Master Sciences de l'information et des Système 1ère année

Présentation de l'option 12 :

Transcription de documents numériques

Hervé GLOTIN

glotin@univ-tln.fr

http://glotin.univ-tln.fr

SIS, USTV

Contexte général

Les programmes d'intelligence artificielle sont aujourd'hui capables de reconnaître des commandes vocales, d'analyser automatiquement des photos satellites, d'assister des experts pour prendre des décisions dans des environnements complexes et évolutifs (analyse de marchés financiers, diagnostics médicaux), de fouiller d'immenses bases de données hétérogènes, telles les innombrables pages du Web ?

Pour réaliser ces tâches, ils sont dotés de modules d'apprentissage leur permettant d'adapter leur comportement à des situations jamais rencontrées, ou d'extraire des lois à partir de bases de données d'exemples.

Transcription de documents numériques

Pré-requit : néant
Usage des méthodes acquise en SSI_1 « Imagerie Numérique »
Dans la continuation de l'option L52 de la licence d'informatique

Cette option ouvre des portes pour :
- L'option M1 2nd semestre
« Recherche d'information dans les documents Multimédias »

- le M2 PRO Réseau de l'USTV option « Biométrie et accès sécurisé »

- le M2 Recherche de Marseille option « Recherche d'information »

- Toutes applications multimédia, système d'information

Répartition = 20h de cours, 16 de TD, 24h de TP

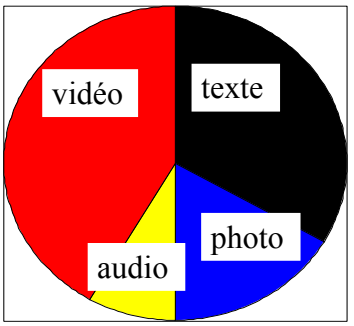
Note finale = 50 % note de TP, 50% note d'examen

Objectif :

- Face à l'augmentation exponentielle de la quantité de données multimédias :
  - Images fixes : presse, médecine, satellites…
  - Archives vidéos
  - Base de texte
- But de l'option : étudier les méthodes *efficaces* d'exploitation du contenu de ces bases
- L'indexation humaine étant trop coûteuse et tributaire de la subjectivité de l'indexeur.

Exemple de répartition des médias : l'Encyclopédia Universalis numérique

Média	unités	Go
texte	50 M mots	2
photo	15 K photos	1
audio	2 h	0.5
Vidéo	4 h	2.5



DYNAMIQUE / STATIQUE

Plan du module

- 1/ Introduction
- 2/ Les fondements de l'apprentissage
- 3/ Environnement méthodologique
- 4/ Apprentissage par optimisation
  Réseaux connexionistes / RNN
- 5/ Application à la parole
- 6/ Application aux images
- 7/ Perspectives Biométriques

Chapitre 1

Introduction générale

Les défis

- Traiter vite de grands ensembles de documents
  - Bien exploiter le couplage entre les médias
  - Combler l'écart entre signal et sémantique
- But : travailler avec des tailles réalistes :
  - 100 K à 10 M images, en contexte bien défini.
  - 100 h de radio
  - 100 h de film…
- Et sur un monde ouvert comme le web ?...

Structuration de texte et d'image fixe

- Description conjointe texte-image :
  - donner une sémantique aux objets contenus dans l'image
  - et interpréter l'ensemble de la scène…
- Problème de dimension : couleurs, textures, formes…
- Applications :
  - recherche d'images
  - désambiguïser les textes

Structuration de l'audio

- Détection des événements sonores
  - musique / parole / Jingle…
  - Identification de locuteur
  - Reconnaissance de locuteur
  - Suivi de locuteur
- Transcription de la parole
- Structuration temporelle
- Thématisation

Quelques Méthodes

Traitement du signal
-Extraction de traits robustes
-Séparation de sources
-Fusion de capteurs

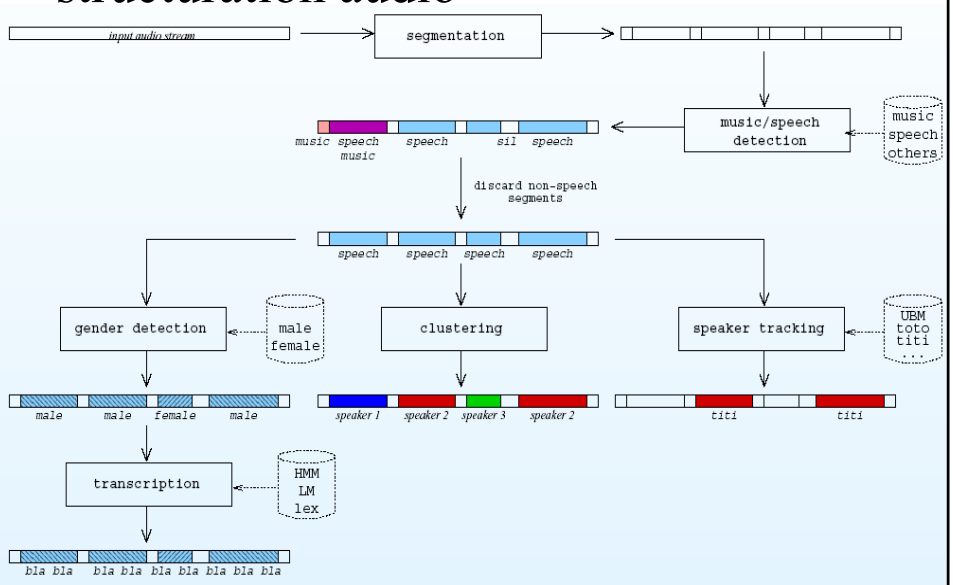
Apprentissage
- MG HMM
- Réseaux de neurones
- ANN récurrent
- Modèles Probabilistes

Fusion de Données MM & Recherche d'information
- Approche Multi-flux
- TAL
-RI textuelle

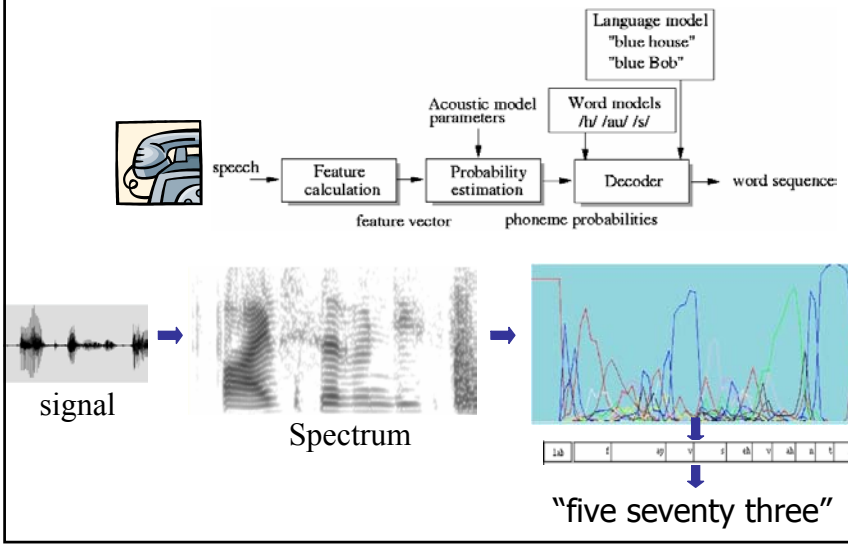
Applications à la reconnaissance automatique de la parole (RAP)

- Principes de base

Architecture d'un système de structuration audio



Automatic speech recognition



PROJET ESTER 2003 2005.
Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques

Équivalent francophone de la campagne de la DARPA, relayé par NIST pour les systèmes de transcription d'émissions de 1996 à 1999.

Intégration d'informations annexes, pour enrichir la transcription :
- extraction automatique de contenu (Automatic Content Extraction, 1999-2002),
- transcriptions automatiques d'émissions,
- détection de thème (Topic Detection and Tracking, 1998-2002).

+ vérification du locuteur et la segmentation selon le locuteur.
+ transcriptions enrichies (Rich Transcription, 2002-2003) :
enrichir la transcription avec des informations concernant le locuteur.

MINISTERE DE LA DEFENSE

DGA technolangue AFCP

Structuration XML d'audio : exemple sur RFI (50 h homme / 1h)

- **<?xml version="1.0"?>**
- **<!DOCTYPE Trans SYSTEM "trans-13.dtd">**
- **<Trans scribe="MTU" audio_filename="fint981207:0700" version="29" version_date="990706" xml:lang="fr">**

- **<Topics>**
- **<Topic id="to1" desc="Les titres du journal"/>**
- **<Topic id="to3" desc="Deux Incendies"/>**
- **…**
- **<Topic id="to41" desc="Point Circulation"/>**
- **</Topics>**

- **<Speakers>**
- **<Speaker id="sp1" name="Nicolas Stoufflet" check="yes" type="male" dialect="native" accent="" scope="global"/>**
- **…**
- **<Speaker id="spk52" name="journaliste 2" check="no" type="female" dialect="native" accent="" scope="local"/>**
- **</Speakers>**

- **<Episode program="France Inter" air_date="981207:0700">**
- **<Section type="filler" startTime="0" endTime="9.632">**
- **<Turn startTime="0" endTime="1.5" speaker="sp1">**
- **<Sync time="0"/>**
- **Patricia Martin , que voici , que**
- **<Event desc="top" extent="instantaneous"/>**
- **voilà !**
- **</Turn>**
- **<Turn speaker="sp53" startTime="1.5" endTime="2.624">**
- **<Sync time="1.5"/>**
- **oh , bonjour**
- **<Event desc="top" extent="instantaneous"/>**
- **Nicolas Stoufflet .**
- **</Turn>**
- **<Turn speaker="sp1" startTime="2.624" endTime="3.765">**
- **<Sync time="2.624"/>**
- **France-Inter**
- **<Event desc="top" extent="instantaneous"/>**
- **, 7 heures .**
- **</Turn>**
- **…**
- **<Background time="10.133" type="music" level="high"/>**
- **lundi 7 décembre . deux incendies cette nuit en région parisienne , dans une maison de retraite de Livry-Gargan en Seine-Saint-Denis ,**

Applications à la reconnaissance automatique de la parole Audio-visuelle

- Principes de base
- Analyse du flux visuel

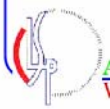
Segmentation des visages par analyse formes & couleurs

(kruppa et al, 2001)

REGION OF INTEREST EXTRACTION

- Face detection and mouth location estimation:
  - Statistical face detection and 26 facial feature localization algorithm (A.W. Senior, IBM Research).
  - Multi-scale (image pyramid) search for faces.
  - Uses Fisher discriminant, prior facial part collocation statistics.
  - Requires training (on about 2000 annotated frames).
  - Face detection acc. = 99.7 %; facial feature acc. ≈ 90 %.
- ROI extraction:
  - Obtain smoothed mouth center and size estimates.
  - Extract a 64 × 64 pixel, size normalized ROI.

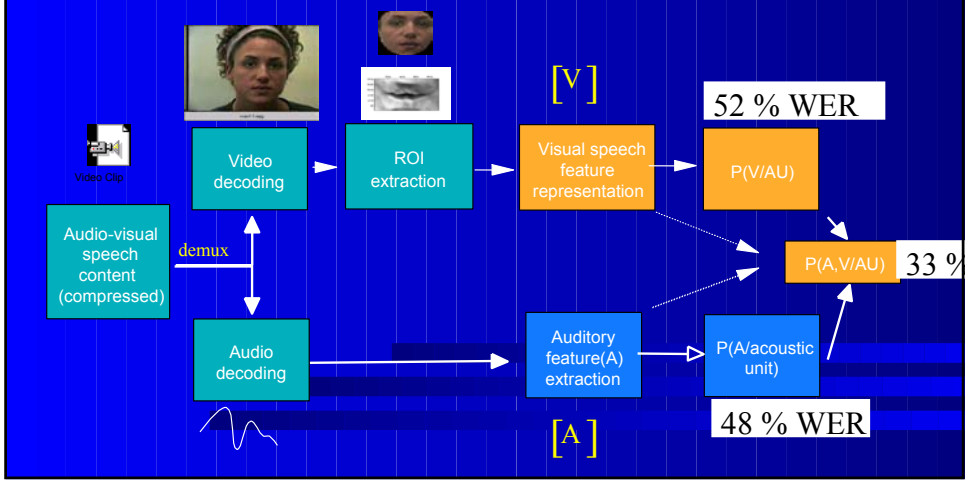
  - ROI failure example:



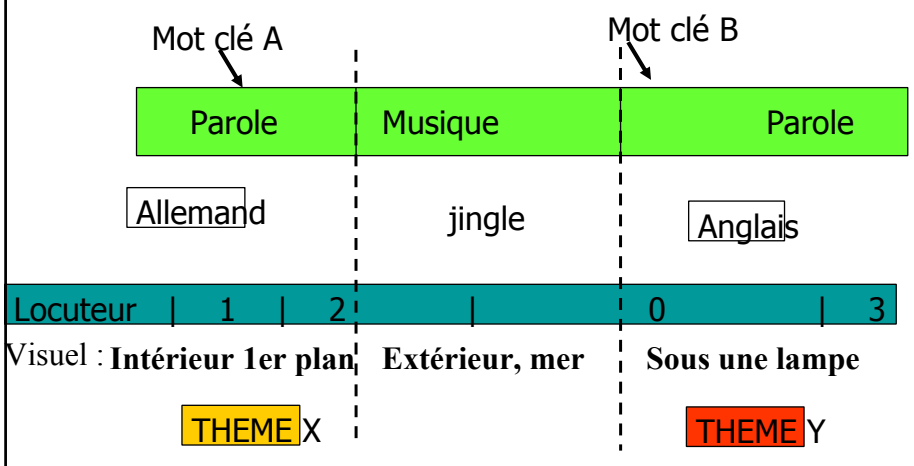
AVSR WS'00

Audio-Visual speech recognition
AV. VIA VOICE Project – IBM 2000

35 h training / 3h test / 10 K mots / 10 dB SNR speech noise



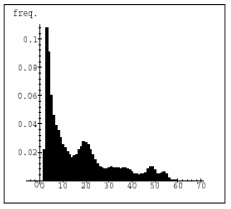
xemple de structuration AV



Applications à la reconnaissance automatique d'images

- Principes de base

Exemples de traits visuels simples

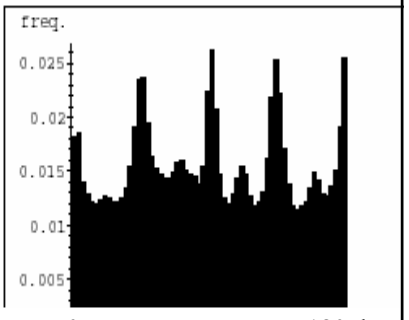


(b)

Histogramme des gris

CANNY (1986)

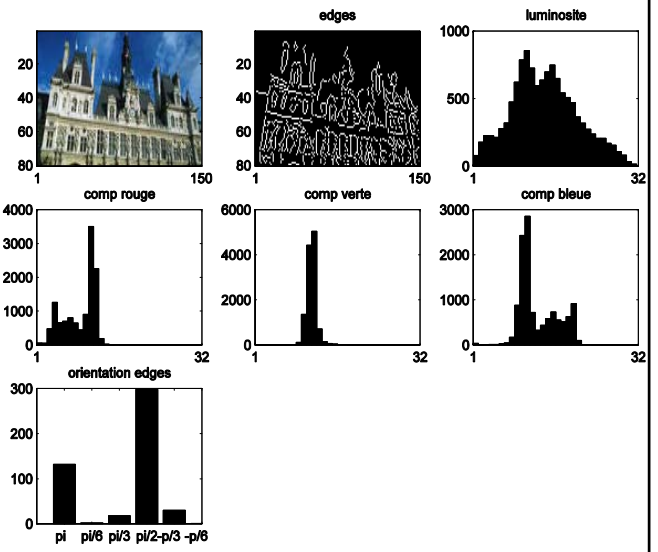
Histo. des directions



(Séléction hautes fréquences)

Exemple de traits

Indexées visuellement par les histogrammes rouge, vert, bleu, luminance et direction (« low level features »)

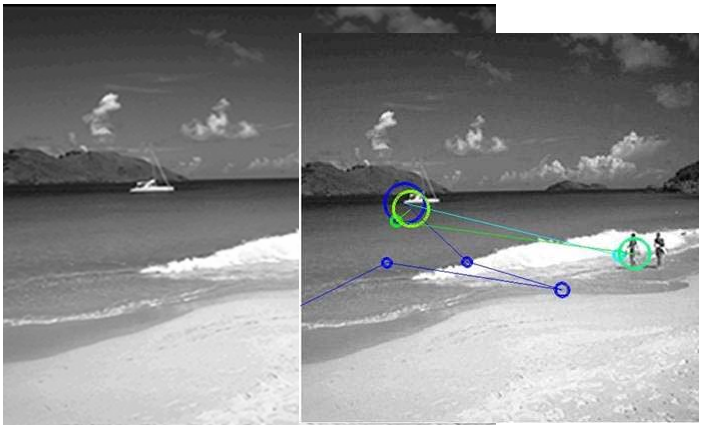


Exemple d'application des Régions d'Intérêt (ROI) lecture de textes sur des films

- Recherche automatique de cadres textuels
  - texture-histogrammes-bordures-dynamique
  - Analyse par parties
- Rehaussement du signal contenu dans les cadres
  - Rehaussement contraste …
- => Reconnaissance sous OCR du commerce

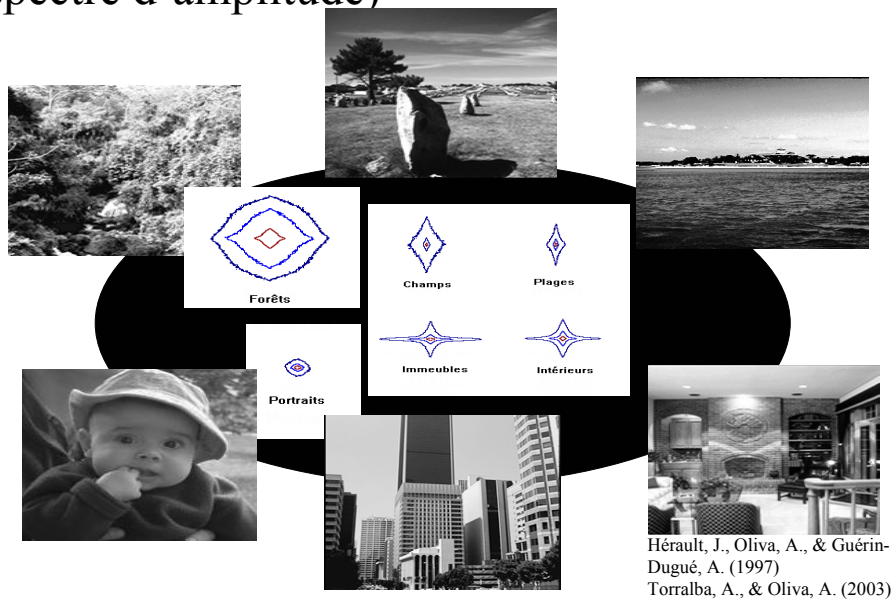
Exemple d 'extraction de texte [Démo IDIAP 2000-2002 www.idiap.ch]

Exploration des scènes naturelles



Est-il possible de prédire les régions d'intérêt à partir des particularités du signal ?

Catégorisation des scènes naturelles (spectre d'amplitude)



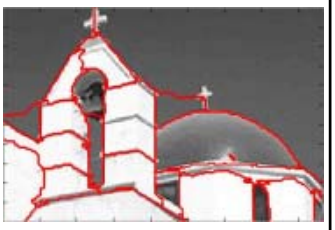
Hérault, J., Oliva, A., & Guérin-Dugué, A. (1997)
Torralba, A., & Oliva, A. (2003)

Exemples de segmentation avec differents algorithmes

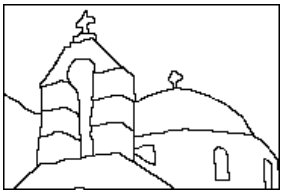
C-means (Bezdek, 1981) / Single-Link (EPRI, 1999) / Graph partition segmentation (Malik, 2000) / Combined Genetic segmentation (DiGesu2002)

CGS

GPS



Human



C-means

Single-link

CGS

GPS

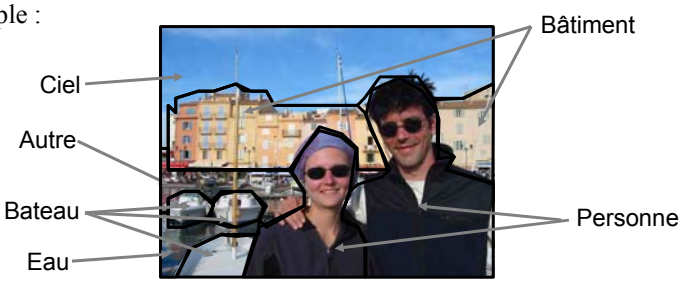
Human

C-means

Single-link

Modèle de représentation d'image

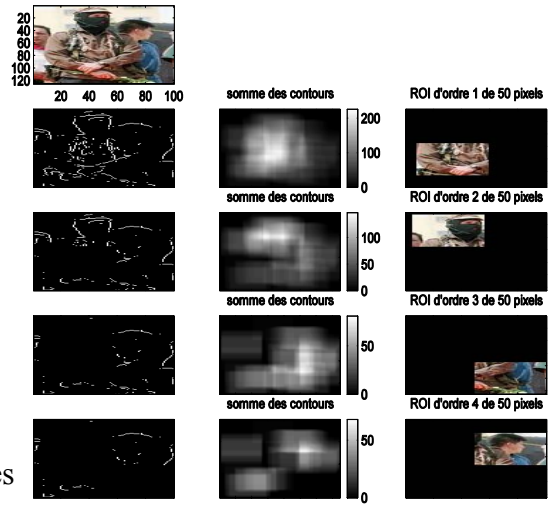
- Segmentation : subdivision de l'image en régions $\{r_i\}$ (*m* régions)
  - Les $\{r_i\}$ forment une partition de l'image
  - Les régions correspondent aux objets réels
- Ensemble de concepts $\{c_j\}$, étiquettes des régions
  - Correspondance région-concept : $\Phi(r_i) = c_j$
  - Exemple : {Personne, Bâtiment, Végétation,…}
  - Objets images non pris en compte = concept 'Autre' (~ mot outil du texte)
- Exemple :



Autre type de segmentation

Segmentation en 4 régions d'intérêts pour éliminer le bruit de fond de l'image

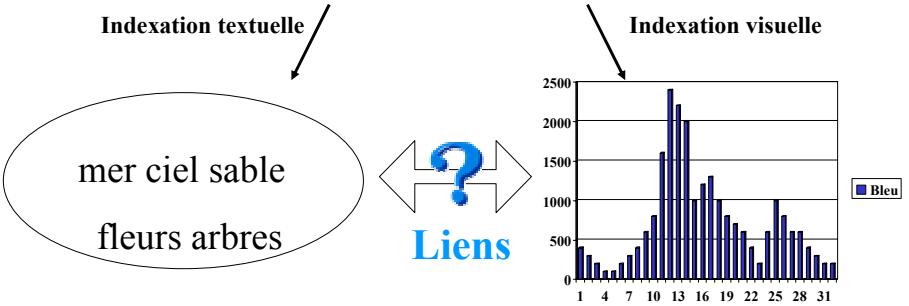
Chaque région d'intérêt possède les 5 mêmes types d'histogrammes que les images entières



APPLICATIONS GENERALES

- RECHERCHE d'INFORMATION
- AUTO-ANNOTATION du son, d'images, de videos…
- IDENTIFICATION
- BIOMETRIE

AUTO-ANNOTATION Problématique



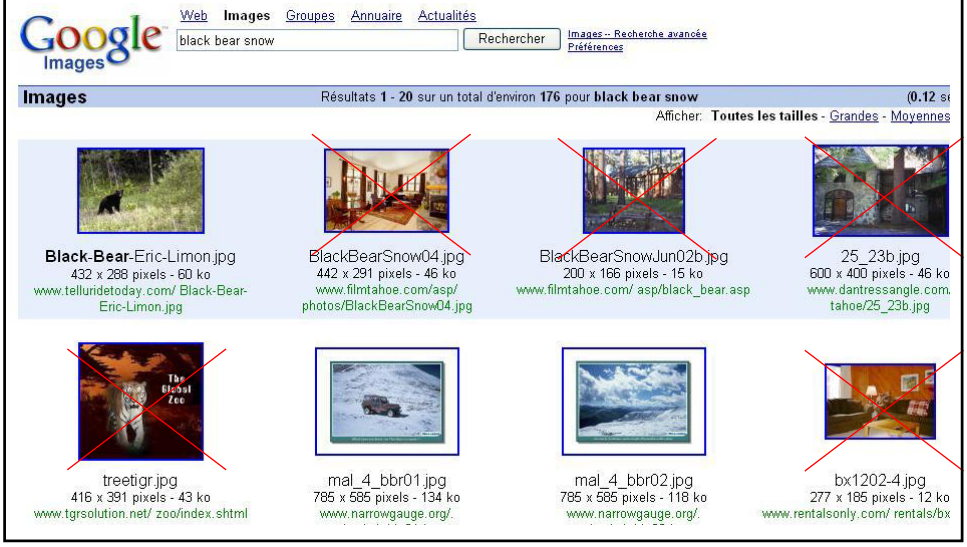
Indexation d'images pour la recherche d'images sur le web

- Indexation textuelle
  - Manuelle : coûteuse, subjective
  - Automatique à partir du nom, de la légende ou du texte entourant l'image : ne décrit pas le contenu de l'image, beaucoup d'erreurs d'indexation car ce texte ne correspond que rarement avec l'image
- Indexation visuelle
  - Couleurs, formes, textures
  - Localisation, régions d'intérêts, segmentation
  - Extraction de la sémantique difficile !

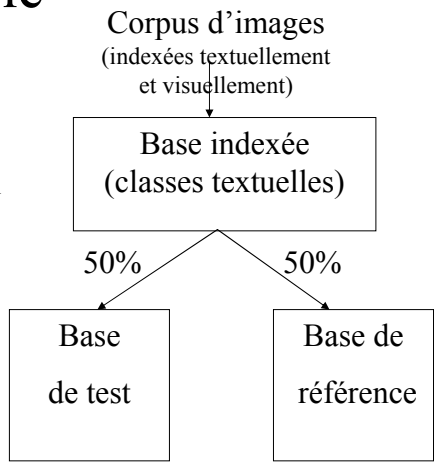
Les systèmes de recherche d'images actuels

- Recherche d'images par des requêtes formées de mot-clés
- Recherche d'images par le contenu d'une image requête
- Peu de systèmes qui utilisent les liens entre un mot et sa (ses) représentation(s) visuelle(s) pour améliorer l'indexation et la recherche d'images

Comment raffiner les résultats d'une recherche d'images ?



Protocole



Étape A

Classer les images à partir de l'indexation textuelle

Étape B

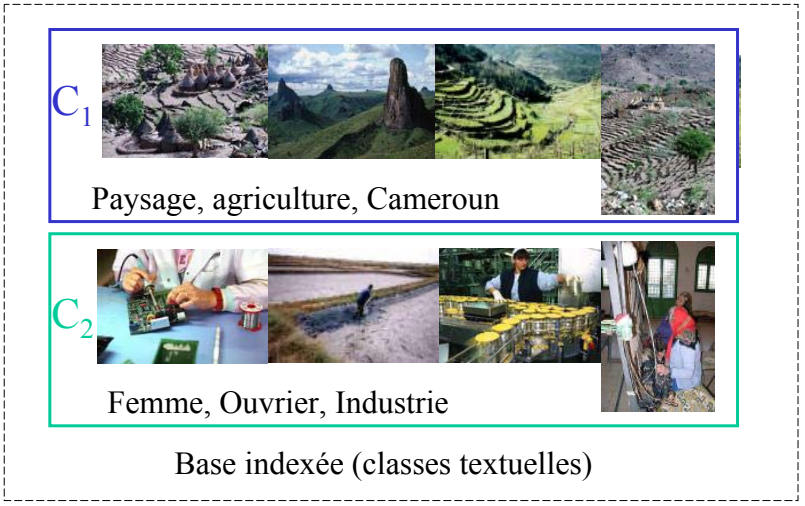
Diviser aléatoirement en deux bases

Étape C

Reclasser les images de la base de test par rapport à l'indexation textuelle, à l'indexation visuelle et par fusion des classifications visuelle et textuelle

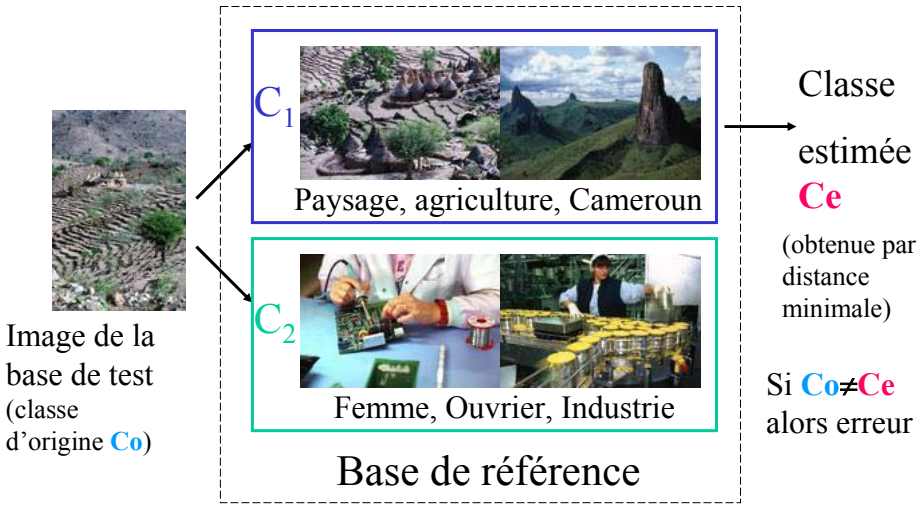
Étape A

Exemple de classes obtenues par CAH sur les vecteurs textuels



Étape C

Classe une image de la base de test



Étape C

Les classifications

1. Classification textuelle pure
2. Classification visuelle pure
3. Classification par fusion des classifieurs visuels et textuels

Résultats de la classification visuelle pure

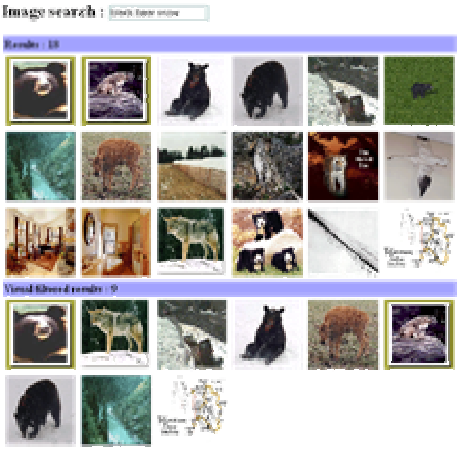
Traits visuels	Taux d'erreurs
Rouge	71.76 %
Vert	76.07 %
Bleu	77.25 %
Luminance	76.07 %
Direction	76.86 %
Aléatoire	91.6 %

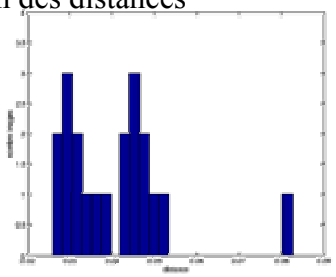
Résultats du rehaussement visuo-textuel

Résultats	Textuelle sans thésaurus	Fusion visuo-textuelle	Gain
Taux d'erreur	13.72%	6.27%	+54.3%

Exemple d'application

Recherche textuelle « classique » sous Google, puis filtrage visuel des images par rapport à la distribution des distances





Distribution des distances pour chacune des images de Google. Cette distribution est bimodale, ce qui permet de considérer que les images du premier mode (distances < 0.04) sont adéquates à la requête, les autres non.

Discussion

- Les résultats dépendent de la qualité du thésaurus.
- Le choix des traits visuels reste un problème ouvert.
- La mise en place d'un système de recherche utilisant ces méthodes posent des problèmes de stockage et d'accès à l'information, notamment par rapport à la segmentation visuelle des images.

Comment améliorer l'indexation automatique ?

Exemple d'indexation automatique

« Notre hôtel au bord de la mer ... »

↓ indexation textuelle

hôtel bord mer

Protocole pour associer automatiquement des mots à une image

Corpus d'images

(indexées textuellement et visuellement)

Base de test

Base de référence

ciel mer fleur arbre

ciel arbre

arbre fleur route

Estimation des probabilités

	Arbre	Ciel	Fleur	Mer
Rouge	2	1	2	1
Vert	3	2	2	1
Bleu	2	2	1	1

Classes visuelles

fusion

arbre (5fois), fleur (4 fois)

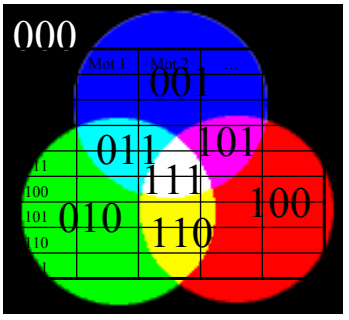
Construction de classes visuelles par méthode dichotomique

- Principe : séparer uniformément l'espace visuel
- Exemple : découpage de l'espace RGB
  - chaque dimension est séparé en deux

0 127 255

- il y a 2^3 = 8 classes

V={230,190,30} → 110



Classement par dichotomie des mots associés à l'image

Calcul du score de la classification

Image de la base de test

Mots de références de l'image	Mots trouvés par le système
fleur arbre lampadaire	arbre fleur ciel

Score : 2/3

Application : correction de l'indexation textuelle d'images

Mots de références de l'image	Mots trouvés par le système
horizon mountain palace tree	sky tree rock building grass

Résultats

- Expériences réalisées sur 10 000 images de Corel
  - 7000 images de référence
  - 3000 images de test

	Aléatoire	Priors (mots les plus fréquents)	LAB_RGS_RGB (18 dimensions)
Score	2 %	27 %	36 %

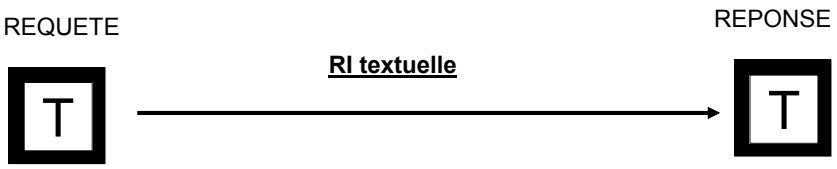
Conclusion

- On peut modéliser les liens entre l'indexation textuelle et visuelle afin de permettre :
  - le rehaussement d'une recherche par mot-clés d'images par leur contenu
  - l'amélioration de l'indexation par mot-clés d'images
- Méthode simple et automatique, donc utilisable sur le web
- Ces systèmes peuvent être utilisés avec n'importe quels types de traits visuels

Perspectives

- Application de ces méthodes de fusion sur les images du web : indexation, recherche par le contenu…
- Applications sur les flux dynamiques : audio, audio-visuel…
- Recherche d'information, identification, biométrie…

RI et Combinaisons d'opérateurs LEGO-AUDIO-VIDEO



Modèle vectoriel, TF / IDF

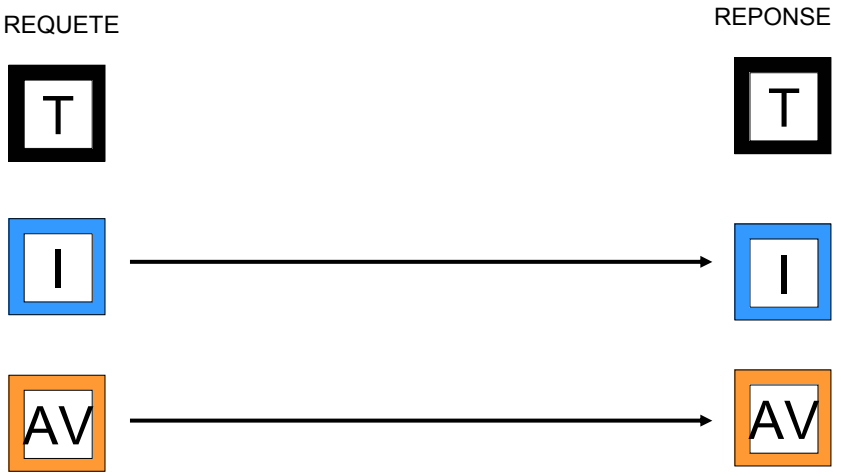
Séparation termes lexicaux qui ont un sens en eux-mêmes indépendamment du contexte des *termes grammaticaux*

A*nalyse morpho-syntaxique*

Difficultés inhérentes à la recherche
en langage naturel : synonymie, polysémie

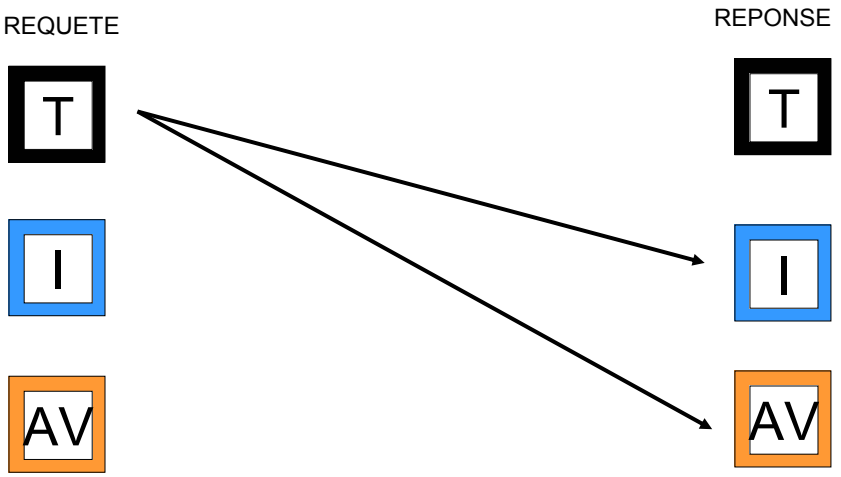
Rem : un texte comporte, en plus de sa fonction principale d'information, la fonction d'auto-indexation

Recherche par similarité

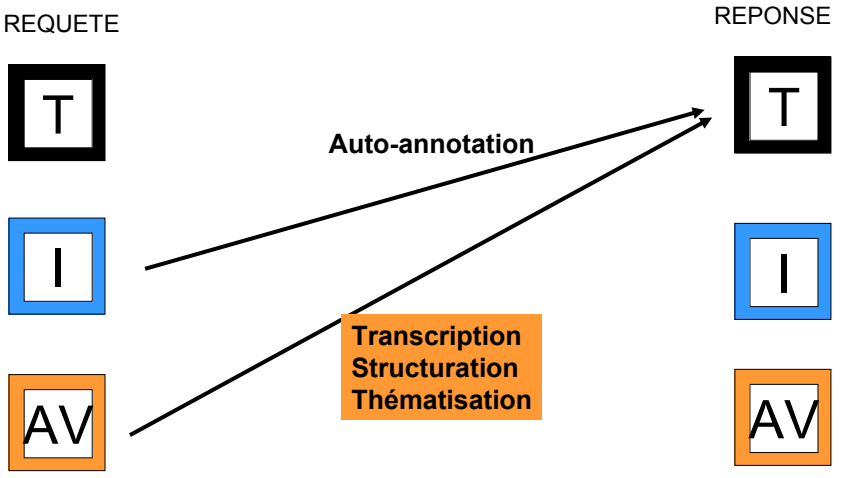


Désambiguisation de textes contenant des images:
Associer des textes qui contiennent des images semblables

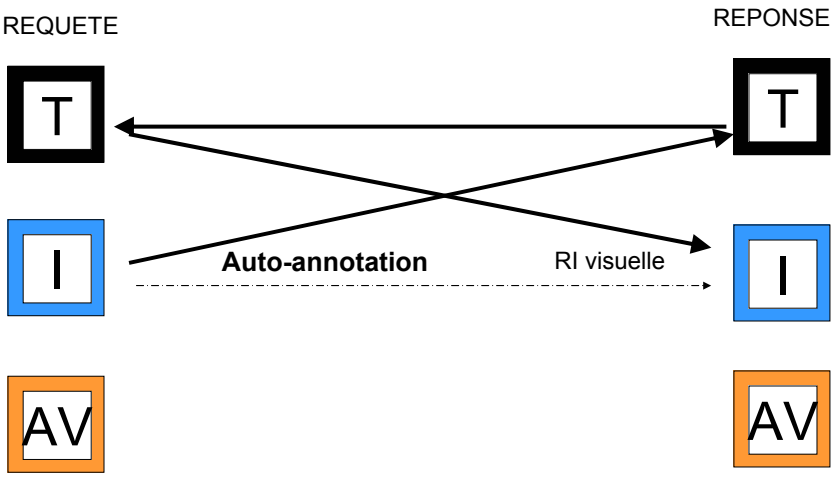
RI visuelle et/ou sonore



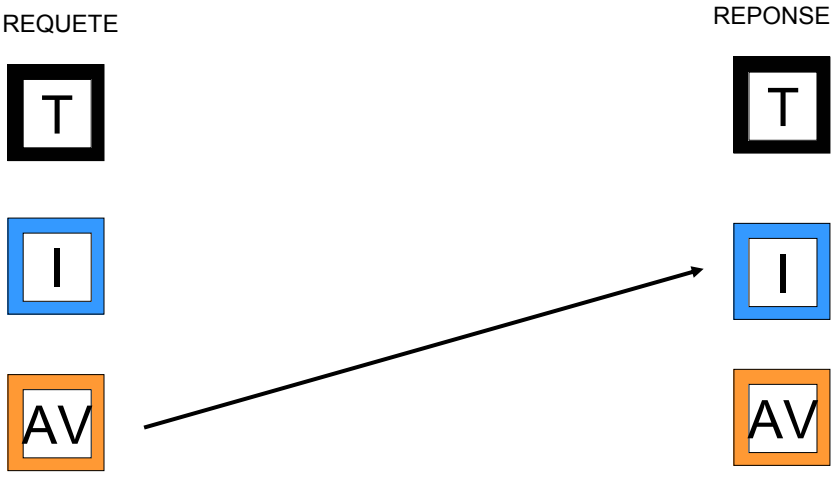
Auto-annotation / Transcription



Recherche d'image par similarité indirecte =
Auto-annotation + RI visuelle

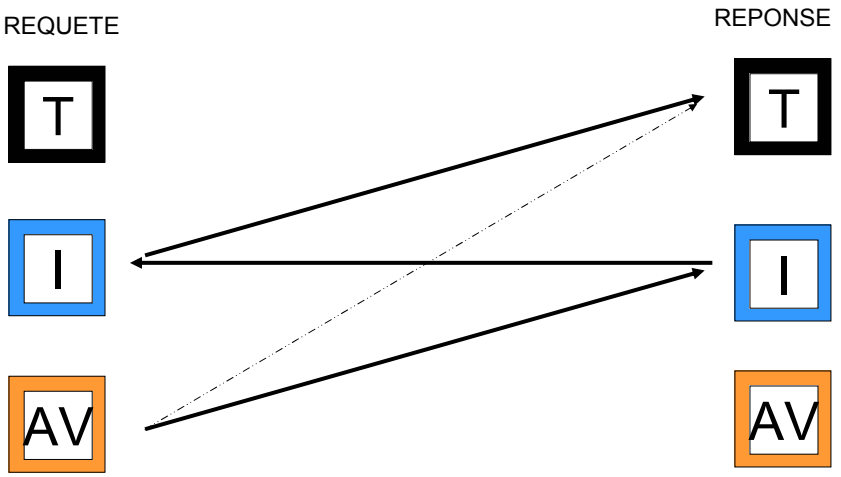


SEGMENTATION

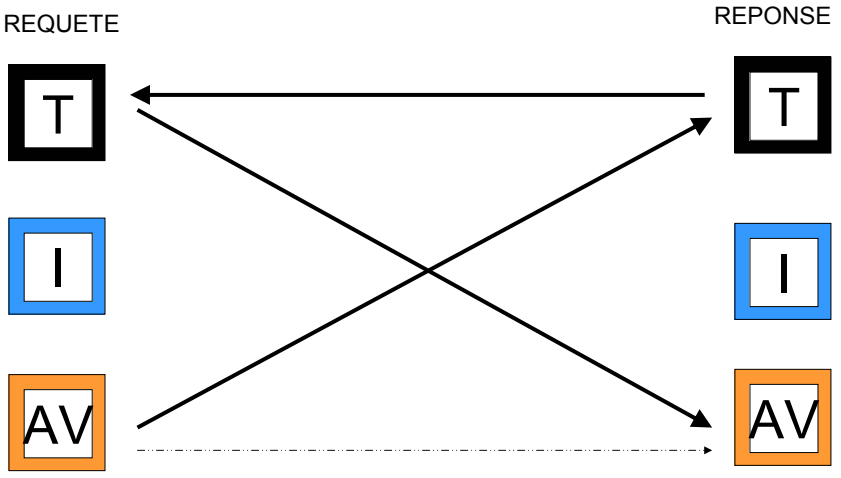


Segmentation par extraction de sous images fixes

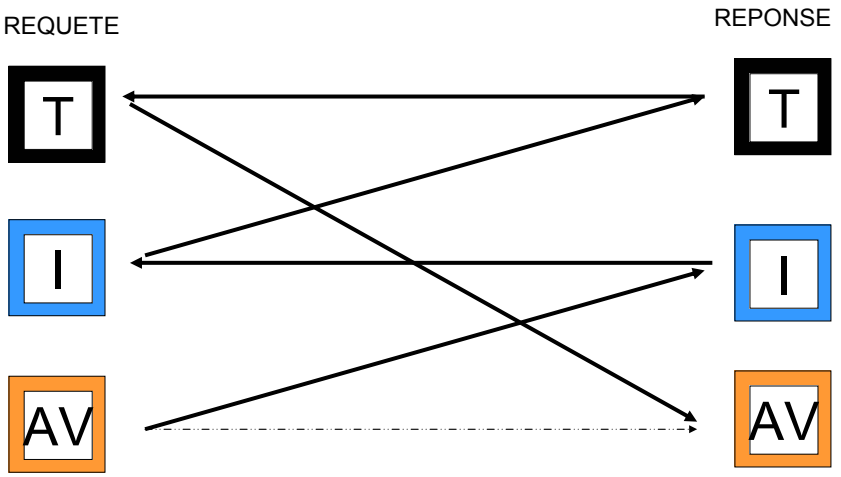
Transcription =
segmentation + auto-annotation

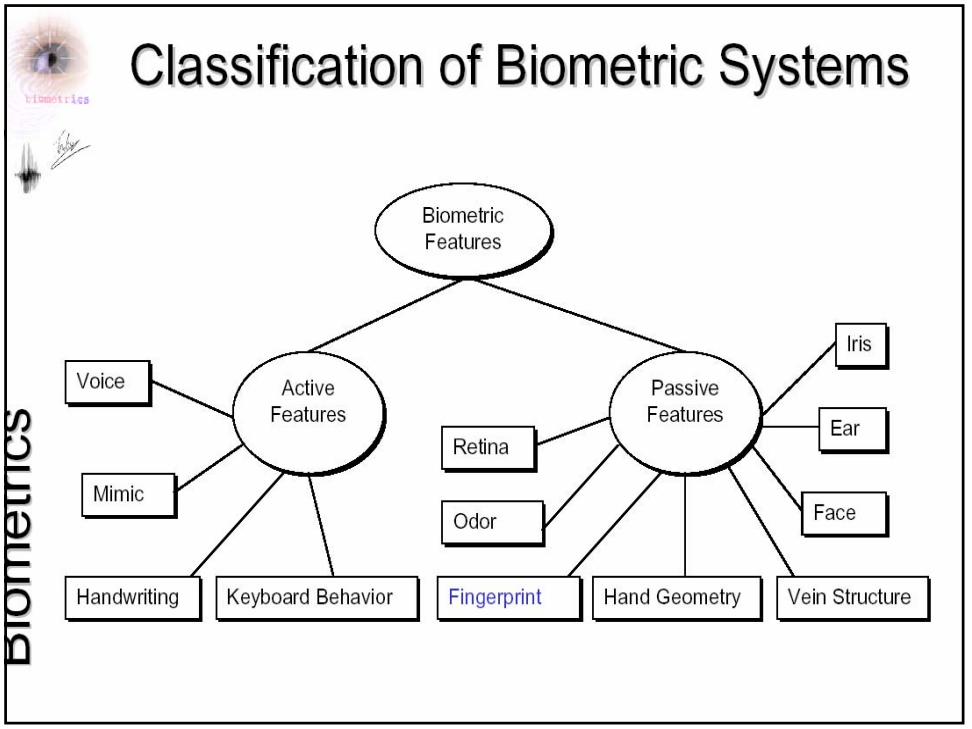


Recherche par similarité indirecte =
Transcription + RI



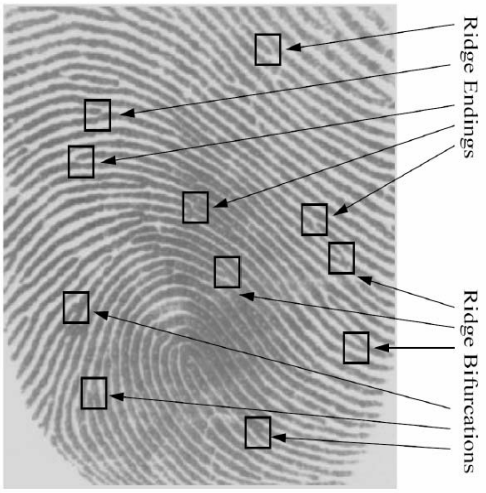
<u>RI par similarité</u> =
segmentation + auto-annotation + RI ?



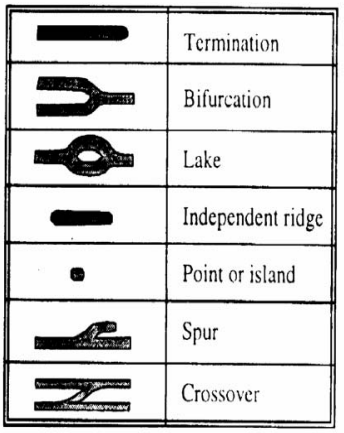


Minutiae features

(local level features)



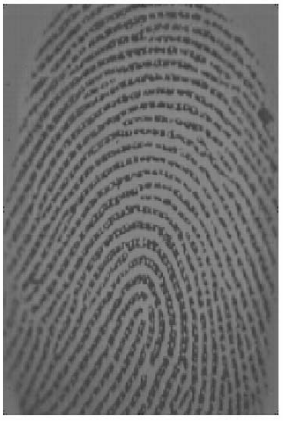
Hong, Wan, Jain



Jain et al

Biometrics

Quality Enhancement



?

Biometrics

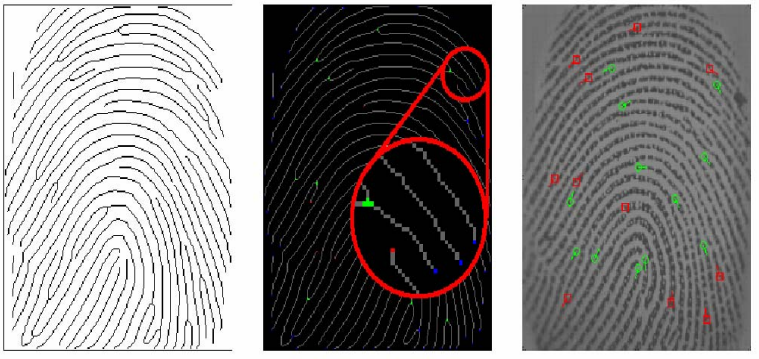
Thinning *ctd.*

Biometrics

Helmholz

Minutiae Feature Extraction *ctd.*

Biometrics



Helmholz

Chapitre 2

Fondement de l'apprentissage

Sources :

Duda & Hart
Cornuéjols & Miclet

Plan

1. Apprentissage naturel
   (a) Définition
   (b) Notions de base
   (c) La simplicité
2. Apprentissage artificiel
   (a) Définition
   (b) Taxonomie
   (c) Les paramètres d'un programme d'apprentissage
   (d) L'induction et les compromis nécessaires
   (e) Une liste raisonnée de concepts et d'algorithmes

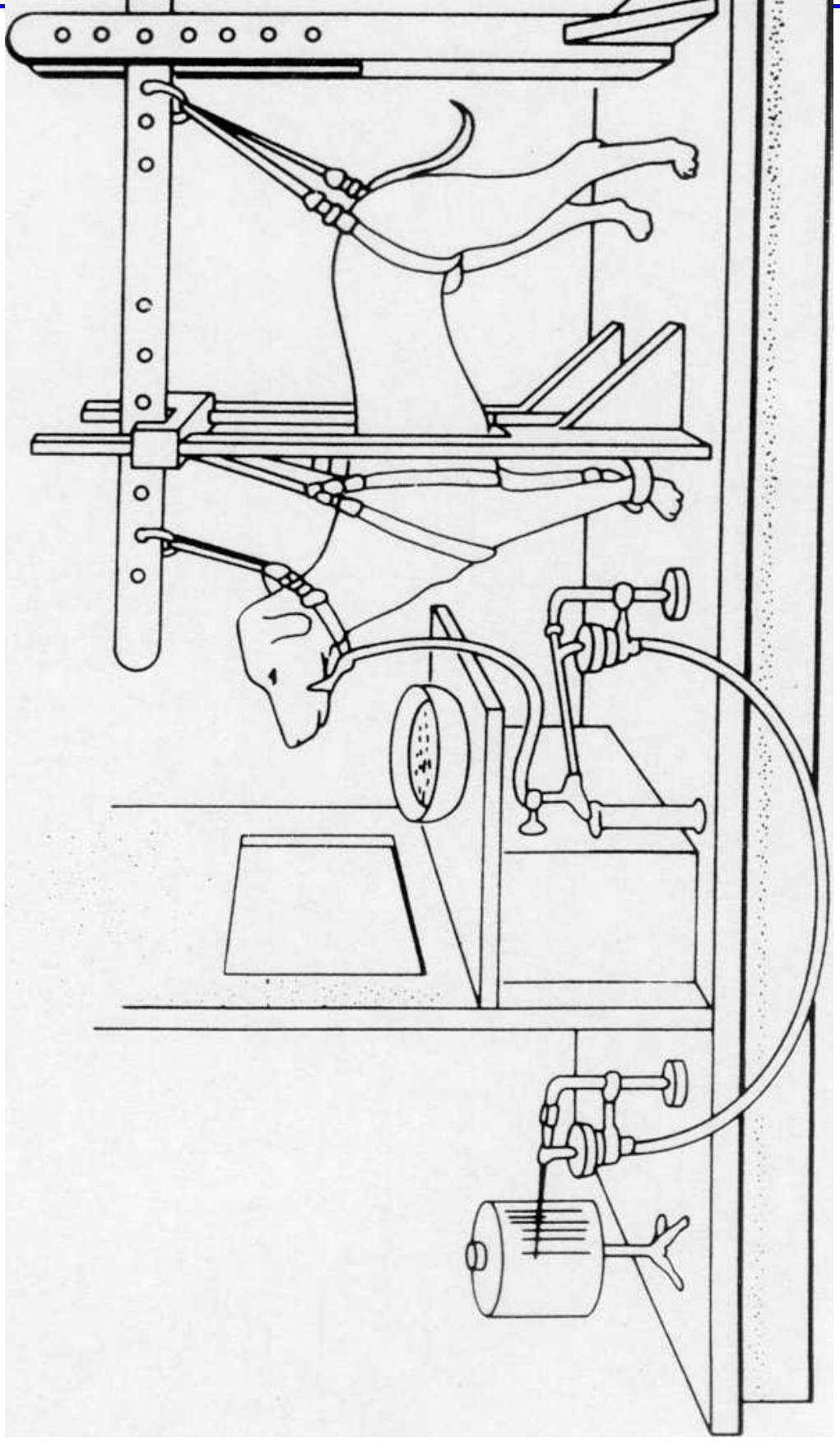
1. (a) : Apprentissage naturel: une définition

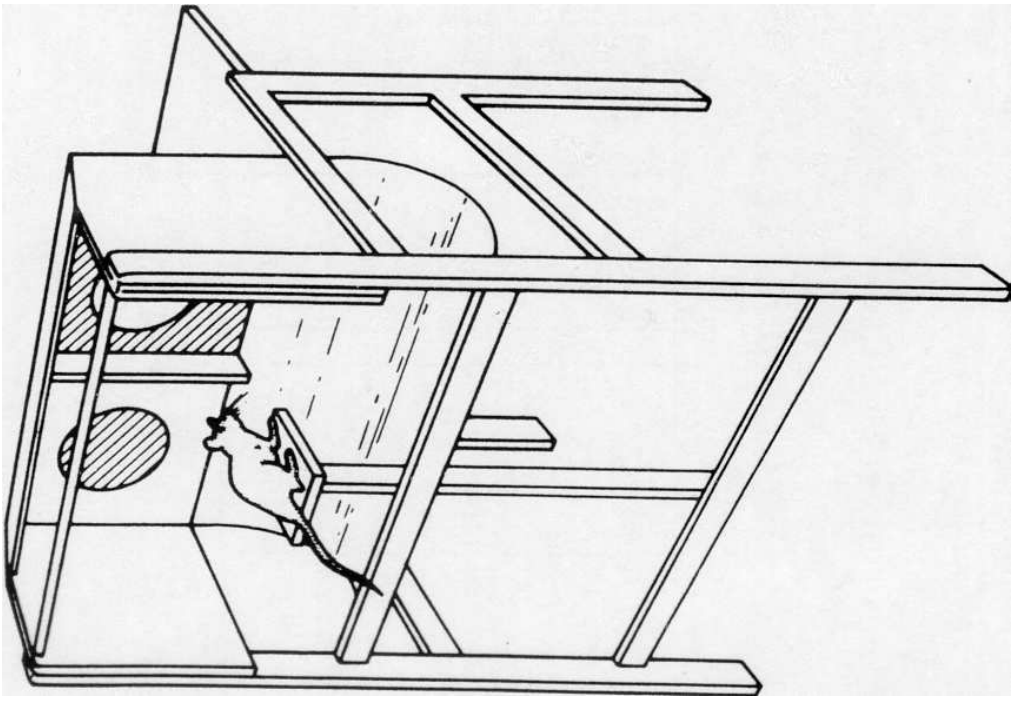
L'apprentissage est une **modification durable** des **potentialités de comportement** résultant d'une **interaction répétée** avec l'environnement.

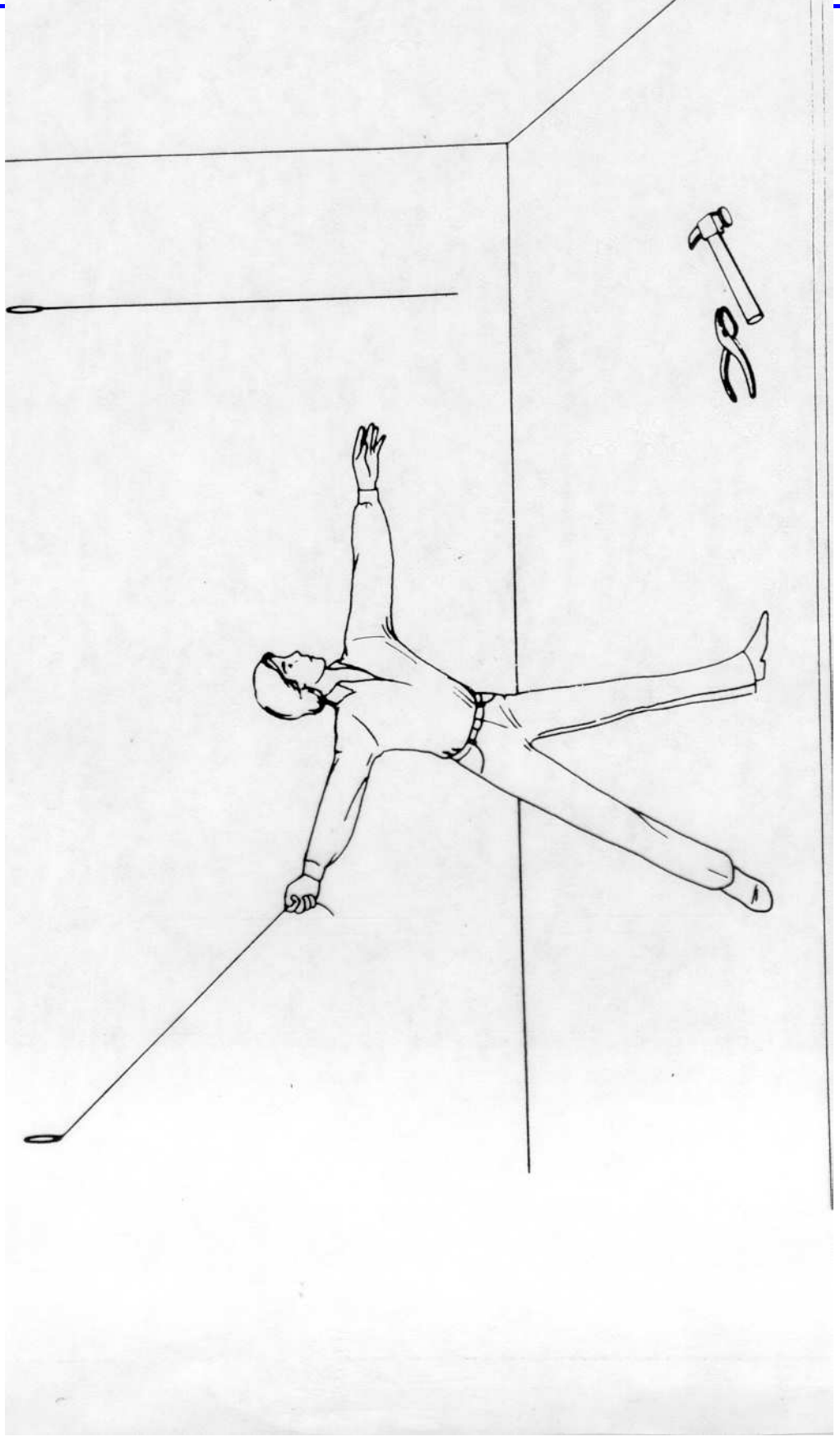
1.(b) : Quelques notions de base en psychologie expérimentale

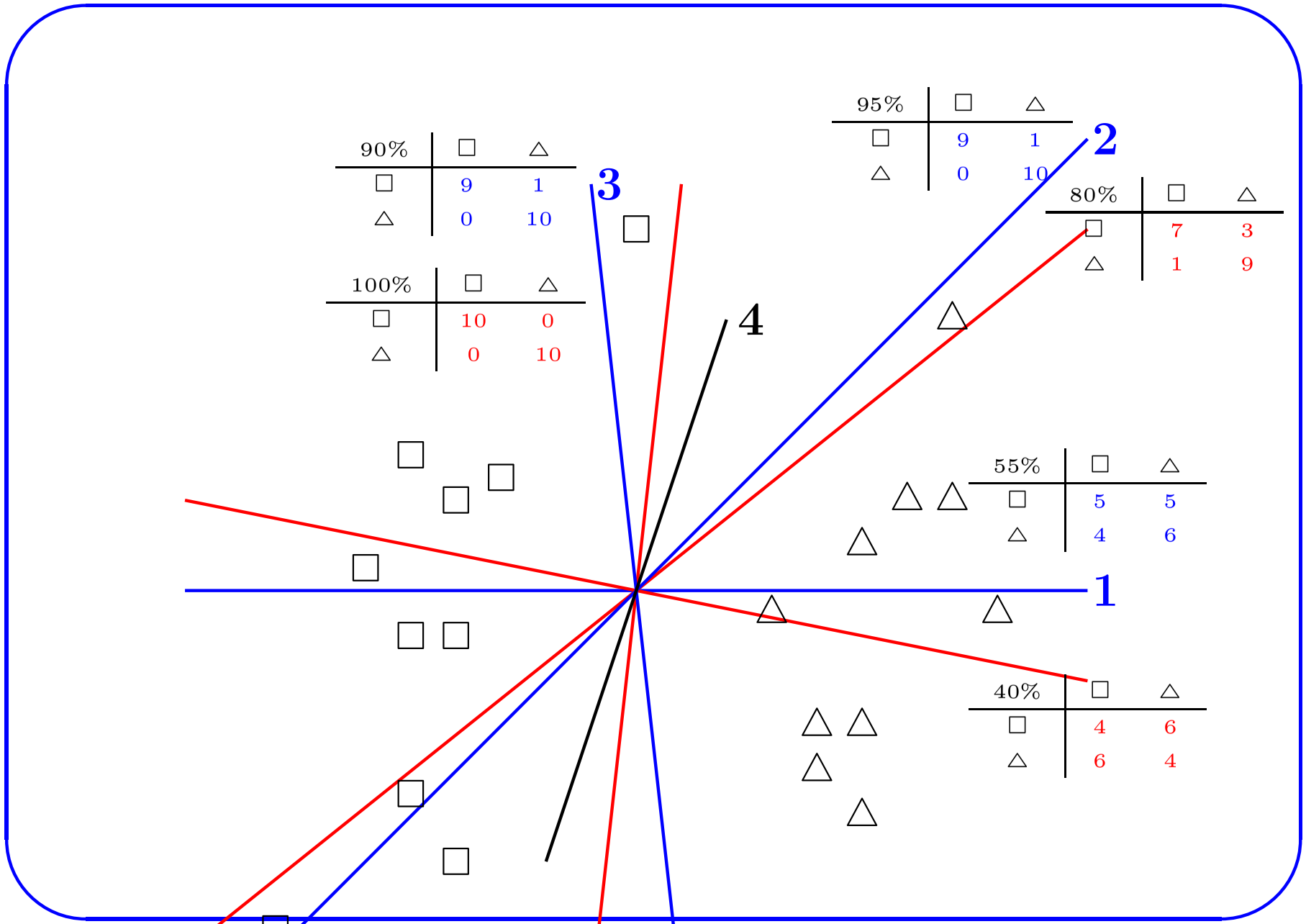
- apprentissage par coeur et conditionnement
- apprentissage par renforcement ou *punition-récompense*
- apprentissage de concept
- résolution de problème
- généralisation et spécialisation
- Le biais de simplicité ou *rasoir d'Occam*

Quelques exemples : le chien de Pavlov, l'apprentissage de la relation "même" par les abeilles, le singe et les bananes.








90%
□
△
□
9
1
△
0
10
95%
□
△
□
9
1
△
0
10
100%
□
△
□
10
0
△
0
10
80%
□
△
□
7
3
△
1
9
55%
□
△
□
5
5
△
4
6
40%
□
△
□
4
6
△
6
4
1
2
3
4

1.(c) : La simplicité : un exemple instructif

Problème Quelle est le chiffre a qui prolonge la séquence :

$$1\ 2\ 3\ 5\ \ \ldots\ a$$

Solution(s) Quelques réponses valides :

- $a = 6$. Argument : c'est la suite des entiers sauf 4.
- $a = 7$. Argument : c'est la suite des nombres premiers.
- $a = 8$. Argument : c'est la suite de Fibonacci.
- $a = 2\pi$ (a peut être *n'importe quel nombre réel* supérieur ou égal 5). Argument : La séquence présentée est la liste ordonnée des racines du polynôme :

  $$P = x^5 - (11+a)x^4 + (41+11a)x^3 - (61-41a)x^2 + (30+61a)x - 30$$

  qui est le développement de

  $$(x-1).(x-2).(x-3).(x-5).(x-a)$$

Généralisation

Il est facile de démontrer ainsi que n'importe quel nombre est une prolongation correcte de n'importe quelle suite de nombres.

Un autre exemple instructif

A	B	C
1	2	3
C	A	B
2	3	1
B	C	?
3	1	?

Question. Quel est le couple de valeurs manquantes que vous mettriez ? Pourquoi ?

A 1	B 2	C 3
C 2	A 3	B 1
B 3	C 1	**A** **2**

Mais on peut construire d'autres carrés latins, par exemple :

A 1	B 2	C 3	D 4	E 5	F 6
C 2	A 3	B 1	F 5	D 6	E 4
B 3	C 1	**D** **4**	E 6	F 2	A 5
D 4	F 5	E 6	A 2	B 3	C 1
E 5	D 6	F 2	B 1	A 4	C 3
F 6	E 4	A 5	C 3	B 1	D 2

Le rasoir d'Occam

La solution la meilleure fait intervenir le moins de concepts.

> "*non sunt multiplicanda entia praeter necessitatem*"

Dans l'exemple mathématique, la solution $a = 8$ est préférable : elle ne nécessite que le concept d'addition.

D'une manière générale, le principe du *rasoir d'Occam* conduit à choisir, pour une valeur explicative égale, la solution la plus simple. On peut ignorer les paramètres qui n'apportent pas d'information.

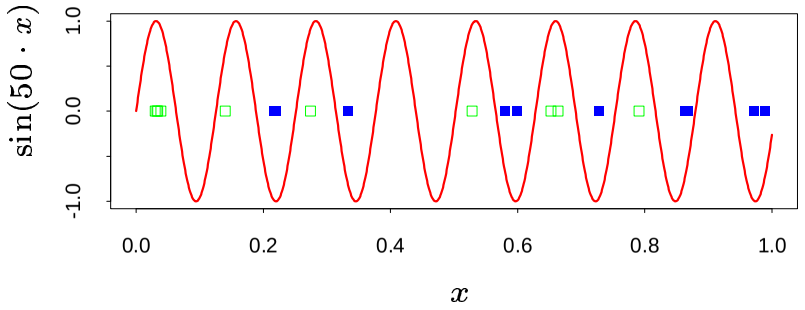


Figure 7.5: *The solid curve is the function* $\sin(50x)$ *for* $x \in [0, 1]$. *The blue (solid) and green (hollow) points illustrate how the associated indicator function* $I(\sin(\alpha x) > 0)$ *can shatter (separate) an arbitrarily large number of points by choosing an appropriately high frequency* α.

La nécéssité d'un biais

Une autre morale peut être tirée de cet exemple : on peut toujours expliquer n'importe quelle solution si se place dans un cadre assez complexe.

On doit donc se fixer une famille de concepts à l'intérieur de laquelle on cherchera la meilleure explication des données.

C'est ce qu'on appelle se donner un *biais d'apprentissage*.

Le compromis simplicité / efficacité devra guider le choix du biais.

2 : Apprentissage automatique

2.(a) : définitions

Un programme possède des capacités d'apprentissage si ses potentialités de comportement sur les données se modifient en fonction de ses performances au fur et à mesure qu'il traite les données.

Un programme possède des capacités d'apprentissage si au cours du traitement d'exemples représentatifs de données il est capable de construire et d'utiliser une représentation de ce traitement en vue de son exploitation.

Quelques notions voisines

- Extraction d'un concept *(extension* $\longrightarrow$ *intension)*
- Catégorisation, classification
- Acquisition de connaissances
- Prédiction
- Généralisation
- Compréhension

Fouille de Données

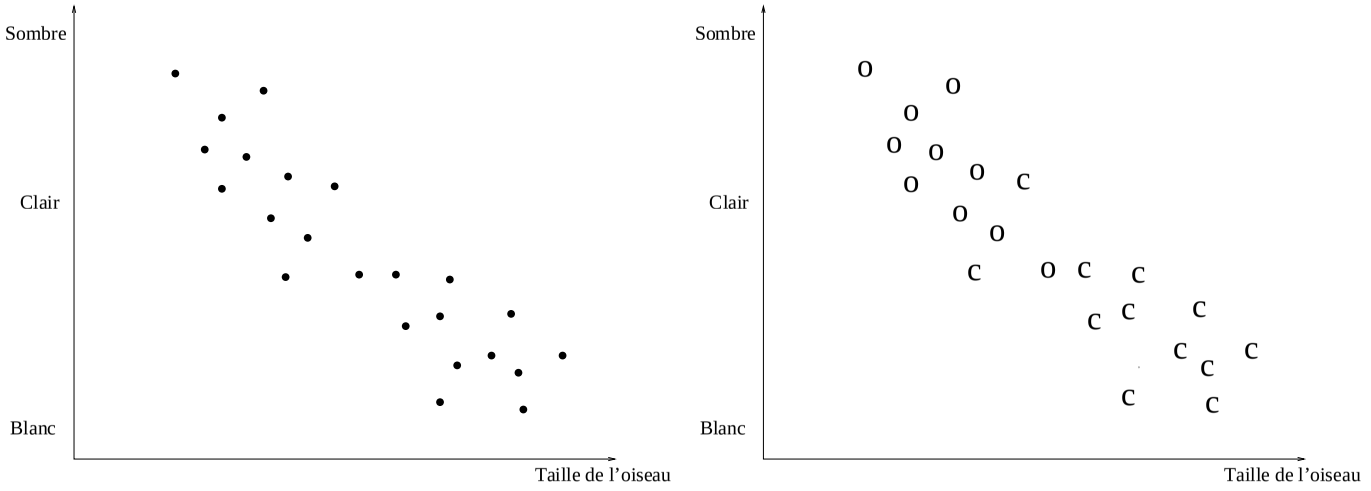
Reconnaissance des Formes

2.(b) : Une taxonomie de base

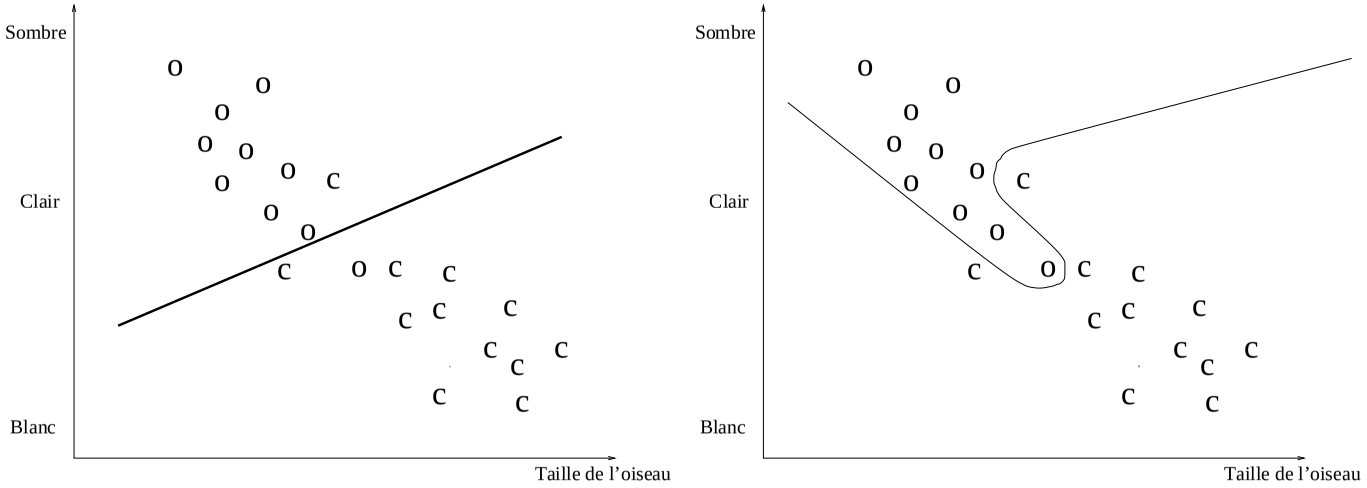
- **Induction**, généralisation
  - Avec professeur : *supervisé*
  - Sans professeur : *non supervisé*
- **Déduction**

  Explication, révision des connaissances

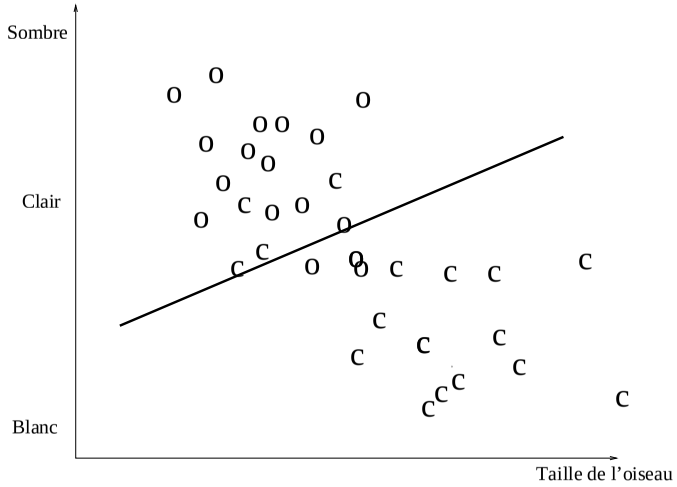
Un exemple ornithologique



Le premier graphique représente les oiseaux observés placés dans l'espace de représentation. Le second graphique représente les mêmes oiseaux, mais il est étiqueté par l'expert. La lettre O signifie que l'oiseau est une oie, C qu'il est un cygne.



Une règle de décision simple et une règle de décision complexe pour séparer les oies des cygnes.



Le test de la règle simple sur d'autres oiseaux.

Un exemple linguistique : l'apprentissage d'une grammaire

Exemples

aab

aaaab

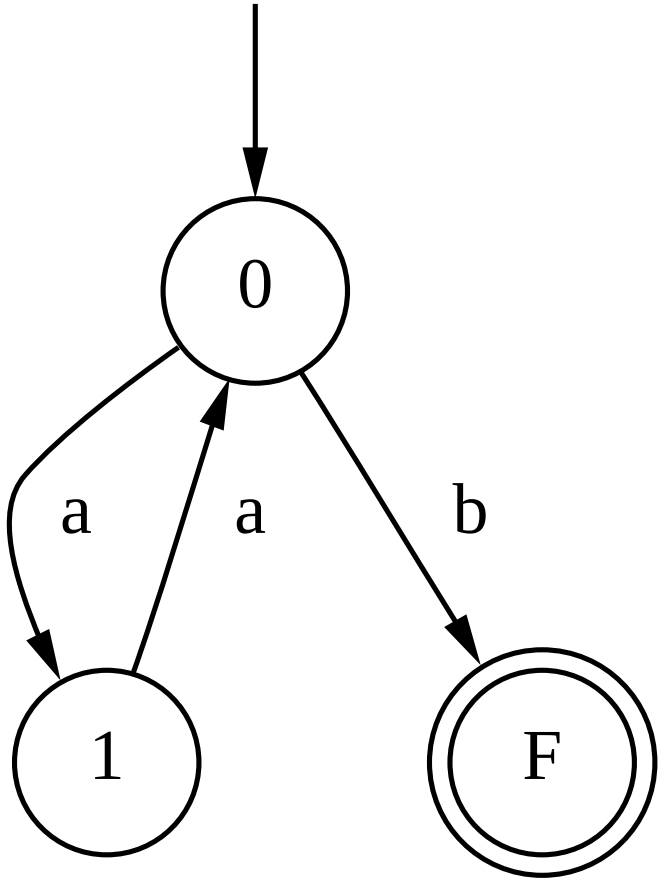
aaaaaaaab

Contre-exemples

aaab

bb

Un concept à apprendre



Un exemple d'extraction de connaissances

Une compagnie d'assurances cherche à lancer un nouveau produit, destiné à couvrir le risque de vol d'objets de valeur à domocile. Elle veut faire une campagne de publicité ciblée auprès d'une partie de ses clients. Cette compagnie ne dispose que de peu de produits du même type et par conséquent sa base de données ne comporte qu'une petite proportion d'enregistrements où un client est déjà associé à une assurance contre le vol à domicile. De plus, comme ces clients possèdent déjà un produit analogue, ce n'est pas vers eux qu'il faut principalement cibler la campagne. Comment savoir si un client qui n'a pas encore d'assurance de ce type sera intéressé par le nouveau produit ?

Une solution est de chercher un profil commun aux clients qui se sont déjà montrés intéressés par un produit de ce type pour viser parmi tous les clients ceux qui ont un profil analogue. Que sera un tel profil ? Dans la base de données, chaque client est décrit par un certain nombre de champs, que l'on peut supposer binaires. Par exemples : "âge inférieur à trente ans", "possède une maison", "a un ou plusieurs enfants", "vit dans une zone à risque de vol", etc. Certains champs peuvent être non remplis : les clients qui ont seulement une assurance automobile n'ont pas été interrogés à la constitution de leur dossier sur l'existence d'un système d'alarme dans leur appartement.

Une façon de constituer un profil consiste à découvrir des associations dans les données, c'est-à-dire des implications logiques approximatives. Disons par exemple que la plupart des clients qui possèdent déjà une assurance contre le vol d'objets de valeur à domicile sont plutôt âgés et n'ont en général qu'une voiture, mais haut de gamme. Il semble raisonnable de démarcher parmi tous les clients ceux qui répondent au même profil. L'hypothèse est donc que posséder une seule voiture (mais de luxe) et être d'âge mûr est un profil qui implique sans doute la possession à domicile d'objets de valeur.

2.(c) : Les paramètres d'un problème d'apprentissage

- L'objectif
- Le protocole
- Le critère de succès
- La nature des entrées *(l'espace de représentation)*
- La nature du résultat *(l'espace des fonctions cibles)*

L'objectif

- Acquisition de connaissances
  - Apprentissage de concepts (*attributs-valeur*)
  - Classification
- Amélioration des performances
  - Adaptation
  - Amélioration en ligne

Le protocole

- Supervisé *vs.* non supervisé
  - Supervision globale *vs.* punition-récompense
  - Interaction avec un oracle
- Présentation des exemples
  - Tous à la fois
  - Un par un, selon un certain tirage

Le critère de succès

- Comportemental :

  Mesure du taux d'erreur en classification (reconnaissance des formes)

  Convergence

- Intrinsèque

  Intelligibilité

  Explication

Les entrées

- Qualité
  - Bruit
  - Origine, représentativité
- Nature
  - Numériques
  - Symboliques (binaires, nominales, séquencielles, logiques, etc.)
  - Mixtes

Les résultats

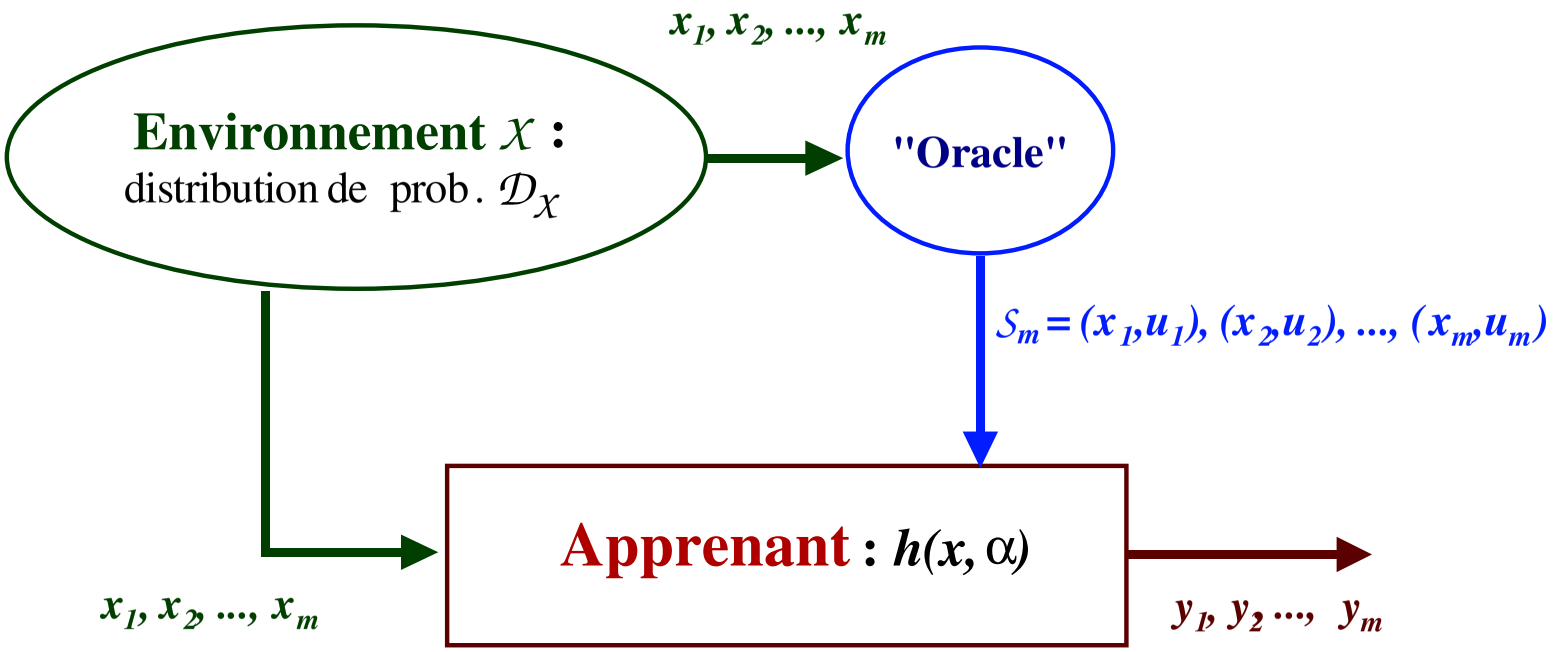
- Décision dans l'espace de représentation des entrées
  - Plus proches voisins (*lazy learning*)
  - Apprentissage par analogie
- Optimisation des paramètres d'une fonction ou d'un algorithme
  - Hyperplans, réseaux connexionnistes
  - HMM
- Optimisation de la structure **et** des paramètres
  - Réseaux bayésiens
- Construction d'un concept (*eager learning*)
  - Arbres de décision
  - Programmation Logique Inductive

2. (d) : L'induction et ses compromis

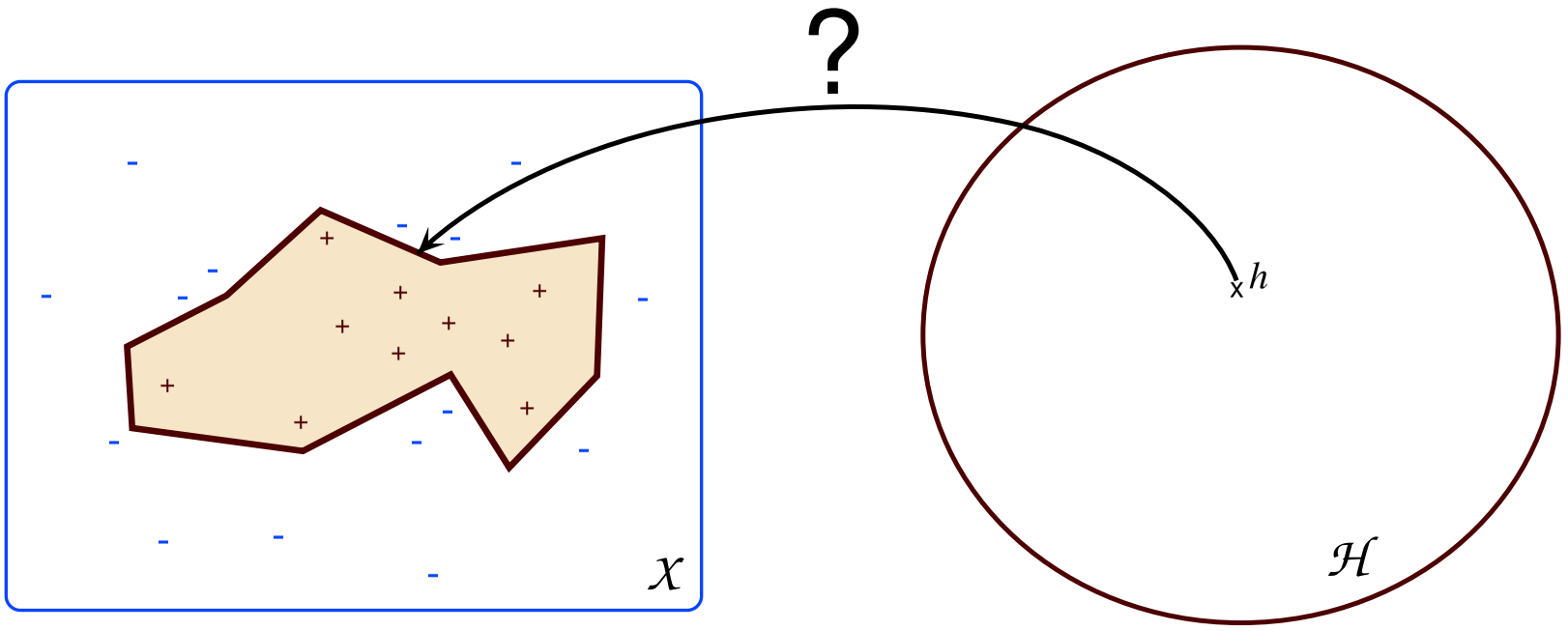
Les sources d'erreur en apprentissage par généralisation sont de trois types :

- Les données peuvent être bruitées, fausses, mal étiquetées
- L'espace $\mathcal{H}$ où l'on cherche une hypothèse est trop restreint
- L'algorithme de recherche dans $\mathcal{H}$ ne fonctionne pas bien

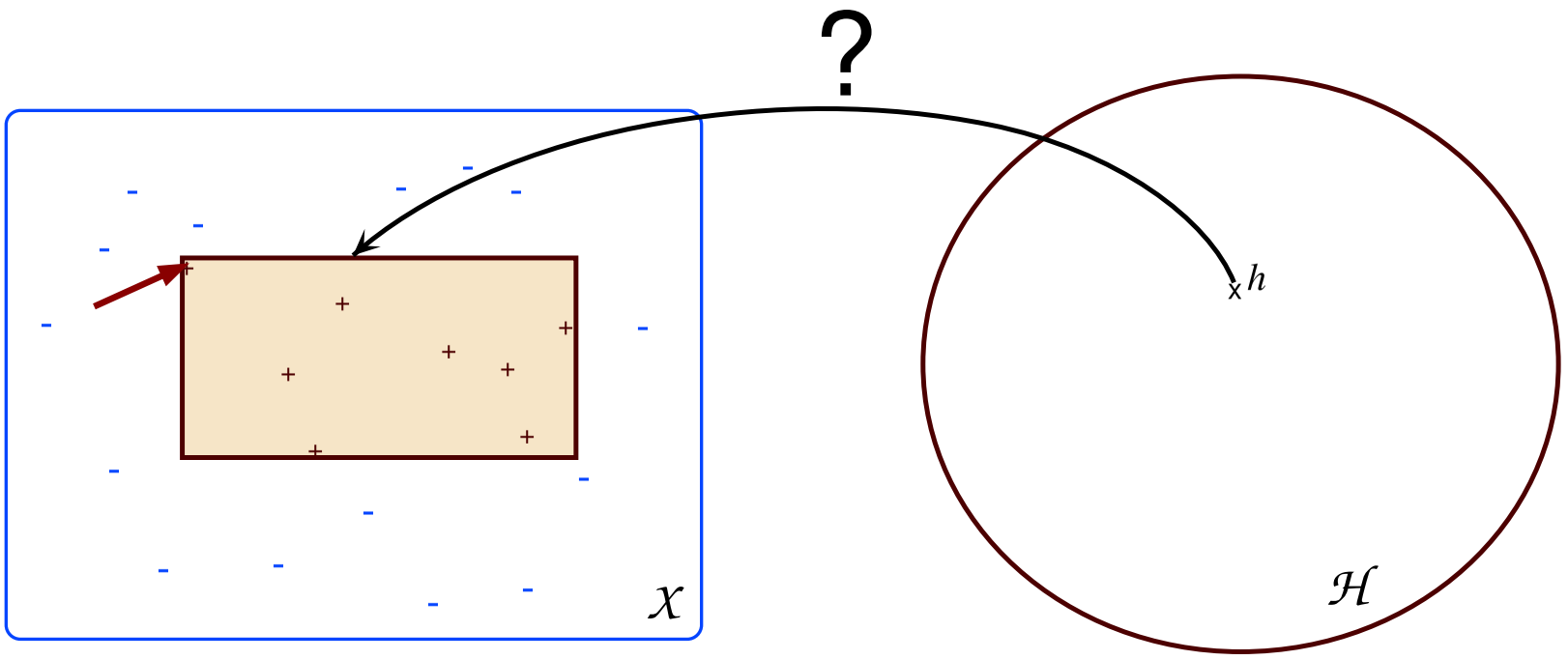
Le scénario de l'induction



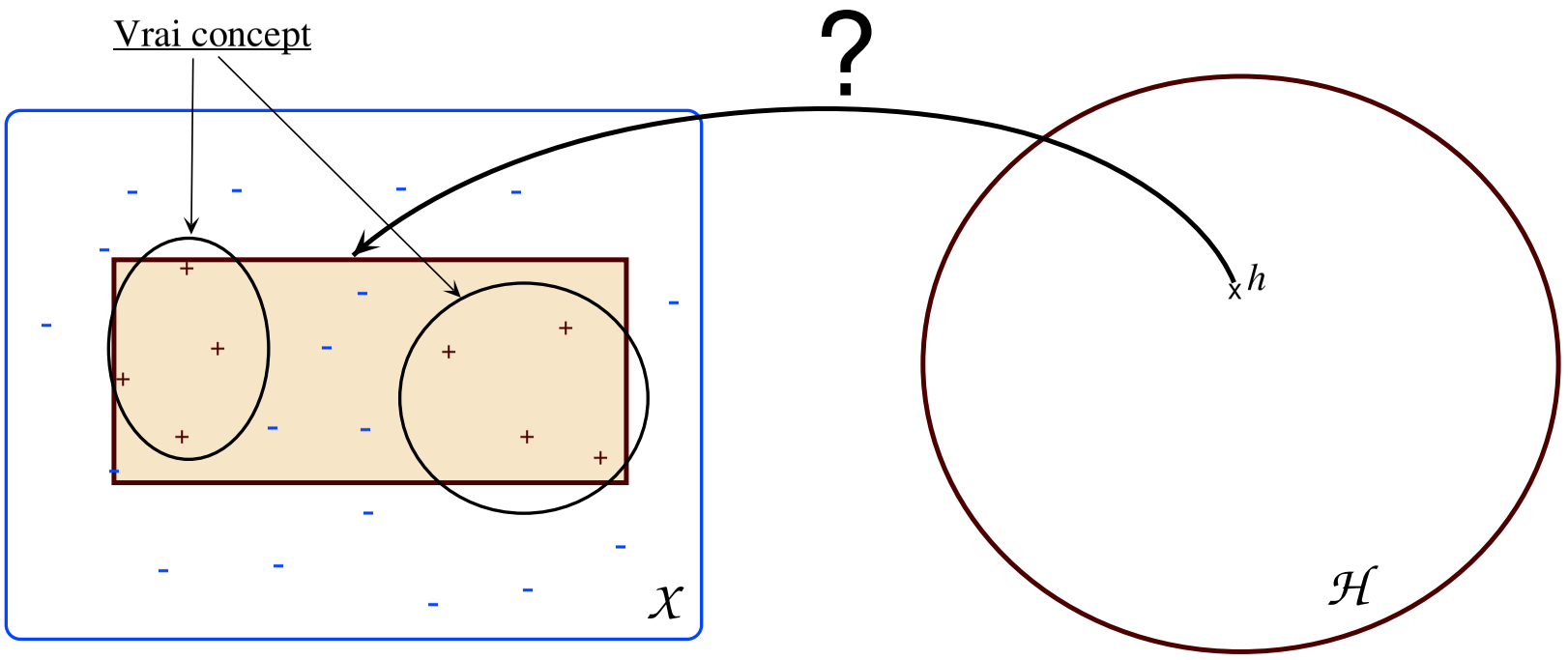
Introduction d'un espace d'hypothèses $\mathcal{H}$



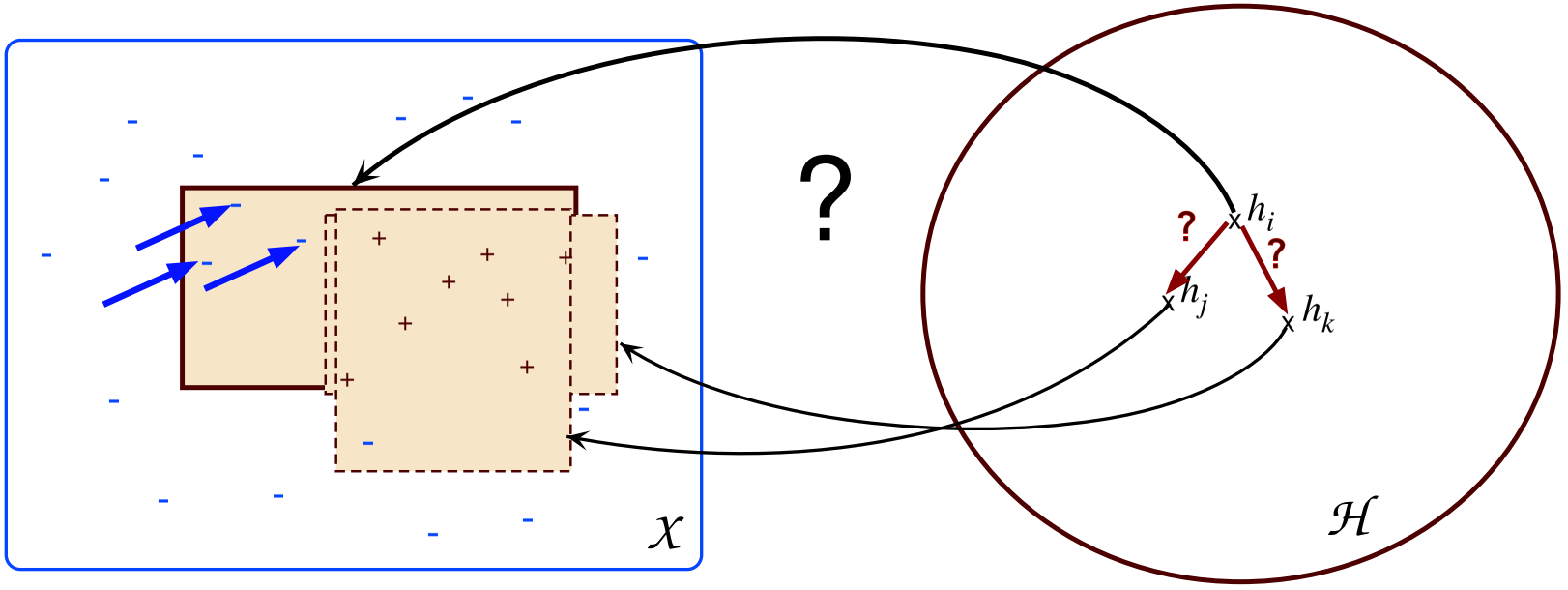
Le langage des hypothèses et la généralisation



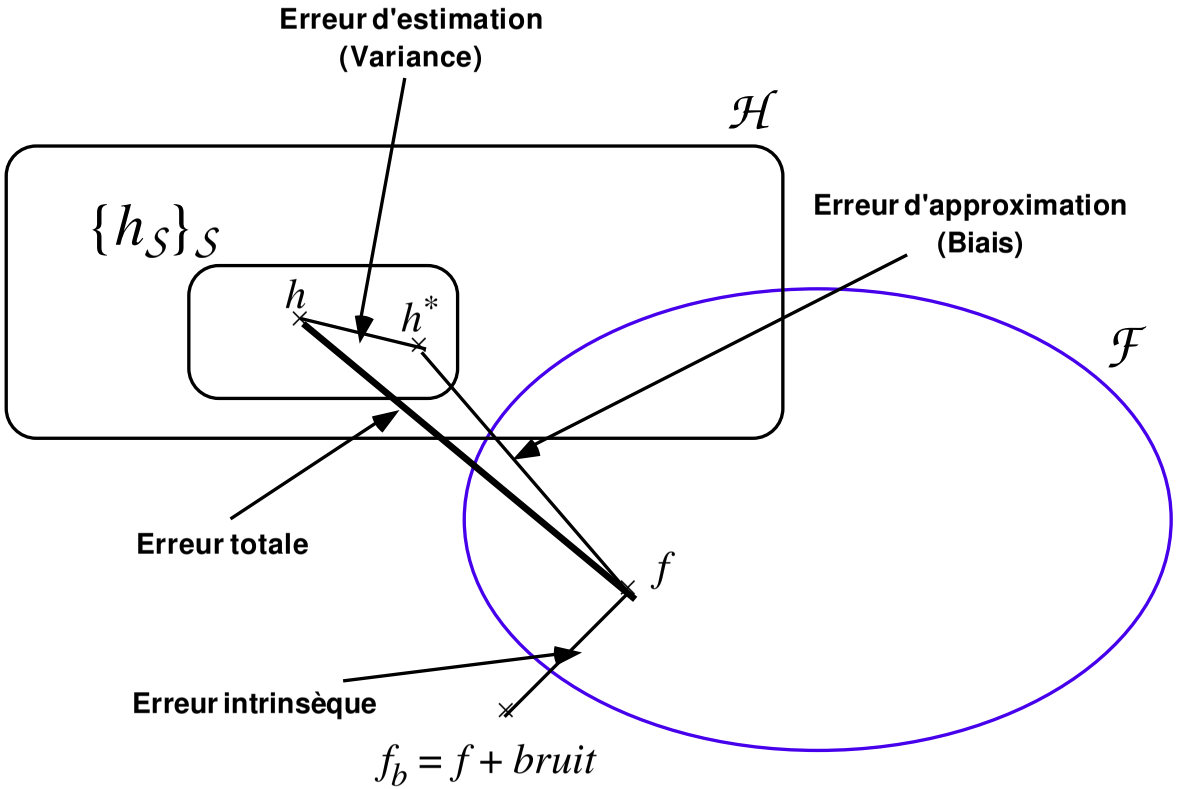
Le langage des hypothèses et la généralisation



L'exploration de l'espace des hypothèses



Les différents types d'erreurs en apprentissage



Le compromis "biais / variance"

Quand $\mathcal{H}$ est restreint :

- La meilleure solution dans $\mathcal{H}$ est facile à trouver
- Mais elle peut être éloignée de la vraie solution

Quand $\mathcal{H}$ est large :

- La meilleure solution dans $\mathcal{H}$ est difficile à trouver
- C'est dommage, car elle est sans doute plus proche de la vraie solution

De plus, $\mathcal{H}$ peut être **trop** large !

Un exemple

On cherche une formule magique pour distinguer les hommes des femmes :

$$a_0 + \Sigma_{i=1}^{d} a_i x_i$$

Les x_i sont des mesures, et la formule est positive si le sujet est une femme, négative sinon.

On dispose de 100 exemples d'hommes et de femmes avec les mesures associées.

Si on prend une seule mesure, *la taille*, un algorithme d'optimisation donnera une valeur très précise à a_0 et a_{taille}. Mais le critère n'est pas très efficace.

Si on prend 50 mesures, *la taille*, *l'âge*, *la longueur des cheveux*, etc... on va trouver une formule qui sépare en effet les exemples, mais...

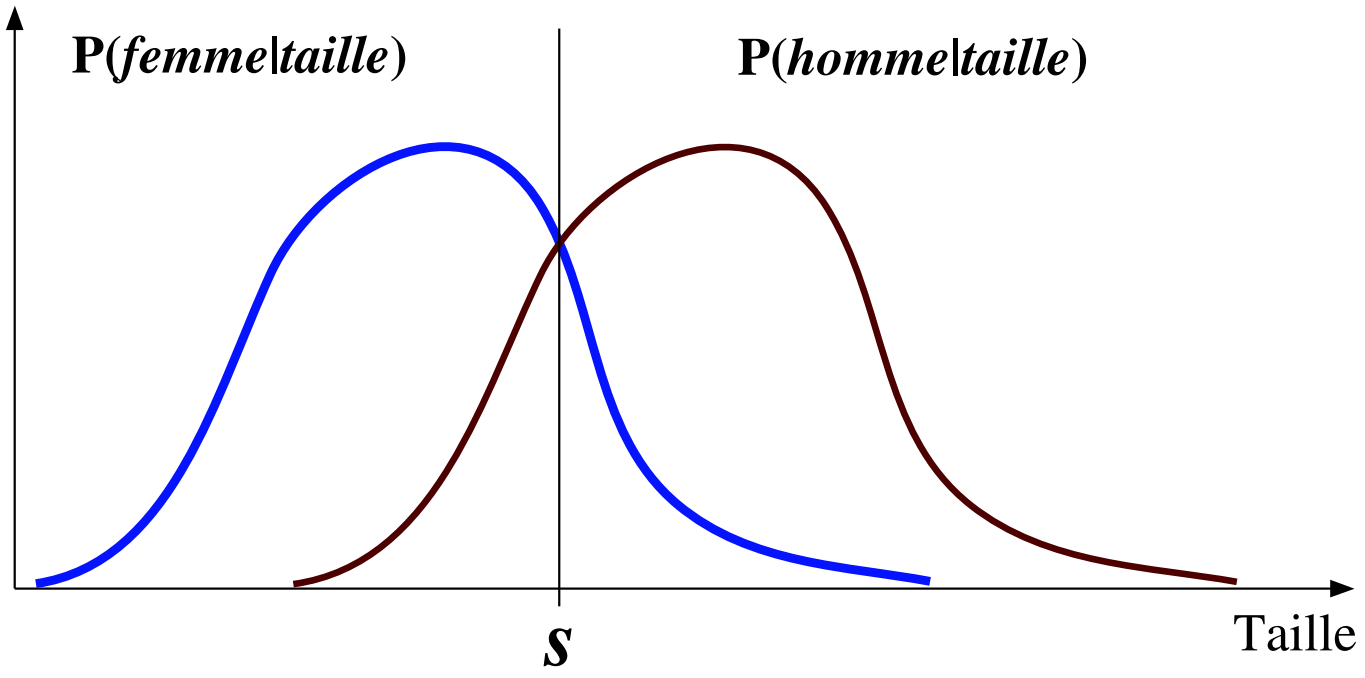
Un exemple (suite)

Pourquoi les deux formules sont-elles si peu magiques ?

- La première est exacte : le seuil trouvé sur *la taille* est excellent, mais cette mesure ne suffit pas à caractériser le sexe.
- La seconde est fausse : trop peu de données d'apprentissage ne permettent pas de fixer les valeurs a_i avec précision.

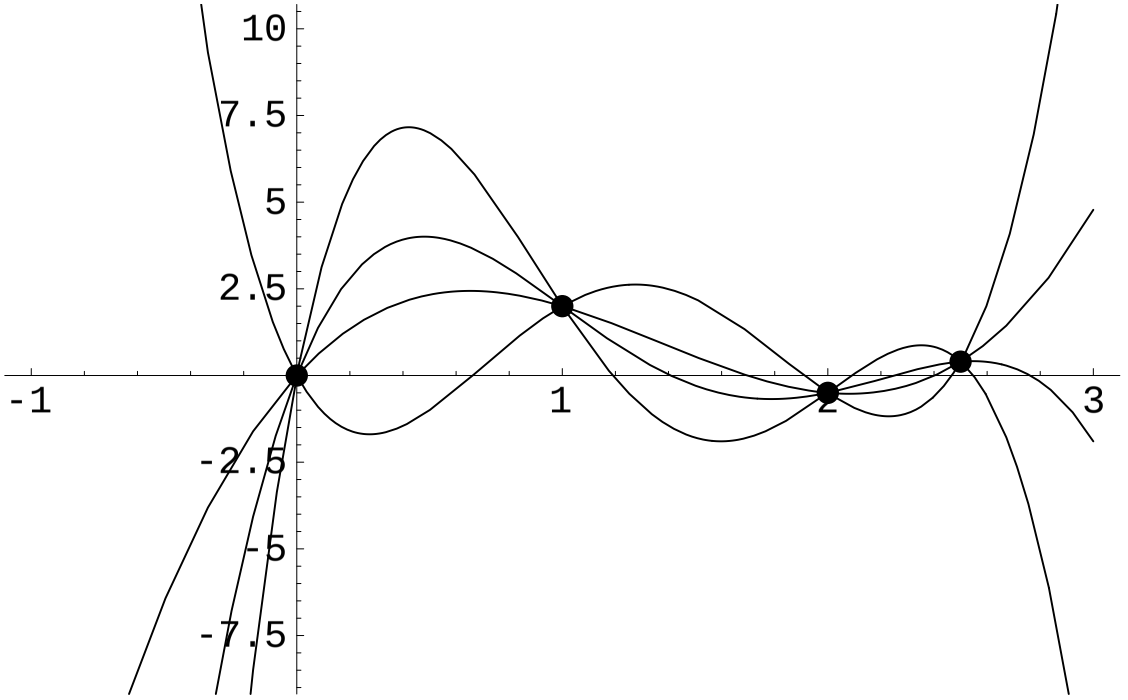
Dans les deux cas, un classement d'un nouvel individu par la formule trouvée doit se lire avec une forte probabilité d'erreur.

Il faut donc trouver un **compromis** efficace...



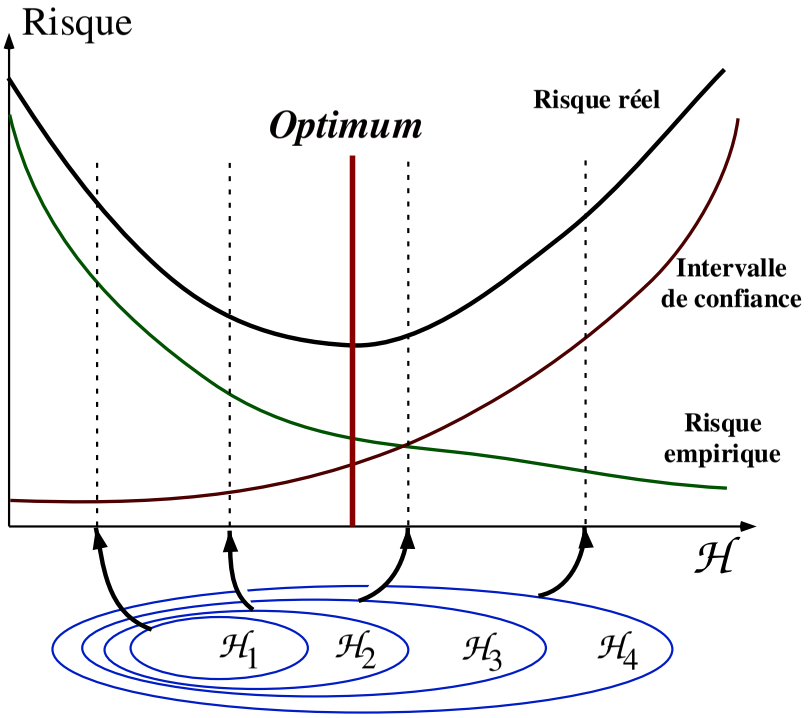
Densités de probabilité des femmes et des hommes en fonction de la taille. Décision en comparant la taille mesurée à un seuil s

Un autre exemple



Quel polynôme doit-on choisir pour interpoler des données ?

Richesse des concepts appris



2.(e) : Une liste raisonnée de concepts et d'algorithmes

Outils

- Définir formellement l'induction.
- Insister sur l'approche bayésienne.
- Décrire les méthodes d'exploration et d'optimisation classiques.
- Donner les moyens statistiques de mesurer la qualité d'un apprentissage. (*théorie pac*)

Apprentissage par exploration

L'espace des Versions

Entrées : données mixtes.

Sortie : un ensemble de concepts syntaxiques et ordonnés.

Utilité : essentiellement méthodologique

La Programmation logique inductive

Entrées : exemples et contre-exemples.

Sortie : un programme PROLOG.

Utilité : grandissante.

Reformulation des connaissances

Une base de règles et un exemple $\longrightarrow$ une base de règles modifié

Inférence grammaticale

Entrées : exemples et contre-exemples de séquences symboliques

Sortie : une grammaire régulière (probabiliste)

Utilité : grandissante

Apprentissage par optimisation

Surfaces discriminantes linéaires et SVM

Outils classiques et robustes.

Réseaux connexionnistes

No comment.

Arbres de décision

Entrées : données mixtes.

Sortie : Un ensemble classificateur de concepts logiques.

Utilité : très employé.

Réseaux bayésiens

Entrées : observations d'attributs.

Sortie : Un réseau de raisonnement probabiliste

Utilité : Original et efficace. Apprentissage difficile.

HMM

Entrées : séquences de vecteurs de $\mathbb{R}^d$.

Sortie : Une probabilité

Utilité : Très bien maîtrisé. Efficace.

Algorithmes et programmation génétiques

Entrées : exemples

Sortie : Un programme LISP

Utilité : Applications variées. Métaphore efficace.

Apprentissage par approximation

Approximation de la décision bayésienne

Robuste

Classification non supervisée et découverte automatique

Couverture de l'espace des exemples

Apprentissage de concepts (en particulier logiques) par approximation successive.

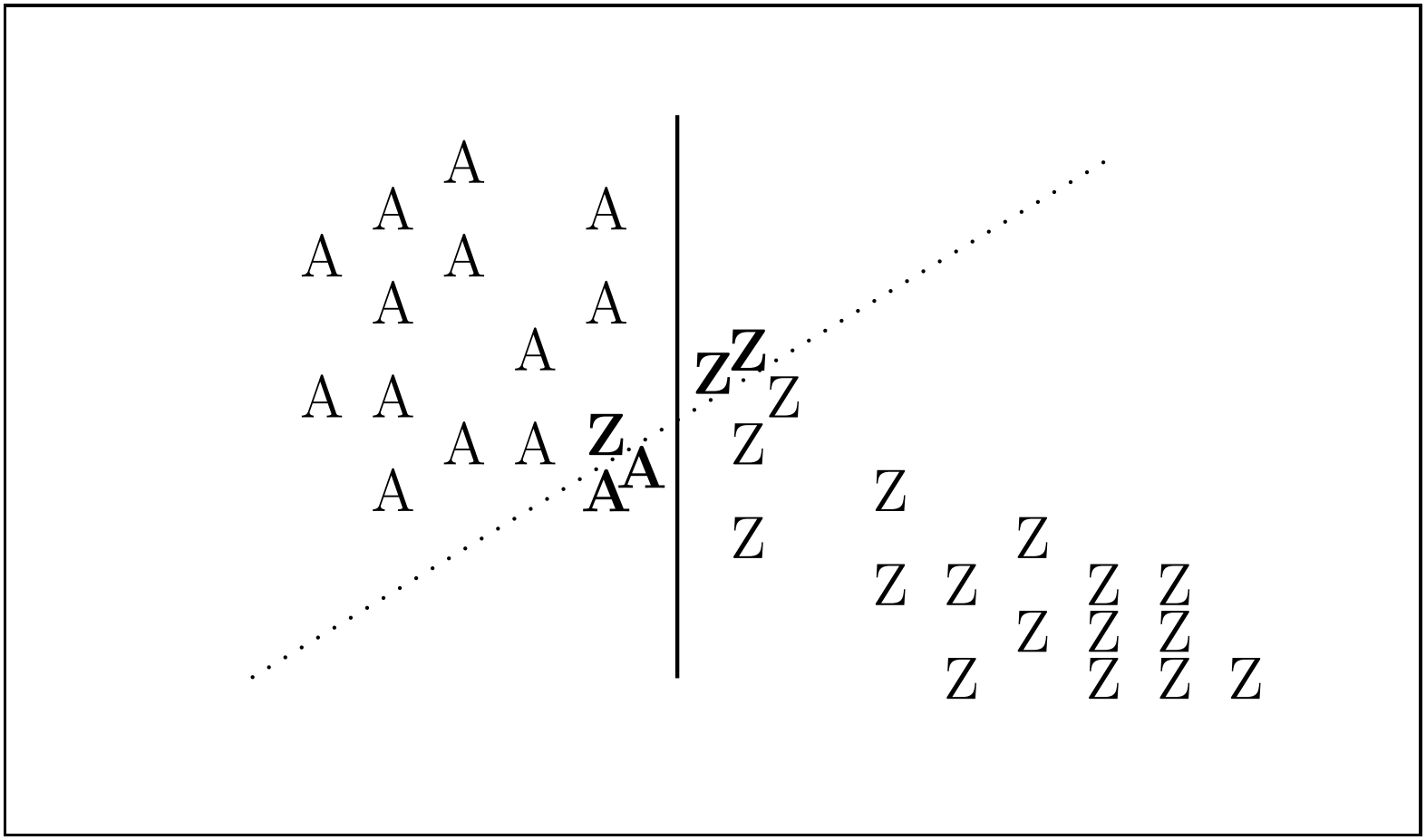
Méthodes de renforcement

Avatar de la commande optimale. Robotique.

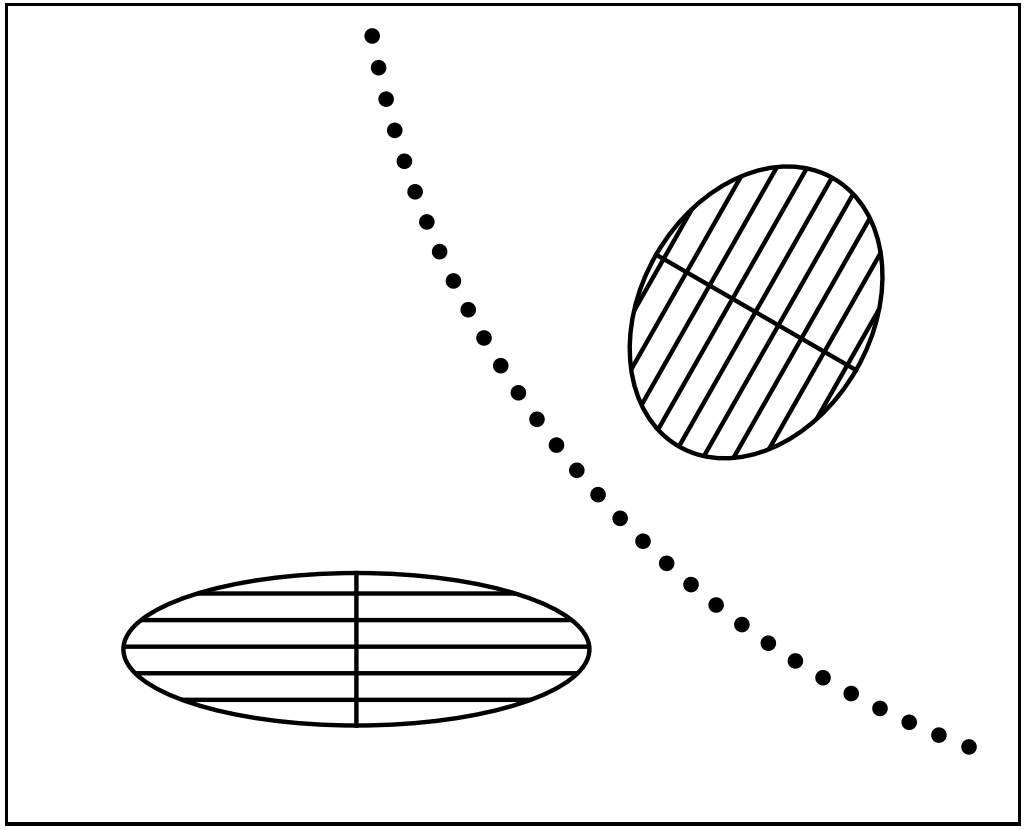
Comités d'experts

Mélanger des méthodes.

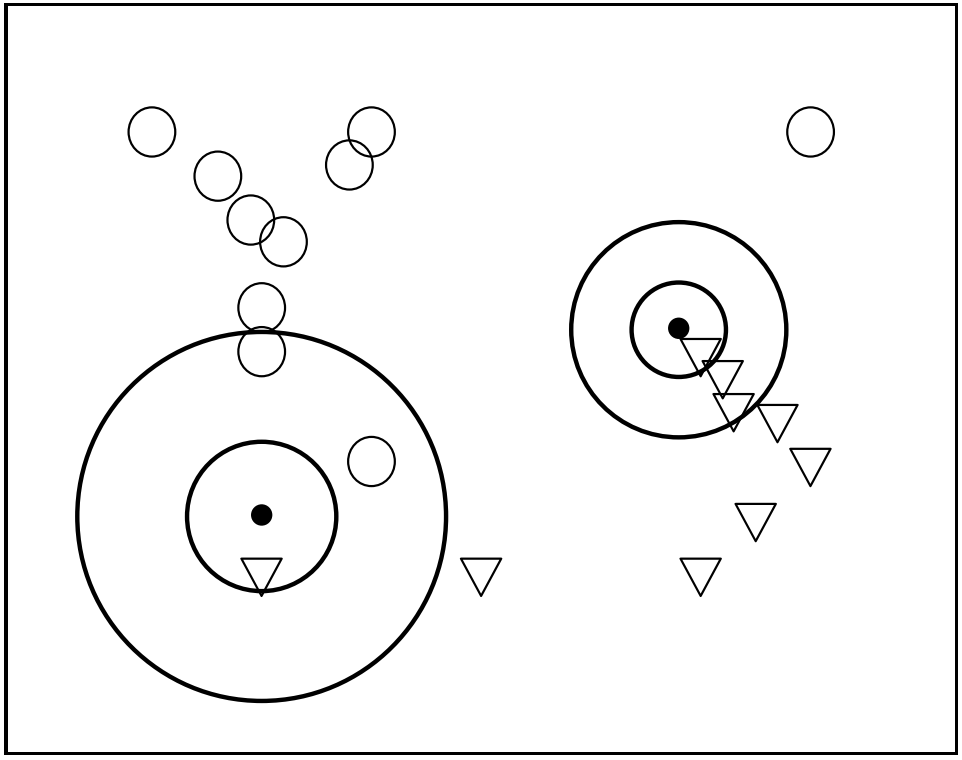
Règle de classification : séparatrice linéaire



Décision bayésienne : modélisation paramétrique des classes



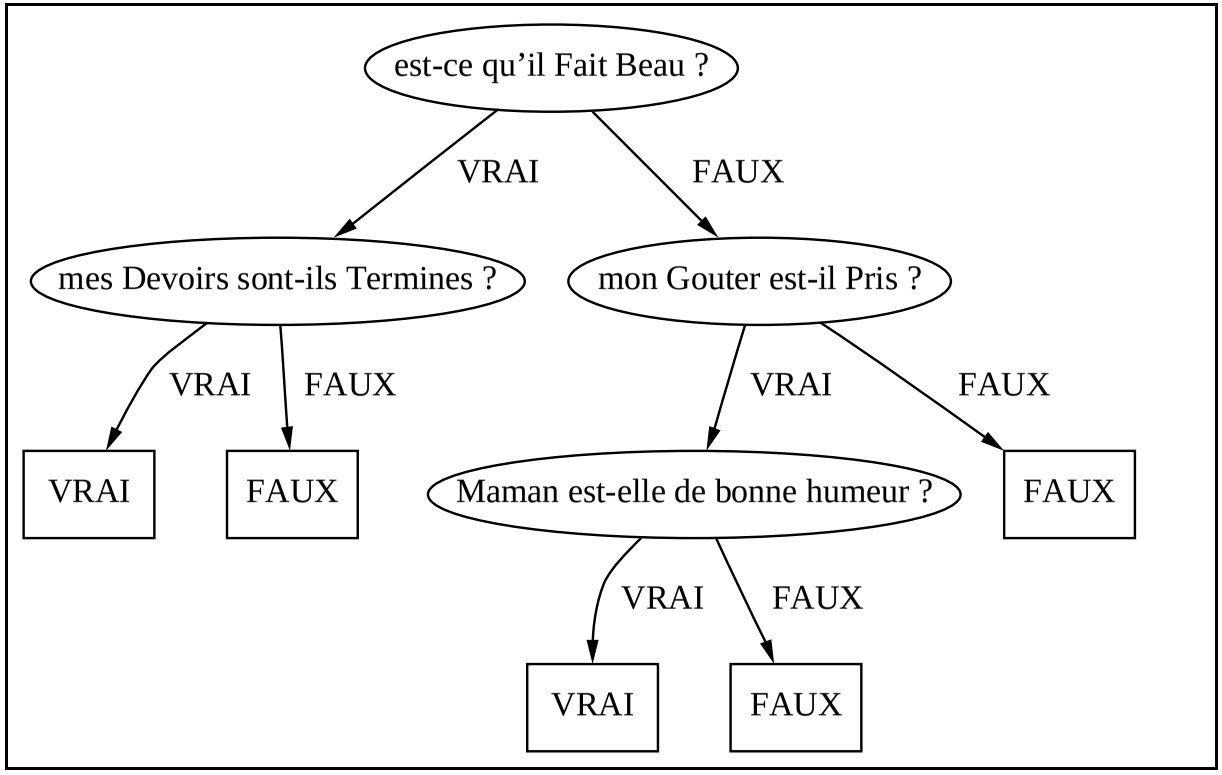
Décision bayésienne : plus proche voisins



Arbres de décision

	Devoirs?	Maman BH?	Fait Beau?	Goûter?	**DECISION**
1	VRAI	FAUX	VRAI	FAUX	**VRAI**
2	FAUX	VRAI	FAUX	VRAI	**VRAI**
3	VRAI	VRAI	VRAI	FAUX	**VRAI**
4	VRAI	FAUX	VRAI	VRAI	**VRAI**
5	FAUX	VRAI	VRAI	VRAI	**FAUX**
6	FAUX	VRAI	FAUX	FAUX	**FAUX**
7	VRAI	FAUX	FAUX	VRAI	**FAUX**
8	VRAI	VRAI	FAUX	FAUX	**FAUX**

Arbres de décision



Réseaux bayésiens pour le diagnostic sur un robot

B la batterie est chargée

L le bloc à soulever n'a pas d'autre bloc posé sur lui (il est *libre*)

M le bras du robot bouge

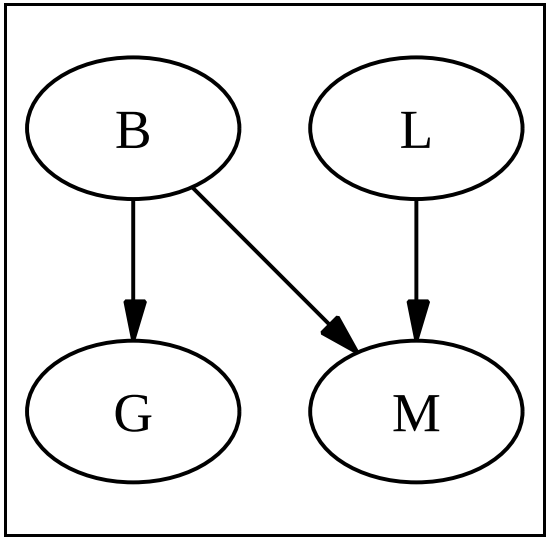
G la jauge de la batterie indique un chargement complet de celle-ci

Réseaux bayésiens : données d'apprentissage

G	M	B	L	Nombre d'exemples
$VRAI$	$VRAI$	$VRAI$	$VRAI$	54
$VRAI$	$VRAI$	$VRAI$	$FAUX$	1
$VRAI$	$FAUX$	$VRAI$	$VRAI$	7
$VRAI$	$FAUX$	$VRAI$	$FAUX$	27
$FAUX$	$VRAI$	$VRAI$	$VRAI$	3
$FAUX$	$FAUX$	$VRAI$	$FAUX$	2
$FAUX$	$VRAI$	$FAUX$	$VRAI$	4
$FAUX$	$FAUX$	$FAUX$	$FAUX$	2
				100

Réseaux bayésiens : graphe et probabilités conditionnelles

$P(B)$	0.95
$P(L)$	0.7
$P(G \mid B)$	0.95
$P(G \mid \neg B)$	0.1
$P(M \mid B,L)$	0.9
$P(M \mid B,\neg L)$	0.05
$P(M \mid \neg B,L)$	0
$P(M \mid \neg B,\neg L)$	0



Bibliographie.

- T. Mitchell *Machine Learning* McGraw-Hill, 1997.
- S. Russel, P. Norvig *Artificial Intelligence : A Modern Approach* Prentice-Hall, 1995.
- N. Nilsson *Artificial Intelligence : A New Synthesis* Morgan-Kaufmann, 1998.
- R. Duda, P. Hart, R. Stork *Pattern Classification* Wiley, 2000.
- A. Cornuéjols, L. Miclet *L'apprentissage artificiel* Eyrolles, 2002.