COURS: DATAMINING

Cours:

- **Chapitre 1: Introduction Datamining**
  - ✓ Définition et processus de Datamining
  - ✓ Tâches et les techniques de Datamining
  - ✓ Exemples d'outils de Datamining
- **Chapitre 2: Les techniques prédictives**
  - ✓ Les techniques pour la tâche d'estimation et de classification
  - ✓ Validation des techniques prédictives
  - ✓ Travaux dirigés et analyse des sorties logiciels
- **Chapitre 3: Les techniques descriptives**
  - ✓ Les techniques pour la tâche de segmentation
  - ✓ Les techniques pour la tâche d'association
  - ✓ Travaux dirigés et analyse des sorties logiciel

Evaluation:

- **Contrôle(s) écrit**

INTRODUCTION DATAMINING

Pr. A. EL OUARDIGHI
jalilardighi@yahoo.fr

PLAN

Motivations

Définition du Datamining

Domaines d'applications

Description du processus de Datamining

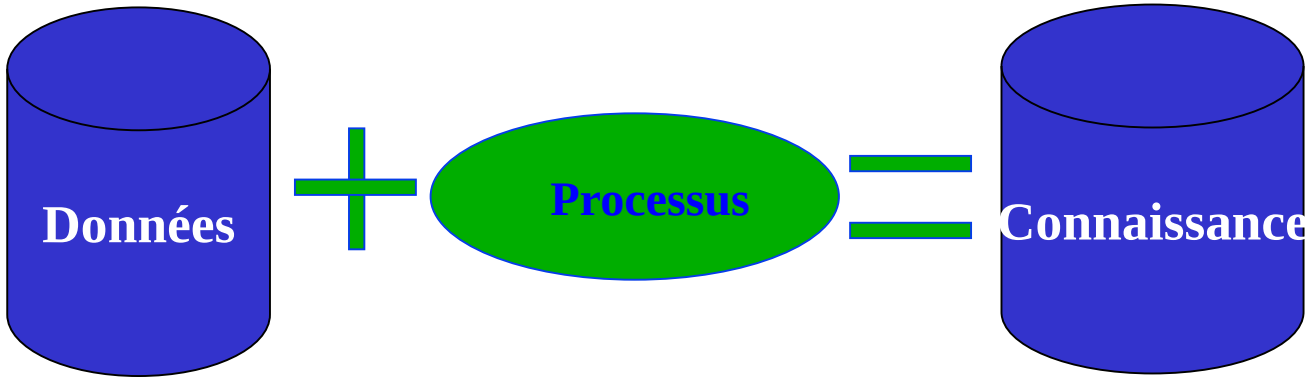
Tâches et Techniques de Datamining

Motivations

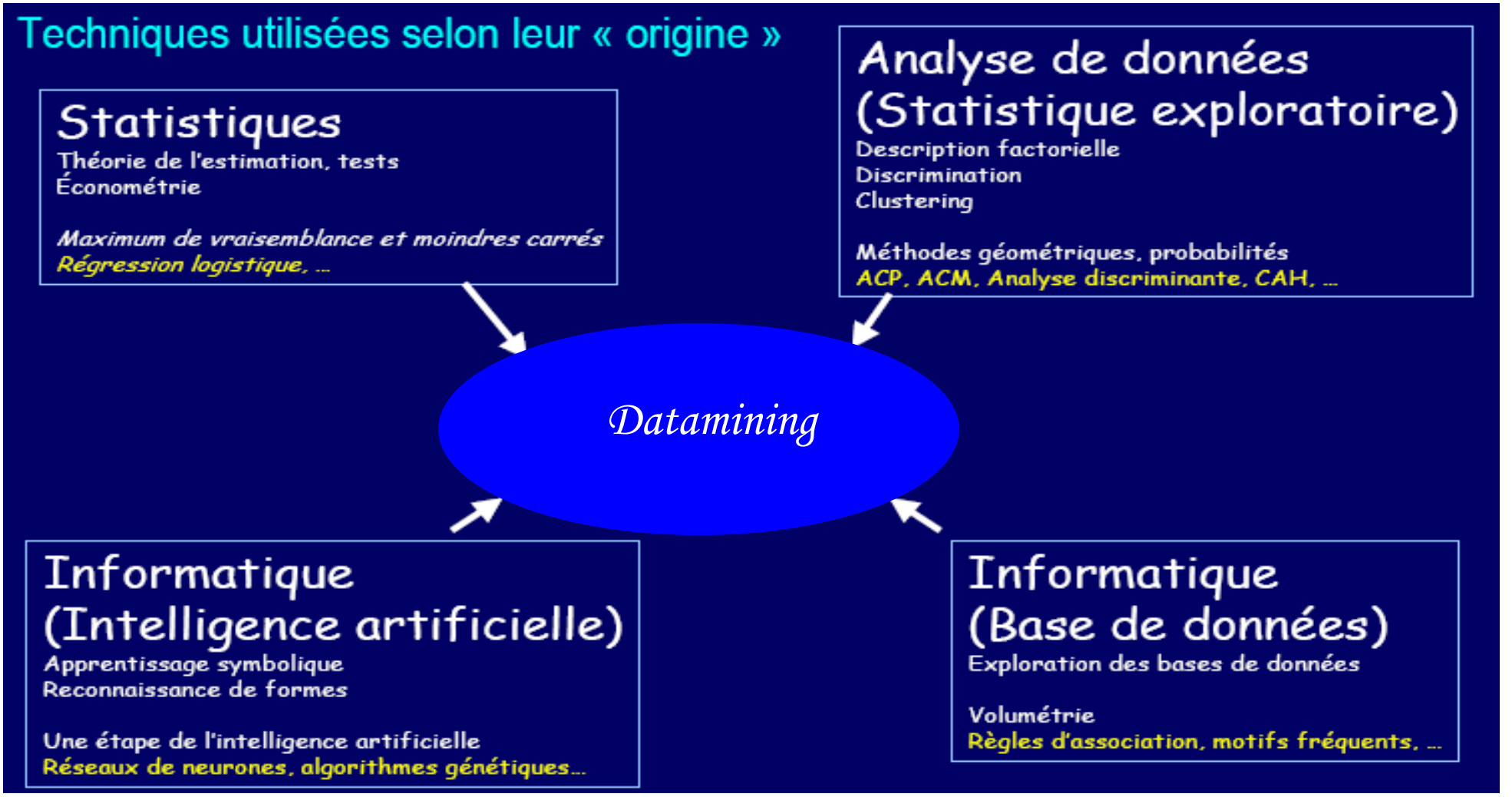
- **Explosion des données**
  - Masse importante de données
  - Données multi-dimensionnelles (milliers d'attributs)
  - Inexploitables par les méthodes d'analyse classiques
  - Besoin de traitement en temps réel de ces données
- **Croissance en puissance des machines capables:**
  - De supporter de gros volumes de données
  - D'exécuter le processus intensif d'exploration
  - De traiter des données Hétérogènes

Définition de Datamining

- **Vocabulaire:**
  - **Extraction de connaissances dans les données (ECD) ou Knowledge discovery in DataBases (KDD) Fouille de données ou Datamining**
- **Définition:**
  - ***« Le terme Datamining correspond à l'ensemble des techniques et des méthodes, qui à partir des données, permettant d'obtenir des connaissances exploitable ».***
- **Équation fondamentale:**



DM: Convergence de plusieurs disciplines

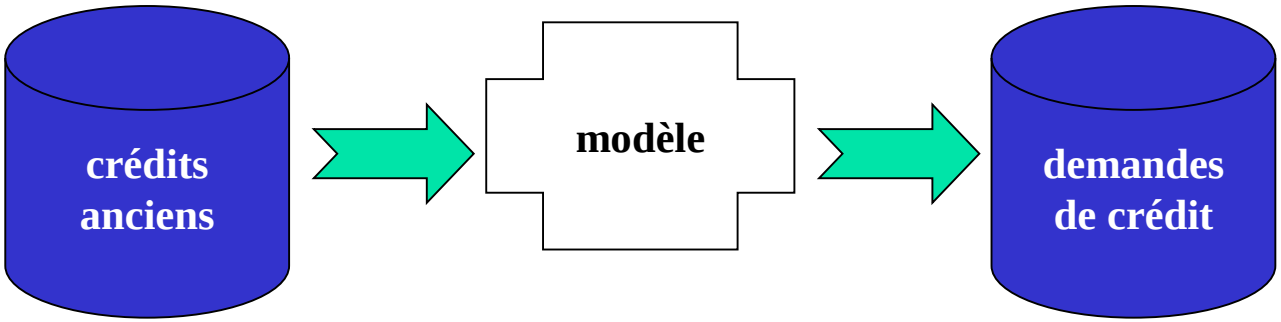


Domaines d'application

- **Marketing:** **Ciblage, Fidélisation, Relation client… .**
- **Gestion et analyse des marchés :** **Profils des consommateurs, modèle d 'achat: « panier de la ménagère »**
- **Gestion et analyse de risque:** **Assurances, Banques (identifiez les clients à risque lors de l'octroi d'un crédit, à la souscription d'un contrat d'assurance…)**
- **Détection de fraudes :** **Télécommunications, utilisation des cartes bancaires...**
- **Gestion de stocks :** **Quand commander un produit, quelle quantité demander,**

Exemple

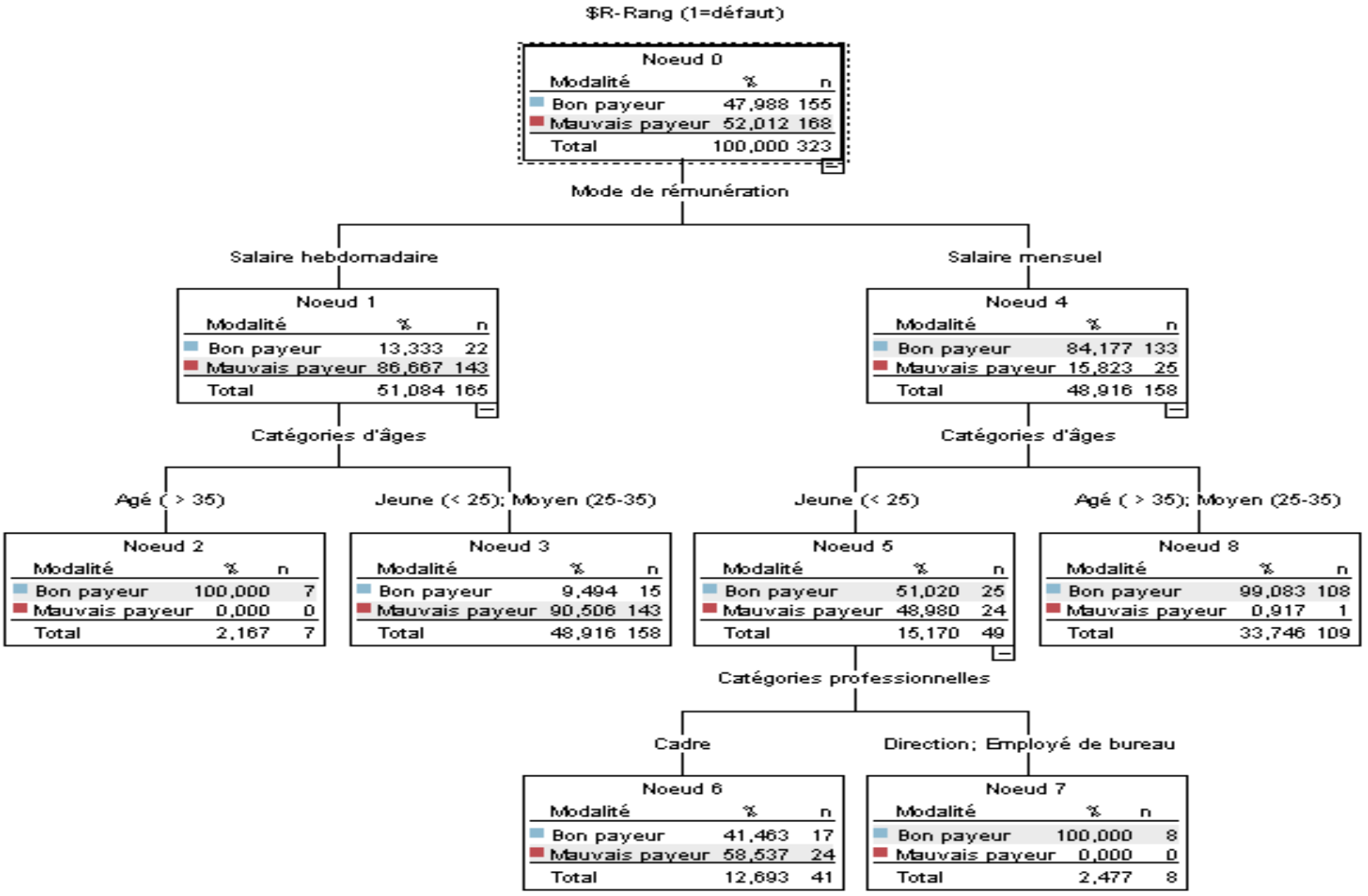
- **Entreprise : banque**
- **Activité : Les prêts**
- **Problème : accepter ou refuser une demande de crédit ?**
- **Solution actuelle : évaluation de la solvabilité du client sur base de critères définis par des gestionnaires expérimentés**
- **Solution Datamining: Analyser la solvabilité observée lors des anciens crédits**



Anciens crédit

	Rang (1=défaut)	Catégories profession...	Mode de rémunération	Catégories d'âges	Possède une carte...
1	Bon payeur	Cadre	Salaire mensuel	Moyen (25-35)	Oui
2	Mauvais payeur	Cadre	Salaire hebdomadaire	Moyen (25-35)	Non
3	Mauvais payeur	Ouvrier	Salaire hebdomadaire	Jeune (< 25)	Oui
4	Bon payeur	Cadre	Salaire mensuel	Moyen (25-35)	Non
5	Bon payeur	Employé de bureau	Salaire mensuel	Jeune (< 25)	Non
6	Bon payeur	Direction	Salaire mensuel	Jeune (< 25)	Oui
7	Bon payeur	Cadre	Salaire mensuel	Agé (> 35)	Non
8	Mauvais payeur	Cadre	Salaire mensuel	Jeune (< 25)	Non
9	Mauvais payeur	Cadre	Salaire hebdomadaire	Jeune (< 25)	Non
10	Mauvais payeur	Employé de bureau	Salaire hebdomadaire	Jeune (< 25)	Non
11	Mauvais payeur	Non qualifié	Salaire hebdomadaire	Jeune (< 25)	Non
12	Mauvais payeur	Ouvrier	Salaire hebdomadaire	Jeune (< 25)	Oui
13	Mauvais payeur	Cadre	Salaire mensuel	Jeune (< 25)	Oui
14	Mauvais payeur	Cadre	Salaire hebdomadaire	Jeune (< 25)	Oui
15	Mauvais payeur	Employé de bureau	Salaire hebdomadaire	Jeune (< 25)	Oui
16	Mauvais payeur	Cadre	Salaire hebdomadaire	Moyen (25-35)	Oui
17	Mauvais payeur	Ouvrier	Salaire hebdomadaire	Jeune (< 25)	Oui
18	Mauvais payeur	Cadre	Salaire mensuel	Jeune (< 25)	Non
19	Mauvais payeur	Employé de bureau	Salaire hebdomadaire	Jeune (< 25)	Oui
20	Mauvais payeur	Non qualifié	Salaire hebdomadaire	Jeune (< 25)	Non
21	Mauvais payeur	Ouvrier	Salaire hebdomadaire	Jeune (< 25)	Oui
22	Mauvais payeur	Employé de bureau	Salaire hebdomadaire	Jeune (< 25)	Oui

Modélisation (Application d'une technique de Datamining)



Modélisation (Application d'une technique de Datamining)

Rang (1=défaut)(a)		B	Erreur std.	Wald	degrés de liberté	Signif.	Exp(B)	Intervalle de confiance 95% pour Exp(B)	
								Borne inférieure	Borne supérieure
Mauvais payeur	Constante	-1.927	1.362	2.002	1	.157			
	[Catégories d'âges=Agé (> 35)]	-2.373	1.174	4.088	1	.043	9.32E-002	9.33E-003	.930
	[Catégories d'âges=Jeune (< 25)]	2.475	.534	21.509	1	.000	11.880	4.174	33.809
	[Catégories d'âges=Moyen (25-35)]	0(b)	.	.	0	.	.	.	.
	[Catégories professionnelles=Cadre]	-.518	1.156	.200	1	.654	.596	6.18E-002	5.747
	[Catégories professionnelles=Direction]	-4.104	1.643	6.243	1	.012	1.65E-002	6.60E-004	.413
	[Catégories professionnelles=Employé de bureau]	-1.130	1.141	.980	1	.322	.323	3.45E-002	3.025
	[Catégories professionnelles=Non qualifié]	-1.118	1.195	.875	1	.349	.327	3.14E-002	3.400
	[Catégories professionnelles=Ouvrier]	0(b)	.	.	0	.	.	.	.
	[Mode de rémunération=Salaire hebdomadaire]	3.310	.525	39.756	1	.000	27.383	9.787	76.614
	[Mode de rémunération=Salaire mensuel]	0(b)	.	.	0	.	.	.	.
	[Possède une carte de crédit=Non]	-.322	.396	.658	1	.417	.725	.333	1.577
	[Possède une carte de crédit=Oui]	0(b)	.	.	0	.	.	.	.

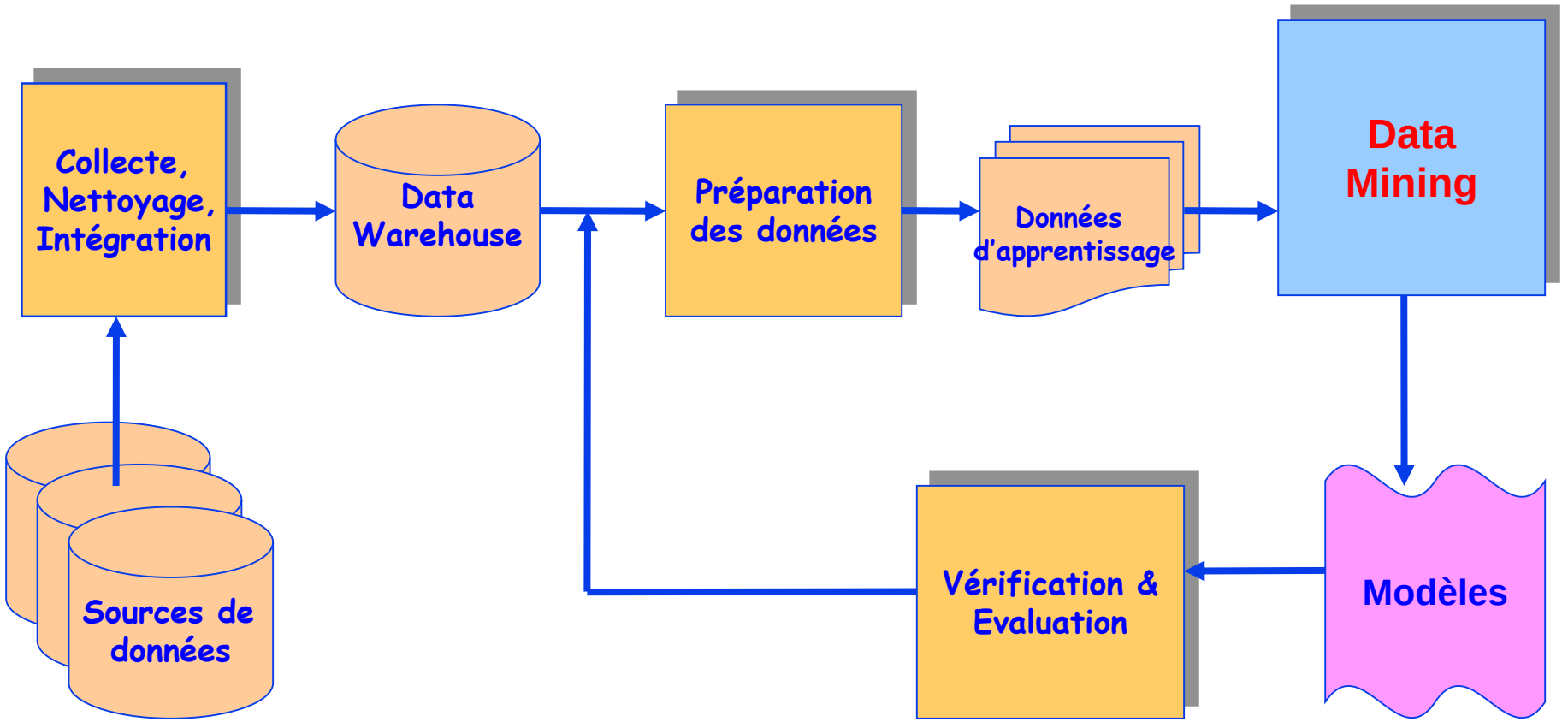
a. La modalité de référence est : Bon payeur.

b. Ce paramètre est remis à zéro parce qu'il est superflu.

Application du modèle sur de nouveaux clients

Mode de rémunération	Catégorie d'âge	Profession	Possède une carte	Rang
Hebdomadaire	Jeune	Cadre	Oui	?
Mensuelle	Agé	Direction	Oui	
Mensuelle	Moyen	Cadre	Oui	

Le processus de Datamining



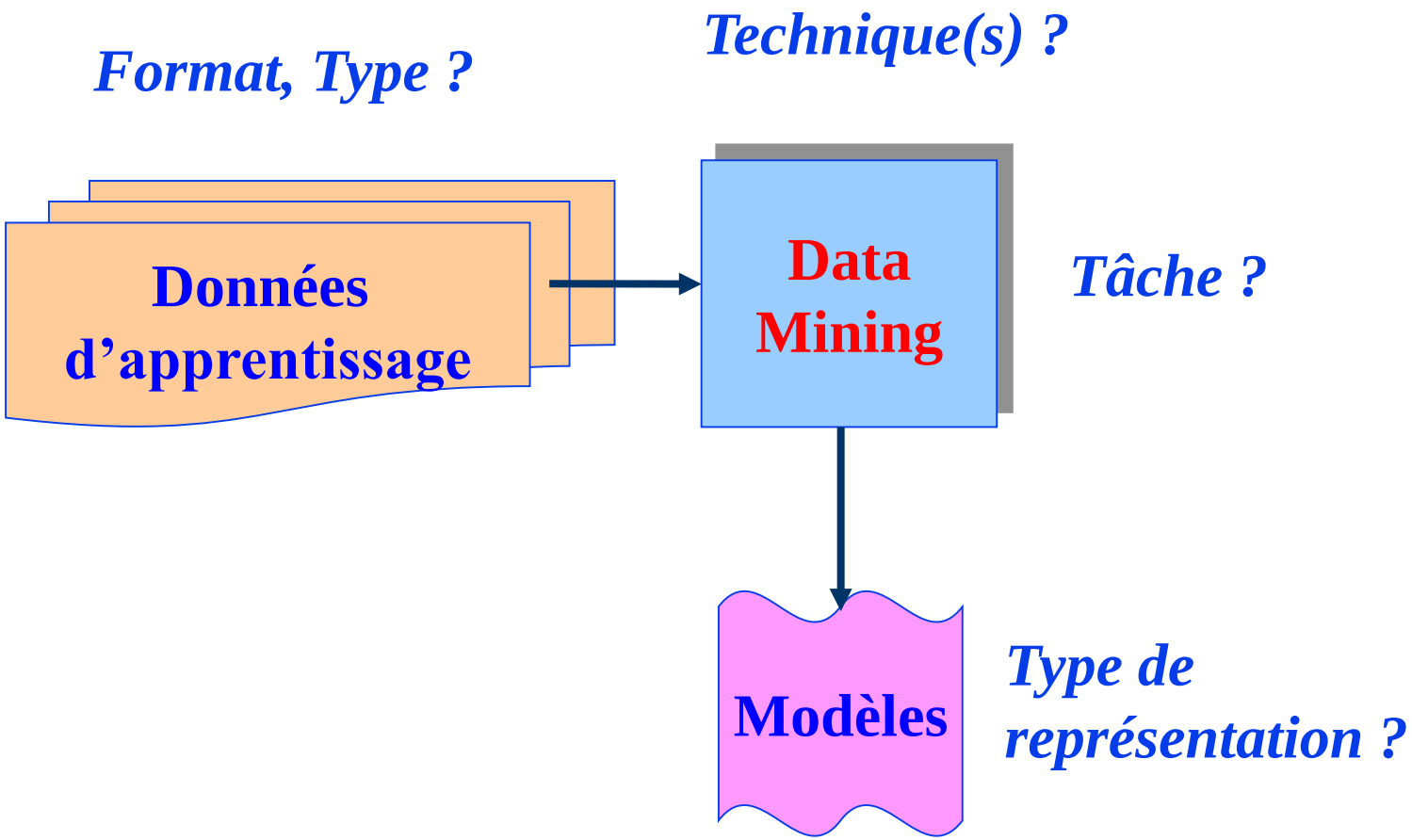
Processus de Data Mining

- **Compréhension du problème**
  - **Détermination des objectives et l'utilité de la connaissance**
  - **Production d'un plan de projet**
- **Compréhension des données**
  - **Accès aux données (Datawarehouse, Datamart, Base de données opérationnelle, Fichiers…)**
  - **Description et l'exploration des données**
  - **Vérification de la qualité des données**
- **Action sur les données**
  - **Nettoyage des données, Données manquantes,**
  - **Sélection de données, réduire la dimension du problème, etc.**

Processus de Data Mining

- **Recherche du modèle**
  - **Sélection de techniques de modélisation**
  - **Création de modèle**
- **Visualiser, évaluer et interpréter les modèles découverts**
  - **Explorer le modèle**
  - **Analyser la connaissance et vérifier sa validité**
  - **Réitérer le processus si nécessaire**
- **Gérer la connaissance découverte**
  - **La mettre à la disposition des décideurs**
  - **L'échanger avec d'autres applications**

Paramètres d'un processus DM



Les différents formats de données

- Les données caractérisent les différents aspects des exemples.
- Une donnée a un type et des valeurs possibles pour ce type.
- Les différents formats de données:
  - Données Quantitatives:
    - Continues (ou d'échelle), exemple : salaire, âge …
    - Discrètes, exemple : nombre d'enfants
  - Données Qualitatives:
    - Boolien (Binaire), exemple: sexe
    - Nominale, exemple: couleur, profession…
    - Ordinale, exemple (chaud, tiède, froid)

Deux grandes familles de méthodes

- **Méthodes prédictives ou supervisées**
- **Méthodes descriptives ou non supervisées**

Les méthodes prédictives

- Visent à extrapoler de nouvelles informations à partir des informations présentes
- Apprentissage supervisé: Il y a une variable « cible » à prédire.
- Deux grandes familles: Classification et prédiction
- Classification: consiste à placer chaque individu de la population dans une classe, parmi plusieurs classes prédéfinies, en fonction des caractéristiques de l'individu indiquées comme variables explicatives.
  - On parle aussi de discrimination ou scoring
- La variable à expliquer est qualitative
- Exemples de techniques de classification:
  - Les réseaux de neurones: Perceptron Multi-couches (PMC)
  - Les arbres de décision: C&RT, C5.0 etc …
  - Les SVM (Support Vector Machine);
  - La régression logistique: Binaire, Multinomiale

Les méthodes prédictives

- **La prédiction: consiste à estimer la valeur d'une variable continue (dite à expliquer) en fonction de la valeur d'un certain nombre d'autres variables (dites explicatives)**
  - **On parle aussi de régression**
- **La variable à expliquer est continue**
- **Exemple de méthodes prédictives:**
  - **Les réseaux de neurones: Perceptron Multi-couches (PMC)**
  - **La régression linéaire**

Les méthodes descriptives

- Visent à mettre en évidence des informations présentes mais cachées par le volume des données
- Apprentissage non supervisé: Il n'y a pas de variable « cible » à prédire
- Trois familles de méthodes: Segmentation, Association, Analyse factorielle
  - La segmentation ou clustering: Trouver dans l'espace de travail des groupes homogènes d'individus ou de variables
    - Techniques de segmentation: K-means, Twostep, Kohonen
  - Association: Trouver des règles d'association entre un ensemble d'éléments avec un bon niveau de probabilité
    - Techniques d'association: Apriori, GRI
  - Analyse factorielle: Projection du nuage de points sur un espace de dimension inférieure pour obtenir une visualisation de l'ensemble des liaisons entre variables tout en minimisant la perte d'information
    - Techniques factorielles: ACP, AFC, AFCM

Les outils logiciels de Datamining

Les outils commerciaux

- **Clementine de SPSS**
  - Le leader de l'analyse prédictive, Clementine propose différentes méthodes de modélisation issues des domaines de l'apprentissage automatique, de l'intelligence artificielle et des statistiques
- **Enterprise Miner de SAS**
  - Enterprise Miner est une solution logicielle intégrée, qui, au travers d'une interface totalement graphique (icônes, fleches…), donne à l'utilisateur l'accès aux différentes étapes de la méthodologie de SAS Institute pour le data mining

Les outils logiciels de Datamining

Les outils commerciaux

Decisia de SPAD

- SPAD est le logiciel français pionnier dans les analyses exploratoires et le data mining. Connu et reconnu pour sa convivialité et son efficacité, il possède les principales techniques statistiques liées au data mining.

Cognos

- La solution couvre toute la chaîne du décisionnel : de l'extraction de données, au reporting, à l'analyse, au scorecarding jusqu'à la diffusion de l'information.

Les outils logiciels de Datamining

- **Les outils libres (opens sources)**
  - **WEKA (Waikato Environment for Knowledge Analysis)**
    - **http://www.cs.waikato.ac.nz/ml/weka/**
  - **TANAGRA**
    - **http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html**
  - **ORANGE**
    - **http://magix.fri.uni-lj.si/orange/**
  - **R**
    - **http://www.r-project.org/**
  - **RapidMiner**
    - **http://rapid-i.com/**

Bibliographie

- « Le datamining », R. Lefebure et G. Venturie, ed. Eyrolles, 2001
- « Datamining et Scoring », S. Tufféry, ed. Dunod, 2002.
- « Datamining: Pratical machine learning tools and techniques », I. Witten and E. Frank, Morgan Kaufman Pub. 2005
- « The elements of statistical learning – Datamining, Inference and Prediction » T. Hastie, R. Tibshirani, J. Friedman, Springer 2001