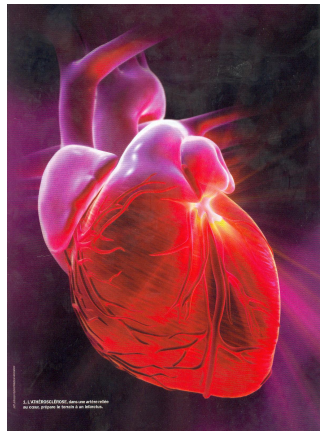
Master 2 Recherche
Apprentissage Statistique et Optimisation

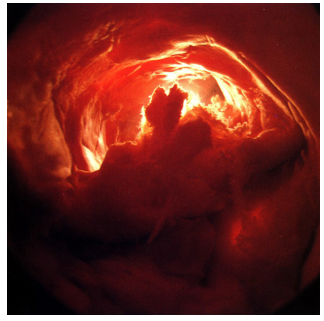
François Yvon – Michèle Sebag
Alexandre Allauzen – Marc Schoenauer

http://www.limsi.fr/Individu/allauzen/wiki/index.php/TSI09 – http://tao.lri.fr/tiki-index.php

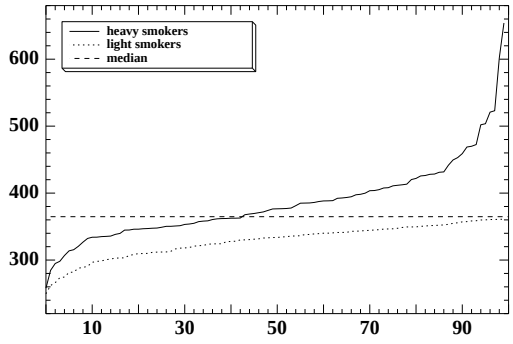
30 septembre 2010

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
  décision et conception optimale: des jeux aux politiques d'énergie



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

L'une des 10 technologies émergentes du 21^e siècle

(MIT Review, 2001)

Le contexte, Master Informatique Paris-Sud

Modules ayant un rapport

- Données semi-structurées et XML ...
- Apprentissage, Optimisation et Applications *O*
- Intégration de données et Web sémantique ...
- Modèle de raisonnement distribué ...
- Reconnaissance vocale, indexation multilingue *O*
- Robotique et agents autonomes *O*
- Systèmes multi-agents *O*
- Traitement automatique des langues *O*
- Interaction Homme-Machine *o*
- Réalité Virtuelle et Augmentée ...

Le contexte, Master Informatique Paris-Sud, suite

Ce module

- Théorie
- Applications
- TP
- Présentations d'articles 10 mn / volontaires

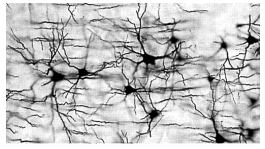
Examen

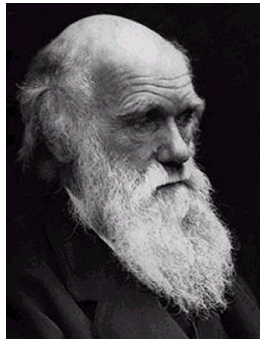
- Questions de cours
- Au choix:
  - Exposé (présentation orale + résumé) d'un article
  - Projet

Plan du Module

1. Introduction
   pourquoi, mise en jambe, définitions,
   gradients, quelques exemples
2. Statistiques
3. Réseaux de neurones
4. Apprentissage de séquences
5. Apprentissage non supervisé
6. Changements de représentation
7. Etre Bayésien
8. Optimiser

Rev. T. Bayes





Quelques bonnes adresses

- Où sont les cours :
  http://tao.lri.fr/tiki-index.php?page=Courses
  http://www.limsi.fr/Individu/allauzen/wiki/index.php/TSI09
- 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. Partie 1. Généralités
2. Partie 2. Apprentissage supervisé, 1

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

$$\text{Monde} \longrightarrow \text{instance } \mathbf{x}_i \longrightarrow \begin{array}{c}\text{Oracle}\\ \downarrow \\ y_i\end{array}$$

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

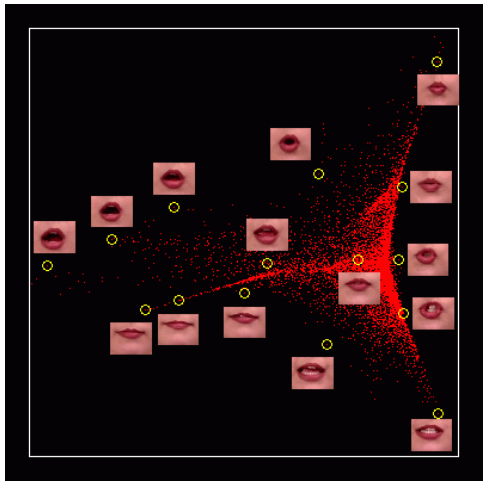
Output: Hypothèse $h : \mathcal{X} \mapsto \mathcal{Y}$

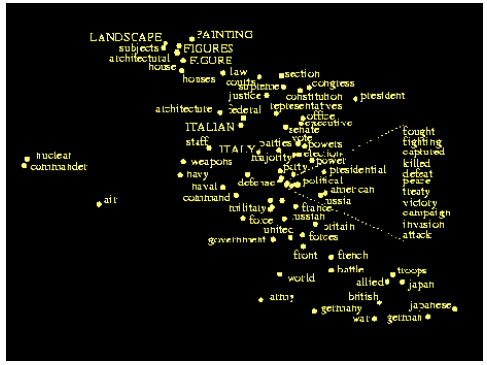
Critère: pas d'erreur (precisions bientot)

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.

...imprécises...

des connaissances { nouvelles / utiles / valides } % 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
  $I = \{$ *Vendredi, bière, couches* $\}$
  $I = \{$ *Pain, beurre, confiture* $\}$
- Règles d'association — seuil de confiance
  Vendredi, couches $\Rightarrow$ *bière*
  Pain, beurre $\Rightarrow$ *Confiture*

Quelques définitions

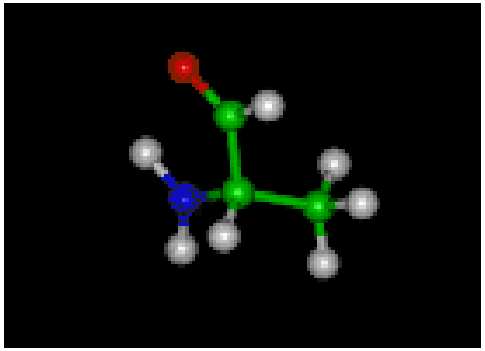
Exemple

- ligne : exemple/ cas/individus/transactions
- colonne : attribut/ feature/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

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 analyse & synthèse/contrôle

Applications industrielles, sciences expérimentales, calcul numérique

Vision, voix, parole, robotique..

Phénomènes sociaux + confidentialité

Hôpitaux, Assurances, Banques, ...

Phénomènes individuels + dynamique rapide

Consumer Relationship Management, User Modelling

Réseaux sociaux, jeux...

PASCAL : 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

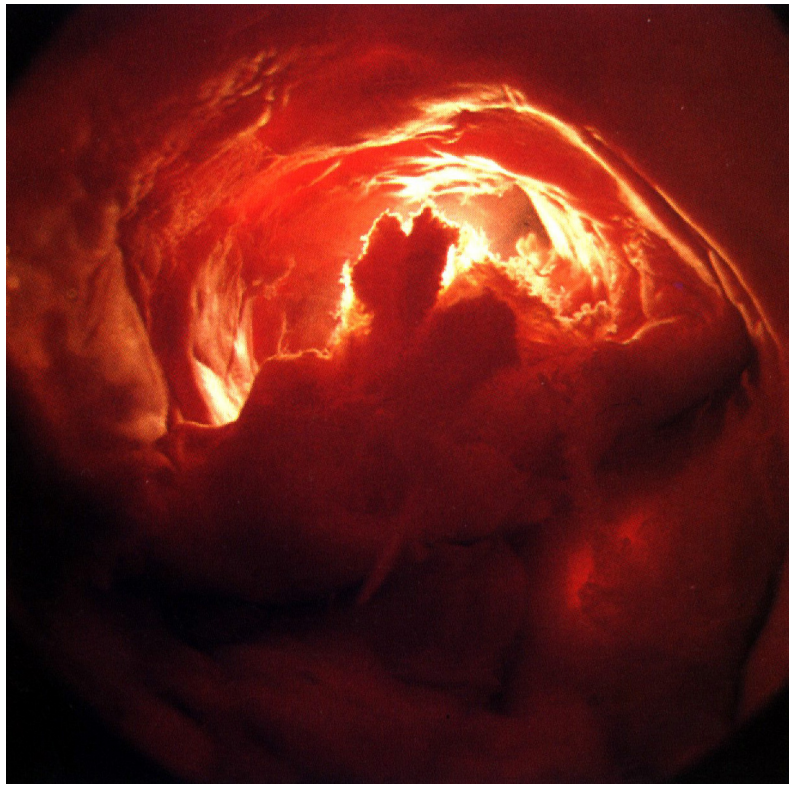
Médecine, Bio-Informatique, Sécurité

Ex: Facteurs de risque

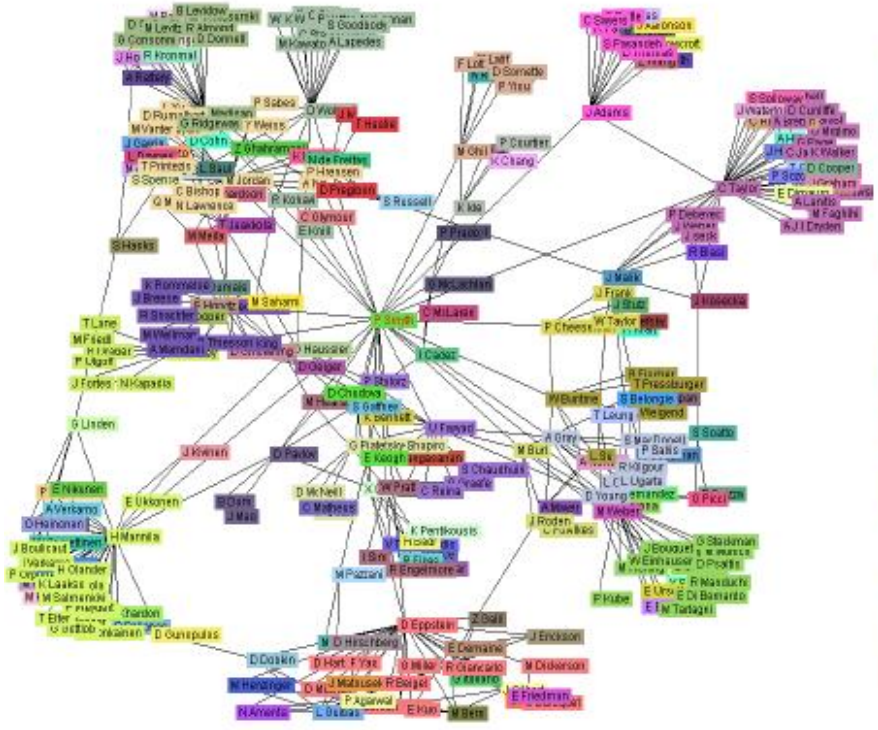
1. Maladies cardio-vasculaires
2. Molécules cancérigènes
3. Gènes de l'obésité...

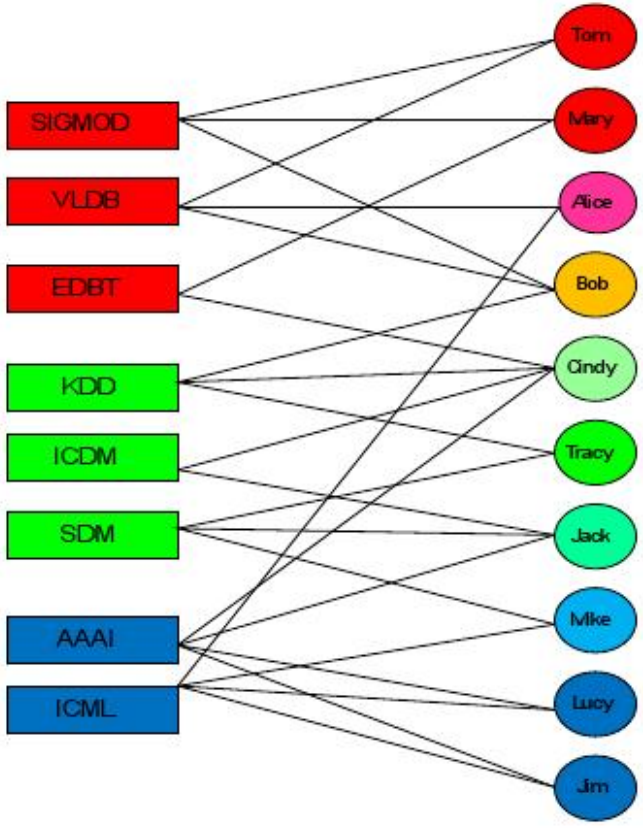
Objectifs

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



Auteurs, conférences, thèmes





Questions

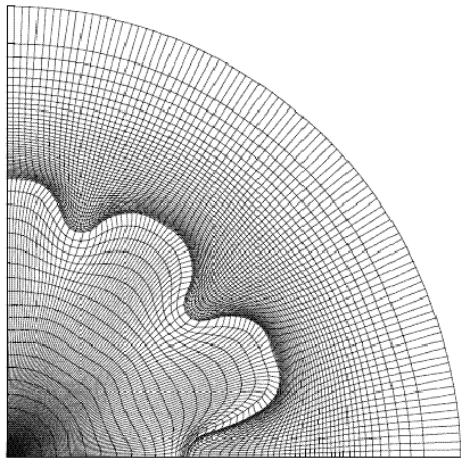
1. Qui fait quoi ?
2. Quelles sont les bonnes conférences ?
3. Quels sont les thèmes qui montent ?
4. Mr Q. Lee est-il le même que Mr Quoc N. Lee ?

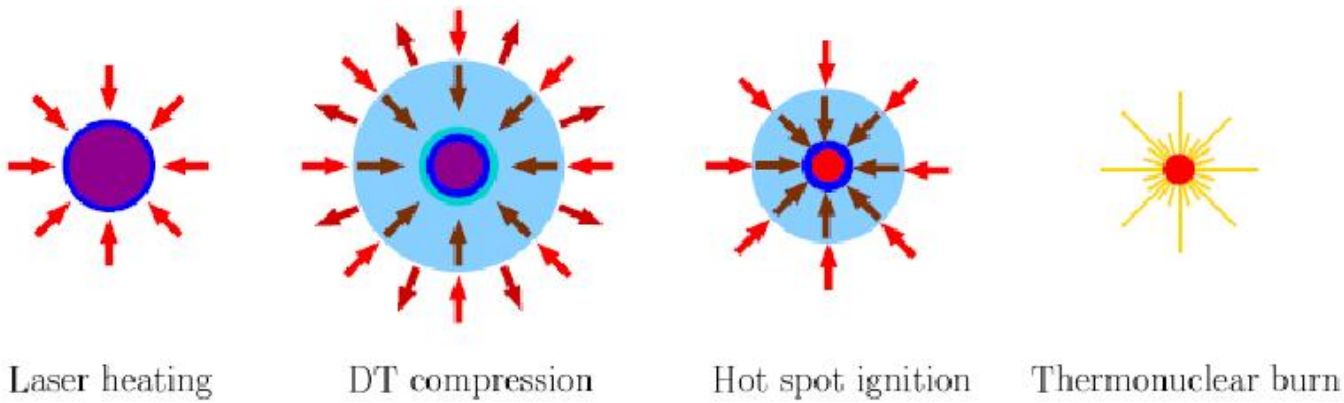
[tr. Jiawei Han, 2010]

e-Science, Conception

Numerical Engineering

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





Fusion par confinement inertiel, ICF

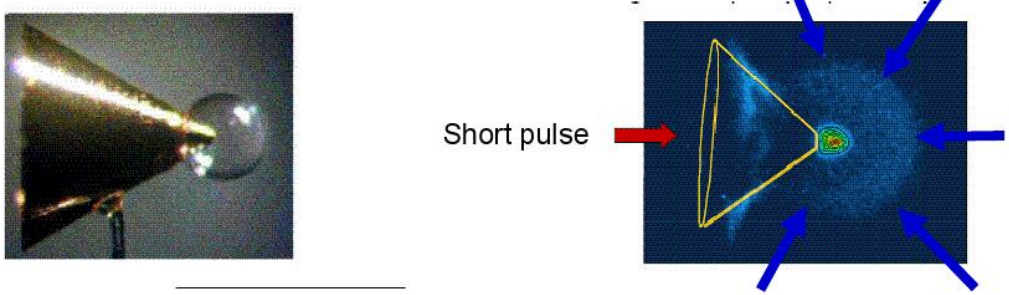
e-Science, Conception (2)

Buts

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

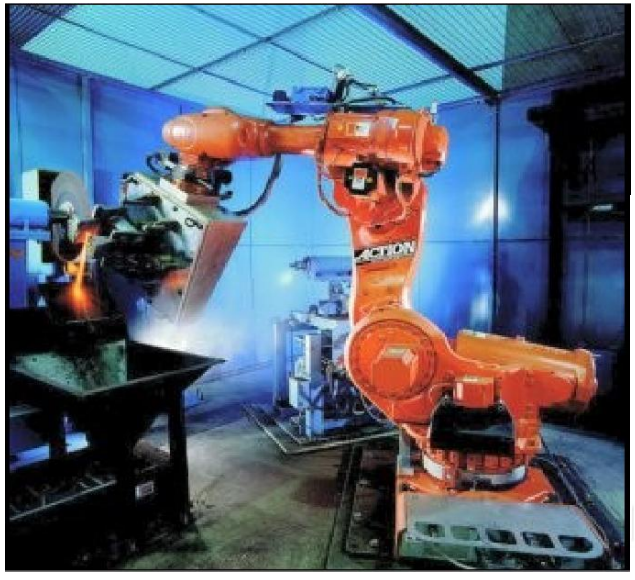
More is Different

Alternative scheme : spherical target with a gold cone*



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

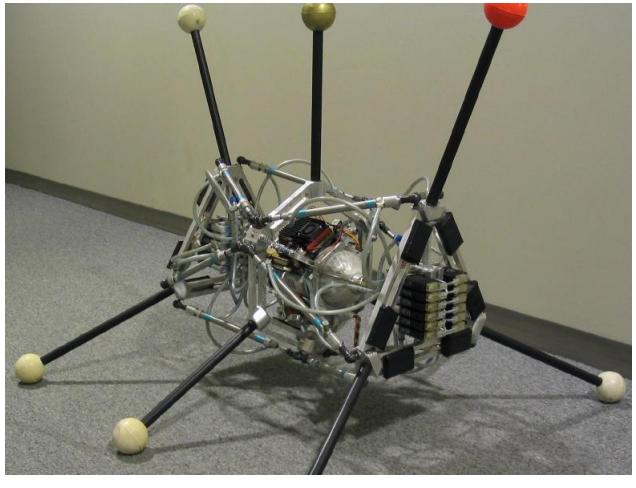
Robotique autonome



Complexe, monde fermé

simple, random

Conception



[tr. Hod Lipson, 2010]

Robotique autonome, 2

Reality Gap

- Concevoir en simulation (in silico)
- Essayer la solution sur le vrai robot (in vivo)

Robotique autonome, 2

Reality Gap

- Concevoir en simulation (in silico)
- Essayer la solution sur le vrai robot (in vivo)
- Catastrophe !

Closing the reality Gap

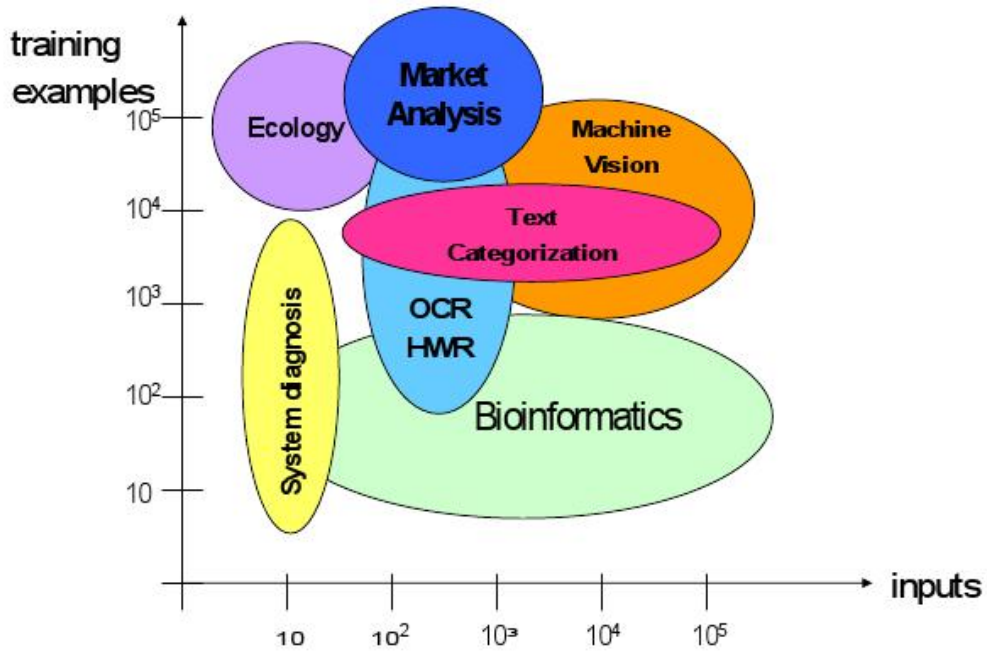
1. Concevoir en simulation
2. Essayer sur le vrai robot environnement protégé
3. Logger les données et mettre à jour le simulateur
4. Goto 1

Apprentissage actif Co-évolution

[tr. Hod Lipson, 2010]

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
- $\searrow$ 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 n 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

Partie 1. Généralités

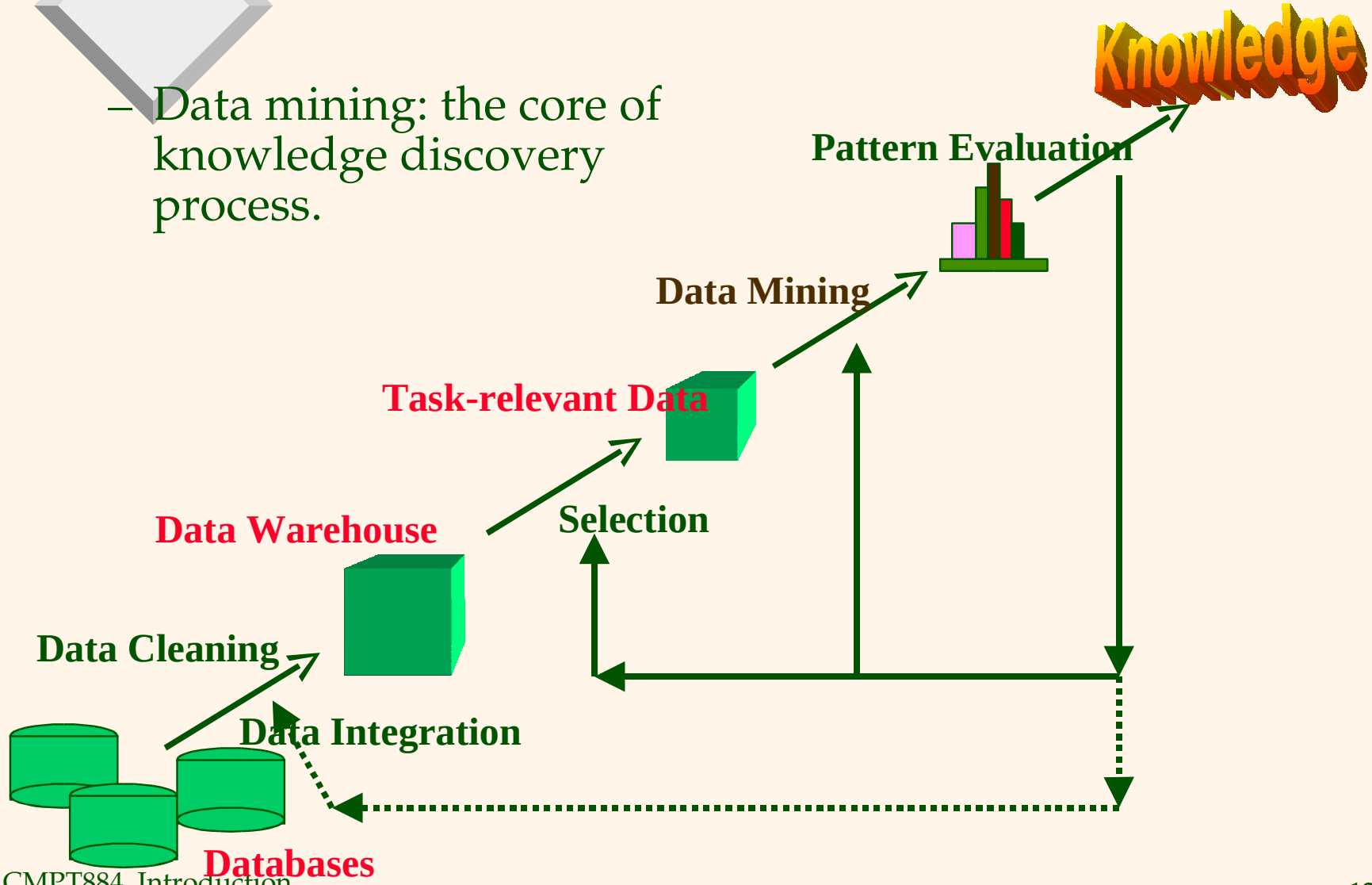
1. Objectifs: Apprentissage supervisé, non supervisé, fouille de données
2. Quelques définitions
3. Domaines d'applications
4. Quelles sont les difficultés ?
5. Méthodologie

Processus (Vision apprentissage)

1. Collecter les données
2. Choisir les attributs — connaissance a priori
3. Choisir le modèle — connaissance a priori
4. Apprendre
5. Valider
6. Fin

Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



Processus (vision Fouille de Données)

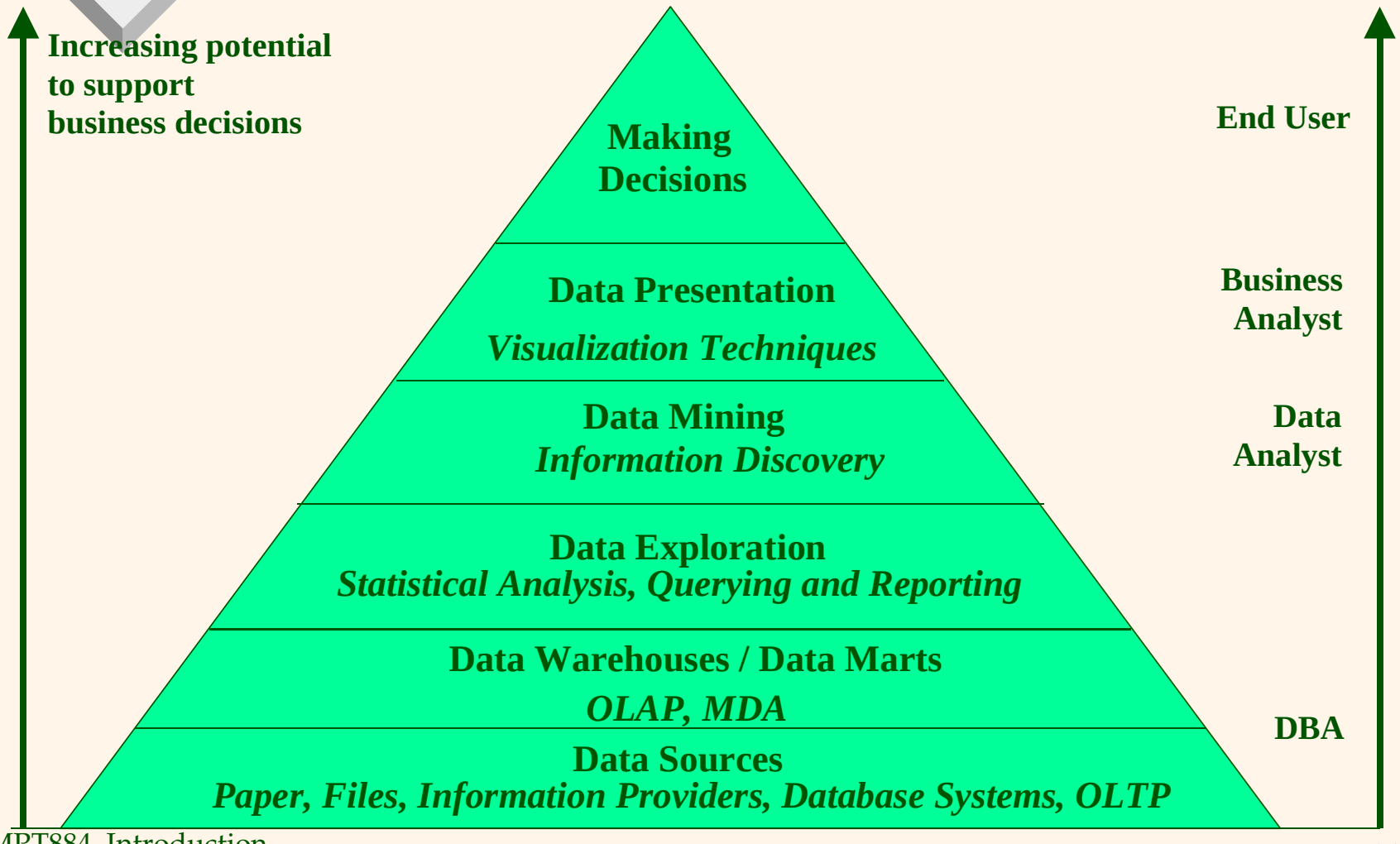
Les étapes

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



Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Le cas le plus simple
4. Quel est le critère ? théorie
   Suffisamment compliqué, mais pas plus
5. Arbres de décision
6. Valider pratique
7. Etude de cas: SKICAT
8. Etude de cas: Stanley

Apprentissage supervisé

Contexte

$$\text{Monde} \longrightarrow \text{instance } \mathbf{x}_i \longrightarrow \begin{array}{c}\text{Oracle}\\ \downarrow \\ y_i\end{array}$$

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

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

Critère: Qualité de h

Apprentissage supervisé, 2

Définitions

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

Tâches

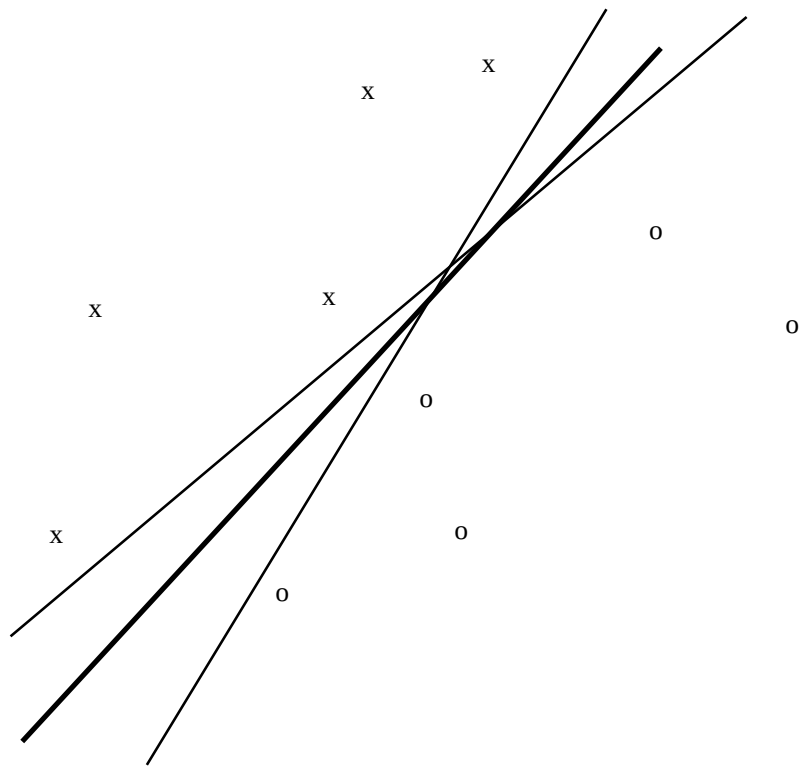
- Choisir $\mathcal{H}$ — sélection de modèle
- Evaluer $h \in \mathcal{H}$ — $score(h)$
- Choisir h^* dans $\mathcal{H}$ — $argmax\ score(h)$

Espace d'hypothèses

Hypothèses numériques

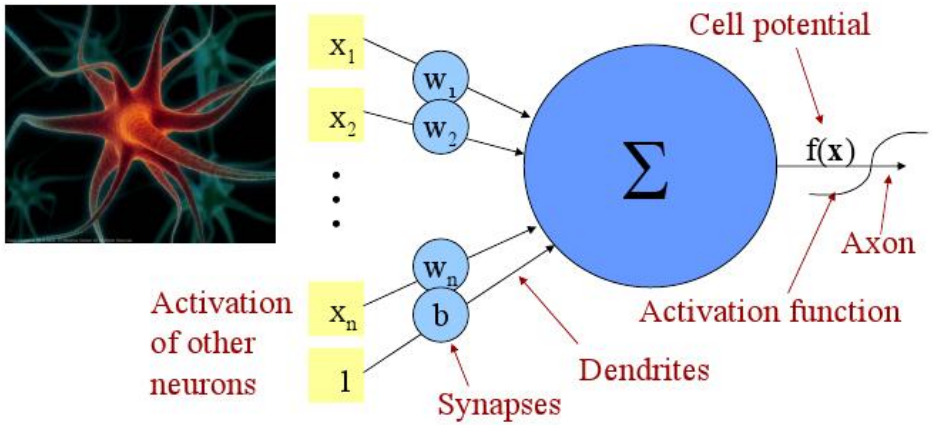
- Fonctions linéaires

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



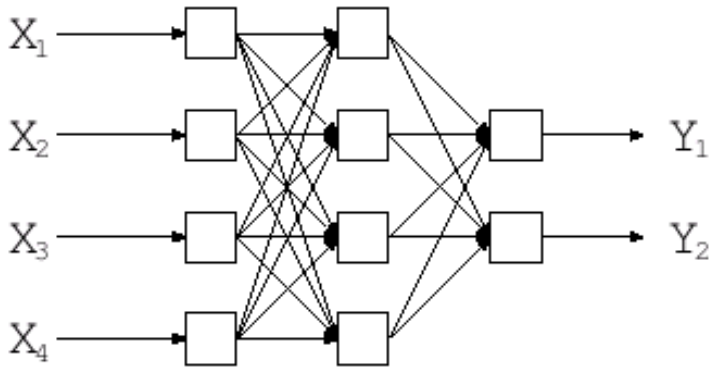
Hypothèses numériques, 2

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

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

Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Le cas le plus simple
4. Quel est le critère ? théorie
   Suffisamment compliqué, mais pas plus
5. Arbres de décision
6. Valider pratique
7. Etude de cas: SKICAT
8. Etude de cas: Stanley

Pourquoi l'optimisation

Cas de la régression $\mathcal{E} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}, i = 1 \dots n\}$

Score: moindre carrés least mean square

Apprentissage = Optimisation
Etant donné $\mathcal{E}$, trouver

$$h^* \in \mathcal{H} \text{ minimisant } \sum_{i=1}^{n} (h(\mathbf{x}_i) - y_i)^2$$

L'espace d'hypothèses le plus simple
$\mathcal{H}$: fonctions linéaires sur $\mathcal{X} = \mathbb{R}^d$

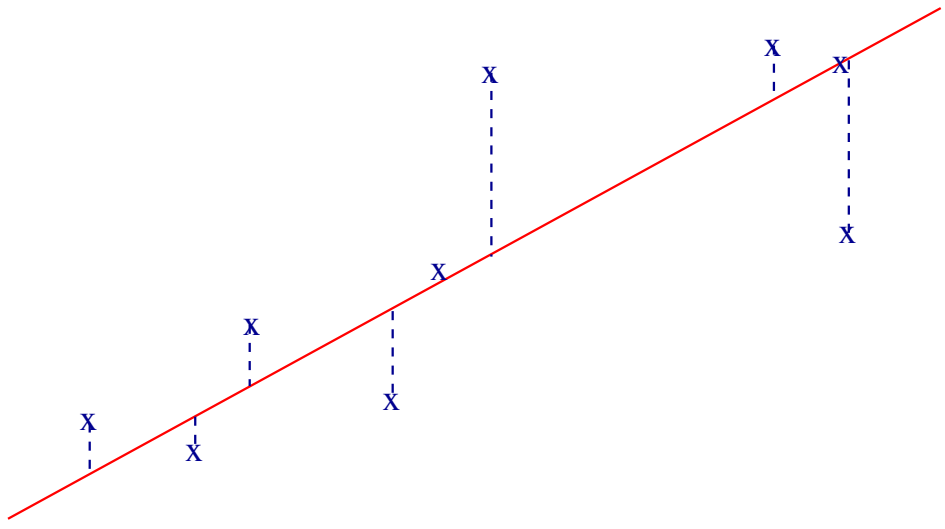
$$h : \mathbb{R}^d \mapsto \mathbb{R}$$

Pourquoi l'optimisation, 2

Fonction linéaire

produit scalaire

$$h(\mathbf{z}) = h(z_1, ..z_d) = \sum_i w_i z_i + w_0 = < w, z > + w_0$$



Erreur quadratique

$$w^* = argmin\{Err(w) = \sum_{i=1}^{n} (< w, \mathbf{x}_i > - y_i)^2\}$$

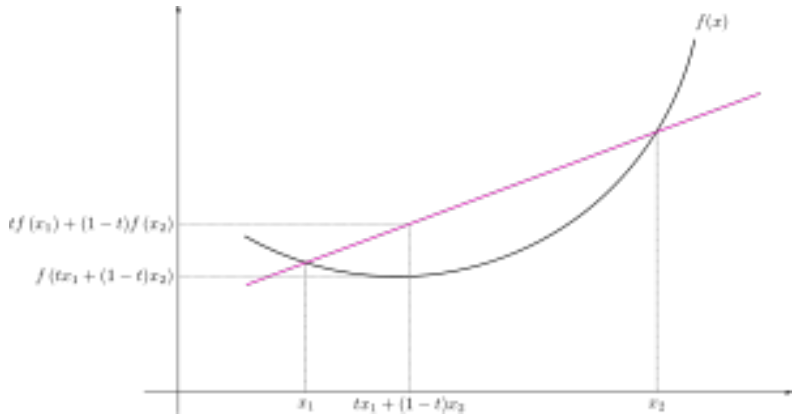
(NB: On sait trouver analytiquement w^* dans ce cas; mais regardons le cas general).

Good News

une fonction quadratique est une fonction convexe

Fonction convexe, définition

$$\forall t \in [0,1], f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2)$$



Propriétés

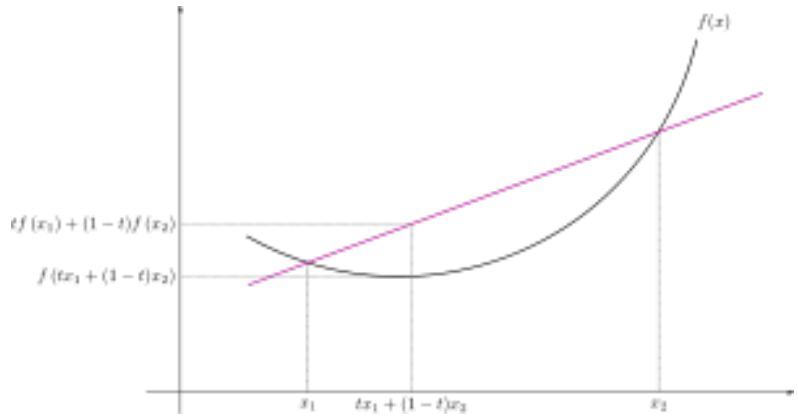
- Une fonction convexe admet un unique minimum global.

Good News

une fonction quadratique est une fonction convexe

Fonction convexe, définition

$$\forall t \in [0,1], f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2)$$



Propriétés

- Une fonction convexe admet un unique minimum global.
- Et on sait le trouver: suivre le gradient...

Gradient – Dérivées

Dérivée

dimension 1

$$\partial f / \partial t = lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

Gradient – Dérivées

Dérivée — dimension 1

$$\partial f / \partial t = lim_{t \to 0} \frac{f(x+t) - f(x)}{t}$$

Gradient — dimension d

$$\nabla f = (\frac{\partial f}{\partial t_1}, \ldots, \frac{\partial f}{\partial t_d})$$

Gradient – Dérivées

Dérivée — dimension 1

$$\partial f/\partial t = lim_{t\to 0}\frac{f(x+t)-f(x)}{t}$$

Gradient — dimension d

$$\nabla f = (\frac{\partial f}{\partial t_1}, \ldots, \frac{\partial f}{\partial t_d})$$

Exemple

$$\frac{\partial Err(w)}{\partial w_1} = 2\sum_{i=1}^{n}(< w, \mathbf{x}_i > -y_i).\mathbf{x}_{i,1}$$

Optimisation, méthodes de gradient

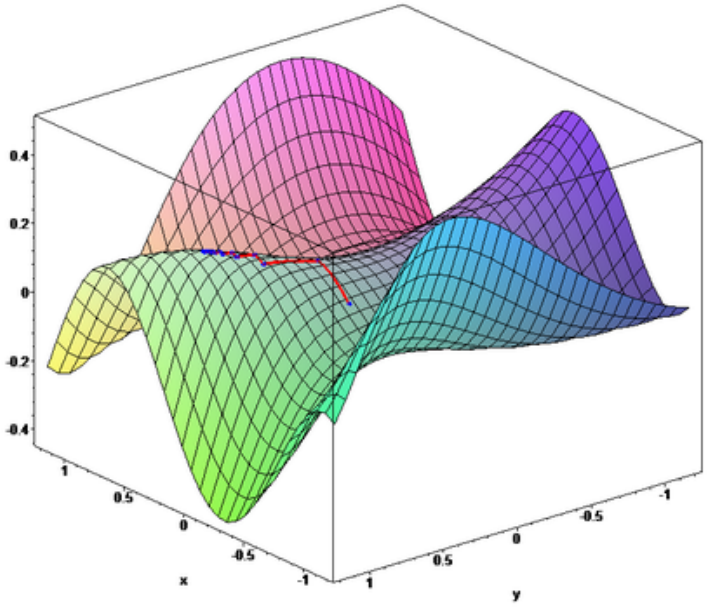
Gradient et optima

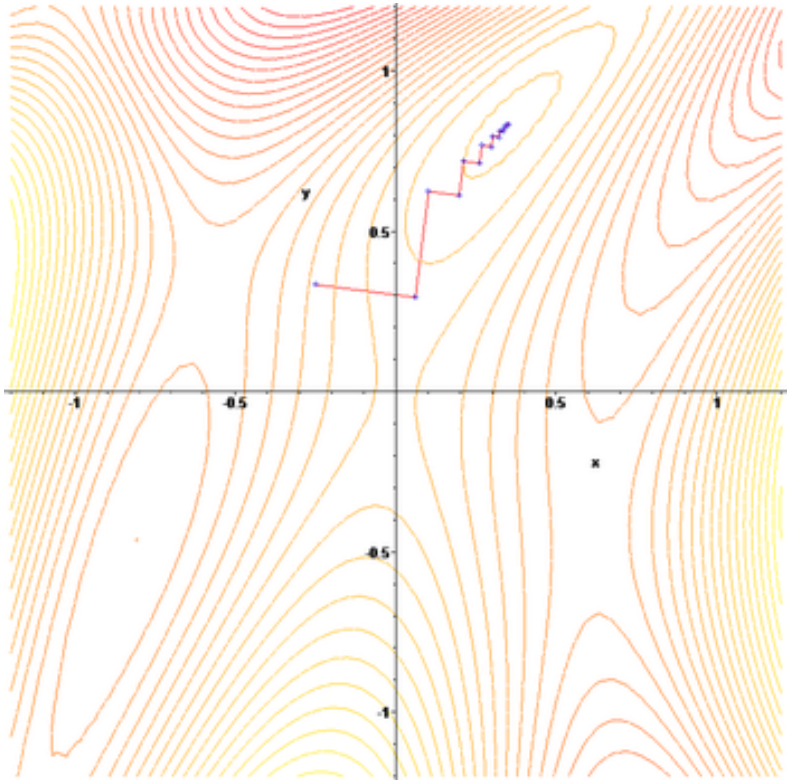
On est à l'optimum $\Leftrightarrow$ le gradient est nul

Méthode de gradient

$$w_t = w_{t-1} - \alpha_t \nabla f(w_{t-1})$$

Arrêt: quand $||w_t - w_{t-1}||$ petit.





Méthodes de gradient

Pour

- En général facile de calculer le gradient
- On peut essayer (long) de trouver le α_t (le pas) optimal

Limites

- Nombre d'itérations possiblement grand.
- Si la fonction présente une grande vallée plate (high condition number), on fait des pas minuscules...

(Remède général: aller chercher la dérivée d'ordre 2: le Hessien).

Quand le problème admet une solution analytique

Cas unidimensionel $\mathcal{E} = \{(x_i, y_i), x_i \in \mathbb{R}, y_i \in \mathbb{R}, i = 1 \dots n\}$

$$Y = wX + w_0 + \text{ bruit}$$

Calculus: moyenne $\bar{x} = \frac{1}{n} \sum_i x_i$

$$\begin{aligned} \hat{w} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{j=1}^n y_j / n}{\sum_{i=1}^n (x_i^2) - (\sum_{i=1}^n x_i)^2 / n} \\ &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\text{Cov}[x,y]}{\text{Var}[x]} = r_{xy} \frac{s_y}{s_x}, \\ \hat{w}_0 &= \bar{y} - \hat{w}\, \bar{x}, \end{aligned}$$

Quand le problème admet une solution analytique

Cas multidimensionel $\mathcal{E} = \{(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in RR, i = 1 \dots n\}$

$$Y = XW + \text{ bruit}$$

avec

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ x_{21} & \cdots & x_{2d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

Calculus:

$$\begin{aligned} X^t Y &= X^t X W \\ W &= (X^t X)^{-1} X^t Y \end{aligned}$$

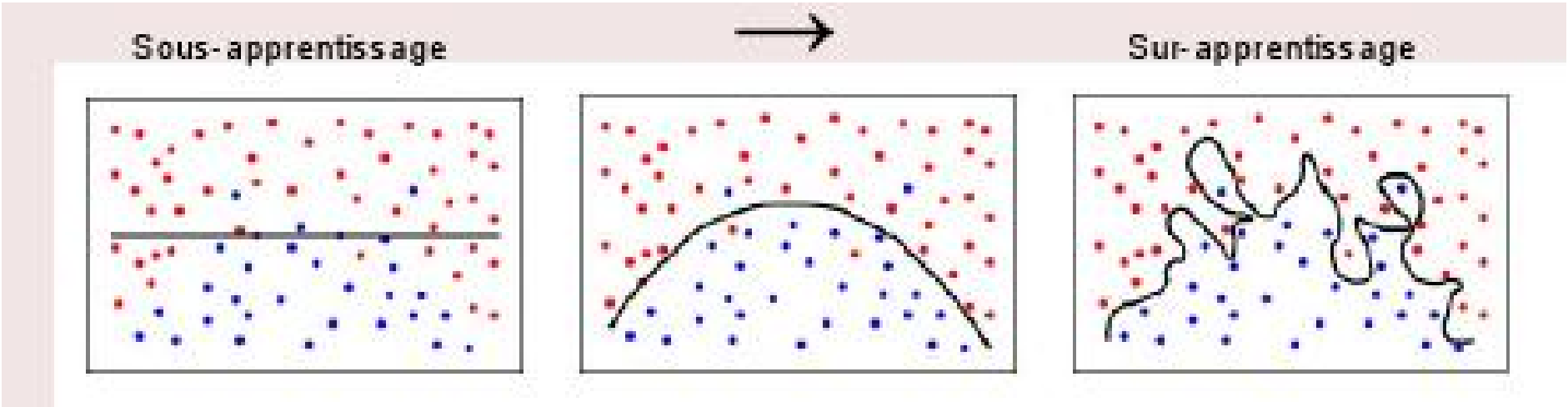
"A mature technology" (45')

`http://videolectures.net/stanfordee364aw08_boyd_lec01/`

Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Le cas le plus simple
4. Quel est le critère ? théorie
   Suffisamment compliqué, mais pas plus
5. Arbres de décision
6. Valider pratique
7. Etude de cas: SKICAT
8. Etude de cas: Stanley

Choisir l'espace d'hypothèses



Remarque: le but n'est pas de ne faire aucune erreur sur l'ensemble d'apprentissage...

Quel est l'objectif ? Qualité de h

Etre bon sur les futurs exemples

Condition nécessaire:

qu'ils ressemblent aux exemples d'entrainement

"même distribution"

Prendre en compte le coût des erreurs $\ell(h(x), y) \geq 0$

toutes les erreurs ne sont pas aussi graves...

Apprentissage Statistique

Minimiser l'espérance du coût de l'erreur

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

Apprentissage Statistique

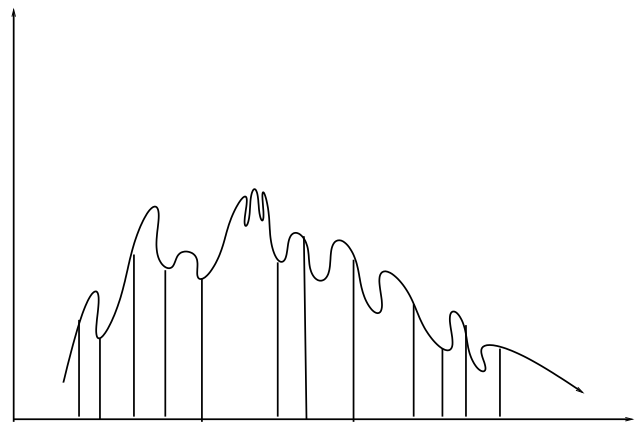
Minimiser l'espérance du coût de l'erreur

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

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

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

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 $\mathcal{X}$.

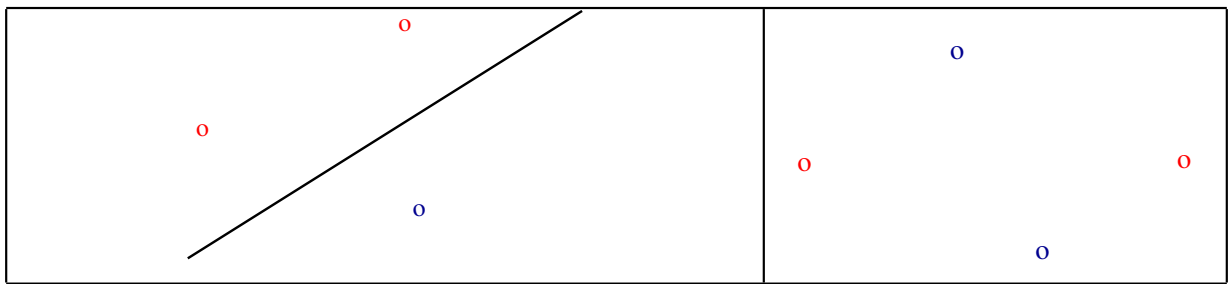
Si, $\forall (y_i)_{i=1}^n \in \{0,1\}^n, \exists h \in \mathcal{H} / h(x_i) = y_i$,

$\mathcal{H}$ 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...



3 pts pulvérisés par une droite	4 points, non

Définition

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

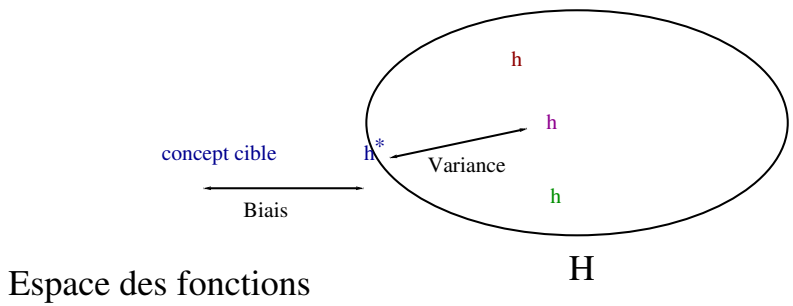
Classification, termes d'erreur

Biais

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

Variance

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



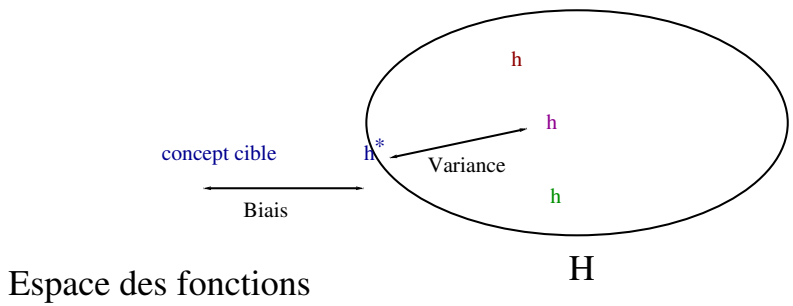
Classification, termes d'erreur

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)

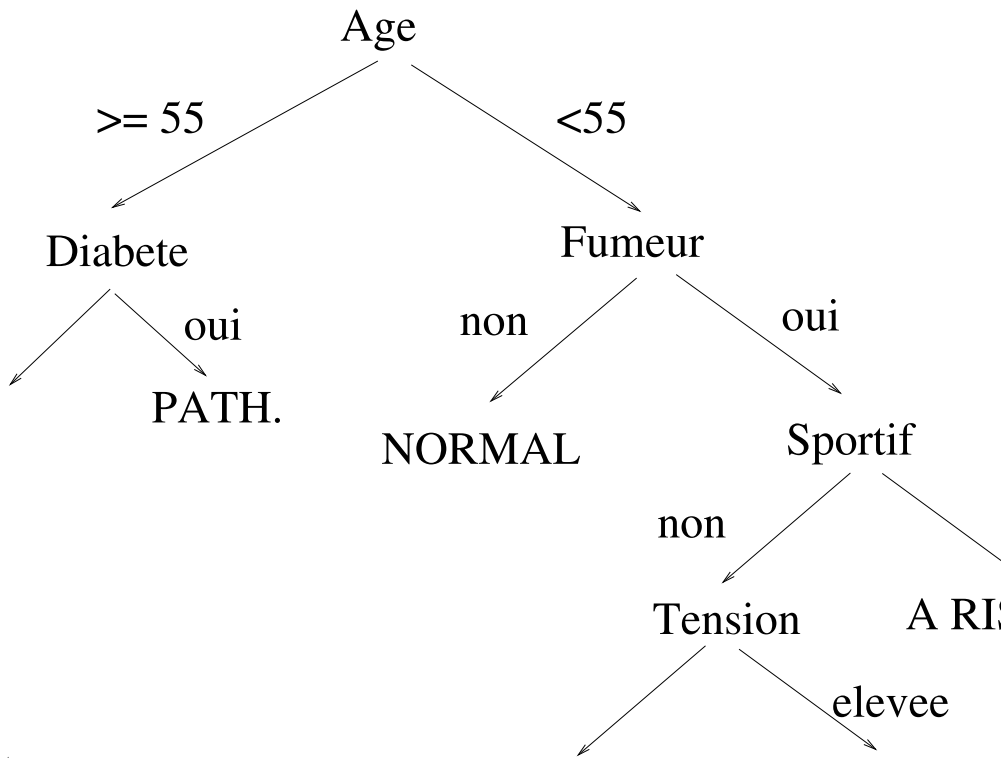
Partie 2. Apprentissage Supervisé

1. Poser le problème
2. Quelques espaces d'hypothèses
3. Le cas le plus simple
4. Quel est le critère ? théorie
   Suffisamment compliqué, mais pas plus
5. Arbres de décision
6. Valider pratique
7. Etude de cas: SKICAT
8. 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

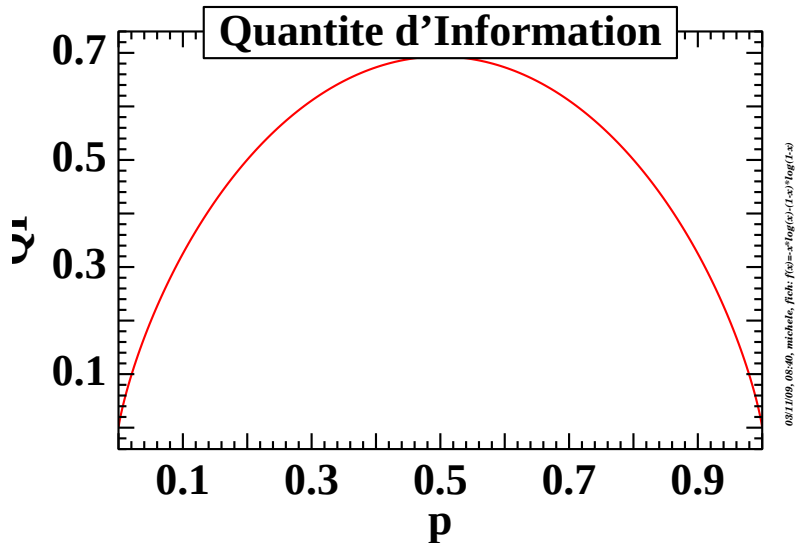
$$\begin{aligned} p &= Pr(Class = 1 | att = val) \\ I([att = val]) &= -p \log p - (1-p) \log (1-p) \\ I(att) &= \sum_i Pr(att = val_i) I([att = val_i]) \end{aligned}$$

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

Quantite d'information



Calcul

value	p(value)	p(poor | value)	QI (value)	p(value) * QI (value)
[0,10[	0.051	0.999	0.00924	0.000474
[10,20[	0.25	0.938	0.232	0.0570323
[20,30[	0.26	0.732	0.581	0.153715

Arbres de décision, 4

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

Complexité

Quantité d'information d'un attribut

$$n \ln n$$

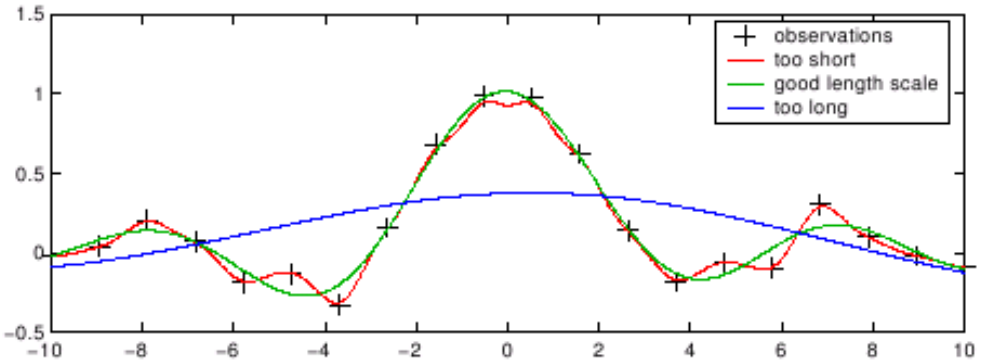
Pour construire un noeud

$$D \times n \ln n$$

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
  erreur empirique/généralité de l'arbre
  (biais / variance)
- Comment : Validation croisée



Validation croisée

param = paramètres de l'algorithme

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. Le cas le plus simple
4. Quel est le critère ? théorie
   Suffisamment compliqué, mais pas plus
5. Arbres de décision
6. Valider pratique
7. Etude de cas: SKICAT
8. Etude de cas: Stanley

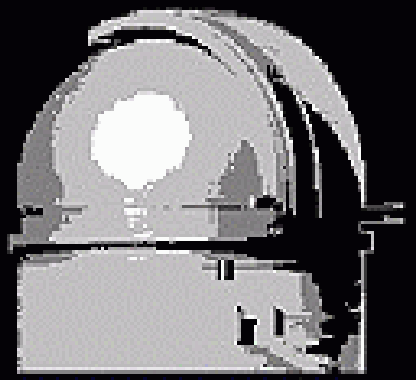
Skicat

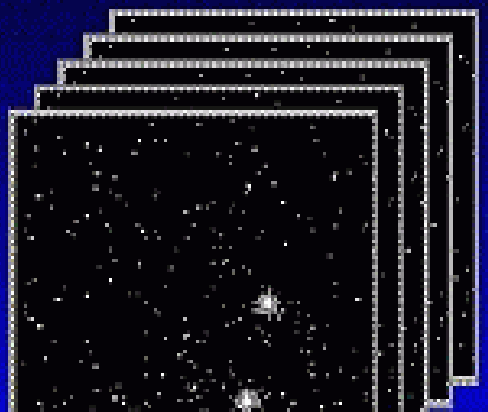
U. M. Fayyad, S. G. Djorgovski, and N. Weir. 1996
Jet Propulsion Lab., Caltech

- 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





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 ?

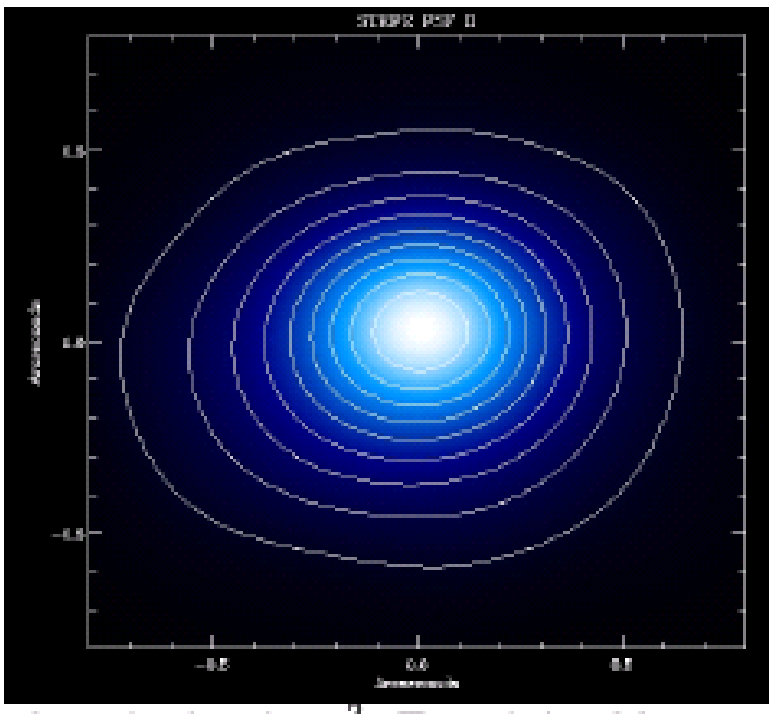
3 Téra bytes.

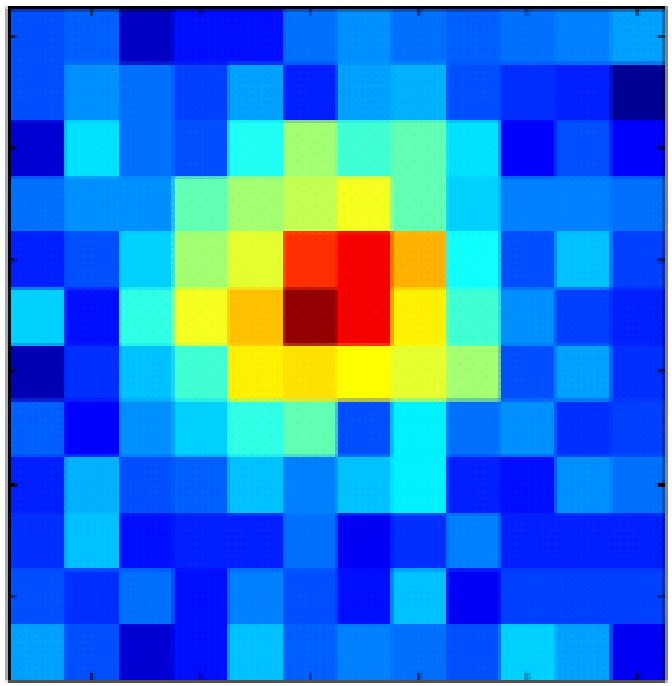
tera Terrorbytes

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

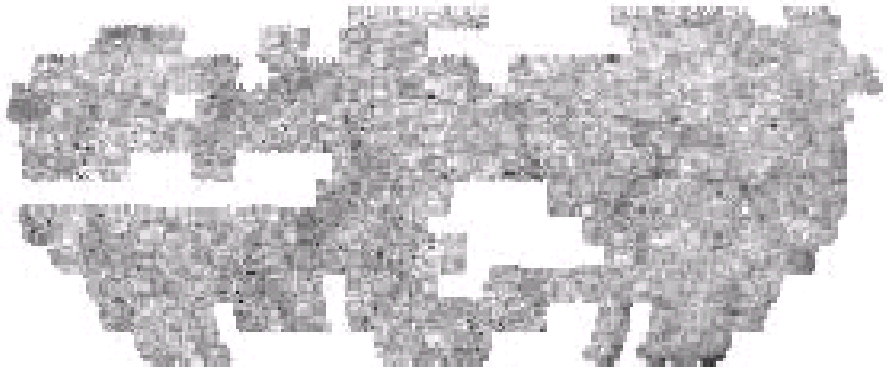
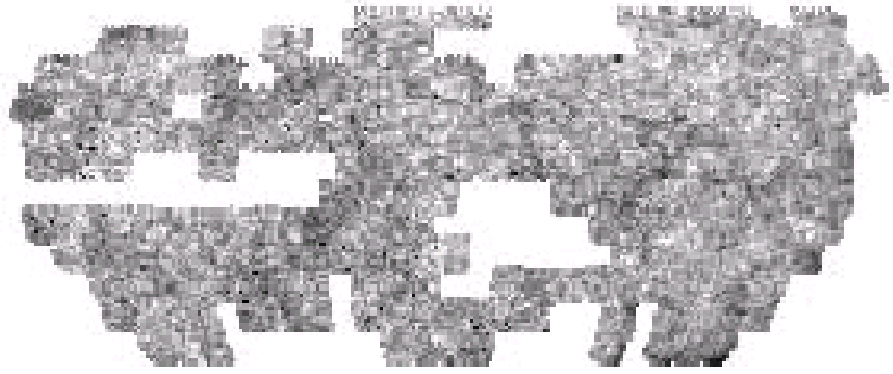
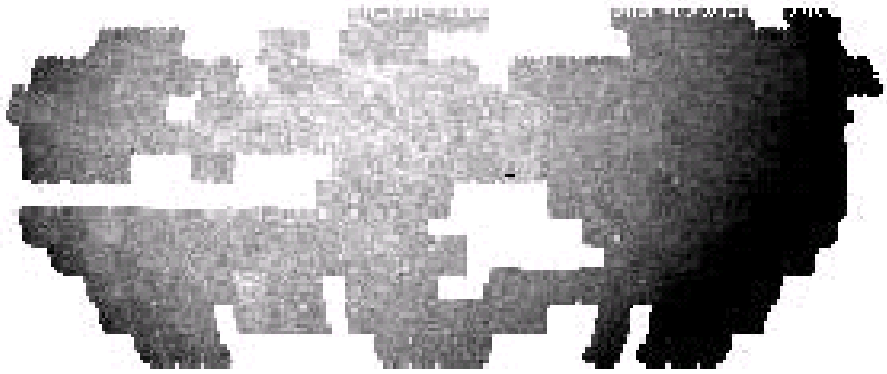
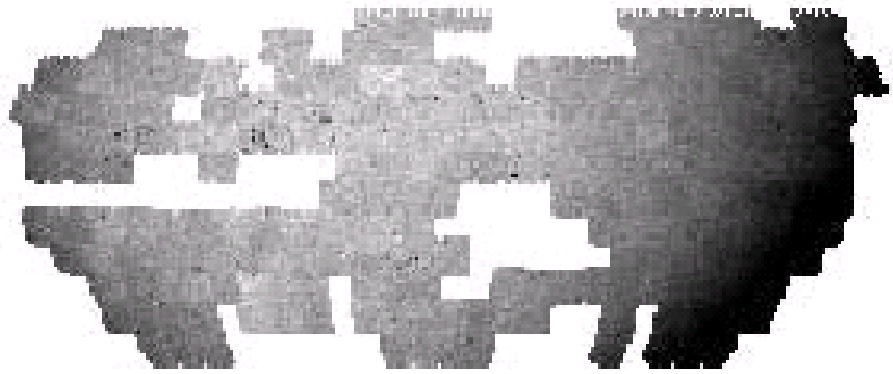
entre géologistes

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

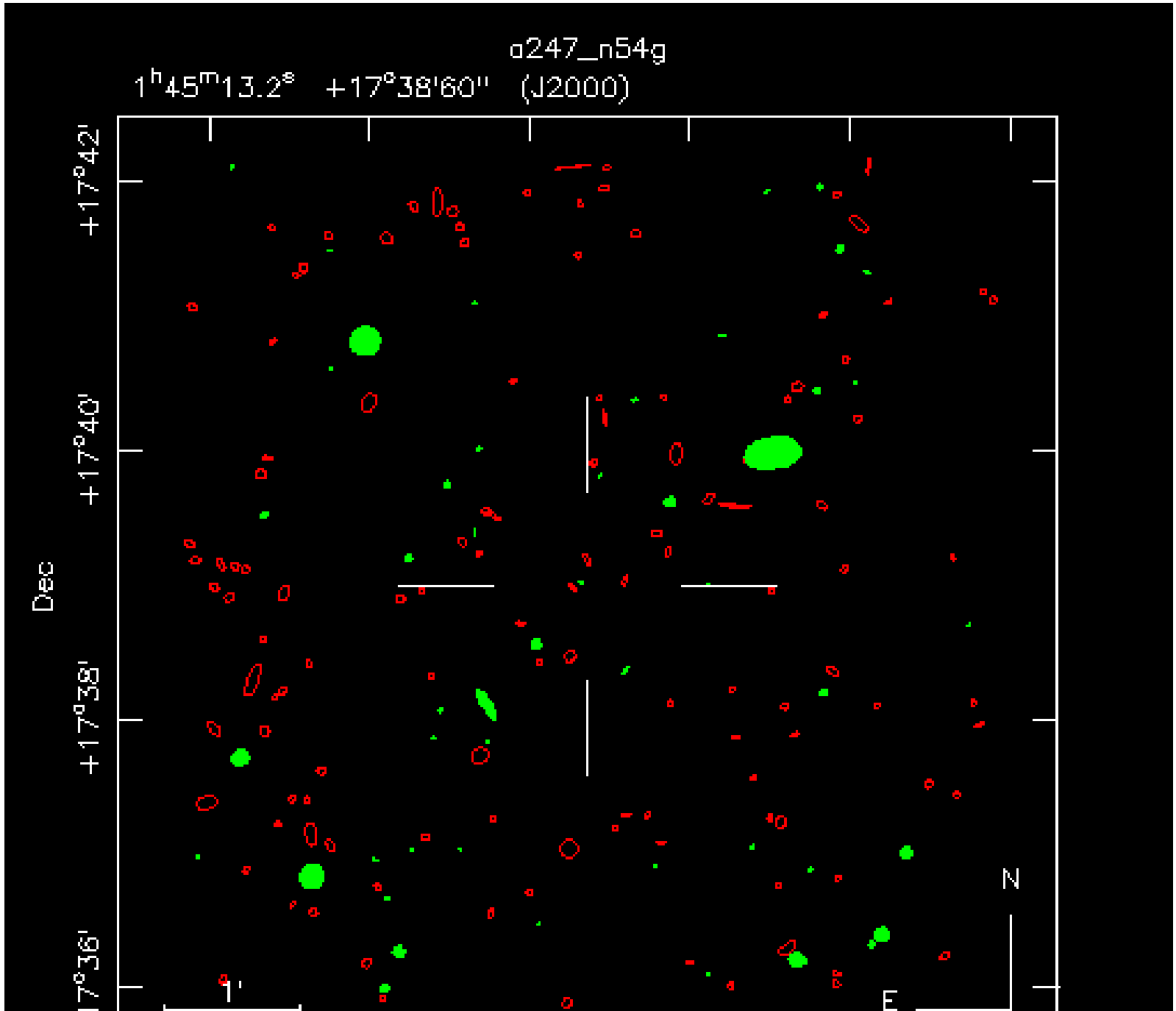
Classified objects: one magnitude fainter than previously

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)
+17°42'
+17°40'
+17°38'
+17°36'
Dec
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

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

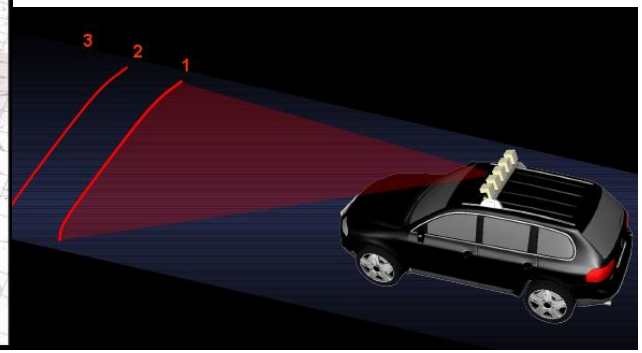
DARPA Challenge - 2005

Navigating: Perception is the issue

The 2005 Darpa Challenge

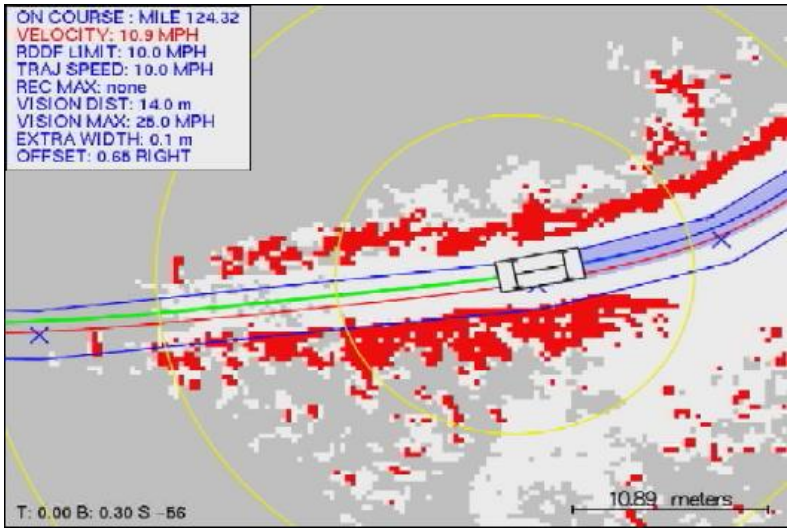
The terrain

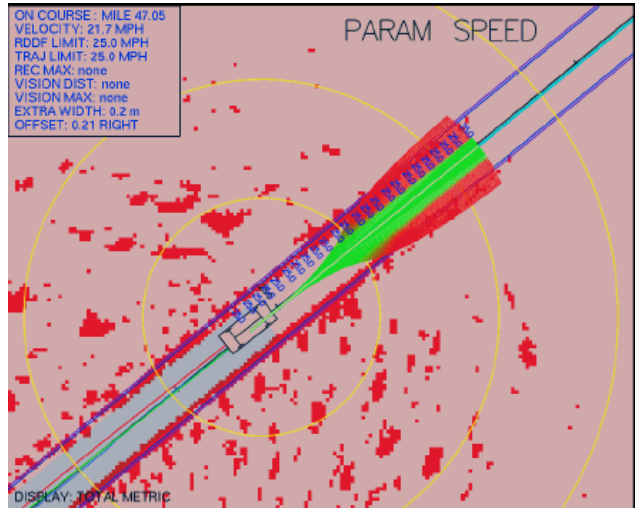
The sensors



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`

Leçons

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

$\Rightarrow$ contiennent des motifs réguliers.

besoin de mesures de qualité et test d'hypothèses.