

# Introduction à l'extraction de connaissances et à la fouille de données

HMIN326M

Pascal Poncelet  
LIRMM

Pascal.Poncelet@lirmm.fr  
<http://www.lirmm.fr/~poncelet>



## Un exemple d'utilisateurs

(une parenthèse avant de rentrer dans le vif du sujet .....



2

## Quelques chiffres

- 3,6 milliards dans le monde (2020)
- 41,2 millions en France
- 32 millions sont inscrits sur au moins un réseau social
- 52 % ont entre 25 et 45 ans
- Facebook : 26 000 000 utilisateurs Français sur 1,6 milliards de membres !!
- En moyenne les utilisateurs ont 177 amis : .... 338 (2019)



(Sources Factory.com (2013)) .... Mis un peu à jour en 2020



3

---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---



---

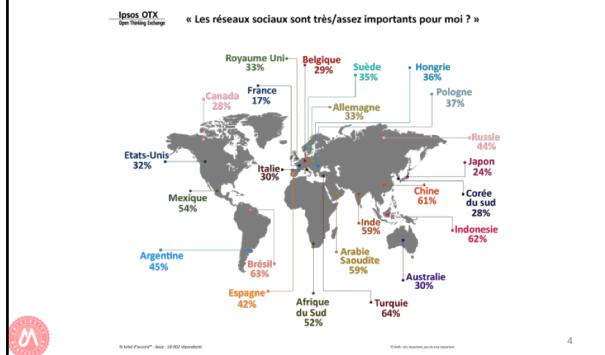


---

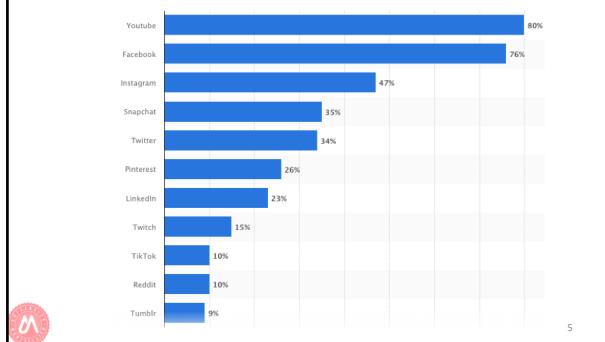


---

De l'utilisation des sites de réseaux sociaux



Les types de réseaux (France - 2019)



## Beaucoup de temps

- 66 % des utilisateurs Français se connectent une fois par jour
  - 22,5 % du temps de connexion est associé aux réseaux sociaux



*How much time will adults in France spend with media in 2019?*

On average, consumers in France ages 18 and older will spend 10 hours, 6 minutes (For simplicity's sake, our style for expressing such a number

**ATTENTION :**  
« Un homme de 42 ans employé dans une entreprise du Maine-et-Loire vient d'être congédié pour un usage abusif de Facebook sur son lieu de travail.

La même sanction avait été retenue début septembre contre une salariée des Pyrénées-Atlantiques. »

## De l'utilisation des sites de réseaux sociaux

- Sondage mené par Harris Interactive,
  - 45% des recruteurs Américains déclarent utiliser les sites de réseaux sociaux (Facebook, MySpace, LinkedIn, Twitter, etc.) pour trouver des informations sur des candidats qui postulent à leurs offres d'emploi
  - 35% ont écarté des candidats en raisons ce qu'ils ont trouvé :
    - 53 % publication du candidat de photos ou d'informations provocantes ou déplacées
    - 44 % parce que l'on voit les candidats buvant ou se droguant
    - 35 % parce qu'ils crachaient sur leurs anciens employeurs, leurs collègues ou leurs clients
    - 29 % parce qu'ils montraient un déficit de communication
    - 26 % parce qu'ils publiaient des propos discriminatoires
    - 24 % parce qu'ils mentaient sur leurs diplômes et
    - 20 % parce qu'ils ont publié des informations confidentielles sur leurs anciens employeurs
- Allemagne : 28% des employeurs (500 entreprises) utilisent Internet pour recueillir des informations dès le début du recrutement



7

---



---



---



---



---



---



---



---

## Les amis de mes amis

- Entretien avec Alex Türk, président de la Cnil (Commission nationale de l'informatique et des libertés).
- « Un de ses copains a pris la photo et l'a balancé sur le réseau social. C'est amusant. Quelques mois plus tard, il était candidat sur un poste et le recruteur lui a glissé sous les yeux la photo de ses fesses en lui demandant s'il était coutumier de ces pratiques ». Source (site Internet du quotidien La Provence)
- « Oh mon dieu ! Je hais mon boulot » ajoutant que son responsable était « pervers » et qu'il ne lui donnait que « du travail de m... »
- ...4 heures plus tard...
  - « Tout d'abord arrêtez de vous flatter, cela ne fait que 5 mois que vous travaillez ici, n'avez pas remarqué que je suis gay. Ensuite le travail de m... comme vous dites est le travail pour lequel je vous paye [...]. Vous semblez avoir oublié qu'il vous restait encore deux semaines de travail en période d'essai. Ne prenez pas la peine de revenir demain. »
- Son patron était en relation sur Facebook



8

---



---



---



---



---



---



---



---

## Notre responsabilité

- Expérience de l'éditeur britannique Sophos (2007)
- Création d'un compte Freddy Staur
- Envoi de Friends à un échantillon de 200 personnes sur FaceBook
- 87 personnes ont répondu en donnant accès à des photos de familles, des informations sur leur goûts, le nom de leur compagnon, compagne, (le nom de jeune fille de leur mère) leur CV ....



9

---



---



---



---



---



---



---



---

## Une expérience

- Take this lollipop:
  - <http://www.youtube.com/watch?v=SnAxsXOcrkw>
  - Vous pouvez essayer :
  - <http://www.takethislolipop.com/>
  - Attention : vous donnez votre adresse facebook ☺  
êtes vous sur ?

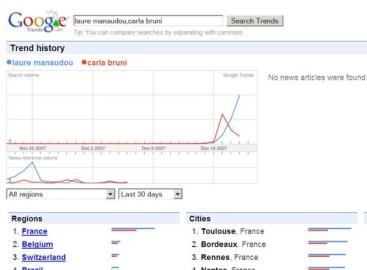


10



## Les moteurs de recherche

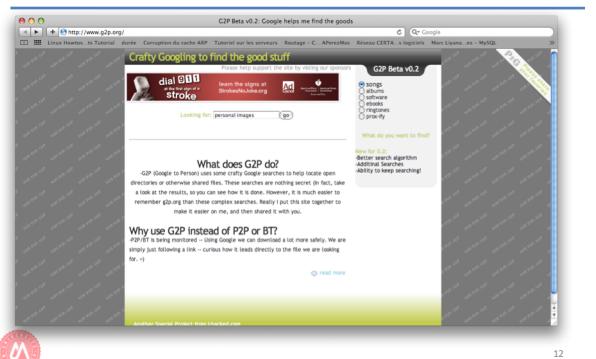
- Les photos de Laure Manaudou - décembre 2007



11



## Difficile ?



12



## Non - google requêtes complexes

La requête google :

```
intitle:index.of +"Last modified "
+"Parent directory " +(XXXXXXXXX)
+(jpeg) +"" -htm -html -php -asp
```

[XXXXMyBestFriendsXXXXXX.jpg](#)

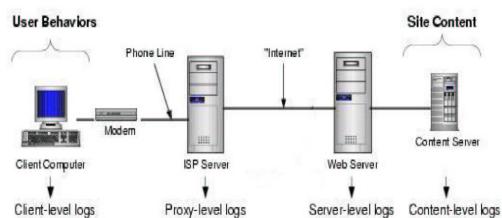


13



## Log ou Logs ?

Information sur les chemins de navigation dans les fichiers logs



14



## Web logs + ?

IP or domain name	User Id	Date and Time	Request
123.456.78.9 - [24/Oct/1999:19:13:44 -0400]		GET /images/tagline.gif HTTP/1.0"	
200 1449	<a href="http://www.teced.com/">http://www.teced.com/</a>	"Mozilla/4.51 [en] (Win98; i)"	
Status			
File Size	Referrer URL	Browser	Cookies

Bases de données des achats  
Bases de données des partenaires  
Géolocalisation  
Cookies

15



## Les exemples de clauses

### 2. COLLECTE ET UTILISATION DES INFORMATIONS

**a. Informations collectées ou reçues de votre part (ou de votre enfant autorisé)**

Nos principaux objectifs dans la collecte d'informations consistent à fournir et à améliorer nos Services, afin d'administrer votre utilisation (ou celle de votre enfant autorisé) et vous permettre (ou à votre enfant autorisé) d'en profiter et d'y naviguer facilement.

**1. Informations relatives aux comptes.**

Pendant le jeu et lorsque vous (ou votre enfant autorisé) vous inscrivez pour créer un compte sur nos Services (« Compte »), nous recueillons certaines informations qui peuvent être ensuite utilisées pour vous identifier ou vous reconnaître (ou votre enfant autorisé) (« Données à caractère personnel »). Plus précisément, du fait que vous devez posséder un compte avec Google, LinkedIn (« PTC »), ou Facebook (« PTC »), lorsque vous vous connectez à nos Services, nous recueillerons les données à caractère personnel (telles que votre adresse e-mail Google, votre adresse e-mail enregistrée sur PTC, et / ou votre adresse e-mail enregistrée sur Facebook) que vous paramétrez de confidentialité sélectionnés sur Google, PTC ou Facebook nous autorisent à accéder.

Lors de l'enregistrement d'un compte PTC, la date de naissance de l'utilisateur et le nom d'utilisateur PTC seront demandés (vous concernant ou votre enfant autorisé). Ces informations nous seront partagées (voir le paragraphe « Comptes pour les enfants » ci-dessous pour plus d'informations à ce sujet).

16

## Les exemples de clauses

nous collecterons certaines informations, telles que votre (ou celui de votre enfant autorisé) nom d'utilisateur et les messages envoyés à d'autres utilisateurs. Ces informations ne permettront pas aux autres utilisateurs de vous identifier (ou votre enfant autorisé), à moins que vous (ou votre enfant autorisé) ne choisissez d'utiliser votre (ou celui de votre enfant autorisé) nom réel et autres informations d'identification. Lorsque vous (ou votre enfant autorisé) créez un compte, nous recueillons également d'autres informations (telles que le pays et la langue) qui ne peuvent pas être utilisées pour vous identifier (ou votre enfant autorisé), à moins qu'elles ne soient combinées avec d'autres informations d'identification.

17

## Récupération d'autres données

Les « Cookies » sont de petits fichiers texte qui sont placés sur votre disque dur par un serveur Web lorsque vous (ou votre enfant autorisé) accédez à nos Services. Nous pouvons utiliser les cookies de session et les cookies persistants pour identifier que vous (ou votre enfant autorisé) vous êtes connecté aux Services et pour nous informer de la manière et la période où vous (ou votre enfant autorisé) interagissez avec nos Services. Nous pouvons également utiliser les cookies pour surveiller l'utilisation globale et le routage du trafic web sur nos services et ainsi personnaliser et améliorer nos Services. Les cookies de session sont supprimés lorsque vous (ou votre enfant autorisé) vous déconnectez des Services et fermez le navigateur. Les cookies persistants restent sur votre ordinateur et permettent d'identifier la façon dont vous utilisez les services au fil du temps. Bien que la plupart des navigateurs acceptent automatiquement les cookies, vous pouvez modifier les options de votre navigateur pour cesser d'accepter automatiquement les cookies ou pour vous avertir avant de les accepter. Sachez néanmoins que, suite à ce refus, vous (ou votre enfant autorisé) pourriez ne pas être en mesure d'accéder à toutes les sections ou caractéristiques des Services. Certains prestataires de services tiers que nous engageons (y compris des annonceurs tiers) peuvent également placer leurs propres cookies sur votre disque dur.

18

## Partage d'informations

Si des bugs, erreurs, ou d'autres incidents ou problèmes surviennent au cours du fonctionnement ou du développement des Services, alors que vous nous inscrivez pour créer un compte, nous pouvons partager vos données à caractère personnel (ou celles de votre enfant autorisé) avec TPCi et / ou TPCI si une telle collaboration s'avère nécessaire pour rechercher, diagnostiquer, corriger et / ou résoudre le problème. Toute information que vous (ou votre enfant autorisé) fournissez directement à TPCi et / ou TPCI est soumise à la Politique de confidentialité de l'entreprise applicable. Nous ne sommes pas responsables des politiques et pratiques en matière de confidentialité, de sécurité et / ou contenu de TPCi ou TPCI.

Les tiers peuvent accéder et utiliser les données



En cas de vente ou fusion

Les informations que nous collections auprès de nos utilisateurs, y compris les données à caractère personnel, sont considérées comme un actif de l'entreprise. Si nous étions rachetés par un tiers à la suite d'une transaction telle qu'une fusion, une acquisition ou une vente d'entreprise, si nos actifs étaient rachetés par un tiers pour cause de faillite ou de cessation de commerce, une partie ou la totalité de nos actifs, y compris vos données à caractère personnel (ou celles de votre enfant autorisé), pourraient être divulguées ou transférées à un tiers acquéreur dans le

Journal of Statistical Software, Volume 10, Issue 10, December 2005.



## Qui suis je ?

- Niantic



- John Hanke (Google Street View) – Récupération des données Wifi
  - Marius Milner (Hacker accusé) travaille à Niantic sous la direction de John Hanke
  - Financement initial : 20 M\$ via Nintendo et .... Google
  - A lire les clauses : <https://www.nianticlabs.com/privacy/pokemongo/fr>

21

### Une vrai valeur commerciale

- Décembre 2007, (Google, Microsoft, MySpace, AOL et Yahoo!), ont enregistré 336 milliards de données personnelles
  - Yahoo! a récolté 110 milliards de transmissions de données, soit en moyenne 811 (1.700 avec l'ensemble de ses partenaires) informations pour chaque internaute ayant visité un de ses sites durant cette période.
  - 110 milliards de données personnelles en un mois !
  - Dresser un portrait-robot fiable de l'internaute consommateur
  - De 10 à 50 euros !!



22

Tout s'achète

- Site de ventes en ligne sur les clients intéressés par la voyance
  - Nom, prénom, adresse, numéro de CB
  - 1 euro par personne  
  - A essayer :)



23

## Les bases clients protégées ?

- Janvier 2009 : 400 000 fiches du fournisseur d'accès à Internet Orange laissées en libre accès sur Internet via une faille de sécurité
  - Octobre 2008 : 30 millions de données de Deutsche Telekom (avec numéros de CB)
  - Août 2008 : les données bancaires d'un million de clients en vente sur eBay (pour 44 euros)
  - Janvier 2009 : 4 millions de comptes visités par des hackers sur Monster
  - Mars 2010 : Fichier SNCF (1 adresse et coordonnées d'un voyageur 8 à 20 euros)



24

## De l'anonymisation

- Expérience d'AOL en 2006
  - Une liste de 20 millions de recherche d'internautes mis en ligne après avoir été anonymisées
  - No. 4417749 a effectué de nombreuses recherches sur « un homme célibataire de 60 ans » et « des informations sur un chiens qui urine partout »
  - En recherchant, localisation (Lilburn, Ga), vue d'un lac, ...
  - Thelma Arnold, a 62-year-old veuve qui vie à Lilburn, Georgie



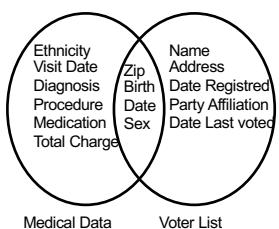
25



## De l'anonymisation

- Fichier anonymisé des soins de santé des fonctionnaires de l'état du Massachusetts mis en ligne (L. Sweeney, 1997)
  - La liste électorale de Cambridge, MA (53 805 inscrits)

□ 69 % d'enregistrements uniques par rapport à code postal, date de naissance



26

## Dossier médical du gouverneur du Massachusetts



## Encore plus loin

- Le big data ...



27

## Un petit exemple

# Un petit exemple

---

 **france inter**  **LE DIRECT**  **RÉÉCOUTER**

[répondre](#)

**Guirec (anonyme),**

mardi 29 mars 2016 à 14:48

Je suis très choqué par ce que nous vend Mathieu Roche. Avant de le vendre à la Chine, à la NSA etc... L'analyse des sentiments exprimés dans les sms...selon moi la modélisation est une atteinte grave à nos libertés sentimentales. Mathieu Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

[répondre](#)

28

---

---

---

---

---

---

---

---

## Une petite analyse manuelle

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.  
L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.  
**M. Roche** est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Quelle est l'opinion exprimée ?
  - M. Guirec n'aime pas les travaux de recherche de M. Roche sur l'analyse des sentiments dans les SMS !



29

---

---

---

---

---

---

---

## Un traitement informatique

*Guirec (Anonyme) Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.  
L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos  
libertés sentimentales.  
M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Comment détecter l'opinion ?
  - Question 1 : le texte est-il positif ou négatif ?
  - Question 2 : qui est l'auteur des critiques et le sujet de la critique ?
  - Question 3 : quel est l'objet de la critique ?



30

---

---

---

---

---

---

---

## Opinion du document

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.

Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Nécessité de faire appel à des ressources externes :

- Senticnet, Babelnet, Dictionnaires spécialisés, Ontologies



31

---



---



---



---



---



---



---



---

## Opinion du document

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.

Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Détection simple des éléments négatifs.  
Même très négatifs (intensité « Très »).  
Possibilité de mettre une fonction de scoring.

32

---



---



---



---



---



---



---



---



---

## Opinion du document

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.

Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes

- Vendre est positif ou négatif ... selon le contexte !
- “Libertés sentimentales, poésie” ... positif

33

---



---



---



---



---



---



---



---



---

## Opinion du document

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Positif, négatif, neutre ?



34

---



---



---



---



---



---



---



---

## Opinion du document

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes

- Pour résumer :
- Quelle fonction d'agrégation pour dire si le document est positif ou négatif ?
- Besoin d'autres informations



35

---



---



---



---



---



---



---



---

## Auteur et sujet de la critique

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.

L'analyse des sentiments exprimés dans les SMS... selon moi est une atteinte grave à nos libertés sentimentales.

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Nécessité de faire appel à des ressources externes :
  - Cible du sentiment, source du sentiment

CoreNLP  
StanfordFrameNet  
Berkeley

Treetagger

36

---



---



---



---



---



---



---



---

## Auteur et sujet de la critique

*Guirec (Anonyme)* Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.  
L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.  
M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Pas forcément évident
- Nous vend ?
- L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.



37

---



---



---



---



---



---



---



---



---

## Auteur et sujet de la critique

*Guirec (Anonyme)* Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.  
L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.  
M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Nécessité d'analyse fine
- Ce que nous vend ?
- L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.



38

---



---



---



---



---



---



---



---



---

## Conclusion partielle

*Guirec (Anonyme)* Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA, etc.  
L'analyse des sentiments exprimés dans les SMS ... selon moi est une atteinte grave à nos libertés sentimentales.  
M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.

- Pour l'instant il semble qu'il y ait une opinion négative exprimée par M. Guirec ?
- Une ontologie spécifique pourrait montrer que poésie et sentiments sont proches et montrer que c'est à propos de la poésie (ressource externe spécifique – généralisation de l'approche ?)



39

---



---



---



---



---



---



---



---



---

## Informations non prises en compte

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA etc.

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une*

*atteinte grave à nos libertés sentimentales.*  
*M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Il est possible de repérer des éléments de géolocalisation ou de détecter des noms propres.
  - Nécessité de ressources externes



45

---

---

---

---

---

---

---

---

---

---

---

#### Informations non prises en compte

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA etc.

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une*

*M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes.*

- Des résultats obtenus de géonames
  - Des patterns spécifiques de géolocalisation « A la, chez les »



---

---

---

---

---

---

---

Attention aux interprétations !

**Guirec (Anonyme)** Je suis très choqué par ce que nous vend M. Roche.  
Avant de le vendre à la Chine, à la NSA etc.

*L'analyse des sentiments exprimés dans les SMS ... selon moi est une*

M. Roche est ennemi de la poésie et trouvera ses ennemis chez les poètes atteinte grave à nos libertés sentimentales.

- M. Guirec exprime une opinion négative !
  - Il s'avère que la cible était difficile à définir
  - Avec des ressources extérieures (souvent nécessité de prendre en compte des aspects de géolocalisation)



42

---

---

---

---

---

---

---

# Les dérives potentielles

- M. Guérec a un avis très négatif sur le Congo (qui est subventionné par la Chine) et il exprime ses sentiments

## **La Chine, premier partenaire du Congo-Brazzaville**

La présence chinoise au Congo, ce n'est pas que des petites boutiques bon marché, loin s'en faut. La Chine est de plus en plus impliquée dans tous les grands travaux de modernisation du pays. D'un coût global de 280 millions de dollars US, la centrale hydraulique d'Imboulo est par exemple le fruit de la coopération sino-congolaise. Idem pour le tout nouvel aérogare et le projet de deuxième piste de

- Nécessité de ressources externes évidentes, nécessité d'autres données, .... nécessité de contrôler



43

## Une entreprise de télécommunications

- Les consommations des utilisateurs sur 3 ans

Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
1	Nicholson Jack	647 224 8984	Paris	2y	12.00
3	Streep Meryl	241 351 3938	London	3y	189.45
4	De Niro Robert	635 345 7799	New York	3y	77.10
6	Pacino Al	854 478 7448	Singapore	3y	369.00
7	Dav Lewis	658 212 8888	Tokyo	3y	131.00
8	Hannibal Dorian	655 872 9984	Tokyo	2y	459.37
11	Monroe Marilyn	613 742 2781	Beijing	3y	830.00
12	Hopkins Anthony	837 378 6380	Cairo	3y	38.78
15	Newman Paul	831 789 7892	Jakarta	3y	299.29
17	Washington Denzel	838 795 2343	Bogota	4y	243.00
18	Wisebacker Harrison	835 789 2343	Paris	50.18	50.18
20	Pennekamper Sam	645 892 8821	Santiago	3y	628.01
21	Blanchett Cate	631 881 1890	Berlin	3y	33.79
22	DiCaprio Leonardo	643 909 8819	Nairobi	3y	8.00
24	Baldwin Robert	671 713 0000	Los Angeles	3y	26.23
26	Depp Johnny	687 713 0000	Moscow	2y	85.11
28	Brooks Jeff	698 382 6514	Toronto	3y	92.75
31	Cruise Russell	689 139 9487	Munich	3y	1,044.48
33	Kidman Nicole	674 270 7824	Tokyo	3y	0.96



44

## Une entreprise de télécommunications

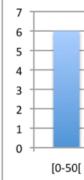
- Un problème de rentabilité
    - Il faut supprimer les utilisateurs non rentables
    - Lesquels faut il garder ?
    - Quel message donner aux autres pour les conserver ?
  - Hypothèses :
    - les utilisateurs sont indépendants
    - Pas de particularité sur la distribution des valeurs de profit



45

## Une approche classique - 1

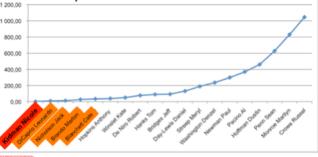
- Un aperçu de la distribution

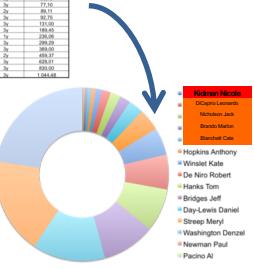


Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
33	Kidman Nicole	674 270 7824	Tokyo	3y	<b>0.96</b>
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	<b>8.00</b>
1	Nicholson Jack	647 224 8984	Paris	2y	<b>12.00</b>
24	Brando Marlon	645 891 1024	Los Angeles	3y	<b>26.23</b>
21	Blanchett Cate	651 891 1059	London	3y	<b>33.79</b>
12	Hopkins Anthony	638 379 6380	Cairo	3y	<b>38.78</b>
18	Winstole Kate	656 980 8793	Hanoi	3y	50.18
4	De Niro Robert	633 345 8799	New York	3y	77.10
26	Hanks Tom	667 017 6390	Montpellier	2y	89.11
28	Bridges Jeff	698 382 8614	Toronto	3y	92.75
7	Day-Lewis Daniel	641 235 8684	Delhi	3y	131.00
3	Streep Meryl	647 231 3938	London	3y	189.45
17	Washington Denzel	624 798 2343	Bogotá	1y	236.06
15	Newman Paul	633 789 7892	Jakarta	3y	299.29
6	Pacino Al	643 804 8884	Singapore	3y	390.00
8	Hoffman Dustin	655 870 9963	Tokyo	2y	459.37
20	Penn Sean	645 892 8921	Santiago	3y	628.01
11	Monroe Marilyn	613 742 7361	Beijing	3y	830.00
31	Crowe Russel	689 139 4947	Munich	3y	1 044.48

 46

## Une approche classique - 2



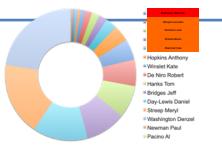


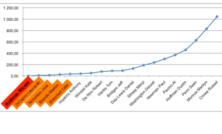
Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
33	Kidman Nicole	674 270 7824	Tokyo	3y	<b>0.96</b>
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	<b>8.00</b>
1	Nicholson Jack	647 224 8984	Paris	2y	<b>12.00</b>
24	Brando Marlon	645 891 1024	Los Angeles	3y	<b>26.23</b>
21	Blanchett Cate	651 891 1059	London	3y	<b>33.79</b>
12	Hopkins Anthony	638 379 6380	Cairo	3y	<b>38.78</b>
18	Winstole Kate	656 980 8793	Hanoi	3y	50.18
4	De Niro Robert	633 345 8799	New York	3y	77.10
26	Hanks Tom	667 017 6390	Montpellier	2y	89.11
28	Bridges Jeff	698 382 8614	Toronto	3y	92.75
7	Day-Lewis Daniel	641 235 8684	Delhi	3y	131.00
3	Streep Meryl	647 231 3938	London	3y	189.45
17	Washington Denzel	624 798 2343	Bogotá	1y	236.06
15	Newman Paul	633 789 7892	Jakarta	3y	299.29
6	Pacino Al	643 804 8884	Singapore	3y	390.00
8	Hoffman Dustin	655 870 9963	Tokyo	2y	459.37
20	Penn Sean	645 892 8921	Santiago	3y	628.01
11	Monroe Marilyn	613 742 7361	Beijing	3y	830.00
31	Crowe Russel	689 139 4947	Munich	3y	1 044.48

 47

## Conclusions

Number	Name	Phone	City	Plan	Avg. 3m Profit in \$
33	Kidman Nicole	674 270 7824	Tokyo	3y	<b>0.96</b>
22	DiCaprio Leonardo	643 909 8918	Nairobi	3y	<b>8.00</b>
1	Nicholson Jack	647 224 8984	Paris	2y	<b>12.00</b>
24	Brando Marlon	645 891 1024	Los Angeles	3y	<b>26.23</b>
21	Blanchett Cate	651 891 1059	London	3y	<b>33.79</b>
12	Hopkins Anthony	638 379 6380	Cairo	3y	<b>38.78</b>





Clients à ne pas retenir  
6 sur 19  
Gain : **119,76 \$**

 48

## Une intuition...

- Big data ... beaucoup de données ?
  - et si il y avait d'autres données ?
    - *Data Linking* et intégration



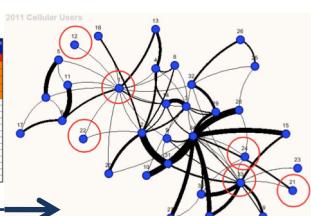
49



## Données additionnelles

- « Inter-call network » avec les fréquences
  - Ceux qui sont connectées avec les 19 personnes

Number	Name	Phone	City	Plan	Avg.	On Profit
1	Kumar Thomas	6373 7825	Bang	3y	\$ 10.00	
22	D'Cunha Leander	6353 8935	Nasik	3y	\$ 8.69	
23	Nihalani Jack	622 226 8944	Paris	2y	\$ 12.00	
24	Shankar Suresh	635 811 1860	Delhi	3y	\$ 10.00	
25	Blankheit Chetan	635 811 1860	Delhi	3y	\$ 33.79	
12	Halden Anthony	636 378 6380	Cairo	3y	\$ 10.00	
13	Shankar Suresh	635 811 1860	Delhi	3y	\$ 10.00	
4	De Niro Robert	635 455 9799	New York	2y	\$ 77.10	
28	Jeffrey Katzenbach	636 382 6641	Toronto	3y	\$ 82.75	
29	Brennan Jim	636 382 6641	Toronto	3y	\$ 82.75	
30	Shane Mory	621 221 3593	London	3y	\$ 189.45	
31	Wynona Judd	635 811 1860	Delhi	3y	\$ 10.00	
15	Neuman Yvonne	635 811 7820	Jakarta	2y	\$ 29.95	
16	McGowan Jennifer	635 811 7820	Jakarta	2y	\$ 29.95	
8	Hoffman Dustin	635 811 9553	Englewood	2y	\$ 49.95	
20	Perry Steve	635 811 9521	Santiago	2y	\$ 628.01	
21	Morrissey Liam	635 811 9521	Santiago	2y	\$ 628.01	

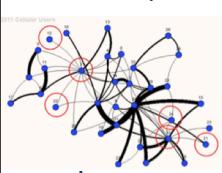


50



## Données additionnelles

- Algorithmes de détection de communautés (*global community detection*)



---

---

---

---

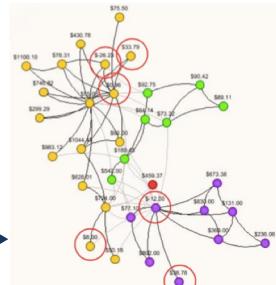
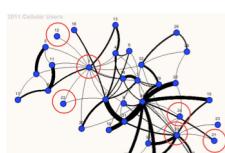
---

---

---

## Données additionnelles

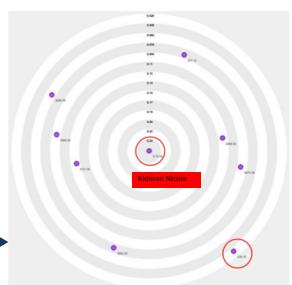
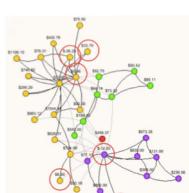
- Algorithme de détection de communautés (*local community mining*)



52

## Données additionnelles

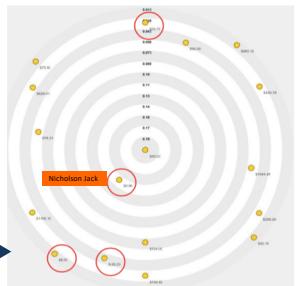
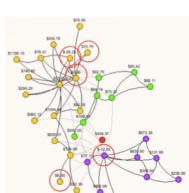
- Centralité par Communauté (Nicole Kidman)



53

## Données additionnelles

- Centralité par Communauté (Jack Nicholson)



54

## Autres conclusions

Risque de perte :  
Nicole Kidman : **3145,32 \$** (0,96 \$)  
Jack Nicholson : **6324,14 \$** (8 \$)

Exploiter des données additionnelles et des techniques d'analyses sophistiquées peuvent offrir de nouvelles perspectives

33	Kidman Nicole	0,96
22	DiCaprio Leonardo	8,00
1	Nicholson Jack	12,00
24	Brando Marlon	26,23
21	Blanchett Cate	33,79
12	Hopkins Anthony	38,78

55

## A l'origine ...

**Application-Controlled Demand Paging for Out-of-Core Visualization**

Michael Cox  
MRJ/NASA Ames Research Center  
Microcomputer Research Lab, Intel Corporation  
[mcbs@nas.nasa.gov](mailto:mcbs@nas.nasa.gov)

David Eberle  
MRJ/NASA Ames Research Center  
[ceberle@nas.nasa.gov](mailto:ceberle@nas.nasa.gov)

**Abstract**  
In the area of scientific visualization, input data sets are often very large. In visualization of Computational Fluid Dynamics (CFD) in particular, input data sets today can surpass 100 Gbytes, and are expected to scale with the ability of supercomputers to solve larger problems. One way to handle already partitioned large data sets into segments, and load appropriate segments as they are needed. However, this does not remove the problem for two reasons: 1) there are data sets

**Publication of:**  
• Conference  
VIS97 IEEE Visualization '97 Conference  
Phoenix, AZ, USA **October 18 - 24, 1997**  
IEEE Computer Society Press Los Alamitos, CA, USA ©1997

56

## Le Big Data s'affiche...

ACM DIGITAL LIBRARY

**Visually exploring gigabyte data sets in real time**  
Full Text: [Get this Article](#)  
Authors: Steve Bryson, Michael Cox, David Eberle, Michael Koenig, NASA Ames Research Center, Moffett Field, CA  
Published in: [Communications of the ACM](#), Volume 42 Issue 8 Aug. 1999  
DOI: <https://doi.org/10.1145/930321087>

**Introduction**  
David N. Koenig  
MBI Technology Solutions  
NASA Ames Research Center  
[dko@nas.nasa.gov](mailto:dko@nas.nasa.gov)

**Big Data**  
There are two distinct types of "big data":  
• big data collections  
• big data objects  
Big data collections are aggregations of many datasets (e.g., sensor data from climate systems).  
Big data objects are single datasets from complex simulations (e.g., complex fluid dynamics, structural analysis).

57

# Numéro spécial dans Nature

## nature International weekly journal of science

Journal home > Archive > Editorial > Full Text

Editorial

Nature 458, 1 (4 September 2008) | doi:10.1038/455001a; Published online 1 September 2008

### Community cleverness required

Researchers need to adapt their institutions and practices in response to the growth of data – and need to complement smart searching.

The Internet search firm Google was incorporated Just 10 years ago this week. Going from a collection of denoted servers housed under a desk to a global network of data centers, Google's growth in information by the petabyte, Google's growth in data is a key theme of this issue. It makes an apt moment for this special issue of *Nature*, which examines what big data sets mean for contemporary science.

'Big', of course, is a moving target. The possibility of the tens of gigabytes we carry around on USB sticks would have seemed like

Journal content

- Journal home
- Advance online publication
- Current issue
- Nature News
- Archive
- Supplements
- Web focuses
- Podcasts
- Videos
- News Specials

Journal Information

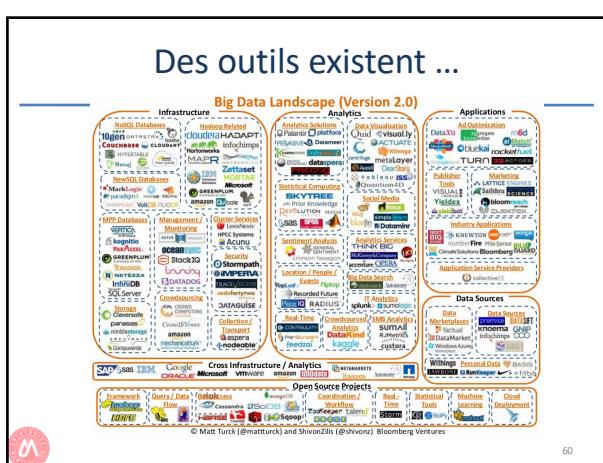
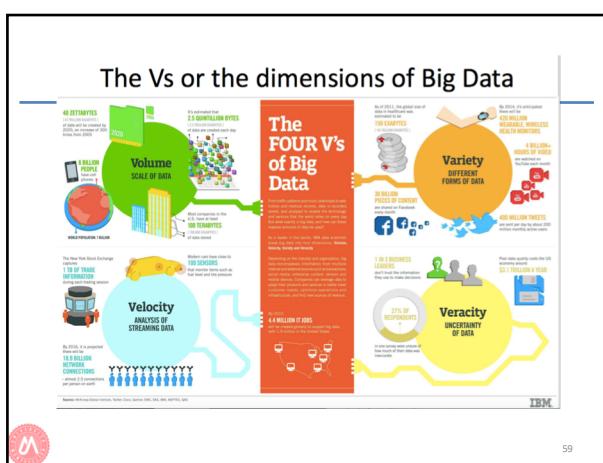
Search  Advanced search

Log in 

subscribe to nature

FULL TEXT

- Previous | Next ▾
- Table of contents
- Download PDF
- View interactive PDF in ReadCube
- Share this article
- CrossRef lists 33 articles citing this article
- Scopus lists 34 articles citing this article



- .... fin de la paranthèse)



61

---

---

---

---

---

---

---

## Plan

- Concrètement ?
- Pourquoi fouiller les données ?
- Le processus d'extraction
- Un aperçu de quelques techniques



62

---

---

---

---

---

---

---

## Pourquoi fouiller les données ?

- De nombreuses données sont collectées et entreposées
  - Données du Web, e-commerce
  - Achats dans les supermarchés
  - Transactions de cartes bancaires
- La pression de la compétition est de plus en plus forte
  - Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)



63

---

---

---

---

---

---

---

## Pourquoi fouiller les données ?

- Les données sont collectées et stockées rapidement (GB/heures)
  - Capteurs : RFID, supervision de procédé
  - Télescopes
  - Puces à ADN générant des expressions de gènes
  - Simulations générant de téraoctets de données



64

---

---

---

---

---

---

---

---

## Pourquoi fouiller les données ?

- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)
 

*Comment explorer des millions d'enregistrements avec des milliers d'attributs ?*
- Besoins de répondre rapidement aux opportunités
- Requêtes traditionnelles (SQL) impossibles
 

*« Rechercher tous les enregistrements indiquant une fraude »*
- Croyance dans la présence de données importantes



65

---

---

---

---

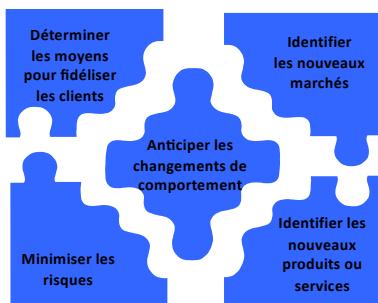
---

---

---

---

## Un enjeu stratégique



66

---

---

---

---

---

---

---

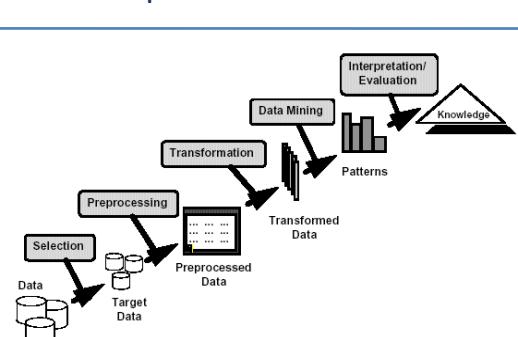
---

## Qu'est ce que le Data Mining ?

- De nombreuses définitions
    - Processus **non trivial** d'extraction de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles, compréhensibles, ....
    - Exploration et analyse, **par des moyens automatiques ou semi-automatiques**, de grandes quantités de données en vue d'extraire des motifs intéressants

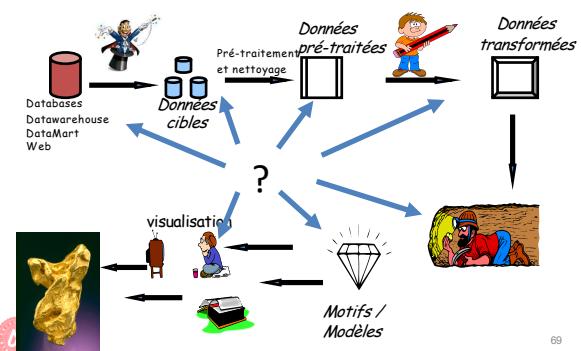


65

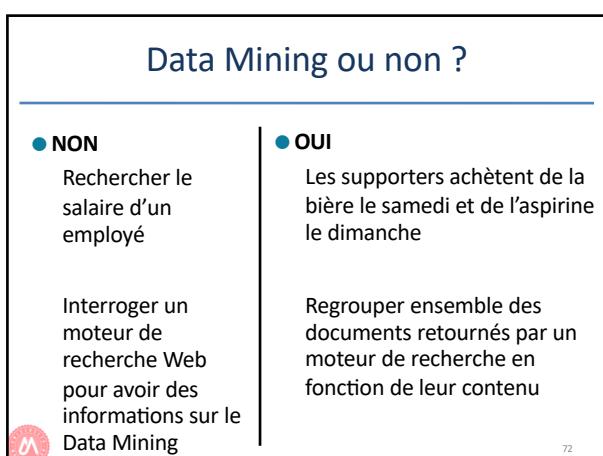
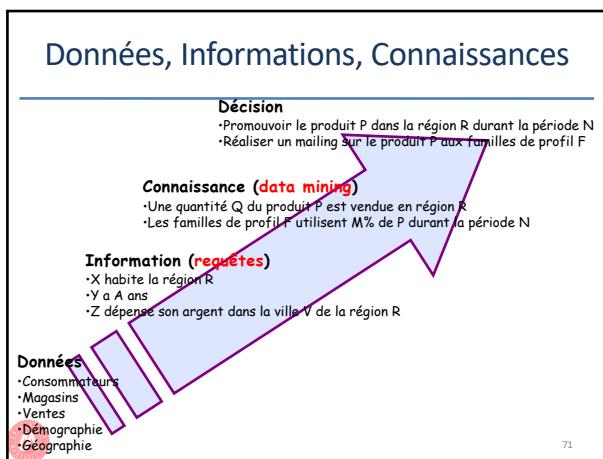
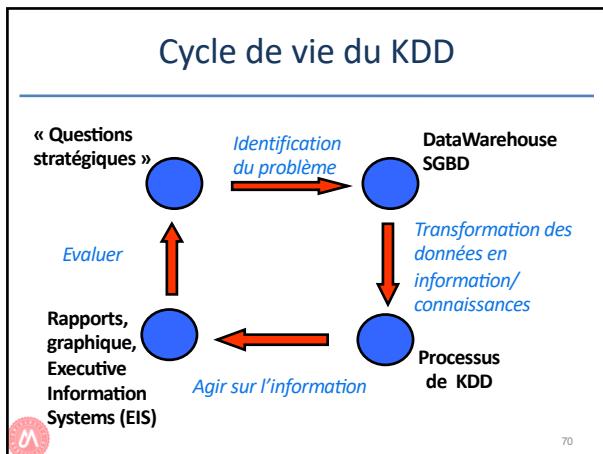


65

## Le processus de KDD



3



## Un exemple d'analyse et de fouille de données

- Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique, cinéma.
  - Il veut étudier ses clients pour découvrir de nouveaux marchés ou vendre plus à ses clients habituels.
  - Questions :
    1. Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
    2. A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
    3. Est-ce que les acheteurs de magazines de musique sont aussi amateurs de cinéma ?
    4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
    5. Peut-on prévoir les pertes de clients et prévoir des mesures pour les diminuer ?

Source Data Mining, Adrian & Zantig 1996

73



## Un exemple d'analyse et de fouille de données

- Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
    - Une requête SQL à partir des données suffit
  - A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
    - Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés. Requêtes multidimensionnelles éventuellement de type OLAP.

74



## Un exemple d'analyse et de fouille de données

- Est-ce que les acheteurs de magazine de musique sont aussi amateurs de cinéma ?
    - Exemple simplifié de problème où l'on demande si les données vérifient une règle : on connaît les acheteurs de magazine de musique on regarde s'ils aiment le cinéma
    - Réponse formulée par une valeur estimant la probabilité que la règle soit vraie. Utilisation d'outils statistiques ou de requêtes sur une base de données

75



## Un exemple d'analyse et de fouille de données

- Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
  - Question ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser.
- Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?
  - Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, ...

Il s'agit de tâches de fouille de données

76

---



---



---



---



---



---



---



---



---

## Applications

- Médecine : bio-médecine, drogue, Sida, séquence génétique, gestion hôpitaux, ...
- Finance, assurance : crédit, prédiction du marché, détection de fraudes, ...
- Social : données démographiques, votes, résultats des élections,
- Marketing et ventes : comportement des utilisateurs, prédiction des ventes, espionnage industriel, ...
- Militaire : fusion de données .. (secret défense)
- Astrophysique : astronomie, « contact » (-))
- Informatique : agents, règles actives, IHM, réseau, Data-Warehouse, Data Mart, Internet (moteurs intelligent, profiling, text mining, ...)

77

---



---



---



---



---



---



---



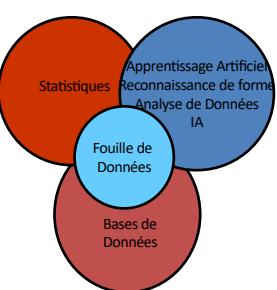
---



---

## Quid des données ?

- Grandes Bases de Données ou non ?
- Faut-il échantillonner ?
  - 100 000 enregistrements, 100 Mo par jour
  - 2 Go par jour, 100 Go par heure
  - ... Déjà les petabyte ( $P^{30}$ ) ...
- Différents domaines
  - Bases de Données
  - Intelligence Artificielle (Machine Learning)
  - Statistiques
  - Algorithmique,
  - Visualisation...



78

---



---



---



---



---



---



---



---



---

# Quid du type de données ?

- Enregistrements
  - Tuples en Relational
  - Matrice de données
  - Données document : textes, vecteur de fréquences des termes
  - Données de transactions
- Graphes et réseaux
  - World Wide Web
  - Réseaux sociaux
  - Structures moléculaires
- Ordonnées
  - Données vidéo : séquences d'images
  - Données temporelles : séries temporelles
  - Données séquentielles : séquences de transactions
  - Données de séquences génétiques
- Spatiales, images et multimedia :
  - Données spatiales : cartes
  - Données Image
  - Données vidéo

	name	age	gender	pid	rel	socia	genre	u.	w.	geo	label	series
Document 1	3	0	5	0	2	6	0	2	0	2		
Document 2	0	7	0	2	1	0	0	0	3	0		
Document 3	0	1	0	0	1	2	2	0	3	0		

<i>TID</i>	<i>Items</i>
1	Pain, Coca, Lait
2	Bière, Pain
3	Bière, Coca, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca, Couches, Lait



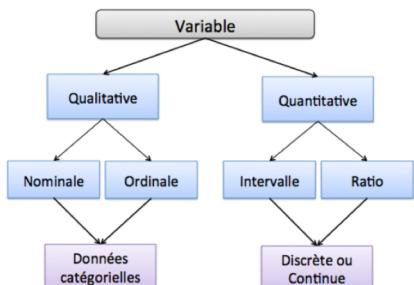
# Les objets données

---

# Attributs

---

## Les différents types



83

---

---

---

---

---

---

---

## Attributs à valeurs nominales

- Les valeurs sont des symboles (des noms)
  - Temps = {Ensoleillé, Pluvieux, Neigeux, Gris}
  - Code Postal, Numéro d'étudiant, couleurs yeux...
- Aucune relation (ordre ou distance) entre les nominaux n'existe
- Seuls des tests d'égalité peuvent être exécutés
- Exemple de règle:
  - If Temps = Pluvieux Then Match = No




---

---

---

---

---

---

---

## Attributs à valeurs nominales

- Cas particulier :
- Valeurs nominales binaires : attribut nominal avec seulement 2 états (0 ou 1)
- Binaire symétrique : les deux résultats ont même importance
  - Genre (Masculin/Féminin)
- Binaire asymétrique : résultat significatif
  - Test médical (positif vs négatif)
  - Traditionnellement 1 quand résultat positif




---

---

---

---

---

---

---

## Attributs à valeur ordinaire

- Une notion d'ordre s'impose sur les ordinaux
- Mais l'amplitude entre les valeurs n'est pas connue
- Les opérations d'addition et de soustraction ne sont pas possibles
- Exemple de règle :
  - Température décrite par les adjectifs {chaud, froid, moyen}, et chaud > moyen > froid
  - If température > froid Then match = Yes

86

---



---



---



---



---



---



---



---

## Attributs à valeur ordinaire

- Ils peuvent être convertis en booléen

	FROID	MOYEN	CHAUD
Canada	TRUE	FALSE	FALSE
France	FALSE	TRUE	FALSE
Seychelles	FALSE	TRUE	TRUE

87

---



---



---



---



---



---



---



---

## Attributs de type intervalle

- Les intervalles impliquent une notion d'ordre, et les valeurs sont mesurées dans des unités spécifiques et fixées
- Le point zéro n'existe pas où ne correspond en rien à l'absence de phénomène
- Exemples :
  - Le calendrier
  - La température exprimée en degrés Celsius ou Fahrenheit (0 en C -> température de congélation de l'eau, 0 en F -> température de solidification d'un mélange à part égal d'eau et de chlorure d'ammonium )

88

---



---



---



---



---



---



---



---

## Attributs de type ratio

- Il existe un point zéro universel
- Toutes les opérations mathématiques sont autorisées sur les attributs de ce type
- Exemple:
  - L'attribut distance : on peut comparer, additionner 2 distances, la distance entre un objet et lui-même est zéro
  - Le poids



89

---

---

---

---

---

---

---

---

## Attributs discrets et continus

- Une variable discrète prend un nombre fini ou dénombrable de valeurs
  - Nombre de mots dans un document, nombre d'habitants
  - Généralement un entier
- Une variable continue peut prendre un nombre infini ou non dénombrable de valeurs
  - Température, poids, taille
  - Généralement un réel



90

---

---

---

---

---

---

---

---

## Quid du type de données ?

- Booléennes, Numériques, Symboliques, Multidimensionnelles, Textuelles, Images, ...
- ... ce n'est pas le monde des bisounours



91

---

---

---

---

---

---

---

---

## Pourquoi pré-traiter les données

- Les données du monde réel sont sales :
    - Incomplètes : manque de valeurs d'attributs, manque d'attributs intéressants, ne contenant que des données agrégées
      - métier=“”
    - Bruitées : contenant des erreurs ou des outliers
      - Salaire=“-10”
    - Inconsistantes : avec des incohérences dans les codes ou les noms
      - Age=“42” Anniversaire=“11/07/1990”
      - Notation initiale “1,2,3”, notation actuelle “A, B, C”
      - Incohérences entre deux enregistrements similaires



92

---

---

---

---

---

---

---

---

---

---

## Pourquoi ?

- Incomplètes
    - Valeur pas applicable au moment de la collecte
    - Temps différent entre la collecte et l'analyse
    - Problème techniques/humain
  - Bruitées (valeurs incorrectes)
    - Défaut d'instrument
    - Erreur humaine ou de l'ordinateur au moment de l'entrée
    - Erreur de transmission de la donnée
  - Inconsistances
    - Différentes sources de données
    - Violation des dépendances fonctionnelles



93

---

---

---

---

---

---

---

## Pourquoi le pré-traitement est important ?

- Sans données de qualité, il n'y a pas de bons résultats de fouille !
  - Toujours regarder les données : 90% des échecs sont liés à la qualité des données !!
  - Les étapes de recherche des données, de nettoyage, de transformation correspondent à la phase la plus longue et la plus importante du processus



94

---

---

---

---

---

---

---

## Pré-traitement des données

- Nettoyer les données
    - Corrections des doublons, des erreurs de saisie
    - Contrôle sur l'intégrité des domaines de valeurs :
      - détection des valeurs aberrantes
      - détection des informations manquantes
  - Intégration des données et transformation
  - Réduction



95



## Pré-traitement des données

- Correction des doublons et des erreurs de saisie

Client	Nom	Adresse	Position	Date Abonnement	Magazine
2807	Dupond	Av du Palais, Paris	Cadre	12/08/2011	Voiture
2807	Dupond	Av du Palais, Paris	Enseignant	11/07/2014	Musique
2807	Dupond	Av du Palais, Paris	Cadre	09/05/2016	BD
3456	Durand	Av de la mer, Nice	Employe	32/02/2222	BD
4356	Duchemin	Rue Principale, Grenoble	Enseignant	13/06/2015	Sport
5832	Dujardin	Place centrale, Lille	Employe	17/07/2016	NULL
2806	Durant	Rue des Chasseurs, ?	Medecin	14/04/2006	Sport
2807	Dupont	Av du Palais, Paris	Cadre	32/02/2226	Maison



96



## Pré-traitement des données

- Intégrité de domaine ou dépendances fonctionnelles non vérifiées

Client	Nom	Adresse	Position	Date Abonnement	Magazine
2807	Dupond	Av du Palais, Paris	Cadre	12/08/2011	Voiture
2807	Dupond	Av du Palais, Paris	Enseignant	11/07/2014	Musique
2807	Dupond	Av du Palais, Paris	Cadre	09/05/2016	BD
3456	Durand	Av de la mer, Nice	Employe	32/02/2222	BD
4356	Duchemin	Rue Principale, Grenoble	Enseignant	13/06/2015	Sport
5832	Dujardin	Place centrale, Lille	Employe	17/07/2016	NULL
2806	Durant	Rue des Chausseurs, ?	Médecin	14/04/2006	Sport
2807	Dupond	Av du Palais, Paris	Cadre	32/02/2222	Maison



97



## Pré-traitement des données

- Information manquante
  - Supprimer l'enregistrement
    - A faire si la classe est manquante car n'aide pas à la classification
  - Remplir manuellement les champs :
    - difficile et long
  - Automatiquement :
    - Remplacer un salaire manquant par le salaire médian des clients
    - Prédire les valeurs manquantes, en le déduisant d'autres paramètres (salaire à partir de l'âge et de la profession)
    - Inférer la valeur avec un algorithme de classification (la valeur à prédire devient la classe recherchée)



95

---

---

---

---

---

---

---

---

---

---

## Pré-traitement des données

- Données bruitées : Plusieurs solutions : lissage, segmentation, régression linéaire
  - Techniques de lissage (*data smoothing*) :
    1. Trier les différentes valeurs de l'attribut considéré : {4, 8, 15, 21, 21, 24, 25, 28, 34}
    2. Partitionner l'ensemble résultat.  
{ {4, 8, 15}, {21, 21, 24}, {25, 28, 34} }
    3. Remplacer les valeurs initiales par de nouvelles valeurs en fonction du partitionnement réalisé :
      - par la valeur moyenne des regroupements réalisés (9, 22, 29)
      - par les min et max des regroupements réalisés. {{4, 4, 15}, {21, 24}, {25, 25, 34}}
  - Implique une perte de précision ou d'information



85

---

---

---

---

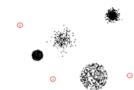
---

---

---

## Pré-traitement des données

- Utilisation de la fouille pour aider à pré-traiter les données
  - Techniques de segmentation (clustering) :
    - Les valeurs similaires sont placées dans une même classe
    - On ne tient pas compte des valeurs isolées (dans une classe comportant trop peu d'éléments)



100

---

---

---

---

---

---

---

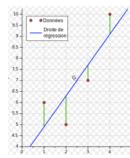
---

---

---

## Pré-traitement des données

- Techniques de régression linéaire :
- Hypothèse : un attribut Y dépend linéairement d'un attribut X
  - Années d'expérience X et salaire Y
- Trouver les coefficients a et b tels que  $Y = aX + b$
- Remplacer les valeurs de Y par celles prédictes
- Données de départ :
  - Un ensemble de couples  $(X_i, Y_j)$
- Détermination des coefficients :
  - Soient  $\bar{X}$  et  $\bar{Y}$  les valeurs moyennes des attributs X et Y.
  - $a = \text{cov}(x,y)/\text{var}(x)$
  - $b = \bar{Y} - a\bar{X}$

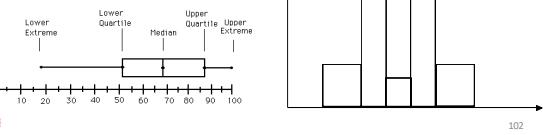


101



## Pré-traitement des données

- Statistiques descriptives sur les données :
- Utiles pour voir la centralité, la dispersion, les variations, les distributions
  - Valeur médiane, moyenne, variance, écart type, mode, quantiles
  - Boxplots, histogrammes,



102



## Attention aux interprétations !



103



## Attention aux interprétations !




---

---

---

---

---

---

---

## Attention aux interprétations!

**ON TEENAGERS, ADULT:**  
**S**tatistics show that  
 teen pregnancy  
 drops off significantly  
 after age 25.  
Mary Axen Tellez, Republican state senator from Colorado Springs  
 (Contributed by Harry F. Power)

**MONDAY DECEMBER 1999**

Data doesn't create meaning, we do. —Susan Etlinger

---

---

---

---

---

---

---

## Attention aux données !

Quartet d'Ascombe								
I		II		III		IV		
x	y	x	y	x	y	x	y	x
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58	
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76	
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71	
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84	
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47	
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04	
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25	
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50	
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56	
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91	
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89	

---

---

---

---

---

---

---

## Attention aux données !

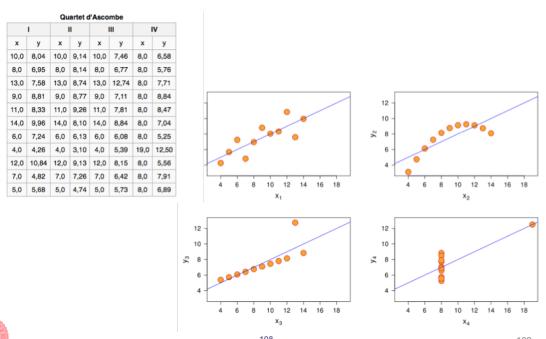
Propriété	Valeur
Moyenne des x	9,0
Variance des x	10,0
Moyenne des y	7,5
Variance des y	3,75
Corrélation entre les x et les y	0,816
Équation de la droite de régression linéaire	$y = 3 + 0,5x$
Somme des carrés des erreurs relativement à la moyenne	110,0



107

107

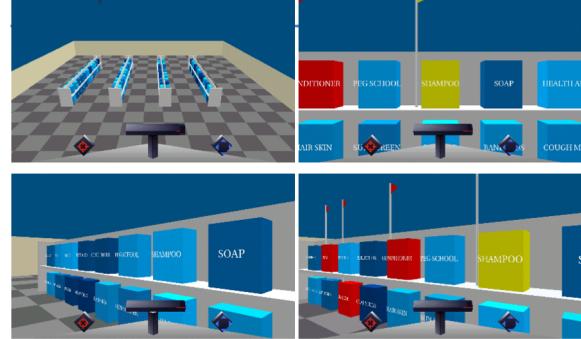
## Importance de la visualisation



108

108

## Importance de la visualisation

Intelligent Miner ([www.ibm.com](http://www.ibm.com))

109

## Pré-traitement des données

- Intégration de données
    - Combiner des données de différentes sources en un seul lieu (ETL/Entrepôt)
  - Intégration de schéma
    - A.cust-id  $\equiv$  B.cust #
  - Identification des entités
    - Bill Clinton = William Clinton
  - Détecter et résoudre les conflits de valeurs dans les données
    - Unités différentes (Km  $\leftrightarrow$  miles)



110

---

---

---

---

---

---

---

---

---

---

## Pré-traitement des données

- Possibilités d'avoir de nouvelles données manquantes ou aberrantes

Client	Date Naissance	Salaire	Propriétaire	Voiture
Dupond	05/07/1973	20K	OUI	OUI
Durand	13/11/1995	2K	NON	OUI
Duchemin	12/09/1987	12K	NON	NON
Durant	01/22/1833	32K	NON	NON



111

---

---

---

---

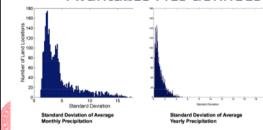
---

---

---

## Transformation des données

- Dépend de l'algorithme de fouille utilisé
  - Regroupements
    - Cas où les attributs prennent un très grand nombre de valeurs discrètes (e.g. adresses que l'on peut regrouper en régions, rapports mensuels en rapports annuels, âge -> jeune, vieux)
    - Agréger des attributs
    - Avantages : les données agrégées ont moins de variations



112

- Pays      15 valeurs différentes
- Ville      3 000 valeurs différentes
- Rues      10 000 valeurs différentes

---

---

---

---

---

---

---

---

---

## Transformation des données

- Attributs discrets
    - Les attributs discrets symboliques prennent leurs valeurs dans un ensemble fini donné (e.g. colonne magazine de l'exemple).
    - Deux représentations possibles : représentation verticale ou représentation horizontale ou éclatée (plus adaptée à la fouille de données)



113

---

---

---

---

---

---

---

---

---

---

## Transformation des données

- Représentation verticale vs éclatée

Client	Magazine
2807	Voiture
2807	Musique
2807	BD
3456	BD
4356	Sport
2806	Sport
2807	Maison

Client	Voiture	Musique	BD	Sport	Maison
2807	1	1	1	0	1
3456	0	0	1	0	0
4356	0	0	0	1	0
2806	0	0	0	1	0



114

---

---

---

---

---

---

---

## Transformation des données

- Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne (e.g. date de naissance)
  - Uniformisation d'échelle
    - Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes



115

---

---

---

---

---

---

---

## Transformation des données

- Un exemple de transformation

Client	Voiture	Musique	BD	Sport	Maison	DN	REV	Prop	Voiture	PN	DA
2807	1	1	1	0	1	45	20	OUI	OUI	1	7
3456	0	0	1	0	0	23	2	NON	OUI	0	NULL
4356	0	0	0	1	0	31	12	NON	NON	0	3
2806	0	0	0	1	0	35	32	NON	NON	NULL	12

Avec

DN : Date de Naissance -> âge  
 REV : Revenu  
 Prop : Propriétaire  
 PN : Paris/Province  
 DA : première date d'abonnement



116

## Similarité ou dissimilarité

- **Similarité**
  - Mesure de la ressemblance de deux objets
  - Plus les objets sont semblables plus grande est la valeur
  - Généralement dans l'intervalle [0,1]
- **Dissimilarité** (e.g., distance)
  - Mesure de la différence entre deux objets
  - Plus la distance est courte plus les objets sont proches
  - La dissimilarité minimale est souvent 0
  - Les limites supérieures sont très variables
- **La proximité fait référence à la similarité ou la dissimilarité**



117

## Matrice et matrice de dissimilarité

- **Matrice de données**
  - N points avec n dimensions

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- **Matrice de dissimilarité**

- N points mais seules les distances sont enregistrées
- Une matrice triangulaire

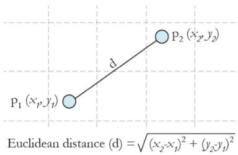
$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$



118

## Attributs numériques

- La distance Euclidienne est la distance « normale » entre deux points



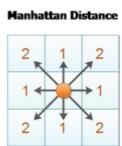
$$\begin{aligned} d(\mathbf{p}, \mathbf{q}) &= d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}. \end{aligned}$$

119



## Attributs numériques

- La distance Manhattan (inspirée des chauffeurs de taxi à Manhattan)



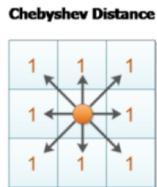
$$d = \sum_{i=1}^n |\mathbf{x}_i - \mathbf{y}_i|$$

120



## Attributs numériques

- La distance de Chebyshev (« le plus long chemin »)



a	b	c	d	e	f	g	h
8	5	4	3	2	2	2	2
7	5	4	3	2	1	1	2
6	5	4	3	2	1	1	2
5	5	4	3	2	1	1	2
4	5	4	3	2	1	1	2
3	5	4	3	2	1	1	2
2	5	4	3	2	1	1	2
1	5	5	5	5	5	5	1
a	b	c	d	e	f	g	h

$$d(i, j) = \lim_{h \rightarrow \infty} \left( \sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

121



## Attributs numérique

- La distance de Minkowski : une généralisation des distances précédentes :

$$d(i, j) = \sqrt{h} |x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \dots + |x_{ip} - x_{jp}|^h$$

Où  $i = (x_{i1}, x_{i2}, \dots, x_{ip})$  et  $j = (x_{j1}, x_{j2}, \dots, x_{jp})$  sont des objets de  $p$  dimensions (La distance est appelée la L- $h$  norme)

Si  $h=1$  → distance Manhattan,  $h=2$  → distance Euclidienne,  $h \rightarrow \infty$  → distance Chebyshev



122

## Attention aux distances

- Rappel : Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

De qui Clara est la plus proche :  
Marie ou Léa ?



123

## Attention aux distances

- Rappel : Attention certains algorithmes sont basés sur des calculs de distance entre enregistrements. Les variations d'échelles entre ces algorithmes peuvent perturber ces algorithmes

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Lea	60	11122
Lucy	60	11074

De qui Clara est la plus proche :  
Marie ou Léa ?

De Léa :  
Diff(Age) Léa = 10, Marie 20  
Diff (Salaire) Léa = 122, Marie 100



124

## Attention aux distances

- Utilisation d'une distance de Manhattan

Nom	Age	Salairé
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

$$\begin{aligned}d(\text{Clara}, \text{Marie}) &= 120 \\d(\text{Clara}, \text{Léa}) &= 132\end{aligned}$$

Clara est plus éloignée de Léa !

## Problème d'échelle des données



125

---

---

---

---

---

---

---

---

---

---

## Normalisation

- Normalisation des attributs : valeurs trop grandes qui pénalisent les distances
  - Normalisation min-max en  $[new\_min, new\_max]$

$$v' = \frac{v - min_{\mathcal{A}}}{max_{\mathcal{A}} - min_{\mathcal{A}}} (new\_max_{\mathcal{A}} - new\_min_{\mathcal{A}}) + new\_min_{\mathcal{A}}$$

Si le salaire varie de 11000 à 11122, la valeur 11100 normalisée entre  $[\emptyset - 1]$  est :

**(11100-11000)/(11122-11000)\*(1-0)+0**  
est transformée 0.81



126

---

---

---

---

---

---

---

## Normalisation

Nom	Age	Salaire
Clara	50	11000
Marie	70	11100
Léa	60	11122
Lucy	60	11074

## Min-max normalisation

Nom	Age	Salaire
Clara	50	0
Marie	70	0,81967
Léa	60	1
Lucy	60	0,60656



127

---

---

---

---

---

---

---

---

---

## Normalisation

- Z-score

$$V' = v - m_A / s_A$$

Où  $m_A$  est la moyenne pour l'attribut A et  $s$  l'écart type pour l'attribut A  
Négatif quand V est en dessous de la moyenne positif autrement

- Alternative : calculer l'écart moyen absolu

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

$$V' = v - m_A / s_f$$

- L'écart moyen absolu est plus robuste que l'écart type notamment en cas d'outliers



128



## Normalisation

- #### • Mise à l'échelle décimale

$$v' = \frac{v}{10^j}$$

<b>Nom</b>	<b>Age</b>	<b>Salaire</b>
Clara	50	110
Marie	70	111
Léa	60	111,22
Lucy	60	110,74



129



## Normalisation

- Z-score normalisation

Nom	Age	Salaire
Clara	-2	-0,5
Marie	2	0,18
Lea	0	0,32
Lucy	0	0

$$d(Clara, Marie) = 4,67$$

$$d(Clara, Léa) = 2,34$$

Clara est plus proche de Léa !

Mean<sub>age</sub>=60 Sage = 5  
Mean<sub>Salairé</sub> = 11074 S<sub>Salairé</sub> = 48



130



## Attributs tous continu

- Echelles différentes :
    - Il y a des attributs dominants
    - Il faut normaliser avant de calculer des distances
    - Tout ramener entre 0 et 1
  - On peut vouloir garder la dissymétrie entre attributs
    - Donner un poids à chaque attribut
    - Calculer la distance en fonction de ce poids

$$\sqrt{w_1(x_1 - y_1)^2 + \cdots + w_n(x_n - y_n)^2}$$

 – Nécessite une très bonne connaissance du domaine !

131

## Attributs binaires

- Une table de contingence pour données binaires

Objet <i>j</i>		<i>sum</i>	a : nombre de positions où $i = 1$ et $j = 1$
1			
Objet <i>i</i>	1	$a$	$b$
	0	$c$	$d$
<i>sum</i>		$a+c$	$b+d$
			$p$
			b : nombre de positions où $i = 1$ et $j = 0$
			c : nombre de positions où $i = 0$ et $j = 1$
			d : nombre de positions où $i = 0$ et $j = 0$

- Exemple  $oi = (1, 1, 0, 1, 0)$  et  $oj = (1, 0, 0, 0, 1)$  :
    - $a=1$ ,  $b=2$ ,  $c=1$ ,  $d=2$



132

## Mesures de distances

- Coefficient d'appariement (matching) simple (invariant pour variables symétriques) :

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

– Ex. pour  $oi=(1,1,0,1,0)$  et  $oj=(1,0,0,0,1)$   $d(oi, oj)=3/5$

- Coefficient de Jaccard

$$d(i,j) = \frac{b+c}{a+b+c}$$

-  $d(oi, oj) = 3/4$



133

## Rappel codage binaire

- Attribut symétrique :**
  - le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 est similaire au codage inverse
- Attribut asymétrique :**
  - Test médical. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre
  - Généralement, on code par 1 la modalité la moins fréquente
    - 2 personnes ayant la valeur 1 pour le test sont plus similaires que 2 personnes ayant 0 pour le test



134

---

---

---

---

---

---

---

## Mesures de distances

Nom	Sexe	Fièvre	Tousse	Test-1	Test-2	Test-3	Test-4
Jacques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- Sexe est un attribut symétrique. Les autres (fièvre, test-1...) sont asymétriques
  - O et P = 1, N = 0. La distance n'est mesurée que sur les asymétriques
- $$d(jacques, marie) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$
- $$d(jacques, jean) = \frac{1 + 1}{1 + 1 + 1} = 0.66$$
- $$d(jean, marie) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$



$$d(i, j) = \frac{b + c}{a + b + c}$$

a : nombre de positions où i a 1 et j a 1  
b : nombre de positions où i a 1 et j a 0,  
c : nombre de positions où i a 0 et j a 1

135

---

---

---

---

---

---

---

## Attributs nominaux

- Une généralisation des attributs binaires, ex: rouge, vert et bleu
- Méthode 1: Matching simple
  - m : nombre d'appariements, p : nombre total de variables

$$d(i, j) = \frac{p - m}{p}$$



136

---

---

---

---

---

---

---

## Attributs ordinaux

- Un attribut ordinal peut être discret ou continu
- L'ordre peut être important (e.g. froid < tiède < chaud)
- Peuvent être traitées comme les variables intervalles
  - remplacer  $x_{if}$  par son rang  $r_{if} \in \{1, \dots, M_f\}$
  - Remplacer le rang de chaque variable par une valeur dans  $[0, 1]$  en remplaçant la variable  $f$  dans l'objet  $I$  par

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- Utiliser une distance pour calculer la similarité

Froid => 1-1/3-1=0  
 Tiède => 2-1/3-1 = 0.5  
 Chaud => 3-1/3-1 = 1

137

## Attributs chaînes

- La distance de Hamming entre deux chaînes de mêmes longueurs est le nombre de positions où les symboles correspondants sont différents
    - En d'autres termes, elle mesure le nombre minimum de substitutions nécessaires pour changer une chaîne en une autre
- 1011101 et 1001001 = 2  
 "Bonjour" et "Bnojour" = 2
- Utilisé en télécommunications, en bioinformatique, en text mining

138

## Attributs mixtes

- Les objets peuvent être décrits avec tous les types de données
  - binaire symétrique, binaire asymétrique, nominale, ordinale, ...
- Utilisation d'une formule pondérée pour combiner leurs effets

$$d(i, j) = \frac{\sum_{k=1}^p w_k d_k(i, j)}{\sum_{k=1}^p w_k}$$

139



## Attributs mixtes

Nom	Age	Prop	Mensualité
Jean	30	1	1000
Pierre	40	0	2200
Paul	45	1	4000

- $d(x,y) = \sqrt{(10/15)^2 + 1^2 + (1200/3000)^2} = 1.27$
  - $d(x,z) = \sqrt{(15/15)^2 + 0^2 + (3000/3000)^2} = 1.41$
  - $d(y,z) = \sqrt{(5/15)^2 + 1^2 + (1800/3000)^2} = 1.21$

Le voisin le plus proche de Jean est Pierre

  - Distance normalisée et sommation  $d(x,y) = d_1(x,y) + d_2(x,y) ..$



140

---

---

---

---

---

---

---

---

---

---

Plan

- Concrètement ?
  - Pourquoi fouiller les données ?
  - Le processus d'extraction
  - Un aperçu de quelques techniques



141

---

---

---

---

---

---

---

## Les tâches du DM

- Les principales tâches
    - Classification (prédictive)
    - Groupement/segmentation (*clustering*) (descriptive)
    - Recherche de règles d'association (descriptive)
    - Recherche de motifs (descriptive)
    - Régression (prédictive)
    - Détection d'anomalies (prédictive)
  - Pour chacune des tâches n méthodes



142

---

---

---

---

---

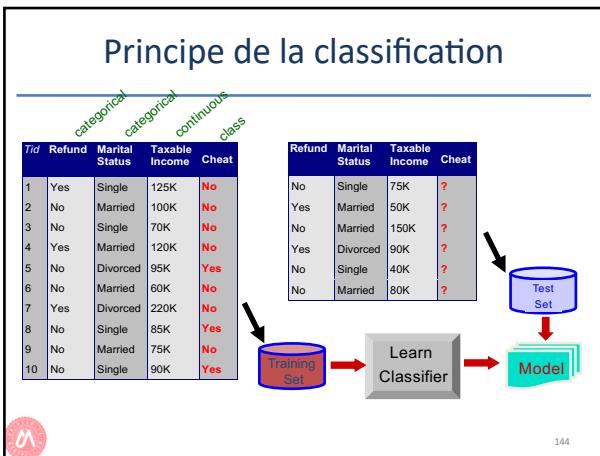
---

---

## Classification

- Soit une collection d'enregistrements (**ensemble d'apprentissage**)
    - Chaque enregistrement contient un ensemble d'attributs, l'un de ces attributs est la **classe**.
  - Rechercher un **modèle** pour l'attribut classe comme une fonction des valeurs des autres attributs
  - But : Affecter de la meilleure manière possible les enregistrement non vues dans la classe.
    - Un **jeu de test** est utilisé pour déterminer l'efficacité du modèle. Généralement le jeu de données est divisé en jeu d'entraînement et en jeu de test. Le jeu d'entraînement est utilisé pour apprendre le modèle et le jeu de test pour valider le modèle.

143



144

## Classification - Exemples

- Marketing direct
    - But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achèteront vraisemblablement un nouveau téléphone portable
    - Fonctionnement :
  - Utiliser des données pour un produit similaire.
    - On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l'attribut classe
    - Collecter diverses informations sur ce type de consommateurs
    - Cette information représente les entrées du classifier.

145

## Le mailing

- Classification... un exemple d 'utilisation

- un cadeau est envoyé par mailing. Un envoi sans réponse coûte 50 € et une réponse assure 100 €.
- Pas d 'envoi de mailing à un client qui aurait répondu : perte de 100 €.



146

---

---

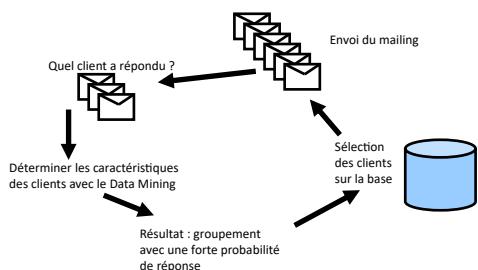
---

---

---

---

## Le mailing



147

---

---

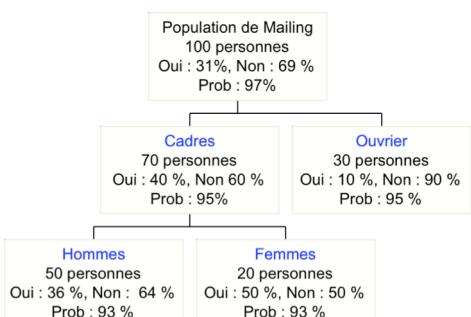
---

---

---

---

## Résultat du mailing



148

---

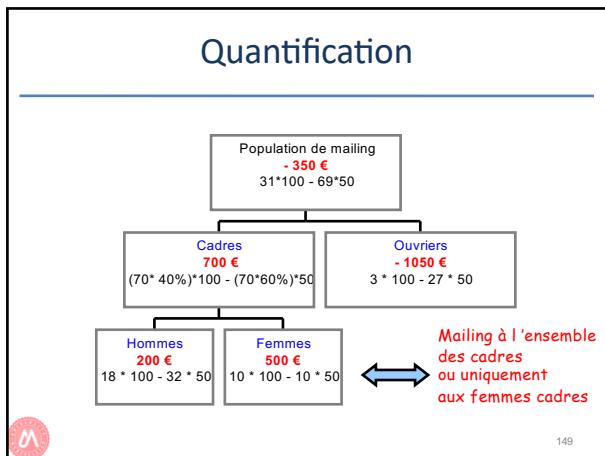
---

---

---

---

---



## Evaluation

Matrice de coûts

Prédit	OBSERVE			TOTAL
	Payé	Retardé	Impayé	
Payé	80	15	5	100
Retardé	1	17	2	20
Impayé	5	2	23	30
TOTAL	86	34	30	150

Validité du modèle : nombre de cas exacts (=somme de la diagonale) divisé par le nombre total :  $120/150 = 0.8$

150

- ## Segmentation(Clustering)
- Soit un ensemble d'objets composés d'un ensemble d'attributs, et une mesure de similarité entre eux, rechercher des clusters tels que :
    - Les objets dans un cluster sont les plus similaires les un des autres
    - Les objets dans des clusters séparés sont les moins similaires entre eux
  - Mesures de similarités :
    - La distance Euclidienne si les attributs sont continus
    - D'autres mesures spécifiques au problème
- 151

## Segmentation(Clustering)

- Soit un ensemble d'objets composés d'un ensemble d'attributs, et une mesure de similarité entre eux, rechercher des clusters tels que :
  - Les objets dans un cluster sont les plus similaires les uns des autres
  - Les objets dans des clusters séparés sont les moins similaires entre eux
- Mesures de similarités :
  - La distance Euclidienne si les attributs sont continus
  - D'autres mesures spécifiques au problème



152

---

---

---

---

---

---

---

## Découverte de règles d'association

- Etant donné un ensemble d'enregistrements qui contiennent des éléments d'une collection
- Générer des règles de dépendance qui prédisent les occurrences d'éléments suivant les occurrences des autres

TID	Items
1	Pain, Coca, Lait
2	Bière, Pain
3	Bière, Coca, Couches, Lait
4	Bière, Pain, Couches, Lait
5	Coca, Couches, Lait

Règles découvertes:  
 {Lait}  $\rightarrow$  {Coca}  
 {Couche, Lait}  $\rightarrow$  {Bière}




---

---

---

---

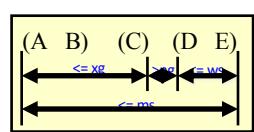
---

---

---

## Découverte de motifs séquentiels

- Étant donné un ensemble d'objets, dans lequel chaque objet est associé à une séquence temporelle, trouver des dépendances séquentielles entre les événements  
 $(A \ B) \ (C) \ (D \ E)$
- Des contraintes temporelles peuvent être prises en compte




---

---

---

---

---

---

---

## Recherche de motifs fréquents

- Analyse des associations
    - Panier de la ménagère, cross marketing, conception de catalogue, analyse de textes
    - Corrélation ou analyse de causalité
  - Clustering et Classification
    - Classification basée sur les associations
  - Analyse de séquences
    - Web Mining, détection de tendances, analyses ADN
    - Périodicité partielle, associations temporelles/cycliques



155

---

---

---

---

---

---

---

---

---

---

## Régression

- Prédire la valeur d'une variable connue en utilisant la valeur d'autres variables en supposant une relation linéaire ou non entre elles
  - Très utilisé en statistiques
  - Exemples:
    - Prédire la quantité de ventes d'un nouveau produit en fonction du budget de publicité
    - Prédire la force du vent en fonction de la température, humidité, pression ...
    - Prédire le cours de la bourse



1

---

---

---

---

---

---

---

---

---

---

## Conclusions

- Ces différents éléments vont être vus dans la suite des cours
  - Objectifs : être capable de mettre en œuvre les algorithmes, les interpréter et les évaluer
  - Dans ce cours : on considère que les données sont propres – peu de prétraitements.  
L’importance c’est de bien comprendre les fonctionnements et surtout les interpréter



157

---

---

---

---

---

---

---

## Conclusions

- Fouille de données de très nombreuses perspectives
  - Données de plus en plus hétérogènes
  - Données de plus en plus volumineuses
  - Données de plus en plus rapides
- Attention à ne pas oublier le pourquoi
  - Savoir bien classer est utile
  - Savoir pourquoi un objet a été mis dans une classe est très utile !

158

---



---



---



---



---



---



---



---



---

## Conclusions

- Des questions de droits :
  - Est ce que j'ai le droit d'utiliser les données ?
  - Est ce que je suis propriétaire des données et des connaissances obtenues ?
  - Est ce que mes connaissances préservent la vie privée ?
- Des questions éthiques ...

159

---



---



---



---



---



---



---



---



---

## Conclusions

- Logiciels
  - Il en existe de nombreux !! Sas, Intelligence Miner, Mineset, ...
  - Beaucoup en open source
  - Weka : utile pour une analyse rapide
  - Scikit-learn : machine learning Python
  - R vs Scikit-learn
- Ne pas oublier le plus important est de comprendre les algorithmes et de savoir comment les utiliser
  - K-means est un algorithme de clustering -> il est disponible sur de nombreuses plateformes

160

---



---



---



---



---



---



---



---



---

## Conclusions

- Quelques pointeurs importants :
  - Kdnuggets = une source d'information sur tout ce qui se fait autour de la fouille : tutoriels, stages, jeux de données, offre d'emploi, news ([www.kdnuggets.com](http://www.kdnuggets.com))
  - Des cours en ligne (e.g. coursera)
  - De très nombreuses ressources ([towardsdatascience.com](http://towardsdatascience.com), [medium.com](http://medium.com), ...)
  - Des associations et listes de diffusions : EGC (Extraction et Gestion de connaissances) ([egc.assoc.fr](http://egc.assoc.fr)), AFIA (Association Française pour l'IA) ([afia.asso.fr](http://afia.asso.fr))
  - De très nombreuses conférences : KDD, ICDM, PKDD, SDM, PAKDD, ...



161

---

---

---

---

---

---

---

- Des questions ?



162

---

---

---

---

---

---

---