

# Fouille de Données

# Data Mining

## **Introduction Générale**

# Plan du cours

1. Définitions et généralités
2. Les types de données à fouiller
3. Modèles de Processus de Data Mining
4. Les tâches du data mining

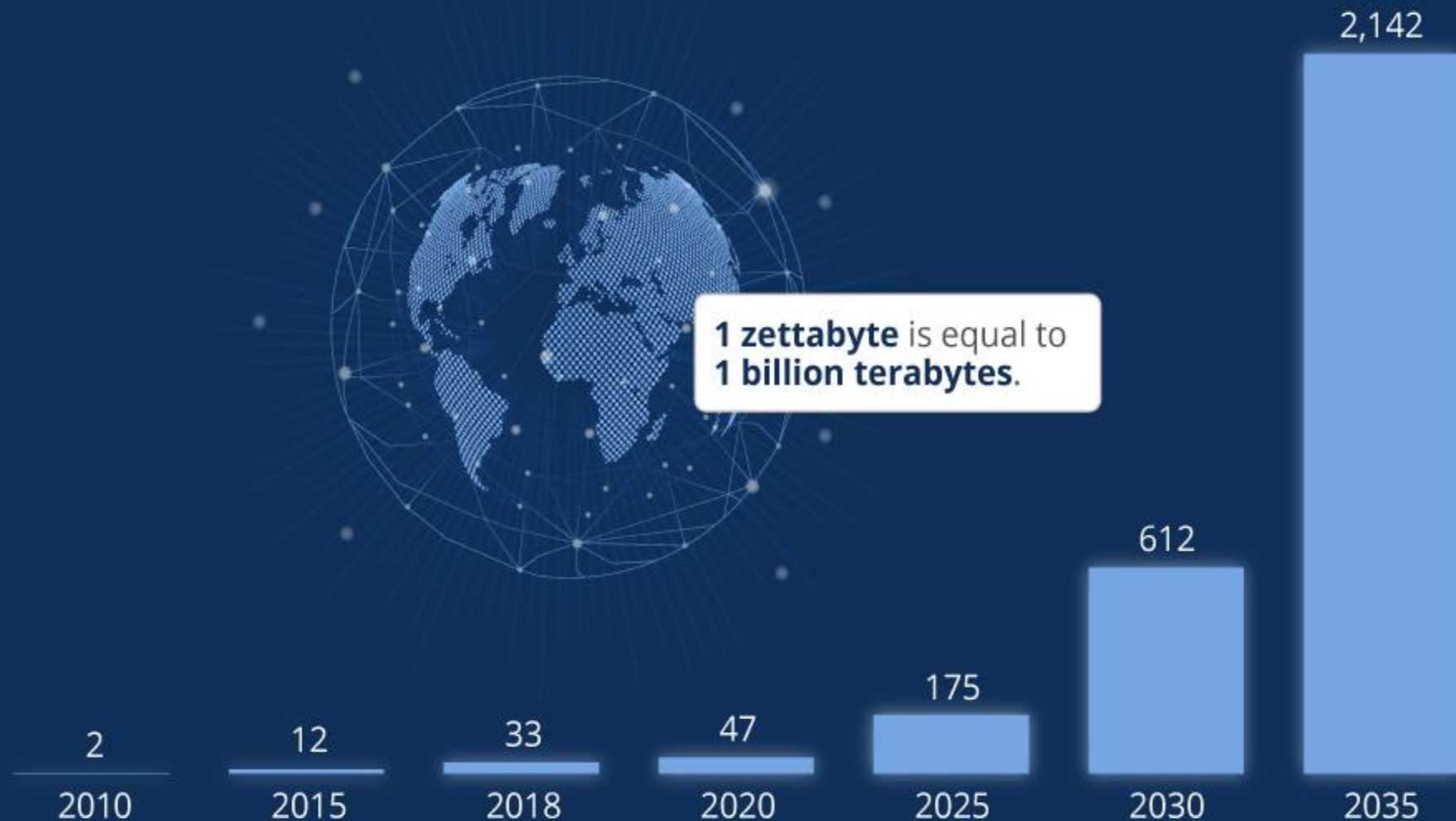
Fouille de Données

Data Mining

« L'humanité produit autant de données en deux jours qu'elle ne l'a fait en milliers d'années. »

# Global Data Creation is About to Explode

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)



# THE WORLD OF DATA

NUMBER  
OF EMAILS  
SENT  
EVERY SECOND

2.9

MILLION

DATA  
CONSUMED BY  
HOUSEHOLDS  
EACH DAY

375

MEGABYTES

VIDEO  
UPLOADED TO  
YOUTUBE EVERY  
MINUTE

20

HOURS

DATA PER  
DAY  
PROCESSED  
BY GOOGLE

24

PETABYTES

TWEETS  
PER  
DAY

50

MILLION

TOTAL MINUTES  
SPENT ON  
FACEBOOK  
EACH MONTH

700

BILLION

DATA SENT  
AND RECEIVED  
BY MOBILE  
INTERNET USERS

1.3

EXABYTES

PRODUCTS  
ORDERED ON  
AMAZON PER  
SECOND

72.9

ITEMS



SOURCES: Cnet, comScore, MapReduce, Radicati Group, Twitter, YouTube

IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

# Fouille de Données - Data Mining

## Contexte :



- ✓ “We are living in the **data age**” .
- ✓ La puissance croissante des nouvelles technologies, ont contribué fortement à l’augmentation des collectes des données, la manipulation, et la capacité de stockage.
- ✓ Les données se sont multipliées en **taille**, en **format**, et en **complexité**.
- ✓ Certains experts estiment que le volume des données double tous les ans.
- ✓ L’humanité produit autant de données en deux jours qu’elle ne l’a fait en milliers d’années.
- ✓ Même la façon de les interroger devient données : requêtes sur BDD, historique de navigation sur le Web, recherches sur Google, etc.

# Fouille de Données - Data Mining

## Besoins:

- ✓ The information paradox. - Timo Lüge
- ✓ **Que** doit-on **faire** avec des données coûteuses à collecter et à conserver ?
- ✓ Le besoin des entreprises de **valoriser** les données qu'elles accumulent dans leurs bases qui croissent de manière exponentielle.
- ✓ Le data mining permet **d'exploiter** ces données au profit de l'activité de l'entreprise.
- ✓ Le data mining permet aussi d'augmenter le retour sur investissement des systèmes d'information.
- ✓ Data mining = Moteur essentiel de du **processus décisionnel**.

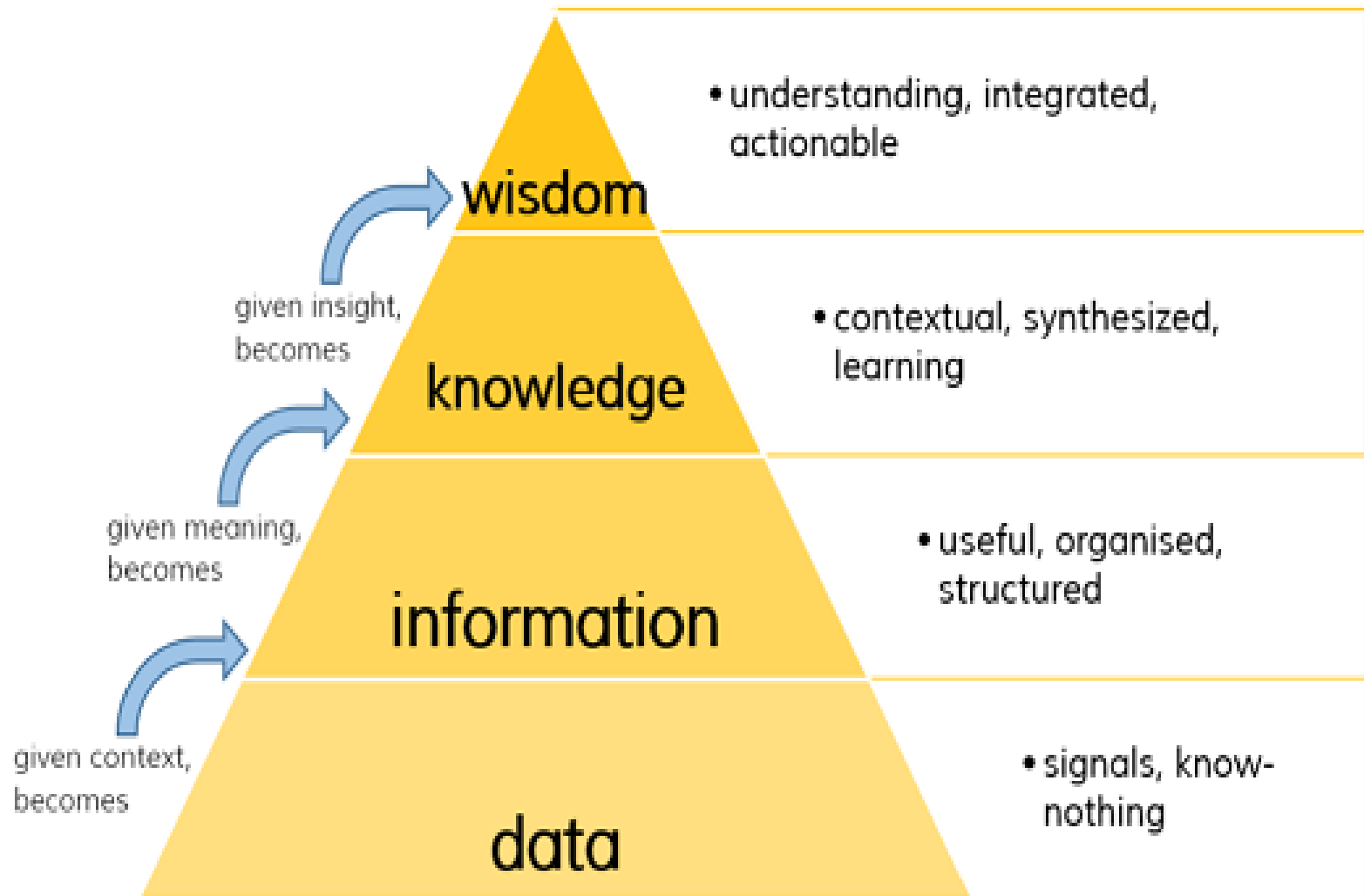
# Fouille de Données



?



Donnée → Information → Connaissance



## Data

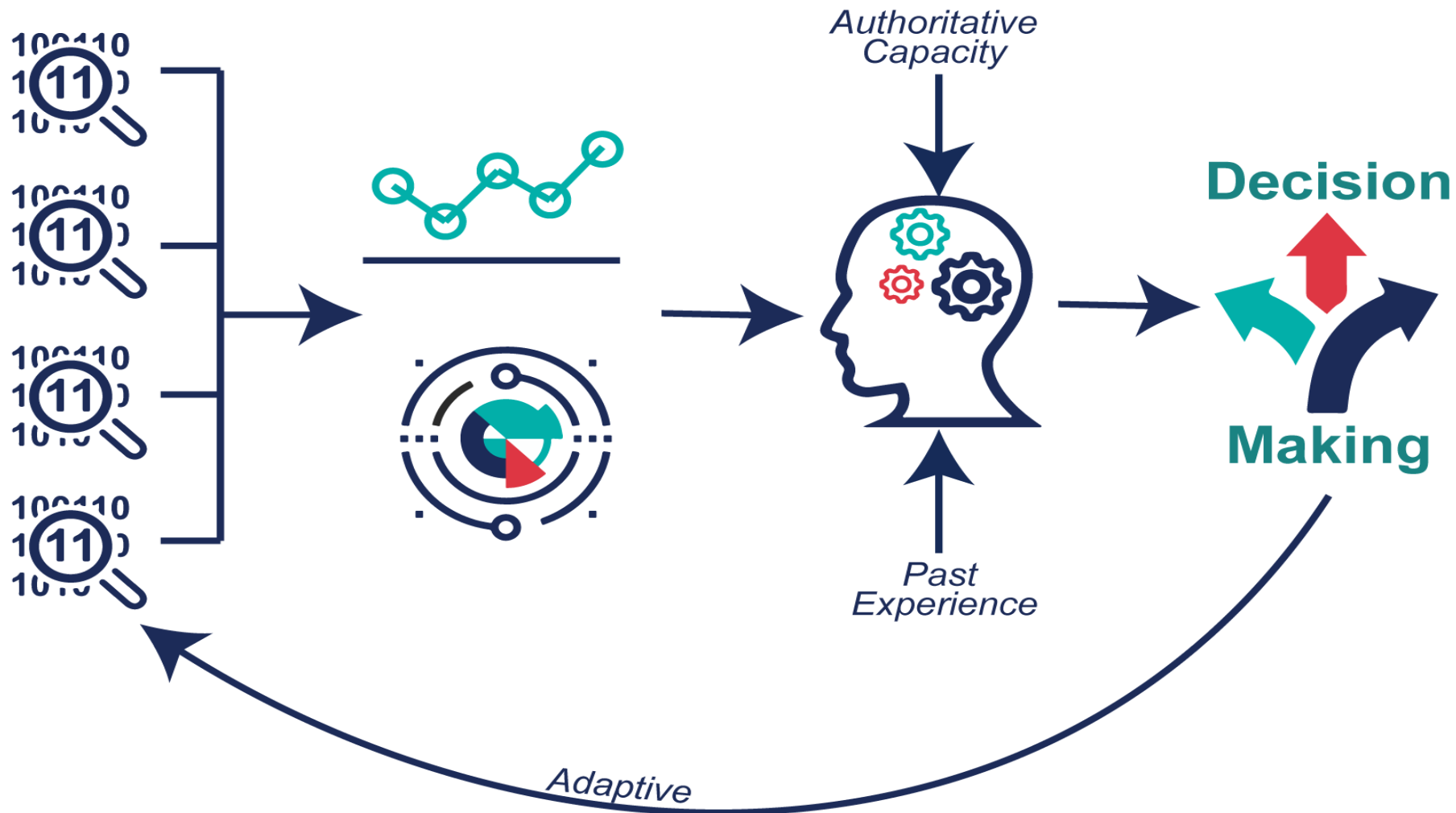
*Raw*

## Information

*Processed*

## Knowledge

*Actionable*



Donnée → Information → Connaissance

- **Donnée** : est le résultat direct d'une **mesure**, collecte, observation, etc.
- **Information** : est une donnée à laquelle un **contexte** lui a été donnée.
- **Connaissance** : est le résultat d'une **réflexion** sur les informations analysées. Est une information à la laquelle un **sens** lui a été donnée.

**Faire parler les données**  
**Pour une meilleure prise de**  
**décisions**

Donnée → Information → Connaissance

Exemple :

- Le nombre d'accidents n'augmente pas quand il fait mauvais.
- 1217 accidents enregistrés durant le mois de mars.
- Une augmentation de 240% du nombre d'incidents par rapport au mois précédent.

Donnée → Information → Connaissance

Exemple :

- Le nombre d'accidents n'augmente pas quand il fait mauvais.

**Connaissance**



- 1217 accidents enregistrés durant le mois de mars.

**Donnée**



- Une augmentation de 140% du nombre d'incidents par rapport au mois précédent.

**Information**

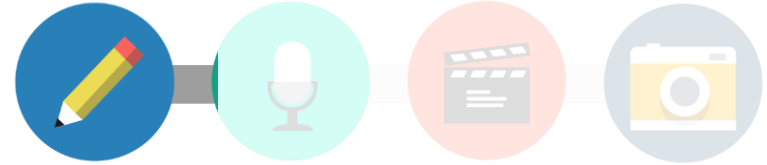


# Fouille de Données



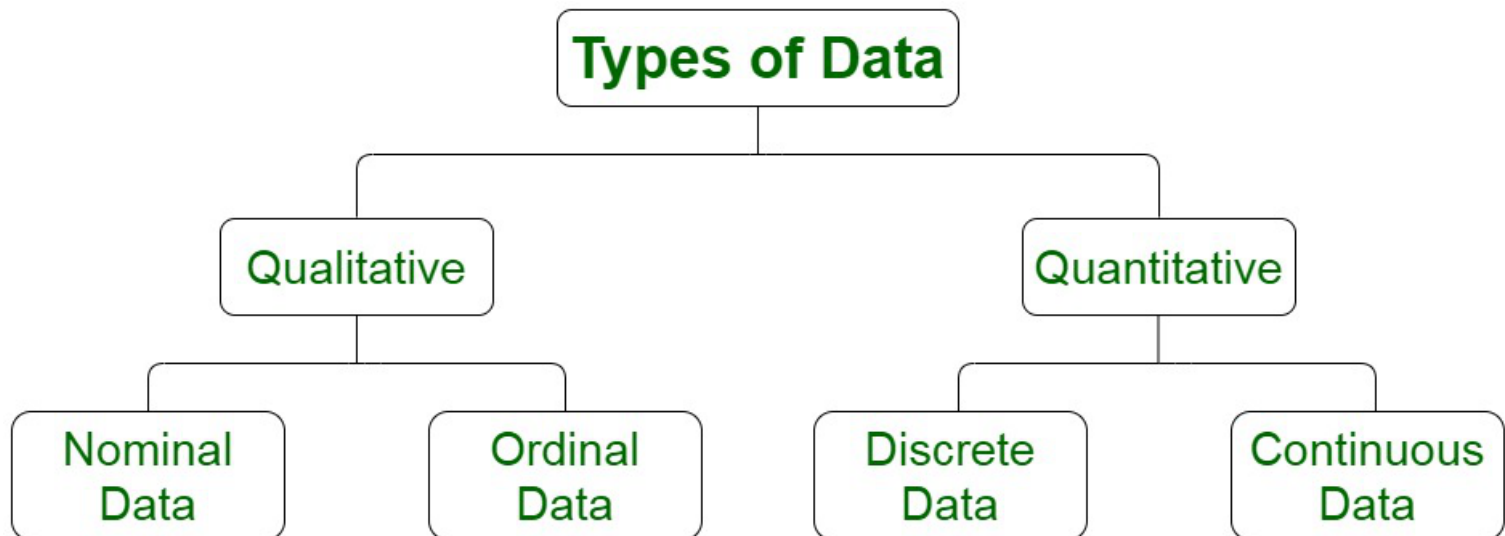
**Quels types ?**

## Quels types de données à fouiller ?

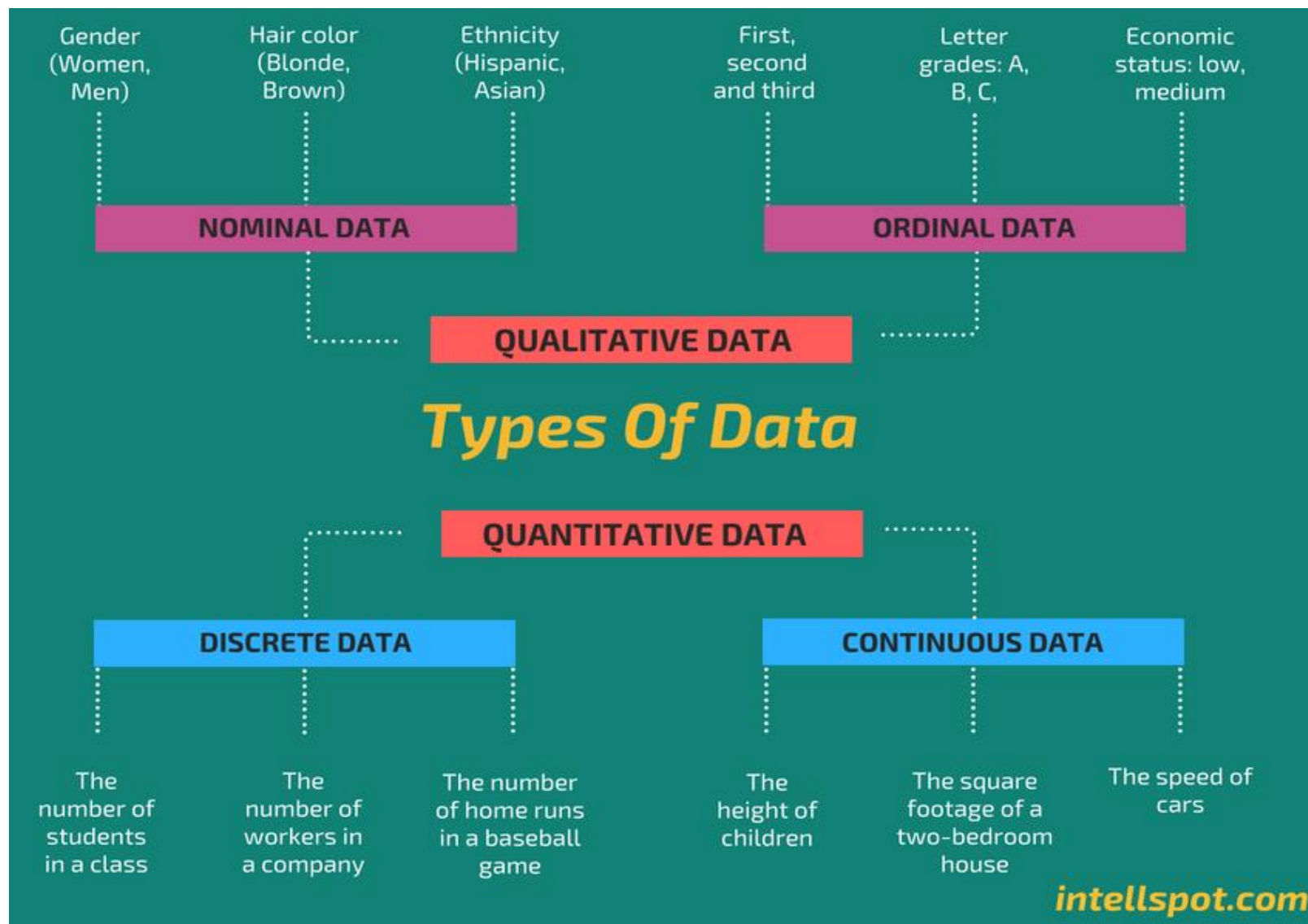


### Deux types d'attributs :

- **Numériques** – Quantitatifs: Comportent les variables réelles ou entières tel que la longueur, le poids, etc. Relation d'ordre ( $5 < 7$ ), mesure de distance, calcul de moyenne, min, max, etc.
- **Catégoriels** - Qualitatifs: Ex : couleur, ou groupe sanguin. Deux variables catégorielles ne peuvent être qu'égales ou différentes.

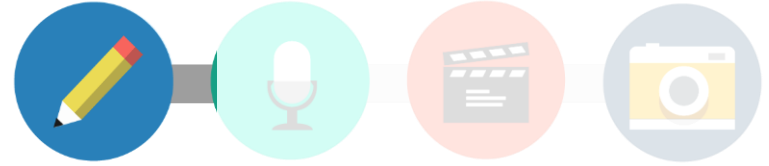


## Quels types de données à fouiller ?

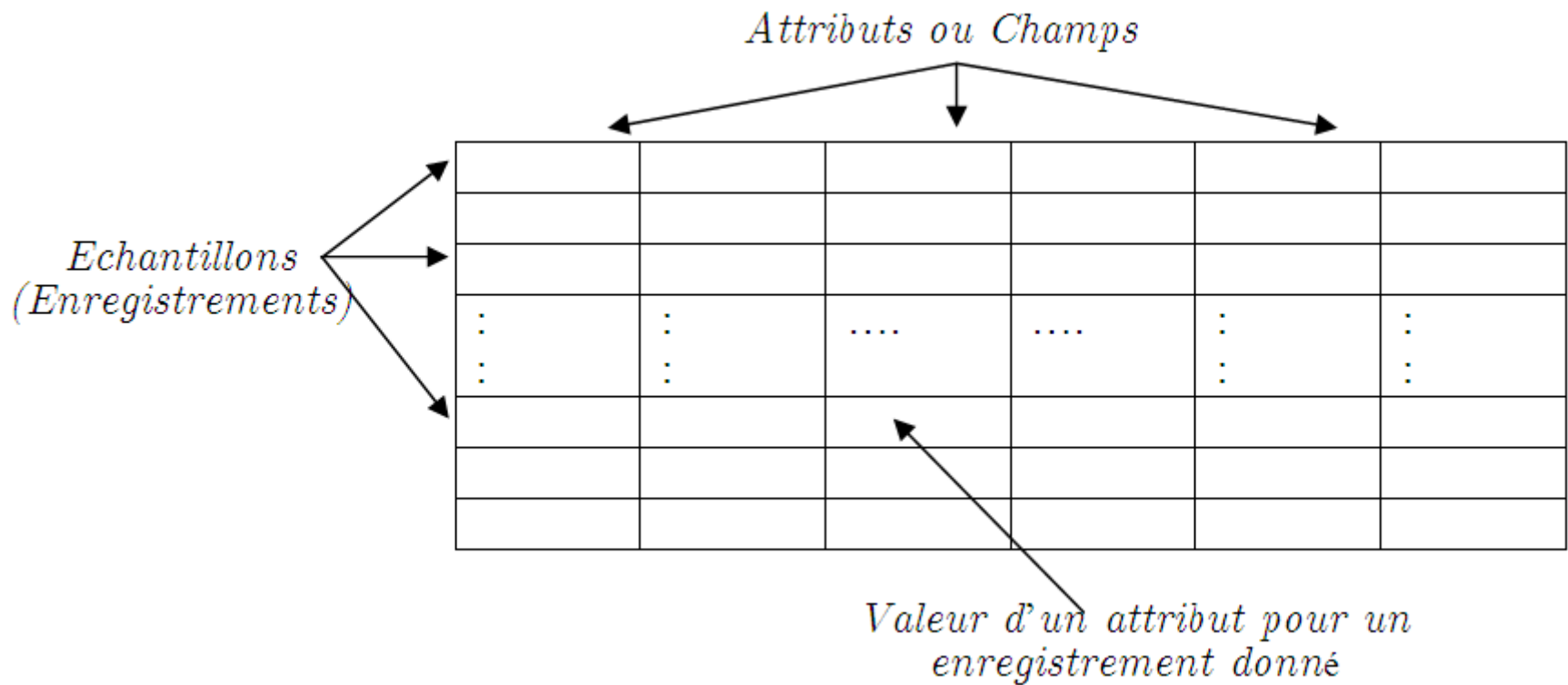




## Quels types de données à fouiller ?



- Données à explorer = Un ensemble d'échantillons / enregistrements
- Echantillon = ensembles d'attributs / champs



## Quels types de données à fouiller ?

- Une donnée est :

- ✓ Enregistrement au sens des bases de données.
- ✓ Individu en statistiques.
- ✓ Instance en POO.
- ✓ Exemple, transaction, etc.

- Une donnée est caractérisée par :

- ✓ Champs en bases de données.
- ✓ Caractéristiques statistiques.
- ✓ Attributs en POO.

ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	85	92	False	No
2	Rainy	80	88	True	No
3	Overcast	83	86	False	Yes
4	Sunny	70	80	False	Yes
5	Sunny	68	?	False	Yes
6	Sunny	65	58	True	No
7	Overcast	64	62	True	Yes
8	Rainy	72	95	?	No
9	Rainy	?	70	False	Yes
10	Sunny	75	72	False	Yes
11	Rainy	75	74	True	Yes
12	?	72	78	True	Yes
13	Overcast	81	66	False	Yes
14	Sunny	71	79	True	No

## Quels types de données à fouiller ?

### **Préparation** des données à fouiller :

- Manuelle : nombre limite d'enregistrements.
- Semi-automatique : nombre moyen d'enregistrements.
- Automatique : immenses bases de données.



### **Qualité** des données à fouiller :

- Données précises, outliers, et non manquantes : noms écrits correctement, valeurs dans les bons intervalles.
- Données enregistrées dans le bon format : numérique/caractère, entière /réelle, etc.
- Redondance minimisée, voire éliminée.

# Fouille de Données



?

## Fouille de Données - Data Mining

**Quoi ?** Explorer, rechercher, et extraire des informations et des connaissances à partir des **données**.

**Pourquoi ?** Condition d'évolution. Inférer des lois. Aboutir à des **connaissances** opérationnelles/actionnables nécessaires à la prise **décisions**.

**Laquelle ?** Les **pertinentes** d'entre elles pouvant aider à comprendre les données ou à prédire le comportement des données futures.

**Depuis où ?** Entrepôts de données (data warehouse), des bases de données distribuées, datasets, Internet.

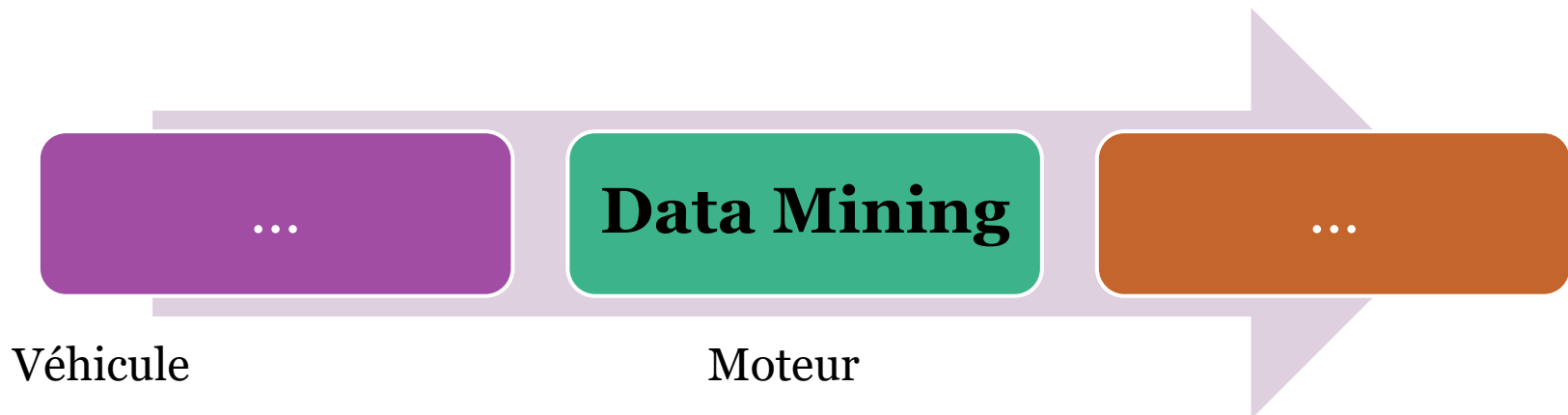
**Avec quoi ?** **Algorithmes**, méthodes, et techniques de l'informatique, l'IA, les statistiques, etc.

**Comment ?** Ces connaissances (qui doivent être validées) sont exprimées sous forme de **modèles** maths, logiques, rapports, graphiques, tendances, etc.

# Fouille de Données - Data Mining

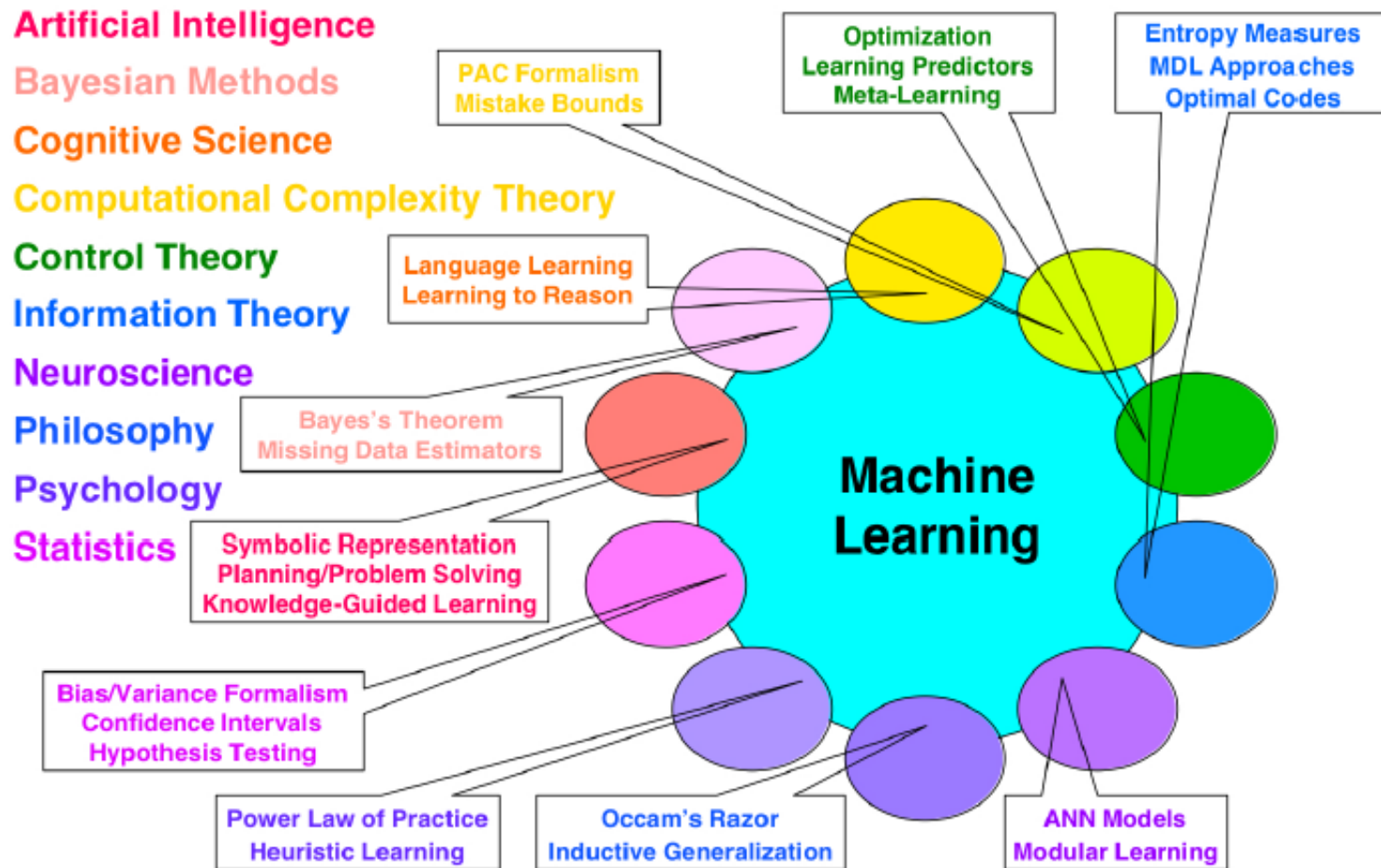
Confusion entre :

- ✓ Data Mining, et
- ✓ Extraction des connaissances à partir des données - ECD
- ✓ Ou Knowledge Discovery in Databases - KDD



# Fouille de Données - Data Mining

Techniques :



# Fouille de Données - Data Mining

Techniques :

Data Mining



A close  
up view

Big Data

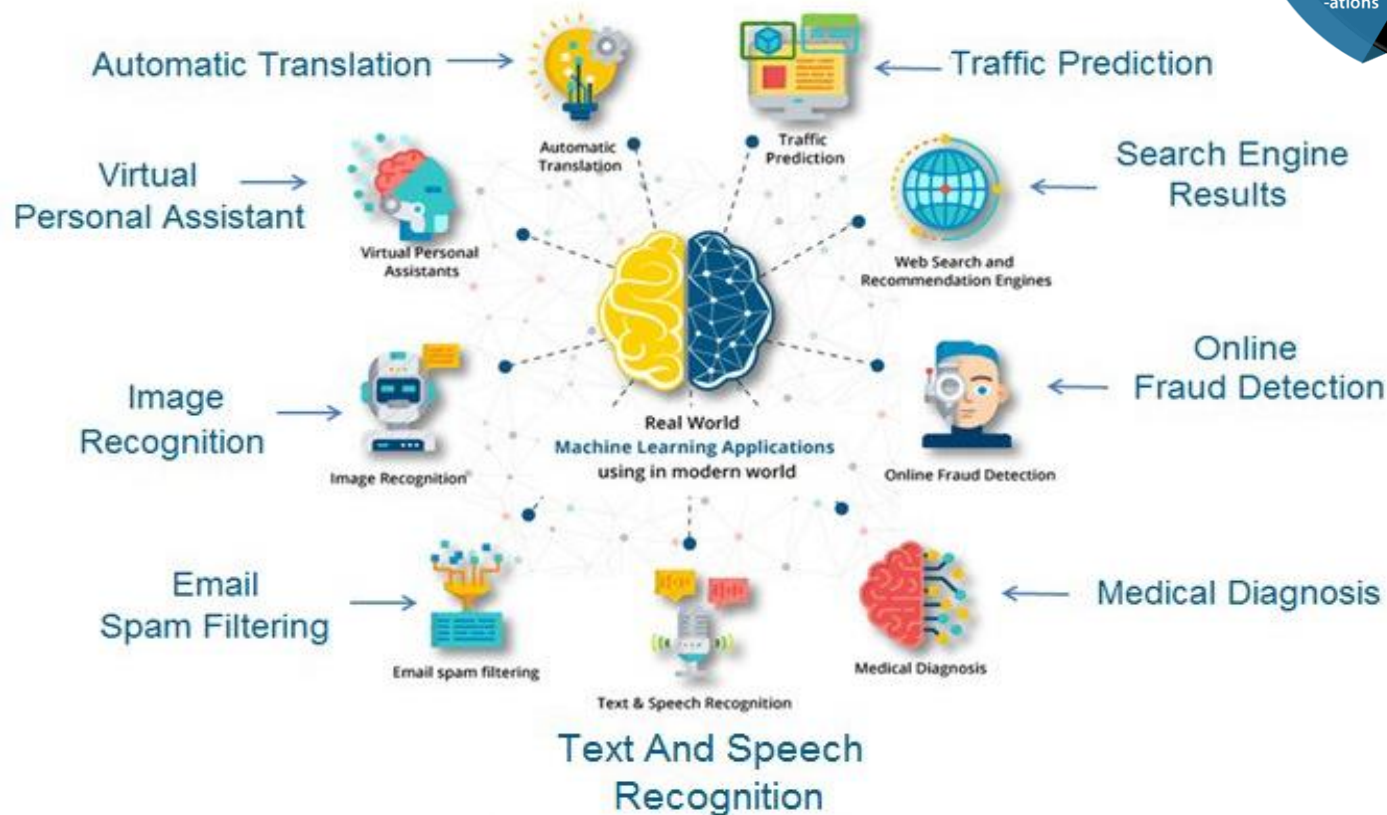
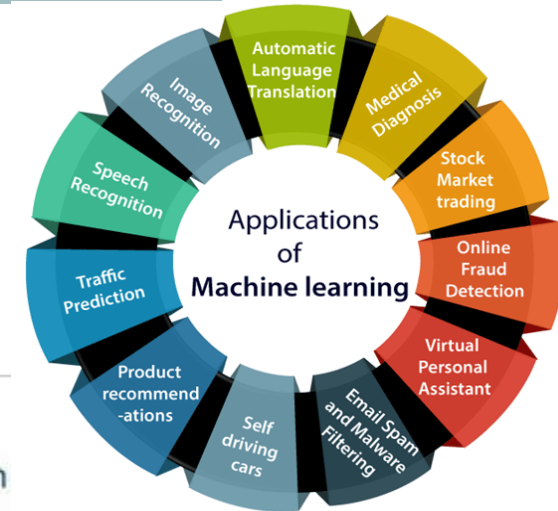


The big  
picture



# Fouille de Données - Data Mining

## Domaines d'utilisation et applications :



# Fouille de Données - Data Mining

## Domaines d'utilisation et applications :

- La gestion de la relation client qui consiste à analyser le comportement de la clientèle pour mieux la fidéliser et lui proposer des produits adaptés.
- organisation des rayonnages dans les supermarchés.
- Organisation de campagne de publicité, emailing, promotions, etc.
- Gestion du risque lié à l'attribution de prêts par le Credit Scoring.
- La détection des fraudes fiscales.
- Détection d'usage frauduleux de cartes bancaires.
- Diagnostiques médicaux, identification des thérapies à succès, etc.
- Web mining, text mining - indexation, image-mining, etc.

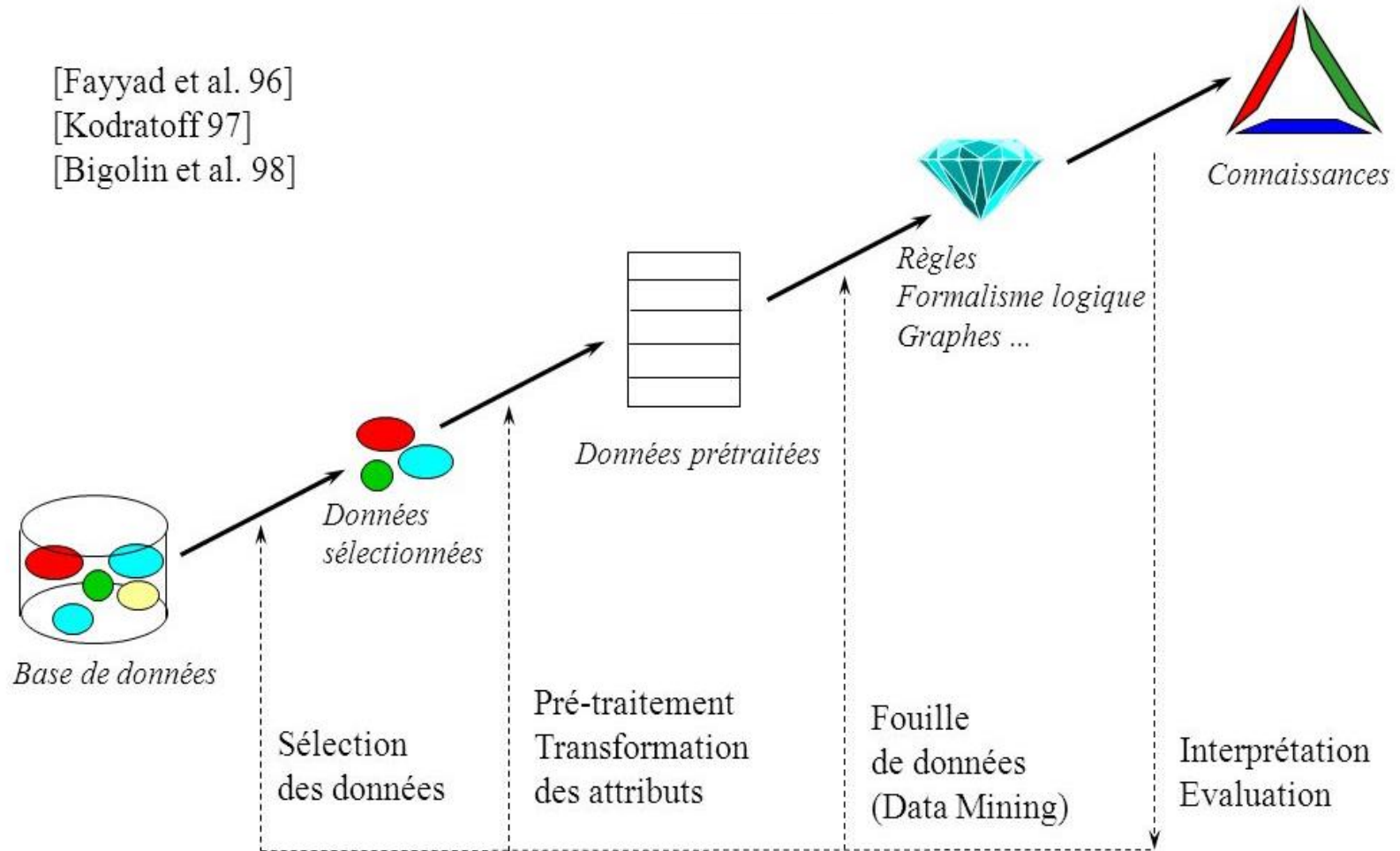
# Modèles de Processus de Data Mining

## KDD – Knowledge Discovery in Databases

[Fayyad et al. 96]

[Kodratoff 97]

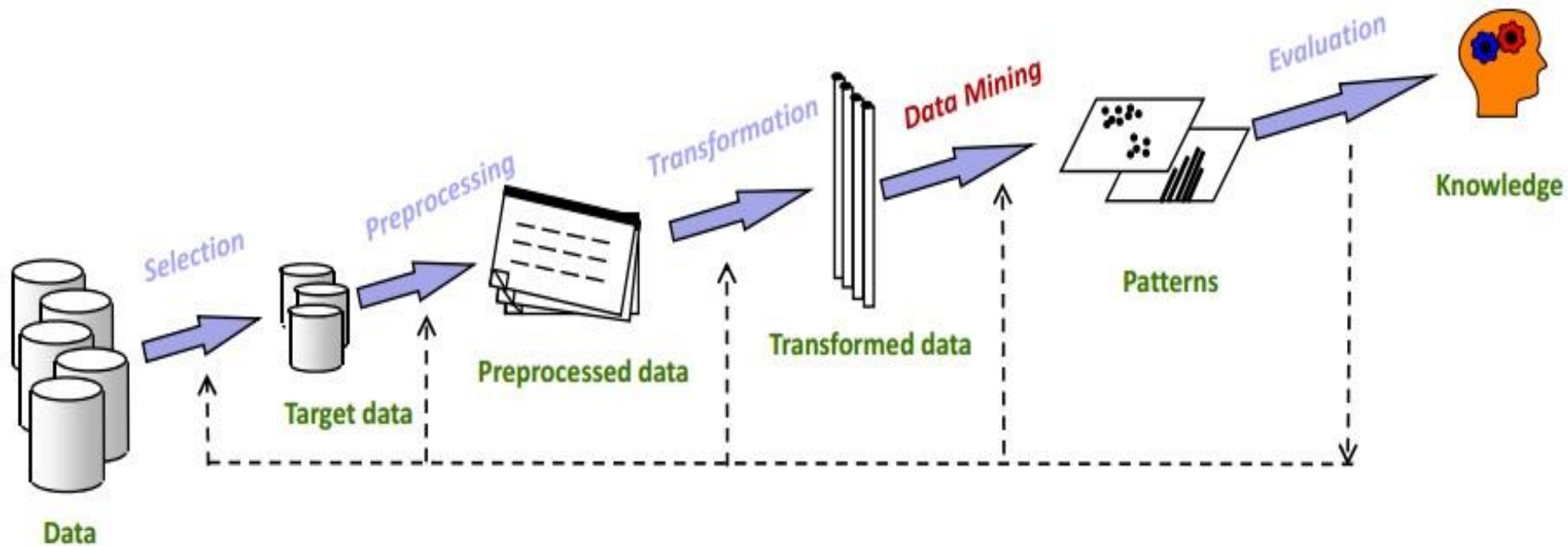
[Bigolin et al. 98]



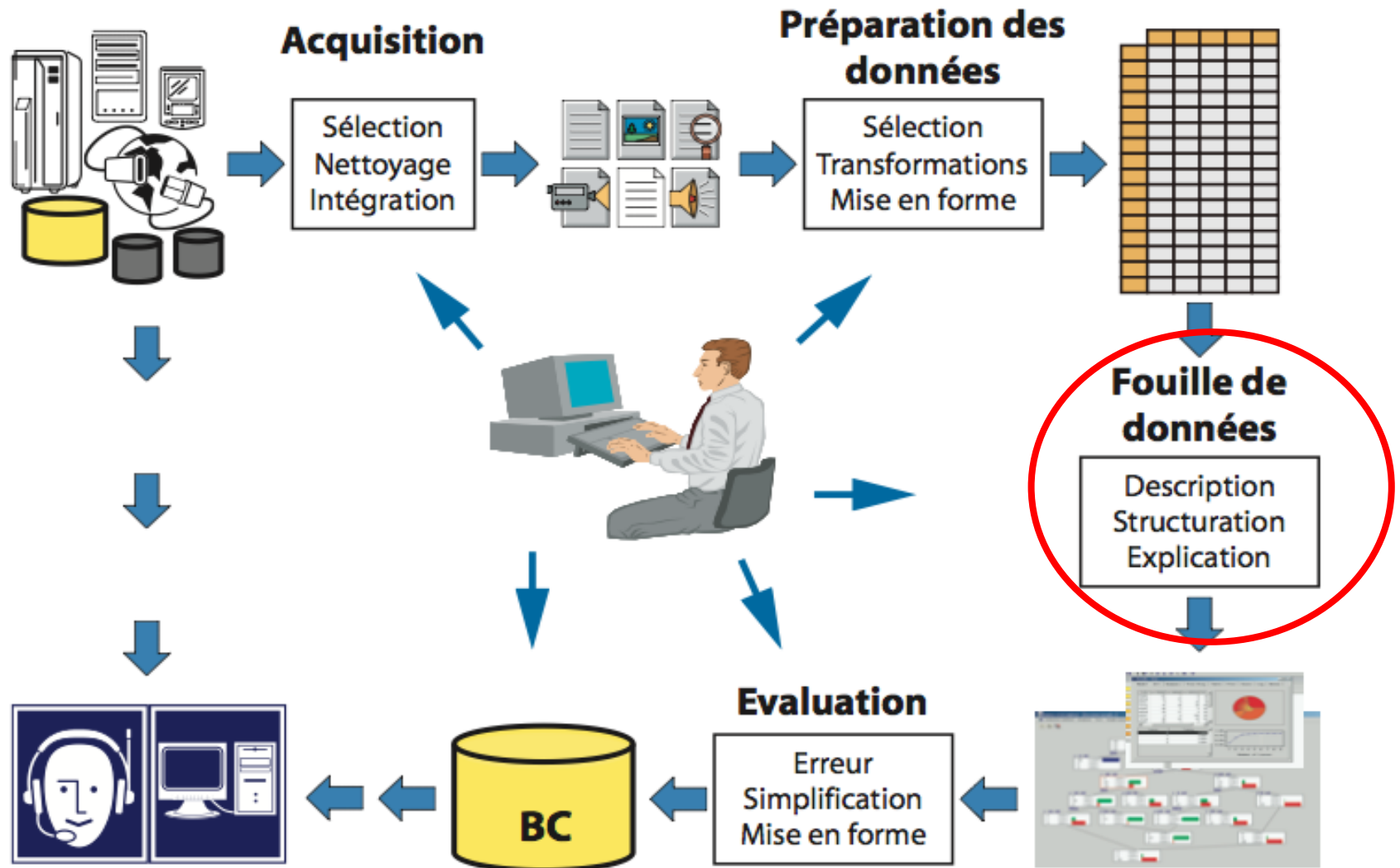
# Modèles de Processus de Data Mining

## KDD – Knowledge Discovery in Databases

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]

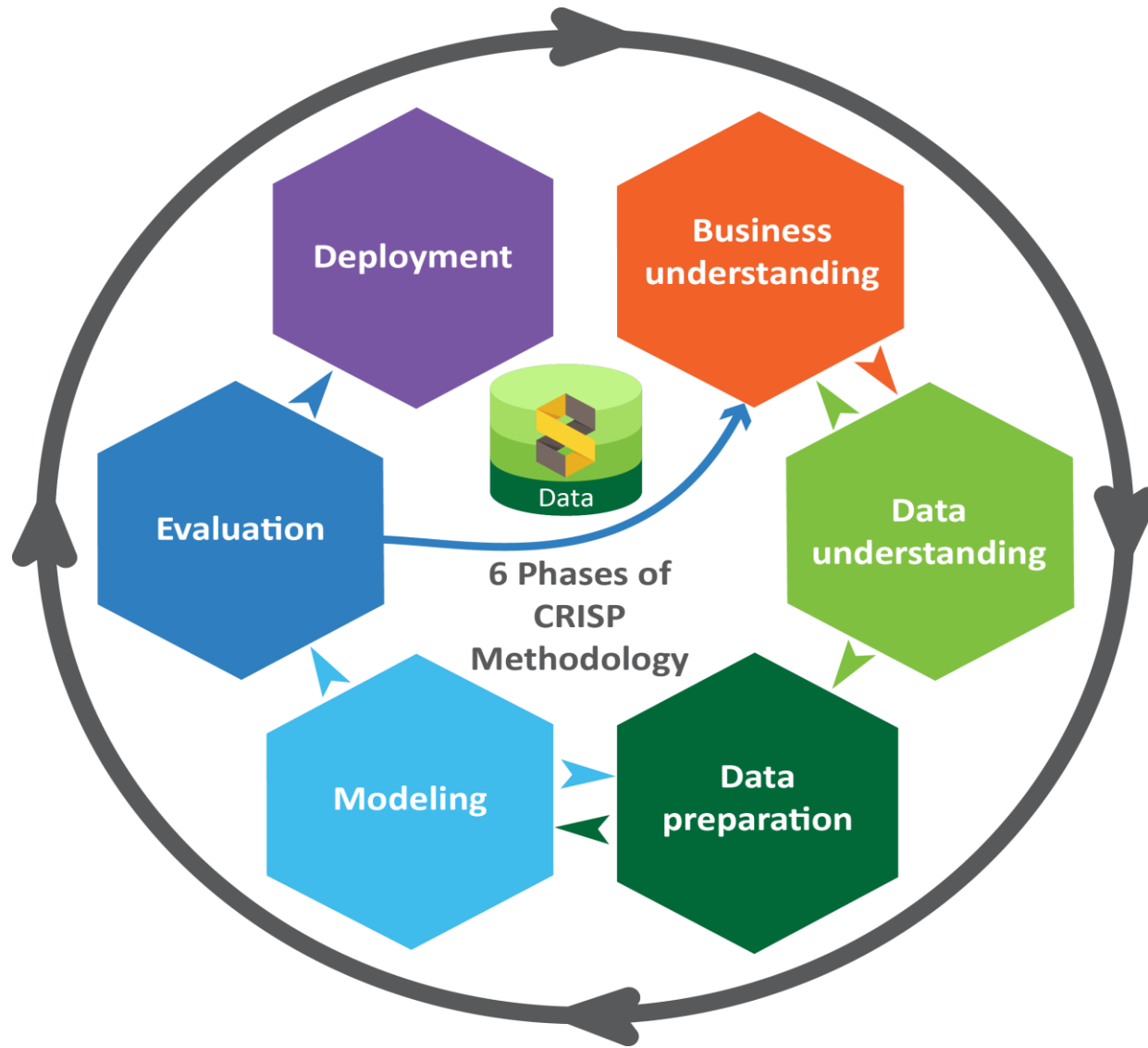


# Extraction de Connaissances à partir des Données



# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining



# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining

## 1 - Définition et compréhension du problème

- Comprendre les **objectifs** et les exigences du projet Data Mining.
- Indispensabilité de comprendre le domaine à explorer.
- Pour un meilleur choix de la technique, et donc de son résultat fiable.
- Meilleure explication et évaluation des résultats obtenus.



# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining

## 2 – Compréhension des données

- Les collecter et comprendre leur signification.
- Déterminer précisément les données à analyser et à identifier la qualité des données.
- Identifier et **sélectionner** les données à utiliser selon le problème défini précédemment.
- Ces données n'ont pas toujours le même format, le même type, et la même structure. (textes, BDD, pages web, images, video, ...etc)





# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining

## 3 – Prétraitement des données



- Regroupe les activités liées à la construction de l'ensemble précis des données à analyser à partir des données brutes.
- Souvent, données bruitées (erreurs de frappe, erreurs système, ...), incohérentes, anomalies, etc.
- Inclut le classement des données en fonction des critères, le **nettoyage** des données, unification des intervalles, lissage, réduction, etc.
- Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données.

# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining

## 4 – Modélisation

- Différents **algorithmes** et techniques sont sélectionnés et appliqués.
- Leurs paramètres sont étalonnés aux valeurs optimales.
- Choisir la bonne **technique** pour extraire les connaissances.



## 5 – Evaluation

- Vise à **vérifier le modèle** ou les connaissances obtenues afin de s'assurer qu'ils répondent aux objectifs formulés au début du processus.
- Elle contribue aussi à la décision de déploiement du modèle, ou si besoin est, à son amélioration.

# Modèles de Processus de Data Mining

Standard **CRISP-DM** : Cross-Industry Standard Process for Data Mining

## 6 – Déploiement

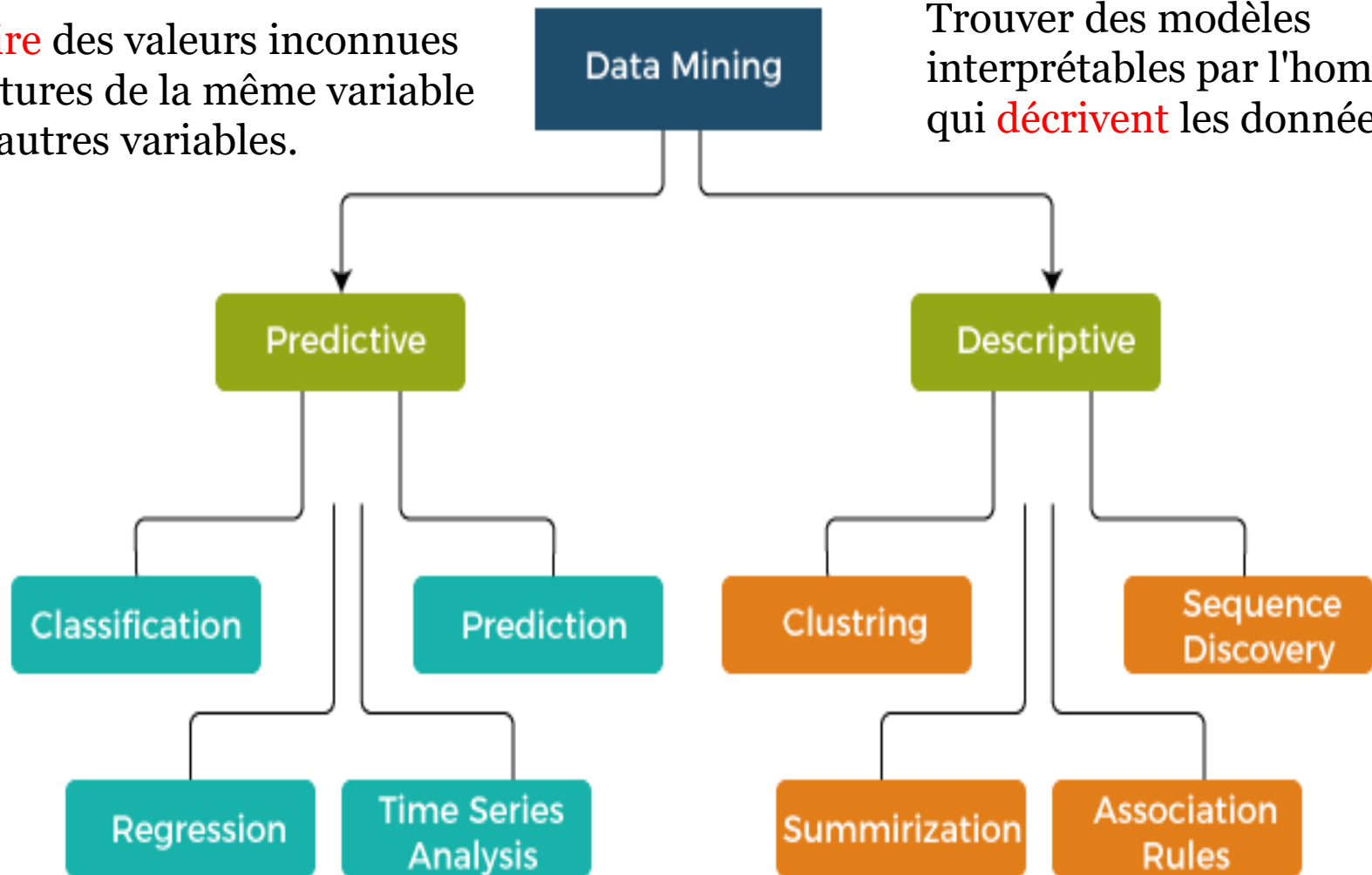


- Son objectif est de mettre la connaissance obtenue par la modélisation, dans une forme adaptée et l'intégrer au processus de prise de décision.
- Aider à la prise de décision en fournissant des modèles et des interprétations compréhensibles aux utilisateurs.
- Le déploiement peut aller, selon les objectifs, de la simple génération d'un **rapport** décrivant les connaissances obtenues jusqu'à la mise en place d'une **application**.

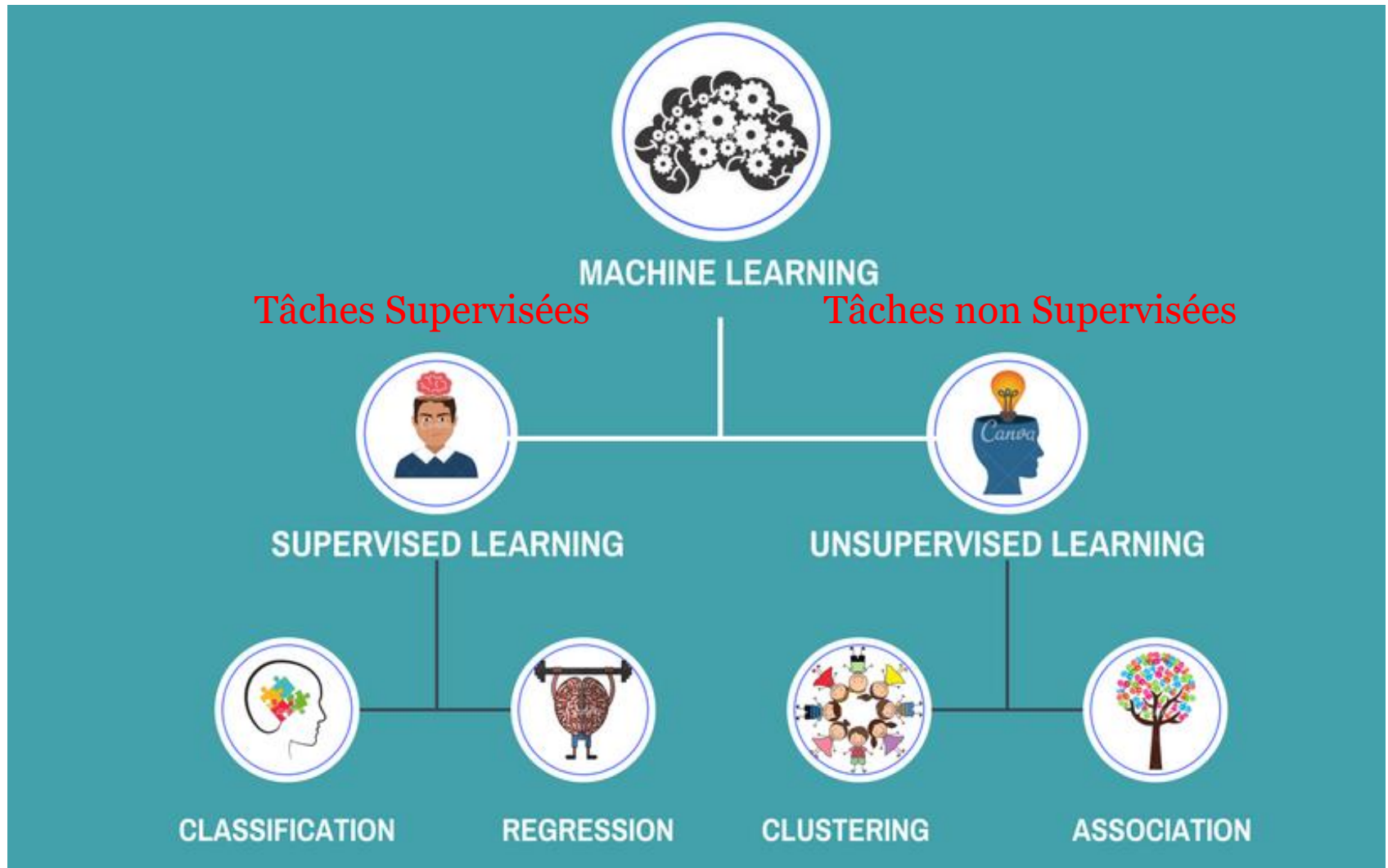
## Les tâches du Data Mining

**Prédire** des valeurs inconnues ou futures de la même variable ou d'autres variables.

Trouver des modèles interprétables par l'homme qui **décrivent** les données.

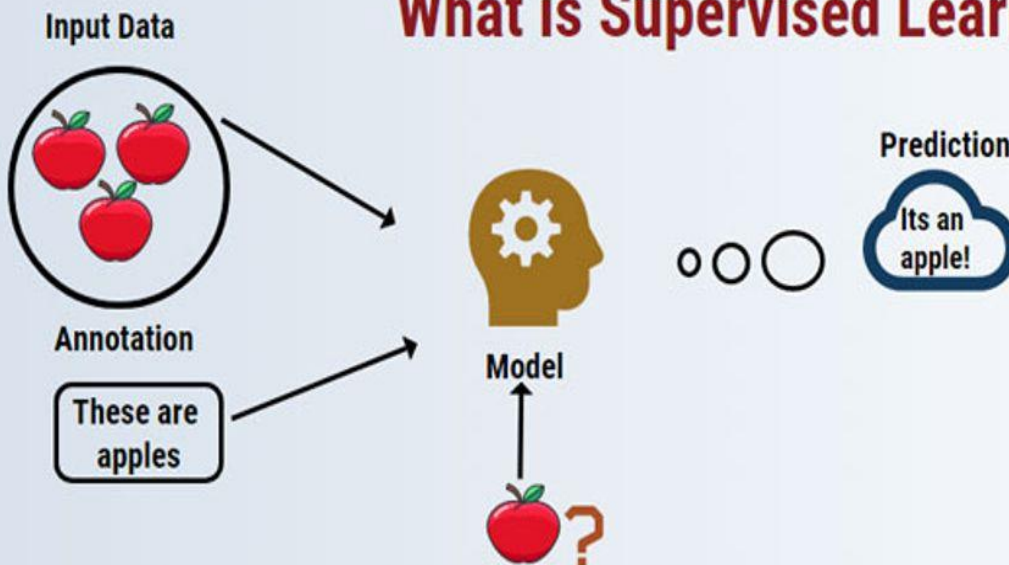


# Les tâches du Data Mining

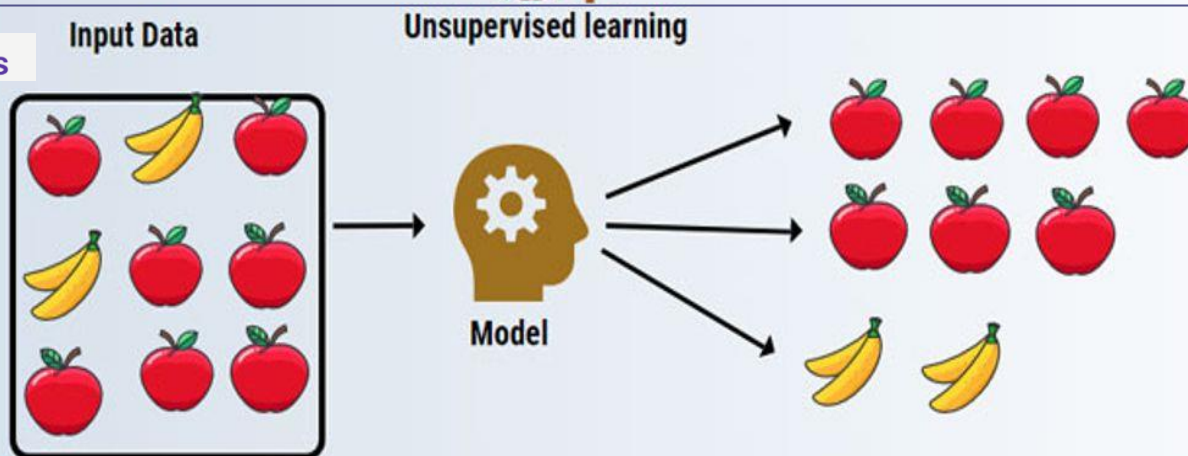


# Les tâches du Data Mining

## What is Supervised Learning?



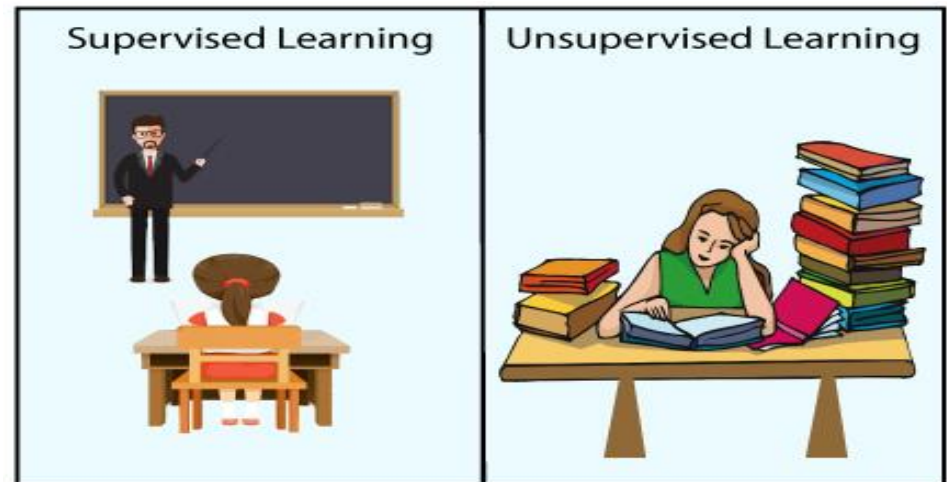
No labels



## Les tâches du Data Mining

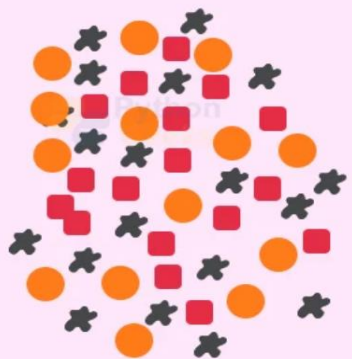
- **Règles d'association**: Consiste à déterminer quels items/attributs "vont ensemble". Les règles d'associations sont de la forme "Si <antécédent>, alors <conséquent>".
- **Classification**: Consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie.
- **Régression**: Similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique.
- **Clustering**: Regroupement en groupes d'objets similaires. Segmenter la totalité de données en des sous groupes relativement homogènes. Maximiser l'homogénéité à l'intérieur de chaque groupe et la minimiser entre les différents groupes.

## Les tâches du Data Mining



## Reinforcement Learning

Input Raw Data



Environment

Reward

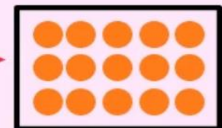
Best Action

State

Selection of Algorithm

Agent

Output

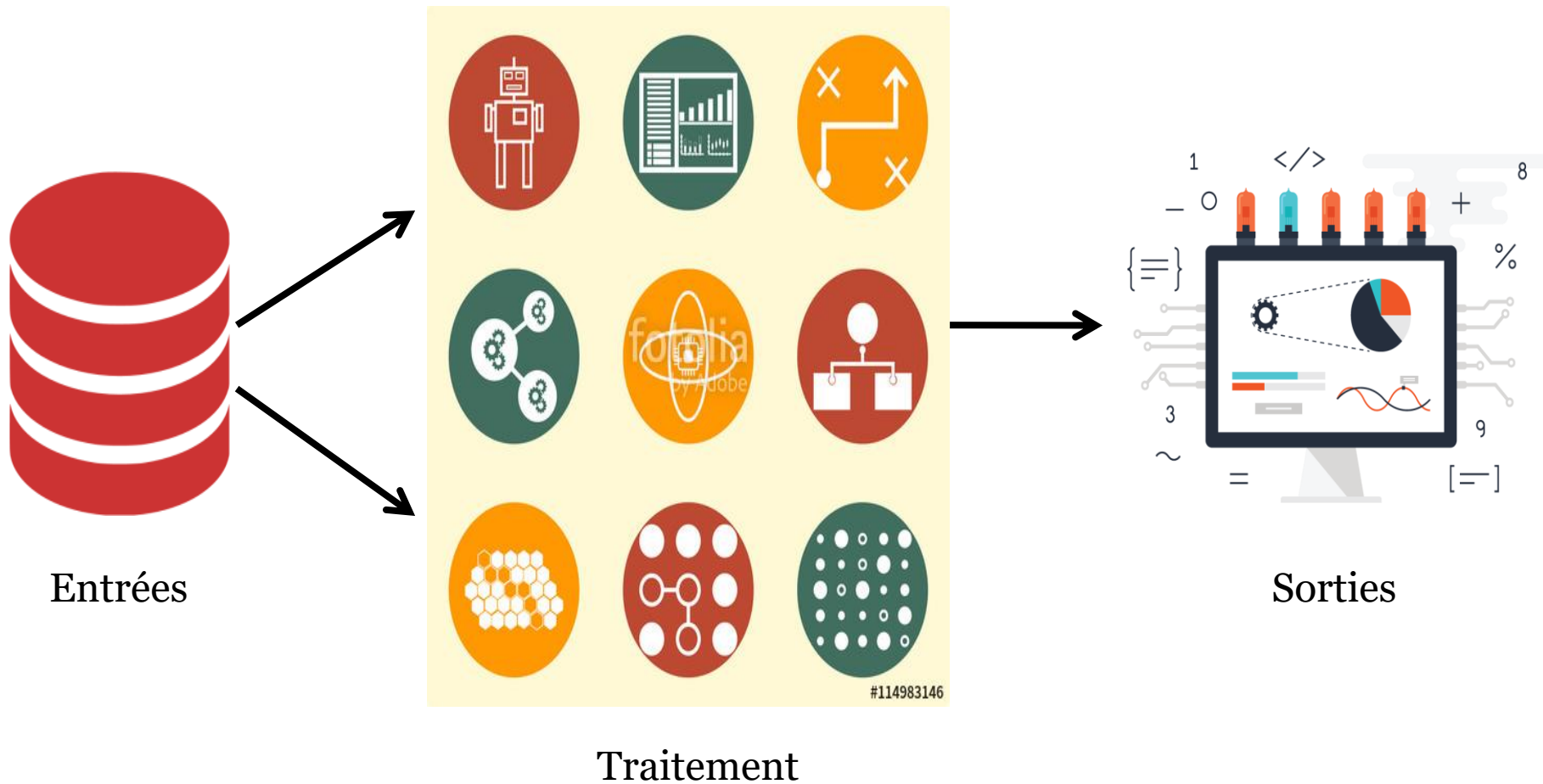




# Fouille de Données - Data Mining

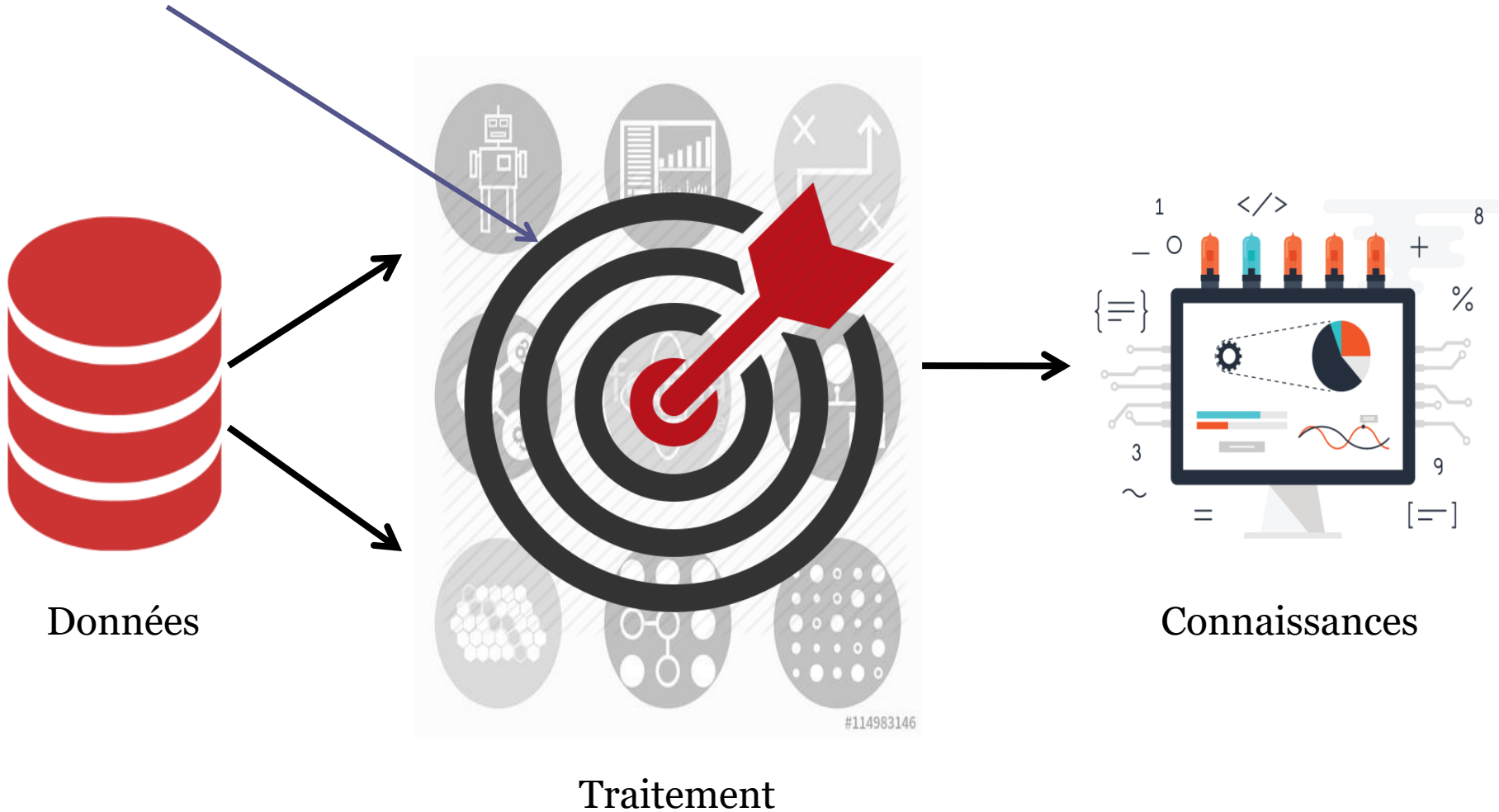
Objectifs de la matière

**SAVOIR – PREDIRE/DECRIRE - DECIDER**



# Fouille de Données - Data Mining

## Objectifs de la matière



# Fouille de Données - Data Mining

## Objectifs de la matière

La fouille de données vise à découvrir, dans les grandes quantités de données, les informations importantes qui peuvent aider à comprendre les données ou à prédire le comportement des données futures.

Le but de ce cours est d'initier les apprenants aux différents algorithmes et techniques utilisés en fouille de données.

**Pré requis** recommandés : Connaissances: *algorithmique, algèbre linéaire*

**Unité d'enseignement** : UEM21

**Crédit** : 5

**Coefficient** : 3

**Mode d'évaluation** : Examen, contrôle continu TD et TP.

# Fouille de Données - Data Mining

## Contenu de la matière :

- 1) **Introduction générale**
- 2) Recherche des modèles fréquents et des règles d'associations
- 3) Classification
- 4) Régression
- 5) Clustering



# Fouille de Données - Data Mining

## Contenu de la matière :



TP : Suite de logiciels libres d'apprentissage et de data mining **Weka et Python.**

**Weka** : pour Waikato Environment for Knowledge Analysis.

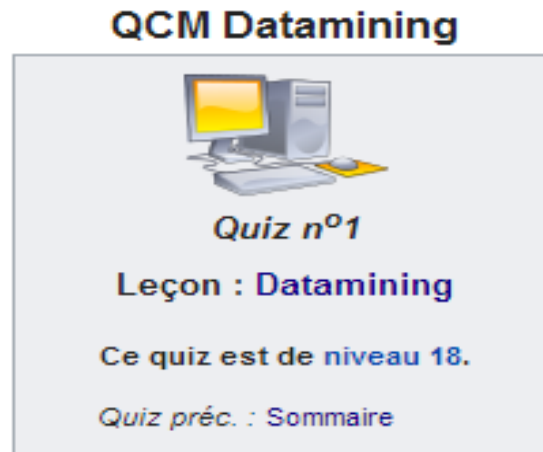


Un projet open source qui implémente plusieurs techniques de fouille de données et de prétraitement, issues de la communauté apprentissage automatique. + API Java

**Scikit-learn** est une bibliothèque libre Python destinée à l'apprentissage automatique.

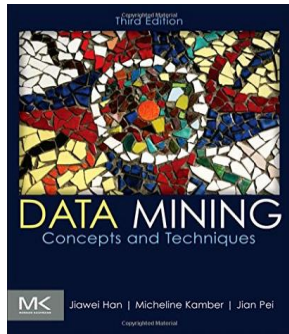
# Quiz

Pour aller plus loin et tester ses acquis :



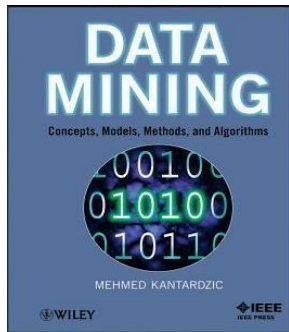
[https://fr.wikiversity.org/wiki/Datamining/Quiz/QCM\\_Datamining](https://fr.wikiversity.org/wiki/Datamining/Quiz/QCM_Datamining)

# Références



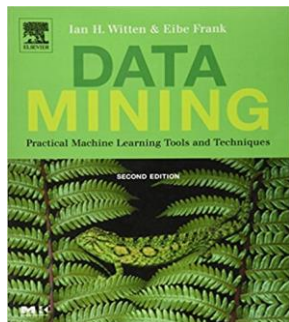
## **Data Mining : concepts and techniques, 3rd Edition**

- ✓ Auteur : Jiawei Han, Micheline Kamber, Jian Pei
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2011 - 744 pages - ISBN 9780123814807



## **Data Mining : concepts, models, methods, and algorithms**

- ✓ Auteur : Mehmed Kantardzi
- ✓ Éditeur : John Wiley & Sons
- ✓ Edition : Aout 2011 – 552 pages - ISBN : 9781118029121



## **Data Mining: Practical Machine Learning Tools and Techniques**

- ✓ Auteur : Ian H. Witten & Eibe Frank
- ✓ Éditeur : Morgan Kaufmann Publishers
- ✓ Edition : Juin 2005 - 664 pages - ISBN : 0-12-088407-0

# Références

Cours – Abdelhamid DJEFFAL – Fouille de données avancée

✓ [www.abdelhamid-djeffal.net](http://www.abdelhamid-djeffal.net)

WekaMOOC – Ian Witten – Data Mining with Weka

✓ <https://www.youtube.com/user/WekaMOOC/featured>

Cours - Laboratoire ERIC Lyon - DATA MINING et DATA SCIENCE

✓ [https://eric.univ-lyon2.fr/~ricco/cours/supports\\_data\\_mining.html](https://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html)

Gregory Piatetsky-Shapiro - KDNuggets

✓ <http://www.kdnuggets.com/>