

Data Mining

Apprentissage et Fouille de données

Dr. Sana Hamdi

sana.hamdi@fst.utm.tn

Plan

- I. Concepts et notions de base
- II. Apprentissage supervisé
- III. Apprentissage non supervisé

Concepts et Notions de Base

- Introduction
- Le processus ECD
- Définition du Data Mining
- Applications du Data Mining
- Les méthodes Data Mining
 - ✓ L'apprentissage non supervisé
 - ✓ L'apprentissage supervisé
- Les méthodologies de travail

Introduction



Exemple introductif (1/3)

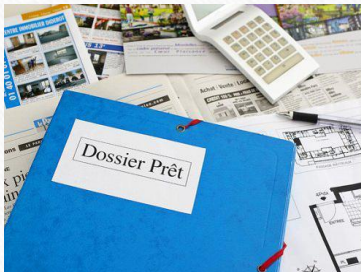
- Demande de crédit bancaire:

- ✓ divorcé
- ✓ 5 enfants
- ✓ chômeur enfin de droits
- ✓ compte à découvert



Exemple introductif (2/3)

- Demande de crédit bancaire:
 - ✓ Analyse de dossier de prêt
 - ✓ Variables explicatives (âge, salaire, statut, nombre d'enfants,...)
 - ✓ Une fonction pour prévoir une prédiction



Exemple introductif (3/3)

- **Expert humain :**
 - Ayant accumulé une expérience considérable
 - Connu par ses décisions justes
- **Système expert :**
 - Cogniticien extrait les connaissances
 - Capable de reproduire les mécanismes cognitifs d'un expert
 - Ne dépasse pas le niveau des experts
- **Système d'apprentissage :**
 - Extrait tout seul l'expertise à partir de données
 - Découverte de nouvelles connaissances

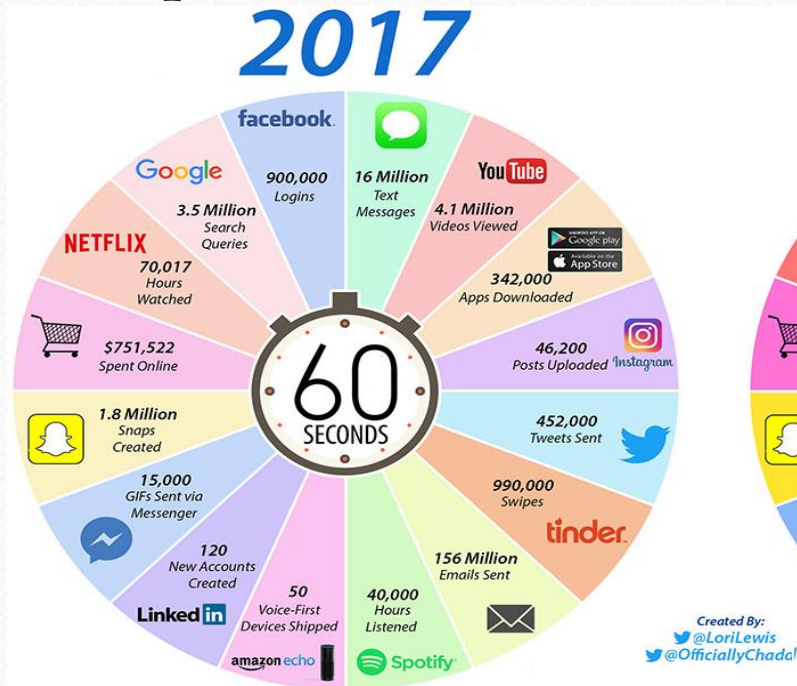


Objectif

- Extraire des nouvelles connaissances
 - Remplacer ou aider l'expert humain ou le décideur dans un domaine spécifique dans le cadre de prise de décision.

Motivations (1/4)

- Explosion des données: Une minute sur le Web, c'est:



2021 This Is What Happens In An Internet Minute



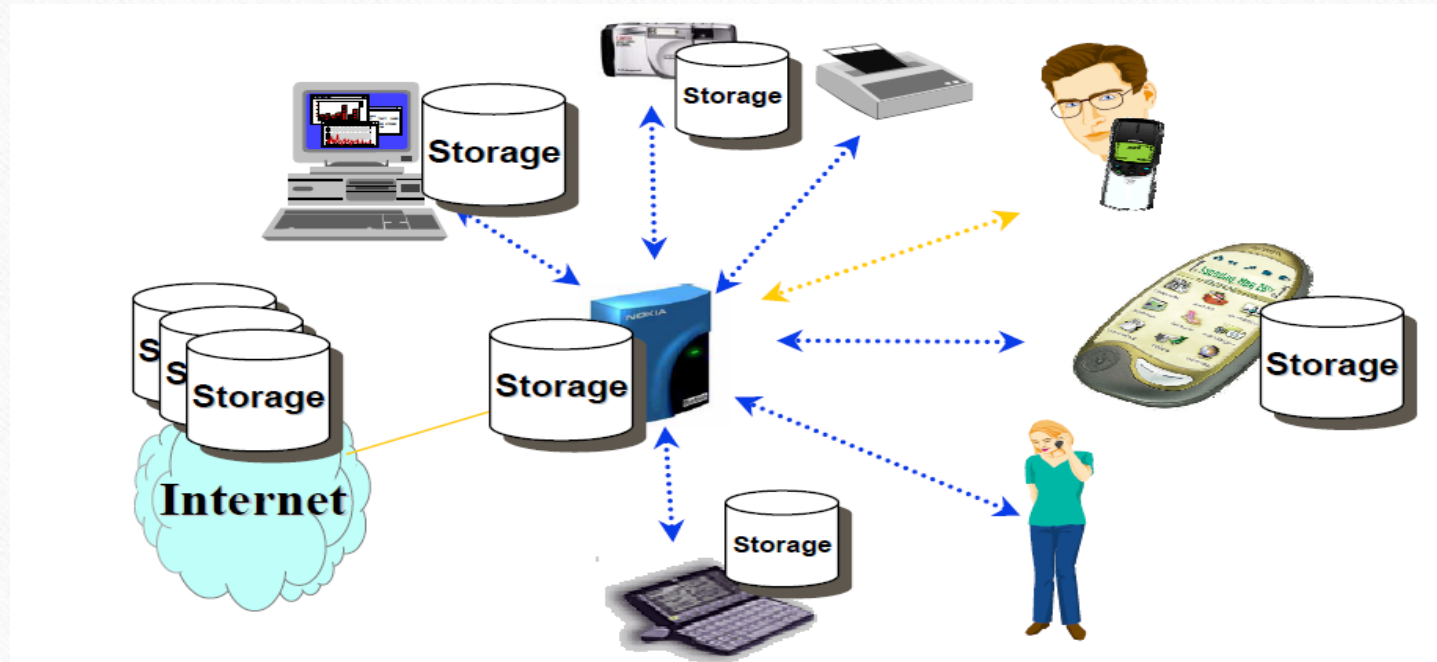
Motivations (2/4)

- **Explosion des données**

- Masse importante de données (millions de milliards d'instances) : elle double tous les 20 mois.
 - ✓ BD très larges: VeryLargeDatabases (VLDB)
- Données multi-dimensionnelles (milliers d'attributs)
 - ✓ BD denses
- **Besoin de traitement en temps réel de ces données**

Motivations (3/4)

- Masse importante de données sur des supports hétérogènes → Manque de connaissances!



Motivations (4/4)

Ce dont on a besoin:

- Automatisation
- Extraction des connaissances des bases de données
- Génération d'hypothèses



Le processus ECD



Solution: Le processus ECD

(Extraction de connaissances à partir de données)

Objectifs :

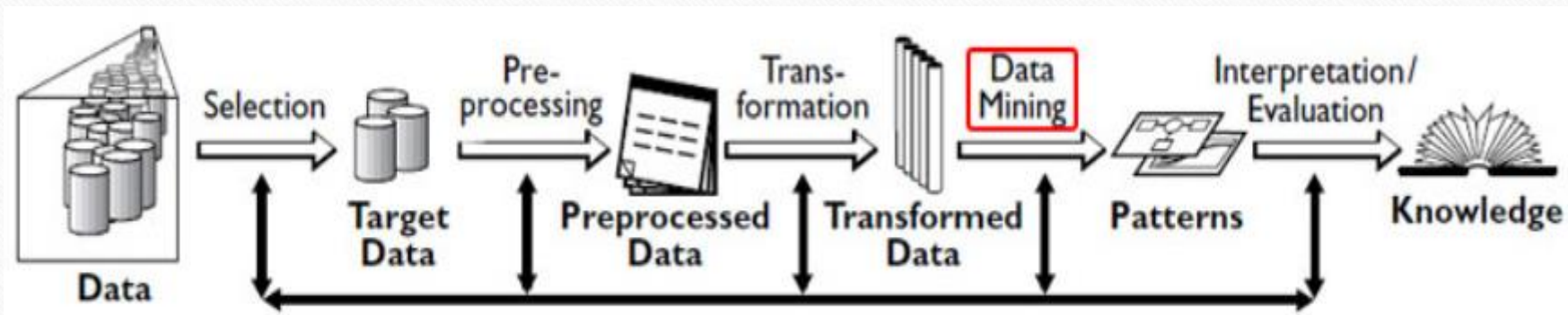
- Limiter l'intervention de l'utilisateur
- Développer des techniques et systèmes efficaces et extensibles pour :
 - l'exploration des données distribuées et des BD larges et multi-dimensionnelles
 - L'extraction des informations cachées
- Découvrir des modèles (« patterns ») difficiles à percevoir

ECD vs Fouille de Données

- **L'ECD (Extraction de Connaissances à partir de Données) est un processus itératif de découverte, dans les BD larges, de modèles de données valides, utiles et compréhensibles.**
 - ✓ Itératif : nécessite plusieurs passes
 - ✓ Valides : valables dans le futur
 - ✓ Utiles : permettent à l'utilisateur de prendre des décisions
 - ✓ Compréhensibles : présentation simple
- L'un de ses traitements est la **Fouille de données** (Data Mining)

Le processus de découverte de connaissances

- Extraction de la connaissance intéressante à partir de données (règles, régularités, patterns, contraintes) à partir de grandes bases de données (**ECD/KDD**)

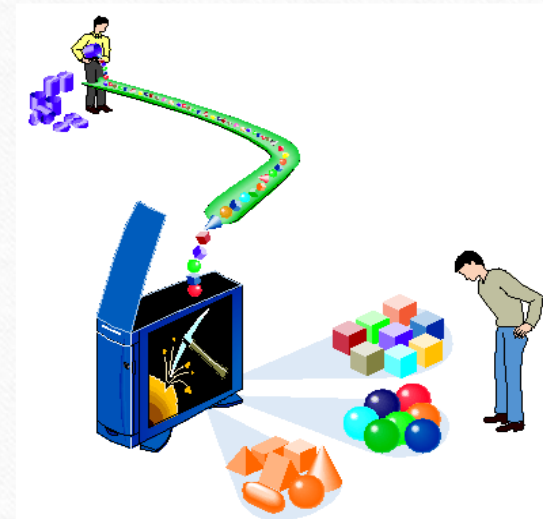


Définition du Data Mining



Définition du Data Mining

- Le data Mining ou la fouille de données présente la découverte d'une connaissance (information intéressante) à partir de grandes quantités de données, par des méthodes automatiques.
 - ✓ Quels types de données?
 - ✓ Qu'est ce qu'une connaissance?
 - ✓ Qu'entend-on par découverte?
- Fortement lié à l'apprentissage automatique!



Quels types de données?

- Tableau de données:
 - N lignes: les individus, les objets d'études
 - M colonnes: les variables, les caractéristiques des objets, les attributs
- Base de données relationnelles:
 - Des tables
 - Des liens entre les tables (un client (dans la table des clients) a acheté des produits (dans la table des produits))
- Entrepôt de données:
 - Mise en commun (fusion) des bases de données
 - Agrégation des valeurs: nombre de commandes par client (par mois (d'un produit...))

Données complexes, hétérogènes, volumineuses et évolutives! Comment interpréter et exploiter toutes ces données?

Les données: type (1/6)

- **Vectérielles**
- **Séquences**
- **Structurées**
- **Temporelles**
- **Spatiales**

Les données: type (2/6)

- Vectorielles

Identifieur	Genre	Age	Niveau études	Marié ?	Nb enfants	Revenu	Profession	A prospecter ?
I_21	M	43	Bac+5	Oui	3	55 000	Architecte	OUI
I_34	M	25	Bac+2	Non	0	21 000	Infirmier	NON
I_38	F	34	Bac+8	Oui	2	35 000	Chercheuse	OUI
I_39	F	67	Bac	Oui	5	20 000	Retraitée	NON
I_58	F	56	CAP	Oui	4	27 000	Ouvrière	NON
I_73	M	40	Bac+3	Non	2	31 000	Commercial	OUI
I_81	F	51	Bac+5	Oui	3	75 000	Chef d'entreprise	OUI

Exemple
(*example, instance*)

Descripteur
Attribut
(*feature*)

Étiquette
(*label*)

Les données: type (3/6)

- Vectorielles
- Séquences
- Structurées
- Temporelles
- Spatiales

La protéine « sp|P00004|CYC_HORSE » est activée par ...

```
1  ttcagttgtg aatgaatgga cgtgccaaat agacgtgccg ccgccgctcg attgcactt
61  tgctttcggt ttgccgtcg ttcacgcgt ttagttccgt tcggttcatt cccagttctt
121 aaataccgga cgtaaaaata cactctaacg gtcccgcgaa gaaaaagata aagacatctc
181 gtagaaatat taaaataaat tcctaaagtc gttggtttct cgttcacttt cgctgcctgc
```

Les données: type (4/6)

- Vectorielles
- Séquences
- **Structurées**
- Temporelles
- Spatiales



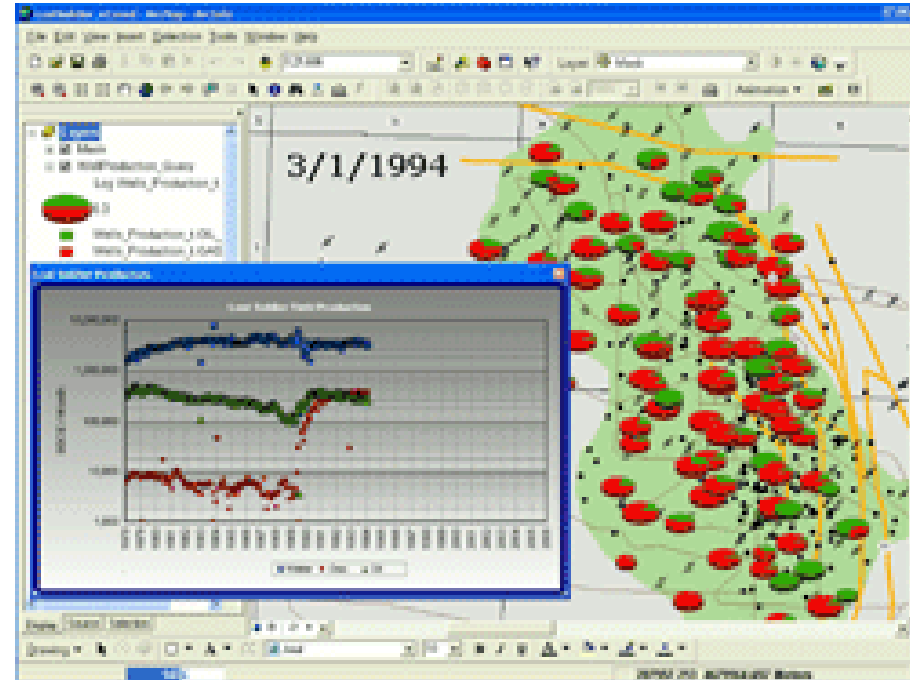
Institut national des sciences appliquées et de technologie
Université à Tunis

L'Institut national des sciences appliquées et de technologie ou INSAT est un établissement universitaire tunisien rattaché à l'Université de Carthage. Formant des techniciens supérieurs et des ingénieurs, il dispense en particulier des formations en informatique dont des masters.
[Wikipédia](#)

Adresse : 676 INSAT Centre Urbain Nord BP• Tunis Cedex 1080
Directeur : [Mohamed Boussaid](#)
Nom officiel : المعهد الوطني للعلوم التطبيقية والتكنولوجيا
Création : 1992

Les données: type (5/6)

- Vectorielles
- Séquences
- Structurées
- **Temporelles**
- Spatiales



Wyoming Oil and Gas Conservation Commission

Les données: type (6/6)

- Vectorielles
- Séquences
- Structurées
- Temporelles
- **Spatiales**



Les données : valeur

- **Types des valeurs des données:**

- ❖ **discrètes:** données binaires (sexe, ...), données énumératives (couleur, ...), énumératives ordonnées (réponses 1:très satisfait, 2:satisfait, ...).
- ❖ **continues:** données entières ou réelles (âge, salaire, ...)
- ❖ Dates
- ❖ Données textuelles
- ❖ Pages/liens web, Multimédia, ...

Les données: exemples

- **Sciences de la vie**
 - médecine : patients et maladies, essais cliniques
 - génomique : gènes, patients, tissus
- **Marketing**
 - fiches clients
 - traces d'usage (site web, communication mobile)
 - Achats
- **Industrie**
 - senseurs : température, vibration
 - Images
 - analyse physico-chimique

Qu'est ce qu'une connaissance?

Données intéressantes

- **liens entre variables**
 - Corrélation
 - dépendance non linéaire
 - capacité de prédiction
- **liens entre individus**
 - interactions significatives
 - groupes homogènes
- **liens entre événements**
 - dépendance logico-temporelle

Données intéressantes : exemples

- **Sciences de la vie**
 - lien entre tabagisme et maladies cardio-vasculaires
 - lien entre tabagisme et cancer du poumon
 - maladies génétiques : mutation , gène détérioré
- **Marketing**
 - évaluation du risque de défaillance pour un crédit
 - typologie des clients
 - recommandation de produits
- **Industrie**
 - identification de modes de fonctionnement normaux d'un matériel
 - qualité d'un produit à partir de mesures objectives

Qu'entend-on par découverte?

- Exploration des données: Savoir produire de la valeur ajoutée à partir de ses propres données:
 - Rapports
 - Outils visuels

Fortement lié à l'apprentissage automatique!

- algorithmes d'apprentissage : inférence à partir d'exemples de résultats voulus
- intervention minimale de l'analyste : choix d'une méthode et analyse des résultats

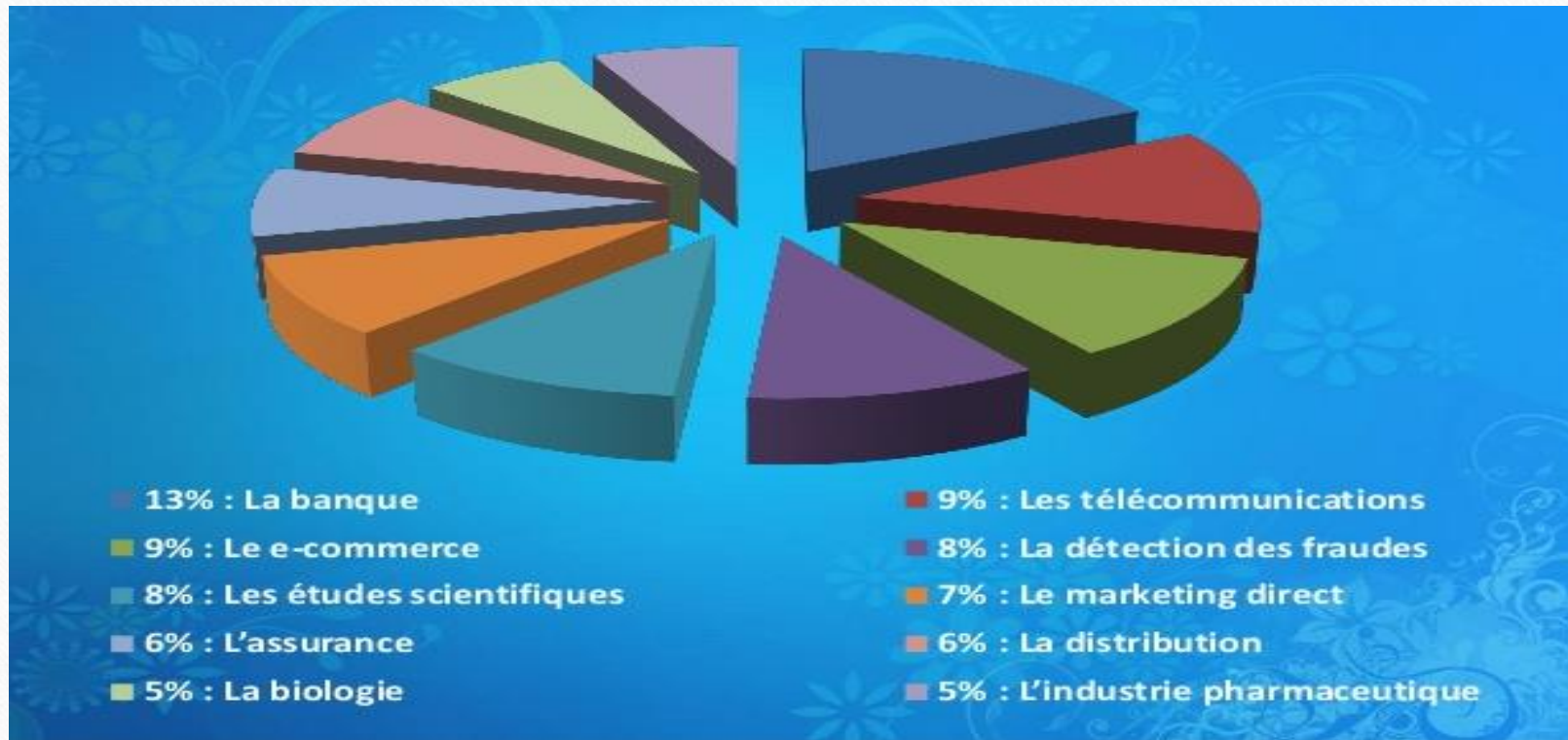
Découverte : exemples

- Exploration
 - statistiques classiques : moyenne, médiane, coefficient de corrélation
 - version visuelle : histogrammes, diagramme à bâtons
- Semi-automatique
 - segmentation d'un ensemble de clients
- Automatique
 - reconnaissance d'empreintes digitales
 - recherche de cooccurrences fréquentes
 - recommandations

Applications du Data Mining



Domaines d'application



Exemple1: Marketing direct

- Vous êtes gestionnaire marketing d'un opérateur de télécommunications mobiles :

Les clients reçoivent un téléphone gratuit (valeur 150€) avec un contrat d'un an; vous payez une commission de vente de 250€ par contrat

- **Problème:** Taux de renouvellement (à la fin du contrat) est de 25%



- ❖ Donner un nouveau téléphone à toute personne ayant expiré son contrat coûte cher.

- ❖ Faire revenir un client après avoir quitter est difficile et coûteux.



Exemple1: Marketing direct

- Trois mois avant l'expiration du contrat, **prédire** les clients qui vont quitter :
 - Si vous voulez les garder, offrir un nouveau téléphone.



Exemple 2: Assurances

- Vous êtes un agent d'assurance et vous devez définir un paiement mensuel adapté à un jeune de 18 ans qui a acheté une Ferrari.

Qu'est ce qu'il faut faire ??



Exemple 2: Assurances

- Analyser les données de tous les clients de la compagnie .
 - ❖ La probabilité d'avoir un accident est basée sur ... ?
 - ✓ Sexe du client (M/F) et l'âge, ...
 - ✓ Modèle de la voiture, âge, ...
 - ✓ etc.
 - ❖ Si la probabilité d'avoir un accident est supérieure à la moyenne:
 - ✓ initialiser la mensualité suivant les risques.



Exemple 3: Banque et Télécom

- Si quelqu'un vous a volé votre carte de crédit ou votre mobile ...



❖ Compagnies bancaires

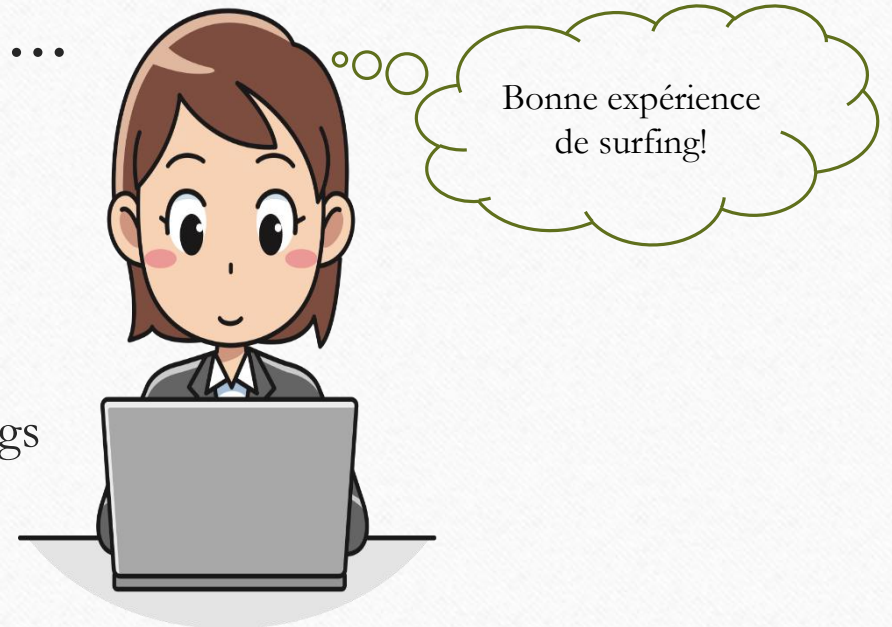
Utiliser les données historiques pour construire un modèle de comportement frauduleux et utiliser le data mining pour identifier des instances similaires.

❖ Compagnies téléphoniques

Analyser les “patterns” qui dérivent du comportement attendu (destinataire, durée, etc.)...

Exemple 4: Web

- **Les logs des accès Web sont analysés pour ...**
 - ✓ Découvrir les préférences des utilisateurs
 - ✓ Améliorer l'organisation du site Web
- **De manière similaire...**
 - ✓ L'analyse de tous les types d'informations sur les logs
 - ✓ Adaptation de l'interface utilisateur/service



Quelques applications concrètes (1/2)

- Visualisation de l'information : liens entre profils facebook (Paul Butler 2010)



Quelques applications concrètes (2/2)

- Amazon, lastfm, netflix...
- Recommandations par co-achats
- Recommandations personnalisées

04/03/13 Amazon.fr : Livres : -5% et livraison gratuite

Chez vous | Promotions | Chèques-cadeaux | Vendre | Aide

Rechercher Livres en français Go

Bonjour Identifiez-vous
Votre compte

Adhérer à Premium

0 Panier

Liste d'envies

Livres Recherche détaillée Nos rubriques Meilleures ventes Précommandes Livres anglais et étrangers Promotions Livres d'occasion Amazon Rachète

Livraison gratuite pour tous les livres sans minimum d'achats

Livres : actu et promotions

- Kindle Fire HD
- Kindle Fire
- Kindle Paperwhite
- Kindle
- 2 livres achetés = 1 gratuit
- Nouveautés Littérature 2013
- Boutique Fitness et minceur
- Nouveautés Polars
- Succès 2012
- Titres en précommande
- Amazon Rachète
- Plus de bonnes affaires

Livres : nos rubriques

- BD et Jeunesse
- BD et Humour
- Jeunesse
- Ados et jeunes adultes
- Manga
- Romans
- Littérature
- Policier et Suspense
- SF, Fantasy et Terreur
- Littérature sentimentale
- Culture et société
- Actu, Politique et Société
- Art, Musique et Cinéma
- Beaux livres
- Dictionnaires et langues
- Esotérisme et Paranormal
- Histoire
- Religions et Spiritualités
- Sciences humaines
- Scolaire
- Scolaire et Parascolaire
- Études supérieures
- Vie pratique
- Calendriers, carnets et agendas
- Cuisine et Vins
- Loisirs créatifs, décoration et bricolage
- Nature, animaux et jardinage
- Santé et Bien-être
- Tourisme et Voyages
- Vie professionnelle
- Droit
- Entreprise et Bourse
- Informatique et Internet
- Sciences, Techniques et Médecine
- Toutes nos rubriques

Livres : nos rubriques

- Livres anglais et étrangers
- Livres d'occasion
- Livres audio

Livres

-5% minimum sur des millions de livres : BD, Manga, Livres pour enfants, Scolaire, Littérature, Romans, Polars, SF, Histoire... et bien plus encore !

Harlan Coben

Ne t'éloigne pas

+ Cliquez ici

A l'honneur cette semaine

Nouveautés 2013

- Les nouveautés
- Littérature
- BD
- Festival d'Angoulême
- Les nouveautés

Les livres adaptés au cinéma

- Lucien
- Zero dark thirty
- L'actualité cinéma

Les nouveautés les plus commandées

- MUSO Littérature
- BD et Humour
- Jeunesse
- Policier et suspense
- Cuisine et vins
- Ados
- Santé et bien-être
- SF, Fantasy et Terreur
- Actu, politique et société

Les recommandations des lecteurs d'Amazon : livres notés 4 étoiles et plus

Amazon Rachète vos livres + Cliquez ici

Nouveauté : "Demain" de Guillaume Musso

Aviez-vous aimé Guillaume Musso ? Découvrez Demain son dernier livre chez XO éditions.

"7 ans après" de Guillaume Musso en Pocket

Vous aimez les romans de Guillaume Musso ? Découvrez 7 ans après en format de poche.

1 livre offert pour 2 livres

10/13 achetés

Nos meilleures ventes Livres

mise à jour toutes les heures

1. 37 jours dans le top 100
Demain
Guillaume Musso
Broché
EUR-23,90 EUR 20,80
2. 115 jours dans le top 100
Indignez-vous l'édition revue et augmentée
Stéphane Hessel
Broché
EUR-9,40 EUR 2,94
3. 48 jours dans le top 100
La France orange mécanique
Laurent Oberthone
Broché
EUR-16,00 EUR 17,10
4. 138 jours dans le top 100
Cinquante nuances plus claires
E. L. James
Broché
EUR-12,90 EUR 16,15
5. 205 jours dans le top 100
Cinquante nuances de Grey
E. L. James
Broché
EUR-12,90 EUR 16,15
6. 138 jours dans le top 100
Cinquante nuances plus sombres
E. L. James
Broché
EUR-12,90 EUR 16,15
7. 12 jours dans le top 100
Belle et bête
Marcia Jacob
Broché
EUR-3,90 EUR 12,83

www.amazon.fr/livre-achat-occasion-litterature-roman/v/ref=sa_menu_lv/275-6811504-74280497_encoding=UTF8&node=301061

Exercice : Data Mining ou non?

	Oui	Non
Rechercher le salaire d'un employé		X
Les hommes achètent de coca et des couches bébé le samedi	X	
Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining		X
Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu	X	

Les Méthodes Data Mining



Définition du Data Mining

- Le data Mining ou la fouille de données présente la découverte d'une connaissance (information intéressante) à partir de grandes quantités de données, par des méthodes automatiques.
 - ✓ Quels types de données?
 - ✓ Qu'est ce qu'une connaissance?
 - ✓ Qu'entend-on par découvrir ou extraire ?
- **Fortement lié à l'apprentissage automatique!**



Apprentissage Automatique

- Apprentissage automatique (Machine Learning)
 - ❖ discipline visant à la construction de règles d'inférence et de décision pour le traitement automatique des données.
 - ❖ Une méthode pour créer un modèle de la réalité à partir de données.
 - ❖ Elle génère ce modèle soit en améliorant un modèle partiel soit en construisant complètement le modèle.
- Cette méthode déployée sur un ordinateur offre à ce dernier la capacité d'apprendre.

Apprentissage Automatique-Définition

« On dit qu'un programme apprend de l'expérience E en ce qui concerne une tâche T et une mesure de performance P , si sa performance sur T , mesurée par P , s'améliore avec l'expérience E ».

Tom Mitchell (1997)

Trois caractéristiques:

- ✓ tâche T
- ✓ mesure de performance à améliorer P
- ✓ source d'expérience E

Apprentissage Automatique-Exemple1

Exemple1: Supposons que votre programme de messagerie électronique surveille les courriels que vous marquez ou non comme indésirable (spam). Ainsi, en fonction de ça, votre programme apprend mieux à filtrer les courriers indésirables. Quelle est la tâche T dans ce cadre?

- Classification des e-mails comme spam et non-spam
- Vous observer marquant vos e-mails comme spam ou non-spam
- Le nombre (la fraction) des e-mails correctement classifiés comme spam ou non spam.
- Aucune de ces réponses, ce n'est pas du machine-learning.



Apprentissage Automatique-Exemple1

Exemple1: Supposons que votre programme de messagerie électronique surveille les courriels que vous marquez ou non comme indésirable (spam). Ainsi, en fonction ça, votre programme apprend mieux à filtrer les courriers indésirables. Quelle est la tâche T dans ce cadre?

- Classification des e-mails comme spam et non-spam —————→ **Task**
- Vous observer marquant vos e-mails comme spam ou non-spam —————→ **Expérience**
- Le nombre (la fraction) des e-mails correctement classifiés comme spam ou non spam. —————→ **Performance**
- Aucune de ces réponses, ce n'est pas du machine-learning.



Apprentissage Automatique-Exemple2

Exemple2: Jeu d'échecs

- **E** = l'expérience de jouer plusieurs fois ce jeu
- **T** = Jouer le jeu d'échecs
- **P** = La probabilité que le programme gagnera le prochain jeu



Apprentissage Automatique-Exemple3

Exemple3: Supposons que nous concevons un algorithme d'apprentissage avec d'historique de nombreuses données météorologiques et que nous lui apprenions à prévoir la météo. Que peut être P?

- La probabilité de prédire correctement le temps d'une date future.
- La tâche de prévision météorologique.
- Le programme examine une grande quantité de données climatiques historiques.
- Aucun d'eux.



Apprentissage Automatique-Exemple3

Exemple3: Supposons que nous concevons un algorithme d'apprentissage avec d'historique de nombreuses données météorologiques et que nous lui apprenions à prévoir la météo. Que peut être P?

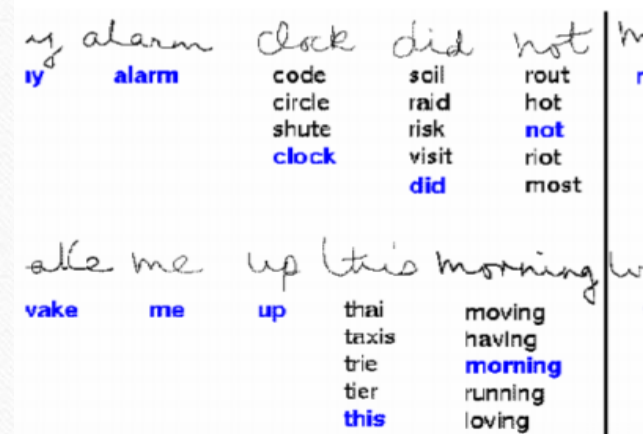
- La probabilité de prédire correctement le temps d'une date future.
- La tâche de prévision météorologique.
- Le programme examine une grande quantité de données climatiques historiques.
- Aucun d'eux.



Apprentissage Automatique-Exemple4

Exemple 4: La reconnaissance de l'écriture manuscrite (handwriting recognition)

- **T** = reconnaissance et classification des mots écrits manuellement à partir des images.
- **E** = une base de données des mots écrits manuellement avec les classifications convenables.
- **P** = Le nombre de mots correctement classifiés.



Apprentissage Automatique-Exemple5

Exemple 5: Conduite d'un robot autonome



- **T** = conduite sur autoroute publique à 4 voies à l'aide de capteurs de vision
- **P** = distance moyenne parcourue avant une erreur (à en juger par le surveillant humain)
- **E** = une séquence d'images et de commandes de pilotage enregistrées lors de l'observation d'un conducteur humain

Apprentissage Automatique-Approches

- Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient :
 - L'apprentissage supervisé
 - L'apprentissage non-supervisé
 - L'apprentissage par renforcement
 - L'apprentissage profond (deep learning)

Apprentissage Automatique-Approches

- Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient :
 - **L'apprentissage supervisé**
 - **L'apprentissage non-supervisé**
 - L'apprentissage par renforcement
 - L'apprentissage profond (deep learning)

Apprentissage supervisé

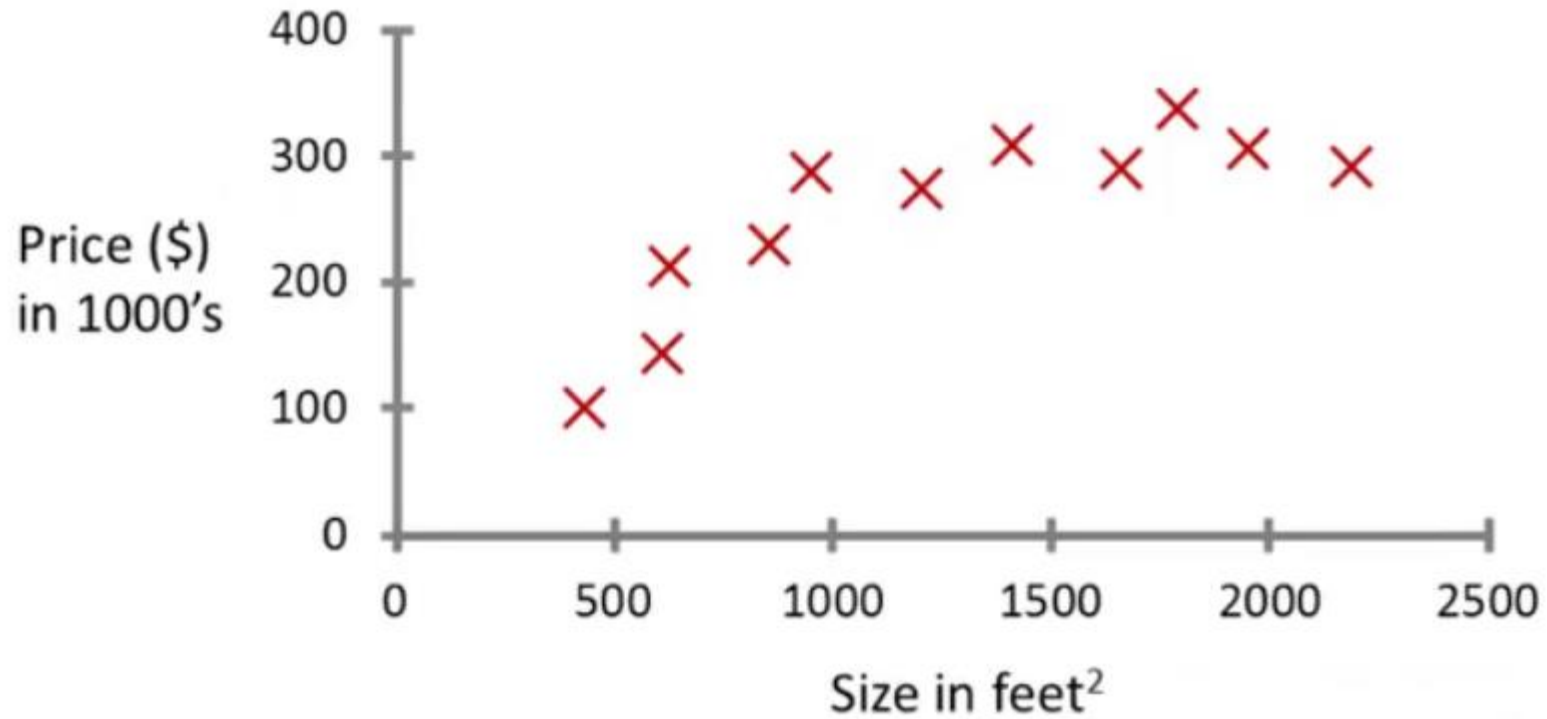
Apprentissage supervisé

- L'apprentissage supervisé se base sur des observations étiquetées ou classées.
- Un expert (superviseur) est employé pour fournir correctement ces étiquettes.
- L'apprenant doit alors trouver ou approximer la fonction qui permet d'affecter la bonne étiquette à ces observations afin de déterminer l'output d'une observation inconnue.
- Il existe 2 types d'algorithmes:
 - Régression (Prédire l'âge d'un embryon à partir de sa taille, son poids, etc)
 - Classification (Prédire qui gagne plus de 50.000\$ à partir de données de recensement)



Apprentissage supervisé: « réponses correctes » offertes

Housing price prediction.

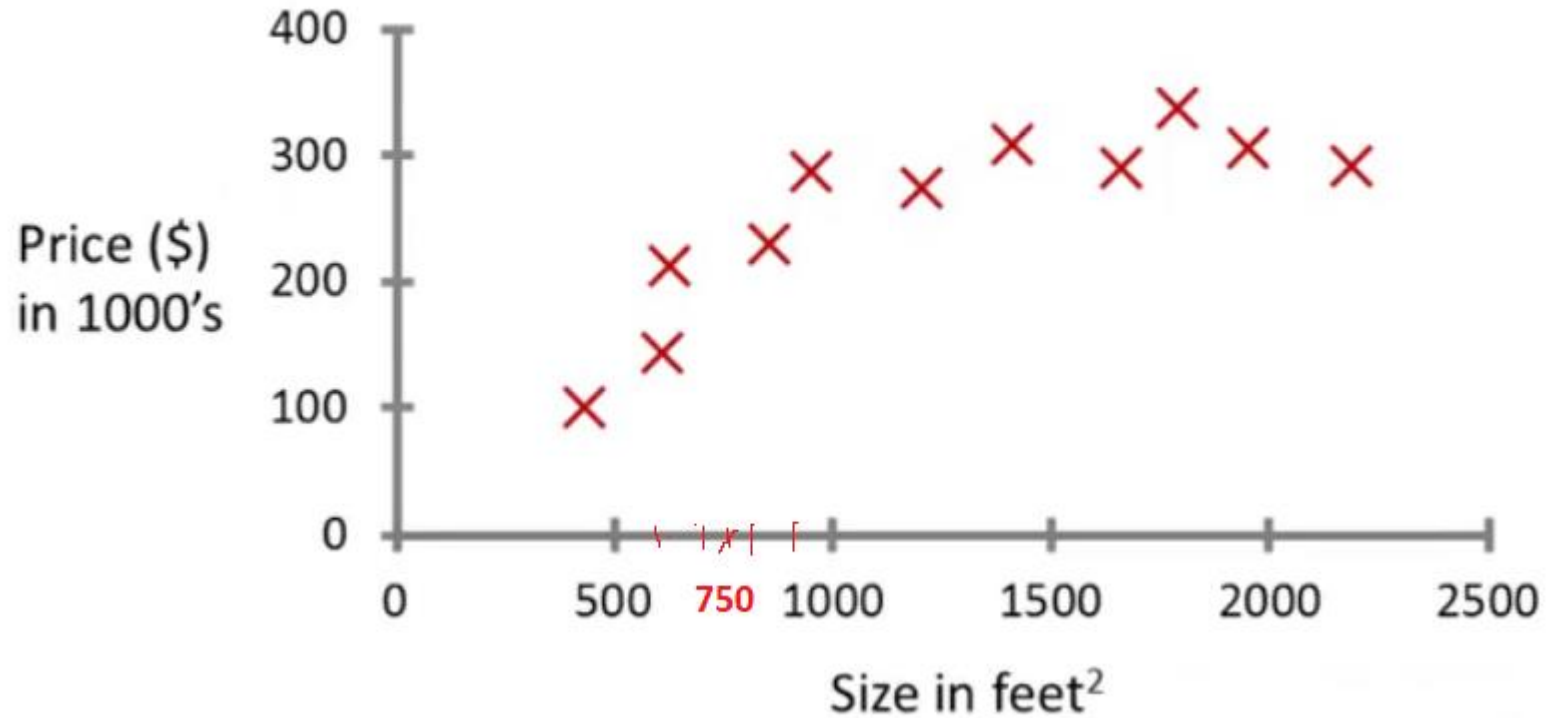


Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

Apprentissage supervisé: « réponses correctes » offertes

Housing price prediction.

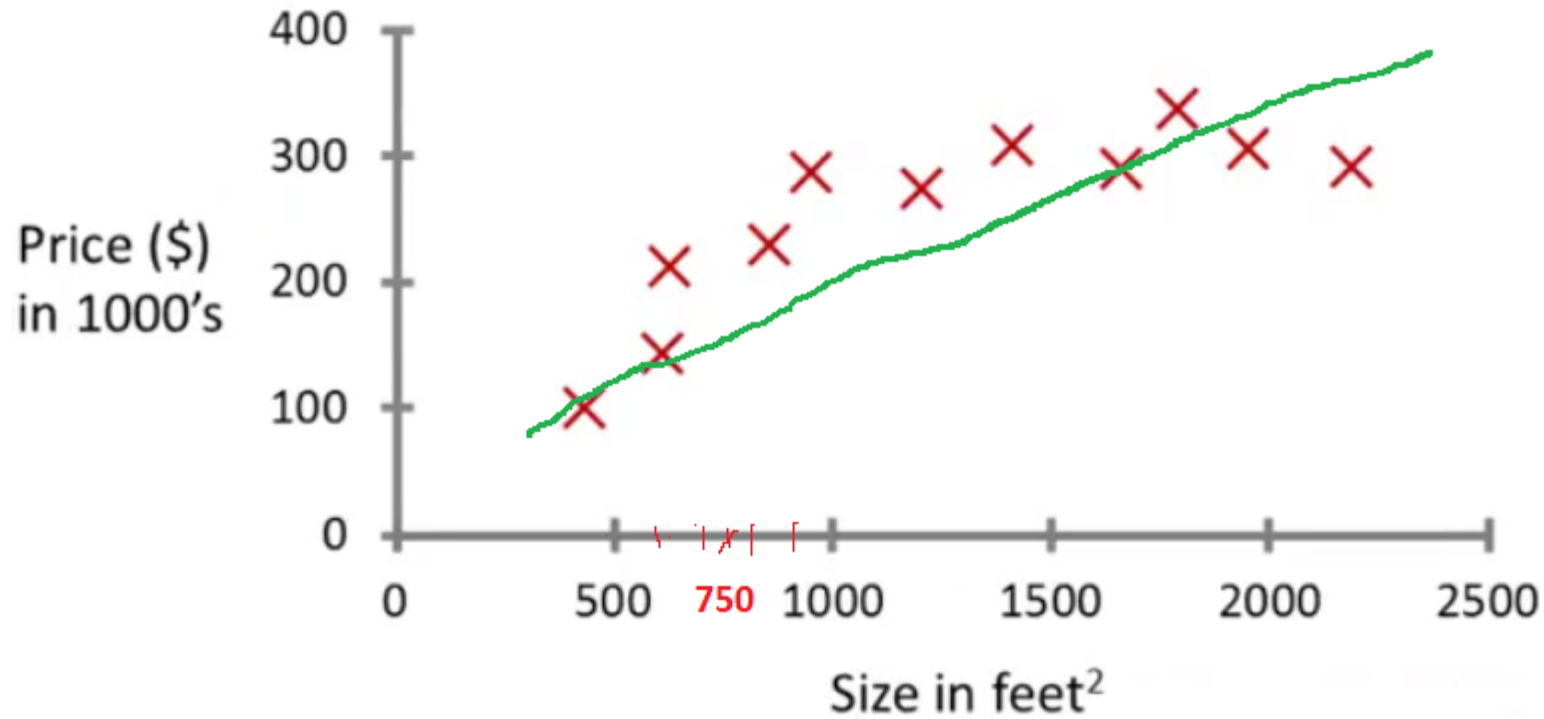


Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

Apprentissage supervisé: « réponses correctes » offertes

Housing price prediction.

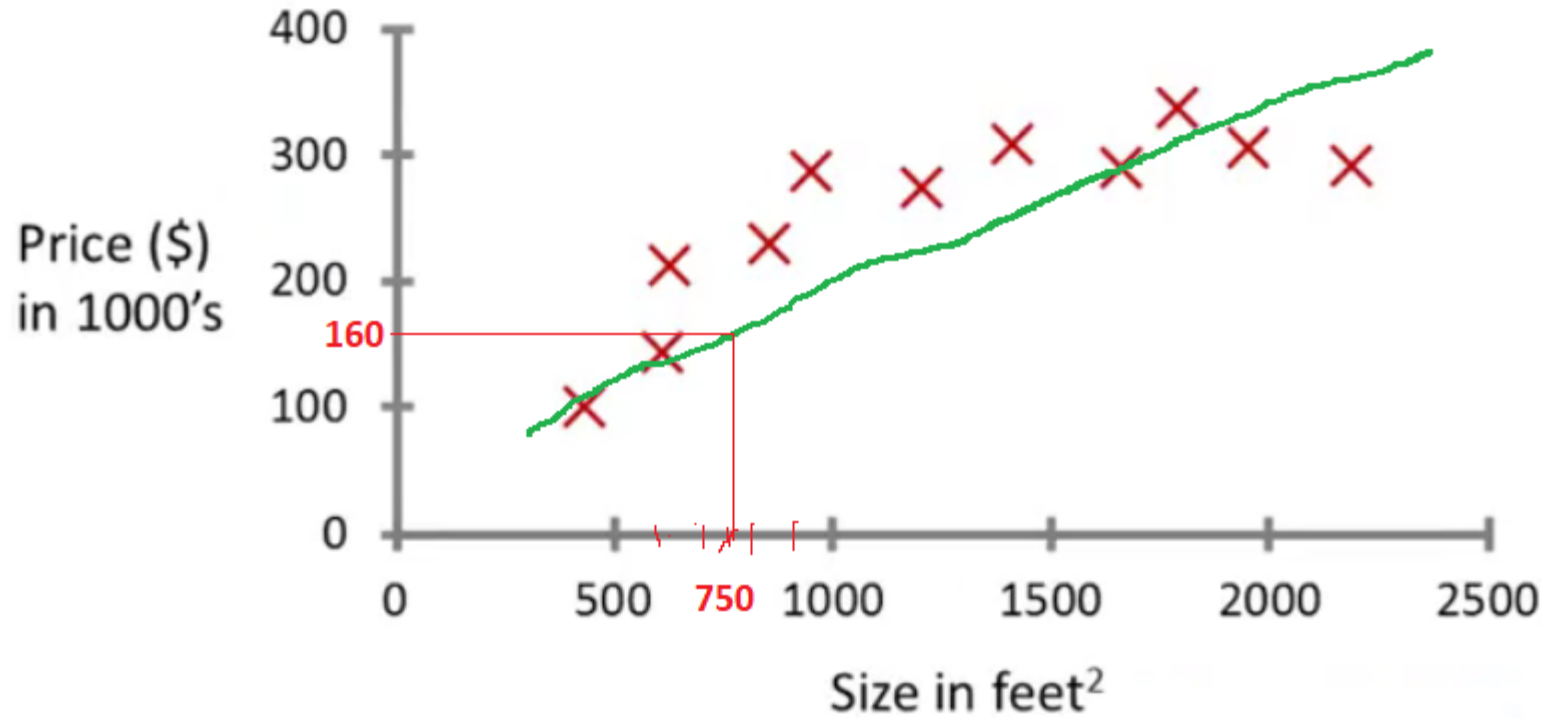


Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

Apprentissage supervisé: « réponses correctes » offertes

Housing price prediction.

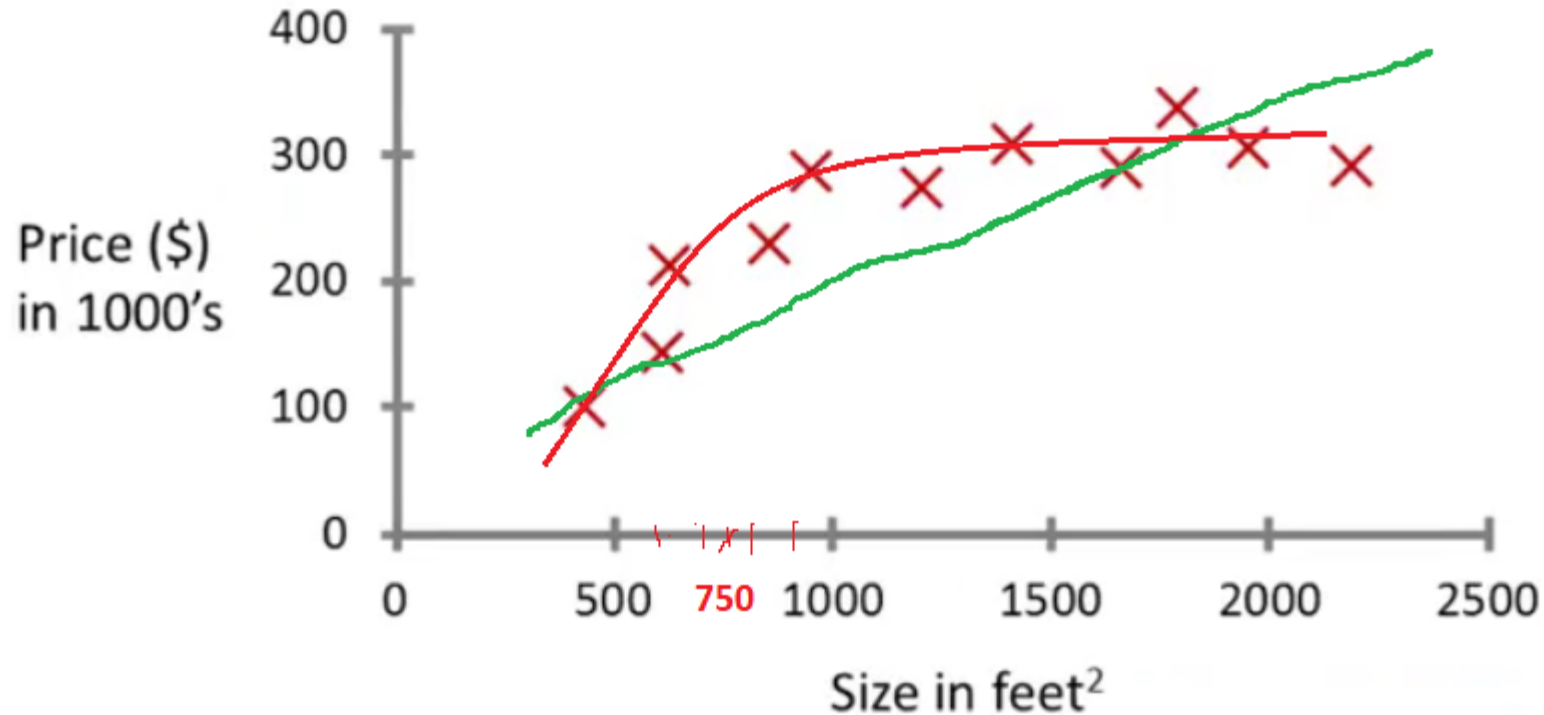


Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

Apprentissage supervisé: « réponses correctes » offertes

Housing price prediction.

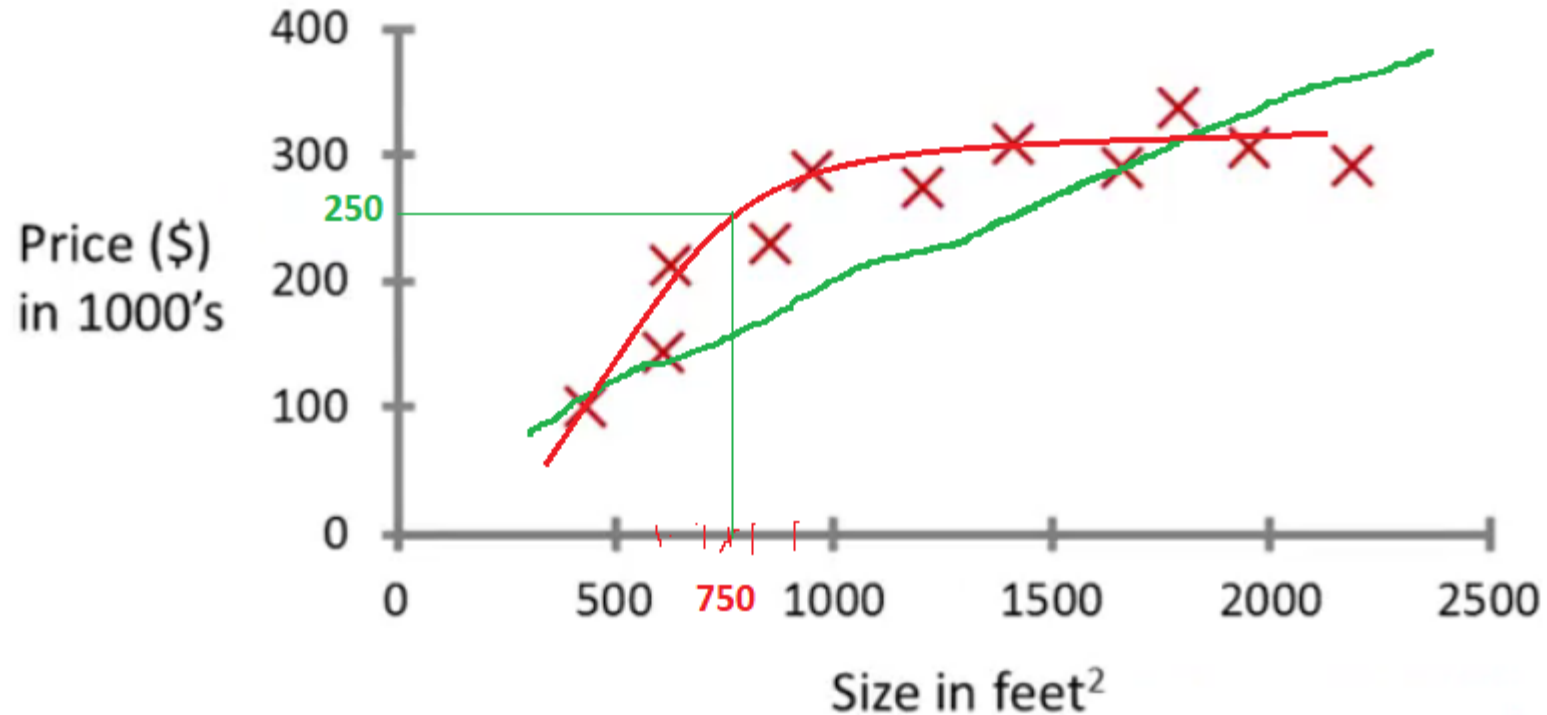


Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

Apprentissage supervisé: « réponses correctes » offertes

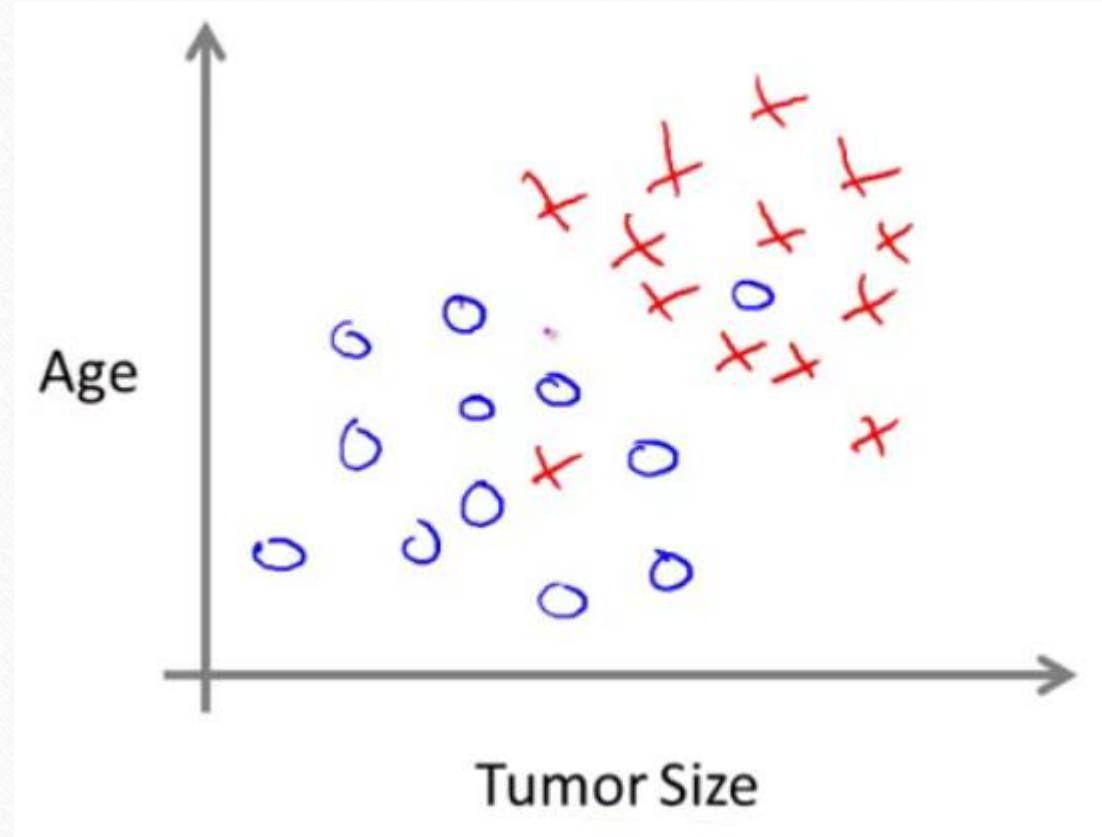
Housing price prediction.



Supposons que vous possédez une maison de **750** pieds carrés et vous espérez vendre la maison. Vous voulez savoir combien vaut votre maison. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Régression: Prédire des valeurs continues (output: prix)

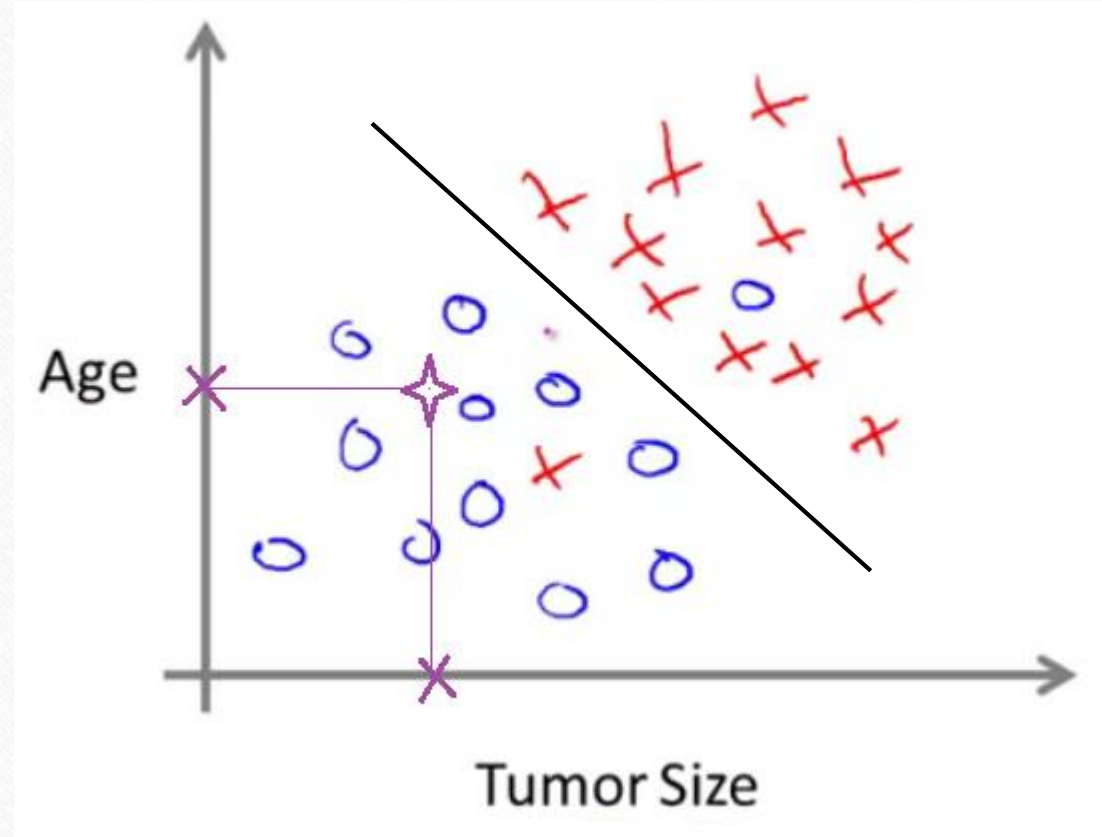
Apprentissage supervisé: « réponses correctes » offertes



Supposons que vous connaissez quelqu'un qui a une tumeur. Vous voulez prédire si c'est malin ou bénin. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Classification: Prédire des valeurs discrètes (0 ou 1 → malin ou bénin)

Apprentissage supervisé: « réponses correctes » offertes



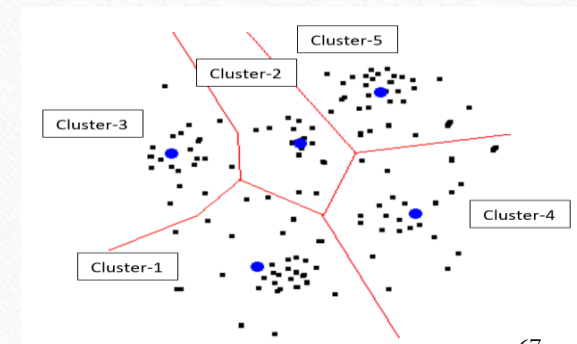
Supposons que vous connaissez quelqu'un qui a une tumeur. Vous voulez prédire si c'est malin ou bénin. Alors, comment l'algorithme d'apprentissage peut-il vous aider ?

Classification: Prédire des valeurs discrètes (0 ou 1 → malin ou bénin)

Apprentissage non supervisé

Apprentissage non supervisé

- Aucun expert n'est disponible.
- L'algorithme doit découvrir par lui-même la structure des données.
 - Par exemple, grouper des exemples de manière à ce que les exemples au sein d'un même groupe se ressemblent suffisamment, et que les exemples de groupes différents soient suffisamment différents.
 - Si on veut automatiquement grouper les utilisateurs du Facebook selon leurs réputations, de sorte que les utilisateurs les plus dignes de confiance (les plus réputés) appartiennent au même cluster → on peut avoir 5 clusters différents (very high, high, medium, low, very low).
- Il existe deux approches de l'apprentissage non supervisé:
 - **Le Clustering (regroupement)**
 - **Les règles d'associations**



Exemple: Google News: chercher des dizaines de milliers de reportages et les regrouper automatiquement

← → ↻ 🔒 https://news.google.com/?hl=fr&gl=FR&ceid=FR:fr

Google Actualités

Rechercher des sujets, des lieux et des sources

À la une

Pour vous

Favoris

www.20minutes.fr/monde/2434471-20190123-venezuela-trump-reconnait-opposant-guaido-comme-president-interim-maduro-riposte

20 MINUTES

ACCUEIL > MONDE


Venezuela: Trump reconnaît l'opposant Guaido comme président par intérim, Maduro riposte

CLASH Contesté, Maduro a donné 72 heures aux diplomates américains pour quitter le pays, alors que des heurts ont éclaté à Caracas...

20 Minutes avec AFP | Publié le 23/01/19 à 22h05 — Mis à jour le 23/01/19 à 22h08

11 COMMENTAIRES 0 PARTAGE

🗨️ 📧 📄 📱 📺 📻 📡 📢 📣 📤 📥 📦 📧 📨 📩 📪 📫 📬 📭 📮 📯 📰 📱 📲 📳 📴 📵 📶 📷 📸 📹 📺 📻 📼 📽 📾 📿



Européennes : des « gilets jaunes » annoncent leur intention de présenter une liste

Sana Hamdi

https://francais.rt.com/international/58310-72-heures-pour-quitter-pays-maduro-venezuela-rompt-relations-etats-unis

Actualité France International Économie Opinions INTERDIT D'INTERDIRE Magazines Documentaires Vidéos RT360 RT vous parle



Nicolas Maduro au balcon du palais présidentiel le 23 janvier.

Enfin de vrais débats !
INTERDIT D'INTERDIRE
animé par FRÉDÉRIC TADDEI

DU LUNDI AU JEUDI
À 19H SUR

RETROUVEZ-NOUS SUR
FREEBOX TV 555

Nicolas Maduro, a déclaré qu'il rompait les relations avec les États-Unis, après qu'ils ont reconnu Juan Guaido, au poste de président par intérim du pays sud-américain. Le chef d'Etat donne 72 heures aux diplomates américains pour quitter le pays.

Le président vénézuélien, Nicolas Maduro, a déclaré le 23 janvier qu'il rompait les relations diplomatiques avec les États-Unis, après que l'administration Trump a reconnu le chef de l'opposition.

Non sécurisé | www.lefigaro.fr/international/2019/01/23/01003-20190123ARTIG00280-venezuela-trump-reconnait-le-leader-de-l-opposition-comme-president-par-interim.php

LE FIGARO · fr

Actualité > International

Venezuela : Guaido se proclame président et obtient la reconnaissance de Trump

Par Le figaro.fr, AFP agence | Mis à jour le 23/01/2019 à 21:54 / Publié le 23/01/2019 à 19:29



CARACAS (VENEZUELA)
23 JANVIER 2019

FIGAROLIVE

Exemple: *La chaîne météo*: prédire le temps pour Mercredi le 13/02/2019 (date de consultation est 28/01/2019) et l'affecter à tous les clusters avec des degrés d'appartenances différentes → Fuzzy clustering

MÉTÉO STATISTIQUE
BIZERTE
Tunisie

MERCREDI
13 février

la chaîne
météo

ALERTE MÉTÉO MONTAGNE ACTUALITÉS VIDÉOS REPORTERS

Les scénarios présentés ici sont les plus probables. ceux correspondants aux pourcentages restants ne sont pas détaillés car trop peu plausibles.

Scénario de probabilité
55% de probabilité

Averses ou pluie intermittente



11°
Température min

12°
Température max

Scénario 2
26% de probabilité

Nuages et soleil



12° Min

14° Max

Scénario 3
17% de probabilité

Ensoleillé



13° Min

14° Max

Logiciels Data Mining

- WEKA
- ORANGE
- TANAGRA
- SIPINA
- Logiciel R
- RapidMiner
- KNIME
- AlphaMiner
- Statistica dataMiner
- http://eric.univ-lyon2.fr/~ricco/tanagra/fr/contenu_tutoriaux_comparaison_logiciels.html
- http://chirouble.univ-lyon2.fr/~ricco/data-mining/logiciels/revue_rapide_des_logiciels_sur_le_site_kdnuggets.pdf

Méthodologies de travail



Démarche Méthodologique KDD/ECD

KNOWLEDGE DATA DISCOVERY
EXTRACTION DE CONNAISSANCES À PARTIR DE DONNÉES

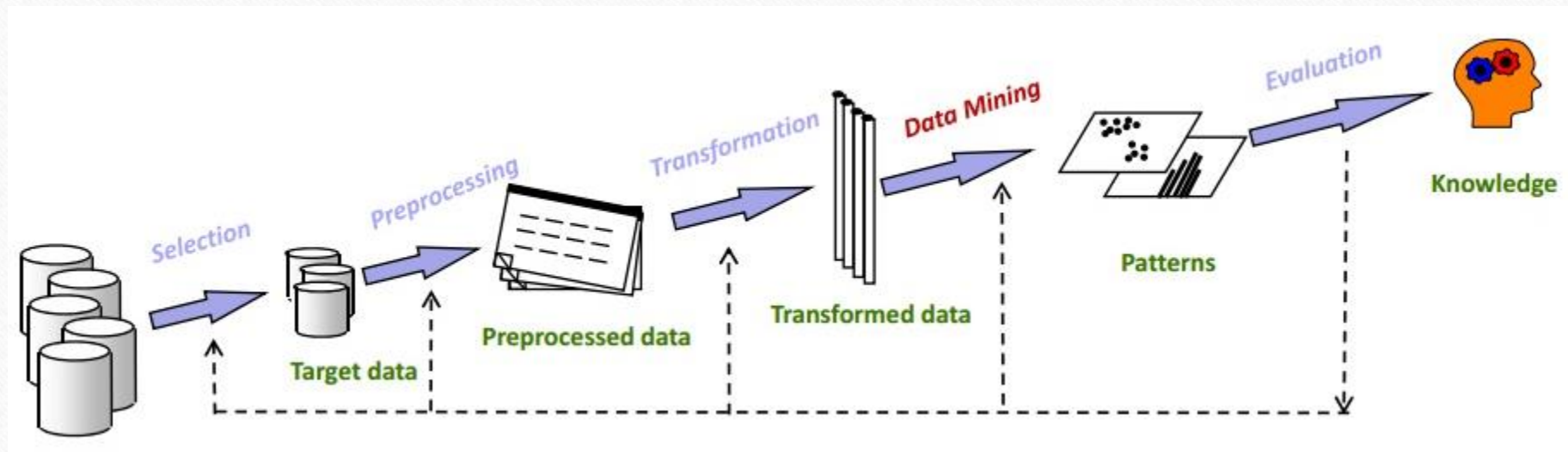
KDD: Définition

- Un processus 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 et des motifs valides, utiles et exploitables à partir des grandes quantités de données et 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: Illustration



KDD: Phases principales

1. Développer et comprendre le domaine de l'application

- C'est le pas initial de ce processus.
- Préparer 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 (changer nominal à numérique par exemple).



Données transformées

- ❖ Une fois que toutes ces étapes seront terminées, les étapes suivantes seront liées à la partie du 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.

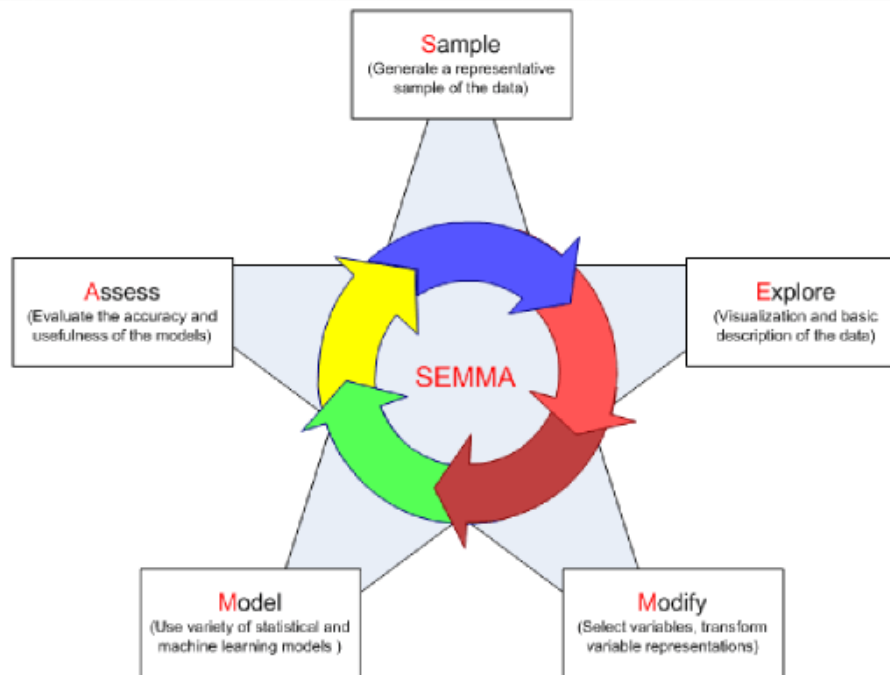
Démarche Méthodologique SEMMA

SAMPLE, EXPLORE, MODIFY, MODEL, ASSESS

SEMMA: 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 « **S**ample, **E**xplore, **M**odify, **M**odel, **A**sses »

SEMMA: Illustration



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, mais assez petit pour être manipulé rapidement.

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)

- 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)

- modélise les 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, etc.

SEMMA: Phases principales

5. Assess (Evaluer)

- évalue l'utilité et la fiabilité des résultats du processus de Data Mining et estime comment il va s'exécuter

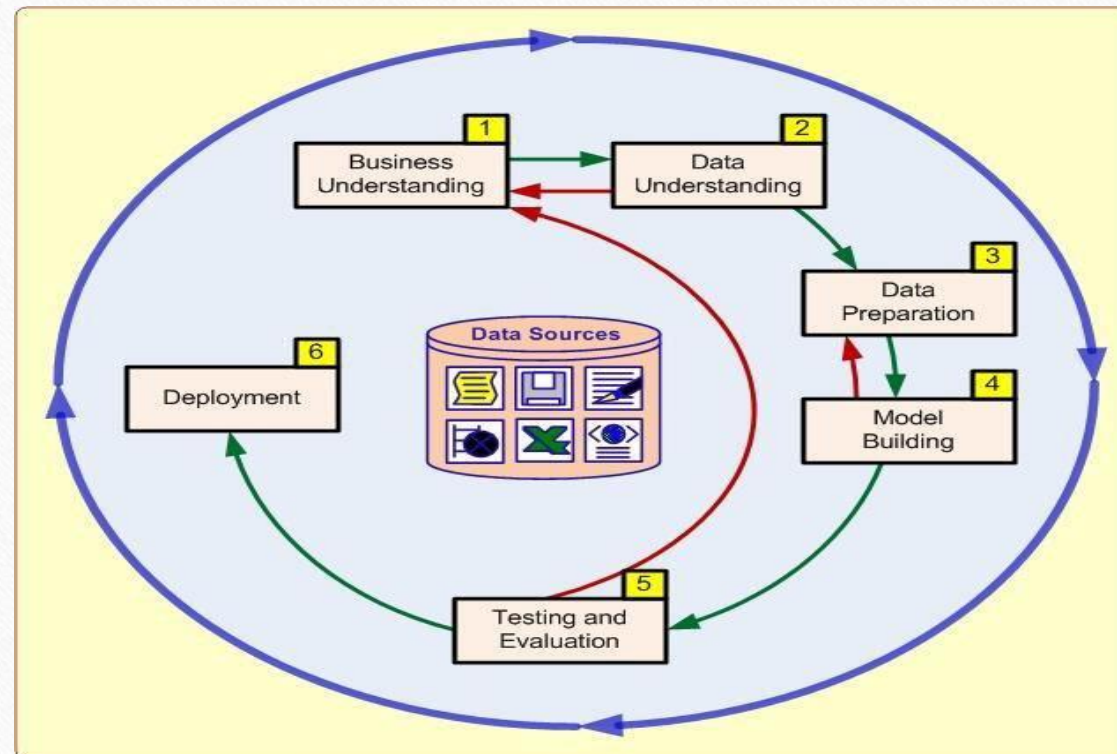
Démarche Méthodologique CRISP-DM

CRoss-Industry Standard Process for Data Mining

CRISP: Démarche méthodologique

CRISP (Cross Industry Standard Process for Data Mining) est développée au départ (années 60) par IBM pour réaliser les projets Data Mining. Elle présente aujourd'hui l'unique méthode utilisable efficacement pour tous les projets Data Mining et Data Science en général.

CRISP: Illustration



CRISP: Étapes

1. La compréhension du problème métier (de l'application)

- La première étape consiste à bien comprendre les éléments métiers et problématiques que le Data Mining vise à résoudre ou à améliorer.

2. La compréhension des données

- Cette phase vise à déterminer précisément les données à analyser (cibles), à identifier la qualité des données disponibles et à faire le lien entre les données et leur signification d'un point de vue métier.

CRISP: Étapes

3- La construction du Data Hub

- Cette phase de **préparation des données** regroupe les activités liées à la construction de l'ensemble précis des données à analyser, faite à partir des données brutes. Elle inclut ainsi le classement des données en fonction de critères choisis, le nettoyage des données, et surtout leur recodage pour les rendre compatibles avec les algorithmes qui seront utilisés.
- La paramétricité des données numériques et leur recodage en données catégorielles sont extrêmement importantes et à réaliser avec soin afin d'éviter que les algorithmes utilisés donnent des résultats faux dans la phase suivante. Toutes ces données doivent en effet être centralisées dans une base de données structurée et qui porte le nom de Data Hub

CRISP: Étapes

4. La modélisation

- C'est la phase de Data Mining proprement dite. La **modélisation** comprend le choix, le paramétrage et le test de différents algorithmes ainsi que leur enchaînement, qui constitue un modèle. Ce processus est d'abord descriptif pour générer de la connaissance, en expliquant pourquoi les choses se sont passées. Il devient ensuite prédictif en expliquant ce qu'il va se passer, puis prescriptif en permettant d'optimiser une situation future.

5. L'évaluation

- L'évaluation vise à vérifier le(s) modèle(s) ou les connaissances obtenues afin de s'assurer qu'ils répondent aux objectifs formulés au début du processus. Elle contribue aussi à la décision de déploiement du modèle ou, si besoin est, à son amélioration. A ce stade, on teste notamment la robustesse et la précision des modèles obtenus.

CRISP: Étapes

6- Le déploiement

- Il s'agit de l'étape finale du processus. Elle consiste en une mise en production pour les utilisateurs finaux des modèles obtenus. Son objectif : mettre la connaissance obtenue par la modélisation, dans une forme adaptée, et l'intégrer au processus de prise de décision.
- Le déploiement peut ainsi aller, selon les objectifs, de la simple génération d'un rapport décrivant les connaissances obtenues jusqu'à la mise en place d'une application, permettant l'utilisation du modèle obtenu, pour la prédiction de valeurs inconnues d'un élément d'intérêt.

Discussion

- Une similitude importante entre les processus est l'aspect itératif
- Les processus se différencient plutôt pour la première étape de « Business Understanding » et la dernière étape de « Using discovery knowledge » qui existent dans CRISP-DM et KDD mais pas dans SEMMA.
- L'utilisation de CRISP-DM et KDD a diminué parce qu'ils définissent ce qu'il faut faire et pas comment le faire. A cause de ça, plusieurs équipes ont commencé à développer leur propre méthodologie de leur côté.

Merci pour votre attention !

Dr. Sana Hamdi

Maitre Assistante en Informatique à l'INSAT

Membre du Laboratoire LIPAH (FST-Tunisie)
et du Laboratoire SAMOVAR (Telecom SudParis-France)

