

DATA MINING

FOUILLE DE DONNÉES

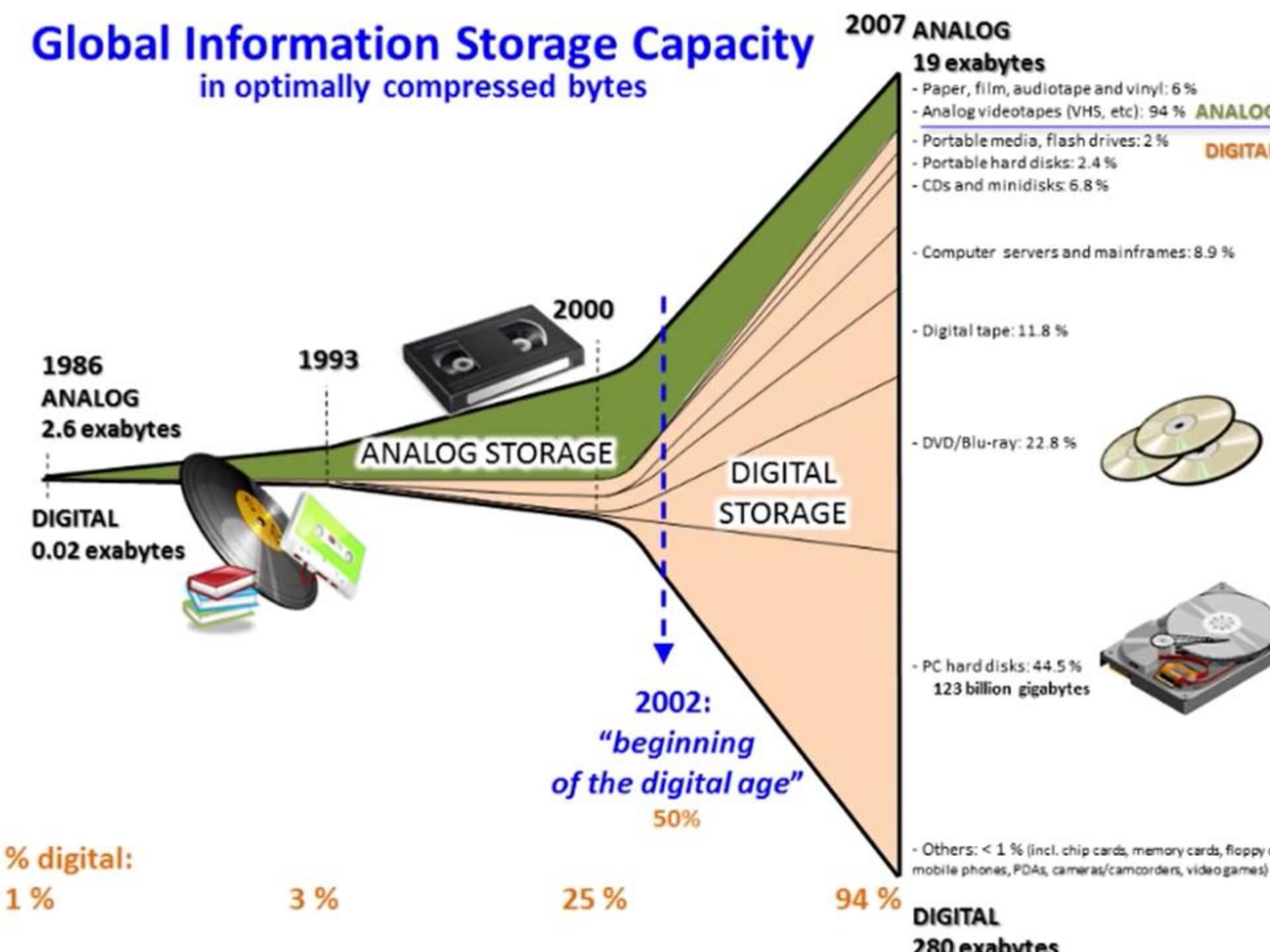
**4 ERP-BI
2022-2023**

Mohamed Hery SELMI- Wiem Trabelsi

- **Les données ? Les sources? Les types ?**
- **Le Big Data**
- **Le Business Intelligence**
- **Analyser, décrire, résumer, prédire, ...Décider ;)**
- **Le Data Mining :**
 - Définition
 - Aspect pluridisciplinaire
 - Applications
 - Finance, Marketing,
 - méthodes et algorithmes
 - Processus et méthodologies
 - À retenir !

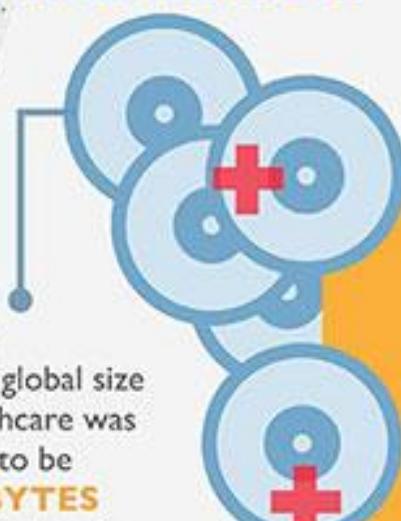
Global Information Storage Capacity

in optimally compressed bytes



By 2014, it's anticipated there will be

**420 MILLION WEARABLE, WIRELESS
HEALTH MONITORS**



As of 2011, the global size
of data in healthcare was
estimated to be
150 EXABYTES
(161 Billion Gigabytes)

Variety

Different
Forms of Data

**4 BILLION+
HOURS OF
VIDEO**

are watched on
YouTube each month



**30 BILLION PIECES OF
CONTENT**

are shared on Facebook every month



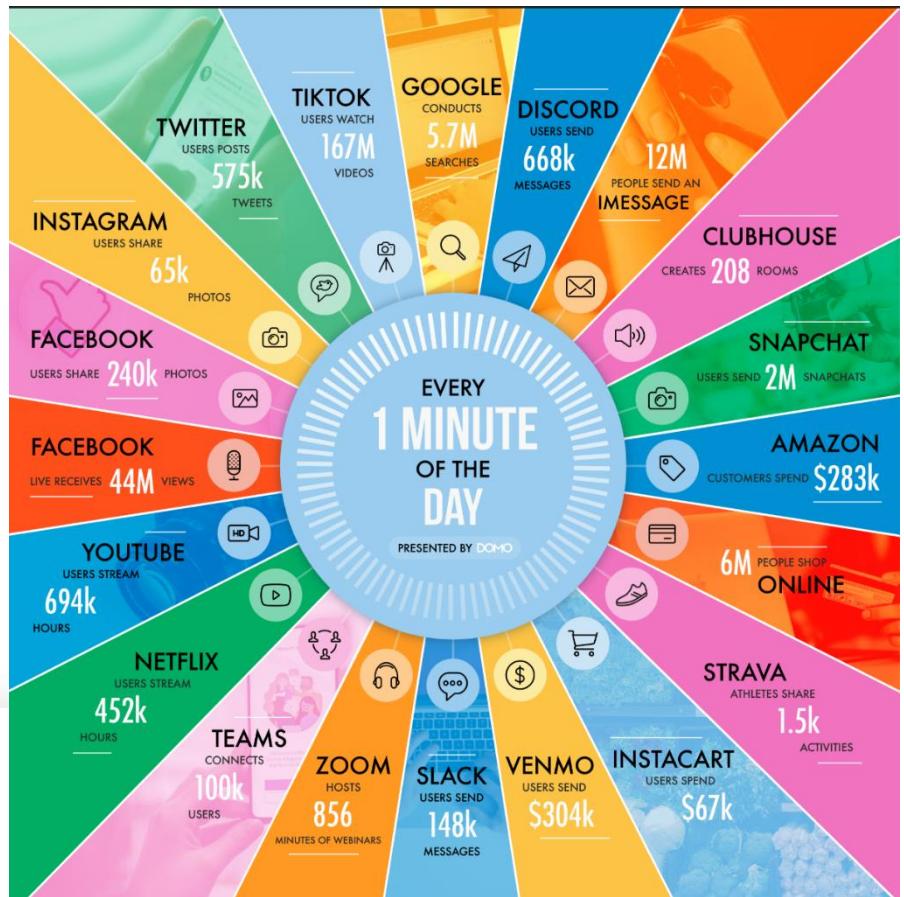
**4 MILLION
TWEETS**

are sent per day by
about 200 million
monthly active users

2017 This Is What Happens In An Internet Minute



2022 One Internet Minute



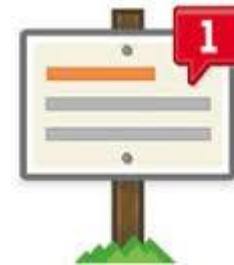
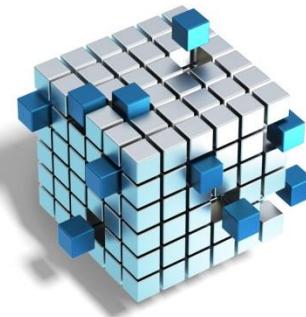
SOURCE <https://www.abondance.com/20221125-50288-infographie-que-se-passe-t-il-sur-internet-en-1-minute-en-2022-et-2021.html>



DIVERSITÉ DE DONNÉES

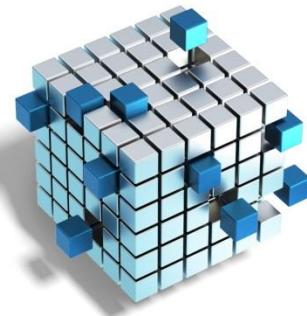


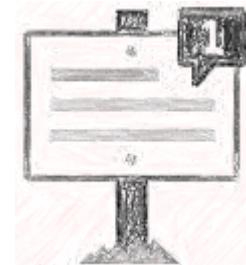
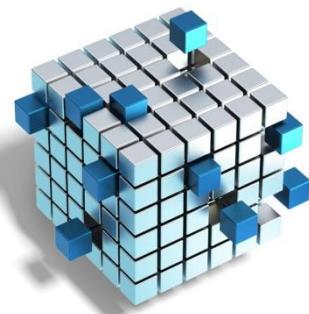
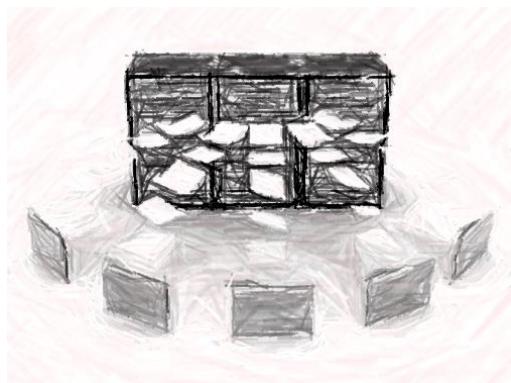
LES APPROCHES « BUSINESS INTELLIGENCE »

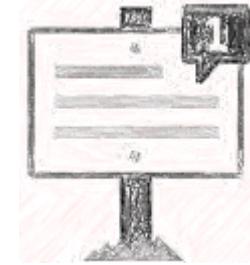
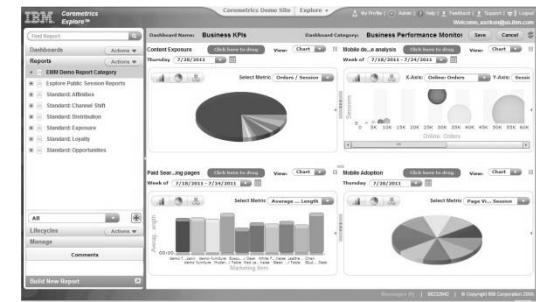
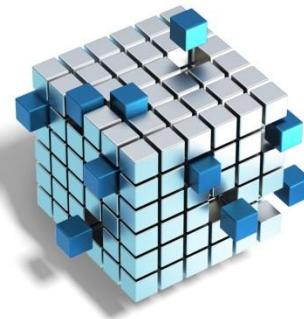




Et si on veut aller plus loin :
Décider !

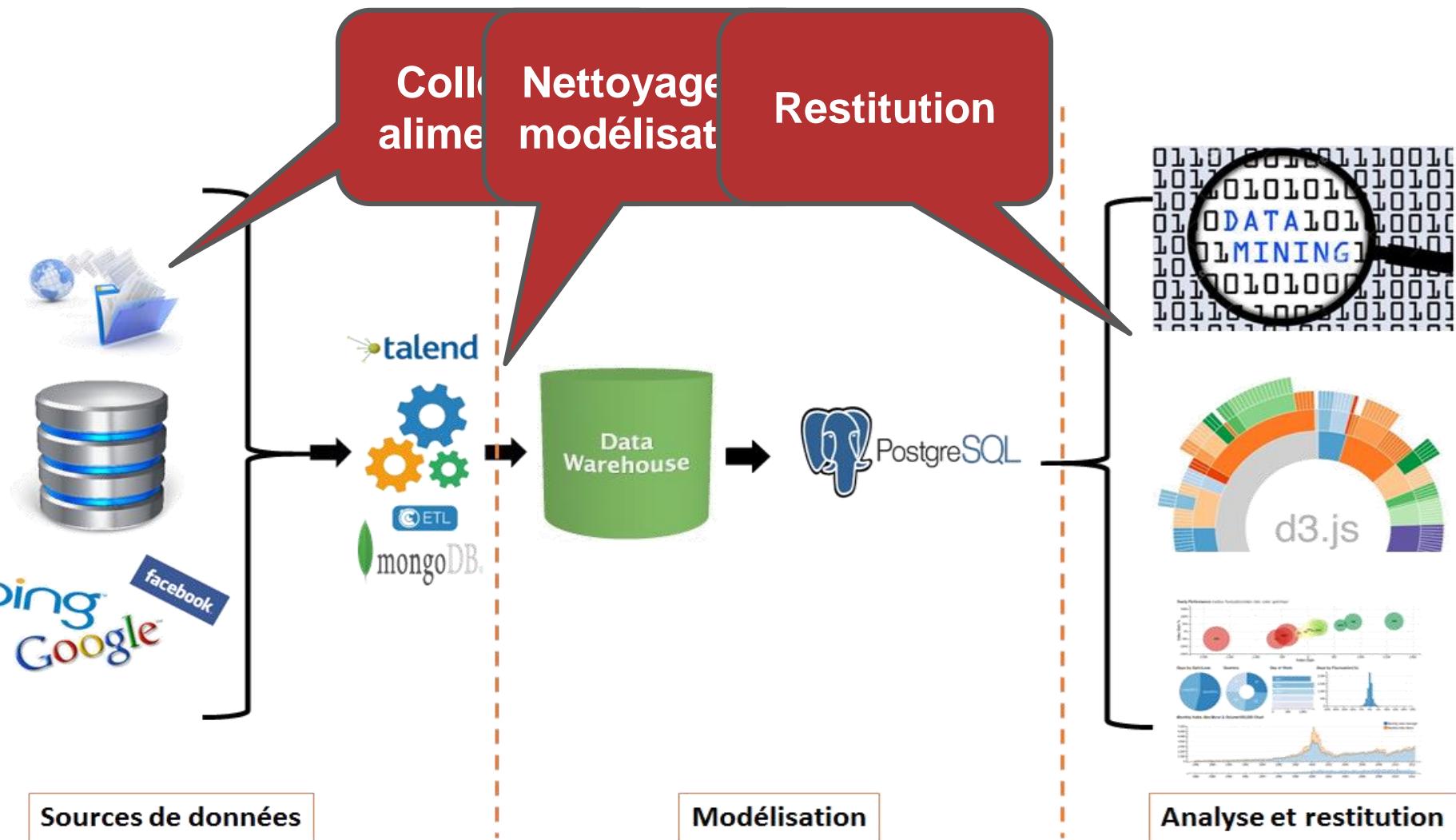






Solution :
Fouiller dans les données

CHAINE DÉCISIONNELLE



DÉFINITION



C'est l'ensemble des algorithmes, méthodes et technologies inspirés de plusieurs autres disciplines, propres ou non au DM pouvant servir à remplacer ou à aider l'expert humain ou le décideur dans un domaine spécifique dans le cadre de prise de décision, et ce en fouillant dans des bases de données décisionnelles des corrélations, des associations, des comportements homogènes, des formules de lien entre indicateurs, des spécification par rapport à une thématique bien déterminée, etc.

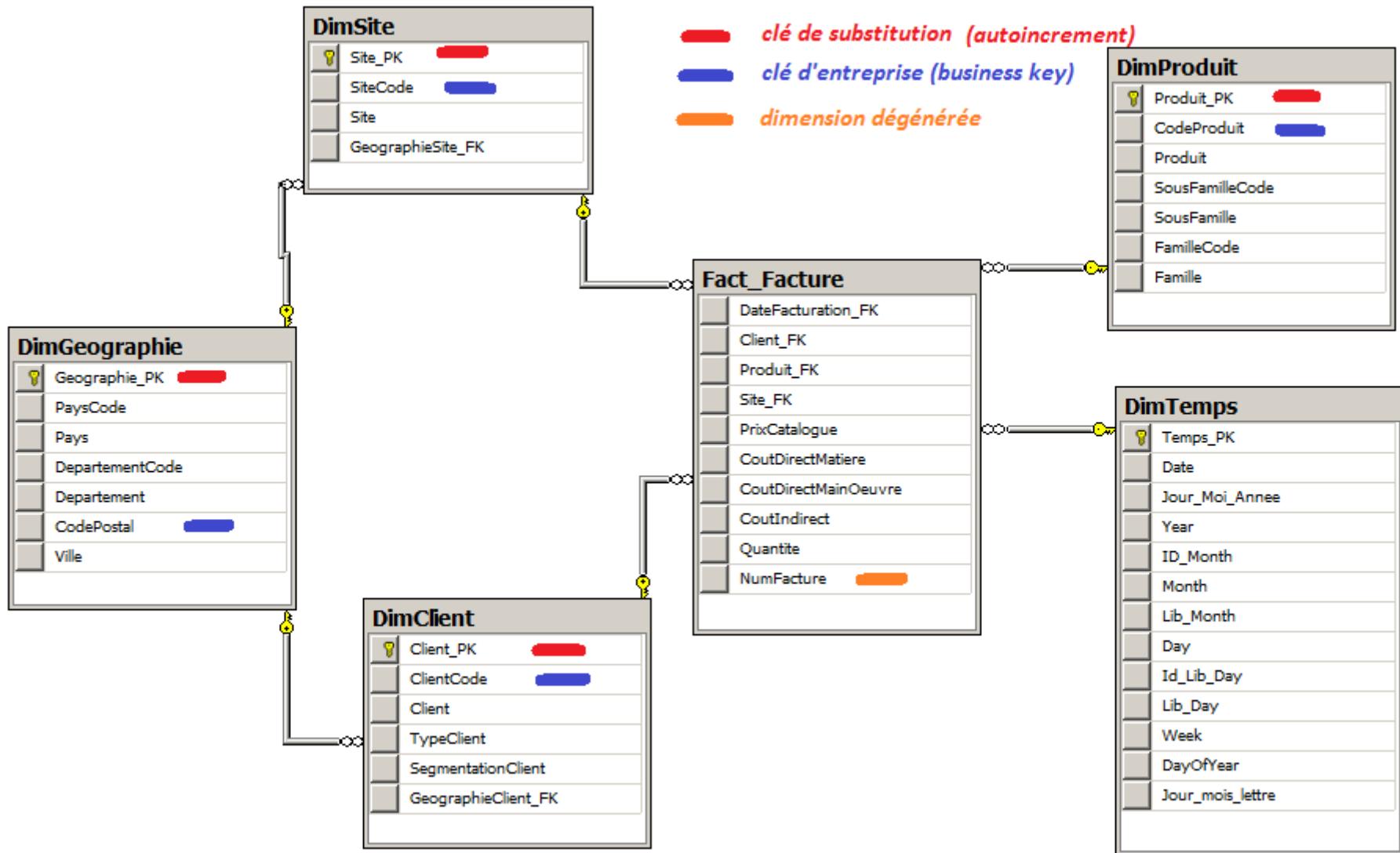
Cet ensemble de techniques peut être une étape fondamentale dans tout processus ECD (KDD) ou bien l'intervention de l'intelligence artificielle, la reconnaissance de forme, et la statistique décisionnelle dans tout processus de Business Intelligence afin de transformer tout système d'aide à la décision en un système DECISIONNEL au vrai sens du mot.

ENJEUX



- Simplifier la production de données ou informations structurées et porteuses de sens
- Créer du sens et des connaissances à partir de données non enrichies et non structurées
- Analyser des tendances sur la durée
- Permettre la création de modèles sur des historiques de données
- Analyse prédictive

STRUCTURATION DE DONNÉES



ENTREPÔT DE DONNÉES

	A	B	C	D	E	F	G	H
1	age	sexe	typedouleur	sucré	tauxmax	angine	depression	coeur
2	70	masculin	D	A	109	non	24	présence
3	67	feminin	C	A	160	non	16	absence
4	57	masculin	B	A	141	non	3	présence
5	64	masculin	D	A	105	oui	2	absence
6	74	feminin	B	A	121	oui	2	absence
7	65	masculin	D	A	140	non	4	absence
8	56	masculin	C	B	142	oui	6	présence
9	59	masculin	D	A	142	oui	12	présence
10	60	masculin	D	A	170	non	12	présence
11	63	feminin	D	A	154	non	40	présence
12	59	masculin	D	A	161	non	5	absence
13	53	masculin	D	A	111	oui	0	absence
14	44	masculin	C	A	180	non	0	absence
15	61	masculin	A	A	145	non	26	présence
16	57	feminin	D	A	159	non	0	absence
17	71	feminin	D	A	125	non	16	absence
18	46	masculin	D	A	120	oui	18	présence
19	53	masculin	D	B	155	oui	31	présence
20	64	masculin	A	A	144	oui	18	absence

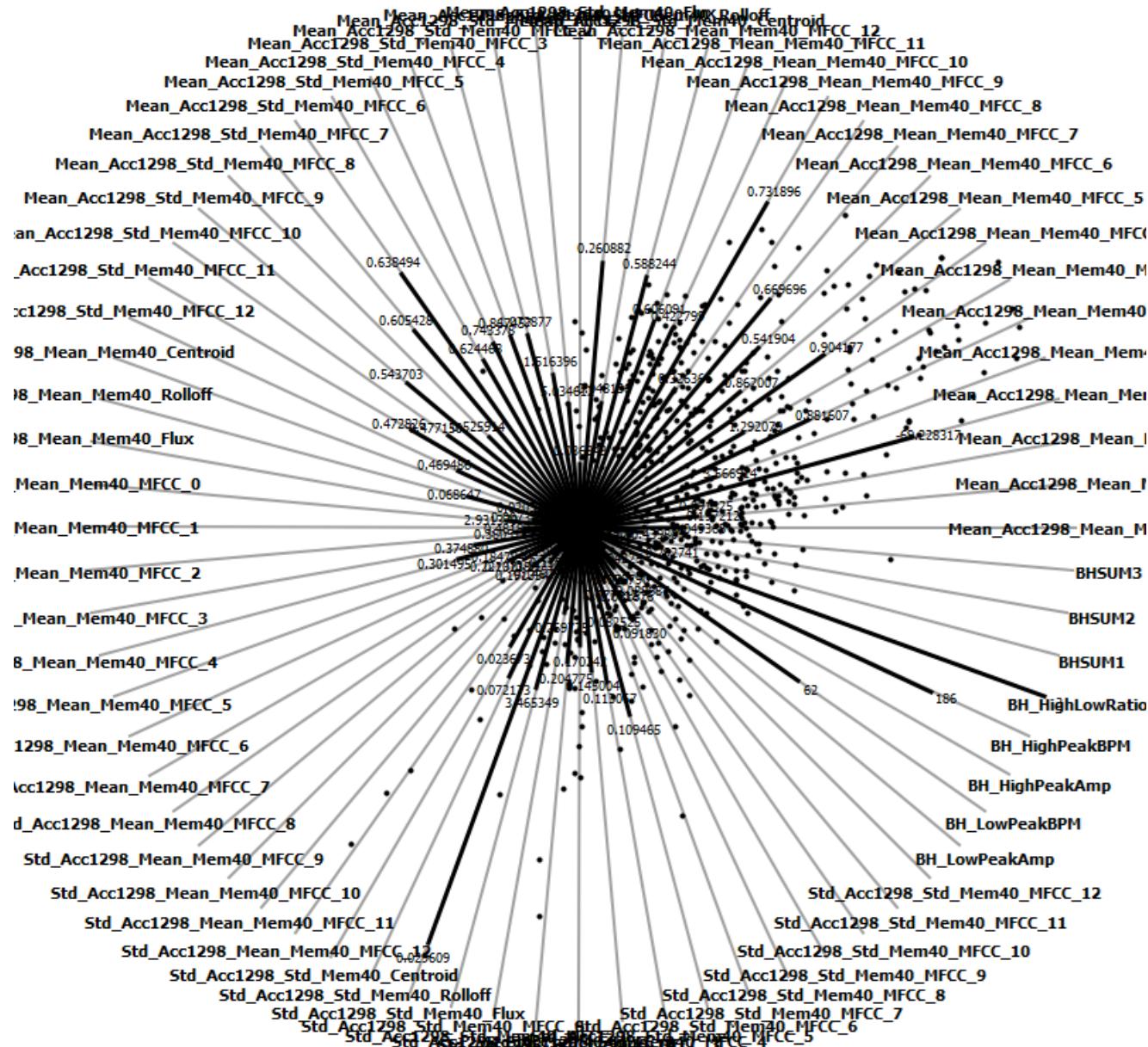
Axes d'analyse

Mesures

Clef de substitution

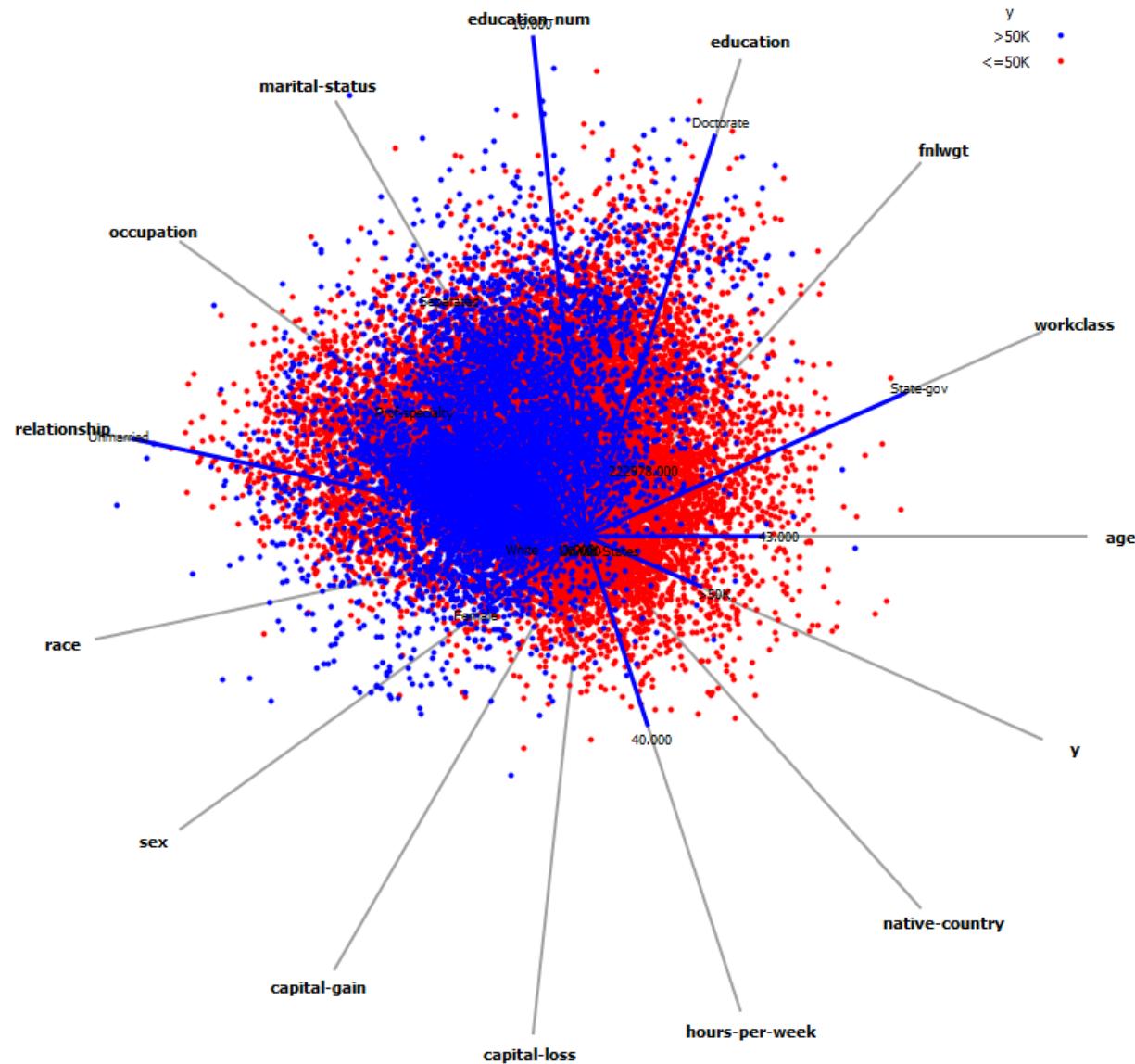
NUAGE DE POINTS

VARIABLES QUANTITATIVES

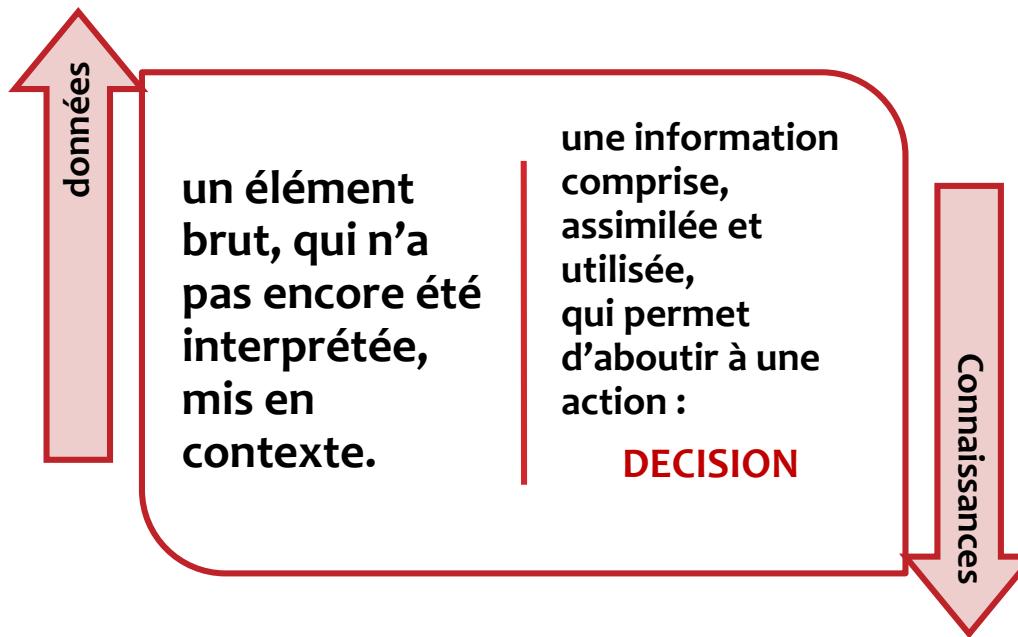


NUAGE DE POINTS

VARIABLES QUALITATIVES



DONNÉES, INFORMATIONS, CONNAISSANCES



client	M	A	R	E	I
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

Donnée :

Client 3 : âge = âgé, Niveau d'études = non

Information :

37,5 % des Clients consultent leurs comptes bancaires sur le Web.

Connaissance :

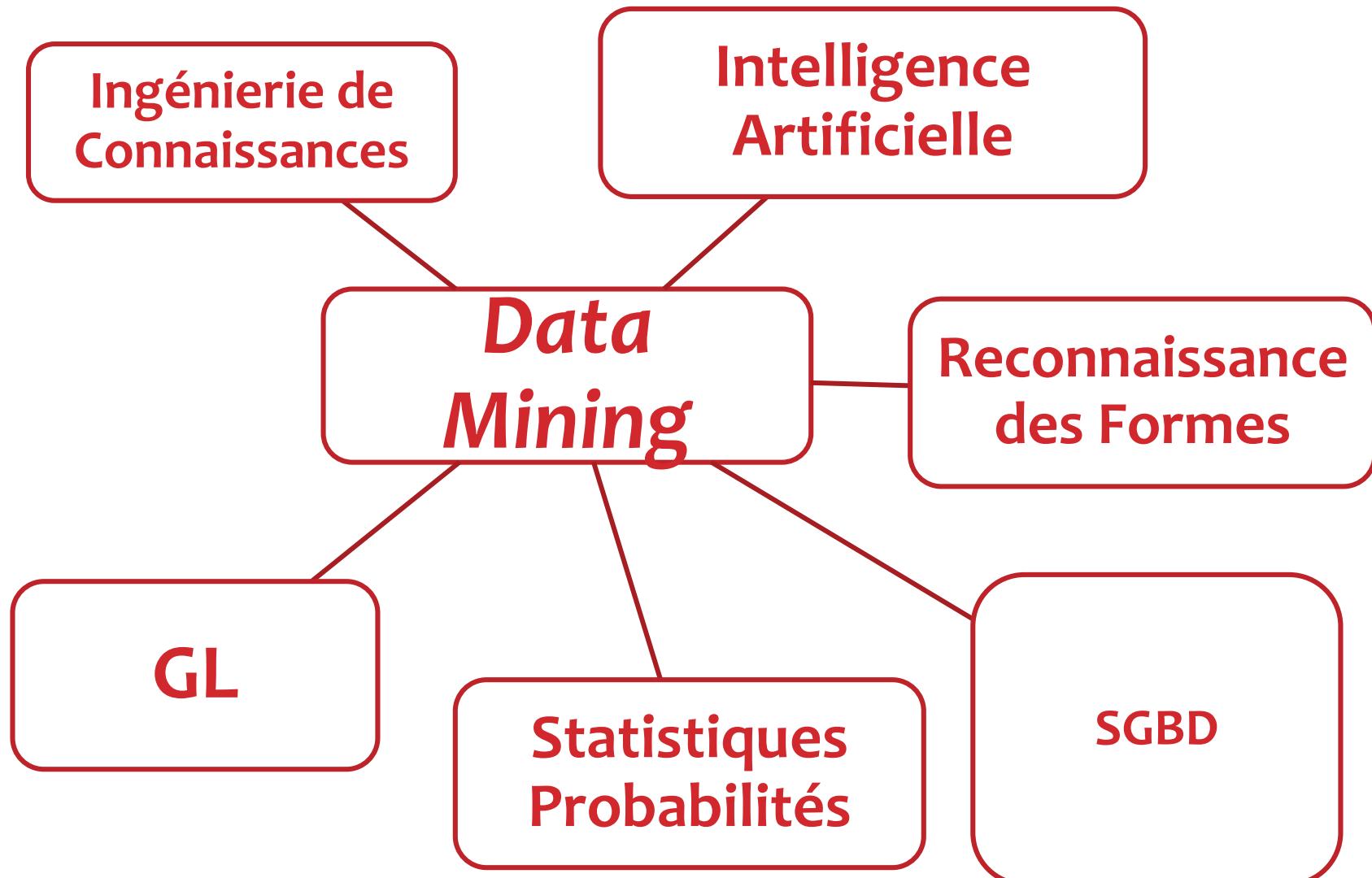
SI E=non ALORS I= non

SI E=oui ET A=moyen ALORS I=oui

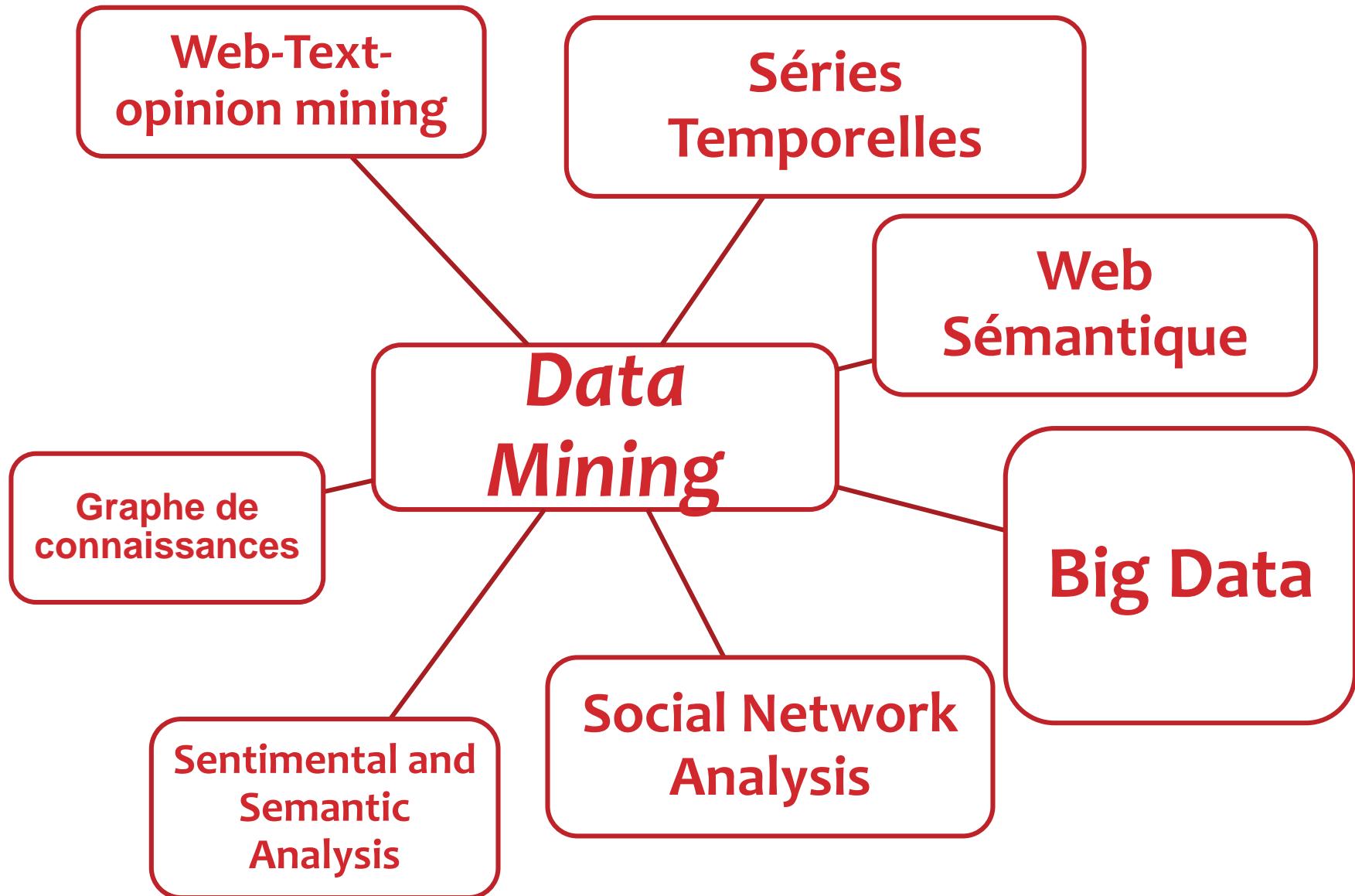
SI E=oui ET A=âgé ALORS I=non

SI A=oui ET A=jeune ALORS I=oui

ASPECT PLURIDISCIPLINAIRE

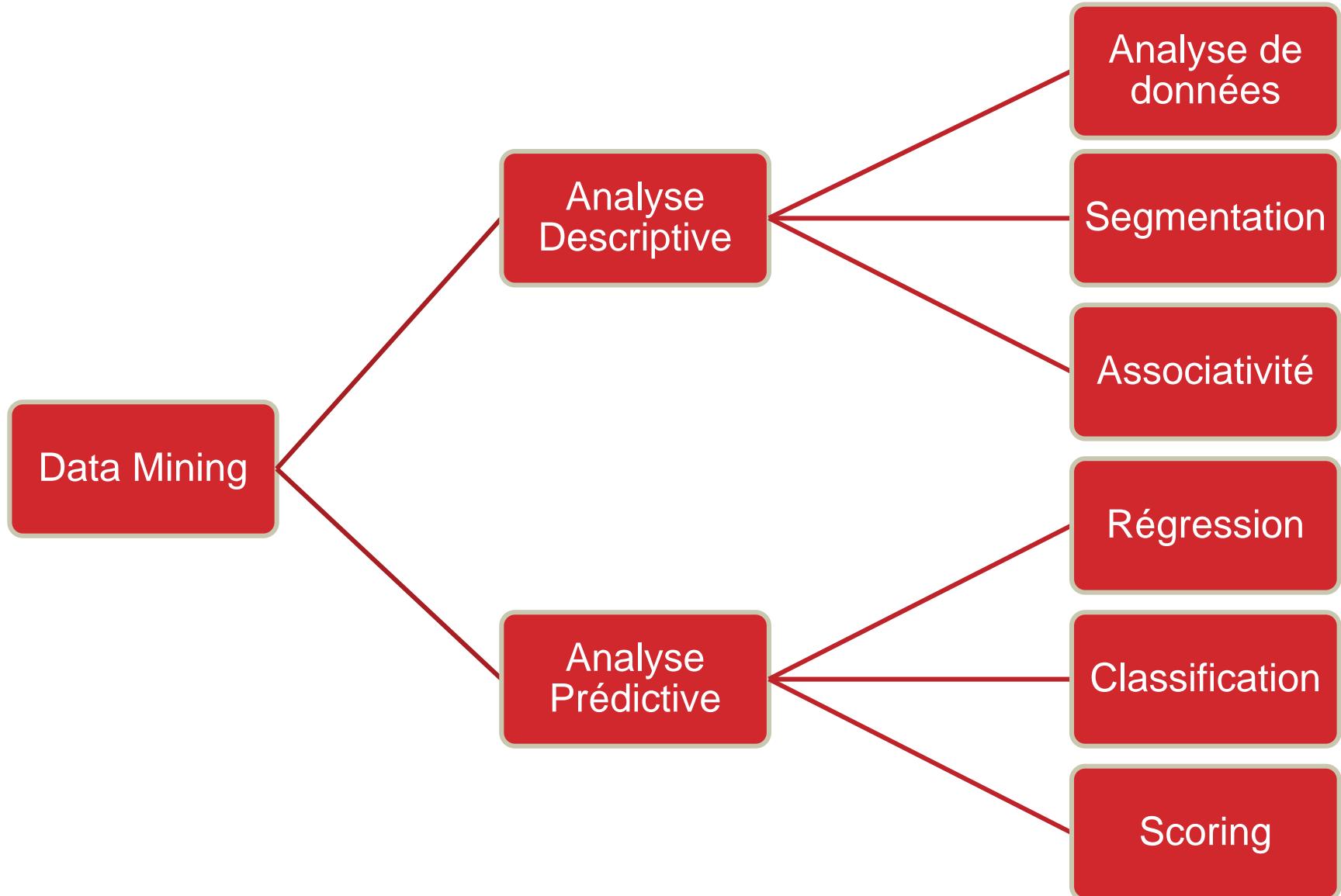


EXTENSION DU DATA MINING

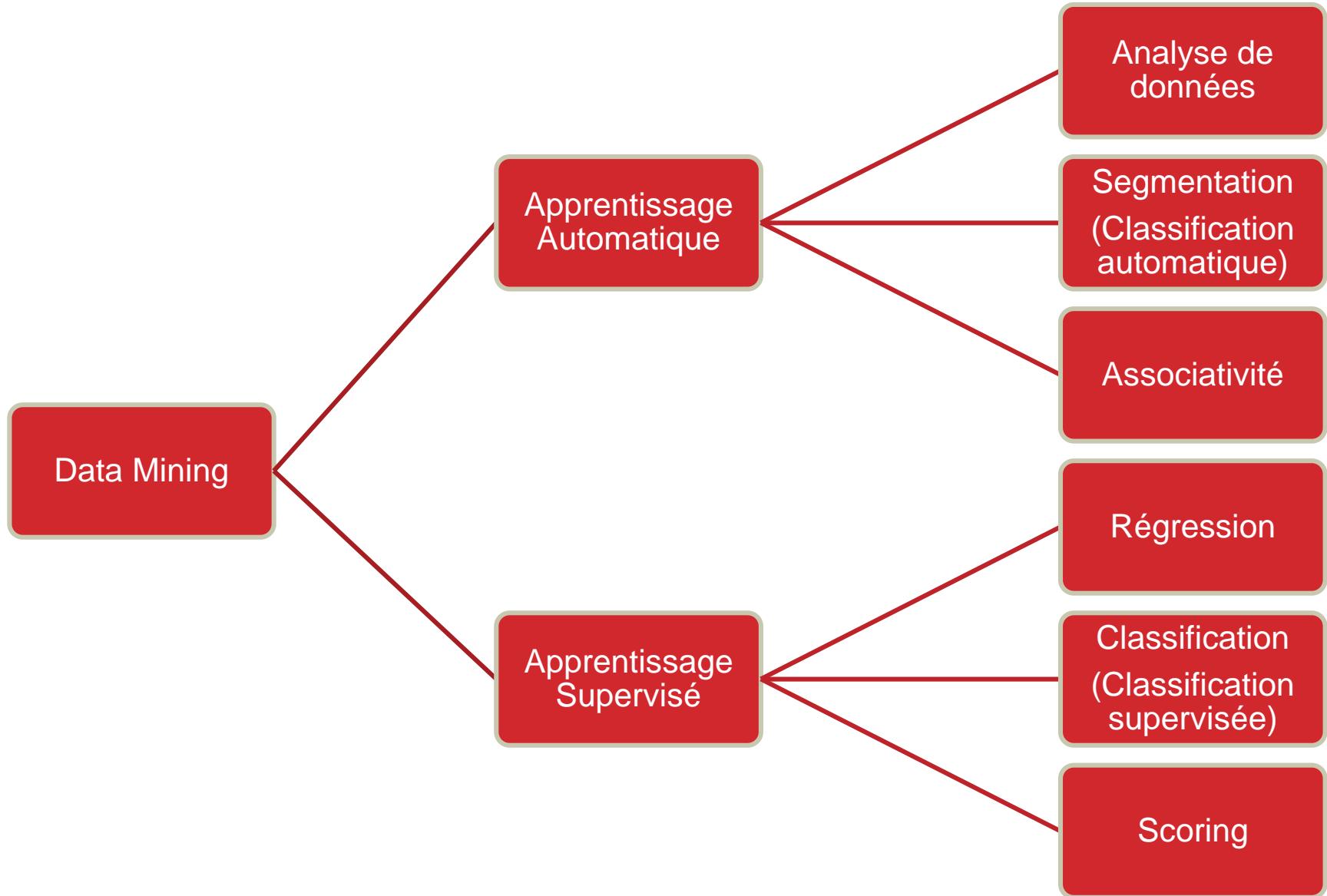


APPLICATIONS DU DATA MINING

APPLICATIONS DE BASE DU DATA MINING



APPLICATIONS DE BASE DU DATA MINING



APPLICATIONS DE BASE DU DATA MINING

Analyse
factorielle

Segmentation
Clustering

Classification

Scoring

Associativité

Régression

REPORTING ANALYTIQUE

Résumer et visualiser l'activité
de l'entreprise à travers
des *indicateurs pertinents*



APPLICATIONS DE BASE DU DATA MINING

Analyse
factorielle

Segmentation
Clustering

Classification

Scoring

Associativité

Régression

SEGMENTATION

Aucune variable décisionnelle, information quantitative



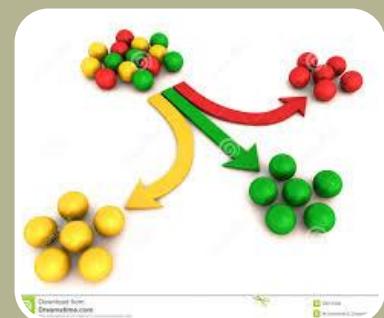
- Les variables d'entrées servent à créer des groupes homogènes
- Les individus de chaque groupe se ressemblent le plus
- Les groupes d'appartenances obtenus se distinguent le plus

La Segmentation a pour objectifs :



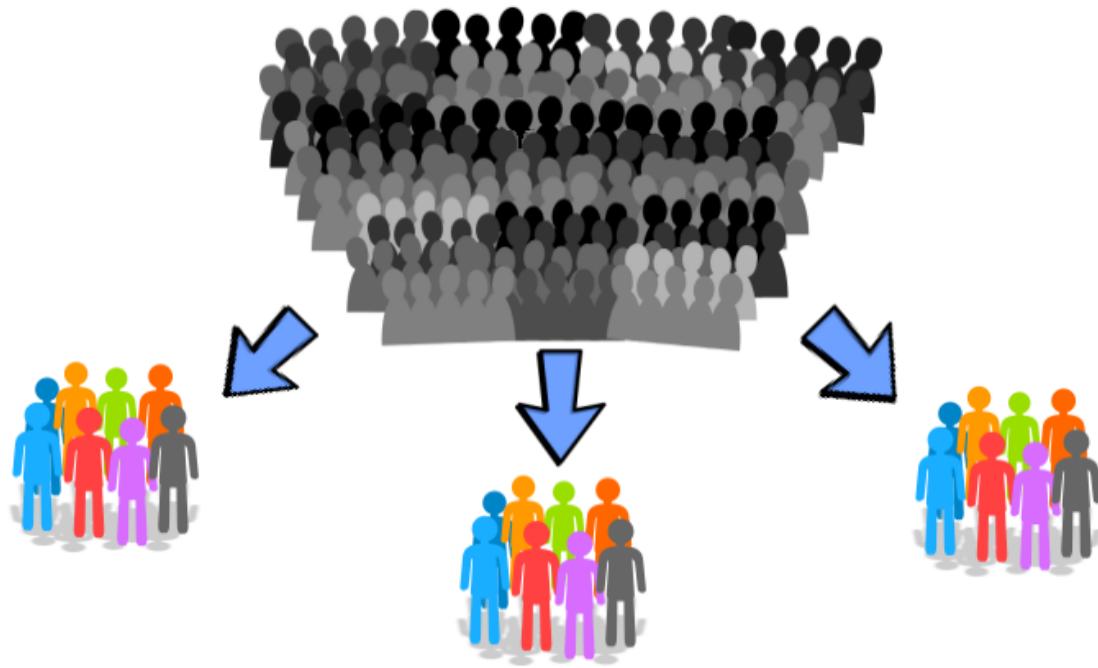
- Trouver les variables métiers influençant la répartition en groupes
- Affecter les individus à leurs nouveaux groupes d'appartenance

Plusieurs méthodes et techniques pour segmenter :



- Partitionnement : k-means
- Hiérarchique : CAH

SEGMENTATION – TYPOLOGIE



Trouver les **comportements typiques** des clients.

APPLICATIONS DE BASE DU DATA MINING

Analyse
factorielle

Segmentation
Clustering

Classification

Scoring

Associativité

Régression

ANALYSE DU CHARIOT



Recherche des articles
les plus/moins **associés**

MÉDECINE

LES SYMPTÔMES ASSOCIÉS



Troubles du sommeil



Fatigue



Dépression



Anxiété



Isolation social



Troubles de l'humeur

APPLICATIONS DE BASE DU DATA MINING

Analyse

Segmentation
Clustering

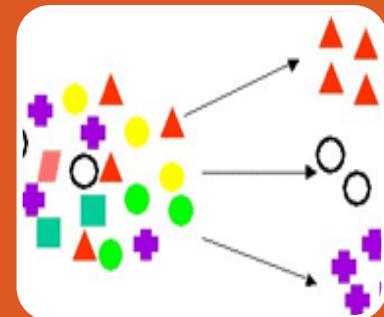
Classification

Scoring

Associativité

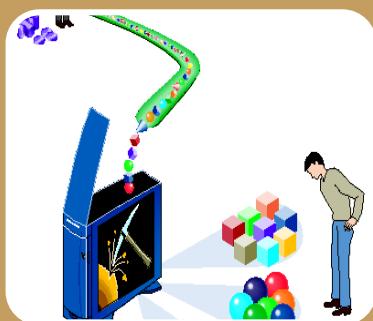
Régression

CLASSIFICATION



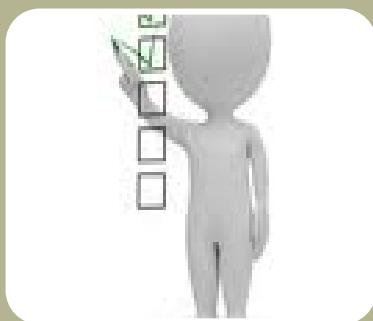
La variable décisionnelle est **qualitative**

- Un dossier de crédit peut être classifié : BON ou MAUVAIS
- Un patient peut présenter un fort risque de maladie cardiaque
- Un client peut présenter un comportement atypique



La Classification a pour objectifs :

- Déetecter les variables possédant un lien fort avec la variable décisionnelle
- Construire un modèle de classification liant ces variables à la décision



Plusieurs méthodes et techniques pour classifier :

- Arbre de décision
- Forêts Aléatoires
- K-NN k-nearest neighbors
- Séparateur à vaste Marge SVM
- Régression Logistique

CLASSIFICATION – PROFILING

Déterminer ce qui
caractérise un groupe
particulier de clients



APPLICATIONS DE BASE DU DATA MINING

Analyse

Segmentation
Clustering

Classification

Associativité

Régression

Scoring

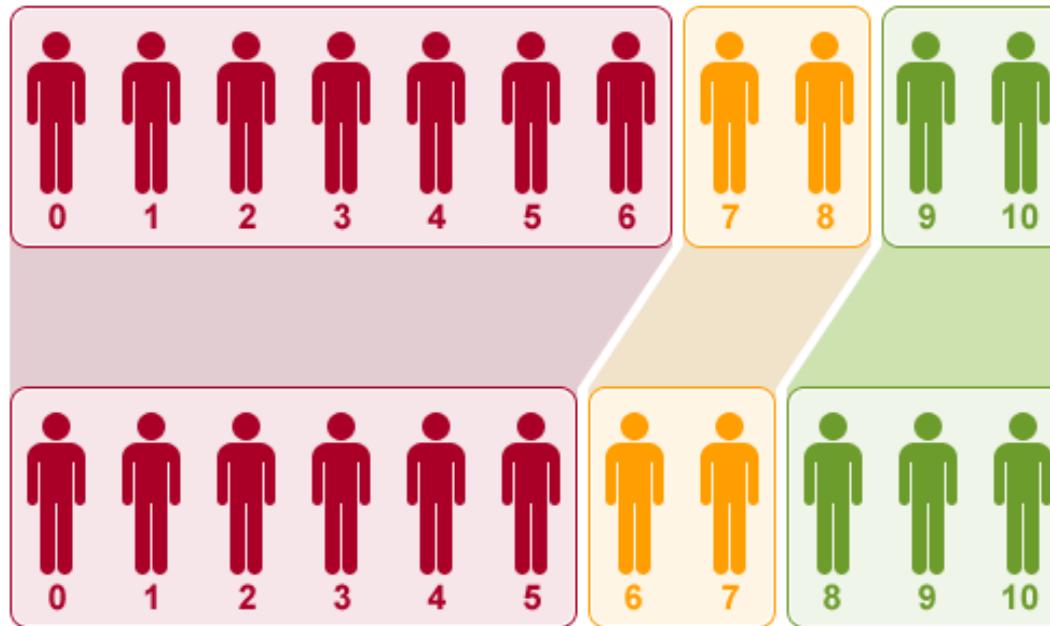
EVALUATION DE CAMPAGNES DE TERRAIN

Déterminer *l'efficacité de la communication* avec les clients.



MARKETING CIBLÉ

Optimiser les **chances d'obtenir des réponses (positives)** de la part des clients à une offre particulière par un ciblage plus précis, mettant en évidence les clients avec une forte probabilité de réponse.



DATA MINING ET CRM

Le datamining a beaucoup d'applications qui peuvent apporter un soutien considérable au marketing et à la gestion de la relation client



- faire correspondre des profils des clients avec des «offres produits» afin d'augmenter le taux de conversion : cross selling

APPLICATIONS DE BASE DU DATA MINING

Analyse

Segmentation
Clustering

Classement
Classification

Scoring

Associativité

Régression

RÉGRESSION LINÉAIRE



La variable décisionnelle est **quantitative**

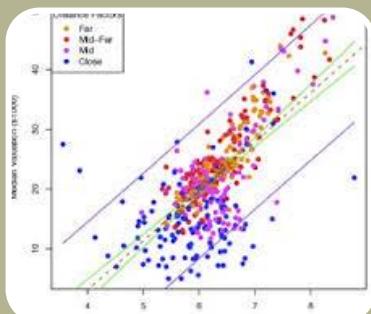
- Prédire les tendances salariales la prochaine année
- Prédire le meilleur pourcentage de réduction de coûts
- Construire un modèle générique pour l'estimation de la consommation de carburant



La régression a pour objectifs :

- Déetecter les variables possédant un lien fort avec la variable cible
- Construire un modèle prédictif avec l'ensemble des variables pertinentes afin de prédire la variable d'intérêt
- Déetecter les individus atypiques ou les aberrances

Régression Linéaire



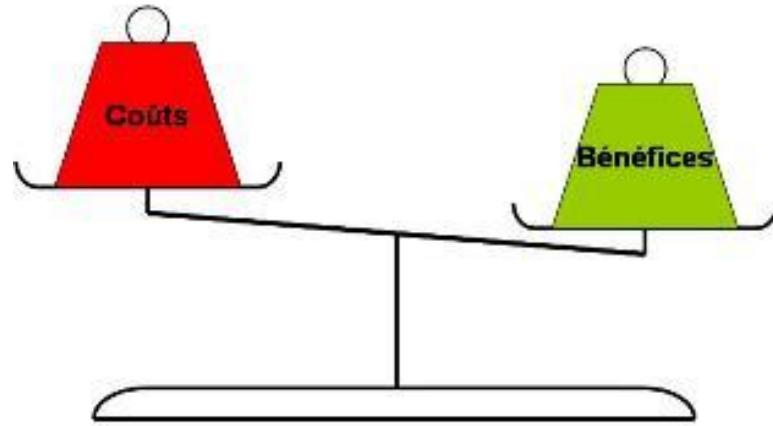
- Méthode des moindres carrés
- Meilleurs prédicteurs

Déterminer le **prix "optimal"**
convenable pour un produit.



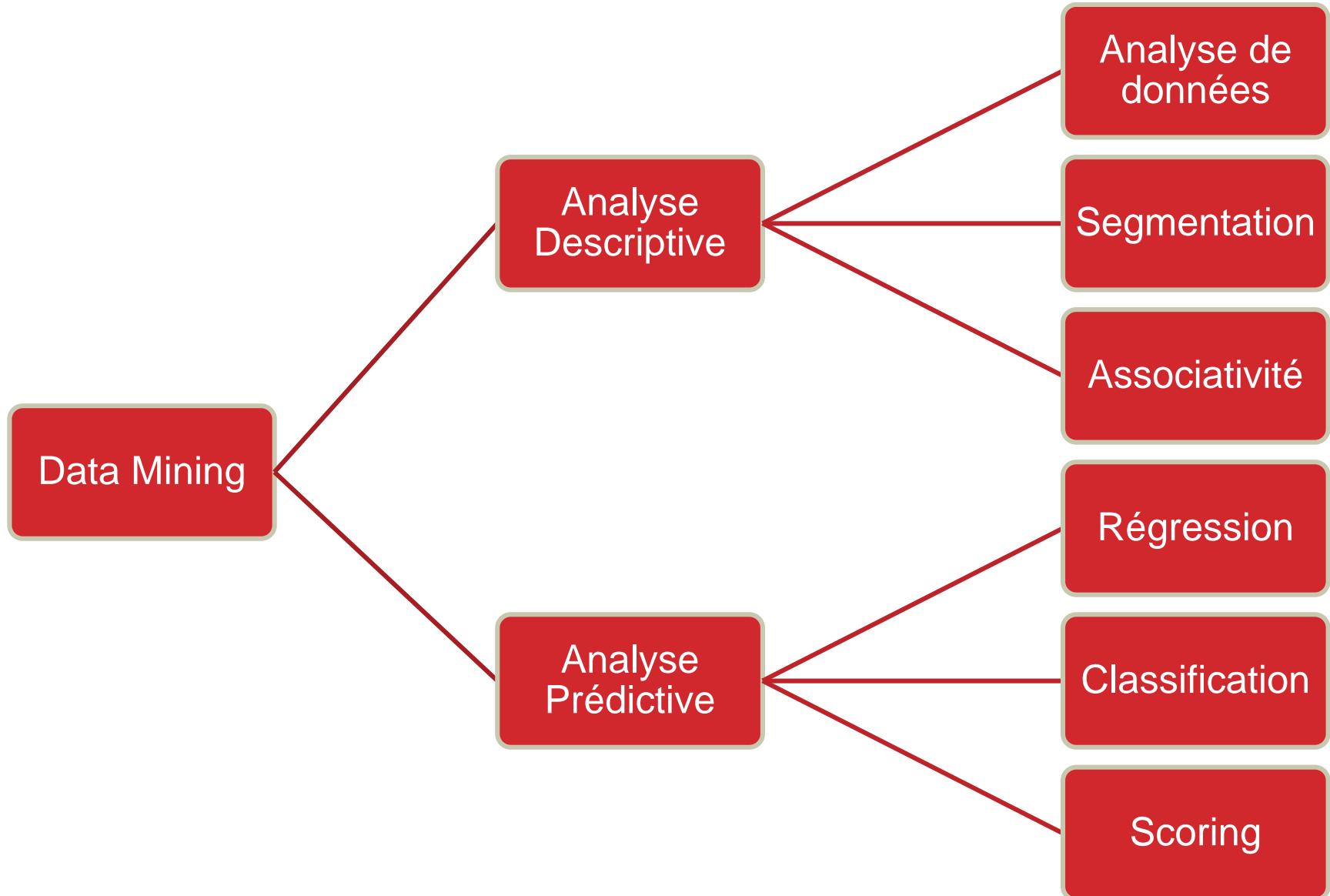
ANALYSE COÛTS-BÉNÉFICES

Identifier les produits et les campagnes les *plus rentables*



LES MÉTHODES DATA MINING

APPLICATIONS DE BASE DU DATA MINING



DEUX FAMILLES DE TECHNIQUES

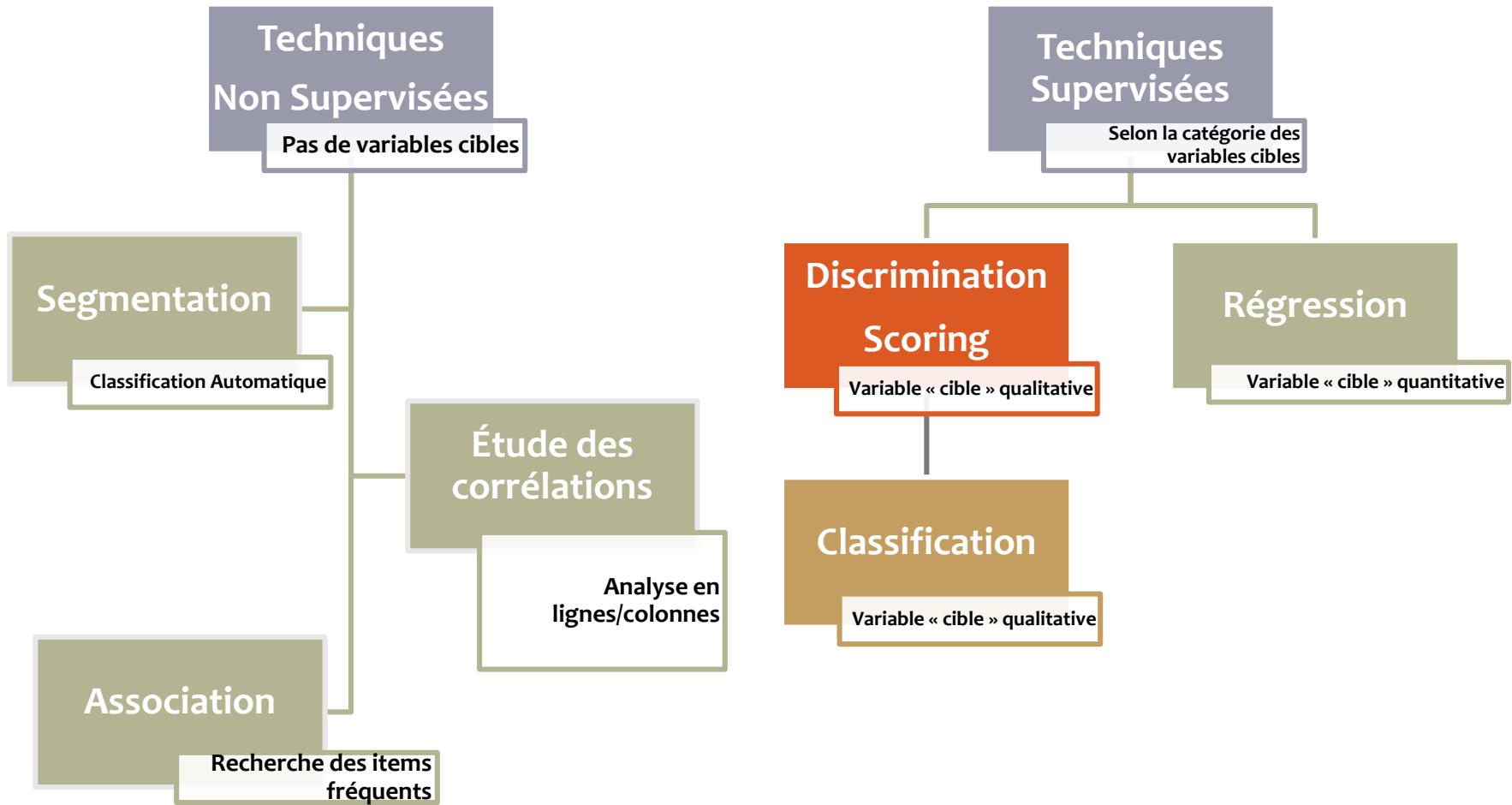
ANALYSE DESCRIPTIVE

- ✓ S'appliquent seulement sur les données quantitatives.
- ✓ Fournissent directement des résultats : à interpréter et à utiliser !
- ✓ visent à mettre en évidence des connaissances **présentes** mais cachées par le volume des données
- ✓ réduisent, résument, synthétisent les masses de données
- ✓ **pas de variable « cible »**

ANALYSE PRÉDICTIVE

- ✓ Fournissent un modèle (et non pas des résultats), créé à partir d'un entrepôt d'apprentissage, testé et validé sur un entrepôt de test, et utilisé dans les problèmes de prise de décision sur des entrepôts de travail
- ✓ visent à découvrir de **nouvelles informations** à partir des informations présentes : connaissances, **décisions**
- ✓ expliquent mieux les données
- ✓ **Une(des) variable(s) « cible(s)»**

TYPES D'APPLICATIONS



DEUX FAMILLES DE TECHNIQUES

Méthodes

Non Supervisées

Analyse en Composantes
Principales
ACP

Méthodes des Centres
Mobiles
K-means

Classification Ascendante
Hiérarchique
CAH

Règles
Associatives
Apriori

Méthodes

Supervisées

Arbres de Décisions

Analyse Linéaire Discriminante

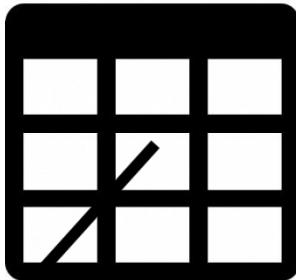
Régression

Plus Proche voisin kNN

Réseaux de Neurones

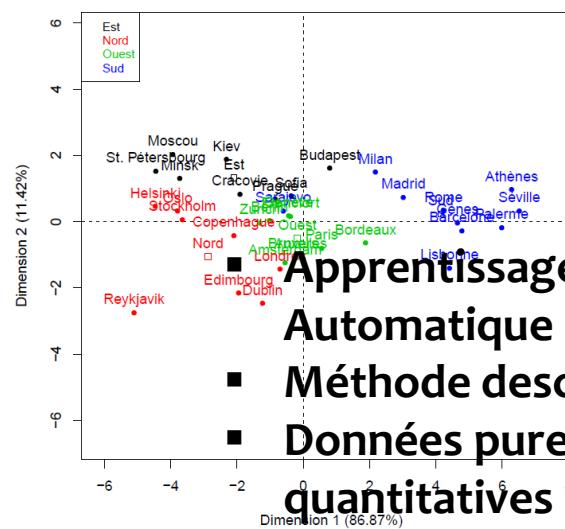
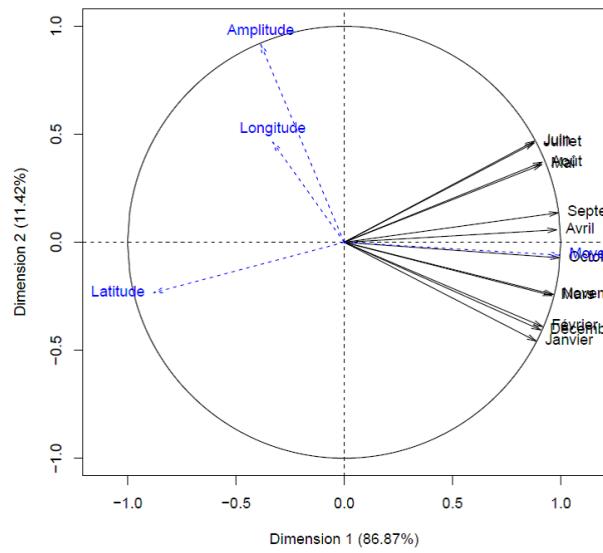
Séparateur à Vaste Marge
SVM

Méthodes Non Supervisées



Méthodes Supervisées

Analyse en Composantes Principales ACP

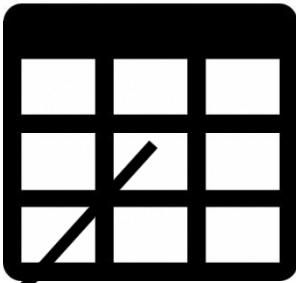


- **Apprentissage Automatique**
- **Méthode descriptive**
- **Données purement quantitatives**
- **Obtention de Corrélations**
- **Relations :**
 - (I_i, I_j) ,
 - (I_i, X_k) ,
 - (X_k, X_p)

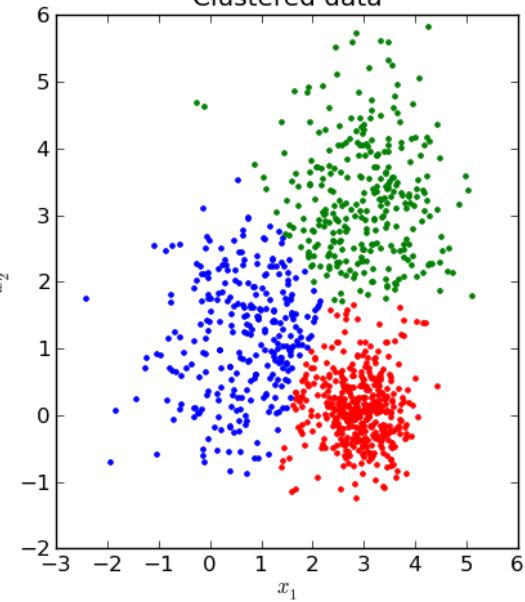
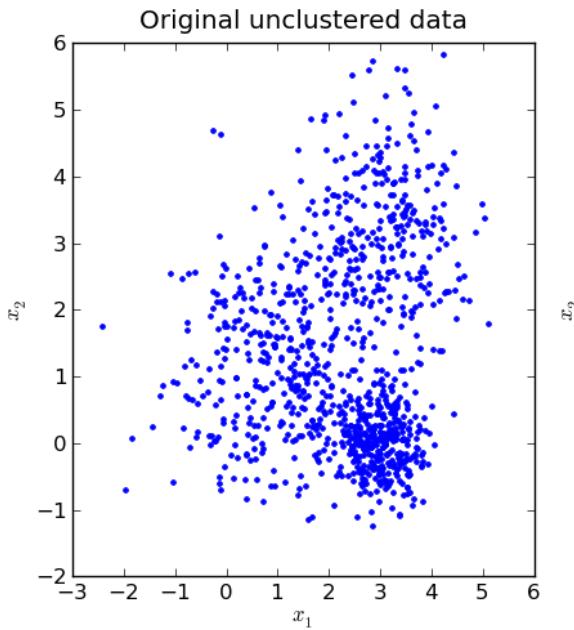
Entrepôt de données

Méthodes Non Supervisées

Méthodes Supervisées



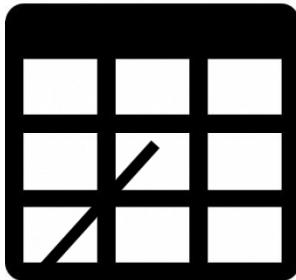
Méthodes de Segmentation (Classification Automatique)



- **Apprentissage Automatique**
- **Méthode descriptive**
- **Obtention d'une nouvelle affection/appartenance**
- **Groupes d'individus possédant des caractéristiques semblables : meilleure classification**
- **Mise à jour de l'entrepôt en Input : ajouter la nouvelle caractéristique des individus**

Méthodes Non Supervisées

Méthodes Supervisées

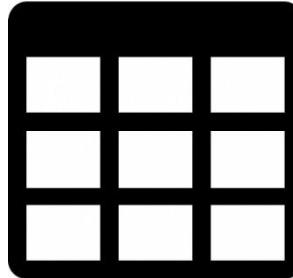


Règles Associatives (algorithme apriori)

```
> inspect(regles.triees[1 :5])
   lhs                      rhs          support confidence      lift
1 {Tomatoes,
  Eggs,
  White_Bread}  => {Sugar_Cookies} 0.02130786  0.80555556 14.618148
2 {Sugar_Cookies,
  Eggs,
  White_Bread}  => {Tomatoes}      0.02130786  0.7837838  11.852553
3 {Onions,
  X2pct_Milk,
  Eggs}           => {Wheat_Bread}   0.02130786  0.7631579  9.891980
4 {Eggs,
  White_Bread,
  Sweet_Relish}  => {Toothpaste}    0.02057311  0.77777778 9.801440
5 {Cola,
  Sweet_Relish}  => {Toothpaste}    0.02130786  0.7631579  9.617203
> |
```

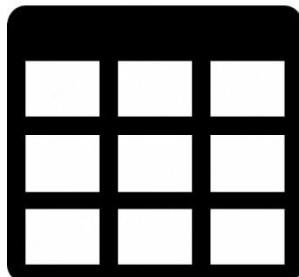
- **Apprentissage Automatique**
- **Méthode descriptive**
- **Données Transactionnelles**
- **Obtention d'un ensemble de règles expliquant l'associativité/correspondance entre les items/produits/profils**
- **Trier les règles selon des indicateurs de performances**

Méthodes Non Supervisées

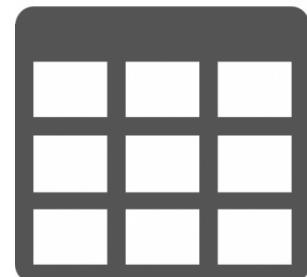


Méthodes Supervisées

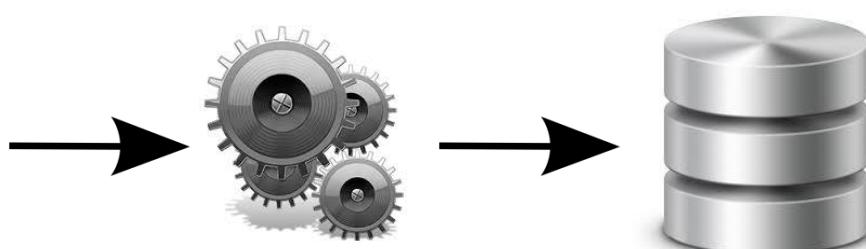
Entrepôt d'apprentissage : présentant une variable décisionnelle $Y_{réelle}$



Entrepôt de test



Entrepôt de travail



Méthodes Supervisées :

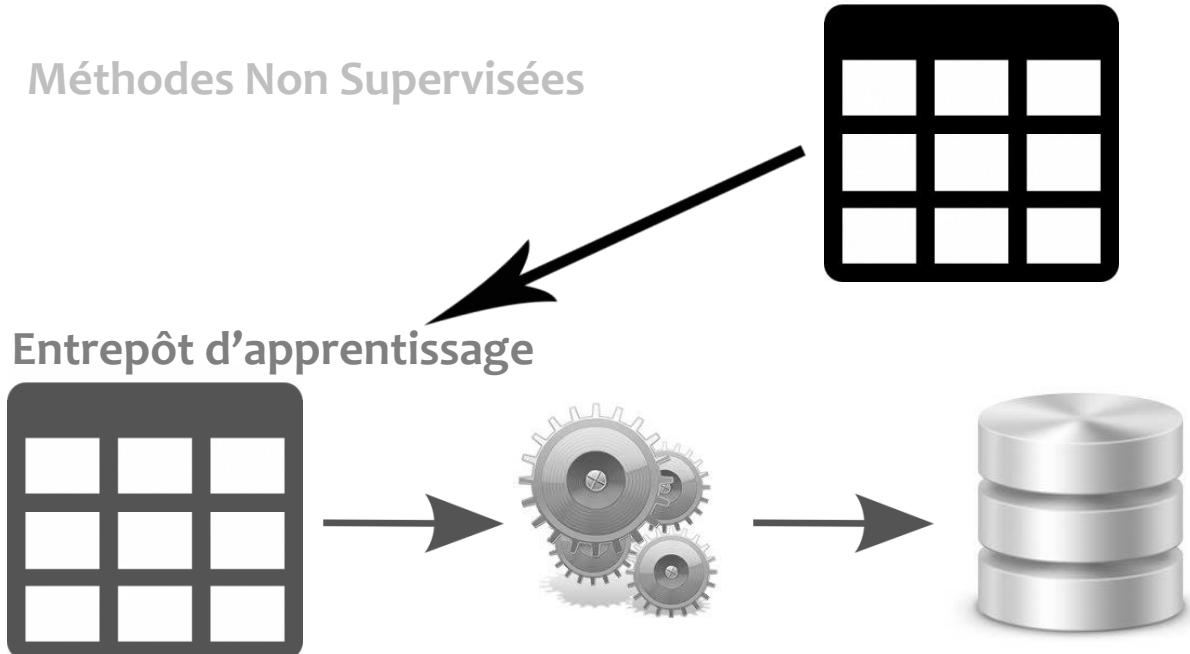
- i. Arbre de Décision
 - ii. Régression
 - iii. Analyse Discriminante LDA
 - iv. Réseaux de Neurones NN
 - v. Séparateur à vaste marge SVM
 - vi. Plus proche voisin k-NN
- Etc.

Modèle de Classification ou de Prédiction :

- ✓ Ensemble de règles décisionnelles
- ✓ Formule mathématique / $Y=f(X_i)$
- Formulations de Score / $Score_A$, $Score_B$, etc.
- Equation d'un plan séparateur
- Résultat d'un algorithme itératif

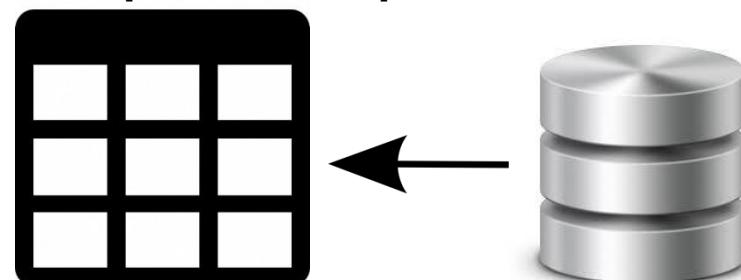
Méthodes Non Supervisées

Méthodes Supervisées

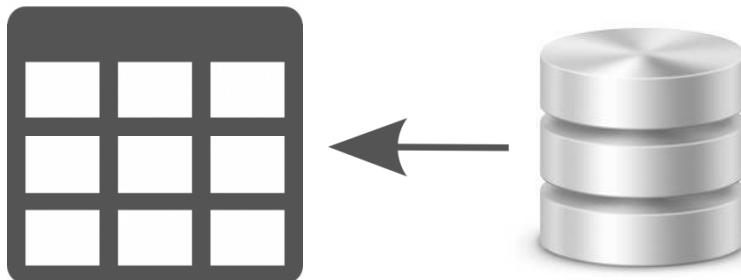


Entrepôt de test : présentant une *variable décisionnelle* $Y_{réelle}$

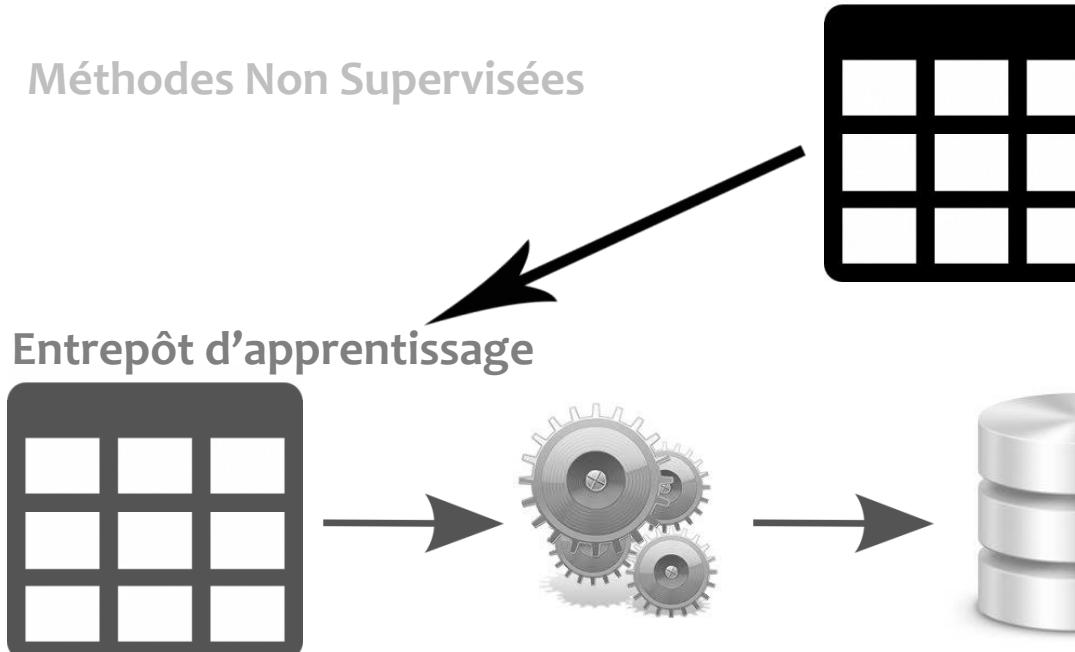
- ✓ Appliquer le modèle construit pour obtenir les valeurs de la variable décisionnelle $Y_{prédictive}$
- ✓ Comparer les deux colonnes $Y_{réelle}$ et $Y_{prédictive}$
- ✓ Valider le modèle



Entrepôt de travail

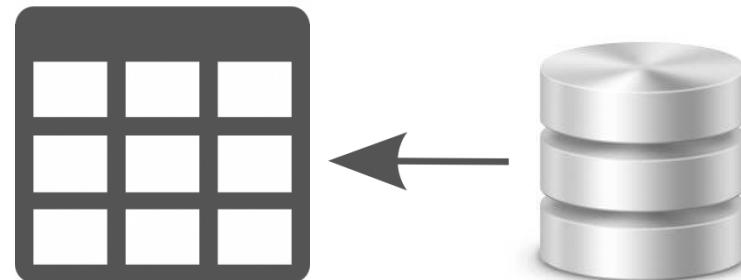


Méthodes Non Supervisées

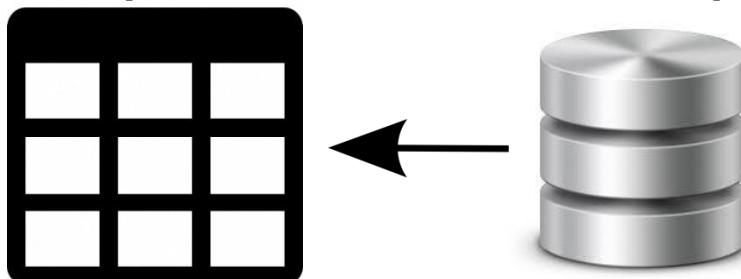


Méthodes Supervisées

Entrepôt de test : présentant une variable décisionnelle $Y_{réelle}$



Entrepôt de travail : ne contenant pas les valeurs de la variable décisionnelle



- ✓ Appliquer le modèle construit pour obtenir les valeurs de la variable décisionnelle $Y_{définitive}$

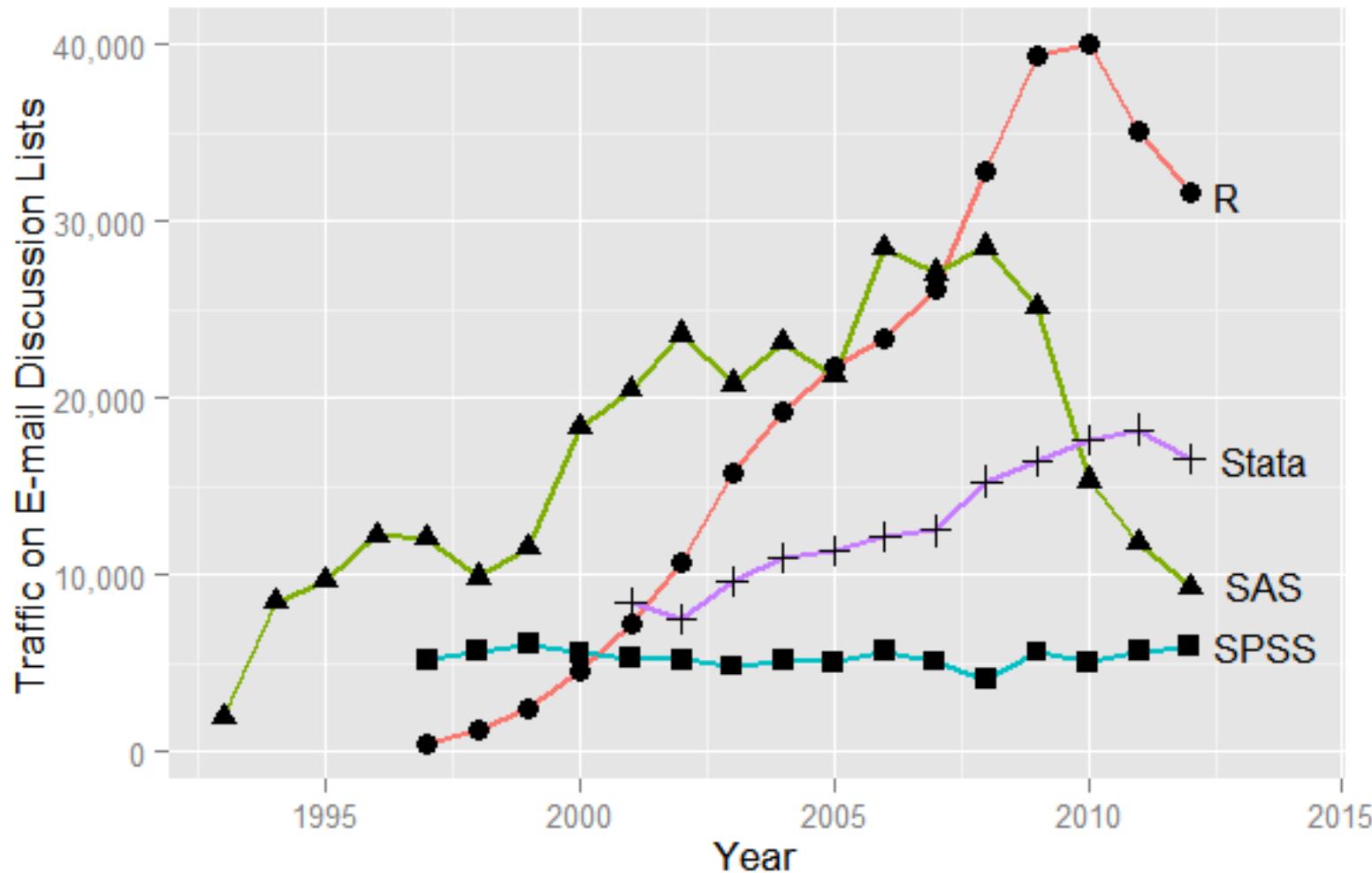
LES LOGICIELS PROFESSIONNELS

Critères de comparaison :

- ✓ Temps de réponse
- ✓ Payant ou gratuit
- ✓ Extensibilité
- ✓ Richesse en terme de méthodes
- ✓ Souplesse du paramétrage des méthodes appliquées



LES LOGICIELS PROFESSIONNELS



MÉTHODOLOGIES DE TRAVAIL



KDD / ECD

**KNOWLEDGE DATA DISCOVERY
EXTRACTION DE CONNAISSANCES À PARTIR DE DONNÉES**

DÉFINITION

PHASES PRINCIPALES

APPLICATION

KDD: DÉFINITION

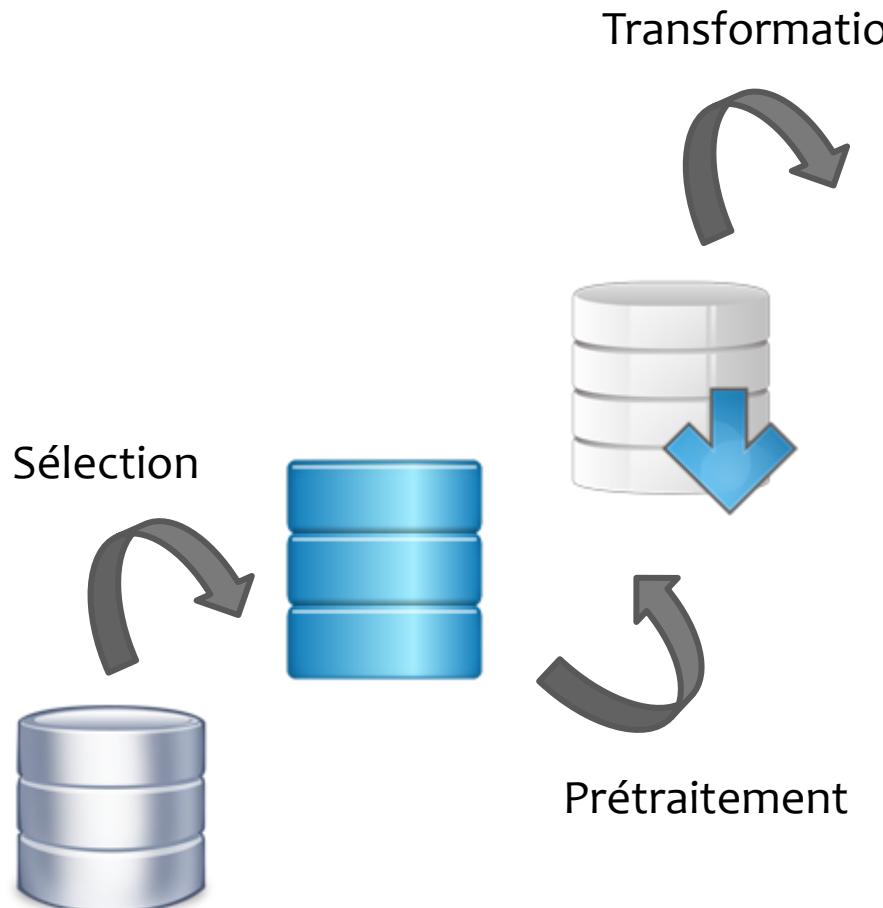
Knowledge Discovery in Databases

- proposé par Ossama Fayyad en 1996
- un processus pour la fouille de données qui a bien répondu aux besoins d'entreprises, et qui est devenu rapidement très populaire.
- KDD a comme but l'extraction des connaissances,
- des motifs valides, utiles et exploitables à partir des grandes quantités de données
- par des méthodes automatiques ou semi-automatiques.

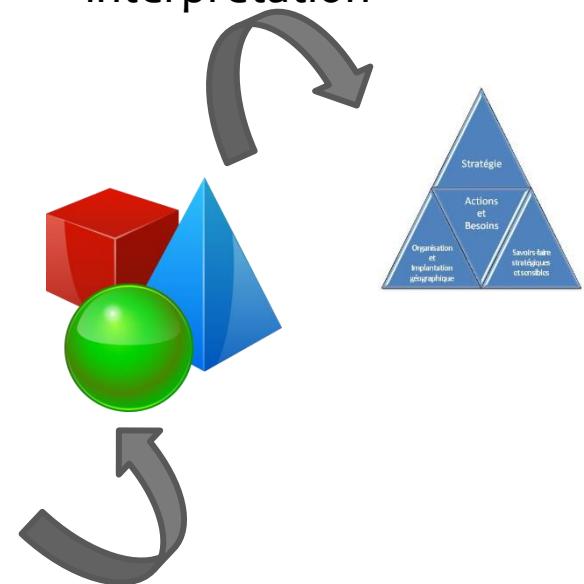
KDD: DÉFINITION

- Le processus de KDD est **itératif** et **interactif**.
- Le processus est itératif : il peut être nécessaire de refaire les pas précédents.
- Le problème de ce processus, comme pour les autres présentés dans la section suivante, est le manque de guidage de l'utilisateur, qui ne choisit pas à chaque étape la meilleure solution adaptée pour ses données.

KDD: PHASES PRINCIPALES



Evaluation et
interprétation



KDD: PHASES PRINCIPALES

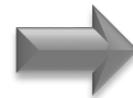
1. Développer et comprendre le domaine de l'application

C'est le pas initial de ce processus. Il prépare la scène pour comprendre et développer les buts de l'application.

KDD: **PHASES PRINCIPALES**

2. Sélection des données

La sélection et la création d'un ensemble de données sur lequel va être appliqué le processus d'exploration.

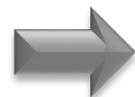


Données ciblées

KDD: **PHASES PRINCIPALES**

3. Le prétraitement et le nettoyage des données

Cette étape inclut des opérations comme l'enlèvement du bruit et des valeurs aberrantes -si nécessaire, des décisions sur les stratégies qui vont être utilisées pour traiter les valeurs manquantes...



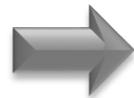
Données prétraitées

KDD: **PHASES PRINCIPALES**

4. La transformation des données

Cette étape est très importante pour la réussite du projet et doit être adaptée en fonction de chaque base de données et des objectifs du projet.

Dans cette étape nous cherchons les méthodes correctes pour représenter les données. Ces méthodes incluent la réduction des dimensions et la transformation des attributs.



Données transformées

◆ Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie de Data mining, avec une orientation sur l'aspect algorithmique.

KDD: **PHASES PRINCIPALES**

5. Choisir la meilleure tâche pour Datamining

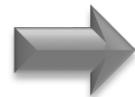
Nous devons choisir quel type de Datamining sera utilisé, en décidant le but du modèle.

- ◆ Par exemple : classification, régression, regroupement...

KDD: **PHASES PRINCIPALES**

6. Choisir l'algorithme de Datamining

Dans cette étape nous devons choisir la méthode spécifique pour faire la recherche des motifs, en décidant quels modèles et paramétrés sont appropriés.



Modèles

KDD: PHASES PRINCIPALES

7. Implémenter l'algorithme de Datamining

Dans cette étape nous implémentons les algorithmes de Datamining choisis dans l'étape antérieure.

Peut être il sera nécessaire d'appliquer l'algorithme plusieurs fois pour avoir le résultat attendu.

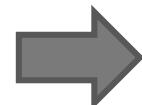
KDD: **PHASES PRINCIPALES**

8. Evaluation

Evaluation et interprétation des motifs découverts.

Cette étape donne la possibilité de:

- Retourner à une des étapes précédentes
- Avoir une représentation visuelle des motifs, enlever les motifs redondants ou non-représentatifs et les transformer dans des termes compréhensibles pour l'utilisateur.



Connaissances

KDD: PHASES PRINCIPALES

9. Utiliser les connaissances découvertes

Incorporation de ces connaissances dans des autres systèmes pour d'autres actions.

Nous devons aussi mesurer l'effet de ces connaissances sur le système, vérifier et résoudre les conflits possibles avec les connaissances antérieures.

KDD: APPLICATION

Le KDD est devenu lui-même un modèle pour les nouveaux modèles.

Le modèle a été utilisé dans plusieurs domaines différentes : ingénierie, médecine, e-business, production, développement du logiciel, etc.



SAMPLE, EXPLORE, MODIFY, MODEL, ASSESS

MISE EN CONTEXTE

DÉFINITION

PHASES PRINCIPALES

APPLICATION

SEMMA: MISE EN CONTEXTE

L’Institut SAS définit le data mining comme le processus utilisé pour révéler des informations précieuses et des relations complexes qui existent dans de grandes quantités de données (**BIG DATA, OPEN DATA**).

→ SAS divise la fouille de données en cinq étapes représentées par l’acronyme **SEMMA**.

SEMMA: DÉFINITION

Prédiction

*Transformation
des variables
prédictives*

*Précision du
modèle*

*Exploration
statistiques*

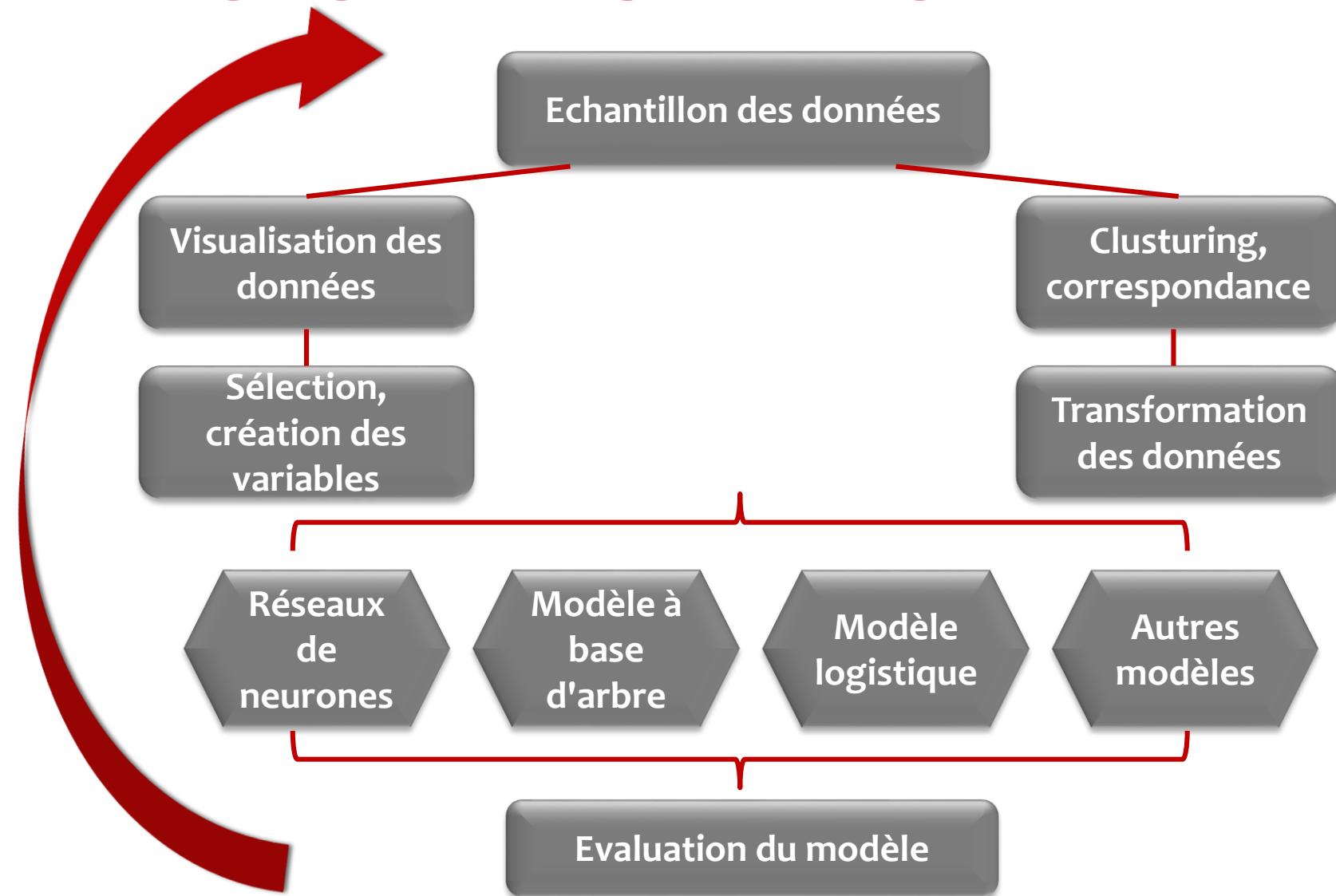
SEMMA

Sélection

*Modélisation
des variables*

*Visualisation
des données*

SEMMA: PHASES PRINCIPALES



SEMMA: PHASES PRINCIPALES

1. Sample: Echantillon des données

Extrait des échantillons d'un vaste ensemble de données, en nombre suffisamment grand pour contenir l'information importante.

SEMMA: PHASES PRINCIPALES

2. Explore: Exploitation des données

Cette étape consiste dans l'exploration des données en recherchant les tendances et les anomalies imprévues afin de mieux comprendre les données.

SEMMA: PHASES PRINCIPALES

3. Modify: Modifier les données

Dans cette étape on modifie les données en créant, en sélectionnant et en transformant les variables afin de s'axer sur le processus de sélection de modèles.

SEMMA: PHASES PRINCIPALES

4. Model: Modélisation des données

Modélisation des données en permettant au logiciel de rechercher automatiquement une combinaison des données qui prédit de façon fiable le résultat souhaité.

Il y a plusieurs techniques de modélisation: les réseaux de neurones, arbres de décision, modèles statistiques - l'analyse en composantes principales, l'analyse de séries temporelles...

SEMMA: PHASES PRINCIPALES

5. Assess: Evaluer le résultat

Cette étape consiste à l'évaluation de l'utilité et la fiabilité des résultats du processus de DataMining et estime comment il va s'exécuter.

En évaluant les résultats obtenus à chaque étape du processus de SEMMA, nous pouvons déterminer la façon de modéliser les nouveaux problèmes déterminés par les résultats précédents, et donc de refaire la phase d'exploration supplémentaire pour le raffinement des données.

SEMMA: APPLICATION

- SAS est définie par SEMMA comme l’organisation logique d’outil SAS Enterprise Miner pour la réalisation des tâches de DataMining.
- Enterprise Miner peut être utilisé comme une partie de n’importe quelle méthodologie itérative de DataMining adoptée par le client.
- Une des différences entre KDD et SEMMA est que SEMMA est *intégré dans l’outil Enterprise Miner* et ils n’utilisent pas d’autres méthodologies, tandis que le KDD est un processus ouvert qui peut être appliqué dans plusieurs environnements.

NCR

SPSS



Mercedes-Benz

CRISP-DM

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

DÉFINITION

PHASES PRINCIPALES

EXEMPLE

esprit

Ecole Supérieure Privée
d'Ingénierie et de Technologies

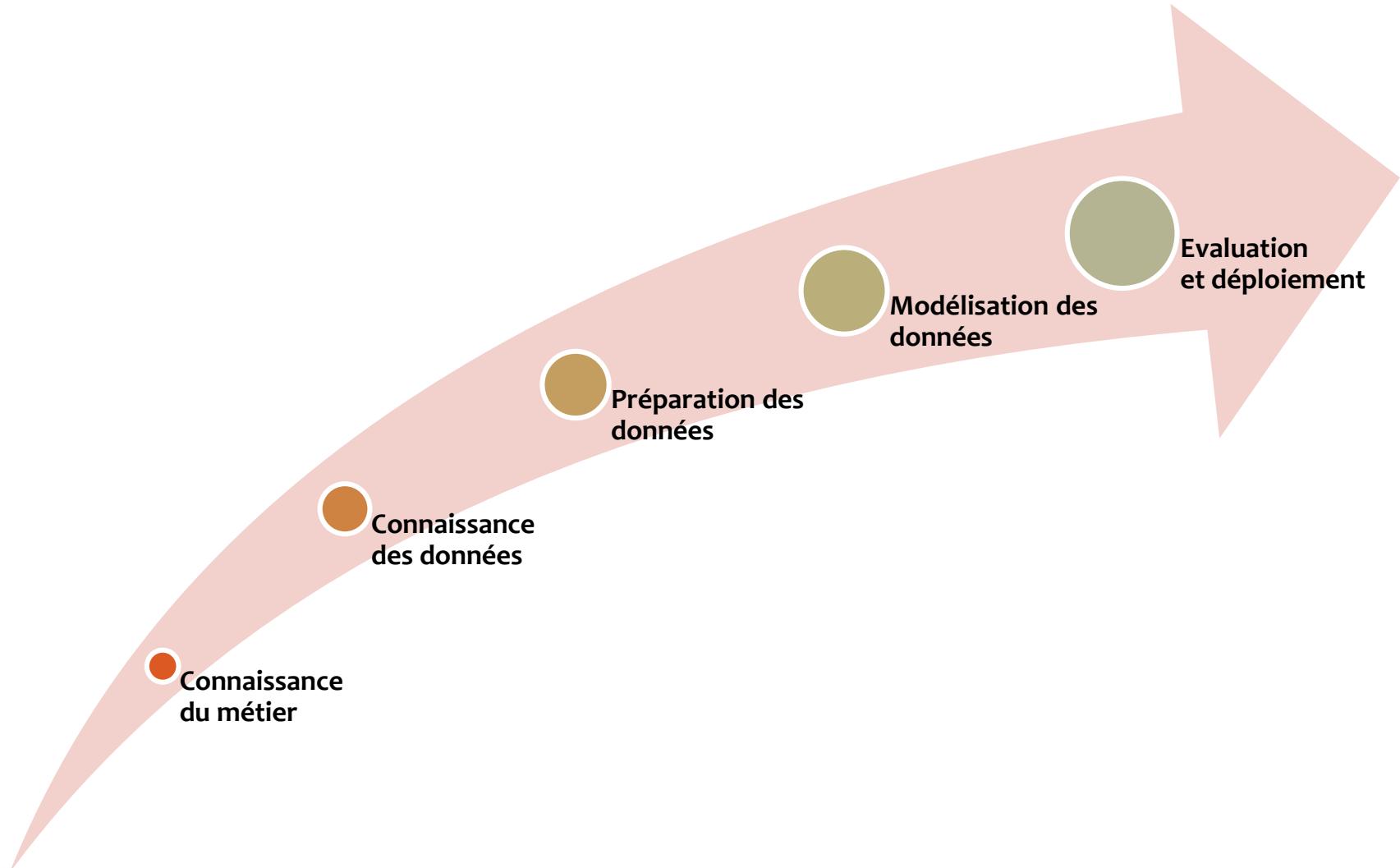
CRISP-DM : DÉFINITION

Cross-Industry Standard Process for Data Mining

Une méthode mise à l'épreuve sur le terrain permettant d'orienter les travaux de Data mining

Processus de data mining qui décrit une approche communément utilisée par les experts pour résoudre les problèmes qui se posent à eux.

PHASES PRINCIPALES



CRISP-DM : DÉFINITION

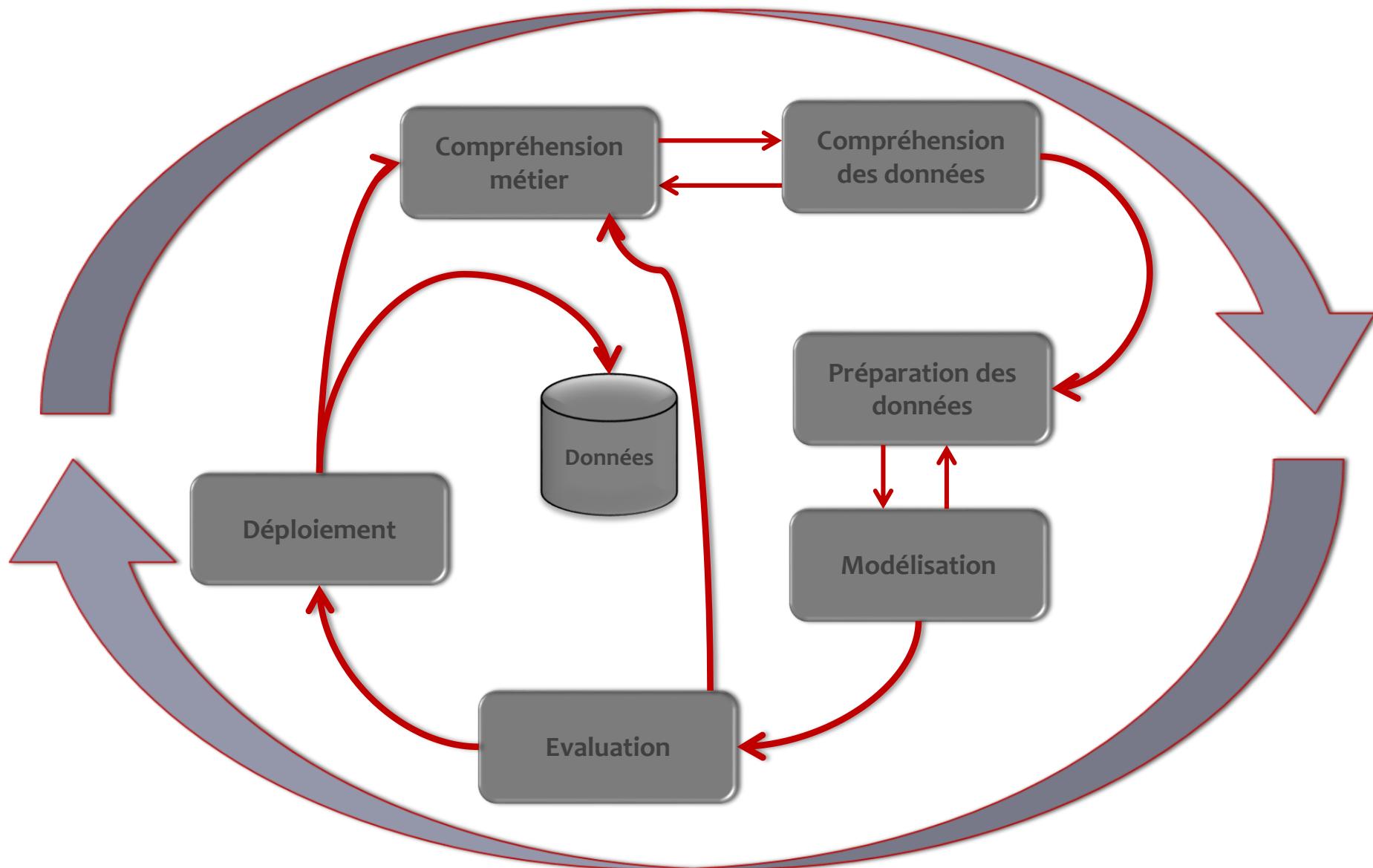
Méthodologie

- ✓ comprend des descriptions des phases typiques d'un projet et des tâches comprises dans chaque phase, et une explication des relations entre ces tâches.

Modèle de processus

- ✓ offre un aperçu du cycle de vie du Data mining.

CRISP-DM : PHASES PRINCIPALES



1. COMPRÉHENSION MÉTIER

Déterminer les objectifs d'affaires

Résoudre un problème spécifique

Evaluer la situation actuelle

Convertir en un problème de data mining

- ✓ Quels types de clients sont intéressés par chacun de nos produits?
- ✓ Quels sont les profils typiques de nos clients?

Élaborer un plan de projet

2. COMPRÉHENSION DES DONNÉES

Collecte de données initiale

Description des données

Exploration des données

Vérification de la qualité des données

Sélection des données

Les données connexes peuvent provenir de nombreuses sources :

- ✓ Interne (ERP, CRM, Data Warehouse...)
- ✓ Externe (données commerciales, données du gouvernement...)
- ✓ Crées (recherche)

LES ENJEUX DE LA SÉLECTION DES DONNÉES

Mettre en place une description concise et claire du problème

- ✓ Identifier les comportements de dépenses des femmes qui achètent des vêtements saisonniers
- ✓ Identifier les modèles de la faillite de détenteurs de cartes de crédit

Identifier les données pertinentes pour la description du problème

- ✓ Données démographiques, données financières...

Les variables sélectionnées pour les données pertinentes doivent être indépendantes les unes des autres.

3. PRÉPARATION DES DONNÉES

Nettoyer les données sélectionnées pour une meilleure qualité

- ✓ Remplissez les valeurs manquantes
- ✓ Identifier ou supprimer les valeurs aberrantes
- ✓ Résoudre la redondance causée par l'intégration des données
- ✓ Les données incohérentes correctes

Transformer les données

- ✓ Convertir des mesures différentes de données dans un échelle numérique unifié en utilisant des formulations mathématiques simples

LES DONNÉES DANS LE MONDE RÉEL!

Incomplètes: manque de valeurs d'attributs, manque de certains attributs d'intérêt ou ne contenant que des agrégats des attributs d'intérêt ou contenant uniquement des données agrégées

- ✓ l'occupation = ""

Bruyantes: contenant des erreurs ou des valeurs aberrantes

- ✓ Salaire = "- 1000"

Incompatibles: contenant des écarts dans les codes ou les noms

- ✓ Age = "42" et anniversaire = "03/07/1993"
- ✓ note =« 1,2,3 » ensuite «A, B, C »

PRINCIPALES CAUSES

Les données incomplètes peuvent provenir de:

- ✓ La valeur de données lors de la collecte «Sans objet»
- ✓ Des considérations différentes entre le moment où les données ont été collectées et lorsqu'elles sont analysées.
- ✓ Problèmes humains / matériels / logiciels

Les données bruyantes (valeurs incorrectes) peuvent provenir de:

- ✓ Les instruments de collecte de données sont erronés
- ✓ L'erreur humaine ou informatique à la saisie de données
- ✓ Les erreurs de transmission de données

Les données incohérentes peuvent provenir de:

- ✓ Les différentes sources de données
- ✓ La violation de la dépendance fonctionnelle (par exemple, de modifier certaines données liées)

TRANSFORMATION DES DONNÉES

Transformez le numérique à des échelles numériques

- ✓ Les échelles salariales de « 100 TND » à « 1000 TND » à un certain nombre de [0.0, 1.0]
- ✓ Le système métrique (par exemple, le mètre, kilomètre) au système anglais (par exemple, des pieds et miles)

Recoder les données catégoriques à des échelles numériques

- ✓ "1" = "oui" et "0" = "No"

4. MODÉLISATION

Traitements des données

- ✓ Ensemble d'apprentissage
- ✓ Ensemble de test...

Les techniques de data mining

- ✓ Association
- ✓ Classification
- ✓ Clustering
- ✓ Prédictions
- ✓ Les motifs séquentiels

5. EVALUATION

Est-ce que le modèle répond aux objectifs métier?

Des objectifs métier importants non résolus?

Est-ce que le modèle est logique?

Est-ce que le modèle est actionnable?

Il devrait être possible de prendre des décisions après cette étape.

Tous les objectifs importants doivent être atteints.

6. DÉPLOIEMENT / IMPLÉMENTATION

En cours de suivi et d'entretien

- ✓ Évaluer la performance par rapport aux critères de réussite
- ✓ La réaction du marché et les changements des concurrents

CRISP-DM: **ETUDE DES FACTURES DE TÉLÉPHONE**

Problème : Les factures de téléphone non payées.

- Le data mining utilisé pour développer des modèles pour prédire le non paiement des factures au plus tôt possible.



CRISP-DM: Exemple

Etude des factures de téléphone

Séquence de période de facturation:

- ✓ Utilisez 2 mois, recevoir la facture, le paiement du mois de facturation, débrancher si la facture n'est pas réglée pendant une période déterminée

1. COMPRÉHENSION MÉTIER

Prédire quels clients seraient insolvables

- ✓ À temps pour l'entreprise pour prendre des mesures préventives (et d'éviter de perdre de bons clients)

hypothèse:

- ✓ Clients insolvables vont changer les habitudes d'appel et l'usage du téléphone pendant une période critique avant et immédiatement après la fin de la période de facturation.

2. COMPRÉHENSION DES DONNÉES

Les informations statiques des clients sont disponibles dans des fichiers

- ✓ Factures, paiements, utilisation...

Un entrepôt de données est utilisé pour recueillir et organiser les données

- ✓ Un codage pour protéger la vie privée des clients

CRÉATION DE L'ENSEMBLE DES DONNÉES CIBLES

Les fichiers des clients:

- ✓ Informations sur les clients
- ✓ Déconnexion
- ✓ Reconnexions

Données dépendantes du temps

- ✓ Factures
- ✓ Paiements
- ✓ Utilisation

100, 000 clients sur une période de 17 mois

L'échantillonnage pour assurer à tous les groupes une représentation appropriée

3. PRÉPARATION DES DONNÉES

Filtrer les données incomplètes

Les appels en promotion supprimés

- ✓ Le volume des données réduit d'environ 50%

Faible nombre des cas de fraude

Vérification croisée avec les déconnexions du téléphone

Les données retardées sont nécessairement synchronisées

5. MODÉLISATION

Analyse discriminante

- ✓ Le modèle linéaire

Les arbres de décision

- ✓ Classificateur à base de règles

Réseaux de Neurones

- ✓ Le modèle non linéaire

5. EVALUATION

Premier objectif est de maximiser la précision de la prédiction des clients insolubles

- ✓ Arbre de décision un classificateur meilleur

Deuxième objectif est de minimiser le taux d'erreur pour les clients de solvants

- ✓ Le modèle Réseau de Neurones proche de l'arbre de décision

Utilisé tous les 3 sur la base de cas par cas.

6. IMPLÉMENTATION

Chaque client a été examiné avec les 3 algorithmes

- ✓ Si tous les 3 sont convenables, utiliser une classification
- ✓ En cas de désaccord, catégorisé comme non classé

Correcte sur les données d'essai avec 0.898

- ✓ Seulement 1 client solvant aurait été débranché

**CE QU'IL
FAUT
REtenir !**

SPÉCIALISTES DU MÉTIER

Définition des objectifs

avant l'élaboration d'un score de risque, il convient de s'entendre sur la définition précise d'un risque

Recensement des données

il est intéressant de connaître les indicateurs considérés comme pertinents par les spécialistes

Analyse des résultats

le spécialiste métier peut dire s'ils paraissent triviaux, nouveaux et intéressants à creuser, ou surprenants et très suspects



DATA MINING – STATISTIQUES ?

certaines techniques n'appartiennent qu'au DATA MINIG

le nombre d'individus étudiés est beaucoup plus important en DM, où l'optimisation des algorithmes est très importante

Data Mining fait beaucoup moins d'hypothèses contraignantes sur les lois statistiques suivies

Data Mining recherche plus la compréhensibilité des modèles que leur précision



APPORTS INCROYABLES DU DATA MINING !



Le Data Mining ne permet pas de faire des découvertes incroyables



LA meilleure façon de partitionner une masse de données

LA meilleure combinaison possible à détecter

LA meilleure interprétation des Individus, des variables...

LE meilleur apprentissage possible à atteindre

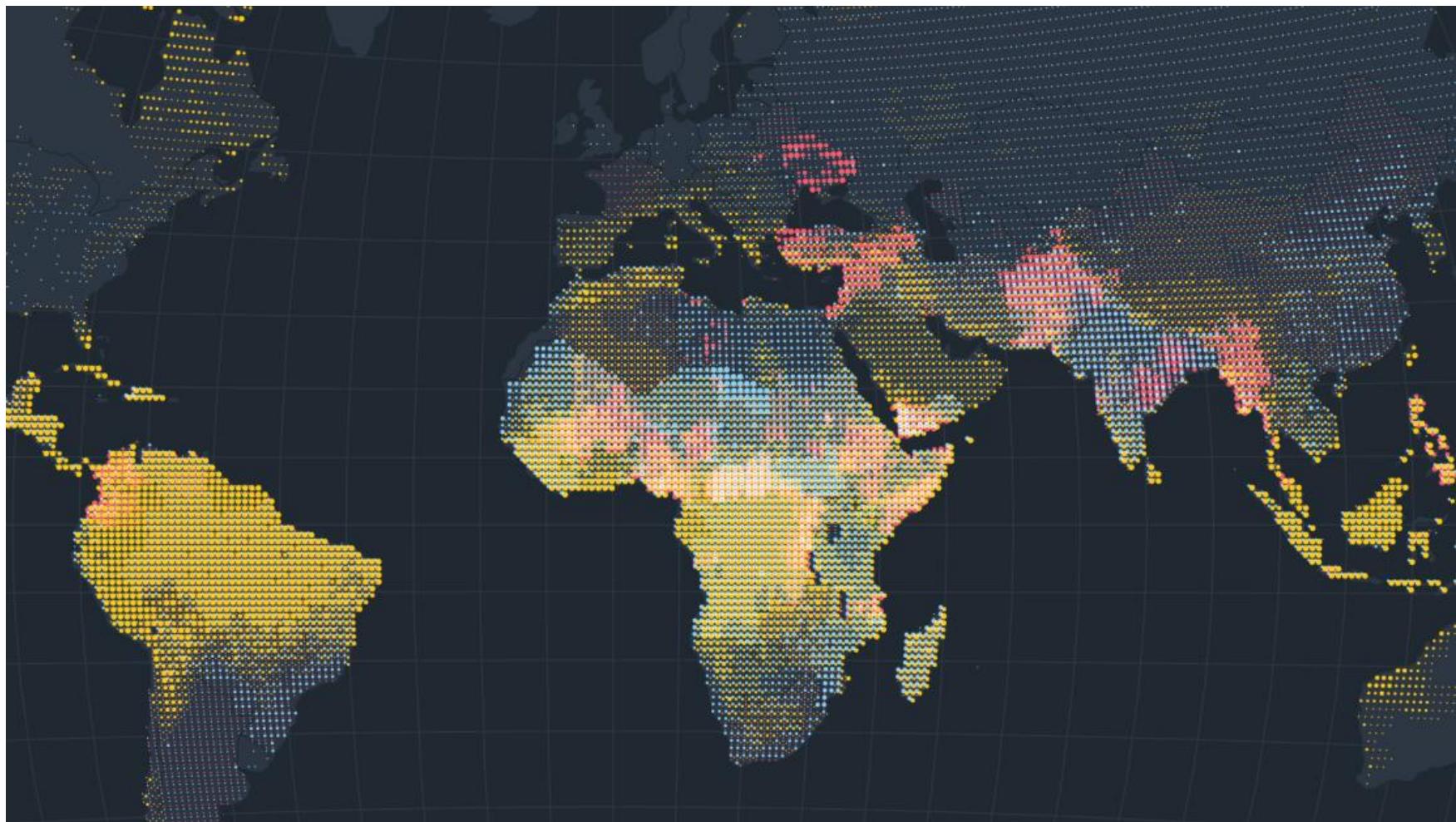
LE meilleur modèle de Classifier

LA meilleure DECISION possible à prendre

DÉROULEMENT DE LA SÉANCE

DATA MINING ;)

THE BEAUTY OF DATAVIZ



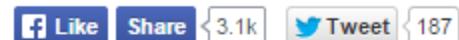
<https://truth-and-beauty.net/>

MODÈLE DE PRÉDICTION COUPE DU MONDE FIFA 2014 !

Prediction model for the FIFA World Cup 2014

June 12, 2014

By Christian Groll



(This article was first published on [Quantifying Information » R bloggers](#), and kindly contributed to R-bloggers)

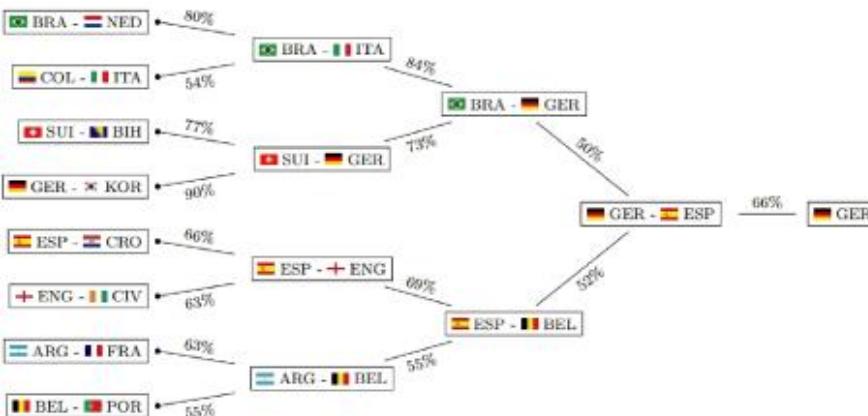
Like a last minute goal, so to speak, Andreas Groll and Gunther Schauberger of Ludwig-Maximilians-University Munich announced their predictions for the FIFA World Cup 2014 in Brazil – just hours before the opening game.

Andreas Groll, with his successful prediction of the European Championship 2012 already experienced in this field, and Gunther Schauberger did set out to predict the 2014 world cup champion based on statistical modeling techniques and R.

A bit surprisingly, Germany is estimated with highest probability of winning the trophy (28.80%), exceeding Brazil's probability (the favorite according to most bookmakers) only marginally (27.65%). You can find all estimated probabilities compared to the respective odds from a German bookmaker in the graphic on their homepage (<http://www.statistik.lmu.de/~schauberger/research.html>), together with the most likely world cup evolution simulated from their model. The evolution also shows the neck-and-neck race between Germany and Brazil: they are predicted to meet each other in the semi-finals, where Germany's probability of winning the game is a hair's breadth above 50%. Although there does not exist a detailed technical report on the results yet, you still can get some insight into the model as well as the data used through a preliminary summary pdf on their homepage (<http://www.statistik.lmu.de/~schauberger/WMGroßSchauberger.pdf>).

team	\hat{p}_{Lasso}	\hat{p}_{Oddset}	team	\hat{p}_{Lasso}	\hat{p}_{Oddset}
1. GER	0.2880	0.1420	17. GHA	0.0022	0.0071
2. BRA	0.2765	0.2028	18. KOR	0.0019	0.0024
3. ESP	0.0900	0.1092	19. ALG	0.0018	0.0071
4. BEL	0.0819	0.0592	20. ECU	0.0017	0.0071
5. ARG	0.0582	0.1420	21. USA	0.0016	0.0071
6. POR	0.0522	0.0237	22. MEX	0.0012	0.0071
7. SUI	0.0413	0.0071	23. JPN	0.0010	0.0047
8. CRO	0.0210	0.0071	24. BIH	0.0008	0.0047
9. ENG	0.0193	0.0355	25. GRE	0.0005	0.0071
10. FRA	0.0135	0.0355	26. RUS	0.0004	0.0118
11. NED	0.0129	0.0355	27. NGA	0.0004	0.0035
12. ITA	0.0094	0.0355	28. AUS	0.0003	0.0024
13. URU	0.0071	0.0284	29. HON	0.0002	0.0005
14. CHI	0.0063	0.0203	30. CRC	0	0.0071
15. COL	0.0052	0.0394	31. CMR	0	0.0024
16. CIV	0.0032	0.0071	32. IRN	0	0.0005

Group A 44%	Group B 24%	Group C 16%	Group D 18%	Group E 22%	Group F 36%	Group G 37%	Group H 26%
1. BRA	1. ESP	1. COL	1. ENG	1. SUI	1. ARG	1. GER	1. BEL
2. CRO	2. NED	2. CIV	2. ITA	2. FRA	2. BIH	2. POR	2. KOR
3. MEX	3. CHL	3. JPN	3. URU	3. ECU	3. NGA	3. GHA	3. RUS
4. CMR	4. AUS	4. GRE	4. CRC	4. HON	4. IRN	4. USA	4. ALG



DATA MINING ET L'ENJEU : DONNÉES !



OBSSESSIONS

QUARTZ



DUNNHUMBY

Why an obscure British data-mining company is worth \$3 billion



Outlook: grey. (Reuters/Toby Melville)

SHARE



WRITTEN BY

Tesco, the troubled British retail group, is starting over. After an accounting scandal, a series of profit warnings, and plunge in its share price, the beleaguered company has launched a major restructuring plan. It will not pay a dividend at the end of this financial

TODAY! SCIENTIST Sciente

Sciente

data minin

1 sur 1

MY AMERICAN SCIENTIST [LOG IN](#) [REGISTER](#)

SEARCH [GO](#)

AMERICAN Scientist

[Current Issue](#) [Past Issues](#) [Scientists' Nightstand](#) [Multimedia](#)

[About](#) [Subscribe](#) [Advertise](#)

HOME > PAST ISSUE > Article Detail

[RAISE FONT SIZE A A A](#) [VIEW PRINTER-FRIENDLY](#)

MACROSCOPE

Science Needs More *Moneyball*

Baseball's data-mining methods are starting a similar revolution in research

Frederick M. Cohan

The *Moneyball* story, in book and film, champions a data-mining revolution that changed professional baseball. On the surface, *Moneyball* is about Billy Beane, the general manager of the Oakland A's, who found a way to lead his cash-strapped club to success against teams with much bigger payrolls. Beane used data to challenge what everyone else managing baseball "knew" to be true from intuition, experience and training. He pioneered methods to identify outstanding players he could afford because they were undervalued by the traditional statistics used by the baseball elite.

This film was marketed as a sports movie. When I saw it, I knew right away what *Moneyball* is really about: the thrill and triumph of data mining. It's an instructive tale of how existing data can be examined for meaning in ways that were never intended or imagined when they were originally collected. Beane and his colleagues challenged the time-honored trinity of batting average, home runs and runs batted in (RBIs) as the essence of the offensive value of a player, replacing these statistics with newer measures based on the same data. They worked off theories developed by baseball writer and historian Bill James, who posited in the 1970s that the traditional stats were really imperfect measurements. James's approach didn't just replace one intuition with another. He let the game decide which stats did the best job of predicting offensive output.

Ce film a été commercialisé comme un film de sport. Lorsque je l'ai vu, je ai su tout de suite ce *Moneyball* est vraiment: le frisson et le triomphe de la fouille de données

This film was marketed as a sports movie. When I saw it, I knew right away what *Moneyball* is really about: the thrill and triumph of data mining. It's an instructive tale of how existing data can be examined for meaning in ways that were never intended or imagined when they were originally collected. Beane and his colleagues challenged the time-honored trinity of batting average, home runs and runs batted in (RBIs) as the essence of the offensive value of a player, replacing these statistics with newer measures based on the same data. They worked off theories developed by baseball writer and historian Bill James, who posited in the 1970s that the traditional stats were really imperfect measurements. James's approach didn't just replace one intuition with another. He let the game decide which stats did the best job of predicting offensive output.

IN THIS SECTION

- [Community Guidelines: Disqus](#)
- [Comments](#)
- [The Art and Science of Communicating Science](#)
- [American Scientist Classics](#)
- [Science Online: Women in Science](#)
- [Purchase a Back Issue](#)
- [Authors](#)
- [2014 Annual Index](#)

This Article from Issue
 May-June 2012
 Volume 100, Number 3
 Page: 182
 DOI: 10.1511/2012.96.182

[PURCHASE PDF](#)
[PRINTER-FRIENDLY VERSION](#)
[SAVE TO LIBRARY](#)

EMAIL TO A FRIEND :

Of Possible Interest

- [Feature Article: The Art of the Jim Henson's, In Food and Water](#)
- [Feature Article: Like Holding a Piece of Sky](#)
- [Feature Article: Curious Chemistry Guides Hydrangea Colors](#)

THE MAGAZINE OF SIGMA XI
 THE SCIENTIFIC RESEARCH SOCIETY

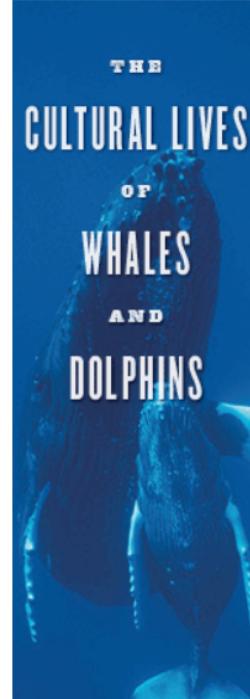


AMERICAN Scientist



SUBSCRIBE

THE CULTURAL LIVES of WHALES AND DOLPHINS



DATA MINING ET LES OSCARS !

Decideo.fr
Twitter RSS Print

MODÉLISATION PRÉdictive : L'ORACLE DES OSCARS AVAIT IL RAISON?

PHILIPPE NIEUWBOURG
4 MARS 2014

100 %... oui, c'en est presque inquiétant... mais Farsite avait prédit exactement les gagnants des principaux Oscars de la soirée 2014. En s'appuyant sur des données structurées et non structurées, Farsite a développé un modèle mathématique qui « prédit » les choix artistiques de l'Académie des Oscars. Info ou intox ?



The famous "selfie" from the 2014 Oscar ceremony that went viral on Twitter.

Grand jeu à l'occasion de la cérémonie des Oscars qui s'est déroulée il y a 2 jours à Los Angeles, la prédiction des résultats. Comme dans tout bon film de science fiction, la question est de savoir si la machine est capable de modéliser le comportement humain; et de prédire des résultats que l'on imaginait irréelables. Quoi de plus subjectif que les choix artistiques d'un jury de professionnels du 7ème art... Plusieurs entreprises se sont attachées à démontrer qu'en réalité leurs choix « artistiques » sont parfaitement prévisibles.

La société ICC avait prédit que :

- Matthew McConaughey remporterait l'Oscar du meilleur acteur pour son rôle dans Dallas Buyers Club... elle avait raison ;
- Alfonso Cuarón remporterait l'Oscar du meilleur réalisateur avec Gravity... elle avait raison ;
- 12 Years a Slave serait récompensé par l'Oscar de la meilleure image... elle avait raison ;
- Jared Leto serait Oscarisé meilleur second rôle masculin dans Dallas Buyers Club... elle avait raison ;
- Cate Blanchett meilleure actrice dans Blue Jasmine... elle avait encore raison ;

ABONNEZ-VOUS À LA NEWSLETTER

Votre email:

Votre nom:

BIG DATA SUMMIT NOV 16-18, 2014 PHOENIX, ARIZONA WWW.BIGDATASUMMIT.US

*Current Decideo Readers
25% RECEIVE DISCOUNT

AGENDA

AUJOURD'HUI

- IEEE VIS 2014 (au 14/11/2014)

MERCREDI 12 NOVEMBRE

- Formation Agile BI (au 14/11/2014)
- "Data analytics" : Des données aux connaissances et à la création de valeur
- Formation en ligne - L'intégration de données d'entreprise pas-à-pas

JEUDI 13 NOVEMBRE

- Talend Roadshow Nantes
- Business Forum Big Data

VENDREDI 14 NOVEMBRE

- BRUSSELS SOFTWARE DAY



Téléchargez
l'application
gratuite

Disponible sur App Store

100 %... oui, c'en est presque inquiétant... mais Farsite avait prédit exactement les gagnants des principaux Oscars de la soirée 2014. En s'appuyant sur des données structurées et non structurées, Farsite a développé un modèle mathématique qui « prédit » les choix artistiques de l'Académie des Oscars

LES DATA MININERs ?



State CIO Priorities for 2016

November 10, 2015

A. Priority Strategies, Management Processes and Solutions Top 10 Final Ranking

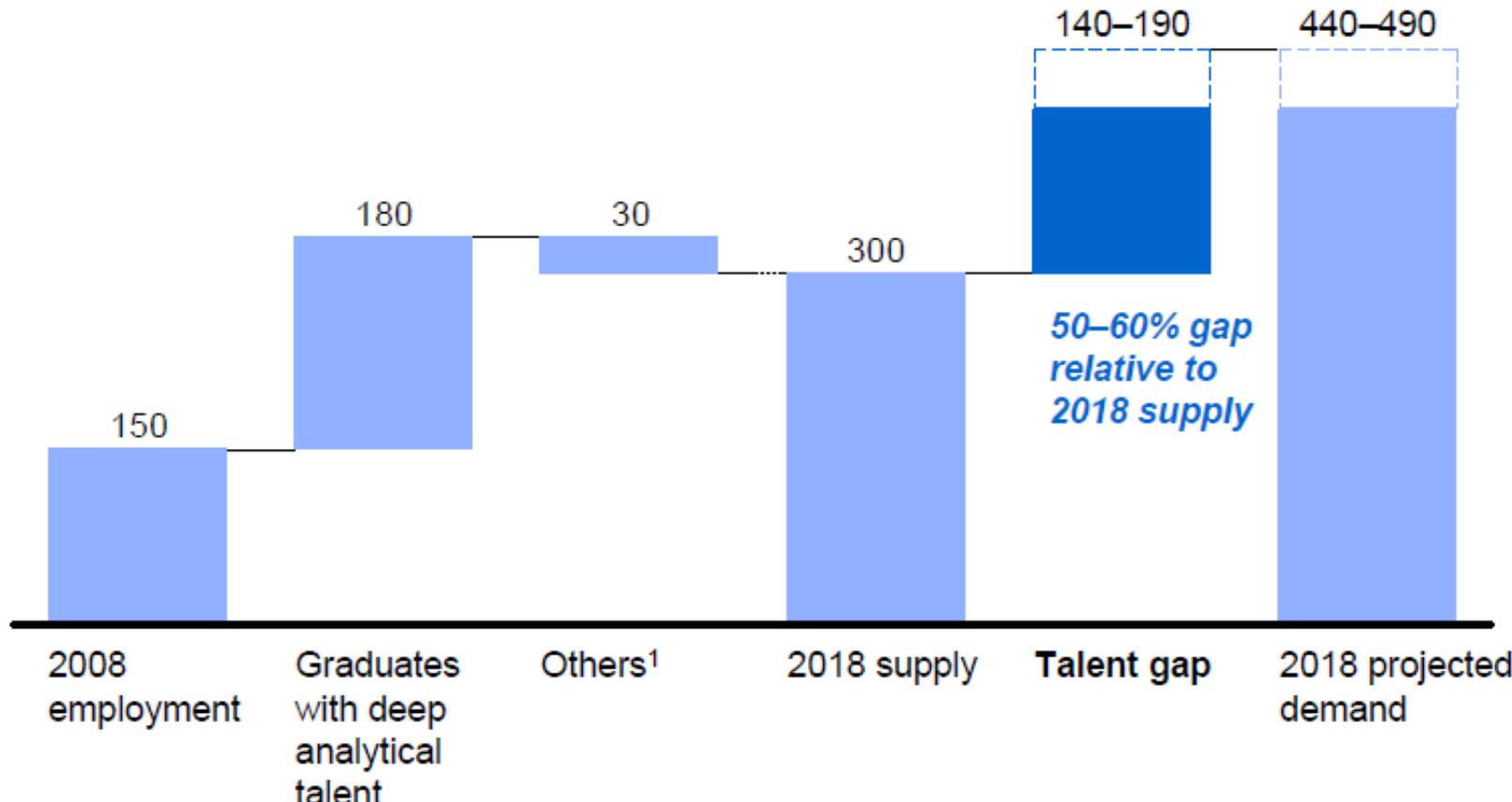
1. **Security and Risk Management:** governance, budget and resource requirements, security frameworks, data protection, training and awareness, insider threats, third party security practices as outsourcing increases, determining what constitutes "due care" or "reasonable"
2. **Cloud Services:** cloud strategy, proper selection of service and deployment models, scalable and elastic IT-enabled capabilities provided "as a service" using internet technologies, governance, service management, service catalogs, platform, infrastructure, security, privacy, data ownership
3. **Consolidation/Optimization:** centralizing, consolidating services, operations, resources, infrastructure, data centers, communications and marketing "enterprise" thinking, identifying and dealing with barriers
4. **Business Intelligence and Data Analytics:** applying BI/BA within the enterprise, communicating the value, building expertise, delivering shared services, exploring big data, data analytics
5. **Legacy Modernization:** enhancing, renovating, replacing, legacy platforms and applications, business process improvement
6. **Enterprise Vision and Roadmap for IT:** vision and roadmap for IT, recognition by administration that IT is a strategic capability, integrating and influencing strategic planning and visioning with consideration of future IT innovations, aligning with Governor's policy agenda
7. **Budget and Cost Control:** managing budget reduction, strategies for savings, reducing or avoiding costs, dealing with inadequate funding and budget constraints
8. **Human Resources/Talent Management:** human capital/IT workforce, workforce reduction, attracting, developing and retaining IT personnel, retirement wave planning, succession planning, support/training, portal for workforce data and trends
9. **Agile and Incremental Software Delivery:** iterative design and incremental development of software solutions, allows for design modifications, prototyping and addition of new capabilities as part of the development process
10. **Disaster Recovery/Business Continuity:** improving disaster recovery, business continuity planning and readiness, pandemic/epidemic and IT impact, testing

GOOD NEWS: DEMAND FOR DATA MINING

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



¹ Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

LES DATA MINERs ?

Top CIO priorities in 2023

- Adopting a single technology vision
- Preparing for next-gen technology services
- Putting cloud costs under a microscope
- Using the cloud for core business applications
- Adopting industry cloud to drive product innovation
- Developing a unified data management architecture
- Turning data into products
- Enabling sustainability with new tech and new programs



LES DATA MINERs ?

LinkedIn Learning

The Skills Companies Need Most in 2020



Top 5 Soft Skills

- ① Creativity
- ② Persuasion
- ③ Collaboration
- ④ Adaptability
- ⑤ Emotional intelligence



Top 10 Hard Skills

- ① Blockchain
- ② Cloud computing
- ③ Analytical reasoning
- ④ Artificial intelligence
- ⑤ UX design
- ⑥ Business analysis
- ⑦ Affiliate marketing
- ⑧ Sales
- ⑨ Scientific computing
- ⑩ Video production