Module Master Recherche
Apprentissage et Fouille

Michele Sebag – Balazs Kegl – Antoine Cornuéjols
http://tao.lri.fr

17 décembre 2007

Pour quoi faire

- Prédiction
  pannes, maladies, achats, préférences,...
- Compréhension
  facteurs de risque, analyse de survie
- Interaction
  les jeux ; “Super-Google”
- 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 spécialistes, dialogue difficile le savoir-faire : des traces dans les données
Le besoin	la gestion humaine des connaissances ne passe pas à l'échelle
L'opportunité	les données sont accessibles

OBJECTIF

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

L'une des 10 technologies émergentes du 21^e siècle
(MIT Review, 2001)

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	reduction 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://videolectures.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. Introduction, définitions, objectifs
2. Méthodologie
3. Classification
4. Etude de cas : SKICAT

Introduction

Fouille de données, une définition... Fayyad et al. 1996

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

...imprécise...

des connaissances
- nouvelles — % au sens commun
- utiles — pour qui
- valides — un pb multi-critères

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

Définitions

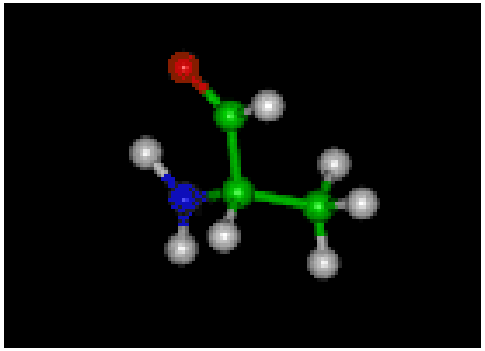
Exemple

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

age	employme	education	edun	marital	...	job	relation	race	gender	hour	country	wealth
					...							
39	State_gov	Bachelors	13	Never_mar	...	Adm_cleric	Not_in_fam	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	...	Handlers_c	Not_in_fam	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	...	Handlers_c	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	...	Other_serv	Not_in_fam	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	...	Prof_speci	Not_in_fam	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_cleric	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acd	12	Never_mar	...	Sales	Not_in_fam	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	...	Craft_repai	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_Sta	poor
33	Private	HS_grad	9	Never_mar	...	Machine_c	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

Espace des instances $\mathcal{X}$

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



molécule alanine

Apprentissage supervisé / non supervisé

Base d'apprentissage $\mathcal{E}$

Apprentissage supervisé

- $\mathcal{E} = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1 \dots n\}$
  - Classification : $\mathcal{Y}$ fini (ex, nom de maladie)
  - Régression : $\mathcal{Y} \subseteq \mathbb{R}$ (ex, durée de survie)
- Espace des hypothèses $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$
  - Choisir $\mathcal{H}$ — sélection de modèle
  - Evaluer $h \in \mathcal{H}$ — $score(h)$
  - Choisir h^* dans $\mathcal{H}$ — $argmax\ score(h)$

Apprentissage non supervisé

- $\mathcal{E} = \{x_i, x_i \in \mathcal{X}, i = 1 \dots n\}$
- Objectif : clustering (distance ou similarité)
- Idem : analyse de données exploratoire

Extraction de régularités

Base d'apprentissage $\mathcal{E}$ variables/items booléens

Extraction d'itemsets fréquents

- Input : seuil de fréquence ϵ
- Output : tous les sous-ensembles d'items $I = i_1, \ldots, i_t$ tq

$$\frac{|\{I \subseteq x_i\}|}{n} > \epsilon$$

Règles d'association

- Input : seuil de fréquence ϵ, seuil de confiance δ
- Output : toutes les règles $I \Rightarrow J$ tq

$$\frac{|\{I \subseteq x_i\}|}{n} > \epsilon \quad \frac{|\{(I \cup J) \subseteq x_i\}|}{|\{I \subseteq x_i\}|} > \delta$$

Extraction de régularités (2)

Exemple

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

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
- $\searrow$ Complexité : n, $nlogn$, 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.

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

- 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

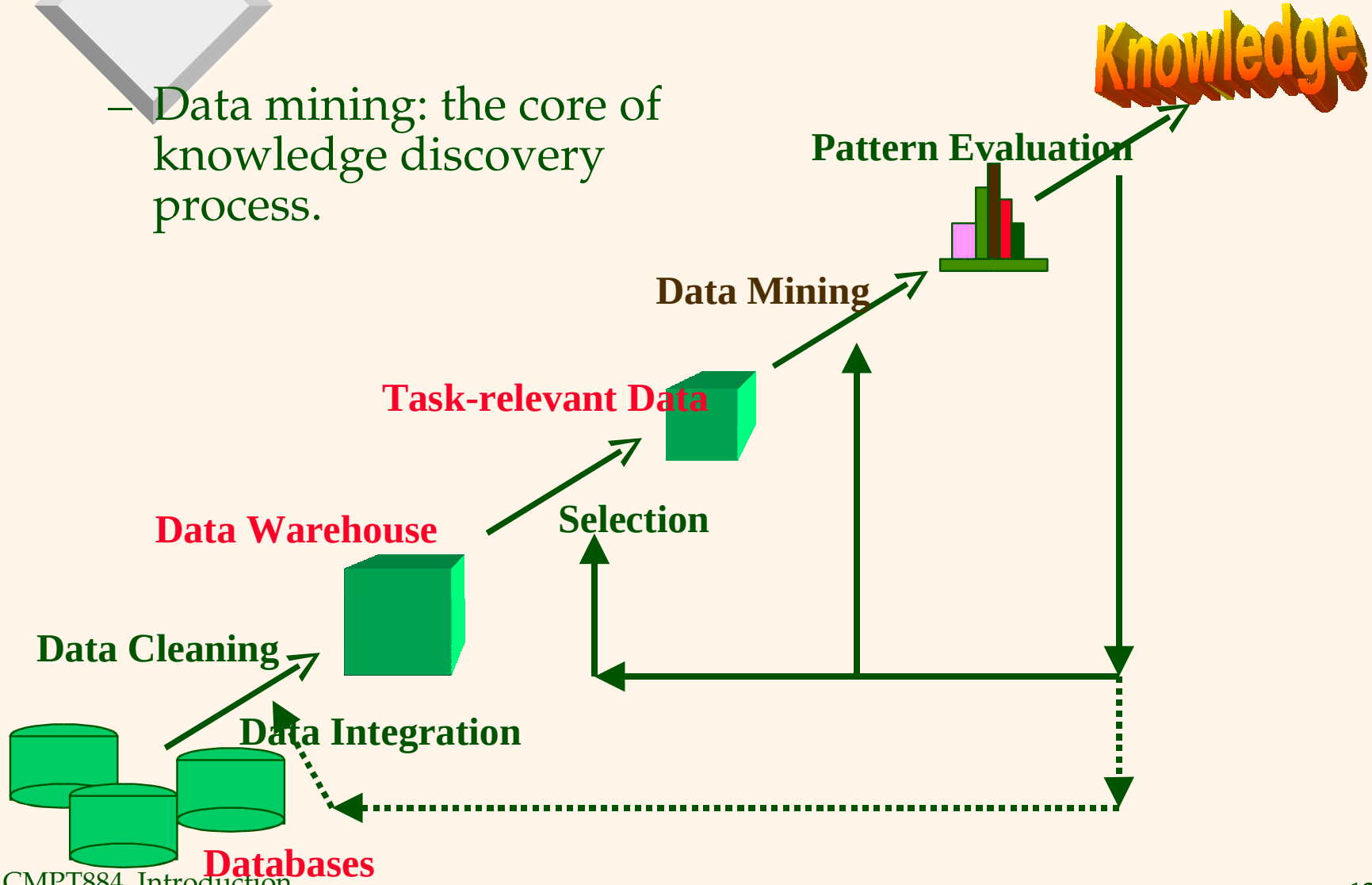
  Données réparties, confidentialité

Plan de ce cours

1. Introduction, définitions, objectifs
2. Méthodologie
3. Classification
4. Etude de cas : SKICAT

Data Mining: A KDD Process

– Data mining: the core of knowledge discovery process.



Méthodologie

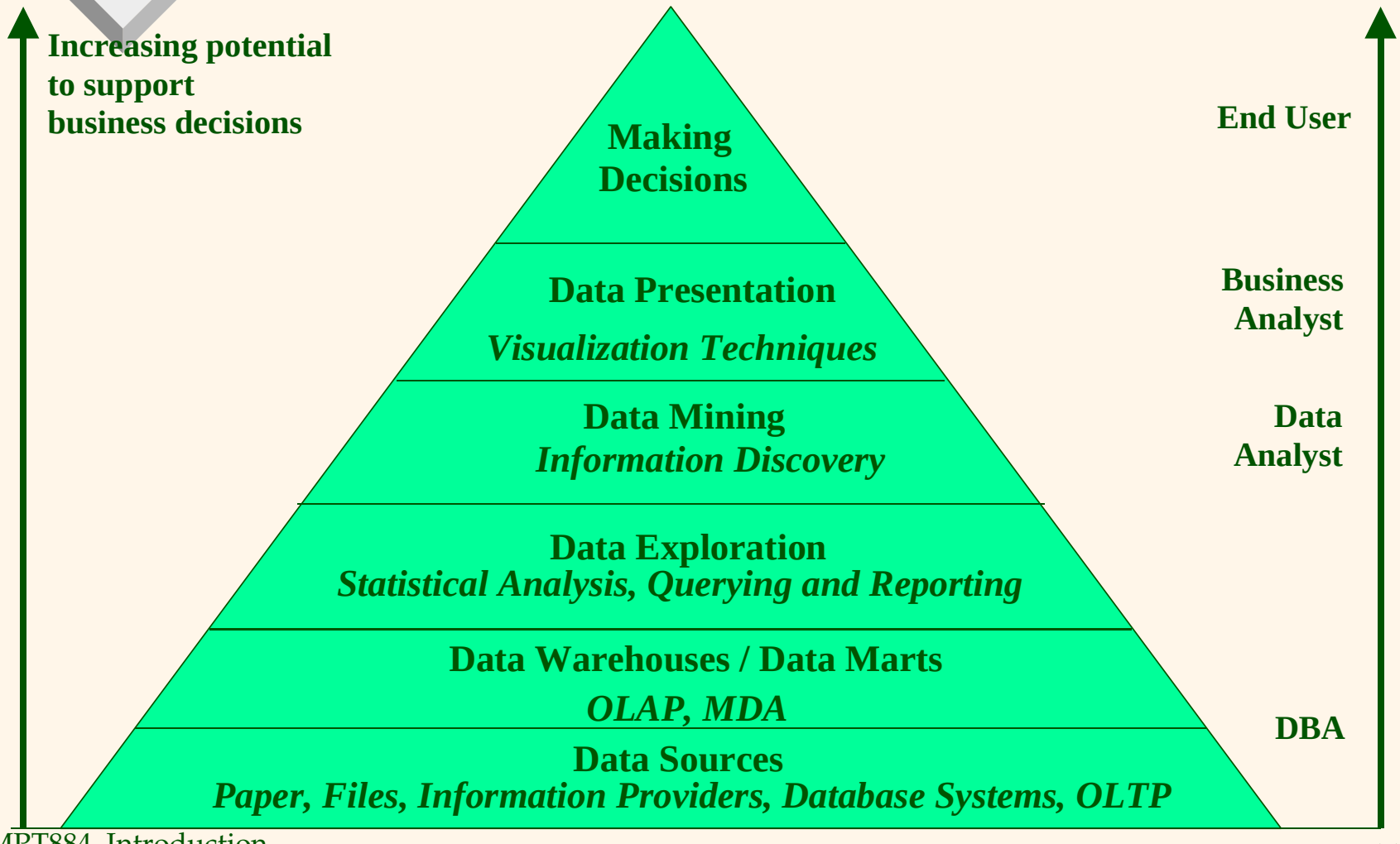
Les étapes

1. Collecte des données — expert, DB
2. Nettoyage — stat, expert
3. Sélection — stat, expert
4. Fouille / Apprentissage
   - Description — *Qu'y a-t-il ds les données ?*
   - Prédiction — *Décider sur un cas*
   - Agrégation — *Prendre une décision globale*
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



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,..

Plan de ce cours

1. Introduction, définitions, objectifs
2. Méthodologie
3. Classification
4. Etude de cas : SKICAT

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

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

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 $x_1, \ldots, x_n$ 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 $\{x_1, \ldots, x_n\}$

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

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

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

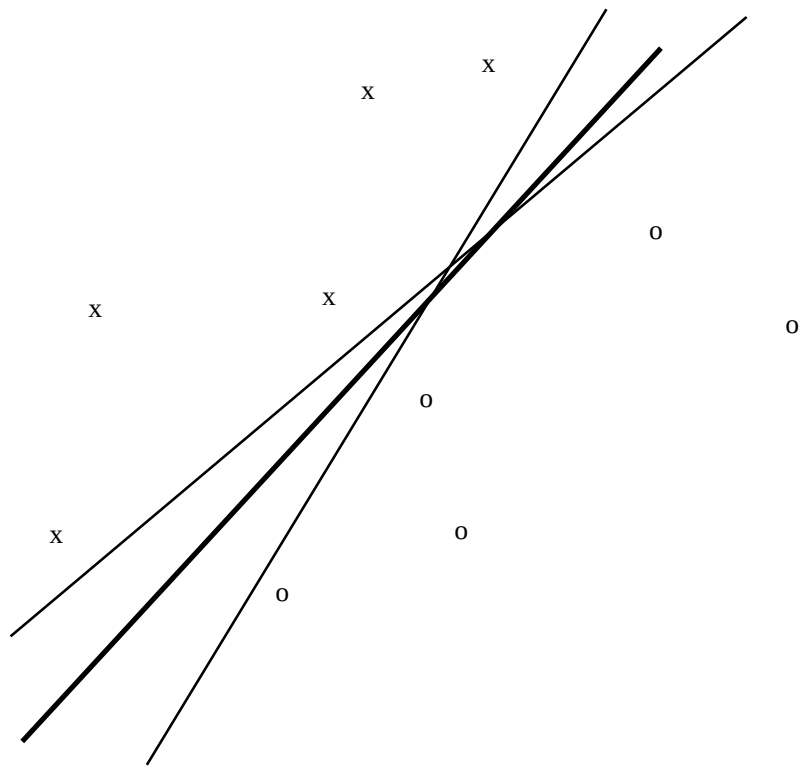
négligeable à la petite échelle
prend le dessus à la grande échelle (Google)

Espace d'hypothèses

Hypothèses numériques

- Fonctions linéaires

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

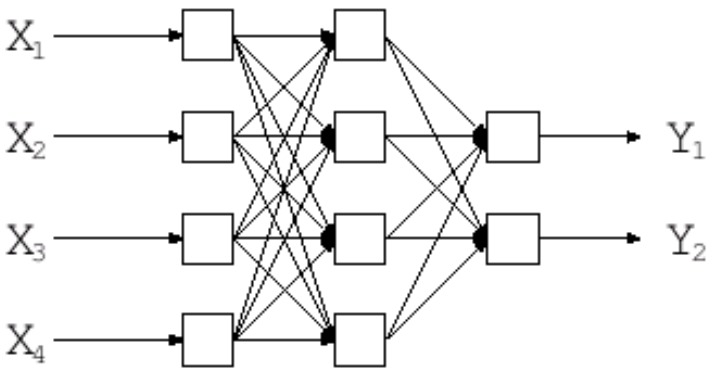


Hypothèses numériques, 2

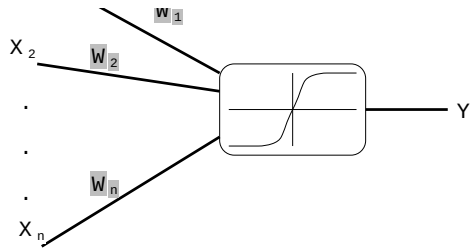
Réseaux neuronaux

$$S(x) = sigma(w_1 z_1 + w_2 z_2 + ... + w_p z_p + w_0)$$

avec $sigma(X) = \frac{1}{1+e^{-X}}$



Muti-layer perceptron



Hypothèses discrètes

Formules booléennes

- Conjonctions

$$Panne\ si\ \neg Essence$$

$$Malade\ si (Temperature > 39.5)$$

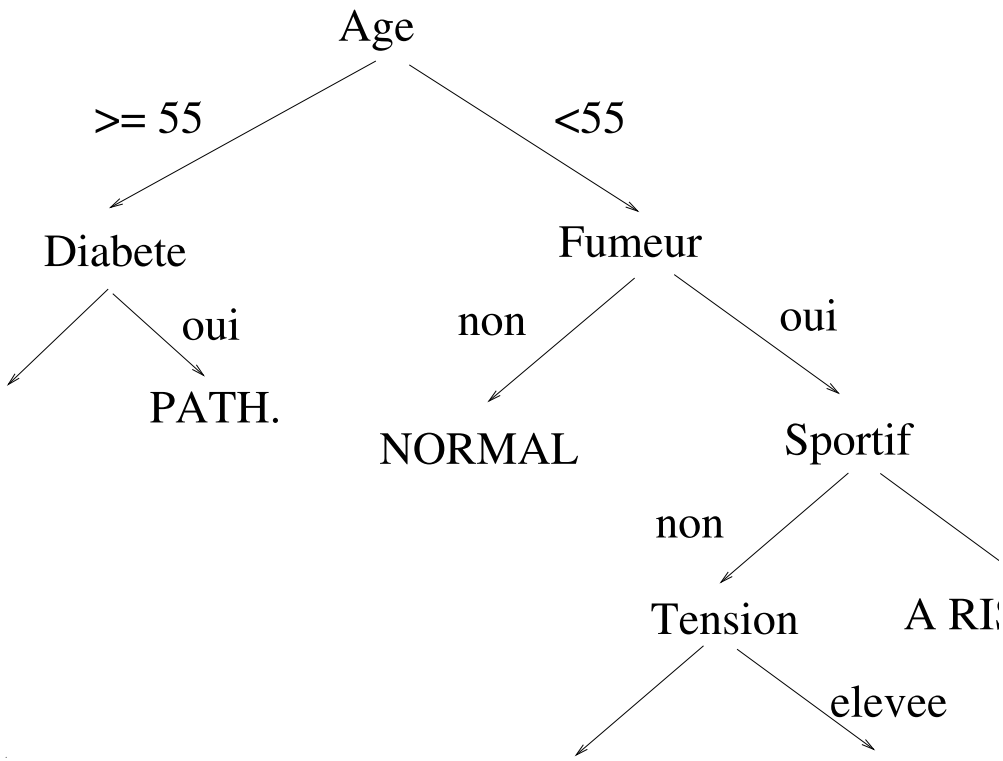
- Liste de décision

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

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 $\mathcal{E} = \{(x_i, y_i)_{i=1}^n,\ x_i \in \mathbb{R}^n,\ y_i \in \{0,1\}\}$
   - 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

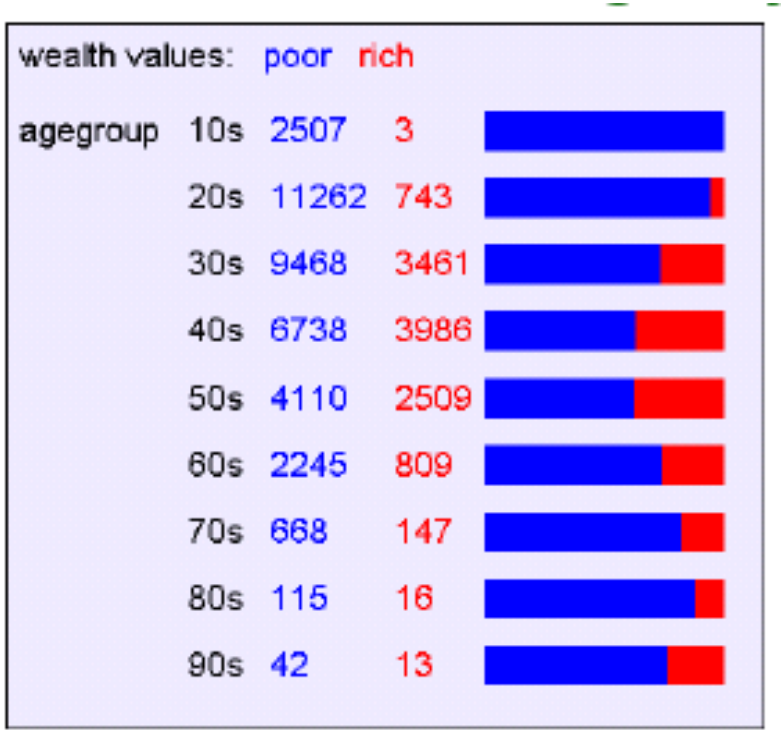
$$n \ln n$$

Pour construire un noeud

$$p \times n \ln n$$

Arbres de décision, 3

Table de contingence



Limitations

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

Limitations

Attributs numériques

- Ordonner les valeurs $val_1 < \ldots < val_t$
- Calculer QI ($[att < val_i]$)
- $\text{QI}(\text{att}) = \max_i \text{QI}([att < val_i])$

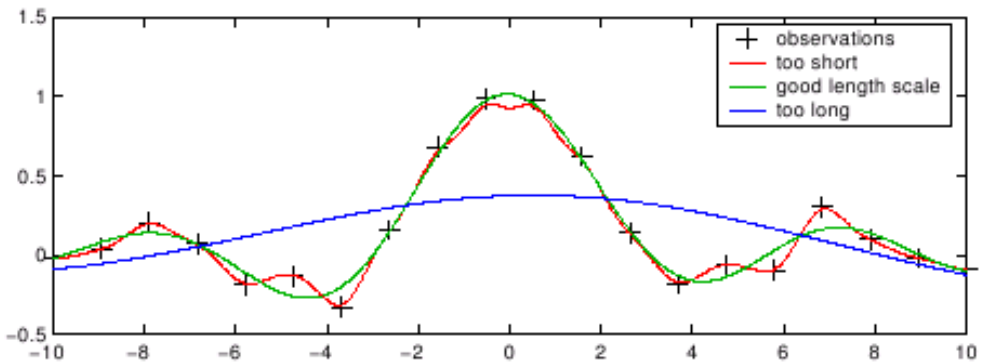
XOR

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

Plan de ce cours

1. Introduction, définitions, objectifs
2. Méthodologie
3. Classification
4. Etude de cas : SKICAT

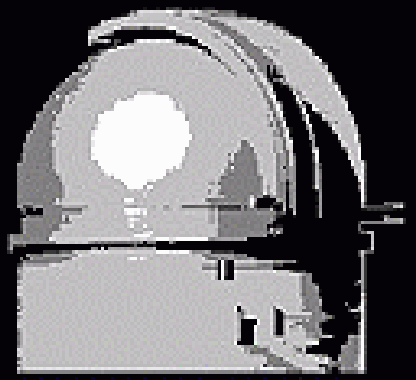
Skicat

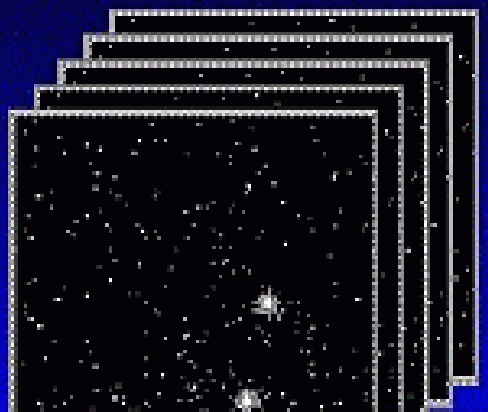
U. M. Fayyad, S. G. Djorgovski, and N. Weir. 1996
Jet Propulsion Lab., Caltech

- 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





Unclassified Sky Objects

Decision Tree Classifier

Image Catalogs

Survey Catalog

Data For Analysis

Attribute Measurement

Image Processing

Learning Algorithm

SKICAT, 2

Objectif final

catalogue du ciel

objets d'un ordre de grandeur moins brillants

Caltech, release 93

$\approx$ 40,000 volumes

Le problème

trop nombreux candidats : artefacts ?

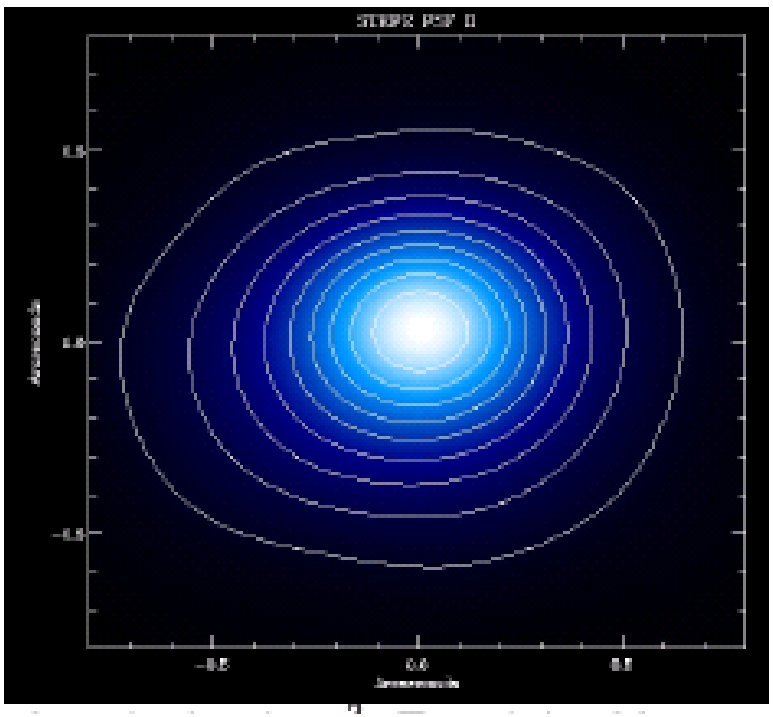
3 Téra bytes.

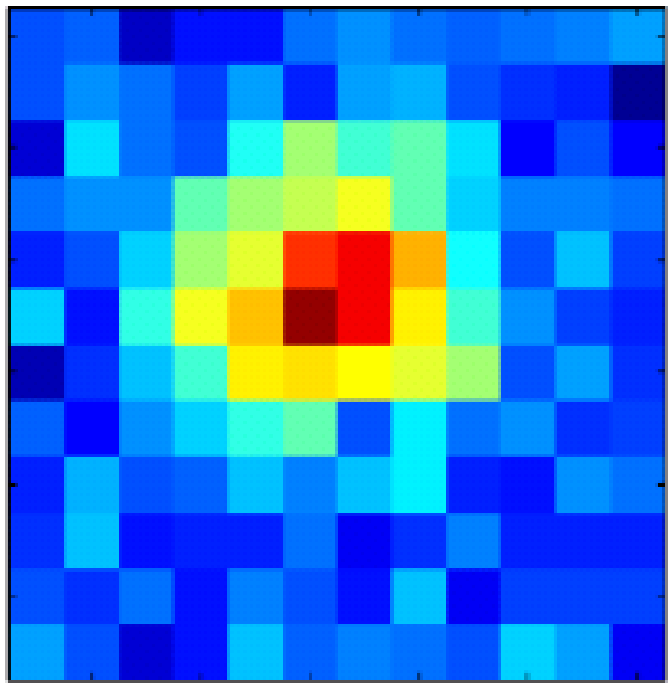
tera Terrorbytes

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

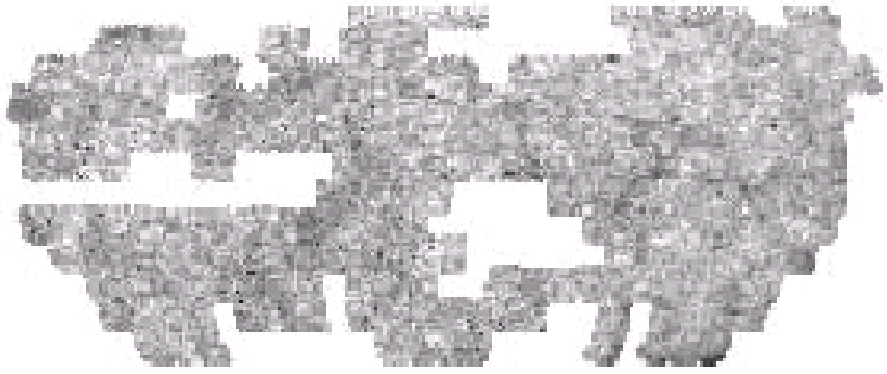
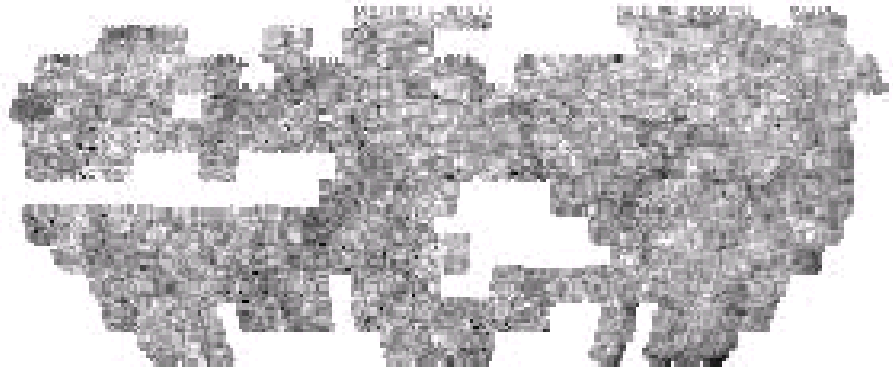
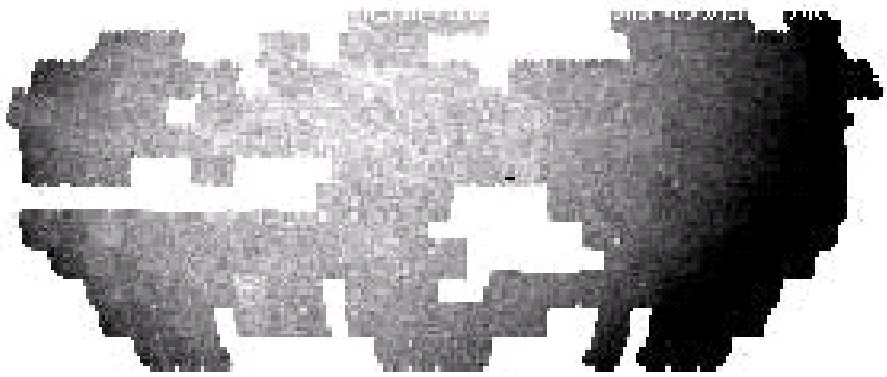
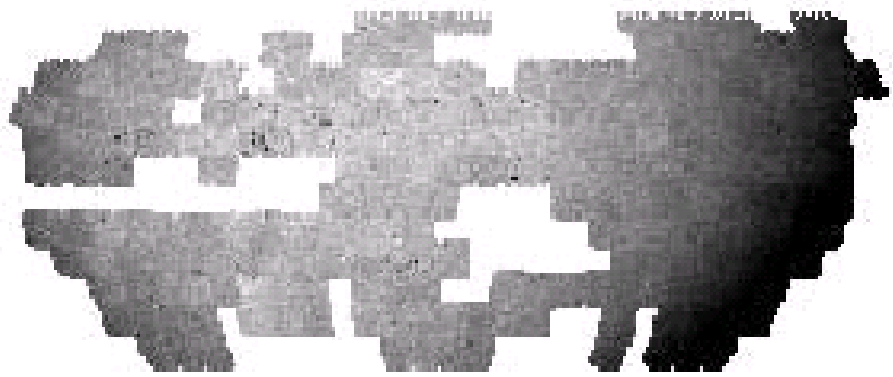
entre géologistes

pour un même géologiste

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

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

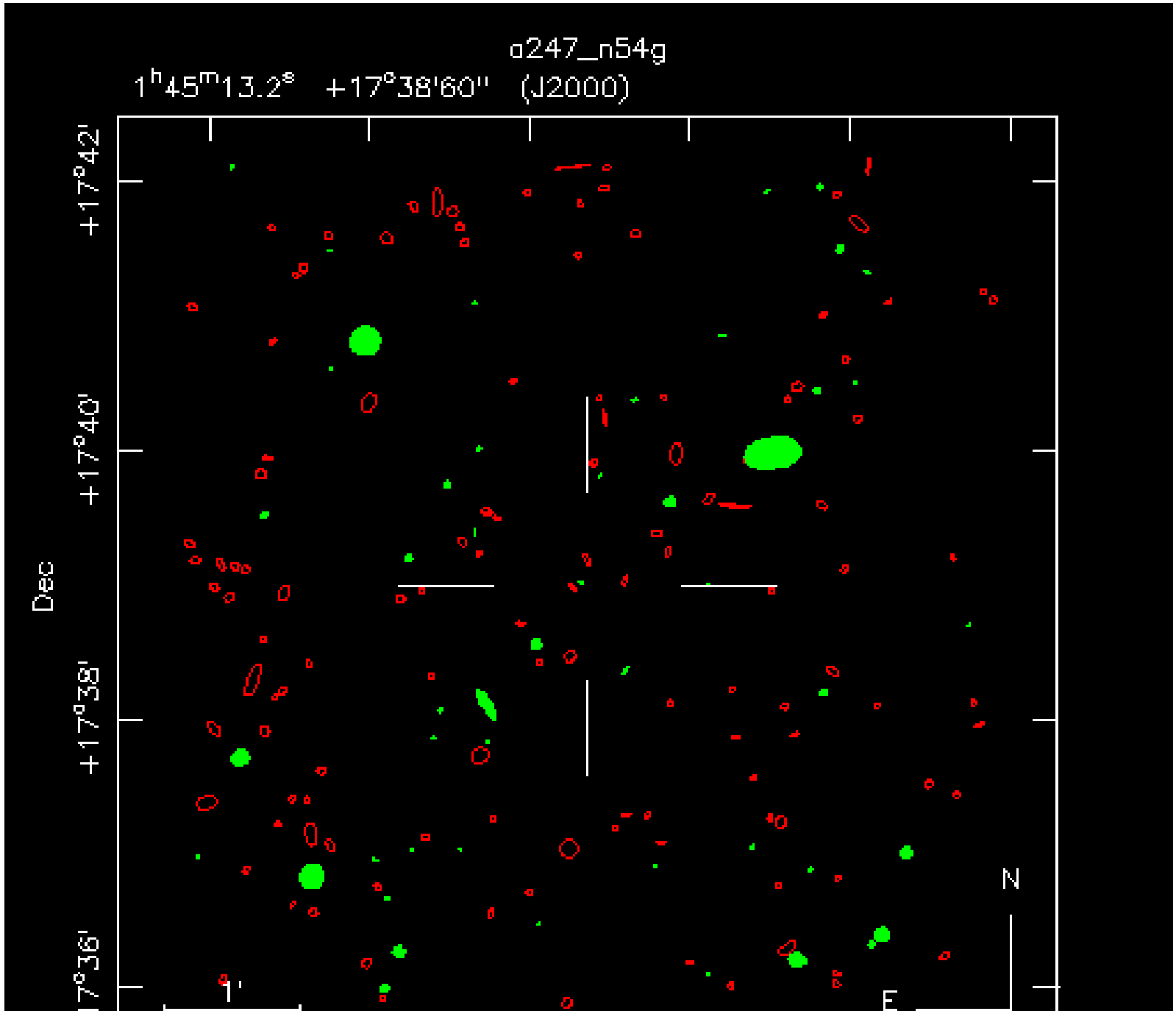
Classified objects: one magnitude fainter than previously

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.


a247_n54g
1h45m13.2s +17°38'60" (J2000)
Dec
+17°42'
+17°40'
+17°38'
+17°36'
1'
N
E

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.

DARPA Challenge - 2005

Vision de près

Apprentissage en ligne et Boostrap

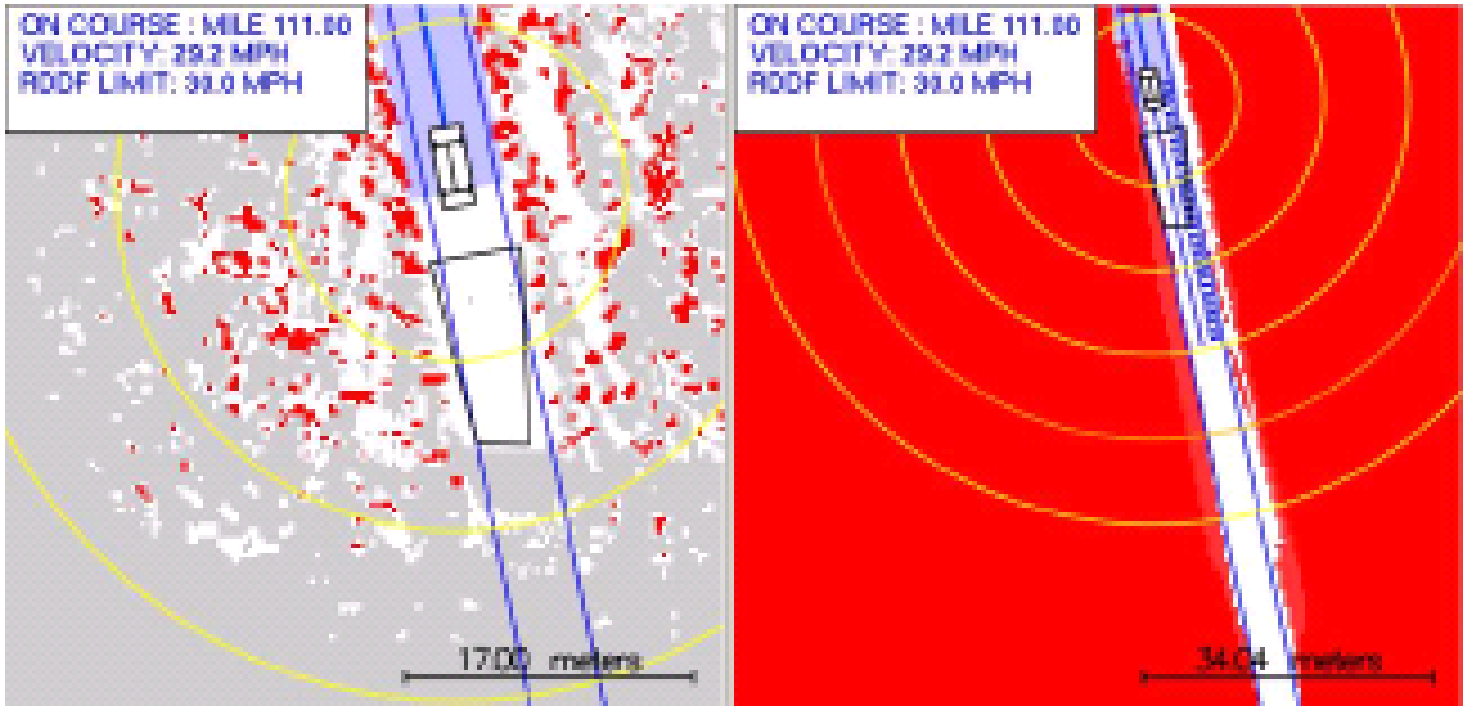
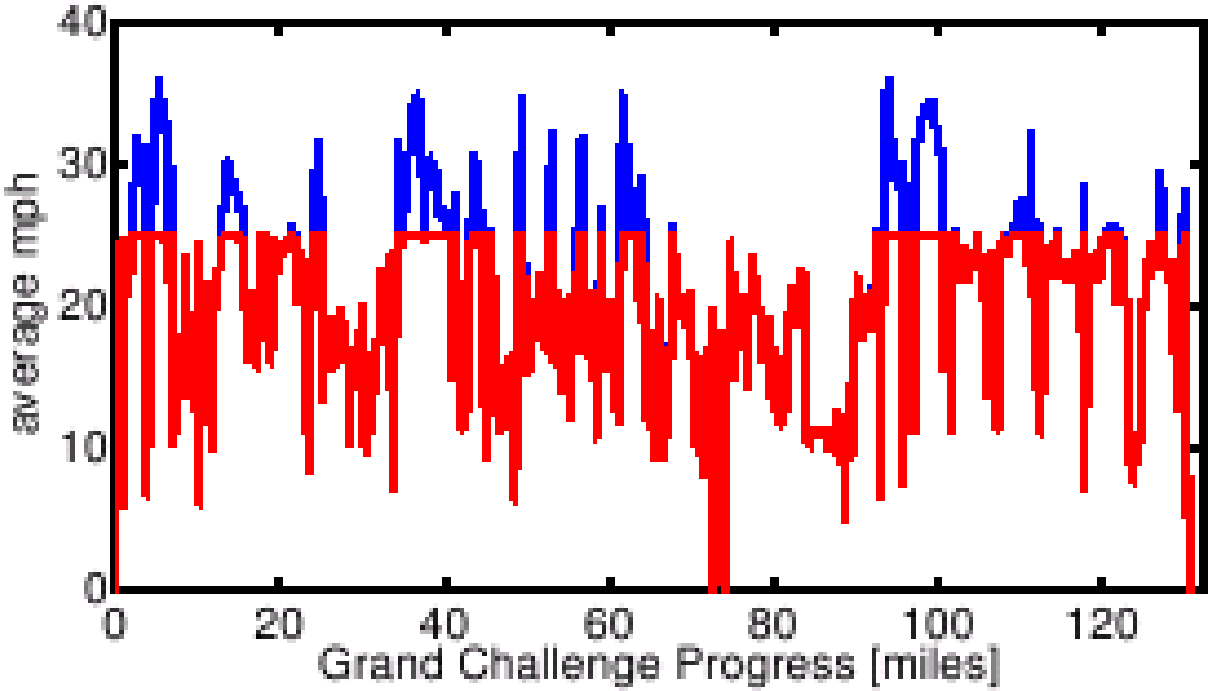


Fig. 2. Real-time generated map of the vehicle vicinity.
The left image shows close-range map as generated by the laser scanners. Red cells are occupied by obstacles, white cells are drivable and grey cells are (as yet) unknown. The black quadrangle is fit into the known empty area and shipped to the computer vision algorithm as training data for drivable road surface.
The right image shows a long-range map as generated by the computer-vision algorithm described in this paper. It can be seen that the road detection range is in this case about 70m, as opposed to only 22m using lasers.

Vitesse



Plus de détails:
http://robots.stanford.edu/papers/dahlkamp.adaptvision06.pdf

Leçons

Le but de la collecte

Les données sont-elles collectées pour l'analyse ?

Usage dérivé, reformulation des données.

Passage à l'échelle

efficacité quand les données ne tiennent pas en mémoire

données (même aléatoires) de grande taille

$\Rightarrow$ contiennent des motifs réguliers.

critères statistiques $\longrightarrow$ comportement asymptotique

besoin de mesures de qualité

Avec du recul

Vision

DM elevates the way we interact with DB

like to see three buttons:

What's new

What's interesting

Predict for me

Myth

DM pervasive; Large datawarehouses; Companies know how

Integrated view

Success means invisibility

Maps findings to actions

Integrating domain knowledge: no AI-hard: deep and narrow.

Technical challenges

- How does the data grow ? Not iid. specially with time.
- Explain how, when, why, a model fails.
- Tuning: complex/understandable
- Interestingness (common sense)
- Scalability/integration DBMS / sampling / abstractions
- Interaction / Visualisation