

# Module Master Recherche Apprentissage et Fouille

Michele Sebag – Balazs Kegel  
<http://tao.lri.fr/tiki-index.php>

30 septembre 2009

# Pour quoi faire

- ▶ Prédiction  
pannes, maladies, achats, préférences,...
- ▶ Compréhension, Modélisation  
facteurs de risque, analyse de survie  
e-Science
- ▶ Interaction  
Jeux ; “Super-Google” ;  
Brain Computer Interface
- ▶ Optimisation—Conception  
apprendre pour décider



**Machine Learning – Apprentissage Artificiel**

**Data Mining – Fouille de Données**

# Le contexte international

L'idéal le siècle des connaissances

La réalité des connaissances pointues et morcelées  
des spécialistes, un dialogue difficile

Le besoin la gestion humaine des connaissances  
ne passe pas à l'échelle

L'opportunité les données sont accessibles  
les connaissances à l'état de traces dans les données

## OBJECTIF

fournir [à l'expert]  
des connaissances nouvelles, utiles, valides

# Le contexte, Master Informatique Paris-Sud

## Autres modules ayant un rapport

- ▶ Traitement statistique de l'information
- ▶ Représentation des connaissances
- ▶ Algorithmes d'évolution et robotique

## Ce module

- ▶ Théorie
- ▶ Etudes de cas
- ▶ Présentations d'articles 10 mn / volontaires

## Examen

- ▶ Questions de cours
- ▶ Au choix:
  - ▶ Exposé (présentation orale + résumé) d'un article
  - ▶ Projet (algos, données, comparaison, rapport)

# Plan du Module

## Etudes de cas – domaines

1. Skicat arbres de décision
2. Reconnaissance de caractères apprentissage d'ensembles, boosting
3. Autonomic Computing apprentissage non supervisé streaming
4. Expérience Auger approches génératives
5. Netflix filtrage collaboratif
6. MoGo apprentissage en-ligne
7. Robotique réduction de dimension apprentissage par renforcement
8. Satisfaction de contraintes apprentissage relationnel

## Quelques bonnes adresses

- ▶ Où sont les cours :  
<http://tao.lri.fr/tiki-index.php?page=Courses>
- ▶ Module Traitement Statistique de l'Information  
<http://www.limsi.fr/Individu/allauzen/wiki/index.php/TSI07>
- ▶ Les cours (transparents) d'Andrew Moore  
<http://www.autonlab.org/tutorials/index.html>
- ▶ Les cours (videos) de PASCAL  
<http://videlectures.net/pascal/>
- ▶ Les tutoriels de NIPS Neuro Information Processing Systems  
<http://nips.cc/Conferences/2006/Media/>
- ▶ Des questions intéressantes  
<http://hunch.net/>

# Plan de ce cours

1. Généralités
2. Classification
3. Etude de cas : SKICAT, Stanley

# Partie 1. Généralités

1. Objectifs: Apprentissage supervisé, non supervisé, fouille de données
2. Quelques définitions
3. Domaines d'applications
4. Quelles sont les difficultés ?
5. Méthodologie



# Apprentissage supervisé

## Contexte

Monde  $\rightarrow$  instance  $\mathbf{x}_i \rightarrow$  Oracle  
 $\downarrow$   
 $y_i$



**Input:** Base d'apprentissage  $\mathcal{E} = \{(\mathbf{x}_i, y_i), i = 1 \dots n, \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

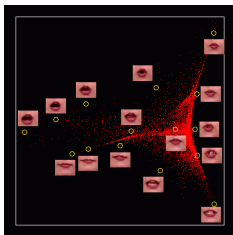
**Output:** Hypothèse  $h : \mathcal{X} \mapsto \mathcal{Y}$

**Critère:** pas d'erreur (precisions bientôt)

# Apprentissage non supervisé

## Base d'apprentissage $\mathcal{E}$

- ▶  $\mathcal{E} = \{x_i, x_i \in \mathcal{X}, i = 1 \dots n\}$
- ▶ Objectif : partitionner les  $x_i$ ; définir des clusters
- ▶ Rôle : analyse de données exploratoire



# Fouille de données

Des spécifications...

Fayyad et al. 1996

Automatic extraction of  
novel, useful and valid knowledge  
from large sets of data.

des connaissances

{ nouvelles  
utiles  
valides

...imprécises...

% au sens commun  
pour qui  
un pb multi-critères

# Extraction de régularités

## Exemple

- ▶ Base de données
  - Tickets de caisse dans un supermarché
  - Dossiers clients d'une compagnie d'assurances
- ▶ Itemsets fréquents seuil de fréquence
  - $I = \{ \textit{Vendredi}, \textit{bière}, \textit{couches} \}$
  - $I = \{ \textit{Pain}, \textit{beurre}, \textit{confiture} \}$
- ▶ Règles d'association seuil de confiance
  - $\textit{Vendredi}, \textit{couches} \Rightarrow \textit{bière}$
  - $\textit{Pain}, \textit{beurre} \Rightarrow \textit{Confiture}$

# Quelques définitions

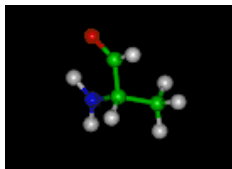
## Exemple

- ▶ ligne : exemple/  
cas/individus/transactions
- ▶ colonne : attribut/  
feature/variables/  
items
- ▶ (optionnel) : attribut  
classe

age	employe	education	edur	marital	...	job	relation	race	gender	hour	country	wealth
39	State_gov	Bachelors	13	Never_mar...	...	Adm_clerks	Not_in_fan	White	Male	40	United_States	poor
51	Self_emp	Bachelors	13	Married	...	Exec_mari	Husband	White	Male	13	United_States	poor
39	Private	HS_grad	9	Divorced	...	Handlers_c	Not_in_fan	White	Male	40	United_States	poor
54	Private	11th	7	Married	...	Handlers_c	Husband	Black	Male	40	United_States	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_mari	Wife	White	Female	40	United_States	poor
50	Private	9th	5	Married_sp...	...	Other_ser	Not_in_fan	Black	Female	16	Jamaica	poor
52	Self_emp	HS_grad	9	Married	...	Exec_mari	Husband	White	Male	45	United_States	rich
31	Private	Masters	14	Never_mar...	...	Prof_speci	Not_in_fan	White	Female	50	United_States	rich
42	Private	Bachelors	13	Married	...	Exec_mari	Husband	White	Male	40	United_States	rich
37	Private	Some_coll	10	Married	...	Exec_mari	Husband	Black	Male	80	United_States	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar...	...	Adm_clerks	Own_child	White	Female	30	United_States	poor
33	Private	Assoc_acc	12	Never_mar...	...	Sales	Not_in_fan	Black	Male	50	United_States	poor
41	Private	Assoc_voc	11	Married	...	Craft_repa	Husband	Asian	Male	40	MissingV	rich
34	Private	7th_8th	4	Married	...	Transport	Husband	Amer_Indi	Male	45	Mexico	poor
26	Self_emp	HS_grad	9	Never_mar...	...	Farming_fi	Own_child	White	Male	35	United_States	poor
33	Private	HS_grad	9	Never_mar...	...	Machine_c	Unmarried	White	Male	40	United_States	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_States	poor
44	Self_emp	Masters	14	Divorced	...	Exec_mari	Unmarried	White	Female	45	United_States	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_States	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

## Espace des instances $\mathcal{X}$

- ▶ Propositionnel :  
 $\mathcal{X} \equiv \mathbb{R}^d$
- ▶ Relationnel : ex.  
chimie.



molécule alanine

# Partie 1. Généralités

1. Objectifs: Apprentissage supervisé, non supervisé, fouille de données
2. Quelques définitions
3. Domaines d'applications
4. Quelles sont les difficultés ?
5. Méthodologie

# Domaines d'application

## Domaine

## But : Modélisation

### Phénomènes physiques

### modélisation et contrôle de process

Applications industrielles, sciences expérimentales, calcul numérique

### Phénomènes sociaux

### + confidentialité

Hôpitaux, Assurances, Banques, ...

### Phénomènes individuels

### + dynamique rapide

*Consumer Relationship Management*

*User Modelling*

PASCAL : <http://www.pascal-network.org>

PASCAL 2: <http://pascallin2.ecs.soton.ac.uk/>

# Banques, Telecom, Vente

Ex: KDD 2009 – Orange

1. Churn
2. Appetency
3. Up-selling

Objectifs

1. Pub plus efficace
2. Moins de fraude
3. Moins de risque





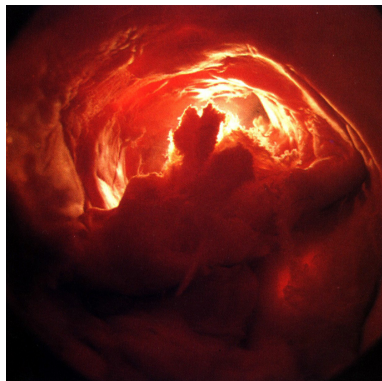
# Médecine, Bio-Informatique, Sécurité

## Ex: Facteurs de risque

1. Maladies cardio-vasculaires
2. Molécules cancérigènes
3. Gènes de l'obésité...

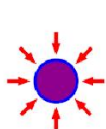
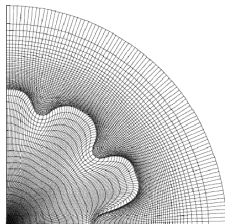
## Objectifs

1. Diagnostic, prévention
2. Médecine personnalisée
3. Identification

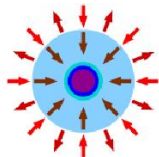


## Numerical Engineering

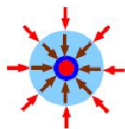
- ▶ De gros codes de calcul
- ▶ Chers en temps calcul
- ▶ Chers en expertise



Laser heating



DT compression



Hot spot ignition



Thermonuclear burn

Fusion par confinement inertiel, ICF

# e-Science, Conception (2)

## Buts

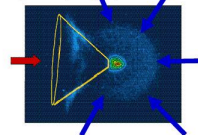
- ▶ Un résultat approché
- ▶ Pour une fraction du temps de calcul
- ▶ Raccourcir le cycle de conception
- ▶ Conception optimale

*More is Different*

Alternative scheme : spherical target with a gold cone\*



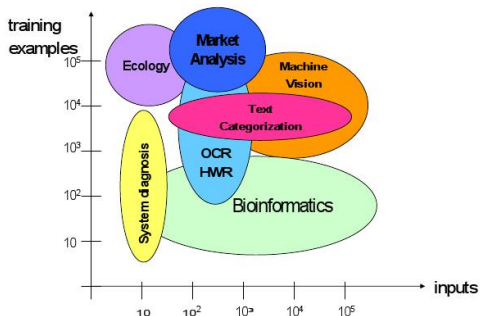
Short pulse



\* Kodama et al. Nature 412 798 (2001); 418 933 (2002);

# Données / Applications

- ▶ Données propositionnelles 80% des applis.
- ▶ Données spatiales, temporelles alarmes, gisements, accidents
- ▶ Données relationnelles chimie, biologie
- ▶ Données semi-structurées texte, Web
- ▶ Données multi-media images, sons, films,..



# Partie 1. Généralités

1. Objectifs: Apprentissage supervisé, non supervisé, fouille de données
2. Quelques définitions
3. Domaines d'applications
4. Quelles sont les difficultés ?
5. Méthodologie

# Sources de difficulté

## Qualité des données / de la représentation

- Bruit ; données manquantes
- + Attributs pertinents Feature extraction
- Données structurées : spatio-temporelles, relationnelles, textes, videos ..

## Distribution des données

- + Exemples indépendants, identiquement distribués
- Autres cas: robotique; flots de données; données hétérogènes...

## Connaissances a priori

- + Critères d'intérêt
- + Contraintes sur la solution

## Sources de difficulté (2)

### Critère d'apprentissage

- + Fonction convexe : un seul optimum
- ↘ Complexité :  $n$ ,  $n \log n$ ,  $n^2$  Passage à l'échelle
- Optimisation combinatoire

H. Simon, 1958:

*In complex real-world situations, optimization becomes approximate optimization since the description of the real-world is radically simplified until reduced to a degree of complication that the decision maker can handle.*

*Satisficing seeks simplification in a somewhat different direction, retaining more of the detail of the real-world situation, but settling for a satisfactory, rather than approximate-best, decision.*

# Critères, suite

## Critères de l'utilisateur

- ▶ Pertinence, Causalité
- ▶ INTELLIGIBILITE
- ▶ Simplicité
- ▶ Stabilité
- ▶ Interactivité, rapidité, visualisation
- ▶ ... Apprentissage de préférences



## Sources de difficulté (3)

### Crossing the chasm

- ▶ Pas de *killer algorithm*
- ▶ Peu de recommandations a priori

### Critères de performance d'un algorithme

- ▶ Consistance

Quand le nombre d'exemples tend vers l'infini  
et que le concept cible  $h^*$  est dans  $\mathcal{H}$   
l'algorithme le trouve.

$$\lim_{n \rightarrow \infty} h_n = h^*$$

- ▶ Vitesse de convergence

$$\|h^* - h_n\| = \mathcal{O}(1/n), \mathcal{O}(1/\sqrt{n}), \mathcal{O}(1/\ln n)$$

# Contexte

## Disciplines et critères

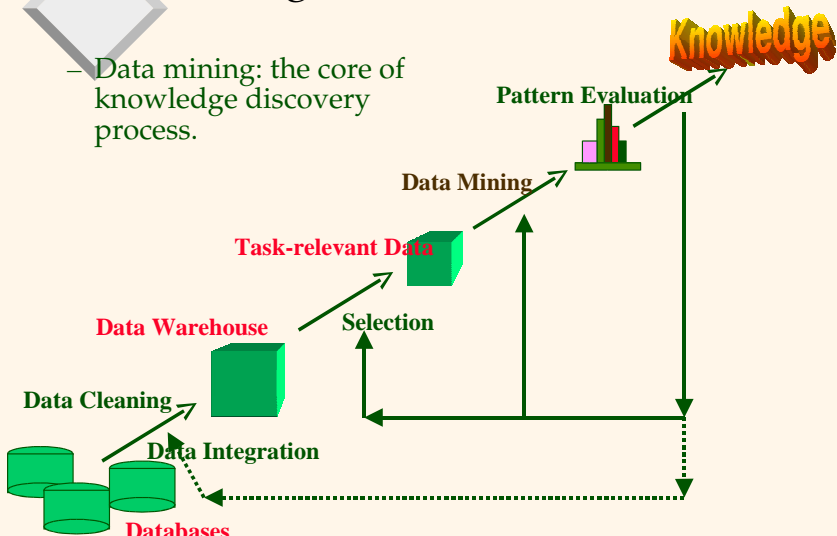
- ▶ Bases/Fouille de Données  
Passage à l'échelle ; au plus près des données
- ▶ Statistiques et analyse de données  
Modèles prédéfinis ; évaluation
- ▶ Apprentissage artificiel  
Connaissances du domaine ; représentations complexes
- ▶ Optimisation  
problèmes bien ou mal posés
- ▶ Interface Homme Machine  
Pas de solution finale : un dialogue
- ▶ Calcul hautes performances  
Données réparties, confidentialité

# Partie 1. Généralités

1. Objectifs: Apprentissage supervisé, non supervisé, fouille de données
2. Quelques définitions
3. Domaines d'applications
4. Quelles sont les difficultés ?
5. **Méthodologie**

# Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



# Méthodologie

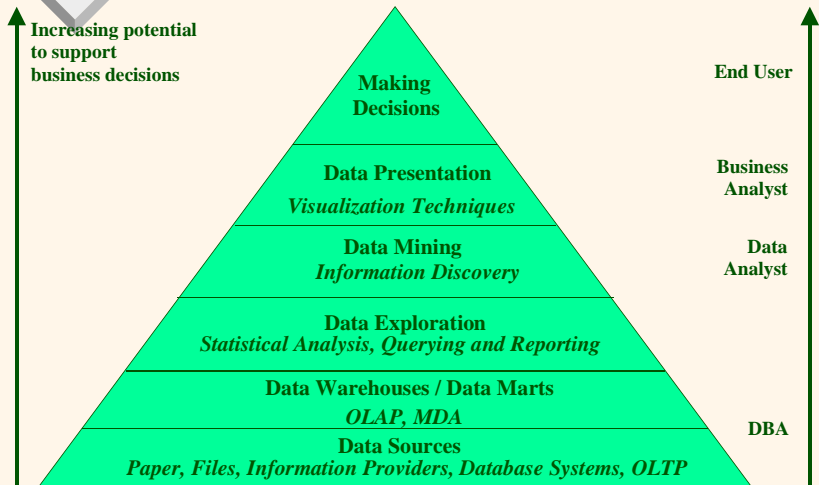
## Les étapes

- |                                   |                                     |
|-----------------------------------|-------------------------------------|
| 1. Collecte des données           | expert, DB                          |
| 2. Nettoyage                      | stat, expert                        |
| 3. Sélection                      | stat, expert                        |
| 4. Fouille / Apprentissage        |                                     |
| ▶ Description                     | <i>Qu'y a-t-il ds les données ?</i> |
| ▶ Prédiction                      | <i>Décider sur un cas</i>           |
| ▶ Agrégation                      | <i>Prendre une décision globale</i> |
| 5. Visualisation                  | chm                                 |
| 6. Evaluation                     | stat, chm                           |
| 7. Recherche de nouvelles données | expert, stat                        |

## Un processus itératif en fonction

des attentes, des données initiales, et des connaissances a priori.

# Data Mining and Business Intelligence



## Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Arbres de décision
4. Etude de cas: SKICAT
5. Etude de cas: Stanley

# Apprentissage supervisé

## Contexte

Monde  $\rightarrow$  instance  $\mathbf{x}_i \rightarrow$  Oracle  
 $\downarrow$   
 $y_i$



**Input:** Base d'apprentissage  $\mathcal{E} = \{(\mathbf{x}_i, y_i), i = 1 \dots n, \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$

**Output:** Hypothèse  $h : \mathcal{X} \mapsto \mathcal{Y}$

Confiance: Estimation  $E[h(x) \neq y]$

**Critère:** Minimiser  $E[\ell(h(x) \neq y)]$  *espérance du coût de l'erreur*



# Apprentissage supervisé, 2

## Définitions

- ▶  $\mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1 \dots n\}$ 
  - ▶ Classification :  $\mathcal{Y}$  fini (ex, nom de maladie)
  - ▶ Régression :  $\mathcal{Y} \subseteq \mathbb{R}$  (ex, durée de survie)
- ▶ Espace des hypothèses  $\mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}$

## Tâches

- ▶ Choisir  $\mathcal{H}$  sélection de modèle
- ▶ Evaluer  $h \in \mathcal{H}$   $score(h)$
- ▶ Choisir  $h^*$  dans  $\mathcal{H}$   $argmax score(h)$

# Apprentissage Statistique

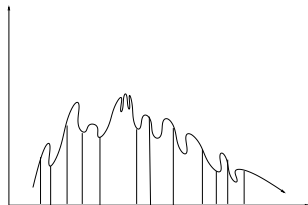
Minimiser l'espérance du coût de l'erreur

$$\text{Minimize } E[\ell(h(x), y)]$$

Principe

Si  $h$  "se comporte bien" sur  $\mathcal{E}$ , et  $h$  est "assez régulier",  $h$  se comporte bien en espérance.

$$E[F] \leq \frac{\sum_{i=1}^N F(x_i)}{n} + c(F, n)$$



# Classification, Problème posé

INPUT

$\sim P(x, y)$

$$\mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \{0, 1\}, i = 1 \dots n\}$$

ESPACE des HYPOTHESES

$$\mathcal{H} \quad h : \mathcal{X} \mapsto \{0, 1\}$$

FONCTION de PERTE

$$\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$$

OUTPUT

$$h^* = \arg \max \{ \text{score}(h), h \in \mathcal{H} \}$$

# Classification, critères

Erreur en généralisation

$$Err(h) = E[\ell(y, h(x))] = \int \ell(y, h(x)) dP(x, y)$$

Erreur empirique

$$Err_e(h) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i))$$

Borne

risque structurel

$$Err(h) < Err_e(h) + \mathcal{F}(n, d(\mathcal{H}))$$

$d(\mathcal{H})$  = dimension de VC de  $\mathcal{H}$ , voir après

# Dimension de Vapnik Cervonenkis

## Principe

Soit  $\mathcal{H}$  un ensemble d'hypothèses:  $\mathcal{X} \mapsto \{0, 1\}$

Soit  $x_1, \dots, x_n$  un ensemble de points de  $\mathcal{X}$ .

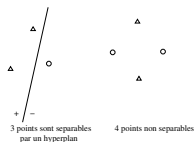
Si,  $\forall (y_i)_{i=1}^n \in \{0, 1\}^n, \exists h \in \mathcal{H} / h(x_i) = y_i,$

$\mathcal{H}$  pulvérise  $\{x_1, \dots, x_n\}$

Exemple:  $\mathcal{X} = \mathbb{R}^p$

$d(\text{hyperplans de } \mathbb{R}^p) = p + 1$

Rq: si  $\mathcal{H}$  pulvérise  $\mathcal{E}$ ,  $\mathcal{E}$  ne nous apprend rien...



## Définition

$$d(\mathcal{H}) = \max \{n / \exists (x_1 \dots, x_n) \text{ pulvérisé par } \mathcal{H}\}$$

# Classification, termes d'erreur

## Biais

Biais ( $\mathcal{H}$ ): erreur de la meilleure hypothèse  $h^*$  de  $\mathcal{H}$

## Variance

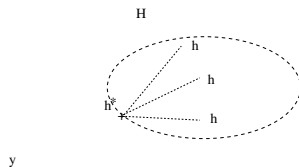
Variance de  $h_n$  en fonction de  $\mathcal{E}$

## Erreur d'optimisation

négligeable à la petite échelle

prend le dessus à la grande échelle

(Google)



BIAIS

VARIANCE

## Partie 2. Apprentissage Supervisé

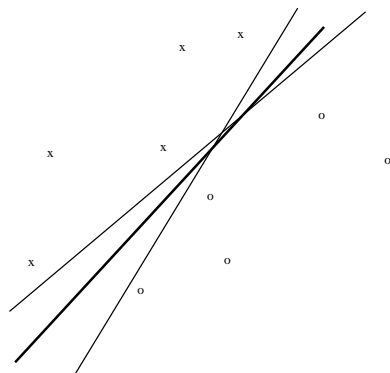
1. Poser le problème
2. Quelques espaces d'hypothèses
3. Arbres de décision
4. Etude de cas: SKICAT
5. Etude de cas: Stanley

# Espace d'hypothèses

## Hypothèses numériques

- ▶ Fonctions linéaires

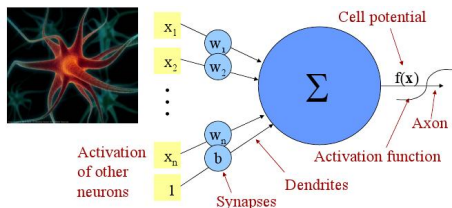
$$h(x) = 3x_1 + 2.17x_2 - 5x_3$$





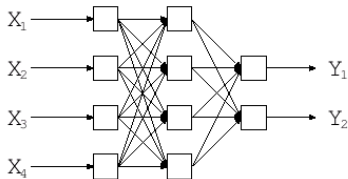
# Hypothèses numériques, 2

## Un neurone



## Un réseau neuronal

- ▶ Fonction d'activation  $f(X) = \frac{1}{1+e^{-X}}$
- ▶ Connexion des neurones



Muti-layer perceptron

# Hypothèses discrètes

## Formules booléennes

- ▶ Conjonctions

*Panne si  $\neg$ Essence*

*Malade si (Temperature > 39.5)*

- ▶ Liste de décision

$L_1 \wedge L_2 \dots$	Panne
$L'_1 \wedge L'_2 \dots$	non Panne
...	
default	non Panne

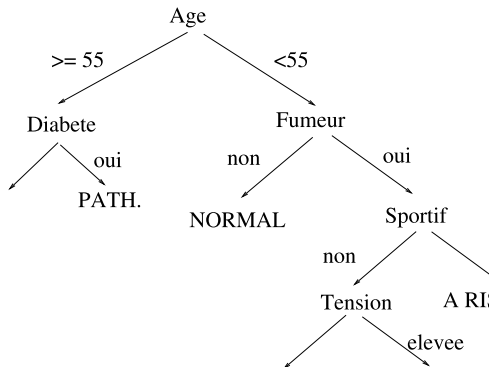
## Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Arbres de décision
4. Etude de cas: SKICAT
5. Etude de cas: Stanley

# Arbres de décision

## C4.5 (Quinlan 86)

- ▶ Parmi les algorithmes les plus utilisés
- ▶ Facile
  - ▶ à comprendre
  - ▶ à implémenter
  - ▶ à utiliser
  - ▶ et peu cher en temps calcul
- ▶ J48, Weka



# Arbres de décision, 2

## Principe

1. Pour  $\mathcal{E} = \{(x_i, y_i)_{i=1}^n, x_i \in \mathbb{R}^D, y_i \in \{0, 1\}\}$ 
  - Si  $\mathcal{E}$  monoclasse ( $\forall i, j, y_i = y_j$ ), stop
  - Si  $n$  trop petit, stop
  - Sinon, trouver l'attribut  $att$  le plus informatif
2. Pour toute valeur  $val$  de  $att$ 
  - Considérer  $\mathcal{E}_{val} = \mathcal{E} \cap [att = val]$ .
  - Goto 1.

## Critère gain d'information

$$\begin{aligned} p &= Pr(Class = 1 | att = val) \\ I([att = val]) &= -p \log p - (1 - p) \log (1 - p) \\ I(att) &= \sum_i Pr(att = val_i) I([att = val_i]) \end{aligned}$$

# Complexité

Quantité d'information d'un attribut

$$n \ln n$$

Pour construire un noeud

$$D \times n \ln n$$

# Arbres de décision, 3

## Table de contingence

wealth values:		poor	rich	
agegroup	10s	2507	3	
	20s	11262	743	
	30s	9468	3461	
	40s	6738	3986	
	50s	4110	2509	
	60s	2245	809	
	70s	668	147	
	80s	115	16	
	90s	42	13	

## Limitations

- ▶ Cas du XOR
- ▶ Attributs avec de nombreuses valeurs
- ▶ Attributs numériques
- ▶ Overfitting

# Limitations

## Attributs numériques

- ▶ Ordonner les valeurs  $val_1 < \dots < val_t$
- ▶ Calculer QI ( $[att < val_i]$ )
- ▶  $QI(att) = \max_i QI([att < val_i])$

## XOR

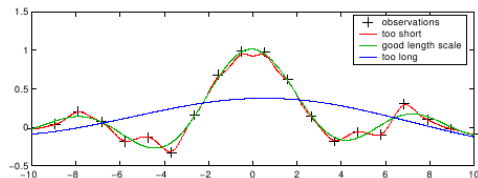
Biaiser la distribution des exemples



# Limitations, 2

## Overfitting

- ▶ Le but n'est pas de coller aux données d'apprentissage
- ▶ ... mais d'être bon en général
- ▶ Il faut ajuster le compromis empirique/généralité
- ▶ Comment : Validation croisée



# Validation croisée

## Principe

- ▶ Découper  $\mathcal{E}$  en 10 sous-ensembles  $\mathcal{E}_i$  stratifiés
- ▶  $\mathcal{E}^i = \mathcal{E} \setminus \mathcal{E}_i$
- ▶ Apprendre  $h_i(param)$  à partir de  $\mathcal{E}^i$
- ▶  $score_i(param)$  : Evaluer  $h_i(param)$  sur  $\mathcal{E}_i$
- ▶  $score(param) = \sum_{i=1}^{10} score_i(param)$

Retenir  $param^* = argmax\{score(param)\}$

## Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Arbres de décision
4. Etude de cas: SKICAT
5. Etude de cas: Stanley

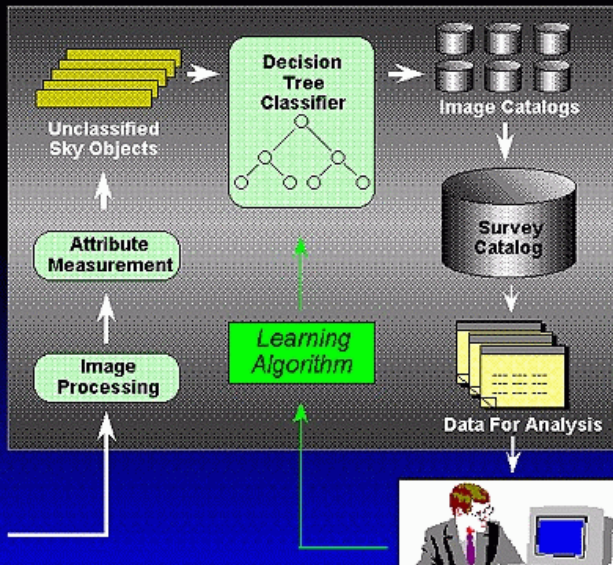
# Skicat

U. M. Fayyad, S. G. Djorgovski, and N. Weir. 1996  
Jet Propulsion Lab., Caltech

- ▶ Quel secteur du ciel regarder ?
- ▶ Térabytes de données
- ▶ Classification : étoiles, étoiles radiantes, galaxies, artefacts
- ▶ Arbres de décision
- ▶ Nb étoiles découvertes/nuit d'observation

Gain d'un facteur 40

# The SKICAT System



# SKICAT, 2

## Objectif final

catalogue du ciel

objets d'un ordre de grandeur moins brillants

Caltech, release 93

≈ 40,000 volumes

## Le problème

trop nombreux candidats : artefacts ?

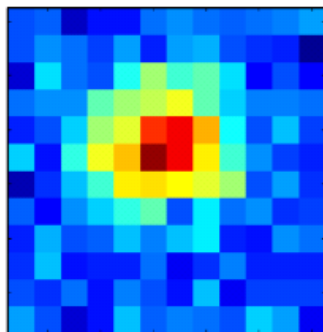
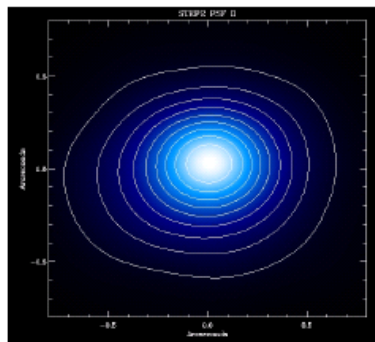
3 Téra bytes.

tera    Terrorbytes

# Skicat, 3

## L'opportunité

- ▶ photos à haute précision (longue mise au point)
- ▶ ... permettant l'étiquetage par les experts
- ▶ photos à basse précision
- ▶ ... inutilisable par les experts



# Skicat, 4

## Variabilité

entre géologues  
pour un même géologue

## Critère

non pas la vérité absolue  
des performances de même ordre





# Skicat, 5

## Mise en œuvre

apprentissage % photos de mise au point.  
pré-traitement

- ▶ brillance, surface, voisinage,...
- ▶ extraction d'un échantillon
- ▶ analyse en composantes principales
- ▶ recodage

Plus d'informations:

[http://www.astro.caltech.edu/~george/dposs/DPOSS\\_III.pdf](http://www.astro.caltech.edu/~george/dposs/DPOSS_III.pdf)

# Skicat, fin

## Impact

Automate a task  $\approx$  tens of man years.

Provide a consistent and objective means for a comprehensive analysis of a scientifically important data set.

## Achievements

94% classification accuracy

Classified objects: one magnitude fainter than previously

200% increase in size of data usable in analysis.

Exceeds human ability in classifying faint objects, solution achieved automatically using learning algorithms on astronomer-provided training data.

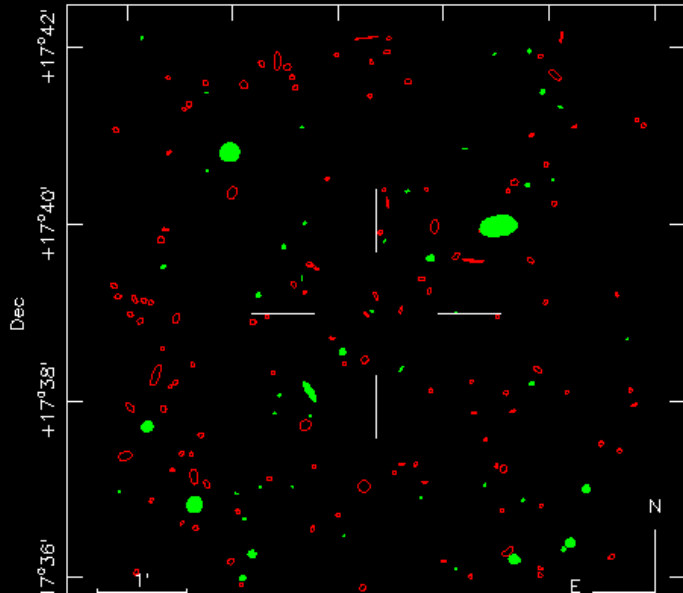
## Skicat, fin, 2

The classification rules produced by the inductive learning techniques form an objective, repeatable, examinable basis for classifying sky objects.

Since the automated approach allows us to classify faint sky objects that cannot be processed visually or by traditional computational techniques, the content of the catalog produced from the survey is increased by three-fold, since the majority of objects in each image are faint.

The training data for faint sky objects was obtained by examining a limited set of higher resolution CCD images covering minute portions of the survey. The learning algorithms are trained to predict the class (only obtainable by humans from higher resolution images) based on measurements from the survey. We thus classify objects that have to date not been classifiable by known techniques.

$1^{\text{h}}45^{\text{m}}13.2^{\text{s}}$   $+17^{\circ}38'60''$  (J2000)  
a247\_n54g



# A posteriori

Fayyad, nov. 2003

Personne ne connaissait les données mieux que les astronomes (30 ans).

Mais le concept (une fois résolu) fait intervenir  
8 variables/attributs parmi 40

## Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Arbres de décision
4. Etude de cas: SKICAT
5. Etude de cas: Stanley

# DARPA Challenge - 2005





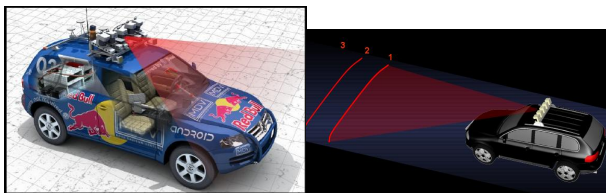
# Navigating: Perception is the issue

## The 2005 Darpa Challenge

### The terrain

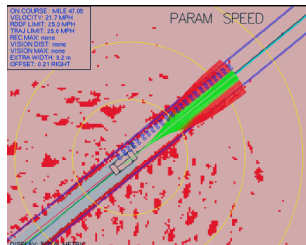
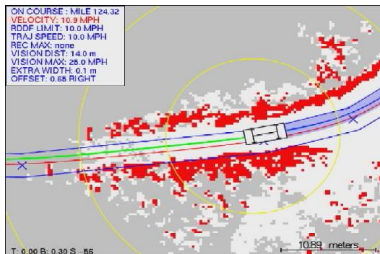


### The sensors



# Lifelong learning

Detection from high-definition, low-range camera: accurate



...used to label long-range sensor data

S. Thrun, Burgard and Fox 2005

<http://sss.stanford.edu/coverage/powerpoints/sss-thrun.ppt>

## Le but de la collecte

Les données sont-elles collectées pour l'analyse ?

NON

Usage dérivé, reformulation des données.

## Passage à l'échelle

efficacité quand les données ne tiennent pas en mémoire

données (même aléatoires) de grande taille

⇒ contiennent des motifs réguliers.

besoin de mesures de qualité et test d'hypothèses.