FOUILLE DE DONNÉES SUPERVISÉE

Master 2 AFD

Nathalie Camelin

sept 2021

Organisation

- Cours communs avec ATAL
  - 6 * CM de 1h20
  - Fouille de Données … focus sur la donnée « texte »
- En pratique
  - 2 CM N. Camelin / 2 CM N. Dugué / 2 CM N. Camelin
  - 4 TPs « Projet guidé »
  - 2 TDs 1h projets + 1 TD 2h restistution
- Supports dispo sur UMTICE
  - M2 INFO AFD - Fouille de données supervisée
  - http://umtice.univ-lemans.fr/course/view.php?id=759
  - Clé : M2FDD

Évaluations

- 1 note de TP (0,5 ECTS)
  - Mise en œuvre d'une tâche de fouille de données
    - Livrables :
      - Modèles prédictifs + prédictions sur corpus de Test
      - Rapport analyse de résultats + Oral
    - Travail en binôme (1 groupe à 3)
    - !! Travail obligatoire en dehors des TPs !!
  - Report en session 2
- 1 examen écrit de 1h30 (1,5 ECTS)

Qu'est-ce que la fouille de données?

- La fouille de données c'est :
  - algorithmes et méthodes
  - pour l'exploration et l'analyse
  - de (souvent très) grandes masses de données
    - numériques
    - hétérogènes
  - afin d'y détecter des
    - règles, associations, tendances, structures

    →restituer de façon concise l'information
  - pour l'aide à la décision

Qu'est-ce que la fouille de données?

- En anglais « Data Mining »
  - Forage de données (« gold mining » ou « diamond mining »)

  → « knowlegde mining from data »

- Domaines connexes
  - Reconnaissance de formes (Pattern analysis)
  - Informatique décisionnelle (Business intelligence)

Qu'est-ce que la fouille de données?

- Application de l'ensemble des techniques pour l'exploration et l'analyse de grandes masses de données afin d'en extraire de la connaissance

- Connaissance ?
  - Information nouvelle (Précédemment inconnue)
  - Non triviale
  - Potentiellement utile

    →*Information intéressante*

- But : Prise de décision

Data mining or not?

- Ce qui n'est pas de la fouille de données
  - Rechercher des numéros de téléphone dans un annuaire
  - Rechercher les pages web contenant le mot « Amazon »
- Ce qui est de la fouille de données
  - Rechercher les noms les plus fréquents de l'annuaire en fonction de zones géographiques
  - Regrouper les pages traitant du site « amazon.com » d'une part et de l'autre celles traitant du fleuve d'amérique du sud

Spectre très large d'applications

- De l'∞ petit à l'∞ grand
- du plus quotidien au moins quotidien
- du plus ouvert au plus sécuritaire
- du plus industriel au plus théorique
- du plus alimentaire au plus divertissant

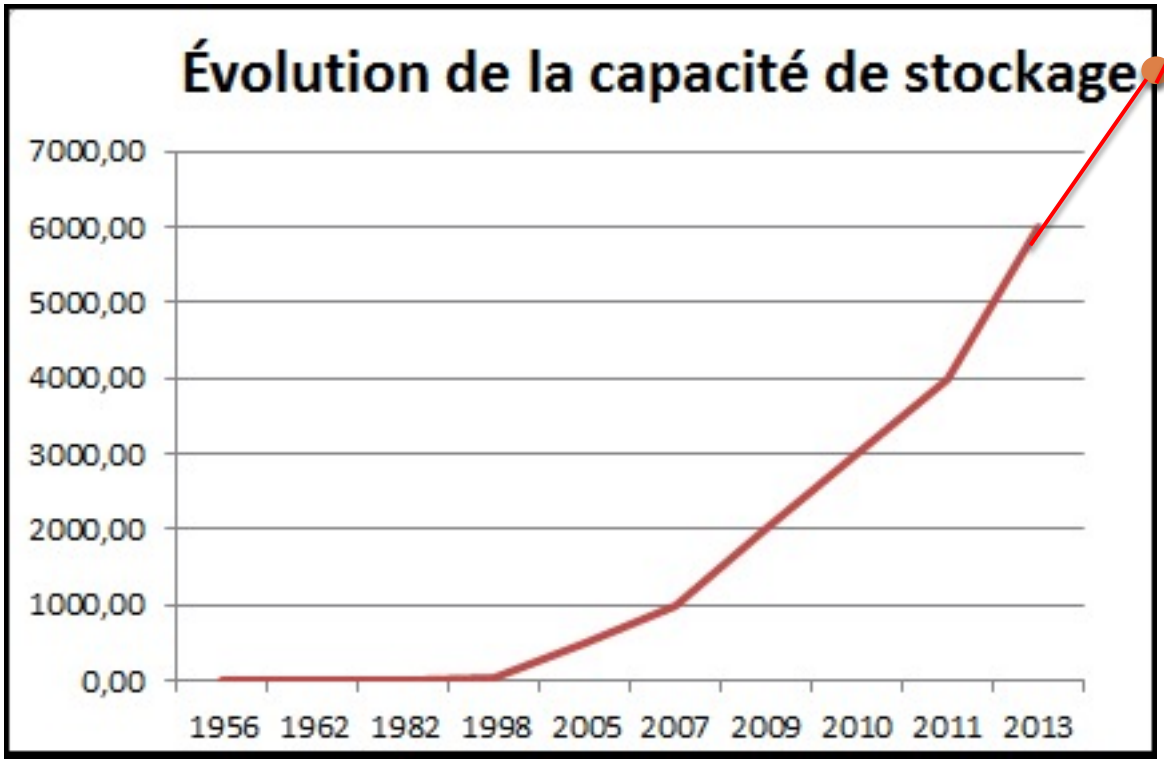
Etat des lieux

« We are data rich, but information poor »

- Explosion des masses de données
  - De très grandes masses de données : bases de données, entrepôts de données, internet, …
- Abondance de données
  - Entrepôts de données … cimetière de données!
  - Certaines données ne sont/seront jamais vues par un humain

Explosion des capacités

- Capacités de stockage et de calcul



Global Information Storage Capacity

in optimally compressed bytes

2007 ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % ANALOG

DIGITAL
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
  123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

DIGITAL
280 exabytes

2000

1993

1986
ANALOG
2.6 exabytes

DIGITAL
0.02 exabytes

ANALOG STORAGE

DIGITAL STORAGE

2002:
"beginning of the digital age"
50%

% digital:
1 %
3 %
25 %
94 %

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Données massivement disponibles ?

- How much information ? 2003
  - http://www2.sims.berkeley.edu/research/projects/how-much-info/
    - The world produces between 1 and 2 exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth.
  - Entre 3 et 5 exa-octets de données originales produites en 2003 (texte, audio, image, vidéo, ...)
  - 1 exa-octets représente 10^{18} octets soit un milliard de gigaOctets

1 000 000 000 000 000 000

Eo Po To Go Mo ko

Données massivement disponibles

- How much information ? 2007
  - Article : http://phys.org/news/2011-02-world-scientists-total-technological-capacity.html

« Looking at both digital memory and analog devices, the researchers calculate that humankind is able to store at least 295 exabytes of information. (Yes, that's a number with 20 zeroes in it.) »

- 2002 : Année charnière : capacité de stockage numérique > capacité de stockage analogique
- 2007 : 94% de notre mémoire digitalisée
- De 1986 à 2007, la capacité de calcul a augmenté de 58% par an et la capacité de stockage de 23% par an …

Données massivement disponibles !

- How much information 2008?
  - http://www.huffingtonpost.com/2011/04/06/world-information-consumption_n_845806.html

    « *an estimate of digital information consumed in 2011 is likely to be far greater than 9.57 zettabytes : 9 570 000 000 000 000 000 000 per year*» !!!

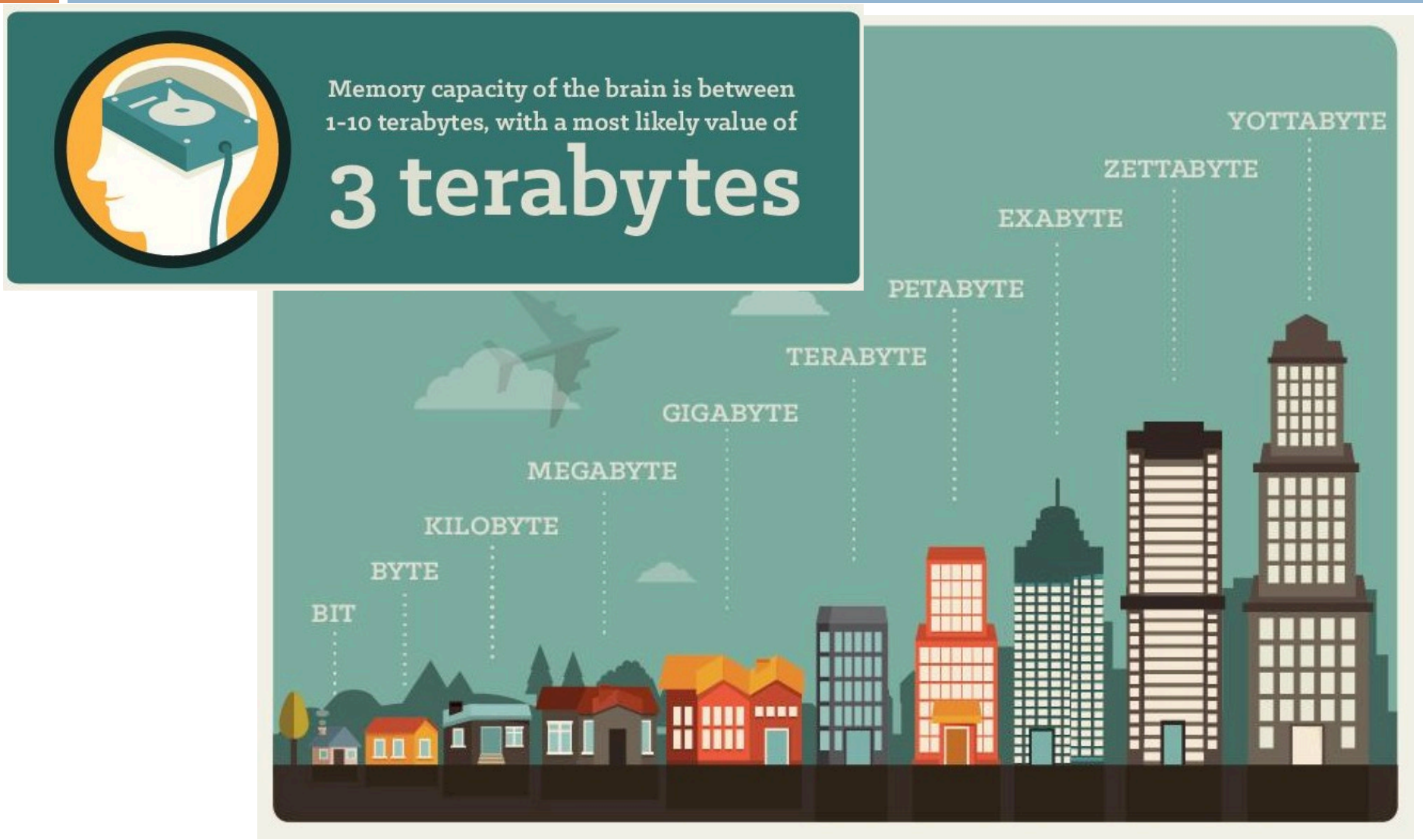
    « Most of this information is incredibly transient: it is created, used, and discarded in a few seconds without ever being seen by a person »…

    « It's the underwater base of the iceberg that runs the world that we see »

Données massivement disponibles !!

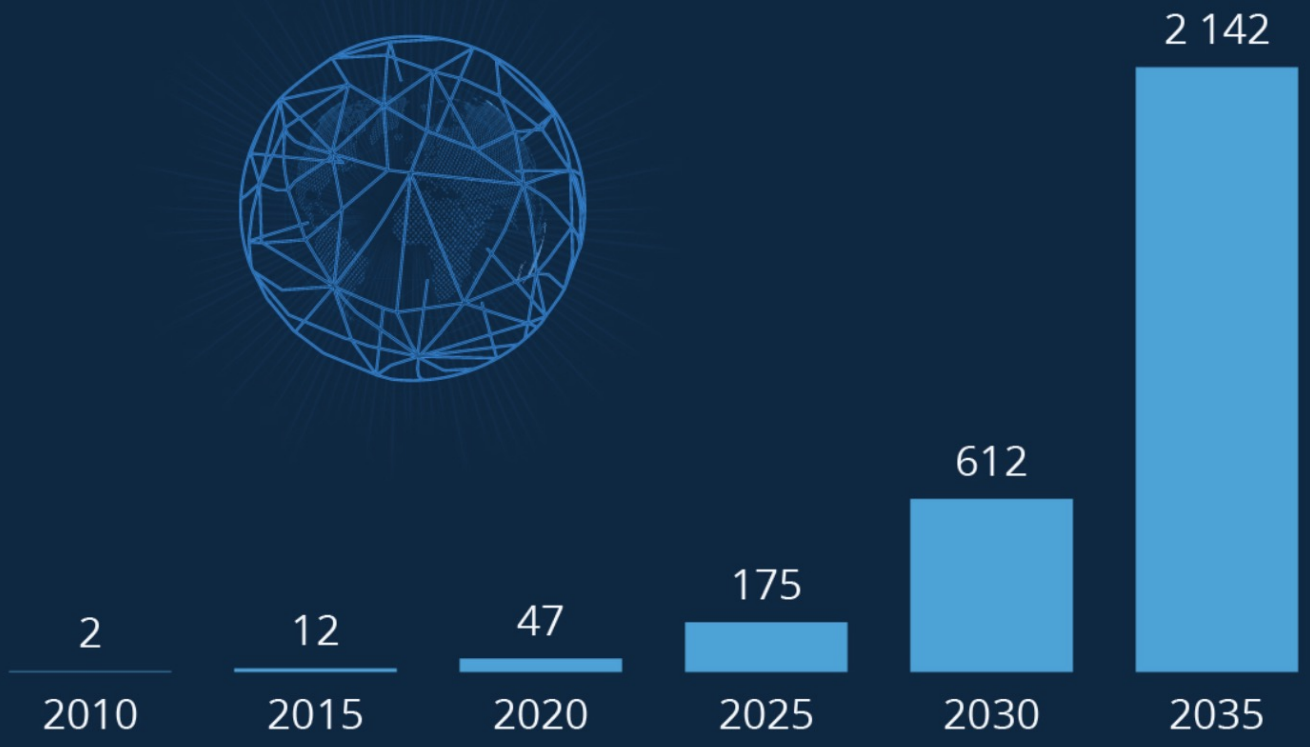
- Quantité de données stockées par an :
  - 2010 : 1,2 Zo
  - 2011 : 1,8 Zo
  - 2012 : 2,8 Zo
- Prédiction de 2012
  - pour 2020 : 40 Zo
  - → finalement a priori environ 60 Zo …

Toujours plus…



Le big bang du big data

Volume annuel de données numériques créées à l'échelle mondiale depuis 2010, en zettaoctets *



* Prévisions de 2020 à 2035. Un zettaoctet équivaut à mille milliards de gigaoctets.
Source : Statista Digital Economy Compass 2019

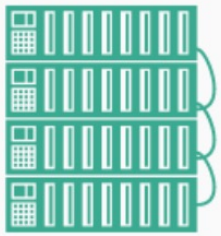
Toutes les données créées en 2018 équivalent à...

33 zettaoctets Quantité de données numériques créées dans le monde en 2018

660 milliards
de disques Blu-ray
(50 gigaoctets)

330 millions
des plus gros
disques durs actuels
(100 téraoctets *)

33 millions
de cerveaux humains
(1 pétaoctet **)



132.000
espaces de stockage
du supercalculateur
le plus rapide
(250 pétaoctets ***)

73 grammes
d'ADN
(455 exaoctets)

Échelle

x 1.000 — Mégaoctet
x 1.000 — Gigaoctet
x 1.000 — Téraoctet
x 1.000 — Pétaoctet
x 1.000 — Exaoctet
Zettaoctet

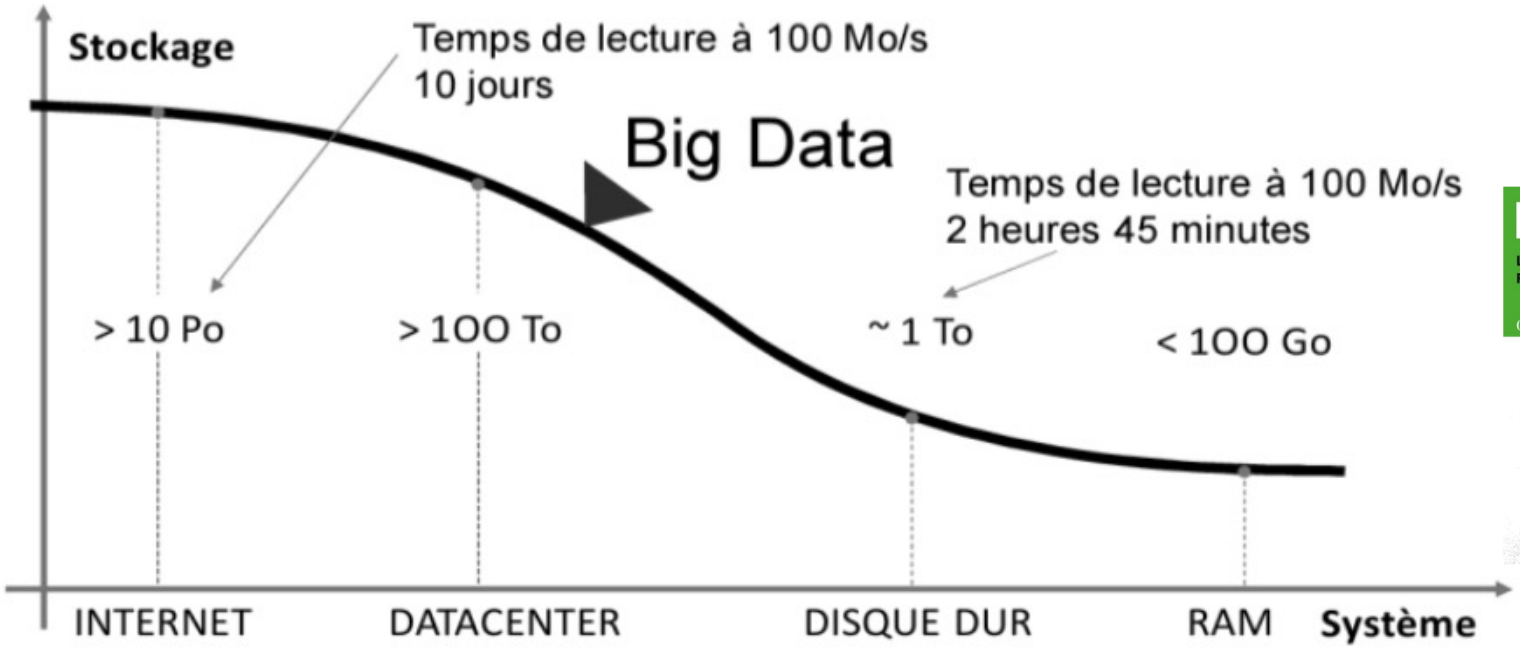
* en date de mars 2019.
** cette donnée varie selon la méthode de calcul.
*** en date de juin 2018.

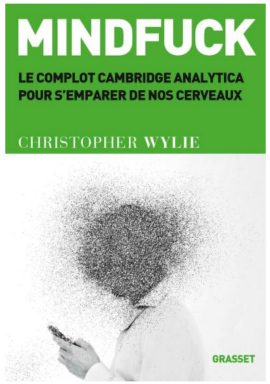
Source : Statista Digital Economy Compass 2019

statista

Big Data ?

- Big Data et Data Mining : objectif commun
  - Extraire de l'information des masses de données
- Quand commence le Big Data ?





Data Mining en vogue …

- Des méthodes de plus en plus performantes capables de traiter toutes sortes de données
  - Lacunaires
  - Aberrantes
  - Hétérogènes
    - Numériques
      - Discrètes
      - Continues
    - Textuelles, Audio, Vidéo, …
      - Text mining, Speech Mining…

Démocratisation du Data Mining

- De plus en plus de solutions logicielles intégrées
  - Traitement des données
  - Algorithmes de statistiques et de data mining
  - Analyse de résultats
- Exemples de logiciels
  - Commerciaux
    - SPAD (Decisia), SAS Entreprise miner, STATISTICA Data Miner, IBM Intelligent Miner
  - Universitaires
    - Tanagra, Orange, ***Weka***
- Comparaison de logiciels gratuits : http://chirouble.univ-lyon2.fr/~ricco/data-mining/logiciels/revue_rapide_des_logiciels_sur_le_site_kdnuggets.pdf

Monde industriel

- Grand intérêt de la part de l'industrie et des sociétés
  - Constitution de gigantesques bases de données pour les besoin de gestion des entreprises

  → Inexploitées!

- De l'informatique décisionnelle au data mining

→ Besoins d'outils performants pour extraire de l'information de valeurs pour la prise de décision

Pourquoi fouiller des données?

- Point de vue commercial
  - → Collecte de données systématique
    - Transactions bancaires
    - Profils des e-acheteurs
    - Liste des achats
    - …
  - Améliorer la compétitivité
    - Analyse du ticket de caisse
    - Etude d'appétence dans les sociétés commerciales
    - Prédiction de l'attrition dans la téléphonie mobile

Pourquoi fouiller des données?

- Point de vue scientifique
  - De très grandes masses de données enregistrées par heure
  - D'où proviennent ces données?
    - Capteurs des satellites
    - Téléscopes scannant le ciel
    - Puces à ADN
    - …
  - Techniques d'analyse statistique classiques insuffisantes
  - Aide à classer et segmenter les données
  - Aide à la formation d'hypothèses scientifiques

Pourquoi fouiller des données?

- Traitement Automatique de la Langue (TAL)
  - Avec le net, de plus en plus de documents contenant du langage naturel
  - Documents Texte, Audio, Vidéos
- Beaucoup d'applications
  - Recherche d'information/Indexation
  - Reconnaissance du locuteur
  - Détection d'opinions
  - Systèmes de question/réponse
  - Résumé automatique
  - …

Quel type de tâches?

- Deux grands types de tâches
  - Fouille de données prédictive
    - Extrapoler de nouvelles informations
    - Prédire
  - Fouille de données descriptive
    - Mettre en évidence des informations « cachées »
    - Comprendre

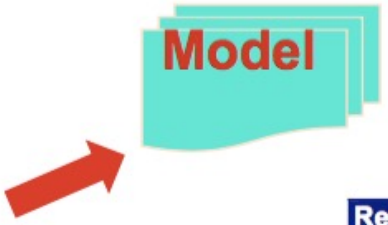
Fouille de données prédictive

- Fouille de données prédictive
  - Inférer des propriétés (concepts) sur de nouvelles données à partir des données actuelles
  - Variable « cible » à prédire
  - 2 grandes familles
    - Classification (en) – Classement (fr)
      - Variable cible de type fini et discret
    - Régression
      - Variable cible de type continu

Classification (en) – Classement (fr)

- Processus qui consiste à construire un modèle qui permet :
  - d'associer une donnée à un concept connu
  - de prédire le concept d'une nouvelle donnée

Id	Rembour sement	Statut marital	Revenu imposable	Fraudeur
1	Oui	Célib.	125K	Non
2	Non	Marié	100K	Non
3	Non	Célib.	70K	Non
4	Oui	Marié	120K	Non
5	Non	Divorcé	95K	Oui
6	Non	Marié	60K	Non
7	Oui	Divorcé	220K	Non
8	Non	Célib.	85K	Oui



Rembour sement	Statut marital	Revenu imposable	Fraudeur
No	Célib.	75K	?
Yes	Marié	50K	?
No	Marié	150K	?
Yes	Divorcé	90K	?

Exemples d'applications prédictives

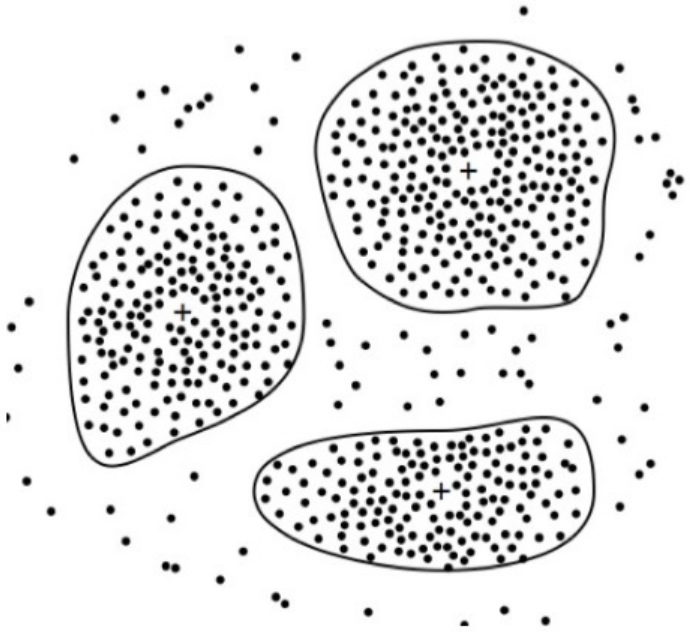
- Marketing
  - Réduire le coût du démarchage téléphonique en ciblant les potentiels clients d'un produit donné
- Finance
  - Détecter les transactions frauduleuses de cartes de crédit
- Société : relation clients
  - Analyser la satisfaction des clients

Fouille de données descriptive

- Fouille de données descriptive
  - Caractériser de manière intelligible les données (humainement interprétable)
  - Tâches
    - Clustering
    - Définition de règles d'association
    - Recherche de motifs séquentiels

Clustering (en) – Classification (fr)

- Processus qui consiste à regrouper des données
  - Les plus homogènes possibles au sein d'un cluster
  - Clusters les plus distants les uns des autres



Définition de règles d'association

- Découverte de règles en fonction d'un ensemble de données définies par plusieurs descripteurs

ID	*Descripteurs*
1	Pain, Soda, Lait
2	Bières, Pain
3	Bières, Soda, Couches, Lait
4	Bières, Pain, Couches, Lait
5	Soda, Couches, Lait

Règles d'association

{Lait} --> {Soda}

{Couches, Lait} --> {Bières}

- Action → positionnement des articles dans le supermarché

Exemples d'application descriptives

- Marketing
  - Créer automatiquement et sans a priori des profils de clients potentiellement intéressés par les mêmes produits
- Recherche d'information
  - Catégorisation thématique de documents en fonction du contenu textuel

Extraction de connaissances

- Un sous-ensemble de

  Knowledge Discovery in Databases (KDD)

  - En français : Extraction de Connaissances à partir de Données (ECD)

Une étape primordiale du KDD

- Étapes du processus de KDD
  1. Définition du problème et ses objectifs
  2. Inventaire/Intégration des données
  3. Sélection/Préparation des données
  4. Fouille de données
  5. Evaluation des performances
  6. Représentation des connaissances pour prise de décisions
  7. Déploiement, enrichissement des modèles
- Souvent : confusion entre Data mining et le KDD

Exemple de charges en temps

Etape	Charge (en jours)	
	Projet Léger	Projet moyen
Définition de la cible et des objectifs	4j	8j
Inventaire des données	7j	10j
Collecte et préparation des données	15j	28j
Elaboration et validation des modèles	15j	25j
Analyse complémentaire, restitution des résultats	9j	12j
Documentation - Présentation	5j	7j
Analyse des premiers test	5j	10j
Total	60j	100j

Qu'est-ce qu'un résultat intéressant?

- Une règle/un modèle est intéressant(e) si :
  - Intelligibilité/interprétable par un humain
  - Véracité sur des nouvelles données
  - Valide une hypothèse à confirmer
  - Inattendu, pas commun
- Est-ce que le data mining ne génère que des résultats intéressants?
- Est-ce que le data mining génère tous les résultats intéressants?

Questions à se poser

- **Avant de mettre en place des techniques de data mining !**

- Est-ce que le problème peut être clairement défini?
- Est-ce que des données potentiellement pertinentes existent?
- Est-ce qu'il existe des connaissances latentes à travers ces données?
- Est-ce que le coût de la mise en oeuvre de la fouille des données peut être "amorti" par le profit de la connaissance acquise?

Facteurs de succès d'un projet

- Objectifs précis, stratégiques et réalistes
- Qualité et richesse des informations collectées
- Maitrise des techniques de data mining utilisées
- Bonne restitution des résultats

En résumé…

- Tâche d'exploration de masses de données afin d'en extraire des modèles/règles *intéressantes*
- Discipline jeune (environ 30-40)
- Aux confluences de plusieurs domaines de recherche scientifique
  - Statistiques, Apprentissage Automatique, Recherche d'Informations, Big Data, …
- Du travail en perspective

Les Data miners

- De nombreuses compétences :
  - Maitrise des outils d'exploitation performants
  - Expertise mathématique pour analyse des résultats
  - Bonne connaissance métier
- Besoin croissant de data miners…

http://archives.lesechos.fr/archives/2012/lesechos.fr/07/15/02021733689 14.htm

http://news.efinancialcareers.com/fr-fr/139910/data-miner-un-job-davenir-en-it-finance-mais-qui-reste-ultra-selectif/

« Data scientist : The Sexiest Job of the 21st Century », T.H. Davenport et D.J Patil, Harward Buziness Review, 2012

Bibliographie

- « Le Data mining », R. Lefebure et G. Venturi, Ed. Eyrolles, 2001.
  - Peu technique, point de vue général, bon recul
- « Data Mining et Statistique décisionnelle », S. Tufféry, Ed. Technip, 2009.
  - Technique, complet
- « Data Mining : Practical machine learning tools and techniques with Java implementations », 3e Ed., I. Witten and E. Frank, Morgan Kaufman Pub., 2011.
  - Très général et complet, logiciel libre accès, technique
- « Big Data et Machine Learning », P.Lemberger, M. Batty, M. Morel, JL. Raffaëlli, Ed. Dunod, 2015
  - Orienté technique du Big Data

Ressources en ligne

- http://chirouble.univ-lyon2.fr/~ricco/data-mining
  - Un portail pour la documentation : liens, supports de cours en ligne, logiciels, donnees
- http://www.kdnuggets.com
  - « Le » portail du DATA MINING, avec toute l'actualité du domaine
- http://data.mining.free.fr
  - Le site de Stéphane Tufféry
- Wikipédia
- Article intéressant sur l'évolution des puissances de calcul : https://www.huffingtonpost.fr/2016/03/28/loi-de-moore-fin-smartphones-ordinateurs-puissance-bonne-nouvelle_n_9547240.html

PARENTHESE (RAPPEL : APPRENTISSAGE AUTOMATIQUE

Trois grands temps

- À partir d'un ensemble de données
  - Description *pertinente* des données
- Mise en œuvre *efficace* d'un algorithme
  - Fonction du type de problème à résoudre
- Evaluation et/ou analyse des résultats obtenus
  - Fonction de l'application visée

Les corpus

- Ensemble des données disponibles
- Corpus d'apprentissage (APP)
  - Entraînement du modèle
- Corpus de développement (DEV) (facultatif)
  - Optimisation des paramètres d'ajustement du modèle (si nécessaire)
- Corpus de test (TEST)
  - Évaluation des performances du modèle en généralisation

!! La taille est critique …

Validation classique des résultats

- Cas classique : Assez de données annotées
  - Ex : 1 APP (70%) et 1 TEST (30%)
  - Estimation de l'erreur de **prédiction**
    - Évaluation du modèle sur l'APP
    - Taux de mauvaise classification sur l'APP
  - Estimation de l'erreur de **généralisation**
    - Evaluation du modèle sur le TEST
    - Taux de mauvaise classification sur le TEST
- Mise en production
  - Ré-apprentissage du modèle sur TOUT le corpus annoté

Mesures classiques globales

- Taux de bonne classification

$$\underset{\text{(Accuracy)}}{\text{Acc}} = \frac{\#\text{ instances bien classées}}{\#\text{ instances classées}}$$

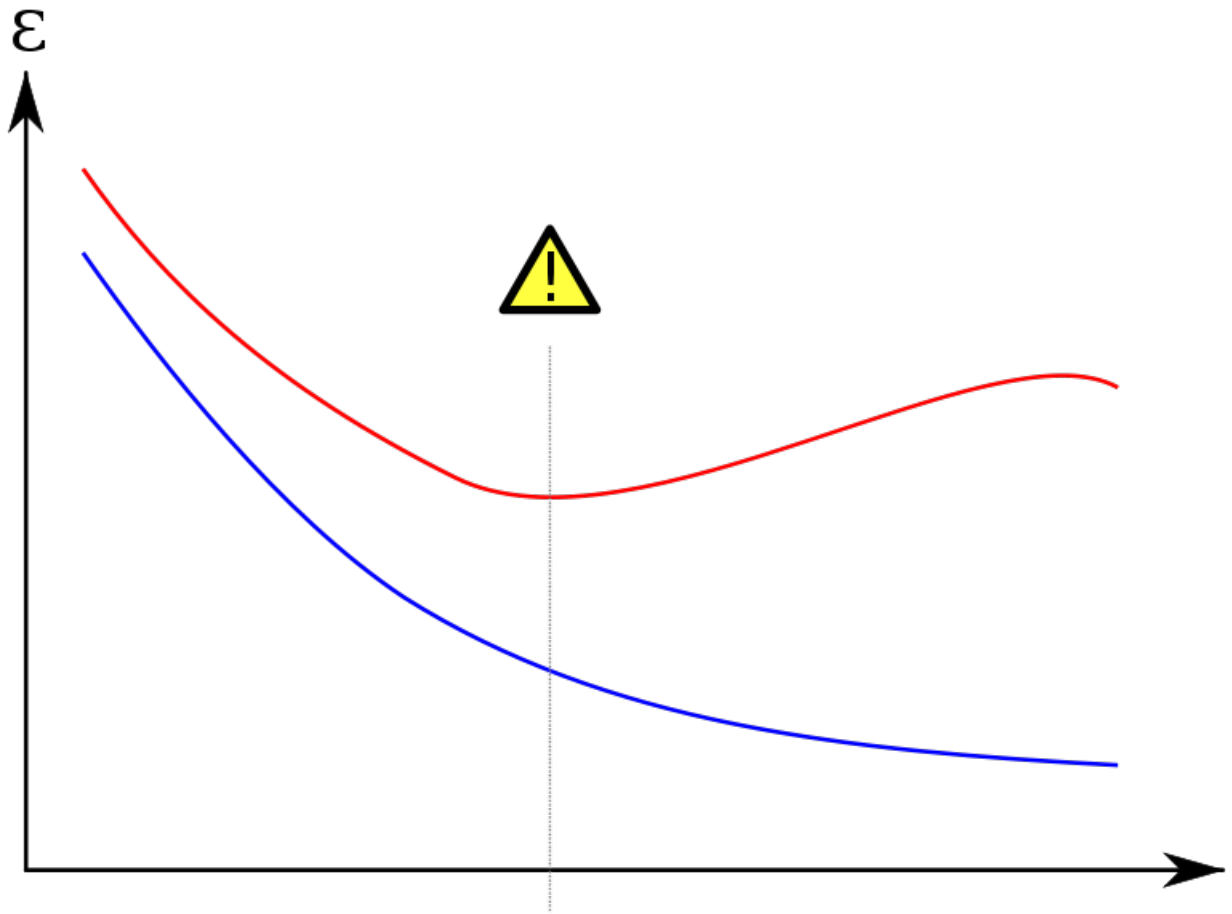
- Taux d'erreur (mauvaise classification)

$$\underset{\text{(Classication Error Rate)}}{\text{CER}} = \frac{\#\text{ instances mal classées}}{\#\text{ instances classées}}$$

Problème de sur-apprentissage

- Deux critères à considérer
  - Erreur de prédiction
  - Erreur de généralisation
- Quand « arrêter » d'apprendre?
  - Erreur de prédiction diminue ET l'erreur de généralisation augmente
- Comment faire?
  - Taille du corpus d'apprentissage
  - Paramètres d'ajustement du modèle

Problème de sur-apprentissage



Mesures en Recherche d'Information

- Précision : pourcentage de documents pertinents

$$\text{précision}_i = \frac{\#\text{ instances correctement classées i}}{\#\text{ instances classées i}}$$

$$\text{précision} = \frac{\sum_i \text{précision}_i}{\text{nombre de classes}}$$

- Précision élevée, moins de bruit

Mesures en Recherche d'Information

- Rappel : pourcentage de documents pertinents retrouvés

$$\text{rappel}_i = \frac{\#\text{ instances correctement classées i}}{\#\text{ instances réellement i}}$$

$$\text{rappel} = \frac{\sum_i \text{rappel}_i}{\text{nombre de classes}}$$

- Rappel élevé, moins de silence

Mesures en Recherche d'Information

- F-mesure : combinaison de la précision et du rappel

$$fmesure = \frac{(1+\beta^2)rappel * précision}{\beta^2(rappel + précision)}$$

généralement $\beta=1$

Confiance dans l'estimation de l'erreur?

- Erreur = variable aléatoire
  - Après classification, 2 valeurs possibles pour la donnée : bien ou mal classée
    - Erreur = probabilité de l'événement « mal classé »
  - En déterminer la moyenne? Un intervalle?

- Calcul de l'Erreur sur *1* corpus de test par le CER
  - Sur 100 exemples de test, 15 sont faux
    - Le taux d'erreur du système est de 15%?

Intervalle de confiance

- Estimation du *taux d'erreur réel* du système à partir du taux d'erreur observé sur un ensemble de test T
  - Approximation de la loi binomiale par la loi normale
    Intervalle de confiance à 95%
  - On estime l'erreur par l'intervalle de confiance :

$$CER \pm 1.96\sqrt{\frac{CER(1-CER)}{N}}$$

!! Nombre d'exemples du jeu de test suffisant

Classification multiclasses

- Macro-mesures
  - Même poids pour toutes les classes
    - \+ Ne pas masquer les classes rares
    - \- Classes rares et très présentes ont même importance

$$\text{mes}_{macro} = \frac{\sum_i \text{mes}_i}{\#\text{ classes}}$$

$$\text{mes} \in \{prec, rapp\}$$

- Micro-mesures
  - Même poids pour tous les documents
  - Classe très présentes masque les résultats sur la classe rare

REF \ HYP	+'	-'
+	TP = ΣTPi	FN = ΣFNi
-	FP = ΣFPi	

$$\text{prec}_{micro} = \frac{TP}{TP + FP}$$

$$\text{rapp}_{micro} = \frac{TP}{TP + FN}$$

FIN PARENTHESE