

Fouille de données 1

Introduction au datamining

Master MEDAS

Béa Arruabarrena – MCF CNAM Paris – DICEN IDF

Sommaire

1. Introduction
2. Définitions
3. Objectifs du Data Mining (DM)
4. Démarche DM
5. Type de données et métrique
6. Exemples de cas d'usages
7. Méthodes de Data Mining
8. Logiciel de DM
9. Ressources et bibliographie

Introduction

Contexte scientifique

Avant 17^e siècle : Science empirique

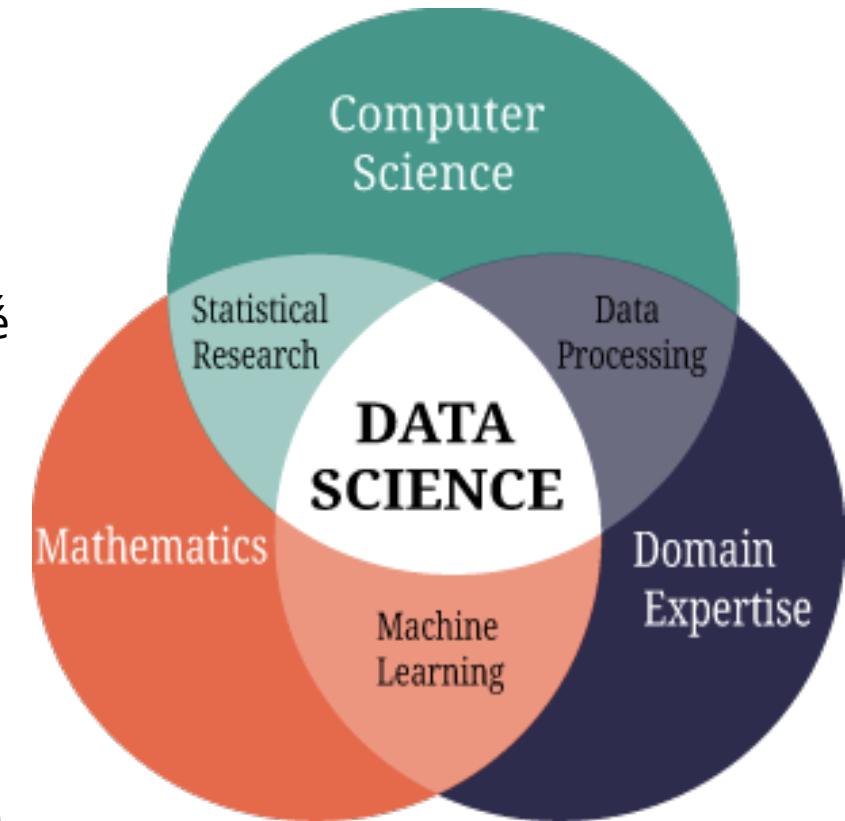
Du 17^e - 1950 : Science théorique

De 1950-90 : Sciences computationnelles («Computational science») : primat au modèle

- Beaucoup de disciplines se sont développées sur le calcul – (physique, biologie, sociologie)
- Logique de simulation : trouver des modèles proches de la réalité
- Développement de la statistique (Fisher, Student,...)
- Données rares et chères, données collectées à partir de plan d'expériences
- Priorité à l'explication des phénomènes observés

De 1990 - Aujourd'hui : Nouveau paradigme : Sciences des données «data science» : primat à la donnée

- Situation d'émergence du datamining Mégadonnées : données omniprésentes, abondantes (nouveaux instruments, simulations)
- Stockage : capacité à gérer et stocker des volumes gigantesques
- Automatisation des processus : réseau connecté /Internet
- Décision quasi en temps réel
- Priorité à la compréhension des phénomènes observés



Contexte sociétal

Données ++

Big Data/Mégadonnées : augmentation sans cesse de données générées par le web et les entreprises

- Twitter : 50M de tweets /jour (=7 téraoctets)
- Facebook : 10 téraoctets /jour
- Youtube : 50h de vidéos uploadées /minute 2.9 million de mail /seconde

Puissance de calcul ++

- Loi de Moore
- Calcul massivement distribué (BDD)

Problématiques nouvelles

- Gestion de gros volume de données
- Crédit à la valeur ajoutée : Intérêt : extraction de connaissances des big data

Contexte- Evolution disciplinaire

Concepts synonymes

- Fouille de données (FD)
- Exploration de données (ED)
- Extraction de connaissances (ECD, KDD)

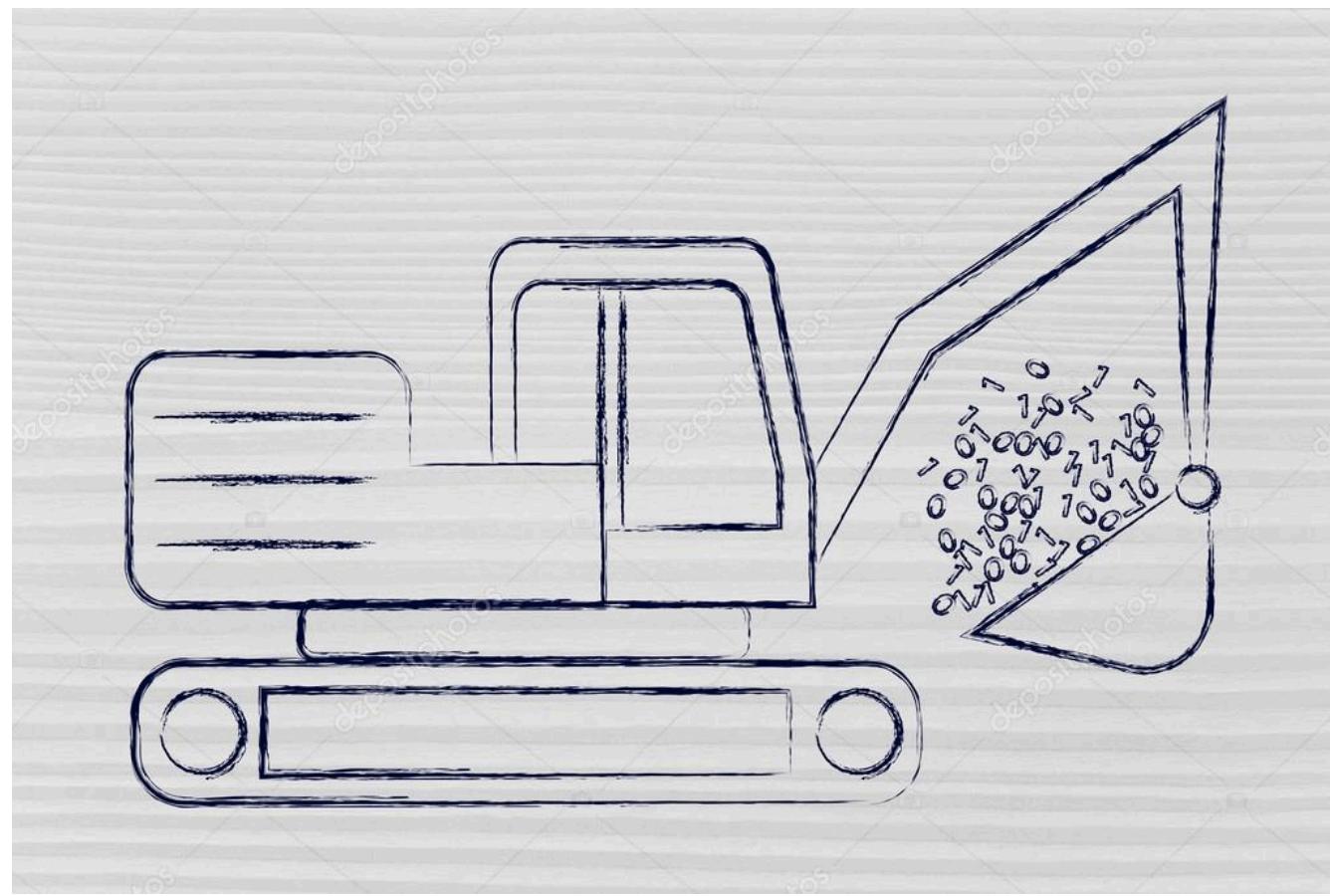
Recherche scientifiques

- Workshops & Conférences internationales depuis 1991
- August 24th-27th 2008 KDD '08: The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas , NV USA
- Data Mining and Knowledge Discovery Journal (1997)
- Special Interest Group Knowledge Discovery in Databases (1999) de l'Association for Computing Machinery (ACM)

Métaphore minière

Par analogie avec la recherche de « pépites d'or » dans un gisement la fouille de données vise :

- à extraire des **connaissances** cachées par analyse globale
- à découvrir des **modèles** (“patterns”) difficiles à percevoir car:
 - le volume de données est très grand
 - le nombre de variables à considérer est important
 - ces “patterns” sont imprévisibles (même à titre d'hypothèse à vérifier)



Définitions

Définitions

« Le datamining, ou fouille de données, est l'ensemble des **méthodes et techniques** destinées à l'**exploration** et l'**analyse** de bases de **données**, souvent de **grande taille**, parfois **hétérogènes**, afin de **détecter**, de façon **automatique** ou semi-automatique des **règles**, des **associations**, des **structures** ou des **tendances** permettant de restituer l'information utile à la prise de **décision** ». Stéphane Tuféry, Data Mining et statistique décisionnelle, L'intelligence des données, 2006.

Fouille de données : en anglais - *Data Mining*

- Données complexes, hétérogènes, évolutives et volumineuses
- Méthode statistiques et d'apprentissage automatique (*machine learning, IA*)
- Formalismes de stockage et de traitement distribués des données (BDDR, NoSQL, Hadoop, MapReduce, Spark ...)

Objectifs du datamining

Objectifs du DM

Décrire :

Mettre en évidence des informations obtenues dans le jeu de données, mais masquées par le volume de données

- Liens entre des variables et représentation synthétique : analyses factorielles
- Groupes d'individus : classification

Prédire :

Extrapoler de nouvelles informations à partir de des données disponibles

- Scoring
- Autres méthodes prédictives, régression

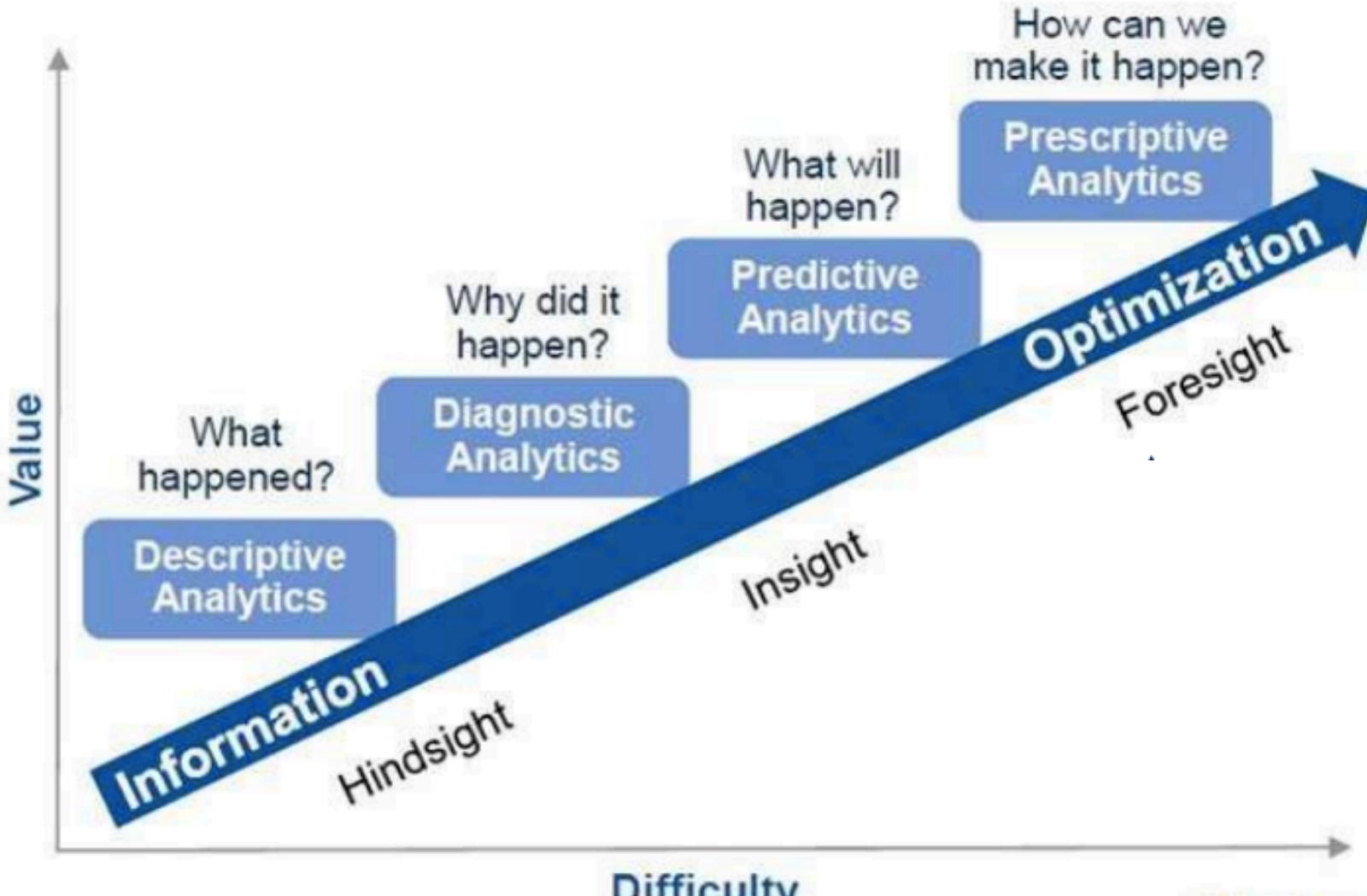
Prescrire :

A partir des comprendre analyse descriptive et prédictive et des analyses mathématiques plus avancées. Les analyses prescriptives anticipent et suggèrent des options de décision sur la manière de tirer parti d'une opportunité future ou d'atténuer un risque futur, et montrent les implications de chaque option de décision.

- Optimisation des process
- Automatisation

Objectifs du DM

le chnam



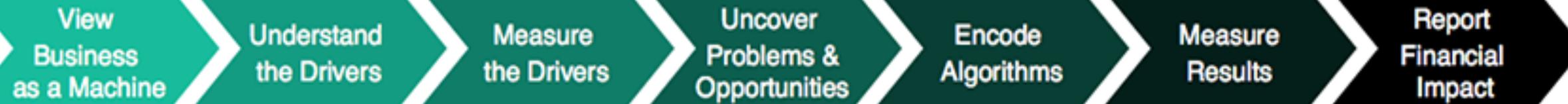
Démarche DM

Démarche DM

CRISP-DM signifie **Cross Industry Standard Process for Data Mining¹**.

« Il s'agit d'un Modèle de Processus de **data mining** qui décrit une approche communément utilisée par les experts en data mining pour résoudre les problèmes qui se posent à eux. Des sondages effectués en 2002, 2004, et 2007 montrent qu'il s'agit de **la méthode principalement utilisée par les data miners**. Cette méthode a été créée par un consortium formé des compagnies NCR, SPSS, et Daimler-Benz. Le processus définit une hiérarchie consistant de phases majeures, de tâches générales, de tâches spécialisées, et d'instances de processus. »

(Wikipédia :
https://fr.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)



Step 1:
Isolate business unit

Step 2:
Define objectives.
Define machine in terms
of people and processes

Step 3:
Collect outcomes in terms
of feedback. Feedback
identifies problems.

Step 1:
Investigate if objectives
are being met

Step 2:
Synthesize outcomes

Step 3:
Hypothesize drivers

Step 1:
Collect Data

Step 2:
Develop KPIs

Step 1:
Evaluate performance vs
KPIs

Step 2:
Highlight potential
problem areas

Step 3:
Review process and
consider what could be
missed or needed to
answer questions

Step 1:
Develop algorithms to
predict and explain
problem

Step 2:
Tie financial value of
individual decisions to
optimize for profit

Step 3:
Use recommendation
algorithms to improve
decisions

Step 1:
Capture outcomes after
decision making system
is implemented

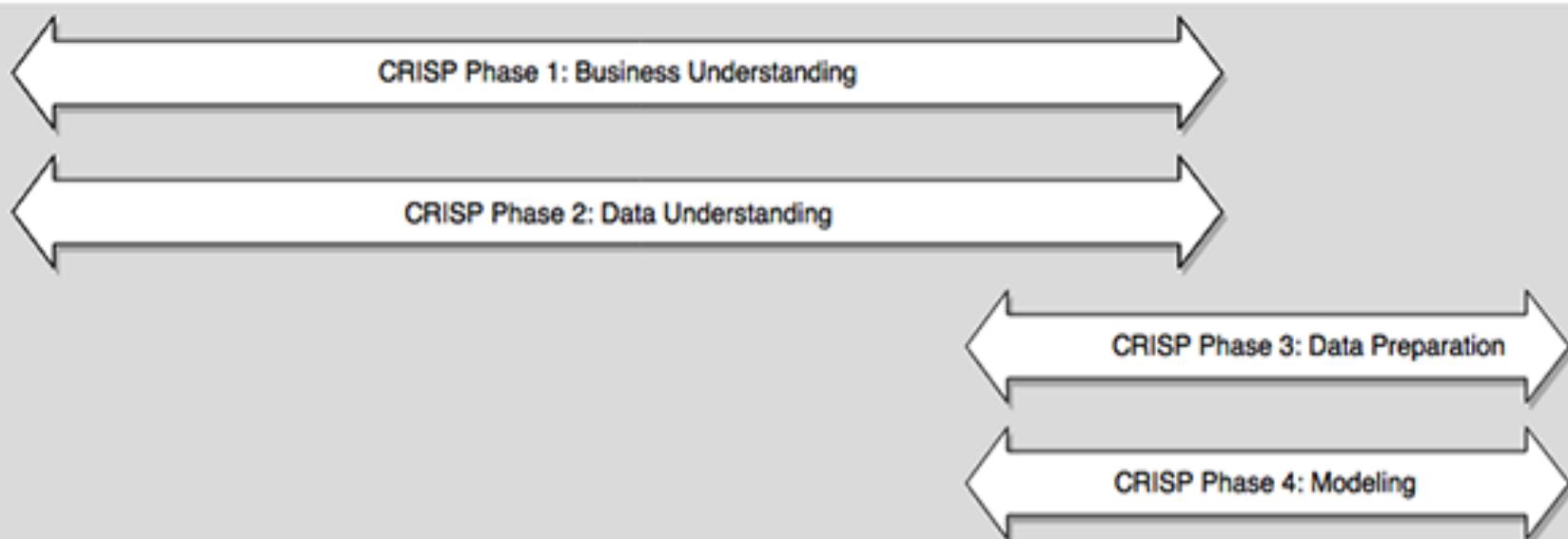
Step 2:
Synthesize results in
terms of good and bad
outcomes identifying what
was done and what
happened

Step 3:
Visualize outcomes over
time to determine
progress

Step 1:
Measure actual results.

Step 2:
Tie to financial benefits

Step 3:
Report financial benefit of
algorithms to key
stakeholders



Démarche DM

1. Identifier le problème

- Cerner les objectifs
- Trouver les sources
- Définir les cibles
- Vérifier les besoins

2. Préparer les données

- Préciser les sources
- Comprendre les données
- Collecter les données
- Nettoyer les données
- Transformer les données
- Intégrer les données

3. Explorer des modèles

- Choisir une méthode
- Echantillonner sur un groupe
- Valider sur le reste
- Calculer le % d'erreurs

4. Utiliser le modèle

- Observer le fonctionnement du modèle
- Recommander des actions

5. Suivre le modèle

- Bâtir des estimateurs
- Corriger et affiner le modèle

Types de données & métriques

Type de données

Les données quantitatives des données qualitatives.

Les données quantitatives sont des valeurs qui décrivent une quantité mesurable/Combien ? (moyenne, des comparaisons (égalité/ différence, infériorité/supériorité, etc.).

- ***les données quantitatives continues*** qui peuvent prendre n'importe quelle valeur dans un ensemble de valeurs : (exemple : la température, le PIB, le taux de chômage)
- ***les données quantitatives discrètes***, qui ne peuvent prendre qu'un nombre limité de valeurs dans un ensemble de valeurs (le nombre d'enfants par famille, le nombre de pièces d'un logement, etc.)

Les données qualitatives décrivent quant à elles des qualités ou des caractéristiques. « quel type » ou « quelle catégorie ». Ces valeurs ne sont plus des nombres, mais un ensemble de modalités.

On ne peut pas faire de calcul sur ces valeurs, même dans l'éventualité où elles prendraient l'apparence d'une série numérique. Elles peuvent toutefois être comparées entre elles et éventuellement triées.

- ***les données qualitatives nominales*** (ou catégorielles), dont les modalités ne peuvent être ordonnées. Par exemple : la couleur des yeux (bleu, vert, marron, etc.), le sexe (homme, femme), la région d'appartenance (68, 38, etc.) ;
- ***les données qualitatives ordinaires***, dont les modalités sont ordonnées selon un ordre « logique ». Par exemple : les tailles de vêtements (S, M, L, XL), le degré d'accord à un test d'opinion (fortement d'accord, d'accord, pas d'accord, fortement pas d'accord).

Sources de données

Capteurs →	variables quantitatives, qualitatives, ordinaires
Texte →	Chaîne de caractères
Parole →	Séries temporelles
Images →	données 2D
Videos →	données 2D + temps
Réseaux →	Graphes
Flux →	Logs, coupons. . .
Etiquettes →	information d'évaluation

Niveaux de structuration des données

Niveau de structuration	Modèle de données	Exemples	Facilité de traitement
Structuré	Système de données relationnel objet/colonne	Base de données d'entreprise...	Facile (indexé)
Semi-structuré	XML, JSON, CSV, logs	API Google, API Twitter, web, logs...	Facile (non indexé)
Non structuré	Texte, image, vidéo	web, e-mails, documents...	Complexé

Les données analysées

Un tableau de données/matrice :

- N lignes : les individus, les objets d'étude
- P colonnes : les variables, les caractéristiques des objets

Une base de données :

- des tables et des tableaux
- des liens entre les tables : un client (dans la table des clients) a acheté des produits (dans la table des produits)

Un entrepôt de données (*data warehouse/Data lake*) :

- mise en commun de bases de données
- agrégation de valeurs : nombre de commandes par enseigne et par mois d'un produit

Métriques des algorithmes

le chnam

Les algorithmes de DM s'appuient sur la notion de similarité dans l'espace X des données.

La similarité est traduite par la notion de distance.

- distance euclidienne : $x, z \in \mathbb{R}^d$, on a

$$d(x, z) = \|x - z\|_2 = \sqrt{\sum_{j=1}^d (x_j - z_j)^2} = \sqrt{(x - z)^T (x - z)}$$

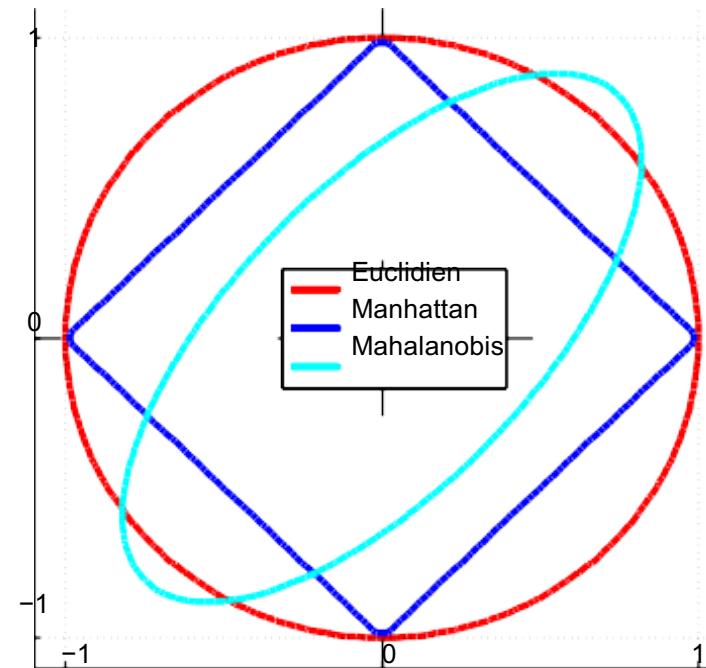
- distance de manhattan

$$d(x, z) = \|x - z\|_1 = \sum_{j=1}^d |(x_j - z_j)|$$

- distance de mahalanobis

$$d(x, z) = \sqrt{(x - z)^T \Sigma^{-1} (x - z)}$$

$\Sigma \in \mathbb{R}^{d \times d}$: matrice carrée définie positive



Exemples

Exemples d'usages

le cnam

Domaine	Usages	Analyse
Santé/ Sciences de la vie	Médecine : patients et maladies, essais cliniques, génomique : gènes, patients, tissus Analyse du génome, mise au point de médicaments Diagnostic	lien entre tabagisme et maladies cardio-vasculaires lien entre tabagisme et cancer du poumon maladies génétiques : mutation → gène détérioré → protéine non produite → maladie Prédiction
Marketing/Entreprise	Gestion de la relation client, fichiers clients, traces d'usage (site web, communication mobile), achats, création de profils clients, ciblage de clients potentiels et nouveaux marchés	évaluation du risque de défaillance pour un crédit typologie des clients recommandation de produits
Industrie	senseurs : température, vibration, images, analyse physico-chimique	fonctionnement et maintenance d'un matériel Prédiction sur une défaillance future
Finances	minimisation de risques financiers	Scoring pour un prêt
Assurance	Détection de fraudes pour les assurances	évaluation du risque
Sécurité	Détection intrusion, spam, e-commerce, détection d'intrusion, recherche d'informations	évaluation du risque

Exemple : e-commerce

Ciblage (Targeting)

- Tracer les séquences de clics des visiteurs (Web analytics)
- Analyser les caractéristiques des acheteurs (âge, sexe, profession, etc.)
- Faire du "targeting" lors de la visite d'un client potentiel
- Prédire la rentabilité d'une campagne marketing

Systèmes de recommandation

- Les clients notent les produits ! Comment tirer profit de ces données pour proposer des produits à un autre client ?
- La technique dit de **filtrage collaboratif** pour regrouper les clients ayant les mêmes “goûts”.

Exemple : analyse des risques

Assurance : Détection de fraudes pour les assurances

- Analyse des déclarations des assurés par un expert afin d'identifier les cas de fraudes.
- Applications de méthodes statistiques pour identifier les déclarations fortement corrélées à la fraude.

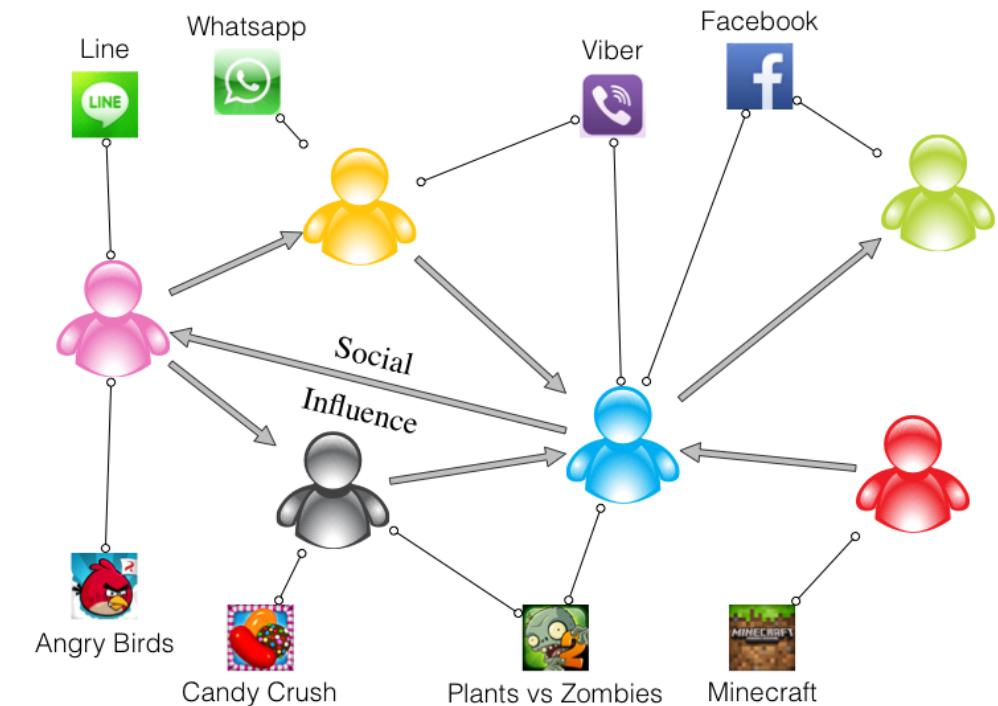
Banque : Prêt Bancaire -Scoring

- Objectif des banques : réduire le risque des prêts bancaires.
- Créer un modèle à partir de caractérisques des clients pour discriminer les clients à risque des autres.
- Crédit de score client

Exemple : marketing

Opinion mining / sentiment analysys

Consiste à analyser l'opinion des usagers sur les produits d'une entreprise à travers les commentaires sur les réseaux sociaux et les blogs



Finance - Trading Advisor/Algorithme de trading

- Application boursière
 - conseil en achat / vente d'actions
- Données de base
 - historique des cours
 - portefeuille client
- Analyse du risque

Introduction aux algorithmes de trading - Xavier Dupré - 14 avril 2013

http://www.xavierdupre.fr/site2013/documents/reports/finance_autos_trat.pdf

Attrition (Churn Analysis)

« Le **taux d'attrition** (ou *churn*, de l'anglais *to churn up* : « brasser », « agiter ») est, au cours d'une période donnée, la proportion de clients perdus ou ayant changé de produit et service de la même entreprise. Ce terme est principalement utilisé dans les secteurs des télécommunications et bancaire, notamment autour de la fidélisation aux offres, mesurée par le taux de fidélité. » (Wikipédia)

- Domaine d'application télécom/bancaire
- Sources de données : Bases de données des clients et des appels et Fichiers des réclamations
- Qui sont les clients le plus susceptibles de partir ?

Exemple : <https://docs.microsoft.com/fr-fr/azure/machine-learning/studio/azure-ml-customer-churn-scenario>

Détection des fraudes en finances

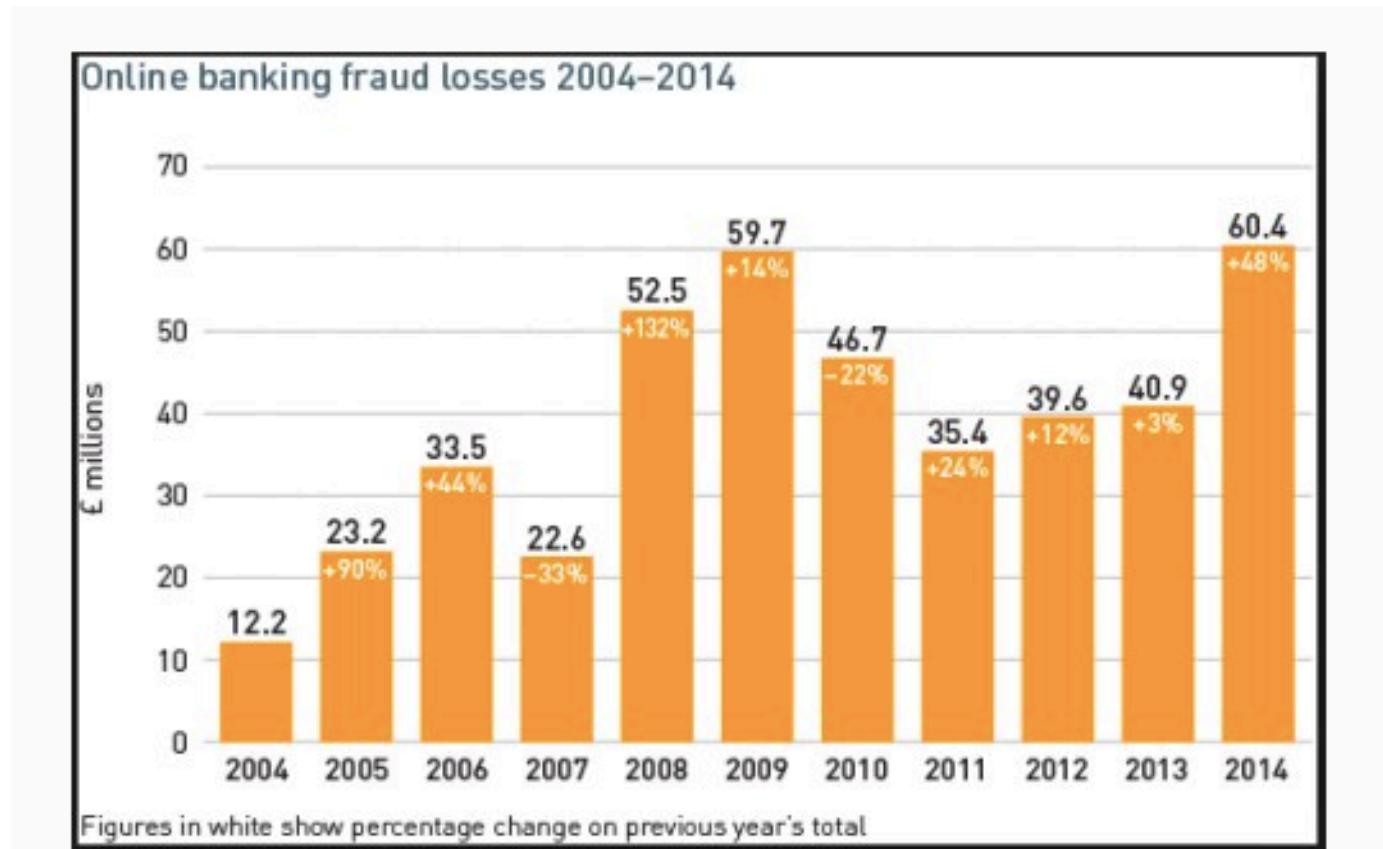
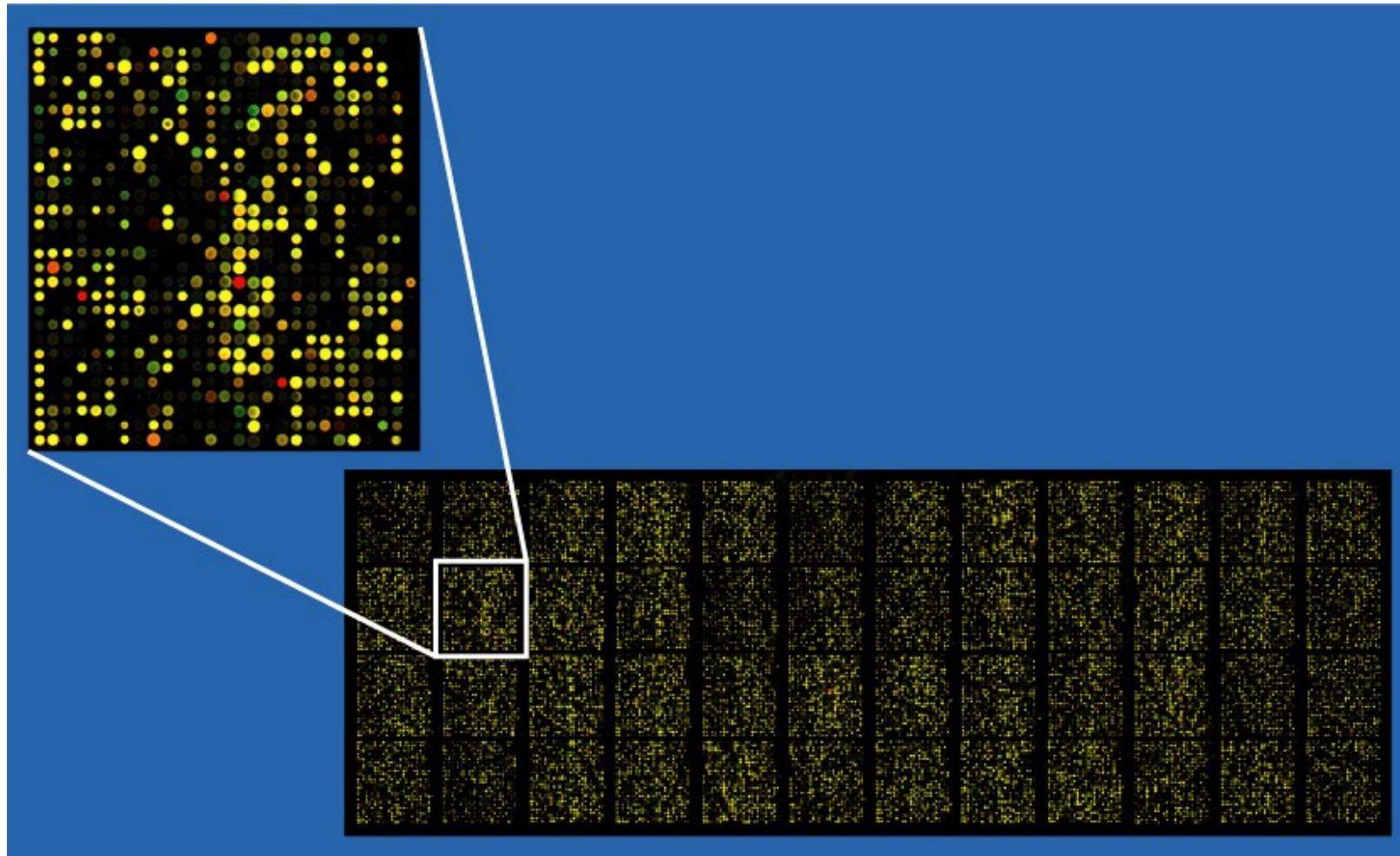


Fig. 1: online banking fraud losses between 2004 and 2014

<https://www.kdnuggets.com/2016/03/combat-financial-fraud-using-big-data.html>

Génomique : Puce à ADN

le cnam



source : Wikipedia

Facebook : analyse des interactions

le cnam



Liens entre profils : [http://www.facebook.com/notes/facebook-
engineering/visualizing-friendships/469716398919](http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919)

Méthodes de DM

Méthodes DM selon les disciplines

le chnam

Ces méthodes reviennent souvent à optimiser les mêmes critères, mais avec des approches et formulations différentes

Statistiques

Théorie de l'estimation,
tests Économétrie
Maximum de vraisemblance et
moindres carrés Régression
logistique, ...

Datamining

Intelligence artificielle

Apprentissage symbolique
Reconnaissance de formes
Réseaux de neurones, algorithmes
génétiques...

Analyse de données (Statistique exploratoire)

Description factorielle
Apprentissage automatique
Discrimination Clustering
Méthodes probabilistes ACP,
ACM, Analyse discriminante,
CAH, ...

Informatique (Base de données)

Exploration des bases de données
Volumétrie
Règles d'association, motifs
fréquents, ...

Méthodes selon les objectifs

Description :

Description/résumé des données pour leur compréhension:

- statistique descriptive (moyenne, médiane, coefficient de corrélation)
- analyse factorielle

Ex : moyenne d'âge des personnes présentant un cancer du sein

Structuration :

Faire ressurgir des groupes « naturels » qui représentent des entités particulières

classification (clustering,

apprentissage non-supervisé)

Ex : découvrir une typologie de comportement des clients d'un magasin

Méthodes de datamining

Explication :

Prédire les valeurs d'un attribut (endogène) à partir d'autres attributs (exogènes)

- régression
- apprentissage supervisé

Ex : prédire la qualité d'un client (rembourse ou non son crédit) en fonction de ses caractéristiques (revenus, statut marital, nombre d'enfants, etc.)

Association :

Trouver les ensembles de descripteurs qui sont le plus corrélés

- règles d'association

Ex : rayonnage de magasins, les personnes qui achètent du poivre achètent également du sel

Relations entre variables

- corrélation
- dépendance non linéaire
- capacité de prédiction

Relations entre individus

- interactions significatives
- groupes homogènes

Relations entre évènements

- co-occurrence
- dépendance logico-temporelle

Apprentissage automatique

le chnam

Il s'agit à partir des observations d'un phénomène de construire un modèle de ce phénomène afin de faire des analyses et des prévisions du phénomène grâce au modèle, et ce de manière automatique (presque) sans intervention humaine

Deux grandes catégories d'apprentissage :

Observations d'un phénomène \Rightarrow des données $z_i \in Z$

1. Apprentissage non supervisé :

- Pas de structure interne aux données
- Classification/clustering, règles d'association, etc.

2. Apprentissage supervisé :

- Modélisation du lien entre x et y
- Pour faire des prévisions : connaissant x , on prédit y
- Classification/régression

Méthodes d'apprentissage

le chnam

Apprentissage non supervisé :

- On ne connaît pas les données (l'analyste intervient peu)
- Statistiques classiques : moyenne, médiane, coefficient de corrélation
- Version visuelle (exploratoire) : histogrammes, diagramme à bâtons

Apprentissage supervisé :

- Intervention minimale de l'analyste pour le choix d'une méthode, indication sur les données par un expert, et analyse des résultats
- Exemples : reconnaissance d'empreintes digitales/ recherche de co-occurrences fréquentes dans un texte.

Apprentissage semi-automatique :

- L'analyste guide le processus
- Algorithmes d'apprentissage : inférence à partir d'exemples de résultats connus
- Exemple : segmentation d'un ensemble de clients/ construction d'un modèle en vue d'une exploitation automatique

Taxonomie des algorithmes

le cham

Algorithme	Mode d'apprentissage	Type de problème à traiter
Régression linéaire univariée	Supervisé	Régression
Régression linéaire multivariée	Supervisé	Régression
Régression polynomiale	Supervisé	Régression
Régression régularisée	Supervisé	Régression
Naïve Bayes	Supervisé	Classification
Régression logistique	Supervisé	Classification
Clustering hiérarchique	Non supervisé	-
Clustering non hiérarchique	Non supervisé	-
Arbres de décision	Supervisé	Régression ou classification
Random forest	Supervisé	Régression ou classification
Gradient boosting	Supervisé	Régression ou classification
Support Vector Machine	Supervisé	Régression ou classification
Analyse en composantes principales	Non supervisé	-

Vocabulaire

Collision du Français et Anglais :

Français	Anglais
Classification	<i>Clustering</i>
Classement	<i>Classification ou ranking</i>
Discrimination	<i>Classification</i>

Apprentissage supervisé

Objectif :

A partir des données $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = \dots, N\}$, estimer les prochaines valeurs.

On parle **d'apprentissage supervisé** car les y_i permettent de guider le processus d'estimation.

Exemples d'application :

Estimer les liens entre habitudes alimentaires et risque d'infarctus pour le diagnostic médical => x_i : d attributs concernant le régime d'un patient, y_i sa catégorie (risque, pas risque).

Techniques :

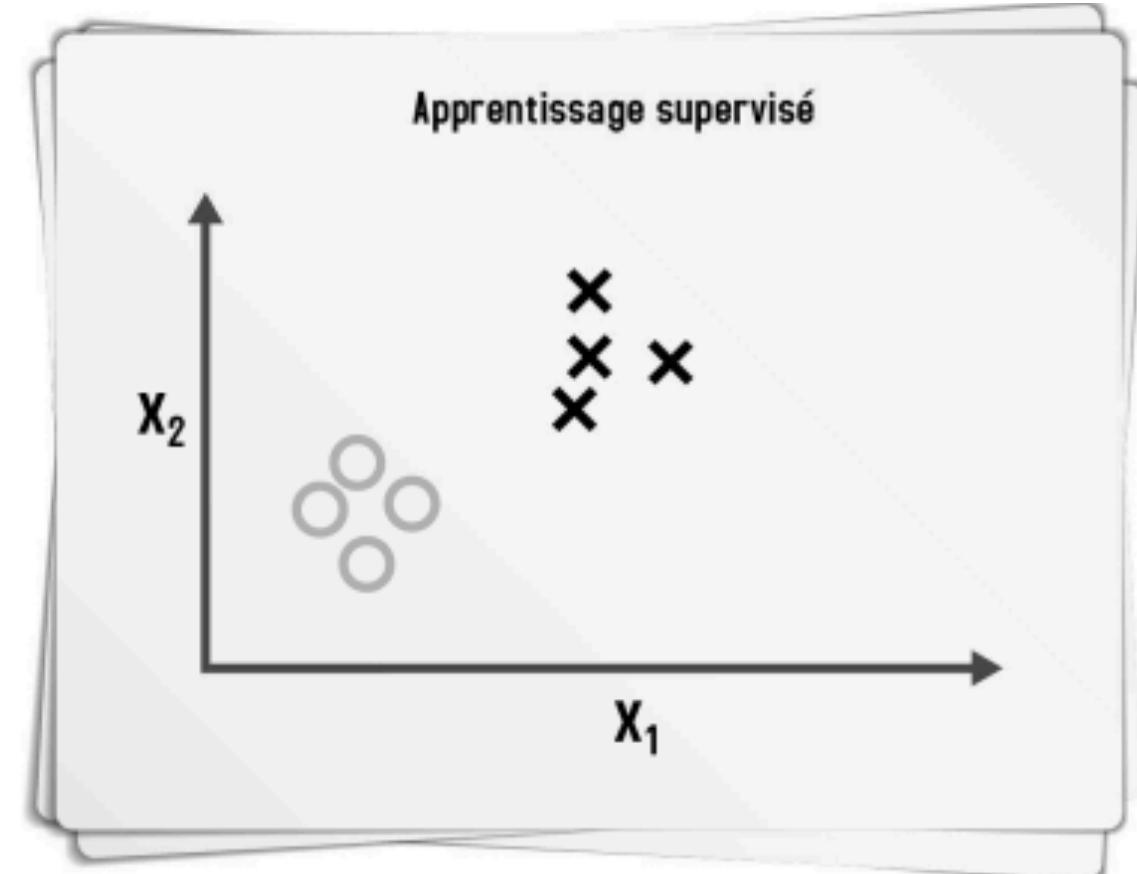
k-plus proches voisins, SVM, régression logistique, arbre de décision ...

Apprentissage supervisé

Les algorithmes supervisés extraient de la connaissance à partir d'un ensemble de données contenant **des couples entrée-sortie**. Ces couples sont déjà « **connus** », dans le sens où les sorties sont définies a priori.

La valeur de sortie peut être une indication fournie par un expert : par exemple, des valeurs de vérité de type OUI/NON ou MALADE/SAIN.

Ces algorithmes cherchent à définir une représentation compacte des associations entrée-sortie, par l'intermédiaire **d'une fonction de prédiction**.



De Eric Biernat, Michel Lutz, 2016

Régression/Classification

On distingue deux types de valeurs de sorties qu'on peut chercher à traiter.

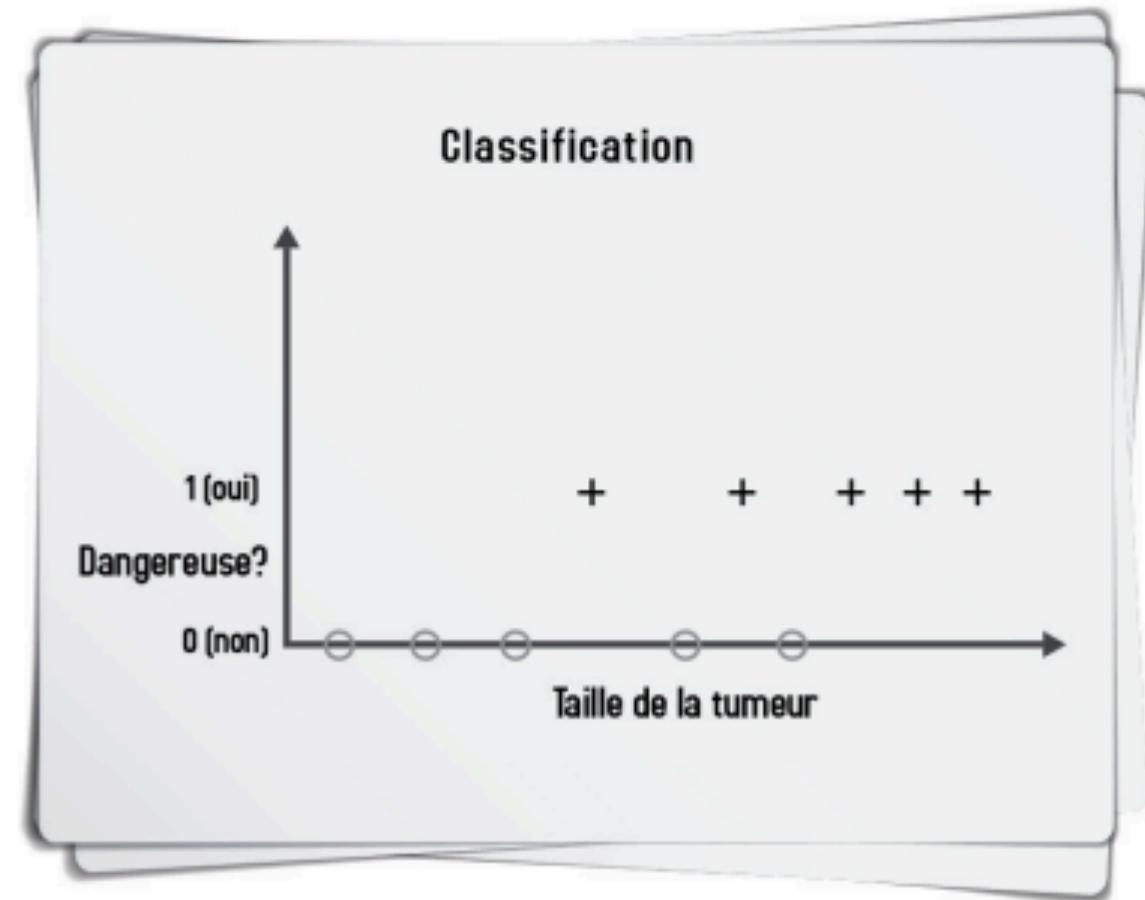
Dans le cadre d'un problème de régression, Y peut prendre une infinité de valeurs dans l'ensemble continu des réels (noté $Y \in \mathbb{R}$). Ce peut être des températures, des tailles, des PIB, des taux de chômage, ou tout autre type de mesure n'ayant pas de valeurs finies a priori.

Dans le cadre d'un problème de classification, Y prend un nombre fini k de valeurs ($Y = \{1, \dots, k\}$). On parle alors d'étiquettes attribuées aux valeurs d'entrée. C'est le cas des valeurs de vérité de type OUI/NON ou MALADE/SAIN évoqués précédemment.

Classification

Selon sa taille, une tumeur est-elle dangereuse ou bénigne ? Ici, on va chercher à classer les observations en fonction de valeurs de réponse possibles en nombre limité : OUI, NON (éventuellement PEUT-ÊTRE).

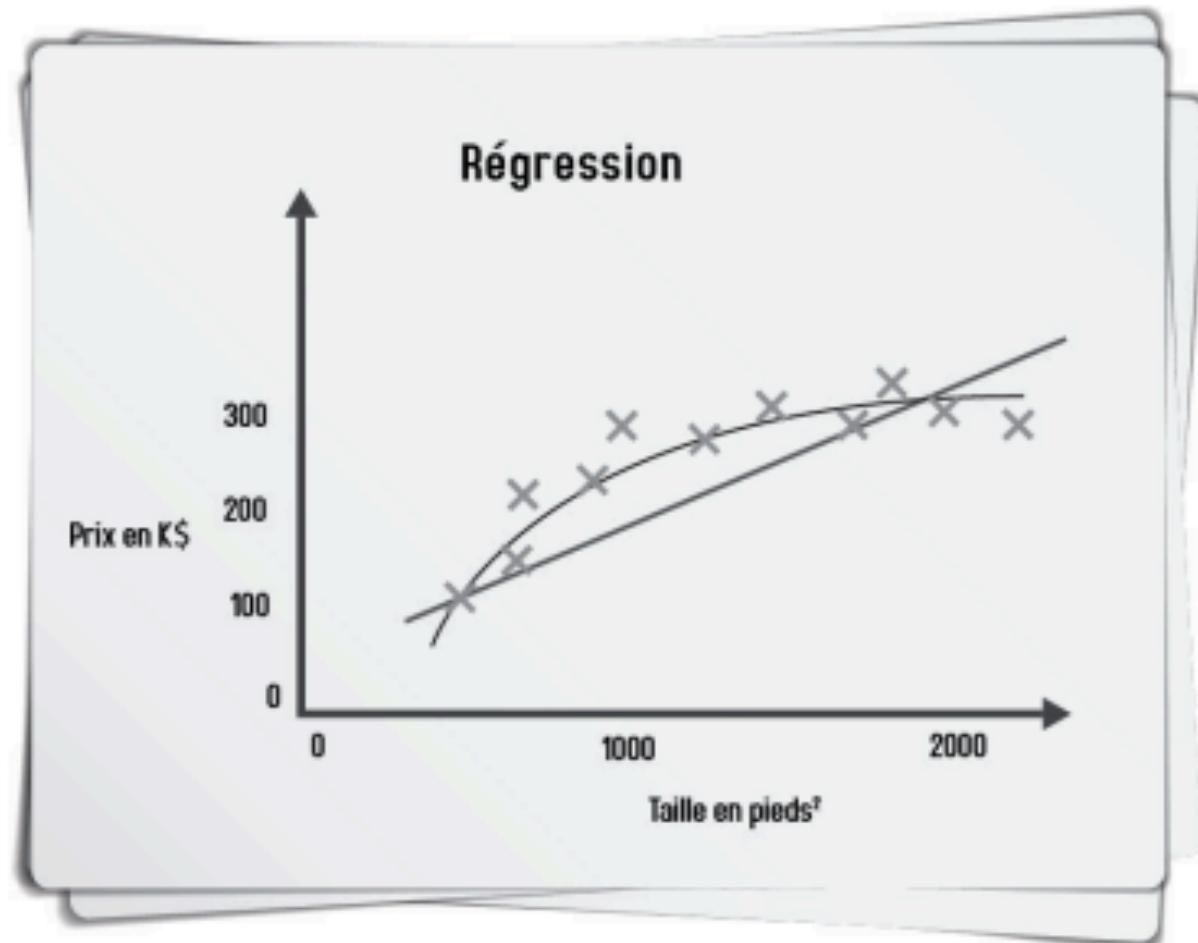
C'est un problème de classification.



Régression

Quel est le prix d'une maison en fonction de sa taille ? Ce prix peut prendre une infinité de valeurs dans \mathbb{R} .

C'est un problème de régression.



Apprentissage non supervisé

Objectif :

Seules les données $\{x_i \in X, i = 1, N\}$ sont disponibles.

On cherche à décrire comment les données sont organisées et en extraire des sous-ensemble homogènes (groupes, classes, etc.).

Exemples d'application :

Catégoriser les clients d'un supermarché. x_i représente un individu (adresse, âge, habitudes de courses ...) pour identification de segments de marchés, catégorisation de documents similaires, segmentation d'images biomédicales, etc.

Techniques :

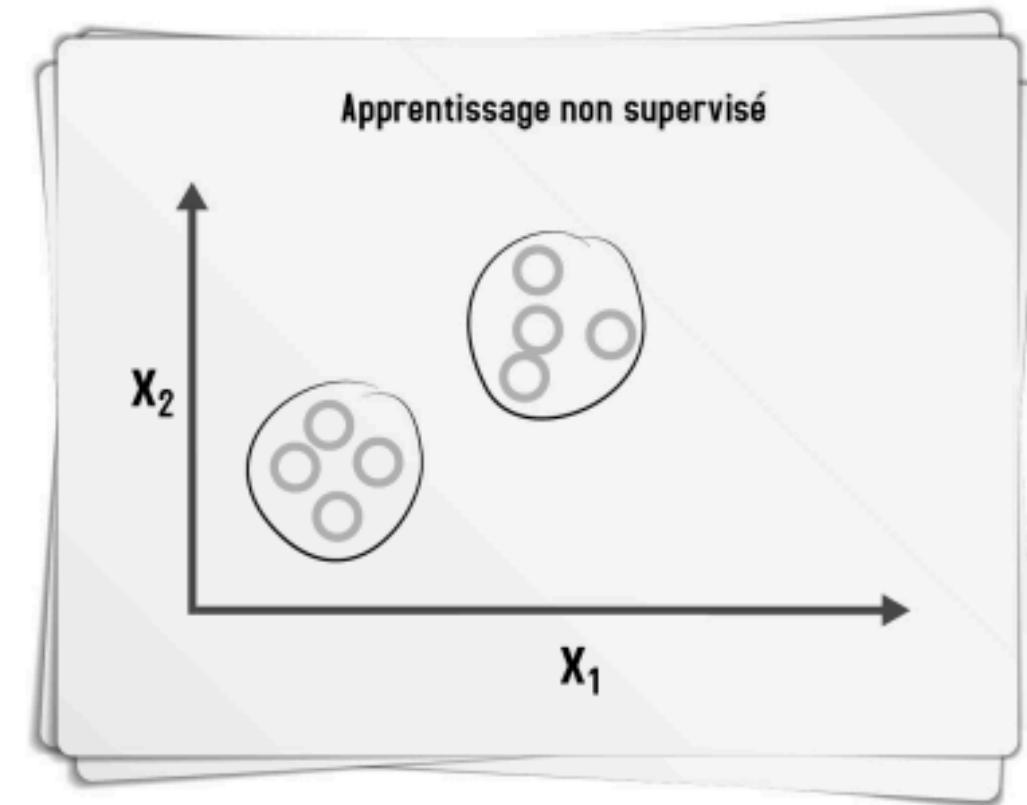
Classification hiérarchique, K-means, Extractions de règles

Apprentissage non supervisé

Les algorithmes non supervisés **n'intègrent pas la notion d'entrée-sortie**. Toutes les données sont équivalentes (on pourrait dire qu'il n'y a que des entrées).

Dans ce cas, les **algorithmes cherchent à organiser les données en groupes**. Chaque groupe doit comprendre des données similaires et les données différentes doivent se retrouver dans des groupes distincts.

Dans ce cas, l'apprentissage se fait à partir d'indications préalablement fournies par un expert.



Apprentissage non supervisé

le chnam

Positionnement :

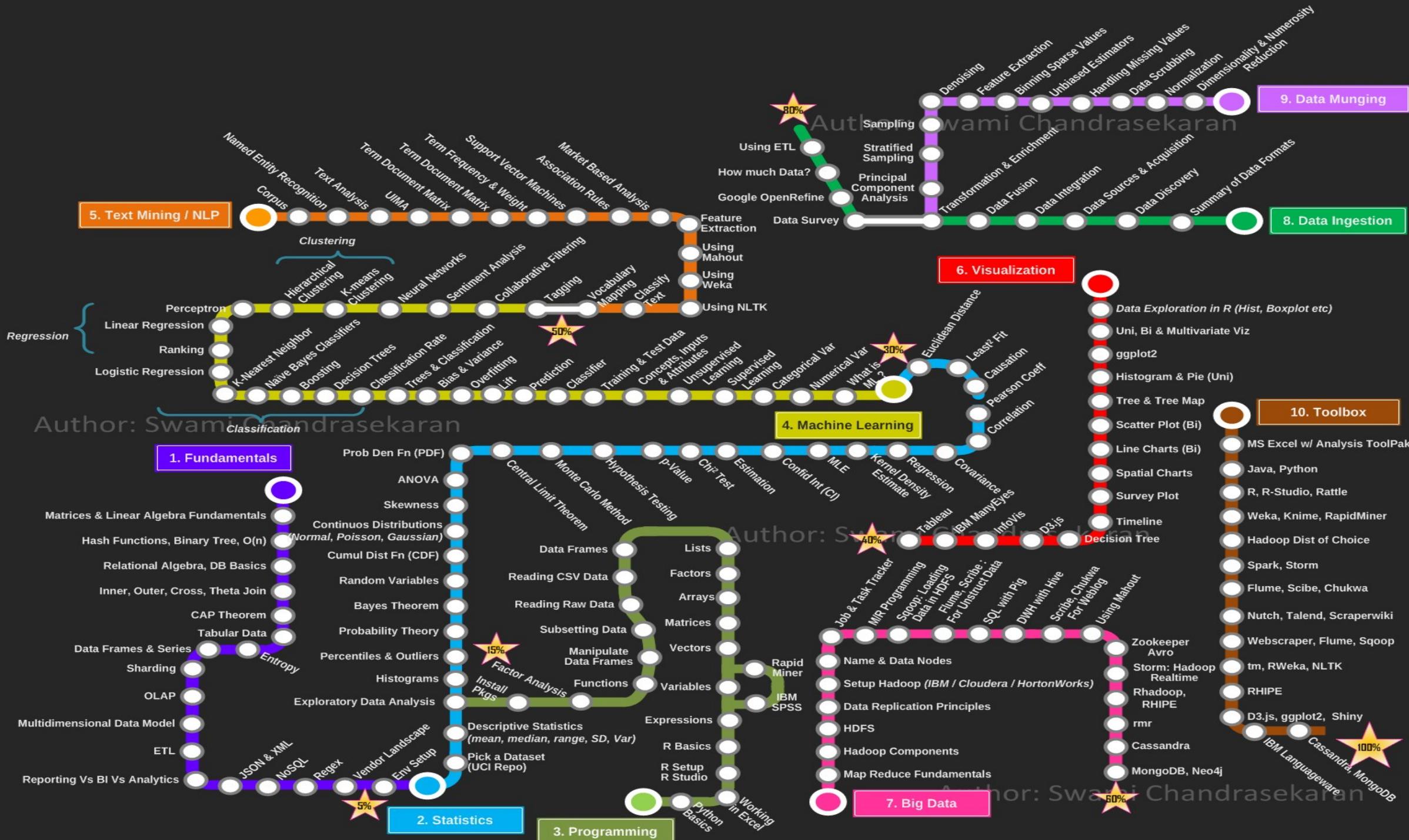
- Représenter des objets dans le plan (un point par objet)
- **Applications** : visualisation globale d'un jeu de données, analyse visuelle (groupes, corrélation, etc.)

Classification (*clustering*) :

- Trouver dans un ensemble d'objets des groupes homogènes (classes) et bien distincts les uns des autres
- S'appuie sur une mesure de **similarité** entre objets
- **Applications** : typologie de clients, regroupement de gènes, regroupement de pages web, etc.

Recherche de schémas fréquents (Patterns):

- Trouver des groupes d'objets fréquemment ensembles
- Trouver des séquences fréquentes d'actions
- **Applications** : recommandations, offres marketing, etc.



Logiciels de DM

le cnam

Commerciaux

- SPAD
- SAS Enterprise miner
- SPSS Clementine
- STATISTICA Data Miner
- IBM Intelligent Miner
- RAPIDMINER
- KNIME

Universitaires

- Logicel R
- TANAGRA
- SIPINA v2.5 & Recherche
- WEKA
- ORANGE

Langage

Python,
Scikit-Learn
Json, etc.

Spécialisés

- Requete : SQL : oracle sql
- Requete Hadoop : Hive, Pig
- Matlab, Octave
- Visualisation : D3JS

Ressources

Site avec outils, tutoriels, ouvrages :

- <https://www.kaggle.com/>
- <http://www.kdnuggets.com>
- <http://www.sas.com/big-data/>
- <http://tutoriels-data-mining.blogspot.com/>
- <https://www.r-bloggers.com/CRAAn>
- <http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html>

Bibliographie

- Le Data mining », R. Lefebure et G. Venturi, ed. Eyrolles, 2001.
- Data Mining et statistique décisionnelle, S. Tufféry, ed. technique, 2006.
- Data Mining : Practical machine learning tools and techniques with Java implementations », I. Witten and E. Frank, Morgan Kaufman Pub., 2000.
- The elements of statistical learning - Data Mining, Inference and Prediction, T. Hastie, R. Tibshirani, J. Friedman, Springer 2001.
- Machine Learning, T. Mitchell, Mc Graw-Hill Editions, 1997.
- Data science : fondamentaux et études de cas: Machine learning avec Python et R - De Eric Biernat, Michel Lutz, 2016

Merci.