to determine the underlying rows. The dialog allows you to select two columns for comparison; the values from the first selected column are represented in the confusion matrix's rows and the values from the second column by the confusion matrix's columns. The output of the node is the confusion matrix with the number of matches in each cell.

**Ports**

**Input Ports**

**KNIME Console**

```
*****  
*** Welcome to KNIME v2.0.0.0019511 - the Konstanz Information Miner ***  
*** Copyright, 2003 - 2008, Uni Konstanz and KNIME GmbH, Germany ***  
*****  
Log file is located at: C:\Program Files\Knime\knime_2.0.0\workspace\metadata\knime\knime.log  
WARN File Reader No Settings available.  
WARN Column Filter All columns retained.  
WARN Partitioning No sampling method selected  
WARN Decision Tree Learner Guessing target column: "Ident_2".  
WARN Scorer No columns selected yet.
```

Orange - Data Mining Fruitf... x

http://www.aillab.si/orange/

Galerie de composant... Hotmail Sites suggérés

Autres favoris

orange

Google™ Custom Search

Features Download Documentation Widgets Scripting

Open source data visualization and analysis for novice and experts. Data mining through visual programming or Python scripting. Extensions for bioinformatics and text mining. Comprehensive, flexible and fast.

Orange 2.0b for Windows

(Downloads for other systems and versions)

Latest News & Blog Entries

- 22 Jul [Orange's new web-site](#)
- 02 Mar [Clustering module](#)
- 07 Nov [Our Fink repository](#)
- 17 Jul [Facelift](#)
- 13 Jun [New documentation being written](#)

A.I. Lab – Université de Lubiana – Slovénie

Culture I.A. - Machine Learning (ICML, ...)  
Code source libre C++

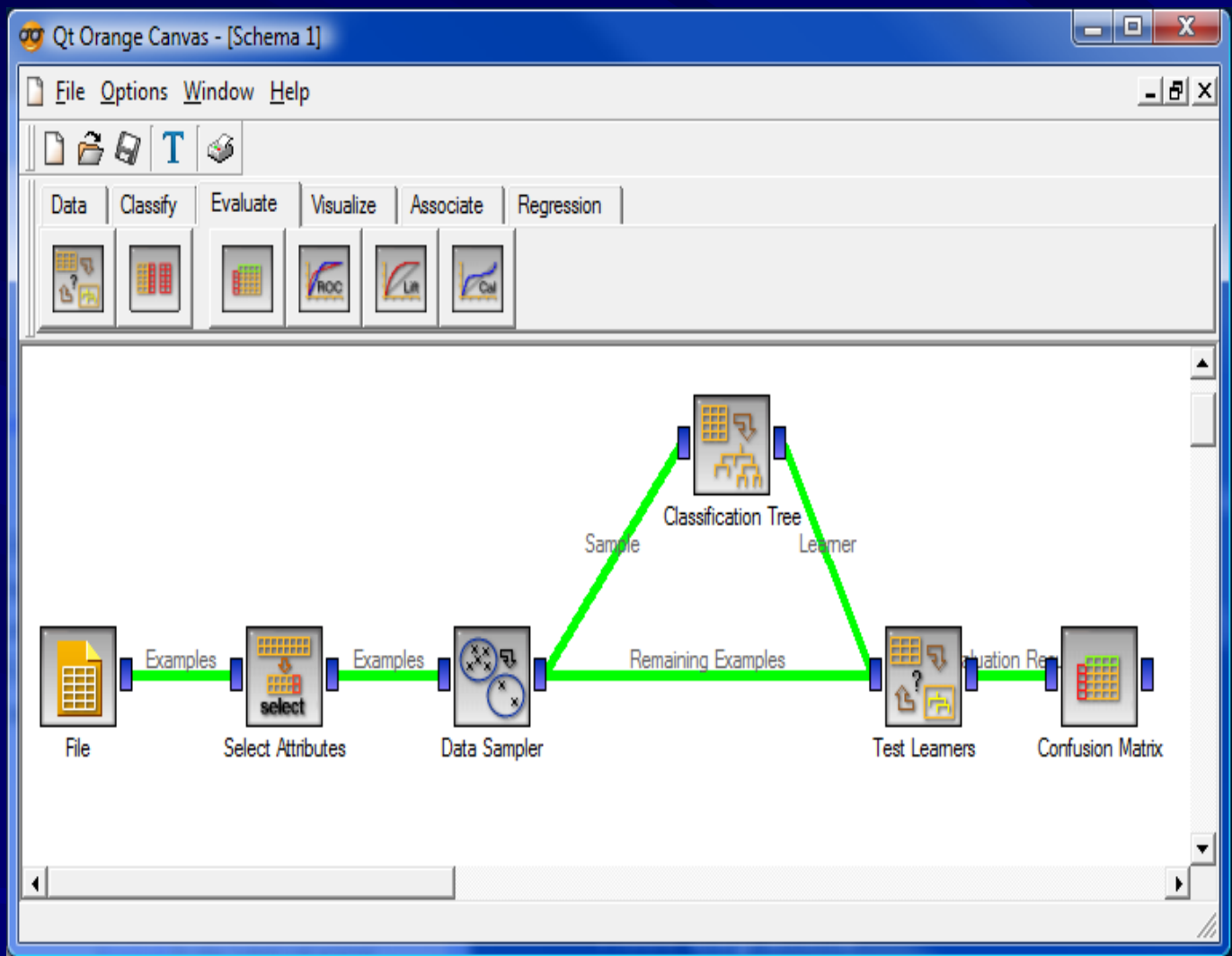
Site Web avec doc en ligne et guide

Mode diagramme  
Programmation en Python

Les méthodes sont des plugins (DLL)

Très user-friendly

# Orange Utilisation



```
# Description: Shows how to construct an orange.ClassifierFromExampleTable
# Category:   classification, lookup classifiers, constructive induction, feature construction
# Classes:   ClassifierByExampleTable, LookupLearner
# Uses:     monk1
# Referenced: lookup.htm

import orange

data = orange.ExampleTable("monk1")
a, b, e = data.domain["a"], data.domain["b"], data.domain["e"]

data_s = data.select([a, b, e, data.domain.classVar])
abe = orange.LookupLearner(data_s)

print len(data_s)
print len(abe.sortedExamples)

for i in abe.sortedExamples[:10]:
    print i
print

for i in abe.sortedExamples[:10]:
    print i, i.getclass().svalue
print

y2 = orange.EnumVariable("y2", values = ["0", "1"])
abe2 = orange.LookupLearner(y2, [a, b, e], data)
for i in abe2.sortedExamples[:10]:
    print i, i.getclass().svalue
print

y2 = orange.EnumVariable("y2", values = ["0", "1"])
abe2 = orange.LookupLearner(y2, [a, b], data)
for i in abe2.sortedExamples:
    print i, i.getclass().svalue
```



The screenshot shows the R Project website with the following content:

- PCA 5 vars**: `princomp(x = data, cor = cor)`. A biplot shows variables: Fertility, Catholic, Agriculture, Examination, Education. A bar chart below shows the variance explained by the first three principal components, with the first three accounting for 60% of the variance.
- Clustering 4 groups**: A dendrogram shows hierarchical clustering of data points into four groups. A bar chart below shows the size of each group: 1 (1), 2 (2), 16 (orange), and 28 (green).
- Factor 1 [41%]** and **Factor 3 [19%]**: A normal distribution curve is shown with a vertical line indicating the position of a data point.
- Getting Started:**
  - R is a free software environment for statistical computing and graphics. It runs on a variety of UNIX platforms, Windows and MacOS. To [download R](#), please visit our [mirror](#).
  - If you have questions about R like how to download and install the software, please read our [answers to frequently asked questions](#) before you send us an email.
- News:**

Navigation links on the left include: About R, What is R?, Contributors, Screenshots, What's new?, Download, Packages, CRAN, R Project Foundation, Members & Donors, Mailing Lists, Bug Tracking, Developer Page, Conferences, Search, Documentation, Manuals, FAQs, The R Journal, Wiki, Books, Certification, Other.

Fondation à but non lucratif

Culture Stat.

CORE R + Packages (plugins)

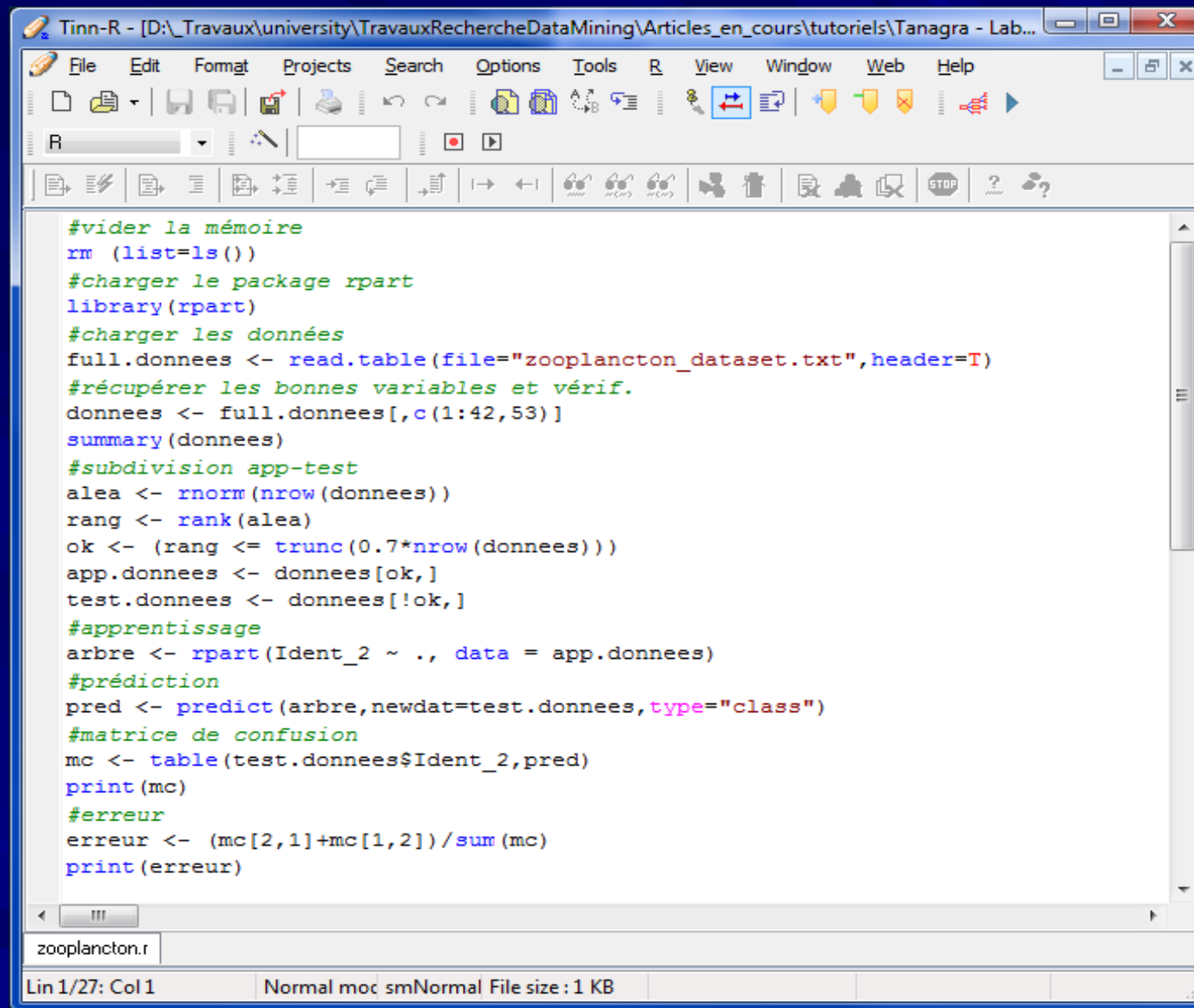
Ex. Package Weka

Doc. des méthodes très organisée

Des tutoriels partout

Mode programmation (langage S)

Quelques tentatives de création d'interfaces plus conviviales



```
#vider la mémoire
rm (list=ls())
#charger le package rpart
library(rpart)
#charger les données
full.donnees <- read.table(file="zooplancton_dataset.txt",header=T)
#récupérer les bonnes variables et vérif.
donnees <- full.donnees[,c(1:42,53)]
summary(donnees)
#subdivision app-test
alea <- rnorm(nrow(donnees))
rang <- rank(alea)
ok <- (rang <= trunc(0.7*nrow(donnees)))
app.donnees <- donnees[ok,]
test.donnees <- donnees[!ok,]
#apprentissage
arbre <- rpart(Ident_2 ~ ., data = app.donnees)
#prédiction
pred <- predict(arbre,newdat=test.donnees,type="class")
#matrice de confusion
mc <- table(test.donnees$Ident_2,pred)
print(mc)
#erreur
erreur <- (mc[2,1]+mc[1,2])/sum(mc)
print(erreur)
```

zooplancton.r

Lin 1/27: Col 1    Normal moc smNormal File size : 1 KB

The screenshot shows a web browser window displaying the Rapid-I website. The browser's address bar shows the URL <http://rapid-i.com/content/blogcategory/38/69/>. The website has a navigation menu with links for HOME, SEARCH, SITEMAP, LEGAL, CONTACT US, and DEUTSCH. The main banner features the text "Rapid-I Report the Future" and "Learn more about the Predictive Analysis and Business Intelligence solutions of Rapid-I". Below the banner is a navigation menu with links for PRODUCTS, DOWNLOADS, SERVICES, COMMUNITY, and ABOUT US. The main content area is divided into several sections: "QUICK LINKS" with links to download RapidMiner Community, order RapidMiner Enterprise, download RapidMiner Plugins, RapidMiner Documentation, All Training Courses, and RapidMiner Interactive Tour; "TESTIMONIALS" with a quote from Michael Van Kleeck, USA; "RANDOM IMAGE" showing a screenshot of the RapidMiner software interface; "RAPIDMINER COMMUNITY EDITION" with a sub-section "OPEN-SOURCE DATA MINING WITH THE JAVA SOFTWARE RAPIDMINER" and a description of RapidMiner as a world-wide leading open-source data mining solution; "RAPIDMINER: ENTERPRISE OPEN" with a description of the modular operator concept; and "OPERATOR OVERVIEW" with a description of RapidMiner (formerly YALE) and its plugins.

Entreprise commerciale  
Community Edition – Gratuite

Dérivée de Yale (Licence GNU)  
Il existe une version commerciale, sans code source  
Code de calcul Weka, mais s'en démarque de plus en plus

Pas de documentation  
Mais une multitude d'exemples « pré-programmées »

Mode diagramme arborescent

Une « profusion » de techniques de data mining

# RapidMiner Utilisation

The screenshot displays the RapidMiner interface for a project named 'zooplancton.xml'. The 'Operator Tree' on the left shows a workflow: Root (Process) -> ArffExampleSource -> SimpleValidation -> CHAID -> ApplierChain -> Test -> Performance. The 'Parameters' tab is active, showing 'keep\_example\_set' (unchecked) and 'use\_example\_weights' (checked). The bottom section shows performance metrics for the 'Test' operator, including precision (83.49%), recall (95.54%), and AUC (0.637) for the 'autre' class. A confusion matrix is also displayed, and a small chart shows a total value of 1.1 CE.

RapidMiner@VGC (zooplancton.xml)

File Edit View Process Tools Help

Operator Tree

- Root  
Process
  - ArffExampleSource  
ArffExampleSource
  - SimpleValidation  
SimpleValidation
    - CHAID  
CHAID
    - ApplierChain  
OperatorChain
      - Test  
ModelApplier
      - Performance  
Performance

Parameters XML Comment New Operator

keep_example_set	<input type="checkbox"/>
use_example_weights	<input checked="" type="checkbox"/>

```
autre: 144 728  
----precision: 83.49% (positive class: autre)  
ConfusionMatrix:  
True: detritus autre  
detritus: 264 34  
autre: 144 728  
----recall: 95.54% (positive class: autre)  
ConfusionMatrix:  
True: detritus autre  
detritus: 264 34  
autre: 144 728  
----AUC: 0.637 (positive class: autre)  
]  
(created by Performance)
```

Max: 1.1 CE  
Total: 1.1 CE


1:51:39 PM



Weka 3 - Data Mining with ...

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

Galerie de composant... Hotmail Sites suggérés Autres favoris

 **WEKA**  
The University of Waikato

**Software**

[project](#) • [software](#) • [book](#) • [publications](#) • [people](#) • [related](#)

Home

**Getting started**

[Requirements](#)

[Download](#)

[Documentation](#)

[FAQ](#)

[Citing Weka](#)

**Further information**

[Datasets](#)

[Related Projects](#)

[Miscellaneous Code](#)

[Other Literature](#)

**Developers**

[Development](#)

[History](#)

[Subversion](#)

[Contributors](#)

**Various**

## Weka 3: Data Mining Software in Java

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or can be used to process data before applying other tools for processing, classification, regression, etc. Weka is also used for developing new machine learning algorithms.

Weka is open source software is licensed under the GNU General Public License.

**Pentaho's live forum for Weka**

The open-source BI software company Pentaho has taken over the administration of Weka. This is a good thing for interaction among Weka project members.

**The Weka mailing list**

Please post Weka-related questions to the Weka mailing list. Please check out the [online documentation](#) and the [mailing list archive](#) (Mirrors: [news.gmane.org](#)) for more information about Weka problems.

University of Waikato  
Licence GNU

Un nombre « monumental » de techniques  
Quasi monopole pendant longtemps

Pas de documentation mais un livre payant  
Tutoriels par les aficionados

Piloté par menu  
Mode diagramme

Mais quel avenir ? cf. version Pentaho

Google

Data Mining Tools Used Poll Pentaho Commercial Open ...

http://weka.pentaho.org/

Galerie de composant... Hotmail Sites suggérés Autres favoris

Customer Portal | Partner Portal | Forum | Contact Us

pentaho™  
open source business intelligence™

Search:

Home Products Services Partners **Community** About Enterprise Edition

Pentaho Reporting Kettle Mondrian **Weka**

## Pentaho Data Mining

Pentaho Data Mining, based on Weka project, is a comprehensive set of tools for machine learning and data mining. Its broad suite of classification, regression, association rules and clustering algorithms can be used to help you understand the business better and also be exploited to improve future performance through predictive analytics.

### Recent News and Releases

- 06/05/09 Weka 3.7.0 is [now available](#).
- 06/05/09 Weka 3.6.1 is [now available](#).
- 06/05/09 Weka 3.4.15 is [now available](#).
- 06/05/09 English documentation for Weka 3.7.0 is [now available](#).
- 06/05/09 English documentation for Weka 3.6.1 is [now available](#).
- 12/19/08 Weka 3.4.14 is [now available](#).
- 12/19/08 English documentation for Weka 3.6.0 is [now available](#).
- 12/19/08 English documentation for Weka 3.4.14 is [now available](#).
- 12/15/08 National Health Service Islington Selects Pentaho Business Intelligence to Improve Patient Services ([press release](#)).
- 09/11/08 Support for importing PMML models into Weka ([press release](#)).
- 12/06/07 Weka Plugins for Pentaho Data Integration 3.0 are [now available](#).
- 12/06/07 Pentaho streamlines delivery of predictive analytics ([press release](#)).

### How to Contribute

You can participate by contributing new code, reporting bugs, testing new releases, answering questions and more; [Email us](#) the proposed contribution and any other relevant details. Welcome to the team.

- [Write a tech tip](#)
- [Report a bug in JIRA](#)
- [Answer posts on the forums](#)
- [Write some code](#)

### Stable

**Weka 3.4.15 (GA) (Release Notes)**  
This is a patch release to Weka 3.4 containing a number of bug fixes. For a detailed list of improvements, please refer to the release notes.

- [Download\(s\)](#) - [Source](#) - [Read me](#)
- [Documentation](#) - [Forum](#)

**New Features since 3.2**

### Whats Next

To suggest a new feature or view our roadmap, [click here](#).

Major features planned in future releases:

- Further PMML support (import/export)
- Pluggable estimators in EM
- Execution of Kettle transforms in KnowledgeFlow
- KnowledgeFlow plugin for Kettle



[Data Mining Tools Used Poll](#) | [Discover why Pentaho Dat...](#)

[http://www.pentaho.com/products/data\\_mining/discover\\_data\\_mining.php](http://www.pentaho.com/products/data_mining/discover_data_mining.php)

[Galerie de composant...](#) | [Hotmail](#) | [Sites suggérés](#) | [Autres favoris](#)

[Products](#) | [Support & Services](#) | [Partners](#) | [Community](#)

**Pentaho BI Suite Enterprise Edition**

**Download**

# Pentaho Data Mining Enterprise Edition

Gain insight into hidden patterns and relationships in your data to discover indicators of future performance.

[Discover](#) | [Try](#) | [Buy](#)

## Pentaho Data Mining

Once you've got analysis, reporting, and dashboards deployed, it's time to take your business intelligence (BI) to the next level by adding data mining and advanced analytics. This is a level of BI excellence that many organizations never manage to evolve to, however the importance of pushing ahead with advanced capabilities cannot be underestimated - they can provide a truly sustainable competitive advantage and enable your organization to maximize both its efficiency and effectiveness.

Data Mining is the process of running data through sophisticated algorithms to uncover meaningful patterns and correlations that may otherwise be hidden. These can be used to help you understand the business better and also exploited to improve future performance through predictive analytics. For example, data mining can warn you there's a high probability a specific customer won't pay on time based on an analysis of customers with similar characteristics.

**Explore and Learn**  
**Resources**

- [Data Sheet](#)
- [User Forum](#)
- [Download](#)
- [Deploying Data Mining Models](#)
- [PMML Support](#)
- [Request a Quote](#)

**Popular Links**

- [BI Economics White Paper](#)
- [More Links](#)

**Version 3**  
 Pentaho BI Suite Enterprise Edition

**User Friendly**  
**Cloud Ready**  
**Community Powered**

**Pentaho BI Suite 3.5**  
 Design. Deploy. Escape.

# Weka

## Utilisation en mode « Knowledge flow » »

The screenshot displays the Weka KnowledgeFlow Environment interface. At the top, there are tabs for DataSources, DataSinks, Filters, Classifiers, Clusterers, Associations, Evaluation, and Visualization. Below these tabs is a toolbar with icons for Data Visualizer, Scatter PlotMatrix, Attribute Summarizer, Model PerformanceChart, Text Viewer, Graph Viewer, and Strip Chart. The main area is titled "Knowledge Flow Layout" and shows a workflow diagram. The workflow starts with an ArffLoader, followed by a ClassAssigner, then a Train Test SplitMaker. The Train Test SplitMaker outputs a training Set and a test Set to a J48 classifier. The J48 classifier outputs text to two Text Viewer components. The J48 classifier also outputs a batch Classifier to a Classifier Performance Evaluator, which then outputs text to another Text Viewer component. Below the workflow diagram is a Status Log table.

Component	Parameters	...	Status
[KnowledgeFlow]		0:...	Welcome to the Weka Knowledge Flow
ArffLoader		-	Finished.
J48	-C 0.25 -M 2	-	Finished.
ClassifierPerforman...		-	Finished.

The Text Viewer window displays the evaluation results for the J48 classifier. The text content is as follows:

```
Result list    Text
14:13:00 - J48  === Evaluation result ===

Scheme: J48
Options: -C 0.25 -M 2Relation: zooplancton.arff

Correctly Classified Instances      1047      89.4872 %
Incorrectly Classified Instances    123       10.5128 %
Kappa statistic                     0.7675
Mean absolute error                  0.1203
Root mean squared error              0.3114
Relative absolute error              26.2522 %
Root relative squared error          65.0585 %
Total Number of Instances           1170
```

# 5. Démonstration

## Arbres de décision

Construction de l'arbre sur un échantillon d'apprentissage  
Évaluation sur un échantillon test

## Règles d'association

Construction et organisation des règles d'association

## Traitement des gros volumes

Tutoriel Arbre + Comparatif diapo suivante

# Performances comparées

## Gros volumes

Logiciel	Temps de traitement (secondes)		Occupation mémoire (Mo)			
	Importation	Induction arbre	Avant lancement	Après importation	Pic traitement	Après induction
KNIME	47	270	92.6	160.4	245.8	245.8
ORANGE	90	130	24.9	259.5	795.7	795.7
R (package rpart)	24	157	18.8	184.1	718.9	718.9
RAPIDMINER	7	298	136.3	228.1	1274.4	1274.4
SIPINA	25	122	7.8	67.1	539.9	539.9
TANAGRA	11	33	7.0	53.1	121.6	73.5
WEKA	10	338	52.3	253.2	699.6	699.6

Arbres de décision, 500.000 obs., 21 descripteurs

Logiciel	Temps de traitement (sec.)		Taux d'erreur en validation croisée (%)	Occupation mémoire (Mo)			
	Importation	Calcul		Au lancement	Avec les données	Durant le traitement	Durant la validation croisée
ORANGE	95	690	4% (6/135)	25	118	317	406
RAPIDMINER JMySVMLearner	5	29	11% (15/135)	124	210	338	608
RAPIDMINER C-SVC (LIBSVM)	5	9	2% (3/135)	124	210	442	870
TANAGRA - SVM	12	130	4% (6/135)	7	337	393	393
TANAGRA C-SVC (LIBSVM)	12	11	4% (6/135)	7	337	406	406
WEKA - SMO	11	12	3% (4/135)	54	243	489	595

**SVM, 135 obs., 31809 descripteurs**

**Les outils se tiennent, tout dépend des méthodes et des caractéristiques des données !!!**



# 6. Conclusion

## Conclusion

Quel logiciel pour quel contexte ?

### Recherche (Data Mining)

Développer de nouvelles techniques  
Les intégrer dans un environnement opérationnel  
Pour des comparaisons à grande échelle  
Les diffuser simplement et largement



Logiciel R  
Avec les Packages

### Utilisateur (Ou Recherche autre que Data Mining)

Contexte d'exploration des données  
i.e. appliquer les techniques à des données,  
Les faire coopérer (ces techniques)  
Interpréter et publier les résultats  
Enseignement



### Les outils se valent

#### Critères de différenciation

Manipulation des données - texte/tableur/sqldb  
Pouvoir les enchaîner (tous)  
Traitement des très gros volumes (Knime ?)  
Profusion des techniques (oui et non)  
Outils graphiques (Knime, Orange)  
Notoriété (Weka)

Et TANAGRA ?

Culture francophone du traitement des données

Machine Learning + Analyse de données et statistique

Un effort constant sur la documentation !!!