



Fouille de données

▷ Introduction

Philippe Lenca et Romain Billot

`philippe.lenca@telecom-bretagne.eu`

Telecom Bretagne
2016-2017





Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours
- 5 Bibliographie



Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours
- 5 Bibliographie



Quelques définitions ...

- un processus d'aide à la décision où les utilisateurs cherchent des modèles d'interprétation dans les données [PCKW89]
- processus complexe permettant l'identification, au sein des données, de motifs valides, nouveaux, potentiellement intéressants et les plus compréhensibles possible [FPSSU96]
- processus interactif et itératif d'analyse d'un grand ensemble de données brutes afin d'en extraire des connaissances exploitables par l'utilisateur-analyste qui y joue un rôle central [KNZ01]



Fouille de données

Introduction

Exemples

Processus

Conclusion

Bibliographie

Processus ECD (1995)

- cycle complet de découverte (des données brutes ... à l'exploitation des modèles, des connaissances)
- récupération, intégration, nettoyage, validation, etc., des données

↪ Grand intérêt industriel.

Fouille de données (1990) - (Data Mining)

- étape de découverte de modèle(s)

↪ Intérêt plutôt académique.



L'apprentissage à partir d'exemples s'est développé bien avant ...

- statistique
- reconnaissance des formes
- intelligence artificielle
- apprentissage

↪ Cependant, beaucoup de choses ont changées ...



Métaphore

Introduction

Exemples

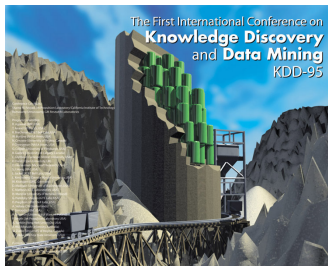
Processus

Conclusion

Bibliographie

Recherche de pépites d'or dans un gisement de pierres :

- informations inconnues, cachées, utiles (orientées *métier*)
- (*sans hypothèse*)
 - volume important de données (individus)
 - nombre important de variables (dimensions)
 - données distribuées, hétérogènes, incomplètes, imprécises, etc.





Des données (de production) à l'utilisateur

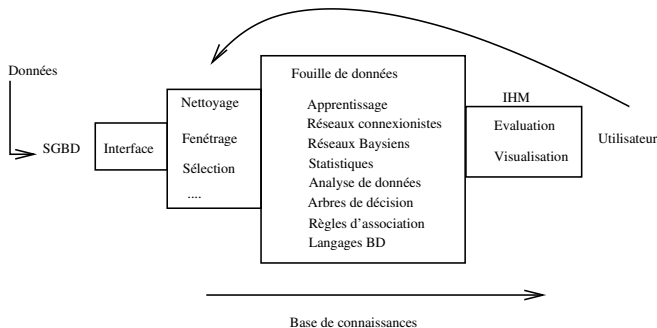
Introduction

Exemples

Processus

Conclusion

Bibliographie



- données réelles telles que fournies par l'utilisateur
- connaissances utiles, intelligibles

↪ Préparation des données, ... , évaluation des connaissances.



L'ECD extrait une connaissance

Introduction

Exemples

Processus

Conclusion

Bibliographie

Connaissances, modèles, etc., issues des données :

- règles
- auparavant inconnue
- phénomènes
- exceptions
- tendances
- etc.

↔ Nécessité de techniques adaptées (hétérogénéité, volumétrie, etc.).



L'ECD extrait une connaissance

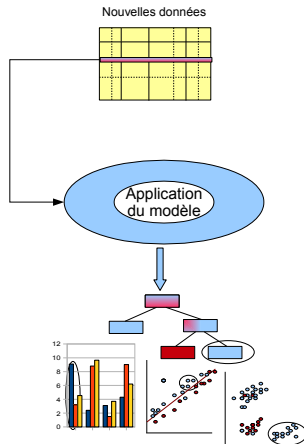
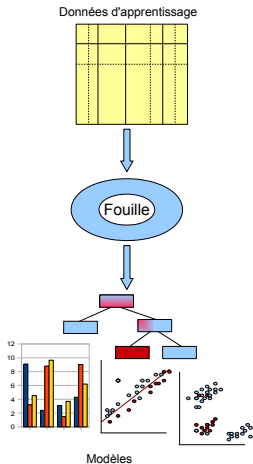
Introduction

Exemples

Processus

Conclusion

Bibliographie





Apprendre ?

Introduction

Exemples

Processus

Conclusion

Bibliographie

- apprentissage supervisé
- apprentissage non-supervisé
- apprentissage semi-supervisé
- apprentissage partiellement supervisé
- apprentissage par renforcement
- etc.



Apprendre ?

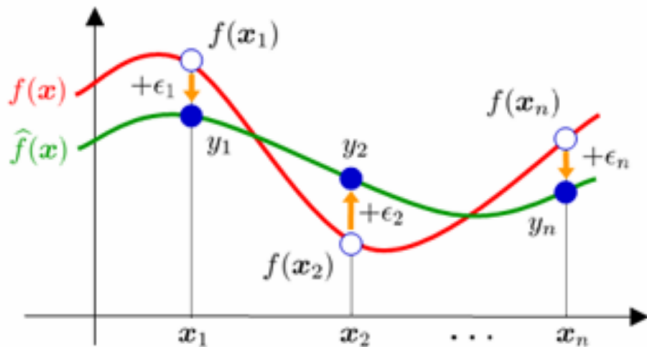
Introduction

Exemples

Processus

Conclusion

Bibliographie





Un modèle [HTF09]

Introduction

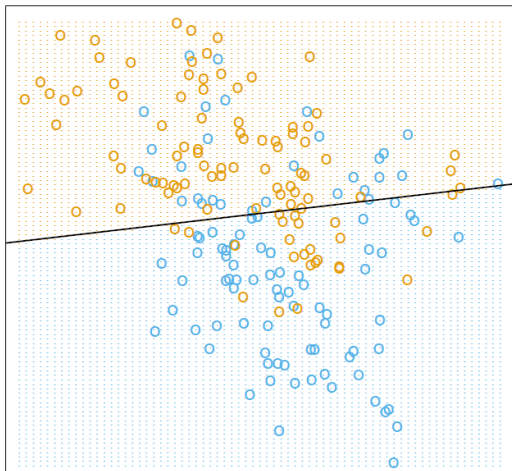
Exemples

Processus

Conclusion

Bibliographie

Linear Regression of 0/1 Response





Un second [HTF09]

Introduction

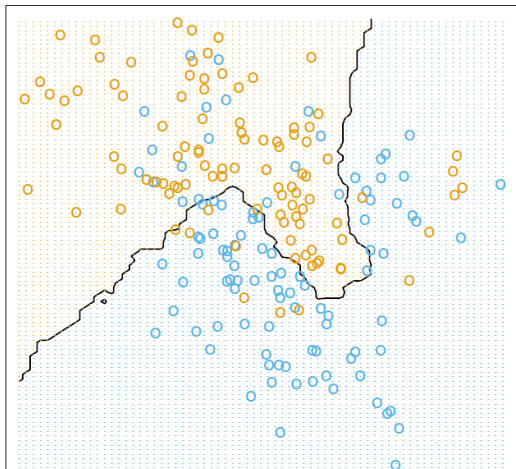
Exemples

Processus

Conclusion

Bibliographie

15-Nearest Neighbor Classifier





Et un troisième [HTF09]

Introduction

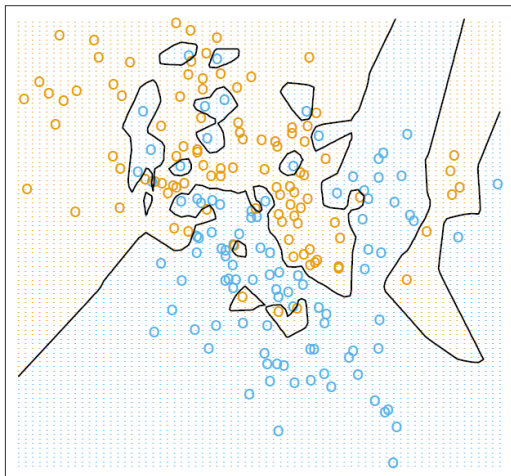
Exemples

Processus

Conclusion

Bibliographie

1-Nearest Neighbor Classifier





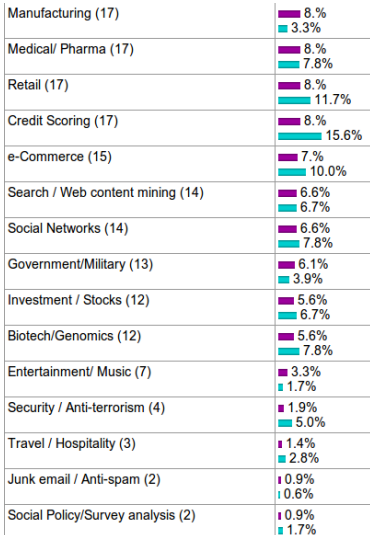
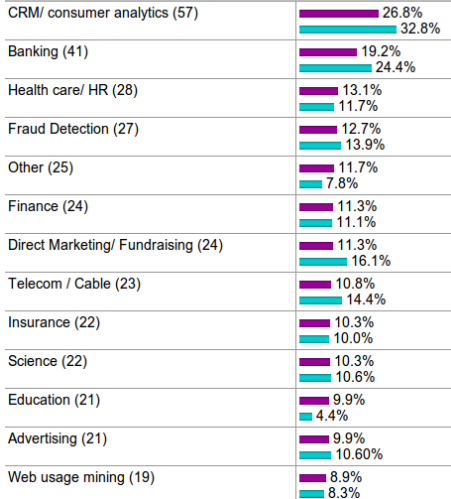
Utilisation de l'information

- avant : utilisation pour contrôle, comptabilité et fiscalité
- désormais : nouvelle ressource, intelligence pour les entreprises, les administrations et la découverte scientifique
 - marketing
 - ressources humaines
 - observation de la concurrence
 - médecine
 - qualité
 - télécom, banque, finance, assurance
 - etc.

↪ L'information doit être utilisée pour une aide à la décision.

Industries / Fields where you applied Data Mining in 2010? [213 voters]

■ 2010 % of voters ■ 2009 % of voters



<http://www.kdnuggets.com/polls/2010/analytics-data-mining-industries-applications.html>



Exigences de l'environnement

Introduction

Exemples

Processus

Conclusion

Bibliographie

Facteur économique

- concurrence croissante, temps de réaction de plus en plus court
- nécessité de gagner en productivité
- nécessité de mesurer de façon fiable, à tout moment, force et faiblesse

↪ Obtenir des informations “décisionnelles” utiles pour les défis à venir, faire évoluer les compétences humaines ...



Constat : richesses inexploitées

Introduction

Exemples

Processus

Conclusion

Bibliographie

Contexte favorable

Développement des Systèmes d'informations décisionnels :

- entrepôts de données - datawarehouse
- reporting
- visualisation graphique
- développements logiciels
- etc.

Croissance exponentielle des d'informations stockées :

- coût élevé
- ... mais peu exploitée (croissance linéaire de son utilisation)

↔ Exploiter ce patrimoine.



Un peu d'histoire ...

Introduction

Exemples

Processus

Conclusion

Bibliographie





Un peu d'histoire ...

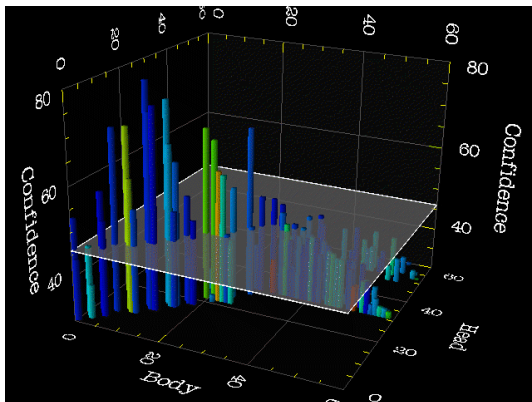
Introduction

Exemples

Processus

Conclusion

Bibliographie



Un peu d'histoire ...

Introduction

Exemples

Processus

Conclusion

Bibliographie

Produits fréquemment achetés ensemble

- ☒ Cet article : Introduction to Algorithms de Thomas H. Cormen Broché EUR 45,65
- ☒ The Art of Computer Programming de Donald E. Knuth - 3 tois EUR 134,83
- ☒ The Art of Computer Programming : Vol 4, Fascicules 0-4 de Donald E. Knuth Broché EUR 64,63

Les clients ayant acheté cet article ont également acheté

Produit	Prix
The Art of Computer Programming de Donald E. Knuth	EUR 134,83
Programming Pearls de Joe Bentley	EUR 30,94
The Algorithms Design Manual de Steve S. Skiena	EUR 41,70
Design patterns: Elements of reusable object orient... de Erich Gamma	EUR 47,98
The Art of Computer Programming : Vol 4, Fascicules... de Donald E. Knuth	EUR 64,63
Computer Organization and Design: The Hardw... de David A. Patterson	EUR 50,30

La recherche de co-occurrences d'événements binaires ou transactionnelles a de très nombreuses applications. L'espace de recherche est en 2^k ce qui nécessite des algorithmes efficaces.

	L1	L2	L3	...	Lk
1	1	1	1		
2	1	1	1		
3	1		1	1	
4	1				1
...	1	1	1		
i	1			1	
j			1		1
k	1				
...		1			
n	1	1	1		



Un peu d'histoire . . . récente

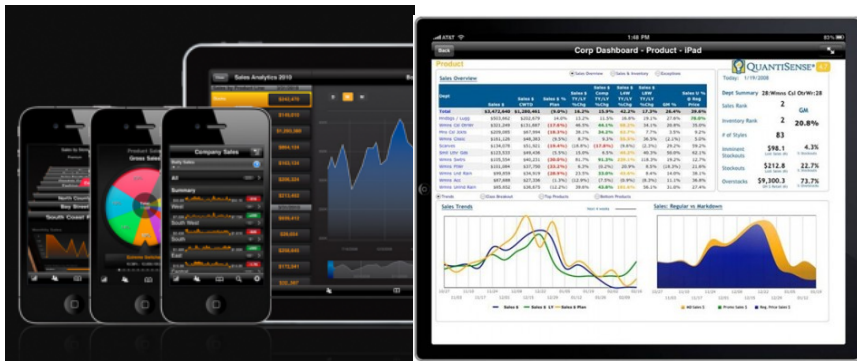
Introduction

Exemples

Processus

Conclusion

Bibliographie





Un peu d'histoire ... récente

Introduction

Exemples

Processus

Conclusion

Bibliographie



SAP prévoit de connecter ses logiciels d'analyse à Google Maps et Earth pour permettre aux utilisateurs d'agrèger des données économiques et sociologiques à des points géographiques à travers le monde, ont annoncé les deux sociétés. Par exemple, une banque pourrait recouper des informations statistiques sur les régions où le marché immobilier est en crise avec la saisie des données sur les demandes de prêt bancaire.

...

Un partenariat signé entre Google et SAP pourrait fournir une alternative aux traditionnels systèmes SIG.

<http://www.lemondeinformatique.fr/> (28/07/2011)



Constat : richesses inexploitées

Introduction

Exemples

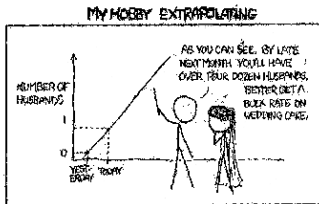
Processus

Conclusion

Bibliographie

Exploiter ce patrimoine de l'entreprise

- comment ?
- dans quel but ? ... pour des systèmes d'aide à la décision



↪ Modèles actionnables pour des décisions justes et rapides.



De la nécessité d'algorithmes efficaces

Introduction

Exemples

Processus

Conclusion

Bibliographie

Même si au cours des dernières décennies des progrès très importants ont été réalisés, tant du côté algorithmique que du côté de la puissance des langages et des machines, on est confronté à toujours plus de données, à toujours plus de besoins, etc :

- YAHOO ! : *25 terabytes of data collected each day ... and growing* (Sources : Mediamark Research, Spring 2004 and comScore Media Metrix, February 2005) ;
- données météorologiques, astronomiques, etc.

↪ Ne pas compter uniquement sur l'augmentation des capacités matérielles ... mais surtout les méthodes d'analyse, sur l'innovation algorithmique, etc.



Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours
- 5 Bibliographie



Exemples - Banque

Introduction

Exemples

Processus

Conclusion

Bibliographie

Recherche de formes caractéristiques d'une fraude :

- à la carte (au milieu de milliers de transactions)

Prédiction :

- des clients qui vont partir
- des clients qui vont augmenter leurs avoirs

Décision :

- en matière de crédit (analyse du risque client)
- des autorisations en crédit-revolving

Aide à l'arbitrage :

- basé sur analyse de formes historiques des cours, sur les valeurs passées



Exemples - Assurance

Introduction

Exemples

Processus

Conclusion

Bibliographie

- modèles de sélection et de tarification
- analyse des sinistres
- recherche des critères explicatifs
 - du risque
 - de fraude
- prévision d'appels sur les plates-formes d'assurance directe



Exemples - ...

Introduction

Exemples

Processus

Conclusion

Bibliographie

- médecine
- production
- télécoms
- droit
- banque, finance, assurance
- etc.

Exemple de modèles :

- analyses prévisionnelles (trafic en fonction de l'heure)
- scores (risque crédit, fidélité d'un client)
- classes (bon/mauvais client)
- règles (facture > 100 et réclamation $> 0.6 \Rightarrow$ départ)



Exploratoire vs. confirmatoire

- Fouille de données devrait être exploratoire
 - chercher tout azimut
 - sans préjugés
- Statistiques sont confirmatoires
 - vérifier une hypothèse
 - vérifier une intuition
- Fouille de données plus une extension qu'une révolution ?

↪ Et les données massives. . .



Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours
- 5 Bibliographie



Un processus en plusieurs étapes

Introduction

Exemples

Processus

Conclusion

Bibliographie

- poser le problème (et compréhension du domaine d'application)
- créer la base d'apprentissage (rechercher les données, sélectionner les données pertinentes)
- (pré)traiter les données brutes (nettoyer les données, réduire les données)
- (pré)traiter les variables (transformer les variables, réduire les variables, créer des variable)
- rechercher le(s) modèle(s) (définir les tâches, choisir les algorithmes appropriés)
- évaluer les résultats (définir les moyens d'évaluer, comparer les modèles, interpréter les résultats)
- intégrer la connaissance (valider les connaissances extraites)



Un processus en plusieurs étapes

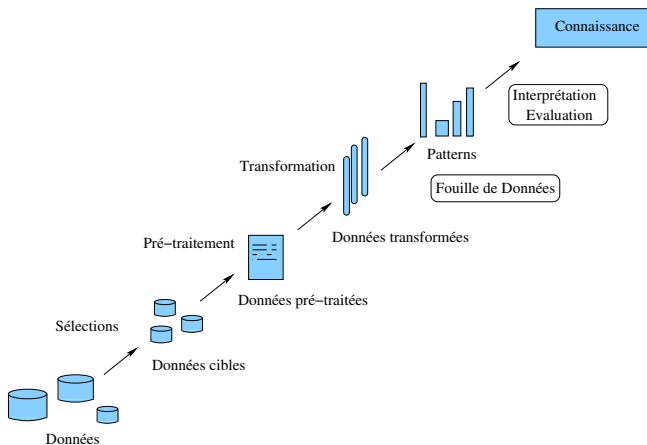
Introduction

Exemples

Processus

Conclusion

Bibliographie





Un processus en plusieurs étapes

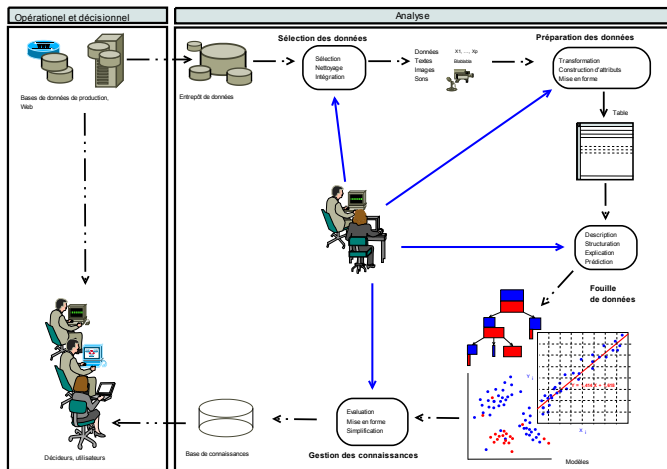
Introduction

Exemples

Processus

Conclusion

Bibliographie





Etape de fouille : exemple de possibilités

Introduction

Exemples

Processus

Conclusion

Bibliographie

Fouille de données

- le Raisonnement à Base de Cas
- les Agents Intelligents
- les Règles d'Association
- les Arbres de Décision
- les Algorithmes Génétiques
- les Réseaux Bayésiens
- les Réseaux Connexionnistes
- les Outils de Visualisation
- l'Analyse de données
- etc.



Compromis entre les techniques

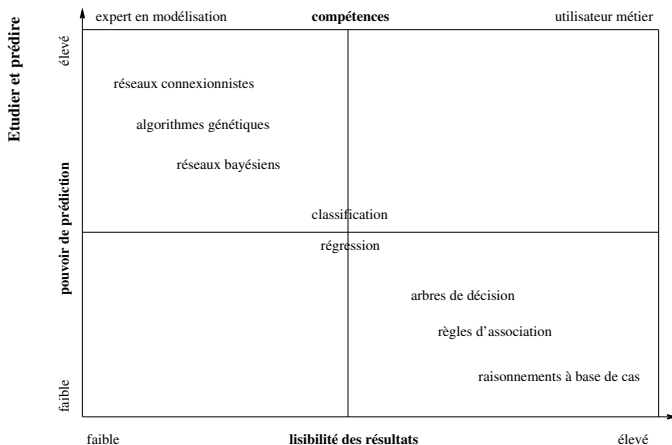
Introduction

Exemples

Processus

Conclusion

Bibliographie



Voir et résoudre



Compromis entre les techniques

Introduction

Exemples

Processus

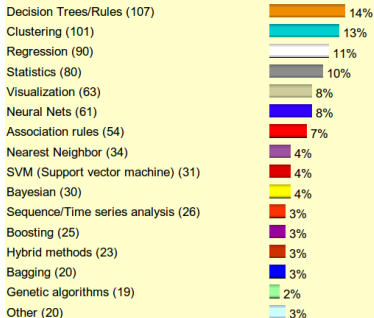
Conclusion

Bibliographie

Autres critères de différenciation :

- ?
- ?
- ...

Data mining/analytic techniques you use frequently: [784 votes total]

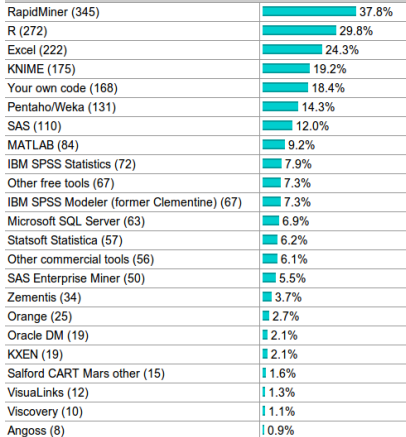


Algorithm	Usage
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %

Nov. 2011, www.kdnuggets.com/2011/11/algorithms-for-analytics-data-mining.html?
k11n27

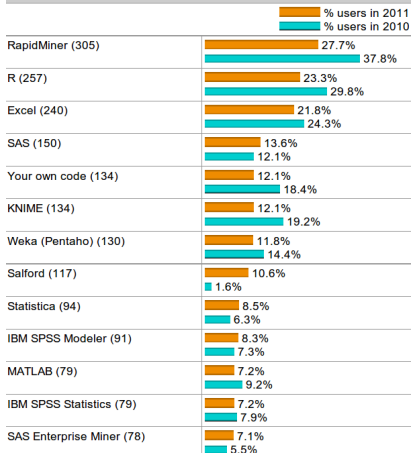
Feb. 2005, www.kdnuggets.com/polls/2005/data_mining_techniques.htm

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [912 voters]



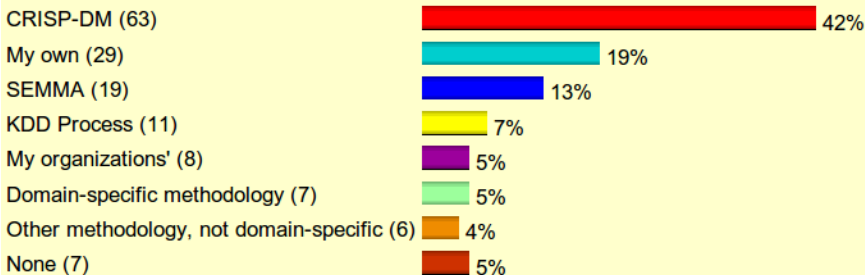
Feb. 2005, www.kdnuggets.com/polls/2005/data_mining_techniques.htm

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [1103 voters]



Nov. 2011, www.kdnuggets.com/2011/05/tools-used-analytics-data-mining.html

What main methodology are you using for data mining? [150 votes total]



Aug. 2007, [http:](http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm)

[//www.kdnuggets.com/polls/2007/data_mining_methodology.htm](http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm)



Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours**
- 5 Bibliographie



Exactitude vérifiable par référence au mode réel, à des catégories abstraites ?














- précises ou non (" Je crois que la température est de 37,2")
- incertaines ou non (" La température est entre 37 et 38")
- ambiguë ou non (" La température est élevée")

Caractéristiques :

- numérique, symbolique, complexe
- volumineuses ou non (dimensions, individus)
- coûteuses ou non (à acquérir, erreur)
- complètes ou non, redondantes ou non
- équilibrées ou non, symétriques ou non

↔ Bien connaître les données.

Types of Data Analyzed/Mined in the past 12 months [144 voters]

table data (fixed # of columns) (102)	 70.8%
time series (56)	 38.9%
itemsets / transactions (52)	 36.1%
text (free-form) (43)	 29.9%
anonymized data (38)	 26.4%
social network data (28)	 19.4%
other (22)	 15.3%
web content (19)	 13.2%
XML data (17)	 11.8%
web clickstream (15)	 10.4%
email (15)	 10.4%
images / video (11)	 7.6%
music / audio (3)	 2.1%

http:

[//www.kdnuggets.com/polls/2010/data-types-analyzed.html](http://www.kdnuggets.com/polls/2010/data-types-analyzed.html)



Caractéristiques :

- lisibles ou non (bon client/mauvais client ; 80% des clients qui achètent le produit *A* achètent aussi le produit *B*)
- couverture, complétude (% d'exemples couverts)
- exactitude (remise en cause possible, pas exacte à 100%)
- expertise (humaine) \Rightarrow formaliser la connaissance

\hookrightarrow Bien connaître les besoins, les limites des modèles.

Des données aux connaissances et à la décision

Introduction

Exemples

Processus

Conclusion

Bibliographie

Le chemin est long ...

- données *brutes*
- récupération, transformation
- alimentation
- modélisation
- recherche de connaissances
- aide à la décision

Des données aux connaissances et à la décision

Introduction

Exemples

Processus

Conclusion

Bibliographie

Objectifs :

- Savoir mettre en œuvre une méthodologie
- Comprendre qualités et limites de plusieurs méthodes de fouille
- Savoir évaluer les performances d'un modèle
- Prendre en main un logiciel
- Mettre en œuvre un processus sur des données



L'ECD ... essentiellement de l'ingénierie.

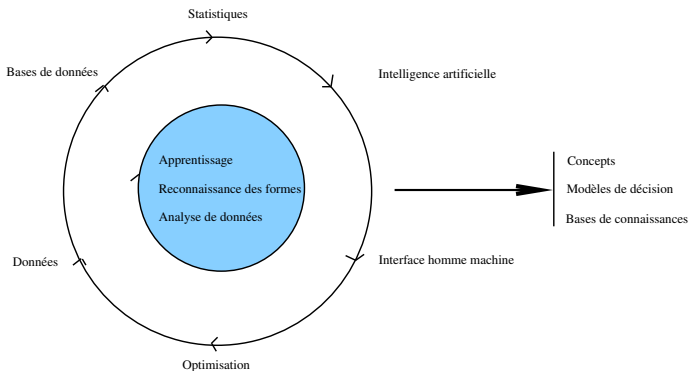
Introduction

Exemples

Processus

Conclusion

Bibliographie



↔ Ensemble de techniques combinées de façon opportuniste pour faire sortir des modèles, des connaissances, des concepts à partir des données ... avec de vraies problématiques de recherche.



Exemple fictif

Données type comportements financiers/Questions métiers

ID	Type	Epargne	Canal	Ordres	Sexe	Age	Salaire	CB	...
1111	J.	10	I.	10	H	35 – 40	40 – 59	T1	...
1012	I.	0	I.	2	H	25 – 30	20 – 29	T1	...
2315	I.	10	G.	25	F	30 – 39	60 – 79	NA	...
2165	I.	20	G.	25	F	25 – 30	60 – 79	T2	...
...

- puis-je déterminer des profils types de différentes catégories d'investisseurs ? (classification)
- puis-je déterminer le type de carte bancaire en fonction des comportements ? (prédiction)
- puis-je prévoir le montant moyen des ordres passés par mois pour un nouveau client ? (régression)



Plan

Introduction

Exemples

Processus

Conclusion

Bibliographie

- 1 Introduction
- 2 Exemples d'application
- 3 L'ECD, un processus en plusieurs étapes
- 4 Conclusion et plan du cours
- 5 Bibliographie**



Bibliographie

Introduction

Exemples

Processus

Conclusion

Bibliographie

- [FPSSU96] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors.
Advances in Knowledge Discovery and Data Mining.
AAAI/MIT Press, 1996.
- [HTF09] T. Hastie, R. Tibshirani, and J. Friedman.
The Elements of Statistical Learning : Data Mining, Inference, and Prediction.
Springer, 2009.
- [KNZ01] Y. Kodratoff, A. Napoli, and D. Zighed.
Bulletin de l'association française d'intelligence artificielle, extraction de connaissances dans des bases de données, 2001.
- [PCKW89] K Parsaye, M. Chignell, S. Khoshafian, and H. Wong.
Intelligent Databases; Object-Oriented, Deductive Hypermedia Technologies.
John Wiley & Sons, 1989.