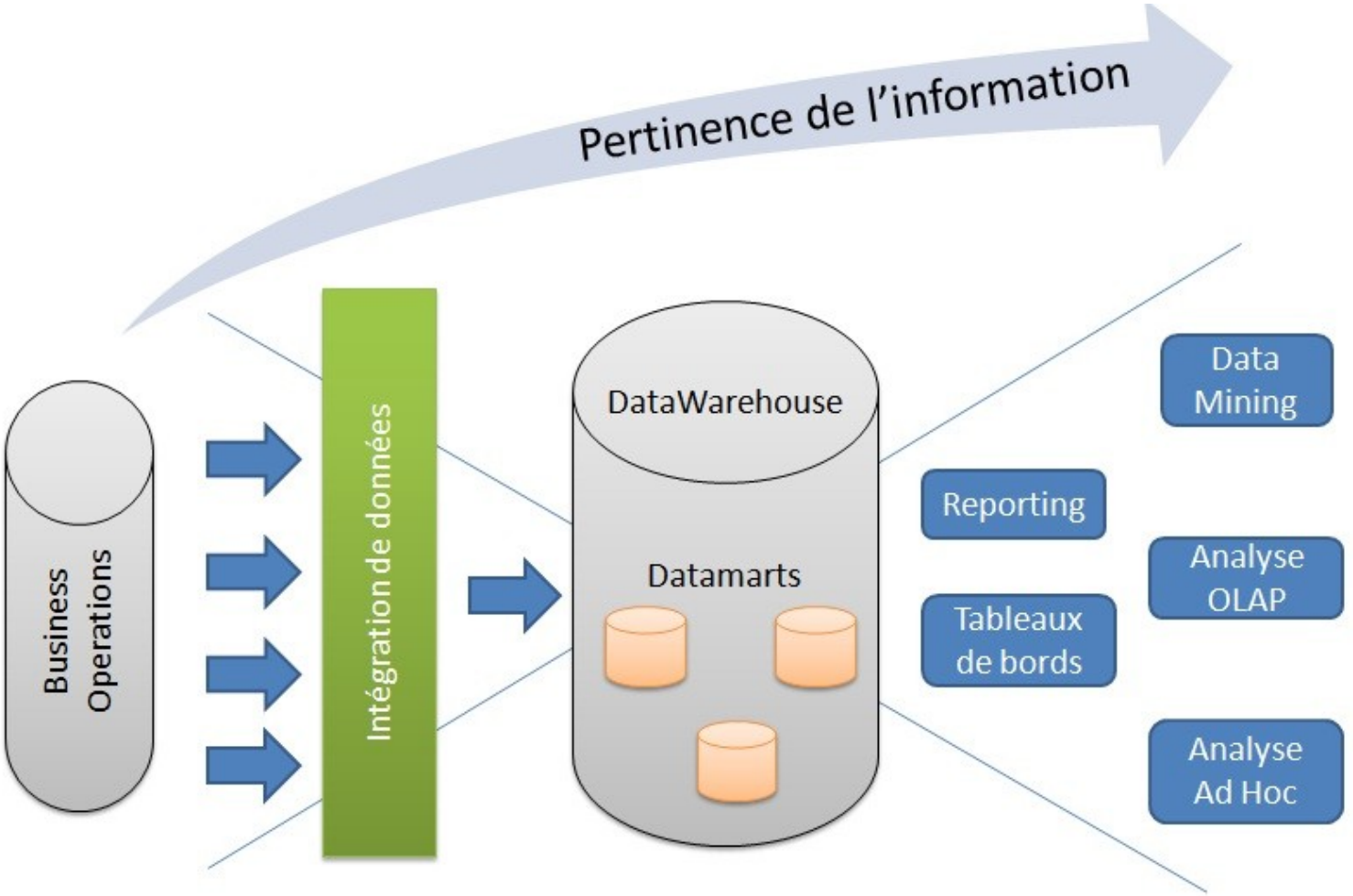
Technologies et Applications du Data Mining

TAYOU DJAMEGNI Clémentin
Professeur
Université de Dschang, IUT-FV de Bandjoun


Pertinence de l'information
Business Operations
Intégration de données
DataWarehouse
Datamarts
Data Mining
Reporting
Analyse OLAP
Tableaux de bords
Analyse Ad Hoc

Plan

Partie I : Informatique Décisionnelle (ID)
Business Intelligence (BI)
Partie II : Fouille de Données
Data Mining

Partie I : Business Intelligence

Plan de la première partie

1. Pilotage de l'entreprise et informatique décisionnelle - ID (Business Intelligence – BI)
2. De l'OLTP (On Line Transactional Processing) à l'OLAP (On Line Analysis Processing)
3. De la Business Intelligence (BI) à la Business Analytics (BA)
4. Un exemple de l'ID/BI dans l'entreprise : la relation client (Customer Relationship Management)

Pilotage de l’entreprise et informatique décisionnelle - ID (Business Intelligence - BI)

Pilotage de l'entreprise et informatique

- **Le pilotage** d'une entreprise dépend de ses **objectifs stratégiques**
- **Ce pilotage** doit prendre en **considération:**
  - une organisation de plus en plus **orientée clients**,
  - des **cycles conception/fabrication** de plus en plus **courts**,
  - de **nouveau canaux de distribution** comme les ventes en ligne sur le Web,
  - l'exigence **d'internationalisation**,
  - ...
- **Dans ce contexte, l'entreprise se doit :**
  - **d'anticiper les besoins des clients**,
  - de **contrôler** l'intégrité et la qualité des **flux de gestion**,
  - d'évaluer la **performance** des différentes entités la composant,
  - ...

Outils informatiques de pilotage des entreprises

- **Outils d'entreposage et d'analyse :**
  - pour **constituer** et **mettre à jour** à partir de diverses sources des « **réservoirs** » de grande quantités de données historisées et multidimensionnelles, …
  - pour en **extraire** selon divers critères des sous-ensembles de données,
  - pour les **analyser** selon différents axes (OLAP), d'**identifie**r des **tendances**, des **corrélations**, faire de la prévision (Data Mining).
- **Outils de veille stratégique :**
  - rattachés à **« l'intelligence économique »** (Competitive Intelligence)
  - pour la collecte sur le Web d'importante quantité de données, leur **filtrage** et en **extraire** les informations pertinentes (Web Mining) pour les analyser ensuite

=> Outils relevant de l'ID ou BI

Informatique décisionnel (ID/BI): Définition

- **L'Informatique Décisionnelle (ID), en anglais Business Intelligence (BI),** est l'informatique à l'usage des décideurs et dirigeants des entreprises
- En **management**, elle permet une **connaissance approfondie de l'entreprise** et la définition et le soutien de **stratégies d'affaires**, par exemple :
  - d'acquérir un avantage concurrentiel,
  - d'améliorer la performance de l'entreprise,
  - de répondre plus rapidement aux changements,
  - d'augmenter la rentabilité, et
  - d'une façon générale la création de valeur ajoutée de l'entreprise.
- Les techniques de ID/BI sont utilisées aussi dans **d'autres domaines que le management** : santé, transport, éducation, énergie, télécommunication, sciences, ...

La pyramide de l'ID/BI

L’ID/BI est crutiale et en pleine croissance

- **Le Web rend l’ID/ BI encore plus nécessaire :**
  - les clients ne sont **pas «physiquement» dans le magasin**
  - les clients peuvent **changer à d'autres magasins plus facilement**
  - comment **connaître ses clients?**
    - analyser les « Web log » pour comprendre le comportement des clients sur le site
    - combiner ces données Web avec les données traditionnelles des clients
- **« Internet sans fil » ajoute à cela :**
  - les **clients sont toujours "en ligne"**
  - la position de la clientèle est connue
  - combiner la position et la connaissance sur le client => très utile

De l'OLTP (On Line Transactional Processing) à l'OLAP (On Line Analysis Processing)

Systèmes d'information opérationnels : OLTP

- Permettent des processus de **traitement en ligne des données – OLTP (On line Transactionnal Processing) :** Interactifs, Concurrents, Nombreux, Répétitifs, Structurés, Simples
- **Supportent en général une ou plusieurs grandes fonctions de l'entreprise** (production, marketing, commercial, ressources humaines, finance, comptabilité, recherche, …)
- Parfois intégrés dans un ERP(Enterprise Ressource Planning traduit Progiciel de Gestion Intégré(PGI)), ils s'appuient sur des **SGBD traditionnels** (Oracle, DB2, ...) pour gérer des **BD** «**opérationnelles**» ou de « **production** » (Mega-Giga octets)
- **Ces processus OLTP concernent :**
  - la mise à jour de données
  - un nombre restreint d'enregistrements
  - des données précises et à jour

Exemple : un supermarché ENREGISTRANT ses ventes

- -

Limites des BD opérationnelles pour la BI

- **BD opérationnelles complexes et inutilisables**
  - souvent difficiles à comprendre
  - ne concernent pas un objectif (sujet) unique d'affaire
- **Données** des BD opérationnelles :
  - identiques dans différentes BD
  - même concept souvent défini différemment
  - adaptées pour les systèmes opérationnels (comptabilité, facturation, ...), pas pour l'analyse des fonctions d'affaires
  - de qualité mauvaise : données manquantes, données imprécises, ...
  - volatiles :
    - supprimées périodiquement dans les systèmes opérationnels (6 mois)
    - modifiées au fil du temps - aucune information historique
-
- -

Nouvelles attentes des SI

- Considérer des **quantités de données HISTORISEES de plus en plus importantes** (Tera, Penta octets), **organisées selon différentes dimensions** (temps, espace géographique, gammes de produit, ...) stockées dans des **ENTREPOTS DE DONNEES**
- Passer du **TRAITEMENT transactionnel en ligne des données** (OLTP) à **l'ANALYSE EN LIGNE** (On Line Analysis Processing - OLAP) de ces entrepôts selon différentes dimensions pour le pilotage de l'entreprise
- Pour prendre de « **bonnes décisions** », il faut accéder en temps réel à ces données, les analyser pour en extraire l'information pertinente, par exemple pour savoir :savoir :
  - Quels sont les résultats des ventes par gamme de produit et par région pour l'année dernière ?
  - Quelle est l'évolution des chiffres d'affaires par type de magasin et par période ?
  - Comment qualifier les acheteurs de mon produit X ?

=> Informatique Décisionnelle - Business Intelligence

Nouvelle technologie informatique de l'ID/BI

Entrepôt de données (Data Warehouse) :

- il récolte, **stocke et gère efficacement des gros volumes de données** pour la prise de décision,
- les données y sont organisées dans des regroupements **homogènes** selon plusieurs **axes d'analyse** et différents **niveaux de détail** (**d'agrégation**).

Analyse en ligne des données OLAP (On Line Analytical processing) :

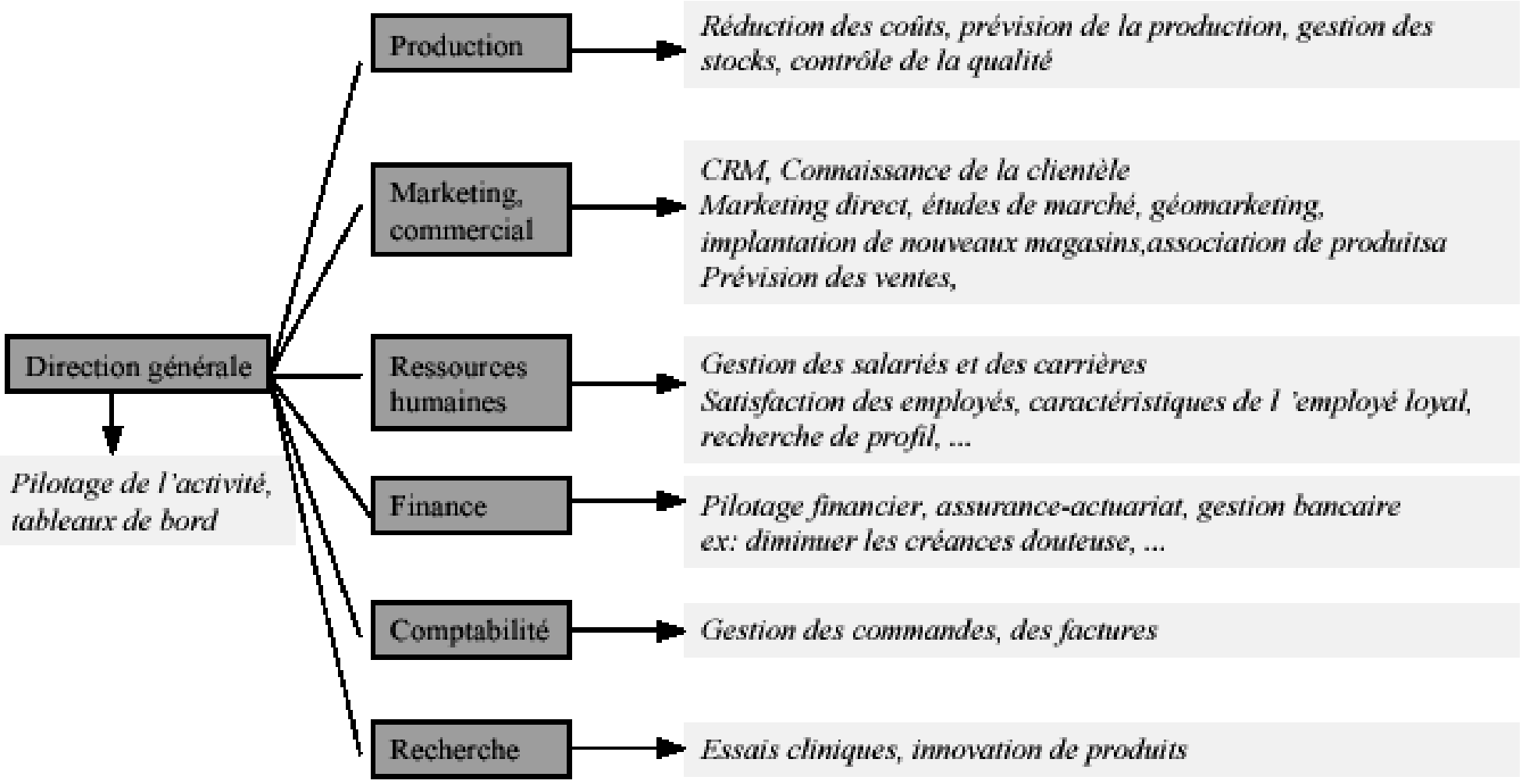
- exploitation d'entrepôts de données permettant **interactivement** de conduire des analyses par **changement de points de vue, de niveau de détail** (**agrégations**),

Fouille de données (Data Mining):

- **extraction automatique de connaissances** (propriétés cachées) dans de grands volumes de données,
- **par des techniques traditionnelles** issues des statistiques et de l'analyse de
- Données,
- **par des techniques plus récentes** issues de l'Intelligence Artificielle (IA)

Tendance à une **intégration croissante des techniques de fouille** dans les **entrepôts de données** (offre commerciale)

L'ID/BI dans l'entreprise : Domaines concernés



De la Business Intelligence (BI) à la Business Analytics (BA)

De la Business Intelligence à Business Analytics (1)

Business Intelligence (BI) :

- s'intéresse à **ce qui s'est passe**
- basée sur les entrepôts de données, l'analyse en ligne (OLAP), le reporting, la
- surveillance et l'alerte automatisées, les tableaux de bord ...

  répondre à des questions telles que : que s'est-il passé , combien, à quelle fréquence, où se situe le problème et quelles sont les actions nécessaires.

- **Business Analytics (BA) :**
  - s'intéresse **aux raisons** pour lesquelles cela s'est produit et si cela se reproduira, en s'appuyant sur la Business Intelligence
  - basée sur l'analyse statistique et quantitative, l'exploration de données, la modélisation prédictive …

    => répondre à des questions telles que : pourquoi cela se produit ?, que se passe-t-il si ces tendances se poursuivent ?, que se passera-t-il ensuite ? (prédictions) et quel est le meilleur résultat possible ? (optimisation).

De la Business Intelligence à Business Analytics (2)

Analyse descriptive & diagnostic analytique (Descriptive Analytics & Diagnostic Analytics) :

- **Comprendre** les données historiques grâce à des rapports, des tableaux de bord, des regroupements : **qu'est-il arrivé ? pourquoi c'est arrivé ?**
  Ex : *changements de prix d'une année à l'autre, croissance des ventes d'un mois à l'autre, le nombre d'utilisateurs ou le revenu total par abonné .*
  => *mesures décrivant tout ce qui s'est produit dans une entreprise pendant une période donnée.*

Analyse prédictive (Prédictive Analytics) :

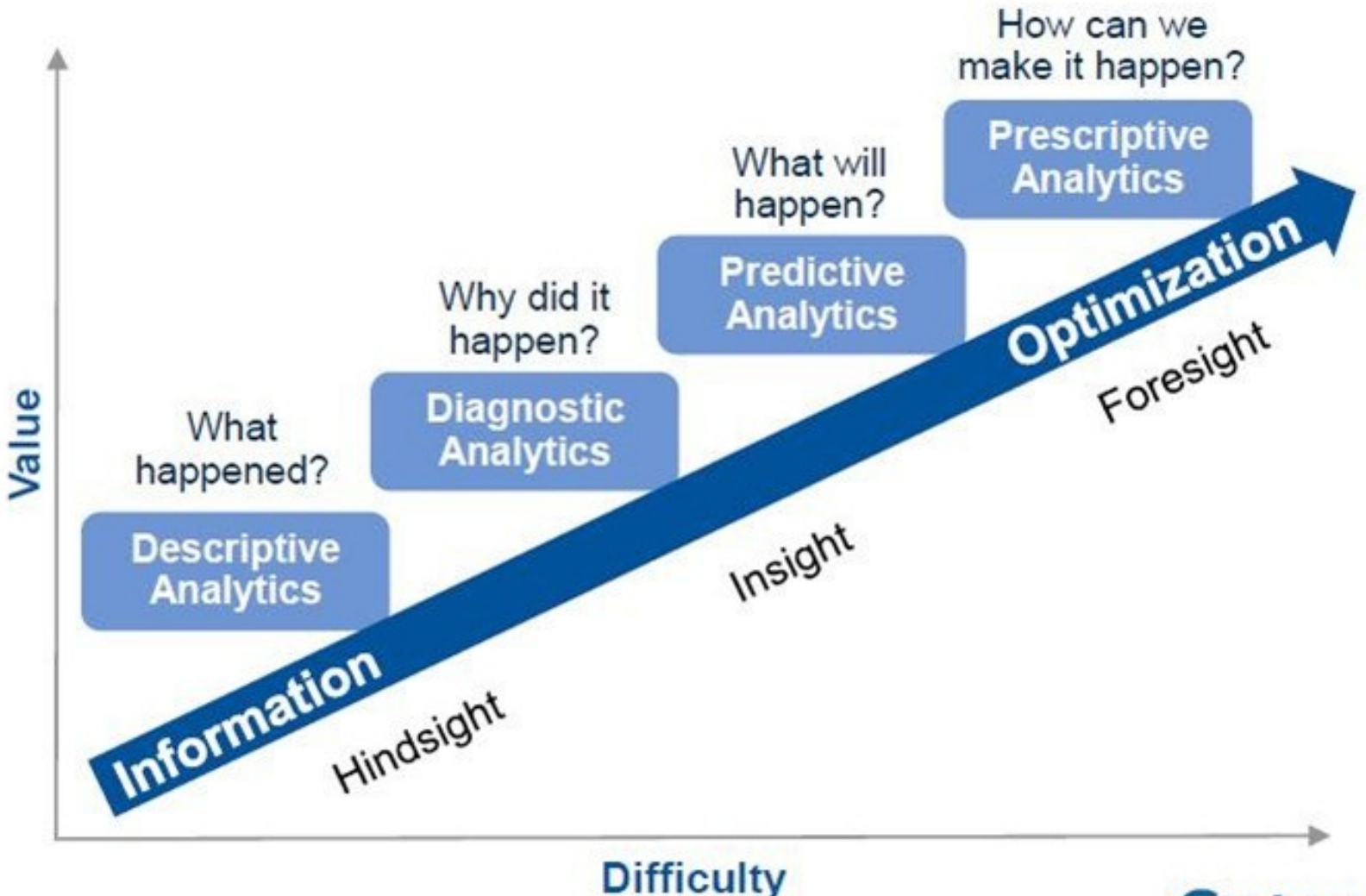
- **Prédire** l'avenir en examinant des données historiques, en détectant des modèles ou des relations dans ces données, puis en extrapolant ces relations dans le temps : quand cela peut arriver ?
- Utilise la modélisation prédictive à l'aide de techniques statistiques et d'apprentissage automatique.,

Analyse prescriptive (Prescriptive Analytics) :
Comment cela peut-il se reproduire ?

- **Identifier les meilleures alternatives** pour minimiser ou maximiser un objectif
- **Recommander** des décisions en utilisant **l'optimisation**, la **simulation**,

Étapes de la Business Analytics

Valeur en fonction du degré d'analyse



Un exemple de l'ID/BI dans l'entreprise : la relation client (Customer Relationship Management)

Un exemple de l'ID/BI dans l'entreprise : la relation client (1)

Contexte économique général :

Dans un climat de concurrence mondiale tendu :

- **Conquérir un nouveau client** coûte **5 fois plus cher que de fidéliser** un client existant
- **5% d'amélioration de la fidélité** des clients entraîne une **augmentation des profits de 10 à 15%**
- ***Tous les clients ne sont pas égaux :*** *30% des clients génèrent 70% du CA*
- Le Client attend un **service personnalisé**, sur-mesure
- La **personnalisation est une source de profit**

=> Gestion de la relation client
(Customer Relationship Management – CRM)

Un exemple de l'ID/BI dans l'entreprise : la relation client (2)

Définition du CRM :

- capacité à **identifier**, à **acquérir** et à **fidéliser** les meilleurs clients dans le **but d'augmenter le chiffre d'affaires et les bénéfices**.
- capacité à **bâtir une relation profitable sur le long terme avec les meilleurs clients** en capitalisant sur l'ensemble des points de contacts

Mieux connaître et comprendre ses clients pour :

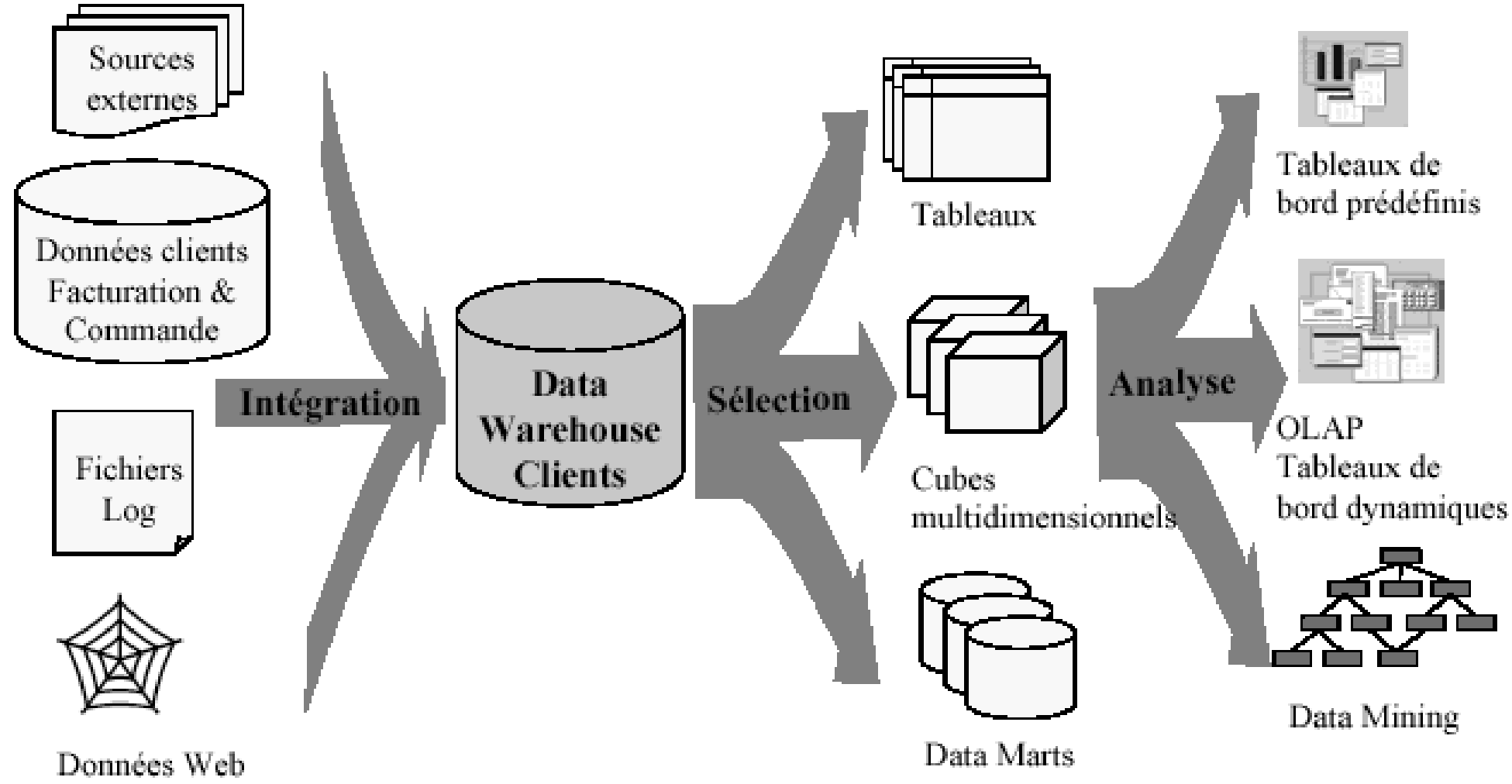
- Réduire les coûts (clients infidèles)
- Comment fidéliser une clientèle ?
- Comment augmenter les profits ?
- Comment identifier les nouvelles opportunités ?

Doit permettre par exemple de répondre aux questions :

- Quels sont les besoins et les attentes des clients? Comment y répondre?
- Quels sont les clients prêts à acheter de nouveaux produits?
- Quels sont les clients les plus profitables, fidèles et pourquoi?
- Quels sont les clients mécontents, et pourquoi?

=> Informatique Décisionnelle - Business Intelligence

Un exemple de l'ID/BI dans l'entreprise : la relation client (3)



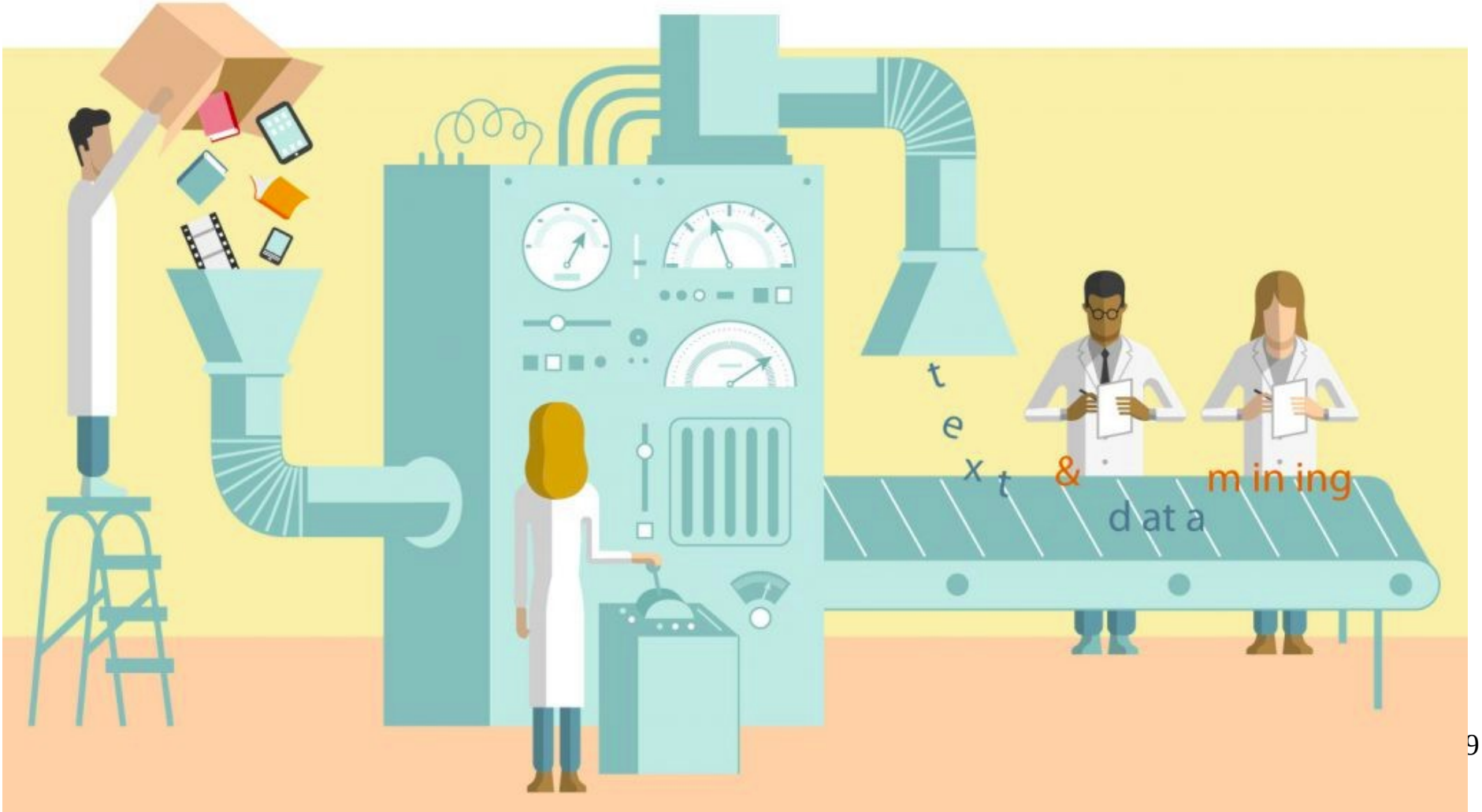
Partie II : Data Mining

Plan de la deuxième partie

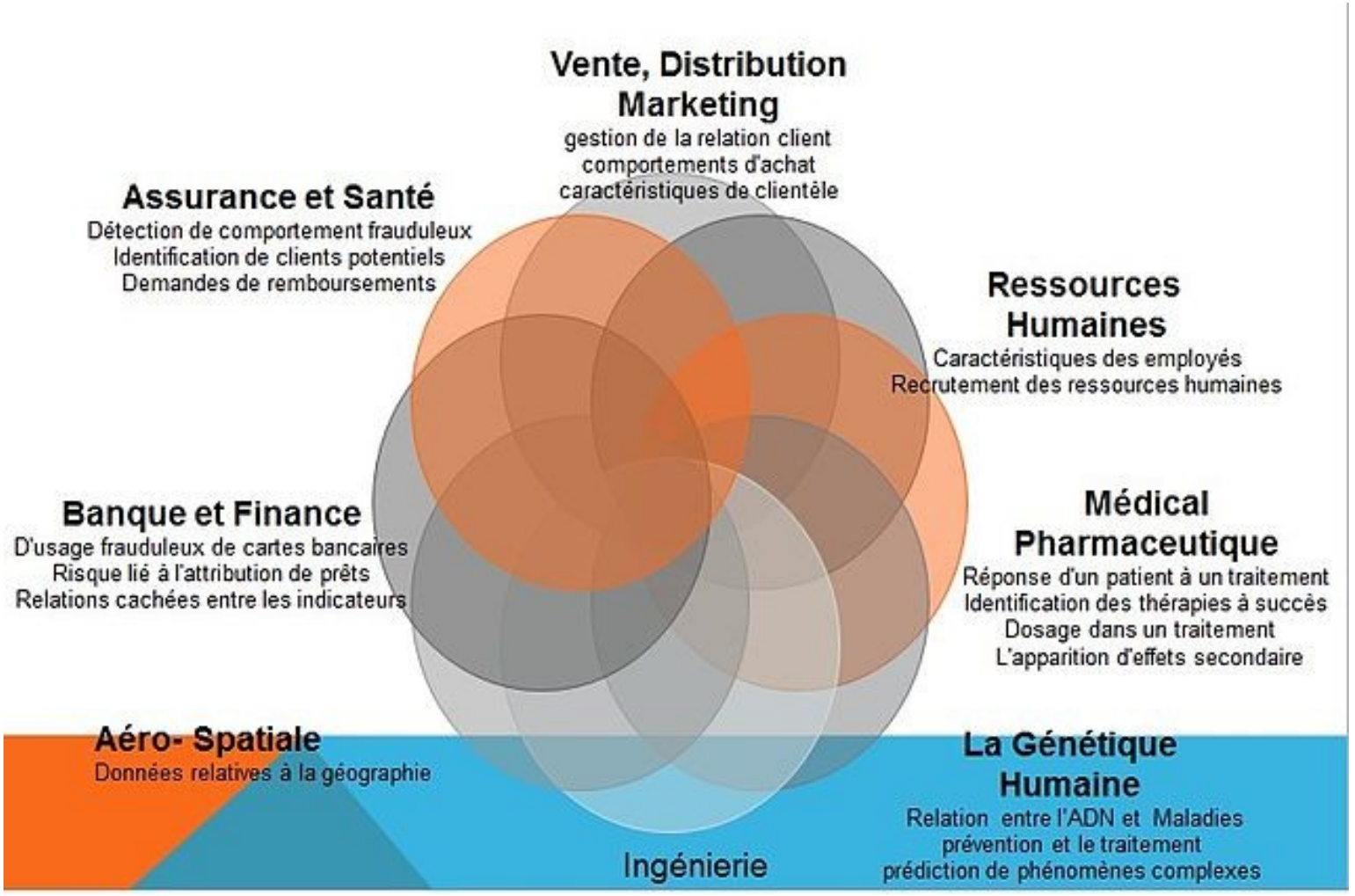
1. Qu’est ce que le data mining ?
2. Technologie de classification : arbre de décision
3. Technologie de classification : K pus proches voisins
4. Technologie de classification : Support Vector Machine (SVM)
5. Technologie d’apprentissage : réseaux de neurones
6. Clustering/segmentation des données
7. Mesure de la qualité de l’apprentissage
8. Découverte des motifs et des règles d’association
9. Quelques exploits concrets du machine learning

Qu’est ce que le Data Mining ?

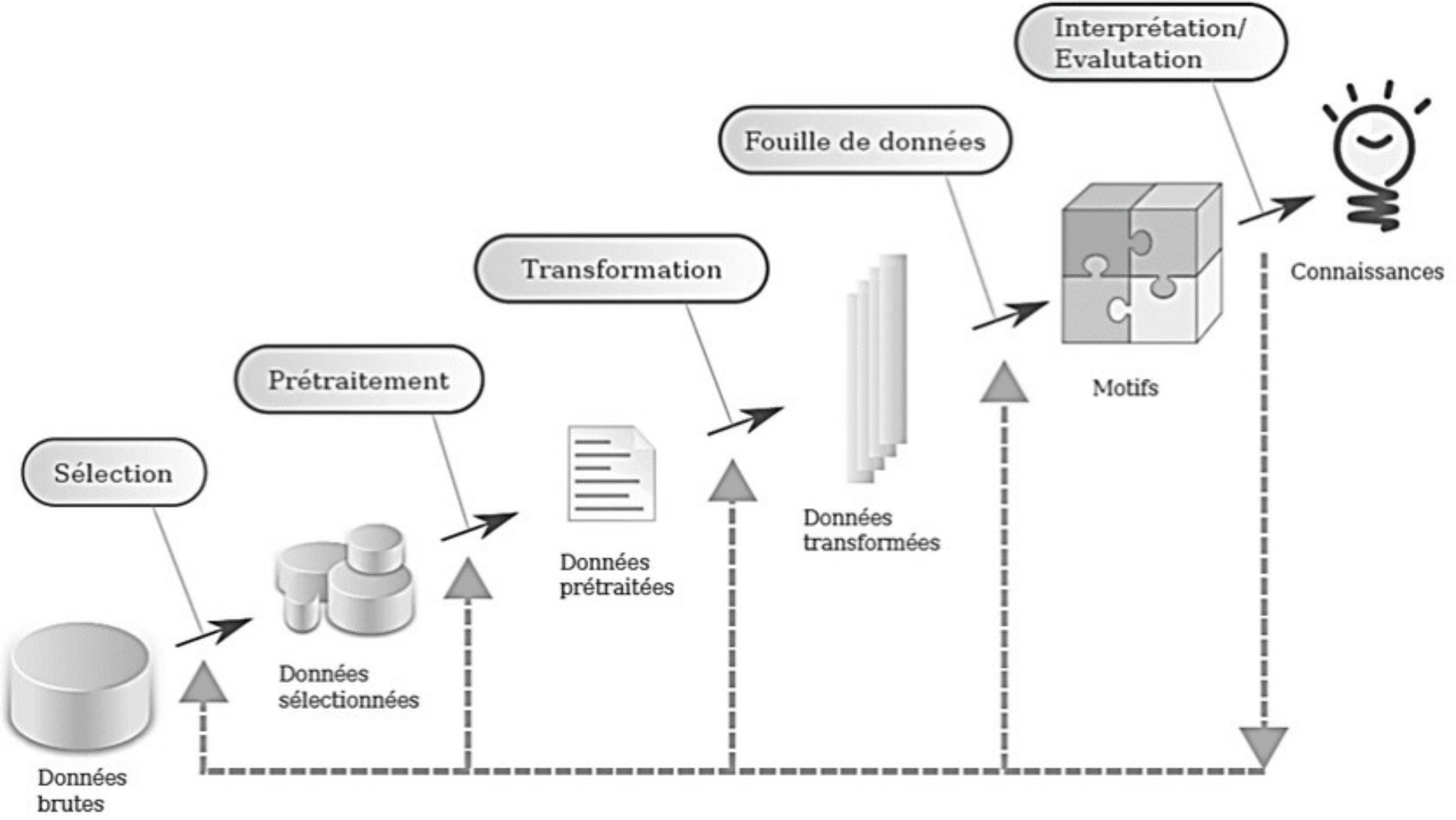
Objectif du Data Mining



Domaines d'Application du Data Mining



Processus d'Extraction des Connaissances



Comparaison : Data Mining et Statistiques (1)

N°	Data mining	Statistiques
1	Le Data mining est le début de la science des données et elle couvre l'ensemble du processus d'analyse des données	Les statistiques sont la base du data mining et constituent la partition principale des algorithmes de data mining
2	Explore et rassemble des donnée qui sont utilisées pour construire un modèle de détection des motifs et faire des théories	Confirme ou infirme des hypothèses. On fournit une théorie à tester à l'aide de statistiques
3	**Les données utilisées sont numériques ou non numériques**	**Les données utilisées sont numériques**

Comparaison : Data Mining et Statistiques (2)

N°	Data mining	Statistiques
4	**processus inductif (Génération d'une nouvelle théorie à partir de données)**	**Processus déductif (aucune prédiction)**
5	**Le nettoyage des données est effectué dans le data mining**	**Les méthodes statistiques sont appliquées sur des données propres (cleaned data)**
6	**La collecte de données est moins importante :** l'exploration de données est un processus de découverte des motifs dans de grands ensembles de données	**La collecte de données est plus importante :**Les données collectées peuvent être des données quantitatives, qualitatives

Comparaison : Data Mining et Statistiques (3)

N°	Data mining	Statistiques
7	Nécessite moins d'interaction avec l'utilisateur pour valider le modèle, donc **facile à automatiser**	Nécessite plus d'interaction avec l'utilisateur pour valider le modèle, donc **difficile à automatiser**
8	**Convient aux grands ensembles de données**	**Convient aux petits ensembles de données**
9	C'est un algorithme qui apprend des données sans aucune règle de programmation	Formalisation de relation dans les données sous forme d'équation mathématique

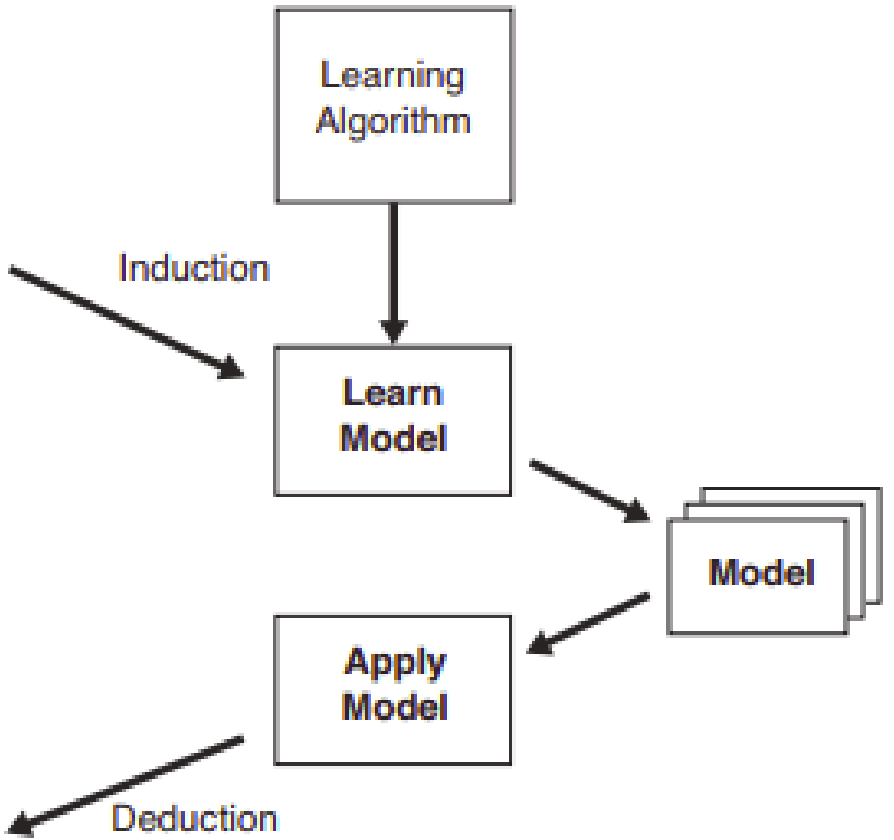
Comparaison : Data Mining et Statistiques (4)

N°	Data mining	Statistiques
10	**Utiliser la pensée heuristique (règles utilisées pour former des jugements et prendre des décisions)**	**N'a pas de place pour la pensée heuristique**
11	Classification, segmentation, réseaux de neurones, règles d'association, motifs, visualisation	Statistique descriptive statistique inférentielle

Processus d'Apprentissage

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes



Test Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Technologie de Classification : Arbre de Décision

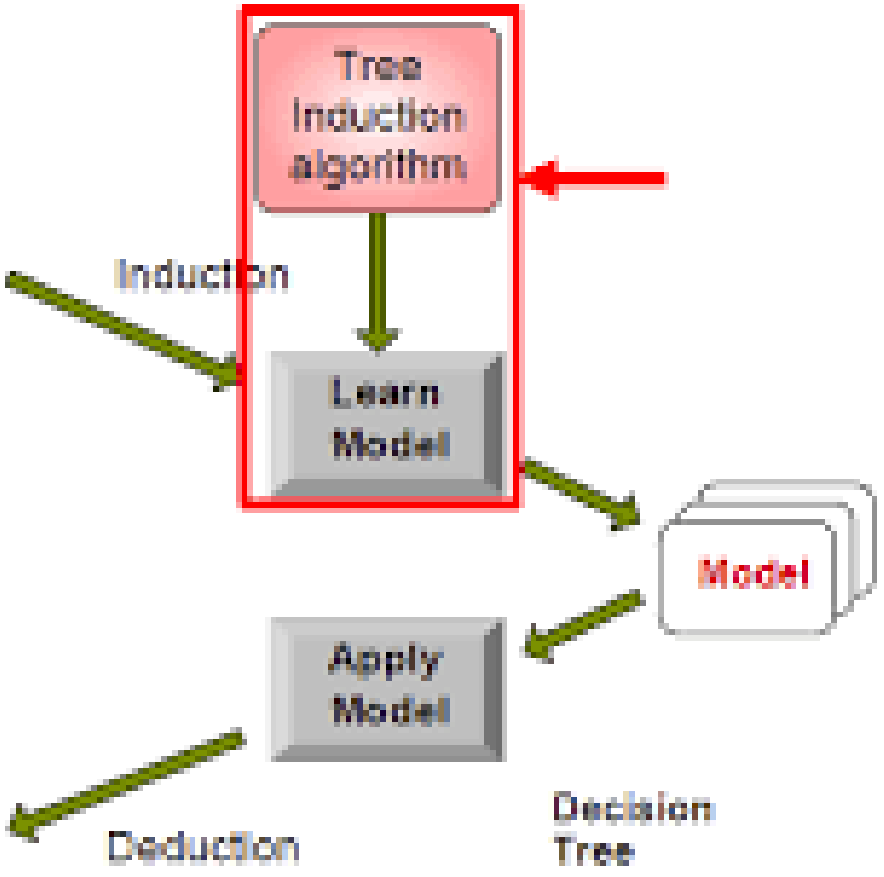
Classification : Arbre de Décision

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

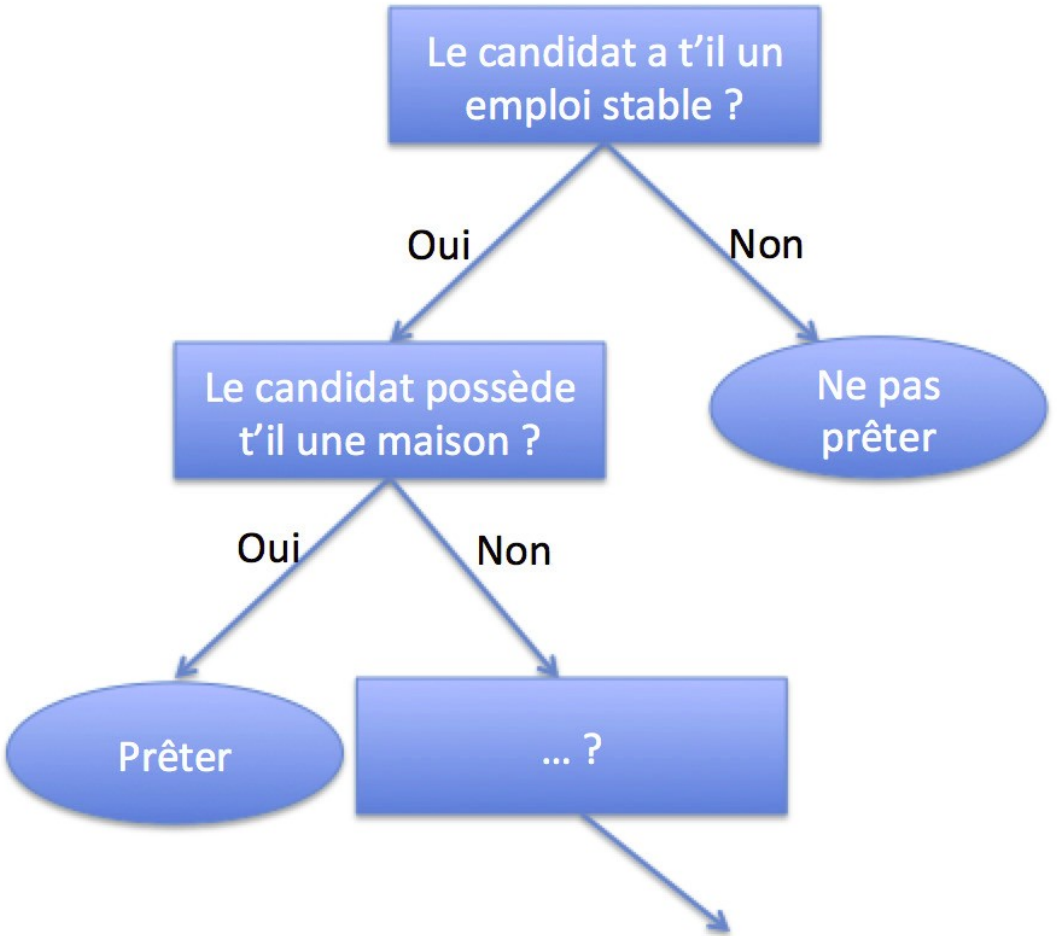
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



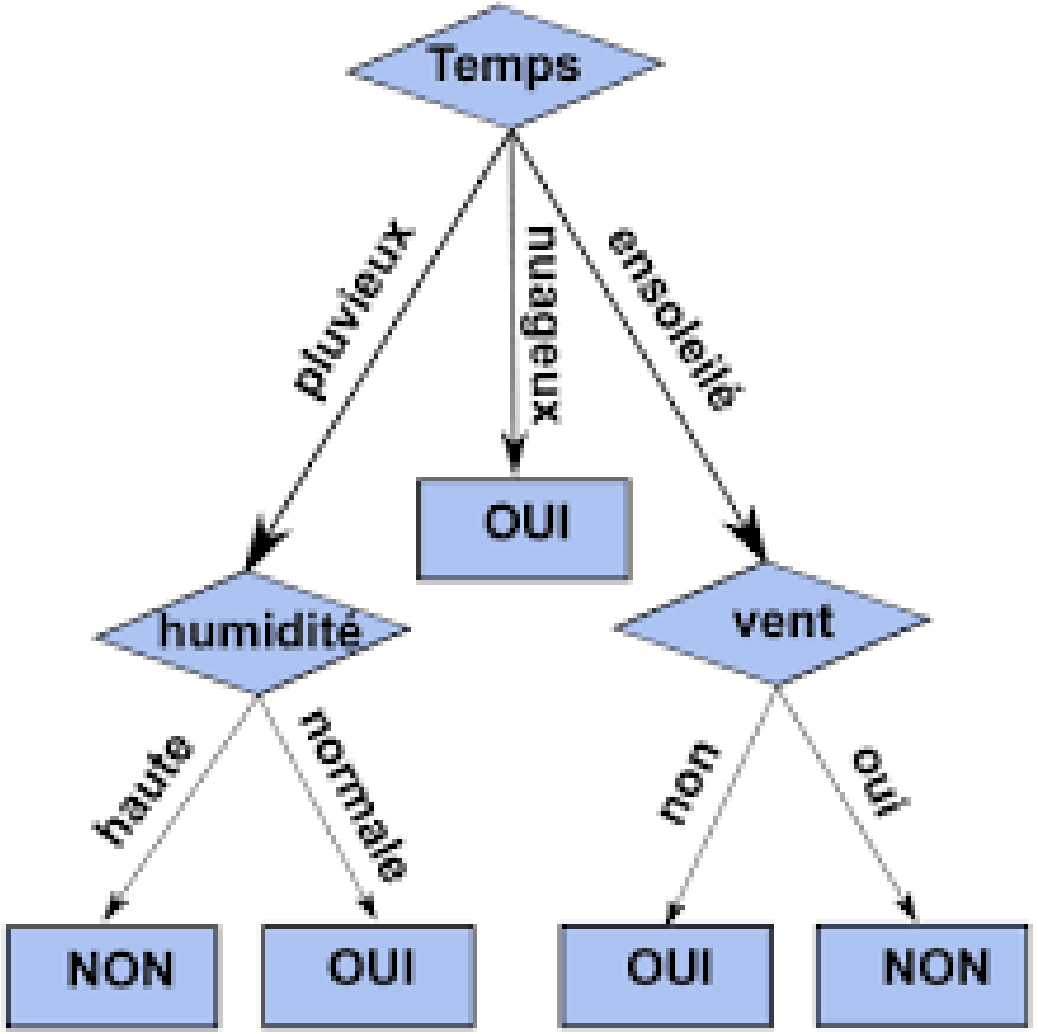
Classification : Arbre de Décision pour l'Accord d'un Crédit



Exemple de Données : Peut-on Jouer au Tennis ?

Day	Outlook	Temperature	Humidity	Wind	Play ball
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Exemple d'Arbre : Peut-on Jouer au Tennis ?



Technologie de Classification : K Plus Proches Voisins

K plus proches voisins : Notion de Similarité

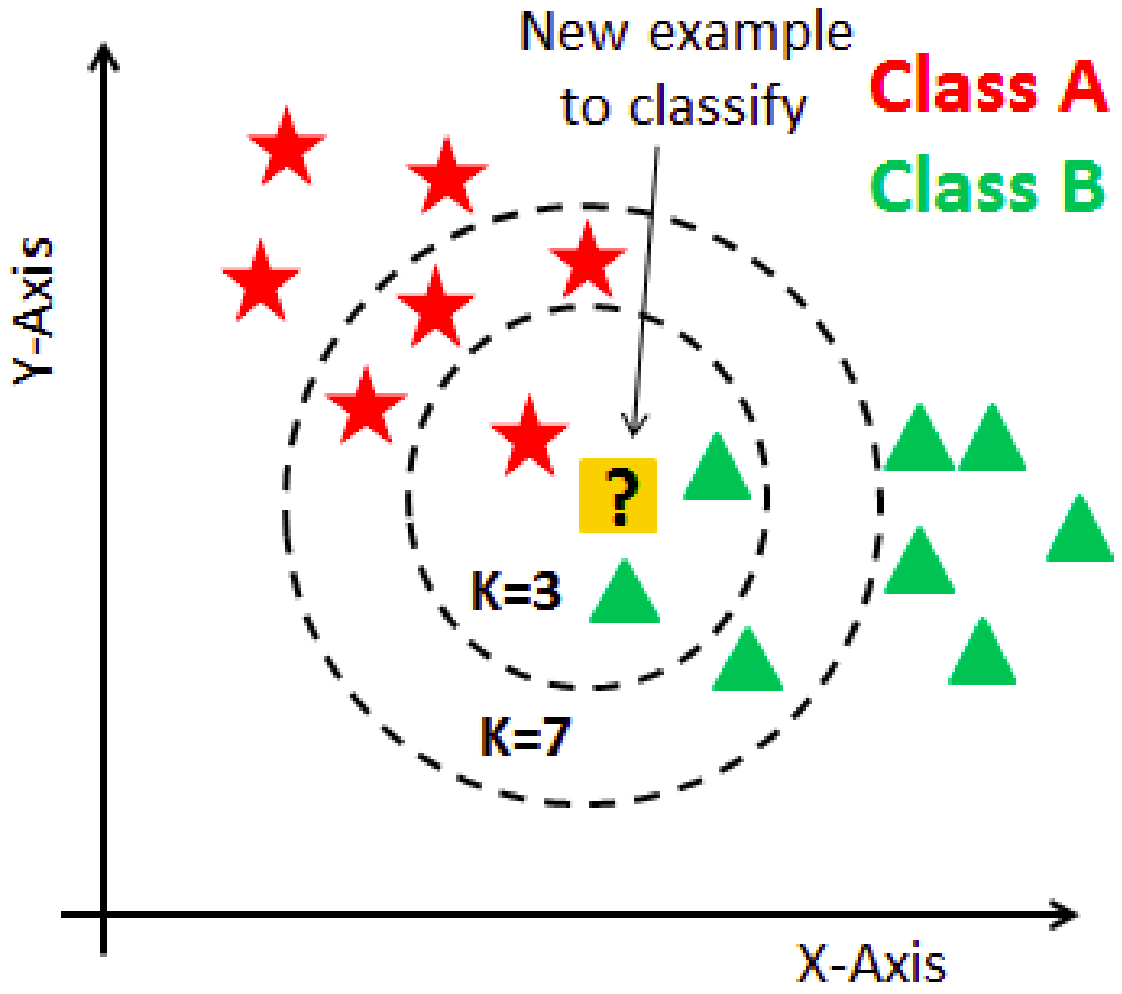
Certaines mesures s'expriment de manière ensembliste. Soient X et Y deux ensembles. **On note |Z| le nombre d'éléments d'un ensemble Z.** Toutes ces mesures de similarité ne conduisent pas à une métrique, elles ne respectent pas l'inégalité triangulaire (ex : le Cosinus)

Coefficient de Dice : **$dice(X , Y) = 2 |X \cap Y| / (|X| + |Y|)$**

Indice et distance de Jaccard ou de Tanimoto :
$jaccard(X , Y) = |X \cap Y| / |X \cup Y|$

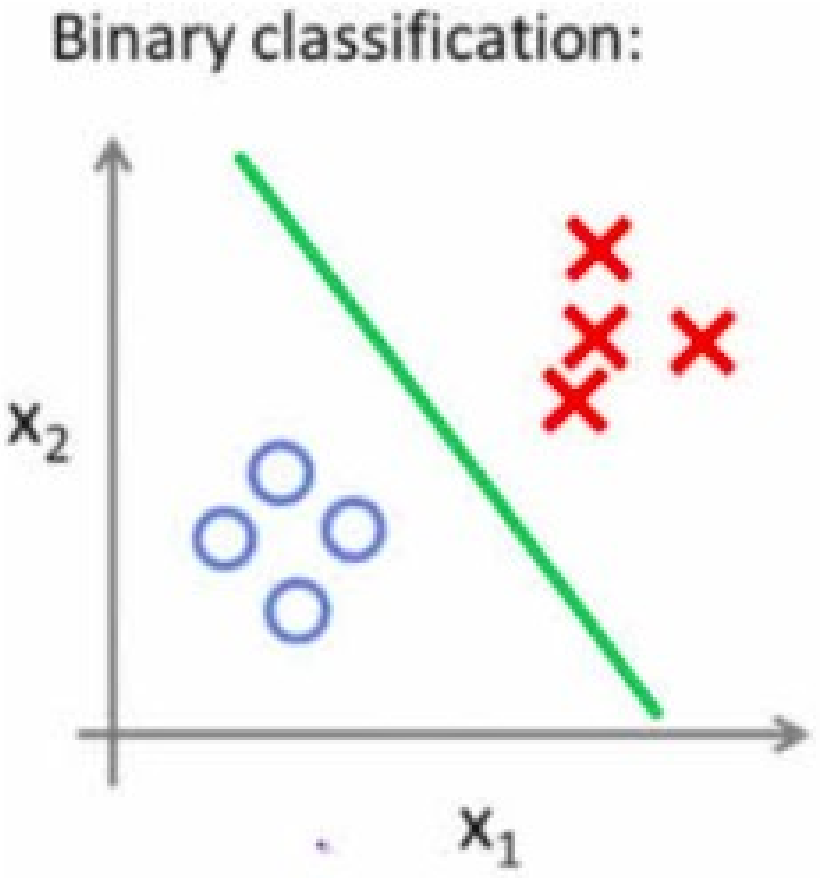
Coefficient de recouvrement
$recouvrement(X , Y) = |X \cap Y| / \min(|X| , |Y|)$

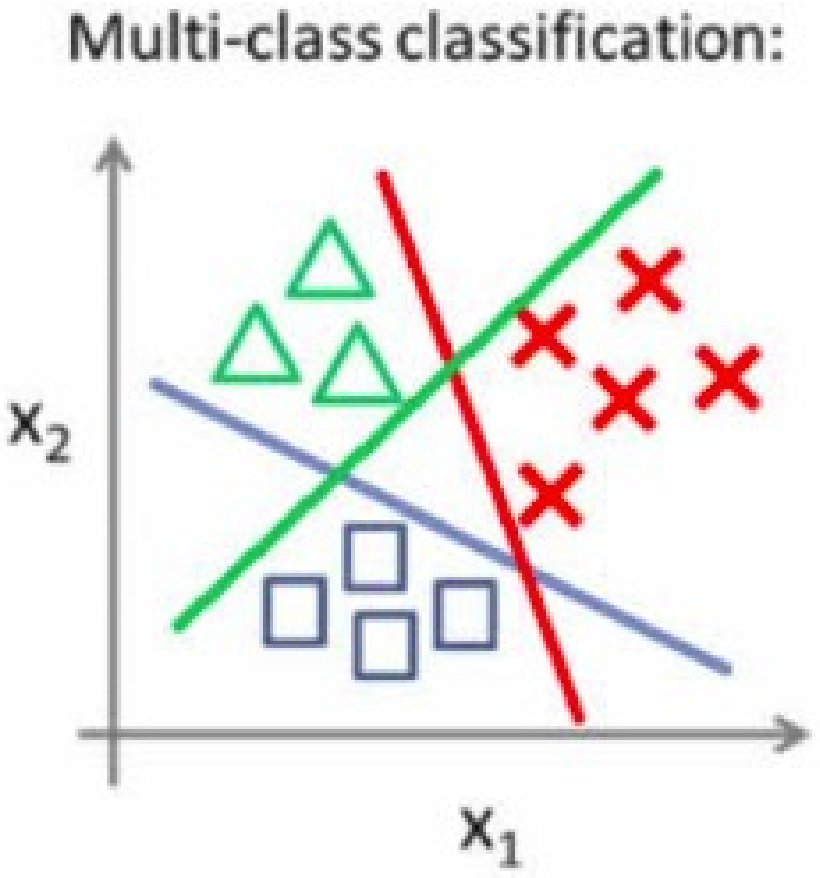
K plus proches voisins : Le Principe



Technologie de Classification : Support Vector Machine (SVM)

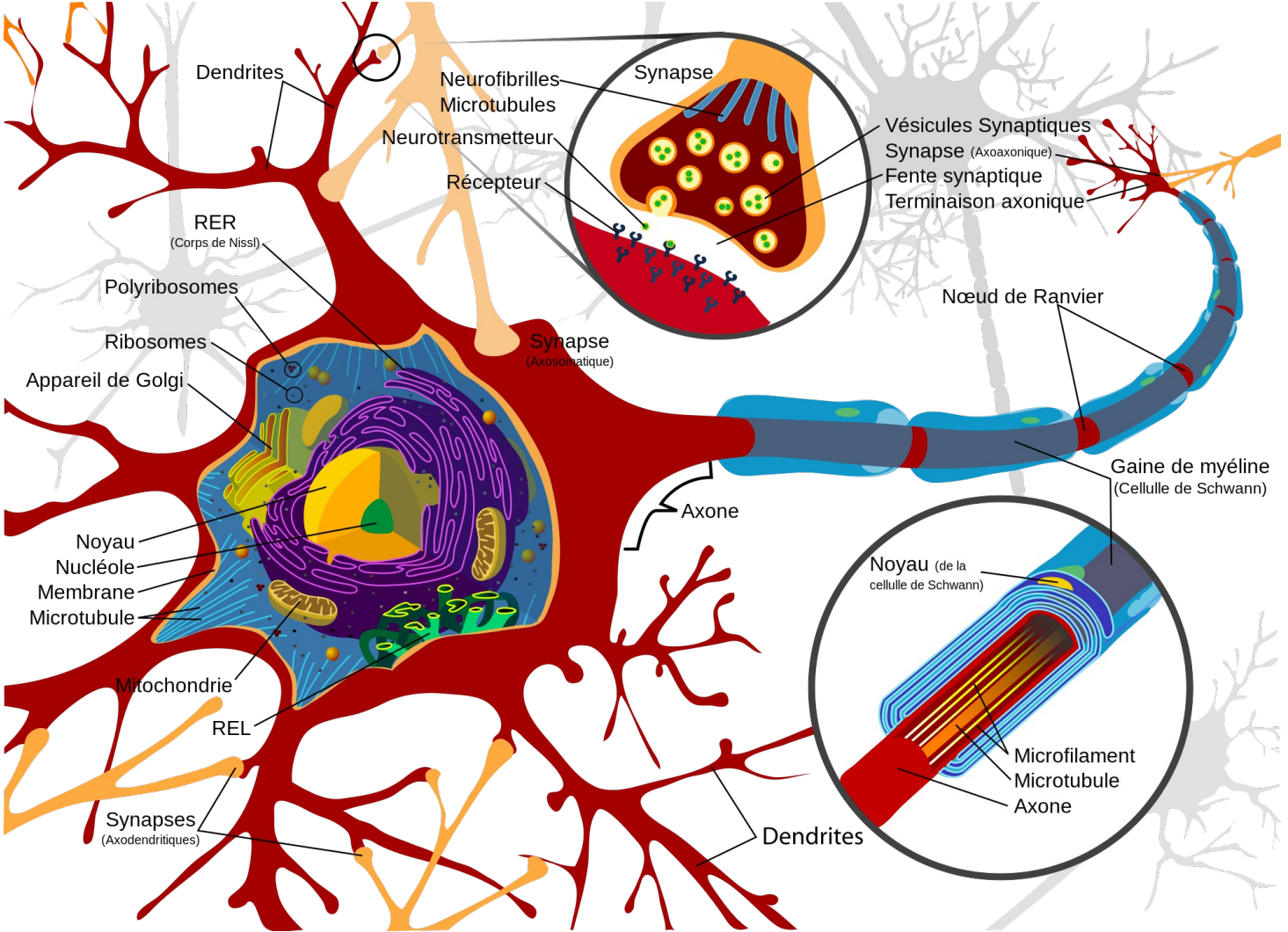
Support Vector Machine : Principe



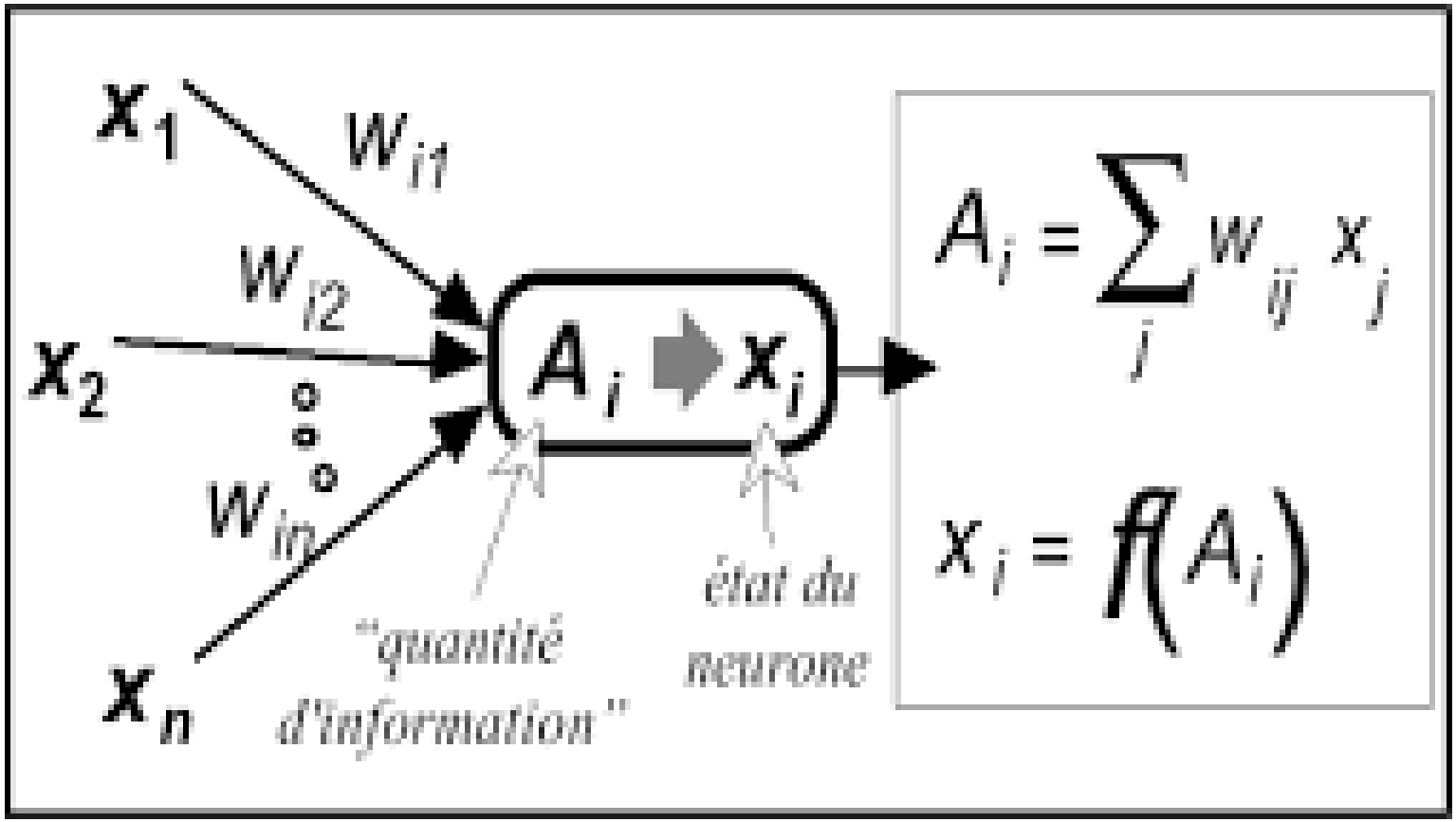


Technologie d'Apprentissage : Réseaux de Neurones

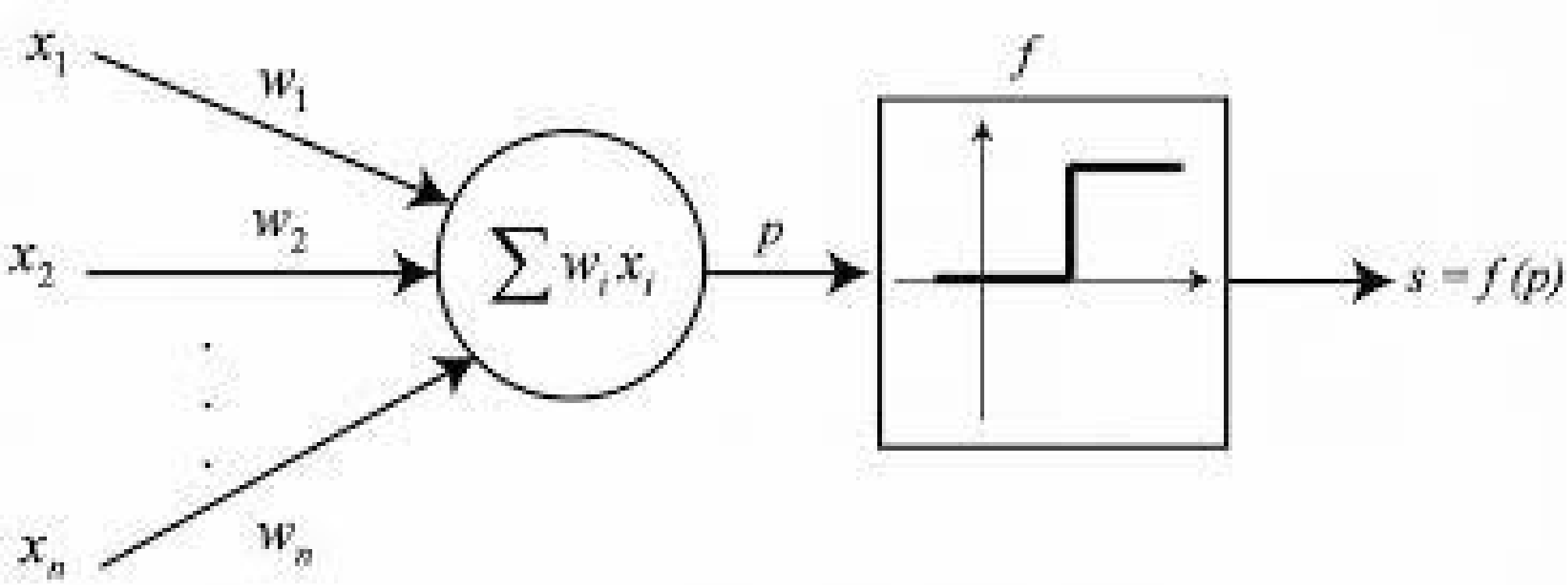
Les Neurones Naturels



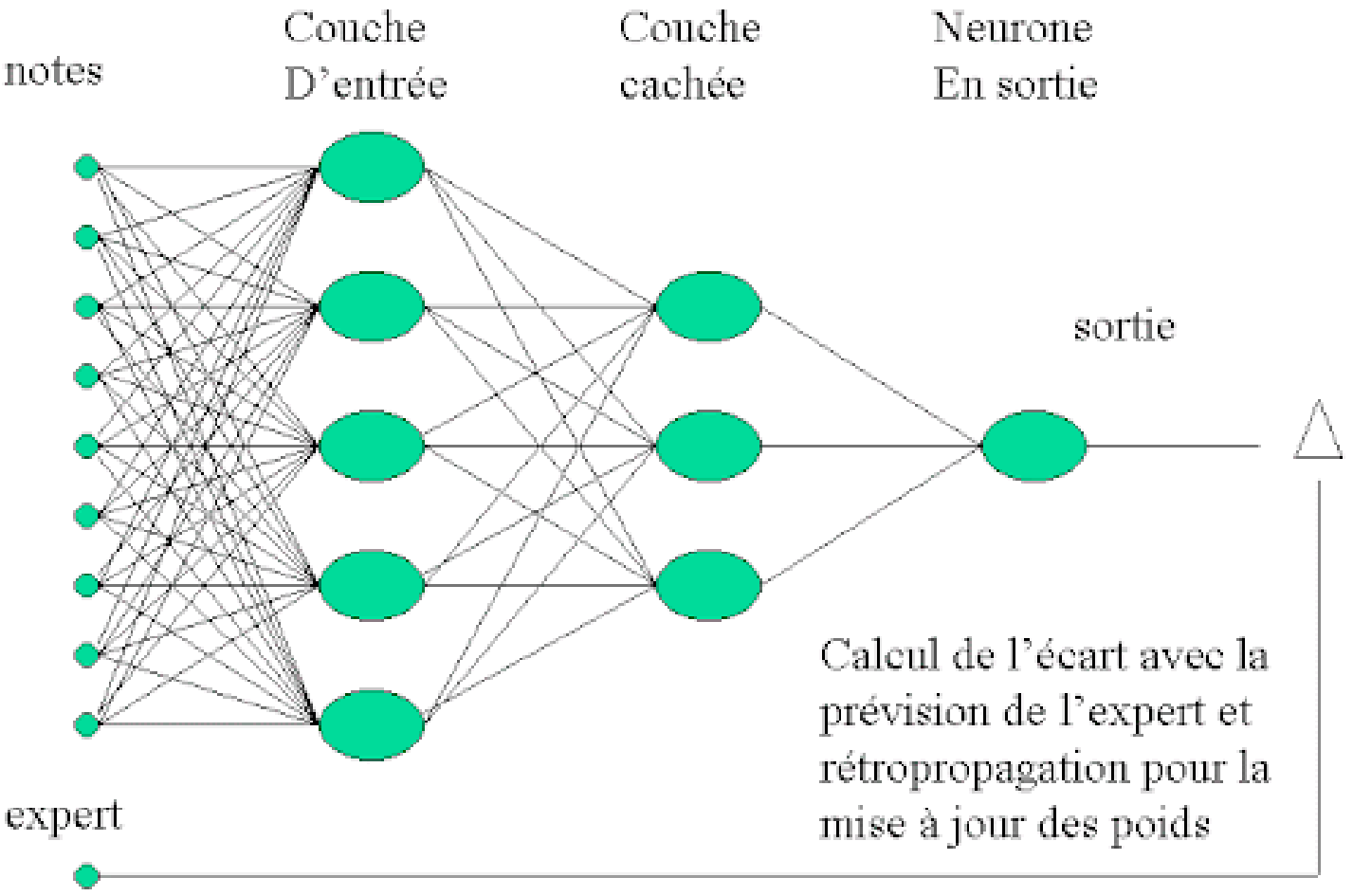
Le Neurone Formel : Principe de Fonctionnement



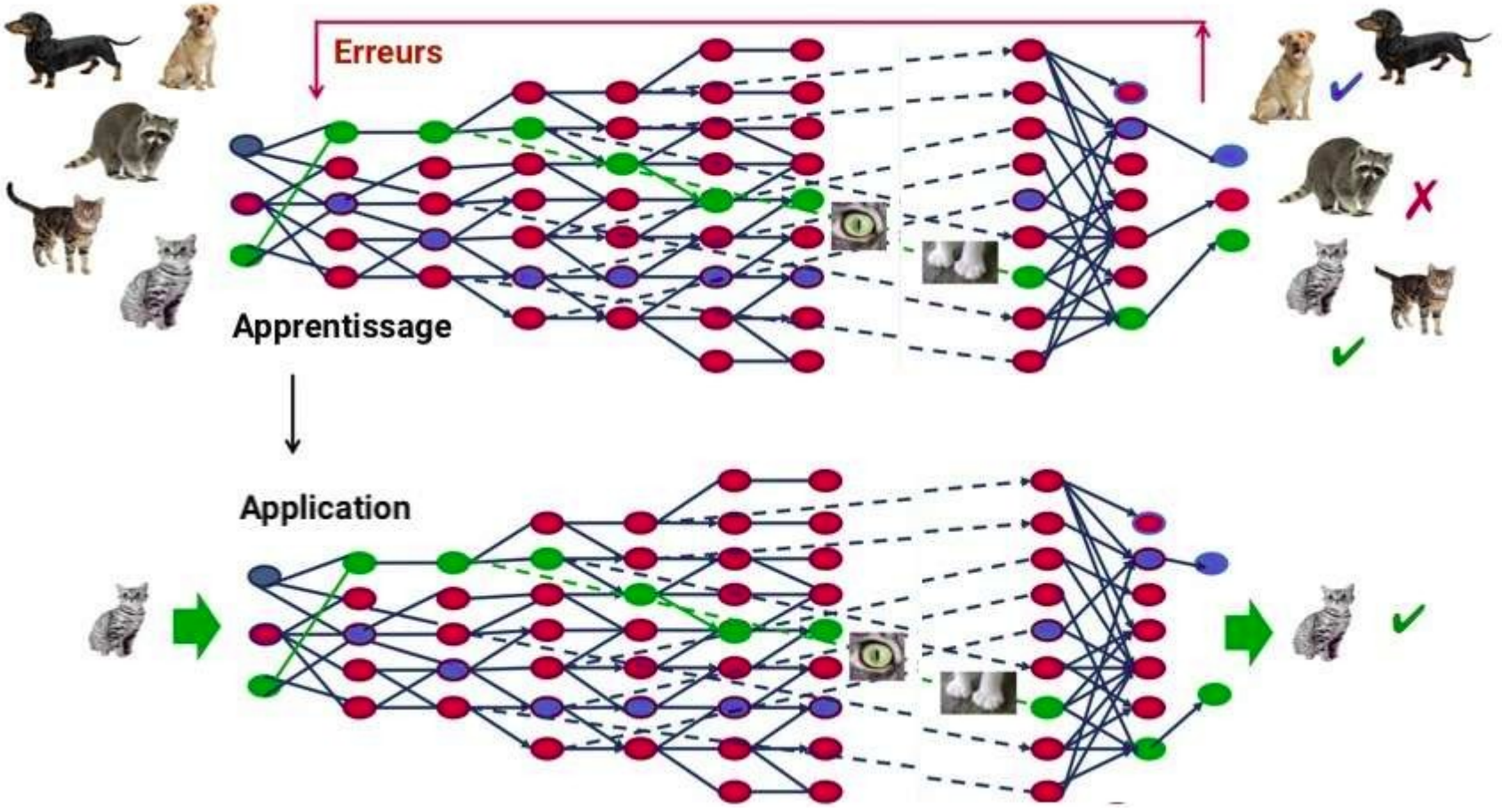
Le Neurone Formel : Un Exemple



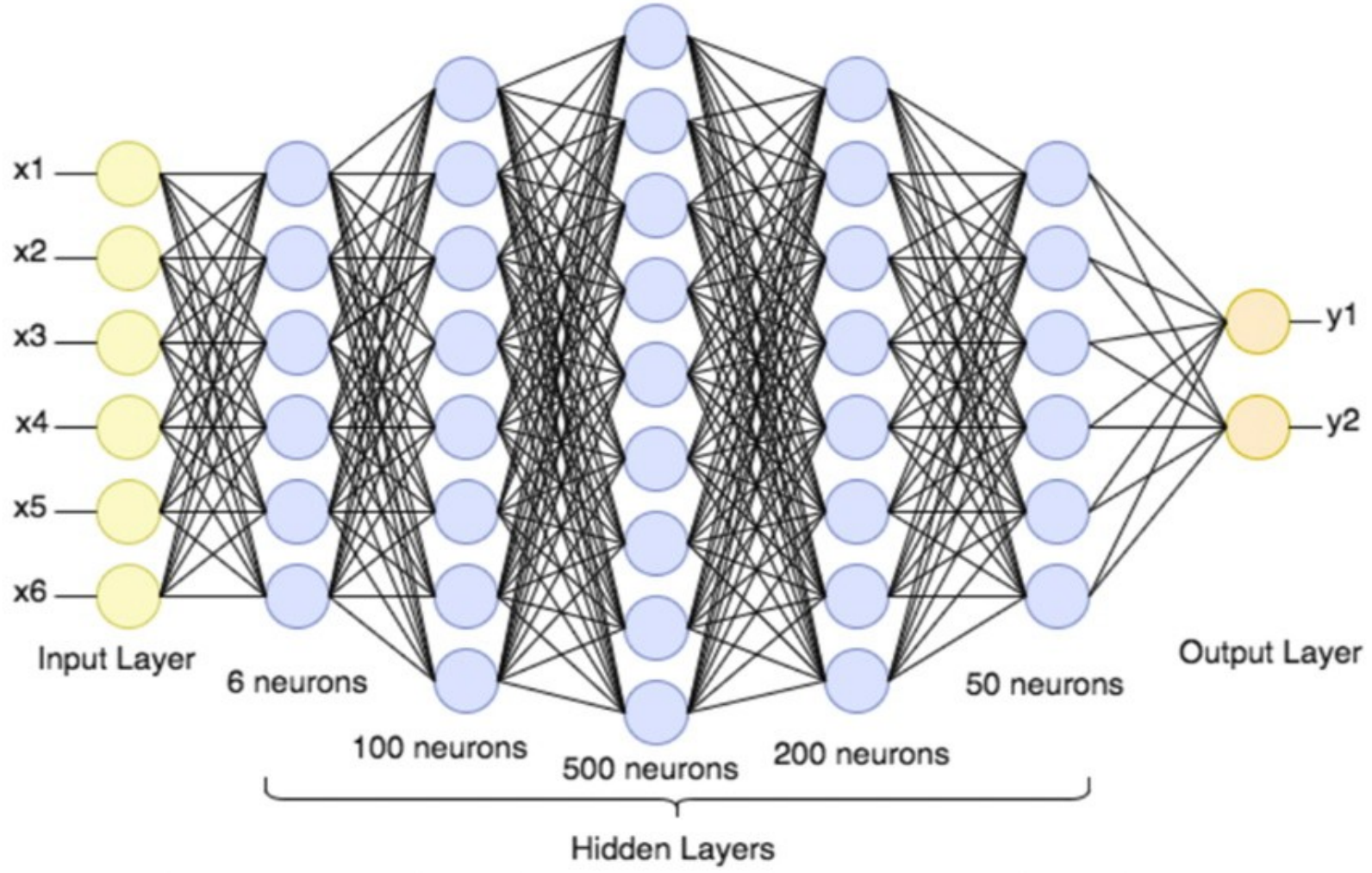
Réseaux de Neurones : Principe



Réseaux de Neurones : Apprentissage et Reconnaissance

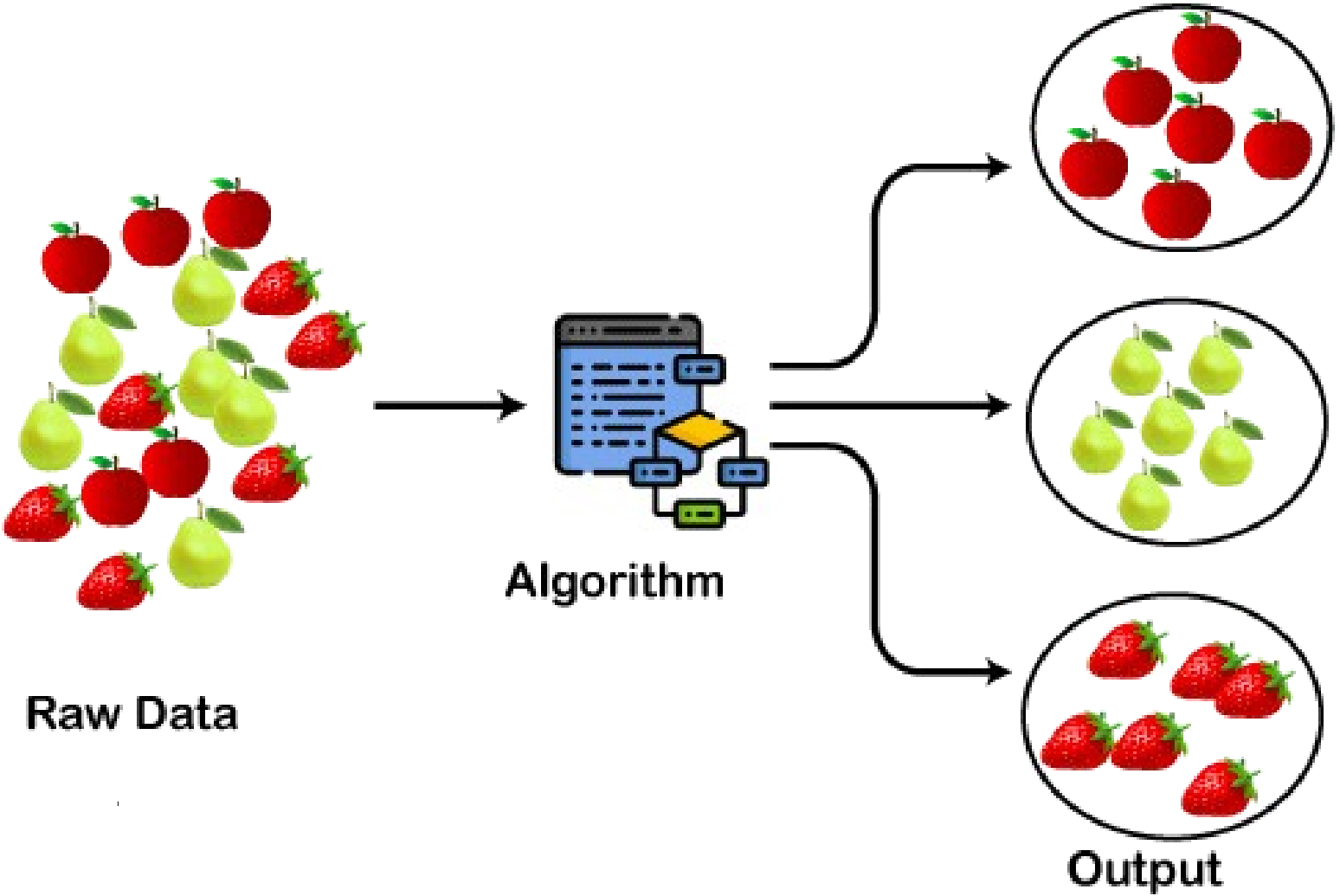


Réseaux de Neurones : Deep Learning

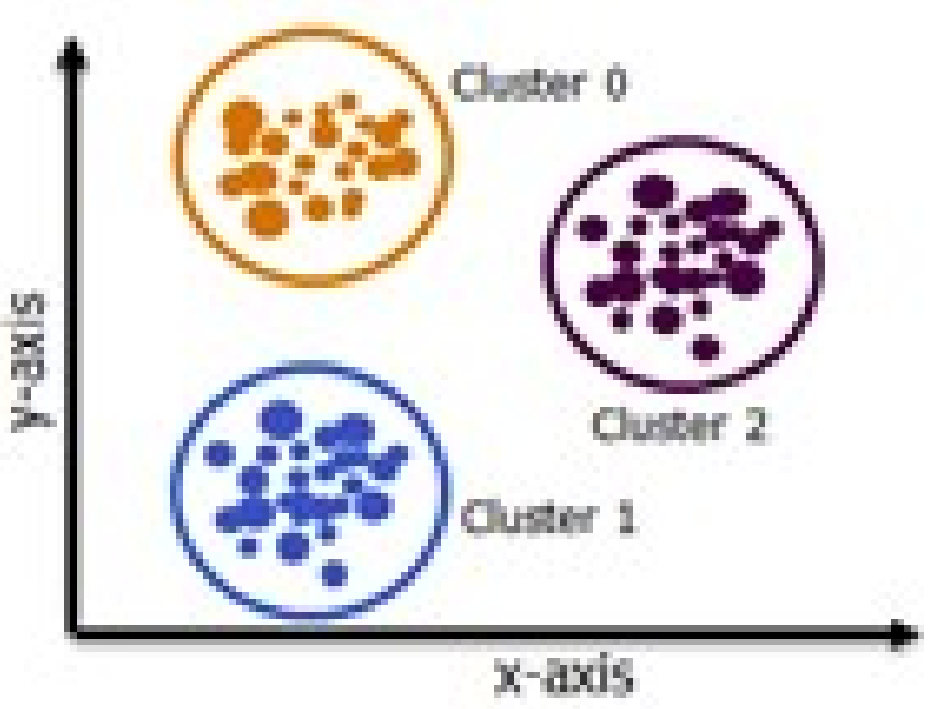


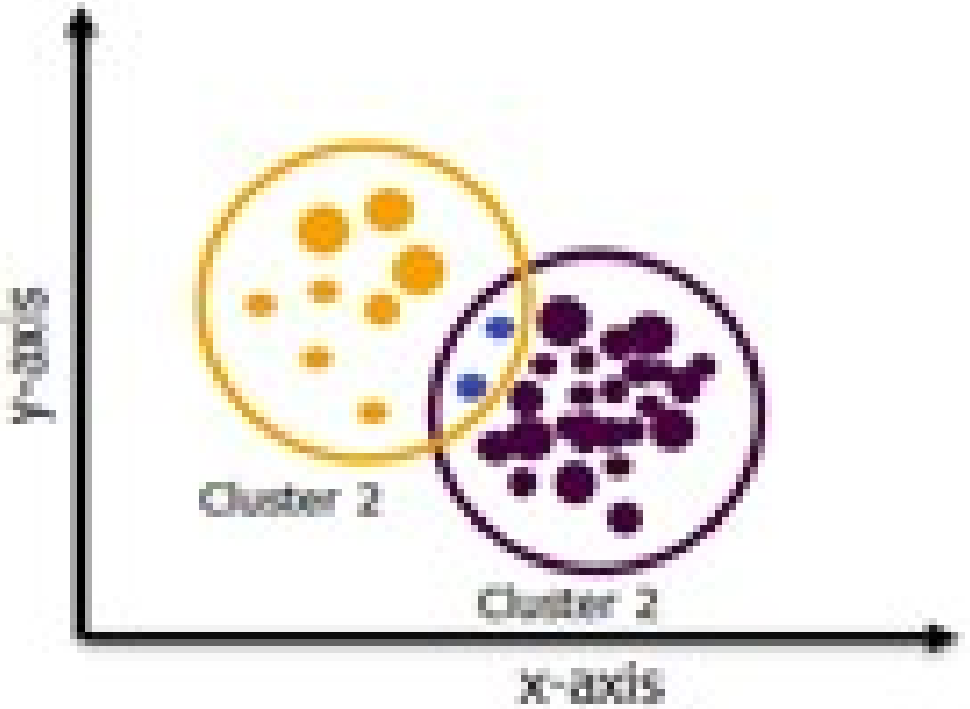
Clustering/Segmentation des Données

Clustering : Principe

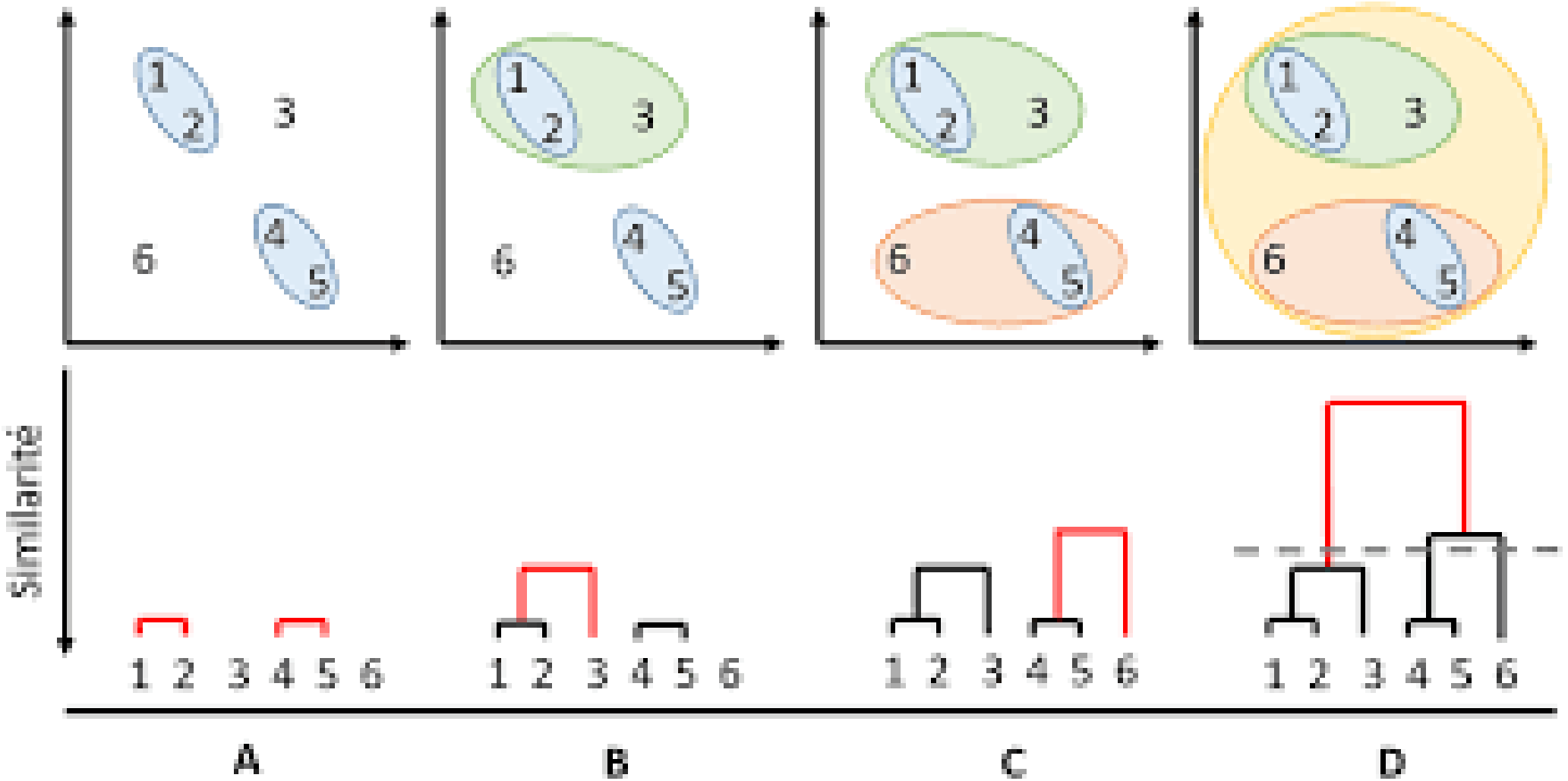


Clustering Exclusif et Clustering en Chevauchement

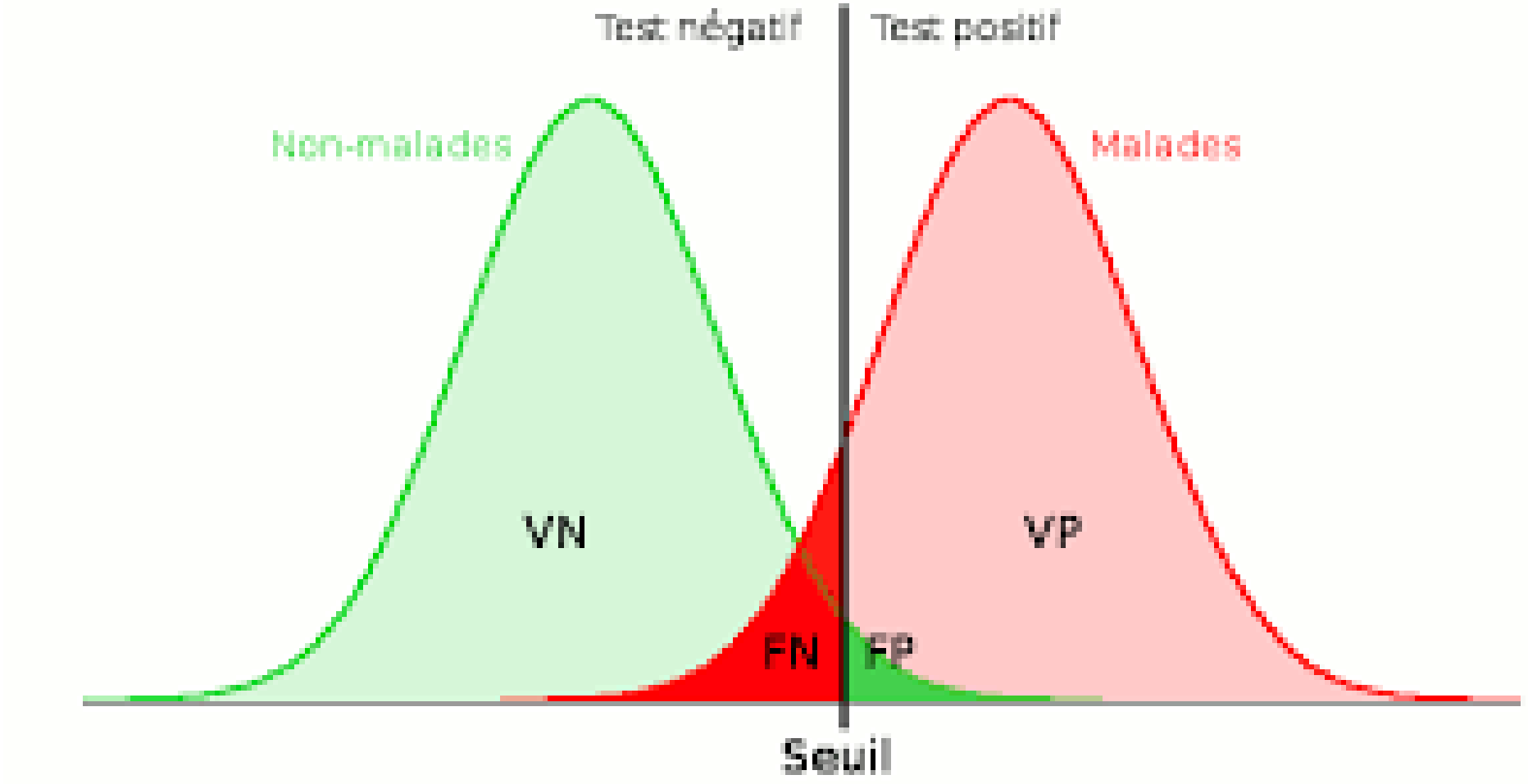




Clustering Hiérarchique



Mesure de La Qualité de l'Apprentissage


Test négatif
Test positif
Non-malades
Malades
VN
VP
FN
FP
Seuil

Vrai/Faux Positif/Negatif

	malade	non malade
test positif	vrai positif (VP)	faux positif (FP)
test négatif	faux négatif (FN)	vrai négatif (VN)

Sensibilité = VP/(VP+FN)

Spécificité = VN/(FP+VN)

Découverte des motifs et des règles d'association

Exemples de règles d'association : Si … Alors …

- Un client achetant des oignons et des pommes de terre simultanément, serait susceptible d'acheter la viande.
- Les clients qui achètent des couches achètent aussi des bières.
- Les clients qui achètent un laptop reviennent environ deux semaines plus tard acheter une imprimante.
- The more the ambient temperature decreases, the more the low temperature decreases (SG = 14%).
- **The more the alcohol consumption increases, the more BMI (Body Mass Index) decreases" (SG = 14.29%).**
- **The more health-percentage expenditure increases, the more GDP (Gross Domestic Product) increases**

Exemples de règles d'association : Suite et fin

- **The more life expectancy increases, the more income composition of resources increases" (SG = 20.41%)**
- **The more adult mortality increases, the more income composition of resources decreases" (SG = 15.51%)**
- **The more weekly death counts by age group in range 0 to 14 increase, the more weekly death counts by age group 85+ increase" (SG = 21.74%)**
- The more government consumption increases, the more regulatory trade barriers increase" (SG = 17.06%).

Quelques exploits concrets du machine learning

IA : BioMind bat des radiologues lors d'une compétition de diagnostics en juillet 2018

225 cas

15 médecins radiologues experts

66% de diagnostics corrects
Prédiction correcte de complication : **63%**

Une intelligence artificielle

BioMind

Entrainée sur les archives de l'hôpital de Beijing Tiantan
87% de diagnostics corrects
Prédiction correcte de complication : **83%**

IA : Predpol lutte contre la criminalité

Prédit le lieu, l'heure et la nature du crime à partir de données historiques.

Utilisée par de nombreuses villes d'Amérique du Nord (Atlanta, Los Angeles …)

Los Angeles Nov. 2011 - mai 2012:

- 33 % d'agressions
- 21 % de crimes violents

IA : Amazon augmente son CA grâce à la recommandation

Emails personnalisés et recommandations sur site:

- Contenu “tendance”
- Articles achetés ensemble
- Recommandations grâce à l’historique d’achat
- Recommandations grâce à l’historique des produits vus
- Nouvelles versions d’un produit déjà possédé

35% du chiffre d’affaire

Les différentes technologies peuvent co-habiter

La Recherche : Défis Technologiques

Quelques questions de recherche

- **Modèles d'apprentissage :** Recherche de nouveaux modèles pertinents...
- **Règles d'association :** Recherche de nouveaux modèles pertinents…
- **Modèles de motifs :** Recherche de nouveaux modèles pertinents
- **Coûts d'exploitation des algorithmes :** Recherches sur l'amélioration des performances
- **Architectures des réseaux de neurones :**
- **Systèmes expert :**
- **Explicabilité :** Recherches pour expliquer « pourquoi ca marche ? »


Merci !


Questions/Réponses