
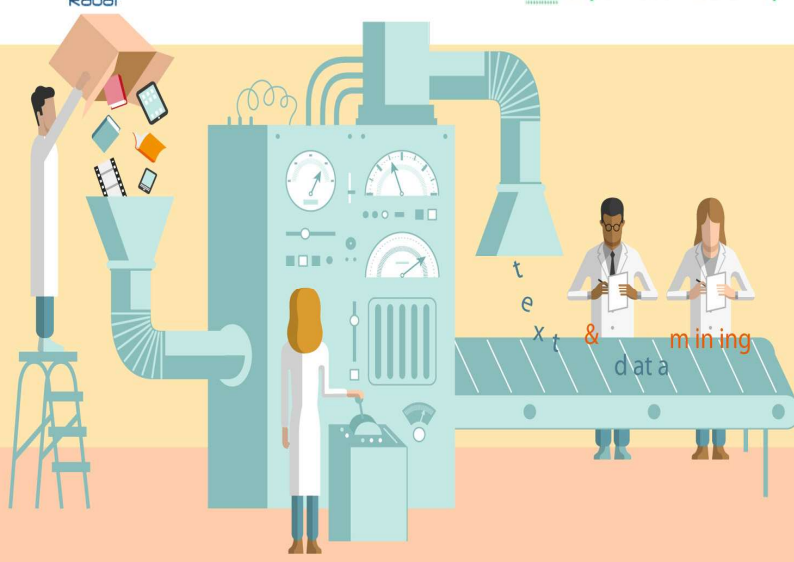


Université Mohammed V
Faculté des Sciences
Rabat

Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences
Département d'Informatique



IPSS
**Intelligent Processing
Systems & Security**



Data Mining & Machine Learning

Master IPS
Faculté des sciences – Rabat
Université Mohamed V

1

Data Mining : Introduction et généralités

De la donnée à la connaissance





ipes.boutyour@gmail.com

2

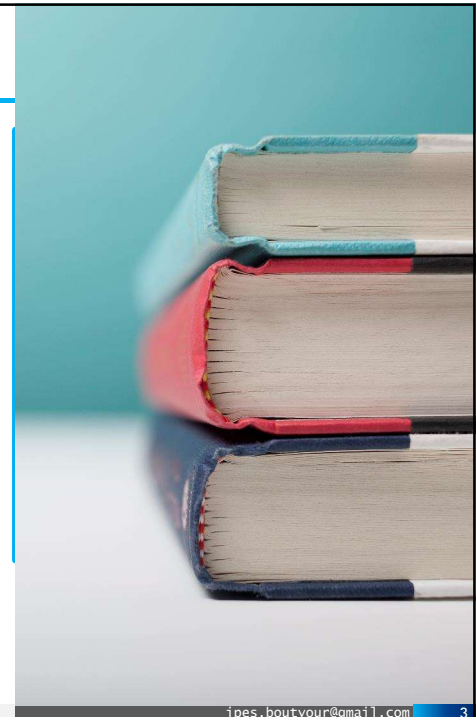
Avant-propos

Syllabus (Data Mining)

- Introduction au data Mining
- Le projet Data Mining
- Compréhension des données
- Préparation des données
- Elaboration des modèles

Evaluation

- 30% : Projet DM & ML (Python)
- 40% : Examen final
- 30% : Exposés/TPs notés (langage R et Tanagra) et Assiduité



ipes.boutyour@gmail.com

3

3

Pourquoi le Data Mining

- **Augmentation exponentielle de la taille des données**

1. La taille des données stockées double tous les deux ans
2. 40.000.000.000 TB en 2020, 50 fois plus qu'en 2000
3. Multi-source: réseaux sociaux, entreprises, IoT, capteurs



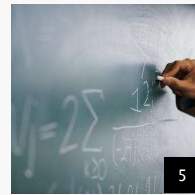
ipes.boutyour@gmail.com

4

4

Pourquoi le Data Mining

2020 This Is What Happens In An Internet Minute



ipes.boutyour@gmail.com

5

5

Pourquoi le Data Mining

- **Augmentation exponentielle de la taille des données**
 - La taille des données stockées double tous les deux ans
 - 40.000.000.000 TB en 2020, 50 fois plus qu'en 2000
 - Multi-source: réseaux sociaux, entreprises, IoT, capteurs, ...
- **Large BDD inexploitable par les méthodes d'analyse classiques** ➔ **Beaucoup de données mais peu de connaissance !**
- **Pression climat économique**
 - Rude concurrence
 - Nécessité d'anticiper et de prédire
 - Individualisation des consommateurs



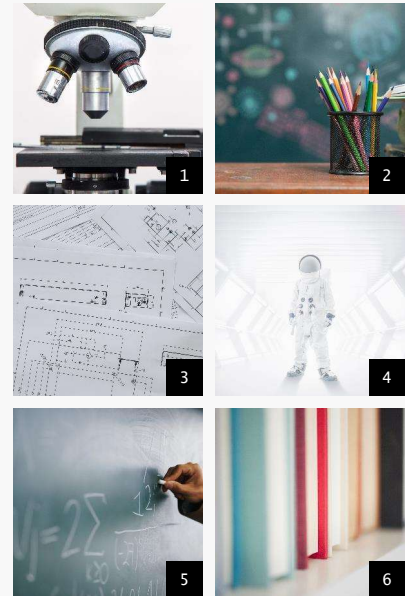
ipes.boutyour@gmail.com

6

6

Pourquoi le Data Mining

- **Données = ressources de valeur**
- **Besoin de techniques pour analyser les données et extraire les informations/connaissances automatiquement**
 - Donnée = fait
 - Information = modèle/motif dans les données
- **Dilemme : extraire les informations intéressantes à partir des données**
- **Solution : Data Mining**



ipes.boutyour@gmail.com

7

7

Qu'est-ce que le Data Mining

- **Data = données**
- **Mining = exploitation minière**



Exploitation des données



ipes.boutyour@gmail.com

8

8

Qu'est-ce que le Data Mining

Définitions

- Selon Piatetski Shapiro :
« L'extraction d'information **originale, auparavant inconnues** et **potentiellement utiles**, à partir de données »
- Selon John Page:
« La découverte de **nouvelles corrélations, tendances** et **modèles** par tamisage d'un large volume de données »

Qu'est-ce que le Data Mining

Définitions

- Selon Kamran Parsaye:
« Un processus **d'aide à la décision** où les utilisateurs cherchent des **modèles d'interprétation** dans les données »
- Selon Michael Berry :
« L'**exploration** et l'**analyse**, par des **moyens automatiques** ou **semi-automatiques**, d'un **large volume de données** afin de **découvrir** des **tendances** ou des **règles** »

Qu'est-ce que le Data Mining

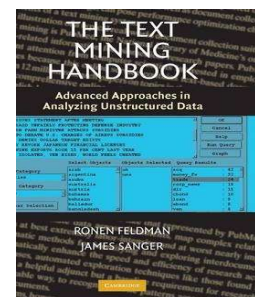
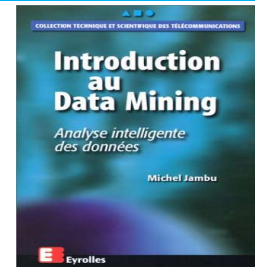
Définitions

- Selon Michel Jambu:

« Un processus non élémentaire de mise à jour de relation, corrélation, dépendances, association, modèles, structure, tendance, classes, facteurs obtenus en naviguant à travers de grands ensembles de données »

- Selon Ronen Feldman:

« Extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de larges bases de données »



ipes.boutyour@gmail.com

11

11

Qu'est-ce que le Data Mining

Synthèse

- « Extraction d'informations intéressantes (non triviales, implicites, préalablement inconnues et potentiellement utiles) à partir de larges entrepôts de données, en utilisant des procédures automatiques ou semi-automatiques pour une prise de décision. »

Autres dénominations:

- Fouille de données
- ECD (Extraction de Connaissances à partir des Données)
- KDD (Knowledge Discovery from Databases)
- Analyse de données

ipes.boutyour@gmail.com

12

12

Qu'est-ce que le Data Mining

Synthèse

- L'objectif principale de Data Mining c'est de créer un processus automatique qui a comme point de départ les données y comme finalité l'aide à la prise des décisions.

Problématique du Data Mining

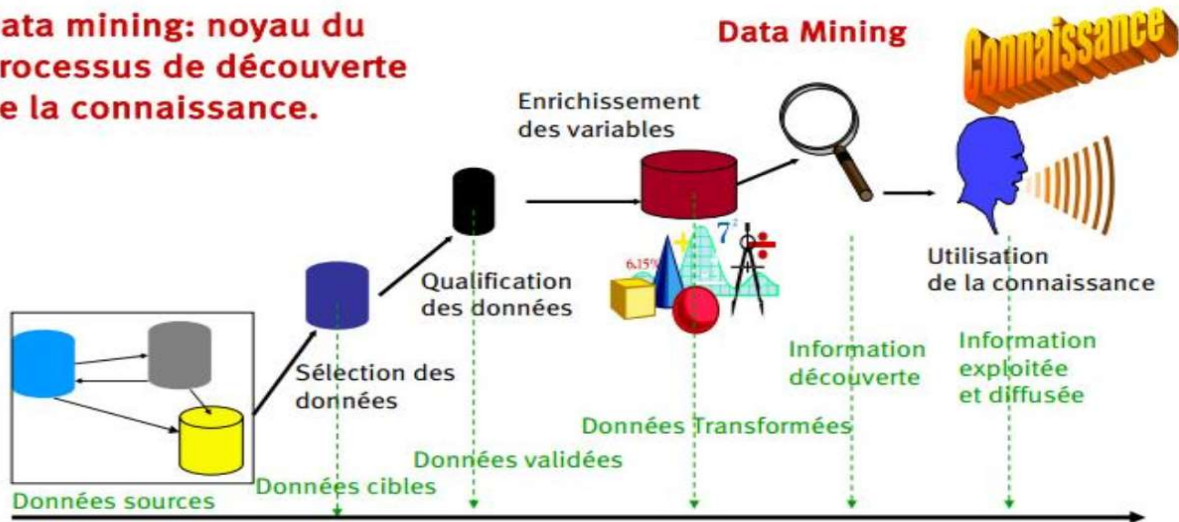
Problématique



Comment gérer la grande quantité des données "brutes" provenant de plusieurs sources pour les rendre accessibles et lisibles par le décideur?

Data Mining : De la donnée à la connaissance

Data mining: noyau du processus de découverte de la connaissance.



ipes.boutyour@gmail.com

15

15

Data Mining : De la donnée à la connaissance

Cycle de vie d'un projet de Data mining

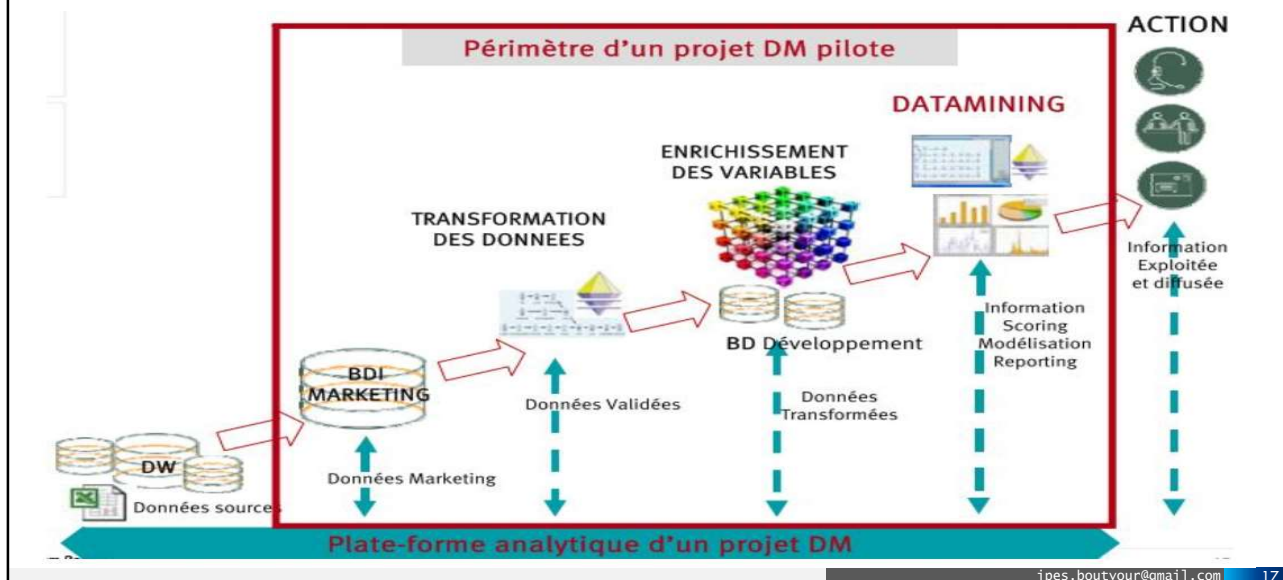
- Compréhension du domaine d'application
- Sélection d'un échantillon de données
- Nettoyage et transformation des données (étape très importante)
- Application des techniques de data mining
- Visualisation des modèles découverts
- Evaluation et interprétation des modèles découverts
- Utilisation de la connaissance extraite

ipes.boutyour@gmail.com

16

16

Data Mining : De la donnée à la connaissance



17

Data Mining : Sur quelles données?

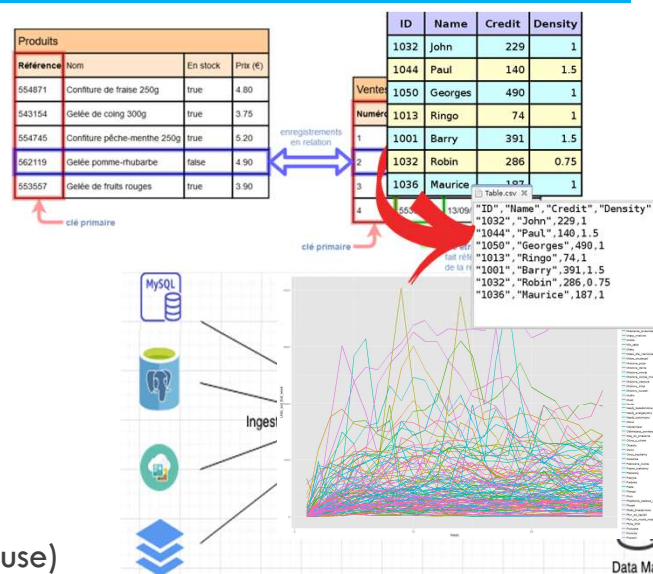
○ Bases de données classiques

- Fichiers plats
- BD Relationnelles
- BD Transactionnelles

○ Bases de données avancées

- Objet et Objet-Relationnelles
- Spatiales
- Séries temporelles ou chronologiques
- Textes et Multimédia
- Hétérogènes
- WWW

○ Entrepôts de données (Data Warehouse)

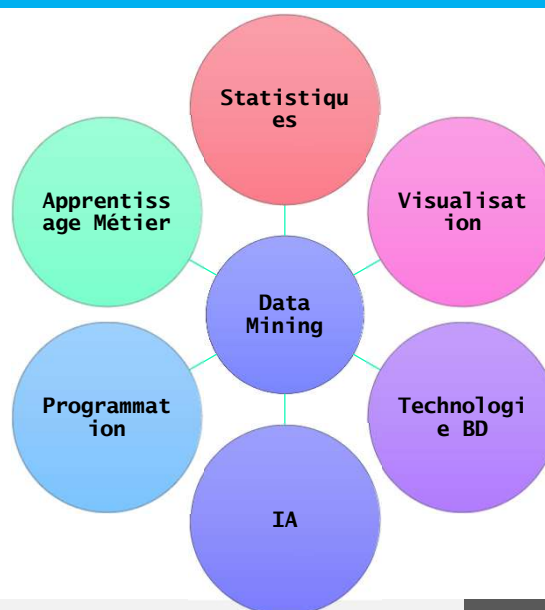


18

Data Mining vs Statistiques

- En statistiques :
 - Quelques centaines d'observations
 - Quelques variables
 - Forte hypothèses sur les lois statistiques
 - Importance accordée au calcul mathématique
 - Echantillon aléatoire
- En Data Mining :
 - Des millions d'observations
 - Des centaines de variables
 - Données recueillies sans étude préalable
 - Nécessité de calcul rapide

Data Mining : Intersection de multiples disciplines



Applications du Data Mining

- Analyse de données et aide à la décision
 - Analyse de marché
 - Marketing ciblé, gestion de relations clients, analyse des achats des clients, ventes croisées, segmentation du marché
 - Analyse de risque
- Autres applications
 - Text Mining : news groups, emails, documents web,...
 - Optimisation des requêtes

Applications du Data Mining

- Analyse de BD de transactions d'un supermarché permet d'étudier le comportement des clients :
 - Réorganiser les rayons
 - Ajuster les promotions
- Regrouper les clients selon certains critères :
 - Cibler les « mailings »
 - Gestion des relation client
- Analyse de données médicales
- Analyse de données financières :
 - Prédire l'évolution des actions
 - Organismes de crédit (dresser des profils de clients)

Applications du Data Mining

○ Détection de fraudes :

- Construire des modèles de comportements frauduleux sur la base des données historiques
- Retrouver les instances similaires en utilisant les techniques de data mining.
- ✓ Assurances auto: détecter les personnes qui collectionnent les accidents et les remboursements

○ Blanchiment d'argent: détecter les transactions suspectes

Applications du Data Mining

○ Astronomie :

- Le laboratoire JPL a découvert 22 quasars (la région compacte entourant un trou noir super massif au centre d'une galaxie massive) en utilisant les techniques de datamining

○ Web :

- IBM a appliqué des algorithmes de data mining pour réorganiser leurs sites WEB afin de faciliter la navigation.
- Améliorer le WEB marketing.
- Personnalisation des pubs affichées
- Optimisation des sites Web
- Profilage et Recommandation

Data Mining : Principales techniques

- **Techniques descriptives** : consiste à trouver les caractéristiques générales relatives aux données fouillées.
 - Classification (Segmentation , Clustering)
 - Typologie
 - Règles d'association
- **Techniques prédictives** : Consiste à utiliser certaines variables pour prédire les valeurs futures inconnues de la même variable ou d'autres variables.
 - Classement
 - Arbre de décision
 - Régression
 - Classification
 - Réseau de neurone

ipes.boutyour@gmail.com

25

25

Data Mining : Principales techniques

1°/ Techniques descriptives

- But :
 - mettre en évidence des informations présentes mais cachées par le volume des données.
 - Réduire, résumer et synthétiser les données

Il n'y a pas de variable à expliquer

ipes.boutyour@gmail.com

26

26

Data Mining : Principales techniques

1°/ Techniques descriptives

- Association (analyse d'affinité): connue comme (Link Analysis) se réfère à découvrir les relations non évidentes entre les données
 - Rapprocher les caractéristiques, comportements ou préférences d'un individu
 - Différentes techniques :
 - Règles d'associations
 - Analyse de corrélation et de causalité
 - Analyse des correspondances (ACM)

Data Mining : Principales techniques

1°/ Techniques descriptives

- Association:
 - Exemple: Cadis d'un super marché

Cadis	Contenu
1	Pain, Lait, Œufs
2	Lait, Pain
3	Pain, Fromage
4	Lait, Fromage
.....	Soda, Œufs, Chips
n	Œufs, Lait, Chocolat, Pain

Data Mining : Principales techniques

1°/ Techniques descriptives

○ Classification :

- Appelée aussi analyse de groupes (clusters) ou Clustering
- Regrouper les données en classes de telle sorte à maximiser la similarité intra-groupe et la minimiser entre groupes distincts
- Différentes techniques :
 - K-means
 - Nuées dynamiques
 - Classification Ascendante Hiérarchique
 - Cartes de Kohonen
 -

ipes.boutyour@gmail.com

29

29

Data Mining : Principales techniques

2°/ Techniques prédictives

○ But :

- extrapoler de nouvelles informations à partir d'informations présentes
- Expliquer les données

Il y a une variable cible à expliquer

ipes.boutyour@gmail.com

30

30

Data Mining : Principales techniques

2°/ Techniques prédictives

○ Classement :

- Prévoir l'appartenance d'un individu à un groupe donné
- Expliquer une caractéristique qualitative à partir d'autres variables (qualitatives ou quantitatives)
- Différentes techniques :
 - Arbres de décisions
 - Régression logistique
 - Réseau de neurones
 -

Data Mining : Principales techniques

2°/ Techniques prédictives

○ **Exemple 1** : Opérateur Télécom

- Les clients reçoivent un téléphone portable gratuit d'une valeur de 3.000 DHs avec un ré-engagement d'un an.
- Donner un téléphone portable gratuit à tous ➡ Coûteux
- Faire revenir un client passé à la concurrence ➡ Difficile et coûteux
- Solution possible: Prédire les clients qui risquent de partir trois mois avant l'expiration de leur contrat et leur offrir un téléphone.

Data Mining : Principales techniques

2°/ Techniques prédictives

○ **Exemple 2** : Opérateur d'assurance

- Comment définir le paiement annuel adapté à un jeune homme de 18 ans qui a acheté une voiture neuve de 500.000dhs?
- Analyser les données de tous les clients
- Extraire les éléments de risque (probabilité d'avoir un accident est basée sur ... probabilité de vol/incendie est basée sur...)
- Solution possible: probabilité d'avoir un accident > moyenne



Augmenter l'annuité

Data Mining : Principales techniques

2°/ Techniques prédictives

○ Variable cible qualitative

- Analyse discriminante / Régression logistique (Scoring)
- Arbres de décisions
- Réseaux de neurones

○ Variable cible quantitative

- Régression linéaire (simple et multiple)
- Arbres de décisions
- Réseaux de neurones

Data Mining : Quelle technique utiliser?

- Différents algorithmes conviennent à différentes tâches
- Différentes forces et faiblesses
- En général, on doit essayer plusieurs algorithmes
- Pour avoir une bonne solution, souvent il est utile de combiner plusieurs algorithmes

Data Mining : Tout est utile?

- Mesure d'intérêt :
 - Un pattern est intéressant s'il est facilement compréhensible, a un degré de certitude, nouveau, peut servir à valider (ou invalider) une hypothèse utilisateur
- Mesure Objective vs. Subjective :
 - Objective : basée sur des mesures statistiques : support, intervalle de confiance, etc.
 - Subjective : basée sur le point de vue de l'utilisateur sur les données (ex: le fait que cela soit inattendu, nouveauté, actionnabilité, etc.)

Data Mining : Tout est utile?

- Trouver tous les patterns intéressants: **Complétude**
 - Association vs. classification vs. Regroupement
 - Trouver que les patterns intéressants: **Optimisation**
 - D'abord les trouver tous puis filtrer
- ou**
- Ne générer que les motifs intéressants

Data Mining : Quels logiciels?

○ **Commerciaux**

- SPAD
- SAS Enterprise Miner
- SPSS Clementine
- STATISTICA DATA Miner
- CORICO
- IBM Intelligent Miner
- RapidMiner
- KNIME
-

○ **Universitaire**

- Langage R - R Studio - Rattle
- Python
- SIPINA
- WEKA
- Orange
-

MÉTHODE CRISP-DM

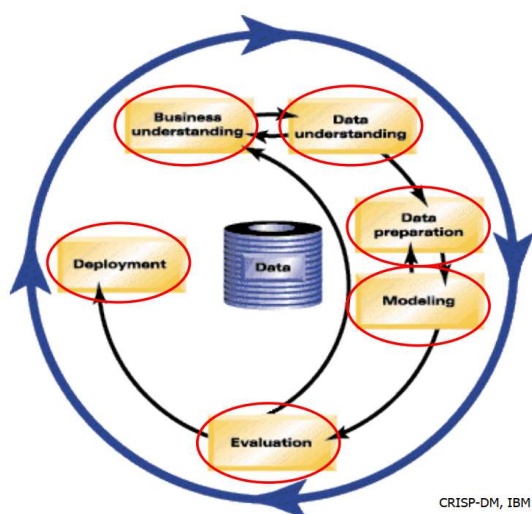
- Cross Industry Standard Process for Data Mining
- Clé de la réussite
- Développée par IBM dans les années 60 pour réaliser les projets data mining
- Seule méthode utilisable efficacement pour tous les projets Data Science.
- Agile et itérative
 - Chaque itération apporte de la connaissance métier supplémentaire qui permet de mieux aborder l'itération suivante.

ipes.boutyour@gmail.com

39

39

MÉTHODE CRISP-DM



1. Compréhension du métier
2. Compréhension des données
3. Constitution du Data Hub
4. Modélisation
5. Evaluation
6. Déploiement

ipes.boutyour@gmail.com

40

40

Facteurs de succès d'un projet DM

- Objectifs définis (précis, stratégiques et réalistes)
- Qualité et la richesse des informations collectées
- Stockage des informations relationnelles sur les clients
- Collaboration des compétences métiers et statistiques
- Maîtrise des techniques de DM utilisées
- Bonne restitution des résultats et l'implication de tous les partenaires chargés de leur mise en œuvre
- Analyse du retour de chaque action pour la suivante

Facteurs de succès d'un projet DM



- En amont, l'entreprise doit :
 - Veiller aux compétences en DM
 - Veiller à la qualité des données recueillies
 - Veiller à une mise en œuvre et un suivi rigoureux des actions
- En aval, l'entreprise s'appuiera sur le DM pour :
 - Adapter éventuellement ses processus marketing
 - Passer du marketing « produit » au marketing « client »
 - Adapter éventuellement ses processus de décision
 - Adapter ses délégations de pouvoir

Facteurs de succès d'un projet DM

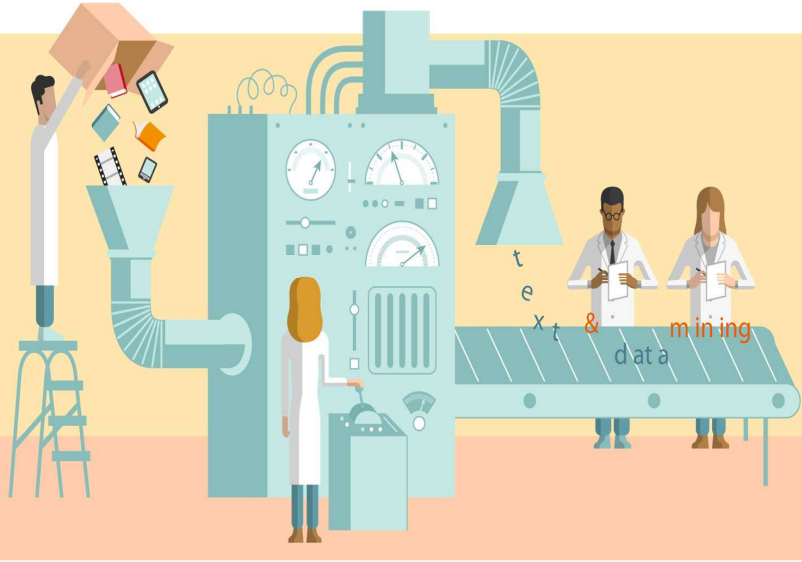
- Vendre le DM :
 - Convaincre les commerciaux/décideurs que le DM ne fournit qu'une aide à la décision, et non la décision elle-même
 - Convaincre les commerciaux de bien alimenter les bases de données marketing
 - Sensibiliser les commerciaux/décideurs au gain de productivité offert par le DM

Freins au succès d'un projet DM

- **Au niveau de l'entreprise :**
 - Méconnaissance/crainte/scepticisme
 - Manque de soutien du top management
 - Difficulté à vulgariser certains résultats
- **Au niveau des données :**
 - Indisponibilité
 - Fréquence de la mise à jour
 - Qualité des données collectées
- **Au niveau des outils:**
 - Complexité des logiciels
 - Coût des applications d DM
- **Au niveau des compétences:**
 - Profil complexe (Informatique, marketing, statistique, etc.)
- **Idées fausses:**
 - On n'a plus besoin de spécialistes du métier
 - Le DM permet de faire des découvertes incroyables
 - Le DM permet de faire des découvertes incroyables



Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences
Département d'Informatique



Data Mining & Machine Learning

Master IPS
Faculté des sciences – Rabat
Université Mohamed V