

# Tanagra

Logiciels libres, spécificités et applications

Ricco RAKOTOMALALA

Université Lyon 2

Laboratoire ERIC

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

---

---

# Ricco ?

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. Une application : classement de planctons

*Comment faire coopérer les techniques dans une seule plate-forme*

5. Comparaison avec les autres logiciels libres

*(Knime, Orange, R, RapidMiner, Weka,...)*

---

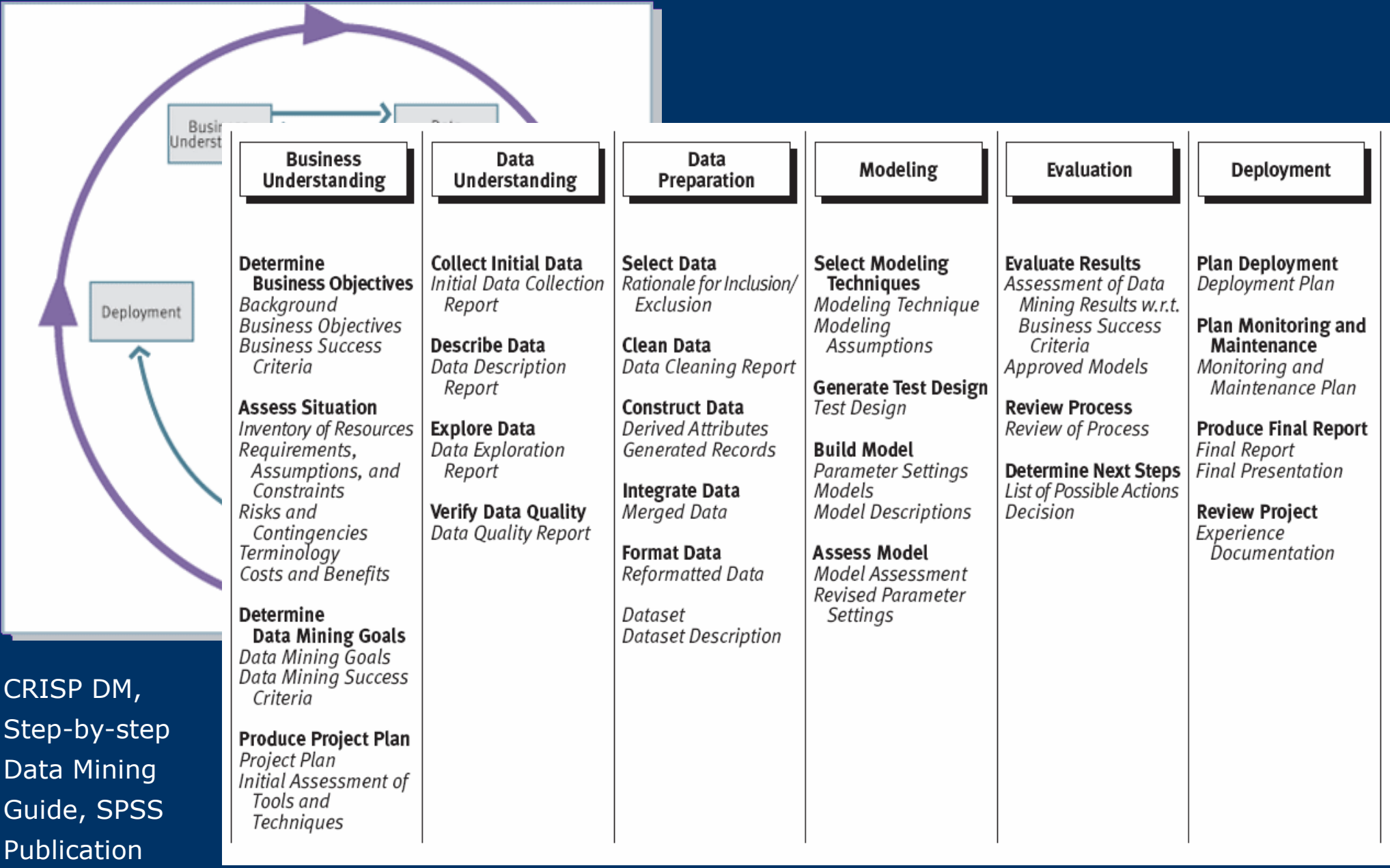
---

# 1. Data Mining



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

# Data Mining



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 ?

<http://cedric.cnam.fr/~saporta/DM.pdf>

L'analyse des données est un outil pour dégager de la gangue des données le pur diamant de la véridique nature.» (J.P.Benzécri1973)

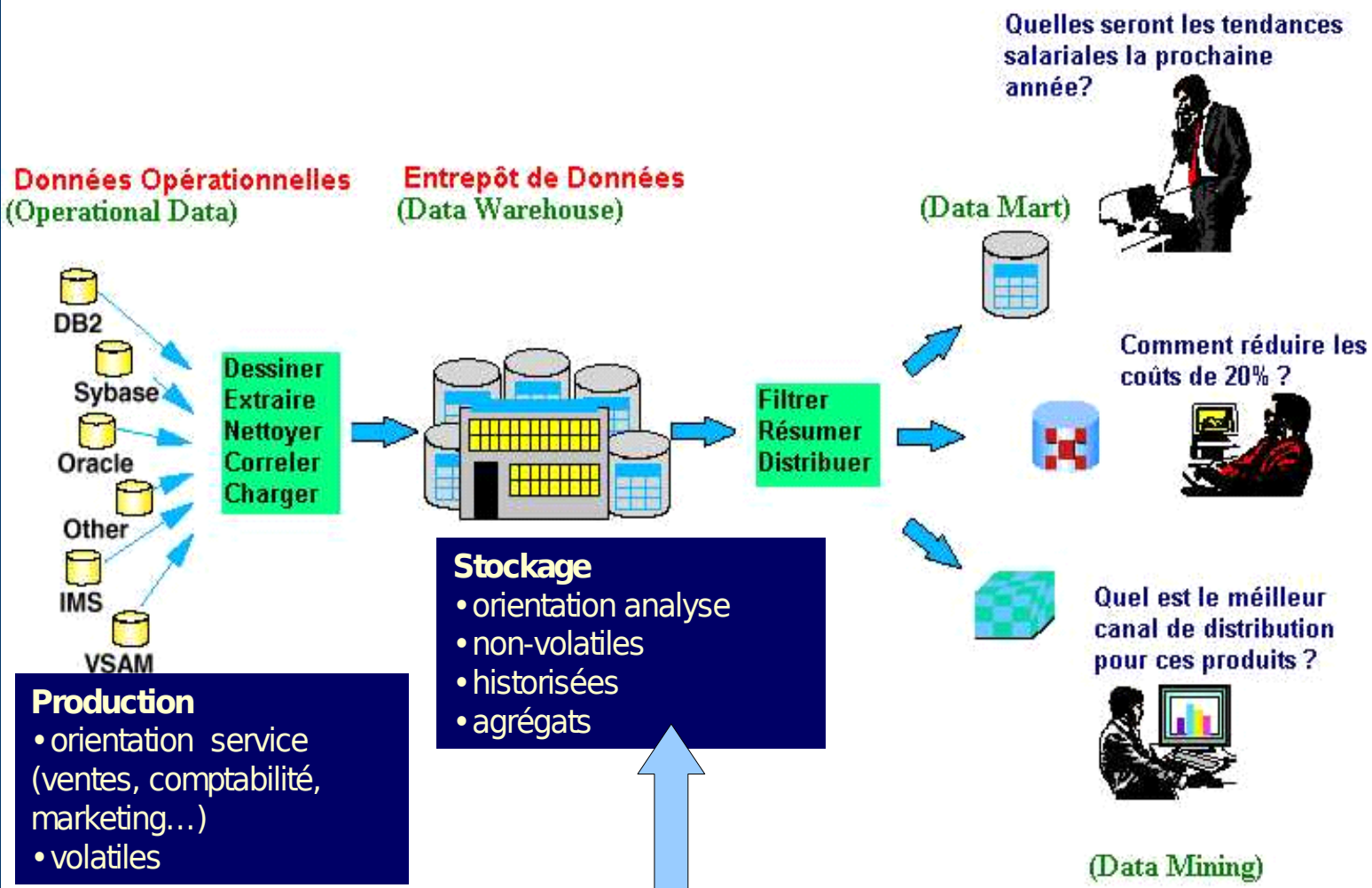
The basic steps for developing an effective process model ?

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

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

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



Problème de volumétrie

## 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				

Informatique  
« Machine Learning »  
Apprentissage symbolique  
Reconnaissance de formes

Une étape de l'intelligence artificielle  
Réseaux de neurones, algorithmes génétiques...

Informatique  
(Base de données)  
Exploration des bases de données

Volumétrie  
Règles d'association, motifs fréquents, ...

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

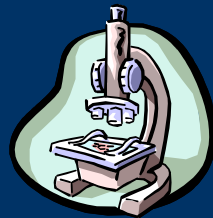


# Traitement des données non structurées

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, ...)

Reconnaissance de la langue à une centrale téléphonique

Analyse des mammographies

Etc.

## 2. Data Mining et logiciel libre

*Attention, les informaticiens 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 »

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

Pouvoir utiliser sans mauvaises surprises

### 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. Relief 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 "Analysis", "Tree management", "View", "Window", and "Help". The "Analysis" menu is open, showing options like "Define class attribute...", "Select active examples...", "Set weight field...", "Set priors...", "Set costs...", "Learning...", "Stop analysis", "Classification", "Test...", "LIFT -- ROC curve...", "Error measurements", "Feature selection", and "Personal tests". The "Classification" option is selected, and a sub-menu is open, showing "on same dataset" and "on other dataset". The "on same dataset" option is further expanded, showing "generate scores" and "on other dataset".

The background shows a data table with columns: sep\_length, sep\_width, pet\_length, pet\_width, and type. The data rows are as follows:

sep_length	sep_width	pet_length	pet_width	type
5.10	3.50	1.40	0.20	Iris-setosa
4.90	3.00	1.40	0.20	Iris-setosa
4.70	3.20	1.30	0.20	Iris-setosa
4.60	3.10	1.50	0.20	Iris-setosa
		1.40	0.20	Iris-setosa
		1.70	0.40	Iris-setosa
				Iris-setosa

The foreground shows a decision tree diagram for classification. The root node is "pet\_length" with a split at 2.45. The left branch (< 2.45) leads to a leaf node with 50 (100%) Iris-setosa. The right branch (>= 2.45) leads to a node with 0 (00%) Iris-setosa and 50 (50%) each of Iris-versicolor and Iris-virginica. This node splits on "pet\_width" at 1.75. The left branch (< 1.75) leads to a leaf node with 0 (00%) Iris-setosa, 49 (91%) Iris-versicolor, and 5 (09%) Iris-virginica. The right branch (>= 1.75) leads to a leaf node with 0 (00%) Iris-setosa and 45 (98%) Iris-virginica.

The bottom left panel shows the "Learning method" section with the following settings:

- MethodName=Improved ChAID (Tschuprow Goodness of Split)
- MethodClassName=TArbreDecision
- Hdl=8
- Merge=0.05
- Split=0.001
- TypeBonferroni=1
- ValueBonferroni=1
- Sampling=0

The bottom right panel shows the "Examples selection" section with the following settings:

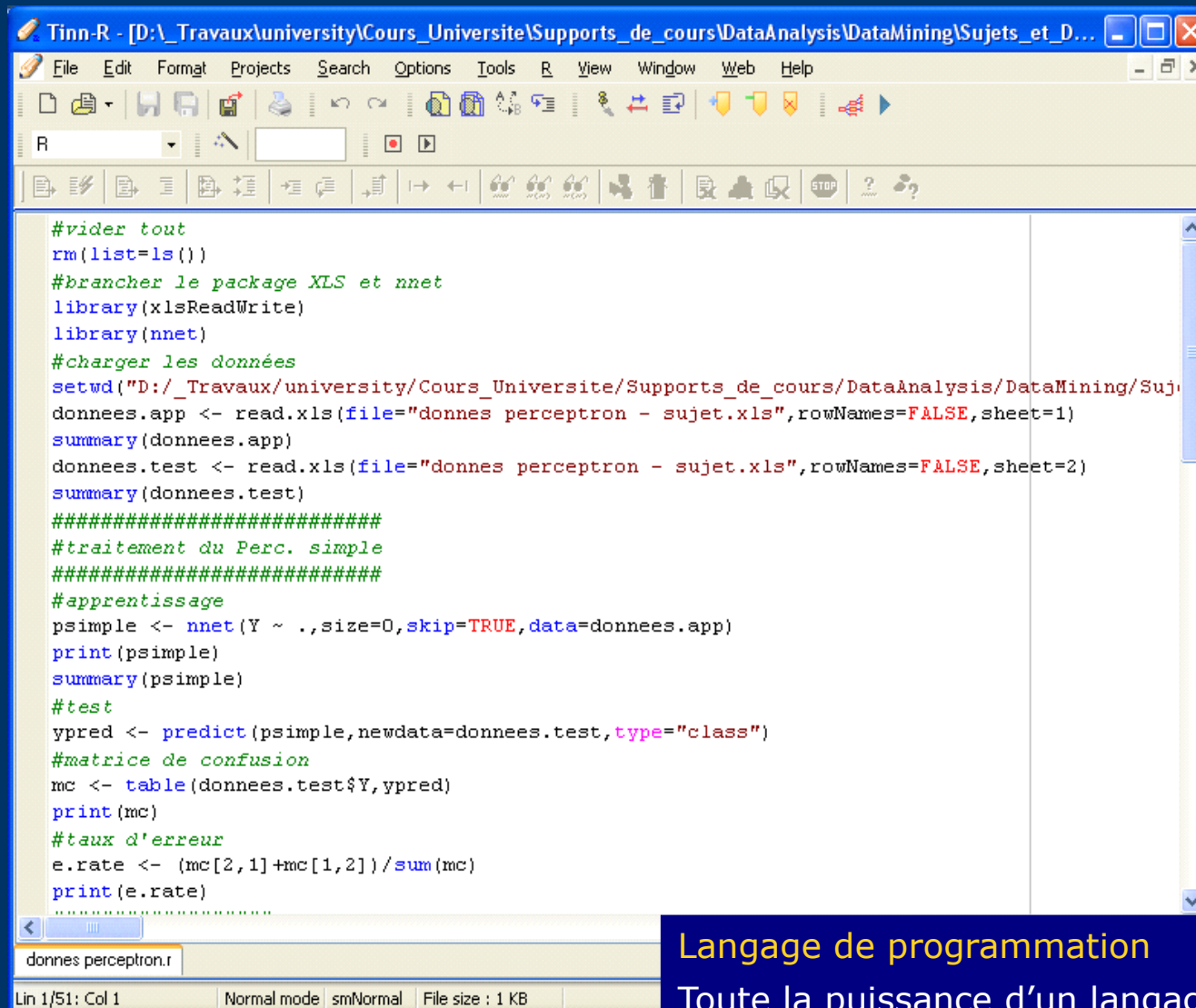
- 150 examples selected
- 0 examples idle

**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



```
#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 is visible under the 'Default title' window. The workflow includes: 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), Cross-validation 4.

On the right, the results for Cross-validation 4 are displayed:

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

At the bottom, a 'Components' panel lists various machine learning and statistical methods:

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

## 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

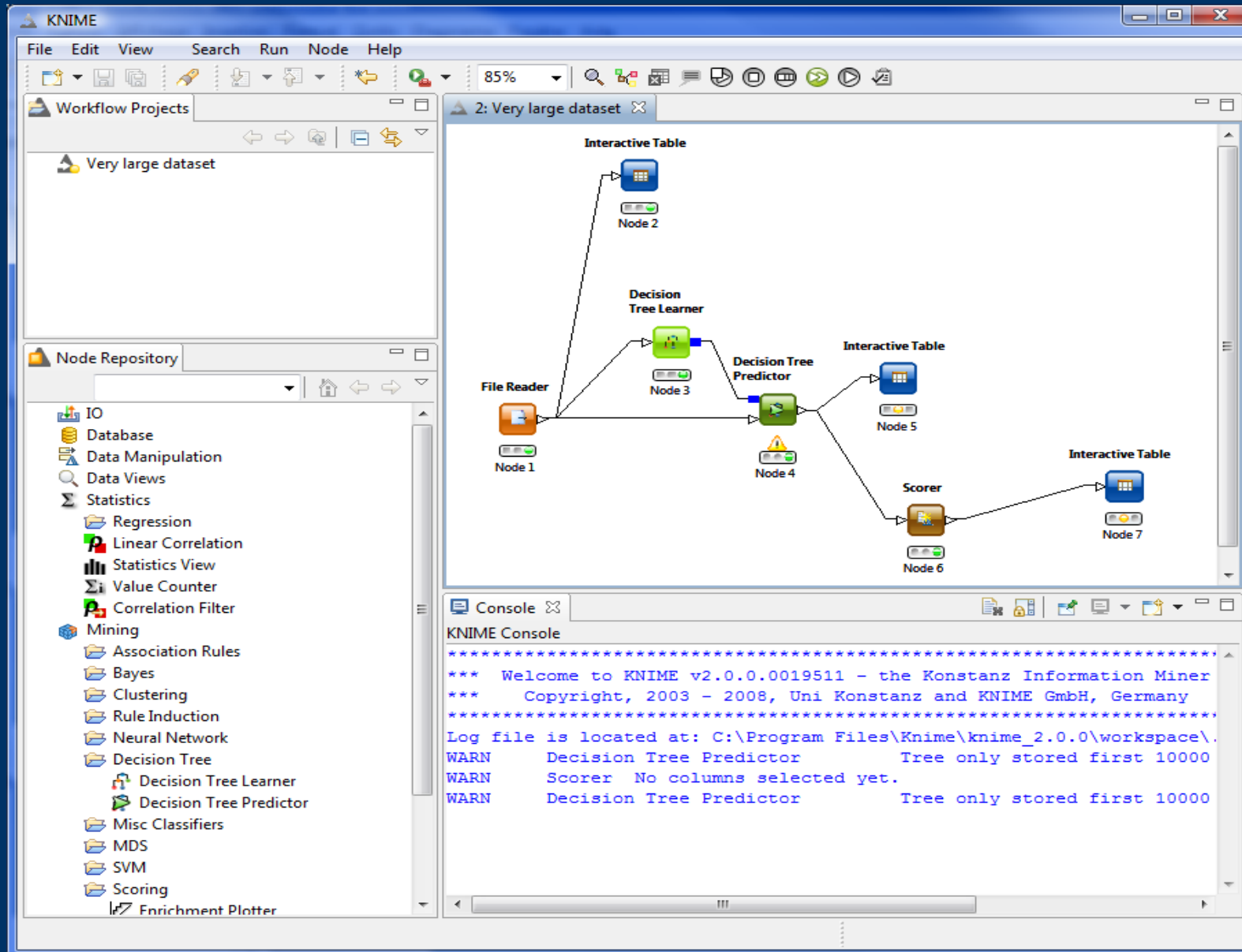


Diagramme de traitements

Une autre manière de présenter les diagrammes de traitements

# 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

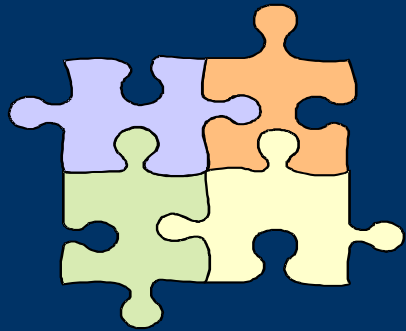
Gestion simplifiée des données - Format texte et accès au format tableur

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

### Réduire les bibliothèques externes

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

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 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). A blue arrow points to this diagram with the text 'Enchaînement des traitements'. On the right, a results window titled 'Overall cross-validation error rate' displays a table with the error rate of 0.0420. A blue arrow points to this window with the text 'Fenêtre de visualisation des résultats'. At the bottom, a 'Components' panel lists various data mining methods, with a blue arrow pointing to it and the text 'Composants de calcul Méthodes de data mining'.

**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**

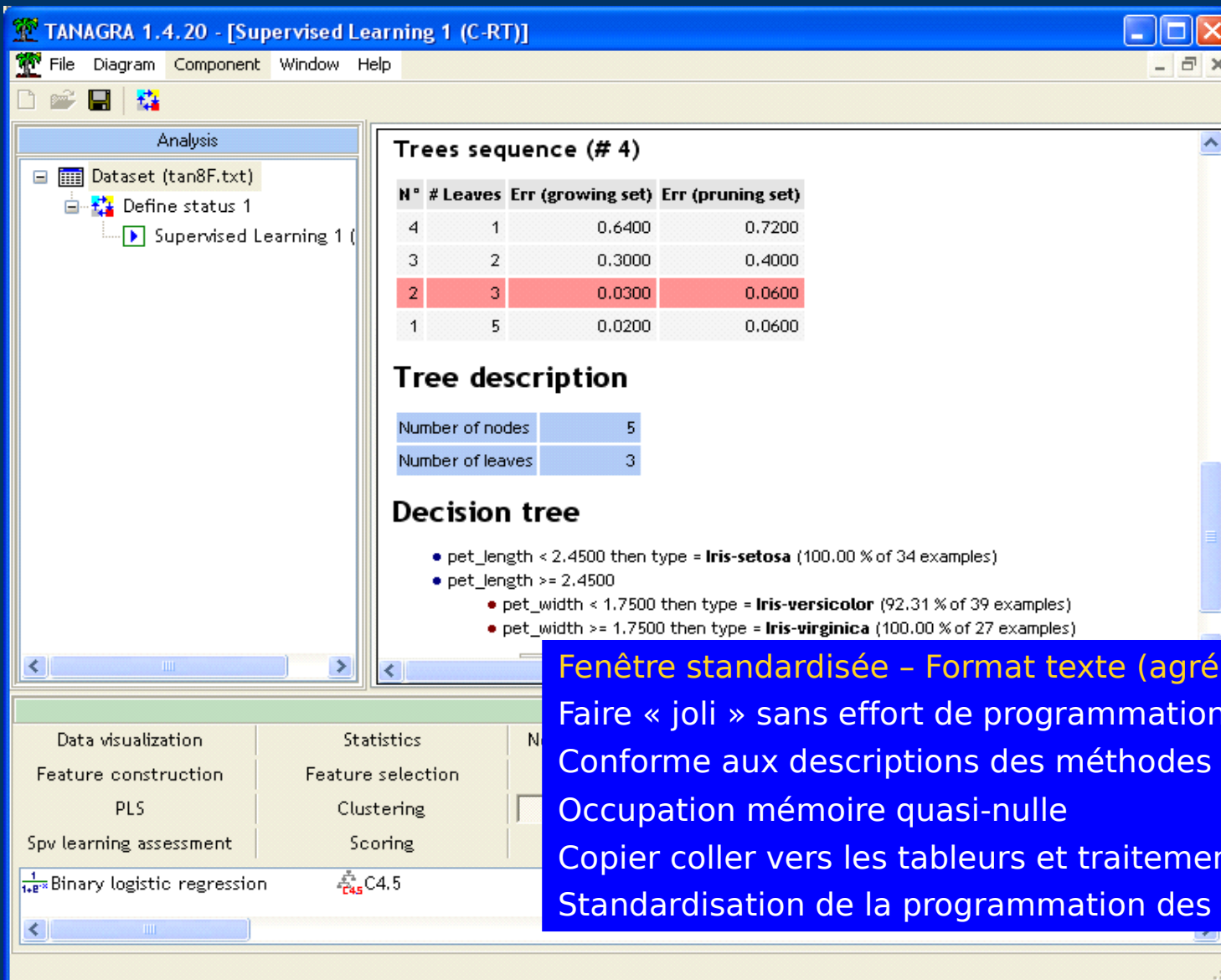
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construction	Feature selection
Regression	Factorial analysis	PLS	Clustering	Spv learning	Meta-spv learning
Spv learning assessment	Scoring	Association			

Binary logistic regression   C-RT   C-SVC  
C4.5   CS-CRT   Decision List   K-NN  
C-PLS   CS-MC4   ID3   Linear discriminant analysis   Multilayer perceptron  
Log-Reg TRIRLS   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 software interface. The main window displays the results of a supervised learning analysis. The left sidebar shows the project structure: Dataset (tan8F.txt) > Define status 1 > Supervised Learning 1. The main area is titled "Trees sequence (# 4)" and contains a table with the following data:

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

Below the table, the "Tree description" section shows:

- Number of nodes: 5
- Number of leaves: 3

The "Decision tree" section shows the following rules:

- 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 "Data visualization" and "Statistics" section with various options like "Feature construction", "PLS", "Spv learning assessment", "Feature selection", "Clustering", and "Scoring".

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 (1/3)

### 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;
```

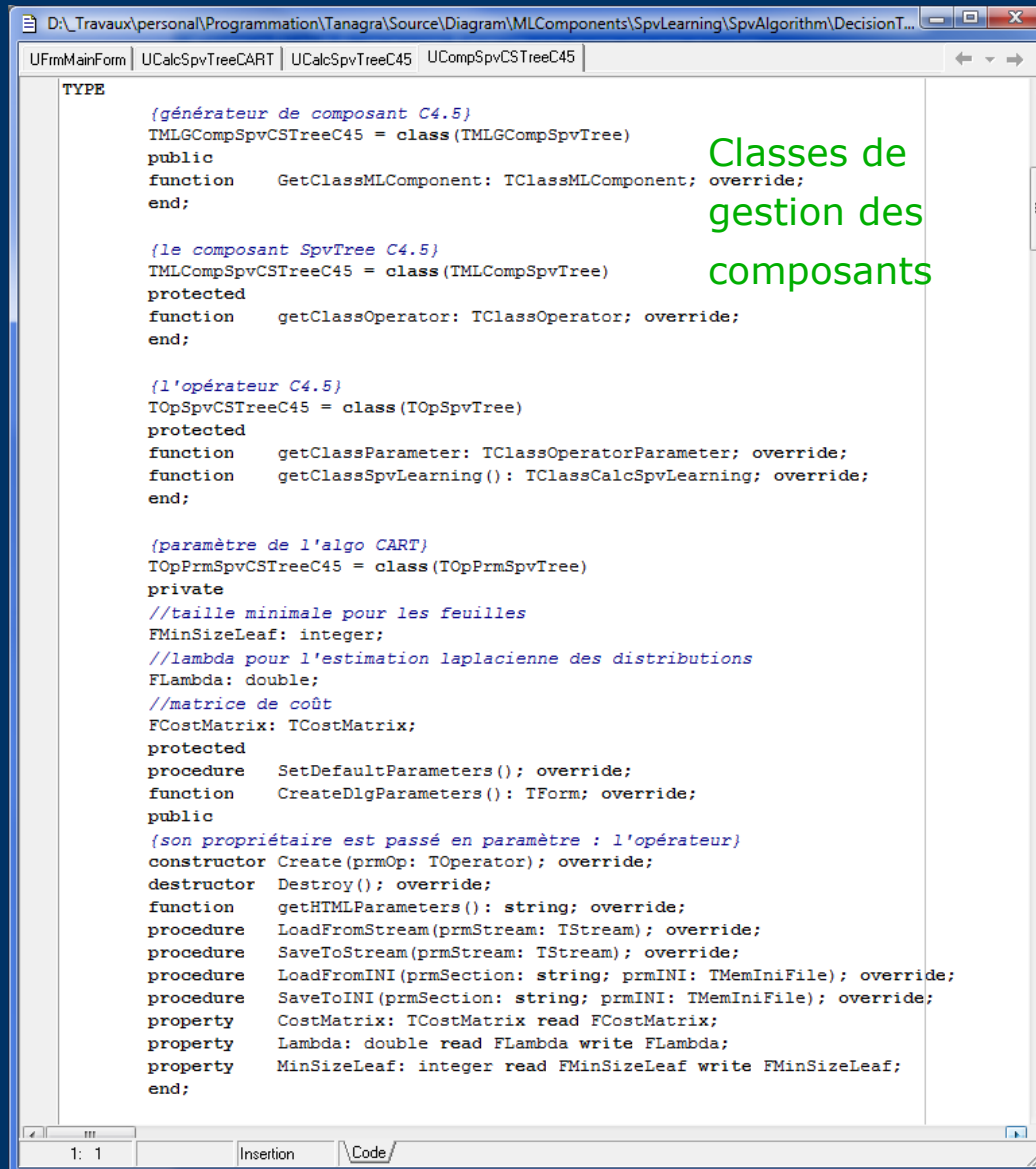
1: 1

Insertion

Code

# Simplicité pour les programmeurs

## Vive la programmation objet (2/3)



```
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(TOPrmSpvTree)
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;
```

Classes de gestion des composants

# Simplicité pour les programmeurs

## Vive la programmation objet (3/3)

```
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

# 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 ?

A l'époque, DELPHI 6.0 PERSO était 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

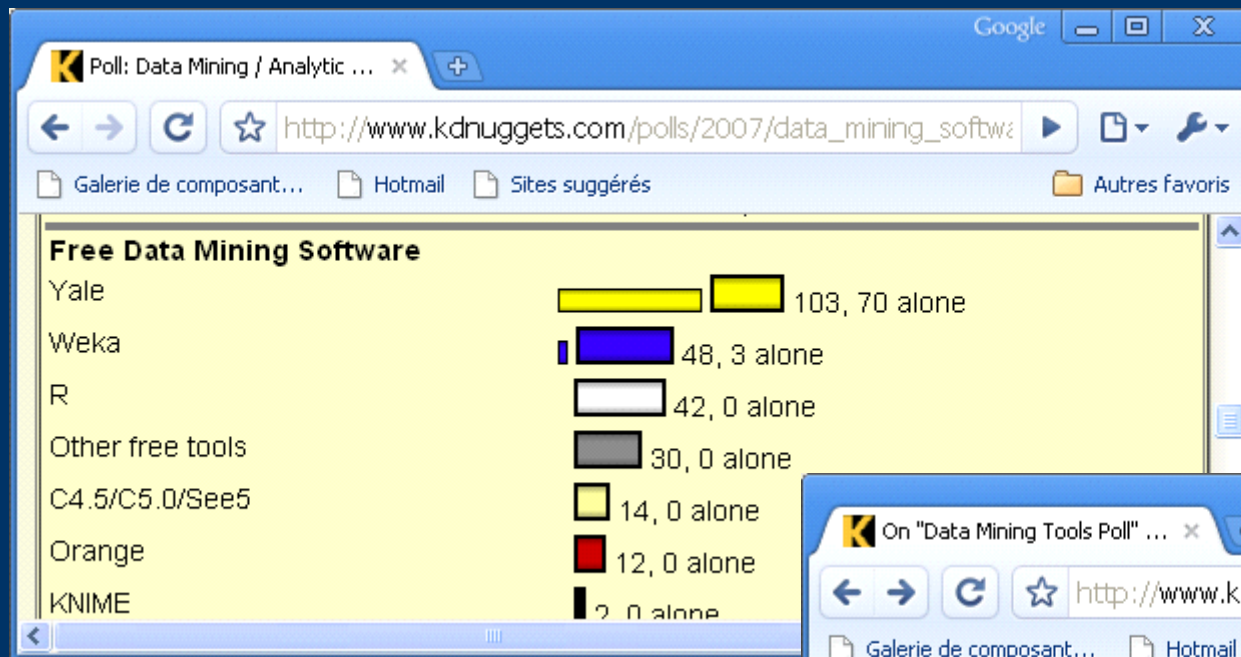
Affinités personnelles...

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](http://www.kdnuggets.com/polls/2007/data_mining_software_tools.html)

This should be accounted for in next years poll.

Dr. Alexander K. Seewald  
alexATseewaldDOTat

## Promotion

Comment faire connaître le logiciel sans tomber dans le « commercial »

### 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



### Documenter le logiciel

Documenter les méthodes : description théorique

Documenter leur mise en oeuvre : les tutoriels

Facilitée par le découpage en « composants » du logiciel

### Monter un site web attrayant (attractif)

La visibilité internet est primordiale

Le téléchargement du logiciel n'est pas le seul enjeu

...et la promotion dans les conférences

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

# Promotion

## Le site web Tanagra



The screenshot shows a web browser window with the following elements:

- Browser Tabs:** "Tutoriels Tanagra pour le D..." and "TANAGRA - Un logiciel de d...".
- Address Bar:** "http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html".
- Navigation:** Back, Forward, Refresh, and Home buttons.
- Site Header:** A blue banner with a palm tree icon and the word "TANAGRA".
- Menu:** A row of icons and labels: "Présentation", "Galerie", "Caractéristiques", "Didacticiels", "Téléchargement", and "Sipina".
- Left Sidebar:** A vertical menu with "Présentation" selected, and other items: "Projet TANAGRA", "Nouveautés **NEW!**", "Historique", "Références", "Autres logiciels", "Portail Data Mining", and "Contact".
- Main Content:**
  - Section Header:** "Le projet TANAGRA".
  - Text 1:** "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."
  - Text 2:** "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."
  - Text 3:** "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."
  - Text 4:** "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."
  - Text 5:** "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."
  - Text 6:** "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..."



# 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 displays 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 icons for "Favoris", "Sites suggérés", "Hotmail", and "Galerie de composants W...". The page content is organized into a sidebar on the left and a main content area on the right.

**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

Google Chrome browser window showing a website with the URL [http://eric.univ-lyon2.fr/~ricco/cours/suppports\\_data\\_mining.html](http://eric.univ-lyon2.fr/~ricco/cours/suppports_data_mining.html). The page contains a table of statistical tests and their associated PDF documents.

Section	Description	PDF Document	Thumbnail	Thumbnail	Thumbnail
<b>NOS FORMATIONS</b> Département Info-Stat Diplômes Licence - Master Formation continue	<b>Test de normalité</b> 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.				
<b>TUTORIELS</b> Portail Data Mining Tutoriels pour le Data Mining	<b>Corrélation et corrélation partielle</b> 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.				
<b>LOGICIELS</b> Tanagra (Open Source) Sipina - Arbres de décision	<b>Mesures d'association pour variables nominales</b> 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).				
<b>VIDÉOS</b> Machine Learning (A. Ng) Statistical Aspects (D. Mease)	<b>Comparaison de populations - Tests paramétriques</b> 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.				
<b>REF. EXTERNES</b> Applied Statistics Data Mining tutorials DM and Analytic Technologies Statnotes : Topics in MVA	<b>Comparaison de populations - Tests non paramétriques</b> 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

Adobe Acrobat Standard window showing a PDF document titled "Comp\_Pop\_Tests\_Nonparametriques.pdf". The document content includes the author name "Ricco Rakotomalala" and the title "Comparaison de populations Tests non paramétriques". The version is noted as "Version 1.0".

# Promotion

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

~130 tutoriels en français à ce jour (09/2009)



### Tutoriels Tanagra pour le Data Mining

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

Affichage des messages en fonction de la requête **régression logistique**.  
[Afficher tous les messages](#)

MARDI 7 OCTOBRE 2008

#### ➤ Régression logistique - Comparaison de logiciels

La régression logistique est une technique prédictive, très populaire dans la communauté statistique. Je ne sais pas si elle est très utilisée parce que très enseignée, ou très enseignée parce que largement utilisée. En tous les cas, on ne peut pas passer à côté si on s'intéresse un tant soit peu au Scoring c.-à-d. aux configurations où l'on souhaite prédire ou expliquer les valeurs d'une variable discrète (nominale ou ordinale) à partir d'une série de descripteurs (de type quelconque).

Les raisons de cet engouement sont nombreuses. La régression logistique s'intègre dans un cadre théorique parfaitement identifié, celui de la régression linéaire généralisée. C'est une technique semi paramétrique. Son champ d'application est large. Par rapport aux techniques issues de l'apprentissage automatique, elle intègre les outils de la statistique inférentielle. Enfin, autre atout fort, la lecture des coefficients sous forme de surcroît de risque (les fameux « odds ratio ») donne aux utilisateurs un outil de choix pour comprendre l'essence de la relation entre les descripteurs et la variable à prédire.

La régression logistique est implémentée dans tous les logiciels de statistique commerciaux. Elle est plus rare en revanche dans les logiciels libres. En partie parce que la méthode est peu connue des informaticiens, ceux qui sont les plus enclins à programmer des outils. La situation change quand même un peu maintenant. Avec le label « data mining », il y a un certain brassage des cultures. On peut parler de « faire une régression » sans que certaines personnes ne s'imaginent que vous êtes en train de retomber en enfance.

#### Supports et tutoriels

- Page principale du blog
- Cours Data Mining
- Portail Data Mining

#### Logiciels

- Site du logiciel Tanagra
- Téléchargement Tanagra
- Site du logiciel Sipina

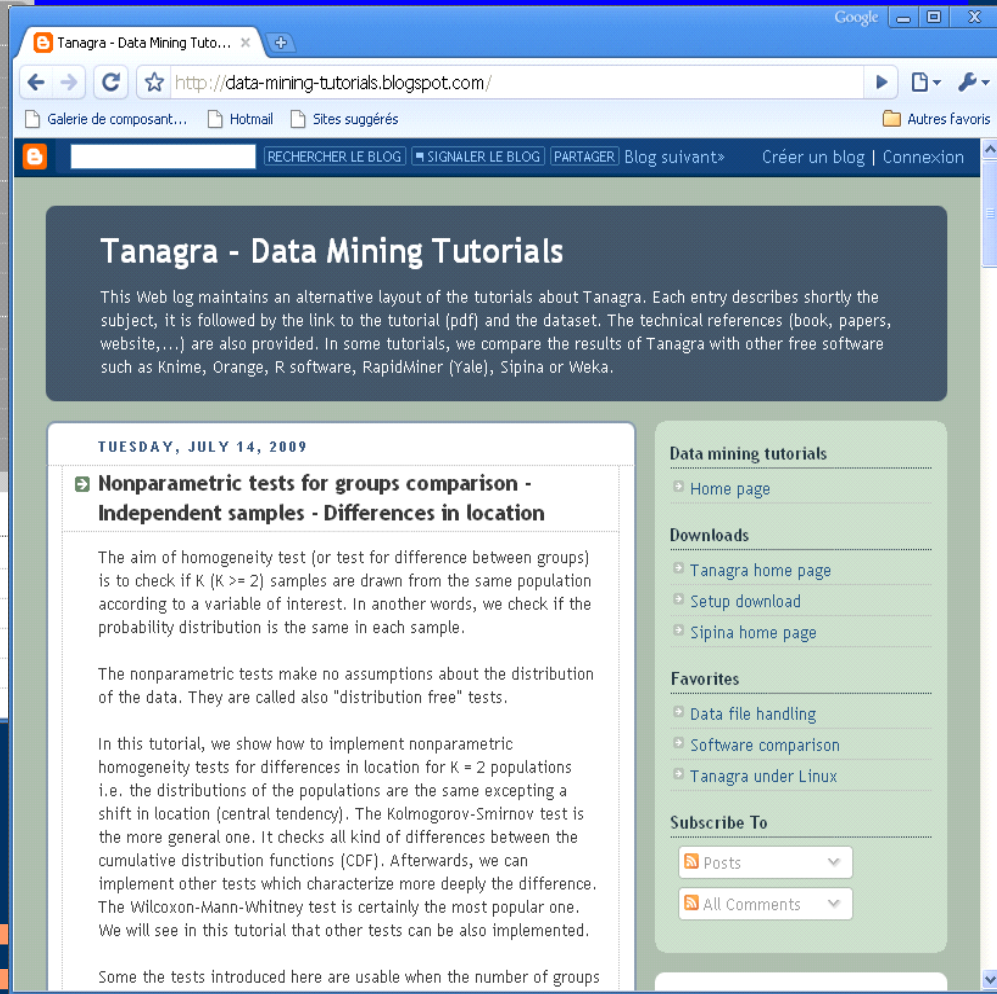
#### Favoris

- Tanagra - Etat des lieux
- Tanagra sous Linux
- Comparaison de logiciels
- Tutoriels en anglais

#### Catégories des tutoriels

- Analyse discriminante (10)
- Analyse factorielle (15)
- App. Supervisé - Scoring (37)
- Arbres de décision (18)
- Classification - Clustering (12)
- Construction de variables (5)

~90 tutoriels en anglais à ce jour (09/2009)



### Tanagra - Data Mining Tutorials

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.

TUESDAY, JULY 14, 2009

#### ➤ Nonparametric tests for groups comparison - Independent samples - Differences in location

The aim of homogeneity test (or test for difference between groups) is to check if  $K$  ( $K \geq 2$ ) samples are drawn from the same population according to a variable of interest. In another words, we check if the probability distribution is the same in each sample.

The nonparametric tests make no assumptions about the distribution of the data. They are called also "distribution free" tests.

In this tutorial, we show how to implement nonparametric homogeneity tests for differences in location for  $K = 2$  populations i.e. the distributions of the populations are the same excepting a shift in location (central tendency). The Kolmogorov-Smirnov test is the more general one. It checks all kind of differences between the cumulative distribution functions (CDF). Afterwards, we can implement other tests which characterize more deeply the difference. The Wilcoxon-Mann-Whitney test is certainly the most popular one. We will see in this tutorial that other tests can be also implemented.

Some the tests introduced here are usable when the number of groups

#### 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



### 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 1.4.32 – Sept. 2009)

164 méthodes stat., analyse de données, data mining

### Documentation libre en ligne (Sept. 2009)

7 ouvrages libres en PDF

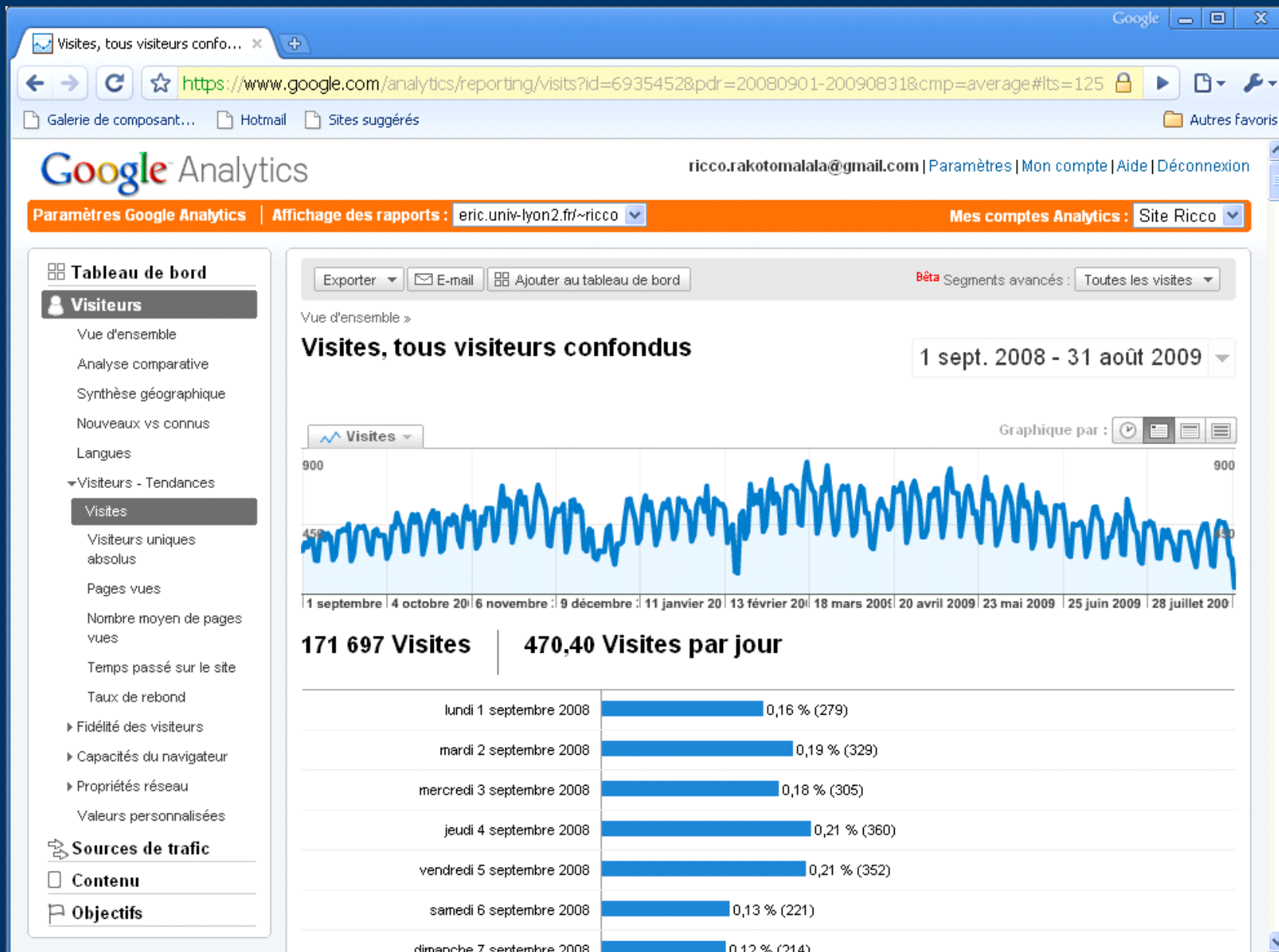
30 « slides » en PDF

130 didacticiels en français

90 didacticiels en anglais

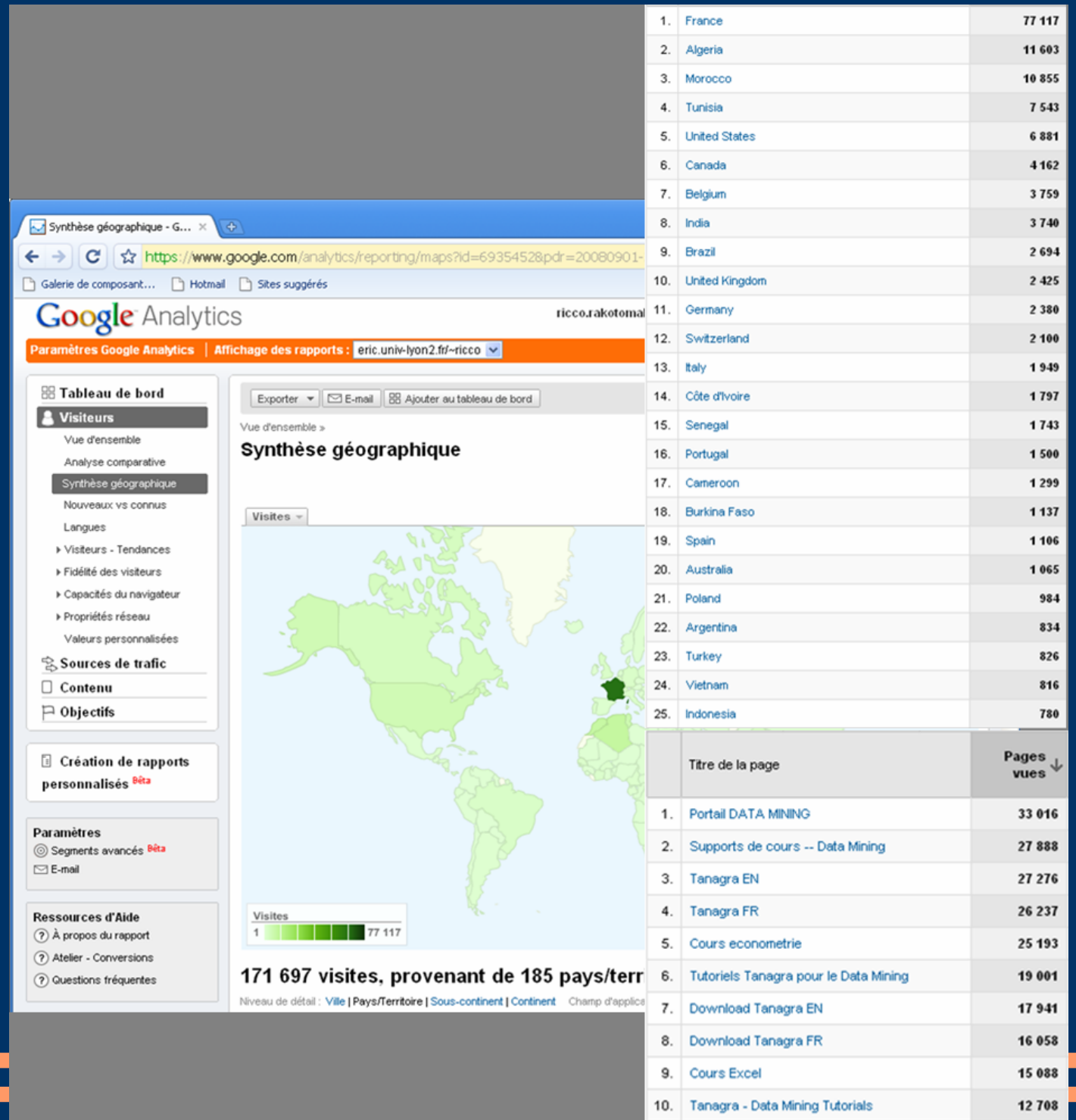
# Tanagra

## Bilan (2) – Diffusion 1/2



# Tanagra

## Bilan (2) – Diffusion 2/2

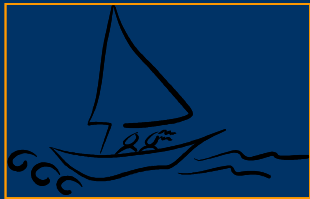


## 4. Classement automatique de planctons

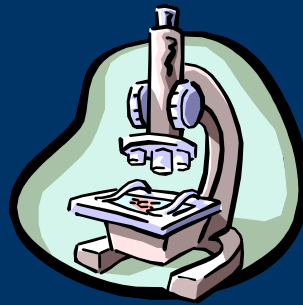


# Le projet ZOOSCAN

## Récupération des données



Campagne  
de pêche



Les prélèvements  
sont scannés :  
ZOOSCAN

Individu = Une  
image de plancton



L'expert  
étiquette  
manuellement  
les objets





# Le projet ZOOSCAN

## Construction des descripteurs



Image originelle  
fournie par le scanner

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

- Paramètres de niveau de gris  
*Mean, Mode, StdDev, etc.*
- Paramètres de taille  
*Area, Perim, etc.*
- Paramètres de forme  
*Circularity, Major, Minor, etc.*
- Paramètres de position  
*X, Y, XM, YM, etc.*

### Références

- Site WEB (Logiciel IMAGEJ)  
<http://rsb.info.nih.gov/ij/docs/menus/analyze.html>
- Voir aussi le fichier IMAGEJ\_Parameters.pdf

1. Classer le plus efficacement possible (avec ce qui est dispo)

2. Regrouper les classes de plancton

3. Produire de nouveaux descripteurs



5. 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.

KNIME supports visual exploration with many interactive charts.

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++

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

The screenshot displays the KNIME software interface with a workflow titled "4: Zooplankton analysis". The workflow consists of the following nodes:

- File Reader (Node 1)**: Reads data from a file.
- Column Filter (Node 2)**: Filters columns based on user-defined criteria.
- Partitioning (Node 3)**: Splits the data into training and testing sets.
- Decision Tree Learner (Node 4)**: Trains a decision tree model on the training data.
- Decision Tree Predictor (Node 5)**: Applies the trained model to the test data.
- Scorer (Node 6)**: Compares two columns by their attribute value pairs and shows the confusion matrix.
- Interactive Table (Node 7)**: Displays the output of the Scorer node in a table format.

The **Node Description** panel for the **Scorer** node is open on the right, providing the following information:

### 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**

The **Console** window at the bottom shows the following output:

```
*****  
*** 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 Fruit... x

Google

http://www.ailab.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.



(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++

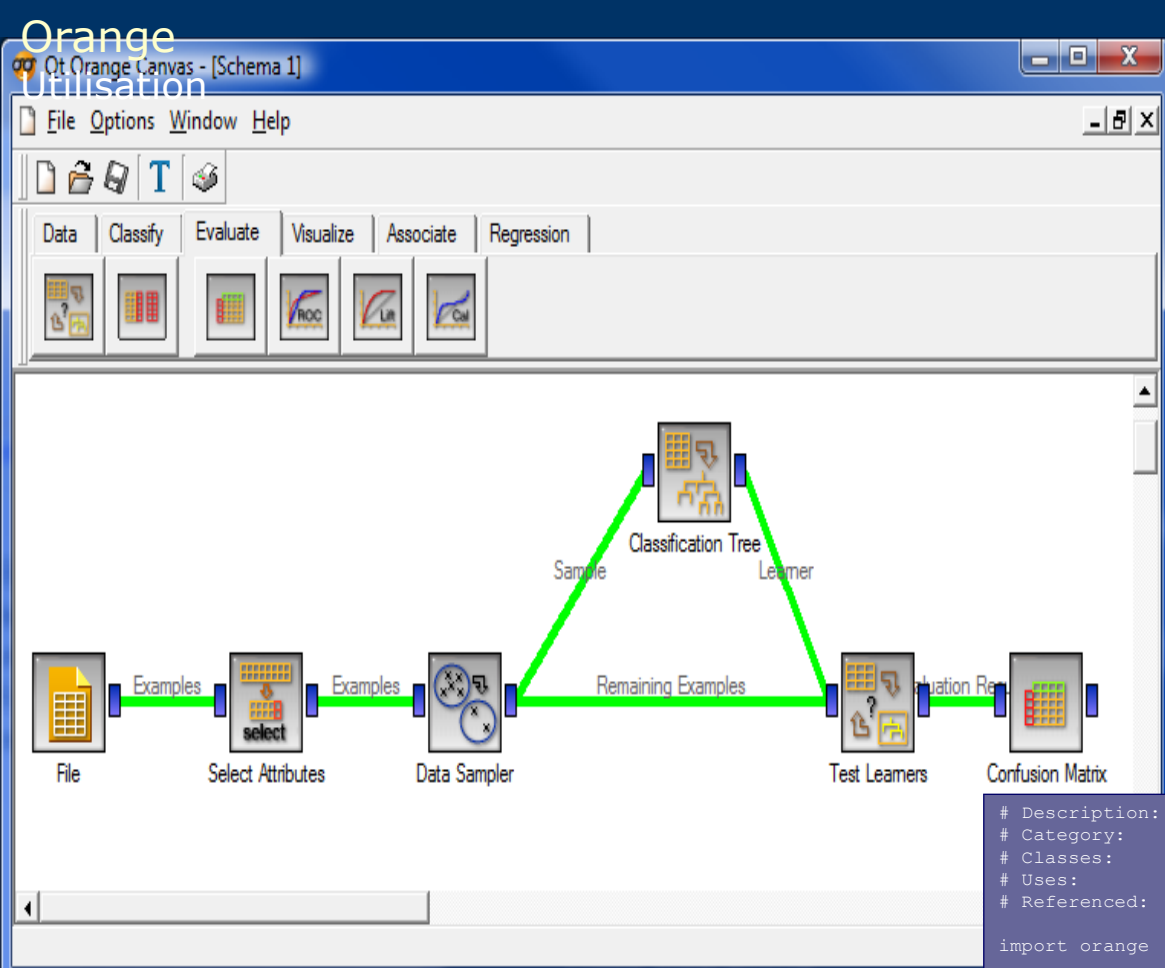
Site Web avec doc en ligne et guide

Mode diagramme

Programmation en Python

Les méthodes sont des plugins (DLL)

Très user-friendly



```

# 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 several statistical analysis results. At the top, the R logo is displayed. The main content area features a PCA plot titled "PCA 5 vars" with the command `princomp(x = data, cor = cor)`. To the left of the PCA plot is a biplot showing variables: Fertility, Examination, Education, Catholic, and Agriculture. Below the biplot is a bar chart with a y-axis from 0.0 to 1.0. To the right of the PCA plot is a scatter plot with orange and green points. Below the scatter plot are three smaller plots: "Clustering 4 groups" with a dendrogram, "Factor 1 [41%]" with a bar chart showing groups 1 (2), 2 (16), and 3 (28), and "Factor 3 [19%]" with a normal distribution curve. A "Getting Started" section is visible at the bottom, containing a list of links and instructions. The left sidebar contains navigation links for "About R", "Download, Packages", "R Project", and "Documentation".

Fondation à but non lucratif

Culture Stat.  
CORE R + Packages (plugins)  
Ex. Package Weka

Doc. des méthodes très organisée  
Des tutoriels partout

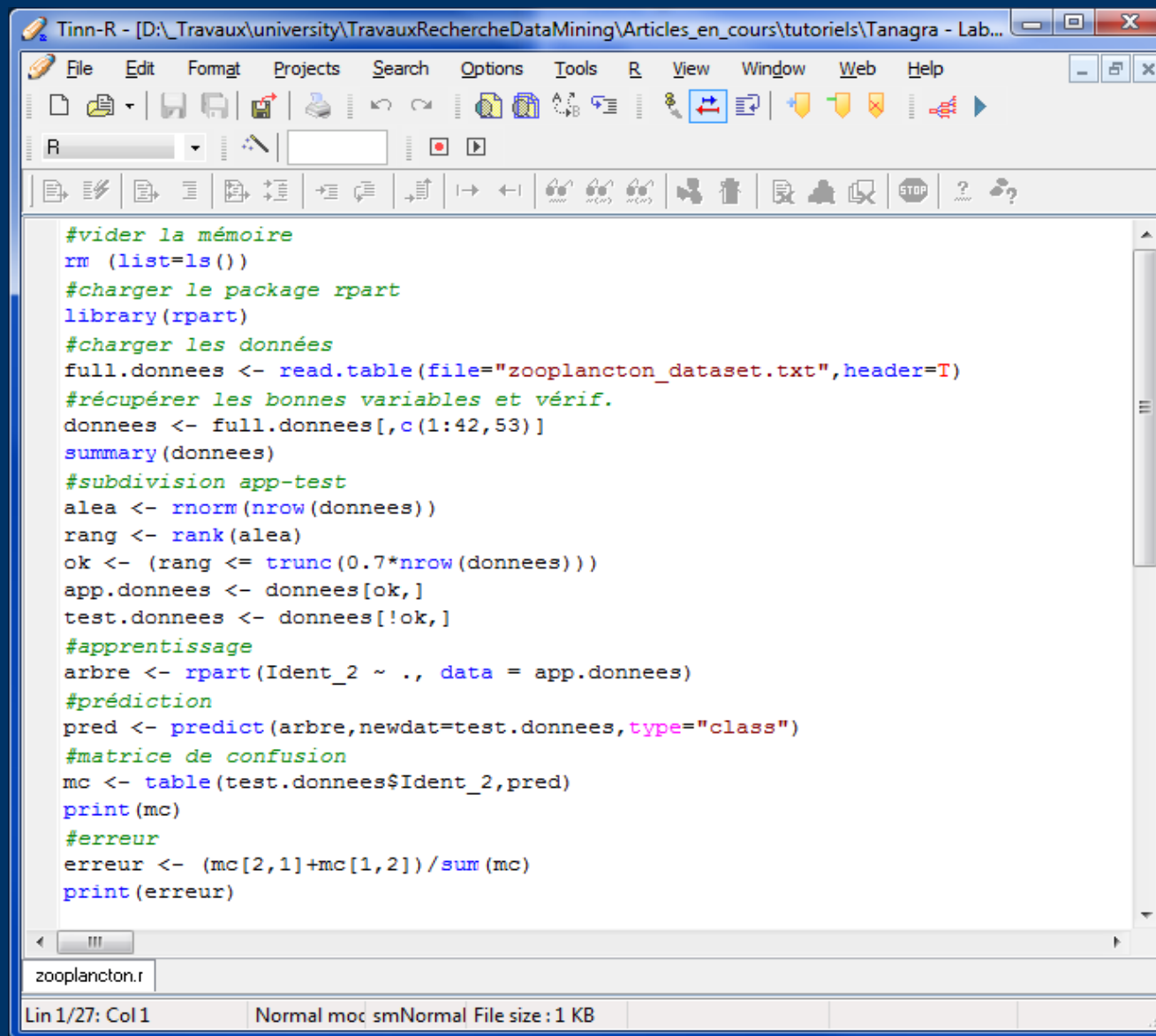
Mode programmation (langage S)

Quelques tentatives de création d'interfaces plus conviviales



# R

## Utilisation



```
#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 the Rapid-I website with the following elements:

- Navigation:** HOME, SEARCH, SITEMAP, LEGAL, CONTACT US, DEUTSCH
- Main Banner:** "Rapid-I Report the Future" with the tagline "Learn more about the Predictive Analysis and Business Intelligence solutions of Rapid-I".
- Secondary Navigation:** PRODUCTS, DOWNLOADS, SERVICES, COMMUNITY, ABOUT US
- QUICK LINKS:** Download RapidMiner Community, Order RapidMiner Enterprise, Download RapidMiner Plugins, RapidMiner Documentation, All Training Courses, RapidMiner Interactive Tour
- TESTIMONIALS:** A quote from Michael Van Kleeck, USA: "RapidMiner is an awesome package. Thank you for making such powerful functionality available in such a convenient form."
- RANDOM IMAGE:** A small screenshot of the RapidMiner software interface.
- CONTENT SECTIONS:** RAPIDMINER COMMUNITY EDITION, OPEN-SOURCE DATA MINING WITH THE JAVA SOFTWARE RAPIDMINER, RAPIDMINER: ENTERPRISE OPEN SOURCE, OPERATOR OVERVIEW.

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 workflow named 'zooplancton.xml'. The 'Operator Tree' on the left shows a process flow: 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 panel shows performance metrics for the 'Test' operator:

```
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)
```


A small line graph on the right shows the AUC curve, with 'Max: 1.1 CE' and 'Total: 1.1 CE' indicated. The system clock in the bottom right corner shows 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 processed by a separate process for processing, classification, regression, or clustering. 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 has taken over the administration of Weka's mailing list for interaction among Weka project members.

**The Weka mailing list**

Please post Weka-related questions to the mailing list. You can also check out the [online documentation](#) or the [mailing list archive](#) (Mirrors: [news.gmane.org](#)).

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

Browser tabs: Data Mining Tools Used Poll | Pentaho Commercial Open ...

Address bar: <http://weka.pentaho.org/>

Navigation: Galerie de composant... | Hotmail | Sites suggérés | Autres favoris

Customer Portal | Partner Portal | Forum | Contact Us



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 x Discover why Pentaho Dat... x

Google

← → ↻ ☆ 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™  
open source business intelligence™

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 » »

Weka KnowledgeFlow Environment

Visualization

Knowledge Flow Layout

```

    graph LR
      ArffLoader -- data Set --> ClassAssigner
      ClassAssigner -- data Set --> TrainTestSplitMaker
      TrainTestSplitMaker -- training Set / test Set --> J48
      J48 -- batch Classifier --> ClassifierPerformanceEvaluator
      J48 -- text --> TextViewer1[Text Viewer]
      ClassifierPerformanceEvaluator -- text --> TextViewer2[Text Viewer]
  
```

Status Log

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

Text Viewer

Result list

14:13:00 - J48

Text

=== Evaluation result ===

Scheme: J48  
Options: -C 0.25 -M 2Relation: zooplankton.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	

# Performances comparées

## Gros volumes (1/2)

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



# Performances comparées

## Gros volumes (2/2)

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**

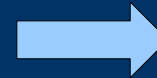
Les outils se tiennent, tout dépend des méthodes et des caractéristiques des données !!!

# Comment choisir

Quel logiciel pour quel contexte ?

## Recherche (Data Mining)

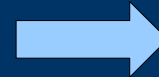
Développer de nouvelles techniques  
Les intégrer dans un environnement opérationnel  
Pour des comparaisons à grande échelle  
Les diffuser simplement et largement



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