**Université Cadi Ayyad Marrakech**
**École Supérieure de Technologie Essaouira**

**DUT IDSD Semestre 4**

**Module**

# *DATA MINING*

**Chapitre I**

# Principes de Data Mining

**Pr: A. Guezzaz**

Département de Génie informatique et Mathématiques (GIM)

**Année scolaire : 2020-2021**

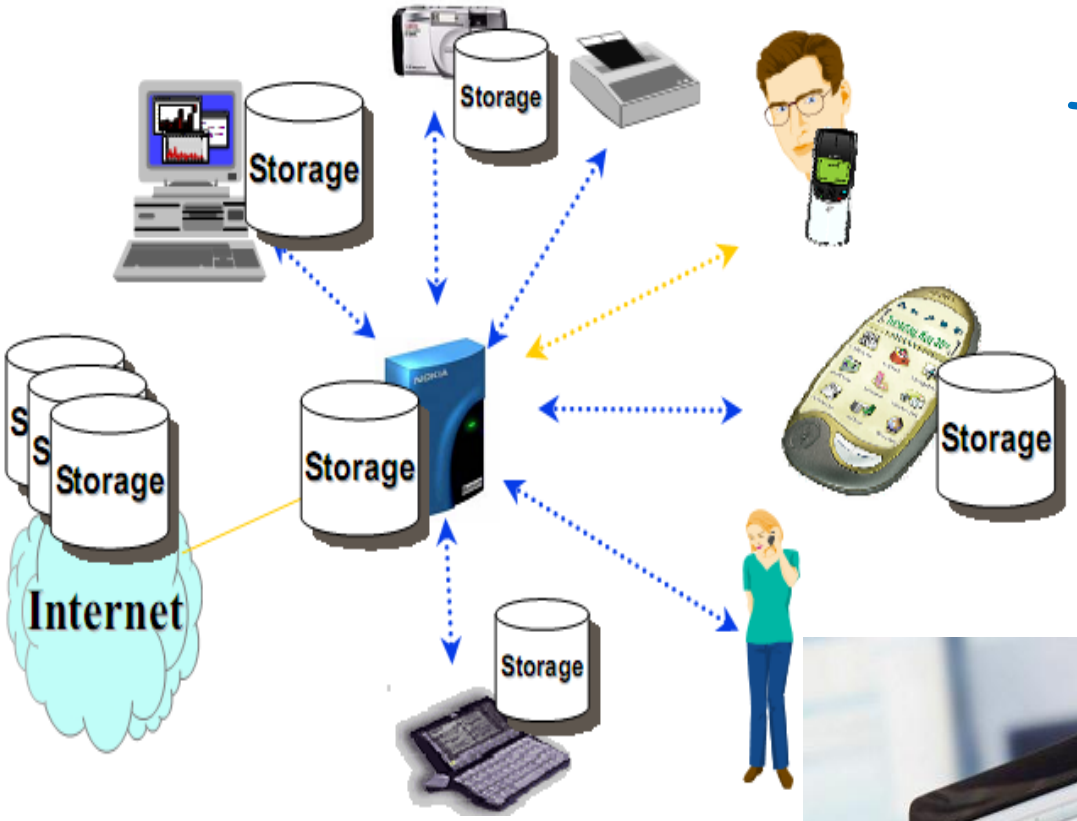
# Objectifs du Module

1. Maîtriser des concepts fondamentaux de la **science des données.**
2. **Découvrir des connaissances** à partir des **données massives** par des **logiciels du Data Mining** comme **WEKA, Rapid Miner, ...**
3. **Interpréter les résultats** effectués et fournis par ces logiciels.
4. Décrire les **méthodes d'apprentissage** supervisé et non supervisé.
5. Choisir et utiliser des **méthodes adéquates** pour résoudre un **problème donné.**
6. Lire et comprendre des **articles scientifiques** de Data Mining.

1. **Introduction au Data Mining**
    1. Processus de découverte des connaissances (KDD)
    2. Démarche méthodologique
    3. Tâches du Data Mining
2. **Prétraitement et visualisation des données**
    1. Variables et attributs
    2. Préparation et Nettoyage des données
    3. Bruit et Données manquantes
    4. Sélection et Réduction des attributs
    5. Visualisation des données
3. **Description de l'environnement WEKA**

# Introduction au Data Mining

# 1. Processus de découverte de connaissances (KDD)




Transfert et stockage des données

Visualisation et interprétation des données

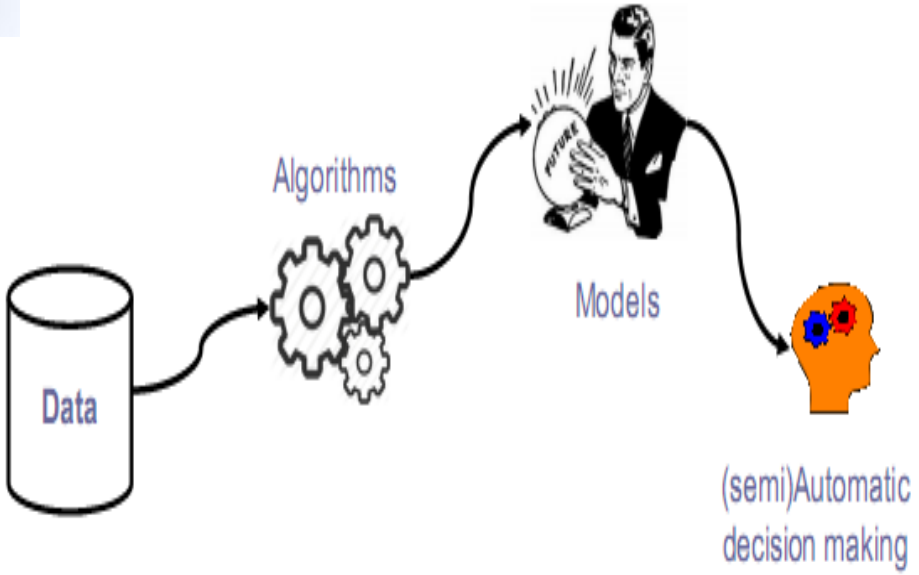
# 1. Processus de découverte de connaissances (KDD)

- **Explosion des données**
  - **Masse importante** de données (*millions de milliards **d'instances***).
  - **VLDB (**Very **L**arge **D**atabases) **:** BD très larges.
  - Données **multi-dimensionnelles** (*milliers d'attributs*)
  - Collecte de masses importantes de données (*Gbytes /heure*).
    - *Données satellitaires, génomiques (micro-arrays, …), simulations scientifiques, etc.*
  - **Problème:** Inexploitables par les méthodes d'analyse classiques.
  - **Besoin: de traitement en temps réel de ces données**.

# 1. Processus de découverte de connaissances (KDD)


BIG DATA

DATA MINING



# 1. Processus de découverte de connaissances (KDD)

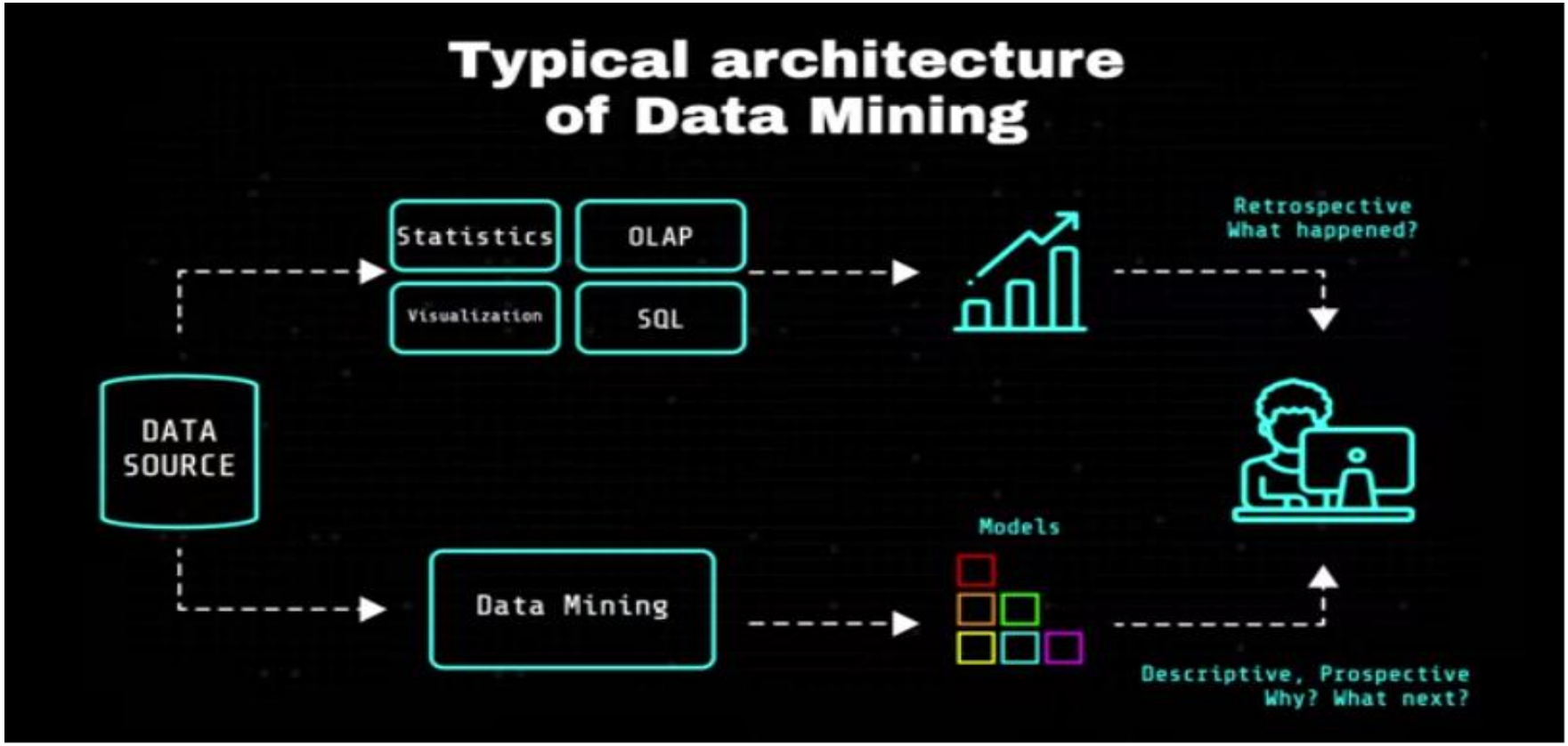
- **Data Mining** (Fouille de données) est un processus **itératif et interactif** de **découverte dans les bases de données** larges.
  - **Itératif** : nécessite plusieurs phases.
  - **Interactif** : l'utilisateur est dans la boucle du processus.
- **Data Mining** permet de découvrir des **modèles de données:**
  - **Valides** : valables dans le futur
  - **Nouveaux** : non prévisibles.
  - **Utiles** : permettent à l'utilisateur de prendre des décisions.
  - **Compréhensibles** : présentation simple.

## ❑ Statistiques vs Data Mining

- **Statistiques: Confirmatoire**, ***User-Driven Modeling.***
  - ➢ **Distribution** d'une seule variable :
    - *moyenne, médiane, variance, écart-type, …*
  - ➢ Explorer **les relation entre variables**:
    - *coefficient de corrélation, …*
  - ➢ Découverte de la cause des relations entre de nombreuses variables est assez complexe.
    - *Réseaux bayésiens (probabilités conditionnelles)*
- **Data Mining: Exploratoire**, ***Data-Driven Modeling***

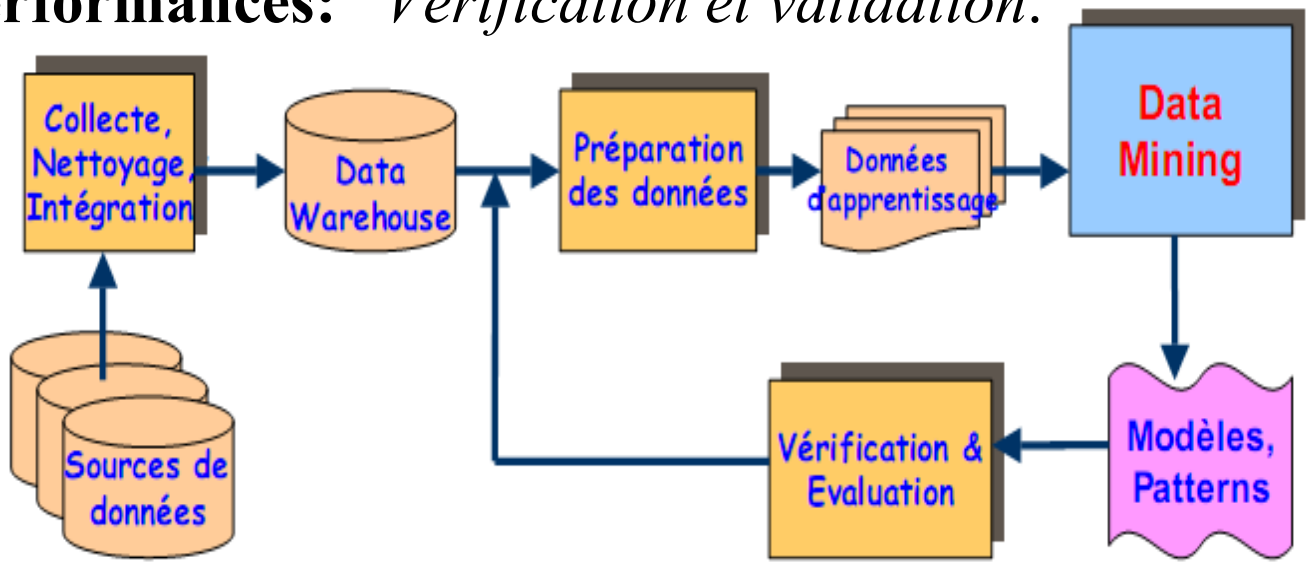
# 1. Processus de découverte de connaissances (KDD)

## ❑ Statistiques vs Data Mining



# 1. Processus de découverte de connaissances (KDD)

- **Sources des données:** *Hétérogénéité, grande masse,*
- **Collecte et prétraitement:** *Extraction, transformation et intégration (ETL).*
- **Data Warehouse:** *Stockage des données.*
- **Préparation des données.**
- **Apprentissage et Test :** *Modèles de données.*
- **Évaluation des performances:** *Vérification et validation.*



# 2. Démarche méthodologique

## 1. Comprendre l'application

- Connaissances a priori, objectifs, etc.

## 2. Sélectionner un échantillon de données

- Choisir une méthode d'échantillonnage

## 3. Nettoyage et transformation des données

- **Supprimer le bruit (Noisy):**
  - *Données superflues, marginales,*
  - *Données manquantes, etc*.
- Effectuer une **sélection d'attributs**, **réduire la dimension** du problème, etc.

# 2. Démarche méthodologique

## 4. Appliquer les techniques de fouille de données

- Choisir le bon algorithme.

## 5. Visualiser, évaluer et interpréter les modèles découverts

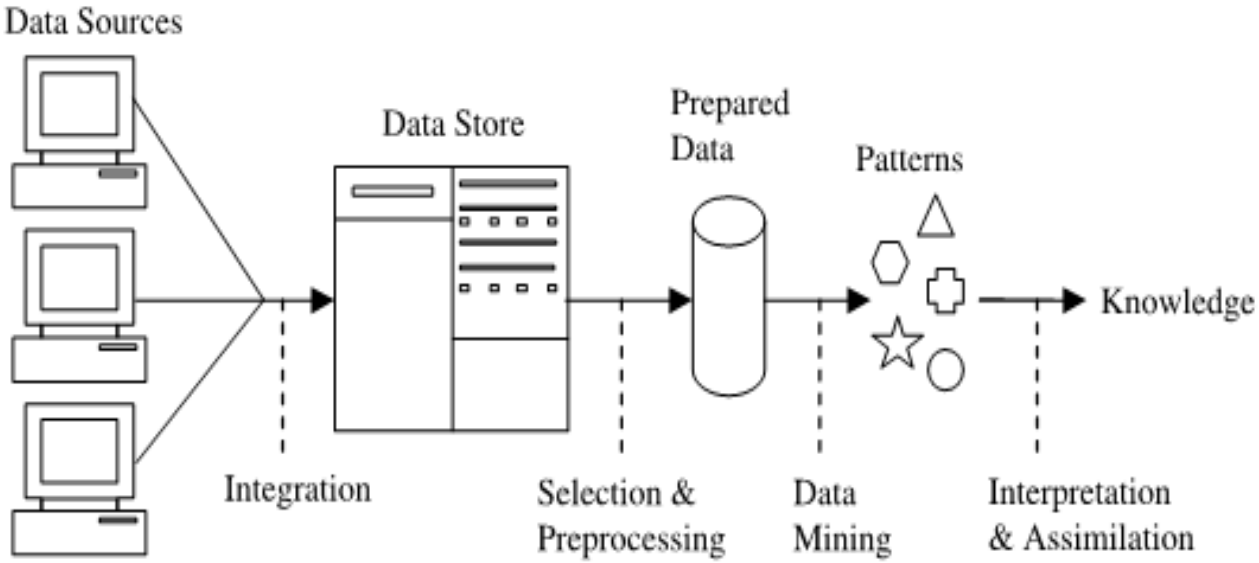
- Analyser la connaissance (intérêt).
- Vérifier sa validité (sur le reste de la base de données).
- Réitérer le processus si nécessaire.

## 6. Gérer la connaissance découverte

- La mettre à la disposition des décideurs.
- L'échanger avec d'autres applications (système expert, …)
- etc

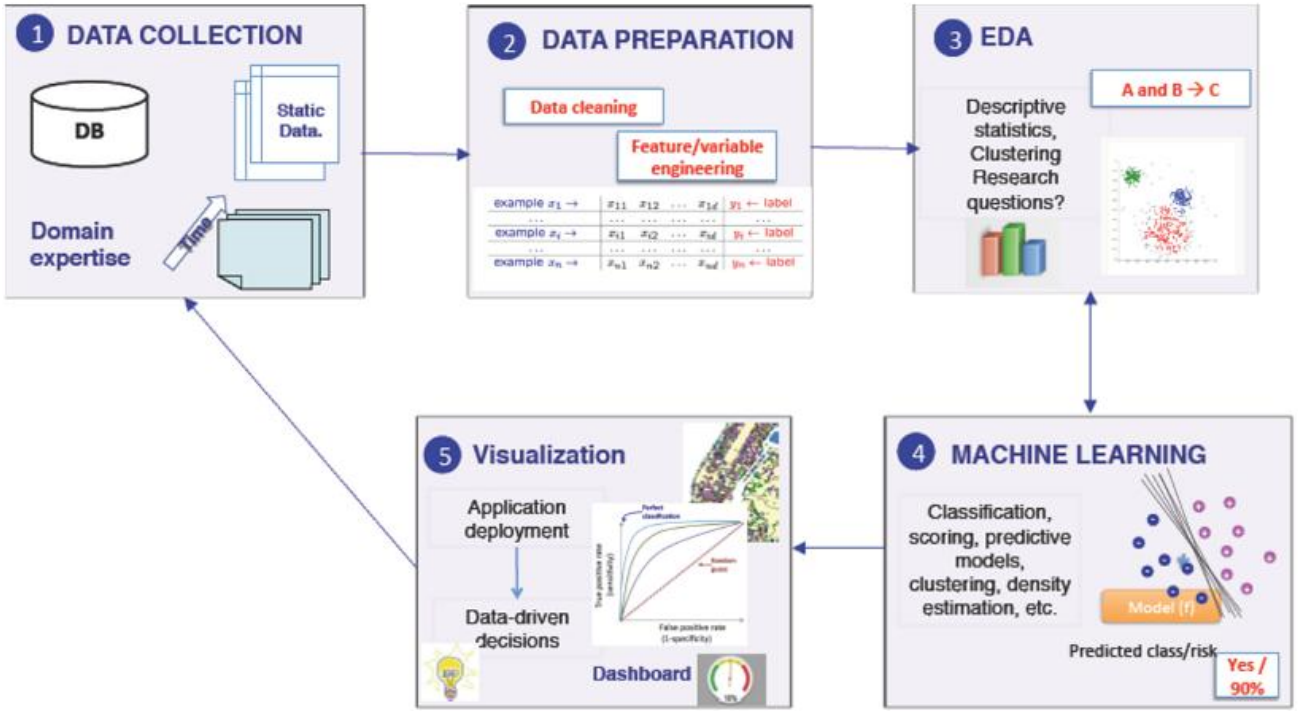
# 2. Démarche méthodologique

- La **découverte de connaissances** a été définie comme «**l'extraction l'information implicite et inconnue et potentiellement utile à partir de données**».
- C'est un **processus** dont l'exploration de données ne constitue qu'une partie, bien que centrale.



# 2. Démarche méthodologique

- Les données arrivent de **nombreuses sources**, **intégrées** et **stockées** dans un **entrepôt de données** (**Data warehouse**).
- Une partie de DW **prétraitée** dans un **format standard.**
- Les **données préparées** sont transmises à **un algorithme de DM**.
- Fournir des **règles** (**modèles**) pour **la prise des décisions.**



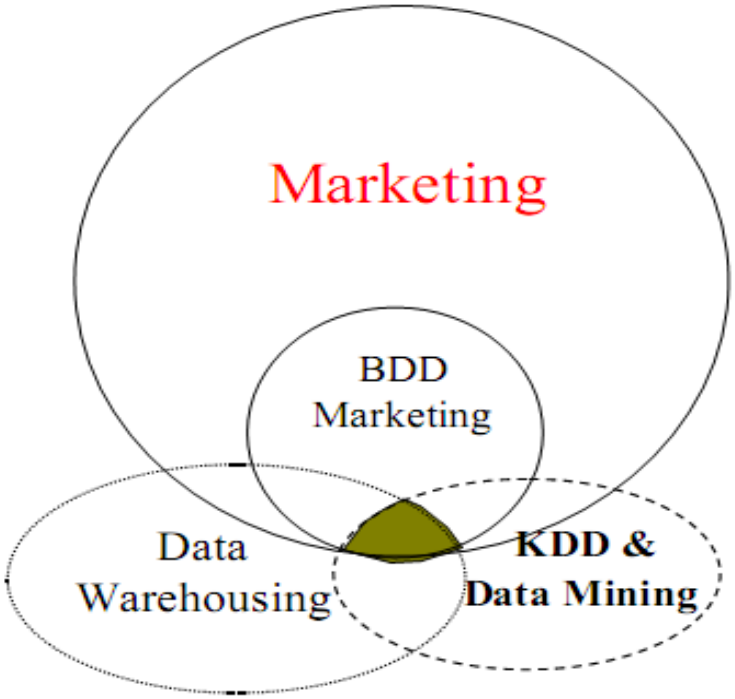
# 2. Démarche méthodologique

## ❑ Domaines d'application

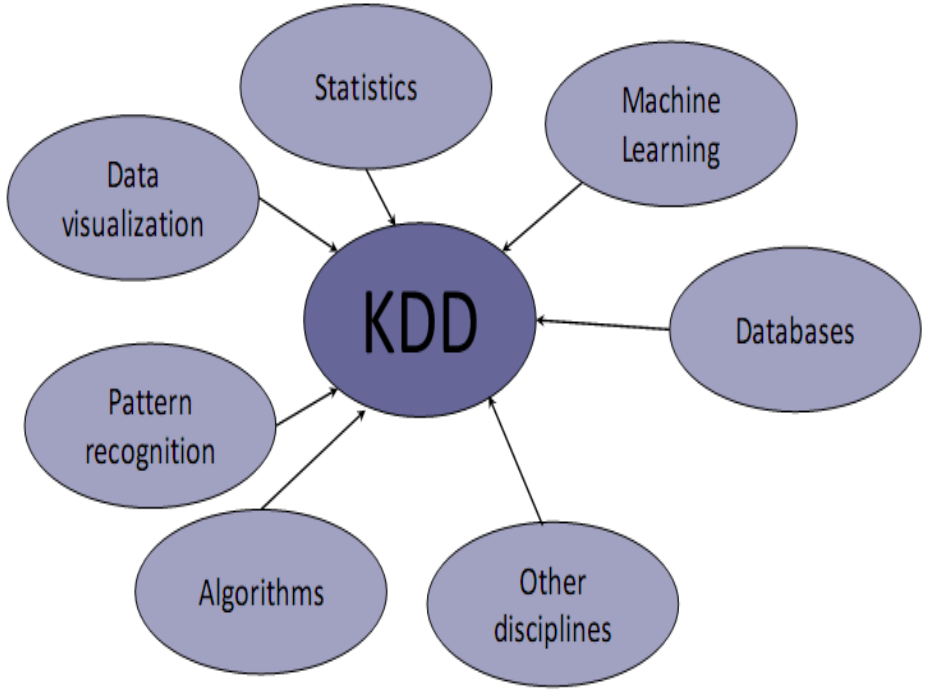
- ✓ Analyse des images satellite.
- ✓ Gestion et analyse de risque:
  - *Assurances,*
  - *Banques (crédit accordé ou non)*
- ✓ **Détection de fraude par carte de crédit**
- ✓ Prévisions financières
- ✓ **Diagnostic médical**
- ✓ Prédire l'audience de la télévision
- ✓ Prévisions météo, …
- ✓ **Web mining, text mining, etc.**

**Thèmes de Recherche scientifique**

## ❑ Domaines d'application



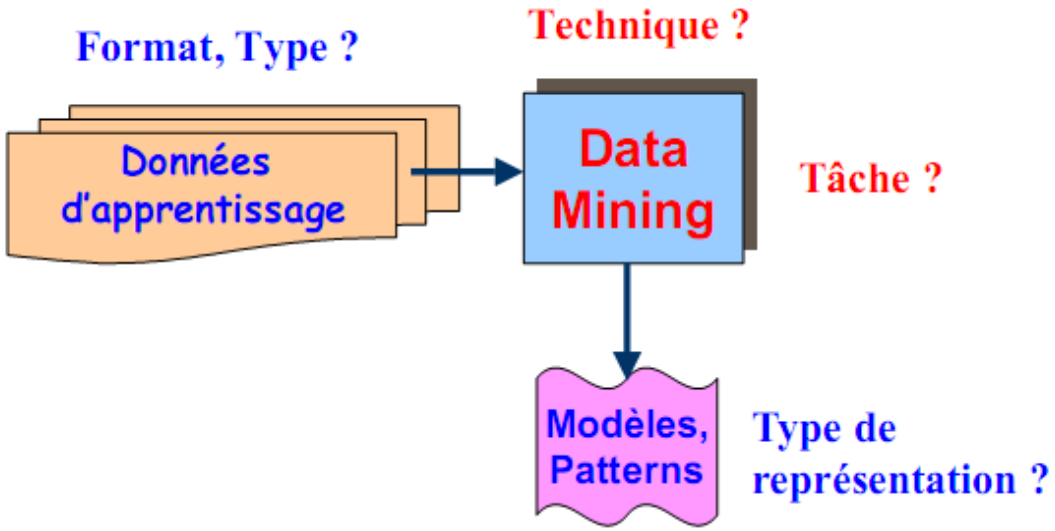

Champs liés au Data Mining




Conception des produits et Marketing ciblé

## ❑ Domaines d'application

- **Les logs des accès Web sont analysés pour …**
  - ➢ *L'analyse de tous les types d'informations sur les logs*
  - ➢ *Découvrir les préférences des utilisateurs*
  - ➢ *Améliorer l'organisation du site Web*



# 3. Tâches du Data Mining

## ❑ Types d'apprentissage

- **Data Set:** Ensemble de données dans un endroit de stockage.
- **Instances**: Exemples (**Samples**) d'un **Data Set**.
- **Attributs**: les valeurs d'un certain nombre de variables.
- Il existe **deux types de données** traitées de manière différente:
  - **Données étiquetées** (**Labelled Data**): un attribut **spécialement** désigné (**cible**). Le but est de **prédire** la valeur de cet attribut pour les **instances** pas encore vues.
  - **Données non étiquetées** (**UnLabelled Data**): qui n'ont pas d'attribut spécialement **désigné**.

# 3. Tâches du Data Mining

## ❑ Types d'apprentissage

- **Apprentissage supervisé (Supervised Learning): la découverte d'un modèle** à l'aide de **données étiquetées** .
  - **Classification:** si **l'attribut** désigné est **catégorique** (prendre l'une d'un certain nombre de **valeurs distinctes).**
    - **Exemple**: *«très bon», «bon» ou «médiocre»*.
  - **Régression:** si **l'attribut** désigné est **numérique**.
    - **Exemple**. le *prix de vente attendu d'une maison*.

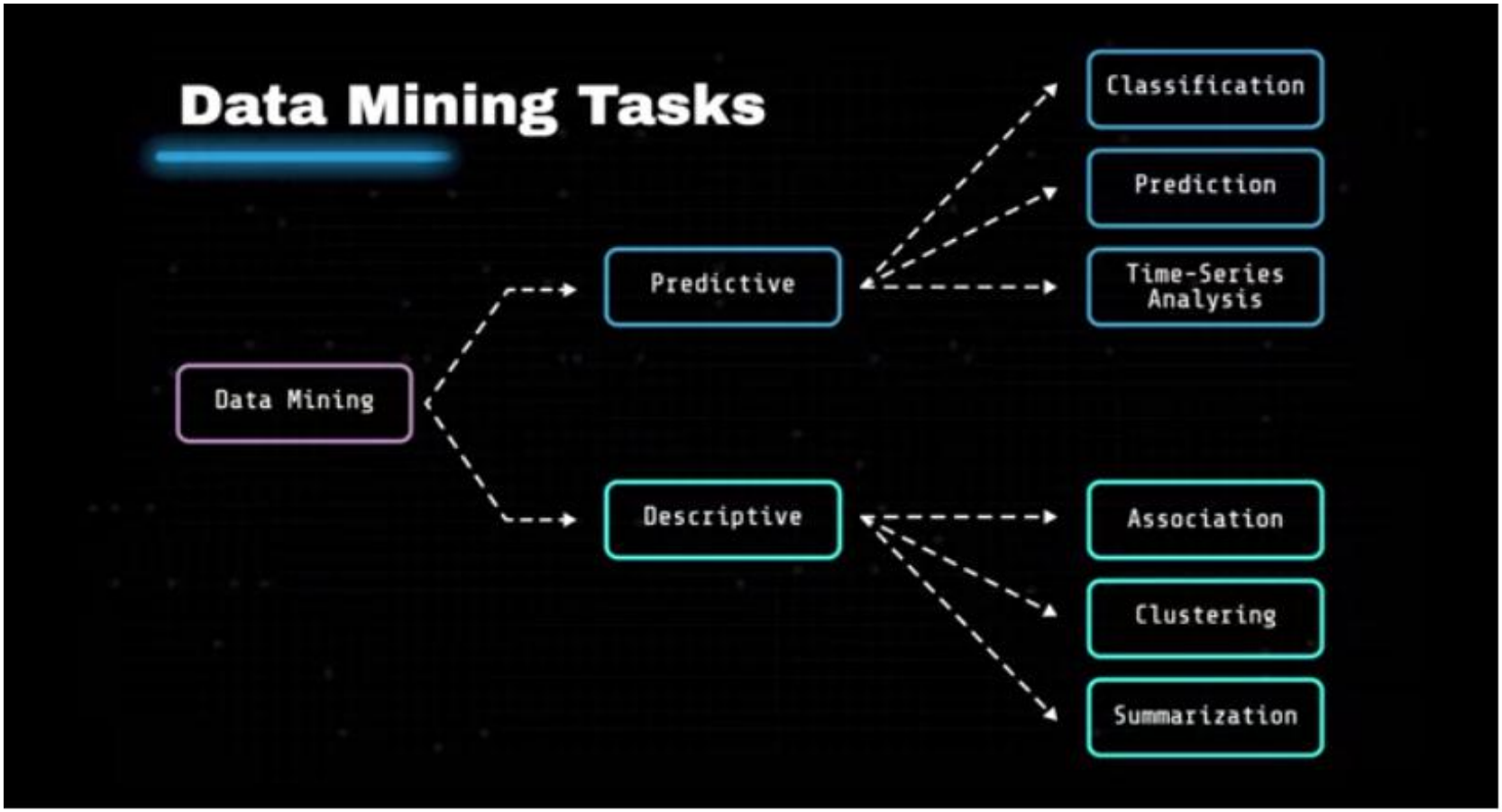
# 3. Tâches du Data Mining

## ❑ Types d'apprentissage

- **Apprentissage non supervisé (Unsupervised Learning): la découverte d'un modèle** à l'aide de **données non étiquetées .**
- Le but est simplement d'extraire **le plus d'informations possible** à partir des données disponibles.
  - **Clustering:** Segmentation, regroupement ou partitionnement.
  - **Règles d'associations** (Associtaion Rules).
  - **Recherche de séquences**
  - **Détection de déviation.**

# 3. Tâches du Data Mining

## ❑ Types d'apprentissage



# 3. Tâches du Data Mining

## ❑ Types d'apprentissage

- ✓ Classification.
- ✓ Régression ou Prédiction numérique (Numerical prediction)
- ✓ Segmentation (Clustering).
- ✓ Règles d'associations (Associtaion Rules).
- ✓ Recherche de séquences (Sequences research).
- ✓ Détection de déviation.

# 3. Tâches du Data Mining

## ❑ Apprentissage supervisé: Classification

- Permet de prédire si **une instance** est un membre d'un groupe ou d'une **classe prédéfinie**.
- **Classes:**
  - **Groupes d'instances** avec des **profils** particuliers.
  - **Apprentissage supervisé:** classes connues à l'avance.
  - **Applications**: *marketing direct* (profils des consommateurs), *médecine* (malades /non malades), etc.

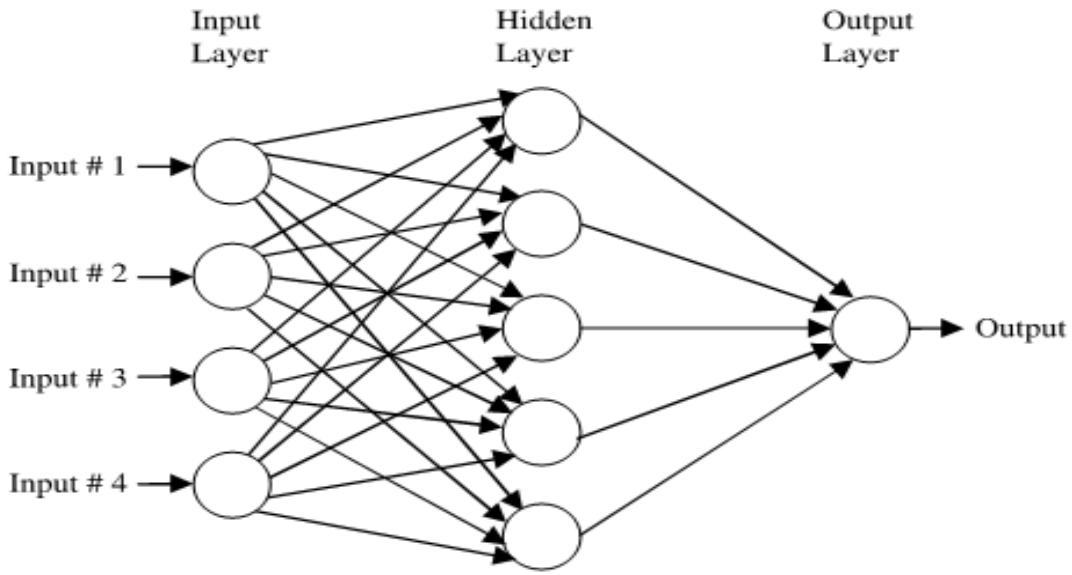
# 3. Tâches du Data Mining

## ❑ Apprentissage supervisé: Régression

- La **classification** est une **forme de prédiction**, où la valeur à prédire est une *étiquette*.
- **La régression (prédiction numérique)** tente à prédire une valeur numérique, telle que le prix d'une marchandise.
- Une manière très populaire est d'utiliser un **réseau de neurones** (**Neural Network**).

## ❑ Apprentissage supervisé: Régression

- **Un réseau de neurone artificiel** est une technique de **modélisation complexe** basée sur un **modèle de neurone humain**.
- Un réseau neuronal **reçoit un ensemble d'entrées** et utilisé pour **prédire une ou plusieurs sorties**.



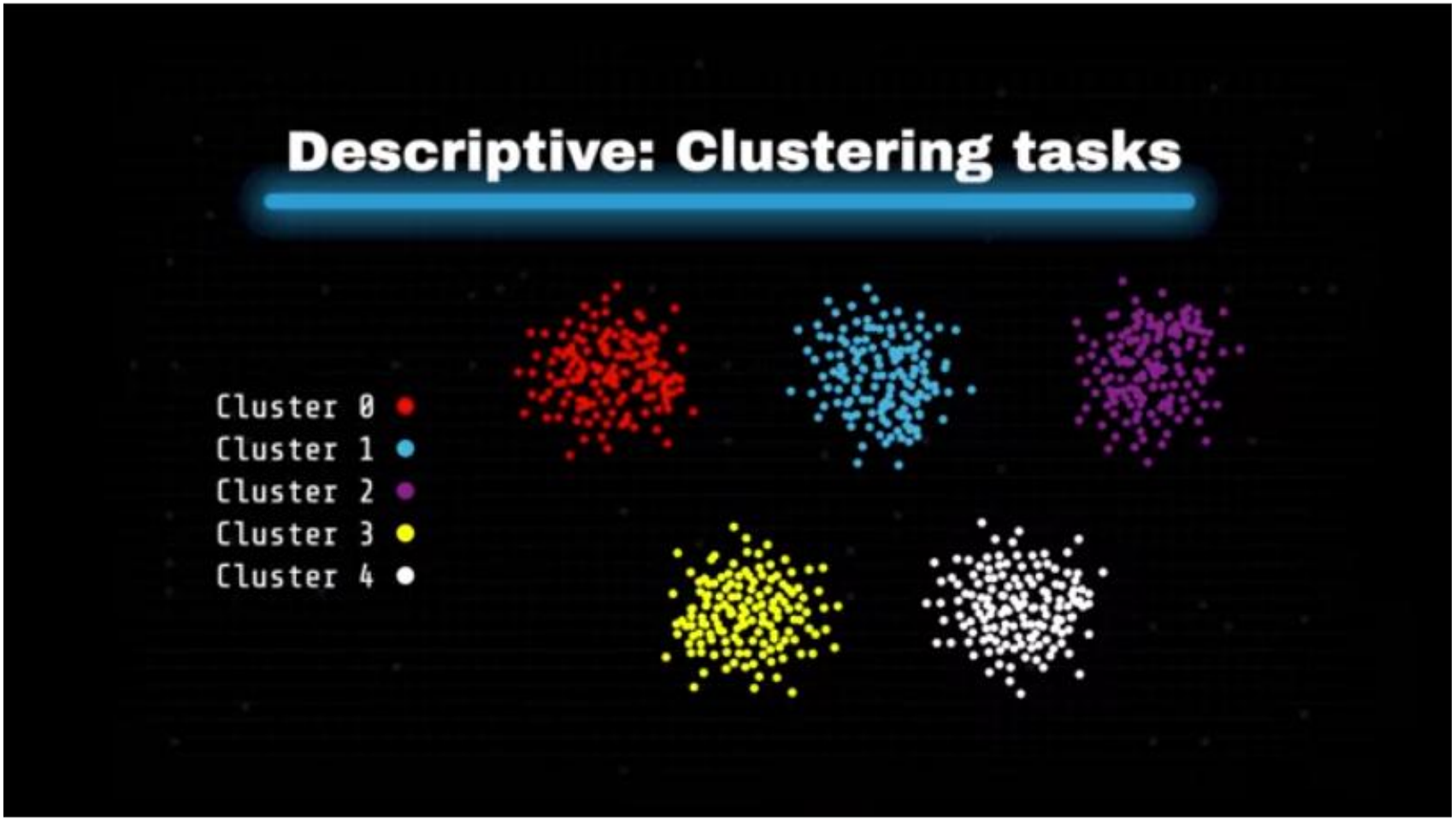

Réseau de neurones artificiel

# 3. Tâches du Data Mining

## ❑ Apprentissage non supervisé: Segmentation

- **La segmentation (Clustering)** est un partitionnement logique d'un ensemble de données (**DataSet**) en groupes (**clusters).**
  - **Clusters**: groupes d'instances ayant mêmes caractéristiques.
  - **Apprentissage non supervisé** (classes inconnues).
  - **Problème**: interprétation des clusters identifiés.
  - **Applications**:
    - ✓ Économie (segmentation de marchés),
    - ✓ Médecine (localisation de tumeurs dans le cerveau),
    - ✓ etc.

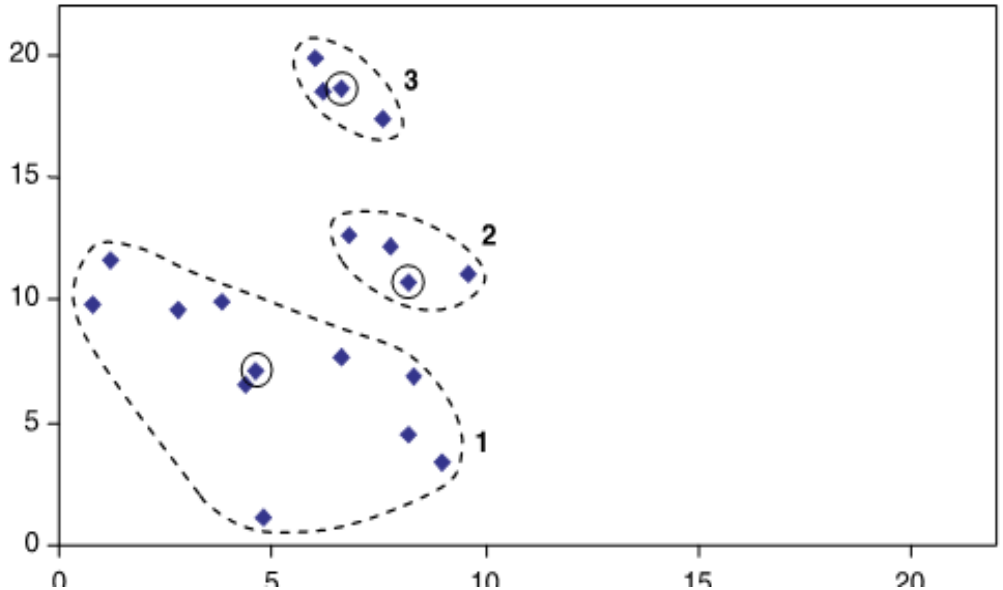
# 3. Tâches du Data Mining



# 3. Tâches du Data Mining

## ❑ Apprentissage non supervisé: Segmentation

- Les **algorithmes de clustering** examinent les données pour trouver des groupes **d'éléments similaires (calcul de similarité)**.
- Par exemple, une compagnie d'assurance peut regrouper ses clients en fonction du *revenu, de l'âge, des types de police souscrite ou de l'expérience des sinistres antérieurs*.

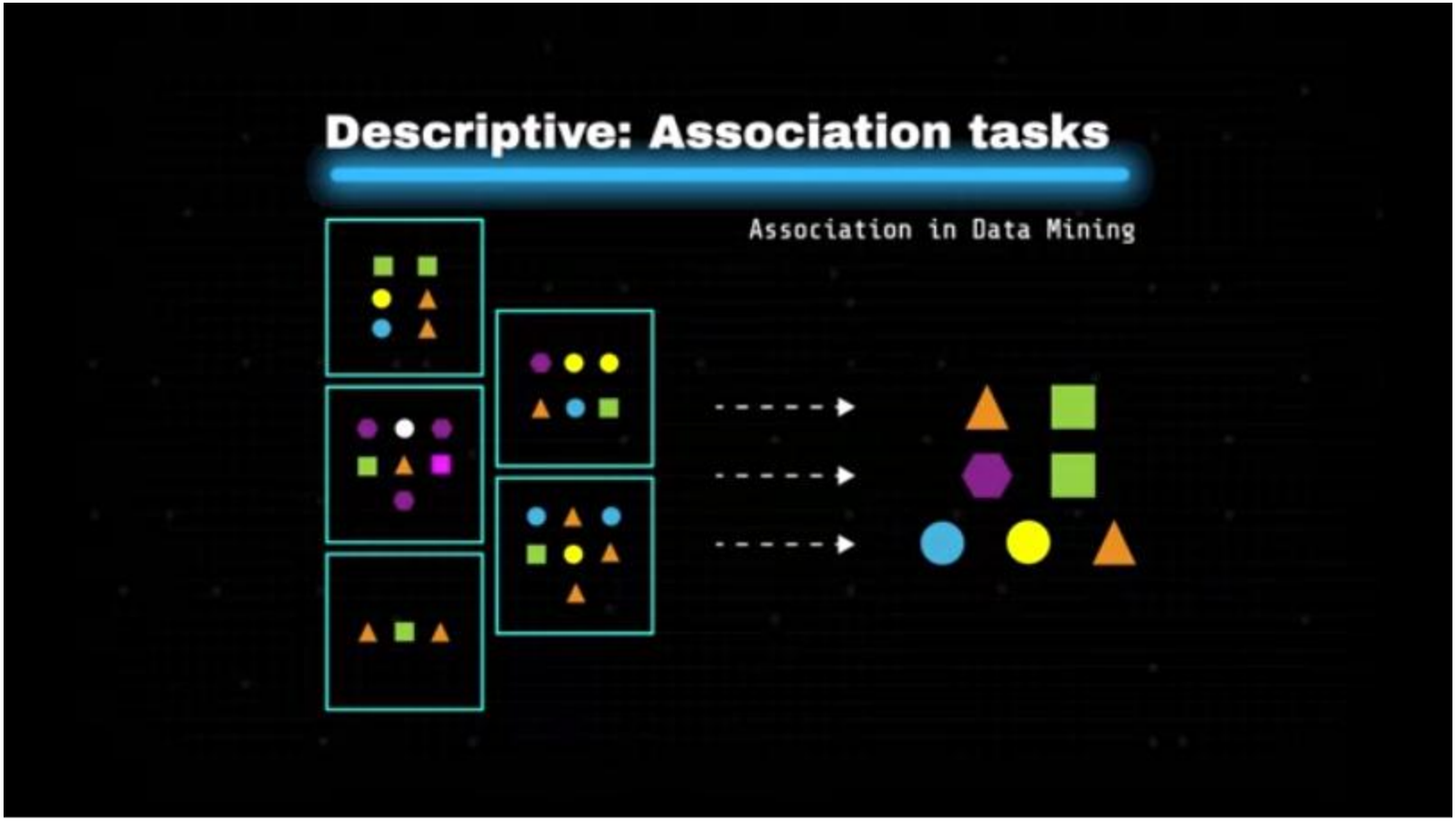


# 3. Tâches du Data Mining

## ❑ Apprentissage non supervisé: Règles d'association

- Les **règles d'association** (***Association rules***) sont des **corrélations** (relations) entre **attributs**.
- **Applications**: *gestion des stocks, web (pages visitées), etc*.
- **Exemple**
  - Articles figurant dans le même ticket de caisse.
  - Ex : achat de produit x + produit y ==> produit z

# 3. Tâches du Data Mining



# 3. Tâches du Data Mining

## ❑ Apprentissage non supervisé:Recherche de séquences

- Prise en compte du temps (**série temporelle**).
- Liaisons entre événements sur une période de temps.
- Extension des règles d'association.
- Achat Télévision ==> Achat Magnétoscope d'ici 5 ans.
- **Applications**:
  - ✓ marketing direct (anticipation des commandes),
  - ✓ Bourse (prédiction des valeurs des actions)

# 3. Tâches du Data Mining

## ❑ Apprentissage non supervisé: Détection et déviation

- **Instances** ayant des caractéristiques les plus différentes des autres.
  - ➢ Basée sur la notion de distance entre instances.
  - ➢ Expression du problème
    - ✓ **Temporelle** : évolution des instances ?
    - ✓ **Spatiale** : caractéristique d'un cluster d'instances ?
- **Applications**
  - ➢ Détection de fraudes (transactions avec une carte bancaire inhabituelle en télémarketing)
- **Caractéristiques**
  - ➢ Problème d'interprétation: bruit ou exception.

# 3. Tâches du Data Mining

## ❑ Apprentissage semi supervisé

- L'apprentissage **semi-supervisé** utilise un ensemble de données **étiquetées** (supervisé) **et non étiquetées** (non supervisé)
- Il est démontré qu'il permet d'améliorer **significativement la qualité** de l'apprentissage.
- L'apprentissage semi-supervisé a un intérêt pratique évident pour des données de grande taille.
- L'apprentissage non supervisés utilisé pour générer automatiquement **des étiquettes** utilisées par l'apprentissage supervisé.
- L'apprentissage semi-supervisé a un grand avantage financier.
- Il permet de baisser le coût d'étiquetage des grandes données

# 3. Tâches du Data Mining

## Algorithmes de Data Mining

- Naîve Bayes (Bayesien naïf)
- K-Nearest neighbour (K plus proche voisin)
- Support Vector Machine (SVM)
- Decision Tree (Arbres de décision)
- Random Forest (Forêt aléatoire)
- Régression linéaire, Régression logistique
- Neural Networks (Réseaux de neurones)
- K-means (K-moyennes),
- Algorithmes génétiques
- Chaînes de Markov cachées
- Soft computing : ensembles flous …

# 3. Tâches du Data Mining

## ❑ Résumé

- **Data Mining**: découverte automatique de **modèles** intéressants à partir **d'ensemble de données** de grande taille.
- **KDD** (knowledge data discovery) est un processus :
  1. **Pré-traitement** (Pre-processing)
  2. **Data mining**
  3. **Post-traitement** (Post-processing)
- Pour le data mining, utilisation de différents …
  - **Source de données**: relationelle, orientée objet, spatiale, WWW, …
  - **Connaissances**: classification, clustering, association, …
  - **Techniques**: apprentissage, statistiques, optimisation, …
  - **Applications**: génomique, télécom, banque, assurance, …

# Pré-traitement et visualisation des données

## ❑ Données

- **Données Discrètes**:
  - ✓ Binaires (Ex. *sexe*, …),
  - ✓ Énumératives (Ex. *Couleur*, …),
  - ✓ Énumératives ordonnées (Ex. *1:très satisfait, 2: satisfait,…*).
- **Données Continues:**
  - ✓ Entières
  - ✓ Réelles (*âge*, *salaire*, …)
- **Dates**
- **Données de type complexe:**
  - ✓ Données textuelles,
  - ✓ images, vidéos,
  - ✓ Pages/ liens web, Multimédia, …

## ❑ Variables

- **Variables nominales:**
  - ➢ Peuvent être de forme **numérique**, mais les valeurs numériques n'ont pas d'interprétation mathématique.
    - **Exemple 1**: le nom ou la couleur d'un objet.
    - **Exemple 2**: étiqueter 10 personnes comme des nombres 1, 2, 3, ..., 10, mais non arithmétique par ex. (1 + 2 = 3).
    - **Une classe** peut être considérée comme une variable nominale.

## ❑ Variables

- **Variable binaire:** un cas particulier d'une **variable nominale** qui prend deux valeurs possibles: **vrai** ou **faux** (**1** ou **0**).
- **Variable ordinale:** similaire aux variables nominales, sauf qu'une variable ordinale a des valeurs qui peuvent être disposées dans un **ordre significatif**.
  - **Exemple**: *petit, moyen, grand.*
- **Variable entière:**
  - L'arithmétique avec des variables entières est **significative**
  - **Exemple**: *(1 étudiant+ 2 étudiants= 3 étudiants).*

## ❑ Variables

- **Variables Interval-scaled** (**à l'échelle par intervalles**) prennent des valeurs numériques qui sont mesurées à intervalles égaux à partir d'un point zéro ou d'une origine.
- Cependant l'origine n'implique pas une véritable absence de la caractéristique mesurée.
- **Ratio-scaled Variables** (**à échelle de rapport**) similaires aux variables à échelle d'intervalle, sauf que le point **zéro reflète l'absence de la caractéristique mesurée**,

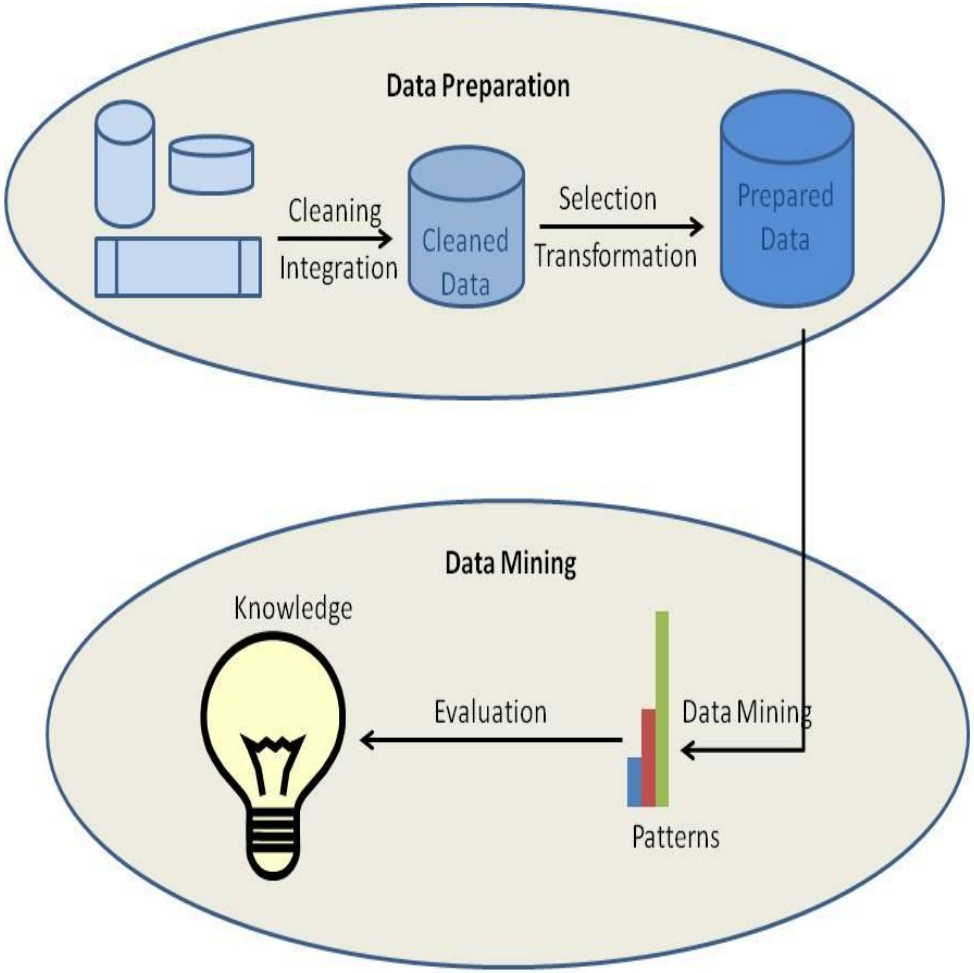
## ❑ Attributes

- Les attributs dont en deux types seulement:
  - ➢ **Catégoriques** correspondant aux variables **nominales**, **binaires** et **ordinales.**
  - ➢ **Continus** correspondant à des variables **entières**, **à l'échelle par intervalle** et **à l'échelle par rapport**

# 2. Préparation et nettoyage des données

- Les bases de données du monde réel sont très sensibles aux données manquantes et incohérentes en raison
  - ➢ De leur **grande taille.**
  - ➢ De leur **hétérogénéité**: sources multiples.
- Les données de **faible qualité** conduiront à des **résultats** d'extraction de **mauvaise qualité**.

# 2. Préparation et nettoyage des données

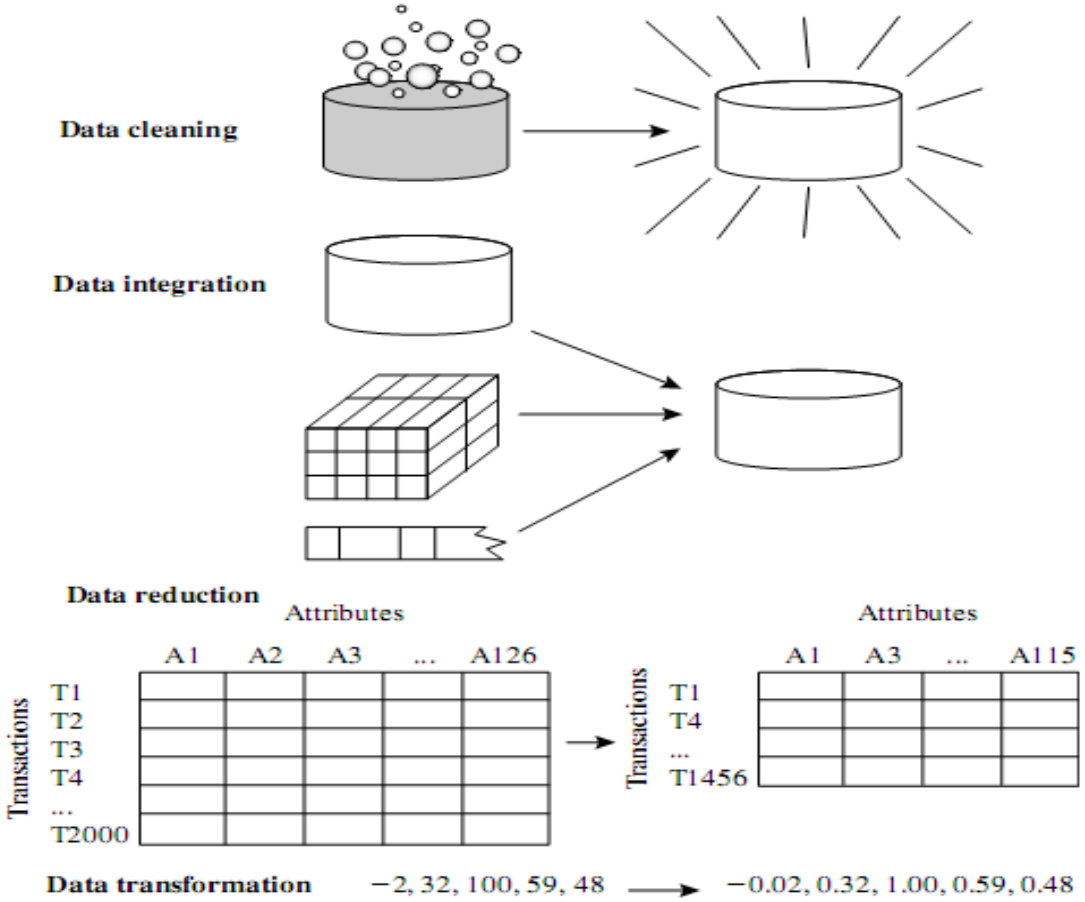


# 2. Préparation et nettoyage des données

- Il existe **plusieurs techniques de prétraitement** des données:
  - ***Nettoyage des données (Data Cleaning) :*** pour supprimer le bruit et corriger les incohérences dans les données.
  - ***Intégration des données (Data Integration)*** fusionne les données de plusieurs sources dans un **entrepôt de données (Data Warehouse)**.
  - ***Réduction des données (Data Reduction) :*** pour réduire la taille des données en agrégeant, en éliminant les fonctionnalités redondantes ou en mettant en cluster.

# 2. Préparation et nettoyage des données

- ***Transformation de données (Data Transformation)*** appliquée pour tomber dans une petite plage de 0,0 à 1,0 pour améliorer la précision des algorithmes d'extraction impliquant des **mesures de distance**.



# 3. Bruit et données manquantes

## ❑ Données manquantes (Missing Values)

- Le nettoyage des données vise à remplir les valeurs manquantes, supprimer les valeurs aberrantes et résoudre les incohérences
- **Comment remplir les valeurs manquantes pour un attribut** ?

1. **Ignorer l'instance (Discard Instance):** lorsque l'étiquette de classe est manquante pour une classification. Cette méthode n'est pas efficace, sauf si l'instance contient plusieurs attributs avec des valeurs manquantes.
2. **Remplir manuellement la valeur manquante:** cette approche prend du temps et peut ne pas être réalisable étant donné un ensemble de données avec de nombreuses valeurs manquantes.

## ❑ Données manquantes (Missing Values)

- On peut remplir une valeur manquante:
  - ➢ Par **une constante globale** (**The most frequent**).
  - ➢ Par **une mesure de tendance centrale** (**mean ou median**).
  - ➢ **Par la valeur probable** déterminée par *régression, probabilité bayésienne, arbre de décision, …*

# 3. Bruit et données manquantes

## ❑ Btuit (Noisy)

- Analyse des **valeurs aberrantes** (**Outliers Analysis**): éléments de données qui ne peuvent pas être regroupés dans une classe ou un cluster donné.

- Bien que les valeurs aberrantes puissent être considérées comme du **bruit et rejetées** dans certaines applications, elles peuvent révéler des connaissances importantes dans d'autres domaines, et peuvent donc être très importantes et leur analyse précieuse.

# Description de l'environement WEKA

# 3. Description de l'environnement WEKA

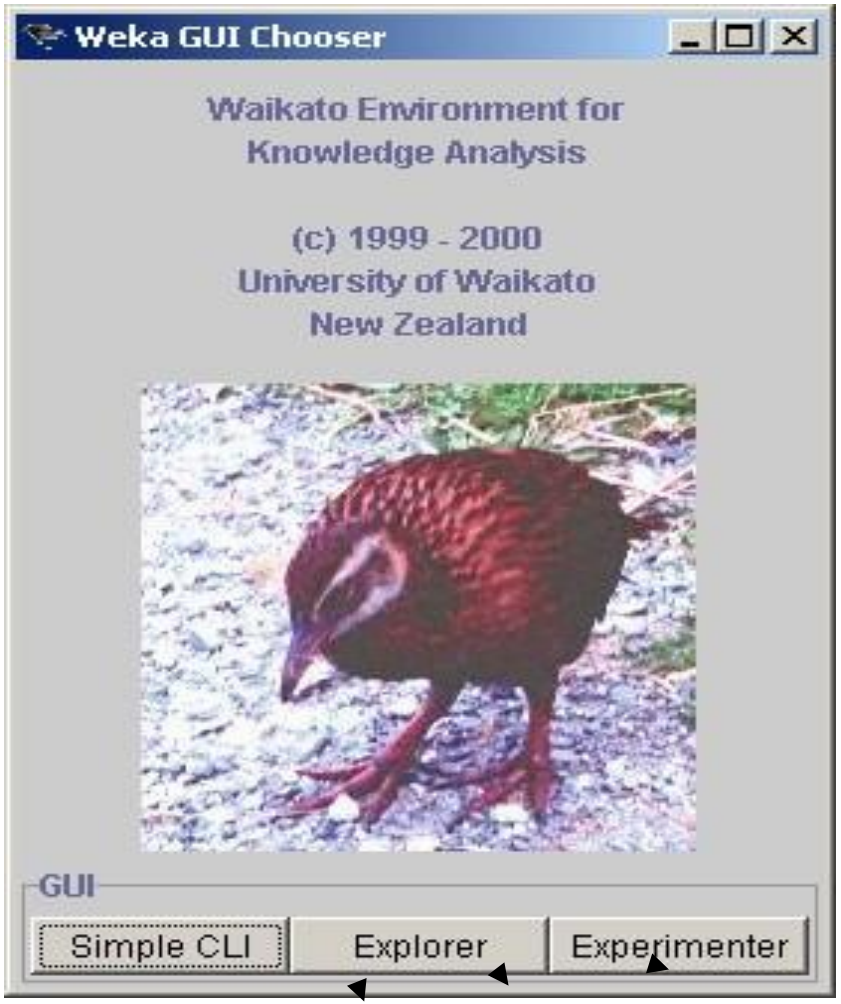
- Une Logiciel gratuit disponible sur le web :
  - **http://www.cs.waikato.ac.nz/ml/weka/**
- Plate forme logicielle en Java tournant sous :
  - **Windows**
  - **Linux**
- Facile à prendre en main


WEKA

**W**aikato **E**nvironment for **K**nowledge **A**nalysis

# 3. Description de l'environnement WEKA

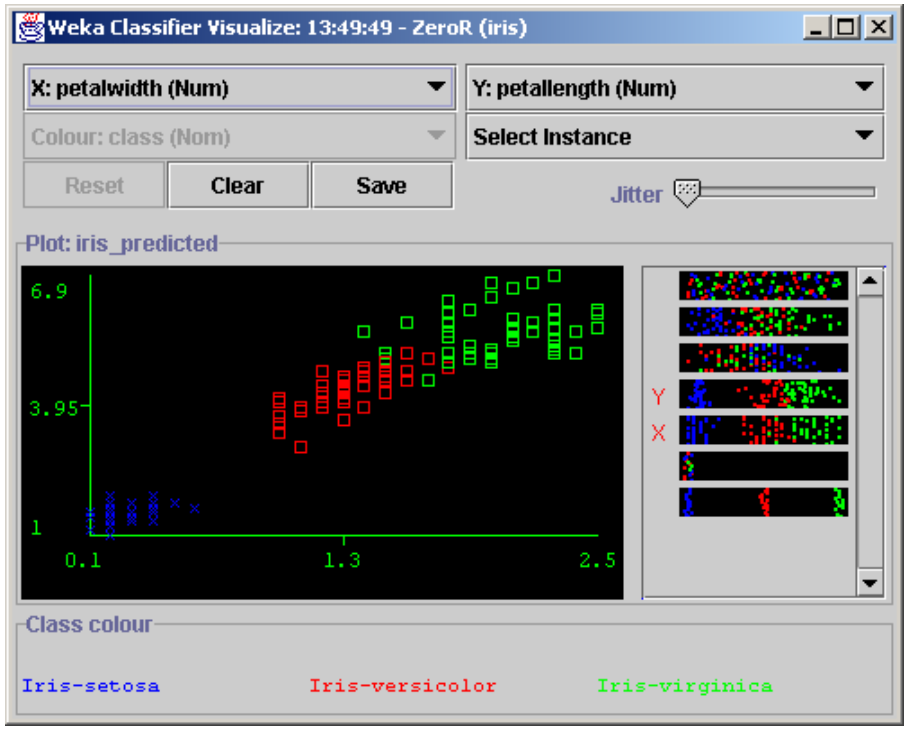
- **Interface en ligne de commande**
- **Explorer (interface graphique)**
  - Filtre
  - Apprentissage (clustering, classification, ...)
  - Sélection d'attributs
  - Visualisation des données.
- **Expérimenter (environnement d'expérience)**
  - Test d'une méthode spécifique sur un ensemble de données avec des critères variés pour la comparaison de résultats



# 3. Description de l'environnement WEKA

- **En entrée :** fichiers, base de données, Url, …
- **En sortie** : affichage des résultats, sortie des résultats dans des fichiers, visualisation graphique …

Exemple de visualisation après une **classification:** une couleur représente une classe



# 3. Description de l'environnement WEKA

## Fonctions disponibles :

- Filtre et Preprocess sur les données
- Classification
- Clustering
- Règles d'association
- Sélection d'attributs
- Visualisateur

# END